

519.6 (075)  
Б30

---

---

Н. С. БАХВАЛОВ  
Н. П. ЖИДКОВ  
Г. М. КОБЕЛЬКОВ

# ЧИСЛЕННЫЕ МЕТОДЫ



Н. С. БАХВАЛОВ  
Н. П. ЖИДКОВ  
Г. М. КОБЕЛЬКОВ

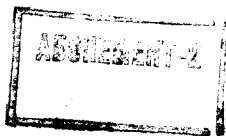
530

---

---

# ЧИСЛЕННЫЕ МЕТОДЫ

*Допущено Министерством высшего  
и среднего специального образования СССР  
в качестве учебного пособия  
для студентов высших учебных заведений*



МОСКВА «НАУКА»  
ГЛАВНАЯ РЕДАКЦИЯ  
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ  
1987

ББК 22.19  
Б30  
УДК 519.6(075.8)

Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы: Учеб. пособие. — М.: Наука. Гл. ред. физ.-мат. лит., 1987. — 600 с.

Написана на основе программы курса численных методов для математических специальностей университетов, отработанной в содружестве педагогических коллективов Московского и Братиславского (ЧССР) университетов.

Методически последовательное изложение численных методов решения задач на ЭВМ содержит, как правило, их строгое теоретическое обоснование. Наряду с этим подробно определена степень практической применимости методов при решении различных классов прикладных задач. Существенное внимание уделено проблеме отбора и оптимизации алгоритмов, закладываемых в основу стандартных программ решения задач.

Для студентов математических факультетов университетов, факультетов прикладной математики, а также для аспирантов и научных работников, связанных с решением задач на ЭВМ.

Табл. 2. Ил. 70.

Рецензенты:

кафедра вычислительной математики Ленинградского государственного университета;

кандидат физико-математических наук В. Г. Сушко

448 210  
НТБ ВНТУ  
м. Вінниця

Б  $\frac{1702070000-140}{053(02)-87}$  54-87

© Издательство «Наука».  
Главная редакция  
физико-математической литературы,  
1987

# ОГЛАВЛЕНИЕ

Предисловие . . . . .	7
Введение . . . . .	9
<b>Глава I. Погрешность результата численного решения задачи . . . . .</b>	<b>17</b>
§ 1. Источники и классификация погрешности . . . . .	17
§ 2. Запись чисел в ЭВМ . . . . .	21
§ 3. Абсолютная и относительная погрешности. Формы записи данных . . . . .	22
§ 4. О вычислительной погрешности . . . . .	24
§ 5. Погрешность функции . . . . .	26
§ 6. Обратная задача . . . . .	31
<b>Глава II. Интерполяция и численное дифференцирование . . . . .</b>	<b>34</b>
§ 1. Постановка задачи приближения функций . . . . .	35
§ 2. Интерполяционный многочлен Лагранжа . . . . .	38
§ 3. Оценка остаточного члена интерполяционного многочлена Лагранжа . . . . .	42
§ 4. Разделенные разности и их свойства . . . . .	42
§ 5. Интерполяционная формула Ньютона с разделенными разностями . . . . .	44
§ 6. Разделенные разности и интерполирование с кратными узлами . . . . .	47
§ 7. Уравнения в конечных разностях . . . . .	50
§ 8. Многочлены Чебышева . . . . .	56
§ 9. Минимизация оценки остаточного члена интерполяционной формулы . . . . .	60
§ 10. Конечные разности . . . . .	63
§ 11. Интерполяционные формулы для таблиц с постоянным шагом . . . . .	66
§ 12. Составление таблиц . . . . .	69
§ 13. О погрешности округления при интерполяции . . . . .	71
§ 14. Применения аппарата интерполирования. Обратная интерполяция . . . . .	72
§ 15. Численное дифференцирование . . . . .	73
§ 16. О вычислительной погрешности формул численного дифференцирования . . . . .	80
§ 17. Рациональная интерполяция . . . . .	82
<b>Глава III. Численное интегрирование . . . . .</b>	<b>83</b>
§ 1. Простейшие квадратурные формулы. Метод неопределенных коэффициентов . . . . .	83
§ 2. Оценки погрешности квадратуры . . . . .	86
§ 3. Квадратурные формулы Ньютона — Котеса . . . . .	91
§ 4. Ортогональные многочлены . . . . .	96
§ 5. Квадратурные формулы Гаусса . . . . .	102
§ 6. Практическая оценка погрешности элементарных квадратурных формул . . . . .	109
§ 7. Интегрирование быстро осциллирующих функций . . . . .	113
§ 8. Повышение точности интегрирования за счет разбиения отрезка на равные части . . . . .	116

§ 9. О постановках задач оптимизации . . . . .	120
§ 10. Постановка задачи оптимизации квадратур . . . . .	125
§ 11. Оптимизация распределения узлов квадратурной формулы . . . . .	127
§ 12. Примеры оптимизации распределения узлов . . . . .	133
§ 13. Главный член погрешности . . . . .	136
§ 14. Правило Рунге практической оценки погрешности . . . . .	139
§ 15. Уточнение результата интерполяцией более высокого порядка точности . . . . .	144
§ 16. Вычисление интегралов в нерегулярном случае . . . . .	146
§ 17. Принципы построения стандартных программ с автоматическим выбором шага . . . . .	153
<b>Глава IV. Приближение функций и смежные вопросы . . . . .</b>	<b>160</b>
§ 1. Наилучшие приближения в линейном нормированном пространстве . . . . .	160
§ 2. Наилучшее приближение в гильбертовом пространстве и вопросы, возникающие при его практическом построении . . . . .	162
§ 3. Тригонометрическая интерполяция. Дискретное преобразование Фурье . . . . .	166
§ 4. Быстрое преобразование Фурье . . . . .	170
§ 5. Наилучшее равномерное приближение . . . . .	173
§ 6. Примеры наилучшего равномерного приближения . . . . .	176
§ 7. Итерационный метод построения многочлена наилучшего равномерного приближения . . . . .	182
§ 8. О форме записи многочлена . . . . .	184
§ 9. О способах вычисления элементарных функций . . . . .	188
§ 10. О скорости приближения функций различных классов . . . . .	194
§ 11. Интерполяция и приближение сплайнами . . . . .	191
§ 12. Энтропия и $\epsilon$ -энтропия . . . . .	204
<b>Глава V. Многомерные задачи . . . . .</b>	<b>208</b>
§ 1. Метод неопределенных коэффициентов . . . . .	209
§ 2. Метод наименьших квадратов и регуляризация . . . . .	210
§ 3. Примеры регуляризации . . . . .	213
§ 4. Сведение многомерных задач к одномерным . . . . .	219
§ 5. Интерполяция функций в треугольнике . . . . .	226
§ 6. Оценка погрешности численного интегрирования на равномерной сетке . . . . .	228
§ 7. Оценка снизу погрешности численного интегрирования . . . . .	231
§ 8. Метод Монте-Карло . . . . .	238
§ 9. Обсуждение правомерности использования недетерминированных методов решения задач . . . . .	242
§ 10. Ускорение сходимости метода Монте-Карло . . . . .	245
§ 11. О выборе метода решения задачи . . . . .	249
<b>Глава VI. Численные методы алгебры . . . . .</b>	<b>254</b>
§ 1. Методы последовательного исключения неизвестных . . . . .	257
§ 2. Метод отражений . . . . .	265
§ 3. Метод простой итерации . . . . .	269
§ 4. Особенности реализации метода простой итерации на ЭВМ . . . . .	271
§ 5. $\delta^2$ -процесс практической оценки погрешности и ускорения сходимости . . . . .	274
§ 6. Оптимизация скорости сходимости итерационных процессов . . . . .	277
§ 7. Метод Зейделя . . . . .	292
§ 8. Метод наискорейшего градиентного спуска . . . . .	296
§ 9. Метод сопряженных градиентов . . . . .	299
§ 10. Итерационные методы с использованием спектрально-эквивалентных операторов . . . . .	299
§ 11. Погрешность приближенного решения системы уравнений и обусловленность матриц. Регуляризация . . . . .	303

§ 12.	Проблема собственных значений . . . . .	309
§ 13.	Решение полной проблемы собственных значений при помощи QR-алгоритма . . . . .	314
<b>Глава VII. Решение систем нелинейных уравнений и задач оптимизации</b>		<b>317</b>
§ 1.	Метод простой итерации и смежные вопросы . . . . .	319
§ 2.	Метод Ньютона решения нелинейных уравнений . . . . .	323
§ 3.	Методы спуска . . . . .	329
§ 4.	Другие методы сведения многомерных задач к задачам меньшей размерности . . . . .	334
§ 5.	Решение стационарных задач путем установления . . . . .	337
§ 6.	Что оптимизировать? . . . . .	345
§ 7.	Как оптимизировать? . . . . .	349
<b>Глава VIII. Численные методы решения задачи Коши для обыкновенных дифференциальных уравнений</b>		<b>357</b>
§ 1.	Решение задачи Коши с помощью формулы Тейлора . . . . .	358
§ 2.	Методы Рунге — Кутты . . . . .	360
§ 3.	Методы с контролем погрешности на шаге . . . . .	365
§ 4.	Оценки погрешности одношаговых методов . . . . .	367
§ 5.	Конечно-разностные методы . . . . .	372
§ 6.	Метод неопределенных коэффициентов . . . . .	375
§ 7.	Исследование свойств конечно-разностных методов на модельных задачах . . . . .	379
§ 8.	Оценка погрешности конечно-разностных методов . . . . .	384
§ 9.	Особенности интегрирования систем уравнений . . . . .	392
§ 10.	Методы численного интегрирования уравнений второго порядка . . . . .	398
<b>Глава IX. Численные методы решения краевых задач для обыкновенных дифференциальных уравнений</b>		<b>403</b>
§ 1.	Простейшие методы решения краевой задачи для уравнения второго порядка . . . . .	403
§ 2.	Функция Грина сеточной краевой задачи . . . . .	409
§ 3.	Решение простейшей краевой сеточной задачи . . . . .	413
§ 4.	Замыкания вычислительных алгоритмов . . . . .	422
§ 5.	Обсуждение постановок краевых задач для линейных систем первого порядка . . . . .	429
§ 6.	Алгоритмы решения краевых задач для систем уравнений первого порядка . . . . .	434
§ 7.	Нелинейные краевые задачи . . . . .	438
§ 8.	Аппроксимации специального типа . . . . .	444
§ 9.	Конечно-разностные методы отыскания собственных значений . . . . .	455
§ 10.	Оптимизация распределения узлов интегрирования . . . . .	458
§ 11.	Построение численных методов с помощью вариационных принципов . . . . .	461
§ 12.	Улучшение сходимости вариационных методов в нерегулярном случае . . . . .	471
§ 13.	Влияние вычислительной погрешности в зависимости от формы записи конечно-разностного уравнения . . . . .	473
<b>Глава X. Методы решения уравнения в частных производных</b>		<b>480</b>
§ 1.	Основные понятия теории метода сеток . . . . .	481
§ 2.	Аппроксимация простейших гиперболических задач . . . . .	490
§ 3.	Принцип замороженных коэффициентов . . . . .	504
§ 4.	Численное решение нелинейных задач с разрывными решениями . . . . .	508
§ 5.	Разностные схемы для одномерного параболического уравнения . . . . .	511
§ 6.	Разностная аппроксимация эллиптических уравнений . . . . .	526

§ 7. Решение параболических уравнений с несколькими пространственными переменными . . . . .	548
§ 8. Методы решения сеточных эллиптических уравнений . . . . .	561
<b>Глава XI. Численные методы решения интегральных уравнений . . . . .</b>	<b>571</b>
§ 1. Решение интегральных уравнений методом замены интеграла квадратурной суммой . . . . .	571
§ 2. Решение интегральных уравнений с помощью замены ядра на вырожденное . . . . .	576
§ 3. Интегральные уравнения Фредгольма первого рода . . . . .	580
<b>Заключение . . . . .</b>	<b>588</b>
<b>Список литературы . . . . .</b>	<b>593</b>
<b>Предметный указатель . . . . .</b>	<b>596</b>

## ПРЕДИСЛОВИЕ

Настоящая книга написана на основе лекций, читавшихся авторами на механико-математическом факультете и на факультете вычислительной математики и кибернетики МГУ. Она возникла в результате переработки и расширения учебного пособия «Численные методы», написанного одним из авторов.

Как и всякое другое пособие по численным методам, эта книга содержит изложение основных положений теории, в данном случае относящихся к приближению функций, интегрированию, задачам алгебры и оптимизации, решению дифференциальных и интегральных уравнений.

Текущий период характерен бурным развитием вычислительной техники; в частности, мощность парка ЭВМ возрастает на порядки в течение десятилетия. Это обстоятельство сопровождается усиленным применением ЭВМ и численных методов для решения резко расширяющегося круга задач. В результате происходит быстрое изменение взглядов на весь комплекс вопросов, связанных с применением ЭВМ, и требований к численным методам решения задач. В частности, в связи с вышесказанным нельзя предложить пособия по численным методам, содержащего рецепты решения всех возможных задач. При выборе метода решения сложной конкретной задачи всякое пособие зачастую играет роль лишь общего руководства, отталкиваясь от которого исследователь рассматривает свои проблемы.

Работа, связанная с созданием и применением численных методов, складывается из чисто теоретических исследований методов решения типичных задач, анализа работы алгоритмов при решении модельных задач, вычислительного эксперимента и ряда других моментов. Существенным моментом при построении и анализе методов решения задач является построение вычислительной модели — простейшей задачи рассматриваемого класса, на примере которой можно провести теоретическое и экспериментальное исследование свойств метода. Большое значение имеют вопросы выбора направления исследования, построения математических моделей рассматриваемых явлений, организации контактов с представителями других наук.

В каждом конкретном случае эти проблемы приобретают свои оттенки и трудно дать какие-либо общие рекомендации, пригодные во всех случаях. Поэтому рассуждения авторов по



затрагиваемым вопросам следует рассматривать не как руководство к действию, а просто как одну из возможных точек зрения.

Эффект от применения ЭВМ и численных методов достигается как в результате решения отдельных трудных задач, так и за счет расширения круга пользователей ЭВМ путем предоставления им бóльших удобств. Для последней цели важно составление стандартных программ решения типичных математических задач и состоящих из них пакетов и библиотек. Поэтому в рассматриваемой книге определенное внимание уделено также вопросам теории численных методов, возникающим при разработке таких стандартных программ.

Авторы признательны Т. Н. Галишниковой, И. П. Мысовских, В. Г. Сушко, Е. Г. Дьяконову, В. Л. Мирошниченко за полезные обсуждения, способствовавшие улучшению рукописи.

## ВВЕДЕНИЕ

Попытаемся определить место теории численных методов в системе других областей знаний и рассказать о проблемах, возникающих в связи с ее применением, прежде чем переходить к непосредственному ее изложению.

Математика как наука возникла в связи с необходимостью решения практических задач: измерений на местности, навигации и т. д. Вследствие этого математика была численной математикой, ее целью являлось получение решения в виде числа.

Численное решение прикладных задач всегда интересовало математиков. Крупнейшие представители прошлого сочетали в своих исследованиях изучение явлений природы, получение их математического описания, как иногда говорят, математической модели явления, и его исследование. Анализ усложненных моделей потребовал создания специальных, обычно численных методов решения задач. Названия некоторых из таких методов — методы Ньютона, Эйлера, Лобачевского, Гаусса, Чебышева, Эрмита — свидетельствуют о том, что их разработкой занимались крупнейшие ученые своего времени.

Настоящее время характерно резким расширением приложений математики, во многом связанным с созданием и развитием средств вычислительной техники. В результате появления ЭВМ с программным управлением менее чем за сорок лет скорость выполнения арифметических операций возросла от 0,1 операции в секунду при ручном счете до  $10^7$  операций на современных серийных ЭВМ, т. е. примерно в  $10^8$  раз.

Рост возможностей в связи с созданием вычислительной техники носит качественный характер и иногда сравнивается с промышленной революцией, вызванной изобретением паровой машины. Уместно вспомнить, что в итоге промышленной революции и последующего на протяжении двух веков развития науки и техники скорость передвижения возросла от скорости пешехода 6 км/ч до скорости космонавта 30 000 км/ч, т. е. в 5000 раз.

Распространенное мнение о всемогуществе современных ЭВМ часто порождает впечатление, что математики избавились почти от всех хлопот, связанных с численным решением задач, и разработка новых методов для их решения уже не столь существенна. В действительности дело обстоит иначе, поскольку потребности эволюции, как правило, ставят перед наукой задачи, находящиеся на грани ее возможностей. Расширение возможностей

приложения математики обусловило математизацию различных разделов науки: химии, экономики, биологии, геологии, географии, психологии, медицины, конкретных разделов техники и др. Процесс математизации состоит в построении математических моделей процессов и явлений и в разработке методов их исследования.

В физике или механике, например, построение математических моделей для описания различных явлений природы и изучение этих моделей с целью объяснения старых или предсказания новых эффектов явлений являются традиционными.

Однако в целом работа в этом направлении зачастую продвигалась относительно медленно, поскольку обычно не удавалось получить решение возникающих математических задач и приходилось ограничиваться рассмотрением простейших моделей. Применение ЭВМ и расширение математического образования резко увеличило возможности построения и исследования математических моделей. Все чаще результаты расчетов позволяют обнаруживать и предсказывать ранее никогда не наблюдавшиеся явления; это дает основания говорить о математическом эксперименте. В некоторых направлениях исследования доверие к результатам численных расчетов так велико, что при расхождении между результатами расчетов и экспериментов в первую очередь ищут погрешность в результатах экспериментов.

Современные успехи в решении таких важных для общества проблем, как атомные, космические, экономические, вряд ли были бы возможны без применения ЭВМ и численных методов.

Требование численного решения новых задач привело к появлению большого количества новых методов. Наряду с этим последние тридцать-сорок лет происходило интенсивное теоретическое переосмысливание и старых методов, а также систематизация всех методов. Эти теоретические исследования оказывают большую помощь при решении конкретных задач и играют существенную роль в наблюдаемом сейчас широком распространении сферы приложений ЭВМ и математики вообще.

Как уже отмечалось, с помощью современных ЭВМ удалось успешно решить ряд важных научно-технических задач. У непосвященного человека может возникнуть превратное впечатление, что успехи в применении ЭВМ обусловлены только повышением их быстродействия. Реально дело обстоит иначе и сложнее.

Правильнее будет сказать, что достижения в области использования ЭВМ обусловлены сочетанием ряда существенных факторов, без пропорционального развития которых они были бы много скромнее:

- 1) увеличение быстродействия ЭВМ, расширение памяти, совершенствование структуры ЭВМ и технических средств общения с ними;

- 2) разработка программных средств общения с ЭВМ, включающая создание операционных систем, языков программирования, библиотек и пакетов стандартных программ, создание но-

вых программных средств общения с ЭВМ, в частности, в диалоговом режиме;

3) рост понимания процессов и явлений науки, техники, природы и общества и создание их математических моделей;

4) совершенствование методов решения традиционных математических и прикладных задач и создание методов решения новых задач;

5) рост понимания возможностей применения ЭВМ среди руководящих работников, ученых, инженеров и, вообще, широких слоев общества; распространение так называемой компьютерной грамотности; координация усилий специалистов разного профиля по использованию вычислительной техники.

Достижения, перечисленные в пп. 3), 5), позволяют ответить на вопрос, какие задачи следует решать с помощью ЭВМ, и организовать их решение, в пп. 2), 4)— как их решать и в п. 1), 2)— дают для этого технические и программные средства.

Просмотр методов решения сложных прикладных задач показывает, что, как правило, эффект, достигаемый за счет совершенствования численных методов, по порядку сравним с эффектом, достигаемым за счет повышения производительности ЭВМ. Трудно сформулировать критерий, по которому можно было бы оценивать эффект применения новых численных методов, и еще труднее дать его достоверную количественную оценку. Все же, если сказать, что эффект от применения новых численных методов при решении прикладных естественнонаучных задач дает 40 % общего эффекта, достигаемого за счет применения новой вычислительной техники и новых численных методов, то эта оценка не будет завышенной.

Рассмотрим пример, иллюстрирующий это утверждение. Решение дифференциальных уравнений в частных производных сводится к решению систем линейных алгебраических уравнений с матрицей, в каждой строке которой имеется 5—10 ненулевых элементов. Накануне появления ЭВМ такие системы уравнений решали в случае числа неизвестных порядка  $10—10^2$ ; сейчас нередки случаи, когда решаются системы с числом неизвестных порядка  $10^5—10^6$ . В гипотетическом случае решения этих задач на современных ЭВМ методами, известными тридцать лет назад, пришлось бы ограничиться системами уравнений с числом неизвестных порядка  $10^3—10^4$  (при тех же затратах времени ЭВМ). Конечность скорости распространения сигнала — 300 000 км/с — ставит уже сейчас существенное ограничение на возможный рост быстродействия однопроцессорных ЭВМ, поэтому значение дальнейшего развития теории численных методов трудно переоценить.

Быстрое проникновение математики во многие области знания, в частности, объясняется тем, что математические модели и методы их исследования применимы сразу ко многим явлениям, сходным по своей формальной структуре. Часто математическая модель, описывающая какое-либо явление, появляется

при изучении других явлений или при абстрактных математических построениях задолго до конкретного рассмотрения данного явления. В частности, и в теории численных методов, так же как в «чистой» математике, полезна разработка общих построений. Однако есть разница в подходе «чистого» и «прикладного» математика к решению какой-либо проблемы. На языке первого понятие «решить задачу» означает доказать существование решения и предложить процесс, сходящийся к решению. Сами по себе эти результаты полезны для прикладника, но, кроме этого, ему нужно, чтобы процесс получения приближения не требовал больших затрат, например времени или памяти ЭВМ. Ему важно не только то, что процесс сходится, но и то, как быстро он сходится. При численном решении задач возникают также новые вопросы, связанные с устойчивостью результата относительно возмущений исходных данных и округлений при вычислениях.

Наряду с теорией численных методов период бурного развития переживает и ряд других разделов математики, непосредственно обязанных ЭВМ своим возникновением. Применение численных методов и ЭВМ к решению естественнонаучных задач оказывает влияние и на традиционные разделы математики.

Математика возникла и развивается как часть естествознания, и долгое время ее развитие существенным образом определялось потребностями физики и механики. Требование математизации новых разделов науки неизбежно приводит к обратному влиянию этих разделов на развитие математики и должно существенно изменить лицо самой математики.

Развитие как теоретических, так и прикладных разделов математики в конечном счете определяется потребностями общества и его материальным вкладом в развитие науки, в частности в образование. Несколько десятилетий назад отношение вложений в науку к общим вложениям в народное хозяйство составляло доли процента. Теперь это отношение настолько велико, что его дальнейший существенный рост невозможен. Поэтому происходит перераспределение вложений в различные направления науки. Это обуславливает еще один канал влияния прикладной стороны математики на развитие ее теоретических разделов. Прикладные исследования имеют непосредственную отдачу; это усиливает доверие общества к математике, расширяет понимание ее проблем и как следствие способствует увеличению вложения средств с целью ее развития.

При реальной работе в области приложений математики возникает большое количество осложнений самого различного, зачастую нематематического характера.

Хотя трудно надеяться, что какие-либо теоретические нравования могут заменить собственный опыт работы, попытаемся обратить внимание на некоторые вопросы общего характера, важные для работы в области приложений математики. Проводимая ниже систематизация этих вопросов является довольно

случайной, условной; по-видимому, можно предложить еще добрый десяток подобных классификаций, имеющих не меньшее право на существование.

1. Первостепенное значение имеет выбор направления исследования. Свобода выбора обычно довольно невелика, так как основные контуры направления исследования обычно задаются «извне».

При выборе направления исследования в пределах имеющихся возможностей полезно иметь в виду следующее «правило трех частей», по своему внешнему виду похожее на шутку. Проблемы делятся на: I — легкие, II — трудные, III — очень трудные. Проблемами I заниматься не стоит, они будут решены в ходе событий и без вашего вмешательства, проблемы III вряд ли удастся решить в настоящее время, поэтому стоит обратиться к проблемам II.

2. Нужно уметь сформулировать на языке математики конкретные задачи физики, механики, экономики, инженерные задачи и т. д., т. е. построить математическую модель рассматриваемого явления.

В теоретической науке исследователь, умеющий правильно формулировать, как говорят, ставить, новые задачи, ценится выше, чем исследователь, умеющий решать кем-то поставленные задачи. Еще более возрастает роль таких ученых в прикладной науке.

Начинающий работу математик часто жалуется на трудности контактов с представителями других наук, которые «даже» не могут сформулировать стоящих перед ними задач. Правильное формулирование задачи — это научная проблема, не менее сложная, чем само решение задачи, и не нужно надеяться, что кто-то другой целиком сделает это за вас. При постановке проблемы первостепенное внимание должно быть уделено выяснению цели исследования; принимаемая математическая модель явления не есть что-то однозначное, раз навсегда связанное с этим явлением, а зависит от цели исследования. Прежде чем выписывать дифференциальные уравнения, выбирать метод решения и обращаться к ЭВМ, стоит подумать, а не будут ли бесполезны все результаты вычислений? В то же время надо воспринимать как должное, что большая часть результатов вычислений будет выброшена сразу же после их получения. Дело в том, что производимая работа зачастую носит исследовательский характер и трудно заранее предсказать, что и в какой форме следует получить, на каком пути нужно искать численное решение задачи. Цель исследования и описание проблемы обычно уточняются в процессе контактов представителей конкретных наук или руководства организаций (*заказчиков*) и математиков (*исследователь* или *исполнителей*).

3. Успех в прикладной науке требует широкой математической подготовки, поскольку только такая подготовка может обеспечить приспособляемость к непрерывно меняющимся типам

задач, предъявляемых к решению. Одной из причин необходимости изучения на первый взгляд «бесполезных» для практики разделов математики является достижение более уверенного и более свободного владения «нужными» разделами математики.

При построении и анализе математических моделей привычка математика «докапываться до конца», подвергать все сомнению, обусловленная его строгим математическим образованием, часто не менее важна, чем интуиция и соображения здравого смысла. Типичное для человека с математическим образованием стремление к общности охвата различных явлений часто помогает выделить наиболее существенные черты явления и отбросить второстепенные.

4. Не следует думать, что совершенное знание математики, численных методов и навыки работы с ЭВМ позволяют сразу решить любую прикладную математическую задачу. Во многих случаях требуется «доводка» методов, приспособление их к решению конкретных задач. При этом типична обстановка, когда используются методы, применение которых теоретически не обосновано, или теоретические оценки погрешности численного метода неприемлемы для практического использования вследствие их громоздкости; при выборе метода решения задачи и анализе результатов приходится полагаться на опыт предшествующего решения задач, на интуицию и сравнение с экспериментом и при этом приходится отвечать за достоверность результатов. Поэтому для успеха в работе необходимы развитое неформальное мышление, умение рассуждать по аналогии, дающие основания ручаться за достоверность результата там, где с позиций логики и математики, вообще говоря, ручаться нельзя.

В рассматриваемом вопросе есть и другая сторона. При численном решении конкретных трудных задач, возникающих в других областях знаний, математик зачастую действует как естествоиспытатель, полагаясь во многом лишь на опыт и «правдоподобные» рассуждения. Крайне желательно, чтобы такая эмпирическая работа подкреплялась теоретическими разработками методов, аккуратной проверкой качества методов на контрольных задачах с известным решением или частным сравнением с экспериментом. При длительном продвижении в каком-то направлении без такого подкрепления может теряться перспектива работы, уверенность в правильности получаемых результатов. Известное высказывание, что хороший теоретик может истолковать в желаемом направлении любые результаты как расчетов, так и эксперимента, содержит большую долю истины.

5. После завершения расчетов наступает этап использования результатов вычислений в практической деятельности, или, как часто говорят, этап внедрения результатов. Правильнее будет сказать, что подготовка к использованию результатов начинается уже с анализа постановки задачи и в процессе ее решения и, по существу, все моменты решения задачи и внедрения не-

разрывно связаны между собой; в процессе формулирования задачи и ее решения заказчик и исполнитель взаимно уточняют постановку задачи и тем самым подготавливают почву для ее внедрения. Поскольку математика в сочетании с ЭВМ используется в самых разнообразных областях, то часто приходится иметь дело с заказчиками, не имеющими опыта применения ЭВМ. В процессе контакта с такими «начинающими» заказчиками особенно важно преодолеть их первоначальное недоверие к вторжению математики в их области исследования; результаты вычислений будут использоваться только тогда, когда заказчик осмыслит их со своих позиций и убедится в том, что их действительно можно и нужно использовать. При правильном подходе к взаимным контактам к концу процесса решения задачи «начинающий» заказчик приходит к пониманию, что ЭВМ и математика могут дать ему не все, но довольно много, а «начинающий» математик — к пониманию того, что он дает заказчику кое-что, но далеко не все нужное для реального решения задачи.

Большое значение имеет наглядность, доступность представления заказчику промежуточных и окончательных результатов исследования: таблицы, графики, вывод информации на экран дисплея; нельзя предполагать наличия или требовать от заказчика большего объема знаний, чем это требуется существом дела. Целесообразнее, чтобы биолог использовал свое умение дифференцировать для построения и исследования математической модели, а не для оценки погрешности метода численного интегрирования.

Математик должен принять во внимание образование и психологию людей, применяющих разработанные им методы и программы. Например, простейшая программа численного интегрирования, предназначенная для широкого круга нематематиков, использующих ЭВМ в своих конкретных исследованиях, должна быть рассчитана на человека, потолок математических знаний которого находится на интуитивном понимании того, что интеграл — это площадь. Чтобы не затруднять пользователя, в описании простейших программ даже ничего не говорится о точности результата. Предполагается, что пользователя удовлетворит невысокая точность результата и программа реализуется, например, так, чтобы в большинстве случаев относительная погрешность результата не превосходила 1 %.

6. Существенным моментом в прикладной работе является необходимость получения результатов в установленный срок. Заказчик, для которого проводятся исследования, расчеты, часто ограничен сроком завершения исследований и принятия решения на их основе. Если исследования не будут завершены к сроку, то решение все равно будет принято, но на основе более грубого, эмпирического или просто «волевого» подхода. Потерянное в таком случае доверие со стороны заказчика часто бывает невозможно восстановить.



В такой ситуации лучше найти по возможности удовлетворительное решение задачи, но в срок, чем получить полное решение задачи к тому времени, когда оно станет бесполезным. В частности, поэтому целесообразно начинать исследование новых задач с рассмотрения простейших моделей, применяя при численном решении простейшие испытанные методы.

7. Также существенным моментом в прикладной работе является то обстоятельство, что работа, как правило, проводится коллективом. Одна из причин этого состоит в том, что построение математической модели, выбор метода решения, непосредственное общение с ЭВМ и анализ результатов требуют различных знаний и квалификации. Другая причина кроется в упомянутой уже необходимости решения задачи в установленный срок. Это требование приводит к необходимости распараллеливания даже однотипной работы между большим числом исполнителей, например путем независимого написания различных блоков программы отдельными исполнителями. Параллельно могут идти отработка различных методов на модельных задачах, обсчет упрощенных моделей, подготовительная работа по написанию окончательной программы решения задачи.

Можно привести много реальных примеров неудачного решения больших вычислительных задач и работ по созданию программного обеспечения, вызванных следующей причиной. Распределение обязанностей между параллельно работающими исполнителями не было в достаточной степени формализовано, т. е. не было выдано однозначного описания окончательного результата работы каждого исполнителя. В результате или основная доля времени уходила на непрерывное согласование отдельных частей работы, или после истечения существенного промежутка времени оказывалось, что эти части работы не стыкуются. Поэтому организаторские способности ученого, осуществляющего общее руководство решением задачи, зачастую не менее важны, чем его математические способности.

Приведенные выше рассуждения в определенной степени иллюстрируют специфику работы в области прикладной математики и показывают, что сложность этой работы в комплексе, как правило, превосходит сложность работы в области теоретической математики, требуя кроме широкой математической эрудиции также других важных свойств человеческого интеллекта и характера.

## Глава I

# ПОГРЕШНОСТЬ РЕЗУЛЬТАТА ЧИСЛЕННОГО РЕШЕНИЯ ЗАДАЧИ

В этой главе объясняются источники возникновения погрешности решения задачи, даются основные правила задания приближенных величин и оценивается погрешность как простейших, так и более сложных функций от приближенно заданных величин.

Конкретные оценки этой главы по существу в дальнейшем не используются, но сам разговор о них необходим для понимания реальной обстановки, в которой используются рассматриваемые в дальнейшем методы решения задач.

### § 1. Источники и классификация погрешности

Погрешность решения задачи обуславливается следующими причинами:

- 1) математическое описание задачи является неточным, в частности неточно заданы исходные данные описания;
- 2) применяемый для решения метод часто не является точным: получение точного решения возникающей математической задачи требует неограниченного или неприемлемо большого числа арифметических операций, поэтому вместо точного решения задачи приходится прибегать к приближенному;
- 3) при вводе данных в машину, при выполнении арифметических операций и при выводе данных производится округления.

Погрешности, соответствующие этим причинам, называют:

- 1) *неустранимой погрешностью*,
- 2) *погрешностью метода*,
- 3) *вычислительной погрешностью*.

Часто неустранимую погрешность подразделяют на две части:

- а) *неустранимой погрешностью* называют лишь погрешность, являющуюся следствием неточности задания числовых данных, входящих в математическое описание задачи;
- б) погрешность, являющуюся следствием несоответствия математического описания задачи реальности, называют *погрешностью математической модели*.

Дадим иллюстрацию этих определений. Пусть имеется маятник (рис. 1.1.1)\*, начинающий движение в момент  $t = t_0$ . Требуется предсказать угол отклонения  $\varphi$  от вертикали в момент  $t_1$ .

Дифференциальное уравнение, описывающее колебание этого маятника, берется в виде

$$l \frac{d^2\varphi}{dt^2} + g \sin \varphi + \mu \frac{d\varphi}{dt} = 0, \quad (1)$$

где  $l$  — длина маятника,  $g$  — ускорение силы тяжести,  $\mu$  — коэффициент трения.

Как только принимается такое описание, решение задачи уже приобретает неустранимую погрешность, в частности, потому, что реальное трение зависит от скорости не совсем линейно; другой источник неустранимой погрешности состоит в погрешностях определения  $l$ ,  $g$ ,  $\mu$ ,  $t_0$ ,  $\varphi(t_0)$ ,  $\varphi'(t_0)$ .

Название этой погрешности — «неустранимая» — соответствует ее существу: она неконтролируема в процессе численного решения задачи и может уменьшиться только за счет более точного описания физической задачи и более точного определения параметров.

Дифференциальное уравнение (1) не решается в явном виде; для его решения требуется применить какой-либо численный метод. Вследствие этой причины и возникает погрешность метода.

Вычислительная погрешность может возникнуть, например, из-за конечности количества разрядов чисел, участвующих в вычислениях.

Введем формальные определения.

Пусть  $I$  — точное значение отыскиваемого параметра (в данном случае — реальный угол отклонения маятника  $\varphi$  в момент времени  $t_1$ ),  $\tilde{I}$  — значение этого параметра, соответствующее принятому математическому описанию (в данном случае — значение  $\varphi(t_1)$  решения уравнения (1)),  $\tilde{I}_h$  — решение задачи, получаемое при реализации численного метода в предположении отсутствия округлений,  $\tilde{I}_h^*$  — приближение к решению задачи, получаемое при реальных вычислениях. Тогда

$$\begin{aligned} \rho_1 &= \tilde{I} - I - \text{неустранимая погрешность,} \\ \rho_2 &= \tilde{I}_h - \tilde{I} - \text{погрешность метода,} \\ \rho_3 &= \tilde{I}_h^* - \tilde{I}_h - \text{вычислительная погрешность.} \end{aligned} \quad (2)$$

\*) Тройная нумерация рисунков и формул указывает главу, параграф, номер формулы или рисунка; двойная, применяемая только для формул, — параграф и номер (в данной главе); одинарная, применяемая также только для формул, — только номер (в данном параграфе).

Полная погрешность  $\rho_0 = \tilde{I}_h^* - I$ , равная разности между реально получаемым и точным решениями задачи, удовлетворяет очевидному равенству

$$\rho_0 = \rho_1 + \rho_2 + \rho_3. \quad (3)$$

Во многих случаях под термином *погрешность* того или иного вида удобно понимать не рассмотренные выше разности между приближениями, а некоторые меры близости между ними. Например, в скалярном случае полагают

$$\rho_0 = |I_h^* - I|, \quad \rho_1 = |\tilde{I} - I|, \quad \rho_2 = |\tilde{I}_h - \tilde{I}|, \quad \rho_3 = |\tilde{I}_h^* - \tilde{I}_h|;$$

при таких обозначениях вместо (3) получаем

$$\rho_0 \leq \rho_1 + \rho_2 + \rho_3. \quad (4)$$

В других случаях решение  $I$  и приближения  $I, I_h, I_h^*$  оказываются элементами некоторых функциональных пространств, часто различных. Например,  $I$  может быть элементом пространства  $F$  непрерывных на  $[0, 1]$  функций, а  $I_h$  — элементом пространства  $F_h$  сеточных функций  $f_h$ , определенных в точках  $x_n = nh, n = 0, 1, \dots, h^{-1}$ ;  $h^{-1}$  — целое. Тогда в качестве меры погрешности вводят некоторую меру близости  $\rho(z_1, z_2)$ , где  $z_1$  и  $z_2$  могут быть элементами как одного, так и различных пространств. Требования на эту меру близости — возможность принять ее за естественную меру погрешности и выполнение неравенства треугольника

$$\rho(z_1, z_3) \leq \rho(z_1, z_2) + \rho(z_2, z_3) \quad (5)$$

при любых  $z_1, z_2, z_3 \in F, F_h$ . При этом не накладывается условие: если  $\rho(z_1, z_2) = 0$ , то  $z_1 \equiv z_2$ ; так что функция  $\rho(z_1, z_2)$  не обязательно является расстоянием в некотором метрическом пространстве.

В рассматриваемом сейчас примере можно положить

$$\rho(f_1, f_2) = \max_{0 \leq n \leq h^{-1}} |f_1(nh) - f_2(nh)|$$

независимо от того, каким пространствам принадлежат  $f_1$  и  $f_2$ .

Может возникнуть такой вопрос по поводу проблемы исследования неустранимой погрешности: зачем изучать неустранимую погрешность решения задачи, если она «неустранима»? По крайней мере такая точка зрения кажется оправданной, если математик получает для численного решения задачи уже готовые уравнения, не участвуя в обсуждении физической постановки задачи.

Это возражение нельзя признать разумным. Часто математик сам занимается исследованием постановки задачи, анализом и упрощением рассматриваемых уравнений. Поскольку все явления в природе взаимосвязаны, в принципе невозможно математически точно описать никакой реальной процесс, происходящий в природе. Однако анализ влияния различных факторов на погрешность решения может позволить получить простейшее описание процесса с допустимой погрешностью. Обычно математик имеет представление о требуемой окончательной точности результата, и, исходя из этого, он может производить необходимые упрощения исходной задачи.

Если математик не участвует в обсуждении физической постановки задачи, то представление о величине неустранимой погрешности ему все равно необходимо по следующей причине. При решении большинства задач нет особого смысла применять метод решения задачи с погрешностью, существенно меньшей, чем величина неустранимой погрешности. Поэтому, имея представление о величине неустранимой погрешности, математик-исследователь может разумно сформулировать требования к точности результата численного решения задачи.

Непомерные требования заказчика к точности результата часто вызваны тем, что он имеет преувеличенные представления о возможностях ЭВМ и поэтому серьезно не продумывает, что все-таки ему нужно.

Такие требования часто снимаются в процессе обсуждения задачи на основе следующих соображений:

- 1) при более детальном подходе к изучению задачи в целом оказывается, что столь высокая точность и не нужна;
- 2а) математическая модель явления настолько груба, что требовать столь высокую точность бессмысленно;
- 2б) параметры модели не могут быть определены с высокой точностью;
- 3) заказчику нужен вообще не количественный, а качественный результат, например такого типа: будет ли работать данное устройство в заданном режиме или нет.

Разберем некоторые встретившиеся нам реальные задачи. К решению была предъявлена система интегральных уравнений с сильно осциллирующими ядрами с числом перемен знака у ядер порядка  $\lambda^{-1} = N = 10^6$ . Требовалось получить решение с относительной погрешностью (определение см. далее) порядка  $10^{-6}$ . Эта система описывала режим работы некоторого оптического устройства. Решение такой системы интегральных уравнений непомерно сложно даже для современных ЭВМ, поэтому был предпринят ее подробный анализ. Оказалось, что относительные погрешности характеристик системы, обусловленные технологией изготовления устройства, являются величинами порядка  $10^{-4}$ , поэтому нет смысла решать задачу со столь высокой точностью, как требовалось вначале. В результате требования к точности искомого решения были снижены до относительной погрешности  $10^{-4}$ . Однако и такая точность все равно еще требовала непомерных затрат машинного времени.

Дальнейший анализ задачи показал, что, по существу, заказчика интересовал ответ только на один вопрос — будет ли данное устройство устойчиво функционировать или нет?

Естественно было предположить две возможности:

1) при малых значениях параметра  $\lambda = 1/N$  решение плавно зависит от этого параметра, поэтому при  $\lambda$ , меньших достаточно малого  $\lambda_0$ , система будет работать в одном режиме — или всегда устойчиво, или всегда неустойчиво;

2) при малых значениях  $\lambda$  решение существенно меняется при изменении этого параметра и интервалы значений параметра  $\lambda$ , где режим работы устойчив, перемежаются с интервалами значений, где режим работы неустойчив.

В связи с этим были предприняты расчеты при довольно крупном значении  $\lambda = 1/10$  с последующим уменьшением значений  $\lambda$  с тем, чтобы понять, какая из двух указанных выше возможностей реализуется. В данном случае реализовывалась первая возможность. В случае второй возможности, вследствие грубости изготовления устройства, вряд ли удалось бы вообще исследовать вопрос об устойчивости работы устройства при  $\lambda = 10^{-6}$ .

Расчеты показывали, что при изменении  $\lambda$  в пределах от  $10^{-1}$  до  $10^{-2}$  имел место устойчивый режим работы устройства и был сделан вывод (подтвержденный потом экспериментально, после конструирования реального устройства) об устойчивости его работы при  $\lambda = 10^{-6}$ .

Другой, на первый взгляд выглядящий курьезным, но на самом деле весьма типичный пример реальной ситуации. Перед математиками была поставлена задача создания алгоритма и программы быстрого (менее чем за 1 с машинного времени) вычисления интегралов специального вида с относительной погрешностью  $10^{-4}$ . Эта задача ими была успешно решена, т. е. был разработан метод вычисления таких интегралов и на его основе создана стандартная программа. В свою очередь исследователи, поставившие задачу (заказчики), не скупясь на затраты машинного времени, для проверки качества предложенного математиками метода и надежности программы сами вычислили приближенно один из таких интегралов с относительной погрешностью, по их мнению,  $10^{-6}$ . Но оказалось, что все попытки решить эту, так называемую *тестовую* задачу с погрешностью, лучшей, чем  $10^{-2}$ , с помощью созданной математиками программы оканчивались неудачей. Возникло предположение о погрешности в самой тестовой задаче. Оказалось, что в ней число  $\pi$  было взято равным 3,14, что вносило в тестовый пример *неустранимую* погрешность, которая, естественно, не могла быть устранена никакими усилиями математиков, создавших алгоритм и программу.

Рассмотрим пример иного рода. При записи интеграла, предъявленного к вычислению, число  $\pi$  для простоты было взято равным 3, и интеграл

$$\int_1^{10} f(x) \frac{\sin x}{x - \pi} dx$$

превратился в расходящийся. Если бы число  $\pi$  было взято с высокой точностью, то при формальном математическом подходе этот интеграл также следовало бы рассматривать как расходящийся; однако при реальном вычислении на ЭВМ искомый интеграл скорее всего был бы вычислен с приемлемой точностью.

## § 2. Запись чисел в ЭВМ

Современные ЭВМ оперируют с числами, записанными в одной из приведенных ниже форм.

Первая форма записи — с фиксированной запятой: все числа в ЭВМ имеют модуль, меньший 1; число знаков после запятой фиксировано. Таким образом, машина оперирует с числами

$$x = \pm \sum_{k=1}^t \alpha_k q^{-k} \equiv \pm (\alpha_1, \dots, \alpha_t); \quad (1)$$

здесь  $q$  — целое — основание системы счисления,  $\alpha_1, \dots, \alpha_t$  — целые в пределах  $0 \leq \alpha_k < q$ .

При операциях над числами  $x$  с  $|x| < 1$  могут появляться числа  $y$  с  $|y| \geq 1$ , и тогда произойдет остановка работы ЭВМ («машинный останов» или «АВОСТ»). Чтобы избежать этого, производится масштабирование задачи — введение новых масштабов. Иногда заранее нельзя указать нужные масштабы; в других случаях введение очень больших масштабов с самого начала приведет к тому, что в исходных данных большое количество первых из  $\alpha_i$  обратится в нуль и произойдет существенная

потеря информации. Поэтому часто предусматривают изменение масштабов уже в процессе решения задачи.

Вторая форма записи, наиболее распространенная в ЭВМ, предназначенных для научных расчетов, — с плавающей запятой: машина оперирует с числами

$$x = \pm q^p \sum_{k=1}^t \alpha_k q^{-k} = \pm q^p (\alpha_1, \dots, \alpha_t); \quad (2)$$

порядок числа  $p$  удовлетворяет неравенству  $|p| \leq p_0$ .

Наиболее распространен случай двоичной системы счисления  $q = 2$ ; случаи  $q = 8$  и  $q = 10$  применяются как вспомогательные на этапах подготовки и выдачи данных. В наиболее совершенных ЭВМ имеются представления чисел как с фиксированной запятой, так и с плавающей; в ряде случаев число разрядов  $t$  может задаваться пользователем по его усмотрению.

При работе в режиме с плавающей запятой пользователь получает дополнительные удобства, так как не надо заботиться о масштабах; однако при этом происходит некоторое замедление работы ЭВМ.

Довольно типичны значения  $q = 2$ ,  $t = 36$ ,  $p_0 = 64$ ; тогда  $q^{-t} = 2^{-36} \approx 1,5 \cdot 10^{-11}$ ,  $2^{p_0} = 2^{64} \approx 7 \cdot 10^{19}$ .

### § 3. Абсолютная и относительная погрешности.

#### Формы записи данных

Если  $a$  — точное значение некоторой величины, а  $a^*$  — известное приближение к нему, то *абсолютной погрешностью* приближенного значения  $a^*$  называют обычно некоторую величину  $\Delta(a^*)$ , про которую известно, что

$$|a^* - a| \leq \Delta(a^*).$$

*Относительной погрешностью* приближенного значения называют некоторую величину  $\delta(a^*)$ , про которую известно, что

$$\left| \frac{a^* - a}{a^*} \right| \leq \delta(a^*).$$

Относительную погрешность часто выражают в процентах.

Если  $a$  — известное число, например  $\pi$ , то иногда говорят об абсолютной  $\Delta(a)$  и относительной  $\delta(a)$  погрешностях задания этого числа. Числа  $\Delta(a)$  и  $\delta(a)$  называют соответственно абсолютной и относительной погрешностью числа  $a$ , если про них известно, что  $|a^* - a| \leq \Delta(a)$ ,  $\left| \frac{a^* - a}{a} \right| \leq \delta(a)$ .

Иногда в литературе абсолютной погрешностью называют величину  $a^* - a$ , а относительной — величину  $\frac{a^* - a}{a^*}$ ; мы будем придерживаться исходных определений, и поэтому у нас всегда  $0 \leq \Delta(a^*)$ ,  $\delta(a^*)$ .

По ходу изложения материала будут употребляться выражения: *большое число, очень большое число, сильный рост функции,*

Чаще всего число  $x$  мы называем *большим*, если  $|x| \gg 1$ , но относительная погрешность результата решения задачи порядка  $|x|2^{-l}$  является допустимой.

Если относительная погрешность порядка  $|x|2^{-l}$  является недопустимо большой, то число  $x$  называем *очень большим*.

Выражение *функция сильно растет* чаще всего означает, что она возрастает в очень большое число раз.

*Значащими цифрами* числа называют все цифры в его записи, начиная с первой ненулевой слева.

Пример. У чисел  $a^* = 0,03045$ ,  $a^* = 0,03045000$  значащими цифрами являются подчеркнутые цифры. Число значащих цифр в первом случае равно 4, во втором — 7.

Значащую цифру называют *верной*, если абсолютная погрешность числа не превосходит единицы разряда, соответствующего этой цифре.

Примеры.  $a^* = 0,03045$ ,  $\Delta(a^*) = 0,000003$ ;  $a^* = 0,03045000$ ,  $\Delta(a^*) = 0,0000007$ ; подчеркнутые цифры являются верными.

Иногда условливаются называть значащую цифру *верной*, если абсолютная погрешность не превосходит половины единиц разряда, соответствующих этой цифре.

Если все значащие цифры верные, то говорят, что число записано *со всеми верными цифрами*.

Пример. При  $a^* = 0,03045$ ,  $\Delta(a^*) = 0,000003$  число  $a^*$  записано со всеми верными цифрами.

Иногда употребляется термин *число верных цифр после запятой*: подсчитывается число цифр после запятой от первой цифры до последней верной цифры. В последнем примере это число равно 5.

Довольно часто информация о некоторой величине задается пределами ее изменения:

$$a_1 \leq a \leq a_2, \quad (1)$$

например

$$1,119 \leq a \leq 1,127.$$

Принято записывать эти пределы с одинаковым числом знаков после запятой; так как обычно достаточно грубого представления о погрешности, то в числах  $a_1$ ,  $a_2$  часто берут столько значащих десятичных цифр, сколько нужно, чтобы разность  $a_2 - a_1$  содержала одну-две значащие цифры.

Многочисленные оговорки «часто», «обычно», «принято» специально употребляются нами, чтобы не создавалось впечатления об обязательности каких-то стандартных форм задания информации о величине погрешности. Эти формы задания информации рассматриваются лишь потому, что они наиболее распространены, а следовательно, наиболее удобны при контактах.

Абсолютную или относительную погрешность обычно записывают в виде числа, содержащего одну или две значащие цифры.



Информацию о том, что  $a^*$  является приближенным значением числа  $a$  с абсолютной погрешностью  $\Delta(a^*)$ , иногда записывают в виде

$$a = a^* \pm \Delta(a^*); \quad (2)$$

числа  $a^*$  и  $\Delta(a^*)$  принято записывать с одинаковым числом знаков после запятой. Например, записи

$$a = 1,123 \pm 0,004,$$

$$a = 1,123 \pm 4 \cdot 10^{-3}$$

относятся к общепринятым и означают, что

$$1,123 - 0,004 \leq a \leq 1,123 + 0,004.$$

Соответственно информацию о том, что  $a^*$  является приближенным значением числа  $a$  с относительной погрешностью  $\delta(a^*)$ , записывают в виде

$$a = a^* (1 \pm \delta(a^*)). \quad (3)$$

Например, записи

$$a = 1,123(1 \pm 0,003), \quad a = 1,123(1 \pm 3 \cdot 10^{-3}), \quad a = 1,123(1 \pm 0,3\%)$$

означают, что

$$(1 - 0,003)1,123 \leq a \leq (1 + 0,003)1,123.$$

При переходе от одной из форм записи к другой надо следить, чтобы пределы изменения, указываемые новой формой записи, были шире старых, иначе такой переход будет незаконным. Например, при переходе от (1) к (2) должны выполняться неравенства

$$a^* - \Delta(a^*) \leq a_1, \quad a_2 \leq a^* + \Delta(a^*),$$

при переходе от (2) к (3) — неравенства

$$a^*(1 - \delta(a^*)) \leq a^* - \Delta(a^*), \quad a^* + \Delta(a^*) \leq a^*(1 + \delta(a^*)),$$

при переходе от (3) к (2) должны выполняться противоположные неравенства (пределы всегда расширяются!).

Следует различать принятую нами выше формально математическую и обиходную терминологию в рассуждениях о величине погрешности. Если в постановке задачи говорится, что требуется найти решение с погрешностью  $10^{-2}$ , то чаще всего не имеется в виду обязательность этого требования. Предполагается лишь, что погрешность имеет такой порядок. Если, например, решение будет найдено с погрешностью  $2 \cdot 10^{-2}$ , то такой результат, скорее всего, также удовлетворит заказчика.

#### § 4. О вычислительной погрешности

Ограничение на порядки чисел в ЭВМ  $|p| \leq p_0$  иногда приводит к прекращению вычислений; в других случаях относительно небольшая разрядность чисел в ЭВМ приводит к недопустимому искажению результата вычислительной погрешностью. Та-

кие алгоритмы, где вследствие ограниченности  $p$  или малости  $t$  возникают подобные эффекты, называют «неустойчивыми».

Построение «устойчивых» алгоритмов, при использовании которых искажение окончательного результата вычислительной погрешностью находится в допустимых пределах, составляют существенную часть теории численных методов.

Рассмотрим пример, показывающий, что повышение точности иногда может быть достигнуто за счет несложного алгебраического преобразования.

Пусть отыскивается наименьший корень уравнения  $y^2 - 140y + 1 = 0$ . Для определенности условимся о следующих правилах округления. Вычисления производятся в десятичной системе счисления, причем в мантиссе числа после округлений удерживается 4 разряда. Имеем

$$y = 70 - \sqrt{4899}, \quad \sqrt{4899} = 69,992 \dots;$$

после округления получаем

$$\sqrt{4899} \approx 69,99, \quad y \approx 70 - 69,99 = 0,01.$$

То же самое значение  $y$  можно, «избавившись от иррациональности в числителе», представить в виде  $y = 1/(70 + \sqrt{4899})$ . Последовательно производя вычисления, получаем  $\sqrt{4899} \approx 69,99$ ;  $70 + 69,99 = 139,99$  и после округления

$$70 + 69,99 \approx 140,0.$$

Наконец,

$$1/140 = 0,00714285 \dots,$$

и после округления

$$y \approx 0,007143.$$

Производя вычисления с дополнительными разрядами, можно проверить, что в обоих случаях все подчеркнутые цифры результатов верные; однако во втором случае точность результата существенно выше. Дело в том, что в первом случае пришлось вычитать близкие большие числа; так как эти числа были большие, то они были округлены с большой абсолютной погрешностью, в результате и ответ получился с большой абсолютной погрешностью. Здесь нам впервые встретилось явление *потери значащих цифр* (или *пропадания значащих цифр*), имеющее место при вычитании близких величин; это явление, например, довольно часто приводит к существенному искажению результата при решении систем линейных алгебраических уравнений.

Рассмотрим другой типичный пример, где порядок выполнения операций влияет на погрешность результата. На машине с плавающей запятой вычисляется значение суммы

$$S_{1\,000\,000} = \sum_{j=1}^{1\,000\,000} \frac{1}{j^2}.$$

Можно вычислять  $S_{1\,000\,000}$  либо по рекуррентной формуле

$$S_n = S_{n-1} + \frac{1}{n^2}, \quad n = 1, \dots, 1\,000\,000, \quad S_0 = 0,$$

либо по рекуррентной формуле

$$\sum_{n-1} = \sum_n + \frac{1}{n^2}, \quad n = 1\,000\,000, \dots, 1, \quad \sum_{1\,000\,000} = 0.$$

Оказывается, во втором случае суммарная вычислительная погрешность будет существенно меньше.

Дело заключается в следующем. В большинстве случаев сложение чисел в ЭВМ осуществляется по следующей схеме. Два числа  $a$  и  $b$  складываются абсолютно точно, а затем происходит отбрасывание последних знаков или округление результата с тем, чтобы осталось  $t$  или  $t-1$  значащих цифр. В результате получается приближенное значение суммы  $a+b$  с погрешностью, не превосходящей  $2^{-t}|a+b|$ , но в неблагоприятном случае большей, чем  $2^{-1-t}|a+b|$ . В первом случае у нас при каждом сложении значение суммы больше 1 и в принципе возможно получение погрешности около  $10^6 \cdot 2^{-t}$ . Во втором случае

$$\sum_n = o\left(\frac{1}{n}\right),$$

поэтому погрешность накапливается существенно медленнее. Можно показать, что погрешность окончательного результата не превосходит  $100 \cdot 2^{-t}$ .

На конкретной ЭВМ было проведено вычисление  $S_{1\,000\,000}$  по обоим алгоритмам и оказалось, что для первого алгоритма погрешность  $\approx 2 \cdot 10^{-4}$ , а для второго —  $\approx 6 \cdot 10^{-8}$ .

## § 5. Погрешность функции

Довольно часто возникает следующая задача. Искомая величина  $y$  является функцией от параметров  $a_1, \dots, a_n$ :  $y = y(a_1, \dots, a_n)$ . Известна область  $G$  в пространстве переменных  $a_1, \dots, a_n$ , которой принадлежат эти параметры. Требуется получить приближение к  $y$  и оценить его погрешность.

Если  $y^*$  — приближенное значение величины  $y$ , то *предельной абсолютной погрешностью*  $A(y^*)$  называют наилучшую при имеющейся информации оценку погрешности величины  $y^*$ ; согласно этому определению в данном случае

$$A(y^*) = \sup_{(a_1, \dots, a_n) \in G} |y(a_1, \dots, a_n) - y^*|; \quad (1)$$

*предельной относительной погрешностью* называют величину

$$\frac{A(y^*)}{|y^*|}.$$

**Задача 1.** Доказать, что предельная абсолютная погрешность  $A(y^*)$  минимальна при  $y^* = (Y_1 + Y_2)/2$ , где  $Y_1 = \inf_G y(a_1, \dots, a_n)$ ,  $Y_2 = \sup_G y(a_1, \dots, a_n)$ .

Рассмотрим наиболее распространенный случай, когда область  $G$  — прямоугольник:

$$|a_j - a_j^*| \leq \Delta(a_j^*), \quad j = 1, \dots, n,$$

и за приближенное значение принимается

$$y^* = y(a_1^*, \dots, a_n^*).$$

Если  $y$  — непрерывно дифференцируемая функция своих аргументов, то, согласно формуле Лагранжа,

$$y(a_1, \dots, a_n) - y^* = \sum_{j=1}^n b_j(\theta)(a_j - a_j^*), \quad (2)$$

где

$$b_j(\theta) = \frac{\partial y}{\partial a_j} \Big|_{(a_1^* + \theta(a_1 - a_1^*), \dots, a_n^* + \theta(a_n - a_n^*))}, \quad 0 \leq \theta \leq 1.$$

Отсюда следует оценка погрешности

$$|y(a_1, \dots, a_n) - y^*| \leq A_0(y^*) = \sum_{j=1}^n B_j \Delta(a_j^*), \quad (3)$$

где

$$B_j = \sup_G \left| \frac{\partial y(a_1, \dots, a_n)}{\partial a_j} \right|.$$

Положим

$$\rho = \sqrt{\sum_{j=1}^n (\Delta(a_j^*))^2}.$$

Если производные  $\frac{\partial y(a_1, \dots, a_n)}{\partial a_j}$  непрерывны, то  $b_j(\theta) = b_j(0) + o(1)$  и

$$B_j = \left| \frac{\partial y(a_1^*, \dots, a_n^*)}{\partial a_j} \right| + o(1).$$

Здесь выражение  $x = y + o(1)$  понимается в смысле:  $x - y \rightarrow 0$  при  $\rho \rightarrow 0$ . Следовательно,  $A_0(y^*) = A^0(y^*) + o(\rho)$ , где

$$A^0(y^*) = \sum_{j=1}^n \left| \frac{\partial y(a_1^*, \dots, a_n^*)}{\partial a_j} \right| \Delta(a_j^*).$$

При практической работе вместо оценки погрешности (3) обычно пользуются более простой, вообще говоря неверной, «оценкой»

$$|y(a_1, \dots, a_n) - y^*| \leq A^0(y^*), \quad (4)$$

называемой *линейной оценкой погрешности*.

Задача 2. Доказать, что  $A_0(y^*) - A(y^*)$ ,  $A^0(y^*) - A(y^*) = o(\rho)$ .

Рассмотрим некоторые примеры определения величин  $A(y^*)$ ,  $A_0(y^*)$  и  $A^0(y^*)$  и произведем сравнение этих величин.

1.  $y = a^{10}$ ,  $a^* = 1$ ,  $\Delta(a^*) = 0,001$ . Тогда

$$y^* = 1, \quad y'_a = 10 \cdot a^9, \quad b(0) = 10,$$

$$B = \sup_{|a-1| \leq 0,001} |10 \cdot a^9| = 10,09 \dots,$$

$$A(y^*) = \sup_{|a-1| \leq 0,001} |a^{10} - 1| = 1,001^{10} - 1 = 0,010045 \dots,$$

$$A_0(y^*) = B\Delta(a^*) = 0,01009 \dots,$$

$$A^0(y^*) = |b(0)|\Delta(a^*) = 0,01.$$

Здесь оценка погрешности через величину  $A_0(y^*)$ , предельно точная оценка (1) и линейная оценка (4) различаются несущественно.

2.  $y = a^{10}$ ,  $a^* = 1$ ,  $\Delta(a^*) = 0,1$ . Тогда

$$y^* = 1, \quad B = \sup_{|a-1| \leq 0,1} 10 \cdot a^9 = 10 \cdot (1,1)^9 = 23, \dots,$$

$$A(y^*) = \sup_{|a-1| \leq 0,1} |a^{10} - 1| = (1,1)^{10} - 1 = 1,5 \dots,$$

$$A_0(y^*) = B\Delta(a^*) = 2,3 \dots, \quad A^0(y^*) = |b(0)|\Delta(a^*) = 1.$$

Здесь различие между этими оценками более заметно.

3. Проведем конкретную оценку погрешности в случае вычисления значений простейших функций. Пусть

$$y = \gamma_1 a_1 + \dots + \gamma_n a_n,$$

где  $\gamma_1, \dots, \gamma_n$  принимают значения  $\pm 1$  или  $-1$ ; пусть известны оценки  $|a_j - a_j^*| \leq \Delta(a_j^*)$ . В данном конкретном случае  $b_j(\theta) \equiv \gamma_j$ ,  $|b_j(\theta)| \equiv 1$ , поэтому

$$A(y^*) = A_0(y^*) = A^0(y^*) = \Delta(a_1^*) + \dots + \Delta(a_n^*).$$

Поскольку, по определению, погрешностью называют любую оценку для  $y - y^*$ , то это соотношение можно также записать в виде

$$\Delta(\pm a_1^* \pm \dots \pm a_n^*) = \Delta(a_1^*) + \dots + \Delta(a_n^*). \quad (5)$$

В случае  $y = a_1 \pm a_2$  это равенство иногда формулируется в виде правила:

*предельная абсолютная погрешность суммы или разности равна сумме предельных абсолютных погрешностей.*

Если погрешности в величинах  $a_j^*$  зависимы, то оценка (5) часто может быть улучшена. Рассмотрим простейший пример:  $a_1 = a$ ,  $a_2 = 1 - a$ ; известно, что в обоих случаях  $a$  одно и то же; тогда независимо от погрешности в значении  $a$  сумма  $a_1 + a_2$  равна 1 и погрешность суммы равна нулю.

Пусть теперь  $y = a_1^{p_1} \dots a_n^{p_n}$ ; тогда при всех  $j$   $b_j(0) = = p_j ((a_j^*)^{-1} \cdot y^*)$  и

$$A(y^*) \approx A^0(y^*) = \sum_{j=1}^n |p_j| |a_j^*|^{-1} |y^*| \Delta(a_j^*).$$

После деления на  $|y^*|$  получаем

$$\frac{A(y^*)}{|y^*|} \approx \frac{A^0(y^*)}{|y^*|} = \sum_{j=1}^n |p_j| \frac{\Delta(a_j^*)}{|a_j^*|} = \sum_{j=1}^n |p_j| \delta(a_j^*). \quad (6)$$

По отношению к частным случаям  $y = a_1 \cdot a_2$  или  $y = a_1 \cdot a_2^{-1}$  соотношение (6) иногда формулируют в виде правила:

*предельная относительная погрешность произведения или частного приближенно равна сумме предельных относительных погрешностей.*

4. Довольно часто возникает задача оценки погрешности функции, заданной неявно уравнением

$$F(y, a_1, \dots, a_n) = 0. \quad (7)$$

Дифференцируя (7) по  $a_j$ , имеем

$$\frac{\partial F}{\partial y} \frac{\partial y}{\partial a_j} + \frac{\partial F}{\partial a_j} = 0,$$

откуда

$$\frac{\partial y}{\partial a_j} = - \left( \frac{\partial F}{\partial a_j} \right) \left( \frac{\partial F}{\partial y} \right)^{-1}. \quad (8)$$

При заданных  $a_1^*, \dots, a_n^*$  можно найти  $y^*$  как корень уравнения (7), а затем значения

$$b_j(0) = - \left( \frac{\partial F}{\partial a_j} \right) \left( \frac{\partial F}{\partial y} \right)^{-1} \Big|_{(y^*, a_1^*, \dots, a_n^*)}. \quad (9)$$

С помощью этих величин можно получить «линейную оценку» погрешности (4).

Вследствие зависимости производных  $\partial y / \partial a_j$  от самого значения  $y$  получение строгих оценок (1), (3) здесь довольно трудоемко.

Часто решение задачи зависит от приближенно задаваемых параметров настолько сложным образом, что получение или использование явных формул для производных по этим параметрам практически неприемлемо из-за своей громоздкости и трудоемкости. В такой ситуации для оценки этих производных целесообразно воспользоваться какими-либо приближенными формулами дифференцирования, например

$$f'(a) \approx \frac{f(a + \varepsilon) - f(a)}{\varepsilon}.$$

Так, производную решения дифференциального уравнения по начальному условию, в принципе, можно вычислить, интегрируя соответствующее уравнение в вариациях, решением которого является эта производная. Однако часто разумнее воспользоваться предыдущей формулой.

5. Рассмотрим один наиболее типичный частный случай из п. 4.

Имеется приближение  $y^*$  к корню уравнения

$$f(y) = a;$$

требуется оценить его погрешность.

Вычислим величину

$$a^* = f(y^*).$$

При малых  $y^* - y$  из равенства

$$f(y) - f(y^*) = a - a^*$$

следует, что

$$f'(y^*)(y - y^*) \approx a - a^*$$

и, таким образом,

$$y - y^* \approx \frac{a - a^*}{f'(y^*)} = \frac{a - f(y^*)}{f'(y^*)}.$$

В часто встречающемся случае  $a = 0$  получаем

$$y - y^* \approx -\frac{f(y^*)}{f'(y^*)}.$$

6. Обратимся к оценке погрешности корней квадратного уравнения

$$F(y, a_1, a_2) = y^2 + a_1 y + a_2 = 0 \quad (10)$$

при заданных приближенных значениях коэффициентов  $a_1^*$ ,  $a_2^*$  и их погрешностях  $\Delta(a_1^*)$ ,  $\Delta(a_2^*)$ .

Пусть  $y^*$  — решение уравнения

$$y^{*2} + a_1^* y^* + a_2^* = 0.$$

Из формулы (9) имеем

$$b_1(0) = \frac{\partial y}{\partial a_1} \Big|_{(a_1^*, a_2^*)} = -\frac{y^*}{2y^* + a_1^*},$$

$$b_2(0) = \frac{\partial y}{\partial a_2} \Big|_{(a_1^*, a_2^*)} = -\frac{1}{2y^* + a_1^*}$$

и, следовательно,

$$A^0(y^*) = \frac{|y^*| \Delta(a_1^*) + \Delta(a_2^*)}{|2y^* + a_1^*|}. \quad (11)$$

Рассмотрим некоторую область  $|a_1| \leq b_1$ ,  $|a_2| \leq b_2$  изменения коэффициентов  $a_1$ ,  $a_2$ . Из явного выражения корней

$$y = -\frac{a_1}{2} \pm \sqrt{\frac{a_1^2}{4} - a_2}$$

следует, что корни являются непрерывными функциями коэффициентов, поэтому

$$|y(a_1, a_2) - y(a_1^*, a_2^*)| \leq \omega(|a_1 - a_1^*|, |a_2 - a_2^*|); \quad (12)$$

при  $(a_1, a_2), (a_1^*, a_2^*)$  из этой области,  $\omega(\lambda_1, \lambda_2) \rightarrow 0$  при  $\lambda_1, \lambda_2 \rightarrow 0$ . Правая часть в (11) стремится к  $\infty$  при  $2y^* + a_2^* \rightarrow 0$ ; поэтому из сравнения (11) и (12) вытекает, что «линейная оценка» погрешности при помощи формулы (4) может оказаться в некоторых случаях сильно завышенной по сравнению с точной оценкой погрешности (3). Дело в том, что ранее предполагалась непрерывная дифференцируемость  $y(a_1, \dots, a_n)$  по аргументам  $(a_1, \dots, a_n)$ . Тогда погрешность  $y^*$  оказывалась величиной того же порядка, что и погрешности аргументов  $\Delta(a_i^*)$ . В случае, когда величина  $y^*$  определяется неявным образом, при некоторых значениях параметров она оказывается недифференцируемой функцией своих аргументов и характер оценки меняется.

Пусть  $y^*$  является двукратным корнем уравнения (7) при  $a_1 = a_1^*, a_2 = a_2^*$ . Разложим левую часть (7) в ряд Тейлора в окрестности точки  $y^*, a_1^*, a_2^*$ . Поскольку

$$F(y^*, a_1^*, a_2^*) = F_y(y^*, a_1^*, a_2^*) = 0$$

при  $y^*$  — двукратном корне уравнения (7), то уравнение (7) примет вид

$$d_{200}(y - y^*)^2 + d_{010}(a_1 - a_1^*) + d_{001}(a_2 - a_2^*) + \dots = 0,$$

где

$$d_{ijk} = \frac{F_{y^i a_1^j a_2^k}(y^*, a_1^*, a_2^*)}{i!j!k!},$$

а отброшенные члены имеют порядок  $o(\rho)$ . В случае уравнения (10) можно показать, что

$$y - y^* = \pm \sqrt{d_{010}(a_1 - a_1^*) + d_{001}(a_2 - a_2^*) + o(\rho)}.$$

Таким образом, погрешность приближенного значения корня оказалась величиной порядка  $O(\sqrt{\rho})$ .

**Задача 3.** Показать, что в случае, когда уравнение имеет корень кратности  $l$ , погрешность корня имеет порядок  $O(\sqrt[l]{\rho})$ .

## § 6. Обратная задача

Часто приходится решать обратную задачу: с какой точностью надо задать значения аргументов  $a_1^*, \dots, a_n^*$  функции  $y = y(a_1, \dots, a_n)$ , чтобы погрешность  $y(a_1^*, \dots, a_n^*)$  не превосходила заданной величины  $\varepsilon$ ?

Пусть точки  $(a_1, \dots, a_n)$  и  $(a_1^*, \dots, a_n^*)$ , соответствующие истинным и приближенным значениям параметров  $a_j$ , принадлежат некоторой выпуклой области  $G$  и  $c_j = \sup_G \left| \frac{\partial y}{\partial a_j} \right|$ . Тогда имеем оценку погрешности

$$|y(a_1, \dots, a_n) - y(a_1^*, \dots, a_n^*)| \leq \sum_{j=1}^n c_j \Delta(a_j^*).$$



Любая совокупность  $(\Delta(a_1^*), \dots, \Delta(a_n^*))$  абсолютных погрешностей, удовлетворяющих неравенству

$$\sum_{j=1}^n c_j \Delta(a_j^*) \leq \varepsilon,$$

обеспечивает требуемую точность.

Если функция  $y$  зависит только от одного аргумента ( $n = 1$ ), то имеем неравенство  $c_1 \Delta(a_1^*) \leq \varepsilon$  и для достижения требуемой точности достаточно взять  $\Delta(a_1^*) = \varepsilon/c_1$ .

В случае  $n > 1$  иногда рекомендуют отвести погрешности каждого аргумента равную долю, т. е. выбрать  $\Delta(a_j^*)$  из условия  $c_j \Delta(a_j^*) = \varepsilon/n$ , т. е.  $\Delta(a_j^*) = \varepsilon/(c_j n)$ . В других случаях предлагают взять все погрешности равными, максимально возможными, т. е. положить

$$\Delta(a_1^*) = \dots = \Delta(a_n^*) = \delta, \text{ где } \delta = \varepsilon/(c_1 + \dots + c_n).$$

В простейших случаях можно последовать этим рецептам, однако в более сложных случаях целесообразно подойти к вопросу о выборе верхних границ для допустимых погрешностей аргументов  $\Delta(a_j^*)$  более аккуратно. Дело заключается в том, что достижение определенной точности в задании аргумента  $a_j$  может существенно зависеть от номера  $j$ . Тогда следует ввести в рассмотрение функцию стоимости  $F(\Delta(a_1^*), \dots, \Delta(a_n^*))$  затрат на задание точки  $(a_1^*, \dots, a_n^*)$  с заданными абсолютными погрешностями координат  $\Delta(a_1^*), \dots, \Delta(a_n^*)$  и найти минимум функции  $F(x_1, \dots, x_n)$  в области  $c_1 x_1 + \dots + c_n x_n \leq \varepsilon$ ,  $0 \leq x_1, \dots, x_n$ . Пусть он достигается в точке  $x_1^0, \dots, x_n^0$ . Далее следует положить  $\Delta(a_j^*) = x_j^0$ ,  $j = 1, \dots, n$ .

В ряде типичных случаев функция  $F(x_1, \dots, x_n)$  имеет вид

$$F(x_1, \dots, x_n) = \sum_{j=1}^n D_j x_j^{-d_j}; \quad D_j, \quad d_j > 0, \quad j = 1, \dots, n.$$

Ясно, что искомое минимальное значение функции  $F(x_1, \dots, x_n)$  достигается в некоторой точке  $(x_1^0, x_2^0, \dots, x_n^0)$  плоскости  $c_1 x_1 + \dots + c_n x_n = \varepsilon$ .

Рассмотрим случай  $n = 2$ . Составляем функцию Лагранжа

$$\Phi(x_1, x_2) = F(x_1, x_2) + \lambda(c_1 x_1 + c_2 x_2 - \varepsilon)$$

и, приравняв нулю производные  $\partial\Phi/\partial x_j$ , получим систему уравнений

$$-d_1 D_1 x_1^{-d_1-1} + \lambda c_1 = 0, \quad -d_2 D_2 x_2^{-d_2-1} + \lambda c_2 = 0.$$

Отсюда

$$x_1 = \left( \frac{D_1 d_1}{\lambda c_1} \right)^{1/(d_1+1)}, \quad x_2 = \left( \frac{D_2 d_2}{\lambda c_2} \right)^{1/(d_2+1)}. \quad (1)$$

Подставляя  $x_i$  в равенство  $c_1 x_1 + c_2 x_2 = \varepsilon$ , получим уравнение относительно  $\lambda$ :

$$c_1 \left( \frac{D_1 d_1}{\lambda c_1} \right)^{1/(d_1+1)} + c_2 \left( \frac{D_2 d_2}{\lambda c_2} \right)^{1/(d_2+1)} = \varepsilon.$$

Видно, что  $\lambda \rightarrow \infty$  при  $\varepsilon \rightarrow 0$ . Пусть  $d_1 > d_2$ . Тогда при больших  $\lambda$  главным членом в левой части является первый; поэтому имеем приближенное равенство

$$c_1 \left( \frac{D_1 d_1}{\lambda c_1} \right)^{1/(d_1+1)} \approx \varepsilon,$$

откуда следует

$$\lambda \approx \frac{D_1 d_1}{c_1} \left( \frac{c_1}{\varepsilon} \right)^{d_1+1}.$$

Подставляя  $\lambda$  в (1), получим

$$\begin{aligned} \Delta(a_1^*) &= x_1 \approx \frac{\varepsilon}{c_1}, \\ \Delta(a_2^*) &= x_2 \approx \left( \frac{D_2 d_2 c_1}{D_1 d_1 c_2} \right)^{1/(d_2+1)} \left( \frac{\varepsilon}{c_1} \right)^{(d_1+1)/(d_2+1)}. \end{aligned} \quad (2)$$

Поскольку  $d_1 > d_2$ , то в рассматриваемом примере стоимость задания первого аргумента при малых  $\varepsilon$  растет быстрее, чем стоимость задания второго аргумента. В соответствии с этим мы получили, что второй аргумент следует задавать с точностью более высокого порядка малости по  $\varepsilon$ , в то время как точность задания первого аргумента практически определяется из равенства  $c_1 \Delta(a_1^*) = \varepsilon$ .

Различный характер зависимости функции стоимости от погрешностей задания аргументов может определяться многими факторами. Если, например, параметры  $a_j$  определяются численным решением некоторых вспомогательных задач, то слагаемые  $D_j \Delta(a_j)^{-d_j}$  характеризуют различную трудоемкость решения этих задач. В других случаях этот характер может определяться сложностью получения экспериментальных данных или трудностью точного измерения тех или иных параметров на приборах.

Литература к гл. I: [5, 35, 37].

## Глава II

### ИНТЕРПОЛЯЦИЯ И ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

В этой главе излагаются наиболее широко используемые способы вычисления приближенных значений функции и ее производных в случае, когда известны значения функции в некоторых фиксированных точках. Множество этих точек иногда задается нам внешними обстоятельствами, а иногда мы можем выбрать их по своему усмотрению.

Такого рода задачи приближения и приближенного дифференцирования часто возникают как самостоятельные в ситуациях, некоторые из которых будут рассмотрены ниже. Кроме того, алгоритмы решения этих задач используются как вспомогательные при построении методов вычисления интегралов, решения дифференциальных и интегральных уравнений. Наличие большого количества методов вызвано историческим развитием теории и практики решения прикладных задач. Многие методы возникли как варианты предшествующих, отличаясь от них формой записи, изменением порядка вычислений, имевшими целью уменьшить влияние погрешности округлений при вычислениях.

Современное бурное развитие теории численных методов, вызванное широким использованием ЭВМ, привело к тому, что многие старые методы перестали использоваться и появились новые.

Некоторые методы вышли из употребления по следующим причинам. Произошло увеличение разрядности чисел в ЭВМ, и как следствие этого оказалось несущественным различие в вычислительной погрешности в зависимости от последовательности арифметических операций; в результате этого в практике вычислений постепенно закрепились простейшие по форме методы. С другой стороны, усложнение модели задачи и требование уменьшения погрешности метода, как правило, требуют и существенного роста числа выполняемых арифметических операций. Несмотря на повышение разрядности чисел в ЭВМ, для ряда методов это обстоятельство приводит к недопустимо большому значению вычислительной погрешности (так называемой не-

устойчивости). Поэтому при повышении требований к точности результата ряд методов был забракован и изъят из вычислительной практики. Тем не менее сохраняется положение, когда для решения каждой конкретной задачи можно применить довольно много методов.

## § 1. Постановка задачи приближения функций

Иногда возникает следующий вопрос. Может быть, наличие большого количества различных способов приближения объясняется просто отсутствием научного подхода к постановке и решению проблемы; если бы такой подход был, то, может быть, удалось бы предложить один оптимальный способ приближения, пригодный во всех случаях? Такой вопрос возникает и при рассмотрении других разделов численного анализа. Сколь бы ни было заманчиво разработать единый подход к решению всех задач, следует все-таки признать, что многообразии методов вызывается существом дела — многообразием различных постановок проблемы. В частности, различные теоретические разделы теории приближений, например интерполяции, можно рассматривать как изучение абстрактных моделей некоторых реальных классов проблем.

1. Простейшая задача, приводящая к приближению функций, заключается в следующем. В дискретные моменты времени  $x_1, \dots, x_n$  наблюдаются значения функции  $f(x)$ ; требуется восстановить ее значения при других  $x$ . Подобная задача возникает и в следующей обстановке. По ходу вычислений на ЭВМ приходится многократно вычислять одну и ту же сложную функцию в различных точках. Вместо ее непосредственного вычисления иногда целесообразно вычислить ее значения в отдельных выбираемых нами по своему усмотрению точках, а в других точках вычислять по каким-либо простым формулам, используя информацию об этих известных значениях.

Иногда из каких-то дополнительных соображений известно, что приближающую функцию целесообразно искать в виде

$$f(x) \approx g(x; a_1, \dots, a_n).$$

Если параметры  $a_1, \dots, a_n$  определяются из условия совпадения  $f(x)$  и приближающей функции в точках  $x_1, \dots, x_n$ , так называемых *узлах интерполяции*,

$$g(x_j; a_1, \dots, a_n) = f(x_j), \quad j = 1, \dots, n,$$

то такой способ приближения называют *интерполяцией* или *интерполированием*.

2. Пусть  $y_1$  — наименьшее из чисел  $x_i$  — узлов интерполяции, а  $y_2$  — наибольшее из них. Если точка  $x$ , в которой вычисляется значение  $f(x)$ , лежит вне отрезка  $[y_1, y_2]$ , то наряду с термином *интерполяция* употребляют термин *экстраполяция*.

Например, известно поведение какой-либо переменной до данного момента времени и требуется высказать какое-то суждение о ее дальнейшем поведении. Это может быть температура, рост производства или потребления какого-либо продукта, рост народонаселения, урожайности и т. п. Задаются какими-то моментами времени, строят интерполирующую функцию и ее значение в какой-то будущей момент принимают за прогнозируемое (экстраполируемое) значение искомой величины.

Если узлы интерполяции выбраны далеко от экстраполируемого момента времени, то тем самым слабо используется существенная информация о поведении переменной в последнее время. Если они выбраны очень близко, то увеличивается роль погрешностей в используемой информации. Таким образом, вопрос о выборе узлов интерполяции и экстраполяции непрост, особенно в задачах, где значения исследуемой функции зависят от многих случайных или трудно учитываемых факторов. Сюда относятся задачи прогноза погоды, урожайности, медицины и т. д., в которых, как правило, требуется применять более сложные (в частности статистические) методы прогнозирования.

3. Наиболее часто используется рассматриваемая ниже интерполяция многочленами. Однако это не единственный возможный вид интерполяции. Иногда удобнее приближать функции тригонометрическими функциями; в других задачах целесообразнее приближать многочленом не  $f(x)$ , а  $\ln f(x)$ , или приближать  $f(x)$  не многочленом от  $x$ , а многочленом от  $\ln x$ .

Последнее время значительно расширяется круг задач, где целесообразна интерполяция дробно-рациональными функциями

$$f(x) \approx g(x) = \frac{\sum_{i=0}^n a_i x^i}{\sum_{j=0}^m b_j x^j}.$$

4. В задачах *планирования экспериментов* возникает следующая проблема. Известен вид хорошего приближения функции, например функция хорошо приближается многочленом второй степени. В то же время измеряемые значения функции содержат большие погрешности. Требуется получить наилучшее в определенной норме приближение при минимальном числе измерений значений функции. Такая задача возникает при планировании экспериментов в биологии, физике, химии, географии, медицине и других областях науки.

5. Задача приближения появляется при составлении стандартных программ вычисления элементарных и специальных функций. Обычно такие функции обладают специальными свойствами, позволяющими резко уменьшить объем вычислений.

Возникающая здесь проблема может быть сформулирована следующим образом. Рассматриваются все функции  $g(x)$ , программа вычисления которых записывается в  $k$  ячейках ЭВМ, такие, что некоторая норма погрешности  $\|f - g\|$  не превосходит  $\varepsilon$ .

Среди всех таких функций нужно выбрать ту, вычисление которой требует минимальных затрат времени ЭВМ.

В зависимости от обстановки норма может выбираться по-разному. В большинстве случаев берется  $\|f\| = \sup_{[a, b]} |f|$ , где

$[a, b]$  — отрезок, на котором приближается функция.

Довольно часто требуется повышенная точность в отдельных точках. Например, одна из стандартных программ вычисления  $\sin x$  обеспечивает малость погрешности в норме

$$\|f\| = \sup_{[0, \pi/2]} |\rho(x)f(x)|, \quad \rho(x) = \min(10^{19}, x^{-1}).$$

Введение множителя  $\rho(x)$  вызывается требованием малости относительной погрешности значений  $\sin x$  при малых  $x$ .

Точно так же по-разному может толковаться требование минимальности затрат времени ЭВМ. Затраты, вообще говоря, могут зависеть от точки, в которой вычисляется значение функции. Обозначим их через  $t(x)$ . Если не имеется информация о том, с какой частотой вычисляются значения функции в тех или иных частях отрезка, то целесообразно в качестве общей меры затрат принять

$$T = \sup_x t(x).$$

Если такая информация есть, то можно принять

$$T = M(t(x)),$$

где  $M(t(x))$  — математическое ожидание случайной величины  $t(x)$ .

6. При задании функции графиком или сложным аналитическим выражением возникают вариационные задачи других типов. Например, пусть решено разбить отрезок  $[a, b]$  на  $l$  частей:

$$[a_{i-1}, a_i], \quad i = 1, \dots, l, \quad a_0 = a, \quad a_l = b,$$

и на каждом отрезке  $[a_{i-1}, a_i]$  приближать функцию многочленом степени  $n_i$ . Среди таких способов приближения отыскивается оптимальный в том или ином смысле. Чаще всего заранее накладывается требование  $n_i \equiv n$ , фиксируется число отрезков разбиения  $l$  и проводится оптимизация метода по  $a_1, \dots, a_{l-1}$ .

В частном случае  $l = 1$  возникает задача наилучшего приближения многочленами. Об этой задаче речь пойдет в гл. IV.

7. Вид приближающей функции существенно зависит от цели, с которой осуществляется приближение. Предположим, что с требуемой точностью функция может быть приближена многочленом десятой степени или выражением  $a_1 \sin \omega_1 x + a_2 \sin \omega_2 x$ . Если полученное приближение используется в теоретических исследованиях, для решения задачи на моделирующем устройстве или в технологическом процессе, то вторая форма записи может быть более удобной. Однако если значения функции вычисляются на ЭВМ, то вторая форма записи потребует при своей реализации большего числа арифметических операций.

Далее будет конкретно рассмотрена задача интерполирования многочленами. Ее выделение вызвано наличием непосредственных многочисленных приложений, а также и следующими обстоятельствами: аппарат интерполирования многочленами является важнейшим аппаратом численного анализа; на его основе строится большинство методов решения других задач; его роль в численном анализе аналогична роли формулы Тейлора в классическом анализе.

Попутно будут рассмотрены некоторые другие вопросы общего характера, имеющие значение для других разделов численного анализа.

## § 2. Интерполяционный многочлен Лагранжа

Среди способов интерполирования наиболее распространен случай линейного интерполирования, когда приближение ищется в виде

$$g(x; a_1, \dots, a_n) = \sum_{i=1}^n a_i \varphi_i(x),$$

где  $\varphi_i(x)$  — фиксированные функции, значения коэффициентов  $a_i$  определяются из условия совпадения с приближаемой функцией в узлах интерполяции  $x_j$ :

$$f(x_j) = \sum_{i=1}^n a_i \varphi_i(x_j), \quad j = 1, \dots, n. \quad (1)$$

Метод решения задачи, при котором коэффициенты  $a_i$  определяются непосредственным решением системы (1), называется *методом неопределенных коэффициентов*.

Как правило, в методе неопределенных коэффициентов число заданных условий равно числу свободных (неизвестных) параметров, подлежащих определению.

Наиболее изучен случай интерполирования многочленами

$$\sum_{i=1}^n a_i x^{i-1}. \quad (2)$$

Тогда

$$\varphi_i(x) = x^{i-1}, \quad i = 1, \dots, n,$$

и система уравнений (1) имеет вид

$$\sum_{i=1}^n a_i x_j^{i-1} = f(x_j), \quad j = 1, \dots, n. \quad (3)$$

Далее мы предполагаем, что все  $x_j$  различные. Определитель этой системы  $\det [x_j^{i-1}]$  отличен от 0 (определитель Вандермонда). Следовательно, система (3) всегда имеет решение, и притом единственное. Таким образом, доказано существование и единственность интерполяционного многочлена вида (2).

Непосредственное нахождение коэффициентов  $a_j$  с помощью решения этой системы уже при сравнительно небольших  $n$ , например при  $n = 20$ , приводит к катастрофическому искажению коэффициентов  $a_j$  вычислительной погрешностью. Кроме того, как мы увидим в гл. IV, уже сама запись многочлена в традиционной форме (2) часто приводит к большой вычислительной погрешности результата. При теоретических исследованиях, например при конструировании алгоритмов решения других задач, эти обстоятельства могут не играть роли. Однако при реальных вычислениях значений функции путем замены их значений значениями интерполяционных многочленов влияние вычислительной погрешности может быть недопустимо большим, поэтому применяются другие виды интерполяционного многочлена и способы его записи.

Нам удастся получить ряд явных представлений интерполяционного многочлена (2), не прибегая к непосредственному решению системы (3). Сразу же отметим, что в других случаях, например при интерполировании функций многих переменных, получение интерполирующей функции в явном виде часто крайне затруднительно, и все-таки приходится прибегать к непосредственному решению системы уравнений типа (1).

Пусть  $\delta_i^j$  есть символ, определяемый соотношениями

$$\delta_i^j = \begin{cases} 1 & \text{при } i = j, \\ 0 & \text{при } i \neq j. \end{cases}$$

Задача интерполирования будет решена, если удастся построить многочлены  $\Phi_i(x)$  степени не выше  $n - 1$  такие, что  $\Phi_i(x_j) = \delta_i^j$  при  $i, j = 1, \dots, n$ . Многочлен

$$g_n(x) = \sum_{i=1}^n f(x_i) \Phi_i(x)$$

будет искомым интерполяционным многочленом. В самом деле,

$$g_n(x_j) = \sum_{i=1}^n f(x_i) \Phi_i(x_j) = \sum_{i=1}^n f(x_i) \delta_i^j = f(x_j);$$

кроме того,  $g_n(x)$  — многочлен степени  $n - 1$ .

Поскольку  $\Phi_i(x_j) = 0$  при  $j \neq i$ , то  $\Phi_i(x)$  делится на  $x - x_j$  при  $j \neq i$ . Таким образом, нам известны  $n - 1$  делителей многочлена степени  $n - 1$ , поэтому

$$\Phi_i(x) = \text{const} \prod_{j \neq i} (x - x_j).$$

Из условия  $\Phi_i(x_i) = 1$  получаем

$$\Phi_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$



Интерполяционный многочлен (2), записанный в форме

$$g_n(x) \equiv L_n(x) = \sum_{i=1}^n f(x_i) \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}, \quad (4)$$

называют *интерполяционным многочленом Лагранжа*.

Существуют другие формы записи того же интерполяционного многочлена (2), например рассматриваемая далее *интерполяционная формула Ньютона* и ее варианты. При точных (без округлений) вычислениях значения, получаемые по различным интерполяционным формулам, совпадают. Наличие же округлений приводит к различию в получаемых по этим формулам значений интерполяционных многочленов. Запись многочлена в форме Лагранжа, как правило, приводит к меньшей величине вычислительной погрешности; запись же многочлена в форме Ньютона более наглядна и позволяет лучше проследить аналогию проводимых построений с основными построениями математического анализа. Кроме того, этим различным формам записи соответствует различное количество арифметических операций при вычислении с их помощью значений интерполяционного многочлена.

Мы употребили термин «количество арифметических операций». Поясним, что имеется в виду. Пусть рассматривается задача вычисления значения многочлена

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$$

в точке  $x$ . Вычисления можно проводить различными способами. Например, можно поступить следующим образом. Вычислить значение  $a_1 x$  и сложить с  $a_0$ . Далее вычислить значение  $a_2 x^2$  и сложить с полученным результатом и т. д. На  $j$ -м шаге, таким образом, вычисляется значение  $a_j x^j$  и складывается с уже вычисленной суммой  $a_0 + a_1 x + \dots + a_{j-1} x^{j-1}$ . Вычисление значения  $a_j x^j$  требует  $j$  операций умножения. Таким образом, описанный выше алгоритм требует для вычисления значения многочлена  $(1 + 2 + \dots + n) = n(n+1)/2$  операций умножения и  $n$  операций сложения. Количество арифметических операций (действий) в данном случае будет равно  $\Phi_1 = n(n+1)/2 + n$ .

Ясно, что количество арифметических операций, необходимых для вычисления значения  $P_n(x)$ , может быть уменьшено. Например, можно последовательно вычислить и запомнить значения  $x^2, x^3, \dots, x^n$ . Это потребует  $n-1$  операции умножения. Далее вычисляем величины  $a_j x^j$  ( $j = 1, \dots, n$ ). Это потребует  $n$  операций умножения. Складывая полученные значения (это требует  $n$  операций сложения), получаем  $P_n(x)$ . В этом случае  $\Phi_2 = (2n-1) + n$  и уже при  $n > 2$  имеет место неравенство  $\Phi_2 > \Phi_1$ .

Можно пойти еще дальше. Запишем  $P_n(x)$  в виде

$$P_n(x) = (((a_n x + a_{n-1}) x + a_{n-2}) x + a_{n-3}) x + \dots + a_0.$$

Для вычисления значения во внутренних скобках  $a_n x + a_{n-1}$  требуется одна операция умножения и одна операция сложения. Для вычисления значения в следующих скобках  $(a_n x + a_{n-1})x + a_{n-2}$  требуется опять одна операция умножения и одна операция сложения, так как  $a_n x + a_{n-1}$  уже вычислено, и т. д. Таким образом, вычисление  $P_n(x)$  при помощи этого алгоритма потребует  $n$  операций умножения и  $n$  операций сложения, т. е.  $\Phi_3 = n + n$ .

Ясно, что  $\Phi_3 < \Phi_2 < \Phi_1$  при  $n > 2$ . Таким образом, вычисление  $P_n(x)$  по последнему алгоритму потребует меньше арифметических операций и, соответственно, меньше времени ЭВМ. Количество арифметических операций, которое требуется для получения результата, является одной из важнейших характеристик метода, по которой происходит сравнение методов.

Так как операция сложения (вычитания) обычно требует существенно меньшего времени на ЭВМ, чем операция умножения (деления), то иногда под количеством арифметических операций понимают только количество операций умножения и деления.

Часто не удается точно вычислить требуемое количество арифметических операций. В этих случаях говорят о *порядке количества арифметических операций* по отношению к какому-либо параметру. В рассматриваемом выше примере

$$\Phi_1 = O(n^2), \quad \Phi_2, \quad \Phi_3 = O(n),$$

$n$  — степень многочлена.

В последнем случае ( $\Phi_2, \Phi_3 = O(n)$ ) говорят, что методы имеют одинаковый порядок количества арифметических операций.

В тех случаях, когда находится порядок количества арифметических операций, бывает важно найти постоянную в главном члене. Например,

$$\Phi_1 = \frac{1}{2} n^2 + o(n^2), \quad \Phi_2 = 3n + o(n), \quad \Phi_3 = 2n.$$

Как правило, метод, требующий меньшего количества арифметических операций, является более быстрым, и поэтому считается лучшим (если не принимать во внимание операции управления программой, требуемый объем оперативной памяти и другие характеристики метода).

Заметим, что значение многочлена  $P_n(x)$  определяется параметрами  $a_0, a_1, \dots, a_n$  и величиной  $x$ . Поэтому в общем случае для вычисления  $P_n(x)$  потребуется не менее  $n$  арифметических операций, т. е. мы имеем оценку снизу для количества арифметических операций. Таким образом, второй и третий методы вычисления  $P_n(x)$  являются *оптимальными по порядку*, так как  $\Phi_2, \Phi_3 = O(n)$  и для любого метода  $\Phi \geq n$ .

В связи с появлением многопроцессорных машин может случиться так, что метод, требующий большего количества арифметических операций, будет

быстрее другого метода с меньшим количеством арифметических операций. Поэтому для случая многопроцессорной ЭВМ нельзя оценивать качество метода только по количеству арифметических операций.

### § 3. Оценка остаточного члена интерполяционного многочлена Лагранжа

В предположении непрерывности  $f^{(n)}(x)$  оценим разность между  $f(x)$  и построенным интерполяционным многочленом  $g_n(x)$ . Положим

$$\varphi(t) = f(t) - g_n(t) - K\omega_n(t),$$

где  $\omega_n(t) = (t - x_1) \dots (t - x_n)$ , а  $K$  выберем из условия  $\varphi(x) = 0$ , где  $x$  — точка, в которой оценивается погрешность. Из уравнения  $\varphi(x) = 0$  получаем

$$K = \frac{f(x) - g_n(x)}{\omega_n(x)}.$$

При таком выборе  $K$  функция  $\varphi(t)$  обращается в нуль в  $(n + 1)$ -й точке  $x_1, \dots, x_n, x$ . На основании теоремы Ролля ее производная  $\varphi'(t)$  обращается в нуль по крайней мере в  $n$  точках. Применяя теорему Ролля к  $\varphi'(t)$ , получаем, что ее производная  $\varphi''(t)$  обращается в нуль по крайней мере в  $(n - 1)$ -й точке. Продолжая эти рассуждения дальше, получаем, что  $\varphi^{(n)}(t)$  обращается в нуль по крайней мере в одной точке  $\xi$ , принадлежащей отрезку  $[y_1, y_2]$ , где

$$y_1 = \min(x_1, \dots, x_n, x), \quad y_2 = \max(x_1, \dots, x_n, x).$$

Поскольку

$$\varphi^{(n)}(t) = f^{(n)}(t) - Kn!,$$

из условия  $\varphi^{(n)}(\xi) = 0$  будем иметь

$$K = \frac{f^{(n)}(\xi)}{n!}.$$

Следовательно, соотношение  $\varphi(x) = 0$  можно переписать в виде

$$f(x) - g_n(x) = \frac{f^{(n)}(\xi) \omega_n(x)}{n!}, \quad \xi \in [y_1, y_2], \quad (1)$$

который и дает представление остаточного члена.

### § 4. Разделенные разности и их свойства

Как будет видно далее, интерполяционный многочлен можно рассматривать как обобщение отрезка ряда Тейлора.

Обобщением понятия производной является понятие *разделенной разности*. Разделенные разности нулевого порядка  $f(x_i)$  совпадают со значениями функции  $f(x_i)$ ; разности первого по-

рядка определяются равенством

$$f(x_i; x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i}, \quad (1)$$

разности второго порядка — равенством

$$f(x_i; x_j; x_k) = \frac{f(x_j; x_k) - f(x_i; x_j)}{x_k - x_i},$$

и, вообще, разности  $k$ -го порядка  $f(x_1; \dots; x_{k+1})$  определяются через разности  $(k-1)$ -го порядка по формуле

$$f(x_1; \dots; x_{k+1}) = \frac{f(x_2; \dots; x_{k+1}) - f(x_1; \dots; x_k)}{x_{k+1} - x_1}. \quad (2)$$

Иногда вместо  $f(x_1; \dots; x_k)$  используют обозначения  $(f)(x_1; \dots; x_k)$  или  $[x_1; \dots; x_k]$ .

*Лемма. Справедливо равенство*

$$f(x_1; \dots; x_k) = \sum_{j=1}^k \frac{f(x_j)}{\prod_{i \neq j} (x_j - x_i)}. \quad (3)$$

Доказательство будем проводить по индукции. При  $k=1$  это равенство превращается в равенство  $f(x_1) = f(x_1)$ , при  $k=2$  совпадает с (1). Пусть (3) доказано при  $k < l$ . Тогда

$$\begin{aligned} f(x_1; \dots; x_{l+1}) &= \frac{f(x_2; \dots; x_{l+1}) - f(x_1; \dots; x_l)}{x_{l+1} - x_1} = \\ &= \frac{1}{x_{l+1} - x_1} \left( \sum_{j=2}^{l+1} \frac{f(x_j)}{\prod_{\substack{i \neq j \\ 2 \leq i \leq l+1}} (x_j - x_i)} - \sum_{j=1}^l \frac{f(x_j)}{\prod_{\substack{i \neq j \\ 1 \leq i \leq l}} (x_j - x_i)} \right). \end{aligned}$$

Если  $j \neq 1, l+1$ , то коэффициент при  $f(x_j)$  в правой части есть

$$\begin{aligned} \frac{1}{x_{l+1} - x_1} \left( \frac{1}{\prod_{\substack{i \neq j \\ 2 \leq i \leq l+1}} (x_j - x_i)} - \frac{1}{\prod_{\substack{i \neq j \\ 1 \leq i \leq l}} (x_j - x_i)} \right) &= \\ &= \frac{(x_j - x_1) - (x_j - x_{l+1})}{(x_{l+1} - x_1) \prod_{\substack{i \neq j \\ 1 \leq i \leq l+1}} (x_j - x_i)} = \frac{1}{\prod_{\substack{i \neq j \\ 1 \leq i \leq l+1}} (x_j - x_i)}, \end{aligned}$$

т. е. имеет требуемый вид; для  $j=1$  или  $j=l+1$  значение  $f(x_j)$  входит только в одно слагаемое в правой части, и коэффициент при нем также имеет требуемый вид. Доказательство закончено.

Непосредственно из (3) вытекает ряд следствий.

1. При фиксированных  $x_1, \dots, x_k$  разделенная разность является линейным функционалом от функции  $f$ :

$$(\alpha_1 f_1 + \alpha_2 f_2)(x_1; \dots; x_k) = \alpha_1 f_1(x_1; \dots; x_k) + \alpha_2 f_2(x_1; \dots; x_k).$$

2. Разделенная разность есть симметрическая функция своих аргументов  $x_1, \dots, x_k$  (т. е. не меняется при любой их перестановке).

Если функция задана в точках  $x_1, \dots, x_n$ , то таблицу

$$\begin{array}{ccccccc} f(x_1) & & & & & & \\ f(x_2) & f(x_1; x_2) & & & & & \\ f(x_3) & f(x_2; x_3) & f(x_1; x_2; x_3) & \dots & & & \\ \vdots & \vdots & \vdots & & & & \\ f(x_n) & f(x_{n-1}; x_n) & \dots & \dots & \dots & \dots & f(x_1; \dots; x_n) \end{array} \quad (4)$$

называют *таблицей разделенных разностей*.

### § 5. Интерполяционная формула Ньютона с разделенными разностями

При помощи разделенных разностей можно получить другую форму записи интерполяционного многочлена (2.4).

Справедливо равенство

$$\begin{aligned} f(x) - L_n(x) &= f(x) - \sum_{i=1}^n f(x_i) \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} = \\ &= \prod_{i=1}^n (x - x_i) \left( \frac{f(x)}{\prod_{i=1}^n (x - x_i)} + \sum_{i=1}^n \frac{f(x_i)}{(x_i - x) \prod_{j \neq i} (x_i - x_j)} \right). \end{aligned}$$

Сравнивая с (4.3), убеждаемся, что выражение в скобках есть  $f(x; x_1; \dots; x_n)$ . Таким образом, можно написать

$$f(x) - L_n(x) = f(x; x_1; \dots; x_n) \omega_n(x), \quad (1)$$

где  $\omega_n(x)$  определено в § 3.

Пусть  $L_m(x)$  — интерполяционный многочлен Лагранжа с узлами интерполяции  $x_1, \dots, x_m$ . Интерполяционный многочлен Лагранжа  $L_n(x)$  можно представить в виде

$$L_n(x) = L_1(x) + (L_2(x) - L_1(x)) + \dots + (L_n(x) - L_{n-1}(x)). \quad (2)$$

Разность  $L_m(x) - L_{m-1}(x)$  есть многочлен степени  $m-1$ , обращающийся в нуль в точках  $x_1, \dots, x_{m-1}$ , поскольку  $L_{m-1}(x_j) = L_m(x_j) = f(x_j)$  при  $1 \leq j \leq m-1$ . Следовательно,

$$L_m(x) - L_{m-1}(x) = A_{m-1} \omega_{m-1}(x), \quad \omega_{m-1}(x) = (x - x_1) \dots (x - x_{m-1}).$$

Полагая  $x = x_m$ , получим

$$f(x_m) - L_{m-1}(x_m) = A_{m-1} \omega_{m-1}(x_m).$$

С другой стороны, полагая в (1)  $n = m - 1$  и  $x = x_m$ , имеем

$$f(x_m) - L_{m-1}(x_m) = f(x_m; x_1; \dots; x_{m-1}) \omega_{m-1}(x_m).$$

Таким образом,  $A_{m-1} = f(x_1; \dots; x_m)$ , поэтому

$$L_m(x) - L_{m-1}(x) = f(x_1; \dots; x_m) \omega_{m-1}(x).$$

Подставляя эти величины в (2), получим

$$L_n(x) = f(x_1) + f(x_1; x_2)(x - x_1) + \dots \\ \dots + f(x_1; \dots; x_n)(x - x_1) \dots (x - x_{n-1}). \quad (3)$$

Интерполяционный многочлен, записанный в такой форме, называется *интерполяционным многочленом Ньютона с разделенными разностями*. Из сравнения (1) с (3.1) следует важное равенство

$$f(x; x_1; \dots; x_n) = \frac{f^{(n)}(\xi)}{n!}, \quad y_1 \leq \xi \leq y_2. \quad (4)$$

В частности, если  $f(x)$  — многочлен

$$P_l(x) = \sum_{j=0}^l b_j x^j$$

степени  $l \leq n$ , то на основании этой формулы имеем

$$P_l(x_0; \dots; x_n) = \begin{cases} b_n & \text{при } l = n, \\ 0 & \text{при } l < n \end{cases}$$

при любых  $x_0, \dots, x_n$ .

Предположим, что точки  $x_1, \dots, x_n$  пронумерованы в порядке возрастания:  $x_1 < x_2 < \dots < x_n$ . Вследствие (4) имеем

$$m! f(x_k; \dots; x_{k+m}) = f^{(m)}(\xi_k), \quad \text{где } x_k \leq \xi_k \leq x_{k+m}.$$

Поэтому величина

$$M_m = \sup_{1 \leq k \leq n-m} m! |f(x_k, \dots, x_{k+m})|$$

может использоваться в качестве приближенной оценки для величины

$$M^{(m)} = \sup_{[x_1, x_n]} |f^{(m)}(x)|.$$

Задача 1. Доказать неравенство

$$|M_m - M^{(m)}| \leq M^{(m+1)} h_m,$$

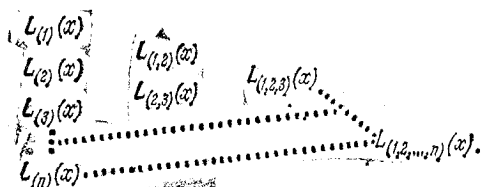
где  $h_m = \max_{1 \leq k \leq n-m} (x_{k+m} - x_k)$ .

Для упрощения вычислений интерполяционного многочлена удобно использовать так называемую *схему Эйткена*.

Пусть  $L_{(k, k+1, \dots, l)}(x)$  — интерполяционный многочлен с узлами интерполяции  $x_k, \dots, x_l$ , в частности  $L_{(k)}(x) = f(x_k)$ . Справедливо равенство

$$L_{(k, k+1, \dots, l+1)}(x) = \frac{L_{(k+1, \dots, l+1)}(x)(x - x_k) - L_{(k, \dots, l)}(x)(x - x_{l+1})}{x_{l+1} - x_k}; \quad (5)$$

действительно, правая часть является многочленом степени  $l - k + 1$  и совпадает с  $f(x)$ , в точках  $x_k, \dots, x_{l+1}$ . Схема Эйткена вычисления значения  $L_{(1, \dots, n)}(x)$  заключается в последовательном вычислении с помощью формулы (5) элементов таблицы значений интерполяционных многочленов



Эта схема положена в основу *стандартной программы* решения следующей задачи.

Дана таблица значений некоторой функции  $f(x)$ ; требуется при каждом значении  $x$  вычислить значение  $f(x)$  с заданной точностью  $E$  или с наилучшей возможной точностью при имеющейся информации.

Трудно дать четкое определение термина *стандартная программа*. По установившейся традиции стандартной программой решения задач некоторого класса называют квалифицированно написанную программу, содержащую описание алгоритма решения задач данного класса. Решение конкретной задачи осуществляется подсоединением к стандартной программе информации об этой конкретной задаче. От стандартной программы также требуется, чтобы допускалось ее использование как элемента программы, предназначенной для решения более сложных задач.

Рассматриваемое ниже построение алгоритма решения задачи является довольно типичным для ситуации, возникающей в реальной практике. Невозможно предложить обоснованный алгоритм решения поставленной задачи для всех функций, поскольку про функцию ничего не известно, кроме ее значений в заданных точках. Однако, предполагая функцию гладкой, мы выводим практический критерий оценки погрешности и, основываясь на нем, строим алгоритм решения задачи.

Пусть  $x$  фиксировано; перенумеруем узлы интерполяции в порядке возрастания  $|x_i - x|$ . Интерполяционные многочлены  $L_{(1, \dots, m)}(x)$  будем обозначать, как обычно,  $L_m(x)$ .

Выше получено представление погрешности (1)

$$f(x) - L_m(x) = f(x; x_1; \dots; x_m) \omega_m(x),$$

а также равенство

$$L_{m+1}(x) - L_m(x) = f(x; \dots; x_{m+1}) \omega_m(x). \quad (6)$$





Зададимся последовательностью совокупностей точек  $x_{ij}^\varepsilon$ ,  $0 < \varepsilon < \varepsilon_0$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, m_i$ , удовлетворяющих следующим условиям: при  $0 < \varepsilon < \varepsilon_0$  все точки  $x_{ij}^\varepsilon$  различные,  $x_{ij}^\varepsilon \rightarrow x_i$  при  $\varepsilon \rightarrow 0$ . В частности, можно положить  $x_{ij}^\varepsilon = x_i + (j-1)\varepsilon$ .

Построим интерполяционный многочлен  $g_s^\varepsilon(x)$  степени  $s-1$ , совпадающий с  $f(x)$  в точках  $x_{ij}^\varepsilon$ . Таблица разделенных разностей, соответствующих этому набору узлов, имеет вид

$$\begin{array}{r}
 f(x_{i1}^\varepsilon) \\
 f(x_{i2}^\varepsilon) \\
 f(x_{i3}^\varepsilon) \\
 \vdots \\
 f(x_{im_i}^\varepsilon) \\
 f(x_{21}^\varepsilon) \\
 \vdots \\
 f(x_{nm_n}^\varepsilon)
 \end{array}
 \begin{array}{l}
 f(x_{i1}^\varepsilon; x_{i2}^\varepsilon) \\
 f(x_{i2}^\varepsilon; x_{i3}^\varepsilon) \\
 \vdots \\
 f(x_{im_i}^\varepsilon; x_{21}^\varepsilon) \\
 \vdots \\
 f(x_{nm_n}^\varepsilon; \dots; x_{im_i}^\varepsilon)
 \end{array}
 \begin{array}{l}
 f(x_{i1}^\varepsilon; x_{i2}^\varepsilon; x_{i3}^\varepsilon) \\
 \vdots \\
 f(x_{i1}^\varepsilon; x_{i2}^\varepsilon; \dots; x_{im_i}^\varepsilon) \\
 \vdots \\
 f(x_{i1}^\varepsilon; x_{i2}^\varepsilon; \dots; x_{im_i}^\varepsilon; \dots; x_{nm_n}^\varepsilon)
 \end{array}
 \quad (2)$$

Выпишем интерполяционную формулу Ньютона с разделенными разностями:

$$g_s^\varepsilon(x) = A_0^\varepsilon + A_1^\varepsilon(x - x_{i1}^\varepsilon) + A_2^\varepsilon(x - x_{i1}^\varepsilon)(x - x_{i2}^\varepsilon) + \dots \\
 \dots + A_{s-1}^\varepsilon(x - x_{i1}^\varepsilon) \dots (x - x_{n, m_n}^\varepsilon),$$

где

$$A_0^\varepsilon = f(x_{i1}^\varepsilon), \quad A_1^\varepsilon = f(x_{i1}^\varepsilon; x_{i2}^\varepsilon), \\
 A_2^\varepsilon = f(x_{i1}^\varepsilon; x_{i2}^\varepsilon; x_{i3}^\varepsilon), \quad \dots, \quad A_{s-1}^\varepsilon = f(x_{i1}^\varepsilon; x_{i2}^\varepsilon; \dots; x_{n, m_n}^\varepsilon).$$

Выражая разделенные разности через производные, имеем

$$f(x_{i1}^\varepsilon; \dots; x_{im}^\varepsilon) = \frac{f^{(m-l)}(\bar{x}_{i1m}^\varepsilon)}{(m-l)!}.$$

Переходя к пределу при  $\varepsilon \rightarrow 0$ , получаем

$$\lim_{\varepsilon \rightarrow 0} f(x_{i1}^\varepsilon; \dots; x_{im}^\varepsilon) = \frac{f^{(m-l)}(x_i)}{(m-l)!}. \quad (3)$$

Таким образом, все разделенные разности таблицы (2) вида  $f(x_{i1}^\varepsilon; \dots; x_{im}^\varepsilon)$  при  $\varepsilon \rightarrow 0$  имеют пределы, которые естественно обозначать  $\underbrace{f(x_i; \dots; x_i)}_{m-l+1 \text{ раз}}$ . Из (3) следует, что

$$\underbrace{f(x_i; \dots; x_i)}_{p+1 \text{ раз}} = \frac{f^{(p)}(x_i)}{p!}. \quad (4)$$

**Задача 1.** Индукцией по порядку разности показать, что все разделенные разности, входящие в таблицу (2), имеют конечные пределы.

Если все элементы таблицы (2) имеют пределы, то на любом отрезке многочлены  $g_s^\varepsilon(x)$  при  $\varepsilon \rightarrow 0$  стремятся к некоторому многочлену

$$\begin{aligned} g_s(x) &= A_0 + A_1(x - x_1) + A_2(x - x_1)^2 + \dots \\ &\dots + A_{s-1}(x - x_1)^{s-1} \dots (x - x_{n-1})^{m_{n-1}}(x - x_n)^{m_n-1} = \\ &= f(x_1) + f(x_1; x_1)(x - x_1) + f(x_1; x_1; x_1)(x - x_1)^2 + \dots; \\ &A_i = \lim_{\varepsilon \rightarrow 0} A_i^\varepsilon. \quad (5) \end{aligned}$$

Многочлен  $g_s(x)$  записывается в виде

$$g_s(x) = \sum_{i=1}^{m_1} \frac{f^{(i-1)}(x_1)}{(i-1)!} (x - x_1)^{i-1} + O((x - x_1)^{m_1}).$$

Отсюда вытекает, что он удовлетворяет условиям, заданным в точке  $x_1$ . Вследствие единственности интерполяционного многочлена многочлен  $g_s^\varepsilon(x)$  не изменится при переобозначении  $x_1 = x_j$ ,  $x_j = x_1$ . Поэтому предельный многочлен будет удовлетворять заданным условиям в любой точке  $x_j$ . Следовательно, этот многочлен является искомым.

**Задача 2.** Доказать равенство

$$f(x) - g_s(x) = \frac{f^{(s)}(\xi)}{s!} \omega_s(x), \quad \omega_s(x) = \prod_{i=1}^n (x - x_i)^{m_i}, \quad y_1 \leq \xi \leq y_2, \quad (6)$$

где  $y_1 = \min(x, x_1, \dots, x_n)$ ,  $y_2 = \max(x, x_1, \dots, x_n)$ .

Согласно (5.1) справедливо равенство

$$f(x) - g_s^\varepsilon(x) = f(x; x_{11}^\varepsilon; \dots; x_{nm}^\varepsilon) \omega_s^\varepsilon(x),$$

где  $\omega_s^\varepsilon(x) = \omega_s(x_\varepsilon)$ . Переходя к пределу при  $\varepsilon \rightarrow 0$ , получим

$$f(x) - g_s(x) = f(x; x_1; \dots; x_n) \omega_s(x).$$

Сравнивая это равенство с (6), имеем

$$f(x; x_1; \dots; x_n) = \frac{f^{(s)}(\xi)}{s!}.$$

Это соотношение остается в силе при предельном переходе  $x \rightarrow x_j$ ,  $j$  — любое. Из этих соотношений следует, что формула (5.4)

$$f(x_1; \dots; x_{N+1}) = \frac{f^{(N)}(\xi)}{N!}$$

(переписанная в других обозначениях) справедлива и в случае, когда не все  $x_1, \dots, x_{N+1}$  — различные.

Мы доказали существование интерполяционного многочлена, удовлетворяющего условиям (1). Задачу интерполяции можно было бы поставить и таким образом.

Задана таблица чисел  $a_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, m_i$ . Требуется построить многочлен  $g_s(x)$  степени  $s - 1$ , удовлетворяющий условиям

$$g_s^{(j-1)}(x_i) = a_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, m_i.$$

Эта задача равносильна исходной задаче, поскольку всегда можно указать гладкую функцию  $f(x)$  такую, что

$$f^{(j-1)}(x_i) = a_{ij}.$$

## § 7. Уравнения в конечных разностях

*Конечно-разностными уравнениями* называют уравнения относительно функций дискретного переменного, возникающие при аппроксимации обыкновенных и многомерных дифференциальных уравнений и в ряде других случаев.

Существует глубокая аналогия между непрерывными и дискретными случаями. В частности, справедливы разностные аналоги формул Грина; если в некоторой задаче применим метод Фурье, то в отношении соответствующей разностной задачи применим дискретный вариант метода Фурье. Практически каждому интегральному тождеству в теории дифференциальных уравнений можно поставить в соответствие некоторый дискретный вариант. В руках квалифицированного математика методы решения конечно-разностных уравнений являются мощнейшим средством исследования чувствительности («устойчивости») алгоритмов к вычислительной погрешности. Если требуется исследовать алгоритм решения некоторой задачи, то подбирают близкую по структуре задачу (например, следуя *принципу замороженных коэффициентов* (см. гл. X)), для которой решение соответствующей конечно-разностной задачи находится в явном виде. Анализируя алгоритм решения исходной задачи на примере этой конечно-разностной задачи, выносят предварительное суждение о его свойствах. Как правило, при практическом решении задач в большинстве случаев полученное на таком пути предварительное суждение дает правильное представление о свойствах алгоритма.

Непосредственно конечно-разностные уравнения потребуются нам в следующем параграфе при описании многочленов Чебышева. Ниже будет проведена аналогия между конечно-разностными уравнениями одного дискретного переменного и обыкновенными дифференциальными уравнениями.

Рассмотрим простейший случай одного линейного уравнения относительно неизвестной функции одного целочисленного аргумента

$$ly = \sum_{i=0}^k a_i(n) y(n+i) = f(n). \quad (1)$$

Это уравнение называется *линейным разностным уравнением*  $k$ -го порядка и является разностным аналогом линейного дифференциального уравнения  $k$ -го порядка

$$\bar{ly} = \sum_{i=0}^k b_i(x) y^{(i)}(x) = f(x). \quad (2)$$

Каждое из уравнений (1) и (2) имеет вид  $Ly = h$ , где  $L$  — линейный оператор. Уравнение  $Ly = 0$  называют *однородным*; формулы

$$y = y(C_1, \dots, C_l, n) \quad \text{или} \quad y = y(C_1, \dots, C_l, x)$$

дают *общее решение* уравнения (1) или (2), если при подстановке значений параметров  $C_i$  можно получить любое решение рассматриваемого уравнения. Если  $v$  — частное решение неоднородного уравнения  $Lv = h$ , то разность  $y - v$  является решением однородного уравнения  $L(y - v) = h - h = 0$ . Таким образом, общее решение неоднородного уравнения представимо в виде суммы частного решения неоднородного уравнения и общего решения однородного уравнения. Решения  $y_1, \dots, y_m$  однородного уравнения  $Ly = 0$  называют *линейно зависимыми* в рассматриваемой области изменения независимого аргумента, если существуют  $C_1, \dots, C_m$ , не все равные нулю, при которых  $C_1 y_1 + \dots + C_m y_m \equiv 0$ . В противном случае эти решения называют *линейно независимыми*. Если функции  $y_i$  являются решениями однородного уравнения  $Ly = 0$ , то любая функция  $\sum_i C_i y_i$  также является решением этого уравнения, поскольку

$$L\left(\sum_i C_i y_i\right) = \sum_i C_i (Ly_i) = 0.$$

Проводимое далее параллельное рассмотрение уравнений (1), (2) подчеркивает общие черты этих уравнений и помогает найти путь исследования уравнения (1) по аналогии с уравнением (2).

Пусть для определенности уравнение (2) рассматривается в области  $x \geq 0$ , а уравнение (1) — в области  $n \geq 0$ .

**Теорема**

Пусть  $b_k(x) \neq 0$  при  $x \geq 0$  и все  $b_l(x)$  непрерывны при  $x \geq 0$  | Пусть  $a_k(n) \neq 0$  при  $n \geq 0$ .

Тогда общее решение однородного уравнения  $Ly = 0$  записывается в виде

$$y = \sum_{i=1}^k C_i y_i,$$

где  $y_1, \dots, y_k$  — линейно независимые решения уравнения  $Ly = 0$ .

**Доказательство**

Согласно теореме существования уравнение  $ly = 0$  имеет решение при любых начальных условиях  $y(0), \dots, y^{(k-1)}(0)$ .

Однородное уравнение  $ly = 0$  можно представить в виде

$$y(n+k) = - \sum_{i=0}^{k-1} \frac{a_i(n)}{a_k(n)} y(n+i). \quad (3)$$

Согласно теореме единственности это решение единственно

Обозначим через  $y_i(x)$  решения уравнения  $ly=0$  при начальных условиях

$$y_i^{(j-1)} = \delta_i^j, \quad i, j = 1, \dots, k.$$

Эти решения образуют линейно независимую систему. В самом деле, если

$$\sum_{i=1}^k C_i y_i(x) \equiv 0,$$

то при  $j = 1, \dots, k$  имеем,

$$\begin{aligned} 0 &= \sum_{i=1}^k C_i y_i^{(j-1)}(0) = \\ &= \sum_{i=1}^k C_i \delta_i^j = C_j. \end{aligned}$$

Следовательно, в случае

$$\sum_{i=1}^k C_i y_i \equiv 0$$

все  $C_i$  обязательно равны нулю, поэтому функции  $y_1, \dots, y_k$  линейно независимы.

Пусть  $y(x)$  — какое-либо решение уравнения  $ly=0$ . Функция

$$z(x) = \sum_{j=1}^k y^{(j-1)}(0) y_j(x)$$

является решением этого уравнения при начальных условиях

$$y(0), \dots, y^{(k-1)}(0).$$

Если мы зададимся  $y(0), \dots, y(k-1)$ , то из (3) сможем вычислить последовательно  $y(k), y(k+1), \dots$ . Таким образом, при любых  $y(0), \dots, y(k-1)$  уравнение  $l(y) = 0$  имеет решение.

Это решение единственно, поскольку значения любого решения удовлетворяют уравнению (3), а из этого уравнения значения  $y(k), y(k+1), \dots$  определяются однозначно.

Обозначим через  $y_i(n)$  решения уравнения  $ly=0$  при начальных условиях

$$y_i(j-1) = \delta_i^j, \quad i, j = 1, \dots, k.$$

$$\sum_{i=1}^k C_i y_i(n) \equiv 0,$$

то при  $j = 1, \dots, k$  имеем

$$\begin{aligned} 0 &= \sum_{i=1}^k C_i y_i(j-1) = \\ &= \sum_{i=1}^k C_i \delta_i^j = C_j. \end{aligned}$$

Пусть  $y(n)$  — какое-либо решение уравнения  $ly=0$ . Функция

$$z(n) = \sum_{j=1}^k y(j-1) y_j(n)$$

$$y(0), \dots, y(k-1).$$

Вследствие единственности решения уравнения

$$\begin{array}{l|l} \bar{ly} = 0 \text{ имеем} & ly = 0 \text{ имеем} \\ y(x) = \sum_{j=1}^k y^{(j-1)}(0) y_j(x) & y(n) = \sum_{j=1}^k y(j-1) y_j(n). \end{array}$$

Теорема доказана.

Далее в курсе дифференциальных уравнений устанавливается следующий факт. Если известны  $k$  линейно независимых решений однородного уравнения  $ly = 0$ , то нахождение решения неоднородного уравнения (2) сводится к решению уравнений

$$dC_j/dx = g_j(x), \quad (4)$$

где  $g_j(x)$  — известные функции, т. е. к отысканию квадратур. Точно так же в случае, когда известны  $k$  линейно независимых решений однородного уравнения  $ly = 0$ , нахождение решения неоднородного уравнения сводится к решению аналогичных (4) разностных уравнений

$$C_j(n+1) - C_j(n) = g_j(n),$$

где  $g_j(n)$  — известные функции.

Перейдем к рассмотрению уравнений с постоянными коэффициентами

$$\begin{array}{l|l} \bar{ly} = \sum_{i=0}^k b_i y^{(i)}(x) = & ly = \sum_{i=0}^k a_i y(n+i) = \\ = f(x), \quad b_k \neq 0, \quad (5) & = f(n), \quad a_k \neq 0, \quad (6) \end{array}$$

и соответствующих однородных уравнений

$$\bar{ly} = \sum_{i=0}^k b_i y^{(i)}(x) = 0. \quad (7) \quad \left| \quad ly = \sum_{i=0}^k a_i y(n+i) = 0. \quad (8)\right.$$

Займемся отысканием частных решений однородного уравнения.

Подставляя в (7) предполагаемый вид частного решения  $\exp\{\lambda x\}$ , получаем уравнение

$$\left( \sum_{i=0}^k b_i \lambda^i \right) \exp\{\lambda x\} = 0.$$

В случае уравнения (8) функцию  $\exp\{\lambda n\}$  удобно записывать в виде  $\mu^n$ ,  $\mu = \exp \lambda$ . Подставляя ее в (8), получаем уравнение

$$\left( \sum_{i=0}^k a_i \mu^i \right) \mu^n = 0.$$

Таким образом, каждому корню уравнения

$$\sum_{i=0}^k b_i \lambda^i = 0, \quad (9) \quad \left| \quad \sum_{i=0}^k a_i \mu^i = 0, \quad (10)\right.$$

называемому *характеристическим*, соответствует частное решение  $\exp\{\lambda x\}$ .

$\mu^n$ .

Если все корни характеристического уравнения простые, мы получаем  $k$  различных решений. Покажем, что каждому  $s$ -кратному корню характеристического уравнения соответствуют  $s$  различных решений однородного уравнения

$$\exp\{\lambda x\}, x \exp\{\lambda x\}, \dots \quad \left| \quad \mu^n, C_1^n \mu^{n-1}, \dots, C_n^{s-1} \mu^{n-s+1}.\right.$$

Пусть для определенности  $\lambda_1 = \dots = \lambda_s, \mu_1 = \dots = \mu_s$ . Разложим на множители характеристический многочлен

$$\sum_{i=0}^k b_i \lambda^i = b_k \prod_{j=1}^k (\lambda - \lambda_j) \quad \left| \quad \sum_{i=0}^k a_i \mu^i = a_k \prod_{j=1}^k (\mu - \mu_j).\right.$$

Зададимся действительным параметром  $\varepsilon > 0, \varepsilon \rightarrow 0$ .

Возьмем  $\lambda_{j\varepsilon}$  такие, что:

а)  $\lambda_{j\varepsilon}$  различны при  $j = 1, \dots, s;$

б) стремятся к  $\lambda_j$  при  $\varepsilon \rightarrow 0$  для всех  $j \leq k$ .

Возьмем  $\mu_{j\varepsilon}$  такие, что:

а)  $\mu_{j\varepsilon}$  различны при  $j = 1, \dots, s;$

б) стремятся к  $\mu_j$  при  $\varepsilon \rightarrow 0$  для всех  $j \leq k$ .

Образует характеристические уравнения, соответствующие этим корням:

$$0 = b_k \prod_{j=1}^k (\lambda - \lambda_{j\varepsilon}) = \sum_{i=0}^k b_{i\varepsilon} \lambda^i \quad \left| \quad 0 = a_k \prod_{j=1}^k (\mu - \mu_{j\varepsilon}) = \sum_{i=0}^k a_{i\varepsilon} \mu^i.\right.$$

Ясно, что

$$b_{i\varepsilon} \rightarrow b_i$$

$$a_{i\varepsilon} \rightarrow a_i \quad (11)$$

при  $\varepsilon \rightarrow 0$ .

Этим характеристическим уравнениям соответствуют уравнения

$$\sum_{i=0}^k b_{i\varepsilon} y_\varepsilon^{(i)}(x) = 0. \quad (12) \quad \left| \quad \sum_{i=0}^k a_{i\varepsilon} y_\varepsilon(n+i) = 0. \quad (13)\right.$$

Пусть при  $\varepsilon > 0$  мы можем указать решения

$y_\varepsilon(x)$  уравнения (12) такие, что при любом  $x \geq 0$  существует предел

$$\lim_{\varepsilon \rightarrow 0} y_\varepsilon(x) = Y(x),$$

$y_\varepsilon(n)$  уравнения (13) такие, что при любом  $n \geq 0$  существует предел

$$\lim_{\varepsilon \rightarrow 0} y_\varepsilon(n) = Y(n).$$

причем  $y_\varepsilon(x)$  сходится к  $Y(x)$  равномерно вместе со всеми производными до порядка  $k$  включительно на любом конечном отрезке  $[x_1, x_2]$ .

Переходя к пределу в (13) с учетом (11), получаем, что предельная функция  $Y(n)$  удовлетворяет уравнению (8).

Переходя к пределу в (12) с учетом (11), получаем, что предельная функция  $Y(x)$  удовлетворяет уравнению (7).

Построим такие последовательности  $y_\varepsilon(x)$  и  $y_\varepsilon(n)$ , которые будут сходиться к частным решениям (7), (8), соответствующим кратным корням. При проведении этих построений удобно использовать разделенные разности. Рассмотрим сначала случай двукратного корня. Положим

$$y_{2\varepsilon}(x) = \exp\{\lambda x\} (\lambda_{1\varepsilon}; \lambda_{2\varepsilon}) = \frac{\exp\{\lambda_{2\varepsilon}x\} - \exp\{\lambda_{1\varepsilon}x\}}{\lambda_{2\varepsilon} - \lambda_{1\varepsilon}} \quad \left| \quad y_{2\varepsilon}(n) = (\mu^n)(\mu_{1\varepsilon}; \mu_{2\varepsilon}) = \frac{\mu_{2\varepsilon}^n - \mu_{1\varepsilon}^n}{\mu_{2\varepsilon} - \mu_{1\varepsilon}} \right.$$

Эти функции являются решениями соответственно уравнений (12), (13). Запишем их в виде

$$y_{2\varepsilon}(x) = x \exp\{\lambda_{1\varepsilon}x\} \times \frac{\exp\{(\lambda_{2\varepsilon} - \lambda_{1\varepsilon})x\} - 1}{(\lambda_{2\varepsilon} - \lambda_{1\varepsilon})x} \quad \left| \quad y_{2\varepsilon}(n) = \mu_{2\varepsilon}^{n-1} + \mu_{2\varepsilon}^{n-2}\mu_{1\varepsilon} + \dots \dots + \mu_{1\varepsilon}^{n-1} \right. \quad (14)$$

Переходя к пределу при  $\varepsilon \rightarrow 0$ , получим

$$y_{2\varepsilon}(x) \rightarrow x \exp\{\lambda_1 x\} \quad \left| \quad y_{2\varepsilon}(n) \rightarrow \mu_1^{n-1} + \dots \dots + \mu_1^{n-1} = n\mu_1^{n-1} \right.$$

Мы получили второе линейно независимое решение, соответствующее двукратному корню.

Случай корня высшей кратности рассмотрим лишь для уравнения (1). Аналогично (14) непосредственно устанавливается, что

$$y_{q\varepsilon} = \mu^n (\mu_{1\varepsilon}; \dots; \mu_{q\varepsilon}) = \sum_{n_1 + \dots + n_q = n+1-q} \mu_{1\varepsilon}^{n_1} \mu_{2\varepsilon}^{n_2} \dots \mu_{q\varepsilon}^{n_q}$$

Общее число слагаемых равно  $C_n^{q-1}$ , поэтому

$$y_{q\varepsilon} \rightarrow Y_q(n) = C_n^{q-1} \mu^{n+1-q}$$

Поскольку в случае  $s$ -кратного корня можно взять  $q = 1, \dots, s$ , то получилось  $s$  частных решений

$$Y_1(n) = \mu^n, \quad Y_2(n) = C_n^1 \mu^{n-1}, \dots, \quad Y_s(n) = C_n^{s-1} \mu^{n+1-s} \quad (15)$$

**Задача 1.** Доказать, что совокупность частных решений (15), соответствующих корням характеристического уравнения (10), образует фундаментальную систему (т.е. они линейно независимы и решение (8) может быть получено как линейная комбинация таких решений).

**Задача 2.** Пусть  $P_{s-1}(n)$  — произвольный многочлен степени  $s-1$ . Доказать, что функция  $P_{s-1}(n)\mu^n$  записывается как линейная комбинация функций (15)

$$P_{s-1}(n)\mu^n = \sum_{j=1}^s C_j Y_j(n)$$



Таким образом, вместо системы решений (15) можно взять систему решений

$$Y_1(n) = \mu_1^n, \quad Y_2(n) = n\mu_1^n, \quad \dots, \quad Y_s(n) = n^{s-1}\mu_1^n.$$

Рассмотрим теперь разностное уравнение

$$\sum_{i=0}^k a_i y(n+i) = \left( \sum_{j=0}^{m-1} C_j n^j \right) \sigma^n. \quad (16)$$

Пусть  $\sigma$  является корнем характеристического уравнения (10) кратности  $s$ ; в частности, если  $\sigma$  не является корнем этого уравнения, то  $s=0$ . Можно показать, что уравнение (16) имеет частное решение вида

$$\left( \sum_{j=s}^{s+m-1} d_j n^j \right) \sigma^n;$$

коэффициенты  $d_j$  могут быть найдены методом неопределенных коэффициентов.

## § 8. Многочлены Чебышева

Рассматриваемые ниже многочлены Чебышева играют фундаментальную роль в теории и практике использования численных методов. С их помощью решается значительная часть задач оптимизации свойств вычислительных алгоритмов. Запись многочленов в традиционной форме часто приводит к большому влиянию вычислительной погрешности, и в этих случаях их целесообразнее записывать в виде линейных комбинаций многочленов Чебышева.

Многочлены Чебышева  $T_n(x)$ , где  $n \geq 0$ , определяются соотношениями

$$\begin{aligned} T_0(x) &= 1, & T_1(x) &= x, \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x) \quad \text{при } n > 0. \end{aligned} \quad (1)$$

Пользуясь рекуррентной формулой (1), получаем, например,

$$\begin{aligned} T_2(x) &= 2x^2 - 1, & T_3(x) &= 4x^3 - 3x, \\ T_4(x) &= 8x^4 - 8x^2 + 1, & T_5(x) &= 16x^5 - 20x^3 + 5x, \dots \end{aligned}$$

Старший член многочлена  $T_{n+1}(x)$  получается из старшего члена многочлена  $T_n(x)$  умножением на  $2x$ , и, следовательно, старший член в  $T_n(x)$  при  $n > 0$  есть  $2^{n-1}x^n$ .

Все многочлены  $T_{2n}(x)$  являются четными функциями, а  $T_{2n+1}(x)$  — нечетными.

При  $n=0$  это утверждение верно. Предположив его справедливость при некотором  $n$ , мы получим, что  $2xT_{2n+1}(x)$  — четная функция и, вследствие (1),  $T_{2n+2}(x)$  — тоже четная функция. Тогда  $2xT_{2n+2}(x)$  и  $T_{2n+3}(x)$ , вследствие (1), — нечетные функции.

При любом  $\theta$  имеем

$$\cos((n+1)\theta) = 2 \cos \theta \cos n\theta - \cos((n-1)\theta).$$

Полагая  $\theta = \arccos x$ , получим

$$\cos((n+1)\arccos x) = 2x \cos(n\arccos x) - \cos((n-1)\arccos x).$$

Функция  $\cos(n\arccos x)$  удовлетворяет тому же разностному уравнению (1) по переменной  $n$ , что и  $T_n(x)$ . Начальные условия при  $n=0$  и  $n=1$  одни и те же:

$$\cos(0 \cdot \arccos x) = 1 = T_0(x),$$

$$\cos(1 \cdot \arccos x) = x = T_1(x);$$

поэтому при всех  $n$

$$T_n(x) = \cos(n\arccos x). \quad (2)$$

Следовательно,

$$|T_n(x)| \leq 1 \quad \text{при} \quad |x| \leq 1. \quad (3)$$

Не нужно думать, что  $|T_n(x)| \leq 1$  при всех вещественных  $x$ . Если  $|x| > 1$ , то  $\arccos x$  не является действительным числом, а косинус такого числа больше 1.

Рекуррентное соотношение (1) является разностным уравнением; ему соответствует характеристическое уравнение

$$\mu^2 - 2x\mu + 1 = 0;$$

отсюда

$$\mu_{1,2} = x \pm \sqrt{x^2 - 1}.$$

При  $x \neq \pm 1$  корни простые, поэтому

$$T_n(x) = c_1 \mu_1^n + c_2 \mu_2^n.$$

Из начальных условий  $T_0(x) = 1$ ,  $T_1(x) = x$  получаем  $c_1 = c_2 = 1/2$ ; таким образом,

$$T_n(x) = \frac{(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n}{2}. \quad (4)$$

Из уравнения

$$T_n(x) = \cos(n\arccos x) = 0$$

получаем, что

$$x_m = \cos\left(\frac{\pi(2m-1)}{2n}\right), \quad m = 1, \dots, n,$$

— нули  $T_n(x)$ . Вследствие (2), (3) точками экстремума  $T_n(x)$  на  $[-1, 1]$  будут точки, где  $|T_n(x)| = 1$ . Решая это уравнение, получим

$$x_{(m)} = \cos\left(\frac{\pi m}{n}\right), \quad m = 0, \dots, n,$$

причем

$$T_n(x_{(m)}) = \cos \pi m = (-1)^m.$$

## Многочлены

$$\bar{T}_n(x) = 2^{1-n} T_n(x) = x^n + \dots$$

называют *многочленами, наименее уклоняющимися от нуля*. Это определение объясняется следующим свойством.

*Лемма.* Если  $P_n(x)$  — многочлен степени  $n$  со старшим коэффициентом 1, то

$$\max_{[-1, 1]} |P_n(x)| \geq \max_{[-1, 1]} |\bar{T}_n(x)| = 2^{1-n}. \quad (5)$$

*Доказательство.* Предположим противное. Многочлен  $\bar{T}_n(x) - P_n(x)$  имеет степень  $n - 1$ ; в то же время

$$\text{sign}(\bar{T}_n(x_{(m)}) - P_n(x_{(m)})) = \text{sign}((-1)^m 2^{1-n} - P_n(x_{(m)})) = (-1)^m,$$

так как, согласно предположению,  $|P_n(x_{(m)})| < 2^{1-n}$  при всех  $m$ . Таким образом, между каждыми двумя точками  $x_{(m)}, x_{(m+1)}$  многочлен  $\bar{T}_n(x) - P_n(x)$  меняет знак. Многочлен  $\bar{T}_n(x) - P_n(x)$  степени  $n - 1$ , отличный от нуля (поскольку он отличен от нуля в точках  $x_{(m)}$ ), имеет  $n$  различных нулей. Мы пришли к противоречию.

**Задача 1.** Доказать более сильное утверждение: если

$$P_n(x) = x^n + \dots \neq \bar{T}_n(x),$$

то

$$\max_{[-1, 1]} |P_n(x)| > 2^{1-n}.$$

Линейной заменой переменных  $x' = \frac{b+a}{2} + \frac{b-a}{2}x$  отрезок  $[-1, 1]$  можно перевести в заданный отрезок  $[a, b]$ . Многочлен  $\bar{T}_n(x)$  при этом преобразуется в многочлен  $\bar{T}_n\left(\frac{2x - (b+a)}{b-a}\right)$  со старшим коэффициентом  $(2/(b-a))^n$ . В соответствии с леммой можно утверждать, что многочлен

$$\bar{T}_n^{[a, b]}(x) = (b-a)^n 2^{1-2n} T_n\left(\frac{2x - (b+a)}{b-a}\right)$$

со старшим коэффициентом 1 является многочленом, наименее уклоняющимся от нуля на отрезке  $[a, b]$ . Это означает, что для любого многочлена  $P_n(x)$  степени  $n$  со старшим коэффициентом 1 справедливо неравенство

$$\max_{[a, b]} |P_n(x)| \geq \max_{[a, b]} |\bar{T}_n^{[a, b]}(x)| = (b-a)^n 2^{1-2n}. \quad (6)$$

Нетрудно проверить, что нулями многочлена  $\bar{T}_n^{[a, b]}(x)$  являются точки

$$x_m = \frac{b+a}{2} + \frac{b-a}{2} \cos\left(\frac{\pi(2m-1)}{2n}\right), \quad m = 1, \dots, n.$$

Многочлены  $\tilde{T}_0(x) = 1/\sqrt{2}$ ,  $\tilde{T}_n(x) = T_n(x)$  при  $n \geq 1$  образуют на  $[-1, 1]$  ортонормированную систему с весом  $2/(\pi\sqrt{1-x^2})$ .

Проверим свойство ортонормированности этих функций. После замены  $x = \cos \theta$  имеем

$$\int_{-1}^1 \frac{2T_n(x) T_m(x)}{\pi \sqrt{1-x^2}} dx = \frac{1}{\pi} \int_{-\pi}^{\pi} \cos n\theta \cos m\theta d\theta = \\ = \frac{1}{2\pi} \int_{-\pi}^{\pi} (\cos((n-m)\theta) + \cos((n+m)\theta)) d\theta = \delta_m^n + \delta_m^{-n}.$$

Второе слагаемое обращается в нуль при  $n, m \geq 0$ , если  $n^2 + m^2 \neq 0$ . Отсюда вытекает требуемое равенство

$$\int_{-1}^1 \frac{2\tilde{T}_n(x) \tilde{T}_m(x)}{\pi \sqrt{1-x^2}} dx = \delta_m^n.$$

Пусть многочлены Чебышева вычисляются по рекуррентной формуле (1). В процессе реальных вычислений вместо значений  $T_n(x)$  получаем приближенные значения  $T_n^*$ , удовлетворяющие соотношениям

$$T_0^* = 1, \quad T_1^* = x + \delta_1, \\ T_{n+1}^* = 2xT_n^* - T_{n-1}^* + \delta_{n+1},$$

где  $\delta_k$  — погрешности, вызванные округлениями.

**Задача 2.** Получить представление погрешности

$$T_N^*(x) - T_N(x) = \sum_{k=1}^N \delta_k \frac{\sin((N+1-k) \arccos x)}{\sqrt{1-x^2}}.$$

**Задача 3.** Получить, используя решение задачи 2, оценку погрешности

$$|T_N^*(x) - T_N(x)| \leq \max_{1 \leq k \leq N} |\delta_k| \cdot N \cdot \min\left(N, \frac{1}{\sqrt{1-x^2}}\right). \quad (7)$$

**Задача 4.** Проверить, что в нулях многочленов Чебышева  $x_m = \cos\left(\frac{\pi(2m-1)}{2N}\right)$  справедливо равенство

$$|T_N'(x_m)| = \frac{N}{\sqrt{1-x_m^2}}.$$

Отсюда заключаем, что при задании  $x$  в окрестности этих нулей с погрешностью  $\delta$  погрешность вычисления значения  $T_n(x)$  будет величиной, близкой к  $\frac{n}{\sqrt{1-x_m^2}} \delta$ ; это означает, что оценка (7) не может быть существенно улучшена.

**Задача 5.** Доказать, что векторы, образованные значениями многочленов  $\tilde{T}_n(x)$ ,  $n \leq N$ , где  $N$  — некоторое фиксированное число, в нулях  $T_N(x)$

образуют некоторую ортогональную систему, а именно

$$\frac{2}{N} \sum_{j=1}^N \tilde{T}_n \left( \frac{\pi(2j-1)}{2N} \right) \tilde{T}_m \left( \frac{\pi(2j-1)}{2N} \right) = \delta_n^m \quad (8)$$

при  $0 \leq n, m < N-1$ .

### § 9. Минимизация оценки остаточного члена интерполяционной формулы

Пусть функция  $f(x)$  приближается на  $[a, b]$  с помощью интерполяционного многочлена степени  $n-1$  с узлами интерполяции  $x_1, \dots, x_n \in [a, b]$ . Согласно (3.1) имеем

$$f(x) - L_n(x) = \frac{f^{(n)}(\xi) \omega_n(x)}{n!},$$

где  $\xi \in [a, b]$ , если  $x \in [a, b]$ . Отсюда следует оценка погрешности интерполяции

$$\|f - L_n\| \leq \frac{\|f^{(n)}\| \cdot \|\omega_n\|}{n!}. \quad (1)$$

Здесь  $\|\cdot\|$  есть обозначение равномерной нормы:  $\|g(x)\| = \sup_{|a, b|} |g(x)|$ .

Займемся минимизацией правой части оценки (1) за счет выбора узлов  $x_1, \dots, x_n$ . Многочлен  $\omega_n(x) = (x-x_1) \dots (x-x_n)$  имеет старший коэффициент 1, поэтому  $\|\omega_n\| \geq (b-a)^n 2^{1-2n}$  согласно (8.6). Если взять в качестве узлов интерполяции

$$x_m = \frac{b+a}{2} + \frac{b-a}{2} \cos \left( \frac{\pi(2m-1)}{2n} \right), \quad m = 1, \dots, n, \quad (2)$$

то

$$\omega_n(x) = (b-a)^n 2^{1-2n} T_n \left( \frac{2x - (b+a)}{b-a} \right)$$

и

$$\|\omega_n\| = (b-a)^n 2^{1-2n}.$$

Следовательно, при таком расположении узлов справедлива наилучшая из оценок, которая может быть получена как следствие оценки (1):

$$\|f - L_n\| \leq \frac{\|f^{(n)}\| (b-a)^n 2^{1-2n}}{n!}. \quad (3)$$

При получении оценки (1) максимум произведения заменен на произведение максимумов сомножителей. Поэтому может возникнуть надежда получить оценку погрешности, лучшую, чем

(3). Однако это не так. Если  $f(x) = a_n x^n + \dots + a_0$  — многочлен степени  $n$ , то

$$f^{(n)}(\xi) = a_n \cdot n! = \text{const},$$

поэтому неравенство (1) превращается в равенство; тогда, вследствие (8.6), при любых узлах интерполяции имеем

$$\|f - L_n\| \geq \frac{\|f^{(n)}\| (b-a)^n 2^{1-2n}}{n!} = |a_n| (b-a)^n 2^{1-2n}.$$

Как уже отмечалось, важной проблемой вычислительной математики является проблема оптимизации методов решения задач некоторого класса. Общая постановка ее такова.

Задается некоторый класс  $P$  решаемых задач  $p$ . Задается некоторое множество  $M$  методов решения. Пусть  $e(p, m)$  — погрешность метода  $m$  при решении задачи  $p$ . Величину

$$e(P, m) = \sup_{p \in P} e(p, m)$$

называют *погрешностью метода на классе задач  $P$* . Величину

$$E(P, M) = \inf_{m \in M} e(P, m)$$

называют *оптимальной оценкой погрешности методов из множества  $M$  на классе задач  $P$* . Если существует метод  $m \in M$ , на котором эта оценка достигается, т. е.  $E(P, M) = e(P, m)$ , то такой метод называют *оптимальным*.

Полученное нами решение задачи об оптимизации узлов распределения интерполяционной формулы можно сформулировать в описанных выше терминах.

Пусть  $P$  — множество задач приближения функций, определенных на  $[a, b]$  и удовлетворяющих условию  $|f^{(n)}(x)| \leq A_n$ . Пусть  $M$  — множество методов приближения, состоящих в том, что функция заменяется ее интерполяционным многочленом  $L_n(x)$  по совокупности узлов  $x_1, \dots, x_n$ ; таким образом, метод решения  $m$  определяется заданием узлов интерполяции  $x_1, \dots, x_n$ . Наконец, пусть мера погрешности  $e(p, m) = \|f - L_n\|$ . Согласно (1) имеем

$$e(P, m) \leq \frac{A_n \|\omega_n\|}{n!}.$$

С другой стороны, для задачи приближения многочлена

$$P_{n+1}(x) = \frac{A_n}{n!} x^n + \dots,$$

относящейся к рассматриваемому классу, имеем

$$e(p, m) = \frac{A_n \|\omega_n\|}{n!}.$$

Следовательно,

$$e(P, m) = \frac{A_n \|\omega_n\|}{n!}.$$

Как мы видели выше,

$$E(P, M) = \inf_m e(P, m) = \inf_{x_1, \dots, x_n} \left( \frac{A_n \|\omega_n\|}{n!} \right) = \frac{A_n (b-a)^n 2^{1-2n}}{n!}.$$

Таким образом, способ интерполяции по узлам многочлена Чебышева (2) является оптимальным в рассматриваемом смысле.

В заключение произведем сравнение оценки (3) с оценкой погрешности разложения функции в ряд Тейлора. Согласно результатам § 6 отрезок ряда Тейлора

$$t_n(x) = \sum_{j=0}^{n-1} \frac{f^{(j)}((a+b)/2)}{j!} \left(x - \frac{a+b}{2}\right)^j$$

совпадает с интерполяционным многочленом Лагранжа при единственном,  $n$ -кратном узле интерполяции  $(a+b)/2$ . Поэтому, естественно, при оптимальном распределении узлов интерполяции (2) мы должны иметь лучшую оценку. В самом деле, оценка

$$\|f - t_n\| \leq \frac{\|f^{(n)}\| (b-a)^n 2^{-n}}{n!}$$

погрешности отрезка ряда Тейлора уступает оценке (3) в  $2^{n-1}$  раз.

Приведем для сведения оценки погрешности интерполяции с узлами в нулях многочленов Чебышева. Для простоты возьмем случай, когда  $[a, b] = [-1, 1]$ .

Оценка 1. Если  $\sup |f^{(m)}(x)| < \infty$ , то  $\|f(x) - L_n(x)\| = O(n^{-m} \ln n)$  при  $n \rightarrow \infty$ .

Оценка 2. Если  $f(x)$  аналитична в каждой точке отрезка  $[-1, 1]$ , то  $\|f(x) - L_n(x)\| = O(q^n)$ , где  $q < 1$ .

Последнюю оценку можно конкретизировать. Пусть  $f(z)$ ,  $z = x + iy$ , — функция, аналитическая в эллипсе на плоскости  $(x, y)$  с фокусами в точках  $-1, 1$ ; тогда  $\|f(x) - L_n(x)\| = O(c^{-n})$ , где  $c > 1$  — сумма полуосей этого эллипса.

Таким образом, при интерполяции по узлам многочлена Чебышева погрешность автоматически уменьшается, если алгоритм применяется к более гладкой функции. Такие алгоритмы иногда называют *ненасыщаемыми*.

Если узлы интерполяции распределены существенно иначе, например равномерно, то даже для аналитической функции погрешность интерполяции может стремиться к бесконечности с ростом числа узлов. Например, для функции  $f(x) = (1 + 25x^2)^{-1}$  имеет место соотношение

$$\|f(x) - L_n(x)\| \geq Aa^n, \quad A > 0, \quad a > 1.$$

**Задача 1.** Пользуясь формулой (8.8), показать, что интерполяционный многочлен с узлами в нулях многочленов Чебышева записывается в виде

$$L_n(x) = \sum_{j=0}^{n-1} a_j \tilde{T}_j(x),$$

$$a_j = \frac{2}{n} \sum_{k=1}^n f(x_j) \tilde{T}_j\left(\frac{\pi(2k-1)}{2n}\right).$$

Такая запись интерполяционного многочлена позволяет быстро и с малой чувствительностью по отношению к вычислительной погрешности вычислять его значения (см. § 4.8).

## § 10. Конечные разности

Пусть узлы таблицы  $x_i$  расположены на равных расстояниях:  $x_i = x_0 + ih$ ,  $f_i$  — соответствующие значения функции; величину  $h$  называют *шагом таблицы*.

Разности  $f_{i+1} - f_i$  называют *разностями первого порядка*. В зависимости от обстановки эту величину обозначают:  $\Delta f_i$  — *разность вперед*,  $\nabla f_{i+1}$  — *разность назад* или  $\delta f_{i+1/2} = f_{i+1/2}^1$  — *центральная разность*. Таким образом,

$$f_{i+1} - f_i = \Delta f_i = \nabla f_{i+1} = \delta f_{i+1/2} = f_{i+1/2}^1. \quad (1)$$

Разности высшего порядка образуют при помощи рекуррентных соотношений:

$$\begin{aligned} \Delta^m f_i &= \Delta(\Delta^{m-1} f_i) = \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i, \\ \nabla^m f_i &= \nabla(\nabla^{m-1} f_i) = \nabla^{m-1} f_i - \nabla^{m-1} f_{i-1}, \\ \delta^m f_i &= \delta(\delta^{m-1} f_i) = \delta^{m-1} f_{i+1/2} - \delta^{m-1} f_{i-1/2}, \\ f_i^m &= f_{i+1/2}^{m-1} - f_{i-1/2}^{m-1}. \end{aligned}$$

Таблицу разностей обычно располагают в виде

$x$	$f$	$f^1$	$f^2$	$f^3$
$x_0$	$f_0$			
$x_1$	$f_1$	$f_{1/2}^1$	$f_1^2$	
$x_2$	$f_2$	$f_{3/2}^1$	$f_2^2$	$f_{3/2}^3$
$x_3$	$f_3$	$f_{5/2}^1$	$f_3^2$	$f_{5/2}^3$
$x_4$	$f_4$	$f_{7/2}^1$		



В некоторых интерполяционных формулах наряду с упомянутыми выше величинами используются средние арифметические двух последовательных величин одного и того же столбца:

$$\begin{aligned} \mu f_i^m &= (f_{i+1/2}^m + f_{i-1/2}^m)/2 \quad \text{при } m \text{ нечетном,} \\ \mu f_{i+1/2}^m &= (f_{i+1}^m + f_i^m)/2 \quad \text{при } m \text{ четном.} \end{aligned}$$

Лемма 1. Разности  $m$ -го порядка выражаются через значения функции по формуле

$$\Delta^m f_i = \sum_{j=0}^m (-1)^j C_m^j f_{i+m-j}, \quad (2)$$

здесь  $C_m^j$  — коэффициенты бинома Ньютона.

Доказательство проводим методом индукции. При  $m = 1$  соотношение (2) выполняется согласно (1). Пусть оно доказано при  $m = l$ . Имеем

$$\Delta^{l+1} f_i = \Delta^l f_{i+1} - \Delta^l f_i = \sum_{j=0}^l (-1)^j C_l^j f_{i+1+l-j} - \sum_{j=0}^l (-1)^j C_l^j f_{i+l-j}.$$

Собирая коэффициенты при одинаковых  $f_k$  и пользуясь равенством

$$C_l^j + C_l^{j+1} = C_{l+1}^{j+1},$$

получим требуемое выражение для величины  $\Delta^{l+1} f_i$ . Лемма доказана.

Из (2) следует, что оператор конечной разности является линейным. В частности, из (2) имеем

$$\begin{aligned} \Delta^2 f_i &= f_{i+2} - 2f_{i+1} + f_i, \\ \Delta^3 f_i &= f_{i+3} - 3f_{i+2} + 3f_{i+1} - f_i, \\ \Delta^4 f_i &= f_{i+4} - 4f_{i+3} + 6f_{i+2} - 4f_{i+1} + f_i. \end{aligned}$$

Лемма 2. При  $x_i \equiv x_0 + ih$  справедливо равенство

$$f(x_i; \dots; x_{i+m}) = \frac{f_{i+m/2}^m}{h^m \cdot m!}. \quad (3)$$

Доказательство проводим по индукции. При  $m = 1$  имеем

$$f(x_i; x_{i+1}) = \frac{f_{i+1} - f_i}{x_{i+1} - x_i} = \frac{f_{i+1/2}^1}{h}.$$

Предположив, что соотношение (3) верно при всех  $m \leq l$ , имеем

$$\begin{aligned} f(x_i; \dots; x_{i+l+1}) &= \frac{f(x_{i+1}; \dots; x_{i+l+1}) - f(x_i; \dots; x_{i+l})}{x_{i+l+1} - x_i} = \\ &= \frac{f_{i+1+1/2}^l - f_{i+1/2}^l}{h^l l! h (l+1)} = \frac{f_{i+(l+1)/2}^{l+1}}{h^{l+1} (l+1)!}; \end{aligned}$$

таким образом, (3) справедливо и при  $m = l + 1$ . Лемма доказана.

Согласно (5.4) имеем  $f(x_i; \dots; x_{i+m}) = \frac{f^{(m)}(\xi)}{m!}$ , где  $x_i \leq \xi \leq x_{i+m}$ . Сопоставляя это равенство с (3), получаем

$$\Delta^m f_i = \nabla^m f_{i+m} = \delta^m f_{i+m/2} = f_{i+m/2}^m = h^m f^{(m)}(\xi). \quad (4)$$

*Следствие. Конечные разности  $n$ -го порядка от многочлена степени  $n$  постоянны, а разности любого более высокого порядка равны нулю.*

Рассмотрим влияние погрешности какого-либо значения  $f_i$  на конечные разности различных порядков; пусть вместо  $f_i$  стоит  $f_i + \varepsilon$ . Тогда имеем таблицу разностей

$$\begin{array}{ccc} & f_{i-2} & \\ & & f_{i-3/2}^1 \\ f_{i-1} & & f_{i-1}^2 + \varepsilon \\ & f_{i-1/2}^1 + \varepsilon & \\ f_i + \varepsilon & & f_i^2 - 2\varepsilon \\ & f_{i+1/2}^1 - \varepsilon & \\ f_{i+1} & & f_{i+1}^2 + \varepsilon \\ & f_{i+3/2}^1 & \\ f_{i+2} & & \end{array}$$

Мы видим, что в соответствии с (2) на разности порядка  $m$  погрешность распространяется с коэффициентами  $(-1)^i C_m^i$ . Если функция достаточно гладкая, то ее разности не очень высокого порядка могут оказаться малыми. В то же время на их фоне величины  $C_m^i \varepsilon$  будут выглядеть достаточно большими. Из наблюдений над таблицей разностей можно указать значение функции, содержащее погрешность, и исправить его.

Точно так же можно обнаруживать погрешности, имевшие место при составлении таблицы разностей. Пусть, например,

$$\begin{aligned} f_{1/2}^3 &= 1 \cdot 10^{-5}, & f_{3/2}^3 &= 2 \cdot 10^{-5}, & f_{5/2}^3 &= 12 \cdot 10^{-5}, \\ f_{7/2}^3 &= -23 \cdot 10^{-5}, & f_{9/2}^3 &= 11 \cdot 10^{-5}, & f_{11/2}^3 &= 1 \cdot 10^{-5}. \end{aligned}$$

Если бы какое-либо значение  $f_i$  содержало относительно большую погрешность  $\varepsilon$ , то в третьих разностях это обстоятельство проявилось бы в наличии величин вида  $\varepsilon$ ,  $-3\varepsilon$ ,  $3\varepsilon$ ,  $-\varepsilon$ . В рассматриваемом случае третьи разности практически равны нулю, за исключением  $f_{5/2}^3$ ,  $f_{7/2}^3$ ,  $f_{9/2}^3$ , которые примерно имеют вид  $\varepsilon$ ,  $-2\varepsilon$ ,  $\varepsilon$ , где  $\varepsilon = 11 \cdot 10^{-5}$ . Это наводит на мысль, что была допущена погрешность при вычислении значения  $f_{7/2}^1$ , которая и

имела следствием эти возмущения в третьих разностях. Этот прием широко использовался при ручном счете для устранения случайных погрешностей расчетчика и на первом этапе использования ЭВМ, когда ЭВМ были малонадежны. Существовавшая на первом этапе использования ЭВМ общая рекомендация по устранению ненадежности заключалась в следующем. Задачу предлагалось решить два раза. В случае несовпадения результатов полагалось просчитать задачу повторно, пока результаты двух расчетов не совпадут.

Описанный выше метод исправления таблиц позволяет в такой ситуации уменьшить объем вычислений примерно вдвое.

Пусть требуется составить таблицу какой-либо гладкой функции, каждое вычисление которой обходится очень дорого. Примером могут служить таблицы сложных специальных функций или таблицы дальности стрельбы при различных углах наклона орудия. Вместо того чтобы считать каждое значение дважды, просчитываем сразу всю таблицу, составляем (вручную или с помощью ЭВМ) таблицу разностей и выявляем значения, которые нужно исправить (или повторным расчетом, или описанным выше приемом исправления таблиц).

## § 11. Интерполяционные формулы для таблиц с постоянным шагом

Поскольку таблицы значений функции с постоянным шагом наиболее употребительны, приведем конкретные расчетные формулы для таких таблиц. Если узлы интерполирования выбираются вблизи точки  $x$ , где вычисляется значение функции, то промежуточная точка  $\xi$  в оценке остаточного члена (3.1) также находится вблизи точки  $x$ . Таким образом, величина  $f^{(n)}(\xi)$  изменится не очень сильно при выборе узлов в окрестности точки  $x$ .

Следовательно, решающее влияние на значение погрешности оказывает величина

$$|\omega_n(x)| = \prod_{j=1}^n |x - x_j|,$$

т. е. произведение расстояний от точки  $x$  до узлов интерполирования. Величина  $|\omega_n(x)|$  будет минимальной, если в качестве узлов интерполирования для нахождения  $f(x)$  мы возьмем  $n$  узлов, ближайших к  $x$ . Для этого при четном  $n = 2l$  следует взять по  $l$  узлов справа и слева от точки  $x$ . При нечетном  $n = 2l + 1$  следует взять узел, ближайший к  $x$ , и по  $l$  узлов слева и справа от него. Если точка  $x$  находится вблизи одного из концов таблицы, то это правило несколько изменится.

При интерполировании в начале или конце таблицы принято записывать интерполяционный многочлен в виде так называемых *формул Ньютона* для интерполирования вперед или назад. Пусть  $L_n(x)$  — интерполяционный многочлен Лагранжа по узлам

$x_0, \dots, x_{n-1}$ . Согласно (5.3) имеем

$$L_n(x) = f(x_0) + f(x_0; x_1)(x - x_0) + \dots \\ \dots + f(x_0; \dots; x_{n-1})(x - x_0) \dots (x - x_{n-2}).$$

Произведем замену переменных  $x = x_0 + ht$  и перейдем согласно (10.3) от разделенных разностей к конечным. Получим

$$L_n(x_0 + ht) = f_0 + f_{1/2}^1 t + \dots + f_{(n-1)/2}^{n-1} \frac{t(t-1)\dots(t-(n-2))}{(n-1)!}. \quad (1)$$

Остаточный член (3.1) представится в виде

$$f(x) - L_n(x) = f^{(n)}(\xi) \frac{t(t-1)\dots(t-(n-1)) h^n}{n!}.$$

Формулу (1) называют *интерполяционной формулой Ньютона для интерполирования вперед*. Если мы произведем такую же замену переменных в интерполяционном многочлене  $L_n(x)$  по узлам  $x_0, x_{-1}, \dots, x_{-(n-1)}$ :

$$L_n(x) = f(x_0) + f(x_0; x_{-1})(x - x_0) + \dots \\ \dots + f(x_0; \dots; x_{-(n-1)})(x - x_0) \dots (x - x_{-(n-2)}),$$

то получим *интерполяционную формулу Ньютона для интерполирования назад*

$$L_n(x_0 + ht) = f_0 + f_{-1/2}^1 t + \dots + f_{-(n-1)/2}^{n-1} \frac{t(t+1)\dots(t+(n-2))}{(n-1)!} \quad (2)$$

с остаточным членом

$$f(x) - L_n(x) = f^{(n)}(\xi) \frac{t(t+1)\dots(t+(n-1)) h^n}{n!}.$$

Эти формулы, в частности, широко используются при построении методов решения дифференциальных уравнений. Таблицы конечных разностей так же, как и таблицы разделенных разностей, используются для оценки производных функций. Если  $f^{(n)}(x)$  непрерывна, то справедливо равенство  $\lim_{h \rightarrow 0} \sigma_n^h = M_n$ ;

здесь  $\sigma_n^h = \max_{a \leq x_m \leq b-nh} \left| \frac{\Delta^n f_m}{h^n} \right|$ ;  $M_n = \max_{[a, b]} |f^{(n)}(x)|$ . Поэтому при малых  $h$  можно принять  $M_n \approx \sigma_n^h$ .

Часто приобретает особо важное значение малость степени полиномов, приближающих функцию. Уменьшения степени таких полиномов без потери точности иногда можно достигнуть, образуя линейные комбинации интерполяционных полиномов. Рассмотрим простейший из таких способов приближения функций.

Требуется приблизить функцию на отрезке  $[x_q, x_{q+1}]$  многочленом второй степени. Выпишем интерполяционную формулу Ньютона 3-й степени по узлам  $x_{q-1}, x_q, x_{q+1}, x_{q+2}$ , взяв узлы в последовательности  $x_q, x_{q+1}, x_{q-1}, x_{q+2}$  и в последовательности  $x_{q+1},$

$x_q, x_{q+2}, x_{q-1}$ . Имеем

$$f(x) = P_3^1(x) + r^1(x); \quad (3)$$

$$P_3^1(x) = f(x_q) + f(x_q; x_{q+1})(x - x_q) + \\ + f(x_q; x_{q+1}; x_{q-1})(x - x_q)(x - x_{q+1}) + \\ + f(x_q; x_{q+1}; x_{q-1}; x_{q+2})(x - x_q)(x - x_{q+1})(x - x_{q-1}); \quad (4)$$

$$r^1(x) = \frac{f^{(4)}(\xi_1)}{4!} (x - x_{q-1})(x - x_q)(x - x_{q+1})(x - x_{q+2});$$

$$f(x) = P_3^2(x) + r^2(x),$$

$$P_3^2(x) = f(x_{q+1}) + f(x_{q+1}; x_q)(x - x_{q+1}) + \\ + f(x_{q+1}; x_q; x_{q+2})(x - x_{q+1})(x - x_q) + \\ + f(x_{q+1}; x_q; x_{q+2}; x_{q-1})(x - x_{q+1})(x - x_q)(x - x_{q+2}); \quad (5)$$

$$r^2(x) = \frac{f^{(4)}(\xi_2)}{4!} (x - x_{q-1})(x - x_q)(x - x_{q+1})(x - x_{q+2}).$$

Поскольку интерполяционный многочлен третьей степени, совпадающий с функцией в 4-х узлах, единствен, то

$$P_3^1(x) = P_3^2(x), \quad r^1(x) = r^2(x) \quad \text{и} \quad f^{(4)}(\xi_1) = f^{(4)}(\xi_2).$$

Образует полусумму равенств (4), (5). Так как  $P_3^1(x) = P_3^2(x)$ , в левой части будет стоять многочлен  $P_3^1(x)$ ; при вычислении правой части будем образовывать полусуммы от соответствующих слагаемых; введем обозначения:  $x_{q+1/2} = x_q + h/2$ ,  $\mu f_{q+1/2}^{2n} = (f_q^{2n} + f_{q+1}^{2n})/2$ . Получим

$$P_3^1(x) = \\ = \mu f_{q+1/2} + f_{q+1/2}^1 h^{-1} (x - x_{q+1/2}) + \frac{1}{2} \mu f_{q+1/2}^2 h^{-2} (x - x_{q+1})(x - x_q) + \\ + \frac{1}{6} f_{q+1/2}^3 h^{-3} (x - x_q)(x - x_{q+1})(x - x_{q+1/2}).$$

Обозначив первые три слагаемые в правой части последнего равенства через  $B_2(x)$ , соотношение (3) запишем в виде

$$f(x) = B_2(x) + R(x),$$

$$B_2(x) = \mu f_{q+1/2} + f_{q+1/2}^1 h^{-1} (x - x_{q+1/2}) + \\ + \mu f_{q+1/2}^2 h^{-2} (x - x_q)(x - x_{q+1}), \quad (6)$$

$$R(x) = f_{q+1/2}^3 h^{-3} (x - x_q)(x - x_{q+1})(x - x_{q+1/2}) + \\ + \frac{f^{(4)}(\xi_1)}{24} (x - x_{q-1})(x - x_q)(x - x_{q+1})(x - x_{q+2}). \quad (7)$$

Многочлен  $B_2(x)$  называют *интерполяционным многочленом Бесселя*. Если подходить формально, то этот многочлен второй степени не является интерполяционным, поскольку он совпадает с  $f(x)$  только в точках  $x_q, x_{q+1}$ .

В следующем параграфе будет видно, что использование многочлена Бесселя дает определенные преимущества по сравнению с непосредственным использованием интерполяционного многочлена второй степени.

## § 12. Составление таблиц

Рассмотрим следующую задачу. Требуется построить таблицу значений некоторой функции так, чтобы погрешность при интерполяции значений функции многочленом заданной степени  $m$  не превосходила  $\epsilon$ . В этом случае говорят, что таблица допускает интерполяцию степени  $m$  (с погрешностью  $\epsilon$ ). Таблицы, выпускаемые для широкого круга пользователей, обычно составляются так, чтобы они допускали интерполяцию первой степени, иначе — *линейную интерполяцию*. Примером таких таблиц могут служить таблицы В. М. Брадиса, известные из школьного курса. В дальнейшем рассматриваем случай таблицы с постоянным шагом.

Для вычисления значения  $f(x)$  при помощи такой таблицы берутся узлы  $x_q$  и  $x_{q+1}$  справа и слева от точки  $x$ :  $x_q < x < x_{q+1}$  (они будут ближайшими к  $x$ ); затем  $f(x)$  заменяется интерполяционным многочленом первой степени по этим узлам (для удобства обозначаем  $x = x_q + th$ ):

$$f(x) \approx L_2(x) = f_q + f_{q+1/2}^1 t.$$

Погрешность этой формулы

$$f''(\xi) h^2 \frac{t(t-1)}{2},$$

где

$$x_q \leq \xi \leq x_{q+1}.$$

Эта величина не превосходит  $\epsilon$ , если

$$(\max |f''(\xi)|) h^2 \max_{[0, 1]} \left| \frac{t(t-1)}{2} \right| \leq \epsilon;$$

$\max_{[0, 1]} \left| \frac{t(t-1)}{2} \right|$  достигается при  $t = 1/2$  и равен  $1/8$ . Таким образом, достаточно выполнения условия

$$\max |f''(\xi)| h^2/8 \leq \epsilon. \quad (1)$$

Пусть мы хотим составить или ввести в машину таблицу  $\sin x$  на  $[0, \pi/2]$  так, чтобы погрешность линейной интерполяции не превосходила  $0,5 \cdot 10^{-6}$ . Поскольку  $\max |(\sin x)''| \leq 1$ , то из (1) вытекает требование на шаг таблицы

$$h^2/8 \leq 0,5 \cdot 10^{-6} \text{ или } h \leq 0,002.$$

Часто требование допустимости линейной интерполяции является слишком жестким и вместо него требуют допустимой *квадратичной интерполяции* (т. е. интерполяции многочленом

второй степени). Простейшим случаем квадратичной интерполяции будет интерполяция многочленом Лагранжа по трем узлам. Пусть  $x_q$  — узел, ближайший к  $x$ , т. е.  $|x - x_q| \leq h/2$ . Имеем

$$f(x) \approx L_3(x) = f_q + f_{q+1/2}^1 t + f_q^2 \frac{t(t-1)}{2}. \quad (2)$$

Остаточный член этой формулы

$$f(x) - L_3(x) = f^{(3)}(\xi) h^3 \frac{t(t^2-1)}{3!}.$$

Чтобы таблица допускала квадратичную интерполяцию (2), достаточно выполнения условия

$$\max |f^{(3)}(\xi)| h^3 \max_{|t| \leq 1/2} \left| \frac{t(t^2-1)}{6} \right| \leq \varepsilon.$$

Так как  $\max_{|t| \leq 1/2} \left| \frac{t(t^2-1)}{6} \right| = \frac{1}{16}$ , то это требование на шаг переписывается в виде

$$\max |f^{(3)}(\xi)| h^3 / 16 \leq \varepsilon. \quad (3)$$

В конкретном случае при  $f(x) = \sin x$ ,  $\varepsilon = 0,5 \cdot 10^{-6}$  получаем  $h \leq 0,02$ .

Рассмотрим другой способ замены функции многочленом второй степени. Пусть  $x \in (x_q, x_{q+1})$ ,  $x = x_q + ht$ ; положим

$$f(x) \approx B_2(x) = \mu f_{q+1/2} + f_{q+1/2}^1 (t-1/2) + \mu f_{q+1/2}^2 \frac{t(t-1)}{2}, \quad (4)$$

т. е. заменим  $f(x)$  многочленом Бесселя второй степени (11.6), написанным по узлам  $x_{q-1}$ ,  $x_q$ ,  $x_{q+1}$ ,  $x_{q+2}$ . Согласно (11.7) остаточный член (4) есть

$$f_{q+1/2}^3 \frac{t(t-1)(t-1/2)}{6} + f^{(4)}(\xi) h^4 \frac{t(t^2-1)(t-2)}{24}.$$

Так как  $f_{q+1/2}^3 = h^3 f^{(3)}(\tilde{\xi})$ , то для допустимости интерполяции по формуле (4) достаточно выполнения соотношения

$$\max |f^{(3)}(\xi)| h^3 \max_{0 \leq t \leq 1} \left| \frac{t(t-1)(t-1/2)}{6} \right| + \max |f^{(4)}(\xi)| h^4 \max_{0 \leq t \leq 1} \left| \frac{t(t^2-1)(t-2)}{24} \right| \leq \varepsilon.$$

Поскольку

$$\max_{0 \leq t \leq 1} \left| \frac{t(t-1)(t-1/2)}{6} \right| = \frac{1}{72\sqrt{3}}, \quad \max_{0 \leq t \leq 1} \left| \frac{t(t^2-1)(t-2)}{24} \right| = \frac{3}{128},$$

то интерполяция по формуле (4) допустима, если

$$\frac{\max |f^{(3)}(\xi)| h^3}{72\sqrt{3}} + \frac{3 \max |f^{(4)}(\xi)|}{128} h^4 \leq \varepsilon. \quad (5)$$

При малых  $h$  главной частью является первое слагаемое; оно меньше, чем левая часть (3), в  $9\sqrt{3}/2 = 7,794\dots$  раз. Следовательно, при малых  $h$  для выполнимости (5) можно взять шаг  $h \approx \sqrt[3]{9\sqrt{3}/2} \approx 1,98$  раз больше, чем для выполнимости (3). В рассматриваемом примере условие (5) имеет вид

$$h^3/72\sqrt{3} + 3h^4/128 \leq 0,5 \cdot 10^{-6}.$$

Решая это неравенство, получим  $h \leq h_0 = 0,038 \dots$ . Заметим, что при ручном счете шаг  $h = 0,038$  неудобен вследствие «некруглости» этого числа. Поэтому при составлении таблиц его заменили бы заведомо на меньшее, но «более круглое» число 0,03.

Если и такой шаг таблицы слишком велик, например таблица не помещается в оперативной памяти ЭВМ, целесообразно увеличить порядок используемого интерполяционного многочлена.

### § 13. О погрешности округления при интерполяции

Предположим, что выбран некоторый способ интерполяции. Выше мы получили представление о погрешности, являющейся следствием замены функции многочленом. Однако существует еще одна причина погрешности, в частности вследствие округления этих значений. Пусть требуется вычислить значение  $f(x)$  по формуле

$$f(x) \approx \sum_{j=1}^n f_j P_j(x),$$

являющейся общим видом рассматриваемых нами интерполяционных формул. Поскольку реально заданы не  $f_j$ , а приближенные значения  $f_j^* = f_j + \eta_j$ , то в результате будет получено значение

$$f^* = \sum_{j=1}^n f_j P_j(x) + \sum_{j=1}^n \eta_j P_j(x).$$

Если известны границы изменения значений  $\eta_j$ , то можно оценить верхнюю грань погрешности

$$\varepsilon = \sum_{j=1}^n \eta_j P_j(x).$$

Например, при условии  $|\eta_j| \leq \eta$  имеем оценку

$$|\varepsilon| \leq \eta \Lambda(x), \quad \Lambda(x) = \sum_{j=1}^n |P_j(x)|.$$

Величина  $\Lambda$  может оказаться очень большой.

**Задача 1.** Пусть  $f(x)$  интерполируется по узлам  $x_j = (j-1)h$ ,  $h = 1/(n-1)$ ,  $j = 1, \dots, n-1$ . Показать, что  $\max_{0 \leq x \leq 1} \Lambda(x) \geq \text{const} \cdot \frac{2^n}{\sqrt{n}}$ .



В случае, когда в качестве узлов интерполяции берутся нули многочленов Чебышева, величина  $\Lambda(x)$  оказывается весьма небольшой.

**Задача 2.** Доказать, что если узлы интерполяции совпадают с нулями многочленов Чебышева, то

$$\max_{[-1, 1]} |\Lambda(x)| \leq \text{const} \cdot \ln n.$$

Если мы вычисляем значение  $f(x)$  интерполяцией по узлам  $x_0, x_1$  при  $x_0 < x < x_1$ , то

$$P_0(x) = \frac{x_1 - x}{x_1 - x_0}, \quad P_1(x) = \frac{x - x_0}{x_1 - x_0}$$

и

$$\eta(|P_0(x)| + |P_1(x)|) = \eta.$$

Таким образом, при линейной интерполяции погрешность, являющаяся следствием округления значений функции, не превосходит погрешности этих значений.

Иногда важно обратить внимание на вычислительную погрешность, вносимую при расчетах по интерполяционным формулам. Многообразие формул интерполирования, применявшихся во времена ручного счета, отчасти объясняется именно поисками алгоритмов, порождающих минимальную вычислительную погрешность.

## § 14. Применения аппарата интерполирования. Обратная интерполяция

Употребление интерполяционных многочленов оказывается полезным при решении такой задачи.

Пусть требуется найти экстремум функции и точку экстремума. Составим таблицу значений функции с крупным шагом. Из рассмотрения этой таблицы можно увидеть место расположения экстремума. В предполагаемой области расположения экстремума приблизим функцию интерполяционным многочленом и найдем его точку экстремума  $P_1$ . В окрестности точки  $P_1$  составим таблицу значений функции с более мелким шагом. Из рассмотрения этой таблицы можно уточнить расположение экстремума и т. д. Степень интерполяционного многочлена берется такой, чтобы точка экстремума определялась в явном виде. В одномерном случае берется интерполяционный многочлен Лагранжа или Бесселя второй степени или интерполяционный многочлен третьей степени. В многомерном случае, как правило, функция приближается многочленом второй степени.

На практике, начиная с окрестности приближения  $P_1$  или следующего приближения  $P_2$ , уже не строят подробной таблицы значений функции, а ограничиваются минимальным числом точек в окрестности имеющегося приближения  $P_n$ , достаточным для построения интерполяционного многочлена,

В одномерном случае иногда после вычисления значения  $f(P_n)$  не вычисляют дополнительно никаких новых значений функции, а проводят интерполяцию, используя это значение и ранее вычисленные значения.

Другой типичной задачей, где может быть применен аппарат интерполирования, является нахождение корня  $X$  уравнения  $f(x) = d$ .

Путь решения этой задачи тот же самый. Составляем таблицу значений функции; определяем по ней грубо, где находится корень уравнения, затем составляем таблицу с более мелким шагом и т. д.

Если вычисление функции относительно нетрудоемко, неразумно применять в процессе вычислений интерполяцию степени выше второй; в противном случае возникает задача нахождения корней многочленов, сама требующая достаточно большого числа арифметических операций. Если вычисление функции трудоемко, может оказаться более выгодным пойти по пути увеличения степени интерполяционного многочлена.

В случае, когда в окрестности  $y = d$  функция  $g(y)$ , обратная к  $f(x)$ , является достаточно гладкой, более эффективным может оказаться применение обратной интерполяции. *Обратной интерполяцией* называется следующий алгоритм. Пусть известны значения функции  $y_i = f(x_i)$  при  $i = 1, \dots, n$ . Эта информация эквивалентна тому, что известны значения  $x_i = g(y_i)$  обратной функции. При условии допустимости интерполяции по переменной  $y$  можно заменить обратную функцию  $g(y)$  интерполяционным многочленом  $L_n(y)$ , удовлетворяющим условиям

$$L_n(y_i) = x_i, \quad i = 1, \dots, n,$$

и положить  $X = g(d) \approx L_n(d)$ . Такой способ особенно удобен, если нас интересуют значения решения уравнений при достаточно большом числе значений  $d$  или желательно получение явного выражения корня уравнения  $f(x) = d$  в зависимости от параметра  $d$ . Если интерполяция по узлам  $y_1, \dots, y_n$  не обеспечивает нужной точности, полагаем  $x_{n+1} = L_n(d)$ ,  $y_{n+1} = f(x_{n+1})$ . Далее, в зависимости от обстановки, целесообразно заменить  $g(d)$  значением интерполяционного многочлена по всем узлам  $y_1, \dots, y_{n+1}$  или по некоторым узлам, ближайшим к  $d$ .

## § 15. Численное дифференцирование

Простейшие формулы *численного дифференцирования* получаются в результате дифференцирования интерполяционных формул.

Пусть известны значения функции в точках  $x_1, \dots, x_n$  и требуется вычислить производную  $f^{(k)}(x_0)$ . Построим интерполяционный многочлен  $L_n(x)$  и положим  $f^{(k)}(x_0) \approx L_n^{(k)}(x_0)$ . Точно так же мы можем заменять значения производных функций зна-

чениями производных других многочленов интерполяционного типа, например Бесселя.

Другой способ построения формул численного дифференцирования, приводящий к тем же формулам, — это метод неопределенных коэффициентов. Наиболее употребителен он в многомерном случае, когда не всегда просто выписывается интерполяционный многочлен. Коэффициенты  $c_i$  формулы численного дифференцирования

$$f^{(k)}(x) \approx \sum_{i=1}^n c_i f(x_i) \quad (1)$$

выбираются из условия, чтобы формула была точна для многочленов максимально высокой степени. Возьмем  $f(x) = \sum_{j=0}^m a_j x^j$  и потребуем, чтобы для такого многочлена соотношение (1) обратилось в равенство

$$\sum_{j=0}^m a_j (x^j)^{(k)} \Big|_{x_0} = \sum_{i=1}^n c_i \left( \sum_{j=0}^m a_j x_i^j \right).$$

Чтобы равенство выполнялось для любого многочлена степени  $m$ , необходимо и достаточно, чтобы коэффициенты при  $a_j$  в правой и левой частях были равны. Поскольку

$$(x^j)^{(k)} = j(j-1)\dots(j-k-1)x^{j-k},$$

то получаем линейную систему уравнений

$$j(j-1)\dots(j-k+1)x_0^{j-k} = \sum_{i=1}^n c_i x_i^j, \quad j=0, \dots, m, \quad (2)$$

относительно неизвестных  $c_i$ . Если  $m = n - 1$ , то число уравнений равно числу неизвестных. Определитель системы является определителем Вандермонда, поэтому отличен от нуля. Таким образом, всегда можно построить формулу численного дифференцирования с  $n$  узлами, точную для многочленов степени  $n - 1$ .

При  $m = n - 1$  и определенном расположении узлов иногда оказывается, что равенство (2) выполнено и при  $j = n$ . Как правило, это будет в случае, когда узлы расположены симметрично относительно точки  $x_0$ .

В приведенных ниже задачах для простоты взято  $x_0 = 0$ . Пусть узлы  $x_i$  расположены симметрично относительно точки  $x_0$ , т. е.  $x_1 = -x_n, x_2 = -x_{n-1}$  и т. д. Если  $n$  нечетно,  $n = 2l + 1$ , то тогда  $x_{l+1} = x_0 = 0$ .

**Задача 1.** Пусть  $k$  четно (в частности,  $k$  может быть равно нулю и тогда речь идет об интерполировании). Доказать, что тогда  $c_n = c_1, c_{n-1} = c_2$  и вообще  $c_{n+1-k} = c_k$ .

Доказать, что вследствие такого свойства симметрии формула (1) автоматически является точной для любой нечетной функции.

В частности, при  $n$  нечетном формула (1) будет точна для  $x^n$ , поэтому она точна и для любого многочлена степени  $n$  (поскольку для любого многочлена степени  $n = 1$  она уже оказалась точной по построению).

**Задача 2.** Пусть  $k$  нечетно. Доказать, что тогда  $c_n = -c_1$ ,  $c_{n-1} = -c_2$ , ... и вообще  $c_{n+1-k} = -c_k$ .

Если  $n$  нечетно,  $n = 2l + 1$ , то при  $k = l + 1$  имеем  $c_{l+1} = -c_{l+1}$  и, следовательно,  $c_{l+1} = 0$ .

Доказать, что вследствие такого свойства симметрии формула (1) автоматически является точной для любой четной функции.

В частности, при  $n$  четном она будет точна для  $x^n$  и поэтому будет точна для любого многочлена степени  $n$ .

Таким образом, при симметричном относительно  $x_0$  расположении узлов,  $k$  четном,  $n$  нечетном или  $k$  нечетном,  $n$  четном формула (1) оказывается точной для многочленов на единицу большей степени.

Свойства симметрии формул численного дифференцирования используются для уменьшения числа уравнений, которые нужно решить при построении формулы.

Пусть требуется построить формулу численного дифференцирования

$$f'(0) \approx c_1 f(-h) + c_2 f(0) + c_3 f(h),$$

точную для многочленов второй степени. Система уравнений (2) в данном случае имеет вид

$$0 = c_1 + c_2 + c_3,$$

$$1 = c_1(-h) + c_3(h),$$

$$0 = c_1(-h)^2 + c_3(h)^2,$$

и, решая ее, получаем  $c_1 = -1/(2h)$ ,  $c_2 = 0$ ,  $c_3 = 1/(2h)$ . Воспользуемся свойством симметрии и сразу возьмем формулу, для которой  $c_3 = -c_1$ ,  $c_2 = 0$ . Тогда первое и третье уравнения выполнены автоматически, а второе приобретает вид  $1 = 2c_3 h$ , т. е.  $c_3 = 1/(2h)$ . Таким образом,  $f'(0) \approx (f(h) - f(-h))/(2h)$ .

**Задача 3.** Если все точки  $x_j$  удалены от точки  $x_0$  на расстояние  $O(h)$ ,  $h$  — малая величина, то при гладкой  $f(x)$  приближенная формула численного дифференцирования (1) имеет порядок погрешности  $O(h^m)$ , где  $m = l + 1 - k$ ,  $l$  — максимальная степень многочленов, для которых точна эта формула.

Построим приближенную формулу вычисления второй производной, использующую те же узлы:

$$f''(0) \approx \frac{c_1 f(-h) + c_2 f(0) + c_3 f(h)}{h^2}.$$

Из условий точности формулы для 1,  $x$ ,  $x^2$  получаем систему уравнений

$$0 = c_1 + c_2 + c_3,$$

$$0 = \frac{c_1(-h) + c_3 h}{h^2},$$

$$1 = \frac{c_1 \frac{h^2}{2} + c_3 \frac{h^2}{2}}{h^2}.$$

Решая эту систему, получим  $c_1 = c_3 = 1$ ,  $c_2 = -2$  и соответствующую приближенную формулу

$$f''(0) \approx \frac{f(h) - 2f(0) + f(-h)}{h^2}. \quad (3)$$

Мы можем не сомневаться в том, что мы получили правильную формулу. Выражение в правой части есть  $f_0^2/h^2 = \Delta^2 f_{-1}/h^2$ , и, согласно (10.4), оно равно значению  $f''(\xi)$ . У многочлена второй степени вторая производная постоянна, поэтому  $f_0^2/h^2 = f''(\xi) \equiv f''(x)$  при любом  $x$ , в частности  $f''(0) = f_0^2/h^2$ .

Построенная формула оказывается точной для любого многочлена третьей степени. Если подставим в левую и правую части (3) функцию  $f(x) = x^3$ , то в обеих частях получим нуль.

Оценим погрешность построенной выше приближенной формулы  $f'(0) \approx (f(h) - f(-h))/(2h)$ . В формуле Тейлора возьмем три члена разложения и остаточный член

$$\begin{aligned} f(h) &= f(0) + hf'(0) + \frac{h^2}{2} f''(0) + \frac{h^3}{6} f'''(\xi_+), \\ f(-h) &= f(0) - hf'(0) + \frac{h^2}{2} f''(0) - \frac{h^3}{6} f'''(\xi_-). \end{aligned}$$

Введем обозначение  $R_k(f) = f^{(k)}(x_0) - \sum_{i=1}^n c_i f(x_i)$ . Имеем

$$\begin{aligned} R_1(f) &= f'(0) - \frac{f(h) - f(-h)}{2h} = \\ &= f'(0) - \left( f(0) + hf'(0) + \frac{h^2}{2} f''(0) + \frac{h^3}{6} f'''(\xi_+) - \right. \\ &\quad \left. - \left( f(0) - hf'(0) + \frac{h^2}{2} f''(0) - \frac{h^3}{6} f'''(\xi_-) \right) \right) / (2h) = \\ &= f'(0) - \left( f'(0) + \frac{h^2}{6} \left( \frac{f'''(\xi_+) + f'''(\xi_-)}{2} \right) \right) = \\ &= -\frac{h^2}{6} \alpha, \quad \alpha = \frac{f'''(\xi_+) + f'''(\xi_-)}{2}. \end{aligned}$$

Значение  $\alpha$  лежит между  $f'''(\xi_+)$  и  $f'''(\xi_-)$ . Поэтому по теореме Ролля найдется  $\xi$  в пределах  $[\xi_-, \xi_+]$  такое, что  $\alpha = f'''(\xi)$ . Таким образом, в итоге имеем

$$R_1(f) = -\frac{h^2}{6} f'''(\xi).$$

Рассмотрим приближенную формулу (3). Предположим сначала, что нам неизвестно, точна ли она для любого многочлена третьей степени. Беря в разложении Тейлора три члена

$$f(\pm h) = f(0) \pm hf'(0) + \frac{h^2}{2} f''(0) \pm \frac{h^3}{6} f'''(\xi_{\pm}),$$

получим

$$R_2(f) = f''(0) - \frac{f(h) - 2f(0) + f(-h)}{h^2} = f''(0) - \\ - \left[ f(0) + hf'(0) + \frac{h^2}{2} f''(0) + \frac{h^3}{6} f'''(\xi_+) - 2f(0) + f(0) - hf'(0) + \right. \\ \left. + \frac{h^2}{2} f''(0) - \frac{h^3}{6} f'''(\xi_-) \right] h^{-2} = -\frac{h}{6} (f'''(\xi_+) - f'''(\xi_-)).$$

Если  $f(x)$  — многочлен третьей степени, то  $f'''(x) = \text{const}$ , поэтому  $R_2(f) \equiv 0$ . Таким образом, из выражения для погрешности мы увидели, что формула (3) точна для всех многочленов третьей степени. По теореме Лагранжа

$$f'''(\xi_+) - f'''(\xi_-) = (\xi_+ - \xi_-) f^{(4)}(\bar{\xi});$$

$\bar{\xi} \in [\xi_-, \xi_+]$ . В то же время  $\xi_+ \in [0, h]$ ,  $\xi_- \in [-h, 0]$ , поэтому  $0 \leq \xi_+ - \xi_- \leq 2h$ . Таким образом,  $\xi_+ - \xi_- = \theta \cdot 2h$ , где  $0 \leq \theta \leq 1$ , и

$$R_2(f) = -\theta \frac{h^2}{6} f^{(4)}(\bar{\xi}), \quad 0 \leq \theta \leq 1.$$

Если в разложении Тейлора взять четыре слагаемых

$$f(\pm h) = f(0) \pm hf'(0) + \frac{h^2}{2} f''(0) \pm \frac{h^3}{6} f^{(3)}(0) + \frac{h^4}{24} f^{(4)}(\xi_{\pm}),$$

то получим выражение для погрешности

$$R_2(f) = -\frac{h^2}{12} \left( \frac{f^{(4)}(\xi_+) + f^{(4)}(\xi_-)}{2} \right).$$

Рассуждая, как и при выводе оценки погрешности для первой производной, имеем

$$R_2(f) = -\frac{h^2}{12} f^{(4)}(\xi).$$

Приведем ряд формул численного дифференцирования функций, заданных на сетке с постоянным шагом  $x_n = x_0 + nh$ :

$$\begin{aligned} f'(x_0) &\approx h^{-1} \sum_{j=1}^n \frac{(-1)^{j-1}}{j} f_{j/2}^j, \\ R_1(f) &= \frac{(-1)^n}{n+1} f^{(n+1)}(\xi) h^n; \\ f'(x_0) &\approx h^{-1} \sum_{j=1}^n \frac{1}{j} f_{-j/2}^j, \\ R_1(f) &= \frac{1}{n+1} f^{(n+1)}(\xi) h^n. \end{aligned} \tag{4}$$

Это так называемые односторонние формулы численного дифференцирования. В первой формуле (4) все узлы удовлетворяют условию  $x_k \geq x_0$ , во второй  $x_k \leq x_0$ . Среди таких формул наиболее употребительны:

$$n = 1, \quad f'(x_0) \approx \frac{f_{1/2}^1}{h} = \frac{f(x_1) - f(x_0)}{h};$$

$$n = 2, \quad f'(x_0) \approx \frac{1}{h} \left( f_{1/2}^1 - \frac{1}{2} f_1^2 \right) = \frac{-f(x_2) + 4f(x_1) - 3f(x_0)}{2h}$$

и соответственно

$$n = 1, \quad f'(x_0) \approx \frac{f_{-1/2}^1}{h} = \frac{f(x_0) - f(x_{-1})}{h};$$

$$n = 2, \quad f'(x_0) \approx \frac{1}{h} \left( f_{-1/2}^1 + \frac{1}{2} f_{-1}^2 \right) = \frac{3f(x_0) - 4f(x_{-1}) + f(x_{-2})}{2h}.$$

Такие приближения производных часто используются при решении дифференциальных уравнений для аппроксимации граничных условий. Приведем примеры симметричных формул:

$$f' \left( x_0 + \frac{h}{2} \right) \approx h^{-1} \sum_{j=0}^{l-1} \frac{(-1)^j \left( \left( \frac{1}{2} \right) \cdot \left( \frac{3}{2} \right) \cdots \left( j - \frac{1}{2} \right) \right)^2}{(2j+1)!} f_{1/2}^{2j+1};$$

$$R_1(f) = (-1)^l \frac{\left( \left( \frac{1}{2} \right) \left( \frac{3}{2} \right) \cdots \left( l - \frac{1}{2} \right) \right)^2}{(2l+1)!} f^{(2l+1)}(\xi) h^{2l}.$$

Наиболее употребительны следующие частные случаи:

$$l = 1, \quad f' \left( x_0 + \frac{h}{2} \right) \approx \frac{f_{1/2}^1}{h} = \frac{f(x_1) - f(x_0)}{h}$$

(уже рассмотренный нами выше в других обозначениях);

$$l = 2, \quad f' \left( x_0 + \frac{h}{2} \right) \approx \frac{1}{h} \left( f_{1/2}^1 - \frac{1}{24} f_{1/2}^3 \right) =$$

$$= \frac{-f(2h) + 27f(h) - 27f(0) + f(-h)}{24h}.$$

Формулы для второй производной записываются в виде

$$f''(x_0) \approx h^{-2} \sum_{j=1}^l \frac{2(-1)^{j-1} ((j-1)!)^2}{(2j)!} f_0^{2j},$$

остаточный член

$$R_2(f) = \frac{2(-1)^l (l!)^2}{(2l+2)!} f^{(2l+2)}(\xi) h^{2l}.$$

Наиболее употребительны частные случаи:

$$l=1, \quad f''(x_0) \approx \frac{f_0''}{h^2} = \frac{f(h) - 2f(0) + f(-h)}{h^2};$$

$$l=2, \quad f''(x_0) \approx \frac{1}{h^2} \left( f_0'' - \frac{1}{12} f_0'''' \right) = \\ = \frac{-f(2h) + 16f(h) - 30f(0) + 16f(-h) - f(-2h)}{12h^2}.$$

Для высших производных простейшую грубую аппроксимацию можно получить, воспользовавшись (1).

При  $m$  фиксированном, не зависящем от  $h$ , имеем

$$\frac{\Delta^k f_m}{h^k} = f^{(k)}(\xi_m), \quad x_0 + mh \leq \xi_m \leq x_0 + (m+k)h,$$

поэтому

$$R_k = f^{(k)}(x_0) - \frac{\Delta^k f_m}{h^k} = f^{(k)}(x_0) - f^{(k)}(\xi_m) = O(h).$$

Наиболее употребительные частные случаи: односторонние формулы численного дифференцирования

$$f^{(k)}(0) \approx \frac{\Delta^k f_0}{h^k} = \frac{f_{k/2}^{(k)}}{h^k},$$

$$f^{(k)}(0) \approx \frac{\nabla^k f_0}{h^k} = \frac{f_{-k/2}^{(k)}}{h^k},$$

имеющие погрешность порядка  $O(h)$ , и симметричные формулы численного дифференцирования. При  $k$  четном

$$f^{(k)}(0) \approx f_0^{(k)}/h^k,$$

при  $k$  нечетном

$$f^{(k)}(0) \approx \frac{f_{1/2}^{(k)} + f_{-1/2}^{(k)}}{2h^k}.$$

Эти формулы имеют погрешность  $O(h^2)$ . При  $k=1, 2$  такими формулами как раз являются формулы, приведенные выше.

При выводе формул численного дифференцирования из приближенного равенства

$$f^{(k)}(x_0) \approx L_n^{(k)}(x_0)$$

оценку погрешности также можно получить, дифференцируя остаточный член в (5.1):

$$f^{(k)}(x_0) = L_n^{(k)}(x_0) + (f(x; x_1; \dots; x_n) \omega_n(x))^{(k)} \Big|_{x_0}.$$

Для получения конкретной оценки надо воспользоваться правилом Лейбница и доказать равенство

$$(f(x; x_1; \dots; x_n))^{(q)} = q! \underbrace{f(x; \dots; x; x_1; \dots; x_n)}_{q+1 \text{ раз}}.$$



## § 16. О вычислительной погрешности формул численного дифференцирования

При решении одной задачи управления имела место следующая ситуация. Управление объектом выбиралось в зависимости от его скорости движения в данный момент; скорость вычислялась по простейшей формуле численного дифференцирования как отношение приращения координат к промежутку времени  $\Delta t$  между двумя последовательными моментами измерения положения объекта.

Перед непосредственным конструированием системы было произведено подробное моделирование ее работы с помощью ЭВМ: координаты объекта брались со случайными погрешностями, соответствующими реальным погрешностям измерения, и т. д.

Оказалось, что объект все время резко меняет направление движения и требуемое управление движением нереализуемо. Однако уменьшение промежутка  $\Delta t$  не приводило к улучшению дела. В данном конкретном примере проблема была решена путем увеличения промежутка  $\Delta t$  в 100 раз по сравнению с предполагавшимся заранее. Попутно это привело к снижению стоимости управляющей системы.

Дело заключается в том, что часто уменьшение погрешности метода, в данном случае формулы численного дифференцирования, сопровождается ростом влияния погрешности исходных данных и вычислительной погрешности. Численное дифференцирование относится к таким задачам, где влияние этих погрешностей сказывается уже при умеренных значениях погрешности метода решения задачи.

Пусть для определенности значение  $f'(x_0)$  определяется из соотношения

$$f'(x_0) \approx (f(x_1) - f(x_0))/h. \quad (1)$$

Согласно (15.4) остаточный член этой формулы имеет вид

$$r_1 = -f''(\xi)h/2.$$

Пусть  $|f''(\xi)| \leq M_2$ , тогда  $|r_1| \leq M_2h/2$ .

Если значения функции  $f(x_i)$  известны с некоторыми погрешностями  $\varepsilon_i$ ,  $|\varepsilon_i| \leq E$ , то погрешность  $f'(x_0)$  будет содержать дополнительное слагаемое

$$r_2 = -(\varepsilon_1 - \varepsilon_0)/h, \quad |r_2| \leq 2E/h.$$

Для простоты пренебрежем округлениями при реальном вычислении правой части (1). Тогда имеем оценку погрешности

$$|r| \leq |r_1| + |r_2| \leq g(h) = M_2h/2 + 2E/h. \quad (2)$$

Для малости погрешности необходима малость  $h$ , но при умень-

шении  $h$  растет второе слагаемое (рис. 2.16.1). Из уравнения  $g'(h) = 0$  получаем точку экстремума  $h_0$  для  $g(h)$ :

$$h_0 = 2 \sqrt{E/M_2},$$

затем значение

$$g(h_0) = 2 \sqrt{M_2 E}.$$

Таким образом, ни при каком  $h$  нельзя гарантировать, что погрешность результата будет величиной порядка  $o(\sqrt{E})$ .

Погрешности  $\epsilon_i$  возникают вследствие погрешностей в задаваемых значениях функций, например, если функция определяется из измерений или вычисляется по некоторой приближенной формуле. Поскольку эти значения округляются дополнительно при вводе в машину, то следует считать, что  $E \geq \text{const} \cdot 2^{-t}$ , где  $t$  — число разрядов. Таким образом, мы можем получить  $f'(x_0)$  в лучшем случае с половиной верных разрядов.

В случае применения формул более высокого порядка точности положение несколько улучшается. Пусть производная  $y^{(k)}(x)$  вычисляется по формуле

$$y^{(k)}(x) \approx \left( \sum_{j=1}^n c_j y_j \right) / h^k, \quad c_j = O(1), \quad (3)$$

с остаточным членом  $O(h^l)$ . Все рассмотренные выше формулы численного дифференцирования могут быть записаны в таком виде: в знаменателе —  $h^k$ , а в числителе — коэффициенты порядка  $O(1)$ . Погрешность, являющаяся следствием погрешностей в правой части (3), оценивается величиной  $\text{const} \cdot E/h^k$ . Таким образом, вместо (2) мы имеем

$$|r| \leq g_1(h) = A_1 h^l + A_2 E/h^k.$$

Минимум правой части достигается при  $h$  порядка  $E^{1/(l+k)}$ , при этом сама правая часть имеет порядок  $E^{l/(l+k)}$ . Таким образом, с ростом  $l$  порядок погрешности по отношению к  $E$  повышается; при этом значение шага, соответствующее минимуму оценки погрешности, становится все больше. Конечно, следует иметь в виду, что величины  $A_1$  и  $A_2$  могут расти с ростом  $l$ , поэтому увеличение  $l$  разумно лишь в определенных пределах.

Иногда складывается обстановка, когда повышение точности формул численного дифференцирования не приводит к требуемому результату. Тогда применяются методы предварительного сглаживания исследуемой функции. Одна группа методов базируется на идеях математической статистики. За счет обработки большого числа наблюдаемых значений функции уменьшается случайная погрешность в ее значениях. Другая группа методов, получающая распространение в последнее время, использует идеи регуляризации. О методах этой группы подробнее будет сказано в гл. V.

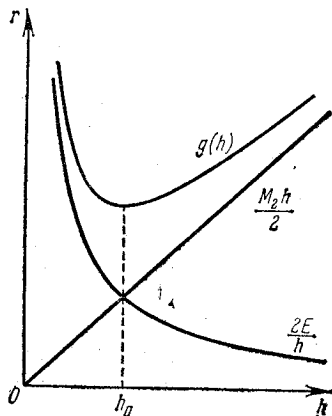


Рис. 2.16.1

## § 17. Рациональная интерполяция

В ряде случаев большую точность приближения можно достигнуть, используя рациональную интерполяцию. При заданных  $f(x_1), \dots, f(x_n)$  приближение к  $f(x)$  ищется в виде

$$R(x) = \frac{a_0 + a_1x + \dots + a_px^p}{b_0 + b_1x + \dots + b_qx^q}, \quad p + q + 1 = n.$$

Коэффициенты  $a_i, b_i$  находятся из совокупности соотношений  $R(x_j) = f(x_j)$  ( $j = 1, \dots, n$ ), которые можно записать в виде

$$\sum_{i=0}^p a_i x_j^i - f(x_j) \sum_{i=0}^q b_i x_j^i = 0, \quad j = 1, \dots, n. \quad (1)$$

Уравнения (1) образуют систему  $n$  линейных уравнений относительно  $n + 1$  неизвестных. Такая система всегда имеет нетривиальное решение.

Функция  $R(x)$  может быть записана в явном виде в случаях  $n$  нечетное,  $p = q$ , и  $n$  четное,  $p - q = 1$ .

Для этого следует вычислить так называемые обратные разделенные разности, определяемые условиями

$$f^-(x_l, x_k) = \frac{x_l - x_k}{f(x_l) - f(x_k)}$$

и рекуррентным соотношением

$$f^-(x_k, \dots, x_l) = \frac{x_l - x_k}{f^-(x_{k+1}, \dots, x_l) - f^-(x_k, \dots, x_{l-1})}.$$

Интерполирующая рациональная функция записывается в виде цепной дроби

$$f(x) = f(x_1) + \frac{x - x_1}{f^-(x_1, x_2) + \frac{x - x_2}{f^-(x_1, x_2, x_3) + \dots + \frac{x - x_{n-1}}{f^-(x_1, \dots, x_n)}}}.$$

Использование рациональной интерполяции по подходящим образом выбранным узлам часто целесообразнее интерполяции многочленами в случае функций с нерегулярным характером поведения (резкое изменение или особенности производных в отдельных точках).

Литература к гл. II: [35, 37].

## Глава III

### ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

Эта глава посвящена методам приближенного вычисления одномерных интегралов. Сначала строятся простейшие формулы для приближенного вычисления интегралов по отрезку. Такие формулы называют *квадратурными*. В многомерном случае (когда размерность интеграла больше единицы) формулы для приближенного вычисления интеграла называют *кубатурными*.

Изучается вопрос о повышении точности вычисления интегралов за счет повышения порядка точности квадратур (т. е. повышения степени полиномов, для которых квадратуры точны), за счет разбиения отрезка интегрирования на части, за счет сведения интегралов от функций с «особенностями» к интегралам от более гладких функций.

На примере численного интегрирования иллюстрируются требования, предъявляемые к стандартным программам и алгоритмам, которые кладутся в их основу. Даются описания ряда стандартных программ численного интегрирования.

#### § 1. Простейшие квадратурные формулы.

##### Метод неопределенных коэффициентов

Простейшие квадратурные формулы можно получить из наглядных соображений. Пусть вычисляется интеграл

$$I = \int_a^b f(x) dx. \quad (1)$$

Если  $f(x) \approx \text{const}$  на рассматриваемом отрезке, то можно положить  $I \approx (b - a)f(\xi)$ ,  $\xi$  — произвольная точка отрезка  $[a, b]$ . Естественно взять в качестве  $\xi$  среднюю точку отрезка; тогда получим *формулу прямоугольников*

$$I \approx (b - a)f\left(\frac{b + a}{2}\right).$$

Предположим, что функция  $f(x)$  на  $[a, b]$  близка к линейной; тогда естественно заменить интеграл площадью трапеции с высотой  $(b - a)$  и основаниями  $f(a)$  и  $f(b)$  (рис. 3.1.1).

Получим формулу трапеций

$$I \approx (b - a) \frac{f(a) + f(b)}{2}.$$

Если функция  $f(x)$  близка к линейной, то из наглядных соображений видно, что формула прямоугольников также должна давать неплохой результат: дело в том, что  $(b - a) f\left(\frac{b+a}{2}\right)$  есть площадь любой трапеции с высотой  $b - a$  и средней линией  $f\left(\frac{b+a}{2}\right)$ ; в частности, она равна площади трапеции, у которой одна из сторон лежит на касательной к графику функции в точке  $\left(\frac{b+a}{2}, f\left(\frac{b+a}{2}\right)\right)$  (рис. 3.1.2).

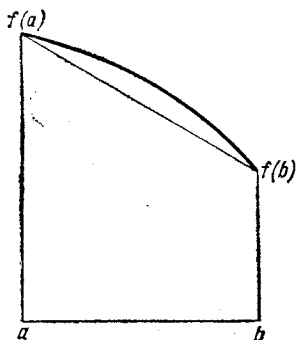


Рис. 3.1.1

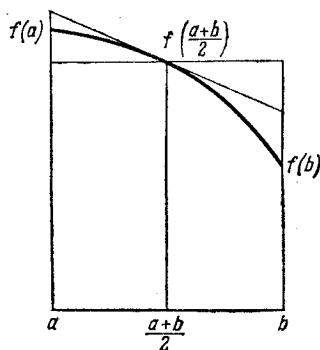


Рис. 3.1.2

Более сложные квадратурные формулы, так же как и формулы численного дифференцирования, строятся методом неопределенных коэффициентов или при помощи аппарата интерполирования.

Рассмотрим простейший пример построения квадратуры методом неопределенных коэффициентов. Строится квадратурная формула

$$\int_{-1}^1 f(x) dx \approx S(f) = c_1 f(-1) + c_2 f(0) + c_3 f(1),$$

точная для многочленов наиболее высокой степени. Погрешность квадратуры

$$R(f) = \int_{-1}^1 f(x) dx - S(f)$$

является линейным функционалом, и при  $f = \sum_{j=0}^m a_j x^j$  имеем

$$R(f) = \sum_{j=0}^m a_j R(x^j).$$

Таким образом, нужно добиться выполнения равенств

$$R(1) = 0, \dots, R(x^l) = 0$$

при возможно большем значении  $l$ . Получаем уравнения

$$R(1) = 2 - (c_1 + c_2 + c_3) = 0,$$

$$R(x) = 0 - (-c_1 + c_3) = 0,$$

$$R(x^2) = \frac{2}{3} - (c_1 + c_3) = 0,$$

$$R(x^3) = 0 - (-c_1 + c_3) = 0,$$

.....

Поскольку нужно определить 3 свободных параметра, то, вообще говоря, можно решить лишь первые 3 уравнения, из которых получаем

$$c_1 = c_3 = 1/3, \quad c_2 = 4/3.$$

В данном конкретном случае четвертое уравнение выполнено автоматически и мы получили квадратуру, точную для многочленов третьей степени, *называемую формулой Симпсона*.

Вообще говоря, требуется вычислять интегралы не по отрезку  $[-1, 1]$ , а по произвольным отрезкам  $[a, b]$ . Переход к отрезку  $[-1, 1]$  удобен тем, что для него арифметические выкладки, выполняемые при построении квадратуры, оказываются короче.

Иногда оказывается, что подынтегральная функция хорошо приближается не многочленами, а так называемыми обобщенными многочленами, т. е. линейными комбинациями вида  $\sum_{j=0}^m b_j \varphi_j(x)$ , где  $\varphi_j(x)$  — какие-то конкретные функции. Тогда методом неопределенных коэффициентов строится квадратура, точная для функций такого вида.

Наиболее часто такие квадратуры используются в случае, когда  $f(x)$  хорошо приближается функциями, представленными в виде произведения некоторой фиксированной функции  $p(x)$  на многочлен, т. е. функциями вида

$$\sum_{j=0}^m a_j p(x) x^j.$$

В этом случае функцию  $p(x)$  называют *весом* или *весовой функцией*, полагают  $F(x) = f(x)/p(x)$  и исходный интеграл записывают в виде

$$\int_a^b F(x) p(x) dx. \tag{2}$$

## Задача построения квадратуры

$$\int_a^b f(x) dx \approx \sum_{j=1}^n c_j f(x_j),$$

точной для всех функций вида  $p(x)P_m(x)$ , где  $P_m(x)$  — многочлен степени  $m$ , заменяется задачей построения квадратуры

$$\int_a^b F(x) p(x) dx \approx \sum_{j=1}^n C_j F(x_j),$$

точной для всех многочленов степени  $m$ . В случае, когда все  $p(x_j)$  отличны от нуля и бесконечности, эти задачи эквивалентны. В дальнейшем подынтегральную функцию в таких интегралах будем обозначать как  $f(x)p(x)$ .

Перейдем к оценке погрешности квадратурных формул.

### § 2. Оценки погрешности квадратуры

Пусть вычисляется интеграл (1.2). Если квадратура точна для многочленов  $P_m(x)$  степени  $m$ , то

$$R(P_m) = I(P_m) - S(P_m) = 0,$$

поэтому

$$R(f) = R(f - P_m) + R(P_m) = R(f - P_m)$$

при любом многочлене  $P_m(x)$  степени  $m$ . Оценивая в  $R(g)$  каждое слагаемое, получим оценку

$$|R(g)| \leq \int_a^b |g(x)| |p(x)| dx + \sum_{j=1}^n |C_j| |g(x_j)| \leq V \sup_{|a, b|} |g(x)|,$$

где

$$V = \int_a^b |p(x)| dx + \sum_{j=1}^n |C_j|.$$

Поэтому

$$|R(f)| = |R(f - P_m)| \leq V \|f - P_m\|_C$$

при любом  $P_m$  — многочлене степени  $m$ ; здесь

$$\|f - P_m\|_C = \sup_{|a, b|} |f(x) - P_m(x)|.$$

Взяв в правой части нижнюю грань по всем многочленам степени  $m$ , получим оценку

$$|R(f)| \leq V E_m(f), \quad (1)$$

где

$$E_m(f) = \inf_{P_m} \|f - P_m\|_C.$$

Построенные выше простейшие квадратурные формулы и ряд более сложных квадратур удовлетворяют условию  $C_j \geq 0$ , если  $p(x) \geq 0$ ; в рассмотренных нами примерах  $p(x) \equiv 1$ .

Условие, что квадратура точна для многочлена нулевой степени, т. е. для функции  $f \equiv 1$ , имеет вид

$$I(1) = \int_a^b p(x) dx = S(1) = \sum_{j=1}^n C_j.$$

При  $p(x) \geq 0$  и  $C_j \geq 0$  имеем

$$\sum_{j=1}^n |C_j| = \sum_{j=1}^n C_j = \int_a^b p(x) dx, \quad (2)$$

поэтому  $V = 2 \int_a^b p(x) dx$ . Обращаясь к (1), получим оценки

$$|R(f)| \leq 2 \left( \int_a^b p(x) dx \right) E_m(f) \leq 2 \left( \int_a^b p(x) dx \right) \|f - P_m\|_C,$$

где  $P_m$  — любой многочлен степени  $m$ . Если в качестве  $P_m$  взять интерполяционный многочлен по нулям многочленов Чебышева, то на основании (2.9.3) имеем

$$\|f - P_m\|_C \leq \frac{\left(\frac{b-a}{2}\right)^{m+1}}{2^m (m+1)!} \|f^{(m+1)}\|_C.$$

В конкретном случае для веса  $p(x) \equiv 1$  и формул прямоугольников, трапеций, Симпсона, где все  $C_j \geq 0$ , имеем  $V = 2(b-a)$  и

$$|R(f)| \leq \frac{(b-a)^{m+2}}{2^{2m} (m+1)!} \|f^{(m+1)}\|_C.$$

В частных случаях, например для формул прямоугольников и трапеций, где  $m = 1$ , отсюда имеем

$$|R(f)| \leq \frac{(b-a)^3}{8} \|f''\|_C;$$

для формулы Симпсона, где  $m = 3$ , имеем

$$|R(f)| \leq \frac{(b-a)^5}{1536} \|f^{(4)}\|_C.$$

Эти оценки одинаковы для всех квадратур, точных для многочленов какой-то определенной степени, например для формул трапеций и прямоугольников. Можно получить и более точные оценки погрешности этих квадратур.

Опишем универсальный способ получения наиболее точных оценок. В качестве  $P_m(x)$  возьмем сумму первых  $(m+1)$ -го членов разложения функции  $f(x)$  по формуле Тейлора в какой-либо точке  $x_0$  отрезка  $[a, b]$ . Для определенности возьмем  $x_0 = a$  и рассмотрим случай, когда все  $x_j \in [a, b]$ .



Пусть  $P_m(x)$  — такая сумма,  $r_m(x)$  — ее остаточный член:

$$f(x) = P_m(x) + r_m(x).$$

Имеем равенство

$$R(f) = R(r_m(x)) = I(r_m(x)) - \sum_{j=1}^n C_j r_m(x_j).$$

Остаточный член формулы Тейлора возьмем в интегральной форме:

$$r_m(x) = \int_a^x \frac{(x-t)^m}{m!} f^{(m+1)}(t) dt.$$

В двукратном интеграле

$$I(r_m(x)) = \int_a^b p(x) \left( \int_a^x \frac{(x-t)^m}{m!} f^{(m+1)}(t) dt \right) dx$$

сделаем интегрирование по  $t$  внешним, а по  $x$  — внутренним. Получим

$$I(r_m(x)) = \int_a^b K_m(t) f^{(m+1)}(t) dt,$$

где

$$K_m(t) = \int_t^b p(x) \frac{(x-t)^m}{m!} dx.$$

Таким образом, получим

$$R(f) = \int_a^b K_m(t) f^{(m+1)}(t) dt - \sum_{j=1}^n C_j \int_a^{x_j} \frac{(x_j-t)^m}{m!} f^{(m+1)}(t) dt.$$

Положим

$$(x_j - t)_+ = \begin{cases} x_j - t & \text{при } a \leq t \leq x_j, \\ 0 & \text{при остальных } t. \end{cases}$$

Используя это обозначение, представим погрешность  $R(f)$  в виде

$$R(f) = \int_a^b Q(t) f^{(m+1)}(t) dt, \quad Q(t) = K_m(t) - \sum_{j=1}^n C_j \frac{(x_j - t)_+^m}{m!}. \quad (3)$$

Отсюда следует оценка погрешности

$$|R(f)| \leq \left( \int_a^b |Q(t)| dt \right) \|f^{(m+1)}\|_C. \quad (4)$$

Если  $Q(t)$  не меняет знака на отрезке  $[a, b]$ , то по теореме о среднем из (3) получаем

$$R(f) = \left( \int_a^b Q(t) dt \right) f^{(m+1)}(\xi), \quad a \leq \xi \leq b. \quad (5)$$

Далее для простоты положим  $p(x) \equiv 1$ .

**Задача 1.** Предположим, что в качестве  $P_m(x)$  берется сумма  $(m+1)$ -го члена разложения функции  $f(x)$  в ряд Тейлора относительно произвольной точки  $x_0 \neq a$ . Доказать, что представление погрешности (3) при этом не изменится.

**Задача 2.** Пусть точка  $a$  фиксирована,  $f^{(m+1)}(x)$  непрерывна в точке  $a$ . Доказать, что при  $b \rightarrow a$ ,  $C_1 = O(b-a)$

$$\int_a^b Q(t) dt = O((b-a)^{m+2}),$$

$$R(f) = \left( \int_a^b Q(t) dt \right) f^{(m+1)}(a) + o((b-a)^{m+2}),$$

где  $Q(t)$  определено в (3).

Рассмотрим для примера формулу трапеций. Тогда

$$\int_t^b (x-t) dt = \frac{(b-t)^2}{2},$$

$$Q(t) = \frac{(b-t)^2}{2} - \frac{(b-a)}{2} (b-t) = (a-t)(b-t) \leq 0$$

и, подставляя  $Q(t)$  в (5), получим

$$\int_a^b Q(t) dt = -\frac{(b-a)^3}{12},$$

т. е.

$$R(f) = -\frac{(b-a)^3}{12} f''(\xi_1)$$

В случае формулы прямоугольников

$$Q(t) = \frac{(b-t)^2}{2} - (b-a) \left( \frac{a+b}{2} - t \right) = \frac{(t - (a+b)/2)^2}{2} \geq 0$$

и, подставляя  $Q(t)$  в (5), получим

$$R(f) = \frac{(b-a)^3}{24} f''(\xi_2).$$

Часто на практике интересуются не оценкой погрешности (5), которая не поддается улучшению, а ее главным (при  $(b-a) \rightarrow 0$ ) членом

$$f^{(m+1)}(a) \int_a^b Q(t) dt.$$

Если для некоторого  $m$  оказалось, что соответствующий интеграл  $\int_a^b Q(t) dt$  равен нулю, то это значит, что квадратура точна для многочленов степени  $m+1$ . В этом случае надо увеличить  $m$  на 1 и провести аналогичные рассуждения для этого нового значения  $m$ .

Для вычисления главного члена погрешности можно поступить следующим образом. Представим  $f(x)$  в виде суммы первых  $(m+2)$ -х членов разложения Тейлора относительно некоторой точки  $x_0 \in [a, b]$  и остаточного члена; при этом, объединив первые  $m+1$  слагаемое в многочлен степени  $m$ , получим

$$f(x) = T_{x_0}^m(x) + \frac{(x-x_0)^{m+1}}{(m+1)!} f^{(m+1)}(x_0) + r_{m+1}(x),$$

где

$$T_{x_0}^m(x) = \sum_{i=0}^m \frac{(x-x_0)^i}{i!} f^{(i)}(x_0), \quad r_{m+1}(x) = o((x-x_0)^{m+1}).$$

Вследствие линейности функционала погрешности имеем равенство

$$R(f) = R(T_{x_0}^m(x)) + \frac{f^{(m+1)}(x_0)}{(m+1)!} R((x-x_0)^{m+1}) + R(r_{m+1}(x)).$$

Первое слагаемое обращается в нуль, так как квадратура точна для многочленов степени  $m$ . Поскольку

$$|R(r_{m+1}(x))| \leq V \|r_{m+1}\|_C = o((b-a)^{m+2})$$

((при условии, что  $V < \infty$ ), то

$$\frac{f^{(m+1)}(x_0)}{(m+1)!} R((x-x_0)^{m+1})$$

является главным членом погрешности  $R(f)$ . Для простоты выкладок при конкретном вычислении  $R((x-x_0)^{m+1})$  часто удобно произвести замену переменных

$$x = (a+b)/2 + ht, \quad h = (b-a)/2,$$

и рассматривать разложение в ряд Тейлора относительно точки  $t=0$ .

**Задача 3.** Проверить, что

$$\frac{1}{(m+1)!} R((x-x_0)^{m+1}) = \int_a^b Q(t) dt.$$

**Задача 4.** Доказать, что  $R((x-x_0)^{m+1})$  не зависит от выбора  $x_0$ .

В частном случае для формулы трапеций имеем

$$R(x-a)^2 = \frac{(b-a)^3}{3} - \frac{1}{2}(b-a)^3 = -\frac{1}{6}(b-a)^3,$$

поэтому погрешность  $R(f)$  с точностью до членов высшего порядка имеет вид

$$R(f) \sim -\frac{1}{12}(b-a)^3 f^{(2)}(a).$$

### § 3. Квадратурные формулы Ньютона — Котеса

Рассмотренные далее квадратуры относятся к большой группе квадратурных формул, полученных с помощью интегрирования интерполяционного многочлена и объединенных под одним названием — квадратурные формулы Ньютона — Котеса.

Зададимся некоторыми  $d_1, \dots, d_n \in [-1, 1]$  и построим интерполяционный многочлен  $L_n(x)$  степени  $n - 1$ , совпадающий с  $f(x)$  в точках  $x_j = \frac{b+a}{2} + \frac{b-a}{2} d_j$ . Положим

$$\int_a^b f(x) p(x) dx \approx \int_a^b L_n(x) p(x) dx.$$

Имеем

$$R_n(f) = \int_a^b f(x) p(x) dx - \int_a^b L_n(x) p(x) dx = \int_a^b p(x) (f(x) - L_n(x)) dx.$$

Разность  $f(x) - L_n(x)$  оценим, воспользовавшись оценкой погрешности интерполяционного многочлена Лагранжа

$$|f(x) - L_n(x)| \leq (\max_{[a, b]} |f^{(n)}(x)|) \frac{|\omega_n(x)|}{n!},$$

где  $\omega_n(x) = (x - x_1) \dots (x - x_n)$ . Отсюда

$$|R_n(f)| \leq (\max_{[a, b]} |f^{(n)}(x)|) \int_a^b \frac{|\omega_n(x)| |p(x)|}{n!} dx.$$

Произведем в последнем интеграле замену переменных, положив  $x = X(t) = \frac{b+a}{2} + \frac{b-a}{2} t$ . Тогда

$$\frac{1}{n!} \int_a^b |\omega_n(x) p(x)| dx = D(d_1, \dots, d_n) \left(\frac{b-a}{2}\right)^{n+1},$$

где

$$D(d_1, \dots, d_n) = \int_{-1}^1 \frac{|\omega_n^0(t) p^0(t)|}{n!} dt,$$

$$\omega_n^0(t) = (t - d_1) \dots (t - d_n), \quad p^0(t) = p\left(\frac{b+a}{2} + \frac{b-a}{2} t\right).$$

Таким образом, справедлива оценка

$$|R_n(f)| \leq D(d_1, \dots, d_n) (\max_{[a, b]} |f^{(n)}(x)|) \left(\frac{b-a}{2}\right)^{n+1}. \quad (1)$$

Пусть все  $d_j$  различны. Тогда

$$L_n(x) = \sum_{j=1}^n f(x_j) \prod_{i \neq j} \frac{x - x_i}{x_j - x_i}.$$

После замены переменных  $x = X(t)$  получим

$$\int_a^b p(x) L_n(x) dx = \left(\frac{b-a}{2}\right) \sum_{j=1}^n D_j f(x_j), \quad (2)$$

где

$$D_j = \int_{-1}^1 p^0(t) \prod_{i \neq j} \frac{t - d_i}{d_j - d_i} dt. \quad (3)$$

Таким образом, построенная квадратурная формула имеет вид

$$\int_a^b f(x) p(x) dx \approx \frac{b-a}{2} \sum_{j=1}^n D_j f\left(\frac{b+a}{2} + \frac{b-a}{2} d_j\right). \quad (4)$$

Как и при численном дифференцировании, можно обнаружить следующее обстоятельство: если задача имеет определенную симметрию, то метод с симметрией того же типа часто обладает дополнительными преимуществами.

Будем называть функцию *четной* относительно точки  $x_0$ , если  $f(x - x_0) = f(x_0 - x)$ , и *нечетной*, если  $f(x - x_0) = -f(x_0 - x)$ .

Можно показать, что для весовой функции  $p(x)$ , четной относительно середины отрезка  $[a, b]$ , и узлов  $x_j$ , расположенных симметрично относительно середины отрезка, т. е.  $d_j = -d_{n+1-j}$ , коэффициенты квадратуры, соответствующие симметричным узлам, равны между собой:

$$D_j = D_{n+1-j}. \quad (5)$$

(Доказать!)

Такие «симметричные» квадратуры обладают следующим дополнительным свойством, которое, формально говоря, не предусматривалось при их построении. Они точны для любой функции, нечетной относительно середины отрезка  $[a, b]$  (т. е. удовлетворяющей условию  $f\left(x - \frac{a+b}{2}\right) = -f\left(\frac{a+b}{2} - x\right)$ ). В са-

мом деле, для таких функций  $\int_a^b p(x) f(x) dx = 0$  вследствие чет-

ности  $p(x)$ , а  $\sum_{j=1}^n D_j f(x_j) = 0$  вследствие (5); поэтому и  $R_n(f) = 0$ .

В частности, квадратуры будут точны для любого одночлена вида  $\text{const} \cdot \left(x - \frac{b+a}{2}\right)^{2l+1}$ . Свойство симметрии (5) помогает также при непосредственном построении формул методом неопределенных коэффициентов.

Рассмотрим теперь симметричную квадратуру, соответствующую нечетному  $n$ . Она точна для  $f(x) = \text{const} \cdot \left(x - \frac{b+a}{2}\right)^n$  и, согласно построению, точна и для любого многочлена степени  $n-1$ . Следовательно, такая квадратура будет точна и для лю-

бого многочлена степени  $n$ . Таким образом, построенные квадратуры с  $2q - 1$  и  $2q$  узлами, с симметричным расположением узлов оказываются точными для многочленов одинаковой степени  $2q - 1$  (для квадратур Гаусса (см. § 5) эта степень выше).

Чтобы получить уточненную оценку погрешности квадратур с нечетным числом узлов не через  $f^{(n)}(x)$ , а через  $f^{(n+1)}(x)$ , следует заменить подынтегральную функцию интерполяционным многочленом Лагранжа, имеющим точку  $(b + a)/2$  двукратным узлом интерполирования. Ниже для случая  $p(x) \equiv 1$  строится ряд элементарных квадратурных формул и дается оценка их погрешностей; при  $n = 1$  и  $n = 3$  для симметричных формул производится оценка погрешности через  $f^{(n+1)}(x)$ .

1. Формула прямоугольников. Пусть  $n = 1$ ,  $d_1 = 0$ . Тогда

$$D = \int_{-1}^1 |t| dt = 1, \quad D_1 = \int_{-1}^1 1 \cdot dt = 2$$

и имеем квадратурную формулу

$$\int_a^b f(x) dx \approx (b - a) f\left(\frac{a + b}{2}\right) \quad (6)$$

с оценкой остаточного члена

$$R(f) = \max_{[a, b]} |f'(x)| \frac{(b - a)^2}{4}.$$

2. Формула прямоугольников как формула с кратным узлом. Пусть  $n = 2$ ,  $d_1 = d_2 = 0$ . Тогда

$$D = \int_{-1}^1 \frac{t^2}{2} dt = \frac{1}{3},$$

$$L_2(x) = f\left(\frac{a + b}{2}\right) + f'\left(\frac{a + b}{2}\right) \left(x - \frac{a + b}{2}\right),$$

$$\int_a^b L_2(x) dx = (b - a) f\left(\frac{a + b}{2}\right).$$

Таким образом, имеем ту же квадратурную формулу

$$\int_a^b f(x) dx \approx (b - a) f\left(\frac{a + b}{2}\right)$$

с оценкой остаточного члена

$$R(f) = \max_{[a, b]} |f''(x)| \frac{(b - a)^3}{24}.$$

В обоих случаях получилась одна и та же квадратурная формула, но с различной оценкой остаточного члена.

3. Формула трапеций. Пусть  $n = 2$ ,  $d_1 = -1$ ,  $d_2 = 1$ . Тогда

$$D = \int_{-1}^1 \frac{|t^2 - 1|}{2} dt = \frac{2}{3}, \quad D_1 = \int_{-1}^1 \frac{1-t}{2} dt = 1,$$

$$D_2 = \int_{-1}^1 \frac{1+t}{2} dt = 1.$$

Получена формула трапеций

$$\int_a^b f(x) dx \approx \frac{(b-a)}{2} (f(a) + f(b)) \quad (7)$$

с оценкой остаточного члена

$$R(f) = \max_{[a, b]} |f''(x)| \frac{(b-a)^3}{12}.$$

4. Формула Симпсона. Пусть  $n = 4$ ,  $d_1 = -1$ ,  $d_2 = d_4 = 0$ ,  $d_3 = 1$ . Тогда

$$D = \int_{-1}^1 \frac{t^2(1-t^2)}{4!} dt = \frac{1}{90}.$$

Согласно формуле интерполирования с кратными узлами, можем написать, что

$$L_4(x) = L_3(x) + f\left(a; b; \frac{a+b}{2}; \frac{a+b}{2}\right) (x-a)(x-b) \left(x - \frac{a+b}{2}\right),$$

где

$$L_3(x) = f(a) + f(a; b)(x-a) + f\left(a; b; \frac{a+b}{2}\right) (x-a)(x-b).$$

Второе слагаемое в выражении  $L_4(x)$  является функцией, нечетной относительно середины отрезка  $[a, b]$ , поэтому

$$\int_a^b L_4(x) dx = \int_a^b L_3(x) dx.$$

Многочлен  $L_3(x)$  является интерполяционным многочленом второй степени, соответствующим  $d_1 = -1$ ,  $d_2 = 0$ ,  $d_3 = 1$  (рис. 3.3.1). Этим значениям  $d_1$ ,  $d_2$ ,  $d_3$  соответствуют  $D_1 = 1/3$ ,  $D_2 = 4/3$ ,  $D_3 = 1/3$ . В результате получаем квадратурную формулу Симпсона

$$\int_a^b f(x) dx \approx \frac{b-a}{6} \left( f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right) \quad (8)$$

с оценкой остаточного члена

$$R(f) = \max_{[a, b]} |f^{(4)}(x)| \frac{(b-a)^5}{2880}.$$

Основной целью настоящей главы является рассмотрение способов вычисления интегралов от функций, заданных аналитическим выражением, и выработка принципов построения стандартных программ интегрирования таких функций. Естественно, что кроме этих задач в теории квадратурных формул имеются и другие задачи, например связанные с обработкой экспериментального материала.

Для примера обратим внимание на квадратурные формулы Чебышева, широко применявшиеся при подсчете водоизмещения судов. Постановка задачи, приведшей к построению этих квадратур, довольно близка к постановке задачи, возникающей при планировании экспериментов (см. гл. II § 1). Вычисляется

интеграл  $\int_{-1}^1 f(x) dx$ , причем известно, что

функция  $f(x)$  с приемлемой точностью может быть приближена многочленом степени  $q$ . Получение каждого значения  $f(x)$ , например путем измерений, обходится довольно дорого, и получаемые значения содержат довольно большие случайные погрешности. Предположим, что погрешности измерения независимы, имеют одинаковую дисперсию  $d$  и математическое ожидание, равное нулю. Тогда дисперсия приближенного значения  $S_n(f)$ , вычисляемого по квадратурной формуле

$$I(f) \approx S_n(f) = \sum_{j=1}^n c_j f(x_j),$$

равна  $d \sum_{j=1}^n c_j^2$ . Условие  $I(f) = S_n(f)$  при  $f = \text{const}$  имеет вид

$$\sum_{j=1}^n c_j = 2. \quad (9)$$

Как нетрудно проверить, минимум величины  $d \sum_{j=1}^n c_j^2$  при условии (9) достигается при  $c_1 = \dots = c_n = 2/n$ . Эти рассуждения привели к следующей постановке задачи: среди всех квадратур

$$I(f) \approx \frac{2}{n} \sum_{j=1}^n f(x_j),$$

точных для многочленов степени  $q$ , найти квадратуру, соответствующую наименьшему  $n$ . При  $q = 0$  и  $q = 1$  искомой будет квадратура прямоугольников  $I(f) = 2f(0)$ ; при  $q = 2$  и  $q \equiv 3 \pmod{2}$  — квадратура Гаусса (см. § 5).

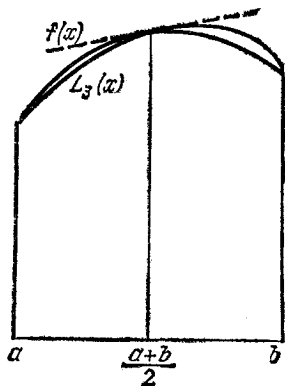


Рис. 3.3.1



## § 4. Ортогональные многочлены

Решения ряда задач математической физики часто исследуют, отыскивая их разложения по ортогональным функциям, в частности по ортогональным многочленам. Наиболее подробно изучены ортогональные системы функций одной переменной. Из ортогональных систем функций многих переменных рассматривают, как правило, лишь системы вида  $\varphi_{n_1}^1(x_1) \times \dots \times \varphi_{n_s}^s(x_s)$ , где  $\varphi_{n_k}^k(x_k)$  — некоторые ортогональные многочлены одной переменной.

Пусть  $H$  — пространство комплекснозначных функций, определенных на  $[a, b]$ , с ограниченным интегралом

$$\int_a^b |f(x)|^2 p(x) dx;$$

скалярное произведение задается равенством

$$(f, g) = \int_a^b f(x) \bar{g}(x) p(x) dx, \quad (1)$$

где  $\bar{g}(x)$  — функция, комплексно сопряженная с  $g(x)$ ,  $p > 0$  почти всюду на  $[a, b]$ ; функции, отличающиеся друг от друга на множестве меры 0, считаются равными.

Система  $\Phi_n$  элементов  $\varphi_1, \dots, \varphi_n$  из  $H$ , т. е.  $\Phi_n = (\varphi_1, \dots, \varphi_n)$ , называется ортогональной, если  $(\varphi_i, \varphi_j) = 0$  при  $i \neq j$ . Система  $\Phi_n = (\varphi_1, \dots, \varphi_n)$  называется линейно независимой, если

$$\sum_{j=1}^n C_j \varphi_j = 0$$

только тогда, когда все  $C_j = 0$ .

Важным аппаратом многих исследований является ортогонализация заданной системы элементов гильбертова пространства.

Лемма 1. Пусть задана линейно независимая система элементов  $\Phi_n = \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_n \end{pmatrix}$ . Можно построить ортогональную линейно

независимую систему  $\Psi_n = \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_n \end{pmatrix}$  элементов вида

$$\psi_j = \sum_{i=1}^j b_{ji} \varphi_i, \quad j = 1, \dots, n, \quad (2)$$

где  $b_{jj} = 1$ .

Мы будем проводить построение такой системы методом индукции. При  $n = 1$  имеем тривиальную систему  $\psi_1 = \varphi_1$ . Пусть

требуемая система  $\Psi_n$  построена при некотором  $n = k$ , тогда элемент  $\psi_{k+1}$  отыскиваем в виде

$$\psi_{k+1} = \varphi_{k+1} - \sum_{i=1}^k a_{ki} \psi_i. \quad (3)$$

Коэффициенты  $a_{ki}$  выбираем из условия ортогональности  $(\psi_{k+1}, \psi_l) = 0$  при  $l \leq k$ . Вследствие ортогональности системы элементов  $\Psi_k$  последнее соотношение представится в виде

$$(\varphi_{k+1}, \psi_l) - a_{kl}(\psi_l, \psi_l) = 0,$$

откуда

$$a_{kl} = \frac{(\varphi_{k+1}, \psi_l)}{(\psi_l, \psi_l)},$$

следовательно, элемент

$$\psi_{k+1} = \varphi_{k+1} - \sum_{i=1}^k \frac{(\varphi_{k+1}, \psi_i)}{(\psi_i, \psi_i)} \psi_i \quad (4)$$

будет ортогонален всем предшествующим. Подставляя в (3)

$$\psi_i = \sum_{q=1}^i b_{iq} \varphi_q$$

при  $i \leq k$ , получим требуемое соотношение.

Совокупность соотношений (2) при  $j \leq n$  можно представить в виде

$$\Psi_n = B_n \Phi_n,$$

где

$$B_n = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots \\ b_{21} & 1 & 0 & 0 & \dots \\ b_{31} & b_{32} & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}.$$

В то же время, перенося все  $\psi_i$  из правой части (4) в левую, получим

$$\Phi_n = A_n \Psi_n,$$

где

$$A_n = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots \\ a_{21} & 1 & 0 & 0 & \dots \\ a_{31} & a_{32} & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}; \quad (5)$$

матрицу  $B_n$  иногда называют *матрицей ортогонализации*. Так как  $\det B_n = 1$ , преобразование, задаваемое матрицей  $B_n$ , является невырожденным и переводит линейно независимую систему элементов  $\Phi_n$  в линейно независимую систему  $\Psi_n$ .

В силу линейной независимости системы  $\Phi_n$  отсюда следует, что  $B_n = A_n^{-1}$ .

При построении ортогональных многочленов в качестве элементов системы функций  $\Phi_j$  берутся функции  $1, \dots, x^{j-1}$  и производится ортогонализация в пространстве со скалярным произведением (1) по описанной выше процедуре. Получаемые многочлены

$$\psi_j(x) = x^{j-1} + \sum_{i=1}^{j-1} b_{ji} x^{i-1} \quad (6)$$

называют *ортогональными многочленами, соответствующими весу  $p(x)$  и отрезку  $[a, b]$* . Иногда ортогональными многочленами, соответствующими весу  $p(x)$ , называют многочлены  $g_j(x) = \alpha_j \psi_j(x)$ , в которых величины  $\alpha_j$  подбирают из каких-либо дополнительных соображений, например из условия  $\|g_j\| = \sqrt{(g_j, g_j)} = 1$ . Систему ортогональных элементов, удовлетворяющих такому условию, называют *ортонормированной*.

Мы уже имели дело с системой многочленов Чебышева

$$T_n(x) = 2^{n-1} x^n + \dots,$$

ортогональных на отрезке  $[-1, 1]$  с весом  $2/(\pi\sqrt{1-x^2})$ . Как отмечалось, значения этих многочленов можно вычислять по рекуррентной формуле

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \quad (7)$$

Вычисление значений ортогональных многочленов Чебышева при помощи формулы (7) более предпочтительно по сравнению с непосредственным вычислением их по явной формуле (6) по следующим причинам.

1. Вычисление по формуле (7) не требует хранения в памяти или вычисления коэффициентов  $b_{ji}$ .

2. Обычно требуется вычислять одновременно значения всех многочленов  $\psi_1(x), \dots, \psi_n(x)$  в одной и той же точке. При независимом вычислении значения каждого многочлена по формуле (6) вычисление значений всех многочленов потребует  $\asymp n^2$  арифметических операций. (Здесь и далее  $a(n) \asymp b(n)$  означает, что  $a(n)$  и  $b(n)$  одного порядка, т. е.  $a(n) = O(b(n))$ ,  $b(n) = O(a(n))$ .)

При одновременном вычислении всех значений при помощи рекуррентного соотношения (7) потребуется  $O(n)$  арифметических операций.

3. Значения  $T_n(x)$ , получаемые при непосредственном вычислении по формуле (6), могут содержать большую вычислительную погрешность.

Дело заключается в следующем: пусть  $T_n(x)$  образуется как сумма слагаемых:

$$T_n(x) = \sum_{j=0}^n d_{nj} x^j. \quad (8)$$

При записи в машине эти слагаемые  $d_n x^j$  смогут приобрести абсолютную погрешность порядка  $|d_n x^j| 2^{-t}$ . Следствием этого может быть погрешность значения  $T_n(x)$  порядка  $D_n(x) 2^{-t}$ , где

$$D_n(x) = \sum_{j=0}^n |d_n x^j|.$$

Оценим снизу  $D_n(x)$ . Из равенства

$$T_n(|x|i) = \sum_{j=0}^n d_n j (|x|i)^j,$$

где  $i$  — мнимая единица, следует оценка

$$|T_n(|x|i)| \leq \sum_{j=0}^n |d_n j x^j| = D_n(x).$$

В то же время, согласно (2.8.4), при действительном  $x$  имеем

$$T_n(|x|i) = \frac{(|x| + \sqrt{1+|x|^2})^n + (|x| - \sqrt{1+|x|^2})^n}{2} i^n.$$

Так как

$$(|x| + \sqrt{1+|x|^2})(|x| - \sqrt{1+|x|^2}) = -1,$$

то отсюда получаем

$$D_n(x) \geq |T_n(|x|i)| \geq \frac{(|x| + \sqrt{1+|x|^2})^n - (|x| - \sqrt{1+|x|^2})^n}{2}.$$

Таким образом, при больших  $n$  и  $x \neq 0$  при непосредственном использовании формулы (8) вычислительная погрешность может достигать значений порядка  $(|x| + \sqrt{1+|x|^2})^n 2^{-t}$ .

Задача 1. Доказать равенство

$$D_n(x) = |T_n(|x|i)|.$$

Мы опять столкнулись здесь с явлением пропадания значащих цифр:  $|T_n(x)| \leq 1$ , но при вычислении значения  $T_n(x)$  из (8) оно получается как сумма больших по модулю слагаемых переменного знака и поэтому приобретает большую погрешность.

В то же время можно показать, что при вычислении по рекуррентной формуле (7) погрешность  $T_n(x)$  имеет порядок  $\min\left(n, \frac{1}{\sqrt{1-x^2}}\right) \cdot O(n2^{-t})$ .

Из изложенного видна важность получения рекуррентных соотношений типа (7), связывающих значения ортогональных многочленов, соответствующих и другим весовым функциям  $p(x)$ .

Справедлива

Теорема (без доказательства). Ортогональные многочлены

$$\psi_j(x) = x^j + \sum_{i=0}^{j-1} b_{ji} x^{i-1}$$

связаны соотношениями

$$\psi_{j+1}(x) = (x + \alpha_j) \psi_j(x) - \beta_j \psi_{j-1}(x), \quad (9)$$

где  $\beta_j > 0$ .

При  $Q_n(x) = \frac{\psi_n(x)}{\|\psi_n(x)\|}$  вместо (9) имеем

$$D_{n+1}Q_{n+1}(x) = (2x + G_n)Q_n(x) - D_nQ_{n+1}(x).$$

Если отрезок  $[a, b]$  конечен, то известно, что  $D_n \rightarrow 1$ ,  $G_n \rightarrow 0$  при  $n \rightarrow \infty$ .

Приведем наиболее употребительные системы ортогональных многочленов, соответствующие различным весовым функциям.

1. Многочлены Якоби. Для отрезка  $[-1, 1]$  и весовой функции  $p(x) = (1-x)^\alpha(1+x)^\beta$ ,  $\alpha, \beta > -1$ , ортогональную систему образуют *многочлены Якоби*

$$P_n^{(\alpha, \beta)}(x) = \frac{(-1)^n}{2^n n!} (1-x)^{-\alpha} (1+x)^{-\beta} \frac{d^n}{dx^n} ((1-x)^{\alpha+n} (1+x)^{\beta+n}).$$

Имеют место соотношения

$$\begin{aligned} \|P_n^{(\alpha, \beta)}(x)\| &= \left( \frac{2^{\alpha+\beta+1} \Gamma(\alpha+n+1) \Gamma(\beta+n+1)}{n! (\alpha+\beta+2n+1) \Gamma(\alpha+\beta+n+1)} \right)^{1/2}, \\ (\alpha+\beta+2n)(\alpha+\beta+2n+1)(\alpha+\beta+2n+2) x P_n^{(\alpha, \beta)}(x) &= \\ &= 2(n+1)(\alpha+\beta+n+1)(\alpha+\beta+2n) P_{n+1}^{(\alpha, \beta)}(x) + \\ &+ (\beta^2 - \alpha^2)(\alpha+\beta+2n+1) P_n^{(\alpha, \beta)}(x) + \\ &+ 2(\alpha+n)(\beta+n)(\alpha+\beta+2n+2) P_{n-1}^{(\alpha, \beta)}(x); \quad (10) \end{aligned}$$

здесь  $\Gamma$  — гамма-функция Эйлера.

Многочлены Якоби удовлетворяют дифференциальному уравнению

$$\begin{aligned} L_{\alpha, \beta} (P_n^{(\alpha, \beta)}(x)) &= (1-x^2) (P_n^{(\alpha, \beta)}(x))_{xx} + \\ &+ ((\beta-\alpha) - (\alpha+\beta+2)x) (P_n^{(\alpha, \beta)}(x))_x = -n(\alpha+\beta+n+1) P_n^{(\alpha, \beta)}(x). \end{aligned}$$

Иначе говоря, они являются собственными функциями дифференциального оператора  $L_{\alpha, \beta}$ .

2. Многочлены Лежандра. Частным случаем многочленов Якоби при  $\alpha = \beta = 0$ , т.е. при весовой функции  $p(x) \equiv 1$ , являются *многочлены Лежандра*

$$L_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n$$

с нормой

$$\|L_n\| = \sqrt{2/(2n+1)},$$

удовлетворяющие рекуррентному соотношению

$$(n+1)L_{n+1}(x) - (2n+1)xL_n(x) + nL_{n-1}(x) = 0.$$

3. Многочлены Чебышева первого рода. При  $\alpha = \beta = -1/2$ ,  $p(x) = 1/\sqrt{1-x^2}$  многочлены Якоби после перенормировки превращаются в *многочлены Чебышева первого рода*  $T_n(x)$ .

4. Многочлены Чебышева второго рода. При  $\alpha = \beta = 1/2$ ,  $p(x) = \sqrt{1-x^2}$  многочлены Якоби после перенормировки превращаются в *многочлены Чебышева второго рода*

$$U_n(x) = (\sin((n+1) \arccos x)) / \sqrt{1-x^2} = T'_{n+1}(x) / (n+1), \quad n = 0, 1, \dots,$$

с нормой  $\|U_n\| = \pi/2$ . Рекуррентное соотношение для многочленов Чебышева второго рода такое же, как для многочленов Чебышева первого рода.

5. Многочлены Эрмита. При  $(a, b) = (-\infty, \infty)$  и  $p(x) = e^{-x^2}$  ортогональную систему образуют *многочлены Эрмита*

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} (e^{-x^2})$$

с нормой

$$\|H_n\| = \sqrt{2^n \cdot n! \sqrt{\pi}},$$

удовлетворяющие рекуррентному соотношению

$$H_{n+1}(x) - 2xH_n(x) + 2nH_{n-1}(x) = 0.$$

Многочлены Эрмита удовлетворяют дифференциальному уравнению

$$H_n'' - 2xH_n' = -2nH_n.$$

6. Многочлены Лагерра. При  $[a, b) = [0, \infty)$  и  $p(x) = x^\alpha e^{-x}$ ,  $\alpha > -1$ , ортогональную систему образуют *многочлены Лагерра*

$$L_n^{(\alpha)}(x) = (-1)^n x^{-\alpha} e^x \frac{d^n}{dx^n} (x^{\alpha+n} e^{-x})$$

с нормой

$$\|L_n^{(\alpha)}\| = \sqrt{n! \Gamma(\alpha + n + 1)}.$$

Для них справедливо рекуррентное соотношение

$$L_{n+1}^{(\alpha)}(x) - (x - \alpha - 2n - 1)L_n^{(\alpha)}(x) + n(\alpha + n)L_{n-1}^{(\alpha)}(x) = 0.$$

Многочлены Лагерра удовлетворяют дифференциальному уравнению

$$x(L_n^{(\alpha)}(x))_{xx} + (\alpha + 1 - x)(L_n^{(\alpha)}(x))_x = -nL_n^{(\alpha)}(x).$$

Рекуррентные соотношения для конкретных ортогональных многочленов, выписанные в пп. 2—4, имеют несколько иной вид, чем (9), поскольку соотношение (9) выписано для ортогональных многочленов, нормированных так, что их старший коэффициент равен 1.

Отметим ряд свойств ортогональных многочленов. Пусть  $P_0(x), \dots, P_n(x)$  — система ортогональных многочленов на отрезке  $[a, b]$  вида

$$P_n(x) = x^n + \sum_{j=0}^{n-1} p_{nj}x^j.$$

*Лемма. Каждый многочлен  $P_n(x)$  имеет ровно  $n$  различных нулей на открытом интервале  $(a, b)$ .*

*Доказательство.* Предположим, что  $P_n(x)$  имеет на  $(a, b)$  только  $l < n$  нулей  $x_1, \dots, x_l$  нечетной кратности. Тогда многочлен

$$P_n(x) \prod_{j=1}^l (x - x_j)$$

не меняет знак на  $[a, b]$ , поэтому

$$\int_a^b P_n(x) \prod_{j=1}^l (x - x_j) p(x) dx \neq 0.$$

С другой стороны, этот интеграл равен нулю, поскольку  $P_n(x)$  ортогонален всем многочленам меньшей степени. Получили противоречие.

**Задача 2.** Пусть  $x_1^{(n)} < \dots < x_n^{(n)}$  — нули  $P_n(x)$ . Тогда нули многочленов  $P_{n-1}(x)$  и  $P_n(x)$  перемежаются, т. е.

$$a < x_1^{(n)} < x_1^{(n-1)} < \dots < x_{n-1}^{(n-1)} < x_n^{(n)} < b.$$

Это свойство ортогональных многочленов используется при составлении таблиц нулей ортогональных многочленов, являющихся нулями квадратур Гаусса.

**Задача 3.** Пусть вес является четной функцией относительно середины отрезка  $[a, b]$  и, для определенности,  $[a, b] = [-1, 1]$ , соответственно  $p(x) = p(-x)$ . Доказать, что все многочлены  $P_{2n}(x)$  четные, многочлены  $P_{2n-1}(x)$  нечетные, т. е.

$$P_j(x) = (-1)^j P_j((-1)^j x) \text{ при всех } j,$$

и рекуррентное соотношение (9) имеет вид

$$P_{j+1}(x) - xP_j(x) + \beta_{j+1}P_{j-1}(x) = 0.$$

При обработке результатов наблюдений возникает задача приближения функций, заданных на множестве точек  $x_1, \dots, x_n$  вещественной оси с помощью многочленов от переменной  $x$ . Эта задача часто решается с помощью ортогональных многочленов дискретной переменной. В теории таких многочленов установлены их свойства, аналогичные свойствам ортогональных многочленов непрерывной переменной; построены дискретные аналоги для всех рассмотренных выше типов ортогональных многочленов непрерывной переменной.

Отметим одно важное свойство распределения нулей ортогональных многочленов. Пусть  $[a, b] = [-1, 1]$ , вес  $p(x)$  почти всюду положителен на  $[-1, 1]$ . Обозначим через  $\omega_n(x_1, x_2)$  число нулей многочлена  $P_n(x)$ , принадлежащих отрезку  $[x_1, x_2]$ . Тогда справедливо соотношение

$$\lim_{n \rightarrow \infty} \frac{\omega_n(x_1, x_2)}{n} = \int_{x_1}^{x_2} \frac{1}{\pi \sqrt{1-x^2}} dx.$$

Таким образом, нули ортогональных многочленов независимо от весовой функции  $p(x)$  распределены асимптотически одинаково, с плотностью  $1/(\pi \sqrt{1-x^2})$ .

## § 5. Квадратурные формулы Гаусса

Из оценки (2.2) следует, что погрешность квадратуры оценивается через погрешность приближения функции многочленами. Функция приближается многочленами более высокой степени точнее, чем многочленами низшей степени:

$$E_0(f) \geq E_1(f) \geq \dots$$

Поэтому есть основания обратить внимание на квадратуры, точные для многочленов по возможности более высокой степени.

Рассмотрим следующую оптимизационную задачу. При заданном числе  $n$  узлов построить квадратуру

$$I(f) = \int_a^b f(x) p(x) dx \approx S_n(f) = \frac{b-a}{2} \sum_{j=1}^n D_j f(x_j), \quad (1)$$

точную для многочленов наиболее высокой степени. Такие квадратуры называют *квадратурами Гаусса*.

Мы видели (§ 2), что квадратура (1) точна для многочленов степени  $m$ , если она точна для всех функций  $x^q$ ,  $q = 0, \dots, m$ . Следовательно, должны выполняться соотношения

$$R_n(x^q) = \int_a^b x^q p(x) dx - \frac{b-a}{2} \sum_{j=1}^n D_j x_j^q = 0, \quad q = 0, \dots, m. \quad (2)$$

Получили систему из  $(m+1)$ -го уравнения относительно неизвестных  $x_1, \dots, x_n, D_1, \dots, D_n$ , где  $x_1, \dots, x_n$  — неизвестные узлы, а  $D_1, \dots, D_n$  — неизвестные коэффициенты квадратурной формулы (1).

При  $m \leq 2n-1$  число уравнений не превосходит числа неизвестных, поэтому можно ожидать, что алгебраическая система (2) имеет решение. Можно попытаться построить квадратурные формулы, соответствующие значению  $m = 2n-1$ , решая эту систему, однако неясно, будут ли узлы квадратур, получаемые из (2), принадлежать отрезку  $[a, b]$ . В противном случае может оказаться, что функция  $f(x)$  не определена в узлах интегрирования и употребление квадратуры невозможно.

Заметим, что в гл. VIII при построении конечно-разностных методов решения обыкновенных дифференциальных уравнений возникнут квадратуры с узлами вне отрезка  $[a, b]$ .

Займемся построением квадратур, соответствующих максимальному значению  $m = 2n-1$ .

**Лемма 1.** Если  $x_1, \dots, x_n$  — узлы квадратуры (1), точной для всех многочленов степени  $2n-1$ , то

$$\int_a^b \omega_n(x) P_{n-1}(x) p(x) dx = 0$$

при  $\omega_n(x) = (x-x_1) \dots (x-x_n)$  и  $P_{n-1}(x)$  — произвольном многочлене степени не выше  $n-1$ .

**Доказательство.** Пусть  $P_{n-1}(x)$  — некоторый многочлен степени не выше  $n-1$ . Вследствие условия леммы квадратура (1) точна для многочлена  $Q_{2n-1}(x) = \omega_n(x) P_{n-1}(x)$  степени



$2n - 1$ . Поэтому

$$\int_a^b \omega_n(x) P_{n-1}(x) p(x) dx = \int_a^b Q_{2n-1}(x) p(x) dx = \\ = \frac{b-a}{2} \sum_{j=1}^n D_j Q_{2n-1}(x_j) = 0.$$

Последнее соотношение вытекает из равенства  $Q_{2n-1}(x_j) = 0$  при всех  $j$ . Лемма 1 доказана.

Далее предполагается, что  $p(x) > 0$  почти всюду на  $[a, b]$ .

Из результатов § 4 вытекает единственность многочлена  $\psi_n(x) = x^n + \dots$ , ортогонального всем многочленам низшей степени, если скалярное произведение задано соотношением

$$(f, g) = \int_a^b f(x) \bar{g}(x) p(x) dx.$$

Поэтому  $\psi_n(x) = \omega_n(x)$  и узлы отыскиваемой квадратуры должны быть нулями  $\psi_n(x)$ . Согласно результатам § 4 многочлен  $\psi_n(x)$  на  $(a, b)$  имеет  $n$  различных нулей.

Лемма 2. Пусть  $x_1, \dots, x_n$  — нули ортогонального многочлена  $\psi_n(x)$  степени  $n$  и (1) — квадратура, точная для многочленов степени  $n-1$ . Тогда квадратура (1) точна для многочленов степени  $2n-1$ .

Доказательство. Произвольный многочлен  $Q_{2n-1}(x)$  степени  $2n-1$  представим в виде

$$Q_{2n-1}(x) = \psi_n(x) g_{n-1}(x) + r_{n-1}(x),$$

где  $g_{n-1}$  и  $r_{n-1}$  — многочлены степени  $n-1$ . Имеем

$$R_n(Q_{2n-1}) = R_n(\psi_n g_{n-1}) + R_n(r_{n-1}) = R_n(\psi_n g_{n-1}),$$

так как  $R_n(r_{n-1}) = 0$  по условию леммы. Далее,

$$R_n(\psi_n g_{n-1}) = \int_a^b \psi_n(x) g_{n-1}(x) p(x) dx - \frac{b-a}{2} \sum_{j=1}^n D_j \psi_n(x_j) g_{n-1}(x_j) = 0,$$

поскольку  $\int_a^b \psi_n(x) g_{n-1}(x) p(x) dx = 0$  вследствие свойства орто-

гональности многочлена  $\psi_n(x)$  многочленам низшей степени, а все  $\psi_n(x_j) = 0$  по предположению леммы. Следовательно,  $R_n(Q_{2n-1}) = 0$ . Лемма 2 доказана.

Теперь можно построить требуемую квадратурную формулу. Для этого зададимся узлами интерполяции  $x_1, \dots, x_n$ , в которых  $\psi_n(x_j) = 0$ , и построим (например, следуя построениям § 3)

квадратуру, точную для многочленов степени  $n - 1$ . В итоге получим требуемую квадратуру

$$\int_a^b f(x) p(x) dx \approx \frac{b-a}{2} \sum_{j=1}^n D_j f(x_j), \quad (3)$$

точную для многочленов степени  $2n - 1$ .

Если почти всюду  $p(x) > 0$ , то не существует квадратуры, точной для всех многочленов степени  $2n$ . В самом деле, возьмем  $Q_{2n}(x) = (x - x_1)^2 \dots (x - x_n)^2$ ; тогда левая часть (1)

$$\int_a^b ((x - x_1) \dots (x - x_n))^2 p(x) dx > 0,$$

а правая равна 0.

*Лемма 3. Коэффициенты  $D_j$  положительны.*

*Доказательство.* Функция  $\left(\frac{\Psi_n(x)}{x - x_l}\right)^2$  является многочленом степени  $2n - 2$ , обращающимся в нуль во всех точках  $x_j \neq x_l$ . Квадратура (3) будет точна для этой функции, поэтому

$$\int_a^b \left(\frac{\Psi_n(x)}{x - x_l}\right)^2 p(x) dx = \frac{b-a}{2} D_l \left(\frac{\Psi_n(x)}{x - x_l}\right)^2 \Big|_{x_l}.$$

Раскрывая выражение  $\Psi_n(x)/(x - x_l)$ , получим

$$D_l = \frac{2}{b-a} \int_a^b \left(\prod_{j \neq l} \frac{x - x_j}{x_l - x_j}\right)^2 p(x) dx > 0.$$

Лемма доказана.

Поскольку все  $D_j > 0$ , то, воспользовавшись (2.1) и (2.2), имеем

$$|R_n(f)| \leq 2 \left(\int_a^b p(x) dx\right) E_{2n-1}(f). \quad (4)$$

Можно получить также оценку погрешности квадратур Гаусса через  $f^{(2n)}(x)$ . Эта оценка имеет вид

$$R_n(f) = f^{(2n)}(\xi) \int_a^b \frac{\Psi_n^2(x)}{(2n)!} p(x) dx. \quad (5)$$

Для практического применения формул Гаусса необходимо иметь в распоряжении узлы и коэффициенты этих квадратур. Можно показать, что для случая  $p(x)$  — четной относительно точки  $(a + b)/2$ , нули ортогональных многочленов, т. е. узлы квадратур Гаусса, расположены симметрично относительно середины отрезка  $[a, b]$ . Вследствие (3.5) коэффициенты квадра-

туры Гаусса (3) будут удовлетворять условию четности  $D_j = D_{n+1-j}$ . Это обстоятельство наполовину уменьшает объем таблиц для формул Гаусса.

Если  $p(x) \equiv 1$ , то коэффициенты  $D_j$  и числа  $d_j = \frac{2x_j - (b+a)}{b-a}$  не зависят от отрезка  $[a, b]$ . В самом деле, если многочлен  $\psi_n(x) = (x-x_1) \dots (x-x_n)$  принадлежит системе многочленов, ортогональных с весом 1 на  $[a, b]$ , то многочлен  $(t-d_1) \dots (t-d_n)$  принадлежит системе многочленов, ортогональных с весом 1 на  $[-1, 1]$ . Поэтому он сам, его нули, а согласно (3.3) и коэффициенты  $D_j$  определяются однозначно, независимо от исходного отрезка  $[a, b]$ .

Приведем для сведения параметры квадратур Гаусса для отрезка  $[-1, 1]$  при  $p(x) \equiv 1$ . В этом случае остаточный член  $R(f)$  для квадратурной формулы (3) есть

$$R(f) = f^{(2n)}(\xi) \frac{2^{2n+1} (n!)^4}{(2n!)^3 (2n+1)}.$$

Вследствие свойства симметрии мы указываем лишь неотрицательные  $d_j$  и коэффициенты при них (табл. 1).

Таблица 1

	$d_1$ $D_1$	$d_2$ $D_2$	$d_3$ $D_3$
1	0,0000000000 2,0000000000		
2	0,5773502692 1,0000000000		
3	0,0000000000 0,8888888888	0,7745966692 0,5555555556	
4	0,8611363115 0,3478548451	0,3399810436 0,6521451549	
5	0,0000000000 0,5688888888	0,9061798459 0,4786286705	0,5384693101 0,2369268851
6	0,9324695142 0,1713244924	0,6612093864 0,3607615730	0,2386191861 0,4679139346

В настоящее время составлены таблицы узлов и весов квадратур Гаусса по крайней мере до  $n = 4096$  с 20 десятичными знаками. Вследствие их большого объема, начиная с некоторого  $n_0$ , их публикуют лишь для  $n$  вида:  $n = 2^k$ ,  $n = 3 \cdot 2^k$ .

Иногда целесообразно видоизменить идею Гаусса построения квадратур, точных на многочленах максимально высокой степени. Например, пусть требуется вычислить  $\int_0^1 f(x) dx$ , а значение  $f(a)$  вычисляется существенно быстрее, чем значения во

внутренних точек отрезка  $[0, 1]$  (или почему-либо заранее известно). Тогда имеет смысл построить квадратуру

$$\int_0^1 f(x) dx \approx \sum_{j=0}^n D_j f(d_j), \quad d_0 = a,$$

точную для многочленов степени  $2n$ . Если вычислять  $\int_{-1}^1 f(x) dx$ , а значения  $f(1)$  и  $f(-1)$  вычисляются существенно быстрее, чем значения во внутренних точках отрезка  $[-1, 1]$ , то имеет смысл построить квадратуру

$$\int_{-1}^1 f(x) dx \approx \sum_{j=0}^n l_j f(d_j), \quad d_0 = -1, \quad d_n = 1, \quad (6)$$

точную для многочленов степени  $2n - 1$ ; в последнем случае оказывается, что  $d_j = d_{n-j}$ ,  $l_j = l_{n-j}$  при всех  $j$ .

Степень многочлена, для которого точна квадратура, определяется числом свободных параметров квадратуры. Квадратура (6) называется *кватратурой Лобатто* или *формулой Маркова*; при  $n = 1$  она совпадает с формулой трапеций, при  $n = 2$  — с формулой Симпсона.

**Задача 1.** Введением весовых функций и заменой переменных  $x = \varphi(t)$  свести построение квадратур (6) к построению некоторых квадратур Гаусса.

**Задача 2.** Пусть  $[a, b] = [-1, 1]$ ,  $p(x) = \frac{1}{\sqrt{1-x^2}}$ . Доказать, что соответствующей квадратурой Гаусса является

$$I(f) = \int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx \approx S_n(f) = \frac{\pi}{n} \sum_{j=1}^n f(x_j),$$

$x_j = \cos \frac{(2j-1)\pi}{2n}$  — нули многочлена Чебышева  $T_n(x)$ .

**Указание.** При проверке точности квадратуры для многочленов степени  $2n - 1$  представить многочлен в виде

$$\sum_{m=0}^{2n-1} a_m T_m(x)$$

и установить, что квадратура точна для  $T_m(x)$  при  $m < 2n$ .

В настоящее время рассчитано много таблиц формул Гаусса и формул типа Лобатто, в частности, при

$$[a, b] = [-1, 1], \quad p(x) = 1,$$

а также в более общем случае при

$$[a, b] = [-1, 1], \quad p(x) = (1+x)^\alpha (1-x)^\beta, \quad \alpha, \beta > -1,$$

и при

$$[a, b] = [0, \infty), \quad p(x) = x^\alpha e^{-x}, \quad \alpha > -1.$$

Если подынтегральная функция интеграла

$$I(f) = \int_0^{\omega} f(x) dx$$

хорошо приближается тригонометрическими многочленами с периодом  $\omega$ , то целесообразно применить квадратуру, являющуюся аналогом квадратуры Гаусса для этого случая вида

$$I(f) \approx S_N(f) = \frac{\omega}{N} \sum_{j=0}^{N-1} f\left(\frac{j\omega}{N}\right). \quad (7)$$

Имеем равенство

$$\begin{aligned} S_N\left(\exp\left\{2\pi m i \frac{x}{\omega}\right\}\right) &= \frac{\omega}{N} \sum_{j=0}^{N-1} \exp\left\{2\pi m i \frac{j\omega}{N}\right\} = \\ &= \begin{cases} \frac{\omega}{N} \sum_{j=0}^{N-1} 1 = \omega & \text{при } \frac{m}{N} \text{ целом,} \\ \frac{\exp\{2\pi m i\} - 1}{\exp\{2\pi m i/N\} - 1} = 0 & \text{при } \frac{m}{N} \text{ не целом.} \end{cases} \end{aligned}$$

В то же время

$$I\left(\exp\left\{2\pi m i \frac{x}{\omega}\right\}\right) = \begin{cases} \omega & \text{при } m = 0, \\ 0 & \text{при } m \neq 0. \end{cases}$$

Следовательно, квадратура (7) точна для функций  $\cos(2\pi m x/\omega)$  при  $m=0$  или при  $m/N$  не целом и для всех функций  $\sin(2\pi m x/\omega)$ . В результате этого оказывается, что квадратура точна для любого тригонометрического многочлена

$$t_N(x) = a_0 + \sum_{m=1}^{N-1} \left( a_m \cos\left(2\pi m \frac{x}{\omega}\right) + b_m \sin\left(2\pi m \frac{x}{\omega}\right) \right) + b_N \sin\left(2\pi N \frac{x}{\omega}\right), \quad (8)$$

следовательно,

$$R_N(f) = R_N(f - t_N) + R_N(t_N) = R_N(f - t_N).$$

Аналогично (4) получаем оценку

$$|R_N(f)| \leq 2\omega \inf_{t_N} \max_{[0, \omega]} |f(x) - t_N(x)|.$$

Нижняя грань берется по множеству всех многочленов вида (8).

Задача 3. Доказать, что не существует квадратур с  $N$  узлами, точных для всех тригонометрических многочленов степени  $N$ .

## § 6. Практическая оценка погрешности элементарных квадратурных формул

Выше получены ряд квадратурных формул и строгие оценки погрешности для них. Однако это не решает всех проблем задачи численного интегрирования. Важнейшей задачей вычислительной математики является создание алгоритмов и их систем, обеспечивающих получение решения задач с заданной точностью при минимальном объеме затрат человеческого труда и работы машины. Практическое применение полученных выше оценок требует аналитических выкладок и поэтому достаточно большого объема работы исследователя; кроме того, эти оценки часто оказываются слишком завышенными. Поэтому при создании таких систем обычно отказываются от использования подобных оценок, зачастую жертвуя строгой гарантией малости погрешности приближенного решения.

Можно говорить, что задача от ее возникновения до получения результата проходит через некоторую систему, состоящую из людей, решающих задачу, и ЭВМ. На первоначальном этапе применения ЭВМ наиболее узким местом, тормозившим работу этой системы, являлось недостаточное количество ЭВМ. Поэтому применение аналитических методов решения или аналитическое проведение оценок погрешности было оправданным.

Однако теперь, с повсеместным распространением вычислительной техники и внедрением ее в различные сферы деятельности общества, обстановка меняется. Узким местом этой системы становятся длительность выбора математической модели, метода решения задачи, программирования и других этапов, предшествующих непосредственному решению задачи на ЭВМ. Прохождение этих этапов особенно замедляется в случае, когда решением задач на ЭВМ занимаются представители конкретных наук, например филологи, медики, экономисты, географы и т. п., мало знакомые с численными методами или программированием. Обучение их тонкостям теории численных методов может превратиться в самоцель, отвлекающую от решения основных задач их науки, и в конечном счете обойтись обществу довольно дорого. Поэтому в настоящее время важнейшей проблемой является создание систем решения задач с максимально простым обращением, предполагающих малую квалификацию пользователя в отношении численных методов и программирования. Например, естественно потребовать, чтобы к программе вычисления интеграла с заданной точностью мог обратиться исследователь, знающий, что такое интеграл, но не умеющий ни интегрировать, ни дифференцировать.

Конечно, в развитии многих областей знания и техники решающая роль математики состоит в создании математической модели явления, а потом уже в применении ЭВМ для ее исследования. При разработке модели от специалиста этой отрасли знания требуется определенная математическая культура, и наше

высказывание не следует понимать как предложение полностью избавиться его от математики.

Подоплекой проводимых здесь рассуждений является следующая известная мысль. Когда мы занимаемся решением каких-то задач, то нужно учитывать эффективность нашей работы не только по совокупным затратам на решение этих задач, но и принимать во внимание убыток, понесенный обществом в результате того, что нами не решены некоторые другие, возможно, более важные задачи.

При практическом анализе погрешности численного интегрирования часто пользуются различными полуэмпирическими приемами. Наиболее распространенным из этих приемов является следующий. Производятся вычисления по двум квадратным формулам

$$I(f) = \int_a^b f(x) dx \approx S^k(f) = \frac{b-a}{2} \sum_{j=1}^{N_k} D_j^k f(x_j^k), \quad k=1, 2;$$

далее некоторая линейная комбинация  $S(f) = S^1(f) + \theta(S^2(f) - S^1(f))$  этих квадратур принимается за приближенное значение интеграла, а величина  $\rho = |S^2(f) - S^1(f)|$  — за меру погрешности приближенной формулы  $I(f) \approx S(f)$ . Довольно типичным является случай  $\theta = 0$ .

Описанный выше подход нельзя считать полностью оправданным вследствие его неоднозначности.

Пусть, например,  $S^1(f)$  — формула Симпсона:

$$\int_{-1}^1 f(x) dx \approx S^1(f) = \frac{f(-1) + 4f(0) + f(1)}{3},$$

$S^2(f)$  — формула трапеций:

$$S^2(f) = f(-1) + f(1)$$

и  $\theta = 0$ . Тогда в качестве меры погрешности выступает величина  $\rho = \frac{2}{3} |f(1) - 2f(0) + f(-1)|$ . Если

$$S^2(f) = 2f(0)$$

— формула прямоугольников, то соответствующее значение  $\rho = \frac{1}{3} |f(1) - 2f(0) + f(-1)|$ . Таким образом, мы получили две различные эмпирические оценки погрешности одной и той же формулы Симпсона.

Попробуем прояснить ситуацию. Выражение  $S(f)$  является некоторой квадратурной суммой

$$S(f) = \frac{b-a}{2} \sum_{j=1}^N D_j f(x_j) \quad (1)$$

по совокупности узлов  $x_j$ , принадлежащих объединению узлов, соответствующих квадратурам  $S^1(f)$  и  $S^2(f)$ . В то же время

$$\rho = |l(f)|,$$

где

$$l(f) = \frac{b-a}{2} \sum_{j=1}^N B_j f(x_j). \quad (2)$$

Возьмем произвольную линейную комбинацию вида (2) и положим

$$S^1(f) = S(f), \quad S^2(f) = S(f) - \theta l(f).$$

Тогда мы получим приближенное значение интеграла  $I(f) \approx S(f)$  с оценкой погрешности  $\rho = \theta |l(f)|$ . Мы видим, что на таком пути можно получить неограниченное множество оценок погрешности одной и той же квадратуры (1).

Рассматриваемую задачу можно формулировать следующим образом. Приближенное значение интеграла вычисляется по формуле

$$I(f) \approx S(f) = \frac{b-a}{2} \sum_{j=1}^N D_j f(x_j). \quad (3)$$

Требуется построить выражение вида (2), дающее представление о погрешности квадратуры (3).

Предположим, что погрешность квадратуры (3) представляется в виде

$$R(f) = \bar{D}(b-a)^{m+1} f^{(m)}(\xi). \quad (4)$$

Рассмотрим случай  $m < N$ . Тогда в качестве  $l(f)$  можно взять величину

$$l(f) = \bar{D}(b-a)^{m+1} m! f(x_{i_1}; \dots; x_{i_{m+1}}),$$

где  $x_{i_1}, \dots, x_{i_{m+1}}$  — различные узлы квадратуры (3). Разделенная разность может быть выражена через производную, поэтому имеем

$$l(f) = \bar{D}(b-a)^{m+1} f^{(m)}(\xi), \quad a < \xi < b.$$

Следовательно, при  $a = \text{const}$ ,  $f^{(m)}(a) \neq 0$ ,  $(b-a) \rightarrow 0$  справедливо соотношение  $R(f) \sim l(f)$  и величину  $\rho = |l(f)|$  можно принять за меру погрешности.

Пусть, например, оценивается погрешность формулы трапеций

$$\int_a^b f(x) dx \approx S(f) = \frac{b-a}{2} \left( f(a) + 0 \cdot f\left(\frac{a+b}{2}\right) + f(b) \right).$$



Согласно оценкам из § 3 имеем

$$R(f) = -\frac{(b-a)^3}{12} f''(\xi).$$

Таким образом, мы можем принять за меру погрешности величину

$$\frac{b-a}{3} \left| f(b) - 2f\left(\frac{a+b}{2}\right) + f(a) \right|.$$

Иначе обстоит дело, когда  $m \geq N$ . Тогда нельзя получить никакого приближения к  $f^{(m)}(\xi)$  через величины  $f(x_1), \dots, f(x_N)$ , и проблема получения эмпирической оценки погрешности в рассматриваемой выше постановке не может быть решена.

Например, мы не можем получить удовлетворительного представления об оценке погрешности формулы Симпсона через значения  $f(a)$ ,  $f\left(\frac{a+b}{2}\right)$ ,  $f(b)$ . Однако можно получить некоторую завышенную оценку погрешности.

Рассмотрим один подход к разрешению возникшей проблемы. Предположим, что нам удалось получить оценку погрешности вида

$$|R(f)| \leq (b-a)^N \bar{D} \max_{[a,b]} |f^{(N-1)}(x)| = \sigma.$$

Положим

$$\rho = (b-a)^N \bar{D} (N-1)! |f(x_1; \dots; x_N)|. \quad (5)$$

При  $a = \text{const}$ ,  $f^{(N-1)}(a) \neq 0$ ,  $(b-a) \rightarrow 0$  имеем  $\sigma \sim \rho$ . Таким образом, величину  $\rho$  можно принять за приближенную оценку погрешности формулы (3). Эта оценка будет сильно завышенной, поскольку при предположении  $m \geq N$  имеем  $R(f) = O((b-a)^{N+1})$ . Однако лучшей оценки погрешности формулы (3) по сравнению с оценкой через  $\sigma$ , по-видимому, нельзя предложить. В случае формулы Симпсона верна оценка (5) при  $\bar{D} = \frac{1}{81}$  и, таким образом, за меру погрешности принимаем величину

$$\frac{4(b-a)}{81} \left| f(b) - 2f\left(\frac{a+b}{2}\right) + f(a) \right|. \quad (6)$$

В случае многомерных интегралов все практические способы оценки погрешности опираются на исходную, раскрытую нами процедуру. Дело в том, что в многомерном случае погрешность оценивается через значения нескольких производных подынтегральной функции. Получение «обоснованных» оценок, подобных (6), для таких формул крайне затруднительно. Поэтому обращаются к исходной процедуре с последующей экспериментальной проверкой результатов ее применения.

## § 7. Интегрирование быстро осциллирующих функций

Пусть требуется вычислить интеграл

$$\int_a^b f(x) \exp \{i\omega x\} dx,$$

где  $\omega(b-a) \gg 1$ ,  $f(x)$  — гладкая функция. Функции

$$\operatorname{Re}(f(x) \exp \{i\omega x\}), \quad \operatorname{Im}(f(x) \exp \{i\omega x\})$$

имеют на рассматриваемом отрезке примерно  $\omega(b-a)/\pi$  нулей. Поскольку многочлен степени  $n$  имеет не более  $n$  нулей, то такие функции могут быть хорошо приближены многочленами степени  $n$  лишь при  $n \gg \omega(b-a)/\pi$ . Поэтому для непосредственного вычисления интегралов от таких функций потребуется применение квадратур, точных для многочленов высокой степени.

Более выгодным может оказаться путь рассмотрения функции  $\exp \{i\omega x\}$  как весовой.

Как и в § 1, зададимся узлами интерполирования

$$x_j = \frac{b+a}{2} + \frac{b-a}{2} d_j, \quad j = 1, \dots, n,$$

и заменим исходный интеграл на  $\int_a^b L_n(x) \exp(i\omega x) dx$ , где  $L_n(x)$  — интерполяционный многочлен с узлами  $x_j$ . Последний интеграл может быть вычислен в явном виде

$$\begin{aligned} \int_a^b L_n(x) \exp \{i\omega x\} dx &= S_n^\omega(f) = \\ &= \frac{b-a}{2} \exp \left\{ i\omega \frac{a+b}{2} \right\} \sum_{j=1}^n D_j \left( \omega \frac{b-a}{2} \right) f(x_j), \end{aligned}$$

где

$$D_j(p) = \int_{-1}^1 \left( \prod_{k \neq j} \frac{\xi - d_k}{d_j - d_k} \right) \exp \{ip\xi\} d\xi. \quad (1)$$

Получилась квадратурная формула

$$\int_a^b f(x) \exp \{i\omega x\} dx \approx S_n^\omega(f) \quad (2)$$

с остаточным членом

$$R_n(f) = \int_a^b (f(x) - L_n(x)) \exp \{i\omega x\} dx.$$

В соответствии с (1.1)

$$|R_n(f)| \leq \int_a^b |f(x) - L_n(x)| dx \leq \leq D(d_1, \dots, d_n) (\max_{[a, b]} |f^{(n)}(x)|) \left(\frac{b-a}{2}\right)^{n+1}.$$

Вычисление интегралов подобного рода является типичной задачей, встречающейся при разложении функций в ряды Фурье, при построении диаграмм направленности антенн и т. д.

В стандартных программах вычисления интегралов от быстро осциллирующих функций используются формулы (1), (2), соответствующие случаям:  $n = 3$ ,  $d_1 = -1$ ,  $d_2 = 0$ ,  $d_3 = 1$  (эту формулу называют *формулой Филона*) или  $n = 5$ ,  $d_1 = -1$ ,  $d_2 = -0,5$ ,  $d_3 = 0$ ,  $d_4 = 0,5$ ,  $d_5 = 1$ .

Если формулы (1), (2) использовать для вычисления интегралов от функций, не являющихся быстро осциллирующими, то может возникнуть следующая неприятность, которую мы проиллюстрируем для  $n = 2$ ,  $d_1 = -1$ ,  $d_2 = 1$ . В этом случае

$$D_1(p) = \int_{-1}^1 \frac{1-\xi}{2} \exp\{ip\xi\} d\xi = \frac{\sin p}{p} + \frac{p \cos p - \sin p}{p^2} i,$$

$$D_2(p) = \int_{-1}^1 \frac{1+\xi}{2} \exp\{ip\xi\} d\xi = \frac{\sin p}{p} - \frac{p \cos p - \sin p}{p^2} i.$$

При  $p \rightarrow 0$  имеем

$$\frac{p \cos p - \sin p}{p^2} = -\frac{p}{3} + O(p^3) \rightarrow 0, \quad \frac{\sin p}{p} \rightarrow 1.$$

Таким образом,  $D_1(p), D_2(p) \rightarrow 1$  при  $p \rightarrow 0$ .

Пусть  $p$  — малое число. Функции  $\cos p$  и  $\sin p$  вычисляются в машине с погрешностями  $O(2^{-t})$  и  $O(p2^{-t})$  соответственно. Вследствие этого коэффициенты  $D_1(p), D_2(p)$  приобретают погрешность  $O(2^{-t}/p)$ . При  $n > 2$  оказывается, что погрешность коэффициентов  $D_i(p)$ , вычисляемых по формулам (1), может оказаться величиной порядка  $2^{-t}/p^{n-1}$ . Например, при  $t = 30$ ,  $n = 5$ ,  $p = 0,01$  такая погрешность уже недопустима. Поэтому стандартные программы вычисления интегралов от быстро осциллирующих функций должны иметь специальный блок, предусматривающий изменение расчетных формул при малых  $p$  с тем, чтобы избежать существенного влияния вычислительной погрешности.

Если  $n$  не очень велико, например  $n = 2$ , то можно пойти по следующему пути: при  $|p| > p_n$ , где некоторое  $p_n$  подбираем экспериментально, вычисления производим по формулам (1), (2); если  $|p| \leq p_n$ , то вычисляем исходный интеграл по формуле трапеций (3.7), рассматривая всю функцию  $f(x)$  ×

$\times \exp\{i\omega x\}$  как подынтегральную. В рассматриваемом случае ( $n = 2, d_1 = -1, d_2 = 1$ ) формула (3.7) приобретает вид

$$\int_a^b f(x) \exp(i\omega x) dx \approx \frac{b-a}{2} [\exp\{i\omega a\} f(a) + \exp\{i\omega b\} f(b)]. \quad (3)$$

Формулы (1), (3) можно объединить в одну формулу

$$\int_a^b f(x) \exp\{i\omega x\} dx = \frac{b-a}{2} \exp\left\{i\omega \frac{a+b}{2}\right\} [D_1(p) f(a) + D_2(p) f(b)], \quad (4)$$

где

$$p = \omega \frac{b-a}{2},$$

$$D_{1,2}(p) = \begin{cases} \frac{\sin p}{p} \pm \frac{p \cos p - \sin p}{p^2} i & \text{при } |p| > p_2, \\ \exp\{\mp pi\} & \text{при } |p| \leq p_2. \end{cases} \quad (5)$$

В стандартных программах вычисления интегралов рассматриваемого типа применяются квадратуры типа (4), (5). Возникает вопрос: зачем усложнять стандартную программу? Может быть, проще написать стандартную программу вычисления по формуле (2), стандартную программу вычисления по формуле трапеций и дать инструкцию: при больших  $|p|$  обращаться к первой из стандартных программ, при малых  $|p|$  — ко второй? Обсудим целесообразность такого подхода. Нашей целью является создание максимальных удобств пользователю ЭВМ. Если в описании правил использования стандартной программы написано слишком много, то пользователь может не понять того, что написано, и

1) обратиться к другой программе, худшей по качеству, но имеющей лучшее описание или более простое обращение;

2) воспользоваться программой неправильно и не получить результатов; например, в рассматриваемом случае, применив метод расчета по формулам (1), (2) при  $p = 0$ , он получит аварийную остановку машины;

3) воспользоваться программой неправильно и получить неверный результат, предполагая, что он верный; например, так случится, если он воспользуется формулами (1), (2) при  $2^{-i}/p^{n-i}$  порядка 1.

Очевидно, что последний случай влечет за собой наиболее неприятные для пользователя последствия.

При выборе метода для стандартной программы надо стремиться к тому, чтобы описание программы, составленной на базе этого метода, было по возможности простым и кратким. Надо иметь в виду, что всякое дополнительное высказывание может быть истолковано неправильно в ущерб для области применения программы.

Приведем один поучительный реальный пример.

В описании одной из стандартных программ вычисления кратных интегралов было написано: «Применение настоящей программы нецелесообразно, если число узлов берется большим 100 000». Спустя короткое время программа была практически изъята из употребления. Оказалось, что среди первых пользователей программ подавляющее большинство сразу задавались числом узлов  $n = 100\,000$ , при этом на вычисление простого интеграла уходило слишком много машинного времени. Весть об этом распространилась повсюду, и вскоре к программе перестали обращаться. На самом деле при ее использовании большинство реальных интегралов вычислялось с приемлемой точностью при числе узлов порядка 1000. Число 100 000 было указано лишь как ориентировочная верхняя граница значений, при которых вычислительная погрешность еще не оказывает катастрофического влияния на результат,

## § 8. Повышение точности интегрирования за счет разбиения отрезка на равные части

Из оценки (2.1) вытекает, что погрешность квадратуры оценивается через погрешность, с которой функция  $f(x)$  может быть приближена многочленами степени  $m$ . Поэтому может показаться естественным добиваться увеличения точности за счет повышения степени многочленов, для которых эта квадратура точна. Однако такой путь содержит свои «подводные камни». При неудачном выборе узлов может оказаться, что величина  $\sum_{j=1}^n |C_j|$  растет вместе с  $n$ . Тогда в оценке (2.1) величина  $V$  также растет с  $n$  и может оказаться, что уменьшение  $E_m(f)$  с ростом  $n$  не компенсирует увеличения  $V$ .

Например, для простейшего равномерного распределения узлов

$$d_j^{(n)} = -1 + \frac{2(j-1)}{n-1}, \quad j=1, \dots, n, \quad [a, b] = [-1, 1],$$

оказывается, что при  $m = n - 1$  имеем  $\log_2 V \sim n$ . В результате этого, например, для аналитической функции  $f(x) = (1 + 25x^2)^{-1}$  при таких узлах

$$|R_n(f)| = \left| \int_{-1}^1 f(x) dx - S_n(f) \right| \rightarrow \infty$$

при  $n \rightarrow \infty$ .

Рассмотрим случай, когда отрезок интегрирования есть  $[-1, 1]$ , и сформулируем общую теорему, указывающую на необходимость осторожного обращения с формулами, точными для многочленов очень высокой степени. Пусть при каждом  $n$  мы имеем квадратуру

$$\int_{-1}^1 f(x) dx \approx \sum_{j=1}^n C_j^{(n)} f(d_j^{(n)}), \quad (1)$$

точную для многочленов степени  $n - 1$ . Обозначим через  $\omega_n(x_1, x_2)$  число узлов квадратуры (1), принадлежащих отрезку  $[x_1, x_2]$ .

**Теорема (без доказательства).** Пусть существует отрезок  $[x_1, x_2] \in \in [-1, 1]$  такой, что

$$\frac{\omega_n(x_1, x_2)}{n} \rightarrow \int_{x_1}^{x_2} \frac{1}{\pi \sqrt{1-x^2}} dx \quad \text{при } n \rightarrow \infty.$$

Тогда можно указать  $b \neq 0$  и  $a$  такие, что для одной из аналитических функций

$$\operatorname{Re} \left( \frac{1}{x - (a + bi)} \right) \quad \text{или} \quad \operatorname{Im} \left( \frac{1}{x - (a + bi)} \right)$$

будет выполняться соотношение

$$\lim_{n \rightarrow \infty} |R_n(f)| = \infty,$$

Таким образом, узлы квадратур (1), точных для многочленов степени  $n-1$ , при больших  $n$  должны располагаться с такой же плотностью, как нули ортогональных многочленов, т. е. как и нули квадратур Гаусса. Иначе такие квадратуры нельзя рассматривать как универсальные.

Перепишем оценку (5.4) погрешности формул Гаусса:

$$|R_n(f)| \leq 2 \left( \int_a^b p(x) dx \right) E_{2n-1}(f). \quad (2)$$

Пусть подынтегральная функция  $f(x)$  непрерывна. Тогда, согласно теореме Вейерштрасса, при любом  $\varepsilon > 0$  найдется многочлен  $P_q(x)$ , для которого  $\max_{[a, b]} |f(x) - P_q(x)| \leq \varepsilon$ , откуда следует, что  $E_m(f) \rightarrow 0$  при  $m \rightarrow \infty$ . Таким образом, для любой непрерывной функции  $f(x)$  погрешность формул Гаусса  $R_n(f) \rightarrow 0$  при  $n \rightarrow \infty$ .

**Задача 1.** Пусть  $f(x)$  — функция, интегрируемая по Риману. Доказать, что для формул Гаусса  $R_n(f) \rightarrow 0$  при  $n \rightarrow \infty$ .

Из сказанного выше видно, что формулы Гаусса могли бы быть положены в основу универсальных программ вычисления интегралов с заданной точностью. При этом придется вводить в ЭВМ или считывать с внешних запоминающих устройств в оперативную память узлы и веса этих квадратур.

Во многих случаях возникает задача вычисления интегралов, где подынтегральная функция или ее производные невысокого порядка имеют участки резкого изменения, например обращаются в бесконечность. Такие функции плохо приближаются многочленами сразу на всем отрезке интегрирования. Здесь часто оказывается более выгодным разбить исходный отрезок на части и на каждой части применять свою квадратурную формулу, Гаусса или какую-либо другую.

Пусть вычисляется интеграл  $I = \int_A^B f(x) dx$ . Разобьем отрезок  $[A, B]$  на  $M$  частей  $[a_{q-1}, a_q]$ , где  $a_0 = A$ ,  $a_M = B$ . Для вычисления интеграла по каждой из частей применим какую-либо квадратурную формулу из § 1,3 вида

$$I_q = \int_{a_{q-1}}^{a_q} f(x) dx \approx s_q = \frac{a_q - a_{q-1}}{2} \sum_{j=1}^n C_j f \left( \frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j \right) \quad (3)$$

с оценкой остаточного члена

$$R_q(f) \leq D \left( \max_{[a_{q-1}, a_q]} |f^{(m)}(x)| \right) \left( \frac{a_q - a_{q-1}}{2} \right)^{m+1}.$$

В результате интеграл по всему отрезку будет аппроксимирован

суммой

$$S_M^n(f) = \sum_{q=1}^M s_q = \sum_{q=1}^M \frac{a_q - a_{q-1}}{2} \sum_{j=1}^n C_j f\left(\frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \quad (4)$$

с оценкой остаточного члена

$$|R_M^n(f)| = |I - S_M^n(f)| \leq D \sum_{q=1}^M \left( \max_{[a_{q-1}, a_q]} |f^{(m)}(x)| \right) \left( \frac{a_q - a_{q-1}}{2} \right)^{m+1}. \quad (5)$$

Выражение (4) часто называют *составной* или *обобщенной квадратурной формулой*.

Рассмотрим наиболее простой для исследования случай, когда отрезки разбиения имеют одинаковую длину  $a_q - a_{q-1} \equiv H$ . Тогда оценка погрешности (5) после замены  $\max_{[a_{q-1}, a_q]} |f^{(m)}(x)|$  на величину  $A_m = \max_{[A, B]} |f^{(m)}(x)|$  приобретает вид

$$|R_M^n(f)| \leq \bar{D} A_m (B - A) H^m, \quad \text{где } \bar{D} = 2^{-(m+1)} D, \quad (6)$$

или

$$|R_M^n(f)| \leq D A_m \frac{(B - A)^{m+1}}{M^m}. \quad (7)$$

Приведем конкретные квадратурные формулы и оценки погрешности для частных случаев формул (3).

1. Составная формула трапеций с постоянным шагом интегрирования. В этом случае при постоянном шаге  $a_q - a_{q-1} \equiv H$  формула (3) приобретает вид

$$\int_{a_0}^{a_M} f(x) dx \approx H \left( \frac{f(a_0)}{2} + f(a_1) + \dots + f(a_{M-1}) + \frac{f(a_M)}{2} \right),$$

а остаточный член оценивается следующим образом:

$$|R_M^2(f)| \leq A_2 \frac{(a_M - a_0) H^2}{12} = A_2 \frac{(a_M - a_0)^3}{12M^2}. \quad (8)$$

2. Составная формула Симпсона с постоянным шагом интегрирования. При постоянном шаге  $a_q - a_{q-1} \equiv H = 2h$  формула (3) приобретает вид

$$\int_{a_0}^{a_M} f(x) dx \approx H \left( \frac{1}{6} f(a_0) + \frac{4}{6} f(a_{1/2}) + \frac{1}{3} f(a_1) + \frac{4}{6} f(a_{3/2}) + \dots + \frac{1}{3} f(a_{M-1}) + \frac{4}{6} f(a_{M-1/2}) + \frac{1}{6} f(a_M) \right),$$

где  $a_j = a_0 + jH$ . Для остаточного члена справедлива оценка

$$|R_M^4(f)| \leq \frac{A_4(a_M - a_0)H^4}{2880} = \frac{A_4(a_M - a_0)^5}{2880M^4} = \frac{A_4(a_M - a_0)}{180} h^4, \quad h = \frac{H}{2}.$$

Последняя запись оценки наиболее употребительна.

Мы получили формулы с порядком погрешности  $O(N^{-m})$  по отношению к общему числу узлов интегрирования  $N = Mn$  в предположении ограниченности  $|f^{(m)}(x)|$ . Заметим, что в случае формул трапеций и Симпсона общее число узлов  $N$  оказалось меньше, чем  $Mn$ , поскольку концы элементарных отрезков  $[a_{q-1}, a_q]$  были узлами интегрирования и значения функции в этих концах использовались для вычисления интегралов по двум соседним элементарным отрезкам.

**Задача 2.** Пусть  $\int_A^B |f^{(q)}(x)| dx < \infty$ ,  $q \leq 2$ ; получить оценку погрешности формулы трапеций

$$|R_M^2(f)| \leq \gamma_q \left( \int_A^B |f^{(q)}(x)| dx \right) H^q, \quad (9)$$

где  $\gamma_q$  — абсолютная постоянная.

**Задача 3.** Пусть  $\int_A^B |f^{(q)}(x)| dx < \infty$ ,  $q \leq 4$ ; получить оценку погрешности формулы Симпсона

$$|R_M^4(f)| \leq \beta_q \left( \int_A^B |f^{(q)}(x)| dx \right) H^q,$$

где  $\beta_q$  — абсолютная постоянная.

Пусть, например, вычисляется

$$\int_0^1 (\sin x)^\alpha dx, \quad 0 < \alpha < 1.$$

Так как  $((\sin x)^\alpha)' \rightarrow \infty$  при  $x \rightarrow 0$ , то мы не можем получить никаких оценок погрешности через  $\max |f'(x)|$ . В то же время функция  $(\sin x)^\alpha$  монотонна на отрезке  $[0, 1]$ , поэтому

$$\int_0^1 |((\sin x)^\alpha)'| dx = (\sin x)^\alpha \Big|_0^1 = 1.$$

Следовательно, при использовании квадратуры (4), соответствующей  $m = 2$  с постоянным шагом  $a_q - a_{q-1} = H$ , согласно (9) имеем оценку погрешности  $O(1/N)$ .

В данном примере из оценки (6) малость погрешности не следовала; в то же время на основании (9) мы заключаем, что эта погрешность порядка  $O(1/N)$ .

Не следует думать, что в случае функций с малым числом ограниченных производных составные формулы численного интегрирования имеют лучший



порядок сходимости по сравнению с формулами Гаусса. Предположим, что подынтегральная функция имеет  $q$  ограниченных производных. Тогда, применяя составную формулу (4), соответствующую  $n = q$ , получим приближенное значение интеграла с погрешностью  $O(N^{-q})$ . С другой стороны, известно, что для такой функции  $E_{2N-1}(f) = O(N^{-q})$ . Поэтому из (5.4) следует оценка погрешности формулы Гаусса с  $N$  узлами

$$R_N(f) = O(N^{-q}).$$

Таким образом, порядок оценки в обоих случаях одинаков.

Обратим внимание еще на одно удобство использования формул Гаусса сразу по всему отрезку интегрирования. Не нужно оценивать число  $q_0$  ограниченных производных подынтегральной функции и в соответствии с этим выбирать наиболее подходящую формулу численного интегрирования по отрезкам разбиения — при применении формул Гаусса порядок погрешности  $O(N^{-q_0})$  обеспечивается автоматически. Конечно, не нужно думать, что формула, имеющая более высокий порядок скорости сходимости, при конкретном числе узлов всегда точнее формулы более низкого порядка скорости сходимости.

## § 9. О постановках задач оптимизации

Мы получили ряд формул численного интегрирования и могли бы получить еще большее количество таких формул. Возникает вопрос: можно ли получить лучшие по порядку оценки погрешности при тех же предположениях о подынтегральной функции или хотя бы улучшить константы в этих оценках? Изучение этого вопроса приводит к задаче построения квадратур с оптимальной оценкой погрешности, или, как говорят, оптимальных квадратур. В связи с этим возникает следующая проблема — чем больше способов решения задачи, тем труднее решиться выбрать какой-то из них, поскольку каждый способ может иметь свои положительные качества: простота программирования, малое время работы ЭВМ, малая загрузка памяти, простота оценки погрешности, применимость к широкому кругу задач. Таким образом следует иметь в виду, что иногда излишняя информация о способах решения задач при большом их количестве может также и затормозить реальное решение задач. Поэтому необходима какая-то систематизация методов решения, их отбор. Естественно пытаться решить задачи наилучшим, оптимальным способом. Мы уже рассматривали некоторые модельные формулировки задач об оптимальных методах решения на примере вычисления значений функций; при этом рассмотрении возникают определенные математические постановки задач, требующих решения.

При рассмотрении каждой новой проблемы желательно получить ее наиболее подробное, наилучшее описание и затем решить возникшую задачу наилучшим образом. Однако обычно это не удается сделать и приходится довольствоваться меньшим: описать проблему наилучшим образом и решить ее удовлетворительно или описать проблему удовлетворительно и затем полностью решить возникшую задачу. При рассмотрении проблемы

оптимизации методов можно говорить о выборе между удовлетворительным решением проблемы оптимизации методов для классов задач, близких к реальным, или полным решением проблемы для эталонных математических классов, подобных рассматриваемому в следующем параграфе. Может вызвать недоумение высказывание об «удовлетворительном решении проблемы оптимизации методов» для каких-то классов — ведь задача оптимизации методов на классе сводится к вполне конкретной математической задаче, которую, по-видимому, можно решить окончательно, а не «удовлетворительно».

В принципе это высказывание верно, однако обычно полностью решить задачу в приемлемое для практики время не удается, так как время, необходимое для построения оптимального метода, обычно существенно превосходит время, в течение которого возникает новое, уточненное описание классов рассматриваемых задач. Также надо иметь в виду, что не всегда удается формализовать это математическое описание класса реальных задач: имеется какое-то качественное представление о классе, имеются неплохие численные методы и интуитивно ясно, что в практическом плане оптимальность методов достигнута; в то же время нет даже четко формализованного описания класса решаемых задач. Например, реально требуется вычислять интегралы от кусочно-гладких функций, однако в течение долгого времени так и не удалось предложить описание множества таких функций, которое соответствовало бы реальной практике вычислений. Точно так же, например, при анализе физических моделей можно описать задачу с помощью сложной системы дифференциальных уравнений и решить эту систему с малой точностью или описать задачу более грубо с помощью простой системы и решить эту систему с большой точностью.

В разных случаях, в зависимости от конкретной обстановки, бывает целесообразен тот или иной подход к решению задачи. Представляется, что для прикладной науки часто более существенно не окончательное решение вариационных задач, а правильная их постановка. Обычно больший эффект дает удовлетворительное решение правильно поставленной в практическом смысле вариационной задачи, чем полное решение упрощенной задачи, не охватывающей все существенные характерные черты исходной проблемы.

Что такое оптимальный метод решения задачи?

Под оптимальным методом решения задачи можно понимать метод, требующий минимальной затраты машинного времени. Но это будет неправильная постановка, поскольку, в принципе, всякую задачу можно решить без применения машин, затратив на это очень большое количество времени.

Под оптимальным можно понимать метод, требующий минимальных затрат времени (или соответственно материальных затрат). При этом следует помнить, что время и затраты на поиск оптимального алгоритма уже входят в решение задачи.

Когда мы ставим вопрос о решении задачи оптимальным образом, мы не учитываем разных возможностей исследователей, решающих задачу. А ведь в зависимости от индивидуальных возможностей исследователя, наличия ЭВМ, библиотеки, лаборантов, сотрудников, обладающих опытом решения подобных задач, оптимальное решение, вообще говоря, будет различным.

Мы говорили об оптимальном решении каждой конкретной задачи. Такая постановка вопроса не совсем правильна также по следующей причине. На самом деле коллектив исследователей сталкивается с целой совокупностью задач. Мы можем поставить перед собой цель решить первую, вторую и т. д. задачи за кратчайшее время. Однако постановка вопроса в корне изменится, если мы зададимся целью решить с минимальными затратами целую совокупность задач. Мы можем решать первую, вторую и т. д. задачи оптимальными методами, но за это время наука уйдет вперед, и если мы не будем изучать новые методы, создавать новые алгоритмы и стандартные программы, рассчитанные на решение будущих задач, то в целом мы проиграем. При наличии большой серии задач, не требующей сверхсрочного решения, выгоднее заняться теоретическими исследованиями, создать новые алгоритмы, а затем уже решать эти задачи.

Нет однозначного ответа также на вопрос о том, кто должен заниматься оптимальным решением задач. Если мы имеем дело с единственной конкретной задачей, заведомо не требующей нового алгоритма и больших затрат машинного времени, ее лучше поручить выполнить работнику низкой квалификации. Увеличение числа задач и их сложности требует привлечения работников высокой квалификации, поскольку здесь их отдача будет наиболее полной. Конечно, важно изучить опыт решения как сложных, так и простых задач подобного типа. Для того чтобы задачу оптимизации методов можно было рассматривать как чисто математическую задачу, необходимо определить целевую функцию исследования, класс рассматриваемых задач и возможности исследователя.

Качество какого-то алгоритма на классе задач мы характеризуем его качеством на самой «плохой» для него задаче этого класса. Поэтому чем уже класс рассматриваемых задач, тем лучше качество алгоритмов на этом классе.

Казалось бы, следует построить оптимальные алгоритмы для решения возможно большего числа классов задач. Однако при излишней детализации слишком много усилий уйдет на отыскание оптимальных методов решения. В этом случае может не хватить сил и времени на реальное решение самих задач, разработку и создание стандартных программ. Конечно, эти замечания против детализации в меньшей мере относятся к случаю специализированных машин и устройств, где следует максимально сузить класс рассматриваемых задач.

При первом взгляде на проблему кажется, что в практике вычислений всегда имеют дело с конкретными задачами и ни-

когда не рассматривают класс задач. По этому поводу можно сказать следующее: формально при выборе метода решения задачи исследователь не относит ее к какому-либо классу; однако метод решения всегда выбирается в зависимости от некоторых типичных свойств задачи: тип дифференциального уравнения, наличие особенности у решения, число конечных производных и т. п. При выборе алгоритма исследователь учитывает эти свойства и тем самым вольно или невольно относит рассматриваемую задачу к классу задач, обладающих этими свойствами.

При анализе задач и разбиении их на классы часто возникает вопрос: какому классу задач уделить первоочередное внимание, в частности, на какой класс задач следует рассчитывать при составлении стандартной программы численного интегрирования? Пусть, например, при помощи этой программы примерно с одинаковой частотой будут вычисляться интегралы как от гладких, так и от не очень гладких функций. На какие функции нужно ориентироваться при выборе алгоритма для этой программы? Ответ на этот вопрос можно получить из следующих соображений. На гладких функциях алгоритм будет работать более эффективно, поэтому время, затрачиваемое на их решение, будет меньше времени, затрачиваемого на решение задач с негладкими функциями. 50 % экономии времени при вычислении интегралов от гладких функций принесет меньшую выгоду, чем 50 % экономии времени при вычислении интегралов от негладких функций. Следовательно, целесообразнее уделить внимание эффективности алгоритма в случае негладких функций. Конечно, вывод будет другим, если доля негладких функций мала.

При отыскании оптимального метода решения класса задач в случае, когда предполагается непосредственное внедрение алгоритма в вычислительную практику, обычно возникает следующая проблема: нам не удастся сразу найти оптимальный метод решения задач рассматриваемого класса. После некоторого времени работы мы находим какой-то метод, иногда близкий к оптимальному, иногда просто лучший, чем ранее известные, а затем постепенно усовершенствуем его. В какой момент надо остановиться в усовершенствовании метода и перейти к составлению стандартных программ решения этого класса задач? При ответе на этот вопрос надо учитывать следующие простые соображения. Если мы поспешим и быстро начнем внедрять только что полученный алгоритм, то, может быть, нам придется вскоре создавать новые алгоритмы из-за плохого качества этого алгоритма. Затягивание времени при составлении стандартных программ также нежелательно: при этом мы допустим большой перерасход машинного времени при решении задач по имеющимся стандартным программам.

Можно возразить, что создание совершенных стандартных программ приведет к большой экономии машинного времени в будущем. Однако при анализе этого возражения в свою очередь следует учесть, что машинное время становится все менее

дефицитным и более дешевым. Кроме того, чем раньше мы решим ту или иную задачу, тем большую реальную пользу получит общество в целом.

Нужно учитывать также важность быстрого внедрения вновь созданных алгоритмов для самой задачи отыскания оптимальных методов решения. Дело в том, что анализ результатов расчетов может указать на необходимость изменения класса рассматриваемых задач и открыть путь для новых теоретических исследований. Мы видим, что подход к проблеме оптимальности должен носить динамический характер: должны строиться оптимальные или близкие к ним методы для все новых классов задач, предъявляемых наукой и техникой, при изменении возможностей, предоставляемых ЭВМ.

При этом необходима как текущая работа, так и работа по доведению предложенных ранее постановок до окончательного решения. Для вычислительной математики, как и для всякой прикладной науки, характерна следующая обстановка. Обычно задачи новых типов предъявляются сначала в незначительном количестве и требуется срочное их решение любой ценой, не считаясь ни с какими затратами. На первом этапе применяется первый попавшийся приемлемый метод. Далее эти задачи поступают в большом количестве, производится более или менее удовлетворительная постановка задачи оптимизации методов и находится некоторое ее решение. Затем задача переводится на поток, т. е. решение задач этого типа производится при помощи систем соответствующих стандартных программ. Не следует думать, что на этом этапе полностью кончается исследование данного типа задач — чтобы создавать эффективные методы решения новых задач, нужно осмысливать те задачи, которые остались несколько позади, и проводить их теоретическое изучение. Иногда бывает целесообразно работу по быстрому решению первых поступивших задач серии и перспективную работу по созданию эффективных методов решения с самого начала организовать параллельно.

Соотношение между текущей и перспективной работой также является важнейшим фактором жизнедеятельности любой научной организации. В каждый момент времени организации предъявляются некоторые требования, выполнение которых необходимо для ее существования: самоокупаемость в случае хозяйственных организаций, своевременная отчетность по годовому плану и т. п. Однако существуют и критические моменты времени, когда предъявляются повышенные требования к работе, например, необходимость своевременного решения новых классов задач. Поэтому при планировании работы должны предусматриваться какие-то теоретические разработки впрок, резервы людей, машинного времени.

Вернемся к вопросу об оптимизации методов.

Часто математик, создав оптимальный или близкий к нему метод решения задачи, сетует на то, что этот метод плохо внед-

руется в вычислительную практику. Ответ на этот вопрос может быть самым разным. Часто это происходит вследствие консерватизма практических работников, их желания работать старыми, привычными методами. Иногда это объясняется недостатками самого метода. Например, случается, что кроме (и даже вместо) оптимальности метода желательны и существенны простота метода и возможность надежного контроля точности получаемых результатов. Может случиться, что сам класс рассмотренных задач не совпадает с классом, к которому относится большинство задач, поступающих для решения.

Конечно, нужно учитывать также вопрос о том, насколько широк возможный круг решаемых задач рассматриваемого класса. Если сейчас и в перспективе ожидается решение небольшого числа задач рассматриваемого класса, разработка оптимальных алгоритмов и создание стандартных программ решения задач этого класса могут себя не оправдать. В то же время изучение задачи оптимизации методов на различных классах часто полезно тем, что при решении возникают новые методы, которые затем находят применение и при решении задач из других классов.

## § 10. Постановка задачи оптимизации квадратур

Рассуждения § 9 показывают важность изучения различных постановок проблемы оптимизации методов на классах задач.

Рассмотрим задачу вычисления интеграла

$$I(f) = \int_{\Omega} f(P) \rho(P) dP.$$

Область интегрирования  $\Omega$  и весовая функция  $\rho(P)$  предполагаются фиксированными. Класс рассматриваемых задач определим заданием класса  $F$  подынтегральных функций. *Погрешностью квадратуры*

$$I(f) \approx S_N(f) = \sum_{j=1}^N D_j f(P_j)$$

на классе  $F$  называют величину

$$R_N(F) = \sup_{f \in F} |R_N(f)|,$$

где, как обычно,

$$R_N(f) = I(f) - S_N(f).$$

Нижняя грань

$$W_N(F) = \inf_{D_j, P_j} R_N(F)$$

называется *оптимальной оценкой погрешности квадратур на рассматриваемом классе*. Если существует квадратура, для которой

$R_N(F) = W_N(F)$ , то такую квадратуру называют *оптимальной* или *наилучшей на рассматриваемом классе*.

В § 2 была получена оценка (2.4) погрешности квадратуры, точной для многочленов степени  $m$ , через  $(m + 1)$ -ю производную функции. Эта оценка является неулучшаемой (см. задачу 2.3). Таким образом, для классов функций

$$F: |f^{(m+1)}(x)| \leq A \quad \text{при } x \in [a, b]$$

величина  $R_N(F)$  известна и задача построения оптимальной квадратуры сводится к нахождению коэффициентов и узлов, на которых достигается нижняя грань  $R_N(F)$ . Для ряда классов функций эту задачу удалось решить. Например, при  $m = 0$  такой квадратурой оказывается составная формула прямоугольников

$$\int_0^1 f(x) dx \approx \frac{1}{N} \sum_{j=1}^N f\left(\frac{j-1/2}{N}\right)$$

с оценкой погрешности

$$|R_N(f)| \leq A/(4N). \quad (1)$$

В настоящее время оптимальные квадратуры получены для небольшого набора классов функций, в основном одной переменной. Непосредственное значение этих квадратур для приложений невелико, однако это не значит, что не нужно заниматься построением таких квадратур.

Построение оптимальных квадратур и дальнейшее их развитие на случай большей гладкости и большего числа переменных оказались ценными не получением конкретных квадратурных формул, а выяснением качественной стороны вопроса: где какие методы лучше, на какую точность можно рассчитывать при использовании определенной информации о подынтегральной функции, какова плотность распределения узлов у «хороших» квадратур.

Пусть, например, при первоначальном анализе задачи мы решили воспользоваться информацией об ограниченности первой производной  $|f'(x)| \leq A$ ; оценка погрешности (7.8) нас не устраивает, поскольку для достижения нужной точности  $\epsilon$  требуется слишком большое число узлов: если  $A/(4N) \leq \epsilon$ , то  $N \geq A/(4\epsilon)$ . Оптимальность оценки (1) указывает на необходимость сужения класса рассматриваемых задач путем учета дополнительной информации о подынтегральной функции (ограниченность второй производной, тип особой точки подынтегральной функции, аналитичность и т. п.) или расширения множества используемых методов интегрирования.

## § 11. Оптимизация распределения узлов квадратурной формулы

Развитие методов численного интегрирования могло бы пойти по пути создания оптимальных методов на различных классах  $C_r(A; [a, b])$  и создания программ на основе этих методов. Здесь и далее  $C_r(A; [a, b])$  — класс функций с кусочно-непрерывной  $r$ -й производной, удовлетворяющей условию

$$|f^{(r)}(x)| \leq A \quad \text{при } x \in [a, b].$$

Вскоре после начала рассмотрения задач оптимизации методов стало ясно, что уже известные на этих классах методы с оценками погрешности (8.7) недалеко от оптимальных. Как увидим в § 7 гл. V, за счет оптимизации квадратурной формулы оценка погрешности (8.7) на рассматриваемых классах не может быть улучшена по порядку. При этом также стало ясно, что некоторые методы, практически совпадающие с методами Эйлера и Грегори (см. § 13), являются асимптотически оптимальными по оценке главного члена погрешности. Имеется в виду следующее. Для этих методов на соответствующих классах функций были получены оценки погрешности

$$c_r/N^r + c_{r+1}/N^{r+1},$$

в то время как для оптимальных на этих классах методов оценка погрешности имеет вид

$$c_r/N^r + O(1/N^{r+1}).$$

Казалось бы, что поскольку есть почти оптимальные методы, то следующим этапом должен стать перевод всех систем интегрирования на использование этих методов.

Однако подобное утверждение нельзя рассматривать как бесспорное. Вследствие большого многообразия задач, требующих решения, при переходе к этапу внедрения методов в практику всегда следует проявлять известную осторожность. Нельзя полностью ручаться, что принятое нами описание классов этих задач наилучшим образом соответствует классу реальных задач. Например, следует признать, что классы  $C_1$  плохо описывают реальные задачи. Конечно, из-за привычки к традиционным методам внедрение новых методов обычно требует энергичных действий со стороны их приверженцев. В то же время надо иметь в виду, что старые методы прошли испытание временем и могут оказаться более пригодными для решения задач некоторых классов, поэтому отбрасывание старых методов должно производиться лишь после достаточного теоретического и практического анализа.

В случае рассматриваемой проблемы оптимизации методов интегрирования на практике еще до полного выяснения вопроса об оптимальности методов пришли к следующим заключениям.



При разбиении исходного отрезка интегрирования на одинаковые элементарные отрезки  $a_q - a_{q-1} \equiv H$  мы получаем информацию о подынтегральной функции равномерно по всему отрезку интегрирования. В то же время подынтегральная функция может быть более гладкой на части отрезка интегрирования, поэтому там следует поместить относительно небольшое количество узлов, т. е. для практически оптимальных методов разбиение отрезка интегрирования на части должно быть приспособлено к специфике поведения подынтегральной функции.

Рассмотрим возможные постановки задачи распределения узлов в зависимости от особенностей поведения производной подынтегральной функции. Эти постановки имеют много общего, однако для конкретных задач тот или иной подход иногда оказывается более удобным. Для простоты изложения мы будем проводить рассмотрение на примере формулы трапеций.

Пусть вычисляется интеграл

$$I(f) = \int_0^1 f(x) dx$$

и подынтегральная функция удовлетворяет условиям  $|f''(x)| \leq A_l$  на отрезках  $[B_{l-1}, B_l]$ ,  $l = 1, \dots, q$ , где  $0 = B_0 < B_1 < \dots < B_q = 1$ . Для вычисления интегралов

$$I_l(f) = \int_{B_{l-1}}^{B_l} f(x) dx$$

применим составную формулу трапеций с равными отрезками разбиения длины  $H_l = b_l/N_l$ , где  $b_l = B_l - B_{l-1}$ :

$$I_l(f) \approx S_l(f) = H_l \left( \frac{f(B_{l-1})}{2} + f(B_{l-1} + H_l) + \dots + \frac{f(B_l)}{2} \right).$$

Согласно результатам § 3 остаточный член оценивается величиной  $A_l b_l^3 / (12N_l^2)$ . Тогда суммарная погрешность при замене  $I(f)$  суммой  $\sum_{l=1}^q S_l(f)$  не превзойдет величины

$$\Phi = \sum_{l=1}^q \frac{A_l b_l^3}{12N_l^2}.$$

Поставим задачу: при заданном числе  $N = N_1 + \dots + N_q$  отрезков разбиения распорядиться выбором величин  $N_l$  так, чтобы суммарная оценка погрешности  $\Phi$  была минимальной.

Найдем минимум  $\Phi$  при условии  $\Psi = N_1 + \dots + N_q - N = 0$ , не предполагая пока, что величины  $N_l$  целые. Приравнявая нулю производные функции Лагранжа  $\Phi + \lambda\Psi$ , получаем систему

уравнений

$$\frac{\partial(\Phi + \lambda\Psi)}{\partial N_l} = -\frac{A_l b_l^3}{6N_l^3} + \lambda = 0.$$

Отсюда

$$N_l = b_l \sqrt[3]{\frac{A_l}{6\lambda}}. \quad (1)$$

Из условия  $\Psi = 0$  получим следующее уравнение относительно  $\lambda$ :

$$\sum_{l=1}^q b_l \sqrt[3]{\frac{A_l}{6\lambda}} = N.$$

Из этого уравнения определяем  $\lambda$  и затем из (1) находим  $N_l$ . Поскольку  $N_l$  должны быть целыми, то возьмем, например,

$$N_l = N_l^0 = \left[ b_l \sqrt[3]{\frac{A_l}{6\lambda}} \right]. \quad (2)$$

Очевидно, что

$$N - q \leq N_1^0 + \dots + N_q^0 \leq N.$$

Мы не нашли настоящего минимума  $\Phi$  по множеству всевозможных целых  $N_1, \dots, N_q$ , удовлетворяющих условию  $N_1 + \dots + N_q = N$ , однако дальнейшее уточнение вряд ли разумно.

Обычно практический интерес представляет другая вариационная задача: найти минимальное значение  $N = N_1 + \dots + N_q$ , при котором оценка погрешности  $\Phi$  не превосходит заданного  $\varepsilon$ . Поскольку  $\Phi$  монотонно убывает с ростом величины  $N_l$ , достаточно ограничиться случаем  $\Phi = \varepsilon$ . Возьмем функцию Лагранжа в виде

$$N_1 + \dots + N_q + \lambda^{-1}(\Phi - \varepsilon).$$

Приравнивая нулю ее производные по  $N_l$ ,  $l = 1, 2, \dots, q$ , получим те же уравнения (1); подставляя значение  $N_l = b_l \sqrt[3]{\frac{A_l}{6\lambda}}$  в уравнение  $\Phi = \varepsilon$ , получим

$$\lambda^{2/3} \sum_{l=1}^q \left( \frac{A_l}{48} \right)^{1/3} b_l = \varepsilon.$$

Определяем  $\lambda$ , а затем соответствующие  $N_l$ . Соотношения для определения величин  $N_l$ , получившиеся при рассмотрении этих двух задач, имеют одинаковый вид. Поэтому для вывода качественных результатов об оптимальном распределении узлов достаточно было бы ограничиться рассмотрением одной из этих задач.

Нашей целью является разработка оптимальных методов решения и разработка на их основе систем программ решения ти-

повых математических задач. Можно представить себе программу вычисления интеграла с заданной точностью, работающую по следующей схеме. Производится вычисление таблицы значений функции на некоторой сетке  $x_0, \dots, x_n$ . По этой таблице составляется таблица разделенных разностей  $f(x_0; x_1; x_2), \dots, f(x_{n-2}; x_{n-1}; x_n)$ . Из рассмотрения этой таблицы делается вывод о наиболее целесообразном разбиении отрезка на части  $[B_{l-1}, B_l]$  и значениях  $A_l$ , соответствующих этим частям. Затем, в соответствии с (1), выбираются значения  $N_l$  и производится интегрирование.

Большинство алгоритмов реально работающих стандартных программ базируется не на таком непосредственном использовании полученных соотношений, а на одном качественном выводе, являющемся следствием (1). Для этого перепишем равенство (1) в виде

$$\frac{1}{12} A_l (b_l/N_l)^3 = \frac{\lambda}{2}.$$

Левая часть этого выражения равна оценке погрешности по элементарному отрезку интегрирования длины  $b_l/N_l$ , на которые был разбит отрезок  $[B_{l-1}, B_l]$ . Таким образом, это соотношение означает, что при оптимальном распределении узлов интегрирования оценки погрешностей, приходящихся на элементарные отрезки интегрирования, должны быть одинаковыми.

Для получения этого вывода достаточно было ограничиться случаем  $q=2$ . Это обстоятельство подтверждает общее свойство качественных характеристик методов решения задач (не обязательно математических): для их получения достаточно ограничиться рассмотрением простейших моделей, учитывающих основные стороны явления.

Как правило, алгоритмы, основанные на качественных выводах о свойствах решения оптимизационной задачи, имеют более широкую область применения, чем алгоритмы, подобные вышеописанному, основывающиеся на количественных соотношениях. Описываемые впоследствии программы вычисления интегралов, основывающиеся на этом качественном выводе, позволяют вычислять с высокой скоростью сходимости интегралы от функций с регулярными особенностями типа  $x^\alpha$ ,  $\alpha > -1$ .

Рассмотрим еще одну, близкую постановку задачи оптимизации распределения узлов интегрирования. Чтобы не утомлять читателя второстепенными деталями, мы не будем проводить подробных оценок членов высшего порядка в оценке погрешности.

Пусть отрезок интегрирования  $[0, 1]$  разбит на части  $[a_{q-1}, a_q]$ ,  $q=1, \dots, N$ ,  $a_0=0$ ,  $a_N=1$ , и интеграл по каждой части вычисляется по формуле трапеций

$$I_q(f) = \int_{a_{q-1}}^{a_q} f(x) dx \approx s_q(f) = \frac{a_q - a_{q-1}}{2} (f(a_{q-1}) + f(a_q)).$$

Тогда интеграл по всему отрезку  $[0, 1]$  вычисляется по формуле

$$I(f) \approx \sum_{q=1}^N s_q(f)$$

с оценкой остаточного члена

$$r = \sum_{q=1}^N \left( \max_{[a_{q-1}, a_q]} |f''(x)| \right) \frac{(a_q - a_{q-1})^3}{12}. \quad (3)$$

Пусть известно, что  $|f''(x)| \leq F(x)$  на  $[0, 1]$ , где  $F(x)$  непрерывна, и пусть в качестве  $a_q$  взяты значения  $\varphi(q/N)$  непрерывно дифференцируемой функции  $\varphi$ , удовлетворяющей условиям  $\varphi(0) = 0$ ,  $\varphi(1) = 1$ . Поскольку

$$a_q - a_{q-1} = \varphi\left(\frac{q}{N}\right) - \varphi\left(\frac{q-1}{N}\right) = \varphi'\left(\frac{q}{N}\right) \frac{1}{N} + o\left(\frac{1}{N}\right)$$

при  $N \rightarrow \infty$ ,

то

$$\max_{[a_{q-1}, a_q]} |f''(x)| \leq \max_{[a_{q-1}, a_q]} F(x) = F(a_q) + o(1) = F\left(\varphi\left(\frac{q}{N}\right)\right) + o(1).$$

Из этих соотношений получаем

$$\begin{aligned} \left( \max_{[a_{q-1}, a_q]} |f''(x)| \right) \frac{(a_q - a_{q-1})^3}{12} &\leq \varepsilon_q = \\ &= \left( \varphi'\left(\frac{q}{N}\right) \right)^3 \frac{F\left(\varphi\left(\frac{q}{N}\right)\right)}{12N^3} + o\left(\frac{1}{N^3}\right). \end{aligned}$$

Подставляя последние соотношения в (3), имеем

$$r \leq \bar{r} = \frac{1}{N^2} \left\{ \sum_{q=1}^N \frac{1}{N} \left( \varphi'\left(\frac{q}{N}\right) \right)^3 \frac{F\left(\varphi\left(\frac{q}{N}\right)\right)}{12} \right\} + o\left(\frac{1}{N^2}\right).$$

Выражение в фигурных скобках является квадратурной суммой Римана для интеграла

$$\int_0^1 (\varphi'(t))^3 \frac{F(\varphi(t))}{12} dt$$

от непрерывной функции. Следовательно,

$$\bar{r} = \frac{1}{N^2} \left( \int_0^1 (\varphi'(t))^3 \frac{F(\varphi(t))}{12} dt \right) + o\left(\frac{1}{N^2}\right). \quad (4)$$

Рассмотрим задачу минимизации первого, главного члена выражения (4). Для удобства решения уравнения Эйлера при-

мем за независимую переменную функцию  $\varphi$ . Тогда коэффициент при  $1/N^2$  в главном члене погрешности запишется в виде

$$\int_0^1 (t'(\varphi))^{-2} \frac{F(\varphi)}{12} d\varphi.$$

Уравнение Эйлера для функции, минимизирующей функционал

$$\int_0^1 G(\varphi, t, t') d\varphi,$$

имеет вид

$$\frac{d}{d\varphi} \left( \frac{\partial G}{\partial t'} \right) - \frac{\partial G}{\partial t} = 0. \quad (5)$$

В рассматриваемом случае  $\partial G/\partial t = 0$ , поэтому из (5) следует, что  $\partial G/\partial t' = \text{const}$ . Подставляя конкретное значение функции  $G$ , получим

$$(t'(\varphi))^{-3} \frac{F(\varphi)}{6} = \text{const}$$

или

$$F(\varphi)(\varphi'(t))^3 = C_1. \quad (6)$$

Общее решение этого уравнения зависит от  $C_1$  и еще от некоторой постоянной  $C_2$ . Значения этих постоянных можно определить из граничных условий  $\varphi(0) = 0$ ,  $\varphi(1) = 1$ . Решение рассмотренной вариационной постановки может практически использоваться различными способами. Например, в случае гладких функций  $f(x)$  программы осуществляют численное интегрирование (6) на сетке с шагом, существенно большим  $1/N$ , и затем распределяют узлы в соответствии с полученным решением.

Из соотношения (6) можно сделать тот же вывод о равенстве оценок погрешностей на элементарных отрезках интегрирования при оптимальном распределении узлов. В самом деле, умножим (6) на  $\frac{1}{12N^3}$ , положим  $t = \frac{q}{N}$  и заменим  $F\left(\varphi\left(\frac{q}{N}\right)\right)$  и  $\frac{1}{N} \varphi'\left(\frac{q}{N}\right)$  соответственно эквивалентными величинами  $\max_{[a_{q-1}, a_q]} F(x)$ ,  $a_q - a_{q-1}$ . В результате получим

$$\left( \max_{[a_{q-1}, a_q]} F(x) \right) \frac{(a_q - a_{q-1})^3}{12} \approx \frac{C_1}{12N^3}. \quad (7)$$

Другой из возможных путей практического использования решения уравнения (6) состоит в следующем. Пусть требуется вычислить большую серию интегралов с одинаковым характерным поведением подынтегральных функций. Выделим простейшую модельную функцию, для которой задача оптимизации узлов может быть решена в явном виде, и далее будем произво-

диль интегрирование с распределением узлов, соответствующим этой функции. Если характер изменения функций из рассматриваемой серии зависит от некоторого параметра, то этот параметр следует учесть при выборе модельной функции; естественно, что модельная функция не обязательно относится к рассматриваемому классу. Чем большее количество задач предъявляется для решения, тем более оправданными могут быть затраты, связанные с удачным выбором и рассмотрением модельной задачи.

## § 12. Примеры оптимизации распределения узлов

Рассмотрим примеры решения уравнения (11.6) для конкретных задач.

Пример 1. Пусть вычисляется серия интегралов

$$I(b) = \int_0^1 f(b, x) dx,$$

где  $b$  — параметр серии,  $f(b, x) = x^b g(b, x)$ ,  $-1 < b < 2$ ,  $g(b, x)$  — гладкая функция,  $g(b, 0) \neq 0$ . Если  $b \neq 0, 1$ , то вторая производная  $f_{xx}(b, x)$  не ограничена в окрестности точки 0, поэтому при выборе модельной задачи следует учесть эту специфику поведения подынтегральной функции. В окрестности точки  $x = 0$  имеем

$$f_{xx}(b, x) = b(b-1)x^{b-2}g(b, 0) + O(x^{b-1}).$$

Таким образом, в окрестности точки  $x = 0$  вторая производная  $f_{xx}$  приблизительно пропорциональна второй производной функции  $y = x^b$ , поэтому функцию  $y = x^b$  естественно рассматривать в качестве модельной. Примем за  $F(x)$  величину  $|b(b-1)|x^{b-2}$ ; тогда уравнение (11.6) запишется в виде

$$|b(b-1)|\varphi^{b-2} \left( \frac{d\varphi}{dt} \right)^3 = C_1,$$

отсюда

$$\left( \frac{3\sqrt[3]{|b(b-1)|}}{b+1} \right) \varphi^{\frac{b+1}{3}} = C_1 t + C_2.$$

Из условия  $\varphi(0) = 0$  получаем, что  $C_2 = 0$ , а из условия  $\varphi(1) = 1$  — что  $\varphi^{\frac{b+1}{3}} = t$ . Таким образом,

$$\varphi(t) = t^{\frac{3}{1+b}} \quad (1)$$

и для модельной задачи вычисления интеграла  $\int_0^1 x^b dx$  оптимальным в рассматриваемом нами смысле является распределе-

ние узлов  $a_q = \left( \frac{q}{N} \right)^{\frac{3}{1+b}}$

Проведенные выше построения, вообще говоря, неприменимы к рассматриваемому случаю, поскольку при получении оценки (11.4) предполагалась ограниченность второй производной функции  $F(x)$ , не имеющая места для данной задачи. Однако можно обосновать применимость оценки (11.4), и в рассматриваемом случае.

**Задача 1.** Пусть для функции

$$f_b(x) = \begin{cases} 0 & \text{при } x=0, \\ x^b & \text{при } x \in (0, 1], \end{cases}$$

где  $-1 < b < 1$ , по формуле трапеций с постоянным шагом  $a_q - a_{q-1} = 1/N$  вычисляется

$$\int_0^1 f_b(x) dx. \quad (2)$$

Доказать, что суммарная погрешность удовлетворяет соотношению  $R_N(f) \sim D_1(b)/N^{1+b}$ , где  $D_1(b) \neq 0$ .

**Задача 2.** Интеграл (2) вычисляется по формуле трапеций с распределением узлов  $a_q = \varphi(q/N)$ ,  $\varphi(t) = t^{\frac{3}{1+b}}$ , определяемым (1). Доказать, что суммарная погрешность удовлетворяет соотношению  $R_N(f) \sim D_2(b)/N^2$ .

**Задача 3.** Интеграл (2) вычисляется по формуле трапеций с распределением узлов  $a_q = \varphi(q/N)$ ,  $\varphi(t) = t^a$ . Показать, что при  $a > 2/(b+1)$  суммарная погрешность  $R_N(f) \sim D(a, b)/N^2$ . Проверить, что  $D(a, b) > D_2(b)$ .

Сравнение результатов решения этих задач показывает, что перераспределение узлов в сторону большей их концентрации вблизи особенности, в частности оптимизация распределения узлов, приводит к увеличению порядка скорости сходимости.

**Пример 2.** Вычисляется серия интегралов

$$\int_0^1 x^b (\ln x) g(b, x) dx,$$

где  $g(b, x)$  — гладкая функция,  $g(b, 0) \neq 0$ ,  $b$  — параметр серии. Поскольку  $\ln x$  имеет особенность в точке 0, то кажется естественным взять в качестве модельной функции  $y = x^b \ln x$ . Ее вторая производная имеет вид

$$y'' = x^{b-2} (b(b-1) \ln x + (2b-1)).$$

При  $F(x) = |(x^b \ln x)''|$  уравнение (10.6) не решается в квадратурах, поэтому упростим задачу. При  $x \rightarrow 0$  функция  $\ln x$  растет медленнее, чем любая степенная функция  $y = x^{-\varepsilon}$ ,  $\varepsilon > 0$ . Исходя из этого в уравнении (11.6) возьмем  $F(x) = \text{const} \cdot x^{b-2}$ .

**Пример 3.** Вычисляется серия интегралов

$$\int_0^1 \exp\left(-\frac{x}{b}\right) g(b, x) dx, \quad (3)$$

где  $g(b, x)$  — гладкая функция,  $g(b, 0) \neq 0$ ,  $b$  — параметр серии, который может принимать очень малые значения. При малых  $b$  подынтегральная функция и ее производные резко меняются в окрестности точки  $x=0$  за счет множителя  $\exp\{-x/b\}$ , поэтому имеет смысл произвести оптимизацию рас-

пределаия узлов интегрирования на модельной задаче вычисления интеграла  $\int_0^1 \exp\{-x/b\} dx$ . Положим  $F(x) = |(\exp\{-x/b\})''|$ . Уравнение (11.6) приобретает вид

$$\frac{1}{b^2} \exp\left\{-\frac{\varphi}{b}\right\} \left(\frac{d\varphi}{dt}\right)^3 = C_1.$$

Отсюда

$$1 - \exp\left\{-\frac{\varphi}{3b}\right\} = C_3 t + C_4.$$

Из условия  $\varphi(0) = 0$  следует, что  $C_4 = 0$ , а из условия  $\varphi(1) = 1$  получаем

$$1 - \exp\left\{-\frac{\varphi}{3b}\right\} = \left(1 - \exp\left\{-\frac{1}{3b}\right\}\right) t,$$

откуда

$$\varphi(t) = 3b \ln \left[ 1 - \left( 1 - \exp\left\{-\frac{1}{3b}\right\} \right) t \right]^{-1}.$$

Пример 4. Вычисляется серия интегралов

$$\int_0^1 \exp\{-x^2/b^2\} g(b, x) dx,$$

где  $g(b, x)$  — гладкая функция,  $g(b, 0) \neq 0$ ,  $b$  — параметр серии, который может принимать малые значения. При малых  $b$  подынтегральная функция и ее производные резко меняются в окрестности точки  $x = 0$  за счет множителя  $\exp\{-x^2/b^2\}$ . Поэтому в качестве модельной возьмем задачу вычисления интеграла

$\int_0^1 \exp\{-x^2/b^2\} dx$ . Положим  $F(x) = |(\exp\{-x^2/b^2\})''|$ , тогда в качестве уравнения (11.6) получим уравнение

$$|1 - 2\varphi^2/b^2| \exp\{-\varphi^2/b^2\} (d\varphi/dt)^3 = C_1 b^2/2,$$

откуда

$$\int \sqrt[3]{|1 - 2\varphi^2/b^2|} \exp\{-\varphi^2/3b^2\} d\varphi = \int \sqrt[3]{C_1 b^2 / \sqrt{2}} dt.$$

Этот интеграл не вычисляется в явном виде, поэтому попытаемся произвести упрощения. Например, можно заменить  $\sqrt[3]{|1 - 2\varphi^2/b^2|}$  на 1. При больших значениях  $\varphi/b$ , когда погрешность такой замены большая, ее влияние не столь значительно из-за малого множителя  $\exp\{-\varphi^2/(3b^2)\}$ . После такого упрощения функция  $\varphi(t)$  будет выражаться через функцию, обратную функции

$$\int_0^{\varphi} \exp(-v^2) dv.$$

Задачи, подобные рассмотренным в примерах 3, 4, возникают довольно часто. Например, при расчетах диаграмм направленности антенн вычисляются

серии интегралов  $\int_0^1 \exp\{ibg(b, x)\} h(b, x) dx$  в широком диапазоне изменения  $b$ ; функции  $g(b, x)$ ,  $h(b, x)$  являются довольно гладкими. При  $b$  не очень



больших эти интегралы могут вычисляться с помощью простейших квадратурных формул. С ростом  $b$  производные подынтегральной функции растут, поэтому требуемое количество узлов интегрирования увеличивается. При очень больших  $b$  можно воспользоваться методом перевала или иными асимптотическими методами. Однако для «промежуточных» значений  $b$  оба эти метода будут плохи: первый — из-за трудоемкости, второй — из-за малой точности. Поэтому иногда применяют следующий метод: контур интегрирования преобразуется так, чтобы он проходил по линиям наискорейшего спуска функции  $\exp\{ibg(b, x)\}$ , как это делается при использовании метода перевала. Получаются интегралы от резко меняющихся функций, аналогичные рассмотренным в примерах 3, 4.

Из приведенных примеров видно, что оптимизация распределения узлов интегрирования на основании уравнения (11.6) требует достаточно высокой квалификации исследователя. Поэтому далее в § 17 будет рассмотрен вопрос о передаче этих функций ЭВМ.

### § 13. Главный член погрешности

Применение формул для оценок погрешности, подобных полученным в § 2, 3, требует достаточно высокой квалификации исследователя, например для получения требуемых оценок производных. При получении ряда из этих оценок, например оценок для составных формул трапеции и Симпсона, возможно существенное загроубление оценки, поскольку общая оценка погрешности равна сумме модулей оценок на отдельных отрезках.

Эти обстоятельства определили интерес к получению выражения для главного члена погрешности. По информации о величине главного члена погрешности можно полноценнее проводить сравнение методов.

Как будет видно далее, сам факт наличия главного члена у погрешности позволяет судить о реальной величине погрешности, не прибегая к теоретическим оценкам. Обратимся к составной формуле трапеций вычисления интеграла

$I(f) = \int_A^B f(x) dx$  с постоянным шагом  $H$ . Для удобства обозначим  $H = (B - A)/M$ ,  $a_q = A + qH$ , в частности  $a_0 = A$ ,  $a_M = B$ . Имеем

$$I(f) \approx S_M(f) = H \left( \frac{f(a_0)}{2} + f(a_1) + \dots + f(a_{M-1}) + \frac{f(a_M)}{2} \right).$$

Согласно § 3.2 справедливо равенство

$$\int_{a_{q-1}}^{a_q} f(x) dx = H \frac{f(a_{q-1}) + f(a_q)}{2} - \frac{f''(\xi_q) H^3}{12}, \quad \xi_q \in [a_{q-1}, a_q].$$

Просуммировав по  $q$ , получим

$$\int_{a_0}^{a_M} f(x) dx = S_M(f) + R(f);$$

$$R(f) = - \sum_{q=1}^M f''(\xi_q) \frac{H^3}{12}.$$

Величину погрешности  $R(f)$  можно записать в виде

$$R(f) = - \frac{H^2}{12} i(f), \quad i(f) = \sum_{q=1}^M H f''(\xi_q).$$

Выражение в правой части есть квадратурная сумма для интеграла  $\int_{a_0}^{a_M} f''(x) dx$ , поэтому при  $H \rightarrow 0$  имеем

$$i(f) \rightarrow \int_{a_0}^{a_M} f''(x) dx.$$

Следовательно,

$$R(f) = - \frac{H^2}{12} \int_{a_0}^{a_M} f''(x) dx + R_1(f), \quad R_1(f) = o(H^2).$$

**Задача 1.** Пусть  $|f^{(3)}(x)| \leq M_3$  на отрезке  $[A, B]$ . Показать, что  $|R_1(f)| \leq c_3 M_3 (B-A) H^3$ .

**Задача 2.** Пусть  $|f^{(4)}(x)| \leq M_4$  на отрезке  $[A, B]$ . Показать, что  $|R_1(f)| \leq c_4 M_4 (B-A) H^4$ .

Полученное соотношение для  $R(f)$  может использоваться в различных целях. Например, его можно представить в виде

$$R(f) = - \frac{H^2}{12} (f'(B) - f'(A)) + o(H^2). \quad (1)$$

После вычисления  $f'(B) - f'(A)$  получаем значение главного члена погрешности. Предположим, что достигнутая точность не является удовлетворительной. Запишем (1) в виде

$$I(f) = S_M^4(f) + o(H^2),$$

где

$$S_M^4(f) = S_M^2(f) - \frac{H^2}{12} (f'(B) - f'(A)).$$

Как следует из решения задачи 2, при  $|f^{(4)}(x)| \leq M_4$  выражение  $S_M^4(f)$  оказывается квадратурной суммой с погрешностью  $O((B-A)H^4)$ , т. е. такой же по порядку, как у формулы Симпсона.

Можно попытаться выделить главный член погрешности по-  
лучившейся формулы. Имеем равенства

$$S_M^4(f) = \sum_{q=1}^M s_q^2(f),$$

$$s_q^2(f) = \frac{H}{2} (f(a_{q-1}) + f(a_q)) - \frac{H^2}{12} (f'(a_q) - f'(a_{q-1})).$$

Величину  $s_q^2(f)$  будем рассматривать как приближенное значе-  
ние интеграла

$$I_q(f) = \int_{a_{q-1}}^{a_q} f(x) dx.$$

Подставляя в разность  $I_q(f) - s_q^2(f)$  представление  $f(x)$  в виде  
отрезка ряда Тейлора, можно получить главный член погреш-  
ности на элементарном отрезке в виде  $\frac{H^5}{720} f^{(4)}(\xi_q^1)$  и т. д.

Продолжая процесс выделения главного члена погрешности,  
приходим к последовательности *квадратурных формул Эйлера*

$$I(f) \approx S_M^{2l}(f), \quad S_M^{2l}(f) = S_M^2(f) - \sum_{j=1}^{l-1} \gamma_{2j} H^{2j} (f^{(2j-1)}(B) - f^{(2j-1)}(A))$$

с оценкой погрешности

$$I(f) - S_M^{2l}(f) = -\gamma_{2l} f^{(2l)}(\xi^{2l})(B - A) H^{2l}. \quad (2)$$

Существует следующее соотношение, которому удовлетворяют числа  $\gamma_j$ :

$$\frac{x}{e^x - 1} = \sum_{j=0}^{\infty} \gamma_j x^j.$$

Обычно принято записывать числа  $\gamma_j$  в виде  $B_j/j!$ , где  $B_j$  — так называемые  
числа Бернулли.

Для сведения приведем несколько значений чисел  $\gamma_j$ :

$$\gamma_2 = \frac{1}{12}, \quad \gamma_4 = -\frac{1}{720}, \quad \gamma_6 = \frac{1}{30\,240}, \quad \gamma_8 = -\frac{1}{1\,209\,600},$$

$$\gamma_{10} = \frac{1}{47\,900\,160}.$$

Использование формул Эйлера неудобно, поскольку необхо-  
димо вычислять не только значения функции, но и значения ее  
производных.

Если же в выражении  $S_M^{2p}(f)$  заменить производные  $f^{(2k-1)}(A)$   
и  $f^{(2k-1)}(B)$  производными интерполяционных многочленов сте-  
пени  $l$  соответственно с узлами  $a_0, \dots, a_l$  и  $a_N, \dots, a_{N-l}$ , то при  
 $l = 2p - 3$  и  $l = 2p - 2$  после проведения промежуточных пре-  
образований получаются *формулы численного интегрирования*

$$I(f) \approx G_M^l(f),$$

$$G_M^l(f) = S_M^2(f) - H \sum_{k=1}^l \beta_k (\nabla^k f(a_M) - (-1)^k \Delta^k f(a_0)),$$

где

$$\beta_k = (-1)^{k+1} \int_0^1 \frac{x(x-1)\dots(x-k)}{(k+1)!} dx.$$

В частности,

$$\beta_1 = \frac{1}{12}, \quad \beta_2 = \frac{1}{24}, \quad \beta_3 = \frac{19}{720}, \quad \beta_4 = \frac{3}{160},$$

$$\beta_5 = \frac{863}{60480}, \quad \beta_6 = \frac{275}{24192}.$$

В случае подынтегральных функций с нерегулярным характером поведения, типа рассмотренных в § 11, применение формул Эйлера и Грегори неэффективно, поскольку производные высших порядков или не ограничены, или очень велики. Поэтому при непосредственном вычислении определенных интегралов эти формулы в настоящее время применяются относительно редко. Однако они используются при интегрировании функций, заданных таблично, при вычислении неопределенных интегралов, при решении интегральных уравнений Вольтерра и в других задачах, где существенно, чтобы значения подынтегральной функции вычислялись именно на равномерной сетке.

**Задача 3.** Доказать, что главный член погрешности квадратуры Грегори  $I(f) \approx G_M^l(f)$  есть  $\beta_{l+1} H^{l+2} (f^{(l+1)}(B) - (-1)^{l+1} f^{(l+1)}(A))$ .

**Задача 4.** Пусть  $\int_0^1 f(x) dx$  вычисляется по составной формуле трапеций с переменным шагом интегрирования:  $a_q = \varphi(q/N)$ , где  $\varphi$  — гладкая функция. Доказать, что главный член погрешности есть

$$-\frac{1}{12N^2} \int_0^1 f''(\varphi(t)) (\varphi'(t))^3 dt.$$

Указание. См. построения § 11.

#### § 14. Правило Рунге практической оценки погрешности

Мы получили, что главный член погрешности формулы трапеций с постоянным шагом интегрирования равен

$$-\frac{1}{12} H^2 (f'(B) - f'(A)).$$

В случае формул более высокого порядка точности можно получить представление главного члена погрешности квадратуры

через производные высших порядков. Непосредственное использование этих выражений для оценки величины главного члена погрешности иногда неудобно, поскольку требует выполнения операции дифференцирования. В других задачах выражение главного члена погрешности может оказаться настолько сложным, что его вычисление требует дополнительного численного интегрирования. Поэтому в вычислительной практике применяется способ практической оценки погрешности, не использующий фактического выражения главного члена погрешности, а опирающийся лишь на факт существования такого главного члена. Для простейших задач типа численного интегрирования этот способ связывается с именем Рунге, в более сложных случаях — с именами Ричардсона и Филиппова. Этот способ основан на выделении главного члена погрешности по результатам расчетов с двумя различными шагами.

Рассмотрим простейший вариант применения этого правила.

Осуществим приближенное вычисление интеграла  $I(f) = \int_A^B f(x) dx$  с помощью формулы трапеций с постоянным шагом  $H_1 = (B - A)/M_1$  и  $H_2 = (B - A)/M_2$ ;  $M_2 = 2M_1$ , т. е.  $H_2 = H_1/2$ . Согласно (12.1) имеем равенство

$$I(f) - S_{M_1}(f) = -\frac{H_1^2}{12}(f'(B) - f'(A)) + o(H_1^2),$$

$$I(f) - S_{M_2}(f) = -\frac{H_2^2}{12}(f'(B) - f'(A)) + o(H_2^2).$$
(1)

Мы стремимся построить алгоритм вычисления главного члена погрешности, не использующий его конкретного выражения. Для этого запишем (1) в виде совокупности приближенных равенств

$$I(f) - S_{M_1}(f) \approx CH_1^2,$$

$$I(f) - S_{M_2}(f) \approx CH_2^2.$$
(2)

Величины  $S_{M_1}(f)$  и  $S_{M_2}(f)$  определены в результате расчетов, поэтому мы имеем два приближенных равенства относительно двух неизвестных  $I(f)$  и  $C$ . Вычитая второе равенство из первого, получим

$$S_{M_2}(f) - S_{M_1}(f) \approx CH_1^2 - CH_2^2 = 3CH_2^2.$$

Таким образом,

$$CH_2^2 = \frac{1}{3}(S_{M_2}(f) - S_{M_1}(f)).$$
(3)

Подставляя приближенное выражение  $CH_2^2$  в (2), получаем приближенное равенство

$$I(f) - S_{M_2}(f) \approx \frac{1}{3}(S_{M_2}(f) - S_{M_1}(f)).$$
(4)

Таким образом, величина  $\frac{1}{3}(S_{M_2}(f) - S_{M_1}(f))$  является главным членом погрешности приближенного значения интеграла  $S_{M_2}(f)$ . Перенос в (4) значение  $S_{M_2}(f)$  в правую часть, получим формулу для более точного по порядку, чем  $S_{M_2}(f)$ , приближения к  $I(f)$ :

$$I(f) \approx S_{M_2}(f) + \frac{1}{3}(S_{M_2}(f) - S_{M_1}(f)). \quad (5)$$

Таким образом, описанный способ построения главного члена погрешности порождает некоторую квадратурную формулу более высокого порядка точности.

**Задача 1.** Доказать, что правая часть в (5) совпадает с составной квадратурной формулой Симпсона.

Информация о величине главного члена погрешности часто используется для аккуратного определения минимального количества узлов, достаточного для достижения заданной точности. Из (3) находим, что

$$C \approx \frac{1}{3H_2^2}(S_{M_2}(f) - S_{M_1}(f)),$$

а затем выбираем шаг интегрирования из условия

$$|CH^2| \leq \varepsilon \quad \text{или} \quad |CH^2| \leq \varepsilon |S_{M_2}(f)|, \quad (6)$$

где  $\varepsilon$  — заданная абсолютная или относительная погрешность результата.

Выписанные выше соотношения (2) — (5) носят асимптотический характер, поэтому значение  $C$ , найденное из (3), будет достоверным (т. е. близким к истинному) лишь при достаточно малом  $H_2$ . В ответственных случаях после решения задачи с шагом  $H$ , удовлетворяющим условиям (6), для контроля над точностью решают задачу с шагом  $2H$  и опять определяют главный член погрешности, соответствующий шагу  $H$ .

Описанный метод уточнения результата по итогам двух расчетов применим к методам любого порядка точности, причем не обязательно брать  $M_2 = 2M_1$ .

**Задача 2.** Имеется некоторый метод решения задачи с погрешностью

$$I(f) - S_M(f) \sim C/M^m.$$

Произведено вычисление интеграла с  $M_1$  и  $M_2 = \lambda M_1$  отрезками разбиения. Показать, что

$$I(f) - S_{M_2}(f) \sim \frac{S_{M_2}(f) - S_{M_1}(f)}{\lambda^m - 1};$$

здесь имеется в виду предельный переход при  $M_2 \rightarrow \infty$ ,  $\lambda = \text{const}$ .

**Задача 3.** Пусть

$$I(f) - S_M(f) = \frac{C}{M^m} + o\left(\frac{1}{M^{m+1}}\right).$$

Показать, что

$$I(f) - S_{M_1}(f) \sim \frac{S_{M_2}(f) - S_{M_1}(f)}{(M_2/M_1)^m - 1}$$

при условии, что  $M_1, M_2 - M_1 \rightarrow \infty$ .

Задача 4. Пусть

$$I(f) - S_M(f) = \frac{C}{M^m} + o\left(\frac{1}{M^{m+2}}\right).$$

Показать, что

$$I(f) - S_{M_2}(f) \sim \frac{S_{M_2}(f) - S_{M_1}(f)}{(M_2/M_1)^m - 1}$$

при условии, что  $M_1 \rightarrow \infty, M_2 > M_1$ .

В случаях, когда вычисляется большое количество интегралов с особенностями определенного вида, без серьезного теоретического анализа нельзя определить порядок сходимости метода на интегралах этого рода (из-за неограниченности производных мы не имеем права пользоваться результатом о существовании главного члена погрешности). В других же случаях порядок погрешности может быть известен, но неясно, каким он оказывается при реально используемых значениях  $M$ .

Рассмотрим вопрос о способах проверки выполнимости соотношения  $R(f) \sim CM^{-m}$  при реально допустимых значениях  $M$ .

Можно постараться подобрать модельную задачу с известным ответом, близкую к рассматриваемой. Тогда после проведения расчета мы будем иметь в распоряжении приближенное значение  $S_M$  и погрешность  $R_M = I - S_M$ . Может случиться, что имеются какие-то предположения о характере поведения этой величины, например, что

$$R_M \sim \text{const} \cdot M^{-m}. \quad (7)$$

В таком случае можно подсчитать для некоторой последовательности  $M_k$  значения  $M_k^m R_{M_k}$  и посмотреть, стабилизируются ли эти величины с ростом  $M$ . Если нет предположений о характере поведения погрешности в данной задаче, то можно применить следующую методику.

Возьмем координатную плоскость  $\ln M, \ln\left(\frac{1}{|R|}\right)$  (рис. 3.14.1), нанесем на нее точки

$$\left(\ln M_k, \ln \frac{1}{|R_{M_k}|}\right). \quad (8)$$

Если эти точки расположены хаотически, то это означает, что числа  $M_k$  не настолько велики, чтобы в погрешности выделился главный член. Предположим, что асимптотическое равенство (7) в данной области изменения параметра  $M$  выполняется с большой точностью. Из (7) следует, что

$$\ln \left| \frac{1}{R_M} \right| \sim m \ln M;$$

после дифференцирования имеем

$$\frac{d \ln \left| \frac{1}{R_M} \right|}{d \ln M} \sim m. \quad (9)$$

Заметим, что операция дифференцирования асимптотических равенств, вообще говоря, незаконна.

Согласно (9) в случае, когда (7) выполняется достаточно точно, точки  $(\ln M_k, \ln(1/|R_{M_k}|))$ , получаемые в результате эксперимента на ЭВМ, должны лежать на кривой, тангенс угла наклона которой стремится к  $m$ . Близость угла наклона этой кривой к постоянному характеризует точность выполнения асимптотического равенства  $R_M \sim \text{const} \cdot M^{-m}$ . Если угол наклона кривой меняется резко, то еще нет оснований применять правило Рунге.

Проверку справедливости предположения о характере поведения погрешности можно осуществлять и таким путем. Если справедливо равенство

$$R_M \sim c/M^m, \quad (10)$$

то

$$S_{M\lambda} - S_M \sim c(1 - \lambda^{-m})/M^m. \quad (11)$$

С другой стороны, если (11)

выполняется при  $M \geq M_0$ , то будет выполняться и (10). Поэтому вместо проверки практической выполнимости (10) можно производить проверку практической выполнимости (11), в частности, при помощи изучения графиков функций  $g(M) = (S_{M\lambda} - S_M)M^m$  или расположения точек

$$\left( \ln M_k, \ln \left| \frac{1}{S_{M_k\lambda} - S_{M_k}} \right| \right). \quad (12)$$

Заметим, что возможности определения значения  $m$  и вообще проверки условия (10) путем численного эксперимента довольно ограничены. Например, случаи  $R_M \sim \text{const} \cdot \frac{\ln M}{M}$  и  $R_M \sim \text{const} \cdot \frac{1}{M}$  практически неразличимы при таком рассмотрении, потому что в обоих случаях  $d \ln \left| \frac{1}{R_M} \right| / d \ln M \rightarrow 1$  при  $M \rightarrow \infty$ .

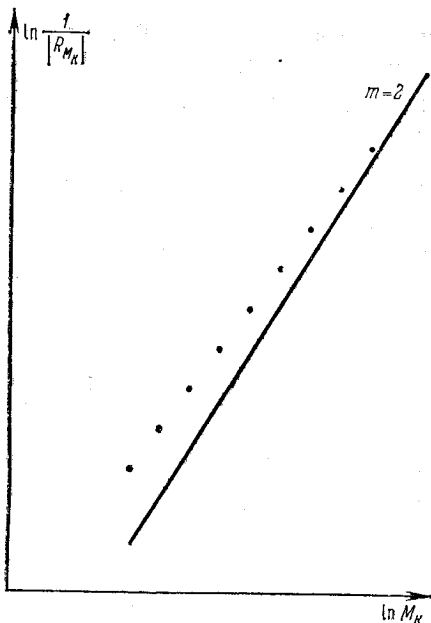


Рис. 3.14.1



### § 15. Уточнение результата интерполяцией более высокого порядка точности

Если подынтегральная функция достаточно гладкая, то, как правило, погрешность квадратурной формулы может быть представлена в виде

$$I(f) - S_M(f) = R_M(f) = \sum_{k=1}^l D_k(f) M^{-i_k} + r(M), \quad (1)$$

где  $i_1 < \dots < i_l$ ,  $r(M) = o(M^{-i_l})$ . Обычно при гладкой подынтегральной функции имеем  $i_2 - i_1 = \dots = i_l - i_{l-1} = s$ , где  $s = 1$  или  $s = 2$ . Например, в предположении ограниченности  $f^{(2m+2)}(x)$  погрешность формулы трапеций, согласно (12.2), представляется в виде

$$R_M(f) = - \sum_{k=1}^m \gamma_{2k} \left( \frac{B-A}{M} \right)^{2k} (f^{(2k-1)}(B) - f^{(2k-1)}(A)) + \\ + o\left(\left(\frac{B-A}{M}\right)^{2m+2}\right).$$

Предположим, что произведено вычисление  $S_M(f)$  при значениях  $M = M_0, \dots, M_l$ . Мы имеем равенства

$$I(f) = S_{M_j}(f) + \sum_{k=1}^l D_k(f) M_j^{-i_k} + r(M_j), \quad j = 0, \dots, l.$$

Образует линейную комбинацию этих соотношений с некоторыми коэффициентами  $c_j$ , потребовав, чтобы

$$\sum_{j=0}^l c_j = 1. \quad (2)$$

Получим соотношение

$$I(f) = \sum_{j=0}^l c_j S_{M_j}(f) + \sum_{k=1}^l D_k(f) \left( \sum_{j=0}^l c_j M_j^{-i_k} \right) + \sum_{j=0}^l c_j r(M_j).$$

Предположим, что выполняются равенства

$$\sum_{j=0}^l c_j M_j^{-i_k} = 0 \quad \text{при } k = 1, \dots, l, \quad (3)$$

тогда

$$I(f) = \sum_{j=0}^l c_j S_{M_j}(f) + \sum_{j=0}^l c_j r(M_j). \quad (4)$$

Если величиной  $\sum_{j=0}^l c_j r(M_j)$  можно пренебречь, то

$$I(f) \approx \sum_{j=0}^l c_j S_{M_j}(f). \quad (5)$$

Система соотношений (2), (3) образует систему из  $l + 1$  линейных алгебраических уравнений с  $l + 1$  неизвестными, поэтому есть основания ожидать, что она имеет решение.

Аналогия между рассматриваемой задачей и задачей интерполяции позволяет найти  $c_j$  в явном виде. Перепишем (1) в виде

$$Q_l(M^{-1}) = S_M(f) + r(M), \quad \text{где} \quad Q_l(y) = I(f) - \sum_{k=1}^l D_k y^k. \quad (6)$$

Из соотношения (6) видно, что задача нахождения  $I(f)$  может формулироваться следующим образом. Заданы значения многочлена  $Q_l(y)$  при  $y = M_0^{-1}, \dots, M_l^{-1}$ ; требуется определить значение  $Q_l(0) = I(f)$ . Согласно интерполяционной формуле Лагранжа (гл. II § 2) имеем

$$Q_l(y) = \sum_{j=0}^l Q_l(y_j) \prod_{i \neq j} \frac{y - y_i}{y_j - y_i},$$

поэтому

$$Q_l(0) = \sum_{j=0}^l c_j Q_l(y_j), \quad \text{где} \quad c_j = \prod_{i \neq j} \frac{y_i}{y_i - y_j}, \quad y_j = M_j^{-1}, \quad (7)$$

и, следовательно,

$$I(f) = Q_l(0) = \sum_{j=0}^l c_j Q_l(y_j) = \sum_{j=0}^l c_j S_{M_j}(f) + \sum_{j=0}^l c_j r(M_j).$$

Мы получили соотношение (4) с выписанными явно значениями  $c_j$ .

Применение описанного метода, иногда называемого *методом Ромберга*, может быть полезным в следующей ситуации. Пусть мы задались какой-то квадратурой, вычислили на ЭВМ и выдали на печать значения  $S_{M \cdot 2^i}(f)$ ,  $i = 0, \dots, l$ , но оказалось, что нужной точности еще не достигли. Тогда можно попытаться получить приближение к интегралу, применив правило Ромберга по некоторой совокупности значений  $S_{M_0 \cdot 2^s}(f)$ ,  $\dots$ ,  $S_{M_0 \cdot 2^p}(f)$ , и для контроля над точностью применить это правило по совокупности значений  $S_{M_0 \cdot 2^{s+1}}(f)$ ,  $\dots$ ,  $S_{M_0 \cdot 2^{p+1}}(f)$ .

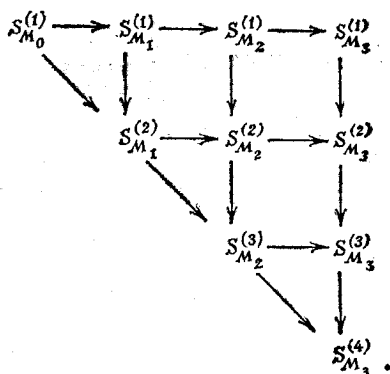
Иногда применяют следующую процедуру численного интегрирования. Задаются некоторым числом  $M_0$  и последовательно вычисляют приближенные значения интеграла по формуле трапеций  $I(f) \approx S_{M_k}^{(1)}(f)$  при  $M_k$  отрезках разбиения:  $M_k = M_0 \cdot 2^k$ . Удобнее всего вести вычисления по формуле

$$S_{M_k}^{(1)}(f) = \frac{1}{2} S_{M_{k-1}}^{(1)}(f) + \frac{B-A}{M_k} \sum_{j=1}^{M_{k-1}} f \left( A + \frac{2j-1}{M_k} (B-A) \right).$$

При каждом  $k$  после вычисления  $S_{M_k}^{(1)}(f)$  последовательно вычисляют  $S_{M_k}^{(2)}(f), \dots, S_{M_k}^{(k+1)}(f)$  по рекуррентной формуле

$$S_{M_k}^{(i)}(f) = S_{M_k}^{(i-1)}(f) + \frac{1}{2^i - 1} (S_{M_k}^{(i-1)}(f) - S_{M_{k-1}}^{(i-1)}(f)).$$

Таким образом, последовательность вычислений определяется схемой



Вычисления значений  $S_{M_k}^{(i)}(f)$  обычно продолжают до тех пор, пока при некотором  $k$  не окажется, что  $\min_i |S_{M_k}^{(i)}(f) - S_{M_{k-1}}^{(i)}(f)| < \epsilon$ .

Как правило, метод Ромберга существенно уступает по эффективности квадратуре Гаусса и методам интегрирования с автоматическим выбором шага (см. § 17).

**Задача 1.** Показать, что  $S_{M_k}^{(i)}(f)$  есть результат применения правила Ромберга к значениям  $S_{M_k}^{(1)}(f), \dots, S_{M_{k-i+1}}^{(1)}(f)$ .

## § 16. Вычисление интегралов в нерегулярном случае

Существенную часть реально встречающихся подынтегральных функций составляют функции с особенностями, причем особенность может содержаться либо в функции, либо в ее производной, или функции, производные которых очень велики. Если такая нерегулярность подынтегральной функции не вызвана колебательным характером ее поведения, то неплохой результат дают стандартные программы с автоматическим выбором шага, которые будут обсуждаться в § 17. В случае расчета малой серии интегралов с особенностями обращение к этим стандартным программам может оказаться наиболее целесообразным способом решения задачи. Для вычисления же большой серии интегралов с особенностями необходимо привлечь исследователей более высокой квалификации. Укажем ряд приемов, которые могут оказаться полезными при рассмотрении этих вопросов.

1. Выделение весовой функции. Пусть вычисляется интеграл  $\int_a^b f(x) dx$ , где пределы интегрирования  $a$  и  $b$  могут быть и бесконечными. Представим подынтегральную функцию в виде  $f(x) = g(x)p(x)$ , где  $p(x)$  — достаточно простая, а  $g(x)$  — гладкая функции. Далее применяем какой-либо из рассмотренных ранее способов вычисления интегралов с весом. Рассмотрим некоторые примеры.

Пусть вычисляется интеграл  $\int_{-1}^1 \frac{dx}{\sqrt{1-x^2}}$ . Представим  $f(x)$  в виде  $\frac{1}{\sqrt{1+x^2}} \cdot \frac{1}{\sqrt{1-x^2}}$ , где функция  $\frac{1}{\sqrt{1+x^2}}$  является гладкой. Функцию  $\frac{1}{\sqrt{1-x^2}}$  можно рассматривать как весовую. Этой весовой функции соответствует квадратура Эрмита.

Пусть вычисляется  $\int_0^1 f(x) dx$ , причем  $f(x)$  может быть представлена в виде  $g(x)x^\alpha$ , где  $-1 < \alpha < 1$ ,  $g(x)$  — гладкая функция,  $g(0) \neq 0$ . При вычислении интеграла по формуле трапеций с постоянным шагом  $H = M^{-1}$  погрешность стремится к нулю медленнее, чем  $M^{-2}$ . Один из возможных способов вычисления интеграла — обращение к квадратурам Гаусса, соответствующим данной весовой функции.

2. Можно пойти по пути разбиения интеграла на части и вычисления интеграла по каждой части при помощи построений из § 7. Представим интеграл в виде

$$I = \sum_{q=1}^{IM} I_q, \quad I_q = \int_{(q-1)H}^{qH} g(x)x^\alpha dx, \quad H = \frac{1}{M}.$$

Заменяв функцию  $g(x)$  на интерполяционный многочлен

$$P_{(q)}(x) = g((q-1)H) + (x - (q-1)H) \frac{(g(qH) - g((q-1)H))}{H},$$

получим

$$\begin{aligned} I_q &\approx \int_{(q-1)H}^{qH} P_{(q)}(x) x^\alpha dx = \\ &= H^{\alpha+1} q^{\alpha+2} \left( \left( \frac{1 - (1-1/q)^{\alpha+1}}{\alpha+1} - \frac{1 - (1-1/q)^{\alpha+2}}{\alpha+2} \right) g(qH) + \right. \\ &\left. + \left( \frac{1 - (1-1/q)^{\alpha+2}}{\alpha+2} - \frac{1-1/q - (1-1/q)^{\alpha+2}}{\alpha+1} \right) g((q-1)H) \right). \quad (1) \end{aligned}$$

Суммируя по  $q$  правые части в (1), получим квадратуру для вычисления исходного интеграла. В ряде случаев будет удобнее

положить  $g(qH) = (qH)^{-\alpha} f(qH)$  и, таким образом, получить квадратуру, имеющую вид

$$I \approx \sum_{q=0}^M D(q, H) f(qH). \quad (2)$$

Задача 1. Для квадратуры (2) получить оценку погрешности вида

$$\text{const} \cdot \max_{[0, 1]} |g''(x)| M^{-2}.$$

Далее будут рассмотрены более простые по виду способы вычисления интегралов от функций с особенностями. Описанный выше способ аппроксимации интеграла по значениям функции на фиксированной, в частности, равномерной сетке обладает определенными преимуществами в случае, когда задача вычисления интеграла представляет часть более сложной задачи, например при решении интегральных уравнений путем сведения к решению системы алгебраических уравнений. Иногда необходимая точность уже достигается при замене функции  $g(x)$  на отрезках разбиения на постоянную. В этом случае полагаем

$$\int_{a_{q-1}}^{a_q} g(x) p(x) dx \approx g(\xi_q) \int_{a_{q-1}}^{a_q} p(x) dx$$

и квадратура для вычисления исходного интеграла приобретает вид

$$\int_0^1 g(x) p(x) dx \approx \sum_{q=1}^M g(\xi_q) \int_{a_{q-1}}^{a_q} p(x) dx. \quad (3)$$

Задача 2. Пусть вычисляется интеграл

$$I = \int_0^1 \frac{g(x) b}{b^2 + x^2} dx, \text{ где } b \text{ — малое число.}$$

Показать, что при использовании формулы трапеций с постоянным шагом  $a_q - a_{q-1} \equiv H = M^{-1}$  погрешность оценивается через

$$\text{const} \cdot \min \left( \frac{1}{Mb}, \frac{1}{(Mb)^2} \right). \quad (4)$$

Задача 3. При  $a_q - a_{q-1} \equiv H$  квадратура (3) для этого интеграла имеет вид

$$I \approx \sum_{q=1}^M g(\xi_q) \left( \text{arctg} \left( \frac{qH}{b} \right) - \text{arctg} \left( \frac{(q-1)H}{b} \right) \right),$$

где  $(q-1)H \leq \xi_q \leq qH$ . Получить оценку погрешности:

$$|R_M| \leq \text{const} \cdot \max_{[0, 1]} |g'(x)| M^{-1}.$$

В рассматривавшихся выше случаях коэффициенты квадратур записываются в виде интегралов

$$\int_{a_{q-1}}^{a_q} P_i(x) \rho(x) dx,$$

где  $P_i(x)$  — некоторые многочлены; эти интегралы вычисляются в явном виде. Для ряда классов задач, где эти интегралы не вычисляются в явном виде, может оказаться разумным найти эти интегралы при помощи численного интегрирования. Эта дополнительная работа оправдывается, если получившиеся формулы используются многократно, например, при вычислении большой серии интегралов, при вычислении кратных интегралов как повторных (см. гл. V), а также при решении интегральных уравнений.

3. Пусть теперь вычисляется

$$I_\omega(f) = \int_0^1 f(x) \exp(i\omega x) dx,$$

где  $\omega$  — большое число,  $f(x)$  — достаточно гладкая функция. Будем рассматривать функцию  $\exp\{i\omega x\}$  как весовую. Представим интеграл в виде

$$I = \sum_{q=1}^M I_q, \quad I_q = \int_{(q-1)H}^{qH} f(x) \exp(i\omega x) dx,$$

для вычисления интеграла  $I_q$  применим квадратуру типа (6.2).

4. В некоторых случаях подынтегральную функцию можно представить в виде  $f(x) = G(x) + g(x)$ , причем  $\int_A^B G(x) dx$  берется в явном виде, а  $g(x)$  — гладкая функция. Пусть вычисляется интеграл

$$I = \int_0^1 f(x) dx, \quad f(x) = \frac{\ln x}{1+x^2}. \quad (5)$$

Возьмем  $G(x) = \ln x$ . Тогда функция  $g(x)$  имеет вид

$$g(x) = -\frac{x^2 \ln x}{1+x^2}.$$

Величина  $\int_0^1 |g'(x)| dx$  будет конечна, и можно показать, что по-

грешность вычисления интеграла  $\int_0^1 g(x) dx$  по формуле трапеций с постоянным шагом  $a_q - a_{q-1} = M^{-1}$  имеет порядок  $O(M^{-2})$ . Чтобы погрешность формулы Симпсона имела порядок  $O(M^{-4})$ , следует взять  $G(x) = (1-x^2) \ln x$ .

В случае  $f(x) = (x^2 + b^2)^{-1}e^x$ ,  $b$  — малое число, целесообразно взять

$$G(x) = (\text{Выч}_{bt} f(z))(x - bi)^{-1} + (\text{Выч}_{-bt} f(z))(x + bi)^{-1}.$$

Достижимое здесь расширение области аналитичности подынтегральной функции особенно эффективно при использовании формулы Гаусса.

5. Другим способом устранения особенности подынтегральной функции является замена переменной интегрирования. При замене переменных  $x = \varphi(t)$ ,  $\varphi(0) = 0$ ,  $\varphi(1) = 1$  исходный интеграл  $I = \int_0^1 f(x) dx$  преобразуется к виду

$$\int_0^1 g(t) dt, \quad (6)$$

где

$$g(t) = f(\varphi(t))\varphi'(t).$$

За счет множителя  $\varphi'(t)$  происходит устранение особенностей подынтегральной функции в отдельных точках. Произведя в интеграле (5) замену переменной  $x = t^k$ , получим

$$I = \int_0^1 \frac{k^2 t^{k-1} \ln t}{1 + t^{2k}} dt.$$

При  $k > 2$  интеграл  $\int_0^1 |g''(t)| dt$  конечен, поэтому погрешность формулы трапеций имеет порядок  $O(M^{-2})$ . С увеличением  $k$  растет порядок производных  $g^{(n)}(t)$ , для которых интеграл

$\int_0^1 |g^{(n)}(t)| dt$  ограничен, поэтому можно применять квадратуры все более высокого порядка точности.

Если  $k$  очень большое, то производные функции  $g(t)$  хотя и конечны, но очень большие, следовательно, должна соблюдаться определенная пропорция между величиной  $k$  и числом узлов  $N$ . Необходимость соблюдения осторожности при употреблении очень больших  $k$  видна хотя бы из следующего. Постоянный шаг  $t_q - t_{q-1} = 1/M$  в интеграле (6) соответствует узлам интегрирования  $a_q = \varphi(q/M) = (q/M)^k$  в исходном интеграле, поэтому при больших  $k$  используется мало значений подынтегральной функции в правой части отрезка  $[0, 1]$  (рис. 3.16.1).

6. Как мы уже видели в § 11, скорость сходимости при вычислении интегралов от функций с особенностями повышается также за счет распределения узлов интегрирования,

7. В некоторых случаях приходится идти по пути сочетания некоторых из описанных способов. Пусть вычисляется интеграл

$$\int_0^1 g(x) x^\alpha \exp\{i\omega x\} dx,$$

где  $\omega$  — большое число,  $g(x)$  — гладкая функция,  $g(0) \neq 0$ ,  $|\alpha| < 1$ . Наличие множителя  $\exp\{i\omega x\}$  требует выделения его

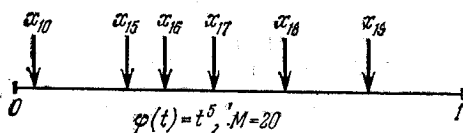


Рис. 3.16.1

как весового. Наличие множителя  $x^\alpha$  требует принятия специальных мер для интегрирования в окрестности нуля. Замена переменных  $x = \varphi(t)$  в данном случае является неприемлемой, поскольку для соответствующей весовой функции  $\exp\{i\omega\varphi(t)\}$  невозможно вычисление в явном виде коэффициентов квадратурных формул. Здесь целесообразнее разбить отрезок интегрирования на неравные части, соответствующие оптимальному распределению узлов при вычислении интеграла от функции  $x^\alpha$ , и применить на каждой части интерполяционные квадратурные формулы (§ 3), соответствующие весовой функции  $\exp\{i\omega x\}$ .

В случае интегралов типа  $\int_0^1 g(x) x^{-\alpha} \sin \omega x dx$  при  $\alpha > 1$ ,  $g(x)$  — гладкой функции,  $g(0) \neq 0$ , такой способ будет неприемлемым, поскольку в окрестности точки  $x = 0$  неинтегрируемая функция  $x^{-\alpha}$  не аппроксимируется многочленами. Здесь целесообразно

разбить исходный интеграл на части  $\int_0^\varepsilon$  и  $\int_\varepsilon^1$ , где  $\varepsilon \sim 1/\omega$ . Для

вычисления второго интеграла разумно применить процедуру, которая описана выше. В первом интеграле функция  $\sin \omega x$  не играет роли осциллирующего множителя, поскольку при таком выборе  $\varepsilon$  она имеет на  $[0, \varepsilon]$  конечное число колебаний. Поэтому этот интеграл можно вычислять, например, распределив узлы интегрирования соответственно оптимальному распределению для функции  $g(0)\omega x^{1-\alpha}$ , аппроксимирующей подынтегральную при малых  $\omega x$ .

8. Упомянем метод Ромберга. Погрешность формулы трапеций с постоянным шагом при вычислении интеграла  $\int_0^1 g(x) x^\alpha dx$  для гладкой функции  $g(x)$ ,  $g(0) \neq 0$ ,  $-1 < \alpha < 1$ , представ-



ляется в виде  $D_1 N^{-1-\alpha} + D_2 N^{-2-\alpha} + \dots$ , и имеются основания для применения приемов, положенных в основу метода Ромберга.

9. Решение ряда задач сводится к вычислению сингулярных интегралов типа

$$I(a) = \int_A^B \frac{g(x)}{x-a} dx,$$

где  $a \in (A, B)$ ,  $g(a) \neq 0$ . Интеграл понимается в смысле главного значения, т. е. как предел

$$\lim_{\varepsilon \rightarrow 0} \left( \int_A^{a-\varepsilon} \frac{g(x)}{x-a} dx + \int_{a+\varepsilon}^B \frac{g(x)}{x-a} dx \right).$$

Интеграл может быть записан как сумма интеграла по отрезку, симметричному относительно точки  $a$ , и интеграла от гладкой функции по оставшейся части. Для простоты предполагаем, что

первый интеграл преобразован к виду  $\int_{-b}^b \frac{g(x)}{x} dx$ . Если  $g(x)$

удовлетворяет условию Гельдера в точке  $x=0$ , т. е.  $|g(x) - g(0)| \leq A|x|^\alpha$ ,  $\alpha > 0$ , то последний интеграл равен несингу-

лярному интегралу  $\int_0^b \frac{g(x) - g(-x)}{x} dx$ . В частности, если  $g(x)$  —

гладкая функция, то новая подынтегральная функция  $\frac{g(x) - g(-x)}{x}$  также является гладкой.

В ряде случаев, например при решении интегральных уравнений с сингулярными ядрами, возникает следующая ситуация. Значения функции  $g(x)$  задаются на некоторой фиксированной сетке. Исходя из информации об этих значениях, требуется вычислять значения интеграла  $I(a)$  при различных  $a$ . Если  $g(x)$  — достаточно гладкая функция, то здесь можно поступить следующим образом.

Разбиваем отрезок  $[A, B]$  на части  $[A_0, A_1], \dots, [A_{M-1}, A_M]$ ,  $A_0 = A$ ,  $A_M = B$ . На каждой из частей  $[A_{q-1}, A_q]$  приближаем функцию  $g(x)$  интерполяционным многочленом  $L^q(x)$ .

При этом требуем, чтобы при всех  $q$  было выполнено условие

$$L^q(A_q) = L^{q+1}(A_q) = g(A_q).$$

Исходный интеграл заменяем суммой интегралов

$$i_q = \int_{A_{q-1}}^{A_q} \frac{L^q(x)}{x-a} dx.$$

Интегралы  $i_q$  вычисляем в явном виде. Если  $a \in (A_{q-1}, A_q)$ , то соответствующий интеграл  $i_q$  следует рассматривать как сингулярный. Если  $a = A_q$ , то

следует объединить интегралы  $i_q$  и  $i_{q+1}$  (расходящиеся) в один сингулярный интеграл

$$i_q + i_{q+1} = \int_{A_{q-1}}^{q+1} \frac{\tilde{L}^q(x)}{x - A_q} dx,$$

$$\tilde{L}^q(x) = \begin{cases} L^q(x) & \text{при } x < A_q, \\ L^{q+1}(x) & \text{при } x > A_q. \end{cases}$$

Получившиеся интегралы вычислим в явном виде.

Задача 5. Пусть отрезок интегрирования разбит на равные части длины  $H$  и на каждой части функция  $g(x)$  аппроксимируется при помощи линейной интерполяции. Таким образом, исходный интеграл аппроксимируется суммой интегралов

$$\sum_{q=1}^M \int_{A+(q-1)H}^{A+qH} \frac{g((q-1)H) + (x - (q-1)H) \frac{g(qH) - g((q-1)H)}{H}}{x - a} dx,$$

где  $H = (B - A)M^{-1}$ . В предположении ограниченности  $|g''(x)|$  получить оценку погрешности  $O(M^{-2} \ln M)$ .

Полезно указать на следующую практически важную деталь. Если решение задачи содержит какие-то неисследованные особенности, ухудшающие сходимость методов, то лучше сразу выделить простейшую модельную задачу, содержащую эти особенности, и провести выбор метода и проверку применимости различных асимптотических критериев на этой модельной задаче. Этот путь обычно приводит к более быстрому пониманию существа вопроса и избавляет от необходимости проведения многочисленных экспериментов на самой задаче. В частности, достигается экономия труда математика при программировании задачи, машинного времени и упрощается исследование за счет возможности построения более содержательных графиков поведения погрешности; здесь имеем возможность получить больше точек (13.8) или (13.12), поскольку для простой задачи их получение менее трудоемко.

## § 17. Принципы построения стандартных программ с автоматическим выбором шага

Как отмечалось в § 12, интегралы от функций с особенностями типа  $x^b$  хорошо вычисляются методами интегрирования с переменным шагом, если узлы интегрирования распределены оптимальным образом. По-видимому, столь же хорошо будут интегрироваться функции с особенностями других типов. Поэтому представляется заманчивым строить стандартные программы численного интегрирования так, чтобы для любой функции распределение узлов являлось оптимальным или близким к нему.

В § 11 указана возможность распределения узлов, близкого к оптимальному, после исследования поведения подынтеграль-

ной функции на редкой сетке. Однако в случае резко меняющихся функций, например функций типа  $x^b$ , реализация этой возможности приводит к недостаточно удовлетворительным результатам. Если первоначальная сетка, на которой исследуется подынтегральная функция, частая, то сильно загружается память ЭВМ; если же она редкая, то не удастся хорошо аппроксимировать оптимальное распределение узлов на участках резкого изменения подынтегральной функции.

Поэтому при разработке стандартных программ интегрирования приняты несколько иные процедуры распределения узлов интегрирования, обеспечивающие лучшее приближение к оптимальному распределению узлов для функций с особенностями. Рассмотрим некоторые из них.

Для вычисления интегралов по элементарным отрезкам разбиения  $[a_{q-1}, a_q]$  выбираются: квадратурная формула

$$I_q(f) \approx \frac{a_q - a_{q-1}}{2} \sum_{j=1}^m D_j f \left( \frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j \right) \quad (1)$$

и мера погрешности

$$\rho_q(f) = \left| \frac{a_q - a_{q-1}}{2} \sum_{j=1}^m B_j f \left( \frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j \right) \right|. \quad (2)$$

Пусть вычисляется

$$\int_A^B f(x) dx, \quad A = a_0.$$

Первая процедура, которую естественно назвать *горизонтальной*, определяется заданием параметров  $\beta$ ,  $\sigma < 1$ ,  $h_0$  и  $\epsilon_0$ . Полагая  $\epsilon_1 = \epsilon_0 \beta$ . Предположим, что каким-то образом уже вычис-

лено приближенное значение интеграла  $\int_{a_0}^{a_q} f(x) dx$ . Программа

располагает в каждый момент времени некоторым значением  $h_q$ , с которым надо начинать считать оставшуюся часть интеграла. Вычисляем величину  $\rho_q(f)$ , соответствующую отрезку  $[a_q, a_q + h_q]$ . Если оказалось, что  $\rho_q(f) \leq \epsilon_0$ , то вычисляем при-

ближенное значение  $\int_{a_q}^{a_q+h_q} f(x) dx$  по формуле (1) и полагаем

$a_{q+1} = a_q + h_q$ . Мы получили приближенное значение величины

$\int_{a_0}^{a_{q+1}} f(x) dx$ . В случае, когда  $\epsilon_1 < \rho_q(f)$ , полагаем  $h_{q+1} = h_q$ ,

в противном случае полагаем  $h_{q+1} = h_q/\sigma$ . Мы готовы к следующему шагу. Если оказалось, что  $\rho_q(f) \geq \epsilon_0$ , то принимаем  $\sigma h_q$  за

новое значение величины  $h_q$  и возвращаемся к исходной позиции: вычислено значение интеграла  $\int_{a_0}^{a_q} f(x) dx$  и задан шаг  $h_q$ . Начальные условия для применения процедуры:

$$q=0, \quad \int_{a_0}^{a_q} f(x) dx = 0, \quad h_0.$$

Процедура должна также иметь блок окончания работы: если оказалось, что  $a_q + h_q > B$ , то следует положить  $h_q = B - a_q$ . Установилась практика брать  $\sigma = 0,5$ .

Другая процедура, которую можно назвать *вертикальной*, определяется заданием числа  $\varepsilon_0$  и заключается в следующем. Пусть на каком-то шаге возникает необходимость вычисления

интеграла по отрезку разбиения  $[c, d]$ , т. е.  $\int_c^d f(x) dx$ . Вычисляет-

ся величина  $\rho(f)$ , соответствующая этому отрезку. Если она оказалась меньше  $\varepsilon_0$ , то этот интеграл вычисляется по соответствующей формуле (1) и программа переходит к следующему справа отрезку разбиения. В противном случае отрезки  $[c, (c+d)/2]$  и  $[(c+d)/2, d]$  объявляются отрезками разбиения и программа сначала обращается к вычислению интеграла по левому из этих отрезков. В начале работы программа обращается к вычислению

исходного интеграла  $\int_A^B f(x) dx$ . Некоторым недостатком этой про-

цедуры является необходимость запоминания отрезков разбиения, интегрирование по которым на данный момент времени не произведено.

Распределение узлов, осуществляемое этими процедурами, не является асимптотически оптимальным в смысле, определяемом в § 11, по следующим причинам: величина  $\rho_q(f)$  не является главным членом погрешности квадратуры (1), а, как правило, есть лишь некоторая грубая, завышенная по порядку оценка для него; отрезки разбиения могут принимать лишь довольно редкий ряд значений (при этом для первой процедуры они имеют вид  $h_0 \sigma^k$ , а для второй —  $(B - A) 2^{-k}$ ).

Рассмотрим некоторые моменты, связанные с практическим использованием описанных выше процедур.

Чтобы в отдельных случаях сделать распределение узлов более близким к оптимальному, иногда величины  $\rho_q(f)$ , соответствующие отрезкам разбиения, сравнивают не с  $\varepsilon_0$ , а с  $\varepsilon_0 h_q^\gamma$ , где  $\gamma$  специально подбирается. На первоначальном этапе применения ЭВМ сложилась традиция производить сравнение величины  $\rho_q(f)$  с величиной  $\varepsilon_0 h_q$ , т. е. брать  $\gamma = 1$ . Оказалось, что для резко меняющихся функций получающееся распределение узлов было далеко от оптимального, а в случае функций с разрывами программа не могла со-

считать интеграл. В самом деле, пусть, например,

$$f(x) = \begin{cases} 0 & \text{при } x < c, \\ 1 & \text{при } x > c. \end{cases}$$

Если при каком-то  $q$  величина  $\rho_q(f)$  содержит одновременно значения  $f$  в узлах, где  $f = 0$  и  $f = 1$ , то  $\rho_q(f)$  имеет порядок  $h_q$ , поэтому нет оснований надеяться, что неравенство  $|\rho_q(f)| \leq \epsilon_0 h_q$  будет выполняться при малых  $h_q$ ; программа будет осуществлять дробление шага до машинного нуля. Если функция меняется на малом участке очень резко, то шаг будет дробиться неоправданно сильно. Это обстоятельство было отмечено пользователями, и после теоретического анализа, в результате которого и была решена задача об оптимальном распределении узлов, было принято полагать  $\gamma = 0$ , если нет особых причин против этого.

В случае резко меняющихся функций следует иметь в виду, что программа может не заметить участка резкого изменения функции. Пусть  $\rho(f) = 0$ , когда  $f(x) = P_s(x)$  — многочлен степени  $s$ , и пусть реальная подынтегральная функция есть

$$f(x) = P_s(x) + g(x),$$

где  $g(x)$  практически отлична от 0 лишь на малом отрезке, например

$$g(x) = \frac{1}{\sqrt{2\pi\epsilon}} \exp\left\{-\frac{(x-C)^2}{2\epsilon}\right\}$$

(рис. 3.17.1). Для определенности обратимся к первой из описанных выше процедур. Если отрезок разбиения  $[a_{q-1}, a_q]$  удален от точки  $C$ , то

$\rho(f) \approx \rho(P_s) = 0$  и программа будет увеличивать шаг. При подходе к точке  $C$  шаг может оказаться настолько большим, что окрестность, где функция  $g(x)$  существенно больше 0, окажется заключенной между узлами. Тогда  $\rho(f)$  будет близко к 0 и на отрезке, содержащем точку  $C$ . В результате в качестве значения  $I(f)$  мы получим значение  $I(P_s)$ ; погрешность этого приближенного значения близка к 1. Для контроля над точностью можно было бы попытаться произвести интегрирование с другим значением  $\epsilon_0$ , однако с большой вероятностью получился бы тот же результат.

Не следует думать, что этот недостаток свойствен только методам интегрирования с автоматическим выбором шага. Если производить вычисления этого интеграла по формулам с постоянным шагом  $H$ , то при  $H \gg \sqrt{\epsilon}$  все равно может оказаться, что  $g(x) \approx 0$  во всех узлах интегрирования, и мы получим приближенное значение  $I(f) \approx I(P_s)$ . Производя численное интегрирование при нескольких шагах и оценивая погрешность по правилу Рунге, можно прийти к неправильному выводу, что приближенное значение интеграла  $I(f) \approx I(P_s)$ . В целом можно сказать, что в случае использования алгоритмов интегрирования с автоматическим выбором шага возможность получения подобных неправильных выводов несколько больше.

При составлении стандартных программ всегда приходится балансировать между двумя крайностями: гарантией требуемой точности для любой подынтегральной функции или быстрым вычислением интеграла с нужной точностью для большинства предъявляемых к решению задач.

По-видимому, можно говорить о каком-то принципе неопределенности по отношению к методам высокой эффективности: дальнейшее повышение скорости работы должно сопровождаться уменьшением надежности. Чтобы избежать «проскакивания» областей резкого изменения функции, можно предусмо-

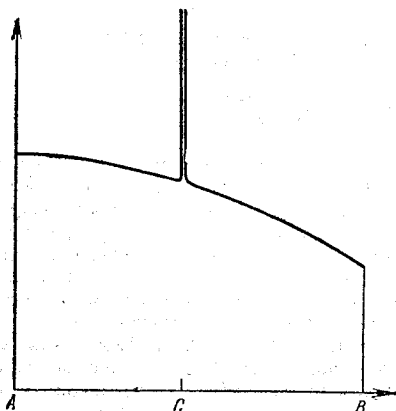


Рис. 3.17.1

треть в программе наличие отрезков, где шаг интегрирования дробится принудительно. Например, в описании стандартной программы можно включить концы некоторого отрезка  $[\alpha, \beta]$ . Если отрезок разбиения  $[a_{q-1}, a_q]$  пересекается с  $[\alpha, \beta]$ , то следующий отрезок разбиения выбирается по более сложному правилу. В качестве  $[\alpha, \beta]$  следует задавать некоторый отрезок, где подынтегральная функция резко меняется.

Обратите внимание на ряд дополнительных моментов. Пусть для гладкой функции  $f(x)$

$$\rho(f) = c_s \left| f^{(s)} \left( \frac{a_{q-1} + a_q}{2} \right) \right| (a_q - a_{q-1})^{s+1} + \varepsilon(a_{q-1}, a_q),$$

$$|\varepsilon(a_{q-1}, a_q)| \leq \bar{c}_s \max_{A, B} |f^{(s+1)}(x)| (a_q - a_{q-1})^{s+2}. \quad (3)$$

Для понимания механизма изменения шага предположим, что  $f^{(s)}(x)$  — кусочно-постоянная функция. В области постоянства  $f^{(s)}(x)$  имеем

$$\rho(f) = c_s |f^{(s)}(x)| (a_q - a_{q-1})^{s+1}.$$

Рассмотрим случай  $\beta < \sigma^{s+1}$ . Пусть для шага  $h^0$ ,  $s$  с которым мы приступаем к интегрированию в данной области постоянства производной, выполняется соотношение  $c_s |f^{(s)}(x)| (h^0)^{s+1} \leq \varepsilon_1$ . Тогда шаг будет увеличиваться каждый раз в  $\sigma^{-1}$  раз, пока величина  $\rho(f)$  не станет больше  $\varepsilon_1$ . Поскольку при шаге, в  $\sigma^{-1}$  раз меньше, выполняется соотношение  $\rho(f) \leq \varepsilon_1$ , то интегрирование будет осуществляться с наименьшим шагом  $h = h^0 \sigma^k$ , для которого  $c_s |f^{(s)}(x)| h^{s+1} > \varepsilon_1$ . Если же для начального в этой области шага выполняется соотношение  $c_s |f^{(s)}(x)| (h^0)^{s+1} > \varepsilon_0$ , то шаг станет наибольшим из шагов  $h^0 \sigma^k$ , для которых выполняется условие  $c_s |f^{(s)}(x)| h^{s+1} \leq \varepsilon_0$ . Если, как мы предположили,  $\varepsilon_1/\varepsilon_0 = \beta < \sigma^{s+1}$ , то может оказаться, что в зависимости от поведения функции при меньших  $x$  в рассматриваемой области выбираются различные шаги. В то же время мы видели, что шаг интегрирования желательно распределять так, чтобы оценка погрешности на всех отрезках разбиения интервала интегрирования была примерно одинаковой. Исходя из сказанного мы делаем вывод о желательности употребления  $\beta \geq \sigma^{s+1}$ .

Рассмотрим случай  $\beta > \sigma^{s+1}$ . Можно указать шаг  $h = h_0 \sigma^k$  такой, что

$$\varepsilon_0 \sigma^{s+1} < c_s |f^{(s)}(x)| h^{s+1} \leq \varepsilon_0. \quad (4)$$

Если оказалось, что

$$\varepsilon_1 < c_s |f^{(s)}(x)| h^{s+1} \leq \varepsilon_0, \quad (5)$$

то при интегрировании по рассматриваемому отрезку будет выбран именно такой шаг. Однако при

$$\varepsilon_0 \sigma^{s+1} < c_s |f^{(s)}(x)| h^{s+1} \leq \varepsilon_1 = \varepsilon_0 \beta \quad (6)$$

программа выработает следующий шаг, равный  $h/\sigma$ . Для этого шага, вследствие (6), окажется, что  $\varepsilon_0 < \rho(f)$ , поэтому шаг  $h/\sigma$  будет признан непригодным. Программа возьмет шаг, равный  $h$ ; поскольку для него выполняется условие (6), то интегрирование по текущему отрезку будет закончено и за исходный шаг для дальнейшего счета будет принят шаг  $h/\sigma$ . Таким образом, фактическое интегрирование происходит с шагом  $h$  и на каждом шаге делается попытка произвести интегрирование с шагом  $h/\sigma$ . Если вычисления с шагами  $h$  и  $h/\sigma$  независимы, то на каждом шаге объем работы будет удваиваться.

В статистике известен следующий факт: если мы зададимся произвольным  $\lambda > 1$  и из различных независимых источников получим некоторые числа  $y$ , то величины  $\{\log_\lambda y\}$  будут асимптотически равномерно распределены на отрезке  $[0, 1]$ .

**Задача 1.** Считая, что  $f^{(s)}(x)$  — числа, происходящие из независимых случайных источников, показать, что при  $\beta > \sigma^{s+1}$  математическое ожидание затрат работ пропорционально  $2 - \log_{\sigma^{s+1}} \beta$ . (Отсюда следует целесообразность выбора  $\beta = \sigma^{s+1}$ .)

Наши рассуждения проведены в предположении, что  $f^{(s)}(x) = \text{const}$ , и поэтому требуют экспериментальной проверки. Исходя из результатов такой проверки на практике  $\beta$  обычно берется равным или несколько большим, чем  $\sigma^{s+1}$ , например, в одной из распространенных стандартных программ  $s = 4$ ,  $\sigma = 1/2$ ,  $\beta = 1/10$ .

Как и в случае интегрирования с постоянным шагом, возникает вопрос практической оценки погрешности. Обратимся ко второй процедуре, имеющей более простое описание. Как и ранее, проведем исследование на примере кусочно-постоянной производной  $f^{(s)}(x)$ . Шаг интегрирования в области постоянства  $f^{(s)}(x)$  для этой процедуры определяется условием: это наибольший шаг  $h$  вида  $(B - A)2^{-k}$ , для которого выполнено условие

$$c_s |f^{(s)}(x)| h^{s+1} \leq \varepsilon_0. \quad (7)$$

Очевидно, что для такого максимального шага справедливо соотношение

$$\varepsilon_0 2^{-(s+1)} < c_s |f^{(s)}(x)| h^{s+1}.$$

Произведем интегрирование с некоторым  $\varepsilon'_0 < \varepsilon_0$ . В области, где

$$\varepsilon'_0 < c_s |f^{(s)}(x)| h^{s+1} \leq \varepsilon_0,$$

«старый» шаг интегрирования будет уже непригоден, поэтому произойдет дробление шага не менее чем вдвое. В области, где выполняется условие

$$\varepsilon_0 2^{-(s+1)} < c_s |f^{(s)}(x)| h^{s+1} \leq \varepsilon'_0,$$

дробления шага интегрирования не произойдет. Таким образом, при  $\varepsilon_0 2^{-(s+1)} < \varepsilon'_0$  может случиться, что дробление шага интегрирования произойдет лишь на части отрезка интегрирования или его вообще не будет. Из сравнения результатов расчетов с  $\varepsilon_0$  и  $\varepsilon'_0$  мы не получим информации о величине погрешности интегрирования в той области, где дробления шага не произошло. При условии

$$\varepsilon'_0 \leq \varepsilon_0 2^{-(s+1)} \quad (8)$$

шаг дробится всюду, однако отсюда не следует, что сравнение результатов расчетов с  $\varepsilon_0$  и любым  $\varepsilon'_0$ , удовлетворяющим (8), дает надежную гарантию точности результата.

Задача 2. Пусть  $\epsilon'_0 < \epsilon_0 2^{-(s+1)}$ . Подобрать постоянные  $c_1, c_2, C$  такие, что при

$$f^{(s)}(x) = \begin{cases} c_1 & \text{на } [A, C], \\ c_2 & \text{на } [C, B] \end{cases}$$

приближенные значения  $S_{\epsilon_0}(f)$  и  $S_{\epsilon'_0}(f)$  интеграла, соответствующие  $\epsilon_0$  и  $\epsilon'_0$  будут одинаковыми, а погрешность отлична от нуля.

Если  $f^{(s)}(x)$  непрерывна, то при  $\epsilon'_0 = \epsilon_0 2^{-(s+1)}$  и при введении в алгоритм условия уменьшения максимального шага, пропорционального  $\epsilon_0^{1/(1+s)}$ , можно предложить и обосновать аналог правила Рунге оценки погрешности

$$I(f) - S_{\epsilon'_0}(f) \sim \frac{S_{\epsilon'_0}(f) - S_{\epsilon_0}(f)}{2^m - 1}, \quad \epsilon_0 \rightarrow 0;$$

здесь  $m + 1$  — порядок главного (при  $\max(a_q - a_{q-1}) \rightarrow 0$ ) члена погрешности квадратуры (1).

Литература к гл. [35—37, 48, 69].



## Глава IV

### ПРИБЛИЖЕНИЕ ФУНКЦИЙ И СМЕЖНЫЕ ВОПРОСЫ

Как мы видели, непрерывная функция не всегда может быть хорошо приближена интерполяционным многочленом Лагранжа. В частности, последовательность интерполяционных многочленов Лагранжа по равноотстоящим узлам не обязательно сходится к функции даже в том случае, если функция бесконечно дифференцируема. В тех случаях, когда сходимость имеет место, часто получение достаточно хорошего приближения требует использования полиномов высокой степени. В то же время если для приближаемой функции удастся подобрать подходящие узлы интерполяции, то степень интерполяционного многочлена, приближающего функцию с заданной точностью, может быть значительно снижена.

В ряде конкретных случаев целесообразно приближать функцию не путем интерполяции, а путем построения так называемого *наилучшего приближения*. Проблемы, связанные с построением элементов наилучшего приближения, и будут рассмотрены в настоящей главе.

#### § 1. Наилучшие приближения в линейном нормированном пространстве

Сформулируем задачу построения наилучшего приближения на абстрактном языке. Пусть имеется элемент  $f$  линейного нормированного пространства  $R$ . Требуется найти его наилучшее приближение линейной комбинацией  $\sum_{j=1}^n c_j g_j$  данных линейно независимых элементов  $g_1, \dots, g_n \in R$ . Это означает: найти элемент  $\sum_{j=1}^n c_j^0 g_j$  такой, что

$$\left\| f - \sum_{j=1}^n c_j^0 g_j \right\| = \Delta = \inf_{c_1, \dots, c_n} \left\| f - \sum_{j=1}^n c_j g_j \right\|.$$

Если такой элемент существует, то он называется *элементом наилучшего приближения*.

**Теорема.** *Элемент наилучшего приближения существует.*

Доказательство. Вследствие соотношений (неравенство треугольника)

$$\left\| f - \sum_{j=1}^n c_j^1 g_j \right\| - \left\| f - \sum_{j=1}^n c_j^2 g_j \right\| \leq \left\| \sum_{j=1}^n (c_j^1 - c_j^2) g_j \right\| \leq \sum_{j=1}^n |c_j^1 - c_j^2| \|g_j\|$$

функция

$$F_f(c_1, \dots, c_n) = \left\| f - \sum_{j=1}^n c_j g_j \right\|$$

является непрерывной функцией аргументов  $c_j$  при любом  $f \in R$ . Пусть  $|c|$  — евклидова норма вектора  $c = (c_1, \dots, c_n)$ . Функция  $F_0(c_1, \dots, c_n) = \|c_1 g_1 + \dots + c_n g_n\|$  непрерывна на единичной сфере  $|c| = 1$  и, следовательно, в некоторой ее точке  $(\tilde{c}_1, \dots, \tilde{c}_n)$  достигает своей нижней грани  $\tilde{F}$  по сфере, причем  $\tilde{F} \neq 0$ , так как равенство  $\tilde{F} = \|\tilde{c}_1 g_1 + \dots + \tilde{c}_n g_n\| = 0$  противоречит линейной независимости элементов  $g_1, \dots, g_n$ . Для любого  $c = (c_1, \dots, c_n) \neq (0, \dots, 0)$  справедлива оценка

$$\|c_1 g_1 + \dots + c_n g_n\| = F_0(c_1, \dots, c_n) = |c| F_0\left(\frac{c_1}{|c|}, \dots, \frac{c_n}{|c|}\right) \geq |c| \tilde{F}.$$

Пусть  $\gamma > 2\|f\|/\tilde{F}$ . Функция  $F_f(c_1, \dots, c_n)$  непрерывна в шаре  $|c| \leq \gamma$ ; следовательно, в некоторой точке шара  $(c_1^0, \dots, c_n^0)$  она достигает своей нижней грани  $F^0$  по шару. Имеем  $F^0 \leq F_f(0, \dots, 0) = \|f\|$ . Вне этого шара выполняются соотношения  $F_f(c_1, \dots, c_n) \geq \|c_1 g_1 + \dots + c_n g_n\| - \|f\| > \tilde{F} 2\|f\|/\tilde{F} - \|f\| = \|f\| > F^0$ .

Таким образом,

$$F_f(c_1, \dots, c_n) \geq F^0 = F_f(c_1^0, \dots, c_n^0)$$

при всех возможных  $c_1, \dots, c_n$ . Теорема доказана.

Элементов наилучшего приближения, вообще говоря, может быть несколько.

Пространство  $R$  называется *строго нормированным*, если из условия

$$\|f + g\| = \|f\| + \|g\|, \quad \|f\|, \|g\| \neq 0,$$

следует  $f = \alpha g, \alpha \neq 0$ .

Задача 1. Доказать, что в случае строго нормированного пространства  $R$  элемент наилучшего приближения единствен.

Задача 2. Доказать, что пространство  $L_p((0, 1), q(x))$ ,  $q(x) > 0$  почти всюду, с нормой

$$\|f\|_p = \sqrt[p]{\int_0^1 |f(x)|^p q(x) dx}$$

строго нормированное при  $1 < p < \infty$ .

Рассмотреть отдельно простейший случай гильбертова пространства  $p=2$ .

## § 2. Наилучшее приближение в гильбертовом пространстве и вопросы, возникающие при его практическом построении

Для гильбертова пространства элемент наилучшего приближения единствен (см. задачу 1.2) и проблема его нахождения формально сводится к решению системы линейных уравнений.

Наиболее простой способ получения этой системы следующий. По определению коэффициенты  $a_j$  элемента наилучшего приближения реализуют минимум выражения

$$\delta^2 = \left\| f - \sum_{j=1}^n a_j g_j \right\|^2 = \left( f - \sum_{j=1}^n a_j g_j, f - \sum_{j=1}^n a_j g_j \right).$$

Приравнивая нулю производные по  $\operatorname{Re} a_i$  и  $\operatorname{Im} a_i$ , получаем искомого систему уравнений для определения  $a_i$ . Вследствие существования и единственности элемента наилучшего приближения, эта система имеет единственное решение.

Построим эту систему и исследуем вопрос о единственности ее решения несколько иным способом.

Для простоты изложения ограничимся случаем вещественных  $f$  и  $g_j$ . Пусть  $a_j = \alpha_j + i\beta_j$ ,  $\alpha_j, \beta_j$  — вещественные числа; имеем  $f - \sum_{j=1}^n a_j g_j = \left( f - \sum_{j=1}^n \alpha_j g_j \right) - i \sum_{j=1}^n \beta_j g_j$ . Положим

$$\Phi(a_1, \dots, a_n) = \left\| f - \sum_{j=1}^n a_j g_j \right\|^2.$$

Если  $f_1$  и  $f_2$  вещественны, то  $\|f_1 + if_2\|^2 = (f_1 + if_2, f_1 + if_2) = (f_1, f_1) + i(f_2, f_1) - i(f_1, f_2) + (f_2, f_2) = \|f_1\|^2 + \|f_2\|^2$ , поэтому

$$\left\| f - \sum_{j=1}^n a_j g_j \right\|^2 = \left\| f - \sum_{j=1}^n \alpha_j g_j \right\|^2 + \left\| \sum_{j=1}^n \beta_j g_j \right\|^2. \quad (1)$$

Согласно (1) имеем равенство

$$\Phi(a_1, \dots, a_n) = \Phi(\alpha_1, \dots, \alpha_n) + \left\| \sum_{j=1}^n \beta_j g_j \right\|^2.$$

Отсюда следует, что

$$\inf_{a_1, \dots, a_n} \Phi(a_1, \dots, a_n) \geq \inf_{\alpha_1, \dots, \alpha_n} \Phi(\alpha_1, \dots, \alpha_n). \quad (2)$$

В то же время

$$\inf_{\alpha_1, \dots, \alpha_n} \Phi(\alpha_1, \dots, \alpha_n) \geq \inf_{a_1, \dots, a_n} \Phi(a_1, \dots, a_n), \quad (3)$$

поскольку в правой части берется нижняя грань по более широкому множеству всевозможных параметров  $a_1, \dots, a_n$ , а в левой — по множеству вещественных параметров. Из (2) и (3)

следует, что исходная задача сводится к нахождению

$$\inf_{\alpha_1, \dots, \alpha_n} \Phi(\alpha_1, \dots, \alpha_n).$$

В точке минимума должны выполняться условия  $\partial\Phi/\partial\alpha_k = 0$ .  
Имеем

$$\begin{aligned} \frac{\partial\Phi}{\partial\alpha_k} &= \left( -g_k, f - \sum_{j=1}^n \alpha_j g_j \right) + \left( f - \sum_{j=1}^n \alpha_j g_j, -g_k \right) = \\ &= -2 \left( f - \sum_{j=1}^n \alpha_j g_j, g_k \right) = 0. \end{aligned}$$

Отсюда получаем систему линейных уравнений относительно коэффициентов  $\alpha_j = a_j$ , соответствующих элементу наилучшего приближения

$$\sum_{j=1}^n a_j (g_j, g_k) = (f, g_k), \quad k = 1, \dots, n. \quad (4)$$

**Задача 1.** Доказать, что коэффициенты  $a_j$ , соответствующие элементу наилучшего приближения, являются решением (4) и в том случае, когда  $f$  и  $g_j$  не обязательно вещественны.

Матрица  $G_n = G(g_1, \dots, g_n) = [(g_j, g_k)]$  называется *матрицей Грама* системы элементов  $g_1, \dots, g_n$ . Поскольку  $(g_j, g_k) = \overline{(g_k, g_j)}$ , то матрица Грама является эрмитовой.

**Лемма.** Если элементы  $g_1, \dots, g_n$  линейно независимы, то матрица  $G_n$  положительно определена.

**Доказательство.** Пусть  $c = (c_1, \dots, c_n)^T$  — произвольный вектор с вещественными компонентами. Имеем равенство

$$\left\| \sum_{j=1}^n c_j g_j \right\|^2 = \left( \sum_{j=1}^n c_j g_j, \sum_{k=1}^n c_k g_k \right) = \sum_{j,k=1}^n c_j c_k (g_j, g_k). \quad (5)$$

Последнее выражение совпадает с  $(G_n c, c)$ , поэтому

$$(G_n c, c) = \left\| \sum_{j=1}^n c_j g_j \right\|^2 \geq 0.$$

Если элементы  $g_j$  линейно независимы, то  $\left\| \sum_{j=1}^n c_j g_j \right\| = 0$  только в том случае, когда все  $c_j = 0$ . Таким образом,  $(G_n c, c) > 0$ , если  $c \neq 0$ , и согласно определению матрица  $G_n$  является положительно определенной. Поскольку матрица  $G_n$  положительно определена, то ее определитель отличен от нуля и, следовательно, система (4) имеет единственное решение.

**Задача 2.** Доказать неравенство

$$G(g_1, \dots, g_{n+1}) \leq G(g_1, \dots, g_n) (g_{n+1}, g_{n+1}).$$

При практическом приближении функций нужно проявлять осторожность при выборе системы функций  $g_j$ . Оказывается, что

при неудачном выборе такой системы вычислительная погрешность коэффициентов  $a_j$  может достигать катастрофических размеров и с добавлением новых функций  $g_n$  получаемое «наилучшее» приближение будет все хуже приближать заданную функцию.

Дело заключается в следующем. Матрица  $G_n$  при неудачном выборе системы функций  $g_j$  имеет большой разброс собственных значений, т. е. отношение максимального (по модулю) собственного значения к наименьшему (по модулю) велико; вычислительная погрешность при решении систем с такой матрицей возрастает по крайней мере пропорционально этому разбросу. Например, в случае отрезка  $[-1, 1]$ , веса  $p(x) \equiv 1$  в системы функций  $g_j = x^{j-1}$  разброс собственных значений матрицы  $G_n$  превосходит  $a(\sqrt{2} + 1)^{2n}/n^b$ , где  $a, b$  — некоторые положительные постоянные. Поэтому в качестве систем функций  $g_j$  целесообразно брать системы ортонормированных функций или в каком-то смысле близких к ним.

Если элементы  $g_k$  образуют ортонормированную систему  $(g_k, g_j) = \delta_{kj}$ , то система (4) приобретает вид

$$a_j = (f, g_j). \quad (6)$$

Тогда наилучшее приближение записывается в форме

$$g = \sum_{j=1}^n (f, g_j) g_j$$

и имеется следующее удобное представление для величины  $\|f - g\|^2$ :

$$\begin{aligned} \|f - g\|^2 &= \left( f - \sum_{j=1}^n a_j g_j, f - \sum_{j=1}^n a_j g_j \right) = \\ &= (f, f) - \sum_{j=1}^n |a_j|^2 = (f, f) - \sum_{j=1}^n |(f, g_j)|^2. \end{aligned}$$

Поскольку  $\|f - g\|^2 \geq 0$ , то из равенства

$$\|f - g\|^2 = (f, f) - \sum_{j=1}^n |(f, g_j)|^2$$

следует, в частности, известное *неравенство Бесселя*

$$(f, f) \geq \sum_{j=1}^n |(f, g_j)|^2.$$

Если исходные элементы не образуют ортонормированной системы, то, вообще говоря, их можно ортогонализировать при помощи рассматривавшегося в гл. III алгоритма ортогонализации. Однако применение этого алгоритма довольно часто приводит к неудовлетворительным результатам. Например, при построении на отрезке  $[-1, 1]$  ортонормированной с весом  $p(x) > 0$ , си-

стемы функций из указанной выше системы функций  $x^j$  будут получены некоторые ортогональные многочлены  $P_j(x)$ , сумма модулей коэффициентов у которых растет не медленнее чем  $(\sqrt{2} + 1)^n n^a$ , где  $a$  определяется через  $\rho(x)$ . Если в дальнейшем значения многочленов вычисляются по явной формуле

$$P_j(x) = \sum_{k=0}^j a_{kj} x^k, \text{ то из-за большой величины суммы модулей}$$

коэффициентов будет большая погрешность в значениях самих многочленов. Для устойчивого вычисления значений многочленов нужно применить какой-то иной алгоритм, например вычислять их по рекуррентным формулам или по явным формулам типа  $T_n(x) = \cos(n \arccos x)$  в случае многочленов Чебышева. Некоторые более детальные сведения по этому вопросу будут приведены в § 8.

Пусть нам требуется приблизить функцию двух переменных  $f(x, y)$  в некоторой области  $G$  плоскости  $(x, y)$ . Явный вид ортогональных функций известен только для простейших областей. Можно применить следующий прием. Возьмем некоторую область  $\Omega \in G$ , для которой известна ортогональная система функций, и продолжим  $f$  в области  $\Omega \setminus G$ . Далее будем приближать  $f$  в области  $G$  с помощью этой системы. Такой прием иногда неэффективен, поскольку не удастся легко построить достаточно гладкое продолжение функции  $f$  в области  $\Omega \setminus G$ , а при недостаточно гладком продолжении приближение будет плохим. Следствием как этого обстоятельства, так и того, что приближение ищется в большей области, часто является неоправданное увеличение значения  $n$ , нужного для достижения заданной точности.

Другой возможный прием сведения к известной ортогональной системе функций состоит в следующем. Возьмем некоторую область  $\Omega$  с известной ортогональной системой  $\varphi_1(\xi, \eta), \dots, \varphi_n(\xi, \eta)$ . Пусть отображение  $x = x(\xi, \eta), y = y(\xi, \eta)$  переводит область  $\Omega$  в  $G, \xi = \xi(x, y); \eta = \eta(x, y)$  — обратное отображение. Будем приближать функцию  $f(x(\xi, \eta), y(\xi, \eta))$  в области

$\Omega$  линейными комбинациями  $\sum_{j=1}^n c_j \varphi_j(\xi, \eta)$ . Это равносильно при-

ближению  $f(x, y)$  в  $G$  линейными комбинациями  $\sum_{j=1}^n c_j \varphi_j(\xi(x, y), \eta(x, y))$ .

При приближении функций большого числа переменных не удается указать методов, пригодных во всех случаях. В каждой конкретной ситуации надо учитывать специфику задачи (вид функции, геометрию области и т. п.).

Рассмотрим для иллюстрации задачу другого рода, при решении которой используются проводимые выше построения. Пусть требуется построить интерполяционный многочлен степени  $N - 1$  с узлами интерполяции  $x_1^N, \dots, x_N^N$ . Пусть эти узлы являются нулями  $Q_N(x)$  степени  $N$  из ортонормированной си-

стемы многочленов  $\{Q_n(x)\}$ , соответствующей весу  $p(x)$ . Например, можно строить интерполяционные многочлены по нулям многочленов Чебышева, о которых шла речь в гл. II. Будем отыскивать интерполяционный многочлен в виде линейной комбинации

$$P_{N-1}(x) = \sum_{j=0}^{N-1} a_j Q_j(x).$$

При  $m, n < N$  многочлен  $Q_m(x)Q_n(x)$  имеет степень не выше  $2N-2$ . Поэтому квадратура Гаусса с  $N$  узлами точна для этого многочлена:

$$\frac{b-a}{2} \sum_{q=1}^N D_q^N Q_m(x_q^N) Q_n(x_q^N) = \int_a^b Q_m(x) Q_n(x) p(x) dx = \delta_n^m. \quad (7)$$

Таким образом, векторы  $Q_m = \{Q_m(x_1^N), \dots, Q_m(x_N^N)\}$  при  $m < N$  образуют ортонормированную систему относительно скалярного произведения

$$(y, z) = \sum_{q=1}^N D_q^N y_q z_q \quad (8)$$

и вектор  $f = (f(x_1^N), \dots, f(x_N^N))^T$  может быть разложен по этой системе векторов:

$$f = \sum_{j=0}^{N-1} d_j Q_j,$$

где  $d_j = (f, Q_j)$ . Многочлен

$$P_{N-1}(x) = \sum_{j=0}^{N-1} d_j Q_j(x) \quad (9)$$

будет искомым.

Таким образом, построение интерполяционного многочлена с узлами интерполяции, соответствующими корням ортогонального многочлена, сводится к вычислению коэффициентов разложения функции  $d_j$  по системе ортогональных многочленов. После этого искомым интерполяционный многочлен вычисляется по формуле (9). Этот алгоритм будет более устойчив по отношению к погрешностям округления по сравнению с непосредственным построением интерполяционного многочлена Лагранжа.

### § 3. Тригонометрическая интерполяция.

#### Дискретное преобразование Фурье

Дискретное преобразование Фурье применяется при решении многих прикладных задач. К ним относятся тригонометрическая интерполяция, вычисление свертки функций, распознавание образов и многие другие. Дискретное преобразование Фурье стало особенно эффективным методом решения прикладных задач после создания быстрого преобразования Фурье (см. § 4).

Пусть  $f(x)$  — периодическая функция с периодом 1 — может быть разложена в ряд Фурье

$$f(x) = \sum_{q=-\infty}^{\infty} a_q \exp \{2\pi i q x\}, \quad (1)$$

причем

$$\sum_{q=-\infty}^{\infty} |a_q| < \infty. \quad (2)$$

Здесь  $i$  — мнимая единица.

Рассмотрим значения этой функции на сетке из точек  $x_l = l/N$ , где  $l, N$  целые,  $N$  фиксировано, и обозначим  $f(x_l) = f_l$ . Если  $q_2 - q_1 = kN$ , где  $k$  целое, то  $q_2 x_l - q_1 x_l = kN x_l = kl$ , где  $kl$  целое. Следовательно,

$$\exp \{2\pi i q_1 x\} = \exp \{2\pi i q_2 x\} \quad (3)$$

в узлах сетки. Поэтому если функция  $f(x)$  рассматривается лишь на сетке узлов  $x_l$ , то в соотношении (1) можно привести подобные члены

$$f_l = \sum_{q=0}^{N-1} A_q \exp \{2\pi i q x_l\}, \quad (4)$$

где

$$A_q = \sum_{s=-\infty}^{\infty} a_{q+sN}. \quad (5)$$

*Лемма. При  $A_q$ , определяемых (5), соотношение (4) остается в силе, если пределы суммирования  $[0, N-1]$  заменить на  $[s, N-1+s]$ , где  $s$  — любое целое.*

В самом деле, если  $q' = q + kN$ , то  $A_{q'} = \sum_{s=-\infty}^{\infty} a_{q+kN+sN}$ . Принимая  $k \pm s$  за новую переменную суммирования  $s'$ , получим

$$A_{q'} = \sum_{s'=-\infty}^{\infty} a_{q+s'N} = A_q.$$

Поскольку в узлах сетки  $\exp \{2\pi i q' x_l\} = \exp \{2\pi i q x_l\}$  согласно (3), то в совокупности имеем

$$A_q \exp \{2\pi i q x_l\} = A_{q'} \exp \{2\pi i q' x_l\}.$$

Таким образом, функция  $A_q \exp \{2\pi i q x_l\}$  является периодической по  $q$  с периодом  $N$ , а сумма

$$\sum_{q=s}^{N-1+s} A_q \exp \{2\pi i q x_l\}$$

не зависит от  $s$  и совпадает с  $f_l$ . Лемма доказана.

Если с самого начала была задана функция, определенная только на сетке, то на этой сетке ее можно также представить



в форме (1). Действительно, такую функцию можно продолжить на всю прямую, доопределив ее между узлами сетки путем линейной интерполяции. Для непрерывной кусочно-дифференцируемой функции выполняется (2), поэтому в точках сетки после приведения подобных членов получим (4).

Определим скалярное произведение для функций на сетке следующим образом:

$$(f, g) = \frac{1}{N} \sum_{l=0}^{N-1} f_l \bar{g}_l.$$

(Множитель  $1/N$  введен для согласованности получаемых соотношений с непрерывным случаем: если  $f(x)$  и  $g(x)$  — непрерывные функции на отрезке  $[0, 1]$ , то

$$(f, g) \rightarrow \int_0^1 f(x) \bar{g}(x) dx$$

при  $N \rightarrow \infty$ .) Функции  $g_q(x) = \exp\{2\pi i q x\}$  при  $0 \leq q < N$  образуют ортонормированную систему относительно так введенного скалярного произведения. Действительно,

$$(g_q, g_j) = \frac{1}{N} \sum_{l=0}^{N-1} \exp\left\{2\pi i \frac{q-j}{N} l\right\}.$$

При  $q \neq j$ , суммируя геометрическую прогрессию, имеем

$$(g_q, g_j) = \frac{1}{N} \frac{\exp\{2\pi i (q-j)\} - 1}{\exp\left\{2\pi i \frac{q-j}{N}\right\} - 1} = 0$$

(при  $0 \leq q, j < N$ ,  $q \neq j$  знаменатель отличен от 0). Поскольку  $(g_q, g_q) = 1$ , то в итоге имеем

$$(g_q, g_j) = \delta_q^j \quad \text{при } 0 \leq q, j < N. \quad (6)$$

Умножая (4) скалярно на  $g_j$ , получим равенство

$$A_j = (f, g_j) = \frac{1}{N} \sum_{l=0}^{N-1} f_l \exp\{-2\pi i j x_l\}. \quad (7)$$

Выражение в правой части образует квадратурную сумму для интеграла

$$\int_0^1 f(x) \exp\{-2\pi i j x\} dx,$$

поэтому

$$A_j \rightarrow a_j = \int_0^1 f(x) \exp\{-2\pi i j x\} dx$$

при  $N \rightarrow \infty$  и фиксированном  $j$ .

Покажем, что соотношение

$$f(x) \approx \sum_{j=0}^{N-1} A_j \exp \{2\pi i j x\} \quad (8)$$

в общем случае не имеет места. Пусть  $f(x) = a_0 + a_{-1} \exp \{-2\pi i x\}$ . Из (4) получаем  $A_0 = a_0$ ,  $A_{N-1} = a_{-1}$ , остальные  $A_j = 0$ . Таким образом, правая часть (8) есть

$$a_0 + a_{-1} \exp \{2\pi i (N-1)x\}.$$

Она совпадает с  $f(x)$  в точках  $x_i$ , но далека от нее вне этих точек.

Воспользовавшись утверждением леммы, перепишем (4) в виде

$$f_i = \sum_{-N/2 < q \leq N/2} A_q \exp \{2\pi i q x_i\}. \quad (9)$$

Если  $f(x)$  — достаточно гладкая функция, то величины  $|a_j|$  с ростом  $j$  убывают быстро, поэтому  $A_q \approx a_q$  при малых  $q$ . Кроме того, при гладкой  $f(x)$  величины  $A_q$  и  $a_q$  малы при больших  $q$ . В итоге есть основание надеяться на справедливость приближенного равенства

$$f(x) \approx \sum_{-N/2 < q \leq N/2} A_q \exp \{2\pi i q x\} \quad (10)$$

во всех точках  $x$ .

Напомним, что это приближенное равенство обращается в точное равенство в точках сетки. Способ аппроксимации функции правой частью (10) носит название *тригонометрической интерполяции*. Соотношение (9) называют *конечным* или *дискретным рядом Фурье*, а коэффициенты  $A_q$  — *дискретными коэффициентами Фурье*.

Игнорирование установленного нами факта о равенстве функций  $\exp \{2\pi i q_1 x\}$  и  $\exp \{2\pi i q_2 x\}$  в узлах сетки при  $q_1 - q_2 = kN$  часто является источником получения неверных соотношений.

При решении одной инженерной задачи потребовалось определить первую собственную частоту колебаний конструкции. Было принято решение написать нестационарное уравнение, описывающее процесс колебаний, вывести на печать график и из рассмотрения графика определить частоту. Соответствующее уравнение, которое мы будем условно обозначать  $x'' = F(x)$ , решалось методом конечных разностей. Для контроля над надежностью результата производился повторный расчет с вдвое более мелким шагом. Графики кривых, полученных в результате расчетов, совпали с точностью 10%. Однако из сравнения с экспериментом оказалось, что полученная частота отличается от истинной в десятки раз. Причина недоразумения заключалась в том, что график решения строился с шагом  $1/N$ , существенно большим периода колебаний решения задачи. Решение было близко к функции  $\text{const} \cdot \exp \{2\pi i q t\}$ , где  $q/N$  близко к четному числу  $2k$ . Поэтому как на сетке с шагом  $1/N$ , так и на вдвое более мелкой с шагом  $1/(2N)$  получался график одной и той же функции  $\text{const} \cdot \exp \{2\pi i (q - 2kN) t\}$ .

В другом случае несоответствие со здравым смыслом возникло при расчете диаграммы направленности антенны. Предпринимавшиеся попытки найти ошибку в программе, методе решения или физическом описании задачи не

приводили к положительному результату. Объяснение оказалось тем же: график сильно колеблющейся функции выдавался на очень редкой сетке. На рис. 4.3.1 сплошной кривой изображен реальный график сечения диаграммы направленности, пунктиром — график, который строился путем интерполяции полученных расчетных значений  $x$  и противоречил эксперименту.

Существует соответствие между задачей приближения функций линейными комбинациями многочленов Чебышева и тригонометрическими многочленами. Пусть на отрезке  $[-1, 1]$

функция  $f(x)$  приближается линейными комбинациями

$$\sum_{j=0}^{m-1} a_j T_j(x).$$

Замена переменных  $x = \cos t$  сводит исходную задачу к задаче приближения функции  $f(\cos t)$  линейной комбинацией

$$\sum_{j=0}^{m-1} a_j T_j(\cos t) = \sum_{j=0}^{m-1} a_j \cos(jt).$$

Справедливо равенство

$$\begin{aligned} (f, g)_1 &= \int_{-1}^1 \frac{f(x) \bar{g}(x)}{\sqrt{1-x^2}} dx = \\ &= \int_0^\pi f(\cos \theta) \bar{g}(\cos \theta) d\theta. \end{aligned}$$

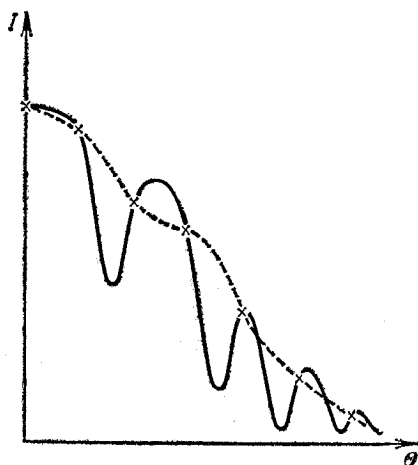


Рис. 4.3.1

Следовательно, задача наилучшего приближения  $f(x)$  в норме, соответствующей скалярному произведению  $(f, g)_1$ , эквивалентна задаче приближения  $f(\cos \theta)$  в норме, соответствующей скалярному произведению  $(f, g)_2 = \int_0^\pi f(\theta) \bar{g}(\theta) d\theta$ . Точно так же суще-

ствует соответствие в случае задач интерполяции и наилучшего приближения в равномерной метрике. Задача интерполирования функции многочленом по узлам  $x_j = \cos\left(\pi \frac{2j-1}{2m}\right)$  — нулям многочлена Чебышева  $T_m(x)$  — после такой замены сводится к задаче интерполирования функции  $f(\cos t)$  при помощи тригонометрического многочлена  $\sum_{j=0}^{m-1} a_j \cos(jt)$  по узлам  $t_j = \pi \frac{2j-1}{2N}$ , образующим равномерную сетку.

#### § 4. Быстрое преобразование Фурье

Осуществление прямого и обратного дискретных преобразований Фурье

$$(f_0, \dots, f_{N-1}) \Leftrightarrow (A_0, \dots, A_{N-1})$$

является составной частью решения многих задач. Непосредственное осуществление этих преобразований по формулам (3.4), (3.7) требует  $O(N^2)$  арифметических операций. Рассмотрим вопрос о возможности сокращения этого числа. Для определенности речь пойдет о вычислении коэффициентов  $A_q$  по заданным значениям функции. Идея построения алгоритмов *быстрого преобразования Фурье* опирается на то, что при составном  $N$  в слагаемых правой части (3.7) можно выделить группы, которые входят в выражения различных коэффициентов  $A_q$ . Вычисляя каждую группу только один раз, можно значительно сократить число операций.

Рассмотрим сначала случай  $N = p_1 p_2$ ;  $p_1, p_2 \neq 1$ . Представим  $q, j$ , лежащие в пределах  $0 \leq q, j < N$ , в виде  $q = q_1 + p_1 q_2$ ,  $j = j_2 + p_2 j_1$ , где  $0 \leq q_1, j_1 < p_1$ ,  $0 \leq q_2, j_2 < p_2$ . Имеем цепочку соотношений

$$\begin{aligned} A_q &= A(q_1, q_2) = \frac{1}{N} \sum_{j=0}^{N-1} f_j \exp \left\{ -2\pi i \frac{qj}{N} \right\} = \\ &= \frac{1}{N} \sum_{j_1=0}^{p_1-1} \sum_{j_2=0}^{p_2-1} f_{j_2+p_2 j_1} \exp \left\{ -2\pi i \frac{(q_1 + p_1 q_2)(j_2 + p_2 j_1)}{p_1 p_2} \right\}. \end{aligned}$$

Из равенства

$$\frac{(q_1 + p_1 q_2)(j_2 + p_2 j_1)}{p_1 p_2} = q_2 j_1 + \frac{q_1 j_2}{N} + \frac{j_1 q_1}{p_1}$$

и предыдущего соотношения получим

$$A(q_1, q_2) = \frac{1}{p_2} \sum_{j_2=0}^{p_2-1} A^{(1)}(q_1, j_2) \exp \left\{ -2\pi i \frac{q_1 j_2}{N} \right\},$$

где

$$A^{(1)}(q_1, j_2) = \frac{1}{p_1} \sum_{j_1=0}^{p_1-1} f_{j_2+p_2 j_1} \exp \left\{ -2\pi i \frac{j_1 q_1}{p_1} \right\}.$$

Непосредственное вычисление всех  $A^{(1)}(q_1, j_2)$  требует  $O(p_1^2 p_2)$  арифметических операций, а последующее вычисление  $A(q_1, q_2)$  — еще  $O(p_1 p_2^2)$  операций. Поэтому при  $p_1, p_2 = O(\sqrt{N})$  общее число операций составит  $O(N^{3/2})$ . Точно так же при  $N = p_1 \dots p_r$  строится алгоритм вычисления совокупности значений  $A_q$ , для которого общее число операций не превосходит  $CN(p_1 + \dots + p_r)$ , здесь  $C$  — постоянная, не зависящая от  $N$ . Выпишем соответствующие расчетные формулы для наиболее употребительного случая  $p_1 = \dots = p_r = 2$ . Представим числа  $q, j$  в виде

$$q = \sum_{k=1}^r q_k 2^{k-1}, \quad j = \sum_{m=1}^r j_{r+1-m} 2^{m-1},$$

где  $q_k, j_m = 0, 1$ . Величину  $q_j 2^{-r}$  можно представить в виде

$$q_j 2^{-r} = \sum_{m=1}^r q_j j_{r+1-m} 2^{m-1-r} = \sum_{m=1}^r \left( \sum_{k=1}^r q_k 2^{k+m-r-2} \right) j_{r+1-m} = \\ = \sum_{m=1}^r \left( \sum_{k=1}^{r-m+1} q_k 2^{k+m-r-2} \right) j_{r+1-m} + s,$$

где  $s$  — целое, равное сумме всех слагаемых вида  $q_k j_{r+1-m} 2^{k+m-r-2}$ , у которых  $k+m-r-2 \geq 0$ . Очевидно, что

$$\exp \left\{ -2\pi i \frac{q_j}{N} \right\} = \exp \left\{ -2\pi i \left( \frac{q_j}{N} - s \right) \right\}, \text{ поэтому}$$

$$A(q_1, \dots, q_r) = A_q = \frac{1}{N} \sum_{j=0}^{N-1} f_j \exp \left\{ -2\pi i \frac{j q}{N} \right\} = \\ = \frac{1}{2} \sum_{j_r=0}^1 \dots \frac{1}{2} \sum_{j_1=0}^1 f_{j_r+2j_{r-1}+\dots+2^{r-1}j_1} \times \\ \times \exp \left\{ -2\pi i \sum_{m=1}^r \left( \sum_{k=1}^{r-m+1} q_k 2^{k+m-r-2} \right) j_{r+1-m} \right\}.$$

После перегруппировки слагаемых имеем

$$A(q_1, \dots, q_r) = \frac{1}{2} \sum_{j_r=0}^1 \exp \left\{ -2\pi i j_r 2^{-r} \sum_{k=1}^r q_k 2^{k-1} \right\} \times \\ \times \left( \frac{1}{2} \sum_{j_{r-1}=0}^1 \exp \left\{ -2\pi i j_{r-1} 2^{1-r} \sum_{k=1}^{r-1} q_k 2^{k-1} \right\} \times \dots \right. \\ \left. \dots \times \left( \frac{1}{2} \sum_{j_1=0}^1 \exp \left\{ -2\pi i j_1 2^{-1} q_1 \right\} f_{j_r+2j_{r-1}+\dots+2^{r-1}j_1} \right) \dots \right).$$

Эго соотношение можно записать в виде последовательности рекуррентных соотношений

$$A^{(m)}(q_1, \dots, q_m; j_{m+1}, \dots, j_r) = \\ = \frac{1}{2} \sum_{j_m=0}^1 \exp \left\{ -2\pi i j_m 2^{-m} \sum_{k=1}^m q_k 2^{k-1} \right\} \times \\ \times A^{(m-1)}(q_1, \dots, q_{m-1}; j_m, \dots, j_r), \quad m = 1, \dots, r,$$

где

$$A^{(0)}(j_1, \dots, j_r) = f_{j_r+2j_{r-1}+\dots+2^{r-1}j_1},$$

$$A^{(r)}(q_1, \dots, q_r) = A(q_1, \dots, q_r).$$

Переход от каждой совокупности  $A^{(m-1)}$  к совокупности  $A^{(m)}$  требует  $O(N)$  арифметических и логических операций; всего таких шагов  $r$ , поэтому общее число операций имеет порядок  $O(Nr) = O(N \log_2 N)$ .

Вычисление при помощи совокупностей  $A^{(m)}$  дает меньшее накопление вычислительной погрешности по сравнению с формулами (3.7). Определенные удобства имеются также при вычислении значений экспонент, входящих в расчетные формулы. При вычислении величин  $A^{(m)}$  используются значения  $\exp\{-2\pi i j 2^{-m}\}$ ,  $j = 0, 1, \dots, 2^m - 1$ . В частности, при  $m = 1$  величина  $\exp\{-\pi i j\}$  принимает значения  $+1$  или  $-1$ . Для вычисления значений  $A^{(m+1)}$  потребуются еще значения  $\exp\{-2\pi i j 2^{-m-1}\}$  при нечетных  $j$ , удовлетворяющих неравенству  $0 \leq j < 2^{m+1}$ . Их можно вычислять через уже вычисленные до этого величины, в частности, при помощи соотношений ( $m \geq 2$ )

$$\exp\{-2\pi i j 2^{-m-1}\} = \frac{\exp\left\{-2\pi i \frac{j+1}{2} 2^{-m}\right\} + \exp\left\{-2\pi i \frac{j-1}{2} 2^{-m}\right\}}{2 \cos(\pi 2^{-m})},$$

где, в свою очередь,

$$\cos(\pi 2^{-m}) = \sqrt{\frac{1}{2}(1 + \cos(\pi 2^{1-m}))}$$

при  $m \geq 1$ .

В ряде случаев удастся еще уменьшить число операций. Один из таких случаев упоминался выше: дана вещественная функция  $f(t) = g(\cos t)$ , известная в точках  $t_l = \pi(2l-1)/(2N)$ ; требуется найти коэффициенты интерполяционного многочлена

$$\sum_{j=0}^{N-1} A_j \cos jt.$$

Другой случай: при четном  $N$  заданы значения функции

$$\sum_{j=1}^{N/2-1} A_j \sin 2\pi jt$$

в точках  $t = l/N$ ,  $0 < l < N/2$ ; нужно определить коэффициенты  $A_j$ .

Задача 1. Найти коэффициенты  $c_j$  произведения двух многочленов

$$\left(\sum_{j=0}^{N-1} a_j x^j\right) \left(\sum_{j=0}^{N-1} b_j x^j\right) = \sum_{j=0}^{2N-2} c_j x^j.$$

Показать, что для их нахождения достаточно  $O(N \log_2 N)$  операций.

## § 5. Наилучшее равномерное приближение

Если норма в линейном нормированном пространстве определяется не через скалярное произведение, то нахождение элемента наилучшего приближения существенно усложняется. Рассмотрим типичную задачу, встречающуюся, в частности, при составлении стандартных программ вычисления функций.

Пусть  $R$  — пространство ограниченных вещественных функций, определенных на отрезке  $[a, b]$  вещественной оси, с нормой

$$\|f\| = \sup_{[a, b]} |f(x)|.$$

Ищется наилучшее приближение вида

$$Q_n(x) = \sum_{j=0}^n a_j x^j.$$

Согласно теореме 1 из § 1 существует элемент наилучшего приближения, т. е. многочлен  $Q_n^0(x)$  такой, что

$$E_n(f) = \|f - Q_n^0\| \leq \|f - Q_n\|$$

при любом многочлене  $Q_n(x)$  степени  $n$ . Такой многочлен  $Q_n^0(x)$  называют *многочленом наилучшего равномерного приближения*. Далее будут установлены необходимые и достаточные условия того, чтобы многочлен являлся многочленом наилучшего приближения для непрерывной функции.

**Теорема Валле—Пуссена.** Пусть существуют  $n + 2$  точки  $x_0 < \dots < x_{n+1}$  отрезка  $[a, b]$  такие, что

$$\text{sign}((f(x_i) - Q_n(x_i))(-1)^i) = \text{const},$$

т. е. при переходе от точки  $x_i$  к следующей точке  $x_{i+1}$  величина  $f(x) - Q_n(x)$  меняет знак. Тогда

$$E_n(f) \geq \mu = \min_{i=0, \dots, n+1} |f(x_i) - Q_n(x_i)|. \quad (1)$$

**Доказательство.** В случае  $\mu = 0$  утверждение теоремы очевидно. Пусть  $\mu > 0$ . Предположим противное, т. е. что для многочлена наилучшего приближения  $Q_n^0(x)$

$$\|Q_n^0 - f\| = E_n(f) < \mu.$$

Имеем

$$\text{sign}(Q_n(x) - Q_n^0(x)) = \text{sign}((Q_n(x) - f(x)) - (Q_n^0(x) - f(x))).$$

В точках  $x_i$  первое слагаемое превосходит по модулю второе, поэтому  $\text{sign}(Q_n(x_i) - Q_n^0(x_i)) = \text{sign}(Q_n(x_i) - f(x_i))$ . Следовательно, многочлен  $Q_n(x) - Q_n^0(x)$  степени  $n$  меняет знак  $n + 1$  раз. Получили противоречие.

**Теорема Чебышева.** Чтобы многочлен  $Q_n(x)$  был многочленом наилучшего приближения непрерывной функции  $f(x)$ , необходимо и достаточно существования на  $[a, b]$  по крайней мере  $m = n + 2$  точек  $x_0 < \dots < x_{n+1}$  таких, что

$$f(x_i) - Q_n(x_i) = \alpha(-1)^i \|f - Q_n\|,$$

где  $i = 0, \dots, n + 1$ ,  $\alpha = 1$  (или  $\alpha = -1$ ) одновременно для всех  $i$ .

Точки  $x_0, \dots, x_{n+1}$ , удовлетворяющие условиям теоремы, принято называть *точками чебышевского альтернанса*.

**Доказательство.** Достаточность. Обозначим через  $L$  величину  $\|f - Q_n\|$ . Применяя (1), имеем  $L = \mu \leq E_n(f)$ , но  $E_n(f) \leq \|f - Q_n\| = L$  вследствие определения величины  $E_n(f)$ .

Следовательно,  $E_n(f) = L$  и данный многочлен является многочленом наилучшего равномерного приближения.

**Необходимость.** Пусть данный многочлен  $Q_n(x)$  является многочленом наилучшего равномерного приближения. Обозначим через  $y_1$  нижнюю грань точек  $x \in [a, b]$ , в которых  $|f(x) - Q_n(x)| = L$ ; из определения  $L$  следует существование такой точки. Вследствие непрерывности  $f(x) - Q_n(x)$  имеем  $|f(y_1) - Q_n(y_1)| = L$ . Для определенности далее рассматриваем случай, когда  $f(y_1) - Q_n(y_1) = +L$ . Обозначим через  $y_2$  нижнюю грань всех точек  $x \in (y_1, b]$ , в которых  $f(x) - Q_n(x) = -L$ , последовательно через  $y_{k+1}$  обозначим нижнюю грань точек  $x \in (y_k, b]$ , в которых  $f(x) - Q_n(x) = (-1)^k L, \dots$  Вследствие непрерывности  $f(x) - Q_n(x)$  при всех  $k$  имеем  $f(y_{k+1}) - Q_n(y_{k+1}) = (-1)^k L$ . Продолжаем этот процесс до значения  $y_m = b$  или  $y_m$  такого, что  $|f(x) - Q_n(x)| < L$  при  $y_m < x \leq b$ . Если  $m \geq n + 2$ , то утверждение теоремы выполнено.

Предположим, что оказалось  $m < n + 2$ . Вследствие непрерывности  $f(x) - Q_n(x)$ , при любом  $k$  ( $1 < k \leq m$ ) можно указать точку  $z_{k-1}$  такую, что  $|f(x) - Q_n(x)| < L$  при  $z_{k-1} \leq x < y_k$ ; положим  $z_0 = a, z_m = b$ . Согласно проведенным выше построениям, на отрезках  $[z_{i-1}, z_i], i = 1, \dots, m$ , имеются точки, в частности точки  $y_i$ , где  $f(x) - Q_n(x) = (-1)^{i-1} L$ , и нет точек, где  $f(x) - Q_n(x) = (-1)^i L$ . Положим

$$v(x) = \prod_{i=1}^{m-1} (z_i - x), \quad Q_n^d(x) = Q_n(x) + dv(x), \quad d > 0,$$

и рассмотрим поведение разности

$$f(x) - Q_n^d(x) = f(x) - Q_n(x) - dv(x)$$

на отрезках  $[z_{i-1}, z_i]$ . Для примера обратимся к отрезку  $[z_0, z_1]$ . На  $[z_0, z_1]$  имеем  $v(x) > 0$ , поэтому

$$f(x) - Q_n^d(x) \leq L - dv(x) < L.$$

В то же время на этом отрезке выполняется неравенство  $f(x) - Q_n(x) > -L$ ; поэтому при достаточно малых  $d$ , например при

$$d < d_1 = \frac{\min_{[z_0, z_1]} |f(x) - Q_n(x) + L|}{\max_{[z_0, z_1]} |v(x)|},$$

на  $[z_0, z_1]$  имеем  $f(x) - Q_n^d(x) > -L$ . Кроме того,

$$|f(z_1) - Q_n^d(z_1)| = |f(z_1) - Q_n(z_1)| < L.$$

Таким образом,  $|f(x) - Q_n^d(x)| < L$  на этом отрезке при достаточно малом  $d$ . После проведения аналогичных рассуждений относительно остальных отрезков  $[z_{i-1}, z_i]$  мы сможем указать малое  $d_0$  такое, что на всех отрезках выполняется неравенство  $|f(x) - Q_n^{d_0}(x)| < L$ . Мы получили противоречие с предположе-



нием, что  $Q_n(x)$  — многочлен наилучшего приближения, а  $m < n + 2$ . Теорема доказана.

*Теорема единственности. Многочлен наилучшего равномерного приближения непрерывной функции единствен.*

*Доказательство.* Предположим, что существуют два многочлена степени  $n$  наилучшего равномерного приближения:

$$Q_n^1(x) \neq Q_n^2(x), \quad \|f - Q_n^1\| = \|f - Q_n^2\| = E_n(f).$$

Отсюда следует, что

$$\left\| f - \frac{Q_n^1 + Q_n^2}{2} \right\| \leq \left\| \frac{f - Q_n^1}{2} \right\| + \left\| \frac{f - Q_n^2}{2} \right\| = E_n(f),$$

т. е. многочлен  $\frac{1}{2} [Q_n^1(x) + Q_n^2(x)]$  также является многочленом наилучшего равномерного приближения. Пусть  $x_0, \dots, x_{n+1}$  — соответствующие этому многочлену точки чебышевского альтернанса; тогда

$$\left| \frac{1}{2} [Q_n^1(x_i) + Q_n^2(x_i)] - f(x_i) \right| = E_n(f), \quad i = 0, \dots, n + 1,$$

или

$$|(Q_n^1(x_i) - f(x_i)) + (Q_n^2(x_i) - f(x_i))| = 2E_n(f).$$

Так как  $|Q_n^k(x_i) - f(x_i)| \leq E_n(f)$ ,  $k = 1, 2$ , то последнее соотношение возможно лишь в том случае, когда

$$Q_n^1(x_i) - f(x_i) = Q_n^2(x_i) - f(x_i).$$

Мы получили, что два различных многочлена  $Q_n^1(x)$  и  $Q_n^2(x)$  степени  $n$  совпадают в точках  $x_0, \dots, x_{n+1}$ , т. е. пришли к противоречию.

**Задача 1.** Функция  $f(x) = \sin 100x$  приближается на отрезке  $[0, \pi]$ . Найти  $Q_{90}(x)$ .

## § 6. Примеры наилучшего равномерного приближения

1. Непрерывная на  $[a, b]$  функция  $f(x)$  приближается многочленом нулевой степени. Пусть

$$\sup_{[a, b]} f(x) = f(x_1) = M, \quad \inf_{[a, b]} f(x) = f(x_2) = m.$$

Многочлен  $Q_0(x) = (M + m)/2$  является многочленом наилучшего приближения, а  $x_1, x_2$  — точками чебышевского альтернанса.

**Задача 1.** Доказать, что наилучшее приближение нулевой степени имеет вид  $Q_0(x) = (M + m)/2$ , если  $f(x)$  не обязательно непрерывна.

**Задача 2.** Построить пример функции и соответствующего многочлена первой степени наилучшего равномерного приближения на  $[a, b]$  так, чтобы среди точек чебышевского альтернанса не было точек  $a$  и  $b$ .

Задача 3. Построить пример функции (естественно, не непрерывной), для которой многочлен наилучшего приближения не удовлетворяет условиям теоремы Чебышева.

2. Выпуклая на отрезке  $[a, b]$  функция  $f(x)$  приближается многочленом первой степени  $Q_1(x) = a_0 + a_1x$ . Вследствие выпуклости  $f(x)$  разность  $f(x) - (a_0 + a_1x)$  может иметь только одну внутреннюю точку экстремума, поэтому точки  $a, b$  являются точками чебышевского альтернанса. Пусть  $d$  — третья точка чебышевского альтернанса. Согласно теореме Чебышева, имеем равенства

$$\begin{aligned} f(a) - (a_0 + a_1a) &= \alpha L, \\ f(d) - (a_0 + a_1d) &= -\alpha L, \\ f(b) - (a_0 + a_1b) &= \alpha L. \end{aligned}$$

Вычитая первое уравнение из третьего, получим  $f(b) - f(a) = a_1(b - a)$ . Отсюда находим  $a_1 = (f(b) - f(a)) / (b - a)$ . Для определения неизвестных  $d, L, a_0$ ,

$a_1$  и  $\alpha = +1$  или  $\alpha = -1$  получено всего три уравнения. Однако следует вспомнить, что точка  $d$  является точкой экстремума разности  $f(x) - (a_0 + a_1x)$ . Если  $f(x)$  — дифференцируемая функция, то для определения  $d$  имеем уравнение  $f'(d) - a_1 = 0$ . Теперь определяем  $a_0$ , например из уравнения, получающегося сложением первого и второго уравнений. Геометрически эта процедура выглядит следующим образом (рис. 4.6.1). Проводим секущую через точки  $(a, f(a))$ ,  $(b, f(b))$ . Для нее тангенс угла наклона равен  $a_1$ . Проводим параллельную ей касательную к кривой  $y = f(x)$ , а потом прямую, равноудаленную от секущей и касательной.

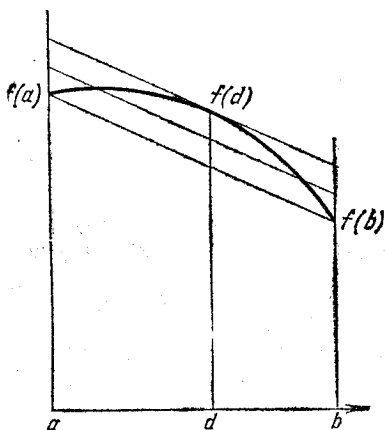


Рис. 4.6.1

Задача 4. Пусть  $f(x) = |x|$ ,  $[a, b] = [-1, 5]$ . Построить многочлен наилучшего равномерного приближения первой степени.

3. Функция  $f(x)$ , удовлетворяющая условию  $f^{(n+1)}(x) \geq 0$ , приближается на  $[a, b]$  многочленом наилучшего приближения степени  $n$ ; требуется оценить величину  $E_n(f)$ . В § 9 гл. II мы имели оценку погрешности интерполяции по узлам, являющимся нулями многочлена Чебышева:

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{\pi(2k-1)}{2(n+1)}\right),$$

а именно

$$|f(x) - P_n(x)| \leq \left(\max_{|a, b|} |f^{(n+1)}(x)|\right) \frac{(b-a)^{n+1}}{2^{2n+1} (n+1)!}. \quad (1)$$

Отсюда следует неравенство

$$E_n(f) \leq \left( \max_{[a, b]} |f^{(n+1)}(x)| \right) \frac{(b-a)^{n+1}}{2^{2n+1} (n+1)!}.$$

Пусть  $Q_n(x)$  — многочлен наилучшего равномерного приближения. Поскольку вследствие теоремы Чебышева разность  $f(x) - Q_n(x)$  меняет знак при переходе от одной точки чебышевского альтернанса к другой, то она обращается в нуль в  $(n+1)$ -й точке  $y_1, \dots, y_{n+1}$ . Поэтому многочлен  $Q_n(x)$  можно рассматривать как интерполяционный с узлами интерполяции  $y_1, \dots, y_{n+1}$ . Согласно (2.3.1) имеем представление для погрешности интерполирования следующего вида:

$$f(x) - Q_n(x) = f^{(n+1)}(\xi) \frac{\omega_{n+1}(x)}{(n+1)!},$$

где  $\omega_{n+1}(x) = (x - y_1) \dots (x - y_{n+1})$ ,  $\xi = \xi(x) \in [a, b]$ . Пусть

$$\max_{[a, b]} |\omega_{n+1}(x)| = |\omega_{n+1}(x_0)|.$$

Имеем

$$\begin{aligned} E_n(f) &= \|f(x) - Q_n(x)\| \geq |f(x_0) - Q_n(x_0)| = \\ &= |f^{(n+1)}(\xi(x_0))| \frac{|\omega_{n+1}(x_0)|}{(n+1)!} \geq \left( \min_{[a, b]} |f^{(n+1)}(x)| \right) \left( \max_{[a, b]} \frac{|\omega_{n+1}(x)|}{(n+1)!} \right). \end{aligned}$$

Согласно (2.8.6) выполняется неравенство

$$\max_{[a, b]} |\omega_{n+1}(x)| \geq (b-a)^{n+1} / 2^{2n+1}.$$

Отсюда следует оценка

$$E_n(f) \geq \left( \min_{[a, b]} |f^{(n+1)}(x)| \right) \frac{(b-a)^{n+1}}{2^{2n+1} (n+1)!}. \quad (2)$$

Таким образом, если  $f^{(n+1)}(x)$  сохраняет знак и меняется не очень сильно, то разность между погрешностью многочлена наилучшего приближения и интерполяционного многочлена по нулям многочленов Чебышева незначительна.

Задача 5. Доказать, что в случае, когда  $f^{(n+1)}(x)$  сохраняет знак на отрезке  $[a, b]$ , чебышевский альтернанс содержит точки  $a$  и  $b$ .

4. Рассмотрим задачу нахождения многочлена наилучшего приближения степени  $n$  в случае, когда

$$f(x) = P_{n+1}(x) = a_0 + \dots + a_{n+1}x^{n+1}, \quad a_{n+1} \neq 0.$$

Тогда  $f^{(n+1)}(x) = a_{n+1}(n+1)!$  и оценки сверху (1) и снизу (2) для  $E_n(f)$  совпадают:

$$E_n(f) = |a_{n+1}| (b-a)^{n+1} 2^{-2n-1}.$$

Таким образом, многочленом наилучшего приближения оказы-

вается интерполяционный многочлен  $Q_n(x)$  с узлами интерполяции

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{\pi(2k-1)}{2(n+1)}\right), \quad k=1, \dots, n+1.$$

Можно получить другое представление этого многочлена наилучшего равномерного приближения, записав его в виде

$$Q_n(x) = P_{n+1}(x) - a_{n+1} T_{n+1}\left(\frac{2x-(a+b)}{b-a}\right) \frac{(b-a)^{n+1}}{2^{2n+1}}. \quad (3)$$

Действительно, выражение в правой части является многочленом степени  $n$ , поскольку коэффициент при  $x^{n+1}$  равен нулю. Точки  $x_i = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{\pi i}{n+1}$ ,  $i=0, \dots, n$ , образуют чебышевский альтернанс.

5. Пусть  $[a, b] = [-1, 1]$  и  $f(x)$  — нечетная функция относительно точки  $\bar{x} = 0$ . Условие нечетности означает, что  $f(x) = -f(-x)$ . Покажем, что здесь многочлен наилучшего приближения любой степени нечетен, т. е. записывается в виде суммы нечетных степеней  $x$ . Действительно, пусть  $Q_n(x)$  — многочлен наилучшего приближения для  $f(x)$ . Имеем  $|f(x) - Q_n(\bar{x})| \leq E_n(f)$ . После замены  $x$  на  $-x$  и умножения выражения под знаком модуля на  $-1$  получим

$$|-f(-x) - (-Q_n(-x))| \leq E_n(f),$$

иначе

$$|f(x) - (-Q_n(-x))| \leq E_n(f).$$

Следовательно, многочлен  $-Q_n(-x)$  также является многочленом наилучшего равномерного приближения. По теореме единственности имеем  $Q_n(x) = -Q_n(-x)$ , что и требовалось доказать.

6. Пусть требуется приблизить функцию  $f(x) = x^3$  на отрезке  $[-1, 1]$  многочленом наилучшего приближения первой степени. Предшествующим результатом можно воспользоваться двояко. Один путь: поскольку искомым многочленом наилучшего приближения будет нечетным, то его достаточно отыскивать среди многочленов вида  $Q_1(x) = \alpha_1 x$ . Второй путь: поскольку многочлен наилучшего приближения второй степени для данной задачи оказывается многочленом первой степени, то исходная задача эквивалентна задаче построения многочлена наилучшего приближения второй степени. Последняя задача наилучшего приближения многочлена многочленом степени, на единицу меньшей, уже нами рассматривалась.

Задача 6. Пусть  $f(x)$  — четная функция относительно середины отрезка приближения  $[-1, 1]$ :  $f(x) = f(-x)$ . Доказать, что многочлен наилучшего приближения  $Q_n(x)$  четен.

Задача 7. Функцию  $f(x) = \exp\{x^2\}$  приблизить на отрезке  $[-1, 1]$  многочленом наилучшего приближения третьей степени,

**Примечание.** Из решения предыдущей задачи следует, что этот многочлен имеет вид  $a_0 + a_2x^2$ . Задача эквивалентна задаче наилучшего приближения функции  $f_1(y) = e^y$  на отрезке  $[0, 1]$  многочленом вида  $a_0 + a_2y$ .

7. Очень часто бывает, что многочлен наилучшего равномерного приближения точно найти не удается. В этих случаях ищется многочлен, близкий к многочлену наилучшего приближения. Рассмотрим примеры такого рода. Для простоты рассматриваем случай приближения на отрезке  $[-1, 1]$ .

Разложим функцию  $f(x)$  в ряд по ортогональной системе многочленов Чебышева:

$$f(x) \sim \sum_{j=0}^{\infty} d_j T_j(x).$$

Отрезок этого ряда

$$\sum_{j=0}^n d_j T_j(x)$$

невысокой степени часто обеспечивает неплохое равномерное приближение.

Иногда бывает затруднительно вычислить явно коэффициенты  $d_j$ , но зато известно разложение Тейлора

$$f(x) = \sum_{j=0}^{\infty} a_j x^j.$$

сходящееся при  $|x| \leq 1$ . Тогда применяют следующий метод (называемый иногда *телескопическим*). Выбирают некоторое  $n$  такое, что погрешность формулы

$$f(x) \approx P_n(x) = \sum_{j=0}^n a_j x^j$$

является достаточно малой. Затем приближают многочлен  $P_n(x)$  многочленом наилучшего равномерного приближения  $P_{n-1}(x)$ . Согласно формуле (3) имеем

$$P_{n-1}(x) = P_n(x) - a_n T_n(x) 2^{1-n}.$$

Поскольку  $|T_n(x)| \leq 1$  на отрезке  $[-1, 1]$ , то

$$|P_{n-1}(x) - P_n(x)| \leq |a_n| 2^{1-n}.$$

Далее приближают многочлен  $P_{n-1}(x)$  многочленом наилучшего равномерного приближения  $P_{n-2}(x)$  и т. д. Понижение продолжается до тех пор, пока погрешность от таких последовательных аппроксимаций остается малой.

Рассматриваемый прием можно описать еще и следующим образом. Разложим многочлен  $P_n(x)$  по многочленам Чебышева:

$$P_n(x) = \sum_{j=0}^n d_j T_j(x).$$

Введем обозначения  $Q_m(x) = \sum_{j=0}^m d_j T_j(x)$  при  $m \leq n$ .

Всякий многочлен  $Q_m(x)$  является многочленом наилучшего равномерного приближения степени  $m$  для многочлена  $Q_{m+1}(x)$ , при этом

$$E_m(Q_{m+1}) = \|Q_{m+1} - Q_m\| = |d_{m+1}|. \quad (4)$$

Это следует, например, из формулы (3) или непосредственно из теоремы Чебышева. Отсюда вытекает, что  $Q_{n-1}(x) = P_{n-1}(x)$ ,  $Q_{n-2}(x) = P_{n-2}(x)$  и т. д.

Таким образом, существо описанного метода заключается в следующем. Исходная функция приближается отрезком ее ряда Тейлора  $P_n(x)$ . Затем

многочлен  $P_n(x)$  раскладывается на многочлены Чебышева и отбрасывается несколько последних членов разложения. Так как

$$|P_n(x) - P_m(x)| \leq \sum_{j=m+1}^n |d_j|,$$

то общая оценка погрешности такова:

$$|f(x) - P_m(x)| \leq \max |f(x) - P_n(x)| + \sum_{j=m+1}^n |d_j|.$$

Рассмотрим задачу приближения функции  $f(x) = \operatorname{arctg} x$  на отрезке  $[-\operatorname{tg}(\pi/8), \operatorname{tg}(\pi/8)]$  с точностью  $0,5 \cdot 10^{-5}$ . Для достижения такой точности при аппроксимации отрезком ряда Тейлора требуется положить

$$\operatorname{arctg} x \approx -\frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \frac{x^{11}}{11}.$$

Приближим полученный многочлен многочленом наилучшего приближения степени, на единицу меньшей. Повторяя данную процедуру три раза, получим многочлен

$$P_5(x) = 0,9999374x - 0,3303433x^3 + 0,1632823x^5,$$

также обеспечивающий требуемую точность. Отметим, что здесь из-за нечетности исходного многочлена показатель степени на каждом шаге уменьшался на 2.

**Задача 8.** Пусть функция  $f(x)$  непрерывна на отрезке  $[-1, 1]$ :

$$Q_{n+1}(x) = a_{n+1}x^{n+1} + \dots + a_0.$$

Доказать, что

$$E_n(f) \geq |a_{n+1}| 2^{-n} - \|f - Q_{n+1}\|.$$

**Задача 9.** Пусть

$$f(x) = \sum_{j=0}^{\infty} d_j T_j(x).$$

Доказать, что

$$E_n(f) \geq |d_{n+1}| - \sum_{j=n+2}^{\infty} |d_j|.$$

Многочлены наилучшего равномерного приближения или близкие к ним используются как важный составной элемент в стандартных программах вычисления элементарных и специальных функций (см. § 9). Часто возникает ситуация, когда функция задается очень сложным явным выражением (например, в виде интеграла  $f(x) = \int_a^b F(x, y) dy$ ), а по ходу решения конкрет-

ной задачи ее значение приходится вычислять в очень многих точках. В этом случае часто полезно вместо непосредственного вычисления значений функции воспользоваться интерполяцией ее значений по таблице или приблизить функцию многочленом. Иногда для этой цели используют многочлены наилучшего приближения в норме  $L_2$  или наилучшего равномерного приближения. Конечно, в каждом конкретном случае полезно посмотреть, оправдают ли себя затраты по построению приближающего многочлена.

## § 7. Итерационный метод построения многочлена наилучшего равномерного приближения

Для нахождения многочленов наилучшего равномерного приближения существует ряд численных итерационных алгоритмов и построенных на их основе стандартных программ. Описываемый ниже алгоритм состоит в получении последовательных приближений к точкам чебышевского альтернанса и соответствующих этим точкам многочленов, сходящихся к многочлену наилучшего равномерного приближения.

В дальнейшем предполагается, что приближаемая непрерывная функция не является многочленом степени  $n$ , т. е.  $E_n(x) \neq 0$ .

Пусть имеются некоторые приближения к точкам чебышевского альтернанса  $x_0^k < \dots < x_{n+1}^k$ . Построим многочлен

$$Q_n^k(x) = \sum_{j=0}^n a_j^k x^j,$$

удовлетворяющий соотношениям

$$f(x_i^k) - Q_n^k(x_i^k) = (-1)^i E^k, \quad (1)$$

где  $E^k$  — некоторая неизвестная величина. Получились уравнения

$$\sum_{j=0}^n a_j^k (x_i^k)^j + (-1)^i E^k = f(x_i^k) \quad (2)$$

относительно  $n+2$  неизвестных  $a_0^k, \dots, a_n^k, E^k$ . Покажем, что определитель этой системы отличен от нуля.

Предположим противное; тогда однородная система уравнений

$$Q_n^k(x_i^k) + (-1)^i E^k = \sum_{j=0}^n a_j^k (x_i^k)^j + (-1)^i E^k = 0$$

имеет ненулевое решение. Если у этого решения  $E^k = 0$ , то хотя бы одно из  $a_j^k$  отлично от нуля, т. е. ненулевой многочлен  $Q_n^k(x)$  степени  $n$  имеет  $n+2$  нуля. Мы пришли к противоречию. Если  $E^k \neq 0$ , то хотя бы одно из  $a_j^k$  также отлично от нуля. Тогда ненулевой многочлен  $Q_n^k(x)$  меняет знак между каждой парой точек  $x_i^k, x_{i+1}^k$ , т. е. имеет не менее  $n+1$  нулей; опять получается противоречие. Таким образом, многочлен  $Q_n^k$  из (2) находится однозначно.

Введем обозначения

$$x_{-1}^k \equiv a, \quad x_{n+2}^k \equiv b, \quad R_n^k(x) = f(x) - Q_n^k(x);$$

зафиксируем  $q_0$  в пределах  $0 < q_0 < 1$ .

За  $x_i^{k+1}$  принимаем точки, принадлежащие отрезкам  $[x_{i-1}^k, x_{i+1}^k]$ ,  $i = 0, \dots, n+1$ , и удовлетворяющие в совокупности условиям:

- а)  $\max_{0 \leq i \leq n+1} |R_n^k(x_i^{k+1})| \geq |E^k| + q_0 (\|R_n^k(x)\| - |E^k|)$ ;
- б)  $R_n^k(x_i^{k+1}) R_n^k(x_{i+1}^{k+1}) \leq 0$  при  $i = 0, \dots, n$ ;
- в)  $|R_n^k(x_i^{k+1})| \geq |E^k|$  при  $i = 0, \dots, n+1$ .

Множество таких точек не пусто и может быть построено, например, следующим образом.

Возьмем произвольную точку  $x^{k+1}$  такую, что

$$|R_n^k(x^{k+1})| \geq E^k + q_0 (\|R_n^k(x)\| - |E^k|).$$

Так как правая часть этого выражения не превосходит  $\|R_n^k(x)\|$ , то такая точка существует. Если  $E^k = 0$ , то требуемым условиям удовлетворяет совокупность, состоящая из точки  $x^{k+1}$  и любых  $n+1$  точек из  $x_0^k, \dots, x_{n+1}^k$ . При  $E^k \neq 0$  все  $|R_n^k(x_i^k)| \neq 0$ . Рассмотрим отдельно случаи:

- а)  $x^{k+1} \in [x_{-1}^k, x_0^k]$ ;
- б)  $x^{k+1} \in [x_{n+1}^k, x_{n+2}^k]$ ;
- в)  $x^{k+1} \in [x_0^k, x_{n+1}^k]$ .

В случае а) полагаем  $x_0^{k+1} = x^{k+1}$ ,

$$x_i^{k+1} = \begin{cases} x_i^k & \text{при } R_n^k(x_0^k) R_n^k(x_0^{k+1}) > 0, \\ x_{i-1}^k & \text{при } R_n^k(x_0^k) R_n^k(x_0^{k+1}) < 0 \end{cases} \text{ для } i = 1, \dots, n+1.$$

В случае б) полагаем  $x_{n+1}^{k+1} = x^{k+1}$ ,

$$x_i^{k+1} = \begin{cases} x_i^k & \text{при } R_n^k(x_{n+1}^k) R_n^k(x_{n+1}^{k+1}) > 0, \\ x_{i+1}^k & \text{при } R_n^k(x_{n+1}^k) R_n^k(x_{n+1}^{k+1}) < 0 \end{cases} \text{ для } i = 0, \dots, n.$$

Обратимся к случаю в). Пусть  $x^{k+1} \in [x_{i_0}^k, x_{i_0+1}^k]$ , где  $0 \leq i_0 \leq n$ . Если  $R_n^k(x^{k+1}) R_n^k(x_{i_0}^k) > 0$ , то положим  $x_{i_0}^{k+1} = x^{k+1}$ ,  $x_i^{k+1} = x_i^k$  при  $i \neq i_0$ . Если же  $R_n^k(x^{k+1}) R_n^k(x_{i_0+1}^k) > 0$ , то положим  $x_{i_0+1}^{k+1} = x^{k+1}$ ,  $x_i^{k+1} = x_i^k$  при  $i \neq i_0 + 1$ .

Во всех случаях полученные совокупности точек  $x_i^k$  удовлетворяют условиям  $\alpha$ ),  $\beta$ ),  $\gamma$ ).

Если не известна дополнительная информация о положении искомым точек чебышевского альтернанса, то в качестве  $x_i^0$  часто берут точки экстремума многочлена Чебышева:

$$x_i^0 = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\pi \frac{n+1-i}{n+1}\right), \quad i = 0, \dots, n+1.$$

**Задача 1.** Пусть  $f(x)$  — многочлен степени  $n+1$ . Показать, что при начальных приближениях  $x_i^0$ , где  $x_i^0$  — точки экстремума многочлена Чебышева, многочлен  $Q_n^0(x)$  окажется искомым многочленом наилучшего равномерного приближения.

Если функция  $f(x)$  близка к четной или к нечетной относительно середины отрезка функции, то может оказаться удобным взять в качестве точки экстремума многочленов Чебышева соответствующей степени. В случае построения многочленов наилучшего приближения для большой серии функций  $f(A, x)$ , зависящих от параметра  $A$ , иногда оказывается удобной следующая процедура, основанная на предположении, что точки чебышевского альтернанса  $x_i$  являются гладкими функциями параметра  $A$ : после нахождения этих точек для некоторых значений  $A_1, \dots, A_m$  начальные приближения при других значениях  $A$  определяются интерполяцией,



Теорема (без доказательства). Пусть

$$\max_{[a, b]} |f^{(l)}(x)| = M_l < \infty \quad \text{при } l = 0, \dots, n+2.$$

Тогда для описанного итерационного процесса выполняется соотношение

$$\|f - Q_n^k\| \leq E_n(f) + \text{const} \cdot q^k, \quad |q| < 1.$$

При этом  $\|Q_n^k - Q_n\| \rightarrow 0$ , где  $Q_n$  — многочлен наилучшего равномерного приближения.

## § 8. О форме записи многочлена

Одна из стандартных программ наилучшего равномерного приближения была отлажена в случае приближения функций многочленами невысокой степени и включена в пакет стандартных программ. Однако при практическом использовании программы для приближения функции многочленами высокой степени в ряде случаев оказывалось, что программа выдает приближение к функции, не обеспечивающее ожидаемой точности, или итерационный процесс, не сходясь, продолжается неограниченно долго, так что приходится прекращать вычисления.

После практического и теоретического анализа возникшей ситуации удалось установить причину происходящего.

Для определенности будем говорить о приближении функции на отрезке  $[-1, 1]$ .

Было установлено, что в случае недостаточно гладкой функции коэффициенты многочлена, приближающего ее с высокой точностью, обязательно будут очень большими. За исключением редко встречающегося случая, когда эти коэффициенты записываются в машине без округления, в этот многочлен вносится погрешность, и он плохо приближает рассматриваемую функцию. Если эти коэффициенты и записываются в машине без округлений, то все равно значения многочлена при  $|x|$ , близком к 1, будут находиться с большой вычислительной погрешностью.

Сформулируем соответствующие строгие утверждения. Предположим, что существует последовательность многочленов

$$P^m(x) = \sum_{j=0}^{n(m)} a_j^m x^j,$$

удовлетворяющих условию

$$|f(x) - P^m(x)| \leq 2^{-m} \quad \text{на } [-1, 1] \quad (1)$$

( $P^m(x)$  — не обязательно многочлены наилучшего равномерного приближения); предположим также, что коэффициенты этих многочленов растут не очень сильно:

$$l_m = \sum_{j=0}^{n(m)} |a_j^m| \leq M 2^{qm}, \quad 0 \leq q. \quad (2)$$

Обозначим через  $G_\varphi$  открытую область, ограниченную дугами, из любой точки которых отрезок  $[-1, 1]$  виден под углом  $\varphi$ . Известна

**Теорема.** Пусть последовательность многочленов  $P^m(x)$  удовлетворяет соотношениям (1), (2). Тогда функция  $f(x)$  может быть аналитически продолжена в открытую область комплексного переменного  $G_{\varphi_0}$ ,  $\varphi_0 = \pi(2q + 1)/(2q + 2)$ .

**Примечание.** Если вместо условия (1) имеем условие  $|f(x) - P^m(x)| \leq 1/m^l$ , то можно показать, что  $\max_{[-1+\delta, 1-\delta]} |f^{(l)}(x)| < \infty$  при любом  $\delta > 0$ .

Какие практические следствия вытекают из этой теоремы? Если число  $a$  записывается в системе с плавающей запятой с  $t$  двоичными разрядами, то его погрешность может оказаться больше  $|a|2^{-t-1}$ ; в худшем случае, когда погрешности коэффициентов  $a_j^n$  одного знака, погрешность значения

$$P^m(1) = \sum_{j=0}^{n(m)} a_j^m,$$

являющаяся следствием этих погрешностей, может превзойти величину

$$\left( \sum_{j=0}^{n(m)} |a_j| \right) 2^{-t-1} = l_m 2^{-t-1}.$$

Чтобы качественно представить характер влияния этой погрешности, будем считать, что число  $m$ , входящее в условия (1), (2), и число разрядов  $t$  велики. Пусть функция  $f(x)$  аналитична в некоторой области  $G_{\varphi_0}$  и не аналитична ни в какой области  $G_\varphi$  при  $\varphi < \varphi_0$ . Пусть  $q_0$  выбирается из соотношения  $q_0 = [2\varphi_0 - \pi]/[2(\pi - \varphi_0)]$ . Возьмем произвольное  $q < q_0$ . Если предположить, что для всех  $m$  выполняется неравенство  $l_m \leq 2^{qm}$ , то, согласно выше сформулированной теореме,  $f(z)$  будет аналитична в соответствующей области  $G_\varphi$  при  $\varphi < \varphi_0$  и получается противоречие с предположением о свойствах функции  $f(z)$ . Таким образом, будут встречаться сколь угодно большие  $m$ , для которых  $l_m \geq 2^{qm}$ . Следовательно, при неблагоприятном стечении обстоятельств погрешность от округления значений  $a_j^m$  может оказаться больше чем  $2^{qm}2^{-t-1}$ . В то же время нельзя рассчитывать на оценку погрешности, лучшую, чем (1). Для суммарной погрешности, получающейся от замены  $f(x)$  на  $P^m(x)$  и от погрешностей в коэффициентах многочлена, мы не можем предложить оценки лучшей, чем  $\varepsilon(m) = 2^{qm}2^{-t-1} + 2^{-m}$ . Уравнение для точки экстремума  $\bar{m}$  функции  $\varepsilon(m)$  имеет вид

$$\varepsilon'_m(\bar{m}) = q \ln 2 \cdot 2^{q\bar{m}}2^{-t-1} - \ln 2 \cdot 2^{-\bar{m}} = 0.$$

Отсюда следует, что  $2^{\bar{m}} = \left(\frac{2}{q}\right)^{1/(q+1)} 2^{t/(q+1)}$ . Подставив это значение в выражение для  $\varepsilon(\bar{m})$ , получим

$$\begin{aligned} \varepsilon(\bar{m}) &= \left(\frac{2}{q}\right)^{q/(q+1)} 2^{-t/(q+1)} 2^{-1} + \left(\frac{2}{q}\right)^{-1/(q+1)} 2^{-t/(q+1)} = \\ &= \frac{q+1}{2} \left(\frac{2}{q}\right)^{q/(q+1)} 2^{-t/(q+1)}. \end{aligned}$$

Таким образом, при вычислениях на машине с  $t$  разрядами нельзя надеяться на получение более точных приближений к значениям функции, чем с  $t/(q+1)$  разрядами. Например, для функции  $f(x) = (1 + 25x^2)^{-1}$  нижней гранью  $\varphi$ , при которой  $f(z)$  аналитична в области  $G_\varphi$ , будет  $\varphi^0 = \operatorname{arctg} 0,2$ ; соответствующее значение  $q+1 \approx 3,2$ . Таким образом, на машине с 60 двоичными разрядами мы в лучшем случае можем рассчитывать на получение приближений к значениям функции примерно с 20 верными двоичными разрядами.

Если производная  $f^{(l)}(x)$  не существует в некоторой внутренней точке отрезка при не очень большом  $l$ , то согласно примечанию к теореме имеет место еще большая потеря точности результата.

Можно привести следующий довод об исключительности подобной обстановки. Обычно приходится приближать целые функции или такие, особенности которых лежат далеко за единичным кругом. Тогда из проведенных выше рассмотрений столь удручающие выводы не следуют. Формально это верно, однако такого рода функции часто, являясь целыми аналитическими функциями, очень быстро возрастают при увеличении мнимой части у аргумента. Хотя для таких функций величины  $l_n$  остаются равномерно ограниченными по  $n$ , они могут принимать столь большие значения, что величина  $|l_n|2^{-t-1}$  окажется существенно больше допустимой погрешности. Примером подобных

функций является функция  $\varphi(x) = \int_0^1 \exp\{i\lambda t^2\} dt$  для больших  $\lambda$ .

Эти обстоятельства побуждают к отысканию иных простых форм записи и вычисления значений многочленов.

Обратимся к форме записи многочленов наилучшего приближения в виде суммы значений многочленов Чебышева

$$P_n(x) = \sum_{j=0}^n d_j T_j(x). \quad (3)$$

Имеем

$$\frac{2}{\pi} \int_{-1}^1 \frac{T_n(x) T_m(x)}{\sqrt{1-x^2}} dx = \begin{cases} 2 & \text{при } n=m=0, \\ \delta_n^m & \text{при } n^2+m^2 > 0, \end{cases}$$

поэтому

$$\begin{aligned} \frac{2}{\pi} \int_{-1}^1 \frac{P_n^2(x)}{\sqrt{1-x^2}} dx &= \frac{2}{\pi} \sum_{j=0}^n |d_j|^2 \int_{-1}^1 \frac{T_j^2(x)}{\sqrt{1-x^2}} dx = \\ &= 2d_0^2 + \sum_{j=1}^n d_j^2 \geq \sum_{j=0}^n d_j^2. \end{aligned}$$

Воспользуемся неравенством Коши — Буняковского

$$\sum_{j=0}^n |d_j| = \sum_{j=0}^n 1 \cdot |d_j| \leq \sqrt{\left(\sum_{j=0}^n 1\right) \cdot \left(\sum_{j=0}^n d_j^2\right)} = \sqrt{(n+1) \sum_{j=0}^n d_j^2}.$$

Отсюда заключаем, что

$$\sum_{j=0}^n |d_j| \leq \sqrt{\frac{2}{\pi} (n+1) \int_{-1}^1 \frac{P_n^2(x)}{\sqrt{1-x^2}} dx}.$$

Если  $\|P_n - f\|_C \rightarrow 0$ , то

$$\frac{2}{\pi} \int_{-1}^1 \frac{P_n^2(x)}{\sqrt{1-x^2}} dx \rightarrow \frac{2}{\pi} \int_{-1}^1 \frac{f^2(x)}{\sqrt{1-x^2}} dx.$$

Таким образом, в этом случае величина  $\sum_{j=0}^n |d_j|$  растет не быстрее чем  $\sqrt{n}$ . Поскольку  $|T_n(x)| \leq 1$ , то погрешность в значении многочлена, являющаяся следствием погрешностей в  $d_j$ , не превзойдет величины порядка  $\left(\sum_{j=0}^n |d_j|\right) 2^{-t} = O(\sqrt{n} 2^{-t})$ . Такая оценка погрешности оказывается приемлемой при реальных вычислениях.

Конечно, не обязательно представлять многочлены, приближающие функцию, в виде линейной комбинации (9) многочленов Чебышева. В зависимости от конкретной обстановки иногда удобнее представлять многочлен как линейную комбинацию многочленов какой-либо другой ортогональной системы, например многочленов Лежандра. (Мы рассмотрели систему многочленов Чебышева вследствие наибольшей простоты рекуррентных соотношений для вычисления их значений.)

Дадим рекомендации по вычислению значений многочлена  $P_n(x)$  на отрезке  $[-1, 1]$  безотносительно к вопросу о наилучшем приближении функций.

Пусть в обычной форме этот многочлен записывается в виде  $\sum_{j=0}^n a_j x^j$  и погрешность порядка  $\left(\sum_{j=0}^n |a_j|\right) 2^{-t}$  является допустимой. Тогда имеет смысл воспользоваться схемой Горнера

$$P_n(x) = a_0 + x(a_1 + x(\dots + x(a_{n-1} + xa_n) \dots)).$$

В противном случае этот многочлен целесообразно представлять в виде линейной комбинации ортогональных многочленов, например многочленов Чебышева:

$$P_n(x) = \sum_{j=0}^n d_j T_j(x).$$

Для уменьшения числа действий при вычислении значений  $P_n(x)$  целесообразно поступить следующим образом. Полагаем

$$D_0 = d_0 - d_0/2, \quad D_j = (d_j - d_{j+2})/2, \quad j = 1, \dots, n-2, \\ D_{n-1} = d_{n-1}/2, \quad v_n = d_n/2.$$

При необходимости вычислить  $P_n(x)$  находим  $y = 2x$ ,  $v_{n-1} = yv_n + D_{n-1}$ , затем последовательно находим  $v_{n-2}, \dots, v_0 = P_n(x)$  по рекуррентной формуле  $v_k = yv_{k+1} - v_{k+2} + D_k$ . Известно, что вычислительная погрешность такого алгоритма есть

$$O\left(n\left(\min\left(n, \frac{1}{\sqrt{1-x^2}}\right)\right)\left(\max_{[-1, 1]}(P_n(x))\right)2^{-t}\right).$$

З а д а ч а 2. Проверить, что действительно  $v_0 = P_n(x)$ .

### § 9. О способах вычисления элементарных функций

Многочлены наилучшего приближения часто используются для приближения функций, возникающих из эксперимента или в результате сложных вычислений в случае, когда не имеется никакой дополнительной информации об их свойствах. Если такая информация имеется, то иногда целесообразнее использовать другой вид приближающей функции. Например, в случае элементарных функций имеется большая информация об их аналитических свойствах, позволяющая резко сократить объем вычислений.

Рассмотрим одну из стандартных программ вычисления  $f(x) = \ln x$ . Аргумент  $x$  представляется в виде  $x = 2^p X$ , где  $0,5 \leq X \leq 1$ , т.е.  $p$  — двоичный порядок числа. Далее  $\ln x$  записываем в виде  $\ln x = p \ln 2 + \ln X$ . Для вычисления  $\ln X$  используется ряд

$$\ln X = 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left(\frac{X-1}{X+1}\right)^{2k+1}.$$

При  $0,5 \leq X \leq 1$  его члены убывают быстрее чем  $3^{-k}$ , в то же время для простейшего ряда

$$\ln X = \sum_{k=0}^{\infty} \frac{(-1)^{k-1}}{k} (X-1)^k$$

имеет место убывание порядка  $2^{-k}/k$ . Для ускорения сходимости полагают  $X = y/\lambda$  и вычисляют  $\ln y$  с помощью ряда

$$\ln y = 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left(\frac{y-1}{y+1}\right)^{2k+1}.$$

Множитель  $\lambda$  подбирается из условия, чтобы величина  $u = \max_{0,5 \leq x \leq 1} |(y-1)/(y+1)|$  была наименьшей. Это дает  $\lambda = \sqrt{2}$  и  $u < 0,172$ , а расчетная формула приобретает вид

$$\ln x = \left(p - \frac{1}{2}\right) \ln 2 + 2v \sum_{k=0}^{\infty} \frac{1}{2k+1} v^{2k}, \quad v = \frac{\sqrt{2} X - 1}{\sqrt{2} X + 1}.$$

Для уменьшения числа учитываемых слагаемых в выражении  $\sum_{k=0}^{\infty} \frac{1}{2k+1} v^{2k}$

это выражение заменяется многочленом наилучшего приближения или близким к нему многочленом на соответствующем участке изменения  $v$ . В рассматриваемом случае формула

$$\ln x \approx (p - 0,5) \ln 2 + v (2,000000815 + 0,666445069v^2 + 0,4150542540v^4)$$

имеет погрешность, не большую чем  $3 \cdot 10^{-8}$ .

В ряде ЭВМ в основу стандартных программ вычисления значений элементарных функций положены представления этих функций в виде цепных дробей или аппроксимация дробно-рациональными функциями. Например, функцию  $\operatorname{tg} x$  можно вычислять, пользуясь представлением в виде цепной дроби

$$\operatorname{tg} x = \frac{x}{1 - \frac{x^2}{3 - \frac{x^2}{5 - \dots}}}$$

Другой распространенный способ представления функций — это дробно-рациональное приближение вида

$$[P_n(x) + Q_m(x)]/[P_n(x) - Q_m(x)].$$

Например, приближение

$$e^x \approx \frac{12(x^2 + 10) + x(x^2 + 60)}{12(x^2 + 10) - x(x^2 + 60)}$$

обеспечивает точность  $9 \cdot 10^{-9}$  при  $|x| \leq 0,5 \ln 2$ .

Трудно представить все многообразие приемов, элементы которых используются при создании стандартных программ вычисления значений элементарных функций. В частности, используемые методы сильно зависят от разрядности ЭВМ, предъявляемых требований точности, длительности выполнения арифметических операций.

Рассмотрим один распространенный прием, применяемый в программах вычисления элементарных функций. Этот прием выделен в основном с целью еще раз обсудить вопрос о вычислительной погрешности.

Простейшие элементарные функции удовлетворяют рекуррентным соотношениям вида

$$f(x) = F(f(G(x))), \quad (1)$$

где функции  $F$  и  $G$  легко вычисляются,  $0 \leq G'(x)$ ,  $0 \leq G(x) \leq qx$  при  $0 \leq x \leq x_0$ ,  $q < 1$ . Воспользовавшись  $n$  раз этой записью, получим

$$f(x) = \underbrace{F(F(\dots F)}_{n \text{ раз}}(f(\underbrace{G(G(\dots (G(x))}_{n \text{ раз}})\dots))). \quad (2)$$

Положим

$$G_n(x) = \underbrace{G(G(\dots (G(x))}_{n \text{ раз}})\dots),$$

$$F_n(x) = \underbrace{F(F(\dots (F(x))}_{n \text{ раз}})\dots).$$

При помощи записи (2) задача приближения функции  $f(x)$  на отрезке  $[0, x_0]$  сводится к задаче приближения на меньшем отрезке  $[0, G_n(x_0)]$ , где  $G_n(x_0) \leq q^n x_0$ . За счет малой длины отрезка высокая точность приближения достигается при значительно меньшей степени аппроксимирующего многочлена.

Укажем ряд соотношений вида (1):

$$e^x = \left( \exp \frac{x}{2} \right)^2, \quad \cos x = 2 \left( \cos \frac{x}{2} \right)^2 - 1,$$

$$\sin x = 2 \left( \sin \frac{x}{2} \right) \sqrt{1 - \left( \sin \frac{x}{2} \right)^2}, \quad \operatorname{tg} x = \frac{2 \operatorname{tg} \frac{x}{2}}{1 - \left( \operatorname{tg} \frac{x}{2} \right)^2},$$

$$\operatorname{ctg} x = \frac{1}{2} \left( \operatorname{ctg} \frac{x}{2} - \frac{1}{\operatorname{ctg} \frac{x}{2}} \right), \quad \operatorname{arctg} x = 2 \operatorname{arctg} \frac{x}{1 + \sqrt{1 + x^2}},$$

$$\operatorname{arcsin} x = 2 \operatorname{arcsin} \left( x / \sqrt{2(1 + \sqrt{1 - x^2})} \right).$$

Наиболее привлекательной по своей простоте выглядит рекуррентная формула для  $\operatorname{ctg} x$ , которая используется также при вычислении значений функций  $\sin x$  и  $\cos x$ : после определения  $u = \operatorname{ctg}(x/2)$  можно вычислить

$$\sin x = \frac{2}{u + 1/u}, \quad \cos x = \frac{u - 1/u}{u + 1/u}, \quad \operatorname{tg} x = \frac{2}{u - 1/u}.$$

Некоторые из этих формул довольно сложны, и их применение не всегда оправдано. Из приведенных построений вытекает следующая схема вычисления значений функций. Находим некоторую аппроксимацию функции  $f(x)$  на  $[0, G_n(x_0)]$ , например, многочленом или отношением двух многочленов:  $f(x) \approx h(x)$ . Последовательно вычисляем

$$G_1(x) = G(x), \quad G_2(x) = G_2(G_1(x)), \quad \dots, \quad G_n(x) = G(G_{n-1}(x)),$$

$$f^* \approx h(G_n(x)),$$

$$F_1(f^*) = F(f^*), \quad \dots, \quad F_n(f^*) = F(F_{n-1}(f^*))$$

и полагаем  $f(x) \approx F_n(x^*)$ .

На примере функции  $e^x$  обратим внимание на опасность нежелательного роста вычислительной погрешности при вычислении значения  $F_n(f^*)$ . Используя равенство  $e^x = 2^{x/\ln 2}$ , где

$$p = \left[ \frac{x}{\ln 2} + \frac{1}{2} \right], \quad X = \left( \frac{x}{\ln 2} - p \right) \ln 2$$

([ ] — знак целой части), сведем задачу к вычислению функции  $e^X$  на отрезке  $\left[ -\frac{1}{2} \ln 2, \frac{1}{2} \ln 2 \right]$ . Имеем

$$e^X = \left( \exp \{X/2^n\} \right)^{2^n} = \left( \dots \left( \exp \{X/2^n\} \right)^2 \dots \right)^2.$$

Как бы ни вычислялись значения  $\exp \{X/2^n\}$ , они будут округлены и их погрешность может превзойти число  $2^{-t-1}$  ( $t$  — разрядность чисел в ЭВМ). Вместо точного значения  $(\exp \{X/2^n\})^{2^n}$  мы получаем приближенное значение  $(\exp \{X/2^n\} + \varepsilon)^{2^n}$ . Согласно формуле Лагранжа имеем  $y_1^N - y_2^N = N \bar{y}^{N-1} (y_1 - y_2)$ , где  $\bar{y}$  лежит между  $y_1$  и  $y_2$ . Поэтому в рассматриваемом случае погрешность равна

$$2^n \varepsilon (\exp \{X/2^n\} + \bar{\varepsilon})^{2^n-1}, \quad \text{где } |\bar{\varepsilon}| \leq \varepsilon.$$

Последний множитель не менее чем 1, и, таким образом, погрешность может оказаться более чем  $2^{n-1-t}$ . К стандартным программам обычно предъявляется требование точности значений функции порядка  $2^{-t}$ , поэтому при больших  $n$  такая величина погрешности является неприемлемой.

В рассматриваемом примере уже с первого взгляда видны резервы повышения точности. Число  $u_k = (\exp\{X/2^n\})^{2^k}$  в двоичной системе записывается в виде  $1, \underbrace{0 \dots 0}_{n-k \text{ раз}} \dots$ . Таким образом, из  $t$  разрядов  $n - k + 1$  трагятся на

$n-k$  раз

запоминание информации, которая и так известна. Поэтому есть надежда, заменив эту информацию на существенную, повысить точность вычислений. Чтобы избавиться от запоминания этих известных двоичных знаков, введем в рассмотрение функцию  $v(X) = e^X - 1$ . Соотношение (1) для этой функции имеет вид  $v(X) = v(X/2)(2 + v(X/2))$ . Значения  $v(X/2^n)$  не превосходят  $2^{1-n}$ . Если они будут известны с погрешностью  $O(2^{-n-t})$ , то погрешность значения  $v(X)$ , а следовательно, и  $e^X = 1 + v(X)$  окажется величиной порядка  $O(n2^{-t})$ . Мы предоставляем читателю проверку справедливости этого утверждения и построение алгоритма вычисления значения  $v(X/2^n)$  с погрешностью  $O(2^{-n-t})$  на машине с плавающей запятой.

Подобная потеря точности возникает всегда, если одно из чисел  $f(0)$  (как в случае  $e^X$ ) или  $\lim_{n \rightarrow \infty} G_n(x)$  (как в случае  $\ln x$ ) конечно и отлично от нуля.

**Задача 1.** Провести подобные построения в случае  $f(x) = \arccos x, \cos x, \ln x$ .

**Задача 2.** Убедиться, что рекуррентные соотношения для вычисления  $\sin x, \operatorname{tg} x, \operatorname{ctg} x, \operatorname{arctg} x, \operatorname{arcsin} x$  в машине с плавающей запятой не нуждаются в подобных преобразованиях.

**Задача 3.** Предположим, что вычисления производятся без округлений и на отрезке  $[0, G_n(x_0)]$  функция  $f(x)$  приближается отрезком ряда Тейлора из  $q$  членов. Доказать, что при малых  $\epsilon$  можно выбрать взаимно согласованные  $n$  и  $q$  порядка  $O(\sqrt{\ln(1/\epsilon)})$ , при которых погрешность приближений к значениям  $f(x)$  на отрезке  $[0, x_0]$  не превосходит  $\epsilon$ .

**Задача 4.** Пусть значения элементарных функций вычисляются в машине с плавающей запятой с  $t$  двоичными разрядами, где  $t$  — большое число. Предположим, что порядки чисел, употребляемых в машине, не ограничены (таких машин не бывает, но ограничения на порядок чисел в современных ЭВМ таковы, что конечность максимального порядка часто можно пренебречь). Доказать, что в случае аппроксимации  $f(x)$  на  $[0, G_n(x_0)]$  отрезком ряда Тейлора длины  $q$  при подходящем выборе  $n$  и  $q$  (порядка  $\sqrt{t}$ ) и подходящей форме записи соотношений (1) можно указать алгоритмы вычисления указанных выше элементарных функций со следующими характеристиками: объем вычислений  $O(\sqrt{t})$ , погрешность приближенных значений  $O(\sqrt{t} 2^{-t})$ .

## § 10. О скорости приближения функций различных классов

Приведем сведения о скорости приближения функций многочленами и дробно-рациональными функциями

$$R_n(x) = \frac{a_0 + \dots + a_n x^n}{b_0 + \dots + b_n x^n}$$

в зависимости от аналитических и дифференциальных свойств приближаемых функций.

Речь пойдет о приближении функций, заданных на отрезке  $[-1, 1]$ , многочленами и дробно-линейными функциями в норме  $\|f\|_1 = \sup_{[-1, 1]} |f(x)|$  и приближении периодических функций с периодом  $2\pi$  тригонометрическими многочленами и тригонометрическими дробями в норме  $\|f\|_2 = \sup_{[0, 2\pi]} |f(x)|$ .

В теоретических исследованиях обычно рассматривают следующие типичные классы функций:  $C_r(M, [-1, 1])$ ,  $r$  целое, — класс функций, у которых на отрезке  $[-1, 1]$  производные  $f^{(l)}$ ,  $l = 0, \dots, r-1$ , непрерывны, производная  $f^{(r)}$  кусочно-непрерывна и  $|f^{(l)}| \leq M$  при  $l = 0, \dots, r$ ;  $H_r(M, 2\pi)$  — класс функций, имеющих период  $2\pi$  и удовлетворяющих тем же условиям на



производные;  $A(M, G)$  — класс функций, допускающих аналитическое продолжение с отрезка  $[-1, 1]$  в открытую односвязную область  $G$  комплексного переменного и удовлетворяющих в  $G$  условию  $|f(z)| \leq M$ ;  $A(M, h, 2\pi)$  — класс функций, периодических с периодом  $2\pi$ , допускающих аналитическое продолжение в область  $|\operatorname{Im} z| < h$  и удовлетворяющих там условию  $|f(z)| \leq M$ .

Через  $\mathcal{E}_H$  будем обозначать эллипсы с суммой полуосей  $H > 1$  и с фокусами в точках отрезка вещественной оси  $[-1, 1]$ . Положим  $\alpha(G) = \sup_{\mathcal{E}_H \in G} H$ ,

т.е.  $\alpha(G)$  — это сумма полуосей наибольшего из эллипсов  $\mathcal{E}_H$ , принадлежащих замкнутой области (рис. 4.10.1).

Существует одно и только одно конформное отображение с неподвижными точками  $-1$  и  $1$ , при котором область  $G$  отображается на внутренность эллипса из семейства  $\mathcal{E}_H$ . Значение суммы полуосей  $H$ , соответствующее этому эллипсу, обозначаем через  $\beta(G)$ .

Введем обозначения:

$t_n(f)$  — величина погрешности наилучшего равномерного приближения тригонометрическими многочленами

$$\begin{aligned} T_n(x) &= a_0 + \\ &= \sum_{j=1}^n (a_j \cos jx + b_j \sin jx), \\ t_n &= \inf_{a_j, b_j} \|f - T_n\|; \end{aligned}$$

$r_n(f)$  — величина погрешности наилучшего равномерного приближения рациональными функциями

$$\begin{aligned} R_n(x) &= \frac{a_0 + \dots + a_n x^n}{1 + b_1 x + \dots + b_n x^n}, \\ r_n(f) &= \inf_{a_j, b_j} \|f - R_n\|; \end{aligned} \quad (1)$$

$\epsilon_n(f) = \|f - L_n^f\|$ , где  $L_n^f(x)$  — интерполяционный многочлен степени  $n$  с чебышевскими узлами интерполяции

$$\cos\left(\pi \frac{2m+1}{2n+2}\right), \quad m=0, \dots, n;$$

$\tau_n(f) = \|f - T_n^f\|_2$ , где  $T_n^f$  — интерполяционный тригонометрический многочлен степени  $n$  с узлами интерполяции

$$x_j = 2\pi j / (2n+1), \quad j=0, \dots, 2n;$$

$\rho_n(f)$  — величина погрешности наилучшего равномерного приближения тригонометрическими дробями вида

$$R_n^T(x) = \frac{a_0 + \sum_{j=1}^n (a_j \cos jx + b_j \sin jx)}{1 + \sum_{j=1}^n (c_j \cos jx + d_j \sin jx)}.$$

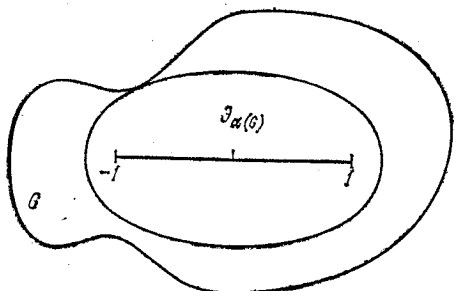


Рис. 4.10.1

Справедливы следующие оценки. Для функций классов  $C_r(M, [-1, 1])$ :

$$E_n(f) = O\left(\frac{1}{n^r}\right), \quad \varepsilon_n(f) = O\left(\frac{\ln n}{n^r}\right), \quad r_n(f) = O\left(\frac{1}{n^r}\right);$$

для функций классов  $H_r(M, 2\pi)$ :

$$t_n(f) = O\left(\frac{1}{n^r}\right), \quad \tau_n(f) = O\left(\frac{\ln n}{n^r}\right), \quad \rho_n(f) = O\left(\frac{1}{n^r}\right);$$

для функций классов  $A(M, G)$ :

$$E_n(f) = O(M(\alpha(G))^{-n}), \quad \varepsilon_n(f) = O(M(\alpha(G))^{-n}), \quad r_n(f) = O(M(\beta(G))^{-n});$$

для функций классов  $A(M, h, 2\pi)$ :

$$t_n(f) = O(Me^{-2hn}), \quad \tau_n(f) = O(Me^{-2hn}), \quad \rho_n(f) = O(Me^{-2hn}).$$

Если область аналитичности — эллипс  $\mathcal{E}_h$  или в периодическом случае полоса  $|\operatorname{Im} z| \leq h$ , то при наилучшем приближении рациональной дробью вида (1) требуемое количество арифметических операций примерно вдвое больше по сравнению с приближениями многочленами, а выигрыша в порядке оценки нет. Если область аналитичности функции имеет существенно другой вид, то приближение рациональными или тригонометрическими рациональными функциями может дать выигрыш.

Особенно большой выигрыш получается в случае, когда  $\alpha(G) = 1$ . А именно, пусть  $f(x)$  — аналитичная функция во всех точках отрезка  $[-1, 1]$ , кроме конечного числа точек, и удовлетворяет условию Гельдера с некоторым  $\lambda > 0$ :

$$|f(x_1) - f(x_2)| \leq C|x_1 - x_2|^\lambda.$$

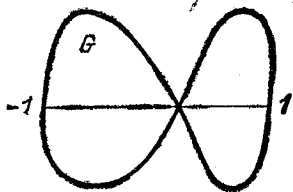


Рис. 4.10.2

На рис. 4.10.2 приведен пример области аналитичности, соответствующей такому случаю. Тогда про величину  $E_n(f)$  можно лишь сказать, что

$$E_n(f) = O(n^{-\sigma}), \quad \sigma = \sigma(\lambda) > 0.$$

В то же время

$$r_n(f) = O(e^{-c\sqrt{n}}), \quad c > 0.$$

Наилучшее приближение рациональными функциями  $R_n(x)$  практически получить трудно, так как  $R_n(x)$  нелинейно зависит от большого числа параметров. Поэтому иногда целесообразно применять некоторые комбинированные способы, где число параметров, от которых функция зависит нелинейно, существенно меньше, например использовать приближения видов:

$P_n(P_m(x))$ , где  $P_n, P_m$  — многочлены;

$P_n(x)/(Q_m(x))^l$ , где  $P_n$  и  $Q_m$  — многочлены,  $n = ml$ ;

$P_n(R_m(x))$ , где  $P_n$  — многочлен,  $R_m$  — рациональная дробь. Погрешность наилучшего приближения такими функциями будем обозначать соответственно  $E_{nm}(f)$ ,  $r_{nm}(f)$ ,  $e_{nm}(f)$ .

Рассмотрим случай, когда граница области  $G$  — аналитическая кривая. Тогда для функций классов  $A(M, G)$  при  $\alpha(G) > 1$  справедливы оценки:

$$e_{nm}(f) \leq E_{nm}(f) = O(M\beta(G)^{-n}) \quad \text{при } m \geq \nu(G) \ln n;$$

$$r_{nm}(f) = O(M\beta(G)^{-n}) \quad \text{при } m \geq \eta(G) \ln n.$$

Если  $\alpha(G) = 1$ , но не все точки интервала  $(-1, 1)$  являются точками аналитичности (рис. 4.10.3), то при  $m \geq \chi(G) \ln n$  справедлива оценка

$$r_{nm}(f) = O(M\delta(G)^{-\sqrt{n}}), \quad \delta(G) > 1.$$

Приведем для сведения хорошо зарекомендовавший себя на практике способ построения хотя и не наилучшего, но достаточно хорошего рационального приближения.

Пусть  $x_1, \dots, x_N$  — некоторая система точек,

$$(f, g) = \sum_{k=1}^N p_k f(x_k) g(x_k), \quad p_k > 0.$$

Пусть  $\varphi_1, \dots, \varphi_n$  — некоторая система функций, удовлетворяющих условию  $(\varphi_i, \varphi_j) = \delta_{ij}$ . Приближение ищется в виде

$$\left( \sum_{j=1}^n a_j \varphi_j(x) \right) \left( \sum_{j=1}^m b_j \varphi_j(x) \right)^{-1};$$

коэффициенты  $a_j, b_j$  определяются из системы линейных уравнений

$$\left( \left( f(x) \sum_{j=1}^m b_j \varphi_j(x) - \sum_{j=1}^n a_j \varphi_j(x) \right), \varphi_q(x) \right) = 0, \\ q = 1, \dots, n + m - 1,$$

или, что то же самое,

$$\sum_{j=1}^m b_j (f \varphi_j, \varphi_q) - \sum_{j=1}^n a_j \delta_j^q = 0.$$

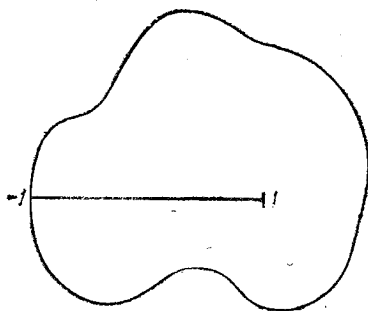


Рис. 4.10.3

Приведем наиболее используемые на практике случаи.

При приближении многочленами на отрезке  $[-1, 1]$  берутся

$$x_j = -1 + 2 \frac{j-1}{N-1},$$

$$p_j = \frac{\Gamma(N+1+\alpha-j) \Gamma(\alpha+j)}{\Gamma(j) \Gamma(N+1-j)}, \quad \alpha > 1,$$

$\varphi_k(x) = h_n^{(\alpha, \alpha)}(x, N)$  — так называемые многочлены Хана. Чаще всего  $\alpha = 0$  и  $\alpha = -1/2$ .

При приближении периодической функции на  $[0, 2\pi]$  берутся  $N$  нечетное,

$$x_j = 2\pi j/N, \quad p_j = 1,$$

$$\varphi_{2k+1}(x) = \cos kx, \quad \varphi_{2k}(x) = \sin kx.$$

## § 11. Интерполяция и приближение сплайнами

Способ приближения сплайнами интересен, кроме всего прочего, отношением к нему специалистов. Одни считают его универсальным методом решения проблем, стоящих перед численным анализом, и ищут применения ему в самых различных направлениях. Другие рассматривают его как очередную дань переменчивой математической моде. По-видимому, истина находится где-то посередине.

Для определенности будем говорить о приближении функции  $f(x)$  на отрезке  $[0, 1]$ . Разобьем отрезок на части  $[x_0, x_1], \dots, \dots, [x_{N-1}, x_N]$ ,  $x_0 = 0$ ,  $x_N = 1$ , и обозначим это разбиение через  $\Delta$ . Назовем сплайном  $S_{\Delta}^m(f, x)$  порядка  $m$  функцию, являющуюся многочленом степени  $m$  на каждом из отрезков

$[x_{n-1}, x_n]$ , т. е.

$$S_{\Delta}^m(f, x) = P_{nm}(x) = a_{n0} + \dots + a_{nm}x^m \quad \text{при } x_{n-1} \leq x \leq x_n, \quad (1)$$

и удовлетворяющую условиям непрерывности производных до порядка  $m - 1$  в точках  $x_1, \dots, x_{N-1}$ :

$$P_{nm}^{(k)}(x_n) = P_{n+1, m}^{(k)}(x_n) \quad \text{при } k = 0, \dots, m - 1, n = 1, \dots, N - 1. \quad (2)$$

Всего имеется в распоряжении  $Q = N(m + 1)$  неизвестных коэффициентов  $a_{nm}$ , и соотношения (2) образуют систему из  $(N - 1)m$  линейных алгебраических уравнений. Другие уравнения для коэффициентов получаются из условия близости сплайна к приближаемой функции и из некоторых дополнительных условий.

Рассмотрим простейшую задачу приближения линейными сплайнами ( $m = 1$ ). Тогда общее число  $Q$  свободных параметров равно  $2N$ . Поставим вопрос о построении сплайна  $S_{\Delta}^1(f, x)$ , совпадающего с функцией  $f(x)$  в точках  $x_0, \dots, x_N$ ; получится система уравнений

$$\begin{aligned} P_{n1}(x_{n-1}) &= f(x_{n-1}), & n &= 1, \dots, N, \\ P_{n1}(x_n) &= f(x_n), & n &= 1, \dots, N. \end{aligned}$$

Эта система распадается на системы уравнений относительно коэффициентов отдельных многочленов

$$\begin{aligned} P_{n1}(x_{n-1}) &= a_{n0} + a_{n1}x_{n-1} = f(x_{n-1}), \\ P_{n1}(x_n) &= a_{n0} + a_{n1}x_n = f(x_n). \end{aligned}$$

Отсюда находим

$$\begin{aligned} a_{n1} &= (f(x_n) - f(x_{n-1})) / (x_n - x_{n-1}), \\ a_{n0} &= f(x_{n-1}) - a_{n1}x_{n-1}. \end{aligned}$$

Многочлен  $P_{n1}(x)$  является многократно рассматривавшимся интерполяционным многочленом первой степени с узлами интерполяции  $x_{n-1}, x_n$ .

Широкое распространение сплайнов во многом вызвано тем, что они являются в определенном смысле наиболее гладкими функциями среди функций, принимающих заданные значения. Сплайны степени выше первой в случае гладкой  $f(x)$  хорошо приближают не только саму функцию, но и ее производные.

Сплайны первой степени  $S_{\Delta}^1(f, x)$  возникают из рассмотрения следующей вариационной задачи. Рассмотрим множество  $S_1$  кусочно-дифференцируемых функций  $s(x)$ , удовлетворяющих условиям

$$s(x_n) = f(x_n), \quad n = 0, \dots, N, \quad I_1(s) = \int_0^1 (s'(x))^2 dx < \infty.$$

Поставим задачу найти функцию  $s_1(x) \in S_1$ , реализующую

$$\inf_{s \in S_1} I_1(s).$$

Уравнение Эйлера для этого функционала имеет вид  $s''(x) = 0$ . Таким образом,  $s_1(x)$  линейна на каждом из отрезков  $[x_{n-1}, x_n]$  и, следовательно,  $s_1(x) = S_\Delta^1(f, x)$ .

Этот подход к приближению функции, при котором естественным образом возник сплайн  $S_\Delta^1(f, x)$ , допускает различные обобщения. Рассмотрим, например, множество  $S_2$  непрерывных, непрерывно-дифференцируемых и дважды кусочно непрерывно-дифференцируемых функций  $s(x)$ , удовлетворяющих условиям

$$s(x_n) = f(x_n), \quad n = 0, \dots, N, \quad I_2(s) = \int_0^1 (s''(x))^2 dx < \infty.$$

Ищется функция  $s_2(x) \in S_2$ , реализующая  $\inf_{s \in S_2} I_2(s)$ . Решением этой задачи оказывается сплайн третьей степени, удовлетворяющий условиям  $P''(x_0) = 0, P''(x_N) = 0$ .

Справедливость этого утверждения непосредственно следует из соответствующих результатов вариационного исчисления. Однако для полноты изложения ниже будет приведено его обоснование.

Уравнением Эйлера для рассматриваемого функционала является уравнение

$$\frac{d^2}{dx^2} \frac{\partial}{\partial s''} (s'')^2 = 0,$$

т. е.  $s^{(4)} = 0$ . Поэтому естественно предположить, что экстремум реализуется на некоторой функции  $s_2(x) \in S_2(x)$ , непрерывной вместе с первой производной, являющейся многочленом третьей степени  $P_{n3}(x) = a_{n0} + a_{n1}x + a_{n2}x^2 + a_{n3}x^3$  на каждом из отрезков  $[x_{n-1}, x_n]$ . Пусть  $\eta(x)$  — непрерывная, непрерывно-дифференцируемая и дважды кусочно непрерывно-дифференцируемая функция, удовлетворяющая условиям

$$\eta(x_0) = \dots = \eta(x_N) = 0. \quad (1)$$

Положим

$$\begin{aligned} F(t) &= \int_0^1 (s_2''(x) + t\eta''(x))^2 dx = \\ &= \int_0^1 (s_2''(x))^2 dx + 2t \int_0^1 s_2''(x) \eta''(x) dx + t^2 \int_0^1 (\eta''(x))^2 dx. \end{aligned}$$

Поскольку по предположению  $F(t) \geq F(0)$ , то

$$F'(0) = 2 \int_0^1 s_2''(x) \eta''(x) dx = 0.$$

Представим величину  $B(\eta) = F'(0)/2$  в виде  $B = b_1 + \dots + b_N$ ,

$$b_n = \int_{x_{n-1}}^{x_n} s_2''(x) \eta''(x) dx = \int_{x_{n-1}}^{x_n} P_{n3}''(x) \eta''(x) dx,$$

и произведем в выражении для  $b_n$  дважды интегрирование по частям. Получим

$$b_n = \int_{x_{n-1}}^{x_n} P_{n3}^{(4)}(x) \eta(x) dx + (P_{n3}''(x) \eta'(x) - P_{n3}'''(x) \eta(x)) \Big|_{x_{n-1}}^{x_n}.$$

Вследствие (1) и равенства  $P_{n3}^{(4)}(x) = 0$  имеем

$$b_n = P_{n3}''(x) \eta'(x) \Big|_{x_{n-1}}^{x_n}.$$

После суммирования по  $n$  получим

$$B(\eta) = P_{N3}''(x_N) \eta'(x_N) + \sum_{n=1}^{N-1} (P_{n3}''(x_n) - P_{n+1,3}''(x_n)) \eta'(x_n) - P_{13}''(x_0) \eta'(x_0) \quad (2)$$

для любой функции  $\eta(x)$  рассматриваемого вида. При любом  $l$  можно подобрать функцию  $\eta_l(x)$  такую, что  $\eta_l'(x_l) = 1$ ,  $\eta_l'(x_n) = 0$  при  $n \neq l$  в пределах  $0 \leq n \leq N$ .

Заметим, что мы не имели права производить интегрирование по частям непосредственно в исходном интеграле  $B(\eta)$ , так как неясен вопрос о существовании производных у  $s_2(x)$  более высокого порядка, чем первый, в точках  $x_1, \dots, x_{N-1}$ .

Подставляя такие  $\eta_l(x)$  в равенство

$$B(\eta_l(x)) = 0, \quad l = 0, \dots, N,$$

получаем  $N + 1$  уравнение

$$P_{13}''(0) = P_{N3}''(1) = 0, \quad P_{n3}''(x_n) - P_{n+1,3}''(x_n) = 0, \quad (3)$$

$$n = 1, \dots, N - 1.$$

Условие, что минимизирующая функция принимает заданные значения в точках  $x_n$ , порождает  $2N$  уравнений

$$P_{n3}(x_n) = P_{n+1,3}(x_n) = f(x_n), \quad n = 1, \dots, N - 1,$$

$$P_{13}(x_0) = f(x_0), \quad P_{N3}(x_N) = f(x_N), \quad (4)$$

а условия непрерывности  $s_2'(x)$  в точках  $x_n$  порождают уравнения

$$P'_{n3}(x_n) = P'_{n+1,3}(x_n), \quad n = 1, \dots, N - 1. \quad (5)$$

В совокупности нами получено  $4N$  уравнений ((3)—(5)) относительно  $4N$  неизвестных коэффициентов многочленов  $P_{n3}(x)$ .

Исследуем вопрос о разрешимости и о практическом решении системы уравнений (3)—(5). Для удобства введем в рассмотрение величины  $M_n = s_2''(x_n)$ . Поскольку функция  $s_2''(x)$  линейна на  $[x_{n-1}, x_n]$ , то

$$s_2''(x) = M_{n-1} \frac{x_n - x}{h_n} + M_n \frac{x - x_{n-1}}{h_n} \quad \text{на } [x_{n-1}, x_n];$$

здесь  $h_n = x_n - x_{n-1}$ . Из этого соотношения и из условий

$$s_2(x_{n-1}) = f(x_{n-1}), \quad s_2(x_n) = f(x_n)$$

можно получить, что

$$s_2(x) = P_{n3}(x) = M_{n-1} \frac{(x_n - x)^3}{6h_n} + M_n \frac{(x - x_{n-1})^3}{6h_n} + \left( f(x_{n-1}) - \frac{M_{n-1}h_n^2}{6} \right) \frac{x_n - x}{h_n} + \left( f(x_n) - \frac{M_n h_n^2}{6} \right) \frac{x - x_{n-1}}{h_n} \quad \text{на } [x_{n-1}, x_n]. \quad (6)$$

Условия

$$P'_{n3}(x_n) = P'_{n+1,3}(x_n), \quad n = 1, \dots, N-1,$$

порождают уравнения

$$\frac{h_n}{6} M_{n-1} + \frac{h_n + h_{n+1}}{3} M_n + \frac{h_{n+1}}{6} M_{n+1} = \frac{f(x_{n+1}) - f(x_n)}{h_{n+1}} - \frac{f(x_n) - f(x_{n-1})}{h_n}. \quad (7)$$

Кроме того, имеем условия

$$P''_{13}(x_0) = 0, \quad P''_{N3}(x_N) = 0,$$

иначе  $M_0 = 0$ ,  $M_N = 0$ . После подстановки  $M_0 = 0$  и  $M_N = 0$  соответственно в первое и последнее уравнения (6) получим систему

$$CM = d \quad (8)$$

из  $(N-1)$ -го уравнения с  $(N-1)$ -м неизвестным:

$$M = (M_1, \dots, M_{N-1})^T, \quad d = (d_1, \dots, d_{N-1})^T.$$

Элементы  $c_{ij}$ ,  $i, j = 1, \dots, N-1$ , матрицы  $C$  согласно (7) задаются соотношениями

$$c_{ij} = \begin{cases} h_i/6 & \text{при } j = i - 1, \\ (h_i + h_{i+1})/3 & \text{при } j = i, \\ h_{i+1}/6 & \text{при } j = i + 1, \\ 0 & \text{при } |i - j| > 1, \end{cases} \quad (9)$$

а элементы  $d_i$  столбца  $d$  — соотношениями

$$d_i = \frac{f(x_{i+1}) - f(x_i)}{h_{i+1}} - \frac{f(x_i) - f(x_{i-1}))}{h_i}.$$

Эта система решается методом прогонки (см. гл. IX) примерно за  $8N$  арифметических операций. После нахождения  $M_j$  по формуле (6) определяем многочлены  $P_{n3}(x)$ .

Покажем, что система уравнений (8) однозначно разрешима. Поскольку число уравнений равно числу неизвестных, то достаточно показать, что однородная система

$$CM = 0 \quad (10)$$

имеет только нулевое решение.

Предположим противное, т. е. что существует ненулевое решение  $M^0 = (M_1^0, \dots, M_{N-1}^0)^T$  системы (10). Пусть  $n_0$  — значение  $n$ , при котором достигается  $\max_{1 \leq n \leq N-1} |M_n^0|$ , т. е.

$$|M_{n_0}^0| = \max_{1 \leq n \leq N-1} |M_n^0| \neq 0.$$

В уравнении  $\sum_{j=1}^{N-1} c_{n_0 j} M_j^0 = 0$  перенесем в левую часть все слагаемые, кроме  $c_{n_0 n_0} M_{n_0}^0$ . Получим

$$c_{n_0 n_0} M_{n_0}^0 = \sum_{j \neq n_0} c_{n_0 j} M_j^0. \quad (11)$$

Из соотношений (9), определяющих  $c_{ij}$ , следует, что при любом  $i$  справедливы соотношения

$$\frac{1}{2} c_{ii} > \sum_{j \neq i} |c_{ij}| > 0,$$

поэтому имеем цепочку неравенств

$$\begin{aligned} |c_{n_0 n_0} M_{n_0}^0| &\leq \sum_{j \neq n_0} |c_{n_0 j}| |M_j^0| \leq \frac{1}{2} |c_{n_0 n_0}| \max_j |M_j^0| = \\ &= \frac{1}{2} |c_{n_0 n_0}| \cdot |M_{n_0}^0| = \frac{1}{2} |c_{n_0 n_0} M_{n_0}^0|. \end{aligned}$$

Вычтем из обеих частей результирующего неравенства

$$|c_{n_0 n_0} M_{n_0}^0| \leq \frac{1}{2} |c_{n_0 n_0} M_{n_0}^0|$$

выражение, стоящее в правой части. Получим  $\frac{1}{2} |c_{n_0 n_0} M_{n_0}^0| \leq 0$ . Поскольку  $c_{n_0 n_0} \neq 0$ , то  $M_{n_0}^0 = 0$ . Мы пришли к противоречию с предположением о существовании у системы (10) ненулевого решения.

Подведем итог проведенным построениям. Доказано существование решения системы (8). Сплайн третьей степени, определяемый соотношениями (6), будет искомым сплайном, удовлетворяющим условиям (3) — (5).

*Лемма.* Полученный сплайн  $s_2(x)$  реализует  $\inf_{s \in S_2} I_2(s)$ .



Доказательство. Пусть  $s(x)$  — произвольная функция из  $S_2$ . Положим  $\eta(x) = s(x) - s_2(x)$ . Имеем

$$F(1) = \int_0^1 (s''(x))^2 dx = \\ = \int_0^1 (s_2''(x))^2 dx + 2B(s(x) - s_2(x)) + \int_0^1 ((s(x) - s_2(x))'')^2 dx. \quad (12)$$

Поскольку  $s_2(x)$  удовлетворяет условиям (3), то в (12) имеем  $B(s(x) - s_2(x)) = 0$ . Таким образом, при любой функции  $s(x) \in S_2$  справедливо соотношение

$$I_2(s) = \int_0^1 (s''(x))^2 dx = \int_0^1 (s_2''(x))^2 dx + \int_0^1 ((s(x) - s_2(x))'')^2 dx = \\ = I_2(s_2) + \int_0^1 ((s(x) - s_2(x))'')^2 dx.$$

Отсюда следует справедливость утверждения леммы.

Из последнего соотношения следует также, что сплайн  $s_2(x)$  является единственной функцией, на которой функционал  $I_2(x)$  достигает своего наименьшего значения.

Задача 1. Доказать, что решение системы (8) удовлетворяет неравенству

$$\max_{1 \leq n \leq N-1} |M_n| \leq \frac{3}{\min_{1 \leq n \leq N-1} h_n} \max_{1 \leq n \leq N-1} |d_n|;$$

получить отсюда однозначную разрешимость системы (3), (4), (5).

При рассматриваемом подходе получаемый сплайн совпадает с  $f(x)$  во всех узлах; такие сплайны называют *интерполяционными*.

Задача 2. Пусть точки  $x_n$  распределены равномерно:  $x_{n+1} - x_n = h$ . Значение интерполяционного сплайна третьей степени  $s_2(x)$  выражается через значения функции  $f_n$  некоторой формулой

$$s_2(x) = \sum_{n=0}^N C_n(x) f_n.$$

Получить оценку

$$|C_n(x)| \leq a_3 e^{-b_3 \frac{|x-nh|}{h}}; \quad (13)$$

$a_3, b_3 > 0$  — абсолютные постоянные, не зависящие от  $f, N, h$ .

Из оценки (13) следует, что значение сплайна  $s_2(x)$  в точке  $x$  слабо зависит от значений  $f_n$  при большом  $|(x - nh)/h|$ .

Описанный выше способ построения сплайна третьей степени страдает следующим недостатком. Согласно (3) имеем  $s_2''(0) = s_2''(1) = 0$ , хотя, как правило,  $f''(0), f''(1) \neq 0$ ; поэтому точность приближенных формул  $f^{(k)}(x) \approx s_2^{(k)}(x)$  вблизи границы ухудшается. Если значения  $f''(0)$  и  $f''(1)$  известны, то в (7) при

$n = 1$  и  $n = N - 1$  следует положить  $M_0 = f''(0)$  и  $M_N = f''(1)$ . Оказывается, что для повышения точности целесообразно задавать в точках  $x_0, x_N$  значения первой производной. Дифференцируя (6), имеем

$$s_2'(x_0) = -\frac{M_0 h_1}{3} - \frac{M_1 h_1}{6} + \frac{f(x_1) - f(x_0)}{h_1}.$$

В случае, если величина  $f'(x_0)$  известна, полагаем правую часть равной  $f'(x_0)$  и получаем дополнительное уравнение, связывающее  $M_0$  и  $M_1$ . Обычно значение  $f'(x_0)$  неизвестно, поэтому поступим следующим образом. Определим интерполяционный многочлен  $Q_0(x)$  третьей степени, совпадающий с  $f(x)$  в точках  $x_0, x_1, x_2, x_3$ . Величину  $f'(x_0)$  заменим выражением  $Q_0'(x_0)$ . Окончательно получим

$$\frac{h_1}{3} M_0 + \frac{h_1}{6} M_1 = \frac{f(x_1) - f(x_0)}{h_1} - Q_0'(x_0).$$

Последняя формула может быть записана в виде

$$\frac{h_1}{3} M_0 + \frac{h_1}{6} M_1 = \frac{h_1}{3} Q_0''(x_0) + \frac{h_1}{6} Q_0''(x_1).$$

Аналогично построим интерполяционный многочлен третьей степени  $Q_N(x)$ , совпадающий с  $f(x)$  в точках  $x_N, x_{N-1}, x_{N-2}, x_{N-3}$ .

Коэффициенты  $M_j$  будем находить из системы уравнений, состоящей из совокупности уравнений (7) при  $n = 1, \dots, N - 1$  и уравнений

$$\begin{aligned} \frac{h_1}{3} M_0 + \frac{h_1}{6} M_1 &= \frac{h_1}{3} Q_0''(x_0) + \frac{h_1}{6} Q_0''(x_1), \\ \frac{h_N}{6} M_{N-1} + \frac{h_N}{3} M_N &= \frac{h_N}{3} Q_N''(x_N) + \frac{h_N}{6} Q_N''(x_{N-1}). \end{aligned} \quad (14)$$

Часто вместо  $Q_0$  и  $Q_N$  лучше брать многочлены четвертой степени, совпадающие с  $f(x)$  в пяти крайних точках.

Задача 3. Пусть  $h_n \equiv h = 1/N$ ,  $|f^{(4)}(x)| \leq A_4$ . Показать, что для сплайна, определяемого системой соотношений (7), (14), выполнены оценки

$$\max_{[0, 1]} |f^{(q)}(x) - s_2^{(q)}(x)| \leq \text{const} \cdot A_4 h^{4-q}, \quad q \leq 3.$$

Описанные выше сплайны часто неудобны из-за своей нелокальности: значение сплайна в точке  $x$  зависит от значений  $f(x_n)$  во всех узлах. Если в процессе работы со сплайнами (а она часто проходит в диалоговом режиме с визуализацией результатов на экране) требуется исправить одно значение, приходится заново решать систему уравнений (7). Особенно эта процедура неприятна в случае приближения функций многих переменных многомерными сплайнами.

Чтобы избежать этого, используем так называемые *локальные* (аппроксимационные) сплайны.

Локальный сплайн первой степени совпадает с построенным выше сплайном  $s_1(x)$ .

Локальные сплайны более высоких степеней, как правило, не совпадают с  $f(x)$  в узлах  $x_n$ . Однако это обстоятельство не носит принципиального характера. Все равно, как правило, значения  $f(x_n)$  известны с некоторой погрешностью  $\delta_n$ , т. е. нам заданы величины  $f_n = f(x_n) + \delta_n$ .

Построение локального сплайна третьей степени опишем на примере случая постоянного шага  $h_n \equiv h = 1/N$ . Для этого используется стандартный сплайн  $B(x)$  третьей степени, определяемый соотношениями

$$B(x) = \begin{cases} \frac{2}{3} - x^2 + \frac{1}{2}|x|^3 & \text{при } |x| \leq 1, \\ \frac{1}{6}(2 - |x|)^3 & \text{при } 1 \leq |x| \leq 2, \\ 0 & \text{при } 2 \leq |x|. \end{cases}$$

Локальные сплайны третьей степени  $B_2^{(1)}$  и  $B_2^{(2)}$  записывают в виде

$$B_2^{(i)}(x) = \sum_{n=-1}^{N+1} \alpha_n^{(i)} B\left(\frac{x - nh}{h}\right), \quad i = 1, 2;$$

способ выбора  $\alpha_n^{(i)}$  будет указан ниже.

**Задача 4.** Доказать, что при любых  $\alpha_n^{(i)}$  функции  $B_2^{(i)}$  являются сплайнами третьей степени, причем  $B_2^{(i)}(x) \equiv 0$  вне отрезка  $[-3h, 1 + 3h]$ .

При  $i = 1$  доопределяют значения  $f_{-1}$  и  $f_{N+1}$  линейной интерполяцией по значениям  $f_0, f_1$  и  $f_N, f_{N-1}$ , соответственно, т. е. берут  $f_{-1} = 2f_0 - f_1$ ,  $f_{N+1} = 2f_N - f_{N-1}$  и полагают  $\alpha_n = f_n$  при  $-1 \leq n \leq N + 1$ .

При  $i = 2$  доопределяют  $f_{-2}, f_{-1}$  и  $f_{N+1}, f_{N+2}$  кубической интерполяцией по значениям  $f_0, f_1, f_2, f_3$  и  $f_N, f_{N-1}, f_{N-2}, f_{N-3}$ , соответственно, и полагают

$$\alpha_n = (8f_n - f_{n+1} - f_{n-1})/6.$$

Конкретные формулы для вычисления величин  $f_{-1}, f_{-2}, f_{N+1}, f_{N+2}$  имеют вид

$$\begin{aligned} f_{-1} &= 4f_0 - 6f_1 + 4f_2 - f_3, \\ f_{-2} &= 10f_0 - 20f_1 + 15f_2 - 4f_3, \\ f_{N+1} &= 4f_N - 6f_{N-1} + 4f_{N-2} - f_{N-3}, \\ f_{N+2} &= 10f_N - 20f_{N-1} + 15f_{N-2} - 4f_{N-3}. \end{aligned}$$

**Задача 5.** Показать, что значение  $B_2^{(1)}(x)$  зависит только от значений  $f_n$  в 4-х ближайших к  $x$  точках  $x_n$ , а значения  $B_2^{(2)}(x)$  — в шести.

В случае, когда погрешности  $\delta_n$  велики, чаще используют сплайны  $B_2^{(1)}(x)$ , а в случае, когда малы — сплайны  $B_2^{(2)}(x)$ ; дело в том, что сплайны  $B_2^{(1)}(x)$

обладают несколько лучшими свойствами сглаживания погрешностей в значениях  $f(x_n)$ , но обеспечивают меньшую точность в случае  $\delta_n \equiv 0$  и гладкой  $f(x)$ .

Задача 6. Показать, что

$$B_2^{(1)}(x_0) = f_0, \quad B_2^{(1)}(x_N) = f_N,$$

$$B_2^{(2)}(x_0) = f_0, \quad B_2^{(2)}(x_1) = f_1, \quad B_2^{(2)}(x_{N-1}) = f_{N-1}, \quad B_2^{(2)}(x_N) = f_N.$$

Совпадение сплайнов  $B_2^{(1)}(x)$  и  $B_2^{(2)}(x)$  с  $f(x)$  в конечных точках существенно упрощает применение таких сплайнов в задачах машинной графики.

Задача 7. Пусть  $|f^{(k)}(x)| \leq A_k$  при  $k = 2, 3, 4$ . Показать, что

$$\max_{[0, 1]} |(B_2^{(1)}(x))^{(k)} - f^{(k)}(x)| \leq \text{const} \cdot h^{2-k} \quad \text{при } k \leq 1,$$

$$\max_{[0, 1]} |(B_2^{(2)}(x))^{(k)} - f^{(k)}(x)| \leq cA_4 h^{4-k} \quad \text{при } k \leq 3.$$

Рассмотрим подход к построению сплайнов, основанный на идее регуляризации. Пусть известно, что погрешности  $\delta_n$  — случайные величины с математическим ожиданием  $M\delta_n = 0$  и дисперсией  $\delta$ . Минимум  $I_2(s)$  ищут при условии

$$\sum_{n=0}^N (s(x_n) - f_n)^2 \leq c(N+1)\delta^2; \quad (15)$$

постоянную  $c$  порядка 1 выбирают, исходя из задаваемого доверительного уровня.

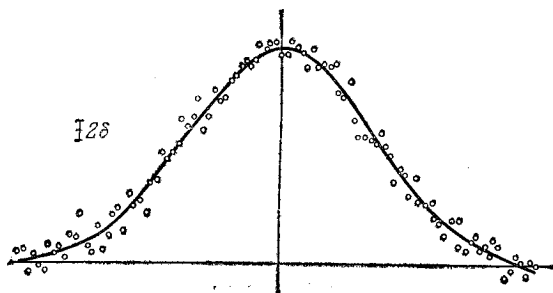


Рис. 4.11.1

Как правило, левая и правая части условия (15) равны в точке минимума  $I_2(s)$ . Поэтому обычно вместо исходной задачи рассматривают задачу безусловной минимизации по  $\lambda$  и  $s$  функционала

$$J_2(\lambda, s) = I_2(s) + \lambda \left( \sum_{n=0}^N (s(x_n) - f_n)^2 - c(N+1)\delta^2 \right).$$

Рис. 4.11.1 соответствует некоторому конкретному случаю приближения при  $N=100$  и  $c=0,9$ . При каждом фиксированном  $\lambda$  величина  $J(\lambda) = \inf_s J_2(\lambda, s)$  определяется с помощью решения системы линейных уравнений. Для нахождения  $\inf_{\lambda} J(\lambda)$  применяется какой-либо метод минимизации функций одной переменной (см. гл. VII).

## § 12. Энтропия и $\epsilon$ -энтропия

Условно можно говорить, что всякий объект задается двоякого рода информацией:

1) информацией о некотором множестве, которому он принадлежит, и

2) информацией, указывающей данный объект среди элементов множества.

Например, конкретный многочлен степени  $m$  задается: 1) информацией о том, что это многочлен степени  $m$ , и 2) таблицей его коэффициентов.

Нашей целью явится обсуждение вопроса о минимальном количестве информации, необходимой для задания элементов множеств. Для ряда задач, например для задачи приближения функций, более традиционным и приспособленным к реальным вычислениям является способ задания функции при помощи таблицы ее значений. Однако рассмотрение вопроса именно о минимальном количестве информации позволяет выработать более общий взгляд на рассматриваемую проблему.

Пусть множество  $G$ , элементы которого требуется задавать, состоит из  $N$  различных элементов  $g_1, \dots, g_N$  и пусть эти элементы задаются совокупностью их  $n$  двоичных знаков  $y_1, \dots, y_n = 0, 1$ .

Для однозначного задания  $g_k \in G$  необходимо, чтобы число элементов  $N$  не превосходило числа таких совокупностей. Всего таких совокупностей  $2^n$ , поэтому необходимым условием однозначности является условие  $N \leq 2^n$ . С другой стороны, при  $N \leq 2^n$  мы можем поставить в соответствие каждой совокупности  $y_1, \dots, y_n$  натуральное число

$$j = 1 + \sum_{i=1}^n 2^{n-i} y_i. \quad (1)$$

При изменении  $y_1, \dots, y_n$  число  $j$  пробегает все значения от 1 до  $2^n$ . Элемент  $g_j$  будем задавать совокупностью  $y_1, \dots, y_n$ , определяющей  $j$  согласно (1). Величину  $\log_2 N$  называют *энтропией*  $H(G)$  множества  $G$ . Минимальное количество двоичных знаков, необходимое для однозначного задания элементов множества  $G$ , будем обозначать  $J(G)$ ; через  $\bar{y}$  будем обозначать наименьшее целое, не меньшее  $y$ . Мы видели, что минимальное количество двоичных знаков, нужное для задания элементов  $G$ , равно наименьшему целому  $n$  такому, что  $N \leq 2^n$ , следовательно,

$$J(G) = \overline{\log_2 N} = \overline{H(G)}. \quad (2)$$

**Задача 1.** Доказать неравенства

$$J\left(\bigcup_{j=1}^q G_j\right) \leq \overline{\log_2 q} + \max_{1 \leq l \leq q} J(G_l), \quad (3)$$

$$J(G_1 \times \dots \times G_q) \leq \sum_{j=1}^q J(G_j), \quad (4)$$

здесь  $\times$  — знак прямого произведения, т. е. элементами  $G_1 \times \dots \times G_n$  являются совокупности  $(g^1, \dots, g^n)$  при  $g^i \in G_i$ .

**Задача 2.** Привести конкретные примеры, где в (3), (4) имеют место равенство или строгое неравенство.

**Задача 3.** Загадывается целое число от 0 до 1000. Отгадывающий может задавать вопросы, на которые будет получать ответы «да» или «нет». Доказать, что 10 вопросов достаточно для угадывания, а 9 — нет.

**Задача 4.** Имеется  $N$  ящиков, одинаковых по виду, из них  $N - 1$  равных по весу, а один тяжелее остальных. Разрешается производить  $n$  взвешиваний на весах без гирь. Доказать, что при  $N \leq 3^n$  можно гарантировать отыскание более тяжелого ящика, а при  $N > 3^n$  — нет.

**Задача 5.** Имеется  $N$  ящиков, одинаковых по виду, любые два из которых различны по весу. Оценить порядок  $n$  — минимального числа взвешиваний на весах без гирь, — при котором возможно упорядочивание ящиков по весу.

Ответ.  $n \sim N \log_2 N$  при  $N \rightarrow \infty$ .

Пусть теперь  $G$  состоит из бесконечного множества элементов. Ясно, что никакого фиксированного конечного числа двоичных знаков недостаточно для однозначного задания элементов из  $G$ . Справедливость этого утверждения также следует из (2).

В практических задачах часто достаточно задания элемента множества с некоторой точностью  $\epsilon$ . Поэтому рассмотрим вопрос о минимальном количестве информации (числа двоичных знаков), необходимом для задания элементов из  $G$  с точностью  $\epsilon$  в метрике некоторого метрического пространства  $F$ , которому принадлежит  $G$ .

Простейшие примеры задач такого рода.

1. Найти минимальное количество двоичных знаков, достаточное для указания точки единичного круга с точностью  $\epsilon$ .

2. Найти минимальное количество двоичных знаков, достаточное для задания точки единичной окружности с точностью  $\epsilon$ .

3. Рассматривается множество функций  $G$ , определенных на отрезке  $[0, X]$ , удовлетворяющих условию  $f(0) = 0$  и условию Липшица  $|f(x_1) - f(x_2)| \leq A|x_1 - x_2|$  при любых  $x_1, x_2 \in [0, X]$ . Требуется найти минимальное количество функций, достаточное для задания функции из этого класса с точностью  $\epsilon$ .

В каждом примере нужно указать метрику, в которой измеряется погрешность элемента, и пространство  $F$ .

В первом примере это может быть либо обычное евклидово расстояние

$$\rho(x, y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}, \quad (5)$$

$(x_1, x_2)$  — координаты точки  $x$ , либо расстояние

$$\rho(x, y) = \max(|x_1 - y_1|, |x_2 - y_2|),$$

либо

$$\rho(x, y) = |x_1 - x_2| + |y_1 - y_2|;$$

$F$  — множество точек плоскости.

Во втором примере это может быть любое из указанных выше расстояний или расстояние между точками единичной ок-

ружности, измеряемое как длина кратчайшей дуги окружности, соединяющей эти точки; при этом  $F$  — множество точек окружности.

В третьем примере метрика в пространстве  $F$  измеримых функций может быть определена, например, соотношениями

$$\rho(f_1, f_2) = \sup_{j_0, x_1} |f_1(x) - f_2(x)| \quad (6)$$

или

$$\rho(f_1, f_2) = \sqrt{\int_0^x |f_1(x) - f_2(x)|^2 dx}.$$

Предполагается следующий способ задания элементов  $g \in G$  с точностью  $\varepsilon$ . Элемент  $g$  определяется некоторой совокупностью из двоичных знаков  $y_1, \dots, y_n$ ; совокупности  $y_1, \dots, y_n$  поставлен в соответствие элемент  $f_{y_1, \dots, y_n} \in F$ , для которого  $\rho(f_{y_1, \dots, y_n}, g) \leq \varepsilon$ . Через  $J_\varepsilon(G, F)$  будем обозначать минимальное количество двоичных знаков, достаточное для задания элементов из  $G$  таким способом.

Множество  $F_\varepsilon \subset F$  назовем  $\varepsilon$ -сетью множества  $G \subset F$ , если для любого  $f \in G$  существует  $f_\varepsilon \in F_\varepsilon$  такое, что  $\rho(f_\varepsilon, f) \leq \varepsilon$ .

**Лемма 1.** Пусть элементы  $f_1, \dots, f_N \in F$  образуют  $\varepsilon$ -сеть для множества  $G$ . Тогда  $J_\varepsilon(G; F) \leq \log_2 N$ .

**Доказательство.** В соответствии с условием леммы каждому элементу  $g \in G$  можно поставить в соответствие некоторый элемент  $f_j \in F$ , для которого  $\rho(g, f_j) \leq \varepsilon$ . Пусть  $y_1, \dots, y_n$  — совокупность двоичных символов, соответствующих  $j$  согласно (1). Задание совокупности  $y_1, \dots, y_n$  определяет элемент  $f_j$ , для которого  $\rho(g, f_j) \leq \varepsilon$ .

Множество элементов  $G'$  называют  $\eta$ -различимым, если расстояние между двумя любыми его элементами более чем  $\eta$ .

**Лемма 2.** Пусть  $\Gamma \subset G$ ,  $\Gamma$  —  $2\varepsilon$ -различно и состоит из  $N$  элементов. Тогда  $\log_2 N \leq J_\varepsilon(G, F)$ .

**Доказательство.** Согласно определению способа задания функций  $g$ , неравенство  $J_\varepsilon(G; F) = n$  означает следующее. Существует  $M \leq 2^n$  элементов  $f_1, \dots, f_M \in F$ , удовлетворяющих условию: любому  $g \in G$  соответствует некоторое  $f_j$ , для которого  $\rho(g, f_j) \leq \varepsilon$ . Различным  $g_i, g_k \in \Gamma$  не может соответствовать один и тот же элемент  $f_j$ . Предположив противное, получим противоречие из цепочки неравенств

$$2\varepsilon < \rho(g_i, g_k) \leq \rho(g_i, f_j) + \rho(g_k, f_j) \leq 2\varepsilon.$$

Таким образом, число элементов  $N$  множества  $\Gamma$  не менее чем  $M$ , т. е.  $N \leq M \leq 2^n = 2^{J_\varepsilon(G; F)}$ . Отсюда следует справедливость утверждения леммы.

При помощи этих лемм можно получать оценки сверху и снизу величин  $J_\varepsilon(G; F)$ .

Диаметром некоторого множества  $Q \in F$  называется  $d(Q) = \sup_{q_1, q_2 \in Q} \rho(q_1, q_2)$ . Пространство  $F$  называется *центрируемым*, если для любого множества  $Q \in F$  конечного диаметра существует центр, т. е. элемент  $q$  такой, что  $\rho(q, f) \leq d(Q)/2$  при любом  $f \in Q$ .

Назовем  $\varepsilon$ -энтропией  $H_\varepsilon(G)$  множества  $G$  величину  $\log_2 N_\varepsilon(G)$ , где  $N_\varepsilon(G)$  — число элементов покрытия  $G$  минимальным числом множеств диаметра  $2\varepsilon$ . Если  $G$  — компакт, а  $F$  центрируемо, то справедливо неравенство

$$H_\varepsilon(G) \leq J_\varepsilon(G; F) \leq [H_\varepsilon(G)] + 1.$$

Рассмотрим примеры оценки величины  $J_\varepsilon(G, F)$ . Пусть  $F$  — вещественная ось,  $G$  — множество точек  $x \in [0, 1]$ ,  $\rho(x_1, x_2) = |x_1 - x_2|$ . Положим  $N_\varepsilon = \overline{0,5\varepsilon^{-1}}$ . Точки  $\varepsilon, 3\varepsilon, \dots, (2N_\varepsilon - 1)\varepsilon$  образуют  $\varepsilon$ -сеть. Поэтому  $J_\varepsilon(G; F) \leq \overline{\log_2 N_\varepsilon}$ . Согласно определению  $N_\varepsilon$  имеем  $(N_\varepsilon - 1)2\varepsilon < 1$ , поэтому  $\varepsilon' = 0,5(N_\varepsilon - 1)^{-1} > \varepsilon$ . Точки  $0, 2\varepsilon', \dots, 1$  образуют  $2\varepsilon$ -различимое множество из  $N_\varepsilon$  элементов, и согласно лемме 2  $J_\varepsilon(G; F) \geq \overline{\log_2 N_\varepsilon}$ . В итоге мы получили равенство

$$J_\varepsilon(G; F) = \overline{\log_2(0,5\varepsilon^{-1})}.$$

**Задача 6.** Пусть  $F$  —  $s$ -мерное евклидово пространство точек  $P(x_1, \dots, x_s)$ ,  $\rho(P_1, P_2)$  равно некоторой норме вектора  $P_2 - P_1$ ,  $G$  — множество точек, принадлежащее шару  $\rho(P, 0) \leq A_2 < \infty$  и содержащее внутри себя шар  $\rho(P, 0) \leq A_1 \neq 0$ . Доказать асимптотическое равенство

$$J_\varepsilon(G) = s \log_2(1/\varepsilon) + O(1).$$

**Задача 7.** Пусть  $G$  — пространство функций с  $(r-1)$ -й непрерывной производной, определенных на отрезке  $[0, X]$  и удовлетворяющих условиям

$$f(0) = \dots = f^{(r-1)}(0) = 0,$$

$$|f^{(r-1)}(x_2) - f^{(r-1)}(x_1)| \leq A|x_2 - x_1|$$

при  $x_1, x_2 \in [0, X]$ , а  $F$  — пространство измеримых на отрезке  $[0, X]$  функций с нормой, определяемой равенством (6). Величина  $J_\varepsilon(G; F)$  является функцией от параметров  $\varepsilon, A, F$ . Доказать, что она является функцией от комбинации  $X \sqrt[r]{A/\varepsilon}$ , т. е.

$$J_\varepsilon(G; F) = \psi\left(X \sqrt[r]{A/\varepsilon}\right).$$

На языке физики комбинация  $X \sqrt[r]{A/\varepsilon}$  называется безразмерной, поскольку она не меняется при изменении масштабов по осям  $x$  и  $f$ . Последнее обстоятельство и должно быть использовано при решении задачи.

**Задача 8.** Доказать, что  $\ln \psi(y) = \ln y + O(1)$ , где  $\psi(y)$  определена в задаче 7.

**Задача 9.** В условиях задачи 7 для случая  $X, A, r = 1$  доказать, что  $J_\varepsilon(G; F) = \overline{1/\varepsilon} - 1$ .

Литература к гл. IV: [4, 23, 29, 60].



## Глава V

### МНОГОМЕРНЫЕ ЗАДАЧИ

Вопросы приближения, интерполирования, численного интегрирования и дифференцирования функций одной переменной, как видно из предшествующего, разработаны достаточно подробно. В настоящее время на основе результатов теоретических исследований созданы довольно развитые системы стандартных программ решения одномерных задач. Значительная часть результатов теоретических исследований для одномерного случая может быть автоматически перенесена на случай функций двух и более переменных, однако при этом часто могут появляться практически недостаточно эффективные методы.

При теоретических построениях задачи разбивают на классы, например выделяют класс задач вычисления интегралов от функций с ограниченными производными, а затем проводят исследования, связанные с этими классами задач. Конечно, принимаемое описание не всегда (скорее даже редко) хорошо описывает класс реально встречающихся задач, однако скорости современных ЭВМ таковы, что, несмотря на грубость описания одномерных задач, удается решать большинство из них, пользуясь стандартными методами, разработанными в результате теоретических исследований.

Трудоемкость решения задач резко возрастает с ростом размерности решаемых задач, и поэтому, как правило, не удается разработать стандартные методы решения широких классов многомерных задач со столь же высокой точностью, как в одномерном случае.

Несколько утешает следующее обстоятельство. Многомерные математические задачи обычно возникают из описания сложных процессов. Обычно уже эти описания являются довольно грубыми, и поэтому значительно реже предъявляется требование решения этих задач с такой же высокой точностью, как в одномерном случае. Например, требования к точности решения уравнений газовой динамики существенно ниже требований, предъявляемых к решению уравнений баллистики и небесной механики.

Не следует, однако, думать, что решение многомерных задач является почти безнадежным делом. В настоящее время многообразие многомерных задач требует, как правило, привлечения

к их исследованию специалистов более высокой квалификации. Развитие теории методов решения многомерных задач и повышение скорости работы ЭВМ, несомненно, повлекут за собой создание стандартных методов решения задач и, как следствие, снижение предъявляемых требований к квалификации пользователей.

В настоящей главе рассмотрены вопросы интерполирования, численного интегрирования и дифференцирования функций в случае нескольких пространственных переменных.

## § 1. Метод неопределенных коэффициентов

Решение ряда многомерных задач часто сводится к решению следующих элементарных задач.

Пусть в некоторой области  $s$ -мерного пространства  $G$  заданы точки  $P_1, \dots, P_N$  и значения функции  $f$  в этих точках. Требуется:

- 1) получить приближение к значению функции  $f(P)$ ;
- 2) получить приближение к значению некоторой производной  $Df$  функции в точке  $P$ ;
- 3) вычислить интеграл

$$I(f) = \int_G f(P) p(P) dP,$$

где  $p(P)$  — некоторая весовая функция.

Простейшим способом решения этих задач является *метод неопределенных коэффициентов*. Пусть из каких-то соображений нам известно, что функция  $f(P)$  хорошо приближается линейными комбинациями вида

$$\sum_{j=1}^N B_j \omega_j(P).$$

Потребуем, чтобы такая линейная комбинация совпадала с  $f(P)$  в заданных узлах, т. е. выполнялись равенства

$$\sum_{j=1}^N B_j \omega_j(P_q) = f(P_q), \quad q = 1, \dots, N. \quad (1)$$

Предположим, что  $\det \|\omega_j(P_q)\| \neq 0$ . Тогда матрица  $\|\omega_j(P_q)\|$  имеет обратную  $A = \|a_{jq}\|$  и решение системы (1) записывается в виде

$$B_j = \sum_{q=1}^N a_{jq} f(P_q). \quad (2)$$

Функция

$$g(P) = \sum_{j=1}^N B_j \omega_j(P)$$

совпадает с функцией  $f(P)$  в точках  $P_q$ . Подставляя в предыдущее соотношение  $B_j$  из (2), получим иное представление  $g(P)$ , а именно:

$$g(P) = \sum_{q=1}^N z_q(P) f(P_q), \quad \text{где } z_q(P) = \sum_{j=1}^N a_{jq} \omega_j(P). \quad (3)$$

Такая форма записи интерполяционной функции является аналогом записи интерполяционного многочлена в форме Лагранжа. Как и в одномерном случае, можно надеяться, что при удачном выборе узлов  $P_q$  и функций  $\omega_j(P)$  будет мала погрешность в приближенных равенствах:

$$f(P) \approx g(P) = \sum_{q=1}^N z_q(P) f(P_q), \quad (4)$$

$$Df(P) \approx Dg(P) = \sum_{q=1}^N Dz_q(P) f(P_q), \quad (5)$$

$$I(f) \approx I(g) = \sum_{q=1}^N C_q f(P_q), \quad (6)$$

где  $C_q = I(z_q)$ .

Как отмечалось в гл. II для одномерного случая, при неудачном выборе большого числа узлов интерполирования погрешность приближенного равенства (4) может оказаться катастрофически большой. Поскольку приближенные равенства (5), (6) являются следствием приближенного равенства (4), то может быть большой и погрешность в приближенных равенствах (5), (6). Поэтому применение соотношений (4) — (6) требует обоснования и выяснения условий законности использования.

## § 2. Метод наименьших квадратов и регуляризация

При построении  $g(P)$  мы добивались равенства с функцией  $f(P)$  в  $N$  точках. Функция  $g(P)$  содержит большое число параметров, поэтому определенная свобода в их выборе позволяет  $g(P)$  сильно отличаться от  $f(P)$  вне этих  $N$  точек.

Повышение качества приближения может достигаться различными способами. Рассмотрим первый из них, называемый *методом наименьших квадратов*. Приближение ищется в виде

$$g(P) = \sum_{j=1}^n B_j \omega_j(P),$$

где  $n \ll N$ . Параметры  $B_j$  определяются из условия

$$\min_{B_1, \dots, B_n} \Phi(g),$$

где, например,

$$\begin{aligned} \Phi(g) = \Phi(B_1, \dots, B_n) &= \sum_{q=1}^N p_q (g(P_q) - f(P_q))^2 = \\ &= \sum_{q=1}^N p_q \left( \sum_{j=1}^n B_j \omega_j(P_q) - f(P_q) \right)^2. \quad (1) \end{aligned}$$

В основе метода наименьших квадратов лежит следующее соображение. Малость величины  $\Phi(g)$  обеспечивает близость функций  $g(P)$  и  $f(P)$  в точках  $P_q$ ; при  $n \ll N$  функция  $g(P)$  содержит относительно мало параметров, поэтому у нее меньше возможностей отличаться от  $f(P)$  вне узлов по сравнению со случаем  $n = N$ .

Числа  $p_q > 0$ , называемые весами, подбираются в зависимости от плотности распределения точек  $P_q$ . Если значения  $f(P_q)$  содержат случайную погрешность, то  $P_q$  выбирают также в зависимости от дисперсии погрешностей измеряемых значений. Там, где точки  $P_q$  распределены плотнее, числа  $p_q$  берутся меньше; значениям  $f(P_q)$  с большей дисперсией погрешности ставят в соответствие меньшие значения  $p_q$ . Такие рекомендации выглядят довольно неопределенными, поскольку нельзя предложить общего правила, пригодного для всех задач. Для конкретных классов задач принципы выбора  $p_q$  и  $n$  вырабатываются с учетом специфических свойств задач на основе статистических критериев и численного эксперимента.

Приравнивая нулю производные  $\partial\Phi/\partial B_i$ , получим систему линейных уравнений для определения  $B_j$ :

$$\frac{1}{2} \frac{\partial\Phi}{\partial B_i} = \sum_{j=1}^n d_{ij} B_j - d_i = 0, \quad (2)$$

$$d_{ji} = d_{ij} = \sum_{q=1}^N p_q \omega_i(P_q) \omega_j(P_q),$$

$$d_i = \sum_{q=1}^N p_q \omega_i(P_q) f(P_q).$$

Раскрывая скобки в (1), получим

$$\Phi(B_1, \dots, B_n) = \sum_{i,j=1}^n d_{ij} B_i B_j - 2 \sum_{j=1}^n d_j B_j + d_0,$$

где

$$d_0 = \sum_{q=1}^N p_q (f(P_q))^2.$$

Так как  $\Phi(B_1, \dots, B_n) \geq 0$ , то симметричная матрица  $D = \|d_{ij}\|$  неотрицательна. Это свойство используется в ряде стандартных программ метода наименьших квадратов. В ряде случаев бывает удобно искать  $B_j$  непосредственно минимизируя  $\Phi$  каким-либо итерационным методом.

Выражая  $B_j$  из (2) через  $d_i$ , а затем через  $f(P_q)$ , получим

$$B_j = \sum_{q=1}^N a_{jq} f(P_q);$$

следовательно,

$$g(P) = \sum_{q=1}^N z_q(P) f(P_q), \quad (3)$$

где

$$z_q(P) = \sum_{i=1}^n a_{iq} \omega_i(P).$$

Воспользовавшись (3), можно получить также формулу численного дифференцирования  $Df \approx Dg$  и квадратурную формулу  $I(f) \approx I(g)$ .

В основе метода регуляризации непосредственно лежат соображения о сглаживании аппроксимирующей функции. Наиболее распространенной формой метода регуляризации является следующая. Приближение отыскивается в виде

$$g(P) = \sum_{j=1}^n B_j \omega_j(P),$$

а коэффициенты  $B_j$  выбираются из условия минимума выражения

$$\Phi(\lambda, g) = \Phi(g) + \lambda \Psi(g), \quad \lambda > 0. \quad (4)$$

Функционал  $\Psi(g)$  подбирается из следующего условия: если значение этого функционала невелико, то функция  $g$  обладает определенной гладкостью. Например,  $\Psi(g)$  может быть некоторым приближением к интегралу  $\int_G |\text{grad } g(P)|^2 dP$ . В приложениях распространен случай  $n = N$ , на котором мы далее и остановимся. Пусть минимум выражения  $\Phi(\lambda, g)$  достигается при некоторых  $B_1^\lambda, \dots, B_N^\lambda$  и

$$g^\lambda(P) = \sum_{j=1}^N B_j^\lambda \omega_j(P).$$

Рассмотрим крайние случаи:  $\lambda = 0$  и  $\lambda$  — очень большое число. Имеем равенство

$$\Phi(0, g) = \sum_{q=1}^N p_q \left( \sum_{j=1}^N B_j \omega_j(P_q) - f(P_q) \right)^2.$$

Если  $\det \|\omega_j(P_q)\| \neq 0$ , то система (1.1) имеет решение и на ее решении правая часть этого равенства обращается в нуль. В то же время выражение  $\Phi(0, g)$  всегда неотрицательно. Таким образом, нижняя грань достигается на значениях  $B_j$ , являющихся решениями системы (1.1). Тогда  $g^0(P)$  совпадает с интерполяционным многочленом с узлами интерполяции  $P_j$ . При больших  $\lambda$  в функционале  $\Phi(\lambda, g)$  определяющим является второе слагаемое, нижняя грань которого достигается на гладкой функции. Следовательно, есть какие-то основания ожидать, что при промежуточных значениях  $\lambda$  функции  $g^\lambda(P)$  будут гладкими и в то же время не очень сильно отличающимися от приближаемой функции в заданных узлах.

Наши рассуждения выглядят довольно расплывчато, однако при такой общей постановке задачи приближения функции без конкретного указания системы функций  $\omega_i$ , распределения точек  $P_i$  и класса рассматриваемых функций вряд ли можно сказать что-либо более определенное. Подобные грубые «физические» соображения часто помогают при конструировании новых методов решения задачи в условиях недостаточной информированности о самой задаче.

Ниже на конкретном примере будет объяснено существо эффекта, достигаемого за счет применения регуляризации.

### § 3. Примеры регуляризации

Пусть  $f(x)$  — действительная периодическая функция с периодом 1. Пусть известны некоторые приближенные значения  $f_q$  величин  $f(x_q)$  при  $x_q = qh$ ,  $q = 0, \dots, N-1$ ;  $Nh = 1$ . Погрешности приближенных значений  $f_q - f(x_q) = d_q$  — независимые случайные величины с математическим ожиданием, равным нулю, и дисперсией  $d^2$ . Требуется получить таблицу приближенных значений производных  $f'(x_q)$ . Погрешность оценивается в норме, соответствующей скалярному произведению

$$(f, g) = \left( \sum_{q=0}^{N-1} f_q \bar{g}_q \right) h.$$

Далее  $f(x)$  рассматривается как периодически продолженная функция.

Для нахождения производной используем простейшую формулу численного дифференцирования

$$f'(x_q) \approx Df_q = \frac{f_{q+1} - f_{q-1}}{2h}. \quad (1)$$

Предположим, что функция  $f(x)$  достаточно гладкая и шаг  $h$  настолько мал, что при отсутствии погрешностей в значениях функции погрешность формулы численного дифференцирования

$$f'(x_q) \approx \frac{f(x_{q+1}) - f(x_{q-1}))}{2h} \quad (2)$$

пренебрежимо мала. Погрешность в значении производной представима в виде

$$R_q = \frac{f_{q+1} - f_{q-1}}{2h} - f'(x_q) = \left( \frac{f(x_{q+1}) - f(x_{q-1}))}{2h} - f'(x_q) \right) + \frac{d_{q+1} - d_{q-1}}{2h} \approx \frac{d_{q+1} - d_{q-1}}{2h} = r_q.$$

Здесь и далее знак приближенного равенства означает равенство с точностью до величин порядка погрешности формулы (2).

Имеем равенство

$$\begin{aligned} M \|R_q\|^2 &\approx M \|r_q\|^2 = M \left( \sum_{q=0}^{N-1} \left( \frac{d_{q+1} - d_{q-1}}{2h} \right)^2 h \right) = \\ &= \frac{N}{4} \sum_{q=0}^{N-1} (M d_{q+1}^2 - 2M d_{q-1} d_{q+1} + M d_{q-1}^2). \end{aligned}$$

Согласно указанным выше свойствам случайных величин  $d_q$ , имеем

$$M d_q d_p = d^2 \delta_p^q. \quad (3)$$

Тогда получаем  $M \|R\|^2 = N^2 d^2 / 2$ , иначе среднеквадратичное значение нормы погрешности равно  $Nd/\sqrt{2}$ . Мы видим, что погрешность приближенной формулы (1) возрастает с уменьшением  $h = 1/N$ ; в то же время для малости погрешности в предположении отсутствия погрешностей в исходной информации (2) требуется достаточная малость  $h$ .

Чтобы выяснить пути уменьшения погрешности, проведем более детальное исследование. Пусть  $A_j^0$  — дискретные коэффициенты Фурье (ДКФ) функции  $f(x_q)$ :

$$f(x_q) = \sum_{-N/2 < j \leq N/2} A_j^0 \exp \{2\pi i j x_q\},$$

а  $A_j$  — ДКФ функции  $f_q$ . Величины  $a_j = A_j - A_j^0$  будут ДКФ функции  $d_q = f_q - f(x_q)$ . Имеем

$$\begin{aligned} d_q &= \sum_{-N/2 < j \leq N/2} a_j \exp \{2\pi i j x_q\}, \\ a_j &= h \sum_{q=0}^{N-1} d_q \exp \{-2\pi i j x_q\}. \end{aligned} \quad (4)$$

Чтобы вычислить ДКФ для  $r_q$ , применим оператор численного дифференцирования  $D$  к функции  $\exp \{2\pi i j x_q\}$ :

$$D \exp \{2\pi i j x_q\} = \frac{\exp \{2\pi i j x_{q+1}\} - \exp \{2\pi i j x_{q-1}\}}{2h} = \frac{i \sin 2\pi j h}{h} \exp \{2\pi i j x_q\}.$$

Применяя оператор  $D$  к обеим частям (4), получим

$$r_q = \sum_{-N/2 < j \leq N/2} \frac{i \sin 2\pi j h}{h} a_j \exp \{2\pi i j x_q\}.$$

Из равенства Парсеваля

$$\|r_q\|^2 = \sum_{-N/2 < j \leq N/2} \left( \frac{\sin (2\pi j h)}{h} \right)^2 |a_j|^2$$

следует, что

$$M \|r_q\|^2 = \sum_{-N/2 < j \leq N/2} \left( \frac{\sin (2\pi j h)}{h} \right)^2 M |a_j|^2.$$

Имеем

$$|a_j|^2 = a_j \bar{a}_j = h^2 \sum_{q, p=0}^{N-1} d_q d_p \exp \{2\pi i j (x_p - x_q)\};$$

мы заменили  $\bar{d}_p$  на  $d_p$ , воспользовавшись тем, что  $d_p$  действительны. Таким образом, мы получили, что

$$M |a_j|^2 = h^2 \sum_{q, p=0}^{N-1} \exp \{2\pi i j (x_p - x_q)\} M (d_q d_p) = h^2 \sum_{q, p=0}^{N-1} d^2 \delta_p^q = h d^2,$$

и

$$M \|r_q\|^2 = \sum_{-N/2 < j \leq N/2} \left( \frac{\sin(2\pi j h)}{h} \right)^2 h d^2. \quad (15)$$

Возьмем некоторую функцию  $\mu(j)$  и положим

$$f^\mu(x) = \sum_{-N/2 < j \leq N/2} A_j^0 \mu(j) \exp \{2\pi i j x\},$$

$$f_q^\mu = \sum_{-N/2 < j \leq N/2} A_j \mu(j) \exp \{2\pi i j x_q\}.$$

При заданных значениях  $f_q$  можно вычислить коэффициенты  $A_j$ , а следовательно, и величины  $f_q^\mu$ . Пусть  $r_q^\mu$  — составляющая погрешности формулы численного дифференцирования, являющаяся следствием погрешностей  $d_q$ :

$$r_q^\mu = D f_q^\mu - D f^\mu(x_q).$$

Аналогично (5) имеем

$$M \|r_q^\mu\|^2 = \sum_{-N/2 < j \leq N/2} \left( \frac{\sin(2\pi j h)}{h} \right)^2 \mu^2(j) h d^2.$$

Если

$$D f(x_q) \approx D f^\mu(x_q), \quad M \|r^\mu\|^2 \ll M \|r\|^2, \quad (6)$$

то имеет смысл положить  $D(x_q) \approx D f^\mu(x_q)$ . Для выполнения первого из соотношений (6) существенна близость коэффициентов Фурье функций  $f(x)$  и  $f^\mu(x)$ ; для выполнения второго — малость  $\mu(j)$  при больших  $j$ .

Попробуем прибегнуть к сглаживанию функции  $f_q$  при помощи метода регуляризации. Рассмотрим функционал

$$\Phi_1^h(\lambda, g) = h \sum_{q=0}^{N-1} (g_q - f_q)^2 + \lambda^2 h \sum_{q=0}^{N-1} \left( \frac{g_{q+1} - g_q}{h} \right)^2,$$

где  $g_N = g_0$ . Если  $g_q$  являются значениями  $g'(x_q)$  гладкой функции  $g(x)$ , то величина

$$h \sum_{q=0}^{N-1} \left( \frac{g_{q+1} - g_q}{h} \right)^2$$

стремится к интегралу  $\int_0^1 (g'(x))^2 dx$ . Условие, что этот интеграл невелик, гарантирует определенную гладкость функции  $g(x)$ . Та-



ким образом, есть какие-то основания считать, что функционал  $\Phi_1^h(\lambda, g)$  удовлетворяет требованиям, накладываемым на функционалы метода регуляризации. Определим сеточную функцию  $g_q^\lambda$  из условия минимума функционала  $\Phi_1^h(\lambda, g)$  и положим

$$f'(x_q) \approx (g_{q+1}^\lambda - g_{q-1}^\lambda)/(2h).$$

Посмотрим, что является аналогом такой регуляризации для случая функций непрерывного аргумента. Пусть

$$f(x) = \sum_{j=-\infty}^{\infty} A_j^0 \exp\{2\pi i j x\},$$

$g_1^\lambda(x)$  — функция, реализующая минимум функционала

$$\Phi_1(\lambda, g) = \int_0^1 (g(x) - f(x))^2 dx + \lambda^2 \int_0^1 (g'(x))^2 dx.$$

Уравнение Эйлера для этого функционала записывается в виде

$$\lambda^2 g'' - (g - f) = 0.$$

Непосредственной проверкой убеждаемся, что функция

$$g_1^\lambda(x) = \sum_{j=-\infty}^{\infty} \frac{A_j^0}{1 + (2\lambda\pi j)^2} \exp\{2\pi i j x\}$$

является решением этого уравнения. Произведем сравнение  $g_1^\lambda(x)$  и  $f(x)$  для малых  $\lambda$ . Если  $|2\lambda\pi j| \ll 1$ , то коэффициенты Фурье этих функций  $A_j^0/(1 + (2\lambda\pi j)^2)$  и  $A_j^0$  близки между собой. Если  $|2\lambda\pi j| \gg 1$ , то коэффициенты Фурье функции  $g_1^\lambda(x)$  много меньше коэффициентов Фурье функции  $f(x)$ . Таким образом, на языке техники регуляризации равносильна некоторой «фильтрации»: несущественно искажая гармоники с малой частотой колебания, она сильно ослабляет гармоники с большой частотой. Если требуется еще меньше исказить амплитуды  $A_j^0$  гармоник с малой частотой и лучше «отфильтровать» высокочастотные колебания, то можно рассмотреть функционал

$$\Phi_n(\lambda, g) = \int_0^1 (g(x) - f(x))^2 dx + \lambda^{2n} \int_0^1 (g^{(n)}(x))^2 dx.$$

Уравнение Эйлера для этого функционала имеет вид

$$\lambda^{2n} g^{(2n)} + (-1)^n (g - f) = 0,$$

отсюда

$$g_n^\lambda(x) = \sum_{j=-\infty}^{\infty} \frac{A_j^0}{1 + (2\lambda\pi j)^{2n}} \exp\{2\pi i j x\}.$$

Рассмотрим графики множителей  $1/(1 + (2\lambda j)^{2n})$ ; при  $|j| < 1/(2\pi\lambda)$  и  $n \rightarrow \infty$  эти множители стремятся к 1, и, таким образом, при больших  $n$  амплитуды соответствующих гармоник искажаются все меньше и меньше. В то же время при  $|j| > 1/(2\pi\lambda)$  эти множители стремятся к 0 и, таким образом,

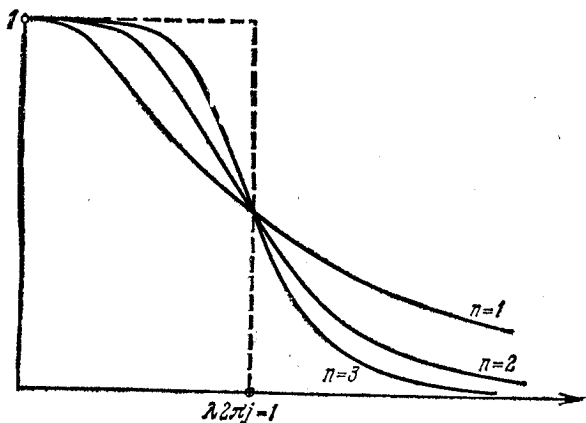


Рис. 5.3.1

соответствующие гармоники умножаются на все меньшие множители (рис. 5.3.1). Таким образом, при  $n \rightarrow \infty$  и  $1/(2\pi\lambda)$  не целом

$$g_n^\lambda(x) \rightarrow \tilde{g}^\lambda(x) = \sum_{|j| < 1/(2\pi\lambda)} A_j^0 \exp\{2\pi i j x\}.$$

Вернемся к дискретному случаю. В выражении  $\Phi_1^h(\lambda, g)$  величина  $g_q$  входит только в сумму слагаемых

$$h(g_q - f_q)^2 + \lambda^2 h \left( \left( \frac{g_{q+1} - g_q}{h} \right)^2 + \left( \frac{g_q - g_{q-1}}{h} \right)^2 \right).$$

Имеем систему уравнений относительно значений  $g_q$  в точке экстремума:

$$\frac{1}{2h} \frac{\partial \Phi_1^h}{\partial g_q} = (g_q - f_q) - \frac{\lambda^2}{h^2} (g_{q+1} - 2g_q + g_{q-1}) = 0. \quad (7)$$

Будем обозначать решение (7) через  $g_q^\lambda$ . Матрица системы уравнений (7) относительно неизвестных  $g_q$  совпадает с матрицей положительно определенной квадратичной формы

$$\sum_{q=0}^{N-1} g_q^2 + \frac{\lambda^2}{h^2} \sum_{q=0}^{N-1} (g_{q+1} - g_q)^2,$$

поэтому ее определитель отличен от нуля, и система (7) имеет, и притом единственное, решение.

Пусть

$$f_q = \sum_{-N/2 < j < N/2} A_j \exp\{2\pi i j x_q\}.$$

Будем отыскивать периодическую функцию  $g_q^\lambda$  в виде

$$g_q^\lambda = \sum_{-N/2 < j < N/2} A_j^\lambda \exp \{2\pi i j x_q\}.$$

Сначала вычислим оператор второй разности  $\delta^2 f_q = f_{q+1} - 2f_q + f_{q-1}$  от функции  $\exp \{2\pi i j x_q\}$ :

$$\begin{aligned} \delta^2 \exp \{2\pi i j x_q\} &= \\ &= \exp \{2\pi i j x_{q+1}\} - 2 \exp \{2\pi i j x_q\} + \exp \{2\pi i j x_{q-1}\} = \\ &= [\exp \{2\pi i j h\} - 2 + \exp \{-2\pi i j h\}] \exp \{2\pi i j x_q\} = \\ &= [2 \cos (2\pi j h) - 2] \exp \{2\pi i j x_q\}; \end{aligned}$$

отсюда получаем

$$\delta^2 \exp \{2\pi i j x_q\} = -4 \sin^2 (\pi j h) \exp \{2\pi i j x_q\}. \quad (8)$$

Подставим представление  $f_q$  и  $g_q$  в виде суммы Фурье в (7), преобразуем вторую разность  $\delta^2 g_q$  с учетом (8), приведем подобные члены и получим

$$\sum_{-N/2 < j < N/2} (A_j^\lambda - A_j + 4\lambda^2 \left(\frac{\sin (\pi j h)}{h}\right)^2 A_j^\lambda) \exp \{2\pi i j x_q\} = 0.$$

Это равенство будет удовлетворяться при

$$A_j^\lambda = A_j \left[1 + 4\lambda^2 \left(\frac{\sin (\pi j h)}{h}\right)^2\right]^{-1}.$$

Таким образом,

$$g_q^\lambda = \sum_{-N/2 < j < N/2} \frac{A_j}{1 + 4\lambda^2 \left(\frac{\sin^2 (\pi j h)}{h}\right)^2} \exp \{2\pi i j x_q\}.$$

Множитель  $\left[1 + 4\lambda^2 \left(\frac{\sin (\pi j h)}{h}\right)^2\right]^{-1}$  убывает с ростом  $j$ . Следовательно, регуляризация с использованием функционала  $\Phi_1^h(\lambda, g)$  также приводит к ослаблению высших гармоник функции.

По аналогии с функционалами  $\Phi_n(\lambda, g)$ , можно осуществлять регуляризацию с помощью функционалов

$$\Phi_n^h(\lambda, g) = h \sum_{q=0}^{N-1} (g_q - f_q)^2 + \lambda^{2n} h \sum_{q=0}^{N-1} \left(\frac{\Delta^n g_q}{h^n}\right)^2.$$

Решение системы уравнений (7) и систем линейных уравнений, возникающих при минимизации функционалов  $\Phi_n^h(\lambda, g)$ , можно получить за  $O(n^2 N)$  арифметических операций методом прогонки решения периодической сеточной краевой задачи (см. гл. IX).

По сравнению с регуляризацией с помощью функционалов  $\Phi_n^h(\lambda, g)$  при  $n > 1$  часто бывает более удобна  $n$ -кратная регу-

ляризация с помощью функционала  $\Phi_1^n(\lambda, g)$ . Гармоники  $\exp\{2\pi i j x\}$  в этом случае умножаются на  $[1 + 4\lambda^2 (\frac{\sin(\pi j h)}{h})^2]^{-n}$ . Преимущество такой последовательной регуляризации состоит в простоте алгоритма, несколько меньшем объеме вычислений (порядка  $O(nN)$ ), и меньшем влиянии вычислительной погрешности.

В случае рассматриваемого примера можно пойти и по такому пути регуляризации: произвести дискретное преобразование Фурье (если возможно — быстрое), найти коэффициенты Фурье  $A_j$ , заменить нулями  $A_j$ , соответствующие гармоникам с высокой частотой, а затем произвести обратное преобразование Фурье.

#### § 4. Сведение многомерных задач к одномерным

Выше рассматривались способы решения многомерных задач, не требующие дополнительной информации о распределении узлов  $x_1, \dots, x_s$ , в которых известны значения функции.

Во многих случаях расположение узлов носит регулярный характер. Если рассматриваемая функция задана аналитически, то узлы можно выбирать по желанию. При удачном расположении узлов приближение, интерполирование, численное дифференцирование и интегрирование функций многих переменных могут быть сведены к последовательному осуществлению этих операций над функциями одной переменной.

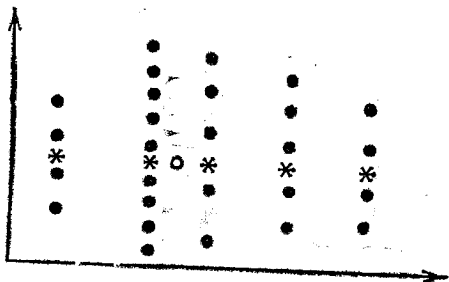


Рис. 5.4.1

Рассмотрим случай расположения узлов, изображенных как  $\bullet$  на рис. 5.4.1. Здесь множество всех узлов  $\Omega$  разбивается на подмножества узлов  $\Omega_1, \dots, \Omega_{m_1}$  (в данном случае  $m_1 = 5$ ), лежащих соответственно на прямых  $x_1 = x_1^1, \dots, x_1 = x_1^{m_1}$ .

Рассмотрим задачу вычисления значения некоторого оператора

$$\mathcal{D}f(P^0) = f^{(r_1, r_2)}(x_1^0, x_2^0).$$

Точка  $(x_1^0, x_2^0)$  изображена на рис. 5.4.1 как  $\circ$ . Частный случай  $r_1 = r_2 = 0$  соответствует задаче вычисления значения функции.

Возьмем какую-либо формулу вычисления производной  $g^{(r_1)}(x_1^0)$  по значениям функции в узлах  $x_1^{l_1}$ :

$$g^{(r_1)}(x_1^0) \approx \sum_{l_1=1}^{m_1} A_{l_1} g(x_1^{l_1}); \quad (1)$$

при этом не имеется в виду, что используются значения функции  $g$  во всех точках  $x_1^i$ . Например, в случае численного дифференцирования речь может идти о простейшей формуле численного дифференцирования по ближайшим к  $x_1^0$  двум узлам.

Подставляя  $g(x_1) = f^{(0, r_2)}(x_1, x_2^0)$ , получим

$$f^{(r_1, r_2)}(x_1^0, x_2^0) \approx \sum_{j_1=1}^{m_1} A_{j_1} f^{(0, r_2)}(x_1^{j_1}, x_2^0)$$

(точки  $(x_1^{m_1}, x_2^0)$  изображены на рис. 5.4.1 как \*).

Задача численного дифференцирования функции двух переменных  $f(x_1, x_2)$  свелась к задаче численного дифференцирования функций  $f(x_1^{m_1}, x_2)$  одной переменной. Такая задача уже рассматривалась ранее.

Пусть  $(x_1^m, x_2^{m_1}), \dots, (x_1^m, x_2^{m, n(m)})$  — узлы, образующие множество  $\Omega_m$ . При каждом  $j_1$  возьмем некоторую формулу вычисления производной  $g^{(r_2)}(x_2^0)$  по значениям функции в узлах  $x_2^{j_1 i_2}$ :

$$g^{(r_2)}(x_2^0) \approx \sum_{i_2=1}^{n(j_1)} A_{j_1 i_2} g(x_2^{j_1 i_2}).$$

Подставляя сюда  $g = f(x_1^{j_1}, x_2)$ , аналогично (1) получим

$$f^{(0, r_2)}(x_1^{j_1}, x_2^0) \approx \sum_{i_2=1}^{n(j_1)} A_{j_1 i_2} f(x_1^{j_1}, x_2^{j_1 i_2}). \quad (2)$$

Воспользовавшись этими соотношениями, из (1) получаем

$$f^{(r_1, r_2)}(x_1^0, x_2^0) \approx \sum_{j_1=1}^{m_1} A_{j_1} \sum_{i_2=1}^{n(j_1)} A_{j_1 i_2} f(x_1^{j_1}, x_2^{j_1 i_2}).$$

Заметим, что формулу (2) нужно строить лишь для значений  $j_1$  таких, что  $A_{j_1} \neq 0$ .

Наиболее часто узлы располагаются в узлах некоторой сетки, являющейся произведением одномерных сеток

$$\Omega_1 = \Omega_1 \times \dots \times \Omega_s, \quad (3)$$

иначе говоря,  $\Omega$  состоит из точек  $(x_1^{j_1}, \dots, x_s^{j_s})$  при  $j_1 = 1, \dots, N_1, \dots, j_s = 1, \dots, N_s$ . В случае расположения узлов в вершинах такой сетки можно было бы поменять местами численное дифференцирование по переменным  $x_1$  и  $x_2$ , т. е. сначала получить некоторую формулу

$$f^{(r_1, r_2)}(x_1^0, x_2^0) \approx \sum_{m_2} A_{m_2} f^{(r_1, 0)}(x_1^0, x_2^{m_2}),$$

а затем воспользоваться формулой

$$f^{(r_1, 0)}(x_1^0, x_2^{m_2}) \approx \sum_{m_1} A_{m_1 m_2} f(x_1^{m_1 m_2}, x_2^{m_2}).$$

Численное дифференцирование (интерполирование) функций большого числа переменных производится аналогично последовательным сведением к численному дифференцированию функций на единицу меньшего числа переменных.

При численном дифференцировании функций многих переменных нужно внимательно следить за величиной отбрасываемых остаточных членов.

Рассмотрим, например, задачу, где применение описанного выше приема последовательного численного дифференцирования может привести к получению неправильной формулы. Пусть значения функции  $f(x_1, x_2)$  известны в узлах сетки  $(m_1 h_1, m_2 h_2)$ . Требуется вычислить значение  $f_{x_1 x_2}(0, 0)$ . Выпишем простейшую формулу численного дифференцирования функции одной переменной

$$f'(x) \sim \frac{f(x+h_1) - f(x)}{h_1}.$$

По формуле Тейлора

$$f(x+h) = f(x) + f'(x)h + f''(x+\theta h)\frac{h^2}{2}, \quad 0 \leq \theta \leq 1,$$

имеем

$$f'(x) = \frac{f(x+h) - f(x)}{h} - f''(x+\theta h)\frac{h}{2}.$$

Поэтому можно написать, что

$$f_{x_1 x_2}(0, 0) = \frac{f_{x_2}(h_1, 0) - f_{x_2}(0, 0)}{h_1} - f_{x_1^2 x_2}(\theta h_1, 0)\frac{h_1}{2}. \quad (4)$$

В свою очередь, возьмем какие-либо аппроксимации производных  $f_{x_2}(h_1, 0)$ ,  $f_{x_2}(0, 0)$ . Имеем равенства:

$$f_{x_2}(h_1, 0) = \frac{f(h_1, h_2) - f(h_1, 0)}{h_2} - f_{x_2 x_2}(h_1, \theta_1 h_2)\frac{h_2}{2}, \quad (5)$$

$$f_{x_2}(0, 0) = \frac{f(0, 0) - f(0, -h_2)}{h_2} + f_{x_2 x_2}(0, \theta_2 h_2)\frac{h_2}{2}. \quad (6)$$

После подстановки (5), (6) в (4) получим

$$f_{x_1 x_2}(0, 0) = \frac{f(h_1, h_2) - f(h_1, 0) - f(0, 0) + f(0, -h_2)}{h_1 h_2} - f_{x_1^2 x_2}(\theta h_1, 0)\frac{h_1}{2} - \frac{1}{2} f_{x_2 x_2}(0, \theta_1 h_2) - \frac{1}{2} f_{x_2 x_2}(h_1, \theta_2 h_2). \quad (7)$$

Здесь при построении формулы численного дифференцирования мы одновременно учитывали остаточный член.

При построении аппроксимации без одновременного учета остаточного члена можно было бы прийти к неверной приближенной формуле

$$f_{x_1 x_2}(0, 0) \approx \frac{f(h_1, h_2) - f(h_1, 0) - f(0, 0) + f(0, -h_2)}{h_1 h_2}.$$

Соотношение (7) можно переписать в виде

$$f_{x_1 x_2}(0, 0) = \frac{f(h_1, h_2) - f(h_1, 0) - f(0, 0) + f(0, -h_2)}{h_1 h_2} - f_{x_2 x_2}(0, 0) + O(h), \quad h = \max(h_1, h_2)$$

Таким образом, если не обращать внимания на величину погрешности, то можно получить приближенную формулу с конечной погрешностью, в данном случае аппроксимирующую не требуемую производную, а выражение  $f_{x_1 x_2}(0, 0) + f_{x_2 x_2}(0, 0)$ .

Если в (5) и (6) использовать одни и те же формулы численного дифференцирования по переменной  $x_2$ , то вместо (7) получается формула с погрешностью, стремящейся к нулю при  $h \rightarrow 0$ . Например, вместо (6) воспользуемся равенством

$$f_{x_2}(0, 0) = \frac{f(0, h_2) - f(0, 0)}{h_2} - f_{x_2 x_2}(0, \theta_3 h_2) \frac{h_2}{2}. \quad (8)$$

После подстановки (5) и (8) в (4) получим

$$f_{x_1 x_2}(0, 0) = \frac{f(h_1, h_2) - f(h_1, 0) - f(0, h_2) + f(0, 0)}{h_1 h_2} - f_{x_1 x_2}(\theta h_1, 0) \frac{h_1}{2} - \frac{1}{2} f_{x_2 x_2}(h_1, \theta_1 h_2) + \frac{1}{2} f_{x_2 x_2}(0, \theta_3 h_2).$$

Если производная  $f_{x_1 x_2}$  непрерывно дифференцируема, то  $f_{x_1 x_2}(0, \theta_3 h_2) - f_{x_1 x_2}(h_1, \theta_1 h) = O(h)$ , поэтому

$$f_{x_1 x_2}(0, 0) = \frac{f(h_1, h_2) - f(h_1, 0) - f(0, h_2) + f(0, 0)}{h_1 h_2} + O(h).$$

Мы получили, что формула численного дифференцирования

$$f_{x_1 x_2}(0, 0) \approx \frac{f(h_1, h_2) - f(h_1, 0) - f(0, h_2) + f(0, 0)}{h_1 h_2} \quad (9)$$

имеет погрешность  $O(h)$ .

Покажем, как получить формулу (9) методом неопределенных коэффициентов. Зададимся видом формулы численного дифференцирования

$$f_{x_1 x_2}(0, 0) \approx l(f) = \frac{af(h_1, h_2) + bf(h_1, 0) + cf(0, h_2) + df(0, 0)}{h_1 h_2}. \quad (10)$$

Коэффициенты  $a, b, c, d$  в правой части взяты в виде  $\text{const} \cdot h_1 \cdot h_2$  из соображений размерности. Пусть  $[x]$  — обозначение размерности некоторой величины; например, если  $x$  — скорость, то  $[x] = \text{м/с}$ . Производная  $f_{x_1 x_2}$  имеет размерность  $[f]/([x_1] \cdot [x_2])$ , функция  $f$  — размерность  $[f]$ ,  $h_1$  — размерность  $[x_1]$ ,  $h_2$  — размерность  $[x_2]$ . Таким образом, величины  $f(P)/(h_1 h_2)$  имеют ту же размерность, что и  $f_{x_1 x_2}$ , поэтому есть основание ожидать, что в разумной формуле численного дифференцирования (10) коэффициенты  $a, b, c, d$  будут безразмерными величинами.

Положим  $R(f) = f_{x_1 x_2}(0, 0) - l(f)$ . Выпишем разложение Тейлора для  $f(x_1, x_2)$  относительно точки  $(0, 0)$  с точностью до членов второго порядка:

$$f(x_1, x_2) = P_2(x_1, x_2) + r(f),$$

$$P_2(x_1, x_2) = f(0, 0) + x_1 f_{x_1}(0, 0) + x_2 f_{x_2}(0, 0) +$$

$$+ \frac{1}{2} x_1^2 f_{x_1 x_1}(0, 0) + x_1 x_2 f_{x_1 x_2}(0, 0) + \frac{1}{2} x_2^2 f_{x_2 x_2}(0, 0).$$

Справедливы соотношения  $(r(f))_{x_1 x_2}|_{(0,0)} = 0$  и  $r(f) = O(h^3)$  в узлах сетки, входящих в выражение  $l(f)$ . Поэтому в предположении, что  $a, b, c, d = O(1)$ , имеем

$$R(f) = (P_2)_{x_1 x_2}|_{(0,0)} + 0 - l(P_2) - l(r(f)) = R(P_2) - O(h^3)/(h_1 h_2).$$

Если  $R(P_2) = 0$  и  $h_1, h_2$  одного порядка, то  $R(f) = O(h)$  и формула численного дифференцирования (10) имеет порядок  $O(h)$ . Выражение  $R(f)$  является линейным функционалом от функции  $f$ , поэтому  $R(P_2) = 0$  для любого многочлена  $P_2$  второй степени, если

$$R(1) = R(x_1) = R(x_2) = R(x_1^2) = R(x_1 x_2) = R(x_2^2) = 0.$$

Получаем систему уравнений

$$R(1) = -(a + b + c + d)/(h_1, h_2) = 0,$$

$$R(x_1) = -(d + b)/h_2 = 0,$$

$$R(x_2) = -(d + c)/h_1 = 0,$$

$$R(x_1^2) = -(d + b) h_1/h_2 = 0,$$

$$R(x_2^2) = -(d + c) h_2/h_1 = 0,$$

$$R(x_1 x_2) = 1 - d = 0.$$

Эта система шести линейных уравнений с четырьмя неизвестными имеет решение  $a = d = 1, b = c = -1$ , соответствующее приближенной формуле (10).

Точно таким же образом строятся формулы численного интегрирования. Пусть вычисляется интеграл

$$I_0 = \int_G f(y_1, \dots, y_s) dy_1 \dots dy_s.$$

Его можно представить в виде

$$I_0 = \int_G I_1(y_1) dy_1, \quad I_1(y_1) = \int_{G_1(y_1)} I_2(y_1, y_2) dy_2,$$

$$I_2(y_1, y_2) = \int_{G_2(y_1, y_2)} I_3(y_1, y_2, y_3) dy_3, \dots, I_{s-1}(y_1, \dots, y_{s-1}) =$$

$$= \int_{G_{s-1}(y_1, \dots, y_{s-1})} I_s(y_1, \dots, y_s) dy_s.$$

Здесь  $I_s(y_1, \dots, y_s) = f(y_1, \dots, y_s)$ ,  $G_0$  — проекция области  $G$  на ось  $y_1$ ,  $G_k(\bar{y}_1, \dots, \bar{y}_k)$  — сечение множества  $G$  плоскостью  $y_1 = \bar{y}_1, \dots, y_k = \bar{y}_k$ . Воспользуемся какой-либо квадратурной формулой для вычисления первого из этих интегралов

$$I_0 \approx \sum_{m_1=1}^{n_1} D^{m_1} I_1(y_1^{m_1}).$$



Задача вычисления исходного интеграла свелась к вычислению интегралов размерности, на 1 меньшей. Полагаем теперь

$$I_1(y_1^{m_1}) \approx \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} I_2(y_1^{m_1}, y_2^{m_1 m_2}),$$

(11)

$$\dots$$

$$I_{s-1}(y_1^{m_1}, \dots, y_{s-1}^{m_{s-1}}) \approx \sum_{m_s=1}^{n_s(m_1, \dots, m_{s-1})} D^{m_1, \dots, m_s} \times$$

$$\times f(y_1^{m_1}, \dots, y_s^{m_s}, \dots, m_s).$$

В итоге получаем

$$I_0 \approx \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \dots$$

$$\dots \sum_{m_s=1}^{n_s(m_1, \dots, m_{s-1})} D^{m_1, \dots, m_s} f(y_1^{m_1}, \dots, y_s^{m_s}, \dots, m_s). \quad (12)$$

В одномерном случае для всех операций численного анализа были получены оценки погрешности через производные рассматриваемой функции. Посмотрим, какие оценки погрешности являются следствием этих оценок для многомерных задач. Значение некоторого оператора от функции приближается значениями других операторов, причем погрешность этой замены можно оценить согласно формулам оценки погрешности одномерных формул. Однако эти новые значения сами уже содержат погрешности, поэтому в суммарную погрешность эти погрешности войдут с некоторыми множителями.

Обратимся к задаче интегрирования. Пусть

$$I_{k-1}(y_1^{m_1}, \dots, y_{k-1}^{m_{k-1}}) =$$

$$= \sum_{m_k=1}^{n_k(m_1, \dots, m_{k-1})} D^{m_1, \dots, m_k} I_k(y_1^{m_1}, \dots, y_k^{m_1 m_k}, \dots, m_k) + E_{m_1, \dots, m_{k-1}}^{k-1}.$$

Последовательно подставляя в равенство

$$I_0 = \sum_{m_1=1}^{n_1} D^{m_1} I_1(y_1^{m_1}) + E^0$$

выражения  $I_1, I_2, \dots$ , получаем цепочку соотношений

$$I_0 = E^0 + \sum_{m_1=1}^{n_1} D^{m_1} I_1(y_1^{m_1}) =$$

$$= E^0 + \sum_{m_1=1}^{n_1} D^{m_1} \left( E_{m_1}^1 + \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} I_2(y_1^{m_1}, y_2^{m_1 m_2}) \right) = \dots$$

$$\dots = R + \sum_{m_1=1}^{n_1} D^{m_1} \left( \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \times \right.$$

$$\times \left( \dots \left( \sum_{m_s=1}^{n_s(m_1, \dots, m_{s-1})} D^{m_1, \dots, m_s} f(y_1^{m_1}, \dots, y_s^{m_s}, \dots, m_s) \right) \dots \right)).$$

где

$$R = E^0 + \sum_{m_1=1}^{n_1} D^{m_1} E^1_{m_1} + \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1, m_2} E^2_{m_1, m_2} + \dots$$

$$\dots + \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1, m_2} \dots \sum_{m_{s-1}=1}^{n_{s-1}(m_1, \dots, m_{s-2})} D^{m_1, \dots, m_{s-1}} E^{s-1}_{m_1, \dots, m_{s-1}}.$$

(13)

Из последнего равенства видно, что погрешность аппроксимации может оказаться существенно больше, чем в одномерном случае, если коэффициенты  $D^{m_1, \dots, m_k}$  — большие.

Формально сведение многомерной задачи к одномерной имеет одинаковый характер и в случае задачи численного дифференцирования (интерполирования), и в случае численного интегрирования. Однако между этими задачами есть такое существенное различие: задача численного дифференцирования (интерполирования) чаще ставится как задача нахождения оператора от функции по значениям на некоторой заданной совокупности узлов  $\Omega$ . Для задачи интегрирования более типичной является возможность распорядиться выбором узлов.

При осуществлении многомерных операций численного дифференцирования (интерполирования и интегрирования) функций, заданных на сетке, являющейся произведением одномерных сеток, разумно использовать одинаковые формулы для аппроксимации промежуточных величин. Имеется в виду, что, например, формулы (11) должны иметь вид

$$I_{k-1}(y_1^{m_1}, \dots, y_{k-1}^{m_{k-1}}) \approx \sum_{m_k=1}^{N_k} D_{k, N_k}^{m_k} I_k(y_1^{m_1}, \dots, y_k^{m_k}), \quad (14)$$

т. е.  $D_{k, N_k}^{m_k}$  и  $y_k^{m_k}$  зависят только от  $m_k$ . Правая часть (12) тогда приобретает вид

$$\sum_{m_1=1}^{N_1} \dots \sum_{m_s=1}^{N_s} D_{1, N_1}^{m_1} \dots D_{s, N_s}^{m_s} f(y_1^{m_1}, \dots, y_s^{m_s}).$$

При построении такой квадратуры мы неявно предполагаем, что область интегрирования — прямоугольный параллелепипед с ребрами, параллельными координатным осям.

Такие формулы численного интегрирования (интерполирования, дифференцирования) называют *прямым произведением* соответствующих одномерных формул численного интегрирования (интерполирования, дифференцирования).

В § 11 будет приведен более сложный пример прямого произведения квадратуры по отрезку на квадратуру по сфере. В случае применения таких аппроксимаций так же, как при вычислении производной  $f_{x_1, x_2}$  по формуле (9), оказывается, что остаточные члены аппроксимации компенсируются.

В случае задачи интегрирования возможна такая ситуация: при гладкой подынтегральной функции может оказаться, что промежуточные интегралы  $I_k(y_1, \dots, y_k)$  не обладают достаточной гладкостью.

Пусть вычисляется интеграл по единичному кругу

$$I = \int_{-1}^1 \left( \int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} f(y_1, y_2) dy_2 \right) dy_1$$

при гладкой функции  $f(y_1, y_2)$ . Если  $f(\pm 1, 0) \neq 0$ , то функция

$$I_1(y_1) = \int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} f(y_1, y_2) dy_2$$

имеет неограниченные производные в точках  $\pm 1$  и поэтому при численном интегрировании по переменной  $y_1$  следует использовать специальные приемы вычисления интегралов от таких функций — переменный шаг интегрирования, в частности интегрирование с автоматическим выбором шага, выделение особенностей и т. д. Целесообразнее записать этот интеграл в виде

$$I = \int_0^1 r I_1(r) dr,$$

$$I_1(r) = \int_0^{2\pi} f(r \cos \varphi, r \sin \varphi) d\varphi.$$

Подынтегральная функция внутреннего интеграла периодическая, поэтому имеет смысл применять квадратуру (3.5.7)

$$\int_0^{2\pi} g(\varphi) d\varphi \approx \frac{2\pi}{n} \sum_{m=0}^{n-1} g\left(\frac{2\pi m}{n}\right).$$

При малом  $r$  производные функции  $f$  по переменной  $\varphi$  содержат малый множитель  $r$ , поэтому при малых  $\varphi$  иногда выгоднее взять число  $n$  поменьше. Для вычисления интеграла по переменной  $r$  можно применить любые из рассматривавшихся ранее квадратурных формул, в частности формулы Гаусса, соответствующие весовой функции  $r$ .

**Задача 1.** Функция задана в узлах сетки  $(m_1 h_1, m_2 h_2)$ . Построить формулы с погрешностями  $O(h_1^2 + h_2^2)$ ,  $O(h_1^4 + h_2^4)$  для вычисления значения  $i_{x_2}(h_1/2, 0)$ .

## § 5. Интерполяция функций в треугольнике

При решении уравнений с частными производными вариационно-разностными методами возникает следующая задача. Имеется некоторый треугольник  $\Delta$ , каждая сторона которого разбита на  $l$  равных частей, и через точки разбиения проведены прямые  $L_q$ , параллельные сторонам треугольника. Стороны треугольника также будем относить к множеству прямых  $L_q$ . Обозначим через  $\Omega$  множество, состоящее из точек пересечения этих прямых, лежащих в замкнутом треугольнике  $\Delta$ . (Таким образом,

$\Omega$  включает также точки разбиения сторон треугольника и вершины треугольника.) Число таких точек равно  $n = 1 + 2 + \dots + (l + 1) = (l + 1)(l + 2)/2$ . Будем обозначать их через  $Q_1(x_1^1, x_2^1), \dots, Q_n(x_1^n, x_2^n)$ .

Ставится задача построения многочлена степени  $l$

$$P(x_1, x_2) = \sum_{m_1 + m_2 \leq l} a_{m_1, m_2} x_1^{m_1} x_2^{m_2},$$

принимающего в этих точках  $Q_j(x_1^j, x_2^j)$  заданные значения

$$P(x_1^j, x_2^j) = f_j, \quad j = 1, \dots, n. \quad (1)$$

Число неизвестных коэффициентов  $a_{m_1, m_2}$  также равно  $n$ , и, таким образом, соотношения (1) образуют систему  $n$  уравнений с  $n$  неизвестными. Если система (1) разрешима, то из нее могут быть найдены коэффициенты  $a_{m_1, m_2}$ . Для их нахождения не обязательно прибегать к описанному выше варианту метода неопределенных коэффициентов, а можно выписать искомым многочлен  $P(x)$  в явном виде.

Возьмем некоторую фиксированную точку  $Q_j$ . Можно показать, что среди прямых  $L_q$  имеется ровно  $l$  прямых, удовлетворяющих следующему условию. Существует не более одной вершины треугольника такой, что  $Q_j$  и эта вершина лежат по одну сторону от такой прямой. При этом оказывается, что каждая точка из  $\Omega$ , отличная от  $Q_j$ , лежит на одной из таких прямых. На рис. 5.5.1 эти прямые обозначены жирными линиями.

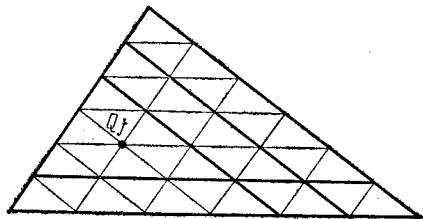


Рис. 5.5.1

Пусть  $L_{j,1}(x_1, x_2) = 0, \dots, L_{j,l}(x_1, x_2) = 0$  — уравнения этих прямых. Функция

$$\varphi_j(x_1, x_2) = \prod_{i=1}^l \frac{L_{j,i}(x_1, x_2)}{L_{j,i}(x_1^j, x_2^j)}$$

является многочленом степени  $l$ , равна 1 в точке  $Q_j$  и 0 в остальных точках  $Q_i$ . Поэтому многочлен степени  $l$

$$P(x_1, x_2) = \sum_{j=1}^n f_j \varphi_j(x_1, x_2)$$

будет искомым.

В случае, когда  $f_j = f(x_1^j, x_2^j)$  при всех  $j$ , многочлен  $P(x_1, x_2)$  будет интерполяционным многочленом по отношению к функции  $f(x_1, x_2)$ .

**Задача 1.** Показать, что значения многочлена  $P(x_1, x_2)$  на каждой из сторон треугольника зависят от значений  $f_j$ , соответствующих точкам  $Q_j$  этой стороны.

**Задача 2.** Пусть  $H$  — длина максимальной из сторон треугольника  $\Delta$ ,  $f_l = f(Q_l)$ ,  $f$  — некоторая гладкая функция

$$M_{l+1} = \max_{r_1+r_2=l+1} \max_{\Delta} |f^{(r_1, r_2)}(x_1, x_2)|$$

$\alpha$  — наименьший из углов треугольника. Получить оценку

$$\max_{\Delta} |f(x_1, x_2) - P(x_1, x_2)| \leq C(l, \alpha) M_{l+1} H^{l+1},$$

где  $C$  — постоянная, зависящая только от  $l$  и  $\alpha$ .

**Задача 3.** При выполнении условия задачи 2 для  $0 < r_1 + r_2 \leq l + 1$  получить оценку

$$\max_{\Delta} |f^{(r_1, r_2)}(x_1, x_2) - P^{(r_1, r_2)}(x_1, x_2)| \leq C(l, r_1, r_2, \alpha) M_{l+1} H^{l+1-r_1-r_2}.$$

**Задача 4.** Исследовать поведение постоянных  $C(l, \alpha)$ ,  $C(l, r_1, r_2, \alpha)$  — при  $\alpha \rightarrow 0$  в задачах 2 и 3 соответственно.

Описанный способ интерполяции широко используется при приближении функций двух переменных. Область  $G$ , где приближается функция, разбивается на треугольники с достаточно малой максимальной стороной. В каждом треугольнике функция приближается соответствующим интерполяционным многочленом  $P(x_1, x_2)$  степени  $l$ . Если разбиение устроено так, что вершина одного треугольника не может быть внутренней точкой стороны другого треугольника, то полученная таким образом приближающая функция будет непрерывна в  $G$  (справедливость последнего утверждения следует из решения задачи 1).

В связи с неограниченным ростом постоянных  $C(l, r_1, r_2, \alpha)$  при  $\alpha \rightarrow 0$  (см. задачу 3) разумное разбиение области  $G$  не должно содержать треугольников  $\Delta$  с очень малыми углами.

**Задача 5.** Построить аналог описанного выше способа интерполяции для задачи интерполяции функции в тетраэдре.

В случае, когда две стороны треугольника  $\Delta$  или соответственно три ребра тетраэдра направлены вдоль координатных осей, интерполирующий многочлен может быть явно записан также с помощью аппарата разделенных разностей для функций многих переменных.

## § 6. Оценка погрешности численного интегрирования на равномерной сетке

Произведем конкретную оценку погрешности численного интегрирования в случае двукратного интеграла от дважды непрерывно дифференцируемой функции

$$I = \int_0^1 \int_0^1 f(x_1, x_2) dx_1 dx_2.$$

Запишем исходный интеграл как повторный:

$$I = \int_0^1 I_1(x_1) dx_1, \quad (1)$$

где

$$I_1(x_1) = \int_0^1 f(x_1, x_2) dx_2.$$

Введем обозначения

$$A_k = \max_{0 \leq x_1, x_2 \leq 1} \left| \frac{\partial^2 f}{\partial x_k \partial x_k} \right|.$$

Применим для вычисления интеграла (1) составную формулу трапеций с постоянным шагом разбиения  $H_1 = 1/N_1$ :

$$I \approx \frac{I_1(0) + I_1(1)}{2N_1} + \sum_{j_1=1}^{N_1-1} \frac{I_1(j_1/N_1)}{N_1}.$$

Поскольку

$$\frac{\partial^2 I_1(x_1)}{\partial x_1^2} = \int_0^1 \frac{\partial^2 f(x_1, x_2)}{\partial x_1^2} dx_2,$$

то

$$\left| \frac{\partial^2 I_1(x_1)}{\partial x_1^2} \right| \leq A_1.$$

Поэтому при  $E^0 = I - S_{N_1}$ , согласно оценке составной формулы трапеций (3.8.8), имеем

$$|E^0| \leq A_1/(12N_1^2). \quad (2)$$

Для вычисления интегралов  $I_1(j_1/N_1)$  применим составную формулу трапеций (3.8.8), имеем

$$I_1\left(\frac{j_1}{N_1}\right) \approx S_{N_2}\left(\frac{j_1}{N_1}\right) = \frac{f(j_1/N_1, 0) + f(j_1/N_1, 1)}{2N_2} + \sum_{j_2=1}^{N_2-1} \frac{f(j_1/N_1, j_2/N_2)}{N_2}.$$

Для погрешности

$$E_{j_1}^1 = I_1(j_1/N_1) - S_{N_2}(j_1/N_1)$$

также согласно (3.8.8) имеем

$$|E_{j_1}^1| \leq A_2/(12N_2^2). \quad (3)$$

Подставляя  $I_1(j_1/N_1) = E_{j_1}^1 + S_{N_2}(j_1/N_1)$  в равенство  $I = E^0 + S_{N_1}$ , получим цепочку соотношений

$$I = E^0 + \frac{(S_{N_2}(0) + E_0^1) + (S_{N_2}(1) + E_{N_1}^1)}{2N_1} + \sum_{j_1=1}^{N_1-1} \frac{S_{N_2}(j_1/N_1) + E_{j_1}^1}{N_1} = S_{N_1 N_2} + R,$$

где

$$S_{N_1 N_2} = \frac{S_{N_2}(0) + S_{N_2}(1)}{2N_1} + \sum_{j_1=1}^{N_1-1} \frac{S_{N_2}(j_1/N_1)}{N_1},$$

$$R = E^0 + \frac{E_0^1 + E_{N_1}^1}{2N_1} + \sum_{j_1=1}^{N_1-1} \frac{E_{j_1}^1}{N_1}. \quad (4)$$

Выражение  $S_{N_1 N_2}$  является квадратурной суммой, вычисляемой по значениям функции  $f$  в  $(N_1 + 1)(N_2 + 1)$  точках  $(j_1/N_1, j_2/N_2)$ ,  $R$  — погрешность численного интегрирования.

Из (4) следует оценка

$$|R| \leq |E^0| + \max_{0 \leq j_1 \leq N_1} |E_{j_1}^1|,$$

и вследствие (2), (3) имеем

$$|R| \leq A_1/(12N_1^2) + A_2/(12N_2^2).$$

При  $N_1 = N_2 = n$  получаем, что

$$|R| \leq (A_1 + A_2)/(12n^2), \quad (5)$$

и по отношению к общему числу узлов интегрирования  $N = (N_1 + 1)(N_2 + 1) = (n + 1)^2$  погрешность имеет порядок  $O(N^{-1})$ .

Предположим, что требуется гарантия, чтобы погрешность не превосходила  $\varepsilon$ . Для этого достаточно выполнения неравенства

$$A_1/(12N_1^2) + A_2/(12N_2^2) \leq \varepsilon. \quad (6)$$

Если  $A_1$  и  $A_2$  отличаются несущественно, то можно взять  $N_1 = N_2 = n$  и исходя из оценки (5) определить минимальное  $n$  из условия

$$(A_1 + A_2)/(12n^2) \leq \varepsilon.$$

Если же  $A_1$  и  $A_2$  различаются существенно, то имеет смысл затратить время на минимизацию вычислительной работы, т. е. искать  $N_1$  и  $N_2$ , минимизирующие  $(N_1 + 1)(N_2 + 1)$  при условии

$$A_1/(12N_1^2) + A_2/(12N_2^2) \leq \varepsilon.$$

Обозначим через  $C_r(\mathbf{A}) = C_{r_1, \dots, r_s}(A_1, \dots, A_s)$  класс функций, у которых в рассматриваемой области определения производные  $f_{x_k^{r_k-1}}, k=1, \dots, s$ , непрерывны, а производные  $f_{x_k^{r_k}}(x_1, \dots, x_s)$  кусочно-непрерывны и удовлетворяют условиям  $|f_{x_k^{r_k}}(x_1, \dots, x_s)| \leq A_k$ .

Задача 1. Пусть для вычисления интеграла

$$\int_0^1 \dots \int_0^1 f(x_1, \dots, x_s) dx_1 \dots dx_s$$

применяются составные формулы точности  $O(N_k^{-r_k})$  по каждой оси, где  $N_k$  — число узлов, соответствующее направлению  $x_k$ . Получить оценку погрешности

$$O\left(\sum_1^k A_k N_k^{-r_k}\right). \quad (7)$$

Минимизируя оценку (7) при заданном общем числе узлов  $N_1, \dots, N_s = N$ , получить оценку погрешности  $O(N^{-r})$ ,  $1/r = 1/r_1 + \dots + 1/r_s$ .

Рассмотрим частный случай:  $r_1 = \dots = r_s = r_0$ ,  $A_1 = \dots = A_s = A_0$ . Тогда  $1/r = 1/r_1 + \dots + 1/r_s = s/r_0$ . Таким образом, рассматриваемые квадратуры обеспечивают оценку погрешности  $O(A_0 N^{-r_0/s})$ . У реальных подынтегральных функций порядок ограниченных производных  $r$  часто оказывается не очень большим, поэтому при больших  $s$ , как показывает полученная оценка, скорость сходимости оказывается плохой.

Возникает вопрос об оптимальных квадратурах на классах многомерных функций. Так как ни для каких реальных классов функций такие квадратуры неизвестны, ограничимся оценкой снизу погрешности оптимальных квадратур.

## § 7. Оценка снизу погрешности численного интегрирования

Напомним постановку задачи оптимизации квадратурных формул на классе функций. Пусть приближенное значение интеграла вычисляется по формуле

$$I(f) = \int_G f(P) p(P) dP \approx S_N(f) = \sum_{j=1}^N D_j f(P_j).$$

Величина

$$R_N(f) = I(f) - S_N(f)$$

называется погрешностью квадратуры, величина

$$R_N(F) = \sup_{f \in F} |R_N(f)|$$

— погрешностью квадратуры на классе функций  $F$ , величина

$$W_N(F) = \inf_{D_j, P_j} R_N(F)$$

— оптимальной оценкой погрешности квадратур на классе  $F$ ; квадратура (если такая существует), на которой эта нижняя грань достигается, называется оптимальной. Мы будем предполагать выполненным условие: существует некоторый куб  $\Delta \in G$ , в котором  $p(P) \geq \gamma > 0$ .

**Теорема.**  $W_N(C_r(\mathbf{A})) \geq d(\Delta, \mathbf{r}, \mathbf{A}) \gamma N^{-r}$ , где  $d(\Delta, \mathbf{r}, \mathbf{A}) > 0$ ,  $r = (r_1^{-1} + \dots + r_s^{-1})^{-1}$ .

Доказательство теоремы осуществляется следующим образом. Будет показано, что для любой совокупности узлов  $P_1, \dots, P_N$  можно построить функцию  $\hat{f}_{P_1, \dots, P_N}(P)$  рассматриваемого класса  $C_r(\mathbf{A})$ , обращающуюся в нуль во всех этих узлах и такую, что  $I(f) \geq d(\Delta, \mathbf{r}, \mathbf{A}) \gamma N^{-r}$ . При этом постоянная  $d > 0$  не зависит от точек  $P_1, \dots, P_N$ . Тогда для любой квадратуры



с этими узлами

$$R_N(C_r(A)) \geq \left| I(f_{P_1, \dots, P_N}) - \sum_{j=1}^N D_j f_{P_1, \dots, P_N}(P_j) \right| = \\ = |I(f_{P_1, \dots, P_N})| \geq d(\Delta, r, A) \gamma N^{-r}.$$

Величина  $R_N(C_r(A))$  оценена снизу постоянной, не зависящей от узлов квадратур  $P_j$  и весов  $D_j$ , поэтому и ее нижняя грань  $W_N(C_r(A))$  по множеству всевозможных квадратур также оценивается снизу этой постоянной  $d(\Delta, r, A) \gamma N^{-r}$ . Таким образом, доказательство теоремы сводится к построению соответствующей функции  $f_{P_1, \dots, P_N}(P)$  для каждой совокупности точек  $P_1, \dots, P_N$ .

Построение такой функции  $f_{P_1, \dots, P_N}(P)$  и доказательство

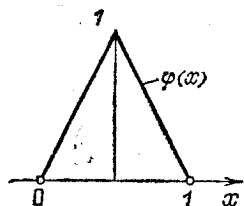


Рис. 5.7.1

утверждения теоремы для простоты будут проведены для случая  $r = (1, \dots, 1)$ ,  $A_1 = \dots = A_s = A_0$ , т. е. для класса непрерывных функций с кусочно-непрерывными первыми производными  $\partial f / \partial x_i$ , ограниченными по модулю постоянной  $A_0$ .

Для простоты выкладки будем проводить построение для случая, когда  $\Delta$  — единичный куб  $0 \leq x_1, \dots, x_s \leq 1$ . Положим  $n = [(2N)^{1/s}] + 1$  и разобьем куб  $\Delta$  на  $n^s$  прямоугольных параллелепипедов  $\Pi_n = \Pi_{n_1 \dots n_s} : (n_k - 1)/n \leq x_k \leq n_k/n, 0 < n_k \leq n, k = 1, \dots, s$ . Пусть (см. рис. 5.7.1)

$$\varphi(x) = \begin{cases} 0 & \text{при } x \leq 0 \text{ или } x \geq 1, \\ 1 - 2|x - 0,5| & \text{при } 0 \leq x \leq 1. \end{cases}$$

Из определения  $n$  следует, что  $(2N)^{1/s} \leq n \leq 2(2N)^{1/s}$ , поэтому

$$2N \leq n^s \leq 2^{s+1}N. \quad (1)$$

Построим функцию  $f_{P_1, \dots, P_N}(P) = f_0(P)$  следующим образом. В тех параллелепипедах  $\Pi_n$ , которые не содержат внутри ни одной из точек  $P_1, \dots, P_s$ , положим

$$f_0(P) = \frac{A_0}{2n} \prod_{k=1}^s \varphi(nx_k - (n_k - 1)). \quad (2)$$

Во всех остальных параллелепипедах положим  $f_0(P) = 0$ .

Функция  $f_0(P)$  согласно своему определению непрерывна в каждом кубе  $\Pi_n$  и обращается в нуль на его границе, поэтому она непрерывна в  $\Delta$ . В точках, где функция  $f_0(P)$  дифференцируема и отлична от 0, имеем

$$\frac{\partial f_0}{\partial x_1} = \frac{A_0}{2} \varphi'(nx_1 - (n_1 - 1)) \prod_{k=1}^s \varphi(nx_k - (n_k - 1)).$$

Поскольку  $|\varphi| \leq 1$ ,  $|\varphi'| \leq 2$ , то  $|\partial f_0 / \partial x_1| \leq A_0$ . Точно так же получаем оценки, что  $|\partial f_0 / \partial x_k| \leq A_0$ ,  $k = 1, 2, \dots, s$ . Согласно построению функции  $f_0(P)$  производные  $\partial f_0 / \partial x_k$  могут иметь разрывы лишь в точках плоскостей  $x_k = m / (2n)$ , где  $m$  — целое; поэтому функция  $f_0(P)$  принадлежит рассматриваемому классу. Из ее построения также следует, что она обращается в нуль со всеми производными во всех узлах  $P_1, \dots, P_N$ .

Оценим снизу значение  $I(f_0)$ . Воспользуемся неравенством  $p(P) \geq \gamma$ , после замены переменных  $x_k = (n_k - 1 + y_k) / n$  получим цепочку соотношений

$$\int_{\Pi_n} p(x_1, \dots, x_s) \frac{A_0}{2n} \prod_{k=1}^s \varphi(nx_k - (n_k - 1)) dx_1 \dots dx_s \geq \\ \geq \frac{A_0 \gamma}{2n^{s+1}} \int_0^1 \dots \int_0^1 \prod_{k=1}^s \varphi(y_k) dy_1 \dots dy_s = \frac{d_1 A_0 \gamma}{n^{s+1}};$$

здесь

$$d_1 = \frac{1}{2} \left( \int_0^1 \varphi(y) dy \right)^s = \frac{1}{2^{s+1}}. \quad (3)$$

Каждая точка  $P_j$  может находиться внутри только одного из параллелепипедов  $\Pi_n$ . Следовательно, по крайней мере в  $n^s - N$  параллелепипедах  $\Pi_n$  функция  $f_0(P)$  отлична от тождественного нуля и определяется равенством (2). Согласно (1) имеем  $N \leq n^s / 2$ , поэтому таких параллелепипедов не менее чем  $n^s / 2$ . С учетом (3) получаем оценку

$$I(f_0) \geq \frac{n^s}{2} \frac{d_1 A_0 \gamma}{n^{s+1}} = \frac{d_1 A_0 \gamma}{2n} \geq \frac{d_1 A_0 \gamma}{2 \cdot 2 \cdot (2N)^{1/s}} = \frac{d A_0 \gamma}{N^{1/s}}, \quad (4)$$

где  $d = d_1 / (4 \cdot 2^{1/s})$ .

Таким образом, построенная функция принадлежит рассматриваемому классу, для нее выполняется неравенство (4) с постоянной  $d$ , не зависящей от узлов интегрирования, она обращается в нуль во всех узлах  $P_j$ . Следовательно, способ построения требуемой функции для любой совокупности узлов  $P_1, \dots, P_N$  указан.

Выше производилась оценка снизу погрешности квадратур, т. е. формул интегрирования

$$I(f) \approx \sum_{j=1}^N D_j f(P_j),$$

где узлы интегрирования  $P_j$  и веса  $D_j$  не зависят от конкретной подынтегральной функции. Для многих задач практически более эффективными являются способы интегрирования, где последующие узлы интегрирования выбираются в зависимости от информации об уже вычисленных значениях функции, например способы интегрирования с автоматическим выбором шага.

Пусть имеется какой-то способ интегрирования, где информация о подынтегральной функции учитывается лишь в виде информации о ее значениях в отдельных точках. Этот способ определяется заданием первого узла интегрирования, правила, по которому отыскиваются следующие узлы, и способа вычисления приближенного значения интеграла. Таким образом, всякий такой способ укладывается в следующую схему: задается некоторый узел  $P_1$ , функции

$$Q_q = \Phi_q(Q_1, \dots, Q_{q-1}; y_1, \dots, y_{q-1}), \quad q = 2, \dots, N,$$

определяющие выбор следующих узлов интегрирования в зависимости от ранее накопленной информации о подынтегральной функции, и функция

$$S_N(Q_1, \dots, Q_N; y_1, \dots, y_N).$$

Здесь  $Q_j$  — точки области  $G$ ,  $y_j$  — числа. При приближенном вычислении конкретного интеграла последовательно вычисляются величины

$$\begin{aligned} f(P_1), P_2 &= \Phi_2(P_1; f(P_1)), f(P_2), \\ P_3 &= \Phi_3(P_1, P_2; f(P_1), f(P_2)), f(P_3), \dots, f(P_N) \end{aligned}$$

и затем полагают

$$I(f) \approx S_N(P_1, \dots, P_N; f(P_1), \dots, f(P_N)).$$

Поскольку точка  $P_1$  задается вместе с заданием функций  $\Phi_j$  и  $S_N$ , зависимость этих функций от  $P_1$  можно было бы опустить.

Так же, как в случае квадратурных формул, можно определить погрешность метода при вычислении данного интеграла

$$\tilde{R}_N(f) = I(f) - S(P_1, \dots, P_N; f(P_1), \dots, f(P_N)),$$

погрешность метода на классе

$$\tilde{R}_N(F) = \sup_{f \in F} |\tilde{R}_N(f)|$$

и оптимальную оценку погрешности на множестве всевозможных методов интегрирования

$$Z_N(F) = \inf_{P_1, \Phi_2, \dots, \Phi_N, S_N} \tilde{R}_N(F).$$

**Теорема** (без доказательства). Пусть класс функций  $F$  — выпуклый центрально симметричный компакт с центром симметрии  $f \equiv 0$  и все функции этого класса равномерно ограничены. Тогда оптимальные оценки погрешности на множествах всевозможных квадратурных формул и всевозможных способов интегрирования совпадают:

$$W_N(F) = Z_N(F).$$

Условие выпуклости класса  $F$  означает, что вместе с любыми  $f_1, f_2 \in F$  этому классу принадлежат также все функции

$f = \theta f_1 + (1 - \theta) f_2$  при  $0 < \theta < 1$ . Условие, что класс является *центрально симметричным* с центром симметрии  $f \equiv 0$ , означает, что вместе со всякой функцией  $f$  этому классу принадлежит также функция  $-f$ .

В частности, все классы  $C_r(A)$  удовлетворяют условиям теоремы и может показаться, что рассмотрение более широкого множества способов интегрирования не представляет интереса. Однако такой вывод нельзя признать правильным. Методы интегрирования с выбором узлов интегрирования в зависимости от полученной в процессе вычислений информации продемонстрировали на практике свою высокую эффективность по сравнению с квадратурными формулами с заранее фиксированными узлами интегрирования. Поэтому более правильным будет вывод о том, что практически встречающиеся задачи более точно описываются некоторыми невыпуклыми классами функций. Например, типичным классом функций, встречающихся в приложениях, является класс кусочно-аналитических функций. Обратимся к одномерному случаю. В случае выпуклости класса функций полусумма двух функций класса также принадлежит этому классу. Полусумма функций, имеющих  $l$  точек, где нарушается аналитичность, может иметь  $2l$  точек с нарушением аналитичности. Таким образом, класс аналитических функций, имеющих не более заданного числа  $l$  точек нарушения аналитичности, не является выпуклым. Если число точек нарушения аналитичности не ограничено сверху, то класс функций не является замкнутым; предел последовательности кусочно-аналитических функций с неограниченно растущим числом точек нарушения аналитичности может оказаться функцией, не являющейся кусочно-аналитической.

Выше производилось сравнение методов интегрирования по верхней грани погрешности на классе функций. Однако возможна следующая ситуация. Два метода имеют одинаковые погрешности на классе функций, в то же время на большинстве функций класса один из методов имеет меньшую погрешность. Ясно, что этот метод является более предпочтительным и сравнение методов по верхней грани погрешности на классе функций в данном случае не дает общей картины.

Из сказанного выше вытекает актуальность решения следующих задач.

**Задача 1.** Как правильно описать класс реально встречающихся функций?

**Задача 2.** Как правильно ввести меру в пространстве реально встречающихся подынтегральных функций? (Ни про одно из известных определений меры в пространстве функций нельзя сказать, что оно правильно описывает обстановку, характерную для приложений.)

Приведем примеры некоторых алгоритмов интегрирования с выбором узлов в зависимости от ранее полученной информации.

Многомерный интеграл записывается как повторный:

$$I(f) = \int_{a_1}^{b_1} I_1(x_1) dx_1,$$

$$I_1(x_1) = \int_{a_1(x_1)}^{b_1(x_1)} I_2(x_1, x_2) dx_2,$$

.....

$$I_{s-1}(x_1, \dots, x_{s-1}) = \int_{a_{s-1}(x_1, \dots, x_{s-1})}^{b_{s-1}(x_1, \dots, x_{s-1})} f(x_1, \dots, x_s) dx_s,$$

и численное интегрирование по некоторым из переменных  $x_j$  производится посредством одномерных алгоритмов интегрирования с автоматическим выбором шага.

Следующий алгоритм вычисления многомерных интегралов имеет другую структуру. Среди известных алгоритмов интегрирования с автоматическим выбором узлов интегрирования этот алгоритм является наиболее эффективным по отношению к задаче вычисления интегралов от функций с особенностями функции или ее производных в изолированных точках.

Пусть вычисляется интеграл

$$\int_G f(X) dX, \quad \Omega = [a_1 \leq x_1 \leq b_1, \dots, a_s \leq x_s \leq b_s], \quad X = (x_1, \dots, x_s)$$

Заменой переменных

$$x_i = 0,5(a_i + b_i) + 0,5(b_i - a_i)t_i, \quad i = 1, \dots, s,$$

интеграл превращается в интеграл по кубу

$$\int_G f(x) dx = \int_G g(t) dt,$$

$$G = [-1 \leq t_1 \leq 1, \dots, -1 \leq t_s \leq 1], \quad t = (t_1, \dots, t_s).$$

В основу метода положены кубатурные формулы

$$\int_G g(t) dt \approx Q_q^s(g), \quad q = 1, 2, 3.$$

Для вычисления интеграла  $\int_G g(t) dt$  вычисляются  $Q_1^s(g)$  и  $Q_2^s(g)$

и проверяется условие

$$|Q_1^s(g) - Q_2^s(g)| \leq \epsilon. \quad (5)$$

Здесь  $\epsilon$  — некоторая условная мера погрешности. Если это условие выполнено, то за приближенное значение интеграла по  $G$

принимается значение, вычисленное по кубатурной формуле  $Q_3^s$ , обычно являющейся линейной комбинацией формул  $Q_1^s$  и  $Q_2^s$ . Если условие (5) не выполняется, то куб  $G$  разбивается на  $2^s$  равных кубов и описанный алгоритм применяется к каждому из этих кубов. Процесс дробления продолжается до тех пор, пока условие (5) не будет выполнено. Если при делении шага  $h$  пополам наступит такой момент, когда  $h^s$  станет машинным нулем, то счет прекращается. Используемые в стандартных программах кубатурные формулы  $Q_q^s$  имеют следующий вид.

I.  $s = 2$ :

$$Q_1^2 = A_1 \sum_{|i|+|j|=1} g(i\alpha, j\alpha) + A_2 \sum_{|i|+|j|=1} g(i\beta, j\beta) + A_3 \sum_{|i|, |j|=1} g(i\gamma, j\gamma),$$

$$Q_2^2 = B_1 \sum_{|i|+|j|=1} g(i\alpha, j\alpha) + B_2 \sum_{|i|+|j|=1} g(i\beta, j\beta) + B_3 \sum_{|i|, |j|=1} g(i\gamma, j\gamma) + B_4 \sum_{|i|, |j|=1} g(i\nu, j\nu),$$

где

$$\alpha = 0,658149897623035910, \quad \beta = 0,549119831921783496,$$

$$\gamma = 0,894427190999915878, \quad \nu = 0,316227766016837933,$$

$$A_1 = 1,06136206790541224, \quad A_2 = -0,234973179016523356,$$

$$A_3 = 0,1736111111111111111,$$

$$B_1 = 3,99942795838189963, \quad B_2 = -6,29803906949301074,$$

$$B_3 = 0,124007936507936507, \quad B_4 = 3,17460317460317460,$$

$$Q_3^2 = Q_2^2.$$

Формула  $Q_1^2$  точна для всех многочленов степени не больше 5.

Формула  $Q_2^2$  точна для всех многочленов степени не больше 7.

II.  $s = 3$ :

$$Q_1^3 = \frac{8}{225} \left( 44g(0, 0, 0) + \frac{121}{8} \sum_{|i|, |j|, |k|=1} g\left(i\sqrt{\frac{5}{11}}, j\sqrt{\frac{5}{11}}, k\sqrt{\frac{5}{11}}\right) + 10 \sum_{|i|+|j|+|k|=1} g(i, j, k) \right),$$

$$Q_2^3 = \frac{8}{1125} \left( -\frac{1552}{5} g(0, 0, 0) + \frac{1573}{40} \sum_{|i|, |j|, |k|=1} g\left(i\sqrt{\frac{5}{11}}, j\sqrt{\frac{5}{11}}, k\sqrt{\frac{5}{11}}\right) + \frac{784}{5} \sum_{|i|+|j|+|k|=1} g\left(i\sqrt{\frac{5}{14}}, j\sqrt{\frac{5}{14}}, k\sqrt{\frac{5}{14}}\right) + 15T_h \right),$$

$$T_h = \sum_{|i|, |j|=1} g(i, j, 0) + \sum_{|i|, |k|=1} g(i, 0, k) + \sum_{|j|, |k|=1} g(0, j, k),$$

$$Q_3^3 = \frac{4}{9} Q_1^3 + \frac{5}{9} Q_2^3.$$

Формулы  $Q_1^3$  и  $Q_2^3$  точны для всех многочленов степени не больше 5. Формула  $Q_3^3$  точна для всех многочленов степени не больше 7.

## § 8. Метод Монте-Карло

При построении квадратурных формул вместе с формулой, как правило, получалась оценка ее погрешности на некотором классе функций. Например, для одномерной формулы трапеций была получена оценка погрешности вида  $\text{const } A_2 N^{-2}$  на классе функций со второй производной, ограниченной постоянной  $A_2$ ; здесь  $N$  — число узлов интегрирования. Такого рода оценки погрешности называют *гарантированными оценками погрешности на классе функций*. На основании этой оценки можно гарантировать, что погрешность приближенного значения интеграла не превосходит определенной величины для всех подынтегральных функций из рассматриваемого класса. При таком подходе, как оценка погрешности метода на классе функций, в случае оценки погрешности конкретного интеграла мы ориентируемся на величину погрешности в случае интегрирования «худшей» функции рассматриваемого класса. Для ряда классов функций эта оценка погрешности на классе настолько плоха, что не дает никакой надежды на получение приближенного значения интеграла с требуемой точностью. Например, согласно теореме из § 7, не существует методов с оценкой погрешности на классе функций  $C_{1, \dots, 1}(A, \dots, A)$ , лучшей, чем  $dAN^{-1/s}$  (здесь  $s$  — кратность вычисляемого интеграла).

Предположим, что требуется гарантировать оценку погрешности, меньшую чем  $0,01dA$ . Тогда число узлов  $N$  должно удовлетворять неравенству  $dAN^{-1/s} \leq 0,01dA$ , т. е. должно выполняться неравенство  $N \geq 100^s$ . Поскольку вычисление каждого значения подынтегральной функции требует обычно большого числа арифметических операций, уже при  $s = 6$  такое требование на число узлов оказывается практически невыполнимым.

Мы оказались в положении, когда на основе указанной выше оценки погрешности нет возможности вычислить значение интеграла с гарантированной оценкой погрешности  $0,01dA$ , поскольку это потребует непомерных затрат времени ЭВМ. Один из способов разрешения создавшегося противоречия состоит в более детальном описании классов подынтегральных функций. Другим выходом из создавшейся обстановки является отказ от получения строгой гарантированной оценки погрешности и получение оценки погрешности лишь с определенной степенью достоверности. В частности, при конструировании методов интегрирования в гл. III и в § 7 мы шли по пути отказа от строгой оценки погрешности: погрешность оценивалась через разность результатов вычисления приближенного значения интеграла при различных методах интегрирования.

Одним из методов приближенного вычисления значений интегралов, при котором погрешность оценивается не гарантированно, а лишь с некоторой степенью достоверности, является метод Монте-Карло.

Пусть требуется вычислить приближенное значение интеграла

$$I(f) = \int_G f(P) dP.$$

Для упрощения выкладок предполагаем, что  $\mu(G)$  — мера области  $G$  — равна 1. Как правило, это условие бывает выполнено, поскольку при практической реализации метода Монте-Карло область интегрирования обычно преобразуется в единичный куб. Предположим, что каким-то образом удается получить  $N$  случайных попарно независимых точек  $P_1, \dots, P_N$ , равномерно распределенных в  $G$ . Далее через  $M(s)$  будем обозначать математическое ожидание случайной величины  $s$ , а через  $D(s)$  — ее дисперсию. Случайные величины  $s_j = f(P_j)$  попарно независимы и одинаково распределены, причем

$$M(s_j) = \int_G f(P) dP = I(f)$$

и

$$D(s_j) = M(s_j^2) - (M(s_j))^2 = D(f),$$

где

$$D(f) = I(f^2) - (I(f))^2.$$

Положим

$$S_N(f) = \frac{1}{N} \sum_{j=1}^N s_j.$$

Вследствие указанных свойств величин  $s_j$  имеем

$$M(S_N(f)) = \frac{1}{N} \sum_{j=1}^N M(s_j) = I(f),$$

$$D(S_N(f)) = \frac{1}{N^2} \sum_{j=1}^N D(s_j) = \frac{1}{N} D(f).$$

С вероятностью  $1 - \eta$  выполняется неравенство (неравенство Чебышева)

$$|S_N(f) - I(f)| \leq \sqrt{D(f)/(\eta N)}. \quad (1)$$

Полагая  $\eta = 0,01$ , получаем: с вероятностью 99 % выполняется неравенство

$$|S_N(f) - I(f)| \leq 10 \sqrt{D(f)/N}.$$



Еще лучшая оценка получается в предположении, что точки  $P_j$  не только попарно независимы, но и независимы в совокупности. Тогда, согласно центральной предельной теореме, случайная величина

$$(S_N(f) - I(f)) / \sqrt{D(f)/N}$$

распределена асимптотически нормально с функцией распределения

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y \exp\left\{-\frac{t^2}{2}\right\} dt.$$

Вероятность того, что случайная величина с такой функцией распределения не превосходит по модулю значение  $y \geq 0$ , асимптотически равна

$$p_0(y) = 1 - \frac{\sqrt{2}}{\sqrt{\pi}} \int_y^{\infty} \exp\left\{-\frac{t^2}{2}\right\} dt.$$

Таким образом, при  $N$  больших с вероятностью, близкой к  $p_0(y)$ , выполняется неравенство

$$|S_N(f) - I(f)| \leq y \sqrt{D(f)/N}.$$

Полагая  $y = 3$  и  $y = 5$ , получаем, что неравенства

$$|S_N(f) - I(f)| \leq 3 \sqrt{D(f)/N} \quad \text{и} \quad |S_N(f) - I(f)| \leq 5 \sqrt{D(f)/N}$$

выполняются соответственно с вероятностями 0,997 и 0,99999. Сформулированные выше утверждения называются иногда правилами «трех сигм» и «пяти сигм» соответственно.

В правой части этих оценок стоит неизвестная величина  $D(f) = I(f^2) - (I(f))^2$ , которую можно оценить на основании информации о вычисленных значениях  $f(P_j)$ .

**Задача 1.** Показать, что

$$D(f) = \frac{N}{N-1} M(s_j(f) - S_N(f))^2. \quad (2)$$

Поскольку на основании закона больших чисел с большой вероятностью выполняется приближенное равенство

$$M(s_j(f) - S_N(f))^2 \approx \frac{1}{N} \sum_{j=1}^N (s_j(f) - S_N(f))^2,$$

то с большой вероятностью выполняется и приближенное равенство

$$D(f) \approx D^{(N)}(f),$$

где

$$D^{(N)}(f) = \frac{1}{N-1} \sum_{j=1}^N (s_j(f) - S_N(f))^2.$$

Приведем схему применения метода Монте-Карло при  $\mu(G) = 1$ . Пусть  $(1 - \eta)$  — уровень надежности, с которым желательно получить приближенное значение интеграла,  $\varepsilon$  — заданная точность. Определяют  $y$  из равенства

$$\eta = \frac{\sqrt{\varepsilon}}{\sqrt{\pi}} \int_y^{\infty} \exp\left\{-\frac{t^2}{2}\right\} dt.$$

Последовательно при  $n = 1, \dots$  получают случайные точки  $P_n$  и вычисляют величины  $t_n(f)$ ,  $S_n(f)$ ,  $d_n(f)$ ,  $D_n(f)$ , пользуясь рекуррентными соотношениями

$$t_n(f) = t_{n-1}(f) + f(P_n), \quad S_n(f) = t_n(f)/n,$$

$$d_n(f) = d_{n-1}(f) + \frac{n}{n-1} (f(P_n) - S_n(f))^2, \quad D_n(f) = d_n(f)/(n-1),$$

а также вычисляют значение

$$\lambda_n = y \sqrt{D_n(f)/n}.$$

Начальные условия рекурсии:

$$t_1(f) = S_1(f) = f(P_1), \quad d_1(f) = D_1(f) = 0.$$

Если оказалось, что  $\lambda_n \leq \varepsilon$ , то вычисления прекращаются; полагают  $I(f) \approx S_N(f)$  и считают, что с вероятностью  $1 - \eta$  выполняется неравенство  $|I(f) - S_N(f)| \leq \varepsilon$ .

Заметим, что в действительности неравенство  $|I(f) - S_N(f)| \leq \varepsilon$  выполняется с вероятностью, несколько меньшей чем  $1 - \eta$ , хотя и очень близкой к ней.

**Задача 2.** Проверить, что  $D_n(f) = D^{(n)}(f)$ .

Для уменьшения погрешности метода Монте-Карло случайные узлы интегрирования берутся распределенными не равномерно, а с некоторой плотностью  $p(P) \neq 1$ ,  $\int_G p(P) dP = 1$ . В этом случае полагают

$$I(f) \approx S_N(f) = \frac{1}{N} \sum_{j=1}^N f(P_j)/p(P_j).$$

**Задача 3.** Показать, что

$$M(S_N(f)) = I(f), \\ D(f) = I(f^2/p) - (I(f))^2.$$

Убедиться в том, что переход к такому выбору случайных величин особенно целесообразен в случае  $f(P)/p(P) \approx \text{const}$ .

Часто в руководствах по использованию метода Монте-Карло говорится следующее. Метод Монте-Карло является универсальным методом вычисления интегралов высокой кратности. Порядок оценки скорости сходимости метода Монте-Карло есть  $O(1/\sqrt{N})$  и не зависит от размерности интеграла, в то время как порядок гарантированных оценок скорости сходимости существенно ухудшается с ростом размерности. Для метода Монте-Карло при каждом  $n$  имеется эффективная оценка погрешности через величину  $\sqrt{D_n(f)/n}$ , и, таким образом, вычисления прекращаются именно в тот момент, когда достигнута требуемая точность.

Отнесемся осторожно к этому заявлению и рассмотрим различные соображения «за» и «против» метода Монте-Карло.

## § 9. Обсуждение правомерности использования недетерминированных методов решения задач

Часть пользователей предубеждена против метода Монте-Карло и отрицает правомерность его использования, поскольку малость погрешности метода обеспечивается лишь с некоторой вероятностью.

Выше мы уже обрисовали картину, складывающуюся при вычислении интегралов очень высокой кратности, как почти полностью безнадежную при получении приближенного значения интеграла с гарантированно малой оценкой погрешности. Такая обстановка и вызвала к жизни применение метода Монте-Карло.

Заметим, что даже программы вычисления однократных интегралов не дают полной гарантии малости погрешности в случае произвольных гладких подынтегральных функций. Кроме того, при решении всякой задачи возможны ошибки в постановке задачи, в программе и т. д. В силу этих и ряда других причин редко можно дать стопроцентную гарантию малости погрешности результата расчета по отношению к реальной модели; некоторая вероятность ошибочности результатов вычислений имеется в любом случае. Все это подчеркивает, что полный отказ от метода Монте-Карло только из-за его вероятностной природы является неоправданным.

С другой стороны, при использовании метода Монте-Карло нужно учитывать следующие отрицательные эффекты.

Для применения метода Монте-Карло необходимо иметь в распоряжении последовательность независимых точек  $P_i$  с заданным законом распределения. Обычно пользователь располагает датчиками случайных или так называемых псевдослучайных чисел, которые выдают последовательности случайных чисел, равномерно распределенных на отрезке  $[0, 1]$ . При помощи преобразований таких случайных величин получаются случайные числа с требуемыми законами распределения. На первых ЭВМ датчиками случайных чисел были некоторые приборы, например использующие явление радиоактивного распада, которые выдавали последовательности случайных величин, иногда даже удовлетворявшие требованию независимости в совокупности. Однако такие приборы работают с малой скоростью, и поэтому с увеличением скорости работы ЭВМ от них отказались. Вместо датчиков случайных чисел используются датчики псевдослучайных чисел — некоторые программы, выдающие последовательности чисел, которые рекомендуется рассматривать как случайные. Использование датчиков псевдослучайных чисел явилось прогрессивным шагом, позволившим широко применять вероятностные методы. Однако при использовании этих датчиков нужно всегда иметь в виду, какими свойствами обладают последовательности чисел, выдаваемые этими датчиками. Например, некоторые датчики псевдослучайных чисел вырабатывают последовательности чисел, которые можно рассматривать лишь как

попарно независимые, а не как независимые в совокупности. В этом случае будет неправомерно пользоваться оценками погрешности, основанными на центральной предельной теореме.

Пусть методом Монте-Карло вычисляется интеграл

$$I(f) = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s) dx_1 \dots dx_s.$$

Предположим, что в качестве узлов мы хотим выбрать последовательность независимых равномерно распределенных точек единичного куба. Если датчик псевдослучайных чисел выдает последовательность  $\xi_1, \dots$  чисел, равномерно распределенных на отрезке  $[0, 1]$ , то можно попытаться в качестве узлов интегрирования взять точки  $(\xi_1, \dots, \xi_s), (\xi_{s+1}, \dots, \xi_{2s}), \dots$

Для законности применения неравенства Чебышева нужно выполнение предположения о независимости распределения любых точек  $P_j$ , т. е. независимости распределения совокупностей

$$(\xi_{(j-1)s+1}, \dots, \xi_{js}), (\xi_{(i-1)s+1}, \dots, \xi_{is}).$$

При увеличении  $s$  это условие накладывает все более жесткие требования на датчики псевдослучайных чисел. Известно много реальных примеров неудачного применения метода Монте-Карло в случае больших  $s$ , вызванных следующей причиной. При использовании метода и оценке погрешности делались допущения о тех или иных статистических свойствах псевдослучайных чисел, в то время как эти предположения на самом деле не выполнялись. В результате делался вывод о малости значения погрешности, который на самом деле не был справедлив.

Таким образом, опасность применения метода Монте-Карло заключается по большей части не в вероятностном характере оценки погрешности, а в том, что вероятностная оценка погрешности производится зачастую в предположении о свойствах датчиков случайных чисел, которые на самом деле не имеют места.

**Задача 1.** Пусть  $\xi_1, \xi_2$  — случайные независимые величины, равномерно распределенные на отрезке  $[0, 1]$ . Пусть  $\{y\}$  — дробная доля числа  $y$ , т. е.  $\{y\} = y - [y]$ , где  $[y]$  — целая часть  $y$ . Положим  $\xi_n = \{\xi_1 + (n-1) \times (\xi_2 - \xi_1)\}$ . Показать что точки  $\xi_n$  попарно независимы, равномерно распределены на  $[0, 1]$  и, согласно построениям предыдущего параграфа,

$$I(f) = \int_0^1 f(x) dx \approx S_N(f) = \frac{1}{N} \sum_{j=1}^N f(\xi_j).$$

**Задача 2.** Пусть вычисляется двукратный интеграл

$$I(f) = \int_0^1 \int_0^1 f(x_1, x_2) dx_1 dx_2.$$

Показать, что при  $P_n = (\xi_{2n-1}, \xi_{2n})$ , где  $\xi_k$  из задачи 1, справедливо соотношение  $M(S_N(f)) = I(f)$ . Вычислить

$$\lim_{N \rightarrow \infty} D(S_N(f)) = d(f)$$

и убедиться, что, как правило,  $d(f) \neq 0$  и поэтому  $S_N(f)$  далеко от  $I(f)$ .

Задача 3. Пусть вычисляется трехкратный интеграл

$$I(f) = \int_0^1 \int_0^1 \int_0^1 f(x_1, x_2, x_3) dx_1 dx_2 dx_3.$$

При  $P_n = (\xi_{3n-2}, \xi_{3n-1}, \xi_{3n})$  вычислить  $M(S_n(f)) = i(f)$  и убедиться, что, как правило,  $i(f) \neq I(f)$  и поэтому  $S_N(f)$  далеко от  $I(f)$ .

Указание. Воспользоваться тем, что  $(\xi_{3n} - 2\xi_{3n-1} + \xi_{3n-2})$  — целое в пределах  $[-1, 1]$ .

Задача 4. Обобщить решение задачи 1 на случай вычисления интеграла по  $s$ -мерному кубу.

К числу достоинств метода Монте-Карло относят независимость порядка оценки от размерности вычисляемого интеграла. Однако, рассуждая только о порядке сходимости метода, можно не заметить следующую немаловажную деталь. Мы получали оценки погрешности приближенного значения интеграла вида

$$|S_N(f) - I(f)| \leq \text{const} \cdot \sqrt{D(f)/N}.$$

Типичным для практики является требование малости относительной погрешности приближенного значения интеграла, что в данном случае означает требование малости величины  $\sqrt{D(f)} / (|I(f)| \sqrt{N})$ . Статистика реально предъявляемых к вычислению интегралов показывает, что величина  $\sqrt{D(f)} / |I(f)|$  имеет тенденцию к резкому росту с ростом размерности интегралов. В качестве иллюстрации приведем интеграл

$$I(f_s) = \int_0^1 \dots \int_0^1 \exp\{-32(x_1^2 + \dots + x_s^2)\} dx_1 \dots dx_s,$$

для которого  $\sqrt{D(f_s)} / |I(f_s)| > 10^{s/2} - 1$ .

Следовательно, практическая трудоемкость метода Монте-Карло существенно возрастает с ростом размерности интегралов (при одинаковой относительной погрешности). При действительном вычислении многократных интегралов методом Монте-Карло перед непосредственным применением метода зачастую с целью уменьшения величины  $\sqrt{D(f)} / |I(f)|$  проводится довольно кропотливое исследование свойств подынтегральной функции, преобразование интегралов с помощью замен переменных и других приемов.

## § 10. Ускорение сходимости метода Монте-Карло

Рассмотрим некоторые приемы повышения практической эффективности метода Монте-Карло.

1. Функция  $f(P)$  представляется в виде

$$f(P) = F(P) + g(P),$$

где функция  $F(P)$  интегрируется явно и содержит в себе все резко меняющиеся компоненты  $f(P)$ , а  $g(P)$  — плавно меняющаяся функция с небольшой дисперсией  $D(g)$ . Иногда область интегрирования разбивается на малые подобласти и в каждой части в качестве  $F(P)$  берут некоторый интерполяционный полином с узлами в этой подобласти.

2. Подходящий подбор плотности распределения узлов  $p(P)$  (см. задачу 3 в § 8) также приводит к уменьшению дисперсии. Мы рассматривали случай, когда все узлы  $P_j$  имеют одинаковую функцию распределения  $p(P)$ . В ряде случаев оказывается целесообразным выбирать узлы интегрирования каждый со своей функцией распределения  $p_j(P)$ .

3. Следующий прием является частным случаем приемов 1 и 2. Исходный интеграл представляется в виде суммы интегралов

$$I(f) = \int_a^b f(P) dP = \sum_{i=1}^m I_i(f),$$

$$I_i(f) = \int_{a_i} b_i f(P) dP,$$

число узлов интегрирования  $N$  представляется в виде  $N = N_1 + \dots + N_m$ , и каждый интеграл  $I_i(f)$  вычисляется по методу Монте-Карло с  $N_i$  узлами интегрирования.

Обратимся к случаю вычисления интеграла от функции, изображенной на рис. 5.10.1. При непосредственном вычислении исходного интеграла в случае  $p(P) \equiv 1$  имеем  $D(f) = 1/4$ . Если  $G_1 = [0, 1/2]$ ,  $G_2 = [1/2, 1]$ , то при любых  $N_1, N_2 \neq 0$  оба интеграла  $I_i(f)$ , а следовательно, и исходный интеграл будут вычислены точно.

Конечно, случай, когда исходную область интегрирования удается разбить на части, где подынтегральная функция постоянна, очень редкий. Однако если все-таки удастся разбить ее на части, где функция меняется мало, то можно получить существенное увеличение точности при том же объеме вычислений.

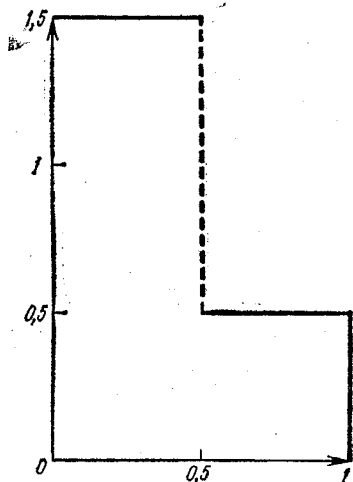


Рис. 5.10.1

Разбиение области интегрирования на части с целью уменьшения дисперсии метода Монте-Карло широко используется, в частности, при обработке естественно-научной информации. Пусть требуется определить водосодержание снега в бассейне некоторой реки. При непосредственном применении метода Монте-Карло выбиралось бы независимо с равномерной плотностью распределения несколько точек, где производилось бы измерение количества водосодержания на единицу поверхности. Участки поверхности с однородными природными условиями (высота над уровнем моря, уровень облачности, заселенность, ориентация склонов гор, осадки, господствующее направление ветра) характеризуются примерно одинаковым водосодержанием. Поэтому удается добиться существенного повышения точности, разбивая бассейн на части с однородными природными условиями и применяя метод Монте-Карло для вычисления интегралов по этим частям.

В случае гладкой подынтегральной функции разбиение области интегрирования на части приводит к увеличению порядка скорости сходимости. Пусть вычисляется интеграл

$$I(f) = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s) dx_1 \dots dx_s.$$

Положим  $N = n^s$  и разобьем исходную область интегрирования на равные кубы  $\Pi_m$ :  $(m_1 - 1)/n \leq x_1 \leq m_1/n, \dots, (m_s - 1)/n \leq x_s \leq m_s/n$ . В каждом кубе выберем случайную точку  $P_m$ . Считаем, что ее плотность распределения — постоянная, равная  $n^s$ , и случайные точки в любых двух кубах выбираются независимо. Положим

$$\bar{S}_N(f) = \frac{1}{n^s} \sum_{m_1, \dots, m_s=1}^n f(P_m).$$

**Задача 1.** Доказать, что  $D(\bar{S}_N(f)) \leq D(S_N(f)) = D(f)/N$ .

Проведем оценку дисперсии в предположении, что функция  $f(P)$  удовлетворяет условию Липшица с постоянной  $A$  по каждой из переменных. Справедливо равенство

$$D(\bar{S}_N(f)) = \sum_{m_1, \dots, m_s=1}^n D\left(\frac{1}{n^s} f(P_m)\right) = \sum_{m_1, \dots, m_s=1}^n \frac{1}{n^{2s}} D(f(P_m)). \quad (1)$$

Имеем равенство

$$M(f(P_m)) = \sigma_m = \int_{\Pi_m} n^s f(P) dP.$$

На основании теоремы о среднем  $\sigma_m = f(\bar{P}_m)$ , где  $\bar{P}_m \in \Pi_m$ , поэтому

$$M(f(P_m)) = f(\bar{P}_m).$$

В то же время

$$\begin{aligned} & f(x_1 + \Delta_1, \dots, x_s + \Delta_s) - f(x_1, \dots, x_s) = \\ & = (f(x_1 + \Delta_1, \dots, x_s + \Delta_s) - f(x_1 + \Delta_1, \dots, x_{s-1} + \Delta_{s-1}, x_s)) + \\ & \quad + (f(x_1 + \Delta_1, \dots, x_{s-1} + \Delta_{s-1}, x_s) - \\ & \quad - f(x_1 + \Delta_1, \dots, x_{s-2} + \Delta_{s-2}, x_{s-1}, x_s)) + \dots \\ & \quad \dots + (f(x_1 + \Delta_1, x_2, \dots, x_s) - f(x_1, \dots, x_s)). \end{aligned}$$

Отсюда следует, что для функций рассматриваемого класса

$$|f(x_1 + \Delta_1, \dots, x_s + \Delta_s) - f(x_1, \dots, x_s)| \leq A(|\Delta_1| + \dots + |\Delta_s|).$$

Если точка  $P_m$  принадлежит  $\Pi_m$ , то каждая из ее координат отличается от соответствующей координаты точки  $\bar{P}_m$  не более чем на  $n^{-1}$ . Поэтому из последнего неравенства следует оценка

$$|f(P_m) - f(\bar{P}_m)| \leq Asn^{-1}$$

при  $P_m \in \Pi_m$ , и, следовательно,

$$|f(P_m) - \sigma_m| \leq Asn^{-1}. \quad (2)$$

Всякая случайная величина  $\xi$  удовлетворяет неравенству

$$|M(\xi)| \leq \sup |\xi|.$$

Следовательно,

$$D(f(P_m)) = M(f(P_m) - \sigma)^2 \leq \sup_{P_m \in \Pi_m} (f(P_m) - \sigma_m)^2 \leq \left(\frac{As}{n}\right)^2.$$

С помощью этой оценки заключаем, что правая часть равенства (1) не превосходит величины

$$n^s n^{-2s} (As/n)^2 = (As)^2 / n^{s+2}.$$

Обозначив общее число узлов  $n^s$  через  $N$ , получаем оценку

$$D(\bar{S}_N(f)) \leq A^2 s^2 / N^{1+2/s}. \quad (3)$$

Отсюда на основании неравенства Чебышева (8.1) заключаем, что с вероятностью  $1 - \eta$  выполняется неравенство

$$|\bar{S}_N(f) - I(f)| \leq AsN^{-1/s} / \sqrt{\eta N}.$$

Полученная оценка погрешности по порядку лучше, чем оценка погрешности  $O(1/\sqrt{N})$  метода Монте-Карло.

Мы получили оценку погрешности по вероятности. Оказывается, что для рассматриваемого метода можно получить и гарантированную оценку погрешности. Умножая (2) на  $n^{-s}$ , получаем неравенство

$$\left| \frac{1}{n^s} f(P_m) - \int_{\Pi_m} f(P) dP \right| \leq As/n^{s+1}.$$



Величина  $S_N(f) - I(f)$  может быть представлена в виде суммы таких слагаемых:

$$\sum_{m_1, \dots, m_s=1}^n \left( \frac{1}{n^s} f(P_m) - \int_{\Pi_m} f(P) dP \right).$$

Суммируя оценки для этих слагаемых, получим

$$|\bar{S}_N(f) - I(f)| \leq As/n = As/N^{1/s}.$$

Сопоставляя эту оценку с теоремой из § 6, заключаем, что рассматриваемый метод имеет гарантированную оценку погрешности, оптимальную по порядку на рассматриваемом классе функций.

Возникает вопрос, можно ли улучшить на этом классе оценку дисперсии (3). Метод Монте-Карло и другие способы интегрирования, подобные рассматриваемому, где приближенное значение интеграла зависит от некоторых случайных параметров, называются *недетерминированными*. Пусть  $S_N(F)$  — приближенное значение интеграла  $I(f)$ , получаемое при применении некоторого недетерминированного способа вычисления.

*Теорема (без доказательства). Существуют  $d_1(r, A)$ ,  $d_2 > 0$ , удовлетворяющие следующему соотношению. Для любого способа вычисления интеграла, где информация о подынтегральной функции используется лишь как информация об ее значениях в  $N$  точках, найдется функция  $f \in C_r(A)$ , для которой*

$$|S_N(f) - I(f)| > d_1(r, A)/N^{r+1/2} \quad (4)$$

*с вероятностью  $d_2$ , где  $1/r = 1/r_1 + \dots + 1/r_s$ , а  $S_N(f)$  — приближенное значение интеграла.*

Следствием неравенства (4) является неравенство

$$\sqrt{D(S_N(f))} \geq d_3(r, A)/N^{r+1/2}, \quad (5)$$

которое, в частности, означает, что оценка (3) не может быть улучшена по порядку.

*Теорема (без доказательства). Можно указать способы интегрирования, для которых имеет место гарантированная оценка погрешности*

$$|S_N(f) - I(f)| \leq d_4(r, A)/N^r \quad (6)$$

*и одновременно*

$$M(S_N(f)) = I(f), \quad \sqrt{D(S_N(f))} \leq d_5(r, A)/N^{r+1/2} \quad (7)$$

*для всех  $f \in C_r(A)$ .*

Мы построили выше такой способ при  $r_1 = \dots = r_s = 1$ . Идея построения таких способов состоит в разбиении исходной области интегрирования на малые части и вычислении интеграла по малой части при помощи некоторой «случайной квадратурной формулы».

**Задача 2.** Пусть  $P_m$  — случайная точка в кубе  $\Pi_m$  (при тех же условиях на распределение и независимость точек  $P_m$ , как в задаче 1). Обозначим через  $P_m^*$  точку, симметричную  $P_m$  относительно центра куба  $\Pi_m$ , и положим

$$s_m(f) = \frac{1}{2n^s} (f(P_m) + f(P_m^*)), \quad S_N(f) = \sum_{m_1, \dots, m_s=1}^n s_{m_i}(f).$$

Доказать, что для функции из класса  $C_{(2, \dots, 2)}(A)$  выполнены одновременно оценки (6), (7). (Заметим, что здесь  $r = 2/s$ .)

## § 11. О выборе метода решения задачи

Обсудим схему решения одной задачи, где оказалось выгодным использование методов Монте-Карло, обсуждавшихся в предыдущем параграфе. Требовалось вычислить серию интегралов

$$I(\alpha_1, \alpha_2, \alpha_3) = \int_G f(P; \alpha_1, \alpha_2, \alpha_3) dP$$

при различных значениях параметров  $\alpha_1, \alpha_2, \alpha_3$ . Область интегрирования  $G$  содержалась в единичном трехмерном кубе; при этом условие принадлежности точки к области  $G$  задавалось громоздкой системой неравенств. Было ясно, что нельзя сделать замену независимых переменных так, чтобы область интегрирования  $G$  приобрела стандартный вид, где было бы возможным применение квадратурных формул высокого порядка точности. Поэтому пришлось вне области  $G$  продолжить  $f$  нулем и рассматривать исходную задачу как задачу вычисления интеграла

$$\bar{I}(\alpha_1, \alpha_2, \alpha_3) = \int_0^1 \int_0^1 \int_0^1 \bar{f}(P; \alpha_1, \alpha_2, \alpha_3) dx_1 dx_2 dx_3,$$

где

$$\bar{f}(P; \alpha_1, \alpha_2, \alpha_3) = \begin{cases} f(P; \alpha_1, \alpha_2, \alpha_3) & \text{при } P \in G, \\ 0 & \text{при } P \notin G. \end{cases}$$

Подынтегральная функция оказалась теперь разрывной. Для вычисления интеграла были применены формулы прямоугольников и Гаусса. С целью контроля над точностью проводились расчеты с различным числом узлов интегрирования. Оказалось, что результаты расчетов медленно устанавливаются (сильно меняются при изменении числа узлов), что указывает на малую точность получаемых приближенных значений. При непосредственном применении метода Монте-Карло установление получаемых приближенных значений было еще хуже. Поэтому было принято решение применить для вычисления описанные выше способы уменьшения дисперсии путем разбиения области на части. Были опробованы оба описанных выше способа. Область интегрирования  $0 \leq x_1, x_2, x_3 \leq 1$  разбивалась на равные кубы с ребрами длиной  $1/n$ , и далее применялись оба метода, рассмотренные в предыдущем параграфе. Как и в случае других квадратур, исследовалось установление результатов вычислений. Оказалось, что в случае вычисления одного значения интеграла оба метода могут обеспечить требуемую точность вычислений при приемлемых затратах машинного времени.

Во многих случаях вычисление интегралов высокой кратности или большой серии интегралов удается осуществить лишь за счет устранения повторения одинаковых вычислений. Решение рассматриваемой задачи представлялось сначала бесперспективным из-за большой трудоемкости проверки условия принадлежности узла  $P$  области  $G$ ; нахождение же каждого значения

$f(P; \alpha_1, \alpha_2, \alpha_3)$  требовало существенно меньшего объема вычислений. Так как все интегралы вычислялись по одной и той же области, то было принято решение применять следующий алгоритм. Куб  $0 \leq x_1, x_2, x_3 \leq 1$  разбивался на кубики с длиной ребра  $1/n$ , в каждом кубике выбиралась случайная точка  $P_m$  и проверялось условие принадлежности точки  $P_m$  области  $G$ , координаты всех точек  $P_m \in G$  были записаны в памяти ЭВМ. Далее все интегралы серии заменяли суммами

$$\sum_{P_m \in G} \frac{1}{n^3} f(P_m; \alpha_1, \alpha_2, \alpha_3)$$

с одними и теми же узлами. Отказ от проверки условия  $P_m \in G$  при вычислении каждого интеграла привел к снижению требований на затраты машинного времени примерно в 100 раз.

Другим фактором, позволившим резко снизить затраты машинного времени, оказался следующий. Подынтегральная функция имела вид

$$g(x_1, \alpha_1) h(x_1, x_2; \alpha_1, \alpha_2, \alpha_3),$$

где вычисление каждого значения функции  $h$  требовало относительно малого числа элементарных операций ЭВМ по сравнению с вычислением каждого значения функции  $g$ . Поэтому была применена следующая схема вычислений. Все интегралы  $T(\alpha_1, \alpha_2, \alpha_3)$ , соответствующие одному и тому же значению параметра  $\alpha_1$ , вычислялись одновременно; благодаря этому каждое значение  $g(x_1, \alpha_1)$  вычислялось при расчете всей серии интегралов только один раз. Такая организация работы позволила довести затраты машинного времени при решении рассматриваемой задачи до приемлемых размеров. Анализ хода решения задачи, проведенный после окончания вычисления серии, показал ряд неиспользованных возможностей, которые предоставили бы дополнительные удобства как исполнителю, так и заказчику. С целью уменьшения затрат машинного времени заказчик старался уменьшить число  $M$  совокупностей значений параметров  $(\alpha_1^i, \alpha_2^i, \alpha_3^i)$ , при которых требовалось вычислить значение интеграла. Наиболее трудоемкую часть при расчетах составляло вычисление значений функции  $g(x_1, \alpha_1)$ , поэтому общие затраты машинного времени были пропорциональны не числу  $M$ , а числу различных значений  $\alpha_1^i$ . Таким образом, общие затраты времени могли бы быть снижены также за счет удачного выбора совокупности  $(\alpha_1^i, \alpha_2^i, \alpha_3^i)$ .

В рассматриваемой задаче имелся еще один неиспользованный резерв повышения точности. Трудоемкость рассматриваемого метода пропорциональна произведению числа различных значений  $\alpha$  на число различных значений координат  $x_1$  узлов интегрирования. Узлы интегрирования выбирались случайно, и поэтому число различных значений  $x_1$  было очень большим. Для уменьшения этого числа можно было бы пойти, например, по следующему пути: аппроксимировать исходный интеграл по формуле прямоугольников

$$I(\alpha_1, \alpha_2, \alpha_3) \approx \sum_{q=1}^m \frac{1}{m} \bar{T}\left(\alpha_1, \alpha_2, \alpha_3; \frac{q-1/2}{m}\right),$$

где

$$\bar{T}\left(\alpha_1, \alpha_2, \alpha_3; \frac{q-1/2}{m}\right) = \int_0^1 \int_0^1 f\left(\alpha_1, \alpha_2, \alpha_3, \frac{q-1/2}{m}, x_2, x_3\right) dx_2 dx_3,$$

и применять метод Монте-Карло с разбиением области интегрирования только для вычисления интегралов  $I(\alpha_1, \alpha_2, \alpha_3; x_1)$ . Другой подход к решению задачи: разбить отрезок  $[0, 1]$  на отрезки  $[(q-1)/m, q/m]$ , на каждом из них

выбрать случайную точку  $\xi_q$  и положить

$$I(\alpha_1, \alpha_2, \alpha_3) \approx \frac{1}{m} \sum_{q=1}^m \bar{I}(\alpha_1, \alpha_2, \alpha_3; \xi_q).$$

Оба указанных выше метода имеют худшую скорость сходимости по сравнению с примененным методом, однако в этих методах вычисляется существенно меньше значений функций  $g(x_1, \alpha_1)$ .

Из последних рассуждений видно, что при вычислении больших серий интегралов (как, впрочем, и в случае больших серий других задач) часто больший эффект достигается не за счет повышения качества метода при решении каждой из задач серии, а за счет лучшей организации вычислений.

При применении процедуры, описанной в § 4, вычисление интеграла высокой кратности как бы сводилось к вычислению большого числа интегралов меньшей кратности. Поэтому при вычислении интегралов высокой кратности иногда можно использовать указанные выше резервы повышения эффективности в случае вычисления серий интегралов.

Обратим внимание на следующую опасность, возникающую при одновременном получении результата сразу по всей серии задач.

Предъявляемая к решению совокупность задач соответствует какому-то реальному явлению. Может случиться, что ранее явление не «обсчитывалось» на ЭВМ и предъявленная математическая модель не является удовлетворительной. Тогда в случае одновременного решения задач серии все результаты вычислений окажутся бесполезными. При последовательном решении задач можно уже после получения первых результатов обнаружить рассогласование математической модели с общей картиной явления. Такое рассогласование типично при решении новых задач, и обычно оно устраняется не сразу, а после большого числа пробных просчетов и совместного обсуждения модели заказчиком и исполнителем.

Исполнитель имеет особые основания быть заинтересованным в обсуждении постановки задачи. Если исходная постановка задачи окажется неразумной, ему придется потратить много времени на выбор нового алгоритма и написание новой программы. Участвуя в обсуждении, исполнитель может уяснить себе возможные варианты изменения постановок задачи и предусмотреть их при составлении программы решения задачи. В связи с этим при решении задач новых типов особенно важно, чтобы программа подразделялась на отдельные функциональные блоки, поддающиеся независимому изменению.

Часто столкновение исходных позиций заказчика (обсчет подробной модели) и исполнителя (минимальный объем работ) приводит к построению упрощенной модели явления, быстрый обсчет которой позволяет решить вопрос о целесообразности рассмотрения более сложной модели.

Рассмотрим другой пример, относящийся к организации работы по выбору метода интегрирования в случае кратных интегралов. Обычно вычисление интегралов по сложной области сводят к вычислению интегралов по области, являющейся прямым произведением областей простейшего вида: отрезков, параллелепипедов, лучей, бесконечных прямых, кругов, сфер, шаров и т. д. Для таких стандартных областей имеется достаточное количество совершенных методов интегрирования, и после такого преобразования области интегрирования можно применить процедуру, описанную в § 4, или какую-либо другую, аналогичную ей.

Пусть вычисляется интеграл

$$I(f) = \int_{1 \leq x_1^2 + x_2^2 + x_3^2 \leq 4} f(x_1, x_2, x_3) dx_1 dx_2 dx_3.$$

Его удобно записать в виде

$$I(f) = \int_1^2 \int_{S_1} g(l, \omega) d\omega dl,$$

где  $S_1$  — единичная сфера,  $d\omega$  — элемент ее площади,

$$g(l, \omega) = l^2 f(l\omega_1, l\omega_2, l\omega_3),$$

$$\omega = (\omega_1, \omega_2, \omega_3), \quad \omega_1^2 + \omega_2^2 + \omega_3^2 = 1.$$

Предположим, что решено вычислять интеграл при помощи квадратур, являющихся прямым произведением квадратур по отрезку  $[1, 2]$  и по сфере  $S_1$ .

Под *прямым произведением* квадратуры

$$\int_1^2 h(l) dl \approx \sum_{j=1}^n d_j h(l_j) \quad (1)$$

и квадратуры

$$\int_{S_1} p(\omega) d\omega \approx \sum_{q=1}^m k_q p(\omega_q) \quad (2)$$

здесь понимается квадратура

$$\int_1^2 \int_{S_1} g(l, \omega) d\omega dl \approx \sum_{j=1}^n \sum_{q=1}^m d_j k_q g(l_j, \omega_q).$$

Исследуем зависимость погрешности интегрирования от способа интегрирования и от числа узлов. Для исследования поведения погрешности численного интегрирования по оси  $l$  выберем какую-то «базовую» квадратуру по единичной сфере:

$$\int_{S_1} p(\omega) d\omega \approx \sum_{q=1}^{m_0} k_q^0 p(\omega_q^0); \quad (3)$$

имеется в виду, что число узлов  $m_0$  мало и в то же время выражения

$$G(l) = \int_{S_1} g(l, \omega) d\omega \quad \text{и} \quad G^0(l) = \sum_{q=1}^{m_0} k_q^0 g(l, \omega_q^0)$$

имеют одинаковый качественный характер поведения по  $l$ . Например, можно попытаться взять в качестве (3) квадратуру

$$\int_{S_1} p(\omega) d\omega \approx \frac{2\pi}{3} \sum p(\omega_1, \omega_2, \omega_3),$$

где суммирование производится по 6 точкам пересечения единичной сферы с координатными осями  $x_1, x_2, x_3$ .

Предположим, что для вычисления интеграла по оси  $l$  принято решение применить или формулу Гаусса, или формулу Симпсона. Последовательно применяя формулы Гаусса

$$\int_1^2 h(l) dl \approx \sum_{j=1}^n d_j^n h(l_j^n).$$

при  $n = n_1, n_2, \dots$  узлах, получаем некоторые величины

$$G_{n_i}^0 = \sum_{j=1}^{n_i} d_j^{n_i} G^0(t_j^{n_i}).$$

Точно так же при числе узлов  $n = n'_1, n'_2, \dots$  получаем приближения по формуле Симпсона  $S_{n_i}^0$ .

Из рассмотрения поведения всех этих величин можно усмотреть значение предела  $I^0$ , к которому они стремятся. Далее, при каждом  $n$  среди всех приближений  $G_{n_i}^0$  и  $S_{n_i}^0$ , с  $n_i, n'_i \leq n$  выберем приближение  $G_n$ , обеспечивающее лучшую точность, и введем функцию  $\varphi_l(n) \equiv |G_n - I^0|$  погрешности численного интегрирования по оси  $l$ .

Точно так же фиксируется некоторая базовая квадратура по переменной  $l$  (часто это квадратура Гаусса с двумя узлами) и строится функция  $\varphi_\omega(m)$  погрешности численного интегрирования по сфере  $\omega$ . Предположим, что суммарная погрешность есть  $R = \varphi_l(n) + \varphi_\omega(m)$ . Если значения функции  $f$  вычисляются независимо, то трудоемкость метода пропорциональна  $nm$ . Минимизируя  $nm$  при заданном требовании на точность, чтобы  $\varphi_l(n) + \varphi_\omega(m) \leq \varepsilon$ , получаем искомые значения числа узлов  $n_0$  и  $m_0$ . В зависимости от того, какой квадратуре — Гаусса или Симпсона — отвечает данное  $n$ , выбираем соответствующую квадратуру. В случае сомнений в правомерности использования этой методики остается возможность проверить правильность результата, проведя дополнительное интегрирование с несколько отличными от  $n_0$  и  $m_0$  значениями  $n$  и  $m$ .

Приведем типичный пример подобной организации выбора способа интегрирования по каждой из осей в случае большой серии интегралов по единичному  $s$ -мерному кубу. По каждой из осей применяется формула Гаусса, при выборе числа ее узлов по каждой из осей в качестве базовых квадратур по остальным осям берутся квадратуры Гаусса с двумя узлами. Дробление числа узлов по каждой рассматриваемой оси продолжается до тех пор, пока разность между двумя последующими приближениями не станет меньше чем  $\varepsilon \cdot s^{-1}$ . После определения таким образом нужного числа узлов  $n^{(i)}$ , соответствующего каждой оси, производится вычисление интеграла с  $n^{(i)}$  узлами по каждой из осей.

Из-за некоторой ненадежности описанного алгоритма обычно производится проверка правомерности его применения в случае данной серии интегралов: выбирается достаточно представительная подсерия интегралов и для нее результаты расчетов по данному алгоритму сравниваются с результатами расчетов при несколько измененных значениях числа узлов по осям.

Л и т е р а т у р а к г л. V: [2, 3, 41, 47—49].

## Глава VI

### ЧИСЛЕННЫЕ МЕТОДЫ АЛГЕБРЫ

К численным методам алгебры традиционно относят численные методы решения систем линейных уравнений, обращения матриц, вычисления определителей, нахождения собственных значений и собственных векторов матриц и нулей многочлена.

При формальном подходе решение этих задач не встречает затруднений: решение системы можно найти, раскрыв определители в формуле Крамера; для нахождения собственных значений матрицы достаточно выписать характеристическое уравнение и найти его корни. Однако эти рекомендации встречают возражения со многих сторон.

Так, при непосредственном раскрытии определителей решение системы с  $m$  неизвестными требует порядка  $m^3$  арифметических операций; уже при  $m=30$  такое число операций недоступно для современных ЭВМ. При сколько-нибудь больших  $m$  применение методов с таким порядком числа операций будет невозможно и в обозримом будущем.

Другой причиной, по которой эти классические способы неприменимы даже при малых  $m$ , является сильное влияние на окончательный результат округлений при вычислениях. Уже при  $m=20$  при расчетах на современных ЭВМ типична аварийная остановка из-за переполнения порядка чисел. Даже если такая остановка не происходит, результат вычислений часто далек от истинного значения из-за влияния вычислительной погрешности. Точно так же обстоит дело при нахождении собственных значений матриц с использованием явного выражения характеристического многочлена.

Методы решения алгебраических задач разделяются на точные, итерационные и вероятностные. Классы задач, для решения которых обычно применяются методы этих групп, можно условно назвать соответственно классами задач с малым, средним и большим числом неизвестных. Изменение объема и структуры памяти ЭВМ, увеличение их быстродействия и развитие численных методов приводят к смещению границ применения методов в сторону систем более высоких порядков. В настоящее время точные методы обычно применяются для решения систем до порядка  $10^4$ , итерационные — до порядка  $10^7$ ,

При изучении итерационных процессов нам понадобятся понятия норм вектора и матрицы. Напомним определения основных норм в пространствах векторов и матриц. Если в пространстве векторов  $\mathbf{x} = (x_1, \dots, x_m)^T$  введена норма  $\|\mathbf{x}\|$ , то согласованной с ней нормой в пространстве матриц  $A$  называют норму

$$\|A\| = \sup_{\mathbf{x} \neq 0} \|A\mathbf{x}\| / \|\mathbf{x}\|. \quad (1)$$

Наиболее употребительны в пространстве векторов следующие нормы:

$$\|\mathbf{x}\|_\infty = \max_{1 \leq j \leq m} |x_j|, \quad (2)$$

$$\|\mathbf{x}\|_1 = \sum_{j=1}^m |x_j|, \quad (3)$$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{j=1}^m |x_j|^2} = \sqrt{(\mathbf{x}, \mathbf{x})}, \quad (4)$$

а согласованными с ними нормами в пространстве матриц являются соответственно нормы

$$\|A\|_\infty = \max_{1 \leq i \leq m} \left( \sum_{j=1}^m |a_{ij}| \right), \quad (5)$$

$$\|A\|_1 = \max_{1 \leq j \leq m} \left( \sum_{i=1}^m |a_{ij}| \right), \quad (6)$$

$$\|A\|_2 = \sqrt{\max_{1 \leq i \leq m} \lambda_{A^*A}^i}; \quad (7)$$

здесь и далее  $\lambda_D^i$  — собственные значения матрицы  $D$ .

Приведем вывод этих соотношений для вещественного случая. Поскольку, согласно (2),

$$\begin{aligned} \|A\mathbf{x}\|_\infty = \max_i \left| \sum_j a_{ij}x_j \right| &\leq \max_i \left( \sum_j |a_{ij}| \max_j |x_j| \right) \leq \\ &\leq \max_i \left( \sum_j |a_{ij}| \right) \max_j |x_j|, \end{aligned}$$

то

$$\frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \leq \max_i \left( \sum_j |a_{ij}| \right).$$

Пусть  $\max_i \left( \sum_j |a_{ij}| \right)$  достигается при  $i = l$ ; для вектора

$$\mathbf{x} = (\text{sign}(a_{l1}), \dots, \text{sign}(a_{lm}))^T$$

имеем  $\|\mathbf{x}\|_\infty = 1$ ,

$$\|A\mathbf{x}\|_\infty \geq \left| \sum_j a_{lj}x_j \right| = \sum_j |a_{lj}| = \left( \max_i \sum_j |a_{ij}| \right) \|\mathbf{x}\|_\infty.$$

Из этих соотношений следует (5).



Точно так же для нормы вектора, определяемой по формуле (3), имеем

$$\|Ax\|_1 = \sum_i \left| \sum_j a_{ij} x_j \right| \leq \sum_i \sum_j |a_{ij}| |x_j| = \sum_j \left( \sum_i |a_{ij}| \right) |x_j| \leq \left( \max_i \sum_i |a_{ij}| \right) \sum_j |x_j|,$$

т. е.

$$\frac{\|Ax\|_1}{\|x\|_1} \leq \max_i \left( \sum_i |a_{ij}| \right).$$

Пусть  $\max_i \sum_i |a_{ij}|$  достигается при  $j = l$ . Для вектора  $x$ , у которого лишь одна компонента  $x_l$  отлична от нуля, имеем

$$\begin{aligned} \|Ax\|_1 &= \sum_i \left| \sum_j a_{ij} x_j \right| = \sum_i |a_{il}| |x_l| = \left( \sum_i |a_{il}| \right) |x_l| = \\ &= \left( \max_i \sum_i |a_{ij}| \right) \sum_j |x_j| = \left( \max_i \sum_i |a_{ij}| \right) \|x\|_1; \end{aligned}$$

отсюда следует (6).

Согласно определению  $\|A\|_2$  и (4), имеем

$$\|A\|_2 = \sup_x \frac{\|Ax\|_2}{\|x\|_2} = \sup_x \sqrt{\frac{(Ax, Ax)}{(x, x)}} = \sqrt{\sup_x \frac{(A^T A x, x)}{(x, x)}}.$$

Матрица  $A^T A$  симметричная, поскольку  $(A^T A)^T = A^T (A^T)^T = A^T A$ .

Пусть матрица  $B$  симметричная,  $e_1, \dots, e_m$  — ортонормированная система ее собственных векторов,  $\lambda_1, \dots, \lambda_m$  — соответствующие собственные значения. Всякий вектор  $x$  представим

в виде  $\sum_{i=1}^m c_i e_i$ . Имеем

$$(Bx, x) = \left( \sum_i \lambda_i c_i e_i, \sum_i c_i e_i \right) = \sum_i \lambda_i |c_i|^2,$$

поэтому

$$(Bx, x) \leq \left( \max_i \lambda_i \right) \sum_i |c_i|^2 = \left( \max_i \lambda_i \right) (x, x) \quad (8)$$

и

$$(Bx, x) \geq \left( \min_i \lambda_i \right) (x, x). \quad (9)$$

В то же время  $(Be_i, e_i) / (e_i, e_i) = \lambda_i$ . Из этих соотношений следует, что

$$\sup_x \frac{|(Bx, x)|}{(x, x)} = \max_i |\lambda_i|. \quad (10)$$

Поскольку  $(A^T A x, x) = (Ax, Ax) \geq 0$ , то все  $\lambda_{A^T A}^i \geq 0$ . Полагая в (10)  $B = A^T A$ , получаем

$$\sup_x \frac{|(A^T A x, x)|}{(x, x)} = \max_i |\lambda_{A^T A}^i| = \max_i \lambda_{A^T A}^i.$$

Из полученных соотношений следует (7).

Отметим важный частный случай.

Если  $A$  — симметричная матрица, то  $\lambda_{A^T A}^i = \lambda_{A^2}^i = |\lambda_A^i|^2$ , поэтому для нее

$$\|A\|_2 = \max |\lambda_A^i|. \quad (11)$$

Если  $Ax = \lambda x$ , то  $\|A\| \|x\| \geq \| \lambda x \| = |\lambda| \|x\|$ . Следовательно, модуль любого собственного значения матрицы  $A$  не больше любой ее нормы.

## § 1. Методы последовательного исключения неизвестных

Рассмотрим точные методы решения системы  $Ax = b$ ; здесь  $A = [a_{ij}]$  — матрицы размерности  $m \times m$ ,  $\det A \neq 0$ ,

$$b = (a_{1, m+1}, \dots, a_{m, m+1})^T.$$

Метод решения задачи относят к классу точных, если в предположении отсутствия округлений он дает точное решение задачи после конечного числа арифметических и логических операций. Если число ненулевых элементов матрицы системы имеет порядок  $m^2$ , то для большинства используемых в настоящее время точных методов решения таких систем требуемое число операций имеет порядок  $m^3$ . Поэтому для применимости точных методов необходимо, чтобы такой порядок числа операций был приемлем для данной ЭВМ; другие ограничения накладываются объемом и структурой памяти ЭВМ.

Оговорка об «используемых в настоящее время методах» имеет следующий смысл. Существуют методы решения таких систем с меньшим порядком числа операций, однако они не используются активно из-за сильной чувствительности результата к вычислительной погрешности.

Наиболее известным из точных методов решения систем линейных уравнений является *метод исключения Гаусса*. Рассмотрим одну из его возможных реализаций. В предположении, что  $a_{11} \neq 0$ , первое уравнение системы

$$\sum_{j=1}^m a_{1j} x_j = a_{1, m+1}, \quad i = 1, \dots, m, \quad (1)$$

делим на коэффициент  $a_{11}$ , в результате получаем уравнение

$$x_1 + \sum_{j=2}^m a_{1j}^1 x_j = a_{1, m+1}^1.$$

Затем из каждого из остальных уравнений вычитается первое уравнение, умноженное на соответствующий коэффициент  $a_{i1}$ . В результате эти уравнения преобразуются к виду

$$\sum_{j=2}^m a_{ij}^1 x_j = a_{i, m+1}^1, \quad i = 2, \dots, m.$$



Таким образом, система (2), получаемая в результате этих преобразований, запишется в виде

$$CAx = Cb, \quad \text{где } C = C_m \dots C_1 C_1.$$

Произведение левых (правых) треугольных матриц является левой (правой) треугольной матрицей, поэтому матрица  $C$  левая треугольная.

Из формулы для элементов обратной матрицы

$$(A^{-1})_{ij} = A_{ji} / \det A$$

следует, что матрица, обратная к левой (правой) треугольной, является левой (правой) треугольной. Следовательно, матрица  $B = C^{-1}$  левая треугольная.

Введем обозначение  $CA = D$ . Согласно построению все  $d_{ii} = 1$  и матрица  $D$  правая треугольная. Отсюда получаем представление матрицы  $A$  в виде произведения левой и правой треугольных матриц:

$$A = C^{-1}D = BD.$$

Равенство  $A = BD$  вместе с условием  $d_{ii} = 1, i = 1, \dots, m$ , образует систему уравнений относительно элементов треугольных матриц  $B$  и  $D$ :  $\sum_{j=1}^m b_{ij}d_{jk} = a_{ik}$ . Поскольку  $b_{ij} = 0$  при  $i < j$  и  $d_{jk} = 0$  при  $k < j$ , эта система может быть записана в виде

$$\sum_{j=1}^{\min(i, k)} b_{ij}d_{jk} = a_{ik} \quad (3)$$

или, что то же самое,

$$\begin{aligned} \sum_{j=1}^k b_{ij}d_{jk} &= a_{ik} \quad \text{при } k \leq i, \\ \sum_{j=1}^i b_{ij}d_{jk} &= a_{ik} \quad \text{при } i < k. \end{aligned}$$

Воспользовавшись условием, что все  $d_{ii} = 1$ , получаем систему рекуррентных соотношений для определения элементов  $b_{ij}$  и  $d_{ij}$ :

$$\begin{aligned} b_{ik} &= a_{ik} - \sum_{j=1}^{k-1} b_{ij}d_{jk} \quad \text{при } k \leq i, \\ d_{ik} &= \frac{a_{ik} - \sum_{j=1}^{i-1} b_{ij}d_{jk}}{b_{ii}} \quad \text{при } i < k. \end{aligned} \quad (4)$$

Вычисления проводятся последовательно для совокупностей  $(i, k) = (1, 1), \dots, (1, m), (2, 1), \dots, (2, m), \dots, (m, 1), \dots, (m, m)$ . Здесь и далее в случае, когда верхний предел сум-

мирования меньше нижнего, считается, что вся сумма равна нулю.

Таким образом, вместо последовательных преобразований системы (1) к виду (2) можно непосредственно произвести вычисление матриц  $B$  и  $D$  с помощью формул (4). Эти вычисления можно осуществить, если только все элементы  $b_{ii}$  окажутся отличными от нуля. Пусть  $A_k, B_k, D_k$  — матрицы главных миноров  $k$ -го порядка матриц  $A, B, D$ . Согласно (3)  $A_k = B_k D_k$ . Поскольку  $\det D_k = 1, \det B_k = b_{11} \dots b_{kk}$ , то  $\det A_k = b_{11} \dots b_{kk}$ . Следовательно,

$$b_{kk} = \det A_k / \det A_{k-1}.$$

Итак, для осуществимости вычислений по формулам (4) необходимо и достаточно выполнение условий

$$\det A_k \neq 0, \quad k = 1, \dots, m. \quad (5)$$

В ряде случаев заранее известно, что условие (5) выполнено. Например, многие задачи математической физики сводятся к решению систем с положительно определенной матрицей  $A$ . Однако в общем случае этого заранее сказать нельзя. Возможен и такой случай: все  $\det A_k \neq 0$ , но среди величин  $b_{kk}$  есть очень малые и при делении на них будут получаться большие числа с большими абсолютными погрешностями. В результате этого решение сильно исказится.

Введем обозначение  $Cb = d = (d_{1, m+1}, \dots, d_{m, m+1})^T$ . Поскольку  $C^{-1} = B$  и  $CA = D$ , то справедливы равенства  $Bd = b, Dx = d$ . Таким образом, после разложения матрицы исходной системы на произведение левой и правой треугольных матриц решение исходной системы сводится к последовательному решению двух систем  $Bd = b, Dx = d$  с треугольными матрицами; это потребует  $N \sim 2m^2$  арифметических операций.

Последовательность операций по разложению матрицы  $A$  на произведение треугольных матриц и по определению вектора  $d$  часто удобно объединить. Уравнения

$$\sum_{j=1}^i b_{ij} d_{j, m+1} = a_{i, m+1}$$

системы  $Bd = b$  можно записать в виде (3)

$$\sum_{j=1}^{\min(i, m+1)} b_{ij} d_{j, m+1} = a_{i, m+1}.$$

Следовательно, значения  $d_{i, m+1}$  могут вычисляться одновременно с остальными значениями  $d_{ij}$  по формулам (4).

При решении практических задач часто возникает необходимость решения систем уравнений с матрицей, содержащей большое количество нулевых элементов. Обычно эти матрицы имеют так называемую *ленточную структуру*. Более точно, матрицу  $A$  называют  $(2q + 1)$ -диагональной или имеющей *ленточную струк-*

туру, если  $a_{ij} = 0$  при  $|i - j| > q$ . Число  $2q + 1$  называют *шириной ленты*. Оказывается, что при решении системы уравнений с ленточной матрицей методом Гаусса число арифметических операций и требуемый объем памяти ЭВМ могут быть существенно сокращены.

**Задача 1.** Исследовать характеристики метода Гаусса и метода решения системы с помощью разложения ленточной матрицы  $A$  на произведение левой и правой треугольных матриц. Показать, что для нахождения решения требуется  $O(mq^2)$  арифметических операций (при  $m, q \rightarrow \infty$ ). Найти главный член числа операций при условии  $1 \ll q \ll m$ .

**Задача 2.** Оценить объем загружаемой памяти ЭВМ в методе Гаусса для ленточных матриц.

При вычислениях без помощи ЭВМ велика вероятность случайных погрешностей. Для устранения таких погрешностей иногда вводят *контрольный столбец системы*  $a_{m+2} = (a_{1, m+2}, \dots, a_{m, m+2})^T$ , состоящий из контрольных элементов уравнений системы

$$a_{i, m+2} = \sum_{j=1}^{m+1} a_{ij}.$$

При преобразовании уравнений над контрольными элементами производятся те же операции, что и над свободными членами уравнений. В результате этого контрольный элемент каждого нового уравнения должен равняться сумме коэффициентов этого уравнения. Большое расхождение между ними указывает на погрешности в вычислениях или на неустойчивость алгоритма вычислений по отношению к вычислительной погрешности.

К примеру, в случае приведения системы уравнений  $Ax = b$  к виду  $Dx = d$  с помощью формул (4) контрольный элемент  $d_{i, m+2}$  каждого из уравнений системы  $Dx = d$  вычисляется по тем же формулам (4). После вычисления всех элементов  $d_{ij}$  при фиксированном  $i$  контроль осуществляется проверкой равенства

$$\sum_{j=i}^{m+1} d_{ij} = d_{i, m+2}.$$

Обратный ход метода Гаусса также сопровождается вычислением контрольных элементов строк системы.

Чтобы избежать катастрофического влияния вычислительной погрешности, применяют *метод Гаусса с выбором главного элемента*. Его отличие от описанной выше схемы метода Гаусса состоит в следующем. Пусть по ходу исключения неизвестных получена система уравнений

$$\begin{aligned} x_i + \sum_{j=i+1}^m a_{ij}^i x_j &= a_{i, m+1}^i, & i = 1, \dots, k, \\ \sum_{j=k+1}^m a_{ij}^k x_j &= a_{i, m+1}^k, & i = k+1, \dots, m. \end{aligned}$$

Найдем  $l$  такое, что  $|a_{k+1, l}^k| = \max_j |a_{k+1, j}^k|$  и переобозначим  $x_{k+1} = x_l$  и  $x_l = x_{k+1}$ ; далее произведем исключение неизвестной  $x_{k+1}$  из всех уравнений, начиная с  $(k+2)$ -го. Такое переобозначение приводит к изменению порядка исключения неизвестных и

Во многих случаях существенно уменьшает чувствительность решения к погрешностям округления при вычислениях.

Часто требуется решить несколько систем уравнений  $Ax = b_q$ ,  $q = 1, \dots, p$ , с одной и той же матрицей  $A$ . Удобно поступить следующим образом: введя обозначения

$$b_q = (a_{1, m+q}, \dots, a_{m, m+q})^T,$$

произведем вычисления по формулам (4), причем элементы  $d_{ik}$  вычислим при  $i < k \leq m + p$ . В результате будут получены  $p$  систем уравнений с треугольной матрицей, соответствующих исходной задаче

$$Dx = d_q, \quad d_q = (d_{1, m+q}, \dots, d_{m, m+q})^T, \quad q = 1, \dots, p.$$

Решаем эти системы каждую в отдельности. Оказывается, что общее число арифметических действий при решении  $p$  систем уравнений таким способом  $N \sim 2m^3/3 + 2pm^2$ .

Описанный выше прием иногда используется для того, чтобы без существенных дополнительных затрат получить суждение о погрешности решения, являющейся следствием погрешностей округления при вычислениях. Задаются вектором  $z$  с компонентами, имеющими по возможности тот же порядок и знак, что и компоненты искомого решения; часто из-за отсутствия достаточной информации берут  $z = (1, \dots, 1)^T$ . Вычисляется вектор  $c = Az$ , и наряду с исходной системой уравнений решается система  $Az = c$ .

Пусть  $x'$  и  $z'$  — реально получаемые решения этих систем. Суждение о погрешности  $x' - x$  искомого решения можно получить, основываясь на гипотезе: относительные погрешности при решении методом исключения систем с одной и той же матрицей и различными правыми частями, которыми являются соответственные величины  $\|x - x'\|/\|x'\|$  и  $\|z - z'\|/\|z'\|$ , отличаются не в очень большое число раз.

Другой прием для получения суждения о реальной величине погрешности, возникающей за счет округлений при вычислениях, состоит в изменении масштабов, меняющем картину накопления вычислительной погрешности. Наряду с исходной системой тем же методом решается система

$$(\alpha A)x' = \beta b, \quad \text{где } \alpha \text{ и } \beta \text{ — числа.}$$

При  $\alpha$  и  $\beta$ , не являющихся целыми степенями двойки, сравнение векторов  $x$  и  $\alpha\beta^{-1}x'$  дает представление о величине вычислительной погрешности. Например, можно взять  $\alpha = \sqrt{2}$ ,  $\beta = \sqrt{3}$ .

Изучение многих задач приводит к необходимости решения систем линейных уравнений с симметричной положительно определенной матрицей. Такие системы возникают, например, при решении дифференциальных уравнений методом конечных элементов или же конечно-разностными методами. В этих случаях матрица системы имеет также и ленточную структуру.

Для решения таких систем, а также систем уравнений более общего вида с эрмитовой не обязательно положительно определенной матрицей применяется метод квадратного корня. Матрица  $A$  представляется в виде

$$A = S^*DS, \quad (6)$$

где  $S$  — правая треугольная матрица,  $S^*$  — сопряженная с ней, т. е.

$$S = \begin{bmatrix} s_{11} & s_{12} & \dots \\ 0 & s_{22} & \dots \\ \dots & \dots & \dots \end{bmatrix},$$

причем все  $s_{ii} > 0$ ,  $D$  — диагональная матрица с элементами  $d_{ii}$ , равными  $+1$  или  $-1$ . Матричное равенство (6) образует систему уравнений

$$a_{ij} = \bar{s}_{ii}s_{ij}d_{ii} + \dots + \bar{s}_{ij}s_{ij}d_{ii} \quad \text{при } i \leq j.$$

Аналогичные уравнения при  $i > j$  отброшены, так как уравнения, соответствующие парам  $(i, j)$  и  $(j, i)$ , эквивалентны. Отсюда получаем рекуррентные формулы для определения элементов  $d_{ii}$  и  $s_{ij}$ :

$$d_{ii} = \text{sign} \left( a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk} \right),$$

$$s_{ii} = \sqrt{\left| a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk} \right|},$$

$$s_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} \bar{s}_{ki}s_{kj}d_{kk}}{\bar{s}_{ii}d_{ii}} \quad \text{при } i < j.$$

Матрица  $S$  является правой треугольной, и, таким образом, после получения представления (6) решение исходной системы также сводится к последовательному решению двух систем с треугольными матрицами. Заметим, что в случае  $A > 0$  все  $d_{ii} = 1$  и  $A = S^*S$ .

**Задача 3.** Оценить число арифметических операций и загрузку памяти ЭВМ (при условии  $a_{ij} = a_{ji}$  объем памяти, требуемый для запоминания матрицы  $A$ , уменьшается) при решении системы с вещественной положительно определенной матрицей  $A$  методом квадратного корня.

Многие пакеты прикладных программ для решения краевых задач математической физики методом конечных элементов организованы по следующей схеме. После формирования матрицы системы  $A$  путем перестановки строк и столбцов (одновременно переставляются  $i$ -я и  $j$ -я строки и  $i$ -й и  $j$ -й столбцы) система преобразуется к виду с наименьшей шириной ленты. Далее применяется метод квадратного корня. При этом с целью уменьшения объема вычислений при решении системы  $Ax = b$  с другими правыми частями матрица  $S$  запоминается.

**Замечание.** Часто этот метод уступает по эффективности итерационным методам.

**Задача 4.** Оценить число арифметических операций и объем требуемой памяти метода квадратного корня в случае матриц ленточной структуры.

Если есть подозрение, что реально полученное решение  $x^1$  сильно искажено вычислительной погрешностью, то можно



поступить следующим образом. Определим вектор  $b^1 = b - Ax^1$ . Погрешность  $r^1 = x - x^1$  удовлетворяет системе уравнений

$$Ar^1 = Ax - Ax^1 = b^1. \quad (7)$$

Решая эту систему в условиях реальных округлений, получаем приближение  $r^{(1)}$  к  $r^1$ . Полагаем  $x^2 = x^1 + r^{(1)}$ . Если точность нового приближения представляется неудовлетворительной, то повторяем эту операцию. При решении системы (7) над компонентами правой части производятся те же линейные операции, что и над компонентами правой части при решении системы (1). Поэтому при вычислениях на ЭВМ с плавающей запятой естественно ожидать, что относительные погрешности решений этих систем будут одного порядка. Поскольку погрешности округлений обычно малы, то  $\|b^1\| \ll \|b\|$ ; тогда  $\|r^1\| \ll \|x^1\|$ , и, как правило, решение (7) определится с существенно меньшей абсолютной погрешностью, чем решение системы (1). Таким образом, применение описанного приема приводит к повышению точности приближенного решения.

Особенно удобно применять этот прием, когда по ходу вычислений в памяти ЭВМ сохраняются матрицы  $B$  и  $D$ . Тогда для каждого уточнения требуется найти вектор  $b^k = b - Ax^k$  и решить две системы с треугольными матрицами. Это потребует всего  $N_1 \sim 4m^2$  арифметических операций, что составит малую долю от числа операций  $N_0 \sim 2m^3/3$ , требующихся для представления матрицы  $A$  в виде  $A = BD$ .

Идея описанного приема последовательного уточнения приближений к решению часто реализуется в такой форме. Пусть матрица  $B$  близка в каком-то смысле к матрице  $A$ , но решение системы  $Bx = c$  требует существенно меньшего объема вычислений по сравнению с решением системы  $Ax = b$ . Решение системы  $Bx = b$  принимаем в качестве первого приближения  $x^1$  к решению. Разность  $x - x^1$  удовлетворяет системе уравнений

$$A(x - x^1) = b - Ax^1.$$

Вместо решения этой системы находим решение системы

$$Br^1 = b - Ax^1$$

и полагаем  $x^2 = x^1 + r^1$ . Таким образом, каждое приближение находится из предыдущего по формуле

$$x^{n+1} = x^n + B^{-1}(b - Ax^n) = (E - B^{-1}A)x^n + B^{-1}b.$$

Если матрицы  $A$  и  $B$  достаточно близки, то матрица  $E - B^{-1}A$  имеет малую норму и такой итерационный процесс быстро сходится (см. также § 10).

Значительно более редкой, чем задача решения системы уравнений, является задача обращения матриц. Для обратной матрицы  $X = A^{-1}$  имеем равенство  $AX = BDX = E$ . Таким образом, для нахождения матрицы  $X$  достаточно последовательно решить две матричные системы  $BX = E$ ,  $DX = Y$ . Нетрудно подсчитать, что при нахождении на таком пути матрицы  $A^{-1}$  общий объем вычислений составит  $N_2 \sim 2m^3$  арифметических операций.

В случае необходимости уточнения приближения к обратной матрице могут производиться при помощи итерационного процесса  $X_k = X_{k-1}(2E - AX_{k-1})$ . Для исследования сходимости итерационного процесса рассмотрим матрицы  $G_k = E - AX_k$ . Имеем равенство

$$G_k = E - AX_k = E - AX_{k-1}(2E - AX_{k-1}) = (E - AX_{k-1})^2 = G_{k-1}^2.$$

Отсюда получаем цепочку равенств

$$G_k = G_{k-1}^2 = G_{k-2}^4 = \dots = G_0^{2^k}.$$

Поскольку

$$A^{-1} - X_k = A^{-1}(E - AX_k) = A^{-1}G_k = A^{-1}G_0^{2^k},$$

то имеем оценку

$$\|A^{-1} - X_k\| \leq \|A^{-1}\| \cdot \|G_0\|^{2^k}.$$

Таким образом, при достаточно хорошем начальном приближении, т. е. если  $\|E - AX_0\| < 1$ , этот итерационный процесс сходится со скоростью геометрической прогрессии.

При решении систем больших порядков величины, необходимые в процессе решения системы, не могут одновременно уместиться в оперативной памяти ЭВМ, поэтому встает проблема размещения этих величин в элементах памяти различных уровней и организации процесса вычислений, при котором передача этих величин из одного места памяти в другое по возможности меньше замедляла бы ход вычислений.

## § 2. Метод отражений

В настоящее время разработано так много точных методов численного решения систем линейных алгебраических уравнений, что даже простое перечисление их затруднительно. Большинство этих методов, как и метод исключения Гаусса, основано на переходе от заданной системы  $Ax = b$  к новой системе  $CAx = Cb$  такой, что система  $Bx = d$ , где  $B = CA$  и  $d = Cb$ , решается проще, чем исходная. При выборе подходящей матрицы  $C$  нужно учитывать по крайней мере следующие два фактора. Во-первых, ее вычисление не должно быть чересчур сложным и трудоемким. Во-вторых, умножение на матрицу  $C$  не должно в каком-то смысле сильно портить матрицу  $A$  (мера обусловленности матрицы не должна сильно меняться (см. § 11)).

Этим условиям в определенной степени удовлетворяет описываемый ниже метод отражений. Среди методов, требующих для своей реализации число операций  $N \sim 2m^3/3$ , этот метод в настоящее время рассматривается как один из наиболее устойчивых к вычислительной погрешности.

Рассмотрим случай вещественной матрицы  $A$ . Если  $w$  — некоторый вектор-столбец единичной длины,  $(w, w) = 1$ , то матрицу

$$U = E - 2ww^T$$

называют *матрицей отражения*. Под  $\boldsymbol{w}\boldsymbol{w}^T$  здесь понимается матрица, являющаяся произведением вектора-столбца  $\boldsymbol{w}$  на вектор-строку  $\boldsymbol{w}^T$ , т. е.  $\boldsymbol{w}\boldsymbol{w}^T = (w_{ij})$ , где  $w_{ij} = w_i w_j$ . Из определения следует, что  $\boldsymbol{w}\boldsymbol{w}^T$  — симметричная матрица.

Непосредственной проверкой убеждаемся, что  $U = U^T$  и  $UU^T = (E - 2\boldsymbol{w}\boldsymbol{w}^T)(E - 2\boldsymbol{w}\boldsymbol{w}^T)^T = E - 2\boldsymbol{w}\boldsymbol{w}^T - 2\boldsymbol{w}\boldsymbol{w}^T + 4\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{w}\boldsymbol{w}^T = E$ ;

здесь мы воспользовались тем, что

$$\boldsymbol{w}^T\boldsymbol{w} = (\boldsymbol{w}, \boldsymbol{w}) = 1. \quad (1)$$

Таким образом, матрица  $U$  — симметричная и ортогональная.

Напомним один факт из алгебры. Пусть  $U$  и  $B$  — две матрицы размерности  $m$ ,  $B$  — многочлен от  $U$ .  $B = P_i(U)$ ; тогда можно переупорядочить их собственные значения так, что  $\lambda_j^B = P_i(\lambda_j^U)$  при  $j = 1, \dots, m$ .

Поскольку  $U$  симметрична и  $U^2 = UU^T = E$ , а все собственные значения матрицы  $E$  равны 1, то все собственные значения матрицы  $U$  удовлетворяют условию  $\lambda_U^2 = 1$ , т. е. равны или +1, или -1.

Собственному значению -1 отвечает собственный вектор  $-\boldsymbol{w}$ . В самом деле,

$$U\boldsymbol{w} = \boldsymbol{w} - 2\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{w} = \boldsymbol{w} - 2\boldsymbol{w} = -\boldsymbol{w}. \quad (2)$$

Все векторы, ортогональные вектору  $\boldsymbol{w}$ , являются собственными. Им соответствует собственное значение, равное +1. Действительно, пусть  $(\boldsymbol{v}, \boldsymbol{w}) = 0$ , тогда имеем

$$U\boldsymbol{v} = \boldsymbol{v} - 2\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{v} = \boldsymbol{v} - 2\boldsymbol{w}(\boldsymbol{w}, \boldsymbol{v}) = \boldsymbol{v}. \quad (3)$$

Представим произвольный вектор  $\boldsymbol{y}$  в виде  $\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{v}$ , где  $\boldsymbol{z} = d\boldsymbol{w}$ ,  $(\boldsymbol{v}, \boldsymbol{w}) = 0$ . Для этого следует взять в качестве  $\boldsymbol{z}$  проекцию вектора  $\boldsymbol{y}$  на вектор  $\boldsymbol{w}$ , т. е.  $\boldsymbol{z} = (\boldsymbol{y}, \boldsymbol{w})\boldsymbol{w}$ , и  $\boldsymbol{v} = \boldsymbol{y} - (\boldsymbol{y}, \boldsymbol{w})\boldsymbol{w}$ . Вследствие (2) и (3) имеем  $U\boldsymbol{y} = -\boldsymbol{z} + \boldsymbol{v}$ . Таким образом,  $U\boldsymbol{y}$  есть зеркальное отражение вектора  $\boldsymbol{y}$  относительно гиперплоскости, ортогональной вектору  $\boldsymbol{w}$ .

Используя эти геометрические свойства матрицы отражений, нетрудно решить следующую задачу: подобрать вектор  $\boldsymbol{w}$  в матрице отражений так, чтобы заданный вектор  $\boldsymbol{y} \neq 0$  имел в результате преобразования  $U\boldsymbol{y}$  матрицей отражения  $U = E - 2\boldsymbol{w}\boldsymbol{w}^T$  направление заданного единичного вектора  $\boldsymbol{e}$ .

Так как  $U$  — ортогональная матрица, а при ортогональных преобразованиях длины векторов сохраняются, то мы должны иметь  $U\boldsymbol{y} = \alpha\boldsymbol{e}$  или  $U\boldsymbol{y} = -\alpha\boldsymbol{e}$ , где  $\alpha = \sqrt{(\boldsymbol{y}, \boldsymbol{y})}$ . Поэтому направление, перпендикулярное плоскости отражения, будет определяться либо вектором  $\boldsymbol{y} - \alpha\boldsymbol{e}$ , либо вектором  $\boldsymbol{y} + \alpha\boldsymbol{e}$  (см. рис. 6.2.1).

Таким образом, векторы  $w_1 = \pm \rho_1^{-1}(y - ae)$  или  $w_2 = \pm \rho_2^{-1} \times (y + ae)$ , где  $\rho_1 = \sqrt{(y - ae, y - ae)}$ ,  $\rho_2 = \sqrt{(y + ae, y + ae)}$ , будут искомыми. Ясно, что данный процесс всегда осуществим. Если векторы  $y$  и  $e$  коллинеарны, а в этом случае либо  $\rho_1$ , либо  $\rho_2$  будет равно нулю, то никаких отражений делать не надо.

Матрицы отражения нашли широкое применение при численном решении различных задач линейной алгебры (в частности, в рассматриваемой нами задаче приведения матрицы системы уравнений к треугольному виду).

**Лемма.** Произвольная квадратная матрица может быть представлена в виде произведения ортогональной и правой треугольной матриц.

**Доказательство.** Пусть дана квадратная матрица порядка  $m$ . Будем приводить ее к правой треугольной матрице путем последовательного умножения слева на ортогональные матрицы. На первом шаге приведения рассмотрим в качестве вектора  $y$  из предыдущего рассуждения первый столбец матрицы  $A$ :

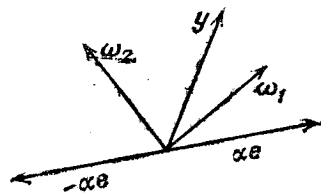


Рис. 6.2.1

$$y_1 = (a_{11}, \dots, a_{m1})^T.$$

Если  $a_{21} = a_{31} = \dots = a_{m1} = 0$ , то переходим к следующему шагу, положив  $A^{(1)} = A$ ,  $U_1 = E$  и введя обозначение  $a_{ij}^{(1)} = a_{ij}$ . В противном случае умножаем матрицу  $A$  слева на матрицу отражения  $U_1 = E_m - 2w_1 w_1^T$ , где  $w_1$  подбирается так, чтобы вектор  $U_1 y_1$  был коллинеарен вектору  $e_1 = (1, 0, \dots, 0)^T$ . Здесь и далее  $E_q$  — единичная матрица размерности  $q$ .

На этом первый шаг закончен, и на следующем шаге будем рассматривать матрицу  $A^{(1)}$  с элементами  $a_{ij}^{(1)}$ , которая либо равна  $A$ , если имеет место первый случай, либо  $A^{(1)} = U_1 A$ , если имеет место второй случай.

Пусть мы уже осуществили  $l - 1 > 0$  шагов и пришли к матрице  $A^{(l-1)}$  с элементами  $a_{ij}^{(l-1)}$  такими, что  $a_{ij}^{(l-1)} = 0$  при  $i > j$ ,  $j = 1, 2, \dots, l - 1$ . В пространстве  $R_{m-l+1}$  векторов размерности  $m - l + 1$  рассмотрим вектор

$$y_l = (a_{l+1, l}^{(l-1)}, a_{l+2, l}^{(l-1)}, \dots, a_{m, l}^{(l-1)})^T.$$

Если  $a_{l+1, l}^{(l-1)} = a_{l+2, l}^{(l-1)} = \dots = 0$ , то переходим к следующему шагу, полагая  $A^{(l)} = A^{(l-1)}$ ,  $U_l = E$ . В противном случае строим матрицу отражения  $V_l = E_{m-l+1} - 2w_l w_l^T$  (размеры матрицы  $V_l$  и вектора  $w_l$  равны  $m - l + 1$ ), переводящую вектор  $y_l$  в вектор, коллинеарный  $e_l = (1, 0, \dots, 0)^T \in R_{m-l+1}$ , и переходим

к матрице

$$A^{(l)} = U_l A^{(l-1)};$$

здесь  $U_l = \begin{bmatrix} E_{l-1} & 0 \\ 0 & V_l \end{bmatrix}$ .

Ясно, что процесс всегда осуществим, и после  $(m-1)$ -го шага мы приходим к матрице

$$A^{(m-1)} = U_{m-1} U_{m-2} \dots U_1 A,$$

имеющей правую треугольную форму.

Если обозначить  $U_{m-1} U_{m-2} \dots U_1 = U$ , то из последнего равенства следует, что  $A = U^T A^{(m-1)}$ , где  $U^T$  — ортогональная, а  $A^{(m-1)}$  — правая треугольная матрицы. Лемма доказана.

Вернемся к решению системы  $Ax = b$ . С помощью указанных преобразований отражения последовательно приводим ее к эквивалентному виду

$$A^{(m-1)} x = Ub,$$

где  $A^{(m-1)}$  — правая треугольная матрица. Если все диагональные элементы  $A^{(m-1)}$  отличны от нуля, то последовательно находим  $x_m, \dots, x_1$ . Если же хотя бы один из диагональных элементов равен нулю, то последняя система вырождена и в силу эквивалентности вырождена и исходная система.

**Задача 1.** Получить асимптотику числа операций метода отражений при  $m \rightarrow \infty$ .

Рассмотрим случай системы уравнений  $Ax = b$  с комплексными  $A$  и  $b$ . Пусть

$$A = A_1 + iA_2, \quad b = b_1 + ib_2, \quad x = x_1 + ix_2.$$

Исходная система уравнений равносильна системе

$$Cy = d \tag{4}$$

с вещественными  $C$  и  $d$ :

$$C = \begin{bmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{bmatrix}, \quad d = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad y = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Поэтому вместо непосредственного решения исходной задачи можно перейти к решению задачи (4) и применить для решения последней метод отражений.

Однако возможен и другой путь — непосредственное применение метода отражений к исходной системе  $Ax = b$ . Здесь матрица отражений  $U = E - 2\omega\omega^*$ ,  $\omega^* = (\bar{\omega}_1, \dots, \bar{\omega}_m)^T$  будет унитарной с собственными значениями вида  $\lambda_U = e^{i\varphi}$ . (Через  $\bar{z}$  обозначено комплексное число, сопряженное с  $z$ .)

**Задача 2.** Перенести метод отражений на случай комплексных матриц.

**Задача 3.** Исследовать метод отражений в случае его применения для решения систем уравнений с ленточной матрицей.

### § 3. Метод простой итерации

Простейшим итерационным методом решения систем линейных уравнений является *метод простой итерации*. Система уравнений

$$Ax = b \quad (1)$$

преобразуется к виду

$$x = Bx + c, \quad (2)$$

и ее решение находится как предел последовательности

$$x^{n+1} = Bx^n + c. \quad (3)$$

Всякая система

$$x = x - D(Ax - b) \quad (4)$$

имеет вид (2) и при  $\det D \neq 0$  эквивалентна системе (1). В то же время всякая система (2), эквивалентная (1), записывается в виде (4) с матрицей  $D = (E - B)A^{-1}$ .

**Теорема** (о достаточном условии сходимости метода простой итерации). *Если  $\|B\| < 1$ , то система уравнений (2) имеет единственное решение и итерационный процесс (3) сходится к решению со скоростью геометрической прогрессии.*

**Доказательство.** Для всякого решения системы (2) имеем  $\|x\| \leq \|B\|\|x\| + \|c\|$ , поэтому справедливо неравенство  $\|x\|(1 - \|B\|) \leq \|c\|$  или  $\|x\| \leq (1 - \|B\|)^{-1}\|c\|$ . Отсюда следует существование и единственность решения однородной системы  $x = Bx$ , а следовательно, и системы (2). Пусть  $X$  — решение системы (2). Из (2) и (3) получаем уравнение относительно погрешности  $r^n = x^n - X$ :

$$r^{n+1} = Br^n. \quad (5)$$

Из (5) получаем равенство

$$r^n = B^n r^0. \quad (6)$$

Отсюда следует, что  $\|r^n\| \leq \|B\|^n \|r^0\| \rightarrow 0$ . Теорема доказана.

Качество итерационного процесса удобно характеризовать скоростью убывания отношения погрешности после  $n$  итераций к начальной погрешности:

$$s_n = \sup_{x^0 \neq X} \frac{\|r^n\|}{\|r^0\|} = \sup_{r^0 \neq 0} \frac{\|B^n r^0\|}{\|r^0\|} = \|B^n\|.$$

Можно гарантировать, что величина  $s_n \leq \varepsilon$ , если  $\|B\|^n \leq \varepsilon$ , т. е. при

$$n \geq n_\varepsilon = \ln(\varepsilon^{-1}) / \ln(\|B\|^{-1}). \quad (7)$$

Если существуют постоянные  $\gamma_{\alpha\beta}$ ,  $\gamma_{\beta\alpha}$  такие, что при  $x \neq 0$

$$\|x\|_\beta / \|x\|_\alpha \leq \gamma_{\alpha\beta}, \quad \|x\|_\alpha / \|x\|_\beta \leq \gamma_{\beta\alpha},$$

то нормы  $\|x\|_\alpha$  и  $\|x\|_\beta$  называются эквивалентными. Имеем

$$\|r^n\|_\beta \leq \gamma_{\alpha\beta} \|r^n\|_\alpha \leq \gamma_{\alpha\beta} \|B\|_\alpha^n \|r^0\|_\alpha \leq \gamma_{\alpha\beta} \gamma_{\beta\alpha} \|B\|_\alpha^n \|r^0\|_\beta.$$

Таким образом, если условие доказанной теоремы выполнено для нормы  $\|\cdot\|_\alpha$ , то утверждение справедливо относительно любой эквивалентной ей нормы.

Любые две нормы в конечномерном пространстве являются эквивалентными. В частности, нормы  $\|x\|_1$ ,  $\|x\|_2$ ,  $\|x\|_\infty$ , вычисляемые соответственно по формулам (2), (3), (4), приведенным во введении к настоящей главе, эквивалентны между собой вследствие справедливости цепочки неравенств

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq m \|x\|_\infty.$$

*Лемма.* Пусть все собственные значения  $\lambda_i$  матрицы  $B$  лежат в круге  $|\lambda| \leq q$ , причем собственным значениям, по модулю равным  $q$ , соответствуют жордановы клетки размерности 1. Тогда существует матрица  $\Lambda = D^{-1}BD$  с нормой  $\|\Lambda\|_\infty \leq q$ .

*Доказательство.* Положим  $\eta = q - \max_{|\lambda_i| < q} |\lambda_i|$ . Собственными значениями матрицы  $\eta^{-1}B$  будут  $\eta^{-1}\lambda_i$ . Преобразуем матрицу  $\eta^{-1}B$  к жордановой форме

$$D^{-1}(\eta^{-1}B)D = \begin{bmatrix} \eta^{-1}\lambda_1 & \alpha_{12} & 0 & \dots \\ 0 & \eta^{-1}\lambda_2 & \alpha_2 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix},$$

где  $\alpha_{i, i+1}$  принимают значения 0 или 1. После умножения на  $\eta$  получим

$$\Lambda = D^{-1}BD = \begin{bmatrix} \lambda_1 & \alpha_{12}\eta & 0 & \dots \\ 0 & \lambda_2 & \alpha_{23}\eta & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}.$$

Если  $|\lambda_i| = q$ , то, согласно условиям леммы,  $\alpha_{i, i+1} = 0$ , поэтому  $|\lambda_i| + |\alpha_{i, i+1}\eta| = q$ . Если  $|\lambda_i| < q$ , то

$$|\lambda_i| + |\alpha_{i, i+1}\eta| \leq \max_{|\lambda_i| < q} |\lambda_i| + \eta = q.$$

Таким образом,  $\|\Lambda\|_\infty = \max_i (|\lambda_i| + |\alpha_{i, i+1}\eta|) \leq q$ .

*Теорема (о необходимом и достаточном условии сходимости метода простой итерации).* Пусть система (2) имеет единственное решение. Итерационный процесс (3) сходится к решению системы (2) при любом начальном приближении тогда и только тогда, когда все собственные значения матрицы  $B$  по модулю меньше 1.

*Доказательство.* Достаточность. Возьмем произвольное  $q$  в пределах  $\max_i |\lambda_i| < q < 1$ . Условие леммы выполнено по отношению к этому  $q$ , поэтому существует матрица  $D$  такая, что  $\|\Lambda\|_\infty \leq q$  при  $\Lambda = D^{-1}BD$ . Поскольку  $B = D\Lambda D^{-1}$ , то

$$B^n = D\Lambda D^{-1}D \dots D^{-1}D\Lambda D^{-1} = D\Lambda^n D^{-1}.$$

Поэтому

$$\|B^n\|_\infty \leq \|D\|_\infty \|D^{-1}\|_\infty q^n \rightarrow 0$$

и

$$\|x^n - X\|_\infty \leq \|D\|_\infty \|D^{-1}\|_\infty q^n \|x^0 - X\|_\infty \rightarrow 0 \quad (8)$$

при  $n \rightarrow \infty$ . Следовательно, и  $\|x^n - X\|_1, \|x^n - X\|_2 \rightarrow 0$ .

Если  $\chi_i$  — координатные орты,  $x = (x_1, \dots, x_m)^T$ , то  $x = \sum_i x_i \chi_i$ . Пусть  $\|\cdot\|$  — некоторая норма, тогда

$$\|x\| \leq \sum_i |x_i| \|\chi_i\| \leq \|x\|_\infty \sum_i \|\chi_i\|.$$

Поэтому при любой норме  $\|\cdot\|$  имеем

$$\|x^n - X\| \leq \left( \sum_i \|\chi_i\| \right) \|D\|_\infty \|D^{-1}\|_\infty q^n \|x^0 - X\|_\infty \rightarrow 0. \quad (9)$$

Соотношения (8), (9) означают также, что любые нормы погрешности убывают быстрее любой геометрической прогрессии со знаменателем, большим  $\max_i |\lambda_i|$ .

**Необходимость.** Пусть  $|\lambda_i| \geq 1$  и  $e_i$  — соответствующий собственный вектор матрицы  $B$ . Тогда при начальном приближении  $x^0 = X + ce_i$ ,  $c \neq 0$ , имеем

$$r^0 = ce_i \quad \text{и} \quad r^n = \lambda_i^n ce_i \not\rightarrow 0 \quad \text{при} \quad n \rightarrow \infty.$$

**Задача 1.** Пусть все собственные значения матрицы  $B$ , за исключением простого  $\lambda_1 = 1$ , лежат внутри единичного круга и система (2) имеет решение  $X$ . Решением системы будут также все  $x = X + ce_1$ . Доказать, что итерационный процесс (3) сходится к одному из таких решений.

#### § 4. Особенности реализации метода простой итерации на ЭВМ

Если все собственные значения матрицы  $B$  лежат внутри единичного круга, то может показаться, что не возникает никаких проблем относительно поведения метода в реальных условиях ограниченности порядков чисел в ЭВМ и присутствия округлений. В обоснование этого иногда приводят следующий довод: возмущения приближений в результате округлений равносильны возмущениям начальных условий итерационного процесса. Поскольку процесс сходящийся, «самоисправляющийся», эти возмущения в конце концов затухнут, и будет получено хорошее приближение к решению исходной задачи.

Однако при решении некоторых систем возникала следующая ситуация. Все собственные значения матрицы  $B$  лежали в круге  $|\lambda| \leq 1/2$ , а итерационный процесс останавливался после некоторого числа итераций из-за переполнения порядков чисел в ЭВМ. В других случаях такого переполнения не происходило, но векторы  $x^n$ , получаемые при вычислениях, не сходились к решению. Последний случай особенно опасен по следующей при-



чине. Можно необоснованно решить, что при условии  $\max |\lambda_i| \leq 1/2$  какое-то определенное число итераций, например 100, заведомо достаточно для получения решения с требуемой точностью. Затем производим эти 100 итераций и рассматриваем полученный результат как требуемый. Поэтому наличие подобных явлений послужило толчком к более детальному исследованию итерационных процессов и формированию новых понятий в теории операторов.

Чтобы понять существо явления, полезно построить пример, где это явление прослеживается в явном виде. В качестве модели выберем итерационный процесс, соответствующий двухдиагональной матрице

$$B_0 = \begin{bmatrix} \alpha & \beta & 0 & \dots & 0 \\ 0 & \alpha & \beta & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \beta \\ 0 & 0 & 0 & \dots & \alpha \end{bmatrix}.$$

При возведении матрицы  $B_0$  в степень  $n$  получается треугольная матрица

$$B_0^n = [b_{ij}^{(n)}] = \begin{bmatrix} \alpha^n & C_n^1 \alpha^{n-1} \beta & C_n^2 \alpha^{n-2} \beta^2 & \dots \\ 0 & \alpha^n & C_n^1 \alpha^{n-1} \beta & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

с элементами  $b_{ij}^{(n)} = C_n^{j-i} \alpha^{n-(j-i)} \beta^{j-i}$ . Если  $r^0 = (0, \dots, 0, 1)^T$ , то

$$r^n = B_0^n r^0 = (b_{1m}^{(n)}, \dots, b_{mm}^{(n)})^T, \quad \|r^n\|_1 = \sum_{i=1}^m |b_{im}^{(n)}|.$$

При  $n < m$  последнее выражение упрощается:

$$\begin{aligned} \|r^n\|_1 &= \sum_{i=1}^m C_n^{m-i} |\alpha|^{n-(m-i)} |\beta|^{m-i} = \\ &= \sum_{k=0}^{m-1} C_n^k |\alpha|^{n-k} |\beta|^k = \sum_{k=0}^n C_n^k |\alpha|^{n-k} |\beta|^k = (|\alpha| + |\beta|)^n. \end{aligned}$$

Рассмотрим случай  $|\alpha| < 1$ ,  $|\alpha| + |\beta| > 1$ ,  $|\beta/(1-\alpha)| < 1$ . Пусть  $c = c^0 = (0, \dots, 0, 1)^T$ . Непосредственно проверяется, что при таком  $c$  решением рассматриваемой системы будет

$$X^0 = \left( \frac{1}{1-\alpha} \left( \frac{\beta}{1-\alpha} \right)^{m-1}, \dots, \frac{1}{1-\alpha} \right)^T.$$

Справедлива оценка

$$\|X^0\|_1 \leq \omega,$$

где

$$\omega = \frac{1}{|1-\alpha|} \sum_{k=0}^{\infty} \left| \frac{\beta}{1-\alpha} \right|^k = \frac{1}{|1-\alpha| \left( 1 - \left| \frac{\beta}{1-\alpha} \right| \right)}.$$

При начальном приближении  $x^0 = X^0 + c^0$  имеем  $r^0 = c^0$  и, согласно проводившимся выше построениям,

$$\|r^n\|_1 = (|\alpha| + |\beta|)^n \text{ для } n < m.$$

Выберем  $m$  таким, чтобы число  $\sigma = [ (|\alpha| + |\beta|)^{m-1} - \omega ] / m$  превосходило пределы, допустимые в ЭВМ. Из полученных ранее соотношений следует, что

$$\|x^{m-1}\|_\infty \geq \|x^{m-1}\|_1 / m \geq (\|r^{m-1}\|_1 - \|x^0\|_1) / m \geq \sigma.$$

Поэтому построенный пример обладает следующими свойствами: норма начального приближения невелика, итерационный процесс сходится при отсутствии округлений и ограничения на порядки чисел в ЭВМ, но останавливается не позднее чем при  $n = m - 1$  из-за недопустимо больших значений компонент приближений.

Обратимся к реальной ситуации, когда на каждом шаге вычислений происходят округления. Рассмотрим подробнее случай, когда переполнение не происходит. Вместо  $x^n$  получаются векторы  $x^{*n}$ , связанные соотношениями

$$x^{*n+1} = Bx^{*n} + c + \rho^n,$$

где  $\rho^n$  — суммарное округление на шаге итерации.

Отсюда и из (3.2) получается уравнение относительно погрешности  $r^{*n} = x^{*n} - X$ :

$$r^{*n+1} = \rho^n + Br^{*n}. \quad (1)$$

Выражая каждое  $r^{*n}$  через предыдущее, получаем

$$\begin{aligned} r^{*n} &= \rho^{n-1} + Br^{*n-1} = \rho^{n-1} + B(\rho^{n-2} + Br^{*n-2}) = \\ &= \rho^{n-1} + B\rho^{n-2} + \dots + B^{n-1}\rho^0 + B^n r^0. \end{aligned} \quad (2)$$

Как мы видели, норма  $\|B_0^n\|$  при  $|\alpha| < 1$ ,  $|\alpha| + |\beta| > 1$  имеет следующий характер поведения: при малых  $n$  она имеет тенденцию к возрастанию, при больших  $n$  стремится к нулю. (Можно показать, что максимальное значение  $\varphi(B_0) = \max_n \|B_0^n\|$  достигается при значении  $n = n_0$  порядка  $m$ .) При таком характере поведения норм  $\|B^n\|$  может возникнуть следующая ситуация. Величина  $\max_n \|x^{*n}\|$  не настолько велика, чтобы происходило переполнение и остановка ЭВМ; в то же время  $\varphi(B)2^{-t} \gg R$ , где  $R$  — максимально допустимая погрешность решения. Поэтому, как правило, при  $n > n_0$  среди слагаемых в правой части (2) присутствует слагаемое  $B^{n_0}\rho^{n-1-n_0}$  с нормой, много большей чем  $R$ . В результате установление приближений  $x^n$  с приемлемой точностью не происходит.

Подведем некоторый итог проведенных построений. Матрицы высокой размерности обладают свойствами, существенно отличными от свойств матриц малой размерности. Кроме собственных значений у таких матриц есть почти собственные значения, т. е.  $\lambda$  такие, что  $\|Ax - \lambda x\| \leq \varepsilon \|x\|$  при  $\|x\| \neq 0$  и очень малом  $\varepsilon$ .

Например, в случае матрицы  $B_0$  при любом  $\lambda$ , лежащем в круге  $|\alpha - \lambda| < |\beta|$ , можно построить вектор  $x_\lambda$  такой, что  $\|B_0 x_\lambda - \lambda x_\lambda\|_\infty \leq \varepsilon_\lambda \times \|x_\lambda\|_\infty$ , где  $\varepsilon_\lambda = |\beta| |(\lambda - \alpha)/\beta|^m$ .

Поведение степеней матрицы  $B^n$  при  $n$  порядка  $m$  определяется во многом такими «почти собственными векторами»  $x_\lambda$  и «почти собственными значениями»  $\lambda$ .

**Задача 2.** Построить «почти собственный вектор»  $x_\lambda$ , соответствующий такому значению  $\varepsilon_\lambda$ .

Суммарная вычислительная погрешность  $\rho_n = \sum_{j=0}^{n-1} B^{n-1-j} \rho^j$  может оказаться большой не только из-за большой величины отдельных слагаемых, но и из-за того, что их много.

Пусть  $B$  — симметричная матрица и  $\|B\|_2 = \max_i |\lambda_B^i| = \lambda_B^1 < 1$ ,  $e^1$  — соответствующий  $\lambda_B^1$  нормированный собственный вектор. Предположим, что на каждом  $j$ -м шаге происходит округление  $\rho^j = \rho e^1$ , где  $\rho$  порядка  $2^{-t}$ . Имеем равенство

$$\rho_n = \rho \sum_{j=0}^{n-1} (\lambda_B^1)^j e^1 = \rho \frac{1 - (\lambda_B^1)^n}{1 - \lambda_B^1} e^1.$$

Поскольку число итераций берется таким, что  $\|B^n\| \ll 1$ , а  $\|B^n\| = (\lambda_B^1)^n$ , то можно считать, что  $\|\rho_n\| \approx \rho / (1 - \lambda_B^1)$ . Таким образом, если  $\lambda_B^1$  близко к 1, то суммарное влияние округлений на шагах интегрирования может оказаться довольно большим.

Покажем, что вычислительная погрешность такого порядка является неизбежной. Предположим, что вместо системы (3.2) решается система  $x = Bx + c + \rho e^1$ . Разность  $\bar{X} - X$  решений этих систем удовлетворяет соотношению  $(\bar{X} - X) = B(\bar{X} - X) + \rho e^1$ , отсюда  $\bar{X} - X = (E - B)^{-1} \rho e^1 = (1 - \lambda_B^1)^{-1} \rho e^1$ . Поэтому погрешность порядка  $(1 - \lambda_B^1)^{-1} \rho$  является неустраняемой; возмущение приближений, создаваемое в ходе итераций, сравнимо с неустраняемой погрешностью.

## § 5. $\delta^2$ -процесс практической оценки погрешности и ускорения сходимости

Рассмотрим вопрос об оценке погрешности приближенного решения системы уравнений. Если  $X^*$  — приближенное решение системы  $Ax = b$ , а  $X$  — точное решение этой системы, то можно написать равенство  $A(X^* - X) = AX^* - b$ . Отсюда следует оценка погрешности

$$\|X^* - X\| = \|A^{-1}(AX^* - b)\| \leq \|A^{-1}\| \|AX^* - b\|,$$

которая редко применяется из-за сложности оценки  $\|A^{-1}\|$ . Поэтому при практическом анализе погрешности приближений, получаемых итерационными методами, обычно вместо этой оценки используется рассматриваемая далее нестрогая, но более простая оценка погрешности, которая строится на основании дополнительной информации, получаемой в процессе вычислений.

Примем следующий критерий разумности практической оценки погрешности:  $v^n$  принимается за практическую погрешность

приближения  $x^n$ , стремящегося к  $X$  при  $n \rightarrow \infty$ , если

$$\|v^n - (x^n - X)\| / \|x^n - X\| \rightarrow 0 \quad \text{при } n \rightarrow \infty. \quad (1)$$

Ясно, что тогда  $\|v^n\| \sim \|x^n - X\|$ .

Рассмотрим метод простой итерации  $x^{n+1} = Bx^n + c$ . Для краткости изложения ограничимся случаем, когда матрица  $B$  простой структуры (т. е. ее жорданова форма диагональна, и поэтому она обладает полной системой собственных векторов).

Пусть  $\lambda_i$ ,  $i = 1, \dots, m$ , — собственные значения матрицы  $B$ , занумерованные в порядке убывания  $|\lambda_i|$ , причем  $1 > |\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_m|$ , а  $e_i$ ,  $\|e_i\| = 1$  — соответствующие собственные векторы, образующие полную систему. Разложим вектор  $r^0$  по базису  $e_i$ :  $r^0 = \sum c_i e_i$ . Тогда

$$x^n - X = B^n r^0 = \sum c_i \lambda_i^n e_i = c_1 \lambda_1^n e_1 + O(|\lambda_2|^n). \quad (2)$$

Здесь и далее выражение  $x_n = y_n + O(\epsilon_n)$  имеет следующий смысл:

$$\|x_n - y_n\| = O(\epsilon_n) \quad \text{при } n \rightarrow \infty.$$

Далее в этом параграфе  $\|x\|$  — это  $\|x\|_2$ .

Укажем способ построения приближения к вектору  $w^n = c_1 \lambda_1^n e_1$  на основании информации, получающейся в ходе вычислений. Согласно (2) имеем

$$\begin{aligned} x^{n-2} - X &= w^n \lambda_1^{-2} + O(|\lambda_2|^n), \\ x^{n-1} - X &= w^n \lambda_1^{-1} + O(|\lambda_2|^n), \\ x^n - X &= w^n + O(|\lambda_2|^n). \end{aligned}$$

Вычитая друг из друга соседние соотношения, получим

$$\begin{aligned} x^{n-1} - x^{n-2} &= w^n (1 - \lambda_1^{-1}) \lambda_1^{-1} + O(|\lambda_2|^n), \\ x^n - x^{n-1} &= w^n (1 - \lambda_1^{-1}) + O(|\lambda_2|^n). \end{aligned} \quad (3)$$

Отсюда

$$\begin{aligned} (x^n - x^{n-1}, x^n - x^{n-1}) &= \|w^n\|^2 |1 - \lambda_1^{-1}|^2 + O(\|w^n\| |\lambda_2|^n), \\ (x^{n-1} - x^{n-2}, x^n - x^{n-1}) &= \|w^n\|^2 |1 - \lambda_1^{-1}|^2 \lambda_1^{-1} + O(\|w^n\| |\lambda_2|^n). \end{aligned} \quad (4)$$

Положим

$$\lambda_1^{(n)} = \frac{(x^n - x^{n-1}, x^n - x^{n-1})}{(x^{n-1} - x^{n-2}, x^n - x^{n-1})}.$$

Воспользуемся соотношениями (4) и в предположении  $c_1 \neq 0$  поделим числитель и знаменатель в выражении для  $\lambda_1^{(n)}$  на  $\|w^n\|^2 \times |1 - \lambda_1^{-1}|^2 \lambda_1^{-1}$ , получим

$$\lambda_1^{(n)} = \frac{\lambda_1 + O\left(\frac{|\lambda_2|^n}{\|w^n\|}\right)}{1 + O\left(\frac{|\lambda_2|^n}{\|w^n\|}\right)}.$$

Поскольку

$$\|w^n\| = |c_1| |\lambda_1|^n, \quad (5)$$

то

$$\lambda_1^{(n)} = \lambda_1 + O(|\lambda_2/\lambda_1|^n). \quad (6)$$

Поделив второе из соотношений (3) на  $1 - (\lambda_1^{(n)})^{-1}$ , получим

$$\begin{aligned} \frac{x^n - x^{n-1}}{1 - (\lambda_1^{(n)})^{-1}} &= w^n \frac{1 - \lambda_1^{-1}}{1 - (\lambda_1^{(n)})^{-1}} + O(|\lambda_2|^n) = \\ &= w^n + w^n \frac{\lambda_1 - \lambda_1^{(n)}}{\lambda_1(\lambda_1^{(n)} - 1)} + O(|\lambda_2|^n). \end{aligned}$$

Из (5), (6) следует  $\|w^n(\lambda_1 - \lambda_1^{(n)})\| = O(|\lambda_2|^n)$ ; поэтому

$$\frac{x^n - x^{n-1}}{1 - (\lambda_1^{(n)})^{-1}} = w^n + O(|\lambda_2|^n).$$

Отсюда и из (2) получаем

$$x^n - X = v^n + O(|\lambda_2|^n),$$

где  $v_n = (x^n - x^{n-1}) / (1 - (\lambda_1^{(n)})^{-1})$ . Заметим, что согласно (3), (6)  $\|v^n\| = |c_1| |\lambda_1|^n + O(|\lambda_2|^n)$ . Из этих равенств вытекает, что  $v^n$  удовлетворяет критерию (1), и поэтому его можно принять за практическую погрешность приближения  $x^n$ .

В случае  $c_1 = \dots = c_l = 0$ ,  $c_{l+1} \neq 0$  проведенные рассуждения останутся в силе, если  $|\lambda_{l+1}| > |\lambda_{l+2}|$ . Во всех соотношениях следует заменить лишь  $\lambda_i$ ,  $c_i$ ,  $e_i$  при  $i = 1, 2$  на  $\lambda_{l+i}$ ,  $c_{l+i}$ ,  $e_{l+i}$ . Описанный способ получения оценки приближенного решения называется  $\delta^2$ -процессом.

Если положить  $y^n = x^n - v^n$ , то  $y^n - X = O(|\lambda_2|^n)$ , и поэтому  $y^n$ , вообще говоря, является лучшим начальным условием для последующих итераций по сравнению с  $x^n$ . Производя время от времени такие уточнения, иногда удается существенно уменьшить общее число итераций.

Для справедливости приближенного равенства

$$x^n - X \approx v^n \quad (7)$$

необходимо, чтобы в правой части равенства

$$x^n - X = \sum_i c_i \lambda_i^n e_i$$

одно из слагаемых преобладало над остальными. Если это так, то векторы  $x^n - x^{n-1}$ ,  $x^{n-1} - x^{n-2}$  приблизительно пропорциональны и

$$\mu_n = \frac{|(x^{n-1} - x^{n-2}, x^{n-1} - x^n)|}{\|x^{n-1} - x^{n-2}\| \|x^{n-1} - x^n\|} \approx 1.$$

Таким образом, условие  $\mu_n \approx 1$  является необходимым для того, чтобы проводившиеся ранее построения были справедливы. Поэтому его можно принять за условие практической применимости (7).

Например, возможна следующая схема метода простой итерации с применением  $\delta^2$ -процесса ускорения сходимости. Задаются некоторым  $\eta'$  в пределах  $1 > \eta' > 0$  и малым  $\eta > 0$ . Если по ходу итераций оказалось, что  $\mu_n \geq 1 - \eta$ , то вычисляется  $v^n$  и вектор  $y^n$  принимается за начальное приближение для последующих итераций. Итерационный процесс прекращается, если  $\mu_n \geq 1 - \eta'$  и  $\|v^n\| \leq \varepsilon$ , где  $\varepsilon$  — требуемая точность.

Если  $\eta$  очень мало, то условие  $\mu_n \geq 1 - \eta$  будет выполняться только после большого числа итераций, ускорение сходимости не будет иметь места. При большом  $\eta$  соотношения, положенные в основу наших построений, выполняются грубо, поэтому не исключено, что применение  $\delta^2$ -процесса сходимости замедлит итерационный процесс. Картина итераций также осложняется наличием погрешностей округлений, так что описанная выше схема требует практической отработки на большом числе примеров с целью выбора оптимальных  $\eta'$ ,  $\eta$  и указания нижней границы значений  $\varepsilon$ , при которых алгоритм применим. Если однородный итерационный процесс подвергается перестройке (в нашем случае при переходе от  $x^n$  к  $y^n$ ), то иногда полезно проверить, не ведет ли эта перестройка к ухудшению. В качестве критерия целесообразности перестройки можно взять некоторое соотношение, связывающее нормы невязок для  $x^n$ ,  $y^n$ , например неравенство вида

$$\|(E - B)y^n - c\| \leq q \|(E - B)x^n - c\|.$$

Замечание о необходимости указания нижней грани значений  $\varepsilon$  вызывается следующим обстоятельством. Пусть для определенности  $\lambda_1 > 0$ . Уже при вычислении  $x^n$  по заданному  $x^{n-1}$  погрешности округления могут возмутить результат на величину  $\delta x^n$  с нормой порядка  $\rho$ . Следствием этого может явиться возмущение  $\delta v^n$ , имеющее норму порядка  $(1 - \lambda_1)^{-1}\rho$ . Отсюда следует, что в случае  $\varepsilon < (1 - \lambda_1)^{-1}\rho$  итерационный процесс может никогда не закончиться.

Проведенные построения показывают, что при реализации метода возникает много таких моментов, разбор которых требует серьезной математической подготовки и проведения большой серии численных экспериментов. Поэтому, несмотря на «простоту» метода простой итерации, будет вполне оправданным создание стандартной программы этого метода.

## § 6. Оптимизация скорости сходимости итерационных процессов

Рассмотрим простейший итерационный способ решения системы уравнений  $Ax = b$ :

$$x^{n+1} = x^n - \alpha(Ax^n - b).$$

Мы видели, что скорость сходимости такого итерационного процесса существенно зависит от максимального модуля собственных значений матрицы  $B = E - \alpha A$ . Если  $\lambda_1, \dots, \lambda_n$  — собственные значения матрицы  $A$ , то  $\max_i |\lambda_i(B)| = \max_i |1 - \alpha \lambda_i|$ .

Из рис. 6.6.1 видно, что при действительных собственных значениях различных знаков этот максимум больше 1 и итерационный процесс расходится.

Обратимся к часто встречающемуся случаю, когда все  $\lambda_i > 0$ . Значения  $\lambda_i$  бывают известны крайне редко, однако довольно типичен случай, когда известна оценка для этих чисел вида  $0 < \mu \leq \lambda_i \leq M < \infty$  при всех  $i$ . Скорость сходимости ите-

рационального процесса можно характеризовать величиной

$$\rho(\alpha) = \max_{\mu \leq \lambda \leq M} |1 - \alpha\lambda|.$$

Рассмотрим задачу минимизации  $\rho(\alpha)$  за счет выбора  $\alpha$ .

Для нахождения  $\min_{\alpha} \rho(\alpha)$  удобно обратиться к геометрической картине (рис. 6.6.2). Ясно, что  $\rho(\alpha) \geq 1$  при  $\alpha \leq 0$ . При  $0 < \alpha \leq M^{-1}$  функция  $1 - \alpha\lambda$  неотрицательна и монотонно убывает на отрезке  $[\mu, M]$ , поэтому  $\rho(\alpha) = 1 - \alpha\mu$ . При  $M^{-1} < \alpha$  величина  $1 - \alpha M$  отрицательна и модуль ее растет

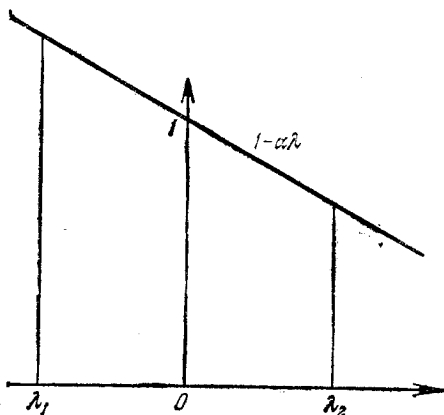


Рис. 6.6.1

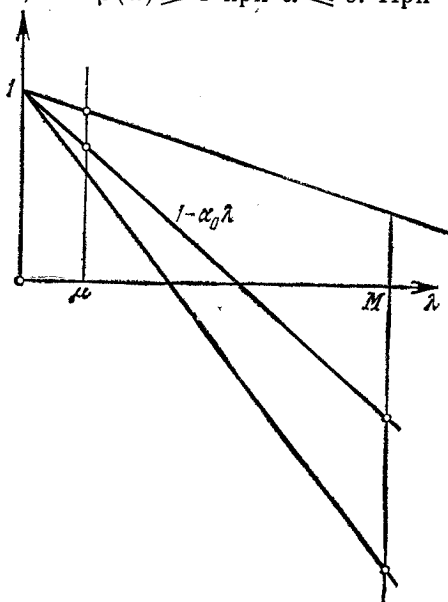


Рис. 6.6.2

с ростом  $\alpha$ . При некотором  $\alpha = \alpha_0$  наступит момент, когда

$$1 - \alpha_0\mu = -(1 - \alpha_0M), \quad (1)$$

и тогда  $\rho(\alpha_0) = |1 - \alpha_0\mu|$ . Если  $\alpha < \alpha_0$ , то  $\rho(\alpha) = 1 - \alpha\mu > 1 - \alpha_0\mu = \rho(\alpha_0)$ ; если  $\alpha_0 < \alpha$ , то  $\rho(\alpha) \geq |1 - \alpha M| = M\alpha - 1 \geq M\alpha_0 - 1 = \rho(\alpha_0)$ . Таким образом, значение  $\alpha = \alpha_0$  является искомым. Решая уравнение (1) относительно  $\alpha_0$ , получим  $\alpha_0 = 2/(M + \mu)$ . Отсюда

$$\rho(\alpha_0) = (M - \mu)/(M + \mu).$$

**Задача 1.** Доказать сходимость итерационного процесса при  $\alpha = \|A\|^{-1}$ .

На примере систем с матрицей  $A > 0$  (здесь и далее неравенство  $A > 0$  означает, что  $A$  — симметричная положительно определенная матрица) рассмотрим более формализованные постановки проблем оптимизации скорости сходимости итерационных процессов,

Если число ненулевых элементов матрицы много больше ее размерности, то операция умножения матрицы на вектор более трудоемка, чем умножение числа на вектор или сложение векторов. Поэтому при оценке трудоемкости итерационных процессов и оптимизации этих процессов далее за меру трудоемкости мы неявно принимаем число умножений матрицы  $A$  на вектор.

Всякая система  $Ax = b$  с  $\det A \neq 0$ , вообще говоря, может быть приведена к виду, где  $A > 0$ , умножением на матрицу  $A^T$  (как говорят, *симметризована*). В самом деле, система  $A^T Ax = A^T b$  эквивалентна исходной, матрица  $A^T A$  симметричная, так как  $A^T A = (A^T A)^T$ , и положительно определена, так как  $(A^T Ax, x) = \|Ax\|^2 > 0$  при  $x \neq 0$ . По возможности стараются избежать симметризации, поскольку, как мы увидим далее, она часто приводит к ухудшению сходимости итерационных процессов.

Пусть  $x^0$  — начальное приближение и каждое следующее приближение  $x^k$  построено по формуле

$$x^k = x^0 - Q_{k-1}(A)(Ax^0 - b), \quad (2)$$

где  $Q_{k-1}(A)$  — многочлен степени  $k-1$  от матрицы  $A$ .

Предположим, известно, что спектр  $S(A)$  матрицы  $A$  (спектром матрицы называется множество ее собственных значений) принадлежит отрезку  $[\mu, M]$ , где  $0 < \mu \leq M < \infty$ . Для решения  $X$  исходной системы имеем

$$X = X - Q_{k-1}(A)(AX - b).$$

Вычитая это равенство из (2), получим уравнение для погрешности

$$r^k = r^0 - Q_{k-1}(A)Ar^0 = P_k(A)r^0,$$

где  $P_k(\lambda) = 1 - \lambda Q_{k-1}(\lambda)$ . Матрица  $P_k(A)$  симметрична как функция симметричной матрицы, поэтому

$$\sup_{r^0} \frac{\|r^k\|_2}{\|r^0\|_2} = \|P_k(A)\|_2 = \max_{\lambda \in S(A)} |P_k(\lambda)| \leq \sigma(P_k(\lambda)), \quad (3)$$

где  $\sigma(g) = \max_{\mu \leq \lambda \leq M} |g(\lambda)|$ . Далее поставим перед собой задачу минимизации  $\sigma(P_k)$  за счет выбора многочлена  $Q_{k-1}$ , т. е. задачу нахождения

$$\sigma_k = \inf_{Q_{k-1}} \sigma(P_k).$$

*Лемма. Справедливо равенство*

$$\sigma_k = \max_{\mu \leq \lambda \leq M} |1 - \lambda Q_{k-1}^0(\lambda)| = |t_k|^{-1}; \quad (4)$$

здесь

$$1 - \lambda Q_{k-1}^0(\lambda) = P_k^0(\lambda) = T_k \left( \frac{2\lambda - (M + \mu)}{M - \mu} \right) \left( T_k \left( -\frac{M + \mu}{M - \mu} \right) \right)^{-1}, \quad (5)$$

как обычно,  $T_k(x)$  — многочлены Чебышева;  $t_k = T_k \left( -\frac{M + \mu}{M - \mu} \right)$ .



Доказательство. Воспользовавшись соотношением  $\max_{|x| \leq 1} |T_k(x)| = 1$ , получаем  $\sigma(P_k^0) = |t_k|^{-1}$ . Предположим, что существует многочлен  $P_k(\lambda) = 1 - \lambda Q_{k-1}(\lambda)$  такой, что

$$\sigma(P_k) < \sigma(P_k^0) = |t_k|^{-1}. \quad (6)$$

Рассмотрим многочлен  $S_k(\lambda) = P_k^0(\lambda) - P_k(\lambda)$ . Пусть

$$\lambda^j = \frac{M+\mu}{2} + \frac{M-\mu}{2} \cos \frac{\pi j}{k}, \quad j=0, \dots, k.$$

Из равенства  $T_k(x) = \cos(k \arccos x)$  имеем

$$P_k^0(\lambda^j) = (-1)^j |t_k|^{-1}.$$

Поскольку  $\lambda^j \in [\mu, M]$ , то, согласно (6),

$$|P_k(\lambda^j)| < |t_k|^{-1}.$$

Отсюда следует, что  $\text{sign } S_k(\lambda^j) = \text{sign } P_k^0(\lambda^j)$ .

Точки  $\lambda^0, \dots, \lambda^k$  расположены монотонно на отрезке  $[\mu, M]$ . Поскольку  $S_k(\lambda)$  меняет знак при переходе от каждой из этих точек к следующей, то  $S_k(\lambda)$  имеет  $k$  корней на  $[\mu, M]$ . Кроме того,

$$S_k(0) = P_k^0(0) - P_k(0) = 1 - 1 = 0.$$

Мы получили многочлен степени  $k$ , который имеет  $k+1$  нуль, следовательно,  $S_k(\lambda) \equiv 0$ ,  $P_k(\lambda) \equiv P_k^0(\lambda)$ ,

$$\max_{[\mu, M]} |P_k(\lambda)| = |t_k|^{-1}.$$

Мы пришли к противоречию с (6). Лемма доказана.

Воспользуемся явным представлением многочленов Чебышева

$$T_n(x) = (\lambda_1^n + \lambda_2^n)/2, \quad \lambda_{1,2} = x \pm \sqrt{x^2 - 1}.$$

При  $x = -\frac{M+\mu}{M-\mu}$  имеем

$$\begin{aligned} x \pm \sqrt{x^2 - 1} &= -\frac{M+\mu}{M-\mu} \pm \sqrt{\left(\frac{M+\mu}{M-\mu}\right)^2 - 1} = -\frac{M+\mu}{M-\mu} \pm \\ &\pm \frac{2\sqrt{M\mu}}{M-\mu} = -\frac{(\sqrt{M} \pm \sqrt{\mu})^2}{M-\mu} = -\frac{(\sqrt{M} \pm \sqrt{\mu})^2}{(\sqrt{M} + \sqrt{\mu})(\sqrt{M} - \sqrt{\mu})}. \end{aligned} \quad (7)$$

Введем обозначение  $\lambda_0 = (\sqrt{M} + \sqrt{\mu})/(\sqrt{M} - \sqrt{\mu})$ . Из (7) имеем  $\lambda_1 = -\lambda_0$ ,  $\lambda_2 = -\lambda_0^{-1}$ . Так как  $|\lambda_2| < 1$ , то при больших  $n$

$$t_n = \left| T_n\left(-\frac{M+\mu}{M-\mu}\right) \right| \sim \frac{1}{2} \lambda_0^n.$$

Поскольку  $\lambda_1^n$  и  $\lambda_2^n$  одного знака, то

$$|t_n| \geq |\lambda_1|^n / 2 = \lambda_0^n / 2. \quad (8)$$

Из представления  $t_k$  следует, в частности, что

$$\sigma_k = \frac{1}{|t_k|} = \frac{1}{\text{ch}(k \ln \lambda_0)} < 1 \text{ при } k > 0.$$

На основе приведенных построений можно предложить несколько типов итерационных процессов.

В одном случае задаются последовательностью значений  $n_0 = 0 < n_1 < n_2 < \dots$ , приближения  $x^{n_i}$  определяют по рекуррентной формуле

$$x^{n_{i+1}} = x^{n_i} - Q_{q_i-1}^0(A)(Ax^{n_i} - b), \quad (9)$$

где  $q_i = n_{i+1} - n_i$ . Имеем

$$r^{n_{i+1}} = P_{q_i}^0(A)r^{n_i}, \quad \|r^{n_{i+1}}\|_2 \leq \sigma_{q_i} \|r^{n_i}\|_2$$

и в итоге

$$\|r^{n_p}\|_2 \leq \sigma_{q_0} \dots \sigma_{q_{p-1}} \|r^0\|_2.$$

Рассмотрим случай  $q_i \equiv k$ , т. е.  $n_i = ik$ . Тогда, обозначив  $x^{n_i} = y^i$ , можно записать итерационный процесс (9) в виде

$$y^{i+1} = y^i - Q_{k-1}^0(A)(Ay^i - b).$$

Соответствующая оценка погрешности имеет вид

$$\|y^i - X\|_2 \leq (\sigma_k)^i \|y^0 - X\|_2.$$

Такой итерационный процесс называют *оптимальным* (по числу итераций) *линейным  $k$ -шаговым итерационным процессом*. В частном случае  $k = 1$ , согласно (5), выполняются соотношения

$$P_1^0(\lambda) = 1 - \lambda Q_0^0(\lambda) = \frac{2\lambda - M - \mu}{\frac{M - \mu}{-M + \mu}} = 1 - \lambda \frac{2}{M + \mu},$$

$$t_1 = \frac{1}{\frac{M + \mu}{-M - \mu}}, \quad \sigma_1 = \frac{1}{|t_1|} = \frac{M - \mu}{M + \mu}.$$

Таким образом, *оптимальный линейный одношаговый итерационный процесс* имеет вид

$$y^{i+1} = y^i - \frac{2}{M + \mu} (Ay^i - b),$$

а погрешность оценивается следующим образом:

$$\|y^i - X\|_2 \leq \left(\frac{M - \mu}{M + \mu}\right)^i \|y^0 - X\|_2 \quad (10)$$

(этот метод мы уже построили выше в этом параграфе).

Можно проверить, что коэффициенты многочленов  $Q_{k-1}^0(A)$  быстро растут с  $k$ , поэтому при больших  $k$  алгоритм вычисления  $x^n$ , использующий информацию о значениях этих коэффициентов, сильно чувствителен к вычислительной погрешности. В связи с этим для вычисления  $x^k$  прибегают к окольным путям.

Запишем соотношение (2) в виде

$$x^k = P_k(A)x^0 - (P_k(A) - E)A^{-1}b$$

и будем рассматривать его как преобразование  $x^0 \rightarrow x^k$ .

Пусть  $P^{(1)}(\lambda)$  и  $P^{(2)}(\lambda)$  — произвольные многочлены и  $P(\lambda) = P^{(1)}(\lambda)P^{(2)}(\lambda)$ . Произведем последовательно преобразования, соответствующие многочленам  $P^{(1)}(\lambda)$  и  $P^{(2)}(\lambda)$ ,

$$\begin{aligned} x^{(1)} &= P^{(1)}(A)x^0 - (P^{(1)}(A) - E)A^{-1}b, \\ x^{(2)} &= P^{(2)}(A)x^{(1)} - (P^{(2)}(A) - E)A^{-1}b. \end{aligned} \quad (11)$$

Имеем равенства

$$\begin{aligned} x^{(2)} &= P^{(2)}(A)P^{(1)}(A)x^0 - P^{(2)}(A)(P^{(1)}(A) - E)A^{-1}b - \\ &- (P^{(2)}(A) - E)A^{-1}b = P^{(2)}(A)P^{(1)}(A)x^0 - (P^{(2)}(A)P^{(1)}(A) - E)A^{-1}b. \end{aligned}$$

Таким образом, последовательное осуществление преобразований (11) равносильно преобразованию, соответствующему многочлену  $P(A)$ . Отсюда следует, что при  $P(\lambda) = P^{(1)}(\lambda) \dots P^{(k)}(\lambda)$  последовательное осуществление преобразований, соответствующих многочленам  $P^{(i)}(\lambda)$ ,  $i = 1, \dots, k$ , приводит к тому же результату, что и преобразование, соответствующее многочлену  $P(\lambda)$ .

Для вычисления  $x^k = x^0 - Q_{k-1}^0(A)(Ax^0 - b)$  поступим следующим образом. Разложим многочлен  $P_k^0(\lambda) = 1 - \lambda Q_{k-1}^0(\lambda)$  на множители  $P_k^0(\lambda) = (1 - \alpha_1^k \lambda) \dots (1 - \alpha_k^k \lambda)$  и произведем последовательно преобразования, соответствующие этим множителям. Если  $P(\lambda) = 1 - \alpha_i^k \lambda$ , то  $Q(\lambda) = (1 - P(\lambda))\lambda^{-1} = \alpha_i^k$  и эти преобразования имеют вид

$$x^{(i)} = x^{(i-1)} - \alpha_i^k (Ax^{(i-1)} - b), \quad i = 1, \dots, k. \quad (12)$$

В результате получим  $x^{(k)}$ , равное  $x^k$ . Нулями многочленов  $P_k^0(\lambda)$  являются  $\lambda_i = (\alpha_i^k)^{-1}$ , нулями многочленов  $T_k\left(\frac{2\lambda - (M + \mu)}{M - \mu}\right)$  являются

$$\lambda_i = \frac{M + \mu}{2} + \frac{M - \mu}{2} \cos \frac{\pi(2i - 1)}{2k}, \quad i = 1, \dots, k.$$

Отсюда

$$\alpha_i^k = 2 / \left( M + \mu + (M - \mu) \cos \frac{\pi(2i - 1)}{2k} \right). \quad (13)$$

При больших  $k$  алгоритм вычисления по формулам (12), (13) также неустойчив к погрешностям округления. Вычитая из (12) равенство  $X = X - \alpha_i^k (AX - b)$ , получим уравнение для погрешности  $r^{(i)} = (E - \alpha_i^k A) r^{(i-1)}$ . При наличии округлений оно запишется в виде

$$r^{(i)} = (E - \alpha_i^k A) r^{(i-1)} + \rho^{i-1}.$$

Последовательно выражая  $r^{(i)}$  через предыдущие, имеем равенство

$$r^{(k)} = \prod_{i=1}^k (E - \alpha_i^k A) r^{(0)} + \sum_{i=0}^{k-1} \prod_{i+2 \leq j \leq k} (E - \alpha_j^k A) \rho^i.$$

Здесь к  $r^{(0)}$  применяется оператор  $P_k^0(A)$  с нормой  $\sigma_k = |t^k|^{-1} < 1$ , в то время как операторы, применяемые к  $\rho_i$ , могут быть с очень большими нормами.

При реальных вычислениях для обеспечения устойчивости алгоритма к округлениям осуществляют «перемешивание» чисел  $\alpha_i^k$ . Алгоритм перемешивания в случае  $k = 2^l$  заключается в следующем. Последовательно, при  $j = 1, \dots, l$  строится «наиболее перемешанная» перестановка чисел  $1, \dots, 2^j$ . При  $j = 1$  она состоит из двух чисел 1, 2. Пусть уже построена перестановка  $(b_1^{j-1}, \dots, b_{2^{j-1}}^{j-1})$ ; следующая перестановка берется в виде  $(b_1^{j-1}, 2^j + 1 - b_1^{j-1}, b_2^{j-1}, 2^j + 1 - b_2^{j-1}, \dots)$ . Например, при  $l = 4$  эта перестановка имеет вид (1, 16, 8, 9, 4, 13, 5, 12, 2, 15, 7, 10, 3, 14, 6, 11). Задаваясь  $k = 2^l$ , строим таблицу чисел  $b_1^l, \dots, b_{2^l}^l$  и производим итерации (12) при значениях

$$\alpha_i^k = 2 / \left[ M + \mu + (M - \mu) \cos \frac{\pi (2b_i^l - 1)}{2k} \right].$$

Зададимся целью построить итерационный процесс, в котором после каждых  $k$  операций умножения матрицы  $A$  на вектор получается приближение

$$y^k = x^0 - Q_{k-1}^0(A) (Ax^0 - b). \quad (14)$$

Согласно (5) имеем

$$Q_{k-1}^0(A) = \left( E - T_k \left( \frac{2A - (M + \mu)E}{M - \mu} \right) t_k^{-1} \right) A^{-1}.$$

Заметим, что  $Q_{-1}^0(A) = (E - E)A^{-1} = 0$  и  $y^0 = x^0$ . Подставив выражение  $Q_{k-1}^0(A)$  в (14) и умножив затем обе части на  $t_k$ , получим

$$t_k y^k = T_k \left( \frac{2A - (M + \mu)E}{M - \mu} \right) (x^0 - A^{-1}b) + t_k A^{-1}b.$$

Из справедливости этого соотношения при  $k = n - 1$ ,  $k = n$ ,  $k = n + 1$  следует равенство

$$\begin{aligned}
 & t_{n+1}y^{n+1} - 2 \frac{2A - (M + \mu)E}{M - \mu} t_n y^n + t_{n-1}y^{n-1} = \\
 & = \left\{ T_{n+1} \left( \frac{2A - (M + \mu)E}{M - \mu} \right) - 2 \frac{2A - (M + \mu)E}{M - \mu} T_n \left( \frac{2A - (M + \mu)E}{M - \mu} \right) + \right. \\
 & \quad \left. + T_{n-1} \left( \frac{2A - (M + \mu)E}{M - \mu} \right) \right\} (x^0 - A^{-1}b) + \\
 & \quad + \left( t_{n+1}E - 2 \frac{2A - (M + \mu)E}{M - \mu} t_n - t_{n-1}E \right) A^{-1}b. \quad (15)
 \end{aligned}$$

Многочлены Чебышева связаны рекуррентным соотношением

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0. \quad (16)$$

Если подставить в (16) вместо  $x$  некоторую матрицу, в частности матрицу  $(M - \mu)^{-1}[2A - (M + \mu)E]$ , то это соотношение также будет выполняться. Поэтому в (15) матрица в фигурных скобках нулевая. Вследствие (16) и равенства  $t_n = T_n \left( -\frac{M + \mu}{M - \mu} \right)$  имеем

$$t_{n+1} - 2 \left( -\frac{M + \mu}{M - \mu} \right) t_n + t_{n-1} = 0 \quad (17)$$

и соотношение (15) может быть преобразовано к виду

$$t_{n+1}y^{n+1} - 2 \left( -\frac{M + \mu}{M - \mu} \right) t_n y^n + t_{n-1}y^{n-1} = 4 \frac{t_n}{M - \mu} (Ay^n - b). \quad (18)$$

Мы получили требуемое рекуррентное соотношение. Приведем его к более удобному виду. Вследствие (17) равенство (18) можно переписать в виде

$$\begin{aligned}
 & t_{n+1}y^{n+1} - (t_{n+1} + t_{n-1})y^n + t_{n-1}y^{n-1} = \\
 & = \frac{2}{M + \mu} (t_{n+1} + t_{n-1})(Ay^n - b)
 \end{aligned}$$

или

$$y^{n+1} = y^n + \omega_n \omega_{n-1} (y^n - y^{n-1}) - \frac{2}{M + \mu} (1 + \omega_n \omega_{n-1})(Ay^n - b), \quad (19)$$

где  $\omega_n = -(t_n/t_{n+1})$ . Разделив обе части (17) на  $t_{n+1}$ , получим

$$1 - 2 \frac{M + \mu}{M - \mu} \omega_n + \omega_n \omega_{n-1} = 0,$$

откуда

$$\omega_n = 1 / \left( 2 \frac{M + \mu}{M - \mu} - \omega_{n-1} \right) \text{ при } n > 0, \quad \omega_0 = -t_0/t_1. \quad (20)$$

Таким образом, можно рекуррентно вычислять величины  $\omega_n$  из (20) и затем векторы  $y^{n+1}$  из (19). Для получения совокуп-

ности векторов  $y^1, \dots, y^n$  потребуется произвести  $n$  умножений матрицы на вектор,  $O(n)$  умножений векторов на числа, сложений векторов и операций с числами. При этом всех  $k$ , вследствие (3)–(5), выполняется оценка

$$\|y^k - X\|_2 \leq |t_k|^{-1} \|x^0 - X\|_2. \quad (21)$$

Квадратное уравнение

$$\omega^2 - 2 \frac{M + \mu}{M - \mu} \omega + 1 = 0$$

имеет два положительных корня

$$\omega = \frac{M + \mu}{M - \mu} \pm \sqrt{\left(\frac{M + \mu}{M - \mu}\right)^2 - 1}.$$

Наименьший из этих корней, равный

$$\frac{M + \mu}{M - \mu} - \sqrt{\left(\frac{M - \mu}{M + \mu}\right)^2 - 1} = \frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}} = \frac{1}{\lambda_0},$$

обозначим через  $\omega$ .

Итерационный процесс (19), (20) называют *оптимальным* (по количеству итераций) *линейным итерационным процессом*. При реализации процесса (19), (20) после каждых  $k$  применений матрицы  $A$  мы получаем оптимальный результат в смысле (3)–(4). Из сказанного видно, что по скорости сходимости итерационный процесс (19), (20) предпочтительнее, чем (9). Однако иногда от него отказываются в пользу (9) из-за соображений экономии памяти ЭВМ.

Получим более наглядную оценку скорости сходимости построенных итерационных процессов. Согласно (21) и (8) для оптимального итерационного процесса выполняется оценка

$$\|y^n - X\|_2 \leq 2\lambda_0^{-n} \|x^0 - X\|_2.$$

Норма погрешности уменьшится по крайней мере в  $\epsilon^{-1}$  раз, если  $2\lambda_0^{-n} \leq \epsilon$ . Отсюда получаем оценку числа итераций, обеспечивающих получение решения с точностью  $\epsilon$ :

$$n \geq n_1 = \log_{\lambda_0} (2/\epsilon) = (\ln \lambda_0)^{-1} \ln (2/\epsilon).$$

Для многих задач число  $M/\mu$  оказывается очень большим. При  $M/\mu \rightarrow \infty$  имеем  $\lambda_0 = 1 + 2\sqrt{\mu/M} + O(\mu/M)$ . Таким образом,

$$\ln \lambda_0 \sim 2\sqrt{\mu/M}, \quad n_1 \sim 0,5\sqrt{M/\mu} \ln (2/\epsilon).$$

Для сравнения рассмотрим оптимальный линейный одношаговый итерационный процесс, имеющий, согласно (9), оценку погрешности

$$\|x^n - X\|_2 \leq \left(\frac{M - \mu}{M + \mu}\right)^n \|x^0 - X\|_2,$$

отсюда получаем оценку числа итераций

$$n \geq n_2 = \left( \ln \frac{M + \mu}{M - \mu} \right)^{-1} \ln \frac{1}{\varepsilon}.$$

При  $M/\mu \rightarrow \infty$  имеем

$$\ln \frac{M + \mu}{M - \mu} \sim \frac{2\mu}{M} \text{ и } n_2 \sim \frac{1}{2} \frac{M}{\mu} \ln \frac{1}{\varepsilon}.$$

Таким образом, в этом сравнении оптимальный итерационный процесс дает выигрыш в числе итераций примерно в  $\sqrt{M/\mu}$  раз.

**Задача 2.** Рассматривается итерационный процесс

$$x^i = x^{i-1} - \alpha_{i-1} (Ax^{i-1} - b), \quad i = 1, 2, \dots$$

Пусть  $p$  нечетное и при всех  $k$  совокупности  $\alpha_0, \dots, \alpha_{p^k-1}$  совпадают с совокупностями

$$\gamma_{i-1} = 2 \left( M + \mu + (M - \mu) \cos \frac{\pi(2i-1)}{2p^k} \right)^{-1}, \quad i = 1, \dots, p^k.$$

Проверить, что при всех  $i = p^k$  приближения  $x^i$  совпадают с приближениями, получаемыми по оптимальному линейному итерационному процессу.

**Задача 3.** Пусть  $\alpha_0 = 2/(M + \mu)$ ,  $\alpha_1 = 1/M$ ,  $\alpha_2 = 1/\mu$  и совокупности  $\alpha_{2^k+1}, \dots, \alpha_{2^{k+1}}$  при каждом  $k > 1$  совпадают с совокупностями величин

$$\gamma_i = 2 \left( M + \mu + (M - \mu) \cos \frac{\pi(2i-1)}{2^{k+1}} \right)^{-1}, \quad i = 1, \dots, 2^k.$$

Показать, что для такого итерационного процесса при всех  $n = 2^{k+1}$  справедлива оценка

$$\|x^n - X\|_2 \leq \frac{4}{(\lambda_0^{n-1} - \lambda_0^{1-n})(\lambda_0 - \lambda_0^{-1})} \|x^0 - X\|_2,$$

т. е.

$$\|x^n - X\|_2 = O(\lambda_0^{-n}) \|x^0 - X\|_2.$$

Рассмотрим типичную задачу математической физики, сводящуюся к решению системы уравнений с большим отношением  $M/\mu$ . Пусть в квадрате  $0 \leq x_1, x_2 \leq 1$  решается уравнение Пуассона  $-\Delta u = f$  при нулевых условиях на границе. Зададимся сеткой с шагами  $h = 1/l$  и напишем систему уравнений, аппроксимирующих дифференциальную задачу:

$$-\frac{u_{m+1,n} - 2u_{mn} + u_{m-1,n}}{h^2} - \frac{u_{m,n+1} - 2u_{mn} + u_{m,n-1}}{h^2} = f_{mn} \quad (22)$$

при  $0 < m, n < l$ ;

$$u_{mn} = 0, \text{ если } m(l-m)(l-n)n = 0.$$

Матрица этой системы является положительно определенной, и для нее

$$\mu = \min \lambda_i = 8l^2 \sin^2 \left( \frac{\pi h}{2} \right) \sim 2\pi^2,$$

$$M = \max \lambda_i = 8l^2 \cos^2 \left( \frac{\pi h}{2} \right) \sim 8l^2,$$

т. е.  $\sqrt{M/\mu} \sim 2l/\pi$ . Например, при шаге  $l = 30$  выигрыш в числе итераций примерно в 20 раз,

В начале этого параграфа было упомянуто, что при симметризации системы ее свойства могут ухудшаться. Действительно, пусть этот процесс применяется к уже симметричной матрице  $A$ , т. е. переходим от системы  $Ax = b$  к системе  $A^2x = Ab$ . Если у старой системы отношение максимального и минимального собственных значений равнялось  $M/\mu$ , то у новой оно будет  $(M/\mu)^2$  и скорость сходимости итерационных процессов будет меньше.

Примечание. При  $n \rightarrow \infty$  имеем

$$\omega_n = \frac{T_n \left( \frac{M + \mu}{M - \mu} \right)}{T_{n+1} \left( \frac{M + \mu}{M - \mu} \right)} \rightarrow \omega = \frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}}.$$

Таким образом, при больших  $n$  итерационная формула (19) близка к формуле

$$z^{n+1} = z^n + \omega^2 (z^n - z^{n-1}) - \frac{2}{M + \mu} (1 + \omega^2) (Az^n - b). \quad (23)$$

Если при  $n > 1$  итерации будут производиться по этой формуле, то при условии  $z^0 = x^0$ ,  $z^1 = y^1$  этот итерационный процесс требует примерно столько же итераций, сколько итерационный процесс (19).

**Задача 4.** Для итерационного процесса (23) получить оценку погрешности

$$\frac{\|r^{n+1}\|_2}{\|r^0\|_2} \leq \frac{2 + (n-1)(1 - \lambda_0^{-2})}{1 + \lambda_0^{-2}} \lambda_0^{-n}. \quad (24)$$

**Указание.** Представить погрешность в виде  $r^n = \sum_{k=1}^m z_k^n e_k$ , где  $\{e_k\}$  — полная ортогональная система собственных векторов матрицы  $A$ . Подстановкой в (23) получить разностное уравнение, связывающее  $z_k^{n-1}$ ,  $z_k^n$ ,  $z_k^{n+1}$ . Получить явное выражение для  $z_k^n$  и с его помощью получить требуемую оценку (24).

**Задача 5.** Показать, что оценка (24) не может быть улучшена на множестве матриц, удовлетворяющих условию  $S(A) \subset [\mu, M]$ .

**Задача 6.** Пусть все собственные значения  $\lambda_i$  матрицы  $A$  вещественны,  $\lambda_i \in [\mu, M]$ , числа  $\mu, M$  являются собственными значениями матрицы  $A$  кратности не выше первой. Доказать, что для оптимального линейного итерационного процесса

$$\|r^n\|_k = O \left( \left( \frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}} \right)^n \right) \|r^0\|_k$$

при  $n \rightarrow \infty$ ,  $k = 1, 2, \infty$ .

## § 7. Метод Зейделя

Пусть решается система уравнений  $Ax = b$ , все диагональные элементы которой ненулевые. В итерационном методе Зейделя последовательно уточняются компоненты решения, причем  $k$ -я компонента находится из  $k$ -го уравнения. Именно, если  $x^n = (x_1^n, \dots, x_m^n)^T$ , то следующее приближение определяет-



ся из системы соотношений

$$\begin{aligned} a_{11}x_1^{n+1} + a_{12}x_2^n + \dots + a_{1m}x_m^n &= b_1, \\ a_{21}x_1^{n+1} + a_{22}x_2^{n+1} + a_{23}x_3^n + \dots + a_{2m}x_m^n &= b_2, \\ \dots &\dots \\ a_{m1}x_1^{n+1} + a_{m2}x_2^{n+1} + \dots + a_{mn}x_m^{n+1} &= b_m. \end{aligned} \quad (1)$$

Систему (1) можно представить в виде

$$Bx^{n+1} + Cx^n = b, \quad (2)$$

где

$$B = \begin{bmatrix} a_{11} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}, \quad C = \begin{bmatrix} 0 & a_{12} & a_{13} & \dots & a_{1m} \\ 0 & 0 & a_{23} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}.$$

Отсюда получаем

$$x^{n+1} = -B^{-1}Cx^n + B^{-1}b. \quad (3)$$

Таким образом, метод Зейделя эквивалентен некоторому методу простой итерации; поэтому для его сходимости при любом начальном приближении необходимо и достаточно, чтобы все собственные значения матрицы  $-B^{-1}C$  по модулю были меньше 1. Вследствие равенства

$$\det(-B^{-1}C - \lambda E) = \det(-B^{-1}) \det(C + B\lambda)$$

собственные значения матрицы  $(-B^{-1}C)$  являются корнями уравнения  $\det(C + B\lambda) = 0$ .

Таким образом, необходимое и достаточное условие сходимости метода Зейделя можно формулировать следующим образом: все корни уравнения

$$\det \begin{bmatrix} a_{11}\lambda & a_{12} & a_{13} & \dots & a_{1m} \\ a_{21}\lambda & a_{22}\lambda & a_{23} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1}\lambda & a_{m2}\lambda & a_{m3}\lambda & \dots & a_{mm}\lambda \end{bmatrix} = 0$$

по модулю меньше 1.

В частных случаях можно предложить более удобные для применения достаточные условия сходимости метода Зейделя.

**Задача 1.** Пусть при всех  $i$

$$\sum_{j \neq i} |a_{ij}| \leq q |a_{ii}|, \quad q < 1.$$

Получить оценку

$$\|x^n - X\|_\infty \leq \dots \leq q^n \|x^0 - X\|_\infty.$$

Проблема решения систем линейных уравнений является модельной относительно более сложных задач решения систем нелинейных уравнений и минимизации функций многих переменных. Для перенесения метода на более сложные задачи важно

понять его наиболее «грубые» качественные свойства, обеспечивающие сходимость. С этой целью наиболее желательно получить геометрическую интерпретацию метода. Обозначим через  $L_i$  плоскость  $\sum_{j=1}^m a_{ij}x_j - b_i = 0$ . При получении приближения  $(x_1^{n+1}, \dots, x_i^{n+1}, x_{i+1}^n, \dots, x_m^n)$  из приближения  $(x_1^n, \dots, x_{i-1}^n, x_i^n, \dots, x_m^n)$  происходит перемещение приближения параллельно оси  $x_i$  до

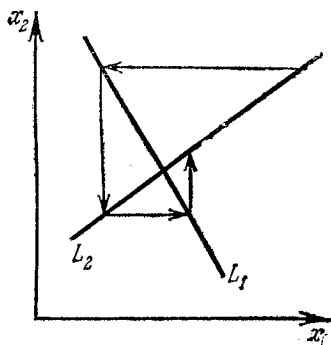


Рис. 6.7.1

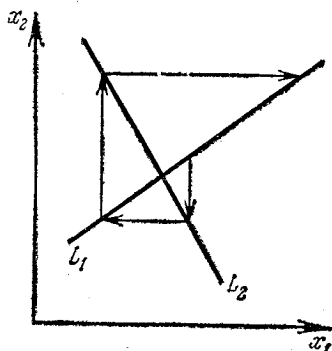


Рис. 6.7.2

пересечения с плоскостью  $L_i$ . Таким образом, геометрически метод Зейделя состоит в циклическом перемещении точек, соответствующих последовательно получаемым приближениям, параллельно координатным осям  $x_i$  до пересечения с плоскостями  $L_i$ . Рис. 6.7.1—6.7.3 иллюстрируют при  $m=2$  случаи, когда метод Зейделя сходится, расходится, имеет цикл (как говорят, «зацикливается»). Сравнение первых двух рисунков показывает, что сходимость метода Зейделя может измениться характер при перестановке уравнений.

Особенно интересная геометрическая картина возникает в случае, когда матрица  $A$  симметричная.

**Теорема.** Пусть  $A$  — вещественная симметричная положительно определенная матрица. Тогда метод Зейделя сходится.

**Доказательство.** При  $A$  симметричной имеем

$$\begin{aligned} F(y) &= (A(y - X), y - X) - (AX, X) = \\ &= (Ay, y) - 2(AX, y) = (Ay, y) - 2(b, y). \end{aligned}$$

Если  $A > 0$ , то  $(A(y - X), y - X) > 0$  при  $y \neq X$ , поэтому функция  $F(y)$  имеет минимум, и притом единственный, при  $y = X$ . Таким образом, задача отыскания решения системы  $Ax = b$  ока-

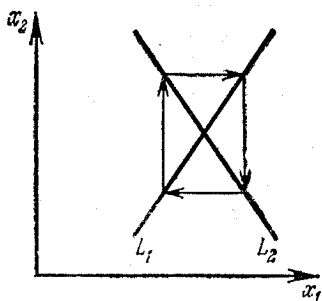


Рис. 6.7.3

залась равносильной задаче отыскания единственного минимума функции  $F(y)$ .

Одним из методов минимизации функций многих переменных является *метод покоординатного спуска*.

Пусть имеется приближение  $(x_1^0, \dots, x_m^0)$  к точке экстремума функции  $F(x_1, \dots, x_m)$ . Рассмотрим функцию  $F(x_1, x_2^0, \dots, x_m^0)$  как функцию переменной  $x_1$  и найдем точку  $x_1^1$  ее минимума. Затем, исходя из приближения  $(x_1^1, x_2^0, \dots, x_m^0)$ , путем минимизации функции  $F(x_1^1, x_2, x_3^0, \dots, x_m^0)$  по переменной  $x_2$  находим следующее приближение  $(x_1^1, x_2^1, x_3^0, \dots, x_m^0)$ . Процесс циклически повторяется.

При уточнении компоненты  $x_k$  происходит смещение по прямой, параллельной оси  $x_k$ , до точки с наименьшим на этой прямой значением  $F(x) = c$ . Ясно, что эта точка будет точкой касания рассматриваемой прямой и линии уровня  $F(x) = c$ . Поэтому в двумерном случае картина приближений выглядит как на рис. 6.7.4.

Применим метод покоординатного спуска для отыскания экстремума функции  $F(y)$ . Обозначим  $F(y) + (AX, X) = (A(y - X), y - X)$  через  $F_0(y)$ . При ми-

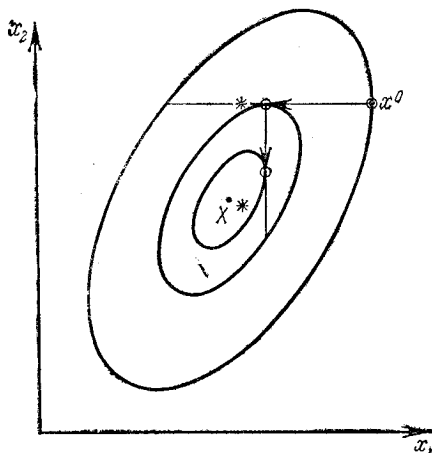


Рис. 6.7.4

нимизации по переменной  $x_k$  происходит перемещение параллельно оси  $x_k$  до точки, где  $F'_{x_k} = 0$ . Следовательно, новое значение  $x_k$  определяется из того же уравнения

$$F'_{x_k} = 2 \left( \sum_{j=1}^m a_{kj} x_j - b_k \right) = 0,$$

что и в случае метода Зейделя. Таким образом, приближения покоординатного спуска минимизации функции  $F(y)$  и метода Зейделя решения исходной системы совпадают.

Если  $x^n \neq X$ , то хотя бы одно из уравнений системы не удовлетворяется и соответствующее значение  $F'_{x_k}(x^n) \neq 0$ . Выберем среди таких  $k$  наименьшее. Тогда при уточнении компонент  $x_1, \dots, x_{k-1}$  мы остаемся в точке  $x^n$ , а при уточнении компоненты  $x_k$  происходит смещение в сторону меньших значений  $F(x)$ ; при уточнении остальных компонент значение  $F(x)$  не возрастает. Таким образом,

$$F(x^{n+1}) < F(x^n), \quad F_0(x^{n+1}) < F_0(x^n);$$

поэтому

$$F_0(\mathbf{x}^{n+1})/F_0(\mathbf{x}^n) < 1 \quad (6)$$

при  $\mathbf{x}^n \neq \mathbf{X}$ . Вследствие (4) имеем равенство  $\mathbf{r}^{n+1} = -B^{-1}C\mathbf{r}^n$ . Соотношение (6) можно переписать в виде

$$\varphi(\mathbf{r}^n) = \frac{(AB^{-1}C\mathbf{r}^n, B^{-1}C\mathbf{r}^n)}{(A\mathbf{r}^n, \mathbf{r}^n)} < 1 \quad (7)$$

при  $\mathbf{r}^n \neq 0$ . На сфере  $\|\mathbf{r}^n\|_2 = 1$  величина  $\varphi(\mathbf{r}^n)$  непрерывна, поэтому она достигает своего наибольшего значения  $\varphi_0$ . Так как  $A > 0$ , то всегда  $\varphi(\mathbf{r}^n) > 0$  и поэтому  $\varphi_0 > 0$ . Положим  $\sqrt{\varphi_0} = \lambda$ . Вследствие (7) имеем  $\lambda^2 < 1$ . Очевидно,  $\varphi(c\mathbf{r}^n) = \varphi(\mathbf{r}^n)$  при любом  $c \neq 0$ , поэтому  $\varphi(\mathbf{r}^n) = \varphi(\mathbf{r}^n/\|\mathbf{r}^n\|_2) \leq \lambda^2$  при любом  $\mathbf{r}^n$ . Отсюда получаем неравенство

$$F_0(\mathbf{x}^{n+1})/F_0(\mathbf{x}^n) \leq \lambda^2, \quad (8)$$

и, таким образом,

$$F_0(\mathbf{x}^n) \leq \lambda^{2n} F_0(\mathbf{x}^0).$$

Из (3.8), (3.9) следуют неравенства

$$(\min \lambda_A^i) \|\mathbf{y} - \mathbf{X}\|_2^2 \leq F_0(\mathbf{y}) \leq (\max \lambda_A^i) \|\mathbf{y} - \mathbf{X}\|_2^2.$$

Отсюда получаем оценку скорости сходимости

$$\begin{aligned} \|\mathbf{x}^n - \mathbf{X}\|_2 &\leq \sqrt{\frac{F_0(\mathbf{x}^n)}{\min \lambda_A^i}} \leq \lambda^n \sqrt{\frac{F_0(\mathbf{x}^0)}{\min \lambda_A^i}} \leq \\ &\leq \lambda^n \sqrt{\frac{\max \lambda_A^i}{\min \lambda_A^i}} \|\mathbf{x}^0 - \mathbf{X}\|_2. \quad (9) \end{aligned}$$

Теорема доказана.

Из рис. 6.7.5 можно усмотреть, что метод Зейделя сходится быстрее, если направление осей эллипсоидов близко к направлению координатных осей, т. е. матрица  $A$  близка к диагональной.

В случае, изображенном на рис. 6.7.4, последовательные приближения смещаются все время монотонно влево и вниз. Такая картина — монотонное смещение отдельных компонент все время в одном и том же направлении — характерна для ряда классов матриц. При этом довольно часто монотонное смещение наблюдается именно у компонент решения, скорость сходимости которых наиболее плохая. В этих случаях для ускорения сходимости прибегают к *методу релаксации*, который заключается в следующем. После уточнения каждой координаты по методу Зейделя производится смещение в том же направлении на  $p$ -ю

часть этого смещения. Таким образом, приближения отыскиваются из соотношения

$$(B - D)x^{n+1} + D\left(\frac{x^{n+1} + px^n}{1+p}\right) + Cx^n = b, \quad (10)$$

$D$  — диагональная матрица с элементами  $a_{ii}$  по диагонали. Как показала практика вычислений, при  $A > 0$  целесообразно брать показатель релаксации  $p$  в пределах  $0 < p < 1$ ; в случае

$0 < p < 1$  метод релаксации обычно называют *методом сверхрелаксации* (или *методом верхней релаксации*). На рис. 6.7.4 изображены символами  $\circ$ ) приближения метода Зейделя, \* — метода сверхрелаксации при  $p = 1/4$ . Например, для уменьшения погрешности в  $\epsilon^{-1}$  раз при решении системы (6.22) методом Зейделя требуется порядка  $ch^{-2} \ln(\epsilon^{-1})$  итераций. Если применить метод сверхрелаксации при  $p = 1 - \omega h$ ,  $\omega > 0$ , то потребуется порядка  $c(\omega)h^{-1} \ln(\epsilon^{-1})$  итераций. Подбор параметра  $\omega$  в этой задаче требует детального рассмотрения. В частности, во многих случаях оптимальное значение параметра релаксации  $p$  определяется экспериментально. Иногда параметр релаксации  $p$  выбирают зависящим от  $n$  и  $i$ .

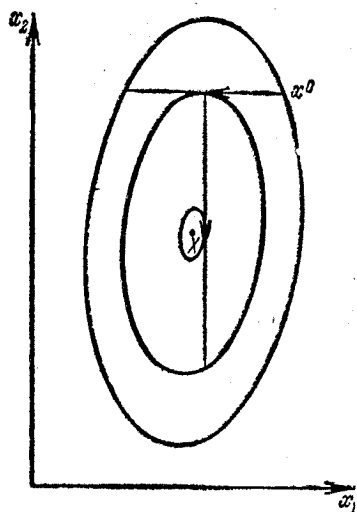


Рис. 6.7.5

Для случая  $A > 0$  еще раз обратимся к геометрической картине (см. рис. 6.7.4). После уточнения компоненты  $x_1$  по методу релаксации при  $-1 < p < 1$  мы попадаем в точку, лежащую внутри или на границе эллипсоида  $F(x) = F(x^0)$ . Рассуждая так же, как при обосновании сходимости метода Зейделя, заключаем, что при  $-1 < p < 1$  всегда  $F_0(x^{n+1}) < F_0(x^n)$ , если  $x^n \neq X$ .

**Задача 2.** Доказать, что при условии  $|p_n| \leq q < 1$  метод релаксации сходится со скоростью геометрической прогрессии.

## § 8. Метод наискорейшего градиентного спуска

Распространенным методом минимизации функций большого числа переменных является *метод градиентного спуска*. Последующее приближение получается из предыдущего смещением в направлении, противоположном градиенту функции  $F(x)$ . Каждое следующее приближение ищется в виде

$$x^{n+1} = x^n - \delta_n \text{grad } F(x^n). \quad (1)$$

Приведенное описание не определяет алгоритм однозначно, поскольку ничего не сказано о выборе параметра  $\delta_n$ . Например,

его можно определять из условия минимума величины

$$F(\mathbf{x}^n - \delta_n \text{grad } F(\mathbf{x}^n)). \quad (2)$$

В этом случае рассматриваемый метод называют *методом наискорейшего градиентного спуска* или просто *методом наискорейшего спуска*.

Для функции  $F(\mathbf{x}) = (\mathbf{A}\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$ , соответствующей системе линейных уравнений с матрицей  $\mathbf{A} = \mathbf{A}^T > 0$ , задача нахождения минимума решается в явном виде. В этом конкретном случае

$$\text{grad } F = 2(\mathbf{A}\mathbf{x} - \mathbf{b})$$

и

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \delta_n 2(\mathbf{A}\mathbf{x}^n - \mathbf{b}).$$

Обозначим  $2\delta_n$  через  $\Delta_n$ , т. е. положим

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Delta_n(\mathbf{A}\mathbf{x}^n - \mathbf{b}). \quad (3)$$

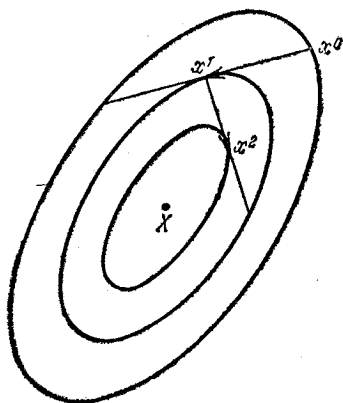


Рис. 6.8.1

Пусть  $\varphi(\Delta_n) = F(\mathbf{x}^{n+1})$ . Вспомогая, что  $\mathbf{A} = \mathbf{A}^T$ , вычислим  $\varphi'(\Delta_n)$ . Имеем  $\varphi(\Delta_n) = F(\mathbf{x}^n) - 2\Delta_n(\mathbf{A}\mathbf{x}^n - \mathbf{b}, \mathbf{A}\mathbf{x}^n - \mathbf{b}) + (\mathbf{A}(\mathbf{A}\mathbf{x}^n - \mathbf{b}), \mathbf{A}\mathbf{x}^n - \mathbf{b})\Delta_n^2$ ; уравнение  $\varphi'(\Delta_n) = 0$  имеет вид  $2((\mathbf{A}(\mathbf{A}\mathbf{x}^n - \mathbf{b}), \mathbf{A}\mathbf{x}^n - \mathbf{b})\Delta_n - (\mathbf{A}\mathbf{x}^n - \mathbf{b}, \mathbf{A}\mathbf{x}^n - \mathbf{b})) = 0$ , откуда

$$\Delta_n = \frac{(\mathbf{A}\mathbf{x}^n - \mathbf{b}, \mathbf{A}\mathbf{x}^n - \mathbf{b})}{(\mathbf{A}(\mathbf{A}\mathbf{x}^n - \mathbf{b}), \mathbf{A}\mathbf{x}^n - \mathbf{b})}. \quad (4)$$

На рис. 6.8.1 изображены последовательные приближения метода наискорейшего спуска и линии уровня функции  $F$ . Итерационный процесс (3), (4) называют *методом наискорейшего спуска* решения рассматриваемой системы линейных уравнений.

Пусть собственные значения матрицы  $\mathbf{A}$  расположены на  $[\mu, M]$ , т. е.  $S_A \subset [\mu, M]$ .

**Теорема.** *Приближения метода наискорейшего спуска удовлетворяют соотношению*

$$F_0(\mathbf{x}^n) \leq \left(\frac{M-\mu}{M+\mu}\right)^{2n} F_0(\mathbf{x}^0), \quad F_0(\mathbf{x}) = (\mathbf{A}(\mathbf{x} - \mathbf{X}), \mathbf{x} - \mathbf{X}). \quad (5)$$

**Доказательство.** При  $\mathbf{y}^n = \mathbf{x}^n$  произведем одну итерацию оптимального одношагового итерационного процесса

$$\mathbf{y}^{n+1} = \mathbf{y}^n - \frac{2}{M+\mu} (\mathbf{A}\mathbf{y}^n - \mathbf{b}). \quad (6)$$

Погрешности итераций  $\mathbf{r}^n = \mathbf{y}^n - \mathbf{X}$  связаны соотношением

$$\mathbf{r}^{n+1} = \left(E - \frac{2}{M+\mu} \mathbf{A}\right) \mathbf{r}^n.$$

Пусть  $e_1, \dots$  — ортонормированная система собственных векторов матрицы  $A$ :  $Ae_i = \lambda_i e_i$ ,  $(e_i, e_j) = \delta_{ij}$ . Поскольку  $\mu \leq \lambda_i \leq M$ , то при всех  $i$  выполняются соотношения

$$-\frac{M-\mu}{M+\mu} \leq 1 - \frac{2}{M+\mu} \lambda_i \leq \frac{M-\mu}{M+\mu}$$

и, таким образом,

$$\left| 1 - \frac{2}{M+\mu} \lambda_i \right| \leq \frac{M-\mu}{M+\mu}. \quad (7)$$

Пусть  $r^n = \sum c_i e_i$ . Справедливы соотношения

$$(Ar^n, r^n) = (\sum c_i \lambda_i e_i, \sum c_i e_i) = \sum \lambda_i c_i^2,$$

$$r^{n+1} = \sum b_i e_i, \text{ где } b_i = \left(1 - \frac{2\lambda_i}{M+\mu}\right) c_i,$$

$$(Ar^{n+1}, r^{n+1}) \leq \sum \lambda_i b_i^2 = \sum \lambda_i \left(1 - \frac{2\lambda_i}{M+\mu}\right)^2 c_i^2.$$

С учетом (7) получаем

$$(Ar^{n+1}, r^{n+1}) \leq \left(\frac{M-\mu}{M+\mu}\right)^2 (Ar^n, r^n).$$

Поскольку  $F_0(y^n) = (Ar^n, r^n)$ , то это означает, что

$$F_0(y^{n+1}) \leq \left(\frac{M-\mu}{M+\mu}\right)^2 F_0(y^n) = \left(\frac{M-\mu}{M+\mu}\right)^2 F_0(x^n).$$

Приближение  $y^{n+1}$  можно записать в виде (1)

$$y^{n+1} = x^n - \alpha \operatorname{grad} F(x^n), \quad \alpha = (M+\mu)^{-1}.$$

Так как на  $x^{n+1}$  достигается минимум  $F(x)$  среди всех приближений вида (1), то  $F(x^{n+1}) \leq F(y^{n+1})$ . Отсюда следует оценка

$$F_0(x^{n+1}) \leq F_0(y^{n+1}) \leq \left(\frac{M-\mu}{M+\mu}\right)^2 F_0(x^n),$$

а поэтому и справедливость утверждения теоремы. Аналогично (7.9) можно получить неравенство

$$\|x^n - X\|_2 \leq \left(\frac{M-\mu}{M+\mu}\right)^n \sqrt{\frac{M}{\mu}} \|x^0 - X\|_2.$$

Хотя на каждом шаге метода наискорейшего спуска уменьшение величины  $F_0(x)$  заведомо не меньше, чем у итерационного процесса (6), мы получили примерно одинаковые оценки скорости сходимости. Однако есть принципиальная разница в этих методах. Для написания итерационного процесса (6) требуется информация о границах спектра  $\mu, M$ . В случае метода (3), (4) такой информации не требуется.

Однако у метода наискорейшего спуска (3), (4) есть следующий недостаток по сравнению с простейшим итерационным процессом (6). При нахо-

ждении каждого следующего приближения он требует не одной, а двух трудоемких операций умножения матрицы на вектор.

Двукратного умножения матрицы на вектор при каждой итерации можно избежать следующим образом. Обозначим  $w^n \equiv Ax^n - b$  и перепишем (3) в виде

$$x^{n+1} = x^n - \Delta_n w^n. \quad (8)$$

Умножая (8) слева на  $A$  и вычитая  $b$ , получим

$$w^{n+1} = w^n - \Delta_n A w^n. \quad (9)$$

Формулу (4) для определения  $\Delta_n$  можно записать в виде

$$\Delta_n = \frac{(w^n, w^n)}{(A w^n, w^n)}. \quad (10)$$

В процессе итерации запоминаются векторы  $x^n$ ,  $w^n$  и на каждом шаге последовательно вычисляются  $A w^n$ ,  $\Delta_n$ ,  $x^{n+1}$ ,  $w^{n+1}$ . В исходном методе наискорейшего спуска (3), (4) погрешность на шаге итерации равносильна возмущению начального приближения и, поскольку процесс сходящийся, ее влияние должно иметь тенденцию к затуханию.

В итерационном процессе (8) — (10) накопление вычислительной погрешности носит более сложный характер.

**Задача 1.** Получить оценку скорости сходимости метода наискорейшего спуска

$$\|x^n - X\|_2 \leq (1 - \mu/M)^n \|x^0 - X\|_2.$$

Реальный выбор итерационного процесса должен производиться с учетом имеющейся информации о границе спектра, объема и состава памяти ЭВМ. Например, при решении сеточных уравнений, аппроксимирующих дифференциальные уравнения в частных производных, иногда идут по следующему пути. Рассматривая задачу на более крупной сетке, проводят вспомогательную работу по возможно более точному определению значений  $\mu$  и  $M$ , соответствующих более мелкой сетке, а затем применяют оптимальный линейный итерационный процесс.

Обратим внимание на интересное обстоятельство. Из геометрической картины итераций метода Зейделя видно, что скорость сходимости метода не меняется при умножении уравнений системы на множители и изменении масштабов по координатным осям, равносильном замене  $x_i = k_i y_i$ .

Иначе обстоит дело в случае метода наискорейшего спуска. Пусть, например,  $A = E$  — единичная матрица. Тогда

$$F(x) = (Ax, x) - 2(b, x) = \sum_{i=1}^m x_i^2 - 2 \sum_{i=1}^m b_i x_i$$

и метод наискорейшего спуска сходится за одну итерацию (доказать!). Произведем замену масштабов  $x_i = k_i y_i$ ,  $k_i > 0$ . Матрица системы  $A$  в данном случае будет диагональной с элементами на диагонали, равными  $k_i$ . Тогда минимизируется функционал

$$\bar{F}(y) = (Ay, y) - 2(b, y) = \sum_{i=1}^m k_i y_i^2 - 2 \sum_{i=1}^m b_i y_i.$$



При большом разбросе  $k_i$  линиями уровня функции  $F$  будут сильно вытянутые эллипсоиды и скорость сходимости метода наискорейшего спуска будет очень медленной.

### § 9. Метод сопряженных градиентов

Этот метод также предназначен для решения систем уравнений

$$Ax = b$$

с положительно определенной матрицей  $A$ .

Пусть  $x^0$  — начальное приближение к решению. Образует вектор

$$x^n = x^0 + \sum_{i=0}^{n-1} \alpha_i A^i r^0, \quad r^0 = Ax^0 - b. \quad (1)$$

Поставим задачу минимизации величины

$$F(x^n) = (Ax^n, x^n) - 2(b, x^n)$$

за счет выбора параметров  $\alpha_i$ . Имеем равенство

$$\frac{dF(x^n)}{d\alpha_j} = 2 \left( Ax^n - b, \frac{dx^n}{d\alpha_j} \right).$$

Так как  $dx^n/d\alpha_j = A^j r^0$  и

$$Ax^n - b = r^0 + \sum_{i=0}^{n-1} \alpha_i A^{i+1} r^0, \quad (2)$$

то условия  $F'_{\alpha_j} = 0$  образуют систему уравнений

$$(Ax^n - b, A^j r^0) = 0, \quad j = 0, \dots, n-1. \quad (3)$$

Подставляя сюда выражение  $Ax^n - b$ , получим систему линейных уравнений

$$(r^0, A^j r^0) + \sum_{i=0}^{n-1} \alpha_i (A^{i+1} r^0, A^j r^0) = 0, \quad j = 0, \dots, n-1, \quad (4)$$

относительно параметров  $\alpha_0, \dots, \alpha_{n-1}$ .

При заданном начальном приближении  $x^0$  можно выбрать какое-то  $n$ , при котором мы ожидаем достигнуть требуемой точности, далее решить систему уравнений (4) относительно неизвестных  $\alpha_i$  и получить приближение  $x^n$ .

Иногда поступают иначе. Как и в § 6, задаются последовательностью значений  $n_0 = 0 < n_1 < n_2 < \dots$ . За  $x^{n_i}$ ,  $i = 1, 2, \dots$ , последовательно принимаем  $x^n$ , получаемое из (4) при  $x^0 = x^{n_i-1}$ ,  $n = n_i - n_{i-1}$ .

Как правило, более целесообразно по аналогии с оптимальным линейным итерационным процессом пойти по пути после-

довательного построения векторов

$$x^n = x^0 + \sum_{i=0}^{n-1} \alpha_i^n A^i (Ax^0 - b),$$

минимизирующих  $F(x)$  при каждом  $n$ . Оказывается, что можно построить рекуррентное соотношение, связывающее эти векторы, аналогичное рекуррентному соотношению оптимального линейного итерационного процесса.

Свойства этого итерационного процесса существенно зависят от минимального числа  $q$  такого, что вектор  $r^0$  можно представить в виде

$$r^0 = \sum_{j=1}^q h_j, \quad h_j \neq 0, \quad Ah_j = \lambda_j h_j;$$

$\lambda_1, \dots, \lambda_q$  — различные собственные значения матрицы  $A$ ,  $h_1, \dots, h_q$  — ее собственные вектора.

Можно показать, что рекуррентные соотношения, связывающие векторы  $x^n$ , минимизирующие  $F(x)$  при каждом  $n$ , имеют вид

$$x^{n+1} = x^n + \alpha_n (x^n - x^{n-1}) + \beta_n (Ax^n - b), \quad (5)$$

где

$$\beta_n = -1/(\sigma_n + \sigma_{n-1}),$$

$$\alpha_n = \beta_n \sigma_{n-1},$$

а  $\sigma_k^n$  определяются по формуле

$$\sigma_k^n = \frac{(A(Ax^n - b), Ax^k - b)}{(Ax^k - b, Ax^k - b)}.$$

При этом в предположении отсутствия округлений  $\sigma_n^n + \sigma_{n-1}^n \neq 0$  при  $n < q$ , а  $x^q$  совпадает с точным решением задачи. Вектор  $x^1$  определяется непосредственно с помощью решения системы (4) и совпадает с первым приближением в методе наискорейшего спуска.

Вместо того чтобы вычислять  $\alpha_n$  и  $\beta_n$  по этим формулам, можно каждый раз определять их, минимизируя по  $\alpha_n$  и  $\beta_n$  функционал  $F(x^{n+1})$  на множестве  $x^{n+1}$  вида (5). Для вычисления  $\alpha_n$  и  $\beta_n$  получатся расчетные формулы

$$\begin{aligned} \alpha_n &= ((Ax^n - b, Ax^n - b)(Ax^n - b, A(x^n - x^{n-1})) - \\ &\quad - (Ax^n - b, x^n - x^{n-1})(A(Ax^n - b), Ax^n - b)) \sigma_n, \\ \beta_n &= ((Ax^n - b, x^n - x^{n-1})(A(x^n - x^{n-1}), Ax^n - b) - \\ &\quad - (Ax^n - b, Ax^n - b)(x^n - x^{n-1}, A(x^n - x^{n-1}))) \sigma_n, \end{aligned} \quad (6)$$

где

$$\sigma_n = [(x^n - x^{n-1}, A(x^n - x^{n-1}))(Ax^n - b, A(Ax^n - b)) - (Ax^n - b, A(x^n - x^{n-1}))^2]^{-1}.$$

При отсутствии округлений выполняется равенство  $(Ax^n - b, x^n - x^{n-1}) = 0$ .

Заметим, что при  $x^{n-1}$  и  $x^n$ , близких к точному решению  $X$ , многие из величин, входящих в выражения  $\alpha_n, \beta_n, \sigma_n$ , малы, поэтому следует обратить внимание на выработку критерия останова итерационного процесса.

При умеренном числе итераций часто целесообразнее использовать иные последовательности рекуррентных соотношений. Например, полагают  $s_1 = r_0 = Ax_0 - b$  и последовательно вычисляют

$$r_n = r_{n-1} - a_n A s_n, \quad x_n = x_{n-1} - a_n s_n, \\ s_{n+1} = r_n + b_n s_n, \quad n = 1, 2, \dots;$$

здесь

$$a_n = (r_{n-1}, r_{n-1}) / (A s_n, s_n),$$

$$b_n = (r_n, r_n) / (r_{n-1}, r_{n-1}).$$

В этом алгоритме объем вычислений на каждом шаге в 2—3 раза меньше по сравнению с исходными формулами (5).

Построенный итерационный процесс имеет много различных названий, в частности *метод Ланцоша* или *метод сопряженных градиентов*.

Произведем оценку скорости сходимости. Для приближений оптимального линейного итерационного процесса  $y^n$  имеем: если

$$x^0 - X = y^0 - X = \sum_i c_i e_i,$$

то

$$y^n - X = \sum_i c_i g_i e_i,$$

где

$$|g_i| \leq |t_n|^{-1} = \left| T_n \left( -\frac{M + \mu}{M - \mu} \right) \right|^{-1}.$$

Отсюда следует оценка

$$F_0(y^n) = (A(y^n - X), y^n - X) = \\ = \sum_i \lambda_i c_i^2 g_i^2 \leq |t_n|^{-2} \sum_i \lambda_i c_i^2 = |t_n|^{-2} (Ax^0 - X, x^0 - X).$$

Приближение метода сопряженных градиентов минимизирует функционалы  $F$  и  $F_0$  среди всех приближений вида (1), которым принадлежит также  $y^n$ . Поэтому

$$F_0(x^n) \leq F_0(y^n) \leq |t_n|^{-2} F_0(x^0).$$

Как и ранее, получаем оценку скорости сходимости в обычной норме:

$$\|x^n - X\|_2 \leq \sqrt{\frac{M}{\mu}} |t_n|^{-1} \|x^0 - X\|_2.$$

*Рассматриваемый процесс имеет ту же по порядку оценку скорости сходимости, что и оптимальный линейный итерационный процесс. Его существенным преимуществом оказывается то, что не требуется знания границ спектра.*

Пусть матрица  $A$  имеет  $l$  ненулевых элементов; естественно,  $l \geq m$ . Для проведения одного шага итерации требуется  $O(l)$  арифметических операций, поэтому при отсутствии округлений для получения точного решения потребуется  $qO(l) = O(ml)$  арифметических операций. Таким образом, в случае матриц с малым числом ненулевых элементов этот метод оказывается также и эффективным точным методом. При этом он предъявляет относительно малые требования к памяти ЭВМ.

Если размерность системы  $m$  велика, часто хорошее приближение к решению получается при числе итераций  $n \ll q$ .

Рассмотрим случай, когда не выполнено условие  $A > 0$ . Система  $A^*Ax = A^*b$  с положительно определенной матрицей  $A^*A$  эквивалентна исходной, но может оказаться, что число ненулевых элементов матрицы  $A^*A$  много больше  $l$ . Тогда будет разумнее не выписывать эту систему в явном виде, а каждый раз вычислять векторы вида  $A^*Au$  последовательным умножением  $u$  на матрицы  $A$  и  $A^*$ .

Метод сопряженных градиентов, как мы видели, обладает более быстрой скоростью сходимости, чем методы простой итерации и наискорейшего спуска. Однако он требует несколько большего объема памяти ЭВМ, так как необходимо хранить в памяти большее число векторов размерности  $m$ . Особенно это заметно в тех случаях, когда матрица  $A$  не хранится в памяти, а для вычисления элементов матрицы  $A$  существует простая формула. В частности, такие системы возникают при решении краевых задач для дифференциальных уравнений.

## § 10. Итерационные методы с использованием спектрально-эквивалентных операторов

Кроме методов простой итерации вида

$$x^{n+1} = x^n - \alpha(Ax^n - b) \quad (1)$$

часто применяются итерационные методы

$$Bx^{n+1} = Bx^n - \alpha(Ax^n - b), \quad (2)$$

где матрица  $B \neq E$  такова, что система уравнений  $Bu = c$  легко может быть решена. Если соотношение (2) умножить на  $B^{-1}$ ,

то получится

$$x^{n+1} = x^n - \alpha B^{-1}(Ax^n - b).$$

Таким образом, итерационный процесс (2) равносильен методу простой итерации с матрицей  $E - \alpha B^{-1}A$ .

Рассмотрим случай, когда  $A > 0$  и отношение максимального  $M$  и минимального  $\mu$  из собственных значений матрицы  $A$  велико. Тогда рассматривавшиеся ранее итерационные процессы сходятся медленно. Пусть  $B \geq 0$  и

$$M_1 = \sup_x \frac{(Ax, x)}{(Bx, x)}, \quad \mu_1 = \inf_x \frac{(Ax, x)}{(Bx, x)}. \quad (3)$$

Предположим, что система уравнений с матрицей  $B$  легко решается и  $M_1/\mu_1 \ll M/\mu$ .

В случае, когда отношение  $M_1/\mu_1$  не очень велико, итерационные методы типа рассматриваемых в настоящем параграфе принято условно называть *итерационными методами с использованием спектрально-эквивалентных операторов*. Покажем, что при удачном подборе матрицы  $B$  метод итераций (2) обладает лучшей сходимостью по сравнению с простейшим методом (1).

Точное решение  $X$  удовлетворяет равенству

$$BX = BX - \alpha(Ax - b);$$

вычитая его из (2), получим уравнение относительно погрешности

$$Br^{n+1} = Br^n - \alpha Ar^n. \quad (4)$$

Приведем матрицу  $B$  при помощи ортогонального преобразования к диагональному виду. Пусть  $B = U^T \Lambda U$ , где

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix},$$

$U$  — ортогональная матрица.

Через  $\sqrt{B}$  принято обозначать матрицу вида

$$U^T \begin{bmatrix} +\sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & +\sqrt{\lambda_m} \end{bmatrix} U.$$

Очевидно,  $\sqrt{B} > 0$ ,  $(\sqrt{B})^* = \sqrt{B}$ ,  $\sqrt{B} \sqrt{B} = B$ . Умножим обе части уравнения (4) слева на  $(\sqrt{B})^{-1}$  и положим  $\sqrt{B}r^n = v^n$ . Получим равенство

$$v^{n+1} = v^n - \alpha C v^n, \quad (5)$$

где  $C = (\sqrt{B})^{-1} A (\sqrt{B})^{-1}$ .

Вследствие соотношений  $A = A^T$ ,  $(\sqrt{B}) = (\sqrt{B})^T$  матрица  $C$  симметрична. Рассмотрим выражение

$$\omega(x) = (Cx, x)/(x, x).$$

Положив  $(\sqrt{B})^{-1}x = y$ , можем написать равенство

$$\omega(x) = \omega(\sqrt{B}y) = (Ay, y)/(By, y).$$

Согласно (3) имеем  $\omega(x) \in [\mu_1, M_1]$ , поэтому все собственные значения матрицы  $C$  также принадлежат отрезку  $[\mu_1, M_1]$ .

При  $\alpha = 2/(M_1 + \mu_1)$  собственные значения матрицы  $E - \alpha C$  по модулю не превосходят величины  $(M_1 - \mu_1)/(M_1 + \mu_1)$ . Поэтому  $\|E - \alpha C\|_2 \leq (M_1 - \mu_1)/(M_1 + \mu_1)$  и аналогично (6.7) имеем

$$\|v^n\|_2 \leq \left(\frac{M_1 - \mu_1}{M_1 + \mu_1}\right)^n \|v^0\|_2. \quad (6)$$

Напомним, что

$$\|v^n\|_2 = \sqrt{(\sqrt{B}r^n, \sqrt{B}r^n)} = \sqrt{(Br^n, r^n)}.$$

Если  $0 < \mu_2 \leq (By, y)/(y, y) \leq M_2$  при всех  $y$ , то из (6) следует оценка

$$\|r^n\|_2 \leq \frac{\|v^n\|_2}{\sqrt{\mu_2}} \leq \left(\frac{M_1 - \mu_1}{M_1 + \mu_1}\right)^n \frac{\|v^0\|_2}{\sqrt{\mu_2}} \leq \left(\frac{M_1 - \mu_1}{M_1 + \mu_1}\right)^n \sqrt{\frac{M_2}{\mu_2}} \|r^0\|_2.$$

Так как функция

$$\frac{M - \mu}{M + \mu} = \frac{1 - (\mu/M)}{1 + (\mu/M)}$$

монотонно убывает с уменьшением  $M/\mu$ , то скорость сходимости нового итерационного процесса быстрее чем у (1).

При окончательном решении вопроса о переходе от итераций по формулам (1) к итерациям по формулам (2) следует учесть количество арифметических операций, выполняемых во время этих итераций. Если итерации по формулам (2) требуют существенно большего количества операций, то такой переход может оказаться нецелесообразным.

**Задача 1.** Пусть  $T(A)$  — количество арифметических операций, выполняемых во время одной итерации по формуле (1),  $T(B)$  — количество арифметических операций, выполняемых во время одной итерации по формуле (2). Привести соображения в пользу того, что переход к итерациям по формуле (2) целесообразен, если

$$T(B) \left(\ln \frac{M_1 + \mu_1}{M_1 - \mu_1}\right)^{-1} < T(A) \left(\ln \frac{M + \mu}{M - \mu}\right)^{-1}.$$

По аналогии с тем, как по итерационному методу (1) был построен метод (2), можно построить аналог любого из рассматривавшихся в § 7, 9, 10 методов.

Пусть имеется итерационный процесс, где погрешности последующих приближений связаны равенством

$$r^{n+1} = \left( E - \sum_{i=1}^k \alpha_i A^i \right) r^n.$$

Напишем равенство

$$v^{n+1} = \left( E - \sum_{i=1}^k \alpha_i C^i \right) v^n, \quad (7)$$

где  $C = (\sqrt{B})^{-1} A (\sqrt{B})^{-1}$ , положим  $v^n = \sqrt{B} r^n$  и умножим (7) слева на  $\sqrt{B}$ . Получим уравнение

$$B r^{n+1} = B r^n - \sum_{i=1}^k \alpha_i (AB^{-1})^{i-1} A r^n.$$

Такое уравнение для погрешности соответствует итерационному процессу

$$B x^{n+1} = B x^n - \sum_{i=1}^k \alpha_i (AB^{-1})^{i-1} (A x^n - b). \quad (8)$$

Если произвести оптимизацию параметров  $\alpha_i$  так, чтобы оценка отношения норм  $\|v^{n+1}\|_2 / \|v^n\|_2$  была наилучшей, то так же, как в § 6, получится наилучший итерационный метод вида (8).

Аналогом оптимального линейного итерационного метода (6.19), (6.20) будет итерационный процесс

$$B y^{n+1} = B y^n + \bar{\omega}_n \bar{\omega}_{n-1} B (y^n - y^{n-1}) - \frac{2}{M_1 + \mu_1} (1 + \bar{\omega}_n \bar{\omega}_{n-1}) (A y^n - b),$$

где  $\bar{\omega}_n$  строятся по  $M_1$  и  $\mu_1$  так же, как  $\omega_n$  по  $M$  и  $\mu$  (см. (6.20)).

Объединением метода (2) с методом наискорейшего спуска является следующий метод. Приближение  $x^{n+1}$  ищется из соотношения

$$B x^{n+1} = B x^n - \alpha_n (A x^n - b), \quad (9)$$

при этом коэффициент  $\alpha_n$  определяется из условия минимума

$$F(x^{n+1}) = (A x^{n+1}, x^{n+1}) - 2(b, x^{n+1}).$$

Из (9) получаем

$$B \frac{d x^{n+1}}{d \alpha_n} = - (A x^n - b).$$

Поскольку

$$\begin{aligned} \frac{d F(x^{n+1})}{d \alpha_n} &= 2 \left( A x^{n+1} - b, \frac{d x^{n+1}}{d \alpha_n} \right) = \\ &= 2 (A (x^n - \alpha_n B^{-1} (A x^n - b)) - b, - B^{-1} (A x^n - b)), \end{aligned}$$

то условие  $d F(x^{n+1}) / d \alpha_n = 0$  является линейным уравнением относительно  $\alpha_n$ .

Аналог итерационного процесса (9.5) имеет вид

$$B x^{n+1} = B x^n + \alpha_n (B x^n - B x^{n-1}) + \beta_n (A x^n - b).$$

Методы, рассмотренные в этом параграфе, получили широкое применение при решении систем уравнений  $Ax = b$  с большим разбросом собственных значений матрицы  $A > 0$ .

**§ 11. Погрешность приближенного решения системы уравнений и обусловленность матриц. Регуляризация**

Предположим, что матрица и правая часть системы заданы неточно и вместо предъявленной к решению системы

$$Ax = b \quad (1)$$

в действительности должна была решаться некоторая система

$$A_1x = b_1, \quad A_1 = A + \Delta, \quad b_1 = b + \eta. \quad (2)$$

Пусть известны оценки  $\|\Delta\|$  и  $\|\eta\|$ . Займемся оценкой погрешности решения.

Сначала выделим главный член погрешности. Будем обозначать решения (1) и (2) через  $X$  и  $X^*$  и разность  $X^* - X$  — через  $r$ . Подставив выражения  $A_1$ ,  $b_1$  и  $X^*$  в (2), будем иметь

$$(A + \Delta)(X + r) = b + \eta.$$

Вычитая из этого равенства (1), получим

$$Ar + \Delta X + \Delta r = \eta,$$

откуда

$$Ar = \eta - \Delta X - \Delta r$$

и

$$r = A^{-1}(\eta - \Delta X - \Delta r). \quad (3)$$

Если  $\|\Delta\|$  и  $\|\eta\|$  малы, то следует ожидать и малости  $\|r\|$ . Тогда слагаемое  $\Delta r$  имеет более высокий порядок малости; отбрасывая это слагаемое, получаем

$$r \approx A^{-1}(\eta - \Delta X).$$

Отсюда следует оценка погрешности

$$\|r\| \leq \sigma \approx \|A^{-1}\|(\|\eta\| + \|\Delta\| \|X\|). \quad (4)$$

Строгая оценка погрешности получается следующим образом. Вследствие (3) выполняется неравенство

$$\|r\| \leq \|A^{-1}\| \|\eta\| + \|A^{-1}\| \|\Delta\| \|X\| + \|A^{-1}\| \|\Delta\| \|r\|.$$

Предположим, что  $\|A^{-1}\| \|\Delta\| < 1$ . Перенеся последнее слагаемое в левую часть и поделив неравенство на коэффициент при  $\|r\|$ , получим оценку

$$\|r\| \leq \frac{\|A^{-1}\|(\|\eta\| + \|\Delta\| \|X\|)}{1 - \|A^{-1}\| \|\Delta\|}. \quad (5)$$

Довольно распространен случай, когда погрешность матрицы системы существенно меньше погрешности правой части. В качестве модели этой ситуации будем рассматривать случай точ-



ного задания матрицы системы. Тогда, полагая в (5)  $\Delta = 0$ , имеем

$$\|r\| \leq \|A^{-1}\| \|\eta\|.$$

Для качественной характеристики связи между погрешностями правой части и решения вводятся понятия *обусловленности системы* и *обусловленности матрицы системы*. Абсолютные погрешности правой части и решения системы зависят от масштабов, которыми измеряются коэффициенты системы. Поэтому удобнее характеризовать свойства системы через связь между относительными погрешностями правой части и решения.

Соответственно этому в качестве *меры обусловленности системы* принимается число

$$\tau = \sup_{\eta} \left( \frac{\|r\|}{\|X\|} : \frac{\|\eta\|}{\|b\|} \right) = \frac{\|b\|}{\|X\|} \sup_{\eta} \frac{\|r\|}{\|\eta\|}.$$

Согласно этому определению, имеем оценку относительной погрешности решения через меру обусловленности и относительную погрешность правой части:

$$\frac{\|r\|}{\|X\|} \leq \tau \frac{\|\eta\|}{\|b\|}. \quad (6)$$

Так как  $r = A^{-1}\eta$ , то

$$\sup_{\eta} \frac{\|r\|}{\|\eta\|} = \|A^{-1}\|$$

и

$$\tau = \frac{\|b\|}{\|X\|} \|A^{-1}\|.$$

Иногда удобнее иметь более грубую характеристику свойств системы только через свойства матрицы  $A$ . Эту характеристику  $\nu(A) = \sup_b \tau$  называют *мерой* (или *числом*) *обусловленности матрицы  $A$* . Согласно этому определению и (6), имеем оценку

$$\frac{\|r\|}{\|X\|} \leq \nu(A) \frac{\|\eta\|}{\|b\|},$$

связывающую относительные погрешности правой части и решения только через свойства матрицы системы. Так как

$$\sup_b \frac{\|b\|}{\|X\|} = \sup_x \frac{\|Ax\|}{\|x\|} = \|A\|,$$

то

$$\nu(A) = \|A\| \|A^{-1}\|.$$

Поскольку любая норма матрицы не меньше ее наибольшего по модулю собственного значения, то  $\|A\| \geq \max |\lambda_A|$ ; поскольку собственные значения матриц  $A$  и  $A^{-1}$  взаимно обратны, то

$$\|A^{-1}\| \geq \max \frac{1}{|\lambda_A|} = \frac{1}{\min |\lambda_A|}.$$

Таким образом,

$$\nu(A) \geq \max |\lambda_A| / \min |\lambda_A| \geq 1.$$

В частности, при  $A = A^T$  имеем  $\|A\|_2 = \max |\lambda_A|$  и

$$\|A^{-1}\|_2 = \max \frac{1}{|\lambda_A|} = \frac{1}{\min |\lambda_A|}.$$

Следовательно, в случае нормы  $\|\cdot\|_2$

$$\nu(A) = \max |\lambda_A| / \min |\lambda_A|.$$

Рассмотрим вопрос о погрешности решения вследствие округления в ЭВМ правой части. Пусть, как обычно,  $t$  — двоичная разрядность чисел в ЭВМ. Каждый элемент  $b_i$  правой части округляется с относительной погрешностью  $O(2^{-t})$ , т. е. с абсолютной погрешностью, равной  $O(|b_i|2^{-t})$ , поэтому

$$\|\eta\| = O(\|b\|2^{-t}) \quad \text{и} \quad \|\eta\|/\|b\| = O(2^{-t}).$$

Следовательно,

$$\|r\|/\|X\| \leq \nu(A) O(2^{-t}).$$

При практической работе вопрос о строгой оценке погрешности полученного приближенного решения системы линейных уравнений с помощью полученных неравенств или каким-либо иным способом возникает редко. Однако информация о порядке погрешности решения часто полезна для получения качественных выводов о том, с какой точностью разумно решать задачу. Соотношения (4), (5) оценивают сверху погрешность решения, являющуюся следствием погрешности исходных данных. Из равенства (3) видно, что оценки (4), (5) довольно точны, поэтому обычно не имеет смысла стремиться получать решение задачи с погрешностью, существенно меньшей чем  $\sigma$ .

Системы уравнений и матрицы с большими значениями мер обусловленности принято называть *плохо обусловленными*, а с малыми — *хорошо обусловленными*. Если правая часть (4), оценивающая погрешность решения через погрешность исходных данных, или оценка вычислительной погрешности недопустимо велики, то полезно принять во внимание какую-то дополнительную информацию о решении рассматриваемой задачи. Подход к решению такой задачи должен быть таким же, как в случае *некорректных задач*.

Рассмотрим простейший случай, когда  $A$  — симметричная матрица. Пусть  $\lambda_1, \dots, \lambda_m \neq 0$  — ее собственные значения, упорядоченные в порядке убывания  $|\lambda_i|$ ; соответствующую ортонормированную систему собственных векторов обозначим через  $e_1, \dots, e_m$ . Решением системы  $Ax = b$  является вектор

$$X = \sum_{i=1}^m c_i e_i, \quad c_i = \frac{(b, e_i)}{\lambda_i}.$$

При реально заданной правой части  $\tilde{b} = b + \eta$  решением будет

$$\tilde{X} = \sum_{i=1}^m \frac{(\tilde{b}, e_i) + (\eta, e_i)}{\lambda_i} e_i.$$

Коэффициент  $(\eta, e_i)/\lambda_i$  может оказаться очень большим при малых  $\lambda_i$ , что ведет к сильному искажению решения. Иногда заранее известно, что в разложении искомого решения задачи  $\sum_{i=1}^m c_i e_i$  коэффициенты  $c_i$ , соответствующие малым по модулю  $\lambda_i$ , малы. В этом случае следует принять какие-то меры с тем, чтобы «отфильтровать» эти составляющие решения.

При небольших  $m$  для решения этой задачи иногда применяется следующий способ: задаются некоторым  $q > 0$ , находят все  $\lambda_i$  и  $e_i$  при  $i \leq q$  и полагают

$$X \approx \sum_{i=1}^q \frac{(\tilde{b}, e_i)}{\lambda_i} e_i;$$

$q$  следует подобрать исходя из дополнительной информации о задаче.

Другие два способа проиллюстрируем на примере матриц, где все  $\lambda_i > 0$ .

Первый способ. Задавая некоторое  $\alpha \geq 0$ , находим решение  $x^\alpha$  системы

$$(\alpha E + A)x^\alpha = \tilde{b}.$$

Оно записывается в виде

$$x_\alpha = \sum_{i=1}^m \frac{(\tilde{b}, e_i)}{\lambda_i + \alpha} e_i.$$

Так как

$$\frac{1}{\lambda_i} - \frac{1}{\lambda_i + \alpha} = \frac{\alpha}{\lambda_i(\lambda_i + \alpha)},$$

то наличие малого параметра  $\alpha$  несущественно изменит слагаемые, соответствующие большим  $\lambda_i$ . В то же время при  $\lambda_i \ll \alpha$  имеем

$$\left| \frac{(\tilde{b}, e_i)}{\lambda_i + \alpha} \right| \ll \left| \frac{(\tilde{b}, e_i)}{\lambda_i} \right|.$$

Это означает, что введение параметра  $\alpha$  приводит к существенному уменьшению роли слагаемых, соответствующих малым  $\lambda_i$ . Подбор оптимального значения  $\alpha$  обычно осуществляют экспериментально, сравнивая результаты расчетов при различных  $\alpha$ .

Второй способ заключается в следующем. Будем решать систему уравнений каким-либо итерационным способом.

Рассмотрим случай итераций по формуле

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \alpha(\mathbf{A}\mathbf{x}^n - \mathbf{b}) \quad (7)$$

при некотором начальном приближении  $\mathbf{x}^0$ . Пусть

$$\tilde{\mathbf{b}} = \sum_{i=1}^m \beta_i \mathbf{e}_i, \quad \mathbf{x}^n = \sum_{i=1}^m z_i^n \mathbf{e}_i.$$

Подставим эти выражения в (7) и, приравнявая коэффициенты при  $\mathbf{e}_i$ , получим соотношения

$$z_i^{n+1} = \alpha\beta_i + (1 - \alpha\lambda_i) z_i^n.$$

Последовательно выражая каждое  $z_i^k$  через предыдущие, имеем

$$z_i^n = \alpha\beta_i + (1 - \alpha\lambda_i) z_i^{n-1} =$$

$$= \alpha\beta_i + (1 - \alpha\lambda_i)(\alpha\beta_i + (1 - \alpha\lambda_i) z_i^{n-2}) = \dots$$

$$\dots = \alpha\beta_i \sum_{k=0}^{n-1} (1 - \alpha\lambda_i)^k + (1 - \alpha\lambda_i)^n z_i^0.$$

Если  $|1 - \alpha\lambda_i| < 1$ , то при  $n \rightarrow \infty$

$$\sum_{k=0}^{n-1} (1 - \alpha\lambda_i)^k \rightarrow \sum_{k=0}^{\infty} (1 - \alpha\lambda_i)^k = \frac{1}{\alpha\lambda_i}; \quad (1 - \alpha\lambda_i)^n \rightarrow 0,$$

поэтому  $z_i^n \rightarrow \beta_i/\lambda_i$ . Пусть  $\alpha \approx (\max \lambda_i)^{-1}$ , т. е. относительно мало. При больших значениях  $\lambda_i$  величина  $(1 - \alpha\lambda_i)^n$  быстро стремится к нулю с ростом  $n$  и  $z_i^n$  близко к своему предельному значению  $\beta_i/\lambda_i$ . В то же время иногда удается подобрать начальное приближение, для которого величины  $z_i^0$  относительно малы при малых  $\lambda_i$ . Тогда при небольшом  $n$  коэффициенты  $z_i^n$ , соответствующие таким  $\lambda_i$ , еще не будут недопустимо большими и получаемое приближение может оказаться приемлемым.

В других случаях решение задачи находят, минимизируя некоторый функционал, близкий к  $F(\mathbf{x}) = (\mathbf{A}\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$ , например функционал  $F(\mathbf{x}) + \alpha(\mathbf{x}, \mathbf{x})$  с малым  $\alpha > 0$ .

Успешность применения описанных приемов в случае несимметричных матриц  $\mathbf{A}$  в существенной степени зависит от структуры жордановой формы и от ряда свойств матрицы. Здесь часто решение находят, минимизируя функционал

$$(\mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{A}\mathbf{x} - \mathbf{b}) + \alpha(\mathbf{x}, \mathbf{x})$$

при малых  $\alpha > 0$ ; значение  $\alpha$  опять-таки подбирается экспериментально из сравнения результатов расчетов при различных  $\alpha$ .

Другая группа методов основана на представлении матрицы системы  $\mathbf{A}$  в виде

$$\mathbf{A} = \mathbf{G}\mathbf{\Lambda}\mathbf{P},$$

где  $G$  и  $P$  — ортогональные матрицы, а  $\Lambda$  — двухдиагональная матрица, у которой могут быть отличными от нуля элементы  $\lambda_{ij}$  при  $j=i$  и  $j=i+1$ .

Большинство из описанных методов решения систем уравнений с плохо обусловленной матрицей относится к методам регуляризации.

Задача 1. Пусть  $A^{(m)} = [a_{ij}^m]$  — матрица размерности  $m \times m$  с элементами

$$a_{ij}^m = \begin{cases} p & \text{при } j=i, \\ q & \text{при } j=i+1, \\ 0 & \text{при } j \neq i, i+1. \end{cases}$$

1. Вычислить матрицу  $(A^{(m)})^{-1}$  и доказать утверждение: при  $|q| < |p|$  матрицы  $A^{(m)}$  хорошо обусловлены, а при  $|q| > |p|$  и  $m$  больших плохо обусловлены.

2. Выписать явно решение системы  $A^{(m)}x = b$  через правую часть.

3. Выписать явно через правую часть  $b$  вектор  $x_{\alpha}$ , минимизирующий функционал

$$(A^{(m)}x - b, A^{(m)}x - b) + \alpha(x, x).$$

4. Попытаться качественно описать эффект, достигаемый за счет применения такой регуляризации.

Объясним еще одну причину, по которой стараются избегать симметризации матриц, предложенной в § 6. Сначала посмотрим, что происходит, когда операция симметризации применяется формально в случае положительно определенной матрицы. Тогда  $A = A^T$  и  $A^T A = A^2$ . При возведении матрицы в квадрат собственные значения возводятся в квадрат, поэтому при  $\|A\| = \|A\|_2$  имеем

$$\nu(A^2) = \frac{\max |\lambda_{A^2}|}{\min |\lambda_{A^2}|} = \frac{(\max |\lambda_A|)^2}{(\min |\lambda_A|)^2} = (\nu(A))^2.$$

Если  $\nu(A) = 1$ , то  $\max |\lambda_A| = \min |\lambda_A|$  и все собственные значения матрицы  $A$  равны между собой по модулю, т. е.  $A = \text{const } E$ . За исключением этого частного случая  $\nu(A) > 1$ , поэтому  $\nu(A^2) > \nu(A)$ . Таким образом, в случае симметричной матрицы применение симметризации увеличивает число обусловленности. Число обусловленности является непрерывной функцией от матрицы, поэтому для матриц, близких к симметричным, применение симметризации также увеличивает число обусловленности матриц. Таким образом, имеется непустой класс несимметричных матриц, по отношению к которым алгоритм симметризации увеличивает число обусловленности.

Задача 2. Существуют ли несимметричные матрицы, для которых  $\nu(A^2) = (\nu(A))^2$ ?

## § 12 Проблема собственных значений

В различных случаях возникают разные требования к информации о собственных значениях и собственных векторах матриц, и это порождает многообразие проблем и приемов решения этой задачи.

1. Для решения ряда задач механики, физики, химии требуется получение всех собственных значений, а иногда и всех собственных векторов некоторых матриц. Эту задачу называют *полной проблемой собственных значений*.

2. В ряде случаев требуется найти лишь максимальное или минимальное по модулю собственное значение матрицы. Проблемы подобного сорта возникают, в частности, при решении некоторых задач ядерной физики. Здесь приходится решать задачи, эквивалентные задачам отыскания собственных значений матриц размерности порядка  $10^3 \div 10^6$  или даже существенно большей. При малых размерностях матриц для решения этих задач чаще применяют итерационные методы, при больших — вероятностные.

3. При исследовании колебательных процессов иногда требуется найти два максимальных по модулю собственных значения матрицы, причем меньшее из них обычно достаточно определить с меньшей точностью.

4. Там же возникает задача отыскания собственного значения матрицы, наиболее близкого к заданному числу  $\lambda^0$ , или отыскания расстояния от данного числа  $\lambda^0$  до спектра матрицы.

Формально рассуждая, можно было бы сказать, что эти задачи, называемые *частичными проблемами собственных значений*, являются частным случаем общей проблемы собственных значений и достаточно ограничиться рассмотрением методов решения общей проблемы. Однако такой подход приведет к неоправданно большому объему вычислений. При обсуждении постановок конкретных задач, связанных с отысканием собственных значений матриц, зачастую значительные усилия тратятся именно на установление минимального объема информации о спектре матрицы, которым можно ограничиться.

Решение задач 2—4 обычно сводят к отысканию максимального по модулю собственного значения некоторой матрицы  $B = g(A)$  такой, что это собственное значение соответствует отыскиваемому собственному значению матрицы  $A$ .

Рассмотрим случай, когда все собственные числа матрицы  $A$  вещественны. Если требуется найти максимальное или минимальное собственное значение матрицы  $A$ , то следует взять  $g(A) = A + cE$ . Очевидно, что при достаточно больших положительных (отрицательных) значениях  $c$  максимальному по модулю собственному значению матрицы  $A + cE$  соответствует максимальное (минимальное) собственное значение матрицы  $A$ . В случае задачи 4 при некотором  $c$  максимальное по модулю

собственное значение матрицы  $E - c(A - \lambda^0 E)^2$  соответствует отыскиваемому собственному значению матрицы  $A$ . Иногда в качестве такой матрицы  $g(A)$  может использоваться матрица  $(A - \lambda^0 E)^{-1}$ . При этом матрица  $(A - \lambda^0 E)^{-1}$  не выписывается в явном виде, а необходимые по ходу вычислений векторы  $(A - \lambda^0 E)^{-1}y$  находят, решая систему уравнений  $(A - \lambda^0 E)x = y$ .

Рассмотрим типичную задачу отыскания двух максимальных по модулю собственных значений матрицы  $A$ . Для простоты предполагаем наличие полной системы собственных векторов  $e_i$ :

$$Ae_i = \lambda_i e_i, \quad (e_i, e_i) = \delta_i^j; \quad |\lambda_1| > |\lambda_2| > |\lambda_3| \geq \dots \geq |\lambda_m|.$$

Зададимся некоторым вектором  $x^0$  и будем последовательно вычислять векторы  $x^{n+1} = Ax^n$ . Представим  $x^0$  в виде  $x^0 = \sum_{i=1}^m c_i e_i$ ; имеем

$$x^n = \sum_{i=1}^m c_i A^n e_i = \sum_{i=1}^m c_i \lambda_i^n e_i.$$

Отсюда следуют соотношения

$$x^n = c_1 \lambda_1^n e_1 + O(|\lambda_2|^n),$$

$$\begin{aligned} (x^n, x^n) &= (c_1 \lambda_1^n e_1 + O(|\lambda_2|^n), c_1 \lambda_1^n e_1 + O(|\lambda_2|^n)) = \\ &= |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n), \end{aligned}$$

$$\begin{aligned} (x^{n+1}, x^n) &= (c_1 \lambda_1^{n+1} e_1 + O(|\lambda_2|^{n+1}), c_1 \lambda_1^n e_1 + O(|\lambda_2|^n)) = \\ &= \lambda_1 |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n). \end{aligned}$$

Положим

$$\lambda_1^{(n)} = (x^{n+1}, x^n) / (x^n, x^n).$$

Из последних соотношений при  $c_1 \neq 0$  получаем

$$\begin{aligned} \lambda_1^{(n)} &= \frac{\lambda_1 |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n)}{|c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n)} = \\ &= \frac{\lambda_1 \left( 1 + O\left(\frac{1}{|c_1|^2} \left| \frac{\lambda_2}{\lambda_1} \right|^n\right) \right)}{1 + O\left(\frac{1}{|c_1|^2} \left| \frac{\lambda_2}{\lambda_1} \right|^n\right)} = \lambda_1 + O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^n\right). \end{aligned} \quad (1)$$

**Задача 1.** Доказать, что в случае симметричной матрицы  $\lambda_1^{(n)} = \lambda_1 + O(|\lambda_2/\lambda_1|^{2n})$ .

Кроме (1) имеем

$$\begin{aligned} \|x^n\| &= |c_1| |\lambda_1|^n + O(|\lambda_2|^n), \\ e_1^{(n)} &= \frac{x^n}{\|x^n\|} = \frac{\sum_{i=1}^m c_i \lambda_i^n e_i}{|c_1 \lambda_1^n| + O(|\lambda_2|^n)} = \\ &= \frac{\frac{c_1 \lambda_1^n}{|c_1 \lambda_1^n|} e_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)}{1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)} = e^{i\varphi_n} e_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right); \end{aligned}$$

здесь  $\varphi_n = \arg(c_1 \lambda_1^n)$ . Таким образом, в ходе этого итерационного процесса также получаем собственный вектор, соответствующий  $\lambda_1$ .

Может случиться, что у матрицы  $A$  имеется два максимальных по модулю собственных значения  $\lambda_1 \neq \lambda_2$ ,  $|\lambda_1| = |\lambda_2| > |\lambda_3| \geq \dots$ . В этом случае величина  $\lambda_1^{(n)}$  будет устанавливаться только в частном случае, когда  $c_1$  или  $c_2$  равно нулю. Если заранее известно, что таких собственных значений два, то их и соответствующие им собственные вектора можно также определить, анализируя поведение  $\lambda_1^{(n)}$  и  $e_1^{(n)}$ .

Рассмотрим типичный случай, когда  $A$ ,  $\lambda_1$ ,  $\lambda_2$ ,  $x^0$  вещественные,  $\lambda_1 > 0$ ,  $\lambda_1 = -\lambda_2$ . Тогда

$$\begin{aligned} x^n &= c_1 \lambda_1^n e_1 + c_2 (-\lambda_1)^n e_2 + O(|\lambda_3|^n), \\ x^{n+2} &= c_1 \lambda_1^{n+2} e_1 + c_2 (-\lambda_1)^{n+2} e_2 + O(|\lambda_3|^n) = \\ &= \lambda_1^2 (c_1 \lambda_1^n e_1 + c_2 (-\lambda_1)^n e_2) + O(|\lambda_3|^n). \end{aligned}$$

Отсюда получаем, что

$$\bar{\lambda}^{(n)} = (x^{n+2}, x^n) / (x^n, x^n) = \lambda_1^2 + O(|\lambda_3 / \lambda_1|^n).$$

При  $\bar{\lambda}^{(n)} > 0$  полагаем  $\lambda_{i,2}^{(n)} = \pm \sqrt{\bar{\lambda}^{(n)}}$ . Имеем

$$z_1^{n+1} = x^{n+1} + \lambda_1^{(n)} x^n = 2c_1 \lambda_1^{n+1} e_1 + O(|\lambda_3|^n),$$

поэтому

$$e_1^{(n)} = z_1^{n+1} / \|z_1^{n+1}\| = e_1 + O(|\lambda_3 / \lambda_1|^n).$$

Точно так же

$$z_2^{n+1} = x^{n+1} + \lambda_2^{(n)} x^n = 2c_2 \lambda_2^{n+1} e_2 + O(|\lambda_3|^n)$$

и

$$e_2^{(n)} = z_2^{n+1} / \|z_2^{n+1}\| = e_2 + O(|\lambda_3 / \lambda_1|^n).$$

Если  $\lambda_i$  — собственные значения матрицы  $A$ , то у сопряженной матрицы  $A^*$  собственным значением будет  $\bar{\lambda}_i$ ; при этом если  $Ae_i = \lambda_i e_i$ ,  $A^*g_j = \bar{\lambda}_j g_j$ ,  $\lambda_i \neq \lambda_j$ , то  $(e_i, g_j) = 0$ . Поэтому при  $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \dots$  для нахождения собственного значения



$\lambda_2$  можно поступить так. Получив приближение  $\tilde{e}_1 \approx e_1$ , аналогично определяем приближение  $\tilde{g}_1$  к вектору  $g_1$ ; перенормируем  $\tilde{g}_1$  из условия  $(\tilde{e}_1, \tilde{g}_1) = 1$ . Далее ведем итерации по формуле  $y^{n+1} = Ay^n$ ; для исключения компоненты, пропорциональной  $e_1$ , время от времени векторы  $y^n$  ортогонализируем по отношению к вектору  $\tilde{g}_1$ , т. е. начальным для последующих итераций вместо  $y^n$  берем вектор

$$\tilde{y}^n = y^n - (y^n, \tilde{g}_1) \tilde{e}_1.$$

Естественно, что сходимость итерационного процесса лучше, если в начальном приближении  $y_0 = \sum_{i=1}^m d_i e_i$  слагаемое  $d_2 e_2$  преобладает над остальными слагаемыми. Возьмем из описанного выше итерационного процесса отыскания  $\lambda_1$  и  $e_1$  некоторое приближение  $x^l$ . Вектор  $y^0 = x^l - (x^l, \tilde{g}_1) \tilde{e}_1$  примем за начальное приближение. Не следует брать  $l$  малым, иначе компоненты  $c_j \lambda_j^l e_j$  при  $j > 2$  не будут малы по сравнению с  $c_2 \lambda_2^l e_2$ ; в то же время не следует  $l$  брать очень большим, поскольку в этом случае компонента  $c_2 \lambda_2^l e_2$  будет малой по сравнению с вычислительной погрешностью.

Часто можно встретить следующее высказывание: если  $c_1 = 0$ , то описанный итерационный процесс, казалось бы, не должен давать приближения, сходящиеся к максимальному по модулю собственному значению; однако в действительности из-за присутствия округлений в процессе итераций может появиться компонента, пропорциональная  $e_1$ , и вследствие этого требуемый результат все равно будет получен.

Реально при использовании современных ЭВМ с большой разрядностью может случиться, что после некоторого числа итераций влияние вычислительной погрешности еще не будет существенно, в то время как величина  $\sum_{i=3}^m c_i \lambda_i^n e_i$  будет мала по сравнению с  $c_2 \lambda_2^n e_2$ . Тогда  $Ax^n \approx \text{const } x^n$  и можно сделать неверный вывод, что искомое первое собственное значение найдено. Поэтому в тех случаях, когда нет уверенности в правильности найденного собственного значения, следует провести еще один или несколько расчетов с другими значениями  $y^0$ .

В ряде случаев появление составляющей, пропорциональной  $e_1$ , не является неизбежным даже в случае присутствия округлений. При решении проблемы собственных значений для дифференциальных и интегральных операторов иногда возникают матрицы  $A$  со специфическими свойствами. Например, часто встречается случай, когда при всех  $i, j$  выполняется равенство

$$a_{ij} = a_{m+1-i, m+1-j} \quad (2)$$

Для определенности рассмотрим случай четного  $m$ . Назовем четными векторы  $x$ , компоненты которых связаны равенством  $x_i = x_{m+1-i}$ ,  $i = 1, \dots, m$ , и нечетными — равенством  $x_i = -x_{m+1-i}$ . При условии (2) вектор  $Ax$  четен, если  $x$  четен, и нечетен, если  $x$  нечетен. Поэтому подпространства четных и нечетных векторов являются собственными для оператора  $A$ . Следовательно, существует полная система собственных векторов, принадлежащих этим подпространствам, т. е. являющихся или четными, или нечетными. Это обстоятельство может быть существенно использовано: если вектор  $x^0$  четен или нечетен, то этим же свойством обладают все векторы  $x^n$ , поэтому при отыскании

каждого последующего вектора  $x^n$  следует ограничиться определением первой половины его компонент. Кроме этого, можно объединить коэффициенты, соответствующие компонентам  $x_i$  и  $x_{m+1-i}$ . Например, при четном  $x^0$  вычисления компонент  $x^n$  можно производить по формулам

$$x_i^{n+1} = \sum_{j=1}^{m/2} (a_{ij} + a_{i, m+1-j}) x_j^n.$$

В случае итераций такого вида мы не выходим за пределы подпространств четных или нечетных векторов соответственно. Поэтому если  $x^0$  и  $e_1$  принадлежат различным подпространствам, составляющая, пропорциональная  $e_1$ , так и не появится.

Кроме непосредственного уменьшения вычислений при отыскании каждого вектора  $x^n$ , использование свойства (2) полезно также по следующей причине. Довольно типичен случай, когда собственные векторы с небольшими нечетными номерами являются четными, а с небольшими четными номерами — нечетными; предположим, что  $|\lambda_1| > |\lambda_2| > |\lambda_3| > \dots$ . Тогда при  $x^0$  четном всегда

$$x^n = c_1 \lambda_1^n e_1 + c_3 \lambda_3^n e_3 + \dots = c_1 \lambda_1^n e_1 + O(|\lambda_3|^n),$$

следовательно,

$$\lambda_1^{(n)} - \lambda_1 = O(|\lambda_3/\lambda_1|^n) \quad \text{при} \quad \lambda^{(n)} = (x^{n+1}, x^n)/(x^n, x^n).$$

Соответственно при  $x^0$  нечетном имеем

$$x^n = c_2 \lambda_2^n e_2 + O(|\lambda_4|^n)$$

и

$$\lambda_2^{(n)} - \lambda_2 = O(|\lambda_4/\lambda_2|^n) \quad \text{при} \quad \lambda^{(n)} = (x^{n+1}, x^n)/(x^n, x^n).$$

При этом не возникает никакой проблемы подавления составляющей, пропорциональной  $e_1$ .

Если  $|\lambda_1| > 1$ , то  $\|x^n\| \rightarrow \infty$  при  $n \rightarrow \infty$ , поэтому при достаточно большом  $n$  произойдет переполнение разрядности чисел и остановка ЭВМ. Если  $|\lambda_1| < 1$ , то  $\|x^n\| \rightarrow 0$ , и вследствие конечности порядков чисел в машине может случиться, что, начиная с некоторого  $n$ ,  $x^n \equiv 0$ . Чтобы избежать этих явлений, полезно время от времени нормировать вектор  $x^n$ , чтобы  $\|x^n\| = 1$ .

Для практической оценки погрешности и ускорения сходимости итерационных процессов может быть применен  $\delta^2$ -процесс и другие приемы, аналогичные методам ускорения сходимости при решении систем линейных уравнений. Например, могут применяться итерации вида  $x^{n+1} = g_k(A)x^n$  со специальным подбором в зависимости от известной информации о спектре матрицы  $A$  многочлена  $g_k(A)$ .

Поскольку

$$\max \lambda_A^i = \sup_x \frac{(Ax, x)}{(x, x)}, \quad \min \lambda_A^i = \inf_x \frac{(Ax, x)}{(x, x)}$$

при  $A = A^T$ , то некоторые приемы отыскания максимального и минимального собственных значений матрицы  $A$  основываются на идее отыскания стационарных точек функционала  $\Phi(x) = (Ax, x)/(x, x)$ .

**Задача 2.** Пусть  $\lambda_1 \approx 5$ ,  $1 \leq \lambda_i \leq 3$  при  $i = 2, \dots, m$ . Построить итерационный процесс вида  $x^{n+1} = (A + cE)x^n$  для получения  $\lambda_1$  с наилучшей при данной информации скоростью сходимости.

Сделать то же самое, если  $\lambda_1 \approx 1$ ,  $2 \leq |\lambda_i| \leq 3$  при  $i = 2, \dots, m$ .

### § 13. Решение полной проблемы собственных значений при помощи QR-алгоритма

Существует ряд тщательно отработанных алгоритмов и программ решения полной проблемы собственных значений. Поэтому в случае, когда возникает такая проблема, в первую очередь рекомендуется использовать стандартные программы решения таких задач. Наиболее совершенные из них основаны на различных модификациях QR-алгоритма, общая схема которого приводится ниже.

Пусть  $A$  — произвольная вещественная матрица. Согласно лемме из § 2 ее можно представить в виде  $A = U^T A_{n-1}$ , где  $U$  — ортогональная, а  $A_{n-1}$  — правая треугольная матрицы. Запишем это равенство в виде

$$A = Q_1 R_1, \quad (1)$$

где  $Q_1$  — ортогональная,  $R_1$  — правая треугольная матрицы. Из (1) имеем  $R_1 = Q_1^{-1} A$ , поэтому матрица  $A_1 = R_1 Q_1 = Q_1^{-1} A Q_1$  подобна матрице  $A$ .

Построим последовательность матриц  $A_n$  по следующему правилу. Матрицу  $A_n$  разлагаем на произведение ортогональной и правой треугольной матриц в виде  $A_n = Q_{n+1} R_{n+1}$  и полагаем  $A_{n+1} = R_{n+1} Q_{n+1}$ . Поскольку  $A_{n+1} = Q_{n+1}^{-1} A_n Q_{n+1}$ , то все матрицы  $A_n$  подобны между собой и подобны исходной матрице  $A$ .

Представим исходную матрицу  $A$  в виде  $A = Q \Lambda Q^{-1}$ , где  $\Lambda$  — правая каноническая форма Жордана, т. е.  $\lambda_{ij} = 0$  при  $j < i$  и  $j > i + 1$ ,  $\lambda_{ii} = \lambda_i$  — собственные значения матрицы  $A$ ,  $\lambda_{i, i+1}$  равны нулю или 1. Всегда можно подобрать матрицу  $Q$  так, чтобы диагональные элементы матрицы  $\Lambda$  были упорядочены в порядке невозрастания модуля

$$|\lambda_1| = \dots = |\lambda_{l_1}| > |\lambda_{l_1+1}| = \dots = |\lambda_{l_2}| > \dots \\ \dots > |\lambda_{l_{s-1}+1}| = \dots = |\lambda_{l_s}|.$$

**Теорема (без доказательства).** Пусть в разложении матрицы  $A$  все диагональные миноры матрицы  $Q$  не вырождены. Тогда последовательность матриц  $A_n$  при  $n \rightarrow \infty$  сходится по форме к клеточному правому треугольному виду  $\bar{A}$ .

Имеется в виду, что после некоторой перестановки строк и одновременно такой же перестановки столбцов матрицы  $A$  будут выполняться соотношения: если  $l_k < i \leq l_{k+1}$ ,  $j < i$  или  $l_{k+1} < j$ ,  $k = 1, \dots, s$ , то  $a_{ij}^{(n)} \rightarrow 0$ .

При реализации описанного алгоритма построения матриц  $A_n$  на практике мы увидим, что некоторые элементы матриц  $A_n$

оказываются малыми. Приравняв их нулю, мы получим клеточную правую треугольную матрицу. Характеристический многочлен этой матрицы равен произведению характеристических многочленов ее диагональных клеток.

Если требуется найти не только собственные значения матрицы  $A$ , но и ее собственные и присоединенные векторы, то в процессе построения последовательности матриц  $A_n$  следует запоминать ортогональные матрицы  $P_n = Q_1 \dots Q_n$ , вычисляемые по рекуррентной формуле  $P_{n+1} = P_n Q_{n+1}$ .

**Задача 1.** Доказать, что каждый шаг  $QR$ -алгоритма требует  $N \sim 10m^3/3$  арифметических операций.

На практике прибегают к различным способам ускорения сходимости. Один из этих способов заключается в следующем. Матрица  $A$  предварительно преобразуется в эквивалентную ей правую почти треугольную матрицу.

Матрица  $A$  называется *правой почти треугольной*, если  $a_{ij} = 0$  при  $j < i - 1$ .

Алгоритм преобразования матрицы  $A$  в правую почти треугольную матрицу заключается в последовательном построении матриц  $A_i$  таких, что первые  $l$  столбцов матрицы  $A_i$  имеют вид первых  $l$  столбцов правой почти треугольной матрицы, т. е.  $a_{ij} = 0$ , если  $j < i - 1$  и  $j \leq l$ . По элементам  $(l + 1)$ -го столбца матрицы  $A_i$  построим матрицу отражения  $U_{l+1}$  (см. § 2) так, чтобы в матрице  $B = U_{l+1} A_i$  элементы  $b_{1, l+1}, \dots, b_{l, l+1}$  были те же, что у матрицы  $A_i$ , а элементы  $b_{l+3, l+1}, \dots, b_{m, l+1}$  были нулевыми. Положим  $A_{l+1} = U_{l+1}^T A_i U_{l+1}$ . Умножение справа на матрицу  $U_{l+1}^T$  не меняет первых  $l + 1$  столбцов матрицы  $B$ , поэтому матрица  $A_{l+1}$  является матрицей требуемого вида. После получения правой почти треугольной матрицы  $A_{m-1}$  применяют  $QR$ -алгоритм в его первоначальной форме.

**Задача 2.** Доказать, что в этом случае каждый шаг  $QR$ -алгоритма требует  $N \sim 6m^2$  арифметических операций.

Для еще большего ускорения сходимости применяется вариант  $QR$ -алгоритма со сдвигом. А именно, строится последовательность ортогональных матриц  $Q_i$  и правых треугольных матриц  $R_i$  по рекуррентным формулам

$$\begin{aligned} A - v_1 E &= Q_1 R_1, & A_1 &= R_1 Q_1 + v_1 E, \\ \dots & \dots & \dots & \dots \\ A_{l-1} - v_l E &= Q_l R_l, & A_l &= R_l Q_l + v_l E. \end{aligned}$$

Матрицы  $A_i$  подобны матрице  $A$ ; за счет введения «сдвигов»  $v_i$  удается добиться ускорения сходимости. Вопрос о наиболее целесообразном выборе параметров  $v_i$  мы рассматривать не будем.

В этой главе было, в частности, приведено большое количество методов решения линейных систем уравнений. Какой же из этих методов все-таки стоит выбрать, решая задачу?

Если порядок системы небольшой и по затратам машинного времени число арифметических операций порядка  $m^3$ , где  $m$  — порядок системы, является приемлемым, то проще всего обратиться к стандартным программам методов отражения (число арифметических действий  $N \sim 2m^3/3$ ) или метода вращения (число арифметических действий  $N \sim 4m^3/3$ , но меньше накопление вычислительной погрешности).

Конечно, при этом надо иметь в виду, что точные методы типа методов отражений или вращений для матрицы общего вида требуют одновременного хранения в памяти ЭВМ порядка  $m^2$  чисел. Если такое количество чисел не помещается в оперативной памяти ЭВМ, а обмен информацией между оперативной и внешней памятью происходит недостаточно быстро или из-за структуры программы, или из-за возможностей ЭВМ, то применение этих программ может оказаться нецелесообразным.

В случае, когда применение этих методов нецелесообразно, имеет смысл проанализировать возможности применения простейших по своей структуре итерационных методов: простой итерации, Зейделя, сверхрелаксации, наискорейшего спуска. Если решается отдельная задача, то вследствие простоты соответствующих программ применение этих методов может быть вполне целесообразным. Если применение этих методов требует больших затрат машинного времени, то следует проанализировать возможности применения более сложных по своей структуре методов: оптимального линейного итерационного процесса, метода с использованием корней многочленов Чебышева, метода сопряженных градиентов, итерационных методов, использующих спектрально эквивалентные операторы.

Если размерность задачи столь велика, что само решение задачи, т. е. вектор  $X$ , не помещается в оперативной памяти ЭВМ, то иногда применяются вероятностные методы решения систем линейных уравнений, которые остались вне нашего рассмотрения.

Литература к гл. VI: [1, 3, 9—11, 18, 19, 25, 33, 45, 50, 63, 64].

## Глава VII

### РЕШЕНИЕ СИСТЕМ НЕЛИНЕЙНЫХ УРАВНЕНИЙ И ЗАДАЧ ОПТИМИЗАЦИИ

Решение задач оптимизации, будь то оптимизация производственных или экономических процессов, оптимизация конструкций или оптимизация численных алгоритмов, сводится в математической формулировке исследуемой задачи к отысканию точек экстремума функционалов. В наиболее типичных случаях возникает задача минимизации функции большого числа переменных в области  $\Omega$ , задаваемой большим числом ограничений типа неравенств или равенств: ищется

$$\inf_{x_1, \dots, x_m} \Phi(x_1, \dots, x_m)$$

при условиях

$$\begin{aligned} \varphi_i(x_1, \dots, x_m) &\geq 0, & i = 1, \dots, l; \\ \psi_i(x_1, \dots, x_m) &= 0, & i = 1, \dots, q. \end{aligned}$$

Задача минимизации функций большого числа переменных возникает также в случае применения вариационных методов к решению задач математической физики и в других разделах прикладной математики.

Системы уравнений

$$f_i(x_1, \dots, x_m) = 0, \quad i = 1, \dots, m,$$

которые мы будем также обозначать

$$F(x) = 0,$$

возникают в случае многих задач указанных выше типов; например, в гл. X будет идти речь о подобных системах, возникающих при решении краевых задач.

Задача минимизации функции и решения системы уравнений сводятся друг к другу. Если  $\Psi(y_1, \dots, y_m) > 0$  при

$$(y_1, \dots, y_m) \neq (0, \dots, 0), \quad \Psi(0, \dots, 0) = 0,$$

то решение системы  $F(x) = 0$  равносильно минимизации функции

$$\Psi(f_1(x_1, \dots, x_m), \dots, f_m(x_1, \dots, x_m)).$$

С другой стороны, пусть  $\inf_G \Phi(x_1, \dots, x_m)$  достигается в точке  $X$ , внутренней по отношению к  $G$ , и функция  $\Phi$  дифференцируема в этой точке. Тогда точка минимума является решением системы уравнений

$$\Phi'_{x_i} = 0, \quad i = 1, \dots, m.$$

Возможность сведения одной из этих задач к другой не означает, что достаточно ограничиться рассмотрением только одной из них; родство этих задач скорее подчеркивает, что они одинаково трудны. О трудности этих задач свидетельствует то, что не существует универсальных алгоритмов решения, практически пригодных при больших  $m$ ; отсутствие таких алгоритмов вызвано существом дела. При решении задач каждого типичного в приложениях класса приходится заниматься теоретической и экспериментальной «доводкой» методов применительно к этому классу задач.

Сведение задачи минимизации функции к системе нелинейных уравнений или наоборот производится на практике с целью снижения трудоемкости решения.

Например, при решении систем нелинейных уравнений иногда поступают следующим образом. Строится функционал, минимум которого достигается на решении системы. Затем, задавшись начальным приближением к точке минимума, проводят итерации каким-либо из методов спуска (см. § 3) и таким путем получают удовлетворительное приближение к решению системы. Исходя из этого приближения производят уточнения при помощи какого-либо итерационного метода, специфического для задачи решения системы уравнений, например метода Ньютона (см. § 2).

Поясним причины, вызывающие такое комбинированное применение методов. Назовем *областью сходимости метода* множества начальных условий, при которых итерации по данному методу сходятся к решению задачи. Применение методов спуска на первоначальном этапе вызвано тем, что обычно они имеют более широкую область сходимости, чем методы, специфические для задачи решения системы уравнений. В то же время последние методы обычно обладают лучшей скоростью сходимости при наличии достаточно хорошего начального приближения; это и обуславливает их применение на заключительном этапе итераций.

На примере решения системы линейных уравнений было также видно, что сведение этой задачи к задаче нахождения минимума функционала приводит к конструированию новых методов решения исходной задачи.

## § 1. Метод простой итерации и смежные вопросы

Так же как в случае линейных уравнений, начнем изучение итерационных методов с *метода простой итерации*.

Этот метод состоит в следующем: система уравнений преобразуется к виду

$$\mathbf{x} = \mathbf{g}(\mathbf{x}), \quad (1)$$

иначе,

$$x_i = g_i(x_1, \dots, x_m), \quad i = 1, \dots, m,$$

и итерации проводятся по формуле

$$\mathbf{x}^{n+1} = \mathbf{g}(\mathbf{x}^n), \quad (2)$$

иначе,

$$x_i^{n+1} = g_i(x_1^n, \dots, x_m^n), \quad i = 1, \dots, m.$$

Подойдем к изучению этого метода с более общих позиций. Пусть  $\mathcal{H}$  — полное метрическое пространство, а оператор  $\mathbf{y} = \mathbf{g}(\mathbf{x})$  отображает  $\mathcal{H}$  в себя. Рассмотрим итерационный процесс

$$\mathbf{x}^{n+1} = \mathbf{g}(\mathbf{x}^n) \quad (3)$$

решения уравнения

$$\mathbf{x} = \mathbf{g}(\mathbf{x}). \quad (4)$$

Если при некотором  $q < 1$  отображение  $\mathbf{y} = \mathbf{g}(\mathbf{x})$  удовлетворяет условию

$$\rho(\mathbf{g}(\mathbf{x}_1), \mathbf{g}(\mathbf{x}_2)) \leq q\rho(\mathbf{x}_1, \mathbf{x}_2) \quad (5)$$

при всех  $\mathbf{x}_1, \mathbf{x}_2$ , то такое отображение называют *сжимающим*.

**Теорема.** Если отображение  $\mathbf{y} = \mathbf{g}(\mathbf{x})$  сжимающее, то уравнение  $\mathbf{x} = \mathbf{g}(\mathbf{x})$  имеет решение  $\mathbf{X}$  и

$$\rho(\mathbf{X}, \mathbf{x}^n) \leq \frac{q^n a}{1 - q};$$

здесь  $a = \rho(\mathbf{x}^1, \mathbf{x}^0)$ ,  $\rho(\mathbf{x}, \mathbf{y})$  — расстояние между  $\mathbf{x}$  и  $\mathbf{y}$ .

**Доказательство.** Согласно (5) имеем

$$\rho(\mathbf{x}^{n+1}, \mathbf{x}^n) = \rho(\mathbf{g}(\mathbf{x}^n), \mathbf{g}(\mathbf{x}^{n-1})) \leq q\rho(\mathbf{x}^n, \mathbf{x}^{n-1}),$$

поэтому  $\rho(\mathbf{x}^{n+1}, \mathbf{x}^n) \leq q^n \rho(\mathbf{x}^1, \mathbf{x}^0) = q^n a$ . При  $l > n$  имеем цепочку неравенств

$$\begin{aligned} \rho(\mathbf{x}^l, \mathbf{x}^n) &\leq \rho(\mathbf{x}^l, \mathbf{x}^{l-1}) + \dots + \rho(\mathbf{x}^{n+1}, \mathbf{x}^n) \leq \\ &\leq q^{l-1}a + \dots + q^n a \leq q^n a \sum_{i=0}^{\infty} q^i = \frac{q^n a}{1 - q}. \end{aligned} \quad (6)$$

Согласно критерию Коши последовательность  $\mathbf{x}^n$  имеет некоторый предел  $\mathbf{X}$ . Переходя к пределу в (6) при  $l \rightarrow \infty$ , получаем

$$\rho(\mathbf{X}, \mathbf{x}^n) \leq \frac{q^n a}{1 - q}.$$







Таким образом, при приближениях, находящихся в малой окрестности решения, погрешности приближений итерационного процесса (2) (а также и процессов (7) и (8)) подчиняются примерно тем же законам, что и погрешности итерационных методов решения систем линейных уравнений. Наличие соотношения (9) позволяет производить ускорение сходимости итерационных процессов.

Рассмотрим случай  $m = 1$  и построим аналог  $\delta^2$ -процесса. При имеющемся приближении  $x^n$  обозначим  $x^{n1} = g(x^n)$ ,  $x^{n2} = g(x^{n1})$ . Согласно (9)

$$\begin{aligned}x^{n1} - X &\approx g'(X)(x^n - X), \\x^{n2} - X &\approx g'(X)(x^{n1} - X).\end{aligned}$$

Из этих соотношений получаем

$$\begin{aligned}g'(X) &\approx \frac{x^{n2} - x^{n1}}{x^{n1} - x^n}, \\X &\approx \frac{x^{n2} - g'(X)x^{n1}}{1 - g'(X)} \approx \frac{x^{n2} - \frac{x^{n2} - x^{n1}}{x^{n1} - x^n}x^{n1}}{1 - \frac{x^{n2} - x^{n1}}{x^{n1} - x^n}} = \frac{x^{n2}x^n - (x^{n1})^2}{x^{n2} - 2x^{n1} + x^n}.\end{aligned}$$

За следующее после  $x^n$  приближение примем

$$x^{n+1} = \frac{x^{n2}x^n - (x^{n1})^2}{x^{n2} - 2x^{n1} + x^n} = \frac{x^n g(g(x^n)) - (g(x^n))^2}{g(g(x^n)) - 2g(x^n) + x^n}. \quad (10)$$

Для характеристики методов решения уравнений вводится понятие *порядка метода*. Говорят, что метод имеет  $k$ -й порядок, если существуют  $c_1 > 0$ ,  $c_2 < \infty$  такие, что

$$\rho(x^{n+1}, X) \leq c_2 (\rho(x^n, X))^k$$

при условии  $\rho(x^n, X) \leq c_1$ . Очевидно, чем больше  $k$ , тем быстрее сходится процесс итераций при малых значениях  $\rho(x^n, X)$ . В вычислительной практике широко распространены методы второго порядка (например, метод, определяемый формулой (10), или метод Ньютона, рассматриваемый в следующем параграфе).

**Примечание.** Иногда в литературе встречается другое, на наш взгляд неразумное, определение порядка метода: говорят, что метод решения системы уравнений  $F(x) = 0$  имеет порядок  $k$ , если при его реализации вычисляются производные функций  $f_i$  до порядка  $k - 1$  включительно.

В § 6.10 были рассмотрены итерационные методы решения линейных систем с помощью спектрально-эквивалентных операторов. Аналогичные методы применяются и для решения нелинейных систем. Выбирается оператор  $G(x)$  такой, что  $x = 0$  является единственным решением уравнения  $G(x) = 0$ . Приближения  $x^{n+1}$  к решению системы  $F(x) = 0$  определяются из соотношения

$$G(x^{n+1} - x^n) = F(x^n). \quad (11)$$

Наиболее распространен случай, когда  $G$  — линейный оператор, например матрица. В ряде случаев оператор  $G$  выбирается зависящим от  $n$ , а также от приближения  $x^n$ . Тогда в схему (11) укладывается также рассматриваемый ниже метод Ньютона решения нелинейных задач.

## § 2. Метод Ньютона решения нелинейных уравнений

Если известно достаточно хорошее начальное приближение к решению системы уравнений

$$F(x) = 0, \quad (1)$$

то эффективным методом повышения точности является *метод Ньютона*.

Идея метода Ньютона заключается в том, что в окрестности имеющегося приближения  $x_n$  задача заменяется некоторой вспомогательной линейной задачей.

Последняя задача выбирается так, чтобы погрешность замены имела более высокий порядок малости, чем первый (в определяемом далее смысле), в окрестности имеющегося приближения. За следующее приближение принимается решение этой вспомогательной задачи.

Рассмотрим случай скалярного уравнения  $f(x) = 0$ . В качестве такой вспомогательной задачи естественно взять линейную задачу

$$f(x_n) + f'(x_n)(x - x_n) = 0.$$

Ее решение  $x = x_n - f(x_n)/f'(x_n)$  принимается за следующее приближение  $x_{n+1}$  к решению исходного уравнения, т. е. итерации ведутся по формуле

$$x_{n+1} = x_n - f(x_n)/f'(x_n).$$

Рассмотрим более общий случай — решение нелинейного функционального уравнения.

Пусть  $F(x)$  — оператор, отображающий линейное нормированное пространство  $\mathcal{X}$  на линейное нормированное пространство  $\mathcal{Y}$ , может быть и совпадающее с  $\mathcal{X}$ . Нормы в этих пространствах соответственно обозначаем  $\|\cdot\|_{\mathcal{X}}$  и  $\|\cdot\|_{\mathcal{Y}}$ . Линейный оператор  $P$ , действующий из пространства  $\mathcal{X}$  в пространство  $\mathcal{Y}$ , назовем производной оператора  $F(x)$  в точке  $x$ , если

$$\|F(x + \eta) - F(x) - P\eta\|_{\mathcal{Y}} = o(\|\eta\|_{\mathcal{X}}) \quad (2)$$

при  $\|\eta\|_{\mathcal{X}} \rightarrow 0$ .

В дальнейшем будем обозначать такой оператор  $P$  через  $F'(x)$ . Пусть, например,

$$x = (x_1, \dots, x_m)^T, \quad F = (f_1, \dots, f_m)^T.$$

Если функции  $f_i$  непрерывно дифференцируемы в окрестности данной точки  $\mathbf{x}$ , то

$$f_i(x_1 + \eta_1, \dots, x_m + \eta_m) = \\ = f_i(x_1, \dots, x_m) + \sum_{j=1}^m \frac{\partial f_i(x_1, \dots, x_m)}{\partial x_j} \eta_j + o(\|\eta\|).$$

Совокупность этих соотношений можно переписать в виде (2), если за  $P$  принять оператор умножения слева на матрицу

$$F'(\mathbf{x}) = \left[ \frac{\partial f_i}{\partial x_j} \right].$$

В простейшем случае  $m = 1$  оператор  $P$  превращается в оператор умножения на производную  $f'_x$ .

Пусть  $\mathbf{X}$  — решение уравнения  $F(\mathbf{X}) = 0$ ,  $\mathbf{x}^n$  — некоторое приближение к  $\mathbf{X}$ . В предположении существования производной  $F'$ , согласно (2), имеем

$$\|F(\mathbf{X}) - F(\mathbf{x}^n) - F'(\mathbf{x}^n)(\mathbf{X} - \mathbf{x}^n)\|_{\mathcal{Y}} = o(\|\mathbf{X} - \mathbf{x}^n\|_{\mathcal{X}}). \quad (3)$$

Если величина  $\|\mathbf{X} - \mathbf{x}^n\|_{\mathcal{X}}$  мала, то можно написать приближенное равенство

$$F(\mathbf{x}^n) + F'(\mathbf{x}^n)(\mathbf{X} - \mathbf{x}^n) \approx F(\mathbf{X}).$$

Поскольку  $F(\mathbf{X}) = 0$ , то

$$F(\mathbf{x}^n) + F'(\mathbf{x}^n)(\mathbf{X} - \mathbf{x}^n) \approx 0.$$

Возьмем в качестве следующего приближения  $\mathbf{x}^{n+1}$  решение уравнения

$$F(\mathbf{x}^n) + F'(\mathbf{x}^n)(\mathbf{x}^{n+1} - \mathbf{x}^n) = 0,$$

если такое решение существует. Между прочим, последнее уравнение имеет вид (1.11). В предположении, что оператор  $F'$  обратим, это решение можно записать в виде

$$\mathbf{x}^{n+1} = \mathbf{x}^n - (F'(\mathbf{x}^n))^{-1} F(\mathbf{x}^n). \quad (4)$$

Такой итерационный процесс называют *методом Ньютона*.

Пусть  $\Omega_a = \{\mathbf{x}: \|\mathbf{x} - \mathbf{X}\|_{\mathcal{X}} < a\}$ . Пусть при некоторых  $a$ ,  $a_1$ ,  $a_2$ ,  $0 < a$ ,  $0 \leq a_1$ ,  $a_2 < \infty$ , выполнены условия:

$$1) \quad \|(F'(\mathbf{x}))^{-1}\|_{\mathcal{Y}} \leq a_1 \quad \text{при } \mathbf{x} \in \Omega_a; \quad (5)$$

$$2) \quad \|F(\mathbf{u}_1) - F(\mathbf{u}_2) - F'(\mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)\|_{\mathcal{Y}} \leq a_2 \|\mathbf{u}_2 - \mathbf{u}_1\|_{\mathcal{X}}^2 \quad (6)$$

при  $\mathbf{u}_1, \mathbf{u}_2 \in \Omega_a$ . Обозначим  $c = a_1 a_2$ ,  $b = \min(a, c^{-1})$ .

Теорема (о сходимости метода Ньютона). При условиях 1), 2) и  $\mathbf{x}^0 \in \Omega_b$  итерационный процесс Ньютона (4) сходится с оценкой погрешности

$$\|\mathbf{x}^n - \mathbf{X}\|_{\mathcal{X}} \leq c^{-1} (c \|\mathbf{x}^0 - \mathbf{X}\|_{\mathcal{X}})^{2^n}. \quad (7)$$

Примечание. Если в рассматривавшемся выше примере в некоторой окрестности решения функции  $f_i$  имеют ограниченные вторые производные, то, согласно формуле Тейлора, имеем

$$f_i(y) = f_i(x) + \sum_{j=1}^n \frac{\partial f_i(x_1, \dots, x_n)}{\partial x_j} (y_j - x_j) + O(\|y - x\|^2),$$

и, таким образом, условие (2) выполнено.

Доказательство. Пусть  $x^0 \in \Omega_b$ . Индукцией по  $n$  докажем, что все  $x^n \in \Omega_b$ . Пусть это утверждение доказано при некотором  $n$ ; так как  $b \leq a$ , то тогда  $x^n \in \Omega_a$ . Подставив в (6)  $u_1 = X$  и  $u_2 = x^n$ , получим

$$\|F(X) - F(x^n) - F'(x^n)(X - x^n)\|_y \leq a_2 \|x^n - X\|_{\mathcal{E}}^2.$$

Поскольку  $F(x^n) = -F'(x^n)(x^{n+1} - x^n)$ , а  $F(X) = 0$ , то это соотношение может быть переписано в виде

$$\|F'(x^n)(x^{n+1} - X)\|_y \leq a_2 \|x^n - X\|_{\mathcal{E}}^2.$$

Воспользовавшись (5), получаем неравенство

$$\|x^{n+1} - X\|_{\mathcal{E}} \leq c \|x^n - X\|_{\mathcal{E}}^2. \quad (3)$$

Отсюда следует, что

$$\|x^{n+1} - X\|_{\mathcal{E}} < cb^2 = (cb)b \leq b,$$

поэтому  $x^{n+1}$  также принадлежит  $\Omega_b$ . Таким образом, при  $x^0 \in \Omega_b$  все  $x^n$  принадлежат  $\Omega_b$  и, следовательно, для них выполняется (8).

Пусть  $q_n = c \|x^n - X\|_{\mathcal{E}}$ . После умножения на  $c$  неравенство (8) запишется в виде  $q_{n+1} \leq q_n^2$ . Индукцией по  $n$  докажем справедливость неравенства

$$q_n \leq q_0^{2^n}.$$

При  $n = 0$  оно очевидно. Предположив его верным при  $n = k$ , получаем

$$q_{k+1} \leq q_k^2 \leq (q_0^{2^k})^2 = q_0^{2^{k+1}}.$$

Таким образом,  $q_n \leq q_0^{2^n}$  при всех  $n$ . Это означает, что

$$c \|x^n - X\|_{\mathcal{E}} \leq (c \|x^0 - X\|_{\mathcal{E}})^{2^n}.$$

Отсюда следует (7). Согласно определению  $c$  и  $b$ ,

$$c \|x^0 - X\|_{\mathcal{E}} < cb \leq 1,$$

и поэтому  $x^n \rightarrow X$ . Теорема доказана.

Обращение оператора  $F'(x^n)$  зачастую оказывается более трудоемкой операцией, чем вычисление значения  $F(x^n)$ . Поэтому метод Ньютона часто модифицируется следующим образом. По ходу вычислений выбирают или заранее задаются некоторой воз-

растающей последовательностью чисел  $n_0=0, n_1, n_2 \dots$ . При  $n_k \leq n < n_{k+1}$  итерации производят по формуле

$$x^{n+1} = x^n - (F'(x^{n_k}))^{-1} F(x^n).$$

Увеличение числа итераций, сопровождающее такую модификацию, компенсируется большей «дешевизной» одного шага итерации. Выбор последовательности  $\{n_k\}$  нужно производить с обоюдным учетом этих факторов.

Рассмотрим геометрическую интерпретацию метода Ньютона в случае решения скалярного уравнения  $f(x)=0$ . Тогда расчетная формула (4) приобретает вид

$$x^{n+1} = x^n - f(x^n)/f'(x^n). \quad (9)$$

Для получения  $x^{n+1}$  геометрически надо найти абсциссу точки пересечения с осью  $x$  касательной к кривой  $y=f(x)$  в точке  $(x^n, f(x^n))$  (рис. 7.2.1). Уже в случае, когда  $f(x)$  — многочлен третьей степени, может случиться, что последовательность  $\{x^n\}$  не сходится к корню при плохом начальном при-

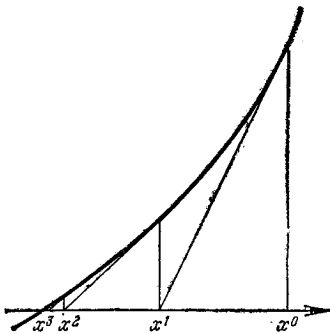


Рис. 7.2.1

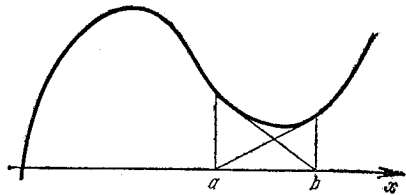


Рис. 7.2.2

ближении. Например, в случае, изображенном на рис. 7.2.2, все четные приближения совпадают с  $a$ , а нечетные — с  $b$ ; метод, как говорят, «зациклился». Для более сложных задач реальное поведение приближений  $x^n$  при плохом начальном приближении становится существенно более запутанным и трудно поддающимся анализу.

Сравним асимптотическую скорость сходимости методов Ньютона и простой итерации. Для последнего мы имели оценку погрешности

$$\|x^n - X\| \leq q^n \|x^0 - X\|, \quad q < 1.$$

Чтобы погрешность стала меньше  $\varepsilon$ , согласно этой оценке достаточно взять

$$n \geq \log_{q^{-1}} \frac{\|x^0 - X\|}{\varepsilon} \sim \log_{q^{-1}} \frac{1}{\varepsilon}.$$

В случае метода Ньютона правая часть (7) будет меньше  $\varepsilon$ , если

$$n \geq -\log_2 \frac{\log_2(c \|x^0 - X\|)}{\log_2(ce)} \sim \log_2 \log_2 \frac{1}{\varepsilon}. \quad (10)$$

Таким образом, асимптотически, при  $\varepsilon \rightarrow 0$ , метод Ньютона требует меньшего числа итераций.

Обратим внимание, что метод Ньютона, записанный в форме (4), сам является разновидностью метода простой итерации. В случае скалярного уравнения  $f(x) = 0$  хорошо видна еще одна особенность метода Ньютона.

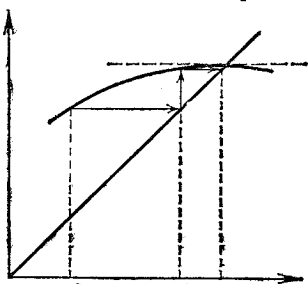


Рис. 7.2.3

Производная правой части (9)  $g(x) = x - f(x)/f'(x)$  по  $x$  равна  $f(x)f''(x)/(f'(x))^2$ . Таким образом,  $g'(X) = 0$ , если  $f'(X) \neq 0$ , и рис. 7.1.1 в этом случае приобретает следующий вид (рис. 7.2.3).

Метод Ньютона оказывается удобным способом извлечения корней целой степени. Задача

извлечения корня  $\sqrt[p]{a}$ ,  $p$  — целое число, равносильна задаче решения уравнений  $x^p - a = 0$ . Расчетная формула метода Ньютона в этом случае приобретает вид

$$x_{n+1} = \frac{p-1}{p} x_n + \frac{a}{p x_n^{p-1}}.$$

**Задача 1.** Рассматривается алгоритм вычисления  $\sqrt{a}$  при  $1 \leq a \leq 4$ ,  $x_0$  полагается равным значению многочлена наилучшего равномерного приближения для  $\sqrt{a}$  на  $[1, 4]$ :  $x_0 = p_1(a) = \frac{17}{24} + \frac{a}{3}$ . Убедиться в справедливости неравенства  $|x_4 - \sqrt{a}| \leq 0,5 \cdot 10^{-25}$ .

В случае решения одного скалярного уравнения  $f(x) = 0$  наряду с методом Ньютона употребителен *метод секущих*.

Простейший вариант этого метода заключается в следующем. В процессе итераций фиксируется некоторая точка  $x^0$ . Приближение  $x^{n+1}$  находится как абсцисса точки пересечения прямой, проходящей через точки  $(x^0, f(x^0))$  и  $(x^n, f(x^n))$ , с осью  $x$  (рис. 7.2.4).

Более эффективен способ, где за  $x^{n+1}$  принимается абсцисса точки пересечения с осью  $x$  прямой, проходящей через точки  $(x^{n-1}, f(x^{n-1}))$  и  $(x^n, f(x^n))$  (рис. 7.2.5). Уравнение этой прямой

$$y_n(x) = f(x^n) + (x - x^n) \frac{f(x^n) - f(x^{n-1})}{x^n - x^{n-1}}.$$

Из условия  $y_n(x^{n+1}) = 0$  получаем

$$x^{n+1} = x^n - \frac{f(x^n)(x^n - x^{n-1})}{f(x^n) - f(x^{n-1})}. \quad (11)$$

Для достижения точности  $\varepsilon$  этим методом, как и в случае метода Ньютона, при достаточно хороших начальных приближениях требуется  $O(\ln \ln(1/\varepsilon))$  итераций. Вычисления прекращают, когда одна из величин  $|x^{n+1} - x^n|$  или  $|f(x^{n+1}) - f(x^n)|$  становится меньше некоторого заранее заданного малого  $\delta > 0$ .



При решении системы  $m$  уравнений

$$F(x) = 0$$

одним из возможных обобщений метода секущих является следующий метод. Пусть определены приближения  $x^{n-m}, \dots, x^n$  и известны значения

$$f_i(x^{n-m}), \dots, f_i(x^n).$$

Пусть  $y = L_i(x)$  — уравнение плоскости, проходящей через точки

$$(x^{n-m}, f_i(x^{n-m})), \dots, (x^n, f_i(x^n));$$

за следующее приближение  $x^{n+1}$  принимаем решение системы уравнений

$$L_i(x) = 0, \quad i = 1, \dots, m.$$

Для  $m > 1$  этот метод применяется редко, обычно в случае, когда можно ограничиться невысокой точностью.

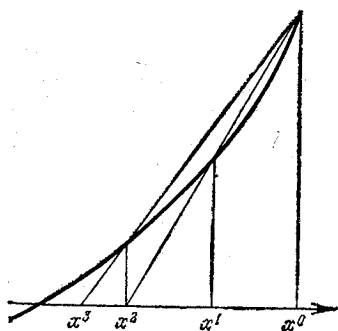


Рис. 7.2.4

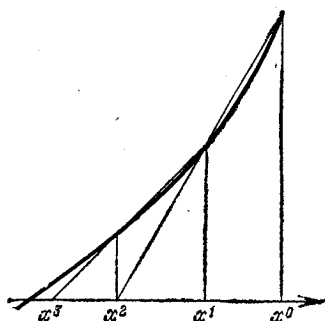


Рис. 7.2.5

Дело в том, что для этого метода при  $n \rightarrow \infty$  характерно «сплющивание»  $m$ -мерного тетраэдра с вершинами в точках  $x^{n-m}, \dots, x^n$ . Следствием этого является быстрое ухудшение обусловленности системы уравнений  $L_i(x) = 0$ . В результате алгоритм вычислений становится неустойчивым к вычислительной погрешности и часто перестает сходиться.

Кроме описанных выше, существует большое число других методов подобного типа, где в окрестности корня функции  $f(x)$  приближается некоторой функцией  $g(x)$ , для которой уравнение  $g(x) = 0$  решается в явном виде. Однако для применения всех этих методов необходимо достаточно хорошее приближение к решению. Иногда для его определения используется *метод вилки*. Определяют  $a_0, b_0$  такие, что  $f(a_0)f(b_0) < 0$ ; выбирают каким-либо образом точку  $c_0 \in (a_0, b_0)$ , например берут  $c_0 = (a_0 + b_0)/2$  или за  $c_0$  берут точку пересечения секущей, проходящей через точки  $(a_0, f(a_0)), (b_0, f(b_0))$ , с осью  $x$ . После вычисления  $f(c_0)$

за  $[a_1, b_1]$  принимают тот из отрезков  $[a_0, c_0]$ ,  $[c_0, b_0]$ , на концах которого  $f(x)$  принимает противоположные знаки, и т. д.

Важной задачей является разработка эффективных методов решения уравнений отдельных типичных классов. Для нахождения корней многочлена  $P(z) = a_0 z^m + \dots + a_m$  как с действительными, так и с комплексными коэффициентами таким методом является *метод парабол*. При заданных приближениях к корню  $z_{n-2}, z_{n-1}, z_n$  приближение  $z_{n+1}$  определяется следующим образом. Строится интерполяционный многочлен второй степени, совпадающий с  $P(z)$  в точках  $z_{n-2}, z_{n-1}, z_n$ . За  $z_{n+1}$  принимается корень этого многочлена, наиболее близкий к  $z_n$ . В стандартных программах метода парабол эта схема подвергнута некоторой модификации; хотя сходимость метода для произвольного многочлена при произвольных начальных условиях не доказана, не зарегистрировано ни одного случая, когда этот метод не сходится или сходится медленно.

### § 3. Методы спуска

Для решения задачи минимизации функционала наиболее часто применяются *методы спуска*. При заданном приближении определяется какое-либо направление, в котором функционал убывает, и производится перемещение приближения в этом направлении. Если величина перемещения взята не очень большой, то значение функционала обязательно уменьшится.

Рассмотрим примеры методов спуска.

При исследовании сходимости метода Зейделя в случае системы уравнений  $Ax = b$  при  $A > 0$  мы описали циклический метод покоординатного спуска минимизации функции  $\Phi(x_1, \dots, x_m)$ . При заданном приближении  $\Phi(x_1^0, \dots, x_m^0)$  отыскивается значение  $x_1 = x_1^1$ , при котором достигается  $\inf_{x_1} \Phi(x_1, x_2^0, \dots, x_m^0)$ , затем отыскивается значение  $x_2 = x_2^1$ , при котором достигается  $\inf_{x_2} \Phi(x_1^1, x_2, x_3^0, \dots, x_m^0)$ , и т. д. Процесс циклически повторяется.

Обозначим через  $P_k x$  приближение, получаемое при спуске из  $x$  по координате  $x_k$ . Присваивая приближению, получающемуся при спуске по очередной координате, следующий номер, можно записать приближения метода циклического покоординатного спуска в виде

$$\begin{aligned} x^1 &= P_1 x, & x^2 &= P_2 P_1 x, \dots, \\ x^m &= P_m \dots P_1 x, & x^{m+1} &= P_1 P_m \dots P_1 x, \dots \end{aligned}$$

При практической реализации этого метода возникает проблема минимизации функций одной переменной. Рассмотрим отдельно задачу минимизации функции одной переменной  $\Phi(x)$  при начальном приближении к точке минимума  $x = x^0$ . Так как эта задача обычно не может быть решена точно, то часто посту-

пают следующим образом: берут некоторые значения  $\bar{x}^0, \bar{\bar{x}}$ , и строят параболу  $y = Q_2(x)$ , удовлетворяющую условиям

$$Q_2(x^0) = \Phi(x^0), \quad Q_2(\bar{x}^0) = \Phi(\bar{x}^0), \quad Q_2(\bar{\bar{x}}^0) = \Phi(\bar{\bar{x}}^0).$$

Абсциссу  $x$  точки минимума  $Q_2(x)$  принимают за следующее приближение  $x^1$ . Уже в одномерном случае можно построить пример, когда последовательность точек, получаемых по описываемому методу, не обязательно сходится к искомой точке экстремума функции  $\Phi(x)$ .

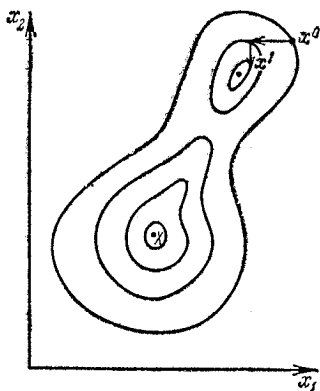


Рис. 7.3.1

Даже если на каждом шаге отыскивается абсолютный экстремум функции  $\Phi(x_1, \dots, x_m)$  по соответствующей координате, то уже при  $m = 2$  может случиться, что итерационный процесс сходится не к искомой точке абсолютного экстремума, а к некоторой точке локального экстремума. На рис. 7.3.1 изображены линии уровня такой функции и получаемые приближения. Спуск в циклическом порядке необязателен. Если из рассмотрения проводившихся ранее вычислений видно, что спуск по каким-либо координатам обеспечивает

наибольшее убывание  $\Phi(x)$ , то иногда целесообразен более частый спуск по этим координатам.

В других случаях при каждом  $n$  после получения приближения  $x^n$  выбирается некоторая совокупность координат

$$x_{i(n,1)}, \dots, x_{i(n,q(n))}$$

производятся независимые спуски по координатам этой группы исходя из приближения  $x^n$ , т. е. находят точки  $P_{i(n,k)}x^n$ . Далее вычисляется

$$\min_{1 \leq k \leq q(n)} \Phi(P_{i(n,k)}x^n)$$

и соответствующая минимуму точка  $P_{i(n,k)}x^n$  принимается за  $x^{n+1}$ .

Иногда номер очередной координаты, по которой осуществляется спуск, выбирается недетерминированно. В этом случае говорят о *случайном покоординатном спуске*.

Другой вариант метода спуска — *метод наискорейшего (градиентного) спуска*. Следующее приближение отыскивается в виде

$$x^{n+1} = x^n - \delta_n \text{grad } \Phi(x^n)$$

(рис. 7.3.2). Значение  $\delta_n$  определяется из условия

$$\min_{\delta_n} \Phi(x^n - \delta_n \text{grad } \Phi(x^n)),$$

т. е. этот алгоритм опять состоит в последовательной минимизации функции одной переменной  $\delta_n$ .

Как и в методе покоординатного спуска, в методе наискорейшего спуска нет необходимости полного решения вспомогательной задачи минимизации функции одной переменной. В окрестности точки своего минимума эта функция меняется мало, и тщательное нахождение ее точки минимума не приводит к существенному эффекту.

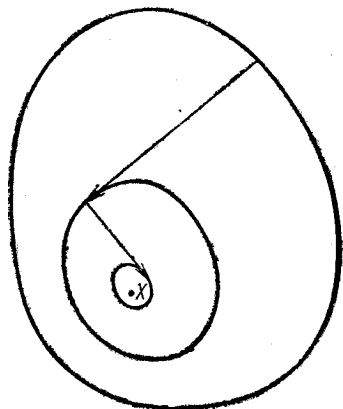


Рис. 7.3.2

В случае метода наискорейшего спуска вопрос об объеме вычислений при минимизации вспомогательных функций одной переменной должен решаться также с учетом относительной трудоемкости вычисления значений функции  $\Phi(x)$  и ее градиента.

Для иллюстрации решения вопроса о выборе метода рассмотрим следующую типичную задачу: решается нелинейная краевая задача для системы обыкновенных дифференциальных уравнений. В гл. IX будет показано, что эта задача сводится к решению нелинейной системы уравнений  $F(x) = 0$ :

$$f_i(x) = f_i(x_1, \dots, x_m) = 0, \quad i = 1, \dots, m,$$

обладающей следующими свойствами.

Количество операций при отыскании одного значения  $f_i(x)$  и при одновременном отыскании всех значений  $f_i(x)$ ,  $i = 1, \dots, m$ , в той же точке одинаково; обозначим его через  $A$ . Количество операций при непосредственном отыскании значений  $\partial f_i(x)/\partial x_j$  и при одновременном отыскании всех значений  $\partial f_i(x)/\partial x_j$ ,  $i = 1, \dots, m$ , в той же точке одинаково. Обозначим его через  $B$ ; обычно  $B \gg A$ .

Тогда при решении задачи методом Ньютона целесообразно вычислять производные  $\partial f_i(x)/\partial x_j$ ,  $j = 1, \dots, m$ , пользуясь приближенной формулой

$$\begin{aligned} \frac{\partial f_i(x_1, \dots, x_m)}{\partial x_j} &\approx \\ &\approx \frac{f_i(x_1, \dots, x_{j-1}, x_j + \Delta, x_{j+1}, \dots, x_m) - f_i(x_1, \dots, x_{j-1}, x_j, \dots, x_m)}{\Delta}. \end{aligned} \quad (1)$$

Область сходимости метода Ньютона обычно невелика, поэтому по крайней мере на начальном этапе итераций целесообразно свести решение этой задачи к минимизации некоторого

функционала и применить какой-либо из методов спуска. Рассмотрим простейший случай функционала

$$\Phi(x) = \sum_{i=1}^m (\lambda_i f_i(x))^2;$$

множители  $\lambda_i = \text{const} \neq 0$ , называемые *масштабными*, подбирают из условий конкретной задачи.

Пусть  $x^n = (x_1^n, \dots, x_m^n)$  — полученное приближение и решено сделать спуск в направлении  $\Delta = (\Delta_1, \dots, \Delta_m)$ ,  $\|\Delta\| = 1$ . Вычисляют приближенные значения производных в этом направлении:

$$l_i = (f_i(x^n + \varepsilon\Delta) - f_i(x^n))/\varepsilon. \quad (2)$$

На прямой  $x = x^n + t\Delta$  имеем

$$f_i(x) \approx f_i(x^n) + tl_i,$$

поэтому следующее приближение  $x^{n+1} = x^n + t\Delta$  определяют из условия

$$\min_t \sum_{i=1}^m (\lambda_i (f_i(x^n) + tl_i))^2.$$

Отметим, что в данном случае существует вопрос о разумном выборе  $\varepsilon$  в формуле (2). (По этому поводу см. § 2.16.)

Задача минимизации функции при наличии ограничений формулируется следующим образом. Ищется величина

$$A = \inf_{x_1, \dots, x_m} \Phi(x_1, \dots, x_m) \quad (3)$$

при условиях

$$\varphi_i(x_1, \dots, x_m) \geq 0, \quad i = 1, \dots, l, \quad (4)$$

$$\psi_i(x_1, \dots, x_m) = 0, \quad i = 1, \dots, q. \quad (5)$$

При решении этой задачи возникают дополнительные трудности по сравнению с решением задачи отыскания безусловного минимума, в которой ищется  $\inf_{(x_1, \dots, x_m) \in \mathbb{R}_m} \Phi(x_1, \dots, x_m)$  — нижняя грань  $\Phi(x_1, \dots, x_m)$  по всему пространству  $\mathbb{R}_m$ .

Непосредственное использование многих из описанных выше методов становится невозможным, и возникает необходимость в их модификации.

В то же время задача минимизации функции при ограничениях типа (4), (5) является весьма актуальной для приложений. Например, существует целый раздел математики — *линейное программирование*, — занимающийся решением задачи (1) — (3) в случае, когда  $\Phi$ ,  $\varphi_i$ ,  $\psi_i$  — линейные функции аргументов  $x_j$ .

Среди других методов, связанных с решением задачи (3) — (5), упомянем *метод штрафа*. Строится последовательность

функций  $\Phi_\lambda(x_1, \dots, x_m)$ , удовлетворяющая следующим условиям:

- 1)  $\Phi_\lambda(x_1, \dots, x_m)$  определена при всех  $(x_1, \dots, x_m)$ ;
- 2)  $\inf_{R_m} \Phi_\lambda(x_1, \dots, x_m) = A_\lambda \rightarrow A$  при  $\lambda \rightarrow \infty$ ;
- 3) если существуют точки  $(\bar{x}_1, \dots, \bar{x}_m)$  и  $(x_1^\lambda, \dots, x_m^\lambda)$  такие, что

$$\Phi(\bar{x}_1, \dots, \bar{x}_m) = A, \quad \Phi_\lambda(x_1^\lambda, \dots, x_m^\lambda) = A_\lambda,$$

то  $(x_1^\lambda, \dots, x_m^\lambda) \rightarrow (\bar{x}_1, \dots, \bar{x}_m)$  при  $\lambda \rightarrow \infty$ .

Вместо решения исходной задачи (3) — (5) решается задача отыскания  $\inf_{R_m} \Phi_\lambda(x_1, \dots, x_m)$  при достаточно больших  $\lambda$ .

Часто вместо первого условия на функцию  $\Phi_\lambda$  требуют выполнения более слабого условия. Функция  $\Phi_\lambda$  определена в точках некоторой односвязной области  $G_\lambda$ ,  $\Phi_\lambda(x_1, \dots, x_m) \rightarrow \infty$  при приближении точки  $(x_1, \dots, x_m)$  к границе области или при  $x_1^2 + \dots + x_m^2 \rightarrow \infty$  (в случае, когда область  $G_\lambda$  неограниченная).

В отдельных случаях условие 3) также несколько модифицируется.

Приведем пример построения такой функции  $\Phi_\lambda$ . Для этого соотношения (4), (5) записываются в таком виде, что  $\varphi_i, \psi_i$  будут определены при всех  $(x_1, \dots, x_m)$  (или при всех  $x_1, \dots, x_m$  из  $G_\lambda$ ).

Вводится некоторая невозрастающая функция  $h(t)$ , определенная при  $-\infty < t < \infty$ , такая, что  $\lim_{t \rightarrow -\infty} h(t) > 0, \lim_{t \rightarrow +\infty} h(t) = 0$ .

Например, можно взять

$$h(t) = \frac{1}{2} - \frac{1}{\pi} \operatorname{arctg} t.$$

В качестве  $\Phi_\lambda(x_1, \dots, x_m)$  берется функция

$$\Phi(x_1, \dots, x_s) + \lambda \sum_{i=1}^m \psi_i^2(x_1, \dots, x_m) + \lambda \sum_{i=1}^q h(\lambda \varphi_i(x_1, \dots, x_m)).$$

Наличие слагаемого  $\lambda \psi_i^2(x_1, \dots, x_m)$  заставляет смещаться точку экстремума  $(x_1^\lambda, \dots, x_m^\lambda)$  в область, где  $\psi_i(x_1, \dots, x_m) = 0$ ; наличие слагаемого  $\lambda h(\lambda \varphi_i(x_1, \dots, x_m))$  заставляет смещаться точку экстремума  $(x_1^\lambda, \dots, x_m^\lambda)$  в область, где  $\varphi_i(x_1, \dots, x_m) \geq 0$ .

Мы придерживаемся точки зрения, что *всякий универсальный метод решения задач минимизации должен обладать существенными недостатками*. Каким недостатком обладает метод штрафа? Оказывается, что при больших  $\lambda$  структура линий уровня  $\Phi_\lambda$ , как правило, такова, что сходимость методов минимизации существенно замедляется.



ния  $\min_{x_{m-1}} \Phi_{m-1}(x_{m-1}, x_m)$  при  $s$  значениях параметра  $x_m$ , т. е.  $s^2$  вычислений значений функции  $\Phi_{m-1}(x_{m-1}, x_m)$ . Это в свою очередь потребует вычисления  $s^3$  значений  $\Phi_{m-2}(x_{m-2}, x_{m-1}, x_m)$  и т. д. В конечном счете потребуются вычисления  $s^m$  значений функции  $\Phi$ . Уже при умеренных  $s$  и  $m$ , например  $s = 10$ ,  $m = 10$ , такой объем вычислений окажется недопустимо большим. Однако при малых  $m$  некоторые модификации рассматриваемой идеи оказываются полезными. Например, возможен такой вариант. Задаются начальным приближением  $(x_1^0, \dots, x_m^0)$ . Реализуют указанный алгоритм при не очень большом значении  $s$ , например  $s = 3$  или  $s = 4$ . При этом значения функции вычисляются в точках некоторого параллелепипеда  $|x_i - x_i^0| \leq \Delta_i^0$ . Получаемую точку минимума  $(x_1^1, \dots, x_m^1)$  принимают за следующее приближение; приближение  $(x_1^2, \dots, x_m^2)$  находят аналогичным образом, но значения функции  $\Phi$  вычисляются в точках параллелепипеда  $|x_i - x_i^1| \leq \Delta_i^1$  и т. д.

Рассмотрим еще один метод, общая структура которого похожа на структуру описанного выше.

Если линии уровня функции  $\Phi$  похожи на сферы, то при применении методов спуска происходит быстрое смещение в направлении минимума (рис. 7.4.1). Однако практически более типичен случай, когда эти линии похожи на эллипсоиды с большим разбросом

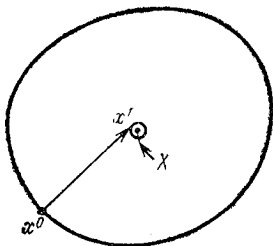


Рис. 7.4.1

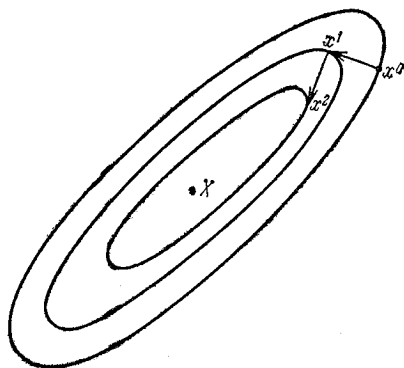


Рис. 7.4.2

полуосей. Тогда при движении по градиенту смещение в направлении точки минимума будет довольно медленным (рис. 7.4.2).

Предположим, что оси этих «эллипсоидов» естественным образом разбиваются на две группы: первая группа состоит из  $m_1$  осей одного порядка и относительно малых, вторая группа состоит из  $m_2$  осей одного порядка и относительно больших.

При решении некоторых задач такого рода хорошо зарекомендовал себя следующий метод, называемый *методом оврагов*. Задаются какими-то приближениями  $x^0$  и  $x^1$  и производят несколько шагов метода спуска, исходя из каждого из этих при-



ближений. Будут получены приближения  $X^0$  и  $X^1$ . Из рис. 7.4.3 видно, что эти приближения будут лежать в многообразии, расположенном в окрестности осей второй группы. Процесс итераций состоит в получении последовательных приближений  $X^0, X^1, \dots$ , лежащих в окрестности этого многообразия. В случае  $m_2 = 1$  приближение  $X^{l+1}$  отыскивается следующим образом. Проведем через  $X^{l-1}$  и  $X^l$  прямую и найдем приближение

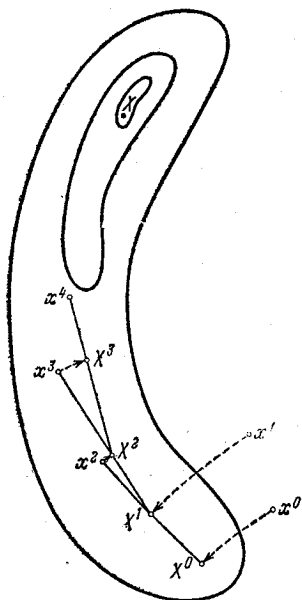


Рис. 7.4.3

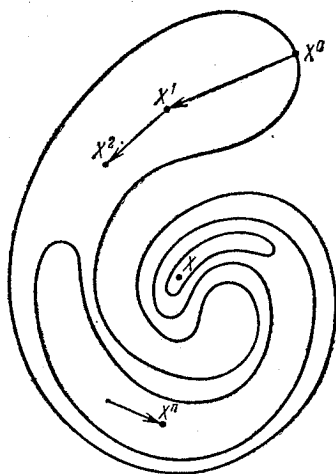


Рис. 7.4.4

$x^{l+1}$  к точке минимума  $\Phi(x)$  на этой прямой. Таким образом, это приближение ищется в виде

$$x^{l+1} = X^l + \alpha(X^l - X^{l-1}).$$

Далее проводим несколько итераций исходя из  $x^{l+1}$  и получаем приближение  $X^{l+1}$ , также лежащее в овраге. В случае  $m_2 > 1$  приближение  $x^{l+1}$  иногда удобно отыскивать в виде

$$x^{l+1} = X^l + \alpha(X^l - X^{l-1}) + \beta \text{grad } \Phi(X^l).$$

Из рис. 7.4.4 видно, что описанный способ оказывается эффективным и в ряде случаев, когда линии уровня функции  $\Phi(X)$  имеют более сложную структуру.

Решение системы уравнений

$$f_i(x_1, \dots, x_m) = 0, \quad i = 1, \dots, m, \quad (2)$$

также формально сводится к последовательному решению уравнений с одним неизвестным. Рассмотрим систему уравнений

$$f_i(x_1, \dots, x_m) = 0, \quad i = 2, \dots, m, \quad (3)$$

относительно неизвестных  $x_2, \dots, x_m$ . Пусть  $x_2(x_1), \dots, x_m(x_1)$  — ее решение. Подставляя выражения  $x_2(x_1), \dots, x_m(x_1)$  в первое из уравнений (2), получим уравнение

$$F_1(x_1) = f_1(x_1, x_2(x_1), \dots, x_m(x_1)) = 0 \quad (4)$$

относительно одной неизвестной  $x_1$ .

Отыскание значений  $x_2(x_1), \dots, x_m(x_1)$  и решение уравнения (4) можно проводить численно. Выбирается какой-то метод решения (4) по значениям функции  $F(x_1)$ ; при каждом требуемом значении  $x_1$  в результате решения (3) получаются значения  $x_2(x_1), \dots, x_m(x_1)$ , которые подставляются затем в правую часть (4). Для решения системы уравнений (3) при каждом значении  $x_1$  применим тот же прием.

Пусть  $x_3(x_1, x_2), \dots, x_m(x_1, x_2)$  — решение системы уравнений

$$f_i(x_1, \dots, x_m) = 0, \quad i = 3, \dots, m, \quad (5)$$

относительно неизвестных  $x_3, \dots, x_m$ . Подставляя  $x_3(x_1, x_2), \dots, x_m(x_1, x_2)$  во второе уравнение системы, получим уравнение

$$F_2(x_1, x_2) = f_2(x_1, x_2, x_3(x_1, x_2), \dots, x_m(x_1, x_2)) = 0.$$

При каждом  $x_1$  это уравнение может быть разрешено относительно  $x_2$ . Его решение  $x_2(x_1)$ , а также  $x_3(x_1, x_2(x_1)), \dots, x_m(x_1, x_2(x_1))$  образуют решение системы (3). Систему (5) при каждом  $x_1, x_2$  опять решаем, сводя к системе, где число неизвестных на единицу меньше, и т. д.

Если для решения каждого вспомогательного уравнения с одной неизвестной потребуется  $s$  вычислений функций, то суммарно этот алгоритм потребует порядка  $s^m$  вычислений правых частей уравнений системы.

По поводу реального применения этого алгоритма можно сказать все то же, что и по поводу применения описанного в начале параграфа метода минимизации.

**Задача 1.** Рассмотреть, во что переходит описанный метод в случае, когда система уравнений (2) линейная.

## § 5. Решение стационарных задач путем установления

Распространенным методом решения стационарных задач является *метод установления*. В этом случае для решения стационарной задачи строится нестационарный процесс, решение которого с течением времени оказывается независимым от него и устанавливается к решению исходной стационарной задачи. Рассмотрим систему дифференциальных уравнений

$$\frac{dx}{dt} + \text{grad } \Phi(x) = 0. \quad (1)$$

Вектор  $dx/dt$  пропорционален градиенту функции  $\Phi(x)$ , т. е. ортогонален ее линиям уровня и направлен в сторону убывания

значений функции  $\Phi(x)$ . Таким образом, при перемещении вдоль траектории системы (1) значение  $\Phi(x)$  не возрастает. Формально справедливость этого утверждения следует из неравенства

$$\frac{d\Phi(x)}{dt} = \left( \text{grad } \Phi(x), \frac{dx}{dt} \right) = -(\text{grad } \Phi(x), \text{grad } \Phi(x)), \quad (2)$$

означающего, что  $d\Phi(x)/dt < 0$  всюду, за исключением стационарных точек функции  $\Phi(x)$ .

Другой нестационарный процесс, решение которого при весьма общих предположениях устанавливается к точке минимума функции  $\Phi(x)$ , описывается системой дифференциальных уравнений

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} + \text{grad } \Phi(x) = 0, \quad \gamma > 0. \quad (3)$$

Для решения этой системы имеем

$$\begin{aligned} \frac{d}{dt} \left( \frac{1}{2} \left( \frac{dx}{dt}, \frac{dx}{dt} \right) + \Phi(x) \right) &= \left( \frac{dx}{dt}, \frac{d^2x}{dt^2} \right) + \\ &+ \left( \text{grad } \Phi(x), \frac{dx}{dt} \right) = -\gamma \left( \frac{dx}{dt}, \frac{dx}{dt} \right) < 0, \quad (4) \end{aligned}$$

если только  $dx/dt \neq 0$ . Функцию  $\Phi(x)$  в первом случае и  $\frac{1}{2} \left( \frac{dx}{dt}, \frac{dx}{dt} \right) + \Phi(x)$  — во втором можно рассматривать как энергию материальной системы, движение которой описывается системами уравнений (1) и (3). Соотношения (2) и (4) показывают, что рассматриваемые нестационарные процессы характеризуются оттоком или, как говорят, *диссипацией* энергии.

Чтобы прояснить вопрос о разумном выборе  $\gamma$ , рассмотрим простейшую модель:  $x$  — скаляр,  $\Phi(x) = \frac{a^2 x^2}{2}$ ; тогда (3) приобретает вид

$$x'' + \gamma x' + a^2 x = 0.$$

Соответствующее характеристическое уравнение

$$\lambda^2 + \gamma \lambda + a^2 = 0,$$

его корни  $\lambda_{1,2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - a^2}$ , и при  $\lambda_1 \neq \lambda_2$ , т. е. при  $\gamma \neq 2a$ , общее решение есть  $c_1 \exp \{\lambda_1 t\} + c_2 \exp \{\lambda_2 t\}$ .

Скорость убывания решений рассматриваемого уравнения определяется величиной

$$\sigma(\gamma) = \max(\text{Re } \lambda_1, \text{Re } \lambda_2).$$

При  $\gamma \leq 2a$  имеем  $\gamma^2/4 - a^2 \leq 0$ , и поэтому

$$\text{Re } \lambda_1 = \text{Re } \lambda_2 = \sigma(\gamma) = -\gamma/2 \geq -a.$$

при  $\gamma > 2a$  величины  $\lambda_1$  и  $\lambda_2$  вещественны и

$$\operatorname{Re} \lambda_1 = -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2} > \operatorname{Re} \lambda_2 = -\frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} - a^2}.$$

Тогда

$$\begin{aligned} \sigma(\gamma) = \operatorname{Re} \lambda_1 &= -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2} = \\ &= -\frac{a^2}{\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2}} > -\frac{a^2}{\frac{\gamma}{2}} > -a. \end{aligned}$$

Таким образом, график  $\sigma(\gamma)$  имеет вид, изображенный на рис. 7.5.1, и  $\min_{\gamma} \sigma(\gamma) = \sigma(2a) = -a$ . Из результатов рассмотре-

ний этой модельной задачи можно сделать следующие качественные выводы.

1. Если коэффициент трения  $\gamma$  очень мал (в нашем случае  $\gamma \ll 2a$ ), то решение системы (3) медленно устанавливается к положению равновесия; при этом

(вследствие условия  $\operatorname{Im} \lambda_1, \operatorname{Im} \lambda_2 \neq 0$ ) происходят колебания около положения равновесия.

2. Если  $\gamma$  велико ( $\gamma \gg 2a$ ), то решение также медленно устанавливается; причина состоит в том, что при большом коэффициенте трения  $\gamma$  движение не может приобрести большой скорости.

3. Оптимальное значение  $\gamma$  лежит где-то посередине и зависит от свойств конкретной функции  $\Phi(x)$ .

Метод установления с помощью решения системы (3) иногда называют *методом тяжелого шарика*. Это название обусловлено следующими соображениями.

Рассмотрим движение материальной точки по поверхности  $y = \Phi(x)$  в поле тяжести, направленном в отрицательном направлении оси  $y$ . Предположим, что трение пропорционально скорости и точка не может отрываться от поверхности. Тогда ее движение опишется системой уравнений

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} + \frac{\operatorname{grad} \Phi(x)}{1 + \|\operatorname{grad} \Phi(x)\|^2} = 0.$$

Ясно, что решение этой системы с течением времени установится к некоторой стационарной точке функции  $\Phi(x)$ . Вблизи экстремума  $\|\operatorname{grad} \Phi(x)\| \ll 1$  и эта система близка к системе (3).

Большинство известных методов установления описывается уравнениями вида

$$A_0\left(x, \frac{dx}{dt}\right) \frac{dx}{dt} + A_1\left(x, \frac{dx}{dt}, \operatorname{grad} \Phi(x)\right) = 0 \quad (5)$$

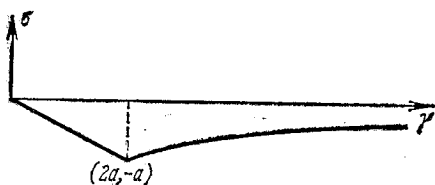


Рис. 7.5.1

или

$$B_0\left(x, \frac{dx}{dt}\right) \frac{d^2x}{dt^2} + B_1\left(x, \frac{dx}{dt}, \text{grad } \Phi(x)\right) = 0, \quad (6)$$

где

$$A_0(X, 0) \neq 0, \quad A_1(X, 0, 0) = 0,$$

$$B_0(X, 0) \neq 0, \quad B_1(X, 0, 0) = 0$$

и выполнены условия диссипативности, обеспечивающие сходимость к точке экстремума  $X$ . Вообще говоря, можно обратить операторы  $A_0$  и  $B_0$  и преобразовать эти уравнения к виду, где  $A_0$  и  $B_0$  — тождественные операторы. Однако исходная форма записи часто практически удобнее.

Может показаться, что построение таких нестационарных процессов, устанавливающихся к решению, уже полностью решает проблему отыскания минимума функции. Осталось «лишь» найти решение получившейся системы дифференциальных уравнений, используя какой-либо из численных методов решения задачи Коши.

В действительности сведение решения стационарной задачи к решению нестационарной не всегда дает удовлетворительное решение проблемы минимизации. Остается еще неясным существенный вопрос о выборе величины шагов численного интегрирования. Предположим, что решение нестационарной задачи устанавливается с требуемой точностью к решению стационарной за некоторый промежуток времени  $T$ . Если интегрирование производится с малым шагом  $\Delta$ , то получаемые расчетные точки будут близки к рассматриваемой траектории и можно рассчитывать на попадание в малую окрестность точки минимума. Однако при этом число шагов  $T/\Delta$  может оказаться недопустимо

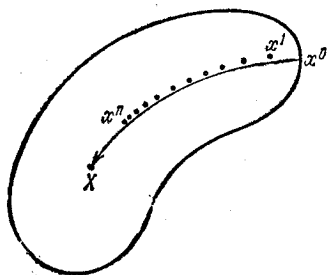


Рис. 7.5.2

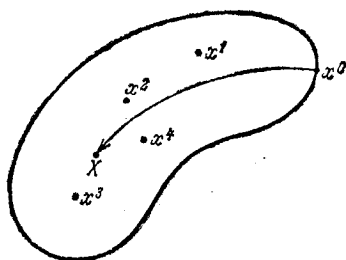


Рис. 7.5.3

большим (рис. 7.5.2). Если шаг интегрирования берется очень большим, то может случиться, что расчетные точки начнут сильно отклоняться от рассматриваемого решения и никогда не попадут в искомую окрестность точки минимума (рис. 7.5.3).

Метод установления применим не только к задачам на экстремум функционала, но и к любым стационарным задачам  $F(x) = 0$ . Строится некоторый процесс вида (5) или (6), где

вместо  $\text{grad } \Phi(x)$  стоит  $F(x)$ , и такой, что  $x(t) \rightarrow X$  при  $t \rightarrow \infty$ ,  $X$  — корень уравнения  $F(X) = 0$ .

Рассмотрим подробнее случай системы линейных уравнений  $Ax - b = 0$  в предположении, что жорданова форма матрицы диагональная и все ее собственные значения  $\lambda_i$  лежат в пределах  $0 < \mu \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m \leq M$ , где  $e_1, \dots, e_m$  — соответствующие собственные векторы, образующие полную систему.

Напишем простейшую аппроксимацию метода установления

$$\frac{dx}{dt} + (Ax - b) = 0 \quad (7)$$

на временной сетке с постоянным шагом

$$\frac{x^{n+1} - x^n}{\tau} + (Ax^n - b) = 0.$$

Соответствующая расчетная формула

$$x^{n+1} = x^n - \tau(Ax^n - b)$$

совпадает с расчетной формулой из § 6.3. Погрешности  $r^n = x^n - X$  удовлетворяют соотношению  $r^{n+1} = (E - \tau A)r^n$ . При  $r^0 = \sum_1^m c_i e_i$  имеем  $r^n = \sum_1^m c_i (1 - \tau \lambda_i)^n e_i$  и скорость убывания погрешности определяется величиной  $\max_{\lambda_i} |1 - \tau \lambda_i|$ . Там уже было

установлено, что наиболее целесообразно брать  $\tau = 2/(\mu + M)$ , и тогда можно утверждать, что погрешность ведет себя как  $O(((M - \mu)/(M + \mu))^n)$ . Метод наискорейшего спуска можно также рассматривать как аппроксимацию метода установления, но уже на сетке с переменным шагом:  $x^{n+1} = x^n - \tau_n(Ax^n - b)$ ; шаг  $\tau_n$  определяется каждый раз из условия  $\min_{\tau_n} \Phi(x^{n+1})$ .

Рассмотрим аппроксимацию метода установления

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} + (Ax - b) = 0$$

на временной сетке с постоянным шагом

$$\frac{x^{n+1} - 2x^n + x^{n-1}}{\tau^2} + \gamma \frac{x^{n+1} - x^{n-1}}{2\tau} + (Ax^n - b) = 0;$$

$0 < \gamma$  — скалярный множитель. Погрешности  $r^n$  удовлетворяют соотношениям

$$\frac{r^{n+1} - 2r^n + r^{n-1}}{\tau^2} + \gamma \frac{r^{n+1} - r^{n-1}}{2\tau} + Ar^n = 0. \quad (8)$$

Разложим векторы  $r^n$  по собственным векторам матрицы  $A$ :

$$r^n = \sum_{i=1}^m c_i^n e_i.$$

Подставим выражения  $r^{n-1}$ ,  $r^n$ ,  $r^{n+1}$  в (8); поскольку векторы  $e_i$  независимы, то коэффициенты при них обращаются в нуль и получается система соотношений

$$\frac{c_i^{n+1} - 2c_i^n + c_i^{n-1}}{\tau^2} + \gamma \frac{c_i^{n+1} - c_i^{n-1}}{2\tau} + \lambda_i c_i^n = 0.$$

Решения этих разностных уравнений записываются в виде

$$c_i^n = C_i^1 (z_1^i)^n + C_i^2 (z_2^i)^n,$$

если  $z_1^i$ ,  $z_2^i$  — простые корни характеристического уравнения

$$\frac{z^2 - 2z + 1}{\tau^2} + \gamma \frac{z^2 - 1}{2\tau} + \lambda_i z = 0, \quad (9)$$

и в виде

$$c_i^n = C_i^1 (z_1^i)^n + C_i^2 n (z_1^i)^n,$$

если корень (9) кратный. Во всех случаях определяющим фактором убывания  $\|r^n\|$  является величина  $\max_i (\max(|z_1^i|, |z_2^i|))$ , которую можно мажорировать сверху величиной

$$\max_{\mu \leq \lambda \leq M} |z(\lambda)|, \quad (10)$$

где  $z(\lambda)$  — максимальный по модулю корень уравнения

$$\frac{z^2 - 2z + 1}{\tau^2} + \gamma \frac{z^2 - 1}{2\tau} + \lambda z = 0.$$

Мы не будем приводить полного решения задачи на экстремум (10), а ограничимся наводящими соображениями и выписыванием ответа.

При  $x^1 = x^0 + \alpha(Ax^0 - b)$  приближения  $x^n$  записываются в виде (6.6.2). Поэтому рассматриваемый итерационный процесс не может дать лучшего приближения, чем оптимальный линейный итерационный процесс (6.6.19). При  $n \rightarrow \infty$  оптимальный линейный итерационный процесс переходит в итерационный процесс (6.6.23), который с учетом явного выражения  $\lambda_0 = (1 + \sqrt{\mu/M}) / (1 - \sqrt{\mu/M})$  может быть записан в виде

$$x^{n+1} = x^n + \left( \frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}} \right)^2 (x^n - x^{n-1}) - \frac{4}{(\sqrt{M} + \sqrt{\mu})^2} (Ax^n - b). \quad (11)$$

Можно подобрать  $\tau$  и  $\gamma$  так, чтобы итерационный процесс совпадал с рассматриваемым. Для этого следует взять

$$\tau = \frac{2}{\sqrt{M} + \gamma}, \quad \gamma = \frac{2\sqrt{M\mu}}{\sqrt{M} + \mu}.$$

Таким образом, в рамках схемы установления с постоянным  $\gamma$  можно получить итерационный процесс, несущественно отличающийся от оптимального линейного итерационного процесса.

Обратим внимание на поведение корней  $z(\lambda)$ , соответствующих (11); имеем характеристическое уравнение

$$z^2 - \left( \left( 1 + \left( \frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}} \right)^2 \right) - \frac{4}{(\sqrt{M} + \sqrt{\mu})^2} \lambda \right) z + \left( \frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}} \right)^2 = 0.$$

Запишем это уравнение в виде

$$z^2 - A(\lambda)z + \mu_0^2 = 0, \quad \mu_0 = (\sqrt{M} - \sqrt{\mu})/(\sqrt{M} + \sqrt{\mu}).$$

Здесь  $A(\lambda)$  — линейная функция от  $\lambda$ , причем  $A(M) = -2\mu_0$ ,  $A(\mu) = 2\mu_0$ . Следовательно,  $|A(\lambda)| < 2\mu_0$  при  $\mu < \lambda < M$ .

Поэтому  $z_{1,2}(\mu) = \mu_0$ ,  $z_{1,2}(M) = -\mu_0$ ,  $z_{1,2}(\lambda) = -\frac{A(\lambda)}{2} \pm \pm \sqrt{\left( \frac{A(\lambda)}{2} \right)^2 - \mu_0^2}$  имеют ненулевую мнимую часть и  $||z_{1,2}(\mu)|| = \mu_0$  при  $\mu < \lambda < M$ .

Таким образом, указана совокупность коэффициента  $\gamma$  и шага  $\tau$ , для которой

$$\max_{\mu < \lambda < M} |z(\lambda)| = (\sqrt{M} - \sqrt{\mu})/(\sqrt{M} + \sqrt{\mu}).$$

Неулучшаемость этой оценки усматривается из оценки скорости сходимости оптимального линейного итерационного процесса.

Как аналог метода сопряженных градиентов в случае минимизации функционала  $\Phi(x)$  общего вида можно рассматривать следующий метод: последующие приближения ищутся в виде

$$x^{n+1} = x^n + \alpha_n(x^n - x^{n-1}) + \beta_n \text{grad } \Phi(x^n),$$

$\alpha_n$  и  $\beta_n$  определяются из условия  $\min_{\alpha_n, \beta_n} \Phi(x^n)$ .

Соответствие между методами решения стационарных задач путем установления и обычными итерационными методами позволяет строить новые итерационные методы или новые процессы установления.

Рассмотрим, например, расчетные формулы Ньютона

$$x^{n+1} - x^n = -(F'(x^n))^{-1} \dot{F}(x^n) \quad (12)$$

решения системы нелинейных уравнений  $F(x) = 0$ . Этим формулам можно дать следующую интерпретацию: введем непрерывное время  $t$  и будем рассматривать величины  $x^n$  как значения некоторой функции в моменты времени  $t_n = n$ . Тогда соотношение (12), переписанное в виде

$$\frac{x^{n+1} - x^n}{t_{n+1} - t_n} = -(F'(x^n))^{-1} F(x^n),$$



можно интерпретировать как получившееся при аппроксимации системы

$$\frac{dx}{dt} = -(F'(x))^{-1} F(x). \quad (13)$$

Решение исходной задачи является стационарной точкой этой системы. При  $x = X + \eta$  имеем

$$\begin{aligned} (F'(x))^{-1} &= (F'(X))^{-1} + O(\|\eta\|), \\ F(x) &= F(X) + F'(X)\eta + O(\|\eta\|^2) = F'(X)\eta + O(\|\eta\|^2). \end{aligned}$$

Таким образом,

$$\begin{aligned} \frac{d\eta}{dt} = \frac{dx}{dt} &= -((F'(X))^{-1} + O(\|\eta\|))(F'(X)\eta + \\ &+ O(\|\eta\|^2)) = -\eta + O(\|\eta\|^2). \end{aligned}$$

Отсюда следует, что  $x = X$  является асимптотически устойчивым решением (13).

Заменяя производную  $dx/dt$  разностным отношением по значениям функции в некоторых точках  $t_0 = 0, t_1, \dots$ , получаем соотношения

$$\frac{x^{n+1} - x^n}{\Delta_n} = -(F'(x^n))^{-1} F(x^n), \quad \Delta_n = t_{n+1} - t_n,$$

иначе,

$$x^{n+1} = x^n - \Delta_n (F'(x^n))^{-1} F(x^n). \quad (14)$$

Таким образом, построение нестационарного процесса, соответствующего методу Ньютона, привело нас к получению итерационного процесса (14) более общего вида.

Рассмотренный пример показывает, что переход от некоторого итерационного алгоритма к соответствующему ему нестационарному процессу имеет много общих черт с построением замыкания вычислительного алгоритма (понятие замыкания алгоритма будет введено в гл. IX). Нами была доказана сходимость метода Ньютона лишь при достаточно хорошем начальном приближении к решению. С целью расширения области сходимости иногда прибегают к следующей модификации метода Ньютона.

Задаются функционалом  $\Phi(x)$ , например  $\|F(x)\|^2$ , нижняя грань которого, равная нулю, достигается на решении задачи. Последовательные приближения отыскиваются в виде (14), причем  $\Delta_n$  определяется из условия

$$\min_{\Delta_n} \Phi(x^n - \Delta_n (F'(x^n))^{-1} F(x^n)).$$

Возникает вопрос о практическом нахождении величины  $\Delta_n$ , при которой достигается этот минимум.

Одна из распространенных процедур определения  $\Delta_n$  состоит в следующем. Задавая  $\lambda \in (0, 1)$ ,  $\theta \in (0, 1]$  и  $l > 0$  — целое, при каждом  $n$  последовательно вычисляют

$$x^{n+1, i} = x^n - \lambda^i (F'(x^n))^{-1} F(x^n), \quad i = 0, \dots, l,$$

и  $q^{n+1, i} = \Phi(x^{n+1, i})$ . Если при некотором  $k \leq l$  оказалось  $q^{n+1, k} \leq \theta \Phi(x^n)$ , то вычисления прекращают и полагают  $x^{n+1} = x^{n+1, k}$ ; в других процедурах находят  $\min_{0 < i \leq l} q^{n+1, i} = q^{n+1, m}$  и полагают  $x^{n+1} = x^{n+1, m}$ .

Если

$$q^{n+1, 0}, \dots, q^{n+1, l} > \theta \Phi(x^n),$$

то возможны, например, такие варианты:

- 1) временный переход к другому методу;
- 2) остановка;
- 3) изменение значений параметров  $l, \lambda, \theta$ .

Отметим в заключение, что методы установления могут применяться и в случае минимизации функции в областях с ограничениями. Тогда уравнения (1), (3) следует дополнить какими-то уравнениями, которым будет подчиняться траектория точки, попавшей на границу.

## § 6. Что оптимизировать?

Выше был рассмотрен ряд способов минимизации функций и решения систем нелинейных уравнений. Нами охвачена лишь небольшая часть известных методов решения этих задач. Вместо дальнейшего изучения методов обратим внимание на другие, не менее важные вопросы.

Как уже упоминалось, к минимизации функционалов (функций) сводятся многие задачи управления отраслями промышленности, сельского хозяйства, транспорта, распределения ресурсов и других областей жизни общества. В этих случаях задачи минимизации принято называть *задачами оптимизации*, поскольку основной целью решения этих задач обычно является достижение какого-то наилучшего, оптимального, режима работы. При этом минимизируемую функцию обычно называют *целевой функцией*.

Решение задач оптимизации складывается из следующих элементов: создание математической модели явления, определение целевой функции и важнейших параметров, подлежащих оптимизации, непосредственная минимизация некоторой функции обычно большого числа переменных, внедрение результатов исследования.

Естественно, что рассмотрение первых двух и последнего вопросов должно проводиться математиками совместно со специалистами конкретной отрасли.

Рассмотрим один реальный пример выбора модели. Производство сахара складывается из следующих этапов: выращивание сахарной свеклы, её транс-

портировка и переработка на заводах. Математики, поставившие перед собой задачу оптимизировать производство сахарной свеклы, в первую очередь решили заняться минимизацией транспортных расходов. Эта задача была успешно решена с помощью ЭВМ, но хозяйственные руководители отказались принять полученный план перевозок к действию. Основой для отказа послужил довод о том, что при оптимизации транспортировки свеклы несколько ухудшались условия транспортировки других сельскохозяйственных культур, а также усложнялась организация переработки: некоторые заводы оказывались перегруженными, другими недогруженными.

По-видимому, за этими возражениями стояли следующие соображения. Себестоимость сахара складывается из следующих составляющих (числа условные):

80 % — стоимость сахарной свеклы,

5 % — транспортные расходы,

15 % — расходы по переработке.

Ожидаемая экономия от минимизации транспортных расходов составляет 10 % этих расходов, т. е. всего 0,5 % себестоимости сахара. В то же время эта перестройка требует больших хлопот и неизвестно, не произойдет ли более существенных потерь за счет каких-то неучтенных моментов.

Из описанной выше структуры себестоимости видно, что за счет оптимизации расходов по транспортировке и переработке трудно достигнуть существенного повышения эффективности. Внимательный анализ проблемы позволил обнаружить следующее обстоятельство. Выход сахара из тонны свеклы пропорционален ее сахаристости: сахаристость изменяется по определенному закону в процессе созревания и падает в процессе хранения. В частности, при позднем начале уборки происходят большие потери из-за того, что на сахарные заводы поступает одновременно большое количество свеклы и она там долго хранится. В результате этого анализа была поставлена новая целевая функция — оптимизация выхода сахара (при заданном оборудовании) и построена математическая модель, учитывающая изменение сахаристости в процессе уборки и хранения свеклы.

Дальнейший анализ показал, что эта модель обладает следующим дефектом. При оптимальном (согласно этой модели) графике работы копка должна начинаться, когда свекла еще находится в периоде интенсивного роста, что экономически невыгодно для непосредственных производителей. Для преодоления этого противоречия нужно, чтобы модель учитывала заинтересованность производителей. Результатом решения оптимизационной задачи должен быть не только производственный план, но и рекомендации по наилучшей организации оплаты труда.

Точно такой же вопрос «Что оптимизировать?» возникает и при решении математических задач. Типичная задача этого класса: функция задана сложным аналитическим выражением или таблицей; требуется получить ее приближение с заданной точностью при помощи выражения наиболее простого вида.

Рассмотрим задачу получения такого приближения для функций

$$f(q) = 1 - \frac{8}{\pi^2} \sum_{k=1}^{\infty} \frac{\exp \left\{ -\frac{2k-1}{2} \pi q \right\}}{(2k-1)^2} \quad \text{при } 0 \leq q \leq \infty.$$

Для удобства сделаем замену переменных  $x = \exp \{-\pi q/2\}$ , т. е. рассмотрим задачу приближения функции

$$g(x) = 1 - \frac{8}{\pi^2} \sum_{k=1}^{\infty} \frac{x^{2k-1}}{(2k-1)^2} \quad \text{на } [0, 1].$$

Попытки приближения многочленами дали неудовлетворительные результаты; анализ показал, что причиной является неограниченность производной  $g'(x)$  в окрестности точки  $x = 1$ . После исследования особенности производной в окрестности точки  $x = 1$  было принято решение отыскивать приближение в виде

$$h(x) = (1-x)(a_0 + a_1x + a_2x^2) \ln \frac{1+x}{1-x} + x(a_3 + a_4x) + 1.$$

За счет выбора параметров  $a_i$  удалось получить приближение с точностью  $10^{-4}$ .

При других попытках приближения оказалось, что  $g(x)$  также может быть приближена при  $x \geq 10^{-4}$  с точностью  $5 \cdot 10^{-4}$  выражением вида

$$\frac{(1-x)(1+x(a_1+a_2x))}{1+a_3x+a_4x^2+a_5x^3}.$$

После выбора класса объектов, в котором отыскивается решение задачи, возникает проблема разумной параметризации этого класса. Например, многочлен

$$Q(x) = \sum_{i=0}^n a_i x^i,$$

приближающий функцию  $f(x)$  на отрезке  $[-1, 1]$ , можно записать также в виде

$$Q(x) = \frac{b_0}{\sqrt{2}} + \sum_{i=1}^n b_i T_i(x).$$

Как отмечалось, первая запись нежелательна вследствие возможной большой вычислительной погрешности. Кроме этого, вторая форма записи имеет следующее преимущество. Пусть приближающий многочлен отыскивается из условия

$$\min_Q \Phi, \quad \text{где} \quad \Phi = \int_{-1}^1 \frac{(f(x) - Q(x))^2}{\sqrt{1-x^2}} dx.$$

При первой форме записи многочлена поверхностями уровня  $\Phi(a_0, \dots, a_n) = \text{const}$  являются эллипсоиды с большим разбросом осей, поэтому итерационные методы минимизации этой функции обладают малой скоростью сходимости. При второй форме записи многочлена имеем

$$\begin{aligned} \Phi(b_0, \dots, b_n) = & \int_{-1}^1 \frac{(f(x))^2}{\sqrt{1-x^2}} dx - \sqrt{2} b_0 \int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx - \\ & - 2 \sum_{i=1}^n b_i \int_{-1}^1 \frac{f(x) T_i(x)}{\sqrt{1-x^2}} dx + \frac{\pi}{2} \sum_{i=0}^n b_i^2; \end{aligned}$$

поверхностями уровня функции  $\Phi(b_0, \dots, b_n)$  являются сферы и итерационные способы минимизации обеспечивают быструю сходимость к точке минимума (здесь эта точка известна).

Удачный выбор параметров и критериев их определения особенно важен при обработке результатов наблюдений.

Как одномерная модель движения снежной лавины может рассматриваться система дифференциальных уравнений

$$\frac{\partial h}{\partial t} + \frac{\partial (hv)}{\partial s} = 0, \quad \frac{\partial v}{\partial t} + v \frac{\partial v}{\partial s} + \frac{g}{2h} \frac{\partial (h^2 \cos \psi)}{\partial s} = \\ = g (\sin \psi - \mu \cos \psi) - k \frac{v^2}{h}. \quad (1)$$

Граничные условия на переднем фронте лавины имеют вид

$$h(\omega - v) = h_0 \omega, \quad h_0 \omega v = \frac{1}{2} g h^2 \cos \psi - \sigma h;$$

здесь  $h$  — высота снега,  $v$  — его скорость,  $\psi$  — угол наклона склона,  $\omega$  — скорость переднего фронта, параметры  $\mu$ ,  $k$ ,  $\sigma$  неизвестны.

Была сделана попытка определения этих параметров из наблюдения за положением переднего фронта лавины, двигавшейся по склону с постоянным наклоном  $\psi = \text{const}$ . Параметры пытались определить из условия минимума функции

$$\Phi(\mu, k, \sigma) = \sum_i (s(\mu, k, \sigma, t_i) - s_i)^2,$$

где  $s_i$  — наблюдаемые положения переднего фронта в моменты  $t_i$ ,  $s(\mu, k, \sigma, t_i)$  — получаемые в результате численного интегрирования системы (1). При непосредственном применении большого числа методов минимизации не обнаруживалось никакой тенденции к установлению получаемых значений  $\mu$ ,  $k$ ,  $\sigma$ . Анализ задачи указал на следующую причину этого явления. Оказывается, что при  $\psi = \text{const}$  с большой точностью выполняется приближенное равенство  $s(\mu, k, \sigma, t) \approx \lambda t$ ; здесь  $\lambda = \left(1 + \frac{1}{\sqrt{\beta}}\right)^{3/2}$ ,  $\beta =$

$\frac{k}{\text{tg} \psi - \mu}$ . Поэтому функция  $\Phi(\mu, k, \sigma)$  с большой точностью близка к неко-

торой функции  $\Phi_0(\lambda)$  одной переменной  $\lambda$ . Таким образом, при минимизации функции  $\Phi$  по переменным  $\mu$ ,  $k$ ,  $\sigma$  можно рассчитывать на получение не всех значений параметров  $\mu$ ,  $k$ ,  $\sigma$ , соответствующих точке минимума, а на получение лишь значения  $\lambda$ , соответствующего точке минимума. Пусть  $\Phi_0(\lambda)$  достигается при  $\lambda = \lambda_0$ . Поверхность  $\lambda = \lambda_0$  является двумерной, в данном случае цилиндрической поверхностью в пространстве переменных  $\mu$ ,  $k$ ,  $\sigma$ . На этой поверхности функция  $\Phi(\mu, k, \sigma)$  оказывается примерно постоянной, и при ее численной минимизации происходит беспорядочное на вид перемещение точек  $(\mu, k, \sigma)$  вдоль этой поверхности. Таким образом, в данном случае истинная причина неудачи состояла не в несовершенстве методов минимизации, а в недостаточности имеющейся информации по отношению к поставленной цели.

Вернемся к замечаниям общего характера.

При построении модели задачи возникает желание создать подробную математическую модель, учтя многие детали задачи, а затем произвести полную оптимизацию проблемы за счет наилучшего выбора всех параметров. Этот путь чреват рядом неприятностей.

Первые трудности возникают при описании модели; стремление учесть слишком многое увеличивает возможность упустить из виду что-либо существенное и тем самым ухудшить модель. При оптимизации по очень большому числу параметров возникает задача минимизации функции большого числа переменных, численное решение которой встречает иногда непреодолимые трудности.

Предположим на мгновение, что такую минимизацию все-таки удалось осуществить. Возникает задача доведения информации до сведения заказчиков с целью реализации результатов математического анализа модели. Заказчик обычно принимает решение исходя из своих собственных качественных критериев, не вдаваясь в детали модели. В случае получения рекомендаций относительно слишком большого числа параметров он не всегда сможет сделать для себя вывод о разумности этих рекомендаций и с большой вероятностью вообще откажется от них.

Следовательно, на первых этапах оптимизации особенно важно построить простейшую модель, учитывающую лишь основные, определяющие, параметры. Это требование вызывается также следующим соображением. Поначалу заказчик часто испытывает недоверие к математикам, вторгающимся в новую для них область. Только после завоевания доверия за счет достижения положительного эффекта при оптимизации важнейших параметров имеет смысл учесть влияние второстепенных параметров. Следует также иметь в виду, что само построение уточненной модели явления обычно становится возможным лишь после совместного анализа заказчиком и исполнителем результатов обсчета упрощенной модели.

Очень важно также, чтобы результаты математических исследований при диалоге с нематематиками подавались в достаточно наглядной, привычной форме, по возможности с привлечением минимального математического аппарата. Неоправданное привлечение абстрактных понятий математики редко приносит реальную пользу.

## § 7. Как оптимизировать?

После выбора модели, целевой функции и параметризации задачи возникает задача минимизации функции обычно большого числа переменных в области, принадлежность к которой задается условием выполнения большого числа ограничений — равенств или неравенств. Наличие ограничений существенно увеличивает сложность задачи минимизации: как правило, точкой экстремума оказывается некоторая граничная точка области. Из-за сложности вопроса и ввиду наличия обширной специальной литературы оставим в стороне многочисленные тонкости, связанные с минимизацией функций при наличии ограничений, и рассмотрим общие вопросы.

Вследствие большой важности решения задач оптимизации для самых различных сторон жизни общества, в настоящее время накопился большой багаж методов и стандартных программ решения задач оптимизации. Поэтому при решении единичной конкретной задачи часто наиболее оправдано обращение к одной из имеющихся стандартных программ.

При решении новых задач оптимизации полезно помнить о следующих моментах. Для рассматриваемых в этой главе методов сходимость приближений всегда доказывалась в предположении наличия достаточно хорошего начального приближения. Это ограничение на метод вызывается существом дела. Так, поверхность уровня многочлена умеренной степени от умеренного числа переменных может содержать весьма много не связанных между собой компонент со сложным взаимным расположением. Поэтому, например, безнадежно рассчитывать на построение алгоритма, позволяющего быстро найти минимум любого многочлена восьмой степени от десяти переменных.

Вышесказанное следует учитывать при использовании стандартных программ, в описании которых указывается на возможность минимизации функций очень большого числа переменных. В лучшем случае оказывается, что:

а) программа эффективно решает задачи минимизации из класса, с которым автор программы обычно имеет дело, или

б) программа, формально говоря, может решить любую задачу минимизации, но требуемое для решения время в подавляющем числе случаев выходит за всякие разумные пределы.

При решении новых задач зачастую приходится разрабатывать специальные, приспособленные именно для этого класса задач, методы отыскания начального приближения к решению. Сложность отыскания приемлемого начального приближения можно проиллюстрировать следующим примером.

Существует ряд стандартных программ наилучшего приближения функций отношением многочленов

$$P_m(x)/Q_n(x) \quad (1)$$

заданных степеней  $m$  и  $n$ . При разработке стандартной программы вычисления интегралов Френеля проводились следующие эксперименты. Каждая из стандартных программ применялась для приближения рассматриваемой функции на некотором отрезке при всех  $m, n$  в пределах  $0 < m + n \leq 12$ . Оказалось, что не менее чем в 3 случаях из 90 возможных каждая программа выдавала ответ, что она не может решить рассматриваемую задачу. У некоторых программ доля таких отказов превосходила половину выданных ответов. В то же время во всех этих программах используются алгоритмы, для которых доказаны теоремы о сходимости при достаточно хорошем начальном приближении.

Чтобы не создало впечатление полной безнадежности решения сколько-нибудь сложных задач оптимизации, рассмотрим тот же вопрос с оптимистической точки зрения.

С этих позиций приведенные выше доводы о сложности минимизации многочленов можно рассматривать как малоубедительные — ведь на самом деле нам никогда не потребуется минимизировать произвольный многочлен. Существует мнение, что задачи минимизации функций с очень сложной структурой линий уровня встречаются довольно редко.

Рассмотрим пример задачи, где сам факт необходимости ее решения может быть поставлен под сомнение.

Неудачи при попытках получения хорошего начального приближения могут быть вызваны следующей особенностью поведения рассматриваемой функции. Точка минимума находится в очень узкой «яме» — при удалении от нее во всех направлениях функция резко возрастает, а потом начинает убывать

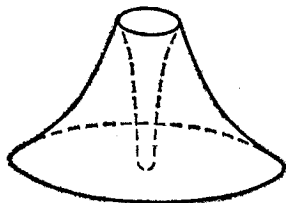


Рис. 7.7.1

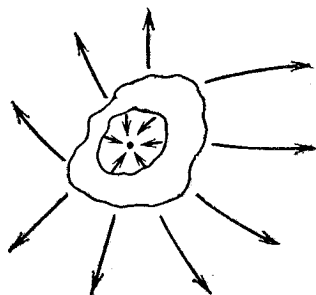


Рис. 7.7.2

(рис. 7.7.1). Предположим, что точка минимума находится путем продвижения в направлении, противоположном градиенту функции  $\Phi(x)$ , иначе — численным интегрированием системы

$$\frac{dx}{dt} + \text{grad } \Phi(x) = 0. \quad (2)$$

Тогда множество начальных условий, исходя из которых мы будем приходить в точку минимума, находится в небольшой окрестности этой точки. Можно говорить, что при рассмотрении окрестности точки минимума в микроскопическом масштабе эта точка оказывается точкой притяжения решений системы (2); при рассмотрении в более крупном масштабе она уже оказывается точкой отталкивания (рис. 7.7.2). Точно такой же характер поведения последовательных приближений к точке минимума будет и у других итерационных методов.

Если «яма», где находится точка минимума, очень узкая, то иногда и не стоит искать эту точку минимума. В самом деле, пусть, например, параметры  $x_i$  отвечают некоторой реальной работающей системе управления. В работе этой системы неизбежны некоторые сбои, т. е. изменения этих параметров. Если точка минимума функции  $\Phi$  находится в такой узкой «яме», то



малые сбои могут существенно испортить характеристики работы системы. В свете вышесказанного выбор точки экстремума в случае, изображенном на рис. 7.7.3, требует дополнительного изучения. Исследователи, обладающие опытом решения практических задач оптимизации, утверждают, что подобные целевые функции с узкими «ямами» возникают обычно в случаях,

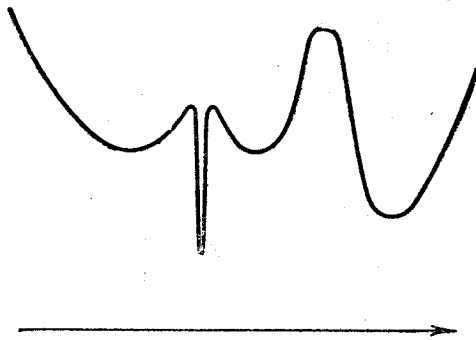


Рис. 7.7.3

когда математическая модель рассматриваемого явления построена неудачно.

Более существенные причины для оптимизма состоят в следующем. Во многих случаях создание математической модели и ее оптимизация зачастую имеют целью улучшение работы уже существующей системы. В этих случаях параметры реальной системы часто яв-

ляются хорошим приближением для дальнейшей оптимизации.

При разработке новой системы типична следующая линия поведения. Сначала строится и оптимизируется простейшая модель, учитывающая важнейшие факторы. Затем модель постепенно усложняется за счет учета все новых и новых факторов. Таким образом, последовательно возникают задачи минимизации функций все большего числа параметров. При удачном построении вспомогательных моделей решение каждой из этих задач оптимизации обычно оказывается хорошим начальным приближением для следующей по сложности задачи.

Это обстоятельство часто используется следующим образом. Пусть перед нами стоит задача минимизации функции большого числа параметров  $x_1, \dots, x_n$ . Построим функцию меньшего числа параметров  $X_1, \dots, X_m$ ,  $m < n$ , приближающую рассматриваемую функцию, иначе — упрощенную модель с определяющими параметрами  $X_1, \dots, X_m$ . Произведем оптимизацию этой функции (модели) и на основе её решения сконструируем начальное приближение. Иногда оказывается полезным произвести несколько шагов такого упрощения функции (модели) и введения новых параметров.

Оптимизация функций (моделей) меньшего числа параметров оказывается более легкой по следующим двум причинам: становится проще структура линий уровня минимизируемой функции; вычисление каждого значения функции обычно требует меньшего объема вычислений.

При построении упрощенных моделей в первую очередь следует учитывать наиболее *важные* параметры задачи.

Что такое важность факторов или параметров? Можно говорить, что важными параметрами являются те, от которых функция сильно зависит. На языке математики это означает, что производные функции по этим параметрам относительно велики. Второстепенными параметрами являются те, от которых рассматриваемая функция зависит слабо, т. е. производные по которым малы. Формально важность параметров можно определить, оценивая производные рассматриваемой функции. Однако для сложных задач такая оценка и особенно математически обоснованный выбор новых параметров  $X_1, \dots, X_m$  весьма трудоемки. Характеристика параметров по более наглядному критерию — их важности — дает возможность руководителю производственной системы подсказать математику первоочередность выбора параметров.

Иногда приемлемый метод оптимизации или способ отыскания хорошего начального приближения можно получить, изучая принципы, которыми руководствуются в своей работе опытный практический работник или руководитель, или приемы их работы.

На одном новом заводе долгое время не удавалось наладить ритмичное производство из-за недостаточного опыта работы операторов. Попытки создания модели производственного процесса, достаточно точной, но в то же время поддающейся анализу средствами математики с помощью ЭВМ, не приводили к успеху. В конце концов пришлось временно отказаться от разработки математической модели и пойти по следующему пути автоматизации и оптимизации производства. В память ЭВМ были записаны режимы работы лучших операторов на родственных предприятиях. Далее, в зависимости от имеющихся на данный момент условий, машина выбирала режим работы, наиболее близкий к режиму работы одного из лучших операторов. Такое мероприятие позволило устранить возникшие трудности.

В другой аналогичной ситуации руководство предприятием не пошло по такому пути. Оно настойчиво требовало от математической группы разработки универсального алгоритма, который по заданным внешним характеристикам конструируемого прибора выдавал бы оптимальный набор внутренних параметров прибора: расположение и размеры деталей конструкции, вес и т. п. Предлагавшиеся математиками алгоритмы оптимизации не оказывались универсальными и в большинстве случаев не приводили к приемлемому решению. Математики предложили подход к решению задачи, имитирующий реальную ситуацию. Конструктор задает компоновку деталей прибора. Машина обчисляет внешние характеристики прибора и выдает их конструктору. На основании полученной информации конструктор вносит изменения в компоновку. Такой диалоговый режим работы позволил бы отказаться от дорогостоящего реального конструирования прибора и его испытаний. К сожалению, в тот момент руководство предприятия отказалось от такого подхода к решению проблемы, а потратило много времени на бесплодные поиски «более квалифицированных» математиков, способных предложить «универсальный» бездиалоговый алгоритм решения казавшейся ему столь простой задачи. Большие материальные затраты и потеря темпов в разработке новой техники в конце концов привели к пониманию того, что предлагавшийся математиками путь решения задачи на данном этапе понимания проблемы был единственно возможным.

Конструкторы приобрели опыт в таком режиме работы. Математикам, анализируя диалог конструкторов с ЭВМ, удалось понять принципы, которыми конструкторы руководствуются при компоновке деталей, и, заложив эти принципы в основу алгоритма решения задачи, создать бездиалоговый, «чисто машинный» алгоритм оптимизации конструкции.

Способы нахождения начального приближения и сами итерационные методы часто имеют аналогию с какими-то реальными процессами и в других явлениях.

Выше упоминалось о методе тяжелого шарика и других методах установления. Работа зрительного аппарата и мозга при отыскании какого-либо предмета, по-видимому, организована по следующей схеме. Сначала производится беглый осмотр всего поля зрения в крупном масштабе, на основании полученной информации выбирается участок для дальнейшего просмотра, затем производится просмотр этого участка в крупном масштабе и т. д. Обратим внимание на сходство с методами из § 3.16 и § 4 данной главы.

При необходимости развивать исследования новых трудных задач важное значение имеет правильная организация научных исследований. Если в основном ясно, в каком направлении нужно развивать исследования, концентрация научных усилий обычно производится в этом направлении. Если наиболее рациональный путь к намеченной цели еще не определился, то часто прибегают к дублированию исследований. Несколько независимых организаций ищут решение, каждая на своем пути, иногда без постоянного обмена информацией. Хотя на первый взгляд кажется, что взаимный обмен информацией всегда полезен, постоянный обмен ею может и помешать возникновению и продвижению оригинальных решений проблемы.

Существует гипотеза (неообщепризнанная), что сходным образом работает мозг при решении какой-либо проблемы: получаемая информация фиксируется не вполне детерминированным образом в различных его участках; в то же время в каждый момент работы над проблемой эта информация извлекается лишь из локализованного участка мозга.

По аналогии со сказанным выше спрашивается следующий подход к решению задачи оптимизации (а также и любых других задач) в случае, если требуется срочное получение результата. Для решения задачи поочередно или независимо используются несколько известных методов решения подобных задач. В обоих случаях, при поочередном или при независимом использовании, алгоритмы  $A_p$ ,  $p = 1, \dots, l$ , работают циклически; длительность  $t_{pq}$  промежутка времени, в течение которого алгоритм  $A_p$  работает  $q$ -й раз, задается пользователем или определяется в процессе работы. При поочередном использовании каждый алгоритм начинает минимизацию с приближения, полученного предшествующим алгоритмом. При независимом использовании каждый алгоритм начинает минимизацию с приближения, полученного в результате предшествующего применения данного алгоритма. Таким образом, в этом случае алгоритмы работают по принципу «кто быстрее».

Режим независимого использования алгоритмов является аналогом параллельной работы организаций при отсутствии взаимного обмена информацией. Как аналог реальной органи-

зации научных исследований с обменом информацией в дискретные моменты времени может рассматриваться следующий режим работы. В начале каждого промежутка времени  $t_{pq}$  алгоритм  $A_p$  просматривает некоторую совокупность приближений из полученных всеми алгоритмами и выбирает наилучшее приближение исходя из своих позиций. Например, он может просматривать все приближения последнего цикла или приближения, полученные алгоритмами в концах всех промежутков  $t_{ij}$ .

Иногда может принести пользу следующая организация работы: после получения алгоритмом  $A_p$  приближения, очень хорошего с позиций алгоритма  $A_s$ , предоставляется время для работы алгоритма  $A_s$ .

Конечно, не следует думать, что непосредственное копирование различных реальных систем всегда позволит наилучшим образом решить рассматриваемую оптимизационную задачу.

Подведем общий итог наших рассуждений. Обычно задачи оптимизации функций большого числа переменных очень трудны; при решении новых задач приходится затрачивать много, иногда бесплодных, усилий, производя пробные просчеты по различным известным и новым алгоритмам. Однако при наличии таких благоприятных факторов, как контакт с практическими работниками и возможность анализа упрощенных моделей, есть все основания сохранять уверенность в благоприятном исходе попыток решения задачи.

Заметим, что в настоящее время достигнут существенный прогресс в области решения задач оптимизации за счет диалогового режима работы пользователя с ЭВМ.

\* \* \*

Не имея перед собой конкретной задачи, невозможно дать рекомендацию, каким методом решения системы нелинейных уравнений или минимизации функций следует воспользоваться. Как уже отмечалось выше, велика возможность столкнуться с ситуацией, когда *область сходимости метода* (множество значений нулевого приближения, при которых метод сходится) очень мала.

Опыт решения подобных задач показывает, что в первую очередь стоит попробовать применить методы, имеющие естественную наглядную интерпретацию, например метод установления, или методы, имитирующие действие человека или животного в подобной ситуации. Для выбора начального приближения надо также привлечь естественные наглядные соображения, имитирующие такие действия. На таком пути часто удается довольно быстро построить алгоритм, позволяющий решить задачу.

При однократном решении простых задач иногда проще всего применить простейший итерационный метод, например Ньютона, с выдачей на экран дисплея хода итерационного процесса.

Задаваясь различными начальными приближениями, часто удается довольно быстро угадать начальное приближение, лежащее в области сходимости метода.

При решении систем уравнений или задач минимизации, возникающих при аппроксимации краевых задач для дифференциальных уравнений, полезно воспользоваться близостью (в соответствующих нормах) решений таких дискретных задач, соответствующих различным шагам сетки. Решение на крупной сетке является хорошим приближением для решения задачи, соответствующей более мелкой сетке. В то же время каждый шаг итерации на крупной сетке менее трудоемок и с теми же затратами можно провести большее число шагов итерации, начиная с одного или со многих начальных приближений. Таким образом, в этом случае имеет смысл решать задачу на последовательности сеток, т. е. последовательно решать несколько систем алгебраических уравнений (порядок системы  $m_j$  возрастает:  $m_{j+1} > m_j$ ). При этом решение  $j$ -й системы  $X_j$  используется для получения начального приближения  $X_{j+1}^0$  к решению  $(j+1)$ -й системы.

Так как векторы  $X_j$  и  $X_{j+1}^0$  имеют, вообще говоря, различную размерность, то для перехода от  $X_j$  к  $X_{j+1}^0$  обычно используют интерполяцию многочленами или сплайнами. Аналогичный подход применим и к другим дискретным задачам, возникающим при аппроксимации задач, связанных с отысканием функций непрерывного аргумента.

В ряде случаев, например при планировании, требуется многократно решать однотипные задачи, причем в режиме реального времени, т. е. решение задачи должно получаться с очень малым отставанием от изменения входных данных задачи. В этом случае для ускорения сходимости следует использовать все многообразие описанных нами методов.

Л и т е р а т у р а к г л. VII: [7, 8, 26, 33].

## Глава VIII

### ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Задача решения обыкновенных дифференциальных уравнений сложнее задачи вычисления однократных интегралов, и доля задач, интегрируемых в явном виде, здесь существенно меньше.

Когда говорят об интегрируемости в явном виде, имеют в виду, что решение может быть вычислено при помощи конечного числа «элементарных» операций: сложения, вычитания, умножения, деления, возведения в степень, логарифмирования, потенцирования, вычисления синуса и косинуса и т. п. Уже в период, предшествовавший появлению ЭВМ, понятия «элементарной» операции и «элементарной» функции претерпели изменение. Решения некоторых частных задач настолько часто встречались в приложениях, что пришлось составить таблицы их значений, в частности таблицы интегралов Френеля, функций Бесселя и ряда других, так называемых *специальных функций*. При наличии таких таблиц исчезает принципиальная разница между вычислением функций  $\sin x$ ,  $\ln x$ , ... и специальных функций. В том и другом случаях можно вычислять значения этих функций при помощи таблиц и те и другие функции можно вычислять, приближая их многочленами, рациональными дробями и т. д. Таким образом, в класс задач, интегрируемых в явном виде, включились задачи, решения которых выражаются через специальные функции. Однако и этот, более широкий, класс составляет относительно малую долю задач, предъявляемых к решению. Существенное расширение класса реально решаемых дифференциальных уравнений, а следовательно, и расширение сферы применения математики произошло с разработкой численных методов и активным повсеместным использованием ЭВМ.

В настоящее время затраты человеческого труда при решении на ЭВМ задачи Коши для обыкновенных дифференциальных уравнений сравнимы с затратами на то, чтобы просто переписать заново формулировку этой задачи. При желании можно непосредственно из оперативной памяти ЭВМ получить с помощью графопостроителя график или изображение решения на экране дисплея. В результате этого для многих категорий науч-

ных работников существенно уменьшился интерес к изучению частных способов интегрирования обыкновенных дифференциальных уравнений в явном виде.

Эта глава посвящена описанию основных методов решения задачи Коши для обыкновенных дифференциальных уравнений, исследованию свойств этих методов и оценке их погрешности.

Мы хотим обратить особое внимание на то обстоятельство, что, как и в других случаях, первоначальный анализ практической пригодности методов и отбрасывание непригодных методов иногда удается произвести, изучая простейшие задачи, где точное и приближенное решения задачи выписываются в явном виде.

## § 1. Решение задачи Коши с помощью формулы Тейлора

Один из простейших по своему описанию методов решения задачи Коши основан на использовании формулы Тейлора.

Пусть требуется найти на отрезке  $[x_0, x_0 + X]$  решение дифференциального уравнения

$$y' = f(x, y) \quad (1)$$

при начальном условии  $y(x_0) = y_0$ ;  $f(x, y)$  — функция, аналитическая в точке  $(x_0, y_0)$ . Дифференцируя (1) по  $x$ , имеем соотношения

$$y'' = f_x(x, y) + f_y(x, y)y', \\ y''' = f_{xx}(x, y) + 2f_{xy}(x, y)y' + f_{yy}(x, y)y'^2 + f_y(x, y)y'', \dots$$

Подставляя  $x = x_0$  и  $y = y_0$  в (1) и в предыдущие соотношения, последовательно получаем значения

$$y'(x_0), \quad y''(x_0), \quad y'''(x_0), \dots$$

Таким образом, можно написать приближенное равенство

$$y(x) \approx \sum_{i=0}^n \frac{y^{(i)}(x_0)}{i!} (x - x_0)^i. \quad (2)$$

Если значение  $|x - x_0|$  больше радиуса сходимости ряда

$$\sum_i \frac{y^{(i)}(x_0)}{i!} (x - x_0)^i,$$

то погрешность (2) не стремится к нулю при  $n \rightarrow \infty$  и предлагаемый метод неприменим.

Иногда целесообразно поступить следующим образом. Разобьем  $[x_0, x_0 + X]$  на отрезки  $[x_{j-1}, x_j]$ ,  $j = 1, \dots, N$ . Будем последовательно получать приближения  $y_j$  к значениям решения  $y(x_j)$ ,  $j = 1, \dots, N$ , по следующему правилу. Пусть значение  $y_j$  уже найдено, вычисляем значения в точке  $x_j$  производных  $y_j^{(i)}$

решения исходного дифференциального уравнения, проходящего через точку  $(x_j, y_j)$ . На отрезке  $[x_j, x_{j+1}]$  полагаем

$$y(x) \approx z_j(x) = \sum_{i=0}^n \frac{y_j^{(i)}}{i!} (x - x_j)^i \quad (3)$$

и соответственно берем

$$y_{j+1} = z_j(x_{j+1}). \quad (4)$$

Рассмотрим случай, когда  $x_{j+1} - x_j \equiv h$ . Если бы значение  $y_j$  совпадало со значением точного значения  $y(x_j)$ , то погрешность от замены  $y_{j+1}$  на  $z_j(x_{j+1})$  имела бы порядок  $O(h^{n+1})$ . Поскольку мы вносим погрешность на  $O(h^{-1})$  отрезках, то можно ожидать, что при уменьшении шага сетки будет выполняться соотношение

$$\max_{0 \leq j \leq N} |y_j - y(x_j)| = O(h^n).$$

В ряде случаев такого рода рассуждения приводят к неправильному заключению о наличии факта сходимости приближенного решения к точному, в то время как в действительности этого нет. Поэтому строгое обоснование сходимости методов при уменьшении шага, а также получение оценки погрешности имеет не только теоретическое, но и важнейшее практическое значение.

При использовании этого метода нужно вычислять значения функции  $f$  и всех ее производных  $f_{x^i y^{m-j}}$  при  $m < n$ , т. е. вычислять  $n(n+1)/2$  значений различных функций. Это требует написания большого числа блоков вычисления производных, что противоречит основной тенденции упрощения отношений между пользователем и ЭВМ.

В настоящее время на некоторых ЭВМ имеются пакеты программ, которые по заданной программе вычисления значений функции строят программу вычисления значений ее производных. Таким образом, при наличии таких пакетов программ, казалось бы, отпадает приведенное выше возражение о сложности использования описанного ранее метода.

Однако этот метод применяется относительно редко. В ряде случаев использование таких пакетов невозможно из-за большой загрузки памяти ЭВМ. Кроме того, как правило, программы, создаваемые с помощью таких пакетов, при той же точности результата требуют существенно больших затрат машинного времени, чем программы, основанные на рассматриваемых далее более простых методах типа Рунге — Кутты и Адамса.

В то же время иногда описанный выше алгоритм может быть полезен. Например, при расчетах траекторий движения небесных тел приходится многократно интегрировать системы дифференциальных уравнений вполне определенного вида при различных начальных условиях и различных значениях параметров в правых частях. То обстоятельство, что все время решается одна



и та же система дифференциальных уравнений, дает следующее преимущество: конкретные формулы для производных правых частей системы имеют много общего; одновременное вычисление всех этих производных требует относительно малого числа арифметических операций, и рассматриваемый метод иногда оказывается эффективнее других методов численного интегрирования.

## § 2. Методы Рунге — Кутты

В частном случае  $n = 1$  расчетная формула (1.3) имеет вид

$$y_{j+1} = y_j + hf(x_j, y_j), \quad h = x_{j+1} - x_j. \quad (1)$$

Этот метод называется *методом Эйлера*. Можно построить другой класс расчетных формул, к которому принадлежит метод Эйлера. Укажем сначала простейшие методы этого класса, получаемые из наглядных соображений. Пусть известно значение решения  $y(x)$  и требуется вычислить значение  $y(x+h)$ . Рассмотрим равенство

$$y(x+h) = y(x) + \int_0^h y'(x+t) dt. \quad (2)$$

При замене интеграла в правой части на величину  $hy'(x)$  погрешность имеет порядок  $O(h^2)$ , т. е.

$$y(x+h) = y(x) + hy'(x) + O(h^2).$$

Поскольку  $y'(x) = f(x, y(x))$ , отсюда имеем

$$y(x+h) = y(x) + hf(x, y(x)) + O(h^2).$$

Отбрасывая член порядка  $O(h^2)$  и обозначая  $x = x_j$ ,  $x+h = x_{j+1}$ , получим расчетную формулу Эйлера (1). Для получения более точной расчетной формулы нужно точнее аппроксимировать интеграл в правой части (2). Воспользовавшись формулой трапеции, получим

$$y(x+h) = y(x) + \frac{h}{2} (y'(x) + y'(x+h)) + O(h^3),$$

иначе,

$$y(x+h) = y(x) + \frac{h}{2} (f(x, y(x)) + f(x+h, y(x+h))) + O(h^3). \quad (3)$$

Соответствующая расчетная формула

$$y_{j+1} = y_j + \frac{h}{2} (f(x_j, y_j) + f(x_{j+1}, y_{j+1})) \quad (4)$$

называется  *неявной формулой Адамса второго порядка точности*. В некоторых случаях, в частности, когда  $f$  линейна по  $y$ , это уравнение может быть разрешено относительно  $y_{j+1}$ . Обычно же

это уравнение неразрешимо явно относительно  $y_{i+1}$ , поэтому произведем дальнейшее преобразование алгоритма.

Заменим  $y(x+h)$  в правой части (3) на некоторую величину

$$y^* = y(x+h) + O(h^2). \quad (5)$$

Тогда правая часть изменится на величину

$$\begin{aligned} \frac{h}{2}(f(x+h, y^*) - f(x+h, y(x+h))) &= \\ &= \frac{h}{2} f_y(x+h, \bar{y})(y^* - y(x+h)), \end{aligned}$$

где  $\bar{y}$  находится между  $y^*$  и  $y(x+h)$ . Вследствие предположения (5) эта величина имеет порядок  $O(h^3)$ . Таким образом, при условии (5) имеет место соотношение

$$y(x+h) = y(x) + \frac{h}{2}(f(x, y(x)) + f(x+h, y^*)) + O(h^3).$$

Условию (5) удовлетворяет результат вычислений по формуле Эйлера

$$y^* = y(x) + hf(x, y(x)).$$

Последние соотношения определяют пару расчетных формул

$$\begin{aligned} y_{i+1}^* &= y_i + hf(x_i, y_i), \\ y_{i+1} &= y_i + \frac{h}{2}(f(x_i, y_i) + f(x_{i+1}, y_{i+1}^*)). \end{aligned} \quad (6)$$

При малых  $h$  выражение в правой части (4) удовлетворяет условию сжимаемости (§ 7.1), поэтому уравнение (4) также можно решать методом простой итерации:

$$y_{i+1}^{k+1} = y_i + \frac{h}{2}(f(x_i, y_i) + f(x_{i+1}, y_{i+1}^k)).$$

Если  $y_{i+1}^0$  вычисляется по методу Эйлера:

$$y_{i+1}^0 = y_i + hf(x_i, y_i),$$

то  $y_{i+1}^1$ , получаемое на первом шаге итерации, совпадает с  $y_{i+1}$ , получаемом по формуле (6). Дальнейшие итерации не приводят к повышению порядка точности по  $h$ ; в то же время иногда главный член погрешности уменьшается при переходе от  $y_{i+1}^k$  к следующему приближению  $y_{i+1}^{k+1}$ . Если такое уменьшение погрешности компенсирует возрастание вычислительных затрат на шаг, то оно целесообразно.

Можно предложить теоретически обоснованный критерий, позволяющий при малых  $h$  выбирать каждый раз наиболее целесообразное число итераций. Однако его использование требует очень большого объема дополнительных вычислений. Поэтому выбор между числом итераций, равным 1 или 2, обычно осуществляется на основе предшествующего опыта, вычислительного эксперимента или просто «волевым» образом.



для которой соответствующее значение  $\varphi^{(s+1)}(0) \neq 0$ . Согласно формуле Тейлора выполняется равенство

$$\varphi(h) = \sum_{i=0}^s \frac{\varphi^{(i)}(0)}{i!} h^i + \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1} = \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1}, \quad (8)$$

где  $0 < \theta < 1$ . Величина  $\varphi(h)$  называется погрешностью метода на шаге, а  $s$  — порядком погрешности метода. При  $q = 1$  имеем

$$\begin{aligned} \varphi(h) &= y(x+h) - y(x) - p_1 h f(x, y), \quad \varphi(0) = 0, \\ \varphi'(0) &= (y'(x+h) - p_1 f(x, y))|_{h=0} = f(x, y)(1 - p_1), \\ \varphi''(h) &= y''(x+h); \end{aligned}$$

здесь и далее  $y = y(x)$ . Равенство  $\varphi'(0) = 0$  выполняется для всех гладких функций  $f(x, y)$  лишь в случае  $p_1 = 1$ . Этому значению  $p_1$  соответствует метод Эйлера. Для погрешности этого метода на шаге, согласно (8), получаем выражение

$$\varphi(h) = \frac{y''(x + \theta h) h^2}{2}.$$

Рассмотрим случай  $q = 2$ . Имеем

$$\varphi(h) = y(x+h) - y - p_1 h f(x, y) - p_2 h f(\bar{x}, \bar{y}),$$

где  $\bar{x} = x + \alpha_2 h$ ,  $\bar{y} = y + \beta_2 h f(x, y)$ .

Вычислим производные функции  $\varphi(h)$ :

$$\begin{aligned} \varphi'(h) &= y'(x+h) - p_1 f(x, y) - p_2 f(\bar{x}, \bar{y}) - p_2 h (\alpha_2 f_x(\bar{x}, \bar{y}) + \\ &\quad + \beta_2 f_y(\bar{x}, \bar{y}) f(x, y)), \\ \varphi''(h) &= y''(x+h) - 2p_2 (\alpha_2 f_{xx}(\bar{x}, \bar{y}) + \beta_2 f_{xy}(\bar{x}, \bar{y}) f(x, y)) - \\ &\quad - p_2 h (\alpha_2^2 f_{xx}(\bar{x}, \bar{y}) + 2\alpha_2 \beta_2 f_{xy}(\bar{x}, \bar{y}) f(x, y) + \beta_2^2 f_{yy}(\bar{x}, \bar{y}) (f(x, y))^2), \\ \varphi'''(h) &= y'''(x+h) - 3p_2 (\alpha_2^2 f_{xx}(\bar{x}, \bar{y}) + \\ &\quad + 2\alpha_2 \beta_2 f_{xy}(\bar{x}, \bar{y}) f(x, y) + \beta_2^2 f_{yy}(\bar{x}, \bar{y}) (f(x, y))^2) + O(h). \end{aligned}$$

Согласно исходному дифференциальному уравнению

$$y' = f, \quad y'' = f_x + f_y f, \quad y''' = f_{xx} + 2f_{xy} f + f_{yy} f^2 + f_y y''.$$

Подставим в выражения  $\varphi(h)$ ,  $\varphi'(h)$ ,  $\varphi''(h)$ ,  $\varphi'''(h)$  значение  $h = 0$ , воспользуемся этими соотношениями и получим

$$\begin{aligned} \varphi(0) &= y - y = 0, \\ \varphi'(0) &= (1 - p_1 - p_2) f(x, y), \\ \varphi''(0) &= (1 - 2p_2 \alpha_2) f_{xx}(x, y) + (1 - 2p_2 \beta_2) f_{xy}(x, y) f(x, y), \\ \varphi'''(0) &= (1 - 3p_2 \alpha_2^2) f_{xx}(x, y) + (2 - 6p_2 \alpha_2 \beta_2) f_{xy}(x, y) f(x, y) + \\ &\quad + (1 - 3p_2 \beta_2^2) f_{yy}(x, y) (f(x, y))^2 + f_y(x, y) y''(x). \quad (9) \end{aligned}$$

Соотношение  $\varphi'(0) = 0$  выполняется при всех  $f(x, y)$ , если

$$1 - p_1 - p_2 = 0; \quad (10)$$

соотношение  $\varphi''(0) = 0$  выполняется, если

$$1 - 2p_2\alpha_2 = 0 \text{ и } 1 - 2p_2\beta_{21} = 0. \quad (11)$$

Таким образом,  $\varphi(0) = \varphi'(0) = \varphi''(0) = 0$  при всех  $f(x, y)$ , если выполнены три указанных выше соотношения (10), (11) относительно четырех параметров. Задавая произвольно один из параметров, получим различные методы Рунге — Кутты с погрешностью второго порядка малости по  $h$ . Например, при  $p_1 = 1/2$  получаем  $p_2 = 1/2$ ,  $\alpha_2 = 1$ ,  $\beta_{21} = 1$ , что соответствует паре расчетных формул (6). При  $p_1 = 0$  получаем  $p_2 = 1$ ,  $\alpha_2 = 1/2$ ,  $\beta_{21} = 1/2$ , что соответствует паре расчетных формул (8). В случае уравнения  $y' = y$ , согласно (9), имеем  $\varphi'''(0) = y$  независимо от значений  $p_1$ ,  $p_2$ ,  $\alpha_2$ ,  $\beta_{21}$ . Отсюда следует, что нельзя построить формул Рунге — Кутты со значениями  $q = 2$  и  $s = 3$ .

В случае  $q = 3$  расчетных формул, соответствующих значению  $s = 4$ , не существует. Наиболее употребительна совокупность расчетных формул при  $q = s = 3$ :

$$k_1 = hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right),$$

$$k_3 = hf(x + h, y - k_1 + 2k_2), \quad \Delta y = \frac{1}{6}(k_1 + 4k_2 + k_3).$$

При  $q = 4, 5$  нельзя построить расчетных формул рассматриваемого вида со значением  $s = 5$ ; при  $q = s = 4$  наиболее употребительна совокупность расчетных формул:

$$k_1 = hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right),$$

$$k_3 = hf\left(x + \frac{h}{2}, y + \frac{k_2}{2}\right),$$

$$k_4 = hf(x + h, y + k_3), \quad \Delta y = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4).$$

Мы использовали выше формулировку «наиболее употребительный». Эта формулировка отражает исторически сложившуюся тенденцию в использовании численных методов. Казалось бы, в руководстве по численным методам следовало не просто отражать тенденцию, а указать, какая формула из данного семейства расчетных формул является наилучшей. Однако ответ на такой вопрос не прост.

У формул одинакового порядка точности по  $h$  главные члены погрешности на шаге часто оказываются непропорциональными. Например, вследствие (8), (9) главный член погрешности формулы (6) равен

$$(B - A)h^3,$$

где

$$B = \frac{1}{6} f_{yy} y'', \quad A = \frac{1}{12} (f_{xx} + 2f_{xy} y' + f_{yy} (y')^2),$$

а у формулы (7) —

$$(B + A/2)h^3.$$

Поэтому можно указать два уравнения таких, что для первого уравнения меньшую погрешность дает метод (6), а для второго уравнения — метод (7).

В подобной ситуации рекомендации в пользу того или другого метода должны основываться на «волевом решении», принятом с учетом традиций и практики использования методов. Понятие практики вычислительной работы является довольно неопределенным. Число различных классов реально встречающихся дифференциальных уравнений существенно превосходит число задач, на которых производится сравнение методов численного интегрирования, поэтому суждения «с позиций практики» не всегда объективны. Однако несмотря на такую неопределенность, критерий практики часто несет в себе определенную положительную информацию, которая зачастую на данном этапе развития науки не может быть формализована или обоснована.

Если исторически первый из методов рассматриваемого класса оказался приемлемым, то в дальнейшем пользователи привыкают к нему. Замена этого метода на другой, даже более эффективный метод требует определенных затрат времени на «привыкание» пользователей к новому методу (а следовательно, и определенных психологических затрат). Чтобы широкий круг пользователей согласился на подобную перестройку, необходимо существенное преимущество нового метода по какой-либо из характеристик.

При дальнейшем рассмотрении для нас будет существенно, что погрешность метода на шаге  $\varphi(h)$  имеет главный член, а именно справедливо представление вида

$$\varphi(h) = \psi(x, y)h^{s+1} + O(h^{s+2}). \quad (12)$$

Наметим основные этапы доказательства этого соотношения. Предположим, что правая часть и все ее производные до порядка  $s+1$  включительно ограничены равномерно в области  $G$ :  $x_0 \leq x \leq x_0 + X$ ,  $-\infty < y < \infty$ . Тогда также будут равномерно ограничены производные всех решений уравнения до порядка  $s+2$  включительно. Согласно формуле Тейлора соотношение (8) можно записать в уточненной форме

$$\varphi(h) = \frac{\varphi^{(s+1)}(0)}{(s+1)!} h^{s+1} + \frac{\varphi^{(s+2)}(\theta h)}{(s+2)!} h^{s+2}.$$

Имеем равенство

$$\varphi^{(s+1)}(0) = y^{(s+1)}(0) - z^{(s+1)}(0).$$

Обе величины  $y^{(s+1)}(0)$  и  $z^{(s+1)}(0)$  явно выражаются через значения в точке  $(x, y)$  функции  $f$  и ее производных порядка не выше  $s$ ; примеры таких явных выражений (при  $s=2$ ) мы уже получали.

Поскольку правая часть дифференцируема  $s+1$  раз, то отсюда следует, что функция  $\psi(x, y)$  дифференцируема в области  $G$  и ее производные  $\psi_x$  и  $\psi_y$  равномерно ограничены в этой области. Аналогично устанавливается, что величина  $\varphi^{(s+2)}(\theta h)$  равномерно ограничена при  $x_0 \leq x < x+h \leq x_0 + X$ . Таким образом, соотношение (12) имеет место.

### § 3. Методы с контролем погрешности на шаге

Часто в ходе расчетов бывает целесообразно изменять шаг интегрирования, контролируя величину погрешности метода на шаге. При практической оценке этой величины можно, напри-

мер, рассуждать следующим образом. Главный член погрешности на шаге интегрирования есть

$$\frac{\varphi^{(s+1)}(0) h^{s+1}}{(s+1)!}.$$

Точка  $(x+h, z(h))$  находится близко от точки  $(x, y)$ , поэтому погрешность на следующем шаге интегрирования будет иметь такой же главный член. В результате двух шагов будет получено приближение  $y^{(1)}$  к значению  $y(x+2h)$  такое, что

$$y^{(1)} - y(x+2h) \sim 2 \frac{\varphi^{(s+1)}(0) h^{s+1}}{(s+1)!}.$$

Если, исходя из точки  $(x, y)$ , применить метод Рунге—Кутты с шагом  $2h$ , то получится приближенное значение  $y^{(2)}$ , для которого

$$y^{(2)} - y(x+2h) \sim \frac{\varphi^{(s+1)}(0) (2h)^{s+1}}{(s+1)!}.$$

Из этих соотношений вытекает представление главного члена погрешности на шаге

$$y^{(1)} - y(x+2h) \sim \frac{y^{(2)} - y^{(1)}}{2^s - 1}.$$

При желании можно уточнить полученное приближенное значение, прибавив к нему величину главного члена погрешности, т. е. положить

$$y(x+2h) \approx y^{(1)} + \frac{y^{(1)} - y^{(2)}}{2^s - 1}. \quad (1)$$

Для более гибкого управления выбором шага интегрирования иногда желательно иметь возможность совершать шаг интегрирования и оценивать погрешность при меньшем количестве вычисляемых значений правых частей.

Примером совокупностей формул с теми же характеристиками точности при меньшем числе обращений к правой части может служить совокупность формул

$$\begin{aligned} k_1 &= hf(x, y), & k_2 &= hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right), \\ k_3 &= hf\left(x + \frac{h}{2}, y + \frac{1}{4}(k_1 + k_2)\right), & k_4 &= hf\left(x + h, y - k_2 + 2k_3\right), \\ k_5 &= hf\left(x + \frac{2h}{3}, y + \frac{1}{27}(7k_1 + 10k_2 + k_4)\right), & & \\ k_6 &= hf\left(x + \frac{h}{5}, y + \frac{1}{625}(28k_1 - 125k_2 + 546k_3 + \right. & & \\ & & & \left. + 54k_4 - 378k_5)\right), \end{aligned} \quad (2)$$

$$\Delta y = \frac{1}{6}(k_1 + 4k_3 + k_4)$$

с главным членом погрешности

$$y(x+h) - z(h) = r + O(h^6),$$
$$r = -\frac{1}{366}(42k_1 + 224k_3 + 21k_4 - 162k_5 - 125k_6).$$

Если положить  $y(x_0+h) \approx z(h) + r$ , то получим метод рассматриваемого типа со значением  $s = 5$  и соответственно погрешностью на шаге порядка  $O(h^6)$ .

В одной из наиболее распространенных стандартных программ управление шагом интегрирования осуществляется по методу, близкому к горизонтальной процедуре из § 3.16. Задаются мерами погрешности на шаге  $\varepsilon_0$  и  $\varepsilon_1 < \varepsilon_0$  и некоторым параметром  $M > 0$  размерности  $y$ . Обычно  $\varepsilon_l/\varepsilon_0 \geq 2^{-l}$ , где  $l$  — порядок величины  $r$  по  $h$ ; часто берут  $\varepsilon_l/\varepsilon_0 = 2^{-l}$ . Если  $\psi_n(h) = |r|/\max(M, |y_n|) > \varepsilon_0$ , то шаг признается слишком большим и делается попытка интегрирования, начиная с тех же значений  $(x_n, y_n)$ , с вдвое более мелким шагом  $(h/2)$ . Если  $\psi_n(h) \leq \varepsilon_0$ , то достигнутая точность признается удовлетворительной. В случае, когда  $\varepsilon_1 \leq \psi_n(h) \leq \varepsilon_0$ , следующий шаг берется равным  $h$ , а в случае, когда  $\psi_n(h) < \varepsilon_1$ , — равным  $2h$ . Такой относительно простой способ выбора переменного шага интегрирования часто позволяет решить задачу с существенно меньшими затратами времени ЭВМ по сравнению со случаем интегрирования с постоянным шагом (при той же точности результата).

#### § 4. Оценки погрешности одношаговых методов

Рассмотрим множество всевозможных методов интегрирования, где последовательно получаются приближения  $y_j$  к значениям  $y(x_j)$ ,  $x_0 < x_1 < \dots < x_N = x_0 + X$ . Пусть в процессе численного интегрирования  $k$  фиксировано и при  $j \geq k$  значения  $y_j$  определяются как значения некоторого функционала

$$y_j = \Phi(f; x_j, \dots, x_{j-k}; y_{j-1}, \dots, y_{j-k}). \quad (1)$$

Такой способ численного интегрирования называют  $k$ -шаговым. Все построенные выше способы интегрирования имеют следующее общее свойство: приближенное значение решения в следующей точке определялось только в зависимости от значения решения в предыдущей точке, следовательно, расчетные формулы, соответствующие этим способам, представимы в виде (1) со значением  $k = 1$ . Такие методы называются *одношаговыми*.

Рассмотрим специальный способ получения оценки погрешности, применимый лишь к одношаговым методам.

Запишем формулу (1) в виде

$$y_{j+1} = \Phi(f, x_j, x_{j+1} - x_j, y_j). \quad (2)$$

Получаемые в процессе реальных вычислений приближения к значениям  $y(x_j)$  связаны не соотношениями (2), а некоторыми



соотношениями

$$y_{j+1} = \Phi(f, x_j, x_{j+1} - x_j, y_j) + \delta_{j+1}. \quad (3)$$

Наличие слагаемого  $\delta_{j+1}$  обусловлено следующими причинами:

- 1) округлением чисел при вычислениях;
- 2) погрешностями в значениях правой части  $f(x, y)$ ; эти погрешности вызваны тем, что рассматриваемая нами функция

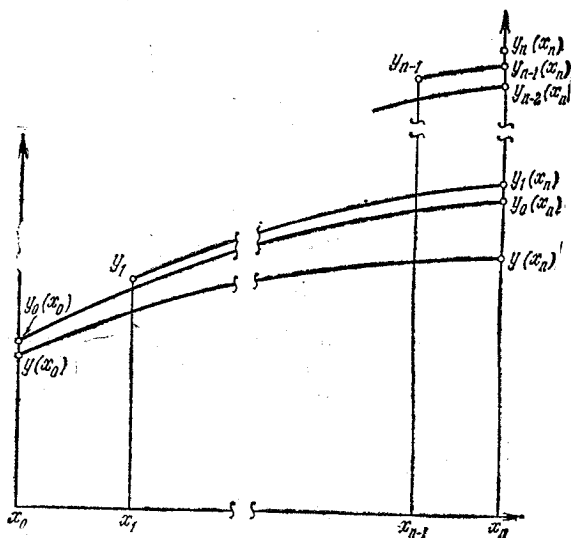


Рис. 8.4.1

$f(x, y)$  является некоторым приближением к правой части реального дифференциального уравнения; кроме того, зачастую в процессе вычисления значений  $f(x, y)$  в ЭВМ эта функция приближается функциями, вычислимыми относительно элементарных операций ЭВМ;

3) в некоторых случаях значение  $y_{j+1}$  определяется из уравнения, эквивалентного (1), но не разрешенного в явном виде относительно переменной  $y_{j+1}$ ; тогда величина  $\delta_{j+1}$  содержит составляющую, являющуюся следствием приближенного решения этого уравнения.

Хотя погрешность  $\delta_{j+1}$  вызвана не только округлением, ее часто называют *вычислительной погрешностью на шаге*.

Точно так же начальное условие  $y_0$  отличается от значения отыскиваемого решения задачи  $y(x_0)$  из-за погрешности в определении исходных данных и округлений. Пусть  $y(x)$  — искомое решение дифференциального уравнения, а  $y_j(x)$  — решения, удовлетворяющие условиям  $y_j(x_j) = y_j$  (рис. 8.4.1). Погрешность

$R_n = y_n(x_n) - y(x_n)$  можно представить в виде

$$R_n = y_n(x_n) - y_0(x_n) + y_0(x_n) - y(x_n) = \\ = \sum_{j=1}^n (y_j(x_n) - y_{j-1}(x_n)) + (y_0(x_n) - y(x_n)). \quad (4)$$

Разность решений дифференциального уравнения в одной точке может быть выражена через разность их в другой точке следующим путем.

*Лемма.* Пусть  $Y_1(x)$  и  $Y_2(x)$  — решения дифференциального уравнения  $y' = f(x, y)$ , где  $f(x, y)$  — непрерывная и непрерывно дифференцируемая по переменной  $y$  функция. Тогда

$$Y_2(\beta) - Y_1(\beta) = (Y_2(\alpha) - Y_1(\alpha)) \exp \left\{ \int_{\alpha}^{\beta} f_y(x, \tilde{y}(x)) dx \right\}, \quad (5)$$

где  $\tilde{y}(x)$  заключено между  $Y_1(x)$  и  $Y_2(x)$ .

*Доказательство.* Вычтем друг из друга равенства

$$Y_2' = f(x, Y_2), \quad Y_1' = f(x, Y_1).$$

Согласно формуле Лагранжа разность  $f(x, Y_2) - f(x, Y_1)$  может быть представлена в виде  $f_y(x, \tilde{y})(Y_2 - Y_1)$ , где  $\tilde{y}$  заключено между  $Y_1$  и  $Y_2$ . В результате получится линейное дифференциальное уравнение относительно  $Y_2 - Y_1$ :

$$(Y_2 - Y_1)' = f_y(x, \tilde{y})(Y_2 - Y_1);$$

отсюда следует соотношение (5). Без дополнительного исследования мы ничего не можем сказать о непрерывности или гладкости функции  $\tilde{y}(x)$ . Однако этого и не требуется. Функция

$$f_y(x, \tilde{y}(x)) = \frac{f(x, Y_2(x)) - f(x, Y_1(x))}{Y_2(x) - Y_1(x)}$$

непрерывна, поскольку числитель и знаменатель — непрерывные функции, а знаменатель отличен от нуля; поэтому все проведенные выше преобразования являются законными. Лемма доказана.

Пусть  $\alpha = x_j$ ,  $\beta = x_n$ ,  $Y_1(x) = y_{j-1}(x)$ ,  $Y_2(x) = y_j(x)$ , тогда

$$y_j(x_n) - y_{j-1}(x_n) = (y_j(x_j) - y_{j-1}(x_j)) \exp \left\{ \int_{x_j}^{x_n} f_y(x, \tilde{y}_j(x)) dx \right\},$$

где  $\tilde{y}_j(x)$  заключено между  $y_{j-1}(x)$  и  $y_j(x)$ .

Точно так же

$$y_0(x_n) - y(x_n) = (y_0(x_0) - y(x_0)) \exp \left\{ \int_{x_0}^{x_n} f_y(x, \tilde{y}_0(x)) dx \right\}.$$

Теперь равенство (4) можно записать в виде

$$R_n = \sum_{j=1}^n \omega_j \exp \left\{ \int_{x_j}^{x_n} f_y(x, \tilde{y}_j(x)) dx \right\} + R_0 \exp \left\{ \int_{x_0}^{x_n} f_y(x, \tilde{y}_0(x)) dx \right\}, \quad (6)$$

где  $\omega_j = y_j(x_j) - y_{j-1}(x_j)$ ,  $j = 1, \dots$

Из (3) вытекает соотношение

$$\omega_j = y_j - y_{j-1}(x_j) = \rho_j + \delta_j,$$

где

$$\rho_j = F(f, x_{j-1}, x_j - x_{j-1}, y_{j-1}) - y_{j-1}(x_j).$$

Посмотрим, какой смысл имеет величина  $\rho_j$ ;  $F(f, x_{j-1}, x_j - x_{j-1}, y_{j-1})$  есть число, получаемое в результате вычислений по расчетной формуле (2),  $y_{j-1}(x_j)$  — значение в точке  $x_j$  точного решения дифференциального уравнения, удовлетворяющего условию  $y_{j-1}(x_{j-1}) = y_{j-1}$ . Таким образом,  $\rho_j$  есть погрешность одного шага рассматриваемого метода, если вычисления начинаются с точки  $(x_{j-1}, y_{j-1})$  и производятся без округлений, а шаг равен  $x_j - x_{j-1}$ . Величина  $\rho_j$  называется *погрешностью метода на шаге*.

Предположим, что при всех  $j$ , соответствующих рассматриваемому отрезку интегрирования  $x_0 < x_j \leq x_0 + X$ , выполняется неравенство

$$|\rho_j| \leq C_1(x_j - x_{j-1})^{s+1}. \quad (7)$$

Пусть

$$L = \sup_{x_0 \leq x \leq x_0 + X} |f_y| < \infty$$

и

$$H = \max_{0 < j \leq N} (x_j - x_{j-1}).$$

Загрубляя (7), имеем

$$|\rho_j| \leq C_1 H^s (x_j - x_{j-1}). \quad (8)$$

При  $x_0 \leq x_j \leq x_n \leq x_0 + X$  справедливы неравенства

$$\exp \left\{ \int_{x_j}^{x_n} f_y(x, \tilde{y}_j(x)) dx \right\} \leq \exp \{L(x_n - x_j)\} \leq \exp \{LX\}.$$

Воспользовавшись этим неравенством для оценки правой части (6), получим

$$|R_n| \leq \exp \{LX\} \left( \sum_{j=1}^n (|\rho_j| + |\delta_j|) + |R_0| \right).$$

Применим теперь к предыдущему неравенству оценку (8), получим

$$|R_n| \leq \exp\{LX\} \left( \sum_{j=1}^n (C_1 H^s (x_j - x_{j-1}) + |\delta_j|) + |R_0| \right) \leq \\ \leq \exp\{LX\} (C_1 H^s (x_n - x_0) + n\delta + |R_0|) \leq \\ \leq \exp\{LX\} (C_1 X H^s + N\delta + |R_0|); \quad (9)$$

здесь  $\delta = \max |\delta_j|$ . Из этого соотношения следует, что  $\max_{x_0 < x_n \leq x_0 + X} |R_n| \rightarrow 0$  при  $H \rightarrow 0$ , если одновременно  $N\delta \rightarrow 0$ ,  $|R_0| \rightarrow 0$ . Таким образом, при достаточно мелком шаге интегрирования и малой вычислительной погрешности приближенное решение, получаемое при употреблении метода Рунге — Кутты, близко к точному решению.

Часто решение дифференциального уравнения отыскивается на большом промежутке. Тогда в полученные оценки погрешности входит как множитель очень большое число  $\exp\{LX\}$ . При  $LX$  большом может оказаться, что достижение нужной точности требует столь мелких шагов и столь малой величины вычислительной погрешности на шаге, что использование рассматриваемого метода будет нецелесообразно. Поэтому характеристика методов только по признаку — сходится ли приближенное решение к точному при измельчении шага и при достаточно быстром уменьшении вычислительной погрешности или не сходится — является еще недостаточной.

Если  $f_y(x, y) \leq -b < 0$ , то в оценке (9) можно избавиться от множителя, резко растущего с увеличением  $X$ . Рассмотрим случай постоянного шага  $x_j - x_{j-1} = h$ . Тогда

$$\exp \left\{ \int_{x_j}^{x_n} f_y(x, \tilde{y}(x)) dx \right\} \leq \exp\{-b(x_n - x_j)\} \leq \exp\{-b(n-j)h\}.$$

Пусть  $|\omega_j| \leq C_1 h^{s+1} + \delta$ . Оценивая правую часть (6), получаем

$$|R_n| \leq \sum_{j=1}^n (C_1 h^{s+1} + \delta) \exp\{-b(n-j)h\} + |R_0| \exp\{-bnh\}. \quad (10)$$

Имеем

$$\sum_{j=1}^n \exp\{-b(n-j)h\} \leq \sum_{k=0}^{\infty} \exp\{-bkh\} = \frac{1}{1 - \exp\{-bh\}}.$$

Таким образом, получаем окончательную оценку погрешности

$$|R_n| \leq \frac{C_1 h^{s+1} + \delta}{1 - \exp\{-bh\}} + |R_0| \exp\{-bnh\}. \quad (11)$$

Поскольку  $1 - \exp\{-bh\} \sim |b|h$ , то верна более простая по виду оценка

$$|R_n| \leq C_2(h^s + \delta/h) + |R_0| \exp\{-bnh\}. \quad (12)$$

Формально эта оценка не зависит от длины промежутка интегрирования  $X$ , однако длина промежутка интегрирования может неявно влиять на значение коэффициента  $C_2$  через оценки производных.

Наличие оценки (11), не ухудшающейся с увеличением промежутка интегрирования, позволяет использовать такие методы для отыскания, например, устойчивых решений дифференциальных уравнений путем установления. Начинаем численное интегрирование с произвольных начальных данных и с течением времени выходим на устойчивое решение. Особенно часто этот прием употребляется при отыскании устойчивых предельных циклов систем обыкновенных дифференциальных уравнений.

В связи с полученной оценкой (12) и возможностью получения аналогичной оценки для случая численного решения задачи Коши для систем дифференциальных уравнений с быстро сближающимися решениями одношаговые методы находят широкое применение в вычислительной практике. В то же время методы, для которых в подобной ситуации погрешность растет неограниченно, практически исчезли из употребления.

Другим достоинством одношаговых методов является удобство изменения шага интегрирования и однотипность вычислений во всех расчетных точках (у конкурирующих с ними методов Адамса изменение шага интегрирования и начало вычислений производятся с помощью некоторых специальных формул, которые мы не рассматриваем из-за их громоздкости).

## § 5. Конечно-разностные методы

Среди  $k$ -шаговых методов наиболее употребительны методы интегрирования на сетке с постоянным шагом  $x_j - x_{j-1} \equiv h = \text{const}$  при помощи соотношений вида

$$\sum_{i=0}^k a_{-i} y_{n-i} - h \sum_{i=0}^k F_i(h, x_{n-i}, y_{n-i}) = 0, \quad (1)$$

где  $a_i$  — постоянные,  $F_i$  — некоторые функции, определяемые функцией  $f(x, y)$ . В свою очередь, среди таких методов традиционно наиболее употребительны методы

$$\sum_{i=0}^k a_{-i} y_{n-i} - h \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) = 0, \quad (2)$$

которые принято называть *конечно-разностными методами* или *конечно-разностными схемами*.

В вычислительной практике применяются формулы (1), (2) со значениями  $a_0 \neq 0$ ,  $b_0 = 0$  — *явные*, или *экстраполяционные*, и формулы с  $a_0 \neq 0$ ,  $b_0 \neq 0$  — *неявные*, или *интерполяционные*.

Формулы (1), (2) при  $a_0 = 0$ ,  $b_0 \neq 0$ , называемые формулами с забеганием вперед, рассматриваться не будут, поскольку они не нашли распространения в вычислительной практике из-за сложности использования. В дальнейшем предполагается, что  $a_0 \neq 0$ . При  $b_0 \neq 0$  уравнение (2) записываем в виде

$$y_n = \Phi_n(y_n), \quad \Phi_n(y_n) = A_n + h \frac{b_0}{a_0} f(x_n, y_n).$$

Здесь

$$A_n = a_0^{-1} \left( - \sum_{i=1}^k a_{-i} y_{n-i} + h \sum_{i=1}^k b_{-i} f(x_{n-i}, y_{n-i}) \right)$$

не зависит от  $y_n$ . Далее решаем это уравнение методом простой итерации:

$$y_n^{k+1} = \Phi_n(y_n^k).$$

Поскольку  $\Phi'_n(y) = h \frac{b_0}{a_0} f_y(x_n, y)$ , то при достаточно малых  $h$  выполнено условие сжимаемости и итерационный процесс сходится. Начальное приближение  $y_n^0$  определяется из какой-либо явной формулы

$$\sum_{i=0}^l a_{-i}^1 y_{n-i} - h \sum_{i=1}^l b_{-i}^1 f(x_{n-i}, y_{n-i}) = 0, \quad a_0^1 \neq 0.$$

Число итераций на шаге определяется из разумного соотношения между трудоемкостью итераций и точностью получаемых в процессе итераций приближений так же, как в случае формулы (2.4), являющейся частным случаем формулы (2).

Простейшие методы типа (2) получаются на основе квадратурных формул. Всякая квадратурная формула

$$\int_{-ph}^0 f(x) dx = h \sum_{i=0}^m b_{-i} f(-ih) + r, \quad m \geq 0, \quad (3)$$

где  $r$  — остаточный член, порождает соответствующую формулу численного интегрирования обыкновенных дифференциальных уравнений. Действительно, подставляя в (3) соотношение  $f(x) = y'(x_n + x)$ , имеем

$$\int_{-ph}^0 f(x) dx = y(x_n) - y(x_{n-p}) = h \sum_{i=0}^m b_{-i} y'(x_{n-i}) + r.$$

Заменяя  $y'(x)$  на  $f(x, y(x))$  и отбрасывая  $r$ , получим

$$y(x_n) - y(x_{n-p}) \approx h \sum_{i=0}^m b_{-i} f(x_{n-i}, y(x_{n-i})). \quad (4)$$

Соответствующая конечно-разностная схема запишется следующим образом:

$$y_n - y_{n-p} - h \sum_{i=0}^m b_{-i} f(x_{n-i}, y_{n-i}) = 0.$$

Например, формуле трапеций

$$\int_{-h}^0 f(x) dx \approx \frac{h}{2} (f(0) + f(-h))$$

соответствует интерполяционная формула

$$y_n - y_{n-1} = \frac{h}{2} (f_n + f_{n-1}), \quad (5)$$

формуле Симпсона

$$\int_{-2h}^0 f(x) dx \approx \frac{h}{3} (f(0) + 4f(-h) + f(-2h))$$

— интерполяционная формула  $y_n - y_{n-2} = \frac{h}{3} (f_n + 4f_{n-1} + f_{n-2})$ ; здесь  $f_m = f(x_m, y_m)$ . Квадратурной формуле прямоугольников, записанной в виде

$$\int_{-2h}^0 f(x) dx \approx 2hf(x),$$

соответствует экстраполяционная формула

$$y_n - y_{n-2} = 2hf_{n-1}. \quad (6)$$

Если остаточный член квадратуры (3) оценивается через  $D(q) \max |f^{(q)}(x)| h^{q+1}$ , то погрешность равенства (4) будет оцениваться через  $D(q) \max |y^{(q+1)}(x)| h^{q+1}$ .

В настоящее время из конечно-разностных методов, как правило, употребляются на практике только методы, соответствующие  $p = 1$ ; их называют *методами Адамса*.

Другие известные методы вида (2) не выдержали испытания на практике по следующим причинам:

1) отсутствует сходимость при уменьшении шага (в предположении отсутствия вычислительной погрешности) даже для бесконечно дифференцируемых правых частей  $f(x, y)$ ;

2) при наличии сходимости происходит экспоненциальный рост (с увеличением  $X$ ) погрешности в рассмотренном в предыдущем параграфе случае, когда  $f_y \leq -b < 0$ ;

3) для некоторых схем при  $p > 1$  возникают дополнительные (по сравнению со случаем  $p = 1$ ) неудобства при изменении шага интегрирования.

Явные формулы Адамса обычно записывают в виде

$$y_n - y_{n-1} = h \sum_{i=0}^m \gamma_i \nabla^i f_{n-1},$$

а неявные — в виде

$$y_n - y_{n-1} = h \sum_{i=0}^m \bar{\gamma}_i \nabla^i f_n;$$

здесь, как и в § 2.10,

$$\nabla^i f_n = \sum_{j=0}^i (-1)^j C_{ij}^i f_{n-j}.$$

Коэффициенты  $\gamma_i, \bar{\gamma}_i$  вычисляются по формулам

$$\gamma_i = \int_0^1 \prod_{k=1}^i \left(1 - \frac{u}{k}\right) du, \quad \bar{\gamma}_0 = 1,$$

$$\bar{\gamma}_i = \gamma_i - \gamma_{i-1} = - \int_0^1 \frac{u}{i} \prod_{k=1}^{i-1} \left(1 - \frac{u}{k}\right) du \quad \text{при } i > 0,$$

Задача 1. Показать, что

$$\gamma_i \sim \frac{\text{const}}{\ln i}, \quad \bar{\gamma}_i \sim \frac{\text{const}}{i \ln i} \quad \text{при } i \rightarrow \infty.$$

Обычно методы Адамса используются по следующей схеме. Сначала вычисляется нулевое приближение по явной формуле Адамса, а затем производятся 1—2 итерации на основе неявной формулы (с тем же значением  $m$ ).

## § 6. Метод неопределенных коэффициентов

Для построения формул численного интегрирования можно также использовать метод неопределенных коэффициентов. Заменяем производную  $y'(x_n)$  и значение  $f(x_n, y(x_n))$  некоторыми выражениями

$$y'(x_n) \approx \sum_{i=0}^k \frac{a_{-i} y(x_{n-i})}{h}, \quad (1)$$

$$f(x_n, y(x_n)) \approx \sum_{i=0}^k b_{-i} f(x_{n-i}, y(x_{n-i})) \quad (2)$$

(предполагается, что  $a_{-i}$  и  $b_{-i}$  не зависят от  $h$ ). Отсюда получаем приближенное равенство

$$\sum_{i=0}^k \frac{a_{-i} y(x_{n-i})}{h} \approx \sum_{i=0}^k b_{-i} f(x_{n-i}, y(x_{n-i})). \quad (3)$$



Ему соответствует конечно-разностная схема

$$\sum_{i=0}^k \frac{a_{-i} y_{n-i}}{h} - \sum_{i=0}^k b_{-i} f_{n-i} = 0. \quad (4)$$

Величина

$$r_n = \sum_{i=0}^k \frac{a_{-i} y(x_{n-i})}{h} - \sum_{i=0}^k b_{-i} f(x_{n-i}, y(x_{n-i}))$$

называется погрешностью аппроксимации исходного дифференциального уравнения схемой (4).

Определение. Разностная задача аппроксимирует дифференциальную на отрезке  $[x_0, x_0 + X]$ , если

$$\|r\| = \max_{x_0 \leq x_n \leq x_0 + X} |r_n| \rightarrow 0 \text{ при } h \rightarrow 0.$$

Вспомогая, что  $f(x_{n-i}, y(x_{n-i})) = y'(x_{n-i})$  и  $x_{n-i} = x_n - ih$  получим

$$r_n = \sum_{i=0}^k \frac{a_{-i} y(x_n - ih)}{h} - \sum_{i=0}^k b_{-i} y'(x_n - ih).$$

Предположим, что все производные решения до порядка  $q$  ограничены:

$$|y^{(p)}(x)| \leq M_p < \infty \text{ при } x_0 \leq x \leq x_0 + X, \quad p = 0, \dots, q.$$

Представим все величины  $y(x_n - ih)$  и  $y'(x_n - ih)$  с помощью формулы Тейлора следующим образом:

$$y(x_n - ih) = \sum_{p=0}^{q-1} y^{(p)}(x_n) \frac{(-ih)^p}{p!} + \beta_n^i,$$

$$y'(x_n - ih) = \sum_{p=1}^{q-1} y^{(p)}(x_n) \frac{(-ih)^{p-1}}{(p-1)!} + \gamma_n^i,$$

где согласно оценке остаточного члена ряда Тейлора имеем

$$|\beta_n^i| \leq M_q (ih)^q / q!, \quad |\gamma_n^i| \leq M_q (ih)^{q-1} / (q-1)!.$$

Подставим выражения  $y(x_n - ih)$  и  $y'(x_n - ih)$  в правую часть представления для  $r_n$  и соберем коэффициенты при  $y^{(p)}(x_n)$ . Получим

$$r_n = E_0 h^{-1} y(x_n) + E_1 y'(x_n) + \dots + E_{q-1} h^{q-2} y^{(q-1)}(x_n) + \varepsilon_n; \quad (5)$$

здесь

$$E_0 = \sum_{i=0}^k a_{-i},$$

$$E_p = \sum_{i=0}^k \frac{a_{-i} (-i)^p}{p!} - \sum_{i=0}^k \frac{b_{-i} (-i)^{p-1}}{(p-1)!}, \quad p > 0, \quad (6)$$

$$e_n = \sum_{i=0}^k \frac{a_{-i} \beta_n^i}{h} - \sum_{i=0}^k b_{-i} \gamma_n^i = O(h^{q-1}).$$

Имеем

$$|e_n| \leq D_q M_q h^{q-1},$$

где

$$D_q = \sum_{i=0}^k |a_{-i}| \frac{i^q}{q!} + \sum_{i=0}^k |b_{-i}| \frac{i^{q-1}}{(q-1)!}.$$

Как правило, производя более аккуратную оценку, можно уменьшить значение  $D_q$  в оценке  $|e_n|$ . Если  $E_0 = \dots = E_m = 0$ , то  $e_n = O(h^m)$  и говорят, что схема (2) имеет  $m$ -й порядок аппроксимации. Всякая схема  $m$ -го порядка аппроксимации является схемой  $q$ -го порядка аппроксимации при  $q < m$ . Если  $E_0 = \dots = E_m = 0$ , а  $E_{m+1} \neq 0$ , то говорят, что порядок аппроксимации строго равен  $m$ .

Согласно (1), (2) для любой гладкой  $y(x)$  имеем соотношения

$$\lim_{h \rightarrow 0} \sum_{i=0}^k \frac{a_{-i} y(x - ih)}{h} = y'(x), \quad (7)$$

$$\lim_{h \rightarrow 0} \sum_{i=0}^k b_{-i} f(x - ih, y(x - ih)) = f(x, y(x)).$$

Лемма. Соотношения (7) выполнены тогда и только тогда, когда

$$E_0 = E_1 = 0, \quad b_0 + \dots + b_{-k} = 1. \quad (8)$$

Согласно формуле Тейлора имеем

$$y(x - ih) = y(x) - ih y'(x),$$

$$f(x - ih, y(x - ih)) = f(x, y(x)) + O(h).$$

Подставляя эти выражения для левых частей в (7), получим

$$\lim_{h \rightarrow 0} \left( \left( \sum_{i=0}^k \frac{a_{-i}}{h} y(x) \right) + \sum_{i=0}^k a_{-i} (-i) y'(x) + O(h) \right) = y'(x),$$

$$\lim_{h \rightarrow 0} \left( \left( \sum_{i=0}^k b_{-i} \right) f(x, y(x)) + O(h) \right) = f(x, y(x)).$$

Для справедливости этих соотношений необходимо и достаточно выполнения условий

$$\sum_{i=0}^k a_{-i} = 0, \quad - \sum_{i=0}^k i a_{-i} = 1, \quad \sum_{i=0}^k b_{-i} = 1.$$

Левая часть первого из этих равенств равна  $E_0$ , разность левых частей второго и третьего равна  $E_1$ . Отсюда следует справедливость утверждения леммы.

Уравнения  $E_0 = \dots = E_m = 0$  образуют однородную линейную систему алгебраических уравнений относительно  $2k + 2$  неизвестных. Если число неизвестных больше числа уравнений  $2k + 2 > m + 1$  (или, что то же самое  $2k \geq m$ ), то эта система имеет ненулевое решение. Можно показать, что при  $2k = m$  эта система имеет однопараметрическое семейство ненулевых решений

$$a_{-i} = c a_{-i}^0, \quad b_{-i} = c b_{-i}^0,$$

причем  $\omega = \sum_{i=0}^k b_{-i}^0 \neq 0$ . Выбирая  $c = \omega^{-1}$ , получим разностную схему  $2k$ -го порядка аппроксимации. Иногда возникает необходимость построить явную схему ( $b_0 = 0$ ). Решая систему уравнений  $b_0 = 0, E_0 = \dots = E_{2k-1} = 0$  и выбирая решение  $\sum_{i=0}^k b_{-i} = 1$ , получаем схему  $(2k - 1)$ -го порядка аппроксимации.

Примеры. При  $k = 2$  общее решение системы уравнений  $E_0 = \dots = E_4 = 0$  имеет вид  $a_0 = c, a_{-1} = 0, a_{-2} = -c, b_0 = b_{-2} = \frac{c}{3}, b_{-1} = \frac{4}{3}c$ . Из условия (8) получаем  $c = 1/2$ ; расчетная схема имеет вид

$$\frac{y_n - y_{n-2}}{2h} - \left( \frac{1}{6} f_n + \frac{2}{3} f_{n-1} + \frac{1}{6} f_{n-2} \right) = 0.$$

Заметим, что выше эта схема была получена из формулы Симпсона.

Если потребовать выполнения равенств  $b_0 = E_0 = E_1 = E_2 = \dots = E_3 = 0$ , то получим схему

$$\frac{y_n + 4y_{n-1} - 5y_{n-2}}{6h} - \left( \frac{2}{3} f_{n-1} + \frac{1}{3} f_{n-2} \right) = 0. \quad (9)$$

Отметим следующее обстоятельство. Если равенства

$$y'(x_n) \approx \sum_{i=0}^k \frac{a_{-i} y(x_{n-i})}{h}, \quad (10)$$

$$f(x_n, y(x_n)) \approx \sum_{i=0}^k b_{-i} f(x_{n-i}, y(x_{n-i})) \quad (11)$$

выполняются с точностью до членов порядка  $O(h^m)$ , то величина  $r_n$ , равная разности погрешностей этих соотношений, имеет порядок  $O(h^m)$  и, следовательно, согласно (3)  $E_0 = \dots = E_m = 0$ . Однако для выполнения условия  $r_n = O(h^m)$  не обязательно требовать, чтобы погрешности приближенных соотношений (10), (11) имели порядок  $O(h^m)$ . Например, для последней схемы  $r_n = O(h^3)$ , в то время как погрешности соотношений (10) порядка  $O(h)$ .

Кроме построенных выше методов типа Рунге — Кутта и конечно-разностных методов следует отметить группу методов, где при нахождении каждого нового значения  $y_n$  используется несколько предшествующих значений  $y_{n-i}$ , как в конечно-разностных методах, но в то же время на каждом шаге производится несколько вычислений правой части, как в методах Рунге — Кутта.

Пример методов этой группы:

$$y_{n-1/2} = y_{n-2} + \frac{h}{8} (9f_{n-1} + 3f_{n-2}),$$

$$y_n^1 = \frac{1}{5} (28y_{n-1} - 23y_{n-2}) + \frac{h}{15} (32f_{n-1/2} - 60f_{n-1} - 26f_{n-2}),$$

$$y_n = \frac{1}{31} (32y_{n-1} - y_{n-2}) + \frac{h}{93} (64f_{n-1/2} + 15f_n^1 + 12f_{n-1} - f_{n-2}),$$

погрешность на шаге порядка  $O(h^6)$ ; здесь  $f_m^k = f(x_m, y_m^k)$ .

## § 7. Исследование свойств конечно-разностных методов на модельных задачах

После конструирования нового метода решения задачи, например метода решения дифференциальных уравнений, целесообразно прежде чем писать программу посмотреть, как будет работать этот метод на простейших модельных задачах, где точное и приближенное решения вычисляются в явном виде. Если для такой задачи метод дает неудовлетворительный результат, то от применения этого метода, скорее всего, стоит отказаться.

Часто первоначально конструируется не один метод, а некоторое семейство методов, зависящих от одного или нескольких параметров. Изучение модельного примера может позволить сравнить эти методы и выбрать оптимальные значения параметров. На примере решаемой в явном виде задачи можно разыграть реальную ситуацию, возникающую при реализации метода.

Такой подход часто более предпочтителен, чем подробное теоретическое исследование, поскольку дает большой выигрыш во времени.

Из формулы (6.4) можно находить значение  $y_n$ , если известны значения  $y_{n-k}, \dots, y_{n-1}$ ; поэтому, чтобы начать вычисления, нужно знать  $y_j$  в  $k$  начальных точках  $y_0, y_1, \dots, y_{k-1}$ . Они могут быть найдены каким-либо образом заранее, например с помощью формулы Тейлора или методом Рунге — Кутта.

Рассмотрим вопрос о том, насколько влияют погрешности в начальных данных разностной задачи  $y_0, \dots, y_{k-1}$  на её решение.

Пусть  $y_n^s$  ( $s=1, 2$ ) — решения разностной задачи

$$\sum_{i=0}^k a_{-i} y_{n-i} - h \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) = 0 \quad (1)$$

при начальных данных  $y_0^s, \dots, y_{k-1}^s$  соответственно. На основании формулы Лагранжа имеем

$$f(x_m, y_m^2) - f(x_m, y_m^1) = l_m \varepsilon_m;$$

здесь  $l_m = f_y(x_m, \tilde{y}_m)$ ,  $\tilde{y}_m$  лежит между  $y_m^1$  и  $y_m^2$ ,  $\varepsilon_m = y_m^2 - y_m^1$ .

Вычтем из соотношения (1) при  $j=2$  то же соотношение при  $j=1$ . Получим уравнение относительно разности  $\varepsilon_m$ :

$$\sum_{i=0}^k (a_{-i} - h b_{-i} l_{n-i}) \varepsilon_{n-i} = 0. \quad (2)$$

Оказывается, что довольно существенную информацию о поведении погрешности можно получить, рассматривая простейшее дифференциальное уравнение  $y' = 0$ . В этом случае (2) превращается в уравнение с постоянными коэффициентами

$$\sum_{i=0}^k a_{-i} \varepsilon_{n-i} = 0. \quad (3)$$

Соответствующее характеристическое уравнение имеет вид

$$F_0(\mu) = \sum_{i=0}^k a_{-i} \mu^{k-i} = 0. \quad (4)$$

Пусть  $\mu_1$  — максимальный по модулю из корней этого уравнения. Поскольку  $E_0 = F_0(1)$ , то условие  $E_0 = 0$  равносильно тому, что  $\mu = 1$  является корнем (4). Поэтому  $|\mu_1| \geq 1$ . Сеточная функция  $\varepsilon_n = \text{const} \cdot \mu_1^n$  является решением уравнения (4).

Пусть  $\mu_1$  вещественно. Рассмотрим ситуацию, когда  $\varepsilon_n = y_n^2 - y_n^1 = \delta \mu_1^{n-k+1}$ ,  $|\mu_1| > 1$ . Разность между значениями решений  $y_n^1$  и  $y_n^2$  в начальных точках не превосходит  $\delta$ , в то время как

$$\max_{nh \leq X} |\varepsilon_n| = \delta |\mu_1|^{X/h - k + 1}$$

экспоненциально растет с ростом числа шагов  $N = X/h$ .

Таким образом, в случае  $\mu_1 = \max |\mu_i| > 1$  малые возмущения начальных данных могут приводить к катастрофическому возмущению решения уже при не очень большом числе шагов.

Пусть  $\mu_1$  не вещественно; сеточные функции  $\varepsilon_{n,1} = \text{Re}(\delta \mu_1^{n-k+1})$  и  $\varepsilon_{n,2} = \text{Im}(\delta \mu_1^{n-k+1})$  будут решениями (3). Поскольку

$\max(|x|, |y|) \geq |x + iy|/\sqrt{2}$ , то

$$\max\left(\max_{nh \leq X} |\varepsilon_{n,1}|, \max_{nh \leq X} |\varepsilon_{n,2}|\right) \geq \frac{\delta}{\sqrt{2}} |\mu_1|^{X/h-k+1}.$$

Таким образом, и в случае  $\mu_1$  не вещественного,  $|\mu_1| > 1$ , решение уравнения (1) может сильно исказиться при малом возмущении начальных данных.

Рассмотрим случай, когда все корни (4) не превосходят по модулю 1, но среди корней, по модулю равных 1, есть  $p$ -кратный корень  $\mu_1$ ,  $|\mu_1| = 1$ ,  $p > 1$ . Для простоты проведем построения для случая  $\mu_1$  вещественного. Сеточная функция  $\varepsilon_n = \delta \mu_1^{n-k+1} \left(\frac{n}{k-1}\right)^{p-1}$  соответствует возмущению начальных данных  $y_0, \dots, y_{k-1}$  не более чем на  $\delta$ , в то время как в конце отрезка возмущение имеет порядок  $\delta(X/h)^{p-1}$ .

Такой степенной (по отношению к числу узлов) рост влияния возмущения начальных данных иногда является допустимым. Однако на примере модельного уравнения  $y' = My$  можно показать, что в случае  $p$ -кратного корня на границе единичного круга возмущение исходных данных сказывается более существенным образом. При  $f(x, y) = My$  все  $l_m = M$  и уравнение (3) относительно  $\varepsilon_n$  является линейным разностным уравнением

$$\sum_{i=0}^k (a_{-i} - Mhb_{-i}) \varepsilon_{n-i} = 0.$$

Соответствующее характеристическое уравнение имеет вид

$$\sum_{i=0}^k (a_{-i} - Mhb_{-i}) \mu^{k-i} = 0. \quad (5)$$

Наложим ограничение,

$$\sum_{i=0}^k b_{-i} \mu_1^{k-i} \neq 0.$$

В противном случае разностное уравнение (1) записывается в виде

$$\left( \sum_{i=0}^{k-1} c_{-i} y_{n-i} - h \sum_{i=0}^{k-1} d_{-i} f(x_{n-i}, y_{n-i}) \right) - \mu_1 \left( \sum_{i=0}^{k-1} c_{-i} y_{n-1-i} - h \sum_{i=0}^{k-1} d_{-i} f(x_{n-1-i}, y_{n-1-i}) \right) = 0.$$

Рассмотрение таких уравнений не представляет интереса, поскольку решение более простого уравнения

$$\sum_{i=0}^{k-1} c_{-i} y_{n-i} - h \sum_{i=0}^{k-1} d_{-i} f(x_{n-i}, y_{n-i}) = 0$$

также оказывается решением (5).

Предположим также, что  $a_0 \neq 0$ . Справедлива  
**Теорема** (без доказательства). 1. Если  $p > 2$ , то среди  
 корней (5) есть корень, удовлетворяющий неравенству

$$|\mu_1(Mh)| \geq \exp\{c|Mh|^{1/p}\}, \quad c > 0.$$

2. Если  $p = 2$ , то или при  $M > 0$ , или при  $M < 0$  среди корней  
 (5) есть корень, удовлетворяющий неравенству

$$|\mu_1(Mh)| \geq \exp\{c|Mh|^{1/p}\}.$$

Таким образом, при  $p \geq 2$  или уравнению  $y' = +|M|y$ , или  
 уравнению  $y' = -|M|y$  соответствует рост возмущения решения в

$$\approx \exp\left\{c|Mh|^{1/p} \frac{X}{h}\right\} = \exp\{c|M|^{1/p} Xh^{1/p-1}\} \text{ раз.}$$

Возмущение решения растет быстрее любой степени числа ша-  
 гов; такой рост возмущения уже при небольшом числе шагов  
 также является недопустимым.

В связи со сказанным практически пригодными могут оказы-  
 ваться лишь схемы, удовлетворяющие следующему условию  $\alpha$ :  
*все корни характеристического уравнения (4) лежат в единич-  
 ном круге и на границе единичного круга нет кратных корней.*

Можно было бы подумать, что все дело только в округлениях  
 и погрешностях исходных данных: если бы их не было, то, мо-  
 жет быть, решение конечно-разностной задачи сходилось бы к  
 решению дифференциальной? На самом деле для любой раз-  
 ностной схемы, не удовлетворяющей условию  $\alpha$ , можно построить  
 пример дифференциального уравнения с бесконечно дифферен-  
 цируемой правой частью, для которого и при отсутствии округ-  
 лений и погрешностей в исходных данных решение конечно-раз-  
 ностной задачи не стремится к решению дифференциальной при  
 измельчении шага.

На первый взгляд может показаться целесообразным строить  
 схемы с возможно большим порядком аппроксимации  $m$ ,  
 ( $E_0 = \dots = E_m = 0$ ). Однако оказывается, что все схемы с боль-  
 шим  $m$  не удовлетворяют условию  $\alpha$ .

**Теорема.** В случаях: а) схема явная,  $m > k$ ; б) схема неяв-  
 ная,  $k$  нечетно,  $m > k + 1$ ; в) схема неявная,  $k$  четно,  
 $m > k + 2$ , среди корней характеристического уравнения (4)  
 имеется корень, по модулю больший 1.

Далее будет показано, что при некоторых дополнительных  
 условиях на погрешности начальных данных разностной задачи  
 и вычислительную погрешность при выполнении условия  $\alpha$  реше-  
 ние разностной задачи (1) сходится к решению дифферен-  
 циальной задачи. Будет приведено выражение главного члена  
 погрешности, из которого видно, что главный член ведет себя  
 примерно одинаково для всех разностных схем одного и того же  
 порядка точности, удовлетворяющих условию  $\alpha$ . Однако это не

означает, что на практике они являются примерно эквивалентными.

Рассмотрим поведение решений двух разностных схем второго порядка аппроксимации:

$$1) \quad \frac{y_n - y_{n-1}}{h} = \frac{f(x_n, y_n) + f(x_{n-1}, y_{n-1})}{2}, \quad (6)$$

$$2) \quad \frac{y_n - y_{n-2}}{2h} = f(x_{n-1}, y_{n-1}) \quad (7)$$

на примере модельного уравнения  $y' = My$ ,  $M = \text{const}$ . Разностные схемы (6), (7) порождают конечно-разностные уравнения

$$\begin{aligned} y_n(1 - Mh/2) - y_{n-1}(1 + Mh/2) &= 0, \\ y_n - 2Mhy_{n-1} - y_{n-2} &= 0. \end{aligned}$$

В первом случае решение уравнения для погрешности имеет вид

$$\varepsilon_n = \left( \frac{1 + Mh/2}{1 - Mh/2} \right)^n \varepsilon_0;$$

это решение растет при  $M > 0$ . Это естественно, поскольку для дифференциальной задачи разность двух решений  $\varepsilon(x)$  с различными начальными условиями записывается в виде

$$\varepsilon(x_n) = \exp\{M(x_n - x_0)\} \varepsilon(x_0)$$

и также растет при  $M > 0$ . При  $M < 0$  как  $\varepsilon_n$ , так и  $\varepsilon(x_n)$  убывают. Решение второго уравнения имеет вид  $c_1 \mu_1^n + c_2 \mu_2^n$ , где  $\mu_1$  и  $\mu_2$  — корни характеристического уравнения  $\mu^2 - 2Mh\mu - 1 = 0$ , т. е.  $\mu_{1,2} = Mh \pm \sqrt{1 + (Mh)^2}$ . Имеем  $\mu_1 = Mh + \sqrt{1 + (Mh)^2} = 1 +$

$+ Mh + \frac{(Mh)^2}{2} + O((Mh)^3) = \exp\{Mh(1 + O(Mh))\}$ . Здесь и далее

мы пользуемся формулой Тейлора  $\sqrt{1 + \varepsilon} = 1 + \frac{\varepsilon}{2} + O(\varepsilon^2)$ .

Отсюда следует равенство  $\mu_1^n = \exp\{Mnh(1 + O(Mh)^2)\}$ . Таким образом, слагаемое  $c_1 \mu_1^n$  соответствует решению разностного уравнения, ведущему себя качественно так же, как решение дифференциального уравнения. Аналогичным образом получаем

$$\begin{aligned} \mu_2 &= Mh - \sqrt{1 + M^2 h^2} = -1 + Mh - \frac{M^2 h^2}{2} + \\ &+ O((Mh)^3) = -\exp\{-Mh(1 + O(Mh))\}. \end{aligned}$$

Имеем  $\mu_2^n = (-1)^n \exp\{-Mnh(1 + O(Mh))\}$ ; слагаемое  $\mu_2^n$  ведет себя качественно иначе, чем решение дифференциального уравнения, и, что особенно существенно, оно *возрастает по модулю при  $M < 0$ , в то время как точное решение убывает*. Рассуждая так же, как выше, заключаем, что вычислительная погрешность может исказить решение на величину порядка  $\delta \exp\{-Mnh(1 + O(Mh))\}$ . При  $M \leq 0$  и большом значении



$|Mnh|$  эта величина может оказаться недопустимо большой, особенно на фоне убывающего решения  $e^{M(x-x_0)}$ .

Поскольку рассматриваемый метод дает неудовлетворительный результат для такой простейшей модельной задачи, его вряд ли можно рекомендовать для широкого употребления, тем более в стандартных программах численного интегрирования дифференциальных уравнений.

Мы отбраковали второй метод на примере задачи, где  $M < 0$  и величина  $\delta e^{M|x|}$  недопустимо большая. В последние двадцать лет в приложениях очень часто стали встречаться задачи с резкими переходными процессами, где решение существенно меняется на малом промежутке времени. Типичной модельной задачей является задача Коши для уравнения  $y' = My$ ,  $M < 0$ , где величина  $|M|X$  настолько велика, что число шагов порядка  $|M|X$  является недопустимым. Если при разумном числе шагов по времени  $|M|h \gg 1$ , то использование первой из рассматриваемых разностных схем также может привести к неудовлетворительным результатам. Для этой схемы имеем

$$\mu_1 = \frac{1 + Mh/2}{1 - Mh/2} = \frac{1 + 2/(Mh)}{-1 + 2/(Mh)} \approx -\exp\left\{\frac{4}{Mh} + O\left(\frac{1}{(Mh)^2}\right)\right\}.$$

Таким образом, решение разностного уравнения имеет вид

$$\begin{aligned} y_n &= (-1)^n \exp\left\{\left(\frac{4}{Mh} + O\left(\frac{1}{(Mh)^2}\right)\right)n\right\} y_0 = \\ &= (-1)^n \exp\left\{\left(\frac{4}{Mh^2} + O\left(\frac{1}{M^2h^3}\right)\right)(x_n - x_0)\right\} y_0 \end{aligned}$$

и существенно отличается от точного решения дифференциальной задачи  $y(x_n) = y_0 e^{M(x_n - x_0)}$ , например, при  $|M|h^2 \gg 1$ .

В качестве итога проводимого выше анализа свойств первого метода можно заключить, что этот метод применим для решения довольно широкого круга задач. В то же время существуют определенные типы задач, называемые *жесткими* (моделируемые случаем  $M < 0$ ,  $|M|X$  очень велико), в которых к применению этого метода нужно отнестись с определенной осторожностью. Для решения таких задач разработаны специальные методы (см. § 9).

## § 8. Оценка погрешности конечно-разностных методов

Произведем оценку погрешности приближенных решений, которые получаются при использовании конечно-разностных методов вида (5.2), удовлетворяющих условию  $\alpha$ . Получаемые в процессе реальных вычислений величины  $y_n$ , являющиеся приближениями к значениям  $y(x_n)$ , связаны на самом деле не соотношением (5.2), а соотношением

$$\sum_{i=0}^k a_{-i} y_{n-i} - h \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) = \delta_n, \quad (1)$$

где величина  $\delta_n$  может быть отлична от нуля (причины появления величины  $\delta_n$  уже были указаны в § 4).

С другой стороны, в § 5, 6 установлено, что значения  $y(x_n)$  точного решения дифференциальной задачи удовлетворяют соотношению

$$\sum_{i=0}^k a_{-i} y(x_{n-i}) - h \sum_{i=0}^k b_{-i} f(x_{n-i}, y(x_{n-i})) = h r_n; \quad (2)$$

при соответствующем подборе коэффициентов  $a_{-i}$  и  $b_{-i}$  оказывалось, что  $r_n = O(h^m)$ , где  $m > 0$ . Вычитая (2) из (1), получим уравнения для погрешности  $R_n = y_n - y(x_n)$ . На основании формулы Лагранжа имеем равенство

$$f(x_{n-i}, y_{n-i}) - f(x_{n-i}, y(x_{n-i})) = l_{n-i} R_{n-i}, \quad (3)$$

где  $l_{n-i} = f_y(x_{n-i}, \tilde{y}_{n-i})$  и  $\tilde{y}_{n-i}$  лежит между  $y_{n-i}$  и  $y(x_{n-i})$ . С учетом (3) разность соотношений (1) и (2) запишется в виде

$$\sum_{i=0}^k a_{-i} R_{n-i} - h \sum_{i=0}^k b_{-i} l_{n-i} R_{n-i} = g_n, \quad (4)$$

где  $g_n = \delta_n - h r_n$ .

*Теорема (об оценке погрешности).* Пусть разностная схема удовлетворяет условию  $\alpha$  и  $|f_y| \leq L$  при  $x_0 \leq x \leq x_0 + X$ . Тогда при  $x_0 \leq x_n \leq x_0 + X$  выполняется неравенство

$$|R_n| \leq c(L, X) \left( \max_{0 \leq i < k} |R_i| + \sum_{j=k}^n |g_j| \right), \quad (5)$$

где  $c(L, X) < \infty$  — некоторая постоянная, зависящая от коэффициентов  $a_{-i}$ ,  $b_{-i}$  и от  $L$  и  $X$ .

*Доказательство.* Сформулируем частный случай леммы из § 3 гл. VI: пусть все собственные значения матрицы  $A$  лежат в круге  $|\lambda| \leq q$  и на границе круга нет кратных корней; тогда можно указать матрицу  $C$  такую, что  $\|D\|_\infty \leq q$ , где  $D = C^{-1}AC$ .

Для удобства оценки преобразуем уравнение (4) в одношаговое векторное уравнение. В дальнейшем для определенности предполагаем  $h$  настолько малым, что  $|hb_0 l_n| \leq |a_0/2|$ , тогда коэффициент  $a_0 - hb_0 l_n$  при  $R_n$  в (4) по модулю не менее  $|a_0/2|$ . Переносим в (4) все слагаемые, не содержащие  $R_n$ , в правую часть и делим на коэффициент при  $R_n$ , получим равенство

$$R_n = \sum_{i=1}^k \frac{-a_{-i} + hb_{-i} l_{n-i}}{a_0 - hb_0 l_n} R_{n-i} + \frac{g_n}{a_0 - hb_0 l_n}. \quad (6)$$

Положим

$$\frac{-a_{-i} + hb_{-i} l_{n-i}}{a_0 - hb_0 l_n} - \left( \frac{-a_{-i}}{a_0} \right) = h v_{in},$$

$$v_{in} = \frac{b_{-i} a_0 l_{n-1} - a_{-i} b_0 l_n}{(a_0 - hb_0 l_n) a_0}, \quad |v_{in}| \leq 2 \frac{|b_{-i} a_0| + |a_{-i} b_0|}{|a_0|^2} L.$$

Введем в рассмотрение векторы

$$\mathbf{Z}_n = (R_n, \dots, R_{n-k+1})^T.$$

Соотношение (6) равносильно равенству

$$\mathbf{Z}_n = A\mathbf{Z}_{n-1} + hV_n\mathbf{Z}_{n-1} + \mathbf{W}_n, \quad (7)$$

где

$$V_n = \begin{bmatrix} v_{1n} & \dots & v_{kn} \\ 0 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & & 0 \end{bmatrix}, \quad \mathbf{W}_n = \begin{pmatrix} \frac{g_n}{a_0 - hb_0 l_n} \\ 0 \\ \dots \\ 0 \end{pmatrix},$$

$$A = \begin{bmatrix} -\frac{a_{-1}}{a_0} & -\frac{a_{-2}}{a_0} & \dots & -\frac{a_{1-k}}{a_0} & -\frac{a_{-k}}{a_0} \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

Действительно, приравняв первые компоненты векторов в правой и левой частях (7), мы получим равенство (6), а приравняв остальные компоненты, получаем тождества

$$R_{n-i} = R_{n-i}, \quad i = 1, \dots, k-1.$$

Вычислим характеристический многочлен матрицы  $A$ :

$$P(\lambda) = \det(A - \lambda E) =$$

$$= \det \begin{bmatrix} -\frac{a_{-1}}{a_0} - \lambda & -\frac{a_{-2}}{a_0} & \dots & -\frac{a_{1-k}}{a_0} & -\frac{a_{-k}}{a_0} \\ 1 & -\lambda & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & -\lambda \end{bmatrix}.$$

Для этого умножим первый столбец на  $\lambda$  и прибавим ко второму, затем — второй на  $\lambda$  и прибавим к третьему и т. д. В результате получим

$$P(\lambda) = \det \begin{bmatrix} p_1(\lambda) & p_2(\lambda) & \dots & p_{k-1}(\lambda) & p_k(\lambda) \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 1 & 0 \end{bmatrix},$$

$$\text{где } p_1(\lambda) = -\frac{a_{-1}}{a_0} - \lambda, \quad p_2(\lambda) = -\frac{a_{-2}}{a_0} + \lambda \left( -\frac{a_{-1}}{a_0} - \lambda \right), \dots$$

$$\dots, \quad p_k(\lambda) = -\frac{a_{-k}}{a_0} + \lambda \left( -\frac{a_{1-k}}{a_0} + \lambda \left( -\frac{a_{2-k}}{a_0} + \dots \right. \right. \\ \left. \left. \dots + \lambda \left( -\frac{a_{-1}}{a_0} - \lambda \right) \dots \right) \right).$$

Раскрывая определитель по последнему столбцу, имеем  $P(\lambda) = (-1)^{k+1} p_k(\lambda)$  или, что то же самое,

$$\begin{aligned} (-1)^k P(\lambda) &= \lambda^k + \frac{a_{-1}}{a_0} \lambda^{k-1} + \dots + \frac{a_{-k}}{a_0} = \\ &= \frac{1}{a_0} (a_0 \lambda^k + a_{-1} \lambda^{k-1} + \dots + a_{-k}). \end{aligned}$$

Характеристическое уравнение матрицы  $A$  оказалось пропорциональным характеристическому уравнению (7.4) разностной схемы. Согласно предположению  $\alpha$  все корни характеристического уравнения матрицы  $A$  лежат в круге  $|z| \leq 1$  и на границе круга нет кратных. Поэтому по отношению к матрице  $A$  условие леммы выполнено со значением  $q = 1$ . Следовательно, существует матрица  $C$  такая, что  $C^{-1}AC = D$  и  $\|D\|_\infty \leq 1$ . Произведем в уравнении (7) замену переменных  $Z_n = CZ_n$ . После умножения слева на  $C^{-1}$  оно приведет к виду

$$z_n = Dz_{n-1} + hv_n z_{n-1} + w_n, \quad (8)$$

где

$$D = C^{-1}AC, \quad v_n = C^{-1}V_n C, \quad w_n = C^{-1}W_n.$$

В матрице  $V_n$  ненулевые элементы находятся только в первой строке, поэтому

$$\|V_n\|_\infty \leq \left( 2 \sum_{i=1}^k \frac{|b_{-i} a_0| + |a_{-i} b_0|}{|a_0|^2} \right) L = vL$$

и

$$\|v_n\|_\infty \leq \|C^{-1}\|_\infty \|V_n\|_\infty \|C\|_\infty \leq vL \|C^{-1}\|_\infty \|C\|_\infty.$$

Имеем

$$\|w_n\|_\infty \leq \|C^{-1}\|_\infty \|W_n\|_\infty = \|C^{-1}\|_\infty \left| \frac{g_n}{a_0 - hb_0 l_n} \right| \leq 2 \|C^{-1}\|_\infty \frac{|g_n|}{|a_0|}.$$

После оценки норм слагаемых в правой части (8) можно записать неравенство

$$\|z_n\|_\infty \leq \beta |g_n| + (1 + \gamma Lh) \|z_{n-1}\|_\infty, \quad (9)$$

где

$$\beta = 2 \frac{\|C^{-1}\|_\infty}{|a_0|}, \quad \gamma = v \|C^{-1}\|_\infty \|C\|_\infty.$$

Для оценки  $\|z_n\|_\infty$  через  $\|z_{k-1}\|_\infty$  и величины  $|g_n|$  поступим следующим образом. Выпишем неравенства (9) при  $j = n, \dots, k$ . В правой части оценки  $\|z_n\|_\infty$  через  $\|z_{n-1}\|_\infty$  оценим  $\|z_{n-1}\|_\infty$  через  $\|z_{n-2}\|_\infty$ , затем в правой части получившегося неравенства оценим  $\|z_{n-2}\|_\infty$  через  $\|z_{n-3}\|_\infty$  и т. д. В результате получим

$$\begin{aligned} \|z_n\|_\infty &\leq \beta |g_n| + (1 + \gamma Lh) (\beta |g_{n-1}| + (1 + \gamma Lh) (\beta |g_{n-2}| + \dots \\ &\dots + (1 + \gamma Lh) (\beta |g_k| + (1 + \gamma Lh) \|z_{k-1}\|_\infty) \dots)), \end{aligned}$$

или, что то же самое,

$$\|z_n\|_\infty \leq \beta \sum_{j=k}^n (1 + \gamma Lh)^{n-j} |g_j| + (1 + \gamma Lh)^{n-k+1} \|z_{k-1}\|_\infty. \quad (10)$$

Упростим эту оценку, одновременно несколько закругив ее. При  $x_0 \leq x_j \leq x_n \leq x_0 + X$  имеем  $(n-j)h \leq X$ ; поэтому

$$(1 + \gamma Lh)^{n-j} \leq \exp \{ \gamma Lh(n-j) \} \leq \exp \{ \gamma LX \}.$$

Теперь из (10) получаем

$$\|z_n\|_\infty \leq \exp \{ \gamma LX \} \left( \beta \sum_{j=k}^n |g_j| + \|z_{k-1}\|_\infty \right). \quad (11)$$

Справедливы неравенства

$$|R_n| \leq \|Z_n\|_\infty \leq \|C\|_\infty \|z_n\|_\infty,$$

$$\|z_{k-1}\|_\infty \leq \|C^{-1}\|_\infty \|Z_{k-1}\|_\infty = \|C^{-1}\|_\infty \max_{0 \leq i < k} |R_i|,$$

и поэтому

$$\|Z_n\|_\infty \leq \|C\|_\infty \exp \{ \gamma LX \} \left( \beta \sum_{j=k}^n |g_j| + \|C^{-1}\|_\infty \|Z_{k-1}\|_\infty \right). \quad (12)$$

Далее получаем оценку погрешности

$$|R_n| \leq \exp \{ \gamma LX \} \left( M_1 \sum_{j=k}^n |g_j| + M_2 \max_{0 \leq i < k} |R_i| \right),$$

где  $M_1 = \beta \|C\|_\infty$ ,  $M_2 = \|C\|_\infty \|C^{-1}\|_\infty$  и  $\gamma$  — некоторые постоянные, зависящие только от коэффициентов  $a_{-i}$ ,  $b_{-i}$  исходной разностной схемы. В частности,  $M_1$  и  $M_2$  зависят только от коэффициентов  $a_{-i}$ . Утверждение теоремы доказано.

Подставляя в (12) значение  $g_j = \delta_j - hr_j$ , получим искомую оценку погрешности

$$|R_n| \leq \exp \{ \gamma LX \} \left( M_1 \sum_{j=k}^n (|\delta_j| + h|r_j|) + M_2 \max_{0 \leq i < k} |R_i| \right). \quad (13)$$

Из оценки (13) видно, что для сходимости решения разностного уравнения к решению дифференциального достаточно выполнения условий

$$\sum_{j=k}^n |\delta_j| \rightarrow 0, \quad h \sum_{j=k}^n |r_j| \rightarrow 0, \quad \max_{0 \leq i < k} |R_i| \rightarrow 0.$$

Оценка погрешности (13) во многих случаях является существенно завышенной. Например, для методов Адамса можно получить оценку погрешности, остающуюся ограниченной в случае  $f_y \leq -b < 0$  при сколь угодно большой длине промежутка интегрирования в предположении, что погрешности округления  $\delta_j$  и погрешности аппроксимации  $r_j$  равномерно ограничены:

$|\delta_j| \leq \delta$ ,  $|r_j| \leq r$ . Заметим, что из оценки погрешности для одношаговых методов (§ 4) следует оценка погрешности метода Эйлера, являющегося частным случаем методов Адамса. В то же время из (13) при таких предположениях нельзя получить равномерной ограниченности погрешности интегрирования. Для получения более реального представления о величине погрешности полезно располагать выражением для главного члена погрешности.

Наметим путь получения этого выражения. При достаточно гладкой функции  $f(x, y)$  согласно (6.3) справедливо равенство

$$r_n = E_{m+1} h^m y^{(m+1)}(x_n) + o(h^m).$$

Поскольку  $\sum_{i=0}^k b_{-i} = E_1 = 1$ , то это выражение можно переписать в более удобном для нас виде

$$r_n = h^m \sum_{i=0}^k b_{-i} E_{m+1} y^{(m+1)}(x_{n-i}) + o(h^m).$$

Предположим, что вычислительная погрешность мала по сравнению с погрешностью аппроксимации, точнее,  $\max_j |\delta_j| = \delta = o(h^{m+1})$ .

При выполнении условий теоремы об оценке погрешности решение разностной задачи сходится к решению дифференциальной, поэтому  $l_n = f_y(x, \tilde{y}_n) = f_y(x_n, y(x_n)) + o(1)$ . С учетом выписанных выше соотношений равенство (4) можно переписать в виде

$$L_h^0 \left( \frac{R_n}{h^m} \right) = \frac{\sum_{i=0}^k a_{-i} \frac{R_{n-i}}{h^m}}{h} - \sum_{i=0}^k b_{-i} (f_y(x_{n-i}, y(x_{n-i})) \frac{R_{n-i}}{h^m} - E_{m+1} y^{(m+1)}(x_{n-i})) = o(1).$$

Таким образом, сеточная функция  $z_n = R_n/h^m$  приближенно удовлетворяет сеточному уравнению  $L_h^0(z_n) = 0$ , которое получается при аппроксимации уравнения

$$z' - (f_y(x, y(x)z - E_{m+1} y^{(m+1)}(x))) = 0. \quad (14)$$

В предположении, что  $z_0, \dots, z_{k-1} = o(1)$ , рассуждая так же, как при доказательстве теоремы об оценке погрешности, получаем, что  $z_n$  близко к решению уравнения (14) при начальном условии  $z(x_0) = 0$ . Это решение можно выписать в явном виде:

$$z_n \approx z(x_n) = -E_{m+1} \int_{x_0}^{x_n} \exp \left\{ \int_x^{x_n} f_y(t, y(t)) dt \right\} y^{(m+1)}(x) dx. \quad (15)$$

Таким образом,

$$R_n \approx h^m z(x_n).$$

Сформулируем аналогичный результат, относящийся к случаю интегрирования системы уравнений. Для системы уравнений  $y' = f(x, y)$ , когда  $y$  и  $f$  — векторы:  $y = (y_1, \dots, y_l)^T$ ,  $f = (f_1, \dots, f_l)^T$ , выражение главного члена погрешности (15) имеет вид

$$y_n - y(x_n) \sim h^m z(x_n),$$

$$z(x_n) = \int_{x_0}^{x_n} \exp \left\{ \int_x^{x_n} f_y(t, y(t)) dt \right\} y^{(m+1)}(x) dx.$$

Здесь  $\exp \left\{ \int_a^b f_y(t, y(t)) dt \right\}$  — матрица  $W(a, b)$ , являющаяся решением матричного дифференциального уравнения

$$W'_b(a, b) = \mathcal{F}_y(b, y(b)) W_b(a, b)$$

при начальном условии  $W(a, a) = E$ , где  $E$  — единичная матрица,  $\mathcal{F}_y(x, u)$  — матрица с элементами  $\partial f_i / \partial y_j |_{x, u}$ ,  $i, j = 1, \dots, l$ .

Проведенные выше рассуждения, каждый этап которых может быть строго обоснован, иногда облекают в более грубую форму. Точное решение дифференциальной задачи удовлетворяет соотношению

$$L_h(y(nh)) \approx E_{m+1} h^m y^{(m+1)}(x),$$

где

$$L_h z_n = \frac{\sum_{i=0}^k a_{n-i} z_{n-i}}{h} - \sum_{i=0}^k b_{-i} f(x_{n-i}, z_{n-i}),$$

а приближенное решение — соотношению

$$L_h(y_n) = 0,$$

поэтому их разность удовлетворяет равенству

$$L_h(y_n) - L_h(y(nh)) \approx -E_{m+1} h^m y^{(m+1)}(x).$$

Поскольку  $y_n$  и  $y(nh)$  близки, это соотношение можно записать в виде

$$L'_h(y(nh))(y_n - y(nh)) \approx -E_{m+1} h^m y^{(m+1)}(x);$$

здесь  $L'_h$  — производная оператора  $L_h$ . Напомним, что поскольку операторы  $L$  и  $L_h$  в определенном смысле близки, то можно написать

$$L'(y(nh))(y_n - y(nh)) \approx -E_{m+1} h^m y^{(m+1)}(x).$$

Известно, что производная оператора  $L$  определяется равенством

$$L'(y(x)) \delta = \lim_{t \rightarrow 0} \frac{L(y(x) + t\delta(x)) - L(y(x))}{t} =$$

$$= \lim_{t \rightarrow 0} \frac{((y + t\delta)' - f(x, y + t\delta)) - (y' - f(x, y))}{t} = \delta' - f_y(x, y(x)) \delta.$$

Отсюда получаем приближенное равенство

$$\delta' - \int_y(x, y(x)) \delta \approx -E_{m+1} h^m y^{(m+1)}(x),$$

где  $\delta = y_n - y(x_n)$ , и затем — (15).

Второй путь вывода выражения для главного члена погрешности уже не поддается непосредственному обоснованию и в принципе может привести к неверным заключениям. Это видно хотя бы из того, что нигде не проявилось условие  $\alpha$ , без которого не имеет места сам факт малости величины  $y_n - y(x_n)$ . Справедливость получаемых на таком пути результатов требует специального обоснования.

В то же время его следует признать крайне полезным, поскольку не известно ни одного противоречащего примера, когда бы его применение приводило к неправильному выражению для главного члена погрешности в случае, если решение разностной задачи сходится к решению дифференциальной.

Для некоторых методов, например Адамса и Рунге — Кутта, выражение главного члена погрешности, как правило, дает реальное представление о величине погрешности. В других случаях, например для метода (7.6), являющегося простейшим случаем так называемого метода Милна, это выражение следует рассматривать лишь как некоторую оценку снизу для реальной величины погрешности.

Рассмотрим случай метода (7.6) и уравнения  $y' = My$ . Точное решение имеет вид

$$y(x) = y_0 e^{M(x-x_0)},$$

а  $E_3 = 1/6$ . Согласно (15) имеем

$$\begin{aligned} z(x_n) &= -\frac{1}{6} \int_{x_0}^{x_n} e^{M(x-x_0)} M^3 y_0 e^{M(x-x_0)} dx = \\ &= -\frac{1}{6} M^3 y_0 e^{M(x_n-x_0)} (x_n - x_0). \end{aligned} \quad (16)$$

Оценим  $|ve^{-v}|$  в области  $v > 0$ . Поскольку эта функция стремится к нулю при  $v \rightarrow 0$  и при  $v \rightarrow \infty$ , то ее наибольшее значение принимается в точке, где ее производная равна нулю, т. е. при  $v = 1$ . Отсюда следует, что  $|ve^{-v}| \leq e^{-1}$  при  $v > 0$ . Рассмотрим случай  $M < 0$ . Подставляя  $v = -M(x_n - x_0)$ , получим  $|Me^{M(x_n-x_0)}(x_n - x_0)| \leq e^{-1}$ . Отсюда и из (16) следует  $|z(x_n)| \leq M^2 |y_0| / (6e)$ , и, таким образом, главный член погрешности равномерно ограничен при  $x_0 < x_n < \infty$ . В то же время из рассмотрения этого модельного примера, проведенного в § 7, вытекает, что реальная величина погрешности довольно сильно растет вследствие большого влияния погрешности начальных значений  $y_0, y_1$ .

Казалось бы, в обрисованной выше ситуации есть какое-то противоречие. Говорится о главном члене погрешности, но в то же время утверждается, что он не является определяющим в реальной величине погрешности. Дело заключается в следующем.



При получении главного члена погрешности имелось в виду, что длина промежутка интегрирования фиксирована, а  $\delta/h$  и  $h$  стремятся к нулю. При рассмотрении модельного примера в § 7 речь шла о поведении погрешности при  $h$  фиксированном,  $x_n - x_0 \rightarrow 0$ .

Как показало рассмотрение этого модельного примера, влияние погрешности исходных данных существенно уже не при очень больших значениях  $|M|(x_n - x_0)$ , поэтому широкое применение метода (7.6) в реальной практике является нецелесообразным, несмотря на малое значение главного члена погрешности при  $\delta/h, h \rightarrow 0$  и  $x_n - x_0$  фиксированном.

## § 9. Особенности интегрирования систем уравнений

Проведившиеся выше построения, в частности, расчетные формулы, применимы без всяких изменений в случае систем уравнений

$$y' = f(x, y). \quad (1)$$

Формальное отличие состоит в том, что в соответствующих соотношениях вместо скалярных величин участвуют некоторые матрицы или тензоры.

Для выявления особенностей, которые могут возникнуть при численном интегрировании систем обыкновенных дифференциальных уравнений, рассмотрим модельный пример линейной системы с постоянными коэффициентами

$$y' = Ay. \quad (2)$$

В случае использования конечно-разностной аппроксимации (8.1) соответствующая система конечно-разностных уравнений имеет вид

$$\sum_{i=0}^k \frac{a_{-i} y_{n-i}}{h} - \sum_{i=0}^k b_{-i} A y_{n-i} = 0. \quad (3)$$

Для простоты предположим, что жорданова форма матрицы простая:  $C^{-1}AC = \Lambda$ ,  $\Lambda$  — диагональная матрица с диагональными элементами  $\lambda_1, \dots, \lambda_l$ . Положим  $C^{-1}y_n = z_n$  и домножим систему (3) слева на  $C^{-1}$ . Получим

$$\sum_{i=0}^k \frac{a_{-i} z_{n-i}}{h} - \sum_{i=0}^k b_{-i} C^{-1} A C z_{n-i} = 0$$

или

$$\sum_{i=0}^k \frac{a_{-i} z_{n-i}}{h} - \sum_{i=0}^k b_{-i} \Lambda z_{n-i} = 0.$$

Эта система распадается на систему скалярных конечно-разностных уравнений относительно компонент  $z_{pn}$  векторов  $z_n =$

$$= (z_{1n}, \dots, z_{ln})^T:$$

$$\frac{\sum_{i=0}^k a_{-i} z_{p, n-i}}{h} - \sum_{i=0}^k b_{-i} \lambda_p z_{p, n-i} = 0; \quad p = 1, \dots, l. \quad (4)$$

Соотношение (4) совпадает с конечно-разностной аппроксимацией для уравнения

$$z'_p = \lambda_p z_p. \quad (5)$$

Если в (2) перейти к новой неизвестной вектор-функции  $z = C^{-1}y$  и умножить (2) слева на матрицу  $C^{-1}$ , то получится система скалярных уравнений (5),  $p = 1, \dots, l$ . В соответствии с определением вектор-функций  $z(x)$  и  $z_n$  имеем равенство

$$z_n - z(x_n) = C^{-1}(y_n - y(x_n)).$$

Следовательно, для получения решения  $y(x)$  с малой погрешностью необходимо и достаточно, чтобы решения  $z_p(x)$  уравнения (5) получались с малой погрешностью в случае интегрирования с помощью аппроксимации (4). При этом имеется в виду, что есть соответствие между аппроксимациями начальных условий

$$y_j = Cz_j, \quad j = 0, \dots, k-1.$$

Проводя аналогичные построения, можно получить тот же вывод и по отношению к методам Рунге — Кутты.

Решение уравнения (5) имеет вид  $z_p = z_p^0 \exp\{\lambda_p(x - x_0)\}$  и существенно изменяется при изменении  $x$  на расстояние  $\Delta x = 1/|\lambda_p|$ , т. е. характерный размер изменения решения порядка  $1/|\lambda_p|$ . Если говорить о векторе  $z(x)$  как о едином целом, то характерный размер его изменения — величина порядка  $\frac{1}{\max_p |\lambda_p|}$ ; точно такой же порядок характерного изменения будет и у вектора  $y(x)$ .

Шаг интегрирования должен быть существенно меньше характерного размера изменения решения, т. е.  $h \ll \frac{1}{\max_p |\lambda_p|}$ .

Отсюда следует оценка снизу для числа шагов интегрирования

$$N = \frac{X}{h} \gg \max_p |\lambda_p| \cdot X.$$

Если число шагов, много большее величины  $\max_p |\lambda_p| \cdot X$ , приемлемо по затратам машинного времени, то желательно применить методы, использующие специфику поведения решения.

В случае, когда

$$|\lambda_p| X \gg 1 \quad \text{при всех } p,$$

для описания решения можно было бы применить асимптотические методы. Однако на практике часто встречаются задачи, когда это условие не выполнено, и поэтому применение асимптотических методов невозможно или крайне затруднительно.

Конечно, возникает вопрос, о каких проблемах идет речь, поскольку решение системы  $y' = Ay$  выписывается в явном виде? Дело в том, что мы говорим об этой задаче как о модельной; реально же метод применяется для решения какой-то сложной задачи, и мы смотрим, как ведет себя метод в применении к простейшей задаче, где все выписывается в явном виде.

Широкий круг прикладных проблем сводится к решению задачи Коши для так называемых *жестких систем* дифференциальных уравнений. В частности, к таким системам относятся системы уравнений, возникающие при применении методов установления

$$\frac{dx}{dt} + \text{grad } F = 0,$$

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} + \text{grad } F = 0$$

в случае функций  $F$ , у которых линии уровня имеют форму эллипсоидов с большим разбросом полуосей.

В качестве модели таких систем берется система уравнений (2), удовлетворяющая определенным условиям на собственные значения матрицы  $A$ . Не существует установившегося определения жестких систем. Часто систему (2) относят к классу жестких, если величины  $(\max_p (\text{Re } \lambda_p))X$ ,  $(\max_p |\text{Im } \lambda_p|)X$  не являются большими положительными числами, а величина  $(\max_p |\lambda_p|)X \gg 1$ . Нелинейную систему  $y' = f(x, y)$  относят к классу жестких систем, если при всех  $x_0$  из некоторого отрезка длины  $X > 0$ , принадлежащего области интегрирования, система уравнений

$$y' = \mathcal{F}_y(x_0, y(x_0))y$$

относится к классу жестких систем в смысле приведенного выше определения.

Из проводившихся выше рассуждений для случая системы (2) видно, что численное решение задачи Коши для таких систем требует разработки специальных методов решения. Такие методы в настоящее время разработаны, и на их основе созданы соответствующие комплексы стандартных программ.

Рассмотрим простейшие варианты наиболее распространенных методов решения жестких систем.

1. Пусть уже найдено приближение  $y_n$  к значению  $y(x_n)$  и ищется приближение к значению  $y(x_{n+1})$ ,  $x_{n+1} - x_n = H$ . Разложим правую часть  $f(x, y)$  в ряд Тейлора в точке  $x_n, y_n$

$$f(x, y) \sim f(x_n, y_n) + \frac{\partial f(x_n, y_n)}{\partial x}(x - x_n) + \\ + \mathcal{F}_y(x_n, y(x_n))(y - y_n) + \dots \quad (6)$$

В прикладных задачах, как правило, возникают такие жесткие системы, что  $f(x, y)$  не зависит от  $x$  или меняется относительно медленно с изменением  $x$ . В этом случае главными членами в правой части являются первый и третий. За  $y_{n+1}$  примем значение в точке  $x_{n+1}$  решения системы

$$z' = f(x_n, y_n) + \mathcal{F}_y(x_n, y(x_n))(z - y_n) \quad (7)$$

при начальном условии  $z(x_n) = y_n$ . Произведем замену переменных  $z(x) - y_n = u(x)$ ,  $x - x_n = t$  и введем обозначения

$$f(x_n, y_n) = b, \quad \mathcal{F}_y(x_n, y(x_n)) = A.$$

Для определения  $z(x_{n+1})$  нужно найти значение  $u(H)$  решения системы  $u' = Au + b$  при начальном условии  $u(0) = 0$ .

Вообще говоря, эта задача относится к классу задач, решаемых в явном виде. Однако для ее решения требуется знать все собственные векторы и собственные значения матрицы  $A$ . Если размерность матрицы  $A$  сколько-нибудь большая, то найти их — довольно трудоемкая задача. Поэтому целесообразнее следующий путь нахождения  $u(H)$ . Решение системы уравнений  $u' = A(t)u + b(t)$  при начальном условии  $u(0) = 0$  записывается в виде

$$u(t) = \int_0^t \exp \left\{ \int_{\tau}^t A(\alpha) d\alpha \right\} b(\tau) d\tau.$$

Матрица  $W(t, \tau) = \exp \left\{ \int_{\tau}^t A(\alpha) d\alpha \right\}$  является решением системы

$$\frac{dW(t, \tau)}{dt} = A(t)W(t, \tau)$$

при начальном условии  $W(\tau, \tau) = E$ ,

В частном случае  $A, b = \text{const}$  имеем  $u(t) = \omega(t)b$ , где матрица  $\omega(t)$  имеет вид

$$\omega(t) = \int_0^t \exp \{ A(t - \tau) \} d\tau.$$

Значение  $\omega(H)$  можно было бы попытаться вычислить, пользуясь разложением в ряд Тейлора

$$\omega(H) \sim H \sum_{i=1}^{\infty} \frac{1}{i!} (AH)^{i-1}, \quad (8)$$

Однако в случае жестких систем при реально допустимых значениях  $H$  величина  $\|A\|H \gg 1$  и

а) для достижения приемлемой точности требуется брать слишком много слагаемых;

б) в тех случаях, когда требуемое для достижения нужной точности число слагаемых допустимо по числу арифметических

действий, использование разложения (8) при  $\|A\|H \gg 1$  может быть неприемлемо по другой причине: среди учитываемых в правой части слагаемых встречаются очень большие, и относительная погрешность, вызванная округлениями, недопустимо велика.

Матрица  $\omega(t)$  удовлетворяет соотношению

$$\omega(t) = \omega\left(\frac{t}{2}\right) \left(2E + A\omega\left(\frac{t}{2}\right)\right), \quad (9)$$

где  $E$  — единичная матрица. Поэтому часто наиболее целесообразно бывает пойти по следующему пути. Выбираем  $s$  такое, что  $\|A\|H \cdot 2^{-s} \ll 1$ ; основываясь на (8), вычисляем

$$\omega\left(\frac{H}{2^s}\right) \approx \frac{H}{2^s} \sum_{i=1}^s \frac{1}{i!} \left(A \frac{H}{2^s}\right)^{i-1},$$

а затем  $\omega(H/2^{s-1})$ , ...,  $\omega(H/2)$ ,  $\omega(H)$  с помощью рекуррентной формулы (9).

В случае линейной системы  $y' = A(x)y$  алгоритм решения задачи может быть несколько упрощен. Здесь в описанном выше алгоритме на каждом шаге возникает необходимость решения системы  $y' = A(x_n)y$  при начальном условии  $y(x_n) = y_n$ . Тогда

$$y_{n+1} = \varphi(H)y_n, \quad \text{где } \varphi(H) = \exp\{AH\}, \quad A = A(x_n).$$

На первый взгляд может показаться разумным следующий путь. Матрица  $\varphi(H)$  удовлетворяет соотношению

$$\varphi(t) = \varphi\left(\frac{t}{2}\right) \varphi\left(\frac{t}{2}\right), \quad (10)$$

поэтому задаемся некоторым  $s$  и вычисляем

$$\varphi\left(\frac{H}{2^s}\right) \approx \sum_{i=0}^k \frac{1}{i!} \left(A \frac{H}{2^s}\right)^i,$$

а затем

$$\varphi\left(\frac{H}{2^{s-1}}\right), \dots, \varphi(H),$$

пользуясь рекуррентной формулой (10). Такой путь при большом  $s$  приводит к существенному накоплению погрешности, что уже показывалось в § 4.9 на примере вычисления  $\exp\{x\}$  с помощью рекуррентной формулы

$$\exp\{x\} = \exp\{x/2\} \exp\{x/2\}.$$

Поэтому поступим, как и в § 4.9. Положим  $\psi(H) = \varphi(H) - E$ , вычислим

$$\psi\left(\frac{H}{2^s}\right) \approx \sum_{i=1}^k \frac{1}{i!} \left(\frac{AH}{2^i}\right)^s$$

и затем  $\psi(H/2^{s-1}), \dots, \psi(H)$ , пользуясь рекуррентной формулой

$$\psi(t) = \psi\left(\frac{t}{2}\right) \left(2E + \psi\left(\frac{t}{2}\right)\right).$$

Далее находим  $\varphi(H) = E + \psi(H)$ .

При  $f_x \neq 0$  метод точности  $O(h^2)$  получается, если в (6) также учесть второе слагаемое. Тогда потребуется аналогичным образом сконструировать точный метод решения вспомогательного уравнения

$$u' = Au + b + ct, \quad c = \frac{\partial f}{\partial x} \Big|_{(x_n, y_n)}.$$

В случае решения линейной задачи  $u' = A(x)u + f(x)$  можно предложить довольно простые методы точности  $O(h^4)$ .

Другая группа методов решения жестких задач строится следующим образом. Зададимся некоторым  $k$  и приблизим производную  $y'(x_n)$  односторонней аппроксимацией  $k$ -го порядка точности

$$y'(x_n) \approx \frac{\sum_{i=1}^k \frac{\nabla^i y_n}{i}}{h} = \sum_{i=0}^k \frac{a_{-i} y_{n-i}}{h}.$$

Выражение  $f(x_n, y(x_n))$  оставим без изменения. Получим конечно-разностную аппроксимацию

$$\sum_{i=0}^k \frac{a_{-i} y_{n-i}}{h} - f(x_n, y_n) = 0. \quad (11)$$

Рассмотрим случай модельного уравнения  $y' = My$ , когда (11) превращается в конечно-разностное уравнение

$$\sum_{i=0}^k \frac{a_{-i} y_{n-i}}{h} - My_n = 0.$$

Решение этого уравнения выписывается через корни характеристического уравнения

$$\sum_{i=0}^k a_{-i} \mu^{k-i} - hM\mu^k = 0.$$

При  $k = 1, 2$  корни этого уравнения удовлетворяют условию  $|\mu| \leq 1$  в области значений  $M$ :  $\operatorname{Re} M < 0$ ; при  $k = 3, 4, 5, 6$  — условию  $|\mu| \leq 1$  в области значений  $M$ :  $|\operatorname{Im} M| \leq -\alpha_k \operatorname{Re} M$ ; здесь  $\alpha_k > 0$ .

В нелинейном случае значение  $y_n$  требуется находить из нелинейной системы (11).

Алгоритмы решения жестких систем различаются способами нахождения начального приближения к решению (11) и алгоритмами приближенного решения (11). Рассмотрим простейший случай  $k = 1$ , когда (11) превращается

в явный метод Эйлера:

$$\frac{y_n - y_{n-1}}{h} - f(x_n, y_n) = 0. \quad (12)$$

Могло бы показаться разумным найти начальное приближение к  $y_n$  с помощью явной формулы Эйлера

$$y_n^0 = y_{n-1} + hf(x_{n-1}, y_{n-1}),$$

однако это не всегда целесообразно. В случае  $f(x, y) = My$  имеем  $y_n^0 = (1 + Mh)y_{n-1}$  и при  $|Mh| \gg 1$  может случиться, что такое приближение будет слишком далеко от истинного решения. Поэтому более безопасный, но не всегда самый эффективный вариант — это положить  $y_n^0 = y_{n-1}$ . Перепишем (12) в виде системы нелинейных уравнений относительно  $y_n$ :

$$y_n - y_{n-1} - hf(x_n, y_n) = 0,$$

и применим итерационный метод Ньютона. В данном конкретном случае интерполяционная формула Ньютона приобретает вид

$$y_n^{k+1} = y_n^k - (E - h\mathcal{F}_y(x_n, y_n^k))^{-1} (y_n^k - y_{n-1} - hf(x_n, y_n^k)), \quad (13)$$

где  $k$  — номер итерации.

В одном из методов решения жестких систем за  $y_n$  принимается значение  $y_n^1$ , получаемое из (13) при  $y_n^0 = y_{n-1}$ . Тогда имеем

$$y_n = y_n^1 = y_{n-1} + h(E - h\mathcal{F}_y(x_n, y_n))^{-1} f(x_n, y_{n-1}).$$

Рассмотрим случай уравнения  $y' = My$ ; тогда

$$y_n = y_{n-1} + \frac{Mh}{1 - Mh} y_{n-1} = \frac{1}{1 - Mh} y_{n-1}.$$

Если  $\text{Re } M < 0$ , в частности, если  $M < 0$ , то при любом  $h$  имеем  $\left| \frac{1}{1 - Mh} \right| < 1$  и погрешности, возникающие в процессе счета, убывают. Поэтому такой метод может быть применен к решению жестких систем.

## § 10. Методы численного интегрирования уравнений второго порядка

Введением новых неизвестных функций дифференциальные уравнения порядка выше первого и их системы сводятся к системам уравнений первого порядка. Таким образом, при формальном подходе вопрос о численном решении задачи Коши для уравнений высших порядков можно было бы считать исчерпанным.

Однако методы, приспособленные специально для решения уравнений высших порядков, часто более эффективны. При разработке таких методов нужно также иметь в виду широкое распространение систем уравнений высоких порядков специального вида, учет специфики которых может еще более повысить эффективность методов. Например, ряд задач небесной механики сводится к интегрированию систем уравнений

$$y'' = f(y).$$

Рассмотрим несколько более широкий класс уравнений

$$y'' = f(x, y).$$

Как и выше, рассуждения проводятся для одного уравнения, поскольку перенос результатов на случай системы осуществляется автоматически. Традиционно наиболее распространенными методами интегрирования таких уравнений являются *явный*

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) \quad (1)$$

и *неявный*

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{i=-1}^k b_{-i} f(x_{n-i}, y_{n-i}) \quad (2)$$

конечно-разностные методы. Эти методы можно было бы получить методом неопределенных коэффициентов, потребовав, чтобы разность

$$\frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2} - \sum_i b_{-i} y''(x_{n-i})$$

была величиной как можно более высокого порядка по  $h$ . Здесь, как обычно, предполагается  $x_{n+1} - x_n = h$ .

Одним из наиболее употребительных среди этих методов является *неявный метод четвертого порядка точности вида (2)*, (часто называемый *методом Нумерова*) при  $k = 1$ :

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \left( \frac{1}{12} f(x_{n-1}, y_{n-1}) + \right. \\ \left. + \frac{10}{12} f(x_n, y_n) + \frac{1}{12} f(x_{n+1}, y_{n+1}) \right),$$

который особенно удобен в случае линейных задач (подробный вывод этого метода для линейного уравнения см. в § 9.1).

Для целей численного интегрирования уравнения

$$y'' = f(x, y, y')$$

употребительными являются следующие методы: значения  $y_n$  и  $z_n$  приближений к  $y(x_n)$  и  $y'(x_n)$  вычисляются из совокупности явных рекуррентных соотношений вида

$$z_{n+1} = z_n + h \sum_{k=0}^l a_k \nabla^k f_n,$$

$$y_{n+1} = y_n + h \sum_{k=0}^l b_k \nabla^k z_{n+1}$$



или неявных вида

$$z_{n+1} = z_n + h \sum_{k=0}^l c_k \nabla^k f_{n+1},$$

$$y_{n+1} = y_n + h \sum_{k=0}^l d_k \nabla^k z_{n+1}.$$

Уравнения более высокого порядка, чем второй, на практике часто сводят к системам уравнений второго или первого порядка (см. также § 9.11).

Как и в случае конечно-разностных схем для уравнений первого порядка, отбросим в (1) и (2) слагаемые, содержащие значения  $f$ , и рассмотрим получившееся разностное уравнение. В обоих случаях оно имеет вид

$$y_{n+1} - 2y_n + y_{n-1} = 0.$$

Его характеристическое уравнение имеет кратный корень  $\mu = 1$ . Не нужно пугаться того, что корень оказался кратным: если в разностных схемах, аппроксимирующих уравнения первого порядка, перед значениями  $f$  стоял коэффициент порядка  $h$ , то здесь стоит коэффициент порядка  $h^2$ .

Все упоминавшиеся ранее разностные схемы интегрирования уравнения  $y'' = f(x, y)$  могут быть представлены в виде

$$\sum_{i=0}^k \frac{a_{-i} y_{n-i}}{h^2} - \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) = 0, \quad a_0 \neq 0; \quad (3)$$

при этом для гладкой функции

$$\frac{\sum_{i=0}^k a_{-i} y(x_{n-i})}{h^2} \rightarrow y''(x_n)$$

при  $x_n$  фиксированном,  $h \rightarrow 0$ ; кроме того,  $\sum_{i=0}^k b_{-i} = 1$ . Пусть  $r_n$  — погрешность аппроксимации (результат подстановки в левую часть решения дифференциальной задачи):

$$r_n = \sum_{i=0}^k \frac{a_{-i} y(x_{n-i})}{h^2} - \sum_{i=0}^k b_{-i} f(x_{n-i}, y(x_{n-i})),$$

и  $|r_n| \leq c_0 h^m$  равномерно на всем рассматриваемом отрезке интегрирования  $[x_0, x_0 + X]$ . Как и в случае уравнений первого порядка, из-за округлений при вычислениях и неточности решения уравнения относительно неизвестной  $y_n$  в случае неявной схемы ( $b_0 \neq 0$ ) реально получаемое приближенное решение  $y_n$  удовлетворяет (3) с некоторой погрешностью. Таким образом,

имеем

$$\frac{\sum_{i=0}^k a_{-i} y_{n-i}}{h^2} - \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) = \delta_n.$$

Вычитая из этого соотношения предыдущее, получим уравнение для погрешности  $R_n = y_n - y(x_n)$ :

$$\frac{\sum_{i=0}^k a_{-i} R_{n-i}}{h^2} - \sum_{i=0}^k b_{-i} l_{n-i} R_{n-i} = g_n, \quad (4)$$

где  $l_j = f_y(x_j, \tilde{y}_j)$ ,  $g_n = \delta_n - r_n$ . Пусть

$$L = \sup_{x_0 \leq x \leq x_0 + X} |f_y(x, y)| < \infty.$$

**Теорема (без доказательства).** Пусть все корни характеристического уравнения разностной схемы

$$\sum_{i=0}^k a_{-i} \mu^{k-i} = 0$$

лежат в единичном круге и на границе круга нет кратных корней, за исключением двукратного корня, равного 1. Тогда при  $x_0 \leq x_n \leq x_0 + X$  справедлива оценка погрешности

$$\max_n \left( |R_n|, \left| \frac{R_n - R_{n-1}}{h} \right| \right) \leq \leq C(L, X) \left( \max_{x_0 \leq x_n \leq x_0 + X} |g_n| + \max_{0 \leq l < k} |R_l| + \max_{0 < l < k} \left| \frac{R_l - R_{l-1}}{h} \right| \right). \quad (5)$$

При численной реализации методов решения уравнений второго порядка с целью уменьшения влияния вычислительной погрешности целесообразно преобразовать расчетные формулы к виду (9.13.7).

Методы численного решения уравнений второго порядка могут быть использованы при численном решении уравнений, где производная решения не выражается явно через само решение и аргумент; для определенности рассмотрим скалярный случай

$$F(x, y, y') = 0.$$

Если бы это уравнение удалось разрешить относительно  $y'$ , то получилось бы некоторое уравнение

$$y' = f(x, y).$$

Первый возможный путь решения задачи состоит в формальном применении методов Рунге — Кутты или конечно-разностных методов; для каждого  $x$  и  $y$  значение  $f$  определяется путем численного решения уравнения

$$F(x, y, f) = 0.$$

Хорошее начальное приближение к искомому значению  $f$  можно получить интерполяцией ранее найденных значений, поэтому число требуемых итераций (например, в методе секущих) обычно оказывается небольшим.

В случае неявных методов бывает целесообразно сразу решать систему уравнений

$$F(x_n, y_n, f_n) = 0,$$

$$\sum_{i=0}^k a_{-i} y_{n-i} - h \sum_{i=0}^k b_{-i} f_{n-i} = 0$$

относительно неизвестных  $y_n, f_n$ .

Иногда (крайне редко) поступают следующим образом: дифференцируя исходное уравнение по  $x$ , получают соотношения

$$\frac{d}{dx} (F(x, y, y')) = F_x(x, y, y') + F_y(x, y, y') y' + F_{y'}(x, y, y') y'' = 0,$$

$$\frac{d^2}{dx^2} (F(x, y, y')) = \dots = 0$$

и т. д. Если значение  $y'(x_0)$  уже найдено, то из этих соотношений можно явным образом определить значения  $y''(x_0), \dots$ , а затем получить  $y(x_0 + h)$  с помощью формулы Тейлора.

Первое из этих соотношений можно переписать в виде

$$y'' = g(x, y, y').$$

Отсюда получается третий путь решения задачи: из уравнения

$$F(x_0, y_0, y'(x_0)) = 0$$

определяется  $y'(x_0)$ , и далее численно интегрируется уравнение

$$y'' = g(x, y, y').$$

Л и т е р а т у р а к г л. VIII: [31, 38, 51, 59, 67, 68].

## Глава IX

### ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ КРАЕВЫХ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

При решении краевых задач возникают дополнительные трудности по сравнению со случаем решения задачи Коши; значительно сложнее исследуется вопрос о существовании решения; после написания сеточной задачи возникает система линейных или нелинейных уравнений, проблема решения которой требует дополнительного изучения.

#### § 1. Простейшие методы решения краевой задачи для уравнения второго порядка

Среди краевых задач для обыкновенных дифференциальных уравнений существенную часть составляют задачи для уравнений и систем второго порядка. В частности, такие задачи возникают в баллистике, теории упругости и т. д.

Начнем изучение вопроса с одной частной, но довольно распространенной краевой задачи. Ищется решение уравнения

$$Ly = y'' - p(x)y = f(x) \quad \text{на } (0, X) \quad (1)$$

при граничных условиях

$$y(0) = a, \quad y(X) = b. \quad (2)$$

Зададимся шагом  $h = XN^{-1}$ ,  $N$  целое; точки  $x_n = nh$  примем за узлы сетки; как обычно,  $y_n$  — приближения к значениям  $y(x_n)$ . После замены производной  $y''(x_n)$  на разностное отношение

$$\frac{\delta^2 y_n}{h^2} \equiv \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2}$$

получаем систему уравнений

$$l(y_n) \equiv \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p_n y_n = f_n, \quad n = 1, \dots, N-1; \quad (3)$$

здесь  $p_n = p(x_n)$ ,  $f_n = f(x_n)$ ; граничные условия заменим соотношениями

$$y_0 = a, \quad y_N = b. \quad (4)$$

Покажем, что при  $p(x) \geq 0$  система уравнений (3), (4) имеет решение, и дадим оценку погрешности.

**Лемма 1.** Пусть  $p(x) \geq 0$ ,  $l(z_n) \leq 0$ ,  $z_0, z_N \geq 0$ . Тогда  $z_n \geq 0$  при всех  $n$ .

**Доказательство.** Обозначим  $\min_{0 \leq n \leq N} z_n$  через  $d$ . Предположим, что  $d < 0$  и, следовательно,  $d \neq z_0, z_N$ . Пусть  $q$  — наименьшее целое такое, что  $z_q = d$ ; из определения  $d$  и  $q$  имеем  $z_{q-1} > d$ ,  $z_{q+1} \geq d$ . Тогда

$$l(z_q) = \frac{(z_{q+1} - d) + (z_{q-1} - d)}{h^2} - p_q z_q \geq \frac{z_{q-1} - d}{h^2} > 0$$

и мы приходим к противоречию с предположением  $d < 0$ .

**Лемма 2.** Если  $p(x) \geq 0$ , то для любой функции  $z_n$  выполняется неравенство

$$\max_{0 \leq n \leq N} |z_n| \leq \max(|z_0|, |z_N|) + Z \cdot \frac{X^2}{8},$$

где

$$Z = \max_{0 < n < N} |l(z_n)|.$$

**Доказательство.** Введем в рассмотрение функцию

$$\omega_n = |z_0| \left(1 - \frac{nh}{X}\right) + |z_N| \frac{nh}{X} + Z \frac{nh(X - nh)}{2}.$$

Для многочлена второй степени величина  $\delta^2 Q/h^2$  совпадает со второй производной, поэтому  $\delta^2 \omega_n/h^2 = -Z$ . Из явного вида  $\omega_n$  следует, что  $\omega_n \geq 0$ , поэтому

$$l(\omega_n) = \delta^2 \omega_n/h^2 - p_n \omega_n \leq -Z, \quad l(\omega_n \pm z_n) \leq -Z \pm l(z_n) \leq 0.$$

Имеем

$$\omega_0 \pm z_0 = |z_0| \pm z_0 \geq 0, \quad \omega_N \pm z_N = |z_N| \pm z_N \geq 0.$$

Функции  $\omega_n \pm z_n$  удовлетворяют условиям леммы 1, поэтому  $\omega_n \pm z_n \geq 0$ . Отсюда следует оценка  $|z_n| \leq |\omega_n| \leq \max_{0 \leq n \leq N} |\omega_n|$ .

Имеем неравенства

$$\begin{aligned} |z_0| \left(1 - \frac{nh}{X}\right) + |z_N| \frac{nh}{X} &\leq \max(|z_0|, |z_N|) \left(\left|1 - \frac{nh}{X}\right| + \left|\frac{nh}{X}\right|\right) = \\ &= \max(|z_0|, |z_N|), \end{aligned}$$

$$nh(X - nh) \leq X^2/4.$$

Поэтому

$$\max_{0 \leq n \leq N} |w_n| \leq \max(|z_0|, |z_N|) + Z \frac{X^2}{8}.$$

Лемма доказана.

Рассмотрим случай, когда функции  $p(x)$  и  $f(x)$  дважды непрерывно дифференцируемы. В курсе дифференциальных уравнений доказывается, что тогда решение  $y(x)$  четырежды непрерывно дифференцируемо.

Пусть  $r_n$  — погрешность аппроксимации, соответствующая конечно-разностной схеме (3):

$$r_n = l(y(x_n)) - f_n = \frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2} - p(x_n)y(x_n) - f(x_n). \quad (5)$$

Поскольку  $p(x)y(x) + f(x) = y''(x)$ , то

$$r_n = \frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2} - y''(x_n).$$

Из оценок погрешности формул численного дифференцирования имеем

$$r_n = \frac{y^{(4)}(\bar{x}_n) h^2}{12}, \quad \text{где } x_{n-1} \leq \bar{x}_n \leq x_{n+1}. \quad (6)$$

Из-за округления получаемые в процессе вычислений приближения  $y_n$  к значениям  $y(x_n)$  удовлетворяют системе (3), (4) с некоторыми погрешностями

$$l(y_n) - f_n = \delta_n. \quad (7)$$

Вычитая (5) из (7), получим уравнение

$$l(R_n) = \delta_n - r_n$$

относительно погрешности приближенного решения  $R_n = y_n - y(x_n)$ . Воспользовавшись леммой 2, получим

$$|R_n| \leq \max(|R_0|, |R_N|) + \frac{X^2}{8} (\max_{0 \leq n \leq N} |r_n| + \max_{0 \leq n \leq N} |\delta_n|).$$

Согласно оценке (6) имеем

$$|r_n| \leq M_4 \frac{h^2}{12}, \quad M_4 = \max_{[0, X]} |y^{(4)}(x)|.$$

Таким образом, окончательная оценка погрешности имеет вид

$$\max_{0 \leq n \leq N} |R_n| \leq \max(|R_0|, |R_N|) + \frac{M_4 X^2 h^2}{96} + (\max_{0 \leq n \leq N} |\delta_n|) \frac{X^2}{8}.$$

Мы видим, что при повышении точности, с которой удовлетворяются граничные условия и разностное уравнение, при одно-

временном стремлении шага к нулю решение сеточной задачи сближается с решением дифференциальной задачи.

**З а м е ч а н и е.** Поскольку  $y_n$  округляются в машине с погрешностью порядка  $2^{-t}$ , величины  $\delta_n$  имеют порядок  $2^{-t}/h^2$ . Если решать сеточную задачу (3), (4) специальным способом (4.9—4.10), то величины  $\delta_n$  будут порядка  $2^{-t}/h^2$ , но их главные части имеют дивергентную форму:

$$\delta_n = \frac{\omega_{n+1/2} - \omega_{n-1/2}}{h} + O(2^{-t}/h),$$

где  $\omega_{n+1/2} = O(2^{-t}/h)$ . Используя это обстоятельство, можно показать, что суммарная вычислительная погрешность метода (9.11.9) имеет порядок  $O(2^{-t}/h)$ .

Описанный метод дает приближенное решение, сходящееся к точному со скоростью  $O(h^2)$ . Займемся построением более точных схем. Будем предполагать, что функции  $p(x)$  и  $f(x)$  непрерывно дифференцируемы 4 раза, тогда решение задачи непрерывно дифференцируемо 6 раз. Еще раз рассмотрим выражение

$$r_n = \frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2} - y''(x_n).$$

Подставим сюда представление  $y(x_{n\pm 1})$  с помощью формулы Тейлора:

$$y(x_{n\pm 1}) = y(x_n) \pm y'(x_n)h + y''(x_n)\frac{h^2}{2} \pm \pm y'''(x_n)\frac{h^3}{6} + y^{(4)}(x_n)\frac{h^4}{24} \pm y^{(5)}(x_n)\frac{h^5}{120} + O(h^6),$$

получим

$$r_n = \frac{y^{(4)}(x_n)h^2}{12} + O(h^4). \quad (8)$$

Вычтем из  $l(y(x_n))$  некоторое слагаемое, аппроксимирующее величину  $y^{(4)}(x_n)h^2/12$ ; получаемой схеме соответствует погрешность аппроксимации более высокого порядка. Например, можно приблизить  $y^{(4)}(x_n)$  выражением

$$\frac{\delta^4 y(x_n)}{h^4} = \frac{y(x_{n+2}) - 4y(x_{n+1}) + 6y(x_n) - 4y(x_{n-1}) + y(x_{n-2}))}{h^4};$$

получится конечно-разностная схема

$$\frac{\delta^2 y_n}{h^2} - \frac{\delta^4 y_n}{12h^4} - p_n y_n = f_n. \quad (9)$$

Эту схему можно также построить непосредственно, заменив производную  $y''(x_n)$  выражением  $h^{-2}(\delta^2 y(x_n) - (1/12)\delta^4 y(x_n))$ , приближающим ее с погрешностью  $O(h^4)$ .

Уравнение (9) содержит 5 неизвестных  $y_n$  с ненулевыми коэффициентами. Решение системы, состоящей из уравнений (9) и уравнений, получающихся при аппроксимации граничных

условий, более трудоемко, чем решение системы (3). Исходя из других соображений построим конечно-разностную схему с погрешностью аппроксимации  $O(h^4)$  такую, что в каждое уравнение входит только три неизвестных.

Дифференцируя дважды исходное уравнение, имеем  $y^{(4)}(x) = (p(x)y + f)''$ , поэтому

$$y^{(4)}(x_n) \approx \frac{\delta^2(p y + f)|_{x_n}}{h^2} = \frac{(p_{n+1}y(x_{n+1}) + f_{n+1}) - 2(p_n y(x_n) + f_n) + (p_{n-1}y(x_{n-1}) + f_{n-1}))}{h^2}.$$

Вычитая из исходной схемы приближение для  $y^{(4)}h^2/12$ , получим схему

$$\frac{\delta^2 y_n}{h^2} - p_n y_n - \frac{1}{12} \delta^2 (p_n y_n + f_n) = f_n$$

или

$$l^{(1)}(y_n) = \frac{\delta^2 y_n}{h^2} - p_n y_n - \frac{1}{12} \delta^2 (p_n y_n) = \bar{l}^{(1)}(f_n) = f_n + \frac{1}{12} \delta^2 f_n.$$

В предположении, что решение непрерывно дифференцируемо 8 раз, рассмотрим погрешность аппроксимации новой схемы  $r_n^{(1)} = l^{(1)}(y(x_n)) - \bar{l}^{(1)}(f_n)$ . Учитывая, что  $p_n y(x_n) + f_n = y''(x_n)$ , получим

$$r_n^{(1)} = l^{(1)}(y(x_n)) - \bar{l}^{(1)}(f_n) = \frac{\delta^2 y(x_n)}{h^2} - y''(x_n) - \frac{1}{12} \delta^2 y''(x_n).$$

Воспользовавшись формулой Тейлора, аналогично (8) получаем равенство

$$r_n^{(1)} = -\frac{y^{(6)}(x_n) h^4}{240} + O(h^6).$$

Построим приближение величины  $y^{(6)}(x_n)$  посредством значений  $y(x_{n-1})$ ,  $y(x_n)$ ,  $y(x_{n+1})$ . Таких приближений можно написать очень много, например следующим образом. Согласно уравнению (1)

$$y'' = p y + f, \quad y''' = (p y + f)', \quad y^{(4)} = (p y + f)'',$$

$$y^{(6)} = (p y + f)^{(4)} = p y^{(4)} + 4p' y^{(3)} + 6p'' y'' + 4p^{(3)} y' + p^{(4)} y + f^{(4)},$$

поэтому справедливо приближенное равенство

$$y^{(6)}(x_n) \approx p_n \frac{\delta^2 (p_n y(x_n) + f_n)}{h^2} + 4p'(x_n) \frac{(p_{n+1}y(x_{n+1}) + f_{n+1}) - (p_{n-1}y(x_{n-1}) + f_{n-1}))}{2h} + 6p''(x_n) (p_n y(x_n) + f_n) + 4p^{(3)}(x_n) \frac{y(x_{n+1}) - y(x_{n-1}))}{2h} + p^{(4)}(x_n) y(x_n) + f^{(4)}(x_n).$$

Прибавляя к  $l^{(1)}(y_n)$  выражение, приближающее  $y^{(6)}(x_n)h^4/240$ , получим конечно-разностную схему с погрешностью аппроксимации  $O(h^6)$ ; при этом в каждое уравнение получившейся алгебраической системы входят только три неизвестных  $y_n$ .



Для практической оценки погрешности решения краевой задачи может применяться правило Рунге. Законность его применения основывается на существовании главного члена погрешности.

**Задача 1.** Пусть функции  $p(x)$  и  $f(x)$  четырежды дифференцируемы. Доказать, что для решения задачи (3), (4) справедливо соотношение

$$\max_n |y_n - y(x_n) - h^2 z(x_n)| = O(h^4);$$

здесь  $z(x)$  — решение краевой задачи

$$Lz = -y^{(4)}(x)/12, \quad z(0) = 0, \quad z(X) = 0.$$

Аналогичный прием последовательного повышения порядка погрешности аппроксимации может быть применен и по отношению к аппроксимациям граничного условия.

Рассмотрим случай граничного условия  $y'(0) - \alpha y(0) = a$ . Дискретное приближение высокой точности к такому граничному условию можно получить непосредственно, заменив производную  $y'(0)$  по какой-либо формуле численного дифференцирования высокой точности:

$$y'(0) \approx \sum_{i=0}^l \frac{c_i y(x_i)}{h}.$$

Однако погрешность аппроксимации будет меньше и решение возникающей алгебраической системы представит меньше трудностей, если идти по описанному выше пути последовательного повышения порядка точности аппроксимации.

Заменим производную  $y'(0)$  отношением  $\frac{y(h) - y(0)}{h}$ , получим  $l_0^{(1)}(y_n) = \frac{y_1 - y_0}{h} - \alpha y_0 - a = 0$ . Подставляя в  $r_0^{(1)} = \frac{y(h) - y(0)}{h} - \alpha y(0) - a$  разложение  $y(h) = y(0) + y'(0)h + O(h^2)$ , получим

$$r_0^{(1)} = y'(0) + \frac{y''(0)h}{2} + O(h^2) - \alpha y(0) - a = \frac{y''(0)h}{2} + O(h^2).$$

Таким образом, погрешность аппроксимации граничного условия есть  $O(h)$ . Поскольку, согласно уравнению (1),

$$y''(0) = p_0 y(0) + f_0,$$

то уравнению

$$l_0^2(y_n) = \frac{y_1 - y_0}{h} - \alpha y_0 - a - (p_0 y_0 + f_0) \frac{h}{2} = 0$$

соответствует второй порядок аппроксимации. Подставляя в  $r_0^{(2)} = l_0^{(2)}(y(x_n))$  разложение  $y(h) = y(0) + y'(0)h + y''(0)h^2/2 + O(h^3)$ , получим  $r_0^{(2)} = \frac{y^{(3)}(0)h^2}{6} + O(h^3)$ . После дифференцирования исходного уравнения (1) имеем

$$y^{(3)}(0) - p_0 y'(0) - p'(0)y(0) - f'(0) = 0;$$

поэтому с учетом граничного условия справедливо равенство

$$y^{(3)}(0) = p_0(\alpha y(0) + a) + p'(0)y(0) + f'(0).$$

Разностному уравнению

$$l_0^{(3)}(y_n) = l_0^{(2)}(y_n) - ((p_0\alpha + p'(0))y_0 + p_0a + f'(0))\frac{h^2}{6} = 0$$

будет соответствовать уже третий порядок аппроксимации.

Можно было бы сразу написать равенство

$$r_0^{(1)} = y''(0)\frac{h}{2} + y^{(3)}(0)\frac{h^2}{6},$$

затем выразить производные  $y''(0)$  и  $y^{(3)}(0)$  через  $y(0)$  и, вычитая из разностной схемы соответствующее выражение, получить уравнение  $l_0^{(3)}(y_n) = 0$ . Однако мы обратили основное внимание именно на способ последовательного повышения порядка точности, поскольку его перенесение на случай уравнений в частных производных является наиболее простым и естественным.

## § 2. Функция Грина сеточной краевой задачи

Функция  $w_n$ , введенная в § 1 при доказательстве леммы 2, носит название *мажорирующей функции*, и метод получения оценок погрешности с использованием этой функции называется *методом мажорант*, или *методом Гершгорина*.

В ряде случаев, когда метод мажорант неприменим, оценку погрешности приближенного решения удается получить, используя так называемую сеточную функцию Грина. Проводимые ниже построения функции Грина сеточной краевой задачи (1.3), (1.4) кроме всего прочего интересны своей аналогией со случаем дифференциальной краевой задачи.

Функция Грина  $G(x, s)$  дифференциальной краевой задачи

$$Ly = y'' - p(x)y = f(x), \quad y(0) = a, \quad y(X) = b,$$

определяется как решение уравнения

$$L(G(x, s)) = G_{xx}(x, s) - p(x)G(x, s) = \delta(x - s) \quad (1)$$

при граничном условии  $G(0, s) = G(X, s) = 0$ ; здесь  $\delta(x)$  —  $\delta$ -функция. Решение задачи (1.1), (1.2) записывается с помощью функции Грина в виде

$$y(x) = \int_0^x G(x, s)f(s)ds + G'_s(x, X)b - G'_s(x, 0)a. \quad (2)$$

Функцию Грина можно задать следующими явными формулами. Пусть  $W^1(x)$ ,  $W^2(x)$  — решения уравнения  $L(W) = 0$  при условиях

$$\begin{aligned} W^1(0) &= 0, & (W^1)'(0) &= 1, \\ W^2(X) &= 0, & (W^2)'(X) &= -1; \end{aligned}$$

тогда

$$G(x, s) = \begin{cases} \frac{W^2(s)W^1(x)}{V^0} & \text{при } 0 \leq x \leq s, \\ \frac{W^1(s)W^2(x)}{V^0} & \text{при } s \leq x \leq X, \end{cases} \quad (3)$$

где  $V^0$  — значение определителя Вронского:

$$V(x) = \begin{vmatrix} W^1(x) & W^2(x) \\ (W^1)'(x) & (W^2)'(x) \end{vmatrix} = \text{const} = V^0.$$

Перейдем к сеточной задаче

$$l(y_n) = \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p_n y_n = f_n, \quad y_0 = a, \quad y_N = b.$$

По аналогии определим функции  $W_n^1, W_n^2$  из соотношений

$$l(W_j^i) = 0, \quad i = 1, 2, \quad j = 1, \dots, N-1, \\ W_0^1 = 0, \quad W_1^1 = h, \quad W_N^2 = 0, \quad W_{N-1}^2 = h.$$

Аналогом определителя Вронского является определитель

$$V_n = \begin{vmatrix} W_n^1 & W_n^2 \\ \frac{W_n^1 - W_{n-1}^1}{h} & \frac{W_n^2 - W_{n-1}^2}{h} \end{vmatrix} = \frac{W_n^2 W_{n-1}^1 - W_n^1 W_{n-1}^2}{h}.$$

Из тождества

$$0 = W_n^1 l(W_n^2) - W_n^2 l(W_n^1) = \\ = \frac{W_n^1 W_{n+1}^2 + W_n^1 W_{n-1}^2 - W_n^2 W_{n+1}^1 - W_n^2 W_{n-1}^1}{h^2} = \frac{V_{n+1} - V_n}{h^2}$$

следует, что величина  $V_n$  не зависит от  $n$ ; будем обозначать ее  $V(h)$ .

Положим

$$G_n^k = \begin{cases} \frac{W_k^2 W_n^1}{V(h)} & \text{при } 0 \leq n \leq k, \\ \frac{W_k^1 W_n^2}{V(h)} & \text{при } k \leq n \leq N. \end{cases} \quad (4)$$

Будем обозначать далее

$$l(G_n^k) = \frac{G_{n+1}^k - 2G_n^k + G_{n-1}^k}{h^2} - p_n G_n^k.$$

Из определения  $W_n^i$  следуют равенства

$$l(G_n^k) = \frac{W_k^2}{V(h)} l(W_n^1) = 0$$

при  $n \leq k$ ,

$$l(G_n^k) = \frac{W_k^1}{V(h)} l(W_n^2) = 0$$

при  $n > k$ ; при  $k = n$  имеем

$$\begin{aligned}
 l(G_n^k) &= \frac{G_{n+1}^n - 2G_n^n + G_{n-1}^n}{h^2} - p_n G_n^n = \\
 &= \left( \frac{W_{n+1}^2 W_n^1 - 2W_n^1 W_n^2 + W_{n-1}^1 W_n^2}{V(h) h^2} - \frac{p_n W_n^1 W_n^2}{V(h)} \right) = \\
 &= \frac{W_n^1}{V(h)} \left( \frac{W_{n+1}^2 - 2W_n^2 + W_{n-1}^2}{h^2} - p_n W_n^2 \right) + \frac{W_n^2 W_{n-1}^1 - W_n^1 W_{n-1}^2}{V(h) h^2}.
 \end{aligned}$$

Согласно определению  $W_n^2$  первая круглая скобка равна 0; вторая скобка равна  $V(h)h$ . В итоге при  $k = n$  получаем  $l(G_n^k) = h^{-1}$ ; объединяя полученные соотношения, имеем

$$l(G_n^k) = \delta_n^k h^{-1}. \quad (5)$$

Функция  $\delta_n^k h^{-1}$  является сеточным аналогом  $\delta$ -функции, а равенство (5) — аналогом (1). Сеточные функции  $W_n^i$ ,  $i = 1, 2$ , являются решениями сеточных задач Коши, соответствующих задачам Коши, определяющим функции  $W^i(x)$ . Предположим, что

$$\|p''\|_{C[0, X]} = \sup_{0 \leq x \leq X} |p''| \leq M_0 < \infty; \quad (6)$$

тогда

$$\|(W^i(x))^{(4)}\|_{C[0, X]} < \infty.$$

Оценив близость начальных данных и воспользовавшись далее теоремой из § 8.10, можно получить оценку близости сеточных и дифференциальных решений задачи Коши вида

$$\max_{0 \leq x_n \leq X} |W_n^i - W^i(nh)| \leq Mh^2, \quad i = 1, 2. \quad (7)$$

Согласно свойствам  $V^0$  и  $V(h)$  имеем

$$\begin{aligned}
 V^0 = V(X) &= \begin{vmatrix} W^1(X) & 0 \\ (W^1)'(X) & -1 \end{vmatrix} = -W^1(X), \\
 V(h) &= \begin{vmatrix} W_N^1 & 0 \\ \frac{W_N^1 - W_{N-1}^1}{h} & -1 \end{vmatrix} = -W_N^1.
 \end{aligned}$$

Поэтому

$$|V(h) - V^0| \leq Mh^2$$

на основании оценки (7).

Далее предполагается, что  $W^1(X) \neq 0$ ; в противном случае однородная краевая задача  $y'' - p(x)y = 0$ ,  $y(0) = y(X) = 0$  имеет ненулевое решение  $W^1(x)$ , а следовательно, неоднородная краевая задача (1.1), (1.2) или не имеет решения, или имеет неединственное решение. Будем также предполагать  $h$  настолько малым, что  $2Mh^2 \leq |W^1(X)| = |V^0|$ . В этом случае имеем

$$V(h) \geq V^0 - Mh^2 \geq |V^0/2| > 0.$$

Сравнивая явные выражения функции Грина дифференциальной (2) и сеточной (4) задач с учетом оценки (7), получаем

$$|G_n^k - G(nh, kh)| \leq Qh^2 \quad (8)$$

при  $0 \leq kh, nh \leq X$ ,

**Лемма.** Если  $V(h) \neq 0$ , то рассматриваемая сеточная задача (1.3), (1.4) однозначно разрешима и ее решение записывается в виде

$$y_n = h \sum_{k=1}^{N-1} G_n^k f_k + \frac{W_n^2}{W_0^2} a + \frac{W_n^1}{W_N^1} b. \quad (9)$$

Формула (9) является сеточным аналогом формулы (2).

**Доказательство.** Поскольку согласно определению функции  $G_n^k$  здесь выполняется равенство  $G_0^k = G_N^k \equiv 0$ , то  $y_0 = a$ ,  $y_N = b$ . Имеем равенство

$$l(y_n) = h \sum_{k=1}^{N-1} l(G_n^k) f_k + \frac{l(W_n^2)}{W_0^2} a + \frac{l(W_n^1)}{W_N^1} b;$$

коэффициенты при  $a$  и  $b$  равны нулю по определению функций  $W_n^l$ . Воспользовавшись равенством (5), получим  $l(y_n) = f_n$ . Таким образом, система линейных уравнений (1.3)–(1.4) при любых правых частях имеет решение, записываемое в виде (9). Но если система линейных уравнений разрешима при любых правых частях, то ее решение единственно. Лемма доказана.

Соотношение (9) может быть двояким способом использовано для оценки близости решений сеточной и дифференциальной задач.

Первый путь состоит в непосредственном сравнении выражений (2) и (9). Для простоты продемонстрируем его в случае  $a = b = 0$ ,  $\delta_n \equiv 0$ . Подставляя  $x = nh$ , перепишем (2) в виде

$$y(nh) = \int_0^{nh} G(nh, s) f(s) ds + \int_{nh}^X G(nh, s) f(s) ds.$$

При условиях  $\|f''\|_{C[0, X]} < \infty$  и (6) можно показать, что подынтегральные функции в обоих интегралах имеют ограниченную вторую производную. Заменяем оба эти интеграла по формуле трапеций с шагом  $h$ . Получим

$$y(nh) = h \left( \frac{G(nh, 0) f(0)}{2} + \sum_{k=1}^{n-1} G(nh, kh) f(kh) + \frac{G(nh, nh) f(nh)}{2} \right) + h \left( \frac{G(nh, nh) f(nh)}{2} + \sum_{k=n+1}^{N-1} G(nh, kh) f(kh) + \frac{G(nh, Nh) f(Nh)}{2} \right) + O(h^2).$$

Разбиение интеграла на две части потребовалось в связи с тем, что погрешность формулы трапеций оценивается через вторую производную подынтегральной функции, а функция  $G(nh, s)$  имеет разрыв первой производной в точке  $nh$ .

Поскольку, по определению функции Грина  $G(x, s)$ , имеем  $G(nh, 0) = G(nh, N) = 0$ , то

$$y(nh) = h \sum_{k=1}^{N-1} G(nh, kh) f_k + O(h^2). \quad (10)$$

Из равенств (9), (10) следует

$$y_n - y(nh) = h \sum_{k=1}^{N-1} (G_n^k - G(nh, kh)) f_k + O(h^2).$$

Воспользовавшись оценкой (7), получим  $y_n - y(nh) = O(h^2)$ . Нетрудно проследить, что эта оценка погрешности равномерна по  $n$  при  $0 \leq x_n \leq X$ , т. е.

$$\max_{0 \leq x_n \leq X} |y_n - y(nh)| = O(h^2).$$

Другой путь состоит в получении уравнения для погрешности и дальнейшей оценки погрешности с помощью формулы Грина. В § 1 было показано, что погрешность  $R_n$  удовлетворяет уравнению

$$l(R_n) = \delta_n - r_n,$$

где  $|r_n| \leq M_4 h^2/12$ ,  $\delta_n$  — погрешность, обусловленная неточностью решения системы (1.3), (1.4). Пусть  $R_0 = R_N = 0$ . Из явного вида функции Грина дифференциальной задачи следует ее равномерная ограниченность; с учетом (7) получаем, что функция  $G_n^k$  также равномерно ограничена:  $|G_n^k| \leq D$  при  $0 \leq kh, nh \leq X$ . Напишем явное представление  $R_n$  с помощью формулы Грина (9):

$$R_n = h \sum_{k=1}^{N-1} G_n^k (\delta_k - r_k). \quad (11)$$

Тогда из (11) следует оценка

$$|R_n| \leq Dh \sum_{k=1}^{N-1} (|\delta_k| + |r_k|).$$

Загрубляя, получим оценку

$$|R_n| \leq DX \left( \frac{M_4}{12} h^2 + \max_{0 < k < N} |\delta_k| \right),$$

имеющую тот же порядок по отношению к  $h$  и  $\max |\delta_k|$ , что и полученная в § 1 оценка.

Эта оценка применима и при  $\inf p(x) < 0$ , когда лемма 1.2 неприменима. Таким образом, использование аппарата функции Грина позволяет расширить множество задач, для которых удается получить оценку погрешности сеточного решения.

### § 3. Решение простейшей краевой сеточной задачи

При численном решении задачи Коши значения решения определяются в последовательных узлах по рекуррентным формулам; в случае краевой сеточной задачи, например (1.3), (1.4), такой возможности нет, поскольку значения решения зависят от граничных условий на обоих концах отрезка интегрирования.

Краевую задачу

$$y'' - p(x)y = f(x), \quad y(0) = a, \quad y(X) = b$$

можно было бы решать следующим способом. Возьмем частное решение  $y_0(x)$  неоднородного уравнения

$$y_0'' - p(x)y_0 = f(x)$$

и два линейно независимых решения однородного уравнения  $y_i'' - p(x)y_i = 0$ ,  $i = 1, 2$ . Общее решение неоднородного уравнения запишется в виде

$$y(x) = y_0(x) + C_1 y_1(x) + C_2 y_2(x);$$

постоянные  $C_1$  и  $C_2$  определяем из граничных условий. Приближения к функциям  $y_i(x)$ ,  $i = 0, 1, 2$ , находим каким-либо численным методом решения задачи Коши, затем определяем  $C_i$  и получаем нужное решение.

Экономнее поступить следующим образом. Находим частное решение неоднородного уравнения  $y_0'' - p(x)y_0 = f(x)$ , удовлетворяющее условию  $y_0(0) = a$ , и частное решение однородного уравнения  $y_1(x)$ , удовлетворяющее условию  $y_1(0) = 0$ . Общее решение неоднородного уравнения, удовлетворяющее условию  $y(0) = a$ , имеет вид  $y_0(x) + Cy_1(x)$ ; значение  $C$  определяется из условия  $y_0(X) + Cy_1(X) = b$ .

Метод решения краевой задачи, следующий этой схеме, принято называть *методом стрельбы* или *методом пристрелки*. Сеточный аналог этого метода заключается в следующем. Задаемся  $y_0^0 = a$ ,  $y_0^1 = 0$ , произвольными  $y_1^i \neq 0$  и  $y_1^0$  и из уравнений

$$l(y_0^n) = \frac{y_{n+1}^0 - 2y_n^0 + y_{n-1}^0}{h^2} - p_n y_n^0 = f_n, \quad (1)$$

$$l(y_1^n) = \frac{y_{n+1}^1 - 2y_n^1 + y_{n-1}^1}{h^2} - p_n y_n^1 = 0 \quad (2)$$

последовательно определяем  $y_2^0, \dots, y_N^0, y_2^1, \dots, y_N^1$ . Затем находим  $C$  из уравнения  $y_N^0 + Cy_N^1 = b$  и полагаем  $y_n = y_n^0 + Cy_n^1$ ; функция  $y_n$  является требуемым решением.

Если все указанные выше величины невозможно поместить в оперативной памяти ЭВМ, то значения  $y_n^i, i=0, 1$ , не хранят в процессе вычислений, а после отыскания  $C$  находят  $y_1 = y_1^0 + Cy_1^1$  и затем последовательно определяют  $y_2, \dots, y_{N-1}$  из уравнения  $l(y_n) = f_n$ . Описанный алгоритм формально применим при любых значениях  $y_1^0$  и  $y_1^1$ ; однако для уменьшения влияния вычислительной погрешности разумнее взять  $y_1^0 = a + O(h)$ .

Многие краевые задачи характерны тем, что среди решений соответствующих им однородных дифференциальных уравнений есть решения с сильным ростом (при убывании или при возрастании  $x$ ).

Рассмотрим модельное уравнение

$$y''(x) - py(x) = 0, \quad p = \text{const} > 0, \quad x \in [0, X]. \quad (3)$$

Решения соответствующего однородного уравнения  $y(x) = \exp\{\pm \sqrt{p}x\}$  сильно возрастают (убывают) на рассматриваемом участке, если величина  $\sqrt{p}X$  достаточно велика. Таким образом, при  $p > 0$  величина  $\sqrt{p}X$  является параметром, существенно влияющим на характер накопления вычислительной погрешности. Рассмотрим накопление вычислительной погрешности на одном из этапов решения сеточной задачи. Пусть уже найдено значение  $y_1$  и далее вычисления происходят по рекуррентной формуле

$$y_{n+1} = 2y_n - y_{n-1} + ph^2 y_n. \quad (4)$$

Получаемые при реальных вычислениях величины  $y_n^*$  связаны равенством

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + p h^2 y_n^* + \delta_n;$$

наличие слагаемого  $\delta_n$  в правой части обусловлено округлениями. Вычитая отсюда (4), получим уравнение относительно погрешности  $\Delta_n = y_n^* - y_n$ :

$$\Delta_{n+1} = 2\Delta_n - \Delta_{n-1} + p h^2 \Delta_n + \delta_n. \quad (5)$$

Рассмотрим следующую модель накопления погрешности:  $\delta_n = \delta = \text{const}$  и погрешности в  $y_0$  и  $y_1$  отсутствуют, т. е.  $\Delta_0 = \Delta_1 = 0$ . Уравнение (5) можно переписать в виде

$$\frac{\Delta_{n+1} - 2\Delta_n + \Delta_{n-1}}{h^2} = p\Delta_n + \omega, \quad \omega = \frac{\delta}{h^2}. \quad (6)$$

При  $\omega = \text{const}$  совокупность соотношений  $\Delta_0 = \Delta_1 = 0$  и (6) образует сеточную задачу, аппроксимирующую задачу Коши

$$\Delta'' = p\Delta + \omega, \quad \Delta(0) = \Delta'(0) = 0.$$

Можно проверить, что условия теоремы из § 8.10 выполнены, поэтому решения этих задач будут близки:

$$\Delta_n \sim \Delta(x_n) = \frac{\text{ch}(\sqrt{p} x_n) - 1}{p} \omega. \quad (7)$$

Асимптотическое равенство понимается здесь в обычном смысле:

$$\Delta_n / \Delta(x_n) \rightarrow 1 \quad \text{при} \quad h \rightarrow 0, \quad \sqrt{p} x_n = \text{const}, \quad x_n \neq 0. \quad (8)$$

При умножении  $\omega$  на какой-либо множитель на этот множитель умножится как  $\Delta_n$ , так и  $\Delta(x_n)$ , поэтому (7) и соответственно (8) справедливы и при  $\omega$ , зависящем от  $h$ , т. е. в рассматриваемом случае

$$\Delta_n \sim \alpha_n = \frac{\text{ch}(\sqrt{p} x_n) - 1}{p h^2} \cdot \delta.$$

Рассмотрим числовой пример:  $p = 10$ ,  $x_n = 10$ ,  $h = 10^{-2}$ ,  $\delta = 10^{-2}$ ; тогда  $|\sigma_n| > 10^4$  и нельзя рассчитывать на получение решения с разумной точностью. Отметим, что значение величины  $\sqrt{p} x_n$ , определившее столь большую величину накопленной вычислительной погрешности, было не очень большим.

При естественной нумерации неизвестных  $y_0, \dots, y_N$  система уравнений (1.3), (1.4) записывается в виде  $Ay = c$ , где матрица  $A$  трехдиагональная. Напомним, что матрица  $A = [a_{ij}]$  называется  $(2m + 1)$ -диагональной, если  $a_{ij} = 0$  при  $|i - j| > m$ .

Для решения систем уравнений такого вида часто наиболее целесообразно применять метод Гаусса при том же естественном порядке исключения неизвестных. В случае, когда этот метод применяется для решения систем уравнений, возникающих



при аппроксимации краевых задач, его называют *методом прогонки*.

Приведем конкретные расчетные формулы метода прогонки в случае системы (1.3), (1.4). Представим граничное условие  $y_0 = a$  в виде  $y_0 = C_0 y_1 + \Phi_0$ , где  $C_0 = 0$ ,  $\Phi_0 = a$ . Подставляя  $y_0 = C_0 y_1 + \Phi_0$  в первое уравнение системы (1.3) — (1.4)

$$\frac{y_0 - 2y_1 + y_2}{h^2} - p_1 y_1 = f_1,$$

получаем уравнение, связывающее значения  $y_1$  и  $y_2$ . Разрешая это уравнение относительно  $y_1$ , имеем

$$y_1 = C_1 y_2 + \Phi_1, \quad (9)$$

где

$$C_1 = \frac{1}{2 + p_1 h^2}, \quad \Phi_1 = C_1 (a - f_1 h^2).$$

Подставляя полученное выражение  $y_1$  через  $y_2$  во второе уравнение системы, получим уравнение, связывающее  $y_2$  и  $y_3$ , и т. д. Пусть уже получено соотношение

$$y_n = C_n y_{n+1} + \Phi_n; \quad (10)$$

подставим выражение  $y_n$  в  $(n + 1)$ -е уравнение системы (1.3) — (1.4):

$$\frac{y_n - 2y_{n+1} + y_{n+2}}{h^2} - p_{n+1} y_{n+1} = f_{n+1}.$$

Разрешая получившееся уравнение

$$\frac{(C_n y_{n+1} + \Phi_n) - 2y_{n+1} + y_{n+2}}{h^2} - p_{n+1} y_{n+1} = f_{n+1}$$

относительно  $y_{n+1}$ , имеем

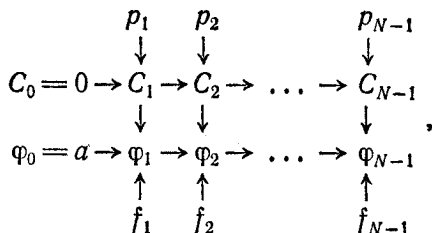
$$y_{n+1} = C_{n+1} y_{n+2} + \Phi_{n+1};$$

здесь

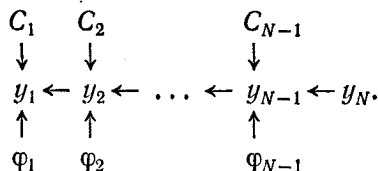
$$C_{n+1} = \frac{1}{2 + p_{n+1} h^2 - C_n}, \quad \Phi_{n+1} = C_{n+1} (\Phi_n - f_{n+1} h^2). \quad (11)$$

Таким образом, коэффициенты уравнений (10), связывающих последующие значения  $y_n$  и  $y_{n+1}$ , можно определять из рекуррентных соотношений (11) при начальных условиях  $C_0 = 0$ ,  $\Phi_0 = a$ . Так как  $y_N$  известно, то после нахождения всех коэффициентов  $C_n$ ,  $\Phi_n$  можно последовательно определять  $y_{N-1}$ , ...,  $y_1$  из соотношений (10). Процесс вычисления коэффициентов  $C_n$ ,  $\Phi_n$  принято называть *прямым ходом прогонки*, а процесс вычисления неизвестных  $y_n$  — *обратным ходом прогонки*. Последовательность производимых вычислений можно изобразить следующей схемой (на этой схеме  $a \rightarrow b$  означает, что значение  $a$  используется при вычислении  $b$ ):

## прямой ход прогонки



## обратный ход прогонки



Названию «метод прогонки» иногда предлагается следующее объяснение. Уравнение

$$y_n = C_n y_{n+1} + \Phi_n$$

получено как следствие граничного условия в точке  $x = 0$  и уравнений системы (1.3), соответствующих точкам  $x_j \leq x_n$ . Таким образом, это равенство выполнено для любого решения системы уравнений  $l(y_j) = f_j$ ,  $j = 1, \dots, n$ , удовлетворяющего левому граничному условию; граничное условие в точке  $x = 0$  «перегоняется» в текущую точку  $x = x_n$ .

**Задача 1.** Выписать расчетные формулы метода квадратного корня в случае решения рассматриваемой системы. Сравнить трудоемкость этого метода с трудоемкостью метода прогонки.

Применим для решения системы (1.3), (1.4) метод Гаусса при порядке исключения неизвестных  $y_0, y_2, \dots, y_N, y_1$ . Из первого уравнения выражаем  $y_2$  через  $y_1$  и подставляем в остальные уравнения. После этого во второе уравнение входят только неизвестные  $y_1$  и  $y_3$ ; выражаем  $y_3$  через  $y_1$  и подставляем в остальные уравнения и т. д. Пусть уже выразили через  $y_1$  неизвестные  $y_2, \dots, y_n$  и получили соотношения  $y_j = \alpha_j y_1 + \beta_j$  при  $2 \leq j \leq n$ ; для единообразия добавим сюда равенства

$$y_0 = \alpha_0 y_1 + \beta_0, \quad \text{где } \alpha_0 = 0, \quad \beta_0 = a,$$

$$y_1 = \alpha_1 y_1 + \beta_1, \quad \text{где } \alpha_1 = 1, \quad \beta_1 = 0.$$

Подставляя выражения  $y_{n-1}$  и  $y_n$  в уравнение

$$y_{n-1} - (2 + p_n h^2) y_n + y_{n+1} = f_n h^2,$$

получим

$$(\alpha_{n-1} y_1 + \beta_{n-1}) - (2 + p_n h^2) (\alpha_n y_1 + \beta_n) + y_{n+1} = f_n h^2,$$

или

$$y_{n+1} = \alpha_{n+1} y_1 + \beta_{n+1},$$

где

$$\alpha_{n+1} = (2 + p_n h^2) \alpha_n - \alpha_{n-1},$$

$$\beta_{n+1} = (2 + p_n h^2) \beta_n - \beta_{n-1} + f_n h^2. \quad (12)$$

Таким образом, коэффициенты  $\alpha_n, \beta_n$  можно последовательно вычислять по рекуррентным формулам (12). После получения соотношения  $y_N = \alpha_N y_1 + \beta_N$

определяем значение  $y_1$ , а затем все  $y_j$  по формулам

$$y_j = \alpha_j y_1 + \beta_j. \quad (13)$$

Вследствие (12) значения  $\alpha_j$  удовлетворяют однородному конечно-разностному уравнению (2) и начальным условиям  $\alpha_0 = 0$ ,  $\alpha_1 = 1$ ; значения  $\beta_j$  — неоднородному конечно-разностному уравнению (1) и начальным условиям  $\beta_0 = a$ ,  $\beta_1 = 0$ . Таким образом,  $\alpha_n$  совпадает с  $y_n^1$ , а  $\beta_n$  — с  $y_n^0$  в методе стрельбы при определенном выборе начальных условий. Функция  $z_n = \alpha_n C + \beta_n$  удовлетворяет неоднородному конечно-разностному уравнению и левому граничному условию при любых  $C$ , причем  $z_1 = \alpha_1 C + \beta_1 = C$ . Следовательно, решая уравнение  $y_n = \alpha_n C + \beta_n$  относительно  $C$ , мы как раз находим значение  $C = y_1$ , которое отыскивалось в методе стрельбы. Вычисления по формуле (13) соответствуют вычислению значений  $y_n$  по формуле  $y_n = y_n^1 C + y_n^0$ . Полученный метод совпадает с методом стрельбы при

$$y_0^0 = a, \quad y_1^0 = 0, \quad y_0^1 = 1, \quad y_1^1 = 0.$$

Решение системы (1.3), (1.4) описанными выше методами требует  $O(N)$  арифметических операций; если при решении этой системы обратиться непосредственно к стандартной программе метода Гаусса, то число операций будет порядка  $O(N^3)$ ; это количество операций состоит из  $O(N)$  *содержательных операций* метода прогонки и  $O(N^3)$  *несодержательных операций* умножения (деления) нуля на некоторое число, и сложения (вычитания) двух нулей. Таким образом, хотя содержательные операции при решении этой системы методом прогонки и по стандартной программе метода Гаусса одни и те же, использование стандартной программы приводит к существенному увеличению затрат на решение задачи.

Текст программы метода Гаусса может быть преобразован в текст программы метода прогонки, если в преобразованиях над строками и столбцами матрицы изменить начало и конец циклов так, чтобы исключить несодержательные операции.

Рассмотрим случай, когда при  $x=0$  имеем граничное условие  $y'(0) - \alpha y(0) = a$ . В § 1 при его аппроксимации возникло уравнение, связывающее значения  $y_0$  и  $y_1$ ; это уравнение может быть записано в виде

$$y_0 = C_0 y_1 + \Phi_0; \quad (14)$$

например, простейшую аппроксимацию  $\frac{y_1 - y_0}{h} - \alpha y_0 = a$  можно переписать в таком виде при  $C_0 = \frac{1}{1 + \alpha h}$ ,  $\Phi_0 = -\frac{\alpha h}{1 + \alpha h}$ . Далее вычисляем  $C_n$ ,  $\Phi_n$  по формулам (11) при  $n = 1, \dots, N-1$  и при обратном ходе прогонки  $y_{N-1}, \dots, y_0$  по формулам (10). Если при  $x = X$  имеем граничное условие  $y'(X) + \beta y(X) = b$ , то аналогично (14) имеем уравнение

$$y_N = \bar{C}_N y_{N-1} + \bar{\Phi}_N. \quad (15)$$

Осуществляя прямой ход прогонки, получим равенство

$$y_{N-1} = C_{N-1} y_N + \Phi_N; \quad (16)$$

решая систему (15), (16), находим  $y_N, y_{N-1}$  и затем последовательно вычисляем  $y_{N-2}, \dots, y_0$  по формулам (10).

При аппроксимации краевых задач для уравнений высших порядков или для систем дифференциальных уравнений появляются системы уравнений  $Ay = c$  с  $(l, s)$ -диагональными матрицами  $A$ ; матрицу  $A = [a_{ij}]$  называют  $(l, s)$ -диагональной, если  $a_{ij} = 0$  при  $j < i - l$  и при  $j > i + s$ . Для решения таких уравнений также довольно часто бывает целесообразно применять метод Гаусса, алгоритм которого может быть записан аналогично методу прогонки в виде совокупности рекуррентных формул.

Если  $l > s$ , то для уменьшения числа арифметических операций целесообразно переобозначить неизвестные и уравнения в обратном порядке, чтобы получить систему с  $(s, l)$ -диагональной матрицей.

**Задача 2.** Подсчитать число операций при решении методом Гаусса системы с  $(l, s)$ -диагональной матрицей при  $l > s, l = s, l < s$ .

В ряде случаев  $(l, s)$ -диагональная матрица системы уравнений записывается естественным образом в виде  $(p, q)$ -диагональной матрицы клеточного вида, т. е.  $A = [A_{ij}]$ ,  $A_{ij}$  — некоторые матрицы такие, что  $A_{ij} = 0$ , если  $j < i - p$  или  $j > i + q$ .

Рассмотрим случай системы уравнений

$$y'' - p(x)y = f(x), \quad (17)$$

где  $y, f$  — векторы размерности  $m$ ,  $p$  — матрица размерности  $m \times m$ , и простейшую аппроксимацию

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p(x_n)y_n = f(x_n). \quad (18)$$

Матрица системы естественным образом записывается в виде (17) с  $p = q = 1$ ;  $A_{ij}$  — матрицы размерности  $m \times m$ . В то же самое время эта матрица является  $(2m - 1, 2m - 1)$ -диагональной или, что то же самое,  $(4m - 1)$ -диагональной. Для решения этой системы может быть применен метод исключения Гаусса в клеточной форме, который аналогично скалярному случаю может быть записан в виде совокупности рекуррентных матричных соотношений типа формул метода прогонки.

**Задача 3.** Подсчитать число арифметических операций для метода Гаусса в клеточной форме и для общей процедуры метода Гаусса с исключением несодержательных операций в применении к системе уравнений (18).

При решении ряда задач возникают системы уравнений с матрицей  $A$ , отличающейся по структуре от  $(l, s)$ -диагональной матрицы наличием ненулевых элементов при  $|i - j| \sim n$ , т. е. вблизи левого нижнего и правого верхнего углов матрицы. Для решения таких систем также целесообразно применять метод Гаусса с исключением несодержательных операций. В случае отыскания периодического решения сеточного уравнения (3) этот вариант метода Гаусса называют *методом циклической прогонки*.

Рассмотрим пример задачи, сводящейся к системе уравнений такого вида.

При сглаживании функций методом регуляризации в § 5.3 возникла следующая задача. Дана периодическая функция  $f_q$  целочисленного аргумента  $q$ ;

период равен  $N$ . Требуется найти периодическую с тем же периодом функцию  $u_q$ , удовлетворяющую системе соотношений

$$\frac{\delta^{2n} u_q}{h^{2n}} + (-\lambda^2)^n u_q = f_q \quad \text{при всех } q.$$

Выпишем эти соотношения при  $q = 1, \dots, N$ . Вследствие условия периодичности заменим значения  $u_q$  при  $q \leq 0$  и при  $q > N$ , входящие в эту систему, соответственно на  $u_{q+N}$  и  $u_{q-N}$ . В результате этого получится система  $N$  уравнений относительно  $N$  неизвестных  $u_1, \dots, u_N$ :  $Au = f$ . Элементы матрицы  $A = [a_{ij}]$  этой системы определяются соотношениями  $a_{ij} = a(|i-j|)$ , где

$$a(0) = (-1)^n (C_{2n}^n h^{-2n} + \lambda^{2n}),$$

$$a(k) = \begin{cases} (-1)^{n-k} C_{2n}^{n-k} h^{-2n} & \text{при } 0 < k \leq n, \\ 0 & \text{при } n < k \leq N-n, \\ (-1)^{k+n-N} C_{2n}^{k+n-N} h^{-2n} & \text{при } N-n \leq k < N; \end{cases}$$

$$A = \begin{bmatrix} a(0) & a(1) & \dots & a(n) & 0 & \dots & 0 & a(N-n) & \dots & a(N-1) \\ a(1) & \cdot & \cdot & \cdot & & & & & & \\ \cdot & \cdot & \cdot & \cdot & & & & & & \\ \cdot & \cdot & \cdot & \cdot & & & & & & \\ a(n) & & & \cdot & \cdot & \cdot & & & & \\ 0 & & & \cdot & \cdot & \cdot & \cdot & & & \\ \cdot & & & \cdot & \cdot & \cdot & \cdot & \cdot & & \\ \cdot & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ 0 & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ a(N-n) & & & & & & & \cdot & \cdot & \\ \cdot & & & & & & & \cdot & \cdot & \\ \cdot & & & & & & & \cdot & \cdot & \\ a(N-1) & & & & & & & & \cdot & \cdot \end{bmatrix}.$$

Матрица  $A$  симметричная и положительно определенная.

Решение этой системы методом Гаусса с исключением несодержательных операций требует  $O(n^2 N)$  арифметических операций.

Задача 4. Выписать расчетные формулы этого метода при  $n = 1$ .

Задача 5. Выписать расчетные формулы метода квадратного корня в конкретном случае решения этой системы при  $n = 1$ .

Выше при рассмотрении метода прогонки выписывались расчетные формулы и подсчитывалось число арифметических действий; был оставлен в стороне вопрос о возможности переполнения ячеек ЭВМ, в частности, вследствие деления на нуль. Кроме того, отсутствие переполнения само по себе не гарантирует избавления от большого влияния вычислительной погрешности.

На модельном примере  $p(x) \equiv p = \text{const}$  рассмотрим поведение прогоночных коэффициентов  $C_n$  при различном знаке  $p$ . Соотношения

$$C_{n+1} = \frac{1}{2 + ph^2 - C_n}, \quad (19)$$

которым удовлетворяют коэффициенты  $C_n$  метода прогонки, совпадают с итерационными формулами решения уравнения

$$C = g(C) = \frac{1}{2 + ph^2 - C}$$

при начальном приближении  $C_0 = 0$ . Уравнение  $C = g(C)$  равносильно квадратному уравнению

$$C^2 - (2 + ph^2)C + 1 = 0; \quad (20)$$

его корни

$$C_{(1)}, C_{(2)} = 1 + \frac{ph^2}{2} \pm \sqrt{ph^2 + \frac{p^2 h^4}{4}}.$$

При  $ph^2 > 0$  очевидно, что  $C_{(1)} > 1$ , поэтому  $0 < C_{(2)} = C_{(1)}^{-1} < 1$ . При  $p < 0$  и  $h$  малом подкоренное выражение отрицательно и уравнение (20) не имеет вещественных корней. Совместное расположение графиков  $y = C$  и  $y = g(C)$  и точек  $(C_n, C_n)$ ,  $(C_n, C_{n+1})$  изображено на рис. 9.3.1, 9.3.2. Видно, что при  $p \geq 0$  значения  $C_n$  лежат в пределах  $0 < C_n \leq 1$  (рис. 9.3.1). При

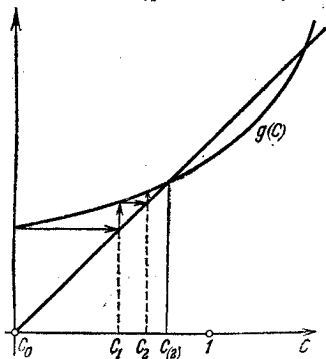


Рис. 9.3.1

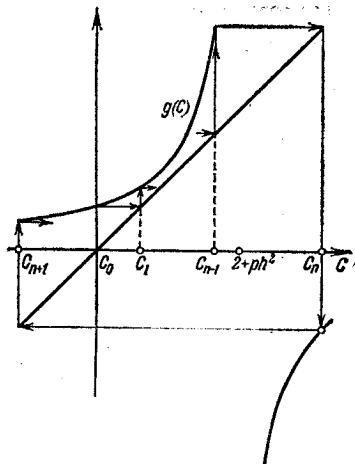


Рис. 9.3.2

$p < 0$  не исключено (рис. 9.3.2), что при достаточно больших  $n$  некоторое значение  $C_n$  окажется близким к  $2 + ph^2$ ; тогда следующее значение  $C_{n+1}$  будет очень большим и может произойти переполнение.

Рассмотрим вопрос о накоплении погрешности при вычислении коэффициентов  $C_n$ . Из (19) получаем, что

$$\frac{\partial C_{n+1}}{\partial C_n} = \frac{1}{(2 + ph^2 - C_n)^2} = (C_{n+1})^2.$$

Следовательно, возмущения в коэффициентах  $C_n$  связаны соотношением  $\delta C_{n+1} \approx C_{n+1}^2 \delta C_n$  и при больших  $C_n$  происходит также большой рост погрешности. Отмеченные выше обстоятельства приводят к необходимости более детального изучения влияния вычислительной погрешности, по крайней мере при  $p \leq 0$ .

## § 4. Замыкания вычислительных алгоритмов

При предварительном анализе алгоритмов каждый раз удавалось многое понять, рассматривая случай решения модельной задачи. Другим эффективным приемом предварительного анализа является исследование замыканий вычислительных алгоритмов.

Чтобы не загромождать изложения, ограничимся рассмотрением существа проблемы; поэтому многие из проводимых в этом параграфе построений не всегда подробно обосновываются.

Пусть решается некоторая задача

$$Lu = f \quad (1)$$

и пусть

$$L^h u^h = f^h \quad (2)$$

— последовательность задач, зависящих от параметра  $h$  (например, шага сетки), решения которых  $u^h$  сходятся при  $h \rightarrow 0$  к решению исходной задачи (1). Предположим, что алгоритм решения задачи (2)  $L^h u^h = f^h$  состоит в последовательном получении некоторых соотношений

$$L_m^h u_m^h = f_m^h, \quad m = 1, \dots, M; \quad (3)$$

при этом  $L_M^h = E$  — единичный оператор,  $u_M^h = f_M^h = (L^h)^{-1} f^h = u^h$ , т. е. на  $M$ -м шаге получается точное решение  $u^h$ . Пусть можно ввести параметр  $z(m, h)$ , монотонно зависящий от  $m$ , такой, что при  $h \rightarrow 0$  и  $z$  фиксированном соотношение (3) переходит в пределе в соотношение

$$L_z y = f_z, \quad 0 \leq z \leq z_0; \quad (4)$$

здесь

$$z_0 = \lim_{h \rightarrow 0} z(M, h).$$

Соотношение (4) называется *замыканием вычислительного алгоритма* (3).

Мы не дали строгого определения замыкания вычислительного алгоритма. Существо дела станет понятнее после рассмотрения замыканий алгоритмов решения краевой задачи.

Определение понятия вычислительного алгоритма не исключает возможности  $M = \infty$ . В этом случае равенства

$$L_M^h = E, \quad f_M^h = (L^h)^{-1} f^h$$

понимаются в том смысле, что

$$L_m^h \rightarrow E, \quad f_m^h \rightarrow (L^h)^{-1} f^h \quad \text{при } m \rightarrow \infty.$$

Случай  $M = \infty$  соответствует итерационным методам решения задачи (1). Если операторы  $L_z$  равномерно ограничены по  $z$ , то говорят, что алгоритм (3) имеет *регулярное замыкание*. В противном случае говорят, что алгоритм имеет *нерегулярное замыкание*.

Если замыкание алгоритма регулярно, то есть основания предположить, что он устойчив к различным возмущениям, в частности к вычислительной погрешности. Поэтому исследование замыкания алгоритма является удобным способом получения предварительного суждения о свойствах нового метода на первоначальной стадии изучения вопроса. Такая предварительная оценка свойств метода не всегда оказывается окончательной. Может случиться, что операторы  $L_z$  равномерно ограничены, но очень большой постоянной, что равносильно практической неограниченности. С другой стороны, возможно, что неограниченность операторов  $L_z$  вызывается неудачной нормировкой уравнений или неудачным выбором норм в пространствах функций  $u$ . Отсутствие равномерной ограниченности операторов  $L_z$  означает неограниченное возрастание  $p(h) = \sup_m \|L_m^h\|$  при  $h \rightarrow 0$ . Не исключена возможность, что величина  $p(h)$  растет очень медленно и соответствующее возрастание влияния вычислительной погрешности при стремлении шага к нулю окажется допустимым. Несмотря на высказанные соображения, изучение замыканий вычислительных алгоритмов приносит большую пользу.

При прямом ходе прогонки получаются соотношения

$$y_n - C_n y_{n+1} = \varphi_n. \quad (5)$$

Попробуем понять, во что переходят эти соотношения при  $h \rightarrow 0$ . Возьмем простейший случай  $p \equiv 0$ . Тогда  $C_0 = 0$ ,  $C_1 = 1/2$ ,  $C_2 = 2/3$ , ... и, как нетрудно установить методом индукции,  $C_n = 1 - 1/(n+1)$ . Пусть фиксирована какая-то точка  $x = nh$ . Тогда левая часть (5), соответствующая этой точке, равна

$$y_n - \left(1 - \frac{1}{x/h + 1}\right) y_{n+1}.$$

При подстановке вместо величин  $y_n$  значений гладкой функции  $y(nh)$  предел левой части равен нулю и нет смысла рассматривать подобное замыкание алгоритма. Чтобы получаемое замыкание было осмысленным, нужно подобрать такой множитель  $\lambda(h)$ , чтобы предел

$$\lim_{h \rightarrow 0} \lambda(h) \left( y(x) - \left(1 - \frac{1}{x/h + 1}\right) y(x+h) \right) \quad (6)$$

был конечен и отличен от нуля. Возьмем  $\lambda(h) = h^{-1}$ ; тогда

$$\begin{aligned} & \frac{1}{h} \left( y(x) - \left(1 - \frac{1}{x/h + 1}\right) y(x+h) \right) = \\ & = \frac{1}{h} \left( y(x) - \left(1 - \frac{h}{x} + O(h^2)\right) (y(x) + y'(x)h + O(h^2)) \right) = \\ & = \frac{y(x)}{x} - y'(x) + O(h). \end{aligned}$$

Таким образом, (6) имеет ненулевой предел. Применим такую же нормировку и в общем случае. Поскольку в рассмотренном



примере существовал предел

$$\lim_{h \rightarrow 0} \frac{C_n - 1}{h} \quad \text{при } nh = x,$$

введем вместо  $C_n$  новую переменную  $\alpha_n = (1 - C_n)/h$ . Напомним, что прогоночные соотношения имеют вид

$$C_{n+1} = \frac{1}{2 + p_{n+1}h^2 - C_n}, \quad \varphi_{n+1} = C_{n+1}(\varphi_n - f_{n+1}h^2). \quad (7)$$

Умножим соотношение (5) на  $-h^{-1}$  и положим  $C_n = 1 - \alpha_n h$ ,  $\varphi_n = -\beta_n h$ . Тогда (5) запишется в виде

$$\frac{y_{n+1} - y_n}{h} - \alpha_n y_{n+1} = \beta_n. \quad (8)$$

Подставив выражения  $C_n$  и  $\varphi_n$  через  $\alpha_n$  и  $\beta_n$  в (7), получим равенства

$$1 - \alpha_{n+1}h = \frac{1}{1 + p_{n+1}h^2 + \alpha_n h},$$

$$\beta_{n+1}h = (1 - \alpha_{n+1}h)(\beta_n h + f_{n+1}h^2);$$

отсюда

$$\alpha_{n+1} = \frac{\alpha_n + p_{n+1}h}{1 + \alpha_n h + p_{n+1}h^2}, \quad (9)$$

$$\beta_{n+1} = \beta_n + (f_{n+1} - \alpha_{n+1}\beta_n)h - \alpha_{n+1}f_{n+1}h^2. \quad (10)$$

Соотношение (9) можно преобразовать к виду

$$\frac{\alpha_{n+1} - \alpha_n}{h} = p_{n+1} - \alpha_n^2 - h \frac{2p_{n+1}\alpha_n - \alpha_n^3 - (p_{n+1}^2 - p_{n+1}\alpha_n^2)h}{1 + \alpha_n h + p_{n+1}h^2}, \quad (11)$$

а соотношение (10) — к виду

$$\frac{\beta_{n+1} - \beta_n}{h} = f_{n+1} - \alpha_{n+1}\beta_n - f_{n+1}\alpha_{n+1}h. \quad (12)$$

Если предположить, что коэффициенты  $\alpha_n$  и  $\beta_n$  при фиксированном  $nh = x$  и  $h \rightarrow 0$  стремятся к некоторым пределам  $\alpha(x)$  и  $\beta(x)$ , то при подстановке в (8) вместо  $y_n$  значений  $y(nh)$  в пределе получается дифференциальное уравнение

$$y' - \alpha(x)y = \beta(x).$$

Соотношения (11), (12) показывают, что величины  $\alpha_n$ ,  $\beta_n$  удовлетворяют уравнениям, напоминающим метод Эйлера численного интегрирования дифференциальных уравнений

$$\alpha' = p(x) - \alpha^2, \quad (13)$$

$$\beta' = f(x) - \alpha\beta. \quad (14)$$

Доказательство того факта, что значения  $\alpha_n$ ,  $\beta_n$  при  $nh = x$  фиксированном,  $h \rightarrow 0$ , стремятся к значениям решений этих диффе-

ренциальных уравнений, затрудняется следующим обстоятельством. Согласно определению

$$\alpha_0 = (1 - C_0)/h = 1/h, \quad \beta_0 = -\varphi_0/h = -a/h. \quad (15)$$

Таким образом, в окрестности точки  $x=0$  начальные условия для численного интегрирования по формулам (11), (12) имеют не совсем обычный характер и развитые нами методы оценки близости решений сеточных и дифференциальных задач в рассматриваемом случае неприменимы. Глядя на соотношения (15), можно было бы предположить, что функция  $\alpha(x)$  является решением (13), ведущим себя вблизи нуля, как  $x^{-1}$ , а  $\beta(x)$  — решением (14), ведущим себя вблизи нуля, как  $-ax^{-1}$ .

Обоснуем это предположение.

Коэффициенты  $C_n$ , а следовательно, и  $\alpha_n$  получаются из рекуррентных соотношений (7), не зависящих ни от правой части дифференциального уравнения, ни от граничного условия в точке  $X$ ; начальное условие рекурсии  $C_0 = 0$  также от них не зависит. Рассмотрим краевую задачу

$$y''(x) - p(x)y(x) = 0, \quad y(0) = 0, \quad y(X) = 1$$

и соответствующую сеточную задачу

$$l(y_l) = 0, \quad y_0 = 0, \quad y_N = 1.$$

Решения этих задач можно записать в виде

$$y(x) = W^1(x)/W^1(X), \quad y_n = W_n^1/W_N^1$$

(см. § 2). Поскольку для этой краевой задачи  $\varphi_0 = 0$  и все  $f_n = 0$ , то  $\beta_n \equiv 0$  и соотношение (8) имеет вид

$$\frac{y_{n+1} - y_n}{h} - \alpha_n y_{n+1} = 0.$$

Подставляя сюда  $y_n = W_n^1/W_N^1$ , получим

$$\alpha_n = \left( \frac{W_{n+1}^1 - W_n^1}{h} \right) (W_{n+1}^1)^{-1}. \quad (16)$$

Пользуясь оценками близости

$$|W_n^1 - W^1(nh)| \leq Mh^2 \quad \text{при } 0 \leq nh \leq X, \quad (17)$$

можно получить оценку

$$\alpha_n = \frac{(W^1)'|_{(n+1)h} + O(h)}{W^1((n+1)h) + O(h^2)}. \quad (18)$$

Функция

$$\alpha(x) = (W^1)'|_x / W^1(x)$$

удовлетворяет дифференциальному уравнению (13), поскольку

$$\left( \frac{(W^1)'}{W^1} \right)' = \frac{(W^1)''}{W^1} - \left( \frac{(W^1)'}{W^1} \right)^2 = p(x) - \left( \frac{(W^1)'}{W^1} \right)^2.$$

Согласно определению  $W^1(x)$ , при малых  $x$  имеем  $W^1(x) \sim x$ ,  $(W^1)'|_x \sim 1$  и, таким образом,  $\alpha(x) = (W^1)'|_x / W^1(x) \sim x^{-1}$ , как и предполагалось.

Функция  $\alpha(x)$  обращается в бесконечность в точках, где  $W^1(x) = 0$ , в частности при  $x = 0$ . В окрестности каждой такой точки  $y$  имеем

$$(W^1)'|_x \sim (W^1)'|_y \neq 0, \quad W^1(x) \sim ((W^1)'|_y)(x - y)$$

и, таким образом,

$$\alpha(x) \sim (x - y)^{-1}. \quad (19)$$

Для точек этой окрестности можно написать цепочки соотношений

$$\frac{W_{n+1}^1 - W_n^1}{h} = (W^1)' |_{(n+1)h} (1 + O(h)),$$

$$\begin{aligned} W_{n+1}^1 &= W^1((n+1)h) + O(h^2) = W^1((n+1)h) \left(1 + O\left(\frac{h^2}{W^1((n+1)h)}\right)\right) = \\ &= W^1((n+1)h) \left(1 + O\left(\frac{h^2}{(n+1)h - y}\right)\right) \end{aligned}$$

и, таким образом,

$$\alpha_n = \alpha((n+1)h) (1 + O(h)) \left(1 + O\left(\frac{h^2}{(n+1)h - y}\right)\right).$$

Отсюда видно, что относительная погрешность равенства  $\alpha_n \approx \alpha((n+1)h)$  является величиной порядка  $O(h)$  вплоть до окрестности порядка  $h$  точек, где  $\alpha(x) = \infty$ . Аналогичным образом исследуется поведение функции  $\beta(x)$ .

Выпишем теперь соотношения (3) и соответствующие замыкания алгоритма. Возьмем  $M = 2N - 1$  и при  $m = 1, \dots, N$  за (3) примем совокупность соотношений

$$\begin{aligned} \frac{y_n - C_n y_{n+1}}{h} &= \frac{\varphi_n}{h}, \quad 0 \leq n < m, \\ \frac{y_{n-1} - 2y_n + y_{n+1}}{h^2} - p_n y_n &= f_n, \quad m \leq n < N, \\ y_N &= b, \end{aligned}$$

а при  $N < m \leq M$  — совокупность соотношений

$$\begin{aligned} \frac{y_n - C_n y_{n+1}}{h} &= \frac{\varphi_n}{h}, \quad 0 \leq n \leq M - m, \\ y_n &= \psi_n, \quad M - m < n \leq N, \end{aligned}$$

где  $\varphi_n$  — точное решение сеточной задачи. Положим  $z = mh$ . Тогда замыкание алгоритма (4) будет выглядеть следующим образом:

$$\begin{aligned} L_z y &= f_z(x), \quad \text{где для } 0 \leq z < 1 \\ L_z y &= \begin{cases} y' - \alpha(x)y & \text{при } 0 < x < z, \\ y'' - p(x)y & \text{при } z \leq x < 1, \\ y & \text{при } x = 1, \end{cases} \\ f_z(x) &= \begin{cases} \beta(x) & \text{при } 0 < x < z, \\ f(x) & \text{при } z \leq x < 1, \\ b & \text{при } x = 1 \end{cases} \end{aligned}$$

и для  $1 \leq z \leq 2$

$$L_z y = \begin{cases} y' - \alpha(x)y & \text{при } 0 < x < 2 - z, \\ y & \text{при } 2 - z \leq x \leq 1, \end{cases}$$

$$f_z(x) = \begin{cases} \beta(x) & \text{при } 0 < x < 2 - z, \\ \psi(x) & \text{при } 2 - z \leq x \leq 1; \end{cases}$$

здесь  $\psi(x)$  — точное решение дифференциальной задачи (равенство  $L_z y|_{x=0} = a$  опущено).

Рассмотрим примеры поведения функции  $\alpha(x)$  и  $\beta(x)$ .

Пусть  $p(x) > 0$ ; тогда, согласно уравнению  $\alpha' = p(x) - \alpha^2$ ,

$$\alpha' < 0 \quad \text{при} \quad \alpha > p_1 = \sqrt{\max_{[0, X]} p(x)}$$

и

$$\alpha' > 0 \quad \text{при} \quad \alpha < p_2 = \sqrt{\min_{[0, X]} p(x)}.$$

Соответствующее поле интегральных кривых изображено на рис. 9.4.1. Поэтому решение  $\alpha(x)$ , удовлетворяющее условию  $\alpha(x) \sim x^{-1}$  при  $x \rightarrow 0$ , монотонно убывает по крайней мере до тех пор, пока не попадет в область  $p_2 < \alpha < p_1$ , где оно и останется. Уравнение (14) линейно относительно  $\beta(x)$ . Так как коэффициент  $\alpha(x)$  конечен при  $x \neq 0$ , то решение этого уравнения остается ограниченным при всех  $x \neq 0$ . Таким образом, были бы все основания признать замыкание алгоритма решения рассматриваемой задачи регулярным, если бы не неограниченность коэффициентов  $\alpha(x)$  и  $\beta(x)$  при  $x \rightarrow 0$ . В рассматриваемом случае значения  $\alpha_0$  и  $\beta_0$  в точке  $n = 0$  являются величинами порядка  $h^{-1}$ ; далее с ростом  $n$  они должны иметь тенденцию к убыванию из-за аналогичной тенденции решений дифференциальных уравнений. Мы уже свыклись с тем, что во многих случаях влияние вычислительной погрешности порядка  $2^{-l}h^{-p}$  является неизбежным и допустимым и возможно, что такая нерегулярность алгоритма не приведет к недопустимо большому влиянию вычислительной погрешности.

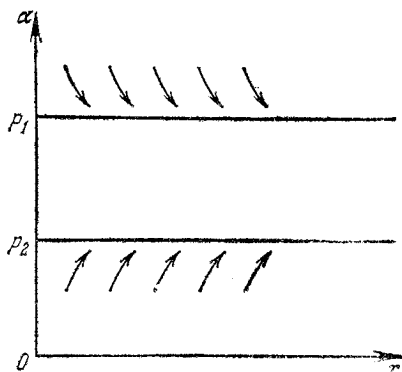


Рис. 9.4.1

Обратимся к случаю, когда не всюду  $p(x) \geq 0$ . Тогда может оказаться, что в некоторой внутренней точке отрезка  $W^1(y) = 0$  и, следовательно,  $\alpha(y) = \infty$ . В этом случае замыкание алгоритма вычислений следует признать нерегулярным; однако и здесь не следует спешить окончательно отказываться от приме-

нения метода прогонки. Если оказалось, что во всех узлах сетки  $W^1(nh)h^{-2} \gg 1$ , то, согласно (16) — (18), имеем

$$\max_n |\alpha_n| = O(h^{-2}). \quad (20)$$

Такого рода нерегулярность алгоритма также не всегда следует считать катастрофической. Функция  $W^1(x)$  — гладкая с производной, отличной от нуля в точках  $y$ , где  $W^1(y) = 0$ . Поэтому часто значение  $\min_{n>0} W^1(nh)$  будет оказываться величиной порядка  $h$  и соотношение (20) будет выполняться. Таким образом, следует ожидать, что и в случае  $p(x) \leq 0$  вычисления по методу прогонки в большинстве случаев будут устойчивыми к различного рода возмущениям.

Оказалось, что замыкание алгоритма прогонки, по существу, регулярно, если  $W^1(x) \neq 0$  при  $x \in (0, X]$ . Условие  $W^1(x_0) \neq 0$  равносильно условию, что краевая задача

$$y(0) = a, \quad y(x_0) = b, \quad y'' - p(x)y = f(x) \quad (21)$$

разрешима на  $[0, x_0]$  при любых  $a$ ,  $b$  и  $f(x)$ . Таким образом, условие регулярности замыкания алгоритма прогонки можно сформулировать в следующем виде.

*Замыкание метода прогонки регулярно, если для любого отрезка  $[0, x_0]$  при  $0 < x_0 \leq X$  краевая задача (21) разрешима.*

Получим этот вывод еще одним путем. Обратимся к формулам обратного хода прогонки. Коэффициенты  $C_j$  и  $\varphi_j$  при  $j < n$  зависят только от  $a$  и значений  $p_j$  и  $f_j$  при  $j < n$ , т. е. в точках  $jh \in [0, nh]$ . Следовательно, после отыскания значения  $y_n$  вычисления происходят так же, как в случае краевой задачи для уравнения  $y'' - p(x)y = f(x)$  на отрезке  $[0, nh]$  при заданных  $y(0) = a$  и  $y(nh) = y_n$ . Если эта краевая задача неразрешима при любых правых частях  $f(x)$ , то возникают сомнения в хороших свойствах соответствующей сеточной задачи. Поэтому в случае, когда при каком-то  $x_0 \in (0, X]$  краевая задача (21) разрешима не при любых правых частях, необходим более детальный анализ метода.

Отметим, что регулярность замыкания еще не обеспечивает малости суммарной вычислительной погрешности в силу следующих обстоятельств. Суммарная погрешность определяется погрешностями в ходе вычислений и множителями пропорциональности, с которыми эти погрешности входят в суммарную погрешность. Регулярность замыкания, как правило, обеспечивает лишь малость погрешностей округления в ходе вычислений. Чтобы множители пропорциональности не были большими, нужна слабая чувствительность решения уравнений замыкания к возмущениям коэффициентов. Однако и этого, вообще говоря, недостаточно. На примере уравнения  $y' = My$  при  $M < 0$  мы видели, что зачастую решение дифференциальной задачи слабо чувствительно к таким возмущениям, в то время как решение сеточной задачи сильно чувствительно.

## § 5. Обсуждение постановок краевых задач для линейных систем первого порядка

Рассмотрим краевую задачу

$$y' - A(x)y = f(x), \quad By(0) = b, \quad Dy(X) = d; \quad (1)$$

здесь  $y, f, b, d$  — векторы размерностей соответственно  $l, l, l-r, r$ , а  $A, B, D$  — матрицы размерностей  $l \times l, (l-r) \times l, r \times l$ . Всюду в дальнейшем предполагается, что ранг матрицы  $B$  равен  $l-r$ , а ранг матрицы  $D$  равен  $r$ .

Прежде чем отыскивать практически пригодные методы решения задачи (1), обсудим вопрос о чувствительности решений краевых задач к разного рода возмущениям. В качестве модели возьмем краевую задачу

$$y' - Ay = 0, \quad A = \text{const}, \quad By(0) = b, \quad Dy(X) = d. \quad (2)$$

Рассмотрим случай, когда все собственные значения матрицы  $A$  различны; в общем случае ход рассуждений изменяется несущественно. Пусть  $\lambda_j = \alpha_j + i\beta_j$  — собственные значения матрицы  $A$ , упорядоченные в порядке возрастания  $\alpha_j$ , а  $e_j$  — соответствующие собственные векторы, причем  $\|e_j\| = 1$ . Общее решение системы  $y' - Ay = 0$  запишется в виде

$$y(x) = \sum_{j=1}^l c_j e_j \exp\{\lambda_j x\}. \quad (3)$$

Собственные значения  $\lambda_j$  разделим на три группы, присваивая собственным значениям каждой группы соответствующий верхний индекс. Собственные значения  $\lambda_j$ , для которых величина  $\exp\{|\alpha_j|X\}$  очень большая, обозначаем как  $\lambda_j^+$ , если  $\alpha_j > 0$ , и как  $\lambda_j^-$ , если  $\alpha_j < 0$ . Остальные собственные значения, т. е. те, для которых величина  $\exp\{|\alpha_j|X\}$  не очень велика, обозначаем как  $\lambda_j^0$ . Соответственно снабдим верхними индексами собственные векторы  $e_j$ ; суммирование по индексам  $j$ , соответствующим этим группам, будем обозначать  $\sum^+, \sum^0, \sum^-$ . Пусть  $l^+, l^0, l^-$  — число собственных значений в соответствующих группах. Форма записи решения (3) ставит в неодинаковое положение концы отрезка интегрирования: все функции  $\exp\{\lambda_j x\}$  в точке  $x=0$  имеют порядок 1, в то время как в точке  $x=X$  одни из них очень большие, другие очень малые. Удобнее другая форма записи общего решения:

$$y(x) = \sum^- c_j^- e_j^- \exp\{\lambda_j^- x\} + \sum^0 c_j^0 e_j^0 \exp\{\lambda_j^0 x\} + \sum^+ c_j^+ e_j^+ \exp\{\lambda_j^+ (x-X)\}.$$

Выпишем систему уравнений  $By(0) = b, Dy(X) = d$ , из которой следует определить постоянные  $c_j$ , соответствующие отыски-

ваемому решению:

$$\begin{aligned} By(0) &= \sum^- c_j^- Be_j^- + \sum^0 c_j^0 Be_j^0 + \sum^+ c_j^+ Be_j^+ \exp\{-\lambda_j^+ X\} = b \\ Dy(X) &= \sum^- c_j^- Be_j^- \exp\{\lambda_j^- X\} + \sum^0 c_j^0 Be_j^0 \exp\{\lambda_j^0 X\} + \\ &\quad + \sum^+ c_j^+ Be_j^+ = d. \end{aligned} \quad (4)$$

Систему уравнений (4) можно записать в форме

$$Gc = g;$$

здесь

$$c = (c_1^-, \dots, c_l^+)^T, \quad g = (b_1, \dots, b_{l-r}, d_1, \dots, d_r)^T,$$

матрица  $G$  имеет клеточный вид:

$$G = \begin{bmatrix} G_1^- & G_1^0 & G_1^+ \\ G_2^- & G_2^0 & G_2^+ \end{bmatrix} \begin{matrix} l-r \\ r \\ r^- \quad l^0 \quad l^+ \end{matrix},$$

в котором клетки записываются следующим образом:

$$\begin{aligned} G_1^- &= [(Be_j^-)_q], \quad 1 \leq j \leq l^-, \quad 1 \leq q \leq l-r, \\ G_1^0 &= [(Be_j^0)_q], \quad l^- < j \leq l^- + l^0, \quad 1 \leq q \leq l-r, \\ G_1^+ &= [(Be_j^+)_q \exp\{-\lambda_j^+ X\}], \quad l^- + l^0 < j \leq l, \quad 1 \leq q \leq l-r, \\ G_2^- &= [(De_j^-)_q \exp\{\lambda_j^- X\}], \quad 1 \leq j \leq l^-, \quad 1 \leq q \leq r, \\ G_2^0 &= [(De_j^0)_q \exp\{\lambda_j^0 X\}], \quad l^- < j \leq l^- + l^0, \quad 1 \leq q \leq r, \\ G_2^+ &= [(De_j^+)_q], \quad l^- + l^0 < j \leq l, \quad 1 \leq q \leq r. \end{aligned}$$

Предположим, что  $\Delta = \det G \neq 0$ , и представим решение системы (4) в виде  $c = G^{-1}g$ . Согласно известным формулам элементы обратной матрицы имеют вид

$$G^{-1} = \left[ \frac{(-1)^{i+j} G_{ij}}{\Delta} \right], \quad (5)$$

где  $G_{ij}$  — миноры матрицы  $G$ . При сформулированных ранее предположениях величины  $\exp\{-\lambda_j^+ X\}$ ,  $\exp\{\lambda_j^- X\}$  ничтожно малы, поэтому ничтожно малы элементы матриц  $G_1^+$  и  $G_2^-$  и определитель

$$\Delta = \det \begin{bmatrix} G_1^- & G_1^0 & G_1^+ \\ G_2^- & G_2^0 & G_2^+ \end{bmatrix}$$

близок к определителю

$$\Delta_0 = \det G^0, \quad G^0 = \begin{bmatrix} G_1^- & G_1^0 & 0 \\ 0 & G_2^0 & G_2^+ \end{bmatrix}.$$

Рассмотрим отдельно случаи:

а) определитель  $\Delta_0$  не мал;

б) определитель  $\Delta_0$  мал, в частности равен нулю.

Поскольку среди элементов матрицы  $G$  нет больших, то, вследствие формулы (5), в случае а) элементы матрицы  $G^{-1}$  обычно не очень большие.

Предположим, что правые части граничных условий  $b$  и  $d$  содержат некоторые погрешности  $\delta b$  и  $\delta d$ . Пусть  $\delta g = (\delta b, \delta d)^T$ , тогда погрешность вектора  $c$  равна  $G^{-1}\delta g$ . Если элементы матрицы  $G^{-1}$  не очень велики, то влияние погрешности  $\delta g$  на коэффициенты  $c_j$  будет не очень большим. Решение задачи является линейной комбинацией с коэффициентами  $c_j$  слагаемых

$$e_j^- \exp\{\lambda_j^- x\}, \quad e_j^0 \exp\{\lambda_j^0 x\}, \quad e_j^+ \exp\{\lambda_j^+(x - X)\}.$$

Поэтому погрешность приближенного решения, являющаяся следствием погрешностей  $\delta b$  и  $\delta d$ , также будет приемлемой.

Заметим, что наши высказывания носят довольно неопределенный характер: «малый», «очень малый», «небольшой», «очень большой»; при анализе конкретной задачи исследователь должен сам решать, насколько приемлем для него тот или иной порядок рассматриваемых величин. В частности, если решается система «большого» порядка, то при «умеренных» значениях коэффициентов системы и «не очень малом» определителе  $\Delta_0$  возможно, что миноры, состоящие из сумм произведений большого числа элементов, окажутся «недопустимо большими».

Если определитель  $\Delta_0$  очень мал или равен нулю, то, вследствие равенства

$$\det G \det G^{-1} = 1,$$

среди элементов матрицы  $G^{-1}$  встретятся большие. Тогда малые возмущения правых частей граничных условий могут приводить к большим возмущениям коэффициентов  $c_j$ , а следовательно, и решения задачи.

Зная элементы матрицы  $G^{-1}$  и собственные значения матрицы  $A$ , из полученных выше соотношений можно получить довольно точную информацию о возмущении решения дифференциальной задачи. Однако получение этой информации само по себе требует большого объема вычислений; перенос этих построений на случай переменной матрицы  $A(x)$  потребует еще большего объема вычислений. Попытаемся поэтому получить критерии устойчивости решения к возмущениям  $\delta b$  и  $\delta d$  качественного характера, требующие меньшей информации о задаче, хотя, может быть, и несколько менее надежные. Таким критерием могут служить соотношения между числами  $l^-$ ,  $l^0$ ,  $l^+$ ,  $l-r$  и  $r$ . Среди элементов первых  $l-r$  строк матрицы  $G^0$  ненулевые элементы могут находиться в первых  $l^- + l^0$  столбцах, соответствующих матрицам  $G_1^-$ ,  $G_1^0$ . Если  $l^+ > r$ , то  $l^- + l^0 < l-r$ , и тогда все миноры порядка  $(l-r) \times (l-r)$ , лежащие в первых  $l-r$  строках, обращаются в нуль. Раскрывая определитель  $\Delta_0$



по первым  $l-r$  строкам, получаем  $\Delta_0 = 0$ . Точно так же если  $l^+ > l-r$ , то все миноры порядка  $r \times r$ , лежащие в последних  $r$  строках, обращаются в нуль, поэтому  $\Delta_0 = 0$ . Если  $l^+ \leq r$ , а  $l^- \leq l-r$ , то определитель  $\Delta_0$  окажется линейной комбинацией произведений элементов матриц  $B$  и  $D$  и координат собственных векторов  $e_j$ , причем коэффициентами при этих произведениях будут произведения чисел  $\exp\{\lambda_j^0 X\}$ , не очень больших и не очень маленьких по модулю. Можно принять гипотезу, что этот определитель оказывается малым числом довольно редко. Тогда решение задачи (2) будет малочувствительно к возмущениям  $\delta b$  и  $\delta d$  правых частей граничных условий.

Мы можем сформулировать полученные выводы в качестве следующего предложения. *Если  $l^+ > r$  или  $l^- > l-r$ , то решение дифференциальной задачи сильно чувствительно к возмущениям правых частей граничных условий; если  $l^+ \leq r$ , а  $l^- \leq l-r$ , то, как правило, решение задачи (2) будет мало чувствительно к изменениям правых частей граничных условий.*

Первую часть этого утверждения можно переформулировать еще в такой форме: *для малой чувствительности задачи к возмущениям граничных условий необходимо, чтобы число независимых частных решений  $e_j \exp\{\lambda_j x\}$ , сильно растущих на  $[0, X]$  с ростом  $x$ , не превосходило числа граничных условий на правом конце, а число частных решений  $e_j \exp\{\lambda_j x\}$ , сильно убывающих на  $[0, X]$  с ростом  $x$ , не превосходило числа граничных условий на левом конце.*

Эта формулировка при определенных уточнениях может быть перенесена и на случай задачи (1) с переменной матрицей  $A(x)$ . Строгая переформулировка этого утверждения будет довольно громоздкой; однако если элементы матрицы  $A(x)$  относительно плавно меняются на  $[0, X]$ , то при первоначальном исследовании устойчивости задачи к возмущениям граничных условий зачастую можно ограничиться подсчетом числа собственных значений матриц  $A(x)$  с большим положительным и большим отрицательным значениями величины  $X \operatorname{Re} \lambda_j(x)$ .

Краевую задачу называют *хорошо обусловленной* (хорошо поставленной), если малые возмущения коэффициентов и правых частей уравнения и граничных условий приводят к столь же малым по порядку изменениям решения задачи. Более аккуратное определение хорошей обусловленности можно дать следующим образом. Наряду с краевой задачей (1) рассмотрим краевые задачи

$$\begin{aligned} \tilde{y}' - (A(x) + \delta A(x)) \tilde{y} &= f(x) + \delta f(x), \\ (B + \delta B) \tilde{y}(0) &= b + \delta b, \quad (D + \delta D) \tilde{y}(X) = d + \delta d \end{aligned}$$

с не очень большой мерой возмущения

$$\varepsilon = \max_{0 \leq x \leq X} (\|\delta A(x)\| + \|\delta f(x)\|) + \|\delta B\| + \|\delta D\| + \|\delta b\| + \|\delta d\|.$$

Если для всех решений таких краевых задач выполняются неравенства

$$\max_{0 \leq x \leq X} \|\tilde{y}(x) - y(x)\| \leq M\varepsilon \quad (6)$$

с не очень большим значением постоянной  $M$ , то исходную задачу называют *хорошо обусловленной*, в противном случае задачу называют *плохо обусловленной*. Минимальное значение  $M(\varepsilon_0)$ , при котором неравенство (6) выполняется при всех  $\varepsilon \leq \varepsilon_0$  ( $\varepsilon_0 > 0$  фиксировано), иногда называют *мерой обусловленности* данной задачи (относительно возмущений с нормой, не большей  $\varepsilon_0$ ). Обусловленность задачи характеризует устойчивость решения к возмущениям исходных данных, например к неточности задания коэффициентов уравнения. Поскольку погрешности от округления при вычислениях эквивалентны возмущениям коэффициентов исходного уравнения, то мера обусловленности характеризует и устойчивость численного решения к возможным округлениям при численном решении. Если известна ориентировочная оценка  $\varepsilon$  возмущения коэффициентов задачи и погрешность порядка  $M(\varepsilon)\varepsilon$  допустима, то имеет смысл непосредственное численное решение задачи.

Рассмотрим в качестве примера задачу Коши для системы  $y'_1 = y_2$ ,  $y'_2 = -y_1$  при  $y_1(0) = 1$ ,  $y_2(0) = 1$  на отрезке  $[0, 30]$ . Собственные значения матрицы  $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$  равны  $\pm 1$ . Величина  $\exp(-30)$  очень малая, а  $\exp(30)$  — очень большая,  $l^- = 1$ ,  $l^0 = 0$ ,  $l^+ = 1$ ,  $r = 0$  и  $l^+ = 1 > r = 0$ . Поэтому малые возмущения граничного условия должны приводить к очень большим изменениям решения. В данном случае проводившиеся нами в этом параграфе построения не имеют особого смысла, поскольку и без них ясно, что, вследствие сильного роста решений исходного линейного уравнения, погрешность решения растет очень быстро.

Пусть для той же системы рассматривается задача с краевыми условиями  $b_1 y_1(0) + b_2 y_2(0) = b$ ,  $d_1 y_1(30) + d_2 y_2(30) = d$ ; собственные векторы, соответствующие собственным значениям  $\lambda_1^- = -1$  и  $\lambda_2^+ = 1$ , равны соответственно  $(1, -1)^T$  и  $(1, 1)^T$ ,  $l^+ = r$ ,  $l^- = l - r$ . Вообще говоря, следует ожидать, что задача будет устойчива к возмущениям  $\delta b$  и  $\delta d$ . Решение отыскивается в виде

$$c_1^- \begin{pmatrix} 1 \\ -1 \end{pmatrix} \exp\{-x\} + c_2^+ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \exp\{x - 30\}.$$

Уравнения (4) имеют вид

$$\begin{aligned} (b_1 - b_2) c_1^- + (b_1 + b_2) c_2^+ \exp\{-30\} &= b, \\ (d_1 - d_2) c_1^- \exp\{-30\} + (d_1 + d_2) c_2^+ &= d. \end{aligned} \quad (7)$$

Отсюда

$$\begin{aligned} c_1^- &= \frac{b(d_1 + d_2) - d(b_1 + b_2) \exp\{-30\}}{\Delta}, \\ c_2^+ &= \frac{d(b_1 - b_2) - b(d_1 - d_2) \exp\{-30\}}{\Delta}, \end{aligned}$$

где

$$\Delta = (b_1 - b_2)(d_1 + d_2) - (d_1 - d_2)(b_1 + b_2) \exp\{-60\}$$

близко к  $\Delta_0 = (b_1 - b_2)(d_1 + d_2)$ . Если коэффициенты  $b_i$  и  $d_i$  небольшие, а  $\Delta_0$  не мало, то коэффициенты  $c_1^-$  и  $c_2^+$  мало изменяются при малых изменениях  $\delta b$  и  $\delta d$ . Если  $\Delta = 0$ , то система (7) имеет ненулевое решение  $c_1^-, c_2^+$  при  $b = d = 0$ . В этом случае говорят, что «задача лежит на спектре», т. е. имеется в виду, что однородная задача имеет ненулевое решение и бессмысленно говорить об устойчивости решения к возмущениям граничных условий. При  $\Delta_0 = 0$  задача плохо обусловлена, поскольку  $\Delta \approx \Delta_0$ .

**Задача 1.** Доказать, что решение хорошо обусловленной краевой задачи единственно.

## § 6. Алгоритмы решения краевых задач для систем уравнений первого порядка

В дальнейшем, в § 6—8, предполагается, что рассматриваемая краевая задача хорошо обусловлена.

Простейшим по форме методом решения краевой задачи (5.1) является *метод стрельбы*. Рассмотрим систему уравнений

$$By(0) = b. \quad (1)$$

Поскольку по предположению ранг матрицы  $B$  равен  $l - r$ , то общее решение системы (1) записывается в виде

$$y_0 + \sum_{i=1}^r c_i y_i;$$

здесь  $y_0$  — произвольное решение неоднородной системы  $By = b$ , а  $y_1, \dots, y_r$  — произвольная система  $r$  линейно независимых решений системы  $By = 0$ . Пусть  $y_1, \dots, y_r, y_0$  — какой-то набор таких векторов. Численным интегрированием найдем частное решение неоднородной системы

$$y'_0 = A(x)y_0 + f(x) \quad (2)$$

при начальном условии  $y_0(0) = y_0$  и решения однородной системы

$$y'_j = A(x)y_j, \quad j = 1, \dots, r, \quad (3)$$

при начальных условиях  $y_j(0) = y_j$ . Пусть  $y(x)$  удовлетворяет (2) и левому граничному условию (1). Вектор  $y(0)$  является решением (1), и поэтому его можно записать в виде

$$y(0) = y_0 + \sum_{j=1}^r c_j y_j;$$

вектор-функция

$$z(x) = y_0(x) + \sum_{j=1}^r c_j y_j(x)$$

удовлетворяет уравнению (2) и совпадает с  $y(x)$  при  $x = 0$ . Следовательно,

$$y(x) = y_0(x) + \sum_{j=1}^r c_j y_j(x). \quad (4)$$

Всякая функция вида (4) удовлетворяет соотношениям (1) и (2). Таким образом, многообразие всех решений (2), удовлетворяющих левому граничному условию (1), задается равенством (4). Чтобы найти искомое решение, надо выделить из этого многообразия решение, удовлетворяющее правому граничному условию. Определим коэффициенты  $c_j$  из системы  $r$  уравнений с  $r$  неизвестными

$$D \left( y_0(X) + \sum_{j=1}^r c_j y_j(X) \right) = d. \quad (5)$$

В предположении однозначной разрешимости задачи (5.1) определитель этой системы отличен от нуля. Действительно, предположив противное, мы получили бы, что однородная крайняя задача ( $f \equiv 0$ ,  $b = 0$ ,  $d = 0$ ) имеет ненулевое решение

$$y(x) = \sum_{j=1}^r c_j y_j(x).$$

Если  $c_1, \dots, c_r$  — решение системы (5), то вектор-функция

$$y(x) = y_0(x) + \sum_{j=1}^r c_j y_j(x)$$

удовлетворяет уравнению и всем граничным условиям и, следовательно, является решением искомой задачи.

При необходимости экономить память следует найти

$$y(0) = y_0(0) + \sum_{j=1}^r c_j y_j(0),$$

или

$$y(X) = y_0(X) + \sum_{j=1}^r c_j y_j(X)$$

и затем решить численно задачу Коши вперед или назад.

Если среди решений однородной системы  $y' = A(x)y$  есть быстро растущие с ростом  $X$ , то столь же быстро может возрастать вычислительная погрешность в решениях  $y_0(x), \dots, y_r(x)$ . В этом случае алгоритм метода стрельбы окажется непригодным для практического использования. По-другому практическую непригодность метода стрельбы в этом случае можно объяснить следующим образом. Пусть метод стрельбы применяется при  $A = \text{const}$ . Обозначим через  $e_j$  и  $\lambda_j$  собственные векторы и соответствующие собственные значения матрицы  $A$ , причем пусть  $\text{Re } \lambda_1 \leq \dots \leq \text{Re } \lambda_{l-1} < \text{Re } \lambda_l$ . Предположим, что

$$y_j = \sum_{k=1}^l \alpha_{jk} e_k;$$

тогда

$$y_j(X) = \sum_{k=1}^l \alpha_{jk} e_k \exp\{\lambda_k X\}. \quad (6)$$

Если  $\alpha_{jl} \neq 0$  и  $\exp\{\operatorname{Re} \lambda_l X\} \gg \exp\{\operatorname{Re} \lambda_{l-1} X\}$ , то

$$y_j(X) \sim \alpha_{jl} e_l \exp\{\lambda_l X\}.$$

Таким образом, все столбцы матрицы системы (5) оказываются приблизительно пропорциональными вектору  $De_l$ , поэтому решение этой системы  $c_1, \dots, c_r$  будет найдено с большой вычислительной погрешностью.

Рассматриваемую задачу можно трактовать как задачу выделения из многообразия

$$y_0(x) + \sum_{j=1}^r c_j y_j(x) \quad (7)$$

вектора, удовлетворяющего правому граничному условию

$$Dy(X) = d.$$

При каждом фиксированном  $x$  множество концов векторов вида (4) образует  $r$ -мерную плоскость в  $l$ -мерном пространстве; плоскость задается концом вектора  $y_0(x)$ , лежащим в этой плоскости, и векторами  $y_1(x), \dots, y_r(x)$ , лежащими в ней. Если бы эти векторы задавались точно, то было бы несущественно, какими векторами задавать плоскость. Однако вследствие погрешностей в значениях этих векторов эта плоскость будет несколько смещена и повернута. Предположим, что исходная краевая задача хорошо поставлена и соответственно норма вектора  $y(x)$  в каждой точке  $x \in [0, X]$  невелика. В рассматриваемом случае при больших значениях  $\operatorname{Re}(\lambda_k x)$  вектор  $y_0(x)$  имеет большую норму, а векторы  $y_j(x)$  примерно пропорциональны. Небольшие возмущения  $y_0(x)$  и векторов  $y_j(x)$  приведут к существенному изменению положения плоскости в области относительно небольших значений  $\|y\|$ , т. е. там, где находятся значения точного решения. Поэтому при таком способе задания многообразия (4) существенно теряется информация о решении.

Чтобы положение плоскости в области небольших значений норм векторов  $\|y\|$ , где находится решение, было устойчиво к погрешностям такого способа ее задания, представляется целесообразным задавать плоскость точкой, являющейся проекцией на нее начала координат, или точкой, близкой к ней, и некоторым множеством векторов, лежащих в этой плоскости и образующих ортогональный репер или близкий к таковому. Существо наиболее распространенных методов прогонки решения задачи (5.1) состоит в непрерывном или дискретном (в отдельных точках отрезка) переходе к заданию многообразия (4) при помощи точки проекции начала координат на эту плоскость и ортогонального репера, лежащего в этой плоскости.

В одном из вариантов метода прогонки поступают следующим образом. Отрезок интегрирования разбивается на части точками  $0 = x_0 < x_1 < \dots < x_m = X$ . Пусть на отрезке  $[x_s, x_{s+1}]$  решение задачи отыскивалось в виде линейной комбинации

$$y(x) = g_0^s(x) + \sum_{j=1}^r c_j g_j^s(x);$$

здесь  $g_0^s(x)$  — решение неоднородной системы (2);  $g_j^s(x)$ ,  $j = 1, \dots, r$ , — решения однородной системы (3). Путем последовательной ортогонализации и нормировки векторов  $g_j^s(x_{s+1})$ ,  $j = 1, \dots, r$ , получают систему ортонормированных векторов  $g_j^{s+1}(x_{s+1})$ ,  $j = 1, \dots, r$ . Полагают

$$g_0^{s+1}(x_{s+1}) = g_0^s(x_{s+1}) - \sum_{j=1}^r (g_0^s(x_{s+1}), g_j^{s+1}(x_{s+1})) g_j^{s+1}(x_{s+1}),$$

т. е. вычитают из вектора  $g_0^s(x_{s+1})$  его проекцию на плоскость, натянутую на векторы  $g_j^{s+1}(x_{s+1})$ ,  $j = 1, \dots, r$ . Далее опять на отрезке  $[x_{s+1}, x_{s+2}]$  ищут  $g_0^{s+1}(x)$  и  $g_j^{s+1}(x)$ ,  $j = 1, \dots, r$ , как решения систем (2) и (3) соответственно. После получения такого представления решения в точке  $X$  находят решение задачи на отрезке  $[x_{m-1}, x_m]$ , воспользовавшись граничным условием в точке  $X = x_m$ . Далее, пользуясь формулами перехода между совокупностями векторов

$$(g_0^s(x_{s+1}), \dots, g_r^s(x_{s+1})) \text{ и } (g_0^{s+1}(x_{s+1}), \dots, g_r^{s+1}(x_{s+1})),$$

последовательно находят решение на отрезках  $[x_{m-2}, x_{m-1}]$ ,  $\dots$ ,  $[x_0, x_1]$ . Точки  $x_1, \dots, x_{m-1}$  выбираются из условия, чтобы система векторов  $g_1^s(x_{s+1}), \dots, g_r^s(x_{s+1})$  была близка ортонормированной, а вектор  $g_0^s(x_{s+1})$  не образовывал слишком малый угол с подпространством, натянутым на этот базис. Поскольку векторы  $g_1^s(x_s), \dots, g_r^s(x_s)$  образуют ортонормированную систему, а вектор  $g_0^s(x_s)$  ортогонален к ним, сформулированные условия выполнены при достаточно малой величине  $\max_s (x_{s+1} - x_s)$ .

В другом варианте метода прогонки ортогонализация базиса, лежащего в плоскости (4), производится в каждой расчетной точке; при этом новый базис строится не путем последовательной ортогонализации, а так, чтобы он в определенном смысле был наиболее близок к старому. Такая ортогонализация может осуществляться в двух вариантах — непрерывном и дискретном. Получающаяся близость базиса к старому усложняет расчетные формулы, но зато позволяет осуществлять интегрирование с более крупным шагом.

Еще в одном варианте метода прогонки подпространство (4) задается в каждой точке вектором нормали и базисом ортогонального ему подпространства. Базис ортогонального подпространства имеет размерность  $l-r$  и при  $l-r < r$  переход к такому базису приводит к уменьшению объема вычислений. Такой подход целесообразен в случае многоточечных задач, напри-

мер, когда решается система уравнений  $l$ -го порядка и задано  $l$  условий на решение в  $l$  различных точках отрезка  $[0, X]$ . Подобного рода задачи возникают, в частности, при определении параметров дифференциального уравнения по результатам наблюдения решения.

Заметим, что в отличие от подробно рассмотренного в § 3, 4 метода прогонки для дифференциального уравнения второго порядка методы прогонки, основанные на идее ортогонализации, при достаточно малой величине  $\max_s (x_{s+1} - x_s)$  оказываются слабо чувствительными к влиянию вычислительной погрешности для любых хорошо определенных краевых задач.

## § 7. Нелинейные краевые задачи

Существует большое сходство между методами решения этих задач и нелинейных алгебраических задач. В частности, так же, как в последнем случае, мы проведем обсуждение различных методов, не приходя в конце концов к конкретной рекомендации по решению произвольных нелинейных краевых задач. По существу, всякий раз решение нелинейной краевой задачи сводится к решению некоторой нелинейной системы уравнений. Различные методы решения нелинейных краевых задач отличаются выбором параметров этих вспомогательных задач и, естественно, методом решения этих задач.

Рассмотрим простейший пример — нелинейную краевую задачу: найти решение уравнения

$$\begin{aligned} x'' - f(t, x) &= 0 \quad \text{при } t \in (0, T), \\ x(0) - a &= 0, \quad x(T) - b = 0. \end{aligned} \quad (1)$$

Предположим, что известно некоторое приближение к решению  $x_n(t)$ ; в окрестности этого приближения справедливо разложение

$$f(t, x) \approx f(t, x_n(t)) + \frac{\partial f}{\partial x}(t, x_n(t))(x - x_n(t)),$$

поэтому представляется целесообразным искать следующее приближение к решению  $x_{n+1}(t)$  как решение краевой задачи

$$\begin{aligned} x''_{n+1} - (f(t, x_n(t)) + \frac{\partial f}{\partial x}(t, x_n(t))(x_{n+1} - x_n(t))) &= 0, \\ x_{n+1}(0) - a &= 0, \quad x_{n+1}(T) - b = 0. \end{aligned} \quad (2)$$

Это же самое уравнение относительно следующего приближения к решению получится при формальном применении схемы метода Ньютона. Рассматриваем (5.1) как операторное уравнение  $F(x) = 0$ ; оператором производной  $F'$  будет оператор  $P$ , определяемый равенствами

$$P\eta = \begin{cases} \eta'' - f_x(t, x(t))\eta & \text{при } t \in (0, T), \\ \eta(0) & \text{при } t = 0, \\ \eta(T) & \text{при } t = T. \end{cases}$$

$$\begin{aligned} \text{Уравнение метода Ньютона относительно } x_{n+1}(t) \text{ записывается в виде} \\ x_n''(t) - l(t, x_n(t)) + (x_{n+1}(t) - x_n(t))'' - l_x(t, x_n(t))(x_{n+1}(t) - x_n(t)) = 0, \\ x_n(0) - a + (x_{n+1}(0) - x_n(0)) = 0, \\ x_n(T) - b + (x_{n+1}(T) - x_n(T)) = 0 \end{aligned}$$

и совпадает с (2).

Рассмотрим сеточную аппроксимацию задачи (1):

$$\begin{aligned} \frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} - f(t_k, x_k) = 0, \quad k = 1, \dots, N-1, \quad (3) \\ x_0 - a = 0, \quad x_N - b = 0; \end{aligned}$$

здесь  $h = T/N$ ,  $x_k$  — приближения к значениям  $x(kh)$ . Пусть  $x_k^n$ ,  $k = 0, \dots, N$ , — совокупность величин, образующих  $n$ -е приближение к решению системы (3). В окрестности этого приближения справедливо соотношение

$$f(t_k, x_k) \approx f(t_k, x_k^n) + f_x(t_k, x_k^n)(x_k - x_k^n).$$

Поэтому следующее приближение отыскиваем из системы уравнений

$$\begin{aligned} \frac{x_{k+1}^{n+1} - 2x_k^{n+1} + x_{k-1}^{n+1}}{h^2} - (f(t_k, x_k^n) + f_x(t_k, x_k^n)(x_{k+1}^{n+1} - x_k^n)) = 0, \\ k = 1, \dots, N-1, \quad (4) \\ x_0^{n+1} - a = 0, \quad x_N^{n+1} - b = 0. \end{aligned}$$

В § 3 мы изучали применение методов стрельбы и прогонки для решения таких линейных уравнений.

**Задача 1.** Выписать расчетные формулы метода Ньютона для системы (3) и убедиться, что они совпадают с (4).

В обоих вариантах, непрерывном и дискретном, мы построили сначала итерационные методы (2) и (4), линеаризуя правую часть, а потом убедились, что эти методы совпадают с итерационным методом Ньютона.

Заметим, что итерационный метод (2), как правило, не может быть реализован из-за невозможности решения уравнения (2) в явном виде и на практике имеют дело именно с методом (4).

Метод (4) может представлять практические неудобства из-за необходимости хранить в памяти ЭВМ все значения  $x_k^n$ ,  $k = 0, \dots, N$ . Поэтому часто прибегают к следующему методу, который также будет изложен в непрерывном и дискретном вариантах.

Исходная задача будет решена, если будет найдено значение  $x'_0$  производной решения в точке 0. Решая задачу Коши для уравнения  $x'' = f(t, x)$  при начальных условиях  $x(0) = a$ ,



$x'(0) = x'_0$ , получим решение на всем отрезке. Если решить задачу Коши при начальных условиях  $x(0) = a$  и произвольном  $x'(0)$ , то получим значение  $x(T)$ , вообще говоря, отличное от  $b$ . Значение  $x(T)$  решения задачи Коши является некоторой функцией от  $x'(0)$ :  $x(T) = \varphi(x'(0))$ . Таким образом, исходная задача сводится к решению нелинейного уравнения

$$F(x'(0)) = \varphi(x'(0)) - b = 0. \quad (5)$$

Нахождение значения  $F(\alpha)$  при каждом  $\alpha$  требует решения задачи Коши

$$x'' = f(t, x), \quad x(0) = a, \quad x'(0) = \alpha. \quad (6)$$

В связи с большой трудоемкостью нахождения значения  $F(\alpha)$  при каждом  $\alpha$  для решения уравнения должны применяться методы, требующие малого вычисления значений функции  $F(\alpha)$ .

В ряде сложных случаев задача решается в диалоговом режиме; значение  $F(\alpha_n)$  для каждого последующего приближения  $\alpha_n$  вычисляется ЭВМ, а выбор  $\alpha_{n+1}$  осуществляется исследователем, решающим задачу. Подобного рода диалоговые системы применяются на практике много столетий. В случае конкретных работающих промышленных систем руководитель часто подбирает значение  $\alpha_{n+1}$ , изучая результат работы системы при ранее взятых значениях  $\alpha_1, \dots, \alpha_n$ . В случае артиллерийской стрельбы вычисление значений  $F(\alpha_n)$  или  $\text{sign } F(\alpha_n)$  осуществляется при помощи посылки снаряда в цель, а выбор  $\alpha_{n+1}$  осуществляется лицом корректирующим стрельбу. Можно найти много аналогов таких алгоритмов в повседневной жизни человека и животных.

Рассмотрим вопрос о решении уравнения (5) методом Ньютона. Согласно формуле дифференцирования решения уравнения второго порядка, справедливо равенство

$$\frac{\partial F(x'(0))}{\partial x'(0)} = \frac{\partial x(T)}{\partial x'(0)} = \eta(T);$$

$\eta(t)$  — решение задачи Коши

$$\begin{aligned} \eta'' - f_x(t, x(t))\eta &= 0, \\ \eta(0) &= 0, \quad \eta'(0) = 1. \end{aligned} \quad (7)$$

Численно или аналитически решая эту задачу, находим значение  $\eta(T)$ .

Поскольку задачу Коши (6), как правило, требуется решать численно, исследуем сразу дискретный вариант этого метода. Рассмотрим сеточную аппроксимацию (3); зададимся произвольным  $x_1$  и, пользуясь рекуррентной формулой, вытекающей из (3),

$$x_{k+1} = 2x_k - x_{k-1} + h^2 f(t_k, x_k), \quad k = 1, \dots, N-1, \quad (8)$$

при начальных условиях  $x_0 = a, x_1$ , находим сеточную функцию, удовлетворяющую сеточному уравнению и левому граничному,

условию. Значение  $x_N$  является некоторой функцией от  $x_1$ ,  $x_N = \varphi(x_1)$ , вычисление каждого значения которой производится с помощью рекуррентного процесса (8). Таким образом, решение задачи (3) сводится к решению скалярного уравнения

$$F(x_1) = \varphi(x_1) - b = 0.$$

Предположим, что для решения этого уравнения используется метод Ньютона. Продифференцировав (8) по  $x_1$ , получим

$$\eta_{k+1} = 2\eta_k - \eta_{k-1} + h^2 \frac{\partial f(t_k, x_k)}{\partial x_k} \eta_k, \quad k = 1, \dots, N-1; \quad (9)$$

здесь  $\eta_k = \partial x_k / \partial x_1$ . Кроме того, имеем

$$\eta_0 = \frac{\partial x_0}{\partial x_1} = \frac{\partial a}{\partial x_1} = 0; \quad \eta_1 = \frac{\partial x_1}{\partial x_1} = 1.$$

Таким образом, при заданном  $x_1$ , производя параллельно вычисления  $x_k$  по формуле (8) и  $\eta_k$  по формуле (9), можно найти  $x_N(x_1)$  и  $\partial x_N / \partial x_1 = \eta_N$  и осуществить следующий шаг метода Ньютона:

$$x_1^{n+1} = x_1^n - \left( \frac{\partial x_N}{\partial x_1^n} \right)^{-1} (\varphi(x_1^n) - b). \quad (10)$$

Если соотношение (9) переписать в виде

$$\frac{\eta_{k+1} - 2\eta_k + \eta_{k-1}}{h^2} - \frac{\partial f(t_k, x_k)}{\partial x_k} \eta_k = 0,$$

то оно превращается в разностную схему, аппроксимирующую уравнение (7).

Выше рассматривается случай, когда (1) — скалярное уравнение. Тогда вместо (10) часто целесообразно вычислять  $\partial x_N / \partial x_1^n$  по приближенной формуле

$$\frac{\partial x^N}{\partial x_1^n} \approx \frac{x^N(x_1^n) - x^N(x_1^n - \Delta_n)}{\Delta_n}, \quad (11)$$

в частности, можно брать  $\Delta_n = x_1^n - x_1^{n-1}$ .

Заметим, что в случае использования формулы (10) и вычисления производной с помощью (11), нужно обращать внимание на разумный выбор величины  $\Delta_n$  (см. § 2.16), поскольку величина погрешности значения  $x^N(x_1)$ , получаемого путем численного интегрирования, часто оказывается довольно большой.

Рассмотрим нелинейную краевую задачу

$$\begin{aligned} y' &= f(x, y), & B(y(0)) &= 0, & D(y(X)) &= 0, \\ y &= (y_1, \dots, y_l)^T, & B &= (b_1, \dots, b_{l-r})^T, & D &= (d_1, \dots, d_r)^T. \end{aligned} \quad (12)$$

Пусть функции  $g_1(y(0)), \dots, g_r(y(0))$  таковы, что система уравнений

$$\begin{aligned} b_1(y(0)) &= 0, \dots, b_{l-r}(y(0)) = 0, \\ g_1(y(0)) &= g_1, \dots, g_r(y(0)) = g_r \end{aligned} \quad (13)$$

однозначно определяет вектор  $y(0) = \omega_0(g)$ ,  $g = (g_1, \dots, g_r)$ . Тогда задача (12) может быть сведена к системе нелинейных уравнений относительно параметров  $g_1, \dots, g_r$ .

Довольно часто граничное условие в точке 0 имеет вид

$$y_1(0) = 0, \dots, y_{l-r}(0) = 0;$$

тогда в качестве  $g_1(y(0)), \dots, g_r(y(0))$  целесообразно взять  $y_{l-r+1}(0), \dots, y_l(0)$ . Таким образом, здесь в качестве искомым параметров выступают неизвестные компоненты решения в точке  $x = 0$ .

Решая систему (13) при каждом  $g_1, \dots, g_r$ , можно определить вектор  $y(0) = \omega_0(g)$ ; решая задачу Коши при начальном условии  $y(0)$ , определяем  $y(X) = \omega_x(g)$  и затем находим  $\psi(g) = D(\omega_x(g))$ . Таким образом, решение задачи (12) сводится к решению нелинейной системы уравнений

$$\psi(g) = 0. \quad (14)$$

Значительную группу методов решения нелинейных систем составляют методы типа метода Ньютона, где наряду со значениями функций  $\psi_i$  используются значения их производных  $\partial\psi_i/\partial g_k$ . Наиболее распространены следующие два способа нахождения этих производных.

1. Для определенности речь пойдет о вычислении производных  $\partial\psi_i/\partial g_1$ . Пусть мы задались значениями  $g_1^0, \dots, g_r^0$ . Из системы уравнений (13) найдем соответствующие  $y_1(0), \dots, y_l(0)$ . Дифференцируя уравнения системы (13) по  $g_1$ , получаем систему уравнений для отыскания производных  $\partial y_j(0)/\partial g_1$ :

$$\sum_{j=1}^l \frac{\partial b_k}{\partial y_j(0)} \frac{\partial y_j(0)}{\partial g_1} = 0, \quad k = 1, \dots, l-r,$$

$$\sum_{j=1}^l \frac{\partial g_m}{\partial y_j(0)} \frac{\partial y_j(0)}{\partial g_1} = \delta_m^1, \quad m = 1, \dots, r.$$

Решение системы  $y' = f(x, y)$  при начальных условиях  $y_1(0), \dots, y_l(0)$  является функцией от параметров  $g_m$ ; вектор производных

$$\frac{\partial y(x)}{\partial g_1} = \left( \frac{\partial y_1(x)}{\partial g_1}, \dots, \frac{\partial y_l(x)}{\partial g_1} \right)^T$$

удовлетворяет системе дифференциальных уравнений

$$\frac{d}{dx} \left( \frac{\partial y(x)}{\partial g_1} \right) = \mathcal{F}_y(x, y) \frac{\partial y(x)}{\partial g_1}, \quad \mathcal{F}_y = \left[ \frac{\partial f_i}{\partial y_j} \right]. \quad (15)$$

После численного решения этой системы получаем вектор  $\partial y(X)/\partial g_1$ . Теперь можно найти производные

$$\frac{\partial \psi_i}{\partial g_1} = \sum_{j=1}^l \frac{\partial d_i}{\partial y_j} \frac{\partial y_j(X)}{\partial g_1} = \left( \text{grad } d_i, \frac{\partial y(X)}{\partial g_1} \right). \quad (16)$$

Этот путь может быть целесообразен, если, например, многократно решается краевая задача (12) при одной и той же правой части  $f(x, y)$ , но при различных граничных условиях.

В качестве стандартного метода определения производных  $\partial\psi_i/\partial g_j$  этот метод неудобен тем, что наряду с написанием программы вычисления правых частей уравнения он требует от пользователя также написания программы вычисления производных  $\partial f_i/\partial y_j$ .

2. Для большинства итерационных методов решения систем уравнений значения производных правых частей требуются лишь с умеренной точностью. Поэтому для нахождения этих производных можно просто воспользоваться какими-либо формулами численного дифференцирования, например простейшей:

$$\frac{\partial\psi_i(g_1, \dots, g_r)}{\partial g_1} \approx \frac{\psi_i(g_1 + \Delta, g_2, \dots, g_r) - \psi_i(g_1, \dots, g_r)}{\Delta}.$$

Вычисление производных  $\partial\psi_i/\partial g_1$  по этой формуле требует дополнительно (по сравнению с нахождением значений  $\psi_i(g_1, \dots, g_r)$ ) еще одного численного решения задачи Коши, соответствующей параметрам  $g_1 + \Delta, g_2, \dots, g_r$ . Поскольку при численном нахождении значений функций  $\psi_i$  может иметь место большая погрешность, которая еще более возрастет при делении на  $\Delta$ , здесь следует обратить внимание на разумный выбор  $\Delta$ . Для описываемого метода типично одновременное нахождение производных всех функций  $\psi_i$  по фиксированной переменной  $g_j$ ; поэтому здесь может оказаться разумным применение итерационных процессов с поочередным уточнением параметров  $g_j$ . Можно представить себе такой итерационный процесс: параметры  $g_j$  уточняются в циклическом порядке; уточненное значение параметра  $\bar{g}_j$  выбирается из условия минимума некоторой функции

$$\Phi_j \left( \psi_1(g_1, \dots, g_r) + \frac{\partial\psi_1(g_1, \dots, g_r)}{\partial g_j} (\bar{g}_j - g_j), \dots, \psi_r(g_1, \dots, g_r) + \frac{\partial\psi_r(g_1, \dots, g_r)}{\partial g_j} (\bar{g}_j - g_j) \right)$$

(см. подробнее § 7.3).

Так же как при решении линейных краевых задач, возникает вопрос о применимости метода в условиях реальных округлений. Если решения  $\eta$  системы уравнений в вариациях  $\eta' = = f_{\eta}$  сильно растут с ростом  $x$ , то погрешности в значениях  $g_j$  и вычислительные погрешности при численном интегрировании приводят к большим погрешностям в значениях функции  $\psi$ ; это в конечном счете приведет к большой погрешности получаемого решения. Если такая погрешность окажется недопустимой, то следует ввести иную параметризацию задачи.

В настоящее время в вычислительной практике получили распространение методы, занимающие промежуточное положение между методами, соответствующими формулам (7.3), где в качестве неизвестных параметров высту-

пают значения решения во всех узлах сетки, и описанным выше методом, где в качестве неизвестных параметров выступают значения решения в одной точке. Задаются точками  $0 = x_0 < x_1 < \dots < x_m = X$  такими, что на отрезках  $[x_{i-1}, x_i]$ ,  $i = 1, \dots, m$ , решения системы уравнений в вариациях не очень сильно возрастают как при продвижении в направлении отрицательных  $x$ , так и при продвижении в направлении отрицательных  $x$ .

В качестве отыскиваемых параметров принимаются неизвестные компоненты решения в точках  $x_0$  и  $x_m$  и значения решения в точках  $x_1, \dots, x_{m-1}$ .

При практическом решении конкретных нелинейных краевых задач для обыкновенных дифференциальных уравнений, как и в случае других нелинейных задач, обычно приходится заниматься «доводкой» метода: предлагается какой-то специальный метод получения начального приближения, который затем модернизируется с целью расширения области начальных условий, при которых он сходится для данного конкретного класса задач. В ряде случаев метод решения строится путем имитации на ЭВМ методов, встречающихся в живой природе, или применяемых практиками для решения задач данного класса. Если рассматриваемая краевая задача является задачей на экстремум некоторого функционала, то исходный функционал приближается функционалом, зависящим от конечного числа параметров, и путем линеаризации последнего получают достаточно хорошее приближение. Примеры подобных функционалов для случая линейных задач будут рассмотрены в § 11.

## § 8. Аппроксимации специального типа

Рассмотрим краевую задачу

$$(k(x)y'(x))' - p(x)y(x) = f(x), \quad y(0) = a, \quad y(x) = b, \quad (1)$$

где  $p(x) \geq 0$ ,  $k(x) \geq k_0 > 0$ ,  $k(x)$  — трижды,  $p(x)$ ,  $f(x)$  — дважды непрерывно дифференцируемые функции, за исключением конечного числа точек, где эти функции или их производные  $k'$ ,  $k''$ ,  $k'''$ ,  $p'$ ,  $p''$ ,  $f'$ ,  $f''$  могут иметь разрывы первого рода.

Пусть  $K$ ,  $P$ ,  $F$  — множества точек разрыва соответственно функций  $k(x)$ ,  $p(x)$ ,  $f(x)$  или их производных,  $\Omega = K \cup P \cup F$ .

В случае уравнений с разрывными коэффициентами или решениями иногда не ясно, как понимать уравнение в точках разрыва. Для решения этого вопроса следует обратиться к интегральным соотношениям, обычно называемым законами сохранения, из которых были получены рассматриваемые дифференциальные уравнения. Уравнение (1) с разрывным  $k(x)$  обычно возникает из интегрального уравнения

$$\int_{x_1}^{x_2} (p(x)y(x) + f(x)) dx = k(x)y' \Big|_{x_1}^{x_2}, \quad (2)$$

которое выполнено для любых  $x_1, x_2 \in [0, X]$ . Если  $x_2 \rightarrow x_0 + 0$ ,  $x_1 \rightarrow x_0 - 0$ , то левая часть стремится к нулю; переходя к пределу в правой части, получаем, что

$$k(x)y' \Big|_{x_0-0}^{x_0+0} = 0 \quad \text{для любых } x_0 \in [0, X].$$

Исходя из этого соотношения решением задачи (1) будем называть функцию  $y(x)$ , удовлетворяющую следующим условиям:

- 1)  $y(x)$  непрерывна на  $[0, X]$ ;  
 2)  $y(x)$  удовлетворяет уравнению всюду на  $[0, X]$ , за возможным исключением точек множества  $\Omega$ ;  
 3) функция  $w(x) = k(x)y'(x)$ , называемая потоком, непрерывна на  $[0, X]$ .

Из условия 3) следует, что функция  $y'(x) = w(x)/k(x)$  непрерывна всюду, за исключением точек разрыва функции  $k(x)$ ; в этих точках  $y'(x)$  будет иметь разрывы первого рода. Для определенности возьмем далее равномерную сетку и при  $n$  целых и полуцелых будем употреблять обозначение  $x_n = nh$ ; далее  $X = Nh$ .

Можно показать, что производные  $y'''(x)$  и  $y^{(4)}(x)$  непрерывны и равномерно ограничены на множестве  $[0, X] \setminus \Omega$ .

Если при построении разностной схемы не учитывается факт разрывности  $y'(x)$ , то может случиться, что решение разностной задачи не сходится к решению задачи (1). Например, такая ситуация возникнет, если раскрыть скобки в первом слагаемом

$$(k(x)y'(x))' = ky''(x) + k'(x)y'(x)$$

и аппроксимировать его выражением

$$k(x_n) \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} + k'(x_n) \frac{y_{n+1} - y_{n-1}}{2h}.$$

Дело состоит в том, что хотя в области гладкости коэффициентов погрешности аппроксимации порядка  $O(h^2)$ , но условие 3) не учтено при построении схемы. Рассмотрим другую аппроксимацию:

$$\begin{aligned} (ky')|_{x_n} &\approx \frac{(ky')|_{x_{n+1/2}} - (ky')|_{x_{n-1/2}}}{h} \approx \\ &\approx \frac{k(x_{n+1/2}) \frac{y_{n+1} - y_n}{h} - k(x_{n-1/2}) \frac{y_n - y_{n-1}}{h}}{h}, \quad (3) \\ (py)|_{x_n} &\approx p(x_n)y_n, \quad \hat{f}(x_n) \approx \hat{f}_n. \end{aligned}$$

Можно показать, что решение соответствующей разностной задачи

$$\frac{k(x_{n+1/2}) \frac{y_{n+1} - y_n}{h} - k(x_{n-1/2}) \frac{y_n - y_{n-1}}{h}}{h} - p(x_n)y_n = \hat{f}_n,$$

$$n = 1, \dots, N-1, \quad y_0 = a, \quad y_N = b,$$

сходится к решению дифференциальной со скоростью  $O(h)$ . В то же время оказывается, что в случае, когда один из интервалов  $(x_{n-1}, x_n)$ ,  $(x_n, x_{n+1})$  содержит точку разрыва функции  $k(x)$ , погрешность аппроксимации в точке  $x_n$  имеет порядок  $h^{-1}$ .

Проверим последнее утверждение и объясним, почему оно не противоречит факту сходимости со скоростью  $O(h)$ . Погрешность аппроксимации запи-

связывается в виде

$$r_n = \frac{v(x_{n+1/2}) - v(x_{n-1/2})}{h} - p(x_n) y(x_n) - f(x_n), \quad (4)$$

где

$$v(x_{n+1/2}) = k(x_{n+1/2}) \frac{y(x_{n+1}) - y(x_n)}{h}.$$

Положим  $v(x_{n+1/2}) - k(x_{n+1/2}) y'(x_{n+1}) = \beta_n$ ; поскольку  $k(x)$  и производная  $y'(x)$  равномерно ограничены на  $[0, X]$ , то величина  $\left| \frac{y(x_{n+1}) - y(x_n)}{h} \right|$  равномерно ограничена по  $n$  и  $h$  и  $\sup_{n, h} |\beta_n| = O(1)$ .

Если производная  $y'''(x)$  ограничена на  $[x_n, x_{n+1}]$ , то

$$|\beta_n| \leq \frac{1}{24} (\sup |k(x)|) \cdot \left( \sup_{[x_n, x_{n+1}]} |y'''| \right) h^2 = O(h^2).$$

Соотношение (4) записывается в виде

$$r_n = \frac{(ky')|_{x_{n+1/2}} - (ky')|_{x_{n-1/2}}}{h} - p(x_n) y(x_n) - f(x_n) + \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h}. \quad (5)$$

Положим

$$\frac{(ky')|_{x_{n+1/2}} - (ky')|_{x_{n-1/2}}}{h} - (ky')'|_{x_n} = \alpha_n;$$

если величина  $(ky')'''$  ограничена, то  $\alpha_n = O(h^2)$ . Окончательное выражение погрешности аппроксимации имеет вид

$$r_n = \alpha_n + \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h} + ((ky')' - py - f)|_{x_n} = \alpha_n + \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h}.$$

Рассмотрим случай, когда на  $(x_n, x_{n+1})$  имеется точка разрыва коэффициента  $k(x)$ ; тогда для  $\beta_{n+1/2}$  не удастся получить оценки лучшей чем  $O(1)$ , а для  $r_n$  — лучшей чем  $O(1/h)$ . Тем не менее погрешность решения имеет порядок  $O(h)$ . Наглядно это можно объяснить следующим образом. Точек, где  $r_n$  имеет порядок  $O(1/h)$ , конечное число; их доля в общем числе узлов порядка  $O(h)$ , поэтому суммарный вклад от погрешности аппроксимации в таких точках будет  $O(h) \cdot O(1/h) = O(1)$ .

Значение  $\beta_{n+1/2}$  входит в выражение погрешности аппроксимации следующим образом:

$$r_n = \alpha_n + \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h},$$

$$r_{n+1} = \alpha_{n+1} + \frac{\beta_{n+3/2} - \beta_{n+1/2}}{h}.$$

Поэтому величина  $\beta_{n+1/2}$  дает вклад в погрешность аппроксимации в точке  $x_n$ , равный  $\beta_{n+1/2}/h$ , а в точке  $x_{n+1}$  — равный  $-\beta_{n+1/2}/h$ . Погрешность решения записывается в виде (см. § 2)

$$y_q - y(x_q) = h \sum G_q^n r_n.$$

Можно показать, что сеточная функция Грина  $G_q^n$  обладает ограниченным разностным отношением  $\left| \frac{G_q^{n+1} - G_q^n}{h} \right| \leq \text{const}$ . Вследствие этого вклад

от величины  $\beta_{n+1/2}$  есть

$$h(G_q^n - G_q^{n+1}) \frac{\beta_{n+1/2}}{h} = O(h) | \beta_{n+1/2} | = O(h).$$

Приведенные выше соображения подтверждают широко распространенное эмпирическое правило:

*при построении разностных схем не следует зря раскрывать скобок и пользоваться формулой дифференцирования произведения.*

Построим более точную разностную схему исходя из закона сохранения (2). Имеем равенство, которое следует из (2) в результате интегрирования в пределах от  $x_{n-1/2}$  до  $x_{n+1/2}$ :

$$\omega(x_{n+1/2}) - \omega(x_{n-1/2}) = \int_{x_{n-1/2}}^{x_{n+1/2}} [p(x)y(x) + f(x)] dx; \quad (6)$$

здесь и далее  $\omega(x) = k(x) \frac{dy}{dx}$ . Так как функция  $y(x)$  кусочно-дифференцируема, то

$$y(x) = y(x_n) + O(h) \quad \text{при} \quad x - x_n = O(h).$$

Поэтому (6) можно переписать в виде

$$\begin{aligned} \omega(x_{n+1/2}) - \omega(x_{n-1/2}) &= \int_{x_{n-1/2}}^{x_{n+1/2}} (p(x)y(x_n) + f(x) + O(h)) dx = \\ &= h(\bar{p}_n y(x_n) + \bar{f}_n) + O(h^2); \end{aligned}$$

здесь

$$\bar{p}_n = \frac{1}{h} \int_{x_{n-1/2}}^{x_{n+1/2}} p(x) dx, \quad \bar{f}_n = \frac{1}{h} \int_{x_{n-1/2}}^{x_{n+1/2}} f(x) dx.$$

После деления на  $h$  получится соотношение

$$\frac{\omega(x_{n+1/2}) - \omega(x_{n-1/2})}{h} - \bar{p}_n y(x_n) = \bar{f}_n + O(h).$$

В случае, если интервал  $(x_{n-1}, x_n)$  содержит точки разрыва  $y'(x)$ , погрешность от непосредственной замены  $\frac{\omega(x_{n-1/2})}{h}$  на  $k(x_{n-1/2}) \frac{y(x_n) - y(x_{n-1}))}{h^2}$  может оказаться величиной порядка  $h^{-1}$ .

Для получения лучшей аппроксимации введем в рассмотрение вспомогательную независимую переменную  $t = \int_0^x \frac{dx}{k(x)}$ . Из



ограниченности производной по  $x$  и равенства  $\frac{du}{dt} = k(x) \frac{du}{dx}$  следует ограниченность производной по  $t$ . Функция  $\omega = k(x) \frac{dy}{dx} = \frac{dy}{dt}$  имеет ограниченную производную по  $x$ , а следовательно, и по  $t$ . Таким образом, вторая производная ограничена и можно написать равенство

$$k(x) \frac{dy}{dx} \Big|_{x_{n-1/2}} = \frac{dy}{dt} \Big|_{t_{n-1/2}} = \frac{y(t_n) - y(t_{n-1})}{t_n - t_{n-1}} + O(t_n - t_{n-1});$$

здесь

$$t_n = \int_0^{x_n} \frac{dx}{k(x)}; \quad t_n - t_{n-1} = \int_{x_{n-1}}^{x_n} \frac{dx}{k(x)} = O(h).$$

Поэтому

$$\omega(x_{n-1/2}) = k(x) \frac{dy}{dx} \Big|_{x_{n-1/2}} = \frac{y(x_n) - y(x_{n-1})}{t_n - t_{n-1}} + O(h).$$

После подстановки  $k(x) \frac{dy}{dx} \Big|_{x_{n \pm 1/2}}$  в левую часть получим

$$r_n = \frac{\frac{y(x_{n+1}) - y(x_n)}{h\bar{k}_{n+1/2}^{-1}} - \frac{y(x_n) - y(x_{n-1})}{h\bar{k}_{n-1/2}^{-1}}}{h} - \bar{p}_n y(x_n) - \bar{f}_n = O(1),$$

где

$$\bar{k}_{j+1/2}^{-1} = \frac{1}{h} \int_{x_j}^{x_{j+1}} \frac{dx}{k(x)}. \quad (7)$$

Соответствующая конечно-разностная схема имеет вид

$$L(y_n) = \frac{\frac{y_{n+1} - y_n}{h\bar{k}_{n+1/2}^{-1}} - \frac{y_n - y_{n-1}}{h\bar{k}_{n-1/2}^{-1}}}{h} - \bar{p}_n y_n = \bar{f}_n. \quad (8)$$

Максимум погрешности аппроксимации у полученной схемы есть  $O(1)$ , поэтому для получения оценки погрешности привлекается ряд дополнительных соображений.

Если отрезок  $[x_{n-1}, x_{n+1}]$  не содержит точек  $\Omega$ , то непосредственной проверкой с помощью разложения в ряд Тейлора устанавливается, что погрешность аппроксимации  $r_n$  есть  $O(h^2)$ . В противном случае погрешность аппроксимации представляется в виде

$$r_n = L_h(y(x_n)) - \bar{f}_n = \frac{\frac{y(x_{n+1}) - y(x_n)}{h\bar{k}_{n+1/2}^{-1}} - \frac{y(x_n) - y(x_{n-1})}{h\bar{k}_{n-1/2}^{-1}}}{h} - \bar{p}_n y(x_n) - \bar{f}_n = \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h} + \alpha_n. \quad (9)$$

где

$$\alpha_n = \frac{w(x_{n+1/2}) - w(x_{n-1/2})}{h} - \bar{p}_n y(x_n) - \bar{f}_n.$$

$$\beta_{n+1/2} = \frac{y(x_{n+1}) - y(x_n)}{h \bar{k}_{n+1/2}} - w(x_{n+1/2}).$$

Если  $(x_n, x_{n+1})$  не содержит точек  $\Omega$ , то разложением в ряд Тейлора устанавливаем, что  $\beta_{n+1/2} = O(h^2)$ ; если  $(x_{n-1}, x_{n+1})$  не содержит точек  $\Omega$ , то так же устанавливаем, что  $\alpha_n = O(h^2)$ ; в противоположных случаях удается получить лишь оценки

$$\beta_{n+1/2} = O(h), \quad \alpha_n = O(h).$$

Далее, следуя намеченному выше способу оценки с помощью аппарата функции Грина, можно получить оценку

$$\max_n |y_n - y(x_n)| = O(h^2).$$

Ниже будет получена другая оценка погрешности.

Из равенств (8) и (9) следует, что

$$r_n = L(y(x_n) - y_n) = L(R_n); \quad (10)$$

здесь  $R_n$  — погрешность приближенного решения. Пусть  $\varphi_n$  — сеточная функция, удовлетворяющая, как и  $R_n$ , условию

$$\varphi_0 = \varphi_N = 0.$$

Умножим (10) на  $h\varphi_n$  и просуммируем в пределах от 1 до  $N-1$ :

$$h \sum_1^{N-1} r_n \varphi_n = h \sum_1^{N-1} L(R_n) \varphi_n. \quad (11)$$

Воспользовавшись выражением (9) для  $r_n$ , перепишем это равенство в виде

$$S_1 + S_2 = S_3 + S_4; \quad (12)$$

$$S_1(\varphi_n) = h \sum_{n=1}^{N-1} \alpha_n \varphi_n, \quad S_2(\varphi_n) = h \sum_{n=1}^{N-1} \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h} \varphi_n,$$

$$S_3(\varphi_n) = -h \sum_{n=1}^{N-1} \bar{p}_n R_n \varphi_n, \quad S_4(\varphi_n) = h \sum_{n=1}^{N-1} \frac{g_{n+1/2} - g_{n-1/2}}{h} \varphi_n,$$

где

$$g_{n+1/2} = \bar{k}_{n+1/2} \frac{R_{n+1} - R_n}{h}.$$

Собрав в выражении  $S_2$  подобные члены при одинаковых слагаемых  $\beta_{n+1/2}$ , получим

$$S_2(\varphi_n) = -h \sum_{n=0}^{N-1} \beta_{n+1/2} \frac{\varphi_{n+1} - \varphi_n}{h}; \quad (13)$$

заметим, что в правой части дописаны нулевые в силу условия  $\varphi_0 = \varphi_N = 0$  слагаемые  $\beta_{N-1/2} \varphi_N$  и  $-\beta_{1/2} \varphi_0$ .

То же самое выражение для  $S_2$  можно было бы получить, применив разностную формулу Абеля суммирования по частям:

$$\sum_{n=0}^{N-1} (a_{n+1} - a_n) b_n = - \sum_{n=1}^N (b_{n+1} - b_n) a_{n+1} + a_N b_N - a_0 b_0. \quad (14)$$

Точно так же получим

$$S_4(\varphi_n) = -h \sum_{n=0}^{N-1} g_{n+1/2} \frac{\varphi_{n+1} - \varphi_n}{h}. \quad (15)$$

Подставим в (12)  $\varphi_n = R_n$ ; имеем

$$S_3(R_n) = -h \sum_{n=1}^{N-1} \bar{p}_n R_n^2 \leq 0,$$

$$S_4(R_n) = -h \sum_{n=0}^{N-1} \bar{k}_{n+1/2} \left( \frac{R_{n+1} - R_n}{h} \right)^2 \leq 0.$$

Поэтому из (12) получаем

$$|S_4(R_n)| \leq |S_3(R_n) + S_4(R_n)| \leq |S_1(R_n)| + |S_2(R_n)|. \quad (16)$$

Из (7) следует, что  $\bar{k}_{n+1/2} \geq k_0$ , поэтому

$$|S_4(R_n)| \geq k_0 S_0(R_n), \quad S_0(R_n) = h \sum_{n=0}^{N-1} \frac{(R_{n+1} - R_n)^2}{h}.$$

При  $R_0 = R_N = 0$  выражение  $(S_0(R_n))^{1/2}$  обозначают как  $\|R_n\|_{\tilde{W}_{2,h}^1}$ .

Видно, что оно является сеточным аналогом нормы в пространстве  $\tilde{W}_2^1$  функций, удовлетворяющих условиям  $\varphi(0) = \varphi(X) = 0$ ,

$$\|\varphi\|_{\tilde{W}_2^1} = \left( \int_0^X \left( \frac{d\varphi}{dx} \right)^2 dx \right)^{1/2}. \quad \text{Нормы } \|\varphi_n\|_{L_{2,h}} = \left( h \sum_{n=1}^{N-1} |\varphi_n|^2 \right)^{1/2} \text{ и } \|\varphi_n\|_{C_h} =$$

$$= \max_{0 < n < N} |\varphi_n| \text{ являются сеточными аналогами норм в пространствах } L_2 \text{ и } C.$$

Теорема (сеточная теорема вложения).

$$\|\varphi_n\|_{L_{2,h}} \leq \sqrt{X} \|\varphi_n\|_{C_h}, \quad (17)$$

$$\|\varphi_n\|_{C_h} \leq \frac{\sqrt{X}}{\sqrt{2}} \|\varphi_n\|_{\tilde{W}_{2,h}^1}. \quad (18)$$

Справедливость первого утверждения непосредственно следует из определения норм и цепочки неравенств

$$\|\varphi_n\|_{L_{2,h}} \leq \sqrt{h(N-1)} \|\varphi_n\|_{C_h}^2 \leq \sqrt{X} \|\varphi_n\|_{C_h}.$$

Пусть  $n_0$  — точка, где достигается наибольшее значение  $|\varphi_n|$ . Рассмотрим случай  $n_0 \leq N/2$ ; случай  $n_0 > N/2$  сводится к рассматриваемому введением

нового индекса  $n = N - n$ . Имеем равенство

$$\varphi_{n_0} = \sum_{n=0}^{n_0-1} (\varphi_{n+1} - \varphi_n) = \sum_{n=0}^{n_0-1} \sqrt{h} \frac{\varphi_{n+1} - \varphi_n}{\sqrt{h}}.$$

Воспользовавшись неравенством для скалярного произведения

$$\left| \sum_{q=1}^l a_q b_q \right| \leq \sqrt{\sum_{q=1}^l |a_q|^2} \cdot \sqrt{\sum_{q=1}^l |b_q|^2},$$

получим

$$|\varphi_{n_0}| \leq \sqrt{n_0 h} \cdot \sqrt{h \sum_{n=0}^{n_0-1} \left( \frac{\varphi_{n+1} - \varphi_n}{h} \right)^2} \leq \sqrt{\frac{X}{2}} \|\varphi\|_{W_{2,h}^1}.$$

Теорема доказана.

Из (17), (18) следует, что

$$\|\varphi\|_{L_{2,h}} \leq \frac{X}{\sqrt{2}} \|\varphi\|_{W_{2,h}^1}. \quad (19)$$

Воспользовавшись (13), получаем оценку

$$\begin{aligned} |S_2(R_n)| &\leq \sqrt{h \sum_{n=0}^{N-1} (\beta_{n+1/2})^2} \cdot \sqrt{h \sum_{n=0}^{N-1} \left( \frac{R_{n+1} - R_n}{h} \right)^2} = \\ &= \sqrt{h \sum_{n=0}^{N-1} (\beta_{n+1/2})^2} \cdot \|R_n\|_{W_{2,h}^1}. \end{aligned}$$

Точно так же с учетом (15) имеем

$$|S_1(R_n)| \leq \|\alpha_n\|_{L_{2,h}} \|R_n\|_{L_{2,h}} \leq \frac{\sqrt{X}}{2} \|\alpha_n\|_{L_{2,h}} \|R_n\|_{W_{2,h}^1}.$$

Таким образом, из (16) следует

$$k_0 \|R_n\|_{W_{2,h}^1}^2 \leq \left( \frac{\sqrt{X}}{2} \|\alpha_n\|_{L_{2,h}} + \sqrt{h \sum_{n=0}^{N-1} (\beta_{n+1/2})^2} \right) \|R_n\|_{W_{2,h}^1},$$

поэтому

$$\|R_n\|_{W_{2,h}^1} \leq \frac{1}{k_0} \sqrt{\frac{X}{2}} \|\alpha_n\|_{L_{2,h}} + \frac{1}{k_0} \sqrt{h \sum_{n=0}^{N-1} (\beta_{n+1/2})^2}.$$

Величина  $\alpha_n$  порядка  $O(h^2)$  за возможным исключением конечного числа  $n$ , соответствующих отрезкам  $[x_{n-1}, x_{n+1}]$ , имеющим общие точки с  $\Omega$ ; для этих точек  $\alpha_n = O(h)$ . Отсюда следует оценка  $\|\alpha_n\|_{L_{2,h}} = O(h^{3/2})$ .

Точно так же выводится, что

$$\sqrt{h \sum_{n=0}^{N-1} (\beta_{n+1/2})^2} = O(h^{3/2}).$$

Таким образом,  $\|R_n\|_{W_{2,h}^1}$ , а следовательно, согласно теореме вложения и

$\|R_n\|_{C_h}$  есть  $O(h^{3/2})$ . Напомним, что на самом деле  $\|R_n\|_{C_h} = O(h^2)$ .

При использовании схемы (8) вычислительный процесс не зависит от положения точек разрыва. Поэтому ее относят к классу *однородных схем*.

Схема (8) на первый взгляд обладает следующим неудобством. Ее коэффициенты  $\bar{k}_{q+1/2}$ ,  $\bar{p}_q$ ,  $\bar{f}_q$  записываются как некоторые интегралы. На самом деле можно показать, что если погрешность в значениях этих коэффициентов есть  $O(h^2)$ , то погрешность приближенного решения оказывается также  $O(h^2)$ . Поэтому если интервал  $(x_{q-1}, x_{q+1})$  не содержит точек разрывов коэффициентов  $k(x)$ ,  $p(x)$ ,  $f(x)$ , то без потери порядка точности можно заменить  $\bar{p}_q$  на  $p(x_q)$ ,  $\bar{f}_q$  на  $f(x_q)$  и  $\bar{k}_{q\pm 1/2}$  на  $k(x_{q\pm 1/2})$ .

В ряде случаев построение разностных схем путем непосредственной аппроксимации производной разностным отношением приводит к недостаточно эффективным разностным схемам. Иногда бывает удобно в окрестности каждого расчетного узла приблизить рассматриваемое уравнение дифференциальным уравнением, интегрируемым в явном виде, и построить разностную схему, точную для его решений.

Рассмотрим дифференциальное уравнение

$$\mu^2 y''(x) + p(x)y(x) = f(x), \quad (20)$$

где  $\mu$  — малое число; для определенности сначала предполагаем  $p(x) > 0$ . В случае  $p = \text{const}$  решения этого уравнения  $\exp\left\{\pm i \frac{\sqrt{p} x}{\mu}\right\}$  колеблются с периодом  $2\pi\mu/\sqrt{p}$ , т. е. очень сильно. Характерный размер изменения решения имеет порядок  $\mu/\sqrt{p}$ , поэтому если не использовать специфику данного уравнения, то для получения высокой точности необходимо выполнение довольно обременительного условия  $h \ll \mu/\sqrt{p}$ .

В окрестности каждого узла  $x_n$  рассматриваемое уравнение близко к уравнению  $\mu^2 y'' + p_n y = f_n$ ,  $p_n = p(x_n)$ ,  $f_n = f(x_n)$ . Общее решение этого уравнения записывается в виде

$$y(x) = D_1 \exp\left\{i \frac{(x - x_n)\sqrt{p_n}}{\mu}\right\} + D_2 \exp\left\{-i \frac{\sqrt{p_n}(x - x_n)}{\mu}\right\} + \frac{f_n}{p_n}; \quad (21)$$

$D_1, D_2$  — произвольные константы. Найдем схему вида

$$a_n y_{n+1} + b_n y_n + c_n y_{n-1} - d_n = 0, \quad (22)$$

точную на всех решениях вида (21). Для этого подставим (21) в соотношение (22). Получим

$$\begin{aligned} a_n \left( D_1 \exp\left\{i \frac{\sqrt{p_n} h}{\mu}\right\} + D_2 \exp\left\{-i \frac{\sqrt{p_n} h}{\mu}\right\} \right) + b_n (D_1 + D_2) + \\ + c_n \left( D_1 \exp\left\{-i \frac{\sqrt{p_n} h}{\mu}\right\} + D_2 \exp\left\{i \frac{\sqrt{p_n} h}{\mu}\right\} \right) + \\ + (a_n + b_n + c_n) \frac{f_n}{p_n} - d_n = 0. \end{aligned}$$

Чтобы это равенство выполнялось при всех  $D_1$  и  $D_2$ , необходимо и достаточно равенства нулю коэффициентов при  $D_1$  и  $D_2$  и свободного члена. Приравняем их к нулю:

$$\begin{aligned} a_n \exp \left\{ i \frac{\sqrt{p_n} h}{\mu} \right\} + b_n + c_n \exp \left\{ -i \frac{\sqrt{p_n} h}{\mu} \right\} &= 0, \\ a_n \exp \left\{ -i \frac{\sqrt{p_n} h}{\mu} \right\} + b_n + c_n \exp \left\{ i \frac{\sqrt{p_n} h}{\mu} \right\} &= 0, \quad (23) \\ (a_n + b_n + c_n) \frac{f_n}{p_n} - d_n &= 0. \end{aligned}$$

Полагая  $a_n = 1$ , получим

$$c_n = 1, \quad b_n = -2 \cos \frac{\sqrt{p_n} h}{\mu}, \quad d_n = \left( 2 - 2 \cos \frac{\sqrt{p_n} h}{\mu} \right) \frac{f_n}{p_n}.$$

Общее решение системы (23) пропорционально полученному частному решению. Умножим все коэффициенты  $a_n, b_n, c_n, d_n$  на  $\mu^2 h^{-2}$ . Тогда получим схему

$$\mu^2 \frac{y_{n+1} - 2 \cos \frac{\sqrt{p_n} h}{\mu} y_n + y_{n-1}}{h^2} - \left( 2 - 2 \cos \frac{\sqrt{p_n} h}{\mu} \right) \frac{\mu^2 f_n}{h^2 p_n} = 0. \quad (24)$$

Это соотношение мы примем за разностное уравнение, соответствующее узлу  $x_n$ . Такая нормировка схемы (24) является наиболее естественной: если вместо  $y_n$  подставить в (24) значение  $y(x_n)$  и при фиксированном  $x_n$  устремить  $h$  к 0, то в пределе получится исходное дифференциальное уравнение

$$\mu^2 y'' - p(x_n) y - f(x_n) = 0.$$

Все приведенные выше построения имеют смысл независимо от знака  $p_n$ ; при  $p_n < 0$  в окончательной расчетной формуле имеем

$$\cos \frac{\sqrt{p_n} h}{\mu} = \cos i \frac{\sqrt{-p_n} h}{\mu} = \operatorname{ch} \frac{\sqrt{-p_n} h}{\mu}.$$

В случае  $p_n \equiv \text{const} > 0$  и  $f \equiv 0$  расчетная формула (24) переписывается в виде

$$y_{n+1} - 2 \cos \frac{\sqrt{p_n} h}{\mu} y_n + y_{n-1} = 0. \quad (25)$$

То, что эта формула является точной на решениях уравнения ((20) при  $p_n \equiv \text{const}$  и  $f \equiv 0$ , можно было бы усмотреть из известной формулы тригонометрии

$$\cos \varphi_1 + \cos \varphi_2 - 2 \cos \frac{\varphi_1 + \varphi_2}{2} \cos \frac{\varphi_1 - \varphi_2}{2} = 0,$$

подставив в нее

$$\varphi_1 = \frac{\sqrt{p} (n-1) h}{\mu} + \alpha, \quad \varphi_2 = \frac{\sqrt{p} (n+1) h}{\mu} + \alpha,$$

$\alpha$  — произвольное число.

Расчетная формула (25) иногда используется для быстрого вычисления таблицы значений  $\cos t$  или  $\sin t$  на равномерной сетке с невысокой точностью. Необходимость в этом может возникнуть, если, например, решается дифференциальное уравнение  $x' = f(x, t)$  с правой частью, содержащей значения  $\cos t$  или  $\sin t$ , и вычисление их значений составляет существенную долю от затрат на вычисление правой части. Заметим, что суммарная вычислительная погрешность при вычислении значений  $\cos t$  и  $\sin t$  по этим формулам порядка  $O(n^2\delta)$  ( $\delta = 2^{-t}$  — погрешность округлений).

Рассмотрим уравнение

$$y'(x) = y^2(x) + a^2(x), \quad (26)$$

решения которого могут обращаться в бесконечность. При аналитической функции  $a(z)$  решения аналитичны в комплексной плоскости в окрестности вещественной оси и иногда представляет интерес найти значения решения уравнения (26) в некоторой точке вещественной оси, отделенной от исходной несколькими полюсами. Один из возможных путей — это численное интегрирование (26) вдоль некоторой кривой, обходящей особые точки. Другой путь — это построение разностных схем, имеющих высокую точность на особенностях решения.

На отрезке  $[x_n, x_{n+1}]$  уравнение (26) приблизим уравнением

$$y'(x) = y^2(x) + a_{n+1/2}^2, \quad a_{n+1/2} = a((x_n + x_{n+1})/2).$$

Общее решение уравнения  $y'(x) = y^2(x) + a^2$  есть  $y(x) = \operatorname{tg}(a(x+c))$ . Воспользуемся формулой

$$\operatorname{tg}(\varphi + \psi) = \frac{\operatorname{tg} \varphi + \operatorname{tg} \psi}{1 - \operatorname{tg} \varphi \operatorname{tg} \psi}$$

при  $\varphi = a(x_n + c)$ ,  $\psi = a(x_{n+1} - x_n)$  и тем, что  $\frac{y(x_n)}{a} = \operatorname{tg} \varphi$ ,  $\frac{y(x_{n+1})}{a} = \operatorname{tg}(\varphi + \psi)$ . Получим равенство

$$\frac{1}{a} y(x_{n+1}) = \frac{\frac{y(x_n)}{a} + \operatorname{tg}(a(x_{n+1} - x_n))}{1 - \frac{y(x_n)}{a} \operatorname{tg}(a(x_{n+1} - x_n))}.$$

Отсюда получаем расчетную формулу для исходного уравнения

$$y_{n+1} = \frac{y_n + a_{n+1/2} \operatorname{tg}(a_{n+1/2}(x_{n+1} - x_n))}{1 - \frac{y_n}{a_{n+1/2}} \operatorname{tg}(a_{n+1/2}(x_{n+1} - x_n))}. \quad (27)$$

За счет некоторого понижения точности ее можно упростить, воспользовавшись приближенным равенством  $\operatorname{tg} \psi \approx \psi$ ; получим

$$y_{n+1} = \frac{y_n + a_{n+1/2}^2(x_{n+1} - x_n)}{1 - y_n(x_{n+1} - x_n)}. \quad (28)$$

Обе расчетные формулы (27) и (28) позволяют получить приближение к решению и после прохождения полюсов, если только случайно не оказалось, что расстояние от одного из узлов до ближайшего полюса много меньше, чем  $\min(x_{n+1} - x_n)^2$ . Этого всегда можно избежать, распорядившись выбором шагов вблизи полюса.

Как примеры расчетных формул подобного рода можно рассматривать рекуррентные формулы метода прогонки (4.9), (4.10).

## § 9. Конечно-разностные методы отыскания собственных значений

Рассмотрим простейшую задачу на собственные значения:

$$y''(x) - p(x)y(x) = \lambda \rho(x)y(x), \quad y(0) = 0, \quad y(X) = 0. \quad (1)$$

Зададимся шагом  $h = XN^{-1}$  и выпишем сеточную задачу

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p_n y_n = \lambda \rho_n y_n, \quad (2)$$

$$n = 1, \dots, N-1, \quad y_0 = y_N = 0; \quad p_n = p(x_n), \quad \rho_n = \rho(x_n).$$

Значения  $\lambda$ , при которых система уравнений (2) имеет ненулевое решение  $y_0, \dots, y_N$ , естественно назвать собственными значениями сеточной задачи. Пусть собственные значения задач (1), (2) упорядочены в порядке убывания, т. е.  $\lambda_1 \geq \lambda_2 \geq \dots$ ;  $\lambda_1^h \geq \lambda_2^h \geq \dots$ .

Рассмотрим модельный пример:  $p(x) \equiv 0$ ,  $\rho(x) \equiv 1$ ; тогда (1) приобретает вид

$$y''(x) = \lambda y(x), \quad y(0) = y(X) = 0.$$

Собственные функции этой задачи есть  $\omega^m(x) = \sin(\pi m x / X)$  и соответствующие собственные значения  $\lambda_m = -(\pi m / X)^2$ . В случае сеточной задачи (2), приобретающей вид

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - \lambda y_n = 0, \quad y_0 = y_N = 0,$$

рассуждаем следующим образом: общее решение разностного уравнения  $y_{n+1} - (2 + \lambda h^2)y_n + y_{n-1} = 0$  записывается в виде

$$y_n = C_1 \mu_1^n + C_2 \mu_2^n,$$

где  $\mu_1, \mu_2$  — корни характеристического уравнения

$$\mu^2 - (2 + \lambda h^2)\mu + 1 = 0. \quad (3)$$

По формуле Виета  $\mu_1 \mu_2 = 1$ , поэтому  $\mu_2 = \mu_1^{-1}$  и  $y_n = C_1 \mu_1^n + C_2 \mu_1^{-n}$ . Условия  $y_0 = y_N = 0$  дают систему уравнений

$$C_1 + C_2 = 0, \quad C_1 \mu_1^N + C_2 \mu_1^{-N} = 0;$$



она имеет ненулевое решение, если ее определитель равен 0, т. е.  $\mu_1^{-N} - \mu_1^N = 0$ . Отсюда

$$\mu_1 = \exp \{ \pi i m / N \}, \quad m = \dots, -1, 0, 1, \dots$$

Из (3) можно выразить значения  $\lambda^h$  через значения  $\mu_1$ :

$$\begin{aligned} \lambda_m^h &= \frac{\mu_1 + \mu_1^{-1} - 2}{h^2} = \frac{\exp \left\{ \frac{\pi i m}{N} \right\} + \exp \left\{ -\frac{\pi i m}{N} \right\} - 2}{h^2} = \\ &= \frac{N^2}{X^2} \left( 2 \cos \frac{\pi m}{N} - 2 \right) = -4 \frac{N^2}{X^2} \sin^2 \frac{\pi m}{2N}. \end{aligned}$$

Для определенности возьмем  $C_1 = (2i)^{-1}$ ; тогда соответствующие собственные функции имеют вид

$$W_n^m = \frac{1}{2i} \left( \exp \left\{ \frac{\pi i m n}{N} \right\} - \exp \left\{ -\frac{\pi i m n}{N} \right\} \right) = \sin \frac{\pi m n}{N}.$$

Собственные значения  $\lambda_1^h, \dots, \lambda_{N-1}^h$  различны между собой, поэтому соответствующие им собственные функции  $\sin \frac{\pi n}{N}, \dots, \sin \frac{\pi(N-1)n}{N}$  также различны. Так как задача (2) является задачей на собственные значения для матрицы размерности  $N-1$ , то мы получили полную систему собственных функций; каждая из функций  $W_n^m$  при  $m \leq 0$  или при  $m \geq N$  равна тождественно нулю или пропорциональна одной из перечисленных выше функций  $W_n^1, \dots, W_n^{N-1}$ . Мы случайно выбрали такой пример, где в узлах сетки  $x_n = nh$  выполняется равенство

$$W_n^m = w^m(nh).$$

В общем случае это равенство не имеет места; однако характер близости собственных значений этих задач типичен и для общего случая. Поскольку, согласно формуле Тейлора,  $\cos x = 1 - \frac{x^2}{2} + \cos(\theta x) \frac{x^4}{24}$ , где  $|\theta| \leq 1$ , то из выражения для  $\lambda_m^h$  получаем

$$\lambda_m^h = -\frac{N^2}{X^2} \left( \frac{\pi^2 m^2}{N^2} - \frac{\cos \left( \theta_m \frac{\pi m}{N} \right)}{12} \left( \frac{\pi m}{N} \right)^4 \right), \quad |\theta_m| \leq 1,$$

или

$$\lambda_m^h = -\left( \frac{\pi m}{X} \right)^2 + \frac{\cos \theta'_m}{12} \left( \frac{\pi m}{X} \right)^4 h^2, \quad \theta'_m = \theta_m \frac{\pi m}{N}. \quad (4)$$

Из этой формулы видно, что  $\lambda_m^h - \lambda_m = O(h^2)$  при фиксированном  $m$ ; в то же время с ростом  $m$  как абсолютная, так и относительная погрешности монотонно возрастают и, например,

$$\frac{\lambda_{N-1}^h}{\lambda_{N-1}^h} = \frac{-\pi^2(N-1)^2}{-4N^2 \sin^2 \frac{\pi(N-1)}{2N}} \sim \frac{\pi^2}{4}.$$

Равенство (4) можно записать в виде оценки  $|\lambda_m^h - \lambda_m| \leq C \lambda_m^2 h^2$  где  $C$  не зависит от  $\lambda_m$  и  $h$ . В случае дважды дифференцируемых функций  $p(x)$ ,  $\rho(x)$  такую оценку можно обосновать.

Для решения задачи (1) с более высокой точностью можно воспользоваться любыми разностными аппроксимациями уравнения  $y''(x) - q(x)y(x) = 0$  более высокой точности. Рассмотрим пример.

В § 1 была построена разностная схема, аппроксимирующая последнее уравнение с погрешностью  $O(h^4)$ :

$$\frac{\delta^2 y_n}{h^2} - q_n y_n - \frac{1}{12} \delta^2 (q_n y_n) = 0 \quad (5)$$

(с несколько иными обозначениями). Уравнение (1) записывается в рассматриваемом виде при  $q(x) = p(x) + \lambda \rho(x)$ . Отсюда получаем сеточную задачу на собственные значения:

$$\begin{aligned} \frac{\delta^2 y_n}{h^2} - (p_n + \lambda \rho_n) y_n - \frac{1}{12} \delta^2 ((p_n + \lambda \rho_n) y_n) = 0, \\ n = 1, \dots, N-1, \quad y_0 = y_N = 0. \end{aligned} \quad (6)$$

Можно показать, что для собственных значений этой задачи выполняется оценка

$$|\lambda_m^h - \lambda_m| = O(\lambda_m^3 h^4).$$

Система уравнений (6) имеет вид  $Ay - \lambda By = 0$ , формально несколько более сложный, чем (2), поскольку матрица  $B$  не диагональная.

При повышении порядка точности могут возникать сеточные задачи, имеющие на первый взгляд еще более сложный вид, например такая: требуется найти  $\lambda$ , при котором система соотношений

$$\begin{aligned} \alpha_n(\lambda, h) y_{n-1} - \beta_n(\lambda, h) y_n + \gamma_n(\lambda, h) y_{n+1} = 0, \\ n = 1, \dots, N-1, \quad y_0 = y_N = 0, \end{aligned} \quad (7)$$

имеет ненулевое решение  $y_n$ .

Рассмотрим случай  $\gamma_n \neq 0$ . Фиксируем некоторое  $\omega_1 \neq 0$ , например  $\omega_1 = h$ ; зададимся произвольным  $\lambda$  и из соотношения

$$\alpha_n(\lambda, h) \omega_{n-1}^\lambda - \beta_n(\lambda, h) \omega_n^\lambda + \gamma_n(\lambda, h) \omega_{n+1}^\lambda = 0 \quad (8)$$

последовательно определим  $\omega_2^\lambda, \dots, \omega_N^\lambda$ . Если  $\omega_N^\lambda = 0$ , то это  $\lambda$  окажется собственным значением и  $\omega_n^\lambda$  — собственной функцией; если  $\omega_N^\lambda \neq 0$ , то это  $\lambda$  не является собственным значением задачи (7).

Для отыскания собственных значений задачи (7), совпадающих с нулями  $\omega_N^\lambda$ , можно применить какой-либо итерационный метод отыскания нулей функции и по ее значениям. Этот процесс облегчается следующим обстоятельством, имеющим место во многих случаях и позволяющим получить «вилку» для искомого корня: если функция  $\omega_n^\lambda$  имеет  $j$  перемен знака на  $(0, X)$ , то  $\lambda_j < \lambda < \lambda_{j+1}$ . Для вычисления значений  $\omega_N^\lambda$  при различных

$\lambda$ , как правило, наиболее рационально воспользоваться непосредственно рекуррентными формулами (8).

Предлагаемый алгоритм вычисления значений  $w_n^\lambda$  совпадает с алгоритмом решения сеточной задачи Коши для уравнения  $y''(x) = (p(x) + \lambda\rho(x))y(x)$ . Если решения дифференциальной задачи Коши сильно растут, то, как и ранее, могут возникать большие вычислительные погрешности.

Для отыскания собственных значений может применяться также и метод прогонки. Не повторяя идеи метода, ограничимся написанием расчетных формул. Положим  $w_n^\lambda/w_{n+1}^\lambda = C_n$ , тогда (8) переписывается в виде

$$\alpha_n C_{n-1} C_n - \beta_n C_n + \gamma_n = 0, \quad (9)$$

откуда

$$C_n = \gamma_n / (\beta_n - \alpha_n C_{n-1}). \quad (10)$$

Если  $w_n^\lambda$  и  $w_{n+1}^\lambda$  одного знака, то  $C_n > 0$ , если разного, то  $C_n < 0$ . Поэтому, наблюдая за переменами знака у  $C_n$ , можно определить число перемен знака у функции  $w_n^\lambda$ . Как мы видели в § 3, коэффициент  $C_n$  может оказываться очень большим, поэтому этот метод чаще применяется лишь для отыскания первого собственного значения.

## § 10. Оптимизация распределения узлов интегрирования

Решения дифференциальных уравнений и систем могут иметь различную гладкость на различных участках отрезка интегрирования. На примере оценки погрешности метода Рунге—Кутта было видно, что вклад от погрешности интегрирования на некотором шаге  $[x_n, x_{n+1}]$  в суммарную погрешность в точке  $x_N = x_0 + X$  равен произведению погрешности на шаге на множитель

$\exp \left\{ \int_{x_n}^{x_N} f_y(x, \tilde{y}(x)) dx \right\}$ , зависящий от  $n$ . Поэтому для некоторых

классов дифференциальных уравнений становится особо актуальной задача оптимизации распределения узлов интегрирования. Произведем анализ этой проблемы, не вникая в тонкости обоснования проводимых построений. Для простоты предполагаем, что начальное условие задано точно и округления отсутствуют. Погрешность результата численного интегрирования по методу Рунге—Кутта в точке  $x_N$  не превосходит

$$S_N = \sum_{n=1}^N |y_n - y_{n-1}(x_n)| \exp \left\{ \int_{x_n}^{x_N} f_y(x, \tilde{y}_n(x)) dx \right\}. \quad (1)$$

Предположим, что  $x_n = \varphi(n/N)$ ,  $\varphi(0) = x_0$ ,  $\varphi(1) = x_0 + X$ ,  $\varphi(t)$  — гладкая функция. Согласно формуле Лагранжа

$$x_n - x_{n-1} = \varphi \left( \frac{n}{N} \right) - \varphi \left( \frac{n-1}{N} \right) = \frac{1}{N} \varphi'(\bar{t}_n),$$

где  $(n-1)/N \leq \bar{t}_n \leq n/N$ , поэтому

$$H = \max_{0 < n \leq N} (x_n - x_{n-1}) \leq \frac{1}{N} \max_{|0, 1|} |\varphi'(t)|, \quad H \rightarrow 0 \quad \text{при} \quad N \rightarrow \infty;$$

следовательно (см. § 8.4), имеем

$$\max_{x_0 \leq x_n \leq x \leq x_0 + X} |y_n(x) - y(x)| \rightarrow 0$$

при  $N \rightarrow \infty$ . Таким образом, можно написать асимптотические равенства (см. также (8.3.12)),

$$\exp \left\{ \int_{x_n}^{x_N} f_y(x, \tilde{y}_n(x)) dx \right\} \sim \exp \left\{ \int_{x_n}^{x_N} f_y(x, y(x)) dx \right\},$$

$$\begin{aligned} y_n - y_{n-1}(x_n) &\sim \psi(x_n, y(x_n))(x_n - x_{n-1})^{k+1} \sim \\ &\sim \psi\left(\varphi\left(\frac{n}{N}\right), y\left(\varphi\left(\frac{n}{N}\right)\right)\right) \left(\frac{1}{N} \varphi'\left(\frac{n}{N}\right)\right)^{k+1}, \end{aligned}$$

используя которые выражение (1) для  $S_N$  представим в виде

$$S_N \sim \frac{1}{N^k} \sum_{n=1}^N \frac{1}{N} \Phi\left(\frac{n}{N}\right), \quad (2)$$

где

$$\Phi(t) = |\psi(\varphi(t), y(\varphi(t)))| (\varphi'(t))^{k+1} \exp \left\{ \int_{\varphi(t)}^{x_0+X} f_y(x, y(x)) dx \right\}. \quad (3)$$

При гладкой функции  $\Phi(t)$  и  $N \rightarrow \infty$  величина  $\sum_{n=1}^N \frac{1}{N} \Phi\left(\frac{n}{N}\right)$  стремится к интегралу

$$I = \int_0^1 \Phi(t) dt \quad (4)$$

и, таким образом,

$$S_N \sim N^{-k} \int_0^1 \Phi(t) dt.$$

Дальнейшие построения являются некоторым усложнением построений из § 3.12. Примем  $\varphi$  за новую переменную в интеграле (4). Тогда он запишется в виде

$$I = \int_{x_0}^{x_0+X} L(\varphi) d\varphi,$$

где

$$L(\varphi) = |\psi(\varphi, y(\varphi))| (\varphi'(\varphi))^{-k} \exp \left\{ \int_{\varphi}^{x_0+X} f_y(x, y(x)) dx \right\}. \quad (5)$$

Задачи минимизации этого интеграла в форме (4) за счет выбора  $\varphi(t)$  и в форме (5) за счет выбора обратной функции  $t(\varphi)$  эквивалентны. Вследствие равенства  $\frac{\partial L}{\partial t} = 0$  уравнение Эйлера

$$\frac{d}{d\varphi} \left( \frac{\partial L}{\partial t'} \right) - \frac{\partial L}{\partial t} = 0$$

в данном случае приобретает вид  $\frac{d}{d\varphi} \left( \frac{\partial L}{\partial t'} \right) = 0$ . Отсюда следует, что

$$\frac{\partial L}{\partial t'} = -k |\psi(\varphi, y(\varphi))| (t'(\varphi))^{-k-1} \exp \left\{ \int_{\varphi}^{x_0+X} f_y(x, y(x)) dx \right\} = \text{const.}$$

Возвращаясь к переменной  $\varphi$ , получаем

$$(\varphi'(t))^{k+1} |\psi(\varphi, y(\varphi))| \exp \left\{ \int_{\varphi}^{x_0+X} f_y(x, y(x)) dx \right\} = C_1. \quad (6)$$

Решение этого дифференциального уравнения зависит от  $C_1$  и еще некоторой постоянной  $C_2$ . Их значения следует определить из граничных условий  $\varphi(0) = x_0$ ,  $\varphi(1) = x_0 + X$ .

Отметим одно неочевидное обстоятельство. Уравнение (6) может быть записано в виде

$$(\varphi'(t))^{k+1} |\psi(\varphi, y(\varphi))| \exp \left\{ - \int_0^{\varphi} f_y(x, y(x)) dx \right\} = C, \quad (7)$$

куда не входят ни начальная, ни конечная точки интегрирования. При  $C = Ca^{k+1}$  уравнению (7) удовлетворяет также функция  $\varphi(at + b)$ . Задавшись некоторым  $N$ , осуществим одновременное интегрирование исходного уравнения и уравнения (7), выбирая каждый раз шаги интегрирования из условия  $x_n - x_{n-1} \sim N^{-1} \varphi'(t_{n-1})$ . Скорее всего, мы придем в конечную точку  $x_0 + X$  с числом узлов  $N_1$ , отличным от  $N$ , и тогда это распределение узлов не будет являться оптимальным распределением, соответствующим данному  $N$ . Это естественно, поскольку, начиная вычисления, мы не могли заранее предвидеть хода поведения решения и сказать, какую величину  $C$  следует взять в правой части (7). Однако можно показать, что, вследствие указанных свойств решения уравнения (7), функция  $\varphi(t)$  будет искомой и распределение узлов  $\varphi(n/N_1)$  — оптимальным (с точностью до погрешности численного интегрирования), соответствующим числу узлов  $N_1$ .

Непосредственное интегрирование уравнений (1) и (7) встречает затруднение из-за необходимости вычисления значений функции  $\psi(x, y(x))$ . Вместо непосредственного вычисления значений этой функции целесообразно использовать величину контрольного члена точности на шаге. При численном интегрировании уравнения (7) следует также иметь в виду, что эти соотношения носят асимптотический характер. В окрестности точек, где  $\psi(x, y(x)) = 0$ , в остаточном члене начинают играть существенную роль слагаемые порядка  $(x_n - x_{n-1})^{k+2}$ .

Дополнительное численное интегрирование уравнения (7) может сильно усложнить решение задачи. Поэтому к вопросу оптимизации распределения узлов часто подходят следующим образом. Пусть решаются задачи из некоторого определенного класса. Рассмотрим модельную для этого класса задачу, для которой можно в явном виде решить уравнение (7). Постараемся на ее примере установить зависимость шага (или меры погрешности на шаге) от поведения решения, при которой распределение узлов близко к оптимальному. Далее все задачи этого класса интегрируем с шагом, соответствующим этой зависимости.

В других случаях заранее задаются некоторой формой такой зависимости. Пусть система уравнений порядка  $m$  интегрируется с контролем точности на шаге. В случае системы уравнений контрольный член  $r$  будет некоторым вектором  $r_n = (r_n^1, \dots, r_n^m)$ . В ряде программ шаг интегрирования выбирается из условия

$$\|r_n\| / \max(M, \|y_n\|) \approx \varepsilon = \text{const}$$

или из условия

$$\max_{1 \leq k \leq m} \frac{|r_n^k|}{\max(|M_k|, |(y_n)_k|)} \approx \varepsilon = \text{const}.$$

Параметры  $M$ ,  $M_k$  подбирают из соображений оптимальности распределения узлов в условиях конкретной задачи.

Рассмотрим пример оптимизации распределения узлов интегрирования.

Пусть задача Коши  $y'(x) = My(x)$  при начальном условии  $y(0)$  решается методом Эйлера. Тогда

$$\psi(x, y(x)) = -\frac{1}{2} y''(x) = -\frac{M^2}{2} y(0) \exp\{Mx\}$$

и уравнение (6) имеет вид

$$(\varphi'_t) \frac{M^2}{2} |y(0)| \exp\{M\varphi\} \exp\{M(X - \varphi)\} = \text{const}.$$

Отсюда получаем  $\varphi'_t = \text{const}$ , т. е. распределение узлов следует взять равномерным.

Наибольшего эффекта решение задачи оптимизации распределения узлов или ее упрощенных вариантов достигает в случае решений с особенностями производных и при решении задач с малыми параметрами при старших производных, например задач типа пограничного слоя.

## § 11. Построение численных методов с помощью вариационных принципов

Часто бывает естественно и целесообразно строить численные методы, исходя из естественной постановки задачи как вариационной или пользуясь определением решения как некоторой функции, удовлетворяющей интегральному тождеству.

1. **Метод Ритца.** Рассмотрим краевую задачу из § 8:

$$Ly \equiv -(k(x)y'(x))' + p(x)y(x) = f(x), \quad y(0) = a, \quad y(X) = b, \quad (1)$$

$$k \geq k_0 > 0.$$

Ее решение является точкой экстремума функционала

$$I(y) = \int_0^X (k(y'(x))^2 + py^2(x) - 2f(x)y(x)) dx \quad (2)$$

на классе функций  $W_2^1[0, X]$ , удовлетворяющих условию  $y(0) = a$ ,  $y(X) = b$ . Напомним, что  $W_2^1[0, X]$  — это класс функций с ограниченным интегралом

$$I^0(y) = \int_0^X [(y'(x))^2 + y^2(x)] dx = \|y\|_{W_2^1[0, X]}^2.$$

Задаются некоторым  $n$  и выбирают совокупность функций  $\varphi_0^n(x), \dots, \varphi_n^n(x)$  с ограниченным интегралом  $I^0(\varphi_k^n)$ , удовлетворяющих условиям

$$\begin{aligned} \varphi_0^n(0) &= a, & \varphi_0^n(X) &= b, \\ \varphi_q^n(0) &= \varphi_q^n(X) = 0, & q &= 1, \dots, n. \end{aligned}$$

Приближенное решение ищется в виде

$$y^n = \varphi_0^n + \sum_{q=1}^n c_q \varphi_q^n.$$

Имеем

$$I(y^n) = \sum_{p, q=1}^n \Lambda(\varphi_p^n, \varphi_q^n) c_p c_q - 2 \sum_{q=1}^n b_q c_q + d_0;$$

здесь

$$\begin{aligned} d_0 &= I(\varphi_0^n), \\ \Lambda(\varphi_p^n, \varphi_q^n) &= \int_0^X (k(\varphi_p^n)'(\varphi_q^n)' + p\varphi_p^n\varphi_q^n) dx, \\ b_q &= \int_0^X (f\varphi_q^n - p\varphi_0^n\varphi_q^n - k(\varphi_0^n)'(\varphi_q^n)') dx. \end{aligned}$$

Находим экстремум функционала  $I(y^n)$  по переменным  $c_1, \dots, c_n$ , и соответствующую функцию  $y^n = \varphi_0^n + \sum_{q=1}^n c_q \varphi_q^n$  принимаем за приближенное решение задачи. При этом нахождение коэффициентов  $c_q$  сведется к решению системы линейных алгебраических уравнений

$$Ac = b, \quad (3)$$

где  $A$  — матрица с элементами  $a_{pq} = \Lambda(\varphi_p^n, \varphi_q^n)$ ,  $b$  — вектор с компонентами  $b_q$ .

Часто бывает удобнее сразу вычесть из решения функцию  $\varphi_0^n$ , удовлетворяющую граничным условиям, т. е. свести исходную задачу к задаче с однородными граничными условиями. В линейном случае (как (1)) обычно  $\varphi_0^n$  берут не зависящим от  $n$ . Часто бывает выполнено следующее условие. Краевая задача

$$Ly + \lambda y = 0, \quad y(0) = y(X) = 0$$

имеет только нулевое решение, если  $\lambda \geq 0$ . Тогда функционал  $I(y)$  ограничен снизу и искомое решение является не просто точкой экстремума, а точкой минимума функционала  $I(y)$ . В этом случае описанный выше метод построения приближенного решения называют *методом Рунца*. Существует ряд моментов, существенно влияющих на сходимость метода Рунца.

Чтобы приближенные решения  $y^n$  сходились к точному в норме  $W_2^1[0, X]$ , т. е. чтобы  $\|y^n - y\|_{W_2^1} \rightarrow 0$  при  $n \rightarrow \infty$ , необходимо и достаточно выполнения следующего условия: для любой функции  $g \in W_2^1$  и любого  $\varepsilon > 0$  существует линейная комбинация

$$g_n = \varphi_0^n + \sum_{q=1}^n c_q \varphi_q^n \quad \text{с} \quad \|g_n - g\|_{W_2^1} \leq \varepsilon.$$

Указанное условие обеспечивает сходимость метода Рунца в предположении, что все вычисления производятся точно. Пусть  $\lambda_n$  и  $\lambda^n$  — наименьшее и наибольшее по модулю собственные значения матрицы системы уравнений (3). Чтобы округления не повлияли на приближение  $y^n$ , существенно выполнение условия

$$|\lambda^n / \lambda_n| \leq M, \quad (4)$$

где  $M$  не зависит от  $n$ .

Довольно часто не удается построить системы функций, удовлетворяющие условию (4). Тогда удовлетворяются использованием систем функций, для которых

$$|\lambda^n / \lambda_n| = O(n^\alpha), \quad (5)$$

где  $\alpha$  — не очень большое число. В случаях (4), (5), как правило, удастся так организовать процесс решения системы (3), что суммарная вычислительная погрешность будет порядка  $O(n^{\beta\delta})$ .

В ряде случаев нетрудно построить системы функций, удовлетворяющие условию (4), но, как правило, для них матрица  $A$  оказывается полностью заполненной. Для задачи (1) такой системой является

$$\varphi_q^n(x) = \sin(\pi x / X), \quad q = 1, \dots, n.$$



В то же время для системы функций, соответствующих вариационно-разностному методу (см. далее),  $\alpha = \beta = 2$ , но зато матрица  $A$  трехдиагональная. Для системы функций  $\varphi_q^n(x) = x^q \times (1-x)$  величина  $|\lambda^n/\lambda_n|$  растет быстрее любой степени  $n$  и матрица заполненная. Если вместо системы функций  $\varphi_q^n(x) = x^q(1-x)$  взять систему

$$\varphi_q^n(x) = x(1-x)T_q(2x/X-1), \quad (6)$$

где  $T_q(x)$  — многочлены Чебышева, то при отсутствии округлений получится одно и то же приближение. В то же время система (6) удовлетворяет условию (5) и при практическом использовании накопление погрешности будет не очень большим.

**З а м е ч а н и е.** Может случиться, что для некоторого набора функций величина  $|\lambda^n/\lambda_n|$  растет с ростом  $n$  очень быстро, но для достижения нужной точности достаточно небольшого значения  $n$ ; тогда такой набор функций приемлем при решении данной задачи.

**2. Метод Бубнова — Галеркина.** Описываемый метод является обобщением метода Ритца и применим в случаях, когда исходная задача не является вариационной. Формально этот метод можно представить следующим образом. Запишем исходную задачу в виде задачи нахождения решения из некоторого интегрального соотношения, справедливого для любой функции  $\psi$  из соответствующего класса:

$$(Ly, \psi) = 0. \quad (7)$$

Под выражением в круглых скобках понимаем скалярное произведение в  $L_2[0, X]$ . Соотношение (7) в дальнейшем будем называть интегральным тождеством. Приближенное решение ищется в виде линейной комбинации

$$y^n = \varphi_0^n + \sum_{q=1}^n c_q \varphi_q^n;$$

задаются некоторой системой функций  $\psi_1^n, \dots, \psi_n^n$  и требуют выполнения интегральных соотношений

$$(Ly^n, \psi_q^n) = 0, \quad q = 1, \dots, n. \quad (8)$$

Так же как в случае метода Ритца, решение исходной задачи сводится к решению системы линейных уравнений (8) относительно неизвестных  $c_1^n, \dots, c_n^n$ ; в матричной форме система уравнений (8) записывается в виде  $Ac = d$ , где  $A = [a_{ij}]$  — матрица размерности  $n \times n$ ,  $c = (c_1^n, \dots, c_n^n)$ ,  $d$  — вектор правой части.

Оба описанных метода применимы в нелинейном случае. Если исходная задача является задачей на экстремум функционала, не являющегося, как (2), квадратичным, то система уравнений (3) относительно значений  $c_1^n, \dots, c_n^n$ , соответствующих точке экстремума  $I(y^n)$ , будет нелинейной.

Точно так же метод Бубнова — Галеркина в случае нелинейного уравнения  $L(y) = 0$  состоит в решении нелинейной системы

$$(L(y^n), \psi_q^n) = 0, \quad q = 1, \dots, n.$$

**3. Вариационно-разностный вариант метода Ритца.** Носителем  $N(f)$  функции  $f$  назовем множество точек, где  $f$  или  $f' \neq 0$ . Если носители функций  $\varphi_i^n$  и  $\varphi_j^n$  не пересекаются, то  $a_{ij} = \Lambda(\varphi_i^n, \varphi_j^n) = 0$ . Наличие большого числа нулевых элементов в матрице может привести к существенному уменьшению объема вычислений при решении системы (3). Это обстоятельство явилось стимулом к разработке *вариационно-разностных методов*, соединяющих в себе преимущества метода Ритца и конечно-разностных методов.

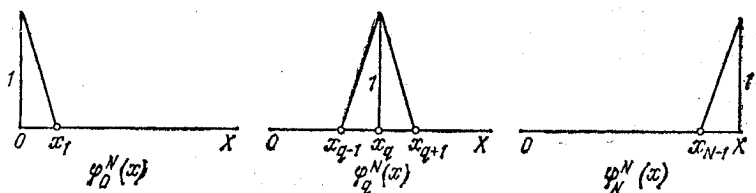


Рис. 9.11.1

Зададимся точками  $0 = x_0 < x_1 < \dots < x_N = X$  и будем отыскивать приближенное решение (1) в виде функции, линейной на каждом из отрезков  $[x_{q-1}, x_q]$  и принимающей заданные значения на концах отрезка  $[0, X]$ . Это равносильно тому, что решение ищется в виде

$$y_N(y) = \bar{\varphi}_0^N(x) + \sum_{q=1}^{N-1} y_q \varphi_q^N(x), \quad (9)$$

$$\bar{\varphi}_0^N(x) = a\varphi_0^N(x) + b\varphi_N^N(x),$$

$$\varphi_0^N(x) = \begin{cases} \frac{x_1 - x}{x_1} & \text{при } 0 \leq x \leq x_1, \\ 0 & \text{при } x_1 \leq x \leq x_N, \end{cases}$$

$$\varphi_N^N(x) = \begin{cases} 0 & \text{при } 0 \leq x \leq x_{N-1}, \\ \frac{x - x_{N-1}}{x_N - x_{N-1}} & \text{при } x_{N-1} \leq x \leq x_N. \end{cases} \quad (10)$$

$$\varphi_q^N(x) = \begin{cases} \frac{x - x_{q-1}}{x_q - x_{q-1}} & \text{при } x_{q-1} \leq x \leq x_q, \\ \frac{x_{q+1} - x}{x_{q+1} - x_q} & \text{при } x_q \leq x \leq x_{q+1}, \\ 0 & \text{в остальных точках} \end{cases}$$

при  $q = 1, \dots, N-1$  (см. рис. 9.11.1).

Система уравнений (3)

$$\partial I(y_N)/\partial y_q = 0, \quad q = 1, \dots, N-1, \quad (11)$$

из которой определяются значения  $y_1, \dots, y_{N-1}$ , в данном случае оказывается системой с трехдиагональной матрицей. Укажем конкретный вид коэффициентов системы, получающейся в случае функций  $\varphi_q^N$  вида (10):

$$a_{qq} = \int_{x_{q-1}}^{x_q} \left( \frac{k(x)}{(x_q - x_{q-1})^2} + p(x) \frac{(x - x_{q-1})^2}{(x_q - x_{q-1})^2} \right) dx +$$

$$+ \int_{x_q}^{x_{q+1}} \left( \frac{k(x)}{(x_{q+1} - x_q)^2} + p(x) \left( \frac{x_{q+1} - x}{x_{q+1} - x_q} \right)^2 \right) dx, \quad q = 1, \dots, N-1;$$

$$a_{q, q+1} = \int_{x_q}^{x_{q+1}} \left( \frac{-k(x)}{(x_{q+1} - x_q)^2} + p(x) \frac{x_{q+1} - x}{x_{q+1} - x_q} \frac{x - x_q}{x_{q+1} - x_q} \right) dx,$$

$$q = 1, \dots, N-2;$$

$$a_{q+1, q} = a_{q, q+1}, \quad q = 2, \dots, N-2,$$

$$b_q = \int_{x_{q-1}}^{x_q} f(x) \frac{x_q - x}{x_q - x_{q-1}} dx + \int_{x_q}^{x_{q+1}} f(x) \frac{x_{q+1} - x}{x_{q+1} - x_q} dx - \beta_q,$$

где

$$\beta_q = \begin{cases} a \int_0^{x_1} \left( p(x) \frac{x_1 - x}{x_1} \cdot \frac{x}{x_1} - \frac{k(x)}{x_1^2} \right) dx, & q = 1, \\ 0, & 1 < q < N-1, \\ b \int_{x_{N-1}}^{x_N} \left( p(x) \frac{x_N - x}{x_N - x_{N-1}} \cdot \frac{x - x_{N-1}}{x_N - x_{N-1}} - \frac{k(x)}{(x_N - x_{N-1})^2} \right) dx, & q = N-1. \end{cases}$$

**4. Вариационно-разностный вариант метода Бубнова — Галеркина.** Если умножим (1) на произвольную функцию  $\psi(x) \in \overset{\circ}{W}_2^1([0, X])$  и проинтегрируем первое слагаемое в выражении

$$\int_0^X (L(y) - f(x)) \psi(x) dx \quad (12)$$

по частям, то получим интегральное тождество

$$\Lambda(y, \psi) = \int_0^X (k(x) y' \psi' + p y \psi - f \psi) dx = 0. \quad (13)$$

Будем искать решение  $y_N$  в виде (9); потребуем, чтобы (13) обращалось в нуль для всех функций  $\psi(x)$  вида

$$\psi(x) = \sum_{q=1}^{N-1} \psi_q \varphi_q^N(x),$$

$\psi_q$  — произвольные числа. Поскольку выражение  $\Lambda(y, \psi)$  линейно по  $\psi$ , то получаем систему уравнений

$$\Lambda(y_N, \psi_q) = 0, \quad q = 1, \dots, N-1. \quad (14)$$

В данном случае эта система уравнений совпадает с (3).

Можно было не производить в (12) интегрирования по частям, а непосредственно потребовать удовлетворения (12) для любой функции вида (9). Однако для функций вида (9) выражение (12) содержит слагаемые типа  $\delta$ -функции, поэтому непосредственное выписывание уравнений системы (14) представило бы дополнительные технические трудности.

Описанный выше метод решения задачи, называемый *проекционно-разностным*, применим и к задачам более общего вида.

Рассмотрим краевую задачу

$$(k(x)y')' + ay' + (by)' + cy = f + F', \quad (15)$$

уже не являющуюся задачей на экстремум некоторого функционала. Умножим (15) скалярно на  $\psi \in \overset{\circ}{W}_2^1([0, X])$  и проинтегрируем некоторые из слагаемых по частям; получим

$$\Lambda_1(y, \psi) = \int_0^X (ky'\psi' - ay'\psi + by\psi' - cy\psi + f\psi - F\psi') dx = 0.$$

Приближение  $y_N$  вида (9) находим из системы уравнений

$$\Lambda_1(y_N, \psi_q) = 0, \quad q = 1, \dots, N-1. \quad (16)$$

Заметим, что интегрирование по частям интеграла от  $(by)'\psi$  является необходимым лишь в случае разрывного коэффициента  $b$ .

Описанные выше методы формально представляют некоторые неудобства, поскольку для построения системы уравнений (16) требуется вычисление некоторых интегралов. Если коэффициенты гладкие, то эти интегралы можно вычислить с помощью квадратурных формул:

$$\begin{aligned} \int_{x_{q-1}}^{x_q} k(x) dx &\approx k\left(\frac{x_{q-1} + x_q}{2}\right)(x_q - x_{q-1}), \\ \int_{x_{q-1}}^{x_q} p(x)(x_q - x)(x - x_{q-1}) dx &\approx p\left(\frac{x_{q-1} + x_q}{2}\right) \int_{x_{q-1}}^{x_q} (x_q - x) \times \\ &\times (x - x_{q-1}) dx = p\left(\frac{x_{q-1} + x_q}{2}\right) \frac{(x_q - x_{q-1})^3}{6}. \end{aligned} \quad (17)$$

$$\int_{x_{q-1}}^{x_q} p(x)(x - x_{q-1})^2 dx \approx p\left(\frac{x_{q-1} + x_q}{2}\right) \int_{x_{q-1}}^{x_q} (x - x_{q-1})^2 dx =$$

$$= p\left(\frac{x_{q-1} + x_q}{2}\right) \cdot \frac{(x_q - x_{q-1})^3}{3},$$

$$\int_{x_{q-1}}^{x_q} f(x)(x - x_{q-1}) dx \approx f\left(\frac{x_q + x_{q-1}}{2}\right) \frac{(x_q - x_{q-1})^2}{2}.$$

Во всех случаях был применен способ вычисления интегралов, при котором подынтегральная функция разлагалась на множители и интеграл от наиболее резко меняющейся функции брался в явном виде. Можно показать, что полученная схема обеспечивает второй порядок точности.

Можно предложить способ построения вариационно-разностных схем более высокого порядка точности; рассмотренный выше способ построения схемы является частным случаем (при  $m = 1$ ) этого способа. Приближение  $y_N(x)$  ищется в виде многочлена  $P_q^m(x)$  степени  $m$  на каждом из отрезков  $[x_{q-1}, x_q]$ , причем значения многочленов, соответствующих отрезкам  $[x_{q-1}, x_q]$  и  $[x_q, x_{q+1}]$ , совпадают в точке  $x_q$ . Минимизируя функционал  $I(y_N)$  на множестве таких многочленов, удовлетворяющих граничным условиям (1), получим систему с клеточно-треугольной матрицей относительно коэффициентов этих многочленов. Иногда удобнее рассматривать как неизвестные не эти коэффициенты, а некоторые другие параметры. Например, можно аналогично случаю приближения сплайнами (§ 4.11) при  $m = 3$  принять за неизвестные значения

$$y_N(x_q), \quad y_N''(x_{q+0}), \quad y_N''(x_{q-0}).$$

Можно на каждом из отрезков  $[x_{q-1}, x_q]$  взять  $m - 1$  дополнительную точку и принять за неизвестные значения  $y_N(x)$  в этих точках.

Вариационно-разностные методы в определенном смысле «технологичнее» разностных. В случае построения вариационно-разностных методов путем минимизации квадратичного функционала возникает система уравнений с положительно определенной матрицей, что обеспечивает определенную «физичность» получаемых приближений и одновременно облегчает решение системы. Переменность шага интегрирования также не оказывает существенного влияния на сложность программ вариационно-разностных методов. Эти преимущества наиболее эффективно проявляются при решении многомерных задач в областях со сложной геометрией. При одном и том же порядке точности использование вариационно-разностных схем часто требует меньшего объема программирования. Все это послужило причиной того, что они взяты за основу, например, при создании

пакетов численных методов решения задач теории упругости. В то же время при решении очень сложных задач, в которых для получения нужной точности требуется число узлов сетки, находящееся на пределе возможностей ЭВМ, часто бывает целесообразно обратиться к сеточным методам.

Заметим, что вариационно-разностные и проекционно-разностные методы называют также *методами конечных элементов*.

**5. Построение разностных схем путем аппроксимации функционала.** При непосредственном построении разностных аппроксимаций в областях сложного вида иногда оказывается, что получается система уравнений со знаконеопределенной матрицей, в то время как исходная задача была знакоопределена. Чтобы преодолеть этот дефект, не прибегая к использованию вариационно-разностных методов, строят конечно-разностные схемы, используя дискретную аппроксимацию минимизируемого функционала, соответствующего задаче. Приближим исходный функционал дискретной аппроксимацией:

$$I(y) \approx I_h(y_h) = \sum_{q=1}^N k(x_{q-1/2}) \left( \frac{y_q - y_{q-1}}{x_q - x_{q-1}} \right)^2 (x_q - x_{q-1}) + \\ + \sum_{q=1}^N \left( \frac{p(x_q)y_q^2 + p(x_{q-1})y_{q-1}^2}{2} \right) (x_q - x_{q-1}) - \\ - \sum_{q=1}^N (f(x_q)y_q + f(x_{q-1})y_{q-1})(x_q - x_{q-1}); \quad (18)$$

здесь  $x_{q-1/2} = (x_q + x_{q-1})/2$ .

Первая сумма получилась при использовании квадратурной формулы прямоугольников, вторая и третья — формулы трапеций. Приравнивая нулю производные  $\partial I_h / \partial y_q$ , получаем систему уравнений

$$2 \left[ k(x_{q-1/2}) \frac{y_q - y_{q-1}}{\Delta x_{q-1}} - k(x_{q+1/2}) \frac{y_{q+1} - y_q}{\Delta x_q} \right] + \\ + p(x_q)y_q \Delta x_{q-1} + p(x_q)y_q \Delta x_q - \\ - f(x_q)(\Delta x_{q-1} + \Delta x_q) = 0, \quad q = 1, \dots, N-1; \quad \Delta x_q = x_{q+1} - x_q.$$

Поделив предыдущее соотношение на  $-2\delta_q$ , где  $\delta_q = (\Delta x_q \pm \Delta x_{q-1})/2$ , получим конечно-разностную схему

$$L_q y = \frac{k(x_{q+1/2}) \frac{y_{q+1} - y_q}{\Delta x_q} - k(x_{q-1/2}) \frac{y_q - y_{q-1}}{\Delta x_{q-1}}}{\frac{\Delta x_q + \Delta x_{q-1}}{2}} - \\ - p(x_q)y_q - f(x_q) = 0, \quad (19)$$

совпадающую со схемой (8.3) в случае равномерной сетки.

Выражение  $I_h(y_h)$  в (18) является многочленом второй степени от переменных  $y_j$ ; оно записывается в виде

$$I_h(y_h) = \sum_{i,j=1}^{N-1} a_{ij} y_i y_j + \sum_{i=1}^{N-1} b_i y_i + c;$$

выше учтено то, что  $y_0 = a$ ,  $y_N = b$ . Из (18) видно, что первые две суммы в  $I_h(y_h)$  неотрицательны. Поэтому

$$I_h(y_h) \geq - \min |f| \left( \sum_{q=0}^N |y_q| \delta_q \right).$$

При таком поведении многочлена второй степени в окрестности бесконечности его главная часть  $\sum_{i,j=1}^{N-1} a_{ij} y_i y_j$  неотрицательна, т. е.  $A = [a_{ij}] \geq 0$ . Поскольку  $a_{ij} = \frac{1}{2} \partial^2 I_h(y_h) / \partial y_i \partial y_j$ , то матрица  $A$  автоматически оказывается симметричной.

**Задача 1.** Доказать, что при  $k > 0$ ,  $p \geq 0$  матрица  $A$  положительно определена:  $A > 0$ .

Чтобы проведенные рассуждения о неотрицательности матрицы  $A$  были справедливы, целесообразно строить аппроксимации функционала, приближая выражения, стоящие под знаком квадрата, не раскрывая скобок.

Пусть, например, исходная задача является задачей на экстремум функционала:

$$I(y) = \int_0^x (k(x)(y' + \lambda(x)y)^2 + p(x)y^2) dx; \quad p \geq 0.$$

Интеграл от первого слагаемого приближаем выражением

$$\sum_{q=1}^N k(x_{q-1/2}) \left( \frac{y_q - y_{q-1}}{\Delta x_q} + \lambda(x_{q-1/2}) \frac{y_q + y_{q-1}}{2} \right)^2 \Delta x_q,$$

от второго — так же, как в (18). Все проведенные выше рассуждения остаются в силе, и поэтому соответствующая матрица  $A$  неотрицательна. Если раскрыть скобки в  $(y' + \lambda y)^2$ , а потом аппроксимировать интеграл, то может случиться, что условие  $A \geq 0$  при крупных шагах будет нарушено.

Подведем итог последних рассуждений, который особенно существен в многомерном случае.

*При построении конечно-разностных методов путем аппроксимации минимизируемого функционала целесообразно записать функционал в виде суммы интегралов от квадратов некоторых выражений и линейной части и аппроксимировать эти выражения, не раскрывая скобок.*

6. Случай невариационной задачи. Для невариационных задач разностные схемы можно получать, аппроксимируя интегральное тождество, из которого определяется решение. Применим этот способ к рассматриваемой нами вариационной задаче. Будем аппроксимировать интегральное тождество

$$\Lambda(y, \psi) = \int_0^x (ky'\psi' + p\psi - f\psi) dx$$

для любой функции  $\psi \in \overset{\circ}{W}_{2,h}^1[0, X]$ . Имеем

$$\begin{aligned} \Lambda(y, \psi) \approx \Lambda_h(y_h, \psi_h) &= \sum_{q=1}^N k(x_{q-1}) \frac{y_q - y_{q-1}}{\Delta x_{q-1}} \frac{\psi_q - \psi_{q-1}}{\Delta x_{q-1}} \Delta x_{q-1} + \\ &+ \sum_{q=1}^N \frac{1}{2} ((p(x_q) y_q - f(x_q)) \psi_q + \\ &+ (p(x_{q-1}) y_{q-1} - f(x_{q-1})) \psi_{q-1}) \Delta x_{q-1} = 0. \end{aligned}$$

Полагая  $\psi_0 = \psi_N = 0$  и собирая коэффициенты при одних и тех же  $\psi_q$ , получим

$$\Lambda_h(y_h, \psi_h) = \sum_{q=1}^{N-1} \delta_q L_q \cdot \psi_q = 0,$$

где  $L_q$  определено формулой (19). Выражение  $\Lambda_h(y_h, \psi_h)$  равно нулю для любой сеточной функции  $\psi$ , если все  $L_q = 0$ . Полученная система уравнений совпадает с системой уравнений (19).

## § 12. Улучшение сходимости вариационных методов в нерегулярном случае

Погрешность вариационных методов существенно зависит от точности, с которой можно приблизить решение функциями из пространства, в котором ищется решение. Рассмотрим некоторые задачи, в которых при использовании вариационных методов имеет смысл несколько усложнить построение системы базисных функций.

1. В случае разрывного коэффициента  $k(x)$  производная  $u'$  также разрывна, поэтому решение плохо приближается кусочно-линейными функциями. В то же время выражение  $ku'$  является дифференцируемой функцией. Поэтому для достижения более высокой точности целесообразно искать приближение в виде функции, которая на каждом из отрезков  $[x_{q-1}, x_q]$  является решением уравнения  $ky' = \text{const}$  или, что то же самое,  $(ky')' = 0$ . В этом случае получается вариационно-разностная схема, имеющая в сеточной норме  $\overset{\circ}{W}_{2,h}^1$  порядок сходимости  $O(h^2)$  даже при измеримых функциях  $k(x)$ ,  $p(x)$ ,  $f(x)$ . Под нормой в  $\overset{\circ}{W}_{2,h}^1$  понимаем выражение

$$\|u_h\|_{\overset{\circ}{W}_{2,h}^1}^2 = \sum_{n=0}^{N-1} h \left( \frac{u_{n+1} - u_n}{h} \right)^2, \quad u_0 = u_N = 0,$$

являющееся сеточным аналогом нормы пространства  $\overset{\circ}{W}_2^1$ .



2. Иногда оказывается, что решение имеет особенность в конечном числе точек, а вдали от них является гладким. Например, относительно решения иногда можно установить, что оно имеет вид

$$y(x) = \sum_{j=1}^l C_j \psi_j(x) + u(x),$$

где  $\psi_j(x)$  — известные функции, а  $u(x)$  — неизвестная гладкая функция. Если некоторые из коэффициентов  $C_j$ , например  $C_{k+1}, \dots, C_l$ , известны, то следует перейти к новой неизвестной функции

$$y^* = y - \sum_{j=k+1}^l C_j \psi_j(x).$$

Далее рассматриваем случай — все  $C_j$  неизвестны. Приближение для гладкой части ищем в виде

$$u(x) = \sum_{q=0}^N c_q \varphi_q^N(x),$$

где  $\varphi_q^N(x)$  те же, что и в (11.9), т. е. решение ищется в виде

$$y(x) = \sum_{j=1}^l C_j \psi_j(x) + \sum_{q=0}^N c_q \varphi_q^N(x)$$

с неизвестными коэффициентами  $C_j, c_q$  и при дополнительном условии на эти коэффициенты, имеющем вид  $y(0) = a, y(X) = b$ . (Здесь для определенности мы приняли, что  $\psi_1, \dots, \psi_l, \varphi_0^N, \dots, \varphi_N^N$  линейно независимы.) Функции  $\psi_j(x)$  в отличие от функций  $\varphi_q^N$  обычно имеют носитель, размеры которого не стремятся к нулю при уменьшении шага сетки. Поэтому строки матрицы  $A$ , соответствующие функциям  $\psi_j$ , как правило, будут полностью заполнены. Матрица системы уже не оказывается трехдиагональной. Перестановкой строк и столбцов ее можно преобразовать к виду, где  $a_{ij} = 0$ , если одновременно  $|i-j| > 1, |i|, |j| > l+1$ . Если решать эту систему методом Гаусса при обратном порядке исключения неизвестных, то общее число арифметических операций оказывается, как и в случае трехдиагональных матриц, порядка  $O(N)$ . В описываемом случае надо особенно внимательно следить за погрешностью метода решения задачи (включающей в себя погрешность приближенного вычисления интегралов) и вычислительной погрешностью.

Рассмотрим в качестве примера краевую задачу

$$\begin{aligned} \varepsilon^2 y''(x) - p(x) y(x) &= f(x), & p(x) > 0, \\ y(0) &= a, & y(X) = b, & \varepsilon \text{ малое.} \end{aligned}$$

Формально говоря, решение не имеет особенности. Однако при малом  $\varepsilon$  имеется пограничный слой, где производные от решения велики и решение плохо приближается функциями вида (11.9). Из теории асимптотических методов известно, что в окрестности точки  $x = 0$  решение хорошо приближается линейными комбинациями функций вида  $x^k \exp\left\{-\frac{\sqrt{p(0)}}{\varepsilon} x\right\}$ , а в окрестности точки  $X$  — функций вида  $(X-x)^k \exp\left\{-\frac{\sqrt{p(X)}}{\varepsilon} (X-x)\right\}$ . Поэтому

приближенное решение имеет смысл отыскивать в виде

$$y(x) = \sum_{k=0}^l C_k x^k \exp \left\{ -\frac{\sqrt{\rho(0)}}{\varepsilon} x \right\} + \\ + \sum_{k=0}^l D_k (X-x)^k \exp \left\{ -\sqrt{\rho(X)} \frac{X-x}{\varepsilon} \right\} + \sum_{q=1}^N c_q \varphi_N^q(x);$$

здесь  $C_k, D_k, c_q$  — неизвестные коэффициенты.

### § 13. Влияние вычислительной погрешности в зависимости от формы записи конечно-разностного уравнения

Как было установлено ранее, вычислительная погрешность имеет различный характер роста для различных способов решения дифференциальных уравнений. Рассмотрим теперь такой частный, но важный вопрос: как зависит вычислительная погрешность от формы записи конечно-разностных уравнений? Хотя все изложение ведется на примере задачи Коши, проводимые соображения относятся в равной мере и к случаю решения краевых задач. Для примера обратимся к методу Эйлера:

$$y_{n+1} = y_n + hf(x_n, y_n). \quad (1)$$

При реальных вычислениях будут получаться величины  $y_n^*$ , связанные соотношением

$$y_{n+1}^* = y_n^* + hf(x_n, y_n^*) + \delta_n; \quad (2)$$

наличие слагаемого  $\delta_n$  является следствием ряда причин — погрешностей при вычислении значений функции  $f(x_n, y_n^*)$ , погрешностей при округлении произведения  $hf(x_n, y_n^*)$  и погрешностей при сложении чисел  $y_n^*$  и округленного значения  $hf(x_n, y_n^*)$ .

Введем обозначение  $y_n^* - y_n = \Delta_n$ ; на основании формулы Лагранжа имеем

$$f(x_n, y_n^*) - f(x_n, y_n) = l_n \Delta_n,$$

где  $l_n = f_y(x_n, \bar{y}_n)$ . Предположим, что всегда  $|f_y(x, y)| \leq L$ . Вычитая (1) из (2), получаем

$$\Delta_{n+1} = (1 + l_n h) \Delta_n + \delta_n,$$

откуда

$$|\Delta_{n+1}| \leq (1 + Lh) |\Delta_n| + \delta, \quad \delta = \max_n \delta_n. \quad (3)$$

Рассмотрим разностное уравнение

$$z_{n+1} = (1 + Lh) z_n + \delta,$$

являющееся *мажорирующим* для (3). Его решение при начальном условии  $z_0 = |\Delta_0|$  есть

$$z_n^0 = |\Delta_0|(1 + Lh)^n + \delta \frac{(1 + Lh)^n - 1}{Lh}.$$

*Лемма.* При всех  $n \geq 0$  справедливо

$$|\Delta_n| \leq z_n^0. \quad (4)$$

*Доказательство.* При  $n = 0$  утверждение (4) очевидно. Пусть оно верно для некоторого  $n$ ; тогда имеем

$$|\Delta_{n+1}| \leq (1 + Lh)z_n^0 + \delta = z_{n+1}^0.$$

Лемма доказана.

Если интегрирование производится на отрезке  $[x_0, x_0 + X]$ , то

$$nh \leq X, \quad (1 + Lh)^n \leq (\exp \{Lh\})^n \leq \exp \{LX\}.$$

Согласно формуле Лагранжа при  $y \geq 0$  имеем  $e^y - 1 = ye^{\theta y} \leq ye^y$ , где  $0 \leq \theta \leq 1$ . Отсюда получаем

$$(1 + Lh)^n - 1 \leq e^{LX} - 1 \leq LX \exp \{LX\}.$$

В итоге имеем оценку

$$|\Delta_n| \leq |z_n^0| \leq |\Delta_0| \exp \{LX\} + \delta h^{-1} X \exp \{LX\}.$$

Рассмотрим случай  $\Delta_0 = 0$ . Тогда погрешность  $\Delta_n = y_n^* - y_n$  при фиксированном  $X$  оценивается сверху через  $O(\delta h^{-1})$ .

Эта оценка неулучшаема по порядку. Например, при  $\Delta_0 = 0$ ,  $f_y = 0$ ,  $\delta_n \equiv \delta$  имеем  $\Delta_{n+1} = \Delta_n + \delta$  и, таким образом,

$$\Delta_N = N\delta = \delta X h^{-1}.$$

Приведем правдоподобные рассуждения, из которых следует, что погрешность округления может оказаться величиной порядка  $\delta h^{-1}$ . Соотношение

$$y_{n+1}^* = y_n^* + hf(x_n, y_n^*) + \delta_n$$

можно переписать в виде

$$y_{n+1}^* = y_n^* + hf^*(x_n, y_n^*),$$

где

$$f^*(x_n, y_n^*) = f(x_n, y_n^*) + \delta_n h^{-1}.$$

Таким образом, результат численного интегрирования уравнения  $y' = f(x, y)$  при наличии округлений будет такой же, как если бы без округлений интегрировалось уравнение со значениями правой части в узлах сетки  $f^*(x_n, y_n^*)$ . Рассмотрим случай  $\delta_n \equiv \delta$ . Разность между решениями дифференциальных уравнений

$$y' = f(x, y) \quad \text{и} \quad y' = f(x, y) + \delta h^{-1}$$

имеет порядок разности между правыми частями этих уравнений, т. е.  $\delta h^{-1}$ . Нет оснований ожидать, что решения разностных уравнений  $y_{n+1} = y_n + hf(x_n, y_n)$  и  $y_{n+1}^* = y_n^* + h(f(x_n, y_n^*) + \delta h^{-1})$  будут отличаться менее, чем на величину того же порядка, т. е. порядка  $\delta h^{-1}$ . Если  $\delta$  порядка  $2^{-t}$ ,  $t$  — разрядность чисел в ЭВМ, то решение разностного уравнения изменится на величину порядка  $2^{-t}h^{-1}$ .

При получении этого вывода сделано допущение  $\delta_n \equiv \delta$ . Рассмотрим задачу вычисления интеграла  $\int_0^1 f(x) dx$ , являющуюся частным случаем рассматриваемой задачи при  $y(0) = 0$ . Пусть  $f(x) = 2/3$ ,  $h = 2^{-k}$ ,  $t - k$  нечетно. При  $x_n > 3/4$   $y_n$  лежит в пределах  $(1/2, 1)$ . Тогда после округления при сложении

$$\begin{aligned} y_n &= 0, & 1 & \alpha_2 \dots \alpha_t, \\ hf_n &= 0, & \underbrace{0 \dots 0}_k & \underbrace{10 \dots 101}_{t-k} \underbrace{0101} \dots \end{aligned}$$

каждый раз происходит отбрасывание подчеркнутой  $\underline{\quad}$  величины, равной  $(1/3)2^{-t}$ . Таким образом, на этом участке действительно  $\delta_n \equiv \delta$  порядка  $2^{-t}$ .

Перейдем к случаю интегрирования уравнения  $y'' = f(x, y)$  при помощи простейшего метода

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} = f(x_n, y_n)$$

по расчетной формуле

$$y_{n+1} = 2y_n - y_{n-1} + h^2 f(x_n, y_n).$$

Реально получаемые значения  $y_n^*$  связаны соотношением

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + h^2 f(x_n, y_n^*) + \delta_n. \quad (5)$$

Значения  $y_n^*$  можно рассматривать как получаемые без округлений при вычислениях по формуле

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + h^2 f^*(x_n, y_n^*),$$

где  $f^*(x_n, y_n^*) = f(x_n, y_n^*) + \delta_n h^{-2}$ .

Рассмотрим случай  $\delta_n \equiv \delta$ . Решения дифференциальных уравнений  $y'' = f(x, y)$  и  $y'' = f(x, y) + \delta h^{-2}$  различаются между собой на величину порядка  $\delta h^{-2}$ .

**Задача 1.** Показать, что в случае уравнения  $y'' = \alpha$ ,  $\alpha = \text{const}$ , возможен случай  $\delta_n \equiv \delta$ ,  $\delta$  порядка  $2^{-t}$ .

Как и для уравнения первого порядка, делаем вывод, что для рассматриваемого алгоритма суммарная вычислительная погрешность может оказаться величиной порядка  $2^{-t}h^{-2}$ . Требование малости этой величины накладывает ограничение снизу на допустимый шаг интегрирования.

Рассмотрим случай, когда такая величина вычислительной погрешности оказывается недопустимо большой. Можно было бы записать рассматриваемое уравнение в виде системы уравнений первого порядка

$$y' = v, \quad v' = f(x, y) \quad (6)$$

и применить какой-либо метод численного интегрирования этой системы. Как уже отмечалось, при этом произошла бы потеря эффективности, поскольку методы общего вида не учитывают специфики этой системы.

Попытаемся записать рассматриваемую расчетную формулу как некоторую расчетную формулу интегрирования системы (6). Введем новую дискретную переменную

$$\frac{y_n - y_{n-1}}{h} = z_n,$$

тогда уравнение (5) запишется в виде

$$\frac{z_{n+1} - z_n}{h} = f(x_n, y_n).$$

Вычисления последовательных значений  $y_n$ ,  $z_n$  будем производить при помощи пары расчетных формул

$$z_{n+1} = z_n + hf(x_n, y_n), \quad y_{n+1} = y_n + hz_{n+1}. \quad (7)$$

При наличии округлений соответственно имеем

$$z_{n+1}^* = z_n^* + hf(x_n, y_n^*) + \alpha_n, \quad y_{n+1}^* = y_n^* + hz_{n+1}^* + \beta_n,$$

где  $\alpha_n, \beta_n = O(2^{-t})$ . Эти соотношения можно представить в виде

$$\begin{aligned} z_{n+1}^* &= z_n^* + hf^*(x_n, y_n^*), \quad f^*(x_n, y_n^*) = f(x_n, y_n^*) + \alpha_n h^{-1}, \\ y_{n+1}^* &= y_n^* + h(z_{n+1}^* + g(x_{n+1})), \quad g(x_{n+1}) = \beta_n h^{-1}. \end{aligned} \quad (8)$$

Если формулы (7) можно трактовать как формулы численного интегрирования системы  $z' = f(x, y)$ ,  $y' = z$ , то формулы (8) соответствуют системе

$$z' = f^*(x, y), \quad y' = z + g(x).$$

Правые части этих систем различаются на величины порядка  $O(2^{-t}h^{-1})$ ; поэтому есть какие-то основания ожидать, что и решения разностных задач, т. е. решения разностной задачи с округлениями и разностной задачи без округлений будут различаться на величину того же порядка.

**Задача 2.** Доказать справедливость этого утверждения.

Конечно, к этому заявлению следует отнестись с осторожностью; мы уже видели, что для некоторых конечно-разностных схем малые погрешности могут приводить к катастрофическому изменению результата.

Приведенные рассуждения о влиянии вычислительной погрешности в конкретных методах интегрирования уравнений первого и второго порядков опираются лишь на учет свойств конечно-разностной схемы, связанных с порядком дифференциального уравнения. Поэтому они переносятся на другие конечно-разностные методы. Например, при интегрировании уравнения  $y^{(k)} = f(x, y)$  и прямом использовании схемы

$$\nabla^k y_n - h^k \sum_{i=0}^m a_{-i} f(x_{n-i}, y_{n-i}) = 0$$

следует ожидать влияния вычислительной погрешности порядка  $2^{-i}h^{-k}$ . Таким образом, в случае уравнений высокого порядка еще более актуальна задача преобразования схемы к форме, где влияние вычислительной погрешности будет меньше. Например, аналогично случаю  $k = 2$  целесообразно ввести вспомогательные переменные

$$z_n^i = \nabla^i y_n h^{-i}, \quad i = 1, \dots, k-1.$$

Попытаемся объяснить улучшение свойств разностной схемы (5) при переходе к расчетным формулам (7), оценивая количество хранимой информации. При использовании расчетной формулы (5) при каждом  $n$  в памяти ЭВМ хранятся величины  $y_{n-1}$  и  $y_n$  и все дальнейшие значения  $y_j$  определяются по этим значениям. Пусть для определенности  $1/2 \leq y_{n-1}, y_n \leq 1$  и  $|y_n - y_{n-1}| \leq Mh$ . В ячейке, содержащей значение  $y_{n-1} = 0, 1\alpha_2 \dots \dots \alpha_t$ , имеется  $t-1$  независимых двоичных знаков  $\alpha_2, \dots, \alpha_t$ ; разряды числа  $y_n = 0, 1\beta_2 \dots \beta_t$  уже не все несут новую информацию.

Дело заключается в следующем. Пусть  $l$  наибольшее целое, такое, что  $Mh < 2^{-l+1}$ . Тогда имеем

$$y_n - y_{n-1} = \pm 0, 0 \dots 0\gamma_{l+1} \dots \gamma_t,$$

и для задания  $y_n - y_{n-1}$ , а следовательно, и  $y_n$  достаточен  $t-l+1$  двоичный знак (знак разности и  $\gamma_{l+1}, \dots, \gamma_t$ ). Поскольку  $l \sim \log_2((Mh)^{-1})$ , то общее количество независимой информации, которое имеется в нашем распоряжении при каждом  $n$ , составляет  $2t - \log_2((Mh)^{-1})$  двоичных разрядов (с точностью до слагаемого, не зависящего от  $t$  и  $h$ ).

В случае вычислений по формуле (7) все разряды чисел  $y_n$  и  $z_n$  независимы, поэтому информация о решении задается независимыми двоичными разрядами.

Тот факт, что количество независимой информации для второго способа больше, конечно, не означает, что этот способ лучше. Не исключено, что эта дополнительная информация не является содержательной и поэтому не позволяет точнее определить решение. Поясним, почему дополнительная информация при втором способе является содержательной,

Для определения решения дифференциального уравнения второго порядка с точностью  $O(\epsilon)$  требуется задание с такой же по порядку точностью значения решения и его производной в некоторой точке. Для разностного уравнения роль этих величин играют  $y_n$  и  $(y_n - y_{n-1})/h$ . При первом способе задание значений  $y_n$  и  $y_{n-1}$  с  $t$  двоичными знаками позволяет определить значение  $(y_n - y_{n-1})/h$  с погрешностью порядка  $O(2^{-t}h^{-1})$ . Таким образом, здесь по известной нам информации мы располагаем возможностью найти дальнейшие значения решения сеточной задачи с погрешностью порядка  $O(2^{-t}h^{-1})$  (если все последующие вычисления будут производиться абсолютно точно). В случае второго способа мы имеем значения  $y_n$  и  $(y_n - y_{n-1})/h$  с  $t$  двоичными знаками, поэтому обладаем возможностью найти решение сеточной задачи с погрешностью порядка  $O(2^{-t})$ . Таким образом, эта дополнительная информация действительно оказывается содержательной. На каждом шаге реального численного интегрирования погрешности округления вносят дополнительную неопределенность в значения вектора  $(y_n, (y_n - y_{n-1})/h)$ : во втором случае порядка  $O(2^{-t})$ , в первом — порядка  $O(2^{-t}/h)$ . Это и приводит к тому, что суммарная вычислительная погрешность решения во втором случае может оказаться величиной порядка  $O(2^{-t}h^{-1})$ , а в первом — порядка  $O(2^{-t}h^{-2})$ .

Рассмотрим метод прогонки решения сеточной краевой задачи (1.3), соответствующей уравнению второго порядка  $y''(x) - p(x)y(x) = f(x)$  при  $p(x) \equiv p = \text{const}$ .

Коэффициенты  $C_n$  вычисляются по рекуррентным формулам  $C_{n+1} = (2 + ph^2 - C_n)^{-1}$ ; в случае  $p(x) \equiv p$  целесообразно заранее вычислить  $Q = 2 + ph^2$  и вести вычисления по формуле  $C_{n+1} = (Q - C_n)^{-1}$ . При вычислении суммы  $2 + ph^2$  произойдет округление, т. е. получится величина  $Q^* = 2 + ph^2 + \delta$ , где  $\delta$  может оказаться величиной порядка  $2^{-t}$ . Это равносильно тому, что без округлений решается уравнение  $y''(x) - p^*y(x) = f(x)$ , где  $p^* = p + \delta h^{-2}$ . Рассуждая как и выше, получаем, что такое возмущение коэффициента  $p$  может привести к возмущению решения на величину порядка  $2^{-t}h^{-2}$ .

Если коэффициенты  $C_n$  существенно больше 1, то при вычислении выражения  $ph^2 + 2 - C_n$  погрешность округления может оказаться величиной порядка  $2^{-t}h^{-2}|C_n|$ . Поскольку вклад от погрешности в одной точке в суммарную погрешность умножается на коэффициент порядка  $h$ , то влияние этого округления на окончательный результат порядка  $2^{-t}h^{-1}|C_n|$ . Если  $|C_n| \gg h^{-1}$ , то это выражение будет существенно больше чем  $2^{-t}h^{-2}$ . Таким образом, наличие точек, где  $|C_n| > h^{-1}$ , приводит к еще большему влиянию вычислительной погрешности.

**З а м е ч а н и е.** Возмущения, вносимые другими округлениями при вычислениях  $C_n$  и  $\varphi_n$ , равносильны дополнительным возмущениям коэффициентов  $p$  и  $f$ .

Чтобы погрешность решения системы (1) была существенно меньше, необходимо по крайней мере задавать ее в форме, где округления коэффициентов системы равносильны существенно меньшим возмущениям коэффициентов

исходной дифференциальной задачи. С этой целью можно, например, перейти к системе

$$\frac{y_n - y_{n-1}}{h} = z_n, \quad \frac{z_{n+1} - z_n}{h} - p_n y_n = f_n. \quad (9)$$

Соответственно при решении этой системы вместо рекуррентных соотношений (3.11) относительно коэффициентов  $C_n$ ,  $\varphi_n$  следует перейти к рекуррентным соотношениям (4.9), (4.10) относительно  $\alpha_n$ ,  $\beta_n$ .

Выше рассматривался случай, когда при всех  $n$  погрешность округления оказывалась одной и той же. Если коэффициент  $p(x) \neq \text{const}$ , то при вычислении величины  $(2 + p_n h^2)$  при различных  $n$  округления могут оказаться различными знаков, и поэтому суммарная погрешность может оказаться по порядку меньшей чем  $2^{-t} h^{-2}$ . В случае задачи Коши для уравнения  $y' = f(x, y)$  при определенных условиях (в частности,  $h = 2^{-t}$ ,  $2^{-t} h^{-2} \ll 1$ ) установлено, что при некоторой мере в пространстве правых частей вычислительная погрешность для большинства задач накапливается медленнее — как  $\delta h^{-1/2}$ . Аналогичный выигрыш в статистической оценке погрешности порядка  $h^{1/2}$ , по-видимому, имеет место и по отношению к уравнениям более высокого порядка.

Отмеченные моменты не влияют на справедливость полученных выше качественных выводов о целесообразности записи расчетных схем для уравнений высших порядков в виде аналогов расчетных схем для уравнений низших порядков.

Обратим внимание на прием практической оценки вычислительной погрешности путем изменения масштабов, применяемый иногда для экспериментального исследования чувствительности метода к вычислительной погрешности. Пусть некоторым методом решается задача Коши

$$\frac{dy}{dx} = f(x, y), \quad y(0) = a.$$

Замена переменных  $x = \mu t$ ,  $y = \lambda z$  сводит эту задачу к задаче

$$\frac{dz}{dt} = \frac{\mu}{\lambda} f(\mu t, \lambda z), \quad z(0) = \frac{a}{\lambda}.$$

Предположим, что первая задача интегрировалась с шагами  $h_i$ ; осуществим численное интегрирование второй задачи тем же методом, но с шагами  $h'_i = h_i/\mu$ .

При отсутствии округлений будет иметь место равенство  $y_i = \lambda z_i$ ; если  $\lambda$  и  $\mu$  оба не являются целыми степенями двойки, то разность между реально получаемыми значениями величин  $y_i$  и  $\lambda z_i$  обычно дает определенное представление о величине вычислительной погрешности. Например, можно взять  $\mu = \sqrt{3}$ ,  $\lambda = \sqrt{2}$ .

В случае интегрирования с автоматическим выбором шага надо обратить внимание на выбор нового значения параметра  $\epsilon$  и начального шага  $h_0$  с тем, чтобы при всех  $i$  имело место равенство  $h'_i = h_i/\mu$ .

Литература к гл. IX: [3, 31, 38].



## Глава X

### МЕТОДЫ РЕШЕНИЯ УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

В случае решения обыкновенных дифференциальных уравнений мы видели следующую картину. Имеется ряд уравнений, интегрируемых в квадратурах. Решение большинства уравнений можно получить, используя лишь численные методы. Принципиально различных постановок задач довольно немного: в основном это задача Коши, краевая задача для линейных уравнений, краевая задача для нелинейных уравнений. Имеется небольшое количество теоретически исследованных и практически отработанных алгоритмов, позволяющих эффективно решать существенную часть задач, связанных с численным решением обыкновенных дифференциальных уравнений. В частности, ряд численных методов решения задачи Коши был разработан еще в прошлом веке.

В настоящее время разработка методов и алгоритмов решения задачи Коши для обыкновенных дифференциальных уравнений продвинута настолько, что зачастую исследователь, имеющий дело с этой задачей, не занимается выбором метода ее решения, а просто обращается к стандартной программе.

В случае уравнений с частными производными число принципиально различных постановок задач существенно больше. В курсе уравнений с частными производными обычно рассматривается незначительная часть таких постановок, главным образом связанных с линейными уравнениями с постоянными коэффициентами. При этом существует очень малое количество задач, решаемых в явном виде. Многообразие постановок в теории уравнений с частными производными связано с многообразием явлений окружающего нас мира.

Большое количество различных постановок задач, связанных с решением уравнений в частных производных, привело к тому, что теория численных методов в этом направлении дробится на большое число направлений. Использование численных методов с применением ЭВМ сильно расширило возможности в исследовании подобных задач. Например, разработанные за последние тридцать лет алгоритмы дают возможность решать с приемлемой затратой машинного времени подавляющее большинство краевых задач, связанных с решением одно- и много-

мерных уравнений параболического типа с переменными, в частности с нелинейно зависящими от решения коэффициентами.

Конечно, здесь иначе, чем в случае обыкновенных дифференциальных уравнений, обстоит дело с обоснованием сходимости численных методов и оценкой погрешности. Для широких классов типовых задач такие исследования проведены. Однако для многих важных классов прикладных задач, предъявляемых математикам для решения, не только не доказан, но часто остается неясным сам факт существования решения.

Например, уже для простой на вид системы

$$v_t + u_x = 0, \quad u_t + \gamma(v^2)_x = 0, \quad (1)$$

описывающей одномерное адиабатическое течение газа в лагранжевых координатах, не доказан при  $|\gamma| > 1$  факт существования решения в целом (т. е. на неограниченном промежутке времени).

При таком состоянии вопроса о существовании решения в настоящее время трудно ожидать получения строгих оценок погрешности и теорем о сходимости сеточных методов при достаточно общих предположениях. Тем не менее, часто пользуясь полуэмпирическими соображениями, аналогиями по сравнению со случаем линейных уравнений и численными экспериментами на задачах с известным точным решением, математики создают численные методы для решения и таких задач. При этом результаты и анализ численных расчетов наравне с экспериментом оказывают существенное влияние на развитие соответствующих разделов теории уравнений с частными производными. Так, например, обстоит дело в случае решения уравнений газовой динамики типа (1).

Несмотря на отсутствие строгих обоснований чисто математической (в частности, алгоритмической) стороны вопроса, математикам, занимающимся решением подобных прикладных задач, часто приходится брать на себя ответственность за достоверность получаемых численных результатов, включая правильность математической постановки задачи.

Конечно, все сказанное не умаляет роли чисто теоретических исследований. Их результаты, в частности теоремы существования, дают уверенность в правильности постановки, подсказывают информацию о качественных свойствах решения, что крайне важно при выборе алгоритма. Наличие известных частных решений, например в задачах газовой динамики, позволяет производить проверку точности предлагаемых методов. Использование известных частных решений простейших задач часто позволяет облегчить численное решение более сложных задач.

## § 1. Основные понятия теории метода сеток

На первых этапах практического решения задач для уравнений с частными производными применялись в основном вариационные и другие методы, где приближенное решение полу-

чается в виде некоторой аналитической формулы. Для решения отдельных классов задач с малой точностью такие методы применяются и в настоящее время.

В последующий период наиболее актуальными для решения явились задачи динамики газа и жидкости, где подобные методы практически неприменимы. На решение этих задач были направлены усилия крупнейших математиков, что имело, в частности, своим следствием создание и широкое продвижение сеточных методов решения уравнений с частными производными. В настоящее время эти методы наряду с вариационно- и проекционно-разностными (метод конечных элементов) являются наиболее распространенными. При решении задач сеточными методами мы получаем совокупность приближенных значений решения в некоторой конечной системе точек. В случае необходимости можно построить формулу (например, интерполяционную) для приближенного представления решения во всей области.

Рассмотрим простейший пример решения задачи сеточным методом. Пусть в полуполосе  $0 \leq x \leq 1$ ,  $0 \leq t < \infty$  решается уравнение

$$u_t - u_{xx} = f(x, t) \quad (2)$$

при начальном и граничных условиях

$$u(x, 0) = \varphi(x), \quad u_x(0, t) + a(t)u(0, t) = b(t), \quad u(1, t) = 0. \quad (3)$$

Задан некоторыми  $h, \tau > 0$  ( $1/h = M$  целое) — шагами сетки. Точки  $(mh, n\tau)$  назовем узлами сетки  $(m, n)$ ; пусть  $u_m^n$  — приближения к значениям  $u(mh, n\tau)$ ,

$$f_m^n = f(mh, n\tau), \quad a^n = a(n\tau), \quad b^n = b(n\tau),$$

$u_h$  и  $f_h$  — функции, определенные на сетке, со значениями в узлах сетки  $u_m^n$  и  $f_m^n$  соответственно. Если искомое решение исходной дифференциальной задачи есть гладкая функция, то выполняются соотношения

$$\begin{aligned} L_h u(mh, n\tau) |_{(m, n)} &\equiv \frac{u(mh, (n+1)\tau) - u(mh, n\tau)}{\tau} - \\ &- \frac{u((m+1)h, n\tau) - 2u(mh, n\tau) + u((m-1)h, n\tau)}{h^2} = \\ &= u_t(mh, n\tau) - u_{xx}(mh, n\tau) + O(h^2 + \tau) = \\ &= f(mh, n\tau) + O(h^2 + \tau), \quad (4) \end{aligned}$$

$$\begin{aligned} l_h u(mh, n\tau) |_{(0, n)} &= \frac{u(h, n\tau) - u(0, n\tau)}{h} + a(n\tau)u(0, n\tau) = \\ &= u_x(0, n\tau) + a(n\tau)u(0, n\tau) + O(h) = b(n\tau) + O(h). \quad (5) \end{aligned}$$

Исходя из этого можно сделать предположение, что решение системы

$$L_h u_n |_{(m, n)} = f_m^n, \quad 0 < m < M, \quad 0 \leq n, \quad (6)$$

$$l_h u_n |_{(0, n)} = b^n, \quad u_M^n = 0, \quad n > 0, \quad (7)$$

$$u_m^0 = \varphi(n; h), \quad 0 \leq m \leq M, \quad (8)$$

является приближением к точному решению исходной задачи. Значения решения системы (6) — (8) можно находить последовательно при каждом  $n$  следующим образом:  $u_m^0$  нам заданы, при каждом  $n$  величина  $u_M^n = 0$ , значения  $u_m^n$  при  $0 < m < M$  находим из (6), а затем  $u_0^n$  — из (7). Приведем простейшую теорему, показывающую, что сходимость приближенного решения сеточной задачи к решению дифференциальной при  $\tau, h \rightarrow 0$  не обязательно имеет место.

Пусть в области  $D$  с границей  $\Gamma = \bigcup_{i=1}^s \Gamma_i$  решается краевая задача

$$L(u) = f \quad (9)$$

при граничных условиях

$$l_i(u) = \varphi_i \quad \text{на } \Gamma_i, \quad i = 1, \dots, s. \quad (10)$$

Фиксируем какую-либо точку  $P$  в пространстве независимых переменных. Значение решения  $u(P)$  зависит от коэффициентов уравнения, правой части  $f$ , функций  $\varphi_i$ .

Рассмотрим случай, когда уравнение (9) и граничные условия (10) линейные, и изучим вопрос о зависимости  $u(P)$  от значений функции  $\varphi_1$ .

(В случае нелинейного уравнения или исследования зависимости от коэффициентов уравнения, которая является нелинейной зависимостью, проводимые далее рассуждения требуют для своего обоснования некоторых дополнительных формальных построений.)

Пусть  $\Phi_1$  — некоторый класс функций  $\varphi_1$ . Областью зависимости  $\Omega_1(P, \Phi_1)$  значения  $u(P)$  по граничному условию  $l_1 u = \varphi_1$  в классе функций  $\Phi_1$  назовем множество точек  $Q \in \Gamma_1$ , удовлетворяющих следующему соотношению: для любой окрестности точки  $Q$  найдутся две функции  $\varphi_1^1, \varphi_1^2 \in \Phi_1$  такие, что

- 1)  $\varphi_1^1 = \varphi_1^2$  вне этой окрестности;
- 2) если  $u_1$  и  $u_2$  — решения задачи (9), (10) при одних и тех же  $f, \varphi_2, \dots, \varphi_s = 0$  и при  $\varphi_1 = \varphi_1^1$  и  $\varphi_1 = \varphi_1^2$  соответственно, то  $u_1(P) \neq u_2(P)$ .

При решении задачи (9), (10) методом сеток решение ищется в узлах некоторой сетки в пространстве независимых переменных. Будем считать, что сетка задается некоторым параметром  $h$ , причем  $h \rightarrow 0$  при измельчении сетки. Если точка  $P$  является узлом сетки, то при данном  $h$  решение сеточной за-

дачи, аппроксимирующей исходную дифференциальную задачу, зависит от значений  $\varphi_1$  в некоторых точках границы  $\Gamma_1$ . Множество этих точек обозначим через  $\Omega_1(P, h)$ .

**Теорема Куранта** (об областях зависимости). *Для того чтобы для некоторых  $f, \varphi_2, \dots, \varphi_s$  и всех  $\varphi_1 \in \Phi_1$  при  $h \rightarrow 0$  значения в точке  $P$  решений сеточной задачи стремились к значению в точке  $P$  решения дифференциальной задачи, необходимо выполнение следующего условия: все точки множества  $\Omega^1(P, \Phi_1)$  являются предельными для точек множеств  $\Omega_1(P, h)$ , иначе говоря,*

$$\Omega_1(P, \Phi_1) \subset \overline{\lim}_{h \rightarrow 0} \Omega_1(P, h).$$

**Доказательство.** Предположим противное, т. е. что существует точка  $Q \in \Omega_1(P, \Phi_1)$ ,  $Q \notin \overline{\lim}_{h \rightarrow 0} \Omega_1(P, h)$ . Тогда имеется окрестность точки  $Q$ , не содержащая точек множеств  $\Omega_1(P, h)$  при  $h < h_0$ . Возьмем функции  $\varphi_1^1, \varphi_1^2 \in \Phi_1$ , соответствующие этой окрестности согласно приведенному выше определению области зависимости. Пусть  $u^1$  и  $u^2$  — решения задачи (9), (10) при  $\varphi_1 = \varphi_1^1, \varphi_1 = \varphi_1^2$  соответственно и  $f, \varphi_2, \dots, \varphi_s = 0$ . Пусть  $u^0$  — решение задачи (9), (10) при  $\varphi_1 = 0$  и заданных  $f, \varphi_2, \dots, \varphi_s$ . Тогда  $u_1 = u^1 + u^0, u_2 = u^2 + u^0$  будут решениями задачи (9), (10) при  $\varphi_1 = \varphi_1^1$  и  $\varphi_1 = \varphi_1^2$  соответственно и тех же  $f, \varphi_2, \dots, \varphi_s$ . Значения решений сеточных задач при соответствующих  $u_1$  и  $u_2$  начальных и граничных условиях будут совпадать при  $h < h_0$ . Поскольку в то же время  $u_1(P) \neq u_2(P)$ , то для одного из этих наборов граничных условий решение сеточной задачи не сойдется к решению дифференциальной. Теорема доказана.

Рассмотрим пример применения этой теоремы. Пусть в полуплоскости  $t \geq 0$  решается задача Коши  $u_t + au_x = 0$  при начальном условии  $u(0, x) = u_0(x)$ . Зададимся сеткой с узлами в точках  $(mh, n\tau)$  и заменим исходную дифференциальную задачу разностной:

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a \frac{u_{m+1}^n - u_m^n}{h} = 0,$$

$$u_m^0 = u_0(mh).$$

Тогда значения  $u_m^n$  при  $n > 0$  определяем последовательно из соотношения

$$u_m^{n+1} = (1 + a\tau/h)u_m^n - (a\tau/h)u_{m+1}^n.$$

Пусть при измельчении сетки  $\tau/h = \kappa = \text{const}$ . Тогда значения решения сеточной задачи в точке  $(x_0, t_0)$  не зависят от начальных значений условий вне отрезка  $[x_0, x_0 + \kappa^{-1}t_0]$ . Точное решение дифференциальной задачи есть  $u(x, t) = u_0(x - at)$ . Поэтому в классе начальных условий, обладающих некоторым ограниченным числом производных, областью зависимости для диф-

ференциальной задачи является точка  $x_0 - at_0$ . Таким образом, на основании теоремы Куранта необходимым условием сходимости в этом классе начальных условий является условие  $x_0 - at_0 \in [x_0, x_0 + \kappa^{-1}t_0]$  или, иначе,  $a \leq 0$ ,  $|a\tau/h| = |a|\kappa \leq 1$ .

Заметим, что в случае рассматриваемой схемы это условие случайно оказывается и достаточным условием сходимости.

Кроме вопроса о сходимости при анализе разностных аппроксимаций, необходимо также исследовать устойчивость получаемого результата относительно погрешностей в исходных данных задачи и при округлениях.

Проиллюстрируем сказанное на этом же примере. Пусть  $a > 0$ ,  $a\tau/h = 1$ . Тогда значения  $u_m^n$  определяются из рекуррентного соотношения

$$u_m^{n+1} = 2u_m^n - u_{m+1}^n.$$

При  $u_m^0 \equiv 0$  решением сеточной задачи будет  $u_m^n \equiv 0$ ; пусть теперь  $u_0^0 = \varepsilon$  и  $u_m^0 = 0$  при  $m \neq 0$ . Пользуясь этим рекуррентным соотношением, получаем, что тогда в узлах сетки  $u_m^n$  принимает следующие значения:

$$u_m^n = \begin{cases} 0 & \text{при } m > 0 \text{ или } m < -n, \\ C_n^{-m} 2^{n+m} (-1)^m \varepsilon & \text{при } -n \leq m \leq 0. \end{cases}$$

Отсюда  $\sum_{m=-\infty}^{\infty} |u_m^n| = 3^n \varepsilon$ , т. е. с ростом  $n$  разность между этим решением и решением  $u_m^n \equiv 0$  катастрофически возрастает; поэтому при  $a\tau/h = 1$  рассматриваемая схема не может быть признана пригодной для решения задачи в случае большого числа шагов.

Таким образом, при замене решения дифференциальной задачи решением разностной его аппроксимации возникают следующие проблемы (аналогичные проблемам, возникавшим ранее при рассмотрении методов решения других задач):

- 1) сходится ли точное решение разностной задачи к решению дифференциальной;
- 2) насколько сильно изменяется решение разностной задачи, если при вычислениях допускаются некоторые погрешности.

Построим формальный математический аппарат, помогающий при решении этих проблем.

Будем рассматривать задачу (9), (10). Относительно  $\Gamma_i$  будем считать, что  $\Gamma_i$  — заданные части границы  $\Gamma$ , причем различные  $\Gamma_i$  могут иметь непустое пересечение;  $l_i$  — некоторые операторы;  $f, \varphi_1, \dots, \varphi_s$  — заданные функции.

Определим некоторое множество точек в пространстве независимых переменных  $D_n$ , которое назовем сеткой (как правило,  $D_n$  выбирают так, чтобы оно принадлежало замкнутой области  $\bar{D}$ ). Обычно сетка, на которой отыскивается решение, зависит от нескольких параметров (в предыдущем примере от  $\tau$  и  $h$ );

однако во многих типичных случаях при дроблении сетки ее шаги связывают между собой каким-то законом вида  $\tau = Ah^\sigma$ . Поэтому в дальнейших общих построениях и определениях для простоты мы указываем зависимость только от одного параметра  $h > 0$ .

Пусть  $U_h$  — пространство функций  $u_h$ , определенных в точках сетки  $D_h$ , которые иначе называют узлами сетки,  $L_h$  — оператор, преобразующий функции из  $U_h$  в функции, определенные на некотором множестве  $D_h^0 \subset D_h$ ; будем предполагать, что  $D_h^0 \subset \bar{D}$ . Множество функций, определенных в точках  $D_h^0$ , будем обозначать через  $F_h$ . Для аппроксимации граничных условий (10) выбираются некоторые множества  $\Gamma_{ih} \subset \Gamma_i$  и в точках этих множеств определяются значения некоторых операторов над пространством функций  $U_h$ ; пусть  $\Phi_{ih}$  — пространство функций, определенных в точках множеств  $\Gamma_{ih}$ . Если  $X \subset Y$  и функция  $v$  определена на множестве  $Y$ , то ее следом на множестве  $X$  называют функцию, определенную на множестве  $X$  и совпадающую там с  $v$ . Если функция  $v$  определена на некотором множестве, содержащем  $D_h$ , то ее след на этом множестве будем обозначать  $[v]_h$ .

Пусть  $U$  — пространство, к которому мы относим решение задачи (9), (10);  $F$  — пространство правых частей  $f$ ;  $\Phi_i$  — пространства функций, определенных на  $\Gamma_i$ . Пусть в пространствах функций  $U, U_h, F, F_h, \Phi_i, \Phi_{ih}$  определены нормы

$$\| \cdot \|_U, \| \cdot \|_{U_h}, \| \cdot \|_F, \| \cdot \|_{F_h}, \| \cdot \|_{\Phi_i}, \| \cdot \|_{\Phi_{ih}}.$$

Эти нормы называют согласованными, если при  $h \rightarrow 0$  для любых достаточно гладких функций  $u \in U, f \in F, \varphi_i \in \Phi_i$  выполняются соотношения

$$\| [u]_h \|_{U_h} \rightarrow \| u \|_U;$$

$$\| [f]_h \|_{F_h} \rightarrow \| f \|_F;$$

$$\| [\varphi_i]_h \|_{\Phi_{ih}} \rightarrow \| \varphi_i \|_{\Phi_i}.$$

Говорят, что сеточная функция  $u_h$  сходится к решению задачи (9), (10), если

$$\| u_h - [u]_h \|_{U_h} \rightarrow 0 \quad \text{при } h \rightarrow 0.$$

Исследование сходимости разностных аппроксимаций имеет смысл производить лишь в нормах, согласованных с некоторыми нормами в пространствах гладких функций. Если отказаться от требования согласованности норм, то условие сходимости иногда может перестать быть содержательным: в случае любой последовательности сеточных функций  $u_h$  путем введения некоторого множителя в определение нормы можно добиться, чтобы эта последовательность сходилась к решению задачи  $u$ .

Рассмотрим некоторую сеточную задачу

$$L_h(u_h) = f_h, \quad (11)$$

$$l_{ih}(u_h) = \varphi_{ih}, \quad i = 1, \dots, s. \quad (12)$$

Говорят, что сеточная задача (11), (12) аппроксимирует дифференциальную задачу (9), (10), если выполняется следующее условие: при любых гладких  $u$ ,  $f$  и  $\varphi_i$

$$z(h) = \|L_h([u]_h) - [L(u)]_h\|_{F_h} + \|f_h - [f]_h\|_{F_h} + \\ + \sum_{i=1}^s (\|l_{ih}([u]_h) - [l_i(u)]_h\|_{\Phi_{ih}} + \|\varphi_{ih} - [\varphi_i]_h\|_{\Phi_{ih}}) \rightarrow 0, \quad \text{если } h \rightarrow 0. \quad (13)$$

Проиллюстрируем приведенные определения на примере рассмотренной аппроксимации уравнения теплопроводности. Через  $D$  будем обозначать множество точек  $0 < x < 1$ ,  $0 < t \leq T$ ; пусть  $\Gamma_1$  — отрезок  $[0, 1]$  оси  $x$ ,  $\Gamma_2$  — полуинтервал  $(0, T]$  оси  $t$ ,  $\Gamma_3$  — полуинтервал  $(0, T]$  прямой  $x = 1$ ; точки  $(0, 0)$ ,  $(1, 0)$  можно было бы отнести и к множествам  $\Gamma_2$ ,  $\Gamma_3$  соответственно. Если  $1/h = M$ ,  $M$  целое,  $N = [T/\tau]$ , то через  $D_h$  обозначим множество точек  $(mh, n\tau)$  — узлов  $(m, n)$ , удовлетворяющих условиям  $0 \leq m \leq M$ ,  $0 \leq n \leq N$ . Определим сеточный оператор  $L_h u_h$  соотношением

$$L_h u_h |_{(m, n)} = \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}. \quad (14)$$

Тогда множество  $D_h^0$  будет состоять из узлов  $(m, n)$  таких, что  $0 < m < M$ ,  $0 \leq n < N$ ; в остальных узлах  $(m, n)$  при  $u_h \in U_h$  значения  $L_h u_h$  не будут определены. Если бы мы положили правую часть (14) равной  $L_h u_h |_{(m, n+1)}$ , то множество  $D_h^0$  состояло бы из узлов  $(m, n)$  таких, что  $0 < m < M$ ,  $0 \leq n \leq N$ . Правую часть сеточной задачи выберем в виде

$$f_h |_{(m, n)} = f(mh, n\tau).$$

Тогда величина  $\|f_h - [f]_h\|_{F_h}$ , входящая в выражение  $z(h)$ , есть нуль. Это соотношение выполняется не для всех схем; например, из соображения повышения точности иногда разумнее было бы полагать правую часть разностной задачи в точке  $(m, n)$  равной  $f(mh, (n + 0,5)\tau)$ . В качестве согласованных норм  $\|\cdot\|_{U_h}$  и  $\|\cdot\|_U$  при исследовании этой задачи обычно выбирают нормы

$$\|u_h\|_{U_h} = \sup_{\substack{0 \leq n \leq N \\ 0 \leq m \leq M}} |u_m^n|, \\ \|u\|_U = \sup_{\substack{0 \leq t \leq T \\ 0 \leq x \leq 1}} |u(x, t)| \quad (15)$$



или же нормы

$$\|u_h\|_{U_h} = \sup_{0 \leq n \leq N} \sqrt{h \sum_{m=0}^M |u_m^n|^2},$$

$$\|u\|_U = \sup_{0 \leq t \leq T} \sqrt{\int_0^1 |u(x, t)|^2 dx}.$$
(16)

В дальнейшем для простоты изложения будем предполагать, что операторы  $L, l, L_h, l_{ih}$  линейные.

В этом случае вводится следующее определение *устойчивости (корректности) сеточной задачи* (11), (12). Эту задачу называют *устойчивой*, если при  $h \leq h_0$  существуют постоянные  $M_0$  и  $M_i$ , не зависящие от  $h$ , такие, что

$$\|u_h\|_{U_h} \leq M_0 \|L_h u_h\|_{F_h} + \sum_{i=1}^s M_i \|l_{ih} u_h\|_{\Phi_{ih}}.$$
(17)

Как видно из определения, в случае линейной задачи в определении устойчивости не входят функции  $f_h$  и  $\Phi_{ih}$ .

Посмотрим, какой смысл в этом определении. Для случая линейных задач *разностная схема* (11), (12) представляет собой систему линейных алгебраических уравнений. Поэтому из (17) следует, что при  $f_h \equiv 0, \Phi_{ih} \equiv 0$  система уравнений (11), (12) имеет лишь нулевое решение; поэтому на основании теоремы Кронекера — Капелли задача (11), (12) разрешима при любых правых частях  $f_h, \Phi_{ih}$ . Таким образом, в случае линейной задачи из условия устойчивости следует однозначная разрешимость системы сеточных уравнений при любых правых частях.

Если  $u_h^1$  и  $u_h^2$  — решения сеточных задач

$$L_h u_h^1 = f_h^1, \quad l_{ih} u_h^1 = \Phi_{ih}^1, \quad i = 1, \dots, s;$$

$$L_h u_h^2 = f_h^2, \quad l_{ih} u_h^2 = \Phi_{ih}^2, \quad i = 1, \dots, s,$$

то при линейных  $L_h$  и  $l_{ih}$  согласно (16) можно написать

$$\|u_h^1 - u_h^2\|_{U_h} \leq M_0 \|L_h u_h^1 - L_h u_h^2\|_{F_h} + \sum_{i=1}^s M_i \|l_{ih} u_h^1 - l_{ih} u_h^2\|_{\Phi_{ih}} =$$

$$= M_0 \|f_h^1 - f_h^2\|_{F_h} + \sum_{i=1}^s M_i \|\Phi_{ih}^1 - \Phi_{ih}^2\|_{\Phi_{ih}}.$$
(18)

Таким образом, в случае выполнения условия устойчивости решения сеточной задачи мало различаются между собой при малом изменении правых частей уравнения и граничных условий.

Пусть  $u \in U$ . Величину  $r_h^0 = L_h[u]_h - f_h$  называют *погрешностью аппроксимации уравнения на решении задачи*, а величины  $r_h^i = l_{ih}[u]_h - \Phi_{ih}$  — *погрешностями аппроксимации граничных условий на решении задачи*. Положим

$$\rho_0(h) = \|L_h[u]_h - f_h\|_{F_h}, \quad \rho_i(h) = \|l_{ih}[u]_h - \Phi_{ih}\|_{\Phi_{ih}}.$$

Если  $u$  — решение задачи (9), (10), то величину  $\rho(h) = \sum_{i=0}^s \rho_i(h)$  называют мерой погрешности аппроксимации разностной схемой (11), (12) дифференциальной задачи (9), (10) на решении. Если  $\rho(h) \rightarrow 0$  при  $h \rightarrow 0$  и  $u$  — решение (9), (10), то говорят, что (11), (12) аппроксимирует (9), (10) на решении задачи. Порядок величины  $\rho(h)$  при  $h \rightarrow 0$  называют порядком аппроксимации на решении.

Выше обсуждалась проблема чувствительности реально получаемого приближенного решения сеточной задачи к округлениям в процессе вычисления этого решения, или, иначе, проблема устойчивости приближенного решения сеточной задачи к погрешностям округления. Решение этой проблемы тесно связано с решением вопроса об устойчивости разностной задачи. Дело в том, что округления, допускаемые при вычислениях, можно рассматривать как возмущения коэффициентов сеточной задачи.

Найдем связь между аппроксимацией, устойчивостью и сходимостью. Предположим, что сеточная аппроксимация (11), (12) удовлетворяет следующим условиям:

1) решение дифференциальной задачи удовлетворяет точно  $(s - k)$ -сеточным граничным условиям

$$l_{ih}[u]_h = \varphi_{ih}, \quad i = k + 1, \dots, s, \quad \text{т. е. } \rho_i(h) = 0, \\ i = k + 1, \dots, s;$$

2) на классе функций из  $U_h$ , удовлетворяющих однородным граничным условиям

$$l_{ih}u_h = 0, \quad i = k + 1, \dots, s,$$

выполняется условие устойчивости

$$\|u_h\|_{U_h} \leq M_0 \|L_h u_h\|_{F_h} + \sum_{i=1}^k M_i \|l_{ih} u_h\|_{\Phi_{ih}}.$$

Теорема Филиппова (о связи устойчивости, аппроксимации и сходимости). При сформулированных выше условиях выполняется неравенство

$$\|u_h - [u]_h\|_{U_h} \leq \sum_{i=0}^k M_i \rho_i(h). \quad (19)$$

Если разностная задача аппроксимирует дифференциальную, то

$$\|u_h - [u]_h\|_{U_h} \rightarrow 0 \quad \text{при } h \rightarrow 0.$$

Доказательство. Поскольку  $l_{ih}(u_h - [u]_h) = 0$  при  $i = k + 1, \dots, s$ , то воспользуемся условием 2), подставив в него  $u_h - [u]_h$  вместо  $u_h$ . Имеем

$$\|u_h - [u]_h\|_{U_h} \leq M_0 \|L_h u_h - L_h [u]_h\|_{F_h} + \sum_{i=1}^k M_i \|l_{ih} u_h - l_{ih} [u]_h\|_{\Phi_{ih}};$$

подставляя сюда  $L_h u_h = f_h$ ,  $l_{ih} u_h = \varphi_h$  и воспользовавшись определением  $\rho_i(h)$ , получаем (19). Если имеет место аппроксимация, т. е.  $\rho_i(h) \rightarrow 0$ ,  $i = 0, \dots, k$ ,  $\rho(h) \rightarrow 0$  при  $h \rightarrow 0$ , то из (19) следует справедливость второго утверждения теоремы:  $\|u_h - [u]_h\|_{U_h} \rightarrow 0$ .

В случае гладких решений исследование аппроксимации схемы на решении является относительно несложной задачей и теорема Филиппова переносит центр тяжести на исследование устойчивости сеточной задачи.

Часто случается, что сеточная задача устойчива в одной норме, согласованной с некоторой дифференциальной нормой, но неустойчива в другой. Так может, например, обстоять дело в случае норм, определяемых равенствами (15), (16). В случае гладких решений для практической приемлемости схемы обычно достаточно устойчивости в какой-либо согласованной норме. В случае разрывных решений к разностным аппроксимациям часто предъявляются некоторые дополнительные требования относительно поведения их решений вблизи мест разрыва решений; в этих случаях часто недостаточно устойчивости в произвольной согласованной норме. Например, в случае решения задач газовой динамики можно указать много сеточных аппроксимаций, устойчивых в метрике (16). Однако при практическом применении стараются использовать схемы, устойчивые не только в метрике (16), но и в метрике (15).

Если выполняется условие согласования, то при гладких  $u$ , переходя к пределу в (17) при  $h \rightarrow 0$ , получаем неравенство

$$\|u\|_U \leq M_0 \|Lu\|_F + \sum_{i=1}^s M_i \|l_i u\|_{\Phi_i}. \quad (20)$$

Из этого соотношения следует корректность постановки дифференциальной задачи (9), (10). Такой путь — получение оценок (17), а из них оценок (20) — часто используется для исследования корректности дифференциальных задач вида (9), (10), для доказательства существования и единственности их решений.

## § 2. Аппроксимация простейших гиперболических задач

Изучение многих важных прикладных задач требует численного решения краевых задач для систем уравнений с частными производными гиперболического типа. Такими системами являются, например, системы уравнений газовой динамики, которые являются квазилинейными системами. Для решений таких систем типично наличие разрывов. В таких задачах в большинстве случаев отсутствует строгое исследование вопросов устойчивости разностных схем и сходимости и реальный отбор разностных схем производится на примере простейших модельных задач, где его проще осуществить и теоретически, и путем численного эксперимента.

Простейшими примерами, на которых производился отбор конечно-разностных методов решения задач газовой динамики, являются уравнения

$$u_t + au_x = 0 \quad (1)$$

и

$$u_t + (\varphi(u))_x = 0. \quad (2)$$

Далее будут рассмотрены некоторые явные аппроксимации уравнения (1), а затем и (2).

При практическом анализе разностных аппроксимаций задачи Коши для гиперболических и параболических уравнений часто руководствуются следующим критерием, называемым *спектральным признаком устойчивости*.

Пусть на сетке с узлами  $(m, n)$  — точками  $(x_m, t_n) = (mh, n\tau)$  — построена некоторая аппроксимация, например вида

$$L_h u_h |_{m,n} = \sum_{i,k} a_{ik} u_{m+i}^{n+k} = 0; \quad (3)$$

выпишем все частные решения уравнения  $L_h u_h = 0$ , имеющие вид

$$u_m^n = (\lambda(\varphi))^n e^{im\varphi}.$$

**Спектральный признак устойчивости (СПУ).**  
Если при заданном законе стремления шагов  $\tau$  и  $h$  к нулю существует  $C < \infty$  такое, что

$$|\lambda(\varphi)| \leq e^{C\tau} \quad \text{для любых } \varphi, \quad (4)$$

то разностная схема может быть применена для численного решения соответствующей задачи Коши.

В противном случае от применения разностной схемы следует воздержаться.

В разумности СПУ можно убедиться, решив следующие задачи.

**Задача 1.** Пусть определена какая-то норма  $\|\cdot\|_n$  на сеточном слое по времени такая, что номер слоя не входит в определение нормы. Этому условию удовлетворяют, например, нормы

$$\|u_m^n\|_n = \sup_m |u_m^n|, \quad (5)$$

$$\|u_m^n\|_n = \sqrt{\sum_n |u_m^n|^2}, \quad (6)$$

но не удовлетворяет норма

$$\|u_m^n\|_n = n \sup_m |u_m^n|.$$

Пусть схема двухслойная и явная, т. е. имеет вид

$$L_h u_h |_{(m,n)} = \alpha u_m^{n+1} + \sum_{|j| \leq k} a_j u_{m+j}^n = 0, \quad \alpha \neq 0, \quad (7)$$

и пусть СПУ не выполнен. Доказать, что ни для какого  $T > 0$  нельзя указать  $Q < \infty$  такое, что при всех  $n\tau \leq T$  выполнено соотношение  $\|u_m^n\|_n \leq Q \|u_m^0\|_0$ .

Задача 2. Пусть  $\tau/h = \text{const}$ , разностная схема (3) двухслойная и явная, т. е. имеет вид (7), и начальные условия финитные, т. е.  $u_m^0 = 0$  при  $|m| \geq M$ . Доказать справедливость оценки

$$\|u_m^n\|_n \leq e^{Cn\tau} \|u_m^0\|_0 \leq e^{CT} \|u_m^0\|_0$$

при  $n\tau \leq T$  в случае нормы (6).

Если краевая задача корректна и условие (4) выполнено, то, как правило, удается построить аппроксимацию граничного условия так, чтобы сеточная задача была устойчива (корректна). В то же время можно привести примеры задач Коши (для систем уравнений), где спектральный признак устойчивости выполнен, а сеточная задача некорректна.

Исследуем при помощи спектрального признака устойчивости ряд сеточных аппроксимаций задачи Коши для уравнения  $u_t + au_x = 0$  в полуплоскости  $t \geq 0$ .

Пример 1. Разностная аппроксимация

$$L_h u_h |_{(m, n)} = \frac{u_m^{n+1} - u_m^n}{\tau} + a \frac{u_{m+1}^n - u_{m-1}^n}{2h} = 0. \quad (8)$$

Подставляем сюда  $u_m^n = \lambda^n(\varphi) \exp\{im\varphi\}$ :

$$\frac{\lambda^{n+1}(\varphi) \exp\{im\varphi\} - \lambda^n(\varphi) \exp\{im\varphi\}}{\tau} + a \frac{\lambda^n(\varphi) \exp\{i(m+1)\varphi\} - \lambda^n(\varphi) \exp\{i(m-1)\varphi\}}{2h} = 0.$$

После сокращения на  $\lambda^n(\varphi) \exp\{im\varphi\}$  получим

$$\lambda(\varphi) = 1 - \frac{a\tau}{2h} (e^{i\varphi} - e^{-i\varphi}) = 1 - ia \frac{\tau}{h} \sin \varphi \quad (9)$$

(см. рис. 10.2.1). На рис. 10.2.1—10.2.4 изображаются наборы узлов (шаблоны), по которым выписываются аппроксимации, и кривые, которые описывает точка  $\lambda(\varphi)$  на комплексной плоскости при изменении  $\varphi$  в пределах  $0 \leq \varphi \leq 2\pi$ . Стрелка означает направление изменения  $\lambda(\varphi)$  при изменении  $\varphi$  от 0 до  $2\pi$  (при  $a > 0$ ). На рис. 10.2.1—10.2.4 цифрой 1 обозначен единичный круг.

Мы имеем  $|\lambda(\varphi)| = \sqrt{1 + (a^2\tau^2/h^2) \sin^2 \varphi}$ ,

$$\max_{0 \leq \varphi \leq 2\pi} |\lambda(\varphi)| = \left| \lambda\left(\frac{\pi}{2}\right) \right| = \sqrt{1 + (a^2\tau^2/h^2)}.$$

Если  $\tau = Ah^2$  при  $\tau, h \rightarrow 0$ , то

$$\max_{0 \leq \varphi \leq 2\pi} |\lambda(\varphi)| = \sqrt{1 + a^2 A^2 \tau} = 1 + a^2 A^2 \tau / 2 + O(\tau^2)$$

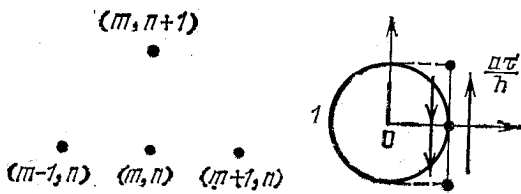


Рис. 10.2.1

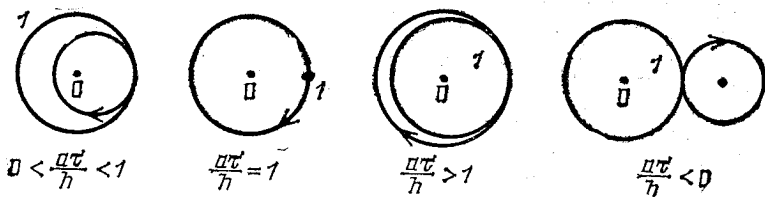


Рис. 10.2.2

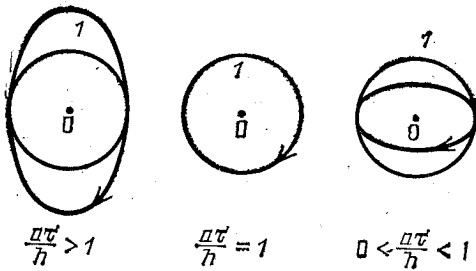


Рис. 10.2.3

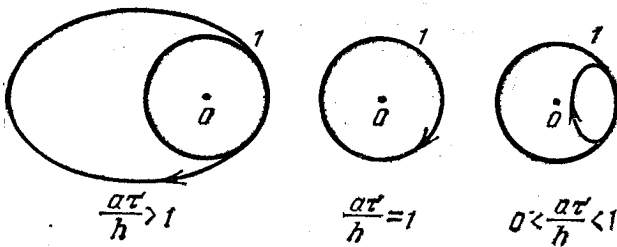


Рис. 10.2.4

и условие (5) выполняется; в этом случае следует ожидать устойчивости. Если  $\lim_{\tau, h \rightarrow 0} \tau/h^2 = \infty$ , то

$$\lim_{\tau, h \rightarrow 0} \left( \left| \lambda \left( \frac{\pi}{2} \right) \right| - 1 \right) / \tau = \infty$$

и условие (4) не выполняется ни при каком  $C$ ; тогда сеточная аппроксимация неустойчива.

Эта аппроксимация практически не употребляется вследствие жесткого ограничения на шаг  $\tau = O(h^2)$ , необходимого для устойчивости, и сильного роста возмущения решения (как  $\exp\{a^2 A^2 n \tau / 2\}$ ).

Пример 2. Разностная аппроксимация:

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a \frac{u_m^n - u_{m-1}^n}{h} = 0 \quad \text{при } a > 0. \quad (10)$$

Аналогично (9) получаем

$$\lambda(\varphi) = 1 - a\tau/h + a\tau/h \exp\{-i\varphi\}.$$

Если  $0 \leq a\tau/h \leq 1$ , то (см. рис. 10.2.2)  $|\lambda(\varphi)| \leq 1 - a\tau/h + a\tau/h = 1$  и следует ожидать устойчивости. Если  $a\tau/h = \kappa > 1$  при  $\tau, h \rightarrow 0$ , то  $\lim_{\tau, h \rightarrow 0} \lambda(\pi) = 1 - 2\kappa < -1$  и аппроксимация неустойчива вследствие СПУ.

Аналогично показывается, что следует ожидать устойчивости схемы

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a \frac{u_{m+1}^n - u_m^n}{h} = 0 \quad (11)$$

при  $a < 0$  и  $|a|\tau/h \leq 1$ .

Пример 3. Разностная аппроксимация:

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a \frac{u_{m+1}^n - u_{m-1}^n}{2h} - \frac{h^2}{2\tau} \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = 0, \quad (12)$$

$$\lambda(\varphi) = \left( \frac{1}{2} - \frac{a\tau}{2h} \right) e^{i\varphi} + \left( \frac{1}{2} + \frac{a\tau}{2h} \right) e^{-i\varphi} = \cos \varphi - \frac{a\tau i}{h} \sin \varphi.$$

При  $|a\tau/h| \leq 1$  имеем (рис. 10.2.3)

$$|\lambda(\varphi)| \leq \left( \frac{1}{2} - \frac{a\tau}{2h} \right) + \left( \frac{1}{2} + \frac{a\tau}{2h} \right) = 1$$

и следует ожидать устойчивости. Если же  $|a\tau/h| = \kappa > 1$  при  $\tau, h \rightarrow 0$ , то  $\lim_{\tau, h \rightarrow 0} \left| \lambda \left( \frac{\pi}{2} \right) \right| = |a\tau/h| = \kappa > 1$  и аппроксимация неустойчива.

Заметим, что эту аппроксимацию можно рассматривать, как (8) с добавленной в нее для устойчивости «вязкостью»:

$$\frac{h^2}{2\tau} \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} \sim \frac{h^2}{2\tau} u_{xx},$$

Пример 4. Аппроксимация «тренога». Приведенные выше аппроксимации имеют первый порядок по совокупности  $\tau$  и  $h$ . Построим аппроксимацию второго порядка; для простоты будем отправляться от аппроксимации (8). Подставляя разложение Тейлора  $u(x, t)$  в точке  $(mh, n\tau)$ , имеем

$$(L_h [u]_h)_{(m, n)} = \tau u_{tt}(mh, n\tau)/2 + O(h^2) + O(\tau^2).$$

Из дифференциального уравнения (1) получаем  $u_{tt} = a^2 u_{xx}$ . Приближим  $a^2 u_{xx}$  выражением  $a^2 (\delta_x^2 u_h)_{m, n}$ . Тогда соответствующая аппроксимация  $\bar{L}_h^{(1)} u_h = 0$  примет вид

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a \frac{u_{m+1}^n - u_{m-1}^n}{2h} - \frac{a^2 \tau}{2} \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = 0. \quad (13)$$

Так же как и ранее, получаем

$$\lambda(\varphi) = 1 - i(a\tau/h) \sin \varphi + (a^2 \tau^2 / h^2) (\cos \varphi - 1).$$

Если мы положим  $\lambda(\varphi) = x + iy$ , то (см. рис. 10.2.4) можно написать

$$\frac{(x - (1 - (a\tau/h)^2))^2}{(a\tau/h)^4} + \frac{y^2}{(a\tau/h)^2} = 1,$$

т. е. точки  $\lambda(\varphi)$  лежат на некотором эллипсе в комплексной плоскости, расположенном симметрично относительно оси  $y = 0$ .

Поскольку  $\lambda(\pi) = 1 - 2a^2 \tau^2 / h^2 < -1$  при  $|a\tau/h| > 1$ , то в случае  $|a\tau/h| = \text{const} > 1$  спектральное условие устойчивости не выполнено.

При  $|a\tau/h| \leq 1$  имеем  $(a^2 \tau^2 / h^2)^2 \leq a^2 \tau^2 / h^2$ ; поэтому можно написать цепочку соотношений

$$\begin{aligned} |\lambda(\varphi)|^2 &= \left(1 + \frac{a^2 \tau^2}{h^2} (\cos \varphi - 1)\right)^2 + \frac{a^2 \tau^2}{h^2} \sin^2 \varphi = \\ &= 1 + \frac{a^2 \tau^2}{h^2} (2 \cos \varphi - 2 + \sin^2 \varphi) + \left(\frac{a^2 \tau^2}{h^2}\right)^2 (\cos \varphi - 1)^2 \leq \\ &\leq 1 + \frac{a^2 \tau^2}{h^2} (2 \cos \varphi - 2 + \sin^2 \varphi) + \frac{a^2 \tau^2}{h^2} (\cos \varphi - 1)^2 = \\ &= 1 + \frac{a^2 \tau^2}{h^2} (2 \cos \varphi - 2 + \sin^2 \varphi + \cos^2 \varphi - 2 \cos \varphi + 1) \equiv 1. \end{aligned}$$

Таким образом, при  $|a\tau/h| \leq 1$  выполнено спектральное условие устойчивости.

Сделаем ряд общих замечаний.

1. Если  $a\tau/h = \kappa$  при  $\tau, h \rightarrow 0$ , то для всех выписанных аппроксимаций выполнялось соотношение

$$\lambda(\varphi) = \exp\{-i\kappa\varphi\} + O(\varphi^{r+1}),$$

где  $r$  — порядок аппроксимации по  $\tau$  и  $h$ . Можно показать, что это условие является необходимым для того, чтобы имела место аппроксимация  $r$ -го порядка точности.



2. Если  $|\alpha\tau/h| = 1$  и  $a > 0$ , то все рассмотренные аппроксимации имеют вид

$$u_m^{n+1} - u_{m-1}^n = 0.$$

Поскольку решение задачи есть  $u = u_0(x - at)$ , то в этих случаях они абсолютно точные; при этом  $\lambda(\varphi) = \exp\{-i(\alpha\tau/h)\varphi\} = e^{-i\varphi}$ .

3. В тех случаях, когда аппроксимации (10) — (13) некорректны вследствие СПУ, их можно было бы также забраковать при помощи теоремы Куранта.

4. В остальных случаях, т. е. для (10) при  $0 \leq \alpha\tau/h \leq 1$ , для (11) при  $-1 \leq \alpha\tau/h \leq 0$  и для (12), (13) при  $|\alpha\tau/h| \leq 1$ , их устойчивость следует из решения задачи 2.

Задача 3. Доказать, что аппроксимации (10) при  $0 \leq \alpha\tau/h \leq 1$ , (11) при  $-1 \leq \alpha\tau/h \leq 0$  и (12) при  $|\alpha\tau/h| \leq 1$  обладают следующим свойством: для их решений выполняется неравенство

$$\sup_m |u_m^{n+1}| \leq \sup_m |u_m^n|, \quad (14)$$

если  $\sup_m |u_m^n|$  существует.

Задача 4. Доказать, что при использовании этих аппроксимаций при любом  $n$  решение сеточной задачи монотонно по  $m$ , если монотонно  $u_m^0$ .

Это свойство монотонности делает схемы, удовлетворяющие условию (14), весьма удобными при интегрировании разрывных решений. Если при интегрировании разрывных решений употреблять аппроксимации, не обладающие таким свойством, то в разностном решении появляются паразитические волны, имитирующие разрывы и иногда мешающие пониманию истинной картины явления.

Сделаем ряд замечаний по поводу практического употребления этих аппроксимаций. Исторически первой была аппроксимация (12); она обладает следующим недостатком.

Если мы подставим в  $L_h[u]_h$  разложение Тейлора  $u(x, t)$  в точке  $(mh, n\tau)$ , то получим

$$\begin{aligned} (L_h[u]_h)|_{(m,n)} &= \tau u_{tt}(mh, n\tau)/2 - (h^2/2\tau)u_{xx}(mh, n\tau) + \dots = \\ &= (\tau a^2/2 - h^2/2\tau)u_{xx}(mh, n\tau). \end{aligned}$$

Условия  $\tau, h \rightarrow 0$  еще недостаточно, чтобы эта схема аппроксимировала исходное уравнение. Необходимо еще добавить условие  $h^2/\tau \rightarrow 0$ . Требование корректности  $|\alpha\tau/h| \leq 1$  и ряд других условий при решении сложных задач приводят к тому, что отношение  $h^2/\tau$  часто остается большим при малых  $\tau$  и  $h$ . Качественно это ухудшение аппроксимации проявляется в проявлении «вязкости» аппроксимации: все неровности решения, включая разрывы, сильно выглаживаются.

Аппроксимации (10), (11) неудобны тем, что при переменном  $a(x, t)$  требуют использования различных узлов в зависи-

мости от знака  $a$ . Еще более осложняется дело в случае систем уравнений. Тем не менее разработаны практически приемлемые варианты этой схемы (схема Годунова), широко применяемые для решения нелинейных систем.

Аппроксимация (13) эффективна в случае гладких решений, но при наличии разрывов дает большое число паразитических волн. Поэтому она подвергается модификации в областях больших градиентов решения.

После построения аппроксимаций (10)—(12) в применении к гиперболическим задачам долгое время казалось, что применение схем высокого порядка точности является неоправданным; такой вывод объяснялся тем, что все известные к тому времени аппроксимации второго порядка превращали разрывные решения в решения с большим числом паразитических волн.

Впоследствии теоретический анализ вопроса о качественных свойствах решений аппроксимаций при наличии разрывов показал, что этим свойством обладают все линейные аппроксимации второго порядка точности. Однако теоретические исследования предсказали, а практика подтвердила наличие аппроксимаций третьего порядка точности с удовлетворительным поведением решений разностных уравнений при наличии разрывов.

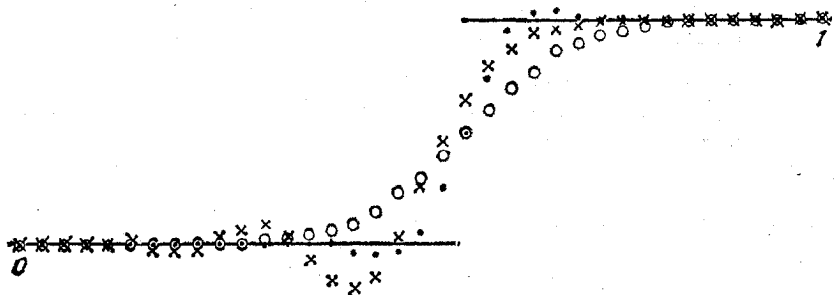


Рис. 10.2.5

На рис. 10.2.5 изображено поведение решений различных разностных аппроксимаций уравнения

$$u_t + u_x = 0 \quad \text{при} \quad u_0(x) = \begin{cases} 0, & \text{если } x < 0, \\ 1, & \text{если } x \geq 1; \end{cases}$$

сплошной линией обозначено точное решение,  $\circ$ ,  $\times$ ,  $\cdot$  — полученные расчетные значения по схемам (10), (13) и третьего порядка соответственно.

Рассмотрим пример применения СПУ в случае решения системы уравнений.

Пусть решается задача Коши для системы уравнений

$$u_t + av_x = 0, \quad v_t + bu_x = 0, \quad ab > 0;$$

условие  $ab > 0$  обеспечивает гиперболичность системы.

Рассмотрим трехслойную разностную схему, аппроксимирующую исходную задачу с погрешностью  $O(h^2 + \tau^2)$ :

$$\begin{aligned} \frac{u_m^{n+1} - u_m^{n-1}}{2\tau} + a \frac{u_{m+1}^n - u_{m-1}^n}{2h} &= 0, \\ \frac{v_m^{n+1} - v_m^{n-1}}{2\tau} + b \frac{v_{m+1}^n - v_{m-1}^n}{2h} &= 0. \end{aligned} \quad (15)$$

Ищем частное решение системы в виде

$$u_m^n = c_1 \lambda^n e^{im\varphi},$$

$$v_m^n = c_2 \lambda^n e^{im\varphi}.$$

После подстановки решения в (15) получим

$$\lambda^{n-1} e^{im\varphi} \left( c_1 \frac{\lambda^2 - 1}{\tau} + 2ac_2 \lambda \frac{i \sin \varphi}{h} \right) = 0,$$

$$\lambda^{n-1} e^{im\varphi} \left( c_2 \frac{\lambda^2 - 1}{\tau} + 2bc_1 \lambda \frac{i \sin \varphi}{h} \right) = 0$$

и, следовательно,

$$c_1 \frac{\lambda^2 - 1}{\tau} + 2c_2 a \lambda \frac{i \sin \varphi}{h} = 0,$$

$$2c_1 b \lambda \frac{i \sin \varphi}{h} + c_2 \frac{\lambda^2 - 1}{\tau} = 0.$$

Эта система линейных уравнений относительно коэффициентов  $c_1, c_2$  имеет ненулевое решение, если определитель системы равен нулю. Получаем уравнение, связывающее  $\lambda$  и  $\varphi$ :

$$\begin{vmatrix} \frac{\lambda^2 - 1}{\tau} & \frac{\lambda a \cdot 2i \sin \varphi}{h} \\ \frac{\lambda b \cdot 2i \sin \varphi}{h} & \frac{\lambda^2 - 1}{\tau} \end{vmatrix} = 0.$$

Отсюда

$$\left( \frac{\lambda^2 - 1}{\tau} \right)^2 + \lambda^2 \frac{4ab \sin^2 \varphi}{h^2} = 0$$

или

$$\lambda^2 \pm 2i \sqrt{ab} \frac{\tau}{h} \sin \varphi \lambda - 1 = 0.$$

Таким образом, окончательно

$$\lambda = \mp i \sqrt{ab} \frac{\tau}{h} \sin \varphi \pm \sqrt{-ab \frac{\tau^2}{h^2} \sin^2 \varphi + 1}.$$

Если  $ab\tau^2/h^2 \leq 1$ , то подкоренное выражение неотрицательно и

$$|\lambda| = ab \frac{\tau^2}{h^2} \sin^2 \varphi + \left( -ab \frac{\tau^2}{h^2} \sin^2 \varphi + 1 \right) = 1.$$

Таким образом, при  $\sqrt{ab} \tau/h \leq 1$  СПУ выполнен.

Задача 5. Показать, что при  $\sqrt{ab} \frac{\tau}{h} = \kappa = \text{const} > 1$  СПУ не выполнен.

Задача 6. С помощью теоремы об областях зависимости показать, что при  $\sqrt{ab} \frac{\tau}{h} = \kappa = \text{const} > 1$  решение сеточной задачи не обязательно сходится к решению дифференциальной задачи.

Пусть в области  $0 \leq m \leq M$  ищется функция  $u_m^n$ , удовлетворяющая (3) и некоторым граничным условиям

$$L_h^1 u_h = 0 \quad (16)$$

относительно значений  $u_m^n$  при  $m$ , близких к нулю, и

$$L_h^2 u_h = 0 \quad (17)$$

относительно значений  $u_m^n$  при  $m$ , близких к  $M$ .

Примечание. Число уравнений относительно значений  $u_m^n$  на каждом слое берется равным числу неизвестных  $M + 1$ , поэтому некоторые из уравнений (3) отбрасываются.

Для краевых сеточных задач с постоянными коэффициентами также имеется СПУ, часто позволяющий довольно просто отбраковывать непригодные для счета разностные схемы. Можно показать, что разностная схема, не удовлетворяющая этому СПУ, неустойчива.

Этот СПУ заключается в следующем.

1. Должен быть выполнен спектральный признак устойчивости задачи Коши (отличный от сформулированного ранее). Ищутся всевозможные частные решения (3) вида

$$u_m^n = \lambda^n \varphi_m, \quad (18)$$

$$\|\varphi\|_0 \equiv \sup_{-\infty < m < \infty} |\varphi_m| < \infty. \quad (19)$$

СПУ задачи Коши состоит в том, что при заданном законе стремления  $\tau, h$  к нулю

$$\overline{\lim}_{\|\varphi\|_0 < \infty} \sup |\lambda| \leq 1, \quad (20)$$

обозначение  $\sup_{\|\varphi\|_0 < \infty}$  введено с целью подчеркнуть еще раз, что верхняя грань берется по множеству всех решений (18), удовлетворяющих условию (19).

2. Должно быть выполнено условие спектральной устойчивости «левой» краевой задачи, состоящее в следующем.

Рассмотрим «левую» краевую задачу

$$L_h u_h|_{(m, n)} = 0 \quad \text{при} \quad 0 \leq m < \infty, \quad L_h^1 u_h = 0$$

(с учетом примечания) и находим ее частные решения вида

$$u_m^n = \lambda^n \varphi_m, \quad \|\varphi\|_+ = \sup_{0 \leq m < \infty} |\varphi_m| < \infty.$$

СПУ левой краевой задачи имеет вид, аналогичный (20):

$$\overline{\lim} \sup_{\|\varphi\|_+ < \infty} |\lambda| \leq 1. \quad (21')$$

3. Точно так же рассматривается «правая» краевая задача

$$L_h u_h = 0 \text{ при } -\infty < m \leq M, \quad L_h^2 u_h = 0.$$

Ищутся ее частные решения вида  $u_m^n = \lambda^n \varphi_m$  такие, что  $\|\varphi\|_- = \sup_{-\infty < m \leq M} |\varphi_m| < \infty$ . СПУ «правой» краевой задачи имеет вид

$$\overline{\lim} \sup_{\|\varphi\|_- < \infty} |\lambda| \leq 1. \quad (21'')$$

Заметим, что замена независимой переменной  $m' = M - m$  переводит «левую» краевую задачу в «правую» и наоборот и соответственно преобразуются друг в друга СПУ «левой» и «правой» задач.

Пример. Рассмотрим сеточную краевую задачу

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = 0, \quad 0 < m < M, \quad Mh = 1, \\ \frac{u_0^n - 4u_1^n + 3u_2^n}{2h} = 0, \quad u_M^n = 0. \quad (22)$$

Эта задача аппроксимирует дифференциальное уравнение

$$u_t - u_{xx} = 0 \text{ при } 0 < x < 1$$

с краевыми условиями  $u_x|_{(0, t)} = 0, u(1, t) = 0$ .

Исследуем спектральную устойчивость сеточной задачи в предположении, что  $\tau/h^2 = \kappa = \text{const}$  при стремлении  $\tau, h$  к нулю.

1. СПУ задачи Коши. Ищем частные решения вида  $u_m^n = \lambda^n \varphi_m$ . После подстановки в (22) и сокращения на  $\lambda^n$  получаем уравнение

$$\frac{\lambda - 1}{\tau} \varphi_m - \frac{\varphi_{m+1} - 2\varphi_m + \varphi_{m-1}}{h^2} = 0$$

или, что то же самое,

$$\varphi_{m+1} - \left(2 + (\lambda - 1) \frac{h^2}{\tau}\right) \varphi_m + \varphi_{m-1} = 0. \quad (23)$$

Функция  $\varphi_m$  является решением одномерного конечно-разностного уравнения с постоянными коэффициентами (23), поэтому  $\varphi_m = c_1 \mu_1^m + c_2 \mu_2^m$ , где  $\mu_{1,2}$  — корни характеристического уравнения

$$\mu^2 - \left(2 + (\lambda - 1) \frac{h^2}{\tau}\right) \mu + 1 = 0. \quad (24)$$

Согласно теореме Виета  $\mu_1\mu_2 = 1$ , поэтому

$$\varphi_m = c_1\mu_1^m + c_2\mu_1^{-m}.$$

Для ограниченности  $\|\varphi\|_0$  при всех  $c_1, c_2$  необходимо, чтобы  $|\mu_1| = 1$ . Полагая  $\mu_1 = e^{i\varphi}$ , из (24) получаем

$$\lambda = 1 - 4 \frac{\tau}{h^2} \sin^2 \frac{\varphi}{2}.$$

Если  $\kappa = \tau/h^2 \leq 1/2$ , то  $\sup |\lambda| = 1$  и СПУ задачи Коши выполнен. В противном случае он не выполнен.

2. Спектральную неустойчивость «левой» краевой задачи мы докажем, «угадав» последовательность частных решений

$$u_m^n = \lambda^n \varphi_m \quad \text{с} \quad \|\varphi_m\|_+ < \infty$$

таких, что  $\lim_{\tau, h \rightarrow 0} |\lambda| > 1$ . Ищем решение «левой» краевой задачи в виде  $u_m^n = \lambda^n \mu^m$ . Подставляя  $u_m^n$  в (22) и сокращая на  $\lambda^n \mu^m$ , получим

$$\frac{\lambda - 1}{\tau} - \frac{\mu - 2 + \mu^{-1}}{h^2} = 0. \quad (25)$$

Подставляя  $u_m^n$  в левое граничное условие (22), получим уравнение

$$1 - 4\mu + 3\mu^2 = 0.$$

Его корни  $\mu_1 = 1$ ,  $\mu_2 = \frac{1}{3}$ ; корню  $\mu_2 = \frac{1}{3}$ , согласно (25), соответствует  $\lambda = 1 + \frac{4}{3} \frac{\tau}{h^2} = 1 + \frac{4}{3} \kappa$ . Для этого частного решения

$$u_m^n = \lambda^n \varphi_m = \left(1 + \frac{4}{3} \kappa\right)^n \left(\frac{1}{3}\right)^m \quad \text{имеем соотношения}$$

$$\|\varphi_m\|_+ = 1 \quad \text{и} \quad |\lambda| = 1 + \frac{4}{3} \kappa > 1.$$

Следовательно, СПУ этой задачи не выполнен и следует взять другую аппроксимацию граничного условия  $u_x = 0$  в точке 0.

**Задача 7.** Доказать, что аппроксимация граничного условия

$$\frac{-u_2^n + 4u_1^n - 3u_0^n}{2h} = 0$$

соответствует «левой» краевой задаче, удовлетворяющей СПУ.

3. Рассмотрим «правую» краевую задачу. Как и в случае задачи Коши, получаем совокупность частных решений вида

$$u_m^n = \lambda^n \varphi_m, \quad \varphi_m = C_1 \mu_1^m + C_2 \mu_2^m,$$

где  $\lambda$  и  $\mu$  связаны соотношением (25), причем, как и там,  $\mu_1\mu_2 = 1$ . Удобнее представлять функцию  $\varphi_m$  в виде

$$\varphi_m = C_1 \mu_1^{m-M} + C_2 \mu_2^{m-M}.$$

Из правого граничного условия (22) получаем  $\varphi_m = C_1 + C_2 = 0$ , поэтому  $\varphi_m = C_1(\mu_1^{m-M} - \mu_2^{m-M})$ . Если  $|\mu_1| > 1$ , то  $|\mu_2| < 1$ , и наоборот; в обоих этих случаях  $|\varphi_m| \rightarrow \infty$  при  $m \rightarrow -\infty$ . Поэтому нас интересуют лишь решения с  $|\mu_1| = |\mu_2| = 1$ . Тогда  $\mu_1 = e^{i\varphi}$ ,  $\mu_2 = e^{-i\varphi}$  и

$$\lambda = 1 - \frac{4\tau}{h^2} \sin^2 \frac{\varphi}{2}.$$

Как и в случае задачи Коши, получаем, что СПУ выполнен при  $\kappa = \tau/h^2 \leq 1/2$ .

Рассмотренные выше разностные схемы относятся к классу *явных*; значения решения на верхнем слое вычисляются по значениям решения на нижних слоях по формулам вида (ср. с (7)):

$$u_m^{n+1} = \sum_{|j| \leq k} a_{m+j}^n u_{m+j}^n;$$

$k$  ограничено при  $h \rightarrow 0$ .

Если в разностное уравнение входят не менее чем два значения решения на верхнем слое, то такую схему относят к классу *неявных схем*. В этом случае значения сеточного решения на верхнем слое находят, решая некоторую систему уравнений

$$Cu^{n+1} = F(u^n, f). \quad (26)$$

В случае, когда матрица  $C$  треугольная (а при решении систем уравнений в частных производных блочно-треугольная) разностную схему называют *полуявной*. Таким образом, полуявные схемы составляют подкласс неявных ( $C$  треугольная матрица), а явные — подкласс полуявных ( $C$  единичная матрица) разностных схем.

В случае неявных (в частности полуявных) схем возникает вопрос о построении алгоритма решения системы уравнений (26), устойчивого к влиянию вычислительной погрешности.

Рассмотрим простейшую полуявную схему для уравнения  $u_t + au_x = 0$ :

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a \frac{u_m^{n+1} - u_{m-1}^{n+1}}{h} = 0. \quad (27)$$

Ищем частные решения вида  $u_m^n = \lambda^n e^{im\varphi}$ ; после подстановки их в (27) получим

$$\left( \frac{\lambda - 1}{\tau} + a \frac{\lambda(1 - e^{-i\varphi})}{h} \right) \lambda^n e^{im\varphi} = 0.$$

Таким образом,  $\lambda(\varphi) = \left( 1 + a \frac{\tau}{h} - a \frac{\tau}{h} e^{-i\varphi} \right)^{-1}$ .

Пусть  $\sigma = \sup_{0 \leq \varphi \leq 2\pi} |\lambda(\varphi)|$ .

Задача 8. Показать, что

1)  $\sigma \leq 1$  при  $a \geq 0$ ;

2)  $\sigma \leq 1$  при  $a \frac{\tau}{h} \leq -1$ ;

3)  $\sigma = \frac{1}{|2\kappa - 1|} > 1$  при  $a \frac{\tau}{h} = \kappa = \text{const}$ ,  $-1 < \kappa < 0$ .

Таким образом, СПУ не выполнен при  $-1 < \kappa < 0$ .

В реальных вычислениях всегда участвует конечное число значений  $u_m^{n+1}$ , соответствующих  $(n+1)$ -му слою. Если мы выпишем уравнение (27) при

$m = 1, \dots, M$ , то получим систему из  $M$  уравнений относительно  $(M + 1)$ -го неизвестных  $u_0^{n+1}, \dots, u_M^{n+1}$ .

Рассмотрим случай, когда решение ищется в прямоугольнике

$$0 \leq x \leq X = Mh, \quad 0 \leq t \leq T.$$

Функция  $u(x, t) = \psi(x - at)$  является решением уравнения  $u_t + au_x = 0$ . Если  $a = 0$ , то решение однозначно определяется заданием начальных условий  $u(x, 0) = u_0(x)$ ; при  $a > 0$  для нахождения решения достаточно задать  $u(x, 0)$  и  $u(0, t)$ ; при  $a < 0$  — задать  $u(x, 0)$  и  $u(X, t)$ .

В случае  $a > 0$  из граничных условий нам известно также  $u_0^{n+1} = u(0, (n + 1)\tau)$  и, таким образом, число уравнений равно числу неизвестных.

Уравнение (27) можно представить в виде рекуррентной формулы для определения  $u_m^{n+1}$ :

$$u_m^{n+1} = \frac{\kappa}{1 + \kappa} u_{m-1}^{n+1} + \frac{u_m^n}{1 + \kappa}, \quad \kappa = \frac{a\tau}{h}, \quad m = 1, \dots, M. \quad (28)$$

Если  $u_{m-1}^{n+1}$  содержит некоторую погрешность  $\delta_{m-1}^{n+1}$ , то вследствие (28) она порождает погрешность  $\delta_m^{n+1}$  в  $u_m^{n+1}$ , равную  $\frac{\kappa}{1 + \kappa} \delta_{m-1}^{n+1}$ . При  $a \geq 0$

имеем  $\kappa \geq 0$  и  $0 \leq \frac{\kappa}{1 + \kappa} < 1$ . Таким образом,  $|\delta_m^{n+1}| < |\delta_{m-1}^{n+1}|$ , т. е. погрешности в значениях  $u_m^{n+1}$  убывают и можно ожидать, что при вычислениях по формулам (28) накопление погрешности не приведет к нежелательным последствиям.

В случае  $a < 0$  известно значение  $u_M^{n+1} = u(X, (n + 1)\tau)$  и вычисления будем вести по рекуррентной формуле, вытекающей из (27):

$$u_{m-1}^{n+1} = \left(1 + \frac{1}{\kappa}\right) u_m^{n+1} - \frac{1}{\kappa} u_m^n, \quad m = M, \dots, 1. \quad (29)$$

Если  $\kappa \leq -1$ , то  $0 \leq 1 + \frac{1}{\kappa} < 1$  и погрешность  $\delta_m^{n+1}$  в значении  $u_{m-1}^{n+1}$  породит погрешность  $\delta_{m-1}^{n+1} = \left(1 + \frac{1}{\kappa}\right) \delta_m^{n+1}$  такую, что  $|\delta_{m-1}^{n+1}| < |\delta_m^{n+1}|$ . Таким образом, следует ожидать, что накопление вычислительной погрешности не будет существенным.

**З а м е ч а н и е.** Для расчетной формулы (28) соотношение  $|\delta_m^{n+1}| \leq |\delta_{m-1}^{n+1}|$  выполнено и при  $-1/2 \leq \kappa < 0$ , а для формулы (29) соотношение  $|\delta_{m-1}^{n+1}| \leq |\delta_m^{n+1}|$  выполнено и при  $-1 < \kappa \leq -1/2$ . В этих случаях, однако, не выполнен спектральный признак устойчивости  $\sigma < 1$ .

Предположим, что в области  $t \leq T$  нас интересует решение задачи Коши при начальном условии  $u(x, 0) = u_0(x)$ , определенном при  $0 \leq x \leq X_0$  и  $a > 0$ . Решение дифференциальной задачи определено в полосе  $at \leq x \leq X_0 + at$ .

Рассмотрим краевую задачу в прямоугольнике  $0 \leq x \leq X = X_0 + aT$ ,  $0 \leq t \leq T$ , задав произвольные гладкие функции  $u(0, t)$  при  $0 \leq t \leq T$  и  $u(x, 0)$  при  $X_0 \leq x \leq X$ , удовлетворяющие лишь условиям согласования  $u(0, 0) = u_0(0)$  и  $u_0(X_0) = u(X_0, 0)$ .

Применим для решения этой задачи полуявную схему (27), проводя вычисления на верхнем слое по рекуррентным формулам (28). Можно показать, что при гладкой функции  $u_0(x)$  в области  $at \leq x \leq X_0 + at$ ,  $0 \leq t \leq T$  полученное сеточное решение будет близко к решению задачи Коши.



При решении задачи Коши и краевых задач, особенно в случае систем уравнений, часто используется неявная схема

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a \frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2h} = 0 \quad (30)$$

с порядком аппроксимации  $O(\tau + h^2)$ . Иногда эта схема используется как вспомогательная для получения решения на полуцелом слое, т.е. находятся  $u_m^{n+1/2}$  из соотношений

$$\frac{u_m^{n+1/2} - u_m^n}{\tau/2} + a \frac{u_{m+1}^{n+1/2} - u_{m-1}^{n+1/2}}{2h} = 0.$$

далее используется явная формула

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a \frac{u_{m+1}^{n+1/2} - u_{m-1}^{n+1/2}}{2h} = 0.$$

Рассмотрим случай той же самой задачи Коши. Если мы выпишем уравнение (30) при  $m = 1, \dots, M-1$ , то получим систему  $(M-1)$ -го уравнения с  $(M+1)$ -м неизвестным. На каждом слое нам не хватает двух уравнений. Значения  $u_0^n$  зададим как и ранее, используя граничные значения, а  $u_M^n$  — по произволу. Систему уравнений (30) относительно значений  $u_1^{n+1}, \dots, u_{M-1}^{n+1}$  будем решать методом прогонки.

Можно показать, что при гладкой функции  $u_0(x)$  и любом  $\varepsilon > 0$  решение сеточной задачи сходится к решению дифференциальной в области, определяемой неравенствами

$$at + \varepsilon \leq x \leq X_0 + at - \varepsilon, \quad 0 \leq t \leq T.$$

Для улучшения сходимости вместо задания  $u_M^{n+1}$  часто целесообразнее взять так называемое «мягкое» граничное условие  $u_M^{n+1} - u_{M-1}^{n+1} = 0$ .

### § 3. Принцип замороженных коэффициентов

Часто не удается произвести теоретическое исследование корректности разностной задачи и доказать сходимость ее решения к решению дифференциальной задачи. В некоторых случаях на данном этапе развития математической теории такое исследование в принципе возможно, но требует от исследователя достаточно высокой квалификации и больших затрат времени.

В такой ситуации иногда ограничиваются исследованием устойчивости схем на основе описываемого ниже принципа замороженных коэффициентов и последующей экспериментальной проверкой полученных выводов путем расчета тестовых задач, по возможности с известным решением.

Принцип замороженных коэффициентов (ПЗК) заключается в следующем.

1. Для разностной схемы пишется уравнение в вариациях (т.е. уравнение, которому удовлетворяет разность двух бесконечно близких решений). Это уравнение является линейным и в случае линейных задач совпадает с исходным уравнением.

2. Фиксируется некоторая точка  $P$  области  $G$ , в которой ищется решение, замораживаются коэффициенты этого уравнения, т. е. все значения коэффициентов уравнения в вариациях берутся равными их значениям в этой точке. Если задача нелинейная, то коэффициенты уравнения в вариациях зависят от неизвестной функции и все значения сеточного решения, входящие в это уравнение, берутся равными их значениям в той же точке  $P$ .

3. Получившаяся сеточная задача  $L_h^P(\delta u_h) = 0$  исследуется на устойчивость методами, которые применяются для исследования устойчивости сеточных задач с постоянными коэффициентами.

Предположим, что сеточная задача устойчива при выполнении условия

$$\varphi(h, P) \geq 0 \quad (1)$$

на шаги сетки; это условие, естественно, может зависеть от выбора точки  $P$ .

4. За условие устойчивости принимают некоторое условие  $\varphi(h) \geq 0$ , из выполнения которого следует выполнение условия  $\varphi(h, P) \geq 0$  для всех точек  $P \in G$ . Часто, особенно в случае нелинейных задач, условие устойчивости  $\varphi(h) \geq 0$  выбирается с некоторым «запасом устойчивости».

Рассмотрим пример применения принципа замороженных коэффициентов.

Пусть решается задача Коши для уравнения

$$u_t + (\varphi(x, t, u))_x - \psi(x, t, u) = 0 \quad (2)$$

при начальном условии  $u(x, 0) = u_0(x)$ . Через  $u_m^n$  будем обозначать приближение к значению решения в точке  $(x, t) = (mh, n\tau)$ . Аппроксимируем уравнение (2) разностной схемой

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{\varphi(mh, n\tau, u_m^n) - \varphi((m-1)h, n\tau, u_{m-1}^n)}{h} - \psi(mh, n\tau, u_m^n) = 0. \quad (3)$$

Пусть  $v_m^n = u_m^n + \delta_m^n$  — другое решение сеточной задачи (3), т. е.  $\delta_m^n$  — разность между двумя решениями задачи (3). Имеем равенство (получаемое при подстановке  $v_m^n$  в (3)):

$$\frac{(u_m^{n+1} + \delta_m^{n+1}) - (u_m^n + \delta_m^n)}{\tau} - \psi(mh, n\tau, u_m^n + \delta_m^n) + \frac{\varphi(mh, n\tau, u_m^n + \delta_m^n) - \varphi((m-1)h, n\tau, u_{m-1}^n + \delta_{m-1}^n)}{h} = 0.$$

Вычитая из этого равенства соотношение (3), получим

$$\frac{\delta_m^{n+1} - \delta_m^n}{\tau} + \frac{\varphi(mh, n\tau, u_m^n + \delta_m^n) - \varphi(mh, n\tau, u_m^n)}{h} -$$

$$- \frac{\varphi((m-1)h, n\tau, u_{m-1}^n + \delta_{m-1}^n) - \varphi((m-1)h, n\tau, u_{m-1}^n)}{h} -$$

$$- (\psi(mh, n\tau, u_m^n + \delta_m^n) - \psi(mh, n\tau, u_m^n)) = 0. \quad (4)$$

С точностью до членов второго порядка малости по величинам  $\delta_j^n$  выполнены приближенные равенства

$$\varphi(mh, n\tau, u_m^n + \delta_m^n) - \varphi(mh, n\tau, u_m^n) \approx \varphi_u(mh, n\tau, u_m^n) \delta_m^n,$$

$$\varphi((m-1)h, n\tau, u_{m-1}^n + \delta_{m-1}^n) - \varphi((m-1)h, n\tau, u_{m-1}^n) \approx$$

$$\approx \varphi_u((m-1)h, n\tau, u_{m-1}^n) \delta_{m-1}^n,$$

$$\psi(mh, n\tau, u_m^n + \delta_m^n) - \psi(mh, n\tau, u_m^n) \approx \psi_u(mh, n\tau, u_m^n) \delta_m^n.$$

Таким образом, бесконечно малое приращение решения  $\delta_m^n$ , его так называемая *вариация*, удовлетворяет уравнению

$$\frac{\delta_m^{n+1} - \delta_m^n}{\tau} + \frac{\varphi_u(mh, n\tau, u_m^n) \delta_m^n - \varphi_u((m-1)h, n\tau, u_{m-1}^n) \delta_{m-1}^n}{h} -$$

$$- \psi_u(mh, n\tau, u_m^n) \delta_m^n = 0;$$

это уравнение называется *уравнением в вариациях* для (4).

Заморозим коэффициенты, взяв все значения  $\varphi_u$  и  $\psi_u$  равными их значениям в некоторой точке

$$\frac{\delta_m^{n+1} - \delta_m^n}{\tau} + a \frac{\delta_m^n - \delta_{m-1}^n}{h} - b \delta_m^n = 0. \quad (5)$$

Отсюда имеем равенство

$$\delta_m^{n+1} = \delta_m^n \left( 1 - \frac{a\tau}{h} + b\tau \right) + \delta_{m-1}^n a \frac{\tau}{h}$$

и затем оценку

$$|\delta_m^{n+1}| \leq \max_m |\delta_m^n| \left( \left| 1 - a \frac{\tau}{h} \right| + |b\tau| + \left| a \frac{\tau}{h} \right| \right). \quad (6)$$

Поскольку (6) выполнено при всех  $m$ , то

$$\max_m |\delta_m^{n+1}| \leq \max_m |\delta_m^n| \left( \left| 1 - a \frac{\tau}{h} \right| + |b\tau| + \left| a \frac{\tau}{h} \right| \right).$$



Типичные значения сужения области устойчивости для различных схем: в 1 раз (сужение не производится); в 1,15 раз; в 1,3 раза; в 1,5 раза; в 2 раза.

Заранее неизвестно, в какой области находятся значения решения сеточной задачи  $u_m^n$ , поэтому до реального решения сеточной задачи нельзя выбрать шаг  $\tau$  таким, чтобы при всех  $m, n$  выполнялось условие устойчивости (8). В частности, в связи с этим при решении нестационарных задач шаг по времени часто берется переменным: ищутся приближения  $u_m^n$  к значениям решения в точках  $(mh, t_n)$ ; шаг  $\tau_n = t_{n+1} - t_n$  зависит от  $n$ . При каждом  $n$  вычисляются

$$A_n^1 = \inf_m a(mh, t_n, u_m^n),$$

$$A_n^2 = \sup_m a(mh, t_n, u_m^n).$$

Если оказалось, что  $A_n^1 < 0$ , то счет по этой схеме прекращается, поскольку ни при каком  $\tau_n > 0$  не удастся добиться удовлетворения условия  $0 \leq a(mh, t_n, u_m^n) \frac{\tau_n}{h}$  при всех  $m$ . Если  $A_n^1 \geq 0$ , то шаг  $\tau_n = t_{n+1} - t_n$  выбирают таким, чтобы выполнялось условие  $A_n^2 \tau_n / h \leq 1$ .

Для линейных и слаболинейных задач во всех известных случаях, когда было проведено строгое исследование устойчивости и коэффициенты уравнения удовлетворяли условию Липшица по всем переменным, имел место следующий факт. Если выполнялся критерий устойчивости по ПЗК, то схема действительно была устойчива.

#### § 4. Численное решение нелинейных задач с разрывными решениями

Рассмотрим задачу Коши для уравнения

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0 \quad (1)$$

при начальном условии

$$u_0(x) = \begin{cases} 1 & \text{при } x < 0, \\ 0 & \text{при } x \geq 0. \end{cases}$$

Эта задача не имеет непрерывного решения, поэтому разностная задача в обычном смысле не аппроксимирует дифференциальную. Однако тем не менее рассмотрим разностные аппро-

КСИМАЦИИ ЭТОЙ ЗАДАЧИ

$$\frac{u_m^{n+1} - u_m^n}{\tau} + u_m^n \frac{u_m^n - u_{m-1}^n}{h} = 0, \quad (2)$$

$$\frac{u_m^{n+1} - u_m^n}{\tau} + u_{m-1}^n \frac{u_m^n - u_{m-1}^n}{h} = 0, \quad (3)$$

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{u_m^n + u_{m-1}^n}{2} \cdot \frac{u_m^n - u_{m-1}^n}{h} = 0 \quad (4)$$

при начальных условиях

$$u_m^0 = \begin{cases} 1 & \text{при } m < 0, \\ 0 & \text{при } m \geq 0. \end{cases} \quad (5)$$

Нетрудно убедиться, что при  $\tau = h$  решением задачи (2), (5) является

$$u_m^n = \begin{cases} 1 & \text{при } m < 0, \\ 0 & \text{при } m \geq 0; \end{cases}$$

решением задачи (3), (5) является

$$u_m^n = \begin{cases} 1 & \text{при } m - n < 0, \\ 0 & \text{при } m - n \geq 0. \end{cases}$$

Решение задачи (4), (5) не выписывается в явном виде. При  $\tau = h$  из (4) имеем

$$u_m^{n+1} = u_m^n - \frac{1}{2} ((u_m^n)^2 - (u_{m-1}^n)^2). \quad (6)$$

Непосредственно вычисляя значения  $u_m^n$ , можно убедиться, что

$$u_m^n \approx \begin{cases} 1 & \text{при } m - n/2 < 0, \quad |m - n/2| \gg 1, \\ 0 & \text{при } m - n/2 > 0, \quad |m - n/2| \gg 1. \end{cases}$$

Более точные теоретические оценки показывают, что

$$u_m^n = \begin{cases} 1 + O(q^{|m-n/2|}) & \text{при } m - n/2 \rightarrow -\infty, \\ 0 + O(q^{|m-n/2|}) & \text{при } m - n/2 \rightarrow +\infty, \end{cases}$$

где  $q < 1$ ; следовательно, решение сеточной задачи (4), (5) близко к разрывной функции

$$u(x, t) = \begin{cases} 1, & x - t/2 < 0, \\ 0, & x - t/2 > 0. \end{cases}$$

Таким образом, решения различных разностных задач, аппроксимирующих на гладком решении одну и ту же дифференциальную задачу, в случае разрывного решения могут сходиться к различным пределам при стремлении шагов сетки к нулю. Заметим, что само решение такой дифференциальной задачи также не определено однозначно, пока ничего не сказано о том, как проходит линия разрыва решения.

В наиболее типичных случаях условия на линии разрыва являются следствием интегральных законов сохранения, из которых возникла данная дифференциальная задача.

Пусть  $u$  — гладкое решение уравнения (1); интегрируя (1) по переменным  $(x, t)$  по некоторой области  $G$ , получим

$$\int_G \left( \frac{\partial u}{\partial t} + \frac{1}{2} \frac{\partial u^2}{\partial x} \right) dx dt = 0$$

или

$$\int_{\Gamma} \left( \frac{1}{2} u^2 dt - u dx \right) = 0; \quad (7)$$

здесь  $\Gamma$  — граница области  $G$ . Если умножить уравнение (1) на  $u$  и проинтегрировать по области  $G$ , то получим

$$\int_G \left( \frac{1}{2} \frac{\partial u^2}{\partial t} + \frac{1}{3} \frac{\partial u^3}{\partial x} \right) dx dt = 0$$

или

$$\int_{\Gamma} \left( \frac{u^3}{3} dt - \frac{u^2}{2} dx \right) = 0. \quad (8)$$

Если для гладкой функции  $u(x, t)$  выполнены условия (7) или (8) для любого контура  $\Gamma$ , то эти условия равносильны. Дело обстоит иначе в случае разрывной  $u(x, t)$ .

Если  $u(x, t)$  — кусочно-гладкая функция, то из (7) можно получить, что в области гладкости функция  $u$  является решением уравнения (1), а вдоль линии разрыва  $X(t)$  выполнено соотношение

$$\frac{dX}{dt} = \omega = \frac{[u^2/2]}{[u]} = \frac{u_+ + u_-}{2};$$

здесь

$$f_+(x, t) = \lim_{\varepsilon > 0, \varepsilon \rightarrow 0} f(x + \varepsilon, t),$$

$$f_-(x, t) = \lim_{\varepsilon < 0, \varepsilon \rightarrow 0} f(x - \varepsilon, t), \quad [f] = f_+ - f_-.$$

Точно так же из (8) следует, что в области гладкости функция  $u$  является решением уравнения (1), а вдоль линии разрыва выполнено соотношение

$$\frac{dX}{dt} = \omega = \frac{[u^3/3]}{[u^2/2]} = \frac{2}{3} \frac{u_+^2 + u_+ u_- + u_-^2}{u_+ + u_-}.$$

Из вышесказанного видно, что для сходимости решения разностной задачи к разрывному решению уравнения (1) существенно определенное соответствие между разностной задачей и законом сохранения, соответствующим дифференциальной задаче.

Соображения здравого смысла, численный эксперимент и теоретические оценки погрешности для случая одной неизвест-

ной функции показали, что разностная схема должна обладать свойством *дивергентности*. Первоначально это свойство формулировалось в следующем виде: если ищется решение уравнения

$$\frac{\partial \varphi(x, t, u)}{\partial t} + \frac{\partial \psi(x, t, u)}{\partial x} = 0,$$

соответствующее закону сохранения  $\int (\psi dt - \varphi dx) = 0$ , то левая часть разностной схемы должна являться линейной комбинацией выражений

$$\varphi(x_m, t_n, u_m^n), \quad \psi(x_m, t_n, u_m^n) \quad (9)$$

или близких к ним. Например, разностная схема (4) удовлетворяет условию дивергентности по отношению к закону сохранения (7). По отношению к закону сохранения (8) условию дивергентности удовлетворяет разностная схема

$$\frac{(u_m^{n+1})^2 - (u_m^n)^2}{2\tau} + \frac{(u_m^n)^3 - (u_{m-1}^n)^3}{3h} = 0.$$

Впоследствии оказалось, что условие дивергентности допускает существенное расширение. Например, такому расширенному условию дивергентности по отношению к уравнению  $u_t + (\varphi(u))_x = 0$  удовлетворяет следующая разностная схема.

Сначала делается полшаг

$$\frac{u_{m+1/2}^{n+1/2} - \frac{u_{m+1}^n + u_m^n}{2}}{\tau} + \varphi' \left( \frac{u_{m+1}^n + u_m^n}{2} \right) \frac{u_{m+1}^n - u_m^n}{h} = 0,$$

а затем полный шаг по формуле

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{\varphi(u_{m+1/2}^{n+1/2}) - \varphi(u_{m-1/2}^{n+1/2})}{h} = 0.$$

Здесь разностная схема записывается в виде линейной комбинации выражений (9) лишь на заключительном шаге.

## § 5. Разностные схемы для одномерного параболического уравнения

После того как мы познакомились с вопросами устойчивости и сходимости для гиперболических задач на нестрогом уровне, перейдем к исследованию разностных схем для параболического уравнения в случае одной пространственной переменной на математическом уровне строгости.

Пусть требуется найти функцию  $u(x, t)$ , являющуюся решением уравнения

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t) \quad (1)$$



в области  $\bar{Q}_T = [0, X] \times [0, T]$  с начальными и краевыми условиями

$$u(x, 0) = u_0(x), \quad u(0, t) = \mu_1(t), \quad u(X, t) = \mu_2(t). \quad (2)$$

Всюду далее будем считать, что функции  $f$ ,  $\mu_i$  и  $u_0$  таковы, что существует достаточно гладкое решение задачи (1), (2).

При построении разностной схемы поступим так же, как и ранее. Разобьем исходную область  $\bar{Q}_T$  прямоугольной сеткой с шагами  $h = X/M$ ,  $\tau = T/N$  соответственно по координатам  $x$  и  $t$ . Будем искать функцию  $u^h$ , определенную в узлах  $(m, n)$  сетки  $\bar{Q}_{h, \tau} = \{(mh, n\tau), 0 \leq m \leq M, 0 \leq n \leq N\}$ , которая является приближением функции  $u$  в  $\bar{Q}_{h, \tau}$ . Обозначим, как и ранее,  $u^h(mh, n\tau) = u_m^n$ .

Заменим производные в (1) разностными отношениями. Производная  $\partial u / \partial t$  в точке  $(mh, n\tau)$  может быть заменена разностным отношением многими способами, например

$$\frac{\partial u}{\partial t} \Big|_{(mh, n\tau)} \approx \frac{u_m^{n+1} - u_m^n}{\tau} \quad \text{или же} \quad \frac{\partial u}{\partial t} \Big|_{(mh, n\tau)} \approx \frac{u_m^n - u_m^{n-1}}{\tau}.$$

В зависимости от способа аппроксимации будут получаться различные разностные схемы. Вторую производную по переменной  $x$  заменим обычным способом:

$$\frac{\partial^2 u}{\partial x^2} \Big|_{(mh, n\tau)} \approx \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}.$$

Подставляя эти соотношения вместо соответствующих производных в (1), получим

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2} + \Phi_m^n; \quad m = 1, \dots, M-1, \\ n = 0, \dots, N-1; \quad (3)$$

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2} + \Phi_m^{n+1}, \quad m = 1, \dots, M-1, \\ n = 0, \dots, N-1 \quad (4)$$

(вторая схема получается после переобозначения  $n \rightarrow n+1$ ). Функция  $\Phi_m^n$  является аппроксимацией  $f(x, t)$ .

Кроме уравнения (1) необходимо аппроксимировать начальные и граничные условия. Положим

$$u_m^0 = u_0(mh), \quad u_0^n = \mu_1(n\tau), \quad u_M^n = \mu_2(n\tau). \quad (5)$$

Таким образом, уравнения (3), (5) и (4), (5) являются разностной аппроксимацией задачи для параболического уравнения (1), (2).

Найдем порядок погрешности аппроксимации разностной схемы (3), (5). Для этого подставим в (3) точное решение

дифференциальной задачи. Так как

$$\frac{u(x, n\tau + \tau) - u(x, n\tau)}{\tau} = \frac{\partial u}{\partial t} \Big|_{(x, n\tau)} + \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2} \Big|_{(x, n\tau + \xi)}, \quad 0 \leq \xi \leq \tau,$$

$$\frac{u((m-1)h, t) - 2u(mh, t) + u((m+1)h, t)}{h^2} = \frac{\partial^2 u}{\partial x^2} \Big|_{(mh, t)} + \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4} \Big|_{(mh+\eta, t)}, \quad 0 \leq \eta \leq h,$$

то

$$\begin{aligned} \frac{u(x, t + \tau) - u(x, t)}{\tau} - \frac{u(x-h, t) - 2u(x, t) + u(x+h, t)}{h^2} - \varphi(x, t) = \\ = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} + \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2} \Big|_{(x, t+\xi)} - \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4} \Big|_{(x+\eta, t)} - \varphi(x, t) = \\ = f(x, t) - \varphi(x, t) + O(h^2 + \tau). \end{aligned}$$

Таким образом, если положить  $\varphi_m^n = f(mh, n\tau)$ , то порядок погрешности аппроксимации разностной схемы (3), (5) будет  $O(h^2 + \tau)$ , так как граничные и начальные условия выполнены точно. Аналогично устанавливается, что порядок погрешности аппроксимации схемой (4), (5) задачи (1), (2) также равен  $O(h^2 + \tau)$ .

Между схемами (3), (5) и (4), (5), однако, имеется принципиальная разница. Выясним ее суть. Из (3) следует соотношение

$$u_m^{n+1} = u_m^n + \frac{\tau}{h^2} (u_{m-1}^n - 2u_m^n + u_{m+1}^n) + \tau \varphi_m^n. \quad (6)$$

В силу того что значения  $u_m^0$  известны, из (6) можно найти значения  $u_m^1$  ( $m=1, \dots, M-1$ ) и т. д. Поэтому если значения  $u_m^n$  известны, то  $u_m^{n+1}$  находятся с помощью явных формул (6). Таким образом, для вычисления функции  $u^n$  в данном случае можно воспользоваться явной формулой (6). Поэтому схема (3), (5) называется *явной*.

Преобразуя (4), имеем

$$-\frac{\tau}{h^2} u_{m-1}^{n+1} + \left(1 + \frac{2\tau}{h^2}\right) u_m^{n+1} - \frac{\tau}{h^2} u_{m+1}^{n+1} = u_m^n + \tau \varphi_m^{n+1},$$

$$m = 1, \dots, M-1;$$

$$u_0^{n+1} = \mu_1^{n+1} \equiv \mu_1((n+1)\tau), \quad u_M^{n+1} = \mu_2^{n+1} \equiv \mu_2((n+1)\tau). \quad (7)$$

При известных  $u_m^n$ ,  $m=1, \dots, M-1$ , соотношения (7) представляют собой систему линейных алгебраических уравнений относительно неизвестных  $u_m^{n+1}$ ,  $m=1, \dots, M-1$ . Поэтому схема (4), (5) называется *неявной*.

Система уравнений (7) относительно вектора неизвестных  $v = \{u_1^{n+1}, \dots, u_{M-1}^{n+1}\}$  может быть записана в виде  $Av = b$ , где



В противном случае

$$\max_m |u_m^{n+1}| \leq \max(|\mu_1^{n+1}|, |\mu_2^{n+1}|).$$

Отсюда следует оценка

$$\|u^{n+1}\| \leq \max(|\mu_1^{n+1}|, |\mu_2^{n+1}|, \|u^n\| + \tau \|\varphi^n\|), \quad (9)$$

связывающая нормы векторов на соседних слоях.

Представим теперь решение  $u^h$  задачи (3), (5) в виде  $u^h = y^h + v^h$ , где  $y^h$  — решение задачи (3), (5) с правой частью  $\varphi^h \equiv 0$ , а  $v^h$  — решение задачи (3), (5) с однородными граничными и начальными условиями. В силу оценки (9) для  $y^h$  имеем

$$\|y^{n+1}\| \leq \max\left(\max_{0 \leq k \leq N} |\mu_1^k|, \max_{0 \leq k \leq N} |\mu_2^k|, \|y^n\|\right) \leq \dots$$

$$\dots \leq \max\left(\max_{0 \leq k \leq N} |\mu_1^k|, \max_{0 \leq k \leq N} |\mu_2^k|, \|u^0\|\right).$$

С другой стороны, для  $v^h$  в силу этой же оценки (9) получаем

$$\|v^{n+1}\| \leq \|v^n\| + \tau \|\varphi^n\| \leq \|v^{n-1}\| + \tau (\|\varphi^n\| + \|\varphi^{n-1}\|) \leq \dots$$

$$\dots \leq \sum_{k=0}^n \tau \|\varphi^k\| \leq T \max_{0 \leq k \leq N} \|\varphi^k\|,$$

если  $(n+1)\tau \leq T$ . Таким образом, окончательно имеем

$$\|u^n\| \leq \|v^n\| + \|y^n\| \leq$$

$$\leq \max\left(\max_{0 \leq k \leq N} |\mu_1^k|, \max_{0 \leq k \leq N} |\mu_2^k|, \|u^0\|\right) + T \max_{0 \leq k \leq N} \|\varphi^k\|.$$

Так как это неравенство справедливо при любом  $n$ ,  $0 \leq n \leq N$ , то это и означает устойчивость разностной схемы в сеточной норме пространства  $S$ . Теорема доказана. (Это соотношение можно получить также непосредственно из (9), не вводя в рассмотрение функции  $y^h$  и  $v^h$ .)

Отметим, что постоянная в (8) получилась зависящей от  $T$ . Если  $\tau/h^2 = \kappa = \text{const}$ , то условие  $\kappa \leq 1/2$  является необходимым и достаточным условием устойчивости.

**Задача 1.** Пусть  $\lim_{h, \tau \rightarrow 0} \frac{\tau/h^2 - 1/2}{\tau} = \infty$ . Доказать, что схема (3), (5) неустойчива. У к а з а н и е. Рассмотреть частные решения

$$u_m^n = \lambda_q^n \sin \frac{\pi m h}{X} q.$$

Доказательство устойчивости разностной схемы (3), (5) было получено при соотношении на шаги сетки  $\tau \leq h^2/2$ . Разностные схемы, которые обладают устойчивостью при определенных соотношениях между шагами сетки, называются *условно устойчивыми*. Соответственно если схема устойчива при любых соотношениях между шагами сетки, то она называется *безусловно устойчивой*.

Покажем, что схема (4), (5) относится к классу безусловно устойчивых схем. Справедлива

**Теорема 2.** При любых  $h$  и  $\tau$  для решения задачи (4), (5) имеет место оценка (8).

Доказательство. По аналогии с доказательством предыдущей теоремы преобразуем (4) к виду

$$u_m^{n+1} + \rho(-u_{m-1}^{n+1} + 2u_m^{n+1} - u_{m+1}^{n+1}) = u_m^n + \tau\varphi_m^{n+1}, \quad 1 \leq m \leq M-1. \quad (10)$$

Из всех значений  $u_m^{n+1}$ , по модулю равных  $\|u^{n+1}\|$ , возьмем то, у которого индекс  $m$  принимает наименьшее значение. Если  $m=0$  или же  $m=M$ , то выполнение (9) очевидно. Пусть теперь  $m$  отлично от 0 и  $M$ . Поскольку  $|u_m^{n+1}| > |u_{m-1}^{n+1}|$  по определению  $m$  и  $|u_m^{n+1}| \geq |u_{m+1}^{n+1}|$ , то  $|2u_m^{n+1}| > |u_{m-1}^{n+1}| + |u_{m+1}^{n+1}|$ ; поэтому  $\text{sign}(2u_m^{n+1} - u_{m-1}^{n+1} - u_{m+1}^{n+1}) = \text{sign} u_m^{n+1}$ . Тогда

$$\begin{aligned} \|u^{n+1}\| &= |u_m^{n+1}| \leq |u_m^{n+1} + \rho(-u_{m-1}^{n+1} + 2u_m^{n+1} - u_{m+1}^{n+1})| = \\ &= |u_m^n + \tau\varphi_m^{n+1}| \leq \|u^n\| + \tau\|\varphi^{n+1}\|; \end{aligned}$$

при любых  $h$  и  $\tau$  для схемы (4), (5) получена оценка (9). Завершение доказательства теоремы совпадает с доказательством предыдущей теоремы. Таким образом, схема (4), (5) является безусловно устойчивой.

Между явной (3), (5) и неявной (4), (5) схемами имеется, таким образом, принципиальное отличие. Явной схеме соответствуют явные формулы для вычисления вектора на слое по известным значениям на предыдущем слое. Однако эта схема является условно устойчивой. Это приводит к тому, что при малом шаге  $h$  мы вынуждены выбирать слишком мелкий шаг по времени ( $\tau \leq h^2/2$ ), чтобы обеспечить устойчивость. Это, в свою очередь, приводит к значительному увеличению затрат времени счета на ЭВМ и не может быть оправдано требованиями точности, если по временной переменной  $t$  решение достаточно гладкое. С другой стороны, при использовании неявной схемы можно значительно увеличить шаг по времени  $\tau$ , однако при переходе от слоя к слою требуется каждый раз решать систему уравнений. Впрочем, в одномерном случае это не представляет проблемы. В частности, используя метод прогонки, можно получить вектор  $u^{n+1}$  при известном  $u^n$  за  $O(M)$  операций, т. е. количество арифметических операций при переходе от слоя к слою по порядку будет тем же, что и в случае явной схемы. Это позволяет сделать вывод о том, что использование неявных схем в одномерном случае часто является более предпочтительным, так как ведет к уменьшению затрат времени счета на ЭВМ.

Перейдем теперь к исследованию устойчивости разностных схем (3), (5) и (4), (5) в других нормах, в частности в сеточных аналогах норм  $L_2$  и  $W_2^1$  по слою. Так как исследование

свойств схем (3), (5) и (4), (5) проводится по одной и той же методике, то имеет смысл объединить эти две схемы. Для наглядности сопоставим разностной схеме рисунок, на котором изображено взаимное расположение узлов сетки, значения в которых связывает разностное уравнение (так называемый *шаблон*). В нашем случае для схем (3), (5) и (4), (5) шаблоны имеют вид, изображенный соответственно на рис. 10.5.1 и 10.5.2.

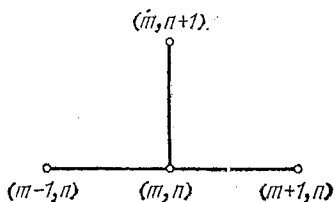


Рис. 10.5.1

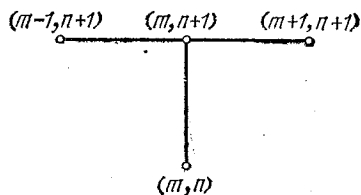


Рис. 10.5.2

Пусть  $v = (v_0, \dots, v_M)$ . Обозначим  $\Delta v_m = (v_{m+1} - 2v_m + v_{m-1})/h^2$ ,  $\Delta u|_{(x,t)} = u(x+h,t) - 2u(x,t) + u(x-h,t)$ . Введем более общий, чем (3), (5) и (4), (5), вид схемы

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \sigma \Delta u_m^{n+1} + (1 - \sigma) \Delta u_m^n + \Phi_m^n, \quad m = 1, \dots, M-1. \quad (11)$$

Постоянная  $\sigma$  в (11) называется *весом* и обычно берется в пределах  $0 \leq \sigma \leq 1$ . В частности, при  $\sigma = 0$  выражение (11) переходит в (3), а при  $\sigma = 1$  получаем (4). Разностную схему (11), (5) называют *схемой с весами*. Она имеет шеститочечный шаблон при  $\sigma \in (0, 1)$  (рис. 10.5.3). Схема (11), (5) является явной лишь при  $\sigma = 0$ . Разностную схему (4), (5), чтобы отличить ее от других неявных схем вида (11) с  $0 < \sigma < 1$ , называют *чисто неявной схемой*.

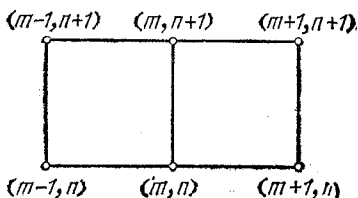


Рис. 10.5.3

Пусть

$$L_{h,\tau}^\sigma u^h|_{(m,n)} = \frac{u_m^{n+1} - u_m^n}{\tau} - \sigma \Delta u_m^{n+1} - (1 - \sigma) \Delta u_m^n.$$

Как и выше, назовем величину

$$r = L_{h,\tau}^\sigma [u]_h - \Phi^h, \quad (12)$$

где  $[u]_h$  — значения решения в узлах сетки  $\bar{Q}_{h,\tau}$ , *погрешностью аппроксимации разностной схемы (11) уравнения (1)*. Заметим, что граничные и начальные условия для функции  $u^h$  выполняются точно. Используя разложение  $u$  в ряд Тейлора в точке

$(x, t + \tau/2) = (mh, n\tau + \tau/2)$ , имеем

$$\frac{u(x, t + \tau) - u(x, t)}{\tau} = \frac{\partial u}{\partial t} \Big|_{(x, t + \tau/2)} + \frac{\tau^2}{24} \frac{\partial^3 u}{\partial t^3} \Big|_{(x, t + \tau/2)}$$

$$\begin{aligned} \Delta u \Big|_{(x, t + \frac{\tau}{2} \pm \frac{\tau}{2})} &= \Delta u \Big|_{(x, t + \frac{\tau}{2})} \pm \frac{\tau}{2} \Lambda \frac{\partial u}{\partial t} \Big|_{(x, t + \frac{\tau}{2})} + O(\tau^2) = \\ &= \frac{\partial^2 u}{\partial x^2} \Big|_{(x, t + \frac{\tau}{2})} + \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4} \Big|_{(x, t + \frac{\tau}{2})} \pm \frac{\tau}{2} \Lambda \frac{\partial u}{\partial t} \Big|_{(x, t + \frac{\tau}{2})} + O(\tau^2 + h^4) \end{aligned}$$

Отсюда

$$\begin{aligned} r_m^n &= L_{h, \tau}^\sigma [u] - \Phi^h \Big|_{(m, n)} = f \left( x, t + \frac{\tau}{2} \right) - \\ &\quad - \varphi_m^n + \tau(\sigma - 0,5) \Lambda \frac{\partial u}{\partial t} \Big|_{(x, t + \frac{\tau}{2})} + O(\tau^2 + h^2) \end{aligned}$$

Таким образом, если  $\varphi_m^n = f(x, t + \tau/2)$ , то  $r = O(h^2 + \tau)$  при  $\sigma \neq 0,5$  и  $r = O(h^2 + \tau^2)$  при  $\sigma = 0,5$ .

Будем исследовать устойчивость по начальным данным, т. е. будем оценивать чувствительность решения к возмущению начальных данных. Положим  $\mu_k(t) \equiv 0$ ,  $\varphi^h \equiv 0$ . Обозначим

$$\|u^n\|_{L_{2, h}} = \left( \sum_{m=1}^{M-1} h (u_m^n)^2 \right)^{1/2}.$$

Назовем разностную схему *устойчивой по начальным данным в норме  $L_{2, h}$* , если существует постоянная  $c_1$ , не зависящая от шагов сетки  $h$  и  $\tau$ , такая, что для решения  $u^h$  задачи (11), (5) с  $\mu_1 = \mu_2 = \varphi_h \equiv 0$  справедлива оценка

$$\max_{0 \leq n \leq N} \|u^n\|_{L_{2, h}} \leq c_1 \|u^0\|_{L_{2, h}}. \quad (13)$$

В теории разностных схем установилась терминология, когда не делают различия между матрицей и порождаемым ею линейным оператором. В дальнейшем мы будем следовать этой терминологии.

Обозначим через  $\Lambda$  оператор (матрицу), который вектору  $(0, v_1, \dots, v_{M-1}, 0)$  ставит в соответствие вектор  $(0, \Lambda v_1, \dots, \Lambda v_{M-1}, 0)$ . В случае  $\mu_1 = \mu_2 = \varphi_h \equiv 0$  уравнение (11) можно записать в виде

$$u^{n+1} = Su^n.$$

Матрица  $S$  называется *матрицей* или *оператором перехода от слоя к слою*. В общем случае  $S$  может зависеть от  $n$ . Пусть  $\{\lambda_i\}$ ,  $i = 1, \dots, M-1$  — собственные числа матрицы  $S$ . Матрица  $S$  симметрична и поэтому  $\|S\|_2 = \max_i |\lambda_i|$ . Функция  $u_m^0$  мо-

жет быть представлена в виде дискретной суммы Фурье

$$u_m^0 = \sum_{k=1}^{M-1} c_k \sin \frac{\pi k (mh)}{X}.$$

Из равенств

$$\Lambda \sin \frac{\pi k mh}{X} = - \frac{4 \sin^2 \frac{\pi k h}{2}}{h^2} \sin \frac{\pi k mh}{X} = - \nu_k \sin \frac{\pi k mh}{X},$$

$$\nu_k = \frac{4 \sin^2 \frac{\pi k h}{2}}{h^2}$$

следует, что числа  $-\nu_k$  являются собственными значениями оператора  $\Lambda$ . Из (11) получаем  $(E - \sigma \tau \Lambda) u^1 = (E + \tau(1 - \sigma) \Lambda) u^0$ , т. е.  $S = (E - \sigma \tau \Lambda)^{-1} (E + \tau(1 - \sigma) \Lambda)$ . Поэтому

$$u_m^1 = S \left( \sum_{k=1}^{M-1} c_k \sin \frac{\pi k (mh)}{X} \right) = \sum_{k=1}^{M-1} \frac{1 - \tau(1 - \sigma) \nu_k}{1 + \tau \sigma \nu_k} c_k \sin \frac{\pi k (mh)}{X}.$$

Таким образом, собственные числа  $\lambda_k$  матрицы  $S$  имеют вид

$$\lambda_k = \frac{1 - \tau(1 - \sigma) \nu_k}{1 + \tau \sigma \nu_k}.$$

Выясним, когда будет выполнено условие  $|\lambda_k| \leq 1$ . Заменяя  $\lambda_k$  его значением, получаем

$$-1 \leq \frac{1 - \tau(1 - \sigma) \nu_k}{1 + \tau \sigma \nu_k} \leq 1.$$

Так как  $\nu_k > 0$ , то при  $\tau > 0$  это неравенство эквивалентно соотношению  $\tau(1 - 2\sigma) \nu_k \leq 2$ . При  $\sigma \geq 1/2$  последнее неравенство будет выполняться при любом  $\tau > 0$ , а при  $\tau < 1/2$  — при

$$\tau \leq \frac{2}{(1 - 2\sigma) \max_k \nu_k} \leq \frac{h^2}{2(1 - 2\sigma)}. \quad (14)$$

Таким образом, нами получены достаточные условия устойчивости схемы (11), (5) по начальным данным. А именно, если  $\varphi \equiv 0$  и  $\mu_1, \mu_2 \equiv 0$ , то при  $\sigma \geq 1/2$  разностная схема (11), (5) безусловно устойчива; при  $\sigma < 1/2$  схема устойчива, если шаги  $h$  и  $\tau$  связаны соотношением (14), т. е. схема (11), (5) в этом случае условно устойчива.

Предположим, что условие  $|\lambda_k| \leq 1$  нарушается:  $|\lambda_{M-1}| \geq 1 + \delta$ , где  $\delta > 0$  и не зависит от  $h$  и  $\tau$ . Положим  $u_m^0 = \sin \frac{\pi(M-1)mh}{X}$ .



Тогда

$$u_m^n = (1 + \bar{\delta})^n \sin \frac{\pi (M-1) m h}{X} \quad \text{и} \quad \|u^n\| = (1 + \bar{\delta})^n \|u^0\|, \quad \text{где} \quad \bar{\delta} \geq \delta > 0.$$

В этом случае  $\max_{n, n\tau \leq T} \|u^n\| \rightarrow \infty$  при  $\tau \rightarrow 0$ , т.е. схема неустойчива.

Иногда используют несколько отличное определение устойчивости по начальным данным. Говорят, что разностная схема устойчива по начальным данным, если собственные числа оператора перехода лежат в круге радиуса  $1 + \sigma$ . Покажем, что в рассматриваемом примере это определение согласуется с (13). Действительно, пусть  $\lambda_k$  — собственные числа матрицы  $S$ .

Тогда  $u^n = S^n u^0 = \sum_{k=1}^{M-1} \lambda_{k^n} c_k \sin \frac{\pi k n}{X}$  и  $\|u^n\| \leq (1 + \sigma \tau)^n \|u^0\|$ . В этом случае при  $n\tau = \text{const}$  и  $\tau \rightarrow 0$  имеем  $\|u^n\| \leq e^{\sigma T} \|u^0\|$ ,  $n\tau \leq T$ .

При исследовании разностных схем для более сложных задач, например при других типах граничных условий, более общем операторе в правой части уравнения (1) и т. п., доказательство устойчивости с использованием принципа максимума или метода Фурье вызывает определенные затруднения, а иногда исследование устойчивости этими приемами является просто невозможным. В этом случае исследование устойчивости разностных схем обычно проводится методом энергетических оценок. Опишем кратко его суть на дифференциальном уровне. Пусть  $u(x, t)$  — решение задачи

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f, \quad u(0, t) = u(X, t) = 0, \quad u(x, 0) = u_0(x). \quad (15)$$

Будем предполагать, что правая часть  $f(x, t)$  и функция  $u_0(x)$  таковы, что при любом  $t \in [0, T]$  существует интеграл

$\int_0^X \left(\frac{\partial u}{\partial x}\right)^2 dx < \infty$  и  $\frac{\partial u}{\partial t}$  непрерывна по  $t$ . Умножим обе части

уравнения (15) на  $u$  и проинтегрируем по  $x$ . Используя формулу интегрирования по частям, получаем энергетическое тождество

$$\frac{1}{2} \int_0^X \frac{\partial}{\partial t} u^2 dx + \int_0^X \left(\frac{\partial u}{\partial x}\right)^2 dx = \int_0^X f u dx. \quad (16)$$

Для функции  $\varphi(x, t)$ , которая при любом  $t \in [0, T]$  принадлежит

пространству  $\dot{W}_2^1[0, X]$ , обозначим  $\|\varphi(t)\|_1^2 = \int_0^X \left(\frac{\partial \varphi}{\partial x}\right)^2 dx$ .

Если функция  $f(t)$  такова, что для любой  $g \in \dot{W}_2^1$  существует интеграл  $\int_0^X f g dx$ , то через  $\|f\|_{-1}$  обозначим норму

$$\|f(t)\|_{-1} = \sup_{g \in \dot{W}_2^1} \frac{1}{\|g\|_1} \int_0^X |f g| dx.$$

Согласно определению  $\| \cdot \|_{-1}$  имеем

$$\left| \int_0^x fu \, dx \right| \leq \|f(t)\|_{-1} \|u(t)\|_1 \leq \varepsilon \|u(t)\|_1^2 + \frac{1}{4\varepsilon} \|f(t)\|_{-1}^2.$$

При получении последней оценки было использовано  $\varepsilon$ -неравенство

$$|ab| \leq \varepsilon a^2 + \frac{1}{4\varepsilon} b^2, \quad (17)$$

которое следует из соотношений

$$0 \leq \left( \sqrt{\varepsilon} a \pm \frac{1}{2\sqrt{\varepsilon}} b \right)^2 = \varepsilon a^2 + \frac{b^2}{4\varepsilon} \pm ab.$$

В данном случае мы обозначили  $\|u(t)\|_1$  через  $a$ , а  $\|f(t)\|_{-1}$  — через  $b$ .

Тогда из (16) следует неравенство

$$\frac{1}{2} \frac{\partial}{\partial t} \int_0^x u^2 \, dx + (1 - \varepsilon) \|u(t)\|_1^2 \leq \frac{1}{4\varepsilon} \|f(t)\|_{-1}^2.$$

Проинтегрируем последнее неравенство по  $t$  в пределах от нуля до  $T$ . В результате получим

$$\frac{1}{2} \|u(T)\|^2 + (1 - \varepsilon) \int_0^T \|u(t)\|_1^2 \, dt \leq \frac{1}{2} \|u_0\|^2 + \frac{1}{4\varepsilon} \int_0^T \|f(t)\|_{-1}^2 \, dt.$$

Здесь  $\|u(t)\| = \left( \int_0^x u^2(x, t) \, dx \right)^{1/2}$ . Последнее неравенство называется

*энергетическим неравенством*. Из него, в частности, следует, что решение  $u(x, t)$  непрерывно зависит от правой части и начальных условий.

Применим несколько похожую схему к исследованию устойчивости сеточной задачи (11), (5). Напомним, что  $u^n$  — значение  $u^h$  на  $n$ -м слое, т. е.  $u^n = (0, u_1^n, \dots, u_{M-1}^n, 0)$ . В этом случае уравнение (11) может быть переписано в виде

$$u_i^n = \sigma \Lambda u^{n+1} + (1 - \sigma) \Lambda u^n + \varphi^n, \quad (18)$$

где  $u_i^n = (u^{n+1} - u^n)/\tau$ . В пространстве векторов такого вида введем скалярное произведение и нормы:

$$(v, w) = \sum_{m=1}^{M-1} h v_m w_m, \quad \|v\|^2 = (v, v),$$

$$\|v\|_1^2 = \sum_{m=0}^{M-1} h \left( \frac{v_{m+1} - v_m}{h} \right)^2.$$

Заметим, что

$$u^{n+1} = \frac{1}{2}(u^{n+1} + u^n) + \frac{\tau}{2}u_i^n, \quad u^n = \frac{1}{2}(u^{n+1} + u^n) - \frac{\tau}{2}u_i^n,$$

поэтому (18) можно преобразовать к виду

$$u_i^n = \frac{1}{2} \Lambda (u^{n+1} + u^n) + \tau(\sigma - 0,5) \Lambda u_i^n + \varphi^n. \quad (19)$$

Умножим обе части (19) скалярно на вектор  $2\tau u_i^n$ . Получим

$$2\tau \|u_i^n\|^2 = (\Lambda(u^{n+1} + u^n), u^{n+1} - u^n) + \\ + 2\tau^2(\sigma - 0,5)(\Lambda u_i^n, u_i^n) + 2\tau(u_i^n \varphi^n), \quad (20)$$

Из формулы суммирования по частям (9.8.14)

$$\sum_{m=1}^{M-1} h \frac{a_{m+1} - a_m}{h} b_m = - \sum_{m=1}^M h \frac{b_m - b_{m-1}}{h} a_m + a_M b_M - a_1 b_0,$$

положив  $a_m = \frac{v_m - v_{m-1}}{h}$ ,  $b_m = v_m$ , имеем

$$(\Lambda v, v) = \sum_{m=1}^{M-1} h \frac{v_{m-1} - 2v_m + v_{m+1}}{h^2} v_m = - \sum_{m=0}^{M-1} h \left( \frac{v_{m+1} - v_m}{h} \right)^2 = \\ = - \|v\|_1^2.$$

Оператор  $\Lambda$  является симметричным (см. гл. X), поэтому

$$(\Lambda(u^{n+1} + u^n), u^{n+1} - u^n) = (\Lambda u^{n+1}, u^{n+1}) - (\Lambda u^n, u^n) = \\ = \|u^{n+1}\|_1^2 - \|u^n\|_1^2.$$

Используя полученные соотношения, преобразуем (20) к виду

$$2\tau \|u_i^n\|^2 + \|u^{n+1}\|_1^2 + 2\tau^2(\sigma - 0,5) \|u_i^n\|_1^2 = \|u^n\|_1^2 + 2\tau(u_i^n, \varphi^n). \quad (21)$$

Полученное равенство по аналогии с непрерывным случаем называют *энергетическим тождеством*.

Оценим скалярное произведение в правой части (21) при помощи  $\varepsilon$ -неравенства  $|(u_i^n, \varphi^n)| \leq \varepsilon \|u_i^n\|_1^2 + \frac{1}{4\varepsilon} \|\varphi^n\|^2$ . Тогда из (21) имеем оценку

$$2\tau [(1 - \varepsilon) \|u_i^n\|_1^2 + \tau(\sigma - 0,5) \|u_i^n\|_1^2] + \|u^{n+1}\|_1^2 \leq \|u^n\|_1^2 + \frac{\tau}{2\varepsilon} \|\varphi^n\|^2. \quad (22)$$

Выясним, при каких  $\sigma$  выражение в квадратных скобках будет неотрицательным. Заметим, что здесь  $\varepsilon$  — произвольное положительное число, которое до сих пор не было фиксировано. Очевидно, что при  $0 < \varepsilon \leq 1$  условие  $\sigma \geq 0,5$  является достаточным, чтобы выражение в квадратных скобках было неотрицательным. В этом случае, фиксируя  $\varepsilon \leq 1$  (например, можно положить

$\varepsilon = 1$ ), из (22) получим

$$\|u^{n+1}\|_1^2 \leq \|u^n\|_1^2 + \frac{\tau}{2\varepsilon} \|\varphi^n\|^2. \quad (23)$$

Проведем более детальное исследование устойчивости при  $\sigma < 0,5$ . С учетом неравенства  $\|v\|_1^2 \leq \frac{4}{h^2} \|v\|^2$ , которое было установлено ранее, из (22) получаем

$$2\tau \left[ 1 - \varepsilon + \frac{4(\sigma - 0,5)\tau}{h^2} \right] \|u_i^n\|^2 + \|u^{n+1}\|_1^2 \leq \|u^n\|_1^2 + \frac{\tau}{2\varepsilon} \|\varphi^n\|^2.$$

Поэтому из выполнения соотношения  $1 - \varepsilon + 4(\sigma - 0,5)\tau/h^2 \geq 0$  следует справедливость (23). Таким образом, для справедливости (23) при  $\sigma < 0,5$  достаточно выполнения соотношения

$$\tau \leq \frac{(1 - \varepsilon)h^2}{4(0,5 - \sigma)}. \quad (24)$$

Используя оценку (23) рекуррентным образом, получаем

$$\|u^n\|_1^2 \leq \|u^0\|_1^2 + \sum_{j=0}^{n-1} \frac{\tau}{2\varepsilon} \|\varphi^j\|^2, \quad n\tau \leq T. \quad (25)$$

Последнее неравенство как раз и означает устойчивость разностной схемы (11), (5) по начальным данным и правой части. При этом сумма в правой части (25) является квадратурной

формулой для интеграла  $\int_0^{n\tau} \frac{\|f(t)\|^2}{2\varepsilon} dt$ .

Отметим, что если соответствующие квадратурные суммы интеграла  $\int_0^\infty \|f(t)\|^2 dt$  ограничены, то из (25) следует ограниченность сеточного решения на бесконечном промежутке времени.

Таким образом, схема (11), (5) по доказанному выше является безусловно устойчивой при  $\sigma \geq 0,5$  и условно устойчивой (шаги  $h$  и  $\tau$  удовлетворяют соотношению (24),  $\varepsilon \geq 0$  не зависит от шагов сетки) при  $\sigma < 0,5$ .

Нами практически не рассматривался вопрос об устойчивости по граничным условиям. Дело заключается в следующем. Возьмем функцию  $s(x, t) = \mu_1(t)(X - x)/X + \mu_2(t)x/X$ . В этом случае функция  $v(x, t) = u(x, t) - s(x, t)$  является решением задачи (1) с однородными граничными условиями и правой частью  $\dot{f} + \partial s/\partial t$ . Таким образом, если функции  $\mu_1$  и  $\mu_2$  имеют производные по  $t$ , то граничные условия в задаче (1) могут быть сняты описанным способом. В сеточном случае можно иначе свести задачу (11) с неоднородными граничными условиями (5) к задаче с однородными условиями. Пусть  $u_m^n$  — решение сеточной задачи.

Положим

$$v_m^n = \begin{cases} u_m^n, & 1 \leq m \leq M-1, \\ 0, & m=0, \quad m=M. \end{cases}$$

В этом случае  $v_m^n$  удовлетворяет однородным граничным условиям (5), начальным условиям (5) и системе уравнений (11) при  $\varphi_m^n$  замененной на  $\psi_m^n$ , где

$$\psi_m^n = \begin{cases} \varphi_m^n, & 2 \leq m \leq M-2, \\ \varphi_1^n - \frac{1-\sigma}{h^2} \mu_1^n - \frac{\sigma}{h^2} \mu_1^{n+1}, & m=1, \\ \varphi_{M-1}^n - \frac{1-\sigma}{h^2} \mu_{M-1}^n - \frac{\sigma}{h^2} \mu_{M-1}^{n+1}, & m=M-1. \end{cases}$$

Таким образом, задача (11), (5) с неоднородными граничными условиями может быть записана как задача с однородными граничными условиями и некоторой измененной правой частью.

Получим оценку скорости сходимости. Будем рассматривать схему (11), (5), так как остальные схемы являются ее частным случаем. Пусть  $u$  — решение дифференциальной задачи (1), а  $u^h$  — решение разностной задачи (11), (5). По определению  $L_{h,\tau}^\sigma[u] - \varphi^h = r$ , где  $r$  — погрешность аппроксимации. Рассмотрим разность  $z = [u] - u^h$ . Она удовлетворяет уравнению

$$\frac{z^{n+1} - z^n}{\tau} = \sigma \Lambda z^{n+1} + (1 - \sigma) \Lambda z^n + \varphi^n - f^n + r^n. \quad (26)$$

Таким образом,  $z^n$  является решением задачи (11) с правой частью  $r$ , равной погрешности аппроксимации, и однородными граничными и начальными условиями (5). Если рассматриваемая схема устойчива, то справедлива оценка (25), откуда

$$\|z^n\|_1^2 \leq \sum_{j=0}^{n-1} \frac{\tau}{2\varepsilon} \|r^j\|^2 \leq \sum_{j=0}^{N-1} \frac{\tau}{2\varepsilon} \|r^j\|^2 \leq \frac{T}{2\varepsilon} \max_{0 \leq j \leq N-1} \|r^j\|^2.$$

В силу того что  $r = O(h^2 + \tau)$  при  $\sigma \neq 0,5$  и  $r = O(h^2 + \tau^2)$  при  $\sigma = 0,5$ , из последней оценки получаем, что решения сеточной задачи (11), (5)  $u^h$  сходятся к решению дифференциальной задачи  $u$  в сеточной норме  $\max \|z^n\|_1$  со скоростью  $O(h^2 + \tau)$ , если  $\sigma \neq 0,5$ , и со скоростью  $O(h^2 + \tau^2)$ , если  $\sigma = 0,5$ . При стремлении шагов сетки  $h$  и  $\tau$  к нулю в случае условно устойчивых схем всегда предполагаем выполненным неравенство, обеспечивающее устойчивость схемы.

Таким образом, решение сеточной задачи сходится к решению дифференциальной задачи со скоростью, по порядку совпадающей с порядком аппроксимации разностной схемы,

Сходимость разностных схем была установлена в сеточной норме пространства  $\overset{\circ}{W}_2^1$  на слое. Для получения оценки скорости сходимости в других нормах следует воспользоваться соответствующими оценками устойчивости либо использовать сеточные теоремы вложения. В частности, из теоремы вложения

$$\max_{1 \leq m \leq M-1} |v_m| \leq \frac{\sqrt{X}}{2} \|v\|_1$$

(см. § 9.8) следует сходимость разностной схемы с порядком, равным порядку аппроксимации, в сеточной норме пространства  $S$ .

Рассмотрим аппроксимацию граничных условий другого типа. Пусть, например,

$$lu \equiv \frac{\partial u}{\partial x} - au \Big|_{(0, t)} = 0. \quad (27)$$

Заменяя в (27) производную  $\partial u / \partial x$  разностным отношением, получаем аппроксимацию граничного условия

$$l_h u^n \equiv \frac{u_1^n - u_0^n}{h} - au_0^n = 0. \quad (28)$$

Оценим погрешность такой аппроксимации. Имеем

$$\begin{aligned} l_h[u] &= \frac{u(h, t) - u(0, t)}{h} - au(0, t) = \\ &= lu + \frac{h}{2} \frac{\partial^2 u}{\partial x^2} \Big|_{(0, t)} + O(h^2) = \frac{h}{2} \frac{\partial^2 u}{\partial x^2} \Big|_{(0, t)} + O(h^2). \end{aligned} \quad (29)$$

Таким образом, построенная аппроксимация (28) граничного условия (27) имеет первый по  $h$  порядок аппроксимации. Для повышения порядка аппроксимации воспользуемся приемом, который применялся при аппроксимации условия (27) в краевых задачах для обыкновенных дифференциальных уравнений. Из уравнения (1) выразим  $\frac{\partial^2 u}{\partial x^2}$ . Получим  $\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} - f$ . Тогда  $\frac{\partial^2 u}{\partial x^2} \Big|_{(0, t)} = \frac{u(0, t + \tau) - u(0, t)}{\tau} - f(0, t) + O(\tau)$  либо  $\frac{\partial^2 u}{\partial x^2} \Big|_{(0, t)} = \frac{u(0, t) - u(0, t - \tau)}{\tau} - f(0, t) + O(\tau)$ . Подставив это выражение в (29), получим

$$l_h[u] = \frac{h}{2} \left( \frac{u(0, t + \tau) - u(0, t)}{\tau} - f(0, t) \right) + O(h^2 + \tau).$$

Таким образом, аппроксимация граничного условия (27) с порядком  $O(h^2 + \tau)$  будет иметь вид

$$\frac{u_1^n - u_0^n}{h} - au_0^n - \frac{h}{2} \left( \frac{u_0^{n+1} - u_0^n}{\tau} - f_0^n \right) = 0 \quad (30)$$

либо

$$\frac{u_1^{n+1} - u_0^{n+1}}{h} - \alpha u_0^{n+1} - \frac{h}{2} \left( \frac{u_0^{n+1} - u_0^n}{\tau} - f_0^{n+1} \right) = 0. \quad (31)$$

Можно также использовать линейную комбинацию этих условий

$$\sigma \left( \frac{u_1^{n+1} - u_0^{n+1}}{h} - \alpha u_0^{n+1} \right) + (1 - \sigma) \left( \frac{u_1^n - u_0^n}{h} - \alpha u_0^n \right) - \frac{h}{2} \frac{u_0^{n+1} - u_0^n}{\tau} = -\frac{h}{2} (\sigma f_0^{n+1} + (1 - \sigma) f_0^n). \quad (32)$$

Из построения следует, что аппроксимация граничного условия (32) имеет порядок  $O(h^2 + \tau)$ . Непосредственно можно убедиться, что при  $\sigma = 0,5$  условие (32) аппроксимирует (27) с порядком  $O(h^2 + \tau^2)$ . Заметим, что шаблон, на котором осуществляется аппроксимация, содержит в общем случае только четыре узла  $(0, n)$ ,  $(1, n)$ ,  $(0, n+1)$ ,  $(1, n+1)$  (см. рис. 10.5.4). Поэтому структура матрицы линейных уравнений для нахождения  $u^{n+1}$  в неявных схемах фактически не изменится; матрица будет иметь трехдиагональный вид. Исследование устойчивости разностных схем с краевыми условиями третьего рода проводится методом энергетических оценок по схеме, описанной выше. По аналогии с построениями, проводившимися для обыкновенных дифференциальных уравнений, можно строить схемы повышенного порядка точности, например  $O(h^4 + \tau)$  или же  $O(h^4 + \tau^2)$ .

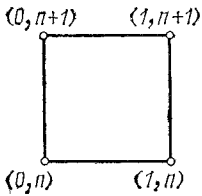


Рис. 10.5.4

## § 6. Разностная аппроксимация эллиптических уравнений

По сравнению с краевыми задачами для обыкновенных дифференциальных уравнений при построении разностных схем в многомерном случае возникают дополнительные трудности, связанные в основном с аппроксимацией граничных условий.

Рассмотрим простейшую краевую задачу — задачу Дирихле для уравнения Пуассона. Пусть область  $\Omega$  представляет единичный квадрат:  $\Omega = \{(x, y), 0 < x, y < 1\}$ ;  $\partial\Omega$  — граница  $\Omega$ . Требуется найти функцию  $u$ , дважды непрерывно дифференцируемую внутри  $\Omega$  и непрерывную в замкнутой области  $\bar{\Omega}$ , которая внутри области удовлетворяет уравнению

$$-\Delta u \equiv -\left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) = f \quad (1)$$

и принимает на границе заданное значение

$$u|_{\partial\Omega} = \alpha(x), \quad x \in \partial\Omega. \quad (2)$$

В дальнейшем независимые переменные будем обозначать как буквами  $(x, y)$ , так и  $(x_1, x_2)$ ,

Опишем построение сетки. Разобьем плоскость  $R$  прямоугольной сеткой с шагами  $h_1$  и  $h_2$ ;  $h_k = 1/N_k$ . Для определенности будем считать, что  $N_1 \leq N_2$ . Точки вида  $(x_m, y_n) = (mh_1, nh_2)$  будем называть узлами сетки и обозначать их  $(m, n)$ . Узлы, лежащие внутри  $\Omega$ , будем называть внутренними узлами и множество таких узлов обозначим  $\Omega_h$ ; узлы, которые лежат на  $\partial\Omega$ , будем называть граничными узлами и множество таких узлов

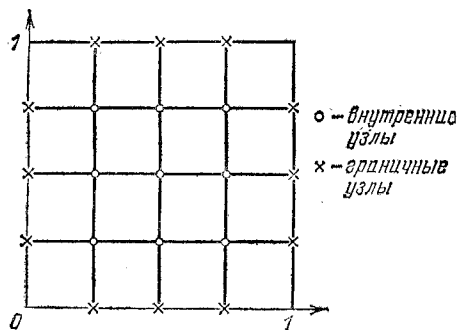


Рис. 10.6.1

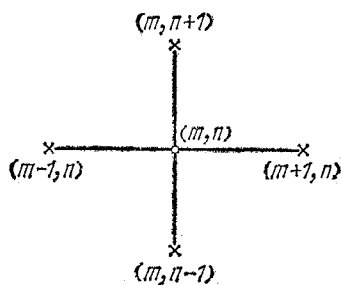


Рис. 10.6.2. Пятиточечный шаблон «крест»;  $\times$  — точки окрестности узла  $(m, n)$

обозначим  $\partial\Omega_h$  (см. рис. 10.6.1). Пусть  $u(x_m, y_n)$  — значение решения задачи в узле  $(m, n)$ .

Сеточную функцию, принимающую в узлах  $(m, n)$  значения  $u_{m, n}$ , будем обозначать  $u^h$ .

Заменяя в (1) производные разностными отношениями, получим систему уравнений

$$-\Delta^h u_{mn} = f_{mn}, \quad m = 1, \dots, N_1 - 1, \quad n = 1, \dots, N_2 - 1; \quad (3)$$

здесь  $\Delta^h = \delta_1^2/h_1^2 + \delta_2^2/h_2^2$ , где операторы  $\delta_1^2$  и  $\delta_2^2$  определены соотношениями

$$\delta_1^2 u_{mn} = u_{m+1, n} - 2u_{mn} + u_{m-1, n}, \quad \delta_2^2 u_{mn} = u_{m, n+1} - 2u_{mn} + u_{m, n-1}.$$

Граничные условия заменим на следующие:

$$u_{mn} |_{\partial\Omega_h} = \alpha(mh_1, nh_2); \quad (mh_1, nh_2) \in \partial\Omega_h. \quad (4)$$

Таким образом,  $u^h$  представляет собой функцию, определенную на сетке  $\Omega_h = \Omega_h \cup \partial\Omega_h$  и удовлетворяющую системе линейных уравнений (3), (4).

Совокупность узлов  $(m, n)$ ,  $(m+1, n)$ ,  $(m-1, n)$ ,  $(m, n+1)$ ,  $(m, n-1)$ , соответствующих значениям  $u^h$ , входящим в уравнение (3), образует шаблон разностной схемы. Уравнение Пуассона (1), таким образом, аппроксимируется на пятиточечном шаблоне «крест». Если  $(m, n)$  — внутренний узел, то окрестностью этого узла будем называть остальные точки шаблона (рис. 10.6.2). Если в окрестности узла есть точки границы, то





поэтому

$$r_{mn} \equiv (-\Delta^h[u]_h)|_{(m,n)} - f_{mn} = -\left(\frac{(\delta_1^2 u)}{h_1^2} + \frac{(\delta_2^2 u)}{h_2^2}\right)\Big|_{(m,n)} - f_{mn} =$$

$$= -\left(\frac{h_1^2}{12} \frac{\partial^4 u}{\partial x_1^4}\Big|_{(m,n)} + \frac{h_2^2}{12} \frac{\partial^4 u}{\partial x_2^4}\Big|_{(m,n)}\right) + o(h_1^2 + h_2^2) = O(h_1^2 + h_2^2).$$

При подстановке точного решения  $u$  в (4) обнаруживаем, что краевые условия (2) выполняются точно;  $r^h$  является погрешностью аппроксимации разностной схемы. Из проведенных рассмотрений следует, что разностная схема (3), (4) имеет второй порядок аппроксимации.

Исследуем разрешимость системы уравнений (3), (4).

**Лемма 1.** Пусть функция  $v^h$  определена на  $\bar{\Omega}_h$  и в узлах  $\Omega_h$  выполняется условие  $\Delta^h v^h \geq 0$ . Тогда хотя бы в одной точке границы  $\partial\Omega_h$  функцией  $v^h$  достигается наибольшее значение.

**Доказательство.** Допустим противное, т. е. что максимальное значение достигается во внутреннем узле (вообще говоря, таких узлов может быть несколько). Среди всех таких узлов выберем тот, у которого наибольшая абсцисса, т. е. узел  $|(m, n)$ , в котором  $v_{mn} > v_{m+1, n}$  и  $v_{mn} = \max_{P \in \bar{\Omega}_h} v^h(P)$ . Тогда, рассматривая  $\Delta^h v^h$  в точке  $(x_m, y_n)$ , получаем

$$\Delta^h v_{mn} = \frac{v_{m-1, n} - 2v_{mn} + v_{m+1, n}}{h_1^2} + \frac{v_{m, n+1} - 2v_{mn} + v_{m, n-1}}{h_2^2} =$$

$$= \frac{(v_{m-1, n} - v_{mn}) + (v_{m+1, n} - v_{mn})}{h_1^2} +$$

$$+ \frac{(v_{m, n+1} - v_{mn}) + (v_{m, n-1} - v_{mn})}{h_2^2} < 0,$$

что противоречит условию леммы; дело заключается в том, что  $v_{m+1, n} - v_{mn} < 0$ , а остальные выражения, стоящие в круглых скобках, неотрицательны, поскольку  $(x_m, y_n)$  — точка максимума. Таким образом, исходная предпосылка является неверной. Утверждение леммы доказано.

Аналогично доказывается

**Лемма 2.** Пусть  $v^h$  определена на  $\bar{\Omega}_h$  и в узлах  $\Omega_h$  выполнено условие  $\Delta^h v^h \leq 0$ . Тогда наименьшее значение достигается функцией  $v^h$ , хотя бы в одной точке границы  $\partial\Omega_h$ .

Из лемм 1 и 2 непосредственно следует

**Теорема 1.** Пусть  $v^h$  определена на  $\bar{\Omega}_h$  и в узлах  $\Omega_h$  удовлетворяет уравнению

$$\Delta^h v_{mn} = 0; \quad m = 1, \dots, N_1 - 1; \quad n = 1, \dots, N_2 - 1.$$

Тогда  $v^h$  достигает своего наибольшего по модулю значения на границе  $\partial\Omega_h$ .

Теорема 1 является разностным аналогом принципа максимума для гармонических функций. Из нее следует, что система уравнений (3), (4) с  $f_{mn} \equiv 0$  и  $\alpha_{mn} \equiv 0$  имеет только нулевое решение, поскольку наибольшее по модулю значение  $u_{mn}$  равно нулю. Следовательно, определитель системы линейных уравнений (3), (4) (число уравнений равно числу неизвестных) отличен от нуля и при любых  $f^h$  и  $\alpha^h$  система (3), (4) имеет единственное решение. Заметим также, что отсюда следует однозначная разрешимость системы уравнений (5) при любой правой части  $\varphi$ .

Повторим и конкретизируем общие построения § 1. Пусть  $U^h$ ,  $F^h$  и  $G^h$  — некоторые пространства функций, определенных на  $\Omega_h$ ,  $\Omega_h$  и  $\partial\Omega_h$ . Введем в них нормы, согласованные с нормами соответствующих пространств в непрерывном случае. Согласно определению (см. § 1) разностная задача (3), (4) будет устойчивой (корректной), если существует постоянная  $c_1$ , не зависящая от  $h_1, h_2$ , такая, что для решения  $u^h$  системы уравнений (3), (4)] справедлива оценка

$$\|u^h\|_{U^h} \leq c_1 (\|f^h\|_{F^h} + \|\alpha^h\|_{G^h}). \quad (7)$$

Исследуем устойчивость схемы (3), (4) и оценим близость  $u$  и  $u^h$ . Пусть нормы в  $U^h$ ,  $F^h$  и  $G^h$  заданы следующим образом:

$$\|u^h\|_{U^h} = \max_{\bar{\Omega}_h} |u_{mn}|, \quad \|f^h\|_{F^h} = \max_{\Omega_h} |f_{mn}|,$$

$$\|\alpha^h\|_{G^h} = \max_{\partial\Omega_h} |\alpha_{mn}|.$$

Вследствие (6) для любого многочлена  $Q(x_1, x_2)$  второй степени выполняется равенство

$$\Delta^h Q_{mn} = \Delta Q|_{(x_m, y_n)},$$

так как четвертые производные, входящие в (6), обращаются в нуль. Возьмем  $R = \sqrt{2}/2 = (\text{diam } \Omega)/2$  и построим вспомогательную функцию

$$Q(x_1, x_2) = \frac{1}{4} \left[ R^2 - \left(x_1 - \frac{1}{2}\right)^2 - \left(x_2 - \frac{1}{2}\right)^2 \right] \|f^h\|_{F^h} + \|\alpha^h\|_{G^h},$$

которую будем рассматривать в узлах сетки  $\bar{\Omega}_h$ . Из вышесказанного следует, что в любом внутреннем узле

$$\Delta^h Q_{mn} = \Delta Q|_{(m, n)} = -\|f^h\|_{F^h},$$

$$m = 1, \dots, N_1 - 1, \quad n = 1, \dots, N_2 - 1.$$

Тогда разность  $v^h = u^h - Q$  в узлах  $\Omega_h$  удовлетворяет неравенству

$$\Delta^h v = f^h + \|f^h\|_{F^h} \geq 0.$$

По лемме 1 функция  $v^h$  принимает наибольшее значение на границе  $\partial\Omega_h$ . Но на  $\partial\Omega_h$  справедливо отношение

$$v^h = \alpha^h - Q = \\ = \alpha^h - \|\alpha^h\|_{G^h} - \frac{1}{4} \left[ R^2 - \left( x_1 - \frac{1}{2} \right)^2 - \left( x_2 - \frac{1}{2} \right)^2 \right] \|f^h\|_{F^h} \leq 0.$$

Таким образом,  $v^h \leq 0$ , т. е.  $u^h \leq Q$  в  $\bar{\Omega}_h$ . Аналогично, рассматривая функцию  $v^h = u^h + Q$  в  $\Omega_h$ , устанавливаем, что

$$\Delta^h v^h \leq 0 \text{ и } v^h|_{\partial\Omega_h} \geq 0.$$

Тогда из леммы 2 следует оценка  $v^h \geq 0$  или же  $u^h \geq -Q$ . Таким образом, всюду в  $\bar{\Omega}_h$  установлено неравенство  $|u^h| \leq Q$  и поэтому

$$\|u^h\|_{U^h} \leq \|Q\|_{U^h} \leq \frac{1}{4} R^2 \|f^h\|_{F^h} + \|\alpha^h\|_{G^h}. \quad (8)$$

Заменяя это неравенство более сильным

$$\|u^h\|_{U^h} \leq (\|f^h\|_{F^h} + \|\alpha^h\|_{G^h}),$$

получаем оценку (7). Таким образом, разностная схема (3), (4) устойчива в сеточном аналоге нормы пространства  $S$ .

Оценим сходимость разностной схемы (3), (4). Для этого запишем уравнение для погрешности  $R_{mn} = u(x_m, y_n) - u_{mn}$

$$-\Delta^h R_{mn} = -\Delta^h [u]_{(m, n)} + \Delta^h u_{mn} = r_{mn};$$

здесь  $r_{mn}$  — погрешность аппроксимации. В силу того что граничные условия выполняются точно, имеем

$$R^h|_{\partial\Omega_h} = 0.$$

Используя оценку (8) при  $R = \sqrt{2}/2$ , получаем

$$\|R^h\|_{U^h} \leq \frac{1}{8} \|r^h\|_{F^h} = O(h_1^2 + h_2^2).$$

Таким образом, решение сеточной задачи (3), (4) сходится к точному решению дифференциальной задачи в сеточной норме пространства  $S$ . Напомним, что все рассуждения проводились в предположении, что решение задачи (1), (2) обладает достаточной гладкостью, а именно что  $u(x, y)$  имеет четвертые производные, ограниченные в  $\bar{\Omega}$ .

Из доказательства сходимости видно, что основным моментом являлось получение оценки (8), характеризующей устойчивость разностной схемы. Сходимость же схемы является следствием аппроксимации и устойчивости, причем порядок скорости сходимости совпадает с порядком аппроксимации. Проведенное выше доказательство сходимости схемы является частным случаем теоремы Филиппова,

Описанный метод дает приближенное решение, сходящееся к точному со скоростью  $O(h^2)$ . По аналогии с одномерным случаем можно построить разностные схемы, обладающие более высоким порядком сходимости. Наметим пути получения более точных схем. Предположим, что решение  $u$  непрерывно дифференцируемо 6 раз в замкнутой области  $\bar{\Omega}$ . Тогда вместо (6) можно написать равенство

$$\frac{\delta_k^2 u}{h_k^2} \Big|_{(m, n)} = \frac{\partial^2 u}{\partial x_k^2} \Big|_{(m, n)} + \frac{h_k^2}{12} \frac{\partial^4 u}{\partial x_k^4} \Big|_{(m, n)} + O(h_k^4).$$

Дифференцируя (1) два раза по  $x$ , получаем

$$\frac{\partial^4 u}{\partial x^4} = -\frac{\partial^4 u}{\partial x^2 \partial y^2} - \frac{\partial^2 f}{\partial x^2}.$$

Таким образом,

$$\frac{\delta_1^2 u}{h_1^2} \Big|_{(m, n)} = \frac{\partial^2 u}{\partial x^2} \Big|_{(m, n)} - \frac{h_1^2}{12} \left( \frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^2 f}{\partial x^2} \right) \Big|_{(m, n)} + O(h_1^4).$$

Заменяя производные в правой части разностными отношениями, имеем

$$\frac{\delta_1^2 u}{h_1^2} \Big|_{(m, n)} = \frac{\partial^2 u}{\partial x^2} \Big|_{(m, n)} - \frac{h_1^2}{12} \left( \frac{\delta_1^2 \delta_2^2 u}{h_1^2 h_2^2} + \frac{\delta_1^2 f}{h_1^2} \right) \Big|_{(m, n)} + O(h_1^4 + h_2^4).$$

Аналогично устанавливаем, что

$$\frac{\delta_2^2 u}{h_2^2} \Big|_{(m, n)} = \frac{\partial^2 u}{\partial y^2} \Big|_{(m, n)} - \frac{h_2^2}{12} \left( \frac{\delta_1^2 \delta_2^2 u}{h_1^2 h_2^2} + \frac{\delta_2^2 f}{h_2^2} \right) \Big|_{(m, n)} + O(h_1^4 + h_2^4).$$

Складывая полученные равенства, получаем

$$\begin{aligned} \left( \frac{\delta_1^2 u}{h_1^2} + \frac{\delta_2^2 u}{h_2^2} \right) \Big|_{(m, n)} &= \Delta u \Big|_{(m, n)} - \frac{h_1^2 + h_2^2}{12} \cdot \frac{\delta_1^2 \delta_2^2 u}{h_1^2 h_2^2} \Big|_{(m, n)} - \\ &\quad - \frac{1}{12} (\delta_1^2 f + \delta_2^2 f) \Big|_{(m, n)} + O(h_1^4 + h_2^4). \end{aligned}$$

Искомая разностная схема четвертого порядка аппроксимации будет иметь вид

$$\begin{aligned} -\Delta^h u_{mn} - \frac{h_1^2 + h_2^2}{12} \cdot \frac{\delta_1^2 \delta_2^2 u_{mn}}{h_1^2 h_2^2} &= \\ = f_{mn} + \frac{1}{12} (\delta_1^2 f_{mn} + \delta_2^2 f_{mn}); & \quad n = 1, \dots, N_1 - 1; \\ & \quad n = 1, \dots, N_2 - 1. \quad (9) \end{aligned}$$

Нетрудно видеть, что шаблон схемы (9) состоит из 9 точек (рис. 10.6.3).

Схемы более высокого порядка, в отличие от одномерного случая, будут содержать тем большее количество узлов, чем выше порядок аппроксимации разностной схемы. В случае, если область является объединением прямоугольников со сторонами, параллельными осям координат, и  $h_1 = h_2$ , при гладком решении рассмотренная схема будет иметь четвертый порядок сходимости. Более точно, если решение  $u$  исходной задачи имеет шестые ограниченные производные в замкнутой области, то справедлива оценка

$$\max_{\bar{\Omega}_h} |u_{mn} - u(x_m, y_n)| \leq ch^4,$$

где  $u_{mn}$  — решение системы сеточных уравнений (9), (4).

Мы рассмотрели построение и исследование разностной схемы в случае, когда  $\Omega$  — квадрат. Построение разностной схемы и исследование ее свойств для случая, когда  $\Omega$

является объединением конечного числа прямоугольников со сторонами, параллельными осям координат, проводится полностью аналогично. Дополнительные трудности возникают, когда область имеет криволинейную границу.

Рассмотрим наиболее простые методы аппроксимации граничных условий в случае области с криволинейной границей. Ограничимся рассмотрением равномерного шага, т. е.  $h_1 = h_2 = h$ . Рассмотрим множество прямых  $x = mh$ ,  $y = nh$ . Точки их пересечения между собой, а также точки их пересечения с  $\partial\Omega$  будем называть узлами. Через  $\Gamma_h$  обозначим узлы, лежащие на  $\partial\Omega$ , а через  $\partial\Omega_h$  — узлы с целыми координатами  $(m, n)$ , лежащие в  $\bar{\Omega}$ , расстояние от которых до границы  $\partial\Omega$ , измеряемое вдоль направления какой-либо из осей координат, меньше  $h$ . Остальные узлы сетки, лежащие в  $\Omega$ , обозначим  $\Omega_h$ . Тогда в каждом узле  $(m, n)$  (т. е. в точке с координатами  $(mh, nh)$ ) сеточной области  $\Omega_h$  можно записать уравнение

$$-\Delta^h u_{mn} = f_{mn}, \quad (m, n) \in \Omega_h. \quad (10)$$

Простейший способ аппроксимации граничных условий заключается в сносе граничных условий в узлы  $\partial\Omega_h$ , т. е. полагаем

$$u_{mn} = \alpha(x, y), \quad (m, n) \in \partial\Omega_h, \quad (11)$$

где  $(x, y)$  — ближайшая к узлу  $(m, n)$  точка границы. В этом случае  $\Omega_h$  — внутренние узлы сеточной области, а  $\partial\Omega_h$  — граничные. Нетрудно видеть, что при таком способе задания граничных условий порядок аппроксимации будет  $O(h)$ . Исследование

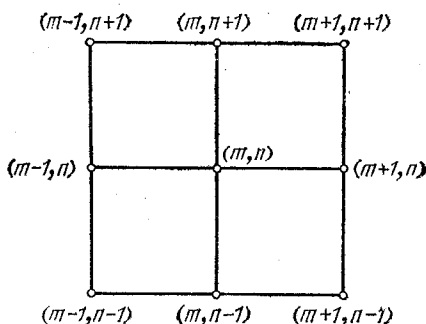


Рис. 10.6.3. Шаблон схемы 4-го порядка

устойчивости и сходимости разностной схемы (10), (11) проводится точно так же, как для схемы (3), (4), однако в данном случае получается порядок сходимости  $O(h)$ . Это является следствием довольно грубой аппроксимации граничных условий.

Рассмотрим еще один способ аппроксимации граничных условий. Назовем узлы  $\partial\Omega_h$  приграничными, а узлы  $\Gamma_h$  — границей сеточной области. В узлах  $\Omega_h$  уравнение (1) заменим уравнением (10); в  $\Gamma_h$  положим  $u^h|_{\Gamma_h} = \alpha^h$ , т. е. в этом случае граничные условия выполняются точно. Пусть  $(m, n)$  — узел  $\partial\Omega_h$ . Для определенности будем считать, что узлы  $(m+1, n)$ ,  $(m, n+1)$ ,  $(m, n-1)$  не лежат на  $\Gamma_h$  (отрезки, соединяющие их с узлом  $(m, n)$ , принадлежат  $\Omega$ ), а узел  $(m-1, n)$  лежит на  $\Gamma_h$ . (В этом случае узел  $(m-1, n)$  соответствует точке  $((m-1)h\theta, nh) \in \partial\Omega$ ;  $0 < \theta < 1$  (см. рис. 10.6.4)). Тогда аппроксимация уравнения (1) в приграничном узле  $(m, n)$  берется в виде

$$\Delta^h u_{mn} = \frac{1}{h} \left( \frac{u_{m+1, n} - u_{mn}}{h} - \frac{u_{mn} - u_{m-1, n}}{h^*} \right) + \frac{u_{m, n+1} - 2u_{mn} + u_{m, n-1}}{h^2}. \quad (12)$$

Здесь  $h^*$  — расстояние между узлами  $(m-1, n)$  и  $(m, n)$ . Аналогично осуществляется аппроксимация уравнения (1) в других узлах  $\partial\Omega_h$ . Итак, в узлах  $\Omega_h$  уравнение (1) аппроксимируется обычным образом, а аппроксимация на неравномерной сетке используется только в узлах  $\partial\Omega_h$ . Поэтому узлы  $\Omega_h$  называют *регулярными*, а узлы  $\partial\Omega_h$  — *нерегулярными*. Аппроксимация уравнения (1) в нерегулярных узлах в данном случае имеет порядок  $O(1)$ .

Для рассмотренной схемы имеет место

**Теорема 2** (без доказательства). *Если решение задачи (1)  $u \in C^4(\bar{\Omega})$ , то разностная схема с аппроксимацией (12) в нерегулярных узлах имеет второй порядок сходимости в сеточной норме пространства  $S$ , т. е.*

$$\|u^h - u\|_{V^h} \leq ch^2.$$

**Замечание.** Довольно распространенным является другой способ аппроксимации уравнения (1) в нерегулярных узлах. А именно, полагают (см. рис. 10.6.4)

$$\Delta^h u_{mn} = \frac{2}{h + h^*} \left( \frac{u_{m+1, n} - u_{mn}}{h} - \frac{u_{mn} - u_{m-1, n}}{h^*} \right) + \frac{\delta_2^2 u_{mn}}{h^2}. \quad (13)$$

В этом случае погрешность аппроксимации в таких узлах имеет порядок  $O(h)$  и теорема 2 будет справедлива. Однако если си-

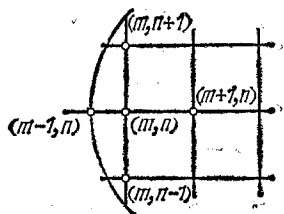


Рис. 10.6.4. Аппроксимация в приграничных узлах

стему линейных уравнений путем исключения граничных значений привести к виду (5), то получится система уравнений с несимметричной матрицей. Таким образом, в данном случае сеточная задача теряет важное свойство, присущее исходной задаче — симметричность оператора.

При исследовании краевых задач с другими граничными условиями и эллиптическими дифференциальными операторами более общего вида, а также краевых задач для систем уравнений в частных производных принцип максимума, разностный аналог которого использовался при исследовании устойчивости и сходимости разностной схемы, вообще говоря, не имеет места. Кроме этого, часто бывает необходимо оценивать не только близость получаемого приближения к точному решению, но и близость их производных. Все это приводит к необходимости создания методов исследования разностных схем, не использующих принцип максимума. Как и ранее, проиллюстрируем сначала методику исследования на модельной задаче Дирихле для уравнения Пуассона в прямоугольнике (1), (2) и на соответствующей разностной схеме (3), (4).

Пусть  $h_1 = h_2 = h$ . Исключая граничные условия (4) из (3), получаем систему уравнений (5). Обозначим через  $H$  линейное пространство сеточных функций, определенных на  $\Omega_h$ . Таким образом, элементы  $H$  можно рассматривать как векторы размерности  $(N_1 - 1) \times (N_2 - 1)$ ; компоненты этих векторов являются значениями функций в узлах  $\Omega_h$ . Матрица  $L_h$  системы уравнений (5) порождает линейный невырожденный оператор в  $H$ . Тогда если  $u^h = \{u_{mn}\}$ ,  $\varphi^h = \{\varphi_{mn}\}$ , то система уравнений (5) может быть записана в операторной форме:

$$L_h u^h = \varphi^h; \quad (14)$$

здесь  $L^h$  — оператор, соответствующий матрице системы уравнений (5). Введем в  $H$  скалярное произведение

$$(u, v) = \sum_{P \in \Omega_h} h^2 u(P) v(P); \quad u, v \in H. \quad (15)$$

Скалярное произведение (15) согласовано со скалярным произведением функций в  $L_2(\Omega)$ , т. е.

$$\lim_{h \rightarrow 0} ([g_1]_h, [g_2]_h) = \int_{\Omega} g_1 g_2 dx$$

для любых непрерывных функций из  $L_2(\Omega)$ .

Пусть  $\|v\|^2 = (v, v)$ . Справедлива

*Лемма 3. Оператор  $L_h$  является симметричным и положительно определенным на  $H$ , и для него выполнена оценка*

$$\gamma_1 \|v\|^2 \leq (L_h v, v) \leq \gamma_2 \|v\|^2, \quad (16)$$

где  $0 < a_1 \leq \gamma_1$ ,  $a \gamma_2 \leq a_2 h^{-2}$ ;  $a_i$  не зависят от  $h$ .

*Доказательство.* Заметим прежде всего, что симметричность оператора  $L_h$  следует из симметричности соответствующей



щей ему матрицы. Проведем, однако, доказательство этого факта другим методом. Пусть  $v \in H$ ; обозначим через  $\tilde{v}$  сеточную функцию, совпадающую на  $\Omega_h$  с  $v$  и равную нулю на  $\partial\Omega_h$ . Тогда из определения  $L_h$  можно записать

$$(L_h v)_{mn} = (-\Delta^h \tilde{v})_{mn}, \quad 1 \leq m, n \leq N-1.$$

Представим  $L_h$  в виде  $L_h = L_1 + L_2$ , где

$$(L_1 v)_{mn} = \frac{-\tilde{v}_{m+1, n} + 2\tilde{v}_{mn} - \tilde{v}_{m-1, n}}{h^2},$$

$$(L_2 v)_{mn} = \frac{-\tilde{v}_{m, n+1} + 2\tilde{v}_{mn} - \tilde{v}_{m, n-1}}{h^2},$$

т. е.  $L_1$  и  $L_2$  являются одномерными сеточными операторами, соответствующими дифференциальным операторам  $-\frac{\partial^2}{\partial x^2}$  и  $-\frac{\partial^2}{\partial y^2}$ . Покажем, что  $L_h$  симметричен и положительно определен. Пусть для определенности  $k=1$ . Используя формулу суммирования по частям (9.8.14), получим

$$(L_1 v, w) = \sum_{m, n=1}^{N-1} h^2 \frac{2\tilde{v}_{mn} - \tilde{v}_{m+1, n} - \tilde{v}_{m-1, n}}{h^2} \tilde{w}_{mn} =$$

$$= \sum_{n=1}^{N-1} h \left( \sum_{m=1}^{N-1} h \frac{2\tilde{v}_{mn} - \tilde{v}_{m+1, n} - \tilde{v}_{m-1, n}}{h^2} \tilde{w}_{mn} \right) =$$

$$= \sum_{n=1}^{N-1} \sum_{m=1}^N h^2 \left( \frac{\tilde{v}_{mn} - \tilde{v}_{m-1, n}}{h} \right) \left( \frac{\tilde{w}_{mn} - \tilde{w}_{m-1, n}}{h} \right).$$

Функции  $\tilde{v}$  и  $\tilde{w}$  входят в правую часть равенства симметричным образом, поэтому

$$(L_1 w, v) = \sum_{n=1}^{N-1} \sum_{m=1}^N h^2 \left( \frac{\tilde{v}_{mn} - \tilde{v}_{m-1, n}}{h} \right) \left( \frac{\tilde{w}_{mn} - \tilde{w}_{m-1, n}}{h} \right) = (L_1 v, w),$$

т. е. оператор  $L_1$  симметричен. С другой стороны, из разностного аналога теоремы вложения (§ 9.8) получаем

$$(L_1 v, v) = \sum_{n=1}^{N-1} \sum_{m=1}^N h^2 \left( \frac{\tilde{v}_{mn} - \tilde{v}_{m-1, n}}{h} \right)^2 \geq$$

$$\geq \frac{1}{4} \sum_{n=1}^{N-1} h \max_{0 < m < N} |v_{mn}|^2 \geq \frac{1}{4} \|v\|^2.$$

Используя аналогичную оценку для оператора  $L_2$ , получаем

$$(L_h v, v) \geq \frac{1}{2} \|v\|^2,$$

т. е. левая часть (16) доказана. В другую сторону оценка получается намного проще. Поскольку  $(\tilde{v}_{mn} - \tilde{v}_{m-1, n})^2 \leq 2(\tilde{v}_{mn}^2 +$

+  $\bar{v}_{m-1, n}^2$ ), то

$$(L_1 v, v) \leq \sum_{n=1}^{N-1} \sum_{m=1}^N h^2 \frac{2(\bar{v}_{mn}^2 + \bar{v}_{m-1, n}^2)}{h^2} = \frac{4}{h^2} \|v\|^2,$$

откуда

$$(L_h v, v) \leq \frac{8}{h^2} \|v\|^2.$$

Лемма доказана.

З а м е ч а н и е. Может показаться, что  $\gamma_1$  и  $\gamma_2$  оценены довольно грубо; однако эти оценки по порядку являются точными. Если на прямоугольнике со сторонами  $l_1$  и  $l_2$  исследовать спектр оператора  $L_h$ , то минимальным и максимальным собственными числами будут соответственно

$$\lambda_{\min} = 4 \left( \frac{\sin^2 \frac{\pi h_1}{2l_1}}{h_1^2} + \frac{\sin^2 \frac{\pi h_2}{2l_2}}{h_2^2} \right),$$

$$\lambda_{\max} = 4 \left( \frac{\cos^2 \frac{\pi h_1}{2l_1}}{h_1^2} + \frac{\cos^2 \frac{\pi h_2}{2l_2}}{h_2^2} \right).$$

Таким образом, (16) выполняется при  $\gamma_1 = \lambda_{\min}$  и  $\gamma_2 = \lambda_{\max}$ , что по порядку при  $h_1 = h_2$  совпадает с оценками, полученными выше.

Из проведенных выше рассуждений вытекает возможность введения в пространстве  $H$  нормы

$$\|v\|_1^2 = (L_h v, v),$$

которую называют *энергетической*. Название связано с тем, что в непрерывном случае при  $u$ , имеющем смысл отклонения мембраны и  $u|_{\partial\Omega} = 0$ , выражение  $\frac{1}{2}(Lu, u)$  пропорционально потенциальной энергии мембраны.

Будем исследовать устойчивость разностной схемы (14) в  $H$ . Умножим обе части (14) скалярно в  $H$  на  $u^h$ ; применяя (16), получим

$$\|u^h\|_1^2 = (\varphi^h, u^h) \leq \|\varphi^h\| \cdot \|u^h\| \leq \gamma_1^{-1} \|\varphi^h\| \|u^h\|_1.$$

Поэтому справедлива оценка

$$\|u^h\|_1 \leq \gamma_1^{-1} \|\varphi^h\|, \quad (17)$$

что означает устойчивость разностной схемы. Отсюда, в частности, следует, что система уравнений (14) при  $\varphi^h \equiv 0$  имеет лишь тривиальное решение  $u^h \equiv 0$ , т. е. мы еще раз доказали, что (14) разрешима при любой правой части единственным образом.

Оценим скорость сходимости разностной схемы (3), (4) в энергетической норме. Как и ранее, будем предполагать, что решение  $u$  дифференциальной краевой задачи (1), (2) имеет ограниченные четвертые производные в замкнутой области. Тогда

погрешность  $R^h = [u]_h - u^h$  будет удовлетворять уравнению

$$L_h R^h = r \equiv L_h [u]_h - L_h u^h, \quad (18)$$

где  $r = O(h^2)$  — погрешность аппроксимации. Применяя оценку (17), имеем

$$\|R^h\|_1 \leq \gamma_1^{-1} \|r\| = O(h^2).$$

Таким образом, рассматриваемая разностная схема имеет второй порядок сходимости по  $h$  в энергетической норме.

При исследовании скорости сходимости в энергетической норме мы предполагали, что решение имеет ограниченные четвертые производные. Оказывается, что это требование является завышенным и тот же результат можно получить в предположении, что решение обладает только третьими ограниченными производными в  $\bar{\Omega}$ . Это связано с тем, что погрешность аппроксимации имеет дивергентный (дипольный) характер. Используя формулу Тейлора, получаем

$$\frac{u(x_{m+1}, y_n) - 2u(x_m, y_n) + u(x_{m-1}, y_n))}{h^2} = \frac{h}{6} \left[ \frac{\partial^3 u}{\partial x^3} \Big|_{(\xi_{m+1}, y_n)} - \frac{\partial^3 u}{\partial x^3} \Big|_{(\xi_m, y_n)} \right] = \frac{h^2}{6} \left( \frac{\psi_{m+1, n}^{(1)} - \psi_{mn}^{(1)}}{h} \right), \quad m = 1, \dots, N-1;$$

здесь  $x_{i-1} \leq \xi_i \leq x_i$ . Аналогичная формула получается и для разности по второй переменной. Следовательно погрешность будет удовлетворять уравнению (18) с правой частью, равной

$$r_{mn} = \frac{h^2}{6} \left( \frac{\psi_{m+1, n}^{(1)} - \psi_{mn}^{(1)}}{h} + \frac{\psi_{m, n+1}^{(2)} - \psi_{mn}^{(2)}}{h} \right).$$

Умножим обе части (18) скалярно в  $H$  на  $R^h$ . Используя формулы суммирования по частям, правую часть преобразуем к виду

$$\begin{aligned} (r, R^h) &= \frac{h^2}{6} \sum_{m, n=1}^{N-1} h^2 \left( \frac{\psi_{m+1, n}^{(1)} - \psi_{mn}^{(1)}}{h} R_{mn} + \frac{\psi_{m, n+1}^{(2)} - \psi_{mn}^{(2)}}{h} R_{mn} \right) = \\ &= \frac{h^2}{6} \sum_{m, n=1}^{N-1} h^2 \left( \frac{\tilde{R}_{mn} - \tilde{R}_{m-1, n}}{h} \psi_{mn}^{(1)} + \frac{\tilde{R}_{mn} - \tilde{R}_{m, n-1}}{h} \psi_{mn}^{(2)} \right). \end{aligned}$$

Таким образом,

$$|(r, R^h)| \leq \frac{h^2}{6} \|R^h\|_1 (\|\psi^{(1)}\| + \|\psi^{(2)}\|) \leq ch^2 \|R^h\|_1;$$

поэтому  $\gamma_1 \|R^h\|_1^2 \leq ch^2 \|R^h\|_1$ , отсюда следует  $\|R^h\|_1 = O(h^2)$ .

При построении разностной схемы, аппроксимирующей уравнение (1) с краевыми условиями другого рода, можно воспользоваться методами, которые применялись в одномерном случае. Пусть для определенности рассматривается задача для уравнения (1) с краевым условием третьего рода

$$\left( \frac{\partial u}{\partial n} + \theta u \right) \Big|_{\partial\Omega} = \alpha, \quad \theta(s) \geq 0, \quad s \in \partial\Omega. \quad (19)$$

Для определенности рассмотрим аппроксимацию условия (19) на участке границы  $x = 1$ . Тогда, заменяя производную в (19) разностным отношением, получаем

$$\frac{u_{Nn} - u_{N-1, n}}{h} + \theta_{Nn} u_{Nn} = \alpha_{Nn}. \quad (20)$$

Найдем погрешность аппроксимации:

$$\begin{aligned} \frac{u(1, nh) - u(1-h, nh)}{h} + \theta(1, nh)u(1, nh) - \alpha(1, n) &= \\ &= \left( \frac{\partial u}{\partial x} - \frac{h}{2} \frac{\partial^2 u}{\partial x^2} + \theta u - \alpha \right) \Big|_{(1, nh)} + O(h^2) = \\ &= -\frac{h}{2} \frac{\partial^2 u}{\partial x^2} \Big|_{(1, nh)} + O(h^2). \end{aligned}$$

Таким образом, погрешность аппроксимации граничных условий (20) имеет первый порядок по  $h$ . Выражая  $\frac{\partial^2 u}{\partial x^2}$  из уравнения (1) получаем  $\frac{\partial^2 u}{\partial x^2} = -f - \frac{\partial^2 u}{\partial y^2}$ . Тогда

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} \Big|_{(1, nh)} &= - \left( f + \frac{\partial^2 u}{\partial y^2} \right) \Big|_{(1, nh)} = \\ &= -f_{Nn} - \frac{u_{N, n+1} - 2u_{Nn} + u_{N, n-1}}{h^2} + O(h^2). \end{aligned}$$

Поэтому если рассмотреть аппроксимацию граничного условия (19) вида

$$\begin{aligned} \frac{u_{Nn} - u_{N-1, n}}{h} + \theta_{Nn} u_{Nn} - \alpha_{Nn} - \\ - \frac{h}{2} \left( f_{Nn} + \frac{u_{N, n+1} - 2u_{Nn} + u_{N, n-1}}{h^2} \right) = 0, \quad (21) \end{aligned}$$

то в силу проведенных выше построений получаем, что соотношение (21) аппроксимирует краевое условие (19) с порядком  $O(h^2)$ . Приводя подобные члены, преобразуем (21) к виду

$$\frac{2u_{Nn} - u_{N-1, n} - 0,5(u_{N, n+1} + u_{N, n-1})}{h} + \theta_{Nn} u_{Nn} = \alpha_{Nn} + \frac{h}{2} f_{Nn}. \quad (22)$$

Отсюда ясно, как будет записываться аппроксимация граничного условия в других узлах  $\partial\Omega_h$ . С помощью подобных рассуждений получаем аппроксимацию в угловом узле  $(N, 0)$ :

$$\frac{2u_{N0} - u_{N-1, 0} - u_{N1}}{h} + 2\theta_{N0} u_{N0} = 2\alpha_{N0} + hf_{N0}. \quad (23)$$

Заметим, что в данном случае граница сеточной области включает угловые точки.

Другой способ аппроксимации граничного условия (19) опирается на то, что берется другая сетка. Рассмотрим множество узлов

$$\bar{\Omega}_h = \{x = (x_1, x_2); x_k^j = jh - h/2, j = 0, \dots, N+1\}, \quad h = N^{-1},$$

и пусть  $\Omega_h$  — множество узлов сетки, лежащих в  $\Omega$ . Тогда, как и ранее, уравнение (1) можно аппроксимировать в узлах  $\Omega_h$  обычной пятиточечной схемой. Краевые условия (19) на такой сетке будем аппроксимировать следующим образом:

$$\frac{u_{N+1, n} - u_{Nn}}{h} + \theta(Nh, nh) \frac{u_{N+1, n} + u_{Nn}}{2} = \alpha(Nh, nh). \quad (24)$$

Предположим, что решение дифференциальной задачи может быть продолжено за пределы области  $\Omega$  с сохранением свойств гладкости. Тогда выражение (24) аппроксимирует краевое условие (19) на участке границы  $x = 1$  с порядком  $O(h^2)$ . Это можно непосредственно проверить, подставляя  $u$  в (24) и воспользовавшись формулой Тейлора в точке  $(1, nh)$ .

**Метод конечных элементов.** До сих пор рассматривалась разностная задача Дирихле для уравнения Пуассона, которая строилась непосредственно путем замены производных в дифференциальном уравнении разностными отношениями. Такой метод построения разностных схем называют *методом конечных разностей*. Аналогично случаю краевых задач для обыкновенных дифференциальных уравнений рассмотрим способы построения дискретных аппроксимаций, основанные на вариационных и проекционных принципах. Будем рассматривать краевую задачу (1), (2) с однородными граничными условиями.

На множестве непрерывно дифференцируемых функций, обращающихся в нуль на границе  $\partial\Omega$ , введем норму

$$\|v\|_{H^1} = \left\{ \int_{\Omega} \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 \right] dx dy \right\}^{1/2} = \left( \int_{\Omega} (\nabla v)^2 dx dy \right)^{1/2}. \quad (25)$$

Замыкание множества таких функций в этой норме является гильбертовым пространством; обозначим его через  $H^1$ .

Рассмотрим задачу о нахождении минимума функционала

$$\min_{v \in H^1} \Phi(v) = \min_{v \in H^1} \left\{ \int_{\Omega} (\nabla v)^2 dx dy - 2 \int_{\Omega} f v dx dy \right\} \quad (26)$$

на пространстве функций  $H^1$ . Если классическое решение  $u$  задачи (1), (2) при  $\alpha(x) \equiv 0$  существует и принадлежит  $H^1$ , то оно дает минимум функционалу (26). Обратное, вообще говоря, неверно; функция, доставляющая минимум на  $H^1$  функционалу (26), не обязательно должна обладать вторыми производными. Таким образом, задачу нахождения решения (1), (2) можно заменить задачей о нахождении минимума квадратичного функционала (26) на  $H^1$ . Решение, получаемое в результате минимизации функционала (26) на  $H^1$ , называют *обобщенным решением краевой задачи (1), (2)*.

Для построения *вариационно-разностной схемы* воспользуемся методом Ритца. Аппроксимируем  $H^1$  некоторым конечномерным подпространством  $V^h$ ; в методе Ритца за приближенное решение задачи (26) принимают функцию  $v \in V^h$ , минимизирующую функционал (26) на подпространстве  $V^h$ .

Подпространство  $V^h$  построим следующим способом. Пусть

$$\bar{\Omega}_h = \{(x, y); x = mh, y = nh; 0 \leq m, n \leq N\}.$$

Разобьем  $\bar{\Omega}$  на квадратные ячейки со стороной  $h$  и вершинами в узлах  $\bar{\Omega}_h$ . Каждую ячейку  $\Omega_{mn} = \{(x, y), mh \leq x \leq (m+1)h, nh \leq y \leq (n+1)h\}$  разобьем диагональю, проходящей через вершины  $(m, n)$ ,  $(m+1, n+1)$ . Таким образом, вся область  $\bar{\Omega}$  будет разбита на прямоугольные треугольники с катетами, равными  $h$ . Эти треугольники будем называть *элементарными*, а разбиение области  $\bar{\Omega}$  на треугольники — *триангуляцией области  $\bar{\Omega}$* . В качестве подпространства  $V^h$  пространства  $H^1$  возьмем пространство непрерывных в  $\bar{\Omega}$  функций, линейных на каждом элементарном треугольнике и обращающихся в нуль на  $\partial\Omega_h$ . Каждая функция из  $V^h$  однозначно определяется своими значениями в узлах  $\bar{\Omega}_h$  и обратно, каждая сеточная функция из  $H$ , которая принимает в узлах сетки заданные значения, однозначно определяет функцию из  $V^h$ . При этом функция из  $V^h$  называется *кусочно-линейным восполнением сеточной функции*. Таким образом, существует взаимно однозначное соответствие между  $H$  и  $V^h$ , где  $H$  — пространство сеточных функций, определенных на  $\bar{\Omega}_h$  и принимающих нулевые значения на  $\partial\Omega_h$ . Функции  $\varphi_{mn}^h \in H$ , где

$$\varphi_{mn}^h = \begin{cases} 1, & (x, y) = (mh, nh), \\ 0, & (x, y) \neq (mh, nh), \end{cases}$$

образуют базис в  $H$ . Соответствующие им кусочно-линейные функции  $\varphi_{mn}^h$  из  $V^h$ , принимающие значения, равные единице в узле  $(m, n)$  и нулю в других узлах, будут образовывать базис в  $V^h$ .

В качестве приближенного решения задачи (26) возьмем функцию  $v^h \in V^h$ , которая минимизирует функционал (26) на подпространстве  $V^h$ , т. е.

$$\min_{v \in V^h} \Phi(v) = \Phi(v^h). \quad (27)$$

Предположим, что  $v^h$  существует; представим ее в виде

$$v^h = \sum_{i, j=1}^{N-1} v_{ij} \varphi_{ij},$$

где  $v_{ij}$  — неизвестные коэффициенты разложения. Отметим, что  $v_{ij}$ , в силу выбора функций  $\varphi_{ij}$ , является значением  $v^h$  в точке  $(i, j)$ . Таким образом, отыскание приближенного решения состоит в определении коэффициентов  $v_{ij}$ . Выпишем уравнения для определения этих коэффициентов. В точке минимума функ-

ции  $\Phi(\sum v_{ij}\varphi_{ij})$  должны выполняться равенства

$$\frac{\partial \Phi(\sum v_{ij}\varphi_{ij})}{\partial v_{mn}} = 0; \quad m, n = 1, \dots, N-1.$$

Вычислим левую часть этого соотношения:

$$\begin{aligned} \frac{\partial \Phi}{\partial v_{mn}} &= \frac{\partial}{\partial v_{mn}} \int_{\Omega} \left[ \left( \sum v_{ij} \frac{\partial \varphi_{ij}}{\partial x} \right)^2 + \left( \sum v_{ij} \frac{\partial \varphi_{ij}}{\partial y} \right)^2 - \right. \\ &\quad \left. - 2f \sum v_{ij} \varphi_{ij} \right] dx dy = 2 \int_{\Omega} \left[ \sum v_{ij} \left( \frac{\partial \varphi_{ij}}{\partial x} \frac{\partial \varphi_{mn}}{\partial x} + \right. \right. \\ &\quad \left. \left. + \frac{\partial \varphi_{ij}}{\partial y} \frac{\partial \varphi_{mn}}{\partial y} \right) - f \varphi_{mn} \right] dx dy. \end{aligned}$$

Следовательно, система уравнений относительно  $v_{ij}$  будет иметь вид

$$\begin{aligned} \sum_{i,j=1}^{N-1} v_{ij} \int_{\Omega} \left( \frac{\partial \varphi_{ij}}{\partial x} \frac{\partial \varphi_{mn}}{\partial x} + \frac{\partial \varphi_{ij}}{\partial y} \frac{\partial \varphi_{mn}}{\partial y} \right) dx dy = \\ = \int_{\Omega} f \varphi_{mn} dx dy, \quad m, n = 1, \dots, N-1. \quad (28) \end{aligned}$$

Количество уравнений в (28) совпадает с количеством неизвестных.

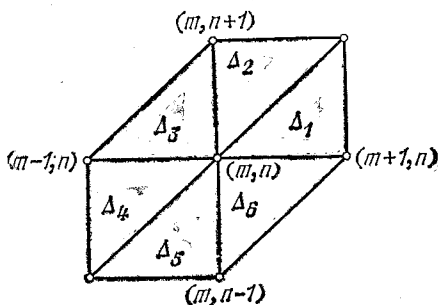


Рис. 10.6.5. Носитель функции  $\varphi_{mn}$

Функция  $\varphi_{mn}$  отлична от нуля лишь в тех элементарных треугольниках, которые имеют узел  $(m, n)$  своей вершиной. Поэтому в каждом из уравнений (28) интегрирование ведется не по всей области  $\Omega$ , а лишь по пересечению таких треугольников с  $\Omega$ . Множество точек, где  $\varphi_{mn} \neq 0$ , образует шестиугольник (см. рис. 10.6.5). Обозначим этот шестиугольник через  $S_{mn}$ , а входящие в него треугольники — через  $\Delta_1, \dots, \Delta_6$ . Положим

$$J_{mn}^1(i, j) = \int_{\Omega} \frac{\partial \varphi_{mn}}{\partial x} \frac{\partial \varphi_{ij}}{\partial x} dx dy;$$

поскольку  $\frac{\partial \varphi_{mn}}{\partial x} \equiv 0$  в  $\Delta_2$  и  $\Delta_5$ , то

$$J_{mn}^1(i, j) = \int_{\Omega} \frac{\partial \varphi_{mn}}{\partial x} \frac{\partial \varphi_{ij}}{\partial x} dx dy = \\ = \int_{\Delta_1 \cup \Delta_4} \frac{\partial \varphi_{mn}}{\partial x} \frac{\partial \varphi_{ij}}{\partial x} dx dy + \int_{\Delta_3 \cup \Delta_4} \frac{\partial \varphi_{mn}}{\partial x} \frac{\partial \varphi_{ij}}{\partial x} dx dy.$$

Отсюда видно, что  $J_{mn}^1 \neq 0$  лишь при  $j = n$  и  $i = m - 1, i = m, i = m + 1$ . Проводя соответствующие вычисления, получаем

$$J_{mn}^1(m, n) = \int_{\Delta_1 \cup \Delta_4} \left( \frac{\partial \varphi_{mn}}{\partial x} \right)^2 dx dy + \int_{\Delta_3 \cup \Delta_4} \left( \frac{\partial \varphi_{mn}}{\partial x} \right)^2 dx dy = 2,$$

$$J_{mn}^1(m+1, n) = \\ = \int_{\Delta_1 \cup \Delta_4} \frac{\partial \varphi_{mn}}{\partial x} \frac{\partial \varphi_{m+1n}}{\partial x} dx dy = -1, \quad J_{mn}^1(m-1, n) = -1.$$

Аналогично для  $J_{mn}^2(i, j) = \int_{\Omega} \frac{\partial \varphi_{mn}}{\partial y} \frac{\partial \varphi_{ij}}{\partial y} dx dy$  получаем

$$J_{mn}^2(m, n) = 2, \quad J_{mn}^2(m, n+1) = -1, \quad J_{mn}^2(m, n-1) = -1;$$

в остальных случаях  $J_{mn}^2(i, j) = 0$ .

Таким образом, уравнение, соответствующее точке  $(m, n)$ , записывается в виде

$$4v_{mn} - v_{m+1, n} - v_{m-1, n} - v_{m, n+1} - v_{m, n-1} = h^2 g_{mn},$$

где  $g_{mn} = \frac{1}{h^2} \int_{\Omega_{mn}} f \varphi_{mn} dx dy$ . Разделив обе части на  $h^2$ , получаем систему сеточных уравнений

$$L_h v_{mn} = g_{mn}, \quad 1 \leq m, \quad n \leq N-1, \quad (29)$$

структура которой полностью совпадает с (5). Единственное отличие заключается в другом способе вычисления правой части. Однако если  $g_{mn}$  вычислять приближенно, полагая

$$\frac{1}{h^2} \int_{\Omega_{mn}} f \varphi_{mn} dx dy \approx f_{mn} \frac{1}{h^2} \int_{\Omega_{mn}} \varphi_{mn} dx dy = f_{mn},$$

то получится разностная схема, полностью идентичная (5).

Так как левая часть системы (29) совпадает с левой частью системы (5), то система уравнений (29) имеет единственное решение при любой  $f$ .

Построение разностных схем таким способом весьма целесообразно в случае уравнений и систем с естественными граничными условиями, когда непосредственная аппроксимация граничных условий вызывает затруднения.



В последнее время получили широкое распространение *проекционно-разностные методы* решения краевых задач (*метод конечного элемента*). Описанный выше метод построения разностных схем с помощью метода Рунца является одной из разновидностей метода конечных элементов.

Опишем в общих чертах суть проекционно-разностного подхода на другой модельной задаче (ср. с § 9.11). За основу метода обычно берется интегральное тождество для определения обобщенного решения. Итак, предположим, что в квадрате  $\bar{\Omega} = \{x = (x_1, x_2), 0 \leq x_1, x_2 \leq 1\}$  требуется найти решение краевой задачи

$$-\Delta u = f, \quad \frac{\partial u}{\partial n} + \alpha u|_{\Gamma} = 0, \quad u|_{\partial\Omega \setminus \Gamma} = 0, \quad \alpha > 0, \quad (30)$$

где  $\Gamma$  — участок границы, лежащий на прямой  $x = 1$ . Предположим, что решение этой задачи существует. Умножим обе части уравнения (30) на функцию  $\varphi$ , частные производные которой являются кусочно-непрерывными,  $\varphi|_{\partial\Omega \setminus \Gamma} = 0$ . Интегрируя по частям и используя краевые условия, получим

$$\int_{\Omega} \nabla u \nabla \varphi \, dx \, dy + \int_{\Gamma} \alpha(s) u(s) \varphi(s) \, ds = \int_{\Omega} f \varphi \, dx \, dy. \quad (31)$$

Соотношение (31) называется *интегральным тождеством*; оно имеет место для любой функции  $\varphi$ , обладающей кусочно-непрерывными производными и обращающейся в нуль на  $\partial\Omega \setminus \Gamma$ . Если  $u$  является классическим решением задачи (30) и имеет суммируемые в квадрате производные, то оно удовлетворяет (31). Обратное, вообще говоря, неверно. Можно указать функции  $u$  и  $f$ , для которых выполнено (31), однако классическое решение задачи (30) существовать не будет. Функцию  $u$ , имеющую первые производные, интегрируемые в квадрате, и удовлетворяющую интегральному тождеству (30), называют *обобщенным решением* задачи (30). Встает вопрос о связи обобщенного решения, определяемого из интегрального тождества (31), с обобщенным решением, определяемым минимизацией соответствующего функционала типа (26). Оказывается, что в данном случае эти обобщенные решения совпадают. Это всегда имеет место в тех случаях, когда исходный дифференциальный оператор является симметричным и положительно определенным. Если эти условия не выполнены, то задача определения обобщенного решения не может быть сформулирована в терминах минимизации некоторого квадратичного функционала. Таким образом, определение обобщенного решения краевой задачи для дифференциального уравнения при помощи интегрального тождества является более общим.

Поступим аналогично предыдущему случаю. Пусть  $H^1$  — пространство, являющееся замыканием множества гладких функций, равных нулю на  $\partial\Omega \setminus \Gamma$  в норме (25). Триангулируем

область  $\bar{\Omega}$  и введем пространство  $V^h$  функций, кусочно-линейных на элементарных треугольниках и обращающихся в нуль на  $\partial\Omega \setminus \Gamma$ . Приближенным решением задачи (31) назовем функцию  $u^h \in V^h$  такую, что для любой  $\varphi \in V^h$  выполняется равенство

$$\int_{\Omega} \nabla u^h \nabla \varphi \, dx \, dy + \int_0^1 \alpha(y) u^h(1, y) \varphi(1, y) \, dy = \int_{\Omega} f \varphi \, dx \, dy. \quad (32)$$

Таким образом, интегральное тождество (31) рассматривается на подпространстве функций из  $H^1$ .

Функция  $u^h$  полностью определяется своими значениями в узлах сетки. Для того чтобы  $u^h$  удовлетворяла (32), необходимо и достаточно, чтобы (32) было справедливо для любой функции  $\varphi \in V^h$ , входящей в базис  $V^h$ . В качестве базисных функций возьмем функции из  $V^h$ , которые равны единице в одном из узлов  $\Omega_h \cup \Gamma_h$  и нулю во всех остальных узлах. Это дает нам систему сеточных уравнений. Обратим внимание на тот факт, что теперь в базис входят функции, которые, вообще говоря, отличны от нуля на  $\Gamma$ .

Представим  $u^h$  в виде  $u^h = \sum u_{mn} \varphi_{mn}$ , где  $u_{mn}$  — неизвестные коэффициенты. Подставляя это выражение в (32), получаем

$$\begin{aligned} \sum_{m, n} u_{mn} \int_{\Omega} \nabla \varphi_{mn} \nabla \varphi_{ij} \, dx \, dy + \sum_{m, n} u_{mn} \int_0^1 \alpha(y) \varphi_{mn}(1, y) \varphi_{ij}(1, y) \, dy = \\ = \int_{\Omega} f \varphi_{ij} \, dx \, dy; \quad 1 \leq i \leq N, \quad 1 \leq j \leq N-1. \end{aligned} \quad (33)$$

Совокупность соотношений (33) образует систему линейных алгебраических уравнений относительно неизвестных коэффициентов  $u_{mn}$  и порождает некоторую проекционно-разностную схему (эта схема могла бы быть получена и как вариационно-разностная). Так как любая функция  $\varphi \in V^h$  может быть разложена по функциям  $\varphi_{mn}$ , то из выполнения (33) будет следовать справедливость (32). Таким образом, выражения (32) и (33) эквивалентны.

Для доказательства разрешимости (33) положим в (32)  $\varphi = u^h$ . Тогда

$$\begin{aligned} \|u^h\|_1^2 &\leq \int_{\Omega} (\nabla u^h)^2 \, dx \, dy + \int_0^1 \alpha(y) [u^h(1, y)]^2 \, dy \leq \\ &\leq \left| \int_{\Omega} f u^h \, dx \, dy \right| = |(f, u^h)| \leq \|f\|_{-1} \cdot \|u^h\|_1, \end{aligned}$$

где

$$\|f\|_{-1} = \sup_{\varphi \in H^1} \frac{|(f, \varphi)|}{\|\varphi\|_1}.$$

Таким образом,  $\|u^h\|_1 \leq \|f\|_{-1}$ . Это означает устойчивость проекционно-разностной схемы, т. е. малым изменениям  $f$  в норме  $\|\cdot\|_{-1}$  соответствует малое изменение  $u^h$  в норме  $\|\cdot\|_1$ .

Рассмотрим структуру матрицы системы уравнений (33). Если  $1 \leq i \leq N-1$ , то  $\varphi_{ij}|_{\Gamma} = 0$ , поэтому интеграл по  $\Gamma$  в (33) равен нулю. В этом случае выражение (33) совпадает с (28), и, следовательно, уравнение, соответствующее функции  $\varphi_{ij}$ , имеет вид

$$4u_{ij} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} = h^2 g_{ij},$$

где

$$g_{ij} = \frac{1}{h^2} \int_{\Omega} f \varphi_{ij} dx dy.$$

В точках  $\Gamma$ , т. е. в узлах вида  $(N, j)$ , имеем

$$\frac{2u_{Nj} - u_{N,j-1} - u_{N,j+1}}{2} + u_{Nj} - u_{N-1,j} + (\alpha'_{j-1} u_{N,j-1} + \alpha'_j u_{Nj} + \alpha'_{j+1} u_{N,j+1}) = h^2 g_{Nj}.$$

Здесь  $\alpha'_n = \int_{\Gamma} \varphi_{Nj} \varphi_{Nn} \alpha(y) dy$ ,  $h^2 g_{Nj} = \int_{\Omega} f \varphi_{Nj} dx dy$ . В случае когда  $\alpha'_n$  и  $g_{ij}$  не вычисляются в явном виде, их можно вычислять приближенно, используя квадратурные формулы. Например,

$$\alpha'_n = \int_{\Gamma} \varphi_{Nj} \varphi_{Nn} \alpha ds \approx \frac{\alpha(jh) + \alpha(nh)}{2} \int_{\Gamma} \varphi_{Nj} \varphi_{Nn} ds,$$

$$h^2 g_{ij} = \int_{\Omega} f \varphi_{ij} dx dy \approx h^2 f_{ij}.$$

В заключение кратко опишем построение проекционно-разностной схемы в случае криволинейной границы и возможные обобщения этого метода. Пусть для простоты рассматривается задача Дирихле для уравнения Пуассона и  $\Omega$  — плоская односвязная область с кусочно-гладкой границей  $\partial\Omega$ , т. е.  $\partial\Omega$  состоит из конечного числа гладких дуг, пересекающихся между собой под ненулевыми углами. Зададимся параметром  $h$  — шагом сетки. Построим ломаную  $\Gamma_h$ , обладающую следующими свойствами:

- область  $\Omega_h$ , ограниченная  $\Gamma_h$ , содержится в  $\bar{\Omega}$ ;
- между точками  $\partial\Omega$  и  $\Gamma_h$  можно установить взаимно однозначное соответствие, т. е. существует взаимно однозначное отображение  $\varphi: \partial\Omega \rightarrow \Gamma_h$ , которое имеет кусочно-непрерывную производную  $\varphi'$ ,  $|\varphi'|, |(\varphi^{-1})'| \leq C$ , где  $C$  не зависит от  $h$ ;
- расстояние от точек  $\Gamma_h$  до  $\partial\Omega$  не превосходит величины  $c_1 h^2$ , где  $c_1 > 0$  — некоторая постоянная, не зависящая от  $h$ ;
- длина звеньев ломаной ограничена снизу величиной  $c_2 h$ ,  $c_2 > 0$ . При принятых допущениях относительно области такое построение всегда возможно.

Разобьем область  $\Omega_h$  на треугольники так, что:

- длины треугольников сторон лежат в пределах  $[c_3 h, c_4 h]$ ,  $c_1 > 0$  — постоянные, не зависящие от  $h$ ;
- площади треугольников лежат в пределах  $[c_5 h^2, c_6 h^2]$ ;
- любые два треугольника либо не пересекаются, либо имеют только одну общую сторону, либо только общую вершину.

Описанное выше разбиение также называется триангуляцией области  $\Omega$ ; вершины треугольников называются узлами сетки. Можно доказать, что такая триангуляция осуществима. Пусть  $H$  — пространство функций, кусочно-линейных над треугольниками  $\Omega_h$  и обращающихся в нуль на  $\Gamma_h$ . Задаче (1) с однородным граничным условием (2) поставим в соответствие проекционно-разностную задачу — найти функцию  $u^h \in H$ , удовлетворяющую при любой  $\varphi \in H$  соотношению

$$\int_{\Omega_h} \nabla u^h \nabla \varphi \, dx \, dy = \int_{\Omega_h} f \varphi \, dx \, dy. \quad (34)$$

Функция  $u^h$  полностью определяется своими значениями в узлах сетки. Поэтому если в качестве  $\varphi$  брать базисные функции пространства  $H$  (равные единице в одном узле и нулю в остальных узлах и кусочно-линейные над треугольниками  $\Omega_h$ ), то (34) является системой линейных алгебраических уравнений относительно значений  $u^h$  в узлах. Доказательство устойчивости проводится так же, как и выше. При исследовании же сходимости следует дополнительно оценить норму решения в  $W_2^1$  в приграничной полосе  $\Omega \setminus \Omega_h$ . С учетом этих оценок можно получить следующее соотношение:

$$\|u - u^h\|_{W_2^1(\Omega_h)} \leq ch,$$

где постоянная  $c$  зависит от нормы решения в  $W_2^2(\Omega)$ .

При построении проекционно-разностной схемы можно использовать более сложные конечные элементы, за счет чего может быть достигнута большая точность. Например, кроме узлов сетки, можно также в качестве узлов рассматривать середины сторон треугольников. Пусть  $H$  — пространство функций, кусочно-квадратичных на элементарных треугольниках  $\Omega_h$  и равных нулю на  $\partial\Omega_h$ , т. е. если функция  $\varphi \in H$ , то она является полиномом второй степени в каждом треугольнике и непрерывна в  $\Omega_h$ . Пусть  $\partial\Omega_h = \partial\Omega$ . В качестве базисных функций в  $H$  можно рассмотреть функции, которые равны единице в одном узле, а в других узлах равны нулю и принадлежат  $H$  (под узлами здесь понимаются узлы сетки и середины сторон треугольников). Тогда можно получить оценку

$$\|u - u^h\|_{W_2^1} \leq ch^2.$$

Аналогично могут быть построены проекционно-разностные схемы с более высоким порядком скорости сходимости. При этом используются способы интерполяции, рассмотренные в § 5.5. Заметим, что структура матрицы линейных алгебраических уравнений при этом ухудшается, а именно возрастает количество ненулевых элементов в строке матрицы и ширина ленты.

Кратко осветим историю вопроса применения описанных методов. Вариационные и проекционные методы при небольшом числе базисных функций применялись издавна, еще до появления ЭВМ. Применение ЭВМ позволило увеличить число базисных функций; при этом часто возрастало суммарное влияние вычислительной погрешности и погрешности, возникающей при аппроксимации интегралов квадратурными суммами.

Это обстоятельство ставило ограничение на точность, с которой могли быть получены решения с помощью вариационных методов. Были проведены теоретические исследования, показавшие, что для устойчивости вариационных методов существенно выполнение некоторого условия на систему базисных функций, называемого условием *сильной минимальности*. Построение

системы базисных функций, удовлетворяющей этому условию, в случае областей сложной формы иногда бывает непросто.

Параллельно шло бурное развитие теории и практики применения конечно-разностных методов. Если при использовании классических вариационных методов для решения линейных задач возникают линейные системы уравнений с полностью заполненной матрицей, то при использовании конечно-разностных уравнений возникают системы уравнений с матрицей, содержащей относительно малое число ненулевых элементов. Это обстоятельство позволяет решать с теми же затратами машинного времени системы уравнений с существенно большим числом неизвестных. Однако в случае областей сложной формы применение конечно-разностных методов представляет определенные неудобства вследствие неоднородности построения разностных уравнений в приграничных точках.

Получивший в последнее время бурное развитие метод конечных элементов свободен от ряда недостатков описанных методов: он не требует специальных усилий по построению системы базисных функций, являющейся сильно минимальной, при его использовании упрощается написание уравнений вблизи границы. Матрица линейной системы уравнений содержит относительно малое число ненулевых элементов. Большая «технологичность» метода позволила создать на его основе ряд *промышленных систем стандартных программ* решения краевых задач, в частности задач теории упругости. При использовании таких систем не требуется знание теории численных методов и тонкостей программирования. Пользователь должен лишь задать триангуляцию области, а часто система и сама осуществляет такую триангуляцию. Эти методы сходятся при меньших требованиях гладкости, чем конечно-разностные методы.

В то же время увеличивается объем работы при вычислении матрицы системы уравнений. Поэтому при решении задач большого объема зачастую применяют все-таки конечно-разностные методы или приходят к составлению систем уравнений с помощью аппроксимации минимизирующего функционала (или интегрального тождества) (см. § 9.12).

Традиционно для решения эллиптических задач применялись методы теории потенциала. С появлением ЭВМ они были практически вытеснены конечно-разностными методами. Однако в последнее время в вычислительную практику стал интенсивно проникать *метод граничных элементов*, имеющий некоторые общие черты с методом потенциала.

## § 7. Решение параболических уравнений с несколькими пространственными переменными

При решении параболических уравнений, как и в случае эллиптических уравнений, переход от одномерного случая к многомерному вызывает качественные затруднения. Поскольку все

принципиальные трудности возникают лишь при переходе от одной пространственной переменной к двум, то в дальнейшем будем рассматривать случай двух пространственных переменных.

Перейдем к построению и исследованию разностных схем. Пусть требуется найти функцию  $u$ , являющуюся решением уравнения

$$\frac{\partial u}{\partial t} = \Delta u + f(x, t), \quad (1)$$

в области  $Q_T = \Omega \times [0, T]$ ,  $\bar{\Omega} = \{x; 0 \leq x_i \leq 1, i = 1, 2\}$  с начальными и граничными условиями

$$u(x, 0) = u_0(x), \quad u(x, t)|_{x \in \Gamma} = \alpha(x, t); \quad (2)$$

здесь  $x = (x_1, x_2)$ ,  $\Gamma = \partial\Omega \times [0, T]$ .

Попытаемся применить к решению задачи (1), (2) методы, разработанные ранее. Введем в рассмотрение квадратную сетку с шагом  $h = 1/M$ :

$$\bar{\Omega}_h = \{x; x = (ih, jh), 0 \leq i, j \leq M\},$$

а на отрезке  $[0, T]$  — сетку с шагом  $\tau = T/N$ . Будем искать приближенное решение задачи (1), (2) в дискретной системе точек (узлов)

$$\bar{Q}_h = \{(x, t); x \in \bar{\Omega}_h, t = n\tau, n = 0, \dots, N\}.$$

Множество  $\bar{Q}_h$  будем называть *сеточной областью*, а множество точек  $\bar{\Omega}_h^n = \{(x, t) \in \bar{Q}_h, x \in \Omega_h, t = n\tau\}$  при фиксированном  $t = n\tau$  будем называть  $n$ -м слоем.

По аналогии с одномерным случаем построим явную и неявную разностные схемы для задачи (1), (2) и попытаемся выяснить, в чем заключается принципиальное отличие от случая одной пространственной переменной. Заменяя  $\frac{\partial u}{\partial t}$  в узле  $(i, j, n)$  разделенной разностью  $\frac{u_{ij}^{n+1} - u_{ij}^n}{\tau}$  или  $\frac{u_{ij}^n - u_{ij}^{n-1}}{\tau}$ , а  $\Delta u$  — выражением (см. § 6)

$$\Delta^h u_{ij}^n = \frac{u_{i+1, j}^n + u_{i-1, j}^n + u_{i, j+1}^n + u_{i, j-1}^n - 4u_{ij}^n}{h^2},$$

получаем разностные схемы: явную

$$\frac{u_{ij}^{n+1} - u_{ij}^n}{\tau} = \Delta^h u_{ij}^n + f_{ij}^n, \quad 1 \leq i, j \leq M-1, \quad (3)$$

$$u_{ij}^k = \alpha(ih, jh, k\tau), \quad (ih, jh) \in \partial\Omega_h, \quad u_{ij}^0 = u_0(ih, jh)$$

и неявную

$$\frac{u_{ij}^{n+1} - u_{ij}^n}{\tau} = \Delta^h u_{ij}^{n+1} + f_{ij}^{n+1}, \quad (4)$$

$$u_{ij}^k = \alpha(ih, jh, k\tau), \quad (ih, jh) \in \partial\Omega_h, \quad u_{ij}^0 = u_0(ih, jh).$$

В схеме (3) счет ведется по явным формулам — по известным значениям  $u_{ij}^n$  из (3) находятся значения  $u_{ij}^{n+1}$ . Поэтому проблем с реализацией алгоритма на ЭВМ не возникает. Остается лишь исследовать устойчивость этой схемы.

По другому обстоит дело в случае схемы (4). Относительно  $u_{ij}^{n+1}$ ,  $1 \leq i, j \leq M-1$ , имеем систему линейных алгебраических уравнений, т. е. схема (4) неявна. Структура матрицы уравнений (4) совпадает со структурой матрицы оператора  $-\Delta^n$  (см. § 6). Поэтому при решении этой системы возникают те же трудности, что и в случае эллиптических уравнений. Напомним, что в одномерном случае проблема численного решения системы уравнений на верхнем слое не возникала, так как можно было воспользоваться методом прогонки.

Введем понятие экономичной разностной схемы. Разностную схему назовем *экономичной*, если она безусловно устойчива и при переходе от слоя к слою требуется количество арифметических операций, пропорциональное числу узлов на слое. (Иногда условие безусловной устойчивости в определении экономичной схемы отсутствует.) Из определения следует, что чисто неявная схема для одномерного уравнения является экономичной.

Перед тем как заниматься построением экономичных разностных схем, исследуем устойчивость разностной схемы в общей постановке.

Введем пространство  $H$  вектор-функций, заданных на  $\Omega_n$ ;  $v_{ij}$  — значение функции  $v \in H$  в узле  $(i, j)$ . Скалярное произведение и норму в  $H$  определим как

$$(v, w) = \sum_{i,j=1}^{M-1} h^2 v_{ij} w_{ij}, \quad \|v\|^2 = (v, v).$$

Разностные схемы (3), (4), рассматривавшиеся выше, связывали два слоя —  $n$ -й и  $(n+1)$ -й, поэтому их естественно называть *двухслойными*. Далее будем рассматривать двухслойные разностные схемы вида

$$B \frac{u^{n+1} - u^n}{\tau} + Au^n = \varphi^n, \quad (5)$$

где  $B$  и  $A$  — симметричные положительно определенные операторы, отображающие  $H$  в себя. Как и в одномерном случае, иногда будем различать устойчивость по начальным данным и по правой части.

Обозначим  $u_i^n = \frac{u^{n+1} - u^n}{\tau}$ . Учитывая равенство  $u^n = u^{n+1} - \tau u_i^n$ , преобразуем (5) к виду

$$(B - \tau A) u_i^n + Au^{n+1} = \varphi^n. \quad (6)$$

Положим  $D = B - \tau A$ . Так как  $u^{n+1} = \frac{u^{n+1} + u^n}{2} + \frac{\tau}{2} u_t^n$  и оператор  $D$  симметричен, то

$$\begin{aligned} 2\tau (Du_t^n, u^{n+1}) &= \tau (Du_t^n, u^{n+1} + u^n) + \tau^2 (Du_t^n, u_t^n) = \\ &= (Du^{n+1}, u^{n+1}) - (Du^n, u^n) + (Du^{n+1}, u^n) - (Du^n, u^{n+1}) + \\ &\quad + \tau^2 (Du_t^n, u_t^n) = (Du^{n+1}, u^{n+1}) - (Du^n, u^n) + \tau^2 (Du_t^n, u_t^n). \end{aligned}$$

Поэтому, умножая обе части (6) скалярно в  $H$  на  $2\tau u^{n+1}$ , получим

$$\begin{aligned} (u^{n+1}, u^{n+1})_D - (u^n, u^n)_D + \tau^2 (u_t^n, u_t^n)_D + \\ + 2\tau \|u^{n+1}\|_A^2 = 2\tau (\varphi^n, u^{n+1}). \end{aligned} \quad (7)$$

Напомним определения:  $(v, w)_D = (Dv, w)$ ,  $\|v\|_A^2 = (Av, v)$ .

Будем считать, что

$$D = B - \tau A > 0. \quad (8)$$

Тогда (7) можно переписать в виде

$$\|u^{n+1}\|_D^2 - \|u^n\|_D^2 + \tau^2 \|u_t^n\|_D^2 + 2\tau \|u^{n+1}\|_A^2 = 2\tau (\varphi^n, u^{n+1}). \quad (9)$$

Соотношение (9) называется *энергетическим равенством*.

Так как пространство  $H$  конечномерно, то существует постоянная  $\kappa$  такая, что  $(Dv, v) \leq \kappa (Av, v)$  для всех  $v \in H$  или, что то же самое,

$$D \leq \kappa A. \quad (10)$$

Так как  $D = D^T > 0$ , то существует оператор (матрица)  $D^{1/2}$ , симметричный и положительно определенный, такой, что  $D^{1/2} D^{1/2} = D$ . Через  $D^{-1/2}$  обозначим оператор  $(D^{1/2})^{-1}$ . В рассматриваемом случае

$$\|u^{n+1}\|_A^2 \geq \kappa^{-1} \|u^{n+1}\|_D^2,$$

$$\begin{aligned} 2\tau |(\varphi^n, u^{n+1})| &= 2\tau |(D^{-1/2} \varphi^n, D^{1/2} u^{n+1})| \leq \\ &\leq 2\tau \|D^{-1/2} \varphi^n\| \|D^{1/2} u^{n+1}\| \leq 2\tau \|u^{n+1}\|_D \|\varphi^n\|_{D^{-1}} \leq \\ &\leq \varepsilon \tau \|u^{n+1}\|_D^2 + \frac{\tau}{\varepsilon} \|\varphi^n\|_{D^{-1}}^2 \end{aligned}$$

и из (9) следует неравенство

$$\left(1 + \frac{2\tau}{\kappa} - \varepsilon \tau\right) \|u^{n+1}\|_D^2 \leq \|u^n\|_D^2 + \frac{\tau}{\varepsilon} \|\varphi^n\|_{D^{-1}}^2.$$

Фиксируя  $\varepsilon$ , например, полагая  $\varepsilon = \kappa^{-1}$ , откуда получаем соотношение

$$\left(1 + \frac{\tau}{\kappa}\right) \|u^{n+1}\|_D^2 \leq \|u^n\|_D^2 + \tau \kappa \|\varphi^n\|_{D^{-1}}^2, \quad (11)$$

связывающее нормы векторов  $u^n$  на соседних слоях.



Имеем последовательность неравенств

$$\begin{aligned} \|u^{n+1}\|_D^2 &\leq \|u^n\|_D^2 + \tau\kappa \|\varphi^n\|_{D^{-1}}^2 \leq \|u^{n-1}\|_D^2 + \\ &+ \sum_{k=n-1}^n \tau\kappa \|\varphi^k\|_{D^{-1}}^2 \leq \dots \leq \|u^0\|_D^2 + \sum_{k=0}^n \tau\kappa \|\varphi^k\|_{D^{-1}}^2. \end{aligned}$$

Предполагая, что  $\kappa$  не зависит от шагов сетки  $h$  и  $\tau$ , при  $n\tau \leq N$  получаем окончательное соотношение

$$\begin{aligned} \|u^n\|_D^2 &\leq \|u^0\|_D^2 + c \sum_{k=0}^{n-1} \tau \|\varphi^k\|_{D^{-1}}^2 \leq \|u^0\|_D^2 + \\ &+ c \sum_{k=0}^{N-1} \tau \|\varphi^k\|_{D^{-1}}^2 \leq \|u^0\|_D^2 + cT \max_{0 \leq k \leq N-1} \|\varphi^k\|_{D^{-1}}^2, \quad (12) \end{aligned}$$

которое означает устойчивость разностной схемы (5) по начальным данным и по правой части. Таким образом, условие  $D = B - \tau A > 0$  является достаточным для устойчивости разностной схемы по правой части и начальным условиям.

Заметим, что выражение  $\sum_{k=0}^{N-1} \tau \|\varphi^k\|_{D^{-1}}^2$ , стоящее в правой части (12), является квадратурной формулой для интеграла  $\int_0^T \|\varphi(t)\|_{D^{-1}}^2 dt$ , а выражение  $\max_{0 \leq k \leq N-1} \|\varphi^k\|_{D^{-1}}^2$  — сеточным аналогом нормы  $\max_{0 \leq t \leq T} \|\varphi(t)\|_{D^{-1}}^2$ .

Найдем необходимое и достаточное условие того, что собственные числа оператора перехода от слоя к слою не превосходили единицы. Это будет означать, что схема устойчива по начальным данным и норма погрешности не возрастает при переходе от слоя к слою. Для этого положим  $\varphi^k \equiv 0$ .

Применим к обеим частям (5) оператор  $B^{-1/2}$ ; получим

$$B^{1/2}u^{n+1} = B^{1/2}u^n - \tau B^{-1/2}Au^n.$$

Обозначим  $B^{1/2}u^n = y^n$ . Тогда  $u^n = B^{-1/2}y^n$  и последнее равенство примет вид

$$y^{n+1} = y^n - \tau B^{-1/2}AB^{-1/2}y^n = (E - \tau B^{-1/2}AB^{-1/2})y^n.$$

Оператор  $S = E - \tau B^{-1/2}AB^{-1/2}$  симметричен, поэтому собственные числа  $S$  лежат на отрезке  $[\gamma_1, \gamma_2]$ , где

$$\gamma_1 = \min_{v \in H} \frac{(Sv, v)}{(v, v)}, \quad \gamma_2 = \max_{v \in H} \frac{(Sv, v)}{(v, v)}.$$

Таким образом, собственные числа  $S$  не превосходят 1, если выполняется соотношение

$$-1 \leq \frac{(Sv, v)}{(v, v)} \leq 1. \quad (13)$$

Преобразуя выражение  $(Sv, v)/(v, v)$  и обозначая  $v = B^{1/2}z$ , получаем

$$\frac{(Sv, v)}{(v, v)} = \frac{(v, v) - \tau (B^{-1/2}AB^{-1/2}v, v)}{(v, v)} = \frac{(Bz, z) - \tau (Az, z)}{(Bz, z)}.$$

Поэтому (13) имеет вид

$$-1 \leq \frac{(Bz, z) - \tau (Az, z)}{(Bz, z)} \leq 1.$$

Правая часть неравенства выполнена всегда, поскольку  $A, B > 0$ ; левая часть эквивалентна выполнению для любого  $z \in H, z \neq 0$ , неравенства

$$2(Bz, z) - \tau (Az, z) \geq 0.$$

Последнее означает, что

$$B \geq \frac{\tau}{2} A. \quad (14)$$

Как видим, условие (14), обеспечивающее устойчивость схемы (5), не совпадает с (8). Однако у схем, обладающих свойством (8), имеется существенное достоинство. В ряде практических задач интервал интегрирования  $T$  уравнения (1) достаточно велик или же требуется вести счет задачи до выхода на стационарный режим. В этих случаях можно воспользоваться более точной оценкой устойчивости схемы (5).

Обозначим  $(1 + \tau/\alpha)^{-1} = q$  и  $\tau\alpha = \gamma$ . Тогда из (11) имеем

$$\begin{aligned} \|u^{n+1}\|_D^2 &\leq q (\|u^n\|_D^2 + \gamma \|\varphi^n\|_{D^{-1}}^2) \leq q^2 \|u^{n-1}\|_D^2 + q\gamma \|\varphi^n\|_{D^{-1}}^2 + \\ &+ \gamma q^2 \|\varphi^{n-1}\|_{D^{-1}}^2 \leq \dots \leq q^{n+1} \|u^0\|_D^2 + \\ &+ q\gamma (\|\varphi^n\|_{D^{-1}}^2 + q \|\varphi^{n-1}\|_{D^{-1}}^2 + \dots + q^n \|\varphi^0\|_{D^{-1}}^2). \end{aligned}$$

Пусть  $\|\varphi\|_\infty = \max_n \|\varphi^n\|_{D^{-1}} < \infty$ . Тогда из последнего неравенства получаем

$$\|u^n\|_D^2 \leq \gamma \frac{q}{1-q} \|\varphi\|_\infty^2 + q^n \|u^0\|_D^2. \quad (15)$$

Отсюда следует при  $\|\varphi\|_\infty < \infty$ , что решение задачи (5) будет ограничено на бесконечном промежутке времени при выполнении условия (8). Отметим, что из (14) ограниченность решения на бесконечном промежутке времени при наличии правой части, вообще говоря, не следует.

Разностную схему (5) можно рассматривать как итерационный процесс для решения уравнения

$$Au = \varphi,$$

где  $\varphi^n = \varphi$ . В этом случае выполнение условия  $B - \tau A > 0$  обеспечивает сходимость итерационного процесса. Действительно, записывая уравнение для погрешности  $z^n = u^n - u$ , имеем

$$Bz_i^n + Az^n = 0, \quad z^0 = u^0 - u.$$

Из условия  $D = B - \tau A > 0$  следует, что  $D$  определяет в  $H$  норму, которую будем обозначать  $\|\cdot\|_D$ . Тогда из (9) получаем

$$\|u^{n+1}\|_D^2 + 2\tau \|u^{n+1}\|_A^2 \leq \|u^n\|_D^2.$$

Отсюда

$$\left(1 + \frac{\tau}{\kappa}\right) \|u^{n+1}\|_D^2 + \tau \|u^{n+1}\|_A^2 \leq \|u^n\|_D^2.$$

Введем норму  $\|u^n\|_1^2 = \|u^n\|_D^2 + \tau \left(1 + \frac{\tau}{\kappa}\right)^{-1} \|u^n\|_A^2$ ; тогда из (16) следует окончательная оценка

$$\|u^{n+1}\|_1^2 \leq \left(1 + \frac{\tau}{\kappa}\right)^{-1} \|u^n\|_1^2 \leq \dots \leq \left(1 + \frac{\tau}{\kappa}\right)^{-n-1} \|u^0\|_1^2.$$

Таким образом, итерационный метод (5) сходится со скоростью геометрической прогрессии. При этом скорость сходимости определяется величинами  $\tau$  и  $\kappa$  и условием  $D > 0$ . Процесс сходится как в норме, определяемой оператором  $D$ , так и в норме, определяемой оператором  $A$ .

Выясним, при каких условиях будут устойчивы схемы (3), (4). Схема (3) уже имеет вид (5), при этом  $B = E$ , а  $A = -\Delta^h$ . Необходимо, таким образом, проверить выполнение условия  $B \geq \frac{\tau}{2} A$ . Имеем (см. § 6)

$$\begin{aligned} (Av, v) = - \sum_{i,j=1}^{M-1} h^2 \Delta^h v_{ij} v_{ij} &= \sum_{i,j=0}^{M-1} h^2 \left[ \left( \frac{\bar{v}_{i+1,j} - \bar{v}_{ij}}{h} \right)^2 + \right. \\ &\left. + \left( \frac{\bar{v}_{i,j+1} - \bar{v}_{ij}}{h} \right)^2 \right] \leq \frac{8}{h^2} \sum_{i,j=1}^{M-1} h^2 v_{ij}^2 = \frac{8}{h^2} (v, v). \end{aligned}$$

Таким образом, чтобы (14) было справедливо, достаточно выполнения неравенства  $2/\tau \geq 8/h^2$ , т. е. явная схема (3) условно устойчива при  $\tau \leq h^2/4$ .

Представим теперь неявную схему (4) в виде (5). В этом случае

$$B = E - \tau \Delta^h, \quad A = -\Delta^h, \quad \text{причем } A > 0.$$

Таким образом, условие (14), равно как и условие (8), выполнено при любых  $\tau$  и  $h$ , т. е. неявная схема (4) безусловно устойчива.

Если для решения системы уравнений относительно значений решения на верхнем слое применяется *марш-алгоритм*, то число арифметических операций при переходе от слоя к слою пропорционально числу неизвестных. Таким образом, неявная схема является экономичной.

Перейдем к изучению других экономичных разностных схем для уравнения (1). Будем рассматривать задачу (1) с однородными граничными условиями, т. е. при  $\alpha \equiv 0$ . Пусть  $\Lambda_1$  и  $\Lambda_2$  — операторы второй разделенной разности по направлению  $x_1$  и  $x_2$  соответственно, т. е.

$$\Lambda_1 v_{ij} = \frac{\bar{v}_{i+1,j} - 2\bar{v}_{ij} + \bar{v}_{i-1,j}}{h^2}, \quad \Lambda_2 v_{ij} = \frac{\bar{v}_{i,j+1} - 2\bar{v}_{ij} + \bar{v}_{i,j-1}}{h^2}.$$

Здесь, как и ранее,  $\tilde{v}$  — функция, совпадающая с  $v$  на  $\Omega_h$  и равная нулю на  $\partial\Omega_h$ .

**Задача 1.** Проверить, что функция  $\varphi_j \sin(\pi mih)$ , где  $\varphi_j$  — произвольная функция аргумента  $j$ , является собственной для оператора  $\Lambda_1$ , а любая функция  $\psi_i \sin(\pi mjh)$ , где  $\psi_i$  — произвольная функция аргумента  $i$ , — собственной для оператора  $\Lambda_2$ .

**Задача 2.** Проверить, что функции  $\varphi_{mn} = \sin(\pi mih) \sin(\pi njh)$  образуют полную систему собственных функций операторов  $\Lambda_1$  и  $\Lambda_2$ .

Положим

$$B = (E - \mu\Lambda_1)(E - \mu\Lambda_2), \quad A = -\Delta^h.$$

Оператор  $B$  является симметричным и положительно определенным как произведение симметричных положительно определенных и коммутирующих между собой операторов. Операторы вида  $B$  обычно называют *расщепляющимися*. Коммутируемость операторов  $(E - \mu\Lambda_1)$  и  $(E - \mu\Lambda_2)$  можно проверить непосредственно; кроме того, она следует из того факта, что эти операторы имеют общую полную систему собственных функций (см. задачу 2) и, следовательно, записываются в виде

$$(E - \mu\Lambda_1) = C^{-1}M_1C, \quad (E - \mu\Lambda_2) = C^{-1}M_2C,$$

где  $M_1, M_2$  — диагональные матрицы, матрица  $C$  одна и та же.

Проверим, при каких  $\mu$  выполняется условие (14). Имеем

$$B = E - \mu(\Lambda_1 + \Lambda_2) + \mu^2\Lambda_1\Lambda_2 = E - \mu\Delta^h + \mu^2\Lambda_1\Lambda_2,$$

поэтому условие (14) приобретает вид

$$E - \mu\Delta^h + \mu^2\Lambda_1\Lambda_2 \geq -\frac{\tau}{2}\Delta^h.$$

Так как оператор  $E + \mu^2\Lambda_1\Lambda_2$  положительно определен, то условие  $\mu \geq \tau/2$  обеспечивает выполнение (14), т. е. при  $\mu \geq \tau/2$  разностная схема (5) безусловно устойчива по начальным данным.

Рассмотрим алгоритм реализации схемы (5) в данном случае. Обозначим  $z = \frac{u^{n+1} - u^n}{\tau}$  и представим (5) в виде

$$(E - \mu\Lambda_1)(E - \mu\Lambda_2)z = \Delta^h u^n + \varphi^n.$$

Последнее уравнение разобьем на два:

$$(E - \mu\Lambda_1)y = \Delta^h u^n + \varphi^n, \quad (E - \mu\Lambda_2)z = y. \quad (17)$$

Функция  $g = \Delta^h u^n + \varphi^n$  может быть вычислена во всех точках  $\Omega_h$ , т. е. можно считать ее известной. Первое уравнение (17) запишем в виде

$$y_{ij} - \mu \frac{y_{i+1,j} - 2y_{ij} + y_{i-1,j}}{h^2} = g_{ij}, \quad i = 1, \dots, M-1; \\ j = 1, \dots, M-1. \quad (18)$$

При фиксированном  $j$  система (18) относительно неизвестных  $y_{1,j}, y_{2,j}, \dots, y_{M-1,j}$  представляет собой систему уравнений с трехдиагональной матрицей, которая может быть решена, например, методом прогонки за  $O(M)$  арифметических операций. Решая (18) при каждом  $j = 1, \dots, M-1$ , найдем функцию  $y$  во всех узлах  $\Omega_h$ . Группы неизвестных, связанные уравнениями (18), объединены на рис. 10.7.1 символом  $\leftrightarrow$ .

Примечание. Если решается неоднородная краевая задача, т. е.  $u|_{\Gamma} \neq 0$ , то, вообще говоря,  $y = (E - \mu\Lambda_2) \frac{u^{n+1} - u^n}{\tau}$  не удовлетворяют граничному условию  $y|_{\Gamma} = 0$  и значения  $y|_{\Gamma}$  требуется определять специальным образом.

Аналогично второе уравнение (17) расписывается в виде

$$z_{ij} - \mu \frac{z_{i,j+1} - 2z_{ij} + z_{i,j-1}}{h^2} = y_{ij}, \quad j = 1, \dots, M-1; \\ i = 1, \dots, M-1. \quad (19)$$

При фиксированном  $i$  (19) является системой уравнений с трехдиагональной матрицей относительно неизвестных

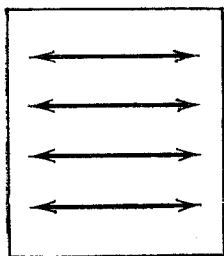


Рис. 10.7.1

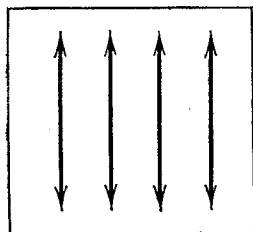


Рис. 10.7.2

$(z_{i,1}, z_{i,2}, z_{i,M-1})$ . Таким образом, функция  $z$  может быть найдена из (19) за  $O(M^2)$  арифметических операций. Группы неизвестных, связанные уравнениями (19), объединены на рис. 10.7.2 символом  $\updownarrow$ . Значения  $u^{n+1}$  находятся по явным формулам:

$$u^{n+1} = u^n + \tau z. \quad (20)$$

Таким образом, суть предполагаемого алгоритма заключается в следующем: при каждом  $j$  решаем систему уравнений (18) с трехдиагональной матрицей. При этом аргумент  $i$  является «горизонтальной» координатной функции  $y$ . Далее, используя найденное значение  $y$ , при каждом  $i$  решаем систему уравнений (19). Аргумент  $j$  в этом случае меняется в «вертикальном» направлении. После этого значение  $u^{n+1}$  находится по явной формуле (20).

Часто возникает ситуация, когда функция  $\alpha(x, t)$  не зависит от  $t$ . В этом случае реализация алгоритма (18) — (20) не требует вычисления граничных значений для  $y$  на каждом шаге. Это связано с тем, что  $z$ , а следовательно, и  $y$  имеют однородные граничные условия.

Описанный алгоритм часто называют методом *расщепления*.

Как мы видели, основная его идея заключается в построении оператора при разностной производной по времени таким образом, чтобы этот оператор являлся произведением операторов  $B = B_1 B_2$ , каждый из которых действует только в одном направлении, и полученная схема аппроксимировала исходную задачу. Так, в данном случае  $B_i = E - \mu \Lambda_i$ .

**Задача 3.** Доказать, что при достаточно гладком решении для рассматриваемой схемы имеет место аппроксимация порядка  $O(\mu + \tau + h^2)$ .

Близким по своей структуре является метод переменных направлений. Суть его заключается в переходе от  $u^n$  к  $u^{n+1}$  по формулам

$$\begin{aligned} \frac{u^{n+1/2} - u^n}{0,5\tau} &= \Lambda_1 u^{n+1/2} + \Lambda_2 u^n + \Phi^n, \\ \frac{u^{n+1} - u^{n+1/2}}{0,5\tau} &= \Lambda_1 u^{n+1/2} + \Lambda_2 u^{n+1} + \Phi^n. \end{aligned} \quad (21)$$

Здесь введен промежуточный вектор неизвестных  $u^{n+1/2}$ ; первое уравнение (21) решается применением прогонки по оси  $x_1$ , а второе — применением прогонки по оси  $x_2$ .

Построение методов такого типа в  $k$ -мерном случае можно провести по точно такой же схеме. Пусть  $A = A_1 + A_2 + \dots + A_k$ , где  $A_i$  — одномерный оператор в  $i$ -м направлении. Положим  $B = (E + \mu A_1) \dots (E + \mu A_k)$ . Нужная разностная схема будет иметь вид

$$(E + \mu A_1) \dots (E + \mu A_k) \frac{u^{n+1} - u^n}{\tau} + Au^n = \Phi^n. \quad (22)$$

Реализация этого метода проводится по такому же алгоритму, что и выше. Параметры  $\mu$  и  $\tau$  выбираются из условий устойчивости и аппроксимации разностной схемы.

Перейдем к рассмотрению методов решения параболических уравнений в случае, когда область  $\Omega$  имеет достаточно произвольную форму. В этом случае реализация описанной выше схемы, основанной на представлении оператора  $B$  на верхнем слое в виде произведения одномерных операторов, встречает существенные затруднения. Однако применима схема (21).

Другая разностная схема может быть получена из следующих соображений. Представим  $A$  в виде  $A = R_1 + R_2$ , где  $R_1^T = R_2$ ,  $R_1$  и  $R_2$  — правая и левая треугольные матрицы. В частности, если  $A = -\Delta^h$ , то  $R_1$  и  $R_2$  можно записать следующим образом:

$$R_1 v|_{(i,j)} = \frac{2\bar{v}_{ij} - \bar{v}_{i+1,j} - \bar{v}_{i,j+1}}{h^2}, \quad R_2 v|_{(i,j)} = \frac{2\bar{v}_{ij} - \bar{v}_{i-1,j} - \bar{v}_{i,j-1}}{h^2}. \quad (23)$$

Рассмотрим разностную схему

$$(E + \sigma\tau R_1)(E + \sigma\tau R_2) \frac{u^{n+1} - u^n}{\tau} + Au^n = \varphi^n. \quad (24)$$

Оператор  $B = (E + \sigma\tau R_1)(E + \sigma\tau R_2)$  будет симметричным и положительно определенным. Поэтому для устойчивости схемы по начальным данным достаточно проверить выполнение условия (14). Параметр  $\sigma$  (весовой множитель) выбирается таким, чтобы схема была устойчива и аппроксимировала исходное уравнение.

Пусть рассматривается пример уравнения (1) в прямоугольнике со стационарными граничными условиями. В этом случае  $A = -\Delta^h$ , а  $R_1$  и  $R_2$  определяются по формулам (23). Тогда  $B = (E + \sigma\tau R_1)(E + \sigma\tau R_2) = E - \sigma\tau\Delta^h + \sigma^2\tau^2 R_1 R_2$ . Оператор  $R_1 R_2$  является положительно определенным, поэтому  $B = B^* > 0$ . Условие  $B \geq \frac{\tau}{2} A$  в данном случае заведомо будет выполнено при  $\sigma \geq 0,5$ . Таким образом, условие  $\sigma \geq 0,5$  обеспечивает безусловную устойчивость схемы (24).

Найдем порядок аппроксимации разностной схемы (24). Так как

$$R_1 v = \frac{2v_{ij} - v_{i+1,j} - v_{i,j+1}}{h^2} = \frac{1}{h} \left( \frac{v_{ij} - v_{i+1,j}}{h} \right) + \frac{1}{h} \left( \frac{v_{ij} - v_{i,j+1}}{h} \right),$$

то, используя формулу Тейлора, получим

$$R_1 [u] = -\frac{1}{h} \left( \frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2} + \frac{h}{2} \frac{\partial^2 u}{\partial x_1^2} + \frac{h}{2} \frac{\partial^2 u}{\partial x_2^2} \right) + O(h).$$

Используя аналогичную оценку погрешности аппроксимации оператора  $R_2$ , приходим к заключению, что выражение  $B = E - \sigma\tau\Delta^h + \sigma^2\tau^2 R_1 R_2$  аппроксимирует единичный оператор  $E$  с порядком  $O(\sigma\tau + \sigma^2\tau^2/h^2)$ . Если считать, что  $\sigma \geq 0,5$  порядка 1, то схема (24) аппроксимирует исходное уравнение (1) с порядком  $O(h^2 + \tau + \tau^2/h^2)$ . Таким образом, величина  $\tau/h$  должна быть достаточно мала. Тем не менее схема (24) все же лучше, чем явная. Для устойчивости явной схемы требуется выполнение условия  $\tau \leq h^2/4$ , в то время как при  $\sigma \geq 0,5$  схема (24) будет безусловно устойчивой и шаг по времени  $\tau$  может быть выбран существенно большим. Например, можно взять  $\tau = ah^2$  или же  $\tau = ah^{3/2}$ . Погрешность аппроксимации по времени в этих случаях будет иметь порядок  $O(h^2)$  и  $O(h)$  соответственно. Следует, однако, отметить, что обычно решение параболического уравнения по  $t$  обладает достаточным количеством производных и эти производные стремятся к нулю. Это оправдывает применение схемы (24) к расчету нестационарных задач, поскольку шаг по времени  $\tau$  можно брать достаточно большим.

Выпишем алгоритм расчета по схеме (24). Обозначим  $(E + \sigma\tau R_2)z$  через  $y$ . Тогда, решая уравнение

$$(E + \sigma\tau R_1)y = -Au^n + \varphi^n, \quad (25)$$

можно найти значение  $y$  по явным формулам. Действительно, в случае, когда  $\Omega$  — квадрат (что несущественно), уравнение (25) в узле  $(i, j)$  имеет вид

$$y_{ij} + \frac{\sigma\tau}{h^2}(2y_{ij} - y_{i+1,j} - y_{i,j+1}) = (-Au^n + \varphi^n)_{i,j}. \quad (26)$$

По известным значениям  $y_{im}, y_{mj}, i, j = 1, \dots, M$ , из (26) по явным формулам можно найти  $y_{i, m-1}$  и  $y_{m-1, j}, i, j = 1, \dots, M-1$ . После этого по тем же формулам находим значения  $y_{i, m-2}, y_{m-2, j}$  и т. д.

Для вычисления значения  $y_{ij}$  требуется число арифметических операций, не зависящее от шагов сетки. Поэтому вычисление значений функции  $y$  во всех узлах потребует  $O(M^2)$  арифметических операций, что по порядку совпадает с количеством узлов на слое. Аналогичным образом решая уравнение

$$(E + \sigma\tau R_2)z = y$$

за  $O(M^2)$  арифметических операций найдем значение  $z$ . Решение  $u^{n+1}$  находится после этого простым пересчетом по формуле

$$u^{n+1} = u^n + \tau z.$$

Таким образом, переход от  $u^n$  к  $u^{n+1}$  в схеме (24) требует числа арифметических операций, пропорционального количеству узлов сетки на слое, т. е. схема (24) экономична.

Во всех рассуждениях относительно схемы (24) предполагалось, что граничные условия не зависят от времени. В этом случае  $y$  и  $z$  на границе  $\partial\Omega_h$  равны нулю.

Наиболее эффективна схема (24), если ее рассматривать как итерационный метод для решения стационарной задачи  $Au = \varphi$ . В этом случае требование аппроксимации по  $t$  не играет никакой роли и параметр  $\tau$  можно выбирать только из соображений наиболее быстрой сходимости итерационного метода. Обычно выбирают  $\tau = O(h)$  и  $\sigma = 1$ , чтобы выполнялось (8). Тогда операторы  $B$  и  $A$  связаны соотношением

$$\gamma_1 h A \leq B \leq \gamma_2 A, \quad (27)$$

где  $\gamma_1, \gamma_2 > 0$  не зависят от  $h$ , и, решая стационарную задачу с помощью итерационного процесса (24), получим решение с точностью  $\varepsilon$  за  $O(h^{-3} \ln(\varepsilon^{-1}))$  арифметических операций.

Задача 4. Доказать оценку (27) при  $\sigma = 1, \tau = O(h)$ .

Задача 5. Показать, что при любых  $\sigma > 0$  и  $\tau$  отношение  $\gamma_2/\gamma_1$  ограничено снизу постоянной, отличной от нуля.

Итерационный процесс можно ускорить, если зафиксировать  $B$  и после этого выбирать параметр  $\tau$  переменным. В частности, если выбирать  $\tau_k$  как указывалось в § 6.6 (чебышевское ускорение), то решение стационарной задачи с точностью  $\varepsilon$  может быть получено за

$$O(h^{-5/2} \ln(\varepsilon^{-1}))$$

арифметических операций.



Для решения параболического уравнения в области достаточно произвольной формы существуют также и другие методы. Рассмотрим метод, который сводит исходную задачу к решению последовательности одномерных задач. Изложение метода проведем на примере уравнения (1).

Представим оператор  $\Delta$  в виде суммы одномерных операторов

$$\Delta = L_1 + L_2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2},$$

а правую часть  $f$  — в виде суммы правых частей:  $f = f_1 + f_2$ . Левая и правая части уравнения (1) равны сумме левых и правых частей уравнений

$$\frac{1}{2} \frac{\partial u}{\partial t} - L_1 u = f_1, \quad \frac{1}{2} \frac{\partial u}{\partial t} - L_2 u = f_2. \quad (28)$$

Опишем переход от  $n$ -го к  $(n+1)$ -му слою. Аппроксимируем первое из уравнений (28) следующим образом:

$$P_h^1 u_{ij}^n \equiv \frac{u_{ij}^{n+1/2} - u_{ij}^n}{\tau} - \frac{u_{i+1,j}^{n+1/2} - 2u_{ij}^{n+1/2} + u_{i-1,j}^{n+1/2}}{h^2} - f_{1,ij}^{n+1/2} = 0. \quad (29)$$

Второе уравнение (28) заменим соотношением

$$P_h^2 u_{ij}^{n+1/2} \equiv \frac{u_{ij}^{n+1} - u_{ij}^{n+1/2}}{\tau} - \frac{u_{i,j+1}^{n+1} - 2u_{i,j}^{n+1} + u_{i,j-1}^{n+1}}{h^2} - f_{2,ij}^{n+1} = 0. \quad (30)$$

Таким образом, алгоритм заключается в последовательном решении уравнений (29), (30). При этом вычисленное значение функции является начальным условием для следующего уравнения.

Ясно, что каждое из уравнений (29), (30) не аппроксимирует исходную задачу. Найдем погрешность аппроксимации. Имеем

$$P_h^1 [u] = \frac{1}{2} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - f_1 + O(h^2 + \tau),$$

$$P_h^1 [u] - \left[ \frac{\partial u}{\partial t} - \Delta u - f \right] = -\frac{1}{2} \frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial y^2} + f_2 + O(h^2 + \tau) \equiv \psi_1.$$

Аналогично

$$P_h^2 [u] = \frac{1}{2} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial y^2} - f_2 + O(h^2 + \tau),$$

$$P_h^2 [u] - \left[ \frac{\partial u}{\partial t} - \Delta u - f \right] = -\frac{1}{2} \frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + f_1 + O(h^2 + \tau) \equiv \psi_2.$$

В общем случае  $\psi_i = O(1)$ , поэтому и уравнения (29), (30) аппроксимируют (1) с порядком  $O(1)$ . Однако

$$\psi_1 + \psi_2 = -\frac{\partial u}{\partial t} + \Delta u + f + O(h^2 + \tau) = O(h^2 + \tau).$$

В этом случае говорят, что *схема* (29), (30) *аппроксимирует задачу* (1) *в суммарном смысле*, т. е. хотя каждое из уравнений (29), (30) не аппроксимирует исходную задачу, сумма погрешностей аппроксимаций этих уравнений равна  $O(h^2 + \tau)$ .

Реализация (29), (30) требует на каждом шаге решения уравнений с трехдиагональной матрицей. Таким образом, этот метод применим для решения уравнения (1), когда область  $\Omega$  имеет достаточно произвольную форму. Остается лишь выявить ее устойчивость и сходимость. Оказывается, имеет место

Теорема (без доказательства). *Схема* (29), (30) *устойчива в сеточной норме пространства*  $C$  *и при достаточно гладком решении*

$$\|u^n - u(n\tau)\|_{C(\Omega_h)} \leq C_1(h^2 + \tau),$$

где  $C_1$  не зависит от  $h$  и  $\tau$ .

Рассмотренный метод получения разностных схем носит название *метода дробных шагов* или же *метода суммарной аппроксимации*. Его можно применять не только в линейных задачах, но и в нелинейных.

В общем случае для уравнения

$$\frac{\partial u}{\partial t} = P^1(u) + \dots + P^k(u), \quad (31)$$

где операторы  $P^i(u)$ , вообще говоря, нелинейные, не обязательно одномерные, схема метода дробных шагов заключается в следующем. Решение на шаге уравнения (31) заменяется последовательным решением на шаге уравнений

$$\frac{\partial u_i}{\partial t} = P^i(u_i), \quad i = 1, \dots, k.$$

При этом в качестве начального условия на шаге для каждого из уравнений берется значение, вычисленное из предыдущего уравнения.

Разбиение исходного уравнения на систему уравнений не обязательно осуществлять, ориентируясь на размерность операторов  $P^i(u)$ .

## § 8. Методы решения сеточных эллиптических уравнений

В этом параграфе будут рассмотрены методы решения систем сеточных эллиптических уравнений. Рассмотрим простейший пример системы такого типа (см. § 6). Пусть

$$\frac{u_{m+1,n} - 2u_{m,n} + u_{m-1,n}}{h_1^2} - \frac{u_{m,n+1} - 2u_{m,n} + u_{m,n-1}}{h_2^2} = \varphi_{mn}, \quad (1)$$

$$u_{m,0} = u_{m,N} = 0, \quad u_{0,n} = u_{M,n} = 0;$$

$$m = 1, \dots, M-1; \quad n = 1, \dots, N-1,$$

система сеточных уравнений относительно неизвестных  $u_{mn}$ ,





Далее последовательно исключаем векторы неизвестных  $u_1, u_2, \dots, u_{N-2}$  и получаем рекуррентные формулы, аналогичные формулам метода прогонки. При реализации этих формул придется произвести порядка  $cN$  операций умножения и обращения матриц размерности  $N - 1$ , что потребует всего порядка  $2cN^4$  арифметических операций. Дополнительно требуется произвести порядка  $O(N)$  операций умножения матрицы на вектор и сложения векторов; для этого потребуется  $O(N^3)$  арифметических операций. Общие вычислительные затраты, таким образом, составят порядка  $2cN^4$  арифметических операций.

Описанные выше методы непосредственно применимы в случае эллиптического уравнения или системы эллиптических уравнений с произвольными, в частности, переменными коэффициентами и во всех этих случаях требуют  $O(N^4)$  арифметических операций.

В конкретном случае системы (2) при  $N = 2^l$  можно предложить метод, требующий  $O(N^3 \ln N)$  арифметических операций. Для этого следует применить метод Гаусса в векторной форме при следующем порядке исключения: сначала исключаются векторы  $u_1, u_3, u_5$ , т. е. все векторы  $u_{2k+1}$ , затем векторы  $u_2, u_6, u_{10}, \dots$ , т. е. все векторы  $u_{2(2k+1)}$ , затем векторы  $u_4, u_{12}, u_{20}, \dots$ , т. е. все векторы  $u_{4(2k+1)}$  и т. д. Оказывается, что при таком способе исключения многократно умножаются и обращаются одни и те же матрицы. Если вместо этого запомнить произведения соответствующих матриц, то общее число операций окажется  $O(N^3 \log_2 N)$ .

Описанные методы требуют в общем случае для запоминания ленты матрицы  $O(N^3)$  слов памяти ЭВМ.

**Задача 1.** Показать, что оценка количества арифметических операций  $O(N^4)$  для метода Гаусса по порядку является неулучшаемой.

**Задача 2.** Показать, что при нумерации неизвестных

$$(u_{11}, u_{12}, u_{21}, \dots, u_{1n}, u_{2, n-1}, \dots, u_{n1}, \dots, u_{N-1, N-1})$$

количество арифметических операций в методе Гаусса по порядку также равно  $O(N^4)$ .

**Задача 3.** Для произвольной области указать порядок исключения неизвестных в системе (1), при котором решение системы получается за  $O(N^3)$  арифметических операций.

Рассмотрим прямой метод решения системы уравнений (1), основанный на использовании дискретного преобразования Фурье в случае прямоугольника. Здесь не будем предполагать, что  $M = N$ . Преобразуем вначале систему уравнений (1) к операторной форме. Пусть  $u = u(mh_1, nh_2)$  — сеточная функция, определенная при  $0 \leq m \leq M, 0 \leq n \leq N$ , которая соответствует решению (1), т. е.  $u(mh_1, nh_2) = u_{mn}$ . Аналогично  $\varphi(mh_1, nh_2) = \varphi_{mn}$ ,  $\varphi_{0n} = \varphi_{mN} = 0$ ,  $\varphi_{m0} = \varphi_{mN} = 0$  — сеточная функция, соответствующая правой части (1). Через  $\Lambda_k$  обозна-

чим операторы

$$(\Lambda_1 v)_{mn} = \begin{cases} \frac{v_{m+1, n} - 2v_{mn} + v_{m-1, n}}{h_1^2}, & 1 \leq m \leq M-1, \\ & 1 \leq n \leq N-1, \\ 0 & \text{в остальных случаях;} \end{cases}$$

$$(\Lambda_2 v)_{mn} = \begin{cases} \frac{v_{m, n+1} - 2v_{mn} + v_{m, n-1}}{h_2^2}, & 1 \leq m \leq M-1, \\ & 1 \leq n \leq N-1, \\ 0 & \text{в остальных случаях.} \end{cases}$$

Тогда система уравнений (1) может быть представлена в операторной форме

$$Au \equiv -(\Lambda_1 + \Lambda_2)u = \varphi. \quad (3)$$

Пусть для определенности  $Mh_1 = 1$ . Заметим, что этого всегда можно достичь, умножая обе части (1) на соответствующий коэффициент. Функция  $u$  как функция переменного  $m$  представима в виде дискретной суммы Фурье по переменной  $m$  (см. § 4.3):

$$u(mh_1, nh_2) = \sum_{j=1}^{M-1} u_j(nh_2) \sin(\pi j m h_1), \quad (4')$$

$$u_j(nh_2) = 2 \sum_{k=1}^{M-1} h_1 u(kh_1, nh_2) \sin(\pi j k h_1).$$

Аналогично представим  $\varphi$  в виде

$$\varphi(mh_1, nh_2) = \sum_{j=1}^{M-1} \varphi_j(nh_2) \sin(\pi j m h_1), \quad (4'')$$

$$\varphi_j(nh_2) = 2 \sum_{k=1}^{M-1} h_1 \varphi(kh_1, nh_2) \sin(\pi j k h_1).$$

Подставляя полученные разложения в (3), получим

$$-(\Lambda_1 + \Lambda_2)u_{mn} = - \sum_{j=1}^{M-1} (u_j(nh_2) \Lambda_1 \sin(\pi j m h_1) + \sin(\pi j m h_1) \Lambda_2 u_j(nh_2)) = \sum_{j=1}^{M-1} \varphi_j(nh_2) \sin(\pi j m h_1). \quad (5)$$

Напомним, что функции  $\sin(\pi j x)$  являются собственными для оператора  $\Lambda_1$ :

$$\Lambda_1 \sin(\pi j m h_1) = \frac{\sin(\pi j(m+1)h_1) - 2\sin(\pi j m h_1) + \sin(\pi j(m-1)h_1)}{h_1^2} = \\
 = \frac{4 \sin^2 \frac{\pi j h_1}{2}}{h_1^2} \sin(\pi j m h_1) \equiv -\lambda_j \sin(\pi j m h_1).$$

Поэтому выражение (5) может быть заменено эквивалентным

$$\sum_{j=1}^{M-1} \left( \lambda_j u_j(nh_2) + \frac{2u_j(nh_2) - u_j((n-1)h_2) - u_j((n+1)h_2)}{h_2^2} \right) \sin(\pi j m h_1) =$$

$$= \sum_{j=1}^{M-1} \varphi_j(nh_2) \sin(\pi j m h_1). \quad (6)$$

Функции  $\sin(\pi j m h_1)$ ,  $j = 1, \dots, M-1$ , являются ортогональными в скалярном произведении

$$(v, \omega) = \sum_{i=1}^{M-1} h_1 v_i \omega_i.$$

Поэтому (6) представляет собой систему независимых уравнений

$$\lambda_j u_j(nh_2) + \frac{2u_j(nh_2) - u_j((n-1)h_2) - u_j((n+1)h_2)}{h_2^2} = \varphi_j(nh_2),$$

$$1 \leq n \leq N-1; \quad u_j(0) = u_j(Nh_2) = 0, \quad j = 1, \dots, M-1. \quad (7)$$

Система уравнений (7) может быть получена из (6) также умножением обеих частей (6) скалярно на  $\sin(\pi j m h_1)$ ,  $j = 1, \dots, M-1$ . Выражение (7) при фиксированном  $j$  представляет собой систему линейных алгебраических уравнений с трехдиагональной матрицей относительно неизвестных  $u_j(h_2)$ ,  $u_j(2h_2)$ , ...,  $u_j((N-1)h_2)$ , которая может быть решена, например, методом прогонки.

Таким образом, алгоритм решения задачи (1) заключается в следующем:

а) находим из (4'') при каждом  $n$ ,  $0 < n < N$ , коэффициенты  $\varphi_j(nh_2)$ ,  $j = 1, \dots, M-1$ ;

б) при  $j = 1, \dots, M-1$  решаем методом прогонки систему уравнений (7). В результате получаем функции  $u_j(nh_2)$ ,  $j = 1, \dots, M-1$ ;

в) из формулы (4') при  $m = 1, \dots, M-1$ ;  $n = 1, \dots, N-1$  вычисляем значения функции  $u(mh_1, nh_2)$ .

Оценим затраченное количество арифметических операций. Пусть  $M = 2^q$ . В этом случае, используя алгоритм быстрого дискретного преобразования Фурье, найдем все  $\varphi_j$  за  $O(MN \log_2 M)$  арифметических операций. При нахождении коэффициентов  $u_j$  потребуется  $O(MN)$  операций. Наконец, при вычислении  $u$  из (4') с использованием быстрого дискретного преобразования Фурье понадобится  $O(MN \log_2 M)$  операций. Поэтому суммарное количество арифметических операций, необходимых для нахождения решения, в данном случае по порядку равно  $O(MN \log_2 M)$ . В частности, при  $M = N$  получаем  $O(N^2 \log_2 N)$ . Напомним, что метод Гаусса для ленточных матриц требовал  $O(N^4)$  операций.

Рассмотренный метод даст решение намного быстрее, чем метод Гаусса. Однако этот метод применим лишь в случае, когда исходная область является прямоугольником, в то время как метод Гаусса применим и в случае областей общей формы.

В случае прямоугольника существует ряд других методов решения с такой же асимптотикой числа действий. Один из вариантов так называемого марш-алгоритма с приемлемой величиной вычислительной погрешности решает задачу за  $O(N^2)$  арифметических операций при  $N = M = 2^k$  \*).

Рассмотрим приближенные методы решения системы уравнений (1), которые применимы не только в случае прямоугольника. В основу класса этих методов положено то свойство системы уравнений (1), что результат применения матрицы системы к вектору вычисляется по простым формулам и не требует запоминания матрицы. Количество арифметических операций, затрачиваемое на вычисление результата применения матрицы к вектору, по порядку равно  $O(MN)$ , т. е. пропорционально длине вектора. Ранее мы упоминали, что матрица  $A$  системы уравнений (1) симметрична и положительно определена и ее собственные числа  $\lambda_{mn}$ ,  $0 < m \leq M$ ,  $0 < n \leq N$ , лежат в промежутке

$$\lambda_{\min} = \frac{4 \sin^2 \frac{\pi h_1}{2}}{h_1^2} + \frac{4 \sin^2 \frac{\pi h_2}{2}}{h_2^2} \leq \lambda_{kl} \leq \frac{4 \cos^2 \frac{\pi h_1}{2}}{h_1^2} + \frac{4 \cos^2 \frac{\pi h_2}{2}}{h_2^2} = \lambda_{\max}.$$

Поэтому для решения системы уравнений

$$Au = \varphi$$

можно применить, например, метод простой итерации

$$\frac{u^{l+1} - u^l}{\tau} + Au^l = \varphi, \quad u^0 = v, \quad (8)$$

где  $v$  — вектор начального приближения. Согласно § 6.6 параметр  $\tau$  целесообразно выбрать из соотношения

$$\tau = \tau_{\text{опт}} = \frac{2}{\lambda_{\max} + \lambda_{\min}}.$$

При этом метод (8) сходится со скоростью геометрической прогрессии и показатель скорости сходимости метода равен

$$q = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = 1 - \frac{2\lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}.$$

При  $h = h_1 = h_2$  имеем  $\lambda_{\min}/\lambda_{\max} = \pi^2 h^2/4 + O(h^4)$ ,

\*) См. Самарский А. А., Капорин И. Е., Кучеров А. Б., Николаев Е. С. Некоторые современные методы решения сеточных уравнений // Изв. вузов. Сер. мат. — 1983. — № 7(254). С. 3—12.



поэтому

$$q = 1 - \frac{\pi^2 h^2}{2} + O(h^4).$$

Пусть  $\epsilon$  — точность, с которой необходимо найти решение системы уравнений (1). Нормы погрешностей  $\mathbf{z}^n = \mathbf{u} - \mathbf{u}^n$  и  $\mathbf{z}^{n-1} = \mathbf{u} - \mathbf{u}^{n-1}$  на соседних слоях связаны соотношением

$$\|\mathbf{z}^n\| \leq q \|\mathbf{z}^{n-1}\| \leq \dots \leq q^n \|\mathbf{z}^0\|.$$

**Задача 4.** Показать, что при  $\tau = \tau_{\text{опт}}$  расчетные формулы (8) приобретают вид:

$$u_{mn}^{l+1} = \frac{1}{4} (u_{m+1, n}^l + u_{m-1, n}^l + u_{m, n+1}^l + u_{m, n-1}^l) + \frac{h^2}{4} \varphi_{mn}$$

как в случае прямоугольника, так и в случае произвольной области.

Поэтому для выполнения неравенства  $\|\mathbf{z}^n\| \leq \epsilon$  достаточно выбрать  $n$  так, чтобы выполнялось неравенство  $q^n \|\mathbf{z}^0\| \leq \epsilon$ . Отсюда  $n |\ln q| \geq \ln(\epsilon^{-1}) + \ln \|\mathbf{z}^0\|$ . Так как  $q = 1 - \pi^2 h^2 / 2 + O(h^4)$ , то  $|\ln q| = |\ln(1 - \pi^2 h^2 / 2 + O(h^4))| = \pi^2 h^2 / 2 + O(h^4)$ . При малых  $h$  имеем

$$n \geq \frac{2}{\pi^2 h^2} \ln(\epsilon^{-1}) + ch^{-2} \ln \|\mathbf{z}^0\|. \quad (9)$$

Так как нас интересует случай малых  $\epsilon$ , т. е. когда точность решения системы достаточно высока, то второй член в правой части (9) может быть отброшен. Поэтому в данном случае количество итераций  $n$  по порядку должно быть равно  $O(h^{-2} \ln(\epsilon^{-1}))$ .

На каждом шаге итерации матрица умножается на вектор. Отсюда может возникнуть впечатление о необходимости запоминания матрицы  $A$ . На самом деле это не так. Нам необходим лишь результат применения матрицы  $A$  к вектору  $\mathbf{v}$ , который вычисляется по формулам

$$(A\mathbf{v})_{mn} = \frac{4v_{mn} - v_{m+1, n} - v_{m-1, n} - v_{m, n+1} - v_{m, n-1}}{h^2}.$$

Поэтому матрицу  $A$  запоминать не надо.

Для вычисления одной компоненты вектора  $A\mathbf{v}$  требуется конечное число арифметических операций, поэтому на каждом шаге итерационного метода (8) затрачивается  $O(h^{-2})$  арифметических операций. Общее количество арифметических операций, необходимых для получения решения с точностью  $\epsilon$ , таким образом, равно  $O(h^{-4} \ln(\epsilon^{-1})) = O(N^4 \ln(\epsilon^{-1}))$ . Для сравнения напомним, что при использовании метода Гаусса для ленточных матриц требовалось  $O(N^4)$  арифметических операций. Таким образом, эти методы различаются по числу арифметических операций несущественно. Правда, следует помнить, что метод Гаусса является прямым, а метод простой итерации — приближенным решением системы (1). Достоинством метода простой

итерации является отсутствие необходимости запоминания матрицы системы, т. е. с точки зрения необходимого объема памяти ЭВМ метод простой итерации является более предпочтительным.

В данном случае была описана схема применения метода простой итерации к решению системы сеточных эллиптических уравнений (1), аппроксимирующих исходную задачу в прямоугольной области; однако все рассуждения справедливы также и для случая произвольной области. Следует отметить, что при этом, вообще говоря, неизвестны точные границы спектра  $\lambda_{\min}$  и  $\lambda_{\max}$ . Эти величины можно оценить, например, следующим образом. Пусть сеточный прямоугольник  $K'$  со сторонами  $l'_1$  и  $l'_2$  содержится в  $\Omega_h$ , а сеточный прямоугольник  $K''$  со сторонами  $l''_1$  и  $l''_2$  содержит  $\bar{\Omega}_h$ . Тогда имеет место соотношение

$$\lambda'_{\min} \leq \lambda_{*l} \leq \lambda''_{\max},$$

где  $\lambda'_{\min}$  — минимальное собственное значение сеточной задачи Дирихле в прямоугольнике  $K'$ , а  $\lambda''_{\max}$  — максимальное собственное значение сеточной задачи Дирихле в прямоугольнике  $K''$ . Выбирая  $\tau = 2/(\lambda'_{\min} + \lambda''_{\max})$ , можно проводить расчет системы уравнений (1) методом простой итерации с той же асимптотикой числа арифметических операций, что и в предыдущем случае.

**Задача 5.** Показать, что для итерационного процесса с чебышевским набором параметров (6.6.13) требуемое число действий  $O(N^3 \ln \varepsilon^{-1})$ .

**Задача 6.** Получить такую же оценку числа действий для оптимального линейного итерационного процесса (6.6.19).

**Задача 7.** Получить такую же оценку для трехслойного итерационного процесса (см. § 6.6, задача 3) с фиксированным параметром  $\omega$ .

**Задача 8.** Показать, что при определенном  $\omega$  средняя трудоемкость и загрузка памяти трехслойного итерационного процесса с фиксированным параметром  $\omega$  может быть снижена вдвое за счет следующего обстоятельства. При нахождении  $u^l$  при  $l$  четном вычисляются и запоминаются только значения в точках с четной суммой  $m+n$ , а при  $l$  нечетном — в точках с нечетной суммой  $m+n$ .

**Задача 9.** Рассмотреть попеременно-треугольный метод как метод решения уравнения  $\Delta u + f = 0$ . Показать, что при выборе переменных шагов по времени  $\tau$  в соответствии с (6.6.13) число затрачиваемых действий будет  $O(N^{5/2} \ln(\varepsilon^{-1}))$ .

**Замечание.** В случае прямоугольника в методе переменных направлений (если его рассматривать как итерационный метод) можно предложить последовательность шагов по времени, при которой общее число операций  $O(N^2 \ln N \ln(\varepsilon^{-1}))$ .

Заметим, что все описанные выше методы укладываются в общую схему решения стационарных задач путем установления, описанную в § 7.5. В частности, двуслойные итерационные методы можно рассматривать как аппроксимации уравнения

$$\frac{\partial u}{\partial t} = \Delta u + f,$$

а трехслойные — как аппроксимации уравнения

$$\frac{\partial^2 u}{\partial t^2} + \gamma \frac{\partial u}{\partial t} = \Delta u + f.$$

В случае необходимости решения более сложных стационарных задач для уравнений с частными производными часто идут по такому пути. Строят нестационарный процесс, сходящийся к ре-

шению задачи, а затем в качестве итерационного процесса берут дискретную аппроксимацию этого нестационарного процесса. Например, в случае уравнения  $\Delta u + f = 0$  один из таких процессов установления описывается уравнением

$$\frac{\partial^2 u}{\partial t \partial x} + \frac{\partial^2 u}{\partial t \partial y} + \alpha \frac{\partial u}{\partial t} = \Delta u + f.$$

**Задача 10.** Доказать, что при определенном соотношении между шагами по  $t$ ,  $x$ ,  $y$  сеточная аппроксимация этого уравнения превращается в метод сверхрелаксации (6.7) для решения сеточного уравнения  $\Delta^h u + f = 0$ . В случае, когда область  $G$  есть квадрат и  $h_1 = h_2 = h$ , указать  $\alpha$ , при котором число итераций порядка  $\frac{2}{\pi} h^{-1} \ln(e^{-1})$ .

Эффективными методами решения сеточных эллиптических уравнений являются интенсивно развиваемые в последнее время метод фиктивных компонент и многосеточный метод.

В итерационном методе фиктивных компонент, предназначенном для решения сеточного уравнения Пуассона в области произвольного вида, на каждом шаге итераций решается первая краевая задача для уравнения Пуассона в некотором прямоугольнике, содержащем внутри себя эту область. Если для решения последней задачи применяется марш-алгоритм, то при любом  $p$  для получения решения сеточной задачи в исходной области с точностью  $O(h^p)$  требуется  $O(h^{-2} \ln h)$  арифметических операций.

В многосеточном методе существенно используется близость решений сеточных задач на сетках с малыми шагами. Упомянем основные идеи, лежащие в основе этого метода. При решении задачи на сетке с шагом  $h$  сначала производится несколько итераций метода простой итерации со значением параметра  $\alpha$ , меньшим  $\alpha_{\text{опт}}$  (часто  $\alpha = (2/3) \alpha_{\text{опт}}$ ); строится уравнение, которому удовлетворяет погрешность приближенного решения. Метод простой итерации с таким значением параметра  $\alpha$  делает погрешность приближенного решения относительно гладкой функцией. Поэтому можно попытаться найти эту погрешность, решая упомянутое уравнение относительно этой погрешности на более крупной сетке. Для решения сеточной задачи с точностью  $O(h^p)$  при малом  $p$  (например, для некоторых задач при  $p \leq 2$ ) требуется  $O(h^{-2})$  арифметических операций; при больших  $p$  достаточно  $O(h^2 \ln(h^{-1}))$  арифметических операций.

Поскольку структура алгоритма не зависит от специфики решаемого уравнения, оказывается, что этот метод применим с равным успехом к знакоопределенным задачам, к нелинейным задачам, к уравнениям, не являющимся уравнениями эллиптического типа, например к системе уравнений Навье—Стокса.

**Литература** к гл. X: [6, 11, 14—17, 19, 21, 22, 24, 27, 28, 32, 43—46, 52—58, 61, 64—66].

## Глава XI

### ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ

В этой главе мы дадим краткие описания алгоритмов решения интегральных уравнений, не вникая подробно в вопросы оценки погрешности.

Задача решения интегральных уравнений возникает как вспомогательная при решении краевых задач для дифференциальных уравнений с частными производными и как самостоятельная при исследовании работы ядерных реакторов, при решении так называемых обратных задач геофизики, при обработке результатов наблюдений и т. п. Мы ограничимся рассмотрением случая интегральных уравнений с одной неизвестной функцией и одной независимой переменной.

#### § 1. Решение интегральных уравнений методом замены интеграла квадратурной суммой

В теории численных методов решения интегральных уравнений рассматриваются следующие типичные задачи. Решения интегрального уравнения Фредгольма первого рода

$$Gy = \int_a^b K(x, s) y(s) ds = f(x), \quad (1)$$

интегрального уравнения Фредгольма второго рода

$$y - \lambda Gy = y(x) - \lambda \int_a^b K(x, s) y(s) ds = f(x), \quad (2)$$

интегрального уравнения Вольтерра первого рода

$$Gy = \int_a^x K(x, s) y(s) ds = f(x), \quad (3)$$

интегрального уравнения Вольтерра второго рода

$$y - \lambda Gy = y(x) - \lambda \int_a^x K(x, s) y(s) ds = f(x) \quad (4)$$

и задачи на собственные значения

$$Gu = \lambda u. \quad (5)$$

В последнем случае ищутся числа  $\lambda$ , при которых задача (5) имеет ненулевое решение.

Воспользуемся какой-либо формулой численного интегрирования

$$J(\psi) = \int_a^b \psi(x) dx \approx S_m(\psi) = \sum_{j=1}^m c_j \psi(x_j^{(m)}), \quad (6)$$

где  $c_j$ , вообще говоря, зависят от  $m$ . Имеем равенство

$$J(\psi) = S_m(\psi) + R_m(\psi), \quad (7)$$

где  $R_m(\psi)$  — остаточный член квадратурной формулы (6).

Рассмотрим для примера уравнение Фредгольма второго рода (2). С помощью соотношения (7) его можно переписать в виде

$$y(x) - \lambda \sum_{j=1}^m c_j K(x, x_j^{(m)}) y(x_j^{(m)}) - R_m(\lambda Ky) = f(x); \quad (8)$$

остаточный член  $R_m(\lambda Ky)$  при вычислении интеграла

$\lambda \int_a^b K(x, s) y(s) ds$  с помощью квадратуры (6) является функцией переменной  $x$ . Полагая в (8)  $x = x_i^{(m)}$ ,  $i = 1, \dots, m$ , получим систему уравнений

$$y(x_i^{(m)}) - \lambda \sum_{j=1}^m c_j K(x_i^{(m)}, x_j^{(m)}) y(x_j^{(m)}) = f(x_i^{(m)}) + R_m(\lambda Ky)|_{x_i^{(m)}}.$$

Отбрасывая остаточный член, приходим к системе линейных алгебраических уравнений

$$\begin{aligned} y_i - \lambda \sum_{j=1}^m c_j K(x_i^{(m)}, x_j^{(m)}) y_j &= f_i, \\ f_i &= f(x_i^{(m)}), \quad i = 1, \dots, m. \end{aligned} \quad (9)$$

Для решения этой задачи могут быть применены стандартные методы решения систем линейных алгебраических уравнений.

Будем рассматривать случай вещественных  $K(x, s)$  и  $f(x)$ . Обратим внимание на следующее обстоятельство. Если ядро

$K(x, s)$  интегрального оператора  $G$  было симметричным ( $K(s, x) \equiv K(x, s)$ ), то оператор  $E - G$ , входящий в левую часть исходного уравнения (2), был симметричным.

Однако матрица системы уравнений (9) не обязательно будет симметричной. Мы видели ранее, что решение систем уравнений с симметричной матрицей в определенном смысле предпочтительнее решения системы уравнений с несимметричной матрицей: шире класс точных и итерационных методов, которые могут быть применены для решения таких систем.

Систему уравнений (9) можно преобразовать к виду, в котором матрица системы будет симметричной. Для этого умножим  $i$ -е уравнение системы (9) на  $c_i$ ; получим систему уравнений

$$c_i y_i - \lambda \sum_{j=1}^m c_i c_j K(x_i^{(m)}, x_j^{(m)}) y_j = c_i f_i, \quad i = 1, \dots, m, \quad (10)$$

уже с симметричной матрицей.

Другой возможный способ симметризации состоит в следующем. Умножим  $i$ -е уравнение в (9) на  $\sqrt{c_i}$  и положим  $\sqrt{c_i} y_i = z_i$ . Получим систему уравнений

$$z_i - \lambda \sum_{j=1}^m \sqrt{c_i} \sqrt{c_j} K(x_i^{(m)}, x_j^{(m)}) z_j = \sqrt{c_i} f_i. \quad (11)$$

В случае  $c_i > 0$  второй способ симметризации является более предпочтительным, поскольку разброс собственных значений у матрицы системы (11), как правило, меньше, чем у матрицы системы (10).

Заметим, что в случае, когда в квадратуре (6) все веса одинаковы:

$$c_i^{(m)} = \dots = c_m^{(m)} = (b - a)/m \quad (12)$$

необходимость в таких симметризациях отпадает.

**Задача 1.** Рассмотреть случай комплексного самосопряженного ядра  $K(x, s) = \overline{K(s, x)}$ .

Проверить, что в этом случае при использовании описанных способов получается система уравнений с самосопряженным ядром.

Конечно, как и в случае решения произвольных систем линейных алгебраических уравнений, при использовании методов Гаусса или квадратного корня в процессе вычислений может возникнуть операция деления на нуль или переполнение.

**Задача 2.** Рассмотреть случай формулы прямоугольников, когда выполнено соотношение (12), и постоянного ядра

$$K(x, s) \equiv K = \text{const.}$$

Применить для решения системы (9) метод исключения Гаусса при естественном порядке исключения неизвестных  $y_1, \dots, y_m$ . Показать, что при  $1 - \lambda K(b - a) \geq \varepsilon > 0$  в ходе исключения по методу Гаусса абсолютные величины всех встречающихся элементов  $a_{ij}^{(l)}$  равномерно ограничены сверху некоторой постоянной  $\kappa(\varepsilon)$ , зависящей только от  $\varepsilon$  и не зависящей от  $m$ :

$$|a_{ij}^{(l)}| \leq \kappa(\varepsilon) < \infty.$$

Показать, что

$$\lim_{m \rightarrow \infty} \sup_{i, j, l} |a_{ij}^{(l)}| = \infty$$

при  $1 - \lambda K(b - a) < 0$ .

Задача 3. Показать, что при условии

$$|\lambda| \sum_{j=1}^m |c_j K(x_i^{(m)}, x_j^{(m)})| \leq 1 - \varepsilon, \quad \varepsilon > 0, \quad (13)$$

при решении системы (9) методом Гаусса всегда

$$|a_{ii}^{(l)}| \leq \kappa_l(\varepsilon) < \infty \quad \text{при любых } m, l, i, l.$$

При рассмотрении задачи на собственные значения (5) при аппроксимации интеграла квадратурной суммой (6) возникает система уравнений

$$\sum_{j=1}^m c_j K(x_i^{(m)}, x_j^{(m)}) y_j = \lambda y_i. \quad (14)$$

Для ее решения могут быть применены стандартные методы решения задач на собственные значения.

В случае  $c_j > 0$  и симметричного ядра  $K(x, s)$ , как и выше, задачу на собственные значения можно преобразовать в задачу на собственные значения симметричной матрицы с помощью введения новых переменных  $\sqrt{c_i} y_i = z_i$ .

Задача 4. Пусть все  $c_j K(x_i^{(m)}, x_j^{(m)}) \geq 0$  и среди них есть ненулевые. Доказать, что максимальное по модулю собственное значение  $\lambda_1$  задачи (14) положительно. Доказать, что среди собственных векторов, соответствующих этому собственному значению, имеется вектор, у которого все компоненты неотрицательны.

Указание. Один из возможных путей решения задачи заключается в рассмотрении итерационного процесса (гл. VI)

$$y_i^{(m+1)} = \sum_{j=1}^m c_j K(x_i^{(m)}, x_j^{(m)}) y_j^{(m)}$$

нахождения собственного вектора при начальном приближении  $(y_1^{(0)}, \dots, y_m^{(0)})^T$  с положительными компонентами  $y_i^{(0)}$ .

Интегральное уравнение Вольтерра второго рода (4) можно записать в виде

$$y(x) - \lambda \int_a^b \tilde{K}(x, s) y(s) ds = f(x)$$

с ядром  $\tilde{K}(x, s) = \begin{cases} K(x, s) & \text{при } s \leq x, \\ 0 & \text{при } s > x \end{cases}$  и применить описанный выше алгоритм решения интегрального уравнения Фредгольма второго рода.

Однако на таком пути мы можем получить методы с плохой сходимостью при  $m \rightarrow \infty$ : погрешность квадратурной формулы (6) может оказаться большой, поскольку при  $s > x$  подынтегральная функция равна нулю и на всем отрезке  $[a, b]$  разрывна (если  $K(x, s) \neq 0$ ) или не обладает высокой гладкостью,

Поэтому целесообразнее поступить следующим образом. Зададимся каким-либо набором точек  $a \leq x_1^{(1)} < \dots < x_m^{(m)} \leq b$ . Выпишем соотношения

$$y(x_i^{(m)}) - \lambda \int_a^{x_i^{(m)}} K(x_i^{(m)}, s) y(s) ds = f(x_i^{(m)})$$

и для вычисления интегралов

$$\int_a^{x_i^{(m)}} K(x_i^{(m)}, s) y(s) ds \quad (15)$$

применим какие-либо квадратурные формулы достаточно высокой точности, использующие значения подынтегральной функции в точках  $x_1^{(m)}, \dots, x_m^{(m)}$ . Если при вычислении интеграла используются значения подынтегральной функции лишь в точках  $x_1^{(m)}, \dots, x_i^{(m)}$ , то матрица системы уравнений (9) будет левой треугольной и отыскание решения системы (9) существенно упростится.

Рассмотрим следующую схему решения задачи (4). Пусть  $K(x, s)$  и  $f(s)$   $l$  раз дифференцируемы. Можно показать, что тогда решение  $y(x)$  также  $l$  раз дифференцируемо. Положим  $x_i^{(m)} = a + (i - 1)H$ ,  $H = (b - a)/(m - 1)$ .

Для вычисления интегралов  $\int_a^{x_i^{(m)}} K(x, s) f(s) ds$  применим формулу Грегори с порядком точности  $O(H^l(x_i^{(m)} - a))$ . Такие формулы определены при  $i \geq l$ . При  $i < l$  для вычисления этих интегралов применим какие-либо квадратурные формулы по узлам  $x_0^{(m)}, \dots, x_{l-1}^{(m)}$  с точностью порядка  $O(H^l)$  или  $O(H^{l+1})$ .

В итоге получим систему уравнений (9), у которой выше главной диагонали ненулевые элементы могут находиться только в первых  $l$  строках и столбцах. Решаем систему из первых  $l$  уравнений системы (9) относительно  $l$  неизвестных  $y_1, \dots, y_l$ ; затем находим остальные  $y_i$ , решая систему уравнений с левой треугольной матрицей.

Если ядро  $K(x, s)$  и правая часть  $f(s)$  — аналитические функции, то иногда целесообразно использовать идею метода Батчера [68] решения дифференциальных уравнений. В этом случае получится система уравнений с полностью заполненной матрицей, но при этом погрешность решения убывает как  $q^m$ ,  $q < 1$ .

Если  $K(x, x) \neq 0$ , то дифференцируя по  $x$  уравнение Вольтерра первого рода (3), мы сведем его решение к решению интегрального уравнения Вольтерра второго рода

$$K(x, x) y(x) + \int_a^x K_x(x, s) y(s) ds = f_x(x). \quad (16)$$

Решение интегральных уравнений Фредгольма первого рода будет предметом дополнительного рассмотрения.



## § 2. Решение интегральных уравнений с помощью замены ядра на вырожденное

Другой классический способ решения интегральных уравнений, применяемый в случаях задач (1.2), (1.5), состоит в замене  $K(x, s)$  — ядра интегрального оператора на вырожденное.

*Вырожденным* называется ядро, представимое в виде

$$K(x, s) = \sum_{j=1}^q c_j(x) d_j(s), \quad q < \infty.$$

Пусть

$$K(x, s) \approx K^0(x, s) = \sum_{j=1}^q c_j(x) d_j(s). \quad (1)$$

Для определенности будем предполагать, что  $c_1(x), \dots, c_q(x)$  линейно независимы и  $d_1(s), \dots, d_q(s)$  также линейно независимы. В противном случае ядро  $K^0(x, s)$  можно записать в виде (1) с меньшим значением  $q$ .

В случае (1) есть основания ожидать, что решение уравнения (1.2) близко к решению интегрального уравнения

$$y(x) - \lambda \int_a^b K^0(x, s) y(s) ds = f(s). \quad (2)$$

Подставляя выражение  $K^0(x, s)$  в (2), получим равенство

$$y(x) = f(x) + \lambda \int_a^b \sum_{j=1}^q c_j(x) d_j(s) y(s) ds. \quad (3)$$

Следовательно,

$$y(x) = f(x) + \lambda \sum_{j=1}^q A_j c_j(x), \quad (4)$$

где

$$A_j = \int_a^b d_j(s) y(s) ds.$$

Таким образом, решение уравнения (2) сводится к определению коэффициентов  $A_j$ .

Подставляя выражение (4) для  $y(x)$  в (3), получим уравнение

$$\lambda \sum_{i=1}^q A_i c_i(x) - \lambda \int_a^b \sum_{i=1}^q c_i(x) d_i(s) \left( f(s) + \lambda \sum_{j=1}^q A_j c_j(s) \right) ds = 0;$$

при получении этого равенства в двух случаях индекс суммирования  $j$  переобозначен через  $i$ . Последнее уравнение можно пе-

реписать в виде

$$\sum_{i=1}^q B_i c_i(x) = 0,$$

где

$$B_i = A_i - \int_a^b d_i(s) f(s) ds - \lambda \sum_{j=1}^q A_j \int_a^b d_i(s) c_j(s) ds.$$

Вследствие линейной независимости функций  $c_i(x)$  имеем  $B_i = 0$ . Мы получили систему уравнений относительно неизвестных  $A_i$ :

$$A_i - \lambda \sum_{j=1}^q (d_i, c_j) A_j = (d_i, f); \quad (5)$$

здесь  $(g, f) = \int_a^b g(x) f(x) dx$  — скалярное произведение. После решения этой системы получаем приближение к решению задачи, которое имеет вид

$$y(x) \approx f(x) + \lambda \sum_{j=1}^q A_j c_j(x).$$

Если применяется метод из § 1, то при больших  $m$  решение системы (1.9) методом Гаусса может оказаться невозможным из-за недопустимо большого (порядка  $m^3$ ) числа операций.

Если

$$\sum_{j=1}^m |\lambda c_j K(x_i^{(m)}, x_j^{(m)})| \leq q < 1, \quad (6)$$

то для решения системы уравнений (1.9) можно применить метод простой итерации

$$y_i^{n+1} = \lambda \sum_{j=1}^m c_j K(x_i^{(m)}, x_j^{(m)}) y_j^n + f_i. \quad (7)$$

Если переписать этот итерационный процесс в векторном виде

$$y^{n+1} = B y^n + f,$$

то вследствие (6) имеем  $\|B\|_\infty \leq q < 1$  и погрешность итерационного процесса убывает как  $O(q^n)$ . Если  $q$  не очень близко к 1, то применение итерационного процесса (7) целесообразнее применения метода Гаусса.

Если итерационный процесс сходится медленно или не сходится, то применяется, например, следующий прием. В уравнение (1.2) вводится новая неизвестная функция

$$z(x) = y(x) - \lambda \int_a^b K^0(x, s) y(s) ds$$

и оно преобразуется к виду

$$z(x) - \int_a^b H(x, s) z(s) ds = g(s). \quad (8)$$

Оказывается, что  $\|H(x, s)\|_\infty \rightarrow 0$ , если  $\|K^0(x, s) - K(x, s)\|_\infty \rightarrow 0$ . Поэтому при достаточно малой величине  $\|K^0(x, s) - K(x, s)\|_\infty$  и разумном выборе квадратурной формулы окажется, что система уравнений (1.9), соответствующая уравнению (8), удовлетворяет условию (6) и итерационный процесс (7) быстро сходится.

Рассмотрим другой способ получения уравнения вида (8) с малой  $\|H\|_\infty$ . Определим интегральный оператор

$$Q_0 g = \int_a^b K_0(x, s) g(s) ds.$$

Обозначим через  $G_0$  оператор, обратный к оператору  $E - \lambda Q_0$ , т.е.  $G_0 f$  равно решению уравнения  $y - \lambda Q_0 y = f$ . Применяя к (1.2) оператор  $G_0$ , получим уравнение

$$G_0(E - \lambda Q) y = G_0 f,$$

которое может быть записано в виде

$$y(x) - \int_a^b R(x, s) y(s) ds = h(s), \quad (9)$$

где опять-таки  $\|R\|_\infty \rightarrow 0$ , если  $\|K - K_0\|_\infty \rightarrow 0$ .

В вычислительной практике наиболее употребителен дискретный вариант описанных алгоритмов. По аналогии со случаем вырожденных ядер можно определить класс вырожденных матриц. Пусть  $M_q^m$  — множество матриц размерности  $m$ , представимых в виде

$$S = \sum_{j=1}^q c_j d_j^T;$$

матрица  $c_j d_j^T$  является квадратной матрицей размерности  $m$  и ранга 1.

Аналогично способу решения интегральных уравнений Фредгольма второго рода можно построить способ решения алгебраических систем с «вырожденной» матрицей

$$y - \lambda S y = f, \quad S \in M_q^m;$$

вектор  $y$  отыскивается в виде  $y = f + \sum_{j=1}^q A_j c_j$ .

Относительно неизвестных  $A_j$  получаем систему уравнений (5), где  $(\cdot, \cdot)$  — обычное скалярное произведение векторов.

Рассмотрим простейший случай, пусть в (1.6) применяется формула прямоугольников

$$\int_a^b \psi(x) dx \approx \sum_{i=1}^m \frac{1}{m} \psi(x_i^{(m)}), \quad x_i^{(m)} = a + \frac{(2i-1)}{2m} (b-a);$$

система уравнений (1.9) имеет вид

$$y_i - \frac{\lambda}{m} \sum_{j=1}^m K(x_i^{(m)}, x_j^{(m)}) y_j = f_i \quad (10)$$

или в векторной форме

$$y - \lambda K^{(m)} y = f.$$

Пусть  $m = sl$ ,  $l$  целое,  $s$  нечетное. Определим матрицу  $S$  размерности  $m \times m$  по правилу: ее элементы  $s_{ij}$  равны

$$\frac{1}{m} K(x_{\alpha}^{(l)}(m, l, f), x_{\beta}^{(l)}(m, l, f)),$$

где  $(x_{\alpha}^{(l)}(m, l, f), x_{\beta}^{(l)}(m, l, f))$  — ближайшая к  $(x_{\alpha}^{(m)}, x_{\beta}^{(m)})$  из точек  $(x_{\alpha}^{(l)}, x_{\beta}^{(l)})$ . Близость измеряется как максимум модулей разностей первых и вторых компонент. Таким образом,

$$|x_{\alpha}^{(l)}(m, l, f) - x_{\alpha}^{(m)}| < \frac{b-a}{2l},$$

$$|x_{\beta}^{(l)}(m, l, f) - x_{\beta}^{(m)}| < \frac{b-a}{2l}$$

(напомним, что  $s$  нечетно).

**Задача 1.** Показать, что  $S \in M_l^m$ .

Далее вводим новый вектор неизвестных  $z = y - \lambda S y$  (аналог перехода к уравнению (8)) или умножаем обе части системы (10) слева на матрицу  $(E - \lambda S)^{-1}$  (аналог перехода к уравнению (9)). В обоих случаях получаем новую систему уравнений вида

$$u - Pu = g. \quad (11)$$

Справедливо следующее утверждение.

*Если ядро  $K(x, s)$  непрерывно, то*

$$\|P\|_{\infty} \leq \omega(l), \quad \omega(l) \rightarrow 0 \quad \text{при} \quad l \rightarrow \infty.$$

Таким образом, получаем систему уравнений, для решения которой может быть эффективно применен метод простой итерации.

При реальном решении задач часто в явном виде системы уравнений (8), (9), (11) не выписываются; на каждом шаге необходимые вспомогательные величины, например в случае (11) значения векторов  $(E - \lambda S)^{-1}h$  при различных  $h$  вычисляются заново. В результате этого трудоемкость метода оказывается довольно малой.

Рассмотрим, например, дискретный вариант перехода к системе (9). Имеем систему

$$(E - \lambda S)^{-1}(E - \lambda K^{(m)})y = (E - \lambda S)^{-1}f.$$

Итерационный процесс записывается в виде

$$y^{n+1} = y^n - (E - \lambda S)^{-1}((E - \lambda K^{(m)})y^n - f).$$

Вычисление вектора  $\varphi^n = (E - \lambda K^{(m)})y^n - f$  требует  $O(m^2)$  арифметических операций.

Вектор  $z^n = (E - \lambda S)^{-1}\varphi^n$  находим на каждой итерации, решая систему уравнений

$$z - \lambda Sz = \varphi^n,$$

с матрицей  $S \in M_l^m$ .

Вычисление коэффициентов системы (5) требует  $O(l^2m)$  операций и вычисление  $z$  при известных  $A_j - O(lm)$  операций. Таким образом, при  $l = O(\sqrt{m})$  на каждом шаге производится  $O(m^2)$  операций, по порядку столько же, сколько и в методе простой итерации.

### § 3. Интегральные уравнения Фредгольма первого рода

Задача решения интегрального уравнения Фредгольма первого рода

$$Qy = \int_a^b K(x, s)y(s) ds = f(x) \quad (1)$$

относится к классу некорректных задач.

Поясним, что это означает. Пусть ядро  $K(x, s)$  вещественно и симметрично, т. е.  $K(s, x) = K(x, s)$ . Предположим также, что  $K(x, s)$  и  $f(s)$  непрерывны. Тогда существует полная ортонормированная система собственных функций  $\varphi_n$  оператора  $Q$ :

$$Q\varphi_n = \int_a^b K(x, s)\varphi_n(s) ds = \lambda_n\varphi_n(x),$$

$$(\varphi_i, \varphi_j) = \int_a^b \varphi_i(s)\varphi_j(s) ds = \delta_{ij},$$

где  $\delta_{ij}$  — символ Кронекера. При этом

$$K(x, s) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(x) \varphi_n(s),$$

сходимость ряда в правой части понимается в норме:

$$\|F(x, s)\| = \sqrt{\int_a^b \int_a^b |F(x, s)|^2 dx ds}.$$

Из последнего соотношения следует, что

$$\|K\|^2 = \sum_{n=1}^{\infty} |\lambda_n|^2$$

и, следовательно,  $\lambda_n \rightarrow 0$  при  $n \rightarrow \infty$ .

Рассмотрим случай, когда  $\lambda_n \neq 0$  при  $1 \leq n \leq n_0$  и все  $\lambda_n = 0$  при  $n > n_0$ . Тогда ядро имеет вид

$$K(x, s) = \sum_{n=1}^{n_0} \lambda_n \varphi_n(x) \varphi_n(s),$$

т. е. является вырожденным. В случае вырожденного ядра

$$\begin{aligned} \int_a^b K(x, s)y(s) ds &= \sum_{n=1}^{n_0} \lambda_n \int_a^b \varphi_n(x) \varphi_n(s) y(s) ds = \\ &= \sum_{n=1}^{n_0} \lambda_n (\varphi_n, y) \varphi_n(x) = f(x). \end{aligned}$$

Следовательно, задача (1) может иметь решение только в том случае, когда  $f(x)$  является линейной комбинацией  $\varphi_1(x), \dots, \dots, \varphi_{n_0}(x)$ , т. е. записывается в виде

$$f(x) = \sum_{n=1}^{n_0} f_n \varphi_n(x).$$

Задача 1. Проверить, что этим решением является

$$y(x) = y_0(x) = \sum_{n=1}^{n_0} \frac{f_n}{\lambda_n} \varphi_n(x).$$

Задача 2. Проверить, что любая функция  $y(x)$ , представимая в виде

$$y(x) = y_0(x) + \sum_{n=n_0+1}^{\infty} c_n \varphi_n(x),$$

где  $\sum_{n=n_0+1}^{\infty} |c_n|^2 < \infty$ , также будет решением уравнения (1).

Таким образом, в рассматриваемом случае задача (1) может не иметь решения; в случае, когда она имеет решение, это решение неединственно.

Рассмотрим случай, когда все  $\lambda_n \neq 0$ . Если  $\|f\| = \sqrt{\int_a^b |f(x)|^2 dx} < \infty$ , то  $f(x)$  представима сходящимся в норме пространства  $L_2$  рядом Фурье

$$f(x) = \sum_{n=1}^{\infty} c_n \varphi_n(x), \quad c_n = (f, \varphi_n),$$

$$\sum_{n=1}^{\infty} c_n^2 = \|f\|^2.$$

Здесь и далее сходимость рядов понимается в смысле нормы пространства  $L_2$ . образуем сумму  $S = \sum_{n=1}^{\infty} \left| \frac{c_n}{\lambda_n} \right|^2$ .

Задача 3. Доказать, что при  $S < \infty$  функция  $y(x) = \sum_{n=1}^{\infty} \frac{c_n}{\lambda_n} \varphi_n(x)$

является решением уравнения (1).

Задача 4. Доказать, что при  $S = \infty$  задача 1 не имеет решения.

Задача 5. Пусть все  $\lambda_n \neq 0$ .

Показать, что задача (1) не может иметь двух различных решений (решения, отличающиеся на множестве меры нуль, считаются совпадающими). Таким образом, возможна следующая ситуация.

Задача (1) может иметь не более одного решения, однако при этом решение существует лишь для множества правых частей, удовлетворяющих условию  $S < \infty$ .

Задача 6. Рассмотреть случай решения уравнения (1) описанным выше методом, когда имеется бесконечное число  $\lambda_n$ , отличных от нуля, и бесконечное число  $\lambda_n$ , равных нулю.

При изучении многих задач, в частности в задачах *интерпретации результатов наблюдений*, или, как говорят, в задачах их *обработки*, часто возникает следующая ситуация. Имеется некоторая функция  $y(x)$ ; мы наблюдаем не ее, а функцию  $f(x) = \int_a^b K(x, s)y(s) ds$ , причем в значения этой функции вносятся возмущения  $\delta f(x)$ .

Таким образом, задача

$$\int_a^b K(x, s)y(s) ds = f(x)$$

имеет решение, но нам реально требуется решать задачу

$$\int_a^b K(x, s)\tilde{y}(s) ds = \tilde{f}(x), \quad (2)$$

где  $\tilde{f}(x) = f(x) + \delta f(x)$ ; норма  $\delta f$ , погрешности измерения  $f(x)$ , мала:

$$\|\delta f(x)\| \leq \varepsilon. \quad (3)$$

Разность между решениями задач (2) и (1), которую можно записать в виде  $\delta y(x) = \tilde{y}(x) - y(x)$ , является решением интегрального уравнения

$$\int_a^b K(x, s)\delta y(s) ds = \delta f(x). \quad (4)$$

Пусть  $\delta f(x) = \sum_{n=1}^{\infty} \alpha_n \varphi_n(x)$ . Условие (3) означает, что  $\sum_{n=1}^{\infty} |\alpha_n|^2 \leq \varepsilon^2$ .

Рассмотрим сначала случай, когда все  $\lambda_n \neq 0$ . Если ряд  $\sum_{n=1}^{\infty} \left| \frac{\alpha_n}{\lambda_n} \right|^2$  расходится, то уравнение (4) не имеет решения. Если даже этот ряд сходится, то мы не можем гарантировать, что погрешность  $\delta y(x)$  будет стремиться к нулю при  $\varepsilon \rightarrow 0$ .

В самом деле, среди всех правых частей  $\delta f$  с  $\|\delta f\| \leq \varepsilon$  имеется правая часть  $\delta f_n = \varepsilon \varphi_n(x)$ , соответствующая такому  $n$ , что  $|\lambda_n| < \varepsilon$ . Тогда  $\delta y = \frac{\varepsilon}{\lambda_n} \varphi_n(x)$ , т. е.  $\|\delta y\| = \varepsilon / |\lambda_n| > 1$ .

Однако никто не обязывает нас непосредственно решать задачу (2) с возмущенной правой частью. Можно попытаться заменить эту задачу некоторой «близкой» задачей, решение которой будет «близко» к  $y(x)$ . Переход к такой «близкой» задаче называют *регуляризацией* (по Тихонову).

Мы уже изучали некоторые способы регуляризации на примере решения систем линейных уравнений.

При решении интегральных уравнений Фредгольма первого рода в качестве такой близкой к (1) задачи рассмотрим уравнение

$$\mu y_\mu(x) + \int_a^b K(x, s) y_\mu(s) ds = \tilde{f}(x), \quad \mu > 0; \quad (5)$$

параметр  $\mu$  иногда называют *параметром регуляризации*.

Теорема. Пусть все  $\lambda_n > 0$ ,

$$y(x) = \sum_{n=1}^{\infty} y_n \Phi_n(x), \quad \|y\|^2 = \sum_{n=1}^{\infty} |y_n|^2 < \infty.$$

Тогда справедливо неравенство

$$\|y_\mu - y\| \leq \omega(\varepsilon, \mu),$$

где  $\omega(\varepsilon, \mu) \rightarrow 0$  при  $\varepsilon/\mu, \mu \rightarrow 0$  ( $\omega(\varepsilon, \mu)$ , вообще говоря, зависит от  $y(x)$ ).

Доказательство. Сравним решение уравнения (5) с точным решением задачи (1). Имеем

$$f(x) = \sum_{n=1}^{\infty} \lambda_n y_n \Phi_n(x) = \sum_{n=1}^{\infty} f_n \Phi_n(x),$$

где  $f_n = \lambda_n y_n$ ,  $y_n = (y(x), \Phi_n(x))$  и

$$\tilde{f}(x) = \sum_{n=1}^{\infty} (f_n + \alpha_n) \Phi_n(x).$$

Подставляя  $y_\mu(x) = \sum_{n=1}^{\infty} y_n^\mu \Phi_n(x)$  в (5), получим

$$\sum_{n=1}^{\infty} (\mu + \lambda_n) y_n^\mu \Phi_n(x) = \sum_{n=1}^{\infty} (f_n + \alpha_n) \Phi_n(x).$$

Таким образом,

$$y_n^\mu = \frac{f_n + \alpha_n}{\mu + \lambda_n},$$

$$y_\mu(x) = \sum_{n=1}^{\infty} \frac{f_n + \alpha_n}{\mu + \lambda_n} \Phi_n(x) = \sum_{n=1}^{\infty} \frac{\lambda_n y_n + \alpha_n}{\mu + \lambda_n} \Phi_n(x).$$

Рассмотрим разность

$$R_\mu(x) = y_\mu(x) - y(x) = \sum_{n=1}^{\infty} \left( \frac{\lambda_n y_n + \alpha_n}{\mu + \lambda_n} - y_n \right) \Phi_n(x).$$

Имеем равенство

$$\frac{\lambda_n y_n + \alpha_n}{\mu + \lambda_n} - y_n = \frac{\alpha_n - \mu y_n}{\mu + \lambda_n}.$$



Таким образом, погрешность  $R_\mu(x)$  можно представить в виде суммы двух слагаемых  $R_\mu^1(x)$  и  $R_\mu^2(x)$ :

$$R_\mu(x) = R_\mu^1(x) + R_\mu^2(x), \quad (6)$$

где

$$R_\mu^1(x) = \sum_{n=1}^{\infty} \frac{\alpha_n}{\mu + \lambda_n} \varphi_n(x),$$

$$R_\mu^2(x) = \sum_{n=1}^{\infty} \frac{-\mu y_n}{\mu + \lambda_n} \varphi_n(x).$$

Вследствие ортонормированности системы функций  $\varphi_n(x)$  имеем

$$\|R_\mu^1\| = \sqrt{\sum_{n=1}^{\infty} \left| \frac{\alpha_n}{\mu + \lambda_n} \right|^2},$$

$$\|R_\mu^2\| = \sqrt{\sum_{n=1}^{\infty} \left| \frac{-\mu y_n}{\mu + \lambda_n} \right|^2}.$$

Поскольку  $\mu, \lambda_n \geq 0$ , то  $\mu + \lambda_n \geq \mu$  и поэтому

$$\|R_\mu^1\| \leq \sqrt{\sum_n \left| \frac{\alpha_n}{\mu} \right|^2} = \frac{1}{\mu} \|\delta f\| \leq \frac{\varepsilon}{\mu}. \quad (7)$$

Перейдем к оценке  $\|R_\mu^2\|$ . Рассмотрим сначала более простой и относительно часто встречающийся случай, когда дополнительно выполнено условие

$$\sqrt{\sum_{n=1}^{\infty} \left| \frac{y_n}{\lambda_n} \right|^2} = F_0 < \infty. \quad (8)$$

Тогда

$$\|R_\mu^2\| \leq \sqrt{\sum_{n=1}^{\infty} \left| \frac{\mu y_n}{\lambda_n} \right|^2} = \mu F_0. \quad (9)$$

Из соотношений (6), (7), (9) следует, что

$$\|R_\mu\| \leq \omega(\varepsilon, \mu) = \frac{\varepsilon}{\mu} + \mu F_0. \quad (10)$$

Следовательно, при дополнительном предположении (8) теорема доказана, поскольку

$$\|R_\mu\| \leq \omega(\varepsilon, \mu) = \frac{\varepsilon}{\mu} + \mu F_0,$$

и  $\omega(\varepsilon, \mu) \rightarrow 0$  при  $\varepsilon/\mu, \mu \rightarrow 0$ .

Таким образом, при достаточно малых  $\mu$ ,  $\varepsilon/\mu$  мы получаем решение задачи с малой погрешностью.

Для получения наилучшей оценки стремления погрешности к нулю найдем  $\min_{\mu} \omega(\varepsilon, \mu)$ . В точке минимума  $\mu = \mu_0$  имеем

$$\omega'_{\mu} = -\frac{\varepsilon}{\mu_0^2} + F_0 = 0, \text{ т. е. } \mu_0 = \sqrt{\varepsilon/F_0}. \text{ Из (10) получаем, что при } \mu = \sqrt{\varepsilon/F_0}$$

$$\|R_{\mu}\| \leq 2 \sqrt{\varepsilon F_0} \rightarrow 0.$$

Проведем теперь доказательство теоремы без предположения (8).

Представим выражение  $\|R_{\mu}^{(2)}\|$  в виде

$$\|R_{\mu}^{(2)}\| = \sqrt{R_{\mu, N}^{(1)} + R_{\mu, N}^{(2)}},$$

где

$$R_{\mu, N}^{(1)} = \sum_{n=1}^N \left| \frac{-\mu y_n}{\mu + \lambda_n} \right|^2,$$

$$R_{\mu, N}^{(2)} = \sum_{n=N+1}^{\infty} \left| \frac{-\mu y_n}{\mu + \lambda_n} \right|^2.$$

Справедливы оценки

$$R_{\mu, N}^{(1)} \leq \sum_{n=1}^N \left| \frac{\mu y_n}{\lambda_n} \right|^2 = \mu^2 F_1^2(N),$$

где  $F_1^2(N) = \sum_{n=1}^N \left| \frac{y_n}{\lambda_n} \right|^2$ , и

$$R_{\mu, N}^{(2)} \leq \sum_{n=N+1}^{\infty} |y_n|^2.$$

Покажем, что  $\|R_{\mu}^2\| \rightarrow 0$  при  $\mu \rightarrow 0$ . Для этого достаточно показать, что для любого  $\delta > 0$  существует  $\mu(\delta)$ , такое, что

$$\|R_{\mu}^2\| \leq \delta \text{ при } \mu \leq \mu(\delta).$$

Возьмем произвольное  $\delta > 0$ . Поскольку ряд  $\sum_{n=1}^{\infty} |y_n|^2$  сходится, то существует  $N(\delta)$  такое, что

$$\sum_{n=N(\delta)+1}^{\infty} |y_n|^2 \leq \frac{\delta^2}{2}.$$

Если  $\mu^2 \leq (\mu(\delta))^2 = \delta^2 / (2(F_1(N(\delta)))^2)$ , то  $R_{\mu, N}^{(1)} \leq \delta^2/2$  и

$$\|R_{\mu}^2\| = \sqrt{R_{\mu, N}^{(1)} + R_{\mu, N}^{(2)}} \leq \sqrt{\delta^2/2 + \delta^2/2} \leq \delta.$$

Таким образом, имеем

$$\begin{aligned} \|R_\mu\| &\leq \|R_\mu^1\| + \|R_\mu^2\|, \\ \|R_\mu^1\| &\leq \varepsilon/\mu, \quad \|R_\mu^2\| \rightarrow 0 \quad \text{при } \mu \rightarrow 0. \end{aligned}$$

Теорема доказана и без предположения (8).

Описанный выше метод регуляризации применим и в случае, когда некоторые из  $\lambda_n$  могут обращаться в нуль.

Пусть  $N_1$  — множество  $n$  таких, что  $\lambda_n > 0$ ,  $N_0$  — множество  $n$  таких, что  $\lambda_n = 0$ . (Каждое из этих множеств может быть как конечным, так и бесконечным; одновременно оба конечными быть не могут.)

Если  $y(x) = \sum_n y_n \Phi_n(x)$ ,  $\sum_n |y_n|^2 < \infty$  — решение уравнения (1), то при  $\sum_{n \in N_0} |a_n|^2 < \infty$  функция  $\sum_n (y_n + a_n) \Phi_n(x)$  также будет решением уравнения (1). Положим  $y^0(x) = \sum_{n \in N_1} y_n \Phi_n(x)$ ; согласно вышесказанному  $y^0(x)$  также является решением уравнения (1).

**Задача 7.** Пусть  $y(x) \neq y^0(x)$  — решение уравнения (1). Показать, что

$$\|y\| > \|y^0\|. \quad (11)$$

Решение уравнения (1) с минимальной нормой (в случае неединственности решения) называется *нормальным*. Из (11) следует, что  $y^0(x)$  — *нормальное решение задачи*.

**Теорема.** Пусть  $y^0(x) = \sum_{n \in N_1} y_n \Phi_n(x)$ ,  $\|y^0\|^2 = \sum_{n \in N_1} |y_n|^2 < \infty$  — решение уравнения (1). Тогда справедливо неравенство  $\|y_\mu - y^0\| \leq \omega(\varepsilon, \mu)$ , где  $\omega(\varepsilon, \mu) \rightarrow 0$  при  $\varepsilon/\mu, \mu \rightarrow 0$ .

Доказательство несущественным образом отличается от проведенного выше в случае, когда все  $\lambda_n > 0$ . Решение  $y^0(x)$  задачи (1) записывается в виде

$$y^0(x) = \sum_{n \in N_1} y_n^0 \Phi_n(x) = \sum_{n=1}^{\infty} Y_n \Phi_n(x),$$

$$\text{где } Y_n = \begin{cases} y_n^0 & \text{при } n \in N_1, \\ 0 & \text{при } n \in N_0. \end{cases}$$

Погрешность  $y_\mu(x) - y^0(x)$  запишется в виде

$$y_\mu(x) - y^0(x) = \sum_{n=1}^{\infty} \left( \frac{Y_n + \alpha_n}{\mu + \lambda_n} - Y_n \right) \Phi_n = S_1 + S_2,$$

где

$$S_1 = \sum_{n \in N_1} \left( \frac{y_n + \alpha_n}{\mu + \lambda_n} - y_n \right) \Phi_n, \quad S_2 = \sum_{n \in N_0} \frac{\alpha_n}{\mu} \Phi_n.$$

Оценка для слагаемого  $S_2$  очевидна:

$$\|S_2\| = \left\| \sum_{n \in N_0} \frac{\alpha_n}{\mu} \Phi_n \right\| = \sqrt{\sum_{n \in N_0} \left| \frac{\alpha_n}{\mu} \right|^2} \leq \frac{1}{\mu} \sqrt{\sum_{n=1}^{\infty} |\alpha_n|^2} = \frac{\varepsilon}{\mu}.$$

Оценка для  $S_1$  производится так же, как в случае доказанной ранее теоремы

Метод регуляризации применяется для решения самых разнообразных задач, в частности нелинейных.

Рассмотрим случай, когда ядро  $K(x, s)$  не симметрично. Определим оператор  $Q^*$  соотношением

$$Q^*y = \int_a^b K(s, x) y(s) ds.$$

Обозначим через  $y_\mu(x)$  решение уравнения

$$\mu y_\mu(x) + Q^*Qy_\mu = Q^*\bar{f}(x).$$

Выберем параметр  $\mu = \mu(\varepsilon)$  из условия

$$\|Qy_{\mu(\varepsilon)} - \bar{f}\| = \varepsilon.$$

Справедлива

Теорема (без доказательства). Пусть уравнение (1) разрешимо и  $y^0(x)$  — нормальное решение уравнения (1), т. е. решение с минимальной нормой. Тогда  $\|y_{\mu(\varepsilon)} - y^0\| \rightarrow 0$  при  $\varepsilon \rightarrow 0$ .

Литература к гл. XI: [6, 34, 62].

## ЗАКЛЮЧЕНИЕ

Мы закончили обсуждение традиционных вопросов теории численных методов. При этом мы, возможно слишком часто, обращали внимание на «подводные камни», встречающиеся при применении того или иного численного метода. У читателя могло возникнуть превратное впечатление, что применение численных методов для решения реальных задач настолько сложная и безнадёжная задача, что от него следует отказаться. Чтобы исправить такое впечатление, посмотрим на этот же вопрос с оптимистических позиций.

Существует большое число задач, где есть хорошо отработанные численные методы и созданные на их основе стандартные программы решения задач. Например, в НИВЦ МГУ разработана Библиотека программ решения типовых задач численного анализа на стандарте языка фортран [12]. В настоящее время выпущено в свет 5-е издание Библиотеки, состоящее из 1200 весьма эффективных стандартных программ, предназначенных для решения широкого круга типичных задач. Сюда относятся задачи вычисления значений специальных функций, приближения функций, вычисления интегралов не очень высокой кратности, задачи линейной алгебры умеренной размерности, задача Коши и линейные краевые задачи для обыкновенных дифференциальных уравнений, некоторые задачи обработки результатов наблюдений, задачи минимизации функций и т. д. Данная Библиотека разработана в версиях для ЭВМ БЭСМ-6, ЕС и СМ.

Стандартные программы решения многих типов прикладных задач входят в математическое обеспечение, поставляемое вместе с ЭВМ.

Если Вам впервые встретилась единичная задача, то, как правило, целесообразнее воспользоваться стандартной программой или самому составить программу, основанную на простейшем методе решения. При решении единичных задач, требующих умеренного объема вычислений, часто идут на более чем 100-кратное увеличение объема вычислений по сравнению с наиболее эффективными методами, лишь бы побыстрее получить результат, воспользовавшись при этом стандартной программой или алгоритмами, для реализации которых можно быстро составить и отладить программу.

Нам регулярно встречались задачи минимизации функций большого числа переменных, в которых применение стандартных программ не приводило к положительному результату. Однако мы не теряли оптимизма и обычно при столкновении с проблемой решения новых типов задач минимизаций начинали именно с попыток использования стандартных программ. Если, например, в 30 % случаев применение стандартных программ оказывалось эффективным, то их применение заведомо можно считать оправданным даже и в том случае, когда использование стандартной программы не приводит к положительному результату; при обращении к стандартной программе будет составлено большое число блоков окончательной программы и зачастую будет накоплена полезная информация о свойствах минимизируемой функции.

Среди «пионеров» использования вычислительной техники встречается убеждение, что чужими стандартными программами пользоваться нельзя и программу всегда следует писать самому, заново. Такое суждение было оправдано на первоначальном этапе использования ЭВМ, когда теория численных методов была развита недостаточно и созданные на ее основе алгоритмы часто были ненадежными. При современном уровне теории численных методов и жестких требованиях к опробованию, так называемому *тестированию*, стандартных программ с такой точкой зрения нельзя согласиться.

Рассмотрим ситуацию, когда требуется путем численного эксперимента исследовать какой-либо физический процесс. Часто, набив много шишек и потратив иногда годы на расчет сложных моделей, начинающий исследователь приходит к пониманию того, что целесообразнее было начинать с расчета простейшей модели и изучать ее с помощью простейших проверенных методов, иногда требующих повышенных вычислительных затрат. Лишь в случае полного доверия к постановке задачи имеет смысл заниматься расчетом сложной модели с применением громоздких по своей структуре методов.

Таким образом, на первоначальном этапе исследования задачи обычно мы имеем дело с простейшей моделью и простейшими методами решения, часто основанными на использовании стандартных программ. Если же мы начинаем исследование задачи с изучения сложных моделей, то скорее всего мы поступаем в чем-то неправильно. Целесообразнее переходить к рассмотрению более сложных задач и применению более сложных методов, имея за плечами опыт использования ЭВМ при решении простейших задач простейшими методами, поскольку при такой последовательности действий использование численных методов и ЭВМ уже не будет казаться чрезмерно трудной проблемой.

Рассмотрение самими математиками постановки прикладной задачи с ее истоков, с простейших моделей, еще важно и в связи со следующим обстоятельством. Часто, стремясь «приспособить»

задачу для численного решения, специалист, не знакомый с численными методами и возможностями ЭВМ, исходит лишь из сложности внешнего вида математической постановки. В результате этого иногда

а) происходит замена исходной задачи задачей, не имеющей к ней отношения,

б) задача, передающаяся численному решению при возможностях современных ЭВМ, становится задачей, не поддающейся такому решению.

Еще один довод в пользу более всестороннего изучения постановки задачи состоит в следующем. Часто в процессе построения простейшей модели мы получаем представление о том, в каком направлении будет идти усложнение метода и программы решения задачи. В результате этого мы сможем заложить в исходный вариант программы, предназначенной для исследования простейшей модели, возможности для дальнейшего усложнения программы.

Во всех случаях целесообразно составлять программу блочно. Дело в том, что при решении сколько-нибудь сложных задач мы заранее часто не можем сказать, будет ли выбранный нами метод решения подходящим для решения этой задачи.

Например, при решении краевой задачи для дифференциальных уравнений у нас может быть неуверенность в разумности выбора разностной схемы во внутренних точках или вблизи границы. Вычисления во внутренних точках и вблизи границы ведутся по различным формулам и поэтому лучше, чтобы они были выделены в отдельные блоки с тем, чтобы можно было производить независимую *отладку методов*, с помощью которых ведутся расчеты уравнения и граничных условий. Также важно предусмотреть возможность быстрого и удобного изменения параметров задачи, например шага сетки при решении дифференциальных уравнений или числа узлов квадратуры. Отладка программы и апробация метода решения задачи при малом числе узлов проходят быстрее и позволяют лучше использовать возможности ЭВМ. Например, при малом числе узлов можно быстро проверить сходимость итерационного процесса.

Иногда профессиональные программисты рекомендуют проводить отладку программы при помощи *тестов*. Берется минимально допустимое число узлов, при котором включаются в работу все блоки программы (и все циклы). При таком числе узлов задача решается без помощи написанной программы, а затем результаты расчета сравниваются с результатами расчета с помощью ЭВМ. Для типичных конечно-разностных методов решения дифференциальных уравнений минимальное число узлов по каждой оси, при котором включаются в работу все блоки и циклы, находится в пределах от 2 до 5.

В процессе работы над отладкой программы в программу вносятся дополнительно команды выдачи на дисплей или на печать промежуточных данных с целью сравнения их с заранее

просчитанным тестом. Такой режим отладки особенно удобен для начинающего программиста.

Специалисты, имеющие большой опыт работы с ЭВМ, предпочитают отлаживать программу крупными модулями, сочетая отладку программы с проверкой качества метода (отладкой метода). Для этого стараются строить программу решения так, чтобы при определенных значениях параметров она превращалась в программу решения задачи с известным решением.

Например, уравнения движения снежной лавины (7.6.1) при определенных значениях параметров превращаются в уравнения мелкой воды, для которых в случае кусочно-постоянных начальных данных известно точное решение (так называемое решение задачи о распаде разрыва). Сравнивая решение, полученное в результате расчетов, с точным решением, можно судить о правильности программы и качестве метода.

Построение решения задачи о распаде разрыва само по себе требует относительно большой вспомогательной работы. Вместо непосредственного построения такого решения для заданной системы часто поступают следующим образом.

Вместо конкретных значений коэффициентов системы ставятся произвольные параметры, производится одновременная подборка этих параметров и системы функций, которая может быть решением задачи о распаде разрыва. В результате получится некоторая задача с известным решением, отличающаяся от интересующей нас задачи лишь значениями числовых параметров. На такой задаче мы можем отладить все наиболее существенные моменты метода и программы.

Проиллюстрируем способы построения задач с известным частным решением на примере дифференциальных уравнений.

Пусть решается краевая задача

$$L(u) = \frac{\partial}{\partial x} \left( (1 + \gamma u^2) \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y} \left( (1 + \gamma) \frac{\partial u}{\partial y} \right) - e^{x+y} = 0$$

в квадрате  $G: 0 < x, y < 1$ ; на границе квадрата  $\Gamma$  ставится краевое условие:

$$\left( \frac{\partial u}{\partial n} + \sigma u - g \right) \Big|_{\Gamma} = 0.$$

Возьмем некоторую функцию  $u^0(x, y)$  и вычислим функции

$$f^0(x, y) = L(u^0(x, y)), \quad g^0 \Big|_{\Gamma} = \left( \frac{\partial u^0}{\partial n} + \sigma u^0 - g \right) \Big|_{\Gamma}.$$

Краевая задача

$$L(u) - f^0(x, y) = 0, \quad \left( \frac{\partial u}{\partial n} + \sigma u - g - g^0 \right) \Big|_{\Gamma} = 0$$

будет задачей с известным точным решением  $u = u^0(x, y)$ .

Часто желательно иметь в распоряжении сеточную задачу с известным точным решением. Если, например, по простейшим



явным формулам решается задача Коши, то такое решение легко вычислить непосредственно.

В случае краевых задач или при решении задачи Коши по неявным формулам часто целесообразно поступить как и выше. Строим сеточную задачу  $L_h(u_h) = 0$ ,  $l_h(u_h) = 0$  (с небольшим числом узлов), берем некоторую сеточную функцию  $u_h^0$  и вычисляем

$$f_h^0 = L_h(u_h^0), \quad g_h^0 = l_h(u_h^0).$$

Сеточной задачей с известным точным решением  $u_h = u_h^0$  является задача

$$L_h(u_h) = f_h^0, \quad l_h(u_h) = g_h^0.$$

Иногда возникает необходимость иметь в распоряжении сеточную задачу с большим числом узлов и с точным решением. В этом случае может оказаться целесообразным поручить ЭВМ вычисление  $f_h^0$  и  $g_h^0$ . Часто такую задачу получают, взяв в качестве  $u_h^0$  множество значений в узлах сетки некоторого многочлена невысокой (обычно 1—4-й) степени. Степень многочлена выбирается из условия, чтобы погрешность аппроксимации дифференциальной задачи сеточной равнялась нулю.

Еще раз выделим основную идею описанного выше приема построения тестов. *Вместо непосредственного построения теста для исходной задачи (построение решения по заданной правой части) строится тест для обратной задачи (по заданному решению строится правая часть) и он оказывается тестом для прямой задачи с той же структурой, но с другими числовыми данными.*

Применение этого приема обычно существенно сокращает затраты на построение теста.

## СПИСОК ЛИТЕРАТУРЫ

1. Абрамов А. А. О численном решении некоторых алгебраических задач, возникающих в теории устойчивости//ЖВМ и МФ. — 1984. — 24, № 3. — С. 339—347.
2. Бахвалов Н. С. Об оптимальных оценках скорости сходимости квадратурных процессов и методов интегрирования типа Монте-Карло на классах функций//Численные методы решения дифференциальных и интегральных уравнений и квадратурные формулы. — М.: Наука, 1964. — С. 5—63.
3. Бахвалов Н. С. Численные методы. — М.: Наука, 1975.
4. Бейкер Дж., Грейвс—Моррис П. Аппроксимации Паде.—М.: Мир, 1986.
5. Березин И. С., Жидков Н. П. Методы вычислений. Т. 1.—М.: Наука, 1966.
6. Березин И. С., Жидков Н. П. Методы вычислений. Т. 2 — М.: Физматгиз, 1962.
7. Васильев Ф. П. Численные методы решения экстремальных задач. — М.: Наука, 1980.
8. Васильев Ф. П. Методы решения экстремальных задач. — М.: Наука, 1981.
9. Воеводин В. В. Численные методы алгебры. Теория и алгоритмы. — М.: Наука, 1966.
10. Воеводин В. В. Вычислительные основы линейной алгебры. — М.: Наука, 1977.
11. Воеводин В. В., Кузнецов Ю. А. Матрицы и вычисления. — М.: Наука, 1984.
12. Воеводин В. В., Арушанян О. Б. Структура и организация Библиотеки численного анализа НИВЦ МГУ//Численный анализ на фортране. Вычислительные методы и инструментальные системы. — М.: Изд-во МГУ, 1979.
13. Волков Е. А. Численные методы. — М.: Наука, 1982.
14. Годунов С. К., Забродин А. В. О разностных схемах второго порядка точности для многомерных задач//ЖВМ и МФ. — 1962. — 2, № 4. С. 706—708.
15. Годунов С. К., Рябенский В. С. Введение в теорию разностных схем. — М.: Наука, 1962.
16. Годунов С. К., Забродин А. В. и др. Численное решение многомерных задач газовой динамики. — М.: Наука, 1976.
17. Годунов С. К., Рябенский В. С. Разностные схемы. — М.: Наука, 1977.
18. Годунов С. К. Решение систем линейных уравнений. — Новосибирск: Наука, 1980.
19. Джордж А., Лю Д. Численное решение больших разреженных систем уравнений. — М.: Мир, 1984.
20. Дробышевский В. И., Дымников В. П., Ривин Г. С. Задачи по вычислительной математике. — М.: Наука, 1980.
21. Дьяконов Е. Г. О построении итерационных методов на основе использования операторов, эквивалентных по спектру//ЖВМ и МФ. — 1966, — 6, № 1, С. 12—34.

22. Дьяконов Е. Г. О некоторых модификациях проекционно-разностных методов//Вестник МГУ. Сер. вычислит. мат. и киберн. — 1977. — 1, № 2, С. 3—19.
23. Завьялов Ю. С., Квасов Б. И., Мирошниченко В. Л. Методы сплайн-функций. — М.: Наука, 1980.
24. Зенкевич О., Морган К. Конечные элементы и аппроксимация. — М.: Мир, 1980.
25. Икрамов Х. Д. Численное решение матричных уравнений. — М.: Наука, 1984.
26. Карманов В. Г. Математическое программирование. — М.: Наука, 1986.
27. Кобельков Г. М. Решение задачи о стационарной свободной конвенции//ДАН СССР. — 1980. — 255, № 2. С. 277—282.
28. Кобельков Г. М. Расчет уравнений Навье — Стокса при больших числах Рейнольдса.//ЖВМ и МФ. — 1984. — 24, № 2. С. 294—304.
29. Колмогоров А. Н., Тихомиров В. М.  $\epsilon$ -энтропия и  $\epsilon$ -емкость множеств в функциональных пространствах//УМН. — 1959. — 14. вып. 2. С. 3—86.
30. Копченова Н. В., Марон И. А. Вычислительная математика в примерах и задачах. — М.: Наука, 1972.
31. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Дифференциальные уравнения. — Минск: Наука и техника, 1982.
32. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Уравнения в частных производных. — Минск: Наука и техника, 1982.
33. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Линейная алгебра и нелинейные уравнения. — Минск: Наука и техника, 1982.
34. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Интегральные уравнения, некорректные задачи и улучшение сходимости. — Минск: Наука и техника, 1984.
35. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Интерполирование и интегрирование. — Минск: Наука и техника, 1983.
36. Крылов В. И., Шульгина А. Т. Справочная книга по численному интегрированию. — М.: Наука, 1966.
37. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т. 1. — М.: Наука, 1976.
38. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т. 2. — М.: Наука, 1977.
39. Ланцош К. Практические методы прикладного анализа. — М.: ГИФМЛ, 1961.
40. Лебедев В. И. Функциональный анализ и вычислительная математика. — М.: Изд-во ВИНТИ, 1987. — (ОБМ АН СССР).
41. Лоусон Ч., Хентон Р. Численное решение задач метода наименьших квадратов. — М.: Наука, 1986.
42. Мак-Кракен Д., Дорн У. Численные методы и программирование на фортране. — М.: Мир, 1977.
43. Марчук Г. И. Методы вычислительной математики. — М.: Наука, 1980.
44. Марчук Г. И., Агошков В. И. Введение в проекционно-сеточные методы. — М.: Наука, 1981.
45. Марчук Г. И., Лебедев В. И. Численные методы в теории переноса нейтронов. — М.: Атомиздат, 1981.
46. Марчук Г. И., Шайдуров В. В. Повышение точности решений разностных схем. — М.: Наука, 1979.
47. Мысовских И. П. Интерполяционные кубатурные формулы. — М.: Наука, 1981.
48. Никольский С. М. Квадратурные формулы. — М.: Наука, 1979.
49. Никифоров А. Ф., Суслев С. К., Уваров В. Б. Классические ортогональные полиномы дискретной переменной. — М.: Наука, 1985.

50. Парлетт Б. Симметричная проблема собственных значений. — М.: Мир, 1983.
51. Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г. Численные методы решения жестких систем. — М.: Наука, 1979.
52. Рябенский В. С., Филиппов А. Ф. Об устойчивости разностных уравнений. — М.: Гостехиздат, 1956.
53. Самарский А. А. Теория разностных схем. — М.: Наука, 1982.
54. Самарский А. А., Андреев В. Б. Разностные методы для решения эллиптических уравнений. — М.: Наука, 1976.
55. Самарский А. А., Гулин А. В. Устойчивость разностных схем. — М.: Наука, 1973.
56. Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений. — М.: Наука, 1978.
57. Самарский А. А., Попов Ю. П. Разностные методы решения задач газовой динамики. — М.: Наука, 1980.
58. Саульев В. К. Интегрирование уравнений параболического типа методом сеток. — М.: Физматгиз, 1960.
59. Современные численные методы решения обыкновенных дифференциальных уравнений//Под ред. Дж. Холла, Дж. Уатта. — М.: Мир, 1979.
60. Стечкин С. Б., Субботин Ю. Н. Сплайны в вычислительной математике. — М.: Наука, 1976.
61. Стрэнг Г., Фикс Дж. Теория метода конечных элементов. — М.: Мир, 1977.
62. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. — М.: Наука, 1986.
63. Уилкинсон Дж., Райни К. Справочник алгоритмов на языке Алгол. Линейная алгебра. — М.: Машиностроение, 1976.
64. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. — М.: Физматгиз, 1963.
65. Форсайт Дж. и др. Машинные методы математических вычислений. — М.: Мир, 1980.
66. Яненко Н. Н. Метод дробных шагов решения многомерных задач математической физики. — Новосибирск: Наука, 1967.
67. Dahlquist Y. Stability and error bounds in the numerical integration of ordinary differential equations. — Uppsala, Almqvist & Wiksells boktr 130 (1959). P. 5—92.
68. Butcher I. G. A modified multistep method for the numerical integration of ordinary differential equations//J. Assoc. Comput. Math. — 1965. — 12, N 1. — P. 124—135.
69. Stroud A. H. and Secrest D. Gaussian, Quadrature Formulas. — Englewood Cliffs, N. Y.: Prentice-Hall, 1966.

## ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

- Алгоритм ненасыщаемый 62  
Аппроксимация 487  
— дифференциального уравнения разностной схемой 403
- Большое число 22
- Валле — Пуссена теорема 174  
Вариация 506  
Верная цифра 23  
Весовая функция 85  
Восполнение 541  
Выделение весовой функции 147
- Гарантированные оценки погрешности на классе функций 238  
Главный член погрешности 136
- Дивергентность разностной схемы 511  
Дискретные коэффициенты Фурье 169  
Дифференцирование численное 73
- Замыкание алгоритма 422  
— — нерегулярное 422  
— — регулярное 422  
Значащие цифры 23
- Интегрирование осциллирующих функций 113  
— системы уравнений 392  
Интерполирование 35  
— с кратными узлами 47  
Интерполяционная формула Лагранжа 40  
— — Ньютона 45  
— — для интерполирования вперед 67  
— — — — — назад 67  
Интерполяция 35  
— квадратичная 69  
— линейная 69  
— тригонометрическая 169
- Итерационные методы построения многочлена наилучшего равномерного приближения 182  
— — с использованием спектрально эквивалентных операторов 299
- Квадратуры Гаусса 103  
— Грегори 139  
— Лобатто 107  
— Ньютона — Котеса 91  
— обобщенные 118  
— прямоугольников 83  
— Ромберга 145  
— Симпсона 85  
— составные 118  
— трапеций 84  
— Филона 114  
— Эйлера 108
- Количество арифметических операций 40  
Конечно-разностное уравнение 50  
Конечно-разностные методы 372  
— — схемы 372
- Ленточная структура 260  
Линейная оценка погрешности 27  
Линейное разностное уравнение 50
- Мажорирующее разностное уравнение 474  
Матрица Грама 163  
— ортогонализация 97  
— отражений 266  
Мера обусловленности матрицы 304  
— — системы 304  
— погрешности аппроксимации 489  
Метод Адамса 374  
— вариационно-разностный 465  
— вилки 329  
— Гаусса 257  
— — с выбором главного элемента 261  
— градиентного спуска 330  
— дробных шагов 561  
— исключения Гаусса 257  
— квадратного корня 262  
— конечных разностей 540  
— — элементов 469  
— Ланцоша 298  
— Монте-Карло 238  
— наименьших квадратов 210  
— наискорейшего спуска 293  
— неопределенных коэффициентов 38, 209

- Метод Ньютона решения нелинейных уравнений 323
- оврагов 335
  - оптимальный 61
  - парабол 329
  - координатного спуска 290
  - пристрелки 414
  - прогонки 416
  - проекционно-разностный 544
  - простой итерации 269, 319
  - регуляризации 212
  - релаксации 291
  - Ритца 463
  - сверхрелаксации 292
  - секущих 327
  - сопряженных градиентов 298
  - спуска 329
  - стрельбы 414
  - суммарной аппроксимации 561
  - установления 337
  - тяжелого шарика 339
  - циклической прогонки 419
  - штрафа 332
  - Эйлера 360
- Методы интерполяционные 373
- Рунге-Кутта 360
  - экстраполяционные 372
- Многочлен наилучшего равномерного приближения 174
- Многочлены Лагерра 101
- Лежандра 100
  - наименее уклоняющийся от нуля 58
  - Чебышева 56
  - — второго рода 100
  - — первого рода 100
  - Эрмита 101
  - Якоби 100
- Наилучшее равномерное приближение 174
- Недетерминированный метод 248
- Некорректные задачи 405
- Неравенство Бесселя 164
- Чебышева 239
- Неустойчивость 34
- Неявная схема, метод 373, 513
- Норма энергетическая 537
- Нормы векторов и матриц 255
- эквивалентные 270
- е-неравенство 521
- Область зависимости 483
- сходимости метода 318
- Обобщенное решение 540
- Обратная интерполяция 73
- Обратный ход метода Гаусса 258
- — прогонки 416
- Обусловленность матрицы системы 304
- системы 304
- Однородные схемы 452
- Односторонние формулы численного дифференцирования 78
- Одношаговые методы численного интегрирования 367
- Оператор расщепляющийся 555
- Оптимальные квадратуры 126
- Оптимальный линейный итерационный процесс 281, 285
- по порядку итерационный процесс 41, 61
- Оптимизация методов 61
- оценки погрешности интерполирования 60
  - распределения узлов интегрирования 125
  - скорости сходимости итерационного процесса 278
- Ортогональная система 96
- Ортогональные многочлены 98
- Ортонормированная система элементов 98
- Остаточный член формулы Лагранжа 42
- Очень большое число 23
- Параметр регуляризации 583
- Планирование эксперимента 36
- Плохо обусловленные системы 305
- Повышение порядка точности разностной схемы 406
- Погрешность абсолютная 22
- аппроксимации дифференциального уравнения разностной схемой 488, 517
  - вычислительная 17
  - квадратуры на классе функций 125
  - математической модели 17
  - метода 17
  - — на классе задач 61
  - — на шаге 368, 370
  - неустранимая 17
  - относительная 26
  - предельная 26
- Порядок метода 322
- погрешности аппроксимации 377
- Почти собственные значения 273
- Преобразование Фурье быстрое 171
- — дискретное 169
- Принцип замороженных коэффициентов 504
- Проблема собственных значений полная 309
- — — частичная 309
- Прямое произведение формул интегрирования, интерполирования, дифференцирования 225
- Прямой ход метода Гаусса 258
- — — прогонки 416
- $\delta^2$ -процесс ускорения сходимости 276

- Разности вперед 63
- высшего порядка 63
- конечные 63
- назад 63
- разделенные 42
- центральные 63
- Разностная схема 488
- — экономичная 550
- Регуляризация 582
- Ряд Фурье дискретный, конечный 169
  
- Сеточная функция Грина 409
- Сжатое отображение 319
- Симметризация системы уравнений 279
- Система жесткая 394
- Слой 514
- Согласованные нормы 486
- Спектральный признак устойчивости 491, 499
- Сплайн 194
- интерполяционный 200
- локальный 201
- Стандартные программы 46
- Строго нормированное пространство 161
- Схема Эйткена 45
- Схемы, точные на решениях специального вида 444
- Сходимость 486
  
- Таблица разделенных разностей 44
- Теорема Валле — Пуссена 174
- Чебышева 174
- Точки чебышевского альтернанса 174
- Триангуляция 541
- Тригонометрическая интерполяция 169
  
- Узел внутренний 527
- граничный 527
- нерегулярный 535
- приграничный 528
- регулярный 535
  
- Узел сетки 486
- Узлы интерполяции 35
- Уравнения в конечных разностях 50
- Условие  $\alpha$  382
- сильной минимальности 547
- Устойчивость 488
- безусловная 515
- по начальным данным 518
- условная 515
- Формула Абеля 450
- прямоугольников 83
- Ромберга 145
- Симпсона 85
- с кратными узлами 93
- с трапеций 84
- Филона 114
- Формулы Адамса 360, 375
- Грегори 138
- численного дифференцирования 74
- Функция сильно растущая 23
  
- Характеристическое уравнение 53
- — разностной схемы 380
- Хорошо обусловленная (поставленная) краевая задача 432
- обусловленные системы 305
  
- Шаблон 517
- Шаг таблицы 63
  
- Экстраполяция 35
- Элемент наилучшего приближения 160
- Элементарный треугольник 541
- Энергетическое неравенство 521
- тождество 520
- Энтропия 204
- е-энтропия 207
  
- Явная схема 372, 513
- Явный метод 372, 513

Николай Сергеевич Бахвалов  
Николай Петрович Жидков  
Георгий Михайлович Кобельков

## ЧИСЛЕННЫЕ МЕТОДЫ

Редактор Т. Н. Галишникова  
Художественный редактор Г. М. Коровина  
Технический редактор В. Н. Кондакова  
Корректоры Е. Ю. Рычагова, Л. С. Сомова

ИБ № 12379

Сдано в набор 23.12.86. Подписано в печать 01.07.87. Т-12157.  
Формат 60×90/16. Бумага типографская № 2. Гарнитура литературная. Печать высокая. Усл. печ. л. 37,5. Усл. кр.-отт. 37,5.  
Уч.-изд. л. 39,22. Тираж 35,500 экз. Заказ 412. Цена 1 р. 60 к.

Ордена Трудового Красного Знамени издательство «Наука»  
Главная редакция физико-математической литературы  
117071 Москва В-71, Ленинский проспект, 15

Ленинградская типография № 2 головное предприятие ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им. Евгении Соколовой Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли, 198052 г. Ленинград, Л-52, Измайловский проспект, 29,



ИЗДАТЕЛЬСТВО «НАУКА»  
ГЛАВНАЯ РЕДАКЦИЯ  
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ  
117071 Москва В-71, Ленинский проспект, 15

*Готовится к печати в 1988 году*

**ВАСИЛЬЕВ Ф. П. Численные методы решения экстремальных задач: Учебное пособие,— 2-е издание, переработанное и дополненное.**

Содержит основные численные методы решения экстремальных задач. Приводятся теоретическое обоснование и краткие характеристики этих методов. Рассматриваются задачи минимизации функций конечного числа переменных и задачи оптимального управления процессами, описываемыми системами обыкновенных дифференциальных уравнений.

Сохранена структура первого издания, но содержание некоторых глав существенно переработано и дополнено.

Первое издание — 1980 г.

Для студентов вузов по специальности «Прикладная математика», а также для специалистов, связанных с решением задач оптимизации.