

# ВВЕДЕНИЕ В ЧИСЛЕННЫЕ МЕТОДЫ

Самарский А. А.

Книга написана на основе курса лекций, читавшихся автором па факультете вычислительной математики и кибернетики МГУ, и предназначается для ознакомления с началами численных методов. Теория численных методов излагается с использованием элементарных математических средств, а для иллюстрации качества методов используются простейшие математические модели.

В книге рассматриваются разностные уравнения, численные методы решения обыкновенных дифференциальных уравнений, линейных и нелинейных алгебраических уравнений, разностные методы для уравнений в частных производных.

Для студентов факультетов и отделений прикладной математики вузов.

## ОГЛАВЛЕНИЕ

Предисловие

Введение	7	§ 3. Консервативные разностные схемы	152
Глава I. Разностные уравнения	23	§ 4. Однородные схемы на	159
§ 1. Сеточные функции	23	неравномерных сетках	
§ 2. Разностные уравнения	26	§ 5. Методы построения разностных	167
§ 3. Решение разностных краевых задач	34	схем	
для уравнений второго порядка		Глава V. Задача Коши для	174
§ 4. Разностные уравнения как	38	обыкновенных дифференциальных	
операторные уравнения		уравнений	
§ 5. Принцип максимума для	55	§ 1. Методы Рунге — Кутта	174
разностных уравнений		§ 2. Многошаговые схемы. Методы	184
Глава II. Интерполяция и численное	61	Адамса	
интегрирование		§ 3. Аппроксимация задачи Коши для	195
§ 1. Интерполяция и приближение	61	системы линейных обыкновенных	
функций		дифференциальных уравнений первого	
§ 2. Численное интегрирование	70	порядка	
Глава III. Численное решение систем	85	§ 4. Устойчивость двухслойной схемы	200
линейных алгебраических уравнений		Глава VI. Разностные методы для	211
§ 1. Системы линейных алгебраических	85	эллиптических уравнений	
уравнений		§ 1. Разностные схемы для уравнения	211
§ 2. Прямые методы	91	Пуассона	
§ 3. Итерационные методы	96	§ 2. Решение разностных	221
§ 4. Двухслойная итерационная схема с	110	уравнений	
чебышевскими параметрами		Глава VII. Разностные методы решения	232
§ 5. Попеременно-треугольный метод	120	уравнения теплопроводности	
§ 6. Вариационно-итерационные методы	126	§ 1. Уравнение теплопроводности с	232
§ 7. Решение нелинейных уравнений	130	постоянными коэффициентами	
Глава IV. Разностные методы решения	137	§ 2. Многомерные задачи	243
краевых задач для обыкновенных		теплопроводности	
дифференциальных уравнений		§ 3. Экономичные схемы	250
§ 1. Основные понятия теории	137	Дополнение	260
разностных схем		Литература	266
§ 2. Однородные трехточечные	149	Предметный указатель	267
разностные схемы		Список обозначений	270

## ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Алгоритм неустойчивый 13	
— условно устойчивый 13	
— экономичный 11	
Аппроксимация разностная (на сотке) 138	
— суммарная 254	
Весовые множители 70	
Вычислительная неустойчивость 115	

Жесткие системы уравнений 192

Задача Дирихле 211

— корректная 14

— Коши 32

— краевая 32

— некорректная 15

— о собственных значениях 42

Интерполянта 61

- Интерполяционный полином 62
  - — Лагранжа 64
  - — Ньютона 64
- Интерполяция эрмитова 65
- Итерационные методы 90
- Итерационный метод двухшаговый (трехслойный) 97
  - — неявный 97
  - — одношаговый (двухслойный) 97
  - — явный 97
- Квадратурная формула 70
  - — Гаусса 82
  - — Котеса 74
  - — прямоугольника 71
  - — Симпсона 72
  - — трапеций 71
  - — Чебышева 83
- Коэффициенты Лагранжа 62
- Краевые условия 33
  - — 1-го рода 33
  - — 2-го рода 33
- Краевые условия 3-го рода 33
- Кубическая сплайн-интерполяция 65
- Линейно независимые векторы 39
  - — решения 27
- Линейное пространство 38
  - — действительное 38
  - — комплексное 38
- Мажорантная функция (мажоранта) 55
- Матрица верхняя треугольная 87
  - — диагональная 86
  - — ленточная 88
  - — нижняя треугольная 86
  - — разреженная 87
- Мера обусловленности 89
- Метод Адамса — Штёрмера 191
  - — баланса (интегро-интерполяционный) 167
  - — Бубнова — Галеркина 173
  - — вариационно-разностный 171
  - — вариационного типа 126
  - — верхней релаксации 101
  - — дихотомии 130
  - — Зейделя 99
  - — касательных 133
  - — конечных элементов 173
  - — линеаризации 133
  - — минимальных невязок 127
  - — Ньютона 133
  - — переменных направлений 251
  - — Пикара (последовательных приближений) 175
  - — попеременно-треугольный 120
  - — поправок 128
  - — прогонки 34
  - — встречной 37
- — левой 37
- — правой 37
- Метод простой итерации 98
  - — прямой 89
  - — прямых 234
  - — разделения переменных 222
  - — Ритца 172
  - — Ричардсона 115
  - — Рунге 82, 165, 178
  - — Рунге — Кутта 174
  - — секущих 136
  - — скорейшего спуска 128
  - — сопряженных градиентов 129
  - — стационарный итерационный 102
  - — сумматорных тождеств 171
  - — Штёрмера 189
  - — энергетических неравенств 144, 207
- Минимизирующий квадратичный функционал 171
- Наилучшее среднеквадратичное приближение 68
- Невязка для разностной схемы на решении 146
- Норма оператора 40
- Обратное интерполирование 67
- Однородная разностная схема 150
- Оператор единичный 41
  - — линейный 40
  - — неотрицательный 41
  - — обратный 40
  - — ограниченный 40
  - — положительный 41
  - — разрешающий 111
  - — самосопряженный 41
  - — сопряженный 41
  - — факторизованный 129, 252
  - — экономичный (экономичность оператора) 119
- Операторное уравнение первого рода 88
- Операторы перестановочные 41
- Ошибка округления 10
- Погрешность аппроксимации для краевого условия 146
  - — в точке,  $m$ -й порядок 139
  - — для уравнения 146
  - — на решении 147
  - — па сетке 140, 185
  - — оператора 139
  - — квадратурной формулы 70
  - — метода 10
- Погрешность неустранимая 10
- Полином обобщенный 68
  - — Чебышева 112, 114
- Принцип максимума 55
- Пространство евклидово (унитарное) 39
  - — нормированное 39

- сеточных функций 46
- энергетическое 45
- Процесс Эйткена 81
- Равенство Парсевала — Стеклова 69
- Равномерное приближение 69
- Разделенные разности 1-го порядка 64
  - 2-го порядка 64
- Размерность линейного пространства 39
- Разностная производная 139
  - левая 139
  - правая 139
  - центральная 139
- схема 141
  - Адамса 188
  - аддитивная 256
  - безусловно устойчивая (пример) 182
  - двухслойная 181, 197
  - Дугласа — Рекфорда 254
  - квазистойчивая 145
  - консервативная 152
  - корректная 142
  - Кранка — Николсона 230
  - крест 212
  - локально-одномерная 258
  - многошаговая 184
  - $m$ -го порядка точности 146
  - $m$ -шаговая ( $m \geq 1$ ) 185
  - неустойчивая 142
  - неявная 198
  - одношаговая 181
  - Писмена — Рекфорда 251
  - предиктор — корректор (счет — пересчет) 180
  - расщепления 258
  - р-устойчивая 201
- Рунге — Куттга 179
- с весами 198
- с опережением 198
- симметричная 198
- условно устойчивая схема (пример) 182
- устойчивая 142, 143
- Разностная схема чебышевская итерационная 112
  - чисто неявная 198
  - Эйлера 141, 176
  - экономичная 250
  - явная 198
- Разностное уравнение линейное с постоянными коэффициентами 26
  - $m$ -го порядка ( $m \geq 1$ ) 26
  - однородное 28
- Разностные неравенства 27
  - формулы Грина 50
  - Сетка квадратная 212
    - неравномерная 16
    - равномерная 16
  - Сеточная функция 16, 138
  - Сплайн порядка  $m$  66
    - Среднеквадратичное уклонение 68
    - Сходимость разностной схемы 146
      - с квадратичной скоростью 134
    - Уравнение теплопроводности 232
    - Устойчивость разностной схемы с весами 182
    - Формула Тейлора 74
      - Формулы бегущего счета 125
      - Численное интегрирование 70
      - Число обусловленности 89
      - Шаблон 139
        - квадратурной формулы 71

## ПРЕДИСЛОВИЕ

Эта книга представляет собой введение в теорию численных методов, использующее минимум сведений из анализа, линейной алгебры и теории дифференциальных уравнений. Книга возникла в результате обработки лекций, которые автор читал в течение нескольких лет для студентов второго курса факультета вычислительной математики ц кибернетики Московского государственного университета им. М. В. Ломоносова.

Содержание книги традиционное — интерполяция и аппроксимация, численное интегрирование, решение нелинейных уравнений, прямые и итерационные методы решения систем линейных алгебраических уравнений, разностные методы решения задач Коши и краевых задач для обыкновенных дифференциальных уравнений.

Автор стремился сделать изложение доступным для первого чтения, обращая внимание на основные понятия теории численных методов и иллюстрируя их простейшими примерами.

В настоящее время при численном решении многих задач физики и техники, описываемых уравнениями математической физики, используется метод конечных разностей. Основные понятия теории разностных методов (аппроксимация, устойчивость, сходимость) мы иллюстрируем на примерах разностных схем для обыкновенных дифференциальных уравнений. При аппроксимации дифференциальных уравнений получаются разностные уравнения, представляющие собой системы линейных уравнений высокого порядка с матрицами специального типа (имеющими много нулевых элементов), например, трехдиагональными. Важную роль играет выбор эффективных методов (прямых

и итерационных) решения таких систем. В связи с этим в книге излагаются основы общей теории итерационных методов. Большое внимание уделено вопросу устойчивости вычислений на электронных вычислительных машинах. В главе V дано простое изложение теории устойчивости задачи Коши для системы разностных уравнений первого порядка. Здесь получены совпадающие необходимые и достаточные условия устойчивости разностных схем, а также исследована асимптотическая устойчивость разностных схем.

В последних двух главах книги (главы VI и VII) рассматриваются разностные методы решения эллиптических уравнений и уравнения теплопроводности. Эти главы являются дополнительными и позволяют осуществить переход к теории разностных схем для уравнений с частными производными.

Колее полное изложение отдельных разделов численных методов можно найти в книгах: Самарский А. А. Теория разностных схем.— М.: Наука, 1977; Самарский Л. А., Николаев Е. С. Методы решения сеточных уравнений.— М.: Наука, 1978, а также в пособиях, список которых приведен в конце книги.

Книга рассчитана па студентов младших курсов, специализирующихся по прикладной математике и математической физике; она может оказаться полезной также для аспирантов и научных сотрудников, изучающих численные методы.

Автор пользуется возможностью выразить глубокую благодарность Л. В. Гулипу, прочитавшему рукопись и сделавшему ряд ценных замечаний, Е. С. Николаеву, оказавшему помошь при написании дополнения, а также М. И. Бакировой и Н. П. Савенковой за помошь в процессе работы над книгой и при подготовке ее к печати.

*А. Л. Самарский*

## ВВЕДЕНИЕ

Появление и непрерывное совершенствование быстрых действующих электронных вычислительных машин (ЭВМ) привело к подлинно революционному преобразованию науки вообще и математики в особенности. Изменилась технология научных исследований, колоссально увеличились возможности теоретического изучения, прогноза сложных процессов, проектирования инженерных конструкций. Решение крупных научно-технических проблем, примерами которых могут служить проблемы овладения ядерной энергией и освоения космоса, стало возможным лишь благодаря применению математического моделирования и новых численных методов, предназначенных для ЭВМ.

Первая крупная проблема — овладение ядерной энергией — требует решения комплекса сложных задач физики и механики (управление работой реактора, использование энергии деления ядер урана, защита от проникающего излучения, охлаждение степок реактора, изучение тепловых полей и упругих напряжений в стеках, решение многих других задач). Все эти задачи необходимо решать до начала работы реактора, используя для них математическое описание (модель) и проводя численные расчеты на ЭВМ. Вторая крупная проблема — освоение космоса — связана с созданием летательных аппаратов и решением для них многих задач аэродинамики и баллистики (например, расчет движения ракеты и управление ее полетом). Здесь также имеется комплекс сложных задач механики, физики и техники, которые могут быть решены только с использованием численных методов.

Укажем еще одну проблему, стоящую перед человечеством, — поиск новых источников энергии. Один из основных проектов получения энергии — использование реакции управляемого термоядерного синтеза ядер дейтерия и трития. Запасы термоядерного горючего на Земле

практически неисчерпаемы, а продукты реакции не загрязняют среду. Однако термоядерная реакция начинается только при экстремальных условиях — при высокой температуре (порядка десятка и сотни миллионов градусов) и огромном сжатии (в тысячи раз) дейтерия и триотия; кроме того, требуется удержать горючее вещество в этом состоянии в течение времени, достаточного для развития реакции горения (синтеза). Создание таких условий — пока еще нерешенная научно-техническая проблема. Существует несколько проектов нагрева, сжатия и удержания термоядерного горючего (плазмы). При их реализации возникает много вопросов, которые надо решать до начала проектирования даже экспериментальных установок. Необходимо прежде всего изучить поведение плазмы при высоких температурах и плотностях, в магнитных полях и выяснить условия, при которых возможна сама реакция термоядерного синтеза.

Такие исследования проводятся на основе математического описания (математической модели) физических процессов и последующего решения соответствующих математических задач на ЭВМ при помощи вычислительных алгоритмов.

В настоящее время можно говорить, что появился новый способ теоретического исследования сложных процессов, допускающих математическое описание, — вычислительный эксперимент, т. е. исследование естественно-научных проблем средствами вычислительной математики. Поясним существование этого способа исследования на примере решения какой-либо физической проблемы. Пусть требуется изучить некоторый физический процесс. Математическому исследованию предшествует выбор физического приближения, т. е. решение вопроса о том, какие факторы надо учесть, а какими можно пренебречь. После этого проводится исследование проблемы методом вычислительного эксперимента, в котором можно выделить несколько основных этапов.

На первом этапе проводится выбор математической модели, т. е. приближенное описание процесса в форме алгебраических, дифференциальных или интегральных уравнений. Эти уравнения обычно выражают законы сохранения основных физических величин (энергии, количества движения, массы и др.). Полученную математическую модель необходимо исследовать методами теории дифференциальных уравнений. Надо установить, правильно ли

поставлена задача, хватает ли исходных данных, не противоречат ли они друг другу, существует ли решение поставленной задачи и единственны ли оно. На этом этапе используются методы классической математики. Следует отметить, что многие физические задачи приводят к таким математическим моделям, разработка теории которых находится в начальной стадии. На практике приходится решать задачи математической физики, для которых не имеется теорем существования и единственности.

Второй этап вычислительного эксперимента состоит в построении приближенного численного метода решения задачи, т. е. в выборе вычислительного алгоритма. Под вычислительным алгоритмом понимают последовательность арифметических и логических операций, при помощи которых находится решение математической задачи, сформулированной на первом этапе. Ниже мы подробнее обсудим требования, предъявляемые к вычислительному алгоритму, предназначенному для использования на современных ЭВМ. По существу вся данная книга посвящена рассмотрению элементарных вычислительных алгоритмов.

На третьем этапе осуществляется программирование вычислительного алгоритма для ЭВМ и на четвертом этапе — проведение расчетов на ЭВМ. Мы не будем останавливаться на вопросах, связанных с программированием, организацией и проведением вычислений на ЭВМ, так как это выходит за рамки данной книги. Отметим лишь, что деятельность по программированию должна быть тесно связана с разработкой конкретных численных алгоритмов.

Наконец, в качестве пятого этапа вычислительного эксперимента можно выделить анализ полученных численных результатов и последующее уточнение математической модели. Может оказаться, что модель слишком груба — результат вычислений не согласуется с физическим экспериментом, или что модель слишком сложна, и решение с достаточной точностью можно получить при более простых моделях. Тогда следует начинать работу с первого этапа, т. е. уточнить математическую модель, и снова пройти все этапы.

Следует отметить, что вычислительный эксперимент — это, как правило, не разовый счет по стандартным формулам, а прежде всего расчет серии вариантов для различных математических моделей.

Остановимся теперь подробнее на некоторых общих характеристиках и требованиях, относящихся к вычислительным алгоритмам. Разработка и исследование вычислительных алгоритмов и их применение к решению конкретных задач составляет содержание огромного раздела современной математики — вычислительной математики.

Вычислительную математику определяют в широком смысле этого термина как раздел математики, включающий круг вопросов, связанных с использованием ЭВМ, и в узком смысле — как теорию численных методов и алгоритмов решения поставленных математических задач. В дальнейшем мы будем иметь в виду вычислительную математику лишь в узком смысле слова.

Общим для всех численных методов является сведение математической задачи к конечномерной. Это чаще всего достигается дискретизацией исходной задачи, т. е. переходом от функций непрерывного аргумента к функциям дискретного аргумента. После дискретизации исходной задачи надо построить вычислительный алгоритм, т. е. указать последовательность арифметических и логических действий, выполняемых на ЭВМ и дающих за конечное число действий решение дискретной задачи. Полученное решение дискретной задачи принимается за приближенное решение исходной математической задачи.

При решении задачи на ЭВМ мы всегда получаем не точное решение исходной задачи, а некоторое приближенное решение. Чем же обусловлена возникающая погрешность? Можно выделить три основные причины возникновения погрешности при численном решении исходной математической задачи. Прежде всего, входные данные исходной задачи (начальные и граничные условия, коэффициенты и правые части уравнений) всегда задаются с некоторой погрешностью. Погрешность численного метода, обусловленную неточным заданием входных данных, принято называть *неустранимой погрешностью*. Далее, при замене исходной задачи дискретной задачей возникает погрешность, называемая *погрешностью дискретизации* или, иначе, *погрешностью метода*. Например, заменив производную  $u'(x)$  разностным отношением  $(u(x + \Delta x) - u(x))/\Delta x$ , мы допускаем погрешность дискретизации, имеющую при  $\Delta x \rightarrow 0$  порядок  $\Delta x$ . Наконец, конечная разрядность чисел, представляемых в ЭВМ, приводит к *ошибкам округления*, которые могут нарастать в процессе вычислений. Естественно требовать, чтобы погрешности

в задании начальной информации и погрешность, возникающая в результате дискретизации, были согласованы с погрешностью решения на ЭВМ дискретной задачи.

Из сказанного видно, что основное требование, предъявляемое к вычислительному алгоритму,— это требование точности. Оно означает, что вычислительный алгоритм должен давать решение исходной задачи с заданной *точностью*  $\epsilon > 0$  за конечное число  $Q(\epsilon)$  действий. Алгоритм должен быть реализуемым, т. е. давать решение задачи за допустимое машинное время. Для большинства алгоритмов время решения задачи (объем вычислений)  $Q(\epsilon)$  возрастает при повышении точности, т. е. при уменьшении  $\epsilon$ . Конечно, можно задать  $\epsilon$  настолько малым, что время счета задачи станет недопустимо большим. Важно знать, что алгоритм дает принципиальную возможность получить решение задачи с любой точностью. Однако на практике величину  $\epsilon$  выбирают, учитывая возможность реализуемости алгоритма на данной ЭВМ. Для каждой задачи, алгоритма и машины есть свое характерное значение  $\epsilon$ .

Естественно добиваться, чтобы число действий (и тем самым машинное время решения задачи)  $Q(\epsilon)$  было минимальным для данной задачи. Для любой задачи можно предложить много алгоритмов, дающих одинаковую по порядку (при  $\epsilon \rightarrow 0$ ) точность  $\epsilon > 0$ , но за разное число действий  $Q(\epsilon)$ . Среди этих, как говорят, эквивалентных по порядку точности алгоритмов надо выбрать тот, который дает решение с затратой наименьшего машинного времени (числа действий  $Q(\epsilon)$ ). Такие алгоритмы будем называть *экономичными*.

Остановимся еще на одном требовании, предъявляемом к вычислительному алгоритму, а именно — требованиях отсутствия аварийного останова (авоста) ЭВМ в процессе вычислений.

Следует иметь в виду, что ЭВМ оперирует с числами, имеющими конечное число значащих цифр и принадлежащих (по модулю) не всей числовой оси, а некоторому интервалу  $(M_0, M_\infty)$ ,  $M_0 > 0$ ,  $M_\infty < \infty$ , где  $M_0$  — машинный нуль,  $M_\infty$  — машинная бесконечность. Если условие  $|M| < M_\infty$  в процессе вычислений нарушается, то происходит аварийный останов ЭВМ вследствие переполнения разрядной сетки, и вычисления прекращаются. Возможность авоста зависит как от алгоритма, так и от исходной задачи.

Если решение исходной задачи выражается через очень большие (очень малые) числа  $|M| > M_\infty$  ( $|M| < M_0$ ), то, как правило, путем изменения масштабов можно привести задачу к виду, содержащему только величины, принадлежащие (по модулю) заданному интервалу  $(M_0, M_\infty)$ . Часто возможность авоста может быть устранена путем изменения порядка действий. Поясним это на простом примере.

**Пример.** Пусть  $M_\infty = 10^p$ ,  $M_0 = 10^{-p}$ ,  $p = 2^n$ ,  $n$  — целое число. Требуется вычислить произведение чисел  $10^{p/2}$ ,  $10^{p/4}$ ,  $10^{-p/2}$ ,  $10^{3p/4}$ ,  $10^{-3p/4}$ .

**1-й способ.** Перенумеруем числа в порядке убывания:  $q_1 = 10^{3p/4}$ ,  $q_2 = 10^{p/2}$ ,  $q_3 = 10^{p/4}$ ,  $q_4 = 10^{-p/2}$ ,  $q_5 = 10^{-3p/4}$  и образуем произведения  $S_{k+1} = S_k q_{k+1}$ ,  $S_1 = q_1$ . Тогда уже на первом шаге мы получим авост, так как  $S_2 = q_1 q_2 = 10^{5p/4} > M_\infty$ .

**2-й способ.** Перенумеруем числа в порядке возрастания:  $q_1 = 10^{-3p/4}$ ,  $q_2 = 10^{-p/2}$ ,  $q_3 = 10^{p/4}$ ,  $q_4 = 10^{p/2}$ ,  $q_5 = 10^{3p/4}$ . В этом случае мы получим на первом шаге

$$S_2 = q_1 q_2 = 10^{-5p/4} < M_0,$$

т. е.  $S_2$  — машинный пуль; нулю равны и все последующие произведения  $S_3$ ,  $S_4$ ,  $S_5$ ; таким образом, здесь происходит полная потеря точности.

**3-й способ.** Перемешаем эти числа, полагая  $q_1 = 10^{-3p/4}$ ,  $q_2 = 10^{p/2}$ ,  $q_3 = 10^{3p/4}$ ,  $q_4 = 10^{-p/2}$ ,  $q_5 = 10^{p/4}$ . Тогда последовательно найдем:

$$\begin{aligned} S_2 &= q_1 q_2 = 10^{-p/4}, \quad S_3 = S_2 q_3 = 10^{p/2}, \\ S_4 &= S_3 q_4 = 10^0, \quad S_5 = S_4 q_5 = 10^{p/4}, \end{aligned}$$

т. е. в процессе вычислений не появляются числа, большие  $10^{p/2}$ , и меньшие  $10^{-p/4}$ . Такой алгоритм является безавостным. В гл. III мы встретимся с итерационным методом решения системы линейных алгебраических уравнений, который может быть авостным или безавостным в зависимости от способа нумерации итерационных параметров, определяющего последовательность вычислений.

При каждом акте вычислений появляются ошибки округления. В зависимости от алгоритма эти ошибки округления могут либо нарастать, либо затухать.

Если в процессе вычислений ошибки округления неограниченно нарастают, то такой алгоритм называют *не-*

*устойчивым* (вычислительно неустойчивым). Если же ошибки округления не накапливаются, то алгоритм является устойчивым.

Примеры. 1. Пусть требуется найти  $y_i$  ( $0 < i \leq i_0$ ) по формуле  $y_{i+1} = y_i + d$  ( $i \geq 0$ ) при заданных  $y_0$ ,  $d$ . Предположим, что при вычислении  $y_i$  внесена ошибка (например, ошибка округления), имеющая величину  $\delta_i$ , т. е. вместо точного значения  $y_i$  получено приближенное значение  $\tilde{y}_i = y_i + \delta_i$ . Тогда вместо точного значения  $y_{i+1}$  получим приближенное значение  $\tilde{y}_{i+1} = (y_i + \delta_i) + d = y_{i+1} + \delta_i$ . Таким образом, ошибка, допущенная на любом промежуточном шаге, не увеличивается в процессе вычислений. Алгоритм устойчив.

2. Рассмотрим уравнение  $y_{i+1} = qy_i$  ( $i \geq 0$ ,  $y_0$  и  $q$  заданы). Пусть, как и в примере 1, вместо  $y_i$  получено значение  $\tilde{y}_i = y_i + \delta_i$ . Тогда вместо  $y_{i+1}$  получим приближенное значение

$$\tilde{y}_{i+1} = q(y_i + \delta_i) = y_{i+1} + q\delta_i.$$

Отсюда видно, что погрешность  $\delta_{i+1} = \tilde{y}_{i+1} - y_{i+1}$ , возникающая при вычислении  $y_{i+1}$ , связана с погрешностью  $\delta_i$  уравнением

$$\delta_{i+1} = q\delta_i, \quad i = 0, 1, 2, \dots$$

Следовательно, если  $|q| > 1$ , то в процессе вычислений абсолютное значение погрешности будет возрастать (алгоритм неустойчив). Если же  $|q| \leq 1$ , то погрешность не возрастает, т. е. алгоритм устойчив. Неустойчивость обычно связывают со свойством экспоненциального нарастания ошибки округления. Если же ошибка округления нарастает по степенному закону при переходе от одной операции к другой («от шага к шагу»), то алгоритм считают *условно устойчивым* (устойчивым при некоторых ограничениях на объем вычислений и требуемую точность). Процесс вычислений можно трактовать так: при переходе от шага к шагу происходит искажение (за счет ошибок округления) последних значащих цифр (от последних значащих цифр справа налево движется «волна ошибки округления»). Нам нужно обычно сохранить верными несколько первых значащих цифр (4—5 знаков), и поэтому вычисления должны быть закончены до того, как до них дойдет «волна ошибки округления». Если ошибка округления  $\epsilon_0$  нарастает от шага к шагу по экспоненциальному закону, то это приводит, как правило, к аварии на про-

межуточном этапе вычислений, если (как в примере 2)  $|q|^i \varepsilon_0 \geq M_\infty$ .

Если  $M_\infty = 10^p$ ,  $\varepsilon_0 = 10^{-k_0}$ , то авост наступает при  $i_0 > (p + k_0)/\lg |q|$ . Иначе обстоит дело при степенном росте ошибки округления. Пусть  $|\delta y_i| \approx i^n \varepsilon_0$  ( $n \geq 1$ ); тогда авост наступит при  $i_0^n \varepsilon_0 \geq M_\infty$ , т. е. при  $i_0 \geq \left(\frac{1}{\varepsilon_0} M_\infty\right)^{1/n} = 10^{(p+k_0)/n}$ .

Отсюда видно, что при  $n = 1$  авоста не будет в силу очевидного ограничения  $i < M_\infty = 10^p$ . Неравенство  $|\delta y_i| \leq \varepsilon$ , где  $\varepsilon = 10^{-k}$  — заданная точность, справедливо при  $i \leq \left(\frac{\varepsilon}{\varepsilon_0}\right)^{1/n} = 10^{(k_0-k)/n} = i_0$ . Если заданы  $\varepsilon$

и  $\varepsilon_0$ , то это неравенство означает ограничение на число уравнений  $i \leq i_0$ . Так, при  $k_0 = 12$ ,  $k = 6$  имеем  $i \leq 10^{6/n}$ , так что  $i \leq 10^3$  при  $n = 2$ . Ясно, что можно указать такое большое  $n$ , что допустимое число уравнений  $i_0$  очень мало. Однако, на практике обычно встречаются случаи небольшого  $n$  (например, для метода прогонки (§ 3 гл. I)  $n = 2$ , т. е. погрешность накапливается по квадратичному закону с ростом числа уравнений).

При решении любой задачи необходимо знать какие-то входные (исходные) данные — начальные, граничные значения искомой функции, коэффициенты и правую часть уравнения и др.

Для каждой задачи ставятся одни и те же вопросы: существует ли решение задачи, является ли оно единственным и как зависит решение от входных данных? Возможны два случая:

Задача поставлена корректно (задача корректна); это значит, что 1) задача разрешима при любых допустимых входных данных; 2) имеется единственное решение; 3) решение задачи непрерывно зависит от входных данных (малому изменению входных данных соответствует малое изменение решения) — иными словами, задача устойчива.

Задача поставлена некорректно (задача некорректна), если ее решение неустойчиво относительно входных данных (малому изменению входных данных может соответствовать большое изменение решения).

Примером корректной задачи может служить задача интегрирования, а примером некорректной задачи — задача дифференцирования.

**Примеры.** 1. Задача интегрирования. Дана функция  $f(x)$ ; найти интеграл

$$J = \int_0^1 f(x) dx.$$

Заменим  $f$  на  $\tilde{f}$  и рассмотрим  $\tilde{J} = \int_0^1 \tilde{f}(x) dx$  и разность

$\delta J = \tilde{J} - J = \int_0^1 \delta f dx$  ( $\delta f = \tilde{f}(x) - f(x)$ ). Отсюда видно, что  $|\delta J| \leq \max_{0 \leq x \leq 1} |\delta f(x)|$ ,  $|\delta J| \leq \epsilon$ , если  $|\delta f| \leq \epsilon$ , т. е.  $J$  непрерывно зависит от  $f$ . Для вычисления интеграла  $J$  воспользуемся квадратурной формулой:

$$J_N = \sum_{k=1}^N c_k f(x_k), \quad c_k > 0, \quad \sum_{k=1}^N c_k = 1.$$

Повторяя рассуждения, приведенные выше, получим

$$\delta J_N = \tilde{J}_N - J_N = \sum_{k=1}^N c_k (\tilde{f}_k - f_k) = \sum_{k=1}^N c_k \delta f_k,$$

$$|\delta J_N| \leq \sum_{k=1}^N c_k \max_{1 \leq k \leq N} |\delta f_k| = \max_{1 \leq k \leq N} |\delta f_k|.$$

Таким образом, задача вычисления интеграла по квадратурной формуле корректна.

2. Задача дифференцирования. Задача дифференцирования функции  $u(x)$ , заданной приближенно, является некорректной. В самом деле, пусть  $\tilde{u}(x) = u(x) + \frac{1}{N} \sin N^2 x$ , где  $N$  достаточно велико. Тогда в метрике  $C$  (на некотором отрезке  $0 \leq x \leq \delta$  ( $\delta > \pi/N^2$ )) имеем  $\|\delta u\|_C = \|\tilde{u} - u\|_C = 1/N \leq \epsilon$  при  $N \geq 1/\epsilon$ . Для погрешности производных  $\delta u' = \tilde{u}' - u' = N \cos N^2 x$  имеем  $\|\delta u'\|_C = N \geq 1/\epsilon$ . Таким образом, малому изменению  $O(\epsilon)$  в  $C$  функции  $u(x)$  соответствует большое изменение  $O(1/\epsilon)$  в  $C$  ее производной.

Поэтому численное дифференцирование также некорректно. Чтобы найти приближенное значение производной по формуле разностной производной с некоторой точностью  $\epsilon > 0$  при условии, что функция задана с погрешностью  $\delta_i$  ( $|\delta_i| \leq \delta_u$ ), необходимо выполнение условий согласования  $\epsilon$ ,  $\delta_u$  и шага  $h$  сетки, например, вида  $\epsilon \geq$

$\geq k\sqrt{\delta_0}$  ( $k = \text{const} > 0$  не зависит от  $h$ ,  $\delta_0$ ), причем шаг сетки ограничен как снизу, так и сверху. Таким образом, достижимая точность численного дифференцирования лимитируется точностью задания самой функции.

В данной книге мы рассматриваем только корректные задачи и корректные численные методы, ориентированные на использование ЭВМ.

Численные методы дают приближенное решение задачи. Это значит, что вместо точного решения  $u$  (функции или функционала) некоторой задачи мы находим решение  $u$  другой задачи, близкое в некотором смысле (например, по норме) к исковому. Как уже указывалось, основная идея всех методов — дискретизация или аппроксимация (замена, приближение) исходной задачи другой задачей, более удобной для решения на ЭВМ, причем решение аппроксимирующей задачи зависит от некоторых параметров, управляя которыми, можно определить решение с требуемой точностью. Например, в задаче численного интегрирования такими параметрами являются узлы и веса квадратурной формулы. Далее, решение дискретной задачи является элементом конечномерного пространства. Остановимся на этом подробнее.

Рассмотрим, например, дискретизацию пространства  $H = \{f(x)\}$  функций  $f(x)$  непрерывного аргумента  $x \in [a, b]$ . На отрезке  $a \leq x \leq b$  введем конечное множество точек  $\omega = \{x_i, i = 0, 1, \dots, N, x_0 = a, x_N = b, x_i < x_{i+1}\}$ , которое назовем *сеткой*. Точки  $x_i$  будем называть *узлами* сетки  $\omega$ . Если расстояние  $h_i = x_i - x_{i-1}$  между соседними узлами постоянно (не зависит от  $i$ ),  $h_i = h$  для всех  $i = 1, 2, \dots, N$ , то сетку  $\omega$  называют *равномерной* (с *шагом*  $h$ ), в противном случае — *неравномерной*. Вместо функции  $f(x)$ , определенной для всех  $x \in [a, b]$ , будем рассматривать *сеточную функцию*  $y_i = f(x_i)$  целочисленного аргумента  $i$  ( $i = 0, 1, \dots, N$ ) или узла  $x_i$  сетки  $\omega$ , а  $H = \{f(x), x \in [a, b]\}$  заменим конечномерным (размерности  $N + 1$ ) пространством  $H_{N+1} = \{y_i, 0 \leq i \leq N\}$  сеточных функций. Очевидно, что сеточную функцию  $y_i = f(x_i)$  можно рассматривать как вектор  $y = (y_0, y_1, \dots, y_N)$ .

Можно провести также дискретизацию и пространства функций  $f(x)$  многих переменных, когда  $x = (x_1, x_2, \dots, x_p)$  — точка  $p$ -мерного евклидова пространства ( $p > 1$ ). Так, на плоскости  $(x_1, x_2)$  можно ввести сетку  $\omega = \{x_i = (i_1 h_1, i_2 h_2), i_1, i_2 = 0, \pm 1, \pm 2, \dots\}$  как множество точек

(узлов) пересечения перпендикулярных прямых  $x_1^{(i_1)} = i_1 h_1$ ,  $x_2^{(i_2)} = i_2 h_2$ ,  $h_1 > 0$ ,  $h_2 > 0$ ,  $i_1, i_2 = 0, \pm 1, \pm 2, \dots$ , где  $h_1$  и  $h_2$  — шаги сетки по направлениям  $x_1$  и  $x_2$  соответственно. Сетка  $\omega$ , очевидно, равномерна по каждому из переменных в отдельности. Вместо функции  $f(x) = f(x_1, x_2)$  будем рассматривать сеточную функцию

$$y_{i_1 i_2} = f(i_1 h_1, i_2 h_2).$$

Если сетка  $\omega$  содержит только те узлы, которые принадлежат прямоугольнику  $(0 \leq x_1 \leq l_1, 0 \leq x_2 \leq l_2)$ , так что  $h_1 = l_1/N_1$ ,  $h_2 = l_2/N_2$ , то сетка имеет конечное число  $N = (N_1 + 1)(N_2 + 1)$  узлов, а пространство  $H_N$  сеточных функций  $y_i = y_{i_1, i_2}$  является конечномерным.

Мы всюду рассматриваем только конечномерное пространство сеточных функций. Заменяя пространство  $H = \{f(x)\}$  функций непрерывного аргумента и исходную задачу пространством  $H_N$  сеточных функций и дискретной аппроксимацией исходной задачи, мы должны быть уверены, что будем лучше приближаться к решению исходной задачи при увеличении числа узлов. Оценка качества приближения и выбор способа аппроксимации — главная задача теории численных методов.

Основное содержание книги в той или иной степени связано с применением разностных методов для решения дифференциальных уравнений. Выделим два главных вопроса:

- получение дискретной (разностной) аппроксимации дифференциальных уравнений и исследование получающихся при этом разностных уравнений;
- решение разностных уравнений.

При получении дискретной аппроксимации (разностной схемы) важную роль играет общее требование, чтобы разностная схема как можно лучше приближала (моделировала) основные свойства исходного дифференциального уравнения. Такие разностные схемы можно получать, например, при помощи вариационных принципов и интегральных соотношений (см. гл. IV). Оценка точности разностной схемы сводится к изучению погрешности аппроксимации и устойчивости схемы. Изучение устойчивости — центральный вопрос теории численных методов и ему уделяется большое внимание в данной книге. Алгоритмы для сложных задач можно представить как последова-

тельность (цепочку) простых алгоритмов (модулей). Поэтому многие принципиальные вопросы теории численных методов можно выяснить на простых алгоритмах.

В главе I рассматриваются одномерные (зависящие от одного целочисленного аргумента) разностные уравнения. Мы ограничиваемся изучением разностных уравнений первого и второго порядков. Разностные уравнения второго порядка представляют собой систему линейных алгебраических уравнений с трехдиагональной матрицей. Для решения краевых задач для этих уравнений применяется так называемый метод прогонки. В I главе даны, в виде справочного материала, сведения о линейных операторах в конечномерном пространстве. В дальнейшем исследуются свойства разностных операторов как линейных операторов в конечномерном пространстве со скалярным произведением. При этом используется простейший математический аппарат — формулы разностного дифференцирования произведения и суммирования по частям.

Во второй главе излагается традиционный материал численного анализа: интерполяция, среднеквадратичная аппроксимация и численное интегрирование.

При аппроксимации дифференциальных уравнений на сетке получаются разностные уравнения, представляющие собой систему линейных алгебраических уравнений высокого порядка (равного числу узлов сетки) со специальной (разреженной, т. е. имеющей много нулевых элементов) матрицей. Простейший пример такой матрицы — трехдиагональная матрица — был указан выше.

В главе III излагаются численные методы решения систем линейных алгебраических уравнений

$$\sum_{j=1}^N a_{ij} u^j = f^i, \quad i = 1, 2, \dots, N, \quad (1)$$

которые можно записать в матричной форме

$$Au = f, \quad (2)$$

где  $A = (a_{ij})$  — квадратная матрица размера  $N \times N$ ,  $u = (u^1, u^2, \dots, u^N)$  — искомый вектор,  $f = (f^1, f^2, \dots, f^N)$  — заданный вектор (правая часть).

Для решения систем уравнений применяются прямые и итерационные методы.

В § 2 гл. III рассматриваются метод исключения Гаусса и метод квадратного корня — прямые методы, требующие для решения системы  $O(N^3)$  арифметических действий.

При изучении итерационных методов систему линейных алгебраических уравнений (2) удобно трактовать как операторное уравнение первого рода с оператором, действующим в  $N$ -мерном пространстве  $H_N$  ( $A : H_N \rightarrow H_N$ ),  $u, f \in H_N$ . Чтобы подчеркнуть эквивалентность матричной и операторной форм записи, будем матрицу и соответствующий оператор обозначать одной и той же буквой  $A$ .

При изложении теории итерационных методов (одношаговых или двухслойных) важную роль играет каноническая форма итерационной схемы

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots \text{ для всех } y_0 \in H_N, \quad (3)$$

где  $A, B : H_N \rightarrow H_N$ ,  $\{\tau_k\}$  — итерационные параметры.

Всюду предполагается, что оператор  $A$  самосопряжен и положительно определен ( $A = A^* > 0$ ). Доказана общая теорема о сходимости стационарного метода с  $\tau_k = \tau = \text{const}$ . Достаточным условием сходимости является неравенство

$$(By, y) > \frac{\tau}{2} (Ay, y) \quad \text{для всех } y \in H, \quad (4)$$

где  $B \neq B^*$  — вообще говоря, несамосопряженный оператор. Отсюда следует сходимость метода простой итерации, метода Зейделя, метода верхней релаксации.

Если известны такие постоянные  $\gamma_1 > 0$ ,  $\gamma_2 > \gamma_1$  что

$$\gamma_1(Bx, x) \leq (Ax, x) \leq \gamma_2(Bx, x) \quad \text{для всех } x \in H_N, \quad (5)$$

где  $B = B^* > 0$ , то можно найти оптимальный чебышевский набор параметров  $\{\tau_k^*\}$ , при которых вычислительный процесс устойчив и безавостен.

Рассматривается универсальный попеременно-трехугольный метод с набором  $\{\tau_k^\alpha\}$  и оператором

$$B = (D + \omega A_1) D^{-1} (D + \omega A_2), \quad (6)$$

где  $D = D^* > 0$ ,  $A_1^* = A_2$ ,  $A_1 + A_2 = A$ , матрицы  $A_1$  и

$A_2$  — треугольные. Получена формула для параметра  $\omega$ . Алгоритм для этого метода очень простой. Всюду приводятся формулы для числа итераций, при которых достигается требуемая точность. Сравнение различных методов проведено на модельной задаче для разностного уравнения второго порядка  $y_{i-1} - 2y_i + y_{i+1} = -h^2 f_i$ ,  $i = 1, 2, \dots, N-1$ ,  $y_0 = y_N = 0$ ,  $h = 1/N$ , соответствующего краевой задаче  $u''(x) = -f(x)$  ( $0 < x < 1$ ),  $u(0) = u(1) = 0$ . Это уравнение есть одномерный аналог уравнения Лапласа. Так как число итераций практически не зависит от числа измерений, то при сравнении можно ограничиться этой одномерной задачей. Попеременно треугольный метод требует  $O\left(\frac{1}{\sqrt{h}} \ln \frac{1}{\varepsilon}\right)$  итераций, где  $\varepsilon > 0$  — заданная точность.

Заметим, что в главе III с помощью простейших математических средств фактически изложена достаточно полная общая теория итерационных методов решения уравнения  $Au = f$  ( $A = A^* > 0$ ).

Основные понятия теории разностных схем: погрешность аппроксимации, устойчивость, сходимость и точность излагаются на примерах краевых задач и задачи Коши для обыкновенных дифференциальных уравнений (гл. IV и гл. V). В главе IV изучаются трехточечные разностные схемы для обыкновенного дифференциального уравнения второго порядка

$$\begin{aligned} \frac{d}{dx} \left( k(x) \frac{du}{dx} \right) - q(x) u &= -f(x), \quad 0 < x < 1, \\ u(0) = u_1, \quad u(1) = u_2, \quad k(x) > 0, \quad q(x) &\geqslant 0. \end{aligned} \quad (7)$$

Исследованы вопросы о скорости сходимости (о порядке точности) однородных разностных схем на неравномерных сетках и для случая разрывных коэффициентов. Это потребовало получения весьма тонких априорных оценок, выражающих устойчивость разностной схемы по правой части.

Для получения разностных схем могут быть использованы различные методы — интегроинтерполяционный метод, метод аппроксимации квадратичного функционала, методы Ритца и Галеркина (§ 5, гл. IV).

Для решения задачи Коши для уравнения первого порядка

$$\frac{du}{dt} = f(t, u), \quad t > 0, \quad u(0) = u_0 \quad (8)$$

применяются методы Рунге — Кутта и Адамса, изложенные в главе V. Эти методы применимы и для системы уравнений, когда  $f$ ,  $u$  — векторы.

Особое место в главе V занимает задача Коши для системы линейных уравнений

$$\frac{du}{dt} + Au = f(t), \quad t > 0, \quad u(0) = u_0, \quad (9)$$

где  $A = (a_{ij})$  — квадратная матрица  $N \times N$ ,  $u(t) = (u^1, u^2, \dots, u^N)$ ,  $f(t) = (f^1, f^2, \dots, f^N)$  — вектор-функции размерности  $N$ .

Такая задача, в частности, возникает, если в уравнении теплопроводности

$$\frac{\partial u}{\partial t} = \Delta u + f(x, t), \quad \Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}, \quad x = (x_1, x_2) \quad (10)$$

заменить оператор Лапласа  $\Delta u$  соответствующим разностным оператором. Тогда (9) можно трактовать как метод прямых для уравнения теплопроводности (10). Используя для решения этой задачи какую-либо одношаговую схему, мы приходим к двухслойной операторно-разностной схеме общего вида, которая записывается в канонической форме

$$B \frac{y_{k+1} - y_k}{\tau} + Ay_k = \varphi_k, \quad k = 0, 1, \dots, \text{для всех } y_0 \in H_N, \quad (11)$$

где  $A$ ,  $B: H_N \rightarrow H_N$  — линейные операторы,  $\tau$  — шаг сетки по  $t$ .

Доказано, что необходимое и достаточное условие устойчивости схемы имеет вид

$$B \geq \frac{\tau}{2} A \text{ или } (Bx, x) \geq \frac{\tau}{2} (Ax, x) \text{ для любых } x \in H_N. \quad (12)$$

Это — основная теорема общей теории устойчивости операторно-разностных схем (ср. А. А. Самарский «Теория разностных схем»), пригодная для исследования устойчивости разностных схем для уравнений с частными производными математической физики (см. гл. VII). Фактически, в § 4 изложены основы общей теории устойчивости разностных схем, включая и асимптотическую устойчивость.

Сведения, полученные в главах III—V, позволяют без труда перейти к изучению теории разностных методов решения уравнений в частных производных. В главе VI такое изучение проведено для разностных схем, аппроксимирующих уравнение Пуассона и эллиптические уравнения в прямоугольнике с краевыми условиями первого рода. Здесь рассмотрены как вопросы сходимости, так и методы решения разностных уравнений.

Наличие общей теории устойчивости двухслойных разностных схем (гл. V) упрощает изложение разностных методов для уравнения теплопроводности с постоянными и переменными коэффициентами, проведенное в главе VII. Здесь рассматриваются также экономичные схемы (переменных направлений, расщепления и т. д.) для многомерных задач, а также общий принцип суммарной аппроксимации, который позволяет проводить расщепление сложных задач на последовательность более простых и существенно упрощать решение многомерных задач математической физики.

Следует отметить, что основное содержание книги излагается с единой точки зрения. Единство достигается за счет трактовки разностных схем как операторных или операторно-разностных уравнений с операторами, действующими в конечномерном пространстве со скалярным произведением. При построении теории итерационных методов и теории устойчивости разностных схем используются простейшие свойства операторов (матриц): знакопредeterminedость, самосопряженность, некоторые свойства собственных значений и собственных векторов; никаких предположений о структуре операторов при этом не делается. Все условия теории оказались очень удобными для проверки в случае конкретных разностных схем. Материал глав VI и VII может служить основой для изучения более полной теории по книгам [6, 9].

# Глава I

## РАЗНОСТНЫЕ УРАВНЕНИЯ

В этой главе изучаются сеточные функции целочисленного аргумента и разностные уравнения второго порядка. Излагается простейший математический аппарат для изучения сеточных функций и разностных операторов. Для решения разностных уравнений второго порядка применяется метод исключения, называемый методом прогонки.

### § 1. Сеточные функции

**1. Сеточные функции и действия над ними.** Как уже упоминалось, в приближенных методах обычно функции непрерывного аргумента заменяются функциями дискретного аргумента — сеточными функциями. *Сеточную функцию* можно рассматривать как функцию целочисленного аргумента:

$$y(i) = y_i, \quad i = 0, \pm 1, \pm 2, \dots$$

Для  $y(i)$  можно ввести операции, являющиеся дискретным (разностным) аналогом операций дифференцирования и интегрирования.

Аналогом первой производной являются разности *первого порядка*:

$\Delta y_i = y_{i+1} - y_i$  — правая разность;

$\nabla y_i = y_i - y_{i-1}$  — левая разность;

$$\delta y_i = \frac{1}{2} (\Delta y_i + \nabla y_i) = \frac{1}{2} (y_{i+1} - y_{i-1}) =$$

центральная разность;

при этом легко заметить, что  $\Delta y_i = \nabla y_{i+1}$ .

Далее можно написать разности *второго порядка*:

$$\Delta^2 y_i = \Delta(\Delta y_i) = \Delta(y_{i+1} - y_i) = y_{i+2} - 2y_{i+1} + y_i,$$

$$\begin{aligned} \Delta \nabla y_i &= \Delta(y_i - y_{i-1}) = (y_{i+1} - y_i) - (y_i - y_{i-1}) = \\ &= y_{i+1} - 2y_i + y_{i-1}, \end{aligned}$$

так что

$$\Delta^2 y_i = \Delta \nabla y_{i+1}.$$

Аналогично определяется разность *m*-го порядка:

$$\Delta^m y_i = \Delta(\Delta^{m-1} y_i),$$

содержащая значения  $y_i, y_{i+1}, \dots, y_{i+m}$ . Очевидно, что

$$\sum_{j=k}^i \Delta y_j = y_{i+1} - y_k, \quad \sum_{j=k}^i \nabla y_j = y_i - y_{k-1}.$$

**2. Разностные аналоги формул дифференцирования произведения и интегрирования по частям.** Пусть  $y_i, v_i$  — произвольные функции целочисленного аргумента. Тогда справедливы формулы

$$\Delta(y_i v_i) = y_i \Delta v_i + v_{i+1} \Delta y_i = y_{i+1} \Delta v_i + v_i \Delta y_i, \quad (1)$$

$$\nabla(y_i v_i) = y_{i-1} \nabla v_i + v_i \nabla y_i = y_i \nabla v_i + v_{i-1} \nabla y_i, \quad (2)$$

которые проверяются непосредственно. Например,

$$\begin{aligned} \Delta(y_i v_i) &= y_{i+1} v_{i+1} - y_i v_i; \\ y_i \Delta v_i + v_{i+1} \Delta y_i &= y_i(v_{i+1} - v_i) + v_{i+1}(y_{i+1} - y_i) = \\ &= y_{i+1} v_{i+1} - y_i v_i = \Delta(y_i v_i). \end{aligned}$$

При выводе формулы для  $\nabla(y_i v_i)$  достаточно учесть, что  $\nabla(y_i v_i) = \Delta(y_{i-1} v_{i-1})$ .

Формулы (1), (2) представляют собой аналоги формулы дифференцирования произведения  $(y(x)v(x))' = yv' + vy'$ .

Аналогом формулы интегрирования по частям является формула суммирования по частям:

$$\sum_{i=0}^{N-1} y_i \Delta v_i = - \sum_{i=1}^N v_i \nabla y_i + (yv)_N - (yv)_0, \quad (3)$$

которую записывают также в виде

$$\sum_{i=1}^{N-1} y_i \Delta v_i = - \sum_{i=1}^{N-1} v_i \nabla y_i + y_{N-1} v_N - y_0 v_1. \quad (4)$$

Для вывода формулы (3) воспользуемся формулой (1); имеем

$$y_i \Delta v_i = \Delta(y_i v_i) - v_{i+1} \Delta y_i = \Delta(y_i v_i) - v_{i+1} \nabla y_{i+1},$$

поскольку  $\Delta y_i = \nabla y_{i+1}$ ; отсюда получаем

$$\begin{aligned} \sum_{i=0}^{N-1} y_i \Delta v_i + \sum_{i=1}^N v_i \nabla y_i &= \\ &= \sum_{i=0}^{N-1} \Delta(y_i v_i) - \sum_{i=0}^{N-1} v_{i+1} \nabla y_{i+1} + \sum_{i=1}^N v_i \nabla y_i = \\ &= y_N v_N - y_0 v_0 - \sum_{i=1}^N v_i \nabla y_i + \sum_{i=1}^N v_i \nabla y_i = (yv)_N - (yv)_0. \end{aligned}$$

Если  $y_0 = 0, y_N = 0$ , то  $\sum_{i=0}^{N-1} y_i \Delta v_i = - \sum_{i=1}^N v_i \nabla y_i$ .

Формулу суммирования по частям можно использовать для вычисления сумм.

Примеры. 1. Вычислить сумму  $S_N = \sum_{i=1}^N i 2^i$ . Положим  $v_i = i, \nabla y_i = 2^i$ , так что

$$y_i = y_{i-1} + 2^i = y_0 + \sum_{j=1}^i 2^j = y_0 + 2^{i+1} - 2.$$

Выберем  $y_0 = 2 - 2^{N+1}$ ; тогда  $y_N = 0$ . Так как  $v_0 = 0, \Delta v_i = 1$ , то из (3) следует

$$\begin{aligned} S_N &= \sum_{i=1}^N v_i \nabla y_i = - \sum_{i=0}^{N-1} y_i \Delta v_i = - \sum_{i=0}^{N-1} y_i = \\ &= -N(y_0 - 2) - \sum_{i=0}^{N-1} 2^{i+1} = N2^{N+1} - (2^{N+1} - 2), \end{aligned}$$

так что  $S_N = (N-1)2^{N+1} + 2$ .

2. Вычислить  $S_N = \sum_{i=1}^N i(i-1) = \sum_{i=1}^{N-1} i(i+1)$ . Положим  $y_i = i, \nabla v_i = i+1$ . Тогда  $v_{i+1} = v_i + (i+1) = v_1 + (2+3+\dots+(i+1)) = (v_1-1) + (i+1)(i+2)/2, v_i = v_1 - 1 + i \times (i+1)/2$ . Выберем  $v_1$  из условия  $v_N = 0$ , т. е.  $v_1 = 1 - N(N+1)/2$ . Применяя формулу (3) и учитывая, что  $y_0 = 0, v_N = 0, \nabla y_i = 1$ , находим

$$\begin{aligned} S_N &= \sum_{i=1}^{N-1} i(i+1) = \sum_{i=0}^{N-1} y_i \Delta v_i = - \sum_{i=1}^N v_i \nabla y_i = - \sum_{i=1}^{N-1} v_i = \\ &= -(N-1)(v_1-1) - \frac{1}{2} \sum_{i=1}^{N-1} i(i+1) = \\ &= -\frac{1}{2} S_N + \frac{(N-1)N(N+1)}{2}. \end{aligned}$$

так что  $S_N = \frac{1}{3} (N - 1) N (N + 1)$ . Отсюда следует, что

$$\sum_{i=1}^N i^2 = 1^2 + 2^2 + \dots + N^2 = S_N + \sum_{i=1}^N i = \frac{N(N+1)(2N+1)}{6}.$$

## § 2. Разностные уравнения

**1. Разностные уравнения.** Линейное уравнение относительно сеточной функции  $y_i = y(i)$  ( $i = 0, \pm 1, \pm 2, \dots$ )

$$a_0(i)y(i) + a_1(i)y(i+1) + \dots + a_m(i)y(i+m) = f(i), \quad (1)$$

где  $a_k(i)$  ( $k = 0, 1, \dots, m$ ),  $f(i)$  — заданные сеточные функции,  $a_0(i) \neq 0$ ,  $a_m(i) \neq 0$ , называется *линейным разностным уравнением*  $m$ -го порядка. Оно содержит  $m + 1$  значений функции  $y(i)$ .

Пользуясь формулами для разностей  $\Delta y_i$ ,  $\Delta^2 y_i$ , ...,  $\Delta^{m-1} y_i$ , можно выразить значения  $y_{i+1}$ ,  $y_{i+2}$ , ...,  $y_{m+1}$  через  $y_i$  и указанные разности:  $y_{i+1} = y_i + \Delta y_i$ ,  $y_{i+2} = \Delta^2 y_i + 2y_{i+1} - y_i = \Delta^2 y_i + 2\Delta y_i + y_i$  и т. д. В результате из (1) получим новую запись *разностного уравнения*  $m$ -го порядка:

$$\alpha_0(i)y_i + \alpha_1(i)\Delta y_i + \dots + \alpha_m(i)\Delta^m y_i = f(i), \quad i = 0, \pm 1, \pm 2, \dots \quad (2)$$

(чем и объясняется термин «разностное уравнение»). Если коэффициенты  $a_0$ ,  $a_1$ , ...,  $a_m$  не зависят от  $i$ ,  $a_0 \neq 0$  и  $a_m \neq 0$ , то (1) называется линейным разностным уравнением  $m$ -го порядка с *постоянными коэффициентами*.

При  $m = 1$  из (1) получаем разностное уравнение *первого порядка*

$$a_0(i)y_i + a_1(i)y_{i+1} = f(i), \quad a_0(i) \neq 0, \quad a_1(i) \neq 0, \quad (3)$$

при  $m = 2$  — разностное уравнение *второго порядка*

$$a_0(i)y_i + a_1(i)y_{i+1} + a_2(i)y_{i+2} = f(i), \quad a_0(i) \neq 0, \quad a_2(i) \neq 0.$$

Мы ограничимся изучением разностных уравнений первого и второго порядков.

**2. Уравнения первого порядка.** Рассмотрим разностное уравнение первого порядка (3). Подставляя  $y_{i+1} = y_i + \Delta y_i$ , получим

$$\bar{a}_0(i)y_i + a_1(i)\Delta y_i = f(i), \quad \bar{a}_0 = a_0 + a_1.$$

Простейшими примерами разностных уравнений первого порядка могут служить уравнения для членов арифметической прогрессии  $y_{i+1} = y_i + d$  и геометрической прогрессии  $y_{i+1} = qy_i$ .

Запишем уравнение (3) в виде

$$y_{i+1} = q_i y_i + \varphi_i, \quad (4)$$

где  $q_i = -a_0(i)/a_1(i)$ ,  $\varphi_i = f(i)/a_1(i)$ . Отсюда видно, что решение  $y(i)$  определено однозначно при  $i > i_0$ , если задано значение  $y(i_0)$ . Пусть при  $i = 0$  задано  $y_0 = y(0)$ . Тогда можно определить  $y_1, y_2, \dots, y_i, \dots$ . Последовательно исключая  $y_i, y_{i-1}, \dots, y_1$  по формуле (4), получим

$$\begin{aligned} y_{i+1} = q_i q_{i-1} \dots q_0 y_0 + \varphi_i + q_i \varphi_{i-1} + q_i q_{i-1} \varphi_{i-2} + \dots \\ \dots + q_i q_{i-1} \dots q_1 \varphi_0, \end{aligned}$$

или

$$y_{i+1} = \left( \prod_{k=0}^i q_k \right) y_0 + \sum_{k=0}^{i-1} \left( \prod_{s=k+1}^i q_s \right) \varphi_k + \varphi_i. \quad (5)$$

Для уравнения с постоянным коэффициентом  $q_i = q$  отсюда получаем

$$y_{i+1} = q^{i+1} y_0 + \sum_{k=0}^i q^{i-k} \varphi_k, \quad i = 0, 1, 2, \dots, \quad (6)$$

т. е. решение разностного уравнения (4) с постоянными коэффициентами.

**3. Неравенства первого порядка.** Если в выражениях типа (1) или (2) знак равенства заменить знаками неравенства  $<, >, \leqslant, \geqslant$ , то получим *разностные неравенства*  $m$ -го порядка. Пусть дано разностное неравенство первого порядка

$$y_{i+1} \leqslant q y_i + f_i, \quad i = 0, 1, 2, \dots, \quad q \geqslant 0; \quad (7)$$

не ограничивая общности, далее всегда считаем  $q > 0$  ( $y_0, q, f_i$  известны). Найдем его решение. Пусть  $v_i$  — решение разностного уравнения

$$v_{i+1} = q v_i + f_i, \quad i = 0, 1, \dots, \quad v_0 = y_0. \quad (8)$$

Тогда справедлива оценка

$$y_i \leqslant v_i. \quad (9)$$

В самом деле, вычитая (8) из (7), находим

$$\begin{aligned} y_{i+1} - v_{i+1} \leqslant q(y_i - v_i) \leqslant q^2(y_{i-1} - v_{i-1}) \leqslant \dots \\ \dots \leqslant q^{i+1}(y_0 - v_0) = 0. \end{aligned}$$

Подставив в (9) явное выражение для  $v_i$ , получим

$$y_i \leq q^i y_0 + \sum_{k=0}^{i-1} q^{i-1-k} f_k, \quad i = 0, 1, 2, \dots, \quad (10)$$

— решение неравенства (7).

**4. Уравнение второго порядка с постоянными коэффициентами.** Рассмотрим разностное уравнение второго порядка

$b y_{i+1} - c y_i + a y_{i-1} = f_i, \quad i = 0, 1, \dots, \quad a \neq 0, \quad b \neq 0,$  (11)  
коэффициенты которого не зависят от  $i$ . Если  $f_i = 0$ , то уравнение

$$b y_{i+1} - c y_i + a y_{i-1} = 0, \quad i = 0, 1, \dots, \quad (12)$$

называется *однородным*. Его решение может быть найдено в явном виде.

Пусть  $\bar{y}_i$  — решение однородного уравнения (12),  $y_i^*$  — какое-либо решение неоднородного уравнения (11). Тогда их сумма  $y_i = \bar{y}_i + y_i^*$  также является решением неоднородного уравнения:

$$\begin{aligned} b(\bar{y}_{i+1} + y_i^*) - c(\bar{y}_i + y_i^*) + a(\bar{y}_{i-1} + y_{i-1}^*) &= \\ &= [b\bar{y}_{i+1} - c\bar{y}_i + a\bar{y}_{i-1}] + [by_{i+1}^* - cy_i^* + ay_{i-1}^*] = f_i. \end{aligned}$$

Это свойство — следствие линейности уравнения (11); оно сохраняет силу для разностного уравнения (1) любого порядка. Очевидно, что если  $\bar{y}_i$  является решением однородного уравнения (12), то и  $c\bar{y}_i$ , где  $c$  — произвольная постоянная, также удовлетворяет этому уравнению.

Пусть  $y_i^{(1)}$  и  $y_i^{(2)}$  — два решения уравнения (12). Они называются *линейно независимыми*, если равенство

$$c_1 y_i^{(1)} + c_2 y_i^{(2)} = 0, \quad i = 0, 1, 2, \dots,$$

возможно только при  $c_1 = c_2 = 0$ . Это эквивалентно требованию, что определитель системы

$$\begin{aligned} c_1 y_i^{(1)} + c_2 y_i^{(2)} &= 0, \\ c_1 y_{i+m}^{(1)} + c_2 y_{i+m}^{(2)} &= 0, \quad m = \pm 1, \pm 2, \dots, \end{aligned}$$

отличен от нуля для всех  $i, m$ . В частности,

$$\Delta_{i,i+1} = \begin{vmatrix} y_i^{(1)} & y_i^{(2)} \\ y_{i+1}^{(1)} & y_{i+1}^{(2)} \end{vmatrix} \neq 0.$$

Так же, как и в теории дифференциальных уравнений, можно ввести понятие *общего решения* разностного уравнения (12) и показать, что если решения  $y_i^{(1)}, y_i^{(2)}$  линейно независимы, то общее решение уравнения (12) имеет вид

$$y_i = c_1 y_i^{(1)} + c_2 y_i^{(2)},$$

где  $c_1$  и  $c_2$  — произвольные постоянные. Общее решение неоднородного уравнения (11) можно представить в виде

$$y_i = c_1 y_i^{(1)} + c_2 y_i^{(2)} + y_i^*, \quad (13)$$

где  $y_i^*$  — какое-либо (частное) решение уравнения (11). Для определения  $c_1$  и  $c_2$ , как и в случае дифференциальных уравнений, надо задать дополнительные условия — начальные или краевые.

Частное решение уравнения (12) можно найти в явном виде. Будем искать его в виде  $y_i = q^i$ , где  $q \neq 0$  — неизвестное пока число. После подстановки  $y_k = q^k$  в (12) получим квадратное уравнение  $bq^2 - cq + a = 0$ , имеющее корни

$$q_1 = \frac{c + \sqrt{c^2 - 4ab}}{2b}, \quad q_2 = \frac{c - \sqrt{c^2 - 4ab}}{2b}. \quad (14)$$

В зависимости от значений дискриминанта  $D = c^2 - 4ab$  возможны три случая:

1)  $D = c^2 - 4ab > 0$ . Корни  $q_1$  и  $q_2$  действительны и различны. Им соответствуют частные решения

$$y_k^{(1)} = q_1^k, \quad y_k^{(2)} = q_2^k;$$

эти решения линейно независимы, так как отличен от нуля определитель:

$$\Delta_{k,k+1} = \begin{vmatrix} q_1^k & q_1^{k+1} \\ q_2^k & q_2^{k+1} \end{vmatrix} = q_1^k q_2^k (q_2 - q_1) \neq 0.$$

Заметим, что  $q_1 \neq 0$  и  $q_2 \neq 0$ , иначе  $a = 0$  и уравнение (12) не является разностным уравнением второго порядка. Общее решение уравнения (12) имеет вид

$$y_k = c_1 q_1^k + c_2 q_2^k. \quad (15)$$

2)  $D = c^2 - 4ab < 0$ . Квадратное уравнение имеет комплексно-сопряженные корни

$$q_1 = \frac{c + i\sqrt{|D|}}{2b}; \quad q_2 = \frac{c - i\sqrt{|D|}}{2b},$$

где  $i$  — мнимая единица. Эти корни удобно представить в виде

$$q_1 = \rho e^{i\varphi}, \quad q_2 = \rho e^{-i\varphi}, \quad \rho = \sqrt{\frac{a}{b}}, \quad \varphi = \arctg \frac{\sqrt{|D|}}{c}.$$

Частными решениями являются не только функции

$$\begin{aligned} q_1^k &= \rho^k e^{ik\varphi} = \rho^k (\cos k\varphi + i \sin k\varphi), \\ q_2^k &= \rho^k e^{-ik\varphi} = \rho^k (\cos k\varphi - i \sin k\varphi), \end{aligned}$$

но и функции

$$y_k^{(1)} = \rho^k \cos k\varphi, \quad y_k^{(2)} = \rho^k \sin k\varphi,$$

которые линейно независимы в силу линейной независимости функций  $\sin k\varphi$  и  $\cos k\varphi$ . Общее решение имеет вид

$$y_k = \rho^k (c_1 \cos k\varphi + c_2 \sin k\varphi). \quad (16)$$

3)  $D = c^2 - 4ab = 0$ . Корни действительны и равны:  $q_1 = q_2 = c/(2b) = q_0$ . Линейно независимыми являются решения

$$y_k^{(1)} = q_0^k, \quad y_k^{(2)} = kq_0^k. \quad (17)$$

Покажем, что  $y_k^{(2)}$  есть решение уравнения (12):

$$\begin{aligned} by_{k+1}^{(2)} - cy_k^{(2)} + ay_{k-1}^{(2)} &= b(k+1)q_0^{k+1} - ckq_0^k + a(k-1)q_0^{k-1} = \\ &= k(bq_0^{k+1} - cq_0^k + aq_0^{k-1}) + (bq_0^2 - a)q_0^{k-1} = 0, \end{aligned}$$

так как  $bq_0^2 - a = b \frac{c^2}{4b^2} - a = \frac{D}{4b} = 0$ . Поскольку

$$\Delta_{k,k+1} = \begin{vmatrix} q_0^k & kq_0^k \\ q_0^{k+1} & (k+1)q_0^{k+1} \end{vmatrix} = q_0^{2k+1} \neq 0, \text{ то решения (17)}$$

линейно независимы, и общее решение имеет вид

$$y_k = c_1 q_0^k + c_2 k q_0^k.$$

**5. Примеры.** Рассмотрим примеры решения разностных уравнений второго порядка (11).

1. Найти общее решение уравнения

$$y_{k+1} - 2py_k + y_{k-1} = 0, \quad a = b = 1, \quad c = 2p > 0.$$

Возможны три случая. 1)  $p < 1$ . Положим  $p = \cos \alpha$ ; тогда  $D = 4(\cos^2 \alpha - 1) = -4 \sin^2 \alpha < 0$ . Частные решения имеют вид

$$y_k^{(1)} = \cos k\alpha, \quad y_k^{(2)} = \sin k\alpha.$$

2)  $p > 1$ . Полагая  $p = \operatorname{ch} \alpha$ , получим для  $q$  квадратное уравнение  $q^2 - 2 \operatorname{ch} \alpha q + 1 = 0$ ; его дискриминант равен  $D = 4(\operatorname{ch}^2 \alpha - 1) = 4 \operatorname{sh}^2 \alpha$ , а корни имеют вид  $q_{1,2} = \operatorname{ch} \alpha \pm \operatorname{sh} \alpha = e^{\pm \alpha}$ . Частными решениями являются функции

$$y_k^{(1)} = \operatorname{ch} k\alpha, \quad y_k^{(2)} = \operatorname{sh} k\alpha.$$

3)  $p = 1$ . В этом случае  $q^2 - 2q + 1 = 0$ ,  $q_{1,2} = 1$ , частные решения имеют вид  $y_k^{(1)} = 1$ ,  $y_k^{(2)} = k$ , а общее решение имеет вид

$$y_k = c_1 + c_2 k.$$

2. Найти решение уравнения

$$y_{k+2} - y_{k+1} - 2y_k = 0.$$

Дискриминант равен  $D = 1 + 8 = 9$ , корнями будут  $q_{1,2} = (1 \pm 3)/2$ ,  $q_1 = 2$ ,  $q_2 = -1$ . Общее решение имеет вид

$$y_k = c_1 2^k + c_2 (-1)^k.$$

3. Найти общее решение уравнения

$$y_{k+1} - y_k - 6y_{k-1} = 2^{k+1}. \quad (18)$$

Общее решение неоднородного уравнения есть сумма  $y_k = \bar{y}_k + y_k^*$  общего решения  $\bar{y}_k$  однородного уравнения и частного решения  $y_k^*$  неоднородного уравнения. Найдем сначала общее решение однородного уравнения. Дискриминант равен  $D = 1 + 24 = 25 > 0$ , и корни квадратного уравнения  $q^2 - q - 6 = 0$  равны  $q_1 = 3$ ,  $q_2 = -2$ , так что  $y_k^{(1)} = 3^k$ ,  $y_k^{(2)} = (-2)^k$ . Частное решение  $y_k^*$  будем искать в виде  $y_k^* = c2^k$ , где  $c = \text{const}$ . Подставляя  $y_k^* = c2^k$  в (18), получим  $c(2^{k+1} - 2^k - 6 \cdot 2^{k-1}) = c \cdot 2^{k-1}(-4) = 2^{k+1}$ ,  $c = -1$ .

Общее решение уравнения (18) имеет вид

$$y_k = c_1 \cdot 3^k + c_2 (-2)^k - 2^k.$$

**6. Разностное уравнение второго порядка с переменными коэффициентами. Задача Коши и краевая задача.** Рассмотрим теперь разностное уравнение с переменными коэффициентами

$$b_i y_{i+1} - c_i y_i + a_i y_{i-1} = f_i, \quad a_i \neq 0, \quad b_i \neq 0, \quad i = 0, 1, 2, \dots \quad (19)$$

Так как  $b_i \neq 0$ , то из (19) получаем следующее рекуррентное соотношение:

$$y_{i+1} = \frac{c_i y_i - a_i y_{i-1} + f_i}{b_i}, \quad b_i \neq 0. \quad (20)$$

Выразим  $y_{i+1}$  и  $y_{i-1}$  через  $y_i$  и разности первого и второго порядков. Тогда уравнение (19) перепишется в виде

$$\Delta^2 y_i + (b_i - a_i) \Delta y_i - (c_i - a_i - b_i) y_i = f_i, \quad a_i \neq 0, \quad b_i \neq 0.$$

Решение разностного уравнения первого порядка зависит от одной произвольной постоянной и определяется однозначно, если задано одно дополнительное условие, например,  $y_0 = c_0$ . Решение уравнения второго порядка определяется двумя произвольными постоянными и может быть найдено, если заданы два дополнительных условия. Если оба условия заданы в двух соседних точках, то это *задача Коши*. Если же два условия заданы в двух разных (но не соседних) точках, то получаем *краевую задачу*. Для нас основной интерес будут представлять краевые задачи. Введем обозначение

$$Ly_i = b_i y_{i+1} - c_i y_i + a_i y_{i-1}$$

и сформулируем эти задачи более подробно.

*Задача Коши:* найти решение уравнения

$$Ly_i = f_i, \quad i = 1, 2, \dots, \quad (21)$$

при дополнительных условиях

$$y_0 = \mu_1, \quad y_1 = \mu_2. \quad (22)$$

Второе условие (22) можно записать иначе:  $\Delta y_0 = y_1 - y_0 = \mu_2 - \mu_1 = \bar{\mu}_1$ , и говорить, что в случае задачи Коши заданы в одной точке  $i = 0$  величины

$$y_0 = \mu_1, \quad \Delta y_0 = \bar{\mu}_1. \quad (22')$$

*Краевая задача:* найти решение уравнения

$$Ly_i = f_i, \quad i = 1, 2, \dots, N-1,$$

при дополнительных условиях

$$y_0 = \mu_1, \quad y_N = \mu_2, \quad N \geq 2. \quad (23)$$

В граничных узлах  $i = 0$  и  $i = N$  можно задать не только значения функций, но и их разности и комбина-

ции, т. е. выражения  $\alpha_1 \Delta y_0 + \beta_1 y_0$  при  $i = 0$  и  $\alpha_2 \nabla y_N + \beta_2 y_N$  при  $i = N$ . Такие условия можно записать в виде

$$y_0 = \kappa_1 y_1 + \mu_1, \quad y_N = \kappa_2 y_{N-1} + \mu_2. \quad (24)$$

Если  $\kappa_1 = \kappa_2 = 0$ , то отсюда получаем *условия первого рода*; при  $\kappa_1 = 1$ ,  $\kappa_2 = 1$  имеем *условия второго рода*

$$\Delta y_0 = -\mu_1, \quad \nabla y_N = \mu_2. \quad (25)$$

Если  $\kappa_{1,2} \neq 0; 1$ , то (24) называют *условиями третьего рода*:

$$\begin{aligned} -\kappa_1 \Delta y_0 + (1 - \kappa_1) y_0 &= \mu_1, \\ \kappa_2 \nabla y_N + (1 - \kappa_2) y_N &= \mu_2. \end{aligned} \quad (26)$$

Кроме того, возможны краевые задачи с комбинацией этих краевых условий: при  $i = 0$  — условия одного типа, при  $i = N$  — условия другого типа.

Решение задачи Коши находится непосредственно из уравнения (21) по рекуррентной формуле (20) с учетом начальных данных  $y_0 = \mu_1$ ,  $y_1 = \mu_2$ . Решение краевых задач находится более сложным методом — методом исключения — и будет изложено ниже.

Для уравнения с постоянными коэффициентами решение краевой задачи может быть найдено в явном виде.

Пример. Найти решение краевой задачи

$$\Delta^2 y_{i-1} = 1, \quad i = 1, 2, \dots, N-1, \quad y_0 = 0, \quad y_N = 0. \quad (27)$$

Однородное уравнение  $\Delta^2 y_{i-1} = y_{i+1} - 2y_i + y_{i-1} = 0$  имеет общее решение  $\bar{y}_i = c_1 + c_2 i$ . Частное решение  $y_i^*$  неоднородного уравнения  $\Delta^2 y_{i-1} = y_{i+1} - 2y_i + y_{i-1} = 1$  ищем в виде  $y_i^* = ci^2$ . Подставляя это выражение в уравнение (27), находим  $\Delta^2 y_{i-1}^* = c((i+1)^2 - 2i^2 + (i-1)^2) = 1$ , т. е.  $c = 1/2$ , так что  $y_i = \bar{y}_i + y_i^* = c_1 + c_2 i + i^2/2$ . Для определения  $c_1$  и  $c_2$  служат краевые условия при  $i = 0$ ,  $i = N$ :  $y_0 = c_1 = 0$ ,  $y_N = c_2 N + N^2/2 = 0$ ,  $c_2 = -N/2$ . Таким образом,

$$y_i = -\frac{1}{2}iN + \frac{1}{2}i^2 = -\frac{1}{2}i(N-i)$$

есть решение задачи (27).

### § 3. Решение разностных краевых задач для уравнений второго порядка

**1. Решение разностных краевых задач методом прогонки. Краевая задача**

$$\begin{aligned} a_i y_{i-1} - c_i y_i + b_i y_{i+1} &= -f_i, \quad a_i \neq 0, \quad b_i \neq 0, \\ i &= 1, 2, \dots, N-1, \quad (1) \\ y_0 &= \kappa_1 y_1 + \mu_1, \quad y_N = \kappa_2 y_{N-1} + \mu_2 \end{aligned}$$

представляет собой систему линейных алгебраических уравнений с трехдиагональной матрицей размера  $(N+1) \times (N+1)$ :

$$A = \begin{bmatrix} 1 - \kappa_1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ a_1 - c_1 & b_1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & a_i - c_i & b_i & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & a_{N-1} - c_{N-1} & b_{N-1} & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & -\kappa_2 & 1 & 0 \end{bmatrix}.$$

Вместо (1) можно написать

$$Ay = f, \quad y = (y_0, y_1, \dots, y_N), \quad f = (\mu_1, -f_1, \dots, -f_{N-1}, \mu_2). \quad (2)$$

В случае первой краевой задачи соответствующая матрица имеет размерность  $(N-1) \times (N-1)$ .

Для решения краевой задачи (1) можно использовать следующий метод исключения, называемый *методом прогонки*. Предположим, что имеет место соотношение

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1} \quad (3)$$

с неопределенными коэффициентами  $\alpha_{i+1}$  и  $\beta_{i+1}$ , и подставим  $y_{i+1} = \alpha_i y_i + \beta_i$  в (1):

$$(a_i \alpha_i - c_i) y_i + b_i y_{i+1} = -(f_i + a_i \beta_i),$$

сравнивая это тождество с (3), находим

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \dots, N-1, \quad (4)$$

$$\beta_{i+1} = \frac{a_i \beta_i + f_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \dots, N-1. \quad (5)$$

Используем краевое условие при  $i = 0$  для определения

$\alpha_1, \beta_1$ . Из формул (3) и (4) для  $i = 0$  находим

$$\alpha_1 = \kappa_1, \quad \beta_1 = \mu_1. \quad (6)$$

Зная  $\alpha_1$  и  $\beta_1$  и переходя от  $i$  к  $i + 1$  в формулах (4) и (5), определим  $\alpha_i$  и  $\beta_i$  для всех  $i = 2, 3, \dots, N$ . Вычисления по формуле (3) ведутся путем перехода от  $i + 1$  к  $i$  (т. е. зная  $y_{i+1}$ , находим  $y_i$ ), и для начала этих вычислений надо задать  $y_N$ . Определим  $y_N$  из краевого условия  $y_N = \kappa_2 y_{N-1} + \mu_2$  и условия (3) при  $i = N - 1$ :  $y_{N-1} = \alpha_N y_N + \beta_N$ . Отсюда находим

$$y_N = \frac{\mu_2 + \kappa_2 \beta_N}{1 - \alpha_N \kappa_2}. \quad (7)$$

Соберем все формулы прогонки и запишем их в порядке применения:

$$\stackrel{(\rightarrow)}{\alpha_{i+1}} = \frac{b_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \dots, N - 1, \quad \alpha_1 = \kappa_1; \quad (8)$$

$$\stackrel{(\rightarrow)}{\beta_{i+1}} = \frac{a_i \beta_i + f_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \dots, N - 1, \quad \beta_1 = \mu_1; \quad (9)$$

$$\begin{aligned} \stackrel{(\leftarrow)}{y_i} &= \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad i = N - 1, N - 2, \dots, 2, 1, 0, \\ y_N &= \frac{\mu_2 + \kappa_2 \beta_N}{1 - \alpha_N \kappa_2}. \end{aligned} \quad (10)$$

Стрелки показывают направление счета:  $(\rightarrow)$  от  $i$  к  $i + 1$ ,  $(\leftarrow)$  — от  $i + 1$  к  $i$ .

Таким образом, краевая задача для уравнения второго порядка сведена к трем задачам Коши для уравнений первого порядка.

**2. Устойчивость метода прогонки.** Формулы прогонки можно применять, если знаменатели дробей (8) и (10) не обращаются в нуль. Достаточными условиями этого являются неравенства

$$\begin{aligned} |c_i| &\geq |a_i| + |b_i|, \quad i = 1, 2, \dots, N - 1, \\ |\kappa_1| &\leq 1, \quad |\kappa_2| \leq 1, \quad |\kappa_1| + |\kappa_2| < 2. \end{aligned} \quad (11)$$

Покажем, что при условиях (11) знаменатели  $c_i - a_i \alpha_i$  и  $1 - \alpha_N \kappa_2$  не обращаются в нуль и

$$|\alpha_i| \leq 1, \quad i = 1, 2, \dots, N. \quad (12)$$

Предположим, что  $|\alpha_i| \leq 1$ , и покажем, что  $|\alpha_{i+1}| \leq 1$ :

тогда отсюда и из условия  $|\alpha_i| = |\chi_1| \leq 1$  будет следовать (12). Рассмотрим разность  $|c_i - a_i\alpha_i| - |b_i| \geq |c_i| - |a_i||\alpha_i| - |b_i| \geq |a_i|(1 - |\alpha_i|) \geq 0$ , так что  $|c_i - a_i\alpha_i| \geq |b_i| > 0$ , и  $|\alpha_{i+1}| = |b_i|/|c_i - a_i\alpha_i| \leq 1$ .

Заметим, что если  $|c_{i_0}| > |a_{i_0}| + |b_{i_0}|$  хотя бы в одной точке  $i = i_0$ , то  $|\alpha_i| < 1$  для всех  $i > i_0$ , и в том числе для  $i = N$ :  $|\alpha_N| < 1$ . Тогда  $|1 - \alpha_N\chi_2| \geq 1 - |\alpha_N||\chi_2| \geq 1 - |\alpha_N| > 0$ , и условие  $|\chi_1| + |\chi_2| < 2$  является лишним. Если  $|\chi_1| < 1$ , то  $|\alpha_N| < 1$ . Если же  $|\chi_1| = 1$ , то  $|\chi_2| < 1$  и  $|\alpha_N| \leq 1$ , и мы имеем  $|1 - \alpha_N\chi_2| \geq 1 - |\alpha_N| \times |\chi_2| \geq 1 - |\chi_2| > 0$ . Таким образом, при выполнении условий (11) задача (1) имеет единственное решение, которое мы находим по формулам прогонки (8)–(10).

Вычисления по формулам (8)–(10) ведутся на ЭВМ приближенно, с конечным числом значащих цифр. В результате ошибок округления фактически находится не функция  $y_i$  — решение задачи (1), — а  $\tilde{y}_i$  — решение той же задачи с возмущенными коэффициентами  $\tilde{a}_i$ ,  $\tilde{b}_i$ ,  $\tilde{c}_i$ ,  $\tilde{\chi}_1$ ,  $\tilde{\chi}_2$  и правыми частями  $\tilde{f}_i$ ,  $\tilde{\mu}_1$ ,  $\tilde{\mu}_2$ . Возникает естественный вопрос: не происходит ли в ходе вычислений возрастание ошибки округления, что может привести как к потере точности, так и к невозможности продолжать вычисления из-за роста определяемых величин. Примером может служить нахождение  $y_i$  по формуле  $y_{i+1} = qy_i$  при  $q > 1$ . Поскольку  $y_n = q^n y_0$ , для любого  $y_0$  можно указать такое  $n_0$ , при котором  $y_{n_0}$  будет машинной бесконечностью. Фактически в силу ошибок округления определяется не точное значение  $y_i$ , а значение  $\tilde{y}_i$  из уравнения  $\tilde{y}_{i+1} = q\tilde{y}_i + \eta$ , где  $\eta$  — ошибка округления. Для погрешности  $\delta y_i = \tilde{y}_i - y_i$  получим уравнение  $\delta y_{i+1} = q\delta y_i + \eta$  ( $i = 0, 1, \dots$ ,  $\delta y_0 = \eta$ ). Из формулы  $\delta y_i = q^i \eta + \eta(q^i - 1)/(q - 1)$  видно, что ошибка  $\delta y_i$  при  $q > 1$  экспоненциально растет с ростом  $i$ .

Вернемся к методу прогонки и покажем, что при  $|\alpha_i| \leq 1$  ошибка  $\delta y_i$  не нарастает. В самом деле, из  $\tilde{y}_i = \alpha_{i+1}\tilde{y}_{i+1} + \beta_{i+1}$ ,  $y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1}$  следует  $\delta y_i = \alpha_{i+1}\delta y_{i+1}$ ,  $|\delta y_i| \leq |\alpha_{i+1}| |\delta y_{i+1}| \leq |\delta y_{i+1}|$ , так как  $|\alpha_{i+1}| \leq 1$ .

Если учесть, что в ходе вычислений возмущаются и коэффициенты  $\alpha_{i+1}$ ,  $\beta_{i+1}$ , то можно показать, что ошибка  $\delta y_i$  пропорциональна квадрату числа узлов  $N$ :

$$\max_{1 \leq i \leq N} |\delta y_i| \leq \varepsilon_0 N^2,$$

где  $\varepsilon_0$  — ошибка округления. Отсюда видна связь между

требуемой точностью в решения задачи, числом  $N$  уравнений и числом значащих цифр ЭВМ, поскольку  $\varepsilon_0 N^2 \approx \varepsilon$ .

**3. Другие варианты метода прогонки.** Рассмотренный выше метод прогонки (8)–(10), при котором определение  $y_i$  производится последовательно справа налево, называют *правой прогонкой*. Аналогично записываются формулы *левой прогонки*:

$$\overset{(-)}{\xi_i} = \frac{a_i}{c_i - b_i \xi_{i+1}}, \quad i = N-1, N-2, \dots, 2, 1, \quad \xi_N = \alpha_2, \quad (13)$$

$$\overset{(-)}{\eta_i} = \frac{b_i \eta_{i+1} + f_i}{c_i - b_i \xi_{i+1}}, \quad i = N-1, N-2, \dots, 2, 1, \quad \eta_N = \mu_2, \quad (14)$$

$$\overset{(+)}{y_{i+1}} = \xi_{i+1} y_i + \eta_{i+1}, \quad i = 0, 1, \dots, N-1, \quad y_0 = \frac{\mu_1 + \alpha_1 \eta_1}{1 - \xi_1 \alpha_1}. \quad (15)$$

В самом деле, предполагая, что  $y_{i+1} = \xi_{i+1} y_i + \eta_{i+1}$ , исключим из (1)  $y_{i+1}$ ; получим

$$-f_i = a_i y_{i-1} + (b_i \xi_{i+1} - c_i) y_i + b_i \eta_{i+1},$$

или

$$y_i = \frac{a_i}{c_i - b_i \xi_{i+1}} y_{i-1} + \frac{f_i + b_i \eta_{i+1}}{c_i - b_i \xi_{i+1}}.$$

Сравнивая с формулой  $y_i = \xi_i y_{i-1} + \beta_i$ , получим (13) и (14). Значение  $y_0$  находим из условия  $y_0 = \alpha_1 y_1 + \mu_1$  и формулы  $y_0 = \xi_1 y_1 + \eta_1$ . Из неравенства  $|c_i - b_i \xi_{i+1}| \geq |c_i| - |b_i| |\xi_{i+1}| \geq |a_i| + |b_i| (1 - |\xi_{i+1}|)$ ,  $|1 - \xi_1 \alpha_1| \geq 1 - |\xi_1| |\alpha_1|$  видно, что условия (11) гарантируют применимость формул левой прогонки и их вычислительную устойчивость, так как  $|\xi_i| \leq 1$  ( $i = 1, 2, \dots, N$ ).

Комбинация левой и правой прогонок дает *метод встречных прогонок*. В этом методе в области  $0 \leq i \leq i_0 + 1$  по формулам (8), (9) вычисляются прогоночные коэффициенты  $\alpha_i$ ,  $\beta_i$ , а в области  $i_0 \leq i \leq N$  по формулам (13), (14) находятся  $\xi_i$  и  $\eta_i$ . При  $i = i_0$  производится сопряжение решений в форме (10) и (15).

Из формул  $y_{i_0} = \alpha_{i_0+1} y_{i_0+1} + \beta_{i_0+1}$ ,  $y_{i_0+1} = \xi_{i_0+1} y_{i_0} + \eta_{i_0+1}$  находим

$$y_{i_0} = \frac{\beta_{i_0+1} + \alpha_{i_0+1} \eta_{i_0}}{1 - \alpha_{i_0+1} \xi_{i_0+1}}.$$

Эта формула имеет смысл, так как хотя бы одна из величин  $|\xi_{i_0+1}|$  или  $|\alpha_{i_0+1}|$  в силу (11) меньше единицы, и, следовательно,  $1 - \alpha_{i_0+1}\xi_{i_0+1} > 0$ . Зная  $y_{i_0}$ , можно по формуле (10) найти все  $y_i$  при  $i < i_0$ , а по формуле (15)—значения  $y_i$  при  $i > i_0$ . Вычисления при  $i > i_0$  и  $i < i_0$  проводятся автономно (имеет место распараллеливание вычислений). Метод встречных прогонок особенно удобен, если, например, требуется найти  $y_i$  лишь в одном узле  $i = i_0$ .

#### § 4. Разностные уравнения как операторные уравнения

**1. Линейное пространство.\*)** Рассмотрим множество  $H$  элементов  $x, y, z, \dots$ , относительно которых известно, что: каждой паре элементов  $x$  и  $y$  из  $H$  каким-то образом сопоставляется третий элемент  $z \in H$ , называемый их суммой и обозначаемый  $z = x + y$ ; каждому элементу  $x \in H$  и каждому числу  $\lambda$  сопоставляется элемент  $u \in H$ , называемый произведением  $x$  на число  $\lambda$  и обозначаемый через  $u = \lambda x$ .

Множество  $H$  называется *линейным пространством*, если операции сложения и умножения на число, определенные для его элементов  $x, y, z, \dots$ , удовлетворяют следующим аксиомам:

- 1)  $x + y = y + x$  для любых  $x, y \in H$  (коммутативность сложения);
- 2)  $(x + y) + z = x + (y + z)$  для любых  $x, y, z \in H$  (ассоциативность сложения);
- 3) существует элемент «пуль», обозначаемый 0, такой, что  $x + 0 = x$  при любом  $x \in H$ ;
- 4) для любого элемента  $x \in H$  существует противоположный элемент  $(-x)$ , такой, что  $x + (-x) = 0$ ;
- 5)  $1 \cdot x = x$ ;
- 6)  $(\lambda\mu)x = \lambda(\mu x)$  (ассоциативность умножения);
- 7)  $\lambda(x + y) = \lambda x + \lambda y$ ;  $(\lambda + \mu)x = \lambda x + \mu x$  (дистрибутивность умножения относительно сложения), где  $\lambda$  и  $\mu$ —любые числа.

Линейное пространство называют *комплексным*, если для его элементов определено умножение на комплексные числа, и *действительным*, если определено умножение только на действительные числа.

---

\*) См., например, Ильин В. А., Позняк Э. Г. Линейная алгебра.— М.: Наука, 1974.

Элементы  $x, y, z, \dots$  линейного пространства  $H$  называют *векторами*.

Векторы  $x_1, x_2, \dots, x_N$  называют *линейно независимыми*, если равенство

$$c_1x_1 + c_2x_2 + \dots + c_Nx_N = 0 \quad (1)$$

возможно только при  $c_1 = c_2 = \dots = c_N = 0$ . Если же найдутся  $c_1, c_2, \dots, c_N$ , не все равные нулю, такие, что имеет место равенство (1), то векторы  $x_1, \dots, x_N$  называют *линейно зависимыми*. Максимальное число (если оно существует) линейно независимых векторов линейного пространства  $H$  называется *размерностью* пространства  $H$ . Пространство, обладающее бесконечным множеством линейно независимых векторов, называется *бесконечно-мерным*.

Пространство  $H$  называется *нормированным*, если для каждого  $x \in H$  определено вещественное число  $\|x\|$ , называемое *нормой*, которое удовлетворяет условиям:

- 1)  $\|x\| > 0$  при  $x \neq 0$ ;  $\|x\| = 0$ , если  $x = 0$ ;
- 2)  $\|x + y\| \leq \|x\| + \|y\|$  (неравенство треугольника);
- 3)  $\|cx\| = |c| \cdot \|x\|$ , где  $c$  — число.

*Евклидовым* (соответственно *унитарным*) пространством называется конечномерное действительное линейное пространство  $H$  (соответственно конечномерное комплексное линейное пространство  $H$ ), в котором каждой паре векторов  $x, y$  поставлено в соответствие вещественное (комплексное) число  $(x, y)$ , называемое *скалярным произведением* этих векторов, причем выполнены условия:

В случае евклидова пространства:

- 1)  $(x, y) = (y, x)$  (симметричность);
- 2)  $(x_1 + x_2, y) = (x_1, y) + (x_2, y)$  (дистрибутивность);
- 3)  $(\lambda x, y) = \lambda(x, y)$  (однородность), где  $\lambda$  — любое действительное число;
- 4) если  $x \neq 0$ , то  $(x, x) > 0$ .

В случае унитарного пространства:

- 1)  $(x, y) = (\overline{y}, \overline{x})$ ;
- 2)  $(x_1 + x_2, y) = (x_1, y) + (x_2, y)$ ;
- 3)  $(\lambda x, y) = \lambda(x, y)$  для любого комплексного числа  $\lambda$ ;
- 4) если  $x \neq 0$ , то  $(x, x) > 0$ .

Заметим, что введенное скалярное произведение  $(x, y)$  порождает в  $H$  норму

$$\|x\| = \sqrt{(x, x)}, \quad (2)$$

Справедливо неравенство Коши — Буняковского

$$|(x, y)|^2 \leq (x, x) \cdot (y, y), \quad (3)$$

которое с учетом (2) можно записать в виде

$$|(x, y)| \leq \|x\| \cdot \|y\|.$$

**2. Линейные операторы в конечномерном пространстве.** Пусть  $H$  — конечномерное линейное пространство со скалярным произведением  $(x, y)$ . Обозначим через  $D$  некоторое подпространство  $H$ . Если каждому вектору  $x \in D$  поставлен в соответствие по определенному правилу вектор  $y = Ax$  из  $H$ , то говорят, что в  $H$  задан *оператор*  $A$ . Множество  $D \subset H$  называется *областью определения* оператора  $A$  и обозначается  $D(A)$ . Множество всех векторов вида  $y = Ax$ ,  $x \in D(A)$  называется *областью значений* оператора  $A$  и обозначается  $R(A)$ . Если  $D(A) = H$ , то говорят, что оператор  $A$  задан на  $H$ .

Оператор  $A$  называют *линейным*, если он а) аддитивен, т. е.  $A(x_1 + x_2) = Ax_1 + Ax_2$  для любых  $x_1, x_2 \in H$ ; б) однороден, т. е.  $A(cx) = cAx$  для любых  $x \in H$  и любых чисел  $c$ . Требования а) и б) эквивалентны условию  $A(c_1x_1 + c_2x_2) = c_1Ax_1 + c_2Ax_2$  для любых  $x_1, x_2 \in H$  и любых чисел  $c_1$  и  $c_2$ .

Линейный оператор называется *ограниченным*, если существует такая постоянная  $M > 0$ , что

$$\|Ax\| \leq M\|x\| \quad \text{для любых } x \in H. \quad (4)$$

Точная нижняя грань множества чисел  $M$ , удовлетворяющих условию (4), называется *нормой* оператора  $A$  и обозначается  $\|A\|$ . Ясно, что

$$\|Ax\| \leq \|A\| \cdot \|x\|. \quad (5)$$

Мы будем всегда рассматривать ограниченные линейные операторы  $A$ , заданные на  $H$  с областью значений  $R(A) \equiv H$ . Такой оператор  $A$  отображает  $H$  в  $H$ , что записывается в виде  $A: H \rightarrow H$ .

В конечномерном пространстве любой линейный оператор ограничен.

Если каждому  $y \in H$  соответствует только один вектор  $x \in H$ , для которого  $Ax = y$ , то этим соответствием определяется оператор  $A^{-1}$ , называемый *обратным*:  $A^{-1}: H \rightarrow H$ . Из определения обратного оператора  $A^{-1}$  следует, что

$$A^{-1}(Ax) = x, \quad A(A^{-1}y) = y \quad \text{для любых } x, y \in H.$$

Оператор  $D$ , действующий по правилу  $Dx = A(Bx)$ , называется *произведением* операторов  $A$  и  $B$  и обозначается  $D = AB$ . Оператор  $E$  называется *единичным* (*тождественным*), если  $Ex = x$  для всех  $x \in H$ . Если существует  $A^{-1}$ , то  $A^{-1}A = AA^{-1} = E$ . Операторы  $A$  и  $B$  называются *перестановочными*, если  $AB = BA$ .

Очевидно, что  $A^{-1}$  — линейный оператор, если линеен оператор  $A$ . Имеет место следующее утверждение:

Для того чтобы линейный оператор  $A: H \rightarrow H$  имел обратный, необходимо и достаточно, чтобы уравнение  $Ax = 0$  имело единственное решение  $x = 0$ .

Оператор  $A^*: H \rightarrow H$  называется *сопряженным* оператору  $A: H \rightarrow H$ , если

$$(Ax, y) = (x, A^*y) \quad \text{для любых } x, y \in H.$$

Оператор  $A$  *самосопряжен* (*симметричен*), если  $A = A^*$  (или  $(Ax, y) = (x, Ay)$  для любых  $x, y \in H$ ). Будем называть линейный оператор  $A$ : *положительным*, если  $(Ax, x) > 0$  ( $x \in H; x \neq 0$ ); *положительно определенным*, если  $(Ax, x) \geq \delta \|x\|^2$  ( $x \in H$ ), где  $\delta > 0$  — число; *неотрицательным*, если  $(Ax, x) \geq 0$  ( $x \in H$ ). Любой оператор  $A$  можно представить в виде суммы:

$$A = A_0 + A_1, \quad A_0 = \frac{1}{2}(A + A^*), \quad A_1 = \frac{1}{2}(A - A^*),$$

где  $A_0 = A_0^*$  — самосопряженный оператор,  $A_1 = -A_1^*$  — кососимметричный оператор, для которого в действительном пространстве  $(A_1x, x) = -(x, A_1x) = -(A_1x, x)$  и, следовательно,  $(A_1x, x) = 0$ . Поэтому для любого оператора  $A$  в действительном пространстве  $H$  выполняется равенство

$$(Ax, x) = (A_0x, x) \quad \text{для любых } x \in H. \quad (6)$$

Мы будем пользоваться операторными неравенствами:

$A \geq 0$ , если  $(Ax, x) \geq 0$ , для всех  $x \in H$ ;

$A > 0$ , если  $(Ax, x) > 0$ , для всех  $x \in H, x \neq 0$ ; (7)

$A \geq \delta E$ , если  $(Ax, x) \geq \delta \|x\|^2$ , для всех  $x \in H$ ,

где  $E$  — единичный оператор.

Неравенство

$$B \geq \alpha A$$

означает, что выполнено условие  $B - \alpha A \geq 0$ , т. е.

$$((B - \alpha A)x, x) \geq 0 \quad (\text{для всех } x \in H).$$

Если  $A \neq A^*$  в действительном пространстве, то неравенство  $A \geq 0$  ( $A > 0$ ) эквивалентно неравенству  $A_0 \geq 0$  ( $A_0 > 0$ ), что следует из (6).

Пусть  $A$  — положительный оператор. Тогда существует обратный оператор  $A^{-1}$ :  $H \rightarrow H$ , причем  $A^{-1} > 0$  при  $A > 0$ ,  $(A^{-1})^* = A^{-1}$  при  $A^* = A$ . В самом деле, оператор  $A^{-1}$  существует, если уравнение  $Ax = 0$  имеет только тривиальное решение. Допустим, что  $Ax = 0$  при  $x \neq 0$ ; тогда  $0 = (Ax, x)$  при  $x \neq 0$ , что противоречит условию  $A > 0$  или  $(Ax, x) > 0$  при  $x \neq 0$ . Таким образом, если  $A > 0$ , то уравнение  $Ax = y$  имеет единственное решение.

**3. Собственные значения линейного оператора.** Пусть  $A$  — самосопряженный оператор в  $N$ -мерном пространстве  $H$  со скалярным произведением  $(\cdot, \cdot)$ . Рассмотрим задачу о собственных значениях оператора  $A$ : требуется найти такие значения параметра  $\lambda$  (собственные значения), при которых однородное уравнение

$$A\xi = \lambda\xi \quad (8)$$

имеет нетривиальные решения (собственные векторы). Приведем основные факты из линейной алгебры о задаче на собственные значения.

1) Самосопряженный оператор  $A$  имеет  $N$  ортонормированных собственных векторов  $\xi_1, \xi_2, \dots, \xi_N$ :

$$(\xi_s, \xi_m) = \delta_{sm}, \quad \delta_{sm} = \begin{cases} 1, & s = m, \\ 0, & s \neq m. \end{cases} \quad (9)$$

2) Соответствующие собственные значения действительны и могут быть расположены в порядке возрастания их абсолютных величин:

$$0 \leq |\lambda_1| \leq |\lambda_2| \leq \dots \leq |\lambda_N|. \quad (10)$$

3) Если  $A$  — положительный оператор, то все собственные числа  $\{\lambda_k\}$  положительны:

$$0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N. \quad (11)$$

В самом деле,  $\lambda_s = (A\xi_s, \xi_s)/\|\xi_s\|^2 = (A\xi_s, \xi_s) > 0$ , так как  $\xi_s \neq 0$ .

4) Произвольный вектор  $x \in H$  можно разложить по собственным векторам оператора  $A = A^*$ :

$$x = \sum_{k=1}^N c_k \xi_k, \quad c_k = (x, \xi_k), \quad (12)$$

причем справедливо равенство

$$\|x\|^2 = \sum_{k=1}^N c_k^2. \quad (13)$$

В самом деле, в силу условия (9) ортонормированности системы  $\{\xi_k\}$  имеем

$$\begin{aligned} \|x\|^2 &= (x, x) = \left( \sum_{k=1}^N c_k \xi_k, \sum_{k'=1}^N c_{k'} \xi_{k'} \right) := \\ &= \sum_{k=1}^N \sum_{k'=1}^N c_k c_{k'} (\xi_k, \xi_{k'}) = \sum_{k=1}^N \sum_{k'=1}^N c_k c_{k'} \delta_{kk'} = \sum_{k=1}^N c_k^2. \end{aligned}$$

5) Если  $A = A^* > 0$ , то решение уравнения  $Ax = f$  можно представить в виде

$$x = \sum_{k=1}^N \frac{f_k}{\lambda_k} \xi_k, \quad (14)$$

где  $f_k = (f, \xi_k)$  — коэффициент Фурье функции  $f$ . Воспользуемся представлениями

$$x = \sum_{k=1}^N c_k \xi_k, \quad f = \sum_{k=1}^N f_k \xi_k$$

и напишем

$$0 = Ax - f = \sum_{k'=1}^N (\lambda_{k'} c_{k'} - f_{k'}) \xi_{k'},$$

Умножая это равенство скалярно на  $\xi_k$  и учитывая, что  $(\xi_k, \xi_{k'}) = \delta_{kk'}$ , найдем  $0 = \lambda_k c_k - f_k$ , т. е.  $c_k = f_k / \lambda_k$ .

6) Норма самосопряженного оператора  $A$  равна модулю его наибольшего собственного значения:

$$\|A\| = \max_{1 \leq k \leq N} |\lambda_k| = |\lambda_N|. \quad (15)$$

В самом деле, пользуясь (12), получим

$$Ax = \sum_{k=1}^N c_k A \xi_k = \sum_{k=1}^N \lambda_k c_k \xi_k,$$

и в силу (10) и (13) имеем

$$\|Ax\|^2 = \sum_{k=1}^N \lambda_k^2 c_k^2 \leq \lambda_N^2 \sum_{k=1}^N c_k^2 = \lambda_N^2 \|x\|^2,$$

т. е.  $\|A\| \leq |\lambda_N|$ . Эта оценка достигается. Действи-

тельно, при  $x = \xi_N$  имеем  $\|Ax\|^2 = \|A\xi_N\|^2 = \|\lambda_N \xi_N\|^2 = |\lambda_N|^2$ , так как  $\|\xi_N\|^2 = 1$ . Отсюда и следует, что  $\|A\| = |\lambda_N|$ .

7) Если  $A = A^*$ , то

$$\|A\| = \sup_{\|x\|=1} |(Ax, x)|. \quad (16)$$

8) Если  $A = A^* > 0$ , то  $\lambda_1 E \leqslant A \leqslant \lambda_N E$ , или

$$\lambda_1 \|x\|^2 \leqslant (Ax, x) \leqslant \lambda_N \|x\|^2, \quad \lambda_1 > 0, \quad x \in H. \quad (17)$$

9) Если оператор  $A$  положителен, то он и положительно определен, т. е. существует такая постоянная  $\delta > 0$ , что из условия  $A > 0$  следует неравенство  $A \geqslant \delta E$ . Для самосопряженного оператора это свойство следует из свойства 8). В общем случае представим  $A$  в виде суммы  $A = A_0 + A_1$ , где  $A_0 = A_0^* > 0$ ,  $A_1 = -A_1^*$  — кососимметрический оператор. Так как  $(A_1 x, x) = 0$ , то  $(Ax, x) = (A_0 x, x) > 0$ . Для  $A_0$  верно свойство 8). Полагая  $\lambda_1 = \lambda_1(A_0) = \delta > 0$ , получаем  $(A_0 x, x) = (Ax, x) \geqslant \delta \|x\|^2$  для всех  $x \in H$ .

10) Если существует  $Q^{-1}$ , то операторные неравенства

$$C \geqslant 0, \quad Q^* C Q \geqslant 0 \quad (18)$$

эквивалентны. Это следует из тождества

$$(Q^* C Q x, x) = (C Q x, Q x) = (C y, y),$$

где  $y = Q x$ ,  $x = Q^{-1} y$ .

11) Пусть  $A_1$  и  $A_2$  — самосопряженные, положительные и нерестановочные операторы в  $H$ :

$$A_1 = A_1^* > 0, \quad A_2 = A_2^* > 0, \quad A_1 A_2 = A_2 A_1. \quad (19)$$

Тогда операторы  $A_1$  и  $A_2$ , их сумма  $A_1 + A_2$  и произведение  $A_1 A_2$  имеют общую систему собственных функций  $\{\xi_k\}$ :

$$\begin{aligned} A_1 \xi_k &= \lambda_k^{(1)} \xi_k, & A_2 \xi_k &= \lambda_k^{(2)} \xi_k, \\ \lambda(A_1 + A_2) &= \lambda(A_1) + \lambda(A_2), \\ \lambda(A_1 A_2) &= \lambda(A_1) \lambda(A_2). \end{aligned}$$

12) Если  $A = A^* > 0$ , то оператор  $A^{-1} = (A^{-1})^* > 0$  также самосопряжен, имеет те же собственные векторы, что и оператор  $A$ , и собственные значения  $\lambda(A^{-1}) = 1/\lambda(A)$ .

В самом деле, из  $A \xi_k = \lambda_k \xi_k$  следует  $\xi_k = \lambda_k A^{-1} \xi_k$ , т. е.  $(A^{-1}) \xi_k = (1/\lambda_k) \xi_k$ . Отсюда заключаем, что неравен-

ства  $\lambda_1 E \leq A \leq \lambda_N E$  и  $(1/\lambda_N)E \leq A^{-1} \leq (1/\lambda_1)E$  эквивалентны.

**4. Обобщенная задача на собственные значения.** Пусть задан самосопряженный положительный оператор  $B$ . Введем новое скалярное произведение  $(x, y)_B = (Bx, y)$  и норму  $\|y\|_B = \sqrt{(By, y)}$ . Пространство  $H$  со скалярным произведением  $(x, y)_B$  называется *энергетическим пространством* и обозначается  $H_B$ .

Рассмотрим обобщенную задачу на собственные значения, состоящую в отыскании нетривиальных решений  $v$  уравнения

$$Av = \mu Bv, \quad v \neq 0, \quad (20)$$

где  $A$  — самосопряженный положительный оператор.

Пусть операторы  $A$  и  $B$  представлены соответственно матрицами  $A = (a_{ij})$ ,  $B = (b_{ij})$  ( $i, j = 1, 2, \dots, N$ ). Операторное уравнение (20) можно записать в виде системы линейных алгебраических уравнений

$$\sum_{j=1}^N a_{ij} v^{(j)} = \mu \sum_{j=1}^N b_{ij} v^{(j)}, \quad i = 1, 2, \dots, N,$$

где  $v^{(0)}, \dots, v^{(N)}$  — компоненты вектора  $v$ . Для определения собственных значений получаем алгебраическое уравнение  $N$ -й степени

$$\det(a_{ij} - \mu b_{ij}) = 0. \quad (21)$$

Для задачи (20) справедливы свойства, аналогичные свойствам обычной задачи на собственные значения, а именно: существуют  $N$  ортонормированных в смысле скалярного произведения  $(x, y)_B$  собственных векторов

$$(v_k, v_m)_B = \delta_{km}, \quad k, m = 1, 2, \dots, N, \quad (22)$$

которым соответствуют собственные значения

$$0 < \mu_1 \leq \dots \leq \mu_N. \quad (23)$$

По аналогии с п. 3 имеем

$$x = \sum_{k=1}^N c_k v_k, \quad c_k = (x, v_k)_B, \quad \|x\|_B^2 = \sum_{k=1}^N c_k^2. \quad (24)$$

Справедливы операторные неравенства

$$\mu_1 B \leq A \leq \mu_N B, \quad (25)$$

причем  $\mu_N$  — норма оператора  $A$  в  $H_B$ . Это значит, что

$$\|Ax\|_B \leq \|A\|_B \|x\|_B.$$

**Замечание.** Неравенства

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0, \quad (26)$$

$$\gamma_1 \leq \mu_k \leq \gamma_2, \quad k = 1, 2, \dots, N, \quad (27)$$

эквивалентны. В самом деле, разложим произвольный вектор  $x = \sum_{k=1}^N c_k v_k$ , найдем  $(A - \gamma B)x = \sum_{k=1}^N c_k (\mu_k - \gamma) Bv_k$  и скалярное произведение

$$((A - \gamma B)x, x) = \sum_{k=1}^N c_k^2 (\mu_k - \gamma) (Bv_k, v_k) = \sum_{k=1}^N (\mu_k - \gamma) c_k^2,$$

где  $\gamma$  — одно из чисел  $\gamma_1$  или  $\gamma_2$ . Полагая  $x = v_k$ , найдем  $((A - \gamma B)v_k, v_k) = \mu_k - \gamma$ . Пусть  $\gamma = \gamma_2$  и выполнено условие  $A \leq \gamma_2 B$ ; тогда  $\mu_k \leq \gamma_2$ . Верно и обратное утверждение. Аналогично проводятся рассуждения при  $\gamma = \gamma_1$ .

**5. Линейные пространства сеточных функций. Разностные операторы.** В дальнейшем мы будем рассматривать функции, заданные на сетке с целочисленными узлами:

$$\omega_N = \{i: i = 0, 1, \dots, N\}.$$

Если на отрезке  $0 \leq x \leq 1$  ввести узлы  $x_i = ih$ ,  $h = 1/N$  ( $i = 0, 1, \dots, N$ ), то получим равномерную сетку с шагом  $h$  как совокупность узлов  $x_i = ih$  с целочисленными индексами:

$$\omega_h = \{x_i = ih: i = 0, 1, \dots, N; h = 1/N\}.$$

Переход от одной сетки к другой очевиден и мы часто не будем делать различия между ними.

Обозначим через  $\Omega_{N+1} = \{y_i, i = 0, 1, \dots, N\}$  пространство сеточных функций, заданных на сетке  $\omega_N$ , через  $\Omega_{N+1}^\circ = \{y_i, i = 0, 1, \dots, N; y_0 = 0, y_N = 0\}$  — подпространство сеточных функций, заданных на  $\omega_N$  и обращающихся в нуль в граничных узлах сетки  $\omega_N$ :  $y_0 = y_N = 0$ . Функции из  $\Omega_{N+1}^\circ$  будем обозначать  $\hat{y}(i) = y_i$ .

Рассмотрим примеры простейших разностных операторов. Для оператора правой разности  $\Delta$  имеем

$$\Delta y_i = y_{i+1} - y_i, \quad i = 0, 1, \dots, N-1;$$

областью определения является  $\Omega_{N+1}$ , областью значений — пространство  $\Omega_N^+ = \{y_i, i = 0, 1, \dots, N-1\}$  размерности  $N$ .

Для оператора левой разности  $\nabla$  имеем

$$\nabla y_i = y_i - y_{i-1}, \quad i = 1, 2, \dots, N;$$

область определения есть  $\Omega_{N+1}$ , область значений — пространство  $\Omega_N^- = \{y_i, i = 1, 2, \dots, N\}$ .

Из формулы

$$\Delta^2 y_{i-1} = \Delta(\Delta y_{i-1}) = \Delta(\nabla y_i) = y_{i+1} - 2y_i + y_{i-1}$$

видно, что оператор второй разности определен для сеточных функций  $y_i$  при  $i = 1, 2, \dots, N-1$ , т. е. отображает  $\Omega_{N+1}$  в пространстве  $\Omega_{N-1} = \{y_i, i = 1, 2, \dots, N-1\}$ . Этим же свойством обладает разностный оператор  $\Lambda$ :

$$\begin{aligned} \Lambda y_i &= b_i y_{i+1} - c_i y_i + a_i y_{i-1} = \\ &= b_i \Delta(\nabla y_i) - (b_i - a_i)(\nabla y_i) - (c_i - a_i - b_i)y_i, \\ &\quad i = 1, 2, \dots, N-1, \end{aligned}$$

т. е.  $\Lambda y_i \in \Omega_{N-1}$ , если  $y_i \in \Omega_{N+1}$ , или, в сокращенной записи,  $\Lambda: \Omega_{N+1} \rightarrow \Omega_{N-1}$ .

Рассмотрим разностную краевую задачу

$$\Lambda y_i = -f_i, \quad i = 1, 2, \dots, N-1, \quad y_0 = \mu_1, \quad y_N = \mu_2 \quad (28)$$

и запишем ее в матричном виде:

$$AY = \Phi, \quad (29)$$

где  $\Phi = (f_1 + a_1 \mu_1, f_2, \dots, f_{N-2}, f_{N-1} + b_{N-1} \mu_2)$  — известный,  $Y = (y_1, y_2, \dots, y_{N-2}, y_{N-1})$  — неизвестные векторы размерности  $N-1$ ,  $A$  — квадратная трехдиагональная матрица размера  $(N-1) \times (N-1)$ :

$$A = - \begin{bmatrix} -c_1 & b_1 & & \cdots & 0 \\ a_2 & -c_2 & b_2 & \cdots & 0 \\ \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{N-1} & -c_{N-1} \end{bmatrix}.$$

Сравнивая (28) и (29), видим, что можно написать

$$\tilde{\Lambda} y_i = -\varphi_i, \quad i = 1, 2, \dots, N-1,$$

$$\tilde{\Lambda} y_1 = -c_1 y_1 + b_1 y_2, \quad \varphi_1 = f_1 + a_1 \mu_1,$$

$$\tilde{\Lambda} y_i = \Lambda y_i, \quad \varphi_i = f_i, \quad i = 2, 3, \dots, N-2, \quad (28')$$

$$\tilde{\Lambda} y_{N-1} = a_{N-1} y_{N-2} - c_{N-1} y_{N-1}, \quad \varphi_{N-1} = f_{N-1} + b_{N-1} \mu_2.$$

Разностный оператор  $\tilde{A}$  отображает  $\Omega_{N-1}$  в  $\Omega_{N-1}$ . Нетрудно заметить, что  $\tilde{A}y_i = \overset{\circ}{\Lambda}y_i$ . Вместо (28') получим

$$\overset{\circ}{\Lambda}y_i = -\varphi_i, \quad i = 1, 2, \dots, N-1.$$

Введем теперь оператор  $A$ , соответствующий матрице (29), полагая

$$Ay_i = -\tilde{A}y_i = -\overset{\circ}{\Lambda}y_i, \quad i = 1, 2, \dots, N-1.$$

Тогда вместо разностной краевой задачи (28) получим операторное уравнение

$$Ay = \varphi,$$

где  $A: \Omega_{N-1} \rightarrow \Omega_{N-1}$ ,  $\varphi \in \Omega_{N-1}$ , т. е. оператор  $A$  действует из  $\Omega_{N-1}$  в  $\Omega_{N-1}$ . Очевидно, что  $A$  — линейный оператор. Заметим, что можно также считать (имея в виду, что  $Ay = -\overset{\circ}{\Lambda}y$ ), что  $A$  отображает  $\overset{\circ}{\Omega}_{N+1}$  в  $\Omega_{N-1}$ .

В пространстве  $H = \Omega_{N-1}$  можно ввести скалярное произведение

$$(y, v) = \frac{1}{N} \sum_{i=1}^{N-1} y_i v_i$$

и норму

$$\|y\| = \sqrt{(y, y)}.$$

Если рассматриваются вторая ( $\kappa_1 = \kappa_2 = 1$ ) или третья ( $\kappa_1 \neq 0, \kappa_2 \neq 0$ ) краевые задачи (см. (1) § 3), то матрица  $A$  есть квадратная матрица размера  $(N+1) \times (N+1)$  и оператор  $A$  определяется следующим образом:

$$\begin{aligned} Ay_i &= -\Lambda y_i = -(b_i y_{i+1} - c_i y_i + a_i y_{i-1}), \quad i = 1, 2, \dots, N-1, \\ Ay_0 &= -(\kappa_1 y_1 - y_0), \quad Ay_N = -(y_N - \kappa_2 y_{N-1}). \end{aligned}$$

В этом случае оператор  $A$  отображает пространство сеточных функций  $H = \Omega_{N+1}$  в себя:  $A: H \rightarrow H$ .

В дальнейшем мы будем рассматривать первую краевую задачу для разностного уравнения второго порядка; в этом случае, как было показано выше,  $H = \Omega_{N-1}$ .

**6. Разностные формулы Грина.** Рассмотрим разностный оператор  $L$ :

$$Ly_i = b_i y_{i+1} - c_i y_i + a_i y_{i-1}, \quad i = 1, \dots, N-1. \quad (30)$$

Если  $b_i \neq a_{i+1}$ , то соответствующая матрица не является

симметричной. Она симметрична только в случае

$$b_i = a_{i+1}, \quad i = 1, 2, \dots, N - 1. \quad (31)$$

Учитывая это условие, перепишем  $Ly_i$  в следующем виде:

$$\begin{aligned} Ly_i &= a_{i+1}y_{i+1} - c_iy_i + a_iy_{i-1} = \\ &= a_{i+1}(y_{i+1} - y_i) - a_i(y_i - y_{i-1}) - (c_i - a_i - a_{i+1})y_i = \\ &= a_{i+1}\nabla y_{i+1} - a_i\nabla y_i - (c_i - a_i - a_{i+1})y_i = \\ &= \Delta(a_i\nabla y_i) - (c_i - a_i - a_{i+1})y_i. \end{aligned} \quad (32)$$

Разобьем отрезок  $[0, 1]$  точками  $x_i$  на  $N$  равных частей, положим  $y(x_i) = y_i = y(i)$  и введем обозначения, которыми будем в дальнейшем всюду пользоваться:

$$\begin{aligned} h &= \frac{1}{N}, \quad x_i = ih, \quad i = 0, 1, \dots, N, \quad x_0 = 0, \quad x_N = 1, \\ y_{x,i} &= \frac{\Delta y_i}{h} = \frac{y_{i+1} - y_i}{h}, \quad y_{\bar{x},i} = \frac{\nabla y_i}{h} = \frac{y_i - y_{i-1}}{h}, \quad (33) \\ y_{\bar{x}x,i} &= y_{\bar{x}x}(i) = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = \frac{\Delta(\nabla y_i)}{h^2}. \end{aligned}$$

Разделим выражение (32) на  $h^2$  и получим разностный оператор

$$\begin{aligned} \Lambda y_i &= (ay_{\bar{x}})_{x,i} - d_i y_i, \\ d_i &= \frac{1}{h^2} (c_i - a_i - a_{i+1}), \quad i = 1, \dots, N - 1. \end{aligned} \quad (34)$$

В § 1 была получена формула суммирования по частям

$$\sum_{i=0}^{N-1} y_i \Delta v_i = - \sum_{i=1}^N v_i \nabla y_i + (yv)_N - (yv)_0. \quad (35)$$

Пользуясь обозначениями (33), перепишем ее в виде

$$\sum_{i=0}^{N-1} y_i v_{x,i} h = - \sum_{i=1}^N v_i y_{\bar{x},i} h + (yv)_N - (yv)_0, \quad (36)$$

так как  $\sum_{i=0}^{N-1} y_i \Delta v_i = \sum_{i=0}^{N-1} y_i \left( \frac{\Delta v_i}{h} \right) h = \sum_{i=0}^{N-1} y_i v_{x,i} h.$

Для дальнейшего изложения нам удобнее в левой части (36) вести суммирование от  $i = 1$  до  $i = N - 1$ ; это

приводит к формуле

$$\sum_{i=1}^{N-1} y_i v_{x,i} h = - \sum_{i=1}^N v_i y_{\bar{x},i} h + (yv)_N - y_0 v_1. \quad (37)$$

Подставим сюда  $v_i = a_i z_{\bar{x},i}$ ; получим

$$\sum_{i=1}^{N-1} y_i (az_{\bar{x}})_{x,i} h = - \sum_{i=1}^N a_i y_{\bar{x},i} z_{\bar{x},i} h + (ayz_{\bar{x}})_N - y_0 (az_{\bar{x}})_1. \quad (38)$$

Это — *первая разностная формула Грина*. Поменяем в ней местами  $y_i$  и  $z_i$ :

$$\sum_{i=1}^{N-1} z_i (ay_{\bar{x}})_{x,i} h = - \sum_{i=1}^N a_i z_{\bar{x},i} y_{\bar{x},i} h + (ay_{\bar{x}}z)_N - z_0 (ay_{\bar{x}})_1 \quad (38')$$

Вычитая (38') из (38), получаем *вторую разностную формулу Грина*

$$\begin{aligned} \sum_{i=1}^{N-1} y_i (az_{\bar{x}})_{x,i} h &= \sum_{i=1}^{N-1} z_i (ay_{\bar{x}})_{x,i} h + a_N (yz_{\bar{x}} - zy_{\bar{x}})_N - \\ &- (y_0 (az_{\bar{x}})_1 - z_0 (ay_{\bar{x}})_1). \end{aligned} \quad (39)$$

Если выполнены условия

$$y_0 = z_0 = 0, \quad y_N = z_N = 0, \quad (40)$$

т. е.  $\overset{\circ}{y} = y$ ,  $\overset{\circ}{z} = z \in \overset{\circ}{\Omega}_{N+1}$ , то в правой части равенства (39) два последних слагаемых обращаются в нуль и

$$\sum_{i=1}^{N-1} \overset{\circ}{y}_i (az_{\bar{x}})_{x,i} h = \sum_{i=1}^{N-1} \overset{\circ}{z}_i (ay_{\bar{x}})_{x,i} h. \quad (41)$$

Вычитая из обеих частей тождества (41) сумму  $\sum_{i=1}^{N-1} d_i \overset{\circ}{y}_i \overset{\circ}{z}_i h_i$  получаем *вторую формулу Грина* для  $y, z \in \overset{\circ}{\Omega}_{N+1}$ :

$$\sum_{i=1}^{N-1} \overset{\bullet}{y}_i \Lambda \overset{\circ}{z}_i h = \sum_{i=1}^{N-1} \overset{\circ}{z}_i \Lambda \overset{\bullet}{y}_i h \quad (42)$$

для разностного оператора

$$\Lambda \overset{\bullet}{y}_i = (ay_{\bar{x}})_{x,i} - d_i \overset{\circ}{y}_i, \quad \text{для всех } \overset{\bullet}{y} \in \overset{\circ}{\Omega}_{N+1}. \quad (43)$$

Пусть  $H = \Omega_{N+1}$  — пространство сеточных функций  $y_i$ , заданных при  $i = 1, 2, \dots, N - 1$ , со скалярным произведением

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h$$

и нормой

$$\|y\| = \sqrt{(y, y)}.$$

Введем оператор  $A$ :

$$Ay = -\overset{\circ}{\Lambda}y, \quad y \in H. \quad (44)$$

Тогда вторую формулу Грина можно записать в виде

$$(y, Az) = (Ay, z). \quad (45)$$

Эта формула выражает свойство самосопряженности оператора  $A$ :  $A^* = A$  и, следовательно,  $\overset{\circ}{\Lambda}^* = \overset{\circ}{\Lambda}$ . При  $\overset{\circ}{z} = \overset{\circ}{y} \in \overset{\circ}{\Omega}_{N+1}$  первая формула Грина (38) дает:

$$-\sum_{i=1}^{N-1} \overset{\circ}{y}_i (\overset{\circ}{a} \overset{\circ}{y}_x)_{x,i} h = \sum_{i=1}^N a_i (\overset{\circ}{y}_{x,i})^2 h > 0$$

при  $\overset{\circ}{y}_i \neq 0, a_i > 0 \quad (46)$

(так как  $\overset{\circ}{y}_0 = \overset{\circ}{y}_N = 0$ , то равенство пулю возможно только при  $y_i = 0$  ( $i = 1, \dots, N - 1$ )). Учитывая определение оператора  $A$ , найдем

$$(Ay, y) = \sum_{i=1}^N a_i (\overset{\circ}{y}_{x,i})^2 h + \sum_{i=1}^{N-1} d_i y_i^2 h > 0, \quad a_i > 0, \quad d_i \geq 0. \quad (47)$$

Таким образом, разностный оператор  $A$ , определенный формулами (43), (44), является самосопряженным и положительным:  $A = A^* > 0$ , если

$$a_i > 0, \quad d_i \geq 0, \quad i = 1, 2, \dots, N - 1, \quad a_N > 0. \quad (48)$$

**7. Условие самосопряженности разностного оператора второго порядка.** Мы убедились в том, что условие (31) достаточно для самосопряженности разностного оператора (30) в пространстве  $H = \overset{\circ}{\Omega}_{N+1}$ . Покажем, что условие (31) необходимо для самосопряженности  $L$ . Представим  $L$

в виде суммы:

$$Ly_i = L_1 y_i + L_2 y_i,$$

$$L_1 y_i = a_{i+1} (y_{i+1} - y_i) - a_i (y_i - y_{i-1}) - (c_i - a_i - b_i) y_i,$$

$$L_2 y_i = (b_i - a_{i+1}) y_{i+1}.$$

Оператор  $L_1 y_i = h^2 \Lambda y_i$ ,  $\Lambda y_i = (ay_x)_{x,i} - d_i y_i$ , как было показано в предыдущем пункте, является самосопряженным в пространстве  $H = \overset{\circ}{\Omega}_{N+1}$  или  $H = \Omega_{N-1}$  со скалярным произведением  $(y, v) = \sum_{i=1}^{N-1} y_i v_i h$ . Поэтому можно написать:

$$\begin{aligned} \left( \frac{1}{h^2} L \overset{\circ}{y}, \overset{\circ}{v} \right) &= \left( \overset{\circ}{y}, \frac{1}{h^2} L \overset{\circ}{v} \right) = \\ &= (\overset{\circ}{\Lambda y}, \overset{\circ}{v}) - (\overset{\circ}{y}, \overset{\circ}{\Lambda v}) + \left( \frac{1}{h^2} L_2 \overset{\circ}{y}, \overset{\circ}{v} \right) - \left( \overset{\circ}{y}, \frac{1}{h^2} L_2 \overset{\circ}{v} \right) = \\ &= \sum_{i=1}^{N-1} \frac{1}{h^2} (b_i - a_{i+1}) (y_{i+1} v_i - y_i v_{i+1}) h. \end{aligned}$$

Отсюда видно, что  $(L \overset{\circ}{y}, \overset{\circ}{v}) = (\overset{\circ}{y}, L \overset{\circ}{v})$ , т. е.  $L = L^*$  только при условии

$$\sum_{i=1}^{N-1} (b_i - a_{i+1}) (y_{i+1} v_i - y_i v_{i+1}) h = 0. \quad (49)$$

В силу произвольности  $y_i$  и  $v_i$  можно взять  $y_i = \delta_{i,i_0+1}$ ,  $v_i = \delta_{i,i_0}$ , где  $i_0$  — любой фиксированный узел ( $i_0 = 1, 2, \dots, N-1$ ),  $\delta_{i,i_0}$  — символ Кронекера. Тогда получим  $y_{i+1} v_i - y_i v_{i+1} = \delta_{i,i_0}$ , и условие (49) дает  $b_{i_0} = a_{i_0+1}$ . Тем самым необходимость условия (31) доказана.

Следует отметить, что уравнение

$$Ly_i = -f_i \quad (50)$$

можно привести к виду

$$Ly_i = \Delta(A_i \nabla y_i) - D_i y_i = -F_i, \quad (51)$$

где  $L$  — самосопряженный оператор. В самом деле, умножим обе части уравнения (50) на  $\mu_i \neq 0$ :

$$Ly_i = \mu_i a_i y_{i-1} - \mu_i c_i y_i + b_i \mu_i y_{i+1} = -\mu_i f_i$$

и потребуем, чтобы для полученного уравнения выпол-

нялось условие (34), т. е.

$$b_i \mu_i = (\mu a)_{i+1} = a_{i+1} \mu_{i+1} = A_{i+1}.$$

Отсюда получаем  $\mu_{i+1} = \left( \frac{b_i}{a_{i+1}} \right) \mu_i = \mu_1 \prod_{k=1}^i b_k / a_{k+1}$  и уравнение (51), где  $A_i = a_i \mu_i$ ,  $D_i = \mu_i(c_i - a_i - b_i)$ ,  $F_i = \mu_i f_i$ .

**8. Собственные значения разностного оператора второго порядка.** Рассмотрим разностную задачу на собственные значения:

$$(ay_x)_{x,i} - d_i y_i + \lambda y_i = 0, \quad i = 1, 2, \dots, N-1, \quad y_0 = y_N = 0, \quad (52)$$

или  $Ay = \lambda y$ ,  $y \in \Omega_{N-1}$ , где  $A$  определяется равенством (44). Оператор  $A$  самосопряжен и положителен, поэтому к нему относится все сказанное в п. 4.

Для простейшего случая  $a_i = 1$ ,  $d_i = 0$  собственные значения и собственные векторы можно найти в явном виде. Итак, требуется найти нетривиальные решения однородного уравнения с однородными краевыми условиями

$$y_{xx,i} + \lambda y_i = 0, \quad i = 1, 2, \dots, N-1, \quad hN = 1,$$

$$y_0 = 0, \quad y_N = 0, \quad y_i \neq 0. \quad (53)$$

Перепишем уравнение (53) в виде

$$y_{i-1} - 2\cos \alpha y_i + y_{i+1} = 0, \quad 2\cos \alpha = 2 - \lambda h^2. \quad (54)$$

Общее решение этого уравнения имеет вид

$$y_i = c_1 \cos i\alpha + c_2 \sin i\alpha. \quad (55)$$

Требуем выполнения краевых условий:  $y_0 = c_1 = 0$ ,  $y_N = c_2 \sin N\alpha = 0$ . Так как ищется нетривиальное решение, то  $c_2 \neq 0$  и  $\sin N\alpha = 0$ , т. е.  $N\alpha = m\pi$  ( $m = 0, 1, 2, \dots$ ),  $\alpha = \alpha_m = m\pi/N = m\pi h$ . Из соотношения  $2\cos \alpha = 2 - \lambda h^2$  находим

$$\lambda h^2 = 2(1 - \cos \alpha) = 4 \sin^2 \frac{\alpha}{2},$$

$$\lambda = \lambda_m = \frac{4}{h^2} \sin^2 \frac{m\pi h}{2}. \quad (56)$$

Этому значению  $\lambda_m$  соответствует собственная функция  $y_m(i) = c \sin \pi m x_i$ ,  $c \neq 0$ ,  $x_i = ih$ ,  $i = 0, 1, 2, \dots, N$ , (57) определенная с точностью до произвольного постоянного

множителя. Нетрудно заметить, что

$$\begin{aligned} y_N(i) &= c \sin \pi N x_i = c \sin \pi i = 0, \quad i = 0, 1, 2, \dots, \\ y_{N+1}(i) &= c \sin \pi(N+1)x_i = c \sin [\pi Nx_i + \pi x_i] = \\ &= c \sin \pi x_i \cos \pi i = (-1)^i y_1(i), \\ y_{N+m+1}(i) &= (-1)^i y_m(i), \quad m = 1, 2, \dots, N-1. \end{aligned}$$

Следовательно, линейно независимы лишь функции  $y_m(i)$  при  $m < N$ . Таким образом, найдено нетривиальное решение (собственные функции  $y_m(i)$ , соответствующие собственным значениям  $\lambda_m$ ).

Выберем множитель  $c$  так, чтобы норма функций  $y_m(i)$  была равна единице:  $\|y_m(i)\| = c \|\sin \pi m x_i\| = 1$ ,  $c > 0$ . Для этого надо вычислить

$$\|\sin \pi m x_k\|^2 = \sum_{k=1}^{N-1} h \sin^2 \pi m x_k = \frac{1}{2} \sum_{k=1}^{N-1} h (1 - \cos 2\pi m x_k).$$

Обозначая  $\alpha = 2\pi m h$  и заменяя  $\cos 2\pi m x_k = \cos \alpha k = \operatorname{Re} e^{i\alpha k}$ , найдем

$$\begin{aligned} \sum_{k=1}^{N-1} h \cos 2\pi m x_k &= \operatorname{Re} \sum_{k=1}^{N-1} h e^{i\alpha k} = h \operatorname{Re} \frac{e^{i\alpha} - e^{i\alpha N}}{1 - e^{i\alpha}} = -h, \\ \|\sin \pi m x_k\|^2 &= \frac{(N-1)h}{2} - \frac{1}{2} \sum_{k=1}^{N-1} h \cos 2\pi m x_k = \\ &= \frac{Nh}{2} = \frac{1}{2}, \end{aligned}$$

$$\|\sin \pi m x\| = 1/\sqrt{2};$$

следовательно,  $c = \sqrt{2}$ . Таким образом, функция

$$y_m(i) = \sqrt{2} \sin \pi m x_i \quad (58)$$

нормирована к единице.

Собственные функции  $y_s(i)$  и  $y_m(i)$ , соответствующие разным собственным значениям  $\lambda_s$  и  $\lambda_m$ , ортогональны в смысле скалярного произведения

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h.$$

Задача (53) является частным случаем задачи (8) с оператором  $Ay(i) = -\hat{y}_{xx}(i)$ . Этот оператор, очевидно,

самосопряжен и положителен, так как

$$(Ay, y) = \sum_{i=1}^{N-1} (y_{x,i})^2 h > 0.$$

Поэтому все сказанное в п. 3 остается в силе и в данном случае.

Собственные значения  $\lambda_s$  возрастают с ростом  $s$ , так как  $\sin \frac{\pi h}{2} s < \sin \frac{\pi h}{2} (s+1) < 1$  при  $s \leq N$ . Наименьшее собственное значение равно  $\lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}$ . Наибольшее собственное значение равно  $\lambda_{N-1} = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}$ ,

$$\text{так как } \sin \frac{\pi h}{2} (N-1) = \sin \left( \frac{\pi}{2} - \frac{\pi h}{2} \right) = \cos \frac{\pi h}{2}.$$

Переписав  $\lambda_1$  в виде  $\lambda_1 = \pi^2 \left( \frac{\sin \xi}{\xi} \right)^2$ ,  $\xi = \pi h/2 \leq \pi/4$  и учитывая, что  $\sin \xi / \xi$  убывает и имеет минимум при  $\xi = \pi/4$ , получаем  $\lambda_1 \geq 8$  при  $h \leq 1/2$ .

Для  $\lambda_{N-1}$  имеем оценку  $\lambda_{N-1} < 4/h^2$  и, следовательно,

$$8 < \lambda_k < 4/h^2, \quad k = 1, 2, \dots, N-1.$$

## § 5. Принцип максимума для разностных уравнений

**1. Принцип максимума и его следствия.** Для разностных уравнений второго порядка с положительными коэффициентами

$$Ly_i = a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -\varphi_i,$$

$$i = 1, 2, \dots, N-1, \quad y_0 = \mu_1, \quad y_N = \mu_2, \quad (1)$$

$$a_i > 0, \quad b_i > 0, \quad c_i \geq a_i + b_i, \quad i = 1, 2, \dots, N-1, \quad (2)$$

имеет место следующий принцип максимума.

**Теорема 1 (принцип максимума).** Пусть разностный оператор  $L$  определен формулами (1), (2). Если для функции  $y_i$ , заданной на сетке  $\omega$  и отличной от постоянной при  $1 \leq i \leq N-1$ , выполнено условие  $Ly_i \geq 0$  ( $Ly_i \leq 0$ ) при всех  $i = 1, 2, \dots, N-1$ , то эта функция не может принимать наибольшего положительного (наименьшего отрицательного) значения во внутренних узлах сетки.

**Доказательство.** Пусть  $Ly_i \geq 0$  ( $i = 1, 2, \dots, N-1$ ). Предположим, что теорема неверна и  $y_i$  в не-

котором внутреннем узле  $i = i_*, 1 \leq i_* \leq N - 1$ , достигает наибольшего положительного значения  $y_{i_*} = \max_{0 \leq i \leq N} y_i = M_0 > 0$ . Так как  $y_i \neq \text{const}$ , то найдется внутренний узел  $i_0$  ( $i_0$  может совпадать с  $i_*$ ), в котором  $y_{i_0} = y_{i_*} = M_0 > 0$ , а в одном из соседних узлов, например, в узле  $i = i_0 - 1$ , выполняется строгое неравенство  $y_{i_0-1} < y_{i_0}$ .

Запишем выражение для  $Ly_i$  в виде  $Ly_i = b_i(y_{i+1} - y_i) - a_i(y_i - y_{i-1}) - (c_i - a_i - b_i)y_i$ . В узле  $i = i_0$  имеем

$$\begin{aligned} Ly_{i_0} = b_{i_0}(y_{i_0+1} - y_{i_0}) - a_{i_0}(y_{i_0} - y_{i_0-1}) - \\ - (c_{i_0} - a_{i_0} - b_{i_0})y_{i_0} < 0, \end{aligned}$$

что противоречит предположению  $Ly_i \geq 0$  для всех  $i = 1, 2, \dots, N - 1$ , в том числе для  $i = i_0$ . Первое утверждение теоремы доказано. Второе утверждение доказывается аналогично (достаточно заменить  $y_i$  на  $-y_i$  и воспользоваться только что доказанным утверждением).

**Следствие 1.** Если выполнены условия (2),  $Ly_i \leq 0$  ( $i = 1, 2, \dots, N - 1$ ),  $y_0 \geq 0$ ,  $y_N \geq 0$ , то  $y_i \geq 0$  ( $i = 0, 1, \dots, N$ ).

*Если  $Ly_i \geq 0$ ,  $y_0 \leq 0$ ,  $y_N \leq 0$ , то  $y_i \leq 0$  ( $i = 0, 1, \dots, N$ ).*

**Доказательство.** Пусть  $Ly_i \leq 0$  и  $y_i < 0$  хотя бы в одном внутреннем узле  $i = i_*$ ; тогда  $y_i$  достигает наименьшего отрицательного значения во внутреннем узле, что невозможно в силу принципа максимума.

**Следствие 2.** Если  $\varphi_i \geq 0$ ,  $\mu_1 \geq 0$ ,  $\mu_2 \geq 0$ , то решение задачи (1)–(2) неотрицательно:  $y_i \geq 0$  ( $i = 0, 1, \dots, N$ ).

**Следствие 3.** Если выполнены условия (2), то задача

$$Ly_i = 0, \quad i = 1, 2, \dots, N - 1, \quad y_0 = 0, \quad y_N = 0 \quad (3)$$

имеет только тривиальное решение и задача (1), (2) однозначно разрешима при любых  $\varphi_i$ ,  $\mu_1$ ,  $\mu_2$ .

**Доказательство.** Предполагая, что решение  $y_i$  задачи (3) отлично от нуля хотя бы в одной точке  $i = i_*$ , мы приходим к противоречию с принципом максимума: если  $y_{i_*} > 0$  ( $y_{i_*} < 0$ ), то  $y_i$  достигает положительного наибольшего (отрицательного наименьшего) значения в некоторой внутренней точке  $i = i_0$ , что невозможно. Следовательно,  $y_i \equiv 0$ .

**Теорема 2 (теорема сравнения).** Пусть  $y_i$  – решение задачи (1), (2),  $\bar{y}_i$  – решение задачи

$$L\bar{y}_i = -\bar{\varphi}_i, \quad i = 1, 2, \dots, N - 1, \quad \bar{y}_0 = \bar{\mu}_1, \quad \bar{y}_N = \bar{\mu}_2.$$

и выполнены условия

$$|\varphi_i| \leq \bar{\varphi}_i, |\mu_1| \leq \bar{\mu}_1, |\mu_2| \leq \bar{\mu}_2.$$

Тогда справедлива оценка

$$|y_i| \leq \bar{y}_i, \text{ для всех } i = 0, 1, \dots, N.$$

**Доказательство.** В силу следствия 2 имеем  $\bar{y}_i \geq 0$ . Для разности  $\bar{y}_i - y_i$  и суммы  $\bar{y}_i + y_i$  получаем уравнение вида (1) с правыми частями  $\bar{\varphi}_i - \varphi_i \geq 0$ ,  $\bar{\mu}_1 - \mu_1 \geq 0$ ,  $\bar{\mu}_2 - \mu_2 \geq 0$  и  $\bar{\varphi}_i + \varphi_i \geq 0$ ,  $\bar{\mu}_1 + \mu_1 \geq 0$ ,  $\bar{\mu}_2 + \mu_2 \geq 0$  соответственно. Так как  $\varphi_i \pm \bar{\varphi}_i \geq 0$  и  $\mu_\alpha \pm \bar{\mu}_\alpha \geq 0$  ( $\alpha = 1, 2$ ), то в силу следствия 2  $\bar{y}_i - y_i \geq 0$ ,  $\bar{y}_i + y_i \geq 0$ , откуда следует  $-\bar{y}_i \leq y_i \leq \bar{y}_i$ ,  $|y_i| \leq \bar{y}_i$ , что и требовалось доказать.

Функцию  $\bar{y}_i$  называют *мажорантой* для решения задачи (1), (2).

**2. Оценка решения краевой задачи.** Решение краевой задачи (1), (2) представим в виде суммы  $y_i = y_i^{(1)} + y_i^{(2)}$ , где  $y_i^{(1)}$  — решение неоднородного уравнения с однородными краевыми условиями:

$$Ly_i = -\varphi_i, \quad i = 1, 2, \dots, N-1, \quad y_0 = y_N = 0, \quad (4)$$

а  $y_i^{(2)}$  — решение однородного уравнения с неоднородными краевыми условиями:

$$Ly_i = 0, \quad i = 1, 2, \dots, N-1, \quad y_0 = \mu_1, \quad y_N = \mu_2. \quad (5)$$

Докажем, что для  $y_i^{(2)}$  справедлива оценка

$$\max_{0 \leq i \leq N} |y_i^{(2)}| \leq \max(|\mu_1|, |\mu_2|). \quad (6)$$

Пусть  $\bar{y}_i$  — решение задачи

$$L\bar{y}_i = 0, \quad i = 1, 2, \dots, N-1,$$

$$\bar{y}_0 = \bar{y}_N = \bar{\mu}, \quad \bar{\mu} = \max(|\mu_1|, |\mu_2|).$$

Тогда по теореме сравнения  $|\bar{y}_i^{(2)}| \leq |\bar{y}_i|$ , а в силу принципа максимума  $\max_{0 \leq i \leq N} |\bar{y}_i| \leq \bar{\mu}$ , так как  $\bar{y}_i \geq 0$  может достигать наибольшего положительного значения только на границе, т. е. при  $i = 0$  или  $i = N$ .

Нетрудно доказать, что величина  $\max_{0 \leq i \leq N} |y_i|$  является нормой. Норму принято обозначать символом  $\|y\|_c$ . Таким образом, мы получили оценку  $\|y^{(2)}\|_c \leq \max(|\mu_1|, |\mu_2|)$ .

**Теорема 3.** Пусть выполнены условия  
 $|a_i| > 0, |b_i| > 0, \bar{d}_i = |c_i| - |a_i| - |b_i| > 0,$   
 $i = 1, 2, \dots, N - 1.$  (7)

Тогда для решения задачи (4) справедлива оценка

$$\|y\|_c \leq \|\varphi/\bar{d}\|_c. \quad (8)$$

**Доказательство.** Для доказательства перепишем (4) в виде

$$c_i y_i = a_i y_{i-1} + b_i y_{i+1} + \varphi_i. \quad (4')$$

Пусть  $|y_i|$  достигает наибольшего значения  $|y_{i_0}| > 0$  при  $i = i_0$  ( $0 < i_0 < N$ ), так что  $|y_i| \leq |y_{i_0}|$  при любом  $i = 0, 1, \dots, N$ . Тогда из (4') при  $i = i_0$  следует

$$\begin{aligned} |c_{i_0}| |y_{i_0}| &= |a_{i_0} y_{i_0-1} + b_{i_0} y_{i_0+1} + \varphi_{i_0}| \leq |a_{i_0}| |y_{i_0-1}| + \\ &+ |b_{i_0}| |y_{i_0+1}| + |\varphi_{i_0}| \leq (|a_{i_0}| + |b_{i_0}|) |y_{i_0}| + |\varphi_{i_0}|, \\ (|c_{i_0}| - |a_{i_0}| - |b_{i_0}|) |y_{i_0}| &\leq |\varphi_{i_0}|, \quad |y_{i_0}| \leq \frac{|\varphi_{i_0}|}{\bar{d}_{i_0}} \leq \left\| \frac{\varphi}{\bar{d}} \right\|_c. \end{aligned}$$

Тем самым оценка (8) доказана.

**Замечание.** Если условие  $d_i = c_i - a_i - b_i > 0$  или  $\bar{d}_i = |c_i| - |a_i| - |b_i| > 0$  не выполнено, например,

$$d_i = c_i - a_i - b_i \geq 0, \quad a_i > 0, \quad b_i > 0, \quad i = 1, 2, \dots, N - 1, \quad (9)$$

т. е.  $d_i$  может в некоторых узлах обращаться в нуль, то теоремой 3 пользоваться нельзя. В этом случае для оценки решения  $y_i$  задачи (4) можно поступать так. Представим  $y_i$  в виде суммы  $y_i = v_i + w_i$ , где  $w_i$  — решение задачи

$$\begin{aligned} \overset{\circ}{L}w_i &= b_i(w_{i+1} - w_i) - a_i(w_i - w_{i-1}) = -\varphi_i, \\ i &= 1, 2, \dots, N - 1, \quad w_0 = w_N = 0. \end{aligned} \quad (10)$$

Тогда  $v_i$  определяется из условий

$$\begin{aligned} Lv_i &= b_i(v_{i+1} - v_i) - a_i(v_i - v_{i-1}) - d_i v_i = -d_i w_i, \\ i &= 1, 2, \dots, N - 1, \quad v_0 = v_N = 0. \end{aligned} \quad (11)$$

В этом можно убедиться, складывая почленно уравнения (10) и (11). Функцию  $w_i$  можно оценить непосредственно

но (см. гл. IV, § 3), написав ее в явном виде, а для оценки  $v_i$  нам понадобится

**Теорема 4.** Для решения задачи (11) при условиях (9) справедлива оценка

$$\|v\|_c \leq \|w\|_c. \quad (12)$$

**Доказательство.** Если  $d_i \equiv 0$ , то в силу следствия 3  $v_i \equiv 0$  и оценка (12) выполнена. Пусть  $d_i \neq 0$  хотя бы в одной точке. Построим мажоранту  $\bar{v}_i$  как решение задачи

$$L\bar{v}_i = -d_i |w_i|, \quad i = 1, 2, \dots, N-1, \quad \bar{v}_0 = \bar{v}_N = 0.$$

Пусть  $\bar{v}_i \geq 0$  достигает наибольшего значения при  $i = i_0$ ; тогда  $\bar{v}_{i_0+1} - \bar{v}_{i_0} \leq 0$ ,  $\bar{v}_{i_0} - \bar{v}_{i_0-1} \geq 0$  и из (4) следует

$$\begin{aligned} d_{i_0} \bar{v}_{i_0} &\leq -b_{i_0} (\bar{v}_{i_0+1} - \bar{v}_{i_0}) + a_{i_0} (\bar{v}_{i_0} - \bar{v}_{i_0-1}) + d_{i_0} \bar{v}_{i_0} = \\ &= d_{i_0} |w_{i_0}|. \end{aligned}$$

Если  $d_{i_0} > 0$ , то  $\bar{v}_{i_0} < |w_{i_0}|$ , и мы сразу получаем оценку (12), так как  $|v_i| \leq \bar{v}_i$ . Если  $d_{i_0} = 0$ , то уравнение (11) принимает вид  $-b_{i_0} (\bar{v}_{i_0+1} - \bar{v}_{i_0}) + a_{i_0} (\bar{v}_{i_0} - \bar{v}_{i_0-1}) = 0$ , и из него следует, что  $\bar{v}_{i_0+1} = \bar{v}_{i_0-1} = \bar{v}_{i_0}$ . Так как  $\bar{v}_i \not\equiv \text{const}$ , то существует такая точка  $i = i_1$ , в которой  $\bar{v}_{i_1} = \bar{v}_{i_0}$ , а в соседней точке, например  $i = i_1 + 1$ ,  $\bar{v}_{i_1+1} < \bar{v}_{i_1}$ ; тогда здесь  $d_{i_1} \neq 0$ , и мы получим рассмотренный выше случай:  $\bar{v}_{i_1} = \|\bar{v}\|_c \leq |w_{i_0}| \leq \|w\|_c$ .

**3. Оценка решения разностного уравнения при помощи формул прогонки.** Для случая, когда  $b_i = a_{i+1}$ , т. е. когда оператор  $Ly_i$  является самосопряженным, можно оценить решение задачи (4) при помощи формул правой прогонки. Уравнение (4) нам удобно записать в форме

$$\begin{aligned} \Lambda y_i = (ay_x)_x, &i = 1, \dots, N-1, \quad y_0 = 0, \quad y_N = 0, \quad (13) \\ a_i &> 0, \quad d_i \geq 0. \end{aligned}$$

Перепишем его в обычном виде:

$$\begin{aligned} a_i y_{i-1} - c_i y_i + a_{i+1} y_{i+1} &= -h^2 \varphi_i, \quad y_0 = y_N = 0, \\ c_i &= a_i + a_{i+1} + h^2 d_i, \quad a_i > 0, \quad i = 1, 2, \dots, N-1. \end{aligned}$$

Рассмотрим формулы прогонки

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad y_N = 0, \quad i = 1, 2, \dots, N-1,$$

$$\alpha_{i+1} = \frac{a_{i+1}}{c_i - a_i \alpha_i}, \quad \alpha_1 = 0, \quad i = 1, 2, \dots, N-1,$$

$$\beta_{i+1} = \frac{a_i \beta_i + \varphi_i h^2}{c_i - a_i \alpha_i}, \quad \beta_1 = 0, \quad i = 1, 2, \dots, N-1.$$

При условиях (7) имеем  $|\alpha_{i+1}| \leq 1$  и

$$|y_i| \leq |y_{i+1}| + |\beta_{i+1}| \leq |y_N| + \sum_{s=i+1}^N |\beta_s| = \sum_{s=i+1}^N |\beta_s|.$$

Вводя функцию  $a_i \beta_i = \eta_i$ , получаем

$$\eta_{i+1} = (\eta_i + h^2 \varphi_i) \alpha_{i+1},$$

$$|\eta_{i+1}| \leq |\eta_i| + h^2 |\varphi_i| \leq |\eta_1| + \sum_{k=1}^i h^2 |\varphi_k|,$$

так что

$$|\beta_{s+1}| \leq \frac{1}{a_{s+1}} h \sum_{k=1}^s h |\varphi_k|.$$

В результате получаем для решения задачи априорную оценку

$$\|y\|_c \leq \sum_{s=1}^N h \frac{1}{a_s} \sum_{k=1}^s h |\varphi_k| \leq \frac{1}{c_1} \sum_{s=1}^N h \sum_{k=1}^s h |\varphi_k| \text{ при } a_s \geq c_1 > 0.$$

Этой оценкой мы воспользуемся при изучении сходимости разностных схем.

## Глава II

# ИНТЕРПОЛЯЦИЯ И ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

## § 1. Интерполяция и приближение функций

**1. Постановка задачи.** Одной из основных задач численного анализа является задача об интерполяции функций. Часто требуется восстановить функцию  $f(x)$  для всех значений  $x$  на отрезке  $a \leq x \leq b$ , если известны ее значения в некотором конечном числе точек этого отрезка. Эти значения могут быть найдены в результате наблюдений (измерений) в каком-то натуральном эксперименте, либо в результате вычислений. Кроме того, может оказаться, что функция  $f(x)$  задается формулой и вычисления ее значений по этой формуле очень трудоемки, поэтому желательно иметь для функции более простую (менее трудоемкую для вычислений) формулу, которая позволяла бы находить приближенное значение рассматриваемой функции с требуемой точностью в любой точке отрезка. В результате возникает следующая математическая задача.

Пусть на отрезке  $a \leq x \leq b$  задана сетка  $\bar{\omega} = \{x_0 = a < x_1 < \dots < x_n = b\}$  и в ее узлах заданы значения функции  $y(x)$ , равные  $y(x_0) = y_0, \dots, y(x_i) = y_i, \dots, y(x_n) = y_n$ . Требуется построить *интерполянту* — функцию  $f(x)$ , совпадающую с функцией  $y(x)$  в узлах сетки:

$$f(x_i) = y_i, \quad i = 0, 1, \dots, n. \quad (1)$$

Основная цель интерполяции — получить быстрый (экономичный) алгоритм вычисления значений  $f(x)$  для значений  $x$ , не содержащихся в таблице данных.

Основной вопрос: как выбрать интерполянту  $f(x)$  и как оценить погрешность  $y(x) - f(x)$ ? Интерполирующие функции  $f(x)$ , как правило, строятся в виде линейных комбинаций некоторых элементарных функций:

$$f(x) = \sum_{k=0}^n c_k \Phi_k(x),$$

где  $\{\Phi_k(x)\}$  — фиксированные линейно независимые функции,  $c_0, c_1, \dots, c_n$  — не определенные пока коэффициенты.

Из условий (1) получим систему  $n+1$  уравнений относительно коэффициентов  $\{c_k\}$ :

$$\sum_{k=0}^n c_k \Phi_k(x_i) = y_i, \quad i = 0, 1, \dots, n.$$

Предположим, что система функций  $\Phi_k(x)$  такова, что при любом выборе узлов  $a = x_0 < x_1 < \dots < x_n = b$  отличен от нуля определитель системы:

$$\Delta(\Phi) = \begin{vmatrix} \Phi_0(x_0) & \Phi_1(x_0) & \dots & \Phi_n(x_0) \\ \Phi_0(x_1) & \Phi_1(x_1) & \dots & \Phi_n(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_0(x_n) & \Phi_1(x_n) & \dots & \Phi_n(x_n) \end{vmatrix}.$$

Тогда по заданным  $y_i$  ( $i = 0, 1, \dots, n$ ) однозначно определяются коэффициенты  $c_k$  ( $k = 0, 1, \dots, n$ ).

В качестве системы линейно независимых функций  $\{\Phi_k(x)\}$  чаще всего выбирают: степенные функции  $\Phi_k(x) = x^k$  (в этом случае  $f = P_n(x)$  — полином степени  $n$ ); тригонометрические функции  $\{\Phi_k(x) = \cos kx, \sin kx\}$  ( $f$  — тригонометрический полином). Используются также рациональные функции

$$\frac{\alpha_0 + \alpha_1 x + \dots + \alpha_m x^m}{\beta_0 + \beta_1 x + \dots + \beta_p x^p}$$

и другие виды интерполирующих функций. Мы рассмотрим интерполяционные полиномы и сплайн-интерполяцию — случай кусочно-полиномиальной интерполяции.

**2. Полиномиальная интерполяция.** Известно, что любая непрерывная на отрезке  $[a, b]$  функция  $f(x)$  может быть хорошо приближена некоторым полиномом  $P_n(x)^*$ :

**Теорема Вейерштрасса.** Для любого  $\varepsilon > 0$  существует полином  $P_n(x)$  степени  $n = n(\varepsilon)$ , такой, что

$$\max_{x \in [a, b]} |f(x) - P_n(x)| < \varepsilon.$$

Однако эта теорема не дает ответа на вопрос о существовании хорошего интерполяционного полинома для заданного множества точек  $\{(x_i, y_i)\}$ .

Итак, будем искать *интерполяционный полином* в виде

$$P_n(x) = \sum_{k=0}^n c_k x^k, \quad (2)$$

---

\* См. например, Ильин В. А., Позняк Э. Г., Основы математического анализа, ч. II.— М.: Наука, 1980, с. 50.

где  $c_k$  — неопределенные коэффициенты. Полагая  $f(x_i) = y_i$ , получаем систему линейных уравнений

$$c_0 + c_1 x_0 + \dots + c_n x_0^n = y_0,$$

$$c_0 + c_1 x_1 + \dots + c_n x_1^n = y_1,$$

$$\dots \dots \dots \dots \dots \dots$$

$$c_0 + c_1 x_n + \dots + c_n x_n^n = y_n.$$

Определителем этой системы является отличный от нуля определитель Вандермонда:

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} = \prod_{n>k>m>0} (x_k - x_m) \neq 0.$$

Отсюда следует, что интерполяционный полином (2) существует и единствен (форм записи его существует много).

В качестве базиса  $\{\Phi_k(x)\}$  мы взяли базис из одночленов 1,  $x$ ,  $x^2$ , ...,  $x^n$ . Для вычислений более удобным является базис полиномов Лагранжа  $\{l_k(x)\}$  степени  $n$  или коэффициентов Лагранжа:

$$l_k(x_i) = \begin{cases} 1, & \text{если } i = k, \\ 0, & \text{если } i \neq k, \quad i, k = 0, 1, \dots, n. \end{cases}$$

Нетрудно видеть, что полином степени  $n$

$$\begin{aligned} l_k(x) &= l_k^{(n)}(x) = \\ &= \frac{(x - x_0)(x - x_1) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_n)}{(x_k - x_0)(x_k - x_1) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_n)} \end{aligned}$$

удовлетворяет этим условиям. Полином  $l_k(x)$ , очевидно, определяется единственным образом. В самом деле, пусть существует еще один полином  $\bar{l}_k(x)$ ; тогда их разность  $\bar{l}_k(x) - l_k(x) = q_n(x)$  есть полином степени  $n$ , обращающийся в нуль в  $n+1$  точках  $x_i$  ( $i = 0, 1, \dots, n$ ). Это возможно только при  $\bar{l}_k(x) - l_k(x) = 0$ .

Полином  $l_k(x)y_k$  принимает значение  $y_k$  в точке  $x_k$  и равен нулю во всех остальных узлах  $x_j$  при  $j \neq k$ . Отсюда следует, что интерполяционный полином

$$P_n(x) = \sum_{k=0}^n l_k(x) y_k = \sum_{k=0}^n y_k \prod_{i \neq k} \frac{x - x_i}{x_k - x_i} \quad (3)$$

имеет степень не выше  $n$  и  $P_n(x_i) = y_i$ . Формулу (3) называют *формулой Лагранжа*. Число арифметических действий для вычисления по (3) пропорционально  $n^2$ . Для оценки близости полинома  $P_n(x)$  к функции  $f(x)$  предполагают, что существует  $n+1$ -я непрерывная производная  $f^{(n+1)}(x)$ . Тогда имеет место формула для погрешности

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{j=1}^{n+1} (x - x_j), \quad \xi \in [a, b].$$

**3. Интерполяционная формула Ньютона.** При вычислениях на ЭВМ удобна *интерполяционная формула Ньютона*. Для ее записи надо ввести так называемые разделенные разности:

разделенная разность первого порядка:  $y(x_i, x_j) = [y(x_i) - y(x_j)]/(x_i - x_j)$ ;

разделенная разность второго порядка:  $y(x_i, x_j, x_k) = [y(x_i, x_j) - y(x_j, x_k)]/(x_i - x_k)$  и т. д. Если  $y(x) = P_n(x)$  — полином степени  $n$ , то для него первая разделенная разность  $P(x, x_0) = [P(x) - P(x_0)]/(x - x_0)$  есть полином степени  $n-1$ , вторая разность  $P(x, x_0, x_1)$  — полином степени  $n-2$  и т. д., так что  $(n+1)$ -я разделенная разность равна нулю.

Из определения разделенных разностей следует:

$$\begin{aligned} P(x) &= P(x_0) + (x - x_0)P(x, x_0), \\ P(x, x_0) &= P(x_0, x_1) + (x - x_1)P(x, x_0, x_1), \\ P(x, x_0, x_1) &= P(x_0, x_1, x_2) + (x - x_2)P(x, x_0, x_1, x_2) \end{aligned}$$

и т. д. Отсюда получим для  $P(x)$  формулу

$$\begin{aligned} P(x) &= P(x_0) + \\ &+ (x - x_0)P(x_0, x_1) + (x - x_0)(x - x_1)P(x_0, x_1, x_2) + \dots \\ &\dots + (x - x_0)(x - x_1) \dots (x - x_n)P(x_0, x_1, \dots, x_n). \quad (4) \end{aligned}$$

Если  $P(x)$  — интерполяционный полином для функции  $y(x)$ , то его значения в узлах сеток совпадают со значениями функции  $y(x)$ , а значит, совпадают и разделенные разности. Поэтому вместо (4) можно написать

$$\begin{aligned} f(x) &= y_0 + \sum_{k=1}^n (x - x_0)(x - x_1) \dots \\ &\dots (x - x_{k-1}) y(x_0, x_1, \dots, x_k) \end{aligned}$$

(*полином Ньютона*). После того как вычислены разделен-

ные разности, вычислять полином Ньютона удобно по схеме Горнера:

$$f(x) = y(x_0) + (x - x_0)[y(x_0, x_1) + (x - x_1)[y(x_0, x_1, x_2) + \dots]].$$

Вычисление  $f(x)$  для каждого  $x$  требует  $n$  умножений и  $2n$  сложений или вычитаний.

Существуют и другие формулы интерполяции. Среди них наиболее употребительна *эрмитова интерполяция*. Здесь задача ставится так. Заданы  $n$  узлов  $\{x_i\}$ ,  $n$  значений функции  $\{y_i\}$  и  $n$  значений производной  $\{y'_i\}$ ; требуется найти полином максимальной степени  $2n - 1$ , такой, что

$$P(x_i) = y_i, \quad P'(x_i) = y'_i, \quad i = 1, 2, \dots, n.$$

Если все  $x_i$  различны, то существует единственное решение, которое находится способом, аналогичным методу Лагранжа.

Следует иметь в виду, что применение полинома высокой степени может приводить к трудным проблемам, связанным с ошибками округления.

**4. Сплайн-интерполяция.** Рассмотрим специальный случай кусочно-полиномиальной интерполяции, когда между любыми соседними узлами сетки функция интерполируется кубическим полиномом (*кубическая сплайн-интерполяция*). Его коэффициенты на каждом интервале определяются из условий сопряжения в узлах:

$$\begin{aligned} f_i &= y_i, \\ f'(x_i - 0) &= f'(x_i + 0), \\ f''(x_i - 0) &= f''(x_i + 0), \quad i = 1, 2, \dots, n - 1. \end{aligned}$$

Кроме того, на границе при  $x = x_0$  и  $x = x_n$  ставятся условия

$$f''(x_0) = 0, \quad f''(x_n) = 0. \quad (5)$$

Будем искать кубический полином в виде

$$\begin{aligned} f(x) &= a_i + b_i(x - x_{i-1}) + c_i(x - x_{i-1})^2 + d_i(x - x_{i-1})^3, \\ &\qquad\qquad\qquad x_{i-1} \leq x \leq x_i. \end{aligned} \quad (6)$$

Из условия  $f_i = y_i$  имеем

$$\begin{aligned} f(x_{i-1}) &= a_i = y_{i-1}, \\ f(x_i) &= a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = y_i, \\ h_i &= x_i - x_{i-1}, \quad i = 1, 2, \dots, n - 1. \end{aligned} \quad (7)$$

Вычислим производные:

$$\begin{aligned} f'(x) &= b_i + 2c_i(x - x_{i-1}) + 3d_i(x - x_{i-1})^2, \\ f''(x) &= 2c_i + 6d_i(x - x_{i-1}), \quad x_{i-1} \leq x \leq x_i, \end{aligned}$$

и потребуем их непрерывности при  $x = x_i$ :

$$\begin{aligned} b_{i+1} &= b_i + 2c_i h_i + 3d_i h_i^2, \\ c_{i+1} &= c_i + 3d_i h_i, \quad i = 1, 2, \dots, n-1. \end{aligned} \tag{8}$$

Общее число неизвестных коэффициентов, очевидно, равно  $4n$ , число уравнений (7) и (8) равно  $4n-2$ . Недостающие два уравнения получаем из условий (5) при  $x = x_0$  и  $x = x_n$ :

$$c_1 = 0, \quad c_n + 3d_n h_n = 0.$$

Выражая из (8)  $d_i = (c_{i+1} - c_i)/3h_i$ , подставляя это выражение в (7) и исключая  $a_i = y_{i-1}$ , получим

$$\begin{aligned} b_i &= [(y_i - y_{i-1})/h_i] - \frac{1}{3} h_i (c_{i+1} + 2c_i), \quad i = 1, 2, \dots, n-1, \\ b_n &= [(y_n - y_{n-1})/h_n] - \frac{2}{3} h_n c_n. \end{aligned}$$

Подставив теперь выражения для  $b_i$ ,  $b_{i+1}$  и  $d_i$  в первую формулу (8), после несложных преобразований получаем для определения  $c_i$  разностное уравнение второго порядка

$$h_i c_i + 2(h_i + h_{i+1}) c_{i+1} + h_{i+1} c_{i+2} = 3 \left( \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right), \quad i = 1, 2, \dots, n-1, \tag{9}$$

с краевыми условиями

$$c_1 = 0, \quad c_{n+1} = 0. \tag{10}$$

Условие  $c_{n+1} = 0$  эквивалентно условию  $c_n + 3d_n h_n = 0$  и уравнению  $c_{i+1} = c_i + d_i h_i$ . Разностное уравнение (9) с условиями (10) решается методом прогонки.

Можно ввести понятие *сплайна порядка  $m$*  как функции, которая является полиномом степени  $m$  на каждом из отрезков сетки и во всех внутренних узлах сетки удовлетворяет условиям непрерывности функции и производных до порядка  $m-1$  включительно. Обычно для интерполяции используются случаи  $m=3$  (рассмотренный выше кубический сплайн) и  $m=1$  (*линейный сплайн*, соответствующий аппроксимации графика функций  $y(x)$  ломаной, проходящей через точки  $(x_i, y_i)$ ).

**5. Применение интерполяции.** Интерполяция применяется во многих задачах, связанных с вычислениями. Укажем некоторые из этих задач.

Обработка физического эксперимента — построение приближенных формул для характерных величин по табличным данным, полученным экспериментально.

Построение приближенных формул по данным вычислительного эксперимента. Здесь возникают нестандартные задачи интерполяции, так как обычно пишутся формулы возможно более простой структуры.

Субтабулирование, т. е. сгущение таблиц; применяется в тех случаях, когда непосредственное вычисление функций трудно или когда имеется мало экспериментальных данных. В машину вводится небольшая таблица, а нужные при расчетах значения функции находятся по мере необходимости по интерполяционной формуле.

Интерполяция применяется также в задаче *обратного интерполирования*: задана таблица  $y_i = y(x_i)$ ; найти  $x_i$  как функцию от  $y_i$ . Примером обратного интерполирования может служить задача о нахождении корней уравнения.

Интерполяционные формулы используются также при вычислении интегралов, при писании разностных аппроксимаций для дифференциальных уравнений на основе интегральных тождеств.

Математическое обеспечение любой ЭВМ содержит стандартные программы интерполяции.

**6. Среднеквадратичная аппроксимация.** До сих пор мы рассматривали построение интерполяционных полиномов  $y(x)$ , совпадающих со значениями исходной функции  $f(x)$  на некотором множестве узлов на сетке  $\omega$ :

$$y(x_i) = f(x_i), \quad x_i \in \omega.$$

Функция  $y(x)$  приближает (аппроксимирует) функцию  $f(x)$  на интервале сетки.

Пусть  $L_2[a, b]$  — пространство вещественных функций со скалярным произведением

$$(f, \varphi) = \int_a^b f(x) \varphi(x) dx$$

и нормой

$$\|f\|_{L_2} = \sqrt{(f, f)}.$$

Рассмотрим общую задачу об аппроксимации функции  $f(x)$  функциями  $L_2$ , заменяя требование  $y_i = f_i$  условием минимума нормы:  $\|f - y\|_{L_2}$  или малости нормы:  $\|f - y\|_{L_2} < \epsilon$ , где  $\epsilon > 0$  — заданная точность.

Отыскание  $\inf \|f - y\|_{L_2}$  есть задача о нахождении *наилучшего среднеквадратичного приближения*. В качестве  $y(x)$  возьмем *обобщенный полином*

$$y(x) = \sum_{k=0}^n c_k \varphi_k(x),$$

где  $\{\varphi_k(x)\}$  — семейство ортонормированных на  $[a, b]$  функций

$$(\varphi_k, \varphi_m) = \delta_{km}, \quad \delta_{km} = \begin{cases} 1, & k = m, \\ 0, & k \neq m, \end{cases}$$

$c_k$  — произвольные коэффициенты. Тогда задача нахождения наилучшего среднеквадратичного приближения сводится к отысканию минимума функции  $n+1$  переменных  $c_0, c_1, \dots, c_n$ :

$$\min_{\{c_k\}} \left\| f(x) - \sum_{k=0}^n c_k \varphi_k(x) \right\|^2 = F(c_0, c_1, \dots, c_n).$$

Вычислим *среднеквадратичное уклонение*

$$\|f - y\|^2 = \|f\|^2 - 2(f, y) + \|y\|^2.$$

Подставляя сюда выражения

$$(f, y) = \sum_{k=0}^n c_k (f, \varphi_k) = \sum_{k=0}^n c_k f_k, \quad f_k = (f, \varphi_k),$$

$$\|y\|^2 = \sum_{k=0}^n c_k^2,$$

получим

$$\|f - y\|^2 = \|f\|^2 + \sum_{k=0}^n (c_k - f_k)^2 - \sum_{k=0}^n f_k^2.$$

Отсюда видно, что минимум погрешности достигается при  $c_k = f_k$ , т. е. на функции

$$\bar{y}(x) = \bar{y}_n(x) = \sum_{k=0}^n f_k \varphi_k(x).$$

В этом случае

$$\|f - \bar{y}_n(x)\|^2 = \|f\|^2 - \sum_{k=0}^n f_k^2. \quad (11)$$

Таким образом, наилучшее среднеквадратичное приближение существует и единственно. Оно приводит к задаче о вычислении интегралов для определения  $c_k = f_k = (f, \varphi_k)$ .

Если функции  $\{\varphi_k\}$  образуют полную ортонормированную систему, то выполняется *равенство Парсеваля — Стеклова*

$$\sum_{k=0}^{\infty} f_k^2 = \int_a^b f^2(x) dx = \|f\|^2. \quad (12)$$

Сравнивая (11) с (12), находим

$$\|f - \bar{y}_n\|^2 = \sum_{k=n+1}^{\infty} f_k^2,$$

т. е.  $\|f - \bar{y}_n\| \rightarrow 0$  при  $n \rightarrow \infty$ ; наилучшее среднеквадратичное приближение сходится к  $f(x)$  и возможна аппроксимация с любой точностью:  $\|f - \bar{y}_n\| \leq \varepsilon$ , если  $n \geq N(\varepsilon)$  ( $n$  достаточно велико).

**Замечание.** Все рассуждения сохраняют силу, если скалярное произведение берется с весом  $\rho(x) > 0$ :

$$(f, \varphi) = \int_a^b f(x) \varphi(x) \rho(x) dx.$$

Возможны и другие критерии аппроксимации, когда уклонение  $f - y$  минимизируется в другой норме, например, в норме пространства  $C$  (*равномерное приближение*). При наилучшем равномерном приближении мы отыскиваем функцию  $y(x)$ , на которой достигается

$$\min_{\{y\}} \max_{a \leq x \leq b} |f(x) - y(x)|.$$

Однако пока не найдено метода нахождения за конечное число действий коэффициентов наилучшего равномерного приближения для функции, заданной на отрезке  $[a, b]$ . Возможны и другие постановки задач аппроксимации — на дискретном множестве, на совокупности отрезков и др. Изучаются также методы нелинейной аппроксимации, например, при помощи рациональных функций. Это оказывается эффективным при обработке экспериментов.

## § 2. Численное интегрирование

**1. Постановка задачи. Простейшие квадратурные формулы.** Задача численного интегрирования состоит в нахождении приближенного значения интеграла

$$J[f] = \int_a^b f(x) dx, \quad (1)$$

где  $f(x)$  — заданная функция. На отрезке  $[a, b]$  вводится сетка  $\omega = \{x_i: x_0 = a < x_1 < \dots < x_i < x_{i+1} < \dots < x_N = b\}$  и в качестве приближенного значения интеграла рассматривается число

$$J_N[f] = \sum_{i=0}^N c_i f(x_i), \quad (2)$$

где  $f(x_i)$  — значения функции  $f(x)$  в узлах  $x = x_i$ ,  $c_i$  — весовые множители (веса), зависящие только от узлов, но не зависящие от выбора  $f(x)$ . Формула (2) называется *квадратурной формулой*.

Задача численного интегрирования при помощи квадратур состоит в отыскании таких узлов  $\{x_i\}$  и таких весов  $\{c_i\}$ , чтобы *погрешность квадратурной формулы*

$$D[f] = \sum_{i=0}^N c_i f(x_i) - \int_a^b f(x) dx = J_N[f] - J[f]$$

была минимальной для функций из заданного класса (величина  $D[f]$  зависит от гладкости  $f(x)$ ). При построении квадратурной формулы обычно представляют интеграл (1) в виде суммы интегралов вида

$$\int_{\alpha}^{\beta} f(x) dx,$$

каждый из которых сводится к стандартному интегралу по отрезку единичной длины:

$$L[f] = \int_0^1 f(s) ds \quad (3)$$

с помощью замены

$$x = \alpha + (\beta - \alpha)s, \quad (4)$$

$$f(x) = f(\alpha + (\beta - \alpha)s) = \bar{f}(s), \quad (5)$$

так что

$$\int_{\alpha}^{\beta} f(x) dx = \kappa \int_0^1 \tilde{f}(s) ds = \kappa L[\tilde{f}], \quad \kappa = \beta - \alpha$$

(черту сверху над  $f(s)$  будем опускать). Будем считать, что  $\omega$  — равномерная сетка. Тогда можно написать

$$J[f] = \sum_{i=1}^N J_i,$$

$$J_i = \int_{x_{i-1}}^{x_i} f(x) dx = h \int_0^1 f(x_{i-1} + hs) ds.$$

Если  $N = 2i_0$  — четное число, то

$$J[f] = \sum_{i=1}^{i_0} J_{2i-1},$$

$$J_{2i-1} = \int_{x_{2i-2}}^{x_{2i}} f(x) dx = 2h \int_0^1 f(x_{2i-2} + 2hs) ds,$$

и т. д.

Таким образом, задача сводится к построению квадратурной формулы для интеграла (3) по единичному отрезку. Выберем на отрезке  $0 \leq s \leq 1$  узлы  $0 \leq s_0 < s_1 < \dots < s_m \leq 1$  (шаблон квадратурной формулы) и поставим интегралу (3) в соответствие формулу

$$\Lambda(f) = \sum_{k=0}^m p_k f(s_k). \quad (6)$$

Рассмотрим простейшие квадратурные формулы:  
формула прямоугольника (шаблон содержит один узел):

$$m = 0, \quad p_0 = 1, \quad s_0 = \frac{1}{2}, \quad \Lambda(f) = f\left(\frac{1}{2}\right);$$

формула трапеции (два узла):

$$m = 1, \quad p_0 = \frac{1}{2}, \quad p_1 = \frac{1}{2}, \quad s_0 = 0, \quad s_1 = 1,$$

$$\Lambda(f) = \frac{1}{2} (f(0) + f(1));$$

формула Симпсона (три узла):

$$m = 2, \quad p_0 = p_2 = \frac{1}{6}, \quad p_1 = \frac{4}{6}, \quad s_0 = 0, \quad s_1 = \frac{1}{2}, \quad s_2 = 1,$$

$$\Lambda(f) = \frac{1}{6} \left( f(0) + 4f\left(\frac{1}{2}\right) + f(1) \right)$$

и другие. На практике, как правило, применяются формулы с небольшим числом узлов шаблона.

Напишем теперь соответствующие формулы для интеграла (1) на равномерной сетке  $\{x_i = ih\}$  с шагом  $h$ . Учитывая замену (4) и (5), получим:

формулу прямоугольника:

$$J_N[f] = \sum_{i=0}^{N-1} h f(x_{i+1/2}), \quad x_{i+1/2} = x_i + \frac{1}{2} h; \quad (7)$$

формулу трапеций:

$$J_N[f] = \sum_{i=0}^N c_i f(x_i) h, \quad c_0 = c_N = \frac{1}{2}, \quad c_i = 1, \quad i = 1, 2, \dots, N-1; \quad (8)$$

формулу Симпсона:

$$J_N[f] = \sum_{i=0}^N c_i f(x_i) \bar{h} = \frac{\bar{h}}{3} (f_0 + 4f_1 + 2f_2 + 4f_3 + \dots + 2f_{N-2} + 4f_{N-1} + f_N) \text{ при } N = 2i_0. \quad (9)$$

**2. Построение квадратурных формул.** В силу сказанного выше изложение достаточно вести для стандартного интеграла (3), которому ставится в соответствие квадратурная формула

$$\int_0^1 f(s) ds \approx \sum_{k=0}^m p_k f(s_k). \quad (10)$$

В общем случае узлы и веса неизвестны и подлежат определению.

Рассмотрим сначала случай, когда узлы заданы и требуется найти веса квадратурной формулы  $\{p_k\}$ . Мы будем пользоваться требованием: формула (10) должна быть точной для любого полинома  $P_r(s)$  степени  $r \leq m$ :

$$\Lambda[P_r] = L[P_r], \quad r \leq m. \quad (11)$$

Для того чтобы полином степени  $r$  удовлетворял (11), достаточно потребовать, чтобы квадратурная формула была точной для любого одночлена  $s^\sigma$  степени  $\sigma$  ( $\sigma = 0, 1, \dots, r$ ). Учитывая, что  $L[s^\sigma] = 1/(\sigma + 1)$ , получаем из (11)  $m + 1$  уравнений

$$\begin{aligned} p_0 + p_1 + \dots + p_m &= 1, \\ p_0 s_0 + p_1 s_1 + \dots + p_m s_m &= 1/2, \\ \vdots &\quad \vdots \\ p_0 s_0^\sigma + p_1 s_1^\sigma + \dots + p_m s_m^\sigma &= 1/(\sigma + 1), \\ \vdots &\quad \vdots \\ p_0 s_0^m + p_1 s_1^m + \dots + p_m s_m^m &= 1/(m + 1). \end{aligned}$$

Эта система имеет единственное решение, так как ее определителем является определитель Вандермонда, отличный от нуля, если нет совпадающих узлов,  $s_0 < s_1 < \dots < s_m$ .

Так, полагая  $m = 2$ ,  $s_0 = 0$ ,  $s_1 = 1/2$ ,  $s_2 = 1$ , имеем систему  $p_0 + p_1 + p_2 = 1$ ,  $p_1/2 + p_2 = 1/2$ ,  $p_1/4 + p_2 = 1/3$ , решением которой являются веса формулы Симпсона:  $p_0 = p_2 = 1/6$ ,  $p_1 = 4/6$ . Таким образом, формула Симпсона является точной для полинома второй степени. Однако, в силу симметрии, она является точной и для всех полиномов третьей степени:

$$P_3(s) = 1 + \alpha_1(s - 1/2) + \alpha_2(s - 1/2)^2 + \alpha_3(s - 1/2)^3,$$

так как она точна для  $f(s) = (s - 1/2)^3$ . В самом деле,

$$\begin{aligned} \Lambda\left[\left(s - \frac{1}{2}\right)^3\right] &= \frac{1}{6}\left(\left(-\frac{1}{2}\right)^3 + 4 \cdot 0 + \left(\frac{1}{2}\right)^3\right) = 0, \\ L\left[\left(s - \frac{1}{2}\right)^3\right] &= \int_0^1 \left(s - \frac{1}{2}\right)^3 ds = 0. \end{aligned}$$

Формулы прямоугольника и трапеции точны для линейной функции, т. е. для полинома первой степени, в чем легко убедиться непосредственно.

В общем случае в качестве  $P_m(s)$  можно выбрать интерполяционный полином Лагранжа

$$P_m(s) = \sum_{k=0}^m l_k^{(m)}(s) f(s_k),$$

где  $l_k^{(m)}(s)$  — интерполяционный коэффициент Лагранжа.

Из равенства

$$L[P_m] = \int_0^1 P_m(s) ds = \sum_{k=0}^m f(s_k) \int_0^1 l_k^{(m)}(s) ds = \sum_{k=0}^m p_k f(s_k)$$

видно, что формула (10) точна для полинома степени  $m$ , если весовые множители  $p_k$  определяются по формуле

$$p_k = \int_0^1 l_k^{(m)}(s) ds. \quad (12)$$

Формулы такого типа называют квадратурными *формулами Котеса*.

Приведем в качестве примеров подобных формул еще две квадратурные формулы:

на четырехточечном шаблоне,  $s_k = k/3$  ( $k = 0, 1, 2, 3$ ),  $m = 3$ :

$$\Lambda(f) = \frac{1}{8} \left( f(0) + 3f\left(\frac{1}{3}\right) + 3f\left(\frac{2}{3}\right) + f(1) \right),$$

$$p_0 = p_3 = \frac{1}{8}, \quad p_1 = p_2 = \frac{3}{8},$$

на пятиточечном шаблоне,  $s_k = k/4$  ( $k = 0, 1, 2, 3, 4$ ),  $m = 4$ :

$$\Lambda(f) = \frac{1}{90} \left( 7f(0) + 32f\left(\frac{1}{4}\right) + 42f\left(\frac{1}{2}\right) + 32f\left(\frac{3}{4}\right) + 7f(1) \right),$$

$$p_0 = p_4 = \frac{7}{90}, \quad p_1 = p_3 = \frac{32}{90}, \quad p_2 = \frac{42}{90}.$$

У всех приведенных выше пяти квадратурных формул шаблоны состоят из узлов, симметричных относительно середины  $s = 1/2$  отрезка  $0 \leq s \leq 1$ .

**3. Формула Тейлора с остаточным членом в интегральной форме.** При исследовании погрешности квадратурной формулы нам понадобится формула Тейлора с остаточным членом в интегральной форме:

$$f(s) = f(0) + sf'(0) + \frac{s^2}{2} f''(0) + \dots + \frac{s^n}{n!} f^{(n)}(0) + R_{n+1}(s), \quad (13)$$

$$R_{n+1}(s) = \int_0^s \frac{(s-t)^n}{n!} f^{(n+1)}(t) dt.$$

Эта формула может быть доказана индукцией по  $n$ . Для  $n = 0$  она верна:

$$f(s) = f(0) + R_1(s), \quad R_1(s) = \int_0^s f'(t) dt.$$

Допустим, что она верна для  $n$ . Интегрированием по частям получаем соотношение

$$\begin{aligned} & \int_0^s -\frac{(s-t)^n}{n!} f^{(n+1)}(t) dt = \\ & = -\frac{(s-t)^{n+1}}{(n+1)!} f^{(n+1)}(t) \Big|_0^s + \int_0^s \frac{(s-t)^{n+1}}{(n+1)!} f^{(n+2)}(t) dt = \\ & = \frac{s^{n+1}}{(n+1)!} f^{(n+1)}(0) + \int_0^s \frac{(s-t)^{n+1}}{(n+1)!} f^{(n+2)}(t) dt, \end{aligned} \quad (14)$$

которое и доказывает формулу (13) для  $n+1$ . Вводя функцию

$$K_n(\xi) = \begin{cases} \xi^n/n! & \text{при } \xi \geq 0, \\ 0 & \text{при } \xi < 0, \end{cases} \quad (15)$$

запишем формулу для остаточного члена  $R_{n+1}$  в виде:

$$R_{n+1}(s) = \int_0^1 K_n(s-t) f^{(n+1)}(t) dt. \quad (16)$$

**4. Формула для погрешности квадратурной формулы.** Переидем к выводу формулы для погрешности квадратурной формулы

$$\Delta(f) = \Lambda[f] - L[f] \quad (17)$$

в классе  $C^{(n+1)}$  функций, имеющих  $(n+1)$ -ю непрерывную производную на отрезке  $0 \leq s \leq 1$ :  $f(s) \in C^{(n+1)}[0, 1]$ . Тогда верна формула (13), или

$$f(s) = P_n(s) + R_{n+1}(s), \quad P_n(s) = \sum_{\sigma=0}^n \frac{s^\sigma}{\sigma!} f^{(\sigma)}(0). \quad (18)$$

Из предыдущего (см. п. 2) ясно, что для полинома  $P_n(s)$  степени  $n$  формула (10) является точной в двух случаях: при  $n \leq m+1 = n_0$ , если  $m$  четно и формула

симметрична; при  $n \leq m = n_0$  во всех других случаях. Мы будем пока предполагать, что

$$\Lambda[P_n] = L[P_n], \quad \text{т. е.} \quad n \leq n_0. \quad (19)$$

Обратимся теперь к разности  $\Delta(f)$  и подставим  $f = P_n + R_{n+1}$  в (17). Учитывая (16) и (19), получим

$$\begin{aligned} \Delta(f) &= \Lambda[f] - L[f] = \\ &= (\Lambda[P_n] - L[P_n]) + (\Lambda[R_{n+1}] - L[R_{n+1}]) = \\ &= \Lambda[R_{n+1}] - L[R_{n+1}] = \sum_{k=0}^m p_k \int_0^1 K_n(s_k - t) f^{(n+1)}(t) dt - \\ &\quad - \int_0^1 \int_0^1 K_n(s - t) f^{(n+1)}(t) dt ds = \\ &= \int_0^1 \left[ \sum_{k=0}^m p_k K_n(s_k - t) - \int_0^1 K_n(s - t) ds \right] f^{(n+1)}(t) dt. \end{aligned}$$

Пользуясь выражением (15) для  $K_n(s - t)$ , находим

$$\int_0^1 K_n(s - t) ds = \int_t^1 \frac{(s - t)^n}{n!} ds = \frac{(1 - t)^{n+1}}{(n + 1)!}.$$

В результате формула для погрешности принимает вид

$$\Delta(f) = \int_0^1 F_{n+1}(t) f^{(n+1)}(t) dt, \quad (20)$$

где

$$F_{n+1}(t) = \sum_{k=0}^m p_k K_n(s_k - t) - \frac{(1 - t)^{n+1}}{(n + 1)!}. \quad (21)$$

Отсюда следует оценка для погрешности

$$|\Delta(f)| \leq M_{n+1} c_{n+1} \quad (22)$$

при  $|f^{(n+1)}(t)| \leq M_{n+1}$ , где  $M_{n+1} > 0$  — постоянная, и при

$$c_{n+1} = \int_0^1 |F_{n+1}(t)| dt.$$

Если  $F_{n+1}(t)$  не меняет знака на отрезке  $0 \leq s \leq 1$ , то в

силу теоремы о среднем имеем

$$\Delta(f) = f^{(n+1)}(\xi) \int_0^1 F_{n+1}(t) dt, \quad \xi \in [0, 1].$$

**5. Оценка погрешности конкретных формул.** Мы получим оценку погрешности  $\Delta(\tilde{f}) = \Lambda(\tilde{f}) - L(\tilde{f})$  квадратурной формулы для стандартного интеграла (3). При переходе к формулам для интегралов (1) и (3) надо учесть, что

$$\frac{d^\sigma \tilde{f}(s)}{ds^\sigma} = \kappa^\sigma \frac{d^\sigma f(x)}{dx^\sigma},$$

$$\tilde{f}(s) = f(x), \quad x = \alpha + (\beta - \alpha)s, \quad dx = \kappa ds, \quad \kappa = \beta - \alpha.$$

Поэтому для погрешности

$$d[f] = \sum_{k=0}^m \kappa p_k f(x_k) - \int_{\alpha}^{\beta} f(x) dx = \kappa \Delta(\tilde{f})$$

в силу (22) верна формула

$$|d[f]| \leq c_{n+1} \kappa^{n+2} \max_{x \in [\alpha, \beta]} |f^{(n+1)}(x)|,$$

$$c_{n+1} = \int_0^1 |F_{n+1}(t)| dt.$$

Для вычисления погрешности  $J_N[f] - J[f]$ , очевидно, надо просуммировать на сетке погрешности  $|D[f]|$ .

Рассмотрим простейшие квадратурные формулы.

1) Формула прямоугольника:  $m = 0$ ,  $p_0 = 1$ ,  $s_0 = 1/2$ ,  $\Lambda(\tilde{f}) = \tilde{f}(1/2)$ . В силу формулы (20) имеем

$$\Delta_1(\tilde{f}) = \int_0^1 F_2(t) \tilde{f}''(t) dt, \quad F_2(t) = K_1 \left( \frac{1}{2} - t \right) - \frac{(1-t)^2}{2},$$

т. е.  $F_2(t) = -(1-t)^2/2 < 0$  при  $t > 1/2$ ,  $F_2(t) = (1/2 - t) - (1-t)^2/2 = -t^2/2 < 0$  при  $t < 0$ , т. е.  $F(t) < 0$  — знакопостоянная функция и

$$\Delta_1(\tilde{f}) = \tilde{f}''(\eta) \int_0^1 F_2(t) dt = -\frac{\tilde{f}''(\eta)}{24}, \quad \eta \in (0, 1),$$

Отсюда следует, что

$$d_i[f] = hf(x_{i-1/2}) - \int_{x_{i-1}}^{x_i} f(x) dx = -\frac{h^3}{24} f''(\xi_i),$$

$$\xi_i \in [x_{i-1}, x_i]. \quad (23)$$

Суммируя по  $i = 1, 2, \dots, N$  и учитывая, что среднее арифметическое равно

$$\sum_{i=1}^N hf''(\xi_i) = \frac{b-a}{N} \sum_{i=1}^N f''(\xi_i) = f''(\xi^*) (b-a), \quad \xi^* \in [a, b],$$

получаем для погрешности формулу прямоугольника:

$$D_N(f) = -\frac{h^2}{24} f''(\xi^*)(b-a).$$

Если  $f(x)$  имеет непрерывные производные по крайней мере четвертого порядка,  $f(x) \in C^{(n)}$  ( $n \geq 4$ ), то можно написать асимптотическое разложение для погрешности:

$$D_N(f) = \alpha_2 h^2 + \alpha_4 h^4, \quad (24)$$

где

$$\alpha_2 = -\frac{1}{24} \int_a^b f''(x) dx = -\frac{1}{24} [f'(b) - f'(a)].$$

В самом деле, подставив в (20) выражение

$$f''(t) = \bar{f}''\left(\frac{1}{2}\right) + \left(t - \frac{1}{2}\right) \bar{f}''' \left(\frac{1}{2}\right) + \frac{1}{2} \left(t - \frac{1}{2}\right)^2 \bar{f}^{IV}(\eta),$$

$$\eta \in (0, 1),$$

после несложных вычислений найдем

$$\Delta_1(\bar{f}) = -\frac{1}{24} \bar{f}''\left(\frac{1}{2}\right) + \frac{1}{960} \bar{f}^{IV}(\eta), \quad \eta \in (0, 1).$$

Отсюда следует, что

$$D_N(f) = -\frac{h^2}{24} \sum_{i=1}^N hf''_{i-1/2} + \frac{h^4}{960} \sum_{i=1}^N hf^{IV}(\xi_i).$$

Учитывая, что в силу (23)

$$\sum_{i=1}^N hf''_{i-1/2} = \int_a^b f''(x) dx - \frac{h^2}{24} f^{IV}(\xi^*) \cdot (b-a), \quad \xi^* \in [a, b],$$

получаем разложение (24).

Из (24) видно, что формула прямоугольника имеет четвертый порядок точности:  $D_N(f) = O(h^4)$ , если функция  $f(x)$  удовлетворяет условию  $f'(a) = f'(b)$ . Если известны  $f'(a)$  и  $f'(b)$ , то можно положить  $f(x) = \varphi(x) + \alpha x + \beta x^2$ , где  $\varphi(x)$  удовлетворяет условию  $\varphi'(a) = \varphi'(b)$ , если выбрать  $\alpha = \frac{bf'(a) - af'(b)}{b - a}$ ,  $\beta = \frac{f'(b) - f'(a)}{2(b - a)}$ . Тогда

$$\int_a^b f(x) dx = \int_a^b \varphi(x) dx + c,$$

$$c = \frac{1}{2} \alpha (b^2 - a^2) + \frac{1}{6} \beta (b^3 - a^3).$$

Интеграл от  $\varphi(x)$  вычисляется по формуле прямоугольника с точностью  $O(h^4)$ .

2) Формула трапеции:  $m = 1$ ,  $p_0 = p_1 = 1/2$ ,  $s_0 = 0$ ,  $s_1 = 1$ ,

$$\Lambda(\bar{f}) = \frac{1}{2}(\bar{f}(0) + \bar{f}(1)).$$

Функция  $F_2(t) = \frac{1}{2}t(1-t) > 0$  знакопостоянна, поэтому верна оценка

$$D_N(f) = \frac{h^2}{12} f''(\xi^*) \cdot (b - a), \quad \xi^* \in [a, b],$$

т. е. коэффициент при  $h^2$  в выражении для погрешности формулы трапеции в 2 раза больше, чем для формулы прямоугольника. Повторяя рассуждения, аналогичные проведенным выше, убеждаемся в том, что верна формула

$$D_N(f) = -2\alpha_2 h^2 + \alpha_4 h^4 \quad \text{при } f \in C^{(n)}, \quad n \geq 4,$$

где  $\alpha_2$  определяется согласно (24),  $\alpha_4 = O(1)$ .

3) Формула Симпсона:  $m = 2$ ,  $s_0 = 0$ ,  $s_1 = 1/2$ ,  $s_2 = 1$ ,  $p_0 = p_2 = 1/2$ ,  $p_1 = 4/6$ ,

$$\Lambda(\bar{f}) = \frac{1}{6} \left( \bar{f}(0) + 4\bar{f}\left(\frac{1}{2}\right) + \bar{f}(1) \right).$$

Так как формула Симпсона точна для полинома третьей степени, то  $n = 3$ , и мы вычисляем:

$$\Delta_3(\bar{f}) = \int_0^1 F_4(t) \bar{f}^{IV}(t) dt,$$

$$F_4(t) = \frac{1}{6}(K_3(0-t) + K_3(1-t)) + \frac{4}{6}K_3\left(\frac{1}{2}-t\right) - \frac{(1-t)^4}{24}.$$

Отсюда находим

$$F_4(t) = \frac{1}{72}(2t^3 - 3t^4), \quad t < \frac{1}{2};$$

$$F_4(t) = \frac{1}{72}(2(1-t)^3 - 3(1-t)^4), \quad t > \frac{1}{2},$$

$$F_4(t) > 0 \text{ для всех } t \in (0, 1),$$

и значит,

$$\int_0^1 F_4(t) dt = \frac{1}{2880},$$

так что верна формула

$$\Delta_3(\bar{f}) = \frac{1}{2880} \bar{f}^{IV}(\eta), \quad \eta \in (0, 1).$$

Переходя к интегралам по  $x$  и учитывая, что  $\kappa = 2h$ ,  $\bar{f}^{IV}(\eta) = (2h)^4 f^{IV}(\xi_1)$ , получим

$$\begin{aligned} D_N(f) &= \sum_{i=0}^{i_0-1} 2h \left\{ \frac{f_{i-1} + 4f_i + f_{i+1}}{6} - \frac{1}{2h} \int_{x_{i-1}}^{x_{i+1}} f(x) dx \right\} = \\ &= \frac{b-a}{180} h^4 f^{IV}(\xi^*), \quad \xi^* \in [a, b], \end{aligned}$$

где  $N = 2i_0$ ,  $h = 1/N$ .

Если  $f(x) \in C^{(n)}$  ( $n \geq 6$ ), то можно получить разложение вида

$$D_N(f) = \alpha_4 h^4 + \alpha_6 h^6, \quad \alpha_6 = O(1),$$

$$\alpha_4 = \frac{1}{180} \int_0^1 f^{IV}(x) dx = \frac{1}{180} (f''(1) - f''(0)).$$

**6. Повышение порядка точности. Метод Рунге.** Для квадратурных формул (по аналогии с предыдущим) можно получить асимптотическое разложение вида

$$D_N(f) = J_N(f) - J(f) = \alpha_2 h^2 + \alpha_4 h^4 + \alpha_6 h^6 + \dots,$$

если  $f(x)$  — достаточно гладкая функция. При этом  $|\alpha_{k+2}|$  значительно меньше  $|\alpha_k|$  ( $k = 2, 4$ ), поэтому повышение порядка точности квадратурной формулы весьма важно.

Проведем расчеты на двух равномерных сетках с шагами  $h_1$  и  $h_2$  соответственно и найдем выражения  $J^{h_1}[f] = J_{N_1}[f]$  и  $J^{h_2}[f] = J_{N_2}[f]$ ,  $h_1 N_1 = h_2 N_2 = b - a$ . Потребуем, чтобы погрешность для их линейной комбинации:

$$\tilde{D}^h(f) = \sigma D^{h_1}(f) + (1 - \sigma) D^{h_2}(f)$$

была величиной более высокого порядка по сравнению с  $D^{h_1}$  и  $D^{h_2}$ . Если для  $D^h = D_N$  имеет место формула вида

$$D^h = J^h[f] - J[f] = \alpha_p h^p + \alpha_q h^q + \dots, \quad q > p,$$

то для  $\tilde{D}^h = (\sigma J^{h_1}[f] + (1 - \sigma) J^{h_2}[f]) - J[f]$  получим  
 $\tilde{D}^h(f) = \alpha_p (\sigma h_1^p + (1 - \sigma) h_2^p) + \alpha_q (\sigma h_1^q + (1 - \sigma) h_2^q) + \dots$

Выберем параметр  $\sigma$  из условия  $\sigma h_1^p + (1 - \sigma) h_2^p = 0$ :

$$\sigma = h_2^p / (h_2^p - h_1^p).$$

Тогда имеем

$$\tilde{D}^h(f) = \alpha_q (\sigma h_1^q + (1 - \sigma) h_2^q) + \dots = O(h^q),$$

$$h = \max(h_1, h_2),$$

причем  $\sigma h_1^q + (1 - \sigma) h_2^q < 0$ . Так, если  $p = 2$ ,  $q = 4$ , то  $\tilde{D}^h(f) = -\alpha_4 h_1^2 h_2^2 + \dots = O(h^4)$ . Таким образом, проводя вычисления на двух сетках с шагами  $h_1$  и  $h_2 \neq h_1$ , мы повысили порядок точности на 2 (на  $q - p$ ) для  $\tilde{J} = \sigma J^{h_1} + (1 - \sigma) J^{h_2}$ .

Заметим, что комбинируя формулу трапеции  $J_{\text{трап}}^{2h}[f]$  и прямоугольника  $J_{\text{прям}}^{2h}[f]$  с шагом  $2h$ , мы получим формулу Симпсона  $J_{\text{симп}}^h$  с шагом  $h$ :

$$\begin{aligned} J_{\text{симп}}^h[f] &= \frac{1}{3} J_{\text{трап}}^{2h}[f] + J_{\text{прям}}^{2h}[f] = \\ &= \frac{h}{6} (f_0 + 4f_1 + 2f_2 + \dots + 2f_{2N-2} + 4f_{2N-1} + f_{2N}), \end{aligned}$$

где  $h = (b - a)/(2N)$ .

Метод расчета на нескольких сетках применяется для повышения порядка точности даже в том случае, когда неизвестен порядок главного члена погрешности (*процесс Эйткена*). Предположим, что для погрешности имеет

место представление

$$D^h(f) = \alpha_p h^p + \alpha_q h^q + \dots, \quad q > p,$$

так что

$$J^h[f] = J[f] + \alpha_p h^p + \alpha_q h^q + \dots$$

Проведем вычисления на трех сетках:  $h_1 = h$ ,  $h_2 = \rho h$ ,  $h_3 = \rho^2 h$  ( $0 < \rho < 1$ ). Определим сначала  $p$ . При этом пренебрегаем членом  $O(h^q)$ . Образуем отношение

$$A = \frac{J^{h_1}[f] - J^{h_2}[f]}{J^{h_2}[f] - J^{h_3}[f]} \approx \frac{h_1^p - h_2^p}{h_2^p - h_3^p} = \frac{1 - \rho^p}{\rho^p (1 - \rho^p)} = \left(\frac{1}{\rho}\right)^p$$

и найдем

$$p \approx \ln A / \ln \frac{1}{\rho}.$$

Зная приближенное значение  $p$ , можно методом Рунге, изложенным выше, повысить порядок точности. Для этого образуем комбинацию  $\tilde{J}^h = \sigma J^{h_1} + (1 - \sigma) J^{h_2}$  и выберем  $\sigma$  так, чтобы  $\sigma h_1^p + (1 - \sigma) h_2^p = (\sigma + (1 - \sigma) \rho^p) h^p = 0$ , т. е.  $\sigma = \rho^p / (\rho^p - 1) = 1/(1 - A)$ . Тогда для погрешности  $\tilde{D}^h = \tilde{J}^h - J$  получаем

$$\tilde{D}^h(f) = O(h^q).$$

Конечно, все эти рассуждения имеют смысл при соответствующей гладкости функции  $f(x)$ .

**7. Другие квадратурные формулы.** Без нарушения общности можно считать

$$J[f] = \int_0^1 f(x) dx. \quad (25)$$

Мы рассматривали до сих пор квадратурные формулы с заданными узлами  $\{x_k\}$ :

$$J_N[f] = \sum_{k=0}^N c_k f(x_k). \quad (26)$$

Эти формулы точны для всех полиномов степени  $N$ . Если считать неизвестными не только  $c_k$ , но и узлы  $x_k$ , то можно требовать, чтобы квадратурная формула (26) была точной для всех полиномов степени  $2N - 1$ . Такая формула называется *формулой Гаусса*. Требуя, чтобы для одночленов  $1, x, x^2, \dots, x^n, \dots, x^N$  формула была точной,

T. e.

$$J_N[x^m] = \sum_{h=0}^N c_h x_h^m = \int_0^1 x^m dx = \frac{x^{m+1}}{m+1} \Big|_0^1 = \frac{1}{m+1},$$

$m = 0, 1, \dots, 2N - 1,$

получим для узлов и весов  $2N + 2$  уравнений

$$c_0 + c_1 + \dots + c_N = 1,$$

$$c_0x_0 + c_1x_1 + \dots + c_Nx_N = 1/2,$$

$$c_0x_0^m + c_1x_1^m + \dots + c_Nx_N^m = 1/(m+1),$$

$$c_0 x_0^{2N+1} + c_1 x_1^{2N+1} + \dots + c_N x_N^{2N+1} = 1/(N+1).$$

Общее число неизвестных равно  $2N + 2$ , т. е.  $N + 1$  неизвестных узлов и  $N + 1$  весовых множителей. Число уравнений также равно  $2N + 2$ . Можно доказать, что написанная система уравнений имеет решение.

Приведем простейшую формулу Гаусса при  $N = 2$ :

$$J_N[f] = \frac{5}{18} f(x_0) + \frac{8}{18} f(x_1) + \frac{5}{18} f(x_2),$$

ГДЗ

$$x_0 = \frac{1 - \sqrt{0.6}}{2}, \quad x_2 = \frac{1 + \sqrt{0.6}}{2}, \quad x_1 = \frac{1}{2}.$$

Формулы Гаусса дают хорошую точность при небольшом числе узлов.

Еще одним примером является квадратурная формула Чебышева, в которой выбираются наилучшие узлы в предположении, что все веса равны. В этом случае

$$J_N[f] = \frac{1}{N} \sum_{k=1}^N f(x_k).$$

Требуя, чтобы формула была точной для  $f(x) = x$ ,  $x^2$ , ...,  $x^N$ , получим  $N$  уравнений для определения  $x_1$ ,  $x_2$ , ...,  $x_N$ :

$$x_1^m + x_2^m + \dots + x_N^m = \frac{1}{m+1}, \quad m = 1, 2, \dots, N.$$

Эти уравнения имеют решения при  $m = 1, 2, \dots, 7, 9$ , а при  $m = 8$  и  $m \geq 10$  не имеют вещественных корней.

При  $m = 3$  формула Чебышева имеет вид

$$\int_0^1 f(x) dx \approx J_3[f] = \frac{1}{3} \left[ f\left(\frac{1}{2} - \frac{1}{4}\sqrt{2}\right) + f\left(\frac{1}{2}\right) + f\left(\frac{1}{2} + \frac{1}{4}\sqrt{2}\right) \right].$$

Для нее коэффициент при  $\|f^{IV}\|_\infty$  в оценке для погрешности в два раза меньше, чем для формулы Симпсона.

**Замечания.** В ряде случаев вычислению интегралов должно предшествовать их преобразование, учитывающее специфику подынтегральной функции. Примеры:

1)  $f(x) = \frac{1}{2\sqrt{x}} f_0(x)$ ,  $f_0(0) \neq 0$ , т. е.  $f(x)$  имеет особенность при  $x = 0$ . Эта особенность устраивается заменой переменной:

$$\int_0^1 f(x) dx = \int_0^1 \frac{f_0(x)}{2\sqrt{x}} dx = \int_0^1 f_0(x) d\sqrt{x} = \int_0^1 f(t^2) dt, \quad t = \sqrt{x}.$$

2) Подынтегральная функция имеет экспоненциальный характер:  $f(x) \approx ce^{ax}$ , т. е. функция  $\ln f(x)$  линейна. Представим  $f(x)$  в виде  $f(x) = \exp\{\ln f(x)\}$ , проинтегрируем  $\ln f(x)$  линейно на отрезке  $[x_{i-1}, x_i]$ :

$$\ln f(x) = \frac{x_i - x}{x_i - x_{i-1}} \ln f_{i-1} + \frac{x - x_{i-1}}{x_i - x_{i-1}} \ln f_i,$$

и затем проинтегрируем по  $x$  от  $x_{i-1}$  до  $x_i$ . Эта формула оказывается полезной на практике.

3) Если  $f(x)$  является быстро осциллирующей функцией, так что ее можно записать в виде  $f(x) = y(x) \cos \omega x$ , где частота  $\omega \gg 1$  велика, то при вычислении интеграла можно воспользоваться следующим приемом. Сначала проинтегрируем по частям:

$$\begin{aligned} \int_{x_{i-1}}^{x_i} f(x) dx &= \int_{x_{i-1}}^{x_i} y(x) \cos \omega x dx = \\ &= \frac{1}{\omega} y \sin \omega x \Big|_{x_{i-1}}^{x_i} - \frac{1}{\omega} \int_{x_{i-1}}^{x_i} y'(x) \sin \omega x dx. \end{aligned}$$

Если  $y(x)$  — линейная на  $[x_{i-1}, x_i]$  функция, то интеграл справа берется в явном виде. Если  $y(x)$  — полином степени  $n$ , то интегрирование по частям производим  $n$  раз.

## Глава III

# ЧИСЛЕННОЕ РЕШЕНИЕ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

В этой главе изучаются методы численного решения систем линейных алгебраических уравнений, т. е. численные методы линейной алгебры. Существуют два типа методов — прямые и итерационные. Мы рассматриваем прежде всего метод исключения Гаусса для систем общего вида и варианты — метод прогонки и методы матричной прогонки для систем специального вида (с трехдиагональной или блочно-трехдиагональной матрицами). Это — *прямые методы*. Их эффективность зависит от порядка системы и структуры матрицы.

При изучении *итерационных* методов мы трактуем систему уравнений как операторное уравнение первого рода  $Au = f$  и излагаем общую теорию итерационных методов для операторных уравнений при минимальных предположениях относительно оператора  $A$ . Общая теория позволяет доказать сходимость итераций для метода Зейделя и метода верхней релаксации при минимальных ограничениях на оператор  $A$ . Рассмотрены два класса методов: 1) для случая, когда известны границы  $\gamma_1 > 0$  и  $\gamma_2 \geq \gamma_1$  спектра оператора  $A$  в некотором энергетическом пространстве  $H_B$ ; 2) для случая, когда границы  $\gamma_1$  и  $\gamma_2$  неизвестны. Весьма эффективным является попеременно-треугольный метод, который изучается в § 5.

## § 1. Системы линейных алгебраических уравнений

**1. Системы уравнений.** Основная задача линейной алгебры — решение системы уравнений

$$Au = f, \quad (1)$$

где  $u = (u^{(1)}, \dots, u^{(N)})$  — искомый вектор,  $f = (f^{(1)}, f^{(2)}, \dots, f^{(N)})$  — известный вектор размерности  $N$ ,  $A = (a_{ij})$  ( $i, j = 1, 2, \dots, N$ ) — квадратная матрица размера  $N \times N$  с элементами  $a_{ij}$ .

Будем предполагать, что матрица  $A$  невырождена;  $\det A \neq 0$ , так что уравнение  $Au = 0$  имеет только тривиальное решение, и система (1) имеет единственное

решение

$$u = A^{-1}f.$$

В курсе линейной алгебры решение системы (1) обычно выражают по формулам Крамера в виде отношений определителей. Для численного решения системы (1) эти формулы непригодны, так как они требуют вычисления  $N+1$  определителей, что требует большого числа действий (порядка  $N!$  арифметических операций). Даже при выборе наилучшего метода вычисление одного определителя требует примерно такого же времени, что и решение системы линейных уравнений современными численными методами. Кроме того, следует иметь в виду, что вычисления по формулам Крамера часто ведут к большим ошибкам округлений.

Особенность большинства численных методов для (1) состоит в отказе от нахождения обратной матрицы. Основное требование к методу решения — минимум числа арифметических действий, достаточных для отыскания приближенного решения с заданной точностью  $\epsilon > 0$  (экономичность численного метода).

**2. Частные случаи систем.** Несложно решить систему (1) в перечисленных ниже частных случаях. Пусть матрица  $A$  — *диагональная*, т. е.  $a_{ij} = 0$ ,  $j \neq i$ ,  $a_{ii} \neq 0$  ( $i, j = 1, 2, \dots, N$ ). Тогда система имеет вид

$$a_{ii}u^{(i)} = f^{(i)},$$

откуда находим

$$u^{(i)} = f^{(i)} / a_{ii}, \quad i = 1, 2, \dots, N.$$

Если  $A$  — *нижняя треугольная* матрица, т. е.  $a_{ij} = 0$  при  $j > i$  ( $i, j = 1, 2, \dots, N$ ),  $a_{ii} \neq 0$ ,

$$A = \begin{bmatrix} -a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{bmatrix},$$

то система уравнений имеет вид

$$\begin{aligned} a_{11}u^{(1)} &= f^{(1)}, \\ a_{21}u^{(1)} + a_{22}u^{(2)} &= f^{(2)}, \\ \vdots &\vdots \\ a_{N1}u^{(1)} + a_{N2}u^{(2)} + \dots + a_{NN}u^{(N)} &= f^{(N)}. \end{aligned}$$

Компоненты вектора  $u$  находятся последовательно при переходе от  $n$  к  $n+1$ , начиная с  $n=1$ :

$$\begin{aligned} u^{(1)} &= \frac{f^{(1)}}{a_{11}}, \quad u^{(2)} = \frac{1}{a_{22}} (f^{(2)} - a_{21}u^{(1)}), \dots, \\ u^{(n+1)} &= \frac{1}{a_{n+1,n+1}} \left( f^{(n+1)} - \sum_{k=1}^n a_{n+1,k}u^{(k)} \right), \\ n &= 2, 3, \dots, N-1. \end{aligned}$$

Чтобы найти вектор  $u = (u^{(1)}, \dots, u^{(N)})$ , требуется  $1 + 3 + 5 + \dots + 2N - 1 = N^2$  арифметических действий.

Если  $A$  — верхняя треугольная матрица, т. е.  $a_{ij} = 0$  при  $j < i$ ,  $a_{ii} \neq 0$  ( $i, j = 1, 2, \dots, N$ )

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & & a_{1N} \\ 0 & a_{22} & \cdots & & a_{2N} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & & a_{N-1,N-1} & \ddots & a_{N-1,N} \\ & & & & a_{NN} \end{bmatrix},$$

то система (1) имеет вид

$$\begin{aligned} a_{11}u^{(1)} + a_{12}u^{(2)} + \dots + a_{1N}u^{(N)} &= f^{(1)}, \\ a_{22}u^{(2)} + \dots + a_{2N}u^{(N)} &= f^{(2)}, \\ \vdots &\vdots \\ a_{N-1,N-1}u^{(N-1)} + a_{N-1,N}u^{(N)} &= f^{(N-1)}, \\ a_{NN}u^{(N)} &= f^{(N)}. \end{aligned}$$

Компоненты вектора  $u = A^{-1}f$  определяются последовательно при переходе от  $n+1$  к  $n$  по формулам

$$\begin{aligned} u^{(N)} &= \frac{f^{(N)}}{a_{NN}}, \quad \dots, \quad u^{(n)} = \frac{1}{a_{nn}} \left( f^{(n)} - \sum_{k=n+1}^N a_{nk}u^{(k)} \right), \dots \\ n &= N-1, N-2, \dots, 2, 1, \quad (2) \end{aligned}$$

что требует также  $N^2$  действий.

Выбор метода и его экономичность зависят от вида матрицы  $A$ , а также от типа компьютера.

Для многих задач  $A$  является разреженной матрицей, большинство элементов которой — нули. Такие матрицы часто появляются при разностной аппроксимации дифференциальных уравнений. Элементы этой матрицы обыч-

по вычисляются по заданным формулам, и их можно не хранить в оперативной памяти машины. Это очень важно, так как порядок таких матриц может достигать нескольких десятков и даже сотен тысяч.

Частным случаем разреженной матрицы является *ленточная матрица*; все ее ненулевые элементы находятся вблизи главной диагонали, т. е.  $a_{ij} = 0$ , если  $|i - j| > l$  где  $l < N$ . Отличные от нуля элементы расположены на  $2l + 1$  диагоналях, включая главную диагональ. Примером является трехдиагональная матрица

$$A = \begin{bmatrix} -c_1 & b_1 & 0 & \dots & 0 \\ a_2 & -c_2 & b_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & & 0 & a_N & -c_N \end{bmatrix}.$$

Соответствующая система (1) имеет вид

$$\begin{aligned} -c_1 u^{(1)} + b_1 u^{(2)} &= f^{(1)}, \\ \dots &\dots \\ a_i u^{(i-1)} - c_i u^{(i)} + b_i u^{(i+1)} &= f^{(i)}, \\ \dots &\dots \\ a_N u^{(N-1)} - c_N u^{(N)} &= f_N, \\ i &= 2, 3, \dots, N-1. \end{aligned}$$

Это разностное уравнение второго порядка, которое мы рассматривали в гл. I, где для его решения применялся метод прогонки.

**3. Операторное уравнение первого рода.** Известно, что всякая матрица  $A = (a_{ij})$  ( $i, j = 1, 2, \dots, N$ ) определяет линейный оператор  $A$ , отображающий пространство  $H^N$  в себя:

$Au \in H^N$  для любого  $u \in H^N$  или  $A: H^N \rightarrow H^N$ .  
Обратно, каждому оператору  $A$  (в некотором базисе  $\xi_1, \dots, \xi_N$ ) соответствует матрица  $A = (a_{ij})$  размера  $N \times N$ , где  $a_{ij}$  — компонента вектора  $A\xi_j$ . Поэтому мы можем рассматривать (1) как *операторное уравнение первого рода*

$$Au = f, \quad u, f \in H^N, \quad (3)$$

с оператором  $A: H^N \rightarrow H^N$ .

Чтобы подчеркнуть эквивалентность задач (1) и (3), мы сохраним одно и то же обозначение  $A$  как для мат-

рицы, так и для оператора. Индекс  $N$  у  $H^N$  будем опускать и писать просто  $H$ . Переход от (1) к операторному уравнению удобен для изложения теории итерационных методов. При этом какая-либо конкретная информация о структуре матрицы  $A$  не используется.

В пространстве  $H$  введем скалярное произведение  $(\cdot)$  и норму  $\|u\| = \sqrt{(u, u)}$ . Будем предполагать, что оператор  $A$  является самосопряженным и положительным:  $A = A^* > 0$ . Будем рассматривать также энергетические пространства  $H_D$  со скалярным произведением  $(u, v)_D = (Du, v)$  и нормой  $\|u\|_D = \sqrt{(Du, u)}$ , где  $D$  — линейный самосопряженный положительный оператор:  $D: H \rightarrow H$ ,  $D = D^* > 0$ .

Обозначим через  $(\xi_s, \lambda_s)$  ( $s = 1, 2, \dots, N$ ) собственные векторы и собственные значения оператора  $A$ :

$$A\xi_s = \lambda_s \xi_s, \quad (\xi_s, \xi_m) = \delta_{sm}, \quad s, m = 1, 2, \dots, N.$$

Так как  $A > 0$ , то  $\lambda_s > 0$ , и можно считать, что  $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$ , а значит, справедливо неравенство

$$\lambda_1 E \leq A \leq \lambda_N E, \quad \lambda_1 = \min_s \lambda_s, \quad \lambda_N = \max_s \lambda_s.$$

Отношение  $\lambda_N/\lambda_1$  называют *числом обусловленности*.

На практике удобнее пользоваться обратным отношением, т. е. параметром  $\xi = \lambda_1/\lambda_N$ , который мы будем называть *мерой обусловленности*. От него, как это будет показано ниже, зависит скорость сходимости итераций. Параметр  $\xi$  для разностных уравнений, аппроксимирующих уравнения математической физики (например, уравнения Лапласа), мал:  $\xi \approx 10^{-2} - 10^{-4}$  (число обусловленности велико).

Из формулы  $u = A^{-1}f$  видно, что

$$\|u\| \leq \|A^{-1}\| \|f\|, \quad \|A^{-1}\| = 1/\lambda_1.$$

Это неравенство выражает устойчивость решения задачи (1) по правой части. Если  $\|A^{-1}\| = 1/\lambda_1$  очень велико, то задача (3) может оказаться некорректной, т. е. неустойчивой по отношению к ошибкам в задании правой части, в том числе к ошибкам округления.

**4. Прямые и итерационные методы.** Численные методы решения системы (1) условно разбивают на две группы, выделяя прямые и итерационные методы (имеются, конечно, и гибридные методы). *Прямые* методы позволяют за конечное число действий получить точное решение

системы уравнений, если входная информация (правая часть уравнения  $f$  и элементы  $a_{ij}$  матрицы  $A$ ) заданы точно и вычисления ведутся без округления. Простейший пример прямого метода — метод прогонки. Конечно, прямые методы также дают решение с определенной точностью, которая зависит от ошибок округления, т. е. от машины, и от характера вычислительной устойчивости, что зависит от самого метода.

*Итерационный метод* позволяет найти приближенное решение системы путем построения последовательности приближений (итераций), начиная с некоторого начального приближения. Само приближенное решение является результатом вычислений, полученным после конечного числа итераций.

Выбор того или иного численного метода зависит от многих обстоятельств — от имеющихся программ, от вида матрицы  $A$ , от типа расчета и др. Поясним слова «тип расчета». Возможны разные постановки задачи:

- 1) найти решение одной конкретной задачи (1);
- 2) найти решение нескольких вариантов задачи (1) с одной и той же матрицей  $A$  и разными правыми частями  $f$ . Может оказаться, что неоптимальный для одной задачи (1) метод является весьма эффективным для многовариантного расчета.

При многовариантном расчете можно уменьшить среднее число операций для одного варианта, если хранить некоторые величины, а не вычислять их заново для каждого варианта. Это, конечно, зависит от машины, от объема ее оперативной памяти.

Из сказанного ясно, что выбор алгоритма должен зависеть от типа расчета, от объема оперативной памяти машины и, конечно, от порядка системы. Качество алгоритма определяется тем машинным временем, которое требуется для нахождения решения системы (1). Естественно выбирать тот метод, для которого время решения минимально по сравнению с другими методами. Однако время расчета зависит от многих факторов: от числа арифметических и логических действий, которые нужно затратить для получения решения с заданной точностью, от быстродействия и оперативной памяти машины, от качества программы. При теоретических оценках качества алгоритмов их сравнение проводится по числу  $Q(\epsilon)$  арифметических действий, достаточных для нахождения решения задачи с заданной точностью  $\epsilon > 0$ .

## § 2. Прямые методы

**1. Метод Гаусса.** Имеется несколько вычислительных вариантов метода Гаусса, основанного на идее последовательного исключения. Процесс решения системы линейных алгебраических уравнений  $Ax = f$ , или

$$\sum_{j=1}^N a_{ij}x_j = f_i, \quad i = 1, 2, \dots, N, \quad (1)$$

по методу Гаусса состоит из двух этапов.

Первый этап (*прямой ход*). Система (1) приводится к треугольному виду

$$x + B^+x = \varphi, \quad (2)$$

где  $x = (x_1, \dots, x_N)$  — неизвестный,  $\varphi = (\varphi_1, \dots, \varphi_N)$  — известный векторы,  $B^+$  — верхняя треугольная матрица.

Второй этап (*обратный ход*). Неизвестные  $x_N, x_{N-1}, \dots, x_1$  определяются по формулам (2) из § 1.

Перейдем к подробному изложению метода. Первый шаг метода Гаусса состоит в исключении из всех уравнений, кроме первого, неизвестного  $x_1$ . Предположим, что  $a_{11} \neq 0$ , разделим первое уравнение (1) ( $i = 1$ ) на  $a_{11}$  и запишем систему (1) в виде

$$x_1 + b_{12}x_2 + \dots + b_{1N}x_N = \varphi_1, \quad b_{1j} = a_{1j}/a_{11}, \quad 2 \leq j \leq N, \quad \varphi_1 = f_1/a_{11}, \quad (3)$$

$$a_{ii}x_1 + a_{i2}x_2 + \dots + a_{iN}x_N = f_i, \quad i = 2, 3, \dots, N. \quad (4)$$

Умножим уравнение (3) на  $a_{ii}$ , где  $i$  — любое из чисел  $i = 2, 3, \dots, N$ , и вычтем полученное уравнение из  $i$ -го уравнения (4):

$$(a_{i2} - a_{11}b_{12})x_2 + \dots + (a_{iN} - a_{11}b_{1N})x_N = f_i - a_{11}\varphi_1, \quad i = 2, 3, \dots, N.$$

Вводя обозначения

$$a_{ij}^{(1)} = a_{ij} - a_{11}b_{1j}, \quad f_i^{(1)} = f_i - a_{11}\varphi_1, \quad i, j = 2, 3, \dots, N, \quad (5)$$

перепишем полученную систему уравнений (эквивалентную системе (1)) в виде

$$x_1 + b_{12}x_2 + \dots + b_{1N}x_N = \varphi_1, \\ a_{i2}^{(1)}x_2 + \dots + a_{iN}^{(1)}x_N = f_i^{(1)}, \quad i = 2, 3, \dots, N.$$

Первый столбец матрицы этой системы состоит из нулей, кроме первого элемента при  $i = 1, j = 1$ , равного единице.

Второй шаг состоит в исключении  $x$ , из системы

$$\begin{aligned} a_{22}^{(1)}x_2 + \dots + a_{2N}^{(1)}x_N &= f_2^{(1)}, \\ \vdots &\quad \vdots \\ a_{N2}^{(1)}x_2 + \dots + a_{NN}^{(1)}x_N &= f_N^{(1)}. \end{aligned} \tag{6}$$

Для этого разделим первое уравнение на  $a_{22}^{(1)}$ :

$$x_2 + b_{23}x_3 + \dots + b_{2N}x_N = \varphi_2,$$

умножим его затем на  $(-a_{i_2}^{(1)})$  и сложим с уравнением

$$a_{i2}^{(1)}x_2 + a_{i3}^{(1)}x_3 + \dots + a_{iN}^{(1)}x_N = f_i^{(1)}, \quad i = 3, 4, \dots, N.$$

В результате получим систему

$$\begin{aligned} x_2 + b_{23}x_3 + \dots + b_{2N}x_N &= \varPhi_2, \\ a_{i3}^{(2)}x_3 + \dots + a_{iN}^{(2)}x_N &= f_i^{(2)}, \quad i = 3, 4, \dots, N, \\ a_{ij}^{(2)} &= a_{ij}^{(1)} - a_{i2}^{(1)}b_{2j}, \quad f_i^{(2)} = f_i^{(1)} - a_{i2}^{(1)}\varPhi_2, \end{aligned} \quad \begin{aligned} (7) \\ i = 3, 4, \dots, N. \quad (8) \end{aligned}$$

Для  $x_3, x_4, \dots, x_N$  имеем систему  $(N - 2)$ -го порядка, аналогичную системе (6)  $(N - 1)$ -го порядка для  $x_1, x_2, \dots, x_N$ .

Продолжая рассуждения, после  $(N-1)$ -го шага (т. е. после исключения  $x_1, x_2, \dots, x_{N-1}$ ) получим

$$a_{NN}^{(N-1)}x_N = f_N^{(N-1)}, \text{ или } x_N = \varphi_N, \quad \varphi_N = f_N^{(N-1)}/a_{NN}^{(N-1)}. \quad (9)$$

В итоге приходим к системе (2) с верхней треугольной матрицей:

Обратный ход метода Гаусса состоит в определении всех  $x_i$  из системы (10) с верхней треугольной матрицей. Нетрудно показать, что изложенный выше метод Гаусса можно применять в том случае, когда все главные миноны отличны от нуля.

Подсчитаем число умножений и делений в методе Гаусса. Рассмотрим сначала прямой ход. На первом шаге требуется  $Q_1 = N^2$  делений и умножений, второй шаг требует  $Q_2 = (N - 1)^2$  действий и т. д. Всего надо сделать  $N$  шагов прямого хода, затратив на это

$$\sum_{k=1}^N (N - k + 1)^2 = \sum_{s=1}^N s^2 = \frac{N(N + 1)(2N + 1)}{6}$$

умножений и делений. Для обратного хода, очевидно, нужно сделать  $N(N - 1)/2$  умножений. Таким образом, для решения системы уравнений (1) требуется  $Q = N(N^2 + 3N - 1)/3$  умножений и делений. Примерно столько же потребуется сложений.

Приведем пример применения метода Гаусса. Рассмотрим систему трех уравнений ( $N = 3$ )

$$2x_1 + 4x_2 + 3x_3 = 4, \quad (11)$$

$$3x_1 + x_2 - 2x_3 = -2, \quad (12)$$

$$4x_1 + 11x_2 + 7x_3 = 7. \quad (13)$$

Прямой ход. Первый шаг. Разделим первое уравнение на  $a_{11} = 2$ :

$$x_1 + 2x_2 + 1,5x_3 = 2. \quad (14)$$

Умножим (14) на  $-3$  и сложим с (12), затем умножим (14) на  $-4$  и сложим с (13):

$$-5x_2 - 6,5x_3 = -8, \quad (15)$$

$$3x_2 + x_3 = 1. \quad (16)$$

Мы получили систему второго порядка.

Второй шаг. Разделим (15) на  $-5$ :

$$x_2 + 1,3x_3 = 1,6. \quad (17)$$

Умножим (17) на  $-3$  и сложим с (16):

$$-2,9x_3 = -5,8. \quad (18)$$

Третий шаг. Делим (18) на  $-2,9$ :

$$x_3 = 2.$$

В результате получили систему

$$x_1 + 2x_2 + 1,5x_3 = 2,$$

$$x_2 + 1,3x_3 = 1,6,$$

$$x_3 = 2$$

с верхней треугольной матрицей

$$\begin{bmatrix} 1 & 2 & 1,5 \\ 0 & 1 & 1,3 \\ 0 & 0 & 1 \end{bmatrix}.$$

**Обратный ход.** Из системы последовательно находим:  $x_3 = 2$ ,  $x_2 = 1,6 - 1,3 \cdot x_3 = 1,6 - 1,3 \cdot 2 = -1$ ,  $x_1 = 2 - 2x_2 - 1,5 \cdot x_3 = 1$ . Таким образом, решение системы (11)–(13) найдено:

$$x_1 = 1, \quad x_2 = -1, \quad x_3 = 2.$$

**2. Метод квадратного корня.** Этот метод пригоден для систем

$$Au = f \quad (19)$$

с эрмитовой (в действительном случае — симметричной) матрицей  $A$ . Матрица  $A$  разлагается в произведение

$$A = S^*DS, \quad (20)$$

где  $S$  — верхняя треугольная,  $D$  — диагональная матрица. Решение уравнения  $Au = f$  сводится к последовательному решению двух систем

$$S^*Dy = f, \quad Su = y. \quad (21)$$

Чтобы получить разложение (20), обозначаем  $S = (s_{ij})$ ,  $D = (d_{ii}\delta_{ij})$  и находим

$$(DS)_{ij} = \sum_{k=1}^N d_{ik}s_{kj} = d_{ii}s_{ij}, \quad (S^*DS)_{ij} = \sum_{k=1}^N \bar{s}_{ki}d_{kk}s_{kj},$$

так как  $S^* = (\bar{s}_{ji})$ , где черта — комплексное сопряжение. В результате получаем уравнение

$$\sum_{k=1}^N \bar{s}_{ki}d_{kk}s_{kj} = a_{ij}. \quad (22)$$

Систему уравнений (22) можно решать рекуррентно. Так как  $S$  — верхняя треугольная матрица, то  $s_{ki} = 0$  при  $k > i$ ,  $\bar{s}_{ik} = 0$  при  $k < i$  и, следовательно,

$$\begin{aligned} \sum_{k=1}^N \bar{s}_{ki}s_{kj}d_{kk} &= \\ &= \sum_{k=1}^{i-1} \bar{s}_{ki}s_{kj}d_{kk} + \bar{s}_{ii}s_{ij}d_{ii} + \sum_{k=i+1}^N \bar{s}_{ki}s_{kj}d_{kk} = \\ &= \sum_{k=1}^{i-1} \bar{s}_{ki}s_{kj}d_{kk} + s_{ii}s_{ij}d_{ii} = a_{ij}. \end{aligned}$$

При  $i = j$  имеем

$$|s_{ii}|^2 d_{ii} = a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk}. \quad (23)$$

Выбирая

$$d_{ii} = \operatorname{sign} \left( a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk} \right), \quad (24)$$

найдем

$$s_{ii} = \sqrt{\left| a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk} \right|}. \quad (25)$$

При  $i < j$  получаем

$$s_{ij} = \frac{a_{ij} - \sum_{h=1}^{i-1} s_{hi} s_{hj} d_{hh}}{s_{ii} d_{ii}}. \quad (26)$$

Полагая  $i = 1, 2, \dots$ , находим последовательно

$$s_{11} = \sqrt{|a_{11}|}, \quad d_{11} = \operatorname{sign} a_{11}, \quad s_{22} = \sqrt{|a_{22} - d_{11} |s_{12}|^2|}, \dots$$

Определитель матрицы, очевидно, равен

$$\det A = \prod_{i=1}^N d_{ii} s_{ii}^2.$$

Метод квадратного корня требует порядка  $N^3/3$  арифметических действий, т. е. при больших  $N$  он вдвое быстрее метода Гаусса и занимает вдвое меньше ячеек памяти. Это обстоятельство объясняется тем, что метод использует информацию о симметрии матрицы.

**3. Связь метода Гаусса с разложением матрицы на множители.** Пусть дана невырожденная матрица  $A$  размера  $N \times N$ . Представим ее в виде произведения

$$A = BC, \quad A = (a_{ij}), \quad B = (b_{ij}), \quad C = (c_{ij}), \quad (27)$$

где  $B$  и  $C$  — треугольные матрицы вида

$$B = \begin{bmatrix} -b_{11} & 0 & \dots & 0 \\ b_{21} & b_{22} & 0 & \dots & 0 \\ b_{31} & b_{32} & b_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{N1} & b_{N2} & b_{N3} & \dots & b_{NN} \end{bmatrix}, \quad C = \begin{bmatrix} 1 & c_{12} & c_{13} & \dots & c_{1N} \\ 0 & 1 & c_{23} & \dots & c_{2N} \\ 0 & 0 & 1 & \dots & c_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix},$$

т. е.  $b_{ik} = 0$  при  $k > i$ ,  $c_{ik} = 0$  при  $k < i$ ,  $c_{ii} = 1$ . Из (27)

следует, что

$$a_{ij} = \sum_{k=1}^N b_{ik} c_{kj}.$$

Преобразуем эту сумму двумя способами:

$$\sum_{k=1}^N b_{ik} c_{kj} = \sum_{k=1}^{i-1} b_{ik} c_{kj} + b_{ii} c_{ij} + \sum_{k=i+1}^N b_{ik} c_{kj} = \sum_{k=1}^{i-1} b_{ik} c_{kj} + b_{ii} c_{ij},$$

$$\sum_{k=1}^N b_{ik} c_{kj} = \sum_{k=1}^{j-1} b_{ik} c_{kj} + b_{ij} c_{jj} + \sum_{k=j+1}^N b_{ik} c_{kj} = \sum_{k=1}^{j-1} b_{ik} c_{kj} + b_{ij} c_{jj}.$$

Отсюда находим

$$b_{ij} = a_{ij} - \sum_{k=1}^{j-1} b_{ik} c_{kj} \quad \text{при } i \geq j, \quad b_{11} = a_{11}, \quad c_{11} = 1,$$

$$c_{ij} = \frac{1}{b_{ii}} \left[ a_{ij} - \sum_{k=1}^{i-1} b_{ik} c_{kj} \right] \quad \text{при } i < j.$$

Матрицы  $B$  и  $C$  найдены.

Решение уравнения  $Au = Bu = f$  сводится к последовательному решению уравнений

$$Bu = f, \quad Cu = \varphi.$$

Построение матриц  $B$  и  $C$  и нахождение  $\varphi = B^{-1}f$  соответствуют прямому ходу, а решение уравнения

$$Cu = \varphi$$

соответствует обратному ходу метода Гаусса.

### § 3. Итерационные методы

**1. Метод итераций для решения системы линейных алгебраических уравнений.** Особое внимание в этой главе мы уделим итерационным методам, так как они широко применяются для решения разностных уравнений математической физики, операторам которых соответствуют ленточные матрицы  $A$  высокого порядка.

Перейдем к общему описанию *метода итераций* для системы линейных алгебраических уравнений

$$Au = f. \quad (1)$$

Для ее решения выбирается некоторое начальное прибли-

жение  $y_0 \in H$  и последовательно находятся приближенные решения (итерации) уравнения (1). Значение *итерации*  $y_{k+1}$  выражается через известные предыдущие итерации  $y_k, y_{k-1}, \dots$ . Если при вычислении  $y_{k+1}$  используется только одна предыдущая итерация  $y_k$ , то итерационный метод называют *одношаговым* (или *двухслойным*) методом; если же  $y_{k+1}$  выражается через две итерации  $y_k$  и  $y_{k-1}$ , то метод называется *двухшаговым* (или *трехслойным*). Мы будем рассматривать в основном одношаговые методы. Будем считать, что  $A: H \rightarrow H$  — линейный оператор в конечномерном пространстве  $H$  со скалярным произведением  $(\cdot, \cdot)$ .

Важную роль играет запись итерационных методов в единой (канонической) форме. Любой двухслойный итерационный метод можно записать в следующей канонической форме:

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \text{ для всех } y_0 \in H, \quad (2)$$

где  $A: H \rightarrow H$  — оператор исходного уравнения (1),  $B: H \rightarrow H$  — линейный оператор, имеющий обратный  $B^{-1}$ ,  $k$  — номер итерации,  $\tau_1, \tau_2, \dots, \tau_{k+1}, \dots$  — итерационные параметры,  $\tau_{k+1} > 0$ . Оператор  $B$  может, вообще говоря, зависеть от номера  $k$ ; для простоты изложения мы предполагаем всюду, что  $B$  не зависит от  $k$ .

Если  $B = E$  — единичный оператор, то метод

$$\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \text{ для всех } y_0 \in H, \quad (3)$$

называют *явным*:  $y_{k+1}$  находится по явной формуле

$$y_{k+1} = y_k - \tau_{k+1}(Ay_k - f).$$

В общем случае, при  $B \neq E$ , метод (2) называют *неявным* итерационным методом: для определения  $y_{k+1}$  надо решить уравнение

$$By_{k+1} = By_k - \tau_{k+1}(Ay_k - f) = F_k, \quad k = 0, 1, \dots \quad (4)$$

Естественно требовать, чтобы объем вычислений для решения системы  $By_{k+1} = F_k$  был меньше, чем объем вычислений для прямого решения системы  $Au = f$ .

Точность итерационного метода (2) характеризуется величиной погрешности  $z_k = y_k - u$ , т. е. разностью между решением  $y_k$  уравнения (2) и точным решением  $u$  исходной системы линейных алгебраических уравнений. Подстановка  $y_k = z_k + u$  в (2) приводит к однородному уравнению для погрешности:

$$B \frac{z_{k+1} - z_k}{\tau_{k+1}} + Az_k = 0, \quad k = 0, 1, \dots, \quad z_0 = y_0 - u. \quad (5)$$

Говорят, что итерационный метод *сходится* в  $H_D$ , если  $\lim_{k \rightarrow \infty} \|z_k\|_D = 0$ , где  $\|z\|_D = \sqrt{Dz, z}$ ,  $D = D^* > 0$ ,  $D: H \rightarrow H$ .

Обычно задают некоторую погрешность (относительную)  $\varepsilon > 0$ , с которой надо найти приближенное решение  $y_k$ , и прекращают вычисления, как только выполняется условие

$$\|y_n - u\|_D \leq \varepsilon \|y_0 - u\|_D. \quad (6)$$

Если  $n = n(\varepsilon)$  — наименьшее из чисел, для которых (6) выполняется, то общее число арифметических действий, которые затрачиваются для нахождения приближенного решения уравнения (1), равно  $Q_n(\varepsilon) = n(\varepsilon)q_0$ , где  $q_0$  — число действий, затрачиваемых для нахождения одной итерации, т. е. для решения уравнения (4). Задача состоит в минимизации  $Q_n(\varepsilon)$  путем выбора  $B$  и параметров  $\{\tau_k\}$ . Начнем с простейших итерационных методов.

**2. Метод простой итерации.** Для решения системы уравнений (1) может быть использован *метод простой итерации*

$$y_{k+1}^{(i)} = y_k^{(i)} - \tau \left( \sum_{j=1}^N a_{ij} y_k^{(j)} - f^{(i)} \right), \quad i = 1, 2, \dots, N, \quad (7)$$

где  $\tau > 0$  — итерационный параметр. Запишем (7) в операторной форме:

$$\frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \dots, \text{для всех } y_0 \in H. \quad (8)$$

Сравнивая с (3), видим, что метод простой итерации задается явной двухслойной схемой с постоянным параметром  $\tau_k \equiv \tau$ .

Существуют и другие варианты метода простой итерации, например такой:

$$y_{k+1}^{(i)} = \frac{1}{a_{ii}} \left( \sum_{j=1}^{i-1} a_{ij} y_k^{(j)} - f^{(i)} \right).$$

Подставляя сюда

$$\sum_{j=1}^{i-1} a_{ij} y_k^{(j)} = \sum_{j=1}^N a_{ij} y_k^{(j)} - a_{ii} y_k^{(i)} = (Ay_k)^{(i)} - (Dy_k)^{(i)},$$

где  $D = (a_{ii} \delta_{ij})$  — диагональная матрица, получаем

$$y_{k+1}^{(i)} = y_k^{(i)} - \frac{1}{a_{ii}} \left( \sum_{j=1}^N a_{ij} y_k^{(j)} - f^{(i)} \right),$$

или, в каноническом виде,

$$D \frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \dots, \tau = 1.$$

Хотя формально эта схема является неявной ( $B = D \neq E$ ), однако  $D = (a_{ii} \delta_{ij})$  — диагональная матрица и потому  $y_{k+1}$  определяется по явным формулам.

**3. Метод Зейделя.** Весьма широко на практике (особенно в тех случаях, когда информации о матрице  $A$  недостаточно) применяется итерационный метод Зейделя в одной из двух форм:

$$\sum_{j=1}^i a_{ij} y_{k+1}^{(j)} + \sum_{j=i+1}^N a_{ij} y_k^{(j)} = f^{(i)}, \quad a_{ii} \neq 0, \quad i = 1, 2, \dots, N, \quad (9)$$

$$\sum_{j=1}^i a_{ij} y_k^{(j)} + \sum_{j=i+1}^N a_{ij} y_{k+1}^{(j)} = f^{(i)}, \quad i = 1, 2, \dots, N. \quad (10)$$

Из обеих формул компоненты вектора  $y_{k+1}$  находятся последовательно. Так, из (9) последовательно определяем  $y_{k+1}^{(1)}, y_{k+1}^{(2)}, \dots, y_{k+1}^{(N)}$ :

$$y_{k+1}^{(1)} = \frac{1}{a_{11}} \left( f^{(1)} - \sum_{j=2}^N a_{1j} y_k^{(j)} \right),$$

$$y_{k+1}^{(i)} = \frac{1}{a_{ii}} \left( f^{(i)} - \sum_{j=i+1}^N a_{ij} y_k^{(j)} - \sum_{j=1}^{i-1} a_{ij} y_{k+1}^{(j)} \right), \\ i = 2, \dots, N.$$

Пользуясь (10), находим последовательно для  $i = N, N-1, \dots, 1$

$$\begin{aligned} y_{k+1}^{(N)} &= \frac{1}{a_{NN}} \left( f^{(N)} - \sum_{j=1}^{N-1} a_{Nj} y_k^{(j)} \right), \\ y_{k+1}^{(i)} &= \frac{1}{a_{ii}} \left( f^{(i)} - \sum_{j=1}^{i-1} a_{ij} y_k^{(j)} - \sum_{j=i+1}^N a_{ij} y_{k+1}^{(j)} \right), \\ &\quad i = N-1, \dots, 1. \end{aligned}$$

Запишем этот метод в матричной (операторной) форме. Для этого представим матрицу  $A$  в виде суммы

$$A = A^- + D + A^+,$$

где  $D = (a_{ii})$  — диагональная матрица размера  $N \times N$ ,  $A^- = (a_{ij}^-)$  — нижняя треугольная (поддиагональная) матрица с нулями на главной диагонали,  $a_{ij}^- = 0$  при  $j \geq i$ ,  $a_{ij}^- = a_{ij}$  при  $j < i$ ,  $A^+ = (a_{ij}^+)$  — верхняя треугольная (наддиагональная) матрица с нулями на главной диагонали,  $a_{ij}^+ = 0$  при  $j \leq i$ ,  $a_{ij}^+ = a_{ij}$  при  $j > i$ . Из определения  $A^-, D, A^+$  следует, что

$$\begin{aligned} Dy^{(i)} &= a_{ii} y^{(i)}, \quad A^- y^{(i)} = \sum_{j=1}^{i-1} a_{ij} y^{(j)}, \\ A^+ y^{(i)} &= \sum_{j=i+1}^N a_{ij} y^{(j)}, \quad (A^+ + D) y^{(i)} = \sum_{j=i}^N a_{ij} y^{(j)}. \end{aligned}$$

Поэтому уравнение (10) можно записать в виде

$$(A^+ + D) y_{k+1}^{(i)} + (A^- y_k)^{(i)} = f^{(i)}, \quad i = 1, 2, \dots, N,$$

или, в векторной форме,

$$(A^+ + D) y_{k+1} + A^- y_k = f.$$

После очевидных преобразований

$$\begin{aligned} (A^+ + D) y_{k+1} + A^- y_k &= (A^+ + D)(y_{k+1} - y_k) + \\ &+ (A^- + (A^+ + D)) y_k = (A^+ + D)(y_{k+1} - y_k) + A y_k \end{aligned}$$

запишем метод Зейделя (10) в каноническом виде:

$$(D + A^+)(y_{k+1} - y_k) + A y_k = f, \quad k = 0, 1, 2, \dots \quad (11)$$

Сравнивая с (2), видим, что метод Зейделя (10) соответ-

ствует

$$B = D + A^+, \quad \tau = 1,$$

т. е. схема (11) является неявной. Однако, так как  $B = D + A^+$  — треугольная матрица, то итерация  $y_{k+1}$  находится по явным формулам. Аналогично записывается и другой вариант метода Зейделя:

$$(D + A^-)(y_{k+1} - y_k) + Ay_k = f, \quad k = 0, 1, \dots, \quad (12)$$

когда  $B = D + A^-$  — нижняя треугольная матрица. Далее, в п. 5 будет показано, что метод Зейделя сходится, если  $A$  — симметричная положительно определенная матрица.

**4. Метод верхней релаксации.** Чтобы ускорить итерационный процесс, можно привести метод Зейделя к методу верхней релаксации, вводя итерационный параметр  $\omega$ , так что

$$(D + \omega A^-) \frac{y_{k+1} - y_k}{\omega} + Ay_k = f, \\ k = 0, 1, \dots, \text{ для всех } y_0 \in H. \quad (13)$$

Сравнивая с (2), видим, что

$$B = D + \omega A^-, \quad \tau = \omega.$$

Преобразуем уравнение (13) к расчетному виду. Учитывая, что

$$(D + \omega A^-) \frac{y_{k+1} - y_k}{\omega} + Ay_k = \left( A^- + \frac{1}{\omega} D \right) y_{k+1} + \\ + \left( A - A^- - \frac{D}{\omega} \right) y_k = \left( A^- + \frac{1}{\omega} D \right) y_{k+1} + \left( A^+ + \left( 1 - \frac{1}{\omega} \right) D \right) y_k,$$

имеем

$$\left( A^- + \frac{1}{\omega} D \right) y_{k+1} + \left( A^+ + \left( 1 - \frac{1}{\omega} \right) D \right) y_k = f.$$

Отсюда находим

$$y_{k+1}^{(i)} = y_k^{(i)} + \frac{\omega}{a_{ii}} \left[ f^{(i)} - \sum_{j=1}^{i-1} a_{ij} y_{k+1}^{(j)} - \sum_{j=i}^N a_{ij} y_k^{(j)} \right], \quad i = 1, 2, \dots, N.$$

При  $\omega = 1$  получаем формулу метода Зейделя.

Скорость сходимости метода верхней релаксации зависит от параметра  $\omega$ . В п. 5 покажем, что для сходимости метода надо потребовать, чтобы  $0 < \omega < 2$ .

**5. Сходимость стационарных итерационных методов.** Метод Зейделя и метод верхней релаксации являются примерами неявных схем вида

$$B \frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \dots, \text{ для всех } y_0 \in H, \quad (14)$$

с несамосопряженным оператором  $B$ , имеющим обратный оператор  $B^{-1}$ . Метод (14) называется *стационарным итерационным методом*, так как  $B$  и  $\tau$  не зависят от номера итерации. Для существования обратного оператора  $B^{-1}$  достаточно потребовать положительности оператора  $B$ . Пусть  $B = D + \omega A^-$ . Так как  $A = A^* > 0$ , то  $(A^-y, y) = (A^+y, y)$ ,  $(A^+)^* = A^-$ , и значит,  $(Ay, y) = (Dy, y) + 2(A^-y, y)$ , т. е.

$$(A^-y, y) = \frac{1}{2} ((A - D)y, y).$$

Подставляя это выражение в формулу  $(By, y) = (Dy, y) + \omega(A^-y, y)$ , находим

$$(By, y) = \left(1 - \frac{1}{2}\omega\right)(Dy, y) + \omega(Ay, y) > 0,$$

если  $0 < \omega < 2$ .

Для погрешности  $z_k = y_k - u$  мы получаем однородное уравнение

$$B \frac{z_{k+1} - z_k}{\tau} + Az_k = 0, \quad k = 0, 1, 2, \dots, z_0 = y_0 - u. \quad (15)$$

**Теорема 1.** Пусть  $A$  — самосопряженный положительный оператор и выполнено условие

$$B > \frac{\tau}{2} A. \quad (16)$$

Тогда метод итераций (14) сходится в  $H_A$ , т. е.

$$\|z_k\|_A = \|y_k - u\|_A \rightarrow 0 \quad \text{при } k \rightarrow \infty.$$

**Доказательство.** Нам понадобится энергетическое тождество

$$2\tau \left( \left( B - \frac{\tau}{2} A \right) \frac{z_{k+1} - z_k}{\tau}, \frac{z_{k+1} - z_k}{\tau} \right) + \|z_{k+1}\|_A^2 = \|z_k\|_A^2, \quad (17)$$

где  $\|z\|_A^2 = (Az, z)$ . Сначала преобразуем уравнение (15)

к виду

$$\left( B - \frac{\tau}{2} A \right) \frac{z_{k+1} - z_k}{\tau} + \frac{1}{2} A (z_k + z_{k+1}) = 0, \quad (18)$$

подставив для этого  $z_k = \frac{1}{2} (z_{k+1} + z_k) - \frac{\tau}{2} \frac{(z_{k+1} - z_k)}{\tau}$ .

Умножая (18) скалярно на  $2\tau \left( \frac{z_{k+1} - z_k}{\tau} \right) = 2(z_{k+1} - z_k)$  и учитывая, что  $(Az_{k+1}, z_k) = (z_{k+1}, Az_k)$ , так как  $A = A^*$  и  $(A(z_k + z_{k+1}), z_{k+1} - z_k) = (Az_{k+1}, z_{k+1}) - (Az_k, z_k) + (Az_k, z_{k+1}) - (Az_{k+1}, z_k) = (Az_{k+1}, z_{k+1}) - (Az_k, z_k)$ , получаем (17).

Пусть выполнено условие  $B > \tau A / 2$ . Тогда первое слагаемое в левой части тождества (17) неотрицательно и  $\|z_{k+1}\|_A^2 \leq \|z_k\|_A^2$ . Отсюда следует, что  $0 \leq \|z_{k+1}\|_A \leq \|z_k\|_A \leq \dots \leq \|z_0\|_A$ , т. е. последовательность  $\{\|z_k\|_A\}$  — невозрастающая и ограниченная снизу нулем. Поэтому в силу теоремы Вейерштрасса  $\{\|z_k\|_A\}$  сходится при  $k \rightarrow \infty$ . Докажем, что  $\lim_{k \rightarrow \infty} \|z_k\|_A = 0$ .

Оператор  $P = B - \frac{\tau}{2} A$  положителен, а  $P_0 = B_0 - \frac{\tau}{2} A = \frac{1}{2} (P + P^*)$  положительно определен, т. е. существует такое число  $\delta > 0$  (см. гл. I § 4), что  $(Py, y) = (P_0y, y) \geq \delta \|y\|^2$  для всех  $y \in H$ .

Поэтому из тождества (17) получаем неравенство

$$\frac{2\delta}{\tau} \|z_{k+1} - z_k\|^2 + \|z_{k+1}\|_A^2 \leq \|z_k\|_A^2. \quad (*)$$

В силу сходимости  $\{\|z_k\|_A\}$  отсюда следует, что существует

$$\lim_{k \rightarrow \infty} \|z_{k+1} - z_k\| = 0. \quad (19)$$

Далее, из уравнения (15) находим

$$\begin{aligned} Az_k &= -\frac{1}{\tau} B (z_{k+1} - z_k), \quad z_k = -\frac{1}{\tau} A^{-1} B (z_{k+1} - z_k), \\ (Az_k, z_k) &= \frac{1}{\tau^2} (A^{-1} B (z_{k+1} - z_k), B (z_{k+1} - z_k)), \\ \|z_k\|_A^2 &\leq \frac{1}{\tau^2} \|A^{-1}\| \|B\|^2 \|z_{k+1} - z_k\|^2. \end{aligned} \quad (**)$$

Отсюда и из (19) заключаем, что  $\lim_{k \rightarrow \infty} \|z_k\|_A = 0$ .

**Замечание.** Из неравенств  $(*)$ ,  $(**)$  следует, что метод итераций (14) сходится при условии (16) со скоростью геометрической прогрессии,  $\|z_{k+1}\|_A^2 \leq \rho^2 \|z_k\|_A^2$ , где

$$\rho^2 = 1 - \frac{2\delta\tau}{\|A^{-1}\| \|B\|^2} < 1.$$

Применим теорему 1 для доказательства сходимости рассмотренных в пп. 2—4 итерационных методов.

**Метод простой итерации,  $B = E$ .** Учитывая, что  $E \geq \frac{1}{\|A\|} A$ , имеем

$$B - \frac{\tau}{2} A = E - \frac{\tau}{2} A \geq \left( \frac{1}{\|A\|} - \frac{\tau}{2} \right) A > 0$$

при  $\frac{1}{\|A\|} - \frac{\tau}{2} > 0$ . Метод простой итерации сходится при всех значениях  $\tau$ , удовлетворяющих неравенству  $\tau < 2/\|A\|$ .

**Метод Зейделя,  $B = D + A^-$ ,  $\tau = 1$ .** В этом случае

$$\begin{aligned} B - \frac{1}{2} A &= D + A^- - \frac{1}{2} (A^- + A^+ + D) = \\ &= \frac{D}{2} + \frac{1}{2} (A^- - A^+), \end{aligned}$$

$$\begin{aligned} \left( \left( B - \frac{1}{2} A \right) y, y \right) &= \frac{1}{2} (Dy, y) + \frac{1}{2} ((A^+ - A^-) y, y) = \\ &= \frac{1}{2} (Dy, y) > 0, \end{aligned}$$

если  $D > 0$ .

**Замечание.** Неравенство  $D > 0$  следует из условия  $A > 0$ . В самом деле, пусть  $A > 0$  и  $\xi = (\xi^1, 0, \dots, 0)$ ; тогда  $(A\xi, \xi) = (D\xi, \xi) = a_{11}(\xi^1)^2 > 0$ , т. е.  $a_{11} > 0$ . Аналогично убеждаемся, что  $a_{ii} > 0$ , и, следовательно,  $D > 0$ . Таким образом, метод Зейделя всегда сходится, если  $A$  — самосопряженный положительный оператор.

Чтобы получить оценку скорости сходимости, надо сделать более сильные предположения. Приведем одну из теорем.

**Теорема 2.** *Метод Зейделя сходится со скоростью геометрической прогрессии со знаменателем  $q < 1$ , если  $A = (a_{ij}) = A^* > 0$ , и*

$$\sum_{j \neq i}^{1+N} |a_{ij}| \leq q |a_{ii}|, \quad i = 1, 2, \dots, N, \quad q < 1. \quad (20)$$

В самом деле, для погрешности  $z_k = y_k - u$  имеем

$$a_{ii}z_{k+1}^{(i)} = - \sum_{j < i} a_{ij}z_{k+1}^{(j)} - \sum_{j > i} a_{ij}z_k^{(j)},$$

$$|a_{ii}| |z_{k+1}^{(i)}| \leq \sum_{j < i} |a_{ij}| |z_{k+1}^{(j)}| + \sum_{j > i} |a_{ij}| |z_k^{(j)}|.$$

Пусть  $\max |z_{k+1}^{(i)}|$  достигается при некотором  $i = i_0$ , так что

$$\|z_{k+1}\|_C = |z_{k+1}^{(i_0)}|, |a_{i_0 i_0}| \cdot \|z_{k+1}\|_C \leq \sum_{j < i_0} |a_{i_0 j}| \cdot \|z_{k+1}\|_C + \\ + \sum_{j > i_0} |a_{i_0 j}| \|z_k\|_C,$$

$$\|z_{k+1}\|_C \leq \left[ \sum_{j > i_0} |a_{i_0 j}| / \left( |a_{i_0 i_0}| - \sum_{j < i_0} |a_{i_0 j}| \right) \right] \|z_k\|_C.$$

В силу условия (20) имеем

$$\sum_{j > i_0} |a_{i_0 j}| \leq q |a_{i_0 i_0}| - \sum_{j < i_0} |a_{i_0 j}| < q \left( |a_{i_0 i_0}| - \sum_{j < i_0} |a_{i_0 j}| \right),$$

и, следовательно,

$$\|z_{k+1}\|_C \leq q \|z_k\|_C \leq q^{k+1} \|z_0\|_C,$$

что и требовалось доказать.

Условие (20) означает, что  $A = (a_{ij})$  — матрица с диагональным преобладанием.

Метод верхней релаксации,  $B = D + \omega A^-$ ,  $\tau = \omega$ . Найдем разность

$$B - \frac{\tau}{2} A = D + \omega A^- - \frac{\omega}{2} (A^- + A^+ + D) = \\ = \left( 1 - \frac{\omega}{2} \right) D + \frac{\omega}{2} (A^- - A^+)$$

и вычислим

$$\left( \left( B - \frac{\tau}{2} A \right) y, y \right) = \left( 1 - \frac{\omega}{2} \right) (Dy, y) > 0 \text{ при } 0 < \omega < 2.$$

Таким образом, метод верхней релаксации сходится при любых значениях  $\omega \in (0, 2)$ , если  $A = A^* > 0$ .

**6. Скорость сходимости неявного метода простой итерации.** Самого факта сходимости итераций недостаточно, чтобы судить о пригодности для практики итерационного метода. Нужна информация о скорости сходимости метода, т. е. фактически о числе итераций  $n = n_0(\varepsilon)$ , доста-

точных для решения задачи с требуемой точностью  $\varepsilon > 0$ . Число итераций  $n_0(\varepsilon)$  зависит от параметра  $\tau$ , который и следует выбирать из условия минимума числа итераций  $n = n(\varepsilon)$ , при котором выполняется условие  $\|y_n - u\|_D \leq \varepsilon \|y_0 - u\|_B$ , где  $D$  — некоторый оператор,  $D = D^* > 0$ .

Мы будем рассматривать неявную стационарную схему (неявную схему простой итерации)

$$B \frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \dots, \text{для всех } y_0 \in H, \quad (21)$$

где  $A$  и  $B$  — положительные самосопряженные операторы.

Методы Зейделя и верхней релаксации не принадлежат этому семейству схем, так как для них оператор  $B$  не является самосопряженным. Для поправки

$$w_k = B^{-1}r_k, \quad r_k = Ay_k - f$$

выполняется (так же, как и для погрешности  $z_k = y_k - u$ ) однородное уравнение

$$B \frac{w_{k+1} - w_k}{\tau} + Aw_k = 0, \quad k = 0, 1, \dots, \quad w_0 = B^{-1}(Ay_0 - f), \quad (22)$$

где  $r_k = Ay_k - f$  — невязка,  $w_k = B^{-1}r_k$  — поправка. В самом деле, из (21) находим

$$\begin{aligned} y_{k+1} &= y_k - \tau B^{-1}(Ay_k - f) = y_k - \tau w_k, \\ Ay_{k+1} - f &= Ay_k - f - \tau Aw_k, \quad r_{k+1} = r_k - \tau Aw_k. \end{aligned}$$

Так как  $r_k = B(B^{-1}r_k) = Bw_k$ , то отсюда следует (22).

Будем предполагать, что выполнены операторные неравенства

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0, \quad \gamma_2 \geq \gamma_1 > 0, \quad (23)$$

или

$$\gamma_1 (Bx, x) \leq (Ax, x) \leq \gamma_2 (Bx, x) \text{ для всех } x \in H, \quad (24)$$

где постоянные  $\gamma_1, \gamma_2$  известны.

**Теорема 3.** Пусть выполнены условия (23), (24). Тогда минимальное число итераций по методу (21) достигается при

$$\tau = \tau_0 = \frac{2}{\gamma_1 + \gamma_2}. \quad (25)$$

При этом выполняется неравенство

$$\|Ay_n - f\|_{B^{-1}} \leq \rho_0^n \|Ay_0 - f\|_{B^{-1}}, \quad n = 1, 2, \dots, \quad (26)$$

$$\rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2. \quad (27)$$

**Доказательство.** Для решения задачи (22) воспользуемся следующей оценкой (доказательство оценки приводится в гл. V)

$$\|w_n\|_B \leq \rho^n \|w_0\|_B \text{ при } \tau \leq \tau_0, \quad (28)$$

где  $\rho = 1 - \tau\gamma_1$ . Минимум  $\rho$  (при котором, число итераций, минимально) достигается при  $\tau = \tau_0$ :  $\rho \geq \rho_0 = 1 - \tau_0\gamma_1 = (1 - \xi)/(1 + \xi)$ . Остается учесть, что  $\|w_n\|_B = \|B^{-1}r_n\|_B = \|r_n\|_{B^{-1}}$ . Теорема доказана.

Требуя, чтобы  $\rho_0^n \leq \varepsilon$ , или  $(1/\rho_0)^n \geq 1/\varepsilon$ , получим оценку для числа итераций:

$$n \geq \ln(1/\varepsilon)/\ln(1/\rho_0). \quad (29)$$

**Замечание.** Функция  $\varphi(\xi) = \ln(1 + \xi)/(1 - \xi) - 2\xi$  положительна для всех  $0 < \xi < 1$ , так как  $\varphi'(\xi) = 2\xi^2/(1 - \xi^2) > 0$ ,  $\varphi(0) = 0$ ; поэтому  $1/\ln(1/\rho_0) < 1/(2\xi)$  и условие (29) выполнено, если

$$n \geq n_0(\varepsilon) = (1/(2\xi)) \ln 1/\varepsilon, \quad \xi = \gamma_1/\gamma_2 \quad (30)$$

( $n_0(\varepsilon)$  — вообще говоря, нецелое). Условие (30) более удобно для оценок. Оценка  $\rho_0^n \leq \varepsilon$ , очевидно, выполнена, если  $n_0(\varepsilon) \leq n < n_0(\varepsilon) + 1$ . Поэтому в качестве  $n$  достаточно брать целую часть числа  $n_0(\varepsilon) + 1$ .

**7. Модельная задача.** Сравнение различных итерационных методов будем проводить на следующей эталонной, или модельной задаче

$$\frac{v_{i-1} - 2v_i + v_{i+1}}{h^2} = -\tilde{f}_i, \quad i = 1, 2, \dots, N-1,$$

$$v_0 = \mu_1, v_N = \mu_2, h = \frac{1}{N}, \quad (31)$$

которая является разностной схемой для краевой задачи

$$\frac{d^2u}{dx^2} = -\tilde{f}(x), \quad 0 < x < 1, \quad u(0) = \mu_1, \quad u(1) = \mu_2.$$

Запишем систему уравнений сначала в матричной форме:

$$Av = f, \quad (32)$$

где  $v = (v^{(1)}, v^{(2)}, \dots, v^{(N-1)})$  — вектор размерности  $N - 1$  и  $A$  — трехдиагональная матрица размера  $(N - 1) \times (N - 1)$ :

$$A = -\frac{1}{h^2} \begin{bmatrix} -2 & 1 & 0 & \dots & & & 0 \\ 1 & -2 & 1 & \dots & & & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & & & 1 & -2 & 1 \\ 0 & \dots & & & 0 & 1 & -2 \end{bmatrix}.$$

Правая часть уравнения (32) имеет компоненты  $f_i = \tilde{f}_i$  при  $i = 2, 3, \dots, N - 2$ ,  $f_1 = \tilde{f}_1 + \mu_1/h^2$ ,  $f_{N-1} = \tilde{f}_{N-1} + \mu_2/h^2$ . Матрице  $A$  соответствует оператор  $A$ , действующий в пространстве  $H = \Omega$  сеточных функций, заданных во внутренних узлах сетки  $\omega_h = \{\bar{x}_i = ih, 0 < i < N\}$ . Пусть  $\Lambda v = v_{xx}$ ,  $v$  — сеточная функция, заданная на сетке  $\bar{\omega}_h = \{\bar{x}_i = ih, 0 \leq i \leq N\}$  и обращающаяся в нуль на границе при  $i = 0, N$ . Тогда можно написать

$$Av = -\Lambda v, \quad v \in \Omega = H, \quad \overset{\circ}{v} \in \overset{\circ}{\Omega}.$$

Введем в  $H = \Omega$ , как обычно, скалярное произведение

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h$$

и воспользуемся формулами (17), (56) из § 4 гл. I, в силу которых

$$(Av, w) = (v, Aw), \quad \text{т. е. } A = A^*,$$

$$(Av, v) \geq \delta \|v\|^2, \quad \delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}, \quad A \geq \delta E.$$

Далее, имеем

$$\|A\| = \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}.$$

Оценим число итераций для явной схемы простой итерации в случае модельной задачи. Имеем  $B = E$ ,  $\delta E \leq A \leq \Delta E$ , т. е.

$$\gamma_1 = \delta, \quad \gamma_2 = \Delta, \quad \xi = \frac{\gamma_1}{\gamma_2} = \operatorname{tg}^2 \frac{\pi h}{2} \approx \frac{\pi^2 h^2}{4}.$$

Для числа итераций имеем

$$n(\epsilon) \geq n_0(\epsilon) = \frac{\ln 1/\epsilon}{2\xi} \approx \frac{2}{10h^2} \ln \frac{1}{\epsilon}.$$

Зададим  $\varepsilon = \frac{1}{2} \cdot 10^{-4} \approx e^{-10}$ ; тогда  $n_0(\varepsilon) \approx \frac{2}{h^2} = 2N^2$ .

В частности, число итераций:

$$\begin{aligned} n_0(\varepsilon) &\approx 200 \quad \text{при } N = 10, \\ n_0(\varepsilon) &\approx 20\,000 \quad \text{при } N = 100. \end{aligned}$$

Метод простой итерации сильно зависит от числа уравнений  $N$  ( $n_0(\varepsilon) \approx N^2$ ). Ниже приводятся методы (см. §§ 4, 5), для которых эта зависимость  $n$  от  $N$  будет более слабой ( $n_0(\varepsilon) \approx N$  и  $n_0(\varepsilon) \approx \sqrt{N}$ ).

Задача (31) является типичной, так как аналогичное разностное уравнение моделирует разностное уравнение Лапласа в двумерном и трехмерном случаях, а число итераций практически не зависит от числа измерений (зависит только от  $h$ ).

**8. Трехслойная схема.** Если  $y_{k+1}$  вычисляется по двум предыдущим итерациям  $y_k$  и  $y_{k-1}$ , то итерационный метод называют *двухшаговым* (или *трехслойным*). Приведем пример трехслойной итерационной схемы. Явная трехслойная схема с постоянными параметрами обычно записывается в виде

$$y_{k+1} = (1 + \alpha)(E - \tau_0 A)y_k - \alpha y_{k-1} + (1 + \alpha)\tau_0 f, \quad k = 1, 2, \dots \quad (33)$$

Первая итерация вычисляется по явному методу простой итерации:

$$y_1 = (E - \tau_0 A)y_0 + \tau_0 f \quad \text{для всех } y_0 \in H, \quad (34)$$

где

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \alpha = \rho_1^2, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \quad (35)$$

$\gamma_1, \gamma_2 > 0$  — границы спектра оператора  $A = A^*$ :  $\gamma_1 E \leqslant A \leqslant \gamma_2 E$ .

Можно показать, что для метода (33), (34) число итераций находится из условия

$$q_n = \rho_1^n \left( 1 + \frac{1 - \rho_1^2}{1 + \rho_1^2} n \right) \leqslant \varepsilon.$$

Отсюда видно, что

$$n_0(\varepsilon) \approx \frac{c_0}{2\sqrt{\xi}} \ln \frac{1}{\varepsilon}, \quad 1 < c_0 < 2. \quad (36)$$

Для модельной задачи  $\sqrt{\xi} \approx \pi h/2$  и

$$n_0(\varepsilon) \approx \frac{c_0}{\pi h} \ln \frac{1}{\varepsilon} \approx c_0 \frac{0,32}{h} \ln \frac{1}{\varepsilon} \approx c_0 \cdot 3,2N \quad \text{при } \varepsilon \approx e^{-10}.$$

Число итераций:

$$n_0(\varepsilon) \approx 32 \div 60 \quad \text{при } N = 10,$$

$$n_0(\varepsilon) \approx 320 \div 620 \quad \text{при } N = 100,$$

т. е. значительно меньше, чем для простой итерации.

Неявная трехслойная схема имеет вид

$$By_{k+1} = (1 + \alpha)(B - \tau_0 A)y_k - \alpha By_{k-1} + (1 + \alpha)\tau_0 f,$$

$$k = 1, 2, \dots,$$

$$By_1 = By_0 - \tau_0 A y_0 + \tau_0 f \quad \text{для всех } y_0 \in H.$$

Если  $B = B^* > 0$ , выполнены неравенства (23), (24) и  $\alpha$ ,  $\tau_0$  вычисляются по формулам (35), то для числа итераций оценка (36) верна и в этом случае.

#### § 4. Двухслойная итерационная схема с чебышевскими параметрами

**1. Постановка задачи.** Пусть дано уравнение

$$Au = f, \quad A: H \rightarrow H. \quad (1)$$

Рассмотрим итерационную схему с переменными параметрами  $\{\tau_k\}$ :

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, 2, \dots, \quad \text{для всех } y_0 \in H. \quad (2)$$

Однородному уравнению

$$B \frac{z_{k+1} - z_k}{\tau_{k+1}} + Az_k = 0, \quad k = 0, 1, 2, \dots, \quad z_0 = y_0 - u, \quad (3)$$

удовлетворяет не только погрешность  $z_k = y_k - u$ , но и поправка  $w_k = B^{-1}(Ay_k - f)$  ( $k = 0, 1, \dots$ ) с начальным условием  $w_0 = B^{-1}(Ay_0 - f)$ . Условие окончания итераций имеет вид

$$\|z_n\|_D \leq \varepsilon \|z_0\|_D, \quad \text{или} \quad \|w_n\|_D \leq \varepsilon \|w_0\|_D. \quad (4)$$

Из (3) видно, что

$$z_{k+1} = S_{k+1}z_k, \quad S_{k+1} = E - \tau_{k+1}B^{-1}A, \quad (5)$$

где  $S_{k+1}$  — оператор перехода со слоя  $k$  на слой  $k+1$ . Исключая  $z_k, z_{k-1}, \dots, z_1$ , найдем при  $k=n-1$ :

$$z_n = T_n z_0, \quad T_n = S_n S_{n-1} \dots S_2 S_1,$$

где  $T_n$  — разрешающий оператор схемы (3). Отсюда следует, что

$$\|z_n\|_D \leq q_n \|z_0\|_D, \quad q_n = \|T_n\|_D. \quad (6)$$

Условие окончания итерации (4) выполнено, если

$$q_n = \|T_n\|_D \leq \varepsilon. \quad (7)$$

Для оценки числа итераций  $n = n(\varepsilon)$  надо получить неравенство (7).

Рассмотрим явную схему (2)

$$\frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, 2, \dots, \quad (8)$$

причем задано любое  $y_0 \equiv H$ , и выберем параметры  $\tau_1, \tau_2, \dots, \tau_n$  из условия  $\min n(\varepsilon)$ . При этом предполагается, что

$$A = A^* > 0, \quad \gamma_1 E \leq A \leq \gamma_2 E, \quad \gamma_1 > 0.$$

Для невязки  $r_k = A y_k - f$  выполняется однородное уравнение

$$\frac{r_{k+1} - r_k}{\tau_{k+1}} + A r_k = 0, \quad k = 0, 1, 2, \dots, r_0 = A y_0 - f,$$

или

$$r_{k+1} = S_{k+1} r_k, \quad S_{k+1} = E - \tau_{k+1} A.$$

Отсюда найдем

$$r_n = T_n r_0, \quad T_n = S_1 S_2 \dots S_n.$$

Разрешающий оператор  $T_n$  есть полином степени  $n$  относительно  $A$ :

$$T_n = P_n(A) = (E - \tau_1 A)(E - \tau_2 A) \dots (E - \tau_n A)$$

с коэффициентами, зависящими только от  $\tau_1, \tau_2, \dots, \tau_n$ .

Для нахождения  $\tau_1, \tau_2, \dots, \tau_n$  получаем оценку

$$\|r_n\| \leq \|P_n(A)\| \|r_0\|.$$

Надо найти такие  $\tau_1, \tau_2, \dots, \tau_n$ , при которых  $\|P_n(A)\|$  минимальна, и оценить эту норму через постоянные  $\gamma_1$  и  $\gamma_2$ . Приведем без доказательства решение этой задачи.

Обозначим через  $\mathfrak{M}_n = \left\{ -\cos \frac{2i-1}{2n} \pi, i = 1, 2, \dots, n \right\}$  множество нулей полинома Чебышева  $T_n(x) = \cos(n \arccos x)$  на отрезке  $-1 \leq x \leq 1$ , а через  $\{\mu_k\}$  — любую последовательность этих нулей,  $\mu_k \in \mathfrak{M}_n$ . Минимальное число итераций  $n(\varepsilon)$  достигается при значениях параметров

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad k = 1, 2, \dots, n,$$

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \quad (9)$$

При этом выполняется оценка

$$\|Ay_h - f\| \leq q_n \|Ay_0 - f\|, \quad q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}. \quad (10)$$

Схема (8) с итерационными параметрами (9) называется *чебышевской итерационной схемой*.

Требование  $q_n \leq \varepsilon$  или  $2\rho_1^n \leq \varepsilon(1 + \rho_1^{2n})$  выполнено, если  $\rho_1^n \leq \varepsilon/2$ , или

$$n(\varepsilon) \geq \ln \frac{2}{\varepsilon} / \ln \frac{1}{\rho_1}. \quad (11)$$

Замечая (ср. § 3, п. 6), что  $\ln \frac{1}{\rho_1} = \ln \frac{1 + \sqrt{\xi}}{1 - \sqrt{\xi}} > 2\sqrt{\xi}$ , заменим (11) более сильным требованием:

$$n(\varepsilon) > n_0(\varepsilon) = \frac{1}{2\sqrt{\xi}} \ln \frac{2}{\varepsilon}, \quad (12)$$

удобным для проверки. Из (12), очевидно, следует (10), и  $q_n \leq \varepsilon$ .

Сравним по числу итераций схему (8) с указанным набором параметров и метод простой итерации на примере модельной задачи из § 3. В этом случае  $\xi \approx \pi^2 h^2/4$ ,  $\sqrt{\xi} = \pi h/2$ .

Для метода простой итерации

$$n_0^{(1)}(\varepsilon) \approx 2/h^2 \quad \text{при} \quad \varepsilon = 10^{-4}.$$

Для чебышевской схемы

$$n_0^{(2)}(\varepsilon) = 3.4/h \quad \text{при} \quad \varepsilon = 10^{-4}.$$

Отсюда видно, что

$$n_0^{(2)} \approx 34, \quad n_0^{(1)} \approx 200 \quad \text{при} \quad N = 10 \quad (h = 1/10).$$

$$n_0^{(2)} \approx 340, \quad n_0^{(1)} \approx 20\,000 \quad \text{при} \quad N = 100 \quad (h = 1/100).$$

**2. Обоснование оптимального выбора параметров.** Переайдем к доказательству оценки (10) в случае итерационных параметров (9). Нам необходимо найти  $\min_{\{\tau_k\}} \|P_n(A)\|$ .

Полином

$$\begin{aligned} P_n(A) &= \prod_{k=1}^n (E - \tau_k A) = \\ &= c_0 + c_1 A + \dots + c_k A^k + \dots + c_n A^n, \quad c_0 = 1, \quad P_n(0) = 1 \end{aligned}$$

является самосопряженным оператором. Пусть  $\xi_s, \lambda_s$  ( $s = 1, 2, \dots, N$ ) — собственные функции и собственные значения оператора  $A$ :

$$A \xi_s = \lambda_s \xi_s, \quad (\xi_s, \xi_m) = \delta_{sm}, \quad s, m = 1, 2, \dots, N.$$

Оператор  $A^k$  имеет те же собственные функции и собственные значения  $\lambda_s^k$ :

$$A^k \xi_s = \lambda_s^k \xi_s. \quad (13)$$

Умножая (13) на  $c_k$  и суммируя по  $k = 0, 1, \dots, n$  ( $c_0 = 1$ ), получим

$$P_n(A) \xi_s = \sum_{k=0}^n c_k A^k \xi_s = \sum_{k=0}^n c_k \lambda_s^k \xi_s = P_n(\lambda_s) \xi_s.$$

Сравнивая это с  $P_n(A) \xi_s = \lambda_s (P_n(A)) \xi_s$ , видим, что

$$\lambda(P_n(A)) = P_n(\lambda(A)).$$

Собственные значения операторного полинома  $P_n(A)$  определяются как полиномы  $P_n(\lambda)$  от соответствующих собственных значений оператора  $A$ , а собственные функции — те же, что и у оператора  $A$ . В силу самосопряженности оператора  $P_n(A)$  его норма равна наибольшему по модулю собственному значению

$$\|P_n(A)\| = \max_{1 \leq s \leq N} |P_n(\lambda_s)|.$$

Собственные значения  $\lambda_s$  оператора  $A$  лежат на отрезке  $[\gamma_1, \gamma_2]$ :  $\gamma_1 \leq \lambda_s \leq \gamma_2$ . Очевидно, что

$$\max_{1 \leq s \leq N} |P_n(\lambda_s)| \leq \max_{\gamma_1 \leq x \leq \gamma_2} |P_n(x)|,$$

где непрерывный аргумент  $x$  принимает все значения на отрезке  $[\gamma_1, \gamma_2]$ , и, следовательно, задача о минимуме  $\|P_n(A)\|$  сводится к задаче о минимаксе полинома  $P_n(x)$ , т. е. о нахождении  $\min_{\{\tau_k\}} \max_{\gamma_1 \leq x \leq \gamma_2} |P_n(x)|$ .

Отобразим отрезок  $[\gamma_1, \gamma_2]$  на отрезок  $[-1, 1]$ , полагая

$$x = \frac{1}{2} [(\gamma_2 - \gamma_1) t + \gamma_2 + \gamma_1], \quad -1 \leq t \leq 1,$$

при  $\gamma_1 \leq x \leq \gamma_2$ . (14)

Тогда  $P_n(t) = \tilde{P}_n(t)$ . Условие нормировки  $P_n(0) = 1$  принимает вид

$$\tilde{P}_n(t_0) = 1, \quad t_0 = -1/\rho_0. \quad (15)$$

Итак, требуется найти полином, наименее уклоняющийся от нуля на отрезке  $-1 \leq t \leq 1$ , так чтобы  $\max |\tilde{P}_n(t)|$  был минимальен при дополнительном условии нормировки (15). Таким полиномом является полином

$$\tilde{P}_n(t) = \frac{T_n(t)}{T_n(t_0)}, \quad (16)$$

где  $T_n(t)$  — полином Чебышева,

$$T_n(t) = \cos(n \arccos t) \quad \text{при } |t| \leq 1, \quad (17)$$

$$T_n(t) = \frac{1}{2} [(t + \sqrt{t^2 - 1})^n + (t - \sqrt{t^2 - 1})^n] \quad \text{при } |t| > 1. \quad (18)$$

Полином Чебышева имеет нули

$$t_i = \cos \frac{2i-1}{2n} \pi, \quad i = 1, 2, \dots, n. \quad (19)$$

Полином  $P_n(x) = (1 - \tau_1 x)(1 - \tau_2 x) \dots (1 - \tau_n x)$  имеет нули  $x_i = 1/\tau_i$ .

Требуя, чтобы корни этих полиномов совпадали, и учитывая связь (14) между  $x$  и  $t$ , получаем  $2 = [(\gamma_1 + \gamma_2) + (\gamma_2 - \gamma_1)t_i]\tau_i$ , откуда следует

$$\tau_i = 2/[\gamma_2 + \gamma_1 + (\gamma_2 - \gamma_1)t_i], \quad i = 1, 2, \dots, n. \quad (20)$$

Эта формула сохраняет силу при любом способе упорядо-

дочения нулей полинома Чебышева; например, вместо (19) можно положить  $t_i = -\cos \frac{2i-1}{2n} \pi$ . Имея это в виду, приходим к формуле (9). Заметим, что если  $n=1$ , то получаем  $\tau_1 = \tau_0$  — оптимальный параметр метода простой итерации.

Итак, параметры  $\tau_1, \tau_2, \dots, \tau_n$  определены согласно (9). Найдем теперь

$$\begin{aligned} q_n &= \max_{\gamma_1 \leq x \leq \gamma_2} |P_n(x)| = \max_{-1 \leq t \leq 1} |\tilde{P}_n(t)| = \\ &= \max_{-1 \leq t \leq 1} \left| \frac{T_n(t)}{T_n(t_0)} \right| = \frac{1}{|T_n(t_0)|}, \end{aligned}$$

так как  $\max_{-1 \leq t \leq 1} |T_n(t)| = 1$ . Имеем  $|t_0| > 1$ ; поэтому для  $|T_n(t_0)|$  воспользуемся формулой (18) при  $t = t_0$ . Преобразуем входящие в нее выражения:

$$\begin{aligned} |t_0| \pm \sqrt{t_0^2 - 1} &= \frac{1}{\rho_0} \pm \sqrt{\frac{1}{\rho_0^2} - 1} = \frac{1}{\rho_0} [1 \pm \sqrt{1 - \rho_0^2}] = \\ &= \frac{1}{\rho_0} \left( 1 \pm \frac{2\sqrt{\xi}}{1-\xi} \right) = \frac{1}{\rho_0} (1 \pm \sqrt{\xi})^2 / (1 - \xi) = \\ &= (1 \pm \sqrt{\xi})^2 / (1 - \xi) = (1 \pm \sqrt{\xi}) / (1 \mp \sqrt{\xi}), \end{aligned}$$

так что  $|t_0| + \sqrt{t_0^2 - 1} = \frac{1}{\rho_1}$ ,  $|t_0| - \sqrt{t_0^2 - 1} = \rho_1$ , и

$$|T_n(t_0)| = \frac{1}{2} \left( \frac{1}{\rho_1^n} + \rho_1^n \right) = \frac{1 + \rho_1^{2n}}{2\rho_1^n} = \frac{1}{q_n}.$$

Оценка (10) доказана.

**3. Вычислительная устойчивость и упорядочение параметров.** Итерационный метод (8) с чебышевским набором параметров  $\{\tau_k\}$  иногда называют *методом Ричардсона*. Известен он давно, однако до недавнего времени почти не использовался на практике из-за *вычислительной неустойчивости*. Поясним это понятие на примере. Возьмем систему уравнений

$$\begin{aligned} u(i-1) - 2u(i) + u(i+1) &= 0, \quad i = 1, 2, \dots, N-1, \\ u(0) = 1, \quad u(N) = 0. \end{aligned} \tag{21}$$

Ее решением является  $u(i) = 1 - x_i$ ,  $x_i = ih$ ,  $h = 1/N$ . Будем искать решение этой задачи чебышевским итера-

ционным методом для  $N = 20$ . Значение  $n_0(\varepsilon)$  мы можем вычислить. Оно может быть нецелым. Выбираем ближайшее целое  $n \geq n_0$ . Для данных  $N$  и  $\varepsilon$  имеем  $n(\varepsilon) = 64$ . Зная

$$\gamma_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}, \quad \gamma_2 = \frac{4}{h^2} \cos^2 \frac{\pi h}{2},$$

$$h = \frac{1}{N}, \quad \xi = \operatorname{tg}^2 \frac{\pi h}{2} \approx \frac{\pi^2 h^2}{4} \approx 0,006,$$

можно вычислить  $\tau_h$  по формуле (20). В качестве начального приближения берется функция

$$y_0^{(i)} = \begin{cases} 1, & i = 0, \\ 0, & i > 0. \end{cases}$$

Оказывается, что для метода (8), (9) небезразлично, в каком порядке берутся нули  $\mu_k$  полинома Чебышева. Рассмотрим два способа нумерации нулей:

$$\alpha_1) \mu_k = \cos \frac{2k-1}{2n} \pi, \quad k = 1, 2, \dots, n,$$

$$t_1 = \cos \frac{\pi}{2n}, \quad t_n = -\cos \frac{\pi}{2n},$$

$$\alpha_2) \mu_k = -\cos \frac{2k-1}{2n} \pi.$$

Результаты расчетов приводятся в табл. 1.

Таблица 1

$k$	Набор $\alpha$ ,		$k$	Набор $\alpha$ ,	
	$\Delta_k = \max_{x_i}  y_k(x_i) - y_{k-1}(x_i) $			$\Delta_k = \max_{x_i}  y_k(x_i) - y_{k-1}(x_i) $	
53	0,12		1	39,6	
55	27		2	$2,6 \cdot 10^3$	
57	$1,9 \cdot 10^4$		4	$8,2 \cdot 10^8$	
59	$3,7 \cdot 10^7$		7	$3,3 \cdot 10^{11}$	
60	$2,6 \cdot 10^9$		9	$1,2 \cdot 10^{14}$	
61	$2,5 \cdot 10^{11}$		11	$1,9 \cdot 10^{16}$	
62	$3,3 \cdot 10^{13}$		12	Авост	
63	$5 \cdot 10^{15}$				
64	Авост				

При меньших значениях  $N$  и  $n$  может оказаться, что рост промежуточных значений  $y_k$  не приводит к авосту, однако происходит накопление ошибок округления, и после  $n$  итераций условие окончания итераций  $\|Ay_k - f\| \leq \varepsilon \|Ay_0 - f\|$  не выполняется.

Эти две особенности вычислительного процесса — рост промежуточных значений, приводящих к авасту, и накопление ошибок округления — мы характеризуем одним термином — *вычислительная неустойчивость*. Причина вычислительной неустойчивости чебышевского метода в том, что нормы  $\|S_{k+1}\|$  оператора перехода  $S_{k+1} = E - \tau_{k+1}A$  для некоторых итераций больше единицы, а вычислительный процесс является реальным, т. е. имеются ограничения снизу и сверху на допустимые числа (есть машинный нуль и машинная бесконечность) и в каждом акте вычислений появляются ошибки округлений.

Вычислим норму для  $S_k = E - \tau_k A$ . Так как  $S_k^* = S_k$ , то  $\|S_{k+1}\| = \sup_{\|x\|=1} |(S_{k+1}x, x)|$ . Из условий  $\gamma_1 E \leq A \leq \gamma_2 E$  следует  $(\tau_{k+1}\gamma_1 - 1)E \leq \tau_{k+1}A - E \leq (\tau_{k+1}\gamma_2 - 1)E$ . Подставляя сюда выражение для  $\tau_{k+1}$  и учитывая, что  $1 - \tau_0\gamma_1 = \tau_0\gamma_2 - 1 = \rho_0$ , получаем

$$-\frac{\rho_0(1 - \mu_k)}{1 + \rho_0\mu_k} E \leq \tau_{k+1}A - E \leq \frac{\rho_0(1 + \mu_k)}{1 + \rho_0\mu_k} E.$$

Отсюда находим

$$\|S_{k+1}\| = \|\tau_{k+1}A - E\| = \begin{cases} \frac{\rho_0(1 + \mu_k)}{1 + \rho_0\mu_k} & \text{при } \mu_k > 0, \\ \frac{\rho_0(1 - \mu_k)}{1 + \rho_0\mu_k} & \text{при } \mu_k < 0, \end{cases}$$

так что  $\|S_{k+1}\| < 1$  для всех  $\mu_k > 0$  и  $\|S_{k+1}\| > 1$  при  $\mu_k < -(1 - \rho_0)/(2\rho_0)$ . Так как

$$-\cos \frac{\pi}{2n} \leq \mu_k \leq -\cos \frac{(2n-1)\pi}{2n} = \cos \frac{\pi}{2n}, \quad k = 1, 2, \dots, n,$$

то для большего числа номеров  $k$  имеем  $\|S_k\| > 1$ , и если подряд используется много параметров  $\tau_k$ , для которых  $\|S_k\| > 1$ , то происходит накопление погрешности округлений и рост итерационных приближений, что и приводит к вычислительной неустойчивости.

Чтобы ослабить этот эффект, естественно попытаться чередовать параметры  $\tau_k$ , для которых  $\|S_k\| > 1$ , с параметрами, для которых  $\|S_k\| < 1$ . На этом пути и проводится построение такой последовательности параметров  $\{\tau_k\}$ , для которой сходимость итераций носит монотонный характер и вычислительная неустойчивость отсутствует. Имеется правило такого упорядочения нулей

$t_i = -\cos \frac{2i-1}{2n} \pi$  полинома Чебышева, а тем самым и параметров  $\{\tau_k\}$ , для любого  $n$ , при котором имеет место вычислительная устойчивость.

Мы приведем это правило для случая, когда  $n$  есть степень числа 2,  $n = 2^p$ ,  $p > 0$  — целое число \*). Обозначим упорядоченное по этому правилу множество нулей  $t_i$  через

$$\mathfrak{M}_n^* = \left\{ -\cos \beta_i, \beta_i = \frac{\pi}{2n} \theta_i^{(n)}, i = 1, 2, \dots, n \right\}, n = 2^p,$$

где  $\theta_i^{(n)}$  — одно из нечетных чисел 1, 3, 5, ...,  $2n-1$ . Задача сводится к упорядочению множества  $n$  нечетных чисел:  $\theta_n = \{\theta_1^{(n)}, \theta_2^{(n)}, \dots, \theta_n^{(n)}\}$ . Исходя из множества  $\theta_1 = \{1\}$ , построим множество  $\theta_n^* = \theta_{2^p}^*$  по формулам

$$\theta_{2i-1}^{(2m)} = \theta_i^{(m)},$$

$$\theta_{2i}^{(2m)} = 4m - \theta_{2i-1}^{(2m)}, \quad i = 1, 2, \dots, m; m = 1, 2, \dots, 2^{p-1},$$

если  $\theta_i^{(m)}$  известны. Соответствующую последовательность параметров  $\{\tau_k^*\}$  будем называть *устойчивым набором*. Пусть, например,  $n = 16 = 2^4$ . Последовательно находим  $\theta_1 = \{1\}$ ,  $\theta_2 = \{1, 3\}$ ,  $\theta_4 = \{1, 7, 3, 5\}$ ,  $\theta_8 = \{1, 15, 7, 9, 3, 13, 5, 11\}$ ,  $\theta_{16} = \{1, 31, 15, 17, 7, 25, 9, 23, 3, 29, 13, 19, 5, 27, 11, 21\}$ . При переходе от  $\theta_m$  к  $\theta_{2m}$  достаточно после каждого  $\theta_{2i-1}^{(m)}$  поставить число, равное  $4m - \theta_{2i-1}^{(m)}$  (нумерация соответствует  $\theta_m$ ). «Устойчивая» последовательность  $\theta_n^*$  не зависит от задачи. Сходимость итераций для этого набора параметров  $\{\tau_k^*\}$  носит немонотонный характер, но колебания здесь невелики и в конечном счете затухают.

Приведем результаты расчетов для задачи (21) по схеме (8), (9) с устойчивым набором параметров  $\{\tau_k^*\}$ :

$k$	1	4	8	16	24	32	48	50	62
$\Delta_k$	39,6	4,7	1,1	0,2	0,1	0,04	$1,5 \cdot 10^{-3}$	$6,7 \cdot 10^{-3}$	$8,7 \cdot 10^{-5}$

**4. Неявные схемы.** Метод Зейделя и метод верхней релаксации сходятся быстрее явного метода простой итерации. Поэтому переход к неявным схемам оправдывает

\*) Правило упорядочения  $\{\tau_k\}$  для любого  $n$  можно найти в [6, 9].

себя. Как нужно выбирать оператор  $B$ ? Основным является общее требование минимума действий  $Q(\varepsilon)$  для нахождения решения с точностью  $\varepsilon > 0$ , которое сводится к двум требованиям: 1) о минимуме числа итераций, которое зависит как от  $B$ , так и от выбора  $\{\tau_k\}$ ; 2) о минимуме числа действий для решения уравнения

$$By_{k+1} = F_k.$$

(экономичность оператора  $B$ ). Примером может быть треугольный оператор, соответствующий треугольной матрице.

Покажем теперь, что результаты, полученные выше для явной схемы, можно перенести на неявную схему. Рассмотрим неявную схему

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \text{для всех } y_0 \in H, \quad (22)$$

где  $A = A^* > 0$ ,  $B = B^* > 0$  и

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0. \quad (23)$$

Выбирая итерационные параметры  $\{\tau_k^*\}$  по формулам (9) и упорядочивая их в соответствии с предыдущим пунктом, получим для решения задачи (22) оценку

$$\|Ay_n - f\|_{B^{-1}} \leq q_n \|Ay_0 - f\|_{B^{-1}}, \quad q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}},$$

$$\rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \quad (24)$$

где  $\gamma_1$  и  $\gamma_2$  — числа, входящие в (23). Для числа итераций  $n = n(\varepsilon)$  верны оценки (11) и (12). Чтобы убедиться в этом, достаточно свести задачу (22) к эквивалентной задаче для явной схемы

$$\frac{x_{k+1} - x_k}{\tau_{k+1}} + Cx_k = 0, \quad k = 0, 1, \dots, \quad x_0 = B^{1/2}w_0, \quad (25)$$

где  $x_k = B^{1/2}w_k$ ,  $C = B^{-1/2}AB^{-1/2}$  — самосопряженный положительный оператор с границами спектра  $\gamma_1$  и  $\gamma_2$ :

$$\gamma_1 E \leq C \leq \gamma_2 E. \quad (26)$$

В самом деле, так как  $B = B^* > 0$ , то существует  $B^{1/2} = (B^{1/2})^* > 0$ . Действуя оператором  $B^{-1/2}$  на уравнении

ние (22), получаем (25) для  $x_k = B^{1/2}w_k$ . Обратный ход рассуждений очевиден. Остается доказать эквивалентность неравенств (23) и (26). Рассмотрим функционал

$$\begin{aligned} J &= ((A - \gamma B)y, y) = (Ay, y) - \gamma(By, y) = \\ &= (AB^{-1/2}(B^{1/2}y), B^{-1/2}(B^{1/2}y)) - \gamma(B^{1/2}y, B^{1/2}y) = \\ &= (Cx, x) - \gamma(x, x) = ((C - \gamma E)x, x), \end{aligned}$$

где  $x = B^{1/2}y$ . Так как  $y$  (а значит, и  $x$ ) — произвольный вектор из  $H$ , то из равенства

$$J = ((A - \gamma B)y, y) = ((C - \gamma E)x, x) \quad (27)$$

следует, что операторы  $A - \gamma B$  и  $C - \gamma E$  имеют одинаковые знаки. Если, например,  $A - \gamma_1 B \geq 0$ , то при  $\gamma = \gamma_1$  равенство (27) дает  $C - \gamma_1 E \geq 0$  и т. д.

Для явной схемы имеем оценку  $\|x_n\| \leq q_n \|x_0\|$ . Подставляя сюда  $x_k = B^{1/2}w_k = B^{-1/2}r_k$ ,  $r_k = Ay_k - f$ , получаем оценку (24).

Для методов Зейделя и верхней релаксации  $B \neq B^*$ , и потому нельзя воспользоваться чебышевским набором параметров.

## § 5. Попеременно-треугольный метод

**1. Попеременно-треугольный метод.** Будем рассматривать неявную итерационную схему

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots \quad (1)$$

Если оператор  $B$  представляет собой произведение конечного числа экономичных операторов, то он также экономичен. Так, экономичным является оператор  $B = B_1 B_2$ , равный произведению треугольных операторов  $B_1$  и  $B_2$ .

Рассмотрим так называемый *попеременно-треугольный* метод — метод (1), для которого оператор  $B$  имеет вид

$$B = (D + \omega R_1)D^{-1}(D + \omega R_2), \quad (2)$$

где  $D = D^* > 0$ ,  $R_1^* = R_2$ ,  $R_1 + R_2 = R$ ,  $R = R^* > 0$ ,  $\omega > 0$  — параметр.

Покажем, что оператор  $B$  является самосопряженным и положительным, т. е. схема (1) с оператором (2) принадлежит исходному семейству схем (2) из § 3 и можно пользоваться всеми полученными ранее результатами об-

щей теории. В самом деле,

$$\begin{aligned} (By, v) &= ((D + \omega R_1) D^{-1} (D + \omega R_2) y, v) = \\ &= ((D + \omega R_2) y, D^{-1} (D + \omega R_2) v) = \\ &= (y, (D + \omega R_1) D^{-1} (D + \omega R_2) v), \end{aligned}$$

а значит,  $(By, v) = (y, Bv)$ , т. е.  $B = B^*$ . Далее, находим  $(By, y) = ((D + \omega R_2) y, D^{-1} (D + \omega R_2) y) = \| (D + \omega R_2) y \|_D^2 - 1 > 0$ , т. е.  $B = B^* > 0$ .

Оператору  $R$  соответствует матрица  $R = (r_{ij})$ . В качестве матриц  $R_1$  и  $R_2$  можно взять нижнюю и верхнюю треугольные матрицы, т. е.

$$\begin{aligned} R_1 = (r_{ij}^-), \quad r_{ij}^- &= \begin{cases} r_{ii}/2, & j = i, \\ r_{ij}, & j < i, \\ 0, & j > i; \end{cases} \\ R_2 = (r_{ij}^+), \quad r_{ij}^+ &= \begin{cases} r_{ii}/2, & j = i, \\ r_{ij}, & j > i, \\ 0, & j < i. \end{cases} \end{aligned}$$

Если  $R$  — симметричная матрица,  $r_{ji} = r_{ij}$ , то  $R_1$  и  $R_2$  взаимно сопряжены,  $R_2 = R_1^*$ .

В качестве  $D = (d_{ij})$  возьмем диагональную матрицу. Тогда  $D + \omega R_1$  — нижняя треугольная, а  $D + \omega R_2$  — верхняя треугольная матрица. Таким образом, процесс итераций сводится к попаременному обращению нижней и верхней треугольных матриц (отсюда и название метода). В самом деле, для каждой итерации надо решать уравнение

$$By_{k+1} = (D + \omega R_1) D^{-1} (D + \omega R_2) y_{k+1} = F_k. \quad (3)$$

Обозначая  $D^{-1} (D + \omega R_2) y_{k+1} = \bar{y}_{k+1}$ , получаем

$$(D + \omega R_1) \bar{y}_{k+1} = F_k, \quad (D + \omega R_2) y_{k+1} = D \bar{y}_{k+1},$$

$$k = 0, 1, \dots \quad (4)$$

Замечая, что  $(R_1 y, y) = (R_2 y, y) = (R y, y)/2$ , находим

$$((D + \omega R_1) y, y) = (D y, y) + \omega (R_1 y, y) =$$

$$= \left( \left( D + \frac{\omega}{2} R \right) y, y \right) = ((D + \omega R_2) y, y) > 0,$$

так как  $D > 0$ ,  $\omega > 0$  и  $R > 0$ ,

Отсюда следует существование обратных операторов  $(D + \omega R_1)^{-1}$ ,  $(D + \omega R_2)^{-1}$ , т. е. разрешимость задачи (4).

**2. Выбор параметра  $\omega$ .** Чтобы пользоваться общей теорией, надо сначала найти параметры  $\gamma_1$  и  $\gamma_2$ , входящие в неравенства

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad (5)$$

которые в силу ограниченности и положительности операторов  $A$  и  $B$  всегда выполнены. Начнем с определения параметра  $\omega > 0$ .

**Лемма.** *Пусть оператор  $B$  определяется по формуле (2), где*

$$R_2^* = R_1, \quad R_1 + R_2 = R, \quad R = R^* > 0$$

и  $R$  удовлетворяет условиям

$$R \geq \delta D, \quad \delta > 0, \quad R_1 D^{-1} R_2 \leq \frac{\Delta}{4} R, \quad \Delta > 0. \quad (6)$$

Тогда справедлива оценка

$$\overset{\circ}{\gamma}_1 B \leq R \leq \overset{\circ}{\gamma}_2 B, \quad \overset{\circ}{\gamma}_1 = \frac{\delta}{1 + \omega\delta + 0,25\omega^2\delta\Delta}, \quad \overset{\circ}{\gamma}_2 = \frac{1}{2\omega}. \quad (7)$$

Отношение  $\xi = \overset{\circ}{\gamma}_1(\omega)/\overset{\circ}{\gamma}_2(\omega)$  имеет наибольшее значение при

$$\omega = \overset{\circ}{\omega} = 2/\sqrt{\delta\Delta}; \quad (8)$$

при этом

$$\overset{\circ}{\xi} = \frac{2\sqrt{\eta}}{1 + \sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}, \quad \overset{\circ}{\gamma}_1 = \frac{\delta}{2(1 + \sqrt{\eta})}, \quad \overset{\circ}{\gamma}_2 = \frac{\delta}{4\sqrt{\eta}}. \quad (9)$$

**Доказательство.** Неравенства (6) означают, что  $(Ry, y) \geq \delta(Dy, y)$ ,  $(D^{-1}R_2 y, R_2 y) \leq \frac{\Delta}{4}(Ry, y)$

для всех  $y \in H$ .

После преобразований

$$\begin{aligned} B &= (D + \omega R_1)D^{-1}(D + \omega R_2) = \\ &= D - \omega(R_1 + R_2) + \omega^2 R_1 D^{-1} R_2 + 2\omega(R_1 + R_2) = \\ &= (D - \omega R_2)D^{-1}(D - \omega R_2) + 2\omega R \end{aligned}$$

получаем

$$\begin{aligned} (By, y) &= (D^{-1}(D - \omega R_2)y, (D - \omega R_2)y) + 2\omega(Ry, y) = \\ &= \| (D - \omega R_2)y \|_{D^{-1}}^2 + 2\omega(Ry, y) \geq 2\omega(Ry, y), \end{aligned}$$

так что

$$B \geqslant 2\omega R, \quad \text{или} \quad R \leqslant \frac{1}{2\omega} B, \quad \overset{\circ}{\gamma}_2 = \frac{1}{2\omega}.$$

Получим теперь оценку для  $B$  сверху. Учитывая (6), найдем

$$\begin{aligned} B = D + \omega R + \omega^2 R_1 D^{-1} R_2 &\leqslant \frac{1}{\delta} R + \omega R + \frac{\omega^2 \Delta}{4} R \leqslant \\ &\leqslant \frac{1}{\delta} \left( 1 + \omega \delta + \frac{\omega^2 \delta \Delta}{4} \right) R, \\ R_1 &\geqslant \overset{\circ}{\gamma}_1 B, \quad \overset{\circ}{\gamma}_1 = \delta \left( 1 + \omega \delta + \frac{\omega^2 \delta \Delta}{4} \right)^{-1}. \end{aligned}$$

Число итераций, необходимое для решения уравнения  $Ry = f$ , зависит от отношения

$$\overset{\circ}{\xi}(\omega) = \overset{\circ}{\gamma}_1 / \overset{\circ}{\gamma}_2 = 2\omega \delta (1 + \omega \delta + \omega^2 \delta \Delta / 4)^{-1}.$$

Выберем  $\omega$  из условия максимума  $\overset{\circ}{\xi}(\omega)$ . Приравнивая нуль производную  $\overset{\circ}{\xi}'(\omega) = 2\delta(1 - \omega^2 \delta \Delta / 4)(1 + \omega \delta + \omega^2 \delta \Delta / 4)^{-2}$ , находим  $\omega = \overset{\circ}{\omega} = 2/\sqrt{\delta \Delta}$ ; при этом  $\overset{\circ}{\xi}''(\overset{\circ}{\omega}) < 0$ . Подставляя это значение  $\omega$  в формулы для  $\overset{\circ}{\gamma}_1$ ,  $\overset{\circ}{\gamma}_2$ ,  $\overset{\circ}{\xi}(\overset{\circ}{\omega})$ , получаем формулу (9). Лемма доказана.

### 3. Скорость сходимости.

**Теорема.** Пусть оператор  $A = A^* > 0$  представлен в виде суммы  $A = A_1 + A_2$ ,  $A_2 = A_1^*$ , и выполнены условия

$$A \geqslant \delta D, \quad A_1 D^{-1} A_2 \leqslant \frac{\Delta}{4} A, \quad \delta > 0, \quad \Delta > 0. \quad (10)$$

Тогда для попаременно-треугольного метода (1) с  $B = (D + \omega A_1) D^{-1} (D + \omega A_2)$ ,  $D = D^* > 0$ , (11)

с параметром  $\omega = 2/\sqrt{\delta \Delta}$  и чебышевским набором параметров

$$\begin{aligned} \tau_k^* &= \frac{\tau_0}{1 + \rho_0 \mu_k^*}, \quad \tau_0 = \frac{2}{\overset{\circ}{\gamma}_1 + \overset{\circ}{\gamma}_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \\ \xi &= \frac{\overset{\circ}{\gamma}_1}{\overset{\circ}{\gamma}_2} = \frac{2\sqrt{\eta}}{1 + \sqrt{\eta}}, \quad (12) \end{aligned}$$

где

$$\overset{\circ}{\gamma}_1 = \frac{\delta}{2(1 + \sqrt{\eta})}, \quad \overset{\circ}{\gamma}_2 = \frac{\delta}{4\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}, \quad \mu_k^* \in \bar{\mathfrak{M}}_n^*, \quad (13)$$

достаточно  $n(\varepsilon)$  итераций:

$$n_0(\varepsilon) \leq n(\varepsilon) < n_0(\varepsilon) + 1, \quad n_0(\varepsilon) < \ln \frac{2}{\varepsilon} / (2 \sqrt{2} \sqrt{\eta}); \quad (14)$$

при этом выполнена оценка

$$\|Ay_n - f\|_{B^{-1}} \leq \varepsilon \|Ay_0 - f\|_{B^{-1}}. \quad (15)$$

**Доказательство.** Воспользуемся предыдущей леммой, полагая  $R = A$ ,  $R_1 = A_1$ ,  $R_2 = A_2$ , а также оценкой (24) из § 4:

$$\|Ay_n - f\|_{B^{-1}} \leq q_n \|Ay_0 - f\|_{B^{-1}}$$

с

$$q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}.$$

Для числа итераций  $n = n(\varepsilon)$  в § 3 была получена оценка  $n_0(\varepsilon) \leq n(\varepsilon) < n_0(\varepsilon) + 1$ , где  $n_0(\varepsilon) < \ln \frac{2}{\varepsilon} / (2 \sqrt{2} \sqrt{\xi})$ .

Подставляя сюда  $\xi = 2\sqrt{\eta}/(1 + \sqrt{\eta})$ , получим (15).

**4. Пример применения попаременно-треугольного метода.** Рассмотрим модельную задачу

$$u_{xx,i} = \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = -f_i, \quad i = 1, 2, \dots, N-1,$$

$$u_0 = 0, \quad u_N = 0. \quad (16)$$

Пусть  $H = \Omega$  — пространство сеточных функций, заданных во внутренних узлах  $i = 1, 2, \dots, N-1$  сетки  $\omega_h$ ; введем скалярное произведение

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h.$$

Оператор  $Ay = -\overset{\circ}{u}_{xx}$  является самосопряженным и положительно определенным:

$$A \geq \delta E, \quad \delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}.$$

Введем операторы  $Dy = y$  ( $D = E$ ) и

$$A_1 y = R_1 y = \frac{y_{\bar{x}, i}}{h} = \frac{y_i - y_{i-1}}{h^2},$$

$$A_2 y = R_2 y = -\frac{y_{x, i}}{h} = -\frac{y_{i+1} - y_i}{h^2}, \quad A_1 + A_2 = A.$$

Итерации  $(y_i)_k = y_k(i)$  находятся по формулам

$$(E + \omega A_1) (\bar{y}_i)_{k+1} = \left( \bar{y}_i + \omega \frac{\bar{y}_i - \bar{y}_{i-1}}{h^2} \right)_{k+1} = F_k(i),$$

$$\bar{y}_{k+1}(i) = \frac{\omega \bar{y}_{k+1}(i-1) + h^2 F_k(i)}{h^2 + \omega},$$

$$(E + \omega A_2) y_{k+1}(i) =$$

$$= y_{k+1}(i) - \frac{\omega}{h^2} (y_{k+1}(i+1) - y_{k+1}(i)) = \bar{y}_{k+1}(i),$$

окончательно имеем

$$y_{k+1}(i) = \frac{\omega y_{k+1}(i-1) + h^2 \bar{y}_{k+1}(i)}{\omega + h^2},$$

$$i = N-1, N-2, \dots, 2, 1.$$

Значения  $\bar{y}_{k+1}(i)$  находятся последовательно при движении слева направо (от  $i-1$  к  $i$ ), а  $y_{k+1}(i)$  — справа налево (от  $i+1$  к  $i$ ); при этом учитываются краевые условия

$$\bar{y}_{k+1}(0) = 0, \quad y_{k+1}(N) = 0.$$

Формулы подобного типа называют *формулами бегущего счета*.

Из равенства  $y_{\bar{x}, i+1} = y_{x, i}$  следует, что  $A_1^* = A_2$ . В самом деле, так как  $v_1 = v_0 + h v_{\bar{x}, 1} = h v_{\bar{x}, 1}$ , то

$$(A_2 y, v) = - \sum_{i=1}^{N-1} y_{x, i} v_i = - y_1 v_1 \frac{1}{h} - \sum_{i=1}^{N-1} y_{i+1} v_{x, i} =$$

$$= y_1 v_{\bar{x}, 1} + \sum_{i=2}^N y_i v_{\bar{x}, i} = \sum_{i=1}^{N-1} y_i v_{\bar{x}, i} = h \sum_{i=1}^{N-1} y_i \frac{v_{\bar{x}, i}}{h} = (y, A_1 v),$$

т. е.  $A_1 = A_2^*$ .

Вычислим постоянную  $\Delta$ :

$$\begin{aligned}
 (A_1 A_2 y, y) &= (A_2 y, A_2 y) = \\
 &= \frac{1}{h^2} \sum_{i=1}^{N-1} (y_{x,i})^2 h = \frac{1}{h^2} \sum_{i=2}^{N-1} (y_{\tilde{x},i})^2 h \leqslant \\
 &\leqslant \frac{1}{h^2} \sum_{i=1}^N h (y_{\tilde{x},i})^2 = \frac{1}{h^2} \sum_{i=1}^{N-1} h (Ay)_i y_i = \frac{1}{h^2} (Ay, y),
 \end{aligned}$$

откуда следует  $\Delta = 4/h^2$ . Таким образом,

$$\begin{aligned}
 \eta &= \frac{\delta}{\Delta} = \sin^2 \frac{\pi h}{2} \approx \frac{\pi^2 h^2}{4}, \quad \sqrt{\eta} \approx \frac{\pi h}{2}, \\
 \xi &= 2 \sqrt{\eta} / (1 + \sqrt{\eta}) \approx 2 \sqrt{\eta} \approx \pi h, \quad \sqrt{\xi} \approx \sqrt{\pi h},
 \end{aligned}$$

так что

$$n_0(\varepsilon) \approx \frac{1}{2 \sqrt{\pi h}} \ln \frac{2}{\varepsilon}.$$

Если  $\varepsilon = 10^{-4}$ , то это дает  $n_0(\varepsilon) \approx 3/\sqrt{h}$ .

Результат таков:

$$\begin{aligned}
 n_0(\varepsilon) &\approx 10 \quad \text{при } h = 1/10 \quad (N = 10), \\
 n_0(\varepsilon) &\approx 30 \quad \text{при } h = 1/100 \quad (N = 100).
 \end{aligned}$$

Напомним, что при  $N = 100$  надо сделать 20 000 итераций по методу простой итерации и 340 итераций по явной чебышевской схеме. Таким образом, попаременно-треугольный метод оказался лучшим среди тех методов, которые мы изучали.

## § 6. Вариационно-итерационные методы

**1. Метод минимальных невязок.** До сих пор при изучении итерационных методов мы всюду предполагали, что постоянные  $\gamma_1$  и  $\gamma_2$  — границы спектра оператора  $A$  в  $H$  или в  $H_\nu$  — известны. Что делать, если такой информации нет? В этом случае можно применять методы, не использующие в явном виде параметры  $\gamma_1$  и  $\gamma_2$ . Это методы вариационного типа. Здесь мы рассмотрим методы минимальных невязок, скорейшего спуска и сопряженных градиентов.

Начнем с метода минимальных невязок для явной схемы

$$\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \text{ для всех } y_0 \in H. \quad (1)$$

Для невязки  $r_k = Ay_k - f$  имеем однородное уравнение

$$\frac{r_{k+1} - r_k}{\tau_{k+1}} + Ar_k = 0, \quad k = 0, 1, \dots, \quad r_0 = Ay_0 - f. \quad (2)$$

Параметр  $\tau_{k+1}$  будем выбирать из условия минимума невязки  $r_{k+1}$  по норме:

$$\begin{aligned} \|r_{k+1}\|^2 &= \|r_k - \tau_{k+1}Ar_k\|^2 = \\ &= \|r_k\|^2 - 2\tau_{k+1}(r_k, Ar_k) + \tau_{k+1}^2 \|Ar_k\|^2 = \varphi(\tau_{k+1}). \end{aligned}$$

Продифференцируем это выражение по  $\tau_{k+1}$ , приравняем производную  $\varphi'(\tau_{k+1})$  нулю:

$$\varphi'(\tau_{k+1}) = -2(r_k, Ar_k) + 2\tau_{k+1}\|Ar_k\|^2 = 0$$

и найдем

$$\tau_{k+1} = \frac{(Ar_k, r_k)}{\|Ar_k\|^2}, \quad k = 1, 2, \dots \quad (3)$$

При этом значении  $\tau_{k+1}$  вторая производная  $\varphi''(\tau_{k+1})$  положительна и, следовательно, достигается  $\min_{\tau_{k+1}} \|r_{k+1}\|^2$ .

До сих пор мы не предполагали, что  $A$  — самосопряженный оператор. Если же  $A = A^* > 0$ , то верны оценки

$$\begin{aligned} \|r_{k+1}\| &\leq \rho_0 \|r_k\|, \quad \|Ay_n - f\| \leq \rho_0^n \|Ay_0 - f\|, \\ \rho_0 &= \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \end{aligned} \quad (4)$$

где  $\gamma_1$  и  $\gamma_2$  — точные границы спектра оператора  $A$ . В самом деле, так как при значении  $\tau_{k+1}$  согласно (3) норма  $\|r_{k+1}\|$  минимальна, то при любом  $\tau \neq \tau_{k+1}$  она должна возрастать, поэтому

$$\|r_{k+1}\|^2 = \|r_k - \tau_{k+1}Ar_k\|^2 \leq \|r_k - \tau_0 Ar_k\|^2 \leq \|E - \tau_0 A\|^2 \|r_k\|^2.$$

С другой стороны, известно, что

$$\|E - \tau_0 A\| = \rho_0 \text{ при } \tau_0 = 2/(\gamma_1 + \gamma_2).$$

Отсюда и следует, что  $\|r_{k+1}\| \leq \rho_0 \|r_k\|$ .

Таким образом, метод минимальных невязок сходится с той же скоростью, что и метод простой итерации (если в методе простой итерации используются точные значения  $\gamma_1$  и  $\gamma_2$ ).

В случае неявного метода невязок, или *метода поправок*, вместо (4) получаем уравнение для поправки:

$$B \frac{w_{k+1} - w_k}{\tau_{k+1}} + Aw_k = 0, \quad k = 0, 1, \dots,$$

$$w_k = B^{-1}r_k, \quad w_0 = B^{-1}(Ay_0 - f), \quad (5)$$

где  $\tau_{k+1}$  определяется по формуле

$$\tau_{k+1} = \frac{(Aw_k, w_k)}{(B^{-1}Aw_k, Aw_k)}, \quad k = 0, 1, \dots \quad (6)$$

Вместо (4) получаем оценку

$$\|Ay_n - f\|_{B^{-1}} \leq \rho_0^n \|Ay_0 - f\|_{B^{-1}}.$$

**2. Метод скорейшего спуска.** Явный метод скорейшего спуска отличается от метода минимальных невязок только формулой для  $\tau_{k+1}$ :

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Ar_k, r_k)}, \quad k = 0, 1, \dots \quad (7)$$

Эта формула получается либо из условия минимума нормы  $\|z_{k+1}\|_A$  погрешности  $z_{k+1} = y_{k+1} - u$ , либо из условия ортогональности невязок  $r_k$  и  $r_{k+1}$ . Умножая скалярно уравнение  $r_{k+1} = r_k - \tau_{k+1}Ar_k$  на  $r_k$ , получаем  $0 = (r_k, r_k) - \tau_{k+1}(Ar_k, r_k)$ , откуда следует формула (7). Поскольку  $Az_k = Ay_k - Au = r_k$ , то

$$(Az_{k+1}, z_{k+1}) = (Az_k - \tau_{k+1}A^2z_k, z_k - \tau_{k+1}Az_k) =$$

$$= (r_k - \tau_{k+1}Ar_k, z_k - \tau_{k+1}r_k) = (r_k, z_k) -$$

$$- 2\tau_{k+1}(r_k, r_k) + \tau_{k+1}^2(Ar_k, r_k).$$

Дифференцируя  $\|z_{k+1}\|_A^2$  по  $\tau_{k+1}$  и приравнивая производную нулю, получим (7).

Далее, имеем

$$\|z_{k+1}\|_A^2 = \|(E - \tau_{k+1}A)z_k\|_A^2 \leq \|(E - \tau_0A)z_k\|_A^2 \leq$$

$$\leq \|E - \tau_0A\|^2 \|z_k\|_A^2 \leq \rho_0^2 \|z_k\|_A^2,$$

т. е.

$$\|z_{k+1}\|_A = \|y_{k+1} - u\|_A \leq \rho_0^n \|y_0 - u\|_A.$$

Метод скорейшего спуска сходится в  $H_A$  с той же скоростью, что и метод простой итерации.

**3. Метод сопряженных градиентов.** Более быстро сходящиеся методы вариационного типа можно найти в классе неявных трехслойных итерационных схем:

$$\begin{aligned} By_{k+1} &= \alpha_{k+1}(B - \tau_{k+1}A)y_k + (1 - \alpha_{k+1})By_{k-1} + \\ &\quad + \alpha_{k+1}\tau_{k+1}f, \quad k = 1, 2, \dots, \end{aligned} \quad (8)$$

$$By_k = (B - \tau_k A)y_{k-1} + \tau_k f.$$

Мы рассмотрим *метод сопряженных градиентов*, широко используемый на практике. Для него итерационные параметры  $\alpha_{k+1}$  и  $\tau_{k+1}$  определяются по формулам

$$\tau_{k+1} = \frac{(r_k, w_k)}{(Aw_k, w_k)}, \quad \alpha_{k+1} = \left( 1 - \frac{\tau_{k+1}}{\tau_k} \frac{(r_k, w_k)}{(r_{k-1}, w_{k-1})} \frac{1}{\alpha_k} \right)^{-1}, \quad (9)$$

где  $k = 0, 1, 2, \dots$ , в предположении, что  $A = A^* > 0$ ,  $B = B^* > 0$ ,  $\gamma_1 B \leq A \leq \gamma_2 B$ ,  $\gamma_1 > 0$ . Формулы для  $\tau_{k+1}$ ,  $\alpha_{k+1}$  получаются из требования минимума нормы разрешающего оператора. При этих оптимальных значениях итерационных параметров верна оценка

$$\begin{aligned} \|y_n - u\|_A &\leq q_n \|y_0 - u\|_A, \quad q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}, \\ \rho_1 &= \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \end{aligned} \quad (10)$$

т. е. скорость сходимости метода сопряженных градиентов — такая же, как и скорость сходимости двухслойного итерационного метода с чебышевскими параметрами (который использует  $\gamma_1$  и  $\gamma_2$  при вычислении параметров  $\tau_{k+1}$ ). Поэтому для числа итераций имеем оценки

$$n_0(\varepsilon) \leq n(\varepsilon) \leq n_0(\varepsilon) + 1, \quad n_0(\varepsilon) = \frac{1}{2\sqrt{\xi}} \ln \frac{2}{\varepsilon}.$$

В качестве оператора  $B$  можно взять факторизованный оператор попеременно-треугольного метода

$$B = (D + \omega A_1) D^{-1} (D + \omega A_2),$$

$$A_1 + A_2 = A > 0, \quad A_1^* = A_2, \quad D = D^* > 0,$$

Расчеты показывают, что число итераций по попарно-треугольному методу в сочетании с методом сопряженных градиентов меньше, чем в случае использования чебышевской схемы.

## § 7. Решение нелинейных уравнений

**1. Итерационные методы.** Рассмотрим нелинейное уравнение

$$f(x) = 0, \quad x \in [a, b], \quad (1)$$

где  $f(x)$  — непрерывная функция. Уравнение может иметь один или несколько корней. Требуется: 1) установить существование корней уравнения; 2) найти приближенные значения корней. Часто обе задачи решаются одновременно. Для нахождения корней применяются итерационные методы.

Простейшим является *метод дихотомии* (деления пополам). Пусть  $f(x_0)f(x_1) \leq 0$ ; тогда на отрезке  $[x_0, x_1]$  лежит не менее одного корня. Найдем  $f(x_2)$ , где  $x_2 = (x_0 + x_1)/2$ , и возьмем  $x_3$  — то из значений  $x_0$  или  $x_1$ , для которого выполняется условие  $f(x_2)f(x_3) \leq 0$ . Отрезок  $[x_2, x_3]$  снова делим пополам, и т. д. Деление продолжим до тех пор, пока длина отрезка станет меньше  $2\epsilon$ , где  $\epsilon$  — точность, с которой надо определить корень. Тогда середина этого отрезка и дает значение корня с требуемой точностью  $\epsilon$ . Процесс, очевидно, сходится со скоростью геометрической прогрессии со знаменателем  $1/2$ . Недостатки метода: выбор начального отрезка  $[x_0, x_1]$  — заранее неясно, к какому корню сойдется процесс (если их несколько на  $[x_0, x_1]$ ).

Второй метод — *метод простой итерации*. Перепишем уравнение (1) в виде

$$x = \varphi(x), \quad (2)$$

где  $\varphi(x)$  можно определить одним из способов:

$$\varphi(x) = x - \alpha f(x), \quad \alpha = \text{const},$$

$\varphi(x) = x + \rho(x)f(x)$ ,  $\rho(x)$  — произвольная функция, не имеющая корней на отрезке  $[a, b]$ .

Метод простой итерации определяется формулой

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, 2, \dots, \quad (3)$$

где  $n$  — номер итерации,  $x_0$  — заданное произвольно начальное приближение. Требуется найти приближенно

решение (корень)  $x = x^*$  уравнения  $x = \varphi(x)$  с относительной погрешностью  $\varepsilon > 0$ , так, чтобы при всех  $n \geq n_0$  выполнялось неравенство

$$|x_n - x^*| \leq \varepsilon |x_0 - x^*|, \quad n \geq n_0(\varepsilon). \quad (4)$$

Это условие может быть выполнено, если последовательность итераций  $\{x_n\}$  сходится при  $n \rightarrow \infty$  к пределу  $x^*$ :  $\lim_{n \rightarrow \infty} x_n = x^*$ . Если (4) имеет место, то вычисления

можно прекратить при  $n = n_0$ . Отсюда видно, что основным является вопрос о сходимости итераций, а также о скорости их сходимости, т. е. о минимальном числе итераций  $n_0(\varepsilon)$ , при котором выполнено (4). Предположим, что в некоторой  $\delta$ -окрестности

$$\Delta = (x_0 - \delta, x_0 + \delta), \quad \delta > 0, \quad (5)$$

точки  $x_0$  функция  $\varphi(x)$  удовлетворяет условию Липшица:  $|\varphi(x'') - \varphi(x')| \leq q|x'' - x'|$  для любых  $x', x'' \in \Delta$  (6) с коэффициентом  $q < 1$ :

$$0 < q < 1 \quad (7)$$

и пусть начальная невязка  $x_0 - \varphi(x_0)$  мала, так что

$$|x_0 - \varphi(x_0)| \leq (1 - q)\delta. \quad (8)$$

Тогда справедливы следующие утверждения:

- все итерации  $x_n$  ( $n = 1, 2, \dots$ ) принадлежат интервалу  $\Delta$ :  $x_n \in \Delta$ ;
- последовательность  $\{x_n\}$  при  $n \rightarrow \infty$  сходится к пределу  $x^*$ , являющемуся корнем уравнения (8);
- уравнение (2) имеет только один корень в  $\Delta$ .

Условие  $x_k \in \Delta$  означает, что

$$|x_k - x_0| < \delta. \quad (9)$$

В силу (8) имеем  $|x_1 - x_0| = |\varphi(x_0) - x_0| \leq (1 - q)\delta < \delta$ , т. е. (9) выполнено при  $k = 1$ . Докажем методом индукции, что (9) справедливо для всех  $k = 1, 2, \dots$ . Предположим, что (9) выполнено при  $k = 1, 2, \dots, n$ , тогда можно вычислить  $\varphi(x_n)$  и  $x_{n+1} = \varphi(x_n)$ . Из (6) следует, что  $|x_{n+1} - x_n| = |\varphi(x_n) - \varphi(x_{n-1})| \leq q|x_n - x_{n-1}|$ , т. е.

$$|x_{n+1} - x_n| \leq q|x_n - x_{n-1}|. \quad (10)$$

Последовательно применяя это неравенство, находим

$$|x_{k+1} - x_k| \leq q^k|x_1 - x_0|, \quad k = 1, 2, \dots, n. \quad (11)$$

Учитывая, что  $x_{n+1} - x_0 = (x_{n+1} - x_n) + (x_n - x_{n-1}) + \dots + (x_2 - x_1) + (x_1 - x_0)$ , получим

$$\begin{aligned}|x_{n+1} - x_0| &\leq (q^n + q^{n-1} + \dots + q + 1)|x_1 - x_0| = \\&= \frac{1 - q^{n+1}}{1 - q} |x_1 - x_0| < \frac{1}{1 - q} |x_1 - x_0| < \delta,\end{aligned}$$

т. е.  $x_{n+1} \in \Delta$ . Так как неравенство (9) в силу (8) верно для  $k = 1$ , то оно выполняется и для  $k = 2, 3, \dots$

Рассмотрим теперь разность  $x_{n+m} - x_n = (x_{n+m} - x_{n+m-1}) + (x_{n+m-1} - x_{n+m-2}) + \dots + (x_{n+2} - x_{n+1}) + (x_{n+1} - x_n)$  и оценим ее:

$$\begin{aligned}|x_{n+m} - x_n| &\leq (q^{m-1} + q^{m-2} + \dots + q + 1)|x_{n+1} - x_n| \leq \\&\leq \frac{1 - q^m}{1 - q} q^n |x_1 - x_0| < q^n \delta,\end{aligned}$$

т. е.  $|x_{n+m} - x_n| \rightarrow 0$  при  $n \rightarrow \infty$  и любом  $m = 1, 2, \dots$  Отсюда, в силу критерия Коши, следует сходимость  $\{x_n\}$ :  $\lim_{n \rightarrow \infty} x_n = x^* \in \Delta$ . Переходя затем в (3) к пределу при  $n \rightarrow \infty$ , убеждаемся, что  $x^*$  есть корень уравнения (2):  $x^* = \varphi(x^*)$ . Этот корень единственный. В самом деле, пусть существует два разных корня  $x'$  и  $x'' \neq x'$ , так что  $x' = \varphi(x')$ ,  $x'' = \varphi(x'')$ . Тогда  $|x'' - x'| = |\varphi(x'') - \varphi(x')| \leq q|x'' - x'| < |x'' - x'|$ , т. е.  $|x'' - x'| < |x'' - x'|$ , что невозможно.

Для погрешности  $z_{n+1} = x_{n+1} - x^*$  имеем

$$\begin{aligned}|z_{n+1}| &= |\varphi(x_n) - \varphi(x^*)| \leq q|x_n - x^*| = \\&= q|z_n| \leq q^{n+1}|z_0|, \quad (12) \\|z_{n+1}| &\leq q^{n+1}|z_0|,\end{aligned}$$

т. е. метод простой итерации сходится со скоростью геометрической прогрессии. Число итераций, при котором выполнено неравенство (4), определяется из условия  $q^n \leq \varepsilon$ , т. е.

$$n \geq \ln \frac{1}{\varepsilon} / \ln \frac{1}{q}.$$

Минимальное число  $n_0(\varepsilon)$  итераций, при которых (4) выполнено, очевидно, равно

$$n_0(\varepsilon) = \left[ \ln \frac{1}{\varepsilon} / \ln \frac{1}{q} \right], \quad (13)$$

где  $[a]$  — целая часть числа  $a > 0$ .

**Замечание.** Если  $\varphi(x)$  имеет производную на  $\Delta$ , то (6) выполнено в случае, когда

$$|\varphi'(x)| \leq q \text{ для всех } x \in \Delta. \quad (14)$$

**2. Метод Ньютона.** Метод определяется формулой

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad f'(x_n) \neq 0, \quad n = 0, 1, 2, \dots \quad (15)$$

Эта формула получается, если в разложении

$$0 = f(x^*) = f(x_n) + (x^* - x_n)f'(x_n) + \frac{1}{2}(x^* - x_n)^2 f''(\xi), \\ \xi = x_n + \theta(x^* - x_n), \quad 0 \leq \theta \leq 1, \quad (16)$$

где  $x^*$  — точное решение уравнения  $f(x) = 0$ , отбросить последний член, заменив  $x^*$  на  $x_{n+1}$ :

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n).$$

Метод Ньютона называют также *методом касательных* или *методом линеаризации*. Его геометрическая интерпретация — участок кривой  $y = f(x)$  при  $x \in [x_n, x_{n+1}]$ , если  $x_n < x_{n+1}$  (или при  $x \in [x_{n+1}, x_n]$ , если  $x_n > x_{n+1}$ ), заменяется отрезком касательной, проведенной из точки  $x = x_n$ .

Записывая  $f(x) = 0$  в виде  $x = \varphi(x)$ , видим, что метод Ньютона можно трактовать как метод простой итерации (3) с правой частью

$$\varphi(x) = x - f(x)/f'(x). \quad (17)$$

Проиллюстрируем метод Ньютона на примере извлечения квадратного корня из числа  $a > 0$ , т. е. решения уравнения  $x^2 = a$  или  $f(x) = x^2 - a = 0$ . Применяя формулу (15), получим

$$x_{n+1} = \frac{1}{2} \left( x_n + \frac{a}{x_n} \right), \quad n = 0, 1, \dots$$

Пусть  $a = 2$ . Выбирая  $x_0 = 1$ , найдем  $x_1 = 1,5$ ,  $x_2 = 1,417$ ,  $x_3 = 1,414, \dots$ , т. е. итерации сходятся очень быстро.

Оценим скорость сходимости итераций. Предположим, что существует вещественный корень  $x^*$  уравнения (1). Рассмотрим некоторую окрестность корня:

$$\Delta_0 = (x^* - \delta_0, x^* + \delta_0), \quad \delta_0 > 0.$$

Будем считать, что функция (17) дважды дифференцируе-

ма в  $\Delta_0$  и ее вторая производная ограничена:

$$|\varphi''(x)| \leq 2q, \quad (18)$$

где  $q > 0$  — постоянная. Разложим  $\varphi(x)$  в строку Тейлора в окрестности  $x = x^*$ :

$$\begin{aligned} \varphi(x) &= \varphi(x^*) + \varphi'(x^*)(x - x^*) + \frac{\varphi''(\xi)}{2}(x - x^*)^2, \\ \xi &= x^* + \theta(x - x^*), \quad 0 \leq \theta \leq 1. \end{aligned} \quad (19)$$

Вычисляя затем

$$\varphi'(x) = f f''/(f')^2 = -f(1/f')', \quad \varphi''(x) = -\left(f\left(\frac{1}{f'}\right)'\right)',$$

и замечая, что  $\varphi'(x^*) = 0$  при  $f'(x^*) \neq 0$ , получим

$$\varphi(x_n) = \varphi(x^*) + \frac{(x_n - x^*)^3}{2} \varphi''(\xi). \quad (20)$$

Для погрешности  $z_{n+1} = x_{n+1} - x^*$  получим формулу:

$$\begin{aligned} z_{n+1} &= x_{n+1} - x^* = \varphi(x_n) - \varphi(x^*) = \frac{1}{2}(x_n - x^*)^2 \varphi''(\xi), \\ z_{n+1} &= \frac{1}{2} \varphi''(\xi) z_n^2. \end{aligned}$$

Отсюда и из (20) следует

$$|z_{n+1}| \leq q z_n^2. \quad (21)$$

Обозначая  $v_n = q |z_n|$ , получаем  $v_{n+1} \leq v_n^2 \leq v_{n-1}^{2^n} \leq \dots \leq v_1^{2^n} \leq v_0^{2^{n+1}}$ , и, следовательно,

$$|z_{n+1}| \leq \frac{1}{q} (q |z_0|)^{2^{n+1}}. \quad (22)$$

Отсюда видно, что итерации (15) сходятся к корню  $x^*$  при  $n \rightarrow \infty$ , если

$$q |z_0| < 1 \text{ или } |z_0| = |x_0 - x^*| < 1/q, \quad (23)$$

т. е. начальное приближение находится в окрестности  $\Delta_0 = (x^* - 1/q, x^* + 1/q)$  с  $\delta_0 = 1/q$  корня  $x = x^*$  уравнения (1). В этом случае метод Ньютона, как принято говорить, сходится с *квадратичной скоростью* (метод простой итерации сходится со скоростью геометрической прогрессии).

Условие окончания итераций  $|z_n| \leq \epsilon |z_0|$ , как следует из (22), или  $|z_n| \leq (q |z_0|)^{2^{n-1}} |z_0|$ , выполнено, если  $n \geq n_0(\epsilon)$ ,

где

$$n_0(\varepsilon) = \left[ \ln \left( 1 + \ln \frac{1}{\varepsilon} / \ln \frac{1}{q|z_0|} \right) \right] \ln 2. \quad (24)$$

Очевидно, что если начальное приближение находится в малой окрестности  $x^*$ , то и все последующие итерации останутся в этой окрестности  $\Delta_0$ . В самом деле, пусть  $|x_0 - x^*| \leq \delta_0$ , причем  $q\delta_0 < 1$ . Тогда будем иметь  $|x_1 - x^*| \leq q|x_0 - x^*|^2 \leq q\delta_0^2 < \delta_0$ ,  $|x_2 - x^*| \leq q|x_1 - x^*|^2 \leq q\delta_0 < \delta_0$  и т. д., так что  $|x_n - x^*| \leq \delta_0$  для любого  $n = 1, 2, \dots$

**Замечания.** 1. Мы не останавливаемся на доказательстве существования корня  $x = x^*$ .

2. Квадратичную сходимость метода Ньютона можно установить и при более слабых ограничениях на  $f(x)$ :

$$|f'(x)| \geq M_1 > 0, \quad |f''(x)| \leq M_2 \text{ для всех } x \in \Delta_0. \quad (25)$$

Используя (15) и (16), получим для погрешности  $z_{n+1} = x_{n+1} - x^*$  выражение

$$z_{n+1} = \frac{f''(\xi)}{2f'(x_n)} z_n^2,$$

из которого в силу условий (25) следует неравенство

$$|z_{n+1}| \leq q|z_n|^2, \quad q = M_2/(2M_1),$$

которое совпадает с (21) (различие только в  $q$ ). Дальнейшие рассуждения приводят к (22), (23) и (24).

**3. Непрерывный метод Ньютона.** Решение уравнения  $f(x) = 0$  можно рассматривать как предел при  $t \rightarrow \infty$  решения задачи Коши:

$$\frac{dx}{dt} + f(x) = 0, \quad x > 0, \quad x(0) = u_0, \quad (26)$$

если этот предел существует. Обозначим через  $x = x(t)$  решение задачи Коши, через  $x_*$  — решение уравнения  $f(x) = 0$ . Для их разности  $z(t) = x(t) - x_*$  имеем

$$\frac{dz}{dt} + (f(x) - f(x_*)) = \frac{dz}{dt} + f'(\xi) \cdot z, \quad \xi = x_* + \theta z, \\ 0 \leq \theta \leq 1,$$

$$\frac{dz}{dt} + \alpha(t)z = 0, \quad t > 0, \quad z(0) = u_0, \quad \alpha(t) = f'(\xi).$$

Отсюда видно, что  $|z(t)| \rightarrow 0$  при  $t \rightarrow \infty$ , если  $f'(x) > 0$ ,

Для решения уравнения (26) надо воспользоваться каким-либо явным методом. Быстрота сходимости  $x(t)$  к  $x_0$  зависит только от величины производной  $f(x)$ .

**4. Метод секущих.** Вычисление производной  $f'(x_n)$  в методе Ньютона может оказаться трудоемким. Если заменить  $f'_n$  разностным отношением  $(f_n - f_{n-1})/(x_n - x_{n-1})$ , то мы получим итерационный метод секущих

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1}) f(x_n)}{f(x_n) - f(x_{n-1})}. \quad (27)$$

Метод секущих сходится медленнее метода Ньютона, однако в (27) вычисляется только функция, а в (15) надо находить и функцию и ее производную. Поэтому объем вычислений на каждой итерации в методе секущих, вообще говоря, меньше.

## Глава IV

# РАЗНОСТНЫЕ МЕТОДЫ РЕШЕНИЯ КРАЕВЫХ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

## § 1. Основные понятия теории разностных схем

Универсальным численным методом решения дифференциальных уравнений является метод конечных разностей. Прежде чем переходить к его изложению, необходимо ввести основные понятия теории разностных схем — аппроксимацию, устойчивость и сходимость.

**1. Простейшие разностные операторы.** Для получения вместо дифференциального уравнения разностного уравнения необходимо:

- заменить область непрерывного изменения аргументов дискретным множеством точек (сеткой);
- заменить (аппроксимировать на сетке) дифференциальное уравнение разностным уравнением.

Вопрос о численном решении дифференциального уравнения сводится к вопросу о решении разностных уравнений. В предыдущих главах мы уже рассматривали примеры сеток:

1) равномерная сетка на отрезке  $0 \leq x \leq 1$  с шагом  $h$ : множество узлов  $\omega_h = \{x_i = ih, i = 0, 1, 2, \dots, N, h = 1/N\}$ ;  $x_0 = 0, x_N = 1$  — граничные узлы;  $\omega_h = \{x_i = ih, i = 1, 2, \dots, N - 1\}$  — множество внутренних узлов;

2) неравномерная сетка: отрезок  $0 \leq x \leq 1$  разбивается на  $N$  частей произвольными точками  $x_1 < x_2 < \dots < x_{N-1}$ ;  $h_i = x_i - x_{i-1}$  — шаг сетки;

$$\bar{\omega}_h = \{x_i, i = 0, 1, \dots, N, x_0 = 0, x_N = 1\}, \quad \sum_{i=1}^N h_i = 1,$$
$$\omega_h = \{x_i, 0 < i < N\};$$

3) сетка на отрезке  $0 \leq t \leq T$ :  $\bar{\omega}_t = \{t_n = n\tau, n = 0, 1, \dots, n_0; n_0\tau = T\}$ .

Вместо функции непрерывного аргумента (например, на отрезке  $0 \leq x \leq 1$ ) рассматривается функция  $y(x_i) = y_i$  дискретного аргумента  $x_i$ , где  $x_i$  — узел сетки  $\omega_h$ , или аргумента  $i$  — номера узла. Эта функция называется се-

точной. Любую сеточную функцию  $y(x_i) = y_i$  можно представить в виде вектора

$$Y = (y_0, y_1, \dots, y_{N-1}, y_N).$$

Поэтому множество сеточных функций образует конечномерное пространство  $H$ , в данном случае размерности  $(N+1)$ . Обычно рассматривается семейство сеток  $\{\bar{\omega}_h\}$ , зависящих от шага как от параметра. Поэтому и сеточные функции  $y = y_h(x)$  зависят от шага как от параметра (или от  $N$ ), если сетка  $\omega_h$  равномерна. Если сетка неравномерна, то под  $h$  понимается вектор  $h = (h_1, h_2, \dots, h_N)$ . Естественно поэтому снабдить пространство сеточных функций индексом  $h$  и писать  $H_h$ . В пространстве  $H_h$  можно ввести норму  $\|\cdot\|_h$ . Укажем простейшие виды норм:

$$\|y\|_C = \max_{x \in \bar{\omega}_h} |y(x)| \quad \text{или} \quad \|y\|_C = \max_{0 \leq i \leq N} |y_i|;$$

$$\|y\| = \left( \sum_{i=1}^{N-1} y_i^2 h_i \right)^{1/2}.$$

Дифференциальный оператор заменяется разностным оператором, действующим в пространстве сеточных функций.

Пусть  $G$  — область евклидова пространства  $R^p$  ( $p = 1, 2, 3$ ) с границей  $\Gamma$ . Например,  $G$  — интервал  $0 < x < 1$ ,  $\Gamma$  — точки  $x = 0, x = 1$ ;  $G$  — прямоугольник  $0 < x_1 < l_1, 0 < x_2 < l_2$ ,  $x = (x_1, x_2) \in G$  ( $p = 2$ ),  $\Gamma$  состоит из отрезков прямых  $x_2 = 0, x_2 = l_2, x_1 = 0, x_1 = l_1$  и т. д. Пусть задан линейный дифференциальный оператор  $L$ , действующий на функцию  $v(x)$ ,  $x \in G$ . Введем на  $\bar{G} = G \cup \Gamma$  сетку  $\bar{\omega}_h$  и будем рассматривать сеточную функцию  $v_h(x)$ ,  $x \in \omega_h$ . Заменим  $Lv$  в точке  $x_i \in \omega_h$  линейной комбинацией значений  $v_h(x)$  сеточной функции на некотором множестве узлов сетки, которое назовем *шаблоном*

$$(L_h v)_i = \sum_{x_j \in \sigma_i} a_{ij}^h v_h(x_j), \quad x_i \in \omega_h(G), \quad (1)$$

где  $a_{ij}^h$  — коэффициенты,  $\sigma_i$  — шаблон,  $\sigma_i \subseteq \bar{\omega}_h$ .

Такая замена  $Lv$  на  $L_h v$  называется *аппроксимацией на сетке* дифференциального оператора  $L$  разностным оператором  $L_h$ , или *разностной аппроксимацией* оператора  $L$ . Изучение разностных аппроксимаций  $L_h$  оператора  $L$  обычно проводится сначала локально, т. е. в любой фиксированной точке сетки. Построение  $L_h$  надо начи-

нать с выбора шаблона  $\sigma(x)$ , т. е. множества узлов, соседних с узлом  $x \in \omega_h$ , в которых значения сеточной функции  $v_h(x)$  могут быть использованы при написании выражения для  $L_h$ .

Рассмотрим несколько примеров построения  $L_h$ .

Пример 1. Первая производная:  $Lv = \frac{dv}{dx} = v'(x)$ .

Возьмем три узла  $(x - h, x, x + h)$ . Можно воспользоваться любым из выражений

$$L_h^+ v = \frac{v(x+h) - v(x)}{h} = v_x \quad (\text{шаблон } (x, x+h));$$

$$L_h^- v = \frac{v(x) - v(x-h)}{h} = v_{\bar{x}} \quad (\text{шаблон } (x-h, x));$$

$$L_h^0 v = \frac{v(x+h) - v(x-h)}{2h} = v_{\circ_x} \quad (\text{шаблон } (x-h, x+h)).$$

Часто применяются названия:  $L_h^+ v = v_x$  — *правая*,  $L_h^- v = v_{\bar{x}}$  — *левая*,  $L_h^0 v = v_{\circ_x} = \frac{1}{2}(L_h^+ v + L_h^- v)$  — *центральная разностные производные*. На трехточечном шаблоне  $(x - h, x, x + h)$  можно определить разностный оператор

$$L_h^{(\sigma)} v = \sigma v_x + (1 - \sigma) v_{\bar{x}},$$

где  $\sigma$  — действительный параметр. Таким образом, существует бесконечное множество разностных аппроксимаций первой производной на трехточечном шаблоне.

*Погрешностью аппроксимации оператора  $L$  оператором  $L_h$  называют разность*

$$\psi = L_h v - Lv.$$

Говорят, что  $L_h$  имеет  $m$ -й порядок аппроксимации в точке  $x$ , если

$$\psi(x) = L_h v(x) - Lv(x) = O(h^m) \quad \text{или} \quad |\psi(x)| \leq Mh^m,$$

где  $M = \text{const} > 0$  не зависит от  $h$ ,  $m > 0$ .

Используя формулу Тейлора

$$\begin{aligned} v(x \pm h) &= v(x) \pm hv'(x) + \\ &\quad + \frac{h^2}{2} v''(x) \pm \frac{h^3}{6} v'''(x) + \frac{h^4}{24} v^{IV}(x) + O(h^5), \end{aligned}$$

нетрудно получить оценки

$$v_x - v' = O(h), \quad v_{\bar{x}} - v' = O(h), \quad v_{\frac{x}{x}} - v' = O(h^2),$$

$$\psi^{(\sigma)} = L_h^{(\sigma)} v - Lv = O\left(\left(\sigma - \frac{1}{2}\right)h + h^2\right).$$

Пример 2. Вторая производная:  $Lv = \frac{d^2v}{dx^2} = v''(x)$ .

Возьмем тот же трехточечный шаблон, что и в примере 1, и напишем разностный оператор

$$L_h v(x) = \frac{v(x+h) - 2v(x) + v(x-h)}{h^2}.$$

Замечая, что  $v(x+h) = v(x) + hv_x$ ,  $v(x-h) = v(x) - hv_{\bar{x}}$ , преобразуем  $L_h v(x)$ :

$$L_h v(x) = \frac{v_x(x) - v_{\bar{x}}(x)}{h} = \frac{v_{\bar{x}}(x+h) - v_{\bar{x}}(x)}{h} = v_{\bar{xx}}(x). \quad (2)$$

Пользуясь формулой Тейлора для  $v(x \pm h)$ , находим

$$\psi = L_h v - Lv = \frac{h^2}{12} v^{IV}(x) + O(h^4) = O(h^2),$$

т. е.  $L_h$  имеет второй порядок аппроксимации.

Обычно требуется оценка погрешности аппроксимации на сетке, т. е. в некоторой сеточной норме  $\|\cdot\|_h$ . Говорят, что  $L_h$  имеет  $m$ -й порядок аппроксимации на сетке, если

$$\|L_h v_h - (Lv)_h\|_h = O(h^m).$$

**2. Разностная схема.** Обычно дифференциальное уравнение  $Lu = f(x)$  решается с некоторыми дополнительными условиями — начальными (задачи Коши), краевыми (краевая задача), либо и с начальными, и с краевыми условиями. Эти дополнительные условия при переходе к разностным уравнениям надо также аппроксимировать.

Пусть задана некоторая область  $G$  с границей  $\Gamma$  и пусть ищется решение  $u = u(x)$ ,  $x \in G$ , линейного дифференциального уравнения

$$Lu = f(x), \quad x \in G, \quad (3)$$

с дополнительным условием на границе:

$$u(x) = \mu(x), \quad x \in \Gamma. \quad (4)$$

Введем в области  $\bar{G} = G + \Gamma$  сетку  $\bar{\omega}_h = \omega_h + \gamma_h$ ,  $\omega_h \in G$ ,  $\gamma_h \in \Gamma$ , и поставим в соответствие задаче (3), (4) разностную задачу с линейным оператором  $L_h$  вида (1):

$$L_h y_h = \phi_h(x), \quad x \in \omega_h; \quad y_h(x) = v_h(x), \quad x \in \gamma_h. \quad (5)$$

Функции  $y_h(x)$ ,  $\phi_h(x)$ ,  $v_h(x)$  зависят от шага  $h$  сетки. Меняя  $h$ , получаем последовательности  $\{y_h\}$ ,  $\{\phi_h\}$ ,  $\{v_h\}$ . Таким образом, мы рассматриваем не одну разностную задачу, а семейство задач, зависящее от параметра  $h$ . Это семейство задач называется *разностной схемой*.

**Пример 1.** Задача Коши:

$$Lu = \frac{du}{dt} + \lambda u = f(t), \quad t > 0, \quad u(0) = u_0.$$

*Разностная схема Эйлера* имеет вид

$$L_\tau y = \frac{y_{n+1} - y_n}{\tau} + \lambda y_n = f_n,$$

$$y_n = y(t_n), \quad t_n = n\tau \in \omega_\tau, \quad n = 0, 1, \dots, y_0 = u_0.$$

**Пример 2.** Первая краевая задача:

$$Lu = u'' = -f(x), \quad 0 < x < 1, \quad u(0) = \mu_1, \quad u(1) = \mu_2. \quad (6)$$

Воспользуемся трехточечным разностным оператором (2):  $L_h y_i = y_{xx,i} = (y_{i+1} - 2y_i + y_{i-1})/h^2$  и получим разностную краевую задачу на сетке  $\bar{\omega}_h = \{x_i = ih, 0 \leq i \leq N, x_N = 1\}$ :

$$L_h y_i = y_{xx,i} = -f_i, \quad i = 1, 2, \dots, N-1,$$

$$y_0 = \mu_1, \quad y_N = \mu_2. \quad (6')$$

**3. Устойчивость.** Нам удобно перейти к записи разностной схемы (5) в операторной форме. Для этого сначала запишем уравнения (5) в матричной форме

$$AY_h = \Phi_h, \quad (7)$$

где  $Y_h$  — искомый конечномерный вектор размерности  $N$ , равной числу узлов сетки, в которых неизвестны значения сеточной функции  $y_h$  (для первой краевой задачи (6') размерность  $Y_h$  равна  $N-1$  — числу внутренних узлов сетки). Значения  $y_h(x_i)$  в узлах  $x_i \in \omega_h$  являются компонентами вектора  $Y_h$ ,  $\phi_h(x_i)$  — компоненты вектора  $\Phi_h$ ,  $A$  — квадратная матрица размера  $N \times N$ .

Введем  $N$ -мерное пространство  $H_h$  сеточных функций, и пусть  $A_h$  — линейный оператор, соответствующий

матрице  $A: A_h: H_h \rightarrow H_h$ . Вместо (7) можно написать

$$A_h y_h = \varphi_h, \quad \varphi_h \in H_h. \quad (8)$$

Пусть  $\|\cdot\|_{(1_h)}$  и  $\|\cdot\|_{(2_h)}$  — некоторые нормы в пространстве  $H_h$ .

Будем говорить, что разностная схема (8) *устойчива*, если существует такая постоянная  $M > 0$ , не зависящая от  $h$  и от выбора  $\varphi_h$ , что для решения  $y_h$  уравнения (8) имеет место оценка

$$\|y_h\|_{(1_h)} \leq M \|\varphi_h\|_{(2_h)} \quad (9)$$

при всех достаточно малых  $h$ :  $|h| \leq h_0$ .

Разностная схема (8) называется *корректной* (*корректно поставленной*), если решение уравнения (8) существует и единствено при любых входных данных  $\varphi_h \in H_h$  и если разностная схема устойчива, т. е. выполнено неравенство (9).

Устойчивость схемы означает непрерывную зависимость решения  $y_h$  от входных данных, причем эта непрерывная зависимость равномерна по  $h$ . Если  $\tilde{y}_h$  — решение уравнения  $A_h \tilde{y}_h = \varphi_h$ , то  $A_h(\tilde{y}_h - y_h) = \varphi_h - \varphi_h$  в силу линейности  $A_h$ ; тогда из (9) следует

$$\|\tilde{y}_h - y_h\|_{(1_h)} \leq M \|\tilde{\varphi}_h - \varphi_h\|_{(2_h)}. \quad (10)$$

Малому изменению входных данных соответствует малое изменение решения.

Если схема (8) разрешима, то существует обратный оператор  $A_h^{-1}$  и

$$y_h = A_h^{-1} \varphi_h, \quad \|y_h\|_{(1_h)} \leq \|A_h^{-1}\| \|\varphi_h\|_{(2_h)}, \quad (11)$$

где  $\|A_h^{-1}\| = \|A_h^{-1}\|_{(2_h \rightarrow 1_h)}$  — норма оператора  $A_h^{-1}$ .

Устойчивость означает равномерную по  $h$  ограниченность обратного оператора

$$\|A_h^{-1}\| \leq M. \quad (12)$$

Схема *неустойчива*, если не существует такой постоянной  $M$ , не зависящей от  $h$ , которая превосходила бы  $\|A_h^{-1}\|$ , т. е.  $\|A_h^{-1}\|$  неограниченно возрастает при  $|h| \rightarrow 0$ .

Может оказаться, что вместо краевого условия первого рода  $u = \mu$  при  $x \in \Gamma$  задано условие

$$lu = \mu(x), \quad x \in \Gamma, \quad (13)$$

где  $l$  — некоторый линейный дифференциальный оператор, например,  $lu = u' - \sigma u$ ,  $\sigma > 0$  или  $lu = u'$  при  $x = 0$  или  $x = 1$ ). Тогда вместо задачи (3), (4) имеем задачу

$$Lu = f(x), \quad x \in G; \quad lu = \mu(x), \quad x \in \Gamma. \quad (14)$$

Соответствующая разностная схема будет иметь вид

$$L_h y_h = \varphi_h \text{ при } x \in \omega_h, \quad l_h y_h = \bar{\mu}_h \text{ при } x \in \gamma_h, \quad (15)$$

где  $l_h$  — линейный разностный оператор, аппроксимирующий оператор  $l$ . Может оказаться, кроме того, что  $\varphi_h$  и  $\bar{\mu}_h$  надо оценивать в разных нормах  $\|\varphi_h\|_{(2h)}$ ,  $\|\bar{\mu}_h\|_{(3h)}$ .

Схема (15) *устойчива*, если для ее решения  $y_h$  справедлива оценка

$$\|y_h\|_{(1h)} \leq M_1 \|\varphi_h\|_{(2h)} + M_2 \|\bar{\mu}_h\|_{(3h)}, \quad (16)$$

где  $M_1 > 0$ ,  $M_2 > 0$  — постоянные, не зависящие от  $h$  и от выбора входных данных  $\varphi_h$  и  $\bar{\mu}_h$ .

Следует отметить, что разностная схема (15) также может быть записана в операторном виде  $A_h y_h = \varphi_h$ , однако при этом  $\|\cdot\|_{(2h)}$  в (9) и (16) могут не совпадать, также как и сами правые части (это ясно уже для первой краевой задачи).

**4. Пример устойчивой схемы.** В качестве примера устойчивой схемы рассмотрим разностную краевую задачу

$$y_{xx,i} = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = -\varphi_i, \quad i = 1, 2, \dots, N-1,$$

$$y_0 = 0, \quad y_N = 0, \quad hN = 1. \quad (17)$$

Следуя § 4 гл. I, определим оператор  $A_h$ . Пусть  $H_h$  — пространство сеточных функций, заданных во внутренних узлах ( $i = 1, 2, \dots, N-1$ ) сетки. Возьмем  $\overset{\circ}{y} \in H_h$  (индекс  $h$  у  $y_h(x)$  пока опускаем) и функцию  $\overset{\circ}{y}$ , совпадающую с  $\overset{\circ}{y}$  во внутренних узлах и равную нулю на границе:  $y_0 = y_N = 0$ . Тогда оператор  $A_h$  определим при помощи тождества

$$(A_h \overset{\circ}{y})_i = -\overset{\circ}{y}_{xx,i}, \quad i = 1, 2, \dots, N-1,$$

и получим вместо (17) операторное уравнение

$$A_h \overset{\circ}{y}_h = \varphi_h. \quad (18)$$

В пространстве  $H_h$  вводим скалярное произведение

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h.$$

Оператор  $A_h$  в  $H_h$  самосопряжен и положительно определен и

$$\delta E \leq A_h \leq \Delta E, \quad \text{или} \quad \delta \|y\|^2 \leq (A_h y, y) \leq \Delta \|y\|^2$$

для всех  $y \in H_h$ , (19)

где  $\delta$  и  $\Delta$  — наименьшее и наибольшее собственные значения оператора  $A$ , равные

$$\delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}, \quad \Delta = \|A_h\| = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}. \quad (20)$$

Обратный оператор  $A_h^{-1}$  самосопряжен, если  $A_h = A_h^*$ . В § 4 гл. I показано, что неравенства (19) эквивалентны операторным неравенствам

$$\frac{1}{\Delta} E \leq A_h^{-1} \leq \frac{1}{\delta} E, \quad \|A_h^{-1}\| = \frac{1}{\delta}. \quad (21)$$

Отсюда следуют равномерная ограниченность нормы обратного оператора  $A_h^{-1}$ :  $\|A_h^{-1}\| \leq 1/\delta < 1/8$  и априорная оценка

$$\|y_h\| \leq \frac{1}{\delta} \|\varphi_h\| \leq \frac{1}{8} \|\varphi_h\|, \quad (22)$$

выражающая устойчивость схемы (18). Эту оценку можно получить методом энергетических неравенств, не прибегая к оценке собственных значений  $\lambda_k(A_h^{-1})$ . В самом деле, умножим уравнение  $A_h y_h = \varphi_h$  скалярно на  $y_h$ :  $(A_h y_h, y_h) = (\varphi_h, y_h)$  и воспользуемся неравенствами  $(\varphi_h, y_h) \leq \|\varphi_h\| \|y_h\|$ ,  $\|y_h\|^2 \leq \frac{1}{\delta} (A_h y_h, y_h)$ ; тогда получим неравенство  $\delta \|y_h\|^2 \leq \|\varphi_h\| \|y_h\|$ , откуда и следует оценка (22).

Схема (17) устойчива также в норме  $\|y\|_c$ :

$$\|y_h\|_c \leq \frac{1}{2} \|\varphi_h\|_c, \quad \|y\|_c = \|y\|_{C_h} = \max_{0 < i < N} |y_i|. \quad (23)$$

Это следует из оценки решения трехточечной разностной краевой задачи, полученной в п. 3 § 5 гл. I. В данном случае оценка имеет вид

$$\|y_h\|_c \leq \sum_{s=1}^{N-1} h \sum_{k=1}^s h |\varphi_k| \leq \|\varphi\|_c \sum_{s=1}^N x_s h < \frac{1}{2} \|\varphi_h\|_c,$$

так как

$$\sum_{s=1}^N x_s h = h^2 \sum_{s=1}^N s = \frac{N(N+1)}{2} h^2 = \frac{1-h}{2} < \frac{1}{2}.$$

5. Пример некорректной схемы. Пусть дана схема

$$A_h y_h = \varphi_h$$

и  $\|A_h\| \rightarrow \infty$  при  $|h| \rightarrow 0$ . Рассмотрим обратную задачу — определить правую часть  $\varphi_h$  по известному решению  $y_h$ :

$$B_h \varphi_h = y_h, \quad B_h = A_h^{-1}.$$

Она является некорректной, так как

$$\|B_h^{-1}\| = \|(A_h^{-1})^{-1}\| = \|A_h\| \rightarrow \infty \text{ при } |h| \rightarrow 0.$$

Это значит, что для любой постоянной  $M$ , не зависящей от  $h$ , можно указать такое  $h_*$ , что  $\|B_h^{-1}\| > M$  при  $|h| \leq h_*$ . Пусть  $\tilde{\varphi}_h$  — решение уравнения  $B_h \tilde{\varphi}_h = \tilde{y}_h$ , а  $\varphi_h$  — решение уравнения  $B_h \varphi_h = y_h$ , тогда

$$\|\tilde{\varphi}_h - \varphi_h\| \leq \|B_h^{-1}\| \|\tilde{y}_h - y_h\|.$$

Если же

$$\|B_h^{-1}\| \leq M \text{ при } |h| \geq h_0,$$

так что справедливо неравенство

$$\|\tilde{\varphi}_h - \varphi_h\| \leq M \|\tilde{y}_h - y_h\|,$$

то будем говорить, что схема *квазистабильна*. Можно ли пользоваться этой схемой для определения  $\varphi_h$  с требуемой точностью  $\epsilon$ , если  $y_h$  задано с некоторой точностью  $\epsilon_0$ :

$$\|\tilde{y}_h - y_h\| \leq \epsilon?$$

Из неравенства  $\|\tilde{\varphi}_h - \varphi_h\| \leq \|B_h^{-1}\| \|\tilde{y}_h - y_h\|$  следует, что решение задачи  $B_h \varphi_h = y_h$  определяется с точностью  $\|B_h^{-1}\| \epsilon_0$ . Пусть требуется найти  $\varphi_h$  с точностью  $\epsilon > 0$ , так что  $\|\tilde{\varphi}_h - \varphi_h\| \leq \epsilon$ ; это возможно при условии

$$\|B_h^{-1}\| \epsilon_0 \leq \epsilon.$$

Отсюда определяем допустимый шаг  $h \geq h_0$ , т. е.  $h_0$ .

Поясним это на конкретной задаче (17). Для неё имеем

$$\|B_h^{-1}\| = \|A_h\| = \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2} \leq \frac{4}{h^2},$$

и условие  $\|B_h^{-1}\|\varepsilon_0 = \Delta\varepsilon_0 \leq \varepsilon$  выполнено, если  $4\varepsilon_0/h^2 \leq \varepsilon$  или

$$h \geq h_0 = 2\sqrt{\varepsilon_0/\varepsilon}.$$

Отсюда видно, что точность задания входных данных  $\varepsilon_0$  должна быть более высокой, чем точность  $\varepsilon$  определения решения.

Пусть, например, заданы погрешность правой части  $\varepsilon_0 = 10^{-8}$  и требуемая точность  $\varepsilon = 10^{-4}$ . Тогда  $h_0 = 2 \cdot 10^{-2} = 1/50$ , т. е. точность  $\varepsilon = 10^{-4}$  можно получить только на сетке с шагом  $h \geq 1/50$ . Если же, например,  $\varepsilon_0 = \frac{1}{4} \cdot 10^{-4}$ ,  $\varepsilon = 10^{-4}$ , то  $h_0 = 1$  и точность  $\varepsilon = 10^{-4}$  нельзя достичь ни на какой сетке при такой точности задания входных данных.

**6. Аппроксимация и сходимость.** При решении задачи (14) разностным методом надо знать, с какой точностью решение разностной задачи приближает решение исходной задачи. Для оценки погрешности, допускаемой при замене (14) разностной схемой (15), надо сравнить решения этих задач. Это сравнение будем проводить в пространстве  $H_h$  сеточных функций. Обозначим через  $u_h(x)$  значения функций  $u(x)$  — точного решения задачи (14) — на сетке  $\omega_h$ :  $u_h \in H_h$ . Рассмотрим погрешность

$$z_h = y_h - u_h,$$

где  $y_h$  — решение задачи (15). Подставляя  $y_h = z_h + u_h$  в (15) и считая  $u = u(x)$  заданной функцией, получим для  $z_h$  разностную задачу

$$L_h z_h = \psi_h, \quad x \in \omega_h; \quad l_h z_h = v_h, \quad x \in \gamma_h, \quad (24)$$

где  $\psi_h = \varphi_h - L_h u_h$  называют *погрешностью аппроксимации для уравнения  $L_h y_h = \varphi_h$  на решении  $u = u(x)$  уравнения  $Lu = f(x)$  (невязка для разностной схемы на решении)*,  $v_h = \mu_h - l_h u_h$  — погрешность аппроксимации для разностного краевого условия  $l_h y_h = \mu_h$  на решении задачи (14).

Будем говорить, что:

разностная схема (15) *сходится*, если

$$\|z_h\|_{(1_h)} \rightarrow 0 \quad \text{при} \quad |h| \rightarrow 0;$$

разностная схема (15) имеет *точность  $m$ -го порядка*

или сходится со скоростью  $O(|h|^m)$ , если

$$\|z_h\|_{(1_h)} = \|y_h - u_h\|_{(1_h)} \leq M |h|^m$$

или

$$\|z_h\|_{(1_h)} = O(|h|^m), \quad m > 0,$$

где  $M > 0$  — постоянная, не зависящая от  $h$ .

Разностная схема (15) имеет  $m$ -й порядок аппроксимации на решении, если

$$\|\psi_h\|_{(2_h)} = O(|h|^m), \quad \|v_h\|_{(3_h)} = O(|h|^m), \quad m > 0. \quad (25)$$

Оценка невязок  $\psi_h$  и  $v_h$  проводится в предположении, что решение исходной задачи существует и имеет столько производных, сколько требуется при получении  $m$ -го порядка аппроксимации.

Приведем два примера оценки  $\psi_h$ .

Примеры. 1. Имеется задача

$$L_h y = -y_{xx} = \varphi(x), \quad x = ih, \quad 1 \leq i \leq N-1, \quad y_0 = y_N = 0, \quad (26)$$

$$Lu = -u'' = f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0.$$

В этом случае краевые условия удовлетворяются точно,  $v_h = 0$  (индекс  $h$  у  $\varphi(x)$ ,  $u(x)$  пока опускаем) и

$$\begin{aligned} \psi_h = \varphi - L_h u &= \varphi + u_{xx} = \varphi + \left( u'' + \frac{1}{2} h^2 u^{IV} + O(h^4) \right) = \\ &= (\varphi + u'') + \frac{h^2}{12} u^{IV} + O(h^4) = \varphi - f + O(h^2), \end{aligned}$$

так как  $u'' = -f(x)$ . Отсюда видно, что  $\|\psi_h\|_c = O(h^2)$ , если положить  $\varphi = f$  или  $\varphi = f + O(h^2)$ .

В п. 1 мы оценивали погрешность  $\psi = L_h v_h - (Lu)_h$  для произвольной функции. При оценке погрешности  $z_h = y_h - u_h$  используется невязка  $\psi_h$ , характеризующая погрешность аппроксимации оператора  $Lu - f$  оператором  $L_h u_h - \varphi_h$  на решении  $u = u(x)$  исходной задачи. Учитывая, что  $f - Lu = 0$ , представим  $\psi_h = \varphi_h - L_h u_h$  в виде

$$\psi_h = (\varphi_h - L_h u_h) - (f - Lu)_h =$$

$$= (\varphi_h - f_h) - (L_h u_h - (Lu)_h) = \psi_h^{(1)} + \psi_h^{(2)},$$

где  $\psi_h^{(1)} = - (L_h u_h - (Lu)_h)$ ,  $\psi_h^{(2)} = \varphi_h - f_h$ ,  $\psi_h^{(1)}$  — погрешность аппроксимации  $L$  оператором  $L_h$  на решении  $u = u(x)$  задачи (6),  $\psi_h^{(2)}$  — погрешность аппроксимации правой части уравнения. Требование  $\|\psi_h\|_{(2_h)} = O(|h|^m)$ ,

очевидно, выполнено, если  $\|\psi_h^{(1)}\|_{(2h)} = O(|h|^m)$ ,  $\|\psi_h^{(2)}\|_{(2h)} = O(|h|^m)$ . Однако эти условия не являются необходимыми для оценки  $\|\psi_h\|_{(2h)} = O(|h|^m)$ , о чем свидетельствует следующий пример.

2. Первая краевая задача (6). Вычислим

$$-\psi_h^{(1)} = u_{xx} - u'' = \frac{1}{12} h^2 u^{\text{IV}} + O(h^4) = O(h^2).$$

Пусть  $\varphi = f + \frac{1}{12} h^2 f_{xx}$ , т. е.  $\varphi - f = O(h^2)$ . Отсюда видно, что  $\psi_h^{(1)} = O(h^2)$  и  $\psi_h^{(2)} = O(h^2)$ , однако схема имеет четвертый порядок аппроксимации, так как

$$\begin{aligned} \psi_h &= \psi_h^{(1)} + \psi_h^{(2)} = \varphi - f + \frac{h^2}{12} u^{\text{IV}} + O(h^4) = \\ &= \frac{h^2}{12} (f_{xx} + u^{\text{IV}}) + O(h^4) = \frac{h^2}{12} (f'' + u^{\text{IV}}) + O(h^4), \end{aligned}$$

$\psi_h = O(h^4)$ , так как  $u^{\text{IV}} + f''(x) = 0$  в силу уравнения  $u'' + f(x) = 0$ .

7. Связь устойчивости и аппроксимации со сходимостью. Рассмотрим линейную разностную схему (15). Если схема устойчива и аппроксимирует исходную задачу, то она сходится (обычно говорят: «из устойчивости и аппроксимации следует сходимость схемы»). В самом деле, для погрешности  $z_h = y_h - u_h$  мы получаем, в силу линейности  $L_h$  и  $l_h$ , задачу (24), аналогичную задаче (15) для  $y_h$ . Поэтому, если схема (15) устойчива, т. е. верна оценка (16), то и для  $z_h$  верна оценка

$$\|z_h\|_{(1h)} \leq M_1 \|\psi_h\|_{(2h)} + M_2 \|v_h\|_{(3h)}. \quad (27)$$

Отсюда видно, что

$$\|z_h\|_{(1h)} = \|y_h - u_h\|_{(1h)} = O(|h|^m),$$

если

$$\|\psi_h\|_{(2h)} = O(|h|^m), \quad \|v_h\|_{(3h)} = O(|h|^m).$$

Таким образом, изучение сходимости и порядка точности разностных схем сводится к изучению погрешности аппроксимации и устойчивости, т. е. к получению априорных оценок (16).

Пример. Для разностной схемы (17) ( $y_{xx,i} = -\varphi_i$ ,  $i = 1, 2, \dots, N-1$ ,  $y_0 = 0$ ,  $y_N = 0$ ) ранее получена

оценка (23). Погрешность аппроксимации, очевидно, есть  $\|\psi_h\|_{C_h} = O(h^2)$  при  $\varphi_i = f_i$ ,  $\|\psi_h\|_{C_h} = O(h^4)$  при  $\varphi_i = f_i + \frac{h^2}{12}f_{xx,i}$ . Так как  $z_{xx,i} = -\psi_{h,i}$  при  $i = 1, 2, \dots, N-1$ ,  $z_0 = 0$ ,  $z_N = 0$ , то и для  $z$  верна оценка  $\|z\|_C \leq \frac{1}{2}\|\psi\|_C$ , откуда следует  $\|y_h - u_h\|_C = O(h^m)$ , где  $m = 2$  при  $\varphi = f$ ,  $m = 4$  при  $\varphi = f + \frac{h^2}{12}f_{xx}$ .

Тем самым изучение схемы (26) завершено (изучение схемы (26) фактически продемонстрировано на последних трех примерах). Это — типичный пример того, как проводится изучение разностных схем.

## § 2. Однородные трехточечные разностные схемы

**1. Исходная задача.** Рассмотрим первую краевую задачу для обыкновенного дифференциального уравнения второго порядка:

$$\begin{aligned} Lu &= \frac{d}{dx} \left( k(x) \frac{du}{dx} \right) - q(x)u = -f(x), \quad 0 < x < 1, \\ k(x) &\geq q > 0, \quad q(x) \geq 0, \quad u(0) = \mu_1, \quad u(1) = \mu_2. \end{aligned} \tag{1}$$

Такое уравнение описывает стационарное, т. е. не меняющееся во времени распределение температуры (стационарное уравнение теплопроводности) или концентрации (уравнение диффузии). Если  $u = u(x)$  — температура, то  $W(x) = -k(x) \frac{du}{dx}$  — тепловой поток ( $k(x)$  — коэффициент теплопроводности).

Задача (1) имеет единственное решение, если  $k(x)$ ,  $q(x)$ ,  $f(x)$  — кусочно-непрерывные функции. Если  $k(x)$  имеет разрыв первого рода в точке  $x = \xi$ , так что  $[k] = k(\xi + 0) - k(\xi - 0) \neq 0$ , то в этой точке должны быть непрерывны температура  $u$  и тепловой поток —  $(ku')$ :

$$[u] = 0, \quad [ku'] = 0 \quad \text{при } x = \xi.$$

Возможны и другие краевые условия при  $x = 0$ ,  $x = 1$ :  $ku' = \sigma_1 u - \mu_1$  при  $x = 0$ ,  $-ku' = \sigma_2 u - \mu_2$  при  $x = 1$ . Если  $\sigma_1 > 0$ , то это условие третьего рода, при  $\sigma_1 = 0$  имеем условие второго рода ( $ku' = -\mu_1$  при  $x = 0$ ). Возможны комбинации различных условий при  $x = 0$  и  $x = 1$ .

**2. Трехточечные разностные схемы.** На отрезке  $0 \leq x \leq 1$  введем равномерную сетку  $\overline{\omega}_h = \{x_i = ih,$

$i = 0, 1, \dots, N$ } с шагом  $h = 1/N$  и выберем трехточечный шаблон  $(x_{i-1}, x_i, x_{i+1})$ , на котором и будем писать разностную схему, аппроксимирующую задачу (1). Любое разностное уравнение на этом шаблоне будет иметь вид

$$b_i y_{i+1} - c_i y_i + a_i y_{i-1} = -h^2 \varphi_i, \quad (2)$$

где  $a_i, b_i, c_i$  — коэффициенты, зависящие от  $k(x), q(x)$  и  $h$ . Они пока не определены. Перепишем (2) иначе:

$$\frac{1}{h} \left( b_i \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) - d_i y_i = -\varphi_i, \quad (3)$$

$$d_i = (c_i - a_i - b_i)/h^2.$$

Будем говорить, что разностная схема *однородна*, если ее коэффициенты во всех узлах сетки для любых коэффициентов дифференциального уравнения вычисляются по одним и тем же формулам. Так, если ввести функционалы  $A[\bar{k}(s)], B[\bar{k}(s)], D[\bar{k}(s)], F[\bar{f}(s)]$ , определенные для любых кусочно-непрерывных функций на отрезке  $-1 \leq s \leq 1$ , и вычислять коэффициенты схемы (3) по формулам

$$a_i = A[k(x_i + sh)], \quad b_i = B[k(x_i + sh)],$$

$$d_i = D[k(x_i + sh)], \quad \varphi_i = F[f(x_i + sh)], \quad \bar{k}(s) = k(x_i + sh),$$

то такая схема будет однородной. Приведем простейшие функционалы

$$A[\bar{k}(s)] = \bar{k}(-0,5), \quad a_i = k_{i-1/2} = k(x_i - 0,5h),$$

$$F(\bar{f}(s)) = f(0), \quad \varphi_i = f_i = f(x_i) \quad \text{и т. д.}$$

Если схема однородна, то удобно пользоваться безиндексной системой обозначений:

$$\Lambda y = \frac{1}{h} (by_x - ay_{\bar{x}}) - dy = -\varphi, \quad x \in \omega_h,$$

$$y(0) = \mu_1, \quad y(1) = \mu_2, \quad (4)$$

где

$$a = a(x), \quad b = b(x), \quad y = y(x), \quad x = ih \in \omega_h,$$

$$y_x = (y(x+h) - y(x))/h, \quad y_{\bar{x}} = (y(x) - y(x-h))/h.$$

Для разрешимости задачи (4) достаточно, чтобы  $a > 0, b > 0, d \geq 0$ , при этом решение можно найти методом прогонки (см. гл. I, § 3).

**3. Условия аппроксимации.** Вычислим погрешность аппроксимации схемы (4):

$$\begin{aligned}\psi &= (\Lambda v + \varphi) - (Lv + f) = (\Lambda v - Lv) + (\varphi - f) = \\ &= \left[ \frac{1}{h} (bv_x - av_{\bar{x}}) - (kv')' \right] - (d - q)v + (\varphi - f),\end{aligned}$$

где  $v(x)$  — произвольная достаточно гладкая функция,  $k, q, f$  также имеют нужное по ходу изложения число производных. Воспользуемся формулой Тейлора:

$$v(x \pm h) = v(x) \pm hv'(x) + \frac{h^2}{2}v''(x) \pm \frac{h^3}{6}v'''(x) + O(h^4)$$

и найдем

$$\begin{aligned}v_x &= v' + \frac{h}{2}v'' + \frac{h^2}{6}v''' + O(h^3), \\ v_{\bar{x}} &= v' - \frac{h}{2}v'' + \frac{h^2}{6}v''' + O(h^3).\end{aligned}$$

Подставим эти выражения для  $v_x$  и  $v_{\bar{x}}$  в формулу для  $\psi$ :

$$\begin{aligned}\psi &= \left( \frac{1}{h}(b-a) - k' \right) v' + \left( \frac{b+a}{2} - k \right) v'' + \frac{h(b-a)}{6}v''' - \\ &\quad - (d-q)v + (\varphi - f) + O(h^2).\end{aligned}$$

Отсюда видно, что схема имеет второй порядок аппроксимации, если выполнены условия

$$\begin{aligned}\frac{b-a}{h} &= k'(x) + O(h^2), \quad \frac{b+a}{2} = k(x) + O(h^2), \\ d &= q(x) + O(h^2), \quad \varphi = f(x) + O(h^2).\end{aligned}\quad (5)$$

В этом случае  $\psi = O(h^2)$ .

Схема (4) с коэффициентами

$$\begin{aligned}b_i &= k_{i+1/2}, \quad a_i = k_{i-1/2}, \quad d_i = q_i, \quad \varphi_i = f_i, \\ b_i &= \frac{k_i + 2k_{i+1/2} + k_{i+1}}{4}, \quad a_i = k_{i-1/2}, \quad d_i = q_i, \quad \varphi_i = f_i\end{aligned}$$

удовлетворяет условиям (5) второго порядка аппроксимации, а схема с коэффициентами

$$b_i = k_{i+1}, \quad a_i = \frac{k_i + k_{i+1}}{2}$$

не удовлетворяет даже условию первого порядка аппроксимации, так как

$$\frac{1}{h} (b_i - a_i) - k'_i = O(1).$$

### § 3. Консервативные разностные схемы

**1. Однородные консервативные схемы.** В § 4 гл. I мы установили, что необходимым и достаточным условием самосопряженности разностного оператора  $\Lambda y$  (симметричности матрицы) является условие  $b_i = a_{i+1}$ . В этом случае задача (2) из § 2 принимает вид

$$\begin{aligned} \Lambda y = \frac{1}{h} \left[ a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right] - d_i y_i = -\varphi_i, \\ i = 1, 2, \dots, N-1, \quad y_0 = \mu_1, \quad y_N = \mu_2. \end{aligned} \quad (1)$$

Уравнение

$$a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} - h d_i y_i = -h \varphi_i \quad (2)$$

является сеточным аналогом уравнения баланса тепла на интервале  $(x_{i-1/2}, x_{i+1/2})$ :

$$w_{i+1/2} - w_{i-1/2} - \int_{x_{i-1/2}}^{x_{i+1/2}} qu \, dx = - \int_{x_{i-1/2}}^{x_{i+1/2}} f(x) \, dx, \quad w = ku',$$

(которое получается при интегрировании уравнения (1) из § 2 по отрезку  $x_{i-1/2} \leq x \leq x_{i+1/2}$ ), и называется *консервативной* схемой, т. е. схемой, для которой выполняются разностные аналоги физических законов сохранения.

Требование  $b_i = a_{i+1}$  для однородной схемы означает, что  $B[k(x+sh)] = A[k(x+(s+1)/h)]$ , или  $B[\bar{k}(s)] = A[\bar{k}(s+1)]$  для любых кусочно-непрерывных функций  $\bar{k}(s)$  на отрезке  $[-1, 1]$ . Это возможно только в том случае, когда функционал  $A[\bar{k}(s)]$  не зависит от значений  $\bar{k}(s)$  при  $0 \leq s \leq 1$ , а  $B[\bar{k}(s)]$  — от значений  $\bar{k}(s)$  при  $-1 \leq s \leq 0$ , так что  $a(x) = A[k(x+sh)]$  при  $-1 \leq s \leq 0$ . Коэффициент  $a(x)$  консервативной схемы зависит только от значений  $k(x)$  на отрезке  $[x-h, x]$ . Условия второго порядка аппроксимации (5) из § 2 для консервативной

схемы (2) принимают вид

$$\frac{a(x+h) - a(x)}{h} = k'(x) + O(h^2), \quad (3)$$

$$\frac{a(x+h) + a(x)}{2} = k(x) + O(h^2),$$

$$d(x) = q(x) + O(h^2), \quad \varphi(x) = f(x) + O(h^2). \quad (4)$$

Отсюда, в частности, следует, что

$$a(x) = k(x) - \frac{1}{2}hk'(x) + O(h^2) = k(x - \frac{1}{2}h) + O(h^2).$$

Запишем консервативную схему (2) в безиндексных обозначениях:

$$(ay_x)_x - d(x)y = -\varphi(x), \\ x = ih \equiv \omega_h, \quad y(0) = \mu_1, \quad y(1) = \mu_2. \quad (5)$$

Будем требовать, чтобы выполнялись также условия

$$a \geq c_1 > 0, \quad d \geq 0. \quad (6)$$

На практике следует пользоваться простыми формулами для  $a$ ,  $d$  и  $\varphi$ , например,  $a_i = k_{i-1/2}$ ,  $d_i = q_i$ ,  $\varphi_i = f_i$ .

Если разрыв функции  $k(x)$  находится в узле  $x = x_i$  сетки, то вычислим коэффициенты однородной схемы:

$$a_i = k_{i-1/2} \text{ или } a_i = \frac{1}{2}(k(x_{i-1} + 0) + k(x_i - 0)), \\ d_i = \frac{1}{2}(q(x_i - 0) + q(x_i + 0)), \quad \varphi_i = \frac{1}{2}(f(x_i - 0) + f(x_i + 0)).$$

В этом случае условия (3) выполнены всюду, а условия (4) заменяются условиями

$$d_i - \frac{1}{2}(q_{i-0} + q_{i+0}) = O(h^2), \quad \varphi_i - \frac{1}{2}(f_{i-0} + f_{i+0}) = O(h^2).$$

Приведем пример схемы, коэффициенты которой вычисляются путем интегрирования по интервалам сетки:

$$a_i = \left( \frac{1}{h} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} \right)^{-1} = \left( \int_{-1}^0 \frac{ds}{k(x_i + sh)} \right)^{-1};$$

$$\varphi_i = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} f(x) dx = \int_{-1/2}^{1/2} f(x_i + sh) ds,$$

$$d_i = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} q(x) dx = \int_{-1/2}^{1/2} q(x_i + sh) ds.$$

Очевидно, что условия (3), (4) выполнены,

**2. Погрешность аппроксимации.** Рассмотрим консервативную схему второго порядка аппроксимации. Пусть  $u = u(x)$  — точное решение задачи

$$Lu = (ku')' - q(x)u = -f(x), \quad 0 < x < 1,$$

$$u(0) = \mu_1, \quad u(1) = \mu_2, \quad (6)$$

а  $y_i = y(x_i)$  — решение разностной краевой задачи (5). Рассмотрим погрешность схемы, т. е. сеточную функцию

$$z(x) = y(x) - u(x), \quad x \in \bar{\omega}_h.$$

Подставляя  $y(x) = z(x) + u(x)$  в уравнение (5) и предполагая, что  $u(x)$  — заданная функция, получим для погрешности  $z(x)$  задачу

$$\Lambda z = (az_x)_x - dz = -\psi(x), \quad x \in \omega_h,$$

$$z(0) = 0, \quad z(1) = 0, \quad a \geq c_1 > 0, \quad d \geq 0, \quad (6')$$

где  $\psi(x) = \Lambda u + \varphi(x) = (au_x)_x - du + \varphi$  — невязка схемы (5) на решении  $u = u(x)$  исходной дифференциальной задачи.

Учитывая, что  $Lu + f = 0$ , напишем

$$\begin{aligned} \psi &= (\Lambda u + \varphi) - (Lu + f) = (\Lambda u - Lu) + (\varphi - f) = \\ &= [(au_x)_x - (ku')'] - (d - q)u + (\varphi - f). \end{aligned}$$

По предложению, схема (5) удовлетворяет условиям второго порядка аппроксимации. Это значит, что  $\psi = O(h^2)$ , если  $k \in C^{(3)}$ ,  $q, f \in C^{(2)}$ ,  $u \in C^{(4)}$ , и, следовательно,

$$\|\psi\|_c = O(h^2).$$

При этих предположениях схема имеет второй порядок точности.

Однако этот же порядок точности сохраняется и при более слабых требованиях гладкости:

$$k(x), \quad q(x), \quad f(x) \in C^{(2)}, \quad u \in C^{(3)}. \quad (7)$$

**Лемма.** *Если выполнены условия (7), то справедлива формула*

$$\frac{(ku')_{i+1/2} - (ku')_{i-1/2}}{h} = (ku')'_i + O(h^2), \quad (8)$$

где  $u = u(x)$  — решение уравнения (6),

**Доказательство.** Воспользуемся формулой Тейлора:

$$v_{i \pm 1/2} = v_i \pm \frac{1}{h} h v'_i + \frac{h^2}{8} v''_i + \frac{h^3}{48} v'''_i (x_i \pm \theta h),$$

$$0 \leq \theta \leq 1, \quad \frac{1}{h} (v_{i+1/2} - v_{i-1/2}) = v'_i + O(h^2).$$

Подставляя сюда  $v = ku'$  и учитывая, что  $(ku')'' = (qu - f)', (ku')''' = (qu - f)''$ , получаем формулу (8).

В силу леммы погрешность аппроксимации  $\psi$  можно представить в виде

$$\psi_i = \eta_{x,i} + \Psi_i^*, \quad \eta_i = (au_x)_i - (ku')_{i-1/2}, \quad \Psi_i^* = O(h^2)$$

при условиях (7).

Учитывая далее, что

$$a_i = k_{i-1/2} + O(h^2) \text{ при } k(x) \in C^{(2)},$$

$$u_{x,i} = \frac{u_i - u_{i-1}}{h} = (u')_{i-1/2} + O(h^2) \text{ при } u \in C^{(3)},$$

получаем  $\eta_i = O(h^2)$ . В самом деле,  $u_i = u_{i-1/2} + \frac{1}{2}hu'_{i-1/2} + \frac{1}{8}h^2u''_{i-1/2} + O(h^3)$ ,

$$u_{i-1} = u_{i-1/2} - \frac{1}{2}hu'_{i-1/2} + \frac{1}{8}h^2u''_{i-1/2} + O(h^3),$$

$$u_{x,i} = u'_{i-1/2} + O(h^2),$$

$$a_i u_{x,i} = (k_{i-1/2} + O(h^2)) (u'_{i-1/2} + O(h^2)) = (ku')_{i-1/2} + O(h^2),$$

$$\eta_i = O(h^2).$$

Ниже будет получена априорная оценка  $\|z\|_C$  непосредственно через  $\eta$  и  $\Psi^*$ .

**3. Априорные оценки.** Переидем к оценке погрешности  $z$  через  $\psi$ . Прежде всего напомним оценку, полученную в § 5 гл. I с помощью метода прогонки:

$$\|z\|_C \leq \frac{1}{c_1} \sum_{i=1}^{N-1} h \sum_{k=1}^i h |\psi_k|,$$

откуда следует

$$\|z\|_C \leq \frac{1}{2c_1} \|\psi\|_C.$$

Покажем, что для решения задачи

$$(az_x)_x - dz = -\mu_x, \quad x \in \omega_h, \quad z(0) = z(1) = 0,$$

$$a \geq c_1 > 0, \quad d \geq 0$$

справедлива оценка

$$\|z\|_C \leq \frac{2}{c_1} (1, |\mu|), \quad (9)$$

где обозначено  $(y, v) = \sum_{i=1}^N y_i v_i h$ .

Представим  $z$  в виде суммы  $z = w + v$ , где  $w$  и  $v$  — решения задач

$$(aw_x)_x = -\mu_x, \quad x \in \omega_h, \quad w(0) = w(1) = 0;$$

$$\Lambda v = (av_x)_x - dv = -dw, \quad x \in \omega_h, \quad v(0) = v(1) = 0.$$

Функцию  $w$  мы найдем в явном виде, а для оценки  $v$  воспользуемся принципом максимума. Из уравнения

$$(aw_x + \mu)_x = 0, \quad (aw_x)_{i+1} + \mu_{i+1} = (aw_x)_i + \mu_i$$

следует, что  $aw_x + \mu = \text{const} = c_0$ . Проведем очевидные преобразования:

$$w_i = w_{i-1} + \frac{(c_0 - \mu_i)h}{a_i} = c_0 \sum_{k=1}^i \frac{h}{a_k} - \sum_{k=1}^i \frac{\mu_k}{a_k} h + w_0,$$

$$0 = w_N = c_0 \sum_{k=1}^N \frac{h}{a_k} - \sum_{k=1}^N \frac{\mu_k}{a_k} h;$$

$$c_0 = \sum_{k=1}^N \frac{\mu_k}{a_k} h \sqrt{\sum_{k=1}^N \frac{h}{a_k}}.$$

Вводя обозначение

$$\alpha_i = \sum_{k=1}^i \frac{h}{a_k} \sqrt{\sum_{k=1}^N \frac{h}{a_k}}, \quad 0 < \alpha_i \leq 1,$$

найдем

$$w_i = \alpha_i \sum_{k=1}^N \frac{h \mu_k}{a_k} - \sum_{k=1}^i \frac{h \mu_k}{a_k}.$$

Отсюда следует

$$\begin{aligned} |w_i| &= \left| - (1 - \alpha_i) \sum_{k=1}^i \frac{h\mu_k}{a_k} + \alpha_i \sum_{k=i+1}^N \frac{h\mu_k}{a_k} \right| \leqslant \\ &\leqslant (1 - \alpha_i) \sum_{k=1}^i \frac{h|\mu_k|}{a_k} + \alpha_i \sum_{k=i+1}^N \frac{h|\mu_k|}{a_k} \leqslant \sum_{k=1}^N \frac{h|\mu_k|}{a_k}. \end{aligned}$$

Далее остается учесть, что  $a_k \geq c_1 > 0$ , и мы получаем

$$\|w\|_C \leq \frac{1}{c_1} \sum_{k=1}^N h|\mu_k| = \frac{1}{c_1} (1, |\mu|). \quad (10)$$

Для оценки  $v$  воспользуемся теоремой 4 из § 5 гл. I:

$$\|v\|_C \leq \|w\|_C. \quad (11)$$

Объединяя неравенства (10) и (11), имеем

$$\|z\|_C = \|w + v\|_C \leq 2\|w\|_C \leq \frac{2}{c_1} (1, |\mu|),$$

т. е. доказана оценка (9).

Обратимся теперь к задаче (6'), где  $\psi = \eta_x + \psi^*$ . Представим  $\psi$  в виде

$$\psi = \mu_x, \text{ где } \mu_i = \eta_i + \sum_{h=1}^{i-1} h\psi_h^*, \quad (12)$$

и воспользуемся оценкой (9). Тогда для решения задачи (6') получим следующие априорные оценки:

$$\begin{aligned} \|z\|_C &\leq \frac{2}{c_1} \left\{ (1, |\eta|) + \sum_{h=1}^N h \left| \sum_{k=1}^{i-1} h\psi_k^* \right| \right\}, \\ \|z\|_C &\leq \frac{2}{c_1} \{ (1, |\eta|) + (1, |\psi^*|) \}. \end{aligned} \quad (13)$$

Остается показать, что имеет место формула (12). В самом деле, обозначая  $\rho_i = \sum_{h=1}^{i-1} h\psi_h^*$ , видим, что  $\rho_{i+1} - \rho_i = h\psi_i^*$ , т. е.  $\psi_i^* = \rho_{x,i}$  и  $\psi = \eta_x + \rho_x = \mu_x$ , где  $\mu_i = \eta_i + \rho_i$ .

**4. Сходимость и точность разностной схемы.** Перейдем к оценке точности разностной схемы. Предполагая, что

$$k(x), q(x), f(x) \in C^{(2)}, \quad u(x) \in C^{(3)},$$

получаем  $\eta(x) = O(h^2)$ ,  $\psi^* = O(h^2)$ . Теперь остается вос-

пользоваться априорной оценкой (13), которую можно заменить и более грубой оценкой

$$\|z\|_C \leq \frac{2}{c_1} (\|\eta\|_C + \|\psi^*\|_C).$$

Отсюда следует, что схема (5) равномерно сходится со вторым порядком, т. е.  $\|z\|_C = \|y - u\|_C \leq Mh^2$ , если выполнены условия (7).

Более трудным является доказательство сходимости схемы в классе разрывных коэффициентов  $k(x)$ ,  $q(x)$ ,  $f(x)$ . Для простоты будем рассматривать случай, когда  $k(x)$  имеет разрыв первого рода в одной точке, а  $q(x)$  и  $f(x)$  непрерывны и принадлежат классу  $C^{(2)}$ .

Обозначим через  $Q^{(k)}[a, b]$  множество кусочно-непрерывных функций, заданных на отрезке  $[a, b]$  и имеющих на  $[a, b]$   $k$  кусочно-непрерывных производных.

Итак, пусть  $k(x) \in Q^{(2)}$ ,  $q(x)$ ,  $f(x) \in C^{(2)}$  и  $k(x)$  имеет разрыв первого рода в точке  $\xi$  на отрезке  $[x_n, x_{n+1}]$ , так что  $\xi = x_n + \theta h$ ,  $0 \leq \theta \leq 1$ . При  $x = \xi$  выполнены условия сопряжения

$$u_- = u_+, \quad (ku')_- = (ku')_+ = w_0,$$

где

$$v_+ = v(\xi + 0), \quad v_- = v(\xi - 0).$$

Тогда  $\eta_i = O(h^2)$  при  $i \neq n+1$ ,  $\psi_i^* = O(h^2)$  при всех  $i = 1, 2, \dots, N-1$ ,  $\eta_{n+1} = a_{n+1}u_{x,n} - (ku')_{n+1/2}$ . Подставляя сюда

$$u_{n+1} = u(\xi) + (1-\theta)hu'_+ + O(h^2),$$

$$u_n = u(\xi) - \theta hu'_- + O(h^2),$$

$$\begin{aligned} u_{x,n} &= (u_{n+1} - u_n)/h = \theta u'_- + (1-\theta)u'_+ + O(h) = \\ &= \theta \frac{(ku')_-}{k_-} + (1-\theta) \frac{(ku')_+}{k_+} + O(h) = \\ &= w_0 \left( \frac{\theta}{k_-} + \frac{1-\theta}{k_+} \right) + O(h), \end{aligned}$$

$$(ku')_{n+1/2} = (ku')_- + O(h) = w_0 + O(h) \quad \text{при } \theta > \frac{1}{2},$$

$$(ku')_{n+1/2} = (ku')_+ + O(h) = w_0 + O(h) \quad \text{при } \theta < \frac{1}{2},$$

получаем

$$\eta_{n+1} = w_0 \left[ a_{n+1} \left( \frac{\theta}{k_-} + \frac{1-\theta}{k_+} \right) - 1 \right] + O(h),$$

т. е.  $\eta_{n+1} = O(1)$  для любой схемы, и только для схемы с коэффициентом

$$\overset{\circ}{a}_i = \left[ \frac{1}{h} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} \right]^{-1}$$

имеем  $\eta_{n+1} = O(h)$ . В самом деле,

$$\frac{1}{\overset{\circ}{a}_{n+1}} = \frac{1}{h} \int_{x_n}^{\xi} \frac{dx}{k(x)} + \frac{1}{h} \int_{\xi}^{x_{n+1}} \frac{dx}{k(x)} = \frac{\theta}{k_-} + \frac{1-\theta}{k_+} + O(h),$$

т. е.  $\overset{\circ}{a}_{n+1} \left( \frac{\theta}{k_-} + \frac{1-\theta}{k_+} \right) = 1 + O(h)$  и, следовательно,  $\eta_{n+1} = O(h)$ . В правую часть неравенства (13) входит величина

$$(1, |\eta|) = \sum_{i=1, i \neq n+1}^N h |\eta_i| + h |\eta_{n+1}|.$$

Тем самым доказана следующая теорема.

**Теорема.** В классе разрывных коэффициентов  $k(x) \in Q^{(2)}$ ,  $q(x), f(x) \in C^{(2)}$  любая однородная разностная схема (5) второго порядка аппроксимации имеет первый порядок точности, а схема с коэффициентом  $a_i = \overset{\circ}{a}_i$  имеет второй порядок точности.

#### § 4. Однородные схемы на неравномерных сетках

**1. Консервативная схема на неравномерной сетке.** Выберем на отрезке  $0 \leq x \leq 1$  произвольную неравномерную сетку

$$\hat{\omega}_h = \{x_i, i = 0, 1, \dots, N, x_0 = 0, x_N = 1\}.$$

Для получения трехточечной консервативной схемы на неравномерной сетке напишем уравнение баланса на отрезке  $[x_{i-1/2}, x_{i+1/2}]$ :

$$w_{i+1/2} - w_{i-1/2} - \int_{x_{i-1/2}}^{x_{i+1/2}} q u dx = - \int_{x_{i-1/2}}^{x_{i+1/2}} f(x) dx, \quad w = k u'.$$

Оно пишется одинаково как для равномерной, так и для

неравномерной сетки. Остается аппроксимировать входящие в уравнение баланса интегралы и производные:

$$w_{i-1/2} = (ku')_{i-1/2} \sim a_i \frac{u_i - u_{i-1}}{h_i}, \quad h_i = x_i - x_{i-1},$$

$$\int_{x_{i-1/2}}^{x_{i+1/2}} f(x) dx \sim \Phi_i \tilde{h}_i, \quad \int_{x_{i-1/2}}^{x_{i+1/2}} qu dx \sim d_i u_i \tilde{h}_i,$$

$$\tilde{h}_i = \frac{1}{2} (h_i + h_{i+1}),$$

где  $d_i$  и  $\Phi_i$  — некоторые сеточные функции. В результате получаем разностную схему

$$\frac{1}{\tilde{h}_i} \left[ a_{i+1} \frac{y_{i+1} - y_i}{h_{i+1}} - a_i \frac{y_i - y_{i-1}}{h_i} \right] - d_i y_i = -\varphi_i, \\ i = 1, 2, \dots, N-1, \quad y_0 = \mu_1, \quad y_N = \mu_2. \quad (1)$$

Для  $d_i$  и  $\varphi_i$  воспользуемся простейшими формулами  $\varphi_i = f_i$ ,  $d_i = q_i$ ,  $i = 1, 2, \dots, N-1$ . Коэффициент  $a_i$  определяется значениями  $k(x)$  на интервале  $(x_{i-1}, x_i)$ , поэтому его можно взять таким же, как и на равномерной сетке; так что  $a_i = k_{i-1/2} + O(h_i^2)$  при  $k(x) \in C^{(2)}$ .

**2. Погрешность аппроксимации.** Введем обозначения

$$y_{\bar{x},i} = \frac{y_i - y_{i-1}}{h_i}, \quad y_{x,i} = \frac{y_{i+1} - y_i}{h_{i+1}}, \quad \hat{y}_{\bar{x},i} = \frac{y_{i+1} - y_i}{\tilde{h}_i}$$

и запишем разностную схему в виде

$$(ay_{\bar{x}})_{\bar{x}} - dy = -\varphi, \quad x = x_i \in \hat{\omega}_h, \quad y_0 = \mu_1, \quad y_N = \mu_2. \quad (1)$$

Полагая  $z = y - u$ , получим для  $z$  уравнение

$$(az_{\bar{x}})_{\bar{x}} - dz = -\psi, \quad x \in \hat{\omega}_h, \quad z_0 = z_N = 0, \quad (2)$$

где

$$\psi = \Lambda u + \varphi = (au_{\bar{x}})_{\bar{x}} - du + \varphi \quad (3)$$

есть невязка для схемы (4) на решении  $u = u(x)$ .

**Лемма 1.** Если  $qu \in C^{(2)}$ ,  $f \in C^{(2)}$ , то для погрешности аппроксимации  $\psi$  справедлива формула

$$\psi = \eta_{\bar{x}} + \psi^*, \quad (4)$$

где  $\eta_i = (au_{\bar{x}})_i - (ku')_{i-1/2} - h_i^2(qu - f)'_i/8$ ,  $\psi_i^* = O(\tilde{h}_i^2)$  при  $\varphi_i = f_i$ ,  $d_i = q_i$ .

Воспользуемся тождеством из п. 1, записав его в виде

$$0 = w_{\tilde{x},i} - \frac{1}{h_i} \int_{x_{i-1/2}}^{x_{i+1/2}} (qu - f) dx, \quad w_i = (ku')_{i-1/2}.$$

Вычтем это тождество из равенства (3):

$$\psi = [(au_{\tilde{x}})_i - (ku')_{i-1/2}] - (du)_i + \varphi_i + \frac{1}{h_i} \int_{x_{i-1/2}}^{x_{i+1/2}} (qu - f) dx. \quad (5)$$

Интеграл, стоящий справа, представим в виде суммы двух интегралов: от  $x_{i-1/2}$  до  $x_i$  и от  $x_i$  до  $x_{i+1/2}$ ; разлагая затем подынтегральную функцию  $\tilde{f} = qu - f$  в окрестности узла  $x = x_i$ , найдем

$$\begin{aligned} \frac{1}{h_i} \int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{f}(x) dx &= \frac{1}{h_i} \left\{ \int_{x_{i-1/2}}^{x_i} [\tilde{f}_i + (x - x_i)\tilde{f}'_i] dx + O(h_i^3) + \right. \\ &\quad \left. + \int_{x_i}^{x_{i+1/2}} [\tilde{f}_i + (x - x_i)\tilde{f}'_i] dx + O(h_{i+1}^3) \right\} = \\ &= \tilde{f}_i + \frac{1}{8h_i} (h_{i+1}^2 - h_i^2) \tilde{f}'_i + O(h_i^2), \end{aligned}$$

так как  $h_i^3 + h_{i+1}^3 < (2h_i)^3$ . Замена  $h_{i+1}\tilde{f}'_i = h_{i+1}\tilde{f}'_{i+1} + O(h_{i+1}^3)$  дает

$$\frac{1}{h_i} \int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{f}(x) dx = \tilde{f}_i + (h^2 \tilde{f}')_{\tilde{x},i} + O(h_i^2).$$

Подставляя это выражение с  $\tilde{f} = qu - f$  в (5), приходим к формуле (4).

Для оценки  $\eta_i$  по порядку рассмотрим разность  $(au_{\tilde{x}})_i - (ku')_{i-1/2}$  при условии  $k \in C^{(2)}$ ,  $u \in C^{(3)}$ . Пользуясь предположением  $a_i = k_{i-1/2} + O(h_i^2)$  и формулами  $u_i = u_{i-1/2} + h_i u'_{i-1/2}/2 + h_i^2 u''_{i-1/2}/8 + O(h_i^3)$ ,  $u_{i-1} = u_{i-1/2} - h_i u'_{i-1/2}/2 + h_i^2 u''_{i-1/2}/8 + O(h_i^3)$ ,  $u_{\tilde{x},i} = (u_i - u_{i-1})/h_i =$

$= u_{i-1/2} + O(h_i^2)$ , получаем

$$(au_x)_i - (ku')_{i-1/2} =$$

$$= (k_{i-1/2} + O(h_i^2))(u'_{i-1/2} + O(h_i^2)) - (ku')_{i-1/2} = O(h_i^2).$$

Таким образом, справедлива оценка

$$\eta_i = O(h_i^2) \text{ при } k(x), q(x), f(x) \in C^{(2)}, u(x) \in C^{(3)}.$$

**Замечание.** Мы предполагали, что  $d_i$  и  $\varphi_i$  определяются по простейшим формулам:  $d_i = q_i$ ,  $\varphi_i = f_i$ . Если же используются более сложные формулы, например

$$\varphi_i = \frac{h_i f_{i-1/2} + h_{i+1} f_{i+1/2}}{2h_i}, \quad \varphi_i = \frac{1}{h_i} \int_{x_{i-1/2}}^{x_{i+1/2}} f(x) dx,$$

то сеточную функцию  $\psi_i^* = O(\hbar_i^2) - (d_i - q_i)u_i + (\varphi_i - f_i)$  можно представить в виде  $\psi_i^* = \rho_{x,i} + \psi_i^{**}$ , где  $\psi_i^{**} = O(\hbar_i^2)$ ,  $\rho_i = O(h_i^2)$ , и заменить в формуле (4)  $\eta_i$  на сумму  $\eta_i + \rho_i$ :

$$\psi = (\rho_i + \eta_i)_{\widehat{x}} + \psi^{**}, \quad (4')$$

$\rho_i = O(h_i^2)$ ,  $\eta_i = O(h_i^2)$ ,  $\psi_i^{**} = O(\hbar_i^2)$  при  $k, q, f \in C^{(2)}$ ,  $u \in C^{(3)}$ .

**3. Оценка скорости сходимости.** Для задачи (2)–(4) справедлива априорная оценка

$$\|z\|_C \leq \frac{1}{c_1} \{(1, |\eta|) + (1, |\psi^*|)\}, \quad (6)$$

где  $(y, v) = \sum_{i=1}^N y_i v_i h_i$ . Если выполнены условия (7) из § 3, то  $\eta_i = O(h_i^2)$ ,  $\psi_i^* = O(\hbar_i^2)$ .

Подставляя  $\eta_i$  и  $\psi_i^*$  в (6), убеждаемся в том, что справедлива следующая теорема.

**Теорема.** В классе гладких коэффициентов  $k, q, f \in C^{(2)}$  любая схема вида (1) сохраняет второй порядок точности на произвольной последовательности неравномерных сеток.

Учитывая замечание п. 2,  $\psi_i^*$  можно представить в виде  $\psi_i^* = \rho_{x,i} + \psi_i^{**}$ , где  $\rho_i = O(h_i^2)$ ,  $\psi_i^{**} = O(\hbar_i^2)$ . Тогда

вместо (6) верна оценка

$$\|z\|_C \leq \frac{2}{c_1} \{(1, |\eta + \rho|) + (1, |\psi^{**}|)\};$$

теорема о втором порядке точности на неравномерной сетке сохраняет силу.

Если коэффициент  $k(x)$  имеет разрывы первого рода в конечном числе точек, то всегда можно выбрать такую неравномерную сетку  $\omega_h(k)$ , что точки разрыва будут узлами этой сетки. Тогда любая схема будет иметь второй порядок точности.

Итак, любая однородная схема второго порядка аппроксимации ( $\psi = O(h^2)$ ) на равномерной сетке и в классе гладких коэффициентов имеет второй порядок точности при специальном выборе неравномерных сеток  $\omega_h(k)$  в классе разрывных коэффициентов.

**4. Точная схема.** Для задачи (1) из § 2 можно построить точную трехточечную схему, решение которой в узлах произвольной сетки совпадает с точным решением  $u = u(x)$  краевой задачи для дифференциального уравнения. Проиллюстрируем возможность построения точной схемы на частном случае задачи при  $q(x) = 0$ :

$$(ku')' = -f(x), \quad 0 < x < 1, \quad u(0) = 0, \quad u(1) = 0. \quad (7)$$

Проинтегрировав уравнение от  $x_i$  до  $x$ , получим уравнение

$$(ku') - (ku')_i + \int_{x_i}^x f(\xi) d\xi = 0.$$

Разделим его на  $k(x)$  и проинтегрируем по  $x$  сначала от  $x_i$  до  $x_{i+1}$ :

$$u_{i+1} - u_i - (ku')_i \int_{x_i}^{x_{i+1}} \frac{dx}{k(x)} + \int_{x_i}^{x_{i+1}} \frac{dx'}{k(x')} \int_{x_i}^{x'} f(\xi) d\xi = 0, \quad (8)$$

затем от  $x_{i-1}$  до  $x_i$ :

$$u_i - u_{i-1} - (ku')_i \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} + \int_{x_{i-1}}^{x_i} \frac{dx'}{k(x')} \int_{x_i}^{x'} f(\xi) d\xi = 0. \quad (9)$$

Введем обозначение

$$a_i^0 = \left[ \frac{1}{h_i} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} \right]^{-1}.$$

Умножим (8) на  $a_{i+1}^0/h_{i+1}$ , (9) — на  $a_i^0/h_i$  и вычтем из первого результата второй. Получим уравнение

$$\frac{1}{h_i} \left[ a_{i+1}^0 \frac{u_{i+1} - u_i}{h_{i+1}} - a_i^0 \frac{u_i - u_{i-1}}{h_i} \right] + \varphi_i = 0,$$

или

$$(a^0 u_x)_{\hat{x}, i} + \varphi_i = 0, \quad (10)$$

где

$$\varphi_i = \frac{a_i^0}{h_i h_i} \int_{x_{i-1}}^{x_i} \frac{dx'}{k(x')} \int_{x'}^{x_i} f(t) dt + \frac{a_{i+1}^0}{h_{i+1} h_i} \int_{x_i}^{x_{i+1}} \frac{dx'}{k(x')} \int_{x_i}^{x'} f(t) dt.$$

Если положить  $x' = x_i + sh_i$  при  $x_{i-1} \leq x' \leq x_i$  и  $x' = x_i + sh_{i+1}$  при  $x_i \leq x' \leq x_{i+1}$ , то эту формулу можно переписать так:

$$\begin{aligned} \varphi_i = & \frac{h_i a_i^0}{h_i} \int_{-1}^0 \frac{ds}{k(x_i + sh_i)} \int_s^0 f(x_i + \lambda h_i) d\lambda + \\ & + \frac{h_{i+1} a_{i+1}^0}{h_i} \int_0^1 \frac{ds}{k(x_i + sh_{i+1})} \int_0^s f(x_i + \lambda h_{i+1}) d\lambda. \end{aligned}$$

Таким образом, схема (10) является точной на произвольной неравномерной сетке и для любых кусочно-непрерывных функций  $k(x)$  и  $f(x)$ . Конечно, практическое использование этой схемы затруднено тем, что коэффициенты выражаются через интегралы от  $k(x)$  и  $f(x)$  и поэтому для их вычисления надо пользоваться квадратурными формулами.

**5. Повышение порядка точности.** Из предыдущего ясно, что для повышения точности приближенного решения надо либо уменьшать шаг сетки  $h$ , либо повышать порядок точности схемы. Однако схемы повышенного порядка точности целесообразно строить лишь для уравнений с постоянными коэффициентами, так как написание таких схем для уравнений с переменными коэффициентами

сопряжено с большими техническими трудностями и часто приводит к трудоемким алгоритмам. Мы уже приводили пример схемы  $O(h^4)$  для уравнения  $u'' = -f(x)$ .

Рассмотрим теперь уравнение

$$u'' - qu = -f(x), \quad q = \text{const} > 0.$$

Напишем разностную схему на равномерной сетке:

$$\Lambda y = y_{xx} - dy = -\varphi(x)$$

и выберем  $d$  и  $\varphi$  так, чтобы она имела аппроксимацию  $O(h^4)$ . Погрешность аппроксимации равна

$$\begin{aligned} \psi = \Lambda u + \varphi &= (\Lambda u - u'') - (d - q)u + \varphi - f = \\ &= \frac{h^2}{12} u^{IV} - (d - q)u + \varphi - f + O(h^4). \end{aligned}$$

Подставив сюда  $u^{IV} = qu'' - f'' = q(qu - f) - f'' = q^2u - qf - f''$ , получим

$$\varphi = -\left(d - q - \frac{h^2}{12}q^2\right)u + \varphi - \left(f + \frac{h^2}{12}qf + \frac{h^2}{12}f''\right) + O(h^4);$$

следовательно,  $\psi = O(h^4)$ , если положить  $d = q + \frac{h^2}{12}q^2$ ,

$\varphi = f + \frac{h^2}{12}(qf + f'')$ . Порядок точности сохранится, если заменить в формуле для  $\varphi$  производную  $f''$  ее разностной аппроксимацией  $f_{xx}$ , так как  $h^2f'' = h^2f_{xx} + O(h^4)$ .

Повышение точности схемы путем уменьшения  $h$  ограничивается также требованием экономичности, т. е. экономии времени получения решения с заданной точностью. Поэтому на практике часто применяется расчет по одной и той же схеме на последовательности сеток, позволяющий повысить точность без существенного увеличения времени счета (метод Рунге), в предположении достаточной гладкости решения.

Предположим, что для решения разностной задачи на любой равномерной сетке справедливо асимптотическое разложение

$$y_i^h = u_i + \alpha(x_i)h^{k_1} + O(h^{k_2}), \quad k_2 > k_1 > 0, \quad (11)$$

где  $\alpha(x_i)$  не зависит от  $h$ . Требуется найти сеточную функцию  $\tilde{y}_i$ , для которой

$$\tilde{y}_i = u_i + O(h^{-2}) \quad (12)$$

на некотором множестве узлов  $\tilde{\omega}_h$ .

Рассмотрим две сетки  $\omega_{h_1}$  и  $\omega_{h_2}$  с шагами  $h_1$  и  $h_2$ , имеющие общие узлы; множество общих узлов обозначим  $\tilde{\omega}_h$ . Пусть  $y_i^{h_1}$  и  $y_i^{h_2}$  — решения разностной задачи на сетках  $\omega_{h_1}$  и  $\omega_{h_2}$  соответственно. Образуем их линейную комбинацию  $\tilde{y}_i = \sigma y_i^{h_1} + (1 - \sigma) y_i^{h_2}$  и подставим сюда разложение (11):

$$\tilde{y}_i = u_i + \alpha(x_i) (\sigma h_1^{k_1} + (1 - \sigma) h_2^{k_1}) + O(h^{k_2}).$$

Приравнивая нуль коэффициент при  $\alpha(x_i)$ , найдем

$$\sigma = h_2^{k_1} / (h_2^{k_1} - h_1^{k_1}); \quad (13)$$

при этом в узлах  $x_i \in \tilde{\omega}_h$  выполняется требование (12).

Таким образом, для повышения точности сеточного решения на некотором множестве узлов  $\tilde{\omega}_h$  надо решить задачу дважды на сетках  $\omega_{h_1}$  и  $\omega_{h_2}$ , пересекающихся по этому множеству, и составить их линейную комбинацию с коэффициентами  $\sigma$  и  $(1 - \sigma)$ , где  $\sigma$  определяется согласно (13).

В частности, можно взять  $h_2 = h_1/2$ ,  $h_1 = h$ ; тогда  $\tilde{\omega}_h = \omega_{h_1}$ . Для схемы второго порядка точности имеем  $k_1 = 2$ ,  $k_2 = 4$  и  $\sigma = -1/3$ ,  $1 - \sigma = 4/3$ .

Возможность получения разложения

$$z_i = y_i - u_i = \alpha(x_i)h^2 + O(h^4)$$

следует из разложения невязки  $\psi_i = \beta(x_i)h^2 + O(h^4)$ , которая является правой частью задачи

$$\Lambda z = -\psi, \quad z_0 = z_N = 0.$$

Использование неравномерных сеток открывает большие возможности эмпирического повышения точности без увеличения числа узлов, если имеется предварительная информация о поведении решения исходной задачи. Так, в области сильного изменения коэффициентов и правой части уравнения естественно сгустить сетку. Вблизи границы разрыва коэффициентов обычно сетку сгущают по закону геометрической прогрессии. Чтобы получить предварительную информацию, можно провести сначала расчет на грубой сетке и после этого окончательный расчет — на специальной сетке.

## § 5. Методы построения разностных схем

Из предыдущего ясно, что разностные схемы для конкретного дифференциального уравнения должны правильно отражать в пространстве сеточных функций основные свойства исходной задачи (такие как самосопряженность, знакопределенность и т. д.). Для рассмотренной нами выше краевой задачи важным требованием оказалось свойство консервативности, эквивалентное свойству самосопряженности разностного оператора. Важной задачей является получение разностных схем с заданным качеством. Для построения таких схем в настоящее время используется ряд методов, о которых рассказывается в этом параграфе.

**1. Интегро-интерполяционный метод.** Обычно дифференциальное уравнение выражает некоторый физический закон сохранения. Этот закон можно написать в интегральной форме для интервала (ячейки) сетки (уравнение баланса). Дифференциальное уравнение получается из уравнения баланса при стремлении шага сетки к нулю в предположении существования непрерывных производных, входящих в уравнение. Входящие в уравнение баланса на сетке производные и интегралы следует заменить приближенными выражениями на сетке. В результате получим однородную схему. Такой метод называется *интегро-интерполяционным методом* или *методом баланса*. Проиллюстрируем его на примере задачи

$$(ku')' - qu = -f(x), \quad 0 < x < 1, \quad (ku') - \sigma_1 u = -\mu_1 \\ \text{при } x = 0, \quad u(1) = \mu_2. \quad (1)$$

Напишем уравнение баланса тепла на отрезке  $0 \leq x \leq 1$ :

$$w_{i+1/2} - w_{i-1/2} + \int_{x_{i-1/2}}^{x_{i+1/2}} f(x) dx = \int_{x_{i-1/2}}^{x_{i+1/2}} q(x) u(x) dx, \\ w = ku', \quad (2)$$

где  $(-w(x))$  — поток тепла,  $q(x)u(x)$  — мощность стоков (при  $q < 0$  — источников) тепла, пропорциональная температуре,  $f(x)$  — плотность распределения внешних источников (стоков) тепла. В левой части этого уравнения записано количество тепла, остающегося за счет тепловых потоков на отрезке  $[x_{i-1/2}, x_{i+1/2}]$  и за счет внешних источников, в правой части — количество тепла, отдаваемое

внешней среде за счет теплообмена на боковой поверхности.

Чтобы получить из (2) трехточечное разностное уравнение, заменим  $w_{i-1/2}$ ,  $w_{i+1/2}$  и интегралы в уравнении (2) линейной комбинацией значений подынтегральных функций в узлах сетки  $(x_{i-1}, x_i, x_{i+1})$ , например,

$$\frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} q(x) u(x) dx \approx d_i u_i, \quad d_i = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} q(x) dx.$$

Проинтегрируем равенство  $u' = w/k$  по  $x$  от  $x_{i-1}$  до  $x_i$ :

$$u_i - u_{i-1} = \int_{x_{i-1}}^{x_i} \frac{w}{k(x)} dx \approx h w_{i-1/2} \frac{1}{a_i},$$

$$a_i = \left[ \frac{1}{h} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} \right]^{-1}.$$

В результате получаем из (2) схему

$$\frac{1}{h} \left[ a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right] - d_i y_i = -\varphi_i,$$

$$\varphi_i = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} f(x) dx.$$

При выводе мы фактически предполагали лишь, что  $u = \text{const}$  при  $x_{i-1/2} \leq x \leq x_{i+1/2}$ ,  $w = \text{const}$  при  $x_{i-1} \leq x \leq x_i$ .

Вместо написанных здесь выражений для  $a_i$ ,  $d_i$ ,  $\varphi_i$  естественно взять более простые формулы, как это и делалось в предыдущих параграфах. Напишем разностную аппроксимацию для краевого условия третьего рода при  $x=0$ . Для этого воспользуемся уравнением баланса при  $0 \leq x \leq x_{1/2} = h/2$

$$w_{1/2} - w_0 - \int_0^{x_{1/2}} qu dx = - \int_0^{x_{1/2}} f(x) dx.$$

Подставляя сюда

$$w_{1/2} = a_1 u_{x,1}, \quad w_0 = (ku')_0 = \sigma_1 u_0 - \mu_1,$$

$$\int_0^{x_{1/2}} qu dx \sim q_0 u_0 \frac{1}{2} h, \quad \int_0^{x_{1/2}} f(x) dx \sim f_0 \frac{1}{2} h$$

и заменяя всюду  $u$  на  $y$ , получим разностное краевое условие

$$a_1 y_{\bar{x},1} - \sigma_1 y_0 + \mu_1 - h q_0 y_0 / 2 = -h f_0 / 2,$$

которое можно записать в виде

$$a_1 y_{\bar{x},1} = \bar{\sigma}_1 y_0 - \bar{\mu}_1, \text{ где } \bar{\sigma}_1 = \sigma_1 + h q_0 / 2, \quad \bar{\mu}_1 = \mu_1 + h f_0 / 2. \quad (3)$$

Оценим на решении  $u = u(x)$  уравнения (1) величину невязки

$$v = a_1 u_{\bar{x},1} - \bar{\sigma}_1 u_0 + \bar{\mu}_1.$$

Подставив сюда  $a_1 = k_{1/2} + O(h^2) = k_0 + \frac{1}{2} h k'_0 + O(h^2)$ ,  $u_1 = u_0 + h u'_0 + h^2 u''_0 / 2 + O(h^3)$ ,  $u_{\bar{x},1} = (u_1 - u_0) / h = u'_0 + h u''_0 / 2 + O(h^2)$ , получим

$$\begin{aligned} v &= (k u')_0 + \frac{1}{2} h (k u')'_0 - \bar{\sigma}_1 u_0 + \mu_1 + O(h^2) = [(k u')_0 - \\ &- \sigma_1 u_0 + \mu_1] + \frac{1}{2} h [(k u')' - q u + f]_0 + O(h^2) = O(h^2), \end{aligned}$$

т. е. разностное краевое условие третьего рода (3) аппроксимирует условие  $k u' = \sigma_1 u - \mu_1$  при  $x = 0$  с погрешностью второго порядка  $v = O(h^2)$ .

Для практического использования краевое условие (3) надо записать в виде

$$y_0 = \chi_1 y_1 - \tilde{\mu}_1, \quad \chi_1 = \frac{a_1}{a_1 + h \bar{\sigma}_1}, \quad \tilde{\mu}_1 = \frac{h \bar{\mu}_1}{a_1 + h \bar{\sigma}_1}.$$

Для повышения точности схемы следует использовать при вычислении интегралов интерполяцию более высокого порядка.

**2. Метод аппроксимации квадратичного функционала.**  
Краевая задача

$$Lu = (k u')' - q u = -f(x), \quad 0 < x < 1, \quad u(0) = 0, \quad u(1) = 0$$

эквивалентна задаче об отыскании минимизирующего элемента квадратичного функционала

$$J[u] = \int_0^1 (k (u')^2 + q u^2) dx - 2 \int_0^1 f u dx.$$

Введем на отрезке  $0 \leq x \leq 1$  сетку  $\bar{\omega}_h = \{x_i = ih, i = 0, 1, \dots, N\}$  и аппроксимируем функционал. Для этого сна-

чала представим его в виде суммы интегралов по интервалам сетки:

$$J[u] = \sum_{i=1}^N J_i[u], \quad J_i[u] = \int_{x_{i-1}}^{x_i} (k(u')^2 + qu^2 - 2fu) dx,$$

после чего аппроксимируем  $J_i$ , например, так:

$$\int_{x_{i-1}}^{x_i} k(u')^2 dx \approx a_i (u_{\bar{x},i})^2 h,$$

$$\int_{x_{i-1}}^{x_i} (qu^2 - 2fu) dx \approx \frac{h}{2} [(qu^2 - 2fu)_i + (qu^2 - 2fu)_{i-1}],$$

где  $a_i$  — некоторый коэффициент, например

$$a_i = \frac{1}{h} \int_{x_{i-1}}^{x_i} k(x) dx.$$

В результате получим функционал

$$J_h[y] = \sum_{k=1}^N h a_k (y_{\bar{x},k})^2 + \sum_{k=1}^{N-1} (q_k y_k^2 - 2f_k y_k) h,$$

где  $y_i = y(i)$  — произвольная сеточная функция, обращающаяся в нуль при  $i = 0, N$ .

Уравнение

$$Ay = \varphi \text{ или } \sum_{j=1}^N a_{ij} y_j = \varphi_i, \quad A = A^* > 0,$$

имеет решение, минимизирующее функционал

$$I_A[y] = (Ay, y) - 2(\varphi, y) = \sum_{i,j=1}^N a_{ij} y_j y_i - 2 \sum_{i=1}^N \varphi_i y_i.$$

В этом можно убедиться, приравняв нулю производную

$$\frac{\partial I_A[y]}{\partial y_{i_0}} = 2 \sum_{j=1}^N a_{i_0 j} y_j - 2 \varphi_{i_0} = 0, \quad \frac{\partial^2 I_A}{\partial y_{i_0}^2} > 0,$$

так как  $a_{ii} > 0$  для всех  $i = 1, 2, \dots, N$  в силу положительности  $A > 0$ .

Вычисляя производные

$$\begin{aligned}\frac{\partial J_h}{\partial y_i} &= 2a_i y_{x,i} - 2a_{i+1} y_{x,i+1} + (2q_i y_i - 2f_i) h, \\ \frac{\partial^2 J_h}{\partial y_i^2} &= \frac{2a_i}{h} + \frac{2a_{i+1}}{h} + 2q_i > 0,\end{aligned}$$

убеждаемся, что элемент  $y = y(x) \in H_h$ , минимизирующий квадратичный функционал, является решением задачи  $(ay_x)_x - q_i y_i = -f_i$ ,  $i = 1, 2, \dots, N-1$ ,  $y_0 = y_N = 0$ .

**3. Метод аппроксимации интегрального тождества (метод сумматорных тождеств).** Пусть

$$(ku')' - qu + f(x) = 0, \quad 0 < x < 1, \quad u(0) = u(1) = 0. \quad (4)$$

Умножая уравнение (4) на произвольную дифференцируемую функцию  $v(x)$ , обращающуюся в нуль при  $x = 0$ ,  $x = 1$  и интегрируя по  $x$  от 0 до 1, получаем тождество

$$I(u, v) = \int_0^1 (ku'v' + quv - fv) dx = 0.$$

Заменяя по аналогии с п. 2 интеграл и производные  $u'$ ,  $v'$ , напишем сумматорное тождество

$$I_h[y, v] = \sum_{i=1}^N a_i y_{x,i} v_{x,i} h + \sum_{i=1}^{N-1} (q_i y_i - f_i) v_i h = 0.$$

Полагая затем, например,  $v_i = \delta_{i,i_0}$ ,  $0 < i_0 < N$ , и учитывая, что  $v_{x,i} = 0$  при  $i < i_0$  и  $i > i_0 + 1$ ,  $v_{x,i_0+1} = -1/h$ ,  $v_{x,i_0} = 1/h$ , получим

$$h \left( \frac{1}{h} a_{i+1} y_{x,i} - \frac{1}{h} a_i y_{x,i} \right) + (q_i - f_i) h = 0 \text{ при } i = i_0,$$

т. е.  $(ay_x)_x - qy = -f$ .

**4. Методы Ритца и Бубнова — Галеркина (вариационно-разностные методы).** Задача о минимуме функционала

$$I[u] = (Au, u) - 2(u, f),$$

где  $A$  — самосопряженный и положительно определенный линейный оператор в гильбертовом пространстве  $H$  со скалярным произведением  $(x, y)$ , эквивалентна задаче о решении уравнения

$$Au = f.$$

Вводится последовательность конечномерных пространств  $V_n$  с базисом  $\{\varphi_i^{(n)}\}$ ,  $i = 1, 2, \dots, n$ .

*Метод Ритца* заключается в том, что ищется элемент  $u_n \in V_n$ , минимизирующий функционал  $I[u]$  в  $V_n$ . Приближенное решение  $u_n$  ищется в виде суммы

$$u_n = \sum_{j=1}^n y_j \varphi_{j1} \quad (5)$$

где  $y_1, \dots, y_n$  — неизвестные коэффициенты. Вычисления дают

$$I[u_n] = \sum_{i,j=1}^n \alpha_{ij} y_i y_j - 2 \sum_{i=1}^n \beta_i y_i,$$

$$\alpha_{ij} = \alpha_{ji} = (A\varphi_i, \varphi_j), \quad \beta_i = (f, \varphi_i);$$

$I[u_n] = \Phi(y_1, y_2, \dots, y_n)$  есть функция  $n$  коэффициентов  $y_i$ . Приравнивая нулю производные  $\partial I[u_n]/\partial y_i$ , получим систему  $n$  уравнений

$$\sum_{j=1}^n \alpha_{ij} y_j - \beta_i = 0, \quad i = 1, 2, \dots, n,$$

для определения  $y_1, y_2, \dots, y_n$ .

Проиллюстрируем метод Ритца на примере задачи (4). В качестве функции  $\varphi_i(x)$  возьмем функцию:

$$\varphi_i(x) = \eta\left(\frac{x-x_i}{h}\right) = \eta_i(x), \quad \eta(s) = \begin{cases} 0, & s < -1, \quad s > 1, \\ 1+s, & -1 < s < 0, \\ 1-s, & 0 < s < 1. \end{cases}$$

Подставляя в формулу для  $\alpha_{ij} A\varphi_i = -(k\varphi_i)' + q\varphi_i$ , имеем

$$\alpha_{ij} = (A\varphi_i, \varphi_j) = \int_0^1 \left( k \frac{d\eta_i}{dx} \frac{d\eta_j}{dx} + q\eta_i \eta_j \right) dx,$$

$$\beta_i = \int_0^1 f(x) \eta_i(x) dx. \quad (6)$$

Вычисления дают

$$\frac{d\eta_i}{dx} = 0 \text{ при } x < x_{i-1}, \quad x > x_{i+1},$$

$$\frac{d\eta_i}{dx} = \begin{cases} 1/h & \text{при } x_{i-1} < x < x_i, \\ -1/h & \text{при } x_i < x < x_{i+1}. \end{cases}$$

Отсюда и из (6) видно, что матрица  $[\alpha_{ij}]$  трехдиагональна, так как от нуля отличны лишь те  $\alpha_{ij}$ , для которых  $j = i - 1, i, i + 1$ . Поэтому для  $y_i$  получаем систему

$$\alpha_{i-1} y_{i-1} + \alpha_{i,i} y_i + \alpha_{i+1} y_{i+1} - \beta_i = 0.$$

Вводя обозначения

$$a_i = -h\alpha_{i-1} + h^2 d_i = h\alpha_{i,i} + h(\alpha_{i-1} + \alpha_{i+1}),$$

$$\beta_i = -h^2 \varphi_i$$

и замечая, что  $\alpha_{i+1,i} = \alpha_{i,i+1}$ , получаем схему

$$a_i y_{i-1} - (a_i + a_{i+1} + h^2 d_i) y_i + a_{i+1} y_{i+1} + h^2 \varphi_i = 0,$$

или

$$(ay_x)_x - dy + \varphi = 0, \quad (7)$$

где

$$a_i = \int_{-1}^0 k(x_i + sh) ds + h^2 \int_{-1}^0 q(x_i + sh) s(1+s) ds,$$

$$d_i = \int_{-1}^0 q(x_i + sh)(1+s) ds + \int_0^1 q(x_i + sh)(1-s) ds,$$

$$\varphi_i = \int_{-1}^0 f(x_i + sh)(1+s) ds + \int_0^1 f(x_i + sh)(1-s) ds.$$

Это — схема второго порядка аппроксимации.

В методе Бубнова — Галеркина решение  $u_n$  также ищется в виде (6), однако коэффициенты  $y_i$  находятся из условия ортогональности невязки  $Au_n - f$  к базисным функциям  $\varphi_i(x)$ :

$$(Au_n - f, \varphi_i) = 0, \quad i = 1, 2, \dots, n; \quad (8)$$

при этом самосопряженность оператора  $A$  не требуется. Для задачи (4) снова выбираем те же базисные функции. Подставляя (6) в (8), получим систему уравнений для  $y_i$ . Вычисляя  $\alpha_{ij}$  и  $\beta_i$ , приходим к той же самой схеме (7), которую мы получили методом Ритца.

При указанном выборе координатных функций  $\varphi_i(x) = \eta\left(\frac{x-x_i}{h}\right)$  методы Ритца и Бубнова — Галеркина совпадают с методом конечных элементов.

## Глава V

# ЗАДАЧА КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

В этой главе мы рассмотрим разностные схемы для решения обыкновенных дифференциальных уравнений (вообще говоря, нелинейных) первого порядка с начальными данными (задачи Коши). Это — классическая область применения численных методов. Имеется много разностных методов, часть из которых возникла в домашнюю эпоху и оказалась пригодной и для современных ЭВМ. Мы ограничимся кратким изложением основных разностных схем, которые широко используются на практике и для которых имеются соответствующие стандартные программы.

### § 1. Методы Рунге — Кутта

**1. Задача Коши для уравнения первого порядка.** Пусть требуется найти непрерывную при  $0 \leq t \leq T$  функцию  $u = u(t)$ , удовлетворяющую дифференциальному уравнению при  $t > 0$  и начальному условию при  $t = 0$ :

$$\frac{du}{dt} = f(t, u(t)), \quad 0 < t \leq T, \quad u(0) = u_0, \quad (1)$$

где  $f(t, u)$  — заданная непрерывная функция двух аргументов.

Если функция  $f(t, u)$  определена в прямоугольнике  $D = \{0 \leq t \leq T, |u - u_0| \leq U\}$  и удовлетворяет в области  $D$  по переменной  $u$  условию Липшица:

$$|f(t, u_1) - f(t, u_2)| \leq K|u_1 - u_2| \quad \text{для всех } (t, u_1), (t, u_2) \in D, \quad (2)$$

где  $K = \text{const} > 0$ , то задача (1) имеет единственное решение.

Для доказательства этого утверждения уравнение (1) интегрируется от 0 до  $t$ :

$$u(t) = u_0 + \int_0^t f(s, u(s)) ds, \quad (3)$$

а полученнное интегральное уравнение решается методом последовательных приближений (методом Пикара):

$$u_{n+1}(t) = u_0 + \int_0^t f(s, u_n(s))ds, \quad (4)$$

где  $n$  — номер приближения (итерации). Метод Пикара сходится и определяет единственное решение уравнения (3) или задачи Коши (1).

Этот метод позволяет найти приближенное решение задачи (1), если в (4) заменить интеграл какой-либо квадратурной формулой. Однако объем вычислений для полученного алгоритма велик, так как для каждой итерации (при фиксированном  $t$ ) необходимо вычислять интеграл.

Иногда для приближенного решения задачи (1) используется аналитический метод, основанный на идее разложения в ряд Тейлора решения задачи Коши (1). Приближенное решение  $u_n(t)$  ищем в виде

$$u_n(t) = \sum_{k=1}^n \frac{t^k}{k!} u^{(k)}(0) + u_0, \quad 0 \leq t \leq T, \quad (5)$$

где  $u^{(1)}(0) = \frac{du}{dt}(0) = f(0, u_0)$ , а значения производных  $u^{(k)}(0)$  ( $k \geq 2$ ) находятся последовательным дифференцированием уравнения (1)

$$\begin{aligned} u^{(2)}(0) &= u''(0) = \left. \frac{d}{dt} f(t, u) \right|_{t=0} = f_t(0, u_0) + f(0, u_0) f_u(0, u_0), \\ u^{(3)}(0) &= u'''(0) = \left. \frac{d^2}{dt^2} f(t, u) \right|_{t=0} = \\ &= f_{t^2}(0, u_0) + 2f_{ut}(0, u_0)f(0, u_0) + f_{u^2}(0, u_0)u''(0), \dots, \\ f_t &= \frac{\partial f}{\partial t}, \quad f_u = \frac{\partial f}{\partial u}, \quad f_{ut} = \frac{\partial^2 f}{\partial u \partial t} \text{ и т. д.} \end{aligned}$$

Для малых  $t$  метод рядов (5) может давать хорошее приближение к точному решению  $u(t)$  при не очень больших  $n$ . Здесь объем вычислений зависит не только от точности  $\varepsilon > 0$  ( $|u(t) - u_n(t)| < \varepsilon$ ) и от  $n = n(\varepsilon)$ , но и от вида функции  $f(t, u)$ , так как нахождение производных  $u^{(k)}(t)$  может оказаться очень трудоемким.

В дальнейшем мы будем предполагать всюду, что функция  $f(t, u)$  является достаточно гладкой, т. е. имеет столько производных (по  $t$  и по  $u$ ), сколько требуется по ходу изложения.

Прежде чем переходить к изложению разностных схем для задачи (1), остановимся на вопросе об устойчивости решения задачи (1). Как изменится решение задачи (1) при изменении начального условия? Пусть  $\tilde{u}(t)$  — решение уравнения (1) с начальным условием  $\tilde{u}(0) = \tilde{u}_0$ . Для погрешности  $z(t) = \tilde{u}(t) - u(t)$  получаем уравнение

$$\frac{dz}{dt} = \alpha(t) z, \quad 0 < t \leq T, \quad z(0) = z_0 = \tilde{u}_0 - u_0, \quad (6)$$

где  $\alpha(t) = [f(t, \tilde{u}) - f(t, u)]/z = f_u(t, u + \theta z)$ ,  $0 \leq \theta \leq 1$ .

Решением (6) является функция

$$z(t) = z(0) \exp \left\{ \int_0^t \alpha(s) ds \right\}.$$

Если  $f_u \leq 0$  для всех  $t, u$ , то

$$|z(t)| \leq |z(0)| \quad \text{или} \quad |\tilde{u}(t) - u(t)| \leq |\tilde{u}_0 - u_0|$$

для всех  $t \in [0, T]$ ,

т. е. решение задачи (1) *устойчиво* по начальным данным (погрешность в начальных данных не нарастает). Задача (1) устойчива также и по правой части:

$$|\tilde{u}(t) - u(t)| \leq |\tilde{u}_0 - u_0| + \varepsilon T \quad \text{при } 0 \leq t \leq T, \quad \text{если } f_u \leq 0,$$

где  $\tilde{u}(t)$  — решение задачи (1) с правой частью

$$\tilde{f} = f(t, \tilde{u}) + \delta f, \quad |\delta f| \leq \varepsilon, \quad \varepsilon = \text{const} > 0.$$

Решение задачи (6) при  $t \rightarrow \infty$  ведет себя аналогично решению линейного уравнения

$$\frac{dz}{dt} + \lambda z = 0, \quad 0 < t \leq T, \quad z(0) = z_0,$$

которое можно рассматривать как модельное уравнение при изучении устойчивости.

**2. Разностная схема Эйлера.** Введем на отрезке интегрирования  $0 \leq t \leq T$  сетку  $\omega_\tau = \{t_n = n\tau, n = 0, 1, \dots\}$ . Будем обозначать через  $y_n = y(t_n)$  сеточную функцию. Простейшим численным методом решения уравнения (1) является *разностная схема Эйлера*:

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n), \quad n = 0, 1, \dots, y_0 = u_0. \quad (7)$$

Значения  $y_n = y(t_n)$  находятся последовательно, начиная

с  $y_0 = u_0$ , по явной формуле

$$y_{n+1} = y_n + \tau f(t_n, y_n), \quad n = 0, 1, \dots, y_0 = u_0.$$

Вместо  $u = u(t)$  мы находим сеточную функцию  $y_n = y(t_n)$  — приближенное решение задачи (1).

Сеточная функция

$$z_n = y_n - u(t_n)$$

является *погрешностью* разностной схемы. Напишем уравнение для  $z_n$ . Для этого подставим  $y_n = z_n + u_n$  в (7) и учтем, что

$$\begin{aligned} y_{n+1} - y_n &= (z_{n+1} - z_n) + (u_{n+1} - u_n), \\ f(t_n, y_n) &= f(t_n, u_n) + [f(t_n, u_n + z_n) - f(t_n, u_n)] = \\ &= f(t_n, u_n) + \alpha_n z_n, \end{aligned}$$

где

$$\alpha_n = f_u(t_n, u_n + \theta z_n), \quad 0 \leq \theta \leq 1.$$

В результате получаем для  $z_n$  задачу

$$\frac{z_{n+1} - z_n}{\tau} = \alpha_n z_n + \psi_n, \quad n = 0, 1, \dots, z_0 = 0, \quad (8)$$

где  $\psi_n$  — *невязка* или *погрешность* аппроксимации схемы (7) на решении  $u = u(t)$  задачи (1), равная

$$\psi_n = f(t_n, u_n) - \frac{u_{n+1} - u_n}{\tau}. \quad (9)$$

Оценим  $\psi_n$  при  $\tau \rightarrow 0$ . Для этого подставим

$$u_{n+1} = u_n + \tau \dot{u}_n + \frac{\tau^2}{2} \ddot{u}_n + \dots \left( \dot{u} = \frac{du}{dt} \right)$$

в (9) и, учитывая, что согласно (1)  $\dot{u}_n = f(t_n, u_n)$ , получим:  $\psi_n = O(\tau)$  или  $\|\psi\|_C = \max_{0 \leq t_n \leq T} |\psi_n| = O(\tau)$ . Это означает,

что схема Эйлера имеет *первый порядок аппроксимации*.

Покажем, что схема Эйлера сходится, т. е.  $\|z_n\|_C = \|y_n - u_n\|_C \rightarrow 0$  при  $\tau \rightarrow 0$ , и имеет *первый порядок точности*, т. е.

$$\|z\|_C = \max_{0 \leq t_n \leq T} |z_n| = O(\tau).$$

Доказательство проведем в предположении, что

$$-K \leq f_u(t, u) \leq K, \quad \tau \leq 2/K. \quad (10)$$

Из (8) определим

$$\begin{aligned} z_{n+1} &= (1 + \tau\alpha_n)z_n + \tau\psi_n, \\ |z_{n+1}| &\leq |1 + \tau\alpha_n||z_n| + \tau|\psi_n| \leq |z_n| + \tau|\psi_n|, \end{aligned}$$

так как  $|1 + \tau\alpha_n| \leq 1$  согласно (10). Отсюда следует, что

$$|z_{n+1}| \leq |z_0| + \sum_{s=0}^n \tau |\psi_s| = \sum_{s=0}^n \tau |\psi_s|, \quad (11)$$

т. е.  $\|z\|_c = O(\tau)$ .

Если условие (10) не выполнено, но  $|f_u| \leq K$ , то вместо (11) получим  $|z_{n+1}| \leq T e^{KT} \|\psi\|_c$ , и утверждение  $\|z\|_c = O(\tau)$  остается в силе.

**3. Повышение порядка точности.** Метод Эйлера весьма прост, однако дает низкую точность. Порядок точности численного решения по  $\tau$  можно повысить, не усложняя алгоритма. Идея метода Рунге повышения точности состоит в следующем. Предположим, что решение  $u = u(t)$  является достаточно гладким и имеет место следующее разложение погрешности  $z_n = y_n - u_n$  по степеням  $\tau$ :

$$y_n = u_n + \alpha(t)\tau + \beta(t)\tau^2 + \dots, \quad (12)$$

где  $\alpha(t)$  и  $\beta(t)$  — функции, не зависящие от  $\tau$ .

Выберем две сетки с шагами  $\tau_1$  и  $\tau_2$ , имеющие общие узлы (например,  $\tau_1 = \tau$ ,  $\tau_2 = \tau/2$ ), решим на каждой сетке задачу (7) и найдем  $y^{(1)}(t_{n_1})$  и  $y^{(2)}(t_{n_2})$  соответственно. Возьмем общий для двух сеток узел  $t_{n*} = t_{n_1} = t_{n_2}$  и напишем (12) при  $n = n^*$ :

$$\begin{aligned} y^{(1)}(t_{n*}) &= u(t_{n*}) + \alpha(t_{n*})\tau_1 + O(\tau_1^2), \\ y^{(2)}(t_{n*}) &= u(t_{n*}) + \alpha(t_{n*})\tau_2 + O(\tau_2^2). \end{aligned}$$

Образуем линейную комбинацию с параметром  $\sigma$ :

$$\begin{aligned} \tilde{y}(t_{n*}) &= \sigma y^{(1)}(t_{n*}) + (1 - \sigma) y^{(2)}(t_{n*}) = \\ &= u(t_{n*}) + [\sigma\tau_1 + (1 - \sigma)\tau_2]\alpha(t_{n*}) + O(\tau_1^2 + \tau_2^2). \end{aligned}$$

Выбирая  $\sigma$  из условия  $\sigma\tau_1 + (1 - \sigma)\tau_2 = 0$ , т. е. полагая  $\sigma = \tau_2/(\tau_2 - \tau_1)$ , получаем

$$\tilde{y}(t_{n*}) = u(t_{n*}) + O(\tau^2), \quad \tau = \max(\tau_1, \tau_2).$$

Сеточная функция  $\tilde{y}$  приближает решение  $u = u(t)$  со вторым порядком точности по  $\tau$ . Таким образом, мы по-

высили точность метода Эйлера, проводя два расчета на сетках с шагами  $\tau_1$  и  $\tau_2$ . Эту процедуру можно продолжить, имея в виду (12). Проводя расчеты по схеме (7) на трех сетках с шагами  $\tau_1$ ,  $\tau_2$ ,  $\tau_3$ , мы определим решение задачи (1) с третьим порядком точности в узлах, являющихся общими для этих трех сеток.

**4. Схемы Рунге — Кутта.** Порядок точности можно повысить путем усложнения разностной схемы. Весьма распространены на практике схемы Рунге — Кутта второго и четвертого порядков точности.

Вычисления по схеме Рунге — Кутта второго порядка точности проводятся в два этапа. На первом этапе находится промежуточное значение  $\bar{y}_n$  по схеме Эйлера с шагом  $\alpha\tau$ :

$$\bar{y}_n = y_n + \alpha\tau f(t_n, y_n);$$

на втором этапе находится значение  $y_{n+1}$  по формуле

$$y_{n+1} = y_n + \tau(1 - \sigma)f(t_n, y_n) + \sigma\tau f(t_n + \alpha\tau, \bar{y}_n),$$

где  $\alpha > 0$ ,  $\sigma > 0$  — параметры. Исключая  $\bar{y}_n$ , получим для  $y_{n+1}$  схему

$$\begin{aligned} \frac{y_{n+1} - y_n}{\tau} &= (1 - \sigma)f(t_n, y_n) + \\ &+ \sigma f(t_n + \alpha\tau, y_n + \alpha\tau f(t_n, y_n)). \end{aligned} \quad (13)$$

Порядок точности схемы зависит от параметров  $\alpha$ ,  $\tau$ .

Найдем выражение для невязки, или погрешности аппроксимации схемы (13). Для этого, по аналогии с п. 2, перенесем  $(y_{n+1} - y_n)/\tau$  в правую часть и подставим  $u_n$ ,  $u_{n+1}$  вместо  $y_n$ ,  $y_{n+1}$ . В результате получим для невязки выражение

$$\begin{aligned} \psi_n &= (1 - \sigma)f(t_n, u_n) + \sigma f(t_n + \alpha\tau, u_n + \alpha\tau f(t_n, u_n)) - \\ &- (u_{n+1} - u_n)/\tau. \end{aligned} \quad (13')$$

Воспользовавшись разложениями по формуле Тейлора, получим

$$\psi_n = \tau(\sigma\alpha - 1/2)u_n'' + O(\tau^2).$$

Отсюда видно, что схема (13) имеет второй порядок аппроксимации  $\psi_n = O(\tau^2)$  при выполнении условия

$$\sigma\alpha = 1/2. \quad (14)$$

Таким образом, существует однопараметрическое семейство схем (13), (14) второго порядка аппроксимации.

Рассмотрим частные случаи.

1)  $\sigma = 1, \alpha = 1/2$ :

$$\frac{\bar{y}_n - y_n}{\tau/2} = f(t_n, y_n), \quad \frac{y_{n+1} - y_n}{\tau} = f\left(t_n + \frac{\tau}{2}, \bar{y}_n\right). \quad (15)$$

Это известная схема *предиктор — корректор*, или *счет — пересчет*. Ее можно переписать иначе:

$$\bar{y}_n = y_n + \frac{\tau}{2} f(t_n, y_n), \quad y_{n+1} = y_n + \tau f\left(t_n + \frac{\tau}{2}, \bar{y}_n\right),$$

или, после исключения  $\bar{y}_n$ , в виде

$$(y_{n+1} - y_n)/\tau = f\left[t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2} f(t_n, y_n)\right]. \quad (15')$$

2)  $\sigma = 1/2, \alpha = 1$ :

$$\frac{y_{n+1} - y_n}{\tau} = \frac{1}{2} [f(t_n, y_n) + f(t_{n+1}, y_n + \tau f(t_n, y_n))]. \quad (16)$$

Эту схему также можно трактовать как схему предиктор — корректор: сначала — схема Эйлера с шагом  $\tau$  (*предиктор*):

$$\bar{y}_n = y_n + \tau f(t_n, y_n);$$

затем — схема с полусуммой (*корректор*):

$$(y_{n+1} - y_n)/\tau = \frac{1}{2}[f(t_n, y_n) + f(t_{n+1}, \bar{y}_n)].$$

Идея метода предиктор — корректор часто используется при написании разностных схем для уравнений математической физики с частными производными.

Приведем формулы для схемы Рунге — Кутта 4-го порядка точности:

$$\frac{y_{n+1} - y_n}{\tau} = \frac{1}{6} [k_1(y_n) + 2k_2(y_n) + 2k_3(y_n) + k_4(y_n)],$$

$$n = 0, 1, \dots, y_0 = u_0, \quad (17)$$

где  $k_1, k_2, k_3, k_4$  — поправки, вычисляемые по формулам

$$k_1 = f(t_n, y_n), \quad k_2 = f(t_n + \tau/2, y_n + \tau k_1/2),$$

$$(18)$$

$$k_3 = f(t_n + \tau/2, y_n + \tau k_2/2), \quad k_4 = f(t_n + \tau, y_n + \tau k_3).$$

При определении  $y_{n+1}$  по заданному  $y_n$  надо четыре раза вычислять правую часть.

Поясним, как вести счет по этой схеме. При  $n = 0$  известно  $y_0 = u_0$ . Можно вычислить последовательно  $k_1, k_2, k_3, k_4$  и найти

$$y_1 = y_0 + \frac{1}{6} \tau (k_1(y_0) + 2k_2(y_0) + 2k_3(y_0) + k_4(y_0)),$$

после чего вычисления повторяются при  $n = 1, 2, \dots$ . Для невязки получаем выражение

$$\begin{aligned} \psi_n = \frac{1}{6} [k_1(u_n) + 2k_2(u_n) + 2k_3(u_n) + k_4(u_n)] - \\ - \frac{u_{n+1} - u_n}{\tau}, \quad (19) \end{aligned}$$

где  $k_i(u_n)$  ( $i = 1, 2, 3, 4$ ) определяются по формулам (18), в которых вместо  $y_n$  подставлено  $u_n$ .

Проводя разложение  $u_{n+1}, k_2(u_n), k_3(u_n), k_4(u_n)$  в окрестности  $t = t_n$ , убеждаемся в том, что  $\psi_n = O(\tau^4)$ , т. е. схема (7), (18) имеет четвертый порядок аппроксимации, если  $u = u(t)$  имеет четыре непрерывных производных.

Все методы Рунге — Кутта являются явными (для определения  $y_{n+1}$  надо провести вычисления по явным формулам) и одношаговыми (для определения  $y_{n+1}$  надо сделать один шаг на сетке от  $t_n$  к  $t_{n+1}$ ).

**5. Устойчивость разностных схем.** В п. 1 мы рассмотрели важное свойство дифференциального уравнения (1) — устойчивость (по начальным данным и по правой части). Для изучения устойчивости по начальным данным нелинейного уравнения (1) будем рассматривать модельное уравнение

$$\frac{du}{dt} + \lambda u = 0, \quad \lambda = \text{const} > 0, \quad t > 0, \quad u(0) = u_0. \quad (20)$$

Его решение  $u(t) = u_0 e^{-\lambda t}$  убывает при  $\lambda > 0$  и

$$|u(t)| \leq |u_0| \quad \text{при } \lambda \geq 0 \quad \text{для всех } t \geq 0, \quad (21)$$

т. е. уравнение (20) *устойчиво* при  $\lambda \geq 0$ , что соответствует условию  $f_u \leq 0$ .

Вводится естественное требование: для разностных схем, аппроксимирующих модельные уравнения, должен выполняться аналог неравенства (21):

$$|y_n| \leq |y_0| \quad \text{для всех } n = 1, 2, \dots \quad (22)$$

Мы увидим ниже, что это не всегда выполняется.

Рассмотрим ряд примеров.

1) Явная схема Эйлера:

$$\frac{y_{n+1} - y_n}{\tau} + \lambda y_n = 0, \quad y_{n+1} = (1 - \tau\lambda) y_n. \quad (23)$$

Отсюда видно, что условие

$$|y_{n+1}| \leq |y_n| \leq \dots \leq |y_0| \quad (24)$$

выполнено при  $|1 - \tau\lambda| \leq 1$  или  $-1 \leq 1 - \tau\lambda \leq 1$ , т. е. при

$$\tau\lambda \leq 2. \quad (25)$$

Если, например,  $\tau\lambda \geq 3$ , то

$$\begin{aligned} |y_{n+1}| &= |\tau\lambda - 1||y_n| \geq 2|y_n| \geq \dots \geq 2^{n+1}|y_0|, \\ |y_n| &\geq 2^n|y_0| \rightarrow \infty \text{ при } n \rightarrow \infty. \end{aligned}$$

Схема неустойчива, условие (24) не выполнено. Таким образом, схема Эйлера (23) *условно устойчива* при  $\tau \leq \leq 2/\lambda$ ,  $\lambda > 0$ .

2) Неявная схема Эйлера:

$$\frac{y_{n+1} - y_n}{\tau} + \lambda y_{n+1} = 0, \quad y_{n+1} = \frac{1}{1 + \tau\lambda} y_n. \quad (26)$$

Так как  $1/(1 + \tau\lambda) \leq 1$  при любых  $\tau\lambda \geq 0$ , то схема *безусловно устойчива*:

$$|y_n| \leq |y_0| \text{ при любых } \tau \text{ и } \lambda \geq 0, \quad n = 0, 1, 2, \dots \quad (27)$$

3) Схема с весами:

$$\frac{y_{n+1} - y_n}{\tau} + \lambda(\sigma y_{n+1} + (1 - \sigma)y_n) = 0, \quad y_{n+1} = qy_n. \quad (28)$$

Схема устойчива при

$$|q| \leq 1, \quad q = \frac{1 - (1 - \sigma)\tau\lambda}{1 + \sigma\tau\lambda}.$$

Видим, что  $|q| \leq 1$ , если  $-1 - \sigma\tau\lambda \leq 1 - (1 - \sigma)\tau\lambda \leq 1 + \sigma\tau\lambda$  или  $1 + \tau(\sigma - 1/2)\lambda \geq 0$ , так что  $1 + \sigma\tau\lambda \geq \tau\lambda/2 > 0$ . Таким образом, схема с весами *безусловно* (при любых  $\tau$ ) *устойчива* при  $\sigma > 1/2$  и *условно устойчива* при  $\sigma < 1/2$ , если  $\tau \leq 1/((1/2 - \sigma)\lambda)$ .

4) Схема Рунге – Кутта второго порядка. Подставляя в формулу (13)  $f = -\lambda y$ , получаем

$$y_{n+1} = qy_n, \quad q = 1 - \tau\lambda + \frac{1}{2}\tau^2\lambda^2. \quad (29)$$

Схема устойчива,  $|y_n| \leq |y_0|$ , если  $|q| = 1 - \tau\lambda + \frac{1}{2}\tau^2\lambda^2 \leq 1$ , что имеет место при

$$\tau\lambda \leq 2. \quad (25)$$

Схема Рунге — Кутта второго порядка устойчива при том же условии, что и явная схема Эйлера.

5) Схема Рунге — Кутта четвертого порядка. Подставляя  $f = -\lambda u$  в (17), (18), получаем

$$\begin{aligned} y_{n+1} &= qy_n, \\ q &= 1 - \tau\lambda + \frac{1}{2}\tau^2\lambda^2 - \frac{1}{6}\tau^3\lambda^3 + \frac{1}{24}\tau^4\lambda^4. \end{aligned} \quad (30)$$

Неравенство  $|q| \leq 1$  выполнено при  $\tau\lambda \leq 2,78$ , т. е. условие устойчивости схемы четвертого порядка немного слабее условия (25) для схемы второго порядка.

Эти примеры показывают, что явные одшаговые схемы условно устойчивы, а среди неявных схем имеются безусловно (абсолютно) устойчивые (например (28) при  $\sigma \geq 1/2$ ). Если  $\lambda > 0$  велико, то шаг  $\tau$ , в силу (25), для явных схем надо выбирать достаточно малым.

**6. О сходимости и точности.** Схема Рунге — Кутта для неоднородного уравнения

$$\frac{du}{dt} + \lambda u = f(t), \quad t > 0, \quad u(0) = u_0 \quad (31)$$

имеет вид

$$y_{n+1} = qy_n + \tau\varphi_n, \quad q = q(\tau\lambda), \quad (32)$$

где выражения для  $q$  и  $\varphi_n$  зависят от порядка схемы. Так, для схемы второго порядка имеем

$$\begin{aligned} q &= 1 - \tau\lambda + \frac{1}{2}\tau^2\lambda^2, \\ \varphi_n &= (1 - \sigma)f(t_n) + \sigma f(t_n + \alpha\tau), \quad \alpha\sigma = \frac{1}{2}. \end{aligned}$$

Для погрешности  $z_n = y_n - u_n$  получаем

$$\frac{z_{n+1} - z_n}{\tau} + \left( \lambda - \frac{\lambda^2\tau}{2} \right) z_n = \psi_n$$

или

$$z_{n+1} = qz_n + \tau\psi_n, \quad n = 0, 1, 2, \dots, z_0 = 0,$$

где  $\psi_n$  — невязка, равная

$$\psi_n = \varphi_n - (u_{n+1} - u_n)/\tau = O(\tau^2).$$

В силу условия устойчивости (25)  $|q| \leq 1$  и

$$|z_{n+1}| \leq |z_n| + \tau |\psi_n| \leq \sum_{k=0}^n \tau |\psi_k|, \quad (33)$$

откуда и следует, что схема (32) сходится и имеет второй порядок точности (сходится со скоростью  $O(\tau^2)$ , или сходится со вторым порядком):

$$\|z\|_c = O(\tau^2).$$

Таким образом, если схема устойчива и аппроксимирует уравнение (1), то она сходится. Это утверждение, доказанное для модельной задачи, имеет общее значение и справедливо для любой из схем второго порядка.

Аналогично доказывается сходимость со скоростью  $O(\tau^2)$  схемы Рунге — Кутта (13) при условии  $f_u \leq 0$ . В этом случае для  $z_n = y_n - u_n$  при  $\sigma\alpha = 1/2$  получаем задачу

$$\frac{z_{n+1} - z_n}{\tau} = \beta_n \left( 1 + \frac{1}{2} \tau \gamma_n \right) z_n + \tau \psi_n, \quad (34)$$

где  $\beta_n = f_u(t_n, u_n + \theta_1 z_n)$ ,  $\gamma_n = f_u(t_n + \tau/2, u_n + \theta_2 z_n)$  ( $0 \leq \theta_i \leq 1$ ,  $i = 1, 2$ ), а  $\psi_n$  определяется по формуле (13'). Перепишем (34) в виде

$$z_n = q_n z_n + \tau \psi_n, \quad q_n = 1 + \tau \beta_n (1 + \tau \gamma_n / 2).$$

Условие устойчивости  $|q_n| \leq 1$ , или  $-1 \leq q_n \leq 1$ , будет выполнено, если  $2 - \tau |\beta_n| + 1/2 \tau^2 |\beta_n| |\gamma_n| \geq 0$ ,  $1/2 \tau |\beta_n| |\gamma_n| \leq |\beta_n|$ , или  $\tau |\gamma_n| \leq 2$ . Первое неравенство выполнено также при  $\tau |\beta_n| \leq 2$ , и, следовательно, достаточно, чтобы

$$\tau K \leq 2, \quad (35)$$

если  $f_u \leq 0$ ,  $|f_u| \leq K$ ,  $(t, u) \in D$ . Условие (35) аналогично (25) и обеспечивает выполнение оценки (33), из которой и следует сходимость схемы (13) со вторым порядком,  $\|z\|_c = O(\tau^2)$ .

## § 2. Многошаговые схемы. Методы Адамса

**1. Многошаговые схемы.** В § 1 мы рассматривали методы Рунге — Кутта для численного решения задачи Коши

$$\frac{du}{dt} = f(t, u), \quad 0 < t \leq T, \quad u(0) = u_0. \quad (1)$$

Эти методы являются *одношаговыми методами*: при определении нового значения  $y_{n+1}$  используется лишь значение  $y_n$ . В общем случае для определения приближенного решения  $y_n$  можно рассмотреть *m-шаговые разностные схемы* ( $m \geq 1$ ), т. е. уравнения вида

$$\sum_{k=0}^m \frac{a_k}{\tau} y_{n-k} = \sum_{k=0}^m b_k f_{n-k}, \quad n = m, m+1, \dots, \quad (2)$$

где  $a_k, b_k$  — числовые коэффициенты,

$$f_{n-k} = f(t_{n-k}, y_{n-k}), \quad a_0 \neq 0, \quad b_m \neq 0.$$

В частности, при  $m = 1$ ,  $b_0 = 0$ ,  $b_1 = -a_0$ ,  $a_1 = -a_0$  получаем схему Эйлера.

Схема (2) называется явной (экстраполяционной), если  $b_0 = 0$  и значения  $y_n$  определяются через предыдущие значения  $y_{n-1}, y_{n-2}, \dots, y_{n-m}$  по явной формуле

$$y_n = \frac{1}{a_0} \sum_{k=1}^m (b_k \tau f_{n-k} - a_k y_{n-k}) = \frac{1}{a_0} F(y_{n-1}, y_{n-2}, \dots, y_{n-m}).$$

Вычисления начинаются с  $n = m$ . Чтобы найти  $y_m$ , надо задать  $m$  начальных значений  $y_0, y_1, \dots, y_{m-1}$ , их можно найти, например, методом Рунге — Кутта, который использует лишь одно начальное значение  $y_0 = u_0$ .

Если  $b_0 \neq 0$ , то схема (2) называется неявной (интерполяционной): для нахождения  $y_n$  при каждом  $n$  надо решать нелинейное уравнение

$$a_0 y_n - b_0 f(t_n, y_n) = F(y_{n-1}, y_{n-2}, \dots, y_{n-m}). \quad (3)$$

Это нелинейное уравнение можно решать, например, методом Ньютона.

Погрешность аппроксимации схемы (2) на решении  $u = u(t)$  уравнения (1), или невязка, определяется по формуле

$$\psi_n = \sum_{k=0}^m b_k f(t_{n-k}, u_{n-k}) - \frac{1}{\tau} \sum_{k=0}^m a_k u_{n-k}. \quad (4)$$

Говорят, что схема (2) имеет *s-й порядок аппроксимации* (или просто, что схема (2) имеет *s-й порядок*), если

$$\|\psi\|_c = O(\tau^s) \quad \text{или} \quad \|\psi\|_c \leq M \tau^s, \quad s > 0, \quad (5)$$

где  $M = \text{const} > 0$  не зависит от  $\tau$ .

Коэффициенты  $a_k, b_k$  подбирают, исходя из требований аппроксимации и устойчивости. Без нарушения общности

можно считать, что

$$\sum_{k=0}^m b_k = 1, \quad (6)$$

так как коэффициенты уравнения (2) определены с точностью до множителя. Разлагая  $\Psi_n$  по степеням  $t$  и требуя, чтобы невязка имела заданный порядок, получаем условия для определения  $a_k, b_k$ . Поскольку  $u = 1$  есть решение уравнения  $u_t = f(t, u)$  при  $f = 0$ , из (2) следует, что

$$\sum_{k=0}^m a_k = 0. \quad (7)$$

Обычно для построения схем (2) применяют другие приемы, использующие интерполяционные и квадратурные формулы. Так, интегрируя дифференциальное уравнение (1) по  $t$  в пределах от  $t_{n-n_0}$  до  $t_n$ , получаем

$$u_n - u_{n-n_0} = \int_{t_{n-n_0}}^{t_n} f(t, u(t)) dt. \quad (8)$$

Чтобы получить отсюда разностную схему, можно использовать для интеграла какую-либо квадратурную формулу.

**2. Метод Адамса.** Каждая квадратурная формула порождает соответствующий метод численного решения обыкновенного дифференциального уравнения (1). Заменим в тождестве

$$u_n - u_{n-1} = \int_{t_{n-1}}^{t_n} f(t, u(t)) dt, \quad (9)$$

которое соответствует тождеству (8) при  $n_0 = 1$ , интеграл квадратурной формулой:

$$\int_{t_{n-1}}^{t_n} f(t, u(t)) dt \approx \tau \sum_{k=0}^m b_k f(t_{n-k}, u_{n-k}). \quad (10)$$

Учитывая (9) и (10), можно написать *разностную схему Адамса*:

$$\frac{y_n - y_{n-1}}{\tau} = \sum_{k=0}^m b_k f(t_{n-k}, y_{n-k}). \quad (11)$$

Она может быть получена из (2), если положить  $a_k = 0$  при  $k = 2, 3, \dots, m$  и  $a_0 = 1, a_1 = -1$ .

Квадратурная формула (10), на основе которой построена схема Адамса, содержит узлы сеток, не принадлежащие интервалу интегрирования  $t_{n-1} \leq t \leq t_n$ . Обычно используется требование, чтобы квадратурная формула была точной для многочлена степени  $m$ . При этом выбирается интерполяционный многочлен с узлами  $t_n, t_{n-1}, \dots, t_{n-m}$ .

При таком построении схемы ее погрешность аппроксимации совпадает с погрешностью квадратурной формулы. В самом деле, невязка для схемы (11) равна

$$\psi_n = \sum_{k=0}^m b_k f(t_{n-k}, u_{n-k}) - \frac{u_n - u_{n-1}}{\tau}.$$

Подставляя сюда из (9) выражение

$$\frac{u_n - u_{n-1}}{\tau} = \frac{1}{\tau} \int_{t_{n-1}}^{t_n} f(t, u(t)) dt,$$

получаем формулу для невязки:

$$\psi_n = \sum_{k=0}^m b_k f(t_{n-k}, u_{n-k}) - \frac{1}{\tau} \int_{t_{n-1}}^{t_n} f(t, u(t)) dt. \quad (12)$$

**3. Явные и неявные схемы.** Если  $b_0 = 0$ , то схема (11) является явной и

$$y_n = y_{n-1} + \tau \sum_{k=1}^m b_k f_{n-k}. \quad (13)$$

Простейшим примером явной схемы Адамса является схема Эйлера

$$y_n - y_{n-1} = \tau f_{n-1} \text{ при } m = 1, \quad b_0 = 0, \quad b_1 = 1. \quad (14)$$

Если положить в (11)  $m = 1, b_0 = 1, b_1 = 0$ , то получим неявную схему Адамса

$$\frac{y_n - y_{n-1}}{\tau} = f_n, \quad \text{или} \quad y_n - \tau f(t_n, y_n) = y_{n-1}. \quad (15)$$

Неявная симметричная одшаговая ( $m = 1$ ) схема

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{2} [f(t_n, y_n) + f(t_{n-1}, y_{n-1})] \quad (16)$$

соответствует значениям  $m = 1$ ,  $b_0 = b_1 = 1/2$  и имеет второй порядок аппроксимации:  $\psi_n = O(\tau^2)$ . Для определения  $y_n$  надо решать (при каждом  $n$ ) нелинейное уравнение  $y_n - \frac{1}{2}\tau f(t_n, y_n) = F_{n-1}$ , где  $F_{n-1} = y_{n-1} + \frac{1}{2}\tau f(t_{n-1}, y_{n-1})$ .

Рассмотрим теперь двухшаговые схемы Адамса, соответствующие  $m = 2$ . Явная двухшаговая ( $m = 2$ ) схема имеет вид

$$\begin{aligned} \frac{y_n - y_{n-1}}{\tau} &= \frac{3}{2} f_{n-1} - \frac{1}{2} f_{n-2}, \\ m = 2, \quad b_0 &= 0, \quad b_1 = \frac{3}{2}, \quad b_2 = -\frac{1}{2}. \end{aligned} \quad (17)$$

Она имеет второй порядок аппроксимации:

$$\psi_n = \frac{3}{2} f(t_{n-1}, u_{n-1}) - \frac{1}{2} f(t_{n-2}, u_{n-2}) - \frac{u_n - u_{n-1}}{\tau} = O(\tau^2).$$

Исследуем устойчивость соответствующей модельной схемы

$$\frac{y_n - y_{n-1}}{\tau} + \lambda \left( \frac{3}{2} y_{n-1} - \frac{1}{2} y_{n-2} \right) = 0. \quad (18)$$

Подставляя сюда  $y_n = q^n$ , получим

$$q^2 - \left( 1 - \frac{3}{2} \mu \right) q - \frac{1}{2} \mu = 0, \quad \mu = \lambda \tau. \quad (19)$$

Так как  $D = 1 - \mu + \frac{9}{4} \mu^2 > 0$  при любых  $\mu$ , то корни  $q_1$ ,  $q_2$  действительны и различны. Устойчивость означает, что  $|q_1| \leq 1$  и  $|q_2| \leq 1$ . Воспользуемся следующим свойством, которое проверяется непосредственно: корни квадратного уравнения  $q^2 + bq + c = 0$  не превосходят по модулю единицу:

$$|q_{1,2}| \leq 1, \text{ если } |b| \leq 1 + c, \quad c \leq 1. \quad (20)$$

Для уравнения (19) имеем  $b = 3\mu/2 - 1$ ,  $c = -\mu/2$ , и условие  $|3\mu/2 - 1| \leq 1 - \mu/2$  выполнено при  $\mu \leq 1$ , или

$$\tau \lambda \leq 1,$$

т. е. схема (18) условно устойчива (шаг  $\tau$  должен быть в 2 раза меньше допустимого шага в схеме Эйлера).

Напишем двухшаговую ( $m = 2$ ) неявную схему Адамса. Требуя, чтобы квадратурная формула (10) была точной для полиномов степени 0, 1, 2, т. е.  $F(t) = f(t, u(t)) =$

$\tau = \{1, t, t^2\}$ , находим коэффициенты  $b_0 = 5/12$ ,  $b_1 = 8/12$ ,  $b_2 = -1/12$ . Схема имеет вид

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{12} (5f_n + 8f_{n-1} - f_{n-2}). \quad (21)$$

Исследуем устойчивость модельной задачи

$$\frac{y_n - y_{n-1}}{\tau} + \frac{\lambda}{12} (5y_n + 8y_{n-1} - y_{n-2}) = 0. \quad (22)$$

Полагая  $y_n = q^n$ , получим характеристическое уравнение  $aq^2 + bq + c = 0$ ,  $a = 1 + \frac{5}{12}\tau\lambda_1$ ,  $b = \frac{8}{12}\tau\lambda - 1$ ,

$$c = -\frac{1}{12}\tau\lambda.$$

Условия (20), при которых  $|q_{1,2}| \leq 1$ , принимают вид  $|b| \leq a + c$ ,  $c \leq a$ . Отсюда следует, что схема (22) устойчива при  $\tau\lambda \leq 6$ .

**4. Задача Коши для уравнения второго порядка.** Рассмотрим задачу Коши:

$$\begin{aligned} \frac{d^2u}{dt^2} &= f(t, u(t)), \quad t > 0, \quad u(0) = u_0, \\ \frac{du}{dt}(0) &= u_1. \end{aligned} \quad (23)$$

Наиболее распространенными являются методы Штёрмера:

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} = \sum_{k=-1}^m b_k f(t_{n-k}, y_{n-k}), \quad m \geq 0, \quad n = 1, 2, \dots, \quad (24)$$

$$y_0 = u_0, \quad y_1 = \bar{u}_1 \quad \text{или} \quad \frac{y_1 - y_0}{\tau} = \tilde{u}_1.$$

Значение  $\bar{u}_1$  (или  $\tilde{u}_1$ ) выбирается так, чтобы погрешность аппроксимации  $v = \frac{1}{\tau}[u(\tau) - u(0)] - \dot{u}(0) - \tilde{u}_1$  имела определенный порядок, например,  $v = O(\tau^p)$ , где  $p$  — порядок аппроксимации схемы (24). Например, при  $p = 2$  находим

$$u(\tau) = u(0) + \tau \dot{u}(0) + \frac{1}{2} \tau^2 \ddot{u}(0) + O(\tau^3),$$

$$\begin{aligned} v &= u_1 + \frac{\tau}{2} \ddot{u}(0) - \tilde{u}_1 + O(\tau^2) = \frac{\tau}{2} f(0, u(0)) + \\ &\quad + O(\tau^2) - \tilde{u}_1 + u_1 = O(\tau^2), \end{aligned}$$

если положить

$$\tilde{u}_1 = u_1 + \frac{1}{2}\tau f(0, u_0), \quad \bar{u}_1 = u_0 + \tau \tilde{u}_1.$$

Если  $b_{-1} = 0$ , то схема (24) — явная, так как в правую часть входят только известные значения  $y_n, y_{n-1}, \dots, y_{n-m}$ . Если  $b_{-1} \neq 0$ , то схема (24) — неявная и для определения  $y_{n+1}$  надо решать уравнение

$$y_{n+1} - b_{-1}f(t_{n+1}, y_{n+1}) = F(y_n, y_{n-1}, \dots, y_{n-m}, t_n).$$

Для получения разностной схемы (24) вычислим интеграл

$$\begin{aligned} \int_{t_{n-1}}^{t_{n+1}} u''v dt &= \int_{t_{n-1}}^{t_n} u''v dt + \int_{t_n}^{t_{n+1}} u''v dt = \\ &= (u'v - uv') \Big|_{t_{n-1}}^{t_n} + (u'v - uv') \Big|_{t_{n-1}}^{t_n} + \int_{t_{n-1}}^{t_n} uv'' dt, \end{aligned} \quad (25)$$

где  $v(t)$  — кусочно-линейная функция

$$v(t) = \begin{cases} (t - t_{n-1})/\tau & \text{при } t_{n-1} \leq t \leq t_n, \\ (t_{n+1} - t)/\tau & \text{при } t_n \leq t \leq t_{n+1}. \end{cases} \quad (26)$$

Подставим (26) в (25) и учтем, что  $v''(t) = 0$ :

$$\int_{t_{n-1}}^{t_{n+1}} u''v dt = \frac{1}{\tau} (u_{n-1} - 2u_n + u_{n+1}). \quad (27)$$

Умножая затем уравнение (23) на  $v(t)$  и учитывая (27), получим тождество

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{\tau^2} = \frac{1}{\tau} \int_{t_{n-1}}^{t_{n+1}} f(t, u(t)) v(t) dt. \quad (28)$$

Погрешность аппроксимации схемы (24) на решении  $u = u(t)$ , или невязка для схемы (24), определяется формулой

$$\psi_n = \sum_{k=-1}^m b_k f(t_{n-k}, u_{n-k}) - \frac{u_{n+1} - 2u_n + u_{n-1}}{\tau^2},$$

которая в силу тождества (28) может быть записана в

виде

$$\psi_n = \sum_{k=-1}^m b_k f(t_{n-k}, u_{n-k}) - \frac{1}{\tau} \int_{t_{n-1}}^{t_{n+1}} f(t, u(t)) v(t) dt. \quad (29)$$

Вводя новую переменную  $s = (t - t_n)/\tau$ , запишем интеграл в более удобном виде:

$$\begin{aligned} \frac{1}{\tau} \int_{t_{n-1}}^{t_{n+1}} F(t) v(t) dt &= \int_{-1}^1 F(t_n + s\tau) \tilde{v}(s) ds, \\ F = f(t, u(t)), \tilde{v}(s) &= \begin{cases} 1+s, & s < 0, \\ 1-s, & s > 0. \end{cases} \end{aligned}$$

Из (29) видно, что первое слагаемое есть квадратурная формула для интеграла от функции  $F(t) = f(t, u(t))$  с весом  $v(t) \geq 0$ . Погрешность аппроксимации схемы полностью определяется погрешностью квадратурной формулы. Построенные на основе этого методы называют также *методами Адамса — Штёрмера*.

Самая простая формула прямоугольника дает схему

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} = f(t_n, y_n),$$

так как  $\frac{1}{\tau} \int_{t_{n-1}}^{t_{n+1}} v(t) dt = 1$ .

Для модельной задачи

$$\frac{d^2u}{dt^2} + \lambda u = 0, \quad t > 0, \quad u(0) = 0, \quad \frac{du}{dt}(0) = u_1$$

имеем

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} + \lambda y_n = 0.$$

Подставляя сюда  $y_n = q^n$ , находим  $q^2 - 2(1 - \tau^2 \lambda/2)q + 1 = 0$ ;  $D < 0$  при  $\lambda \tau^2 \leq 4$ ,  $\tau \leq 2/\sqrt{\lambda}$ ; при этом  $|q_1| = |q_2|$  и схема устойчива при условии  $\tau \leq 2/\sqrt{\lambda}$  или  $\tau \sqrt{\lambda} \leq 2$ .

**5. Системы уравнений.** Многие методы переносятся без изменения на задачу Коши для системы уравнений

$$\frac{du}{dt} = f(t, u), \quad t > 0, \quad u(0) = u_0, \quad (30)$$

где  $u = (u^1(t), u^2(t), \dots, u^N(t))$  — искомый,  $f = (f^1, f^2, \dots, f^N)$  — заданный векторы. Запишем (30) покомпонентно:

$$\frac{du^i}{dt} = f^i(t, u), \quad t > 0, \quad u^i(0) = u_0^i, \quad i = 1, 2, \dots, N. \quad (31)$$

Пусть  $u, v$  — два решения задачи (30) с начальными данными  $u(0) = u_0, v(0) = v_0$ . Для их разности  $z = u - v$  получим систему линейных уравнений

$$\frac{dz^i}{dt} = \sum_{j=1}^n \alpha_{ij}(t) z^j,$$

где  $\alpha_{ij}$  — значение производной  $\partial f^i / \partial u^j$  в некоторой средней точке  $(t, \bar{u}_j)$ ,  $\bar{u}_j = (v^1, v^2, \dots, v^{j-1}, u^j + \theta, z^j, u^{j+1}, \dots, u^N)$  ( $0 \leq \theta_j \leq 1, j = 1, 2, \dots, N$ ). Поэтому линейной моделью системы нелинейных уравнений (30) является линейная система

$$\frac{du^i}{dt} + \sum_{j=1}^N a_{ij} u^j = f^i(t) \quad (32)$$

или, в векторной форме,

$$\frac{du}{dt} + Au = f(t), \quad A = (a_{ij}). \quad (33)$$

Для устойчивости этого уравнения по начальным данным достаточно, чтобы матрица  $A$  была неотрицательной. В следующем параграфе будут найдены необходимые и достаточные условия устойчивости схем для систем линейных уравнений (33).

На практике часто встречаются системы уравнений, которые называются жесткими и решение которых обычными методами представляет большие трудности. Пусть  $\{\lambda_k\}$  — собственные числа матрицы  $A$  (если  $A$  — несимметричная, то  $\lambda_k$  могут быть комплексными). Систему уравнений (33) называют *жесткой*, если  $\operatorname{Re} \lambda_k > 0$  ( $k = 1, 2, \dots, N$ ) и если отношение  $\xi = \max_k \operatorname{Re} \lambda_k / \min_k \operatorname{Re} \lambda_k$  велико.

Если матрица  $A$  симметрична, то все собственные числа вещественны и жесткость системы (33) означает, что матрица  $A$  положительна и что система (33) плохо обусловлена, т. е.

$$\xi = \frac{\max_k \lambda_k}{\min_k \lambda_k} \gg 1.$$

Жесткими, в частности, являются уравнения, получающиеся при сведении уравнений с частными производными к системе обыкновенных дифференциальных уравнений путем разностной аппроксимации оператора, содержащего производные по пространственным переменным (например, оператора Лапласа в случае уравнения теплопроводности).

Явные методы оказались непригодными для численного решения жестких систем, так как они приводят к большим ограничениям на шаг из-за требований устойчивости в ущерб требованиям точности. Поясним это на примере системы двух уравнений

$$\frac{du_1}{dt} + a_1 u_1 = 0, \quad \frac{du_2}{dt} + a_2 u_2 = 0, \quad A = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}, \quad t > 0, \\ a_2 > 0, \quad a_1 > 0, \quad a_2 \gg a_1. \quad (34)$$

Решение этой системы есть вектор

$$u(t) = (u_1(t), u_2(t)), \\ u_1(t) = u_1(0) e^{-a_1 t}, \quad u_2(t) = u_2(0) e^{-a_2 t};$$

его компоненты убывают с ростом  $t$ , причем  $|u_2(t)| \ll |u_1(t)|$  при достаточно большом  $t$ .

Возьмем явную схему

$$\frac{y_1^{n+1} - y_1^n}{\tau} + a_1 y_1^n = 0, \quad \frac{y_2^{n+1} - y_2^n}{\tau} + a_2 y_2^n = 0, \\ n = 0, 1, \dots, y_i^n = y_i(t_n), \quad i = 1, 2. \quad (35)$$

Система распадается на два уравнения, каждое из которых можно решать отдельно, однако они связаны выбором общего шага  $\tau$ . Схема устойчива, если одновременно выполняются два условия  $a_1 \tau \leq 2$  и  $a_2 \tau \leq 2$ . Так как  $a_2 \gg a_1$ , то оба условия выполнены, если  $\tau \leq 2/a_2$ . Допустимый шаг  $\tau$  определяется фактически той компонентой  $u_2(t)$  решения, которая быстрее убывает.

Для решения системы (34) пригодна неявная схема

$$\frac{y_1^{n+1} - y_1^n}{\tau} + a_1 y_1^{n+1} = 0, \quad \frac{y_2^{n+1} - y_2^n}{\tau} + a_2 y_2^{n+1} = 0,$$

которая устойчива при любых  $\tau$  и  $a_1 \geq 0, a_2 \geq 0$ .

В последнее время появился ряд новых неявных схем, алгоритмов для них и программ, пригодных для решения

жестких систем линейных и нелинейных дифференциальных уравнений.

**6. Общие замечания.** 1. При выборе того или иного численного метода учитывается много обстоятельств, таких, как объем вычислений, требуемый объем оперативной памяти ЭВМ, порядок точности, устойчивость по отношению к ошибкам округления и др. Мы рассматривали всюду методы с постоянным шагом  $\tau = t_{n+1} - t_n$ . Переход к переменному шагу  $\tau_{n+1} = t_{n+1} - t_n$  носит формальный характер и для одношаговых схем не приводит к каким-либо новым принципиальным вопросам. Для многошаговых ( $m \geq 2$ ) схем формулы меняются.

В общем случае решение может быть сильно меняющейся немонотонной функцией. Естественно пользоваться неравномерной сеткой и уменьшать шаг (сгущать сетку) в области быстрого изменения функции  $u(t)$ , чтобы обеспечить более точное приближение  $u(t)$  сеточным решением. Однако заранее нам неизвестно поведение решения  $u = u(t)$ . Поэтому на практике поступают так: проводят сначала расчет на равномерной сетке; если видно, что решение  $u = u(t)$  сильно меняется на некотором интервале  $t_* < t < t^*$ , то сетка сгущается на  $[t_*, t^*]$  и проводится решение задачи на такой неравномерной сетке. Вообще рекомендуется проводить расчеты на нескольких сгущающихся сетках. Если при сгущении сетки решение мало меняется, то нужная точность достигнута. Для повышения порядка точности применим метод Рунге, использующий расчеты на разных сетках (если решение  $u = u(t)$  обладает достаточной гладкостью). В ходе расчета может оказаться необходимым использовать схемы разного порядка точности в разных областях изменения аргумента.

2. Часто приходится решать уравнения с сильно меняющимися коэффициентами, например,

$$\frac{du}{dt} = \alpha(t) u, \quad t > 0, \quad u(0) = u_0. \quad (36)$$

Такое уравнение встречается при описании задач химической кинетики. Его решением является функция

$$u(t) = u_0 \exp \left\{ \int_0^t \alpha(s) ds \right\}.$$

Если  $\alpha(t) \geq 0$ , то можно пользоваться схемой Эйлера при

любом  $\tau$ :

$$y_{n+1} = y_n + \tau \alpha_n y_n = (1 + \tau \alpha_n) y_n. \quad (37)$$

Если же  $\alpha(t) < 0$ , то может оказаться, что  $1 + \tau \alpha_1 < 0$  при некотором  $n = n_1$  и  $y_{n_1+1} < 0$ , т. е. решение теряет смысл. В этом случае можно пользоваться неявной схемой

$$\begin{aligned} y_{n+1} &= y_n + \tau \alpha_n y_{n+1}, \\ y_{n+1} &= y_n / (1 - \tau \alpha_n), \quad 1 - \tau \alpha_n > 1, \end{aligned} \quad (38)$$

которая устойчива при любых  $\tau$ . Если  $\alpha(t)$  меняет знак при некоторых значениях  $t$ , то в тех узлах, где  $\alpha(t) > 0$ , надо использовать явную схему (37), а в узлах, где  $\alpha(t) < 0$  — неявную схему (38).

Методы Адамса являются менее трудоемкими по сравнению с методами Рунге — Кутта. Недостатком методов Адамса является нестандартное начало вычислений; для определения  $y_1, y_2, \dots, y_{m-1}$  обычно используется метод Рунге — Кутта. Для двухшаговых (и тем более многошаговых) схем Адамса изменение шага  $\tau$  требует усложнения формул, в отличие от метода Рунге — Кутта. На практике используется комбинация методов Рунге — Кутта и Адамса с программой автоматического выбора шага для получения заданной точности.

### § 3. Аппроксимация задачи Коши для системы линейных обыкновенных дифференциальных уравнений первого порядка

**1. Задача Коши.** В этом параграфе мы будем изучать линейные разностные схемы (одношаговые или двухшаговые), которые появляются при аппроксимации задачи Коши для системы линейных обыкновенных дифференциальных уравнений первого порядка, а также при аппроксимации дифференциальных уравнений с частными производными (метод прямых).

Рассмотрим задачу Коши

$$\frac{du^i}{dt} + \sum_{j=1}^N a_{ij} u^j = f^i(t), \quad t \geq 0, \quad u^i(0) = u_0^i, \quad i = 1, 2, \dots, N. \quad (1)$$

Обозначая через  $A = (a_{ij})$  квадратную матрицу размера  $N \times N$  с элементами  $a_{ij}$ , не зависящими от  $t$ , через  $u(t) = (u^1(t), u^2(t), \dots, u^N(t))$  — искомый, а через  $f(t) = (f^1(t),$

$f^2(t), \dots, f^N(t)$  — заданный векторы размерности  $N$ , запишем систему в виде

$$\frac{du}{dt} + Au = f(t), \quad t \geq 0, \quad u(0) = u_0. \quad (2)$$

Сохраним то же обозначение  $A$  и для соответствующего оператора, действующего в пространстве  $H^N$  размерности  $N$  ( $A: H^N \rightarrow H^N$ ). В пространстве  $H^N$  введем скалярное произведение  $(u, v)$  и норму  $\|u\| = \sqrt{(u, u)}$ . Будем предполагать, что оператор  $A$  положителен:

$$A > 0, \text{ или } (Ax, x) > 0 \text{ для всех } x \in H^N, \quad x \neq 0.$$

Задача Коши (1) при условии (2) имеет единственное решение. В самом деле, пусть существуют два решения  $\bar{u}(t)$  и  $\tilde{u}(t)$  задачи (2). Тогда их разность удовлетворяет однородным условиям

$$\frac{dz}{dt} + Az = 0, \quad t > 0, \quad z(0) = 0, \quad z(t) = \bar{u}(t) - \tilde{u}(t). \quad (3)$$

Умножая (3) скалярно на  $z$  и учитывая, что  $\left(z, \frac{dz}{dt}\right) = \frac{1}{2} \frac{d}{dt} (z, z)$ , получаем

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|z\|^2 + (Az, z) &= 0, \\ \|z(t)\|^2 + \int_0^t (Az(t'), z(t')) dt' &= \|z(0)\|^2. \end{aligned}$$

Так как  $A > 0$ ,  $z(0) = 0$ , то отсюда следует, что

$$\|z(t)\|^2 = 0, \quad z(t) \equiv 0, \quad \bar{u}(t) \equiv \tilde{u}(t).$$

Отметим одно важное свойство решения задачи (2) при  $f(t) \equiv 0$ :

$$\|u(t)\| \leq e^{-\lambda_1 t} \|u(0)\|, \text{ если } A = A^* > 0, \quad (4)$$

где  $\lambda_1$  — наименьшее собственное значение оператора  $A$ :

$$A\xi_k = \lambda_k \xi_k, \quad k = 1, 2, \dots, N, \quad 0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N.$$

Для доказательства (4) будем искать решение  $u(t)$  задачи (2) в виде

$$u(t) = \sum_{k=1}^N \alpha_k(t) \xi_k, \quad \|u(t)\|^2 = \sum_{k=1}^N \alpha_k^2(t).$$

После подстановки этого выражения в уравнение (2) с  $f(t) \equiv 0$  найдем

$$\sum_{k=1}^N \left( \frac{d\alpha_k}{dt} + \lambda_k \alpha_k \right) \xi_k = 0,$$

и, следовательно,  $\frac{d\alpha_k}{dt} + \lambda_k \alpha_k = 0$ ,  $\alpha_k(t) = \alpha_k(0) e^{-\lambda_k t}$ ,

так что

$$\|u(t)\|^2 = \sum_{k=1}^N \alpha_k^2(0) e^{-2\lambda_k t} \leq e^{-2\lambda_1 t} \sum_{k=1}^N \alpha_k^2(0) = e^{-2\lambda_1 t} \|u(0)\|^2.$$

**2. Разностные схемы.** Введем сетку с шагом  $\tau$  по переменному  $t$ :  $\omega_\tau = \{t_n = n\tau, n = 0, 1, 2, \dots\}$  и обозначим через  $y_n = y(t_n)$  сеточную функцию аргумента  $t_n = n\tau$  (или  $n$ ) со значениями в  $H^N$ . Напишем явную схему

$$\frac{y_{n+1} - y_n}{\tau} + Ay_n = f_n, \quad n = 0, 1, 2, \dots, \quad y_0 = u_0, \quad (5)$$

так что  $y_{n+1}$  находится по явной формуле

$$y_{n+1} = y_n - \tau(Ay_n - f_n), \quad n = 0, 1, 2, \dots, \quad y_0 = u_0. \quad (5')$$

Решение  $y_n$  задачи (5) зависит не только от  $\tau$ , но и от  $N$  или от параметра  $h = 1/N$ :  $y_n = y_{n,\tau,h}$ . Фактически мы рассматриваем не одну задачу (5), а совокупность задач  $\{5_{\tau,h}\}$  для всевозможных  $\tau$  и  $h$ . Это и есть разностная схема. Ее решением является семейство функций  $\{y_{n,\tau,h}\}$ . Чтобы не усложнять запись, мы будем в тех случаях, когда это не вызывает недоразумений, индексы  $\tau$  и  $h$  опускать. Схема (5) является *одношаговой* (или *двухслойной*) *разностной схемой*.

Вообще под *двухслойной схемой* понимают уравнение, связывающее значения вектора  $y(t)$  для двух значений аргумента  $t = t_n$  и  $t = t_{n+1}$  (для двух слоев):

$$By_{n+1} = Cy_n + F_n, \quad n = 0, 1, \dots,$$

где  $B, C$  — квадратные матрицы  $N \times N$  (линейные операторы  $B, C: H^N \rightarrow H^N$ ),  $y_n, F_n$  — векторы размерности  $N$ . Это уравнение можно всегда переписать в следующем каноническом виде:

$$B \frac{y_{n+1} - y_n}{\tau} + Ay_n = \varphi_n, \quad n = 0, 1, 2, \dots, \quad y_0 = u_0. \quad (6)$$

Для определения  $y_{n+1}$  надо решить уравнение

$$By_{n+1} = \Phi_n, \quad \Phi_n = By_n - \tau(Ay_n - \varphi_n).$$

Будем всюду предполагать существование обратного оператора  $B^{-1}$ .

Если  $B = E$  — единичный оператор, то мы получаем явную схему (5). В случае  $B \neq E$  схему (6) называют неявной. Часто встречаются схемы

$$\frac{y_{n+1} - y_n}{\tau} + Ay_{n+1} = \varphi_n \quad (\text{чисто неявная схема}), \quad (7)$$

$$\frac{y_{n+1} - y_n}{\tau} + \frac{1}{2} A(y_n + y_{n+1}) = \varphi_n \quad (\text{симметричная схема}). \quad (8)$$

Они являются частными случаями (при  $\sigma = 1$  и  $\sigma = 1/2$ ) схемы с весами

$$\frac{y_{n+1} - y_n}{\tau} + A(\sigma y_{n+1} + (1 - \sigma)y_n) = \varphi_n, \quad n = 0, 1, \dots, \quad (9)$$

которую можно записать в каноническом виде (6) с

$$B = E + \sigma\tau A, \quad (10)$$

если учесть, что  $\sigma y_{n+1} + (1 - \sigma)y_n = y_n + \sigma\tau(y_{n+1} - y_n)/\tau$ .

**3. Погрешность аппроксимации.** Пусть  $u = u(t)$  — решение задачи (2),  $y_n = y(t_n)$  — решение задачи (6); подставляя в (6)  $y_n = u_n + z_n$ , для погрешности  $z_n = y_n - u_n$ ,  $u_n = u(t_n)$ , получаем

$$B \frac{z_{n+1} - z_n}{\tau} + Az_n = \psi_n, \quad n = 0, 1, 2, \dots, \quad z_0 = 0, \quad (11)$$

где

$$\psi_n = \varphi_n - Au_n - B \frac{u_{n+1} - u_n}{\tau} \quad (12)$$

есть невязка, или погрешность аппроксимации для схемы (6) на решении  $u = u(t)$  исходной задачи (2).

Пусть  $\|u\|_{(1)}, \|v\|_{(2)}$  — некоторые нормы в  $H^N = H_h$ . Схема (6) сходится, если  $\|z_n\|_{(1)} \rightarrow 0$  при  $\tau \rightarrow 0$  для всех  $n = 1, 2, \dots$ . Схема (6) имеет  $m$ -й порядок точности, или сходится со скоростью  $O(\tau^m)$ , если

$$\|z_n\|_{(1)} = O(\tau^m), \quad \text{т. е. } \|z_n\|_{(1)} \leq M\tau^m, \quad (13)$$

где  $M = \text{const}$  не зависит от  $\tau$ .

Напомним, что схема (6) имеет  $m$ -й порядок аппроксимации на решении уравнения (1), если для невязки  $\psi_n$  выполняется оценка

$$\|\psi_n\|_{(2)} = O(\tau^m). \quad (14)$$

Выясним условия аппроксимации схемы (6) с  $m = 1, 2$ . Предполагая, что  $u = u(t)$  имеет нужное по ходу изложения число производных, находим

$$\begin{aligned} u_{n+1} &= \left( u + \frac{\tau}{2} \dot{u} + \frac{\tau^2}{8} \ddot{u} \right)_{n+1/2} + O(\tau^3), \\ \dot{u}_n &= \left( \frac{du}{dt} \right)_n, \quad \ddot{u}_n = \left( \frac{d^2u}{dt^2} \right)_n, \\ u_n &= \left( u - \frac{\tau}{2} \dot{u} + \frac{\tau^2}{8} \ddot{u} \right)_{n+1/2} + O(\tau^2), \\ \frac{1}{\tau} (u_{n+1} - u_n) &= \dot{u}_{n+1/2} + O(\tau^2), \\ \psi_n &= \varphi_n - (Au + B\dot{u})_{n+1/2} + \frac{\tau}{2} A\dot{u}_{n+1/2} + O(\tau^2) = \\ &= \varphi_n - f_{n+1/2} + (f - Au - \dot{u})_{n+1/2} + \\ &\quad + \left( E - B + \frac{\tau}{2} A \right) \dot{u}_{n+1/2} + O(\tau^2) = \\ &= \varphi_n - f_{n+1/2} + \left( E - B + \frac{\tau}{2} A \right) \dot{u}_{n+1/2} + O(\tau^2). \end{aligned}$$

Отсюда видно, что условие (14) будет выполнено, если

$$\begin{aligned} \|\varphi_n - f_{n+1/2}\|_{(2)} &= O(\tau^m), \\ \left\| \left( E - B + \frac{\tau}{2} A \right) \dot{u} \right\|_{(2)} &= O(\tau^m), \quad m = 1, 2. \end{aligned} \quad (15)$$

В частности, для явной схемы (в случае  $B = E$ ) имеем

$$\left\| \frac{\tau}{2} A\dot{u} \right\|_{(2)} = O(\tau),$$

и  $\|\psi_n\|_{(2)} = O(\tau)$  при  $\|\varphi_n - f_{n+1/2}\| = O(\tau)$ , например, при  $\varphi_n = f_n$ .

Для симметричной схемы ( $\sigma = 1/2$ )  $B = E + \tau A/2$ , если  $\|\varphi_n - f_{n+1/2}\|_{(2)} = O(\tau^2)$ , то  $\|\psi_n\|_{(2)} = O(\tau^2)$ , поскольку  $\|(E - B + \tau A/2)\dot{u}\|_{(2)} = 0$ ; при этом можно взять, например,  $\varphi_n = f_{n+1/2}$ .

*Схема с опережением* ( $\sigma = 1$ ) имеет 1-й порядок аппроксимации, так как  $\|(E - B + \tau A/2)\dot{u}\|_{(2)} = \tau \|A\dot{u}\|_{(2)}/2 = O(\tau)$ .

**4. Устойчивость и сходимость.** Как отмечалось выше, схема (6) *устойчива* (по начальным данным и по правой части), если ее решение непрерывно зависит от входных данных (от  $y_0$  и от  $\varphi_n$ ), причем эта зависимость непрерывна по  $\tau$  и  $N$ , или по  $h$ . Для оценки решения задачи пользуемся нормой  $\|u\|_{(1)}$ , а для оценки правой части — нормой  $\|v\|_{(2)}$ . Воспользуемся более строгим определением устойчивости.

Схема (6) будет *устойчивой*, если для любых  $y_0$ ,  $\varphi_n$  существуют такие постоянные  $M_1 > 0$  и  $M_2 > 0$ , не зависящие ни от  $\tau$ , ни от  $N$ ,  $y_0$ ,  $\varphi_n$ , что для решения задачи (6) выполняется неравенство

$$\|y_n\|_{(1)} \leq M_1 \|y_0\|_{(1)} + M_2 \max_{0 \leq k \leq n} \|\varphi_k\|_{(2)}. \quad (16)$$

Если схема (6) устойчива и обладает аппроксимацией  $\|\psi_n\|_{(2)} \rightarrow 0$  при  $\tau \rightarrow 0$ , то она сходится:

$$\|y_n - u_n\|_{(1)} \rightarrow 0 \text{ при } \tau \rightarrow 0, \quad n = 1, 2, \dots \quad (17)$$

(из аппроксимации и устойчивости следует сходимость схемы). В самом деле, если схема (6) устойчива, то для решения  $z_n = y_n - u_n$  задачи (11), согласно (16), выполняется оценка

$$\|z_n\|_{(1)} \leq M_1 \max_{0 \leq k \leq n} \|\psi_k\|_{(2)}. \quad (18)$$

Отсюда и следует, что  $\|z_n\|_{(1)} \rightarrow 0$ , если  $\|\psi_n\|_{(2)} \rightarrow 0$  при  $\tau \rightarrow 0$ .

Изучение сходимости и порядка точности сводится к изучению погрешности аппроксимации и устойчивости разностной схемы (6).

#### § 4. Устойчивость двухслойной схемы

**1. Устойчивость по начальным данным.** Будем рассматривать двухслойную схему в канонической форме

$$B \frac{y_{n+1} - y_n}{\tau} + Ay_n = \varphi_n, \quad n = 0, 1, \dots,$$

задано начальное значение  $y_0 \in H$ , (1)

где  $A, B: H \rightarrow H$  ( $H = H^N$ ). Решение задачи (1) можно представить в виде суммы  $y = y^{(1)} + y^{(2)}$  решений двух

задач

$$B \frac{y_{n+1} - y_n}{\tau} + Ay_n = 0, \quad n = 0, 1, \dots, \quad y_0 = u_0, \quad (2)$$

$$B \frac{y_{n+1} - y_n}{\tau} + Ay_n = \varphi_n, \quad n = 0, 1, \dots, \quad y_0 = 0, \quad (3)$$

( $y^{(1)}$  — решение задачи (2),  $y^{(2)}$  — решение задачи (3)).

Схема (1) устойчива по начальным данным, если для решения задачи (2) верна оценка

$$\|y_n\|_{(1)} \leq M_1 \|y_0\|_{(1)}. \quad (4)$$

Схема (1) устойчива по правой части, если для решения задачи (3) верна оценка

$$\|y_n\|_{(1)} \leq M_2 \max_{0 \leq k < n} \|\varphi_k\|_{(2)}. \quad (5)$$

Здесь  $M_1, M_2$  не зависят от  $N, \tau, n$ .

Мы будем пользоваться более простым условием устойчивости по начальным данным:

$$\|y_n\|_{(1)} \leq \|y_{n-1}\|_{(1)}, \dots, \|y_1\|_{(1)} \leq \|y_0\|_{(1)} \quad (M_1 = 1), \quad (6)$$

а также условием  $\rho$ -устойчивости:

$$\|y_n\|_{(1)} \leq \rho \|y_{n-1}\|_{(1)} \leq \dots \leq \rho^n \|y_0\|_{(1)}, \quad \rho > 0. \quad (7)$$

Очевидно, что схема устойчива в смысле определения (4), если  $\rho = e^{c_0 \tau}$ , где  $c_0 = \text{const}$  не зависит от  $n, \tau, N$ . В этом случае  $\rho^n = e^{c_0 t_n} \leq e^{c_0 T} = M_1$  при  $0 \leq t_n \leq T, c_0 > 0$  или  $\rho^n \leq 1$  при  $c_0 \leq 0$ .

В пространстве  $H$  введем скалярное произведение (,) и норму  $\|x\| = \sqrt{(x, x)}$ . Пусть  $D = D^* > 0$  — самосопряженный положительный оператор. В качестве нормы  $\|y\|_{(1)}$  выберем энержетическую норму

$$\|y\|_{(1)} = \|y\|_D = \sqrt{(Dy, y)}. \quad (8)$$

В частности,  $D = A, D = E$  или  $D = B$  (при  $B = B^* > 0$ ). Из (2) следует, что

$$y_{n+1} = Sy_n, \quad S = E - \tau B^{-1}A, \quad (9)$$

где  $S$  — оператор перехода со слоя на слой.

Схема (2) устойчива в  $H_D$ , если справедлива оценка

$$\|y_{n+1}\|_D \leq \|y_n\|_D. \quad (10)$$

Из оценки  $\|y_{n+1}\|_D = \|Sy_n\|_D \leq \|S\|_D \|y_n\|_D$  следует, что

неравенство (10) эквивалентно условию

$$\|S\|_D \leq 1. \quad (11)$$

Это условие, в свою очередь, эквивалентно условию

$$J_D = \|y\|_D^2 - \|Sy\|_D^2 = (Dy, y) - (DSy, Sy) \geq 0 \quad \text{для всех } y \in H. \quad (12)$$

Таким образом, (10), (11) и (12) эквивалентны, т. е. выполнение любого из них влечет за собой выполнение двух других.

## 2. Необходимое и достаточное условие устойчивости. Основная теорема.

**Теорема 4.** Если  $A = A^*$  — самосопряженный положительный оператор и существует оператор  $B^{-1}$ , то для устойчивости схемы (2) в  $H_A$ :

$$\|y_{n+1}\|_A \leq \|y_n\|_A \quad (13)$$

необходимо и достаточно, чтобы выполнялось неравенство

$$(By, y) - \frac{\tau}{2} (Ay, y) \geq 0 \quad \text{для всех } y \in H, \text{ или } B \geq \frac{\tau}{2} A. \quad (14)$$

**Доказательство.** Достаточно убедиться в эквивалентности (14) и неравенства  $J_A \geq 0$ , где

$$\begin{aligned} J_A &= (Ay, y) - (ASy, Sy) = \\ &= (Ay, y) - (Ay - \tau AB^{-1}Ay, y - \tau B^{-1}Ay) = \\ &= 2\tau(AB^{-1}Ay, y) - \tau^2(AB^{-1}Ay, B^{-1}Ay). \end{aligned}$$

Обозначив  $B^{-1}Ay = x$ ,  $Ay = Bx$ , получим

$$J_A = 2\tau \left( (Bx, x) - \frac{\tau}{2} (Ax, x) \right) \geq 0 \quad \text{для всех } x \in H, \quad (15)$$

т. е. неравенства (14), (15) и, следовательно, (13), (14) эквивалентны. Это значит, что из (14) следует (11), (12) при  $D = A$  и (13) (условие (14) достаточно для устойчивости). Если же схема устойчива, т. е. выполнено (13) или  $\|S\|_A \leq 1$ , то  $J_A \geq 0$  и, следовательно,  $B \geq \tau A/2$  (необходимость условия (14)).

**Замечание.** Условие (14) можно пояснить на примере разностной схемы

$$b \frac{y_{n+1} - y_n}{\tau} + ay_n = 0, \quad n = 0, 1, 2, \dots, \quad a > 0, \quad b > 0$$

с числовыми коэффициентами  $a, b$ . Эта схема соответствует задаче Коши

$$bu'(t) + au(t) = 0, \quad t > 0, \quad u(0) = u_0.$$

Из формулы  $y_{n+1} = (1 - \tau a/b)y_n$  видно, что схема устойчива, т. е.  $|y_{n+1}| \leq |y_n| \leq \dots \leq |y_0|$ , если  $|1 - \tau a/b| \leq 1$ ,  $-1 \leq 1 - \tau a/b \leq 1$ , т. е.  $b \geq \tau a/2$ . Аналогия с операторным неравенством  $B \geq \tau A/2$  очевидна.

**3. Примеры применения основной теоремы.** Пример 1. Явная схема:  $B = E$ ,  $A = A^* > 0$ . Из неравенства Коши — Буняковского  $(Ax, x) \leq \|Ax\| \|x\| \leq \|A\| \|x\|^2$  следует  $A \leq \|A\|E$ , или

$$E \geq \frac{1}{\|A\|} A. \quad (16)$$

Рассмотрим теперь разность  $B - \frac{1}{2}\tau A = E - \frac{1}{2}\tau A \geq \geq \frac{1}{\|A\|}A - \frac{1}{2}\tau A = \left(\frac{1}{\|A\|} - \frac{\tau}{2}\right)A$ . Так как  $A > 0$ , то условие  $B - \frac{1}{2}\tau A \geq 0$  будет выполнено при  $\frac{1}{\|A\|} - \frac{\tau}{2} \geq \geq 0$ , т. е. при

$$\tau \leq 2/\|A\|. \quad (17)$$

Это необходимое и достаточное условие устойчивости явной схемы в  $H_A (\|y_n\|_A \leq \|y_0\|_A)$ .

Пример 2. Схема (9) из § 3 с весами,  $A = A^* > 0$ . Для нее  $B = E + \sigma\tau A$  и  $B - \frac{1}{2}\tau A = E + \left(\sigma - \frac{1}{2}\right)\tau A \geq \geq \left(\frac{1}{\|A\|} + \left(\sigma - \frac{1}{2}\right)\tau\right)A \geq 0$ , если

$$1 + \left(\sigma - \frac{1}{2}\right)\tau\|A\| \geq 0. \quad (18)$$

Отсюда видно, что схема с весами устойчива в  $H_A$  при любых  $\tau > 0$  (безусловно устойчива), если  $\sigma \geq 1/2$ , и условно устойчива при  $\tau \leq 1/[(1/2 - \sigma)\|A\|]$ , если  $\sigma < 1/2$ .

Пример 3. Устойчивость в  $H$  (при  $D = E$ ) схемы с весами (9) из § 3:

$$(E + \sigma\tau A) \frac{y_{n+1} - y_n}{\tau} + Ay_n = 0, \quad n = 0, 1, 2, \dots, \quad (19)$$

$$B = E + \sigma\tau A.$$

Применяя оператор  $A^{-1}$  к обеим частям уравнения (19),

получаем

$$\begin{aligned} \tilde{B} \frac{y_{n+1} - y_n}{\tau} + \tilde{A} y_n &= 0, \quad n = 0, 1, 2, \dots, \\ \tilde{B} &= A^{-1} + \sigma \tau E, \quad \tilde{A} = E. \end{aligned} \tag{20}$$

Эта схема устойчива в силу теоремы 1 в  $H_{\tilde{A}} = H (\tilde{A}^* = \tilde{A} = E > 0)$  при  $\tilde{B} = \frac{1}{2} \tau \tilde{A} = A^{-1} + \left( \sigma - \frac{1}{2} \right) \tau E \geqslant \left( \frac{1}{\|A\|} + \left( \sigma - \frac{1}{2} \right) \tau \right) E \geqslant 0$ , т. е. при выполнении (18) (при этом мы учли оценку  $A^{-1} \geqslant \frac{1}{\|A\|} E$ , которая следует из (16)). Таким образом, из (18) следует, что для (19) верна оценка (10) при  $D = \tilde{A}$ , т. е.

$$\|y_n\| \leq \|y_0\|. \tag{21}$$

Схему (19) можно записать в виде

$$\begin{aligned} y_{n+1} &= S y_n, \quad S = (E + \sigma \tau A)^{-1} (E - (1 - \sigma) \tau A), \\ A &= A^* > 0. \end{aligned} \tag{22}$$

Поэтому для нее при условии (18) верна оценка (21), что означает

$$\begin{aligned} \| (E + \sigma \tau A)^{-1} (E - (1 - \sigma) \tau A) \| &\leq 1, \\ \text{если } 1 + \left( \sigma - \frac{1}{2} \right) \tau \|A\| &\geq 0. \end{aligned} \tag{23}$$

Эта оценка понадобится в дальнейшем.

#### 4. Устойчивость в $H_B$ .

**Теорема 2.** *Если  $A = A^* > 0$ ,  $B = B^* > 0$ , то для устойчивости схемы (2) в  $H_B$ :*

$$\|y_{n+1}\|_B \leq \|y_n\|_B \tag{24}$$

*необходимо и достаточно, чтобы выполнялось условие (14).*

**Доказательство.** Схему (2) запишем в виде (9) и покажем, что условие

$$\|S\|_B \leq 1 \tag{25}$$

эквивалентно неравенству (14), т. е. из (14) следует (25) и, обратно, из (25) следует (14).

Пусть  $y$  — произвольный вектор из  $H$ ; представим его в виде

$$y = \sum_{k=1}^N \alpha_k \xi_k,$$

где  $\{\xi_k\}$  — собственные векторы задачи:

$$A\xi_k = \lambda_k B\xi_k, \quad \lambda_k > 0,$$

$$(B\xi_k, \xi_m) = \delta_{km} = \begin{cases} 1, & k = m, \\ 0, & k \neq m, \end{cases} \quad k, m = 1, 2, \dots, N. \quad (26)$$

Учитывая, что  $S\xi_k = \xi_k - \tau B^{-1}A\xi_k = (1 - \tau\lambda_k)\xi_k$ ,  $BS\xi_k = (1 - \tau\lambda_k)B\xi_k$ , найдем

$$\begin{aligned} (By, y) &= \sum_{k=1}^N \alpha_k^2, \quad (Ay, y) = \sum_{k=1}^N \lambda_k \alpha_k^2, \\ (BSy, Sy) &= \sum_{k=1}^N \alpha_k^2 (1 - \tau\lambda_k)^2 \leq \|S\|_B^2 \sum_{k=1}^N \alpha_k^2 = \|S\|_B^2 (By, y), \end{aligned} \quad (27)$$

где

$$\|S\|_B^2 = \max_{1 \leq k \leq N} (1 - \tau\lambda_k)^2. \quad (28)$$

Неравенство (25) эквивалентно условию

$$\tau\lambda_k \leq 2, \quad k = 1, 2, \dots, N, \quad (29)$$

которое, в свою очередь, эквивалентно неравенству (14), так как

$$(By, y) - \frac{\tau}{2} (Ay, y) = \sum_{k=1}^N \alpha_k^2 \left(1 - \frac{\tau\lambda_k}{2}\right).$$

Тем самым эквивалентность (24) и (14) доказана.

### 5. $\rho$ -устойчивость.

**Теорема 3.** Если  $A = A^* > 0$ ,  $B = B^* > 0$ , то необходимым и достаточным условием  $\rho$ -устойчивости схемы (2) с любым  $\rho > 0$ :

$$\|y_{n+1}\|_D \leq \rho \|y_n\|_D, \quad D = A, B, \quad (30)$$

являются операторные неравенства

$$\frac{1 - \rho}{\tau} B \leq A \leq \frac{1 + \rho}{\tau} B. \quad (31)$$

**Доказательство.** Неравенства (31) эквивалентны условиям (см. гл. I, § 4, п. 4):

$$\frac{1-\rho}{\tau} \leq \lambda_k \leq \frac{1+\rho}{\tau}, \quad k = 1, 2, \dots, N, \quad (32)$$

где  $\lambda_k$  — собственные числа задачи (26).

Допустим, что  $D = B$  и верны (31) или (32). Из (32) следует  $-\rho \leq \tau\lambda_k - 1 \leq \rho$ ,  $|1 - \tau\lambda_k| \leq \rho$ , и в силу (27)  $\|S\|_B \leq \rho$  (так как  $\|S\|_B^2$  — наименьшая постоянная, для которой верно неравенство  $(BSy, Sy) \leq M(By, y)$ ), т. е. справедлива оценка (30) (достаточность). Если верна оценка (30), то  $|1 - \tau\lambda_k| \leq \rho$  и, следовательно, выполнены (32) и (31) (необходимость).

Аналогично доказывается теорема и при  $D = A$ , если учсть, что

$$(ASy, Sy) = \sum_{k=1}^N \alpha_k^2 \lambda_k (1 - \tau\lambda_k)^2 \leq \max_{1 \leq k \leq N} (1 - \tau\lambda_k)^2 (Ay, y).$$

Из (30) следует

$$\|y_n\|_D \leq \rho^n \|y_0\|_D.$$

Возникает вопрос, при каких условиях имеет место априорная оценка (30) с  $\rho < 1$ ? Ответ на него дает следующая теорема.

**Теорема 4.** *Пусть выполнены условия*

$$A = A^* > 0, \quad B = B^* > 0, \quad \gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0. \quad (33)$$

*Тогда для решения задачи (2) верна оценка*

$$\|y_{n+1}\|_D \leq \rho \|y_n\|_D, \quad \rho = 1 - \tau\gamma_1, \quad D = A, B, \quad (34)$$

*если*

$$\tau \leq \tau_0, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}. \quad (35)$$

Для доказательства надо вычислить норму  $\|S\|_B = \|S\|_A = \max_{1 \leq k \leq N} |1 - \tau\lambda_k|$  при условии, что  $\gamma_1 \leq \lambda_k \leq \gamma_2$ ,  $0 < \gamma_1 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N = \gamma_2$ . Рассмотрим разность

$$\varphi_k = (1 - \tau\lambda_1)^2 - (1 - \tau\lambda_k)^2 = 2\tau(\lambda_k - \lambda_1) \left(1 - \frac{\tau}{2}(\lambda_k + \lambda_1)\right).$$

Отсюда видно, что  $\varphi_k \geq 0$  при  $1 - \frac{\tau}{2}(\lambda_k + \lambda_1) \geq 1 - \frac{\tau}{2} \times \times (\gamma_2 + \gamma_1) \geq 1 - \frac{\tau_0}{2}(\gamma_1 + \gamma_2) = 0$ , т. е.  $\max_{1 \leq k \leq N} |1 - \tau\lambda_k| = 1 - \tau\gamma_1$ , если  $\tau \leq \tau_0$ . Теорема доказана.

**6. Устойчивость по правой части. Метод энергетических неравенств.** Рассмотрим задачу (3) и перепишем ее в виде

$$y_{n+1} = S y_n + \tau B^{-1} \varphi_n, \quad n = 0, 1, \dots, \\ S = E - \tau B^{-1} A, \quad y_0 = 0. \quad (36)$$

Воспользуемся неравенством треугольника

$$\|y_{n+1}\|_D \leq \|S y_n\|_D + \tau \|B^{-1} \varphi_n\|_D \leq \|S\|_D \|y_n\|_D + \tau \|B^{-1} \varphi_n\|_D. \quad (37)$$

Если выполнены условия теоремы 2, то  $B = B^* > 0$ ,  $D = B$  и  $\|S\|_D = \|S\|_B \leq 1$  при  $B \geq \frac{\tau}{2} A$ ,  $\|B^{-1} \varphi_n\|_B^2 = (B(B^{-1} \varphi_n), B^{-1} \varphi_n) = (B^{-1} \varphi_n, \varphi_n) = \|\varphi_n\|_{B^{-1}}^2$ , и из (37) следует

$$\|y_{n+1}\|_B \leq \|y_n\|_B + \tau \|\varphi_n\|_{B^{-1}}.$$

Суммируя по  $n = 0, 1, 2, \dots$  и учитывая, что  $y_0 = 0$ , получим

$$\|y_n\|_B \leq \sum_{k=0}^{n-1} \tau \|\varphi_k\|_{B^{-1}}. \quad (38)$$

Эта априорная оценка выражает устойчивость схемы (1) по правой части при том же условии (14).

Можно получить и другие оценки. Для этого воспользуемся весьма общим методом энергетических неравенств.

Подставим  $y_n = \frac{1}{2} (y_n + y_{n+1}) - \frac{\tau}{2} \frac{y_{n+1} - y_n}{\tau}$  в (1):

$$\left( B - \frac{\tau}{2} A \right) \frac{y_{n+1} - y_n}{\tau} + \frac{1}{2} A (y_{n+1} + y_n) = \varphi_n.$$

Умножим это уравнение скалярно на  $2(y_{n+1} - y_n)$  и учтем, что  $(A(y_{n+1} + y_n), y_{n+1} - y_n) = (Ay_{n+1}, y_{n+1}) + (Ay_n, y_{n+1}) + (Ay_{n+1}, y_n) - (Ay_n, y_n) = (Ay_{n+1}, y_{n+1}) - (Ay_n, y_n)$ , так как  $(Ay_n, y_{n+1}) = (Ay_{n+1}, y_n)$  в силу самосопряженности  $A$ . В результате получим «энергетическое тождество»

$$2\tau \left( \left( B - \frac{\tau}{2} A \right) \frac{y_{n+1} - y_n}{\tau}, \frac{y_{n+1} - y_n}{\tau} \right) + (Ay_{n+1}, y_{n+1}) = \\ = (Ay_n, y_n) + 2(\varphi_n, y_{n+1} - y_n). \quad (39)$$

Отсюда видно, что при  $\varphi_n = 0$  и  $B \geq \frac{\tau}{2} A$  верна оценка (13).

Преобразуем  $2(\varphi_n, y_{n+1} - y_n) = 2\tau \left( \varphi_n, \frac{y_{n+1} - y_n}{\tau} \right)$ . Для этого воспользуемся неравенством:

$$|ab| = (\sqrt{2\varepsilon}a)\left(\sqrt{\frac{1}{2\varepsilon}}b\right) \leq \varepsilon a^2 + \frac{1}{4\varepsilon}b^2,$$

где  $a, b, \varepsilon > 0$  — любые числа. В нашем случае

$$\begin{aligned} 2(\varphi_n, y_{n+1} - y_n) &\leq 2\tau \|\varphi_n\| \left\| \frac{y_{n+1} - y_n}{\tau} \right\| \leq \\ &\leq 2\tau\varepsilon \left\| \frac{y_{n+1} - y_n}{\tau} \right\|^2 + \frac{\tau}{2\varepsilon} \|\varphi_n\|^2. \end{aligned}$$

Подставляя эту оценку в тождество (39), получим

$$\begin{aligned} 2\tau \left( \left( B - \varepsilon E - \frac{\tau}{2}A \right) \frac{y_{n+1} - y_n}{\tau}, \frac{y_{n+1} - y_n}{\tau} \right) + \|y_{n+1}\|_A^2 &\leq \\ &\leq \|y_n\|_A^2 + \frac{\tau}{2\varepsilon} \|\varphi_n\|^2. \quad (40) \end{aligned}$$

Если выполнено неравенство

$$B \geq \varepsilon E + \frac{\tau}{2} A, \quad \varepsilon > 0, \quad (41)$$

то из (40) следует (с заменой  $n$  на  $k$ )

$$\|y_{k+1}\|_A^2 \leq \|y_k\|_A^2 + \frac{\tau}{2\varepsilon} \|\varphi_k\|^2.$$

Суммируя по  $k = 0, 1, 2, \dots, n-1$ , получаем оценку

$$\|y_n\|_A^2 \leq \|y_0\|_A^2 + \frac{1}{2\varepsilon} \sum_{k=0}^{n-1} \tau \|\varphi_k\|^2, \quad (42)$$

которая выражает устойчивость схемы (1) по правой части и по начальным данным в  $H_A$ .

**Пример.** Схема с весами (1):  $B = E + \sigma\tau A$ . Для нее условие (41) означает, что

$$(1 - \varepsilon)E + \left( \sigma - \frac{1}{2} \right) \tau A \geq 0.$$

В частности, оценка (42) верна при  $\varepsilon = 1$  и  $\sigma \geq 1/2$ .

**7. Асимптотическая устойчивость.** Для задачи Коши

$$\frac{du}{dt} + Au = 0, \quad t > 0, \quad u(0) = u_0$$

в § 3, п. 1 была получена оценка

$$\|u(t)\| \leq e^{-\gamma_1 t} \|u(0)\|,$$

где  $\lambda_1 = \min_k \lambda_k(A)$ .

Найдем условия, при которых аналогичная оценка имеет место для схемы (2). Воспользуемся теоремой 4. Пусть выполнены условия (33). Тогда в силу (34), (35)

$$\|y_n\|_A \leq \rho^n \|y_0\|_A, \quad \rho = 1 - \tau\gamma_1, \quad \tau \leq \tau_0 = \frac{2}{\gamma_1 + \gamma_2}. \quad (43)$$

Отсюда следует оценка, выражающая свойство асимптотической устойчивости

$$\|y_n\|_A \leq e^{-\gamma_1 t_n} \|y_0\|_A \quad (44)$$

(при этом учтено, что  $\rho = 1 - \tau\gamma_1 < e^{-\tau\gamma_1}$ ).

Рассмотрим схему с весами и предположим, что

$$\delta E \leq A \leq \Delta E, \quad \delta = \lambda_1 > 0, \quad \Delta = \lambda_N > 0. \quad (45)$$

Вычислим  $\gamma_1$  и  $\gamma_2$ . Учитывая (45), имеем

$$\begin{aligned} B = E + \sigma\tau A &\geq \left(\frac{1}{\Delta} + \sigma\tau\right) A = \frac{1}{\gamma_2} A; \\ B &\leq \left(\frac{1}{\delta} + \sigma\tau\right) A = \frac{1}{\gamma_1} A, \\ \gamma_1 &= \frac{\delta}{1 + \sigma\tau\delta}, \quad \gamma_2 = \frac{\Delta}{1 + \sigma\tau\Delta}. \end{aligned} \quad (46)$$

Для явной схемы  $\gamma_1 = \delta$ ,  $\gamma_2 = \Delta$  условие асимптотической устойчивости

$$\tau \leq 2/(\delta + \Delta) \quad (47)$$

близко к условию обычной устойчивости с  $\rho = 1$ . При  $\sigma \neq 0$  условие  $\tau \leq 2/(\gamma_1 + \gamma_2)$  приводит к неравенству

$$2 + 2(\sigma - 1/2)\tau(\delta + \Delta) - 2\sigma(1 - \sigma)\tau^2\delta\Delta \geq 0.$$

При  $\sigma = 1$  оно выполнено для любого  $\tau$ , т. е. чисто неявная схема с  $\sigma = 1$  безусловно асимптотически устойчива. Симметричная схема

$$\frac{y_{n+1} - y_n}{\tau} + \frac{1}{2} A (y_{n+1} + y_n) = 0, \quad \sigma = \frac{1}{2}, \quad (48)$$

асимптотически устойчива при условии

$$\tau \leq \tau^*, \quad \tau^* = 2/\sqrt{\delta\Delta} \quad (49)$$

и безусловно устойчива в обычном смысле. В этом случае

$$\rho = e^{-\lambda_1 t + \zeta(\tau^3)} < e^{-\lambda_1 t}$$

и верна оценка

$$\|y_n\| \leq e^{-\lambda_1 t_n} \|y_0\| \quad \text{при } \tau \leq \tau^*, \quad \sigma = 1/2. \quad (50)$$

Что произойдет, если условие  $\tau \leq \tau_0$  не выполнено, т. е.  $\tau > \tau_0$ ? Тогда  $\max_k |1 - \tau \lambda_k|$  достигается не при  $k = 1$ , а при  $k = N$  и  $\rho = \tau \gamma_2 - 1$ . Асимптотика (при больших  $t_n$ ) решения разностной задачи не имеет ничего общего с асимптотическим решением исходной задачи. Таким образом, нарушение асимптотической устойчивости приводит к потере точности схемы при больших  $t$ .

## Глава VI

# РАЗНОСТНЫЕ МЕТОДЫ ДЛЯ ЭЛЛИПТИЧЕСКИХ УРАВНЕНИЙ

В этой главе мы рассмотрим разностные схемы и методы решения разностных уравнений для уравнения Пуассона и эллиптических уравнений с переменными коэффициентами.

### § 1. Разностные схемы для уравнения Пуассона

**1. Исходная задача.** Рассмотрим уравнение Пуассона

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x_1, x_2). \quad (1)$$

Будем искать его решение, непрерывное в прямоугольнике

$$\bar{G} = G \cup \Gamma = \{x = (x_1, x_2): 0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$$

и принимающее на границе  $\Gamma$  заданные значения:

$$u|_{\Gamma} = \mu(x). \quad (2)$$

Задача, определяемая уравнением (1) и условием (2), называется *задачей Дирихле (первой краевой задачей)*.

**2. Разностная схема «крест».** Для численного решения задачи (1), (2) введем в  $\bar{G}$  сетку  $\omega_h = \omega_h \cup \gamma_h = \{x_i = (i_1 h_1, i_2 h_2), i_\alpha = 0, 1, \dots, N_\alpha, h_\alpha = l_\alpha / N_\alpha, \alpha = 1, 2\}$  и обозначим через  $y_i = y_{i_1 i_2} = y(i_1, i_2) = y(x_i)$  сеточную функцию, заданную на  $\omega_h$ ;  $h_1$  и  $h_2$  — шаги сетки по координатам  $x_1$  и  $x_2$ .

Чтобы написать разностную схему для (1), (2), аппроксимируем каждую из производных  $\partial^2 u / \partial x_\alpha^2$  на трехточечном шаблоне, полагая

$$\frac{\partial^2 u}{\partial x_1^2} \sim \frac{u(x_1 - h_1, x_2) - 2u(x_1, x_2) + u(x_1 + h_1, x_2)}{h_1^2} = u_{x_1 x_1},$$

$$\frac{\partial^2 u}{\partial x_2^2} \sim \frac{u(x_1, x_2 - h_2) - 2u(x_1, x_2) + u(x_1, x_2 + h_2)}{h_2^2} = u_{x_2 x_2},$$

знак  $\sim$  означает аппроксимацию. Пользуясь этими выражениями, заменим (1) разностным уравнением

$$\begin{aligned} & \frac{y(i_1 - 1, i_2) - 2y(i_1, i_2) + y(i_1 + 1, i_2)}{h_1^2} + \\ & + \frac{y(i_1, i_2 - 1) - 2y(i_1, i_2) + y(i_1, i_2 + 1)}{h_2^2} = -f(i_1, i_2), \quad (3) \end{aligned}$$

или, в сокращенной записи,

$$y_{\bar{x}_1 x_1}(i_1, i_2) + y_{\bar{x}_2 x_2}(i_1, i_2) = -f(i_1, i_2).$$

В безындексных обозначениях имеем

$$y_{\bar{x}_1 x_1}(x) + y_{\bar{x}_2 x_2}(x) = -f(x), \quad x = (i_1 h_1, i_2 h_2) \in \omega_h(G). \quad (4)$$

К этому уравнению надо присоединить краевые условия

$$y = \mu(x), \quad x = (i_1 h_1, i_2 h_2) \in \gamma_h. \quad (5)$$

Граница  $\gamma_h$  сетки состоит из всех узлов  $(0, i_2), (N_1, i_2), (i_1, 0), (i_1, N_2)$ , кроме вершин прямоугольника  $(0, 0), (0, N_2), (N_1, 0), (N_1, N_2)$ , которые не используются. Разностное уравнение (3) записано на пятиточечном шаблоне

$$(i_1 - 1, i_2), (i_1 + 1, i_2), (i_1, i_2), (i_1, i_2 - 1), (i_1, i_2 + 1).$$

Схему (4) часто называют схемой *крест*. Если  $h_1 = h_2 = h$ , т. е. сетки по  $x_1$  и  $x_2$  совпадают, то сетку  $\omega_h$  называют *квадратной*. На такой сетке разностную схему (4) можно записать в виде

$$\begin{aligned} y(i_1, i_2) = & \\ = & \frac{y(i_1 - 1, i_2) + y(i_1 + 1, i_2) + y(i_1, i_2 - 1) + y(i_1, i_2 + 1) + h^2 f(i_1, i_2)}{4}. \end{aligned}$$

Для однородного уравнения ( $f = 0$ ) получаем

$$\begin{aligned} y(i_1, i_2) = & \frac{1}{4} [y(i_1 - 1, i_2) + y(i_1 + 1, i_2) + \\ & + y(i_1, i_2 - 1) + y(i_1, i_2 + 1)], \end{aligned}$$

т. е. значение в центре шаблона определяется как среднее арифметическое значений в остальных узлах шаблона.

**3. Погрешность аппроксимации.** Пусть  $u = u(x)$  — решение задачи Дирихле (1), (2), а  $y = y(i_1, i_2)$  — решение разностной задачи (4), (5). Рассмотрим погрешность

$$z(x) = y(x) - u(x), \quad x = (i_1 h_1, i_2 h_2) \in \omega_h.$$

Подставляя  $y = z + u$  в (4), (5), получаем для погрешности  $z = z(x)$  неоднородное уравнение

$$\Lambda z = z_{\bar{x}_1 \bar{x}_1} + z_{\bar{x}_2 \bar{x}_2} = -\psi(x), \quad x \in \omega_h(G), \quad (6)$$

с однородным краевым условием

$$z = 0 \quad \text{при} \quad x \in \gamma_h. \quad (7)$$

Здесь

$$\psi(x) = \Lambda u + f(x) = u_{\bar{x}_1 \bar{x}_1} + u_{\bar{x}_2 \bar{x}_2} + f(x) \quad (8)$$

есть невязка или погрешность аппроксимации для схемы (4) на решении  $u = u(x)$  уравнения (1).

Покажем, что

$$|\psi| \leq M_4 \frac{h_1^2 + h_2^2}{24}, \quad (9)$$

где

$$M_4 = \max_{x \in G} \left( \left| \frac{\partial^4 u}{\partial x_1^4} \right|, \left| \frac{\partial^4 u}{\partial x_2^4} \right| \right).$$

В самом деле, учитывая формулы

$$\begin{aligned} u(x_1 \pm h_1, x_2) &= u(x_1, x_2) \pm h_1 \frac{\partial u}{\partial x_1}(x_1, x_2) + \\ &\quad + \frac{h_1^2}{2} \frac{\partial^2 u}{\partial x_1^2}(x_1, x_2) \pm \frac{h_1^3}{6} \frac{\partial^3 u}{\partial x_1^3}(x_1, x_2) + \\ &\quad + \frac{h_1^4}{24} \frac{\partial^4 u}{\partial x_1^4}(x_1, x_2), \quad \bar{x}_1 = x_1 + \theta_1 h_1, \quad 0 \leq \theta_1 \leq 1, \\ u(x_1, x_2 \pm h_2) &= u(x_1, x_2) \pm h_2 \frac{\partial u}{\partial x_2}(x_1, x_2) + \frac{h_2^2}{2} \frac{\partial^2 u}{\partial x_2^2}(x_1, x_2) \pm \\ &\quad \pm \frac{h_2^3}{6} \frac{\partial^3 u}{\partial x_2^3}(x_1, x_2) + \frac{h_2^4}{24} \frac{\partial^4 u}{\partial x_2^4}(x_1, x_2), \quad \bar{x}_2 = x_2 + \theta_2 h_2, \\ &\quad 0 \leq \theta_2 \leq 1, \end{aligned}$$

находим

$$\psi = \left( \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} - f(x) \right) + \frac{h_1^2}{24} \frac{\partial^4 u}{\partial x_1^4}(\bar{x}_1, x_2) + \frac{h_2^2}{24} \frac{\partial^4 u}{\partial x_2^4}(x_1, \bar{x}_2).$$

Отсюда и из (1) следует (9).

Таким образом, схема (4) имеет второй порядок аппроксимации.

**4. Схема повышенного порядка точности.** Используя девятиточечный шаблон  $(x_1, x_2), (x_1 \pm h_1, x_2), (x_1, x_2 \pm h_2), (x_1 \pm h_1, x_2 \pm h_2)$ , можно построить схему, имеющую четвертый порядок аппроксимации (и точности), если предположить, что решение задачи (1)–(2)  $u = u(x) \in C^{(6)}(\bar{G})$ . Эта схема имеет вид

$$\begin{aligned} \Lambda' y &= \left( \Lambda_1 + \Lambda_2 + \frac{h_1^2 + h_2^2}{12} \Lambda_1 \Lambda_2 \right) y = -\varphi(x), \quad x \in \omega_h, \\ y(x) &= \mu(x), \quad x \in \gamma_h, \\ \Lambda_1 y &= y_{\bar{x}_1 x_1}, \quad \Lambda_2 y = y_{\bar{x}_2 x_2}, \\ \varphi &= f + \frac{h_1^2}{12} \Lambda_1 f + \frac{h_2^2}{12} \Lambda_2 f. \end{aligned} \quad (10)$$

Непосредственная проверка показывает, что невязка равна

$$\psi = \Lambda' u + \varphi = O(|h|^4). \quad (11)$$

Для погрешности  $z = y - u$ , где  $y$  — решение задачи (10), получаем

$$\Lambda' z = -\psi(x), \quad x \in \omega_h; \quad z = 0, \quad x \in \gamma_h. \quad (12)$$

**5. Свойства разностного оператора.** Пусть  $\overset{\circ}{y}(x)$  — сеточная функция, заданная на сетке  $\overset{\circ}{\omega}_h = \overset{\circ}{\omega}_h(\bar{G})$  и равная нулю на границе  $\overset{\circ}{\gamma}_h$  сетки, и пусть  $\overset{\circ}{\Omega}$  — множество сеточных функций  $\overset{\circ}{y}$ .

Определим оператор  $A$  следующим образом:

$$Ay = -\Lambda \overset{\circ}{y} = -\overset{\circ}{y}_{\bar{x}_1 x_1} - \overset{\circ}{y}_{\bar{x}_2 x_2} \text{ для всех } y \in \overset{\circ}{\Omega}, \quad (13)$$

где  $\overset{\circ}{\Omega}$  — пространство сеточных функций, заданных во внутренних узлах сетки  $\overset{\circ}{\omega}_h$  и совпадающих там с  $y$ ,  $y(x) = \overset{\circ}{y}(x)$  при  $x \in \overset{\circ}{\omega}_h$ . Обозначая

$$\begin{aligned} \varphi &= f + \frac{\mu(l_1, x_2)}{h_1^2} \text{ при } x_1 = l_1 - h_1, \quad 0 < x_2 < l_2, \\ \varphi &= f + \frac{\mu(0, x_2)}{h_1^2}, \quad x_1 = h_1, \quad 0 < x_2 < l_2, \end{aligned}$$

$$\varphi = f + \frac{\mu(x_1, l_2)}{h_2^2}, \quad 0 < x_1 < l_1, \quad x_2 = l_2 - h_2,$$

$$\varphi = f + \frac{\mu(x_1, 0)}{h_2^2}, \quad 0 < x_1 < l_1, \quad x_2 = h_2,$$

$\varphi(x) = f(x)$  в остальных точках  $x \in \omega_h$ , запишем разностную схему (4), (5) в операторном виде:

$$A\varphi = \varphi, \quad y, \quad \varphi \in H, \quad (14)$$

где  $H = \Omega$ .

Введем в  $H$  скалярное произведение

$$(y, v) = \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2-1} \overset{\circ}{y}(i_1, i_2) \overset{\circ}{v}(i_1, i_2) h_1 h_2$$

и покажем, что оператор  $A$  самосопряжен. Представим  $A$  в виде суммы  $A = A_1 + A_2$ , где  $A_1 y = -\overset{\circ}{y}_{x_1 x_1}$ ,  $A_2 y = -\overset{\circ}{y}_{x_2 x_2}$ , и покажем, что каждый из «одномерных» операторов  $A_1$  и  $A_2$  является самосопряженным. Достаточно показать это для оператора  $A_1$ . Рассмотрим скалярное произведение

$$(A_1 y, v) = - \sum_{i_2=1}^{N_2-1} h_2 \left( \sum_{i_1=1}^{N_1-1} \overset{\circ}{y}_{x_1 x_1}(i_1, i_2) \overset{\circ}{v}(i_1, i_2) h_1 \right). \quad (15)$$

Воспользуемся одномерной формулой Грина (гл. I, § 4):

$$\sum_{i_1=1}^{N_1-1} \overset{\circ}{y}_{x_1 x_1}(i_1, i_2) \overset{\circ}{v}(i_1, i_2) h_1 = \sum_{i_1=1}^{N_1-1} \overset{\circ}{y}(i_1, i_2) \overset{\circ}{v}_{x_1 x_1}(i_1, i_2) h_1.$$

Подставляя это выражение в (15), получаем

$$(A_1 y, v) = - \sum_{i_2=1}^{N_2-1} h_2 \left( \sum_{i_1=1}^{N_1-1} \overset{\circ}{y}(i_1, i_2) \overset{\circ}{v}_{x_1 x_1}(i_1, i_2) h_1 \right) = (y, A_1 v).$$

Аналогично убеждаемся в том, что  $A_2^* = A_2$ , и, следовательно,

$$\begin{aligned} (Ay, v) &= ((A_1 + A_2)y, v) = (A_1 y, v) + (A_2 v, y) = \\ &= (y, A_1 v) + (y, A_2 v) = (y, Av), \end{aligned}$$

т. е.  $A^* = A$ ,

Если воспользоваться первой разностной формулой Грина

$$\sum_{i_1=1}^{N_1-1} \overset{\circ}{y}_{x_1 x_1}(i_1, i_2) \overset{\circ}{y}(i_1, i_2) h_1 = - \sum_{i_1=1}^{N_1} (\overset{\circ}{y}_{x_1}(i_1, i_2))^2 h_1,$$

то получим

$$(A_1 y, y) = \sum_{i_2=1}^{N_2-1} h_2 \sum_{i_1=1}^{N_1} (\overset{\circ}{y}_{x_1}(i_1, i_2))^2 h_1 > 0,$$

и, аналогично,  $(A_2 y, y) > 0$ , так что  $A > 0$ , т. е.  $A$  — самосопряженный и положительно определенный оператор.

Нетрудно найти границы  $\delta$  и  $\Delta$  оператора  $A$ , т. е. числа, для которых выполнены неравенства  $\delta E \leq A \leq \Delta E$ , где  $E$  — единичный оператор. В самом деле, в § 4 гл. I показано, что

$$\delta_1 \sum_{i_1=1}^{N_1-1} (\overset{\circ}{y}(i_1, i_2))^2 h_1 \leq \sum_{i_1=1}^{N_1} (\overset{\circ}{y}_{x_1}(i_1, i_2))^2 h_1 \leq \Delta_1 \sum_{i_1=1}^{N_1-1} (\overset{\circ}{y}(i_1, i_2))^2 h_1,$$

где

$$\delta_1 = \frac{4}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1}, \quad \Delta_1 = \frac{4}{h_1^2} \cos^2 \frac{\pi h_1}{2l_1}.$$

Суммируя эти неравенства по  $i_2 = 1, 2, \dots, N_2 - 1$ , получим  $\delta_1(y, y) \leq (A_1 y, y) \leq \Delta_1(y, y)$ .

Аналогично находим  $\delta_2(y, y) \leq (A_2 y, y) \leq \Delta_2(y, y)$ , где

$$\delta_2 = \frac{4}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2}, \quad \Delta_2 = \frac{4}{h_2^2} \cos^2 \frac{\pi h_2}{2l_2}.$$

Отсюда следует

$$\delta \|y\|^2 \leq (Ay, y) \leq \Delta \|y\|^2, \quad (16)$$

где

$$\delta = \delta_1 + \delta_2 = \frac{4}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2}, \quad (17)$$

$$\Delta = \Delta_1 + \Delta_2 = \frac{4}{h_1^2} \cos^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \cos^2 \frac{\pi h_2}{2l_2}.$$

В квадрате ( $l_1 = l_2 = 1$ ) на квадратной сетке ( $h_1 = h_2 = h$ ) имеем

$$\delta = \frac{8}{h^2} \sin^2 \frac{\pi h}{2}, \quad \Delta = \frac{8}{h^2} \cos^2 \frac{\pi h}{2}, \quad \delta + \Delta = \frac{8}{h^2}. \quad (18)$$

**6. Разностная задача на собственные значения.** Рассмотрим задачу: найти такие значения параметра  $\lambda$  (собственные значения), при которых однородная задача

$$y_{x_1 x_1} + y_{x_2 x_2} + \lambda y = 0, \quad x \in \omega_h, \quad y = 0, \quad x \in \gamma_h \quad (19)$$

имеет нетривиальные решения (собственные функции). Воспользуемся методом разделения переменных и будем искать решение задачи (19) в виде произведения

$$y(x_1, x_2) = v(x_1)w(x_2) \neq 0 \quad (20)$$

функции  $v(x_1)$ , зависящей только от  $x_1$ , и функции  $w(x_2)$ , зависящей только от  $x_2$ . Подставив (20) в (19) и разделив на  $y = vw$ , получим

$$\frac{v_{x_1 x_1}}{v} = -\frac{w_{x_2 x_2}}{w} - \lambda, \quad (x_1, x_2) \in \omega_h. \quad (21)$$

Левая часть зависит только от  $x_1$ , а правая — только от  $x_2$ ; равенство (21) возможно только при условии

$$\frac{v_{x_1 x_1}}{v} = \lambda^{(1)}, \quad -\frac{w_{x_2 x_2}}{w} - \lambda = \lambda^{(1)},$$

где  $\lambda^{(1)} = \text{const}$ . Отсюда получаем две одномерные задачи на собственные значения для отрезков  $0 \leq i_1 h_1 \leq l_1$  и  $0 \leq i_2 h_2 \leq l_2$  соответственно:

$$v_{x_1 x_1} + \lambda^{(1)} v = 0, \quad 0 < x_1 = i_1 h_1 < l_1, \quad v = 0, \quad i_1 = 0, N_1, \quad (22)$$

$$w_{x_2 x_2} + \lambda^{(2)} w = 0, \quad 0 < x_2 = i_2 h_2 < l_2, \quad w = 0, \quad i_2 = 0, N_2, \quad (23)$$

где  $\lambda^{(2)} = \lambda - \lambda^{(1)}$ , или  $\lambda = \lambda^{(1)} + \lambda^{(2)}$ .

Обращаясь к п. 8 § 4 гл. I, выпишем решение задач (22), (23) в виде

$$\lambda_{k_1}^{(1)} = \frac{4}{h_1^2} \sin^2 \frac{\pi k_1 h_1}{2l_1},$$

$$v_{k_1}^{(1)}(x_1) = \sqrt{\frac{2}{l_1}} \sin \frac{\pi k_1 x_1}{l_1}, \quad k_1 = 1, 2, \dots, N_1 - 1,$$

$$\lambda_{k_2}^{(2)} = \frac{4}{h_2^2} \sin^2 \frac{\pi k_2 h_2}{2l_2},$$

$$w_{k_2}^{(2)}(x_2) = \sqrt{\frac{2}{l_2}} \sin \frac{\pi k_2 x_2}{l_2}, \quad k_2 = 1, 2, \dots, N_2 - 1,$$

где  $x_\alpha = i_\alpha h_\alpha$ ,  $i_\alpha = 0, 1, \dots, N_\alpha$ ,  $\alpha = 1, 2$ .

Отсюда следует, что задача (19) имеет собственные значения

$$\lambda_{k_1 k_2} = \frac{4}{h_1^2} \sin^2 \frac{\pi k_2 h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi k_2 h_2}{2l_2},$$

$$k_\alpha = 1, 2, \dots, N_\alpha - 1, \alpha = 1, 2, \quad (24)$$

и соответствующие собственные функции  $y_k = v_{k_1}^{(1)}(x_1) w_{k_2}^{(2)}(x_2)$ :

$$y_k = y_{k_1 k_2}(x_1, x_2) = \sqrt{\frac{4}{l_1 l_2}} \sin \frac{\pi k_1 x_1}{l_1} \sin \frac{\pi k_2 x_2}{l_2},$$

$$x_\alpha = i_\alpha h_\alpha, \quad i_\alpha = 0, 1, \dots, N_\alpha, \quad k_\alpha = 1, 2, \dots, N_\alpha - 1,$$

$$\alpha = 1, 2. \quad (25)$$

Эти собственные функции ортонормированы:

$$(y_{k_1 k_2}, y_{m_1 m_2}) = \delta_{k_1 m_1} \delta_{k_2 m_2}.$$

Из (17) и (25) видно, что

$$\delta = \min \lambda_{k_1 k_2} = \lambda_{1,1}, \quad \Delta = \max \lambda_{k_1 k_2} = \lambda_{N_1-1, N_2-1},$$

где  $\delta$  и  $\Delta$  определяются по формулам (17). Для  $\delta$  и  $\Delta$  верны оценки

$$\delta \geqslant 8 \left( \frac{1}{l_1^2} + \frac{1}{l_2^2} \right), \quad \Delta < \frac{4}{h_1^2} + \frac{4}{h_2^2}. \quad (26)$$

**7. Оценка скорости сходимости схемы «крест».** Принцип максимума. Для погрешности  $z = y - u$  схемы в п. 3 получена задача (6), (7), где

$$\psi(x) = O(|h|^2), \quad |h|^2 = h_1^2 + h_2^2 \quad (27)$$

в предположении достаточной гладкости решения  $u = u(x) \in C^{(4)}(\bar{G})$  исходной задачи (1), (2). Докажем, что схема (4) сходится со скоростью  $O(|h|^2)$  (имеет второй порядок точности) в сеточной норме  $C$ , т. е.  $\|z\|_C = O(|h|^2)$ , где  $\|z\|_C = \max_{x \in \omega_h} |z(x)|$ . Для этого нам понадобится оценка решения задачи (6), (7) через правую часть  $\psi$ . Разност-

ная задача Дирихле (4) является частным случаем задачи

$$\begin{aligned} \mathcal{L}[y] = & a_{i_1 i_2} y_{i_1 i_2} - b_{i_1-1, i_2} y_{i_1-1, i_2} - b_{i_1+1, i_2} y_{i_1+1, i_2} - \\ & - b_{i_1, i_2-1} y_{i_1, i_2-1} - b_{i_1, i_2+1} y_{i_1, i_2+1} = \varphi_{i_1 i_2}, \\ x = (i_1 h_1, i_2 h_2) \in & \omega_h; \quad y = \mu, \quad x \in \gamma_h, \quad (28) \end{aligned}$$

где  $a = a_{i_1 i_2}$ ,  $b = b_{i_1, i_2}$  — коэффициенты. В случае (4) имеем

$$\begin{aligned} a_{i_1 i_2} &= 2 \left( \frac{1}{h_1^2} + \frac{1}{h_2^2} \right), \\ b_{i_1 \pm 1, i_2} &= \frac{1}{h_1^2}, \quad b_{i_1, i_2 \pm 1} = \frac{1}{h_2^2}, \\ b_{i_1, i_2} &= 0. \end{aligned} \quad (29)$$

Оператор  $\mathcal{L}[y]$  можно записать иначе:

$$\begin{aligned} \mathcal{L}[y] = & d_{i_1 i_2} y_{i_1 i_2} + b_{i_1-1, i_2} (y_{i_1, i_2} - y_{i_1-1, i_2}) + \\ & + b_{i_1+1, i_2} (y_{i_1, i_2} - y_{i_1+1, i_2}) + b_{i_1, i_2-1} (y_{i_1, i_2} - y_{i_1, i_2-1}) + \\ & + b_{i_1, i_2+1} (y_{i_1, i_2} - y_{i_1, i_2+1}), \quad (30) \end{aligned}$$

где  $d_{i_1 i_2} = a_{i_1 i_2} - b_{i_1-1, i_2} - b_{i_1+1, i_2} - b_{i_1, i_2-1} - b_{i_1, i_2+1}$ .

Будем предполагать, что выполнены условия

$$d = d_{i_1 i_2} \geq 0, \quad b_{i_1 \pm 1, i_2} > 0, \quad b_{i_1, i_2 \pm 1} \geq 0. \quad (31)$$

Для задачи (4) имеем  $d = 0$ .

**Теорема 1.** Пусть выполнены условия (31) и  $\varphi(x) \geq 0$ ,  $y|_{\gamma} \geq 0$ . Тогда решение уравнения (28) неотрицательно, т. е.  $y(x) \geq 0$  во всех узлах сетки  $\omega_h = \omega_h(\bar{G})$ .

**Доказательство.** Предположим, что утверждение теоремы неверно и существует по крайней мере один узел  $x_{i_0} = (i_1^0 h_1, i_2^0 h_2)$ , в котором  $y(x_{i_0}) < 0$ . Тогда функция  $y(x)$  в некотором внутреннем узле сетки должна принимать наименьшее отрицательное значение  $\min_{x \in \omega_h} y(x) = y(x_*)$ .

В этом узле выполняется уравнение (28). Если  $d(x_*) = 0$  и  $\varphi(x_*) = 0$ , то уравнение (28) выполнено только при условии  $y(x) = y(x_*)$  во всех узлах шаблона. Однако, так как  $\varphi(x) \neq 0$ , то существует узел  $x_{i_{**}}$ , в котором  $y(x_{i_{**}}) = y(x_*) = \min y(x) = c_0 < 0$  и по крайней мере в одном узле, например при  $x = x_{i_1+1}$ , имеем  $y_{i_1+1} > c_0$ , и, сле-

довательно,  $\mathcal{L}[y]|_{x=x_{i**}} < 0$ , что противоречит условию  $\mathcal{L}[y] = \varphi(x) \geq 0$ . Полученное противоречие и доказывает теорему.

**Теорема 2 (теорема сравнения).** *Пусть  $\bar{y}(x)$  — решение задачи*

$$\mathcal{L}[\bar{y}] = \bar{\varphi}, \quad x \in \omega_h, \quad \bar{y} = \bar{\mu}, \quad x \in \gamma_h \quad (32)$$

*и выполнены условия (31). Если*

$$|\varphi(x)| \leq \bar{\varphi}(x), \quad x \in \omega_h, \quad |\mu(x)| \leq \bar{\mu}(x), \quad x \in \gamma_h, \quad (33)$$

*то для решения задачи (28) верна оценка*

$$|y(x)| \leq \bar{y}(x) \quad \text{для всех } x \in \bar{\omega}_h.$$

Достаточно убедиться, что для функций  $u = \bar{y}(x) + y(x)$ ,  $v = \bar{y}(x) - y(x)$  выполнены условия теоремы 1, и, следовательно,  $u(x) \geq 0$ ,  $v(x) \geq 0$ , или  $y(x) \geq -\bar{y}(x)$ ,  $y(x) \leq \bar{y}(x)$ , т. е.  $|y| \leq \bar{y}$ .

Итак, функция  $\bar{y}(x)$  является мажорантой. Если мажоранта  $\bar{y}(x)$  найдена, то решение задачи (28) оценивается согласно теореме 2. Для задачи (4) в качестве мажоранты выберем функцию

$$\bar{y}(x) = C [L^2 - (x_1^2 + x_2^2)], \quad L^2 = l_1^2 + l_2^2. \quad (34)$$

Вычислим сначала  $\bar{\varphi} = \mathcal{L}[\bar{y}] = -\Lambda \bar{y} = C \Lambda (x_1^2 + x_2^2) = C (\Lambda_1 x_1^2 + \Lambda_2 x_2^2) = 4C$ , так как  $(x_1^2)_{x_1 x_1} = \frac{1}{h_1^2} ((x_1 + h_1)^2 - 2x_1^2 + (x_1 - h_1)^2) = 2$ . Из формулы (34) видно, что  $\bar{\mu} = \bar{y}(x) > 0$  на границе  $\gamma_h$ . Обратимся теперь к задаче (6), (7) для погрешности  $z = y - u$  схемы (4). Выбирая  $4C = |\psi|_c$  и учитывая, что  $z|_{\gamma_h} = 0$ , получаем  $|z(x)| < \bar{y}(x) < CL^2$ , так что

$$\|z\|_C \leq \frac{L^2}{4} \|\psi\|_C. \quad (35)$$

Отсюда и из (9) следует равномерная сходимость схемы (4) со вторым порядком точности.

**Замечание.** Уравнение (28) можно заменить уравнением более общего вида

$$\mathcal{L}[y] = a(x)y(x) - \sum_{\substack{\xi=\sigma(x) \\ \xi \neq x}} b(x, \xi)y(\xi) = \varphi(x),$$

где  $a(x) > 0$ ,  $b(x, \xi) > 0$ ,  $\sigma(x)$  — множество узлов  $\xi \neq x$  шаблона с центром в узле  $x$ , причем

$$d(x) = a(x) - \sum_{\xi \in \sigma(x)} b(x, \xi) \geqslant 0.$$

Для уравнения (36) верны теоремы 1 и 2. В случае схемы повышенного порядка точности шаблон состоит из девяти узлов, множество  $\sigma(x)$  — из восьми узлов, причем  $a = \frac{5}{3} (h_1^{-2} + h_2^{-2})$ , а в правой части имеются коэффициенты  $\frac{1}{6} (5h_1^{-2} - h_2^{-2})$ ,  $\frac{1}{6} (5h_2^{-2} - h_1^{-2})$ , которые положительно только при условии

$$1/\sqrt{5} \leq h_1/h_2 \leq \sqrt{5},$$

и, следовательно, оценка вида (35) будет получаться при этом условии.

## § 2. Решение разностных уравнений

**1. Прямые методы. Метод разделения переменных.** Система разностных уравнений для задачи Дирихле из § 1;

$$\Lambda y = \overset{\circ}{y}_{x_1 x_1} + \overset{\circ}{y}_{x_2 x_2} = -f(x), \quad x \in \omega_h, \quad y = \mu, \quad x \in \gamma_n \quad (1)$$

имеет матрицу высокого порядка  $(N_1 - 1)(N_2 - 1)$ . Обычно берут  $N_1, N_2 \sim 50 - 100$ , так что число уравнений в системе (1) равно  $10^3 - 10^4$ . Решение систем столь высокого порядка методом Гаусса потребовало бы числа действий порядка  $(N_1 - 1)^3(N_2 - 1)^3$ , т. е.  $10^9 - 10^{12}$  действий, если бы у системы (1) не было одного хорошего качества: матрица системы является слабо заполненной и имеет лишь  $\sim 5N_1N_2$  отличных от нуля элементов. Поэтому для решения системы разностных уравнений удается построить методы, требующие  $O(N \ln N)$  и даже  $O(N)$  действий, где  $N = (N_1 - 1)(N_2 - 1)$ . Опишем один из прямых методов решения разностной задачи Дирихле уравнения Пуассона в прямоугольнике.

Перепишем задачу (1) в виде

$$\overset{\circ}{\Lambda} y = \overset{\circ}{y}_{x_1 x_1} + \overset{\circ}{y}_{x_2 x_2} = -\varphi(x), \quad x \in \omega_h, \quad \overset{\circ}{y}|_{\gamma_h} = 0, \quad (2)$$

где  $\overset{\circ}{y}(x) = y(x)$  при  $x \in \omega_h$ , а  $\varphi(x)$  определяется по формулам (14) из § 1.

Ее решение можно найти методом разделения переменных. Пусть  $\{v_{k_2}^{(2)}(x_2), \lambda_{k_2}^{(2)}\}$  ( $k = 1, 2, \dots, N_2 - 1$ ) — собственные функции и собственные значения задачи

$$\Lambda_2 v + \lambda v = 0, \quad x \in \omega_k; \quad v(0) = v(l_2) = 0. \quad (3)$$

Выражения для  $\lambda_{k_2}^{(2)}$  и  $v_{k_2}^{(2)}(x_2)$  даны в п. 6, § 1.

Разложим решение  $\overset{\circ}{y}(x_1, x_2)$  и правую часть  $\varphi(x_1, x_2)$  по собственным функциям  $\{v_{k_2}^{(2)}\}$ :

$$\overset{\circ}{y}(x_1, x_2) = \sum_{k_2=1}^{N_2-1} c_{k_2}(x_1) v_{k_2}(x_2), \quad (4)$$

$$\varphi(x_1, x_2) = \sum_{k_2=1}^{N_2-1} \varphi_{k_2}(x_1) v_{k_2}(x_2), \quad (5)$$

где  $x_\alpha = i_\alpha h_\alpha$ ,  $i_\alpha = 1, 2, \dots, N_\alpha - 1$ ,  $\alpha = 1, 2$ ,  $c_{k_2}(x_1)$  и  $\varphi_{k_2}(x_1)$  — коэффициенты Фурье, например,

$$\varphi_{k_2}(x_1) = \sum_{i_2=1}^{N_2-1} h_2 \varphi(x_1, i_2 h_2) v_{k_2}(i_2 h_2).$$

Применим оператор  $\Lambda = \Lambda_1 + \Lambda_2$  к произведению  $c_{k_2} v_{k_2}$ :

$$\begin{aligned} \Lambda c_{k_2}(x_1) v_{k_2}(x_2) &= \\ &= v_{k_2}(x_2) \Lambda_1 c_{k_2}(x_1) + c_{k_2}(x_1) \Lambda_2 v_{k_2}(x_2) = \\ &= v_{k_2}(x_2) \Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1) v_{k_2}(x_2) = \\ &= [\Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1)] v_{k_2}(x_2). \end{aligned}$$

Подставляя затем это выражение в (2) и учитывая (5), получим

$$\sum_{k_2=1}^{N_2-1} [\Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1) + \varphi_{k_2}(x_1)] v_{k_2}(x_2) = 0. \quad (6)$$

В силу ортогональности  $\{v_{k_2}(x_2)\}$  это тождество возможно только при равенстве нулю выражения в фигурных скобках:

$$\begin{aligned} \Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1) &= -\varphi_{k_2}(x_1), \quad k_2 = 1, 2, \dots, N_2 - 1, \\ x_1 = i_1 h_1, \quad 0 < i_1 < N_1, \quad c_{k_2}(i_1 h_1) &= 0, \quad i_1 = 0, N_1. \end{aligned} \quad (7)$$

В самом деле, умножая (6) скалярно на  $v_{k_2}(x_2)$ , имеем

$$0 = \sum_{k=1}^{N_2-1} \{\cdot\}_k (v_k, v_{k_2}) = \sum_{k=1}^{N_2-1} \{\cdot\}_k \delta_{kk_2} = \{\cdot\}_{k_2} = 0,$$

где  $\{\cdot\}_{k_2}$  — содержимое фигурной скобки (6).

Задачи (7) решаются методом прогонки; всего требуется  $N_2 - 1$  раз использовать алгоритм прогонки для  $k_2 = 1, 2, \dots, N_2 - 1$ . Зная  $c_{k_2}(x_1)$ , найдем по формуле (4) решение задачи (2). Для этого надо сначала вычислить коэффициенты Фурье  $\Phi_{k_2}(x_1)$  ( $k_2 = 1, 2, \dots, N_2 - 1$ ). Из формул (4) и (5) видно, что  $y(x_1, x_2)$  и  $\Phi_{k_2}(x_1)$  вычисляются по формулам одного и того же вида:

$$w_i = \sum_{k=1}^{N-1} \alpha_k \sin \frac{k\pi i}{N}, \quad i = 1, 2, \dots, N - 1. \quad (8)$$

Разработан специальный алгоритм быстрого преобразования Фурье для вычисления сумм, который позволяет вычислить сумму (8) за  $5N \log_2 N$  арифметических действий (при  $N = 2^n$ ,  $n$  — целое число) вместо  $O(N^2)$  при обычном способе суммирования. Этот алгоритм позволяет найти решение исходной задачи (2) за  $O(N_1 N_2 \log_2 N_2)$  действий. Метод разделения переменных можно комбинировать с методом редукции или декомпозиции, являющимся модификацией метода Гаусса. В результате получим алгоритм с числом действий  $Q \approx 5N_1 N_2 \log_2 N_2$ , что в два раза меньше, чем для алгоритма разделения, приведенного выше.

**2. Итерационные методы.** Для решения разностной задачи Дирихле для уравнения Пуассона в прямоугольнике наиболее экономичными являются прямые методы. В настоящее время имеются стандартные программы на алгоритмических языках фортран и алгол для решения уравнений Пуассона в прямоугольнике с краевыми условиями трех типов, а также со смешанными краевыми условиями. Однако в случае, когда область не является прямоугольником или рассматриваются уравнения с переменными коэффициентами, применяются итерационные методы. Фактически прямые методы экономичны лишь в случае, когда переменные разделяются.

В гл. III рассматривалась теория итерационных методов для уравнения

$$Ay = \Phi,$$

где  $A = A^* > 0$ . Сравнение различных методов проводилось для модельной одномерной задачи на отрезке  $0 \leq x \leq 1$ :

$$y_{xx} = -f(x), \quad x = ih, \quad 0 < i < N, \quad y_0 = y_N = 0.$$

Для нее оператор  $A$  имеет вид  $Ay = -\overset{\circ}{y}_{xx}$ . Границы оператора  $A$  определяются постоянными

$$\delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}, \quad \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}.$$

Число итераций для рассмотренных в гл. III методов зависит от отношения

$$\eta = \frac{\delta}{\Delta} = \operatorname{tg}^2 \frac{\pi h}{2} \approx \frac{\pi^2 h^2}{4}. \quad (9)$$

Рассмотрим теперь в качестве модельной двумерную задачу Дирихле в единичном квадрате ( $l_1 = l_2 = 1$ ) на квадратной сетке с шагом  $h = h_1 = h_2$ :

$$Ay = -\overset{\circ}{y}_{x_1 x_1} - \overset{\circ}{y}_{x_2 x_2} = \varphi, \quad \varphi, y \in H. \quad (10)$$

Число интервалов по каждому из направлений равно  $N$ , так что  $h = 1/N$ .

Границы  $\delta$  и  $\Delta$  оператора  $A$  найдены в § 1 (см. (18) из § 1), отношение  $\eta = \delta/\Delta$  совпадает с (9). Отсюда следует, что число итераций не зависит от числа измерений (если  $h_1 \neq h_2$ ,  $l_1 \neq l_2$ , то слабо зависит). Поэтому те оценки числа итераций различных итерационных методов, которые мы получили для одномерной модельной задачи, справедливы и для двумерного случая.

В случае неквадратной сетки число итераций для двумерной задачи может несколько отличаться от числа итераций для одномерной задачи.

Мы рассмотрим здесь лишь попеременно-треугольный итерационный метод для решения разностной задачи Дирихле (10).

**3. Попеременно-треугольный метод.** Для решения операторного уравнения

$$Au = f, \quad A = A^* > 0, \quad A: H \rightarrow H, \quad (11)$$

в гл. III рассматривались двуслойные одношаговые итерационные методы, которые записывались в следующей

канонической форме:

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, n,$$

для всех  $y_0 \in H$ , (12)

где  $B: H \rightarrow H$ ,  $B = B^* > 0$ . Для  $A$  и  $B$  выполнены условия

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0, \quad (13)$$

где  $\gamma_1, \gamma_2$  — постоянные.

Минимальное число итераций  $\min_{\{\tau_k\}} n(\varepsilon)$  при заданных  $\gamma_1, \gamma_2$  достигается при выборе чебышевских параметров

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \sigma_k}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

$$k = 1, 2, \dots, n, \quad (14)$$

где  $\sigma_k$  принадлежит некоторому специально упорядоченному множеству пулей полинома Чебышева; при таком упорядочении метод (12) является вычислительно устойчивым.

Для определения  $(k+1)$ -й итерации имеем уравнение

$$By_{k+1} = F_k, \quad F_k = By_k - \tau_{k+1}(Ay_k - f).$$

Число действий при вычислении  $y_{k+1}$  зависит от  $B$ . Выбирая

$$B = (D + \omega A_1)D^{-1}(D + \omega A_2), \quad (15)$$

где  $A_1$  и  $A_2$  — операторы с треугольными матрицами  $A_1^* = A_2$ ,  $A_1 + A_2 = A$ , а  $D = D^* > 0$  — произвольный оператор, получаем попаременно-треугольный метод. Обычно  $D = (d_{i_1 i_2})$  — диагональная матрица. В гл. III дана теория этого метода и найдены постоянные  $\gamma_1, \gamma_2$  и  $\omega$  при заданных условиях

$$A \geq \delta D, \quad A_1 D^{-1} A_2 \leq \frac{\Delta}{4} A, \quad \delta > 0, \quad \Delta \geq \delta > 0, \quad (16)$$

которые можно записать в эквивалентном виде:

$$(Ay, y) \geq \delta (Dy, y), \quad (D^{-1} A_2 y, A_2 y) \leq \frac{\Delta}{4} (Ay, y).$$

В этом случае имеем

$$\omega = \frac{2}{\sqrt{\delta \Delta}}, \quad \xi = \frac{2 \sqrt{\eta}}{1 + \sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}, \quad (17)$$

а для числа итераций верна оценка

$$n(\varepsilon) \approx n_0(\varepsilon) = \frac{1}{2\sqrt{2}\sqrt[4]{\eta}} \ln \frac{2}{3}. \quad (18)$$

**4. Попеременно-треугольный метод для разностной задачи Дирихле.** Обратимся к задаче (10). Оператор  $A$  представим в виде суммы  $A = A_1 + A_2$ , где

$$A_1 y = \frac{y_{x_1}}{h_1} + \frac{y_{x_2}}{h_2}, \quad A_2 y = -\frac{y_{x_1}}{h_1} - \frac{y_{x_2}}{h_2},$$

и положим  $D = E$ . Сопряженность  $A_1$  и  $A_2$ :  $A_2 = A_1^*$  устанавливается сравнением их матриц или с помощью первой разностной формулы Грина:  $(A_1 y, v) = (y, A_1^* v) = (y, A_2 v)$ .

Для определения  $y_{k+1}$  получаем уравнение

$$\begin{aligned} By_{k+1} &= (E + \omega A_1)(E + \omega A_2)y_{k+1} = F_k, \\ F_k &= \overset{\circ}{By}_k + \tau_{k+1}(\Lambda y_k + \varphi) \quad (y_k = \mu, \overset{\circ}{y}_k = 0 \text{ при } x \in \gamma_h). \end{aligned}$$

Значения  $y_{k+1}$  находятся последовательно из уравнения

$$(E + \omega A_1) \overset{\circ}{y}_k^{(1)} = F_k, \quad (E + \omega A_2) \overset{\circ}{y}_{k+1} = \overset{\circ}{y}_k^{(1)}.$$

Отсюда получаем формулы

$$\begin{aligned} \overset{\circ}{y}_k^{(1)}(i_1, i_2) &= \left[ \frac{\chi_1 \overset{\circ}{y}_k^{(1)}(i_1-1, i_2) + \chi_2 \overset{\circ}{y}_k^{(1)}(i_1, i_2-1) + F_k(i_1, i_2)}{(1 + \chi_1 + \chi_2)} \right], \\ \chi_1 &= \frac{\omega}{h_1^2}, \quad \chi_2 = \frac{\omega}{h_2^2}, \\ \overset{\circ}{y}_{k+1}(i_1, i_2) &= \\ &= \left[ \frac{\chi_1 \overset{\circ}{y}_{k+1}(i_1+1, i_2) + \chi_2 \overset{\circ}{y}_{k+1}(i_1, i_2+1) + \overset{\circ}{y}_k^{(1)}(i_1, i_2)}{(1 + \chi_1 + \chi_2)} \right]. \quad (19) \end{aligned}$$

Чтобы определить  $\overset{\circ}{y}_k^{(1)}(i_1, i_2)$ , выбираем узел  $i_1 = 1, i_2 = 1$  в левом углу прямоугольника; тогда остальные два узла  $(i_1-1, i_2)$  и  $(i_1, i_2-1)$  шаблона  $\{(i_1, i_2), (i_1-1, i_2), (i_1, i_2-1)\}$  лежат на границе и, следовательно,  $\overset{\circ}{y}^{(1)}(i_1-1, i_2) = \overset{\circ}{y}^{(1)}(i_1, i_2-1) = 0$  известны. Зная  $\overset{\circ}{y}_k^{(1)}$  при  $i_1 = 1, i_2 = 1$ , последовательно находим  $\overset{\circ}{y}_k^{(1)}$  при  $i_1 = 2, 3, \dots, N_1 - 1$  и  $i_2 = 1$  (на первой строке). Далее, полагаем

$i_2 = 2$  и находим последовательно  $\overset{\circ}{y}_k^{(1)}$  па второй строке при  $i_1 = 1, 2, \dots, N - 1$ . Для определения  $y_{k+1}$  проводим вычисления на шаблоне  $\{(i_1, i_2), (i_1 + 1, i_2), (i_1, i_2 + 1)\}$  по столбцам сверху вниз: фиксируем  $i_1 = N_1 - 1, N_1 - 2, \dots, 2, 1$ , и при каждом  $i_1$  меняем  $i_2 = N_2 - 1, N_2 - 2, \dots, 2, 1$ . Начинаем счет  $\overset{\circ}{y}_{k+1}$  с узла  $(i_1 = N_1 - 1, i_2 = N_2 - 1)$  в верхнем правом углу. Следует отметить, что счет  $y_{k+1}$  можно также вести по строкам справа налево: фиксируем  $i_2 = N_2 - 1, N_2 - 2, \dots, 2, 1$  и при каждом  $i_2$  меняем  $i_1 = N_1 - 1, N_1 - 2, \dots, 2, 1$ . Впрочем, вычисление  $\overset{\circ}{y}_k^{(1)}$  можно вести не по строкам, а по столбцам снизу вверх. Это видно из самих формул.

Вычисления ведутся по рекуррентным формулам (19); счет, очевидно, устойчив. Алгоритм подобного типа, как уже отмечалось, называют *алгоритмом бегущего счета*.

Подсчитаем число арифметических действий на один узел сетки: вычисление  $F_k$  требует 10 операций сложения и 10 операций умножения; вычисление  $y_{k+1}$  при заданном  $F_k$  требует 4 операции сложения и 6 операций умножения.

Итого требуется для определения  $y_{k+1}$  в одном узле провести 14 операций сложения и 16 операций умножения. Число действий можно уменьшить, если хранить в оперативной памяти не одну, а две последовательности  $\{y_k\}$  и  $\{w_{k+1}\}$  и для определения  $y_{k+1}$  пользоваться алгоритмом

$$(E + \omega A_1) \overset{\circ}{w}_{k+1/2} = \Lambda y_k + f, \quad (E + \omega A_2) \overset{\circ}{w}_{k+1} = \overset{\circ}{w}_{k+1/2}, \\ y_{k+1} = y_k + \tau_{k+1} \overset{\circ}{w}_{k+1}.$$

В этом случае для перехода от  $y_k$  к  $y_{k+1}$  достаточно 10 операций сложений и 10 операций умножения на один узел.

**5. Выбор параметров попаременно-треугольного метода для разностной задачи Дирихле.** Чтобы воспользоваться общей теорией гл. III (см. § 5 гл. III), надо найти постоянные  $\delta$  и  $\Delta$ , входящие в условие (16). В нашем случае  $A = A_1 + A_2 \geq \delta E$ , где  $\delta$  — наименьшее собственное значение оператора  $A$ , равное

$$\delta = 4 \left( \frac{1}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{1}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2} \right). \quad (20)$$

Рассмотрим оператор  $A_1 D^{-1} A_2 = A_1 A_2$ . Учитывая, что

$$A_1^* = A_2, (a_1 b_1 + a_2 b_2)^2 \leq (a_1^2 + a_2^2)(b_1^2 + b_2^2),$$

находим

$$\begin{aligned} (A_1 A_2 y, y) &= (A_2 y, A_2 y) = \\ &= \left( \left( \frac{1}{h_1} y_{x_1} + \frac{1}{h_2} y_{x_2} \right)^2, 1 \right) \leq \left( \frac{1}{h_1^2} + \frac{1}{h_2^2} \right) ((y_{x_1})^2 + (y_{x_2})^2, 1) = \\ &= \left( \frac{1}{h_1^2} + \frac{1}{h_2^2} \right) \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2-1} [(y_{x_1})^2 + (y_{x_2})^2]_{i_1 i_2} h_1 h_2 \leq \\ &\leq \left( \frac{1}{h_1^2} + \frac{1}{h_2^2} \right) (Ay, y), \end{aligned}$$

так как (см. § 1 гл. V)

$$(Ay, y) = \sum_{i_2=1}^{N_2-1} h_2 \sum_{i_1=0}^{N_1-1} (y_{x_1})_{i_1 i_2}^2 h_1 + \sum_{i_1=1}^{N_1-1} h_1 \sum_{i_2=0}^{N_2-1} (y_{x_2})_{i_1 i_2}^2 h_2.$$

Сравнивая равенства

$$(A_1 A_2 y, y) \leq \left( \frac{1}{h_1^2} + \frac{1}{h_2^2} \right) (Ay, y) \text{ и } A_1 A_2 \leq \frac{\Delta}{4} A,$$

заключаем, что

$$\Delta = 4 \left( \frac{1}{h_1^2} + \frac{1}{h_2^2} \right). \quad (21)$$

Зная  $\delta$  и  $\Delta$ , находим  $\eta = \delta/\Delta$  и по формулам § 5 гл. V находим параметры  $\gamma_1, \gamma_2, \xi$ , после чего оцениваем число итераций по формуле

$$n(\varepsilon) \approx \ln \frac{2}{\varepsilon} / \ln \frac{1}{\rho_1}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}.$$

Пользуясь  $n(\varepsilon)$ , выбираем устойчивый набор чебышевских параметров  $\sigma_h, \tau_{k+1}$  и  $\omega = 2/\sqrt{\delta\Delta}$ .

Приведем результат сравнения методов решения по числу итераций  $n_0(\varepsilon)$ : метода простой итерации ( $n_0^{(1)}(\varepsilon)$ ), явной схемы с чебышевским набором ( $n_0^{(2)}(\varepsilon)$ ) и попаременно-треугольного метода ( $n_0^{(3)}(\varepsilon)$ ) для двумерной модельной задачи (10), пользуясь приближенными формулами  $n_0^{(1)}(\varepsilon) \approx 2/h^2, n_0^{(2)}(\varepsilon) \approx 3,2/h, n_0^{(3)}(\varepsilon) \approx 2,9/\sqrt{h}$  при  $\varepsilon = 10^{-4}$  (табл. 2).

Таблица 2

$h$	$n_0^{(1)}(\varepsilon)$	$n_0^{(2)}(\varepsilon)$	$n_0^{(3)}(\varepsilon)$
1/10	200	32	9
1/50	5 000	160	24
1/100	20 000	320	29

**6. Разностные уравнения с переменными коэффициентами.** Пусть требуется в прямоугольнике  $\bar{G} = \{(x_1, x_2) : 0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$  решить задачу Дирихле для эллиптического уравнения с переменными коэффициентами:

$$Lu = L_1 u + L_2 u = -f(x),$$

$$x = (x_1, x_2) \in G, u = \mu(x), \quad x \in \Gamma, \quad (22)$$

$$L_\alpha u = \frac{\partial}{\partial x_\alpha} \left( k_\alpha(x) \frac{\partial u}{\partial x_\alpha} \right), \quad 0 < c_1 \leq k_\alpha(x) \leq c_2, \alpha = 1, 2,$$

где  $c_1$  и  $c_2$  — постоянные. При  $k_1 = k_2 = 1$  получаем уравнение Пуассона  $\Delta u = -f$ .

Разностная схема строится на сетке  $\omega_h = \{x_i = (i_1 h_1, i_2 h_2) | i_\alpha = 0, 1, \dots, N_\alpha, h_\alpha = l_\alpha / N_\alpha, \alpha = 1, 2\}$ . Каждый оператор  $L_\alpha$  заменяется на трехточечном шаблоне  $(x_\alpha - h_\alpha, x_\alpha, x_\alpha + h_\alpha)$  разностным оператором:

$$\Lambda_\alpha u = (a_\alpha u_{\tilde{x}_\alpha})_{x_\alpha} = \frac{1}{h_\alpha} \left[ \frac{a_\alpha^{(+1)_\alpha} (u_\alpha^{(+1)_\alpha} - u)}{h_\alpha} - \frac{a_\alpha (u - u^{(-1)_\alpha})}{h_\alpha} \right],$$

где  $u^{(\pm 1)_1} = u((i_1 \pm 1)h_1, i_2 h_2)$ ,  $u^{(\pm 1)_2} = u(i_1 h_1, (i_2 \pm 1)h_2)$ . Для  $a_1$  и  $a_2$  можно выбрать простейшие выражения

$$a_1(x_1, x_2) = k_1(x_1 - 1/2h_1, x_2) = k_1^{(-1/2)_1},$$

$$a_2(x_1, x_2) = k_2(x_1, x_2 - 1/2h_2) = k_2^{(-1/2)_2},$$

обеспечивающие второй порядок аппроксимации:

$$\Lambda_\alpha u - L_\alpha u = O(h_\alpha^2).$$

В результате оператору  $Lu$  ставится в соответствие разностный оператор на пятиточечном шаблоне:

$$\Lambda u = \Lambda_1 u + \Lambda_2 u = (a_1 u_{\tilde{x}_1})_{x_1} + (a_2 u_{\tilde{x}_2})_{x_2}.$$

Напишем разностную схему

$$\begin{aligned} \Lambda y &= -f(x), \quad x \in \omega_h, \quad y = \mu(x), \quad x \in \gamma_h, \\ 0 < c_1 &\leq a_\alpha \leq c_2, \quad \alpha = 1, 2, \end{aligned} \quad (23)$$

соответствующую задаче (22).

Введем в пространстве сеточных функций  $H = \Omega_N$  оператор

$$\begin{aligned} Ay &= -\overset{\circ}{\Lambda}y, \quad A = A_1 + A_2, \\ A_1y &= -\overset{\circ}{\Lambda}_1y, \quad A_2y = -\overset{\circ}{\Lambda}_2y \end{aligned}$$

и запишем (23) в операторной форме:

$$Ay = \varphi, \quad y, \varphi \in H,$$

где  $\varphi$  отличается от  $f$  только в 4 приграничных узлах ( $i_1 = 1, N_1 - 1, 0 < i_2 < N_2$ ) и ( $0 < i_1 < N_1, i_2 = 1, N_2 - 1$ ).

Оператор  $A$ , очевидно, является самосопряженным:  $(Ay, v) = (y, Av)$ .

Из формулы

$$-\sum_{i_1=1}^{N_1-1} (a_1 \overset{\circ}{y}_{x_1})_{x_1, i_1} \overset{\circ}{y}_{i_1} h_1 = \sum_{i_1=1}^{N_1} (a_1 (\overset{\circ}{y}_{x_1})^2)_{i_1} h_1$$

и неравенства  $0 < c_1 \leq a \leq c_2$  следует, что

$$c_1(Ry, y) \leq (Ay, y) \leq c_2(Ry, y) \text{ или } c_1R \leq A \leq c_2R, \quad (24)$$

где  $R$  есть изученный выше оператор Лапласа

$$Ry = -\overset{\circ}{y}_{x_1 x_1} - \overset{\circ}{y}_{x_2 x_2}. \quad (25)$$

Отсюда заключаем, что

$$c_1 \overset{\circ}{\delta}E \leq A \leq c_2 \overset{\circ}{\Delta}E,$$

где  $\overset{\circ}{\delta}$  и  $\overset{\circ}{\Delta}$  определяются формулами (20), (21).

Для решения задачи (23) можно воспользоваться по-переменно-треугольным методом с оператором

$$B = (E + \omega R_1)(E + \omega R_2), \quad R_1 + R_2 = R, \quad R_1^* = R_2$$

при  $D = E$ .

В этом случае имеем  $\overset{\circ}{\gamma}_1 B \leq A \leq \overset{\circ}{\gamma}_2 B$ , где  $\overset{\circ}{\gamma}_1 = c_1 \overset{\circ}{\gamma}_1$ ,  $\overset{\circ}{\gamma}_2 = c_2 \overset{\circ}{\gamma}_2$ , а постоянные  $\overset{\circ}{\gamma}_1$  и  $\overset{\circ}{\gamma}_2$  найдены для оператора

(25). Для числа итераций имеем оценку

$$n_0(\varepsilon) \approx \sqrt{\frac{c_2}{c_1}} \tilde{n}_0(\varepsilon), \quad \tilde{n}_0(\varepsilon) = \frac{1}{2\sqrt{2}\sqrt[4]{\eta}} \ln \frac{2}{\varepsilon}.$$

Для уравнения с переменными коэффициентами требуется в  $\sqrt{c_2/c_1}$  раз больше итераций, чем для уравнения Пуассона.

Таблица 3

$\frac{c_2}{c_1}$	$h = 1/32$		$h = 1/128$	
	$D = E$	$D = d(x)E$	$D = E$	$D = d(x)E$
2	23	20	45	39
8	46	23	90	47
32	92	25	180	53
128	184	26	360	57
512	367	26	720	59

Однако можно не вводить оператор  $R$ , соответствующий оператору Лапласа, а сразу представить оператор с переменными коэффициентами в виде

$$\begin{aligned} A &= A_1 + A_2, \\ A_1 y &= \frac{1}{h_1} \left( a_{11} y_{x_1} + \frac{1}{2} y a_{x_1} \right) + \frac{1}{h_2} \left( a_{21} y_{x_2} + \frac{1}{2} y a_{x_2} \right), \\ A_2 y &= -\frac{1}{h_1} \left( a_{11}^{(+1_1)} y_{x_1} + \frac{1}{2} y a_{x_1} \right) - \frac{1}{h_2} \left( a_{21}^{(+1_2)} y_{x_2} + \frac{1}{2} y a_{x_2} \right). \end{aligned}$$

Оператор  $B$  выбирается в форме

$$B = (D + \omega A_1) D^{-1} (D + \omega A_2), \quad (26)$$

где  $D = d(x)E$  — диагональная матрица. Для применения общей теории надо найти постоянные  $\delta$  и  $\Delta$ , входящие в условия  $A \geq \delta D$ ,  $A_1 D^{-1} A_2 \leq \frac{\Delta}{4} A$ . Коэффициент  $d(x)$  выбирается из условия максимума отношения  $\eta = \delta/\Delta$ , и, следовательно, максимума  $\xi = \gamma_1/\gamma_2$ . В результате получается алгоритм, у которого число итераций  $n_0(\varepsilon)$  слабо зависит от отношения  $c_2/c_1$ . Об этом свидетельствует табл. 3.

## Глава VII

# РАЗНОСТНЫЕ МЕТОДЫ РЕШЕНИЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ

В этой главе рассмотрены разностные схемы для решения уравнения теплопроводности. Детально исследовано одномерное уравнение с постоянными коэффициентами. Приведены разностные схемы для многомерного уравнения теплопроводности с переменными коэффициентами.

## § 1. Уравнение теплопроводности с постоянными коэффициентами

**1. Исходная задача.** Процесс распространения тепла в одномерном стержне  $0 < x < l$  описывается уравнением теплопроводности

$$c\rho \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( k \frac{\partial u}{\partial x} \right) + f_0(x, t), \quad (1)$$

где  $u = u(x, t)$  — температура в точке  $x$  стержня в момент  $t$ ,  $c$  — теплоемкость единицы массы,  $\rho$  — плотность,  $c\rho$  — теплоемкость единицы длины,  $k$  — коэффициент теплопроводности,  $f_0$  — плотность тепловых источников. В общем случае  $k$ ,  $c$ ,  $\rho$ ,  $f_0$  могут зависеть не только от  $x$  и  $t$ , но и от температуры  $u = u(x, t)$  (квазилинейное уравнение теплопроводности) и даже от  $\partial u / \partial x$  (нелинейное уравнение). Если  $k$ ,  $c$ ,  $\rho$  постоянны, то (1) можно записать в виде

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} + f, \quad f = \frac{f_0}{c\rho}, \quad (2)$$

где  $a^2 = k/(c\rho)$  — коэффициент температуропроводности. Без ограничения общности можно считать  $a = 1$ ,  $l = 1$ .

В самом деле, вводя переменные  $x_1 = \frac{x}{l}$ ,  $t_1 = \frac{a^2 t}{l^2}$ ,  $f_1 = \frac{l^2}{a^2} f$ , получим

$$\frac{\partial u}{\partial t_1} = \frac{\partial^2 u}{\partial x_1^2} + f_1, \quad 0 < x_1 < 1.$$

Мы будем рассматривать первую краевую задачу (иногда говорят: начально-краевую задачу) в области  $\bar{D} = \{0 \leq x \leq 1, 0 \leq t \leq T\}$ . Требуется найти непрерывное в  $\bar{D}$  решение  $u = u(x, t)$  задачи

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < 1, \quad 0 < t \leq T,$$

$$\begin{aligned} u(x, 0) &= u_0(x), \quad 0 \leq x \leq 1, \quad u(0, t) = u_1(t), \\ u(1, t) &= u_2(t), \quad 0 \leq t \leq T. \end{aligned} \quad (3)$$

**2. Некоторые свойства решений уравнения теплопроводности.** В силу принципа максимума для решения задачи (3) имеет место оценка

$$\begin{aligned} \max_{0 \leq x \leq 1, 0 \leq t \leq T} |u(x, t)| &\leq \max \left( \max_{0 \leq x \leq 1} |u_0(x)|, \max_{0 \leq t \leq T} |u_1(t)|, \right. \\ &\quad \left. \max_{0 \leq t \leq T} |u_2(t)| \right) + \int_0^T \max_{0 \leq x \leq 1} |f(x, t)| dt. \end{aligned} \quad (4)$$

Рассмотрим однородное уравнение с однородными краевыми условиями:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < T, \\ u(0, t) &= u(1, t) = 0, \quad 0 \leq t \leq T, \\ u(x, 0) &= u_0(x), \quad 0 \leq x \leq 1. \end{aligned} \quad (5)$$

Решение этой задачи находится методом разделения переменных в виде

$$u(x, t) = \sum_{k=1}^{\infty} c_k e^{-\lambda_k t} X_k(x), \quad (6)$$

где  $\lambda_k$  и  $X_k(x)$  — собственные значения и ортонормированные собственные функции задачи

$$X'' + \lambda X = 0, \quad 0 < x < 1, \quad X(0) = X(1) = 0,$$

равные

$$\lambda_k = k^2 \pi^2, \quad X_k(x) = \sqrt{2} \sin k \pi x, \quad (7)$$

причем

$$(X_k, X_m) = \int_0^1 X_k(x) X_m(x) dx = \delta_{km},$$

$$\delta_{km} = \begin{cases} 1, & k = m, \\ 0, & k \neq m. \end{cases}$$

В самом деле, все частные решения (гармоники)  $u_k(x, t) = c_k e^{-\lambda_k t} X_k(x)$  удовлетворяют уравнению и краевым условиям (5). Из начального условия

$$u(x, 0) = u_0(x) = \sum_{k=1}^{\infty} c_k X_k(x) \quad (8)$$

находятся коэффициенты  $c_k = (u_0, X_k)$ .

Из (6) и (8) следует

$$\|u(t)\|^2 = (u(x, t), u(x, t)) =$$

$$= \sum_{k=1}^{\infty} c_k^2 e^{-2\lambda_k t} \|X_k\|^2 \leq e^{-2\lambda_1 t} \sum_{k=1}^{\infty} c_k^2 = e^{-2\lambda_1 t} \|u_0\|^2,$$

так как

$$\|u_0\|^2 = \sum_{k=1}^{\infty} c_k^2, \quad \lambda_k > \lambda_{k-1} > \dots > \lambda_1 = \pi^2.$$

Таким образом, для решения задачи (5) верна оценка

$$\|u(t)\| \leq e^{-\lambda_1 t} \|u_0\|, \quad \lambda_1 = \pi^2, \quad (9)$$

выражающая свойство асимптотической (при  $t \rightarrow \infty$ ) устойчивости задачи (5) по начальным данным (§ 4 п. 7 гл. V). В силу возрастания  $\lambda_k = k^2 \pi^2$  с ростом  $k$ , начиная с некоторого момента  $t$ , в сумме (6) будет преобладать первое слагаемое (первая гармоника), т. е. будет иметь место приближенное равенство

$$u(x, t) \approx c_1 e^{-\lambda_1 t} X_1(x).$$

Эта стадия процесса называется *регулярным режимом*.

**3. Разностные схемы.** В области  $\bar{D}$  введем сетку

$$\begin{aligned} \bar{\omega}_{h\tau} = \{(x_i t_j): x_i = ih, t_j = j\tau, i = 0, 1, \dots, N, h = 1/N, \\ j = 0, 1, \dots, L, \tau = T/L\} \end{aligned}$$

с шагами:  $h$  по  $x$  и  $\tau$  по  $t$ . Заменив производную по  $x$  разностным выражением

$$\left( \frac{\partial^2 u}{\partial x^2} \right)_i \sim \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = u_{xx,i} = \Lambda u_i,$$

вместо (3) получим систему дифференциально-разностных уравнений (*метод прямых*)

$$\frac{dv_i}{dt} = \Lambda v_i + f_i, \quad i = 1, 2, \dots,$$

с краевыми и начальными условиями

$$v_0(t) = u_1(t), \quad v_N(t) = u_2(t), \quad v_i(0) = u_0(x_i).$$

Для численного решения этой задачи, по аналогии с гл. V, заменим производную по  $t$  разностным отношением

$$\frac{dv_i}{dt} \sim \frac{v_i(t_{j+1}) - v_i(t_j)}{\tau} = \frac{v_i^{j+1} - v_i^j}{\tau} = (v_t)_i^j,$$

правую часть возьмем в виде линейной комбинации значений при  $t = t_j$  (на  $j$ -м слое) и  $t = t_{j+1}$  (на  $(j+1)$ -м слое):

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \sigma \Lambda y_i^{j+1} + (1 - \sigma) \Lambda y_i^j + \varphi_i^j, \quad (10)$$

где  $\sigma$  — параметр, а  $\varphi_i^j$  — некоторая правая часть, например,  $\varphi_i^j = f_i^j$ ,  $\varphi_i^j = f_i^{j+1/2}$  и т. д. Сюда надо присоединить дополнительные условия

$$y_0^j = u_1(t_j), \quad y_N^j = u_2(t_j), \quad y_i^0 = u_0(x_i), \quad (11)$$

$$j = 0, 1, 2, \dots, 0 \leq i \leq N.$$

Схема (10) определена на 6-точечном шаблоне

$$\begin{array}{ccc} (x_{i-1}, t_{j+1}) & (x_i, t_{j+1}) & (x_{i+1}, t_{j+1}) \\ \times & \times & \times \\ (x_{i-1}, t_j) & (x_i, t_j) & (x_{i+1}, t_j) \end{array}$$

Рассмотрим явную схему ( $\sigma = 0$ ) на 4-точечном шаблоне:

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \frac{y_{i-1}^j - 2y_i^j + y_{i+1}^j}{h^2} + \varphi_i^j. \quad (12)$$

Значения на  $(j+1)$ -м слое находятся по явной формуле

$$y_i^{j+1} = \left(1 - \frac{2\tau}{h^2}\right) y_i^j + \frac{\tau}{h^2} (y_{i-1}^j + y_{i+1}^j) + \tau \varphi_i^j.$$

В случае  $\sigma = 1$  получаем полностью неявную схему — схему с опережением на шаблоне  $\begin{smallmatrix} \times & \times & \times \end{smallmatrix}$ :

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \frac{y_{i-1}^{j+1} - 2y_i^{j+1} + y_{i+1}^{j+1}}{h^2} + \varphi_i^j. \quad (13)$$

Для определения  $y_i^{j+1}$  из (13) получаем краевую задачу

$$\frac{\tau}{h^2} y_{i-1}^{j+1} - \left(1 + \frac{2\tau}{h^2}\right) y_i^{j+1} + \frac{\tau}{h^2} y_{i+1}^{j+1} = -F_i^j, \quad 0 < i < N,$$

$$F_i^j = y_i^j + \tau \varphi_i^j, \quad y_0^{j+1} = u_1(t_{j+1}), \quad y_N^{j+1} = u_2(t_{j+1}),$$

которая решается методом прогонки.

Часто используется симметричная неявная схема (иногда ее называют *схемой Кранка — Николсона*) с  $\sigma = 1/2$  и шаблоном  $\begin{array}{cccc} \times & \times & \times \\ \times & \times & \times \end{array}$ :

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \frac{1}{2} \left( \frac{y_{i-1}^{j+1} - 2y_i^{j+1} + y_{i+1}^{j+1}}{h^2} + \frac{y_{i-1}^j - 2y_i^j + y_{i+1}^j}{h^2} \right) + \varphi_i^j. \quad (14)$$

Значения  $y_i^{j+1}$  на новом слое и в этом случае находятся методом прогонки для краевой задачи:

$$\begin{aligned} \frac{\tau}{2h^2} y_{i-1}^{j+1} - \left(1 + \frac{\tau}{h^2}\right) y_i^{j+1} + \frac{\tau}{2h^2} y_{i+1}^{j+1} &= -F_i^j, \quad 0 < i < N, \\ y_0^{j+1} &= u_1(t_{j+1}), \quad y_N^{j+1} = u_2(t_{j+1}), \\ F_i^j &= \left(1 - \frac{\tau}{h^2}\right) y_i^j + \frac{\tau}{2h^2} (y_{i-1}^j + y_{i+1}^j) + \tau \varphi_i^j. \end{aligned} \quad (15)$$

В общем случае (при любом  $\sigma$ ) схема (10) называется *схемой с весами*. При  $\sigma \neq 0$  она неявная и  $y_i^{j+1}$  определяется методом прогонки как решение задачи

$$\begin{aligned} \sigma \tau \Lambda y_i^{j+1} - y_i^{j+1} &= -F_i^j, \quad 0 < i < N, \\ y_0^{j+1} &= u_1(t_{j+1}), \quad y_N^{j+1} = u_2(t_{j+1}), \quad j = 0, 1, \dots \end{aligned} \quad (16)$$

Перейдем к изучению свойств схемы (10) с любым  $\sigma$ .

**4. Оценка погрешности аппроксимации.** Чтобы оценить порядок точности схемы с весами (10), надо сначала оценить погрешность аппроксимации (невязку) и найти априорные оценки, выражающие устойчивость схемы по правой части. Разностная схема (10), (11) учитывает начальные и граничные данные точно. Перепишем схему (10) в безиндексной форме. Вводя обозначения

$$\begin{aligned} y &= y_i^j, \quad \hat{y} = y_i^{j+1}, \quad \Lambda y = y_{xx} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}, \\ y_t &= \frac{y_i^{j+1} - y_i^j}{\tau}, \quad y^{(\sigma)} = \sigma y_i^{j+1} + (1 - \sigma) y_i^j, \end{aligned}$$

получаем

$$\begin{aligned} y_t &= \Lambda y^{(\sigma)} + \varphi, \quad (x_i t_j) \in \omega_{h\tau}, \quad y(x, 0) = u_0(x), \\ y_0 &= \mu_1(t), \quad y_N = u_2(t) \quad (t = t_j = j\tau, \quad j = 0, 1, \dots). \end{aligned} \tag{17}$$

Пусть  $u = u(x_i, t_j)$  — точное решение исходной задачи (3),  $y$  — решение разностной задачи (17). Подставляя в (17)  $y = z + u$ , получим для погрешности  $z = y - u$  следующие условия:

$$\begin{aligned} z_t &= \Lambda z^{(\sigma)} + \psi, \quad (x, t) \in \omega_{h\tau}, \\ z(x, 0) &= 0, \quad z(0, t) = z(1, t) = 0, \end{aligned} \tag{18}$$

где

$$\psi = \Lambda u^{(\sigma)} + \varphi - u_t \tag{19}$$

есть погрешность аппроксимации схемы (17) на решении  $u = u(x, t)$  задачи (3) (невязка схемы).

Найдем разложение  $\psi$  по степени  $h$  и  $\tau$  в окрестности точки  $(x_i, \bar{t} = t_j + \frac{1}{2}\tau)$ . Учитывая, что

$$\begin{aligned} u^{(\sigma)} &= \sigma \hat{u} + (1 - \sigma) u = \frac{u + \hat{u}}{2} + \left(\sigma - \frac{1}{2}\right) \tau u_t, \\ \hat{v} &= \bar{v} + \frac{1}{2} \tau \frac{\partial \bar{v}}{\partial t} + \frac{\tau^2}{8} \frac{\partial^2 \bar{v}}{\partial t^2} + \frac{\tau^3}{48} \frac{\partial^3 \bar{v}}{\partial t^3} + O(\tau^4), \\ \bar{v} &= v \left(x, t_j + \frac{1}{2} \tau\right), \\ v &= \bar{v} - \frac{1}{2} \tau \frac{\partial \bar{v}}{\partial t} + \frac{\tau^2}{8} \frac{\partial^2 \bar{v}}{\partial t^2} - \frac{\tau^3}{48} \frac{\partial^3 \bar{v}}{\partial t^3} + O(\tau^4), \\ \Lambda u &= u_{xx} = Lu + \frac{h^2}{12} u^{IV} + O(h^4), \quad Lu = \frac{\partial^2 u}{\partial x^2}, \end{aligned}$$

получаем

$$\begin{aligned} \psi &= \left(L\bar{u} + \bar{f} - \frac{\partial \bar{u}}{\partial t}\right) + \varphi - \bar{f} + \left(\sigma - \frac{1}{2}\right) \tau L \frac{\partial u}{\partial t} + \\ &\quad + \frac{h^2}{12} L^2 u + O(\tau^2 + h^4). \end{aligned}$$

Так как в силу уравнения (3) имеем

$$L\bar{u} + \bar{f} - \frac{\partial \bar{u}}{\partial t} = 0,$$

то

$$L \frac{\partial u}{\partial t} = L^2 u + L f$$

и

$$\begin{aligned} \psi = & \left( \varphi - \bar{f} + \left( \sigma - \frac{1}{2} \right) \tau L \bar{f} \right) + \left( \frac{h^2}{12} + \left( \sigma - \frac{1}{2} \right) \tau \right) L^2 \bar{u} + \\ & + O(\tau^2 + h^4). \end{aligned}$$

Отсюда видно, что

$$\psi = O(\tau + h^2) \text{ при } \varphi = f \text{ и } \sigma \neq \frac{1}{2},$$

$$\psi = O(\tau^2 + h^2) \text{ при } \varphi = \bar{f} \text{ и } \sigma = \frac{1}{2}.$$

Если выбрать  $\sigma$  так, чтобы коэффициент при  $L^2 \bar{u}$  был равен нулю:

$$\sigma = \sigma_* = \frac{1}{2} - \frac{h^2}{12\tau}, \quad (20)$$

а  $\varphi$  положить равным

$$\varphi = \bar{f} + \frac{h^2}{12} L \bar{f} \text{ или } \varphi = \bar{f} + \frac{h^2}{12} \Lambda \bar{f} \quad (21)$$

(оба выражения отличаются на величину  $O(h^4)$ , так как  $\Lambda \bar{f} - L \bar{f} = O(h^2)$ ), то мы получим схему повышенного (по  $x$ ) порядка аппроксимации:  $\psi = O(h^4 + \tau^2)$  при  $\sigma = \sigma_*$ . Эта схема также неявная, и поэтому  $y_i^{j+1}$  находится из уравнения  $\sigma_* \tau \Lambda \hat{y} - \hat{y} = -F$  методом прогонки.

**5. Устойчивость схемы.** Обратимся к изучению устойчивости и сходимости схемы (17). Рассмотрим сначала явную схему ( $\sigma = 0$ ) и чисто неявную схему ( $\sigma = 1$ ). Уравнение (17) для явной схемы запишем в виде

$$\begin{aligned} y_i^{j+1} = & \left( 1 - \frac{2\tau}{h^2} \right) y_i^j + \frac{\tau}{h^2} (y_{i-1}^j + y_{i+1}^j) + \tau \varphi_i^j, \quad 0 < i < N, \\ y_0^{j+1} = 0, \quad y_N^{j+1} = 0, \quad y_i^0 = & u_0(x_i), \quad 0 \leq i \leq N. \end{aligned} \quad (22)$$

Если коэффициент при  $y_i^j$  неотрицателен, т. е.

$$\tau \leq h^2/2, \quad (23)$$

то из (22) следует, что

$$\|y^{j+1}\|_C \leq \|y^j\|_C + \tau \|\varphi^j\|_C, \quad (24)$$

где  $\|y\|_C = \max_{0 \leq i \leq N} |y_i|$ . Суммирование по  $k$  от 0 до  $j-1$

дает

$$\|y^j\|_C \leq \|y^0\|_C + \sum_{k=0}^{j-1} \tau \|\varphi^k\|_C. \quad (25)$$

Это неравенство и выражает устойчивость в сеточной норме  $C$  явной схемы по начальным данным и по правой части при условии (23) (явная схема условно устойчива).

Неявную схему (17) при  $\sigma = 1$  перепишем в виде

$$\tau \Lambda y_i^{j+1} - y_i^{j+1} = -F_i, \quad F_i = y_i^j + \tau \varphi_i^j$$

или

$$\begin{aligned} \frac{\tau}{h^2} y_{i-1}^{j+1} - \left(1 + \frac{2\tau}{h^2}\right) y_i^{j+1} + \frac{\tau}{h^2} y_{i+1}^{j+1} &= -F_i^j, \quad 0 < i < N, \\ y_0^{j+1} = y_N^{j+1} &= 0. \end{aligned}$$

Воспользуемся теперь теоремой 3 из § 5 гл. I: для решения задачи

$$\begin{aligned} A_i y_{i-1} - C_i y_i + A_{i+1} y_{i+1} &= -F_i, \\ C_i = A_i + A_{i+1} + D_i, \quad 0 < i < N, \quad y_0 = y_N &= 0 \end{aligned}$$

верна оценка

$$\|y\|_C \leq \left\| \frac{F}{D} \right\|_C.$$

В нашем случае  $A_i = A_{i+1} = \tau/h^2$ ,  $D_i = 1$ ,

$$\|y^{k+1}\|_C \leq \|F^k\|_C \leq \|y^k\|_C + \tau \|\varphi^k\|_C. \quad (26)$$

Отсюда суммируем по  $k = 0, 1, \dots, j-1$ , получаем оценку (25). Таким образом, чисто неявная схема безусловно устойчива, т. е. устойчива при любых  $\tau$  и  $h$ . В случае произвольного  $\sigma$  разностное уравнение имеет вид

$$\begin{aligned} \frac{\sigma\tau}{h^2} y_{i-1}^{j+1} - \left(1 + \frac{2\sigma\tau}{h^2}\right) y_i^{j+1} + \frac{\sigma\tau}{h^2} y_{i+1}^{j+1} &= -F_i^j, \\ 0 < i < N, \quad y_0^{j+1} = y_N^{j+1} &= 0, \\ F_i^j = \left(1 - \frac{2(1-\sigma)\tau}{h^2}\right) y_i^j + \frac{(1-\sigma)\tau}{h^2} (y_{i-1}^j + y_{i+1}^j) + \tau \varphi_i^j. \end{aligned}$$

Отсюда видно, что коэффициент при  $y_i^j$  неотрицателен, если

$$\tau \leq \frac{h^2}{2(1-\sigma)} \text{ или } \sigma \geq 1 - \frac{h^2}{2\tau}. \quad (27)$$

При этом условии  $\|F\|_c \leq \|y\|_c + \tau \|\varphi\|_c$ ; пользуясь затем теоремой 3 из § 5 гл. I, получим оценку (25) при условии (27). В частности, для симметричной схемы устойчивость в  $C$  имеет место при  $\tau \leq h^2$ . Фактически же схема (17) с  $\sigma \geq 1/2$  безусловно устойчива в  $C$  по начальным данным, так что

$$\|y^j\|_c \leq M_0 \|\overset{\circ}{y}\|_c,$$

где  $M_0 = \text{const} > 1$ . Однако это неравенство доказывается довольно сложным способом.

Ниже будет показано, что в другой норме условие устойчивости схемы с весами имеет вид

$$\sigma \geq \sigma_0 = \frac{1}{2} - \frac{h^2}{4\tau}, \quad (28)$$

так что схема с  $\sigma \geq 1/2$  безусловно устойчива, а при  $\sigma < 1/2$  вместо (27) ставится условие устойчивости

$$\tau \leq \frac{h^2}{4(1/2 - \sigma)}. \quad (29)$$

Указанный результат (29) получается на основе общей теории устойчивости.

По аналогии с § 4 гл. I введем оператор  $A$ :

$$Ay = -\overset{\circ}{A}\overset{\circ}{y}, \quad y \in \Omega, \quad \overset{\circ}{y} \in \overset{\circ}{\Omega},$$

где  $\overset{\circ}{\Omega}$  — множество функций  $\overset{\circ}{y}$ , заданных на сетке  $\bar{\omega}_h = \{x_i: x_i = ih, i = 0, 1, \dots, N, h = 1/N\}$  и равных нулю на границе при  $i = 0, N$ , а  $y$  — множество функций, заданных во внутренних узлах сетки  $x \in \omega_h = \{x_i: x_i = ih, i = 1, 2, \dots, N-1, h = 1/N\}$ .

Запишем схему с весами в канонической форме:

$$Bz_t + Az = \psi(t), \quad t \in \bar{\omega}_\tau, \quad Z(0) = 0, \quad B = E + \sigma A. \quad (30)$$

Для этого достаточно подставить  $z^{(\sigma)} = \hat{z} + (1 - \sigma)z = z + \sigma(\hat{z} - z) = z + \sigma t z_t$  в (18).

Оператор  $A$ , как показано в гл. I, самосопряжен и положителен:  $A = A^* > 0$ , если скалярное произведение в  $H$  определить по формуле

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h.$$

Устойчивость схемы (30) исследована в гл. V, где показано, что схема (30) устойчива в  $H_A$  при

$$\sigma \geqslant \sigma_0 = \frac{1}{2} - \frac{1}{\tau \|A\|}. \quad (31)$$

В данном случае  $\|A\| = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}$ . Отсюда следует, что схема (17) устойчива при любых  $\tau$  и  $h$ , если  $\sigma \geqslant 1/2$ . Если  $\sigma < 1/2$ , то схема устойчива при

$$\tau \leqslant \frac{1}{(1/2 - \sigma) \|A\|}.$$

Подставляя сюда  $\|A\| \approx 4/h^2$ , получаем

$$\tau \leqslant \frac{h^2}{4(1/2 - \sigma)}, \text{ и } 4(1/2 - \sigma)\tau \leqslant h^2.$$

В частности, при  $\sigma = \sigma_*$  имеем  $4(1/2 - \sigma_*)\tau = h^2/3 < h^2$ , т. е. схема повышенного порядка аппроксимации безусловно устойчива.

**6. Сходимость схемы.** Для доказательства сходимости схемы (17) надо получить априорную оценку для задачи (30). Воспользуемся неравенством для  $z$ , полученным при исследовании сходимости схем в гл. V, в силу которого для (30) и (48) верна оценка погрешности

$$\|z^j\|_A \leqslant \sum_{h=0}^{j-1} \tau \|\psi^h\| \text{ при } \sigma \geqslant 0, \quad \sigma \geqslant \sigma_0. \quad (32)$$

Подставив сюда  $Az = \overset{\circ}{z}_{\bar{x},x}$ , найдем,

$$\|z\|_A^2 = (Az, z) = - \left( \overset{\circ}{z}_{\bar{x},x}, \overset{\circ}{z} \right) = \left[ \overset{\circ}{z}_{\bar{x}}, \overset{\circ}{z}_{\bar{x}} \right] = \sum_{i=1}^N h (z_{\bar{x},i})^2$$

и воспользуемся оценкой

$$\|z\|_C = \max_{x \in \omega_n} |z| \leqslant \frac{1}{2} \left( \sum_{i=1}^N h (z_{\bar{x},i})^2 \right)^{1/2} = \frac{1}{2} \|z\|_A.$$

В результате получаем

$$\|z^j\|_C \leqslant \frac{1}{2} \sum_{h=0}^{j-1} \tau \|\psi^h\|, \quad (33)$$

т. е. схема (17) сходится в сеточной норме  $C$  со скоростью

$$\|y^j - u^j\|_C = \|z^j\|_C = O(h^2 + \tau) \text{ при } \sigma \neq 1/2, \quad \sigma \geqslant \sigma_0,$$

$\|z^j\|_C = O(h^2 + \tau^2)$ ,  $\sigma = 1/2$ . Если  $\sigma_* \geqslant 0$ , т. е.  $\tau \geqslant h^2/\sigma$ ,

то и для схемы  $\sigma = \sigma_*$  верна оценка (33) и

$$\|z^j\|_C = O(h^4 + \tau^2) \text{ при } \sigma = \sigma_*.$$

**7. Асимптотическая устойчивость.** Свойство асимптотической (при  $t \rightarrow \infty$ ) устойчивости задачи (5) по начальным данным выражает оценка (9). При больших  $t$  решение задачи (5) определяется первой гармоникой

$$u(x, t) \approx c_1 e^{-\lambda_1 t} X_1(x)$$

(регулярный режим). Естественно требовать, чтобы решение разностной задачи

$$y_t = \sigma \hat{\Lambda} y + (1 - \sigma) \Lambda y; \quad x = ih, \quad t = j\tau, \quad (34)$$

$$i = 1, 2, \dots, N-1, \quad j = 0, 1, \dots,$$

$$y(0, t) = 0, \quad y(1, t) = 0, \quad y(x, 0) = u_0(x),$$

обладало аналитическими свойствами.

В гл. V для операторно-разностной схемы с весами

$$\begin{aligned} By_t + Ay = 0, \quad t \in \omega, \quad y(0) = y_0, \quad B = E + \sigma t A, \\ \delta E \leq A \leq \Delta E, \quad \delta > 0, \quad A = A^* > 0 \end{aligned}$$

установлена асимптотическая устойчивость схемы с весами

$$\|y^j\| \leq e^{-\tau t_j} \|y\|^\circ$$

при дополнительном условии

$$\tau \leq \tau_0(\sigma),$$

где  $\tau_0 = 2/(\delta + \Delta)$  для явной схемы ( $\delta = 0$ ),  $\tau_0 = \infty$  ( $\tau$  — любое) для неявной схемы ( $\sigma = 1$ ) и  $\tau_0 = 2/\sqrt{\delta\Delta}$  для симметричной схемы ( $\sigma = 1/2$ ). Для схемы (34) имеем

$$\delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}, \quad \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}, \quad \delta + \Delta = \frac{4}{h^2}.$$

Для явной схемы ( $\sigma = 0$ )  $\tau_0 = h^2/2$  и условие асимптотической устойчивости совпадает с условием обычной устойчивости; неявная схема  $\sigma = 1$  по-прежнему безусловно устойчива. Однако симметричная схема ( $\sigma = 1/2$ ), будучи безусловно устойчивой в обычном смысле, асимм-

тотически устойчива при условии

$$\tau \leq \tau_0, \quad \tau_0 = \frac{h^2}{\operatorname{tg} \pi h} \approx \frac{h}{\pi}.$$

В этом случае решение разностной задачи (34) с  $\sigma = 1/2$  при больших  $t$  определяется первой гармоникой:

$$y_i^j \approx c_1 \rho^j \sin \pi x_i \approx c_1 e^{-\lambda_1 t_j} \sin \pi x.$$

Здесь  $\rho = (1 - 1/2\tau\delta)/(1 + 1/2\tau\delta) = e^{-\lambda_1 \tau}(1 + O(\tau^2))$ .

Если условие  $\tau \leq \tau_0$  нарушено, т. е.  $\tau > \tau_0$ , то при больших  $t$  преобладает не первая, а последняя гармоника:

$$y_i^j \approx c_1 \rho^j \sin \pi (N - 1) x_i \approx c_1 \rho^j (-1)^i \sin \pi x_i,$$

где  $\rho = \frac{1/2\tau\Delta - 1}{1/2\tau\Delta + 1} < e^{-\lambda_1 \tau}$ , что, конечно, не имеет ничего общего с решением дифференциального уравнения.

Требование асимптотической устойчивости тесно связано с точностью схемы и фактически означает и требование асимптотической точности. Особенно четко это проявляется при расчетах на реальных сетках для больших  $t$ . Отметим, что условие  $\tau \approx h/\pi$  для симметричной схемы не является обременительным. Доказывается, что чисто неявная схема ( $\sigma = 1$ ) может обеспечить приемлемую точность в случае больших значений  $t$  только при шаге  $\tau$ , сравнимом с шагом явной схемы, что лишает чисто неявную схему при проведении расчетов для больших  $t$  ее основного преимущества — устойчивости при любых  $\tau$  и  $h$ .

## § 2. Многомерные задачи теплопроводности

**1. Разностные схемы с весами.** На плоскости  $x = (x_1, x_2)$  рассмотрим область  $G$  с границей  $\Gamma$ . Будем искать решение задачи теплопроводности в области  $\bar{G} = G + \Gamma$  для всех  $0 \leq t \leq T$ . Требуется найти функцию  $u(x, t)$ , определенную в цилиндре  $\bar{Q}_T = \bar{G} \times [0, T] = \{(x, t): x \in G, 0 \leq t \leq T\}$ , удовлетворяющую в  $Q_T = G \times (0, T) = \{(x, t): x \in G, 0 < t \leq T\}$  уравнению теплопроводности

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}, \quad (1)$$

краевым условиям первого рода на границе  $\Gamma$  области  $G$

$$u = \mu(x, t), \quad x \in T, \quad 0 \leq t \leq T, \quad (2)$$

и начальному условию при  $t = 0$ :

$$u(x, 0) = u_0(x), \quad x \in \bar{G}. \quad (3)$$

Предположим, что  $\bar{G}$  — прямоугольник:

$$\bar{G} = \{x = (x_1, x_2): 0 \leq x_1 \leq l_1, 0 \leq x_2 \leq l_2\}.$$

Введем в  $\bar{G}$  прямоугольную сетку

$$\begin{aligned} \bar{\omega}_h = \{x_1 = (x_1^{(i_1)}, x_2^{(i_2)}): & x_\alpha^{(i_\alpha)} = i_\alpha h_\alpha, \quad i_\alpha = 0, 1, \dots, N_\alpha, \\ & h_\alpha = l_\alpha / N_\alpha, \quad \alpha = 1, 2\} \end{aligned}$$

с границей

$$\begin{aligned} \gamma_h = \{x_i = (i_1 h_1, i_2 h_2): & i_1 = 0, N_1, 0 < i_2 < N_2; \\ & i_2 = 0, N_2, 0 < i_1 < N_1\}. \end{aligned}$$

Аппроксимируем оператор Лапласа  $Lu = \Delta u$  разностным оператором на пятиточечном шаблоне (см. гл. VI, § 1)

$$Lu \sim \Lambda u = u_{\bar{x}_1 \bar{x}_1} + u_{\bar{x}_2 \bar{x}_2}.$$

Задачу (1)–(3) заменим дифференциально-разностной задачей (методом прямых):

$$\begin{aligned} \frac{dv_i(t)}{dt} = \Lambda v_i(t) + f_i(t), \quad i = (i_1, i_2), \quad v_i(0) = u_0(x_i), \\ x_i \in \omega_h, \quad v_i(t)|_{\gamma_h} = \mu_i(t), \quad 0 \leq t \leq T. \end{aligned} \quad (4)$$

Введем на отрезке  $0 \leq t \leq T$  сетку  $\bar{\omega}_\tau = \{t_j = j\tau: 0 \leq t_j \leq T\}$  с шагом  $\tau$ . Напишем схему с весами

$$\frac{y^{j+1} - y^j}{\tau} = \Lambda(\sigma y^{j+1} + (1 - \sigma)y^j) + \varphi^j, \quad j = 0, 1, \dots, \quad (5)$$

где  $y^j = y(x_i, t_j) = y(i_1 h_1, i_2 h_2; t_j)$ ,  $x = (i_1 h_1, i_2 h_2) \in \omega_h$ . Присоединим к уравнениям (5)

$$y(x, 0) = u_0(x), \quad x = (i_1 h_1, i_2 h_2) \in \bar{\omega}_h, \quad (5')$$

$$y(x_i, t) = \mu_i(t), \quad x \in \gamma_h, \quad t = j\tau \in \omega_h.$$

Отсюда видно, что для определения  $\hat{y} = y^{j+1}$  на новом слое

$t = t_{j+1}$  надо решить разностное уравнение

$$\begin{aligned} \hat{y} - \sigma\tau\Lambda\hat{y} &= F, \quad F = y + (1 - \sigma)\tau\Lambda y + \tau\varphi, \quad x \in \omega_h, \\ \hat{y} &= \mu, \quad x \in \gamma_h. \end{aligned} \quad (6)$$

Разрешимость этой задачи следует из того, что оператор  $(E - \sigma\tau\Lambda)$  является положительно определенным при  $\sigma > -1/(\tau\|A\|)$ , в силу того, что  $(E - \sigma\tau\Lambda)\hat{y} = (E + \sigma\tau A)\hat{y}$  в пространстве сеточных функций  $\hat{y}$ , заданных на сетке  $\omega_h$  и обращающихся в нуль на границе  $\gamma_h$  (ср. гл. VI). Покажем это.

Вводя скалярное произведение

$$\begin{aligned} (y, v) &= \sum_{x_i \in \omega_h} y(x_i) v(x_i) h_1 h_2 = \\ &= \sum_{i_1=1}^{N_1-1} h_1 \sum_{i_2=1}^{N_2-1} h_2 y(i_1 h_1, i_2 h_2) v(i_1 h_1, i_2 h_2) \end{aligned} \quad (7)$$

и учитывая, что  $(Ay, y) \leq \|Ay\|\|y\| \leq \|A\|\|y\|^2$ , находим  
 $((E - \sigma\tau\Lambda)\hat{y}, \hat{y}) = ((E + \sigma\tau A)y, y) =$   
 $= \|y\|^2 + \sigma\tau(Ay, y) \geq \left( \frac{1}{\|A\|} + \sigma\tau \right) (Ay, y) > 0,$

так как  $(Ay, y) \geq \delta\|y\|^2 > 0$  (см. гл. VI, § 4, п. 5).

Запишем подробно в индексной форме разностное уравнение

$$\begin{aligned} \sigma\gamma_1 (\hat{y}_{i_1-1, i_2} + \hat{y}_{i_1+1, i_2}) - (1 + 2\sigma(\gamma_1 + \gamma_2)) \hat{y}_{i_1 i_2} + \\ + \sigma\gamma_2 (\hat{y}_{i_1 i_2-1} + \hat{y}_{i_1 i_2+1}) = -F_{i_1 i_2}, \end{aligned} \quad (8)$$

где

$$\begin{aligned} y_{i_1 i_2} &= y(i_1 h_1, i_2 h_2), \quad \gamma_1 = \tau/h_1^2, \quad \gamma_2 = \tau/h_2^2, \\ F_{i_1 i_2} &= (1 - 2(1 - \tau)(\gamma_1 + \gamma_2)) y_{i_1 i_2} + (1 - \sigma) \gamma_1 (y_{i_1-1, i_2} + \\ &+ y_{i_1+1, i_2}) + (1 - \sigma) \gamma_2 (y_{i_1, i_2-1} + y_{i_1, i_2+1}) + \varphi_{i_1 i_2}, \\ \hat{y}_{i_1 i_2} &= \hat{\mu}_{i_1 i_2}, \quad x_i = (i_1 h_1, i_2 h_2) \in \gamma_h. \end{aligned}$$

Это разностная краевая задача решается относительно  $\hat{y}$  теми же методами, что и разностная задача Дирихле для уравнения Пуассона (см. гл. VI, § 2). Здесь коэффициен-

ты уравнения постоянны, область  $\bar{G}$  — прямоугольник, поэтому наиболее экономичными являются прямые методы решения разностных уравнений (8). Итерационные методы менее экономичны.

**2. Устойчивость и сходимость.** Пользуясь определенным выше в гл. VI оператором  $A$ :

$$Ay = -\overset{\circ}{\Lambda}y = -\overset{\circ}{y}_{x_1 x_1} - \overset{\circ}{y}_{x_2 x_2}, \quad \overset{\circ}{y} \in \overset{\circ}{\Omega}, \quad y \in \Omega = H,$$

запишем схему (5) в канонической форме:

$$\begin{aligned} B \frac{y^{j+1} - y^j}{\tau} + Ay^j &= \varphi^j, \quad j = 0, 1, \dots, \quad y^0 = u_0, \quad y \in H, \\ B &= E + \sigma \tau A. \end{aligned} \tag{9}$$

Оператор  $A$  изучен в гл. VI. Он является самосопряженным и положительно определенным в пространстве  $H = \Omega$  размерности

$$(N_1 - 1)(N_2 - 1), \quad A = A^*, \quad \delta_0 E \leq A \leq \Delta_0 E,$$

где

$$\begin{aligned} \delta_0 &= \frac{4}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2}; \\ \Delta_0 &= \frac{4}{h_1^2} \cos^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \cos^2 \frac{\pi h_2}{2l_2}, \quad \Delta_0 = \|A\|. \end{aligned} \tag{10}$$

В силу общей теории (см. гл. V) схема (9) устойчива в  $H_A$  при

$$\sigma \geq \sigma_0, \quad \sigma_0 = \frac{1}{2} - \frac{1}{\tau \|A\|}. \tag{11}$$

В частности, для явной схемы имеем условие

$$\tau \leq \frac{2}{\Delta_0}, \quad \text{или} \quad \tau < \left( \frac{2}{h_1^2} + \frac{2}{h_2^2} \right)^{-1}. \tag{12}$$

На квадратной сетке ( $h_1 = h_2 = h$ ) условие устойчивости явной схемы имеет вид

$$\tau < h^2/4$$

(ср. с условиями  $\tau < h^2/2$  для одномерной задачи). Из (11) видно, что схемы с

$$\sigma \geq 1/2,$$

в том числе чисто неявная ( $\sigma = 1$ ) и симметричная ( $\sigma = 1/2$ ), безусловно устойчивы. Явную схему ( $\sigma = 0$ )

можно записать в виде

$$\begin{aligned} y_{i_1 i_2}^{j+1} = & (1 - 2(\gamma_1 + \gamma_2)) y_{i_1 i_2}^j + \gamma_1 (y_{i_1-1, i_2}^j + y_{i_1+1, i_2}^j) + \\ & + \gamma_2 (y_{i_1, i_2-1}^j + y_{i_1, i_2+1}^j) + \tau \varphi_{i_1 i_2}^j. \end{aligned} \quad (13)$$

Сумма коэффициентов при  $y$  в правой части (12) равна единице. Если все коэффициенты неотрицательны, т. е. выполнено условие  $\gamma_1 + \gamma_2 \leq 1/2$ ,  $\gamma_1 = \tau/h_1^2$ ,  $\gamma_2 = \tau/h_2^2$ , эквивалентное условию устойчивости (12), то из (13) следует неравенство

$$\|y^{j+1}\|_c \leq \|y^j\|_c + \tau \|\varphi^j\|_c.$$

Суммируя по  $k = 0, 1, \dots, j-1$ , получаем оценку (ср. § 4)

$$\|y^j\|_c \leq \|y^0\|_c + \sum_{k=0}^{j-1} \tau \|\varphi^k\|_c, \quad (14)$$

которая сохраняет силу при любых шагах сетки для чисто неявной схемы ( $\sigma = 1$ ). Во всех других случаях оценка (14) имеет место при  $\sigma \geq 1 - 1/\tau\Delta_0$ . Для доказательства сходимости надо, как обычно, исследовать невязку

$$\psi = \Lambda(\widehat{\sigma u} + (1-\sigma)u) + \varphi - u_t.$$

Учитывая, что  $\Lambda u = Lu + O(|h|^2)$ ,  $|h|^2 = h_1^2 + h_2^2$ , по аналогии с одномерным случаем находим

$$\psi = O(|h|^2 + \tau^2) + \left(\sigma - \frac{1}{2}\right) O(\tau).$$

Для погрешности  $z = y - u$  имеем задачу

$$B \frac{z^{j+1} - z^j}{\tau} + Az^j = \psi^j, \quad j = 0, 1, \dots, \quad z^0 = z(0) = 0.$$

Отсюда и из априорных оценок следует сходимость в  $C$  схемы (5) со скоростью  $O(\tau + |h|^2)$  при  $\sigma \neq 1/2$  и  $O(\tau^2 + |h|^2)$  при  $\sigma = 1/2$  (полная аналогия с одномерным случаем), если  $\sigma \geq 1 - \frac{1}{\tau\Delta_0}$ .

Для решения задачи, в силу оценки, полученной в гл. V, выполняется неравенство

$$\|z^{j+1}\|_A \leq \sum_{k=0}^j \tau \|\psi^k\| \quad \text{при } \sigma \geq \sigma_0 = \frac{1}{2} - \frac{1}{\tau\Delta_0}, \quad \sigma \geq 0,$$

где

$$\|z\|_A^2 = \|z\|_{A_1}^2 + \|z\|_{A_2}^2, \quad A_1 y = -y_{\bar{x}_1 x_1}, \quad A_2 y = -y_{\bar{x}_2 x_2},$$

$$\|z\|_A^2 = \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2-1} h_2 \left( z_{\bar{x}_1}(i_1, i_2) \right)^2 + \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2} h_2 \left( z_{\bar{x}_2}(i_1, i_2) \right)^2.$$

Отсюда следует безусловная устойчивость сходимости схемы (5) в  $H_A$  со скоростью  $O(\tau + |h|^2)$  при  $\sigma \neq 1/2$ ,  $\sigma \geq 1/2$  и  $O(\tau^2 + |h|^2)$  при  $\sigma = 1/2$ .

Проведенное выше исследование надо дополнить условиями асимптотической устойчивости. Поскольку эти условия  $\tau \leq \tau_0$  были получены для операторно-разностной схемы с весами с произвольным оператором

$$A = A^* > 0, \quad \delta_0 E \leq A \leq \Delta_0 E,$$

то ими можно воспользоваться и для нашей схемы (5). Пользуясь выражениями (10) для  $\delta_0$  и  $\Delta_0$ , получаем условия асимптотической устойчивости  $\tau \leq \tau_0^{(1)}$ ,  $\tau_0^{(1)} = 2 \left( \frac{4}{h_1^2} + \frac{4}{h_2^2} \right)^{-1}$  для явной схемы ( $\sigma = 0$ ),  $\tau \leq \tau_0^{(2)}$ ,  $\tau_0^{(2)} = \frac{2}{\sqrt{\delta_0 \Delta_0}}$ ,  $\delta_0, \Delta_0$  из (10), для симметричной схемы ( $\sigma = 1/2$ ).

В частности, при  $h_1 = h_2 = h$ ,  $l_1 = l_2 = l$  имеем

$$\delta_0 = \frac{8}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \Delta_0 = \frac{8}{h^2} \cos^2 \frac{\pi h}{2l}, \quad \tau_0^{(1)} = \frac{h^2}{4},$$

$$\tau_0^{(2)} = \frac{h^2}{2} \left( \sin \frac{\pi h}{l} \right)^{-1} \approx \frac{hl}{2\pi}.$$

Предельное значение  $\tau_0^{(2)}$  в два раза меньше, чем для одномерной схемы (5) из § 1.

Чисто неявная схема  $\sigma = 1$  безусловно асимптотически устойчива.

**3. Переменные коэффициенты.** Рассмотрим задачу (1), предполагая, что  $L$  есть эллиптический оператор второго порядка с переменными коэффициентами и без смешанных производных:

$$Lu = L_1 u + L_2 u, \quad L_1 u = \frac{\partial}{\partial x_1} \left( k_1(x, t) \frac{\partial u}{\partial x_1} \right),$$

$$L_2 u = \frac{\partial}{\partial x_2} \left( k_2(x, t) \frac{\partial u}{\partial x_2} \right),$$

$$c_1 \leq k_\alpha(x, t) \leq c_2, \quad (x, t) \in \bar{Q}_T = G \times (0, T].$$

Каждый из операторов  $L_1$  и  $L_2$  аппроксимируем разностным трехточечным оператором:

$$\begin{aligned} L_1 &\sim \Lambda_1, & L_2 &\sim \Lambda_2, \\ \Lambda_1 v &= (a_1 v_{\bar{x}_1})_{x_1}, & \Lambda_2 v &= (a_2 v_{\bar{x}_2})_{x_2}, \end{aligned}$$

где  $a_1 = a_1(i_1 h_1, i_2 h_2, t)$ ,  $a_2 = a_2(i_1 h_1, i_2 h_2, t)$  — некоторые функционалы от значений  $k_1$  и  $k_2$  соответственно; в простейшем случае  $a_1 = k_1((i_1 - 1/2)h_1, i_2 h_2, t)$ ,  $a_2 = k_2(i_1 h_1; (i_2 - 1/2)h_2, t)$ , что обеспечивает второй порядок аппроксимации:  $\Lambda_\alpha u - L_\alpha u = O(h_\alpha^2)$ ,  $\alpha = 1, 2$ . Оператору  $L$  ставится в соответствие разностный оператор  $\Lambda$ :

$$\Lambda v = \Lambda_1 v + \Lambda_2 v = (a_1 v_{\bar{x}_1})_{x_1} + (a_2 v_{\bar{x}_2})_{x_2}. \quad (45)$$

Запишем  $\Lambda_1 v$  и  $\Lambda_2 v$  в индексной форме

$$\begin{aligned} \Lambda_1 v &= \frac{1}{h_1} \left[ a_1 ((i_1 + 1) h_1, i_2 h_2; t) \frac{v_{i_1+1, i_2} - v_{i_1, i_2}}{h_1} - \right. \\ &\quad \left. - a_1 (i_1 h_1, i_2 h_2; t) \frac{v_{i_1, i_2} - v_{i_1-1, i_2}}{h_1} \right], \\ \Lambda_2 v &= \frac{1}{h_2} \left[ a_2 (i_1 h_1, (i_2 + 1) h_2; t) \frac{v_{i_1, i_2+1} - v_{i_1, i_2}}{h_2} - \right. \\ &\quad \left. - a_2 (i_1 h_1, i_2 h_2; t) \frac{v_{i_1, i_2} - v_{i_1, i_2-1}}{h_2} \right]. \end{aligned}$$

Разностная схема с весами имеет тот же вид (5), что и в п. 1. Берется то же сеточное пространство  $H = \Omega$  со скалярным произведением (7) и вводится оператор  $A$ :

$$Ay = -\overset{\circ}{\Lambda}y = -(a_1 \overset{\circ}{y}_{\bar{x}_1})_{x_1} - (a_2 \overset{\circ}{y}_{\bar{x}_2})_{x_2}.$$

Учитывая, что для одномерного случая оператора

$$A: Ay = -(\overset{\circ}{ay}_{\bar{x}})_x$$

$$c_1 (\overset{\circ}{A}y, y) \leq (Ay, y) \leq c_2 (\overset{\circ}{A}y, y), \quad \overset{\circ}{A}y = -\overset{\circ}{y}_{\bar{x}x},$$

$$0 < c_1 \leq a \leq c_2,$$

нетрудно убедиться, что такие же неравенства выполняются и для двумерного оператора (45):

$$c_1 \overset{\circ}{A} \leq A \leq c_2 \overset{\circ}{A}, \quad \overset{\circ}{A}y = -\overset{\circ}{y}_{\bar{x}_1 x_1} - \overset{\circ}{y}_{\bar{x}_2 x_2}.$$

Отсюда видно, что  $\delta E \leq A \leq \Delta E$ ,  $\delta = c_1 \delta_0$ ,  $\Delta = c_2 \Delta_0$ , где  $\delta_0$  и  $\Delta_0$  определяются по формулам (10). Для определения  $\hat{y} = y^{j+1}$  на новом слое получаем задачу (6), где  $\Lambda$  определяется из (15). В случае явной схемы  $\hat{y}$  определяется в каждом узле  $x \in \omega_h$  по формуле

$$\hat{y} = y + (1 - \sigma) \tau \Lambda y + \tau \varphi.$$

Для неявных схем ( $\sigma \neq 0$ ) надо решать пятиточечное разностное уравнение с переменными коэффициентами. Здесь используются итерационные методы, наиболее экономичным из них является попеременно-треугольный метод (см. гл. V, § 5), число итераций для которого есть величина  $O\left(\frac{1}{\sqrt[4]{h}} \ln \frac{1}{\epsilon}\right)$ , если  $\tau \approx O(h)$ . Описание попеременно-треугольного метода для разностных уравнений с переменными коэффициентами дано в гл. VI; применительно к уравнению (6) с оператором  $\Lambda$  вида (15) его следует несколько видоизменить.

### § 3. Экономичные схемы

**1. Метод переменных направлений.** Сравним явные и неявные схемы (5) по двум характеристикам: объем вычислений для определения  $y^{j+1}$  и ограничение на шаг  $\tau$ .

**Явная схема:** для определения  $y^{j+1}$  на сетке  $\omega_h$  надо затратить число действий, пропорциональное числу узлов, т. е. число действий, приходящихся на один узел, не зависит от сетки  $\omega_h$ . Однако шаг  $\tau$  жестко ограничен сверху условием

$$\tau \leq \tau_0(h): \tau \leq h^2/4 \quad \text{при } h_1 = h_2 = h \text{ для схемы (13).}$$

**Неявная схема ( $\sigma \geq 1/2$ ):** для определения  $y^{j+1}$  надо решить систему  $(N_1 - 1)(N_2 - 1)$  пятиточечных разностных уравнений; для этого, по крайней мере в случае переменных коэффициентов, требуется число действий на один узел сетки  $\omega_h$ , возрастающее при  $|h| \rightarrow 0$ .

Возникает задача — построить схемы, сочетающие лучшие качества явных и неявных схем: безусловно устойчивые, с числом действий на каждом слое, пропорциональным числу узлов сетки  $\omega_h$ . Такие схемы принято называть **экономичными**. Конечно, мы должны сделать оговорку: безусловно устойчивые в обычном смысле схемы должны быть асимптотически устойчивы, что приводит

к ограничению на шаг, значительно более слабому (например,  $\tau \leq h/(2\pi)$  при  $\sigma = 1/2$ ,  $h_1 = h_2 = h$ ,  $l_1 = l_2 = l$ ), чем условие устойчивости ( $\tau \leq h^2/4$ ) для явной схемы. Кстати, условие  $\tau = O(h)$  естественно для схемы  $O(\tau^2 + |h|^2)$ .

Первые экономичные схемы появились в 1955—1956 гг. и были названы *методами переменных направлений*. Основная алгоритмическая идея их экономичности состоит в том, что для перехода со слоя  $t_j$  на слой  $t_{j+1}$  надо решать методом прогонки трехточечные разностные уравнения спачала вдоль строк, а затем — вдоль столбцов сетки  $\omega_h$ .

Приведем формулы метода переменных направлений (продольно-поперечной схемы Писмена — Рекфорда) для задачи (1) с оператором  $L$ :  $Lu = L_1 u + L_2 u$ , где  $L_\alpha$  — один из операторов:

$$L_\alpha u = \frac{\partial^2 u}{\partial x_\alpha^2} \text{ или } L_\alpha u = \frac{\partial}{\partial x_\alpha} \left( k_\alpha(x, t) \frac{\partial u}{\partial x_\alpha} \right), \quad \alpha = 1, 2.$$

Пусть  $\Lambda_2$ ,  $\Lambda_1$  — соответствующие трехточечные операторы и  $\Lambda = \Lambda_1 + \Lambda_2$ . Вводя промежуточное значение  $\bar{y} = y^{j+1/2}$ , формулируем разностную схему переменных направлений:

$$\frac{y^{j+1/2} - y^j}{\tau/2} = \Lambda_1 y^{j+1/2} + \Lambda_2 y^j + \varphi^j, \quad x \in \omega_h, \quad y^{j+1/2} = \bar{\mu} \quad \text{при } i_1 = 0, N_1, \quad (1)$$

$$\frac{y^{j+1} - y^{j+1/2}}{\tau/2} = \Lambda_1 y^{j+1/2} + \Lambda_2 y^{j+1} + \varphi^j, \quad x \in \omega_h, \quad y^{j+1} = \bar{\mu}^{j+1} \quad \text{при } i_2 = 0, N_2, \quad y^0 = u_0(x), \quad x \in \bar{\omega}_h, \quad (2)$$

где  $\bar{\mu}$  — промежуточное значение функции  $\mu(x, t)$ , равное

$$\bar{\mu} = \frac{\mu^j + \mu^{j+1}}{2} - \frac{\tau}{4} \Lambda_2 (\mu^{j+1} - \mu^j).$$

Для определения  $y^{j+1/2}$  и  $y^{j+1}$  имеем разностные краевые задачи

$$\begin{aligned} & \frac{1}{2} \tau \Lambda_1 y^{j+1/2} - y^{j+1/2} = -F^j, \\ & F^j = y^j + 1/2 \tau (\Lambda_2 y^j + \varphi^j), \quad x \in \omega_h, \\ & y^{j+1/2} = \bar{\mu}, \quad i_1 = 0, N_1, \\ & \frac{1}{2} \tau \Lambda_2 y^{j+1} - y^{j+1} = -F^{j+1/2}, \end{aligned} \quad (3)$$

$$\begin{aligned} F^{j+1/2} &= y^{j+1/2} + \frac{1}{2}\tau(\Lambda_1 y^{j+1/2} + \varphi^j), \quad x \in \omega_h, \\ y^{j+1} &= \mu^{j+1}, \quad i_2 = 0, N_2. \end{aligned}$$

Первая задача решается прогонкой по строкам ( $i_2 = 1, 2, \dots, N_2 - 1$ ), вторая — прогонкой по столбцам ( $i_1 = 1, 2, \dots, N_1 - 1$ ). Число действий на один узел конечно и не зависит от сетки.

Схема (3) устойчива как по начальным данным, так и по правой части при любых  $\tau$  и  $|h|$  и имеет точность  $O(\tau^2 + |h|^2)$ . В этом можно убедиться путем исключения  $y^{j+1/2}$  и сведения схемы (1), (2), к эквивалентной двухслойной схеме с факторизованным оператором  $B_j$ :

$$B \frac{y^{j+1} - y^j}{\tau} + Ay^j = \Phi^j, \quad j = 0, 1, \dots, \quad y^0 = u_0 \in H, \quad (4)$$

$$B = \left( E + \frac{\tau}{2} A_1 \right) \left( E + \frac{\tau}{2} A_2 \right), \quad A_\alpha y = -\Lambda_\alpha \overset{\circ}{y} = -\overset{\circ}{y}_{x_\alpha x_\alpha}, \quad \alpha = 1, 2,$$

где  $H = \Omega$  — пространство сеточных функций, заданных во внутренних узлах сетки  $\omega_h$ .

Очевидно, что  $A_\alpha = A_\alpha^* > 0$ ,  $\alpha = 1, 2$ ,  $A_1 A_2 = A_2 A_1$ . Поэтому  $B = E + \tau A/2 + \tau^2 A_1 A_2 / 2 \geq E + \tau A/2 > \tau A/2$ , и схема устойчива.

**2. Факторизованные схемы.** Оператор  $B$ , представленный в виде произведения нескольких операторов  $B = B_1 B_2 \dots B_p$ , будем называть *факторизованным*, а соответствующую схему

$$B \frac{y^{j+1} - y^j}{\tau} + Ay^j = \varphi^j, \quad j = 0, 1, \dots, \quad y^0 = y(0), \quad (5)$$

— *факторизованной схемой*.

Если для решения задачи

$$B_\alpha v = F_\alpha, \quad \alpha = 1, 2,$$

с заданной правой частью  $F_\alpha$  требуется  $O(N_1 N_2)$  число действий, то и для определения  $y^{j+1}$  по известному  $y^j$  надо  $O(N_1 N_2)$  действий (оператор  $B$  «экономичен»). Так как

$$By^{j+1} = B_1 B_2 y^{j+1} = F^j,$$

то алгоритм сводится к последовательному решению уравнений

$$B_1 y^{j+1/2} = F^j, \quad B_2 y^{j+1} = y^{j+1/2}.$$

Опираясь на теорию устойчивости двухслойных схем, нетрудно, отправляясь от схемы с весами, построить экономичную факторизованную схему (методом регуляризации).

Итак, пусть

$$A = A_1 + A_2, \quad B = E + \sigma\tau A = E + \sigma\tau(A_1 + A_2),$$

$$A_1 = A_1^*, \quad A_2 = A_2^*.$$

Тогда схема (9) из § 2 устойчива при  $\sigma \geq \sigma_0 = \frac{1}{2} - \frac{1}{\tau \|A\|}$ .

Заменим в (9) оператор  $B$  факторизованным оператором

$$\tilde{B} = (E + \sigma\tau A_1)(E + \sigma\tau A_2),$$

отличающимся от  $B$  членом  $\sigma^2\tau^2 A_1 A_2$ ,

$$\tilde{B} = B + \sigma^2\tau^2 A_1 A_2.$$

В результате получим факторизованную схему

$$\tilde{B} \frac{y^{j+1} - y^j}{\tau} + Ay^j = \Phi^j, \quad j = 0, 1, \dots, y^0 = u_0 \in H, \quad (6)$$

того же порядка аппроксимации  $O((\sigma - 1/2)\tau + \tau^2)$ , что и исходная схема с весами. Так как исходная схема с весами устойчива ( $\sigma \geq \sigma_0$ ), то и факторизованная схема (6) устойчива в силу условия

$$\tilde{B} > B \geq \tau A/2,$$

которое выполнено, если  $A_1$  и  $A_2$  перестановочны и  $A_\alpha^* = -A_\alpha > 0$ ,  $\alpha = 1, 2$ .

Для определения  $y^{j+1}$  мы получаем уравнение  $\tilde{B}y^{j+1} = F^j$ , или

$$(E + \sigma\tau A_1)(E + \sigma\tau A_2)y^{j+1} = F^j,$$

$$F^j = \tilde{B}y^j + \tau(\Phi^j - Ay^j),$$

которое решается последовательно:

$$(E + \sigma\tau A_1)\bar{y} = F^j, \quad (E + \sigma\tau A_2)y^{j+1} = \bar{y}$$

(с соответствующими краевыми условиями). Более экономичным (экономия на вычислении правой части  $F^j$ ) является следующий алгоритм:

$$(E + \sigma\tau A_1)w^{j+1/2} = F^j = \Phi^j - Ay^j, \quad (7)$$

$$(E + \sigma\tau A_2)w^{j+1} = w^{j+1/2}, \quad y^{j+1} = y^j + \tau w^{j+1}.$$

Однако при этом надо хранить не один, а два вектора

( $w^{j+1/2}$  или  $w^{j+1}$  и  $y^j$ ). При  $\sigma = 1$  из (7) следует вторая схема переменных направлений (*схема Дугласа — Рекфорда*)

$$\frac{y^{j+1/2} - y^j}{\tau} + A_1 y^{j+1/2} + A_2 y^j = \Phi^j,$$

$$(E + \tau A_2) \frac{y^{j+1} - y^{j+1/2}}{\tau} = \frac{y^{j+1/2} - y^j}{\tau}.$$

**3. Метод суммарной аппроксимации.** Чтобы получить экономичные схемы для широкого класса задач (уравнения с переменными коэффициентами, области сложной формы и т. д.), необходимо изменить понятие разностной схемы.

Мы отказываемся от обычного понятия аппроксимации, которое мы рассматривали выше, и заменяем его более слабым понятием *суммарной аппроксимации*. Поясним его. Пусть переход от слоя  $j$  к слою  $j+1$  осуществляется в несколько этапов, на каждом из которых используется обычная двухслойная схема, не аппроксимирующая исходное уравнение, однако сумма певязок для каждой промежуточной схемы

$$\psi = \sum_{\alpha=1}^p \psi_\alpha \quad (8)$$

стремится к нулю при стремлении к нулю шага  $\tau$  по переменному  $t$ .

Идею метода суммарной аппроксимации можно изложить на примере задачи Коши для обыкновенного дифференциального уравнения

$$\frac{du}{dt} + au = f(t), \quad t > 0, \quad u(0) = u_0, \quad (9)$$

где  $a > 0$  — число. Предположим, что

$$a = a_1 + a_2, \quad a_1 > 0, \quad a_2 > 0, \quad f(t) = f_1(t) + f_2(t). \quad (10)$$

Очевидно, что такое представление возможно всегда.

Введем сетку  $\omega_\tau = \{t_j = j\tau, j = 0, 1, \dots\}$  и на каждом шаге  $(t_j, t_{j+1})$  будем решать вместо (9) последовательно два уравнения

$$\frac{1}{2} \frac{dv_{(1)}}{dt} + a_1 v_{(1)} = f_1(t), \quad t_j \leq t \leq t_{j+1/2} = t_j + \frac{\tau}{2}, \quad (11)$$

$$\frac{1}{2} \frac{dv_{(2)}}{dt} + a_2 v_{(2)} = f_2(t), \quad t_{j+1/2} \leq t \leq t_{j+1}$$

с начальными данными

$$\begin{aligned} v_{(1)}(t_j) &= v(t_j), \quad v_{(2)}(t_{j+1/2}) = v_{(1)}(t_{j+1/2}), \\ j &= 0, 1, \dots, v_{(1)}(0) = u_0. \end{aligned} \quad (12)$$

Решением задачи (11)–(12) является функция

$$v(t) = v_{(2)}(t). \quad (13)$$

Каждое из уравнений (11) аппроксимируем двухслойной разностной схемой с шагом  $\tau/2$ . Например, возьмем неявную схему

$$\frac{y^{j+1/2} - y^j}{\tau} + a_1 y^{j+1/2} = f_1^j, \quad (14)$$

$$\frac{y^{j+1} - y^{j+1/2}}{\tau} + a_2 y^{j+1} = f_2^j.$$

Вычислим невязки  $\psi_1$  и  $\psi_2$  для схем (11). Подставим в (11)

$$\begin{aligned} y^j &= z^j + u^j, \quad y^{j+1/2} = z^{j+1/2} + u^{j+1/2}, \quad y^{j+1} = z^{j+1} + u^{j+1}, \\ \frac{z^{j+1/2} - z^j}{\tau} + a_1 z^{j+1/2} &= -\psi_1^j, \\ \frac{z^{j+1} - z^{j+1/2}}{\tau} + a_2 z^{j+1} &= -\psi_2^j, \quad j = 0, 1, \dots, \\ z^0 &= 0, \quad \psi_1^j = \frac{u^{j+1/2} - u^j}{\tau} + a_1 u^{j+1/2} - f_1^j, \\ \psi_2^j &= \frac{u^{j+1} - u^{j+1/2}}{\tau} + a_2 u^{j+1} - f_2^j. \end{aligned}$$

Подставляя сюда

$$u^{j+1} = (u + \overset{\circ}{\tau u}/2)^{j+1/2} + O(\tau^2), \quad u^j = (u - \overset{\circ}{\tau u}/2)^{j+1/2} + O(\tau^2),$$

получаем

$$\begin{aligned} \psi_1^j &= (\overset{\circ}{u}/2 + a_1 u - f_1)^{j+1/2} + O(\tau), \\ \psi_2^j &= (\overset{\circ}{u}/2 + a_2 u - f_2)^{j+1/2} + O(\tau). \end{aligned} \quad (15)$$

Отсюда видно, что  $\psi_1^j = O(1)$ ,  $\psi_2^j = O(1)$ , однако

$$\psi_1^j + \psi_2^j = O(\tau) \rightarrow O \quad \text{при} \quad \tau \rightarrow 0. \quad (16)$$

Все проведенные выше рассуждения, начиная с (10), (11), (14), сохраняют силу, если  $a_1$  и  $a_2$  — матрицы или операторы, а  $u$ ,  $f$ ,  $y$  — векторы.

Таким образом, схема (11), (12) аппроксимирует задачу (9) в суммарном смысле (16) (такие схемы мы называем *аддитивными*).

Для доказательства сходимости схемы (11), (12) надо получить оценку для погрешности  $z^{j+1} = y^{j+1} - u^{j+1}$ , учитывающую свойство (16) суммарной аппроксимации. Положим

$$\begin{aligned}\psi_\alpha &= \overset{\circ}{\psi}_\alpha + \psi_\alpha^*, \\ \overset{\circ}{\psi}_\alpha &= (\overset{\circ}{u}/2 + a_\alpha u - f_\alpha)^{j+1/2}, \quad \psi_\alpha^* = O(\tau), \quad \alpha = 1, 2, \\ z^{j+1/2} &= \eta_{j+1/2} + \xi_{j+1/2}, \quad z^{j+1} = \eta_{j+1} + \xi_{j+1},\end{aligned}$$

где  $\eta_{j+1}$ ,  $\xi_{j+1}$  — решения задач

$$\begin{aligned}\eta_{j+1/2} &= \eta_j + \tau \overset{\circ}{\psi}_1, \quad \eta_{j+1} = \eta_{j+1/2} + \tau \overset{\circ}{\psi}_2, \\ j &= 0, 1, \dots, \eta_0 = 0.\end{aligned}\quad (17)$$

$$\begin{aligned}(1 + a_1 \tau) \xi_{j+1/2} &= \xi_j + \tau \tilde{\psi}_1, \quad (1 + a_2 \tau) \xi_{j+1} = \xi_{j+1/2} + \tau \tilde{\psi}_2, \\ j &= 0, 1, \dots,\end{aligned}\quad (18)$$

$$\begin{aligned}\xi_0 &= 0, \\ \tilde{\psi}_1^j &= \psi_1^{*j} - a_1 \tau \eta_{j+1/2}, \quad \tilde{\psi}_2^j = \psi_2^{*j} - a_2 \tau \eta_{j+1}.\end{aligned}\quad (19)$$

Отсюда находим  $\eta_{j+1} = \eta_j + \tau (\overset{\circ}{\psi}_1^j + \overset{\circ}{\psi}_2^j) = \eta_j = \dots = \eta_0 = 0$ , т. е.  $\eta_j = 0$  для всех  $j = 0, 1, \dots$ , и  $z^j = \xi_j$ .

$$\eta_{j+1/2} = \tau \overset{\circ}{\psi}_1 = O(\tau), \quad \tilde{\psi}_\alpha = O(\tau). \quad (20)$$

Из (16) получаем

$$\begin{aligned}|\xi_{j+1/2}| &\leq |\xi_j| + \tau |\tilde{\psi}_1^j|, \\ |\xi_{j+1}| &\leq |\xi_{j+1/2}| + \tau |\tilde{\psi}_2^j| \leq |\xi_j| + \tau (|\tilde{\psi}_1^j| + |\tilde{\psi}_2^j|),\end{aligned}$$

так что справедлива оценка

$$|z^{j+1}| \leq \sum_{h=0}^j \tau (|\tilde{\psi}_1^h| + |\tilde{\psi}_2^h|), \quad (21)$$

из которой в силу (17), и следует сходимость со скоростью  $O(\tau)$  аддитивной схемы (14).

Вместо (11) можно взять другую систему уравнений:

$$\begin{aligned} \frac{dv_{(1)}}{dt} + a_1 v_{(1)} &= f_1(t), \quad t_j \leq t \leq t_{j+1}, \quad v_{(1)}(t_j) = v(t_j), \\ \frac{dv_{(2)}}{dt} + a_2 v_{(2)} &= f_2(t), \quad t_j \leq t \leq t_{j+1}, \quad v_{(2)}(t_j) = v_{(1)}(t_{j+1}), \\ j &= 0, 1, \dots, \quad v_{(1)}(0) = u_0. \end{aligned} \quad (22)$$

Решением этой задачи является функция

$$v(t) = v_{(2)}(t). \quad (23)$$

В отличие от (11) здесь оба уравнения интерпретируются на всем отрезке  $t_j \leq t \leq t_{j+1}$ , и поэтому аппроксимация этих уравнений проводится с шагом  $\tau$  (а не  $\tau/2$ , как в случае (11)) и дает те же схемы (14). Оба способа сведения задачи (9) к системе задач (11) или (22) используют одно и то же свойство

$$a = a_1 + a_2 \quad (24)$$

и условие  $f = f_1 + f_2$ , которому всегда можно удовлетворить.

Рассмотрим в качестве примера уравнение теплопроводности

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad x = (x_1, x_2), \quad (25)$$

$$Lu = \Delta u = L_1 u + L_2 u, \quad l_\alpha u = \frac{\partial^2 u}{\partial x_\alpha^2}, \quad \alpha = 1, 2,$$

$L_1$  и  $L_2$  — «одномерные» операторы. Решение уравнения

$$\frac{\partial v_{(\alpha)}}{\partial t} = L_\alpha v_{(\alpha)} + f_\alpha, \quad (26)$$

очевидно, является более простой задачей, чем решение уравнения (25). Условия  $L = L_1 + L_2$ ,  $f = f_1 + f_2$  гарантируют суммарную аппроксимацию для схемы, получающейся при обычной аппроксимации, например, с помощью двухслойной схемы с весами каждого из уравнений системы

$$\frac{dv_{(1)}}{dt} = L_1 v_{(1)} + f_1, \quad t_j \leq t \leq t_{j+1}, \quad v_{(1)}^j = v^j,$$

$$\frac{dv_{(2)}}{dt} = L_2 v_{(2)} + f_2, \quad t_j \leq t \leq t_{j+1}, \quad v_{(2)}^j = v_{(1)}^{j+1},$$

$$v^{j+1} = v_{(2)}^{j+1}.$$

В результате мы получим аддитивную схему, локально-одномерную схему или схему расщепления

$$\begin{aligned} \frac{y^{j+1/2} - y^j}{\tau} &= \Lambda_1 (\sigma_1 y^{j+1/2} + (1 - \sigma_1) y^j) + \varphi_1^j, \quad x \in \omega_h, \\ \frac{y^{j+1} - y^{j+1/2}}{\tau} &= \Lambda_2 (\sigma_2 y^{j+1} + (1 - \sigma_2) y^{j+1/2}) + \varphi_2^j, \\ y^0 &= u_0(x), \quad x \in \omega_h, \\ y^{j+1/2}|_{\gamma_h} &= \mu^{j+1/2}, \quad y^{j+1}|_{\gamma_h} = \mu^{j+1}. \end{aligned} \quad (27)$$

Здесь  $\Lambda_1 y = y_{x_1 x_1}$ ,  $\Lambda_2 y = y_{x_2 x_2}$ . Параметры  $\sigma_1$  и  $\sigma_2$  определяются из условий устойчивости и аппроксимации. Например, при  $\sigma_1 = \sigma_2 = 1$  получаем схему с опережением

$$\begin{aligned} \frac{y^{j+1/2} - y^j}{\tau} &= \Lambda_1 y^{j+1/2} + \varphi_1^j, \\ \frac{y^{j+1} - y^{j+1/2}}{\tau} &= \Lambda_2 y^{j+1} + \varphi_2^j, \quad j = 0, 1, \dots \end{aligned}$$

Подставляя сюда  $y^j = z^j + u^j$ ,  $y^{j+1/2} = z^{j+1/2} + (u^j + u^{j+1})/2$ ,  $y^{j+1} = z^{j+1} + u^{j+1}$ , получим для погрешности  $z$  уравнения

$$\begin{aligned} \frac{z^{j+1/2} - z^j}{\tau} &= \Lambda_1 z^{j+1/2} + \psi_1^j, \\ \frac{z^{j+1} - z^{j+1/2}}{\tau} &= \Lambda_2 z^{j+1} + \psi_2^j, \end{aligned}$$

где  $u$  — решение исходной задачи (25),  $\psi_1$  и  $\psi_2$  — невязки, равные

$$\begin{aligned} \psi_1^j &= \Lambda_1 \frac{u + \hat{u}}{2} - \frac{1}{2} \frac{\hat{u} - u}{\tau} + \varphi_1^j, \quad \psi_2^j = \Lambda_2 \hat{u} - \frac{1}{2} \frac{\hat{u} - u}{\tau} + \varphi_2^j, \\ \hat{u} &= u^{j+1}, \quad u = u^j. \end{aligned}$$

Отсюда видно, что  $\psi_1 = O(1)$ ,  $\psi_2 = O(1)$ , т. е. каждое из уравнений (27) в отдельности не аппроксимирует уравнение (25). Возьмем сумму невязок

$$\begin{aligned} \psi &= \psi_1 + \psi_2 = \Lambda_1 \frac{u + \hat{u}}{2} + \Lambda_2 \hat{u} - \frac{\hat{u} - u}{\tau} + \varphi_1 + \varphi_2 = \\ &= (L_1 + L_2) \bar{u} - \frac{\partial \bar{u}}{\partial t} + \varphi_1 + \varphi_2 + O(\tau + |h|^2), \end{aligned}$$

где  $\tilde{u} = u^{j+1/2}$ . Учитывая уравнение (25) при  $t = t_{j+1/2}$ , получим

$$\psi = \varphi_1 + \varphi_2 - f^{j+1/2} + O(\tau + |h|^2) = O(\tau + |h|^2),$$

$$|h|^2 = h_1^2 + h_2^2,$$

если

$$\varphi_1 + \varphi_2 = f^{j+1/2} + O(\tau^2).$$

Этого можно достигнуть, полагая, например,

$$\varphi_1 = 0, \varphi_2 = f^{j+1/2} \text{ или } \varphi_1 = \varphi_2 = f^j/2.$$

Можно показать, что схема (27) сходится равномерно со скоростью

$$O(\tau + |h|^2), \text{ т. е. } \|y^{j+1} - u^{j+1}\|_c = O(\tau + |h|^2).$$

Из приведенных примеров видно, что метод суммарной аппроксимации позволяет проводить расщепление сложных задач на последовательность более простых и существенно упрощать решение многомерных задач математической физики.

## ДОПОЛНЕНИЕ

### Марш-алгоритм и метод редукции для решения системы линейных уравнений с трехдиагональной матрицей

Во многих приложениях встречаются задачи, приводящие к решению систем линейных алгебраических уравнений специального вида (с разреженной матрицей, имеющей много нулевых элементов) высокого порядка. Такие системы возникают при разностной аппроксимации эллиптических уравнений или при использовании неявных схем для уравнения теплопроводности и др.

После аппроксимации обыкновенного дифференциального уравнения второго порядка на трехточечном шаблоне в гл. IV было получено разностное уравнение второго порядка, которое представляет собой систему линейных алгебраических уравнений порядка  $N - 1$  ( $N - 1$  — число внутренних узлов) с трехдиагональной матрицей. В § 3 гл. I для решения такой системы был построен метод, для реализации которого требуется  $O(N)$  арифметических операций.

При аппроксимации двумерного уравнения Пуассона на пятиточечном шаблоне в гл. VI была получена разностная схема, которой соответствует система линейных алгебраических уравнений с пятидиагональной матрицей порядка  $N = (N_1 - 1)(N_2 - 1)$ , где  $N_1 - 1$ ,  $N_2 - 1$  — число внутренних узлов по каждому направлению. При разбиении вектора неизвестных на блоки, содержащие по  $N_1 - 1$  элементов, мы получим заштрихованые системы с блочно-трехдиагональной матрицей, число блоков которой равно  $N_2 - 1$ . Для такой системы в § 2 гл. VI был рассмотрен метод разделения переменных с оценкой  $O(N \log N)$  для числа операций. При многократном решении систем подобного типа важное значение приобретает экономичность вычислительных алгоритмов.

Ниже будет построен прямой метод решения специальных систем с трехдиагональной матрицей, для которого требуется всего  $O(N)$  операций как в случае, когда элементы матрицы суть скаляры, так и в случае блочной матрицы.

**1. Марш-алгоритм.** Сначала рассмотрим случай, когда элементы матрицы — скаляры. Зашифтуем систему с трехдиагональной матрицей в виде трехточечной разностной задачи:

$$-y_{i-1} + Cy_i - y_{i+1} = F_i, \quad 1 \leq i \leq N - 1, \quad y_0 = 0, \quad y_N = 0, \quad (1)$$

где  $C$  — число, и предположим, что  $N = 2k + 1$ . Если разностное уравнение второго порядка (1) записать в виде рекуррентных соотношений

$$y_{i+1} = Cy_i - y_{i-1} - F_i, \quad i \geq 1, \quad y_0 = 0, \quad (2)$$

то нетрудно заметить, что все неизвестные  $y_i$  можно найти последовательно по формуле (2), если каким-либо способом вычислить значение  $y_1$ . При этом любое  $y_i$  будет линейно выражаться через  $y_0$  и  $y_1$ . Сказанное дает нам основание записать для любого

$i \geq 1$  соотношение

$$y_{i+1} = \alpha_i y_1 - \beta_{i-1} y_0 - p_i \quad (3)$$

с неопределенными пока коэффициентами  $\alpha_i$ ,  $\beta_i$ ,  $p_i$ . Если положить

$$\alpha_0 = 1, \quad \beta_{-1} = 0, \quad p_0 = 0, \quad (4)$$

то (3) будет справедливо и при  $i = 0$ . Итак, решение задачи (1) будем искать в виде (3) для любого  $i \geq 0$ .

Записывая (1) в виде рекуррентных соотношений

$$y_{i-1} = Cy_i - y_{i+1} - F_i, \quad i \leq N-1, \quad y_N = 0 \quad (5)$$

и проводя аналогичные рассуждения, получим, что решение задачи (1) для любого  $i \leq N$  можно искать в виде

$$y_{i-1} = \xi_{N-i} y_{N-1} - \eta_{N-i-1} y_N - q_{N-i}, \quad (6)$$

если положить

$$\xi_0 = 1, \quad \eta_{-1} = 0, \quad q_0 = 0. \quad (7)$$

Заметим, что если  $y_{N-1}$  будет найдено, то все  $y_i$  можно вычислить последовательно по формуле (5).

Найдем  $y_1$  и  $y_{N-1}$ . Для этого определим коэффициенты  $\alpha_i$ ,  $\beta_i$ ,  $\xi_i$ ,  $\eta_i$ ,  $p_i$ ,  $q_i$ . Сравнивая (2) и (3) при  $i = 1$ , а (5) и (6) при  $i = N-1$ , получим

$$\alpha_1 = \xi_1 = C, \quad \beta_0 = \eta_0 = 1, \quad p_1 = F_1, \quad q_1 = F_{N-1}. \quad (8)$$

Найдем теперь рекуррентные формулы для определения искомых коэффициентов. Подставим (3), а также вытекающие из него выражения для  $y_i$  и  $y_{i-1}$ :

$$y_i = \alpha_{i-1} y_1 - \beta_{i-2} y_0 - p_{i-1}, \quad y_{i-1} = \alpha_{i-2} y_1 - \beta_{i-3} y_0 - p_{i-2}$$

в уравнения (1). Получим

$$-(\alpha_{i-2} - Ca_{i-1} + \alpha_i) y_1 + (\beta_{i-3} - C\beta_{i-2} + \beta_{i-1}) y_0 + \\ + p_{i-2} - Cp_{i-1} + p_i = F_i, \quad i \geq 2.$$

Для того чтобы эти равенства были тождественными для всех  $i$ , достаточно положить для  $i \geq 2$

$$p_i = Cp_{i-1} - p_{i-2} + F_i, \quad (9)$$

$$\alpha_i = Ca_{i-1} - \alpha_{i-2}, \quad \beta_{i-1} = C\beta_{i-2} - \beta_{i-3}. \quad (10)$$

Аналогично, используя (6) и (1), получим для  $i \leq N-2$  рекуррентные соотношения

$$q_{N-i} = Cq_{N-i-1} - q_{N-i-2} + F_i,$$

$$\xi_{N-i} = C\xi_{N-i-1} - \xi_{N-i-2}, \quad \eta_{N-i-1} = C\eta_{N-i-2} - \eta_{N-i-3}.$$

Заменив здесь  $N-i$  на  $i$ , получим для  $i \geq 2$  формулы

$$q_i = Cq_{i-1} - q_{i-2} + F_{N-i}, \quad (11)$$

$$\xi_i = C\xi_{i-1} - \xi_{i-2}, \quad \eta_{i-1} = C\eta_{i-2} - \eta_{i-3}. \quad (12)$$

Итак, формулы (4), (7)–(12) полностью определяют искомые коэффициенты. Сравнивая (10) и (12) при условиях (4), (7), (8),

получим, что  $\beta_i = \eta_i = \xi_i = a_i$  для  $i \geq 0$ . Таким образом, формулы (3), (6) принимают вид

$$y_{i+1} = a_i y_i - a_{i-1} y_0 - p_i, \quad i \geq 0, \quad (13)$$

$$y_{i-1} = a_{N-i} y_{N-1} - a_{N-i-1} y_N - q_{N-i}, \quad i \leq N, \quad (14)$$

где

$$p_i = C p_{i-1} - p_{i-2} + F_i, \quad i \geq 2, \quad p_0 = 0, \quad p_1 = F_1, \quad (15)$$

$$q_i = C q_{i-1} - q_{i-2} + F_{N-i}, \quad i \geq 2, \quad q_0 = 0, \quad q_1 = F_{N-1}, \quad (16)$$

$$a_i = C a_{i-1} - a_{i-2}, \quad i \geq 2, \quad a_0 = 1, \quad a_1 = C. \quad (17)$$

Найдем теперь  $y_1$  и  $y_{N-1}$ . Для этого положим в (13)  $i = k$ , а в (14)  $i = k + 2$ . Учитывая, что  $N = 2k + 1$ , получим

$$y_{k+1} = a_k y_1 - a_{k-1} y_0 - p_k, \quad y_{k+1} = a_{k-1} y_{N-1} - a_{k-2} y_N - q_{k-1}.$$

Вычитая из второго равенства первое, получим уравнение относительно  $y_1$  и  $y_{N-1}$ :

$$a_{k-1} y_{N-1} - a_k y_1 + a_{k-1} y_0 - a_{k-2} y_N = q_{k-1} - p_k. \quad (18)$$

Подучим еще одно уравнение для  $y_1$  и  $y_{N-1}$ , полагая  $i = k - 1$  в (13) и  $i = k + 1$  в (14) и вычитая из первого равенства второе,

$$-a_k y_{N-1} + a_{k-1} y_1 - a_{k-1} y_0 + a_{k-1} y_N = p_{k-1} - q_k. \quad (19)$$

Учитывая, что  $y_0 = y_N = 0$ , сложим и вычтем (18) и (19). Получим эквивалентную систему

$$(a_{k-1} - a_k)(y_{N-1} + y_1) = q_{k-1} - p_k + p_{k-1} - q_k, \quad (20)$$

$$(a_{k-1} + a_k)(y_{N-1} - y_1) = q_{k-1} - p_k - p_{k-1} + q_k,$$

решая которую, найдем искомые значения  $y_1$  и  $y_{N-1}$ :

$$y_1 = (\alpha_{k-1}^2 - \alpha_k^2)^{-1} [\alpha_k (q_{k-1} - p_k) + \alpha_{k-1} (p_{k-1} - q_k)], \quad (21)$$

$$y_{N-1} = (\alpha_{k-1}^2 - \alpha_k^2)^{-1} [\alpha_{k-1} (q_{k-1} - p_k) + \alpha_k (p_{k-1} - q_k)].$$

Таким образом, алгоритм решения задачи (1) состоит в вычислении по формулам (15)–(17) коэффициентов  $p_{k-1}$ ,  $p_k$ ,  $q_{k-1}$ ,  $q_k$ ,  $\alpha_{k-1}$ ,  $\alpha_k$ , по формулам (21) – значений  $y_1$ ,  $y_{N-1}$  и неизвестных  $y_i$ ,  $i = 2, 3, \dots, k$ , по формуле (2), а для  $i = N - 2, N - 3, \dots, k + 1$  по формуле (5) при заданных  $y_0$ ,  $y_N$  и вычисленных  $y_1$ ,  $y_{N-1}$ . Описанный алгоритм получил название *марш-алгоритма*. Число подсчетов, что для его реализации требуется примерно  $8N$  операций. Можно показать, что если  $C \neq 2 \cos m\pi/N$ ,  $m$  – целое число, то задача (1) разрешима при любой правой части и  $\alpha_{k-1}^2 \neq \alpha_k^2$ . Следовательно, в этом случае формулы (21) не содержат деления на нуль.

Описанный выше марш-алгоритм можно использовать и в случае, когда  $C$  – квадратная матрица,  $F_i$  – заданные, а  $y_i$  – искомые векторы. Заметим, что рассмотренная нами в гл. VI разностная задача Дирихле для уравнения Пуассона на прямоугольной равномерной по каждому направлению сетке, введенной в прямоугольник, может быть записана в виде (1). В этом случае компонентами вектора являются значения искомой сеточной функции, соответствующие  $i$ -й строке сетки, а матрица  $C$  – трехдиагональная и ее порядок равен числу внутренних строк сетки.

Пусть  $M$  — порядок матрицы  $C$ . Тогда векторы  $p_i, q_i$  имеют размер  $M$  и для вычисления  $p_{k-1}, q_{k-1}, p_k, q_k$  по формулам (15), (16) потребуется  $O(MN)$  операций. Очевидно, что такое же количество операций потребуется и для нахождения векторов  $y_i$ ,  $2 \leq i \leq N-2$ , по формулам (2), (5). Рассмотрим теперь вопрос о вычислении  $y_1$  и  $y_{N-1}$ .

Из формулы (17) следует, что  $\alpha_k$  есть полином степени  $k$  от  $C$ , причем, если  $C$  — число, то  $\alpha_k$  — алгебраический полином, а если  $C$  — матрица, то  $\alpha_k$  — матричный полином. Для полинома, удовлетворяющего рекуррентному соотношению (17), есть явное представление:  $\alpha_k = U_k(C/2)$ , где  $U_k(x)$  — полином Чебышева второго рода степени  $k$ :

$$U_k(x) = \begin{cases} \frac{\sin(k+1)\arccos x}{\sin\arccos x}, & |x| \leq 1, \\ \frac{(x + \sqrt{x^2 - 1})^{k+1} - (x - \sqrt{x^2 - 1})^{k+1}}{2\sqrt{x^2 - 1}}, & |x| \geq 1. \end{cases}$$

Используя явное выражение для  $\alpha_k$ ,  $k \geq 0$ , и учитывая, что  $\alpha_k$  — полином с единичным коэффициентом при старшей степени, можно получить следующие разложения:

$$\begin{aligned} \alpha_k - \alpha_{k-1} &= \prod_{l=1}^k \left( C - 2 \cos \frac{(2l-1)\pi}{2k+1} E \right), \\ \alpha_k + \alpha_{k-1} &= \prod_{l=1}^k \left( C - 2 \cos \frac{2l\pi}{2k+1} E \right). \end{aligned} \quad (22)$$

Используя (22) и (20), построим следующий алгоритм для нахождения  $y_1$  и  $y_{N-1}$ :

$$\begin{aligned} v_0 &= p_k - q_{k-1} - p_{k-1} + q_k, \quad w_0 = q_{k-1} - p_k - p_{k-1} + q_k, \\ &\quad \left( C - 2 \cos \frac{(2l-1)\pi}{2k+1} E \right) v_l = v_{l-1}, \\ &\quad \left( C - 2 \cos \frac{2l\pi}{2k+1} E \right) w_l = w_{l-1}, \quad l = 1, 2, \dots, k, \\ y_1 &= 0,5(v_k - w_k), \quad y_{N-1} = 0,5(v_k + w_k). \end{aligned} \quad (23)$$

Так как каждая из систем (23) имеет трехдиагональную матрицу (число таких систем  $2k$ ) и может быть решена методом прогонки с затратой  $O(M)$  операций, то для нахождения  $y_1$  и  $y_{N-1}$  потребуется  $O(NM)$  арифметических операций.

Итак, для решения системы (1) с трехдиагональной матрицей построен метод с числом арифметических операций, пропорциональным числу неизвестных.

Обратим внимание на то, что построенный марш-алгоритм может быть численно неустойчивым. Действительно, если число  $C$  удовлетворяет условию  $|C| > 2$ , то для алгоритма характерен экспоненциальный по  $N$  рост погрешности, поскольку среди корней характеристического уравнения  $q^2 - Cq + 1 = 0$  имеется один, по модулю большие единицы. Такого же типа неустойчивость

имеет место и в том случае, когда матрица  $C$  имеет собственные значения, превосходящие по модулю 2. Для таких задач в настоящее время построен вариант марш-алгоритма, устойчивый в том смысле, что погрешность растет по степенному закону при росте  $N$ .

**2. Метод редукции.** В ряде случаев при решении систем линейных алгебраических уравнений с трехдиагональной матрицей большое значение имеет точность полученного решения. Анализ формул метода прогонки, который применяется для решения таких систем, показывает, что источником погрешности могут служить формулы для вычисления прогоночных коэффициентов. Эти формулы содержат операцию деления на разность близких по значению величин. Ниже будет рассмотрен метод редукции решения указанных систем, свободный от этого недостатка.

Итак, пусть требуется найти решение трехточечной разностной задачи

$$-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, \quad 1 \leq i \leq N-1, \quad (24)$$

$$y_0 = 0, \quad y_N = 0,$$

где  $c_i = a_i + b_i + d_i$ ,  $a_i > 0$ ,  $b_i > 0$ ,  $d_i \geq 0$ ,  $N = 2^n$ . Идея метода редукции состоит в последовательном исключении из системы (24) неизвестных сначала с нечетными номерами, затем с номерами, кратными 2, и т. д.

Выпишем три идущие подряд уравнения системы (24) с номерами  $i-1$ ,  $i$ ,  $i+1$ , где  $i$  — четное число:

$$-a_{i-1} y_{i-2} + (a_{i-1} + b_{i-1} + d_{i-1}) y_{i-1} - b_{i-1} y_i = f_{i-1}, \quad (25)$$

$$-a_i y_{i-1} + (a_i + b_i + d_i) y_i - b_i y_{i+1} = f_i, \quad (26)$$

$$-a_{i+1} y_i + (a_{i+1} + b_{i+1} + d_{i+1}) y_{i+1} - b_{i+1} y_{i+2} = f_{i+1}. \quad (27)$$

Умножая уравнение (25) на  $\alpha_i^{(1)} = a_i (a_{i-1} + b_{i-1} + d_{i-1})^{-1}$ , уравнение (27) на  $\beta_i^{(1)} = b_i (a_{i+1} + b_{i+1} + d_{i+1})^{-1}$  и складывая полученные уравнения с (26), найдем

$$-a_i^{(1)} y_{i-2} + (a_i^{(1)} + b_i^{(1)} + d_i^{(1)}) y_i - b_i^{(1)} y_{i+2} = f_i^{(1)}, \quad (28)$$

$$i = 2, 4, 6, \dots, N-2, \quad y_0 = 0, \quad y_N = 0,$$

где  $a_i^{(1)} = \alpha_i^{(1)} a_{i-1}$ ,  $b_i^{(1)} = \beta_i^{(1)} b_{i+1}$ ,  $d_i^{(1)} = \alpha_i^{(1)} d_{i-1} + d_i + \beta_i^{(1)} d_{i+1}$ ,  $f_i^{(1)} = \alpha_i^{(1)} f_{i-1} + f_i + \beta_i^{(1)} f_{i+1}$ . Если неизвестные с четными номерами будут найдены (они удовлетворяют системе (28)), то остальные неизвестные определяются по формуле

$$y_i = \frac{f_i + a_i y_{i-1} + b_i y_{i+1}}{a_i + b_i + d_i}, \quad i = 1, 3, 5, \dots, N-1.$$

Описанный процесс исключения неизвестных может быть, очевидно, применен к системе (28), из которой на втором шаге будут исключены неизвестные с номерами, кратными 2, но не кратными 4. В результате  $l$ -го шага процесса исключения получим

систему

$$\begin{aligned} -a_i^{(l)}y_{i-2^l} + (a_i^{(l)} + b_i^{(l)} + d_i^{(l)})y_i - b_i^{(l)}y_{i+2^l} &= f_i^{(l)}, \\ i = 2^l, 2 \cdot 2^l, 3 \cdot 2^l, \dots, N - 2^l, \quad y_0 = 0, \quad y_N = 0, \end{aligned} \quad (29)$$

где

$$\begin{aligned} a_i^{(l)} &= \alpha_i^{(l)}a_{i-2^{l-1}}^{(l-1)}, \quad b_i^{(l)} = \beta_i^{(l)}b_{i+2^{l-1}}^{(l-1)}, \\ d_i^{(l)} &= \alpha_i^{(l)}d_{i-2^{l-1}}^{(l-1)} + a_i^{(l-1)} + \beta_i^{(l)}d_{i+2^{l-1}}^{(l-1)}, \\ f_i^{(l)} &= \alpha_i^{(l)}f_{i-2^{l-1}}^{(l-1)} + f_i^{(l-1)} + \beta_i^{(l)}f_{i+2^{l-1}}^{(l-1)}, \\ \alpha_i^{(l)} &= a_i^{(l-1)} \left( a_{i-2^{l-1}}^{(l-1)} + b_{i-2^{l-1}}^{(l-1)} + d_{i-2^{l-1}}^{(l-1)} \right)^{-1}, \\ \beta_i^{(l)} &= b_i^{(l-1)} \left( a_{i+2^{l-1}}^{(l-1)} + b_{i+2^{l-1}}^{(l-1)} + d_{i+2^{l-1}}^{(l-1)} \right)^{-1}, \\ i &= 2^l, 2 \cdot 2^l, 3 \cdot 2^l, \dots, N - 2^l, \quad l \geq 1. \end{aligned} \quad (30)$$

Здесь использованы обозначения  $a_i^{(0)} = a_i$ ,  $b_i^{(0)} = b_i$ ,  $d_i^{(0)} = d_i$ ,  $f_i^{(0)} = f_i$ .

Процесс исключения закончится на  $(n-1)$ -м шаге, когда система (29) будет состоять из одного уравнения относительно неизвестного  $y_{N/2} = y_{2^{n-1}}$ . Из этого уравнения найдем

$$y_{2^{n-1}} = \frac{f_{2^{n-1}}^{(n-1)} + a_{2^{n-1}}^{(n-1)}y_0 + b_{2^{n-1}}^{(n-1)}y_N}{a_{2^{n-1}}^{(n-1)} + b_{2^{n-1}}^{(n-1)} + d_{2^{n-1}}^{(n-1)}}, \quad y_0 = y_N = 0. \quad (31)$$

Остальные неизвестные определяются по формулам

$$y_i = \frac{f_i^{(l)} + a_i^{(l)}y_{i-2^l} + b_i^{(l)}y_{i+2^l}}{a_i^{(l)} + b_i^{(l)} + d_i^{(l)}}, \quad i = 2^l, 3 \cdot 2^l, 5 \cdot 2^l, \dots, N - 2^l, \quad (32)$$

где  $l = n-2, n-3, \dots, 0$ ,  $y_0 = y_N = 0$ . Заметим, что формула (32) включает в себя формулу (31) при  $l = n-1$ .

Итак, в методе редукции на прямом ходе по формулам (30) для  $l = 1, 2, \dots, n-1$  вычисляются  $a_i^{(l)}$ ,  $b_i^{(l)}$ ,  $d_i^{(l)}$ ,  $f_i^{(l)}$ , а на обратном ходе по формуле (32) для  $l = n-1, n-2, \dots, 0$  находится искомое решение. Отметим, что метод не требует дополнительной памяти, так как величины  $a_i^{(l)}$ ,  $b_i^{(l)}$ ,  $d_i^{(l)}$ ,  $f_i^{(l)}$  могут быть размещены соответственно на месте  $a_{i-2^{l-1}}^{(l-1)}$ ,  $b_{i-2^{l-1}}^{(l-1)}$ ,  $d_i^{(l-1)}$ ,  $f_i^{(l-1)}$ . Для реализации метода требуется  $12N$  сложений,  $8N$  умножений и  $3N$  делений.

## ЛИТЕРАТУРА

1. Бахвалов Н. С. Численные методы.— М.: Наука, 1975.
2. Березин И. С., Жидков Н. П. Методы вычислений.— М.: Наука, 1966, ч. 1; Физматгиз, 1962, ч. 2.
3. Вoeводин В. В. Численные методы алгебры; теория и алгоритмы.— М.: Наука, 1966.
4. Годунов С. К., Рябенский В. С. Разностные схемы.— М.: Наука, 1977.
5. Калинин И. Н. Численные методы.— М.: Наука, 1978.
6. Ляшко И. И., Макаров В. Л., Скоробогатько А. А. Методы вычислений.— Киев: Высшая школа, 1977.
7. Марчук Г. И. Методы вычислительной математики.— М.: Наука, 1980.
8. Никольский С. М. Квадратурные формулы.— М.: Наука, 1979.
9. Самарский А. А. Теория разностных схем.— М.: Наука, 1977.
10. Самарский А. А., Андреев В. Б. Разностные методы для эллиптических уравнений.— М.: Наука, 1976.
11. Самарский А. А., Гулин А. В. Устойчивость разностных схем.— М.: Наука, 1973.
12. Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений.— М.: Наука, 1978.
13. Самарский А. А., Попов Ю. П. Разностные методы газовой динамики.— М.: Наука, 1980.
14. Тихонов А. Н., Самарский А. А. Уравнения математической физики.— М.: Наука, 1972.
15. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры.— М.: Физматгиз, 1963.
16. Яненко Н. Н. Метод дробных шагов решения многомерных задач математической физики.— Новосибирск: Наука, 1967.

## ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

- Алгоритм неустойчивый 13
  - условно устойчивый 13
  - экономичный 11
- Аппроксимация разностная (на сетке) 138
  - суммарная 254
- Весовые множители 70
- Вычислительная неустойчивость 115
- Жесткие системы уравнений 192
- Задача Дирихле 211
  - корректная 14
  - Коши 32
  - краевая 32
  - некорректная 15
  - о собственных значениях 42
- Интерполянта 64
- Интерполяционный полином 62
  - — Лагранжа 64
  - — Ньютона 64
- Интерполяция Эрмитова 65
- Итерационные методы 90
- Итерационный метод двухшаговый (трехслойный) 97
  - — неявный 97
  - — одношаговый (двухслойный) 97
  - — явный 97
- Квадратурная формула 70
  - — Гаусса 82
  - — Котеса 74
  - — прямоугольника 71
  - — Симпсона 72
  - — трапеции 71
  - — Чебышева 83
- Коэффициенты Лагранжа 62
- Краевые условия 33
  - 1-го рода 33
  - 2-го рода 33
- Краевые условия 3-го рода 33
- Кубическая сплайн-интерполяция 65
- Линейно независимые векторы 39
  - — решения 27
- Линейное пространство 38
  - — действительное 38
  - — комплексное 38
- Мажорантная функция (мажоранта) 55
- Матрица верхняя треугольная 87
  - диагональная 86
  - ленточная 88
  - нижняя треугольная 86
  - разреженная 87
- Мера обусловленности 89
- Метод Адамса — Штёрмера 191
  - баланса (интегро-интерполяционный) 167
  - Бубнова — Галеркина 173
  - вариационно-разностный 171
  - вариационного типа 126
  - верхней релаксации 101
  - дихотомии 130
  - Зейделя 99
  - касательных 133
  - конечных элементов 173
  - линеаризации 133
  - минимальных невязок 127
  - Ньютона 133
  - переменных направлений 251
  - Пикара (последовательных приближений) 175
  - nonпеременно-треугольный 120
  - поправок 128
  - прогонки 34
  - — встречной 37
  - — левой 37
  - — правой 37

- Метод простой итерации 98
  - прямой 89
  - прямых 234
  - разделения переменных 222
  - Ритца 172
  - Ричардсона 115
  - Рунге 82, 165, 178
  - Рунге — Кутта 174
  - секущих 136
  - скорейшего спуска 128
  - сопряженных градиентов 129
  - стационарный итерационный 102
  - сумматорных тождеств 171
  - Штёрмера 189
  - энергетических неравенств 144, 207
- Минимизирующий квадратичный функционал 171
- Наилучшее среднеквадратичное приближение 68
- Невязка для разностной схемы на решении 146
- Норма оператора 40
- Обратное интерполирование 67
- Однородная разностная схема 150
- Оператор единичный 41
  - линейный 40
  - неотрицательный 41
  - обратный 40
  - ограниченный 40
  - положительный 41
  - разрешающий 111
  - самосопряженный 41
  - сопряженный 41
  - факторизованный 129, 252
  - экономичный (экономичность оператора) 119
- Операторное уравнение первого рода 88
- Операторы перестановочные 41
- Ошибка округления 10
- Погрешность аппроксимации для краевого условия 146
  - в точке,  $m$ -й порядок 139
  - для уравнения 146
  - на решении 147
  - на сетке 140, 185
  - оператора 139
  - квадратурной формулы 70
  - метода 10
- Погрешность неустранимая 10
- Нолином обобщенный 68
  - Чебышева 112, 114
- Принцип максимума 55
- Пространство евклидово (унитарное) 39
  - нормированное 39
  - сеточных функций 46
  - энергетическое 45
- Процесс Эйткена 81
- Равенство Парсеваля — Стеклова 69
- Равномерное приближение 69
- Разделенные разности 1-го порядка 64
  - — 2-го порядка 64
- Размерность линейного пространства 39
- Разностная производная 139
  - — левая 139
  - — правая 139
  - — центральная 139
  - схема 141
  - — Адамса 186
  - — аддитивная 256
  - — безусловно устойчивая (пример) 182
  - — двухслойная 181, 197
  - — Дугласа — Рекфорда 254
  - — квазистабильная 145
  - — консервативная 152
  - — корректная 142
  - — Кранка — Николсона 236
  - — крест 212
  - — локально-одномерная 258
  - — многошаговая 184
  - —  $m$ -го порядка точности 146
  - —  $m$ -шаговая ( $m \geq 1$ ) 185
  - — неустойчивая 142
  - — пятивенная 198
  - — одношаговая 181
  - — Писсена — Рекфорда 251
  - — предиктор — корректор (счет — пересчет) 180
  - — расщепления 258
  - —  $\rho$ -устойчивая 201
  - — Рунге — Кутта 179
  - — с весами 198
  - — с опережением 198
  - — симметричная 198
  - — условно устойчивая схема (пример) 182
  - — устойчивая 142, 143

## ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

269

- |                                       |                              |
|---------------------------------------|------------------------------|
| Разностная схема чебышев-             | Среднеквадратичное уклонение |
| ская итерационная 112                 | 68                           |
| — — чисто неявная 198                 | Сходимость разностной схемы  |
| — — Эйлера 141, 176                   | (со скоростью $O(h^m)$ ) 146 |
| — — экономичная 250                   | — с квадратичной скоростью   |
| — — явная 198Р                        | 134                          |
| Разностное уравнение линей-           | Уравнение теплопроводности   |
| ное с постоянными коэффи-             | 232                          |
| циентами 26                           | Устойчивость разностной схе- |
| — — $m$ -го порядка ( $m \geq 1$ ) 26 | мы с весами 182              |
| — — однородное 28                     |                              |
| Разностные неравенства 27             |                              |
| — формулы Грина 50                    | Формула Тейлора 74           |
|                                       | Формулы бегущего счета 125   |
| Сетка квадратная 212                  | Численное интегрирование 70  |
| — неравномерная 16                    | Число обусловленности 89     |
| — равномерная 16                      |                              |
| Сеточная функция 16, 138              | Шаблон 139                   |
| Сплайн порядка $m$ 66                 | — квадратурной формулы 74    |

## СПИСОК ОБОЗНАЧЕНИЙ

- $\omega_N = \{i; i = 0, 1, \dots, N\}$  — сетка с целочисленными узлами  
 $\overline{\omega}_h = \{x_i = ih, h = 1/N, 0 \leq i \leq N\}$  — равномерная сетка с шагом  $h$  на отрезке  $[0, 1]$   
 $h$  — шаг сетки  $\omega_h$   
 $y_i = y(x_i) = y(i)$  — значение сеточной функции в  $i$ -м узле сетки  
 $\widehat{\omega}_h$  — неравномерная сетка  
 $h_i = x_i - x_{i-1}$  — шаг неравномерной сетки  $\widehat{\omega}_h$ :  
 $\tilde{h}_i = \frac{1}{2} (h_i + h_{i+1})$   
 $v_{i_1 i_2} = v(x_{i_1}^1, x_{i_2}^2)$  — значение сеточной двумерной функции в узле  $(i, j)$   
 $v_{i_1 i_2}^n = v(x_{i_1}^1, x_{i_2}^2, t_n)$  — значение сеточной функции в узле  $(i, j)$  на  $n$ -м временном слое  
 $v_{ij}^{n+1} = \widehat{v}$  — значение сеточной двумерной функции в узле  $(i, j)$  на  $(n+1)$ -м временном слое  
 $\Delta y_i = y_{i+1} - y_i$  — правая разность в  $i$ -м узле  
 $\nabla y_i = y_i - y_{i-1}$  — левая разность в  $i$ -м узле  
 $\delta y_i = \frac{1}{2} (\nabla y_i + \Delta y_i)$  — центральная разность в  $i$ -м узле  
 $\Delta^2 y_{i+1} = \Delta(\nabla y_{i+1}) = \Delta(\Delta y_i)$  — разность второго порядка  
 $y_{x,i} = (y_{i+1} - y_i)/h$  — правая разностная производная в узле  $x_i$   
 $y_{x,i}^- = (y_i - y_{i-1})/h$  — левая разностная производная в узле  $x_i$   
 $y_{x,i}^o = (y_{i+1} - y_{i-1})/(2h)$  — центральная разностная производная в узле  $x_i$   
 $y_{xx,i} = (y_{i+1} - 2y_i + y_{i-1})/h^2$  — вторая разностная производная  
 $H$  — гильбертово пространство  
 $(y, v)$  — скалярное произведение элементов  $y, v \in H$ ,  $\|y\| = \sqrt{(y, y)}$   
 $E$  — единичный оператор  
 $A^*$  — оператор, сопряженный оператору  $A$   
 $A^{-1}$  — оператор, обратный оператору  $A$   
 $A > 0$  — положительный оператор  
 $A \geq 0$  — неотрицательный оператор  
 $A \geq \delta E$ ,  $\delta > 0$  — положительно определенный оператор  
 $\|y\|_A = \sqrt{(Ay, y)}$ ,  $y \in H$ , — энергетическая норма

СПИСОК ОБОЗНАЧЕНИЙ

274

Пространство сеточных функций:

$$\Omega_{N+1} = \{y_i, i = 0, \dots, N\}$$

$$\overset{\circ}{\Omega}_{N+1} = \{y_i, i = 0, \dots, N; y_0 = 0, y_N = 0\}$$

$y_i$  — функция из  $\overset{\circ}{\Omega}_{N+1}$

$$\Omega_N^+ = \{y_i, i = 0, 1, \dots, N - 1\}$$

$$\Omega_N^- = \{y_i, i = 1, 2, \dots, N\}$$

Скалярные произведения и нормы на сстке:

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h, \|y\| = \sqrt{(y, y)}$$

$$(y, v) = \sum_{i=1}^N y_i v_i h, \|y\| = \sqrt{(y, y)}$$

$$\|y\|_C = \max_{x_i \in \bar{\Phi}_h} |y(x_i)| = \max_{0 \leq i \leq N} |y(x_i)|$$