

Э. Э. Шноль

7

СЕМЬ ЛЕКЦИЙ

по

ВЫЧИСЛИТЕЛЬНОЙ  
МАТЕМАТИКЕ



**Э. Э. Шноль**

**СЕМЬ ЛЕКЦИЙ  
ПО ВЫЧИСЛИТЕЛЬНОЙ  
МАТЕМАТИКЕ**

**Издание третье, дополненное**



**URSS**  
**МОСКВА**

**Шноль Эммануил Эльевич**

**Семь лекций по вычислительной математике.** Изд. 3-е, доп. — М.: Издательство ЛКИ, 2008. — 112 с.

Предмет этих лекций — первоначальное знакомство с вычислительной математикой. Они рассчитаны на широкий круг читателей, не специализирующихся в этой области.

Студентам, обучающимся по специальностям «Математика», «Прикладная математика», «Физика», и лицам, преподающим вычислительную математику, могут быть интересны отдельные замечания, сделанные в лекциях 1–3 и 7.

Издательство ЛКИ. 117312, г. Москва, пр-т Шестидесятилетия Октября, д. 9.  
Формат 60×90/16. Печ. л. 7. Зак. № 1209.

Отпечатано в ООО «ЛЕНАНД».

117312, г. Москва, пр-т Шестидесятилетия Октября, д. 11А, стр. 11.

**ISBN 978–5–382–00378–8**

© Э. Э. Шноль, 1992, 2007

© Издательство ЛКИ, 2007

НАУЧНАЯ И УЧЕБНАЯ ЛИТЕРАТУРА	
	E-mail: <a href="mailto:URSS@URSS.ru">URSS@URSS.ru</a>
	Каталог изданий в Интернете:
	<a href="http://URSS.ru">http://URSS.ru</a>
	Тел./факс: 7 (495) 135–42–16
URSS	Тел./факс: 7 (495) 135–42–46

5399 ID 62480



## ПРЕДИСЛОВИЕ К ТРЕТЬЕМУ ИЗДАНИЮ

Эти вводные лекции были прочитаны почти 40 лет назад. Фактическое их содержание невелико и могло быть изложено на вдвое меньшем числе страниц. Я думаю, однако, при знакомстве с любой научной дисциплиной эмоциональная сторона предмета не менее важна, чем фактическая.

В серьезных курсах вычислительной математики пропорции меняются.

Но и там обсуждение основных идей и рассмотрение простых примеров часто играют большую роль, чем общие теоремы.

За прошедшие десятилетия электронные вычислительные машины (компьютеры) чрезвычайно развились, и общение человека с ними стало совсем другим. Вычислительная математика за этот громадный срок тоже сильно изменилась.

Однако стиль работы людей, решающих трудные научные или технические задачи, остался в основном таким, каким он был в 70-е годы (об этом немного сказано в первой лекции).

Работа протекает примерно так. Для изучаемой проблемы выбирается ее математическая модель, и производятся расчеты — ставится численный эксперимент. Сопоставляются данные этих экспериментов с наблюдениями (данными опытов) и с теорией, если она разработана. Подправляется модель, снова производятся расчеты и т. д. При этом может возникать много «внутренних», вычислительных вопросов, с которыми нужно разбираться. Например, перестает работать какой-нибудь алгоритм, реализованный в стандартной программе или программном комплексе; или же он работает, но очень медленно. В каком-нибудь упрощенном варианте, где ответ известен, получается плохая

точность и т. п. Наконец, если вычислительный эксперимент дает неожиданные интересные результаты, нужно убедиться, что они надежны и не являются артефактом. Разумеется, элементарных сведений, содержащихся в этих лекциях, недостаточно, чтобы разбираться с подобными вопросами.

Цель их гораздо скромнее: объяснить, что не всегда можно полагаться на готовые компьютерные программы, и бывает нужно знать, какая вычислительная математика в них заложена.

Э. Э. Шноль, июнь 2007 г.

## ПРЕДИСЛОВИЕ

В 1970—71 годах в Пущине и Москве проводилась подготовительная работа по созданию специализированного вычислительного центра «для нужд биологов». Базой для его развертывания могла служить лаборатория математического моделирования Института биологической физики, которой руководил А. М. Молчанов. Ему и было предложено организовать вычислительный центр.

Я был приглашен А. М. Молчановым включиться в «математическое просвещение» сотрудников Пушкинского научного центра биологических исследований. При поддержке тогдашнего Совета Научного центра и директора ИБФ Г. М. Франка был организован цикл лекций об элементах вычислительной математики. Аудитория состояла преимущественно из молодых людей, имеющих самые разные вузовские дипломы — физиков, химиков, биологов. Этот состав в основном определил стиль и содержание этого (вводного) курса. Просмотрев сейчас текст «свежим взглядом» спустя много лет, я, конечно, увидел в нем много недостатков. Один из крупных — полное отсутствие упоминания о задачах минимизации.

В нескольких местах текста указывается, что тот или иной вопрос будет подробно разобран позже, в более систематическом курсе. Эти намерения мне удалось осуществить лишь в небольшой мере. Но потребность в таком, более продвинутом курсе для естествоиспытателей, на мой взгляд, сохраняется. Его бы следовало, наверное, прочесть, записать и опубликовать. Надеюсь на реализацию этого замысла, я не стал ничего менять в записках лекций 1970 года.

(август 1988 г., Пущино)

# ЛЕКЦИЯ 1

## ОБЩЕЕ ВВЕДЕНИЕ

1. Я начну с нескольких общих замечаний на тему: в каком виде сейчас представляется положение дел в вычислительной математике специалистам-математикам.

Вычислительная математика существует столь же давно, что и математика вообще, но за последние 20—25 лет произошел качественный скачок, связанный с появлением электронных вычислительных машин.

Этот скачок привел к появлению нового метода изучения природы, а именно — к возможности проведения численного эксперимента. Можно сказать, что сейчас, кроме теоретического и естественно-экспериментального метода, существует еще третий — метод численного эксперимента. Это и есть то новое, что возникло в последние 20—25 лет.

В методологическом смысле все три подхода равноправны, т. е. в зависимости от задачи, от области науки и, наконец, от сложившейся ситуации, первый, второй или третий подход к проблеме оказывается более успешным, более важным, необходимым и т. д.

Соотношение между ними можно изобразить следующим образом:



Стрелки изображают связи: теория может проверяться численным экспериментом, численный эксперимент — физическим или наоборот.

2. Можно кратко обрисовать те случаи, когда численный эксперимент становится особенно важным или просто необходимым. Я попытался наметить три таких области, границы между которыми, конечно, можно провести лишь условно.

Первая область задач охватывает те вопросы, для которых, можно сказать, теория разработана безупречно. Говоря языком математиков — уравнения абсолютно достоверны, но эксперимент поставить трудно или вообще невозможно. Роль численного эксперимента при этом необычайно высока. Самый, пожалуй, яркий пример этому — проблемы небесной механики, или же, в современном аспекте — проблемы космической динамики. Уравнения, с которыми имеем дело, это уравнения динамики. Они бесспорны. Но, скажем, если речь идет о выработке программы полета к Луне, то раньше, чем полетит к ней реальный объект, проводятся расчеты множества траекторий с оптимизацией нужных параметров и т. д., и без подобного численного эксперимента были бы невозможны успешные полеты.

В эту же группу задач, очевидно, входит и знаменитая проблема трех тел\*. Здесь численный эксперимент оказывается полезным уже не для чисто практических целей, а для выяснения сложных теоретических вопросов.

Вторая область задач охватывает те вопросы, когда явления в общем понятны, но имеются неизвестные нам параметры. Мы же, имея данные эксперимента, хотим выяснить, каковы эти параметры. Один из ярких примеров этому — проблема внутреннего строения звезд. Там, внутри звезд, происходят ядерные реакции, которые в настоящее время детально изучены, есть теория механической устойчивости звезд, но вместе с тем есть и множество неясностей. Например, мы не знаем химический состав звезды (от него зависит очень много!), не известны условия в центре и т. д., словом — нет характеристик, которые не поддаются прямому измерению. В нашем распоряжении имеется тем не менее, достаточно много данных: спектры, поверхностная температура, масса. Чтобы достигнуть понимания того, что происходит в звездах, — нам необходимо построить такие модели внутреннего строения звезд, которые давали бы нужные значения внешних наблюдаемых параметров.

Здесь, по-видимому, более 95% известных сведений получено из численных экспериментов. В ходе таких экспериментов выясняются тонкие детали строения звезд. Например, различные ядерные реакции могут протекать в различных слоях (зонах). Есть случаи, когда таких зон было найдено до 10. Довольно успешно удавалось также проследить и эволюцию звезды во времени.

---

\* Т. е. задача о движении трех материальных точек, связанных Ньютоновским тяготением.



Третья область задач — изучение очень сложных явлений, когда протекают одновременно и взаимодействуют множество элементарных процессов. Грубо говоря, теория известна лишь в общих чертах, и с точки зрения формального математика тут нечего решать, т. к. не написаны сами уравнения. В этих случаях можно не пытаться сразу объяснить явление в целом, а вместо этого получить некоторые частные сведения, используя упрощенные «модели». При рассмотрении ряда моделей можно будет, хотя бы в принципе, добиться нужного понимания наблюдаемых явлений. К этой области задач относятся современные задачи газовой динамики, и в качестве примера рассмотрим стандартную газодинамическую задачу: тело сложной формы летит с большой скоростью. Во время полета оно сильно разогревается, начинается диссоциация молекул газа, затем, при более высоких температурах — ионизация, может статься, что тело начнет плавиться и т. д. Картина явления, в целом, необычайно сложна, и описать ее математически точно не представляется возможным. Что делают в таком случае? Берут несколько упрощенных задач, заведомо зная, что они не отвечают действительности, и получают для них решения. Ну, например, тело имеет «хитрую» форму и рассмотрение его движения строго математически исследователи дать не могут. Тогда они заменяют его конусом и говорят: «Эту задачу мы можем решить точно и это решение может быть полезно при рассмотрении движения тел другой формы». Вторую разновидность задач этой области (не отделенную резко от первой) составляют проблемы, в которых мы не знаем всех элементарных процессов и не ставим себе целью их полное изучение. Вместо этого мы строим (по возможности простые) математические модели, правильно описывающие то, что происходит в природе. Сюда относятся, в частности, большое количество задач химической кинетики. Таким образом, в этой области задач речь может идти:

- а) о расчете идеализированных вариантов поставленной задачи;
- б) о подборе механизма, дающего совпадение с наблюдаемыми эффектами. Постановка численного эксперимента сводится при этом к построению моделей и решению уравнений, описывающих эти модели.

Вот таковы, приблизительно, те проблемы, в которых численный эксперимент играет важную роль. Естественно, можно привести примеры и противоположного свойства, когда от численного эксперимента никакой пользы нет, но на этом я останавливаться не буду.

3. Сама вычислительная математика с появлением возможности проведения численного эксперимента изменила свой дух: рассматриваемые проблемы настолько усложнились, что точные теоремы в реальных задачах доказываются лишь в исключительных случаях.

Вкратце положение дел сейчас таково. Математикам удастся, следуя своим математическим традициям, строго доказать теоремы для простейших моделей, и поэтому вся область собственно вычислительной математики состоит как бы из двух подобластей: малой подобласти, где рассматриваются сравнительно простые задачи и где результаты рассмотрения можно представить в виде серии теорем, и громадной подобласти, связанной с решением практических задач, где никаких теорем нет. Здесь успехи в решении задачи связаны с проведением численных экспериментов. Придумаем модель. Посчитаем. Получилось — хорошо, не получилось — подумаем в чем дело, пересмотрим исходную модель и т. д. Такова стандартная схема работы современного вычислителя.

Чтобы понять «до какой степени нет теорем» в реальных задачах, я приведу единственный пример. Сейчас значительный процент машинного времени во всех странах тратится на решение задач газовой динамики и гидродинамики. Между тем у математиков нет теоремы, которая бы утверждала очень скромную вещь: решение этих задач существует. Уравнения (в простейшем варианте) написаны около 200 лет назад, а теоремы, которая бы утверждала, что решение существует, нет и сегодня, т. е. нет того, с чего, казалось бы, математик должен бы был начинать.

Резюмируя, можно сказать, что вычислительная математика является сейчас полужэкспериментальной наукой. Во всяком случае считать ее просто разделом математики, я думаю, уже нельзя. Вместе с тем не надо забывать, что вычислительная математика опирается на все громадное здание современного Математического Анализа (и алгебры), используя подчас глубокие и абстрактные математические теоремы.

4. То, что я думаю рассказать, наполовину или даже более можно было бы изложить в виде точных теорем. Но я буду больше стремиться объяснить суть дела, чем доказывать теоремы, даже тогда, когда это возможно сделать.

В целом я думаю, что разумно принять порядок, который был когда-то в средней школе — порядок концентрического изложения. Мне кажется, что для курса вычислительной математики стоит сделать 3 концентрических круга. Сначала — основные идеи и общие соображения, затем — более глубоко — основы, а затем отдельные избранные вопросы со всеми подробностями.

В первом концентре, с которого я начну, я буду доказывать очень мало, и даже точные формулировки будут приводиться не всегда. Позже можно будет сделать необходимые уточнения.

Первую тему можно назвать так: интегрирование обыкновенных дифференциальных уравнений. Более точно, мы будем заниматься сегодня начальной задачей, или задачей Коши, для обыкновенных дифференциальных уравнений.

При этом я немного нарушу традиции. Ни в одном курсе вычислительной математики эта тема не является первой, она в луч-

шем случае излагается где-нибудь «в серединке». По степени же практической важности эта тема является одной из важнейших. Учитывая это, а также то, что для понимания сути дела нам потребуются минимальные сведения из математики (только то, что дифференциальные уравнения — это вещь полезная, что их нужно решать, и сверх того — лишь формула Тейлора), я решил начать изложение с этой темы.

## ЗАДАЧА КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

### 1. Постановка задачи Коши

Рассмотрим, для примера, систему уравнений:

$$\begin{cases} \frac{d^2 u}{dt^2} = f(u, T) \\ \frac{dT}{dt} = g\left(u, \frac{du}{dt}, T\right). \end{cases} \quad (1)$$

Это система двух дифференциальных уравнений для определения двух неизвестных функций  $u(t)$  и  $T(t)$ . Функции эти неравноправны: в систему входит дважды продифференцированная функция  $u$  и однажды — функция  $T$ . Чтобы найти  $u(t)$  и  $T(t)$ , нужно задать 3 условия при  $t=0$ :

$$u(0) = u_0; \quad u'(0) = u'_0; \quad T(0) = T_0. \quad (2)$$

Данные (2) называют начальными данными. В общем случае для определения функций из дифференциальных уравнений надо задать столько начальных данных для каждой функции, каков порядок ее наивысшей производной.

Отметим, затем, что в системе (1) старшие производные для каждой из функций выделены, или — можно сказать иначе — система разрешена относительно старших производных. В таких случаях говорят, что система написана в нормальной форме. Будем считать, что в дальнейшем система дифференциальных уравнений записывается в нормальной форме.

Итак, постановка задачи такова: задана система дифференциальных уравнений в нормальной форме и необходимое число начальных данных при  $t=0$ . Требуется найти все функции, входящие в систему, на некотором интересующем нас промежутке времени  $0 \leq t \leq t_{\max}$ .

### 2. Теоретическое отступление

В теории дифференциальных уравнений доказывается «теорема существования и единственности». Она гарантирует, что ре-

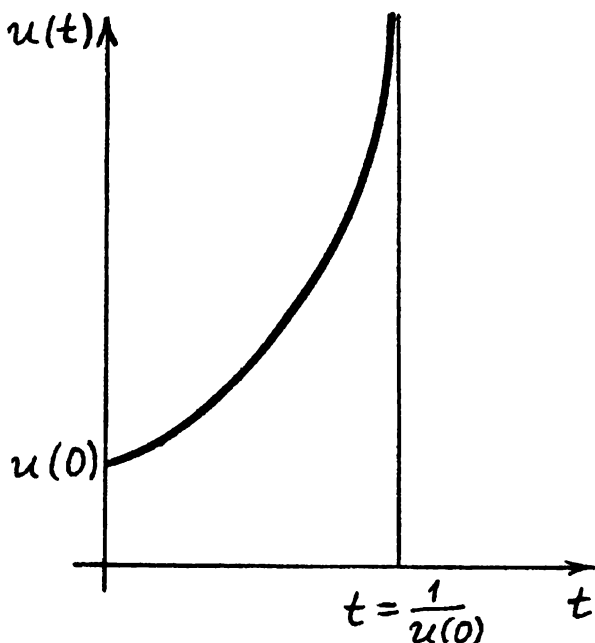
шение интересующей нас задачи существует (и притом только одно). Но! — только на некотором участке времени  $0 \leq t \leq t_*$ .

Это ограничение — не является дефектом математики. Это — *ограничение по существу*. Для системы  $n$  линейных уравнений

$$\frac{du^\alpha}{dt} = \sum_{\beta=1}^n a_{\beta}^{\alpha}(t) u^{\beta} \quad \alpha=1, \dots, n$$

такого ограничения нет: решение задачи Коши существует при всех  $t$ .

Иначе обстоит дело в случае нелинейных уравнений. Рассмотрим простейшее нелинейное дифференциальное уравнение:  $u' = u^2$ . Его решение:  $u(t) = 1/(c - t)$ . Считая, что начальное значение  $u(0) = 1/c$  задано, получим:  $u(t) = \frac{u(0)}{1 - u(0)t}$ . При  $t_* = c$  знаменатель обращается в ноль. Если  $u(0) > 0$ , то удастся найти решение задачи Коши только до момента  $t_* = 1/u(0)$ . На графике это решение выглядит следующим образом:



### 3. Стандартная запись

Возвратимся к рассмотрению системы (1). Перепишем ее, используя стандартную форму записи — в виде системы дифференциальных уравнений 1-го порядка.

Пусть  $u' = v$ . Тогда задача Коши для системы (1) может быть представлена таким образом:

$$\begin{cases} u' = v \\ v' = f(u, T) \\ T' = g(u, v, T) \end{cases} \quad \begin{array}{l} u(0), v(0), T(0) \text{ — заданы;} \\ u(t), v(t), T(t) \text{ — ищутся.} \end{array} \quad (3)$$

Для простоты записи будем использовать далее векторные обозначения

$$u = (u^1, u^2, \dots, u^n); \quad f = (f^1, f^2, \dots, f^n). \quad (4)$$

Тогда общую систему дифференциальных уравнений первого порядка можно записать таким образом

$$\frac{du}{dt} = f(u, t). \quad (5)$$

Итак, общая задача, стоящая перед нами, такова: дана система уравнений (5), задано  $u(0)$  (т. е. все  $u^i(0)$ ), нужно найти  $u(t)$ . Для решения такой «стандартной задачи» на большинстве вычислительных машин существуют стандартные программы.

Последнее замечание. Часто интересующая нас система (5) является автономной, т. е. в правую часть ее не входит (явно) независимое переменное  $t$ :

$$\frac{du}{dt} = f(u). \quad (6)$$

Так будет, например, во всех тех случаях, когда  $t$  есть время и система (5) описывает процесс, происходящий в неизменных внешних условиях. При теоретических рассмотрениях бывает удобно и неавтономную систему (5) записать в автономном виде\*. Чтобы максимально упростить формулы, я буду считать, что это приведение уже сделано и вместо (5) буду (не изменяя обозначений) рассматривать систему (6).

#### 4. Метод Эйлера

Переходя к построению приближенных решений, мы будем для простоты писать формулы для одного уравнения, указывая там, где это потребуется, изменения в формулах для случая системы.

---

\* Этого всегда можно добиться введением нового независимого переменного  $s$  и формальным увеличением числа неизвестных функций на единицу.

Итак, пусть задано уравнение

$$\frac{du}{dt} = f(u) \quad (7)$$

и начальное значение  $u(0) = u_0$ . Начнем с того, что перейдем от непрерывного изменения  $t$  к дискретному: выберем некоторое маленькое число  $\tau$  и будем рассматривать лишь значения  $t = t_k = k\tau$ . Постараемся найти  $u(t_k)$ . Положим в качестве приближенного значения

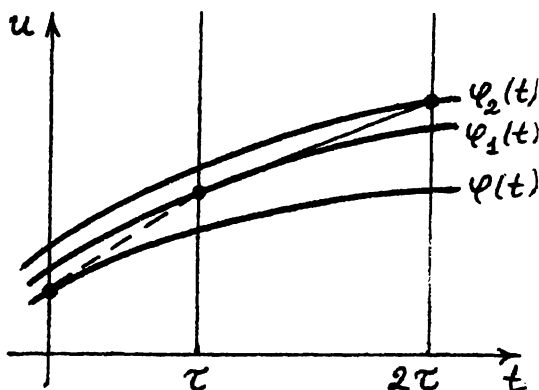
$$u(\tau) \approx u_1 = u_0 + \frac{du}{dt}(0)\tau = u_0 + f(u_0)\tau.$$

Положим далее

$$\begin{aligned} u(2\tau) &\approx u_2 = u_1 + f(u_1)\tau \\ \dots\dots\dots &\dots\dots\dots \\ u(k\tau) &\approx u_k = u_{k-1} + f(u_{k-1})\tau. \end{aligned} \quad (8)$$

Такой способ приближенного решения уравнения (7)' называется методом Эйлера.

Насколько отличаются значения  $u_1, u_2, \dots, u_k$  от точных значений интересующей нас функции  $u = \varphi(t)$ ? На первом шаге мы вместо точной формулы Тейлора воспользовались первыми двумя членами ее. Иначе говоря, на промежутке  $(0, \tau)$  мы двигались не по кривой  $u = \varphi(t)$ , а по касательной к ней. Точка с координатами  $(\tau, u_1)$  лежит поэтому не на кривой  $u = \varphi(t)$ , а на некотором соседнем решении уравнения (7)  $u = \varphi_1(t)$ . На следующем шаге мы, двигаясь по касательной к кривой  $u = \varphi_1(t)$ , попадаем на кривую  $u = \varphi_2(t)$  и т.д. Ясно, что метод Эйлера не является особенно точным. Пренебрегая на каждом шаге в формуле Тейлора всеми членами, кроме первых двух, мы делаем ошибку



порядка  $\tau^2$ . Ошибки эти должны каким-то образом суммироваться, так что к концу интересующего нас интервала мы можем ожидать ошибки порядка  $\tau$ . В таких случаях говорят, что метод имеет первый порядок точности. Для практических нужд он слишком груб и почти никогда не употребляется.

Займемся теперь уточнением метода Эйлера.

### 5. Более точные методы: использование старших производных

Для того, чтобы получить более точные формулы, следовало бы написать формулу Тейлора с бóльшим числом членов. На первый взгляд, кажется, что это невозможно, так как уравнение доставляет нам только первую производную  $u' (= \frac{du}{dt})$ . Однако можно из этого же уравнения извлечь старшие производные. В самом деле, если  $u = \varphi(t)$  решение, то  $\varphi'(t) = f(\varphi(t))$  есть тождество. Продифференцировав его по  $t$ , мы получим:

$$\varphi'' = f'(\varphi) \varphi' = f'(\varphi) f(\varphi).$$

Таким образом, следствием из уравнения  $u' = f(u)$  является уравнение

$$\frac{d^2 u}{dt^2} = f_2(u), \quad f_2(u) = f'(u) f(u). \quad (9)$$

Для системы уравнений (6) эти формулы имеют вид

$$\frac{d^2 u^i}{dt^2} = \sum_k \frac{\partial f^i}{\partial u^k} f^k(u). \quad (9a)$$

Аналогично можно было бы получить уравнение  $u''' = f_3(u)$  и т. д. Если теперь положить

$$u_{k+1} = u_k + \tau f(u_k) + \frac{\tau^2}{2} f_2(u_k), \quad (10)$$

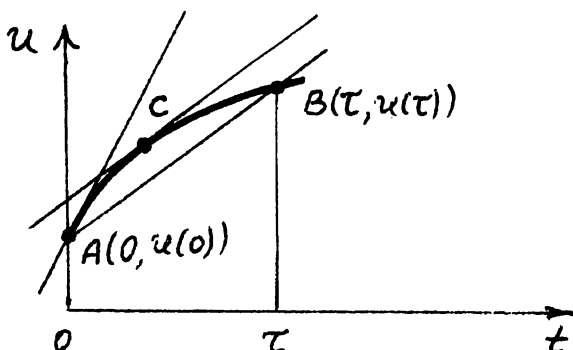
т. е. учесть в формуле Тейлора 3 слагаемых, мы получим метод, дающий на каждом шаге ошибку порядка  $\tau^3$ .

### 6. Более точные методы: многократное использование заданного уравнения

Во многих случаях (особенно для систем уравнений) выражения для функций  $f_2$ ,  $f_3$  очень громоздки; поэтому обычно стараются избежать вычисления этих функций и получают несколько членов формулы Тейлора косвенно, используя на каждом

шаге несколько раз исходное уравнение. Простейшие формулы такого рода, называемые уточненным методом Эйлера (или методом Эйлера с пересчетом), мы сейчас выведем.

Наводящие соображения можно получить из рассмотрения рисунка.



По нему видно, что к направлению хорды  $AB$  гораздо ближе направление касательной в средней точке  $C$ , чем в точке  $A$ . Если мы сумеем (достаточно точно) найти  $u'(\frac{\tau}{2})$ , то можно надеяться более точно найти  $u(\tau)$ . Поскольку  $u'(\frac{\tau}{2}) = f(u(\frac{\tau}{2}))$ , нам достаточно найти  $u(\frac{\tau}{2})$ . Итак, положим сначала:  $u_{1/2} = u_0 + \frac{\tau}{2} f(u_0)$ . А теперь  $u_1 = u_0 + \tau f(u_{1/2})$ . Оценим, насколько отличается  $u_1$  от точного значения  $u(\tau)$  и проверим, оправдались ли наши ожидания. По формуле Тейлора:

$$f(u_{1/2}) = f\left(u_0 + \frac{\tau}{2} f(u_0)\right) = f(u_0) + f'(u_0) \frac{\tau}{2} f(u_0) + O(\tau^2).$$

И дальше:

$$u_1 = u_0 + \tau f(u_{1/2}) = u_0 + \tau f(u_0) + \frac{\tau^2}{2} f(u_0) f'(u_0) + O(\tau^3). \quad (11)$$

Теперь вспомним, что  $f(u_0) = u'(0)$ , а  $f(u_0) f'(u_0) = u''(0)$ . Значит в формуле (11) первые три члена совпадают с формулой Тейлора и отличие  $u(\tau)$  от  $u_1$  имеет порядок  $\tau^3$ . Таким образом, мы получили более точную формулу за счет двукратного вычисления функции  $f$ . При этом мы косвенно (без прямого дифференцирования) учли вторую производную решения. Перепишем еще раз формулы уточненного метода Эйлера. При известном  $u_k$  мы находим последовательно:



$$u_{k+1/2} = u_k + \frac{\tau}{2} f(u_k), \quad u_{k+1} = u_k + \tau f(u_{k+1/2}). \quad (12)$$

На промежутке изменения  $t$  порядка 1 этот метод дает ошибку порядка  $\tau^2$  или, как говорят, имеет 2 порядок точности. В большинстве случаев при решении дифференциальных уравнений нужен более высокий порядок точности. Следующие красивые формулы Рунге-Кутты дают метод 4 порядка точности. При их использовании косвенно используются производные решения до 4 порядка включительно. Достигается это вычислением 4 значений  $f$  от разных аргументов:

$$1 \text{ этап} \quad f^{(1)} = f(u_k),$$

$$2 \text{ этап} \quad f^{(2)} = f(u_k + \frac{\tau}{2} f^{(1)}),$$

$$3 \text{ этап} \quad f^{(3)} = f(u_k + \frac{\tau}{2} f^{(2)}),$$

$$4 \text{ этап} \quad f^{(4)} = f(u_k + \tau f^{(3)}),$$

$$u_{k+1} = u_k + \frac{\tau}{6} (f^{(1)} + 2f^{(2)} + 2f^{(3)} + f^{(4)}).$$

Я не знаю хорошего способа объяснить, как придумать такие формулы. Но когда они уже написаны, оценить ошибку на одном шаге нетрудно. Нужно снова воспользоваться формулой Тейлора, удержав необходимое число слагаемых. Эту мало поучительную выкладку я проводить не буду.

## 7. Идея Адамса:

**использование уже найденного участка решения для приближенного вычисления старших производных**

Существует другой путь для построения более точных методов. Именно, для приближенного вычисления старших производных можно использовать уже найденные значения  $u_k$  или  $u'_k = f(u_k)$ . Простейшая реализация этой идеи такова. Положим  $u''_k = \frac{1}{\tau} (u'_k - u'_{k-1})$ . Тогда можно рассчитывать, что формула

$$u_{k+1} = u_k + \tau u'_k + \frac{\tau^2}{2} u''_k = u_k + \tau \left( \frac{3}{2} u'_k - \frac{1}{2} u'_{k-1} \right) \quad (14)$$

будет точнее, чем метод Эйлера. И в самом деле, можно показать, что формула (14) (где  $u'_k = f(u_k)$ ) дает метод 2 порядка точности. Аналогично, используя большее число предыдущих значений, можно получить еще более точные формулы. Например, формула часто употребляемого *метода Адамса 4 порядка* имеет вид

$$u_{k+1} = u_k + \frac{\tau}{24} (55u'_k - 59u'_{k-1} + 37u'_{k-2} - 9u'_{k-3}). \quad (15)$$

Здесь  $u'_s = f(u_s)$ . Такого рода формулы (в отличие от формул Рунге-Кутты!) легко выводятся, если предположить, что все используемые значения  $u'_s$  относятся к одному (точному) решению дифференциального уравнения. Достаточно разложить  $u'_s$  по формуле Тейлора (в окрестности точки  $t_k$ ) и подобрать подходящие коэффициенты при  $u'_s$  \*. Однако простота эта обманчива. Дело в том, что при практическом применении метода Адамса значения  $u_k, u_{k-1}, u_{k-2}, \dots$  никогда не лежат на одной интегральной кривой \*\*. Поэтому само понятие «ошибка на одном шаге» и использование формулы Тейлора для оценки этой ошибки становятся не ясными.

При следующем, более глубоком рассмотрении темы «Обыкновенные дифференциальные уравнения» мы вернемся к этим вопросам.

И последнее, очевидное замечание. При использовании метода Адамса (и ему аналогичных) первые несколько значений нужно найти каким-то другим способом, например, методом Рунге-Кутты.

## 8. Исторические и другие замечания

По-видимому, Эйлер уже владел всеми изложенными идеями, кроме идеи Адамса. Он понимал, как высшие производные получать из данного уравнения прямым дифференцированием, а также то, что можно обойтись и без прямого дифференцирования. В его курсе «Интегрального исчисления» (1768) указывалось, что можно вычислять высшие производные косвенным образом.

Изложение метода Адамса было дано в 1883 г. в книге Башфорта и Адамса по теории капиллярных сил (придуман он был, по-видимому, гораздо раньше, в 1855 г.). Но в дальнейшем этот метод был забыт настолько, что даже в очень популярной в начале нашего века энциклопедии математических знаний, опубликованной на немецком языке, ничего не говорится о методе Адамса.

В конце XIX века была опубликована работа Рунге, идейно восходящая к Эйлеру, а затем через несколько лет работа Кутты, где был приведен целый набор формул. Идеи Адамса, возродил, по-видимому, академик А. Н. Крылов, и сейчас метод Адамса получил широкое применение и воспринимается, как равноправно сосуществующий с методом Рунге-Кутты.

\* Ход вычислений намечен в примечании 2.

\*\* Аккуратнее было бы говорить о точках  $(t_k, u_k); (t_{k-1}, u_{k-1}); \dots$

Сейчас, действительно, кажется неразумным (ну, по крайней мере — неэкономным!) считать, не используя полученную ранее информацию: применение метода Адамса (и аналогичных ему) часто уменьшает время счета в 3—4 раза.

С другой стороны, кажется оправданным применение метода Рунге-Кутты в случаях, когда решение ведет себя «непредсказуемо», т.е. интегральная кривая имеет «неожиданные» всплески или сбросы (участки резких изменений в ходе  $u(t)$ ). В частности, необходимое на таких участках измельчение шага метод Рунге-Кутты переносит безболезненно.

Очевидно, что в каждом конкретном случае нужно специально рассматривать вопрос о применимости того или иного метода счета. Нужно думать, Не надо быть настолько оптимистом, чтобы предполагать, что, имея вычислительную машину и программы Рунге-Кутты или Адамса (или те и другие), достаточно уметь вычислить правые части уравнения  $u' = f(u, t)$ . Часто встречаются задачи, которые без размышлений и аналитического исследования «напрямую» не могут быть решены. Это могут быть задачи, где трудности заранее очевидны. Скажем, нужно выходить (при  $t=0$ ) или проходить (где-то при  $t>0$ ) через особые точки системы. Далее в системе могут быть большие или малые параметры. Например, система дифференциальных уравнений может описывать одновременно протекающие процессы, сильно различающиеся по скорости. Ясно, что их совместное рассмотрение может оказаться очень трудным. Однако, встречаются и такие задачи, где трудности заранее неочевидны, завуалированы. Грубо говоря, особенности «проявляются» лишь в процессе решения.

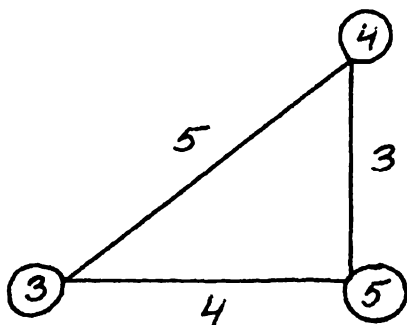
В качестве примера рассмотрим один вариант задачи трех тел.

## 9. Пример

Условие: имеется 3 материальные точки с массами 3, 4 и 5, расположенные в углах Пифагорова треугольника. Их скорости в начальный момент равны 0, а взаимодействие между ними определяется тяготением. Эти точки как-то движутся в плоскости, после того, как «включено тяготение и время». Для решения задачи необходимо 6 уравнений Ньютона (для каждой точки пишется по 2 уравнения 2 порядка). Правые части уравнений просты — их легко запрограммировать, но тем не менее эта задача не была решена до 1967 г. \*. Дело в том, что задача существенно хитрее, чем кажется на первый взгляд. Точки начинают двигаться, и в какой-то момент две из них начинают быст-

---

\* См. V. Szebehely and C. F. Peters. Complete solution of a general problem of three bodies. *Astron. Journ.*, 1967, v. 72, No 7. Рисунки из этой статьи перепечатаны в книге: В. Г. Демин. Судьба солнечной системы. «Наука», 1969.



ро сближаться и подходят друг к другу на расстояние  $\sim 10^{-4}$ . Если вспомнить, что Ньютонова сила тяготения меняется с расстоянием как  $1/r^2$ , то ясно, что при сближении взаимодействие изменяется на 8 порядков! Это и есть трудное место в задаче — даже ничтожные ошибки, полученные при обсчете конфигурации, когда 2 частицы сближены, могут повести в дальнейшем к очень большим ошибкам. Через это место в задаче, фигурально выражаясь, чрезвычайно трудно пробиться, тут надо думать. Авторам пришлось решать эту задачу полуаналитически. Они ввели новые переменные, переписали ряд формул в другом виде, разобрались с особенностями, и только тогда удалось «пробиться». Результаты оказались следующими.

Частицы описывают очень замысловатые траектории и время от времени близко сходятся. Через время около 30 выбранных единиц частицы подходят почти к своим исходным позициям и скорости их близки к нулевым. Казалось бы, что после этого движение должно повториться, но ничтожных отличий достаточно, чтобы этого не получилось\*. Спустя время порядка 60 единиц получается следующее: два более тяжелых тела начинают вращаться вокруг общего центра масс и удаляться в одну сторону, тогда как третье тело движется в противоположную. Центр масс всей системы, естественно, остается «на месте». Система распалась, и далее считать уже лишено смысла, далее надо «писать формулы» для задачи двух тел (надо отметить, что возможность такого распада системы была показана задолго до этой работы у нас в Союзе).

Последние два замечания.

1. При счете использовался один из вариантов метода Рунге-Кутты.

2. Ошибка при расчете критической ситуации даже в восьмом знаке может превратиться в ошибку в первом знаке! Поэтому

---

\* Это наводит на мысль, что, если при  $t=0$  частицы расположить несколько по-иному, то можно получить точное повторение. Другими словами, в окрестности найденного решения, возможно, есть периодическое решение.

здесь особенно необходимо контролировать точность вычислений. В данной задаче критерий точности был косвенным — это закон сохранения энергии. При точном решении этот закон должен выполняться точно, при численном решении — приблизительно. В задаче получено сохранение энергии с точностью до 10-го десятичного знака, и это является косвенным свидетельством точности вычислений.

## Дополнения и примечания к 1 лекции

### 1. (к п. 3) Сведение произвольной системы к автономной

Вместо системы (5)  $\frac{du}{dt} = f(u, t)$  напомним:

$$\frac{du}{ds} = f(u, t), \quad \frac{dt}{ds} = 1. \quad (5')$$

Исходная задача Коши (задано  $u(0) = u_0$ ; ищется  $u(t)$ ) перейдет в эквивалентную: заданы  $u(0) = u_0$ ,  $t(0) = 0$ ; ищутся  $u(s)$ ,  $t(s)$ . Система (5') является автономной,  $t$  есть теперь одна из неизвестных функций, а новая независимая переменная  $s$  — в правую часть не входит.

Обозначив  $v = (u^1, \dots, u^n, t)$ ;  $g = (f^1, \dots, f^n, 1)$ , можно записать систему (5') в виде:  $\frac{dv}{ds} = g(v)$ . Отметим, что процедура сведения неавтономной системы к автономной, столь удобная для вывода общих формул, не всегда является безобидной. Например, линейное уравнение  $\frac{du}{dt} = a(t)u$  перейдет при этом в нелинейную систему:

$$\begin{cases} \frac{du}{ds} = a(t)u \\ \frac{dt}{ds} = 1. \end{cases}$$

2. (к п. 7). Вычисления, приводящие к формулам метода Адамса, проиллюстрируем на примере формулы 3-го порядка точности. Ищем формулу метода в виде:

$$u_{k+1} = u_k + \tau(au'_k + bu'_{k-1} + cu'_{k-2}); \quad u'_k = f(u_k).$$

Неопределенные коэффициенты  $a$ ,  $b$ ,  $c$  мы подберем так, чтобы получить нужный порядок точности. По формуле Тейлора

$$u'_{k-1} = u'_k - \tau u''_k + \frac{\tau^2}{2} u'''_k + O(\tau^3); \quad u''_k = u''(t_k), \quad u'''_k = u'''(t_k)$$

$$u'_{k-2} = u'_k - 2\tau u''_k + 2\tau^2 u'''_k + O(\tau^3)$$

$$u_{k+1} = u_k + \tau \left[ (a+b+c)u'_k - \tau(b+2c)u''_k + \right. \\ \left. + \tau^2 \left( \frac{b}{2} + 2c \right) u'''_k \right] + O(\tau^4).$$

Для точного решения  $u(t)$ :

$$u(t_{k+1}) = u(t_k + \tau) = u(t_k) + \tau u'(t_k) + \\ + \frac{\tau^2}{2} u''(t_k) + \frac{\tau^3}{6} u'''(t_k) + O(\tau^4).$$

Приравнивая коэффициенты при одинаковых степенях  $\tau$ , получаем систему уравнений:

$$a+b+c=1; \quad b+2c=-\frac{1}{2}; \quad \frac{b}{2}+2c=\frac{1}{6}.$$

Найдя  $a, b, c$  из этой системы, получаем искомую формулу метода Адамса 3-го порядка:

$$u_{k+1} = u_k + \frac{\tau}{12} (23u'_k - 16u'_{k-1} + 5u'_{k-2}).$$

Аналогичные рассуждения приводят к формуле (15) в тексте. Напомним еще раз, что при выводе мы считали значения  $u_k, k_{k-1}, u_{k-2}$  принадлежащими (некоему) точному решению  $u(t)$  уравнения  $u'=f(u)$ :  $u_s = u(t_s), u'_s = u'(t_s); s=k, k-1, u-2$ .

### 3. (к п. 6). Формулы метода Рунге-Кутты для общей системы 1 порядка

$$\text{Пусть } \frac{du}{dt} = f(u, t).$$

По известному  $u_k$  находятся последовательно:

- 1)  $f^{(1)} = f(u_k, t_k)$
- 2)  $f^{(2)} = f\left(u_k + \frac{\tau}{2} f^{(1)}, t_k + \frac{\tau}{2}\right)$
- 3)  $f^{(3)} = f\left(u_k + \frac{\tau}{2} f^{(2)}, t_k + \frac{\tau}{2}\right)$
- 4)  $f^{(4)} = f(u_k + \tau f^{(3)}, t_k + \tau)$

$$\text{Затем } u_{k+1} = u_k + \frac{\tau}{6} (f^{(1)} + 2f^{(2)} + 2f^{(3)} + f^{(4)}).$$

#### 4. Литературные замечания

Из книг «домашнего» периода особый интерес представляет книга: «Лекции о приближенных вычислениях» академика А. Н. Крылова (5 издание, 1950 г.). В разделе, посвященном интегрированию обыкновенных уравнений, там содержатся, в частности, исторические замечания п. 8. Из более поздних трудов стоит ознакомиться с книгой Милна «Численное решение дифференциальных уравнений» (ИЛ; 1955). Это ясно написанная книга относится к началу «машинного периода».

Наконец, современная точка зрения (в несколько крайнем своем выражении) изложена в книге Р. В. Хемминга «Численные методы» («Наука»; 1968).

## ЛЕКЦИЯ 2

### СИСТЕМЫ ЛИНЕЙНЫХ УРАВНЕНИЙ. МЕТОД ИСКЛЮЧЕНИЯ

В этой лекции в первом приближении будет рассмотрен вопрос о решении систем линейных алгебраических уравнений. Системы линейных уравнений часто встречаются в вычислительной практике. Особенно часто — как составная часть задачи. Так, например, можно сказать (это будет не очень далеко от истины), что в задачах математической физики мы вообще никаких уравнений кроме линейных не решаем: решение сложных нелинейных задач сводится к многократному решению линейных систем. Забегая вперед, я хочу подчеркнуть следующее. О задаче решения линейных систем как о единой задаче имеет смысл говорить лишь до тех пор, пока число уравнений невелико. В таком случае это есть сравнительно простая задача\*. Чем больше уравнений в системе, тем более специальную, специфическую структуру она обычно имеет, и тем более необходимо для решения этой системы использовать ее специальные свойства. Некоторые виды таких систем и специфические способы их решения мы рассмотрим позже.

#### 1. Метод исключения

Простейший способ решения линейных систем по существу всем известен из школьного курса. Обычно, когда школьники решают систему линейных уравнений, они исключают какое-либо неизвестное, получают систему меньшего порядка. Потом исключают еще одно неизвестное и т. д. — пока не доберутся до одного уравнения с одним неизвестным. Оформив несколько более культур-

---

\* Стандартные программы для решения линейных систем есть на всех вычислительных машинах.



но этот процесс, мы и получим метод исключения. Я поясню его сначала на простом числовом примере. Пусть дана система 4 уравнений:

$$\begin{aligned} 2x_1 + 3x_2 + 6x_3 + 8x_4 &= 7 & [1] \\ 3x_1 + 7x_2 + 3x_3 + 6x_4 &= 3 & [2] \\ 2x_1 + 4x_2 + 7x_3 + 7x_4 &= 2 & [3] \\ 2x_1 + 5x_2 + 3x_3 + 7x_4 &= 3 & [4] \end{aligned} \quad (1)$$

Из первого уравнения можно выразить  $x_1$  через  $x_2, x_3, x_4$ . Если подставить это выражение в остальные уравнения, то получится система 3 уравнений с 3 неизвестными. Практически удобнее провести это исключение, комбинируя уравнения. Например, так. Разделим первое уравнение на коэффициент при  $x_1$ . Получим:

$$x_1 + \frac{3}{2}x_2 + 3x_3 + 4x_4 = \frac{7}{2} \quad [1']$$

Теперь из уравнения [2] вычтем [1'], умноженное на 3, из [3] вычитаем [1'], умноженное на 2 и т. д. В результате новые уравнения [2'—4'] составляют отдельную систему с 3 неизвестными ( $x_2, x_3, x_4$ ), а вся система приобретает вид:

$$x_1 + \frac{3}{2}x_2 + 3x_3 + 4x_4 = \frac{7}{2} \quad [1']$$

$$\frac{5}{2}x_2 - 6x_3 - 6x_4 = -\frac{15}{2} \quad [2']$$

$$x_2 + x_3 - x_4 = -5 \quad [3']$$

$$2x_2 - 3x_3 - x_4 = -4. \quad [4']$$

Исключая аналогично из уравнений [3'] и [4'] 2 неизвестное ( $x_2$ ), с помощью уравнения [2'] (уравнение [1'] просто переписывается), мы получим систему:

$$x_1 + \frac{3}{2}x_2 + 3x_3 + 4x_4 = \frac{7}{2}$$

$$x_2 - \frac{12}{5}x_3 - \frac{12}{5}x_4 = -3$$

$$\frac{17}{5}x_3 + \frac{7}{5}x_4 = -2$$

$$\frac{9}{5}x_3 + \frac{19}{5}x_4 = 2.$$

Наконец, на последнем шаге получается следующая система, эквивалентная исходной:

$$\begin{aligned}
 x_1 + \frac{3}{2}x_2 + 3x_3 + 4x_4 &= \frac{7}{2} \\
 x_2 - \frac{12}{5}x_3 - \frac{12}{5}x_4 &= -3 \\
 x_3 + \frac{7}{17}x_4 &= -\frac{10}{17} \\
 \frac{52}{17}x_4 &= \frac{52}{17}.
 \end{aligned}
 \tag{2}$$

Это по-прежнему система 4 уравнений с 4 неизвестными, но она очень легко решается: из последнего уравнения находим  $x_4$ , подставляем его в 3 уравнение и находим  $x_3$ , затем  $x_2$  из 2 уравнения и, наконец,  $x_1$  из 1-го ( $x_4 = 1$ ;  $x_3 = -1$ ;  $x_2 = -3$ ;  $x_1 = 7$ ). Таким образом, можно сказать, что метод исключения состоит в переходе от заданной системы к эквивалентной ей «треугольной системе», которая легко решается.

## 2. Трудности в методе исключения

В рассмотренном примере мы использовали простейшую схему метода исключения:  $x_1$  исключается с помощью первого уравнения,  $x_2$  — с помощью второго и т. д. Ясно, что такая схема не всегда приводит к цели. Так, например, если в 1 уравнение  $x_1$  совсем не входит (коэффициент при нем равен 0), то мы не сможем сделать даже первого шага. В этом случае можно, конечно, взять уравнение, содержащее  $x_1$ , и ему присвоить номер 1.

Аналогичная неприятность (обращение в 0 нужного нам коэффициента) подстерегает нас на каждом шаге метода исключения.

Подчеркнем, что это может случиться в системах вполне благополучных на вид. Например, в такой, у которой все коэффициенты отличны от нуля:

$$\begin{aligned}
 x_1 + 2x_2 + 3x_3 + 4x_4 &= 2 \\
 x_1 + 3x_2 + x_3 + 2x_4 &= -1 \\
 2x_1 + 3x_2 + 8x_3 + 7x_4 &= 10 \\
 2x_1 + 5x_2 + 3x_3 + 7x_4 &= 3.
 \end{aligned}
 \tag{3}$$

Здесь обращается в 0 коэффициент при  $x_3$  после 2 шагов исключения.

Таким образом, на каждом шаге метода исключения может оказаться необходимым выбор уравнения, с помощью которого мы будем исключать очередное неизвестное. Поскольку уже использованные (переставленные на первые места) уравнения мы больше не используем, выбор наш на каждом шаге становится

все более бедным. Естественно возникает вопрос, не окажется ли, что на каком-то шаге очередное неизвестное ( $x_k$ ) не входит ни в одно из оставшихся уравнений (с номерами  $k, k+1, \dots, n$ ). Можно доказать, что если заданная система имеет единственное решение, такой случай невозможен. Таким образом, если бы мы занимались алгеброй, то вопрос о решении системы линейных уравнений можно было бы считать исчерпанным: для всякой системы, имеющей (единственное) решение, указан рецепт его нахождения. Реально дело обстоит сложнее, поскольку все вычисления производятся с конечной точностью.

### 3. Трудности, связанные с приближенностью вычислений. Исключение с выбором главного элемента

При практических вычислениях с конечным числом десятичных знаков мы можем иногда по методу исключения получить результат, далекий от точного решения. Это происходит в тех случаях, когда (при выбранном порядке исключения) некоторые из «ведущих» коэффициентов\* очень малы (хотя и не равны нулю). Чтобы избежать потери точности, можно для исключения очередного неизвестного использовать то уравнение, в котором это неизвестное имеет наибольший коэффициент\*\*. Так видоизмененный способ называется методом исключения с выбором главного элемента. Эффект, достигаемый выбором главного элемента, я покажу на «модельном» примере системы 2 уравнений.

Пусть дана система:

$$\begin{aligned} \varepsilon x_1 + x_2 &= 1 & [1] \\ x_1 + 2x_2 &= 4. & [2] \end{aligned} \quad (4)$$

Возьмем  $\varepsilon = -10^{-5}$  и предположим, что все вычисления производятся с сохранением 4 значащих цифр (или, как коротко говорят, с 4 знаками). Решим сначала эту систему по простейшей схеме метода исключения, т. е. исключим  $x_1$  с помощью первого уравнения:

$$\begin{aligned} x_1 - 10^5 x_2 &= -10^5 & [1'] \\ (2 + 10^5) x_2 &= 4 + 10^5. & [2'] \end{aligned}$$

Если вычисления производятся с 4 знаками, то мы получим из [2']:  $x_2 = 1$ ; подставляя в [1'], получим  $x_1 = 0$ . Теперь поступим по-другому: исключим  $x_1$  с помощью второго уравнения (это и будет для нашего примера исключение с выбором главного элемента). Получим

\* Ведущий коэффициент — коэффициент при исключаемом неизвестном в уравнении, с помощью которого это неизвестное исключается.

\*\* Имеется в виду, конечно, наибольший по абсолютной величине.

$$x_1 + 2x_2 = 4 \quad [1'] \quad (= [2])$$

$$(1 + 2 \cdot 10^{-5})x_2 = 1 + 4 \cdot 10^{-5} \quad [2'] \quad (= [1] + 10^{-5} \cdot [2])$$

Отсюда:  $x_2 = 1$ ;  $x_1 = 2$  (вместо  $x_1 = 0!$ ). Легко проверяется, что точные значения  $x_1$  и  $x_2$  отличаются от  $x_1 = 2$ ,  $x_2 = 1$  меньше чем на  $10^{-4}$ , так что второй раз мы решили систему достаточно точно. Возвращаясь к общему методу, заметим, что до сих пор мы исключали неизвестные в их первоначальном порядке ( $x_1, x_2, \dots$ ) и меняли, при необходимости, лишь порядок уравнений. Возможны, конечно, и другие варианты метода исключения. Например, такой. На очередном шаге выбирается не уравнение, а неизвестное, которое в очередном уравнении стоит с наибольшим коэффициентом. Этот способ называется методом исключения с выбором главного элемента по строке; в отличие от него наш первоначальный выбор называют выбором главного элемента по столбцу\*.

#### 4. Всегда ли метод исключения с выбором главного элемента дает разумный результат?

На этот естественный вопрос, по-видимому, нет четкого ответа. Практика показывает, что если система «не очень плохая» (что это значит — надо определять в каждом конкретном случае), то метод исключения с выбором главного элемента «хорошо работает». Однако я хочу обратить ваше внимание на то, что метод исключения с выбором главного элемента можно «обмануть», умножив одно из уравнений на очень большое число. Покажем это на том же модельном примере (4). Мы убедились, что в этой системе (при вычислениях с 4 знаками) нельзя исключать  $x_1$  с помощью первого уравнения, куда это неизвестное «почти не входит». Умножив первое уравнение на  $N = 10^6$ , мы получим систему, формально (с точки зрения алгебры) эквивалентную исходной:

$$-10x_1 + 10^6x_2 = 10^6$$

$$x_1 + 2x_2 = 4.$$

Выбирая главный элемент в столбце  $\begin{pmatrix} -10 \\ 1 \end{pmatrix}$ , мы должны будем как раз исключать  $x_1$  с помощью первого уравнения! Аналогично можно изменить порядок исключения при выборе главного элемента по строке, разделив одно из неизвестных, скажем  $x_1$ , на большое число  $N$  (т. е. введя новое неизвестное  $y_1 = \frac{x_1}{N}$ ).

---

\* Поскольку выбирается наибольший элемент в строке или столбце матрицы коэффициентов.

Таким образом, умножение на число отдельных уравнений или неизвестных не есть безобидное действие с точки зрения вычислительной математики: вычислительный алгоритм (в частности, программа для машины) может при этом сработать по-другому и дать совсем другой результат. При практическом решении систем линейных уравнений желательна поэтому разумная нормировка уравнений и неизвестных. Обычно при подходящем выборе единиц измерения коэффициенты в уравнениях бывают одного порядка. Если это оказалось не так, то стоит перед решением сделать вот что: отнормировать уравнения так, чтобы в каждом уравнении максимальный коэффициент был порядка 1\*.

## 5. О трудоемкости метода исключения для общей системы $l$ линейных уравнений

Решение общей системы  $n$  линейных уравнений с  $n$  неизвестными по методу исключения состоит из 2 этапов.

1. Процесс исключения: сведение заданной системы к эквивалентной ей «треугольной» (уравнения «треугольной» системы есть некие комбинации исходных уравнений).

II. Решение треугольной системы. Забегая вперед, замечу, что основную часть вычислительной работы поглощает I этап — процесс исключения неизвестных. Его трудоемкость еще несколько увеличивается, если нужно производить выбор главного элемента\*\*. Чтобы подсчитать число арифметических действий, опишем еще раз ход вычислений для простейшей схемы метода исключения. Итак, пусть задана система

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = f_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = f_2 \\ \dots\dots\dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = f_n \end{cases} \quad (5)$$

Как обычно, первый индекс у коэффициента — номер уравнения, а второй — номер неизвестного. Опишем процесс исключения неизвестных. При этом для сокращения записи будем писать только коэффициенты перед неизвестными и свободные члены. Исходная система тогда будет иметь вид:

\* Стандартные программы обычно не включают в себя нормировку.

**\*\* Для вычислительной машины даже перестановка двух уравнений есть «работа»: пересылка числа часто требует времени, сравнимого со временем арифметических операций.**

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{vmatrix} f_1 \\ f_2 \\ \dots \\ f_n \end{vmatrix}.$$

Поступаем следующим образом:

- 1) первое уравнение делим на  $a_{11}$ :  $[1] \rightarrow [1']$ ;
- 2) полученное уравнение  $1'$  умножаем на  $a_{21}$  и вычитаем его из второго уравнения;
- 3) уравнение  $1'$  умножаем на  $a_{31}$  и вычитаем его из 3-го уравнения и т. д.

Схематически это выглядит так:

$$\begin{array}{ccc} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{vmatrix} f_1 \\ f_2 \\ \dots \\ f_n \end{vmatrix} & \rightarrow & \begin{pmatrix} 1 & b_{12} & \dots & b_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{vmatrix} g_1 \\ f_2 \\ \dots \\ f_n \end{vmatrix} \rightarrow \\ \begin{pmatrix} 1 & b_{12} & \dots & b_{1n} \\ 0 & a'_{22} & \dots & a'_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{vmatrix} g_1 \\ f'_2 \\ f_3 \\ \dots \\ f_n \end{vmatrix} & \rightarrow & \begin{pmatrix} 1 & b_{12} & \dots & b_{1n} \\ 0 & a'_{22} & \dots & a'_{2n} \\ 0 & a'_{32} & \dots & a'_{3n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{vmatrix} g_1 \\ f'_2 \\ f'_3 \\ \dots \\ f_n \end{vmatrix} \rightarrow \\ \rightarrow & \rightarrow & \begin{pmatrix} 1 & b_{12} & b_{13} & \dots & b_{1n} \\ 0 & a'_{22} & a'_{23} & \dots & a'_{2n} \\ 0 & a'_{32} & a'_{33} & \dots & a'_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & a'_{n2} & a'_{n3} & \dots & a'_{nn} \end{pmatrix} \begin{vmatrix} g_1 \\ f'_2 \\ f'_3 \\ \dots \\ f'_n \end{vmatrix}. \end{array}$$

Следующий шаг аналогичен первому. Первое уравнение теперь трогать не будем, а со вторым и последующим поступаем как с заданной системой на первом шаге:

- 1) второе уравнение делим на  $a'_{22}$  ( $[2'] \rightarrow [2'']$ );
- 2) полученное уравнение  $2''$  умножаем на  $a'_{31}$  и вычитаем его из третьего уравнения и т. д. (всего  $n-2$  комбинации).

В результате получим:

$$\begin{pmatrix} 1 & b_{12} & b_{13} & \dots & b_{1n} \\ 0 & 1 & b_{23} & \dots & b_{2n} \\ 0 & 0 & a''_{33} & \dots & a''_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & a''_{n3} & \dots & a''_{nn} \end{pmatrix} \begin{vmatrix} g_1 \\ g_2 \\ f''_3 \\ \dots \\ f''_n \end{vmatrix}.$$

**Продолжая эту процедуру, через  $n$  шагов мы получим:**

$$\begin{pmatrix} 1 & b_{12} & b_{13} & \dots & b_{1n} \\ 0 & 1 & b_{23} & \dots & b_{2n} \\ 0 & 0 & 1 & \dots & b_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{vmatrix} g_1 \\ g_2 \\ g_3 \\ \dots \\ g_n \end{vmatrix}. \quad (6)$$

## 6. Пример: трехдиагональная система

При решении краевых задач для дифференциальных уравнений часто приходится дело иметь с линейными системами, большая часть коэффициентов в которых равна нулю. Простейшая (и очень часто встречающаяся) система такого рода имеет следующий вид:

$$\begin{array}{rcl} b_0 x_0 + c_0 x_1 & = & f_0 \\ a_1 x_0 + b_1 x_1 + c_1 x_2 & = & f_1 \\ a_2 x_1 + b_2 x_2 + c_2 x_3 & = & f_2 \\ \dots\dots\dots & & \\ a_n x_{n-1} + b_n x_n & = & f_n. \end{array} \quad (10)$$

В матрице коэффициентов этой системы отличны от 0 лишь коэффициенты, стоящие на главной диагонали и 2 соседних с ней:

$$\begin{pmatrix} b_0 & c_0 & 0 & 0 & \dots & 0 & 0 \\ a_1 & b_1 & c_1 & 0 & \dots & 0 & 0 \\ 0 & a_2 & b_2 & c_2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & a_n & b_n \end{pmatrix}. \quad (11)$$

Ясно, что такую систему глупо решать по стандартной программе: машина будет большую часть времени оперировать с нулями. Но сам метод исключения здесь вполне пригоден.

Особенно простые и красивые формулы получаются, если можно использовать простейшую схему метода исключения (без выбора главного элемента). Чтобы их вывести, заметим сначала, что на каждом шаге исключения нам придется исключать определенное неизвестное  $x_i$  лишь из одного уравнения. По окончании процесса исключения мы получим треугольную систему очень специального вида:

$$\begin{pmatrix} 1 & p_0 & 0 & \dots & 0 \\ 0 & 1 & p_1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{vmatrix} r_0 \\ r_1 \\ \vdots \\ r_n \end{vmatrix} \quad (12)$$

Из этой системы легко последовательно (рекуррентно) найти все  $x_i$ :

$$x_n = r_n; \quad x_k = r_k - \rho_k x_{k+1} \quad (k = n-1, \dots, 0). \quad (13)$$

Теперь проследим, как возникают коэффициенты  $p_k$  и  $r_k$  в про-



цессе исключения. На  $k$ -ом шаге процесса исключения картина такова. Первые  $k-1$  уравнений уже в окончательном виде,  $k$ -ое уравнение изменилось при исключении  $x_{k-1}$ . Все остальные — не тронуты:

$$\left( \begin{array}{ccccccc} 1 & p_0 & 0 & \dots & 0 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & p_{k-1} & 0 & \dots \\ 0 & 0 & \dots & 0 & b'_k & c_k & 0 \dots \\ 0 & 0 & \dots & 0 & a_{k+1} & b_{k+1} & c_{k+1} \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{array} \right) \left| \begin{array}{c} r_0 \\ \dots \\ r_{k-1} \\ f'_k \\ f_{k+1} \\ \dots \end{array} \right|. \quad (14)$$

Дальше можно писать только две строки. Отнормировав уравнение  $[k']$ , получим

$$\left( \begin{array}{ccccccc} 0 & \dots & 1 & p_k & 0 & \dots & 0 \\ 0 & \dots & a_{k+1} & b_{k+1} & c_{k+1} & \dots & 0 \end{array} \right) \left| \begin{array}{c} r_k \\ f_{k+1} \end{array} \right|.$$

Исключив  $x_k$  из уравнения  $[k+1]$ , получим:

$$\left( \begin{array}{ccccccc} 0 & \dots & 1 & p_k & 0 & \dots & 0 \\ 0 & \dots & 0 & b'_{k+1} & c_{k+1} & \dots & 0 \end{array} \right) \left| \begin{array}{c} r_k \\ f'_{k+1} \end{array} \right|.$$

Здесь  $b'_{k+1} = b_{k+1} - a_{k+1} p_k$ ,  $f'_{k+1} = f_{k+1} - a_{k+1} r_k$ . Поделив на  $b'_{k+1}$ , получим

$$p_{k+1} = \frac{c_{k+1}}{b_{k+1} - a_{k+1} \cdot p_k}; \quad r_{k+1} = \frac{f_{k+1} - a_{k+1} \cdot r_k}{b_{k+1} - a_{k+1} \cdot p_k}. \quad (15)$$

Итак, решение системы (10) простейшим методом исключения сводится к следующему:

### I этап

Находятся рекуррентно в порядке  $k=0, 1, \dots, n-1$   $p_{k+1}$  и  $r_{k+1}$  по формулам (15).

$$\text{При этом } p_0 = \frac{c_0}{b_0}, \quad r_0 = \frac{f_0}{b_0}.$$

### II этап

Находятся рекуррентно в порядке  $k=n-1, \dots, 1, 0$  неизвестные  $x_k$  по формуле (13)

$$x_k = r_k - p_k x_{k+1}. \quad (16)$$

При этом  $x_n = r_n$ . Формулы (15) и (16) очень часто употребляются и носят выразительное название «прогонка» (I этап — «прямая прогонка», II этап — «обратная прогонка»). Различная «направленность» (по индексам) этих двух этапов не есть, конеч-

но, особое свойство этих формул — это общее свойство метода исключения. А вот простота вычислений специфична для системы (10). Заметьте, что с ростом  $n$  количество действий здесь растет пропорционально  $n$  (а не  $\sim \frac{1}{3}n^3$ , как в общем случае).

Поэтому без труда решаются трехдиагональные системы даже при  $n \sim 100\text{—}200$ . В заключение еще раз напомним, что прогонка не есть какой-то новый способ решения. Это есть лишь удобная формульная запись самого простого варианта метода исключения.

## 7. Метод исключения и определители

Идея исключения неизвестных очень естественна и элементарна. Поэтому особенно интересно взглянуть на метод исключения с более «ученой» точки зрения\*. Вы знаете, что в алгебре существуют универсальные формулы для решения любой невырожденной системы линейных уравнений. По этим формулам (называемым формулами Крамера) каждое неизвестное представляется в виде отношения двух определителей  $n$ -го порядка:

$$x_k = \Delta_k / \Delta. \quad (17)$$

В свою очередь определитель  $n$ -го порядка может быть задан явной формулой. Точный вид ее для нас сейчас не важен. Важно лишь следующее: в этой формуле  $n!$  ( $= 1 \cdot 2 \cdot \dots \cdot n$ ) слагаемых, а каждое слагаемое состоит из  $n$  множителей. Что произойдет, если решение линейной системы находить прямо по формулам Крамера и определители при этом считать «в лоб» по задающим их формулам?

Всего нужно вычислить  $n+1$  определитель. Вычисление каждого из них требует  $n \cdot n!$  умножений. Итого  $(n+1)n \cdot n! \approx (n+2)! \cdot n!$  умножений. Прикинем, что это означает, скажем, для  $n=20$ . Воспользовавшись приближенной формулой  $k! \approx \sqrt{2\pi k} (k/e)^k$ , получим:  $22! > (22/e)^{22} > 8^{22} > 10^{19}$ . Предположим, что мы располагаем машиной, делающей  $10^5$  арифметических действий в секунду. Тогда нам понадобится более  $10^{14}$  секунд или более 3 миллионов лет непрерывной работы\*\*. Реально, по стандартной программе система 20 порядка решается за доли секунды! Каким образом получается, что прекрасные общие формулы непригодны для практического использования? На этот вопрос я не могу дать ясного ответа. Замечу только следующее. По первоначальному смыслу неизвестные (и уравнения)

\* Всерьез алгебранескую структуру метода исключения мы обсудим позже, во 2 концентре. Здесь же затрагивается лишь арифметическая сторона дела (число необходимых арифметических операций).

\*\* 1 год  $\approx 3 \cdot 10^7$  сек.

в системе линейных уравнений равноправны. Формулы Крамера сохраняют это равноправие (при перестановке неизвестных соответствующие определители могут лишь поменять знак). При решении системы методом исключения мы устанавливаем некоторый вполне определенный порядок неизвестных. Нарушив имевшуюся симметрию, мы чрезвычайно выигрываем в скорости!

Решив систему по методу исключения, мы тем самым (косвенно) вычислим отношение определителей (17). Легко сообразить, что попутно можно найти и каждый из них в отдельности. Самый интересный — определитель системы  $\Delta^*$ . Вспомним, что определитель не меняется, если к одной из его строк добавляется другая, умноженная на любое число. Значит, в процессе исключения мы меняем определитель системы лишь при нормировке уравнений: при делении строки на очередной ведущий коэффициент определитель делится на то же число \*\*. В полученной треугольной системе определитель равен произведению диагональных элементов, т. е. 1. Обозначим ведущий элемент  $k$ -го шага  $c_k$  (в простейшей схеме исключения  $c_1 = a_{11}$ ,  $c_2 = a_{22}$  и т. д.).

Тогда  $\frac{\Delta}{c_1 \cdot c_2 \cdot \dots \cdot c_n} = 1$  или  $\Delta = c_1 \cdot c_2 \cdot \dots \cdot c_n$ . Значит, мы можем попутно в процессе исключения найти определитель  $\Delta$  — для этого нужно  $\approx \frac{1}{3} n^3 + \frac{1}{2} n^2$  умножений и делений. Несмотря на простоту, вывод этот все-таки удивителен. Подумайте сами. Определитель, скажем, 10-го порядка

$$\Delta = \begin{vmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,10} \\ \dots & \dots & \dots & \dots \\ a_{10,1} & a_{10,2} & \dots & a_{10,10} \end{vmatrix}$$

есть функция 100 независимых переменных  $a_{i,k}$ . Формула, задающая эту функцию в явном виде, содержит более 3 миллионов слагаемых, в каждом из которых перемножаются 10 из 100 переменных. Тем не менее значение этой функции при заданных значениях аргументов можно вычислить менее, чем за 1000 арифметических действий. Попробуйте, глядя на разложение определителя, до этого додуматься!

## 8. Решение системы линейных уравнений и обращение матриц

Теперь мы немного заглянем на более высокий уровень и сопоставим задачу решения системы линейных уравнений

$$Ax = f \tag{18}$$

\* Т.е. определитель, составленный из коэффициентов системы.

\*\* Если уравнения переставляются, то нужно будет еще учесть изменение знака  $\Delta$  при перестановке.

с задачей обращения матрицы  $A$ . Заметим, прежде всего, что при известной обратной матрице  $A^{-1}$  решение (18) есть:  $x = A^{-1}f$ . Умножение матрицы на вектор требует немного вычислений. Значит, зная  $A^{-1}$ , мы можем решить систему (18) с любой правой частью  $f$  сравнительно легко.

Заметим дальше, что столбцы обратной матрицы сами являются решениями системы (18) при специальном выборе правых частей. Скажем, первый столбец отвечает вектору  $f = (1, 0, \dots, 0)$  — это есть прямое следствие равенства  $AA^{-1} = I$  ( $I$  — единичная матрица).

Таким образом, знание  $A^{-1}$  эквивалентно знанию решений системы (18) при  $n$  различных правых частях. Поэтому на первый взгляд кажется, что обращение матрицы  $A$  должно быть в  $n$  раз более трудоемким, чем решение системы (18). Небольшое размышление показывает, однако, что это не так. Дело в том, что основная работа в процессе исключения производится над коэффициентами системы и не зависит от правых частей. В результате оказывается, что обращение матрицы лишь в фиксированное число раз — приблизительно вдвое — более трудоемко, чем решение отдельной системы. Поэтому, если нужно решить одну систему несколько раз ( $>3-4$ ) при разных правых частях, может оказаться разумным использовать обратную матрицу.

Алгоритм обращения матрицы легко извлечь из того же метода исключения, и мы позже это сделаем. А пока заметим, что число умножений  $N \sim n^3$  для обращения матрицы с некоторой точки зрения является совершенно удивительным. В самом деле. Рассмотрим произведение двух матриц  $n$ -го порядка  $C = AB$ . Для нахождения элементов  $C$  согласно определению нужно произвести  $n^3$  умножений — по  $n$  на каждый из элементов  $c_{ik}$ :

$$c_{ik} = \sum_{a=1}^n a_{ia} b_{ak}.$$

Интуитивно обращение матрицы кажется гораздо более трудной задачей, чем перемножение двух матриц. Между тем оно требует не большого количества действий! Стоит все же реабилитировать нашу интуицию. Обращение матриц, конечно, *есть* более сложная задача, но в *алгоритмическом* смысле. Алгоритмы здесь нетривиальны, для данной матрицы не все алгоритмы пригодны и т. д. Таким образом, наша интуитивная настороженность при обращении матриц — правильна. Но не следует все же забывать и вторую сторону дела: обращение матрицы невысокого порядка — не столь уж трудоемкая процедура.

1) (к п. 1). Убедимся, что после  $k$  шагов метода исключения очередное неизвестное  $x_{k+1}$  обязательно входит в одно из оставшихся  $n-k$  уравнений. Для этого нам понадобится следующее утверждение из теории систем линейных уравнений. Если система при заданных правых частях имеет единственное решение, то она обязательно имеет решение при любых правых частях.

Предположим теперь, что  $x_{k+1}$  не входит в (преобразованные) уравнения  $[k+1], \dots, [n]$ . Неизвестные  $x_1, x_2, \dots, x_k$  из этих уравнений уже исключены. Стало быть, мы имеем систему  $n-k$  уравнений с  $n-k-1$  неизвестным. Ясно, что такая (переопределенная) система не может иметь решения при *любых* правых частях (легко дать этому и формальное доказательство). Полученное противоречие и доказывает утверждение.

2) (к п. 3). Приведенный в тексте модельный пример хорош лишь своей простотой. Однако ситуация, когда потеря точности происходит за счет одного очень малого коэффициента — исключительна. Гораздо чаще можно получить неверный ответ за счет наличия многих умеренно малых ведущих коэффициентов. Примеры такого рода сложнее и мы построим их позже.

3) (к п. 5). Выбор главного элемента по столбцу (т. е. большего из  $n$  чисел) требует на первом шаге  $\sim n$  арифметических и логических действий. Всего необходимое количество действий имеет порядок  $n^2$  и уже при  $n \sim 5-6$  замедление, связанное с выбором главного элемента, становится несущественным\*.

4) (к п. 8). Некоторой сенсацией была недавняя работа Strassen'a \*\*, где был предложен алгоритм перемножения матриц, требующий  $\sim n^\alpha$ ,  $\alpha < 3$  действий. На сегодня открыт вопрос: каково минимальное  $\alpha$ ? Нельзя ли его сделать близким к 2? Независимо от практической ценности подобных алгоритмов их изучение очень интересно для теории.

5) Некоторые литературные указания.

Геометрические аспекты решения систем линейных уравнений (в смысле геометрии  $n$ -мерного пространства), в частности, вопрос о правильной нормировке, рассмотрены в книге:

К. Ланцош. Практические методы прикладного анализа. Физматгиз, 1961.

Много различных методов для решения систем линейных урав-

---

\* Иногда на каждом шаге используют выбор наибольшего элемента во всей матрице. Такой выбор требует  $\sim \frac{1}{3} n^3$  арифметических действий (скажем, вычитаний). Это уже по порядку сравнимо с основными действиями процесса исключения.

\*\* V. Strassen, Gaussian elimination is not optimal. Numer. Math. 13, 354—356, 1969. Переведено в сб. «Математика», № 14: 3, 1970.

нений, обращения матриц и нахождения собственных значений описаны в книгах:

*В. В. Воеводин.* Численные методы алгебры. М., «Наука», 1966.

*Д. К. Фаддеев, В. Н. Фаддеева.* Вычислительные методы линейной алгебры. М., Физматгиз, 1963.

Некоторые идеи, обсуждающиеся в этих книгах, мы рассмотрим позже.

## ЛЕКЦИЯ 3

### НЕЛИНЕЙНЫЕ УРАВНЕНИЯ. МЕТОД ПОСЛЕДОВАТЕЛЬНЫХ ПРИБЛИЖЕНИЙ

В этой теме есть две основные идеи:

1) идея линеаризации, т. е. приближенной замены нелинейной системы уравнений — линейной;

2) идея последовательных приближений: искомое решение нелинейной задачи ищется как предел последовательности, члены которой получаются друг из друга по заданному правилу.

Эти две фундаментальные идеи математического анализа приобретают в вычислительной математике особое значение, проникая почти во все ее разделы.

#### 1. Локальная постановка задачи

Мы будем заниматься решением системы уравнений:

$$\begin{cases} f_1(x_1, \dots, x_n) = 0 \\ \dots \\ f_n(x_1, \dots, x_n) = 0 \end{cases} \quad \text{или} \quad f(x) = 0.$$

Мы убедимся в том, что последовательные приближения сходятся к интересующему нас решению, *если* исходное значение  $x^{(0)}$  выбрано достаточно близко к решению. Далее мы будем заниматься именно такой «локальной» задачей.

Что касается нахождения подходящего *начального* приближения, — исходного для метода последовательных приближений, то для такой задачи не существует универсальных алгоритмов, и она может иногда быть очень трудной. Мы начнем со случая одного уравнения — многое можно понять уже в этом простом случае.

## Линеаризация

Итак, рассмотрим уравнение  $f(x)=0$ , и пусть  $a$  — его решение. Пусть  $x^{(0)}$  — приближенное значение решения, найденное графически или каким-нибудь другим способом. Требуется найти  $a$  с заданной точностью. Основная идея, используемая для решения такой локальной задачи, состоит в следующем: всякая функция  $f(x)$  локально, на небольшом промежутке изменения аргумента, близка к линейной функции\*.

Если приближенно заменить  $f(x)$  на линейную функцию  $l(x)=kx+b$  в окрестности  $x^{(0)}$  (содержащей  $a$ ), то можно надеяться, что корень  $x^{(1)}$  линейного уравнения  $l(x)=0$  будет мало отличаться от  $a$ . Его мы и примем за приближенное значение корня уравнения  $f(x)=0$ .

### 3. Метод Ньютона

Выражение « $l(x)$  мало отличается от  $f(x)$ », не является вполне определенным.

Как бы ни был уточнен его смысл, подходящих линейных функций существует бесконечно много и строить их можно по-разному.

Наиболее универсальным (одинаково применимым как к случаю одного уравнения, так и к случаю системы) является способ, предложенный Ньютоном. По методу Ньютона в качестве  $l(x)$  выбирают первые два члена в формуле Тэйлора для  $f(x)$  в точке  $x^{(0)}$ :

$$l(x) = f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)}).$$

Уравнение  $l(x)=0$  имеет решение  $x^{(1)} = x^{(0)} - f(x^{(0)})/f'(x^{(0)})$ . Приняв  $x^{(1)}$  за исходное значение, можно аналогично найти  $x^{(2)}$ . Повторяя этот процесс, получим последовательность  $x^{(1)}, x^{(2)}, \dots, x^{(s)}, \dots$ . При этом

$$x^{(s+1)} = x^{(s)} - \frac{f(x^{(s)})}{f'(x^{(s)})} = g(x^{(s)}); \quad g(x) = x - \frac{f(x)}{f'(x)}. \quad (1)$$

Метод Ньютона имеет простую геометрическую интерпретацию. В самом деле, прямая  $y=l(x)$ , где  $l(x)=f_0+f'_0(x-x^{(0)})$ , есть касательная к графику  $y=f(x)$ , а значение  $x^{(1)}$  есть абсцисса точки пересечения этой касательной с осью  $x$ . Поэтому метод Ньютона для одного уравнения называют иногда методом касательных.

---

\* Более точно — всякая дифференцируемая функция. Заметьте, что это та самая идея, которая лежит в основе дифференциального исчисления.



#### 4. Поведение последовательности $x^{(s)}$

Будут ли значения  $x^{(1)}, x^{(2)}, \dots, x^{(s)}, \dots$ , полученные по методу Ньютона, неограниченно приближаться к корню  $a$ ? Это зависит от того, насколько удачно выбрано  $x^{(0)}$ . Для случая одного уравнения, которым мы пока занимаемся, представление о поведении последовательности  $x^{(s)}$  можно получить из рассмотрения рисунков. Типичные случаи таковы (рис. 1—5).

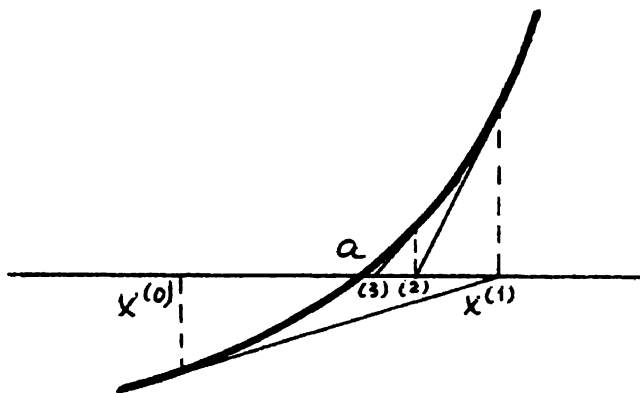


Рис. 1

Функция монотонна и выпукла в одну сторону (на рисунке  $f' > 0$ ,  $f'' > 0$ ). Последовательность  $x^{(s)} \rightarrow a$  при *любом*  $x^{(0)}$ . Если  $x^{(0)} > a$ , то последовательность монотонно убывает. При  $x^{(0)} < a$ , получим  $x^{(1)} > a$ ; последующие  $x^{(s)}$  приближаются к корню справа.

Случай 1 редко имеет место при всех  $x$ , но в *некоторой окрестности* корня он является нормальным: если  $f'(a) \neq 0$  и  $f''(a) \neq 0$ , то на небольшом отрезке около  $a$   $f'$  и  $f''$  сохраняют знак.

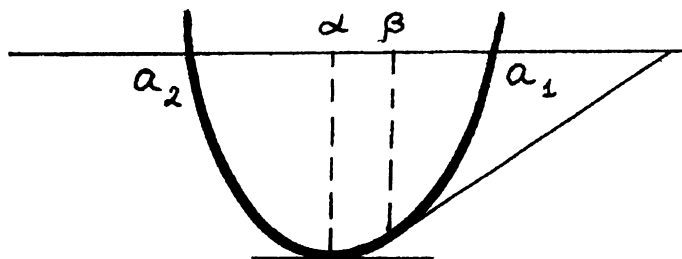


Рис. 2

Здесь при  $x^{(0)} > a$  (т. е. при  $f'(x^{(0)}) > 0$ ) последовательность  $x^{(s)}$  сходится к правому корню — это уже рассмотренный случай 1. При  $x^{(0)} < a$   $x^{(s)}$  сходятся к левому корню. Таким обра-

зом, есть зона притяжения правого корня и зона притяжения левого корня. Границей между ними служит точка  $a$ , где  $f'(a)=0$  (в самой этой точке метод Ньютона отказывает!). Заметим, что если  $x^{(0)}$  близко к  $a$  (на рисунке  $x^{(0)}=\beta$ ), то  $x^{(1)}$  будет отстоять далеко от корня — дальше, чем  $x^{(0)}$ . Все последующие  $x^{(s)}$  будут приближаться к корню (рис. 2).

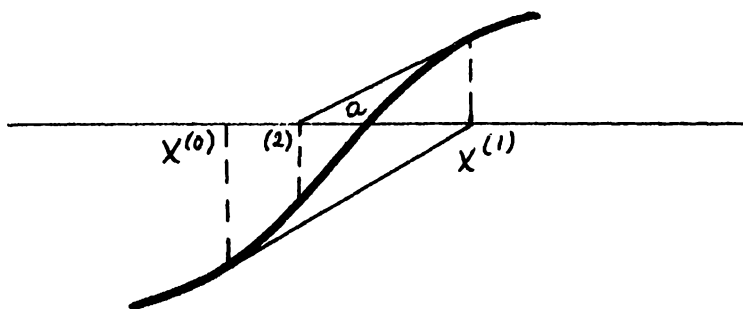


Рис. 3,а

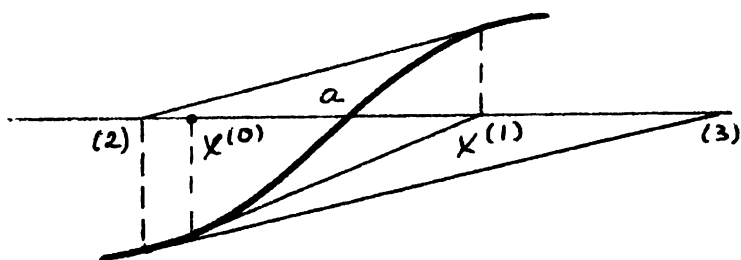


Рис. 3,б

График  $y=f(x)$  имеет перегиб при  $x=a$ , т. е.  $f''(a)=0$ . При  $x^{(0)}$ , близком к  $a$ , последовательность  $x^{(s)}$  будет колебаться вокруг  $a$ , все более приближаясь к корню. Если  $x^{(0)}$  выбрать подальше, то последовательность  $x^{(s)}$  может совсем не иметь предела (рис. 3,а,б).

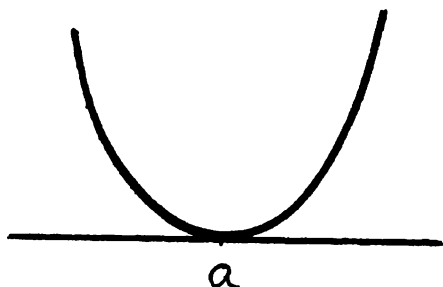


Рис. 4

Двукратный корень:  $f(a) = f'(a) = 0$ ;  $f''(a) \neq 0$ . Здесь будет сходимость для любого  $x^{(0)}$  (в области, где  $f''$  не меняет знак). Забегая вперед, заметим, что сходимость здесь медленнее, чем в случае простого корня (рис. 4).

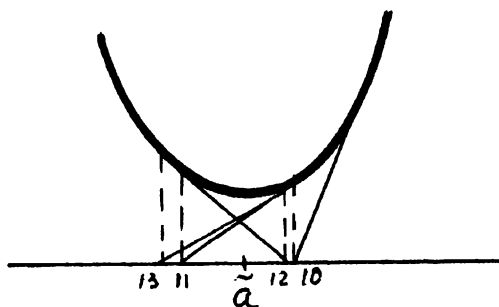


Рис. 5

В вычислительной практике встречаются и такие случаи: был кратный корень, но уравнение слегка возмутилось и корень исчез. В таком случае метод Ньютона дает последовательность, которая (с некоторого номера) колеблется и, конечно, нигде не сходится. При этом с некоторого момента становятся приблизительно равными все четные и все нечетные члены последовательности (скажем,  $x_{10} \approx x_{12} \approx x_{14}, \dots, x_{11} \approx x_{13} \approx x_{15}, \dots$ ) (рис. 5).

## 5. Алгебраические уравнения

В случае, когда  $f(x)$  — многочлен, уравнение  $f(x) = 0$  называется алгебраическим. Для алгебраического уравнения обычно бывает интересна задача в целом: нахождение *всех* корней, установление области, где лежат все корни и т. д. Для решения таких задач существуют специальные методы, использующие алгебраичность уравнения. Сейчас мы ими заниматься не будем и заметим только вот что. Если  $f(x)$  — многочлен, то  $g(x)$  в (1) — рациональная функция.

Таким образом, при решении алгебраического уравнения любой степени по методу Ньютона вычисления сводятся лишь к рациональным операциям (сложению, вычитанию, умножению и делению). Извлекать корни (т. е. находить  $\sqrt{c}$ ,  $\sqrt[3]{c}$ , ...) при этих вычислениях не требуется. Наоборот, метод Ньютона можно применить для приближенного вычисления корней.

Например, вычисление  $\sqrt{c}$  ( $c > 0$ ) есть нахождение положительного решения уравнения  $x^2 - c = 0$ . Формула Ньютона для этого уравнения выглядит так:

$$x^{(s+1)} = \frac{1}{2} \left( x^{(s)} + \frac{c}{x^{(s)}} \right). \quad (2)$$

Каково бы ни было  $x^{(0)} > 0$ , процесс сходится (это случай, изображенный на рис. 2). Вычислительные машины обычно не пользуются таблицами: значения квадратных корней вычисляются каждый раз, когда они нужны, как раз по формуле (2) \*.

## 6. Огрубленный метод Ньютона

Способ Ньютона можно несколько изменить с тем, чтобы избежать вычисления  $f'(x)$  на каждом шаге. (Это оказывается особенно полезным при решении системы уравнений). Именно, если  $x^{(0)} = c$  достаточно близко к корню, то можно на всех шагах вместо  $f'(x^{(s)})$  использовать  $f'(c)$ .

Последовательные приближения будут вычисляться в этом случае по формуле:

$$x^{(s+1)} = g_c(x^{(s)}); \quad g_c(x) = x - \frac{f(x)}{f'(c)}. \quad (3)$$

## 7. Метод Ньютона как метод итераций

Последовательные приближения по методу Ньютона находятся по формуле  $x^{(s+1)} = g(x^{(s)})$ . Если последовательность  $x^{(s)}$  сходится к  $a$ , то, переходя в обеих частях этого равенства к пределу, получим  $a = g(a)$ , т.е.  $a$  является корнем уравнения  $x = g(x)$ . Легко проверяется, что это уравнение эквивалентно исходному  $f(x) = 0$ , как для точного, так и для огрубленного метода Ньютона \*\*.

Таким образом, мы можем взглянуть на метод Ньютона со следующей общей точки зрения. Исходное уравнение  $f(x) = 0$  заменяется эквивалентным ему уравнением специального вида:  $x = g(x)$ . Это новое уравнение мы пытаемся решить «итерациями»:  $x^{(s+1)} = g(x^{(s)})$ .

Специфика метода проявляется в выборе  $g$  (для точного метода Ньютона одна функция  $g$ , для огрубленного — другая и т.д.). Сама процедура итераций во всех случаях одинакова. Разумно поэтому рассмотреть вопрос о сходимости итераций отдельно.

## 8. Достаточное условие сходимости итераций

**Т е о р е м а.** Предположим, что уравнение  $x = g(x)$  имеет решение  $x = a$ .

Если

$$|g'(x)| \leq q < 1, \quad (4)$$

\* Формула (2) последовательных приближений для  $\sqrt{c}$  была известна еще древним грекам.

\*\* Это утверждение нуждается в некотором уточнении, если уравнение  $f(x) = 0$  имеет «кратный» корень  $x$ . ( $f(x_*) = f'(x_*) = 0$ ).

то при произвольном  $x^{(0)}$  последовательность  $x^{(s+1)} = g(x^{(s)})$  сходится к  $a$ .

**Уточнение.** Если условие (4) выполнено в окрестности корня  $x=a$ , то теорема остается верной, если  $x^{(0)}$  лежит в этой же окрестности.

**Доказательство теоремы.** Оценим разность  $x^{(s+1)} - a$ . По формуле конечных приращений (т. к.  $a = g(a)$ )

$$g(x^{(s)}) - g(a) = g'(\xi^{(s)})(x^{(s)} - a); \xi^{(s)} = a + \theta(x^{(s)} - a), \quad 0 \leq \theta \leq 1.$$

Таким образом

$$|x^{(s+1)} - a| \leq q |x^{(s)} - a| \leq q^{s+1} |x^{(0)} - a| \rightarrow 0 \quad \text{при } s \rightarrow \infty.$$

Теорема доказана.

**З а м е ч а н и е.** При доказательстве мы не только установили факт сходимости, но и оценили ее скорость. Именно, разность  $x^{(s)} - a$  стремится к 0 не медленнее, чем геометрическая прогрессия со знаменателем  $q$ :

$$|x^{(s)} - a| \leq \text{const} \cdot q^s.$$

В таких случаях говорят коротко о сходимости со скоростью геометрической прогрессии.

## 9. Добавления и уточнения к теореме о сходимости

1) Предположим, что  $g'(a) = \lambda \neq 0$  (и выполнено условие (4)). Тогда можно утверждать несколько больше, чем было доказано. Именно, скорость сходимости будет точно определяться числом  $\lambda$ :

$$x^{(s)} - a \sim k\lambda^s \quad (\text{т. е. } x^{(s)} - a = k\lambda^s + o(\lambda^s)).$$

В частности, при  $\lambda > 0$  мы будем (с некоторого номера  $s$ ) приближаться к корню с одной стороны. Если же  $\lambda < 0$ , то сходимость будет осциллирующей:  $x^{(s)}$  попеременно справа и слева от  $a$ . Заметим, что разности соседних приближений ведут себя аналогично:

$$\Delta x^{(s)} = x^{(s+1)} - x^{(s)} \sim k(\lambda - 1)\lambda^s = k_1\lambda^s.$$

Наблюдая за этими разностями, можно приближенно определить  $\lambda$ .

2) (после первого — очевидное замечание). Если  $g'(a) = 0$ , то  $x^{(s)} - a$  стремится к 0 быстрее, чем любая геометрическая прогрессия.

3) Если  $|g'(a)| > 1$  (строго больше!), то сходимости итераций заведомо нет. В случае, если  $|g'(a)| < 1$  (и  $g'(x)$  — непрерыв-

на), то в некоторой окрестности  $a$   $|g'(x)| \leq q < 1$  и теорема гарантирует сходимость.

Таким образом, не ясным остается лишь случай  $|g'(a)| = 1$ . Как и всегда, этот промежуточный случай тоньше. Итерации в этом случае могут сходиться, а могут и не сходиться. Но даже если сходимость есть, она очень медленная и практически мало интересна.

4) Основное условие теоремы — малость производной — можно заменить несколько более общим: для любых  $u$  и  $v$  должно выполняться неравенство

$$|g(u) - g(v)| \leq q|u - v|, \quad q < 1. \quad (5)$$

Если  $g$  имеет непрерывную производную, то (5) эквивалентно (4).

5) Если условие (4) или (5) выполнено при всех  $x$ , то легко доказать не только сходимость, но и *существование* единственного решения (в теореме существование решения предполагается заранее). Единственность доказывается совсем просто. Если  $a_1$  и  $a_2$  — два решения, то:

$$|a_1 - a_2| = |g(a_1) - g(a_2)| \leq q|a_1 - a_2| < |a_1 - a_2| \text{ — противоречие.}$$

**Последнее замечание.** Простота доказательства теоремы не делает ее тривиальной и второстепенной. Напротив, эта теорема играет большую роль и допускает важные обобщения.

## 10. Сходимость огрубленного метода Ньютона для некратного корня

Последовательные приближения для этого метода вычисляются по формуле

$$x^{(s+1)} = g_c(x^{(s)}), \quad \text{где} \quad g_c(x) = x - \frac{f(x)}{f'(c)}.$$

Чтобы применить теорему о сходимости итераций, вычислим  $g'_c(x)$ :

$$g'_c(x) = 1 - \frac{f'(x)}{f'(c)} = \frac{f'(c) - f'(x)}{f'(c)}.$$

Предположим теперь, что в точке  $a$  ( $a$  — решение уравнения  $f(x) = 0$ ) производная  $f'(a) = \lambda \neq 0$ . Тогда в некоторой окрестности этой точки  $|f'(x)| \geq \frac{1}{2} |\lambda| = k$ . Выберем теперь еще меньшую окрестность  $U(a)$  так, чтобы для любых  $x$  и  $\xi$ , лежащих

в  $U$ , было:  $|f'(x) - f'(\xi)| < \varepsilon^*$ . Если  $c$  выбрана в этой окрестности ( $c \in U$ ), то для  $x \in U$

$$|g'_c(x)| < \frac{\varepsilon}{k}.$$

Выбором достаточно малой окрестности  $U$  мы можем сделать  $q = \frac{\varepsilon}{k} < 1$  ( $k = \frac{1}{2} |f'(a)|$  — фиксировано!). В такой окрестности  $U(a)$  можно гарантировать сходимость огрубленного метода Ньютона \*\*. Обычно

$$g'_c(a) = \frac{f'(c) - f'(a)}{f'(c)} \neq 0$$

и огрубленный метод Ньютона сходится с «нормальной» скоростью — скоростью геометрической прогрессии. При этом чем меньше окрестность  $U$ , тем меньше  $q$ , и тем быстрее сходимость.

## 11. Сходимость метода Ньютона

В этом случае

$$g(x) = x - \frac{f(x)}{f'(x)}; \quad g'(x) = \frac{f(x) \cdot f''(x)}{(f'(x))^2}.$$

Если  $f(a) = 0$ ,  $f'(a) \neq 0$ , то  $g'(a) = 0$  — можно ожидать очень быстрой сходимости (см. п. 9). Действительно

$$|x^{(s+1)} - a| = |g(x^{(s)}) - g(a)| \leq C |x^{(s)} - a|^2, \quad (6)$$

где

$$C = \frac{1}{2} \max |g''(x)|. \quad (6')$$

Если обозначить  $x^{(s)} - a = \delta_s$ , то мы получим:  $|\delta_{s+1}| \leq C \delta_s^2$ . Грубо говоря, ошибка на каждом шаге возводится в квадрат. Эта характерная для метода Ньютона очень быстрая «ньютоновская» сходимость наблюдается на практике редко, поскольку точным методом Ньютона удастся воспользоваться в сравнительно небольшом числе случаев.

Замечательно, что точный метод Ньютона (для одного урав-

\* Здесь, как и раньше, мы предполагаем, что все встречающиеся функции непрерывны.

\*\* Легко оценить величину такой окрестности через  $\max |f''(x)|$ .

нения) оказывается достаточно сильным, чтобы обеспечить сходимость и в случае кратного корня\*.

Рассмотрим для примера случай двукратного корня:

$$f(a) = f'(a) = 0; \quad f''(a) = m \neq 0.$$

Тогда

$$f(x) = \frac{1}{2} m(x-a)^2 + o(x-a)^2$$

$$f'(x) = m(x-a) + o(x-a)$$

$$g(x) - a = \frac{1}{2} (x-a) + o(x-a).$$

Отсюда  $x^{(s+1)} - a \approx \frac{1}{2} (x^{(s)} - a)$  — имеет место сходимость со скоростью геометрической прогрессии  $\left(\frac{1}{2}\right)^s$ . Скорость сходимости не зависит от выбора функции  $f(x)$ !

## 12. Система двух уравнений

При переходе к системам любого числа уравнений нам не понадобится никаких новых идей. Чтобы не затемнять индексами прозрачное существо дела, рассмотрим сначала систему двух уравнений:

$$f_1(x, y) = 0, \quad f_2(x, y) = 0.$$

Предположим, что  $(x^{(0)}, y^{(0)})$  близко к решению этой системы  $(a, b)$ . Положим  $x = x^{(0)} + \Delta x$ ,  $y = y^{(0)} + \Delta y$ , подставим в систему и линеаризуем ее:

$$\begin{aligned} f_1(x^{(0)}, y^{(0)}) + \frac{\partial f_1}{\partial x}(x^{(0)}, y^{(0)}) \Delta x + \frac{\partial f_1}{\partial y}(x^{(0)}, y^{(0)}) \Delta y &= 0 \\ f_2(x^{(0)}, y^{(0)}) + \frac{\partial f_2}{\partial x}(x^{(0)}, y^{(0)}) \Delta x + \frac{\partial f_2}{\partial y}(x^{(0)}, y^{(0)}) \Delta y &= 0. \end{aligned} \quad (7)$$

Из этой системы двух линейных уравнений находим  $\Delta x$  и  $\Delta y$ ; затем:  $x^{(1)} = x^{(0)} + \Delta x$ ,  $y^{(1)} = y^{(0)} + \Delta y$ . Это — первое приближение метода Ньютона. Заменяя в (7)  $(x^{(0)}, y^{(0)})$  на  $(x^{(s)}, y^{(s)})$ , получим формулы для приращений на  $s$ -ом шаге метода Ньютона.

Если частные производные на всех шагах сохраняются такими, какими они были на первом шаге, получим снова видоизмененный (огрубленный) метод Ньютона.

---

\* В этом случае сама идея линеаризации в окрестности  $a$  кажется неприменимой: линейный член в формуле Тэйлора в точке  $a$  отсутствует!



Как и в случае одного уравнения, для нахождения подходящего нулевого приближения не существует универсального способа, но, конечно, для системы эта задача гораздо труднее.

К счастью, довольно распространен случай, когда решение систем нелинейных уравнений входит как составная часть в большую вычислительную задачу. Тогда на каждом этапе большой задачи приходится решать систему, мало отличающуюся от предыдущей. Решение предыдущей системы и является в этом случае подходящим начальным приближением для следующей системы.

### 13. Система уравнений. Метод Ньютона

Рассмотрим теперь общий случай. Пусть дана система уравнений:  $f_i(x_1, \dots, x_n) = 0$ ;  $i = 1, 2, \dots, n$ , или, коротко,  $f(x) = 0$  и пусть  $a$  — решение этой системы. Предположим, что  $x^{(0)}$  близко к  $a$ . Напишем  $x = x^{(0)} + \Delta x$  и будем искать  $\Delta x$ .

Уравнение  $f(x) = 0$  или  $f(x^{(0)} + \Delta x) = 0$ , мы заменяем по методу Ньютона на линейное уравнение (точнее линейную систему), оставив два члена в формуле Тэйлора:

$$f(x^{(0)} + \Delta x) \approx f(x^{(0)}) + f'(x^{(0)}) \cdot \Delta x.$$

Итак,  $\Delta x$  есть решение линейной системы

$$\frac{df}{dx}(x^{(0)}) \cdot \Delta x = -f(x^{(0)}) \quad (8)$$

или в полной записи

$$\sum_{k=1}^n \frac{\partial f_i}{\partial x_k}(x_1^{(0)}, \dots, x_n^{(0)}) \Delta x_k = -f_i(x_1^{(0)}, \dots, x_n^{(0)}). \quad (8a)$$

Найдя  $\Delta x$ , мы получим приближенное решение системы:

$$x^{(1)} = x^{(0)} + \Delta x.$$

Аналогично строятся следующие приближения. Таким образом, на каждом шаге метода Ньютона нужно решать систему  $n$  линейных уравнений. Как и раньше, можно загрузить метод, используя на всех шагах значение  $f'(c)$  вместо  $f'(x^{(s)})$ . Тогда на каждом шаге придется решать одну и ту же систему линейных уравнений (будут меняться лишь правые части). Такое видоизменение метода Ньютона обычно тем более полезно, чем больше  $n$  (число уравнений): точный метод Ньютона требует на

каждом шаге вычисления  $n^2 + n$  функций, а огрубленный (начиная со второго шага) — лишь  $n$  функций\*.

#### 14. О сходимости метода Ньютона для системы уравнений

Положение дел здесь, в общем, такое же, как для одного уравнения. Именно. Последовательные приближения, отвечающие точному или огрубленному методу Ньютона, всегда сходятся, если начальное приближение  $(x^{(0)})$  выбрано достаточно близко к решению. При этом метод Ньютона дает быструю, квадратичную сходимость, а огрубленный метод — сходимость со скоростью геометрической прогрессии. Для справедливости этих утверждений снова нужно предположить, что решение  $x = a$  является невырожденным: определитель

$$\det(f_{ik}) \neq 0, \quad f_{ik} = \frac{\partial f_i}{\partial x_k}(a).$$

(Для случая одного уравнения это есть условие  $f'(a) \neq 0$ ). Далее. Последовательные приближения по методу Ньютона снова можно рассматривать как решение итерациями эквивалентной системы. Именно. Обозначим решение линейной системы

$$f'(x) \cdot \Delta x = -f(x) \quad \text{так:} \quad \Delta x = -[f'(x)]^{-1} f(x).$$

Тогда мы можем записать последовательные приближения по методу Ньютона в виде, аналогичном формуле (1):

$$x^{(s+1)} = g(x^{(s)}); \quad g(x) = x - \left( \frac{df}{dx}(x) \right)^{-1} \cdot f(x) **. \quad (9)$$

#### 15. Итерации в системах

Итак, снова возникает вопрос о сходимости итераций:

$$x^{(s+1)} = g(x^{(s)}). \quad (10)$$

Качественный ответ может быть таким.

Для сходимости итераций (10) достаточно, чтобы частные производные  $\frac{\partial g_i}{\partial x_k}$  были малы. Существует несколько

---

\* Если необходимо вычислить несколько (скажем, больше трех) приближений, то при использовании огрубленного метода Ньютона можно еще «экономить» на решении линейных систем (см. предыдущую лекцию).

\*\* Буквально прочитанная формула (9) требует вычисления обратной матрицы и умножения ее на вектор. Подчеркнем, однако, еще раз, что необходимым является лишь решение линейной системы, а не обращение матрицы.

способов, как уточнить это утверждение. Один из вариантов дает следующая теорема.

### Теорема

Пусть известно:

- 1) уравнение  $x = g(x)$  имеет решение  $x = a$ ;
- 2) для любых 2 векторов  $u$  и  $v$  выполняются условия

$$\|g(u) - g(v)\| \leq q \|u - v\|, \quad q < 1. \quad (11)$$

(Здесь  $\|c\|$  — длина вектора  $c$ ). Тогда последовательность  $x^{(s)}$ , определенная формулой  $x^{(s+1)} = g(x^{(s)})$ , при любом  $x^{(0)}$  сходится к  $a$ .

### Уточнение

Как и для одного уравнения, достаточно потребовать выполнения условия (11) в некоторой окрестности  $a$  ( $\|x - a\| < \delta$ ). Утверждение теоремы остается верным, если  $x^{(0)}$  лежит в этой окрестности.

Доказательство — почти очевидно:

$$\|x^{(k+1)} - a\| = \|g(x^{(k)}) - g(a)\| \leq q \|x^{(k)} - a\| \leq C q^k \rightarrow 0.$$

Видно, что имеет место сходимость не хуже, чем со скоростью геометрической прогрессии; чем меньше  $q$ , тем быстрее будет сходимость.

### Замечание

Условие (11) имеет простой геометрический смысл. Будем рассматривать векторную функцию  $y = g(x)$  как *отображение*, переводящее каждый вектор  $x$  в некоторый вектор  $y$ . Тогда (11) означает, что при отображении  $g$  расстояние между векторами уменьшается; или, говорят, отображение  $g$  — *сжимающее*.

Посмотрим, как можно проверить условие (11). Для этого оценим  $g(u) - g(v)$  по координатам:

$$|g_i(u) - g_i(v)| \leq M_i \|u - v\|,$$

где  $M_i = \max \|\text{grad } g_i\|$  \*\*. Значит

$$\|g(u) - g(v)\| = \left[ \sum (g_i(u) - g_i(v))^2 \right]^{1/2} \leq (\sum M_i^2)^{1/2} \|u - v\|.$$

Если  $M = \max_i M_i$ , то<sup>†</sup>

$$\|g(u) - g(v)\| \leq M \sqrt{n} \|u - v\|.$$

\* Для одного уравнения условие (11) есть  $|g(u) - g(v)| \leq q |u - v|$ ; см. п. 9.

\*\* Действительно, производная функции по любому направлению не превосходит длины ее градиента.

Итак, если для всех  $g_i$   $\|\text{grad } g_i\| \leq q/\sqrt{n}$  ( $q < 1$ ), то гарантирована сходимость итераций  $x^{(s+1)} = g(x^{(s)})$  со скоростью геометрической прогрессии  $q^s$ . Величину  $\text{grad } g_i$  легко оценить через частные производные  $g_i^*$ ; мы получим тогда достаточное условие сходимости, требующее некоторой малости всех частных производных  $\frac{\partial g_i}{\partial x_k}$ . Для  $n=1$  оно совпадает с условием  $|g'(x)| \leq q < 1$ . Однако на этом аналогия с одним уравнением кончается! Для одного уравнения условие  $|g'(x)| < 1$  является по существу необходимым и достаточным для сходимости итераций (более точно: необходимо для сходимости условие  $|g'(x)| \leq 1$ ). Для системы это не так: итерации могут сходиться, несмотря на то, что некоторые из производных будут очень велики!

## 16. Процесс итераций в линейном приближении

Предполагая, что  $x^{(s)}$  мало отличается от  $a$ , заменим  $g(x)$  на линейную функцию  $l(x)$ . Тогда

$$x^{(s+1)} = g(x^{(s)}) \approx a + A(x^{(s)} - a), \quad a = \frac{dg}{dx}(a).$$

Обозначив  $r^{(s)} = x^{(s)} - a$ , получим (в том же приближении) простую формулу:

$$\bar{r}^{(s+1)} = A r^{(s)}$$

или

$$r^{(s)} = \underbrace{A \cdot A \cdot \dots \cdot A}_s \cdot r^{(0)} = A^s \cdot r^{(0)}.$$

Итак, в этом приближении вопрос сводится к поведению степеней матрицы  $A$ . Оказывается, поведение последовательности  $r^{(s)} = A^s r^{(0)}$  определяют собственные значения и собственные векторы матрицы  $A$ , а не непосредственно величина ее элементов. Вполне может случиться, что некоторые элементы  $A$  велики, но процесс итераций быстро сходится. Например, для матрицы

$$A = \begin{pmatrix} 0 & 100 & 1000 \\ 0 & 0 & 100 \\ 0 & 0 & 0 \end{pmatrix}$$

уже на третьем шаге получим  $\bar{r}^{(3)} = 0$  (у этой матрицы все три собственных значения равны 0). Возвращаясь к общему случаю, предположим для простоты, что матрица  $A$  имеет  $n$  ве-

---

\*  $\|\text{grad } g\| \leq \sqrt{n} \cdot \max_i \left| \frac{\partial g}{\partial x_i} \right|$ .

ественных собственных значений  $\lambda_1, \dots, \lambda_n$  и соответствующих им собственных векторов  $e^{(1)}, \dots, e^{(n)}$ . Возьмем некоторый исходный вектор  $r$  и разложим его по собственным векторам:  $r = \sum c_k e^{(k)}$ . Применим  $A$ :

$$Ar = \sum_k c_k A e^{(k)} = \sum_k \lambda_k c_k e^{(k)}.$$

Аналогично для  $A^s$ :

$$A^s r = \sum_{k=1}^n \lambda_k^s c_k e^{(k)}. \quad (12)$$

Формула (12) при всей ее простоте содержит богатую информацию. Из нее видно, в частности, что для стремления к нулю *достаточно* выполнения условия

$$|\lambda_k| < 1; \quad k = 1, 2, \dots, n. \quad (13)$$

Условие (13) является *необходимым*, если мы хотим, чтобы  $A^s r \rightarrow 0$  при *любом*  $r$ . (Действительно, при  $r = e^{(k)}$   $A^s r = \lambda_k^s e^{(k)}$ ). Можно показать, что условие (13) сохраняет свою силу и в общем случае, при наличии комплексных (и кратных) собственных значений. Более точно, справедлива следующая теорема.

Для того, чтобы итерации  $x^{(s+1)} = Ax^{(s)} + b$  при произвольном  $x^{(0)}$  стремились к пределу\*, необходимо и достаточно, чтобы для всех собственных значений матрицы  $A$  выполнялось неравенство  $|\lambda(A)| < 1$ .

### З а к л ю ч и т е л ь н ы е з а м е ч а н и я

#### Последовательные приближения для решения линейных систем

До сих пор мы говорили о последовательных приближениях для решения нелинейных систем, предполагая, что линейные системы мы можем, если понадобится, решить каким-нибудь *прямым* методом (типа метода исключения).

Однако очень большие системы линейных уравнений тоже часто приходится решать последовательными приближениями. Это есть отдельная большая тема для будущего. Пока я замечу только, что проблема состоит не в том, чтобы построить сходящийся процесс последовательных приближений\*\*, а в том, чтобы приближения сходились *достаточно быстро*.

Для случая, когда последовательные приближения сводятся

\* Очевидно, этот предел  $a$  будет решением системы  $a = Aa + b$ .

\*\* Такой процесс для любой (невыврожденной) линейной системы можно построить.

к простым итерациям  $x^{(s+1)} = Ax^{(s)} + b$ , скорость сходимости (при наличии полной системы собственных векторов) указывает формула (12). Именно, если  $|\lambda_k| \leq q < 1$ , то разность  $x^{(s)} - a$  убывает при  $s \rightarrow \infty$  не хуже геометрической прогрессии  $Cq^s$ \*. На самом деле, формула (12) дает больше: она указывает механизм сходимости. Поэтому часто эта же формула может подсказать, как ускорить сходимость. Ограничимся одним примером. Пусть все  $|\lambda_k| < 0,8$ , кроме  $|\lambda_1|$ , близкого к 1 (скажем,  $\lambda_1 = 0,9999$ ). Тогда после 20—40 итераций все компоненты  $r^{(s)}$ , кроме компоненты  $c_1 \lambda_1^s$  вдоль  $e^{(1)}$ , будут малы. При дальнейших итерациях мы будем очень медленно двигаться «по направлению  $e^{(1)}$ »:

$$x^{(s+1)} - x^{(s)} = r^{(s+1)} - r^{(s)} \approx (\lambda_1^{s+1} - \lambda_1^s) c_1 e^{(1)}.$$

Ясно, что это неразумно: нужно сразу совершить большой шаг в этом направлении. Практически можно поступить так. Возьмем вектор  $u = x^{(40)} - x^{(30)}$  и положим  $\tilde{x} = x^{(30)} + \alpha u$ ; постоянную  $\alpha$  выберем так, чтобы сделать минимальным  $x - (A\tilde{x} + b)$ . Один такой шаг может заменить несколько десятков или даже сотен итераций по формуле

$$x^{(s+1)} = Ax^{(s)} + b !$$

---

\* Это утверждение нуждается в уточнении в нескольких пунктах. Один из них:  $\|x^{(s)} - a\|$  в начале может даже возрастать.

## ЛЕКЦИЯ 4

### ПРИБЛИЖЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ И ИНТЕГРИРОВАНИЕ

Обе основные операции анализа — дифференцирование и интегрирование — содержат в себе предельный переход и, тем самым, требуют знания функции на целом отрезке. Нам нужны приближенные дискретные аналоги этих операций, такие, которые бы использовали лишь конечное число значений функции.

Традиционный способ получать такие формулы состоит в следующем. Берут функцию и строят многочлен, который в нескольких точках совпадает с этой функцией; такой многочлен называется интерполяционным, а сама процедура называется интерполированием\*. Затем вместо того, чтобы интегрировать или дифференцировать функцию, интегрируют или дифференцируют этот многочлен. Поскольку при построении интерполяционного многочлена используются только несколько значений функции, то окончательные формулы содержат только эти значения.

Таков традиционный путь. Если следовать этому пути, то нужно сначала развить теорию интерполирования многочленами. Это требует довольно много времени. Поэтому для начала мы поступим не так, а получим нужные нам простые формулы элементарными средствами.

Идея интерполирования тоже будет использована, но не в общем виде, а для нужных нам конкретных примеров. Вот таков план.

#### 1. Приближенное нахождение производной

Для того, чтобы получить простейшую приближенную формулу для производной, нужно знать только ее определение:

---

\* Более точно: интерполированием посредством многочленов заданной степени.

$$f'(x) = \lim_{h \rightarrow 0} \frac{1}{h} (f(x+h) - f(x)). \quad (1)$$

При малом  $h$  можно положить:

$$f'(x) \approx \frac{1}{h} (f(x+h) - f(x)). \quad (2)$$

Это и есть простейшая приближенная формула.

В определении (1)  $h$  может принимать значения обоих знаков.

В дискретной записи принято обозначать через  $h$  положительное число, так что можно написать еще одну формулу:

$$f'(x) \approx \frac{1}{h} (f(x) - f(x-h)), \quad h > 0. \quad (2')$$

Какую ошибку мы совершаем, заменяя производную разностным отношением по формуле (2)? Это легко сообразить. Напишем:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(\xi) \quad (x < \xi < x+h).$$

Отсюда

$$\frac{m_2}{2} h \leq \left| f'(x) - \frac{f(x+h) - f(x)}{h} \right| \leq \frac{M_2}{2} h,$$

где  $m_2 = \min |f''(x)|$ ,  $M_2 = \max |f''(x)|$ . При  $h \rightarrow 0$  ошибка стремится к нулю со скоростью  $h$  или, как говорят, формула (2) имеет *первый порядок точности*. Сложением формул (2) и (2') получается симметричная формула:

$$f'(x) \approx \frac{1}{2h} (f(x+h) - f(x-h)). \quad (3)$$

Формула (3), как легко проверить, точнее формулы (2), а именно, ошибка здесь имеет порядок  $h^2$  — это есть формула второго порядка точности\*.

Это увеличение точности получилось, так сказать, «задаром» — только за счет симметрии. Это случается очень часто. В этой лекции мы будем иметь еще несколько случаев, когда симметрия формулы повышает ее точность.

Таким образом, как правило, нужно пользоваться формулой (3) — она точнее. У этой формулы есть лишь один очевидный недостаток: она неприменима на конце отрезка  $(a, b)$ , на котором задана функция.

---

\*Ошибка формулы (3) не превосходит  $\frac{1}{6} M_3 h^2$ , где  $M_3 = \max |f'''(x)|$ .



Часто тем не менее бывает нужно сосчитать производную в концевой точке более точно, чем по формуле (2). Я покажу сейчас, как это можно сделать

$$\begin{array}{c} a+2h \\ \hline a \quad a+h \quad \quad \quad b \end{array}$$

Рассмотрим наряду с точкой  $a$  еще 2 точки:  $a+h$  и  $a+2h$ . Обозначим значения функции в этих точках  $f_0, f_1, f_2$  и построим формулу следующего вида:

$$f'(a) \approx \alpha f_0 + \beta f_1 + \gamma f_2.$$

Коэффициенты  $\alpha, \beta, \gamma$  подберем так, чтобы получить второй порядок точности. Воспользуемся опять формулой Тэйлора:

$$f_1 = f_0 + hf'_0 + \frac{h^2}{2} f''_0 + \dots; \quad f_2 = f_0 + 2hf'_0 + 2h^2 f''_0 + \dots$$

Отсюда

$$\alpha f_0 + \beta f_1 + \gamma f_2 = (\alpha + \beta + \gamma) f_0 + (\beta + 2\gamma) hf'_0 + \left(\frac{1}{2} \beta + 2\gamma\right) h^2 f''_0 + O(h^3).$$

Потребуем, чтобы:

$$\alpha + \beta + \gamma = 0, \quad (\beta + 2\gamma)h = 1, \quad \frac{1}{2} \beta + 2\gamma = 0.$$

Найдя из этой системы  $\alpha, \beta, \gamma$ , получим нужную формулу:

$$f'_0 \approx \frac{1}{2h} (-f_2 + 4f_1 - 3f_0)$$

или

$$f'(a) = \frac{1}{2h} (-f(a+2h) + 4f(a+h) - 3f(a)) + O(h^2). \quad (4)$$

Таким же способом неопределенных коэффициентов можно было бы получить еще более точные формулы, но в большинстве случаев для вычисления первой производной достаточно формул (2), (3) и (4).

**З а м е ч а н и е.** То, что формула (3) точнее, чем (2), можно выразить еще несколько иначе. Запишем разностное отношение

$$\frac{1}{h} (f(x+h) - f(x)) \quad \text{в виде} \quad \frac{1}{h} \left( f\left(\xi + \frac{h}{2}\right) - f\left(\xi - \frac{h}{2}\right) \right),$$

где  $\xi = x + \frac{h}{2}$ .

Тогда:

$$\frac{1}{h}(f(x+h) - f(x)) = f'\left(x + \frac{h}{2}\right) + O(h^2). \quad (5)$$

Таким образом, правая часть формулы (2) гораздо точнее дает  $f'\left(x + \frac{h}{2}\right)$ , чем  $f'(x)$ .

## 2. Вычисление второй производной

Чтобы получить приближенную формулу для второй производной, можно поступить так же, как при выводе формулы (4): а именно, записать выражение  $f''_0$  с неопределенными коэффициентами через  $f_0$ ,  $f_1$ ,  $f_2$  и подобрать эти коэффициенты. Если это проделать, то получится следующая формула:

$$f''_0 \approx \frac{1}{h^2}(f_2 - 2f_1 + f_0) \quad \text{или} \quad (6)$$

$$f''(x) \approx \frac{1}{h^2}(f(x+2h) - 2f(x+h) + f(x)).$$

Такая формула имеет первый порядок точности. Если эту формулу «сдвинуть», используя симметрично расположенные точки, то получится более точное выражение:

$$f''(x) \approx \frac{1}{h^2}(f(x+h) - 2f(x) + f(x-h)). \quad (7)$$

Формула (7) имеет 2 порядок точности и является самой употребительной формулой для приближенного вычисления второй производной. Если точности этой формулы нехватает (что бывает сравнительно редко), то применяются формулы, использующие большее число значений функции и имеющие 3 или 4 порядок точности. Такие формулы легко выводятся тем же способом неопределенных коэффициентов.

**З а м е ч а н и е.** В анализе вторая производная определяется посредством двух предельных переходов. Сначала находится

$$f'(x) = \lim_{h \rightarrow 0} \frac{1}{h}(f(x+h) - f(x))$$

и затем

$$f''(x) = \lim_{a \rightarrow 0} \frac{1}{a}(f'(x+a) - f'(x)).$$

Формула (7) (или (6)) позволяет совершить оба предельных перехода одновременно:

$$f''(x) = \lim_{h \rightarrow 0} \frac{1}{h^2}(f(x+h) - 2f(x) + f(x-h)).$$

### 3. Приближенное нахождение определенного интеграла

Для приближенного нахождения интеграла  $\int_a^b f(x)dx$  можно так же, как и для производной, исходить из определения. По определению интеграл есть предел интегральных сумм:

$$\int_a^b f(x)dx = \lim_{\delta \rightarrow 0} \sum f(\xi_i) \Delta x_i. \quad (8)$$

Здесь  $\Delta x_i = x_{i+1} - x_i$ ,  $\delta = \max(\Delta x_i)$ ,  $x_i \leq \xi_i \leq x_{i+1}$ . В анализе доказывается, что от того, как именно разбит отрезок  $(a, b)$  на отрезки  $(x_i, x_{i+1})$  и где выбраны  $\xi_i$  — предел не зависит\*.

Чтобы получить приближенное значение интеграла, можно не совершать предельный переход и записать при малом  $\delta$ :

$$\int_a^b f(x)dx \approx \sum f(\xi_i) \Delta x_i. \quad (9)$$

Остается большой произвол в том, какие взять отрезочки  $\Delta x_i$  и где взять точки  $\xi_i$ . Проще всего, все отрезки  $(x_i, x_{i+1})$  взять одинаковыми и  $\xi_i$  — в серединах этих отрезков. Обозначив

$\Delta x = h = \frac{b-a}{n}$ , можно написать:

$$\int_a^b f(x)dx \approx h \sum_{k=0}^{n-1} f(\xi_k), \quad \xi_k = a + kh + \frac{h}{2}. \quad (10)$$

Это и есть простейшая формула для приближенного вычисления интегралов, или простейшая *квадратурная формула*.

Формула (10) имеет простой геометрический смысл. Написав

$$\int_a^b f(x)dx = \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f(x)dx \quad (x_0=a, x_n=b),,$$

мы можем считать, что в (10) произведена замена  $\int_{x_k}^{x_{k+1}} f(x)dx$  на  $f(\xi_k) \cdot h$ . Это означает, что площадь криволинейной фигуры на каждом участке заменена площадью прямоугольника высоты  $f(\xi_k)$ .

Чтобы оценить ошибку формулы (10), напомним формулу Тэйлора в окрестности  $\xi_k$ :

$$f(x) = f(\xi_k) + f'(\xi_k)(x - \xi_k) + r(x)$$

---

\* В предположении, что  $f(x)$  — непрерывна.

$$|r(x)| \leq \frac{1}{2} M_2(x - \xi_k)^2;$$

$$M_2 = \max |f''(x)|.$$

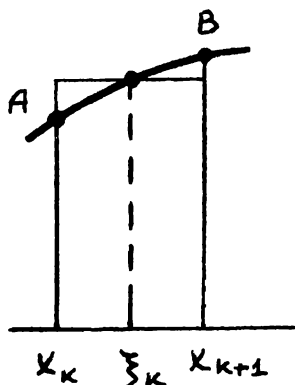
Теперь имеем:

$$\int_{x_k}^{x_{k+1}} f(x) dx = f(\xi_k)h + \int_{x_k}^{x_{k+1}} r(x) dx$$

(интеграл от второго слагаемого равен нулю в силу симметрии!).

Отсюда:

$$\left| \int_{x_k}^{x_{k+1}} f(x) dx - f(\xi_k) \cdot h \right| \leq \frac{M_2}{24} h^3.$$



Суммируя, получаем:

$$\left| \int_a^b f(x) dx - \sum_k f(\xi_k) \cdot h \right| \leq \frac{M_2}{24} (b-a) h^2. \quad (11)$$

Таким образом, при  $h \rightarrow 0$  ошибка при вычислении интеграла по формуле (10) для дважды дифференцируемой функции стремится к 0 со скоростью  $h^2$ , или как говорят, эта квадратурная формула имеет 2 *порядок точности*.

Часто вместо значений  $f$  в серединах отрезков используют полусумму значений в концах. Просуммировав по всем отрезкам длины  $h$ , получим:

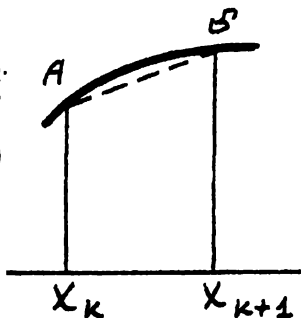
$$\int_a^b f(x) dx \approx \frac{h}{2} (f_0 + 2f_1 + 2f_2 + \dots + 2f_{n-1} + f_n); \quad f_k = f(x_k). \quad (12)$$

Эта формула называется обычно «формулой трапеций»: здесь на каждом участке площадь криволинейной фигуры  $\int_{x_k}^{x_{k+1}} f(x) dx$  заменяется на площадь трапеции:

$$\frac{1}{2} (f_k + f_{k+1}) \cdot h.$$

Формула трапеций имеет тот же, второй-порядок точности и (при малых  $h$ ) дает приблизительно вдвое большую ошибку.

Заметим, что формула (10), где  $f(x)$  на каждом участке приближенно заменяется константой и формула (12), где для замены используется линейная функция, дают одинаковый порядок точности (и даже (10) несколько точнее!). Эта неожиданно высокая точность формулы (10)



объясняется тем, что точки  $\xi_k$  расположены точно в серединах интервалов  $(x_k, x_{k+1})$ . Снова симметрия повышает точность!

Заметим, далее, что формулу трапеций тоже можно рассматривать как некоторую интегральную сумму. Для этого достаточно разбить отрезок  $(a, b)$  на  $(n+1)$  частей: два участка длины  $\frac{h}{2}$  по краям, остальные длины  $h$ . После этого:  $\xi_0 = a$ ;  $\xi_{n+1} = b$ , все остальные  $\xi_k = a + kh$  расположатся в серединах своих отрезков.

Однако, полезнее другая точка зрения на формулу трапеций, легко допускающая обобщения. В формуле трапеций мы на каждом участке  $x_k \leq x \leq x_{k+1}$  функцию  $f(x)$  заменили линейной функцией, а  $\int_{x_k}^{x_{k+1}} f(x) dx$  на интеграл от этой линейной функции.

Если вместо линейной функции, т. е. многочлена первой степени, рассмотреть многочлены более высокой степени, получатся более точные квадратурные формулы. Реализуем эту идею.

#### 4. Формула Симпсона

Разобьем участок интегрирования  $(a, b)$  на  $n$  равных частей. Для каждого из полученных отрезков подберем квадратный многочлен, так чтобы он совпал с функцией в концах отрезка и в его середине. В качестве приближенного значения интеграла на этом отрезке примем значение интеграла от квадратного многочлена. Таким образом, нам понадобятся значения  $f(x_k)$  в точках  $x_k$  с интервалом  $h = \frac{b-a}{2n}$ .

Вычисления мы проведем для одного интервала, расположив его симметрично относительно 0. (Ясно, что от положения интервала на оси  $x$  результат не зависит). Итак, должно быть:  $f(x) = \alpha x^2 + \beta x + \gamma$  при  $x = -h, 0, h$ . Отсюда

$$\gamma = f(0), \quad \beta = \frac{1}{2h}(f(h) - f(-h)), \quad \alpha = \frac{1}{2h^2}(f(h) - 2f(0) + f(-h))$$

$$\int_{-h}^h (\alpha x^2 + \beta x + \gamma) dx = \frac{1}{3} h [f(-h) + 4f(0) + f(h)].$$

Написав такие формулы для каждого из  $n$  отрезков длины  $2h$  и просуммировав, получим

$$\begin{aligned} \int_a^b f(x) dx &\approx \frac{h}{3} [(f_0 + 4f_1 + f_2) + (f_2 + 4f_3 + f_4) + \dots + (f_{2n-2} + 4f_{2n-1} + f_{2n})] \\ \int_a^b f(x) dx &\approx \frac{h}{3} [f_0 + 4f_1 + 2f_2 + 4f_3 + \dots + 2f_{2n-2} + 4f_{2n-1} + f_{2n}]. \end{aligned} \quad (13)$$

Это и есть квадратурная формула Симпсона; она является одной из самых употребительных.

## 5. Свойства формулы Симпсона

Если  $f(x)$  — многочлен второй степени, то формула Симпсона (при любом  $h$ ) дает точный ответ. Это, конечно, не удивительно, так как при выводе этой формулы мы  $f(x)$  заменяли многочленом второй степени\*.

Замечательно, что и для многочлена 3 степени эта формула является тоже точной! Это легко может быть проверено прямым вычислением, однако поучительно получить результат без вычислений. Если участок интегрирования расположен симметрично относительно 0, то для  $f(x) = x^3$  точное значение интеграла

$\int_{-c}^c x^3 dx$  и значение по формуле Симпсона равны 0 (оба в силу

симметрии). Значит, на таком участке формула Симпсона точна для любого многочлена 3 степени. Осталось заметить, что любой

отрезок  $[a, b]$  при замене  $x \rightarrow x - r$ ,  $r = \frac{a+b}{2}$  переходит в симмет-

ричный отрезок, а многочлен 3 степени снова переходит в многочлен 3 степени. Теперь ясно, что формула Симпсона имеет 4 порядок точности: на каждом участке длины  $2h$  функцию можно заменить некоторым многочленом 3 степени\*\* с ошибкой  $\sim h^4$ , а для полученной кусочномногочленной функции формула, очевидно, является точной.

Ясно, далее, что ошибка зависит от величины четвертой производной функции. Однако, заранее совсем не очевидно, что в выражение для ошибки будет входить малый числовой множитель. Именно, справедлива следующая оценка

$$\left| \int_a^b f(x) dx - J_h[f] \right| \leq \frac{M_4}{180} (b-a) h^4; \quad M_4 = \max |f^{(4)}(x)|. \quad (14)$$

Здесь  $J_h[f]$  — приближенное значение интеграла по формуле Симпсона. Вы видите, что формула Симпсона оказывается гораздо точнее, чем можно было ожидать.

## \* 6. Старшие производные

Чтобы получить формулы для приближенного нахождения старших производных (пусть даже не очень точные), будем рас-

\* При фиксированном  $h$  формула Симпсона дает точный ответ и для кусочномногочленной функции, совпадающей с разными многочленами 2 степени на каждом из участков  $[x_{2k}, x_{2k+2}]$  (длины  $2h$ ).

\*\* Например, отрезком формулы Тэйлора.

суждать следующим образом.  $m$ -ая производная от  $f(x)$  есть результат применения  $m$  раз операции дифференцирования  $\frac{d}{dx} = D$ :

$$f'(x) = D[f], \quad f''(x) = D[D[f]] = D^2[f], \dots, \quad f^{(m)}(x) = D^m[f].$$

Простейшим дискретным аналогом дифференцирования является операция  $D_h$ :

$$D_h[f] = \frac{1}{h} (f(x+h) - f(x)) = \frac{1}{h} \Delta_h[f].$$

Мы обозначили здесь через  $\Delta_h$  «первую разность» функции:

$$\Delta_h[f] = f(x+h) - f(x).$$

Естественно предположить, что простейшим дискретным аналогом  $m$ -й производной является результат  $m$ -кратного применения  $D_h$ . По определению производной

$$\lim_{h \rightarrow 0} D_h[f] = D[f].$$

Можно доказать, что при любом  $m$

$$\lim_{h \rightarrow 0} D_h^m[f] = D^m[f].$$

При конечном  $h$  получаем приближенную формулу, имеющую первый порядок точности

$$f^{(m)}(x) \approx D_h^m[f] = \frac{1}{h^m} \Delta_h^m[f].$$

Выражение  $\Delta_h^m[f]$  — результат  $m$ -кратного применения  $\Delta_h$  — называется  $m$ -ой (правой) разностью функции  $f$ .  $m$ -ая разность  $f$  есть линейная комбинация значений:  $f(x)$ ,  $f(x+h)$ , ...,  $f(x+mh)$ . При  $m = 1, 2, 3$ , получается:

$$\Delta_h f(x) = f(x+h) - f(x)$$

$$\Delta_h^2 f(x) = f(x+2h) - 2f(x+h) + f(x)$$

$$\Delta_h^3 f(x) = f(x+3h) - 3f(x+2h) + 3f(x+h) - f(x)$$

и т. д. (аналогия с биномиальной формулой очевидна).

Используя симметрично расположенные значения аргумента можно (как и в случае второй производной) получить более точные формулы. Например, для четвертой производной можно написать следующие выражения:

$$f^{IV}(x) \approx \frac{1}{h^4} (f_4 - 4f_3 + 6f_2 - 4f_1 + f_0)$$

$$f^{IV}(x) \approx \frac{1}{h^4} (f_3 - 4f_2 + 6f_1 - 4f_0 + f_{-1}); \quad f_k = f(x + kh)$$

$$f^{IV}(x) \approx \frac{1}{h^4} (f_2 - 4f_1 + 6f_0 - 4f_{-1} + f_{-2}).$$

Последняя формула имеет второй порядок точности.

## 7. Примеры дискретных задач, отвечающих задачам анализа

Умея заменять дифференцирование и интегрирование их дискретными аналогами, мы можем сводить к дискретным задачам различные задачи анализа. При этом линейным дифференциальным и интегральным уравнениям ставятся в соответствие системы (конечного числа) линейных алгебраических уравнений.

**Пример 1.** Разностная система, отвечающая дифференциальному уравнению 2 порядка.

Пусть требуется найти функцию  $y(x)$  ( $a \leq x \leq b$ ), удовлетворяющую дифференциальному уравнению

$$y''(x) + q(x)y(x) = f(x)$$

и двум граничным условиям:

$$y(a) = \alpha, \quad y(b) = \beta.$$

Для приближенного решения этой задачи разделим отрезок  $[a, b]$  на  $n$  равных частей. Точки деления обозначим  $x_k = a + kh$ ,  $h = \frac{1}{n}(b - a)$ . Будем искать значения функции  $y(x)$  только при  $x = x_k$ . Записав  $y''(x_k)$  при  $k = 1, 2, 3, \dots, n-1$  по формуле (7), мы получим вместо дифференциального уравнения линейные алгебраические уравнения

$$\frac{1}{h^2} (y_{k+1} - 2y_k + y_{k-1}) + q_k y_k = f_k.$$

Здесь  $q_k = q(x_k)$ ,  $f_k = f(x_k)$  ( $k = 1, 2, 3, \dots, n-1$ ) — известные величины,  $y_k$  — неизвестные. Добавив уравнения  $y_0 = \alpha$  и  $y_n = \beta$ , мы получим систему  $n+1$  уравнений с  $n+1$  неизвестными. Матрица коэффициентов этой системы имеет следующий «тредиагональный» вид:



$$\begin{pmatrix} 1 & 0 & 0 & 0 & \dots & \dots & 0 \\ A & B_1 & A & 0 & \dots & \dots & 0 \\ 0 & A & B_2 & A & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & A & B_{n-1} & A \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix} \quad A = \frac{1}{h^2} \quad B_k = -\frac{2}{h^2} + q_k.$$

Такая система легко решается методом исключения. Решение особенно просто для случая, когда пригоден *простейший* метод исключения\*. При этом в любом случае количество необходимых арифметических действий пропорционально  $n$ .

**Пример 2.** Сведение линейного интегрального уравнения к системе линейных алгебраических уравнений.

Пусть требуется найти функцию  $y(x)$ , удовлетворяющую на отрезке  $a \leq x \leq b$  линейному интегральному уравнению

$$y(x) + \int_a^b K(x, s) y(s) ds = f(x), \quad (15)$$

где  $K(x, s)$  (ядро интегрального уравнения) и  $f(x)$  — заданные функции\*\*. При каждом фиксированном  $x$  в уравнение входит определенный интеграл

$$\int_a^b Q(s) ds, \quad \text{где } Q(s) = Q_x(s) = K(x, s) y(s).$$

Заменим этот интеграл какой-нибудь квадратурной формулой

$$\int_a^b Q(s) ds \approx \sum_{j=1}^n \mu_j Q(\xi_j) \quad (\text{например, формулой Симпсона}).$$

Тогда уравнение (15) приближенно заменяется уравнением

$$y(x) + \sum_j \mu_j K(x, \xi_j) y(\xi_j) = f(x). \quad (16)$$

Если бы нам были известны все значения  $y(\xi_j)$ , то по формуле (16) можно было бы найти  $y(x)$  для любого значения  $x$ .

---

\* Формульная реализация простейшего метода исключения для такой системы («прогонка») приведена в лекции 2.

\*\* Решение более простого на вид уравнения

$$\int_a^b K(x, s) y(s) ds = f(x)$$

(уравнение первого рода) есть гораздо более трудная задача.

Чтобы найти  $y(\xi_j)$ , будем рассматривать уравнение (16) только для значений  $x = \xi_1, \xi_2, \dots, \xi_n$ . Мы получим тогда систему  $n$  линейных алгебраических уравнений с  $n$  неизвестными. В отличие от системы, соответствующей дифференциальному уравнению в первом примере, здесь получается система «общего вида»: все коэффициенты ее, вообще говоря, отличны от нуля.

Как решить такую систему? Если  $n$  невелико (скажем,  $n \sim 15-20$ ), то можно воспользоваться каким-нибудь стандартным методом решения линейных систем, например, методом исключения. Если же требуется высокая точность (и потому  $n$  велико), то для разумного решения этой системы обязательно нужно использовать ее «непрерывное» происхождение (т. е. то, что  $f_j$  и  $y_j$  есть значения непрерывных функций в близких друг к другу точках). Как это делается, мы подробно рассмотрим позже.

## 8. Линейные дифференциальные и линейные разностные уравнения

Теперь мы оставим в стороне собственно вычислительные проблемы и проследим аналогию между линейными дифференциальными и линейными разностными уравнениями.

Для простоты рассмотрим сначала линейное уравнение второго порядка:

$$a_2(x)u'' + a_1(x)u' + a_0(x)u = f(x). \quad (17)$$

Заменим  $u''$  и  $u'$  их дискретными аналогами по формулам (2) и (6) и будем рассматривать только значения  $x = kh$ . Мы получим тогда *линейное разностное уравнение*:

$$b_2(k)u_{k+2} + b_1(k)u_{k+1} + b_0(k)u_k = f_k. \quad (18)$$

Нетрудно показать, что, подбирая  $a_s(x)$ , можно при фиксированном  $h$  получить *любые*  $b_s(k)$ . Поэтому *любую* систему уравнений вида (18), независимо от ее происхождения, мы будем называть *линейным разностным уравнением 2 порядка*. При этом предполагается, что  $b_2$  и  $b_0$  не обращаются в 0, так что при каждом  $k$  уравнение захватывает 3 точки ( $b_1(k)$  может обращаться в 0).

В том случае, когда  $b_s$  не зависят от  $k$ , мы будем говорить о разностном уравнении с *постоянными коэффициентами*.

Свойства разностных уравнений во многом аналогичны свойствам дифференциальных уравнений. Проследим несколько пунктов сходства.

$$a_2 u'' + a_1 u' + a_0 u = f \quad (19a) \quad b_2 u_{k+2} + b_1 u_{k+1} + b_0 u_k = f_k \quad (19b)$$

1) Задача Коши:  $u(x)_1$  при  $x \geq 0$  однозначно определяется заданием  $u(0)$  и  $u'(0)$ \*. Постановка такой задачи обосновывается теоремой существования и единственности решения.

2) Для однородного уравнения ( $f(x) \equiv 0$ ) общее решение есть линейная комбинация двух (не пропорциональных) частных решений:

$$u(x) = C_1 u^{(1)}(x) + C_2 u^{(2)}(x).$$

3) Для уравнения (19а) с постоянными коэффициентами существуют решения вида:

$$u(x) = e^{\lambda x} = \mu^x,$$

где  $\lambda$  — корень уравнения

$$a_2 \lambda^2 + a_1 \lambda + a_0 = 0. \quad (20a)$$

Если (20а) имеет некрлатные корни, то общее решение *одно-  
родного* уравнения (19а) имеет вид:

$$u(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x}.$$

Аналогию эту можно продолжить и углубить, рассмотрев вместо задачи Коши — краевые задачи, случай кратных корней характеристического уравнения и т. д.

1) Здесь естественно называть задачей Коши задание  $u_0$  и  $u_1$ . Существование и единственность решения при  $k \geq 0$  такой задачи очевидны из записи:

$$u_{k+2} = \frac{1}{b_2} (f_k - b_1 u_{k+1} - b_0 u_k).$$

2) Для разностного уравнения при  $f_k \equiv 0$  это тоже справедливо. Произвольное решение

$$u_k = C_1 u_k^{(1)} + C_2 u_k^{(2)}.$$

Здесь  $u_k^{(1)}$  и  $u_k^{(2)}$  — любые два решения (18). Например, можно определить  $u^{(1)}$  и  $u^{(2)}$  условиями Коши:

$$\begin{aligned} u_0^{(1)} &= 1, & u_0^{(2)} &= 0 \\ u_1^{(1)} &= 0, & u_1^{(2)} &= 1. \end{aligned}$$

Тогда

$$C_1 = u_0, \quad C_2 = u_1.$$

3) Для уравнения (19б) с постоянными коэффициентами существуют решения вида  $u_k = \mu^k$ , где  $\mu$  — корень уравнения

$$b_2 \mu^2 + b_1 \mu + b_0 = 0 \quad (20б)$$

Если уравнение (20б) имеет не кратные корни, то общее решение *однородного* уравнения (19б) имеет вид:

$$u_k = C_1 \mu_1^k + C_2 \mu_2^k.$$

\* Разумеется, вместо 0 можно взять любое значение  $x_0$ .

Пока мы этого делать не будем и лишь кратко обсудим общий случай — линейное разностное уравнение  $n$ -го порядка. Общий вид такого уравнения

$$\sum_{s=0}^n b_s(k) u_{k+s} = f_k \quad (b_0(k) \neq 0, b_n(k) \neq 0).$$

При каждом  $k$  в уравнении участвуют  $n+1$  значение  $u_k$  \*. При постоянных  $b_s(k) = b_s$  и  $f_k = 0$  опять существуют частные решения вида  $u_k = \mu^k$ , где  $\mu$  — корень характеристического уравнения:

$$\sum_{s=0}^n b_s \mu^s = 0.$$

«Задача Коши» для такого разностного уравнения ставится так. Заданы  $u_0, u_1, \dots, u_{n-1}$ . Ищутся  $u_k$  при всех  $k > n-1$ . Задача эта, очевидно, имеет решение и притом только одно.

## 9. Три примера

1) Рассмотрим последовательность целых чисел, определенную следующим правилом. Первые два равны 1; каждое следующее есть сумма двух предыдущих:

$$u_0 = u_1 = 1; \quad u_{k+2} = u_k + u_{k+1}.$$

Эта знаменитая последовательность обладает многими интересными свойствами и называется «числа Фибоначчи».\*\*

С нашей сегодняшней точки зрения соотношение  $u_{k+2} = u_k + u_{k+1}$  есть разностное уравнение 2 порядка с постоянными коэффициентами. Условие  $u_0 = u_1 = 1$  есть задача Коши для этого уравнения. Решив эту задачу, мы получим явные формулы для  $u_k$ .

Общее решение уравнения  $u_{k+2} = u_k + u_{k+1}$  имеет вид:  $u_k = C_1 \mu_1^k + C_2 \mu_2^k$ , где  $\mu_j$  — корни характеристического уравнения  $\mu^2 - \mu - 1 = 0$ :

$$\mu_1 = \frac{1 - \sqrt{5}}{2}, \quad \mu_2 = \frac{1 + \sqrt{5}}{2}.$$

\* Впрочем, внутренние коэффициенты —  $b_1, b_2, \dots, b_{n-1}$  вполне могут обращаться в 0.

\*\* Последовательность эта впервые была рассмотрена Леонардо Фибоначчи в 13 веке, в связи с (сильно идеализированной) задачей о размножении кроликов. Это есть, таким образом, первое применение математики к биологии!

Из начальных данных находим

$$C_1 = -\frac{1}{\sqrt{5}} \cdot \frac{1 - \sqrt{5}}{2}; \quad C_2 = \frac{1}{\sqrt{5}} \cdot \frac{1 + \sqrt{5}}{2}.$$

Таким образом,  $k$ -ое число Фибоначчи задается следующей формулой:

$$u_k = \frac{1}{\sqrt{5}} \left[ \left( \frac{1 + \sqrt{5}}{2} \right)^{k+1} - \left( \frac{1 - \sqrt{5}}{2} \right)^{k+1} \right].$$

Замечательно, что для задания последовательности целых чисел нужно использовать иррациональные числа!

2) В качестве второго примера рассмотрим разностное уравнение, которое произошло из дифференциального. Пусть

$$u'(x) = au(x), \quad u(0) = 1. \quad (21)$$

Решение этой задачи Коши есть  $u(x) = e^{ax}$ . Перейдем к разностной задаче Коши:

$$\frac{1}{h}(u_{k+1} - u_k) = au_k; \quad u_0 = 1. \quad (22)$$

Из уравнения (22)  $u_{k+1} = (1 + ah)u_k$ , т. е.

$$u_k = (1 + ah)^k = \mu^k.$$

Попробуем установить соответствие между решениями дифференциального и разностного уравнения. Для этого запись (22) неудобна — в ней не указано значение  $x$ .

Определим функцию  $u(x) = u_h(x)$  для дискретного набора значений  $x = x_k = kh$ :  $u(x_k) = u_k$ . Уравнение (22) переписывается в виде:

$$\frac{1}{h} [\tilde{u}(x+h) - \tilde{u}(x)] = a\tilde{u}(x), \quad (22)$$

а решение  $u_k$  в виде  $\tilde{u}(x) = (1 + ah)^{x/h}$ .

Теперь зафиксируем  $x$  и устремим  $h$  к 0:  $\tilde{u}_h(x) = (1 + ah)^{x/h} \rightarrow e^{ax} = u(x)^{**}$ . Мы видим, что в этом примере все обстоит «бла-

\* Эта формула была впервые написана через несколько столетий после Фибоначчи!

\*\*  $\frac{x}{h}$  не обязано быть целым. Поэтому точнее писать так:

$\max_x |\tilde{u}_h(x) - u(x)| \rightarrow 0$  при  $h \rightarrow 0$ .

Здесь максимум берется по всем  $x$ , в которых определена функция  $\tilde{u}(x)$ .

гополучно»: решение разностного уравнения  $\tilde{u}_h(x)$  при уменьшении  $h$  стремится к решению дифференциального.

Последнее замечание. Формула (22) есть запись простейшего метода интегрирования дифференциальных уравнений — метода Эйлера — для уравнения (21). Мы убедились, что здесь при  $h \rightarrow 0$  приближенное решение, полученное по методу Эйлера, сходится к точному. Разность  $u - u_h$ , как легко убедиться, имеет порядок  $h$ .

3) Пусть снова  $u'(x) = au(x)$ .

Для построения разностного уравнения воспользуемся более точной формулой для производной:

$$u'(x) \approx \frac{1}{2h}(u(x+h) - u(x-h)).$$

Получим

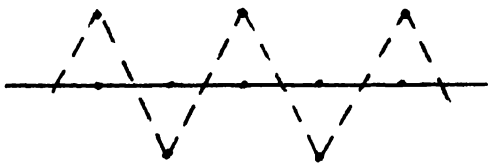
$$\frac{1}{2h}(u_{k+1} - u_{k-1}) = au_k, \quad u_0 = 1 \quad (23)$$

или

$$\frac{1}{2h}(\tilde{u}(x+h) - \tilde{u}(x-h)) = a\tilde{u}(x), \quad \tilde{u}(0) = 1. \quad (23)$$

Обратите внимание на радикальное отличие формулы (23) от (22). Уравнение (23) имеет второй порядок, стало быть — два линейно независимых решения. Возьмем в качестве таких решений  $u_k^{(1)} = \mu_1^k$  и  $u_k^{(2)} = \mu_2^k$ , где  $\mu_1$  и  $\mu_2$  — корни характеристического уравнения  $\mu^2 - 2ah\mu - 1 = 0$ . При малых  $h$   $\mu_1 \approx 1$ ,  $\mu_2 \approx -1$ .

Рассмотрим сначала частное решение  $u_k^{(2)} = \mu_2^k$  ( $\mu_2 < 0$ ) или  $\tilde{u}^{(2)}(x) = \mu_2^{x/h}$ . При малом  $h$   $\tilde{u}^{(2)}(x)$  сильно осциллирует и при  $h \rightarrow 0$  ни к чему не стремится! В то же время легко проверяется, что  $u_1(x) = \mu_1^{x/h}$  при  $h \rightarrow 0$  сходится к  $u(x) = e^{ax}$ .



Мы видим, что из-за повышения порядка разностное уравнение богаче дифференциального. У него есть решение, переходящее в пределе в гладкую функцию, удовлетворяющую дифференциальному уравнению, но есть и «паразитическое» решение, которому ничего не отвечает в пределе (при  $h \rightarrow 0$ ).

Заметим в заключение, что, если мы захотим решить задачу Коши для (23), то нужно задать кроме  $u_0$  еще  $u_1$ . Если наугад задать  $u_1$ , то решение (23), конечно, не будет при  $h \rightarrow 0$  приближаться к  $e^{ax}$ . Чтобы такое приближение имело место, нужно задать  $u_1 = u_0$ , или при  $h \rightarrow 0$   $u_1$  приближать к  $u_0$ .

## ЛЕКЦИЯ 5

### УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ. ПРОБЛЕМА УСТОЙЧИВОСТИ РАЗНОСТНЫХ АППРОКСИМАЦИЙ

#### 1. Вводные замечания

Мы будем в этой лекции рассматривать два простейших уравнения с частными производными: волновое уравнение и уравнение теплопроводности. Мы ограничимся рассмотрением уравнений с двумя независимыми переменными  $x, t$ ; уже случай 3 переменных является для нас сейчас непосильной темой. Таким образом, все будет происходить на прямой  $-\infty < x < \infty$  (или на отрезке этой прямой) и в некотором интервале времени  $t$ .

Мы будем, как и в случае обыкновенных уравнений, рассматривать задачу Коши, или, в случае конечного отрезка, смешанную задачу: кроме начальных данных (при  $t=0$ ) задаются подходящие краевые условия на концах рассматриваемого отрезка. Однако, ситуация здесь много сложнее, чем для обыкновенных дифференциальных уравнений по двум причинам. Первая из них очевидна: задача гораздо более громоздка, нужно превратить в дискретные две переменные  $x$  и  $t$  и в каждый момент времени  $t_n$  находить много значений неизвестной функции  $u(x_k, t_n)$ .

Вторая (и более важная) причина состоит в следующем. Простейшие дискретные (разностные) аппроксимации часто оказываются непригодными из-за возникающей неустойчивости. Поэтому вначале нужно думать не о том, чтобы найти более точный алгоритм, а о том, чтобы найти хоть какой-нибудь пригодный способ приближенного решения. Наконец, не нужно забывать, что уравнения с частными производными гораздо более «индивидуальны», чем обыкновенные уравнения — сама постановка правильной задачи зависит от уравнения. Вы увидите, что эта индивидуальность проявляется и при приближенном решении этих уравнений.

## 2. Уравнения

Одномерное волновое уравнение выглядит так:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}. \quad (1)$$

Общее решение этого уравнения записывается в виде

$$u(x, t) = f(x + ct) + g(x - ct),$$

т. е. зависит от 2 произвольных функций  $f$  и  $g$  одной переменной. Слагаемое  $f(x + ct)$  соответствует волне, распространяющейся влево,  $g(x - ct)$  — вправо\*. Всякое решение (1) есть суперпозиция двух таких волн. Простейшее уравнение «необратимого типа» имеет вид:

$$\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2}, \quad D > 0. \quad (2)$$

Если  $u$  — температура, то это уравнение описывает теплопроводность, если  $u$  — концентрация, то диффузию и т. д.

Полное волновое уравнение (1) является для нас все еще слишком сложным и для начала вместо (1) мы рассмотрим более простое уравнение, описывающее волны, распространяющиеся в одном направлении:

$$\frac{\partial u}{\partial t} = c \frac{\partial u}{\partial x}. \quad (3)$$

Общее решение его есть

$$u(x, t) = f(x + ct). \quad (4)$$

Если при  $t=0$  задать (для всех  $x$ )  $u(x, 0) = f(x)$ , то в момент  $t$  получим  $u(x, t) = f(x + ct)$  — первоначальный профиль «волны» без искажения передвигается со скоростью  $|c|$  (при  $c > 0$  влево, а при  $c < 0$  — вправо).

Как правило, нас интересует  $u(x, t)$  не на всей прямой  $-\infty < x < \infty$ , а на отрезке  $a \leq x \leq b$ . Тогда, кроме начальных данных  $u(x, 0)$ , нужно (при  $c > 0$ ) задать еще  $u(b, t)$  — значения решения на правом конце (рис. 1).

На левом конце ничего задавать не нужно. Это легко понять, заметив, что решение  $u(x, t)$  постоянно на прямых  $x + ct = \text{const}$  — характеристиках уравнения (3). Значит, на левую границу  $x=a$  значения  $u$  «приносятся» характеристиками снизу

---

\* Мы считаем здесь  $c > 0$ .



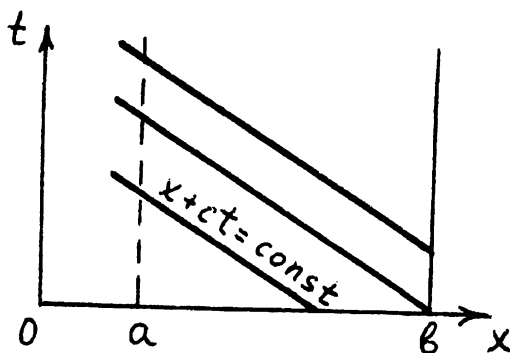


Рис. 1.

(с отрезка  $t=0; a \leq x \leq b$ ) или справа (с отрезка  $x=b; 0 \leq t \leq T$ ). Если в уравнении  $u_t = cu_x$   $c$  будет переменным ( $c=c(x, t)$ ), то характеристики уже не будут прямыми\*, но решение по-прежнему будет постоянно на характеристиках.

Наконец, для уравнения  $u_t = c(x, t)u_x + F(x, t)$  решение уже не будет постоянным на характеристиках, но и здесь (при  $c(x, t) > 0$ ) все влияния передаются справа налево. Более точно это означает вот что. Изменим  $u(x, t)$  в некоторый момент  $t_0$  на участке  $a \leq x \leq b$ . Тогда при  $t > t_0$  решение изменится лишь в полоске между характеристиками (см. рис. 2). Правее же  $x=b$  никаких изменений не произойдет.

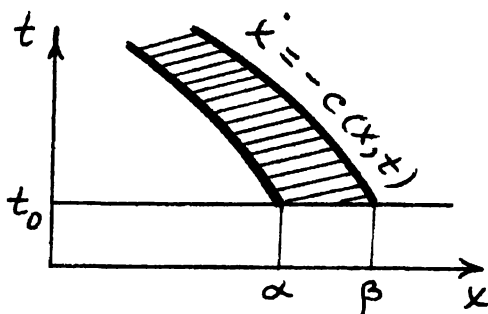


Рис. 2.

Для уравнения теплопроводности (2) нужно задавать граничные условия на обоих концах отрезка  $(a, b)$ . Если внести изменение в  $u(x, t)$  в некоторый момент времени, то в следующие

\* Характеристики будут здесь решения обыкновенного уравнения

$$\frac{dx}{dt} = -c(x, t).$$

моменты  $u(x, t)$  изменится на всем отрезке. Таким образом, влияния передаются здесь в обе стороны и притом формально с бесконечной скоростью.

### 3. Простейшие разностные аппроксимации

Пусть требуется найти решение уравнения  $u_t = cu_x$  (3) при  $-\infty < x < \infty$  по известным начальным данным:  $u(x, 0) = \varphi(x)$ . Выберем некоторые малые  $\tau$  и  $h$  и будем искать приближенное решение  $u(x, t)$  лишь для значений  $x_k = kh$ ,  $t_n = n\tau$ . Обозначим  $u(x_k, t_n) = u_k^n$  и заменим производные в уравнении (3) какими-нибудь приближенными дискретными выражениями. Это можно, конечно, сделать многими способами.

Например, так:

<p>I     <math>\frac{1}{\tau} (u_k^{n+1} - u_k^n) = \frac{c}{h} (u_{k+1}^n - u_k^n)</math></p> <p>II    <math>\frac{1}{\tau} (u_k^{n+1} - u_k^n) = \frac{c}{h} (u_k^n - u_{k-1}^n)</math></p> <p>III   <math>\frac{1}{\tau} (u_k^{n+1} - u_k^n) = \frac{c}{2h} (u_{k+1}^n - u_{k-1}^n)</math></p>	<p>Схемы:</p> <p style="text-align: right;">(5)</p>
---	---

Любое из уравнений (5) позволяет последовательно найти  $u_k^1, u_k^2, \dots, u_k^n, \dots$ , если положить  $u_k^0 = \varphi(x_k) = \varphi(kh)$ . (Если ищется  $u(x, t)$  при  $0 \leq t \leq T$ , то можно остановиться при  $n = N = T/\tau$ .)

Однако, не всеми формулами (5) можно пользоваться. При  $c > 0$  формула II — совершенно непригодна. Формула III — практически непригодна. Формулой I можно пользоваться, если

$$\frac{c\tau}{h} \leq 1.$$

Легче всего забраковать формулу II. Если начальные данные отличны от нуля только на отрезке  $[a, b]$ , то картинка, соответствующая вычислениям по схеме II, будет выглядеть так (рис. 3; крестики — точки сетки, где  $u_k^n$  могут не равняться нулю): «Приближенное решение»  $u(x, t)$  будет все время равно нулю левее  $x = a$ . Между тем, *точное* решение отлично от нуля в полосе между характеристиками, изображенными на рисунке. Вы видите, что  $u(x, t)$  даже качественно не похоже на точное решение  $u(x, t)$ . Причина ясна: в дифференциальном уравнении (3) все влияния передаются *налево*, а в разностной системе II — *направо*.

\* Иногда возникает опасность спутать верхний индекс  $n$  с показателем степени. В таких случаях мы будем писать  $u_k^{(n)}$ .

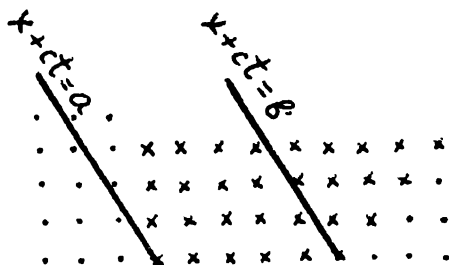


Рис. 3.

Аналогичные рассуждения показывают, что схема I может дать решение  $u(x, t)$ , близкое к  $u(x, t)$ , лишь при условии  $ct \leq h$ . На этот раз ненулевой участок решения будет со временем двигаться налево, как и должно быть. Посмотрим, однако, с какой скоростью движется граница, отделяющая невозмущенную область ( $u \equiv 0$ ) от возмущенной ( $u \neq 0$ ). За шаг  $\tau$  эта граница передвигается влево на  $h$ . Значит, скорость ее движения  $c_1 = h/\tau$ . Скорость движения возмущения в точном уравнении  $c$ . Если  $ct > h$ , то  $c > c_1$ . Это означает, что граница ненулевого участка движется медленней, чем нужно. Ясно, что это приводит к ошибочному результату и потому условие  $\frac{c\tau}{h} \leq 1$  необходимо для разумности применения схемы I. Мы покажем потом, что это же условие является и достаточным. Итак, в этих двух случаях мы можем из самых простых и наглядных соображений откинуть непригодные аппроксимации. Что касается схемы III, то ее непригодность не очевидна. Действительно, возмущения здесь распространяются в обе стороны и при  $\frac{c\tau}{h} \leq 1$  скорость их распространения достаточна. Тем не менее, вычислять по этой схеме нельзя\*: ничтожные изменения в начальных данных или в процессе счета приводят к большим изменениям в дальнейшем.

Забегая вперед, я подчеркну, что дело здесь не только в приближенности вычислений. Даже если мы могли бы производить арифметические действия абсолютно точно, мы не получили бы по схеме III (при произвольных начальных данных) результата, близкого к точному. Позже мы это аккуратно сформулируем и докажем, а пока займемся вторым уравнением.

Для уравнения теплопроводности  $u_t = Du_{xx}$  простейшая разностная аппроксимация имеет вид

\* Если только не пользоваться очень малым шагом по времени (см. лекцию 6).

$$\frac{1}{\tau}(u_k^{n+1} - u_k^n) = \frac{D}{h^2}(u_{k+1}^n - 2u_k^n + u_{k-1}^n). \quad (6)$$

Перепишем:

$$u_k^{n+1} = \gamma u_{k+1}^n + (1-2\gamma)u_k^n + \gamma u_{k-1}^n; \quad \gamma = \frac{D\tau}{h^2}. \quad (6a)$$

Теперь ясно, что свойства этой разностной аппроксимации (или, как говорят, разностной схемы) зависят от (безразмерной) величины  $\gamma$ , а не от  $\tau$  и  $h$  в отдельности. Оказывается, схема (6) пригодна при  $\gamma \leq \frac{1}{2}$  и непригодна при  $\gamma > \frac{1}{2}$  по той же причине — из-за возникающей неустойчивости\*.

Нам следовало бы теперь определить точно, что такое устойчивость и неустойчивость, и выяснить, почему разностные схемы в одном случае устойчивы, в другом нет. Однако, раньше чем это делать, мы «докажем» одну неверную теорему. Я думаю, это поможет нам оценить деликатность ситуации и понять существо дела. Итак, следите за доказательством.

#### 4. Теорема

Пусть для приближенного решения задачи Коши:  $u_t = u_x$ ,  $u(x, 0) = \varphi(x)$  используется аппроксимация

$$\frac{1}{\tau}(u_k^{n+1} - u_k^n) = \frac{1}{h}(u_k^n - u_{k-1}^n); \quad u_k^0 = \varphi(x_k). \quad (7)$$

Предположим, что  $\tau$  и  $h$  стремятся к 0 так, что  $\frac{\tau}{h} = \gamma$  остается постоянным. Тогда  $\tilde{u}(x, t) \rightarrow u(x, t)$ , т. е. решения разностных уравнений стремятся к решению дифференциального.

Раньше, чем доказывать, уточним формулировку. Функции  $\tilde{u}(x, t)$  заданы не при всех  $x$  и  $t$ , а на дискретном множестве точек (которое к тому же меняется при изменении  $\tau$  и  $h$ ). Как понимать сходимость  $\tilde{u}(x, t)$  к  $u(x, t)$ ? Проще всего так:

$$\max |\tilde{u}(x, t) - u(x, t)| \rightarrow 0.$$

Здесь  $\max$  берется по всем  $x = x_k$  и для всех  $t = t_n$  из интервала  $0 \leq t_n \leq T$ .

Перейдем теперь к доказательству.

\* О том, что  $\gamma = \frac{1}{2}$  является «критическим» значением, можно догадаться:

при  $\gamma = \frac{1}{2}$  один из коэффициентов в (6a) меняет знак.

1 этап. Рассмотрим начальные данные специального вида:

$$u(x, 0) = e^{isx}.$$

Точное решение при таких начальных данных есть

$$u(x, t) = e^{is(x+t)}.$$

Найдем приближенное решение по формуле (7):

$$u_k^{n+1} = u_k^n + \frac{\tau}{h} (u_k^n - u_{k-1}^n) = (1 + \gamma) u_k^n - \gamma u_{k-1}^n; \quad \gamma = \frac{\tau}{h}.$$

Найдем  $u_k^n$  на первом слое (при  $n=1$ ):

$$\begin{aligned} u_k^1 &= (1 + \gamma) u_k^0 - \gamma u_{k-1}^0 = (1 + \gamma) e^{isx_k} - \gamma e^{is(x_k - h)} = \\ &= e^{isx_k} [1 + \gamma - \gamma e^{-ish}]. \end{aligned}$$

Таким образом, переход от нулевого к первому слою сводится к умножению на  $1 + \gamma - \gamma e^{-ish}$ . Соответственно, на  $n$ -ом слое получим:  $u_k^n = e^{isx_k} (1 + \gamma - \gamma e^{-ish})^n$ ; или, так как  $n = t_n / \tau$

$$\tilde{u}(x, t) = e^{isx} [1 + \gamma(1 - e^{-ish})]^{t/\tau}. \quad (8)$$

Теперь

$$\begin{aligned} \lim_{h \rightarrow 0, \tau \rightarrow 0} [1 + \gamma(1 - e^{-ish})]^{t/\tau} &= \lim_{h \rightarrow 0} [1 + \gamma(1 - e^{-ish})]^{t/\gamma h} = \\ &= \lim_{h \rightarrow 0} (1 + is\gamma h + o(h))^{t/\gamma h} = e^{ist} \quad (\gamma = \text{const!}). \end{aligned}$$

Итак:  $\tilde{u}(x, t) \rightarrow e^{isx} \cdot e^{ist} = u(x, t)$ , т. е. для начальных данных специального вида сходимость доказана.

2 этап. Пусть теперь  $u(x, 0) = \varphi(x)$  — произвольная функция, достаточно гладкая и равная 0 вне конечного интервала. Разложим ее в интеграл Фурье:

$$u(x, 0) = \varphi(x) = \int_{-\infty}^{\infty} g(s) e^{isx} ds.$$

Точное решение будет иметь вид

$$u(x, t) = \varphi(x+t) = \int_{-\infty}^{\infty} g(s) e^{isx} e^{ist} ds.$$

---

\* Вас не должно смущать то, что мы рассматриваем здесь комплексные решения. Как и всегда при изучении *линейных* уравнений, можно все вычисления производить в комплексной форме, а в окончательном ответе, если это нужно, перейти к вещественным величинам.

Для приближенного решения в силу линейности разностных уравнений тоже получим интеграл:

$$\tilde{u}(x, t) = \int_{-\infty}^{\infty} g(s) e^{isx} [1 + \gamma(1 - e^{-ish})]^{t/vh} ds. \quad (9)$$

Теперь устремим  $h$  к 0, переходя в (9) к пределу под знаком интеграла; получим:  $\lim_{h \rightarrow 0} \tilde{u}(x, t) = u(x, t)$  — результат явно аб-

сурдный. Ведь (7) — это та самая схема (5, II), которая здесь (при  $c=1$ ) совершенно не годится.

В чем дело? Вы, вероятно, уже догадались. На первом этапе все правильно\*. А вот предельного перехода под знаком интеграла здесь совершать нельзя! Мы ошибаемся как раз в том месте, в котором нематематик склонен не задумываться (предоставляя излишнее, как он полагает, «наведение строгости», математикам).

Итак, если взять начальные данные в виде отдельной гармоник ( $u(x, 0) = \cos sx$  или  $\sin sx$ ) или суммы конечного числа гармоник, и если вычисления производить абсолютно точно, то эта глупая и бессмысленная схема дает в пределе правильный ответ. Однако, чем выше будет частота (чем больше  $s$ ), тем меньше  $h$  понадобятся, тем «позже» будет хорошее приближение. Можно сказать, что трудности вызываются высокими частотами. Это есть полезный урок из этой неверной теоремы. Второй полезный урок — прием использования специального вида начальных данных. Этот прием нам очень пригодится.

## 5. Определение устойчивости

Поясним сначала суть дела. Пусть нам дано дифференциальное уравнение

$$D_{x,t}[u] = 0 \quad (10)$$

( $D_{x,t}$  — дифференциальный оператор, включающий в себе дифференцирование по  $x$  и по  $t$ ). Требуется решить задачу Коши при заданных начальных условиях  $u(x, 0) = \varphi(x)$ .

Рассмотрим разностное уравнение  $\bar{D}_{x,t}[u] = 0$ , соответствующее дифференциальному уравнению (10),  $\bar{D}_{x,t}$  — разностный оператор зависящий от выбора  $\tau$  и  $h$ ; пусть  $u(x, 0) = \varphi(x)$ . Нас интересует  $u(x, t)$  при  $-\infty < x < \infty$ ;  $0 \leq t \leq T$ . Поэтому нужно найти  $u(x, t)$  для тех же пределов изменения  $x$  и  $t$ . Мы хотим, чтобы эта задача была «хорошей», т. е., чтобы небольшим изменениям в начальных данных отвечали бы небольшие изменения решения в целом.

\* Хотя наши рассуждения и не были безупречно строгими.

Сосредоточимся теперь на линейных задачах. Для линейных уравнений есть смысл выделить в записи однородную и неоднородную часть:

$$L[u] = f \quad \text{и} \quad \tilde{L}[\tilde{u}] = \tilde{f}.$$

Если уравнение линейное, то изменение начальных данных равносильно добавлению к решению некоторого решения однородного уравнения. Действительно, если  $L[u_1] = f$  и  $L[u_2] = \tilde{f}$ , то разность  $u_1 - u_2$  удовлетворяет уравнению  $L[u_1 - u_2] = 0$ . Итак, однородное линейное уравнение  $L[u] = 0$  должно иметь «малое» решение при «малых» начальных данных. Однако для однородной линейной задачи понятие «малости» решения теряет смысл (можно умножить начальные данные на миллион и решение тоже умножится на миллион). Разумно требовать, чтобы решение не очень превосходило начальные данные. Вот к чему сводится для этого случая требование устойчивости решения задачи Коши.

### Определение

Задача Коши для линейного разностного уравнения

$$L_{\tau,h}[\tilde{u}] = \tilde{f}; \quad \tilde{u}(x, 0) = \varphi(x); \quad 0 \leq t \leq T$$

называется устойчивой, если для решения соответствующего однородного уравнения  $L_{\tau,h}[\tilde{w}] = 0$  справедливо неравенство:

$$|\tilde{w}(kh, n\tau)| \leq C \max_k |\tilde{w}(kh, 0)| \quad (11)$$

при всех допустимых  $\tau, h$ ;  $-\infty < k < \infty$ ;  $0 \leq n\tau \leq T$ . Постоянная  $C$  не должна зависеть от  $\tau$  и  $h$ .

## 6. Пояснения и замечания к определению устойчивости

1) Определение, строго говоря, относится не к одной разностной задаче Коши, а к совокупности таких задач (при разных  $\tau, h$ ). Предполагается, что все разностные уравнения построены по одной схеме. Поэтому чаще говорят об устойчивости разностной схемы.

2) Неравенство (11) является ограничением лишь тогда, когда  $\tau$  и  $h$  могут быть сколь угодно малы: для фиксированных  $\tau$  и  $h$  константу  $C$  всегда можно подобрать (так как число слоев по времени тоже будет фиксировано). Таким образом, понятие устойчивости разностной схемы есть предельное, асимптотическое понятие. Однако, в определении не требуется, чтобы  $\tau$  и  $h$  могли стремиться к 0 произвольно, независимо друг от друга. Такое требование (абсолютная устойчивость) обычно бывает слишком сильным. С типичными ограничениями мы уже встречались:

$$\frac{c\tau}{h} \leq 1 \quad \text{или} \quad \frac{D\tau}{h^2} \leq \frac{1}{2}.$$

3) По смыслу определения постоянная  $C$  не должна возрастать с уменьшением  $\tau$  (и соответствующим увеличением  $N = T/\tau$ ). Однако, с ростом  $T$   $C$  вполне может расти, так что, вообще говоря,  $C = C(T)$ .

4) Задача Коши на бесконечной прямой встречается сравнительно редко. Чаще мы имеем дело с конечным отрезком. В этом случае нужно определять устойчивость смешанной задачи.

Обозначим через  $\Gamma$  ту часть границы, где задаются краевые условия (включая сюда и начальные условия). Например, для уравнения  $\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2}$  это 3 отрезка. Мы будем требовать, чтобы (рис. 4)

$$|\bar{u}(x, t)| \leq C \max_{\Gamma} |\bar{u}(x, t)|.$$

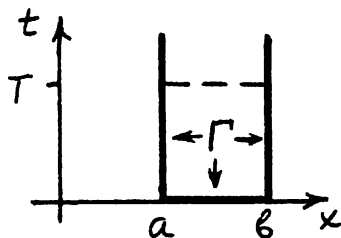


Рис. 4

Максимум берется теперь по начальным и граничным условиям.

Последнее, методологическое замечание. Определение устойчивости почти дословно повторяется для обыкновенных (одномерных) разностных уравнений. Правда, проблема устойчивости при решении обыкновенных дифференциальных уравнений не стоит столь остро — можно использовать готовые, проверенные формулы — но почти все идеи можно понять на этом случае. В то же время запись сильно упрощается. Поэтому, когда мы будем более глубоко рассматривать эту тему, то начнем с обыкновенных дифференциальных и, соответственно, одномерных разностных уравнений.

## 7. Обнаружение неустойчивости разностных аппроксимаций по методу Фурье

Рассмотрим снова задачу Коши для уравнения теплопроводности

$$\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2}; \quad -\infty < x < \infty, \quad 0 \leq t \leq T; \quad u(x, 0) = \varphi(x).$$

Покажем, что простейшая разностная схема (см. 6а)

$$u_k^{(n+1)} = \gamma u_{k-1}^{(n)} + (1-2\gamma) u_k^{(n)} + \gamma u_{k+1}^{(n)}, \quad (12)$$

неустойчива при  $\gamma = \frac{D\tau}{h^2} > \frac{1}{2}$ . Для этого выберем такие на-



чальные данные, при которых решение легко находится в явном виде. Вычисления удобно провести в комплексной форме. Задав начальные данные в виде  $u(x, 0) = e^{isx}$ , ищем решение уравнений (12) по методу Фурье в виде

$$\tilde{u}(x, t) = e^{isx + \lambda t} \quad \text{или} \quad u_k^{(n)} = \tilde{u}(kh, n\tau) = e^{iskh + \lambda n\tau}. \quad (13)$$

Обозначим  $e^{\lambda\tau} = \mu$ , подставим в (12) и сократим на  $\mu^n e^{iskh}$ , получим

$$\mu = \gamma e^{-ish} + 1 - 2\gamma + \gamma e^{ish} = 1 - 2\gamma + 2\gamma \cos sh.$$

Поскольку  $|u_k^{(n)}| = \mu^n$ , оценим  $|\mu|$ . Очевидно,  $1 - 4\gamma \leq \mu \leq 1$ . Значит, при  $\gamma \leq \frac{1}{2}$   $|\mu| \leq 1$  и  $|\tilde{u}(x, t)| \leq 1$ .

При  $\gamma > \frac{1}{2}$  выберем  $s = \pi/h$ , тогда  $|1 - 4\gamma| > 1$ .

При максимальном значении  $t = T = n\tau$ , получим  $|\tilde{u}(x, t)| = \mu^n$ . Эта величина при  $\tau \rightarrow 0$  (т.е.  $N \rightarrow \infty$ ) неограниченно растет. Вы видите, что требование  $|\tilde{u}(x, t)| \leq C$  не выполнено — неустойчивость схемы (12) при  $\gamma > \frac{1}{2}$  доказана. Вполне аналогично исследуется методом Фурье устойчивость задачи Коши для других разностных уравнений с постоянными коэффициентами. Особенно замечательно, что исследование устойчивости методом Фурье для уравнения  $u_t = cu_x$  автоматически бракует те разностные схемы, (5, I) и (5, II), непригодность которых мы выясняли, рассматривая распространение возмущений. Мы могли вовсе ничего не знать о том, как «передаются влияния» этими схемами и отбросить их в силу неустойчивости. Смотрите, как красиво это получается. Пусть для уравнения  $\frac{\partial u}{\partial t} = c \frac{\partial u}{\partial x}$  использована схема (5, II):

$$u_k^{(n+1)} = (1 + \gamma) u_k^{(n)} - \gamma u_{k-1}^{(n)}, \quad \gamma = \frac{c\tau}{h}.$$

Положим  $u_k^{(n)} = e^{isk\lambda^n}$ , тогда  $\lambda(s) = 1 + \gamma - \gamma e^{-is}$ . Когда  $s$  меняется от 0 до  $2\pi$ ,  $\lambda(s)$  описывает окружности, изображенные на рис. 5. При  $\gamma > 0$   $|\lambda(s)| \geq 1$  ( $|\lambda|_{\max} = 1 + 2\gamma$ ) — схема неустойчива. При  $c < 0$  и  $|\gamma| < 1$  окружность  $\Omega$  лежит внутри единичного круга — схема устойчива (точнее, не обнаруживает неустойчивости по методу Фурье). Наконец, при  $c < 0$  и  $|\gamma| > 1$  (рис. 5в) схема неустойчива —  $\max |\lambda(s)| = 2|\gamma| - 1 > 1$ . Последнее замечание. Чтобы обнаружить неустойчивость, достаточно ее обнаружить хоть на чем-нибудь. Чтобы доказать устойчивость, нужно провести доказательство для всех начальных данных. Это, конечно, труднее, но для тех простых примеров, которые рассматриваются в этой лекции, мы в состоянии это сделать.

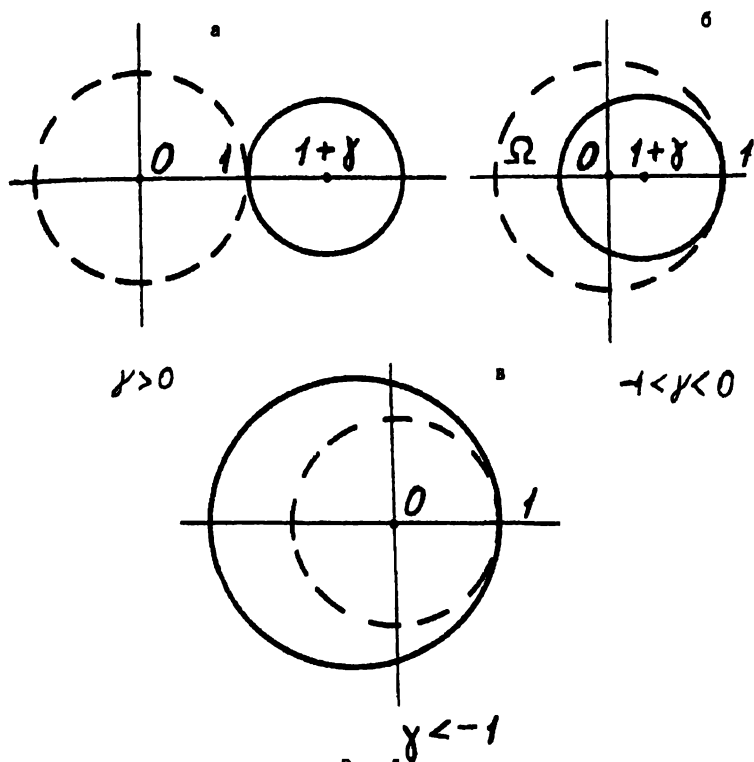


Рис. 5

## 8. Пример доказательства устойчивости разностной схемы

Покажем, что для уравнения  $u_t = Du_{xx}$  схема (12) является устойчивой при  $\gamma = \frac{D\tau}{h^2} \leq \frac{1}{2}$ . Если  $\gamma \leq \frac{1}{2}$ , то  $1 - 2\gamma \geq 0$ . Значит

$$\begin{aligned}
 |u_k^{(n+1)}| &\leq \gamma |u_{k-1}^{(n)}| + (1 - 2\gamma) |u_k^{(n)}| + \gamma |u_{k+1}^{(n)}| \leq \\
 &\leq [2\gamma + (1 - 2\gamma)] \max_k |u_k^{(n)}| = \max_k |u_k^{(n)}|.
 \end{aligned}$$

Или:  $\max_k |u_k^{(n+1)}| \leq \max_k |u_k^{(n)}|$ , т.е. при переходе от момента  $t = n\tau$  к моменту  $t = n\tau + \tau$   $\max_k |\bar{u}|$  не возрастает.

Итак, при всех  $k$  и  $n$   $|u_k^{(n)}| \leq \max_k |u_k^{(0)}|$  или

$$|\bar{u}(x, t)| \leq \max |\bar{u}(x, 0)| \quad (13).$$

Определение устойчивости выполнено. Постоянная  $C$  здесь

оказалась равной 1 — не зависящей не только от  $\tau$  и  $h$  (что требуется), но и от промежутка изменения времени. Как и должно быть, при  $\gamma > \frac{1}{2}$  наше рассуждение не проходит: в оценку войдет величина  $|1 - 2\gamma| + 2\gamma$ , которая  $> 1$ .

К этому доказательству полезно сделать несколько замечаний.

1. Доказательство дословно повторяется для случая краевой задачи, если на границах заданы значения  $u$ ; при этом вместо (13) получается

$$|\tilde{u}(x, t)| \leq \max_{\Gamma} |\tilde{u}| \quad (14)$$

$\Gamma$  — та часть границы, где задаются краевые условия (см. рис. 4).

2. Свойство (14) — решение не превосходит максимума исходных данных — называют «принципом максимума». Как известно, исходное уравнение  $u_t = Du_{xx}$  обладает этим свойством.

Разностное уравнение (12) обладает им при  $\gamma \leq \frac{1}{2}$  и не обладает при  $\gamma > \frac{1}{2}$ . (В первом мы убедились, второе легко проверить).

3. В проведенном доказательстве нигде не используется постоянство  $D$ . Если  $D = D(x, t)$ , схема будет устойчивой при

$$\bar{\gamma} = \frac{\tau}{h^2} \max D(x, t) \leq \frac{1}{2}^* \quad (15)$$

4. Для уравнения  $u_t = cu_x$  тоже справедлив принцип максимума. При  $c > 0$  разностная система I (схема :.) сохраняет этот принцип, если  $\gamma = \frac{c\tau}{h} \leq 1$ . Отсюда вытекает устойчивость этой схемы. Коэффициент  $c$  снова может быть переменным.

---

\* Впрочем, при переменном коэффициенте теплопроводности правильнее писать  $\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( D \frac{\partial u}{\partial x} \right)$ . И здесь для устойчивости простейшей разностной схемы условия (15) достаточно.

## ЛЕКЦИЯ 6


### УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ. НЕЯВНЫЕ РАЗНОСТНЫЕ АППРОКСИМАЦИИ. ВОПРОС О ТОЧНОСТИ ПРИБЛИЖЕНИЙ

#### 1. Простейшая неявная схема

В предыдущей лекции мы выяснили, что при приближенном решении уравнений с частными производными на первый план выдвигается вопрос об устойчивости разностных аппроксимаций.

Ограничения из-за устойчивости часто бывают весьма жесткими. Классический пример — уравнение теплопроводности

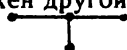
$$\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2}. \quad (1)$$

Если  $\tau$  — шаг по  $t$  и  $h$  — шаг по  $x$ , то условие устойчивости простейшей разностной схемы  таково:

$$\frac{D\tau}{h^2} \leq \frac{1}{2} \quad \text{или} \quad \tau \leq \frac{h^2}{2D}. \quad (2)$$

Это условие во многих случаях слишком ограничительно (требует слишком малого  $\tau$ ). Поэтому почти сразу, как только разностные аппроксимации стали всерьез применяться к уравнениям с частными производными, был предложен другой способ.

Его можно схематически изобразить так или формульно:



$$\frac{1}{\tau} (u_k^{n+1} - u_k^n) = \frac{D}{h^2} (u_{k+1}^{n+1} - 2u_k^{n+1} + u_{k-1}^{n+1}). \quad (3)$$

Здесь  $u_k^n = \tilde{u}(kh, n\tau)$ ,  $\tilde{u}(x, t)$ , как и раньше, приближенное решение (определенное только для дискретного набора точек).

Вы видите, что здесь нет «явной» формулы для нахождения значений на следующем слое по известным значениям на предыдущем слое — для значений  $u_k^{n+1}$  мы записали лишь систему уравнений, которую еще нужно решить. В этом и состоит радикальное отличие этой *неявной* аппроксимации от простейших аппроксимаций, рассмотренных в предыдущей лекции.

Немедленно возникает два вопроса.

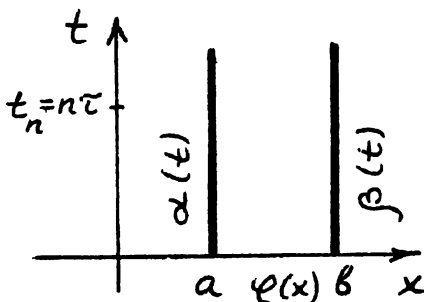
- 1) Как решить систему уравнений на очередном,  $n$ -ом слое?
- 2) Будет ли схема (3) более устойчива, чем простейшая *явная* схема — только в этом случае ее имеет смысл использовать. Забегая вперед, скажу ответ на второй вопрос: схема (3) устойчива при *любых*  $\tau$  и  $h$ . Это очень важно: мы можем выбирать такие  $\tau$  и  $h$ , какие требует сама задача — большие при плавном изменении  $u$  и меньшие при резком.

## 2. Решение уравнений на слое. Прогонка

Итак, пусть задано уравнение (1) и краевая задача для него:  $u(x, 0) = \varphi(x)$ ;  $u(a, t) = \alpha(t)$ ;  $u(b, t) = \beta(t)$ . Для нахождения  $u_k^n$  по известным  $u_k^{n-1}$  мы получим из (3) систему линейных уравнений:

$$\gamma u_{k+1}^n - (1 + 2\gamma) u_k^n + \gamma u_{k-1}^n = -u_k^{n-1}, \quad \gamma = \frac{D\tau}{h^2}. \quad (4)$$

Если добавить к (4) граничные условия  $u_0^n = \alpha(n\tau)$ ,  $u_K^n = \beta(n\tau)$ , то будет столько же уравнений, сколько неизвестных.



Заметим теперь, что система (4) имеет знакомый нам «треугольный» вид. Мы рассматривали такую систему во 2 лекции и убедились, что простейший метод исключения (без выбора главного элемента) допускает удобную формульную запись. Напишем эти формулы еще раз. Пусть система линейных уравнений имеет вид:

$$a_k y_{k-1} + b_k y_k + c_k y_{k+1} = f_k; \quad k = 0, 1, \dots, K; \quad (a_0 = c_K = 0) \quad (5)$$

Тогда:

1 этап

Находятся рекуррентно в порядке  $1, \dots, K-1, K$   $p_k$  и  $r_k$  по формулам

$$p_k = \frac{c_k}{b_k - a_k p_{k-1}}; \quad r_k = \frac{f_k - a_k r_{k-1}}{b_k - a_k p_{k-1}}. \quad (6)$$

При этом  $p_0 = c_0/b_0$ ;  $r_0 = f_0/b_0$ .

2 этап

Находятся рекуррентно в порядке  $K, K-1, \dots, 0$  неизвестные по формулам

$$y_k = r_k - p_k y_{k+1}, \quad y_K = r_K. \quad (7)$$

Может быть, не вредно еще раз подчеркнуть, что этот рецепт (называемый «прогонкой») не есть какой-то специальный метод решения линейных систем. Напротив, это есть простейший способ последовательного исключения неизвестных, записанный (для этой системы) рекуррентными формулами.

Итак, систему «на слое» (4) мы можем легко решить прогонкой (формулы (6) и (7)). Возникает лишь один вопрос: применим ли здесь этот способ (т. е. дает ли он решение с достаточной точностью)?

Ответить на него можно двумя способами:

1) Непосредственно показать, что при решении системы (4) по формулам (6) и (7) не придется делить на 0 и ошибки от округлений не будут сильно возрастать.

2) Убедиться, что для нашей системы *простейший* метод исключения совпадает с *полным* (включающим в себя выбор главного элемента). Иначе говоря, показать, что выбор главного элемента здесь не нужен, и после этого сослаться на недоказанную (и не сформулированную) теорему: всякая «порядочная» система может быть решена методом исключения с выбором главного элемента.

Логически первый путь лучше, так как не придется использовать недоказанных теорем. Второй способ, пожалуй, поучительнее (это доказательство намечено в дополнении).

### 3. Об устойчивости неявной схемы (4)

Начнем с исследования методом Фурье. Для этого мы поступим так. Отбросим границы и рассмотрим задачу по всей прямой  $-\infty < x < \infty$ . Теперь зададим начальные данные  $u(x, 0) = e^{isx}$  и ищем решение в виде:

$$\tilde{u}(x, t) = e^{i\sigma x + \lambda t} \quad \text{или} \quad u_k^{(n)} = e^{i\sigma k} \mu^n (\sigma = h\sigma, \mu = e^{\lambda\tau}). \quad (8)$$

Нас интересует, каким будет  $\mu$ . Подставив в (4) и сократив на  $\mu^{n-1}e^{i\sigma k}$ , получим  $\mu[\gamma e^{i\sigma} - (1+2\gamma) + \gamma e^{-i\sigma}] = -1$ , или

$$\mu = \frac{1}{1 + 4\gamma \sin^2 \sigma/2}. \quad (9)$$

При всех (вещественных)  $\sigma$  и всех  $\gamma > 0$   $\mu \leq 1$ , условие устойчивости снова выполнено в усиленном виде:

$$|\tilde{u}(x, t)| \leq |\tilde{u}(x, 0)|.$$

Итак, метод Фурье не обнаруживает неустойчивости ни при каких  $\tau$  и  $h$ . Конечно, мы не доказали устойчивости полностью, при любых начальных данных, но каждая отдельная «гармоника» не возрастает. Аккуратное доказательство абсолютной устойчивости схемы (4) (и притом сразу для краевой задачи и переменного  $D$ ) можно снова получить из принципа максимума\*. А пока мы вернемся к методу Фурье. При внимательном рассмотрении здесь вовсе не все так ясно.

Заметим, во-первых, что отказавшись от краевых условий, мы должны на очередном слое решать *бесконечную* систему линейных уравнений. Правда, эта система устроена просто — в каждое уравнение входит лишь 3 неизвестных, но все же это бесконечная система. Во-вторых, мы нашли *некоторое* решение задачи Коши для наших специальных начальных данных. Но нет ли другого? Собственно, этим вопросом следовало бы заняться заранее, до исследования устойчивости; попробуем разобратся.

Система уравнений на очередном слое (4) есть (при известном  $u_k^{(n-1)}$ ) линейное одномерное разностное уравнение второго порядка с постоянными коэффициентами. Обозначив временно  $u_k^n = u_k$ ,  $-u_k^{(n-1)} = f_k$ , мы можем записать его общее решение в виде  $u_k = u_k^{(n)} + v_k$ , где  $u_k^{(n)}$  *какое-нибудь* решение уравнения

$$\gamma u_{k+1} - (1+2\gamma)u_k + \gamma u_{k-1} = f_k \quad (10)$$

и  $v_k$  — общее решение соответствующего *однородного* уравнения

$$\gamma v_{k+1} - (1+2\gamma)v_k + \gamma v_{k-1} = 0. \quad (11)$$

Значит, решение задачи Коши для разностной системы (4) неединственно: на каждом слое ( $n = 1, 2, \dots$ ) можно добавить лю-

---

\* См. дополнение к этой лекции.

бое решение уравнения (11). Стало быть, мы поступили неаккуратно — просто «отбросить» границы нельзя! Однако, можно границы «отодвинуть на  $\infty$ » и потребовать *ограниченности*  $\bar{u}(x, t)$  при  $|x| \rightarrow \infty$ .

Покажем, что *ограниченное* решение (10) может быть только одно. Действительно, если бы было два таких решения, то их разность удовлетворяла бы уравнению (11). Общее решение (11) есть  $u_k = C_1 \mu_1^k + C_2 \mu_2^k$ , где  $\mu_1$  и  $\mu_2$  — корни характеристического уравнения

$$\gamma \mu^2 - (1 + 2\gamma) \mu + \gamma = 0.$$

Значит,  $|u_k| \leq \text{const}$  при всех  $k$  только при  $C_1 = C_2 = 0$ .

Итак, если мы хотим рассматривать неявную схему (3) (или (4)) на всей прямой, то нужно добавить требование: при каждом  $n$  для всех  $k$   $|u_k^n| \leq C$ . При таком понимании «задачи Коши» ее решение единственно\*, и удобное для метода Фурье рассмотрение задачи на всей прямой приобретает точный смысл.

#### 4. Еще один пример

Рассмотрим уравнение  $\frac{\partial u}{\partial t} = c \frac{\partial u}{\partial x}$  и две похожих разностных аппроксимации — явную и неявную:

$$\frac{1}{\tau} (u_k^{(n+1)} - u_k^{(n)}) = \frac{c}{2h} (u_{k+1}^{(n)} - u_{k-1}^{(n)}) \quad \text{схема } \begin{array}{c} | \\ \hline \end{array} \quad (12)$$

$$\frac{1}{\tau} (u_k^{(n+1)} - u_k^{(n)}) = \frac{c}{2h} (u_{k+1}^{(n+1)} - u_{k-1}^{(n+1)}) \quad \text{схема } \begin{array}{c} \hline | \end{array} \quad (13)$$

Рассмотрим снова (как в п. 3) задачу Коши на всей прямой в классе ограниченных функций и исследуем устойчивость методом Фурье. Положив  $u_k^{(n)} = \mu^n e^{i\sigma k}$ , мы получим

$$\mu = 1 + \frac{1}{2} \gamma (e^{i\sigma} - e^{-i\sigma}) = 1 + i\gamma \sin \sigma \quad \text{для схемы } \begin{array}{c} | \\ \hline \end{array}$$

$$\mu = (1 - i\gamma \sin \sigma)^{-1} \quad \text{для схемы } \begin{array}{c} \hline | \end{array}; \quad \gamma = \frac{c\tau}{h}.$$

Вы видите, что неявная схема (13) устойчива при любых  $\tau$  и  $h$ , точнее — не проявляет неустойчивости при проверке методом Фурье. Для явной схемы (12)  $|\mu_{\max}|^2 = 1 + \gamma^2 > 1$  при любом  $\gamma$ . Здесь метод Фурье, как будто бы, выявляет неустойчивость. Однако, давайте посмотрим более аккуратно.

\* Мы, однако, не доказывали, что при любых ограниченных начальных данных решение такой задачи Коши *существует*.



## 5. О необходимом условии устойчивости

У начинающих вычислителей часто складывается мнение, что при исследовании устойчивости разностной задачи Коши методом Фурье необходимо требовать, чтобы  $|\mu| \leq 1$ . Но, вообще, это не так: требование  $|\mu| \leq 1$  не является необходимым для устойчивости.

Рассмотрим для примера уравнение  $\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2} + au$ ;  $a > 0$ .

Перейдем к разностному уравнению:

$$\frac{1}{\tau} (u_k^{(n+1)} - u_k^{(n)}) = \frac{D}{h^2} (u_{k+1}^{(n)} - 2u_k^{(n)} + u_{k-1}^{(n)}) + au_k^{(n)}.$$

Подставив  $u_k^{(n)} = \mu^n e^{ik\sigma}$ , получим

$$\mu(\sigma) = 1 - 4\gamma \sin^2 \frac{\sigma}{2} + a\tau, \quad \gamma = \frac{D\tau}{h^2}.$$

Ни при каком  $\tau$  мы не можем добиться того, чтобы  $|\mu(\sigma)| \leq 1$  при всех  $\sigma$ . Таким образом, обязательно будут растущие решения задачи Коши.

В чем дело? Может быть нужно забраковать эту схему? Нет, схема не виновата: в исходном дифференциальном уравнении есть экспоненциально растущее решение  $u(x, t) = e^{at}$ . Ясно, что в приближающем его разностном уравнении тоже должны быть растущие решения.

Чтобы получить правильное ограничение на  $\mu$ , вернемся к определению устойчивости разностной задачи Коши. При фиксированных начальных данных мы требуем:  $|u(x, t)| \leq C$  при  $0 \leq t \leq T$  ( $C$  не зависит от  $\tau$  и  $h$ ). При  $u(x_k, t_n) = u_k^{(n)} = \mu^n e^{ik\sigma}$  получим  $|u(x, t)| = |\mu|^n$ . Обратите внимание:  $n$  пробегает не все возможные значения, а подчинено условию  $0 \leq n \leq T/\tau$ .

Итак, должно быть при  $\tau \rightarrow 0$  и  $h \rightarrow 0$   $|\mu(h, \tau)|^{T/\tau} \leq C$ . Это требование эквивалентно неравенству:

$$|\mu(h, \tau)| \leq 1 + a\tau^* \quad (\text{для допустимых } \tau \text{ и } h) \quad (14)$$

Вот это условие (а не условие  $|\mu| \leq 1$ ) действительно необходимо для устойчивости.

Теперь вернемся к схеме (12). Используя условие (14), мы видим, что эта схема формально устойчива\*\*, если в качестве допустимых взять  $\tau$  и  $h$ , подчиненные неравенству  $\gamma^2 \leq a\tau$ , или:  $\tau/h^2 \leq b$  (с любым  $b$ ). Так что, строго говоря, высказывание

\* При этом  $C = \exp(aT)$ .

\*\* Точнее, выдерживает проверку на устойчивость методом Фурье.

«схема (12) неустойчива» — неправильно. Тем не менее, обычно так говорят. При этом имеют в виду следующее:

1) Условие  $\tau/h^2 \leq b$  практически неудобно: при малых  $h$  диктует слишком малые  $\tau$ .

2) Произвол в выборе  $b$  свидетельствует о формальном характере этого условия — оно никак не отражает свойств дифференциального уравнения  $u_t = cu_x$ . Для контраста рассмотрим знакомое нам условие  $c\tau/h \leq 1$ . Это условие означает, что  $c\tau \leq h$ , т. е. за время  $\tau$  возмущения, движущиеся в силу дифференциального уравнения, должны проходить путь не больше  $h$  — ясное и наглядное требование.

3) Разностное уравнение (12) имеет растущие по времени решения  $u(x, t)$  (рост тем быстрее, чем больше  $b$ ). Исходное дифференциальное уравнение  $u_t = cu_x$  растущих решений не имеет.

## 6. Типичная картина при возникновении неустойчивости

Предположим, что по ошибке (или в целях эксперимента) при вычислениях на машине использована неустойчивая схема. Что получается? Если неустойчивость сильная, то довольно скоро получится «авост»\* — машина остановится (или займется другими задачами). Если же неустойчивость слабая, то она проявится в возникновении характерных осцилляций. Картинка в некоторый момент времени будет примерно такая



Почему возникает такая «пила»? Я сейчас это объясню, а потом извлеку из этой неприятной для вычислителя картины некоторые позитивные выводы. Вернемся к классическому примеру  $u_t = Du_{xx}$ . Возьмем снова начальные данные  $u(x, 0) = \exp(isx)$  и простейшую разностную схему  $\frac{u_k^{(n)} - u_k^{(n-1)}}{\tau} = \frac{u_{k+1}^{(n-1)} - 2u_k^{(n-1)} + u_{k-1}^{(n-1)})}{h^2}$ . Тогда  $u_k^{(n)} = u(kh, n\tau) = (\mu(s))^n e^{ik\sigma}$ , где  $\sigma = hs$  и  $\mu(s) = 1 - 4\gamma \sin^2 \sigma/2$ . Если зафиксировать  $s$  и  $\gamma$ , а  $\tau$  и  $h$  устремить к 0, то с некоторого момента будет  $|\mu(s)| \leq 1$ .

Откуда же берется неустойчивость при  $\gamma > 1/2$ ? При  $\gamma > 1/2$  наиболее неустойчивой будет гармоника, для которой  $\sin \sigma/2 = 1$ , т. е.  $\sigma = \pi$ ,  $s = \pi/h$ . Это отвечает начальным данным  $u(x, 0) = e^{i\pi x/h}$  или  $u_k^{(0)} = (-1)^k$ . Таков формальный ответ.

По существу же — это самая «высокая гармоника» на сетке. Как понимать такое высказывание? Конечно, можно взять  $e^{isx}$  при  $s > \pi/h$ , но на сетке  $x_k = kh$

$$e^{isx} = e^{is_1 x},$$

где уже  $|s_1| \leq \pi/h$  ( $s_1 = s - 2\pi m/h$  при целом  $m$ ).

\* То есть вычисления прекратятся из-за появления слишком больших чисел.

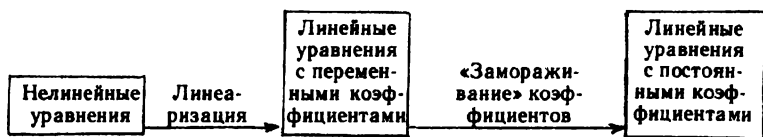
Иначе говоря, при наличии минимального расстояния  $h$  по  $x$  есть максимальная частота  $s_{max} = \pi/h$ . При уменьшении  $h$  она увеличивается. Такая двойственность между малыми расстояниями и высокими частотами играет очень большую роль в квантовой механике и во многих разделах математического анализа. Итак, в нашем примере самыми неустойчивыми являются гармоника с самой высокой частотой. Произвольные начальные данные содержат высокие гармоники (кроме того, они появляются за счет округления)\*. Если есть неустойчивость, эти гармоники вырастают быстрее других и появляются частые колебания по  $x$ . Это — сигнал для вычислителя: «ищи неустойчивость».

Пример с уравнением теплопроводности является типичным. Каждая фиксированная гармоника  $e^{isx}$  при  $t$  и  $h$ , достаточно малых, изменяется в разностном уравнении почти так же, как в дифференциальном\*\*. Но с уменьшением  $h$  «вступают в игру» все более высокочастотные колебания по  $x$ . Они-то и могут сильно вырасти на заданном промежутке времени.

## 7. Линеаризация и «замораживание» коэффициентов

При исследовании устойчивости задачи Коши для разностных уравнений наиболее «ходовой» рабочий инструмент — метод Фурье. Но он легко применяется лишь к линейным уравнениям с постоянными коэффициентами. Практически мы имеем обычно дело с переменными коэффициентами, а еще чаще с нелинейными уравнениями.

Чтобы исследовать устойчивость в такой задаче, совершают следующие две процедуры. Во-первых, линеаризуют уравнение (и граничные условия) вблизи начальных данных или данных на очередном слое. Во-вторых, «замораживают» коэффициенты, т. е. берут вместо переменных коэффициентов — постоянные (равные значениям их в некоторой точке).



\* Немного более аккуратно. Нужно записать

$$u_k^{(0)} = \frac{1}{2\pi} \int_{-\pi}^{\pi} a(\sigma) e^{ik\sigma} d\sigma \quad (\text{при этом } a(\sigma) = \sum_{m=-\infty}^{\infty} u_m^{(0)} e^{-im\sigma}).$$

$$\text{Тогда } u_k^{(n)} = \frac{1}{2\pi} \int_{-\pi}^{\pi} (\mu(\sigma))^n a(\sigma) e^{ik\sigma} d\sigma.$$

\*\* Мы уже обсуждали это в 5 лекции при «доказательстве» теоремы.

Наконец, исследуют получившуюся простую задачу. Если она оказывается устойчивой, то полагают устойчивой и исходную задачу.

Первый шаг более или менее понятен. Мы двигаемся малыми (для данной задачи) шагами по времени. За малое время система ведет себя приблизительно как линейная. Но почему можно заменить переменные коэффициенты — постоянными? Коэффициенты вовсе не являются слабо меняющимися по  $x$ ! Картинки возникновения неустойчивости проливают некоторый свет на этот вопрос. Дело в том, что неустойчивость возникает «на высоких частотах», за счет осцилляций  $u$ , сравнимых с масштабом сетки  $h$ . Если  $h$  достаточно мало, то по сравнению с этими быстрыми осцилляциями коэффициенты меняются очень медленно и плавно. Вот почему при исследовании устойчивости столь много ценных сведений может дать такой грубый прием, как замена переменных коэффициентов — постоянными. Разумеется, все эти описательные рассуждения не есть доказательство, но во многих случаях, аккуратно оговорив условия, можно дать «принципу замораживания коэффициентов» строгое доказательство.

## 8. Понятие о точности разностных аппроксимаций дифференциальных уравнений

В отличие от проблемы устойчивости этот вопрос есть смысл обсудить в самом общем виде. Ненадолго мы забудем даже, что нас интересуют дифференциальные уравнения.

Итак, пусть есть уравнение

$$I \quad F[u] = 0.$$

Вместо этого уравнения мы решаем некоторое приближенное уравнение II

$$\tilde{F}[\tilde{u}] = 0.$$

Как узнать, насколько мы ошибаемся? Проще всего было бы из  $u$  вычесть  $\tilde{u}$  и оценить величину разности. Но решение I нам неизвестно. Можно попробовать поступить так. Подставить  $\tilde{u}$  в точное уравнение I. Получится  $F[\tilde{u}] = \delta$ . Если  $\delta$  — мало, то, наверное, приближение хорошее. Для интересующего нас случая, когда I — дифференциальное уравнение, а II — разностное, возникает следующее затруднение. Функция  $\tilde{u}(x, t)$  задана только для дискретного набора точек. Раньше, чем подставить  $\tilde{u}$  в дифференциальное уравнение, нужно превратить ее в достаточно хорошую функцию непрерывного аргумента. Это можно сделать, но требует некоторых хлопот.

Поэтому обычно поступают наоборот. Берут (конечно, мысленно) точное решение  $u(x, t)$  и подставляют в приближенное уравнение II.

Итак, рассмотрим дифференциальное уравнение для опреде-

ления функции  $u(x, t): F[u] = 0$  и разностное уравнение для определения функции дискретного аргумента  $\tilde{u}(x, t)$

$$(x = kh, t = n\tau): \tilde{F}[\tilde{u}] = 0.$$

Пусть  $\tau$  и  $h$  стремятся к 0, подчиняясь, возможно, некоторым ограничениям. Если при этом для любой достаточно гладкой функции  $\varphi(x, t)$   $\tilde{F}_{\tau, h}[\varphi] \rightarrow F[\varphi]^*$ , то говорят, что  $\tilde{F}_{\tau, h}$  аппроксимирует  $F$ .

Для точного решения  $u(x, t)$  уравнения  $F[u] = 0$  мы получим:  $\tilde{F}_{\tau, h}[u] = r_{\tau, h} \rightarrow 0$ .

По скорости стремления  $r$  к 0 судят о точности аппроксимации. В частности, если  $r_{\tau, h} = C_1\tau^p + C_2h^q + o(\tau^p + h^q)$ , то говорят, что рассматриваемая разностная аппроксимация имеет порядок  $p$  по  $t$  и порядок  $q$  по  $x$ .

**Важное замечание.**

Условие  $\tilde{F}_{\tau, h}[\varphi] \rightarrow F[\varphi]$  требует определенной записи разностного уравнения. Скажем, нельзя записать  $\tilde{F}$  для

$$F[\varphi] = \frac{\partial \varphi}{\partial t} - D \frac{\partial^2 \varphi}{\partial x^2}$$

в виде

$$\begin{aligned} \tilde{F}_{\tau, h}[\varphi] &= \varphi(x, t + \tau) - \varphi(x, t) - \\ &- \frac{D\tau}{h^2} (\varphi(x + h, t) - 2\varphi(x, t) + \varphi(x - h, t)). \end{aligned}$$

(Тогда при  $\tau$  и  $h \rightarrow 0$   $\tilde{F}_{\tau, h}[\varphi] \rightarrow 0$ ). Правильная запись такова:

$$\begin{aligned} \tilde{F}_{\tau, h}[\varphi] &= \frac{1}{\tau} (\varphi(x, t + \tau) - \varphi(x, t)) - \\ &- \frac{D}{h^2} (\varphi(x + h, t) - 2\varphi(x, t) + \varphi(x - h, t)). \end{aligned}$$

Рассмотрим пример.

$$\text{Пусть } L[u] = \frac{\partial u}{\partial t} - c \frac{\partial u}{\partial x},$$

$$\tilde{L}_{\tau, h}[\tilde{u}] = \frac{1}{\tau} (\tilde{u}(x, t + \tau) - \tilde{u}(x, t)) - \frac{c}{h} (\tilde{u}(x + h, t) - \tilde{u}(x, t)).$$

Для произвольной (дифференцируемой) функции  $\varphi(x, t)$  при  $\tau$  и  $h \rightarrow 0$

$$\tilde{L}_{\tau, h}[\varphi] \rightarrow \frac{\partial \varphi}{\partial t} - c \frac{\partial \varphi}{\partial x}.$$



Таким образом,  $\tilde{L}$  аппроксимирует  $L$ . Пусть теперь  $\frac{\partial u}{\partial t} - c \frac{\partial u}{\partial x} = 0$ . Вычисляя  $r_{\tau, h} = \tilde{L}_{\tau, h}[u]$ , получим:

\* Это не совсем аккуратная запись, но ее нетрудно уточнить.

$$r_{\tau,h} = u_t(x,t) + \frac{\tau}{2} u_{tt}(x,t) - cu_x(x,t) - \frac{ch}{2} u_{xx}(x,t) + \\ + O(\tau^2 + h^2) = \frac{\tau}{2} u_{tt}(x,t) - \frac{ch}{2} u_{xx}(x,t) + O(\tau^2 + h^2).$$

Таким образом, схема имеет первый порядок точности по  $x$  и по  $t^*$ . Заметьте, что точность аппроксимации одинакова как для случая  $c > 0$ , когда эта разностная схема пригодна, так и для случая  $c < 0$ , когда она совершенно не годится!

Аналогично можно разобрать и другие примеры. Я ограничусь еще двумя, которые приведу без вычислений.

1. Для уравнения  $u_t = Du_{xx}$  простейшая явная схема  и простейшая неявная схема  обе имеют первый порядок точности по  $t$  и второй по  $x$ , т. е. при малых  $\tau$  и  $h$

$$r_{\tau,h} \approx A(x,t) \cdot \tau + B(x,t) \cdot h^2. \quad (15)$$

Заметьте, что в случае явной схемы мы вынуждены брать  $\tau \sim h^2$ , так что оба слагаемых в (15) имеют один порядок малости. В случае неявной схемы (когда вполне допустимо  $\tau \sim h$ ) желательно уменьшить первое слагаемое. Это можно сделать, если использовать симметрично 6 точек

$$\begin{array}{ccccccc} n+1 & & & & & & \\ & \cdot & \cdot & \cdot & & & \\ n & & & & & & \\ & \cdot & \cdot & \cdot & & & \\ & k-1 & k & k+1 & & & \end{array}$$

2. Для волнового уравнения  $u_{tt} = c^2 u_{xx}$  простейшая разностная схема такова 

$$\frac{1}{\tau^2} (u_k^{n+1} - 2u_k^n + u_k^{n-1}) = \frac{c^2}{h^2} (u_{k+1}^n - 2u_k^n + u_{k-1}^n).$$

Она имеет второй порядок точности по  $x$  и по  $t$ . Условие устойчивости такой схемы дает знакомое нам неравенство, имеющее простой наглядный смысл:  $|c| \tau \leq h$ .

### 9. Определение порядка точности при приближенном интегрировании обыкновенных дифференциальных уравнений

Напомню, что для обыкновенных дифференциальных уравнений точность приближенного метода мы оценивали по ошибке, совершаемой на одном шаге. Такое определение приспособлено к задаче Коши и даже в этом случае не годится для много-

\* Если  $\tau$  и  $h \rightarrow 0$  так, что  $\frac{c\tau}{h} = 1$ , то точность схемы повышается. Проверьте это!

точечных методов (типа метода Адамса). Теперь мы можем легко дать общее определение.

### О п р е д е л е н и е

Пусть обыкновенное дифференциальное уравнение  $D[u]=0$  заменено при каждом  $h$  системой уравнений  $\bar{D}_h[u]=0$  для определения приближенных значений  $u(kh)=u_k$ . Это приближение имеет (локальный) порядок точности  $m$ , если:

- а) при  $h \rightarrow 0$   $\bar{D}_h[\varphi] \rightarrow D[\varphi]$  для произвольной функции  $\varphi^*$ ;
- б) при подстановке точного решения  $u(x)$  интересующей нас задачи в приближенное уравнение получается

$$\bar{D}_h[u] = A(x)h^m + o(h^m).$$

### П р и м е р: Метод Адамса

Для уравнения  $\frac{du}{dx} = f(u)$  формула метода Адамса «второго порядка» имеет вид:

$$u_{k+1} = u_k + h \left( \frac{3}{2} f(u_k) - \frac{1}{2} f(u_{k-1}) \right). \quad (16)$$

Запишем:

$$\begin{aligned} D[u] &= \frac{du}{dx} - f(u) \\ \bar{D}_h[\tilde{u}] &= \frac{1}{h} [\tilde{u}(x+h) - \tilde{u}(x)] - \frac{3}{2} f(\tilde{u}(x)) + \frac{1}{2} f(\tilde{u}(x-h)). \end{aligned} \quad (17)$$

При  $h \rightarrow 0$  для любой функции  $\varphi(x)$   $\bar{D}_h[\varphi] \rightarrow \frac{d\varphi}{dx} - f(\varphi)$ , т. е. запись (17) правильная. Пусть теперь

$$D[u] = 0, \quad \text{т. е.} \quad u'(x) = f(u(x)).$$

Тогда

$$D_h[u] = \frac{1}{h} (u(x+h) - u(x)) - \frac{3}{2} u'(x) + \frac{1}{2} u'(x-h) = O(h^2).$$

Итак, формула (15) дает метод второго порядка точности в смысле нашего общего определения.

## 10. Заключительные замечания

Заканчивая предварительное рассмотрение уравнений с частными производными, заметим, что мы оставили без ответа (и даже без обсуждения) основной вопрос: насколько отличается найденное нами приближенное решение  $u(x, t)$  от точного? В описательных терминах ответ дать нетрудно. Именно, если раз-

\* Имеющей достаточное число производных.

ностная схема *устойчива* и *аппроксимирует* дифференциальное уравнение, то при уменьшении  $\tau$  и  $h$   $u(x, t)$  приближается к  $u(x, t)$ .

Скорость приближения определяется точностью аппроксимации.

Для аккуратной формулировки нужно несколько расширить требования устойчивости: необходима устойчивость не только относительно изменения начальных (и краевых) условий, но и относительно изменения правой части (линейного) уравнения. После такого расширения соответствующая теорема (для линейных уравнений!) легко доказывается.

## Дополнения и примечания к 6 лекции

1) (к п. 2). Рассмотрим систему линейных уравнений с трехдиагональной матрицей

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ a_1 & b_1 & c_1 & 0 & \dots & 0 & 0 & 0 \\ 0 & a_2 & b_2 & c_2 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}$$

Покажем, что метод исключения с выбором главного элемента (начиная со 2 шага) будет совпадать с простейшим методом исключения, если:

$$(1) \quad |b_k| > |a_k| + |c_k|; \quad (2) \quad |b_k| > |a_k| + |a_{k+1}|.$$

В частности при постоянных, не зависящих от  $k$  коэффициентах достаточно:

$$(1') \quad |b| > |a| + |c|; \quad (2') \quad |b| > 2|a|.$$

Доказательство проведем индукцией по числу шагов исключения. После первого шага мы получим

$$\begin{pmatrix} 1 & 0 & 0 & \dots \\ 0 & 1 & p_1 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix} \quad \text{где} \quad |p_1| = \frac{|c_1|}{|b_1|} < 1.$$

Пусть теперь  $|p_{k-1}| < 1$  и на  $k$ -м шаге не понадобилось переставлять строки. Докажем, что на  $(k+1)$ -ом шаге также не понадобится переставлять строки и, что  $|p_k| < 1$ .

$$\begin{array}{ccccccccc} \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & p_{k-1} & 0 & 0 & & 0 & 1 & p_{k-1} & & 0 & 0 \\ 0 & a_k & b_k & c_k & 0 & \longrightarrow & 0 & 0 & b_k - a_k p_{k-1} & c_k & 0 \\ 0 & 0 & a_{k+1} & b_{k+1} & c_{k+1} & & 0 & 0 & a_{k+1} & & b_{k+1} & c_{k+1} \\ \dots & \dots & \dots & \dots & \dots & & \dots & \dots & \dots & \dots & \dots \end{array}$$



В силу предположения  $|p_{k-1}| < 1$  имеем  $|b_k - a_k p_{k-1}| > |b_k| - |a_k| > |a_{k+1}|$  — перестановка строк не нужна. Теперь

$$|p_k| = \left| \frac{c_k}{b_k - a_k p_{k-1}} \right| < \frac{|c_k|}{|b_k| - |a_k|} < 1.$$

Индукция проведена. Заметим, что для схемы (4) условия (1'), (2') выполнены.

2) (к п. 3). Покажем, что разностная система (3), дополненная краевыми условиями:

заданы  $u_0^n = \alpha_n$  и  $u_K^n = \beta_n$ , удовлетворяет принципу максимума, т.е.

$$|u_k^n| \leq \max(|u_j^0|, |\alpha_n|, |\beta_n|); \quad 0 \leq k \leq K, \quad 0 \leq n \leq N. \quad (18)$$

Предположим, что  $\max_k u_k^n$  достигается во внутренней точке  $k_0$  ( $1 \leq k_0 \leq K-1$ ). Тогда  $u_{k_0-1}^n - 2u_{k_0}^n + u_{k_0+1}^n \leq 0$  и из (3) получим:

$$u_{k_0}^n - u_{k_0}^{n-1} \leq 0, \quad \text{т.е.} \quad \max_k u_k^n \leq \max_k u_k^{n-1}.$$

Аналогично для внутреннего минимума  $\min_k u_k^n \geq \min_k u_k^{n-1}$ .

Если  $\max_k |u_k^n|$  достигается на границе, то он равен  $|\alpha_n|$  или  $|\beta_n|$ .

Окончательно  $\max_k |u_k^n|$  не превосходит наибольшего из чисел:  $\max_k |u_k^{n-1}|$ ,  $|\alpha_n|$ ,  $|\beta_n|$ . Отсюда, двигаясь по  $n$  от 0 до  $N$ , получаем (18).

## ЛЕКЦИЯ 7

### КРАЕВЫЕ ЗАДАЧИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ \*

#### 1. Предварительные замечания

Название этой темы требует некоторых пояснений. Мы не будем заниматься общими краевыми задачами для систем дифференциальных уравнений — они гораздо труднее, чем задача Коши. Даже для одного уравнения второго порядка, если оно нелинейно, краевая задача может оказаться очень трудной. Мы ограничимся поэтому линейными дифференциальными уравнениями. При этом все основные идеи хорошо видны уже на примере одного дифференциального уравнения 2 порядка. Его мы и рассмотрим.

#### 2. Постановка задачи

Пусть дано линейное уравнение

$$y''(x) + q(x)y(x) = f(x), \quad a \leq x \leq b. \quad (1)$$

Требуется найти решение этого уравнения, удовлетворяющее двум (линейным) краевым условиям — в точке  $a$ :

$$\alpha_a y(a) + \beta_a y'(a) = \gamma_a \quad (1a)$$

в точке  $b$ :

$$\alpha_b y(b) + \beta_b y'(b) = \gamma_b. \quad (1b)$$

---

\* Текст этой лекции труднее предыдущих. Пункты, отмеченные \*, могут быть при первом чтении пропущены.

Такую задачу мы будем для краткости называть «задача I». Что известно о решении такой задачи? Рассмотрим наряду с задачей I соответствующую однородную задачу II:

$$z''(x) + q(x)z(x) = 0 \quad (2)$$

$$\alpha_a z(a) + \beta_a z'(a) = 0 \quad (2a)$$

$$\alpha_b z(b) + \beta_b z'(b) = 0. \quad (2b)$$

Справедливы следующие два утверждения.

1. Если однородная задача (II) имеет только нулевое решение  $z(x) \equiv 0$ , то неоднородная задача (I) имеет решение (при любом  $f$  и любых  $\gamma$ ) и притом только одно.

2. Если однородная задача (II) имеет ненулевое решение ( $z(x) \not\equiv 0$ ), то соответствующая задача I имеет решение не всегда, а лишь тогда, когда  $f(x)$  и  $\gamma$  удовлетворяют некоторым условиям (в конечном числе).

Таким образом, положение дел здесь вполне аналогично тому, которое наблюдается для систем (конечного числа) линейных алгебраических уравнений: если однородная система имеет только тривиальное решение, (т.е. определитель системы не равен нулю), то неоднородная система всегда имеет единственное решение. Если же однородная система имеет нетривиальное решение (т.е. определитель равен нулю — система вырождена), то для существования решения нужно наложить несколько (линейных) условий на правые части неоднородной системы\*.

Со времени работ Фредгольма по интегральным уравнениям такого рода ситуация называется в математике альтернативной Фредгольма. Мы будем предполагать, что решение интересующей нас задачи (I) существует и единственно (т.е. имеет место первый случай альтернативы). Наша задача — найти это решение. Заметьте, что в теории дифференциальных уравнений такой вопрос специально не рассматривается — он кажется совершенно тривиальным.

### 3. Способ решения краевой задачи, вытекающей из теории дифференциальных уравнений

Известно, что общее решение неоднородного уравнения (1) есть сумма какого-либо одного частного решения и общего решения однородного уравнения:  $y(x) = y_n(x) + z(x)$ . В свою очередь, общее решение (2) есть линейная комбинация 2 частных решений.

Итак:

$$\text{Общее решение (2): } z(x) = C_1 z_1(x) + C_2 z_2(x) \quad (3)$$

---

\* Число таких условий  $m$  равно степени вырождения системы, т.е. числу уравнений однородной системы, которые являются комбинациями остальных уравнений.

$$\text{общее решение (1): } y(x) = y_n(x) + C_1 z_1(x) + C_2 z_2(x). \quad (4)$$

Чтобы выбрать из всех решений (1) то единственное, которое удовлетворяет краевым условиям, нужно найти  $C_1$  и  $C_2$ .

Пусть, для простоты, краевые условия имеют вид:

$$y(a) = A, \quad y(b) = B.$$

Тогда для определения  $C_1$  и  $C_2$  мы получим следующую систему 2 линейных уравнений

$$\begin{cases} A = y_n(a) + C_1 z_1(a) + C_2 z_2(a) \\ B = y_n(b) + C_1 z_1(b) + C_2 z_2(b). \end{cases} \quad (5)$$

Итак, достаточно найти одно (любое!) решение уравнения (1) и два (непропорциональных) решения уравнения (2). Это можно сделать, решая задачу Коши — один раз для (1) и два раза для (2). Таким образом, решение краевой задачи (1) сводится в основном к решению *трех* задач Коши.

Нетрудно сообразить, что можно обойтись решением двух задач Коши. Для этого нужно  $y_n(x)$  и  $z(x)$  выбирать не совсем произвольно, а так, чтобы они уже удовлетворяли одному из граничных условий. Например, при граничном условии  $y(a) = A$  можно взять в качестве начальных данных для задач Коши:

$$\begin{aligned} y_n(a) &= A, & y'_n(a) &= 0; \\ z(a) &= 0, & z'(a) &= 1. \end{aligned}$$

Всякое решение уравнения (1), у которого  $y(a) = A$ , есть комбинация этих двух:

$$y(x) = y_n(x) + Cz(x). \quad (6)$$

Для определения  $C$  нужно опять использовать правое граничное условие. Если задано  $y(b) = B$ , то

$$C = \frac{1}{z(b)} (B - y_n(b)). \quad (7)$$

Заметьте, что  $z(b) \neq 0$ . В самом деле. Если бы  $z(b) = 0$ , то функция  $z(x)$  удовлетворяла бы однородному уравнению (2) и двум однородным граничным условиям  $z(a) = z(b) = 0$ , т.е. должно быть (как мы предположили с самого начала)  $z(x) \equiv 0$ . Однако, это невозможно, так как  $z'(a) = 1$ .

#### 4. К чему может привести «сокращение знаков»

Коль скоро мы умеем численно решать задачу Коши, то способ, который был только что описан, позволяет решить краевую задачу (1): она сводится при этом к двум задачам Коши. Во

многих случаях этот способ дает прекрасный результат. Но, пожалуй, чаще этот теоретически безупречный рецепт практически непригоден.

Дело здесь в следующем. Возникшая из конкретной проблемы краевая задача имеет обычно (при подходящем выборе единиц измерения) решение  $y(x) \sim 1$ . Но частные решения, которые мы выбираем ( $y_n(x)$  и  $z(x)$ ) могут оказаться быстро растущими. Если  $y_n(x)$  и  $z(x)$  велики на части отрезка  $[a, b]$ , то при нахождении  $y(x)$ :  $y(x) = y_n(x) + Cz(x)$  произойдет вычитание двух больших величин, имеющих разность  $\sim 1$ . При этом неизбежно сокращение верных значащих цифр.

Таким образом, естественный способ решения краевой задачи пригоден лишь в тех случаях, когда частные решения  $y_n(x)$  и  $z(x)$  не являются быстро растущими. Для уравнения (1) это обычно бывает так при  $q(x) > 0$  (решение однородного уравнения (2) имеют тогда «колебательный характер»). При  $q(x) < 0$  точность решения, получаемого по формуле (6), будет тем хуже, чем больше  $|q(x)|$  (при заданном интервале  $(a, b)$ ).

Возникает вопрос. Если естественный способ решения краевой задачи отказывается, то, может быть, сама краевая задача «плохая» и ее вообще не следует решать? Мы покажем сейчас на примере постоянного  $q(x)$ , что в некотором смысле все обстоит наоборот: Чем больше  $|q(x)|$  при  $q(x) < 0$ , тем «лучше» краевая задача!

## 5. \* Отступление.

### Что значит «хорошая» краевая задача?

Вычислительную задачу естественно считать хорошо поставленной, если небольшим изменениям входных данных отвечают небольшие изменения решения.

Рассмотрим следующую краевую задачу:

$$y''(x) - k^2 y(x) = f(x), \quad y(0) = y(2\pi) = 0. \quad (8)$$

Ввиду линейности этой задачи достаточно потребовать, чтобы малым  $f(x)$  отвечали небольшие  $y(x)$ :

$$\max |y(x)| \leq M \max |f(x)|. \quad (9)$$

Оценим  $y(x)$ . Пусть в точке  $x_0$  находится положительный абсолютный максимум  $y(x)$ . Тогда  $y''(x_0) \leq 0$ . Из уравнения:  $k^2 y(x_0) = y''(x_0) - f(x_0)$ . Значит,  $f(x_0) \leq 0$ ;  $k^2 y(x_0) = |f(x_0)| + y''(x_0) \leq |f(x_0)|$ .

Итак,  $y(x_0) \leq \frac{1}{k^2} |f(x_0)|$ . Аналогичная оценка справедлива и для отрицательного минимума  $y(x)$ , и мы можем написать:  $\max |y(x)| \leq \frac{1}{k^2} \max |f(x)|$ , т. е. в формуле (9)  $M = \frac{1}{k^2}$ . Что касается ре-

шения задачи Коши, то для однородного уравнения  $z'' - k^2 z = 0$  при начальных условиях  $z(0) = 0$ ,  $z'(0) = 1$  получим  $z(x) = \frac{1}{k} \operatorname{sh} kx$ .

Чем больше  $k$ , тем быстрее растет эта функция.

Итак, имеет место следующая парадоксальная на первый взгляд ситуация: чем больше  $k$ , тем «лучше» краевая задача (8) и тем «хуже» соответствующая задача Коши. При больших  $k$  свести решение краевой задачи к задачам Коши нельзя\*, но отказаться от решения краевой задачи мы не имеем права — математически она хорошо поставлена.

## 6. Разностная задача, отвечающая краевой задаче (I)

Заменяем дифференциальное уравнение  $y'' + q(x)y = f(x)$  (1) разностным уравнением

$$\frac{1}{h^2} (y_{k+1} - 2y_k + y_{k-1}) + q_k y_k = f_k; \quad h = \frac{b-a}{n};$$

$$q_k = q(x_k), \quad f_k = f(x_k); \quad x_k = a + kh; \quad k = 0, \dots, n. \quad (10)$$

Добавим к этому граничные условия (в простейшем случае:  $y_0 = A$ ,  $y_n = B$ ); мы получаем «трехдиагональную» систему из  $n+1$  линейных уравнений для определения неизвестных

$$y_0, y_1, \dots, y_n.$$

Такая система легко решается методом исключения — совсем легко, если применим простейший метод исключения («прогонка») и немного сложнее, если нужно выбирать главный элемент\*\*. В любом случае проблемы здесь нет.

Трехдиагональная система получится не только для уравнения (1), но и для более общего уравнения

$$y'' + c(x)y' + q(x)y = f(x).$$

Если точность аппроксимации производных (по трем точкам) покажется недостаточной, можно использовать более точные формулы. Мы получим тогда, скажем 5-диагональную систему линейных уравнений, которую тоже нетрудно решить методом исключения.

Таким образом, с чисто практической точки зрения вопрос о решении краевых задач для линейных дифференциальных уравнений второго порядка можно считать решенным вполне удовлетворительно. Однако, остается некоторое чувство неудовлет-

\* Конкретно для (8) уже при  $k \sim 4-5$ .

\*\* Кстати, легко проверить, что как раз для «трудного» случая  $q(x) < 0$  применима прогонка (см. добавление к 6 лекции).

воренности. Почему все-таки мы не могли наметить план решения, рассматривая дифференциальные, а не разностные уравнения?

## 7. \* Идея переноса граничных условий

Проанализируем еще раз сложившуюся ситуацию. Мы знаем, что искомое решение задачи I находится среди функций вида

$$y(x) = y_n(x) + Cz(x), \quad (6)$$

где  $y_n(x)$  и  $z(x)$  удовлетворяет граничному условию при  $x=a$  (соответственно неоднородному и однородному.)

Обратите внимание: в записи (6) использовано только левое граничное условие (правое понадобится для нахождения  $C$ ). Таким образом, левое граничное условие из *всех* решений неоднородного уравнения  $y'' + q(x)y = f(x)$  выбирает некоторое семейство решений  $S_-$  — тех, которые представлены в виде (6). Сделаем теперь следующий шаг. Заметим, что само это семейство решений  $S_-$  не зависит от того, как именно выбраны  $y_n(x)$  и  $z(x)$ :  $S_-$  состоит просто из *всех* решений уравнения (1), удовлетворяющих левому граничному условию. Мы можем задавать  $S_-$  с помощью другой пары функций  $y_n(x)$  и  $z(x)$ , можем использовать различные  $y_n(x)$  и  $z(x)$  на различных частях отрезка, или задавать  $S_-$  еще каким-нибудь способом. Конкретно мы поступим так: установим связь между значениями  $y(x)$  и  $y'(x)$  (при одном  $x$ ) для функций, принадлежащих этому семейству.

Зафиксируем некоторое значение  $x=x_0$ . Тогда заданием  $y(x_0)$  и  $y'(x_0)$  однозначно\* определится некоторое решение уравнения  $y'' + q(x)y = f(x)$ . Если задавать *произвольно*  $y(x_0)$  и  $y'(x_0)$ , мы получим *все* решения этого уравнения, а не только те, которые удовлетворяют левому граничному условию. Значит, для  $y$  из  $S_-$  есть связь между  $y(x_0)$  и  $y'(x_0)$ . Легко показать, что эта связь будет линейной\*\*.

Итак, если  $y \in S_-$ , то для каждого  $x$ :

$$\alpha_-(x)y(x) + \beta_-(x)y'(x) = \gamma_-(x). \quad (11)$$

Соотношение (11) вполне аналогично левому граничному условию

$$\alpha_a y(a) + \beta_a y'(a) = \gamma_a.$$

и совпадает с ним при  $x=a$ . Таким образом, *следствием* левого граничного условия (1a) является такого же вида связь при каждом  $x$ . Образно говоря, если мы сумеем достаточно точно найти  $\alpha_-(x)$ ,  $\beta_-(x)$ ,  $\gamma_-(x)$ , то «пронесем» граничное условие с левого

\* В силу теоремы единственности для задачи Коши.

\*\* Продифференцировав (6) и исключив  $C$ .

конца через весь отрезок\*. На правой границе мы получим тогда два условия: принесенное  $\alpha_{\rightarrow}(b)y(b) + \beta_{\rightarrow}(b)y'(b) = \gamma_{\rightarrow}(b)$  и заданное с самого начала  $\alpha_b y(b) + \beta_b y'(b) = \gamma_b$ . Теперь кажется, что мы избавились от главной трудности в краевой задаче — получили два условия на *одном* конце и можем решать задачу Коши с данными при  $x=b$ . Однако, эта задача Коши ничуть не лучше тех, от решения которых мы отказались! Чтобы избежать решения задачи Коши, можно действовать симметрично. Исходя из правого граничного условия найдем в каждой точке соотношение

$$\alpha_{\leftarrow}(x)y(x) + \beta_{\leftarrow}(x)y'(x) = \gamma_{\leftarrow}(x). \quad (12)$$

Затем из (11) и (12) найти  $y(x)$ . Чтобы эти соображения превратить в практический рецепт, нужно указать способ для нахождения  $\alpha$ ,  $\beta$ ,  $\gamma$ . В общем виде мы это сейчас делать не будем, а рассмотрим один важный частный случай, когда удобнее действовать несимметрично: найдя  $\alpha_{\rightarrow}(x)$ ,  $\beta_{\rightarrow}(x)$ ,  $\gamma_{\rightarrow}(x)$ , рассматривать соотношение

$$\alpha_{\rightarrow}(x)y(x) + \beta_{\rightarrow}(x)y'(x) = \gamma_{\rightarrow}(x)$$

как дифференциальное уравнение первого порядка для определения  $y(x)$ .

### 8.\* «Прогонка» для дифференциального уравнения

$$y'' + q(x)y = f(x)$$

Предположим, что в соотношении (11) коэффициент  $\beta_{\rightarrow}(x) \neq 0$ . Разделив на  $\beta_{\rightarrow}(x)$ , запишем (11) в виде

$$y'(x) = p(x)y(x) + r(x). \quad (13)$$

Мы должны подобрать  $p(x)$  и  $r(x)$  так, чтобы всякая функция, удовлетворяющая соотношению (13), принадлежала  $S_{\rightarrow}$ , т.е. удовлетворяла дифференциальному уравнению  $y'' + q(x)y = f(x)$  (1) и граничному условию  $\alpha_a y(a) + \beta_a y'(a) = \gamma_a$  (1a). Граничному условию мы удовлетворим, если потребуем, чтобы

$$p(a) = -\frac{\alpha_a}{\beta_a}, \quad r(a) = \frac{\gamma_a}{\beta_a}.$$

Попробуем удовлетворить уравнению. Продифференцируем

---

\* Разумеется, мы должны научиться находить  $\alpha$ ,  $\beta$ ,  $\gamma$ , минуя вычисление  $y_+(x)$  и  $z(x)$ : вычисление  $\alpha$ ,  $\beta$ ,  $\gamma$  через  $y_+(x)$  и  $z(x)$  из формулы (6) снова привело бы к «сокращению знаков».



соотношение (13):  $y'' = p'y + py' + r'$ . Используя уравнение (1), получим

$$-qu + f = p'y + py' + r'$$

или

$$y' = -\frac{p' + q}{p}y + \frac{f - r'}{p}. \quad (14)$$

Соотношение (14), выполняющееся для *всех* функций  $S_{\omega}$ , должно *совпадать* с соотношением (13). Поэтому получаем

$$p' = -p^2 - q \quad (15a)$$

$$r' = -rp + f. \quad (15b)$$

Формулы (15) позволяют нам предложить следующий рецепт для решения краевой задачи I.

*1 этап.* Решаем дифференциальные уравнения для  $p(x)$  и  $r(x)$ :

$$p' = -p^2 - q; \quad r' = -rp + f$$

с начальными условиями

$$p(a) = -\frac{\alpha_a}{\beta_a}; \quad r(a) = -\frac{\gamma_a}{\beta_a}$$

( $q(x)$  и  $f(x)$  — известные функции).

*2 этап.* Находим  $y(x)$  из соотношения (13)  $y'(x) = p(x)y(x) + r(x)$ , рассматривая его как дифференциальное уравнение первого порядка. При решении этого уравнения используем правое граничное условие  $\alpha_b y(b) + \beta_b y'(b) = \gamma_b$  и коэффициенты  $p(x)$  и  $r(x)$ , найденные на 1 этапе.

Уравнения (15) обладают многими интересными свойствами. Я отмечу здесь только одно, самое главное. Уравнение для  $p(x)$

$$\frac{dp}{dx} = -p^2(x) - q(x) \text{ — нелинейное!}$$

Таким образом, для решения линейной краевой задачи предлагается решать нелинейное уравнение. Это так называемое уравнение Риккати. В теории дифференциальных уравнений доказывается, что его решение может быть сведено к решению линейного уравнения 2 порядка. Очень поучительно, что здесь теория «работает в обратную сторону»: линейное уравнение сводится к нелинейному.

## 9. \* Аналогия с дискретной задачей

Дискретная (разностная) задача (10), если можно исполь-

зовать простейший метод исключения (прогонку), решается в 2 этапа\*.

1) Находим  $p_k$  и  $r_k$  по рекуррентным формулам

$$m_k = \frac{c_k}{b_k - a_k p_{k-1}}; \quad r_k = \frac{f_k - a_k r_{k-1}}{b_k - a_k p_{k-1}}. \quad (16)$$

Здесь  $a_k$ ,  $b_k$ ,  $c_k$  — коэффициенты трехдиагональной системы;  $k$  меняется «слева направо»: от 1 до  $n-1$ .

2) Находим  $y_k$  по формуле

$$y_k = -p_k y_{k+1} + r_k, \quad (17)$$

двигаясь «справа налево»:  $k = n-1, n-2, \dots, 0$ . На первом этапе используется только левое граничное условие. В начале второго — правое. Аналогия, как вы видите, полная. Более того, если отнормировать подходящим образом  $p_k$  и  $r_k$ , то в пределе, при  $h \rightarrow 0$  «дискретные» формулы (16) и (17) перейдут в «непрерывные» — (15) и (13)!

Таким образом, обнаруживается замечательная связь. С одной стороны была «чисто алгебраическая» задача о решении системы линейных алгебраических уравнений, и мы применили «чисто алгебраический» метод исключения (в его простейшей форме). С другой стороны была дифференциальная задача, и довольно сложные рассуждения о переносе граничных условий привели нас к системе двух дифференциальных уравнений первого порядка (15) (и еще одному уравнению (13) отдельно). По существу же, это близкие вещи, и (говоря нестрого) при  $h \rightarrow 0$  дискретная «прогонка» переходит в непрерывную. Пожалуй, стоит подчеркнуть, что сама идея переноса граничных условий (выделение семейства решений, удовлетворяющих одному граничному условию) носит общий характер. Эту идею (даже для разностных уравнений) совсем не обязательно связывать с методом исключения.

Последнее замечание. Рецепт п. 8 предлагает для решения краевой задачи решать 2 задачи Коши. Правда, не для исходного уравнения 2 порядка, но чем это лучше?

Пока не доказано, что эти задачи Коши являются вычислительно устойчивыми, рецепт п. 8, собственно, ничего не стоит. Оказывается, во многих случаях (их можно точно указать) задачи Коши (15) и (13) являются «хорошими». Например, это будет так при  $q(x) < 0$  (как раз «трудный случай»!) и граничных условиях  $y'(a) = y'(b) = 0$ . Но сейчас мы не будем проводить соответствующие доказательства.

---

\* См. лекцию 4.

## ДОПОЛНЕНИЕ

### Некоторые книги по вычислительной математике

1. Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы. М.: Наука, 1987; 3-е изд.: М.: Бином. Лаборатория знаний, 2004.
2. Воеводин В. В. Вычислительные основы линейной алгебры. М.: Наука, 1977.
3. Годунов С. К., Рябенький В. С. Разностные схемы. 2-е изд. М.: Наука, 1977.
4. Дьяченко В. Ф. Основные понятия вычислительной математики. 2-е изд. М.: Наука, 1977.
5. Калиткин Н. Н. Численные методы. М.: Наука, 1978.
6. Ортега Дж., Пул У. Введение в численные методы решения дифференциальных уравнений. М.: Наука, 1986.
7. Марчук Г. И. Методы вычислительной математики 3-е изд. М.: Наука, 1989.
8. Самарский А. А. Введение в численные методы 3-е изд. М.: Наука, 1997.
9. Самарский А. А., Гулин А. Н. Численные методы. М.: Наука, 1989.

## СОДЕРЖАНИЕ

Предисловие к третьему изданию . . . . .	1
Предисловие . . . . .	3
<b>Лекция 1. Общее введение . . . . .</b>	<b>4</b>
<b>Задача Коши для обыкновенных дифференциальных уравнений . . . . .</b>	<b>8</b>
1. Постановка задачи Коши . . . . .	8
2. Теоретическое отступление . . . . .	8
3. Стандартная запись . . . . .	9
4. Метод Эйлера . . . . .	10
5. Более точные методы: использование старших производных . . . . .	12
6. Более точные методы: многократное использование заданного уравнения . . . . .	12
7. Идея Адамса: использование уже найденного участка решения для приближенного вычисления старших производных . . . . .	14
8. Исторические и другие замечания . . . . .	15
9. Пример . . . . .	16
Дополнения и примечания к 1 лекции . . . . .	18
<b>Лекция 2. Системы линейных уравнений. Метод исключения . . . . .</b>	<b>21</b>
1. Метод исключения . . . . .	21
2. Трудности в методе исключения . . . . .	23
3. Трудности, связанные с приближенностью вычислений. Исключение с выбором главного элемента . . . . .	24
4. Всегда ли метод исключения с выбором главного элемента дает разумный результат? . . . . .	25
5. О трудоемкости метода исключения для общей системы $n$ линейных уравнений . . . . .	26
6. Пример: трехдиагональная система . . . . .	29
7. Метод исключения и определители . . . . .	31
8. Решение системы линейных уравнений и обращение матриц . . . . .	32
Примечания ко 2 лекции . . . . .	34
<b>Лекция 3. Нелинейные уравнения. Метод последовательных приближений . . . . .</b>	<b>36</b>
1. Локальная постановка задачи . . . . .	36
2. Линеаризация . . . . .	37
3. Метод Ньютона . . . . .	37
4. Поведение последовательности $x^{(s)}$ . . . . .	38
5. Алгебраические уравнения . . . . .	40
6. Огрубленный метод Ньютона . . . . .	41
7. Метод Ньютона как метод итераций . . . . .	41
8. Достаточное условие сходимости итераций . . . . .	41
9. Добавления и уточнения к теореме о сходимости . . . . .	42
10. Сходимость огрубленного метода Ньютона для некротного корня . . . . .	43
11. Сходимость метода Ньютона . . . . .	44
12. Система двух уравнений . . . . .	45
13. Система уравнений. Метод Ньютона . . . . .	46
14. О сходимости метода Ньютона для системы уравнений . . . . .	47
15. Итерации в системах . . . . .	47
16. Процесс итераций в линейном приближении . . . . .	49
Заключительные замечания. Последовательные приближения для решения линейных систем . . . . .	50

<b>Лекция 4. Приближенное дифференцирование и интегрирование</b>	<b>52</b>
1. Приближенное нахождение производной	52
2. Вычисление второй производной	55
3. Приближенное нахождение определенного интеграла	56
4. Формула Симпсона	58
5. Свойства формулы Симпсона	59
6. Старшие производные	59
7. Примеры дискретных задач, отвечающих задачам анализа	61
8. Линейные дифференциальные и линейные разностные уравнения	63
9. Три примера	65
<b>Лекция 5. Уравнения в частных производных. Проблема устойчивости разностных аппроксимаций</b>	<b>68</b>
1. Вводные замечания	68
2. Уравнения	69
3. Простейшие разностные аппроксимации	71
4. Теорема	73
5. Определение устойчивости	75
6. Пояснения и замечания к определению устойчивости	76
7. Обнаружение неустойчивости разностных аппроксимаций по методу Фурье	77
8. Пример доказательства устойчивости разностной схемы	79
<b>Лекция 6. Уравнения в частных производных. Неявные разностные аппроксимации. Вопрос о точности приближений</b>	<b>81</b>
1. Простейшая неявная схема	81
2. Решение уравнений на слое. Прогонка	82
3. Об устойчивости неявной схемы (4)	83
4. Еще один пример	85
5. О необходимом условии устойчивости	86
6. Типичная картина при возникновении неустойчивости	87
7. Линеаризация и «замораживание» коэффициентов	88
8. Понятие о точности разностных аппроксимаций дифференциальных уравнений	89
9. Определение порядка точности при приближенном интегрировании обыкновенных дифференциальных уравнений	91
10. Заключительные замечания	92
Дополнения и примечания к 6 лекции	93
<b>Лекция 7. Красивые задачи для обыкновенных дифференциальных уравнений</b>	<b>95</b>
1. Предварительные замечания	95
2. Постановка задачи	95
3. Способ решения краевой задачи, вытекающий из теории дифференциальных уравнений	96
4. К чему может привести «сокращение знаков»	97
5. Отступление. Что значит «хорошая» краевая задача?	98
6. Разностная задача, отвечающая краевой задаче I	99
7. Идея переноса граничных условий	100
8. «Прогонка» для дифференциального уравнения $y'' + q(x)y = f(x)$	101
9. Аналогия с дискретной задачей	102
<b>Дополнение. Некоторые книги по вычислительной математике</b>	<b>104</b>

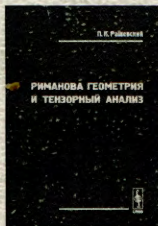
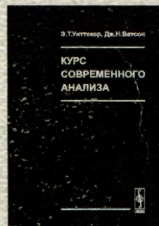


## Об авторе Эммануил Эльевич ШНОЛЬ



Несколько лет преподавал в средней школе и в Московском физико-техническом институте. Около четверти века Э. Э. Шноль был сотрудником Института прикладной математики АН СССР, который носит теперь имя М. В. Келдыша. С 1970 года автор работает в Пушинском Центре биологических исследований Российской академии наук.

Наше издательство предлагает следующие книги:



5399 ID 62480

НАУЧНАЯ И УЧЕБНАЯ ЛИТЕРАТУРА



9 785382 003788 &gt;

Тел./факс: 7 (495) 135-42-16  
Тел./факс: 7 (495) 135-42-46



URSS

E-mail:  
URSS@URSS.ru  
Каталог изданий  
в Интернете:  
<http://URSS.ru>