

Министерство образования и науки Российской Федерации
Федеральное агентство по образованию

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ

М.Ю.Белевич

Математические методы решения океанологических задач

Основные вычислительные идеи и методы



Санкт-Петербург
2008

УДК 551.509.313

ISBN 5-86813-010-3

Белевич М.Ю. Математические методы решения океанологических задач. Основные вычислительные идеи и методы. Учебное пособие. СПб., изд. РГГМУ, 2008.— 191 с.

Рецензенты: лаборатория волновых процессов СПО ГОИН (проф. И.Н.Давидан); проф. Б.А.Каган (СПбФ ИОРАН).

Учебное пособие призвано дать представление о методах решения дифференциальных уравнений в частных производных. Вводятся основные понятия из функционального анализа. Обсуждаются отдельные темы теории дифференциальных уравнений в частных производных, непосредственно связанные с используемыми разностными аппроксимациями. Рассматриваются основные вопросы решения уравнений разностными методами.

Курс адресован студентам и аспирантам, специализирующимся в области математического моделирования океанологических процессов.

The main goal of these lecture notes is to give fundamental knowledge of methods for solving the partial differential equations. Some basic notions of the functional analysis are introduced. Selected topics concerning the partial differential equations connected with difference approximations to be used are discussed. Most important questions of the finite-difference methods are considered.

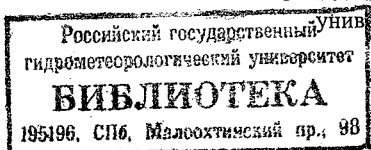
The book is addressed to graduate and post-graduate students who specialize in mathematical modelling of oceanographical processes.

В $\frac{-1}{2K2(03)-99}$ Без объявления

ISBN 5-86813-010-3

© М.Ю. Белевич, 2008

© Российский государственный гидрометеорологический университет (РГГМУ), 2008



Содержание

Предисловие	6
Часть I. Общие вопросы	8
1. Отображения и векторные пространства	9
1.1. Введение	9
1.1.1. Жить — значит моделировать	9
1.1.2. Универсальные средства описания	10
1.1.3. Меры и плотности	12
1.1.4. Примеры моделей	13
1.2. Отображения	15
1.3. Векторное пространство	17
1.4. Базис и размерность векторного пространства	23
2. Линейные операторы	27
2.1. Линейные операторы	27
2.2. Матрица линейного оператора	29
2.3. Векторное пространство линейных операторов	31
3. Метод последовательных приближений	33
3.1. Идея метода	33
3.2. Метрика	35
3.3. Норма вектора	37
3.4. Сходимость и банаховы пространства	39
3.5. Эквивалентные нормы	42
4. Сходимость метода последовательных приближений	43
4.1. Норма линейного оператора	43
4.2. Сходимость метода последовательных приближений	44
4.2.1. Существование обратного оператора	45
4.2.2. Сходимость приближенных решений	45
4.3. Пример приложения метода последовательных приближений	48
4.4. Спектральный радиус оператора	49
4.5. Задача на собственные значения	50

5. Гильбертово пространство и ряд Фурье	52
5.1. Гильбертово пространство	52
5.2. Ортогональность	54
5.3. Ортогонализация	55
5.4. Ряд Фурье	57

Часть II. Простейшие непрерывные модели **60**

6. Уравнения 1-го порядка	61
6.1. Однородное уравнение	61
6.2. Задача Коши для уравнения 1-го порядка	66
6.3. Неоднородное и квазилинейное уравнения	69
7. Системы дифференциальных уравнений	73
7.1. Общие соображения	73
7.2. Линейная система с двумя независимыми переменными .	74
7.2.1. Замена переменных	75
7.2.2. Исследование типа системы	76
8. Некоторые классические задачи	78
8.1. Задача Коши для волнового уравнения	79
8.2. О постановке задач математической физики	85

Часть III. Разностные методы **88**

9. Методы построения разностных операторов	89
9.1. Введение	89
9.2. Дискретное представление функций	90
9.3. Методы построения разностных операторов	96
10. Методы построения разностных схем	103
10.1. Метод неопределенных коэффициентов	103
10.2. Интегральный метод построения разностных схем	105
10.3. Нестационарные задачи	113

11. Основы теории разностных схем	116
11.1. Структура разностного оператора	116
11.2. Сходимость, аппроксимация, устойчивость	119
12. Устойчивость разностных схем	125
12.1. Введение	125
12.2. Примеры использования прямого метода	127
12.3. Спектральный метод (метод Неймана)	129
12.4. Примеры использования спектрального метода	131
12.5. Физический смысл условий устойчивости	136
13. Ошибки разностных схем	139
13.1. Амплитудные и фазовые ошибки	139
13.2. Природа амплитудной и фазовой ошибок	143
14. Организация вычислений	146
14.1. Понятие о методах расщепления	146
14.2. Вычислительные граничные условия	150
15. Обзор методов решения систем линейных алгебраических уравнений	154
15.1. Итерационные методы	154
15.1.1. Общие сведения	154
15.1.2. Обзор основных методов	155
15.2. Точные методы	158
15.2.1. Метод Гаусса	158
15.2.2. Методы прогонки	159
15.2.3. Циклическая прогонка	161
Литература	164
Часть IV. Практические занятия	166
Предметный указатель	188

Предисловие

Цель настоящего учебного пособия — дать студентам представление о современных методах решения дифференциальных уравнений в частных производных — основе подавляющего большинства математических моделей.

За последние десятилетия математизация охватила многие прикладные дисциплины, в том числе и гидрометеорологию. Если в недавнем прошлом большинство гидрометеорологов обходилось общими курсами математического анализа, восходящими к Коши и Лапласу, то в настоящее время этого уже явно недостаточно. Широкое внедрение ЭВМ привело к повсеместному использованию математических моделей, создание которых требует более глубоких математических знаний. Данная книга призвана в какой-то мере восполнить пробелы в математическом образовании студентов-океанологов и дать им общее представление о методах решения задач, возникающих в процессе математического моделирования.

Первая часть целиком посвящена избранным темам из функционального анализа и является фундаментом всего следующего материала. Здесь вводятся основные понятия и определения, используемые затем на протяжении всего курса. При этом я старался не упускать из виду основной цели — решение дифференциальных уравнений — и новые понятия вводить лишь тогда, когда это продиктовано логикой исследования. Не все введенные понятия используются впоследствии одинаково широко, но в любом случае, я надеюсь, они обогатят словарь студента и помогут ему читать специальную литературу. В этой же части рассматривается один из фундаментальных методов решения уравнений — метод последовательных приближений. Он важен не только сам по себе, но и как канва, в которую удастся органично вплести часть нового материала.

Во второй части обсуждаются отдельные вопросы теории дифференциальных уравнений в частных производных, которые непосредственно связаны с используемыми далее разностными аппроксимациями. Я думаю, этот материал позволит студентам лучше понять, что нужно требовать и что можно ожидать от конструируемых разностных схем. Третья часть посвящена разностным методам, которые наиболее широко используются в численном моделировании. В последней части

собраны упражнения и задачи, решаемые со студентами на практических занятиях.

В заключение я хочу поблагодарить всех своих друзей и коллег, кто так или иначе способствовал написанию этой книги. Прежде всего это С.Фокин, частые дискуссии и споры с которым помогли мне многое понять и продумать. Р.Романов, В.Горчаков, А.Сафрай прочитали первоначальный вариант рукописи и сделали немало полезных и ценных замечаний.

М.Белевич

Хоть оно непонятно... а все-таки слово этакое... на улице ты его не услышишь... Непонятно оно, а чувствуешь, что это слово для души.

М.Горький. Дело с застежками.

Часть I. Общие вопросы

1. Отображения и векторные пространства

1.1. Введение

1.1.1. Жить — значит моделировать

Задача живого организма — выжить, а для этого необходимо правильно взаимодействовать с окружающей средой и адекватно реагировать на ее изменения. Намеревающийся выжить организм должен, таким образом, придерживаться некой оптимальной стратегии поведения, для чего ему необходимо представлять, что его окружает, и уметь прогнозировать возможные изменения этого окружения.

Природу не спросишь, как она собирается себя вести, а прогнозы делать нужно. Живые организмы научились обходить эту трудность: вопросы адресуются не самой природе, а ее суррогату (заменителю, модели), который создается организмами именно с такой целью¹. Пока модель явления удовлетворительно его описывает, ею пользуются для выработки оптимальной стратегии. В противном случае модель подправляют или, наконец, строят другую. Те организмы, которым не удалось научиться создавать модели окружающей среды и работать с ними, вымерли.

Таким образом, без моделирования не обойтись: оно всеобъемлюще и всепроникающе. Это не только способ познания окружающей действительности, но и способ самого существования. Мы настолько свыклись с этим, что постоянное моделирование воспринимается как нечто само собой разумеющееся. Человеческое сознание наполнено всевозможными моделями разного уровня сложности. Без особой натяжки можно сказать, что человек живет в мире моделей, им же самим созданных. Мы постоянно используем одни модели, модернизируем другие, создаем третьи. Сама окружающая действительность воспринимается нами в терминах имеющихся моделей, так что всякая серьезная замена одних моделей другими порождает большие психологические проблемы, справиться с которыми под силу не каждому.

¹Казалось бы, здесь можно сделать исключение для ученых-натуралистов, естествоиспытателей, всех тех, кто, как раз и задает вопросы Природе. Однако даже у них профессиональная деятельность далеко не исчерпывает всех жизненных ситуаций, так что ответы Природы могут лишь способствовать усовершенствованию моделей, но не заменить их.

Простейшие модели — статические картины окружающего мира. Помимо того что хранится у нас в памяти, к ним следует отнести карты, схемы, фотографии, зарисовки и пр. Без помощи таких моделей любое действие нам пришлось бы начинать с разведки, каждый день был бы первым в жизни.

Более сложные модели — динамические, учитывающие возможную эволюцию окружающего нас мира. Без них нам не поймать брошенный предмет, не попасть в движущуюся мишень, не спланировать день и т.п.

Все эти модели совершенно необходимы. Однако они способны воспроизводить только качественную сторону изучаемого процесса, почему их и называют *качественными*. Многие тысячи лет человек пользуется такими моделями и с их помощью не только выжил, но и занял в природе лидирующее положение. Вместе с тем качественные модели не свободны от недостатков, главным из которых является невозможность точного описания различных сторон процесса. Знание лишь качественных свойств явлений или процессов (близко—далеко, быстро—медленно, тепло—холодно) может привести к значительным ошибкам при выборе оптимальной стратегии. Даже один и тот же человек по-разному оценивает, например, температуру окружающего воздуха в зависимости от настроения, самочувствия и т.п. Что уж говорить о разных людях.

1.1.2. Универсальные средства описания

Потребность в объективных оценках привела к поиску универсальных средств описания, не зависящих от времени, места или наблюдателя. Таким универсальным средством стали числа. Число само по себе, разумеется, ничего не может сказать о физическом процессе или объекте, но вкупе с заранее выбранными эталонами позволяет указать, во сколько раз рассматриваемое свойство у изучаемого объекта больше или меньше, чем у эталона. Переход к числам происходит по следующей схеме:

- 1) выбирается эталонный (отсчетный, базисный) объект,
- 2) рассматриваемому свойству эталона ставится в соответствие число 1 (т.е. полагается, что значение данного свойства эталона равно

единице),

- 3) сравнивая данное свойство у всех остальных объектов из некоторого множества с эталонным, мы ассоциируем с этими объектами те или иные числа в зависимости от результатов сравнения.

Первые два пункта есть не что иное, как выбор системы единиц измерения, а третий — собственно измерение. Таким образом, с каждым свойством физического тела путем измерения мы связываем некоторое число имеющее смысл количества этого свойства в терминах эталонного свойства. Именно поэтому физические величины размерны. Даже наша обиходная речь насыщена различными числами с указанием соответствующих эталонов. Каждый может привести множество таких примеров (*приведите их*).

Модели, позволяющие работать с этими числовыми эквивалентами свойств тел, называются *количественными* или *математическими* моделями, поскольку именно математика и есть та дисциплина, что работает с числами. С помощью математических моделей можно предсказывать конкретные численные значения характеристик явления, и эти значения позволят одинаково описать явление любому наблюдателю, независимо от места, времени или национальных черт. Таким образом, ясно, что модель погоды, основанная на народных приметах, является качественной. Модели же, используемые в гидрометцентрах и позволяющие предсказывать численные значения характеристик погоды (температуры, скорости ветра и т.п.), относятся к количественным моделям.

В общем случае математическая модель — это совокупность уравнений (алгебраических, дифференциальных, интегральных), описывающих процессы, происходящие в моделируемом явлении. Каждое такое уравнение само по себе есть модель процесса более низкого уровня. Модели же сложных явлений часто строятся, как из кирпичиков, из моделей более простых.

Результатом математического моделирования являются формулы (если удается получить аналитическое решение уравнений модели), позволяющие рассчитать характеристики моделируемого процесса, или таблицы значений этих характеристик (если аналитическое решение получить не удастся либо оно неудобно). Чаще всего встречается именно последний случай, когда решение можно получить лишь в виде та-

блицы с помощью каких-либо вычислительных процедур. Такого рода модели носят специальное название *численных моделей*. Специфика этих моделей состоит в том, что они позволяют получить принципиально приближенное решение, что связано с дискретностью вычислительных устройств, ограниченным объемом памяти, конечным быстродействием и т.п.

К основным этапам математического моделирования можно отнести:

- 1) формализацию физической задачи, т.е. перевод ее в математическую форму,
- 2) решение математической задачи с помощью подходящих математических средств,
- 3) интерпретацию результата в исходных физических терминах.

Наша основная задача (т.е. цель данного курса) связана со вторым этапом — научиться строить и исследовать алгоритмы получения численного решения уравнений модели.

1.1.3. Меры и плотности

Переход к числам и далее к численным моделям тесно связан с измерениями. Все измеряемые свойства тел описываются так называемыми *мерами* или *интегральными параметрами*. Масса, длина, объем, заряд, количество тепла — все это примеры измеряемых (интегральных) параметров. Они характеризуют тело или его фрагмент в целом.

Есть физические задачи, в которых достаточно знать только значение интегральных параметров тела. Сами тела мыслятся как физические точки, снабженные указанными параметрами. Изучается эволюция этих параметров при перемещении тел в пространстве и взаимодействии их друг с другом. Таковы задачи механики. В случае механики сплошных сред (а именно такие задачи мы будем рассматривать в данном курсе) интерес представляет, кроме всего прочего, еще и распределение изучаемых свойств в самом теле.

Сплошная среда появляется в результате применения *гипотезы сплошности* к достаточно большой совокупности частиц, микроскопи-

ческих по сравнению с масштабами изучаемого явления. Каждое тело, рассматриваемое как сплошная среда, включает по определению бесконечно много точек. Дискретное физическое тело (например, совокупность молекул или звездное скопление) заменяется континуумом точек, что дает возможность описывать свойства тела гладкими функциями (т.е. имеющими нужное число производных), определенными на этом континууме. Индивидуальность реальных частиц, составляющих тело, при этом теряется, и эволюция их свойств, таким образом, не изучается.

Любой интегральный параметр всегда характеризует континуум точек. Значение его в точке равно нулю. Поэтому для математического описания свойств тел в точках вместо интегральных параметров приходится вводить их *плотности* или *дифференциальные (локальные) параметры*. Любая плотность определяется как предел отношения двух интегральных параметров, т.е. по существу является производной одной меры по другой. Вот примеры локальных параметров:

- скорость (метр/секунда),
- давление (ньютон/метр²),
- плотность массы (килограмм/метр³),
- удельный объем (метр³/килограмм).

В этих примерах в скобках приведены размерности локальных параметров. Все они записываются как отношения единиц измерения (эталонов) соответствующих интегральных параметров. В отличие от измеряемых интегральных параметров среды плотности не измеряются в принципе. Они всегда вычисляются.

Связав с каждой точкой изучаемого объекта некоторое число, характеризующее какое-то его свойство, мы получим *поле локального параметра (поле плотности)*, соответствующего этому свойству. Таковы, например, поля скалярных величин — температуры, давления, плотности массы.

1.1.4. Примеры моделей

Задача математического моделирования состоит в прогнозировании с помощью имеющейся модели хода развития интересующих нас собы-

тий, основываясь на уже известных событиях и каких-то априорных соображениях. Математическая модель A , таким образом, ставит в соответствие имеющемуся набору данных (свойствам объекта, выраженным в числовой форме) g другой набор данных (свойств) f — прогноз. Формально это обычно записывается так:

$$Af = g. \quad (1.1)$$

Если рассматривается одно свойство, выражение (1.1) представляет собой одно уравнение. Если изучается несколько свойств — имеем систему уравнений.

Нетрудно привести примеры. Далее мы будем рассматривать простейшие модели, описываемые одним, реже — двумя уравнениями, с тем чтобы сложность модели не затемняла существа метода решения. Такие уравнения вам знакомы из курса математической физики. Это, например, уравнение теплопроводности

$$\partial_t T = \mu \Delta T + Q,$$

волновое уравнение

$$\partial_{tt} U = \alpha \partial_{xx} U$$

или уравнение переноса некоторой величины C со скоростью v вдоль оси x :

$$\partial_t C + v \partial_x C = 0.$$

Если обозначить через A те действия

$$A = \begin{cases} \partial_t - \mu \Delta, \\ \partial_{tt} - \alpha \partial_{xx}, \\ \partial_t + v \partial_x, \end{cases}$$

которые производятся над неизвестной функцией f (в наших примерах — это T , U или C), уравнения примут вид (1.1). Правая часть g в примерах равна Q или 0 .

Наша основная задача — понять в каком случае уравнение (1.1) имеет решение, и если это решение существует, научиться его строить, т.е. находить f по известным A и g .

Сначала посмотрим, каким общим требованиям должно отвечать A , чтобы решение существовало. Дело в том, что возможна ситуация,

когда задача сформулирована так, что ее решение отсутствует. Легко привести пример. Пусть требуется найти вещественный корень уравнения $x^2 + 1 = 0$. Известно, что на множестве вещественных чисел решение такой задачи не существует (*а на каком множестве существует?*).

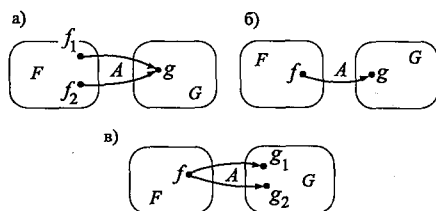


Рис. 1. Варианты соответствий элементов множеств: а) — различным элементам из F допускается ставить в соответствие один и тот же элемент из G ; б) — взаимнооднозначное соответствие; в) — одному элементу из F допускается ставить в соответствие совокупность элементов из G .

1.2. Отображения

Будем рассматривать значения свойств f и g как элементы некоторых множеств F и G . Например, если, скажем, g — плотность массы и $g \in G$, то G — это множество всех мыслимых значений плотности массы. Если же под f подразумевается концентрация, то в качестве F разумно рассматривать множество всех возможных значений концентрации, и т.д. В таком случае, модель A можно считать правилом, устанавливающим соответствие между элементами множеств F и G . Возможные варианты соответствий показаны на рис.1.

Если правило A включает только варианты соответствий «а» и «б» и не включает «в», оно называется *отображением* множества F в множество G . В этом случае мы пишем $A: F \rightarrow G$ и говорим, что A есть отображение F в G . Желая к тому же указать связь элементов множеств, мы пишем $f \mapsto g$, $f \in F$, $g \in G$ и говорим, что элемент f из F отображается в элемент g из G . Если A включает лишь соответствия

типа «б», оно называется *взаимнооднозначным отображением*.

Обратите внимание на то, что в определении специально оговаривается единственность элемента g . Это значит, что если некоторое правило (вариант «в») ставит в соответствие элементу f сразу несколько элементов $(g_1, g_2, \dots)$, оно не является отображением. В свою очередь, сразу несколько элементов $(f_1, f_2, \dots)$ могут отображаться в один и тот же элемент g (вариант «а»).

Множество F , для любого элемента которого отображение A указано, мы называем *областью определения* отображения A , обозначаем $D(A)$ и говорим, что A *определено* на множестве $D(A) = F$. Если $f \in D(A)$, то Af называется *образом элемента f* . Аналогично определяется образ множества. Пусть P является подмножеством $D(A)$. Множество образов всех элементов из P образует *образ множества P* (рис.2). Образ области определения $D(A)$ называется *множеством значений отображения A* и обозначается через $R(A)$.

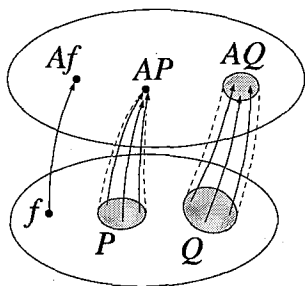


Рис. 2. Образы элемента и множества.

Можно рассмотреть отображение I , которое ставит в соответствие элементу f тот же самый элемент f , т.е. $If = f$. Такое отображение называется *единичным*, и для него $D(I) = R(I)$.

Если имеются два отображения A и B , таких, что $A: M \rightarrow N$ и $B: N \rightarrow P$, можно определить новое отображение $C: M \rightarrow P$, которое называется *композицией* отображений A и B и обозначается через $B \circ A$ (см. рис. 3). Если $t \in M$, то A отображает его в элемент $At \in N$, а B , в свою очередь, переводит его в элемент $B(At) \in P$. Таким образом, $(B \circ A)t = B(At)$.

Пусть Q является подмножеством $R(A)$. Множество элементов $f \in D(A)$, таких, что $Af \in Q$, называется *прообразом множества Q* и обозначается $A^{-1}Q$. Если в качестве множества Q взять единственный элемент g , то $A^{-1}g$ будет обозначать прообраз g . В общем случае прообразом элемента служит некоторое подмножество области определения отображения, и символ A^{-1} не обозначает никакого отображения.

Существует, однако, уже упоминавшийся важный класс отображений, ставящих в соответствие разным элементам из области определе-

ния различные элементы из области значений (вариант «б» на рис.1). В этом случае каждый элемент $g \in R(A)$ имеет прообраз, состоящий из единственного элемента $A^{-1}g \in D(A)$. Такое отображение A , как уже говорилось, называется взаимнооднозначным, и в этом случае соответствие A^{-1} может рассматриваться как другое взаимнооднозначное отображение, называемое *обратным* к A . Их композиция есть единичное отображение $A \circ A^{-1} = A^{-1} \circ A = I$.

Эти отображения имеют непосредственное отношение к нашей основной задаче — решению уравнения (1.1). Уравнение (1.1) имеет единственное решение только, если отображение A является взаимнооднозначным. Решение уравнения сводится тогда к построению обратного отображения A^{-1} . Действительно, применим обратное отображение A^{-1} к уравнению (1.1)

$$A^{-1}(Af) = A^{-1}g \Rightarrow (A^{-1} \circ A)f = A^{-1}g$$

и вычислим композицию в левой части $(A^{-1} \circ A)f = If = f$. В результате получаем решение задачи

$$f = A^{-1}g.$$

Таким образом, находя обратное отображение, мы решаем задачу.

1.3. Векторное пространство

Для того чтобы решить физическую задачу, нам надо сформулировать ее в виде задачи математической, выбрав систему единиц и выразив рассматриваемые свойства тел числами относительно этой системы единиц. К физической постановке задачи мы возвращаемся, когда математическая задача решена и требуется интерпретировать полученный результат.

Числа удобны. Нам известно, как их складывать, умножать и пр. Если какое-либо свойство тела можно охарактеризовать одним числом, то работа с

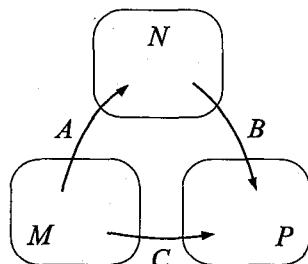


Рис. 3. Композиция отображений.

ним также относительно проста. Скажем, в некоторой точке тела температура, которой поставлено в соответствие число T , увеличилась на величину, с которой связано число ΔT , тогда результирующую температуру мы ассоциируем с числом $(T + \Delta T)$. Если исходное значение температуры уменьшилось в α раз, то в этом случае с результатом мы связываем число $\frac{1}{\alpha}T$. Используя, далее, выбранную систему единиц, мы всегда можем указать, какой же стала температура в данной точке.

Работать с другими свойствами может оказаться сложнее. Пусть, например, требуется описать движение точки среды, для чего надо проследить эволюцию ее скорости. Скорость есть предел отношения перемещения точки в пространстве к соответствующему временному интервалу, и для того чтобы ее однозначно определить, мало указать величину v этого отношения, надо еще сообщить направление. Можно, например, задать вертикальный α и горизонтальный β углы относительно некоторого базисного направления, под которыми это перемещение происходит. Таким образом, при рассмотрении скорости точки с ней связывается не одно число, а совокупность чисел (v, α, β) . То же происходит, если в каждой точке определена совокупность свойств и ассоциированных с ними чисел.

В общем случае точке континуума ставится в соответствие совокупность чисел $(a_1, \dots, a_n)$, и нужно придумать способ работы с такими объектами. Во-первых, чтобы не перепутать эти числа, следует договориться о порядке их следования. Другими словами, каждой точке объекта надо ставить в соответствие *упорядоченную совокупность чисел* $(a_1, \dots, a_n)$. Во-вторых, следует указать, что понимать под *суммой* таких совокупностей, например, $(a_1, \dots, a_n) + (b_1, \dots, b_n)$, а также под *произведением совокупности на какое-либо число* α , т.е. придумать правила, указывающие, как суммировать элементы множеств и умножать их на числа.

Чем руководствоваться при выборе таких правил? Есть два абстрактных соображения: простота использования и непротиворечивость. С одной стороны хорошо бы придумать правила, которые позволяли бы обращаться с элементами множеств также просто, как с отдельными числами. С другой стороны надо, чтобы результат применения этих правил существовал, т.е. принадлежал тому же множеству, что и исходные элементы. Физическая интерпретация таких правил может быть различной и обычно диктуется решаемой задачей.

Поясним все это двумя примерами. Пусть рассматриваемое множество — это совокупность точек плоскости. Отметим на этой плоскости любые две точки p и q . Какую точку плоскости следует ассоциировать с суммой $p + q$? Без привлечения каких-либо внешних соображений ответ на этот вопрос дать нельзя: годится любая точка. Внешним же мотивом обычно является желание максимально упростить какую-то конкретную задачу.

Например, физическая задача сложения двух сил или скоростей принимает наиболее простой вид, если сумму элементов множества интерпретировать следующим (хорошо известным) способом. Выберем на плоскости некоторую точку, назовем ее нулем и обозначим знаком 0 . После этого с любой точкой плоскости свяжем вектор (такой, каким его представляют в школе: направленным отрезком, стрелочкой, идущей из точки 0 в любую точку плоскости). Теперь множество точек плоскости можно трактовать как множество векторов, имеющих общее начало в точке 0 .

Пусть p и q — два произвольных вектора. Назовем их суммой вектор r , направленный по диагонали параллелограмма, построенного на этих векторах, длина которого равна длине диагонали, т.е. $r = p + q$ (рис.4 б). Пусть, далее, две точки p и q (или, что то же, два вектора) лежат на одной прямой с точкой 0 . Допустим, каким-то образом мы умеем измерять длину. Обозначим длину вектора через ℓ . Если $\ell_p/\ell_q = \alpha$, то будем говорить, что $p = \alpha q$, когда p и q лежат по одну сторону от точки 0 , и $p = -\alpha q$, когда они лежат по разные стороны (рис.4 а). Таким образом, мы указали геометрическую интерпретацию умножению вектора на число.

Приведенная интерпретация суммы и умножения на число хороша для определенного круга физических задач, но отнюдь не единственна и не универсальна. Пусть, например, изучается движение обода колеса, который ассоциируется с окружностью. Множеством точек в этом случае является окружность, т.е. подмножество только что рассмотренного множества, и старый способ интерпретации суммы и произведения на число перестает работать. Действительно, выберем на окружности точку 0 , а также точки p и q . Легко видеть, например, что сумма любых, отличных от нуля p и q , определенная согласно старому правилу, дает точку, не лежащую на окружности. Требуется новая интерпретация суммы, подходящая для данной конкретной задачи (*какая?*).

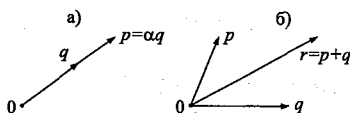


Рис. 4. Действия над векторами.

Мало, однако, просто установить соответствие между слагаемыми и их суммой, надо также указать, как эту сумму вычислять, т.е. опреде-

лить свойства операций суммирования и умножения на число. К таким свойствам удобно в частности относить коммутативность, дистрибутивность и ассоциативность сложения и умножения, и ряд других. Эти свойства формулируются в виде аксиом, которым указанные операции должны удовлетворять. Вот эти аксиомы:

- 1) существует элемент 0, называемый *нулем*, такой, что $f + 0 = f$;
- 2) для любого f существует *противоположный* элемент $(-f)$ такой, что $f + (-f) = 0$;
- 3) умножение на единицу $1 \cdot f = f$;
- 4) коммутативный закон
$$f + g = g + f;$$
- 5) ассоциативные законы
$$(f + g) + h = f + (g + h),$$
$$(\alpha\beta)f = \alpha(\beta f);$$
- 6) дистрибутивные законы
$$\alpha(f + g) = \alpha f + \alpha g,$$
$$(\alpha + \beta)f = \alpha f + \beta f.$$

В этих аксиомах $1, \alpha, \beta$ — числа. Хотя количество аксиом и внушительно, они являются лишь формализацией хорошо известных и часто используемых свойств определяемых операций (типа: от перемены мест слагаемых сумма не меняется — аксиома 4, или общий множитель можно выносить за скобку — аксиомы 7 и 8) и нужны нам для удобства проведения вычислений.

Множество, для элементов которого определены операции сложения и умножения на число, обладающие указанными свойствами, мы будем называть *векторным пространством*. Формально же все сказанное выше определяется так.

Пусть V — непустое множество и f, g, h — некоторые его элементы. Это множество называется *векторным* (или *линейным*) *пространством*, если указано правило, по которому любым двум элементам из V ставится в соответствие третий элемент из V , называемый суммой

элементов, и правило, по которому любому элементу из V и любому числу ставится в соответствие элемент из V , называемый произведением элемента на число, и эти правила подчиняются аксиомам (1)–(8). Элементы $f, g, h, \dots \in V$ называются *точками* (или *векторами*).

Примеры.

- 1) $\mathbb{R}^1$ — множество вещественных чисел. Выполнение аксиом (1)–(8), для стандартным образом определенных сложения и умножения, нетрудно проверить. $\mathbb{R}^1$ — это векторное пространство, точками или векторами которого служат вещественные числа. Кстати, если «разместить» все вещественные числа на прямой (т.е. выбрать нулевую точку, а точку p связать с числом α , если расстояние от 0 до p равно α , то и здесь векторы можно представить в виде стрелочек, направленных из точки 0 в точку p .
- 2) $\mathbb{R}^n$ — множество, элементом которого является любая упорядоченная² совокупность из n чисел $(x^1, x^2, \dots, x^n)$. Число x^i (значок над x — не степень, а индекс) будем называть i -й компонентой элемента. Определим сложение элементов $\mathbb{R}^n$ и умножение их на число покомпонентно: если $f = (f^1, f^2, \dots, f^n)$ и $g = (g^1, g^2, \dots, g^n)$ — элементы $\mathbb{R}^n$ и α — число, то

$$f + g = (f^1 + g^1, f^2 + g^2, \dots, f^n + g^n)$$

и

$$\alpha f = (\alpha f^1, \alpha f^2, \dots, \alpha f^n).$$

Нулевым элементом назовем элемент $(0, 0, \dots, 0)$. Выполнение аксиом (1)–(8) легко проверяется. Таким образом, и множество $\mathbb{R}^n$ с определенными операциями является векторным пространством.

Сделаем попутно небольшое добавление к примеру 2, показывающее связь множеств $\mathbb{R}^1$ и $\mathbb{R}^n$. Пусть P и Q — два произвольных множества, состоящих из элементов p_i и q_j соответственно. Можно образовать новое множество, элементами которого будут всевозможные упорядоченные пары (p_i, q_j) . Это новое множество

²Как уже говорилось, упорядоченность означает, что элементы $\mathbb{R}^n$ различаются не только численными значениями x^i , но и порядком их расположения.

называется *прямым произведением* множеств P и Q и обозначается через $P \times Q$.

Пусть теперь V и W — векторные пространства. Прямое произведение $V \times W$ можно также превратить в векторное пространство, если сложение и умножение на число определить следующим образом:

$$(f, g) + (p, q) = (f + p, g + q),$$

$$\alpha(f, g) = (\alpha f, \alpha g).$$

Здесь $f, p \in V$, $g, q \in W$, $(f, g), (p, q) \in V \times W$ и α — вещественное или комплексное число.

Очевидно, пространство $\mathbb{R}^n$ можно трактовать как прямое произведение n векторных пространств $\mathbb{R}^1$

$$\mathbb{R}^n = \underbrace{\mathbb{R}^1 \times \mathbb{R}^1 \times \dots \times \mathbb{R}^1}_n.$$

- 3) $\mathbb{C}$ — множество комплексных чисел $(\alpha + i\beta)$, где $\alpha, \beta \in \mathbb{R}^1$, а $i \equiv \sqrt{-1}$. Сложение и умножение на число определим следующим образом:

$$(\alpha + i\beta) + (\gamma + i\delta) = ((\alpha + \gamma) + i(\beta + \delta)),$$

$$\varepsilon(\alpha + i\beta) = ((\varepsilon\alpha) + i(\varepsilon\beta)).$$

Нулевым назовем элемент $(0 + i0)$. Аксиомы (1)–(8) выполняются и здесь, откуда следует, что и $\mathbb{C}$ с указанными операциями сложения и умножения на число также является векторным пространством.

- 4) Множество $n \times n$ матриц также будет векторным пространством, если сумму матриц и умножение матрицы на число определить так, как это делается в линейной алгебре, т.е. покомпонентно. Нулевым элементом этого пространства будет нулевая матрица, все элементы которой равны нулю.

Можно привести и множество других примеров.

Если некоторое подмножество S векторного пространства V само образует векторное пространство, то оно называется *подпространством* векторного пространства V . Другими словами, подмножество S

таково, что применение к его элементам операций сложения и умножения на число не выводит за границы этого подмножества. Например, любая плоскость, проходящая через точку 0 (почему именно такая?) в $\mathbb{R}^3$, является подпространством $\mathbb{R}^3$, так как сама является векторным пространством $\mathbb{R}^2$. Аналогично любая прямая, проходящая через точку 0, является подпространством $\mathbb{R}^3$. Кроме того, данная прямая является подпространством тех плоскостей $\mathbb{R}^2$, в которых она лежит и через точку 0 которых проходит.

1.4. Базис и размерность векторного пространства

Будем далее полагать, что на множестве значений интересующих нас свойств точки объекта введена структура векторного пространства³. Обозначим такое векторное пространство буквой V . Совокупность свойств f есть вектор (точка, элемент)⁴ этого пространства: $f \in V$.

Если нас интересует одно свойство f физического тела, описываемое скалярной величиной, то для сопоставления f числа достаточно выбрать один базисный элемент e (базисный вектор) с эталонным значением этого свойства. Такому значению ставится в соответствие число 1, а остальным значениям f — другие числа α в соответствии с формулой $f = \alpha e$. Такая связь называется *линейной*.

Если свойство f описывается тензорной величиной ранга $r > 0$, то ему ставится в соответствие уже не одно число, а упорядоченная совокупность n^r чисел, где n — геометрическая размерность⁵ пространства

³В простейшем случае речь идет о скалярных величинах (тензорах нулевого ранга), например, о температуре или плотности в данной точке тела. Более сложные примеры — векторные величины (тензоры первого ранга), например, положение точки тела в пространстве или ее скорость. Далее идут свойства, описываемые тензорами более высоких рангов.

⁴В силу исторических причин со словом «вектор» связано два понятия. Одно частное — тензор первого ранга, и другое — общее — элемент векторного пространства. Вектор как элемент векторного пространства не обязательно является тензором первого ранга, но, разумеется, может быть и им. Обычно путаницы не возникает, и из контекста видно о чем идет речь. В крайнем случае, можно последовательно использовать синонимы (линейное пространство вместо векторного и точка или элемент вместо вектор), исключаяющие указанную двусмысленность.

⁵Опять же в силу исторических причин, в физике слово «размерность» многозначно. С одной стороны, размерностью физической величины (или формулой

мест.

В общем случае число базисных элементов должно равняться минимальному числу эталонов, достаточных для описания изучаемых свойств тела. Если для этих целей требуется n эталонов, т.е. совокупность $\{e_i\}_{i=1}^n$ базисных векторов, то числовой характеристикой указанной совокупности свойств в каждой точке тела будет упорядоченная совокупность n чисел.

Введение структуры векторного пространства на множестве приводит к тому, что все элементы множества оказываются связанными друг с другом соотношениями вида

$$\alpha f + \beta g + \gamma h + \dots = 0, \quad (1.2)$$

где $\alpha, \beta, \gamma, \dots$ — числа, а $f, g, h, \dots$ — векторы. Сумма, стоящая в левой части равенства (1.2), называется *линейной комбинацией* векторов $f, g, h, \dots$. Если равенство (1.2) выполняется при ненулевом векторе, скажем αf , то вектор f может быть выражен через остальные векторы рассматриваемой линейной комбинации

$$f = -\frac{\beta}{\alpha}g - \frac{\gamma}{\alpha}h - \dots$$

Векторы, входящие в такую комбинацию, называются *линейно зависимыми*. На возможности подобной записи основано сопоставление вектору характеризующей его совокупности чисел: если некоторый вектор h линейно зависит от базисных векторов $\{e_i\}$, то он может быть записан в виде линейной комбинации этих векторов

$$h = \sum_i h^i e_i. \quad (1.3)$$

размерности) называется соотношением, определяющее связь между единицей измерения этой величины и используемыми эталонами. Она является функцией, определяющей, во сколько раз изменится численное значение этой величины при переходе от исходной системы единиц измерения к другой, т.е. от одних эталонов к другим.

С другой стороны, под размерностью векторного пространства (см. ниже стр.26) понимается минимальное число эталонов (базисных элементов), достаточное для описания всех элементов этого пространства (в смысле указания численных значений, однозначно характеризующих эти элементы).

В первом случае, термин относится к физической величине, т.е. к некоторому свойству тела или его точки (масса, скорость, плотность энергии, ...). Во втором — речь идет о множестве значений физической величины (множество значений массы тел, множество значений скорости изменения места точки тела и т.п.)

Таким образом, базисных векторов в совокупности $\{e_i\}$ должно быть столько, чтобы все остальные векторы линейно от них зависели.

Кроме того, базисные векторы (т.е. эталонные элементы) должны обеспечивать взаимнооднозначную связь векторов (в том числе и самих себя) с совокупностями характеризующих их чисел.⁶ Какие ограничения накладывает это требование на выбор базисных векторов?

Допустим, свойства f характеризуются эталонами неоднозначно, т.е. существует два (или более) разных набора чисел $\{\alpha_i\}$ и $\{\beta_i\}$ таких, что

$$f = \sum \alpha_i e_i = \sum \beta_i e_i, \quad (1.4)$$

т.е. имеется хотя бы одно значение $i = k$, при котором $\alpha_k \neq \beta_k$. В этом случае разность представлений (1.4) такова

$$0 = \sum (\alpha_i - \beta_i) e_i = \sum \gamma_i e_i, \quad \gamma_i \equiv \alpha_i - \beta_i, \quad (1.5)$$

и хотя бы для одного значения индекса $i = k$ коэффициент $\gamma_k \neq 0$. Тогда значение k -го эталона может быть выражено через значения остальных эталонов

$$e_k = - \sum_{i \neq k} \frac{\gamma_i}{\gamma_k} e_i, \quad (1.6)$$

и это говорит о том, что эталоны являются зависимыми друг от друга, точнее — линейно зависимыми, так как связь (1.6) является линейной.

Очевидно, для обеспечения указанной однозначности описания совокупность эталонов, т.е. базисных векторов, должна быть *линейно независимой*. Это значит, что равенство (1.5) должно выполняться лишь в том случае, когда все коэффициенты $\gamma_i = \alpha_i - \beta_i = 0$, т.е. $\alpha_i = \beta_i$ для всех i . Отсюда же следует, что если базисные векторы линейно независимы, нулевому вектору всегда ставится в соответствие совокупность чисел, состоящая из одних нулей.

Если при описании совокупности n независимых свойств использовать $m < n$ базисных векторов, то описать все векторы из V линейными комбинациями (1.3) не удастся. Если для описания элементов V суммами (1.3) воспользоваться $p > n$ базисными векторами, мы утратим

⁶В физической интерпретации это обеспечивает нам однозначные числовые характеристики изучаемых свойств объекта или явления относительно выбранной системы единиц.

единственность описания. Таким образом, существует некоторое минимально необходимое число базисных векторов n , которое называется *размерностью* векторного пространства.

Максимальная совокупность линейно независимых базисных элементов $\{e_i\}_{i=1}^n$ называется *базисом*. Базисные векторы только в том случае составляют базис, если их число равно размерности пространства.

Выбрав базис, каждый элемент f пространства V можно однозначно охарактеризовать упорядоченной совокупностью чисел $(f^1, \dots, f^n)$. Впредь мы будем обозначать эти числа той же буквой, что и вектор, нумеруя их верхними индексами. Каждый базисный вектор e_i также характеризуется совокупностью чисел. Для того чтобы равенство

$$e_i = \alpha e_1 + \dots + \omega e_n$$

выполнялось, только одно i -ое число совокупности $(\alpha, \dots, \omega)$ должно быть отлично от нуля и равно единице⁷, т.е.

$$e_1 = (1, 0, \dots, 0), \quad e_2 = (0, 1, \dots, 0), \dots, e_n = (0, 0, \dots, 1).$$

Теперь все элементы f пространства V характеризуются упорядоченными совокупностями чисел и могут быть записаны в виде линейной комбинации базисных векторов

$$f = \sum_{i=1}^n f^i e_i.$$

Числа из совокупности $\{f^i\}_{i=1}^n$, характеризующие вектор f , называются *компонентами вектора относительно выбранного базиса*.

⁷ Отсюда, кстати, следует, что базисный вектор не может быть нулевым вектором.

2. Линейные операторы

2.1. Линейные операторы

Рассмотрим уравнение

$$Af = g, \quad (2.1)$$

где f и g — элементы множества V , а A — отображение множества V в себя. В тех случаях, когда множество V наделено структурой векторного пространства, отображение A обычно называют *оператором*. Наша задача, зная g и оператор A , определить f , если это возможно.

Рассмотрим некоторые простые примеры операторов $A: \mathbb{R}^1 \rightarrow \mathbb{R}^1$.

- 1) Тожественный оператор $y = Ax \equiv x$. Область определения оператора и область его значений есть, очевидно, все $\mathbb{R}^1$, и оператор A взаимнооднозначен.
- 2) Оператор возведения в квадрат $y = Ax \equiv x^2$. Область определения $D(A) = \mathbb{R}^1$, а область значений $R(A) = \{y: y \geq 0\}$. Так как A отображает точки $\pm x$ в точку x^2 , то оператор не взаимнооднозначен.
- 3) $y = Ax \equiv x(x^2 - 1)$. Здесь и $D(A)$ и $R(A)$ суть $\mathbb{R}^1$. Поскольку точки $0, \pm 1$ отображаются в точку 0 , оператор не взаимнооднозначен.
- 4) $y = Ax \equiv \exp(x)$. И здесь $D(A) = \mathbb{R}^1$, а $R(A) = \{y: y > 0\}$; оператор взаимнооднозначен.

В приведенных примерах те операторы, которые не являются взаимнооднозначными можно таковыми сделать, ограничив так или иначе их область определения (сделайте это).

Простейшим оператором является *линейный оператор* (будем обозначать его буквой L). По определению он отображает линейную комбинацию элементов

$$h = \alpha f + \beta g, \quad \alpha, \beta \in \mathbb{R}^1, \quad f, g, h \in D(L)$$

в линейную же комбинацию

$$Lh = L(\alpha f + \beta g) = \alpha Lf + \beta Lg, \quad Lf, Lg, Lh \in R(L).$$

Легко видеть, что это определение имеет смысл, лишь если и $R(L)$ и $D(L)$ являются векторными пространствами.

Рассмотрим несколько примеров линейных операторов.

- 1) Часто встречающаяся задача — решение системы линейных алгебраических уравнений

$$\sum_{j=1}^n a_{ij}x_j = b_i.$$

Она сводится к виду $Lf = g$, если в качестве оператора L рассматривать матрицу, элементами которой являются коэффициенты a_{ij} , а в качестве f и g взять элементы $(x_1, x_2, \dots, x_n)$ и $(b_1, b_2, \dots, b_n)$ соответственно. В этом случае L есть оператор, отображающий $\mathbb{R}^n$ в $\mathbb{R}^n$. Линейность его легко проверяется.

- 2) Оператор дифференцирования

$$Lf \equiv f'.$$

Ясно, что этот оператор определен лишь на множестве функций, имеющих производную. Он линеен, что также легко проверить.

- 3) Рассмотрим, наконец, два специальных оператора.

- Пусть f — элемент векторного пространства V . Оператор I , такой, что $If = f$, называется *единичным (тождественным)* оператором ($I: f \mapsto f$).
- Оператор O , отображающий произвольный элемент $f \in V$ в нулевой элемент $Of = 0$, называется *нулевым оператором* ($O: f \mapsto 0$).

Пусть задано уравнение

$$Lf = g, \quad f \in D(L) \subset V, \quad g \in R(L) \subset V.$$

Формально проблема его решения, как указано на стр.17, сводится к отысканию оператора L^{-1} , такого, что

$$L^{-1}(Lf) = If = f \quad \text{или} \quad L^{-1} \circ L = I.$$

Если такой оператор существует, то он называется *обратным* (к исходному оператору L).

В том случае, когда оператор L не взаимнооднозначен, построить обратный оператор L^{-1} нельзя. Можно, однако, попробовать ограничить область определения оператора таким образом, чтобы он стал взаимнооднозначным. Это позволит сформулировать и, возможно, решить близкую задачу.

2.2. Матрица линейного оператора

Если линейный оператор определен на пространстве с базисом, то он может быть записан в виде квадратной матрицы $n \times n$, где n — размерность пространства, на котором он действует. Действительно, пусть V — векторное пространство, $\{e_i\}$ — базис в нем, а L — линейный оператор, $L: V \rightarrow V$. Вектор $f \in V$ отображается оператором L в некий вектор $g \in V$. Аналогично и базисные векторы $e_i \in V$ будут отображаться в некоторые векторы $\ell_i \in V$. Поскольку $\{e_i\}$ есть базис, образы ℓ_i базисных векторов e_i могут быть записаны через свои компоненты $(\ell_i^1, \ell_i^2, \dots, \ell_i^j, \dots)$ в этом базисе:

$$Le_i = \ell_i = \sum_j \ell_i^j e_j, \quad \ell_i^j \in \mathbb{R}^1. \quad (2.2)$$

У каждого образа базисного вектора n штук компонент ℓ_i^j , столько же, сколько и самих базисных векторов. Располагая числа ℓ_i^j в виде таблицы, связывая с индексом i — номер столбца, а с индексом j — номер строки, получим квадратную матрицу Λ , которая называется *матрицей оператора L в базисе $\{e_i\}$* .

Пусть векторы f и g в этом же базисе имеют вид

$$f = \sum f^i e_i, \quad g = \sum g^k e_k, \quad f^i, g^k \in \mathbb{R}^1, \quad (2.3)$$

где f^i, g^k — компоненты векторов f и g в базисе $\{e_i\}$. Тогда, пользуясь линейностью оператора, уравнение

$$Lf = g$$

можно записать в так называемой *компонентной форме*, связывающей компоненту вектора g с компонентами вектора f . Вначале, подставляя

в уравнение выражения (2.2), (2.3), получим

$$\begin{aligned} Lf &= L \sum_i f^i e_i = \sum_i f^i L e_i = \sum_i f^i \left(\sum_k \ell_i^k e_k \right) = \\ &= \sum_k \left(\sum_i \ell_i^k f^i \right) e_k = \sum_k g^k e_k. \end{aligned}$$

Для того чтобы найти выражение k -й компоненты вектора g через компоненты вектора f и матрицу Λ оператора L , достаточно вычесть друг из друга обе части последнего равенства

$$\sum_k \left(\sum_i \ell_i^k f^i - g^k \right) e_k = 0.$$

Поскольку в любом базисе нулевой вектор имеет нулевые компоненты, находим

$$\sum_i \ell_i^k f^i = g^k, \quad \forall k.$$

В итоге имеем систему линейных алгебраических уравнений:

$$\begin{cases} \ell_1^1 f^1 + \ell_2^1 f^2 + \ell_3^1 f^3 + \dots = g^1 \\ \ell_1^2 f^1 + \ell_2^2 f^2 + \ell_3^2 f^3 + \dots = g^2 \\ \dots \dots \dots \end{cases}$$

Вводя полужирные обозначения для векторов (в смысле линейной алгебры)

$$\mathbf{f} = (f^1, f^2, \dots)^T, \quad \mathbf{g} = (g^1, g^2, \dots)^T,$$

где верхний индекс T означает транспонирование, окончательно в компонентной форме имеем

$$\Lambda \mathbf{f} = \mathbf{g}.$$

Таким образом, исходная задача свелась к задаче линейной алгебры.

В качестве примера нахождения матрицы оператора рассмотрим оператор вычисления второй производной, определенный на векторном пространстве дважды дифференцируемых функций с базисом $\{e^{inx}\}_{n=1}^{\infty}$, где $i = \sqrt{-1}$.

Пусть $Lf \equiv f''(x)$, а f — дважды дифференцируемая функция. Для того чтобы найти матрицу оператора L , вычислим элемент $\ell_n = Le_n$:

$$\ell_n = Le_n = (e^{inx})'' = -n^2 e^{inx}.$$

Поскольку $\ell_n = \sum_j \ell_n^j e_j$, очевидно, что

$$\ell_n^j = \begin{cases} -n^2, & n = j, \\ 0, & n \neq j. \end{cases}$$

Таким образом, матрицей оператора является бесконечная диагональная матрица.

2.3. Векторное пространство линейных операторов

Часто удобно бывает рассматривать сложные операторы как сумму более простых. Для того чтобы это можно было делать, нужно на множестве линейных операторов ввести структуру векторного пространства. Требуется дать определение сумме двух операторов и произведению оператора на число.

Пусть V — векторное пространство, а L , L_1 и L_2 — линейные операторы, действующие на нем $L, L_1, L_2: V \rightarrow V$. Определим оператор $(L_1 + L_2)$, называемый *суммой операторов* L_1 и L_2 , как

$$(L_1 + L_2)f \equiv L_1f + L_2f, \quad \forall f \in V,$$

и оператор (αL) , $\alpha \in \mathbb{R}^1$, называемый *произведением оператора на число*, как

$$(\alpha L)f \equiv \alpha(Lf), \quad \forall f \in V, \quad \forall \alpha \in \mathbb{R}^1.$$

Нулевым элементом будем считать нулевой оператор O . Проверив выполнение требуемых аксиом (см. стр. 20), убедимся в том, что мы получили векторное пространство линейных операторов⁸. Обозначим его через $\mathcal{L}(V)$. Теперь линейные операторы, действующие на векторном пространстве V , сами могут рассматриваться как векторы, но уже другого векторного пространства $\mathcal{L}(V)$.

⁸Поясним смысл введенных определений на примере оператора $(L_1 + L_2)$. Считается, что определение оператору уже дано и что такие операторы L_1 и L_2 — известны. Требуется решить, какой оператор следует поставить в соответствие их сумме $(L_1 + L_2)$. Делается это так.

Поскольку вектор f считается определенным, определенными являются и век-

торы L_1f и L_2f . Суммирование векторов нами уже было определено и, значит известно, какой вектор ставится в соответствие сумме векторов $L_1f + L_2f$. Будем, поэтому, считать оператор $(L_1 + L_2)$ таким, что вектор $(L_1 + L_2)f$ равен сумме $L_1f + L_2f$.

3. Метод последовательных приближений

Мощным средством решения уравнения $Lf = g$ является метод, основанный на построении последовательности операторов L_n^{-1} , которые в некотором смысле все лучше и лучше приближают оператор L^{-1} . Этот метод носит название *метода последовательных приближений*. Обратный оператор L^{-1} интерпретируется как предел последовательности операторов, а сама последовательность строится в ходе решения задачи.

Для нахождения решения f организуется так называемая *итерационная*⁹ процедура, позволяющая по ранее вычисленным приближенным значениям решения f_{n-1} (часто называемым просто *приближениями* или *итерациями*) находить следующее приближение f_n . Все вычисления проводятся по так называемой *рекуррентной*¹⁰ формуле, связывающей новое приближение f_n с прежними. Чаще других применяются простейшие формулы, позволяющие вычислить f_n по f_{n-1} . Для запуска всего итерационного процесса требуется выбрать начальное приближение f_0 (или несколько первых приближений, если используется более сложный вариант рекуррентной формулы). Решение задачи f понимается как предел последовательности итераций $\{f_n\}_{n=0}^{\infty}$.

3.1. Идея метода

Идея метода очень проста. Представим оператор L в виде разности¹¹

$$L = I - M.$$

Здесь I — тождественный оператор, а M — некий другой линейный оператор (*почему он линейный?*). Теперь исходное уравнение выглядит так

$$Lf \equiv (I - M)f = g$$

или если раскрыть скобки, так

$$f = g + Mf. \quad (3.1)$$

⁹От лат. iteratio — повторение.

¹⁰От лат. recurrens (recurrentis) — возвращающийся.

¹¹Для того чтобы это стало возможным, нам требуется векторное пространство линейных операторов.

Уравнение в форме (3.1) можно записать как рекуррентную формулу

$$f_n = g + M f_{n-1}. \quad (3.2)$$

В качестве нулевого приближения f возьмем некоторую точку f_0 . Далее, подставляя нулевое приближение в правую часть уравнения (3.1) вместо f_{n-1} , получим новую точку

$$f_1 = g + M f_0,$$

которую интерпретируем как первое приближение. Продолжая процесс далее, можно получить любое n -е приближение.

Мы ожидаем, что последовательность $\{f_n\}$ имеет предел $\lim_{n \rightarrow \infty} f_n = f$. В этом случае говорят, что итерации сходятся, метод последовательных приближений работает и позволяет получить решение задачи (3.1), т.е. фактически построить обратный оператор L^{-1} . Для того чтобы убедиться в сходимости метода или ее отсутствии, мы сначала найдем связь n -го приближения с начальным, а затем рассмотрим возможные способы оценки близости итераций к точному решению.

Выразим $(n-1)$ -е приближение по формуле (3.2) через $(n-2)$ -е приближение и подставим в ту же формулу:

$$\begin{aligned} f_n &= g + M(g + M f_{n-2}) = \\ &= (I + M)g + M(M f_{n-2}). \end{aligned} \quad (3.3)$$

В последнем выражении у нас получилась композиция двух операторов M , которую принято называть степенью оператора. Это частный случай произведения операторов, т.е. композиции двух разных операторов.

Если $L, N : V \rightarrow V$ — линейные операторы и V — векторное пространство, то *произведением операторов* NL называется оператор $P : V \rightarrow V$, ставящий в соответствие элементу $v \in V$ элемент $w = N(Lv) \in V$. Произведение NN будем называть *степенью оператора* и обозначать N^2 . Точно также можно определить любую целую степень оператора. Кроме того, удобно считать нулевую степень любого оператора равной единичному оператору: $N^0 = I$.

Запишем выражение (3.3) для f_n через степень оператора M и продолжим подстановку приближений вплоть до f_0 :

$$\begin{aligned} f_n &= (I + M)g + M^2 f_{n-2} = \\ &= (I + M)g + M^2(g + M f_{n-3}) = (I + M + M^2)g + M^3 f_{n-3} = \dots = \\ &= \left(I + \sum_{i=1}^{n-1} M^i \right) g + M^n f_0. \end{aligned}$$

Выбирая в качестве начального приближения $f_0 = g$ и учитывая, что $M^0 = I$, окончательно получим

$$f_n = \left(\sum_{i=0}^n M^i \right) g.$$

Если последовательность $\{f_n\}$ имеет предел, то он дает решение задачи

$$f = \left(\sum_{i=0}^{\infty} M^i \right) g. \quad (3.4)$$

Тем самым мы строим обратный оператор

$$L^{-1} = \sum_{i=0}^{\infty} M^i.$$

Чтобы иметь возможность использовать описанный метод, последовательность итераций должна сходиться, расстояния между ними должны сокращаться и в пределе давать решение задачи. Теперь, таким образом, надо заняться способами оценки близости элементов множеств (итераций и точного решения — в нашем случае).

3.2. Метрика

Выше мы уже раз использовали понятие расстояния, не оговаривая, как его измерить. Рассмотрим этот вопрос подробнее. В случае с точками на прямой, плоскости или в трехмерном пространстве все интуитивно понятно. Однако если под точками понимать элементы векторных

пространств, то что называть расстоянием между ними? Воспользуемся имеющимся интуитивным представлением расстояния d между двумя точками f и g , скажем на плоскости, проанализируем его и в общем случае придумаем нечто аналогичное.

Во-первых, расстояние зависит лишь от взаимного расположения точек и не зависит от их положения относительно любой другой точки. Таким образом, аргументов у функции d всего два $d = d(f, g)$. Во-вторых, неважно, как мы измеряем расстояние: от первой точки ко второй или наоборот, т.е. функция d симметрична $d(f, g) = d(g, f)$. В-третьих, расстояние между двумя разными точками есть положительное число, значит, и функция d положительна $d(f, g) > 0$. В-четвертых, если точки совпадают, то расстояние между ними равно нулю, откуда $d(f, f) = d(g, g) = 0$. И, наконец, если мы рассматриваем расстояния между тремя точками, то любое из них не превышает суммы двух других (длина стороны треугольника не превышает суммы длин двух других сторон).

Вот такими же свойствами мы наделим и расстояние между точками любого векторного пространства или даже просто произвольного множества. Формально все, что было сказано, запишется так.

Пусть X — произвольное множество и $f, g, h \in X$. Поставим каждой паре f и g в соответствие неотрицательное число $d(f, g)$, такое, что для любых f, g и h из X справедливо:

- 1) $d(f, g) = d(g, f)$,
- 2) $d(f, g) > 0$, если $f \neq g$,
- 3) $d(f, f) = 0$,
- 4) $d(f, g) \leq d(f, h) + d(h, g)$.

Очевидно, что $d(f, g)$ есть функция (отображение), определенная на любой паре векторов из X со значениями в числах (в $\mathbb{R}^1$). Такая функция называется *метрикой* на множестве X , а само X , снабженное метрикой (или, как говорят, пара (X, d)), — *метрическим пространством*.

Обратите внимание, что для того чтобы некоторое множество стало метрическим пространством, на нем необходимо ввести метрику и

только. При этом оно, естественно, не становится векторным пространством, ибо на нем еще не определены сложение элементов и умножение их на число.

3.3. Норма вектора

В том случае, когда мы работаем с векторным пространством V , можно воспользоваться уже имеющейся выделенной точкой 0 и ввести более удобное понятие нормы вектора. По сути, это расстояние между рассматриваемым вектором и нулевым, или, другими словами, длина вектора. Для обозначения нормы используется специальный значок $\|\cdot\|$, т.е. норма вектора f обозначается $\|f\|$. Формальное определение нормы таково.

Пусть V — векторное пространство и $f \in V$. *Нормой вектора f* называется неотрицательная числовая функция $\|f\|$, определенная на V , такая, что для любых $f, g \in V$ и $\alpha \in \mathbb{C}$ выполняются условия:

- 1) $\|f\| > 0, \quad f \neq 0,$
- 2) $\|0\| = 0,$
- 3) $\|\alpha f\| = |\alpha| \cdot \|f\|,$
- 4) $\|f + g\| \leq \|f\| + \|g\|$ (неравенство треугольника).

Заменой элемента g на $(h - f)$ и последующим переобозначением легко получить и другие варианты неравенства треугольника:

$$\begin{aligned} \left| \|f\| - \|g\| \right| &\leq \|f + g\| \leq \|f\| + \|g\|, \\ \left| \|f\| - \|g\| \right| &\leq \|f - g\| \leq \|f\| + \|g\|. \end{aligned}$$

Линейное пространство, снабженное нормой (или пара $(V, \|\cdot\|)$), называется *нормированным векторным пространством*. Если для каждой точки V известна ее норма, то легко условиться и о том, как измерять расстояния между точками из V , т.е. получить метрику. Для этого расстоянием между двумя точками f и g можно считать норму их разности

$$d(f, g) = \|f - g\|. \quad (3.5)$$

Данное выше определение нормы не задает нам её единственным образом. Часто для одного и того же векторного пространства можно ввести несколько норм. Получающиеся при этом нормированные пространства считаются разными.

Примеры.

- 1) Линейное пространство $\mathbb{R}^1$ становится нормированным, если нормой $\|x\|$ элемента $x \in \mathbb{R}^1$ считать его модуль $|x|$. Легко проверить, что $|x|$ действительно норма.
- 2) В пространстве $\mathbb{R}^n$ норму вектора $x = (x^1, x^2, \dots, x^n)$ можно ввести многими способами. Наиболее часто используются следующие нормы: октаэдрическая норма, или норма $\|\cdot\|_1$:

$$\|x\|_1 = \sum_{i=1}^n |x^i|; \quad (3.6)$$

сферическая (евклидова) норма, или норма $\|\cdot\|_2$:

$$\|x\|_2 = \left(\sum_{i=1}^n |x^i|^2 \right)^{\frac{1}{2}}; \quad (3.7)$$

нормы $\|\cdot\|_p$, где p — натуральное число (нормы $\|\cdot\|_1$, $\|\cdot\|_2$ — частные случаи норм $\|\cdot\|_p$):

$$\|x\|_p = \left(\sum_{i=1}^n |x^i|^p \right)^{\frac{1}{p}}; \quad (3.8)$$

кубическая норма, или норма $\|\cdot\|_\infty$:

$$\|x\|_\infty = \max_i |x^i|. \quad (3.9)$$

- 3) Пусть $F([a, b])$ — множество вещественных функций, определенных на отрезке $[a, b]$. Введем на этом множестве структуру векторного пространства. Пусть $f, g \in F$ и $\alpha \in \mathbb{R}^1$. Определим новые функции $f + g$ и αf , считая, что для всех $x \in [a, b]$

$$(f + g)(x) = f(x) + g(x),$$

$$(\alpha f)(x) = \alpha f(x).$$

Иначе говоря, значение суммы функций в точке x равно сумме значений функций-слагаемых в той же точке. Аналогично для αf . Аксиомы векторного пространства, очевидно, выполняются (проверьте это).

Как правило, столь общие линейные пространства функций не рассматривают, а изучают подмножества F , которые в свою очередь также образуют линейные пространства, т.е. являются подпространствами F . Например, пространство $C^0([a, b])$ непрерывных функций, определенных на $[a, b]$, или пространство $C^k([a, b])$ функций, обладающих k ограниченными непрерывными производными. Функциональные линейные пространства, как правило, бесконечномерны. Эти пространства также можно сделать нормированными, если построить для них норму. Так, $C^0([a, b])$ часто снабжается так называемой *равномерной нормой*:

$$\|f\| = \max_{x \in [a, b]} |f(x)|. \quad (3.10)$$

Очевидно, что это бесконечномерный аналог нормы $\|\cdot\|_\infty$. Другая возможная норма — норма $\|\cdot\|_1$:

$$\|f\|_1 = \int_a^b |f(x)| dx. \quad (3.11)$$

Существуют и иные варианты норм.

Равномерной нормой можно дать наглядную интерпретацию. Пусть M — множество функций, таких, что $\|f(x)\| < \alpha$, для всех $f(x) \in M$ и $\alpha > 0$, $\alpha \in \mathbb{R}^1$. Тогда все f из M должны «укладываться» в полосу $\pm\alpha$ относительно оси абсцисс (рис.5).

Если P — множество функций f , для которых расстояние от заданной функции g не превышает β , т.е. $\|f - g\| \leq \beta$, то все изменения f должны быть заключены в полосу шириной 2β , охватывающей функцию g . Для нормы $\|\cdot\|_1$ это не так, поскольку ограничиваться будет лишь интеграл, а значения функции в отдельных точках могут этому ограничению не удовлетворять. Ограничение $\|f\| < \alpha$, таким образом, более сильное, нежели $\|f\|_1 < \alpha$, и из первого следует второе, но не наоборот.

3.4. Сходимость и банаховы пространства

Метод последовательных приближений — итерационный метод, который заключается в построении последовательности приближений

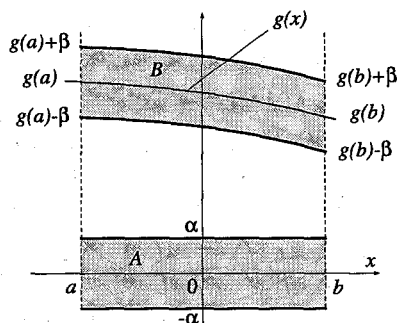


Рис. 5. Значения функций, норма которых не превосходит α , должны лежать в полосе A, а значения функций f , расстояние которых до функции g не превышает $\beta > 0$ (т.е. $\|f - g\| \leq \beta$), располагаются только в полосе B.

$\{f_n\}$, являющихся (по замыслу) с ростом номера n все более и более точными аппроксимациями решения f . Последовательность $\{f_n\}$, как говорят, должна *сходиться* к точному решению f .

Что понимать под этой сходимостью? Если функции f_n интерпретировать как векторы (точки) векторного пространства, то нам надо определить сходимость последовательности векторов. Естественно воспользоваться здесь опытом исследования сходимости числовых последовательностей и придумать что-нибудь в том же духе.

Говорят, что числовая последовательность $\{a_n\}_{n=1}^{\infty}$ сходится к некоторому числу a , когда модуль разности $|a_n - a|$ стремится к нулю при стремлении n к ∞ . Если члены последовательности $\{f_n\}$ принадлежат произвольному множеству X , то воспользоваться сходимостью, определенной для числовых последовательностей, непосредственно в терминах f_n не удастся. Однако если ввести на X метрику или норму, то мы сможем определить сходимость $\{f_n\}$ буквально так же, как это было сделано для чисел.

Рассмотрим некоторое линейное пространство V с метрикой $d(f, g)$, $f, g \in V$ (если V нормировано, то $d(f, g) = \|f - g\|$). Пусть $\{f_n\}_{n=1}^{\infty}$ есть последовательность точек из V , т.е. некоторое счетное¹² подмножество

¹²Счетное потому, что элементы пронумерованы целыми числами, т.е. сосчитаны.

V. Если существует такая точка $f \in V$, что с ростом n расстояние между f_n и f сокращается в пределе до нуля

$$\lim_{n \rightarrow \infty} d(f_n, f) = 0 \quad \text{или} \quad \lim_{n \rightarrow \infty} \|f_n - f\| = 0,$$

то последовательность $\{f_n\}$ называется *сходящейся*, а элемент f называется *пределом последовательности* $\{f_n\}$. Это записывается так:

$$\lim_{n \rightarrow \infty} f_n = f, \quad \text{или} \quad f_n \rightarrow f.$$

Для того чтобы оценить, является ли последовательность сходящейся, мы должны знать предел и убедиться в том, что расстояние от точек последовательности до предела стремится к нулю. Это, вообще говоря, неудобно, так как предел, как правило, неизвестен, а в нашем распоряжении имеются лишь члены последовательности. Вместе с тем кажется очевидным, что с ростом номеров точек f_i и f_j сходящейся последовательности расстояние между ними стремится к нулю. При этом возрастающие номера i и j вовсе не должны быть между собой как-либо связаны, нужно только, чтобы оба номера увеличивались. Действительно, пусть $f_i, f_j \in \{f_n\}$ и $f_n \rightarrow f$, тогда в силу неравенства треугольника

$$d(f_i, f_j) \leq d(f_i, f) + d(f_j, f).$$

Если $i \rightarrow \infty$ и (независимо от i) $j \rightarrow \infty$, то и $d(f_i, f) \rightarrow 0$ и $d(f_j, f) \rightarrow 0$, откуда $d(f_i, f_j) \rightarrow 0$. Последовательность $\{f_n\}$, обладающая свойством $d(f_i, f_j) \rightarrow 0$ при $i, j \rightarrow \infty$, называется *фундаментальной*, или *последовательностью Коши*¹³. И мы только что показали, что всякая сходящаяся последовательность является последовательностью Коши.

Верно ли обратное? То есть, всякая ли последовательность Коши сходится? Для конечномерного случая обратное утверждение верно. В бесконечномерном случае (а множества функций, как правило, бесконечномерны) это уже не всегда так, и сходимость последовательностей может зависеть от выбранной нормы.

Нормированное векторное пространство, в котором всякая последовательность Коши сходится по норме (т.е. предел является элементом

¹³Коши (Cauchy) Огюстен Луи (1789–1857) — французский математик.

того же пространства), называется *полным нормированным векторным пространством*, или *банаховым пространством*¹⁴. В бесконечномерном случае полнота пространства зависит от выбора нормы.

3.5. Эквивалентные нормы

Если векторное пространство, наделенное одной нормой полно, а наделенное другой — нет, то естественно задаться вопросом, какие же нормы не приводят к нарушению полноты векторного пространства? Ответ известен — это так называемые эквивалентные нормы. Две нормы $\|\cdot\|_a$ и $\|\cdot\|_b$ на векторном пространстве V называются *эквивалентными*, если найдутся числа $c_1 > 0$ и $c_2 > 0$, такие, что для любого $f \in V$

$$c_1 \|f\|_a \leq \|f\|_b \leq c_2 \|f\|_a. \quad (3.12)$$

Конечно, с равным успехом можно ограничить и норму $\|f\|_a$:

$$\frac{1}{c_2} \|f\|_b \leq \|f\|_a \leq \frac{1}{c_1} \|f\|_b,$$

или, в других обозначениях:

$$k_1 \|f\|_b \leq \|f\|_a \leq k_2 \|f\|_b, \quad k_1, k_2 = \text{const} > 0.$$

Если нормы $\|\cdot\|_a$ и $\|\cdot\|_b$ эквивалентны, то из сходимости по норме $\|\cdot\|_a$ следует сходимость по норме $\|\cdot\|_b$, и наоборот. Действительно, пусть $\|f_n - f\|_a \rightarrow 0$ при $n \rightarrow \infty$. Тогда из (3.12) следует, что числовая последовательность $\|f_n - f\|_b$ мажорируется последовательностью, стремящейся к нулю и, следовательно, также сходится к нулю.

В конечномерном случае все нормы эквивалентны. В бесконечномерном случае это не так.

¹⁴По имени польского математика Стефана Банаха (Banach) (1892–1945).

4. Сходимость метода последовательных приближений

4.1. Норма линейного оператора

Поскольку в методе последовательных приближений оператор L^{-1} интерпретируется как предел некоторой последовательности операторов и эта последовательность строится в ходе решения задачи, необходимо определить понятие сходимости операторов и степень близости их друг к другу. Для этого введем в векторном пространстве линейных операторов норму. Определим норму оператора, обобщив уже известное нам понятие равномерной нормы (см. стр. 39).

Рассмотрим линейный оператор $L: D(L) \rightarrow R(L)$, отображающий $D(L) \subset B$ — подмножество банахова пространства B в $R(L) \subset B$ — подмножество того же пространства. Мы выбираем здесь банахово пространство для того, чтобы векторы из $D(L)$ и $R(L)$ обладали нормой, а последовательности Коши сходились. Оператор L отображает точку f из $D(L)$ в точку Lf из $R(L)$.

Назовем оператор L *ограниченным*, если существует конечное число $m < \infty$ такое, что каким бы ни был вектор f из $D(L)$, норма вектора Lf из $R(L)$ никогда не превзойдет норму f более чем в m раз:

$$\|Lf\| \leq m\|f\|, \quad \forall f \in D(L). \quad (4.1)$$

Если такого числа не существует, т.е. оператор L не ограничен на $D(L)$, он называется *неограниченным*.

В случае, когда L — ограниченный оператор, для каждого вектора $f \in D(L)$ существует некоторое минимальное значение $m_{\min}(f)$, при котором неравенство (4.1) остается справедливым. Максимальное же значение всех таких $m_{\min}$ будем называть *нормой оператора L* и обозначать $\|L\| \equiv \max_f (m_{\min}(f))$. Если m в (4.1) равно $\|L\|$, неравенство справедливо для всех $f \in D(L)$. Далее под $\mathcal{L}(B)$ мы будем понимать векторное пространство линейных ограниченных операторов.

Разберем пример. Пусть $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$ — n -мерный линейный оператор, который, как мы уже знаем, относительно базиса $\{e_j\}$ может быть задан

квадратной матрицей (ℓ_i^j) :

$$Lf = \sum_j \sum_i \ell_i^j f^i e_j.$$

Предположим, что $\mathbb{R}^n$ наделено нормой $\|\cdot\|_\infty$, тогда

$$\begin{aligned} \|Lf\|_\infty &= \max_j \left| \sum_{i=1}^n \ell_i^j f^i \right| \leq \max_j \sum_i |\ell_i^j| \cdot |f^i| \leq \\ &\leq \left(\max_j \sum_i |\ell_i^j| \right) \left(\max_i |f^i| \right) = m \|f\|_\infty, \end{aligned}$$

где $m = \max_j \sum_i |\ell_i^j|$. Отсюда $\|L\| \leq m$. Покажем, что $\|L\| = m$. Попробуем найти такое f , для которого $\|Lf\|_\infty = m\|f\|_\infty$. Из определения m видно, что существует такое целое k , при котором $m = \sum_i |\ell_i^k|$. В качестве f выберем вектор с нормой, равной единице, и координатами $\ell_i^k/|\ell_i^k|$. Тогда цепочка неравенств обращается в равенства

$$\|Lf\|_\infty = \max_j \left| \sum_i \ell_i^j f^i \right| = \left| \sum_i \ell_i^k f^i \right| = \sum_i |\ell_i^k| = m \|f\|_\infty,$$

и в итоге мы получаем требуемый результат.

Введенная выше норма оператора задает на $\mathcal{L}(B)$ структуру нормированного пространства. Более того, можно показать, что пространство $\mathcal{L}(B)$ полно по операторной норме, т.е. является банаховым.

4.2. Сходимость метода последовательных приближений

Вернемся, теперь к исследованию сходимости метода последовательных приближений. Если для оператора L существует обратный оператор L^{-1} , то по нашему построению он равен

$$L^{-1} = (I - M)^{-1} = \sum_{i=0}^{\infty} M^i. \quad (4.2)$$

Выясним при каких условиях этот ряд сходится, т.е. в каком случае метод последовательных приближений работает.

4.2.1. Существование обратного оператора

Пусть B — банахово пространство и оператор L (а значит, и M) есть элемент $\mathcal{L}(B)$. Если сумма ряда (4.2) существует, то оператор L^{-1} ограничен по норме $\|L^{-1}\| < \infty$ и может рассматриваться как оператор, обратный оператору L . Из формулы (4.2) и в силу неравенства треугольника имеем:

$$\|L^{-1}\| = \left\| \sum_n M^n \right\| \leq \sum_n \|M^n\|. \quad (4.3)$$

Использование неравенства треугольника в последнем неравенстве позволило получить числовой ряд, для исследования которого существуют развитые средства.

Чтобы продвинуться дальше требуется понять, как оценить норму произведения операторов. Можно показать, что если $L : V \rightarrow U$ и $N : U \rightarrow W$ — ограниченные операторы и V, U, W — нормированные пространства, то оператор $P = NL$ также ограничен и его норма

$$\|P\| \leq \|N\| \cdot \|L\|. \quad (4.4)$$

Используя (4.4), продолжим цепочку неравенств (4.3)

$$\sum_n \|M^n\| \leq \sum_n \|M\|^n.$$

Ряд $\sum \|M\|^n$ устроен совсем просто — это геометрическая прогрессия со знаменателем $\|M\|$. Относительно него известно, что он сходится, т.е. $\sum \|M\|^n < \infty$, если $\|M\| < 1$. Последнее неравенство и есть условие сходимости ряда (4.2) и существования обратного оператора.

4.2.2. Сходимость приближенных решений

Легко установить, что при $\|M\| < 1$ последовательность $\{f_n\}$ имеет предел f . Для этого вычислим норму разности точного решения исходного уравнения f и n -го приближения f_n и выясним, при каких условиях она стремится к нулю. Найдем разность исходного уравнения (3.1) и рекуррентной формулы (3.2):

$$f - f_n = M(f - f_{n-1}). \quad (4.5)$$

Вычислим норму обеих частей полученного уравнения и, воспользовавшись определением нормы, оценим расстояние между точным решением f и n -й итерацией f_n в терминах расстояния между f и предыдущей итерацией f_{n-1} :

$$\|f - f_n\| = \|M(f - f_{n-1})\| \leq \|M\| \cdot \|f - f_{n-1}\|.$$

Применяя $n-1$ раз формулу (4.5) в правой части полученного неравенства, найдем оценку расстояния между f и f_n в терминах расстояния между f и начальным приближением f_0 :

$$\|f - f_n\| \leq \|M\|^n \|f - f_0\|. \quad (4.6)$$

Отсюда при $\|M\| < 1$ вытекает, что

$$\lim_{n \rightarrow \infty} \|f_n - f\| = 0, \quad \text{или} \quad \lim_{n \rightarrow \infty} f_n = f.$$

Неравенство (4.6) характеризует скорость сходимости итерационного процесса. Если для последовательных приближений справедливо неравенство

$$\|f_n - f\| \leq cq^n, \quad c = \text{const},$$

то говорят, что *приближения f_n сходятся к f со скоростью геометрической прогрессии со знаменателем q* . В рассматриваемом методе скорость сходимости последовательных приближений именно такая, а знаменателем прогрессии служит норма $\|M\|$.

При известном f неравенство (4.6) можно было бы использовать для определения необходимого числа итераций n для достижения заданной точности ε

$$\|f - f_n\| \leq \|M\|^n \|f - f_0\| < \varepsilon \Rightarrow n > \frac{\ln \frac{\varepsilon}{\|f - f_0\|}}{\ln \|M\|}.$$

Если начальное приближение не является с указанной точностью искомым решением, т.е. $\|f - f_0\| > \varepsilon$, величина $\ln \frac{\varepsilon}{\|f - f_0\|}$ отрицательна. Величина $\ln \|M\|$ отрицательна в силу того, что $\|M\| < 1$. Таким образом, n должно быть больше некоторого положительного вещественного числа. Однако, поскольку f неизвестно, полученная оценка числа итераций неопределенна.

Чтобы получить содержательную оценку числа n , неравенство (4.6) надо переформулировать так, чтобы норма разности $\|f - f_0\|$ выражалась в терминах известных величин. Используя исходное уравнение $f = g + Mf$ запишем норму разности $\|f - f_0\|$ в виде

$$\|f - f_0\| = \|g + Mf - f_0\|.$$

Прибавляя к правой части и отнимая от нее вектор Mf_0 , получим оценку

$$\begin{aligned}\|f - f_0\| &= \|g + Mf_0 - Mf_0 + Mf - f_0\| \leq \\ &\leq \|g + Mf_0 - f_0\| + \|M\| \cdot \|f - f_0\|.\end{aligned}$$

Откуда

$$\|f_0 - f\| \leq \frac{\|g + Mf_0 - f_0\|}{1 - \|M\|}.$$

Величина $g + Mf_0 - f_0 = g - (I - M)f_0 = g - Lf_0$, стоящая в числителе, известна. Она называется *невязкой* начального приближения и равна разности правой и левой частей исходного уравнения при подстановки в него f_0 вместо точного решения f .

Подставляя найденную оценку в (4.6), окончательно получим

$$\|f_n - f\| \leq \frac{\|M\|^n}{1 - \|M\|} \|Lf_0 - g\|. \quad (4.7)$$

В правой части этого неравенства стоят известные величины, поэтому не составляет труда, зная $\|M\|$, вычислить необходимое для достижения заданной точности число итераций (*найдите его*). Помимо $\|M\|$, это число также зависит от правой части g исходного уравнения и начального приближения f_0 .

Интересно рассмотреть, как меняется расстояние между двумя соседними членами последовательности $\{f_n\}$ под действием оператора M с ростом числа итераций n . Оценим расстояние между парой соседних векторов f_{n+1} и f_n в терминах расстояния между предыдущей парой векторов f_n и f_{n-1} , вычислив нормы обеих частей разности рекуррентных формул $f_n = g + Mf_{n-1}$ и $f_{n+1} = g + Mf_n$:

$$\|f_{n+1} - f_n\| = \|Mf_n - Mf_{n-1}\| \leq \|M\| \cdot \|f_n - f_{n-1}\|.$$

При $\|M\| > 1$ расстояния между соседними членами последовательности растут. Если же $\|M\| < 1$, расстояние между двумя соседними векторами с ростом n сокращается — члены последовательности сближаются. В этом случае оператор M называют *оператором сжатия* (или *сжимающим*). Термин очень нагляден.

4.3. Пример приложения метода последовательных приближений

В качестве примера использования метода последовательных приближений рассмотрим приложение его к решению систем линейных алгебраических уравнений и получим ограничения, вытекающие из условия $\|M\| < 1$. Пусть требуется решить уравнение

$$Af = g, \quad (4.8)$$

где $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ — оператор, записанный в виде квадратной $n \times n$ матрицы, а $f, g \in \mathbb{R}^n$ — векторы неизвестных и правых частей соответственно. Представим оператор A в виде $(I - M)$, где I — единичная $n \times n$ матрица, а M — некая новая $n \times n$ матрица (оператор). Уравнение (4.8) запишется в виде:

$$(I - M)f = g.$$

Если $\|M\| < 1$, можно организовать итерационную процедуру

$$f_{n+1} = g + Mf_n$$

и вычислить $f = \lim_{n \rightarrow \infty} f_n$. Это хорошо известный метод решения систем линейных уравнений — *метод Якоби*¹⁵. Выясним, какие ограничения на элементы матрицы A накладывает требование $\|M\| < 1$.

Пусть a_{ij} — элементы матрицы A , тогда числа

$$(\delta_{ij} - a_{ij})$$

будут элементами матрицы M (δ_{ij} — символы Кронекера¹⁶; здесь они выполняют роль элементов единичной матрицы). Будем работать с нормой $\|\cdot\|_\infty$,

¹⁵ Якоби (Jacobi) Карл Густав Якоб (1804–1851) — немецкий математик.

¹⁶ Кронекер (Kronecker) Леопольд (1823–1891) — немецкий математик. Символы, носящие его имя, определяются так:

$$\delta_{ij} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

и тогда

$$\|M\| = \max_j \sum_{i=1}^n |\delta_{ij} - a_{ij}|.$$

Если $\|M\| < 1$, то для любого j выполняется $\sum_{i=1}^n |\delta_{ij} - a_{ij}| < 1$. Запишем в этой сумме j -е слагаемое отдельно (только в нем символ Кронекера будет ненулевым):

$$\sum_{i=1}^n |\delta_{ij} - a_{ij}| = \sum_{i \neq j} |a_{ij}| + |1 - a_{jj}| < 1.$$

Отсюда с помощью неравенства $1 - |a_{jj}| \leq |1 - a_{jj}|$ найдем

$$\sum_{i \neq j} |a_{ij}| < 1 - |1 - a_{jj}| \leq |a_{jj}|.$$

Таким образом, каждый диагональный элемент по модулю должен быть больше суммы модулей всех остальных элементов в данном столбце. Можно показать, что то же справедливо и для элементов в строке. Этот результат — хорошо известное условие сходимости метода Якоби — так называемое *условие диагонального преобладания* матрицы A .

4.4. Спектральный радиус оператора

Как мы убедились, метод последовательных приближений работает в том случае, когда $\|M\| < 1$ или оператор M есть оператор сжатия. Ну а если $\|M\| \geq 1$? Значит ли это, что метод последовательных приближений не применим? Не обязательно. Не исключено, что выбор нормы оказался неудачен и с другой нормой наш оператор стал бы оператором сжатия. Чтобы весь предыдущий анализ остался справедливым, новую норму следует искать среди множества эквивалентных норм (см. с.42), поскольку при переходе к эквивалентным нормам сходящиеся последовательности остаются сходящимися. Окончательный «диагноз» относительно сходимости метода последовательных приближений можно, таким образом, ставить лишь в том случае, если найдена эквивалентная норма, в которой оператор M является оператором сжатия, либо построена минимально возможная эквивалентная норма и M по-прежнему не является оператором сжатия. Минимальная норма тесно связана с так называемым спектральным радиусом оператора. Рассмотрим это понятие.

Процедура установления признака существования оператора L^{-1} состояла в том, что мы записывали ряд усиливающих друг друга неравенств, пока не получили геометрическую прогрессию, где все стало очевидным. Можно, однако, ограничиться предпоследним шагом и рассмотреть оценку

$$\|L^{-1}\| \leq \sum_n \|M^n\|. \quad (4.9)$$

Применим к последнему ряду признак Коши сходимости числовых рядов, для чего вычислим предел

$$r = \lim_{n \rightarrow \infty} \sqrt[n]{\|M^n\|}.$$

Число r называется *спектральным радиусом* оператора M . Поскольку $\|M^n\| \leq \|M\|^n$, ясно, что $r \leq \|M\|$. Согласно признаку Коши, если $r < 1$, то ряд (4.9) сходится и, значит, оператор L^{-1} существует; если $r > 1$, то ряд (4.9) расходится и у оператора L нет обратного оператора. При $r = 1$ мы не можем ничего сказать относительно сходимости ряда.

Зная спектральный радиус r , можно построить эквивалентную норму в пространстве B , при которой норма линейного оператора M сколь угодно близка к его спектральному радиусу. Нетрудно, далее, сформулировать признак сходимости метода последовательных приближений в терминах спектрального радиуса оператора M , а именно: если спектральный радиус $r(M) < 1$, то последовательные приближения сходятся к решению уравнения. Поскольку мы можем построить эквивалентную норму $\|M\|_*$, такую, что

$$\|M\|_* \leq r + \varepsilon, \quad \forall \varepsilon > 0,$$

то полученную оценку скорости сходимости (4.7) можно переписать в виде

$$\|f - f_n\| \leq \frac{(r(M) + \varepsilon)^n}{1 - (r(M) + \varepsilon)} \|g - Lf_0\|.$$

4.5. Задача на собственные значения

Все сформулированные критерии и оценки предполагают, что мы умеем вычислять норму оператора или его спектральный радиус. Однако пока в нашем распоряжении только такое неконструктивное определение нормы, как максимальное значение m константы $m_{\min}$, при

котором для любого $f \in B$

$$\|Mf\| \leq m\|f\|.$$

Чтобы получить конструктивную оценку нормы оператора M , рассмотрим уравнение

$$Mf = \lambda f, \quad (4.10)$$

где λ — некое число. Вычислив норму векторов, стоящих в правой и левой частях равенства, получим:

$$\|Mf\| = |\lambda| \cdot \|f\| \leq \|M\| \cdot \|f\| \quad \Rightarrow \quad \|M\| \geq \max(|\lambda|).$$

Откуда норма $\|M\|$ оказывается не меньше максимального значения $|\lambda|$, которое удовлетворяет уравнению (4.10). Таким образом, задача оценки нормы оператора сводится к поиску значений λ , при которых уравнение (4.10) или уравнение

$$(\lambda I - M)f = 0 \quad (4.11)$$

имеет нетривиальное, т.е. отличное от нуля, решение¹⁷. Последнее означает, что при этих λ оператор $(\lambda I - M)$ перестает быть взаимнооднозначным и, следовательно, обратный оператор $(\lambda I - M)^{-1}$ отсутствует. Действительно, если $f \neq 0$ есть решение (4.11), то и αf есть также решение (4.11) для любого действительного или комплексного α . Тем самым взаимная однозначность оператора нарушается.

Значения λ , при которых уравнение (4.10) имеет нетривиальное решение, называются *собственными значениями линейного оператора M* . Векторы f , являющиеся решением (4.10) (с точностью до постоянного множителя) при некотором собственном значении λ , называются *собственными векторами оператора M* , отвечающими данному λ . Если B — функциональное пространство, то собственные векторы часто называют *собственными функциями*.

Множество всех собственных значений оператора M называется его *точечным* (или *дискретным*) *спектром*. Для конечномерных операторов спектр совпадает с точечным спектром. Задача (4.11) поиска собственных значений оператора M называется *задачей на собственные значения*, или *спектральной задачей*.

¹⁷Нетрудно видеть, что функция $f = 0$ всегда является решением (тривиальным) уравнения (4.10).

5. Гильбертово пространство и ряд Фурье

5.1. Гильбертово пространство

Часто удобным способом введения нормы является ее определение через *скалярное произведение* векторов. Скалярным произведением $(\cdot, \cdot)$ называется числовая функция двух аргументов, обладающая следующими свойствами:

- 1) $(f, g) = (g, f)$ — симметрия,
- 2) $(f, g + \alpha h) = (f, g) + \alpha(f, h)$ — линейность,
- 3) $(f, f) \geq 0$, $(f, f) = 0 \Leftrightarrow f = 0$ — положительная определенность.

Здесь f, g, h — векторы из векторного пространства V , а α — вещественное число. Любой паре векторов скалярное произведение ставит в соответствие число. Векторное пространство с введенным в нем скалярным произведением называется *евклидовым*¹⁸.

Примеры.

- 1) В пространстве $\mathbb{R}^n$, элементами которого являются упорядоченные совокупности чисел $f = (f^1, \dots, f^n)$, $g = (g^1, \dots, g^n)$, ..., скалярное произведение определяется формулой

$$(f, g) = f^1 g^1 + f^2 g^2 + \dots + f^n g^n.$$

Легко проверить, что все аксиомы скалярного произведения выполняются.

- 2) Пусть $V = C^0([0, 1])$ — пространство непрерывных функций, определенных на отрезке $[0, 1]$. Скалярное произведение в этом случае можно задать так:

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

¹⁸Евклид (Ευκλείδης) — древнегреческий математик. По наиболее распространенной версии он работал в Александрии в III веке до н. э.

Определив скалярное произведение, мы можем ввести норму следующим образом:

$$\|f\| = \sqrt{(f, f)}. \quad (5.1)$$

То, что это выражение есть норма, нетрудно показать. Первые два свойства нормы очевидно удовлетворяются, так что остается проверить неравенство треугольника. Чтобы не возиться с корнями, его удобно записать в виде:

$$\|f + g\|^2 \leq (\|f\| + \|g\|)^2.$$

Пользуясь определением (5.1), а также свойствами скалярного произведения, левую часть этого неравенства представим в виде:

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

В правой части того же неравенства имеем

$$(\|f\| + \|g\|)^2 = \|f\|^2 + 2\|f\| \cdot \|g\| + \|g\|^2.$$

Если $(f, g) \leq \|f\| \cdot \|g\|$, то справедливым окажется и неравенство треугольника. Мы покажем это, доказав справедливость более жесткого неравенства $|(f, g)| \leq \|f\| \cdot \|g\|$.

Рассмотрим норму вектора $f + \alpha g$, где α — некоторое число. Очевидно, что

$$\begin{aligned} 0 \leq \|f + \alpha g\|^2 &= \|f\|^2 + 2(f, \alpha g) + \|\alpha g\|^2 = \\ &= \|f\|^2 + 2\alpha(f, g) + \alpha^2\|g\|^2. \end{aligned}$$

Откуда следует

$$-2\alpha(f, g) \leq \|f\|^2 + \alpha^2\|g\|^2. \quad (5.2)$$

Теперь, выбрав число α равным

$$\alpha = -\frac{(f, g)\|f\|}{|(f, g)|\|g\|},$$

получим требуемое неравенство. А именно, после подстановки α в (5.2), найдем

$$2\frac{(f, g)^2\|f\|}{|(f, g)|\|g\|} \leq \|f\|^2 + \frac{(f, g)^2\|f\|^2}{(f, g)^2\|g\|^2}\|g\|^2,$$

или, сокращая,

$$|(f, g)| \leq \|f\| \cdot \|g\|. \quad (5.3)$$

Неравенство (5.3) называется *неравенством Коши-Буняковского*¹⁹. Оно и доказывает неравенство треугольника.

Таким образом, всякое евклидово пространство можно сделать нормированным, определив норму равенством (5.1). Если евклидово пространство оказывается полным по норме (5.1), то оно называется *гильбертовым*²⁰ (такие пространства будем обозначать буквой H). Другими словами, гильбертовым называется банахово пространство, в котором норма определена через скалярное произведение.

5.2. Ортогональность

Из неравенства Коши-Буняковского (5.3) следует, что функция $\frac{(f, g)}{\|f\| \cdot \|g\|}$ не превосходит по модулю единицы. Это обстоятельство позволяет отождествить ее с косинусом некоторого угла φ , а скалярное произведение записать в виде:

$$(f, g) = \|f\| \cdot \|g\| \cos \varphi.$$

Угол φ достаточно условно называют углом между векторами f и g . Числа $f_g \equiv \|f\| \cos \varphi$ и $g_f \equiv \|g\| \cos \varphi$ называются проекциями f на g и g на f соответственно. В том случае, когда норма одного из векторов, например g , равна единице, имеем $(f, g) = \|f\| \cos \varphi = f_g$, т.е. скалярное произведение вектора f на единичный вектор g равно проекции f на g .

Если $(f, g) = 0$ и $\|f\| \neq 0$, $\|g\| \neq 0$, угол φ равен $\pm \frac{\pi}{2}$. В этом случае векторы называются *ортогональными*, что обозначается как $f \perp g$. Если имеется последовательность ненулевых векторов $\{f_n\}$ и для каждой пары векторов f_i, f_j справедливо равенство $(f_i, f_j) = 0$, $i \neq j$, то такая последовательность называется *ортогональной*.

Ортогональные векторы обладают важным свойством линейной независимости. Легко показать, что если $\{f_i\}_{i=1}^n$ — ортогональные векторы, то

$$\alpha_1 f_1 + \alpha_2 f_2 + \dots + \alpha_n f_n = 0$$

¹⁹Буняковский Виктор Яковлевич (1804–1889) — русский математик.

²⁰По имени немецкого математика Давида Гильберта (Hilbert) (1862–1943).

только, если $\alpha_1 = \alpha_2 = \dots = \alpha_n = 0$. Действительно, допустим, что векторы $\{f_i\}_{i=1}^n$ линейно зависимы, т.е. существует ненулевой набор чисел $\{\alpha_i\}$ такой, что $\sum \alpha_i f_i = 0$ — нулевой вектор. Возьмем любой вектор $f_j \in \{f_i\}_{i=1}^n$ и умножим на него скалярно обе части равенства

$$(f_j, \sum \alpha_i f_i) = (f_j, 0) = 0.$$

В то же время из ортогональности векторов следует, что

$$(f_j, \sum_i \alpha_i f_i) = \sum_i \alpha_i (f_j, f_i) = \alpha_j \|f_j\|^2 = 0.$$

По предположению последовательность $\{f_i\}_{i=1}^n$ не содержит нулевого вектора и, следовательно, $\|f_j\| \neq 0$, откуда $\alpha_j = 0$. То же справедливо и для любого другого элемента из $\{f_i\}_{i=1}^n$, а значит, все $\alpha_i = 0$.

Если последовательность $\{f_i\}$ такова, что

$$(f_i, f_j) = \delta_{ij} = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases} \quad (5.4)$$

она называется *ортонормированной*. В этом случае нормы всех векторов f_i равны единице. Будем обозначать такую систему векторов через $\{e_i\}$.

5.3. Ортогонализация

Существует процедура, позволяющая строить из произвольной последовательности векторов ортонормированную последовательность. Она называется *процессом ортогонализации Грама-Шмидта*²¹ и заключается в следующем.

Пусть в гильбертовом пространстве H задана последовательность векторов $\{h_i\}_{i=1}^n$, которую нужно ортогонализировать. Для того чтобы построить первый вектор e_1 ортонормированной последовательности $\{e_i\}_{i=1}^n$, достаточно взять первый ненулевой вектор h_1 последовательности $\{h_i\}$ и пронормировать его, т.е. положить e_1 равным

$$e_1 = \frac{h}{\|h_1\|},$$

²¹Грамм (Gram) Йорген Педерсен (1850–1916) — датский математик. Шмидт (Schmidt) Эрхард (1876–1959) — немецкий математик.

поскольку нормы ортонормированных векторов равны единице.

Следующий вектор e_2 должен быть ортогональным e_1 . Попробуем из элементов e_1 и h_2 построить такую линейную комбинацию g_2 , чтобы она была ортогональна e_1 . Пусть $g_2 = h_2 - \alpha_{21}e_1$, где α_{21} — число, выбором которого обеспечивается ортогональность $g_2 \perp e_1$. Действительно, варьируя число α_{21} , мы изменяем не только длину вектора g_2 , но и его направление. Всегда можно подобрать такое α_{21} , чтобы $g_2 \perp e_1$ (см. рис.6). В этом случае

$$\begin{aligned} 0 = (e_1, g_2) &= (e_1, h_2 - \alpha_{21}e_1) = \\ &= (e_1, h_2) - \alpha_{21}\|e_1\|^2 = (e_1, h_2) - \alpha_{21}. \end{aligned}$$

Откуда $\alpha_{21} = (e_1, h_2)$ и $g_2 = h_2 - (h_2, e_1)e_1$. Нормированный вектор e_2 должен совпадать с g_2 по направлению, поэтому выбираем его равным $e_2 = \frac{g_2}{\|g_2\|}$.

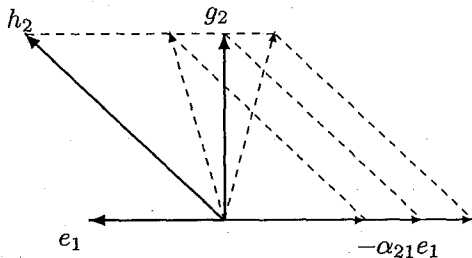


Рис. 6. Построение вектора, ортогонального данному.

Аналогично выбираем вектор e_3 , который уже должен быть ортогонален двум ранее найденным векторам e_1 и e_2 . Сначала строим вспомогательный вектор $g_3 = h_3 - \alpha_{32}e_2 - \alpha_{31}e_1$. Коэффициенты α_{31} и α_{32} находим из условия ортогональности g_3 векторам e_1 и e_2 , т.е. решаем систему:

$$\begin{aligned} 0 = (g_3, e_1) &= (h_3, e_1) - \alpha_{32}(e_2, e_1) - \alpha_{31}(e_1, e_1) = (h_3, e_1) - \alpha_{31}, \\ 0 = (g_3, e_2) &= (h_3, e_2) - \alpha_{32}(e_2, e_2) - \alpha_{31}(e_1, e_2) = (h_3, e_2) - \alpha_{32}. \end{aligned}$$

Откуда $\alpha_{31} = (h_3, e_1)$, $\alpha_{32} = (h_3, e_2)$. Наконец, нормировкой получаем ортонормированный вектор $e_3 = \frac{g_3}{\|g_3\|}$.

И так далее. Все векторы h_k , линейно зависящие от уже построенных векторов $\{e_i\}_{i=1}^{k-1}$, обнуляются (нулевым оказывается соответствующий вектор g_k) и отбрасываются. В результате получаем последовательность $\{e_i\}$, элементы которой ортонормированы и, следовательно, линейно независимы.

5.4. Ряд Фурье

Обладая ортонормированной последовательностью $\{e_i\}$ конечной или бесконечной в зависимости от размерности пространства H , мы можем использовать ее в качестве базиса и записать любой вектор f из H в виде ряда по e_i :

$$f = \sum_i f^i e_i, \quad (5.5)$$

где f^i — некоторые числа. Коэффициенты f^i легко найти, пользуясь тем, что $\{e_i\}$ — ортонормированная последовательность. Умножим скалярно обе части равенства (5.5) на какой-либо вектор e_j

$$(f, e_j) = \left(\sum_i f^i e_i, e_j \right) = \sum_i f^i (e_i, e_j)$$

и, учитывая (5.4), получим

$$f^j = (f, e_j). \quad (5.6)$$

Число f^j , таким образом, равно проекции f на единичный вектор e_j . Ряд (5.5), коэффициенты которого находятся по формулам (5.6), называется *рядом Фурье*²² для f , а коэффициенты f^i — *коэффициентами Фурье*.

Любой отрезок ряда Фурье, т.е. его частичная сумма, обладает замечательным свойством *наилучшего приближения*. Допустим, мы хотим

²²Фурье (Fourier) Жан Батист Жозеф (1768–1830) — французский физик и математик.

аппроксимировать некоторый элемент $f \in H$ функцией $\hat{f}$, являющейся линейной комбинацией m векторов $\{e_i\}$:

$$f \approx \hat{f} = \sum_{i=1}^m \alpha_i e_i, \quad (5.7)$$

Требуется подобрать числа $\alpha_i \in \mathbb{R}^1$ так, чтобы это приближение было наилучшим. Наилучшим естественно считать то приближение, для которого

$$\|f - \hat{f}\| = \min.$$

Норма разности $\|f - \hat{f}\|$ является функцией α_i , $i = 1, \dots, m$ и, как оказывается, достигает минимума в том случае, если коэффициенты α_i есть коэффициенты Фурье, т.е. находятся по формуле (5.6). Иначе говоря, наилучшая линейная комбинация (5.7) представляет собой отрезок ряда Фурье.

Действительно, рассмотрим квадрат нормы разности $\|f - \hat{f}\|^2$

$$\|f - \hat{f}\|^2 = \left\| f - \sum \alpha_i e_i \right\|^2 = \left(f - \sum \alpha_i e_i, f - \sum \alpha_i e_i \right) =$$

Вычисляя скалярное произведение, получим

$$= \|f\|^2 - 2 \left(f, \sum \alpha_i e_i \right) + \left(\sum \alpha_i e_i, \sum \alpha_i e_i \right) =$$

Второй член, очевидно, равен $2 \sum \alpha_i (f, e_i) = 2 \sum \alpha_i f^i$, где f^i — коэффициенты Фурье, вычисленные по формуле (5.6), а третий член в силу ортонормированности $\{e_i\}$, равен $\sum \alpha_i^2$. Далее имеем

$$= \|f\|^2 + \sum \alpha_i^2 - 2 \sum \alpha_i f^i = \|f\|^2 + \sum (\alpha_i^2 - 2\alpha_i f^i) =$$

Добавим и вычтем из скобки величину $(f^i)^2$. Тогда второе слагаемое можно представить как квадрат разности

$$= \|f\|^2 - \sum (f^i)^2 + \sum (\alpha_i - f^i)^2.$$

Последнее слагаемое — единственное, зависящее от коэффициентов α_i . Нетрудно видеть, что все выражение минимально, если последнее слагаемое равно нулю, т.е. при $\alpha_i = f^i$, что и требовалось доказать. Таким

образом, имеем

$$\|f - \hat{f}\| = \min \quad \Rightarrow \quad \|f - \hat{f}\|^2 = \|f\|^2 - \sum (f^i)^2.$$

Так как $\|f - \hat{f}\|^2 \geq 0$, то из последнего равенства следует, что

$$\sum_{i=1}^m (f^i)^2 \leq \|f\|^2.$$

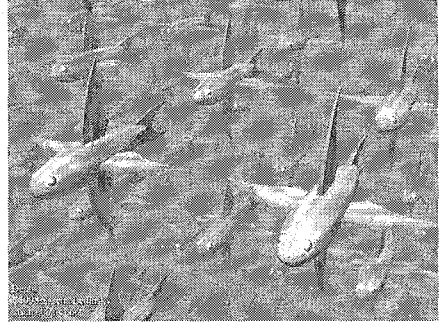
Это неравенство справедливо при любом m , и поскольку $\|f\|^2$ не зависит от m , ряд $\sum_{i=1}^{\infty} (f^i)^2$ сходится и мы получаем *неравенство Бесселя*²³:

$$\sum_{i=1}^{\infty} (f^i)^2 \leq \|f\|^2. \quad (5.8)$$

Геометрически это означает, что сумма квадратов проекций вектора f на ортогональные направления не превосходит квадрата длины самого вектора. Это бесконечномерный аналог теоремы Пифагора²⁴.

²³Бессель (Bessel) Фридрих Вильгельм (1784–1846) — немецкий астроном и геодезист.

²⁴Пифагор (Πυθαγόρας) Самосский (VI век до н. э.) — древнегреческий философ, религиозный и политический деятель, основатель пифагореизма, математик.



Часть II. Простейшие непрерывные модели

6. Уравнения 1-го порядка

Прежде чем заниматься изучением методов решения уравнений, рассмотрим те операторы, с которыми мы будем сталкиваться чаще всего при решении конкретных задач. Обсуждать будем линейные дифференциальные операторы, если не указано противное. Пусть L — дифференциальный оператор, тогда выражение

$$Lf = g$$

называется *дифференциальным уравнением*. Иногда говорят, что дифференциальное уравнение есть соотношение между неизвестной функцией f и ее производными. Если в это соотношение входят частные производные, то уравнение будет *уравнением в частных производных*. Максимальный порядок производной определяет *порядок дифференциального уравнения*. Часто уравнения записываются так, чтобы в левой части находились неизвестные слагаемые. В правой части, в таком случае, оказываются известные члены уравнения. Если правая часть равна нулю, $g = 0$, уравнение называется *однородным*, иначе — *неоднородным*.

6.1. Однородное уравнение

Пусть Ω^n — некоторая область n -мерного векторного пространства V^n и χ — отображение, ставящее в соответствие каждой точке $P \in \Omega^n$ точку $x_P = (x_P^1, \dots, x_P^n) \in \mathbb{R}^n$. Числа $x_P^1, \dots, x_P^n$ — координаты точки P относительно отображения χ , которое называется в таком случае *координатным отображением*.

Пусть в Ω^n задана гладкая кривая φ (рис.7). Свяжем взаимно однозначно каждую точку кривой с вещественным числом t , т. е. построим взаимнооднозначное отображение τ интервала (a, b) вещественной прямой $\mathbb{R}^1$ на кривую φ . Число t назовем *параметром*, а кривую, снабженную параметром, — *параметризованной кривой*. Возможны различные параметризации. Например, если t интерпретировать как время, то мы можем измерять его в часах, минутах, секундах и т. п. и будем при этом получать разные параметризации²⁵. Композиция отображений $\chi \circ \tau$ пе-

²⁵Иными словами одной и той же точке при разных параметризациях будут поставлены в соответствие разные числа. Скажем, это может быть 1, если время измеряется в часах, 60 — если в минутах, $\frac{1}{24}$ — если в сутках, и т.д.

реводит точку из $\mathbb{R}^1$ в точку из $\mathbb{R}^n$ и дает часто применяемый параметрический способ задания кривой φ :

$$\mathbf{x}(t) = (x^1(t), \dots, x^n(t)). \quad (6.1)$$

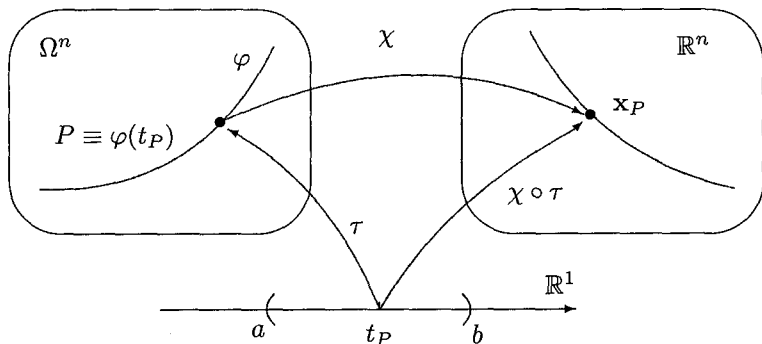


Рис. 7. Координатное отображение $\chi : P \mapsto x_P$ и параметризация кривой $\tau : t_P \mapsto P$.

При изменении параметра t его образ перемещается по кривой φ . Скорость $\vec{v}$ этого перемещения есть предел

$$\vec{v} \equiv \lim_{\Delta t \rightarrow 0} \frac{\varphi(t + \Delta t) - \varphi(t)}{\Delta t} = d_t \varphi.$$

Используя координатное отображение, то же можно записать иначе:

$$\vec{v} \equiv d_t \mathbf{x} = (d_t x^1, \dots, d_t x^n) \equiv (v^1, \dots, v^n).$$

Числа $v^i = d_t x^i$, $i = 1, \dots, n$ называются компонентами касательного вектора к кривой φ в точке P с координатами $\mathbf{x}$.

Если через точку P мы проведем другую кривую, то получим в этой точке еще один касательный вектор. Иной касательный вектор мы также получим, если перепараметризуем кривую φ (т. е. воспользуемся какой-либо другой параметризацией кривой). Поскольку через каждую точку Ω^n можно провести бесконечное множество кривых и каждую

параметризовать бесконечным числом способов, то в каждой точке Ω^n существует бесконечно много касательных векторов. Все касательные векторы не являются элементами Ω^n , а принадлежат *касательному к Ω^n в точке P пространству*. Это пространство обозначается $T\Omega^n_P$, или T_P . Так, если Ω^n — двумерная поверхность, скажем, сфера S^2 , то TS^2 — касательная к ней плоскость. В каждой точке Ω^n имеется свое касательное пространство, часть которых показана на рис.8.

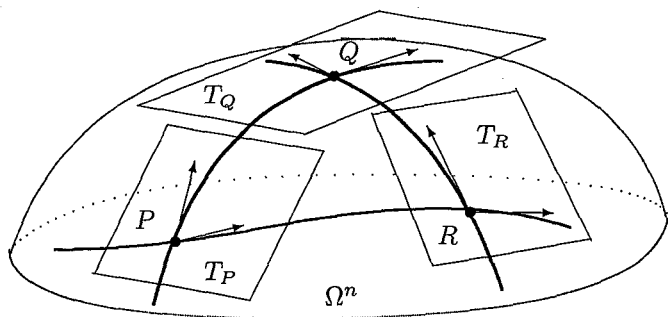


Рис. 8. Примеры касательных пространств.

К каждой точке параметризованной кривой, заданной в Ω^n , «прикреплено» касательное пространство. Любая такая кривая в каждой своей точке «выбирает» из соответствующего касательного пространства касательный вектор. Сама кривая задает касательную прямую, а параметризация — направление и длину вектора. Если указать семейство параметризованных кривых, заполняющих Ω^n так, чтобы через каждую точку Ω^n проходила одна и только одна кривая, то на Ω^n окажется заданным и *касательное векторное поле*.

Таким образом, семейство параметризованных кривых определяет на Ω^n векторное поле. Важно, однако, то, что при достаточно общих условиях верно и обратное: векторное поле, заданное в области Ω^n , определяет семейство кривых. Они являются решениями обыкновенного дифференциального уравнения

$$d_t \mathbf{x} = \vec{v}, \quad (6.2)$$

что в компонентной форме эквивалентно системе уравнений:

$$\begin{aligned} d_t x^1 &= v^1 \\ &\vdots \\ d_t x^n &= v^n. \end{aligned} \quad (6.3)$$

Эти кривые называются *фазовыми кривыми* поля $\vec{v}$.

При изменении параметра t меняться может не только положение точки на кривой, но и значения каких-либо функций, на этой кривой определенных. Определим в области Ω^n дифференцируемую функцию h , т. е. каждой точке Ω^n поставим в соответствие число — значение функции h в этой точке. Тогда на кривой φ также будет задана функция h . С изменением параметра t меняется точка кривой φ , а следовательно, и значение функции $h(\varphi(t)) \equiv h(t)$, которое параметру t соответствует. Скорость этого изменения v_h можно определить как предел

$$\vec{v}_h \equiv \lim_{\Delta t \rightarrow 0} \frac{h(t + \Delta t) - h(t)}{\Delta t} = d_t h.$$

Поскольку при координатном отображении χ точка Ω^n отображается в точку $x = (x^1, \dots, x^n)$ из $\mathbb{R}^n$, функция h , определенная на Ω^n , оказывается эквивалентной функции n переменных $h(x^1, \dots, x^n)$, определенной на $\mathbb{R}^n$. Функция $h(\varphi(t))$, заданная на кривой $\varphi(t)$, таким образом, эквивалентна функции $h(x^1(t), \dots, x^n(t))$, заданной на кривой, определяемой уравнениями:

$$\begin{aligned} x^1 &= x^1(t) \\ &\vdots \\ x^n &= x^n(t). \end{aligned} \quad (6.4)$$

Дифференцируя h по t как сложную функцию, получаем

$$\vec{v}_h \equiv d_t h(x(t)) = \partial_{x^1} h d_t x^1 + \dots + \partial_{x^n} h d_t x^n.$$

Если на Ω^n определить другую функцию, например g , то скорость ее изменения v_g вдоль той же кривой φ выразится очень похожей формулой:

$$\vec{v}_g \equiv d_t g = \partial_{x^1} g d_t x^1 + \dots + \partial_{x^n} g d_t x^n.$$

Сравнивая обе формулы, находим, что оператор дифференцирования по параметру кривой имеет вид:

$$\begin{aligned} d_t &= d_t x^1 \partial_{x^1} + \dots + d_t x^n \partial_{x^n} = \\ &= v^1 \partial_{x^1} + \dots + v^n \partial_{x^n}. \end{aligned} \quad (6.5)$$

Величины $v^i = d_t x^i$ являются компонентами касательного вектора $\vec{v}$ и определяются кривой и ее параметризацией. По этой причине выражение (6.5) называется *оператором дифференцирования по направлению вектора $\vec{v}$* в точке x и обозначается $L_{\vec{v}}|_x$, а число $L_{\vec{v}}h|_x$ называется *производной функции h по направлению вектора $\vec{v}$* в точке x .

Если на Ω^n задано касательное векторное поле, то в каждой точке Ω^n можно вычислить производную функции f по направлению соответствующего этой точке касательного вектора, и, таким образом, задать на Ω^n новую функцию

$$L_{\vec{v}}f|_x = v^1 \partial_{x^1} f + \dots + v^n \partial_{x^n} f.$$

Если $f \in C^1(\Omega^n)$ — дифференцируемая функция, то $L_{\vec{v}}f \in C^0(\Omega^n)$ — непрерывная функция. Оператор $L_{\vec{v}}: C^1(\Omega^n) \rightarrow C^0(\Omega^n)$

$$L_{\vec{v}} = v^1 \partial_{x^1} + \dots + v^n \partial_{x^n}$$

называется *линейным дифференциальным оператором 1-го порядка*. Линейность этого оператора очевидна. Можно показать, что он обладает также следующими свойствами:

- 1) $L_{\vec{v}}(fg) = fL_{\vec{v}}g + gL_{\vec{v}}f$,
- 2) $L_{\vec{u}+\vec{v}} = L_{\vec{u}} + L_{\vec{v}}$,
- 3) $L_f \vec{v} = fL_{\vec{v}}$.

Здесь f, g — произвольные дифференцируемые функции.

Уравнение вида

$$L_{\vec{v}}f = 0, \quad (6.6)$$

где f — дифференцируемая функция, определенная в области Ω^n , а $\vec{v}$ — векторное поле, заданное на Ω^n , называется *линейным однородным дифференциальным уравнением 1-го порядка*. Решением уравнения (6.6) является множество дифференцируемых функций, производные которых по направлению векторного поля $\vec{v}$ равны нулю. Поле $\vec{v}$

называется *характеристическим векторным полем* уравнения (6.6). Фазовые кривые этого поля называются *характеристиками*, а уравнение

$$d_t \mathbf{x} = \vec{v}, \quad (6.7)$$

аналогичное уравнению (6.2), — *уравнением характеристик* для уравнения (6.6).

Уравнения (6.6) и (6.7) тесно связаны. Действительно, поле $\vec{v}$ определяет на Ω^n кривые — характеристики, которые являются решениями уравнения (6.7). Дифференцируя решение f уравнения (6.6) вдоль этих кривых, получим

$$d_t f = L_{\vec{v}} f = 0.$$

Функция f постоянна на характеристике, т.е. на решении уравнения (6.7). Функции, постоянные на решении какого-либо уравнения, называются *первыми интегралами* этого уравнения. В нашем случае f есть первый интеграл уравнения (6.7), а связь уравнений (6.6) и (6.7) оказывается следующей: решением однородного дифференциального уравнения в частных производных (6.6) является первый интеграл уравнения характеристик (6.7).

6.2. Задача Коши для уравнения 1-го порядка

Если функция f постоянна на характеристике, то очевидно, что функции $f + \text{const}$ или $f \cdot \text{const}$ также будут постоянными. Оператор $L_{\vec{v}}: C^1(\Omega^n) \rightarrow C^0(\Omega^n)$, таким образом, не является взаимнооднозначным, а уравнение (6.6) не имеет единственного решения.

В том случае, когда уравнение (6.6) моделирует какое-либо явление (например, перенос некоторого свойства f полем скорости $\vec{v}$), что требует нахождения единственного решения, в формулировку задачи необходимо включать дополнительные условия. Они должны так ограничить область определения оператора $L_{\vec{v}}$, чтобы он стал взаимнооднозначным. Часто эти условия формулируются в виде задания на некотором множестве $\Gamma \subset \Omega^n$ значений функции $f|_{\Gamma} = f_0$. Множество Γ должно быть выбрано так, чтобы задача имела одно и только одно решение. Это возможно не при любом выборе Γ . Дело в том, что поскольку f постоянна на характеристиках, мы можем задавать произвольные значения f_0 лишь в тех случаях, когда характеристика имеет с множеством

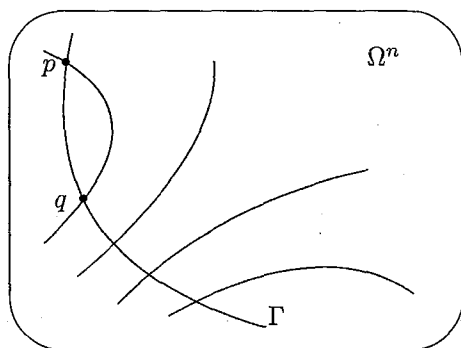


Рис. 9. Пример неудачного расположения множества Γ : в точках p и q задаваемые значения должны быть одинаковы.

Γ только одну общую точку. Иначе, значения f_0 , принадлежащие одной и той же характеристике, не будут независимыми и должны равняться друг другу (рис.9). Обычно в качестве Γ берут именно такое подмножество Ω^n , которое с каждой характеристикой имеет только одну общую точку. Размерность Γ оказывается на единицу меньше размерности Ω^n .

В итоге формулируется задача

$$L_{\vec{v}} f = 0, \quad (6.8)$$

$$f|_{\Gamma} = f_0, \quad (6.9)$$

которая называется *задачей Коши* для уравнения (6.8). Условие (6.9) носит название *начального условия*. Такое название связано с тем, что параметризация характеристик может быть выбрана такой, что нулевое значение параметра $t = 0$ (начало моделируемого процесса) для каждой характеристики будет лежать на множестве Γ .

Вдоль характеристик $\varphi(t)$ задача (6.8)–(6.9) сводится к задаче Коши для обыкновенного дифференциального уравнения, описывающего изменение f с изменением параметра t :

$$d_t f = 0, \quad (6.10)$$

$$f|_{t=0} = f_0. \quad (6.11)$$

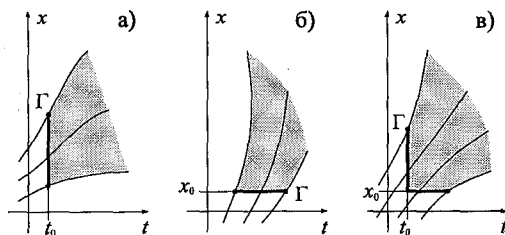


Рис. 10. Варианты задания множества Γ . Области влияния условий, заданных на Γ (начальных/краевых условий) заштрихованы.

Интегрируя (6.10) по параметру от 0 до t , получаем $f(\varphi(t)) = f_0$, т. е. решение постоянно на характеристике и равно начальному значению.

Часто одна из функций координатного отображения $\Omega^n \rightarrow \mathbb{R}^n$ интерпретируется как время t , а кривые на Ω^n параметризуются значениями времени. Остальные координаты в таком случае считаются пространственными. Уравнение (6.6) в этом случае принимает вид

$$\partial_i f + u^i \partial_{x^i} f = 0, \quad i = 1, \dots, n-1 \quad (6.12)$$

и называется *уравнением переноса* величины f полем скорости $\vec{u} = (u^1, \dots, u^{n-1})$. Система уравнений, определяющая характеристики, содержит теперь $(n-1)$ уравнение

$$d_t x^i = u^i, \quad i = 1, \dots, n-1,$$

поскольку первое уравнение превращается в тождество $d_t t = 1$. Если уравнение (6.12) трактовать как закон изменения во времени функции f в точке пространства в результате движения среды из невзаимодействующих частиц, то характеристики оказываются траекториями движения частиц среды, а само уравнение (6.12) — *уравнением движения в эйлеровых*²⁶ *координатах*. С другой стороны, уравнение (6.10), справедливое на характеристике, есть то же *уравнение движения*, за-

²⁶По имени Леонарда Эйлера (Euler) (1707–1783) — математика, физика и астронома. Он родился и учился в Швейцарии, работал в Германии и России.

писанное в терминах лагранжеских²⁷ координат, в качестве которых, как правило, берут значения пространственных координат на множестве Γ .

Хотя множество Γ в известном смысле произвольно, его предпочтительный вид обычно диктуется физикой моделируемого явления. Часто в качестве Γ выбирают подмножество точек Ω^n , для которых фиксируется либо значение параметра $t = 0$ (собственно *начальное условие*), либо какая-либо пространственная координата $x^i = x_0^i$ (*краевое условие*), либо их комбинация (*смешанное краевое условие*). Для одномерного случая все три варианта задания множества Γ показаны на рис.10.

6.3. Неоднородное и квазилинейное уравнения

Если правая часть уравнения (6.6) отлична от нуля, оно называется *неоднородным линейным дифференциальным уравнением*:

$$L_{\vec{v}} f = g. \quad (6.13)$$

Здесь $\vec{v} = \vec{v}(\mathbf{x})$; $g = g(\mathbf{x})$. В компонентной форме то же уравнение запишется так

$$v^i \partial_{x^i} f = g.$$

Характеристики неоднородного уравнения также являются решениями уравнения (6.7) и, значит, совпадают с характеристиками однородного уравнения. Эволюция функции f вдоль характеристик описывается обыкновенным дифференциальным уравнением

$$d_t f = g. \quad (6.14)$$

Решение f изменяется со скоростью g , а не остается постоянным, как это было в однородном случае.

Выбирая нехарактеристическое подмножество $\Gamma \subset \Omega^n$ и задавая на нем начальные значения функции $f|_{\Gamma} = f_0$, ограничивающие область определения оператора $L_{\vec{v}}$, мы можем поставить задачу Коши. Вдоль характеристик (уравнение (6.14)) она легко интегрируется, и решение имеет вид:

$$f(\varphi(t)) = f_0 + \int_0^t g(\varphi(\tau)) d\tau.$$

²⁷Лагранж (Lagrange) Жозеф Луи (1736–1813) — французский математик и механик.

До сих пор поле $\vec{v}$ считалось не зависящим от решения. Сняв это ограничение, мы получим *квазилинейное уравнение*

$$L_{\vec{v}}f = g,$$

где $\vec{v} = \vec{v}(x, f)$. Вдоль характеристик решение f меняется со скоростью g

$$d_t f = g(x). \quad (6.15)$$

Как и прежде, векторное поле $\vec{v}$ определяет характеристики уравнения (6.15)

$$d_t x = \vec{v}(x, f). \quad (6.16)$$

Однако в квазилинейном случае для нахождения характеристик одного уравнения (6.16) недостаточно: поле $\vec{v}$ само зависит от решения f . Для построения характеристик необходимо решать систему уравнений (6.15), (6.16).

Задача Коши для квазилинейного уравнения ставится аналогично разобранным выше случаям: выбирается нехарактеристическое подмножество $\Gamma \subset \Omega^n$ размерности $(n-1)$, на котором задаются начальные условия f_0 .

В качестве примера рассмотрим простейший вид квазилинейного уравнения — так называемое *невязкое уравнение Бюргерса*²⁸:

$$\partial_t u + u \partial_x u = 0. \quad (6.17)$$

Оно может служить простой моделью движения жидкости со скоростью u в трубе вдали от стенок в случае пренебрежимо малой вязкости. Пусть $\Omega^2 = \{t \geq 0, x \in [a, b]\}$. Поставим для уравнения (6.17) задачу Коши. В качестве множества Γ выберем множество точек, лежащих на отрезке $\Gamma = \{t = 0, x \in [a, b]\}$, и зададим на Γ значения скорости

$$u|_{t=0} = U(x), \quad x \in [a, b],$$

где $U(x)$ — известная функция. Характеристики уравнения (6.17) определяются системой уравнений:

$$\begin{aligned} d_t x &= u, \\ d_t u &= 0. \end{aligned} \quad (6.18)$$

²⁸Бюргерс (Burgers) Иоханнес Мартинус (1895–1981) — голландский физик.

Здесь последнее уравнение есть уравнение (6.17), записанное на характеристике $\varphi(t)$. Если x_0 — точка пересечения характеристики φ с множеством Γ , то $U(x_0)$ — значение скорости на этой характеристике в момент времени $t = 0$. Таким образом, начальными условиями для системы (6.18) являются равенства:

$$\begin{aligned} x|_{t=0} &= x_0, \\ u|_{t=0} &= U(x_0). \end{aligned} \quad (6.19)$$

Интегрируя систему (6.18) от $t = 0$ до t с учетом начальных условий (6.19), легко находим:

$$u(\varphi(t)) = U(x_0), \quad (6.20)$$

$$x = x_0 + U(x_0)t. \quad (6.21)$$

Согласно (6.20) скорость u на характеристике постоянна и равна начальному значению $U(x_0)$. Уравнение же (6.21) есть уравнение характеристики $\varphi(t)$. В области Ω^2 этой характеристике соответствуют точки с координатами $(t, x) = (t, x_0 + U(x_0)t)$. Таким образом, характеристиками уравнения (6.17) являются прямые линии с тангенсами углов наклона, равными $1/U(x_0)$ (рис.11). Из уравнений (6.20) и (6.21) следует, что решение уравнения (6.17) может быть записано в виде

$$u(t, x + U(x)t) = U(x).$$

Обратите внимание на две особенности полученного решения. Во-первых, начальные данные, заданные на отрезке $[a, b]$, могут определять решение не во всей области Ω^2 , а только в сравнительно небольшой ее части. Так, если $U(x) > 0$ для любого $x \in [a, b]$, то начальные данные позволяют найти решение только на множестве точек, ограниченном координатными линиями $t = 0$, $x = b$ и характеристикой $x = a + U(a)t$. На рис.11 а и рис.11 б показана типичная ситуация. Область, лежащая слева от характеристики, проходящей через точку a , никак не связана с начальными значениями. Решения же в области справа от той же характеристики полностью определяются начальным профилем скорости $U(x)$ на отрезке $[a, b]$. Эта область называется *областью влияния* начальных данных (на рисунке она заштрихована).

Во-вторых, если начальный градиент скорости на отрезке $[a, b]$ отрицателен (рис.11 в), то возможна ситуация, когда характеристики в области Ω^2 пересекутся. К моменту времени t_c первоначально пологий профиль скорости трансформируется в скачок скорости в точке x_c , градиент скорости станет бесконечным и наступит так называемая *градиентная катастрофа*. Решение

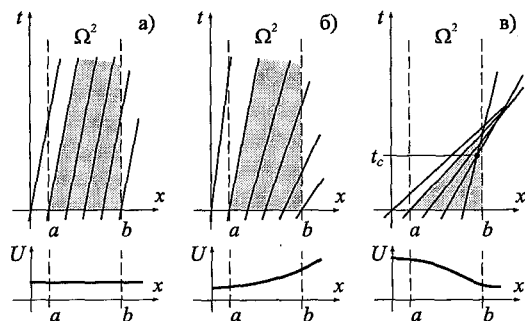


Рис. 11. Характеристики уравнения 1-го порядка и соответствующие им начальные профили скорости.

задачи существует локально во времени от 0 до t_c , где оно перестает быть единственным.

7. Системы дифференциальных уравнений

7.1. Общие соображения

Обобщением одного дифференциального уравнения 1-го порядка можно считать как системы дифференциальных уравнений 1-го порядка, так и уравнения более высоких порядков. Однако между этими двумя вариантами обобщений нет взаимнооднозначного соответствия. Лишь в частном случае линейной системы с постоянными коэффициентами верна следующая

Теорема. Любую систему n линейных дифференциальных уравнений с постоянными коэффициентами для n неизвестных функций можно свести к одному дифференциальному уравнению с постоянными коэффициентами для любой из неизвестных функций.

Напротив, любое дифференциальное уравнение порядка выше первого эквивалентно системе дифференциальных уравнений 1-го порядка. Рассмотрим простейший случай, когда Ω^n представляет собой двумерную область Ω^2 , а искомое решение является функцией двух аргументов. Случай многих независимых переменных можно найти в руководствах по теории дифференциальных уравнений в частных производных (см. Список литературы).

Вводя координаты, отображающие точку $P \in \Omega^2$ в точку $x \in \mathbb{R}^2$, запишем уравнение 2-го порядка общего вида:

$$F(x, y, f, \partial_x f, \partial_y f, \partial_{xx} f, \partial_{xy} f, \partial_{yy} f) = 0. \quad (7.1)$$

Здесь $x \equiv (x, y)$ — координаты точки P из области Ω^2 . Для того чтобы получить эквивалентную систему уравнений, введем дифференцируемые функции g и h такие, что $g \equiv \partial_x f$ и $h \equiv \partial_y f$. Подстановка этих функций в уравнение приводит его к виду:

$$F(x, y, f, g, h, \partial_x g, \partial_y g, \partial_y h) = 0.$$

Здесь смешанная производная $\partial_{xy} f$ выражена через производную функции g .

Рассматривая далее определения функций g и h как уравнения, связывающие эти функции с производными функции f , получим систему дифференциальных уравнений. Для того, чтобы эта система стала эквивалентной уравнению (7.1), необходимо учесть то, что смешанную производную $\partial_{xy}f$ можно выразить и через производную функции h . Другими словами, должны выполняться равенства $\partial_y g = \partial_x h = \partial_{xy}f$. В итоге получаем систему уравнений:

$$\begin{cases} F(x, y, f, g, h, \partial_x g, \partial_y h) = 0, \\ \partial_x f - g = 0, \\ \partial_y f - h = 0, \\ \partial_y g - \partial_x h = 0, \end{cases} \quad (7.2)$$

эквивалентную уравнению 2-го порядка с двумя независимыми переменными общего вида. Любое решение (7.1) является решением (7.2), и наоборот.

7.2. Линейная система с двумя независимыми переменными

В линейном случае систему m уравнений для m -мерного вектора неизвестных функций f можно записать в виде:

$$A_1 \partial_x f + A_2 \partial_y f = g. \quad (7.3)$$

Здесь $f = (f^1, \dots, f^m)$; $g = (g^1, \dots, g^m)$; A_1 и A_2 — квадратные матрицы $m \times m$. Компоненты матриц, как и компоненты вектора g , в общем случае могут быть функциями точки x и решения f . Для простоты, однако, будем считать их постоянными.

По крайней мере, одна из двух матриц неособая²⁹. Пусть это будет матрица A_1 и, значит, существует обратная ей матрица A_1^{-1} . Матрица A_2 может быть любой. Умножая слева обе части уравнения (7.3) на A_1^{-1} , приходим к эквивалентному уравнению

$$\partial_x f + A \partial_y f = h, \quad (7.4)$$

где введены обозначения: $A \equiv A_1^{-1} A_2$ и $h \equiv A_1^{-1} g$.

²⁹В противном случае исходная система сводится к системе меньшего числа уравнений.

7.2.1. Замена переменных

Если матрица A диагональная, то система уравнений (7.4) распадается на m независимых дифференциальных уравнений 1-го порядка для компонент вектора неизвестных f . Такие уравнения мы рассматривали в предыдущих параграфах. В этом случае в каждой точке Ω^2 имеется m характеристик, по которым распространяются компоненты решения, каждая компонента по своей характеристике.

Если же A — матрица общего вида, можно попробовать произвести замену переменных так, чтобы уравнение (7.4) относительно нового вектора f' содержало диагональную матрицу. Преобразуем вектор f с помощью пока неизвестной матрицы преобразования Λ

$$f = \Lambda f' \quad (7.5)$$

и подставим в уравнение (7.4)

$$\partial_x \Lambda f' + A \partial_y \Lambda f' = h.$$

Дифференцируя и перегруппировывая слагаемые, получим

$$\Lambda \partial_x f' + \Lambda A \partial_y f' = h - (\partial_x \Lambda + A \partial_y \Lambda) f'.$$

Поскольку нас интересует невырожденное преобразование Λ , существует обратное преобразование Λ^{-1} . Умножим полученное уравнение слева на Λ^{-1}

$$\partial_x f' + \Lambda^{-1} \Lambda A \partial_y f' = \Lambda^{-1} (h - (\partial_x \Lambda + A \partial_y \Lambda) f') \equiv h'. \quad (7.6)$$

Матрица $\Lambda^{-1} \Lambda A$ подобна матрице A в том смысле, что имеет те же собственные числа, что и матрица A (такое преобразование матрицы A называется *преобразованием подобия*). Выбором матрицы Λ можно превратить матрицу $\Lambda^{-1} \Lambda A$ в диагональную. Из курса алгебры известно, что если $(m \times m)$ матрица A имеет m различных действительных собственных значений, то ее собственные векторы линейно независимы и матрица, составленная из этих собственных векторов, будет как раз искомой матрицей Λ . В этом случае система (7.4) сводится к виду

$$\partial_x f' + D \partial_y f' = h', \quad (7.7)$$

где $D \equiv \Lambda^{-1} A \Lambda$ — диагональная матрица (единственные отличные от нуля элементы матрицы D суть стоящие на главной диагонали собственные числа матрицы A).

Характеристики, по которым распространяются компоненты вектора f' , находятся из уравнений:

$$\begin{aligned} d_t x &= 1, \\ d_t y &= d_{ii} = \nu_i, \quad i = 1, \dots, m. \end{aligned} \quad (7.8)$$

Здесь d_{ii} — i -й диагональный элемент матрицы D , а ν_i — i -е собственное число матрицы A .

7.2.2. Исследование типа системы

По определению ν называется *собственным числом* матрицы A , если существует ненулевой вектор ξ , такой, что

$$A\xi = \nu\xi. \quad (7.9)$$

Вектор ξ в этом случае называется *собственным вектором* матрицы A . Собственный вектор, как видно из (7.9), определяется с точностью до постоянного множителя.

Для нахождения собственных чисел заметим, что уравнение (7.9) эквивалентно однородной системе уравнений

$$(A - \nu I)\xi = 0, \quad (7.10)$$

которая имеет нетривиальное решение только в том случае, если

$$\det(A - \nu I) = 0. \quad (7.11)$$

Уравнение (7.11) называется *характеристическим уравнением* матрицы A , является уравнением m -й степени относительно ν и имеет m корней. Если $m = 2$, характеристическое уравнение имеет два корня, и в зависимости от вида корней (собственных чисел) система (7.4) принадлежит к одному из трех типов: *гиперболическому* — если оба корня действительны и отличны от нуля, *эллиптическому* — корни комплексны и *параболическому* — один из корней равен нулю.

Если система состоит из большего числа уравнений, возможны и промежуточные типы (например, если имеется n действительных и m пар комплексных корней), которые, однако, в математической физике практически не встречаются. В случае уравнений с переменными коэффициентами и нелинейных уравнений тип системы может зависеть от рассматриваемой области, а в последнем случае — и от решения. Такие системы называются *системами смешанного типа*. И, наконец, если система уравнения эквивалентна дифференциальному уравнению высшего порядка, то это уравнение считается принадлежащим тому же типу, к которому относится и система уравнений.

Вернемся, однако, к системам двух уравнений с двумя независимыми переменными. В *гиперболическом* случае характеристическое уравнение имеет в каждой точке области определения Ω^2 два вещественных корня, которым соответствуют два семейства характеристик, удовлетворяющих уравнению (7.11). Дополнительные условия можно ставить на любой кривой, не являющейся характеристикой и пересекающей каждую характеристику не более одного раза. В противном случае эти дополнительные условия не могут задаваться произвольно. Так же как и в случае одного уравнения, могут быть поставлены задача Коши, краевая и смешанная задачи.

Если корнями характеристического уравнения является пара комплексно сопряженных чисел, мы имеем систему *эллиптического* типа. В этом случае вещественных характеристик нет, и формально мы можем ставить задачу Коши на любой кривой $\Gamma \in \Omega^2$. Однако такая задача для эллиптических систем оказывается некорректной (см. стр.87) и потому не ставится. Обычными для систем эллиптического типа являются краевые задачи.

И, наконец, *параболические* системы имеют один, отличный от нуля, корень характеристического уравнения и, следовательно, одно семейство характеристик. В этом случае возможна постановка как задачи Коши, так и смешанной краевой задачи.

Резюмируя, можно сказать, что с эллиптическими уравнениями связаны краевые задачи, а для гиперболических и параболических уравнений характерны задачи Коши и смешанные задачи.

8. Некоторые классические задачи

Общее уравнение 2-го порядка с постоянными коэффициентами и двумя независимыми переменными

$$F(x, y, f, \partial_x f, \partial_y f, \partial_{xx} f, \partial_{xy} f, \partial_{yy} f) = 0$$

преобразованием координат $\xi = \xi(x, y)$, $\eta = \eta(x, y)$ может быть приведено к одному из следующих, так называемых *канонических*, видов (способ приведения излагается в любом учебнике математической физики):

— в эллиптическом случае

$$\partial_{\xi\xi} f + \partial_{\eta\eta} f = Q_1(\xi, \eta, f, \partial_{\xi} f, \partial_{\eta} f), \quad (8.1)$$

— в гиперболическом случае

$$\partial_{\xi\xi} f - \partial_{\eta\eta} f = Q_2(\xi, \eta, f, \partial_{\xi} f, \partial_{\eta} f), \quad (8.2)$$

или

$$\partial_{\xi\eta} f = Q_3(\xi, \eta, f, \partial_{\xi} f, \partial_{\eta} f), \quad (8.3)$$

— в параболическом случае

$$\partial_{\xi\xi} f = Q_4(\xi, \eta, f, \partial_{\xi} f, \partial_{\eta} f). \quad (8.4)$$

Как правило, в задачах с гиперболическими и параболическими уравнениями одна из независимых переменных интерпретируется как время, а остальные — как пространственные координаты. В задачах с эллиптическими уравнениями обе независимые переменные считаются пространственными координатами. При такой трактовке задачи первого вида описывают развивающиеся во времени процессы и называются *нестационарными*. Задачи второго вида описывают не зависящие от времени распределения величин в пространстве и потому называются *стационарными*. Будем обозначать время через t , а пространственные координаты через x, y и обсудим некоторые типичные задачи для уравнений 2-го порядка.

8.1. Задача Коши для волнового уравнения

Рассмотрим простейшее уравнение типа (8.2), так называемое *одномерное волновое уравнение*:

$$\partial_{tt}f - c^2 \partial_{xx}f = 0, \quad (x, t) \in \Omega = \{t > 0, x \in \mathbb{R}^1\}. \quad (8.5)$$

Задача Коши состоит в отыскании функции $f(x, t)$, дважды дифференцируемой по своим аргументам, при всех $t > 0$ по значениям f и $\partial_t f$, заданным на множестве $\{x, t = 0\}$. Начальными условиями, таким образом, являются

$$f|_{t=0} = f_0(x), \quad (8.6)$$

$$\partial_t f|_{t=0} = l_0(x). \quad (8.7)$$

- 1) *Представление уравнения (8.5) в виде системы уравнений 1-го порядка.*

Введем обозначения³⁰

$$g \equiv \partial_t f, \quad h \equiv -c \partial_x f. \quad (8.8)$$

После их подстановки в уравнение (8.5) оно принимает вид:

$$\partial_t g + c \partial_x h = 0. \quad (8.9)$$

Второе уравнение, связывающее функции g и h , получим из условия согласованности введенных определений (8.8). В силу того, что смешанную производную $\partial_{tx}f$ можно вычислить и как $\partial_x g$, и как $\frac{1}{c} \partial_t h$, две последние производные должны быть равны друг другу, т.е.

$$\partial_t h + c \partial_x g = 0. \quad (8.10)$$

Таким образом, уравнение 2-го порядка (8.5) эквивалентно системе уравнений 1-го порядка (8.8)–(8.10). Заметим при этом, что подсистема (8.9)–(8.10) не зависит от уравнений (8.8) и может

³⁰Традиционное введение коэффициента $-c$ в определение функции h обеспечивает симметричный вид получаемых уравнений. Вполне можно обойтись и без этого.

быть решена отдельно. Начальные условия легко получаются из (8.6). Для функции g из условия (8.7) имеем

$$g|_{t=0} = l_0(x), \quad (8.11)$$

а дифференцируя условие (8.6), получим

$$h|_{t=0} = -c \partial_x f_0. \quad (8.12)$$

2) *Запись системы (8.9), (8.10) в матричной форме.*

Введем в рассмотрение вектор неизвестных $\mathbf{f} = \begin{pmatrix} g \\ h \end{pmatrix}$ и запишем систему в виде:

$$\partial_t \mathbf{f} + A \partial_x \mathbf{f} = 0. \quad (8.13)$$

Матрица A легко находится из условия

$$A \partial_x \mathbf{f} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \partial_x g \\ \partial_x h \end{pmatrix} = \begin{pmatrix} c \partial_x h \\ c \partial_x g \end{pmatrix} \Rightarrow A = \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix}.$$

3) *Определение типа системы.*

Для того чтобы определить тип системы (8.13) и, следовательно, уравнения (8.5), надо знать собственные числа матрицы A . Они находятся при решении задачи на собственные значения из условия существования нетривиального решения ξ однородной системы линейных алгебраических уравнений

$$A\xi = \nu\xi \quad \text{или} \quad (A - \nu I)\xi = 0. \quad (8.14)$$

Тривиальное решение $\xi = 0$ у системы (8.14) существует всегда, а нетривиальное — только в том случае, когда система имеет нулевой определитель. Равенство нулю определителя системы называется *характеристическим уравнением*

$$\det(A - \nu I) = \begin{vmatrix} -\nu & c \\ c & -\nu \end{vmatrix} = \nu^2 - c^2 = 0.$$

Решим его и получим два вещественных и различных собственных числа $\nu_1 = c$ и $\nu_2 = -c$. Система (8.13), таким образом, называется *гиперболической*.

4) Приведение к эквивалентной системе с диагональной матрицей.

В силу гиперболичности система (8.13) может быть приведена к эквивалентной ей системе (7.7) с диагональной матрицей D , которая связана с исходной матрицей A преобразованием подобия $D = \Lambda^{-1}A\Lambda$. Столбцами преобразующей матрицы Λ служат собственные векторы ξ матрицы A . Эти векторы являются решениями однородной системы линейных алгебраических уравнений (8.14). Каждому собственному числу отвечает собственный вектор. Различным собственным числам соответствуют линейно независимые собственные векторы. Подставляя найденные ранее значения ν_1 и ν_2 , получим две системы для определения векторов ξ :

$$\begin{cases} -c\xi^1 + c\xi^2 = 0, \\ c\xi^1 - c\xi^2 = 0, \end{cases} \quad \begin{cases} c\xi^1 + c\xi^2 = 0, \\ c\xi^1 + c\xi^2 = 0. \end{cases}$$

При подстановке в систему собственных чисел она становится линейно зависимой, имеет бесконечно много решений

$$\xi^1 = \xi^2 \quad \text{и} \quad \xi^1 = -\xi^2 \quad (8.15)$$

и позволяет однозначно определить лишь направления собственных векторов. Сами же векторы определяются с точностью до постоянного множителя. Таким образом, если одна из компонент вектора выбрана произвольно, уравнения (8.15) позволяют найти вторую. Например, если положить $\xi^1 = 1$, тогда собственными векторами будут

$$\xi_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \xi_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

а матрица

$$\Lambda = (\xi_1 | \xi_2) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

будет матрицей преобразования. Возможен и иной выбор ξ^1 . Иной тогда будет и матрица Λ . (Можно ли положить $\xi^1 = 0$?) Обратная матрица Λ^{-1} легко вычисляется и равна

$$\Lambda^{-1} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Дальнейшие действия очевидны (см. предыдущую лекцию). В итоге получаем систему уравнений с диагональной матрицей

$$\partial_t \mathbf{f}' + D \partial_x \mathbf{f}' = 0, \quad (8.16)$$

эквивалентную (8.13). Здесь

$$\mathbf{f}' = \Lambda^{-1} \mathbf{f} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} g \\ h \end{pmatrix} = \frac{1}{2} \begin{pmatrix} g+h \\ g-h \end{pmatrix} \equiv \begin{pmatrix} g' \\ h' \end{pmatrix},$$

$$D = \Lambda^{-1} A \Lambda = \begin{pmatrix} c & 0 \\ 0 & -c \end{pmatrix}. \quad (8.17)$$

В компонентной форме система (8.16) выглядит так:

$$\begin{cases} \partial_t g' + c \partial_x g' = 0, \\ \partial_t h' - c \partial_x h' = 0. \end{cases} \quad (8.18)$$

Начальные условия нетрудно получить исходя из вида компонент вектора $\mathbf{f}'$:

$$g'|_{t=0} = \frac{1}{2}(g+h)_{t=0} = \frac{1}{2}(l_0 - c \partial_x f_0) \equiv g'_0,$$

$$h'|_{t=0} = \frac{1}{2}(g-h)_{t=0} = \frac{1}{2}(l_0 + c \partial_x f_0) \equiv h'_0. \quad (8.19)$$

5) Решение задачи (8.18), (8.19).

Система (8.18) состоит из двух независимых уравнений, которые были рассмотрены ранее. Характеристики первого суть семейство прямых $x = x_0 + ct$, второго — прямые $x = x_0 - ct$ (рис.12).

На характеристиках решения обоих уравнений постоянны, и начальные профили $g'_0(x)$, $h'_0(x)$ распространяются независимо друг от друга, каждый по своим характеристикам (рис.12 б). Через любую точку (x, t) проходят две характеристики: $x = x_{01} + ct$ и $x = x_{02} - ct$ (рис.12 а). Таким образом, решение задачи (8.18), (8.19), которое есть вектор с компонентами $g'(x, t)$ и $h'(x, t)$, легко находится: в точку (x, t) вдоль характеристики $x = x_{01} + ct$ попадает значение $g'_0(x_{01})$, а вдоль характеристики $x = x_{02} - ct$ —

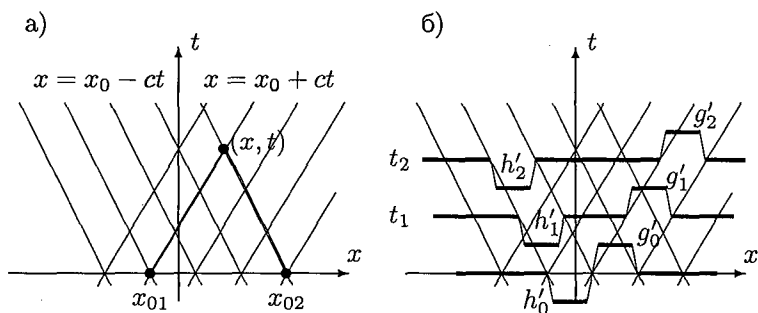


Рис. 12. Характеристики и решение системы (8.18) при $c > 0$. а) — решение в точке (x, t) определяется начальными условиями в точках x_{01} и x_{02} ; б) — начальные профили g'_0, h'_0 распространяются вдоль характеристик независимо друг от друга.

значение $h'_0(x_{02})$. Поскольку $x_{01} = x - ct$, а $x_{02} = x + ct$, окончательно получаем

$$\begin{aligned} \mathbf{f}' &= \begin{pmatrix} g'(x, t) \\ h'(x, t) \end{pmatrix} = \begin{pmatrix} g'_0(x_{01}) \\ h'_0(x_{02}) \end{pmatrix} = \begin{pmatrix} g'_0(x - ct) \\ h'_0(x + ct) \end{pmatrix} = \\ &= \frac{1}{2} \begin{pmatrix} l_0(x - ct) - c\partial_x f_0(x - ct) \\ l_0(x + ct) + c\partial_x f_0(x + ct) \end{pmatrix}. \end{aligned} \quad (8.20)$$

6) Построение решения исходной задачи.

Для построения решения задачи (8.5)–(8.7) воспользуемся каким-либо из уравнений (8.8) — можно взять любое. Пусть это будет уравнение $\partial_t f = g$. Из равенства $\mathbf{f} = \Lambda \mathbf{f}'$ имеем

$$g(x, t) = g'(x, t) + h'(x, t) = \partial_t f.$$

Используя выражения для компонент вектора $\mathbf{f}'$ (8.20), найдем

$$\begin{aligned} \partial_t f &= \frac{1}{2} (l_0(x - ct) - c\partial_x f_0(x - ct)) + \\ &+ \frac{1}{2} (l_0(x + ct) + c\partial_x f_0(x + ct)). \end{aligned}$$

Проинтегрировав это выражение от 0 до t , получим

$$f(x, t) - f(x, 0) = I_1 + I_2,$$

где

$$I_1 = \frac{1}{2} \int_0^t l_0(x - ct) dt + \frac{1}{2} \int_0^t l_0(x + ct) dt$$

и

$$I_2 = \frac{c}{2} \int_0^t \partial_x f_0(x + ct) dt - \frac{c}{2} \int_0^t \partial_x f_0(x - ct) dt.$$

Для вычисления интегралов в I_1 сделаем замену переменных $\zeta = x + ct$, $\eta = x - ct$, и учтем, что $d\zeta = cdt$, $d\eta = -cdt$ (в каждом интеграле интегрирование ведется по t , а переменная x выступает в качестве параметра). В результате получим:

$$\begin{aligned} I_1 &= \frac{1}{2c} \int_x^{x+ct} l_0(\zeta) d\zeta - \frac{1}{2c} \int_x^{x-ct} l_0(\eta) d\eta = \\ &= \frac{1}{2c} \int_x^{x+ct} l_0(\zeta) d\zeta + \frac{1}{2c} \int_{x-ct}^x l_0(\eta) d\eta = \\ &= \frac{1}{2c} \int_{x-ct}^{x+ct} l_0(\eta) d\eta. \end{aligned}$$

При вычислении I_2 мы также воспользуемся заменой переменных: $\eta = x - ct$, $\zeta = x + ct$, и учтем, что $\partial_x f = \partial_\eta f = \partial_\zeta f$ (при вычислении частных производных все аргументы, кроме того, по которому осуществляется дифференцирование, фиксируются). В итоге получим:

$$\begin{aligned} I_2 &= \frac{1}{2} \int_x^{x-ct} \partial_\eta f_0(\eta) d\eta + \frac{1}{2} \int_x^{x+ct} \partial_\zeta f_0(\zeta) d\zeta = \\ &= \frac{1}{2} (f_0(x - ct) - f_0(x)) + \frac{1}{2} (f_0(x + ct) - f_0(x)) = \\ &= \frac{1}{2} (f_0(x - ct) + f_0(x + ct)) - f_0(x). \end{aligned}$$

Наконец, учтем, что $f(x, 0) = f_0(x)$, и окончательно найдем:

$$f(x, t) = \frac{1}{2} (f_0(x - ct) + f_0(x + ct)) + \frac{1}{2c} \int_{x-ct}^{x+ct} l_0(\eta) d\eta. \quad (8.21)$$

Решение волнового уравнения в виде (8.21) известно под названием *формулы Даламбера*³¹.

Исследование типа одномерного уравнения теплопроводности и двумерного уравнения Лапласа³² проведем в упражнениях.

8.2. О постановке задач математической физики

В первой части курса мы рассматривали уравнения, записанные в виде $Af = g$, где под A понимался оператор $A: V \rightarrow V$, а неизвестное f и правая часть g рассматривались как элементы векторного пространства $f, g \in V$. Определение оператора включает задание его области определения $D(A) \subseteq V$ и области значений $R(A) \subseteq V$. Если в общем случае оператор оказывается не взаимнооднозначным, то сужением его области определения можно попытаться добиться взаимной однозначности, что даст возможность построить обратный оператор и тем самым решить задачу.

Это так называемый *операторный формализм*. Физические же задачи, как правило, ставят в другой более традиционной форме: в виде собственно *дифференциального уравнения* (или *системы уравнений*) и совокупности *дополнительных условий* — *начальных, краевых или граничных*. Отсутствие взаимной однозначности означает, что уравнение не обладает единственным решением. Дополнительные условия как раз и позволяют нам из множества решений, удовлетворяющих дифференциальному уравнению, выбрать то единственное, которое отвечает существу задачи. Таким образом, эти условия фактически ограничивают область определения оператора, делая его взаимнооднозначным.

Например, простейшая краевая задача для уравнения 2-го порядка

$$d_{xx}y + q(x)y = f(x),$$

$$x \in [a, b], \quad q, f \in C^0([a, b]), \quad y \in C^2([a, b]),$$

$$y(a) = 0, y(b) = 0$$

³¹ Даламбер (D'Alembert) Жан Лерон (1717–1783) — французский математик, механик и философ-просветитель.

³² Лаплас (Laplace) Пьер Симон (1749–1827) — французский астроном, математик, физик.

в операторной форме будет выглядеть так:

$$Ly \equiv (d_{xx} + qI)y = f,$$

$$L: C^2([a, b]) \rightarrow C^0([a, b]).$$

Область определения оператора $D(L)$ ограничена теми элементами $y \in C^2([a, b])$, которые на концах отрезка $[a, b]$ принимают нулевые значения.

Вид граничных условий, как правило, диктуется физикой задачи. Наиболее часто встречаются два типа граничных условий: *одноточечные* и *двухточечные* (многоточечные). В первом случае условия, задающие область определения оператора, ставятся в одной точке (например, на одном конце отрезка, если задача одномерна) и включают задание значений функции и ее производных. Такие задачи называются *задачами Коши* и характерны для обыкновенных дифференциальных уравнений. Другое название этих задач — *задачи с начальными условиями*. Дело в том, что отрезок, на конце которого ставятся условия, можно трактовать как область времени, а решение дифференциального уравнения рассматривать как меняющийся во времени процесс. В этом случае точку задания граничных условий естественно называть началом отсчета времени, а сами граничные условия — *начальными условиями*.

Второй тип граничных условий ставится на границе (краях) области задания решения. Отсюда и название таких задач — *краевые задачи*. Если одним из физических аргументов решения является время и часть условий ставится в начальный момент времени, то такие задачи также называются *задачами Коши*, *задачами с начальными условиями*, или *смешанными краевыми задачами*. Это наиболее часто встречающийся тип задач. Более того, решения многих краевых задач можно рассматривать как предел решения задачи Коши при $t \rightarrow \infty$.

К числу задач Коши в первую очередь относятся так называемые *эволюционные задачи*, описывающие изменение во времени некоторой физической системы. Дифференциальный оператор в таких задачах содержит производную по времени только первого порядка.

Типичная постановка эволюционной задачи такова.

Пусть состояние системы описывается функцией $u(r, t)$, где t — время, $t \in [0, T]$ и r — радиус-вектор точки в области пространства $r \in \Omega$.

Состояние системы в начальный момент времени $u_0 = u(r, 0)$ известно во всей области Ω и в любой момент времени на границе области $\partial\Omega$. Требуется найти функцию u , удовлетворяющую уравнению

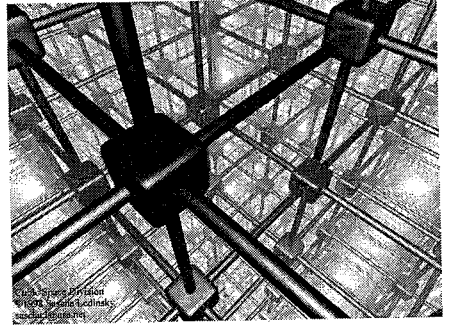
$$\partial_t u = Au,$$

где A — некий дифференциальный оператор. Состояние системы в любой фиксированный момент времени естественно рассматривать как точку некоторого функционального пространства B (функции, принадлежащие этому пространству, зависят только от пространственных координат). Траектория движения этой точки в пространстве B определяет собой эволюцию системы. В качестве пространства B удобно рассматривать банахово пространство (т.е. искать решения задачи, принадлежащие банахову пространству, с тем, чтобы любая сходящаяся последовательность в B сходилась в точке из B).

Выбор пространства B и области определения оператора является важной частью *корректной* постановки задачи. Задача называется *корректной*, если для любых начальных данных из заданного множества решение задачи: 1) существует, 2) единственно и 3) устойчиво. Иначе говоря, математическая задача, соответствующая физическому явлению, должна удовлетворять следующим требованиям:

- 1) среди ограничений на решение (дифференциальное уравнение и граничные условия) не должно быть противоречивых (*существование*);
- 2) если неоднозначность не присуща физике явления, она должна быть исключена (*единственность*);
- 3) можно считать, что математическая задача только тогда правильно описывает физическое явление, когда малые изменения данных задачи приводят к малому изменению решения (*устойчивость*).

Для определения малости требуется ввести нормы в пространстве данных задачи и в пространстве решений. Устойчивость в этом случае можно сформулировать так: малым по норме изменениям данных задачи должны соответствовать малые по норме изменения решения. Если это требование не выполняется, задача называется *некорректной*.



Часть III. Разностные методы

9. Методы построения разностных операторов

9.1. Введение

Когда задача сформулирована и поставлена, наступает черед ее решения. В редких случаях удастся найти решение в виде формулы. Значительно чаще уравнения модели не имеют аналитического решения, и его приходится искать с помощью различных численных процедур, что позволяет получить лишь приближенное решение задачи.

Дело в том, что любой вычислитель, будь то человек или компьютер, обладает ограниченной памятью и ограниченным быстродействием. Эти особенности вычислителя отсекают всякие бесконечности, приходящие к задаче и возможному методу ее решения. Так, множество хранящихся в памяти чисел конечно, сами числа (за малым исключением) представляются в памяти лишь приближенно (все бесконечные десятичные дроби заменяются на конечные) и т.п. Любые функции должны описываться конечным набором чисел (например, конечным числом коэффициентов Фурье). Чаще всего такое описание будет приближенным. Ограниченность быстродействия приводит к тому, что бесконечные процессы (суммирование бесконечных рядов, вычисление пределов, итерационные процедуры и т.п.) заменяются конечными. Каждая из указанных операций порождает ошибки и это отражается на точности результата.

В конечном счете краевая задача (дифференциальные уравнения и граничные условия) трансформируется в задачу линейной алгебры, т.е. заменяется системой линейных алгебраических уравнений, которая затем и решается. Существует две основные группы методов численного решения дифференциальных уравнений: проекционные методы и конечно-разностные. О первой группе пару слов скажем сейчас, а второй будем заниматься остальное время.

Пусть надо решить краевую задачу $Lf = g$, где L — линейный оператор. Произвольная функция, определенная на некотором отрезке, описывается несчетным набором своих значений. Для того чтобы функции f и g можно было описать счетным набором чисел (например, коэффициентов Фурье), будем считать их принадлежащими гильбертову пространству H . Иными словами, мы будем искать только такие

решения (другие решения, которые может быть тоже существуют, мы искать не планируем).

Итак, $L : H \rightarrow H$ и $f, g \in H$. Выбрав ортонормированный базис $\{e_i\}_{i=1}^{\infty}$, можем записать $f = \sum_i f^i e_i$ и $g = \sum_i g^i e_i$. Проекционные методы решения уравнения заключаются в том, что приближенное решение $\hat{f}$ ищется в виде отрезка ряда разложения f по базисным элементам (функциям):

$$f \approx \hat{f} = \sum_{i=1}^n \hat{f}^i e_i. \quad (9.1)$$

Исходная краевая задача сводится к системе линейных алгебраических уравнений относительно конечной совокупности коэффициентов Фурье $\{\hat{f}^j\}_{j=1}^n$. Приближенное решение краевой задачи $\hat{f}$ вычисляется затем по формуле (9.1). Требуемая точность решения достигается за счет выбора того или иного базиса, варьирования длины n отрезка ряда (9.1) и т.п.

Другой подход к построению приближенного решения заключается в том, что вместо конечного набора коэффициентов Фурье ищется конечный набор значений решения в некоторых заранее выбранных точках области определения Ω — узлах. Само приближенное решение, определенное во всей области Ω , находится по узловым значениям с помощью какого-либо способа восполнения (*интерполяции*). В этом случае область определения решения Ω заменяется дискретным набором узлов — так называемой *сеткой*, а функции, определенные на Ω , — сеточными функциями (векторами), размерность которых определяется числом узлов сетки. Здесь также дифференциальная задача сводится к линейноалгебраической, но система линейных уравнений записывается не для коэффициентов Фурье, а для значений решения в узлах сетки. Методы решения дифференциальных уравнений, основанные на этом подходе, называются *разностными*.

9.2. Дискретное представление функций

Пусть интересующие нас функции определены в области Ω . Выберем в ней множество точек Ω^n общим числом n , которое назовем *сеткой*, а сами точки $x_i \in \Omega^n$ назовем *узлами* сетки. Выбор сетки диктуется обычно либо предполагаемым видом решения (для хорошо-

го описания больших градиентов требуется больше узлов сетки), либо соображениями удобства (при равномерном расположении узлов расчетные формулы получаются проще), либо, наконец, конфигурацией области Ω . Будем для простоты считать, что рассматриваемая область одномерна $\Omega = [x_{\min}, x_{\max}]$, а узлы упорядочены так, что

$$x_{\min} \leq x_1 < x_2 < \dots < x_{n-1} < x_n \leq x_{\max}.$$

Расстояния между узлами $\Delta x_i = x_i - x_{i-1}$ называются *шагами* сетки. Если все они одинаковы ($\Delta x_i = \Delta x_j = \Delta x$, $\forall i, j \leq n$), сетка называется *равномерной*. Для равномерной сетки, очевидно, справедливо $x_i = x_1 + (i-1)\Delta x$, $i = 1, \dots, n$.

Функции, определенные на Ω^n , т.е. только в узлах сетки, называются *сеточными функциями*. В точках Ω , не принадлежащих Ω^n , сеточные функции не определены. Пусть $f(x)$ — некоторая непрерывная функция, заданная в области Ω , тогда функция $\mathbf{f} = (f_1, f_2, \dots, f_n)$, где $f_j = f(x_j)$ есть сеточная функция, определенная на множестве Ω^n .

Вектор $\mathbf{f}$ дает нам некоторую информацию о поведении непрерывной функции. Сколь полна эта информация? Иначе говоря, насколько точно мы можем описать поведение функции f в точках $x \in \Omega$ по ее значениям в узлах? К сожалению, приходится констатировать, что информации, заключенной в $\mathbf{f}$, недостаточно для описания поведения функции $f(x)$ в произвольных точках области Ω . Дело в том, что $\mathbf{f}$ есть отображение (проекция) элемента f бесконечномерного векторного пространства $V(\Omega)$ на конечномерное векторное пространство $\mathbb{R}^n$. Такое отображение не является взаимнооднозначным, и прообразом $\mathbf{f}$ оказывается не элемент, а подмножество $V(\Omega)$.

Единственным выходом из этой ситуации является использование неких дополнительных соображений о поведении функции f . Они должны настолько сузить нам множество отображаемых функций, что проектирование станет взаимнооднозначным и прообраз $\mathbf{f}$ превратится в точку из $V(\Omega)$. Как правило, эти дополнительные соображения сводятся к допущению того, что функция f может быть описана линейной комбинацией функций из некоторого конечного набора элементов $V(\Omega)$. Другими словами, вместо функции $f \in V(\Omega)$ рассматривается ее аппроксимация $\hat{f}$, принадлежащая конечномерному подмножеству $V(\Omega)$.

Пусть, например, $f \in L_2(\Omega)$, $L_2(\Omega)$ — гильбертово пространство функций, определенных на $\Omega = [x_{\min}, x_{\max}] \subset \mathbb{R}^1$, с нормой $\|f\|^2 = \int_{\Omega} f^2 dx$. Если $\{e_j\}_{j=1}^{\infty}$ — ортонормированный базис в $L_2(\Omega)$, то функция f может быть представлена своим рядом Фурье относительно этого базиса:

$$f = \sum_j f^j e_j.$$

Пусть далее на Ω задана сетка $\Omega^n = \{x_k\}_{k=1}^n$, в узлах которой определена сеточная функция $\mathbf{f} = \{f^k\}$ такая, что $f^k = f(x_k)$. В каждом узле значение $f_k \in \mathbb{R}^1$, а сама функция $\mathbf{f}$, следовательно, принадлежит пространству $\mathbb{R}^n$.

Зададим на $\mathbb{R}^n$ систему ортонормированных базисных векторов $\{e_k\}_{k=1}^n$. Тогда

$$\mathbf{f} = \sum_{k=1}^n f^k e_k. \quad (9.2)$$

Базис в $L_2(\Omega)$ выберем так, чтобы значения первых n базисных функций в узлах x_j совпадали с компонентами базисных векторов e_k . Это всегда можно сделать, например, так, как показано на рис.13.

Теперь, когда выбран базис $\{e_j\}$, мы можем приближенно записать функцию f в виде отрезка ряда Фурье:

$$f \approx \hat{f} = \sum_{j=1}^n \hat{f}^j e_j, \quad (9.3)$$

а компоненты разложения $\hat{f}^j$ определить из условия $f(x_k) = f_k$. Действительно, подставляя в (9.3) значения $x = x_k$, получаем систему линейных алгебраических уравнений относительно $\{\hat{f}^j\}_{j=1}^n$:

$$\mathbf{f} = E\mathbf{F}, \quad (9.4)$$

где $\mathbf{f} = (f_1, f_2, \dots, f_n)^T$, $\mathbf{F} = (\hat{f}^1, \hat{f}^2, \dots, \hat{f}^n)^T$. Элементами матрицы E являются значения базисных функций, вычисленные в узлах сетки $e_k^j = e_k(x_j)$, т.е. строками матрицы E являются базисные векторы: элементы k -й строки матрицы E совпадают с элементами k -го базисного вектора e_k . Базис $\{e_k\}$ — ортонормированный, так что и строки матрицы E попарно ортонормированы. Матрицы, строки которых обладают

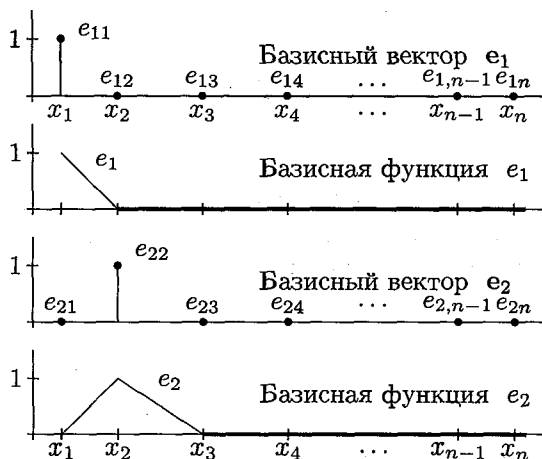


Рис. 13. Пример выбора базисных векторов $\{e_k\}$ и базисных функций $\{e_k\}$ таких, что $e_k(x_j) = e_k^j$.

таким свойством, называются *ортонормальными*. Рассмотрим произведение EE^T , где E^T — транспонированная матрица E . Произведение двух квадратных матриц есть квадратная матрица, (jk) -й элемент которой определяется как

$$\sum_{m=1}^n e_j(x_m) e_k(x_m) = (e_j, e_k) = \delta_{jk}$$

в силу ортонормированности базиса $\{e_k\}$. Таким образом,

$$EE^T = I, \quad \text{или} \quad E^T = E^{-1},$$

а это известное свойство ортогональных матриц: обратная матрица равна транспонированной. Поскольку E — ортогональная матрица, система уравнений (9.4) легко решается:

$$F = E^T f, \quad (9.5)$$

или

$$\hat{f}^j = \sum_{j=1}^n f^j e_{ij}. \quad (9.6)$$

Формулы (9.2), (9.4) и (9.5), (9.6) носят название *прямого и обратного дискретного преобразования Фурье*. Для комплексных базисных функций эти формулы имеют вид:

$$\hat{f} = \sum_{j=1}^n \hat{f}^j e_j, \quad \hat{f}^j = \sum_{k=1}^n f^k e_j^*(x_k),$$

где e_j^* — есть функция, комплексно сопряженная функции e_j . Таким образом, выбрав базисные векторы e_k и базисные функции e_k и построив по формулам (9.2) и (9.5) функцию $\hat{f}$, мы можем сравнить ее с исходной функцией f и оценить с помощью, скажем, нормы разности $\|f - \hat{f}\|$, насколько хороша наша аппроксимация. Эта оценка будет зависеть не только от числа и расположения узлов, т.е. множества Ω^n , но и от выбранных базисных векторов и функций.

В качестве примера рассмотрим аппроксимацию функции отрезком тригонометрического ряда Фурье. Ограничим рассмотрение частным случаем $n = 2N + 1$ равномерно расположенных узлов. В качестве базиса в $\mathbb{R}^n$ возьмем элементы вида:

$$\begin{aligned} e_1 &= \sqrt{\frac{1}{n}} (1, \dots, 1), \\ e_2 &= \sqrt{\frac{2}{n}} (\dots, \cos \frac{2\pi}{D} x_j, \dots), \\ e_3 &= \sqrt{\frac{2}{n}} (\dots, \sin \frac{2\pi}{D} x_j, \dots), \\ &\vdots \\ e_{2N} &= \sqrt{\frac{2}{n}} (\dots, \cos N \frac{2\pi}{D} x_j, \dots), \\ e_{2N+1} &= \sqrt{\frac{2}{n}} (\dots, \sin N \frac{2\pi}{D} x_j, \dots), \end{aligned}$$

где $D = x_{\max} - x_{\min}$, $\Delta x = \frac{D}{2N}$, $x_j = j\Delta x$, $j = 0, \dots, 2N$. Эта система элемен-

тов ортонормирована, что можно показать, вычислив суммы трех видов:

$$\begin{aligned} \sum_{j=0}^{2N} \sin k \frac{2\pi}{D} x_j \sin m \frac{2\pi}{D} x_j, \\ \sum_{j=0}^{2N} \sin k \frac{2\pi}{D} x_j \cos m \frac{2\pi}{D} x_j, \\ \sum_{j=0}^{2N} \cos k \frac{2\pi}{D} x_j \cos m \frac{2\pi}{D} x_j. \end{aligned}$$

В выбранном базисе сеточная функция f может быть представлена своим рядом Фурье

$$f = \sum_{k=1}^n \hat{f}^k e_k, \quad (9.7)$$

а j -я компонента сеточной функции — записана в виде суммы

$$f_j = \frac{1}{\sqrt{n}} \hat{f}^1 + \sqrt{\frac{2}{n}} \sum_{k=1}^N \left(\hat{f}^{2k} \cos k \frac{2\pi}{D} x_j + \hat{f}^{2k+1} \sin k \frac{2\pi}{D} x_j \right).$$

С помощью формул Эйлера

$$\cos x = \frac{1}{2} (e^{ix} + e^{-ix}), \quad \sin x = \frac{1}{2i} (e^{ix} - e^{-ix})$$

ряд (9.7) для j -й компоненты элемента f преобразуется к виду:

$$f_j = \sum_{k=-N}^N \hat{c}_k \exp(ik \frac{2\pi}{D} x_j),$$

где

$$\hat{c}_k = \begin{cases} \frac{1}{2} \sqrt{\frac{2}{n}} (\hat{f}^{2k} - i \hat{f}^{2k+1}), & k > 0, \\ \frac{1}{\sqrt{n}} \hat{f}^1, & k = 0, \\ \frac{1}{2} \sqrt{\frac{2}{n}} (\hat{f}^{-2k} + i \hat{f}^{-2k+1}), & k < 0, \end{cases}$$

$$\hat{c}_k = \hat{c}_{-k}^*.$$

Звездочкой помечено комплексно сопряженное число. В качестве базиса в $L_2(\Omega)$ возьмем систему функций

$$\frac{1}{\sqrt{D}}, \dots, \sqrt{\frac{2}{D}} \cos k \frac{2\pi}{D} x, \sqrt{\frac{2}{D}} \sin k \frac{2\pi}{D} x, \dots,$$

ортонормированных на Ω и принимающих в узлах x_j те же значения, что и компоненты элементов $\{e_k\}$. Тогда непрерывная на Ω функция f в этом базисе может быть представлена следующим рядом Фурье:

$$f = \frac{1}{\sqrt{D}} f^1 + \sqrt{\frac{2}{D}} \sum_{k=1}^{\infty} \left(f^{2k} \cos k \frac{2\pi}{D} x + f^{2k+1} \sin k \frac{2\pi}{D} x \right),$$

или в комплексной форме

$$f = \sum_{k=-\infty}^{\infty} c_k \exp(ik \frac{2\pi}{D} x).$$

Члены этого ряда $c_k \exp(ik \frac{2\pi}{D} x)$ называются *модами*, *гармониками*, или *волнами Фурье*; k — *волновым числом*, а коэффициент c_k — *амплитудой k -й моды*. Поскольку значения f в узлах сетки совпадают со значениями сеточной функции, мы можем вычислить амплитуды первых N мод. Эти моды соответствуют первым N волновым числам и имеют длины волн $\lambda_k = \frac{D}{k} \geq \frac{D}{N}$. Таким образом, информации, содержащейся в узлах сетки, нам достаточно лишь для построения длинноволновой аппроксимации функции f .

9.3. Методы построения разностных операторов

Приступая к решению задачи разностными методами, мы прежде всего попытаемся заменить непрерывный оператор L его разностным аналогом Λ_n . Этап выбора и построения разностной схемы является одним из наиболее ответственных. Неправильно или наугад выбранная схема может повлечь за собой грубые, а то и просто неверные результаты, вычислительную неустойчивость и массу других неприятностей. Добавьте сюда труд по написанию компьютерной программы, реализующей алгоритм, усилия и время, затраченное на отладку и поиск причины неудач, и т.д. и т.п. Наиболее важным свойством «хорошей» разностной схемы является ее способность аппроксимировать исходную дифференциальную задачу. В этом разделе мы рассмотрим некоторые употребительные методы построения аппроксимирующих разностных схем. Начнем с представления производных разностными операторами, а затем обсудим способы аппроксимации краевых задач.

Рассмотрим некоторую бесконечное число раз дифференцируемую функцию f , определенную в области Ω . Пусть для простоты эта область одномерна. С помощью координатного отображения поставим в

соответствие каждой точке P из Ω ее координату $x(P)$ из $\mathbb{R}^1$ и определим в $\mathbb{R}^1$ наиболее простую равномерную сетку с постоянным шагом Δx по оси x . Занумеруем узлы полученной сетки так, чтобы номер 0 приходился на начало координат. Тогда узел с координатами $i\Delta x \equiv x_i$ будет иметь номер i . Значение функции f в узле i будем обозначать через f_i . Совокупность значений $\{f_i\}_{i=0}^n$ есть сеточная функция $\mathbf{f}$, определенная на Ω^n . Наша задача заключается в построении разностных аналогов производных функции f в узлах сетки.

Займемся построением разностного аналога первой производной $d_x f$. Воспользуемся определением производной функции $f(x)$ в точке x . Функция f называется *дифференцируемой в точке x* , если существуют и равны друг другу пределы

$$d_x f = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{f(x) - f(x - \Delta x)}{\Delta x}, \quad (9.8)$$

или, что то же, существует предел

$$d_x f = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x - \Delta x)}{2\Delta x}. \quad (9.9)$$

Величина $d_x f$ называется *производной* функции f в точке x .

В случае сеточных функций приходится, во-первых, строить аналоги производных лишь в точках сетки x_i и, во-вторых, отказываться от вычисления пределов и заменять их приближенными выражениями. В соответствии с приведенными определениями, таких приближенных выражений может быть три:

$$\frac{f_{i+1} - f_i}{\Delta x}, \quad \frac{f_i - f_{i-1}}{\Delta x}, \quad \frac{f_{i+1} - f_{i-1}}{2\Delta x}.$$

Все они имеют одинаковую структуру, отношения разностей, почему и называются *разностными формулами*. Множество узлов, участвующих в определении разностного оператора, называется *шаблоном*. Различают двух-, трех- и более точечные шаблоны.

Первые две разностные формулы часто называют *направленными разностями* (или *разностями вперед/назад*³³). Они строятся на двух-точечных шаблонах, включающих точки x_i , x_{i+1} в первом случае и

³³Относительно рассматриваемой точки x_i .

точки x_{i-1} , x_i — во втором. Последнюю формулу называют *центральной (симметричной) разностью*. Она строится на трехточечном шаблоне, включающем точки x_{i-1} , x_i , x_{i+1} (первая и последняя точки входят в разностное выражение, которое аппроксимирует производную в средней точке).

Для того чтобы оценить погрешность таких приближенных выражений, воспользуемся разложением в ряд Тейлора³⁴ значений функции f в узлах $i \pm 1$ относительно узла i . Рассматривая, например, узел $i + 1$, найдем

$$f_{i+1} = f_i + d_x f_i \Delta x + \frac{1}{2} d_{xx} f_i \Delta x^2 + O(\Delta x^3). \quad (9.10)$$

Здесь и далее символами $d_x f_i$, $d_{xx} f_i$, ... обозначены производные, вычисленные в точке x_i . Символом $O(\alpha^n)$ традиционно обозначаются суммы, члены которых пропорциональны целым, положительным степеням α , начиная со степени n , или, как говорят, *порядка* α^n . В нашем случае — это слагаемые, пропорциональные Δx , начиная с Δx^3 . Разрешая формулу (9.10) относительно $d_x f_i$, получим

$$\begin{aligned} d_x f_i &= \frac{1}{\Delta x} (f_{i+1} - f_i) - \frac{1}{2} \Delta x d_{xx} f_i + O(\Delta x^2) = \\ &= \frac{1}{\Delta x} (f_{i+1} - f_i) + O(\Delta x). \end{aligned} \quad (9.11)$$

Откуда находим, что порядок отбрасываемых членов равен первой степени Δx . Аналогичную формулу получим, если рассмотрим узел $i - 1$

$$f_{i-1} = f_i - d_x f_i \Delta x + \frac{1}{2} d_{xx} f_i \Delta x^2 + O(\Delta x^3) \quad (9.12)$$

и далее

$$d_x f_i = \frac{1}{\Delta x} (f_i - f_{i-1}) + O(\Delta x). \quad (9.13)$$

Для того чтобы оценить погрешность последней разностной формулы, соответствующей определению (9.9), рассмотрим разность разложений (9.10) и (9.12), разрешенную относительно производной $d_x f_i$

$$d_x f_i = \frac{1}{2\Delta x} (f_{i+1} - f_{i-1}) + O(\Delta x^2). \quad (9.14)$$

³⁴Тейлор (Taylor) Брук (1685-1731) — английский математик.

Здесь порядок отбрасываемых членов оказался равным второй степени Δx , и погрешность последней формулы, таким образом, меньше, чем у предыдущих формул³⁵. Связано это с тем, что мы рассматриваем частный случай равномерной сетки, в силу чего в обоих разложениях (9.10) и (9.12) слагаемые, пропорциональные Δx^2 , равны и в разность $(f_{i+1} - f_{i-1})$ не входят.

Все три разностные формулы имеют простую геометрическую интерпретацию (рис.14): производная в узле (тангенс угла наклона касательной) аппроксимируется тангенсами углов наклона секущих.

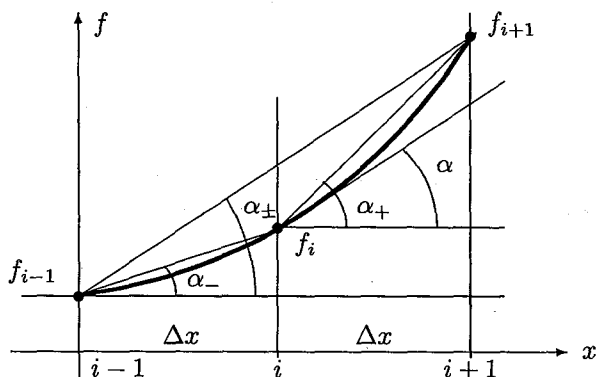


Рис. 14. Геометрическая интерпретация разностных формул, аппроксимирующих первую производную $d_x f_i$, равную тангенсу угла наклона касательной к f в точке x_i : $\operatorname{tg} \alpha = d_x f_i$, $\operatorname{tg} \alpha_- = \frac{1}{\Delta x}(f_i - f_{i-1})$, $\operatorname{tg} \alpha_+ = \frac{1}{\Delta x}(f_{i+1} - f_i)$, $\operatorname{tg} \alpha_{\pm} = \frac{1}{2\Delta x}(f_{i+1} - f_{i-1})$.

То, каким образом повысился порядок отбрасываемых членов в формуле центральных разностей, показывает, как можно строить более точные разностные формулы. Для этого требуется привлекать большее число узлов и комбинировать с нужными весами соответствующие им ряды Тейлора. Как правило, этого, однако, не делают ввиду возника-

³⁵Имеется в виду, что при стремлении $\Delta x \rightarrow 0$ члены второго порядка малости $O(\Delta x^2)$ уменьшаются быстрее, чем члены первого порядка малости $O(\Delta x)$.

ющих трудностей, связанных с подбором весов, а также с аппроксимацией производных на границах области.

Разностные формулы для второй производной строятся аналогично. Снова используем простейшую равномерную сетку, и рассмотрим ряд (9.10). Если мы разрешим относительно второй производной непосредственно это выражение, то получим:

$$d_{xx}f_i = \frac{2}{\Delta x} \left(\frac{1}{\Delta x} (f_{i+1} - f_i) - d_x f_i \right) + O(\Delta x).$$

Помимо разностного выражения, первое слагаемое в правой части содержит производную $d_x f_i$. Аналогичное выражение, полученное из ряда (9.12)

$$d_{xx}f_i = \frac{2}{\Delta x} \left(\frac{1}{\Delta x} (f_{i-1} - f_i) + d_x f_i \right) + O(\Delta x),$$

содержит ту же производную с противоположным знаком. Складывая оба выражения получим разностный аналог второй производной:

$$d_{xx}f_i = \frac{1}{\Delta x^2} (f_{i+1} - 2f_i + f_{i-1}) + O(\Delta x^2). \quad (9.15)$$

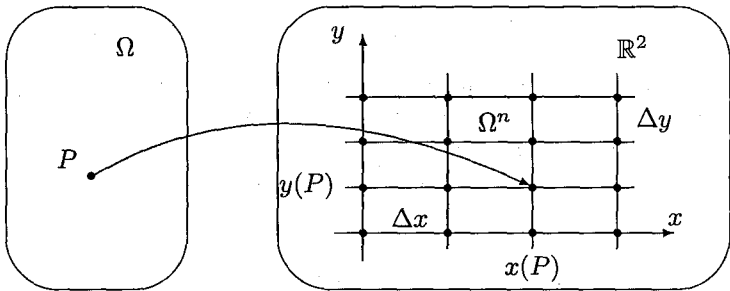


Рис. 15. Отображение области Ω на сетку Ω^n .

Обратите внимание, что здесь, как и в случае с центрально-разностной аппроксимацией первой производной, порядок отбрасываемых членов также увеличивается благодаря использованию равномерной сетки.

При сложении (9.10) и (9.12) члены, пропорциональные Δx^3 , сокращаются, отчего и повышается порядок отбрасываемых членов в (9.15).

Задав сетку, выбрав шаблон и определив разностные аналоги производных, можно построить разностный аналог дифференциального уравнения, комбинируя разностные операторы. Рассмотрим, например, двумерную область. С помощью координатных отображений $x(P)$, $y(P)$, $P \in \Omega$, поставим в соответствие каждой точке P из Ω ее координаты $x(P)$, $y(P)$ из $\mathbb{R}^2$ и определим в $\mathbb{R}^2$ сетку Ω^n . Рассмотрим простую равномерную сетку с постоянным шагом Δx по оси x и Δy по оси y (рис.15).

Занумеруем узлы полученной сетки так, чтобы номер $(0, 0)$ приходился на начало координат. Тогда узел с координатами $(i\Delta x, j\Delta y) \equiv (x_i, y_j)$ будет иметь номер (i, j) . Значение функции f в узле (i, j) будем обозначать через f_{ij} . Совокупность значений $\{f_{ij}\}$, $i = 0, \dots, I$, $j = 0, \dots, J$, $n = (I+1)(J+1)$, есть сеточная функция f , определенная на Ω^n .

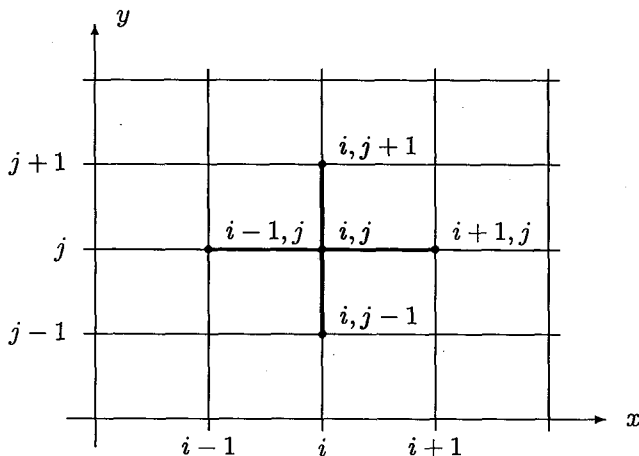


Рис. 16. Пятиточечный шаблон для уравнения Лапласа.

Используя введенную сетку и пятиточечный шаблон, показанный на рис.16, нетрудно построить разностный аналог двумерного уравнения

Лапласа:

$$\Delta f \equiv \partial_{xx}f + \partial_{yy}f = 0.$$

Аппроксимируя производные центральными разностями по формуле (9.15), получим

$$\begin{aligned} & \frac{1}{\Delta x^2}(f_{i+1,j} - 2f_{ij} + f_{i-1,j}) + O(\Delta x^2) + \\ & + \frac{1}{\Delta y^2}(f_{i,j+1} - 2f_{ij} + f_{i,j-1}) + O(\Delta y^2) = 0. \end{aligned} \quad (9.16)$$

Записывая разностное уравнение (9.16) для всех узлов сетки, мы сводим дифференциальную задачу к линейно алгебраической.

10. Методы построения разностных схем

10.1. Метод неопределенных коэффициентов

Разностный аналог производной, как мы видели, оказывается суммой значений функции в узлах шаблона, взятых с некоторыми весами. Заниматься подбором весов можно разве что в простейших случаях. Более общим в этом отношении является *метод неопределенных коэффициентов*, используя который мы пытаемся аппроксимировать не каждую производную в отдельности, а сразу весь дифференциальный оператор. Для этого, выбрав шаблон, мы записываем разностный оператор в виде суммы узловых значений функций с некоторыми, подлежащими определению, коэффициентами. Применяя далее разложение в ряды Тейлора для значений в соседних узлах относительно центрального (i, j) -го узла и группируя члены разложения, мы строим систему уравнений относительно неопределенных коэффициентов из условия аппроксимации с заданным порядком точности. Решая систему, находим неизвестные коэффициенты и тем самым разностный оператор.

Аппроксимируем, например, одномерное уравнение теплопроводности

$$\partial_t f = \kappa \partial_{xx} f, \quad \kappa = \text{const} > 0, \quad (x, t) \in \Omega$$

со вторым порядком точности по времени и пространству. В области Ω построим равномерную сетку Ω^n , выберем пятиточечный шаблон G (см. рис.16) и попытаемся найти веса узловых значений функции, такие, чтобы выполнялось равенство

$$(\partial_t f - \kappa \partial_{xx} f)_j^n = \sum_{k, l \in G} a_{kl} f_{j+l}^{n+k} + O(\Delta t^2, \Delta x^2), \quad (10.1)$$

где a_{kl} — подлежащие определению коэффициенты (веса). Записывая сумму разложений в ряд Тейлора значений функции в узлах шаблона

относительно узла (n, j) и группируя слагаемые, получим

$$\begin{aligned}
 \sum_{k,l \in G} a_{kl} f_{j+l}^{n+k} &= \sum_{k,l} a_{kl} \left(f_j^n + \partial_t f_j^n (t_{n+k} - t_n) + \partial_x f_j^n (x_{j+l} - x_j) + \right. \\
 &\quad + \frac{1}{2} (\partial_{tt} f_j^n (t_{n+k} - t_n)^2 + \partial_{xx} f_j^n (x_{j+l} - x_j)^2) + \\
 &\quad \left. + \partial_{tx} f_j^n (t_{n+k} - t_n)(x_{j+l} - x_j) + O(\Delta t^3, \Delta x^3) \right) = \\
 &= f_j^n \sum_{k,l} a_{kl} + \\
 &\quad + (\partial_t f_j^n) \sum_{k,l} a_{kl} (t_{n+k} - t_n) + \\
 &\quad + (\partial_x f_j^n) \sum_{k,l} a_{kl} (x_{j+l} - x_j) + \\
 &\quad + \frac{1}{2} (\partial_{tt} f_j^n) \sum_{k,l} a_{kl} (t_{n+k} - t_n)^2 + \\
 &\quad + \frac{1}{2} (\partial_{xx} f_j^n) \sum_{k,l} a_{kl} (x_{j+l} - x_j)^2 + \\
 &\quad + (\partial_{tx} f_j^n) \sum_{k,l} a_{kl} (t_{n+k} - t_n)(x_{j+l} - x_j) + \\
 &\quad + \sum_{k,l} a_{kl} O(\Delta t^3, \delta x^3). \tag{10.2}
 \end{aligned}$$

Для того чтобы сумма (10.2) удовлетворяла (10.1), веса a_{kl} должны быть решением следующей системы уравнений:

$$\left\{ \begin{array}{l} \sum a_{kl} = 0, \\ \sum a_{kl} (t_{n+k} - t_n) = 1, \\ \sum a_{kl} (x_{j+l} - x_j) = 0, \\ \sum a_{kl} (t_{n+k} - t_n)^2 = 0, \\ \sum a_{kl} (t_{n+k} - t_n)(x_{j+l} - x_j) = 0, \\ \sum a_{kl} (x_{j+l} - x_j)^2 = -2\kappa, \end{array} \right. \tag{10.3}$$

или

$$\left\{ \begin{array}{cccccc} a_{-1,0} & +a_{0,-1} & +a_{0,0} & +a_{0,1} & +a_{1,0} & = 0, \\ a_{-1,0} & & & & -a_{1,0} & = -\frac{1}{\Delta t}, \\ & a_{0,-1} & & -a_{0,1} & & = 0, \\ a_{-1,0} & & & & +a_{1,0} & = 0, \\ & a_{0,-1} & & +a_{0,1} & & = -\frac{2\kappa}{\Delta x^2}. \end{array} \right. \quad (10.4)$$

Обратите внимание, что пятое уравнение в (10.3) удовлетворяется тождественно, так как либо Δt , либо Δx равны нулю. Система (10.4) легко решается, и мы получаем:

$$-a_{-1,0} = a_{1,0} = \frac{1}{2\Delta t}, \quad a_{0,-1} = a_{0,1} = -\frac{\kappa}{\Delta x^2}, \quad a_{0,0} = \frac{2\kappa}{\Delta x^2}.$$

Остаточный член в сумме разложений (10.2) оказывается равным $O(\Delta t^2, \Delta x^2)$, а не $O(\Delta t^2, \Delta x)$, как может показаться на первый взгляд. В этом легко убедиться, если исследовать следующий член разложения. При найденных значениях весов a_{kl} члены разложения, содержащие Δx^3 , сокращаются. Таким образом, окончательно получаем:

$$(\partial_t f - \kappa \partial_{xx} f)_j^n = \frac{f_j^{n+1} - f_j^{n-1}}{2\Delta t} - \frac{f_{j+1}^n - 2f_j^n + f_{j-1}^n}{\Delta x^2} + O(\Delta t^2, \Delta x^2).$$

10.2. Интегральный метод построения разностных схем

Часто с исходными дифференциальными уравнениями связаны некоторые интегральные соотношения. Например, так называемые дифференциальные законы сохранения являются следствием интегральных законов сохранения. Если конечноразностный аналог дифференциального закона сохранения, будучи просуммированным по всем узлам сетки, не даст аппроксимацию интегрального закона сохранения, значит в процессе решения будет накапливаться ошибка, связанная с наличием фиктивных источников или стоков исследуемой характеристики. Не все разностные аналоги дифференциальных законов сохранения обеспечивают выполнение интегральных законов сохранения. В каждом конкретном случае выполнение интегральных законов приходится проверять отдельно. Существует, однако, метод построения раз-

ностных схем, обеспечивающий автоматическое выполнение интегральных законов. Он основан на том, что дифференциальное уравнение переписывается в виде соответствующего интегрального закона сохранения, который затем аппроксимируется. Как уже говорилось выше, дифференциальному уравнению

$$\partial_t f + \operatorname{div} \vec{F}(f) = 0 \quad (10.5)$$

соответствует интегральный закон сохранения

$$\int_V \partial_t f dV + \int_S (\vec{F}(f), \vec{n}) dS = 0,$$

где $\vec{F}(f)$ — поток величины f , V — произвольный объем, S — поверхность, этот объем ограничивающая, а $\vec{n}$ — внешняя нормаль к поверхности S . Если объем V не изменяется во времени, то

$$\partial_t \int_V f dV + \int_S (\vec{F}(f), \vec{n}) dS = 0. \quad (10.6)$$

Изменение во времени интегрального по объему V значения величины f балансируется интегральным по поверхности S объема V потоком $\vec{F}(f)$, а сам поток $\vec{F}(f)$ на S определяется граничными условиями.

Если разбить объем V на k непересекающихся объемов (будем называть их *боксами*) V_k : $V = \bigcup_k V_k$, ограниченных поверхностями S_k , то закон сохранения (10.6) для уравнения (10.5) можно записать в виде

$$\sum_k \left(\partial_t \int_{V_k} f dV + \int_{S_k} (\vec{F}(f), \vec{n}) dS \right) = 0.$$

Откуда в силу произвольности разбиения

$$\partial_t \int_{V_k} f dV + \int_{S_k} (\vec{F}(f), \vec{n}) dS = 0. \quad (10.7)$$

При суммировании боксов потоки $\vec{F}(f)$ на границах между ними взаимно компенсируют друг друга. Нескомпенсированными остаются

лишь потоки на S , и мы приходим, таким образом, к уравнению (10.6). Эта черта уравнения (10.7) называется свойством *дивергентности*. На представлении уравнения (10.5) в виде закона сохранения (10.7) и основан интегральный метод построения разностных схем.

В области Ω задается сетка Ω^n , и объем $V \in \Omega$ разбивается на боксы V_k так, чтобы внутри каждого бокса находился один и только один узел. Далее интегралы вычисляются по тем или иным приближенным формулам (если их невозможно вычислить точно), использующих только значения функций в узлах сетки. В итоге мы получаем разностный аналог закона сохранения, обладающий, как и (10.7), свойством дивергентности. Такие разностные схемы называются *дивергентными* или *консервативными*.

Надо иметь в виду, что, по крайней мере для гладких решений, консервативная и неконсервативная формы дифференциальных уравнений эквивалентны друг другу, чего нельзя сказать о соответствующих им разностных схемах. И если разностное уравнение, полученное при аппроксимации закона сохранения, является дивергентным, то из этого вовсе не следует, что и разностное уравнение, полученное из других соображений, будет также обладать этим свойством.

Разберем применение интегрального метода на двух примерах.

1) Рассмотрим одномерное стационарное уравнение теплопроводности:

$$\partial_x(\kappa \partial_x \vartheta) - q\vartheta + g = 0, \quad x \in (0, L), \quad (10.8)$$

$$\begin{aligned} (\kappa \partial_x \vartheta - a\vartheta)_{x=0} &= b, \\ \vartheta|_{x=L} &= c. \end{aligned} \quad (10.9)$$

Здесь x — пространственная координата, $\kappa(x)$, $q(x)$, $g(x)$ — заданные гладкие функции, $\kappa > 0$, $q \geq 0$, a , b , c — константы, $a \geq 0$. Краевая задача (10.8), (10.9) может служить моделью стационарного распределения температуры $\vartheta(x)$ в стержне длины L в том случае, когда на одном конце стержня происходит обмен теплом с окружающей средой по закону Ньютона, а на другом — поддерживается постоянная температура. Величина

$$\kappa \partial_x \vartheta = F(\vartheta) \quad (10.10)$$

есть плотность потока тепла в точке x . В нашем одномерном случае объем $\mathcal{V}$ — это отрезок $(0, L)$ и равен области Ω . Зададим в Ω равномерную сетку Ω^n с шагом Δx (рис. 17) и разобьем $\mathcal{V}$ на n боксов (каждый внутренний бокс также длиной Δx). Первый и последний узлы лежат на границах области Ω . Обозначим границы боксов дробными индексами. Например, границы i -го бокса обозначим через $x_{i-\frac{1}{2}}$ и $x_{i+\frac{1}{2}}$.



Рис. 17.

Запишем уравнение (10.8) в виде интегрального закона сохранения, справедливого в пределах i -го бокса, для чего проинтегрируем его по i -му боксу. Получим

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (\partial_x (\kappa \partial_x \vartheta) - q \vartheta + g) dx = 0. \quad (10.11)$$

Первое слагаемое в подынтегральном выражении интегрируется точно и дает разность потоков на границах бокса $F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}$. В данном одномерном случае точное интегрирование по i -му боксу есть результат применения теоремы Гаусса-Остроградского³⁶. Подынтегральное выражение, т.е. производная $\partial_x (\kappa \partial_x \vartheta) = \partial_x F$, представляет собой дивергенцию плотности потока F . В одномерном случае — это частная производная $\partial_x F$. Результат интегрирования, по существу, есть интеграл от нормальной составляющей плотности потока F по поверхности i -го бокса. Поскольку эта поверхность состоит из двух точек $x_{i-\frac{1}{2}}$ и $x_{i+\frac{1}{2}}$, а нормальные составляющие имеют разный знак (так как нормали в обеих точках внешние по отношению к боксу), мы получаем в итоге суммарный поток через границу $F_{i+\frac{1}{2}} + (-F_{i-\frac{1}{2}})$.

³⁶Гаусс (Gauss) Карл Фридрих (1777–1855) — немецкий математик.
Остроградский Михаил Васильевич (1801–1861) — русский математик.

Последние два слагаемых могут быть проинтегрированы приближенно, для чего используются простейшие квадратурные формулы прямоугольников или трапеций. Поскольку i -й бокс содержит лишь один узел с номером i , интеграл от второго слагаемого можно приближенно записать в виде произведения ϑ_i на интеграл от известной функции q . Интегралы от известных функций q и g аппроксимируются произведениями средних значений этих функций в пределах рассматриваемого бокса, $\bar{q}_i$ и $\bar{g}_i$, на размер этого бокса, т.е. на длину шага:

$$-\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \vartheta q dx + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} g dx \approx (-\vartheta_i \bar{q}_i + \bar{g}_i) \Delta x.$$

В результате, нормируя на Δx , получаем

$$\frac{1}{\Delta x} (F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}) - \vartheta_i \bar{q}_i + \bar{g}_i = 0. \quad (10.12)$$

Потоки $F_{i\pm\frac{1}{2}}$ через границы бокса можно приближенно вычислить, если проинтегрировать производную $\partial_x \vartheta$, полученную из (10.10), в пределах от x_{i-1} до x_i (или от x_i до x_{i+1})

$$\int_{x_{i-1}}^{x_i} \frac{F(\vartheta)}{\kappa} dx = \vartheta_i - \vartheta_{i-1},$$

и интеграл в правой части аппроксимировать величиной

$$\int_{x_{i-1}}^{x_i} \frac{F(\vartheta)}{\kappa} dx \approx F_{i-\frac{1}{2}} \int_{x_{i-1}}^{x_i} \frac{1}{\kappa} dx = F_{i-\frac{1}{2}} \frac{\Delta x}{\bar{\kappa}_{i-\frac{1}{2}}}.$$

Отсюда

$$F_{i-\frac{1}{2}} = \bar{\kappa}_{i-\frac{1}{2}} \frac{\vartheta_i - \vartheta_{i-1}}{\Delta x} \quad (10.13)$$

и аналогично

$$F_{i+\frac{1}{2}} = \bar{\kappa}_{i+\frac{1}{2}} \frac{\vartheta_{i+1} - \vartheta_i}{\Delta x}. \quad (10.14)$$

Подставляя (10.12), (10.13) и (10.14) в (10.11) и нормируя на Δx , получим

$$\frac{1}{\Delta x} \left(\bar{\kappa}_{i+\frac{1}{2}} \frac{\vartheta_{i+1} - \vartheta_i}{\Delta x} - \bar{\kappa}_{i-\frac{1}{2}} \frac{\vartheta_i - \vartheta_{i-1}}{\Delta x} \right) - \vartheta_i \bar{q}_i + \bar{g}_i = 0. \quad (10.15)$$

Это и есть разностный аналог уравнения (10.8) без учета граничных условий. Он содержит $n-1$ уравнение относительно $n+1$ неизвестных. Два недостающих уравнения дают граничные условия. Первое граничное условие можно записать в виде

$$F_0 = a\vartheta_0 + b. \quad (10.16)$$

Чтобы его учесть, модернизируем формулу (10.12), для чего проинтегрируем исходное уравнение (10.8) по половине бокса от 0 до $x_{\frac{1}{2}}$. Получим

$$F|_0^{\frac{1}{2}} - \int_0^{\frac{1}{2}} (q\vartheta - g)dx = (F_{\frac{1}{2}} - F_0) - (\vartheta_0 \bar{q}_0 - \bar{g}_0) \frac{\Delta x}{2} = 0. \quad (10.17)$$

Здесь интеграл заменен на его приближенное значение

$$\int_0^{\frac{1}{2}} (q\vartheta - g)dx = (\vartheta_0 \bar{q}_0 - \bar{g}_0) \frac{\Delta x}{2}.$$

Поток $F_{\frac{1}{2}}$ определим далее из (10.13),

$$F_{\frac{1}{2}} = \frac{\vartheta_1 - \vartheta_0}{\int_{x_0}^{x_1} \frac{dx}{\kappa}} = \bar{\kappa}_{\frac{1}{2}} \frac{\vartheta_1 - \vartheta_0}{\Delta x}, \quad (10.18)$$

а поток F_0 получим из граничного условия (10.16). Подстановка найденных потоков в формулу (10.17) дает одно из недостающих уравнений

$$\bar{\kappa}_{\frac{1}{2}} \frac{\vartheta_1 - \vartheta_0}{\Delta x} - \vartheta_0 \left(\frac{\Delta x}{2} \bar{q}_0 + a \right) = b - \frac{\Delta x}{2} \bar{g}_0. \quad (10.19)$$

Второе граничное условие в этой задаче используется для нахождения решения в n -м узле.

$$\vartheta_n = c. \quad (10.20)$$

Таким образом, мы аппроксимировали краевую задачу (10.8), (10.15) разностной схемой (10.15), (10.19), (10.20).

Легко показать, что сумма (10.15) по всем внутренним узлам вместе с выражениями (10.19) и (10.20) приводит к аппроксимации интегрального закона сохранения

$$\int_0^L (\partial_x(\kappa \partial_x \vartheta) - q\vartheta + g) dx = 0.$$

Действительно, суммирование по внутренним узлам выражения (10.15), умноженного на Δx дает

$$\bar{\kappa}_{n-\frac{1}{2}} \frac{\vartheta_n - \vartheta_{n-1}}{\Delta x} - \bar{\kappa}_{\frac{1}{2}} \frac{\vartheta_1 - \vartheta_0}{\Delta x} - \sum_{i=1}^{n-1} (\vartheta_i \bar{q}_i - \bar{g}_i) \Delta x = 0.$$

Потоки через внутренние границы боксов взаимно скомпенсировались и не вошли в результат. Учитывая теперь выражения (10.19) и (10.20), аппроксимирующие граничные условия, окончательно получим

$$\left(\bar{\kappa}_{n-\frac{1}{2}} \frac{c - \vartheta_{n-1}}{\Delta x} - (\vartheta_0 a + b) \right) - \left(\sum_{i=1}^{n-1} (\vartheta_i \bar{q}_i - \bar{g}_i) \Delta x - (\vartheta_0 \bar{q}_0 - \bar{g}_0) \frac{\Delta x}{2} \right) = 0.$$

Здесь первая скобка аппроксимирует интеграл $\int_0^L \partial_x(\kappa \partial_x \vartheta) dx$, а вторая — интеграл $\int_0^L (q\vartheta - g) dx$.

- 2) Консервативность — важное свойство разностной схемы. Однако его наличие еще не гарантирует работоспособности алгоритма. В качестве иллюстрации построим с помощью интегрального метода разностную схему для невязкого уравнения Бюргерса

$$\partial_t u + u \partial_x u = 0, \quad \Omega = \{x \in (0, L); t \geq 0\}. \quad (10.21)$$

Уравнение (10.21) можно записать в дивергентной форме (в форме закона сохранения):

$$\partial_t u + \partial_x F = 0, \quad (10.22)$$

введя поток $F = \frac{1}{2}u^2$. Определим в области Ω равномерную сетку Ω^n с шагом по времени Δt и шагом по пространству Δx . Разобьем отрезок $[0, L]$ на боксы (см. рис.18) и проинтегрируем уравнение (10.22) в пределах i -го бокса

$$\partial_t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u dx + F|_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} = 0.$$

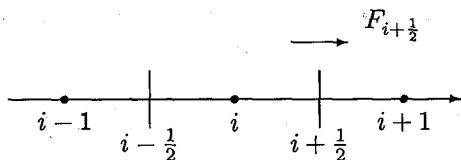


Рис. 18.

Вычислив интеграл по формуле центральных прямоугольников, получим

$$\Delta x \partial_t u_i + F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = 0.$$

Для приближенного вычисления потоков $F_{i\pm\frac{1}{2}}$ на границах боксов наиболее естественным на первый взгляд кажется использовать формулы

$$F_{i\pm\frac{1}{2}} = \frac{1}{2}(F_i + F_{i\pm 1}).$$

В этом случае получается консервативная центрально-разностная схема

$$\partial_t u_i + \frac{1}{2\Delta x}(F_{i+1} - F_{i-1}) = 0,$$

аппроксимирующая исходное уравнение со вторым порядком точности. Однако дальнейшее изучение этой разностной схемы (см. стр.129) показывает, что на ее основе нельзя построить устойчивого двуслойного алгоритма интегрирования уравнения Бюргерса.

10.3. Нестационарные задачи

Процессы, меняющиеся во времени, моделируются задачами Коши или смешанными задачами для нестационарных (гиперболических или параболических) уравнений. Решение строится для каждого момента времени (или, как говорят, *временного слоя*) с использованием граничных условий и данных с предыдущих временных слоев. Область интегрирования по времени часто формально бесконечна ($t > 0$), и ограничивается либо машинными ресурсами, либо достижением квазистационарного или квазипериодического режима. Сетка во временной области выбирается равномерной с шагом $\Delta t = t_{n+1} - t_n$, а производные по времени, как правило, аппроксимируются простейшими двух-, трехточечными разностными формулами. Получающиеся разностные схемы называются соответственно *двух-* и *трехслойными*. Для аппроксимации эволюционного уравнения можно также использовать интегральный метод. Рассмотрим, например, следующее эволюционное уравнение:

$$\partial_t f + Lf = g, \quad f|_{t=0} = f_0, \quad (10.23)$$

где L — дифференциальный оператор, содержащий лишь пространственные производные. В области $t \geq 0$ введем равномерную сетку с шагом Δt . Будем обозначать $f^n \equiv f(t_n)$. Чтобы получить двуслойную разностную схему, проинтегрируем уравнение (10.23) по t от t_n до t_{n+1} :

$$\int_{t_n}^{t_{n+1}} (\partial_t f + Lf) dt = \int_{t_n}^{t_{n+1}} g dt,$$

откуда

$$f^{n+1} - f^n + L \int_{t_n}^{t_{n+1}} f dt = \int_{t_n}^{t_{n+1}} g dt. \quad (10.24)$$

В зависимости от способа вычисления интегралов различают *явные*, *неявные схемы* и *схему Кранка–Николсона*³⁷. Если интегралы вычисляются по формуле левых прямоугольников, то мы получаем *явную*

³⁷Кранк (Crank) Джон (1916–2006) и Николсон (Nicolson) Филлис (1917–1968) — английские математики.

разностную схему, или схему Эйлера:

$$\frac{f^{n+1} - f^n}{\Delta t} + Lf^n = g^n. \quad (10.25)$$

Значения решения в каждой пространственной точке $(n+1)$ -го временного слоя находятся по уже известным значениям с предыдущего слоя. При использовании формулы правых прямоугольников получается *явная разностная схема*

$$\frac{f^{n+1} - f^n}{\Delta t} + Lf^{n+1} = g^n. \quad (10.26)$$

Здесь значение f^{n+1} зависит от значений решения во всех пространственных точках $(n+1)$ -го же временного слоя. И, наконец, если мы применим формулу трапеций, то получим *схему Кранка-Николсон*

$$\frac{f^{n+1} - f^n}{\Delta t} + L \frac{f^{n+1} + f^n}{2} = g^n, \quad (10.27)$$

которую можно трактовать как сумму формул (10.25) и (10.26). Схемы (10.25)–(10.27) используют для аппроксимации производных по времени значения функций лишь на двух временных слоях, и потому называются *двуслойными*. Первые две схемы имеют первый порядок точности по t , а третья – второй.

Разрешая формулы (10.25)–(10.27) относительно f^{n+1} , получаем следующее рекуррентное соотношение:

$$f^{n+1} = Gf^n + \Delta t Sg^n, \quad (10.28)$$

где G — оператор перехода (шага), а S — оператор источника. Начальные данные дают нам значение $f^0 = f_0$. Используя далее (10.28), мы можем рассчитать решение в любой момент времени. Выражение (10.28) называется *каноническим видом* разностной схемы.

Аналогично конструируются и трехслойные разностные схемы, где для вычисления решения в момент времени t_{n+1} используются значения функций на двух предыдущих слоях t_n и t_{n-1} . Рассмотрим снова уравнение (10.23), проинтегрируем его в пределах (t_{n-1}, t_{n+1}) и вычислим интегралы по формуле центральных прямоугольников. Тогда

получим трехслойную разностную схему второго порядка точности, известную под названием *четарда*:

$$\frac{f^{n+1} - f^{n-1}}{2\Delta t} + Lf^n = g^n. \quad (10.29)$$

Комбинируя схему (10.29) со схемой Кранка–Николсон, записанной для (n) -го временного слоя, получим трехслойную *схему Адамса–Башфорта*³⁸

$$\frac{f^{n+1} - f^n}{\Delta t} + L \left(\frac{3}{2}f^n - \frac{1}{2}f^{n-1} \right) = 2g^n - g^{n-1}, \quad (10.30)$$

имеющую также второй порядок точности по времени. Существуют и другие варианты.

³⁸Адамс (Adams) Джон Коуч (1819–1892) и Башфорт (Bashforth) Фрэнсис (1819–1912) — английские математики.

11. Основы теории разностных схем

11.1. Структура разностного оператора

Рассмотрим уравнение

$$Lf = g, \quad (11.1)$$

где $f, g \in H(\Omega)$, L — линейный оператор, $L: H(\Omega) \rightarrow H(\Omega)$. Приближенным решением $\hat{f}$ задачи (11.1) будем называть проекцию функции f на некоторое, выбранное нами, n -мерное подпространство $H_f^n(\Omega) \subset H(\Omega)$. Оператор проектирования (проектор), отображающий $H(\Omega)$ в $H_f^n(\Omega)$, обозначим через Π_f . Поскольку размерность H_f^n меньше размерности H , оператор Π_f не является взаимно однозначным, и обратное отображение Π_f^{-1} не существует. Областью определения $D(L)$ и областью значений $R(L)$ оператора L мы считаем все пространство $H(\Omega)$. Назовем оператор $\hat{L}$ *сужением* оператора L на область $D(\hat{L})$, если $D(\hat{L}) \subset D(L)$ и $Lf = \hat{L}f$ для всех f из $D(\hat{L})$. Иначе говоря, в области $D(\hat{L})$ эти операторы равны, а вне этой области оператор $\hat{L}$ не определен.

Рассмотрим сужение $\hat{L}$ оператора L на подпространство H_f^n , тогда $D(\hat{L}) = H_f^n(\Omega)$. Область значений $R(\hat{L})$ оператора $\hat{L}$ обозначим через $H_g^n(\Omega)$ (рис.19). В силу линейности L множество H_g^n является подпространством H . Оно не обязательно совпадает с $H_f^n(\Omega)$. Их пересечение, вообще говоря, может состоять из одного нуля. Проектор, отображающий H в H_g^n , обозначим через Π_g . Спроектируем уравнение (11.1) на подпространство H_g^n с помощью оператора Π_g и обозначим $\Pi_g g \equiv \hat{g} \in H_g^n$:

$$\Pi_g Lf = \hat{g}. \quad (11.2)$$

Посмотрим, что представляет собой композиция $\Pi_g \circ L$. Для этого выберем в пространстве $H(\Omega)$ ортонормированный базис $\{e_i\}_{i=1}^{\infty}$, такой, чтобы совокупность элементов $\{e_i\}_{i=1}^n$ образовывала базис подпространства H_f^n .

Представим функцию f в виде ряда Фурье по выбранному базису и подставим в (11.2):

$$\Pi_g Lf = \Pi_g L \sum_{i=1}^{\infty} f^i e_i = \Pi_g \sum_{i=1}^{\infty} f^i (Le_i). \quad (11.3)$$

Мы считаем, что оператор L^{-1} существует и, кроме того, $D(L) = R(L) = H$. Известно, что в этом случае последовательность векторов $\{Le_i\}_{i=1}^{\infty}$ образует базис в H , если e_i — базисные векторы. Поскольку первые n элементов e_i лежат в H_f^n , для любой функции φ из H_f^n справедливо равенство $L\varphi = \hat{L}\varphi$, где $\hat{L}: H_f^n \rightarrow H_g^n$ — сужение оператора L на H_f^n . Таким образом, получаем, что $Le_i = \hat{L}e_i$, $i = 1, \dots, n$, и $\{\hat{L}e_i\}_{i=1}^n$ образует базис в H_g^n . Проектор Π_g по определению отображает H в H_g , и раз $\{Le_i\}_{i=1}^n$ является базисом H_g , то (11.3) можно записать далее так:

$$\Pi_g Lf = \sum_{i=1}^n f^i (\hat{L}e_i) = \hat{L} \sum_{i=1}^n f^i e_i. \quad (11.4)$$

Последняя сумма в (11.4) есть проекция f на H_f^n , так что окончательно получаем

$$\Pi_g Lf = \hat{L}\Pi_f f = \hat{L}\hat{f},$$

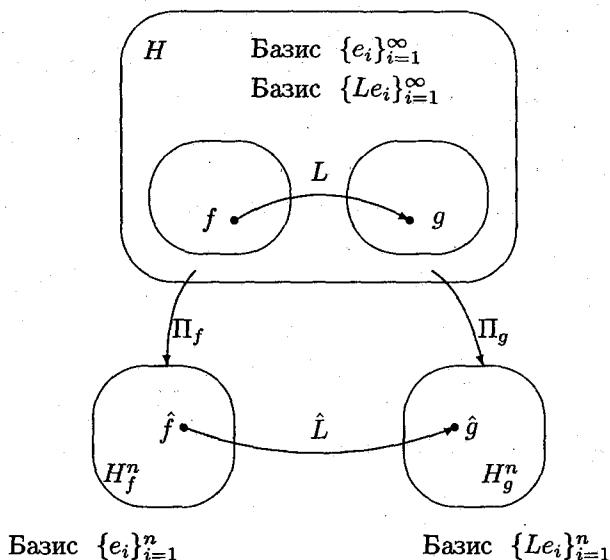
и уравнение (11.2) записывается в виде

$$\hat{L}\hat{f} = \hat{g}. \quad (11.5)$$

Переход от исходного уравнения (11.1) к его приближенному аналогу (11.5) существенно упрощает задачу, так как бесконечномерный случай мы сводим к конечномерному (увы, ценой невосполнимых утрат, ибо обратный переход невозможен).

Последнее преобразование, которому должно подвергнуться исходное уравнение (11.1), заключается в том, что путем выбора специального базиса в H_f^n , такого, что коэффициентами Фурье для функций служат значения $\hat{f}_i$ в узлах сетки, мы сводим задачу (11.5) к системе алгебраических уравнений относительно вектора $f = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_n)$. Обозначим такой базис в пространстве H_f^n через $\{\varepsilon_i\}_{i=1}^n$. В пространстве H_g^n в качестве базиса возьмем функции $\{\hat{L}\varepsilon_i\}_{i=1}^n$, которые обозначим через $\{\xi_i\}_{i=1}^n$. Формулы

$$\begin{aligned} \hat{f} &= \sum_{i=1}^n \hat{f}^i \varepsilon_i = \sum_{i=1}^n \hat{f}_i \varepsilon_i, & \hat{f}_i &= \hat{f}(x_i), \\ \hat{g} &= \sum_{i=1}^n \hat{g}^i \xi_i = \sum_{i=1}^n \hat{g}_i \xi_i, & \hat{g}_i &= \hat{g}(x_i), & x_i &\in \Omega^n \end{aligned}$$

Рис. 19. Связь операторов L и $\hat{L}$.

задают линейные взаимно однозначные отображения $S_n: H_g^n \rightarrow \mathbb{R}^n$ и $Q_n: H_f^n \rightarrow \mathbb{R}^n$. Функции $\hat{f}$ и $\hat{g}$ соответственно отображаются в n -мерные векторы $\mathbf{f}$ и $\mathbf{g}$ (сеточные функции):

$$Q_n \hat{f} = \mathbf{f}, \quad S_n \hat{g} = \mathbf{g}. \quad (11.6)$$

Подставляя выражения (11.6) в уравнение (11.5), получаем требуемую систему линейных алгебраических уравнений

$$S_n \hat{L} Q_n^{-1} \mathbf{f} = \mathbf{g},$$

или

$$\Lambda_n \mathbf{f} = \mathbf{g}, \quad \Lambda_n \equiv S_n \hat{L} Q_n^{-1}. \quad (11.7)$$

Система (11.7) называется *разностной схемой*, или *разностной задачей*, соответствующей исходной задаче (11.1), а оператор $\Lambda_n: \mathbb{R}^n \rightarrow \mathbb{R}^n$

называется *разностным оператором*, или *разностным аналогом* оператора $\hat{L}$. Отображение $Q_n^{-1}: \mathbb{R}^n \rightarrow H_f^n$, позволяющее восстанавливать приближенное решение $\hat{f}$ по дискретному набору его значений, называется *интерполяцией*.

Как видим, разностный оператор Λ_n далеко не однозначно определяется видом оператора L . И способы проектирования, задаваемые проекторами Π_f , Π_g , и операторы S_n , Q_n выбираются нами в соответствии с некоторыми дополнительными соображениями. Это и просто-та оператора Λ_n (чем он проще, тем легче вычислить обратный оператор), и желание наделить разностный аналог многими свойствами, присущими исходному оператору L , что порождает большое разнообразие разностных аналогов для, в общем-то, скромного числа исходных уравнений.

11.2. Сходимость, аппроксимация, устойчивость

Исходную краевую задачу

$$Lf = g \quad (11.8)$$

мы заменяем ее аналогом — разностной задачей

$$\Lambda_n f = g, \quad (11.9)$$

решив которую, получим вектор f . Сколь хорошо найденное решение? Очевидно, что непосредственно сравнивать f и $\hat{f}$ нельзя. Они являются элементами различных векторных пространств (бесконечномерного и конечномерного), так что, оценить норму их разности невозможно.

Для получения требуемых оценок надо иметь два элемента одного и того же векторного пространства. Так, можно с помощью какой-либо интерполяционной процедуры построить из конечномерного вектора f функцию непрерывного аргумента — приближенное решение $\hat{f}$. Тогда, близость приближения $\hat{f}$ к точному решению f может быть оценена величиной

$$\|s\| = \|f - \hat{f}\|,$$

где s — функция, называемая *погрешностью решения*.

Можно, напротив, спроектировать точное решение f на сетку и получить конечномерный вектор $P_m f$. Далее, его можно сравнить с решением разностной задачи $\mathbf{f}$, и их близость оценить величиной

$$\|\mathbf{r}\|_m = \|P_m f - \mathbf{f}\|_m.$$

Здесь $\mathbf{r}$ — сеточная функция, которую называют *погрешностью разностной схемы*³⁹.

Если при измельчении сетки (т.е. уменьшении размеров боксов и увеличении числа узлов) величина $\|s\| \rightarrow 0$, то говорят, что имеет место *сходимость приближенных решений*. Если при тех же условиях $\|\mathbf{r}\|_m \rightarrow 0$, то говорят, что имеет место *сходимость решений разностной задачи*.

Для того чтобы записать все это формально, введем понятие характерного размера бокса. Будем характеризовать бокс Ω_i его *диаметром*, который определим как

$$\text{diam} \Omega_i = \max_{j,k} \|x_j - x_k\|, \quad \forall x_j, x_k \in \Omega_i.$$

Тогда величина $h = \|\text{diam} \Omega_1, \dots, \text{diam} \Omega_n\|_n$ есть характерный размер бокса данной сетки. Очевидно, что при $h \rightarrow 0$ число узлов $n \rightarrow \infty$, но обратное неверно.

Таким образом, сходимость приближенных решений имеет место при $\lim_{h \rightarrow 0} \|s\| = 0$, а решения разностной задачи сходятся, если $\lim_{h \rightarrow 0} \|\mathbf{r}\|_m = 0$. Последнее означает, что при измельчении сетки решение разностной задачи должно все лучше и лучше описывать проекцию точного решения на сетку Ω^m .

Здесь надо иметь в виду, что поскольку члены последовательности $\{\|\mathbf{r}\|_m\}_{m=1}^\infty$ есть нормы в различных пространствах $\mathbb{R}^m$ (так называемые *сеточные нормы*), то они должны быть согласованы между собой и с нормой пространства $H(\Omega)$, к которому принадлежит решение f . Это означает, что для любой функции $u \in H(\Omega)$ должно выполняться равенство

$$\lim_{h \rightarrow 0} \|P_m u\|_m = \|u\|.$$

³⁹Мы предполагаем, что, и пространство $H(\Omega)$, и пространство $\mathbb{R}^m$ являются нормированными с нормами $\|\cdot\|$ и $\|\cdot\|_m$ соответственно.

Например, пусть $\mathbf{u} = (u_1, u_2, \dots, u_n)$, $u_i = u(x_i)$. Тогда сеточная норма

$$\|\mathbf{u}\|_m = \left(\sum_{i=1}^m h |u_i|^2 \right)^{\frac{1}{2}}$$

согласована с нормой

$$\|u\| = \left(\int_{\Omega} |u(x)|^2 dx \right)^{\frac{1}{2}},$$

а сеточная норма

$$\|\mathbf{u}\|_m = \max_{1 \leq i \leq m} |u_i|$$

согласована с нормой

$$\|u\| = \max_{x \in \Omega} |u|.$$

Напротив, сеточная норма

$$\|\mathbf{u}\|_m = \left(\sum_{i=1}^m |u_i|^2 \right)^{\frac{1}{2}}$$

не является согласованной с нормами для функций непрерывного аргумента, так как для произвольных $\mathbf{u}$ ряд $\sum_{i=1}^{\infty} |u_i|^2$ может расходиться.

Непосредственно оценить нормы погрешностей нельзя, поскольку обе они содержат неизвестные функции. Одна — решение задачи f , другая — его проекцию на сетку $P_m f$. В этом случае приходится прибегать к косвенным оценкам, и проще такие оценки получить для вектора $\mathbf{r}$.

Если существуют положительные, не зависящие от h , константы c_1 и l , такие, что

$$\|\mathbf{r}\|_n \leq c_1 h^l,$$

то говорят, что разностная схема имеет l -й порядок точности.

В каком случае мы можем рассчитывать на сходимость решений разностной задачи? По определению погрешность разностной схемы есть разность $\mathbf{r} = P_m f - \mathbf{f}$. Вектор $\mathbf{f}$ обращает разностную задачу в

тождество. Если же подставить в разностное уравнение проекцию точного решения $P_n f$, получим

$$\Lambda_n P_n f = \Lambda_n f + \Lambda_n r = g + \Lambda_n r.$$

Сеточная функция

$$d \equiv \Lambda_n r = \Lambda_n(P_n f) - \Lambda_n f = \Lambda_n P_n f - g \quad (11.10)$$

называется *погрешностью аппроксимации разностной задачи на решении исходной задачи*. Будем говорить, что уравнение (11.9) *аппроксимирует* уравнение (11.8), а оператор Λ_n *аппроксимирует оператор* L на элементе $f \in D(L)$, если

$$\lim_{h \rightarrow 0} \|d\|_n = 0. \quad (11.11)$$

При этом, если существуют положительные, не зависящие от h , константы c_2 и k , такие, что

$$\|d\|_n \leq c_2 h^k,$$

то говорят, что разностная схема имеет k -й *порядок аппроксимации*.

Поскольку из (11.10) следует, что норма погрешности разностной схемы

$$\|r\|_n \leq \|\Lambda_n^{-1}\|_n \cdot \|d\|_n,$$

то, очевидно, наличие аппроксимирующего свойства у оператора Λ_n необходимо для сходимости. Этого, однако, недостаточно, так как оператор Λ_n^{-1} может не быть ограниченным. Поэтому вторым важным свойством оператора Λ_n является ограниченность его обратного оператора. Это свойство непосредственно связано с корректностью постановки разностного аналога исходной краевой задачи.

Разностная задача называется *корректной*, если для любых начальных данных из некоторого заданного множества ее решение: 1) существует, 2) единственно и 3) устойчиво. Первые два условия по сути означают существование обратного оператора Λ_n^{-1} . Что касается устойчивости, то под этим понимается непрерывная и равномерная относительно h зависимость решений разностной задачи от правой части. Таким образом, если решение устойчиво, то найдется такая положительная, не зависящая от h , константа c_3 , что

$$\|f\|_n \leq c_3 \|g\|_n. \quad (11.12)$$

Поскольку $f = \Lambda_n^{-1}g$, то из (11.12) следует, что

$$\|\Lambda_n^{-1}\|_n \leq c_3, \quad \forall n, h. \quad (11.13)$$

Иначе говоря, последовательность операторов $\{\Lambda_n^{-1}\}_n$ должна быть равномерно ограничена по h . Обратите внимание, что здесь речь идет о последовательности разностных операторов $\{\Lambda_n^{-1}\}_n$, соответствующих разным сеткам, образующимся при $n \rightarrow \infty$ ($h \rightarrow 0$). Устойчивость разностной схемы (11.9), как она определена выше, вообще говоря, не связана напрямую с исходной краевой задачей (11.8), а является свойством самой разностной схемы. Для линейного случая, который мы рассматриваем, все сказанное выше подытоживается в теореме, связывающей устойчивость разностной схемы со сходимостью ее решений.

Теорема. Если разностная задача (11.9) корректна и аппроксимирует корректную краевую задачу (11.8), тогда решение разностной задачи сходится к решению исходной краевой задачи, а порядок точности совпадает с порядком аппроксимации.

Доказательство. Погрешность разностной схемы r , согласно выражению (11.10), удовлетворяет уравнению

$$\Lambda_n r = d,$$

которое с точностью до правой части совпадает с уравнением (11.9). В силу корректности разностной задачи найдется положительная константа c_3 , такая, что

$$\|\Lambda_n^{-1}\|_n \leq c_3,$$

откуда

$$\|r\|_n \leq c_3 \|d\|_n. \quad (11.14)$$

Поскольку разностная схема аппроксимирует исходную краевую задачу, то при измельчении сетки $\|d\|_n \rightarrow 0$, а значит, и $\|r\|_n \rightarrow 0$, т.е. имеет место сходимость решений разностной задачи. Если разностная схема имеет k -й порядок аппроксимации

$$\|d\|_n \leq c_2 h^k,$$

то из (11.14) получаем

$$\|r\|_n \leq c_2 c_3 h^k = c_1 h^k.$$

Таким образом, разностная схема имеет k -й порядок точности, совпадающий с порядком аппроксимации, что и требовалось доказать.

Часто эту теорему и ей подобные называют *теоремами об эквивалентности*, имея в виду, что при указанных условиях аппроксимация и устойчивость (корректность) разностной схемы эквивалентны сходимости ее решений к решению исходной задачи.

Наличие сходимости дает нам уверенность в том, что при необходимости любая требуемая точность решения может быть достигнута проведением расчетов на сетке с соответствующим числом узлов.

12. Устойчивость разностных схем

12.1. Введение

Выше мы определили устойчивость разностной схемы

$$\Lambda_m f = g \quad (12.1)$$

как равномерную относительно размера сеточного бокса h ограниченность последовательности операторов $\{\Lambda_m^{-1}\}_m$:

$$\|\Lambda_m^{-1}\| \leq c, \quad c = \text{const} > 0, \quad \forall m, h,$$

где Λ_m — разностный оператор, аппроксимирующий как собственно дифференциальный оператор L , так и граничные условия. Отсюда

$$\|f\| \leq c\|g\|. \quad (12.2)$$

Если схема (12.1) является разностным аналогом эволюционного уравнения, то решение строится как последовательность векторов $\{f^n\}$, начиная с некоторого f^0 , определяемого начальными условиями. В качестве нормы $\|f\|$ удобно брать

$$\|f\| = \max_n \|f^n\|,$$

где n — номер временного шага. Тогда условие устойчивости (12.2) равносильно выполнению условия

$$\|f^n\| \leq c\|g\|, \quad \forall n. \quad (12.3)$$

Рассмотрим эволюционное уравнение

$$\partial_t f + Lf = g,$$

и пусть правая часть не зависит от времени. Тогда двуслойная явная разностная схема, записанная в каноническом виде (см. формулу (10.28) на стр.114), такова:

$$f^{n+1} = G_m f^n + \Delta t g. \quad (12.4)$$

Здесь $G_m = I - \Delta t \Lambda_m$. Записывая аналогичное выражение для n -го временного слоя и подставляя в (12.4), получим

$$f^{n+1} = G_m(G_m f^{n-1} + \Delta t g) + \Delta t g = G_m^2 f^{n-1} + (G_m + I)\Delta t g.$$

Продолжая подстановки вплоть до f^0 , будем иметь

$$f^{n+1} = G_m^{n+1} f^0 + \left(\sum_{i=0}^n G_m^i \right) \Delta t g.$$

Для нормы $\|f^{n+1}\|$ справедливо неравенство

$$\|f^{n+1}\| \leq \|G_m^{n+1}\| \cdot \|f^0\| + \Delta t \left(\sum_{i=0}^n \|G_m^i\| \right) \|g\|.$$

С подобной ситуацией мы уже сталкивались при изучении метода последовательных приближений. Рассуждая аналогично найдем, что равномерная ограниченность норм степеней оператора G_m будет иметь место, если

$$1) \quad \|G_m\| < 1,$$

или

$$2) \quad r = \lim_{n \rightarrow \infty} \sqrt[n]{\|G_m^n\|} < 1.$$

Здесь r — спектральный радиус оператора G_m . Полагая $n\Delta t \leq T$, где верхняя граница временного интервала T фиксирована, строгие неравенства могут быть заменены нестрогими. При этих условиях, очевидно, разностная схема (12.4) будет устойчива.

Практические методы исследования устойчивости разностных схем линейных дифференциальных уравнений основаны на этих двух критериях. Первый критерий реализуется в так называемом *прямом методе*, а второй — в *спектральном методе*, или *методе Неймана*.

Представим оператор перехода в виде

$$G_m = (G_m)_0 + (G_m)_b,$$

где $(G_m)_0$ соответствует краевой задаче с нулевыми граничными условиями, а оператор $(G_m)_b$ определяется аппроксимацией ненулевых

граничных условий. Тогда, допуская, что $\|(G_m)_b\|$ мала, определение устойчивости относительно оператора перехода G_m часто заменяют определением устойчивости относительно оператора $(G_m)_o$, или *устойчивости по начальным данным*.

12.2. Примеры использования прямого метода

Использование прямого метода проиллюстрируем двумя примерами.

- 1) Исследуем устойчивость явной разностной схемы для одномерного уравнения теплопроводности

$$\frac{f_i^{n+1} - f_i^n}{\Delta t} = \kappa \frac{f_{i-1}^n - 2f_i^n + f_{i+1}^n}{\Delta x^2}, \quad \kappa = \text{const} > 0. \quad (12.5)$$

Если (12.5) привести к каноническому виду

$$f_i^{n+1} = df_{i-1}^n + (1 - 2d)f_i^n + df_{i+1}^n,$$

где $d = \frac{\kappa \Delta t}{\Delta x^2}$, оператор $(G_m)_o$ окажется квадратной трехдиагональной матрицей. Будем рассматривать устойчивость только по начальным данным без учета оператора $(G_m)_b$, тогда

$$(G_m)_o = \begin{pmatrix} (1-2d) & d & & & 0 \\ & \ddots & \ddots & \ddots & \\ & & d & (1-2d) & d \\ & & & \ddots & \ddots & \ddots \\ 0 & & & & d & (1-2d) \end{pmatrix}_{m \times m}.$$

Норму $\|(G_m)_o\|$ определим как $\max_j \sum_{i=1}^m |G_{ij}|$, что дает $\|(G_m)_o\| = 2d + |1 - 2d|$. Требование $\|(G_m)_o\| \leq 1$ приводит к неравенству

$$2d + |1 - 2d| \leq 1, \quad (12.6)$$

которое удовлетворяется при $d \leq \frac{1}{2}$. В этом случае разностная схема (12.5) оказывается устойчивой. Однако, ее устойчивость зависит от величины d , или от соотношения шагов сетки Δt и Δx . Такие схемы называются *условно устойчивыми* в отличие от *абсолютно устойчивых* схем, у которых $\|(G_m)_o\| \leq 1$ независимо от соотношения шагов сетки.

Часто прямой метод применяют, минуя этап построения оператора перехода G_m . Определим норму $\|f^n\| = \max_i |f_i^n|$, тогда

$$\begin{aligned}\|f^{n+1}\| &= \max_i |df_{i-1}^n + (1-2d)f_i^n + df_{i+1}^n| \leq \\ &\leq d \max_i |f_{i-1}^n| + |1-2d| \max_i |f_i^n| + d \max_i |f_{i+1}^n| = \\ &= (2d + |1-2d|) \|f^n\|.\end{aligned}$$

Если $\|f^{n+1}\| \leq \|f^n\|$, то решение не возрастает с течением времени, и разностная схема, таким образом, оказывается устойчивой. Это условие снова приводит к неравенству (12.6) и дает то же условие устойчивости.

- 2) В качестве второго примера рассмотрим явную разностную схему для одномерного уравнения переноса

$$\frac{f_i^{n+1} - f_i^n}{\Delta t} + u \frac{f_i^n - f_{i-1}^n}{\Delta x} = 0, \quad (12.7)$$

или

$$f_i^{n+1} = (1 - \mu) f_i^n + \mu f_{i-1}^n, \quad \mu = \frac{u \Delta t}{\Delta x}.$$

Снова определим норму $\|f^n\| = \max_i |f_i^n|$, тогда

$$\begin{aligned}\|f^{n+1}\| &= \max_i |(1 - \mu) f_i^n + \mu f_{i-1}^n| \leq \\ &\leq |1 - \mu| \max_i |f_i^n| + |\mu| \max_i |f_{i-1}^n| = \\ &= (|1 - \mu| + |\mu|) \|f^n\|.\end{aligned}$$

Схема (12.7) условно устойчива, если $0 \leq \mu \leq 1$. В этом случае $\|f^{n+1}\| \leq \|f^n\|$, и $\|f^{n+1}\|$ не возрастает с ростом n (т.е. во времени).

Обратите внимание, что в отличие от схемы (12.5) для уравнения теплопроводности, где устойчивость определялась только соотношением Δt и Δx , устойчивость схемы (12.7) зависит еще и от знака величины u (направления переноса). Если $u < 0$, схема (12.7) перестает быть устойчивой при любых Δt и Δx (становится *абсолютно неустойчивой*) и применять ее нельзя. Такое свойство двухточечных разностных аппроксимаций $\frac{1}{\Delta x}(f_i^n - f_{i-1}^n)$ и $\frac{1}{\Delta x}(f_{i+1}^n - f_i^n)$ первой производной породило соответствующую терминологию: устойчивая разностная аппроксимация носит название «разности против потока», а неустойчивая называется «разностью по потоку».

В том случае, когда поток может менять направление с течением времени приходится либо отказываться от аппроксимации в помощью направленных разностей, либо отслеживать направление потока и менять в соответствии с ним разностную аппроксимацию. В качестве примера рассмотрим невязкое уравнение Бюргерса (10.21), записанное в дивергентной форме

$$\partial_t u + \partial_x F = 0,$$

где поток $F = \frac{1}{2}u^2$. Интегрирование этого уравнения в пределах i -го бокса дает

$$\Delta x \partial_t u_i + F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = 0.$$

Для получения устойчивой аппроксимации направленными разностями нужно аппроксимировать значения потока на границах бокса значениями в близлежащих узлах, сообразуясь со знаком потока на границе. Знак потока можно приближенно оценить по знаку суммы потоков в соседних узлах

$$\text{sign}(F_{i\pm\frac{1}{2}}) = \text{sign}(F_i + F_{i\pm 1}).$$

Таким образом, следующая разностная аппроксимация первого порядка точности

$$F_{i+\frac{1}{2}} = \begin{cases} F_i, & (F_i + F_{i+1}) > 0, \\ F_{i+1}, & (F_i + F_{i+1}) < 0, \end{cases}$$

$$F_{i-\frac{1}{2}} = \begin{cases} F_{i-1}, & (F_{i-1} + F_i) > 0, \\ F_i, & (F_{i-1} + F_i) < 0 \end{cases}$$

оказывается счетно устойчивой и вполне согласуется с физикой процесса переноса.

12.3. Спектральный метод (метод Неймана)

Согласно второму критерию, нормы степеней оператора перехода G_m^n будут равномерно ограничены относительно n , если спектральный радиус оператора G_m не превышает единицы, или, что то же, спектр оператора G_m лежит в круге единичного радиуса на комплексной плоскости. Поскольку конечномерный оператор (матрица) перехода G_m имеет дискретный спектр, состоящий из m комплексных собственных чисел, то его спектральный радиус определяется максимальным по модулю собственным числом λ_m , а значит, разностная схема (12.4) будет

устойчива, если $|\lambda_m| \leq 1$. Это условие называется *необходимым спектральным условием устойчивости Неймана*⁴⁰. Для определения λ_m надо решить следующую *спектральную задачу (задачу на собственные значения)*:

$$G_m \mathbf{f} = \lambda \mathbf{f}, \quad \mathbf{f} = (f_1, \dots, f_m). \quad (12.8)$$

Набор собственных чисел (спектр) полностью определяется оператором перехода

$$G_m = I - \Delta t \Lambda_m.$$

Матрица Λ_m , а следовательно, и матрица G_m , в одномерном случае имеет ленточную структуру. Все строки ленты одинаковы, за исключением начальных и конечных строк, которые зависят от вида граничных условий и выбранной аппроксимации. Как и в случае прямого метода, будем изучать только устойчивость по начальным данным, т.е. ограничимся рассмотрением оператора $(G_m)_0$.

В одномерном случае равномерной по x сетки на отрезке D с числом узлов $(2M + 1)$ следующие функции

$$\left\{ \exp \left(i \frac{2\pi}{D} k \Delta x j \right) \right\}_{k=-M}^M = \{ e^{i\alpha j} \}, \quad \alpha = \frac{2\pi}{D} k \Delta x, \quad i = \sqrt{-1}$$

являются собственными функциями (векторами) оператора $(G_m)_0$. Каждому $\alpha(k)$ соответствует свое собственное число $\lambda(\alpha)$.

Если начальные данные, заданные в узлах сетки, с помощью дискретного преобразования Фурье представить в виде ряда по собственным функциям оператора $(G_m)_0$

$$\mathbf{f}_0 = (f_{01}, \dots, f_{0m}), \quad f_{0j} = \sum_{\alpha} f_0^{\alpha} e^{i\alpha j}, \quad j = 1, \dots, m,$$

то спектральную задачу

$$(G_m)_0 \mathbf{f} = \lambda \mathbf{f}$$

можно заменить задачей определения собственного числа λ как функции от α :

$$(G_m)_0 \begin{pmatrix} \dots \\ e^{i\alpha j} \\ \dots \end{pmatrix} = \lambda \begin{pmatrix} \dots \\ e^{i\alpha j} \\ \dots \end{pmatrix}.$$

⁴⁰ Джон (Янош) фон Нейман (Neumann) (1903–1957) — американский математик и физик. Родился в Будапеште, с 1930 в США.

Разным α (т.е. разным гармоникам) соответствуют разные собственные числа, и разностная схема будет устойчивой в том случае, если модуль $|\lambda(\alpha)| \leq 1$ при любых α .

12.4. Примеры использования спектрального метода

Рассмотрим применение метода Неймана на нескольких примерах.

- 1) Явная разностная схема для уравнения теплопроводности $\partial_t f = \kappa \partial_{xx} f$. В качестве начальных данных возьмем гармонику $f_0 = \{e^{i\alpha j}\}_j$, где j — номер узла сетки. Подставляя f_0 в разностную схему, после первого шага получаем изменение амплитуды в λ раз:

$$(G_m)_0 f_0 = \lambda f_0,$$

а после n -го шага — в λ^n раз. Заменяя в разностном уравнении

$$\frac{f_j^{n+1} - f_j^n}{\Delta t} = \kappa \frac{f_{j+1}^n - 2f_j^n + f_{j-1}^n}{\Delta x^2} \quad (12.9)$$

f_j^n на $\lambda^n e^{i\alpha j}$, получим:

$$\lambda^{n+1} e^{i\alpha j} - \lambda^n e^{i\alpha j} = d \left(\lambda^n e^{i\alpha(j+1)} - 2\lambda^n e^{i\alpha j} + \lambda^n e^{i\alpha(j-1)} \right),$$

или

$$\lambda = 1 + d(e^{-i\alpha} - 2 + e^{i\alpha}) = 1 + 2d(\cos \alpha - 1) = 1 - 4d \sin^2 \frac{\alpha}{2}.$$

Величина λ характеризует изменение амплитуды k -й гармоники за один шаг по времени и часто называется *множителем перехода*. Для устойчивости необходимо, чтобы

$$|\lambda(\alpha)| \leq 1, \quad \forall \alpha.$$

Таким образом, условие устойчивости получаем из неравенства

$$\left| 1 - 4d \sin^2 \frac{\alpha}{2} \right| \leq 1,$$

которое сводится к системе неравенств

$$d \sin^2 \frac{\alpha}{2} \geq 0, \quad d \sin^2 \frac{\alpha}{2} \leq \frac{1}{2}.$$

Если $d > 0$, первое неравенство выполняется для любых α , а второе — лишь при $d \leq \frac{1}{2}$. Таким образом, необходимым условием устойчивости схемы (12.9) является $d \in [0, \frac{1}{2}]$.

В случае неявной схемы 12.10

$$\frac{f_j^{n+1} - f_j^n}{\Delta t} = \kappa \frac{f_{j+1}^{n+1} - 2f_j^{n+1} + f_{j-1}^{n+1}}{\Delta x^2} \quad (12.10)$$

имеем

$$\lambda(1 - 2d(\cos \alpha - 1)) = 1,$$

или

$$\lambda = \frac{1}{1 + 2d(1 - \cos \alpha)},$$

откуда следует, что $\lambda \leq 1$ при любых α при $d > 0$. В обоих случаях условие устойчивости выражается в терминах $d = \frac{\kappa \Delta t}{\Delta x^2}$. Поскольку величины $\kappa, \Delta t, \Delta x$ существенно положительные, то схема (12.10) оказывается абсолютно устойчивой. В случае явной схемы (12.9) условие устойчивости $d \leq \frac{1}{2}$, по сути, оказывается условием, накладываемым на соотношения шагов сетки по времени и пространству. Как правило, пространственная сетка фиксирована, так что для шага по времени получаем ограничение

$$\Delta t \leq \frac{\Delta x^2}{2\kappa}.$$

- 2) Явная разностная схема с направленными разностями для уравнения переноса

$$\frac{f_j^{n+1} - f_j^n}{\Delta t} + u \frac{f_j^n - f_{j-1}^n}{\Delta x} = 0$$

при подстановке $\lambda^n e^{i\alpha j}$ вместо f_j^n дает

$$\lambda = 1 - \mu(1 - e^{-i\alpha}) = (1 - \mu) + \mu e^{-i\alpha}. \quad (12.11)$$

Заметив, что $|\lambda| \leq 1 \Leftrightarrow |\lambda|^2 = \lambda \lambda^* \leq 1$, умножим (12.11) на сопряженное комплексное число. Получаем

$$|\lambda|^2 = (1 - \mu + \mu e^{-i\alpha})(1 - \mu + \mu e^{i\alpha}) = 1 - 4\mu(1 - \mu) \sin^2 \frac{\alpha}{2},$$

откуда из условия $|\lambda|^2 \leq 1$ имеем

$$\mu(\mu - 1) \leq 0,$$

что при $\mu > 0$ соответствует условию

$$\mu \leq 1.$$

Аналогично для неявной схемы

$$\frac{f_j^{n+1} - f_j^n}{\Delta t} + u \frac{f_j^{n+1} - f_{j-1}^{n+1}}{\Delta x} = 0$$

получаем

$$|\lambda|^2 = \frac{1}{1 + 2\mu(1 + \mu)(1 - \cos \alpha)}.$$

При $\mu > 0$ неравенство $|\lambda| \leq 1$ оказывается справедливым при любых α . Неравенство $\mu < 1$, полученное выше для явной схемы, можно также рассматривать как ограничение шага по времени

$$\Delta t < \frac{\Delta x}{u}.$$

- 3) В качестве примера исследования устойчивости многослойных разностных схем рассмотрим центрально-разностную схему для волнового уравнения

$$\partial_{tt} f = u^2 \partial_{xx} f.$$

Явная разностная схема записывается следующим образом:

$$\frac{f_j^{n+1} - 2f_j^n + f_j^{n-1}}{\Delta t^2} = u^2 \frac{f_{j+1}^n - 2f_j^n + f_{j-1}^n}{\Delta x^2}.$$

После подстановки $f_j^n = \lambda^n e^{i\alpha j}$ получаем

$$\lambda^{n-1}(\lambda^2 - 2\lambda + 1) = \mu^2 \lambda^n (e^{i\alpha} - 2 + e^{-i\alpha}), \quad \mu = \frac{u\Delta t}{\Delta x},$$

или, сокращая на λ^{n-1} и используя формулу Эйлера для косинуса,

$$\lambda^2 - 2\left(1 - 2\mu^2 \sin^2 \frac{\alpha}{2}\right)\lambda + 1 = 0.$$

Введем обозначение $a \equiv 2\mu^2 \sin^2 \frac{\alpha}{2}$ и запишем полученное уравнение в виде

$$\lambda^2 - 2(1 - a)\lambda + 1 = 0.$$

Это квадратное уравнение в зависимости от знака дискриминанта $(1 - a)^2 - 1$ имеет либо два действительных корня, либо два комплексно сопряженных.

- а) Если $(1-a)^2 - 1 > 0$, уравнение имеет действительные корни

$$\lambda_{1,2} = 1 - a \pm \sqrt{(1-a)^2 - 1}.$$

В этом случае $a > 2$, откуда $|\mu| > 1$, и, поскольку ось симметрии параболы $y = \lambda^2 - 2(1-a)\lambda + 1$ проходит через точку $1-a$, один из корней по модулю всегда превосходит единицу. Таким образом, при $|\mu| > 1$ разностная схема неустойчива.

- б) Теперь, пусть $(1-a)^2 - 1 \leq 0$, и уравнение имеет два комплексно сопряженных корня

$$\lambda_{1,2} = 1 - a \pm i\sqrt{1 - (1-a)^2}.$$

В этом случае $a \leq 2$, $|\mu| \leq 1$, корни по модулю равны единице, а разностная схема оказывается устойчивой (при условии $|\mu| \leq 1$).

- 4) Применим метод Неймана к исследованию устойчивости двумерной разностной схемы. Рассмотрим двумерное уравнение переноса

$$\partial_t f + u \partial_x f + v \partial_y f = 0, \quad u, v > 0.$$

Для аппроксимации производных используем направленные против потока разности

$$\frac{f_{jk}^{n+1} - f_{jk}^n}{\Delta t} + u \frac{f_{jk}^n - f_{j-1,k}^n}{\Delta x} + v \frac{f_{jk}^n - f_{j,k-1}^n}{\Delta y} = 0,$$

где j соответствует координате x , а k — координате y . Подставляя решение в виде

$$f_{jk}^n = \lambda^n e^{i(\alpha j + \beta k)},$$

получаем уравнение для определения множителя перехода λ :

$$\lambda^{n+1} - \lambda^n + \frac{u \Delta t}{\Delta x} (1 - e^{-i\alpha}) \lambda^n + \frac{v \Delta t}{\Delta y} (1 - e^{-i\beta}) \lambda^n = 0,$$

или, сокращая на λ^n и вводя обозначения $\mu_u = \frac{u \Delta t}{\Delta x}$, $\mu_v = \frac{v \Delta t}{\Delta y}$,

$$\lambda = (1 - \mu_u - \mu_v) + (\mu_u e^{-i\alpha} + \mu_v e^{-i\beta}).$$

Квадрат модуля $|\lambda|^2$ найдем, умножая λ на сопряженное число

$$|\lambda|^2 = (1 - \mu_u - \mu_v + \mu_u e^{-i\alpha} + \mu_v e^{-i\beta})(1 - \mu_u - \mu_v + \mu_u e^{i\alpha} + \mu_v e^{i\beta})$$

Производя умножение и используя формулы Эйлера, получим:

$$(1 - \mu_u - \mu_v)^2 + \mu_u^2 + \mu_v^2 + 2\mu_u\mu_v \cos(\alpha - \beta) + \\ + 2(1 - \mu_u - \mu_v)(\mu_u \cos \alpha + \mu_v \cos \beta)$$

Раскрывая далее скобки и приводя подобные члены, с учетом тождества $1 - \cos \alpha = 2 \sin^2 \frac{\alpha}{2}$, находим:

$$1 - 4\mu_u(1 - \mu_u) \sin^2 \frac{\alpha}{2} - 4\mu_v(1 - \mu_v) \sin^2 \frac{\beta}{2} + \\ + 4\mu_u\mu_v \left(\sin^2 \frac{\alpha}{2} + \sin^2 \frac{\beta}{2} - \sin^2 \frac{\alpha - \beta}{2} \right),$$

или

$$|\lambda|^2 = 1 - 4\mu_u(1 - (\mu_u + \mu_v)) \sin^2 \frac{\alpha}{2} - \\ - 4\mu_v(1 - (\mu_u + \mu_v)) \sin^2 \frac{\beta}{2} - 4\mu_u\mu_v \sin^2 \frac{\alpha - \beta}{2}.$$

Требуя $|\lambda|^2 \leq 1$, определим условие устойчивости двумерной разностной схемы. Имеем

$$(1 - (\mu_u + \mu_v)) \left(\mu_u \sin^2 \frac{\alpha}{2} + \mu_v \sin^2 \frac{\beta}{2} \right) + \mu_u\mu_v \sin^2 \frac{\alpha - \beta}{2} \geq 0.$$

Это неравенство выполняется при любых α и β , если $\mu_u + \mu_v \leq 1$. В противном случае всегда найдутся такие α и β , при которых неравенство не имеет места. Действительно, при $\mu_u + \mu_v > 1$ первое слагаемое меньше нуля и для $\alpha = \beta$ неравенство не выполняется. Окончательно получаем условие устойчивости в виде

$$\mu_u + \mu_v \leq 1.$$

Оно оказалось более жестким, нежели в одномерном случае. Для шага по времени имеем ограничение

$$\Delta t \leq \frac{1}{\frac{u}{\Delta x} + \frac{v}{\Delta y}}.$$

- 5) Рассмотрим, наконец, одномерное уравнение переноса при наличии диффузии

$$\partial_t f + u \partial_x f = \kappa \partial_{xx} f$$

и исследуем устойчивость следующей разностной схемы ВВЦП (разность Вперед по Времени и Центральная разность по Пространству):

$$f_j^{n+1} - f_j^n + \mu(f_j^n - f_{j-1}^n) = d(f_{j+1}^n - 2f_j^n + f_{j-1}^n).$$

Для множителя перехода получаем выражение

$$\lambda = (1 - \mu) - 4d \sin^2 \frac{\alpha}{2} + \mu e^{-i\alpha}.$$

Квадрат модуля $|\lambda|^2$ равен

$$|\lambda|^2 = 1 - 4(2d + \mu(1 - \mu)) \sin^2 \frac{\alpha}{2} + d(d + \mu) (4 \sin^2 \frac{\alpha}{2})^2.$$

Из условия $|\lambda|^2 \leq 1$ получаем неравенство

$$d(d + \mu) \left(4 \sin^2 \frac{\alpha}{2}\right) - (2d + \mu(1 - \mu)) \leq 0, \quad (12.12)$$

которое должно выполняться для любых α . Если имеет место

$$4d(d + \mu) - (2d + \mu(1 - \mu)) = (\mu + 2d)(\mu + 2d - 1) \leq 0,$$

т.е. случай $\sin^2 \frac{\alpha}{2} = 1$, то при $\mu \geq 0$ и $d \geq 0$ неравенство (12.12) выполняется и для всех прочих α . Таким образом, неравенство $\mu + 2d \leq 1$ определяет условие устойчивости рассматриваемой разностной схемы. Из него можно также получить ограничение на шаг по времени для заданной пространственной сетки и конкретных значений κ и u . Подставляя выражения для d и μ , находим

$$\Delta t \leq \frac{1}{\frac{u}{\Delta x} + 2 \frac{\kappa}{\Delta x^2}}.$$

При $\kappa = 0$ либо $u = 0$ получаем два частных предельных случая:

$$\begin{aligned} \Delta t &\leq \frac{\Delta x}{u} && \text{— для уравнения переноса,} \\ \Delta t &\leq \frac{\Delta x^2}{2\kappa} && \text{— для уравнения теплопроводности.} \end{aligned}$$

12.5. Физический смысл условий устойчивости

При аппроксимации уравнения теплопроводности разностным мы получили условие устойчивости

$$d \equiv \frac{\kappa \Delta t}{\Delta x^2} \leq \frac{1}{2}, \quad (12.13)$$

откуда следует, что максимальный шаг по времени для данной пространственной сетки не должен превышать

$$\Delta t_m = \frac{1}{2} \frac{\Delta x^2}{\kappa}.$$

Эта величина есть характерное время, за которое приращение решения Δf диффундирует на характерное расстояние Δx . Скорость распространения возмущений от узла к узлу в рассмотренной явной разностной схеме равна отношению $\frac{\Delta x}{\Delta t}$ и является конечной величиной, в отличие от бесконечной скорости, определяемой уравнением теплопроводности. В случае, когда $\Delta t > \Delta t_m$, скорость становится слишком малой и разностная схема перестает воспроизводить процесс теплопередачи, который оказывается слишком быстрым. В свою очередь неявная разностная схема, так же как и исходное уравнение, обладает бесконечной скоростью распространения возмущений и потому не имеет ограничения на шаг по времени, связанного с устойчивостью.

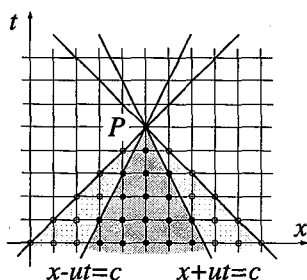


Рис. 20. Области зависимости решения волнового уравнения (выделена более темным цветом) и устойчивого разностного уравнения.

Параметр μ часто называется *числом Куранта*, а условие (12.14) — *условием Куранта-Фридрихса-Леви*⁴¹ (КФЛ). Аналогичное условие устойчивости получается и для явной центрально-разностной схемы второго порядка точности, аппроксимирующей волновое уравнение (15.23).

Физический смысл условия КФЛ достаточно очевиден. Вспомним, что волновое уравнение

$$\partial_{tt}f - u^2 \partial_{xx}f = 0 \quad (12.15)$$

обладает семейством характеристик, описываемых уравнениями

$$x \pm ut = \text{const.}$$

⁴¹Курант (Courant) Рихард (1888–1972), Фридрихс (Friedrichs) Курт (1901–1982) и Леви (Lewy) Ганс (1904–1988) — американские математики. Родились и учились в Германии, перед II мировой войной эмигрировали в США.

Обратите внимание, что условие (12.13) ограничивает, вообще говоря, не шаг по времени, а отношение $\frac{\Delta t}{\Delta x^2}$. Но поскольку на практике шаг пространственной сетки обычно ограничивается наличными машинными ресурсами, неравенство (12.13) часто рассматривают как ограничение шага по времени для явной разностной схемы.

Перейдем теперь к рассмотрению уравнения переноса. Явная конечно-разностная аппроксимация устойчива, если

$$\mu \equiv \frac{u \Delta t}{\Delta x} \leq 1. \quad (12.14)$$

Решение уравнения (12.15) в некоторой точке P зависит лишь от значений f в точках области зависимости, ограниченной двумя характеристиками, проходящими через рассматриваемую точку P (рис. 20). Тангенсы углов наклона характеристик волнового уравнения к оси x равны соответственно $\pm \frac{1}{u}$. Из условия КФЛ (12.14) вытекает, что для устойчивости явной разностной схемы необходимо выполнение неравенства

$$\frac{\Delta t}{\Delta x} \leq \frac{1}{|u|}.$$

Прямые $x \pm \frac{\Delta x}{\Delta t} t = \text{const}$, проходящие через точку P (на рис. 20 эта точка выбрана так, что является узлом сетки), ограничивают область зависимости разностного решения. Условие КФЛ означает, что для устойчивости решения область зависимости разностного уравнения должна охватывать область зависимости волнового уравнения.

13. Ошибки разностных схем

Замена дифференциального уравнения разностным приводит к тому, что получаемое решение не совпадает с решением исходного уравнения. Отличия разностного решения от точного принято объяснять *ошибками разностных схем*. Изучение и классификация этих ошибок основывается на представлении решения в виде тригонометрического ряда Фурье и определении изменения амплитуды и фазы каждой гармоники за один шаг по времени для разностного уравнения и соответствующий интервал времени для исходного дифференциального уравнения.

Если амплитуда гармоник разностного решения уменьшается быстрее, чем в случае точного решения, то это явление называют *амплитудной ошибкой*, вызванной присущей разностной схеме дополнительной *диффузией*, или *схемной вязкостью*. Изменение фазы гармоник разностного решения, отличное от изменения фазы гармоник точного решения, называют *фазовыми ошибками*, вызванными различием в фазовых скоростях распространения гармоник (иначе говоря, наличием *вычислительной дисперсии*).

Все ошибки порождаются аппроксимацией дифференциального оператора разностным и вкуче составляют уже упоминавшуюся *ошибку аппроксимации*. Не пытаясь строить общую теорию, рассмотрим ошибки, присущие разностным схемам на отдельных примерах.

13.1. Амплитудные и фазовые ошибки

Рассмотрим, как изменяется амплитуда и фаза k -й гармоники под действием дифференциального оператора и аппроксимирующего его разностного.

- 1) В качестве первого примера возьмем уравнение переноса

$$\partial_t f + u \partial_x f = 0, \quad t \geq 0, \quad x \in [0, L], \quad u > 0.$$

Подставляя решение в виде k -й гармоники

$$f = A_k(t) e^{i \frac{2\pi}{L} kx},$$

получаем уравнение для амплитуды A_k

$$d_t \ln A_k = -iu \frac{2\pi}{L} k,$$

откуда

$$A_k(t) = A_k(0)e^{-iu\frac{2\pi}{L}kt}.$$

Таким образом, амплитуда во времени не меняется и, следовательно, модуль коэффициента перехода от шага к шагу равен единице:

$$\left| \frac{A_k(\Delta t)}{A_k(0)} \right| = 1.$$

Фаза k -й гармоники за время Δt (за один шаг по времени) меняется на величину

$$\varphi_k = -u\frac{2\pi}{L}k\Delta t = -\alpha\mu, \quad \alpha = \frac{2\pi}{L}k\Delta x, \quad \mu = \frac{u\Delta t}{\Delta x}.$$

Что же получится при замене дифференциального оператора разностным? Аппроксимируем уравнение переноса направленными разностями по времени и по пространству (разности против потока):

$$f_j^{n+1} - f_j^n + \mu(f_j^n - f_{j-1}^n) = 0.$$

Так же как при исследовании устойчивости методом Неймана, подставляем решение в виде

$$f_j^n = A_k^n e^{i\frac{2\pi}{L}jk\Delta x}$$

и находим

$$A_k^{n+1} = A_k^n ((1 - \mu + \mu \cos \alpha) - i\mu \sin \alpha), \quad \alpha = \frac{2\pi}{L}k\Delta x.$$

Модуль коэффициента перехода

$$\left| \frac{A_k^{n+1}}{A_k^n} \right| = \sqrt{1 - 4\mu(1 - \mu) \sin^2 \frac{\alpha}{2}}$$

оказывается меньшим единицы и зависящим (через α) от номера гармоники k . Таким образом, в отличие от точного решения имеет место затухание и тем большее, чем выше номер гармоники (рис. 21). Максимальным затуханием обладает волна длиной $2\Delta x$ (так называемая двухшаговая волна), так как в этом случае $k = \frac{L}{2\Delta x}$ и $\sin^2 \frac{\alpha}{2} = 1$.

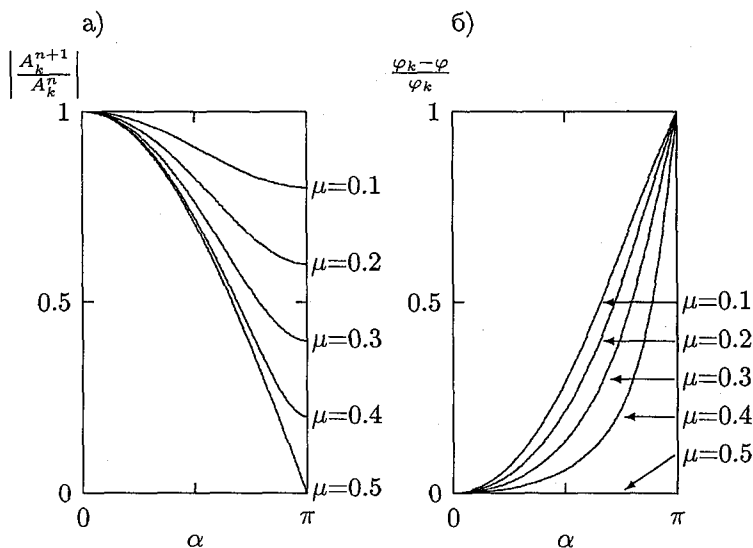


Рис. 21. Амплитудные (а) и фазовые (б) ошибки разностного аналога уравнения переноса.

На каждом шаге по времени фаза k -й гармоники меняется на величину

$$\varphi = \arctg \frac{\operatorname{Im}(f_j^{n+1})}{\operatorname{Re}(f_j^{n+1})} = \arctg \frac{-\mu \sin \alpha}{1 - \mu + \mu \cos \alpha}.$$

Относительная фазовая ошибка разностного решения соответственно равна

$$\frac{\varphi_k - \varphi}{\varphi_k} = 1 - \frac{1}{\mu \alpha} \arctg \frac{\mu \sin \alpha}{1 - \mu + \mu \cos \alpha}.$$

При $\mu = \frac{1}{2}$ эта ошибка равна нулю. В остальных случаях имеет место либо запаздывание ($\mu < \frac{1}{2}$), либо ускорение ($\mu > \frac{1}{2}$). Фазовая ошибка в зависимости от номера гармоники показана на рис.21. Как и амплитудная ошибка, она растет с увеличением волнового числа и достигает наибольшего значения для двухшаговой волны.

2) Рассмотрим теперь уравнение теплопроводности

$$\partial_t f = \kappa \partial_{xx} f, \quad t \geq 0, \quad x \in [0, L], \quad \kappa > 0.$$

Подстановкой решения в виде $f = A_k(t)e^{i\frac{2\pi}{L}kx}$ получаем закон изменения во времени амплитуды k -й гармоники

$$A_k(t) = A_k(0)e^{-\kappa(\frac{2\pi}{L}k)^2 t}.$$

Отсюда следует, что модуль коэффициента перехода равен

$$\left| \frac{A_k(\Delta t)}{A_k(0)} \right| = e^{-\alpha^2 d}, \quad \alpha = \frac{2\pi}{L} k \Delta x, \quad d = \frac{\kappa \Delta t}{\Delta x^2},$$

т.е. имеет место экспоненциальное затухание амплитуды, тем большее, чем выше номер гармоники. Фаза гармоник при этом во времени не меняется.

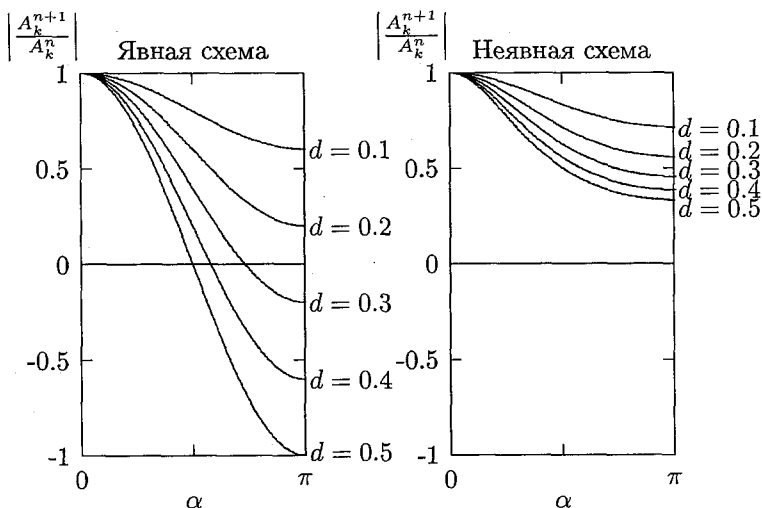


Рис. 22. Зависимость коэффициента перехода при различных значениях d для случая явной и неявной разностных схем уравнения теплопроводности.

Аппроксимируем теперь дифференциальное уравнение явной центрально-разностной схемой

$$f_j^{n+1} - f_j^n = d(f_{j+1}^n - 2f_j^n + f_{j-1}^n).$$

Коэффициент перехода есть вещественная функция номера гармоники

$$\lambda = 1 - 4d \sin^2 \frac{\alpha}{2} = 1 - 4d \sin^2 \frac{\pi}{L} k \Delta x ,$$

и, следовательно, фазовая ошибка отсутствует. На рис. 22 приведены зависимости коэффициентов перехода разностного уравнения при различных значениях d . Там же показаны аналогичные кривые для случая неявной разностной схемы.

13.2. Природа амплитудной и фазовой ошибок

Проведенный выше анализ показал, что замена дифференциального оператора разностным приводит к несвойственному исходной задаче затуханию Фурье-компонент решения, а в случае уравнения переноса — и к появлению фазовой ошибки. Эти ошибки оказываются зависящими от номера гармоники. Какова же природа ошибок, приносимых разностными операторами? Для ответа на этот вопрос попробуем вывести дифференциальное уравнение, решение которого в узлах сетки совпадает с решением конечно-разностного уравнения. Воспользуемся примерами из предыдущего пункта и начнем с уравнения переноса.

1) Заменяем в разностном уравнении

$$f_j^{n+1} - f_j^n + \mu(f_j^n - f_{j-1}^n) = 0 \quad (13.1)$$

значения f_j^{n+1} и f_{j-1}^n разложениями в ряд Тейлора относительно узла $x = j\Delta x$, $t = n\Delta t$. В результате получим

$$\begin{aligned} \partial_t f + u \partial_x f = & -\frac{\Delta t}{2} \partial_{tt} f + u \frac{\Delta x}{2} \partial_{xx} f - \\ & -\frac{\Delta t^2}{6} \partial_{ttt} f - u \frac{\Delta x^2}{6} \partial_{xxx} f + \dots \end{aligned} \quad (13.2)$$

Правая часть уравнения (13.2), в общем случае отличная от нуля, есть погрешность аппроксимации. Таким образом, переходя к разностной схеме, мы заменяем исходное уравнение переноса уравнением (13.2). Ограничившись членами первого порядка малости, получим так называемое *первое дифференциальное приближение разностной схемы*:

$$\partial_{tt} f = \frac{u \Delta x}{\Delta t} \partial_{xx} f - \frac{2}{\Delta t} (\partial_t f + u \partial_x f) , \quad (13.3)$$

которое является гиперболическим уравнением. Оно обладает двумя семействами характеристик, задаваемых уравнениями

$$x = \pm t \sqrt{\frac{u \Delta x}{\Delta t}} + \text{const}.$$

Наличие двух характеристических направлений объясняет возникновение ошибки: в отличие от исходного уравнения переноса с одним характеристическим направлением начальные значения переносятся в противоположных направлениях вдоль двух семейств характеристик. Такое поведение решения похоже на процесс диффузии. И действительно, погрешность аппроксимации можно интерпретировать как появление дополнительного вязкого члена в исходном уравнении переноса. Для этого удобно заменить производные по времени производными по пространству. Дифференцируя (13.2) по t и по x и исключая затем из двух полученных уравнений смешанную производную $\partial_{xt}f$, находим

$$\partial_{tt}f = u^2 \partial_{xx}f + O(\Delta t, \Delta x). \quad (13.4)$$

Аналогично можно заменить и другие производные. В итоге получаем так называемое *модифицированное уравнение*:

$$\partial_t f + u \partial_x f = \frac{u \Delta x}{2} (1 - \mu) \partial_{xx}f + O(\Delta t^2, \Delta x^2). \quad (13.5)$$

Как видим, конечно-разностное уравнение моделирует процесс переноса при наличии диффузии (схемной вязкости) с коэффициентом диффузии, зависящим от шага сетки по пространству Δx , а также от соотношения шагов по времени и по пространству μ . Поскольку процесс диффузии сопровождается уменьшением амплитуд Фурье-гармоник решения, не имеющим места для уравнения переноса, замена исходного уравнения уравнением (13.5) приводит к появлению амплитудных ошибок.

Из уравнения (13.5) вытекает еще одна трактовка условия устойчивости. Краевая задача для параболического уравнения (13.5) оказывается корректной лишь при положительном коэффициенте диффузии

$$\frac{u \Delta x}{2} (1 - \mu) \geq 0.$$

Если нарушается условие устойчивости $\mu \leq 1$, коэффициент диффузии становится отрицательным и краевая задача перестает быть корректной.

- 2) Рассмотрим далее явную разностную схему для уравнения теплопроводности

$$f_j^{n+1} - f_j^n = d(f_{j+1}^n - 2f_j^n + f_{j-1}^n).$$

Первое дифференциальное приближение имеет вид

$$\partial_{tt}f = \frac{2\kappa}{\Delta t} \partial_{xx}f - \frac{2}{\Delta t} \partial_t f. \quad (13.6)$$

Как и в случае уравнения переноса, мы получаем гиперболическое (волновое) уравнение и процесс теплопередачи фактически моделируем процессом распространения волн. Переходя от уравнения (13.6) к модифицированному уравнению, находим, что с точностью до $O(\Delta t^2, \Delta x^4)$ оно совпадает с исходным уравнением теплопроводности

$$\partial_t f = \kappa \partial_{xx} f.$$

Таким образом, амплитудные ошибки порождаются главным образом за счет замены уравнения теплопроводности, обладающего бесконечной скоростью распространения возмущений (характеристики параллельны оси x), волновым уравнением, имеющим конечную скорость распространения возмущений $\frac{2\kappa}{\Delta t}$ (характеристики образуют с осью x углы $\pm \arctg \frac{\Delta t}{2\kappa}$).

14. Организация вычислений

14.1. Понятие о методах расщепления

Применение разобранных методов, как правило, не вызывает затруднений в одномерных задачах. На практике, однако, такие случаи встречаются редко, а рассмотрение двух- и трехмерных задач заметно усложняет алгоритмы. Особенно это касается неявных разностных схем. В этой связи очень полезными оказываются методы, позволяющие свести шаг интегрирования сложной многомерной задачи к последовательному интегрированию более простых одномерных задач. Такие методы носят название *методов расщепления*.

Будем рассматривать часто встречающуюся ситуацию, когда в эволюционном уравнении

$$\partial_t f + Lf = 0, \quad f|_{t=0} = f_0 \quad (14.1)$$

оператор L , содержащий лишь пространственные производные, может быть записан в виде суммы более простых операторов L_1 и L_2 . Например, двумерный оператор $u\partial_x + v\partial_y$ естественно представляется в виде суммы одномерных операторов $u\partial_x$ и $v\partial_y$. Итак

$$L = L_1 + L_2. \quad (14.2)$$

Сначала рассмотрим явную первого порядка точности по времени аппроксимацию уравнения (14.1)

$$\frac{f^{n+1} - f^n}{\Delta t} + Lf^n = 0, \quad f^0 = f_0. \quad (14.3)$$

С учетом (14.2) это выражение можно записать в виде

$$f^{n+1} = (I - \Delta t (L_1 + L_2)) f^n. \quad (14.4)$$

Поскольку разностная схема (14.3) аппроксимирует уравнение (14.1) с первым порядком, добавление произвольных членов более высокого порядка малости не изменит аппроксимирующих свойств схемы (14.3). Этим можно воспользоваться для того, чтобы преобразовать (14.4) к виду, содержащему композицию новых операторов, зависящих от L_1 и L_2 порознь.

Добавим к (14.4) член $\Delta t^2 L_1 L_2 f^n$ второго порядка малости, тогда (14.4) может быть записано в виде

$$f^{n+1} = (I - \Delta t L_1) (I - \Delta t L_2) f^n, \quad (14.5)$$

и алгоритм вычисления f^{n+1} расщепляется на два этапа:

- 1) вычисление промежуточного значения φ

$$\varphi = (I - \Delta t L_2) f^n = f^n - \Delta t L_2 f^n,$$

- 2) вычисление решения на $(n+1)$ временном слое

$$f^{n+1} = (I - \Delta t L_1) \varphi = \varphi - \Delta t L_1 \varphi.$$

Из (14.5) следует, что оператор перехода имеет вид

$$G = (I - \Delta t L_1) (I - \Delta t L_2),$$

и для устойчивости алгоритма в целом требуется

$$\|G\| \leq \|I - \Delta t L_1\| \|I - \Delta t L_2\| \leq 1.$$

Подробно вопросы устойчивости алгоритмов расщепления рассматриваются в монографиях Г.И.Марчука (см. список литературы).

Аналогичные рассуждения применимы и к неявной разностной схеме вида

$$\frac{f^{n+1} - f^n}{\Delta t} + L f^{n+1} = 0, \quad f^0 = f_0, \quad (14.6)$$

для которой получаем

$$(I + \Delta t L_1) (I + \Delta t L_2) f^{n+1} = f^n. \quad (14.7)$$

Вычисление f^{n+1} также осуществляется в два этапа:

- 1) вычисление промежуточного значения φ

$$(I + \Delta t L_1) \varphi = f^n, \quad \text{или} \quad \varphi = (I + \Delta t L_1)^{-1} f^n,$$

2) вычисление f^{n+1}

$$(I + \Delta t L_2) f^{n+1} = \varphi \quad \text{или} \quad f^{n+1} = (I + \Delta t L_2)^{-1} \varphi.$$

В обоих случаях необходимо обращаться операторы более простые, нежели исходный оператор $(I + \Delta t L) = (I + \Delta t (L_1 + L_2))$.

Теперь для решения задачи (14.1) воспользуемся разностной схемой Кранка-Николсон второго порядка точности по времени:

$$\frac{f^{n+1} - f^n}{\Delta t} + \frac{1}{2} L (f^{n+1} + f^n) = 0, \quad f^0 = f_0. \quad (14.8)$$

Раскроем скобки и приведем подобные члены.

$$\left(I + \frac{\Delta t}{2} (L_1 + L_2) \right) f^{n+1} = \left(I - \frac{\Delta t}{2} (L_1 + L_2) \right) f^n. \quad (14.9)$$

Если теперь к левой части полученного выражения добавить член $\frac{1}{4} \Delta t^2 L_1 L_2 (f^{n+1} - f^n)$, имеющий третий порядок точности по времени (так как согласно (14.8) разность $f^{n+1} - f^n = O(\Delta t)$), то порядок аппроксимации схемы (14.9) не изменится, а разностную схему можно будет записать в виде:

$$\left(I + \frac{\Delta t}{2} L_1 \right) \left(I + \frac{\Delta t}{2} L_2 \right) f^{n+1} = \left(I - \frac{\Delta t}{2} L_1 \right) \left(I - \frac{\Delta t}{2} L_2 \right) f^n,$$

откуда следует четырехэтапный алгоритм расчета очередного шага по времени:

1) вычисление первого промежуточного значения φ^1

$$\varphi^1 = \left(I - \frac{\Delta t}{2} L_2 \right) f^n,$$

2) вычисление второго промежуточного значения φ^2

$$\varphi^2 = \left(I - \frac{\Delta t}{2} L_1 \right) \varphi^1,$$

3) вычисление третьего промежуточного значения φ^3

$$\left(I + \frac{\Delta t}{2} L_1\right) \varphi^3 = \varphi^2,$$

4) получение решения на $(n+1)$ временном слое

$$\left(I + \frac{\Delta t}{2} L_2\right) f^{n+1} = \varphi^3$$

На первых двух этапах используется явный алгоритм, а на двух последних необходимо обращать операторы.

На базе схемы Кранка-Николсон можно построить и другой алгоритм расщепления, если (14.9) переписать в виде

$$\left(I + \frac{\Delta t}{2} (L_1 + L_2)\right) f^{n+1} = \left(I + \frac{\Delta t}{2} (L_1 + L_2)\right) f^n - \Delta t L f^n,$$

или

$$\left(I + \frac{\Delta t}{2} (L_1 + L_2)\right) (f^{n+1} - f^n) = -\Delta t L f^n.$$

Откуда получаем

$$\left(I + \frac{\Delta t}{2} L_1\right) \left(I + \frac{\Delta t}{2} L_2\right) \left(\frac{f^n - f^{n+1}}{\Delta t}\right) = L f^n.$$

Алгоритм интегрирования на шаг по времени состоит из следующих этапов:

- 1) $\varphi^1 = L f^n$,
- 2) $\left(I + \frac{\Delta t}{2} L_1\right) \varphi^2 = \varphi^1$,
- 3) $\left(I + \frac{\Delta t}{2} L_2\right) \varphi^3 = \varphi^2$,
- 4) $f^{n+1} = f^n - \Delta t \varphi^3$.

Здесь, так же как и в предыдущем случае, два этапа содержат обращение оператора, а в двух других вычисления проводятся по явным формулам.

14.2. Вычислительные граничные условия

Сведение дифференциальной задачи к разностной порождает некоторые новые проблемы и явления, связанные с аппроксимацией начальных и граничных условий. Пусть, например, мы аппроксимируем следующую задачу для уравнения переноса:

$$\partial_t f + u \partial_x f = 0, \quad x \in [0, L], \quad u > 0, \quad t > 0,$$

$$f|_{t=0} = f^0(x),$$

$$f|_{x=0} = f_0(t)$$

на пятиточечном шаблоне (см. рис. 16) центральными разностями по времени и по пространству

$$f_j^{n+1} - f_j^{n-1} + \mu(f_{j+1}^n - f_{j-1}^n) = 0, \quad \mu = \frac{u \Delta t}{\Delta x}. \quad (14.10)$$

На временном слое $(n+1)$ значение в первой расчетной точке f_1^{n+1} вычисляется с учетом граничного условия

$$f_1^{n+1} = f_1^{n-1} - \mu(f_2^n - f_0^n),$$

далее используется формула (14.10), а для последнего расчетного узла имеем

$$f_L^{n+1} = f_L^{n-1} - \mu(f_{L+1}^n - f_{L-1}^n),$$

где индексом L отмечено значение f в узле сетки, совпадающим с концом отрезка $[0, L]$. Для вычисления f_L^{n+1} , таким образом, необходимо задание величины f_{L+1}^n .

Это требование не связано с дифференциальной задачей, а диктуется исключительно используемой конечно-разностной аппроксимацией. Более того, как мы знаем, второе независимое граничное условие на конце L отрезка $[0, L]$ ставить нельзя, ибо так сформулированная краевая задача, вообще говоря, не имеет решения. Для разностного уравнения (14.10) ситуация иная, и для определения f_L^{n+1} необходимо дополнительное граничное условие, которое называется *вычислительным граничным условием*. Если воспользоваться схемой первого порядка точности с направленными разностями, то такой проблемы

не возникает. Однако в ряде случаев первый порядок точности оказывается недостаточным, и при использовании направленных разностей второго порядка точности снова необходимо ставить вычислительное граничное условие. На этот раз требуется задавать значение в фиктивном узле f_{-1}^n . Во избежание подобных затруднений значения вдоль границ часто рассчитывают по схемам первого порядка точности, что, разумеется, снижает общую точность расчета.

Аналогичная ситуация возникает при интегрировании уравнения (14.10) шагами по времени. Для нахождения вектора f^2 помимо физического начального условия f^0 необходимо задать *вычислительное начальное условие* f^1 . Такое дополнительное начальное условие обычно получают с помощью какой-либо двуслойной разностной схемы, т.е. схемы более низкого порядка точности.

Постановка вычислительного начального условия приводит к появлению в решении так называемой *вычислительной моды* в отличие от *физической моды* — искомого решения. Рассмотрим, как это происходит. Коэффициент перехода λ для схемы (14.10) является решением квадратного уравнения

$$\lambda^2 + (i2\mu \sin \alpha)\lambda - 1 = 0.$$

Из условия устойчивости $|\lambda|^2 \leq 1$ следует ограничение $\mu \leq 1$. Если теперь ввести обозначение $\sin \beta \equiv \mu \sin \alpha$, то легко получить выражение для корней $\lambda_{1,2}$ в терминах β :

$$\lambda_{1,2} = \pm \cos \beta - i \sin \beta,$$

или

$$\lambda_1 = e^{-i\beta}, \quad \lambda_2 = -e^{i\beta}.$$

Решения, связанные с λ_1 , являются приближением к решению задачи и называются *физическими модами*. Решения же, связанные с λ_2 , на каждом шаге по времени меняют знак и потому оказываются паразитными *вычислительными модами*.

Выбор вычислительного начального условия определяет долю вычислительной моды в решении. Допустим скорость переносащего потока $u = 0$, тогда уравнение сводится к $\partial_t f = 0$, решением является $f = \text{const}$, $\mu = 0$, и коэффициенты перехода соответственно равны $\lambda_1 = 1$, $\lambda_2 = -1$. Векторы решения на двух соседних временных слоях,

в частности на первом и втором слое, связаны между собой зависимостью

$$f^2 = \lambda f^1,$$

где λ — коэффициент перехода. Рассмотрим два крайних случая выбора f^1 .

- 1) Пусть $f^1 = f^0$. Из разностного уравнения следует, что $f^2 = f^0$. Сравнивая оба выражения, находим, что $\lambda = 1 = \lambda_1$, или $f^2 = \lambda_1 f^1 = \lambda_1^2 f^0$. Вычисляя, далее, для n -го временного слоя, получаем $f^n = \lambda_1^n f^0$, т.е. решение постоянно во времени, как и должно быть. В этом случае оно состоит только из физической моды, а вычислительная мода отсутствует.
- 2) Пусть теперь $f^1 = -f^0$. Аналогичные рассуждения показывают, что $\lambda = -1 = \lambda_2$, а $f^n = \lambda_2^n f^0$. На этот раз решение на каждом шаге по времени меняет знак, что не наблюдается у точного решения, и, следовательно, состоит лишь из вычислительной моды.

В общем случае решение будет включать в себя комбинацию обеих мод:

$$f^n = a\lambda_1^n f^0 + b\lambda_2^n f^0.$$

Это решение должно удовлетворять физическому и вычислительному начальным условиям

$$af^0 + bf^0 = f^0, \quad a\lambda_1 f^0 + b\lambda_2 f^0 = f^1,$$

откуда

$$a = \frac{f^1 - \lambda_2 f^0}{f^0(\lambda_1 - \lambda_2)}, \quad b = (1 - a).$$

Таким образом,

$$f^n = \lambda_1^n \frac{f^1 - \lambda_2 f^0}{(\lambda_1 - \lambda_2)} + \lambda_2^n \frac{\lambda_1 f^0 - f^1}{(\lambda_1 - \lambda_2)} = \lambda_1^n f_{\text{ф}} + \lambda_2^n f_{\text{в}},$$

где $f_{\text{ф}}$ — физическая мода, а $f_{\text{в}}$ — вычислительная.

Ясно, что если $f^1 = \lambda_1 f^0$, т.е. значения f^1 находятся с помощью двуслойной схемы

$$f_j^1 - f_j^0 + \frac{\mu}{2}(f_{j+1}^0 - f_{j-1}^0) = 0,$$

решение будет содержать только физическую моду. Но это лишь в идеальном случае, когда все вычисления проводятся без ошибок. Ошибки, однако, всегда есть (например, ошибки округления), что рано или поздно приведет к появлению вычислительной моды. Стандартным методом борьбы с такими паразитными решениями при использовании трехслойных схем является периодическое включение двуслойной схемы, т.е. проведение расчета очередного шага с помощью алгоритма, не обладающего вычислительной модой.

15. Обзор методов решения систем линейных алгебраических уравнений

Рассмотренные в третьей главе методы сводили краевую задачу к линейной системе алгебраических уравнений. Этой теме посвящена обширная литература от основополагающих фундаментальных работ до справочников и школьных учебников. Созданы и опубликованы многочисленные программы для ЭВМ, предназначенные для решения задач линейной алгебры. Наряду с этим сплошь и рядом встречаются ситуации, когда системы линейных уравнений приходится решать самому, не прибегая к готовым программам. Но даже если и используются библиотечные программы, следует понимать, как они работают. По этой причине рассмотрим бегло основные методы решения систем линейных алгебраических уравнений.

Традиционно эти методы делятся на *точные* (дающие точный ответ без учета ошибок округления) и *итерационные* (точное решение есть предел бесконечной последовательности приближений). Начнем наш обзор с итерационных методов.

15.1. Итерационные методы

15.1.1. Общие сведения

Итак, требуется решить уравнение

$$Af = g, \quad (15.1)$$

где A — квадратная $n \times n$ матрица с элементами a_{ij} , а $f = (f_1, \dots, f_n)$, $g = (g_1, \dots, g_n)$ — n -мерные векторы. Итерационные методы решения системы (15.1) позволяют строить последовательность *приближений* (*итераций*) $\{f^k\}_{k=0}^{\infty}$. Если эта последовательность имеет предел, то он называется *решением системы* (15.1):

$$f = \lim_{k \rightarrow \infty} f^k.$$

Очередное приближение строится с помощью рекуррентной формулы

$$f^{k+1} = f^{k+1}(A, g, f^0, \dots, f^k),$$

где k — номер итерации. Преимущественно используется простейшая рекуррентная формула

$$\mathbf{f}^{k+1} = B\mathbf{f}^k + \mathbf{h}, \quad (15.2)$$

где B — квадратная $n \times n$ матрица, а $\mathbf{h}$ — n -мерный вектор. Для организации счета по такой формуле требуется задание начального приближения $\mathbf{f}^0$. Вид матрицы B и вектора $\mathbf{h}$ определяет тот или иной итерационный метод.

Многие итерационные методы сводятся к *методу последовательных приближений*, который был проанализирован в главе 1. Там же было получено условие сходимости метода Якоби — диагональное преобладание матрицы A — и произведена оценка погрешности.

В общем случае итерационный метод сходится, если $\|B\| < 1$, т.е.:

- 1) если все собственные числа матрицы B по модулю меньше единицы;
- 2) если имеет место диагональное преобладание элементов матрицы B .

Все итерационные методы строятся по одной схеме: система (15.1) преобразуется к виду:

$$B_1\mathbf{f} = B_2\mathbf{f} + \mathbf{g}, \quad (15.3)$$

где $B_1 - B_2 = A$, и B_1 легко обратимая матрица: единичная, диагональная или треугольная. В этом случае из (15.2) без труда получается выражение:

$$\mathbf{f} = B_1^{-1}(B_2\mathbf{f} + \mathbf{g}) = B\mathbf{f} + \mathbf{h},$$

которое далее используется в качестве рекуррентной формулы (15.2)

15.1.2. Обзор основных методов

Рассмотрим наиболее часто встречающиеся методы.

Метод последовательных приближений (метод Якоби). Исходная система (15.1) преобразуется к виду:

$$A\mathbf{f} = (A + I - I)\mathbf{f} = \mathbf{g},$$

где I — единичная матрица, и далее

$$\mathbf{f} = (I - A)\mathbf{f} + \mathbf{g}.$$

Счет организуется по формуле

$$\mathbf{f}^{k+1} = (I - A)\mathbf{f}^k + \mathbf{g}. \quad (15.4)$$

Таким образом $B = I - A$, а $\mathbf{h} = \mathbf{g}$. В компонентной форме рекуррентная формула метода выглядит так:

$$f_i^{k+1} = \sum_j (\delta_{ij} - a_{ij}) f_j^k + g_i = f_i^k - \sum_j a_{ij} f_j^k + g_i.$$

Метод простой итерации. Представим матрицу A в виде суммы

$$A = L + D + R, \quad (15.5)$$

где D — диагональная матрица с элементами a_{ii} , L и R — нижняя и верхняя треугольные матрицы, состоящие из соответствующих элементов исходной матрицы A . Система (15.1) подвергается далее следующему преобразованию:

$$A\mathbf{f} = (A + D - D)\mathbf{f} = D\mathbf{f} - (D - A)\mathbf{f} = \mathbf{g},$$

откуда получаем

$$\mathbf{f}^{k+1} = D^{-1}(D - A)\mathbf{f}^k + D^{-1}\mathbf{g}, \quad (15.6)$$

$$B = D^{-1}(D - A) = -D^{-1}(L + R), \quad \mathbf{h} = D^{-1}\mathbf{g}.$$

Матрица D^{-1} , обратная диагональной матрице D , вычисляется просто, поскольку также является диагональной с элементами $1/a_{ii}$. В компонентной форме рекуррентная формула метода такова:

$$f_i^{k+1} = \frac{1}{a_{ii}} \left(\sum_j (\delta_{ij} - a_{ij}) f_j^k + g_i \right) = \frac{1}{a_{ii}} \left(f_i^k - \sum_j a_{ij} f_j^k + g_i \right).$$

Метод Гаусса-Зейделя⁴². Снова запишем матрицу в виде (15.5), а систему (15.1) преобразуем к виду

$$Af = (L + D)f + Rf = g.$$

Отсюда получаем рекуррентную формулу

$$f^{k+1} = -(L + D)^{-1}Rf^k + (L + D)^{-1}g,$$

$$B = -(L + D)^{-1}R, \quad h = (L + D)^{-1}g.$$

Компонентную форму рекуррентной формулы метода можно получить так. Запишем сначала в компонентной форме выражение $Lf^{k+1} + Df^{k+1} + Rf^k = g$. Получим

$$\sum_{j < i} a_{ij}f_j^{k+1} + a_{ii}f_i^{k+1} + \sum_{j > i} a_{ij}f_j^k = g_i.$$

Переносим теперь первое и третье слагаемые в правую часть и нормируя на диагональный элемент, найдем выражение для f_i^{k+1} :

$$f_i^{k+1} = \frac{1}{a_{ii}} \left(- \sum_{j < i} a_{ij}f_j^{k+1} - \sum_{j > i} a_{ij}f_j^k + g_i \right).$$

Метод последовательной релаксации Пусть $\tau > 0$ — число, которое мы будем называть параметром релаксации. Умножим на него систему (15.1)

$$\tau Af = \tau g$$

и преобразуем ее следующим образом:

$$\begin{aligned} \tau Af &= (\tau L + \tau D + \tau R + D - D)f = \\ &= (\tau L + D)f + ((\tau - 1)D + \tau R)f = \tau g. \end{aligned}$$

В результате получаем рекуррентную формулу (запишите ее в компонентной форме):

$$f^{k+1} = -(\tau L + D)^{-1}((\tau - 1)D + \tau R)f^k + (\tau L + D)^{-1}\tau g,$$

⁴²Фон Зейдель (Seidel) Филипп Людвиг (1821–1896) — немецкий математик.

$$B = -(\tau L + D)^{-1}((\tau - 1)D + \tau R), \quad \mathbf{h} = (\tau L + D)^{-1}\tau \mathbf{g}.$$

Параметр релаксации используется для настройки метода на максимальную скорость сходимости и обычно подбирается эмпирически. При этом $\tau \in (0, 2)$. Если $\tau > 1$, получаем метод верхней релаксации, если $\tau < 1$ — метод нижней релаксации.

15.2. Точные методы

15.2.1. Метод Гаусса

Из точных методов рассмотрим один из лучших методов решения систем линейных уравнений — *метод Гаусса* и некоторые его модификации. Метод Гаусса, или метод последовательных исключений, базируется на так называемой *LU-теореме*, согласно которой при определенных условиях (проверяемых, как правило, *a posteriori*) матрица A в уравнении (15.1) может быть единственным образом представлена в виде произведения (*факторизована*) $A = LU$, где L — нижняя треугольная матрица с единицами на главной диагонали, а U — верхняя треугольная матрица:

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ \ell_{21} & 1 & 0 & \cdots \\ \ell_{31} & \ell_{32} & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots \\ 0 & u_{22} & u_{23} & \cdots \\ 0 & 0 & u_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

В этом случае система (15.1) записывается как

$$LU\mathbf{f} = \mathbf{g}$$

и сводится к двум системам с треугольными матрицами

$$\begin{aligned} L\mathbf{h} &= \mathbf{g}, \\ U\mathbf{f} &= \mathbf{h}. \end{aligned}$$

Решая первую систему, находим промежуточный вектор $\mathbf{h}$ (эта процедура носит название *прямого хода*, или *исключения*). Из второй системы (с помощью *обратного хода*, или *подстановки*) определяем решение $\mathbf{f}$. Обе системы решаются элементарно, так как матрицы L и U треугольные.

Часто метод Гаусса используется без предварительной факторизации. Матрица U и вектор h строятся без явного построения матрицы L и решения соответствующей системы уравнений. Эта процедура состоит в том, что очередная строка (начиная с первой) используется для обнуления части столбца, стоящего под главной диагональю, для чего она умножается на некоторое число α и вычитается из нижестоящих строк. Коэффициент α выбирается таким, чтобы элемент в соответствующем столбце обратился в нуль. В результате такого преобразования (прямого хода) мы сразу получаем уравнение

$$Uf = h.$$

Обратным ходом, начиная с последней компоненты вектора f_n , мы восстанавливаем решение f .

В таком виде метод Гаусса работает лишь при условии, что главная диагональ матрицы A не содержит нулевых элементов. В противном случае применяют его модификацию — *метод Гаусса с выбором главного (ведущего) элемента* — которая состоит в том, что перед обнулением очередного столбца нижележащие уравнения переставляются так, чтобы на главной диагонали оказался максимальный по модулю (главный или ведущий) элемент.

15.2.2. Методы прогонки

Решение больших систем уравнений довольно времязатратная задача, и отсутствие быстрых алгоритмов долго сдерживало широкое использование таких вычислительных методов, как преобразование Фурье, метод конечных элементов, неявные разностные методы и пр. Надо сказать, что в общем случае (матрица системы произвольна) альтернативы методу Гаусса, по сути, нет. Другое дело, если матрица системы обладает некоторыми заранее известными свойствами — положительно определенная, симметричная, ортогональная, ленточная и т.п. В этом случае, используя априорную информацию о структуре матрицы, можно построить алгоритм решения систем уравнений, существенно более быстрый по сравнению с методом Гаусса. Цена этому быстродействию — потеря универсальности.

Численные методы решения дифференциальных уравнений часто приводят к системам линейных алгебраических уравнений с ленточ-

ной (в простейшем случае — трехдиагональной) матрицей. Для решения систем с подобной матрицей разработан специальный алгоритм, называемый *методом прогонки*, который является ни чем иным, как методом Гаусса, учитывающим структуру матрицы. Суть метода прогонки состоит в том, что, как и в методе Гаусса, с помощью очередной строки производят исключение (обнуление) поддиагональных элементов столбца. Однако, поскольку известно, что матрица имеет ненулевые элементы, расположенные вдоль диагонали на ленте шириной 2ℓ , процесс Гаусса не доводят до конца, а ограничиваются лишь обнулением первых ℓ поддиагональных элементов.

Рассмотрим детальнее простейший и наиболее часто встречающийся случай трехдиагональной матрицы. Система уравнений имеет вид:

$$\begin{pmatrix} a_{11} & a_{12} & & & 0 \\ a_{21} & a_{22} & a_{23} & & \\ & a_{32} & a_{33} & a_{34} & \\ & 0 & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \\ g_3 \\ \vdots \end{pmatrix}.$$

Прямым ходом прогонки мы обнуляем расположенные под главной диагональю элементы $a_{21}, a_{32}, \dots, a_{i,i-1}, \dots, a_{n,n-1}$ и получаем двухдиагональную матрицу:

$$\begin{pmatrix} a_{11} & a_{12} & & & 0 \\ 0 & \tilde{a}_{22} & a_{23} & & \\ & 0 & \tilde{a}_{33} & a_{34} & \\ & 0 & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} g_1 \\ \tilde{g}_2 \\ \tilde{g}_3 \\ \vdots \end{pmatrix}.$$

Элементы, помеченные тильдой, преобразуются в результате прямого хода. Формулы преобразования имеют вид

$$\tilde{a}_{ii} = a_{ii} - a_{i-1,i}\beta_i, \quad \tilde{g}_i = g_i - \tilde{g}_{i-1}\beta_i,$$

$$\beta_i = \frac{a_{i,i-1}}{\tilde{a}_{i-1,i-1}}, \quad i = 2, \dots, n.$$

С помощью обратного хода по формулам

$$f_n = \frac{\tilde{g}_n}{\tilde{a}_{nn}},$$

$$f_i = \frac{\bar{g}_i - a_{i,i+1}f_{i+1}}{\bar{a}_{ii}}$$

получаем решение системы линейных уравнений.

15.2.3. Циклическая прогонка

Рассмотренный вариант метода прогонки может применяться для систем линейных алгебраических уравнений, получающихся, в частности, при аппроксимации дифференциальных уравнений с краевыми условиями Неймана или Дирихле⁴³. В задачах гидромеханики часто встречается, однако, еще один тип краевых условий — так называемые *периодические граничные условия*. При аппроксимации таких задач конечными разностями матрица системы линейных алгебраических уравнений теряет ленточную структуру. Например, если для условий Неймана/Дирихле мы получаем трехдиагональную матрицу, то при постановке периодических граничных условий сохраняются три ненулевые диагонали и появляются ненулевые элементы a_{1n} и a_{n1} :

$$\begin{pmatrix} a_{11} & a_{12} & & & a_{1n} \\ a_{21} & a_{22} & a_{23} & & \\ & a_{32} & a_{33} & a_{34} & \\ & & & \ddots & \ddots \\ a_{n1} & & & \ddots & \ddots \end{pmatrix}.$$

Для систем уравнений с матрицами такого вида разработан специальный алгоритм, называемый *циклической прогонкой*. Пусть надо решить систему вида (здесь указаны только отличные от нуля элементы):

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & & & a_{1n} \\ a_{21} & a_{22} & a_{23} & & a_{2n} \\ & & \ddots & \ddots & \\ & & & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ a_{n1} & & & a_{n,n-1} & & a_{nn} \end{array} \right) \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_{n-1} \\ g_n \end{pmatrix}. \quad (15.7)$$

Запишем эту систему в виде системы матричного и скалярного уравнений:

$$\begin{cases} \bar{A}\bar{f} + f_n A_n = \bar{g}, \\ A^n \bar{f} + f_n a_{nn} = g_n. \end{cases} \quad (15.8)$$

⁴³ Дирихле (Dirichlet) Петер Густав Лежен (1805–1859) — немецкий математик.

Здесь $\bar{A}$, A_n , A^n , a_{nn} — выделенные части матрицы A , а $\bar{f}$, f_n и $\bar{g}$, g_n — соответствующие части векторов f и g :

$$A = \left(\begin{array}{c|c} \bar{A} & A_n \\ \hline A^n & a_{nn} \end{array} \right), \quad f = \left(\begin{array}{c} \bar{f} \\ f_n \end{array} \right), \quad g = \left(\begin{array}{c} \bar{g} \\ g_n \end{array} \right).$$

Поскольку исходная система линейна, вектор $\bar{f}$ должен линейно зависеть от компоненты f_n . Это видно из (15.8). Таким образом, можно записать

$$\bar{f} = -f_n \bar{p} + \bar{q}, \quad (15.9)$$

где $\bar{p}$ и $\bar{q}$ — некоторые векторы той же размерности, что и $\bar{f}$ (знаки выбраны для удобства). Подставляя (15.9) в первое уравнение (15.8), получаем

$$(\bar{A}\bar{q} - \bar{g}) - f_n(\bar{A}\bar{p} - A_n) = 0. \quad (15.10)$$

Равенство (15.10) тождественно удовлетворяется, если

$$\begin{cases} \bar{A}\bar{q} = \bar{g}, \\ \bar{A}\bar{p} = A_n. \end{cases} \quad (15.11)$$

Решив (15.11) и вычислив промежуточные векторы $\bar{p}$ и $\bar{q}$, из второго уравнения (15.8) найдем компоненту решения f_n :

$$f_n = \frac{A^n \bar{q} - g_n}{\bar{A}\bar{p} - a_{nn}}. \quad (15.12)$$

После этого остальные компоненты решения легко рассчитываются с помощью первого уравнения (15.8):

$$\bar{A}\bar{f} = (\bar{g} - f_n A_n). \quad (15.13)$$

Таким образом, решение f системы (15.7) ищется в три приема:

- 1) нахождение промежуточных векторов $\bar{p}$ и $\bar{q}$ из систем уравнений (15.11);
- 2) вычисление n -й компоненты решения f_n ;
- 3) получение всего вектора решения f путем решения системы (15.13).

Обратите внимание, что везде, где нам приходится обращаться матрицу (решать систему уравнений), мы имеем дело с трехдиагональной матрицей $\bar{A}$. Иначе говоря, циклическая прогонка сводится к случаю, разобранным в предыдущем пункте.

Литература к части I

- Коллатц Ф.* Функциональный анализ и вычислительная математика.— М.: Мир, 1969.
- Колмогоров А.Н., Фомин С.В.* Элементы теории функций и функционального анализа.— М.: Наука, 1968.
- Рихтмайер Р.* Принципы современной математической физики. Ч.1.— М.: Мир, 1982.
- Хатсон В., Пим Дж.* Приложения функционального анализа и теории операторов.— М.: Мир, 1983.

Литература к части II

- Арнольд В.И.* Обыкновенные дифференциальные уравнения.— М.: Наука, 1975.
- Арнольд В.И.* Дополнительные главы теории обыкновенных дифференциальных уравнений.— М.: Наука, 1978.
- Зельдович Я.Б., Мишкис А.Д.* Элементы математической физики.— М.: Наука, 1973.
- Курант Р.* Дифференциальные уравнения с частными производными.— М.: Наука, 1964.
- Тихонов А.Н., Самарский А.А.* Уравнения математической физики.— М.: Наука, 1978.

Литература к части III

Теории и применению разностных схем посвящена большая литература. Рассмотрение общих вопросов можно найти, например, в следующих книгах:

- Годунов С.К., Рябенский В.С.* Разностные схемы.— М.: Наука, 1977.

Марчук Г.И. Методы вычислительной математики.— М.: Наука, 1980.

Рунтмайер Р., Мортон К. Разностные методы решения краевых задач.— М.: Мир, 1972.

Прикладные аспекты отражены в монографиях:

Андерсон Д., Таннехил Дж., Плетчер Р. Вычислительная гидромеханика и теплообмен.— М.: Мир, 1990.

Белов П.Н. Численные методы прогноза погоды.— Л.: Гидрометеиздат, 1975.

Гордин В.А. Математические задачи гидродинамического прогноза погоды. Вычислительные аспекты.— Л.: Гидрометеиздат, 1987.

Марчук Г.И. Численные методы в прогнозе погоды.— Л.: Гидрометеиздат, 1967.

Марчук Г.И. Численное решение задач динамики атмосферы и океана.— Л.: Гидрометеиздат, 1974.

Мезингер Ф., Аракава А. Численные методы, используемые в атмосферных моделях.— Л.: Гидрометеиздат, 1979.

Поттер Д. Вычислительные методы в физике.— М.: Мир, 1975.

Рунч П. Вычислительная гидромеханика.— М.: Мир, 1980.

Численные методы решения задач динамики атмосферы и океана.— Л.: Гидрометеиздат, 1968.

Чему учить? Всякий знает, что ему нужно. Которые умнее, те берут что есть, которые поглупее — те ничего не получают, и всякий сам учится...

М. Горький. Макар Чудра.

Часть IV. Практические занятия

Практические занятия к части I

Упражнения 1

- 1) С помощью аксиом векторного пространства показать, что если V — векторное пространство, $f \in V$ и $\alpha, \beta \in \mathbb{R}^1$, то верны утверждения:

а) $f + g = h \Leftrightarrow f = h - g$,

б) $\alpha 0 = 0$,

в) $0f = 0$.

Решение.

- а) Прибавляя к обеим частям равенства вектор, противоположный вектору g , в соответствии с аксиомой (4) находим требуемый результат.

- б) Из аксиом (3) и (6) получим

$$\alpha(f + 0) = \begin{cases} \alpha f + \alpha 0 \\ \alpha f \end{cases} \Rightarrow \alpha 0 = 0.$$

- в) Из свойств чисел и аксиом (7) и (3) имеем

$$\beta f = (0 + \beta)f \stackrel{(7)}{=} 0f + \beta f \stackrel{(3)}{\Rightarrow} 0f = 0.$$

- 2) Рассмотрите множество точек окружности некоторого радиуса. Выберите точку нуль и определите сложение точек и умножение их на число так, чтобы множество точек окружности образовало векторное пространство.
- 3) Из каких элементов состоит множество, являющееся подпространством $\mathbb{R}^1, \mathbb{R}^2, \mathbb{R}^3$ и не совпадающее ни с одним из них?
- 4) Построить векторное пространство $n \times n$ матриц.
- 5) Ввести структуру векторного пространства на следующих множествах. Построить таблицу сложения и придумать способ умножения на вещественное число.

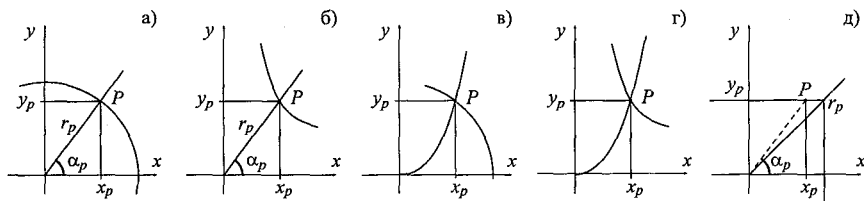


Рис. 23. Преобразования координат.

- а) Множество целых чисел $0, \dots, 9$.

Указание. В качестве примера и подсказки рассмотреть колесико счетчика.

- б) Множество состояний игральной кости (предполагается, что кость всегда лежит на одной из своих граней).

Указание. С каждым состоянием свяжем упорядоченную тройку чисел (по числу возможных направлений вращения кости). Выберем некоторое состояние за нулевое и поставим ему в соответствие элемент $(0, 0, 0)$. С поворотом кости на $\frac{\pi}{2}$ по трем возможным направлениям свяжем сложение одним из трех элементов: $(0, 0, 1)$, $(0, 1, 0)$ или $(1, 0, 0)$. Четыре поворота в одном направлении возвращают кость в исходное состояние.

Упражнения 2

- 1) Рассмотреть системы координат в первом квадранте:

- декартовы; координатные линии — перпендикулярные прямые,
- полярные; координатные линии — концентрические окружности и лучи, проходящие через центр окружностей,
- «гиперболические»; координатные линии — равнобокие гиперболы с общим центром и проходящие через него лучи,
- модифицированные полярные — 1; координатные линии — концентрические окружности и проходящие через их центр параболы,

- д) модифицированные «гиперболические»; координатные линии — равнобокие гиперболы с общим центром и проходящие через него параболы,
- е) модифицированные полярные - 2; координатные линии — концентрические квадраты и лучи, проходящие через их центр.

Вывести формулы преобразования координат произвольной точки P в первом квадранте (см. рис. 23)

- а) декартовы (x_p, y_p) — полярные (r_p, α_p) ,
- б) декартовы (x_p, y_p) — «гиперболические» (r_p, α_p) ,
- в) декартовы (x_p, y_p) — модифицированные полярные - 1 (r_p, α_p) ,
- г) декартовы (x_p, y_p) — модифицированные «гиперболические» (r_p, α_p) ,
- д) декартовы (x_p, y_p) — модифицированные полярные - 2 (r_p, α_p) .

Решение.

- а) Уравнение прямой, проходящей через начало координат под углом α к оси x : $y = x \operatorname{tg} \alpha$,
уравнение окружности радиуса r с центром в начале координат: $y = \sqrt{r^2 - x^2}$. Решая систему из этих двух уравнений сначала относительно (x, y) , а затем — (r, α) , получим формулы преобразования:

$$\begin{aligned} x_p &= r_p \cos \alpha_p, & r_p &= \sqrt{x_p^2 + y_p^2}, \\ y_p &= r_p \sin \alpha_p, & \alpha_p &= \arctg \frac{y_p}{x_p}. \end{aligned}$$

- б) Аналогично вписываем систему уравнений прямой $y = x \operatorname{tg} \alpha$ и равнобокой гиперболы $y = \frac{r}{x}$, решаем относительно

(x, y) и (r, α) и получаем:

$$\begin{aligned} x_p &= \sqrt{\frac{r_p}{\operatorname{tg} \alpha_p}}, & r_p &= x_p y_p, \\ y_p &= \sqrt{r_p \operatorname{tg} \alpha_p}, & \alpha_p &= \operatorname{arctg} \frac{y_p}{x_p}. \end{aligned}$$

- в) Записываем систему уравнений окружности $y = \sqrt{r^2 - x^2}$ и параболы $y = \alpha x^2$, решаем относительно (x, y) и (r, α) и находим:

$$\begin{aligned} x &= \sqrt{\frac{\sqrt{1 + \alpha^2 r^2} - 1}{2\alpha^2}}, & r &= \sqrt{x^2 + y^2}, \\ y &= \frac{\sqrt{1 + \alpha^2 r^2} - 1}{2\alpha}, & \alpha &= \frac{y}{x^2}. \end{aligned}$$

- г) Решаем систему уравнений равнобокой гиперболы $y = \frac{r}{x}$ и параболы $y = \alpha x^2$ сначала относительно (x, y) , а затем — (r, α) , что дает

$$\begin{aligned} x_p &= \sqrt[3]{\frac{r_p}{\alpha_p}}, & r_p &= x_p y_p, \\ y_p &= \sqrt[3]{\alpha_p r_p^2}, & \alpha_p &= \frac{y_p}{x_p^2}. \end{aligned}$$

- д) Координаты (r, α) как функции (x, y) :

$$\begin{aligned} \alpha_p &= \operatorname{arctg} \frac{y_p}{x_p}, \\ r_p &= \max(x_p, y_p) \sqrt{2}. \end{aligned}$$

Координаты (x, y) как функции (r, α) :

Случай $\alpha_p < \frac{\pi}{4}$. Координата x_p не зависит от угла α_p и равна $x_p = \frac{r_p}{\sqrt{2}}$. Поскольку $\operatorname{tg} \alpha_p = \frac{y_p}{x_p}$, находим $y_p = \frac{r_p}{\sqrt{2}} \operatorname{tg} \alpha_p$.

Аналогично рассматриваем случай $\alpha_p > \frac{\pi}{4}$. Получаем $x_p = \frac{r_p}{\sqrt{2} \operatorname{tg} \alpha_p}$ и $y_p = \frac{r_p}{\sqrt{2}}$.

Собирая все вместе, получаем:

$$\begin{aligned}x_p &= \frac{r_p}{\sqrt{2}} \min(1, \frac{1}{\operatorname{tg} \alpha_p}), \\y_p &= \frac{r_p}{\sqrt{2}} \min(1, \operatorname{tg} \alpha_p).\end{aligned}$$

- 2) Покажите, что если взаимнооднозначный оператор L — линейен, то L^{-1} — также линейный оператор.

Решение. В соответствии с определением оператор L^{-1} линейен, если для любых чисел α и β и любых векторов f и g из области определения оператора выполняется равенство

$$L^{-1}(\alpha f + \beta g) = \alpha L^{-1}f + \beta L^{-1}g.$$

Допустим, что это равенство имеет место и приведем его к тождеству. Подействуем на обе части равенства оператором L . В левой части тогда получим вектор $\alpha f + \beta g$. В правой же части имеем $L(\alpha L^{-1}f + \beta L^{-1}g)$. В силу линейности L последнее выражение легко преобразуется

$$L(\alpha L^{-1}f + \beta L^{-1}g) = \alpha L(L^{-1}f) + \beta L(L^{-1}g) = \alpha f + \beta g.$$

Полученное тождество доказывает линейность обратного оператора.

- 3) Какие операторы из примеров на стр.27 линейны?
4) Покажите, что оператор из примера 1 стр.28 линейен.
5) Как выглядят матрицы единичного и нулевого операторов?

Решение. Рассматриваем векторное пространство с базисом $\{e_i\}$, тогда произвольный вектор $f = \sum_i f^i e_i$ и далее для $If = f$ имеем:

$$If = \begin{cases} I(\sum_j f^j e_j) = \sum_i f^i (Ie_i) = \sum_i f^i I_i = \sum_i f^i \sum_j I_i^j e_j = \sum_j e_j \sum_i f^i I_i^j, \\ f = \sum_j f^j e_j. \end{cases}$$

Отсюда

$$f^j = \sum_i f^i I_i^j \quad \Rightarrow \quad I_i^j = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

- 6) Пусть (f^1, f^2) — компоненты вектора f из двумерного векторного пространства относительно базиса (e_1, e_2) . Найти его компоненты относительно базиса (e_1, e_2) , если $e_1 = e_2$ и $e_2 = e_1$.
- 7) $y = Ax = x(x^2 - 1)$, $R(A) = ?$ Ограничить область определения оператора так, чтобы он стал взаимнооднозначным.
- 8) Показать, что оператор A в системе линейных алгебраических уравнений 1-й степени $Ax = b$ линеен.

Упражнения 3

- 1) Выведите неравенства

$$\begin{aligned} \left| \|f\| - \|g\| \right| &\leq \|f + g\|, \\ \left| \|f\| - \|g\| \right| &\leq \|f - g\|. \end{aligned}$$

Решение. Рассмотрим неравенство треугольника

$$\|f + g\| \leq \|f\| + \|g\|.$$

Заменой элемента g на $(h - f)$ и последующим переобозначением

- 2) Наглядное представление о нормах $\|x\|_1, \|x\|_2, \dots, \|x\|_\infty$ дает множество элементов $x \in \mathbb{R}^n$, для которых $\|x\| = 1$, или так называемая *единичная сфера*. Постройте ее для плоского случая, т.е. для $x \in \mathbb{R}^2$.

Решение. см. рис.24.

- 3) Покажите, что величина $\|f\| = \max_{x \in [a, b]} |f(x)|$ (это формула (3.10)) действительно является нормой.
- 4) Покажите, что нормы $\|x\|_1 = \sum_i |x^i|$ и $\|x\|_\infty = \max_i |x^i|$, $x \in \mathbb{R}^n$ эквивалентны.

Ответ:

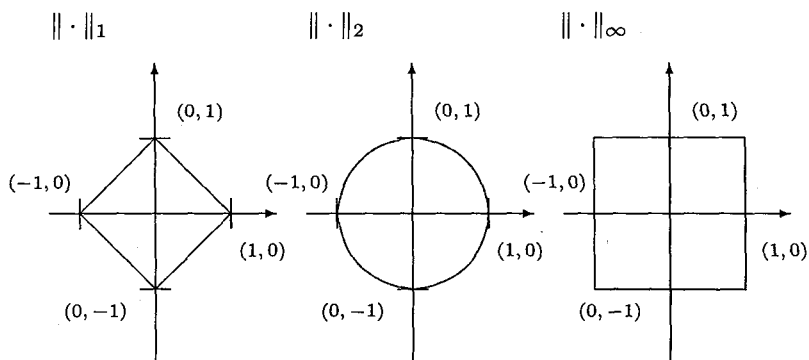


Рис. 24. Вид единичной сферы на $\mathbb{R}^2$ для разных норм.

Рассмотрим неравенство $c_1 \|x\|_1 \leq \|x\|_\infty \leq c_2 \|x\|_1$ и попробуем найти положительные константы c_1 и c_2 , при которых это неравенство выполняется. Подставив выражения для норм, получим:

$$c_1 \sum_i |x^i| \leq \max_i |x^i| \leq c_2 \sum_i |x^i|. \quad (15.14)$$

Если обозначить $x_{\min} \equiv \min_i |x^i|$, $x^i \neq 0$ и $x_{\max} \equiv \max_i |x^i|$ и учесть, что

$$nx_{\min} \leq \sum_i |x^i| \leq nx_{\max},$$

то неравенства (15.14) можно записать так:

$$c_1 nx_{\min} \leq x_{\max} \leq c_2 nx_{\max}.$$

Отсюда находим $c_1 \leq \frac{x_{\max}}{nx_{\min}}$ и $c_2 \geq \frac{1}{n}$

- 5) Покажите, что равномерная норма $\|f\| = \max_{x \in [a, b]} |f(x)|$ и норма $\|f\|_1 = \int_a^b |f(x)| dx$ в пространстве непрерывных функций неэквивалентны.

Решение.

По теореме о среднем норму $\|f\|_1$ можно записать так:

$$\int_a^b |f(x)| dx = (b-a) |f(\xi)|, \quad \xi \in (a, b).$$

Далее из определения эквивалентных норм имеем

$$c_1 \|f\| \leq \|f\|_1 \leq c_2 \|f\| \Rightarrow c_1 \leq (b-a) \frac{|f(\xi)|}{\max_x |f(x)|} \leq c_2.$$

Отсюда $c_2 \geq (b-a)$, но $c_1 = 0$, так как $\frac{|f(\xi)|}{\max_x |f(x)|}$ — любое, сколь угодно малое положительное число.

- 6) Покажите, что произведение линейных операторов есть линейный оператор.
- 7) Проверьте, удовлетворяет ли $\|f - g\|$ требованиям, предъявляемым к метрике $d(f, g)$.

Решение. Обсудим неравенство треугольника (остальное — тривиально). Покажем, что из неравенства треугольника $\|f + g\| \leq \|f\| + \|g\|$ (см. определение нормы) следует неравенство $\|f - g\| \leq \|f - h\| + \|h - g\|$, которое можно записать в терминах метрики $d(f, g) \leq d(f, h) + d(h, g)$.

Всегда можно подобрать векторы $\hat{f}$, $\hat{g}$ и h такие, что $f = \hat{f} - h$ и $g = h - \hat{g}$. Подставляя эти выражения в неравенство треугольника, получим $\|\hat{f} - \hat{g}\| \leq \|\hat{f} - h\| + \|h - \hat{g}\|$. То же неравенство можно записать используя метрику $d(\hat{f}, \hat{g}) \leq d(\hat{f}, h) + d(h, \hat{g})$. Оно выполняется для любой пары векторов $\hat{f}$ и $\hat{g}$, ч.т.д.

Упражнения 4

- 1) Покажите, что $\mathcal{L}(V)$ — векторное пространство.
- 2) Проведите для нормы $\|\cdot\|_1$: $\|f\|_1 = \sum_i |f^i|$ рассуждения, аналогичные приведенным на стр.43.
- 3) Покажите, что для любого линейного ограниченного оператора предел $r = \lim_{n \rightarrow \infty} \sqrt[n]{\|M^n\|}$ существует.

- 4) Покажите, что спектральный радиус оператора не зависит от выбора нормы из множества эквивалентных?

Решение. Требуется установить, верно ли равенство

$$\lim_{n \rightarrow \infty} \sqrt[n]{\|M^n\|_a} = \lim_{n \rightarrow \infty} \sqrt[n]{\|M^n\|_b}?$$

Запишем цепочку соотношений, где $r_a \equiv \lim_{n \rightarrow \infty} \sqrt[n]{\|M^n\|_a}$ и нормы эквивалентны, т.е. существует константа $c > 0$ такая, что $\|M^n\|_b \geq c\|M^n\|_a$:

$$\begin{aligned} \lim_{n \rightarrow \infty} \left(\sqrt[n]{\|M^n\|_a} - \sqrt[n]{\|M^n\|_b} \right) &= \lim_{n \rightarrow \infty} \sqrt[n]{\|M^n\|_a} \left(1 - \frac{\sqrt[n]{\|M^n\|_b}}{\sqrt[n]{\|M^n\|_a}} \right) = \\ &= r_a \left(1 - \lim_{n \rightarrow \infty} \sqrt[n]{\frac{\|M^n\|_b}{\|M^n\|_a}} \right) \leq \\ &\leq r_a \left(1 - \lim_{n \rightarrow \infty} \sqrt[n]{\frac{c\|M^n\|_a}{\|M^n\|_a}} \right) = \\ &= r_a (1 - \lim_{n \rightarrow \infty} \sqrt[n]{c}) = 0. \end{aligned}$$

Что и требовалось доказать.

- 5) Покажите, что множество собственных векторов для некоторого λ образует векторное пространство. Такое векторное пространство является подмножеством пространства B . Следовательно, собственные векторы, отвечающие некоторому собственному значению λ , образуют векторное подпространство в B . Оно называется *собственным подпространством*, отвечающим λ .
- 6) Покажите, что $\|L\|$ — норма.
- 7) Показать, что аксиомы нормы (кроме неравенства треугольника) выполняются для функции (f, f) .
- 8) Определите, сколько надо сделать итераций n в методе последовательных приближений, чтобы $\|f_n - f\| < \delta$, если $\delta > 0$ задано?
- 9) Получите условия сходимости метода Якоби, пользуясь нормой $\|\cdot\|_1$.

Упражнения 5

- 1) Почему при определении скалярного произведения не используется $\sin$?
- 2) Что произойдет, если описанный процесс ортогонализации применить к множеству линейно зависимых векторов?
- 3) Рассмотрите последовательность $\{f_i\}$:

$$1, t, t^2, t(t-1), t^3, t^4, \dots$$

в пространстве функций с нормой

$$\|f\| = \sqrt{\int_a^b f(t)^2 dt}, \quad t \in (a, b).$$

С помощью процесса Грама-Шмидта получите первые пять членов ортонормированной последовательности $\{e_i\}$. Постройте графики для членов обеих последовательностей и сравните их. Ортогонализацию провести для двух областей определения функций:

- а) $a = -1, b = 1$,
- б) $a = 0, b = 1$.

Решение. см. рис.25 и рис.26 .

- 4) Выполните ортогонализацию последовательности $\{f_i\}$:

$$1, \sin \varphi, \cos \varphi, \sin 2\varphi, \cos 2\varphi, \dots$$

в пространстве функций с нормой

$$\|f\| = \sqrt{\int_a^b f(x)^2 dx}, \quad x \in (a, b).$$

Постройте графики для первых пяти членов обеих последовательностей и сравните их. Ортогонализацию провести для двух областей определения функций

- а) $a = -\frac{\pi}{2}, b = \frac{\pi}{2}$,
- б) $a = 0, b = \frac{\pi}{2}$.

Решение. см. рис.27 и рис.28 .

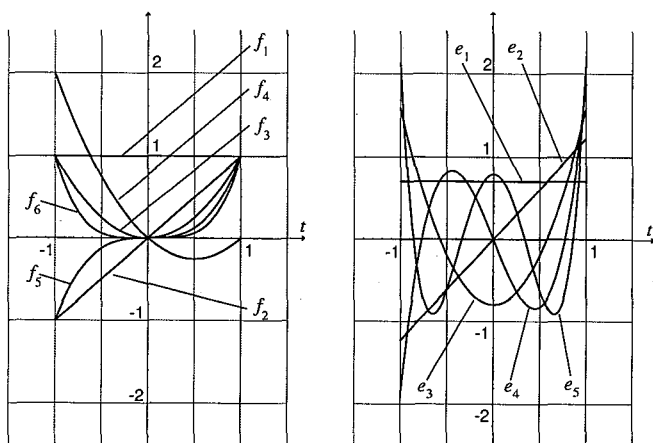


Рис. 25. Ортогонализация степенного ряда на отрезке $[-1, 1]$.

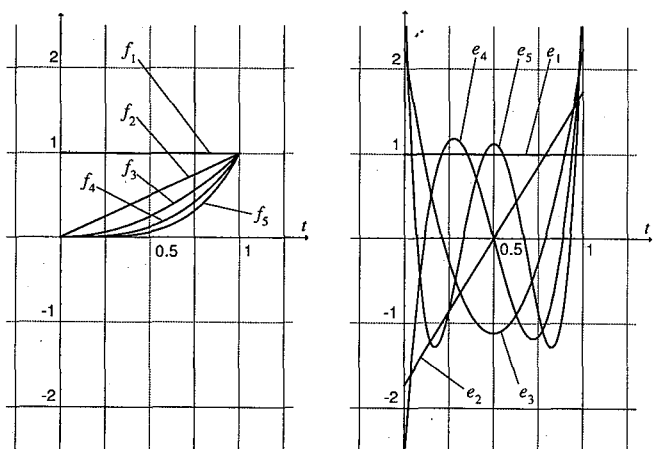


Рис. 26. Ортогонализация степенного ряда на отрезке $[0, 1]$.

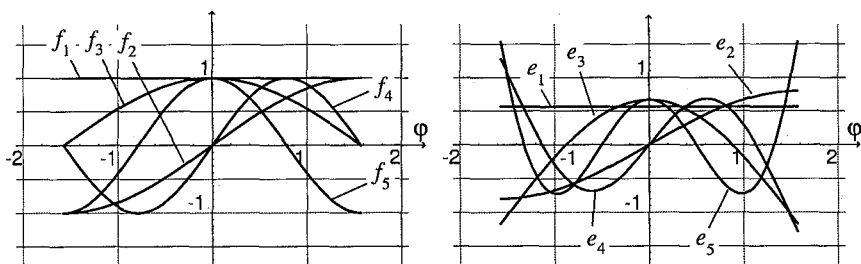


Рис. 27. Ортогонализация тригонометрического ряда на отрезке $[-\pi, \pi]$.

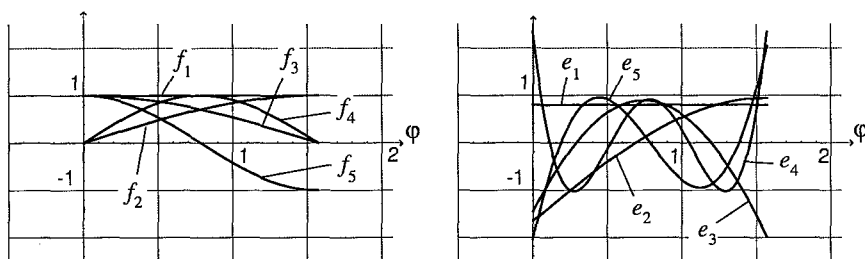


Рис. 28. Ортогонализация тригонометрического ряда на отрезке $[0, \pi]$.

Практические занятия к части II

Упражнения 1

- 1) Покажите, что указанные свойства выполняются
 - а) $L_{\vec{v}}(fg) = fL_{\vec{v}}g + gL_{\vec{v}}f$,
 - б) $L_{\vec{u}+\vec{v}} = L_{\vec{u}} + L_{\vec{v}}$,
 - в) $L_{f\vec{v}} = fL_{\vec{v}}$. Здесь f, g — дифференцируемые функции.
- 2) Покажите, что производная $L_{\vec{w}} = L_{\vec{u}}L_{\vec{v}} - L_{\vec{v}}L_{\vec{u}}$ является оператором первого порядка.

Упражнения 2

- 1) Постройте характеристики для одномерного уравнения переноса (6.12), покажите, что характеристики образуют семейство прямых с одинаковым наклоном (рис.11 а) и найдите решение задачи Коши (6.10), (6.11).
- 2) По достаточно быстро убывающему начальному профилю скорости постройте характеристики и с их помощью ряд профилей скорости вплоть до момента времени наступления градиентной катастрофы.
- 3) Приведите уравнение движения для u компоненты

$$\partial_t \rho + u \partial_x \rho + \rho \partial_x u = 0$$

к виду

$$\partial_t p + u \partial_x p + \rho c^2 \partial_x u = 0,$$

воспользовавшись уравнением состояния $\rho = \rho(p, s)$, записанного в терминах плотности массы, давления и энтропии. Примите во внимание, что производная сложной функции ρ по некоторой переменной α дает

$$\partial_\alpha \rho = \partial_p \rho \partial_\alpha p + \partial_s \rho \partial_\alpha s = \frac{1}{c^2} \partial_\alpha p + (\partial_s \rho) \partial_\alpha s.$$

Упражнения 3

1) Одномерная динамика идеальной жидкости.

Течение идеальной жидкости со скоростью u , давлением p , плотностью ρ и энтропией s описывается в декартовых координатах следующей системой дифференциальных уравнений:

$$\begin{aligned}\partial_t \rho + u \partial_x \rho + \rho \partial_x u &= 0, \\ \partial_t u + u \partial_x u + \frac{1}{\rho} \partial_x p &= 0, \\ \partial_t s + u \partial_x s &= 0.\end{aligned}\tag{15.15}$$

Для замыкания этой системы трех уравнений относительно четырех неизвестных функций используется уравнение состояния, связывающее значения плотности, давления и энтропии:

$$\rho = \rho(p, s).\tag{15.16}$$

Пользуясь уравнением состояния (15.16) и уравнением баланса энтропии (15.15₃), исключить из системы производные от плотности (в уравнении неразрывности (15.15₁)), записать систему в матричной форме, исследовать тип системы, построить эквивалентную систему независимых друг от друга дифференциальных уравнений и найти ее характеристики.

Решение.

- а) *Исключение ρ .* С помощью уравнения состояния (15.16) коэффициенты системы, содержащие ρ , всегда могут быть записаны в терминах p и s . Для замыкания системы необходимо выразить также и производные от ρ через производные от p и s . Это можно сделать, продифференцировав (15.16) по некоторой переменной a . Получим

$$\partial_a \rho = \partial_p \rho \partial_a p + \partial_s \rho \partial_a s.$$

В нашем случае переменная a принимает значения t и x . Подставляя соответствующие выражения в уравнение неразрывности, вводя обозначение $\partial_p p \equiv c^2$ (c — скорость звука в

среде) и приводя подобные члены, найдем

$$\begin{aligned} 0 &= (\partial_p \rho \partial_t p + \partial_s \rho \partial_t s) + u (\partial_p \rho \partial_x p + \partial_s \rho \partial_x s) + \rho \partial_x u = \\ &= \partial_s \rho (\partial_t s + u \partial_x s) + \frac{1}{c^2} (\partial_t p + u \partial_x p) + \rho \partial_x u. \end{aligned}$$

Поскольку первое слагаемое в силу уравнения баланса энтропии равно нулю, окончательно уравнение неразрывности можно записать в следующей эквивалентной форме:

$$\partial_t p + u \partial_x p + \rho c^2 \partial_x u = 0. \quad (15.17)$$

- б) *Матричная форма.* Упорядочим неизвестные функции, например, так (p, u, s) . Будем их интерпретировать как компоненты вектора f и запишем полученную систему в матричной форме:

$$\partial_t f + A \partial_x f = 0. \quad (15.18)$$

Чтобы найти вид матрицы A учтем, что $\partial_x f = (\partial_x p, \partial_x u, \partial_x s)$ и произведение $A \partial_x f$ равно

$$A \partial_x f = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} \partial_x p \\ \partial_x u \\ \partial_x s \end{pmatrix} = \begin{pmatrix} u \partial_x p + \rho c^2 \partial_x u \\ u \partial_x u + \frac{1}{\rho} \partial_x p \\ u \partial_x s \end{pmatrix}.$$

Отсюда получаем

$$A = \begin{pmatrix} u & \rho c^2 & 0 \\ \frac{1}{\rho} & u & 0 \\ 0 & 0 & u \end{pmatrix},$$

- в) *Исследование типа системы.* Определим тип системы (15.18), для чего решим следующую задачу на собственные значения:

$$A \xi = \nu \xi \quad \text{или} \quad (A - \nu I) \xi = 0.$$

Эта однородная система линейных алгебраических уравнений имеет нетривиальное решение в том случае, когда ее определитель равен нулю

$$\det(A - \nu I) = ((u - \nu)^2 - c^2) (u - \nu) = 0,$$

откуда

$$\nu_1 = u - c, \quad \nu_2 = u + c, \quad \nu_3 = u.$$

Поскольку все собственные числа вещественные, различные и отличны от нуля, система (15.18) является гиперболической.

- г) *Приведение к эквивалентной системе с диагональной матрицей.* Преобразованием подобия $\Lambda^{-1}A\Lambda = D$, где $\Lambda \equiv (\xi_1|\xi_2|\xi_3)$, матрица A может быть приведена к диагональной матрице D . Собственные векторы $\xi_i \equiv (\xi_i^1, \xi_i^2, \xi_i^3)$ матрицы A являются решениями следующих однородных систем уравнений:

$$\begin{cases} c\xi_1^1 + \rho c^2\xi_1^2 = 0, \\ \frac{1}{\rho}\xi_1^1 + c\xi_1^2 = 0, \\ c\xi_1^3 = 0, \end{cases} \quad \begin{cases} -c\xi_2^1 + \rho c^2\xi_2^2 = 0, \\ \frac{1}{\rho}\xi_2^1 - c\xi_2^2 = 0, \\ c\xi_2^3 = 0, \end{cases} \quad \begin{cases} \rho c^2\xi_3^2 = 0, \\ \frac{1}{\rho}\xi_3^1 = 0, \\ 0 \cdot \xi_3^3 = 0, \end{cases}$$

которые сводятся соответственно к недоопределенным системам:

$$\begin{cases} \xi_1^1 = -\rho c\xi_1^2, \\ \xi_1^3 = 0, \end{cases} \quad \begin{cases} \xi_2^1 = \rho c\xi_2^2, \\ \xi_2^3 = 0, \end{cases} \quad \begin{cases} \xi_3^1 = 0, \\ \xi_3^2 = 0, \\ 0 \cdot \xi_3^3 = 0. \end{cases}$$

Выбирая $\xi_1^1 = \rho$, $\xi_2^1 = \rho$, $\xi_3^3 = 1$, получим три собственных вектора матрицы A :

$$\xi_1 = \begin{pmatrix} \rho \\ -\frac{1}{c} \\ 0 \end{pmatrix}, \quad \xi_2 = \begin{pmatrix} \rho \\ \frac{1}{c} \\ 0 \end{pmatrix}, \quad \xi_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Возможен любой выбор, не дающий нулевого вектора. Получим другие собственные векторы, имеющие те же направления. Матрицу Λ строим в виде $\Lambda = (\xi_1|\xi_2|\xi_3)$:

$$\Lambda = \begin{pmatrix} \rho & \rho & 0 \\ -\frac{1}{c} & \frac{1}{c} & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Обратная матрица вычисляется элементарно и равна

$$\Lambda^{-1} = \begin{pmatrix} \frac{1}{2\rho} & -\frac{c}{2} & 0 \\ \frac{1}{2\rho} & \frac{c}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Вычисление правой части h' , согласно выражению (7.6), дает

$$h' = -\frac{p}{2\rho^2} \begin{pmatrix} \partial_t \rho + (u - c) \partial_x \rho \\ \partial_t \rho + (u + c) \partial_x \rho \\ 0 \end{pmatrix}.$$

В итоге система (15.18) преобразуется в систему с диагональной матрицей относительно вектора неизвестных $f' = \Lambda^{-1}f$:

$$\partial_t f' + \Lambda^{-1} A \Lambda \partial_x f' = h', \quad f' = \begin{pmatrix} \frac{1}{2} \left(\frac{p}{\rho} - cu \right) \\ \frac{1}{2} \left(\frac{p}{\rho} + cu \right) \\ s \end{pmatrix} \quad (15.19)$$

или

$$\begin{aligned} \partial_t \left(\frac{p}{\rho} - cu \right) + (u - c) \partial_x \left(\frac{p}{\rho} - cu \right) &= -\frac{p}{\rho^2} (\partial_t \rho + (u - c) \partial_x \rho), \\ \partial_t \left(\frac{p}{\rho} + cu \right) + (u + c) \partial_x \left(\frac{p}{\rho} + cu \right) &= -\frac{p}{\rho^2} (\partial_t \rho + (u + c) \partial_x \rho), \\ \partial_x s + u \partial_x s &= 0. \end{aligned}$$

Система (15.19) имеет три характеристики, определяемые уравнениями:

$$\begin{aligned} d_t x &= u \pm c \Rightarrow x = x_0 + (u \pm c)t, \\ d_t x &= u \Rightarrow x = x_0 + ut. \end{aligned}$$

Первые две характеристики называются *звуковыми*. Третья характеристика называется *энтропийной*.

2) Одномерное уравнение теплопроводности

$$\partial_t f = \kappa^2 \partial_{xx} f, \quad (x, t) \in \Omega = \{t > 0, x \in [a, b]\}. \quad (15.20)$$

является простейшим уравнением параболического типа. Покажите это.

Решение. Для того чтобы убедиться в том, что уравнение (15.20) действительно имеет параболический тип, представим его в виде системы уравнений. Обозначая $\kappa \partial_x f \equiv g$, получим

$$\begin{cases} \partial_t f - \kappa \partial_x g = 0, \\ \kappa \partial_x f = g, \end{cases} \quad (15.21)$$

или в матричной форме

$$\begin{aligned} A_1 \partial_t f + A_2 \partial_x f &= g, \\ A_1 &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & -\kappa \\ \kappa & 0 \end{pmatrix}, \\ f &= \begin{pmatrix} f \\ g \end{pmatrix}, \quad g = \begin{pmatrix} 0 \\ g \end{pmatrix}. \end{aligned}$$

Поскольку из двух матриц неособой является матрица A_2 , то, умножая, обе части уравнения на A_2^{-1} , приведем его к виду (7.4):

$$A \partial_t f + \partial_x f = h, \quad A = \begin{pmatrix} 0 & 0 \\ -\frac{1}{\kappa} & 0 \end{pmatrix}, \quad h = \begin{pmatrix} \frac{g}{\kappa} \\ 0 \end{pmatrix}.$$

Решая характеристическое уравнение для матрицы A , найдем, что оба собственных числа равны нулю. Таким образом, система (15.21) и уравнение (15.20) параболического типа, и матрица A не может быть приведена к диагональной форме. Характеристиками, как следует из уравнения (7.8), являются прямые $t = \text{const}$, параллельные оси x (Почему? Обратите внимание на параметризацию характеристик).

3) Двумерное уравнение Лапласа

$$\partial_{xx} f + \partial_{yy} f = 0 \quad (x, y) \in \Omega^2 \quad (15.22)$$

служит простейшим примером уравнения эллиптического типа. Покажите это.

Практические занятия к части III

Упражнения 1

- 1) Выведите разностную формулу для смешанной производной $\partial_{xy}f$ и оцените порядок отбрасываемых членов.
- 2) Получите разностную формулу для второй производной в случае неравномерной сетки.

Упражнения 2

- 1) Постройте разностную аппроксимацию краевой задачи для уравнения теплопроводности (10.8), (10.9) в случае, когда все узлы сетки расположены в центрах боксов. Остальные условия те же, что в примере, разобранным на лекции.

Решение. Поскольку во внутренних точках аппроксимация прежняя, рассматриваем лишь приграничные точки. Граничные условия

$$(\kappa \partial_x \vartheta - a \vartheta)_{x=0} = b \Rightarrow F_{\frac{1}{2}} - a \vartheta_{\frac{1}{2}} = b, \quad \vartheta|_{x=L} = \vartheta_L,$$

используются для построения аппроксимации потоков $F_{\frac{1}{2}}$ и $F_{n+\frac{1}{2}}$.

- а) 1-е граничное условие. Ищем поток $F_{\frac{1}{2}}$.

$$\frac{\vartheta_{\frac{1}{2}} - \vartheta_1}{\vartheta_2 - \vartheta_1} = \frac{0 - \frac{\Delta x}{2}}{\frac{3}{2}\Delta x - \frac{\Delta x}{2}} = -\frac{1}{2} \Rightarrow \vartheta_{\frac{1}{2}} = \vartheta_1 - \frac{(\vartheta_2 - \vartheta_1)}{2} = \frac{3}{2}\vartheta_1 - \frac{1}{2}\vartheta_2.$$

Отсюда имеем

$$F_{\frac{1}{2}} = \frac{a}{2}(3\vartheta_1 - \vartheta_2) + b.$$

- б) 2-е граничное условие. Ищем поток $F_{n+\frac{1}{2}}$.

$$\int_{x_n}^{x_{n+\frac{1}{2}}} \frac{F}{\kappa} dx = \vartheta_{n+\frac{1}{2}} - \vartheta_n = \vartheta_L - \vartheta_n.$$

Интеграл аппроксимируем формулой $F_{n+\frac{1}{2}} \frac{\Delta x}{2\kappa_{n+\frac{1}{2}}}$ и в итоге получаем выражение для потока

$$F_{n+\frac{1}{2}} = 2\kappa_{n+\frac{1}{2}} \frac{\vartheta_L - \vartheta_n}{\Delta x}.$$

Упражнения 3

- 1) Найдите порядок точности разностных аппроксимаций уравнения теплопроводности: трехслойная схема «чехарда», двуслойная схема Кранка-Николсон, трехслойная схема Адамса-Башфорта.
Указание. При определении порядка точности разностных схем мы действуем в порядке, обратном тому, которого придерживались при их построении. Если при построении разностной схемы мы используем нелокальную аппроксимацию (конечные разности) локальных величин (производные), то при оценке точности, напротив, мы пользуемся рядами Тейлора для описания поведения функций в точках шаблона с помощью локальных величин.

Упражнения 4

- 1) Исследуйте устойчивость схемы ВВЦП (Вперед по Времени и Центральная по Пространству) для уравнения переноса.
- 2) Покажите, что условие устойчивости центрально-разностной схемы для одномерного волнового уравнения имеет вид

$$\mu^2 \leq 1, \quad \text{или} \quad \left| \frac{u \Delta t}{\Delta x} \right| \leq 1. \quad (15.23)$$

- 3) Найдите условие устойчивости явной центрально-разностной схемы для двумерного уравнения теплопроводности.
- 4) Исследуйте устойчивость двуслойной разностной схемы с симметричными разностями по пространству для уравнения переноса с диффузией.

Упражнения 5

- 1) Рассмотрим метод последовательной релаксации. Какой метод мы получим, если $\tau = 1$?
- 2) Выведите расчетные формулы, связывающие компоненты вектора f^{k+1} с компонентами f^k , B и h , для рассмотренных четырех

итерационных методов. Например, для метода Гаусса–Зейделя должна получиться формула:

$$f_i^{k+1} = \frac{1}{a_{ii}} \left(- \sum_{j=1}^{i-1} a_{ij} f_j^{k+1} - \sum_{j=i+1}^n a_{ij} f_j^k + g_i \right).$$

- 3) Выведите формулы метода прогонки для случая пятидиагональной матрицы; для случая ленточной матрицы с шириной ленты 2ℓ .

Предметный указатель

- амплитуда моды, 96
- вектор, 21
 - касательный, 62
- векторное поле, 63
 - характеристическое, 66
- волна Фурье, 96
- волновое число, 96
- временной слой, 113
- вычислительная дисперсия, 139
- гармоника, 96
- дивергентность, 107
- дискретное преобразование Фурье, 94
- задача
 - Коши, 67, 86
 - краевая, 86
 - корректная, 87
 - на собственные значения, 51
 - некорректная, 87
 - нестационарная, 78
 - разностная, 118
 - корректная, 122
 - с начальными условиями, 86
 - спектральная, 51
 - стационарная, 78
 - эволюционная, 86
- интерполяция, 119
- итерация, 154
- канонический вид разностной схемы, 114
- композиция отображений, 16
- коэффициенты Фурье, 57
- матрица
 - оператора, 29
 - ортгональная, 93
- метрика, 36
- метод
 - Гаусса, 158
 - Гаусса-Зейделя, 157
 - исследования устойчивости Неймана, 126
 - прямой, 126
 - спектральный, 126
 - последовательной релаксации, 157
 - последовательных приближений, 33, 155
 - прогонки, 160
 - простой итерации, 156
 - циклической прогонки, 161
 - Якоби, 48, 155
- методы
 - разностные, 90
 - итерационные, 154
 - точные, 154
- множество значений, 16
- множитель перехода, 131
- мода, 96
 - вычислительная, 151
 - физическая, 151
- модели
 - качественные, 10
 - количественные, 11
 - математические, 11
 - численные, 12
- неравенство
 - Бесселя, 59
 - Коши-Буняковского, 54
- норма
 - равномерная, 39
 - вектора, 37
 - оператора, 43
 - сеточная, 120
 - эквивалентная, 42
- нуль, 20
- область
 - влияния, 71
 - зависимости, 138
 - определения отображения, 16
- образ
 - множества, 16
 - элемента, 16
- обратный ход (подстановка), 158
- оператор
 - дифференциальный однородный, 65
 - дифференцирования по направлению вектора, 65
 - источника, 114
 - неограниченный, 43
 - ограниченный, 43
 - перехода (шага), 114
 - разностный, 119
 - сжатия, 48

- ортогонализация, 55
- ортогональные векторы, 54
- отображение, 16
 - взаимно однозначное, 16
 - единичное, 16
 - обратное, 17
- ошибки
 - амплитудная, 139
 - фазовая, 139
- параметр кривой, 61
- параметризованная кривая, 61
- первое дифференциальное приближение, 142
- первый интеграл, 66
- погрешность
 - аппроксимации на решении, 122
 - разностной схемы, 120
 - решения, 119
- подпространство, 23
- порядок
 - аппроксимации, 122
 - дифференциального уравнения, 61
 - точности, 121
- последовательность
 - Коши, 41
 - сходящаяся, 41
 - фундаментальная, 41
- предел последовательности, 41
- приближение, 154
- приближенное решение, 116
- произведение
 - оператора на число, 31
 - операторов, 34
 - прямое, 22
 - скалярное, 52
- производная по направлению вектора, 65
- прообраз
 - множества, 17
- пространство
 - банахово, 42
 - векторное, 21
 - гильбертово, 54
 - евклидово, 52
 - касательное, 63
 - линейное, 21
 - метрическое, 36
 - нормированное векторное, 37
- противоположный элемент, 20
- процесс Грама-Шмидта, 55
- прямой ход (исключение), 158
- разности
 - вперед/назад, 97
 - направленные, 97
 - симметричные, 98
 - центральные, 98
- разностная схема
 - Адамса-Башфорта, 115
 - двуслойная, 113
 - Кранка-Николсон, 114
 - неявная, 114
 - трехслойная, 114
 - чехарда, 115
 - Эйлера, 114
 - явная, 114
- ряд Фурье, 57
- свойство наилучшего приближения, 57
- сетка, 90
 - равномерная, 91
- смешанное краевое условие, 69
- собственная функция, 51
- собственное значение, 51
- собственное число, 76
- собственный вектор, 51, 76
- спектр
 - дискретный, 51
 - точный, 51
- спектральный радиус, 50
- степень оператора, 34
- сужение оператора, 116
- сумма операторов, 31
- схема
 - консервативная, 107
 - разностная, 118
- схемная вязкость, 139
- сходимость
 - приближенных решений, 120
 - решений разностной задачи, 120
- точка, 21
- узел, 90
- уравнение
 - в частных производных, 61
 - волновое, 79
 - дифференциальное, 61, 65
 - квазилинейное, 70
 - неоднородное, 69
 - модифицированное, 143
 - переноса, 68
 - характеристик, 66
 - характеристическое, 76
- условие
 - граничное вычислительное, 150
 - Куранта-Фридриха-Леви, 137

- краевое, 69
- Неймана спектральное, 130
- начальное, 67, 69, 86
 - вычислительное, 151
- устойчивость
 - абсолютная, 127
 - по начальным данным, 127
 - условная, 127
- фазовая кривая, 64
- факторизация, 158
- формула Даламбера, 85
- функция сеточная, 91
- характеристика, 66
- число Куранта, 137
- шаблон, 97
- шаг сетки, 91

Учебное издание

Михаил Юрьевич Белевич

**Математические методы решения
океанологических задач**

Основные вычислительные идеи и методы

Учебное пособие

Редактор Л.В. Ковель

Компьютерная верстка Н.И. Афанасьевой

ЛР № 020309 от 30.12.96.

Подписано в печать 30.10.08. Формат 60 × 90 1/16. Гарнитура Times New Roman.
Бумага офсетная. Печать офсетная. Усл.-печ. л. 12,1. Тираж 300 экз. Заказ № 51/08.
РГГМУ, 195196, Санкт-Петербург, Малоохтинский пр., 98.
ЗАО «НПП «Система», 195112, Санкт-Петербург, Малоохтинский пр., 80/2.

107-80