В. И. КОСАРЕВ

12 ЛЕКЦИЙ

по вычислительной математике

В. И. КОСАРЕВ

12 ЛЕКЦИЙ ПО ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКЕ

ВВОДНЫЙ КУРС

Издание 3-е, исправленное и дополненное

Рекомендовано Учебно-методическим объединением высших учебных заведений Российской Федерации по образованию в области прикладных математики и физики в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению подготовки «Прикладные математика и физика»

Москва
ФИЗМАТКНИГА
2013

ББК 22.31
К71
УДК 519.63 (074.8)

Рецензенты:
кафедра вычислительных методов факультета вычислительной математики и кибернетики МГУ (зав. кафедрой академик *А. А. Самарский*),
д. ф.-м. н., проф. *В. Ф. Дьяченко*

КОСАРЕВ В. И. 12 лекций по вычислительной математике (вводный курс). — Изд. 3-е, испр. и доп. — М.: Физматкнига, 2013. — 240 с.
ISBN 978-5-89155-214-2.

Учебное пособие написано на основе лекций, которые на протяжении многих лет автор читал студентам Московского физико-технического института (государственного университета).

Пособие содержит необходимые начальные представления о средствах, терминологии и возможностях вычислительной математики. В книге освещены следующие темы: методы вычисления решений нелинейных уравнений и систем уравнений; прямые и итерационные методы решения систем линейных уравнений; интерполяция и среднеквадратичное приближение для функций, задаваемых таблицей своих значений; численное дифференцирование и численное интегрирование; численное решение обыкновенных дифференциальных уравнений (задача Коши, краевые задачи); элементы теории разностных схем (аппроксимация, устойчивость, сходимость); разностные схемы для модельных уравнений математической физики (уравнения переноса, теплопроводности, Пуассона).

Книга адресована студентам различных технических специальностей, для которых вычислительные методы не являются профилирующим предметом.

Ил. 60.

ISBN 978-5-89155-214-2

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ К ТРЕТЬЕМУ ИЗДАНИЮ

Я снова обращаюсь со словами признательности в адрес издательства, которое сочло возможным еще раз выпустить мой учебник.

В настоящем издании внесены назревшие редакционные правки и исправлены замеченные недочеты.

Естественно, вычислительная математика, как любая ветвь математики, развивается, возникают новые идеи, решения. В нашей области — новые алгоритмы, их обоснования, что отражается в потоке публикаций, монографий, учебников.

Я, ассоциируя себя как автора вводного учебника по теме «Методы вычислений», не ввожу в обсуждение новых тем, алгоритмов и т. д., обозначая соответствующие ссылки в библиографии.

ПРЕДИСЛОВИЕ К ПЕРВОМУ ИЗДАНИЮ

Предлагаемый учебник имеет вполне определенного адресата. Он ориентирован на студентов научно-естественных и технических специальностей, не связанных с профессиональным использованием численных методов, и на соответствующих специалистов, эпизодически сталкивающихся с необходимостью самостоятельно вычислять на ЭВМ решение не слишком сложных задач. (Сказанное, конечно, не означает, что эта книга противопоказана студентам, специализирующимся по машинным вычислениям. Автор надеется, что будущим вычислителям-профессионалам она может оказаться полезной в качестве книги для первого чтения.)

Надо отметить, что при всем обилии учебников и монографий (в том числе хороших и очень хороших) по методам вычислений книг, представляющих собой своего рода введение в специальность, ориентированных на читателя, совершенно неподготовленного (в плане численного решения тех или иных задач), не так уж много. Среди них я отметил бы в первую очередь две (конкретизируя тем самым «жанр» предлагаемых «Лекций»): это курс А. Н. Тихонова и Д. П. Костомарова «Вводные лекции по прикладной математике» (1984 г.) и учебник В. Ф. Дьяченко «Основы вычислительной математики» (1972 г.).

Обозначив «жанровую» общность данной книги с упомянутыми выше, отмечу и различия. От «Вводных лекций» А. Н. Тихонова и Д. П. Костомарова предлагаемые «Лекции» отличаются отбором и объемом излагаемого материала. Что касается «Основ...» В. Ф. Дьяченко, то, как мне представляется, они предполагают у читателя в большей степени математический склад мышления, нежели тот, на который ориентирована эта книга.

Изложение численных методов не может не опираться на материал из математического анализа и линейной алгебры. Требуемые сведения в данном случае не выходят за рамки обязательных курсов для студентов, которые были названы в качестве адресата «Лекций». Из математического анализа это формула (разложение в ряд) Тейлора для функции одной и многих переменных, теорема Лагранжа о среднем, теорема Ролля. Из линейной алгебры — представление о нормированных пространствах, о собственных значениях и векторах матриц. Впрочем, справочные сведения по линейной алгебре, в порядке напоминания, приводятся в Лекции 2.

Лекции 9–12 опираются на элементарные сведения из теории уравнений в частных производных и предполагают знакомство с простейшими постановками задач для классических модельных уравнений.

Коротко о построении излагаемого материала.

Каждая лекция посвящена отдельному разделу (теме, группе родственных методов для решения определенного класса задач), что видно из содержания. Каждая лекция (кроме предпоследней) сопровождается дополнением, в определенной степени расширяющим и углубляющим основной материал. В практике автора содержание дополнений, как правило, использовалось для обсуждения на семинарских занятиях — так же, как вопросы и упражнения, которыми завершается обсуждение каждого раздела.

Подробная библиография по теме каждой лекции имеет целью сориентировать читателя, которому потребуется более глубоко разобраться с теми или иными вопросами.

В дополнительном списке литературы приведены учебники, монографии и сборники, посвященные развитию и применению численных методов для решения конкретных задач из различных разделов физики и математики.

В заключение этого краткого предисловия считаю своим долгом выразить искреннюю признательность своим коллегам по кафедре вычислительной математики МФТИ, в творческом взаимодействии с которыми сформировалась эта книга.

Я особо благодарен профессорам В. С. Рябенькому, А. С. Холодову, доценту А. И. Лобанову, которые познакомились с предлагаемым учебником на стадии рукописи и сделали ряд существенных замечаний.

Хотелось бы здесь выразить самую глубокую признательность профессору Л. А. Чудову, который не только инициировал работу над этой книгой, но постоянным, доброжелательным вниманием стимулировал доведение этой работы до конца.

Я благодарен также Н. В. Пулькиной за помощь в технической обработке рукописи.

ПРЕДВАРИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

Краткое обсуждение вопросов, связанных со спецификой машинных вычислений, влиянием на вычисляемые результаты ошибок во входных данных и погрешностей округления при выполнении арифметических операций. Примеры.

Приступая к изложению азов вычислительной математики, хотелось бы отметить, что численные методы, которые мы будем здесь обсуждать, являются в большинстве случаев приближенными, т. е. позволяют получить лишь некоторое (требуемое, разумное) приближение к искомому решению рассматриваемой задачи. Реализация численных методов сводится к выполнению многих (тысяч или миллионов) простейших арифметических операций типа: «сложить», «умножить», «разделить». При этом следует помнить, раз и навсегда уяснить («зарубить себе на носу»), что получить точные результаты посредством выполнения этих операций на ЭВМ *принципиально нельзя* даже в том случае, если численный метод является точным.

Результаты вычислений всегда содержат *неустранимые погрешности*. Источником их являются, во-первых, погрешности исходных данных, во-вторых, погрешности округлений, неизбежные при выполнении арифметических операций на ЭВМ. (Ошибки округления отсутствуют, если вычисления ведутся по правилам арифметики целых чисел, однако при решении задач физического содержания целыми числами в подавляющей массе случаев никак нельзя обойтись.)

В зависимости от способа вычислений неустранимые погрешности искажают результат в большей или в меньшей степени. С этим обстоятельством необходимо считаться, так как при неправильной (неграмотной, нерациональной) организации вычислений можно получить абсурдные результаты.

В подтверждение сказанного приведем простые примеры.

Пример 1. Пусть надо решить систему двух линейных уравнений

$$-10^{-7}x_1 + x_2 = 1,$$
$$x_1 + 2x_2 = 4.$$

Первый метод. Исключая x_1 из первого уравнения: $x_1 = 10^7 x_2 - 10^7$, и подставляя это выражение во второе уравнение, получаем $x_2 = \dfrac{10^7 + 4}{10^7 + 2}$.

Проведя вычисления с семью значащими цифрами (что типично для большинства современных компьютеров при работе в режиме с

ординарной точностью), получаем $x_2 = 1.000000$, $x_1 = 0.000000$, что совершенно неверно, как видно из второго уравнения.

Второй метод. Исключая x_1 из второго уравнения: $x_1 = 4 - 2x_2$, получаем для x_2 формулу $x_2 = \dfrac{1 + 4 \cdot 10^{-7}}{1 + 2 \cdot 10^{-7}}$. После вычислений получаем $x_2 = 1.000000$, $x_1 = 2.000000$ — правильное (с точностью до шести десятичных цифр) решение.

Погрешности входных данных также могут сильно повлиять на результаты.

Пример 2*). Пусть требуется найти объем шара, касающегося цилиндра радиуса R и двух касательных к нему взаимно перпендикулярных плоскостей (см. рис. 1).

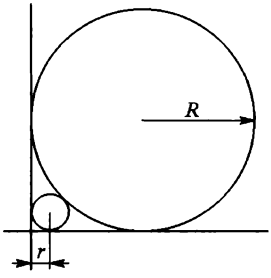


Рис. 1.

Легко усмотреть, что радиус шара r равен

$$r = R\frac{\sqrt{2} - 1}{\sqrt{2} + 1},$$

и, следовательно, объем

$$V = \frac{4}{3}\pi R^3 \left(\frac{\sqrt{2} - 1}{\sqrt{2} + 1}\right)^3.$$

Займемся вычислениями последнего множителя $S = \left(\dfrac{\sqrt{2} - 1}{\sqrt{2} + 1}\right)^3$. Очевидно, избавившись от знаменателя, можно записать

$$S = (\sqrt{2} - 1)^6 = (3 - 2\sqrt{2})^3 = 99 - 70\sqrt{2}.$$

Рассматривая последние три формулы как три метода вычисления величины S, приведем результаты расчетов в таблице, привлекая для иррационального числа $\sqrt{2}$ следующие приближения $\sqrt{2} \approx 7/5 = 1.4$ (относительная погрешность примерно 1%) и $\sqrt{2} \approx 17/12 = 1.41(6)\ldots$ (относительная погрешность примерно 0.18%):

$\sqrt{2}$	$(\sqrt{2} - 1)^6$	$(3 - 2\sqrt{2})^3$	$99 - 70\sqrt{2}$
7/5	0.004096	0.008000	1
17/12	0.005233	0.004630	$-0.1(6)\ldots$

Вычисления проводились с большим количеством значащих цифр, так что различие результатов, к которым приводят эти три «метода», обу-

*) Этот пример заимствован из [3, т. 1, с. 39]

словлено различным влиянием погрешности в приближенном представлении исходной константы ($\sqrt{2}$) при разных способах расчета.

Вообще говоря, нетрудно понять, каким образом в рассмотренном случае формируются абсурдные результаты. Представим себе два близких по величине числа:

$$x = 0.\mathrm{abcdefg}, \quad y = 0.\mathrm{abcde_1f_1g_1}.$$

Здесь a, b, ... — цифры после десятичной точки в записи чисел. Таким образом, в рассмотренном случае числа x, y отличаются друг от друга пятой и далее (после десятичной точки) цифрами. Разность этих чисел будет величиной пятого порядка малости: $\sim 10^{-5}(\mathrm{efg} - \mathrm{e_1f_1g_1})$. Допустим, что цифры f, $\mathrm{f_1}$ и далее искажены за счет каких-то ошибок (исходных или погрешностей округления, если величины x, y получены в процессе предшествующих вычислений), тогда в вычисленной разности $(x - y)$ мы получим только одну правильную цифру $(\mathrm{e} - \mathrm{e_1})$, то есть ошибка результата на данной элементарной арифметической операции резко возрастает (если и цифры e, $\mathrm{e_1}$ неверны, то результат $(x - y)$ просто не имеет ничего общего с правильным. Именно таков «механизм» формирования результатов в наших примерах. В первом случае при вычислении x_1 по формуле $10^7 x_2 - 10^7$ все учитываемые цифры совпадали, в итоге — результат абсолютно неверен. Во втором примере в последнем столбце таблицы точный результат определяется пятой и последующими значащими цифрами чисел 99 и $70\sqrt{2}$ (третьей и последующими цифрами после десятичной точки), в то время как вторая значащая цифра при $\sqrt{2} \approx 7/5$ (третья при $\sqrt{2} \approx 17/12$) искажена за счет погрешности $\sqrt{2}$, — результаты также абсурдны.

Как можно понять из приведенных примеров, операция вычитания близких (и тем более больших) по величине чисел чревата неприятностями при машинных вычислениях. И, следовательно, помня об этом, при планировании вычислений (в частности, при выборе алгоритма) следует по возможности избегать ситуаций, когда возникает необходимость в выполнении этих операций. В первом примере в этой связи следует выбрать второй способ решения. Во втором примере предпочтительным является первый «метод» вычисления результата — $(\sqrt{2} - 1)^6$.

Почему так — можно понять, привлекая аппарат математического анализа к исследованию влияния ошибок входных данных на результат вычислений.

Пусть вычисляется величина $\varphi(x, y)$. При неточных входных данных: $\widetilde{x} = x \pm \Delta x$, $\widetilde{y} = y \pm \Delta y$, ошибка результата оценивается сверху следующим образом:

$$|\Delta\varphi| \leqslant |\varphi(\widetilde{x}, \widetilde{y}) - \varphi(x, y)| \leqslant \left|\frac{\partial\varphi}{\partial x}\right||\Delta x| + \left|\frac{\partial\varphi}{\partial y}\right||\Delta y|.$$

Объективной мерой погрешности является относительная ошибка — отношение погрешности величины к самой величине (по модулю). В данном случае

$$\delta = \left|\frac{\Delta\varphi}{\varphi}\right| \leqslant \frac{1}{|\varphi|}\left(\left|\frac{\partial\varphi}{\partial x}\right||\Delta x| + \left|\frac{\partial\varphi}{\partial y}\right||\Delta y|\right).$$

Например, для $\varphi = (\sqrt{2}-1)^6$, $\widetilde{x} = 7/5$ $(\Delta x \ll 0.0143)$, $y = \widetilde{y} = 1$ $(\Delta y = 0)$, имеем

$$\delta \leqslant \frac{1}{|\varphi|}\left|\frac{\partial\varphi}{\partial x}\right|\Delta x = \frac{1}{(\sqrt{2}-1)^6}\cdot 6(\sqrt{2}-1)^5\Delta x = \frac{6}{\sqrt{2}-1}\Delta x \approx 0.207,$$

т. е. погрешность результата (0.004096) в этом случае составляет $\approx 21\%$. Точно так же можно проанализировать погрешность остальных результатов и прийти тем самым к уже упомянутому выводу о том, что первый «метод» вычислений — предпочтительней.

З а м е ч а н и е. При получении оценки для относительной погрешности надо, согласно определению относительной погрешности, привлекать точное значение результата $\varphi(x, y)$, но оно-то как раз и неизвестно и является целью вычислений. Тем не менее, оценку неустранимой погрешности можно получить, исходя из следующего приближенного определения относительной погрешности:

$$\delta \approx \left|\frac{\Delta\varphi}{\widetilde{\varphi}}\right| \leqslant \frac{1}{|\varphi(\widetilde{x},\widetilde{y})|}\left(\left|\frac{\partial\varphi}{\partial x}\right||\Delta x| + \left|\frac{\partial\varphi}{\partial y}\right||\Delta y|\right),$$

где $\widetilde{\varphi} = \varphi(\widetilde{x}, \widetilde{y})$ — вычисленный результат. ▲*)

Отмеченные обстоятельства предопределяют, что при использовании численных методов наряду с конечными результатами столь же важное значение будут иметь оценка их достоверности и оценка возможного влияния на результаты разного рода погрешностей.

Более подробную информацию о влиянии неустранимых погрешностей на результаты выполнения арифметических операций на ЭВМ можно найти в [1, с. 23–72], [12, с. 16–24], [16, с. 2–42].

*) Здесь и далее значком ▲ будем обозначать конец замечания.

ЛЕКЦИЯ 1

ЧИСЛЕННОЕ РЕШЕНИЕ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

Формулировка рассматриваемой задачи. Локализация корней. Метод половинного деления. Метод Ньютона, скорость сходимости. Метод простых итераций, достаточное условие сходимости, скорость сходимости. Метод релаксаций. Геометрическая интерпретация. Метод секущих.

Пусть задана функция $f(x)$. Нужно найти корни уравнения

$$f(x) = 0. \tag{1.1}$$

Ограничимся обсуждением проблемы поиска действительных корней.

Задача разбивается на два этапа:

локализация корней, т. е. предварительный анализ расположения корней на оси x, в результате которого выявляются такие отрезки оси x, каждому из которых принадлежит не более одного корня;

вычисление с требуемой точностью корня (корней), принадлежащих заданному отрезку (заданным отрезкам).

Локализация корней выходит за рамки собственно вычислительной математики. Это скорее предмет математического анализа. (Впрочем, на этом этапе можно воспользоваться ЭВМ для того, чтобы вычислить таблицу значений функции $f(x)$ или построить график этой функции.) Поэтому, полагая, что корни локализованы в оговоренном выше смысле, будем заниматься следующей конкретной задачей.

Вычислить с заданной точностью ε корень уравнения (1.1), принадлежащий отрезку $[a, b]$. Предполагается, что на отрезке $[a, b]$ находится единственный корень. Кроме того, мы будем полагать, что искомый корень является простым (некратным). Некоторые замечания по поводу вычисления кратных корней будут сделаны ниже, по ходу изложения материала данной лекции.

З а м е ч а н и е. При решении задач физического содержания зачастую отрезок $[a, b]$, содержащий требуемый корень, известен из физических соображений. ▲

Что касается малого параметра ε, характеризующего требуемую точность, или, иными словами, допустимый уровень погрешности, то смысл его в том, что вычисленное значение корня $\bar{x}$ должно отличаться от точного x_* не более, чем на ε:

$$|\bar{x} - x_*| \leqslant \varepsilon.$$

Метод половинного деления. Этот метод состоит в следующем. Делим отрезок $[a, b]$ пополам: $c = (a + b)/2$. Проверяя знаки $f(a)$, $f(c)$, $f(b)$, выясняем, какой половине исходного отрезка принадлежит корень. Последнюю снова делим пополам и т. д., до тех пор, пока не выполнится условие $(b - a)/2^n \leqslant \varepsilon$, где n — число проведенных делений исходного отрезка пополам. После этого любую точку последнего отрезка можно взять в качестве ответа $\overline{x}$.

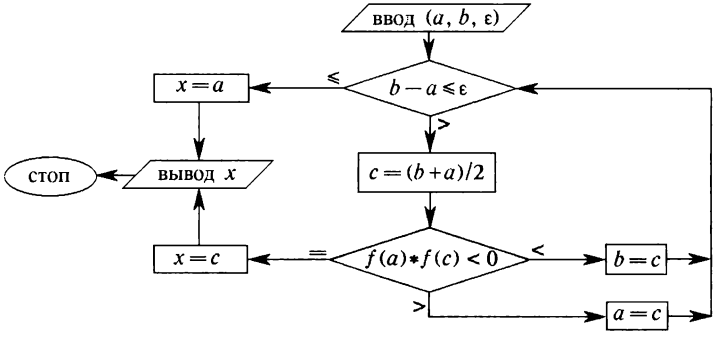


Рис. 1.1.

Приведем формализованное (в виде фрагмента блок-схемы, рис. 1.1) описание сформулированного алгоритма решения рассматриваемой задачи.

Замечание 1. Данный метод можно рассматривать как последовательное уточнение локализации искомого корня: на каждом шаге размер окрестности, которой он принадлежит, уменьшается вдвое. ▲

Замечание 2. Очевидно, что метод половинного деления работает, если при переходе через корень $f(x)$ меняет знак, т. е. он может быть использован для вычисления корней нечетной кратности: однократный (простой) корень, трехкратный и т. д. ▲

Метод Ньютона. Перейдем к обсуждению еще одного простого (многие из вас уже сталкивались с ним, может быть, даже в школе), но более эффективного, как мы увидим, метода вычисления корня.

Зададимся некоторым начальным (нулевым) приближением вычисляемого корня $x_0 \in [a, b]$. Линеаризуем функцию $f(x)$ в окрестности точки x_0, представив ее в виде отрезка ряда Тейлора:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0),$$

и вместо нелинейного уравнения (1.1) решим линеаризованное уравнение

$$f(x_0) + f'(x_0)(x - x_0) = 0,$$

трактуя его решение как следующее (первое) приближение к искомому значению корня. Получим

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Продолжая этот процесс, приходим к формуле Ньютона

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{1.2}$$

для вычисления последовательности приближений к искомому решению.

Геометрическая интерпретация этого процесса, показанная на рис. 1.2, делает понятным другое название этого метода — метод касательных. Из формулы (1.2) вытекает условие применимости метода: функция $f(x)$ должна быть дифференцируемой и $f'(x)$ в окрестности корня не должна менять знак.

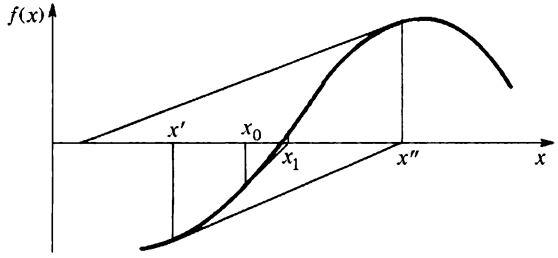


Рис. 1.2.

Замечание. Если первая или ряд последующих производных обращаются в ноль в точке корня, а в окрестности корня $f'(x) \neq 0$, то формула (1.2) сохраняет смысл, и, таким образом, метод Ньютона можно использовать и для вычисления кратных корней, но при этом «качество» метода, которое мы выявим применительно к однократному корню, теряется (см. упражнение 4). В этой связи я бы порекомендовал (в порядке упражнения) попробовать с хорошей точностью вычислить на машине корень уравнения $(x-1)^6 = 1$. ▲

Каковы же условия сходимости последовательности значений $\{x_n\}$, вычисляемых по формуле (1.2) к корню уравнения (1.1)?

Предполагая, что $f(x)$ дважды непрерывно дифференцируема, напишем формулу Тейлора для $f(x_*)$ (x_* — искомый корень) в окрестности n-го приближения

$$f(x_*) = 0 = f(x_n) + f'(x_n)(x_* - x_n) + \frac{1}{2} f''(\xi)(x_* - x_n)^2,$$

где $\xi \in [x_*, x_n]$.

Разделив последнее соотношение на $f'(x_n)$ и перенеся первые два слагаемых из правой части в левую, получим

$$\left[x_n - \frac{f(x)}{f'(x_n)}\right] - x_* = \frac{1}{2}\frac{f''(\xi)}{f'(x_n)}(x_n - x_*)^2,$$

что, учитывая (1.2), переписываем в виде

$$x_{n+1} - x_* = \frac{1}{2}\frac{f''(\xi)}{f'(x_n)}(x_n - x_*)^2.$$

Отсюда

$$|x_{n+1} - x_*| = \frac{1}{2}\frac{|f''(\xi)|}{|f'(x_n)|}|x_n - x_*|^2. \tag{1.3}$$

Из (1.3) следует оценка

$$|x_{n+1} - x_*| \leqslant \frac{1}{2}\frac{M_2}{m_1}|x_n - x_*|^2, \tag{1.4}$$

где $M_2 = \max\limits_{[a,b]}|f''(x)|$, $m_1 = \min\limits_{[a,b]}|f'(x)|$.

Очевидно, ошибка на каждом шаге убывает, если

$$\frac{1}{2}\frac{M_2}{m_1}|x_0 - x_*| < 1. \tag{1.5}$$

Полученное условие означает, что сходимость зависит от выбора начального приближения; это можно усмотреть и из геометрической интерпретации метода (см. рис. 1.2): если в качестве начального приближения взять значение x', то на сходимость последующих приближений (x'' и т. д.) рассчитывать не приходится. С другой стороны, как видно, всегда можно добиться выполнения условия (1.5) за счет более точного выбора начального приближения x_0, т. е. за счет более аккуратной локализации искомого корня, которой можно достичь, например, с помощью предыдущего метода (см. замечание к нему).

Оценка (1.4) характеризует скорость убывания погрешности для метода Ньютона: на каждом шаге погрешность пропорциональна квадрату предыдущей. Это очень высокий темп. Например, если в некотором приближении получена одна точная цифра после запятой, то в следующем можно ожидать две точные цифры, далее — четыре и т. д. Для сравнения в методе половинного деления ошибка на каждом шаге уменьшается в два раза, т. е.

$$|x_{n+1} - x_*| \leqslant \frac{1}{2}|x_n - x_*| \leqslant \ldots \leqslant \frac{1}{2^{n+1}}(b - a).$$

Здесь, чтобы получить четыре точные цифры результата, надо выполнить столько делений исходного отрезка пополам, чтобы

$$\frac{1}{2^{n+1}}(b-a) \leqslant \frac{1}{2}\cdot 10^{-4}.$$

Отсюда $n \geqslant [4+\lg(b-a)]/\lg 2$. Например, при $b-a=1$ получим $n \geqslant 14$. Правда, надо учесть, что в методе деления отрезка пополам для перехода к очередному приближению достаточно вычислить одно значение функции $f(x)$, в то время как метод Ньютона требует еще вычисления производной от нее.

Метод простых итераций. Теперь рассмотрим более общий итерационный процесс, т. е. метод последовательных приближений (частным случаем которого является и метод Ньютона). Представим уравнение (1.1) в виде

$$x = \varphi(x). \tag{1.6}$$

Очевидно, искомый корень при подстановке его в (1.6) превращает последнее в тождество

$$x_* \equiv \varphi(x_*).$$

Рассмотрим последовательность значений x, которая определяется следующим образом:

$$x_{n+1} = \varphi(x_n), \quad x_0 \text{ задано } (x_0 \in [a,b]). \tag{1.7}$$

Оказывается, при определенных свойствах функции $\varphi(x)$ последовательность (1.7) сходится к корню уравнения (1.1). Достаточные условия сходимости формулируются в теореме.

Теорема. *Пусть в окрестности искомого корня* $U_* = \{|x - x_*| \leqslant r\}$ *функция* $\varphi(x)$ *удовлетворяет условию Коши–Липшица с константой, меньшей единицы, т. е. для любых* $x', x'' \in U_*$:

$$|\varphi(x') - \varphi(x'')| \leqslant q|x' - x''|, \quad q = \text{const} < 1. \tag{1.8}$$

Тогда последовательность (1.7) с $x_0 \in U_*$ *сходится к корню, т. е.* $x_n \to x_*$ *при* $n \to \infty$, *и имеет место оценка погрешности*

$$|x_n - x_*| \leqslant q^n |x_0 - x_*|. \tag{1.9}$$

Покажем сначала, что все x, определяемые в (1.7), принадлежат окрестности U_*. В самом деле, если $x \in U_*$ и $y = \varphi(x)$, то

$$|y - x_*| = |\varphi(x) - \varphi(x_*)| \leqslant q|x - x_*| \leqslant |x - x_*| \leqslant r.$$

Далее, $|x_n - x_*| = |\varphi(x_{n-1}) - \varphi(x_*)| \leqslant q|x_{n-1} - x_*|$, откуда $|x_n - x_*| \leqslant q^n|x_0 - x_*|$, т. е. доказано утверждение (1.9). Из него непосредственно следует сходимость $x_n \underset{n\to\infty}{\longrightarrow} x_*$.

Замечание. На практике вместо условия (1.8) обычно требуют выполнения условия

$$|\varphi'(x)|_{U_*} \leqslant q < 1, \tag{1.8'}$$

из которого следует (1.8) в силу теоремы о среднем из математического анализа. ▲

В силу очевидной неоднозначности перехода от (1.1) к (1.6) практически всегда можно подобрать $\varphi(x)$ таким образом, чтобы в некоторой окрестности корня выполнялось условие (1.8). Может быть, при этом окажется необходимым вернуться к этапу локализации, чтобы сузить рассматриваемую окрестность.

Что касается скорости сходимости метода простых итераций, то из (1.9) следует, что на каждом шаге погрешность убывает в q раз. (Напомним, что для метода половинного деления $q = 1/2$.) Таким образом, погрешности ведут себя как члены убывающей геометрической прогрессии со знаменателем q. В этой связи говорят, что метод простых итераций сходится со скоростью геометрической прогрессии. Или, замечая, что последовательные погрешности связаны друг с другом линейным неравенством

$$|x_{n+1} - x_*| \leqslant q|x_n - x_*|,$$

говорят, что метод простых итераций представляет собой *линейный итерационный процесс* (или *метод первого порядка*).

Замечание. Последнее определение имеет смысл, если в любой сколь угодно малой окрестности корня константа q ограничена снизу, т. е. $q > q_0 > 0$. Например, очевидно, что при

$$\varphi(x) = x - \frac{f(x)}{f'(x)}$$

мы получаем метод Ньютона (как частный случай метода простых итераций). Легко устанавливается, что для него $|\varphi'(x)|\Big|_{x\to x_*} \longrightarrow 0$ т. е. q не ограничено снизу. Как следствие мы получили соотношение (1.4) между последовательными погрешностями метода Ньютона, которое позволяет трактовать последний как *квадратичный итерационный процесс* (или *метод второго порядка*). ▲

Используя неравенство (1.9), можно получить априорную оценку количества итераций (приближений), которые нужно выполнить, что-

бы получить результат с требуемой точностью. Для этого, очевидно, достаточно решить относительно n неравенство

$$q^n|x_0 - x_*| \leqslant \varepsilon.$$

Мажорируя неизвестную величину $|x_0 - x_*|$, например, размером рассматриваемой окрестности $(b-a)$, получаем:

$$n \geqslant \frac{\lg[\varepsilon/(b-a)]}{\lg q}. \tag{1.10}$$

При решении практических задач метод типа (1.7) (отличный от метода Ньютона) может оказаться предпочтительней метода Ньютона, потому что последний требует вычислений $f'(x)$ на каждой итерации. Подходящую функцию $\varphi(x)$, обеспечивающую переход от (1.1) к (1.6), можно искать, например, в виде $\varphi(x) = x - g(x)f(x)$, подбирая функцию $g(x)$ так, чтобы на $[a, b]$ выполнилось условие (1.8′).

Остановимся несколько подробней на оценке погрешности для метода простых итераций.

Прежде всего отметим, что наряду с (1.9) можно получить оценку, не зависящую от точного значения корня:

$$|x_n - x_*| \leqslant \frac{q^n}{1-q}|x_1 - x_0|$$

(см. упражнения).

На практике, однако, найти $q = \max\limits_{[a,b]} |\varphi'(x)|$ удается далеко не всегда. Тем более, для метода Ньютона может оказаться затруднительным получить оценки M_2 и m_1, которые можно было бы (в соответствии с (1.4)) использовать для контроля за текущей погрешностью в процессе вычислений. Поэтому при проведении расчетов довольно часто контроль достигнутой точности осуществляется на основе следующего приближенного условия: проверяется неравенство $|x_{n+1} - x_n| \leqslant \varepsilon$, как только оно выполнится, считается, что последнее вычисленное значение (x_{n+1}) является требуемым результатом.

З а м е ч а н и е. При сходимости последовательных приближений к корню с разных сторон, что имеет место, если $\varphi' < 0$ в рассматриваемой окрестности (см. рис. 1.3*б*), величина $|x_{n+1} - x_n|$ мажорирует истинную погрешность, т. е. $|x_{n+1} - x_n| > |x_{n+1} - x_*|$, и поэтому данный критерий окончания вычислений является вполне объективным. Если же $\varphi' > 0$, то сходимость к корню носит односторонний характер (как на рис. 1.3*а*), и условие $|x_{n+1} - x_n| \leqslant \varepsilon$ может выполнить-

ся гораздо раньше нужного требования $|x_{n+1} - x_*| \leqslant \varepsilon$. В этом случае контроль достигнутой точности лучше осуществлять по проверке неравенства $\dfrac{1}{1-q}|x_{n+1} - x_n| \leqslant \varepsilon$, которое вытекает из легко устанавливаемой оценки $|x_n - x_*| \leqslant \dfrac{1}{1-q}|x_{n+1} - x_n| \leqslant \dfrac{q}{1-q}|x_n - x_{n-1}|$ (см. упражнения). Правда, при этом все-таки надо предварительно получить оценку q (см. [1, с. 99]). ▲

Влияние неустранимых погрешностей на вычисляемые приближения будет рассмотрено в более общем случае (для систем уравнений) в Лекции 4.

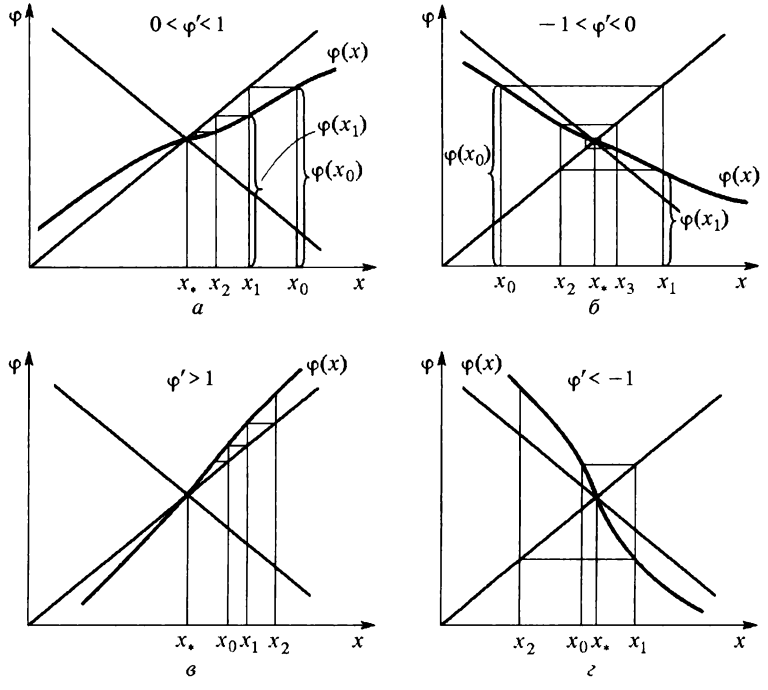


Рис. 1.3.

Метод релаксаций. Как было отмечено выше, для перехода от (1.1) к (1.6) функцию $\varphi(x)$ можно искать в следующем виде: $\varphi(x) = x - g(x)f(x)$. Часто полагают $g(x) = \tau = \text{const}$, записывая (1.1) в виде

$$x = x - \tau f(x) = \varphi(x) \tag{1.11}$$

и подбирая параметр τ так, чтобы в рассматриваемой окрестности выполнялось условие сходимости

$$|\varphi'(x)| = |1 - \tau f'(x)| < 1. \tag{1.12}$$

Таким образом, построенный метод называется методом релаксаций; τ — релаксационный параметр. (См. упражнение 2 к данной лекции.)

Геометрическая интерпретация рассмотренных методов. Метод простых итераций допускает наглядную геометрическую трактовку. В самом деле, решение уравнения $x = \varphi(x)$ — это точка пересечения прямой $y = x$ с кривой $y = \varphi(x)$ в плоскости (x, y).

На рис. 1.3*а–г* иллюстрируются различные ситуации: сходимость к корню односторонняя, с разных сторон, расходящиеся итерационные процессы.

Метод релаксаций (1.11) по сути означает, что функцию $f(x)$ в исходном уравнении (1.1) мы заменяем в окрестности всех x_n прямой $y = f(x_n) + \dfrac{1}{\tau}(x - x_n)$ с постоянным наклоном, и в качестве очередного приближения рассматриваем точку пересечения ее с осью абсцисс (рис. 1.4).

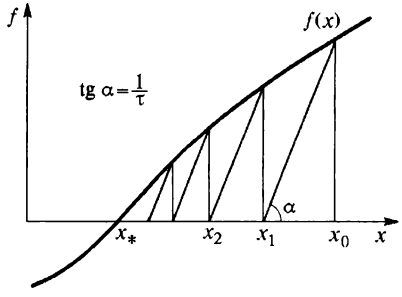


Рис. 1.4.

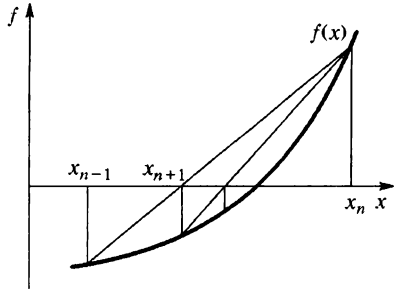


Рис. 1.5.

Иногда вместо метода Ньютона, чтобы избавиться от необходимости вычислять $f'(x)$ на каждом шаге итераций, используют так называемый *модифицированный* (или *огрубленный*) *метод Ньютона*:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}.$$

Очевидно, это есть частный случай метода релаксаций с релаксационным параметром

$$\tau = \frac{1}{f'(x_0)}.$$

Метод секущих. Метод Ньютона требует вычисления производной $f'(x)$ при вычислении каждого приближения. Это может заметно понизить его реальную эффективность (в смысле затрат машинного времени).

Если на $(n+1)$-м шаге (резонно полагая, что величина $|x_n - x_{n-1}|$ достаточно мала) использовать в формуле Ньютона приближение

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}},$$

то мы придем к формуле

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1}) f(x_n)}{f(x_n) - f(x_{n-1})},$$

которая определяет метод секущих. Название связано с геометрической интерпретацией метода (рис. 1.5). Нетрудно убедиться, что хорда к кривой $f(x)$, проведенная через точки $(x_{n-1}, f(x_{n-1}))$ и $(x_n, f(x_n))$, пересекает ось абсцисс в точке x_{n+1}. Можно доказать ([9], с. 146), что метод секущих сходится медленнее, чем метод Ньютона, но быстрее, чем какой-либо линейный итерационный процесс (в частности, модифицированный метод Ньютона).

ДОПОЛНЕНИЕ К ЛЕКЦИИ 1

Об условиях сходимости метода простых итераций. Можно доказать справедливость следующих утверждений.

Утверждение 1. *Пусть в окрестности* $U_a = |x - a| \leqslant r$ *функция* $\varphi(x)$ *удовлетворяет условиям*:

1) *Коши–Липшица*: *для любых* $x', x'' \in U_a$ $|\varphi(x') - \varphi(x'')| \leqslant q(x' - x'')$, *причем* $q = \text{const} < 1$;

2) $|\varphi(a) - a| \leqslant (1-q)r$.

Тогда

1) *в* U_a *существует одно и только одно решение уравнения* (1.6);

2) *если* $x_0 \in U_a$, *то* x_n, *вычисляемые согласно* (1.7), *сходятся к корню уравнения* (1.6);

3) *справедлива оценка* (1.9).

Очевидно, что это более сильное утверждение, нежели теорема о достаточном условии сходимости итераций, сформулированная в Лекции. Окрестность, определенная в данном утверждении, никоим образом не связана с неизвестным решением x_*. Более того, само существование решения является следствием условий 1), 2).

Обратим внимание на следующее обстоятельство. Если известно, что искомый корень принадлежит рассматриваемой окрестности, одного условия 1) может оказаться недостаточно для сходимости итераций. При нарушенном условии 2) вычисляемые приближения могут в этом случае выйти за пределы окрестности, где выполнено условие сходимости 1). С этим надо считаться, так как на практике, решая нелинейное уравнение, мы ищем конкретный корень, принадлежащий окрестности, найденной на стадии локализации решения. Последняя, однако, не обязана совпадать с окрестностью U_* из теоремы о достаточном условии сходимости (центром которой является искомый корень). В итоге можно столкнуться с ситуацией, показанной на рис. 1.6: на отрезке $[a, b]$ выполнено условие 1), но уже x_1 выходит за пределы $[a, b]$; сходимость, как видно, отсутствует. В этом случае целесообразно вернуться к этапу локализации, чтобы уменьшить размер окрестности (например, методом половинного деления), из которой выбирается начальное приближение. Конструктивность такого подхода опирается на следующее утверждение.

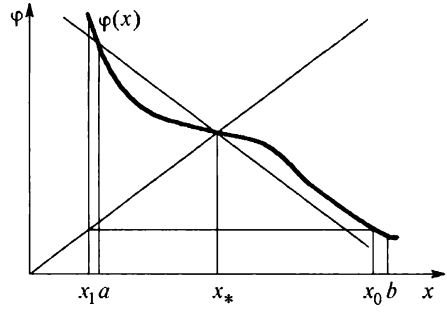


Рис. 1.6.

Утверждение 2. *Пусть функция $\varphi(x)$ в окрестности $U_b = \{|x - b| \leqslant r\}$ удовлетворяет условию Коши–Липшица и известно, что $x_* \in U_b$. Тогда существует такая окрестность $U \in U_b$, содержащая искомое решение ($x_* \in U$), что если выбрать $x_0 \in U$, то последовательные приближения (1.7) сходятся к решению x_*, и имеет место оценка (1.9).*

В заключение укажем дополнительную литературу, где вопросы решения нелинейных уравнений обсуждаются более подробно: [1, с. 80–12], [3, т. 2, с. 128–150], [5, с. 11–14], [7, с. 176–183, 189–195], [9, с. 138–150], [12, с. 190–207].

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Получить следующую оценку погрешности метода простых итераций

$$|x_n - x_*| \leqslant \frac{q}{1-q}|x_n - x_{n-1}| \leqslant \frac{q^n}{1-q}|x_1 - x_0|$$

при выполненных условиях теоремы о сходимости.

2. Найти диапазон τ, для которых метод релаксации сходится, если на $[a, b]$ $f'(x) > 0$ ($f'(x) < 0$) и $M_1 = \max\limits_{[a,b]} |f'(x)|$.

3. Получить априорную оценку количества приближений для метода Ньютона, гарантирующих получение результата с точностью ε (выразить искомое количество итераций через M_2, m_1, используемые в (1.4), и ε).

4. Получить связь между погрешностями двух последовательных приближений метода Ньютона при вычислении кратного корня уравнения (1.1).

5. Предполагая, что для метода секущих последовательные погрешности связаны соотношением $\Delta_{n+1} \approx \text{const} \cdot \Delta_n^p$ ($\Delta = |x_n - x_*|$), проверить на основе расчетов для различных уравнений:

справедливо ли сделанное предположение?

если да, то получить оценку для показателя степени p.

Указание. Последовательные пары вычисленных значений (Δ_n, Δ_{n+1}) отобразить на графике в логарифмическом масштабе.

6. Показать, что при достаточном условии сходимости метода Ньютона (см. (1.5)) имеет место $|x_n - x_*| < |x_n - x_{n-1}|$.

ЛЕКЦИЯ 2

ЧИСЛЕННОЕ РЕШЕНИЕ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ. ПРЯМЫЕ МЕТОДЫ

Необходимые сведения из линейной алгебры. Формулировка рассматриваемой задачи. Простейшая схема метода исключения (Гаусса). Варианты метода исключения с выбором главного элемента. Метод прогонки для систем с трехдиагональной матрицей. Неустранимые погрешности при решении системы методом исключения. Число обусловленности.

Необходимые сведения из линейной алгебры. Предварительно вспомним материал линейной алгебры, относящийся к нормам векторов и матриц, а также некоторые свойства матриц.

Замечание. Мы будем полагать далее, что компоненты векторов и элементы матриц суть действительные числа. ▲

Нормой вектора $\mathbf{X}$ называется число, обозначаемое $\|\mathbf{X}\|$ и удовлетворяющее условиям:

1) $\|\mathbf{X}\| \geqslant 0$, $\|\mathbf{X}\| = 0 \iff \mathbf{X} = 0$,

2) $\|\alpha\mathbf{X}\| = |\alpha| \cdot \|\mathbf{X}\|$, α — скаляр,

3) $\|\mathbf{X} + \mathbf{Y}\| \leqslant \|\mathbf{X}\| + \|\mathbf{Y}\|$.

Примеры:

1. $\|\mathbf{X}\|_1 = \sum |x_i|$,

2. $\|\mathbf{X}\|_2 = \left(\sum x_i^2\right)^{1/2}$ — евклидова норма,

3. $\|\mathbf{X}\|_\infty = \|\mathbf{X}\|_c = \max\limits_i |x_i|$ — равномерная норма.

Векторное пространство с введенной в нем нормой называют *нормированным*. Одновременно оно является метрическим, так как норма определяет метрику — расстояние между элементами пространства:

$$\rho(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|.$$

(Впредь, как равнозначными, мы будем пользоваться, например, терминами: пространство с равномерной нормой, пространство с равномерной метрикой.)

Нормой квадратной матрицы A называется число, обозначаемое $\|A\|$ и удовлетворяющее свойствам:

1) $\|A\| \geqslant 0$, $\|A\| = 0 \iff A = 0$,

2) $\|\alpha A\| = |\alpha| \cdot \|A\|$, α — скаляр,

3) $\|A+B\| \leqslant \|A\| + \|B\|$,

4) $\|AB\| \leqslant \|A\| \cdot \|B\|$.

Норма матрицы $\|A\|$ *согласована с нормой вектора* $\|\mathbf{X}\|$, если $\|A\mathbf{X}\| \leqslant \|A\| \cdot \|\mathbf{X}\|$. Именно использование согласованных норм позволит нам далее получать требуемые оценки для погрешности методов последовательных приближений, которые мы будем рассматривать.

Норма матрицы A называется *подчиненной* норме вектора $\mathbf{X}$, если $\|A\|$ вводится следующим образом:

$$\|A\| = \sup_{\mathbf{X}\neq 0} \frac{\|A\mathbf{X}\|}{\|\mathbf{X}\|} = \sup_{\|\mathbf{X}\|=1} \|A\mathbf{X}\|.$$

Нетрудно видеть, что подчиненная норма согласована с соответствующей метрикой векторного пространства. В самом деле:

$$\frac{\|A\mathbf{X}\|}{\|\mathbf{X}\|} \leqslant \sup_{\mathbf{X}\neq 0} \frac{\|A\mathbf{X}\|}{\|\mathbf{X}\|} = \|A\|,$$

отсюда $\|A\mathbf{X}\| \leqslant \|A\| \cdot \|\mathbf{X}\|$.

Чтобы получить конкретное выражение подчиненной нормы матрицы через ее элементы, надо найти $\sup\limits_{\|\mathbf{X}\|=1} \|A\mathbf{X}\|$. Найдем, например, $\|A\|_c$ — норму матрицы, подчиненную равномерной метрике векторного пространства. Итак, пусть $\mathbf{Y} = A\mathbf{X}$. Тогда

$$\|A\mathbf{X}\|_c = \|\mathbf{Y}\|_c = \max_i |y_i| = \max_i \left|\sum_j A_{ij}x_j\right|.$$

Далее,

$$\left|\sum_j A_{ij}x_j\right| \leqslant \sum_j |A_{ij}| \cdot |x_j| \leqslant \max_j |x_j| \sum_j |A_{ij}| = \|\mathbf{X}\|_c \sum_j |A_{ij}|.$$

Следовательно, для векторов с $\|\mathbf{X}\|_c = 1$

$$\|A\mathbf{X}\|_c \leqslant \max_i \sum_j |A_{ij}|.$$

Покажем, что существует вектор, на котором найденная верхняя оценка достигается, т. е. что она является точной верхней гранью для оцениваемых величин.

Пусть максимум правой части последнего неравенства достигается при $i = i_0$. Тогда, очевидно, для вектора

$$\mathbf{X} = \{\operatorname{sign}(A_{i_0 j}),\quad j = 1, 2, \ldots, n\}$$

получим

$$y_{i_0} = (A\mathbf{X})_{i_0} = \sum_j A_{i_0 j} x_j = \sum_j A_{i_0 j}\,\mathrm{sign}(A_{i_0 j}) = \\ = \sum_j |A_{i_0 j}| = \max_i \sum_j |A_{ij}|.$$

Так как согласно полученной выше оценке это максимально возможная величина компоненты вектора $\mathbf{Y}$, то по определению равномерной нормы для выбранного вектора $\mathbf{X}$

$$\|A\mathbf{X}\|_c = \|\mathbf{Y}\|_c = \max_i |y_i| = y_{i_0} = \max_i \sum_j |A_{ij}|.$$

Итак, $\|A\|_c = \max\limits_i \sum_j |A_{ij}|$.

Аналогично легко устанавливается, что $\|A\|_1 = \max\limits_j \sum_i |A_{ij}|$ — норма, подчиненная метрике $\|X\|_1$ векторного пространства.

Можно показать также, что норма матрицы, подчиненная евклидовой метрике векторного пространства, определяется следующим образом:

$$\|A\|_{\mathrm{sp}} = \|A\|_2 = [\rho(A^{\mathrm{T}}A)]^{1/2},$$

где $\rho(A^{\mathrm{T}}A) = \max(\lambda_{A^{\mathrm{T}}A})$ — спектральный радиус матрицы $A^{\mathrm{T}}A$ (A^{T} — транспонированная матрица A), $\lambda_{A^{\mathrm{T}}A}$ — собственные значения матрицы $A^{\mathrm{T}}A$. Определенная таким образом норма называется *спектральной*. Если при этом матрица A — симметричная, т. е. $A^{\mathrm{T}} = A$, то $A^{\mathrm{T}}A = A^2$, $\max(\lambda_{A^2}) = [\max(\lambda_A)^2]$ и, следовательно, в этом случае

$$\|A\|_{\mathrm{sp}} = \|A\|_2 = \max|\lambda_A|,$$

(λ_A — собственные числа самой матрицы A).

Следует отметить, что в конечномерном линейном пространстве все нормы эквивалентны в том смысле, что, если имеет место $\|x_n\|_\alpha \underset{n\to\infty}{\longrightarrow} 0$ *(x_n — последовательность элементов пространства, α — признак нормы), то по любой другой норме также* $\|x_n\|_\beta \underset{n\to\infty}{\longrightarrow} 0$.

В заключение этого краткого экскурса в линейную алгебру отметим еще одно свойство матриц, которое нам понадобится.

Матрица A называется *положительно определенной* ($A > 0$) (неотрицательно определенной, $A \geqslant 0$), если $(A\mathbf{X}, \mathbf{X}) > 0$ ($(A\mathbf{X}, \mathbf{X}) \geqslant 0$) для любых $\mathbf{X} \neq 0$. Для положительно определенной симметричной матрицы справедливо:

$$\min_i \lambda_i > 0,$$

где λ_i — собственные значения матрицы A.

Прямые методы решения линейных систем уравнений. Возвращаясь к теме лекции, для начала отметим, что задачи линейной алгебры являются самыми распространенными с прикладной точки зрения, и поэтому вызывают неизменно возрастающий интерес и со стороны математиков-вычислителей.

Согласно информации из [8, с. 3], в библиографическом указателе библиотеки МИАНа*) сделана попытка собрать сведения о всех работах *вычислительного* характера по линейной алгебре, начиная от статьи Гаусса по обработке результатов наблюдений по методу наименьших квадратов (1828 г.):

с 1828 г. по 1974 г. (т. е. за 147 лет) — 4000 наименований;

с 1975 г. по 1980 г. (т. е. за 5 лет) — 3000;

с 1981 г. по 1984 г. (т. е. за 3 года) — 4000.

Задачи линейной алгебры — это:

— решение систем линейных уравнений,

— вычисление определителей и обращение матриц,

— вычисление собственных значений и собственных векторов матриц.

В практических приложениях наиболее часто приходится иметь дело с первой задачей. (Мы увидим, что необходимость решать линейные системы уравнений возникает при реализации едва ли не всех методов, с которыми мы будем сталкиваться). Именно об этой задаче пойдет речь ниже. По поводу остальных отметим лишь, что решение второй задачи, в конечном счете, сводится к первой; последняя же связана с проблемой поиска корней алгебраического полинома n-й степени (n — порядок матрицы).

З а м е ч а н и е. В конце последнего абзаца обозначен формальный подход к проблеме поиска собственных чисел матрицы. Современные методы не требуют выписывания характеристического многочлена. См. [1, с. 211; 2, с. 309; 9, с. 156]. ▲

Итак, пусть требуется найти решение системы

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= f_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= f_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= f_n \end{aligned} \tag{2.1}$$

или в компактной (векторной) форме

$$A\mathbf{X} = \mathbf{f}.$$

---

*) Математический институт им. В. А. Стеклова Российской академии наук.

Мы будем считать, что $\Delta = \det A \neq 0$, т. е. решение (2.1) существует и единственно.

В принципе, известны формулы, дающие в явной форме решение задачи (2.1). Это формулы Крамера:

$$x_i = \frac{\Delta_i}{\Delta},$$

где Δ_i — определитель матрицы, которая получается из матрицы A заменой столбца с номером i столбцом правых частей (2.1). Определители при этом предлагается вычислять по формулам, рассматриваемым в курсах линейной алгебры. Например, для Δ:

$$\Delta = \sum_{P_n} (-1)^{[p_1, \ldots, p_n]} a_{p_1 1} a_{p_2 2} \ldots a_{p_n n},$$

где последовательности $p_1, p_2, \ldots, p_n$ представляют собой различные перестановки натуральных чисел от 1 до n, $P_n = n!$ — число возможных перестановок, а $[p_1, p_2, \ldots, p_n]$ — число так называемых «беспорядков» в перестановке.

Однако в качестве конкретного метода решения системы (2.1) данные формулы совершенно неприменимы, так как при подсчете каждого определителя по приведенной выше формуле надо вычислить $n!$ слагаемых, что нереально при весьма умеренных n. Например, уже при $n = 100$ имеем $100! \gg 10^{90}$. Если одно слагаемое вычисляется, скажем, за 10^{-6} с, что вполне допустимо для современных машин, то время расчета составит совершенно фантастическую цифру:

$$T \geqslant 10^{90} \cdot 10^{-6}\,\text{с} = \frac{10^{84}}{86\,400}\,\text{суток} \approx 3 \cdot 10^{76}\,\text{лет!}$$

Фактически же в настоящее время с использованием подходящих методов решаются системы гораздо более высокого порядка (до $n \approx 10^4$). Что же это за методы? Их множество. В целом они разделяются на две группы: прямые и итерационные.

Прямые (или *конечные*) *методы* позволяют теоретически (в предположении, что вычисления проводятся без ошибок округления) получить точное решение задачи (2.1) за *конечное* число арифметических операций.

Итерационные методы, другими словами, *методы последовательных приближений*, позволяют вычислять последовательность векторов $\{\mathbf{X}^{(n)}\}$, сходящуюся при $n \to \infty$ к решению задачи (2.1). На практике при использовании итерационных методов ограничиваются вычислением конечного числа приближений в зависимости от допустимого уровня погрешности.

Простейшая схема метода исключения (Гаусса). Варианты. В данной лекции речь пойдет о прямых методах. Это, как правило, различные варианты метода последовательного исключения неизвестных. Мы рассмотрим простейшие.

Очевидно, что сложность системы (2.1) определяется структурой матрицы A. Если A — диагональная матрица

$$A = D = \begin{pmatrix} d_1 & & & 0 \\ & d_2 & & \\ & & \ddots & \\ 0 & & & d_n \end{pmatrix},$$

то система распадается на n линейных уравнений, каждое из которых содержит одну неизвестную величину, и проблем с вычислениями не возникает.

Просто решается задача (2.1) и в случае, когда матрица A является треугольной. Пусть, например,

$$A = A_+ = \begin{pmatrix} a_{11} & a_{12} & & a_{1n} \\ & a_{22} & & a_{2n} \\ & & \ddots & \\ 0 & & & a_{nn} \end{pmatrix},$$

Очевидно, для всех i $a_{ii} \neq 0$, так как $\det A \neq 0$. Тогда из последнего уравнения

$$x_n = \frac{f_n}{a_{nn}};$$

далее, $x_m = \dfrac{1}{a_{mm}} (f_m - a_{mn}x_n - a_{mn-1}x_{n-1} - \ldots - a_{mm+1}x_{m+1})$ для $m = n{-}1, n{-}2, \ldots, 2, 1$.

Замечание. Ввиду большого многообразия методов решения задачи (2.1) (да и других задач, с которыми мы будем иметь дело) в дальнейшем важную роль будет играть такой критерий отбора, как трудоемкость метода, выраженная в количестве требуемых арифметических операций. ▲

Оценим объем вычислений, связанный с решением системы с треугольной матрицей. Чтобы вычислить x_n, надо выполнить одну операцию, для вычисления x_{n-1} — три, x_{n-2} — пять и т. д. Нетрудно усмотреть, что общее число операций равно

$$\Omega = \sum_{m=1}^{n} (2m - 1) = n^2.$$

Перейдем теперь к обсуждению вариантов метода исключения для системы (2.1) с матрицей общего вида.

Типичная схема реализации метода исключения разделяется на два этапа: *прямой ход* — приведение исходной матрицы к треугольному виду и *обратный ход* — решение системы с треугольной матрицей. Рассмотрим реализацию прямого хода.

Пусть $a_{11} \neq 0$. В противном случае всегда можно переставить уравнения в (2.1), так чтобы это условие выполнялось (ввиду того, что $\det A \neq 0$, все элементы первого столбца не могут одновременно быть нулевыми). Умножая первое уравнение (2.1) на $l_i^{(1)} = -a_{i1}/a_{11}$ и складывая результат с i-м уравнением, исключаем из i-го уравнения неизвестную x_1. Проделав эту операцию со всеми уравнениями, начиная со второго ($i = 2, 3, \ldots, n$), придем к системе уравнений следующего вида

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= f_1, \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n &= f_2^{(1)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n &= f_n^{(1)}, \end{aligned}$$

где $a_{ij}^{(1)} = a_{ij} + l_i^{(1)}a_{1j}$, $f_i^{(1)} = f_i + l_i^{(1)}f_1$ для $i = 2, 3, \ldots, n$, $j = 2, 3, \ldots, n$.

Предполагая, что $a_{22}^{(1)} \neq 0$ (если это не так, то достаточно переставить должным образом уравнения для $i = 2, 3, \ldots, n$), и, используя коэффициенты $l_i^{(2)} = -a_{i2}^{(1)}/a_{22}^{(1)}$, исключаем x_2 из всех уравнений для $i = 3, 4, \ldots, n$. Аналогичным образом из остающихся уравнений исключаются x_3, x_4 и так далее. Очевидно, выполнив $(n-1)$ шагов подобного рода, придем к системе уравнений с треугольной матрицей, которая решается так, как было сказано выше (это уже обратный ход).

Какова трудоемкость этого метода? Обратный ход, как мы видели, требует выполнения n^2 арифметических операций. В то же время из общих соображений ясно, что полное число операций должно быть пропорционально n^3 (мы ищем n неизвестных, каждая из которых должна зависеть от n^2 элементов матрицы). В самом деле, аккуратный подсчет операций, выполняемых на первом этапе исключения, приводит к следующей оценке объема вычислительной работы в рамках прямого хода: $\Omega = \frac{1}{6}n(n-1)(4n+7)$.

При больших n количество операций $\Omega \sim \frac{2}{3}n^3$. Это вполне приемлемая величина (при $n \sim 10^3$, быстродействии ЭВМ порядка 10^6 операций в секунду требуемое для расчета время порядка одного часа).

Мы рассмотрели простейшую схему исключения (этот метод называют также методом Гаусса) и далеко не лучшую. Обратимся еще раз к примеру, который приводился в «Предварительных замечаниях» к лекциям. Рассматривалась система

$$-10^{-7}x_1 + x_2 = 1,$$
$$x_1 + 2x_2 = 4.$$

Мы видели, что в одном из вариантов метода исключения результаты получались совершенно неверными. Напомним «механизм» возникновения больших погрешностей: деление на малые числа, появление больших (по величине) промежуточных результатов, потеря точности при вычитании больших (близких друг к другу) чисел.

Таким образом, порядок последовательного исключения неизвестных может сильно сказаться на результатах расчетов (тем более для систем высокого порядка такой исход весьма вероятен). Уменьшить опасность подобного рода, т. е. уменьшить в процессе выкладок вероятность деления на малые числа, позволяют варианты метода Гаусса с выбором *главного элемента*.

Выбор главного элемента по столбцам. Перед исключением x_1 отыскивается $\max\limits_i |a_{i1}|$. Допустим, максимум соответствует $i = i_0$. Тогда первое уравнение в исходной системе (2.1) меняем местами с i_0-м уравнением. (Для ЭВМ эта процедура связана с перестановкой двух строк расширенной матрицы (2.1).) После этого осуществляется первый шаг исключения. Затем перед исключением x_2 из оставшихся уравнений отыскивается $\max\limits_{2 \leqslant i \leqslant n} |a_{i2}^{(1)}|$, осуществляется соответствующая перестановка уравнений и т. д.

Выбор главного элемента по строке. Перед исключением x_1 отыскивается $\max\limits_j |a_{1j}|$. Пусть максимум достигается при $j = j_0$. Тогда поменяем взаимно номера у неизвестных x_1 и x_{j_0} (максимальный по величине из коэффициентов 1-го уравнения окажется в позиции a_{11}) и приступим к процедуре исключения x_1, и т. д. Наиболее надежным является метод исключения с *выбором главного элемента по всей матрице* коэффициентов на каждом шаге исключения.

Рассмотренные модификации метода Гаусса позволяют, как правило, существенно уменьшить неблагоприятное влияние погрешностей округления на результаты расчета.

Впрочем, в прикладных задачах довольно часто приходится сталкиваться с линейными системами, при решении которых можно не заботиться о «вредном» воздействии неустранимых погрешностей на решение, спокойно применяя простейшую схему гауссова исключения

(без выбора главного элемента). Это системы, для матриц которых выполнено *условие диагонального преобладания*:

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}| \text{ для всех } i = 1, 2, \ldots, n.$$

Можно показать (см. упражнение 4), что условие диагонального преобладания остается справедливым после каждого шага исключений в процессе приведения матрицы к треугольному виду, т. е.

$$|a_{ii}^{(k)}| > \sum_{\substack{j=k \\ j\neq i}}^{n} |a_{ij}^{(k)}|, \quad i = k, k+1, \ldots, n$$

для всех $k = 1, 2, \ldots, n-1$. Это означает, что перед каждым исключением очередной неизвестной главный элемент будет находиться в «нужной позиции».

Более подробно вопросы, связанные с поведением *неустранимых погрешностей* при решении линейных систем, будут затронуты в дополнении к лекции.

В заключение рассмотрим важный случай *систем специального вида* (*с трехдиагональной матрицей*):

$$\begin{aligned}
b_1x_1 + c_1x_2 \qquad\qquad\qquad &= f_1,\\
a_2x_1 + b_2x_2 + c_2x_3 \qquad\quad &= f_2,\\
a_3x_2 + b_3x_3 + c_3x_4 \quad &= f_3,\\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots &\\
a_{n-1}x_{n-2} + b_{n-1}x_{n-1} + c_{n-1}x_n &= f_{n-1},\\
a_nx_{n-1} + b_nx_n &= f_n.
\end{aligned} \tag{2.2}$$

Мы увидим в дальнейшем, что необходимость решать подобные системы довольно часто возникает в качестве составного элемента при реализации различных численных методов решения дифференциальных уравнений.

Применим для решения системы (2.2) простейшую схему метода Гаусса (без выбора главного элемента). Поделив первое уравнение на b_1, перепишем его в виде

$$x_1 + p_1x_2 = q_1, \quad \left(p_1 = \frac{c_1}{b_1},\ q_1 = \frac{f_1}{b_1}\right).$$

Умножая это уравнение на a_2 и вычитая его из второго уравнения системы (2.2), исключаем из последнего x_1:

$$(b_2 - a_2p_1)x_2 + c_2x_3 = f_2 - a_2q_1.$$

Переписываем последнее уравнение в виде

$$x_2 + p_2 x_3 = q_2, \quad \left(p_2 = \frac{c_2}{b_2 - p_1 a_2},\ q_2 = \frac{f_2 - q_1 a_2}{b_2 - p_1 a_2}\right).$$

Исключаем с помощью этого соотношения x_2 из третьего уравнения системы (2.2) и т. д. В итоге приходим к системе уравнений с двухдиагональной матрицей

$$\begin{gathered} x_i + p_i x_{i+1} = q_i \quad (i = 1, 2, \ldots, n-1,) \\ x_n = q_n, \end{gathered} \tag{2.3}$$

коэффициенты и правые части которой вычисляются по формулам

$$\begin{gathered} p_1 = \frac{c_1}{b_1}, \quad p_i = \frac{c_i}{b_i - p_{i-1} a_i}, \quad i = 2, 3, \ldots, n-1, \\ q_1 = \frac{f_1}{b_1}, \quad q_i = \frac{f_i - q_{i-1} a_i}{b_i - p_{i-1} a_i}, \quad i = 2, 3, \ldots, n. \end{gathered} \tag{2.4}$$

На основе проделанных выкладок можно сформулировать алгоритм решения задачи (2.2), состоящий из двух этапов:

1) по формулам (2.4) вычисляются массивы p_i и q_i;

2) по формулам

$$x_n = q_n, \quad x_i = q_i - p_i x_{i+1}, \quad i = n-1, n-2, \ldots, 1, \tag{2.5}$$

вытекающим из (2.3), вычисляется искомое решение.

Этот алгоритм известен под названием *метода прогонки* для систем с трехдиагональной матрицей. Первый этап — так называемая *прямая прогонка*, второй — *обратная прогонка*.

Ввиду опасности обращения знаменателя в формулах (2.4) в ноль возникает вопрос об условиях применимости метода прогонки. Таковыми являются, в частности, условия диагонального преобладания в исходной матрице: $|b_i| > |a_i| + |c_i|$ для всех i. В самом деле, в этом случае $|p_1| < 1$.

Используя логику индукции, покажем, что $|p_i| < 1$ для всех $i = 2, 3, \ldots, n-1$. Итак, пусть $|p_{i-1}| < 1$; тогда

$$|p_i| = \frac{|c_i|}{|b_i - p_{i-1} a_i|} \leqslant \frac{|c_i|}{\|b_i| - |p_{i-1}\| a_i\|} < \frac{|c_i|}{\|b_i| - |a_i\|} < \frac{|c_i|}{|c_i|} = 1.$$

Утверждение доказано. Приведенная цепочка неравенств показывает одновременно, что $|b_i - p_{i-1} a_i| > |c_i| \geqslant 0$, то есть знаменатель в формулах (2.4) всегда отличен от нуля.

Замечание. Точно также можно обосновать применимость прогонки, организованной в обратном порядке («снизу вверх»). ▲

Отметим, что метод прогонки относится к классу *экономичных* методов. *Экономичными называются методы, для которых число требуемых арифметических операций пропорционально числу неизвестных.* Нетрудно видеть, что расчет по формулам (2.4), (2.5) действительно требует выполнения порядка $8n$ элементарных операций. Экономичность в данном случае (в противовес общему случаю (2.1)) достигнута за счет того, что при реализации метода исключения, приведшего к формулам (2.4) и (2.5), мы не выполняли операций над нулевыми элементами исходной матрицы.

ДОПОЛНЕНИЕ К ЛЕКЦИИ 2

О неустранимой погрешности при решении линейных систем методом исключения. Пример, рассмотренный в Лекции 2, показывает, что при неудачной последовательности действий в процессе исключения неизвестных может произойти недопустимое искажение результатов. «Механизм» этого явления: деление на малое число — вычитание больших чисел — потеря точности. В качестве превентивных мер, позволяющих снизить неблагоприятный эффект, предлагалось использовать модификации метода гауссова исключения с выбором главного элемента. При этом ограничивается рост (по величине) пересчитываемых элементов матрицы на этапе прямого хода, и вероятность значительной потери точности существенно понижается (см. упражнения 2, 3).

Остановимся несколько подробней на оценке возможной *неустранимой* погрешности решения системы линейных уравнений. Известно, что источниками неустранимой погрешности являются не только округления при выполнении машинных операций, но также ошибки, содержащиеся в исходных данных. Разберемся сначала с последними, предполагая, что арифметические операции выполняются точно.

Итак, пусть вместо системы

$$A\mathbf{X} = \mathbf{f} \tag{2.6}$$

решается задача

$$(A + \delta A)(\mathbf{X} + \delta\mathbf{X}) = \mathbf{f} + \delta\mathbf{f}. \tag{2.7}$$

Здесь δA — матрица возмущений, моделирующих ошибки коэффициентов исходных уравнений (2.6), $\delta\mathbf{f}$ — соответственно возмущения правых частей, $\delta\mathbf{X}$ — обусловленный этими возмущениями вектор «ошибок», отличающий решение (2.7) от решения (2.6).

Переписывая (2.7) в виде

$$A\mathbf{X} + A\,\delta\mathbf{X} + \delta A\,\mathbf{X} + \delta A\,\delta\mathbf{X} = \mathbf{f} + \delta\mathbf{f}$$

и вычитая из последнего соотношение (2.6), приходим к системе уравнений

$$A\delta\mathbf{X} + \delta A\,\delta\mathbf{X} = \delta\mathbf{f} - \delta A\,\mathbf{X}, \tag{2.8}$$

которая описывает зависимость $\delta\mathbf{X}$ от возмущений (ошибок) исходных данных.

Далее будем полагать, что возмущения коэффициентов уравнений δA и погрешности решения $\delta\mathbf{X}$ в достаточной мере малы, так что в уравнениях (2.8) можно пренебречь квадратичными членами $\delta A\,\delta\mathbf{X}$. Тогда интересующую нас ошибку $\delta\mathbf{X}$ можно представить в виде

$$\delta\mathbf{X} \simeq A^{-1}(\delta\mathbf{f} - \delta A\,\mathbf{X}).$$

Вводя в рассмотрение нормы векторов и согласованные с ними нормы матриц, получим оценку величины погрешности:

$$\|\delta\mathbf{X}\| \simeq \|A^{-1}(\delta\mathbf{f} - \delta A\,\mathbf{X})\| \leqslant \|A^{-1}\|\,(\|\delta\mathbf{f}\| + \|\delta A\| \cdot \|\mathbf{X}\|) =$$

$$= \|A^{-1}\|\left(\|\mathbf{f}\|\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|} + \|A\|\frac{\|\delta A\|}{\|A\|}\|\mathbf{X}\|\right).$$

Учитывая, что $\|\mathbf{f}\| = \|A\mathbf{X}\| \leqslant \|A\| \cdot \|\mathbf{X}\|$, получаем далее

$$\|\delta\mathbf{X}\| \leqslant \|A^{-1}\|\left(\|A\| \cdot \|\mathbf{X}\|\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|} + \|A\| \cdot \|\mathbf{X}\|\frac{\|\delta A\|}{\|A\|}\right) =$$

$$= \|A^{-1}\| \cdot \|A\| \cdot \|\mathbf{X}\|\left(\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|} + \frac{\|\delta A\|}{\|A\|}\right).$$

В итоге оценка для относительной погрешности решения может быть записана в виде

$$\frac{\|\delta\mathbf{X}\|}{\|\mathbf{X}\|} \lesssim \mu_A\left(\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|} + \frac{\|\delta A\|}{\|A\|}\right), \tag{2.9}$$

где $\mu_A = \|A^{-1}\| \cdot \|A\|$. Значение μ_A называется *числом обусловленности матрицы* A. Именно эта величина определяет, насколько сильно погрешности входных данных могут повлиять на решение системы (2.6).

Всегда $\mu_A \geqslant 1$. В самом деле, имеем $E = A^{-1}A$. Отсюда $1 = \|E\| = \|A^{-1}A\| \leqslant \|A^{-1}\| \cdot \|A\| = \mu_A$. Если значение μ_A является умеренным ($\mu_A \sim 1 \div 10$), ошибки входных данных слабо сказываются на решении; система (2.6) в этом случае называется *хорошо обусловленной*. Если μ_A велико ($\mu_A \geqslant 10^3$), система (2.6) *плохо обусловлена*, решение ее сильно зависит от ошибок в правых частях и коэффициентах.

Замечание 1. Вообще говоря, более точное представление о хорошей или плохой обусловленности системы должно опираться на

требования, предъявляемые к решению. Если, к примеру, погрешность входных данный $\sim 10^{-6}$, а допустимая погрешность решения $\sim 10^{-2}$, то даже при $\mu \sim 10^4$ систему можно считать хорошо обусловленной. ▲

Замечание 2. Хотелось бы подчеркнуть, что данное свойство (обусловленность), выражаемое неравенством (2.9), никак не связано с предполагаемым методом решения системы, а является изначальной характеристикой решаемой задачи. ▲

Пример. Рассмотрим систему

$$\begin{aligned} 100x_1 + 99x_2 &= 199, \\ 99x_1 + 98x_2 &= 197. \end{aligned}$$

Ее решение: $x_1 = x_2 = 1$.

Исказим теперь слегка ее правые части:

$$\begin{aligned} 100x_1 + 99x_2 &= 198.99, \\ 99x_1 + 98x_2 &= 197.01. \end{aligned}$$

Решение «искаженной» системы: $x_1 = 2.97$, $x_2 = -0.99$.

Чтобы сопоставить полученные результаты с оценкой (2.9), будем пользоваться следующими согласованными нормами для векторов и матриц:

$$\|\mathbf{X}\| = \max_i |x_i|, \quad \|A\| = \max_i \sum_j |a_{ij}|.$$

Для рассмотренного примера имеем

$$\mathbf{f} = \begin{pmatrix} 199 \\ 197 \end{pmatrix}, \quad \delta\mathbf{f} = \begin{pmatrix} -0.01 \\ 0.01 \end{pmatrix}, \quad \text{т. е. } \|\mathbf{f}\| = 199, \quad \|\delta\mathbf{f}\| = 0.01.$$

Относительная погрешность $\dfrac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|} \approx \dfrac{1}{2} \cdot 10^{-4} = 0.005\%$. Это очень малая величина.

Далее, $\|A\| = 199$,

$$\det A = -1, \quad A^{-1} = \begin{pmatrix} -98 & 99 \\ 99 & -100 \end{pmatrix}, \quad \|A^{-1}\| = 199,$$

$$\mu_A = (199)^2 = 39601 \approx 4 \cdot 10^4.$$

Согласно (2.9) $\dfrac{\|\delta\mathbf{X}\|}{\|\mathbf{X}\|} \leqslant 4 \cdot 10^4 \cdot \dfrac{10^{-4}}{2} = 2(!)$, что, как видно, согласуется с результатами решения рассмотренных систем.

Замечание. Если бы для величин $\frac{\|\delta \mathbf{f}\|}{\|\mathbf{f}\|}$, μ_A мы использовали точные значения, то получили бы, что в данном примере оценка (2.9) для ошибки решения достигается. Следовательно, эта оценка не является завышенной. ▲

Возвращаясь к вопросу об оценке влияния на решение системы погрешностей округления при выполнении арифметических операций, будем краткими и, следуя [8], лишь сошлемся на результаты известных исследований.

Можно показать [6, 8], что машинное решение $\mathbf{X}' = \mathbf{X} + \delta\mathbf{X}$ линейной системы, вычисленное методом Гаусса (с той или иной схемой выбора главного элемента или вообще без выбора), точно удовлетворяет уравнениям с определенным образом возмущенными коэффициентами

$$(A + \delta A)\mathbf{X}' = \mathbf{f}.$$

Для нормы матрицы так называемых эквивалентных возмущений δA справедлива оценка вида

$$\|\delta A\| \approx ng(A)\|A\|p^{-t}.$$

Здесь n — порядок системы, p — основание машинной арифметики (как правило, $p = 2$), t — число значащих цифр, учитываемых при выполнении арифметических операций, $g(A) = \max\limits_k \dfrac{\max\limits_{i,j} |a_{ij}^{(k)}|}{\max\limits_{i,j} |a_{ij}|}$, где k — номер шага на этапе прямого хода метода исключений. (Таким образом, величина $g(A)$ показывает, насколько могут возрасти пересчитываемые элементы матрицы на стадии приведения ее к треугольному виду. Отсюда ее название — коэффициент роста.)

Привлекая далее (2.9), получаем оценку ошибок решения системы (2.6) за счет погрешностей вычислений:

$$\frac{\|\delta \mathbf{X}\|}{\|\mathbf{X}\|} \approx \mu_A n g(A) p^{-t}. \tag{2.10}$$

Если, например, решается система из 10^3 уравнений и $\mu_A \approx 1$, $g(A) \approx 1$, то при счете с ординарной точностью на 32-разрядных машинах ($p = 2$, $t = 26$) нельзя рассчитывать на точность лучшую, нежели $\frac{\|\delta \mathbf{X}\|}{\|\mathbf{X}\|} \sim n \cdot 2^{-26} \sim n \cdot 10^{-7}|_{n=10^3} \sim 10^{-4}$. Если при этом $\mu_A \approx 10^4$, то может произойти полная потеря точности.

Замечание. Плохо обусловленные системы вызывают определенные трудности при решении. Из (2.9) следует, что решение их

сильно зависит от ошибок входных данных, а из (2.10) вытекает, что даже при отсутствии ошибок во входных величинах может произойти значительная (если не полная) потеря точности на стадии вычислений по методу Гаусса за счет погрешностей округлений. ▲

Более подробные сведения по рассмотренным здесь вопросам можно найти в [1, с. 122–173], [2, с. 257–268, 303–308], [3, т. 2, с. 9–53], [4, с. 15–70], [8], [9, с. 126–138], [11, с. 85–95], [12, с. 48–81], [16].

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Предполагается, что для решения системы

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 = f_1,\\
&a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = f_2,\\
&\ldots\ldots\ldots\ldots\ldots\\
&a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mm}x_m + a_{mm+1}x_m + 1 = f_m,\\
&\ldots\ldots\ldots\ldots\ldots\\
&a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn-1}x_{n-1} + a_{nn}x_n = f_n.
\end{aligned}
$$

используется метод Гаусса без выбора главного элемента, но с учетом специфики матрицы (над нулевыми элементами арифметические действия не производятся).

Оценить количество элементарных арифметических операций, необходимых для вычисления решения, если исключение элементов производится

а) «сверху вниз»;

б) «снизу вверх».

2. Показать, что при использовании метода исключения с выбором главного элемента по столбцам максимальный по модулю элемент пересчитанной матрицы после каждого шага исключения не может более чем в два раза превосходить максимальный по модулю элемент предыдущей матрицы.

3. То же для метода исключения с выбором главного элемента по строкам.

4*. Показать, что свойство диагонального преобладания матрицы сохраняется при использовании для решения системы метода Гаусса без выбора главного элемента.

5. Как, используя метод исключения, вычислить определитель заданной матрицы?

6. Как вычислить элементы матрицы, обратной заданной?

7. Для системы

$$
\begin{aligned}
10^{-3}x_1 + x_2 &= f_1,\\
x_1 + x_2 &= f_2
\end{aligned}
$$

ответить на следующие вопросы:

а) каково число обусловленности матрицы этой системы;

б) какова допустимая относительная погрешность правых частей, при которой относительная погрешность решения не превосходит 10^{-2}?

ЛЕКЦИЯ 3

ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ ЛИНЕЙНЫХ СИСТЕМ

Метод простых итераций. Достаточное условие сходимости. Примеры итерационных процессов: метод Якоби, метод Зейделя, однопараметрический метод итераций. Оптимизация параметра. Возможные обобщения. Влияние неустранимых погрешностей на последовательные приближения.

Метод простых итераций. Мы переходим к обсуждению итерационных методов (то есть методов последовательных приближений) решения линейных систем уравнений:

$$A\mathbf{X} = \mathbf{f}. \tag{3.1}$$

(Как и в прошлой лекции, мы считаем, что решение (3.1) существует и единственно.)

Различные варианты метода простых итераций связаны с переходом от системы (3.1) к эквивалентной системе

$$\mathbf{X} = P\mathbf{X} + \mathbf{g}. \tag{3.2}$$

Итерационный процесс, опираясь на (3.2), строим очевидным образом:

$$\mathbf{X}^{(k)} = P\mathbf{X}^{(k-1)} + \mathbf{g}, \qquad \mathbf{X}^{(0)} \text{ задано}. \tag{3.3}$$

Здесь k — номер приближения.

Условия сходимости метода последовательных приближений (3.3) формулируются в следующих теоремах.

Теорема 1. *Для сходимости итераций (3.3) к решению системы (3.2) достаточно, чтобы в какой-либо норме выполнялось условие* $\|P\| \leqslant q < 1$.

Тогда независимо от выбора $\boldsymbol{X}^{(0)}$

$$\|\mathbf{X}^{(k)} - \mathbf{X}^*\| \leqslant q^k \|\mathbf{X}^{(0)} - \mathbf{X}^*\|, \tag{3.4}$$

где $\mathbf{X}^*$ — точное решение (3.2).

Доказательство. Подстановка точного решения в (3.2) обращает последнее в тождество:

$$\mathbf{X}^* \equiv P\mathbf{X}^* + \mathbf{g}.$$

Вычитая его из (3.3), получим

$$\mathbf{X}^{(k)} - \mathbf{X}^* = P(\mathbf{X}^{(k-1)} - \mathbf{X}^*),$$

где $\mathbf{X}^{(k)} - \mathbf{X}$ — вектор погрешности (или просто погрешность) k-го приближения.

Оценивая величину погрешности по какой-либо норме (с которой согласована норма матрицы, фигурирующая в условии теоремы), получаем

$$\|\mathbf{X}^{(k)} - \mathbf{X}^*\| \leqslant \|P\| \cdot \|\mathbf{X}^{(k-1)} - \mathbf{X}^*\| \leqslant$$
$$\leqslant q\|\mathbf{X}^{(k-1)} - \mathbf{X}^*\| \leqslant \ldots \leqslant q^k\|\mathbf{X}^{(0)} - \mathbf{X}^*\|.$$

Очевидно, что при $q < 1$ $\lim\limits_{k\to\infty} \mathbf{X}^{(k)} = \mathbf{X}^*$.

Теорема 2 (без доказательства). *Для сходимости итераций (3.3) к решению системы (3.2) необходимо и достаточно, чтобы все собственные значения матрицы P по абсолютной величине были меньше единицы.*

Опираясь на (3.4), можно получить априорную оценку числа приближений, гарантирующую, что вычисленное решение будет отличаться от точного не более, чем на малое число ε.

$$\|\mathbf{X}^{(k)} - \mathbf{X}^*\| \leqslant q^k\|\mathbf{X}^{(0)} - \mathbf{X}^*\| \leqslant \varepsilon.$$

Разрешая последнее неравенство относительно k, получаем

$$k \geqslant \left\lceil \frac{\lg \dfrac{\varepsilon}{\|\mathbf{X}^{(0)} - \mathbf{X}^*\|}}{\lg q} \right\rceil. \tag{3.5}$$

Квадратные скобки означают ближайшее целое (сверху) к значению выражения, заключенного в эти скобки. Оценкой (3.5) можно воспользоваться, если мажорировать каким-то разумным образом неизвестную начальную погрешность $\|\mathbf{X}^{(0)} - \mathbf{X}^*\|$ (см. упражнение 3).

Теперь самое время осознать, зачем, собственно, нужны итерационные методы, если мы умеем вычислять решение, пользуясь, например, какой-либо модификацией метода Гаусса. Вопрос становится ясным, если оценить эффективность различных подходов с точки зрения вычислительных затрат.

Метод Гаусса (в простейшей интерпретации), как мы видели, при $n \gg 1$ требует выполнения приблизительно $(2/3)n^3$ арифметических операций. Метод итераций (3.3) реализуется приблизительно

за $(2n^2)K$ операций ($2n^2$ умножений и сложений связано с умножением матрицы P на вектор $\mathbf{X}^{(k-1)}$, K — число приближений). Если допустимая погрешность достигается при $K < n/3$, то метод итераций становится предпочтительней. В задачах, с которыми практически приходится иметь дело, зачастую $K \ll n$.

Кроме того, методы итераций могут оказаться предпочтительней с точки зрения устойчивости вычислений, в смысле влияния вычислительных погрешностей на результаты расчетов (см. дополнения к лекции).

Примеры итерационных процессов.

Метод Якоби. Запишем каждое уравнение (3.1) в виде, разрешенном относительно неизвестного с коэффициентом на главной диагонали матрицы A:

$$x_m = \frac{1}{a_{mm}}(f_m - a_{m1}x_1 - a_{mm-1}x_{m-1} - a_{mm+1}x_{m+1} - \ldots - a_{mn}x_n),$$

$m = 1, 2, \ldots, n$.

То есть мы переписали (3.1) в виде (3.2) с матрицей

$$P = -\begin{pmatrix} 0 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \ldots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & 0 & \frac{a_{23}}{a_{22}} & \ldots & \frac{a_{2n}}{a_{22}} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \frac{a_{n1}}{a_{nn}} & \frac{a_{n2}}{a_{nn}} & \frac{a_{n3}}{a_{nn}} & \ldots & 0 \end{pmatrix}.$$

Если ввести в рассмотрение диагональную матрицу

$$D = \begin{pmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{nn} \end{pmatrix}, \quad \text{то} \quad \begin{aligned} P &= -D^{-1}(A - D), \\ \mathbf{g} &= D^{-1}\mathbf{f}. \end{aligned}$$

Итерационный процесс (3.3) с определенной таким образом матрицей P называется *методом Якоби*. Фактически вычисления проводятся по формулам

$$x_m^{(k)} = \frac{1}{a_{mm}}(f_m - a_{m1}x_1^{(k-1)} - \ldots - a_{mm-1}x_{m-1}^{(k-1)} - a_{mm+1}x_{m+1}^{(k-1)} - \ldots - a_{mn}x_n^{(k-1)}), \quad m = 1, 2, \ldots, n. \qquad (3.6)$$

Для сходимости метода Якоби достаточно, чтобы для исходной матрицы A имело место диагональное преобладание, т. е. чтобы ко-

эффициенты исходных уравнений удовлетворяли условиям

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \text{для всех } i.$$

В самом деле, тогда условие теоремы 1 выполнено для нормы $\|P\|_c$:

$$\|P\|_c = \max_i \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} = \max_i \frac{\sum\limits_{j \neq i} |a_{ji}|}{|a_{ii}|} < 1.$$

Метод Зейделя. Этот метод отличается от метода Якоби только тем, что при вычислении k-го приближения m-й компоненты используются уже вычисленные k-е приближения предыдущих (1-й, 2-й, ..., $(m-1)$-й) компонент:

$$\begin{aligned}
x_1^{(k)} &= \frac{1}{a_{11}}(f_1 - a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)} - \ldots - a_{1n}x_n^{(k-1)}),\\
x_2^{(k)} &= \frac{1}{a_{22}}(f_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k-1)} - \ldots - a_{2n}x_n^{(k-1)}),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
x_m^{(k)} &= \frac{1}{a_{mm}}(f_m - a_{m1}x_1^{(k)} - a_{m2}x_2^{(k)} - \ldots - a_{mm-1}x_{m-1}^{(k)} -\\
&\qquad - a_{mm+1}x_{m+1}^{(k-1)} - \ldots - a_{mn}x_n^{(k-1)}),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{3.7}$$

Если представить матрицу A в виде суммы $A = A_- + D + A_+$ (A_-, A_+, D — нижняя треугольная, верхняя треугольная и диагональная матрицы с элементами исходной матрицы A), то методу Зейделя соответствует матрица

$$P = -(A - A_+)^{-1}A_- = -(A_- + D)^{-1}A_-.$$

Можно доказать, что метод Зейделя гарантировано сходится, если:

— выполнено условие диагонального преобладания матрицы A; или

— матрица A является симметричной и положительно определенной.

В одинаковых условиях (при наличии диагонального преобладания) метод Зейделя сходится примерно в два раза быстрее метода Якоби.

Однопараметрический метод итераций. Перепишем (3.1) в виде $\mathbf{X} = \mathbf{X} - \tau(A\mathbf{X} - \mathbf{f}) = (E - \tau A)\mathbf{X} + \tau\mathbf{f}$, где τ — пока неопределенный параметр.

Мы фактически привели (3.1) к форме (3.2) с $P = (E - \tau A)$, $\mathbf{g} = \tau\mathbf{f}$. Чтобы обеспечить сходимость итераций

$$\mathbf{X}^{(k)} = P\mathbf{X}^{(k-1)} + \mathbf{g}, \tag{3.8}$$

параметр τ надо подобрать так, чтобы по какой-то из норм было выполнено условие $\|P\| \leqslant q < 1$.

Далее будем предполагать, что матрица исходной системы симметрична и положительно определена (т.е. $A^{\mathrm{T}} = A$ и $A > 0$) и что известны границы спектра матрицы A (минимальное и максимальное собственные значения). При этих предположениях мы не только определим диапазон значений τ, гарантирующих сходимость, но и найдем оптимальное τ, при котором величина погрешности приближений убывает с номером приближения наиболее быстро.

Итак, замечая, что

$$\mathbf{X}^* \equiv P\mathbf{X}^* + \mathbf{g}, \tag{3.9}$$

$\mathbf{X}^*$ — точное решение (3.1), и вводя в рассмотрение вектор ошибки k-го приближения $\mathbf{r}^{(k)} = \mathbf{X}^{(k)} - \mathbf{X}^*$, получим, вычитая (3.9) из (3.8):

$$\mathbf{r}^{(k)} = P\mathbf{r}^{(k-1)}. \tag{3.10}$$

Так как матрица A — симметричная, то существует ортонормированный базис из собственных векторов $\{\mathbf{e}_i, i = 1, 2, \ldots, n\}$, таких что $A\mathbf{e}_i = \lambda_i\mathbf{e}_i$, λ_i — соответствующие собственные значения. Причем в силу положительной определенности матрицы A $\min\limits_i \lambda_i > 0$. Будем далее полагать, что $\{\lambda_i\}$ упорядочены по возрастанию и обозначим $\overline{\lambda} = \min\limits_i \lambda_i$ и $\Lambda = \max\limits_i \lambda_i$.

Пусть $\mathbf{r}^{(k-1)} = \sum\limits_{i=1}^{n} c_i\mathbf{e}_i$ — разложение по элементам базиса вектора ошибки $(k-1)$-го приближения. (Пользуясь евклидовой метрикой, имеем $\|\mathbf{r}^{(k-1)}\|_2^2 = (\mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)}) = \sum c_i^2$.)

Подставляя это разложение в (3.10), получим

$$\mathbf{r}^{(k)} = P\left(\sum_i c_i\mathbf{e}_i\right) = \sum_i c_i(P\mathbf{e}_i) = \sum_i c_i(E - \tau A)\mathbf{e}_i =$$

$$= \sum_i c_i(1 - \tau\lambda_i)\mathbf{e}_i = \sum_i c_i\mu_i\mathbf{e}_i. \tag{3.11}$$

Как следует из цепочки соотношений (3.11), μ_i – i-е собственное значение матрицы P; оно связано с i-м собственным значением матрицы A равенством

$$\mu_i = 1 - \tau\lambda_i.$$

Далее из (3.11) следует, что

$$\|\mathbf{r}^{(k)}\|_2^2 = \sum_i c_i^2 \mu_i^2 \leqslant \max_i \mu_i^2 \|\mathbf{r}^{(k-1)}\|_2^2. \tag{3.12}$$

Таким образом, если $\max\limits_i |\mu_i| \leqslant q < 1$, то погрешность будет убывать с номером приближения как член геометрической прогрессии со знаменателем q.

Замечание 1. Мы могли бы и сразу (без выкладок, связанных с привлечением разложения вектора ошибок по базису) сформулировать этот вывод, заметив, что P — симметричная матрица ($P^{\mathrm{T}} = P$), и вспомнив, что норма матрицы P, подчиненная евклидовой норме вектора, определяется в этом случае как $\|P\|_2 = \max\limits_i |\mu_i|$. ▲

Замечание 2. Проделанные выкладки по существу представляют собой доказательство теоремы 2 о сходимости итерационного процесса для частного случая $A^{\mathrm{T}} = A > 0$. В самом деле, достаточность следует из (3.12). Необходимость вытекает из следующего рассуждения.

Пусть $\max\limits_i \mu_i^2$ достигается при $i = i_0$ и $|\mu_{i_0}| \geqslant 1$. Пусть далее вектор ошибки нулевого приближения имеет лишь одну компоненту $\mathbf{r}^{(0)} = c_{i_0}\mathbf{e}_{i_0}$. Очевидно, тогда $\mathbf{r}^{(k)} = c_{i_0}\mu_{i_0}^k \mathbf{e}_{i_0}$; следовательно, $\|\mathbf{r}^{(k)}\| = c_{i_0}|\mu_{i_0}|^k \geqslant c_{i_0}$ и сходимость отсутствует. ▲

Оптимизация параметра. Чтобы разобраться, при каких значениях параметра τ будут выполнены найденные ограничения ($\max\limits_i |\mu_i| \leqslant q < 1$), обеспечивающие сходимость последовательных приближений, обратимся к рис. 3.1, на котором иллюстрируется расположение собственных чисел λ_i и μ_i в плоскости (λ, μ).

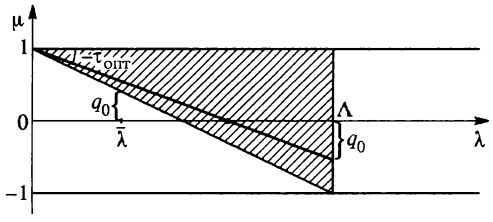


Рис. 3.1.

Очевидно, собственные числа $\mu_i = 1 - \tau\lambda_i$ лежат на прямых $\mu = 1 - \tau\lambda$ с тангенсом угла наклона $-\tau$. Из рисунка видно, что условие

$|\mu_i| < 1$ выполняется при всех i (т. е. при всех $\overline{\lambda} \leqslant \lambda_i \leqslant \Lambda$), когда соответствующие прямые лежат внутри заштрихованного сектора, т. е. когда

$$0 < \tau < \frac{2}{\Lambda}. \tag{3.13}$$

Из разложения (3.11) следует, что величину $|\mu_i|$ можно трактовать как коэффициент «подавления» составляющей вектора ошибки вдоль орта $\mathbf{e}_i$ при переходе от одного приближения к следующему. Из рис. 3.1 видно, что $\max\limits_i \mu_i$ достигается либо при $\lambda_i = \overline{\lambda}$, либо при $\lambda_i = \Lambda$; соответствующая компонента ошибки при увеличении числа итераций убывает наиболее медленно.

Оптимальным среди всех возможных значений (3.13) является $\tau = \tau_{\text{опт}}$, при котором достигается

$$\min_{0<\tau<\frac{2}{\Lambda}} \{\max_{\overline{\lambda}<\lambda_i<\Lambda} |1 - \tau\lambda_i|\}.$$

Замечание. Это вариант так называемой задачи о минимаксе, одной из основных задач теории оптимального управления. ▲

Из геометрических соображений (см. рис. 3.1) ясно, что решение этой задачи доставляет прямая, которая наименее уклоняется от нуля на отрезке $[\overline{\lambda}, \Lambda]$, т. е. прямая, проходящая через середину этого отрезка. Соответствующее значение $\tau_{\text{опт}} = 2/(\overline{\lambda} + \Lambda)$, а минимальный (по модулю) коэффициент подавления погрешности достигается одновременно при $\lambda = \overline{\lambda}$ и $\lambda = \Lambda$ и равен

$$q_0 = \min_{0<\tau<\frac{2}{\Lambda}} \max_{[\overline{\lambda}, \Lambda]} |1 - \tau\lambda_i| = |1 - \tau_{\text{опт}}\overline{\lambda}| =$$

$$= |1 - \tau_{\text{опт}}\Lambda| = 1 - \frac{2\overline{\lambda}}{\overline{\lambda} + \Lambda} = \frac{\Lambda - \overline{\lambda}}{\Lambda + \overline{\lambda}}.$$

Именно величина q_0 определяет реальный темп убывания погрешности с номером приближения в рамках оптимального однопараметрического итерационного процесса. Имеет место оценка

$$\|\mathbf{r}^{(k)}\| \leqslant q_0^k \|\mathbf{r}^{(0)}\|. \tag{3.14}$$

Пусть $\eta = \overline{\lambda}/\Lambda$, тогда $q_0 = (1 - \eta)/(1 + \eta)$.

Замечание. В данном случае величина η обратно пропорциональна числу обусловленности матрицы A. В самом деле, $\mu_A = \|A^{-1}\| \cdot \|A\|$. При сделанных предположениях относительно матри-

цы A ($A^{\mathrm{T}} = A > 0$) подчиненная евклидовой метрике векторного пространства спектральная норма для матриц приводит к $\|A\|_2 = \Lambda$ и $\|A^{-1}\|_2 = 1/\bar{\lambda}$, т. е. $\mu_A = \Lambda/\bar{\lambda}$. ▲

Используя (3.14), получаем априорную оценку числа приближений, гарантирующих достижение заданной точности ε:

$$k \geqslant k_0 = \frac{\ln \dfrac{\varepsilon}{\|\mathbf{r}^{(0)}\|}}{\ln q_0}.$$

Если $\eta \ll 1$ (число обусловленности велико), то

$$\ln q_0 = \ln \frac{1-\eta}{1+\eta} = \ln(1-\eta) - \ln(1+\eta) \approx -2\eta$$

и

$$k_0 \approx \frac{1}{2\eta} \ln \frac{\|\mathbf{r}^{(0)}\|}{\varepsilon}. \tag{3.15}$$

З а м е ч а н и е. Здесь и далее приводятся некоторые оценки для систем с большим числом обусловленности. Это не случайно. Дело в том, что с такими системами приходится иметь дело довольно часто при численном решении уравнений с частными производными, как мы увидим в Лекциях 11 и 12. ▲

П р и м е р. Пусть надо решить систему из $n = 10^4$ линейных уравнений с симметричной положительно определенной матрицей. Метод Гаусса требует в этом случае выполнения порядка $n^3 \sim 10^{12}$ элементарных операций. В итерационных методах для перехода от одного приближения к следующему необходимо выполнить порядка n^2 операций. Таким образом, объем вычислений, который сопряжен с методом итераций с одним оптимальным параметром ($k_0 n^2$), согласно (3.15) порядка $\dfrac{1}{\eta} \ln \dfrac{1}{\varepsilon} \cdot 10^8$, и при не слишком больших числах обусловленности этот метод эффективнее метода Гаусса.

ДОПОЛНЕНИЯ К ЛЕКЦИИ 3

Возможные обобщения. Подробный анализ однопараметрического метода итераций приоткрывает (в методическом плане) пути построения более эффективных итерационных процессов. Например, если при переходе от одного приближения к другому использовать различные итерационные параметры, то, оказывается, можно найти такую

их последовательность, что скорость сходимости существенно возрастает сравнительно с однопараметрическим методом. Например, в m-параметрическом (или, как иногда говорят, m-шаговом) итерационном процессе последовательные приближения вычисляются по формулам

$$\mathbf{X}^{(k)} = (E - \tau_k A)\mathbf{X}^{(k-1)} + \tau_k \mathbf{f}, \quad k = 1, 2, \ldots, m$$

(Затем m-е приближение принимается за нулевое, цикл повторяется и т. д., до сходимости с требуемой точностью.)

Последовательные ошибки (в пределах одного цикла) $\mathbf{r}^{(k)} = \mathbf{X}^{(k)} - \mathbf{X}^*$ ($\mathbf{X}^*$ — точное решение исходной системы) в этом случае связаны между собой соотношением

$$\mathbf{r}^{(m)} = (E - \tau_m A)\mathbf{r}^{(m-1)} =$$

$$= (E - \tau_m A)(E - \tau_{m-1} A)\mathbf{r}^{(m-2)} = \ldots = \left[\prod_{k=1}^{m} (E - \tau_k A)\right]\mathbf{r}^{(0)};$$

$$\|\mathbf{r}^{(m)}\| \leqslant \|S_m\| \cdot \|\mathbf{r}^{(0)}\|, \quad \text{где} \quad S_m = \prod_{k=1}^{m} (E - \tau_k A).$$

В предположении, что $A^{\mathrm{T}} = A > 0$, имеет место равенство

$$\|S_m\| = \max_i \left|\prod_{k=1}^{m} (1 - \tau_k \lambda_i)\right|,$$

где λ_i — собственные значения матрицы A.

Далее поиск оптимальной последовательности параметров $\{\tau_k^{\text{опт}}\}$ сводится к отысканию полинома m-й степени $\mu(\lambda)$:

$$\mu = \prod_{k=1}^{m} (1 - \tau_k \lambda),$$

наименее уклоняющегося от нуля при $\bar{\lambda} \leqslant \lambda \leqslant \Lambda$. Решение этой задачи выражается через корни полинома Чебышёва (см. Приложения I, III). При этом уменьшение ошибки за m шагов итерационного цикла определяется величиной

$$q_m^{\text{опт}} = \|S_m^{\text{опт}}\| = \min_{\{\tau_k\}} \left\{ \max_{\bar{\lambda} \leqslant \lambda \leqslant \Lambda} \prod_{k=1}^{m} \left|(1 - \tau_k \lambda_i^{\text{опт}})\right| \right\}.$$

Например, для плохо обусловленных систем ($\eta \ll 1$) можно получить

$$q_m^{\text{опт}} \sim 2\left(\frac{1 - \sqrt{\eta}}{1 + \sqrt{\eta}}\right)^m.$$

Оценивая, за сколько шагов ошибка будет меньше заданной величины ε, получим

$$m \approx \frac{1}{2\sqrt{\eta}} \ln \frac{\|\mathbf{r}^{(0)}\|}{\varepsilon}.$$

Сравнительно с оценкой (3.15) для одношагового процесса с оптимальным параметром это существенно лучше. Применительно к рассмотренному в конце лекции примеру данный метод требует вычисления порядка $\frac{1}{\sqrt{\eta}} \ln \frac{1}{\varepsilon} \cdot 10^8$ операций. Таким образом, даже для системы с числом обусловленности $\mu_A \sim 10^4$ ($\eta \sim 10^{-4}$) он будет намного эффективнее метода Гаусса (не говоря уже о том, что метод Гаусса в такой ситуации, согласно оценкам, приведенным в дополнении к Лекции 2, может оказаться неприменимым при использовании 32-разрядных машин из-за практической потери точности).

О погрешности округлений при использовании итерационных методов. Мы будем говорить здесь о влиянии ошибок округления на результаты вычислений. Что касается ошибок входных данных, то возможное их влияние на искомое решение линейной системы было исследовано в дополнении к предыдущей лекции *независимо от метода*, который применяется для расчета.

Анализ влияния вычислительных погрешностей на последовательные приближения опирается на тот факт, что суммарный эффект ошибок округлений при выполнении одного шага итерационного процесса

$$\mathbf{X}^{(k)} = P\mathbf{X}^{(k-1)} + \mathbf{g}$$

можно трактовать как возмущение правой части (вектора $\mathbf{g}$).

Итак, реальный вычислительный процесс запишем в виде

$$\widetilde{\mathbf{X}}^{(k)} = P\widetilde{\mathbf{X}}^{(k-1)} + \mathbf{g} + \delta\mathbf{g}^{(k)}.$$

Тогда для разности между реально вычисляемыми приближениями $\widetilde{\mathbf{X}}^{(k)}$ и «идеальными» $\mathbf{X}^{(k)}$, которые получали бы, если бы вычисления проводились без ошибок округления, будем иметь

$$\widetilde{\mathbf{X}}^{(k)} - \mathbf{X}^{(k)} = P(\widetilde{\mathbf{X}}^{(k-1)} - \mathbf{X}^{(k-1)}) + \delta\mathbf{g}^{(k)}.$$

Считая, что в какой-то норме выполнено достаточное условие сходимости $\|P\| \leqslant q < 1$, и оценивая разности по соответствующим со-

гласованным нормам, получим

$$\|\widetilde{\mathbf{X}}^{(k)} - \mathbf{X}^{(k)}\| \leqslant q\|\widetilde{\mathbf{X}}^{(k-1)} - \mathbf{X}^{(k-1)}\| + \|\delta\mathbf{g}^{(k)}\| \leqslant$$
$$\leqslant q^2\|\widetilde{\mathbf{X}}^{(k-2)} - \mathbf{X}^{(k-2)}\| + q\|\delta\mathbf{g}^{(k-1)}\| + \|\delta\mathbf{g}^{(k)}\| \leqslant \ldots \leqslant$$
$$\leqslant q^k\|\widetilde{\mathbf{X}}^{(0)} - \mathbf{X}^{(0)}\| + (\max_k \|\delta\mathbf{g}^{(k)}\|)(1 + q + \ldots + q^{k-1}).$$

Естественно считать, что $\|\widetilde{\mathbf{X}}^{(0)} - \mathbf{X}^{(0)}\| = 0$ (так как начальное приближение задается, а не вычисляется).

Обозначая максимально возможную суммарную (в пределах одной итерации) неустранимую погрешность $\delta = \max_k \|\delta\mathbf{g}^{(k)}\|$ и заменяя сумму геометрической прогрессии по известной формуле, получим окончательную оценку отличия реально вычисленного n-го приближения от «идеального»

$$\|\widetilde{\mathbf{X}}^{(k)} - \mathbf{X}^{(k)}\| \leqslant \delta\frac{q^k - 1}{q - 1} < \frac{\delta}{1 - q}.$$

Одновременно это мера максимально достижимой точности в рамках итерационного процесса. Очевидно, бессмысленно задаваться в качестве допустимого уровня погрешности ε величиной, существенно меньшей, нежели $\delta/(1 - q)$.

Разумеется, последнее соображение можно (и следует) принимать в расчет, только если известна хотя бы грубая оценка величины δ.

Представление о порядке δ можно получить на основе следующих рассуждений. Из статистического анализа погрешностей округления при большом объеме вычислений известно [9, с. 23], что выполнение N однотипных арифметических операций приводит к суммарной ошибке округлений порядка $N^{1/2}\Delta_0$, где Δ_0 — погрешность округления при выполнении одной элементарной операции.

Далее, в рассматриваемых итерационных методах переход от одного приближения к следующему сводится к умножению матрицы на вектор. При этом для вычисления каждой компоненты результата требуется выполнить порядка n умножений и сложений, где n — порядок системы. Следовательно, можно считать, что погрешность вычисления каждой компоненты нового приближения, а тем самым и δ суть величины порядка $n^{1/2}\Delta_0$.

Мы получили достаточно благоприятную оценку для неустранимых погрешностей вычислений при использовании итерационных методов. Даже если число обусловленности, как в примере, рассмотренном в предыдущем разделе, велико: $\mu_A \sim 10^4$, $\eta \sim 10^{-4}$,

$\frac{\delta}{1-q} \sim \frac{\delta}{\eta} \sim 10^4\,\delta$, то полной потери точности не происходит (как в методах Гаусса) по крайней мере при $\delta \approx 10^{-6}$ и меньших, что, согласно изложенным выше соображениям, вполне реально для суммарной погрешности округлений за один шаг итерационного цикла даже для систем очень высокого порядка (вплоть до 10^4) при стандартной точности машинных вычислений ($\Delta_0 \sim 10^{-8}$).

З а м е ч а н и е. На практике итерационные методы часто применяются для решения систем с сильно разреженной матрицей (с числом ненулевых элементов порядка n). В этом случае $\delta \sim \Delta_0$, соответственно применительно к рассматриваемому примеру оценка неустранимой погрешности уменьшается на два порядка. ▲

О канонической записи итерационных процессов. Множество итерационных схем, сконструированных для решения различных линейных систем, укладывается в каноническую форму записи:

$$B_{k+1}\frac{\mathbf{X}^{(k+1)}-\mathbf{X}^{(k)}}{\tau_{k+1}} + A\mathbf{X}^{(k)} = \mathbf{f};$$

$B_k = E$ — явные итерационные процессы, $B_k \neq E$ — неявные; $B_k = B$, $\tau_k = \tau$ — стационарные процессы.

Мы рассмотрели здесь следующие частные случаи:

а) $B_k = E$, $\tau_k = \tau$ — однопараметрический метод;

б) $B_k = E, \tau_k \in \{\tau_k, k = 1, 2, \ldots, m\}$ — m-шаговый итерационный процесс;

в) $B_k = D$, $\tau_k = 1$ — метод Якоби;

г) $B_k = D + A_-$, $\tau_k = 1$ — метод Зейделя.

Более полные сведения об итерационных методах решения линейных систем уравнений можно найти в [1, с. 174–189], [2, с. 269–308], [3, т. 2, с. 44–75], [9, с. 153–155, 301–423], [10, с. 17–35, 420], [11, с. 96–129], [12, с. 82–126], [13].

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Доказать, что

а) $\max\limits_{i,j}|a_{ij}|$ не является нормой матрицы;

б) $\|A\|_E = \left(\sum\limits_{i,j} a_{ij}^2\right)^{1/2}$ является нормой матрицы;

в) $\|A\|_M = n\left(\max\limits_{i,j}|a_{ij}|\right)$ является нормой матрицы.

2. Доказать, что норма $\|A\|_M$ согласована с нормами $\|\mathbf{X}\|_{1,2,c}$.

3. Показать, что в условиях теоремы 1 справедлива оценка

$$\|\mathbf{X}^{(n)} - \mathbf{X}^*\| \leqslant \frac{q^n}{1-q}\|\mathbf{X}^{(1)} - \mathbf{X}^{(0)}\|.$$

4. Дана система уравнений

$$\frac{1}{h^2}(U_{i,j+1} + U_{i,j-1} + U_{i+1,j} + U_{i-1,j} - 4U_{i,j}) = f_{i,j}, \quad i,j = \overline{1, N-1},$$
$$U_{0,j} = U_{N,j} = U_{i,0} = U_{i,N} = 0, \quad i,j = \overline{0, N}.$$

а) Найти число арифметических операций, необходимых для выполнения одной итерации метода Якоби при решении этой системы.

б) То же для метода Зейделя.

5. Для решения системы

$$2x_1 + x_2 = f_1,$$
$$x_1 + 2x_2 = f_2$$

предполагается применить метод Якоби.

а) Оценить количество приближений, уменьшающих ошибку начального приближения в 10^3 раз.

б) То же для однопараметрического метода итераций с $\tau = 0.3$.

в) То же для оптимального значения итерационного параметра (найти требуемое оптимальное значение).

ЛЕКЦИЯ 4

ЧИСЛЕННОЕ РЕШЕНИЕ СИСТЕМ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

Формулировка задачи. Метод Ньютона. Метод простых итераций. Теорема о достаточном условии сходимости. Скорость сходимости. Варианты итерационных схем. Каноническая форма записи итерационных схем. Влияние погрешностей округления.

При обсуждении методов вычисления решения нелинейных систем мы, в основном, будем следовать плану, реализованному в Лекции 1, где речь шла об отыскании корней нелинейных уравнений.

Итак, рассматривается система

$$f_i(x_1, x_2, \ldots, x_n) = 0, \quad i = 1, 2, \ldots, n, \tag{4.1}$$

или в векторной записи

$$\mathbf{F}(\mathbf{X}) = 0, \tag{4.1'}$$

где $\mathbf{X}$ — вектор неизвестных величин, а $\mathbf{F}(\mathbf{X})$ — вектор-функция, определяющая структуру уравнений системы:

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}, \qquad \mathbf{F}(\mathbf{X}) = \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ f_2(x_1, \ldots, x_n) \\ \ldots\ldots\ldots \\ f_n(x_1, \ldots, x_n) \end{pmatrix}.$$

Будем считать, что система (4.1) имеет по крайней мере одно решение и что интересующее нас решение $\mathbf{X}^*$ принадлежит известной окрестности U_*, выявленной на стадии локализации решения.

Замечание. Как и в случае одного уравнения, локализация решения осуществляется либо на основе физических соображений (если задача имеет физическое содержание), либо с привлечением методов математического анализа (например, для системы двух уравнений можно приближенно оценить месторасположение корней, анализируя на плоскости (x_1, x_2) поведение кривых, задаваемых уравнениями $f_1(x_1, x_2) = 0$, $f_2(x_1, x_2) = 0$). ▲

Метод Ньютона как метод линеаризации исходной задачи. Пусть $\mathbf{X}^{(0)} \in U_*$ — выбранное начальное приближение. В окрестно-

сти $\mathbf{X}^{(0)}$ линеаризуем уравнения (4.1), представив каждую из функций левой части (4.1) в виде отрезка ряда Тейлора для функции n переменных:

$$f_i(x_1, x_2, \ldots, x_n) \approx f_i(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}) + \sum_{j=1}^{n} \left.\frac{\partial f_i}{\partial x_j}\right|^{(0)} (x_j - x_j^{(0)}) =$$

$$= f_i^{(0)} + \sum_{j=1}^{n} \left.\frac{\partial f_i}{\partial x_j}\right|^{(0)} (x_j - x_j^{(0)}) = 0, \quad i = 1, 2, \ldots, n. \qquad (4.2)$$

Вместо задачи (4.1) решим линейную систему (4.2), применяя известные методы (например, метод Гаусса). Найденное решение будем считать первым приближением и т. д.

Существует теорема, которая устанавливает, что при некоторых требованиях, предъявляемых к $\dfrac{\partial f_i}{\partial x_j}$, $\dfrac{\partial^2 f_i}{\partial x_l \partial x_m}$ в рассматриваемой окрестности U_*, последовательность вычисляемых таким образом приближений сходится к решению $\mathbf{X}^*$ ([2], [3, т. 2]).

Система уравнений для перехода от $(k-1)$-го приближения к k-му в рамках данного метода согласно изложенному выше имеет вид

$$\sum_{j=1}^{n} \left.\frac{\partial f_i}{\partial x_j}\right|^{(k-1)} (x_j^{(k)} - x_j^{(k-1)}) = -f_i^{(k-1)}, \quad i = 1, 2, \ldots, n, \qquad (4.3)$$

где $f_i^{(k-1)} = f_i(x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_n^{(k-1)})$.

Систему невысокого порядка ($n = 2, 3$) можно записать в разрешенном виде относительно $x^{(k)}$, т. е. в виде расчетных формул. Например, для $n = 2$

$$x_1^{(k)} = x_1^{(k-1)} - \frac{f_1^{(k-1)} \left.\dfrac{\partial f_2}{\partial x_2}\right|^{(k-1)} - f_2^{(k-1)} \left.\dfrac{\partial f_1}{\partial x_2}\right|^{(k-1)}}{\left.\left(\dfrac{\partial f_1}{\partial x_1}\dfrac{\partial f_2}{\partial x_2} - \dfrac{\partial f_2}{\partial x_1}\dfrac{\partial f_1}{\partial x_2}\right)\right|^{(k-1)}},$$

$$x_2^{(k)} = x_2^{(k-1)} - \frac{f_2^{(k-1)} \left.\dfrac{\partial f_1}{\partial x_1}\right|^{(k-1)} - f_1^{(k-1)} \left.\dfrac{\partial f_2}{\partial x_1}\right|^{(k-1)}}{\left.\left(\dfrac{\partial f_1}{\partial x_1}\dfrac{\partial f_2}{\partial x_2} - \dfrac{\partial f_2}{\partial x_1}\dfrac{\partial f_1}{\partial x_2}\right)\right|^{(k-1)}} \qquad (4.4)$$

Для систем более высокого порядка ($n \geqslant 4$) реализация перехода от одного приближения к следующему, как уже отмечалось, состоит в решении известными методами системы линейных уравнений (4.3).

Если ввести в рассмотрение матрицу Якоби для вектор-функции $\mathbf{F}(\mathbf{X})$:

$$D = \frac{D\mathbf{F}}{D\mathbf{X}} = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{pmatrix}, \tag{4.5}$$

то систему (4.3) можно записать компактным образом в векторном виде

$$D^{(k-1)}\left(\mathbf{X}^{(k)} - \mathbf{X}^{(k-1)}\right) = -\mathbf{F}^{(k-1)}. \tag{4.6}$$

Формально разрешая (4.6) относительно $\mathbf{X}^{(k)}$, приходим к записи метода Ньютона для системы нелинейных уравнений, схожей с формулой Ньютона для скалярного случая:

$$\mathbf{X}^{(k)} = \mathbf{X}^{(k-1)} - \left[D^{(k-1)}\right]^{-1}\mathbf{F}^{(k-1)}. \tag{4.7}$$

Очевидно, что (4.4) есть не что иное, как покомпонентная запись (4.7) для $n = 2$.

Метод простых итераций. Допустим, что уравнения (4.1) каким-то образом приведены к виду

$$x_i = \varphi_i(x_1, x_2, \ldots, x_n), \quad i = 1, 2, \ldots, n, \tag{4.8}$$

или в векторной форме

$$\mathbf{X} = \mathbf{\Phi}(\mathbf{X}) \tag{4.8′}$$

с вектор-функцией $\mathbf{\Phi}(\mathbf{X}) = \begin{pmatrix} \varphi_1(x_1, \ldots, x_n) \\ \varphi_2(x_1, \ldots, x_n) \\ \cdots\cdots\cdots\cdots \\ \varphi_n(x_1, \ldots, x_n) \end{pmatrix}$.

Опираясь на (4.8′), можно вычислять последовательность

$$\mathbf{X}^{(k)} = \mathbf{\Phi}(\mathbf{X}^{(k-1)}) \tag{4.9}$$

(при заданном $\mathbf{X}^{(0)}$). При некоторых условиях эта последовательность может сходиться к решению (4.1). Эти условия формулируются в следующей теореме.

Теорема. *Пусть* $\boldsymbol{X}^*$ *— искомое решение задачи* (4.1) *или, что то же, задачи* (4.8′). *Если в окрестности* $U_* = \{\|\boldsymbol{X} - \boldsymbol{X}^*\| \leqslant r\}$

вектор-функция $\mathbf{\Phi}(\boldsymbol{X})$ *удовлетворяет условию Коши–Липшица, т. е. для любых* $\boldsymbol{X}', \boldsymbol{X}'' \in U_*$ *имеет место*

$$\|\mathbf{\Phi}(\mathbf{X}') - \mathbf{\Phi}(\mathbf{X}'')\| \leqslant q\|\mathbf{X}' - \mathbf{X}''\|, \quad q = \text{const}, \tag{4.10}$$

и если при этом $q < 1$, *то последовательность* $\boldsymbol{X}^{(k)}$, *вычисляемых согласно (4.9) с* $\boldsymbol{X}^{(0)} \in U_*$, *сходится к решению, т. е.*

$$\|\mathbf{X}^{(k)} - \mathbf{X}^*\| \underset{k\to\infty}{\longrightarrow} 0.$$

При этом имеет место оценка

$$\|\mathbf{X}^{(k)} - \mathbf{X}^*\| \leqslant q^k\|\mathbf{X}^{(0)} - \mathbf{X}^*\|. \tag{4.11}$$

Доказательство дословно совпадает с доказательством соответствующего утверждения в Лекции 1 (единственное отличие: здесь используется оценка значений по норме, а там — по абсолютной величине).

В случае одного нелинейного уравнения (в скалярном случае) вместо условия Коши–Липшица в качестве достаточного критерия сходимости мы проверяли в рассматриваемой окрестности решения выполнение неравенства $|\varphi'(x)| < 1$ (см. Лекцию 1). Сейчас мы обобщим этот критерий применительно к данной ситуации.

Запишем выражение для разности между компонентами k-го приближения, вычисленного по методу (4.9), и точным (искомым) решением рассматриваемой задачи:

$$x_i^{(k)} - x_i^* = \varphi_i(x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_n^{(k-1)}) - \varphi_i(x_1^*, x_2^*, \ldots, x_n^*) =$$

(по теореме о среднем для функции многих переменных)

$$= \sum_{j=1}^{n} \left.\frac{\partial\varphi_i}{\partial x_j}\right|_{\vec{\xi}_i \in U_*} (x_j^{(k-1)} - x_j^*), \quad i = 1, 2, \ldots, n.$$

Введем в рассмотрение матрицу Якоби вектор-функции $\mathbf{\Phi}(\mathbf{X})$:

$$M_\varphi = \frac{D\mathbf{\Phi}}{D\mathbf{X}} = \begin{pmatrix} \dfrac{\partial\varphi_1}{\partial x_1} & \cdots & \dfrac{\partial\varphi_1}{\partial x_n} \\ \cdots\cdots\cdots\cdots\cdots \\ \dfrac{\partial\varphi_n}{\partial x_1} & \cdots & \dfrac{\partial\varphi_n}{\partial x_n} \end{pmatrix}.$$

С ее помощью совокупность выписанных соотношений можно записать в векторной форме:

$$\mathbf{X}^{(k)} - \mathbf{X}^* = M_\varphi(\mathbf{X}^{(k-1)} - \mathbf{X}^*) \tag{4.12}$$

(имея в виду, что элементы различных строк матрицы вычисляются, вообще говоря, в различных точках U_*).

Пусть, далее, M — «мажорирующая» матрица с элементами

$$m_{ij} = \max_{U_*} |\partial\varphi_i/\partial x_j|, \quad \text{так что} \quad \|M_\varphi\| \leqslant \|M\|.$$

Тогда

$$\|\mathbf{X}^{(k)} - \mathbf{X}^*\| \leqslant \|M_\varphi\| \cdot \|\mathbf{X}^{(k-1)} - \mathbf{X}^*\| \leqslant \|M\| \cdot \|\mathbf{X}^{(k-1)} - \mathbf{X}^*\|.$$

И если

$$\|M\| \leqslant q < 1, \tag{4.13}$$

то $\|\mathbf{X}^{(k)} - \mathbf{X}^*\| \leqslant q\|\mathbf{X}^{(k-1)} - \mathbf{X}^*\|$ и последовательные приближения сходятся к решению X^*.

Именно условие (4.13) привлекается при анализе сходимости конкретных итерационных схем.

Пример. Рассмотрим систему

$$x^3 + y^2 - 6x + 3 = 0,$$
$$x^3 - y^2 - 6y + 2 = 0.$$

Записав ее в виде $x = \frac{1}{6}(x^3 + y^2 + 3)$, $y = \frac{1}{6}(x^3 - y^2 + 2)$, исследуем на сходимость соответствующий метод итераций для корня, принадлежащего области $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$.

В данном случае

$$M_\varphi = \begin{pmatrix} \frac{x^2}{2} & \frac{y}{3} \\ \frac{x^2}{2} & -\frac{y}{3} \end{pmatrix}, \qquad M = \begin{pmatrix} \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix},$$

$$\|M\|_c = \frac{1}{2} + \frac{1}{3} = \frac{5}{6} < 1.$$

Следовательно, согласно (4.11) можно рассчитывать на сходимость, так что

$$\|\mathbf{X}^{(k)} - \mathbf{X}^*\|_c = \max\{|x_k - x_*|, |y_k - y_*|\} \leqslant \left(\frac{5}{6}\right)^k \cdot \frac{1}{2}.$$

(В предположении, что $x_0 = y_0 = 0.5$, имеем $\|\mathbf{X}^{(0)} - \mathbf{X}^{(*}\|_c \leqslant \frac{1}{2}$.)

Замечание. В последнем абзаце сказано, что «можно рассчитывать на сходимость». Дело вот в чем. В теореме рассматривалась окрестность U_* с центром в точке $\mathbf{X}^*$. Реально, как в этом примере, условие сходимости проверяется в некоторой окрестности, относительно которой лишь известно, что искомое решение принадлежит ей. Строго говоря, мы не гарантированы в этом случае от того, что какое-то приближение на начальной стадии итераций не выйдет за пределы окрестности, где выполнено условие (4.13). (Иллюстрация подобной ситуации применительно к одному нелинейному уравнению приведена на рис. 1.6 в дополнениях к Лекции 1.) В этом случае необходимо вернуться к этапу локализации решения, чтобы уточнить («сузить») рассматриваемую окрестность. ▲

Варианты итерационных схем. Перепишем систему (4.1') в виде

$$\mathbf{X} = \mathbf{X} + \tau\mathbf{F}(\mathbf{X}), \tag{4.14}$$

где τ — некая константа.

Требуя, чтобы в рассматриваемой окрестности решения U_* выполнялось достаточное условие сходимости,

$$\|M_\varphi\| = \left\|\frac{D\mathbf{\Phi}}{D\mathbf{X}}\right\| = \left\|E + \tau\frac{D\mathbf{F}}{D\mathbf{X}}\right\| < 1,$$

подберем параметр τ. Это *метод релаксации* с одним параметром.

Замечание. Велика вероятность, что последнее неравенство после конкретизации выбора нормы сведется к несовместной (для одного τ) последовательности неравенств. Например, если использовать равномерную метрику, то необходимо удовлетворить n неравенствам для одного параметра

$$\left|1 + \tau\frac{\partial f_i}{\partial x_i}\right| + \sum_{j\neq i}^{n}\left|\tau\frac{\partial f_i}{\partial x_j}\right|_{U_*} < 1, \quad i = 1, 2, \ldots, n. \tag{4.15}$$

Обобщение этого подхода состоит в том, что вводится в рассмотрение не один, а множество свободных параметров. В общем случае n^2 элементов матрицы $T = \{\tau_{ij}\}$:

$$\mathbf{X} = \mathbf{X} + T\mathbf{F}(\mathbf{X}) = \mathbf{\Phi}(\mathbf{X}).$$

Тогда $M_\varphi = E + T\dfrac{D\mathbf{F}}{D\mathbf{X}} = E + TD$ и условие сходимости $\|M_\varphi\| < 1$ приводится к n неравенствам типа (4.15) для n^2 свободных параметров. Это задача, в которой можно «погрязнуть» и не выбраться из нее.

Впрочем, если не требовать постоянства параметров τ_{ij}, а считать их функциями $\mathbf{X}$ и выбрать последние так, что $T(\mathbf{X}) = -\left(\dfrac{D\mathbf{F}}{D\mathbf{X}}\right)^{-1}$, то мы придем к методу Ньютона (см. (4.7)). Подметив это обстоятельство, можно набор из n^2 постоянных параметров определить следующим образом:

$$T = -\left[\left.\frac{D\mathbf{F}}{D\mathbf{X}}\right|_{\mathbf{X}=\mathbf{X}^{(0)}}\right]^{-1},$$

т. е. один раз обратить матрицу $D = \frac{D\mathbf{F}}{D\mathbf{X}}$ в точке начального приближения и далее использовать ее при вычислении каждого следующего. Это так называемый *огрубленный метод Ньютона* для системы нелинейных уравнений.

В рамках поиска подходящей схемы метода релаксаций разумным подходом представляется введение n свободных параметров в виде элементов диагональной матрицы

$$T = \begin{pmatrix} \tau_1 & & & 0 \\ & \tau_2 & & \\ & & \ddots & \\ 0 & & & \tau_n \end{pmatrix}.$$

Тогда условие сходимости $\left\|E + T\frac{D\mathbf{F}}{D\mathbf{X}}\right\| < 1$ при использовании нормы $\|\cdot\|_c$ сводится к системе n неравенств

$$\left|1 + \tau_i \frac{\partial f_i}{\partial x_i}\right| + \sum_{j \neq i}^{n} \left|\tau_i \frac{\partial f_i}{\partial x_j}\right| \Bigg|_{U_*} < 1, \quad i = 1, 2, \ldots, n$$

для n параметров τ_i.

Каноническая запись одношаговых итерационных процессов. Рассмотренные методы итераций можно записать в единообразной, канонической форме:

$$B_k \frac{\mathbf{X}^{(k)} - \mathbf{X}^{(k)(k-1)}}{\tau_k} + \mathbf{F}(\mathbf{X}^{(k-1)}) = 0;$$

$B_k^{(k)}$ — заданные матрицы, τ_k — заданные скалярные параметры.

Частные случаи:

$B \equiv E$, $\tau_k \equiv \tau$ — метод релаксаций с одним параметром;

$B_k = \begin{pmatrix} \frac{1}{\tau_1} & & & 0 \\ & \frac{1}{\tau_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\tau_n} \end{pmatrix}$, $\tau_k \equiv 1$ — n-параметрический метод релаксаций;

$B_k = \left.\frac{D\mathbf{F}}{D\mathbf{X}}\right|_{\mathbf{X}^{(k)}}$, $\tau_k \equiv 1$ — метод Ньютона;

$B_k = \left.\frac{D\mathbf{F}}{D\mathbf{X}}\right|_{\mathbf{X}^{(0)}}$, $\tau_k \equiv 1$ — «огрубленный» метод Ньютона.

Отметим также возможность использования зейделевских вариантов итерационных схем для решения системы (4.1).

Перепишем, например, (4.1) в виде соотношений, связывающих значения k-х и $(k-1)$-х компонент решения:

$$\begin{aligned} f_1(x_1^{(k)}, x_2^{(k-1)}, x_3^{(k-1)}, \ldots, x_n^{(k-1)}) &= 0, \\ f_2(x_1^{(k)}, x_2^{(k)}, x_3^{(k-1)}, \ldots, x_n^{(k-1)}) &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ f_i(x_1^{(k)}, \ldots, x_i^{(k)}, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)}) &= 0. \end{aligned}$$

Из первого уравнения находится $x_1^{(k)}$, из второго — $x_2^{(k)}$ (при уже известном значении $x_1^{(k)}$) и т. д. При такой организации вычислений каждое уравнение содержит лишь одну неизвестную, соответственно для расчета привлекаются методы из арсенала средств решения скалярных уравнений, обсуждавшихся в Лекции 1.

Аналогично систему (4.8) можно записать в виде

$$x_i^{(k)} = \varphi_i(x_1^{(k)}, x_2^{(k)}, \ldots, x_i^{(k)}, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)}), \quad i = 1, 2, \ldots, n$$

и для вычисления i-й компоненты «итерировать» фактически скалярное i-е соотношение (ибо $x_1^{(k)}, \ldots, x_{i-1}^{(k)}$ к этому моменту уже вычислены).

Зейделевскими эти варианты итерационных схем названы ввиду их очевидной аналогии с методом Зейделя для линейных систем уравнений (см. Лекцию 3).

С точки зрения программной организации подобные схемы будут представлять собой вложенные итерационные циклы:

— внутренние итерации для решения скалярных уравнений для каждой компоненты (пока, например, не выполнится условие

$|(x_i^{(k)})^{(s)} - (x_i^{(k)})^{(s-1)}| < \varepsilon$, где s — номер приближения для $x_i^{(k)}$ во внутреннем итерационном цикле);

— внешний итерационный цикл замыкается проверкой условия $\|\mathbf{X}^{(k)} - \mathbf{X}^{(k-1)}\| < \varepsilon$ после того, как вычислены все $x_i^{(k)}$ $(i = 1, 2, \ldots, n)$.

Замечание. Вопросы сходимости методов подобного типа мы здесь не рассматриваем. ▲

ДОПОЛНЕНИЕ К ЛЕКЦИИ 4

Влияние неустранимых погрешностей на вычисляемые приближения. Влияние погрешностей округления на вычисляемые приближения анализируется так же, как для итерационных схем решения систем линейных уравнений.

Пусть $\mathbf{X}^{(k)} = \mathbf{\Phi}(\mathbf{X}^{(k-1)})$ — «идеальный» итерационный процесс, при реализации которого не принимаются в расчет погрешности округлений при выполнении элементарных арифметических операций. $\widetilde{\mathbf{X}}^{(k)} = \mathbf{\Phi}(\widetilde{\mathbf{X}}^{(k-1)}) + \boldsymbol{\delta}^{(k)}$ — «реальный» вычислительный процесс, соответствующий рассматриваемому методу итераций. Здесь $\widetilde{\mathbf{X}}^{(k)}$ — реально вычисляемые приближения, $\boldsymbol{\delta}^{(k)}$ — совокупное влияние погрешностей округлений в пределах одного шага (связанного с переходом от $(k-1)$-го приближения к k-му).

Вычитая из второго соотношения первое, получим

$$\widetilde{\mathbf{X}}^{(k)} - \mathbf{X}^{(k)} = \mathbf{\Phi}(\widetilde{\mathbf{X}}^{(k-1)}) - \mathbf{\Phi}(\mathbf{X}^{(k-1)}) + \boldsymbol{\delta}^{(k)}.$$

Предполагая, что выполнено достаточное условие сходимости, оценим неустранимую погрешность k-го приближения по какой-либо норме:

$$\begin{aligned}
\|\widetilde{\mathbf{X}}^{(k)} - \mathbf{X}^{(k)}\| &\leqslant \|\mathbf{\Phi}(\widetilde{\mathbf{X}}^{(k-1)}) - \mathbf{\Phi}(\mathbf{X}^{(k-1)})\| + \|\boldsymbol{\delta}^{(k)}\| \leqslant \\
&\leqslant q\|\widetilde{\mathbf{X}}^{(k-1)} - \mathbf{X}^{(k-1)}\| + \|\boldsymbol{\delta}^{(k)}\| \leqslant \\
&\leqslant q(q\|\widetilde{\mathbf{X}}^{(k-2)} - \mathbf{X}^{(k-2)}\| + \|\boldsymbol{\delta}^{(k-1)}\|) + \|\boldsymbol{\delta}^{(k)}\| = \\
&= q^2\|\widetilde{\mathbf{X}}^{(k-2)} - \mathbf{X}^{(k-2)}\| + q\|\boldsymbol{\delta}^{(k-1)}\| + \|\boldsymbol{\delta}^{(k)}\| \leqslant \ldots \leqslant \\
&\leqslant q^k\|\widetilde{\mathbf{X}}^{(0)} - \mathbf{X}^{(0)}\| + q^{k-1}\|\boldsymbol{\delta}^{(1)}\| + q^{k-2}\|\boldsymbol{\delta}^{(2)}\| + \ldots + q\|\boldsymbol{\delta}^{(k-1)}\| + \|\boldsymbol{\delta}^{(k)}\|.
\end{aligned}$$

Очевидно, $\widetilde{\mathbf{X}}^{(k)}(0) = \mathbf{X}^{(k)}(0)$, так как начальное приближение не вычисляется, а задается.

Вводя в рассмотрение максимальную погрешность, накопленную за счет округлений на одном шаге итерационного процесса

$$\delta = \max_i \|\boldsymbol{\delta}^{(i)}\|,$$

получим окончательную оценку погрешности на k-м шаге

$$\|\widetilde{\mathbf{X}}^{(k)} - \mathbf{X}^{(k)}\| \leqslant \delta(q^{k-1} + q^{k-2} + \ldots + 1) = \frac{\delta(1-q^k)}{1-q} \leqslant \frac{\delta}{1-q},$$

откуда следует, что влияние погрешности округлений умеренно, если q не слишком близко к единице.

Дополнительные сведения о вопросах, рассмотренных в Лекции 4 и в дополнении к ней, можно найти в [1, с. 191–210], [2, с. 317–356], [5, с. 11–17], [7, с. 183–189, 193, 196], [9, с. 150–155], [12, с. 207–213].

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Используя идею метода секущих для одного уравнения, построить итерационный процесс для решения системы

$$\begin{aligned} f(x,y) &= 0,\\ g(x,y) &= 0. \end{aligned}$$

Проверить, выполнено ли условие сходимости заданных итерационных процессов для приведенных ниже систем (задачи 2–6) в окрестности решений (точнее, в указанных точках, принадлежащих упомянутым окрестностям).

2. $\begin{cases} x_{k+1} = \sqrt{2(x_k + y_k)}, \\ y_{k+1} = \sqrt[4]{1 - x_k^{4(k)}}, \end{cases}$ для системы $\begin{cases} x^4 + y^4 - 1 = 0, \\ x^2/2 - x - y = 0 \end{cases}$

в точках: а) $x_0 = 1$, $y_0 = -0.5$; б) $x_0 = -0.7$, $y_0 = 0.9$.

3. $\begin{cases} x_{k+1} = \sqrt[3]{23.1 + y_k}, \\ y_{k+1} = \sqrt{25.9 - x_k^{2(k)}}, \end{cases}$ для системы $\begin{cases} x^2 + y^2 = 25.9, \\ y - x^3 + 23.1 = 0 \end{cases}$

в точках: а) $x_0 = 3$, $y_0 = 4$; б) $x_0 = 2.6$, $y_0 = -4.3$.

4. $\begin{cases} x_{k+1} = \left(\dfrac{y_k^2 - 1.98}{2}\right)^{1/3}, \\ y_{k+1} = x_k + \dfrac{1.03}{x_k}, \end{cases}$ для системы $\begin{cases} xy - x^2 - 1.03 = 0, \\ 2x^3 - y^2 + 1.98 = 0 \end{cases}$

в точке $x_0 = 1$, $y_0 = 2$.

5. $\begin{cases} x_{k+1} = \left(\dfrac{1 + y_k^2}{2}\right)^{1/3}, \\ y_{k+1} = \left(\dfrac{y_k + 4}{x_k}\right)^{1/3}, \end{cases}$ для системы $\begin{cases} 2x^3 - y^2 - 1 = 0, \\ xy^3 - y - 4 = 0 \end{cases}$

в точке $x_0 = 1.2$, $y_0 = 1.7$.

6. $\begin{cases} x_{k+1} = 0.1 - x_k^2 + 2y_k z_k, \\ y_{k+1} = y_k^2 - 3x_k z_k - 0.2, \\ z_{k+1} = 0.3 - z_k^2 - 2x_k y_k \end{cases}$ для системы $\begin{cases} x + x^2 - 2yz = 0.1, \\ y - y^2 + 3xz = -0.2, \\ z + z^2 + 2xy = 0.3 \end{cases}$

в точке $x_0 = 0$, $y_0 = -0.2$, $z_0 = 0.2$.

ЛЕКЦИЯ 5

ПРИБЛИЖЕНИЕ ФУНКЦИЙ

Постановка задачи об интерполировании функции степенными полиномами. Полиномы Лагранжа, Ньютона. Ошибка интерполяции. Кусочная интерполяция. Среднеквадратичное приближение. Метод наименьших квадратов. Неустранимые погрешности при интерполяции. Возможные обобщения задачи об интерполировании.

Пусть функция $f(x)$ задана множеством своих значений для дискретного набора точек (т. е. таблицей):

x_0	x_1	...	x_n	...
f_0	f_1	...	f_n	...

Здесь $f_i = f(x_i)$. Требуется найти приближенное значение $f(x)$ для $x \neq x_i$.

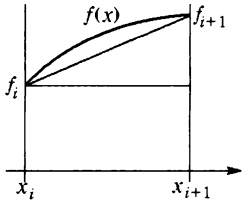


Рис. 5.1.

Это одна из самых часто встречающихся прикладных задач. (Табличные значения могут быть получены либо в результате расчетов, проведенных на ЭВМ, либо в процессе замеров, осуществленных в рамках какого-либо эксперимента.)

Очевидно, для достаточно подробной таблицы (когда $|x_i - x|$ — малые величины) для $x \approx x_i$ можно положить $f(x) \approx f_i$. Погрешность этого приближения $f(x) - f_i \approx f'(x)(x - x_i)$. Наверное, более точное приближение получим, если для $x \in [x_i, x_{i+1}]$ заменим функцию $f(x)$ отрезком прямой, проходящим через точки (x_i, f_i), (x_{i+1}, f_{i+1}), как это показано на рис. 5.1:

$$f(x) \approx f_i + \frac{f_{i+1} - f_i}{x_{i+1} - x_i}(x - x_i).$$

Итак, вырисовывается идея приближения функции степенными полиномами, принимающими в заданных точках заданные (табличные) значения. Эта идея лежит в основе теории интерполирования.

Приближение функций интерполяционными полиномами. Итак, пусть функция $f(x)$ задана таблицей

$$\{f_i = f(x_i), \quad i = 0, 1, 2, \ldots, n\},$$

содержащей значения в $(n + 1)$-й точке, причем x_i попарно различны.

Будем искать полином от x, проходящий через табличные точки:

$$P_n(x) = a_0 + a_1 x + \ldots + a_n x^n. \tag{5.1}$$

Требуя, чтобы в каждой табличной точке значение полинома совпадало с заданным значением функции, получим замкнутую систему линейных уравнений относительно неопределенных коэффициентов $\{a_k\}$:

$$a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_n x_i^n = f_i, \quad i = 0, 1, \ldots, n. \tag{5.2}$$

Определителем этой системы является рассматриваемый в курсах математического анализа определитель Вандермонда для системы несовпадающих точек:

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix} = \prod_{0 \leqslant j < i \leqslant n}^{n} (x_i - x_j) \neq 0,$$

т. е. искомый полином существует и единственен, если координаты табличных точек попарно различны (что соответствует известной теореме алгебры о том, что через $(n+1)$ точку можно провести единственную параболу n-й степени).

Замечательно то, что решение поставленной задачи можно выписать в явном виде:

$$\begin{aligned} P_n(x) = {} & f_0 \frac{(x - x_1)(x - x_2)\ldots(x - x_n)}{(x_0 - x_1)(x_0 - x_2)\ldots(x_0 - x_n)} + \ldots + \\ & + f_k \frac{(x - x_0)\ldots(x - x_{k-1})(x - x_{k+1})\ldots(x - x_n)}{(x_k - x_0)\ldots(x_k - x_{k-1})(x_k - x_{k+1})\ldots(x_k - x_n)} + \ldots + \\ & + f_n \frac{(x - x_0)(x - x_1)\ldots(x - x_{n-1})}{(x_n - x_0)(x_n - x_1)\ldots(x_n - x_{n-1})} = \sum_{k=0}^{n} f_k \frac{L_n^{(k)}(x)}{L_n^{(k)}(x_k)}, \end{aligned} \tag{5.3}$$

где $L_n^{(k)}(x) = (x - x_0)\ldots(x - x_{k-1})(x - x_{k+1})\ldots(x - x_n)$ — полиномы n-й степени специального вида.

В самом деле, очевидно, что (5.3) представляет собой полином n-й степени и что при подстановке в (5.3) значения $x = x_k$ получим $P_n(x_k) = f_k$ для любого k.

Полином, проходящий через табличные точки и записанный в форме (5.3), называется *интерполяционным полиномом Лагранжа*. Значения $\{x_i;\, i = 0, 1, \ldots, n\}$ называют *узлами интерполяции*. Если узлы упорядочены по величине, т. е. $x_{i+1} > x_i$ для всех i, то величины

$\{h_i = x_{i+1} - x_i;\ i = 0, 1, \ldots, n-1\}$ называют *шагами интерполяции*. Если $h_i = h = \text{const}$, так что $x_i = x_0 + hi$, $i = 0, 1, \ldots, n$, то говорят об *интерполяции по равноотстоящим узлам*. Отрезок $[x_0, x_n]$ — *отрезок интерполяции*.

Введем в рассмотрение еще один полином специального вида $(n+1)$-й степени:

$$\omega_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n) = \prod_{i=0}^{n}(x - x_i).$$

Тогда, очевидно, $L_n^{(k)}(x_k) = \omega'_{n+1}(x_k)$, а $L_n^{(k)}(x) = \omega_{n+1}(x)/(x - x_k)$, и полином Лагранжа можно записать с использованием $\omega_{n+1}(x)$ в виде:

$$P_n(x) = \omega_{n+1}(x) \sum_{k=0}^{n} \frac{f_k}{\omega'_{n+1}(x_k)(x - x_k)}. \tag{5.4}$$

Заметим, что коэффициент при x^n, как следует из (5.3), равен

$$a_n = \sum_{k=0}^{n} \frac{f_k}{L_n^{(k)}(x_k)}. \tag{5.5}$$

Приведем еще одну форму записи интерполяционного полинома:

$$P_n(x) = A_0 + A_1(x - x_0) + A_2(x - x_0)(x - x_1) + \ldots + \\ + A_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}). \tag{5.6}$$

Требование совпадения значений полинома с заданными значениями функции приводит к системе линейных уравнений с треугольной матрицей для неопределенных коэффициентов $\{A_i;\ i = 0, 1, \ldots, n\}$:

$$\begin{cases} A_0 = f_0, \\ A_0 + A_1(x_1 - x_0) = f_1, \\ A_0 + A_1(x_2 - x_0) + A_2(x_2 - x_0)(x_2 - x_1) = f_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases} \tag{5.7}$$

численное решение которой, как отмечалось в Лекции 2, не составляет труда.

Интерполяционный полином, записанный в форме (5.6), называется *полиномом Ньютона*. Он интересен тем, что каждая частичная

сумма его первых $(m+1)$ слагаемых представляет собой интерполяционный полином m-й степени, построенный по первым $(m+1)$ табличным данным.

Замечание. Принимая во внимание (5.5), можно выписать в явном виде решение системы (5.7):

$$A_m = \sum_{k=0}^{m} \frac{f_k}{L_m^{(k)}(x_k)}, \quad m = 0, 1, \ldots, n. \blacktriangle \tag{5.8}$$

Напомним, что в силу единственности решения задачи о построении интерполяционного полинома (5.1), (5.3), (5.6) — это различные формы записи одного полинома. Небезынтересно сопоставить эти формы записи с точки зрения удобства использования при практическом интерполировании. Однако сначала сделаем следующее замечание.

Замечание. Если надо вычислить приближенное значение функции при некотором $x \neq x_i$ $(i = 0, 1, \ldots, n)$, то это вовсе не означает, что надо привлекать интерполяционный полином, построенный по всем табличным точкам. Для большого числа табличных точек это бессмысленно, и мы еще коснемся этого вопроса ниже. Поступают так: строят полином невысокой степени по узлам, ближайшим к точке x, и используют его для вычисления $f(x)$. ▲

Возвращаемся к вопросу об удобствах использования той или иной формы записи интерполяционного полинома.

Что касается полинома Лагранжа, то он удобней других, если требуется приближать различные функции, заданные табличными значениями в одних и тех же точках. Если же в качестве результата нужна непосредственно формула, приближающая функцию $f(x)$, то, конечно, предпочтительней многочлен в форме (5.1) или полином Ньютона (с позиций нашего обычного восприятия формулы).

Но чтобы найти коэффициенты (5.1), необходимо решить систему линейных уравнений общего вида (5.2), в то время как коэффициенты полинома Ньютона отыскиваются из простой системы (5.7) или вычисляются по формулам (5.8). Опишем в качестве примера следующую ситуацию.

Учитывая последнее замечание, допустим, что вблизи $x = x_0$ мы построили полином третьей степени (по точкам x_0, x_1, x_2, x_3), а затем выяснилось, что точность, которую он обеспечивает, недостаточна и надо использовать интерполяцию четвертой степени. Для полинома Ньютона повышение его порядка на единицу сводится к добавлению одного слагаемого, т. е. в нашем случае к вычислению коэффициента A_4 (например, по формуле (5.8)), в то время как использование полинома в форме (5.1) требует при повышении порядка интерполя-

ции решения системы линейных уравнений (5.2) (в рассматриваемом случае — четвертого порядка).

З а м е ч а н и е. Разумеется, полином в форме Ньютона можно записать в окрестности любой табличной точки, «назвав» ее узлом x_0, а ближайшие табличные точки (с любой стороны и в любом порядке) — узлами x_1, x_2 и т. д. ▲

Погрешность интерполяции. Ошибка приближения функции интерполяционным полиномом n-й степени в точке x — это разность

$$R_n(x) = f(x) - P_n(x).$$

Оценить величину погрешности позволяет следующая теорема.

Т е о р е м а. *Пусть на отрезке* $[a, b]$, *таком, что* $[x_0, x_n] \subset [a, b]$, *функция* $f(x)$ $(n+1)$ *раз непрерывно дифференцируема. Тогда*

$$R_n(x) = \frac{f^{(n+1)}(x')}{(n+1)!}\omega_{n+1}(x), \tag{5.9}$$

где $x' \in [a, b]$.

Доказательство. Будем искать погрешность в виде

$$R_n(x) = C(x)\omega_{n+1}(x), \tag{5.10}$$

где $C(x)$ — функция, ограниченная на $[a, b]$ (при такой форме записи выражения для погрешности гарантируется, что она обращается в ноль в узлах интерполяции).

Чтобы получить представление о $C(x)$, рассмотрим вспомогательную функцию

$$\varphi(x) = f(x) - P_n(x) - C(\xi)\omega_{n+1}(x), \tag{5.11}$$

где ξ — некоторое фиксированное значение на отрезке $[a, b]$ такое, что $\xi \neq x_i$ для $\forall i$. Очевидно, на $[a, b]$ функция $\varphi(x)$ имеет $(n+2)$ нуля. Это узлы интерполяции и точка $x = \xi$. Согласно теореме Ролля, существует точка $x' \in [a, b]$, в которой $\varphi^{(n+1)}(x') = 0$. Продифференцировав (5.11) $(n+1)$ раз и подставив $x = x'$, получим

$$0 = \varphi^{(n+1)}(x') = f^{(n+1)}(x') - (n+1)!C(\xi).$$

Отсюда $C(\xi) = \dfrac{f^{(n+1)}(x')}{(n+1)!}$. (Ясно, что x' в теореме Ролля зависит от расположения нулей функции $\varphi(x)$; тем самым x' представляет собой некоторую неявную зависимость $x' = x'(\xi)$ и полученное отношение действительно определяет функцию от ξ.)

Так как при $x = \xi$ $\varphi(\xi) = 0$, то (5.11) можно записать в виде

$$\frac{f^{n+1}(x')}{(n+1)!}\omega_{n+1}(\xi) = f(\xi) - P_n(\xi) = R_n(\xi). \qquad (5.11')$$

В силу произвольности $\xi \in [a, b]$ ($\xi \neq x_i$ для $\forall i$), заменяя в (5.11') ξ на x, получаем (5.9), т. е. утверждение теоремы.

Из (5.9) следует оценка погрешности интерполяции

$$|R_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)|, \quad M_{n+1} = \max_{[a,b]}|f^{(n+1)}(x)|. \qquad (5.12)$$

Конкретная величина погрешности в точке x зависит, очевидно, от значения полинома $\omega_{n+1}(x)$ в этой точке. Качественный характер графика $\omega_{n+1}(x)$ иллюстрируется на рис. 5.2.

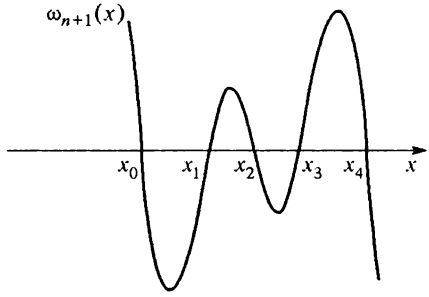


Рис. 5.2.

За пределами отрезка интерполяции (т. е. при *экстраполяции*) $|\omega_{n+1}(x)|$ быстро растет, экстремальные значения меньше в окрестности середины отрезка интерполяции.

Для равноотстоящих узлов ($x_i = x_0 + ih$) для $x \in [x_0, x_n]$:

$$\max_x|\omega_{n+1}(x)| \approx |\omega_{n+1}(x_0 + h/2)| \leqslant h \cdot h \cdot (2h) \cdot (3h)\ldots(nh) = n!h^{n+1}.$$

Поэтому на отрезке интерполяции ($x \in [x_0, x_n]$):

$$|R_n(x)| \leqslant \frac{M_{n+1}}{(n+1)}h^{n+1}.$$

Это сильно завышенная оценка ошибки. Для получения точной оценки надо искать экстремумы $\omega_{n+1}(x)$.

Замечание 1. Оценка (5.12) для погрешности интерполирования не является завышенной. Можно показать, что она достигается, например, при интерполировании полиномом n-й степени полинома $(n+1)$-й степени. ▲

З а м е ч а н и е 2. Можно за счет выбора узлов интерполирования минимизировать величину

$$\max_{[x_0, x_n]} |\omega_{n+1}(x)|$$

и тем самым добиться уменьшения погрешности интерполяции (см. Приложение 2). ▲

П р и м е р ы. Приблизим функцию $f(x)$ для $x \in [x_0, x_1]$ по равноотстоящим узлам полиномами различной степени и оценим погрешность.

1) $f(x) \approx P_1(x) = f_0 + \dfrac{f_1 - f_0}{h}(x - x_0)$;

$R_1(x) = \dfrac{f''(x')}{2}(x - x_0)(x - x_1)$, $|R_1(x)| \leqslant \dfrac{M_2}{8}h^2$, так как

$$\max_{[x_0, x_1]} |\omega_2(x)| = \max |(x - x_0)(x - x_1)| = \frac{h^2}{4}.$$

2) $f(x) \approx P_2(x) = f_0 + \dfrac{f_1 - f_0}{h}(x - x_0) + \dfrac{f_2 - 2f_1 + f_0}{2h^2}(x - x_0)(x - x_1)$;

$$R_2(x) = \frac{f'''(x')}{6}(x - x_0)(x - x_1)(x - x_2), \quad |R_2(x)| \leqslant \frac{M_3}{9\sqrt{3}^{(k)}}h^3 \leqslant \frac{M_3}{15}h^3.$$

3) $f(x) \approx P_3(x) = P_2(x) + \dfrac{f_3 - 3f_2 + 3f_1 - f_0}{6h^3}(x - x_0)(x - x_1)(x - x_2)$;

$$R_3(x) = \frac{f^{(k)}(\mathrm{IV})(x')}{24}(x - x_0)(x - x_1)(x - x_2)(x - x_3), \quad |R_3(x)| \leqslant \frac{M_4}{24}h^4.$$

Кусочная интерполяция. Оценка погрешности интерполяции (5.12) получена в предположении существования $(n + 1)$-й непрерывной производной $f(x)$. Далеко не всегда приходится иметь дело с очень гладкими функциями, это обстоятельство является аргументом в пользу так называемой *кусочной интерполяции*. Суть ее в том, что для приближения функции в точке x строится полином невысокой степени по данным в табличных точках, ближайшим к x. По сути дела рассмотренные примеры представляют собой кусочную интерполяцию в окрестности $[x_0 \div x_1]$. Приведем дополнительные примеры.

Пусть надо вычислить $f(x)$ для $x \in [x_i, x_{i+1}]$.

а) *Кусочно-линейная интерполяция.* Используется линейное приближение

$$f(x) \approx f_i + \frac{f_{i+1} - f_i}{h}(x - x_i).$$

б) *Кусочно-квадратичная интерполяция.* Привлекается еще одна табличная точка (x_{i-1} или $x_i + 2$) и строится полином второй степени.

Например,

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

Замечание. Это полином в форме Ньютона. Его можно представить в другой (запоминающейся) форме:

$$f(x) \approx f_i + \frac{f_{i+1} - f_{i-1}}{2h}(x - x_i) + \frac{f_{i+1} - 2f_i + f_{i-1}}{2h^2}(x - x_i)^2. \quad \blacktriangle$$

в) Кусочно-кубическая интерполяциая на отрезке (например, $[x_{i-1}, x_{i+2}]$) и т. д.

Среднеквадратичное приближение. Рассмотрим коротко принципиально иной способ приближения функций, заданных таблицей своих значений $\{f_i; i = 0, 1, \ldots, n\}$ в точках x_i. Будем искать приближение в виде полинома степени m:

$$P_m(x) = a_0 + a_1 x + \ldots + a_m x^m,$$

такого, который минимизирует сумму квадратов отклонений полинома от заданных значений функции:

$$\delta(a_0, a_1, \ldots, a_m) = \sum_{i=0}^{n} [P_m(x_i) - f_i]^2. \tag{5.13}$$

Очевидно, при $m = n$ решением поставленной задачи является интерполяционный полином, ибо на нем достигается абсолютный минимум (5.13): $\delta \equiv 0$. Известно (см., например, [3, т. 1]), что при $m \leqslant n$ поставленная задача имеет единственное решение (при $m > n$, очевидно, бесконечно много решений доставляют абсолютный минимум величине δ: произвольные $(n + 1)$ коэффициентов определяются из условий интерполяции, остальные полагаются равными нулю). Итак, остановимся на случае $m < n$.

Выпишем известные из математического анализа условия минимума (5.13):

$$\frac{\partial \delta}{\partial a_k} = 2 \sum_{i=0}^{n} [P_m(x_i) - f_i] x_i^k = 0, \quad k = 0, 1, \ldots, m,$$

или после подстановки выражения для $P_m(x_i)$ и перегруппировки слагаемых

$$a_0 \sum_{i=0}^{n} x_i^k + a_1 \sum_{i=0}^{n} x_i^{k+1} + \ldots + a_m \sum_{i=0}^{n} x_i^{k+m} = \sum_{i=0}^{n} f_i x_i^k, \quad k = 0, 1, \ldots, m. \tag{5.14}$$

Для неопределенных коэффициентов $\{a_0, a_1, \ldots, a_m\}$ мы получили замкнутую систему линейных алгебраических уравнений с симметричной матрицей

$$\begin{pmatrix} (n+1) & \Sigma x_i & \Sigma x_i^2 & \ldots & \Sigma x_i^m \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 & \ldots & \Sigma x_i^{m+1} \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 & \ldots & \Sigma x_i^{m+2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix},$$

элементы которой вычисляются через координаты табличных точек. В свою очередь правые части (5.14), как видно, определяются заданными табличными значениями функции.

Полином степени $m < n$ с коэффициентами, найденными из (5.14), называется *среднеквадратичным приближением функции, заданной таблицей* (иногда его называют также *наилучшим среди полиномов степени m приближением к функции по табличным данным*). Соответствующая погрешность приближения характеризуется среднеквадратичным отклонением $\Delta = \left(\frac{1}{n+1} \sum_i [P_m(x_i) - f_i]^2 \right)^{1/2}$.

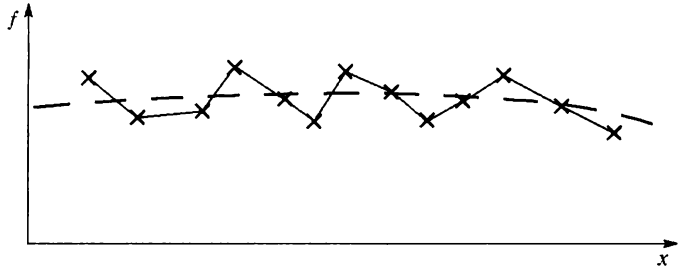


Рис. 5.3.

Основная сфера его применения — обработка экспериментальных данных (построение эмпирических формул). Дело в том, что результаты экспериментальных измерений, как правило, характеризуются заметным разбросом (рис. 5.3). Это следствие ошибок измерений, экспериментальный «шум». Интерполяционный полином, построенный по (отмеченным на рис. 5.3 крестиками) значениям функции, испытывает слишком сильное влияние «шума» и плохо приближает истинную зависимость $f(x)$, в то время как среднеквадратичный полином, минимизируя сумму квадратов отклонений и проходя между отмеченными значениями, обладает очевидным преимуществом: «сглаживает шум» (штриховая линия на рис. 5.3).

З а м е ч а н и е. Минимизация (5.13) называется методом наименьших квадратов. В данном случае этот метод использован для решения задачи о приближении функции. ▲

ДОПОЛНЕНИЯ К ЛЕКЦИИ 5

Неустранимые погрешности при интерполировании. Ошибка интерполяции $R_n^{(k)}(x)$ — это ошибка метода, и, как это свойственно методической погрешности, она «управляема»: ее можно уменьшать, сгущая узлы интерполирования или (при некоторых требованиях к $f(x)$) повышая степень полинома. Источником неустранимой погрешности интерполяции служат ошибки входных данных, т. е. погрешности табличных значений приближаемой функции.

Итак, пусть в точках $\{x_i\}$ известны значения $\{\widetilde{f}_i^{(k)} = f_i \pm \Delta_i\}$, где f_i — точные значения $f(x_i)$, $\pm\Delta_i$ — погрешности табличных данных, причем величину $\Delta = \max\limits_i \Delta_i$ можно считать известной (это, например, погрешность измерений при эксперименте). $\widetilde{P}_n^{(k)}(x)$ — интерполяционный полином, который реально можно построить по наличным данным, $P_n(x)$ — «идеальный» интерполяционный полином, отвечающий точным значениям $f_i = f(x_i)$. Тогда реальная погрешность

$$|\widetilde{R}_n(x)| = |f(x) - \widetilde{P}_n(x)| = |f(x) - P_n(x) + P_n(x) - \widetilde{P}_n(x)| \leqslant$$
$$\leqslant |f(x) - P_n^{(k)}(x)| + |P_n(x) - \widetilde{P}_n(x)| = |R_n(x)| + \Delta_p$$

представляет сумму методической погрешности $|R_n(x)|$ и погрешности, обусловленной ошибками табличных данных,

$$\Delta_p = |\widetilde{P}_n(x) - P_n(x)|.$$

Чтобы оценить последнюю, удобно использовать лагранжеву форму записи интерполяционного полинома:

$$\Delta_p = \left| \sum_{i=0}^{n} \widetilde{f}_i \frac{L_n^{(i)}(x)}{L_n^{(i)}(x_i)} - \sum_{i=0}^{n} f_i \frac{L_n^i(x)}{L_n^{(i)}(x_i)} \right| \leqslant \sum_{i=0}^{n} \Delta_i \left| \frac{L_n^{(i)}(x)}{L_n^{(i)}(x_i)} \right| \leqslant L_n \Delta,$$

где $L_n = \max\limits_{[x_0, x_n]} \Phi_n(x)$, а $\Phi_n(x) = \sum\limits_{i=0}^{n} \left| \dfrac{L_n^{(i)}(x)}{L_n^{(i)}(x_i)} \right|$.

Исследовать поведение $\Phi_n(x)$ в общем случае затруднительно. Приведем (для ориентировки) результаты в простейших случаях.

а) *Кусочно-линейная интерполяция.* Пусть $x \in [x_i, x_{i+1}]$ и

$$f(x) \approx \widetilde{P}_1(x) = \widetilde{f}_i \frac{x - x_{i+1}}{x_i - x_{i+1}} + \widetilde{f}_{i+1} \frac{x - x_i}{x_{i+1} - x_i}.$$

Тогда

$$\Phi_1(x) = \left|\frac{x - x_{i+1}}{x_i - x_{i+1}}\right| + \left|\frac{x - x_i}{x_{i+1} - x_i}\right| = \frac{x - x_{i+1}}{x_i - x_{i+1}} + \frac{x - x_i}{x_{i+1} - x_i} \equiv 1,$$

т. е. $\Delta_p = \Delta$(!).

б) *Кусочно-квадратичная интерполяция.* Пусть $x \in [x_{i-1}, x_{i+1}]$ и

$$f(x) \approx \widetilde{P}_2(x) = \widetilde{f}_{i-1} \frac{(x - x_i)(x - x_{i+1})}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})} + \\ + \widetilde{f}_i \frac{(x - x_{i-1})(x - x_{i+1})}{(x_i - x_{i-1})(x_i - x_{i+1})} + \widetilde{f}_{i+1} \frac{(x - x_{i-1})(x - x_i)}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)}.$$

Тогда

$$\Phi_2(x) = \left|\frac{(x - x_i)(x - x_{i+1})}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})}\right| + \\ + \left|\frac{(x - x_{i-1})(x - x_{i+1})}{(x_i - x_{i-1})(x_i - x_{i+1})}\right| + \left|\frac{(x - x_{i-1})(x - x_i)}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)}\right|.$$

Рассматривая по отдельности случаи, когда $x \in [x_{i-1}, x_i]$ и $x \in [x_i, x_{i+1}]$ (чтобы избавиться от модулей), можно получить, что $\Phi_2(x) \leqslant 1.25 \approx L_2$.

Заключая этот раздел, отметим, что в целом влияние погрешностей табличных данных на приближение функции с помощью интерполяции полиномами невысокой степени сравнительно невелико.

Но при увеличении n — степени интерполяционного полинома $L_n = \max \Phi_n(x)$ — константа Лебега быстро растет (см. Приложения 2, 4), что служит еще одним фактором, побуждающим отказываться от интерполяции полиномами большой степени ($n \gtrsim 5 \div 7$).

Возможные обобщения приближения функций с помощью интерполяции.

1. По аналогии с (5.1) можно искать приближение в виде обобщенного интерполяционного полинома

$$P_n(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x)$$

по системе линейно независимых функций $\{\varphi_k(x);\ k=0,1,\ldots,n\}$. Исходя из условий интерполяции (совпадения значений полинома с табличными значениями функции), для неопределенных коэффициентов $\{a_i\}$ получаем систему линейных уравнений

$$a_0\varphi_0(x_i)+a_1\varphi_1(x_i)+\ldots+a_n\varphi_n(x_i)=f_i,\quad i=0,1,\ldots,n.$$

Для существования и единственности решения необходимо [3, т. 1], чтобы детерминант удовлетворял условию

$$\Delta=\begin{vmatrix}\varphi_0(x_0) & \varphi_1(x_0) & \ldots & \varphi_n(x_0)\\ \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_n(x_1)\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_n(x_n)\end{vmatrix}\neq 0.$$

Например, периодическую функцию может оказаться удобным приближать в виде полинома по системе функций $\{\varphi_k(x)=1,\ \cos kx,\ \sin kx\ (k=1,2,\ldots,m,$ так что $(2m+1)=n)$. Это так называемая *тригонометрическая интерполяция* (см. Приложение 4).

2. Если в табличных точках заданы не только значения функции, но и ее производные, то для приближения функции используются *полиномы Эрмита*, для построения которых используются условия:

— в табличных точках полином принимает заданные значения;

— производные от полинома также принимают заданные значения.

Естественно, что полиномы Эрмита обеспечивают лучшее приближение сравнительно с обычной интерполяцией.

3. *Интерполяция кубическими сплайнами.* В отличие от кусочно-кубичной лагранжевой интерполяции, здесь при переходе от одного участка интерполяции к другому не претерпевают разрыва не только первые производные, но и вторые. Это значит, что сплайновая интерполяция обеспечивает *сквозное* (на всем отрезке интерполирования) гладкое приближение к функции $f(x)$ в виде полиномов третьей степени.

Строятся они следующим образом. На каждом отрезке $[x_k,\ x_{k+1}]$ ищется полином третьей степени $P_3^{(k)}=a_k+b_kx+c_kx^2+d_kx^3$ со своими коэффициентами. Требуя, чтобы все эти полиномы в табличных точках принимали заданные значения и чтобы полиномы на соседних отрезках гладким (вплоть до вторых производных) образом сопрягались друг с другом, получаем замкнутую линейную систему для $4n$ неопределенных коэффициентов, которая решается относительно просто.

4. *Интерполяция функций двух переменных.* Пусть на множестве точек $\{(x_i, y_i)\}$ заданы значения функции $f_{ij} = f(x_i, y_j)$. Ищем приближение функции в виде полинома

$$P(x, y) = a + bx + cy + dx^2 + exy + fy^2 + \ldots$$

Требуя, чтобы на некотором множестве табличных точек (в окрестности точки, в которой нужно найти приближение функции) выполнялись бы равенства $P^{(k)}(x_i, y_j) = f_{ij}$, получим систему уравнений для коэффициентов $a, b, c, \ldots$

Пример. Кусочно-линейное приближение в окрестности точки (x_i, y_j) (см. рис. 5.4)

$$f(x) \approx P_1(x, y) = f_{ij} + \frac{f_{i+1j} - f_{ij}}{h_x}(x - x_i) + \frac{f_{ij+1} - f_{ij}}{h_y}(y - y_j),$$

построенное по табличным данным в точках (x_i, y_j), (x_{i+1}, y_j), (x_i, y_{j+1}).

5. *Обратная интерполяция.* Иногда возникает необходимость определить, при каком значении x функция $f(x)$ принимает заданное значение. Учитывая, что при заданной таблице (см. начало лекции), с формальной точки зрения, вообще говоря, безразлично, что считать функцией и что аргументом, в этом случае представляется удобным приближать с помощью интерполяционного полинома зависимость $x(f)$. Это и есть обратная интерполяция. Очевидно, не составляет труда выписать соответствующие выражения для полиномов в форме Лагранжа или Ньютона.

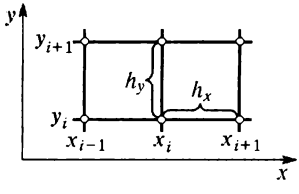


Рис. 5.4.

Более подробные сведения по вопросам, рассмотренным в Лекции 5 и дополнениях к ней, можно найти в [1, с. 292–363], [2, с. 34–73], [3, т. 1, с. 77–216], [5, с. 17–25], [9, с. 27–70], [12, с. 127–160], [15, с. 78–98, 210–259].

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Предложить экономичный способ вычисления значений интерполяционного полинома n-й степени, записанного в форме: а) (5.1); б) Ньютона.

Предполагается, что коэффициенты полиномов ($\{a_i\}$, соответственно $\{A_i\}$) известны (уже вычислены).

Замечание. Под экономичным здесь понимается способ вычислений, требующий числа арифметических действий, пропорционального степени полинома. ▲

2. С какой точностью можно вычислить $\sin 20°$ по известным значениям $\sin 0°$, $\sin 30°$, $\sin 45°$, $\sin 60°$, используя интерполяцию: а) линейную, б) квадратичную?

3. Таблица значений $f(x) = e^x$ дана на отрезке $[0, 1]$ с шагом 0.1. Какова наибольшая погрешность интерполяции: а) линейной, б) квадратичной?

4. Таблица интеграла вероятности $\dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^x e^{-x^2}\, dx$ на отрезке $[0, 3]$ дана с шагом $h = 0.01$. Какова наибольшая погрешность линейной интерполяции?

5. С какой точностью имеет смысл задавать значения таблицы $f(x) = \sin x$ на отрезке $[0, \pi]$, если шаг таблицы $h = \pi/20$ и предполагается использовать квадратичную интерполяцию для приближения функции в нетабличных точках?

6. В точках $x_k = k$ $(k = 0, 1, \ldots, 20)$ дана таблица функции $f(x) = x^3$. С какой точностью можно вычислить кубический корень из 1200, используя обратную кусочно-квадратичную интерполяцию по этой таблице?

7. Найти среднеквадратичное приближение среди полиномов степени не выше второй для функции, заданной значениями $-1, 0, 1$ в точках $-0.5, 0, 0.5$.

Сравнить найденное приближение с интерполяционным полиномом 2-й степени, построенным по тем же данным. Сравнить также значение построенного полинома в точке $x = 0.25$ со значением функции $\sin \pi x$.

8. Функция $f(x) = 11^x$ задана таблицей

x	0	1	2
$f(x)$	1	11	121

а) Найти среднеквадратичное приближение в виде обобщенного полинома $f(x) = P_1(x) = a + b \cdot 10^x$.

б) Найти интерполяционный обобщенный полином, приближающий эту функцию:

$$f(x) \simeq P_2(x) = a + b \cdot 10^x + c \cdot 10^{2x}.$$

в) Сравнить найденные приближения с $f(x)$ при $x = 0.5$, $x = 1.5$. Изобразить качественную картину на графике.

ЛЕКЦИЯ 6

ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ И ИНТЕГРИРОВАНИЕ

Построение формул для приближенного вычисления производных, анализ погрешности. Неустойчивость численного дифференцирования. Квадратурные формулы прямоугольников, трапеций, Симпсона. Погрешность квадратурных формул. Вычисление несобственных интегралов.

Численное дифференцирование. В этом разделе мы обсудим способы конструирования формул, позволяющих вычислять приближенные значения производных от функции, заданной таблицей значений.

Итак, пусть в точках $x_i, i = 0, 1, \ldots, n$ известны значения функции $f(x) : \{f_i, i = 0, 1, \ldots, n\}$. Простейший способ построения формул численного дифференцирования состоит в следующем. По табличным данным приближаем функцию интерполяционным полиномом. Дифференцируя полином нужное число раз, получаем требуемую формулу.

Например, при использовании полинома n-й степени имеем

$$f(x) = P_n(x) + R_n(x);$$

$R_n(x)$ — ошибка интерполяции (или остаточный член интерполирования). Соответственно для k-й производной от функции $f(x)$ на отрезке интерполирования $[x_0, x_n]$ получаем приближенную формулу

$$\frac{d^k f}{dx^k} \approx \frac{d^k P_n(x)}{dx^k}, \tag{6.1}$$

погрешность которой характеризуется k-й производной от ошибки интерполяции:

$$\frac{d^k R_n(x)}{dx^k}.$$

Замечание. Анализ погрешности формул численного дифференцирования, опирающийся на дифференцирование остаточного члена, затруднен тем, что выражение для последнего, как следует из способа его вывода (см. Лекцию 5),

$$R_n(x) = \frac{f^{n+1}(x')}{(n+1)!}(x - x_0)(x - x_1)\ldots(x - x_n)$$

содержит неявную зависимость от x через $x' = x'(x) \in [x_0, x_n]$. Тем не менее, именно такой способ исследования погрешности реализован в [9]. Мы ниже познакомимся с другим подходом к анализу точности формул численного дифференцирования. ▲

Примеры. Пусть надо приблизить производные функции $f(x)$ в окрестности табличной точки $x \sim x_i$. Далее мы будем полагать, что табличные точки равноотстоят друг от друга, т. е. $x_i = x_0 + ih$. Это непринципиально с точки зрения намеченного способа конструирования приближенных формул для производных, просто формулы будут выглядеть проще.

а) Приблизим в рассматриваемой окрестности функцию полиномом первой степени:

$$f(x) \approx p_1(x) = f_i + \frac{f_{i+1} - f_i}{h}(x - x_i).$$

Тогда

$$\left.\frac{df}{dx}\right|_{x \sim x_i} \approx \frac{dp_1(x)}{dx} = \frac{f_{i+1} - f_i}{h}. \tag{6.2}$$

Используя кусочно-линейную интерполяцию, можно было бы приблизить $f(x)$ в рассматриваемой окрестности и так:

$$f(x) \approx \overline{p}_1(x) = f_i + \frac{f_i - f_{i-1}}{h}(x - x_i).$$

Отсюда

$$\left.\frac{df}{dx}\right|_{x \sim x_i} \approx \frac{d\overline{p}_1(x)}{dx} = \frac{f_i - f_{i-1}}{h}. \tag{6.2'}$$

Мы получили простейшие приближенные формулы для первой производной от функции, заданной таблично: (6.2) — *правое разностное отношение*, (6.2') — *левое разностное отношение*. Очевидно, можно было бы их написать сразу, опираясь на определение производной от функции и не привлекая интерполяцию в качестве промежуточного этапа. Ясно также, что для получения приближенных формул для второй и высших производных линейного приближения функции недостаточно.

б) Приблизим в рассматриваемой окрестности функцию полиномом второй степени:

$$f(x) \approx p_2(x) = f_i + \frac{f_{i+1} - f_{i-1}}{2h}(x - x_i) + \frac{f_{i+1} - 2f_i + f_{i-1}}{2h^2}(x - x_i)^2.$$

(Легко проверяется, что это интерполяционный полином, построенный по значениям функции в точках $\{x_{i-1}, x_i, x_{i+1}\}$. Мы привлекли здесь такую форму записи потому, что она несколько компактней

сравнительно с ньютоновской, тем более с лагранжевой формами.) Дифференцируя один раз, получаем новую приближенную формулу для первой производной:

$$\left.\frac{df}{dx}\right|_{x\sim x_i} \approx \frac{f_{i+1}-f_{i-1}}{2h} + \frac{f_{i+1}-2f_i+f_{i-1}}{h^2}(x-x_i). \tag{6.3}$$

Здесь, в отличие от (6.1), (6.2), приближение зависит от x. В частности, для $x = x_i$ имеем

$$\left.\frac{df}{dx}\right|_{x=x_i} \approx \frac{f_{i+1}-f_{i-1}}{2h}. \tag{6.4}$$

Это так называемое *центральное разностное отношение.*

Дифференцируя полином два раза, получаем приближенную формулу для второй производной:

$$\left.\frac{d^2f}{dx^2}\right|_{x\sim x_i} \approx \frac{f_{i+1}-2f_i+f_{i-1}}{h^2}. \tag{6.5}$$

Аналогичным образом, привлекая интерполяцию более высокого порядка, можно получать новые формулы для первой и второй производных и формулы для высших производных.

Погрешность формул численного дифференцирования. Отметим, что любая формула для приближения любой производной *в конкретной точке* имеет следующую структуру:

$$\frac{d^kf}{dx^k} \approx \sum c_i f_i, \tag{6.6}$$

(c_i — постоянные коэффициенты), где суммирование производится по некоторому диапазону табличных данных. Анализ погрешности формулы (6.6) сводится к следующему. Предполагая необходимую гладкость функции $f(x)$, заменяем все входящие в правую часть (6.6) значения f_i по формулам Тейлора *относительно точки, для которой рассматривается приближение* (6.6). После проведения простых арифметических выкладок в качестве главного члена правой части получим приближенное значение производной. Остальные члены будут характеризовать погрешность.

Пример. Оценить погрешность формулы (6.2) для точки $x \in [x_i, x_{i+1}]$. Заменяем f_i и f_{i+1} по формулам Тейлора относительно

точки x:

$$f_{i+1} = f(x) + f'(x)(x_{i+1} - x) + \frac{1}{2} f''(x)(x_{i+1} - x)^2 + \\ + \frac{1}{6} f'''(\xi)(x_{i+1} - x)^3,$$

$$f_i = f(x) + f'(x)(x_i - x) + \frac{1}{2} f''(x)(x_i - x)^2 + \frac{1}{6} f'''(\eta)(x_i - x)^3.$$

(Последние слагаемые представляют собой остаточные члены, так что $\xi \in [x, x_{i+1}]$, $\eta \in [x_i, x]$.)

Подставляя эти выражения в правую часть (6.2), получаем

$$\frac{f_{i+1} - f_i}{h} = \frac{1}{h}\left\{f'(x)(x_{i+1} - x_i) + \frac{1}{2} f''(x)\left[(x_{i+1} - x)^2 - (x_i - x)^2\right] + \right. \\ \left. + \frac{1}{6} f'''(\xi)(x_{i+1} - x)^3 - \frac{1}{6} f'''(\eta)(x_i - x)^3\right\} = f'(x) + \\ + \frac{1}{2} f''(x)(x_{i+1} + x_i - 2x) + \frac{1}{6} f'''(\xi)\frac{(x_{i+1} - x)^3}{h} - \frac{1}{6} f'''(\eta)\frac{(x_i - x)^3}{h}.$$

Таким образом, погрешность формулы (6.2) для произвольной точки $x \in [x_i, x_{i+1}]$

$$\Delta = \left| f'(x) - \frac{f_{i+1} - f_i}{h} \right| \leqslant M_2 \left| \frac{x_{i+1} + x_i}{2} - x \right| + \frac{M_3}{3} h^2, \qquad (6.7)$$

где $M_2 = \max\limits_{[x_i, x_{i+1}]} |f''(x)|$, $M_3 = \max\limits_{[x_i, x_{i+1}]} |f'''(x)|$.

Для $x = \dfrac{x_i + x_{i+1}}{2}$ $\Delta \sim O(h^2)$. Для остальных x $\Delta \leqslant \dfrac{M_2 h}{2}$, т. е. формула (6.2) является в общем случае формулой первого порядка точности, но для середины интервала она обеспечивает второй порядок.

Замечание. Вообще говоря, второй порядок точности достигается для всех точек x, таких, что $\left| x - \dfrac{x_i + x_{i+1}}{2} \right| = O(h^2)$. ▲

Численное интегрирование. Речь пойдет о методах вычисления значения интеграла

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

Мы рассмотрим здесь простейшие, но в то же время широко используемые в практических вычислениях формулы: прямоугольников (с центральной точкой), трапеций, Симпсона.

Способ их получения состоит в следующем. Разобьем отрезок интегрирования $[a, b]$ на N элементарных шагов. Точки разбиения $x_n (n = 0, 1, \ldots, N)$; $h_n = x_{n+1} - x_n$, так что $\sum\limits_{n=0}^{N-1} h_n = b - a$. В дальнейшем будем называть x_n узлами, h_n — шагами интегрирования. (В частном случае шаг интегрирования может быть постоянным $h = (b - a)/N$.) Также будем пользоваться обозначением $f_n = f(x_n)$.

Искомое значение интеграла представим в виде

$$J = \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} f(x)\,dx = \sum_{n=0}^{N-1} J_n, \tag{6.8}$$

где $J_n = \int\limits_{x_n}^{x_{n+1}} f(x)\,dx$.

Формула прямоугольников. Считая h_n малым параметром, заменим J_n в (6.8) площадью прямоугольника с основанием h_n и высотой $f_{n+1/2} = f(x_n + h_n/2)$. Тогда придем к локальной формуле прямоугольников

$$\widetilde{J}_n = h_n f_{n+1/2}. \tag{6.9}$$

Суммируя в соответствии с (6.8) приближенные значения по всем элементарным отрезкам, получаем формулу прямоугольников для вычисления приближения к J:

$$\widetilde{J} = \sum_{n=0}^{N-1} h_n f_{n+1/2}. \tag{6.10}$$

В частном случае, когда $h_n = h = \text{const}$, формула прямоугольников записывается в виде

$$\widetilde{J} = h \sum_{n=0}^{N-1} f_{n+1/2}. \tag{6.10'}$$

З а м е ч а н и е. Можно конструировать аналогичные формулы, используя в качестве высоты элементарных прямоугольников значение $f(x)$ не в середине отрезков, а, например, на границе (левой или правой). Но в этом случае существенно ухудшается точность вычисляемого результата (см. ниже «Погрешность квадратурных формул»). ▲

Формула трапеций. На элементарном отрезке $[x_n, x_{n+1}]$ заменим подынтегральную функцию интерполяционным полиномом пер-

вой степени:

$$f(x) \approx f_n + \frac{f_{n+1} - f_n}{x_{n+1} - x_n}(x - x_n).$$

Выполняя интегрирование по отрезку, приходим к локальной формуле трапеций:

$$\widetilde{J}_n = \frac{1}{2}(x_{n+1} - x_n)(f_{n+1} + f_n) = \frac{1}{2}h_n(f_{n+1} + f_n). \qquad (6.11)$$

Замечание. Название связано с тем, что интеграл по элементарному отрезку заменяется площадью трапеции с основаниями, равными значениям $f(x)$ на краях отрезка, и высотой, равной h_n. ▲

Суммируя (6.11) по всем отрезкам, получаем формулу трапеций для вычисления приближения к J:

$$\widetilde{J} = \frac{1}{2}\sum_{n=0}^{N-1} h_n(f_n + f_{n+1}). \qquad (6.12)$$

В случае постоянного шага интегрирования формула принимает вид

$$\widetilde{J} = \frac{h}{2}\sum_{n=0}^{N-1}(f_n + f_{n+1}) = \frac{h}{2}\left[f_0 + 2f_1 + \ldots + 2f_{N-1} + f_N\right].$$

Формула Симпсона. На элементарном отрезке $[x_n, x_{n+1}]$, привлекая значение функции в середине, заменим подынтегральную функцию интерполяционным полиномом второй степени

$$f(x) \approx P_2(x) = f_{n+1/2} + \frac{f_{n+1} - f_n}{h_n}\left[x - \frac{x_{n+1} + x_n}{2}\right] + \\ + \frac{f_{n+1} - 2f_{n+1/2} + f_n}{2(h_n/2)^2}\left[x - \frac{x_{n+1} + x_n}{2}\right]^2$$

(напомним, что $h_n = x_{n+1} - x_n$, $f_n = f(x_n)$, $f_{n+1/2} = f(x_n + h_n/2)$). Вычисляя интеграл от полинома по отрезку $[x_n, x_{n+1}]$, приходим к локальной формуле Симпсона

$$\widetilde{J}_n = \frac{h_n}{6}(f_n + 4f_{n+1/2} + f_{n+1}). \qquad (6.13)$$

Суммируя (6.13) по всем отрезкам, получаем формулу Симпсона для вычисления приближения к J:

$$\widetilde{J} = \frac{1}{6}\sum_{n=0}^{N-1} h_n(f_n + 4f_{n+1/2} + f_{n+1}). \qquad (6.14)$$

Для постоянного шага интегрирования $h_n = \text{const} = h = (b-a)/N$ формула принимает вид

$$\widetilde{J} = \frac{1}{6} h \sum_{n=0}^{N-1} (f_n + 4f_{n+1/2} + f_{n+1}) =$$

$$= \frac{h}{6}(f_0 + 4f_{1/2} + 2f_1 + 4f_{3/2} + \ldots + 2f_{N-1} + 4f_{N-1/2} + f_N). \quad (6.15)$$

Замечание. Иногда последнюю формулу записывают без использования дробных индексов в виде

$$\widetilde{J} = \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \ldots + 2f_{N-2} + 4f_{N-1} + f_N). \quad (6.15')$$

К этой записи приходим, если под локальной формулой понимать результат интегрирования по паре элементарных отрезков

$$\widetilde{J}_n = \int_{x_{n-1}}^{x_{n+1}} \widetilde{P}_2(x)\,dx = \frac{h}{3}(f_{n-1} + 4f_n + f_{n+1}),$$

где $\widetilde{P}_2(x)$ — интерполяционный полином второй степени для $f(x)$ на $[x_{n-1}, x_{n+1}]$, построенный по значениям в точках x_{n-1}, x_n, x_{n+1}. (Шаг h здесь предполагается постоянным.) Суммируя локальные приближения по всем парам, получим (6.15′). Разумеется, число пар на $[a, b]$ в этом случае должно быть целым, т. е. N — четным. ▲

Формулы, используемые для приближенного вычисления интеграла, называются *квадратурными*.

Погрешность квадратурных формул. Один из возможных способов оценки точности построенных формул состоит в следующем. Рассмотрим интеграл по элементарному отрезку

$$J_n = \int_{x_n}^{x_{n+1}} f(x)\,dx.$$

Выберем на этом отрезке какую-либо «опорную» точку $x = z$ и представим подынтегральную функцию по формуле Тейлора относительно этой точки:

$$f(x) = f(z) + f'(z)(x-z) + \frac{1}{2} f''(x-z)^2 + \ldots + R(x-z),$$

где $R(x-z)$ — остаточный член используемой формулы Тейлора. Вычисляя интеграл от последней суммы, получаем представление J_n в виде

$$J_n = f(z)h_n + A_2h_n^2 + A_3h_n^3 + \ldots + \int_{x_n}^{x_{n+1}} R(x-z)\,dz, \qquad (6.16)$$

где коэффициенты A_2, A_3, ... зависят от производных $f'(z)$, $f''(z)$, ...

Заметим далее, что каждая из рассматриваемых квадратурных формул (прямоугольников, трапеций и Симпсона) в пределах элементарного отрезка $[x_n, x_{n+1}]$ может быть представлена в виде

$$\widetilde{J}_n = h[rf_n + sf_{n+1/2} + qf_{n+1}] \qquad (6.17)$$

со своими коэффициентами r, s, q. Заменяя в (6.17)каждое из значений функции f по формуле Тейлора относительно той же точки z, получим представление приближенного значения $\widetilde{J}_n$ также в виде, аналогичном (6.16):

$$\widetilde{J}_n = f(z)h_n + B_2h_n^2 + B_3h_n^3 + \ldots + \widetilde{R}. \qquad (6.18)$$

Сравнивая представления (6.16) и (6.18), обнаруживаем, что наряду с первым слагаемым в (6.16), (6.18) совпадает еще некоторое количество $(p-1)$ слагаемых, так что $A_2 = B_2$, $A_3 = B_3$, ... Разность несовпадающих слагаемых будет, очевидно, характеризовать ошибку квадратурной формулы. Оценивая величину этой разности, приходим к оценке для локальной (на интервале $[x_n, x_{n+1}]$) погрешности:

$$|\widetilde{J}_n - J_n| \leqslant D \max_{[x_n, x_{n+1}]} |f_n^{(p)}|\, h_n^{p+1}, \qquad (6.19)$$

где D — числовая константа, а $f^{(p)}$ — p-я производная функции $f(x)$.

Суммируя локальные погрешности по всем интервалам, получим требуемую оценку погрешности рассматриваемой формулы:

$$|\widetilde{J} - J| \leqslant D(b-a)M_p\,\overline{h}^p, \qquad (6.20)$$

где $M_p = \max|f^{(p)}|$ по всему отрезку $[a, b]$, а $\overline{h} = \max\limits_n h_n$, если $h_n \neq$ $=$ const. Степень p принято называть порядком точности квадратурной формулы.

Для рассмотренных здесь квадратурных формул получаемые таким образом оценки погрешности имеют вид:

формула прямоугольников — $|\widetilde{J} - J| \leqslant \dfrac{1}{24}(b-a)M_2\overline{h}^2$,

формула трапеций — $|\widetilde{J} - J| \leqslant \dfrac{1}{12}(b-a)M_2\overline{h}^2$, (6.20′)

формула Симпсона (6.15) — $|\widetilde{J} - J| \leqslant \dfrac{1}{2880}(b-a)M_4\overline{h}^4$,

формула Симпсона (6.15′) — $|\widetilde{J} - J| \leqslant \dfrac{1}{180}(b-a)M_4h^4$.

Замечание 1. От выбора «опорной» точки результат не зависит. Например, для оценки погрешности формул прямоугольников и Симпсона (6.15) целесообразно взять в качестве z середину отрезка $[x_n, x_{n+1}]$ (выкладки при этом упрощаются). ▲

Замечание 2. Полученные оценки, как следует из их вида, зависят от гладкости подынтегральной функции. Например, если $f(x)$ только трижды непрерывно дифференцируема на $[a, b]$, то оценка погрешности формулы Симпсона ухудшается на порядок. ▲

Если известны оценки для абсолютных величин соответствующих производных, то, используя (6.20), можно априори (до проведения расчета) определить шаг интегрирования $h = \text{const}$, при котором погрешность вычисленного результата гарантировано не превышает допустимого уровня погрешности ε. Для этого, как следует из (6.20), достаточно решить относительно h неравенство $D(b-a)\,M_p\,h^p \leqslant \varepsilon$.

Однако типичной является ситуация, когда величины нужных производных не поддаются оценке. Тогда контроль за точностью вычисляемых результатов можно организовать, проводя вычисления на последовательно сгущающейся сетке узлов интегрирования.

Контроль за точностью вычисляемого значения интеграла. Пусть шаг интегрирования $h = \text{const}$, $J^{(h)}$ — вычисленное с шагом h приближение к J. Если, далее, вычислено также приближенное значение $J^{(h/2)}$ с шагом $h/2$, то в качестве приближенной оценки погрешности последнего вычисленного значения можно рассматривать величину

$$|J^{(h/2)} - J| \approx |J^{(h/2)} - J^{(h)}|.$$

При необходимости вычислить результат с требуемой точностью (ε) вычисления повторяются с последовательно уменьшающимся (вдвое) шагом до тех пор, пока не выполнится условие

$$|J^{(h/2)} - J^{(h)}| \leqslant \varepsilon.$$

Счет с автоматическим выбором шагов интегрирования. Можно применять указанное правило для контроля за локальной погрешностью на каждом элементарном интервале. При этом длина оче-

редного интервала $h_n = x_{n+1} - x_n$, посредством последовательного уменьшения (или увеличения!) начальной длины вдвое, устанавливается такой, чтобы локальная погрешность удовлетворяла неравенству

$$|\widetilde{J}_n - J_n| \approx |J_n^{(h)} - J_n^{(h/2)}| \leqslant \varepsilon h_n/(b-a).$$

Тогда в худшем случае ошибка вычисленного значения по всему отрезку интегрирования не будет превосходить сумму локальных погрешностей $\sum_n \varepsilon h_n/(b-a) = \varepsilon$, т. е. не будет превосходить заданного уровня погрешности.

Способ вычисления с автоматическим выбором имеет то преимущество, что он «приспосабливается» к особенностям подынтегральной функции: в областях резкого изменения функции шаг уменьшается, а там, где функция меняется слабо, — увеличивается.

ДОПОЛНЕНИЯ К ЛЕКЦИИ 6

Неустойчивость численного дифференцирования. Речь пойдет о влиянии погрешностей входных данных на результат вычисления производных по формулам численного дифференцирования. Разберемся с сутью вопроса на конкретном примере.

Пусть в точках $\{x_i, i = 1, 2, \ldots, n\}$ заданы значения функции $\{\widetilde{f}_i\}$, которые отличаются от точных значений $f_i = f(x_i)$:

$$\widetilde{f}_i = f_i \pm \delta_i,$$

δ_i — погрешности табличных данных. Как правило, оценка абсолютных погрешностей исходных данных

$$\delta = \max_i \delta_i \tag{6.21}$$

является известной величиной.

Пусть в точке $x = x_i$ нужно приблизить первую производную от функции $f(x)$. Используем для этой цели какую-либо приближенную формулу, для определенности — простейшую:

$$\left.\frac{df}{dx}\right|_{x=x_i} \approx \frac{\widetilde{f}_{i+1} - \widetilde{f}_i}{h}. \tag{6.22}$$

Ее погрешность

$$\Delta = \left|\frac{df}{dx} - \frac{\widetilde{f}_{i+1} - \widetilde{f}_i}{h}\right| = \left|\left[\frac{df}{dx} - \frac{f_{i+1} - f_i}{h}\right] + \left[\frac{f_{i+1} - f_i}{h} - \frac{\widetilde{f}_{i+1} - \widetilde{f}_i}{h}\right]\right| \leqslant$$

$$\leqslant \left|\frac{df}{dx} - \frac{f_{i+1} - f_i}{h}\right| + \left|\frac{f_{i+1} - \widetilde{f}_{i+1}}{h}\right| + \left|\frac{f_i - \widetilde{f}_i}{h}\right|.$$

Первое слагаемое представляет собой методическую ошибку и согласно оценке, полученной в лекции, не превосходит $\frac{M_2}{2}h$, где $M_2 = \max\limits_{[x_i, x_{i+1}]} |f''(x)|$.

Учитывая (6.21), получаем

$$\Delta \leqslant \frac{M_2}{2}h + \frac{2\delta}{h} = \Phi(h). \tag{6.23}$$

Очевидно, при $h \sim \delta$ неустранимая погрешность может стать величиной порядка $O(1)$. В этом проявляется неустойчивость численного дифференцирования: погрешности табличных данных могут как угодно сильно исказить вычисляемый по приближенной формуле результат.

Как видно из (6.23), существует оптимальный шаг $h_{\text{опт}}$, при котором оценка для возможной ошибки минимальна. Найдем его:

$$\frac{d\Phi(h)}{dh} = \frac{M_2}{2} - \frac{2\delta}{h^2} = 0.$$

Отсюда $h_{\text{опт}} = 2\sqrt{\frac{\delta}{M_2}}$. При этом $\Phi(h_{\text{опт}}) = \frac{M_2}{2}2\sqrt{\frac{\delta}{M_2}} + 2\delta\frac{1}{2}\sqrt{\frac{M_2}{\delta}} = 2\sqrt{M_2\delta}$.

Таким образом, в лучшем случае (при оптимальном шаге) можно лишь гарантировать, что производная от $f(x)$, вычисленная по формуле (6.22), может отличаться от точного значения не более, чем на $2\sqrt{M_2\delta}$. Если, например, $\delta \leqslant 0.01$, то $\Delta \sim 0.1$ (при $M_2 \sim 1$). Точно так же для любой другой формулы численного дифференцирования при известных оценках для величин производных от $f(x)$, требуемых при анализе, может быть найден вклад в итоговую ошибку неустранимых погрешностей входных данных и оптимальный шаг, при котором суммарная погрешность минимизируется.

Устойчивость квадратурных формул. При численном интегрировании влияние погрешностей входных данных на результат вычислений вполне умеренно. В самом деле, любая из рассмотренных нами квадратурных формул может быть представлена в виде

$$\widetilde{J} = \sum c_i f_i \ \ (c_i > 0 \text{ — постоянные коэффициенты}),$$

где суммирование производится по всем узлам интегрирования. Нетрудно понять, что $\sum c_i \equiv b - a$. (В самом деле, любая квадратурная формула дает точный результат, если $f = \text{const}$, в частности при

$f = 1$.) Таким образом, для $\widetilde{f}_i = f_i \pm \delta_i$ имеем

$$|J - \widetilde{J}| = \left|J - \sum c_i \widetilde{f}_i\right| = \left|\left(J - \sum c_i f_i\right) + \left(\sum c_i f_i - \sum c_i \widetilde{f}_i\right)\right| \leqslant$$
$$\leqslant \left|J - \sum c_i f_i\right| + \sum c_i |f_i - \widetilde{f}_i|.$$

Первое слагаемое в правой части неравенства — ошибка метода, второе — неустранимая погрешность, обусловленная погрешностями входных данных. Для последней получаем оценку

$$\sum c_i \left|f_i - \widetilde{f}_i\right| \leqslant \delta \sum c_i = \delta(b - a),$$

т. е. неустранимая погрешность при вычислениях по квадратурным формулам остается ограниченной (и является величиной порядка погрешности самих входных данных, если длина отрезка интегрирования порядка единицы).

Приемы вычисления несобственных интегралов. Имеются в виду сходящиеся интегралы двух типов:

1) $\int\limits_a^b f(x)\,dx$, причем $f(x) \to \infty$ при $x \to a$, 2) $\int\limits_a^\infty f(x)\,dx$.

З а м е ч а н и е. Второй интеграл, вообще говоря, может быть сведен к первому заменой переменной интегрирования $t = 1/x$. Поэтому пока будем говорить об интегралах первого типа. ▲

Очевидно, непосредственное использование квадратурных формул трапеций и Симпсона для вычисления таких интегралов невозможно (так как точка $x = a$, в которой подынтегральная функция не определена, является для этих формул узлом интегрирования). По методу прямоугольников вычисления формально провести можно, но результат будет сомнительным, так как оценка погрешности теряет смысл (производные подынтегральной функции не ограничены).

Продемонстрируем приемы, которые позволяют получать надежные результаты в подобных случаях, на примере интеграла

$$J = \int_0^1 \frac{\cos x}{\sqrt{x}}\,dx.$$

а) Иногда подходящая замена переменной интегрирования позволяет вообще избавиться от особенности. В рассматриваемом примере

после замены $x = t^2$ получаем

$$J = 2\int_0^1 \cos(t^2)\,dt,$$

и интеграл вычисляется с требуемой точностью по любой из квадратурных формул.

б) Та же цель (избавление от особенности) достигается иногда предварительным интегрированием по частям:

$$J = \int_0^1 \frac{\cos(x)}{\sqrt{x}}\,dx = 2\sqrt{x}\cos(x)\Big|_0^1 + 2\int_0^1 \sqrt{x}\sin(x)\,dx.$$

Последний интеграл формально может быть вычислен стандартным образом, но оценка погрешности для любой квадратурной формулы будет иметь лишь первый порядок, так как при $x = 0$ не существует второй производной от подынтегральной функции. Проводя еще пару раз интегрирование по частям, придем к интегралу от дважды непрерывно дифференцируемой функции, который с гарантированной точностью может быть вычислен по формулам трапеций или прямоугольников.

в) Если упомянутыми простыми средствами избавиться от особенности не удается, то прибегают к универсальному методу выделения особенности. В нашем случае представим интеграл в виде суммы двух:

$$J = J_1 + J_2, \quad J_1 = \int_0^\delta \frac{\cos(x)}{\sqrt{x}}\,dx, \quad J_2 = \int_\delta^1 \frac{\cos(x)}{\sqrt{x}}\,dx.$$

Второй интеграл особенности не содержит и вычисляется по любой квадратурной формуле. (Вопрос о выборе δ обсуждается ниже.) Первый интеграл с требуемой точностью вычисляем аналитически, используя представление подынтегральной функции в окрестности особой точки $(x = 0)$ в виде отрезка ряда по степеням x, который получим после замены $\cos(x)$ соответствующим рядом Тейлора:

$$J_1 = \int_0^\delta \frac{1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} + \ldots + (-1)^m \dfrac{x^{2m}}{(2m)!} + \ldots}{\sqrt{x}}\,dx = 2\sqrt{\delta} - \frac{1}{2!}\frac{2}{5}\delta^{5/2} +$$

$$+ \frac{1}{4!}\frac{2}{9}\delta^{9/2} + \ldots + (-1)^m \frac{1}{(2m)!}\frac{1}{(2m+1/2)}\delta^{2m+1/2} + \ldots$$

Допустим, что мы решили ограничиться первыми m слагаемыми в полученном представлении. При этом погрешность не превосходит последнего приведенного в записи для J_1 слагаемого (в силу знакопеременности полученного ряда). Следовательно, для выбора двух параметров (δ и m) имеем условие

$$\frac{1}{(2m)!}\,\frac{1}{(2m+1/2)}\,\delta^{2m+1/2} \leqslant \frac{\varepsilon}{2}.$$

(Еще $\varepsilon/2$ отводится в качестве допустимого уровня погрешности при вычислении J_2.)

Таким образом, один из параметров (m или δ) можно задавать по своему усмотрению, второй — определяется из неравенства. При этом нужно принять в расчет следующее соображение.

Если $\delta \ll 1$, то существенно ухудшается оценка погрешности для любой квадратурной формулы, которую мы планируем использовать для вычисления J_2, так как в качестве коэффициента при h^p (где p — порядок точности выбранной формулы) фигурирует максимальное на $[\delta, 1]$ значение p-й производной от подынтегральной функции, которое при $x = \delta$ в нашем случае имеет порядок $\delta^{-(p+1/2)}$. Следовательно, целесообразно задавать «не слишком малое» δ (например, $\delta = 0.1$), а m найти затем из приведенного выше неравенства.

Замечание. Конечно, если поиск последовательных членов разложения подынтегральной функции затруднен, то приходится ограничиваться доступными членами, а из упомянутого неравенства определять δ. ▲

Пример. Вычислим интеграл второго типа: $\int\limits_0^\infty e^{-x^2}\,dx$. Можно, как отмечалось, свести его к интегралу первого типа. Но мы применим непосредственно универсальный прием выделения особенности. (Здесь особенность в том, что верхний предел интегрирования — бесконечность.) Представим интеграл в виде суммы двух: $J = J_1 + J_2$; J_1 — интеграл по конечному отрезку $[a, A]$, а J_2 — по $[A, \infty]$. Вычисление J_1 при заданном A вопросов не вызывает. Выберем теперь A так, чтобы в пределах допустимой погрешности вторым интегралом можно было пренебречь, т. е. так, чтобы $|J_2| \leqslant \varepsilon/2$. Например, учитывая, что при $A \geqslant 1$

$$\int\limits_A^\infty e^{-x^2}\,dx \leqslant \int\limits_A^\infty xe^{-x^2}\,dx = \frac{1}{2}\,e^{-A^2}$$

и требуя, чтобы выполнялось $\frac{1}{2}e^{-A^2} \leqslant \frac{1}{2}\varepsilon$, найдем, что для этого достаточно взять $A \geqslant \sqrt{|\ln \varepsilon|}$.

Более подробные сведения по вопросам численного дифференцирования и интегрирования можно найти в следующих источниках: [1, с. 364–408], [3, т. 1, с. 217–330], [9, с. 70–125], [12, с. 161–186].

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Показать, что если табличные значения $f_k = f(x_k)$ функции $f(x)$ даны с постоянным шагом ($x_k = x_0 + kh$), то приближенную формулу для n-й производной на отрезке $[x_0, x_n]$ (полученную дифференцированием интерполяционного полинома n-й степени) можно представить в виде

$$\frac{d^n f}{dx^n} \approx \frac{1}{h^n} \sum_{k=0}^{n} (-1)^k C_n^k f_{n-k}.$$

2. По таблице $\{f_k = f(x_k),\ x_k = kh,\ k = 0, 1, \ldots, K\}$ построить приближенную формулу для вычисления третьей производной от функции $f(x)$ в точке $x = x_k$. Какова ее погрешность (получить выражение для главного члена погрешности)?

3. По таблице $\{f_k = f(x_k),\ x_k = kh,\ k = 0, 1, \ldots, K\}$ построить приближенные формулы для вычисления третьей производной от функции $f(x)$ в крайних точках таблицы: $x = x_0$ и $x = x_K$. Какова погрешность построенных формул (получить выражение для главного члена погрешности)?

4. По таблице $\{f_k = f(x_k),\ x_k = kh,\ k = 0, 1, \ldots, K\}$ построить формулу второго порядка точности (относительно h) для вычисления df/dx при $x = x_0$; ($x = x_K$). Обосновать (получить выражение для главного члена погрешности).

5. Дана таблица значений функции $f(x) : \{f_k = f(x_k),\ x_k = kh,\ k = 0, 1, \ldots, K\}$. Найти оптимальное значение шага h, при котором приближенная формула для первой производной от функции в точке $x = x_k$: $(f(x_k + h) - f(x_k - h))/(2h)$ будет иметь наилучшую точность в предположении, что ошибки табличных данных не превышают $e = 0.0001$, а $\max\limits_{[x_0, x_K]} |f'''(x)| \leqslant M_3$. Какова погрешность формулы для производной при оптимальном шаге?

6. Дана таблица значений функции $f(x) : \{f_k = f(x_k),\ x_k = kh,\ k = 0, 1, \ldots, K\}$. Найти оптимальное значение шага h, при котором приближенная формула для второй производной от функции в точке $x = x_k$: $(f(x_k + h) - 2f(x_k) + f(x_k - h))/h^2$ будет иметь наилучшую точность в предположении, что погрешность табличных данных $e = 0.0001$, а $\max\limits_{[x_0, x_K]} |f^{\mathrm{IV}}(x)| \leqslant M_4$. Какова погрешность при оптимальном шаге?

7. Для полиномов какой степени формула Симпсона дает точный результат?

8. Получить оценки погрешности рассмотренных здесь квадратурных формул.

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Задача Коши для системы уравнений первого порядка, разрешенных относительно производных. Методы Эйлера (явный и неявный). Представление о методах как о разностных схемах, аппроксимирующих исходную задачу. Метод Эйлера с пересчетом, предиктор-корректор, методы Рунге–Кутты, методы Адамса. Доказательство сходимости метода Эйлера.

Задача Коши. Мы приступаем к рассмотрению методов численного решения задач для обыкновенных дифференциальных уравнений. Начнем с задачи Коши для системы

$$\begin{cases} \dfrac{dy^{(i)}}{dx} - f^{(i)}(x, y^{(1)}, y^{(2)}, \ldots, y^{(n)}) = 0 \text{ для } a < x \leqslant b, \ i = 1, 2, \ldots, n, \\ y^{(i)}(a) = y_a^{(i)} \ \text{ — заданные значения.} \end{cases} \tag{7.1}$$

Система (7.1) может быть записана в компактной (векторной) форме

$$\begin{cases} \dfrac{d\mathbf{y}}{dx} - \mathbf{f}(x, \mathbf{y}) = 0, \quad a < x \leqslant b, \\ \mathbf{y}(a) = \mathbf{y}_a, \end{cases} \tag{7.1′}$$

где

$$\mathbf{y} = \begin{pmatrix} y^{(1)}(x) \\ y^{(2)}(x) \\ \ldots \\ y^{(n)}(x) \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f^{(1)}(x, y^{(1)}, \ldots, y^{(n)}) \\ f^{(2)}(x, y^{(1)}, \ldots, y^{(n)}) \\ \ldots\ldots\ldots\ldots \\ f^{(3)}(x, y^{(1)}, \ldots, y^{(n)}) \end{pmatrix}. \tag{7.2}$$

Замечание. В форме (7.1) может быть записана задача Коши для системы уравнений произвольного порядка, разрешенных относительно старших производных. Например, задача

$$\begin{cases} \dfrac{d^m y}{dx^m} - F\left(x, y, \dfrac{dy}{dx}, \ldots, \dfrac{d^m - 1y}{dx^m - 1}\right) = 0, \ a < x \leqslant b, \\ y(a) = \alpha_0, \\ \left.\dfrac{dy}{dx}\right|_{x=a} = \alpha_1, \\ \ldots\ldots\ldots\ldots \\ \left.\dfrac{d^m - 1y}{dx^m - 1}\right|_{x=a} = \alpha_{m-1} \end{cases}$$

сводится к виду (7.1), если ввести новые неизвестные функции:

$$y^{(1)} \equiv y, \quad y^{(2)} = \frac{dy^{(1)}}{dx} = \frac{dy}{dx},$$

$$y^{(3)} = \frac{dy^{(2)}}{dx} = \frac{d^2y}{dx^2}, \quad \ldots, \quad y^{(m)} = \frac{dy^{(m-1)}}{dx} = \frac{d^{(m-1)}y}{dx^{m-1}}. \qquad \blacktriangle$$

Мы будем предполагать, что на отрезке $[a, b]$ для задачи (7.1) выполнены условия существования и единственности решения (при анализе точности различных численных алгоритмов будем предполагать также необходимую гладкость функций $f^{(i)}$ по своим аргументам).

Обсуждение методов решения задачи Коши ради простоты будем проводить на примере задачи для одного уравнения

$$\begin{cases} \dfrac{dy}{dx} - f(x, y) = 0, \quad a < x \leqslant b, \\ y(a) = y_a. \end{cases} \tag{7.3}$$

При этом не будем забывать о том, что *любой из численных алгоритмов для решения задачи (7.3) без особых затруднений может быть применен для решения системы (7.1′)*, если правильно учесть векторный характер искомого решения и функций f, выраженный соотношениями (7.2).

Построение численных алгоритмов опирается на дискретизацию задачи. Введем в области расчета $x \in [a, b]$ дискретный набор точек $\omega^{(h)} = \{x_k = a + hk, \quad k = 0, 1, \ldots, K\,(Kh = b - a)\}$, в которых будет вычисляться приближенное решение. Точки x_k будем называть *узлами интегрирования* или *узлами сетки*, расстояние h между ними — *шагом интегрирования* или *шагом сетки*, а совокупность узлов $\omega^{(h)}$ — *сеточной областью* или *сеткой узлов*.

Другие обозначения, которыми мы будем далее пользоваться:

$y^{(h)} = \{y_k, \quad k = 0, 1, \ldots, K\}$ — совокупность искомых приближенных значений решения задачи (7.3) в узлах сетки;

$[y]^{(h)} = \{y(x_k), \quad k = 0, 1, \ldots, K\}$ — совокупность точных значений решения задачи (7.3) в узлах сетки (проекция решения исходной задачи на сеточную область);

$f^{(h)} = \{f_k = f(x_k, y_k), \quad k = 0, 1, \ldots, K\}$ — значения правой части (7.3) в узлах.

Различные совокупности величин, отнесенных к узлам сетки, будем называть *сеточными функциями*. Очевидно, введенные таким образом сеточные функции можно трактовать как элементы $(K + 1)$-мерного векторного пространства. Опираясь на это представление,

определим погрешность численного решения:

$$\delta^{(h)} = y^{(h)} - [y]^{(h)} = \{y_k - y(x_k), \quad k = 0, 1, \ldots, K\}$$

и, привлекая какую-либо норму в упомянутом векторном пространстве, оценим величину погрешности как $\|\delta^{(h)}\|$. (Например, $\|\delta^{(h)}\| = \max\limits_k |y_k - y(x_k)|$.)

Определения. Будем говорить, что численное решение сходится к точному ($y_k \to y(x_k)$), если $\|\delta^{(h)}\| \underset{h\to 0}{\longrightarrow} 0$.

Будем также говорить, что метод, по которому получено численное решение, является методом p-го порядка точности, если

$$\|\delta^{(h)}\| \leqslant \text{const} \cdot h^p.$$

Переходим теперь к обсуждению конкретных методов вычисления решения задачи (7.3) в узлах сетки. Простейший способ их конструирования опирается на замену производной в уравнении (7.3) в окрестности каждого узла сетки по формулам численного дифференцирования, использующим значения искомого решения в узлах сетки.

Метод Эйлера (явный). Приближая производную в окрестности каждого k-го узла правым разностным отношением, приходим к методу Эйлера:

$$\begin{cases} \dfrac{y_{k+1} - y_k}{h} - f(x_k, y_k) = 0, \quad (k = 0, 1, \ldots, K-1), \\ y_0 = y_a. \end{cases} \tag{7.4}$$

Очевидно, в данном случае искомая интегральная кривая AB приближается ломаной $ACDEF$ (рис. 7.1), наклон которой на элементарном участке $[x_k, x_{k+1}]$ определяется наклоном интегральной кривой уравнения (7.3), выпущенной из точки (x_k, y_k). Последовательные значения y_k вычисляются по формуле $y_{k+1} = y_k + hf_k$, которая немедленно получается из верхнего соотношения (7.4).

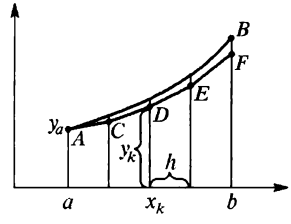


Рис. 7.1.

Метод Эйлера (неявный). К этому методу приходим, приближая производную в окрестности k-го узла левым разностным отношением

$$\begin{cases} \dfrac{y_k - y_{k-1}}{h} - f(x_k, y_k) = 0, \quad (k = 1, 2, \ldots, K) \\ y_0 = y_a. \end{cases} \tag{7.4'}$$

При вычислении решения по этому методу возникают некоторые трудности, ибо с учетом «направления счета» (слева направо) неизвестная величина (y_k) в каждое (k-е) уравнение входит, вообще говоря, нелинейным образом. Но трудности эти непринципиальны, достаточно вспомнить о методах решения нелинейных уравнений. (Например, можно предложить следующий итерационный процесс для вычисления решения в очередном (k-м) узле:

$$y_k^{(s+1)} = y_{k-1} + hf(x_k, y_k^{(s)}) = \varphi(y_k^{(s)}),\ y_k^{(0)} = y_{k-1}.$$

Здесь s — номер приближения. Достаточное условие сходимости итераций

$$|\varphi'| = h|f'_y|\Big|_{x=x_k\ y\sim y_k} < 1$$

удовлетворяется, очевидно, при достаточно малом h.)

Замечание 1. Алгебраические соотношения между компонентами сеточной функции, которыми заменяются исходные дифференциальные уравнения в окрестности каждого узла сетки, будем называть впредь *разностными уравнениями (соотношениями)*.

Замкнутую систему разностных уравнений вместе с дополнительными условиями (начальными или краевыми) называют *разностной схемой*. Таким образом, метод (7.4) — это явная разностная схема Эйлера, (7.4′) — неявная разностная схема Эйлера. ▲

Замечание 2. Приведем дополнительный комментарий к терминам «явный» («неявный») метод (схема). Метод называется явным, если система уравнений определяющая этот метод может быть записана непосредственно в виде расчетных формул для вычисления приближенных значений решения в узлах сетки. Если же дело сводится к необходимости решать систему уравнений (линейных или нелинейных), схему называют неявной. ▲

Обратимся к вопросу о погрешности численного решения. Что касается первых двух методов, которые введены в рассмотрение, то погрешность их можно оценить непосредственно (см. дополнения к данной лекции). Однако для многих других методов это затруднительно. И потому мы, опираясь на явный метод Эйлера, изложим пока предварительные соображения, касающиеся общего подхода к анализу точности численных результатов.

Используя определение погрешности, выразим вычисляемое решение через точное:

$$y_k = y(x_k) + \delta_k,\quad (k = 0, 1, \ldots, K).$$

Подставляя это выражение в (7.4), получаем

$$\frac{y(x_{k+1}) - y(x_k)}{h} + \frac{\delta_{k+1} - \delta_k}{h} - f(x_k, y(x_k) + \delta_k) = 0.$$

По теореме Лагранжа (из математического анализа)

$$f(x_k, y(x_k) + \delta_k) = f(x_k, y(x_k)) + (\widetilde{f}_y)_k \delta_k,$$

где $(\widetilde{f}_y{}')_k$ — производная от f по y в точке $(x_k, \widetilde{y})$ такой, что $\widetilde{y} \in [y(x_k),\ y(x_k)+\delta_k]$. Перепишем предыдущее соотношение в виде

$$\frac{\delta_{k+1} - \delta_k}{h} - (\widetilde{f}_y)_k \delta_k = -\left[\frac{y(x_{k+1}) - y(x_k)}{h} - f(x_k, y(x_k))\right]. \qquad (7.5)$$

Мы получили разностные уравнения, которые описывают поведение погрешности метода Эйлера при переходе от узла к узлу. Определяющее значение при этом имеют правые части этих уравнений, представляющие собой результат подстановки точного решения исходной задачи (7.3) в разностные уравнения метода Эйлера (7.4).

Решением системы уравнений (7.4) является сеточная функция $y^{(h)}$. При подстановке в разностные уравнения какой-то другой сеточной функции (например, $[y]^{(k)}$) уравнения не удовлетворяются, возникает так называемая *невязка*, или *погрешность аппроксимации*. В рассматриваемом случае ошибка аппроксимации

$$\psi^{(h)} = \left\{\psi_k = \frac{y(x_{k+1}) - y(x_k)}{h} - f(x_k, y(x_k)),\ k = 0, 1, \ldots, K-1\right\}. \qquad (7.6)$$

Как видно, именно эти величины находятся в правых частях уравнений (7.5).

Ошибка аппроксимации в каждом узле (ψ_k) в определенной степени показывает, насколько разностное уравнение в окрестности данного узла отличается от исходного, дифференциального, а величина погрешности $\|\psi^{(h)}\|$ характеризует в целом «близость» исходной задачи (7.3) и соответствующей разностной, из которой находится приближенное решение $y^{(h)}$ (в данном случае задачи (7.4)). Важно то, что (в Лекции 8 мы докажем соответствующую теорему) погрешность численного решения, получаемого по какому-либо сходящемуся методу, определяется погрешностью аппроксимации разностных уравнений, соответствующих этому методу.

Величину ошибки аппроксимации нетрудно оценить. Например, для метода Эйлера (7.4) согласно (7.6)

$$\psi_k = \left[\frac{y(x_{k+1}) - y(x_k)}{h} - f(x_k, y(x_k))\right].$$

Предполагая необходимую (для нижеследующих выкладок) гладкость решения исходной задачи, заменим значение функции $y(x_{k+1})$ по формуле Тейлора относительно узла x_k:

$$y(x_{k+1}) = y(x_k) + hy'(x_k) + \frac{h^2}{2}\widetilde{y}_k'' \quad \left(\widetilde{y}_k'' = y''\big|_{x\in[x_k, x_k+h]}\right).$$

Подставляя в формулу для ψ_k, получаем после элементарных арифметических преобразований

$$\psi_k = [y'(x_k) - f(x_k, y(x_k))] + \frac{h}{2}\widetilde{y}_k''.$$

Или, согласно уравнению (7.3):

$$\psi_k = \frac{h}{2}\widetilde{y}_k'',$$

в предположении, что y'' существует и ограничена на $[a, b]$. Стало быть, $\|\psi^{(h)}\| \leqslant \dfrac{M_2}{2}h$, где $M_2 = \max\limits_{[a,b]}|y''|$ (если использовать равномерную метрику). Следовательно, явный метод Эйлера представляет собой метод первого порядка точности. Аналогично устанавливается, что метод (7.4′) также имеет первый порядок точности.

Вернемся к обсуждению других методов численного решения задачи (7.3).

Приблизив производную $\frac{dy}{dx}$ в окрестности k-го узла с помощью центрального разностного отношения, приходим к системе соотношений

$$\begin{cases} \dfrac{y_{k+1} - y_{k-1}}{2h} - f(x_k, y_k) = 0, \quad k = 1, 2, \ldots, K-1, \\ y_0 = y_a. \end{cases} \tag{7.7}$$

Легко убедиться, что верхние соотношения (7.7) аппроксимируют исходное дифференциальное уравнение со вторым порядком точности (т. е. $|\psi_k| \leqslant \text{const} \cdot h^2$ для всех k). Однако, система (7.7) пока не замкнута — необходимо каким-то образом доопределить значение y_1. Делая это, например, с помощью метода Эйлера: $y_1 = y_0 + hf(x_0, y_0)$ и записывая уравнения (7.7) в виде $y_{k+1} = y_{k-1} + 2hf(x_k, y_k)$, получаем расчетную формулу, по которой можно вычислить решение в каждой узловой точке.

Чтобы понять, как можно строить иные методы второго порядка точности, обратимся к рис. 7.2. Здесь в пределах отрезка $[x_k, x_{k+1}]$ в утрированно увеличенном масштабе изображена ин-

тегральная кривая OP, выпущенная из точки O с координатами (x_k, y_k); y_A — значение при (x_{k+1}), которое получается по явному методу Эйлера; y_B — значение, вычисляемое по неявному методу Эйлера (звено соответствующей ломаной выпускается из точки O с наклоном, равным наклону интегральной кривой в точке P); y_C — значение, которое соответствует пересечению $x = x_{k+1}$ с прямой, выпущенной из точки O с наклоном, равным наклону интегральной кривой в середине отрезка $[x_k, x_{k+1}]$. Исходя из приведенной картинки можно предположить, что точность y_C больше, нежели y_A или y_B. Опираясь на это предположение, сконструируем следующие расчетные алгоритмы.

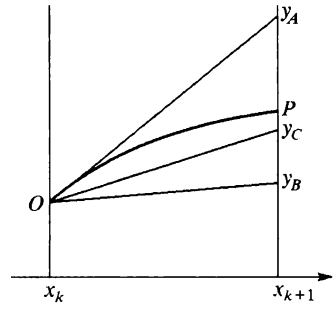


Рис. 7.2.

Модифицированный метод Эйлера. Запишем систему уравнений

$$\frac{y_{k+1} - y_k}{h} - f_{k+1/2} = 0 \quad (f_{k+1/2} = f(x_k + \frac{h}{2}, y_{k+1/2})); \qquad y_0 = y_a, \tag{7.8}$$

при этом значение $y_{k+1/2}$, относящееся к середине отрезка $[x_k, x_{k+1}]$, приближенно вычисляется по методу Эйлера

$$y_{k+1/2} = y_k + \frac{h}{2} f(x_k, y_k), \; k = 0, 1, \ldots, K-1, \tag{7.8'}$$

Исключая из (7.8) промежуточное значение $y_{k+1/2}$, можно записать этот метод в виде следующей системы разностных уравнений:

$$\frac{y_{k+1} - y_k}{h} - f(x_k + \frac{h}{2}, y_k + \frac{h}{2} f(x_k, y_k)) = 0, \quad k = 0, 1, \ldots, K-1; \qquad y_0 = y_a. \tag{7.9}$$

Нетрудно убедиться, что система (7.9) аппроксимирует задачу (7.3) со вторым порядком точности.

Метод Эйлера с пересчетом. Другой способ расчета, опирающийся на изложенные выше геометрические соображения, состоит в замене исходной задачи (7.3) системой

$$\begin{cases} \dfrac{y_{k+1} - y_k}{h} - \dfrac{1}{2}(f(x_{k+1}, y_{k+1}) + f(x_k, y_k)) = 0, \; k = 0, 1, \ldots, K-1, \\ y_0 = y_a. \end{cases} \tag{7.10}$$

Здесь, как видно, наклон интегральной кривой посредине отрезка $[x_k, x_{k+1}]$ приближенно заменяется средним арифметическим наклонов на границах этого отрезка. Снова нетрудно проверить, что (7.10) аппроксимирует задачу (7.3) со вторым порядком точности.

Как вычислять решение в этом случае? Система (7.10) является *неявной*, т. е. каждое уравнение нельзя сразу записать в виде расчетной формулы, так как неизвестные входят в уравнения, вообще говоря, нелинейным образом. Но, так же как и для (7.4′), можно обратиться, например, к итерационному процессу (при вычислении y_{k+1} из k-го уравнения)

$$y_{k+1}^{(s+1)} = y_k + \frac{h}{2}\left(f(x_{k+1}, y_{k+1}^{(s)}) + f(x_k, y_k)\right) \tag{7.11}$$

(s — номер приближения).

Остановимся чуть подробнее на следующем варианте этого метода, вытекающем из итерационного процесса (7.11). Вычислим начальное («нулевое») приближение $y_{k+1}^{(0)}$ по явному методу Эйлера

$$y_{k+1}^{(0)} = y_k + hf(x_k, y_k). \tag{7.11′}$$

Для первого приближения, согласно (7.11), получаем формулу

$$y_{k+1} = y_k + \frac{h}{2}\left(f(x_{k+1}, y_{k+1}^{(0)}) + f(x_k, y_k)\right). \tag{7.11″}$$

Если с помощью (7.11′) исключить из (7.11″) $y_{k+1}^{(0)}$, получим

$$y_{k+1} = y_k + \frac{h}{2}\left[f(x_{k+1}, y_k + hf(x_k, y_k)) + f(x_k, y_k)\right],$$

или

$$\frac{y_{k+1} - y_k}{h} = \frac{1}{2}\left[f(x_{k+1}, y_k + hf(x_k, y_k)) + f(x_k, y_k)\right]. \tag{7.12}$$

Можно убедиться, что уравнения (7.12) аппроксимируют исходное дифференциальное уравнение (7.3) со вторым порядком точности. Таким образом, в рамках метода итераций (7.11) второй порядок точности достигается уже на первой итерации, если нулевое приближение вычислять по методу Эйлера (7.11′).

Алгоритм вычислений, описываемый формулами (7.11′) и (7.11″), называется методом *предиктор-корректор*: предиктор (предсказание результата) — $y_{k+1}^{(0)}$, вычисляемое по формуле (7.11′), корректор (уточнение результата) — y_{k+1}, формула (7.11″).

Замечание. Методы (7.8), (7.8′) и (7.11′), (7.11″) являются частными вариантами однопараметрического семейства схем Рунге–Кутты второго порядка точности (см. дополнения к данной лекции). ▲

Метод Рунге–Кутты четвертого порядка точности (без вывода). В заключение приведем расчетные формулы одного из наиболее часто используемых для численного решения задачи Коши метода.

При вычисленном значении y_k расчет y_{k+1} осуществляется по формулам

$$\begin{cases} y_{k+1} = y_k + \dfrac{h}{6}(p_1 + 2p_2 + 2p_3 + p_4), \\ p_1 = f(x_k, y_k), \\ p_2 = f(x_k + \dfrac{h}{2}, y_k + \dfrac{h}{2}p_1), \\ p_3 = f(x_k + \dfrac{h}{2}, y_k + \dfrac{h}{2}p_2), \\ p_4^{(k)} = f(x_k + h, y_k + hp_3), \end{cases} \tag{7.13}$$

где p_i — вспомогательные величины.

ДОПОЛНЕНИЯ К ЛЕКЦИИ 7

О сходимости явного метода Эйлера. Согласно (7.4), приближенное решение в узле x_{k+1} связано с решением в узле x_k формулой

$$y_{k+1} = y_k + hf(x_k, y_k). \tag{7.14}$$

Представим точное решение задачи (7.3) в узле x_{k+1} по формуле Тейлора относительно узла x_k:

$$y(x_{k+1}) = y(x_k) + h\left.\frac{dy}{dx}\right|_{x_k} + \frac{h^2}{2}\left.\frac{d^2y}{dx^2}\right|_{\widetilde{x}}, \tag{7.15}$$

где $\widetilde{x} \in [x_k, x_{k+1}]$. Согласно уравнению (7.3) $\left.\dfrac{dy}{dx}\right|_{x_k} = f(x_k, y(x_k))$, $\left.\dfrac{d^2y}{dx^2}\right|_{\widetilde{x}} = \widetilde{f}'\big|_{\widetilde{x}} = (f_x + y'f_y)\big|_{\widetilde{x}} = (f_x + ff_y)\big|_{\widetilde{x}}$. Вычитая (7.15) из (7.14), получим соотношение между погрешностями в соседних точках:

$$\delta_{k+1} = \delta_k + h[f(x_k, y_k) - f(x_k, y(x_k))] - \frac{h^2}{2}\widetilde{f}'. \tag{7.16}$$

Напомним, что согласно предположению о том, что в области расчета выполнены условия существования и единственности решения (7.3), имеем

$$|f| < C,\ |f_y| < C_2 \text{ при любых } x \in [a, b].$$

Предположим, кроме того, что $|\widetilde{f_x}| < C_1$. Тогда $|\widetilde{f'}| \leqslant |\widetilde{f_x}| + |\widetilde{f}| \cdot |\widetilde{f_y}| \leqslant C_1 + CC_2 = C_3$, а согласно теореме Лагранжа о среднем $|f(x_k, y_k) - f(x_k, y(x_k))| \leqslant C_2|\delta_k|$.

Оценивая δ_{k+1}, определяемое формулой (7.16), по абсолютной величине, получим

$$|\delta_{k+1}| \leqslant (1 + C_2 h)|\delta_k| + C_3 \frac{h^2}{2} \leqslant$$

(в силу рекуррентности)

$$\leqslant (1 + C_2 h)^2 |\delta_{k-1}| + (1 + C_2 h) C_3 \frac{h^2}{2} + C_3 \frac{h^2}{2} \leqslant \ldots \leqslant$$

$$\leqslant (1 + C_2 h)^{k+1} |\delta_0| + \frac{C_3 h^2}{2} \left[\sum_{m=0}^{k} (1 + C_2 h)^m \right] =$$

(используя формулу для суммы геометрической прогрессии и заменяя $k + 1$ на $\frac{x_{k+1}}{h}$)

$$= (1 + C_2 h)^{\frac{x_{k+1}}{h}} |\delta_0| + \frac{C_3 h^2}{2} \left[\frac{(1 + C_2 h)^{k+1} - 1}{1 - (1 + C_2 h)} \right] =$$

$$= (1 + C_2 h)^{\frac{x_{k+1}}{h}} |\delta_0| + \frac{C_3 h}{2C_2} \left[(1 + C_2 h)^{\frac{x_{k+1}}{h}} - 1 \right].$$

Отсюда получаем оценку

$$|\delta_k| \leqslant e^{C_2(b-a)} |\delta_0| + \frac{C_3 h}{2C_2} \left(e^{C_2(b-a)} - 1 \right), \tag{7.17}$$

справедливую при любом k. Естественно считать, что $|\delta_0| = 0$ (поскольку y_0 — заданное значение). Тогда из (7.17) следует вывод о сходимости при $h \to 0$ и о первом порядке точности метода Эйлера, так как $\|\delta^{(h)}\| = \max\limits_k |\delta_k| \leqslant \text{const} \cdot h$.

Однопараметрическое семейство методов Рунге–Кутты второго порядка точности. Обобщая записи введенных в этой лекции алгоритмов (7.9) и (7.12), рассмотрим способ перехода от решения в k-м

узле к решению в $(k+1)$-м, представляемый разностным уравнением

$$\frac{y_{k+1}-y_k}{h} - \alpha f(x_k, y_k) - \beta f(x_k + \gamma h,\ y_k + \delta h f(x_k, y_k)) = 0, \quad (7.18)$$

где α, β, γ, δ — пока неопределенные параметры, которые найдем из требования, чтобы соотношение (7.18) аппроксимировало исходное дифференциальное уравнение (7.3) со вторым порядком точности.

Итак, ошибка аппроксимации в k-м узле как результат формальной подстановки решения исходной задачи (7.3) в разностное уравнение (7.18) имеет вид

$$\psi_k = \frac{y(x_{k+1}) - y(x_k)}{h} - \alpha f(x_k, y(x_k)) - \\ - \beta f(x_k + \gamma h, y(x_k) + \delta h f(x_k, y(x_k)))$$

(под $y(x)$ понимается точное решение исходной задачи).

Выбирая в качестве «опорной» точки $x = x_k$ и заменяя все слагаемые в выражении для ψ_k по формуле Тейлора относительно этой точки, получим (аргумент x_k далее опущен ради компактности последующих записей):

$$\psi_k = \frac{y + hy' + \frac{h^2}{2}y'' + \frac{h^3}{6}\widetilde{y}''' - y}{h} - \alpha f - \\ - \beta\left[f + \gamma h f_x + (\delta h f) f_y + \frac{\gamma^2 h^2}{2}\widetilde{f}_{xx} + \frac{\delta^2 h^2 f^2}{2}\widetilde{f}_{yy} + \gamma\delta h^2 f \widetilde{f}_{xy}\right].$$

Знак $\sim$ над обозначением функций означает, что указанные производные от них вычисляются в окрестности точки x_k для $\widetilde{y}'''$ и точки $(x_k, y(x_k))$ для производных от функции f.

Далее,

$$\psi_k = y' + \frac{h}{2}y'' + \frac{h^2}{6}\widetilde{y}''' - (\alpha+\beta)f - \beta h(\gamma f_x + \delta f f_y) - \\ - \frac{\beta h^2}{2}(\gamma^2 \widetilde{f}_{xx} + \delta^2 \widetilde{f}^2 \widetilde{f}_{yy} + 2\gamma\delta \widetilde{f}\widetilde{f}_{xy}). \quad (7.19)$$

Заметим, что в силу исходного уравнения

$$y' - f = 0, \quad y'' - (f_x + f f_y) = 0, \\ y''' - [(f_{xx} + 2f f_{xy} + f^2 f_{yy}) + (f_x f_y + f f_y^2)] = 0. \quad (7.20)$$

Учитывая (7.20), видим, что ψ_k будет величиной второго порядка малости, если

$$\alpha + \beta = 1,\ \beta\gamma = \beta\delta = \frac{1}{2}.$$

Исходя из этих трех условий для четырех параметров, выразим, например, α, γ, δ через β:

$$\alpha = (1 - \beta),\ \gamma = \delta = \frac{1}{2\beta},$$

(сопоставление последнего соотношения из (7.20) с соответствующими слагаемыми из (7.19) приводит к выводу, что ни при каком β аннулировать члены второго порядка малости в (7.19) не удается).

Таким образом, мы пришли к однопараметрическому семейству схем (методов) второго порядка точности, называемых *методами Рунге–Кутты*:

$$\frac{y_{k+1} - y_k}{h} = (1 - \beta)p_1 + \beta p_2, \tag{7.21}$$

где вспомогательные значения p_1, p_2 вычисляются по формулам:

$$p_1 = f(x_k, y_k),\ p_2 = f\left(x_k + \frac{h}{2\beta},\ y_k + \frac{h}{2\beta}p_1\right). \tag{7.21′}$$

Нетрудно видеть, что при $\beta = 1$ (7.21)–(7.21′) переходят в формулы модифицированного метода Эйлера (7.8)–(7.8′), а при $\beta = \frac{1}{2}$ в формулы метода предиктор-корректор (7.11′)–(7.11″).

З а м е ч а н и е. Похожим образом можно получить формулы Рунге–Кутты высших порядков точности (в том числе формулы (7.13)). ▲

Представление о многоточечных методах. Многоточечные (иногда их называют *многошаговыми*) методы решения задачи Коши характерны тем, что вычисляемое значение решения в текущем узле зависит от данных не только в одном предыдущем узле, но и в ряде предшествующих. Мы уже сталкивались с методом подобного рода (см. (7.7)). Отталкиваясь от этого примера, легко понять, что можно получить другие варианты многошаговых методов, если производную в исходном уравнении (7.3) заменять по многоточечным формулам численного дифференцирования.

Можно также строить такого рода методы, используя метод неопределенных коэффициентов. Запишем разностное соотношение

в окрестности k-го узла в виде

$$\frac{a_0 y_k + a_1 y_{k-1} + \ldots + a_m y_{k-m}}{h} - (b_0 f_k + b_1 f_{k-1} + \ldots + b_l f_{k-l}) = 0, \tag{7.22}$$

а коэффициенты $a_0, a_1, \ldots, a_m$, $b_0, b_1, \ldots, b_l$ подберем, требуя, чтобы уравнение (7.22) аппроксимировало исходное дифференциальное уравнение с максимально возможным порядком точности.

Если положить, $b_0 = 0$ будем иметь *явный* метод (соотношение (7.22) можно непосредственно записать в виде расчетной формулы для y_k.

При $b_0 \neq 0$ метод (7.22) является *неявным*.

Если $a_0 = -a_1 = 1$, $a_2 = a_3 = \ldots = a_m = 0$, то соотношения (7.22) называют *методами Адамса*.

Коснемся коротко подхода, который позволяет конструировать методы Адамса различной точности. Заметим, что решение уравнения $\frac{dy}{dx} = f(x, y)$ удовлетворяет интегральному соотношению

$$y_{k+1} - y_k = \int_{x_k}^{x_{k+1}} f\,dx. \tag{7.23}$$

Если решение в узлах вплоть до k-го уже вычислено, то по известным значениям $f_i^{(k)} = f(x_i, y_i)$, $i = k, k-1, \ldots$ можно интерполировать подынтегральную функцию полиномами различной степени. Вычисляя интеграл от выбранного интерполяционного полинома, будем получать различные формулы Адамса:

а) заменяя подынтегральную функцию ее значением в точке $x_k^{(k)}$ (полиномом нулевой степени), получим

$$\int_{x_k}^{x_{k+1}} f\,dx = f_k h + O(h^2)$$

или $y_{k+1} = y_k + h f_k$ — явный метод Эйлера;

б) $f\big|_{[x_k, x_{k+1}]} \approx f_{k+1}$, $\int_{x_k}^{x_{k+1}} f\,dx = f_{k+1} h + O(h^2)$,

$y_{k+1} = y_k + h f(x_k, y_{k+1})$ — неявный метод Эйлера;

в) $f\big|_{[x_k, x_{k+1}]} \approx f_{k+1/2} = f(x_k + \frac{h}{2}, y_{k+1/2})$,
$\int\limits_{x_k}^{x_{k+1}} f\,dx = f_{k+1/2}h + O(h^3)$, $y_{k+1} = y_k + hf_{k+1/2}$, доопределяя значение $y^{(k)}_{k+1/2}$ (например, как в (7.8′), приходим к модифицированному методу Эйлера;

г) $f\big|_{[x_k, x_{k+1}]} \approx f_k + \dfrac{f_{k+1} - f_k}{h}(x - x_k)$,
$\int\limits_{x_k}^{x_{k+1}} f\,dx = \dfrac{h}{2}(f_k + f_{k+1}) + O(h^3)$, $y_{k+1} = y_k + \dfrac{h}{2}[f_k + f_{k+1}]$ — получаем метод (7.10).

Замечание. Иногда его называют методом трапеций, что совершенно понятно в силу представленного здесь способа его получения. ▲

Пока это были примеры уже знакомых нам двухточечных методов (полученные новым способом). Посмотрим, как выглядят *многоточечные* методы, которые можно получить таким образом.

Приблизим функцию f на отрезке $[x_k, x_{k+1}]$ полиномом, записанным в форме Ньютона:

$$f \approx f_k + \frac{f_k - f_{k-1}}{h}(x - x_k) + \\ + \frac{f_k - 2f_{k-1} + f_{k-2}}{2h^2}(x - x_k)(x - x_{k-1}) + \ldots \qquad (7.24)$$

Учитывая первые два слагаемых при вычислении интеграла, получим $\int\limits_{x_k}^{x_{k+1}} f\,dx = h\left[\frac{3}{2}f_k - \frac{1}{2}f_{k-1}\right] + O(h^3)$. Отсюда приходим к методу второго порядка точности:

$$\frac{y_{k+1} - y_k}{h} - \frac{3}{2}f_k + \frac{1}{2}f_{k-1} = 0. \qquad (7.25)$$

Замечание. Вывод о точности метода ясен без проверки на аппроксимацию из самого способа получения соотношения (7.25) (и последующих). В самом деле, поскольку здесь решение задачи Коши рассматривается как вычисление интеграла, то погрешность метода не превосходит суммы локальных погрешностей использованных

квадратурных формул по всем элементарным отрезкам, на которые разбивается расчетная область. ▲

Учитывая в (7.24) три слагаемых, придем к методу третьего порядка точности

$$\frac{y_{k+1} - y_k}{h} = \frac{1}{12}(23f_k - 16f_{k-1} + 5f_{k-2}). \tag{7.26}$$

Наконец, так же может быть получен метод Адамса четвертого порядка

$$\frac{y_{k+1} - y_k}{h} = \frac{1}{24}(55f_k - 59f_{k-1} + 37f_{k-2} - 9f_{k-3}). \tag{7.27}$$

Любопытен вопрос: какой из двух теперь известных нам методов четвертого порядка точности предпочтительней — метод Адамса (7.27) или метод Рунге–Кутты (7.13)? При ответе на этот вопрос нужно принимать в расчет следующие соображения: метод Адамса требует меньших затрат (арифметических операций) при определении очередного значения y_{k+1}, так как при счете по формуле (7.27) нужно лишь один раз вычислять значение функции — f_k, другие требующиеся значения — $f_{k-1}, f_{k-2}, f_{k-3}$ — к этому моменту уже вычислены (достаточно их сохранять в памяти ЭВМ), в то время как определение y_{k+1} по формулам Рунге–Кутты требует в обязательном порядке вычислять четыре вспомогательных значения f (см. формулы (7.13)).

С другой стороны, чтобы начать вычисления по формулам Адамса, необходимо помимо заданного значения y_0 как-то определить (например, по тем же формулам Рунге–Кутты) значения y_1, y_2, y_3 в первых трех узлах интегрирования.

Кроме того (и это более важно), формулы Рунге–Кутты позволяют без затруднений проводить вычисления с переменным шагом интегрирования (например, с шагом, автоматически выбираемым из соображений требуемой точности), по формулам Адамса — это сложно.

Наконец, при использовании многошаговых методов есть бóльшая вероятность неблагоприятного поведения вычислительных погрешностей, так называемой *неустойчивости*. Но о проблемах устойчивости решения разностных уравнений мы будем говорить в следующих лекциях.

Дополнительные сведения по теме данной лекции можно найти в следующих источниках: [1, с. 410–482], [4, с. 71–170], [5, с. 25–34], [9, с. 237–260], [11, с. 174–211], [12, с. 214–258]. Достаточно полный обзор современных методов численного решения обыкновенных дифференциальных уравнений приводится в книге [32] (с приложением конкретных программ на Фортране).

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Найти погрешность аппроксимации уравнений (7.7).
2. То же для уравнений (7.9).
3. То же для уравнений (7.10).
4. То же для уравнений (7.12).
5. Выписать расчетные формулы метода Рунге–Кутты четвертого порядка точности для задач:

а) $\dfrac{dy}{dx} = f_1(x, y, z)$, $\dfrac{dz}{dx} = f_2(x, y, z)$, $a < x \leqslant b$; $y(a)$, $z(a)$ заданы;

б) $\dfrac{du}{dt} = f(t, u, v, w)$, $\dfrac{dv}{dt} = g(t, u, v, w)$, $\dfrac{dw}{dt} = r(t, u, v, w)$ при $0 < t \leqslant T$; $u(0)$, $v(0)$, $w(0)$ заданы.

Указание. В этих случаях, отталкиваясь от представления о векторном характере соотношений (7.13), надо правильно записать их скалярные следствия для вычисляемых компонент вектор-функции $\mathbf{Y}$ в рассматриваемых задачах:

а) $\mathbf{Y} = (y, z)$;

б) $\mathbf{Y} = (u, v, w)$.

ЛЕКЦИЯ 8

ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ (ПРОДОЛЖЕНИЕ)

Численное решение краевых задач. Линейный случай: непосредственная аппроксимация исходной задачи, сведение ее к решению последовательности задач Коши. Нелинейные задачи: прогонка с итерациями (для уравнения второго порядка), метод «пристрелки». Разностные схемы для обыкновенных дифференциальных уравнений. Аппроксимация, устойчивость, сходимость. Теорема о сходимости численного решения к решению исходной задачи. Элементы теории разностных уравнений. Примеры аналитических решений разностных задач. Примеры неустойчивого метода (разностной схемы) для задачи Коши.

В этой лекции мы остановимся на методах решения краевых задач для обыкновенных дифференциальных уравнений. Для начала рассмотрим задачу для линейного уравнения второго порядка

$$\begin{cases} y'' + p(x)y' + q(x)y = f(x), \quad a < x < b; \\ y(a) = y_a, \quad y(b) = y_b. \end{cases} \tag{8.1}$$

Далее всегда будем полагать, что решение этой и других краевых задач, которые мы будем рассматривать, существует и единственно.

Непосредственная разностная аппроксимация исходной краевой задачи. Линейный случай. Вводим в расчетной области (как мы это уже делали в Лекции 7) сетку узлов интегрирования:

$$\omega^{(h)} = \{x_k = a + kh, \quad k = 0, 1, \ldots, K; \; hK = (b - a)\}.$$

Точно так же под искомым решением будем понимать множество подлежащих вычислению значений решения в узлах сетки $\omega^{(h)}$, т. е. сеточную функцию

$$y^{(h)} = \{y_k, \quad k = 0, 1, \ldots, K\}.$$

Рассматривая дифференциальное уравнение (8.1) в окрестности k-го узла и заменяя производные по формулам численного дифференцирования сеточной (табличной) функции $y^{(h)}$, придем к следующей замкнутой системе линейных алгебраических уравнений для компонент

$y^{(h)}$ (разностной схеме):

$$\begin{cases} y_0 = y_a, \\ \dfrac{y_{k+1}-2y_k+y_{k-1}}{h^2} + p_k \dfrac{y_{k+1}-y_{k-1}}{2h} + q_k y_k = f_k, \quad k = 1, 2, \ldots, K-1, \\ y_K = y_b; \end{cases} \tag{8.2}$$

здесь $p_k = p(x_k)$, $q_k = q(x_k)$, $f_k = f(x_k)$.

Нетрудно видеть, что для внутренних узлов каждое уравнение задачи (8.2) приближает дифференциальное уравнение со вторым порядком точности. В самом деле, ошибка аппроксимации в k-м узле, в том смысле, как она была введена в прошлой лекции, составляет

$$\psi_k = \frac{y(x_{k+1}) - 2y(x_k) + y(x_{k-1})}{h^2} + p(x_k)\frac{y(x_{k+1}) - y(x_{k-1})}{2h} + \\ + q(x_k)y(x_k) - f(x_k),$$

где $y(x)$ — точное решение исходной задачи (8.1). Заменяя значения $y(x_{k+1})$ и $y(x_{k-1})$, входящие в выражение для ψ_k, по формулам Тейлора относительно k-го узла, получим

$$\psi_k = \frac{y+hy'+\frac{h^2}{2}y''+\frac{h^3}{6}y'''+\frac{h^4}{24}\widetilde{y}^{(\mathrm{IV})}-2y+y-hy'+\frac{h^2}{2}y''-\frac{h^3}{6}y'''+\frac{h^4}{24}\widetilde{\widetilde{y}}^{(\mathrm{IV})}}{h^2} + \\ + p\frac{y+hy'+\frac{h^2}{2}y''+\frac{h^3}{6}y'''_* - \left(y-hy'+\frac{h^2}{2}y''-\frac{h^3}{6}y'''_{**}\right)}{2h} + qy - f|_{x=x_k} = \\ = (y'' + py' + qy - f)\big|_{x=x_k} + \frac{h^2}{24}(\widetilde{y}^{(\mathrm{IV})} + \widetilde{\widetilde{y}}^{(\mathrm{IV})}) + p\frac{h^2}{12}(y'''_* + y''_{**}).$$

Под $\widetilde{y}^{(\mathrm{IV})}, \widetilde{\widetilde{y}}^{(\mathrm{IV})}, y'''_*, y'''_{**}$ понимаются соответствующие производные в точках из окрестности $[x_{k-1}, x_{k+1}]$. Для внутренних узлов в силу уравнения (8.1)

$$|\psi_k| \leqslant h^2\left(\frac{1}{12}\max_{[a,b]}|y^{(\mathrm{IV})}| + \frac{1}{6}\max_{[a,b]}|p(x)y'''|\right).$$

Следовательно, $\|\psi^{(h)}\| = \max\limits_k |\psi_k| \leqslant \mathrm{const} \cdot h^2$. И, принимая во внимание утверждение, высказанное в Лекции 7 (оно будет доказано ниже), о том, что точность численного решения при определенных условиях определяется ошибкой аппроксимации, можно рассчитывать на второй порядок точности метода (8.2) для решения задачи (8.1). Естественно, этот вывод справедлив только в предположении, что y''' и $y^{(\mathrm{IV})}$ существуют и ограничены на $[a, b]$.

З а м е ч а н и е. Если бы при конструировании разностных уравнений производная y' заменялась односторонним разностным отношением, полученная схема имела бы первый порядок аппроксимации. Вполне резонно использовать для приближения разных производных в исходном уравнении равноточные формулы, что мы и сделали выше. ▲

Отметим (без обоснования), что при достаточно малом h решение (8.2) существует и единственно. Подметив, кроме того, что система (8.2) является трехдиагональной, приходим к выводу, что естественным способом вычисления решения является метод прогонки. Условия, гарантирующие устойчивость счета по формулам прогонки $|b_k| > |a_k| + |c_k|$, применительно к системе уравнений вида $a_k x_{k-1} + b_k x_k + c_k x_{k-1} = f_k$ (см. Лекцию 2) в данном случае сводятся к требованию $|-2 + q_k h^2| > \left|1 + \frac{p_k}{2} h\right| + \left|1 - \frac{p_k}{2} h\right|$, которое, очевидно, выполняется для всех k, если $q_k < 0$ и шаг h достаточно мал, так что

$$\left|1 + \frac{p_k h}{2}\right| + \left|1 - \frac{p_k h}{2}\right| = \left(1 + \frac{p_k h}{2}\right) + \left(1 - \frac{p_k h}{2}\right) = 2.$$

(Заметим, что при $q(x) < 0$ частные решения уравнения (8.1) имеют экспоненциальный характер.)

Сведение решения линейной краевой задачи к решению задачи Коши. Из теории *линейных* дифференциальных уравнений следует, что общее решение уравнения (8.1) может быть представлено в виде

$$y(x) = \bar{y}(x) + C_1 Y_1(x) + C_2 Y_2(x), \tag{8.3}$$

где $\bar{y}(x)$ — частное решение неоднородного уравнения, а $C_1 Y_1(x) + C_2 Y_2(x)$ — общее решение однородного уравнения ($Y_1(x)$, $Y_2(x)$ — два линейно независимых решения; C_1, C_2 — произвольные постоянные).

Рассмотрим следующие задачи Коши:

$$\begin{cases} \bar{y}'' + p(x)\bar{y}' + q(x)\bar{y} = f(x), \quad a < x \leqslant b, \\ \bar{y}(a) = 0,\ \bar{y}'(a) = 0; \end{cases} \tag{8.4}$$

$$\begin{cases} Y_1'' + p(x)Y_1' + q(x)Y_1 = 0, \quad a < x \leqslant b, \\ Y_1(a) = 1,\ Y_1'(a) = 0; \end{cases} \tag{8.5}$$

$$\begin{cases} Y_2'' + p(x)Y_2' + q(x)Y_2 = 0, \ a < x \leqslant b, \\ Y_2(a) = 0, \ Y_2'(a) = 1. \end{cases} \tag{8.6}$$

Решив их численно с использованием любого из методов, обсуждавшихся в Лекции 7, мы будем располагать (в памяти ЭВМ) тремя массивами

$$\{\bar{y}_k, Y_{1k}, Y_{2k}, \quad k = 0, 1, \ldots, K\},$$

представляющими приближенно функции $\bar{y}(x), Y_1(x)$ и $Y_2(x)$, входящие в (8.3). При этом, в силу специального выбора начальных условий в задачах (8.5) и (8.6), массивы (векторы) $\{Y_{1k}\}$ и $\{Y_{2k}\}$ линейно независимы. Учитывая (8.3) и краевые условия из (8.1), получаем соотношения для C_1 и C_2:

при $x = a$ $\quad \bar{y}_0 + C_1(Y_1)_0 + C_2(Y_2)_0 = y_a$, откуда $C_1 = y_a$ (8.7)

при $x = b$ $\quad \bar{y}_K + C_1(Y_1)_K + C_2(Y_2)_K = y_b$. Следовательно,

$$C_2 = \frac{y_b - C_1(Y_1)_K - \bar{y}_K}{(Y_2)_K} = \frac{y_b - y_a(Y_1)_K - \bar{y}_K}{(Y_2)_K}. \tag{8.8}$$

Найденные значения C_1, C_2 позволяют из трех предварительно найденных массивов скомбинировать по формуле (8.3) результирующий массив, который будет представлять приближенное решение краевой задачи (8.1) в узлах сетки. Точность этого решения определяется точностью метода, использованного для решения задач Коши (8.4)–(8.6).

Замечание 1. В рамках данного подхода мы ищем функции, которые удовлетворяют нужным уравнениям. Выбор начальных данных в определенной степени произволен (лишь бы решения задач (8.5) и (8.6) были независимы).

Если в задаче (8.4)в качестве начального значения выбрать $\bar{y}(a) = y_a$, то из (8.7) следует, что $C_1 = 0$, т. е. искомое решение (8.3) будет зависеть от решения только двух вспомогательных задач Коши (8.4) и (8.6). Для C_2 из (8.8) тогда получим

$$C_2 = \frac{y_b - \bar{y}_K}{(Y_2)_K}. \ \blacktriangle \tag{8.9}$$

Замечание 2. Данный подход (сведение краевой задачи к последовательности задач Коши) может быть распространен на решение краевой задачи для линейного дифференциального уравнения произвольного порядка (или для системы линейных дифференциальных уравнений). $\blacktriangle$

Относительно сферы применимости такого подхода к решению краевой задачи (8.1) ограничимся общим утверждением, что при машинной реализации он может отказать, если частные решения исходного уравнения (которые, собственно, и находятся из вспомогательных задач Коши) представляют собой быстро растущие функции. В этом случае либо решение задачи Коши может оказаться нереализуемым из-за выхода вычисляемых значений за пределы диапазона представимых в ЭВМ чисел, либо значительная погрешность может возникнуть на стадии вычисления C_2 (за счет вычитания больших, близких по величине чисел в (8.8) или, соответственно, деления в (8.9)). Напомним, однако, что для задачи (8.1) частные решения могут иметь вид быстро растущих экспонент, когда $q(x) < 0$ и $|q(x)| \geqslant 1$, но в этом случае, как мы видели, решение может быть получено по первому методу (с использованием прогонки). В свою очередь, если $q(x) > 0$ и условие устойчивости вычислений по формулам прогонки не выполнено, то резонно использовать метод сведения исходной задачи к задаче Коши, как это было здесь описано.

Замечание 3. Приведенные выше (а также последующие) рассуждения о сфере применимости того или иного подхода при решении краевых задач, разумеется, имеют ориентировочный характер. Аргументированный выбор расчетного алгоритма зачастую требует более глубокого качественного анализа возможных решений рассматриваемой задачи. ▲

Непосредственная разностная аппроксимация дифференциального уравнения. Нелинейный случай. Посмотрим теперь, как эти подходы распространяются на нелинейный случай. Рассмотрим задачу

$$\begin{cases} y'' + \varphi(x, y, y') = 0, \quad a < x < b, \\ y(a) = y_a, \quad y(b) = y_b. \end{cases} \tag{8.10}$$

Непосредственная аппроксимация дифференциального уравнения разностными приводит к нелинейной разностной схеме:

$$\begin{cases} y_0 = y_a, \\ \dfrac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + \varphi\left(x_k, y_k, \dfrac{y_{k+1} - y_{k-1}}{2h}\right) = 0, \quad k = 1, 2, \ldots, K-1, \\ y_K = y_b. \end{cases} \tag{8.11}$$

Так же как и для линейной задачи, имеет место второй порядок аппроксимации. Для выбора конкретного способа расчета надо обратиться к методам решения нелинейных систем уравнений. Если, на-

пример, обозначить левую часть уравнений (8.11), умноженных на h^2, при $k = 1, \ldots, K-1$ через $F(y_{k+1}, y_k, y_{k-1})$, а через $y_k^{(s)}$ — s-е приближение для решения в k-м узле ($\delta y_k = y_k^{(s+1)} - y_k^{(s)}$ — разность между s-м и $(s+1)$-м приближениями), то метод Ньютона для (8.11) приведет к следующей системе линейных трехточечных соотношений для перехода из s-го к $(s+1)$-му приближению:

$$\begin{cases} \delta y_0 = 0, \\ \left(1 + \dfrac{h}{2}\varphi_{y'}\right)\delta y_{k+1} - (2 - h^2\varphi_y)\delta y_k + \left(1 - \dfrac{h}{2}\varphi_{y'}\right)\delta y_{k-1} = \\ \qquad = -F\left(y_{k-1}^{(s)}, y_k^{(s)}, y_{k+1}^{(s)}\right), \\ \delta y_K = 0. \end{cases} \tag{8.12}$$

Легко выясняются требования, которые гарантируют устойчивость прогонки.

Вот еще пример возможного итерационного процесса:

$$\begin{cases} y_0 = y_a, \\ \dfrac{y_{k+1}^{(s+1)} - 2y_k^{(s+1)} + y_{k-1}^{(s+1)}}{h^2} + \varphi\left(x_k, y_k^{(s)}, \dfrac{y_{k+1}^{(s)} - y_{k-1}^{(s)}}{2h}\right) = 0, \\ \qquad k = 1, 2, \ldots, K-1, \\ y_K = y_b. \end{cases} \tag{8.13}$$

Метод «пристрелки». В отличие от линейного случая теперь представление общего решения в виде (8.3) не имеет места. Тем не менее, метод редукции нелинейных краевых задач к последовательности задач Коши с успехом используется и применительно к задаче (8.10) сводится к следующим действиям. Заметим, что решение обсуждаемой задачи имеет простой геометрический смысл. Нужно найти интегральную кривую («траекторию»), проходящую через точки A, B (рис. 8.1). Если в (8.10)заменить условие $y(b) = y_b$ на $y'(a) = \alpha$, мы будем иметь задачу Коши, решить численно которую не составляет труда с помощью любого из методов, рассмотренных в Лекции 7 (кривая AB' на рис. 8.1). В силу известной из теории обыкновенных дифференциальных уравнений теоремы о непрерывной зависимости решения задачи

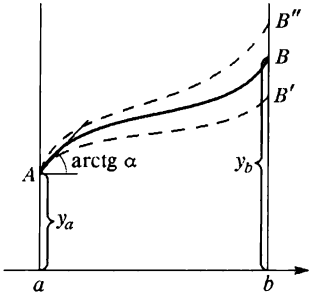


Рис. 8.1.

Коши от начальных данных вычисленное значение в точке $x = b$ является функцией заданного при $x = a$ параметра α: $y(b) \equiv F(\alpha)$. Возвращаясь к исходной задаче (8.10), с учетом этого обстоятельства нам надо подобрать такое значение α, чтобы удовлетворялось уравнение

$$F(\alpha) - y_b = 0. \tag{8.14}$$

Мы пришли к задаче вычисления корня нелинейного уравнения.

Необычность ситуации в том, что функция $F(\alpha)$ задана непривычным пока для нас образом — алгоритмически: чтобы найти значение функции F при заданном значении аргумента, надо решить задачу Коши

$$\begin{cases} y'' + \varphi(x, y, y') = 0, \; a < x \leqslant b, \\ y(a) = y_a, \; y'(a) = \alpha. \end{cases} \tag{8.15}$$

Но с точки зрения методов численного решения уравнений (Лекция 1) не важно, как задана функция, достаточно уметь вычислять ее значения. Если, например, из каких-то соображений (или в результате предварительных расчетов) известно, что искомое решение лежит между двумя кривыми AB' и AB'' с начальными наклонами α_0 и α_1, то простейшим методом решения уравнения (8.14) является метод половинного деления. Можно также использовать упоминавшуюся в Лекции 1 модификацию метода Ньютона:

$$\alpha_{n+1} = \alpha_n - \frac{F(\alpha_n) - y_b}{\dfrac{1}{\Delta}\left(F(\alpha_n + \Delta) - F(\alpha_n)\right)}, \quad \alpha_0 \text{ задано.} \tag{8.16}$$

(Производная $F'(\alpha_n)$ в формуле Ньютона здесь заменена простейшей формулой численного дифференцирования, Δ — малый параметр.) Очевидно, при использовании метода (8.16) вычисления корня (8.14) для перехода от n-го приближения к $(n+1)$-му необходимо два раза решить задачу Коши (8.15): один раз с $y'(a) = \alpha_n$ и другой — с $y'(a) = \alpha_n + \Delta$.

Описанный метод сведения нелинейной краевой задачи (8.10) к задачам Коши (метод «пристрелки») может быть распространен на другие задачи, в частности, для систем нелинейных уравнений (см. «Вопросы и упражнения»). Что касается сферы применимости метода «пристрелки», то, как и в линейном случае, здесь могут возникнуть непреодолимые трудности, если частные решения дифференциальных уравнений являются быстро растущими функциями. В этом случае предпочтительным может оказаться метод, основанный на непосредственной аппроксимации исходной задачи (о котором шла речь выше).

Аппроксимация. Устойчивость. Сходимость численного решения задач для дифференциальных уравнений. Мы обсудим здесь общие вопросы, касающиеся условий сходимости численного решения к решению исходной задачи. В частности, докажем теорему о сходимости, в которой эти условия формулируются. Рассмотрим задачу в абстрактной формулировке (привлечение которой придаст определенную общность последующим выводам).

Итак, пусть надо найти решение задачи

$$LU = F(x), \tag{8.17}$$

принадлежащее области $\overline{\omega} = \omega \cup \gamma$ (ω — внутренние точки расчетной области, γ — граничные). Мы будем считать, что решение задачи существует и единственно. Оператор L определяет вид дифференциальных уравнений в ω и вид дополнительных условий на γ.

Пример 1. Для задачи Коши

$$\begin{cases} \dfrac{dy}{dx} - f(x, y) = 0, \quad a < x \leqslant b, \\ y(a) = y_a \end{cases} \tag{8.18}$$

область $\overline{\omega}$ совпадает с отрезком $[a, b]$, граница γ состоит из одной точки $x = a$; оператор L можно определить так:

$$Ly = \begin{cases} \dfrac{dy}{dx} - f(x, y), & 0 < x \leqslant b, \\ y(a), & x = a; \end{cases}$$

правая часть

$$F(x) = \begin{cases} 0, & 0 < x \leqslant b, \\ y_a, & x = a. \end{cases}$$

Пример 2. Для краевой задачи (8.1) область $\overline{\omega}$ снова совпадает с отрезком $[a, b]$, граница γ состоит из двух точек, $x = a$ и $x = b$:

$$Ly = \begin{cases} \dfrac{d^2y}{dx^2} + p(x)\dfrac{dy}{dx} + q(x)y, \ a < x < b, \\ y(a), \ x = a, \\ y(b), \ x = b; \end{cases}$$

$$F(x) = \begin{cases} f(x), \ a < x < b, \\ y_a, \ x = a, \\ y_b, \ x = b. \end{cases}$$

Введем, как это делалось в примерах, рассмотренных в прошлой и настоящей лекциях, в области расчета $(\overline{\omega})$ совокупность расчетных точек (узлов интегрирования, сетку узлов): $\overline{\omega}^{(h)} = (\omega^{(h)} \cup \gamma^{(h)})$, $\omega^{(h)}$ — внутренние узлы, $\gamma^{(h)}$ — граничные узлы. Вместо исходной задачи (8.17) будем рассматривать задачу вычисления сеточной функции (вектора) $U^{(h)}$, определенной своими компонентами в узлах сетки, являющимися (в этих узлах) приближением для решения.

Пусть эта новая задача (разностная схема) записывается в виде

$$L_h U^{(h)} = F^{(h)}; \tag{8.19}$$

L_h — «разностный оператор», определяющий вид системы соотношений (8.19).

Пример 3. Явный метод Эйлера для задачи Коши (8.18):

$$\begin{cases} \dfrac{y_k - y_{k-1}}{h} - f(x_{k-1}, y_{k-1}) = 0, \quad k = 1, 2, \ldots, K, \\ y_0 = y_a. \end{cases} \tag{8.20}$$

Здесь $\overline{\omega}^{(h)} = \left\{x_k = a + kh, \quad k = 0, 1, 2, \ldots, K \ \left(h = \dfrac{b-a}{K}\right)\right\}$; $\gamma^{(h)}$ состоит из одной точки — x_0.

Оператор L_h может быть определен записью:

$$L_h y^{(h)} = \begin{cases} \dfrac{y_k - y_{k-1}}{h} - f(x_{k-1}, y_{k-1}), \quad k = 1, 2, \ldots, K, \\ y_0, \ k = 0. \end{cases}$$

Правые части $(F^{(h)})$:

$$F^{(h)} = \begin{cases} 0, \quad k = 1, 2, \ldots, K, \\ y_a, \quad k = 0. \end{cases}$$

Пример 4. Непосредственная разностная аппроксимация (8.2) краевой задачи (8.1):

$$\overline{\omega}^{(h)} = \left\{x_k = a + kh, \quad k = 0, 1, \ldots, K \ \left(h = \frac{b-a}{K}\right)\right\},$$

$$\gamma^{(h)} = \{x_0, x_K\},$$

$$L_h y^{(h)} = \begin{cases} y_0, & k = 0 \\ \dfrac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + p_k \dfrac{y_{k+1} - y_{k-1}}{2h} + q_k y_k, & 1 \leqslant k \leqslant K-1, \\ y_K, & k = K; \end{cases}$$

$$F^{(h)} = \begin{cases} y_a, & k = 0, \\ f_k, & 1 \leqslant k \leqslant K-1, \\ y_b, & k = K. \end{cases}$$

Погрешностью приближенного (разностного) решения $U^{(h)}$ называется функция, определенная в узлах сетки (сеточная функция) следующим образом: $\delta^{(h)} = U^{(h)} - [U]^{(h)}$,

где $[U]^{(h)}$ — совокупность значений точного решения задачи (8.17) в узлах сетки (проекция точного решения на сеточную область). Например, для задач (8.1), (8.2):

$$\delta^{(h)} = \{\delta_k, \quad k = 0, 1, \ldots, K\},$$

где $\delta_k = y_k - y(x_k)$ — погрешность в k-м узле.

Если при этом $\|\delta^{(h)}\| \leqslant \text{const} \cdot h^p$, то метод является методом p-го порядка точности (сходится со скоростью h^p). Так как функции, определенные на сеточной области, могут интерпретироваться как векторы, то, выбирая норму в соответствующем векторном пространстве, определим величину погрешности как $\|\delta^{(h)}\|$. Говорят, что имеет место *сходимость* численного решения к точному, если $\|\delta^{(h)}\| \underset{h \to 0}{\longrightarrow} 0$.

Мы уже привлекали в наших предыдущих обсуждениях представление о погрешности *аппроксимации*. Приведем теперь ее формальное (общее) определение:

$$\psi^{(h)} = L_h [U]^{(h)} - F^{(h)}. \tag{8.21}$$

Как видно из (8.21), ошибка аппроксимации — это совокупность «невязок», к которым приводят соотношения (8.19) при формальной подстановке в них точного решения задачи (8.17) (спроецированного на сеточную область). Величина погрешности аппроксимации — $\|\psi(h)\|$. (В Лекции 7 в отдельных случаях мы приводили примеры вычисления погрешности аппроксимации.) Будем говорить, что *метод* (разностная схема) (8.19) *аппроксимирует задачу* (8.17), если $\|\psi^{(h)}\| \underset{h \to 0}{\longrightarrow} 0$. Если

$$\|\psi^{(h)}\| \leqslant C h^p, \quad C = \text{const}, \tag{8.22}$$

то говорят, что имеет место *аппроксимация p-го порядка*.

Введем новое, важнейшее (особенно для дальнейших обсуждений) свойство численного решения.

Пусть решение задачи (8.19) существует и единственно для любых $F^{(h)}$, определенных на $\overline{\omega}^{(h)}$. Будем называть это решение *устойчивым* (соответственно схему (8.19) — устойчивой), если для любого $\varepsilon > 0$ найдется $h_0 > 0$, такое что для решения «возмущенной» задачи

$$L_h Z^{(h)} = F^{(h)} + \delta f^{(h)} \quad \text{с} \quad \|\delta f^{(h)}\| \leqslant \varepsilon \tag{8.23}$$

имеет место

$$\|Z^{(h)} - U^{(h)}\| \leqslant C_1 \|\delta f^{(h)}\| \tag{8.24}$$

при любом $0 \leqslant h \leqslant h_0$ с константой C_1, не зависящей от h.

Теперь мы можем сформулировать условия, при которых имеет место сходимость численного решения (8.19) к точному решению исходной задачи (8.17).

Теорема. *Пусть*:

1) задача (8.19) аппроксимирует задачу (8.17), причем имеет место оценка (8.22);

2) решение задачи (8.19) устойчиво, т. е. имеет место (8.24). Тогда численное решение $U^{(h)}$ сходится к решению (8.17), и имеет место следующая оценка погрешности

$$\|\delta^{(h)}\| \leqslant \text{const} \cdot h^p. \tag{8.25}$$

Доказательство. Согласно (8.19) $L_h U^{(h)} = F^{(h)}$, а ввиду (8.21) $L_h[U]^{(h)} = F^{(h)} + \psi^{(h)}$. Рассматривая в последнем соотношении $\psi^{(h)}$ как возмущение правой части, в силу (8.24) и (8.22) будем иметь

$$\|[U^{(h)}] - U^{(h)}\| \leqslant C_1 \|\psi^{(h)}\| \leqslant C_1 C h^p = \text{const} \cdot h^p.$$

Мы пришли к выводу (8.25), из которого следует сходимость.

Из оценки (8.25) следует, что метод (8.19) является *методом p-го порядка точности.*

Замечание 1. Доказанная теорема имеет важное (общее) значение, так как она определяет условия, достаточные для сходимости численного решения не только для задач с обыкновенными дифференциальными уравнениями, но и для задач с уравнениями в частных производных. (В этом смысле она послужит нам мостиком, преодолев который, мы приступаем — в следующей лекции — к обсуждению проблем численного решения уравнений в частных производных.) ▲

Замечание 2. Мы не акцентировали внимание на вопросах устойчивости численного решения задач для обыкновенных дифференциальных уравнений, где эти вопросы не очень актуальны

(см., впрочем, раздел «Пример неустойчивого метода» в дополнениях к данной лекции). Зато в последующем (для уравнений в частных производных) проблема устойчивости численного решения выйдет на первый план. ▲

Имея в виду, что данный курс лекций является, как было отмечено в предисловии, *введением в специальность*, здесь уместно сделать дополнительные комментарии в связи с доказанной теоремой.

Напомним, что с самого начала предполагалось, что исходная задача корректна, т. е. имеет единственное решение, которое устойчиво по отношению к возмущениям входных данных: при малых возмущениях последних решение меняется мало.

Почему же проблема устойчивости возникает заново при переходе к разностной задаче, которой мы заменяем исходную? Дело вот в чем.

Для исходной задачи значения решения в разных точках «жестко» связаны между собой дифференциальными уравнениями, которым решение обязано удовлетворять. После дискретизации, т.е. перехода к разностной схеме, значения разностного решения в каждом узле хотя и взаимосвязаны (через разностные уравнения), но каждое из них вычисляется индивидуально, отдельно! В расчет каждого привносится тем самым своя погрешность (за счет округлений при выполнении арифметических операций). Коррелируя друг с другом, эти погрешности при значительном количестве расчетных точек (узлов) могут быстро нарастать. Это и есть, грубо говоря, механизм возникновения неустойчивости при переходе к разностной задаче.

Повторяем, что особенно актуальными проблемы устойчивости становятся при численном решении уравнений в частных производных, и это понятно, так как сеточные области там становятся многомерными, каждый узел приобретает большее число соседних точек, соответственно повышается вероятность неблагоприятной корреляции ошибок в этих точках.

ДОПОЛНЕНИЕ К ЛЕКЦИИ 8

Простейшие элементы теории разностных уравнений. В Лекции 7 было введено представление о разностном уравнении как о соотношении, связывающем компоненты некоторой сеточной функции. В этом смысле запись метода Эйлера (8.20) представляет собой систему нелинейных разностных уравнений (компоненты сеточной функции входят в уравнения, в общем случае, нелинейно), а схема (8.2) является системой линейных разностных уравнений.

В этом разделе мы остановимся на некоторых вопросах, связанных с линейными разностными уравнениями. Приведем сначала формаль-

ное определение. Соотношение

$$\sum_{m=-M_1}^{M_2} a_m U_{k+m} = f_k, \tag{8.26}$$

где $k = 0, \pm 1, \pm 2, \ldots$; a_m — коэффициенты, не зависящие от U_k, (причем $a_{M_1}, a_{M_2} \neq 0$), называется *линейным разностным уравнением порядка* $M = M_1 + M_2$ для сеточной функции $U^{(h)} = \{U_k\}$. В общем случае коэффициенты a_m могут быть функциями индекса k. Если $a_m = \text{const}$, то (8.26) называется *уравнением с постоянными коэффициентами*; если $f_k \equiv 0$, то уравнение называется *однородным*. Например, уравнение

$$a_0 U_k + a_1 U_{k+1} = f_k \quad k = 0, \pm 1, \pm 2, \ldots \tag{8.27}$$

является линейным разностным уравнением первого порядка (заметим, что при каждом k в это уравнение входят значения сеточной функции с *двумя соседними индексами*), а уравнение

$$a_{-1} U_{k-1} + a_0 U_k + a_1 U_{k+1} = f_k \quad k = 0, \pm 1, \pm 2, \ldots \tag{8.28}$$

— уравнением второго порядка (при каждом k в него входят значения компонент сеточной функции с *тремя соседними индексами*).

Существует далеко идущая аналогия между теорией линейных разностных уравнений и теорией линейных дифференциальных уравнений. В частности, общее решение (8.26) может быть представлено в виде суммы частного решения неоднородного уравнения и общего решения однородного уравнения. Для однородного уравнения с постоянными коэффициентами общее решение представляет собой линейную комбинацию независимых решений, которые можно искать в виде

$$U_k = \lambda^k. \tag{8.29}$$

Пример 1. Для однородного уравнения первого порядка, подставляя (8.29) в (8.27) с $f_k \equiv 0$, получаем $a_0 \lambda^k + a_1 \lambda^{k+1} = 0$, откуда $\lambda = -a_0/a_1$. Таким образом, общее решение однородного уравнения (8.27) имеет вид

$$U_k = C\left(-\frac{a_0}{a_1}\right)^k. \tag{8.30}$$

Если при некотором значении $k = k_0$ значение U_{k_0} задано, то из совокупности (8.30) выделяется конкретное решение

$$U_k = U_{k_0}\left(-\frac{a_0}{a_1}\right)^{k-k_0},$$

удовлетворяющее заданному условию.

Пример 2. Для однородного уравнения второго порядка ((8.28) при $f_k \equiv 0$) подстановка (8.29) приводит к *характеристическому уравнению* для λ:

$$a_1\lambda^2 + a_0\lambda + a_{-1} = 0. \tag{8.31}$$

Двум различным корням этого уравнения соответствуют два линейно независимых решения (см. упражнения) λ_1^k и λ_2^k, а общее решение однородного уравнения может быть представлено в виде

$$U_k = C_1\lambda_1^k + C_2\lambda_2^k \tag{8.32}$$

с произвольными константами C_1 и C_2. Чтобы из (8.32) выделить конкретное решение (т. е. определить конкретные значения C_1 и C_2), надо задать значения компонент U_k в двух различных узлах.

Замечание. Если корни равны (кратный корень), то второе (независимое) решение ищется в специальном виде по аналогии с тем, как это делается для линейных дифференциальных уравнений с постоянными коэффициентами. ▲

Сравнение аналитических решений дифференциального и разностных уравнений. Из предыдущего раздела следует, что, если дифференциальная задача аппроксимируется системой линейных однородных разностных уравнений с постоянными коэффициентами, то решение последних может быть представлено в аналитической форме. Сравнивая это представление с формулой, дающей решение исходной дифференциальной задачи, мы получаем дополнительную возможность судить о свойствах рассматриваемого алгоритма численного решения.

Замечание. Следует подчеркнуть, что аналитическое выражение решения разностных уравнений — это не что иное, как (с точностью до машинных ошибок округления) формула для последовательных численных значений, получаемых при реализации алгоритма, представляемого системой этих разностных соотношений (разностной схемой). ▲

Рассмотрим следующую задачу Коши:

$$\begin{cases} y' + Ay = 0, \quad 0 < x \leqslant 1 \ (A > 0), \\ y(0) = 1. \end{cases} \tag{8.33}$$

Очевидно, решение (8.33): $y = e^{-Ax}$.

Пример 1. Пусть для численного решения задачи (8.33)на сетке $\{x_k = kh,\ k = 0, 1, \ldots, K;\ Kh = 1\}$ используется явный метод Эйлера

$$\begin{cases} \dfrac{y_{k+1} - y_k}{h} + Ay_k = 0, \quad k = 0, 1, \ldots, K-1, \\ y_0 = 1. \end{cases} \tag{8.34}$$

Верхние соотношения (8.34), если распространить их для $k = 0, \pm 1, \pm 2, \ldots$, представляют собой однородное разностное уравнение первого порядка с постоянными коэффициентами, которое мы запишем в виде $y_{k+1} - (1 - Ah)y_k = 0$. Общее его решение, согласно (8.30): $y_k = C(1 - Ah)^k$. Используя заданное при $k = 0$ условие $y_0 = 1$, находим аналитическое представление данных, вычисляемых по методу Эйлера (при $k = 0, 1, \ldots, K-1$):

$$y_k = (1 - Ah)^k = (1 - Ah)^{x_k/h}. \tag{8.35}$$

Из (8.35) следует, что $y_k \underset{h\to 0}{\longrightarrow} y(x_k) = e^{-Ax_k}$. Чтобы получить более подробное представление о характере сходимости, преобразуем выражение (8.35) (считая, что $Ah \ll 1$):

$$y_k = e^{(x_k/h)\ln(1-Ah)} = e^{(x_k/h)[-Ah - A^2h^2/2 + O(h^3)]} =$$

$$= e^{-Ax_k} \exp^{-A^2x_kh/2 + O(h^2)} = e^{-Ax_k}\left[1 - \frac{A^2x_kh}{2} + O(h^2)\right].$$

Отсюда следует, что

$$\delta_k = y_k - y(x_k) = e^{-Ax_k}\left[-\frac{A^2x_kh}{2} + O(h^2)\right], \tag{8.36}$$

т. е. метод Эйлера является методом первого порядка точности. Главный член погрешности решения в k-м узле определен формулой (8.36).

Пример 2. Рассмотрим теперь для решения задачи (8.33) метод (7.7) (Лекция 7):

$$\begin{cases} \dfrac{y_{k+1} - y_{k-1}}{2h} + Ay_k = 0, \quad k = 1, \ldots, K-1, \\ y_0 = 1, \\ y_1 = 1 - Ah. \end{cases}$$

Верхние соотношения, если распространить их для $k = 0, \pm 1, \pm 2, \ldots$, представляют собой разностное уравнение второго порядка

с постоянными коэффициентами. Соотношения во второй и третьей строчках задают значения компонент сеточной функции при $k = 0, 1$. Общее решение разностного уравнения определяется в соответствии с (8.32) корнями характеристического уравнения

$$\lambda^2 + 2Ah\lambda - 1 = 0,$$

откуда

$$\lambda_{1,2} = -Ah \pm \sqrt{1 + (Ah)^2}|_{Ah \ll 1} = -Ah \pm \left(1 + \frac{A^2h^2}{2}\right) + O(h^4).$$

Таким образом, общее решение

$$y_k = C_1\lambda_1^k + C_2\lambda_2^k,$$

где

$$\lambda_1 = 1 - Ah + \frac{A^2h^2}{2} + O(h^4), \quad \lambda_2 = -\left(1 + Ah + \frac{A^2h^2}{2} + O(h^4)\right).$$

Учитывая заданные при $k = 0, 1$ условия, получаем соотношения для C_1, C_2:

$$C_1 + C_2 = 1, \quad C_1\lambda_1 + C_2\lambda_2 = 1 - Ah,$$

из которых находим

$$C_1 = \frac{1 - Ah - \lambda_2}{\lambda_1 - \lambda_2} = \frac{1 + \dfrac{A^2h^2}{4} + O(h^4)}{1 + \dfrac{A^2h^2}{2} + O(h^4)} =$$

$$= \left(1 + \frac{A^2h^2}{4}\right)\left(1 - \frac{A^2h^2}{2}\right) + O(h^4) = 1 - \frac{A^2h^2}{4} + O(h^4),$$

$$C_2 = \frac{\dfrac{A^2h^2}{4} + O(h^4)}{1 + \dfrac{A^2h^2}{2} + O(h^4)} = \frac{A^2h^2}{4} + O(h^4).$$

Далее, подобно тому, как это было сделано в примере 1, преобразуем выражения λ_1^k и λ_2^k:

$$\lambda_1^k = \left[1 - Ah + \frac{A^2h^2}{2} + O(h^4)\right]^k =$$

$$= \exp\left\{\frac{x_k}{h}\ln\left[1 - Ah + \frac{A^2h^2}{2} + O(h^4)\right]\right\} =$$

$$= \exp\left\{\frac{x_k}{h}\left[\left(-Ah + \frac{A^2h^2}{2} + O(h^4)\right) - \frac{1}{2}\left(Ah - \frac{A^2h^2}{2} + \ldots\right)^2 - \right.\right.$$

$$\left.\left. - \frac{1}{3}\left(Ah - \frac{A^2h^2}{2} + \ldots\right)^3 + \ldots\right]\right\} =$$

$$= \exp\left\{\frac{x_k}{h}\left[-Ah + \frac{1}{2}A^3h^3 - \frac{1}{3}A^3h^3 + \ldots\right]\right\} =$$

$$= \exp\left\{\frac{x_k}{h}\left[-Ah + \frac{1}{6}A^3h^3 + O(h^4)\right]\right\} =$$

$$= e^{-Ax_k}\exp\left[\frac{1}{6}A^3x_kh^2 + O(h^3)\right] =$$

$$= e^{-Ax_k}\left[1 + \frac{1}{6}A^3x_kh^2 + O(h^3)\right].$$

Аналогично,

$$\lambda_2^k = (-1)^k e^{Ax_k}\left[1 - \frac{1}{6}A^3x_kh^2 + O(h^3)\right].$$

Итак, окончательно

$$y_k = \left(1 - \frac{A^2h^2}{4}\right)e^{-Ax_k}\left[1 + \frac{1}{6}A^3x_kh^2\right] +$$

$$+ (-1)^k\frac{A^2h^2}{4}e^{Ax_k} + O(h^3) =$$

$$= e^{-Ax_k} - e^{-Ax_k}\frac{A^2h^2}{4}\left(1 - \frac{2}{3}Ax_k\right) + (-1)^k e^{Ax_k}\frac{A^2h^2}{4} + O(h^3),$$

или $y_k = e^{-Ax_k} + \delta_k$, где

$$\delta_k = -e^{-Ax_k}\frac{A^2h^2}{4}\left[\left(1 - \frac{2}{3}Ax_k\right) - (-1)^k e^{2Ax_k}\right] + O(h^3). \qquad (8.37)$$

Главный член погрешности (8.37) является величиной второго порядка малости, что соответствует замечанию, сделанному в Лекции 7, о втором порядке точности этого метода.

Фактически, в рамках проведенного анализа, найдя выражение для главного члена погрешности, мы получаем довольно подробное представление о свойствах рассматриваемого метода. Например, из (8.37) следует, что ошибка носит колебательный характер из-за второго слагаемого в квадратных скобках. Более того, при $A \geqslant 1$ данный метод второго порядка точности может оказаться не только хуже метода Эйлера (первого порядка), но даже в принципе нереализуемым на ЭВМ.

Пусть, к примеру, $A = 20$. Относительная погрешность метода Эйлера при $x = 1$ согласно (8.36):

$$\delta y = \frac{|\delta_K|}{e^{-A}} \approx \frac{A^2 x_k h}{2} = 200h.$$

Таким образом, чтобы получить при $x = 1$ погрешность примерно 10% ($\delta y \approx 0.1$), надо вести вычисления с шагом $h = 0.0005$.

Для разобранного в примере 2 метода в выражении для погрешности (8.37) при $A = 20$ будет превалировать, очевидно, второе слагаемое в квадратных скобках, т. е.

$$\delta_k \approx (-1)^k \frac{A^2 h^2}{4} e^{Ax_k}.$$

Чтобы получить при $x = 1$ результат вычислений с той же точностью, надо выбрать шаг h так, чтобы

$$\frac{|\delta_k|}{e^{-A}} = \frac{A^2 h^2}{4} e^{2A} = 100 h^2 e^{40} \approx 0.1.$$

Этому требованию удовлетворяет $h \approx 6.5 \cdot 10^{-11}$. Чтобы это значение могло быть представлено в ЭВМ, необходимо проводить вычисления с двойной точностью. Допустим, что расчет по схеме Эйлера в указанных условиях требует времени в одну секунду. Тогда при счете по методу примера 2 с найденным шагом требуется пройти примерно в 10^7 раз больше расчетных точек, соответственно машинного времени потребовалось бы приблизительно 10^7с $\sim$ 1 год.

Замечание. Системы линейных разностных уравнений с постоянными коэффициентами возникают, в частности, в результате записи тех или иных методов решения задач для обыкновенных линейных дифференциальных уравнений с постоянными коэффициентами. Важно, что свойства решений разностных уравнений, т. е. свойства соответствующих методов, выявляемые в рамках анализа, продемонстрированного в данном разделе, проявляются и на других, более сложных (например, нелинейных) задачах. ▲

Пример неустойчивого метода (разностной схемы). В основной части лекции было отмечено, что проблема устойчивости численных методов решения задач для обыкновенных дифференциальных уравнений не является актуальной. В самом деле, например, в [2] доказано, что двухточечные методы (все методы Рунге–Кутты) обеспечивают сходимость (т. е. устойчивы) при условиях, гарантирующих существование и единственность решения исходной задачи. С многоточечными методами, однако, могут возникать осложнения.

Рассмотрим следующий метод решения задачи (8.33):

$$\begin{cases} \dfrac{-y_{k+1} + 4y_k - 3y_{k-1}}{2h} + Ay_{k-1} = 0, \quad k = 1, 2, \ldots, K-1, \\ y_0 = 1, \\ y_1 = 1 - Ah. \end{cases} \tag{8.38}$$

Верхние соотношения (8.38) аппроксимируют уравнение (8.33) во внутренних узлах интегрирования. Это становится очевидным, если представить разностное отношение в виде

$$\frac{-y_{k+1} + 4y_k - 3y_{k-1}}{2h} = 2\frac{y_k - y_{k-1}}{h} - \frac{y_{k+1} - y_{k-1}}{2h}.$$

Ясно, что после подстановки сюда значений точного решения $y(x_k)$ при $h \to 0$ получится нужная производная dy/dx. (Аккуратное же исследование приводит к выводу, что схема (8.38) аппроксимирует задачу (8.33) со вторым порядком точности.)

Заметим, что разностные уравнения (8.38) суть однородные уравнения с постоянными коэффициентами, и, следовательно (см. первый раздел дополнений), мы можем извлечь дополнительную информацию о разностном решении. Характеристическое уравнение для (8.38) имеет вид

$$\lambda^2 - 4\lambda + (3 - 2Ah) = 0.$$

Корни уравнения $\lambda_{1,2} = 2 \mp \sqrt{1 + 2Ah}$. В предположении, что $Ah \ll 1$, $\lambda_{1,2} \approx 2 \mp (1 + Ah) + O(h^2)$, т. е. $\lambda_1 \approx (1 - Ah) + O(h^2)$, $\lambda_2 \approx (3 + Ah) + O(h^2)$. Общее решение $y_k = C_1\lambda_1^k + C_2\lambda_2^k$. Используя дополнительные условия при $k = 0, 1$, находим C_1, C_2:

$$C_1 = 1 + O(h^2), \quad C_2 = O(h^2).$$

Таким образом, численное решение по методу (8.38) может быть представлено в аналитической форме:

$$y_k = [1 + O(h^2)](1 - Ah)^{x_k/h} + O(h^2)3^{x_k/h}\left(1 + \frac{Ah}{3}\right)^{x_k/h} =$$
$$= e^{-Ax_k}[1 + O(h^2)] + O(h^2)3^{x_k/h}[e^{Ax_k/3} + O(h)]. \tag{8.39}$$

Очевидно, что при $h \to 0$ сходимость к решению задачи (8.33) отсутствует, так как $|O(h^2)|3^{x_k/h} \underset{h\to 0}{\longrightarrow} \infty$. Мы здесь имеем дело с неустойчивой схемой: условие устойчивости (8.24) не выполнено.

Замечание. Можно было не искать решение в виде (8.39), а, заметив, что одно из частных решений $|\lambda_2^k| \underset{h\to 0}{\longrightarrow} \infty$, констатировать неустойчивость алгоритма (8.38), так как любое конкретное решение представляет собой линейную комбинацию этих частных решений. Кстати, один из распространенных способов анализа устойчивости алгоритмов состоит в том, что, выписав разностные уравнения применительно к «тестовой» задаче (для линейного дифференциального уравнения с постоянными коэффициентами), находят частные решения λ^k и смотрят, когда (при каких условиях) $|\lambda^k| \leqslant \text{const}$ при $|h^k| \to 0$. ▲

Дополнительную информацию по рассмотренным здесь вопросам можно найти в [1, с. 483–524], [2, с. 379–480], [4, с. 15–70, 71–170], [7, с. 197–220], [9, с. 261–279], [11, с. 23–60, 139–175], [12, с. 25–47], [32].

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Найти погрешность аппроксимации уравнений (8.11).

2. Описать алгоритм пристрелки для решения задач:

а) $u' = f(x, u, v)$, $v' = g(x, u, v)$ $(a < x < b)$, $u(a) = \alpha$, $v(a) + v(b) = \beta$ (α, β — заданные значения);

б) $u' = f(x, u, v, w)$, $v' = g(x, u, v, w)$, $w' = h(x, u, v, w)$ $(a < x < b)$, $u(a) = u_a$, $v(b) = v_b$, $w(b) = w_b$ (u_a, v_b, w_b — заданные значения).

3. Показать, что при $\lambda_1 \neq \lambda_2$ решения λ_1^k, λ_2^k однородного уравнения (8.28) линейно независимы.

ЛЕКЦИЯ 9

РАЗНОСТНЫЕ СХЕМЫ ДЛЯ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

Модельные уравнения (переноса, теплопроводности, Пуассона). Эволюционные задачи, типичные формулировки задач для уравнений переноса и теплопроводности. Аппроксимация. Примеры разностных схем для модельных задач. Явные и неявные схемы. Интегро-интерполяционный метод построения разностных схем, аппроксимирующих законы сохранения.

Мы приступаем к обсуждению методов численного решения задач для уравнений (и систем уравнений) в частных производных. Источником таких задач большей частью является математическое моделирование физических процессов. Поэтому раздел математики, связанный с изучением свойств возможных решений уравнений в частных производных называется *математической физикой* (сами же уравнения, о которых идет речь, зачастую называют уравнениями математической физики).

Что касается методов численного решения задач для уравнений математической физики, то существует множество монографий, учебников, в том или ином объеме освещающих эти вопросы. В рамках вводного курса наша цель — получить первоначальные представления о простейших способах конструирования численных алгоритмов (разностных схем) для решения задач математической физики, о подходах, применяемых при анализе свойств полученных разностных схем.

Модельные уравнения переноса, волновое, теплопроводности и Пуассона. Мы ограничимся здесь рассмотрением численных методов решения задач для простых (модельных) уравнений:

$$\frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} = f(t,x) \qquad \text{— уравнение переноса,}$$

$$\frac{\partial U}{\partial t} = \mu\frac{\partial^2 U}{\partial x^2} \quad (\mu > 0) \qquad \text{— уравнение теплопроводности (или диффузии),}$$

$$\frac{\partial^2 U}{\partial t^2} + a^2\frac{\partial^2 U}{\partial x^2} = f(t,x) \qquad \text{— волновое уравнение (уравнение малых колебаний),}$$

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f(x,y) \qquad \text{— уравнение Пуассона,}$$

Первые три уравнения описывают эволюцию состояния $U(t,x)$ во времени (t) и пространстве (x). Их (эти уравнения) называют *эволю-*

ционными или *нестационарными*, в отличие от четвертого, которое описывает установившееся (*стационарное*) состояние $U(x, y)$ в пространстве (x, y). Модельные уравнения привлекательны в методическом плане по той причине, что при относительной своей простоте они несут в себе существенные черты (особенности, свойства) сложных уравнений, описывающих многие реальные физические процессы.

Отметим, что при наличии двух независимых переменных для каждого из перечисленных уравнений возможно большее (сравнительно с обыкновенными дифференциальными уравнениями) многообразие различных формулировок задач. Вместе с тем заметно расширяется и арсенал возможных методов численного решения этих задач.

Итак, начнем учиться ориентироваться в множестве численных алгоритмов для задач математической физики.

Рассмотрим общую формулировку исходной задачи: найти функцию U, удовлетворяющую во внутренних точках ω — области изменения независимых переменных (рис. 9.1) — некоторому уравнению

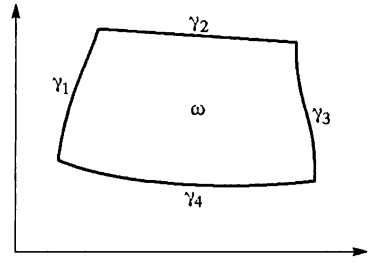


Рис. 9.1.

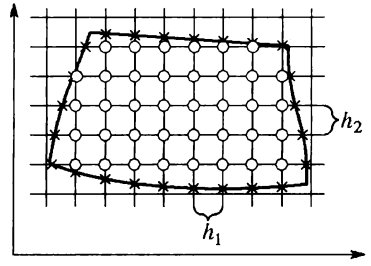


Рис. 9.2.

(системе) в частных производных, а на участках границы Γ — γ_1, $\gamma_2, \ldots$ — необходимым условиям, обеспечивающим корректность задачи (решение существует, единственно и устойчиво по отношению к малым возмущениям входных данных).

Эту задачу будем обозначать впредь записью

$$LU = F, \tag{9.1}$$

где L — оператор, определяющий вид дифференциальных уравнений в ω, дополнительных соотношений на Γ; F — соответствующие правые части (входные данные).

Применяя разностный метод решения задачи (9.1), в области расчета вводим сетку узлов (сеточную область) $\omega^{(h)} \bigcup \Gamma^{(h)}$. В простейшем случае — это точки пересечения линий сетки, параллельных осям координат (рис. 9.2), удаленных друг от друга на расстояния

$(h_1, h_2, \ldots)$, называемые шагами сетки по соответствующим направлениям и представляющие собой малые параметры. При этом узлы, лежащие внутри ω (кружочки на рис. 9.2), образуют $\omega^{(h)}$ — совокупность внутренних точек сеточной области. Точки пересечения линий сетки с границей Γ образуют $\Gamma^{(h)}$ — совокупность граничных узлов сеточной области (крестики на рис. 9.2).

Замечание 1. Это не единственный способ определения совокупности внутренних и граничных узлов сетки. В Лекции 12 будет приведен пример другого определения сеточной области. ▲

Замечание 2. В общем случае сетка может быть неравномерной ($h_1, h_2, \ldots \neq \text{const}$), криволинейной (рис. 9.3). В последнем случае шаги сетки (сеточные параметры) — это расстояния между соседними узлами, лежащими на одной линии. ▲

При стремлении шагов сетки к нулю сетка *сгущается*, узлы $\omega^{(h)}$, $\Gamma^{(h)}$ равномерно покрывают расчетную область ω и границу Γ.

Приближенное решение задачи (9.1) будем вычислять в узлах сетки. Искомые значения вкупе образуют сеточную функцию $U^{(h)}$. Заменяя дифференциальные уравнения и дополнительные соотношения задачи (9.1) некоторыми соотношениями для искомых компонент сеточной функции $U^{(h)}$ (как это делается, будет показано ниже на конкретных примерах), приходим к задаче (разностной схеме) для $U^{(h)}$:

$$L_h U^{(h)} = F^{(h)}, \qquad (9.1')$$

которая, собственно, и решается численно.

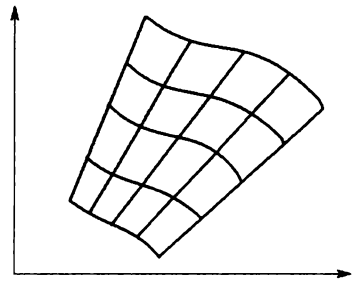


Рис. 9.3.

Под ошибкой найденного решения естественно понимать

$$\delta^{(h)} = U^{(h)} - [U]^{(h)},$$

где $[U]^{(h)}$ — сеточная функция, представляющая собой совокупность значений решения исходной задачи (9.1) в узлах сетки (проекция решения (9.1) на сеточную область).

Величина ошибки

$$\|\delta^{(h)}\| = \|U^{(h)} - [U]^{(h)}\|. \qquad (9.2)$$

Использование абстрактных формулировок (9.1), (9.1′) для исходной и разностной задач позволяет выявить общие, не зависящие от конкретной задачи требования к разностной схеме (9.1′), выполнение

которых гарантирует малость ошибки разностного решения. Они содержатся в теореме, доказанной в конце предыдущей лекции: если разностная схема (9.1′) аппроксимирует исходную задачу (9.1), т. е.

$$\|\psi^{(h)}\| = \|L_h[U]^{(h)} - F^{(h)}\| \underset{(h)\to 0}{\longrightarrow} 0,$$

и решение задачи (9.1′) устойчиво, то имеет место сходимость разностного решения $U^{(h)}$ к решению исходной задачи, т. е.

$$\|\delta^{(h)}\| \underset{(h)\to 0}{\longrightarrow} 0.$$

Здесь обозначение (h) имеет смысл совокупности сеточных параметров (шагов сетки), соответственно $(h) \to 0$ означает, что все шаги сетки стремятся к нулю.

Перейдем к конкретным примерам. Пока будем рассматривать различные схемы для решения эволюционных задач. При этом в данной лекции внимание будет акцентироваться на способах построения схем, аппроксимирующих исходную задачу.

Задача Коши для уравнения переноса.

$$\begin{cases} U_t + aU_x = f(x,t), & t > 0,\ -\infty < x < \infty \\ U(0,x) = \varphi(x). \end{cases} \tag{9.3}$$

Замечание. Сопоставляя эту конкретную формулировку задачи с записью абстрактной задачи (9.1), можем отметить, что здесь

$$LU = \begin{cases} \dfrac{\partial U}{\partial t} + a\dfrac{\partial U}{\partial x}, & t > 0, \quad -\infty < x < \infty \\ U(0,x), & t = 0; \quad -\infty < x < \infty \end{cases}$$

$$F = \begin{cases} f(t,x), & t > 0, \\ \varphi(x), & t = 0. \end{cases} \blacktriangle$$

На полуплоскости $t > 0$ вводим прямоугольную равномерную сетку узлов с координатами $\{t_n = n\tau,\ n = 0, 1, \ldots;\ x_k = kh,\ k = 0, \pm 1, \pm 2, \ldots\}$ (рис. 9.4). Обозначим через U_k^n компоненту сеточной функции в (n,k)-м узле, являющимся точкой пересечения n-го слоя по t с k-м слоем по x. Тогда

$$U^{(h)} = \{U_k^n,\ n = 0, 1, \ldots;\ k = 0, \pm 1, \pm 2, \ldots\}$$

Построим сеточную задачу. Для этого в каждом внутреннем узле ($n > 0$), привлекая компоненты сеточной функции в этом узле и ближайших, заменим производные в дифференциальном уравнении по простейшим формулам численного дифференцирования:

$$\frac{U_k^n - U_k^{n-1}}{\tau} + a\frac{U_{k+1}^{n-1} - U_k^{n-1}}{h} = f_k^{n-1}, \quad \begin{matrix} n = 1, 2, \dots \\ k = 0, \pm 1, \pm 2, \dots . \end{matrix} \tag{9.3'}$$

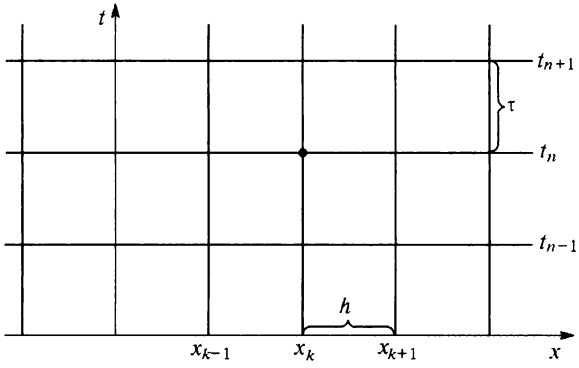


Рис. 9.4.

Добавляя соотношения $U_k^0 = \varphi_k$, $k = 0, \pm 1, \pm 2, \dots$, получаем систему соотношений (разностную схему) для неизвестных компонент сеточной функции.

Замечание 1. Сопоставляя (9.3') с (9.1'), отмечаем, что

$$L_h U^{(h)} = \begin{cases} \dfrac{U_k^n - U_k^{n-1}}{\tau} + a\dfrac{U_{k+1}^{n-1} - U_k^{n-1}}{h}, & n > 0;\ k = 0, \pm 1, \pm 2, \dots \\ U_k^0, & n = 0;\ k = 0, \pm 1, \pm 2, \dots \end{cases}$$

$$F^{(h)} = \begin{cases} f_k^{n-1}, & n > 0,\ k = 0, \pm 1, \pm 2, \dots \\ \varphi_k, & n = 0,\ k = 0, \pm 1, \pm 2, \dots \end{cases}$$

Записав верхнее соотношение в виде, разрешенном относительно U_k^n,

$$U_k^n = \left(1 + \frac{a\tau}{h}\right)U_k^{n-1} + \frac{a\tau}{h}U_{k+1}^{n-1},$$

получим формулу, последовательно определяющую решение во всех узлах по k при $n = 1, 2, \dots$ ▲

З а м е ч а н и е 2. Разумеется, эта схема имеет чисто методическое значение, поскольку немыслимо проведение вычислений в бесконечном количестве точек каждого слоя по n. ▲

Если ограничены вторые производные решения задачи (9.3), то схема (9.3′) аппроксимирует задачу (9.3). В самом деле, ошибка аппроксимации в $(n-1, k)$-м узле (под $U(t, x)$ понимается значение точного решения задачи (9.3) в точке (t, x)):

$$\psi_k^{n-1} = \frac{U(t_n, x_k) - U(t_{n-1}, x_k)}{\tau} + a\frac{U(t_{n-1}, x_{k+1}) - U(t_{n-1}, x_k)}{h} - f(t_{n-1}, x_k) =$$

(после замены входящих сюда значений $U(t, x)$ по формулам Тейлора относительно узла (t_{n-1}, x_k))

$$= \left(\frac{\tau}{2}\widetilde{U}''_{t^2} + \frac{ah}{2}\widetilde{U}''_{x^2}\right)\Big|_k^{n-1}, \quad n > 1.$$

При $n = 0$

$$\psi_k^0 = U(0, x_k) - \varphi(x_k) \equiv 0.$$

Следовательно,

$$\|\psi^{(h)}\| = \max_{n,k} |\psi_k^n| \leqslant \frac{1}{2}(M_{2t}\tau + M_{2x}h) \underset{\tau, h \to 0}{\longrightarrow} 0, \tag{9.4}$$

где $M_{2t} = \max\limits_{t,x} |U''_{t^2}|$, $M_{2x} = \max\limits_{t,x} |U''_{x^2}|$.

Когда для погрешности аппроксимации имеет место оценка типа (9.4), говорят, что соответствующая схема является схемой *первого порядка аппроксимации по t и по x.*

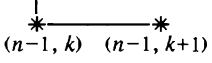


Рис. 9.5.

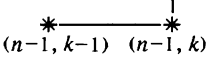


Рис. 9.6.

Конфигурация узлов, значения сеточной функции в которых определяют вид разностных уравнений, записываемых для внутренних точек сетки, называется *шаблоном разностной схемы.* Для схемы (9.3′) шаблон имеет вид, показанный на рис. 9.5. Ссылка на шаблон удобна тем, что он, как правило (не всегда однозначно), определяет вид разностных уравнений схемы, о которой идет речь. Например,

для той же задачи схема с шаблоном, показанным на рис. 9.6, записывается следующим образом:

$$\begin{cases} \dfrac{U_k^n - U_k^{n-1}}{\tau} + a\dfrac{U_k^{n-1} - U_{k-1}^{n-1}}{h} = f_k^{n-1}, & n = 1, 2, \dots \\ U_k^0 = \varphi_k, & k = 0, \pm 1, \pm 2, \dots \end{cases} \qquad (9.3'')$$

Перейдем теперь к рассмотрению более реальной (с точки зрения возможности проведения расчетов) задачи.

Краевая задача для уравнения переноса. Пусть надо найти решение уравнения переноса $U_t + aU_x = f(t, x)$ (здесь и далее считается $a = \text{const} > 0$) в прямоугольнике $0 \leqslant t \leqslant T$, $0 \leqslant x \leqslant 1$.

Обсудим предварительно вопросы, связанные с корректной формулировкой исходной задачи.

Рассмотрим уравнение $dx/dt = a$. Решением его является семейство линий (в данном случае прямых) $x - at = C$, где — произвольная константа. Вдоль каждой линии этого семейства исходное уравнение переноса может быть записано в виде обыкновенного дифференциального уравнения:

$$\left.\frac{dU}{dt}\right|_C = f(t, x)\big|_C, \qquad (9.5)$$

где $\left.\dfrac{dU}{dt}\right|_C = \dfrac{\partial U}{\partial t} + \dfrac{\partial U}{\partial x}\dfrac{dx}{dt} = U_t + aU_x$ — так называемая производная по направлению, определяемому уравнением $dx/dt = a$.

Замечание. Линии (в пространстве двух независимых переменных), вдоль которых уравнение в частных производных переходит в обыкновенное, называются *характеристиками* уравнения. Наличие действительных характеристик является признаком гиперболичности данного уравнения. ▲

Таким образом, чтобы однозначно определить решение уравнения переноса в какой-либо внутренней точке A (рис. 9.7), можно, проведя через эту точку характеристику, решать далее задачу Коши для уравнения (9.5) по t в сторону возрастания, либо в обратном направлении. В первом случае «начальное» условие должно быть задано в точке O, во втором — в точке O_1. Ассоциируя t со временем (т.е. принимая, что t может лишь возрастать), а x — с пространственным направлением, приходим к выводу, что для однозначного определения решения во всех внутренних точках (A, B), *необходимо задать значения U на отрезке* $[0, 1]$ *оси x (начальные данные) и на отрезке* $[0, T]$ *оси t (краевые условия)*. Перенос начального возмущения $(\varphi(x), \psi(t))$ осуществляется слева направо (по x).

Таким образом, математическая формулировка задачи для этого случая имеет вид:

$$\begin{cases} U_t + aU_x = f(t,x) & (t > 0, 0 < x \leqslant 1), \\ U(0,x) = \varphi(x) & (0 \leqslant x \leqslant 1), \\ U(t,0) = \psi(t) & (0 < t \leqslant T). \end{cases} \tag{9.6}$$

Замечание. Очевидно, при $a < 0$ надо было бы задавать дополнительное условие не на левой границе, а на правой. Перенос возмущения справа налево (по x). ▲

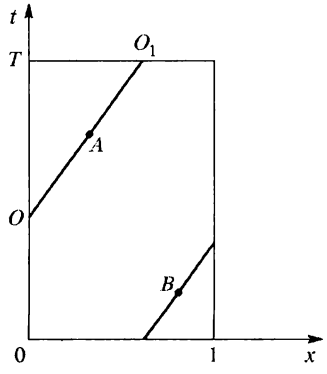


Рис. 9.7.

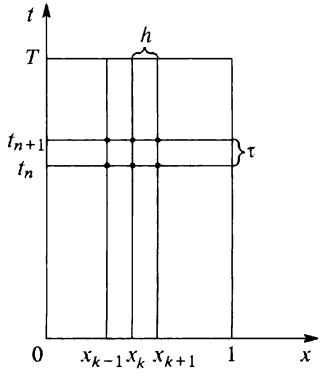


Рис. 9.8.

Обратимся теперь к проблемам численного решения задачи (9.6). Введем в расчетной области множество узлов сетки (рис. 9.8):

$$\omega^{(h)} = \{(t_n, x_k),\ t_n = n\tau,\ n = 0, 1, \ldots, N = \frac{T}{\tau},$$
$$x_k = kh,\ k = 0, 1, \ldots, K = \frac{1}{h}\}.$$

Здесь τ — шаг сетки по времени (t); h — шаг по пространству (x); N — номер последнего слоя по t, для которого надо вычислить решение; K — номер слоя по x, совпадающего с правой границей расчетной области.

Искомыми величинами будем считать компоненты сеточной функции

$$U^{(h)} = \{U_k^n;\ n = 0, 1, \ldots, N; k = 0, 1, \ldots, K\}.$$

(Компонента U_k^n относится к (n, k)-му узлу сетки, лежащему, напомним, на пересечении n-го слоя по t с k-м слоем по x.)

Пример 1. Рассмотрим схему с шаблоном, показанным на рис. 9.9.

Разностные уравнения для внутренних точек сетки строятся так же, как в схеме (9.3″) для задачи Коши. В итоге получим

$$\begin{aligned}
&\frac{U_k^n - U_k^{n-1}}{\tau} + a\frac{U_k^{n-1} - U_{k-1}^{n-1}}{h} = f_k^{n-1}; \quad n = 1, 2, \ldots, N; \quad k = 1, \ldots, K,\\
&U_k^0 = \varphi(x_k), \quad k = 0, 1, \ldots, K,\\
&U_0^n = \psi(t_n), \quad n = 1, 2, \ldots, N
\end{aligned} \tag{9.6′}$$

(предполагается, что $\varphi(0) = \psi(0)$).

Очевидно, это, как и (9.3′), схема первого порядка аппроксимации, как по t, так и по x. Главные члены погрешности аппроксимации в узле $(n-1, k)$: $\psi_k^{n-1} = \left[\frac{\tau}{2}U_{tt} - \frac{ah}{2}U_{xx}\right]_k^{n-1} + O(\tau^2, h^2)$ — во внутренних точках, $\psi_k^0 = \psi_0^n = 0$ — в граничных узлах.

Остановимся на *последовательности вычислений*. Из соотношений второй строки (9.6′) находятся значения сеточной функции во всех узлах начального слоя. Затем вычисление величин на каждом последующем слое по t осуществляется стандартным образом:

— с помощью соотношения из третьей строчки (9.6′) (при нужном n) находится значение в узле, принадлежащем левой границе расчетной области;

— разностные уравнения (9.6′), будучи переписаны в виде

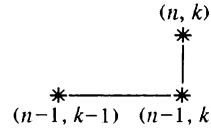


Рис. 9.9.

$$U_k^n = U_k^{n-1} - \frac{a\tau}{h}(U_k^{n-1} - U_{k-1}^{n-1}) + \tau f_k^{n-1},$$

дают расчетную формулу для вычисления решения в остальных узлах расчетного слоя (временного слоя с номером n).

Замечание. Процесс вычисления решения эволюционных задач «расслаивается»: счет ведется посредством перехода от одного слоя по t к другому. Это обстоятельство естественным образом связано со спецификой эволюционных задач: их решение в каждый момент времени однозначно определяется состоянием в предшествующий момент. ▲

В свете этого замечания в разностных уравнениях величины, относящиеся к верхнему слою, при планировании вычислений трактуются

как неизвестные, а величины с предшествующего слоя считаются известными.

Разностная схема называется *явной*, если в каждое уравнение входит не более одной неизвестной величины (в указанном смысле). В противном случае схема называется *неявной*.

Очевидно, схема (9.6′) — явная.

Пример 2. Рассмотрим схему с шаблоном, показанным на рис. 9.10:

$$\frac{U_k^n - U_k^{n-1}}{\tau} + a\frac{U_{k+1}^{n-1} - U_k^{n-1}}{h} = f_k^{n-1}, \quad n = 1, 2, \ldots, N; \quad k = 0, 1, \ldots, K-1; \qquad (9.6'')$$

$$U_k^0 = \varphi(x_k), \quad k = 0, 1, \ldots, K;$$
$$U_0^n = \psi(t_n), \quad n = 1, 2, \ldots, N.$$

Это снова явная схема первого порядка аппроксимации по t и по x. Главные члены погрешности аппроксимации в узле $(n-1, k)$:

$$\psi_k^{n-1} = \left[\frac{\tau}{2}U_{tt} + \frac{ah}{2}U_{xx}\right]_k^{n-1} + O(\tau^2, h^2), \ n > 0,$$
$$\psi_k^0 \equiv 0.$$

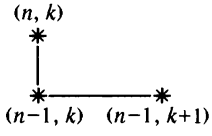


Рис. 9.10.

При планировании последовательности вычислений сталкиваемся с противоречием. Решение в точках левой границы определяется неоднозначно: во-первых, соотношениями из третьей строчки (9.6″), а, во-вторых, уравнениями первой строки при $k = 0$. В то же время в (9.6″) не хватает соотношений для определения величин в узлах правой границы (при $k = K$).

Таким образом, несмотря на наличие аппроксимации, эта схема непригодна для численного решения задачи (9.6). (Мы увидим в дальнейшем, что она не удовлетворяет условию устойчивости.)

Краевая задача для уравнения теплопроводности. Типичная формулировка задачи для уравнения теплопроводности на конечной по координате x области (для определенности будем считать $0 \leqslant x \leqslant 1$):

$$\begin{cases} U_t = \mu U_{xx} \ (\mu > 0) \ \text{ при } \ t > 0, \ 0 < x < 1, \\ U(0, x) = \varphi(x), \ \ U(t, 0) = \psi_1(t), \ \ U(t, 1) = \psi_2(t). \end{cases} \qquad (9.7)$$

Если, например, U — температура, то задача (9.6) описывает теплоперенос вдоль одномерного стержня единичной длины при заданном

начальном распределении температуры — $\varphi(x)$ и заданных температурных режимах на концах стержня: $\psi_1(t)$ и $\psi_2(t)$.

З а м е ч а н и е. Если U — концентрация, то задача (9.7) описывает диффузию вещества, поэтому уравнение $U_t = \mu U_{xx}$ иногда называют *уравнением диффузии.* ▲

Вообще, это простейшее модельное уравнение параболического типа содержит в себе характерные черты сложных параболических уравнений, описывающих различные диссипативные процессы (типа теплопроводности или диффузии) в плазме, в магнитной гидродинамике, в биологии и т. д.

Использование самых простых формул численного дифференцирования при конструировании возможных разностных схем здесь требует привлечения по крайней мере трех точек по x и двух по t. Соответственно, элементарные допустимые шаблоны показаны на рис. 9.11, 9.12.

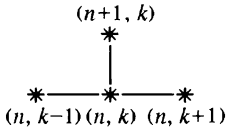


Рис. 9.11.

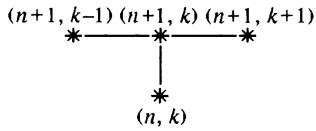


Рис. 9.12.

П р и м е р 1. Явная схема с шаблоном, изображенным на рис. 9.11. После введения сетки (так же, как в предыдущем разделе) разностная схема записывается в виде

$$\begin{aligned} &\frac{U_k^{n+1} - U_k^n}{\tau} = \mu \frac{U_{k+1}^n - 2U_k^n + U_{k-1}^n}{h^2}, \quad n = 0, 1, \ldots, N-1, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad k = 1, 2, \ldots, K-1, \\ &U_k^0 = \varphi(x_k), \quad k = 0, 1, \ldots, K, \\ &U_0^n = \psi_1(t_n), \quad U_K^n = \psi_2(t_n), \quad n = 1, 2, \ldots, N. \end{aligned} \tag{9.7′}$$

Это схема первого порядка аппроксимации по t и второго — по x. Главные члены погрешности аппроксимации во внутренних узлах (n, k):

$$\psi_k^n = \left[\frac{\tau}{2} U_t t - \mu \frac{h^2}{12} U_{x^4}^{(4)}\right]_k^n + O(\tau^2, h^4).$$

В граничных узлах и в точках начального слоя: $\psi_0^n = \psi_k^n = \psi_k^0 = 0$.

Последовательность вычислений вполне аналогична той, что была описана для схемы (9.6′), с той оговоркой, что расчетная формула для решения во внутренних точках текущего $(n+1)$-го слоя здесь имеет

вид

$$U_k^{n+1} = U_k^n + \frac{\mu\tau}{h^2}(U_{k+1}^n - 2U_k^n + U_{k-1}^n),$$

а значения в узлах правой границы вычисляются по заданной формуле — $\psi_2(t_n)$.

Пример 2. Неявная схема с шаблоном, показанным на рис. 9.12:

$$\frac{U_k^{n+1} - U_k^n}{\tau} = \mu \frac{U_{k+1}^{n+1} - 2U_k^{n+1} + U_{k-1}^{n+1}}{h^2}, \quad \begin{matrix} n = 0, 1, \ldots, N-1, \\ k = 1, 2, \ldots, K-1, \end{matrix}$$

$$U_k^0 = \varphi(x_k), \quad k = 0, 1, \ldots, K,$$

$$U_0^n = \psi_1(t_n), \quad U_K^n = \psi_2(t_n), \quad n = 1, 2, \ldots, N.$$

Так же, как и явная, это схема первого порядка по t и второго — по x:

$$\psi_k^{n+1} = -\left[\frac{\tau}{2} U_t t + \mu \frac{h^2}{12} U_{x^4}^{(4)}\right]_k^{n+1} + O(\tau^2, h^4)$$

во внутренних узлах сетки (в граничных узлах ошибка аппроксимации равна нулю).

Вычисление решения в узлах $(n+1)$-го слоя при известных данных на n-м слое, как нетрудно видеть, сводится к решению системы линейных уравнений с трехдиагональной матрицей и реализуется по формулам прогонки (см. Лекцию 2).

ДОПОЛНЕНИЕ К ЛЕКЦИИ 9

О консервативности и других свойствах разностных уравнений. Для решения конкретной задачи естественно использовать разностную схему, которая (как это следует из основной теоремы о сходимости) аппроксимирует исходную задачу и является корректной, т. е. приводит к единственному решению, устойчивому по отношению к малым возмущениям входных данных. Мы видели, что для одной и той же задачи часто можно построить множество аппроксимирующих разностных схем (в том числе и устойчивых, как будет показано в Лекции 10). Таким образом, возникает необходимость выбора конкретной схемы расчета из множества возможных. В качестве критериев отбора довольно естественно привлекать такие, как более высокая точность, простота организации вычислений, экономичность вычислений. Однако это отнюдь не исчерпывающий набор требований, предъявляемых к разностным схемам.

При решении задач физического содержания зачастую априори известны некоторые (важные!) свойства искомых функций. Элементарный пример: одной из искомых характеристик в задачах о процессах в сплошной среде (газе, плазме) является плотность. Очевидно, плотность не может быть отрицательной. При конструировании разностной схемы можно исходить из дополнительного (наряду с аппроксимацией и устойчивостью) требования: расчетные значения плотности в узлах сетки должны принимать неотрицательные значения. (На этом пути приходят к так называемым *положительным разностным схемам*.)

Другой пример. Пусть известно, что некоторая физическая характеристика (та же плотность или температура, к примеру) распределена по пространственной координате монотонным образом. Можно строить схемы, которые гарантируют монотонное распределение соответствующей расчетной величины (так называемые *монотонные схемы*).

Особенно важную роль играют *консервативные разностные схемы*. Дело в том, что при решении физических задач дифференциальные уравнения, расчет которых планируется, как правило, выражают собой математическую запись тех или иных законов сохранения. С точки зрения математики дифференциальные уравнения являются следствием некоторых интегральных соотношений и выражают закон сохранения локально (в данной точке пространства, в данный момент времени). Упомянутые интегральные соотношения являются общей формой записи законов сохранения (в конечной области пространства, на конечном интервале по времени). Было бы весьма желательным потребовать от разностной схемы, аппроксимирующей дифференциальные уравнения (т. е. локальные законы сохранения), чтобы следствием разностных уравнений была аппроксимация интегральных соотношений, выражающих, как сказано, общую форму соответствующих законов сохранения. Схемы, удовлетворяющие этому требованию, называются *консервативными*. Важность консервативности разностных схем обусловлена, во-первых, общим (универсальным) характером законов сохранения (тем, что на какой-то из них или на ряд из них мы опираемся, решая практически любую задачу физического содержания). Во-вторых, тем, что, как показывает и практика расчетов и соответствующие исследования, во многих случаях только консервативные схемы позволяют получить достоверные результаты.

Учет дополнительных требований к искомому разностному решению зачастую достигается использованием более «изощренных» способов конструирования разностных схем, нежели те, которые мы использовали до сих пор (замена производных в дифференциальных

уравнениях разностными отношениями по тому или иному шаблону). Остановимся коротко на одном способе, который приводит к консервативным разностным схемам.

Интегро-интерполяционный метод построения разностных схем. Рассмотрим задачу о распространении тепла вдоль тонкого (одномерного) стержня. Соответствующий закон сохранения в данном случае представляет собой закон сохранения энергии — приращение энергии за время Δt на отрезке Δx равно разности потоков тепла на границах этого отрезка (мы полагаем ниже, что теплоемкость равна единице и энергия однозначно определяется температурой):

$$\int_0^{x+\Delta x} [U(t+\Delta t, x) - U(t,x)]dx = -\int_t^{t+\Delta t} [W(t, x+\Delta x) - W(t,x)]dt \tag{9.8}$$

Здесь $U(t,x)$ — распределение температуры в момент времени t, Δx — рассматриваемый отрезок стержня, Δt — рассматриваемый временной интервал; $W(t,x)$ — поток тепла (в единицу времени) в сечении x. (Очевидно, $\int_0^{x+\Delta x} U(t,x)\,dx$ — энергия стержня длины Δx в момент времени t; $\int_0^{t+\Delta t} W(t,x)\,dt$ — поток тепла через левую границу отрезка стержня Δx за время Δt.)

Рассматривая закон сохранения (9.8) для бесконечно малых Δx и Δt, приходим к дифференциальной (локальной) его форме:

$$\frac{\partial U}{\partial t} = -\frac{\partial W}{\partial x}. \tag{9.9}$$

Далее, поток тепла между соседними бесконечно малыми частицами стержня в единицу времени согласно закону Фурье пропорционален градиенту температуры:

$$W = -\varkappa(t,x,U)\frac{\partial U}{\partial x}, \tag{9.10}$$

$\varkappa$ — коэффициент теплопроводности (в общем случае функция независимых переменных и температуры).

Таким образом, приходим к следующей математической формулировке рассматриваемой задачи: найти решение уравнения

$$\frac{\partial U}{\partial t} = \frac{\partial}{\partial x}\left[\varkappa(t,x,U)\frac{\partial U}{\partial x}\right], \quad t>0,\ 0<x<L \tag{9.11}$$

(L — длина стержня), при заданном начальном распределении температуры $U(0,x) = \varphi(x)$, и заданном граничном режиме (при $x = 0, L$ заданы либо температура, либо тепловой поток).

Используя прежний подход к построению разностной схемы, мы бы переписали дифференциальное уравнение (9.11) в виде

$$\frac{\partial U}{\partial t} = \varkappa \frac{\partial^2 U}{\partial x^2} + \frac{\partial U}{\partial x}\left[\frac{\partial \varkappa}{\partial x} + \frac{\partial \varkappa}{\partial U}\frac{\partial U}{\partial x}\right], \tag{9.12}$$

и приближали бы производные $\dfrac{\partial U}{\partial t}$, $\dfrac{\partial U}{\partial x}$, $\dfrac{\partial^2 U}{\partial x^2}$ подходящими разностными отношениями.

Теперь поступим иначе. Введем в рассмотрение наряду с сеткой, к узлам которой отнесены искомые значения U_k^n (кружочки на рис. 9.13) вспомогательную сетку (крестики на рис. 9.13), к узлам которой отнесем значения $W_{k\pm1/2}^n$ потока тепла при $t = t_n$ между элементом стержня длиной h (выделенным жирной линией на рис. 9.13) и соседними (справа и слева) такими же элементами.

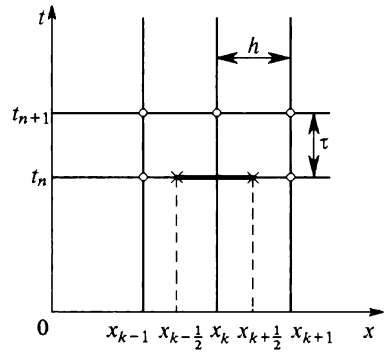


Рис. 9.13.

Далее, пользуясь введенными обозначениями, аппроксимируем не дифференциальное уравнение (9.12), а непосредственно закон сохранения (9.8) для обозначенного локального элемента в пределах временного интервала τ, заменяя интегралы в (9.8) по простейшим квадратурным формулам:

$$(U_k^{n+1} - U_k^n)h + (W_{k+1/2}^n - W_{k-1/2}^n)\tau = 0. \tag{9.13}$$

Разделив (9.5) на $h\tau$, получаем разностное уравнение, аппроксимирующее (9.9). Определив (с учетом закона Фурье (9.10)) локальный поток тепла:

$$W_{k+1/2}^n = -\varkappa_{k+1/2}^n \frac{U_{k+1}^n - U_k^n}{h}, \tag{9.14}$$

приходим к системе разностных уравнений, которые, исключая с помощью (9.14) $W_{k\pm1/2}^n$, можно записать в виде

$$\frac{U_k^{n+1} - U_k^n}{\tau} = \frac{1}{h}\left[\varkappa_{k+1/2}^n \frac{U_{k+1}^n - U_k^n}{h} - \varkappa_{k-1/2}^n \frac{U_k^n - U_{k-1}^n}{h}\right] \tag{9.15}$$

и которые во внутренних узлах сетки аппроксимируют дифференциальные уравнения (9.11) с первым порядком по τ и вторым — по h, в

чем можно убедиться, используя стандартную технику исследования на аппроксимацию.

Замечание 1. В (9.15) надо определить способ вычисления $\varkappa^n_{k+1/2}$, если коэффициент теплопроводности зависит от U. Это можно сделать, не нарушая второго порядка аппроксимации по x, например, одним из следующих способов:

а) $\varkappa^n_{k+1/2} = \frac{1}{2}[\varkappa_{k+1} + \varkappa_k] = \frac{1}{2}[\varkappa(t_n, x_k, U^n_k) + \varkappa(t_n, x_{k+1}, U^n_{k+1})]$,

б) $\varkappa^n_{k+1/2} = \varkappa(t_n, x_{k+1/2}, U^n_{k+1/2}) = \varkappa\left(t_n, x_k + \frac{h}{2}, \frac{U^n_k + U^{(k)}_{k+1}{}^n}{2}\right)$. ▲

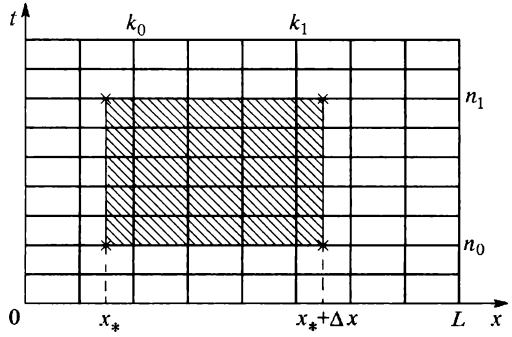


Рис. 9.14.

Замечание 2. При получении (9.13) из (9.8) мы использовали формулу прямоугольников с центральной точкой для вычисления энергии рассматриваемого интервала в момент времени t_n и t_{n+1} и формулу прямоугольников с «высотой», относящейся к нижней границе интервала по времени, для вычисления потока тепла через границы отрезка за время τ. Если бы для последнего использовать формулу трапеций (например, на правой границе: $(\tau/2)(W^{n+1}_{k+1/2} + W^n_{k+1/2})$), то мы пришли бы к схеме второго порядка точности как по h, так и по τ, но это была бы неявная схема и, в общем случае, нелинейная. ▲

Замечание 3. При $\varkappa = \text{const}$ на этом пути получаются разностные схемы, с которыми мы уже сталкивались (конструируя их обычным образом). В частности, соотношения (9.15) переходят в явные четырехточечные разностные уравнения

$$\frac{U^{n+1}_k - U^n_k}{\tau} = \varkappa \frac{U_{k+1} - 2U^n_k + U^n_{k-1}}{h^2}.$$ ▲

Главное, чего мы достигли на этом пути, — мы построили разностные уравнения, удовлетворяющие условию консервативности. В самом деле, просуммировав уравнения (9.13) по k (от k_0 до k_1) и по n (от n_0 до $n_1 - 1$), приходим к соотношению

$$\sum_{k=k_0}^{k_1} (U_k^{n_1} - U_k^{n_0})h + \sum_{n=n_0}^{n_1-1} (W_{k_1+1/2}^n - W_{k_0-1/2}^n)\tau = 0, \qquad (9.16)$$

которое, очевидно, аппроксимирует интегральный закон сохранения (9.8) по конечной области $x \in [x_*, x_* + \Delta x]$ $(x_* = x_{k_0} - \frac{h}{2})$, $x_* + \Delta x = x_{k_1} + \frac{h}{2})$, $t \in [t_*, t_* + \Delta t)$ $(t_* = t_{n_0}$, $t_* + \Delta t = t_{n_1})$, принадлежащей области расчета (заштрихована на рис. 9.14).

Продемонстрированный здесь способ конструирования разностных уравнений, основанный на аппроксимации интегральных законов сохранения по элементарной ячейке разностной сетки, называется *интегро-интерполяционным методом*.

Дополнительную и более подробную информацию по рассмотренным здесь вопросам можно найти в [2, с. 480–490], [4, с. 171–220], [7, с. 221–239], [9, с. 290–311], [12, с. 34–47, 259–290].

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Выписать для задачи (9.6) разностные схемы с шаблонами, показанными на рис. 9.15 и 9.16. Исследовать на аппроксимацию. Описать последовательность вычислений.

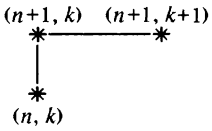


Рис. 9.15.

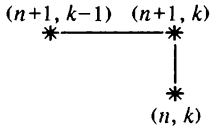


Рис. 9.16.

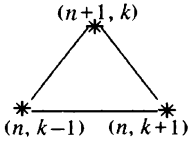


Рис. 9.17.

2. Исследовать на аппроксимацию схему Лакса для задачи Коши (9.3) (шаблон на рис. 9.17):

$$\frac{U_k^{n+1} - \frac{1}{2}(U_{k+1}^n + U_{k-1}^n)}{\tau} + a\frac{U_{k+1}^n - U_{k-1}^n}{2h} = f_k^n, \quad \begin{matrix} n = 0, 1, \ldots, \\ k = 0, \pm 1, \pm 2, \ldots, \end{matrix}$$

$$U_k^0 = \varphi(x_k).$$

Замечание. Это пример так называемой *негибкой* или, иначе, *условно аппроксимирующей схемы* (аппроксимирующей исходную задачу не для произвольного отношения между шагами сетки). ▲

3. Построить для задачи (9.3) разностные схемы второго порядка точности с шаблонами, показанными на рис. 9.18 и 9.19.

4. Для задачи (9.7) может быть выписано однопараметрическое семейство схем с шеститочечным шаблоном (рис. 9.20):

$$\frac{U_k^{n+1}-U_k^n}{\tau}=\mu\left[\sigma\frac{U_{k+1}^{n+1}-2U_k^{n+1}+U_{k-1}^{n+1}}{h^2}+(1-\sigma)\frac{U_{k+1}^n-2U_k^n+U_{k-1}^n}{h^2}\right],$$

$$0\leqslant\sigma\leqslant 1,$$

$$n=0,1,\ldots;\ k=0,\pm1,\pm2,\ldots,$$

$$U_k^0=\varphi(x_k).$$

Провести исследование на аппроксимацию данной схемы при произвольном σ. Каков результат при $\sigma=0.5$?

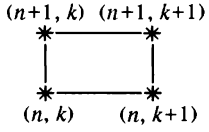


Рис. 9.18.

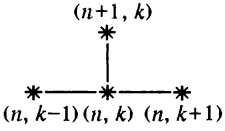


Рис. 9.19.

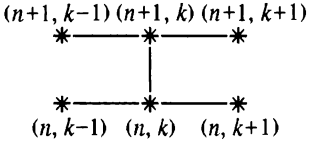


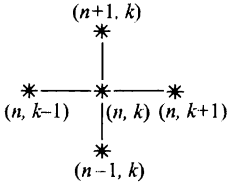
Рис. 9.20.

5. Построить разностную схему для решения смешанной задачи для волнового уравнения

$$U_{tt}-a^2U_{xx}=f(t,x)\ (a=\text{const})\quad t>0,\ \ 0<x<1,$$
$$U(0,x)=\varphi(x),\ \ U_t(0,x)=\varphi_1(x),$$
$$U(t,0)=\psi_1(t),\ \ U(t,1)=\psi_2(t).$$

Исследовать предложенную схему на аппроксимацию.

Замечание. Здесь естественным шаблоном для аппроксимации является трехслойный (по времени) шаблон



ЛЕКЦИЯ 10

УСТОЙЧИВОСТЬ РАЗНОСТНЫХ СХЕМ

Устойчивость линейных разностных схем. Устойчивость по начальным данным, правым частям, краевым условиям. Примеры анализа устойчивости простейших схем. Метод гармоник. Принцип «замороженных коэффициентов». Пример явной схемы для системы гиперболических уравнений. Пример исследования устойчивости нелинейной схемы.

Устойчивость линейных разностных схем. Пусть для решения исходной (дифференциальной) задачи:

$$LU = F \tag{10.1}$$

используется разностная схема

$$L_h U^{(h)} = F_h. \tag{10.1'}$$

Из теоремы, доказанной в Лекции 8, следует, что решение $U^{(h)} \underset{(h)\to 0}{\longrightarrow} [U]^{(h)}$, если (10.1′) аппроксимирует задачу (10.1) и $U^{(h)}$ устойчиво. Вопросам построения аппроксимирующих схем для исходной задачи была посвящена Лекция 9.

В данной лекции, занимаясь проблемами устойчивости разностного решения, мы ограничимся рассмотрением *линейных* задач, для которых операторы, определяющие вид левых частей уравнений (10.1) и (10.1′) линейны, т. е.

$$L(U + V) = LU + LV$$

и

$$L_h(U^{(h)} + V^{(h)}) = L_h U^{(h)} + L_h V^{(h)}.$$

Отметим, что в линейном случае определение устойчивости, данное в Лекции 8, может быть сформулировано в следующем виде.

Разностная схема (10.1′) устойчива, если для любой функции $F^{(h)}$ (10.1′) имеет единственное решение $U^{(h)}$, такое, что

$$\|U^{(h)}\| \leqslant C\|F^{(h)}\| \tag{10.2}$$

с константой C, не зависящей от параметров сетки (h). (См. «Вопросы и упражнения».)

Пример. Краевая задача для уравнения теплопроводности:

$$\begin{cases} U_t - \mu(t,x)U_{xx} = f(t,x)\ (\mu > 0) \text{ при } 0 < x < 1,\ t > 0, \\ U(0,x) = \varphi_(x), \\ U(t,0) = \psi_1(t),\ U(t,1) = \psi_2(t). \end{cases} \tag{10.3}$$

Замечание. Устанавливая связь данной задачи с абстрактной формулировкой 10.1, замечаем, что здесь

$$LU = \begin{cases} \dfrac{\partial U}{\partial t} - \mu(t,x)\dfrac{\partial^2 U}{\partial x^2} & \text{для } t > 0 \text{ и } 0 < x < 1, \\ U(0,x) & \text{для } t = 0,\ 0 \leqslant x \leqslant 1, \\ U(t,0) & \text{для } t > 0,\ x = 0, \\ U(t,1) & \text{для } t > 0,\ x = 1. \end{cases}$$

Соответственно,

$$F = \begin{cases} f(t,x), & t > 0,\ 0 < x < 1, \\ \varphi(x), & t = 0,\ 0 \leqslant x \leqslant 1, \\ \psi_1(t), & t > 0,\ x = 0, \\ \psi_2(t), & t > 0,\ x = 1. \end{cases}$$

Введем в рассмотрение сеточную область (обычным образом, как в Лекции 9) и рассмотрим для задачи (10.3) явную разностную схему (шаблон представлен на рис. 10.1):

$$\begin{cases} \dfrac{U_k^n - U_k^{n-1}}{\tau} - \mu_k^{n-1}\dfrac{U_{k+1}^{n-1} - 2U_k^{n-1} + U_{k-1}^{n-1}}{h^2} = f_k^n & n > 0,\ 0 < k < K = \dfrac{1}{h}, \\ U_k^0 = \varphi(x) & \text{для } 0 \leqslant k \leqslant K, \\ U_0^n = \psi_1(t_n),\ U_K^n = \psi_2(t_n) & \text{для } n = 1, 2, \ldots \end{cases} \tag{10.3'}$$

Сопоставляя (10.3′) с (10.1′), отмечаем, что $L_h U^{(h)}$ определяется левыми частями уравнений (10.3′). Соответственно,

$$F^{(h)} = \begin{cases} f^{(h)} & \text{во внутренних узлах сетки,} \\ \varphi^{(h)} & \text{в узлах нижнего слоя } (n = 0), \\ \psi^{(h)} & \text{в точках левой } (k = 0) \text{ и правой } (k = K) \text{ границ,} \end{cases}$$

где $f^{(h)} = \{f_k^n,\ n > 0,\ 0 < k < K\}$, $\varphi^{(h)} = \{\varphi(x_k),\ n = 0,\ 0 \leqslant k \leqslant K\}$, $\psi^{(h)} = \{\psi_1(t_n)$ при $k = 0$, $\psi_2(t_n)$ при $k = K$ $(1 \leqslant n \leqslant N)\}$.

Очевидно, для любой из векторных норм, которые можно привлечь для оценки величины $\|F^{(h)}\|$, справедливо

$$\|F^{(h)}\| \leqslant \|f^{(h)}\| + \|\varphi^{(h)}\| + \|\psi^{(h)}\|.$$

Существование и единственность решения (10.3′) не вызывает сомнений, поскольку очевиден способ его однозначного вычисления (он аналогичен тому, что обсуждался в Лекции 9, с той оговоркой, что здесь величины μ и f не постоянны). Следовательно, проверка устойчивости сводится к выяснению, при каких условиях выполнено неравенство, вытекающее из определения 10.2:

$$\|U^{(h)}\| \leqslant \text{const} \cdot (\|f^{(h)}\| + \|\varphi^{(h)}\| + \|\psi^{(h)}\|). \qquad (10.2')$$

Запись условия устойчивости в виде (10.2′) детализирует представление о факторах, от которых зависит устойчивость схемы: наличие $f^{(h)}$ в правой части (10.2′) означает *устойчивость по правым частям разностных уравнений*, $\varphi^{(h)}$ — *по начальным данным*, $\psi^{(h)}$ — *по граничным условиям*. Ясно, что устойчивость по каждому из упомянутых факторов является в общем случае необходимым условием устойчивости в целом (в смысле выполнения неравенства (10.2′) со всеми слагаемыми в правой части). С другой стороны, если исходная задача представляет собой задачу Коши для эволюционных уравнений, то определяющими факторами являются устойчивость по правым частям и по начальным данным. (Краевых условий в формулировке задачи Коши нет.) Напротив, для стационарной задачи (не зависящей от времени) нет начальных данных, и устойчивость определяется правыми частями уравнений и краевыми условиями. Это общие соображения.

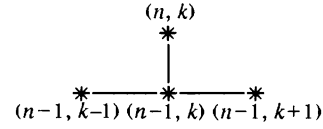


Рис. 10.1.

В дальнейшем на конкретных примерах мы убедимся, что проверка неравенства (10.2′) является наиболее актуальным моментом при анализе разностных схем с целью выбора конкретной схемы расчета из множества аппроксимирующих исходную задачу. Иногда удается установить справедливость неравенства (10.2′) безо всяких предположений о сеточных параметрах. В этом случае схема называется *абсолютно устойчивой*.

Если неравенство (10.2′) удовлетворяется в предположении о некотором соотношении между шагами сетки, то это предположение явля-

ется *достаточным условием устойчивости*. Если удается установить, что при некотором ограничении на сеточные параметры устойчивость имеет место лишь по отдельному фактору, то данное ограничение на шаг сетки трактуется как *необходимое условие устойчивости*.

Вернемся к нашему примеру, упростив его: будем считать, что рассматривается задача Коши для уравнения теплопроводности без правой части ($f^{(k)}(t, x) \equiv 0$), с постоянным коэффициентом $\mu > 0$. Соответствующая разностная схема имеет вид

$$\begin{cases} \dfrac{U_k^n - U_k^{n-1}}{\tau} - \mu \dfrac{U_{k+1}^{n-1} - 2U_k^{n-1} + U_{k-1}^{n-1}}{h^2} = 0, & n > 0;\ k = 0, \pm 1, \ldots, \\ U_k^0 = \varphi(x_k), & k^{(k)} = 0, \pm 1, \ldots \end{cases} \tag{10.3''}$$

Записав верхнее соотношение в виде

$$U_k^n = (1 - 2\rho)U_k^{n-1} + \rho(U_{k+1}^{n-1} + U_{k-1}^{n-1})$$

с $\rho = \mu\tau/h^2$, получаем формулу, определяющую решение в узлах n-го слоя по t через данные на предыдущем слое. В частности, для специального начального распределения $\varphi_k = \varphi(x_k) = (-1)^k\varepsilon$ будем иметь: на 1-м слое по t

$$U_k^1 = (1 - 2\rho)\varepsilon(-1)^k + \rho[(-1)^{k+1}\varepsilon + (-1)^{k-1}\varepsilon] = \\ = (1 - 4\rho)(-1)^k\varepsilon = (1 - 4\rho)\varphi_k,$$

на n-м слое

$$U_k^n = (1 - 4\rho)^n\varphi_k = (1 - 4\rho)^{(t_n/\tau)}\varphi_k. \tag{10.4}$$

При $\tau \to 0$ решение в каждом узле сетки будет оставаться ограниченным, только если $|1 - 4\rho| \leqslant 1$, т. е. для $-1 \leqslant 1 - 4\rho \leqslant 1$ или для $\rho = \mu\tau/h^2 \leqslant 1/2$. При выполнении этого условия в рассматриваемом случае имеет место устойчивость по начальным данным:

$$\|U^{(h)}\| = \max_{n,k} |U_k^n| \leqslant \max_k |\varphi_k| = \|\varphi^{(h)}\| = \varepsilon.$$

Очевидно, что это условие, полученное для упрощенной задачи, при специальных начальных данных можно рассматривать лишь в качестве необходимого условия устойчивости в более общем случае, для схемы (10.3′).

На примере этой схемы мы покажем теперь, как может быть иногда доказано выполнение непосредственно неравенства (10.2), означающего устойчивость рассматриваемой схемы при некоторых априорных предположениях.

Итак, пусть шаги сетки, с которыми используется схема (10.3′), удовлетворяют условию

$$\rho = \max_{n,k} \mu_k^n \frac{\tau}{h^2} \leqslant \frac{1}{2} \tag{10.4′}$$

(при $\mu = \text{const}$ это условие, как видно, совпадает с найденным выше необходимым условием устойчивости схемы). Перепишем верхнее соотношение (10.3′) в виде

$$U_k^n = \left(1 - \frac{2\mu_k^{n-1}\tau}{h^2}\right) U_k^{n-1} + \frac{\mu_k^{n-1}\tau}{h^2}\left(U_{k+1}^{n-1} + U_{k-1}^{n-1}\right) + \tau f_k^{(k)} n.$$

Оценивая величину U_k^n по модулю, получим

$$|U_k^n| \leqslant \left|1 - \frac{2\mu_k^{n-1}\tau}{h^2}\right| |U_k^{n-1}| + \frac{\mu_k^{n-1}\tau}{h^2}\left(|U_{k+1}^{n-1}| + |U_{k-1}^{n-1}|\right) + \tau |f_k^n| \leqslant$$

$$\leqslant \left(1 - \frac{2\mu_k^{n-1}\tau}{h^2}\right) \max_{0\leqslant k\leqslant K} |U_k^{n-1}| + 2\frac{\mu_k^{n-1}\tau}{h_{(k)2}} \max_{0\leqslant k\leqslant K} |U_k^{n-1}| + \tau \max_k |f_k^n| \leqslant$$

$$\leqslant \max_{c\leqslant k\leqslant K} |U_k^{n-1}| + \tau \max_k |f_k^n|.$$

В силу произвольности номера ($0 < k < K$), для которого получена оценка, можно написать

$$\max_{0<k<K} |U_k^n| \leqslant \max_{0\leqslant k\leqslant K} |U_k^{n-1}| + \tau \max_k |f_k^n|,$$

или, усиливая неравенство, за счет второго слагаемого в правой части:

$$\max_{0<k<K} |U_k^n| \leqslant \max_{0\leqslant k\leqslant K} |U_k^{n-1}| + \tau \max_{k,n} |f_k^n| = \max_k |U_k^{n-1}| + \tau \|f^{(h)}\|.$$

Наконец, учитываем краевые условия на n-м слое:

$$\max_{0\leqslant k\leqslant K} |U_k^n| \leqslant$$

$$\leqslant \max\{\max_{0\leqslant k\leqslant K} |U_k^{n-1}| + \tau \|f^{(h)}\|,\ \max\{|\psi_1(t_n)|, |\psi_2(t_n)|\}\}. \tag{10.5}$$

Неравенство (10.5) называется *принципом максимума* для схемы (10.3′). Из него следует условие устойчивости, так как, учитывая рекуррентность неравенства 10.5 по n, можно переписать его в виде

$$\max_k |U_k^n| \leqslant \max\{\max_k |U_k^{n-2}|\} + 2\tau \|f^{(h)}\|,$$

$$\max\{|\psi_1(t_n)|, |\psi_1(t_{n-1})|, |\psi_2(t_n)|, |\psi_2(t_{n-1})|\} \leqslant \ldots \leqslant$$

$$\leqslant \max\{\max_k |U_k^0| + n\tau \|f^{(h)}\|,\ \max\{\max_n (|\psi_1(t_n)|, \max_n |\psi_2(t_n)|)\}\}.$$

В силу произвольности номера n из последнего неравенства следует

$$\max_{n,k} |U_k^n| \leqslant \max\{\|\varphi^{(h)}\| + N\tau\|f^{(h)}\|, \|\psi^{(h)}\|\},$$

где

$$\|\varphi^{(h)}\| = \max_k |U_k^0| = \max_k |\varphi_k|,$$

$$\|\psi^{(h)}\| = \max\{(\max_n |\psi_1(t_n)|, \max_n |\psi_2(t_n))|\}.$$

И окончательно:

$$\|U^{(h)}\| \leqslant \max\{\|\varphi^{(h)}\| + T\|f^{(h)}\|, \|\psi^{(h)}\|\} \leqslant$$
$$\leqslant C(\|f^{(h)}\| + \|\varphi^{(h)}\| + \|\psi^{(h)}\|),$$

где

$$C = \begin{cases} T, & \text{при } T > 1, \\ 1, & \text{при } T \leqslant 1. \end{cases}$$

Мы получили неравенство (10.2′) для рассматриваемой схемы, т. е. доказали достаточность условия (10.4′) для устойчивости схемы (10.3′).

Подобного рода выкладки (связанные с установлением принципа максимума) приводят иногда к успеху при исследовании устойчивости некоторых схем для рассматриваемой и других задач. Но мы далее остановимся на обсуждении простого и достаточно универсального метода, который позволяет находить *необходимые условия устойчивости разностных схем для эволюционных задач.*

Метод гармоник. Как уже отмечалось, разностные схемы для эволюционных задач специфичны в том смысле, что решение разностных уравнений в узлах фиксированного слоя по времени определяется однозначно по данным на предыдущем слое, — двухслойные схемы (или по данным на нескольких предыдущих слоях — многослойные схемы).

Замечание. Для отдельных задач (например, для волнового уравнения) естественными являются трехслойные схемы. Для них решение в рассматриваемый момент времени определяется данными на двух предыдущих слоях. ▲

Напомним, что мы рассматриваем линейные задачи.

Оказывается, при некоторых дополнительных упрощающих предположениях решение соответствующих разностных уравнений может быть выписано в явном виде. Анализируя поведение этих решений, приходят к условиям, которые трактуются как необходимые при распространении их на схемы более общего вида.

Упрощающие предположения состоят в следующем. В рамках обсуждаемого сейчас метода гармоник при исследовании устойчивости разностных схем мы будем отвлекаться от правых частей разностных уравнений (полагая их нулевыми) и краевых условий, т.е. будем рассматривать схемы, аппроксимирующие задачу Коши для однородных дифференциальных уравнений. Кроме того, будем считать постоянными коэффициенты этих уравнений («замораживая» их, если фактически они не постоянны в исходной задаче).

В этих условиях разностные уравнения имеют частные решения вида гармоник произвольной частоты ω:

$$u_k^n = C\lambda^n e^{i\omega x_k} = C\lambda^n e^{i\omega kh} = C\lambda^n e^{ik\alpha}, \tag{10.6}$$

где $C = \text{const}$, $\alpha = \omega h$, ω — произвольное натуральное число, $\lambda = \lambda(\alpha, \tau, h)$ подлежит отысканию для каждой конкретной схемы.

Мы будем убеждаться в наличии решений вида (10.6) после оговоренных упрощений разностных схем во всех конкретных примерах, которые будем рассматривать.

После сделанных упрощений входными данными для получения схемы являются лишь начальные условия.

Вытекающее из (10.2) условие устойчивости по начальным даным для решений (10.6) сводится к требованию ограниченности амплитуды этих гармоник:

$$|\lambda^n| \leqslant \text{const}. \tag{10.7}$$

Далее мы увидим, что для каждой схемы будет получаться своя зависимость амплитуды от параметров сетки и параметра α: $\lambda = \lambda(\tau, h, \alpha)$. Требуя выполнения неравенства (10.7) при произвольном α (т.е. для произвольной гармоники), находим необходимое условие устойчивости рассматриваемой схемы в виде некоторого ограничения на шаги сетки τ, h.

Учитывая, что $n = t_n/\tau$, нетрудно убедиться, что проверка неравенства (10.7) эквивалентна проверке более простого условия:

$$|\lambda| \leqslant 1 + A\tau, \quad \text{где} \quad A = \text{const}. \tag{10.7'}$$

В самом деле, если выполнено (10.7'), то

$$\begin{aligned} |\lambda^n| = |\lambda|^n \leqslant (1 + A\tau)^{t_n/\tau} &= \exp\{(t_n/\tau)\ln(1 + A\tau)\} = \\ &= \exp\{(t_n/\tau)[A\tau - A^2\tau^2/2 + O(\tau^3)]\} \leqslant \\ &\leqslant \exp\{(t_n/\tau)[A\tau]\} = \exp\{At_n\}, \end{aligned} \tag{10.7''}$$

и, стало быть, при любом конечном значении t_n выполнено требование (10.7).

Неравенство (10.7′) называется *условием Неймана устойчивости разностных схем для эволюционных задач.*

Фактически с учетом оговоренных выше упрощающих предположений, лежащих в основе метода гармоник, мы, как уже отмечалось, находим условие устойчивости по начальным данным. Анализ показывает, что большей частью оно является определяющим для устойчивости разностных схем, аппроксимирующих эволюционные задачи.

Перейдем к исследованию устойчивости конкретных схем.

Пример 1. Схема (10.3′) после упрощений, на которые опирается метод гармоник, переходит в (10.3″). Будем искать частные решения разностных уравнений в виде (10.6). Подставляем (10.6) в уравнения и, сокращая на $C\lambda^{n-1}e^{ik\alpha}$, получаем

$$\frac{\lambda-1}{\tau}-\mu\frac{e^{-i\alpha}-2+e^{-i\alpha}}{h^2}=0.$$

Отсюда $\lambda=1+\rho(e^{i\alpha}-2+e^{-i\alpha})$, где $\rho=\mu\tau/h^2$, или

$$\lambda=1+2\rho(\cos\alpha-1)=1-4\rho\sin^2\frac{\alpha}{2}.$$

Привлекаем условие Неймана.

Замечание. В тех случаях, когда выражение для λ не зависит явно от τ (т.е. τ входит только в некотором отношении с h), неравенство (10.7′) можно заменить более простым условием $|\lambda|\leqslant 1$. ▲

Учитывая сделанное замечание, получаем

$$-1\leqslant 1-4\rho\sin^2\frac{\alpha}{2}\leqslant 1,$$

и окончательно

$$\rho=\frac{\mu\tau}{h^2}\leqslant\frac{1}{2}. \tag{10.8}$$

Это классическое условие устойчивости явной разностной схемы для параболического уравнения.

Замечание 1. Мы уже получали это условие (выше). Но тогда это было сделано на основе анализа существенно более узкого класса частных решений (10.4). ▲

Замечание 2. *Принцип «замороженных» коэффициентов.* Для уравнений с непостоянными коэффициентами метод гармоник применяется в предположении, что коэффициенты «заморожены» (постоянны). Затем в окончательном условии коэффициенты «размораживаются». Применительно к рассматриваемой схеме последнее означает переход от условия (10.8) к условию $\rho=\max\limits_{t,x}\mu(t,x)\tau/h^2\leqslant 1/2$.

Выше была доказана достаточность этого условия для схемы (10.3′) без каких-либо упрощающих предположений. Это подтверждает достоверность условий устойчивости, получаемых по методу гармоник. ▲

Пример 2. Неявная схема для задачи (10.3):

$$\begin{cases} \dfrac{U_k^{n+1}-U_k^n}{\tau}-\mu_k^{n-1}\dfrac{U_{k+1}^{n+1}-2U_k^{n+1}+U_{k-1}^{n+1}}{h^2}=f_k^n, & n>0,\ 0<k<K=\dfrac{1}{h}, \\ U_k^0=\varphi(x_k), & \text{для } 0\leqslant k\leqslant K, \\ U_0^n=\psi_1(t_n),\ U_K^n=\psi_2(t_n), & \text{для } n=1,2,\ldots \end{cases}$$

Подставляя после стандартных упрощающих предположений (10.6) в разностные уравнения, после несложных выкладок находим:

$$\lambda=\frac{1}{1+4\frac{\mu\tau}{h^2}\sin^2\frac{\alpha}{2}}.$$

Очевидно, $|\lambda|=\lambda\leqslant 1$ всегда, независимо от шагов сетки. Соответственно схема *абсолютно устойчива*.

Замечание. Строго говоря, в рамках данного подхода полученный результат следует трактовать так: необходимых условий устойчивости, которые могут быть получены при допущениях, сопутствующих методу гармоник, для данной схемы нет. Но, как уже отмечалось, выводы относительно устойчивости, к которым приводит метод гармоник, подтверждаются, большей частью, практическими расчетами. В частности, неявные схемы, как правило, являются абсолютно устойчивыми. ▲

Пример 3. Краевая задача для уравнения переноса:

$$\begin{cases} U_t+aU_x=f(t,x), & t>0,\ 0<x\leqslant 1,\ a>0, \\ U(0,x)=\varphi(x), & 0\leqslant x\leqslant 1, \\ U(t,0)=\psi(t), & 0<t\leqslant T. \end{cases} \tag{10.9}$$

Рассмотрим схему с шаблоном на рис. 10.2:

$$\begin{cases} \dfrac{U_k^{n+1}-U_k^n}{\tau}+a\dfrac{U_k^n-U_{k-1}^n}{h}=f_k^n, & n=0,1,\ldots,N-1;\ k=1,2,\ldots,K; \\ U_k^0=\varphi(x_k), & k=0,1,\ldots,K; \\ U_0^n=\psi(t_n), & n=1,2,\ldots,N. \end{cases} \tag{10.9′}$$

Применяя метод гармоник, проводим необходимые выкладки:

$$\frac{\lambda-1}{\tau}+a\frac{1-e^{-i\alpha}}{h}=0,$$

$$\lambda=(1-\rho)+\rho e^{-i\alpha},\quad \rho=\frac{a\tau}{h}=\text{const}.$$

В общем случае в рамках метода гармоник для λ получается комплексное выражение. В таком случае, можно выписать формулу для $|\lambda|^2$ и, рассматривая неравенство $|\lambda|^2\leqslant 1$, искать условия, при которых оно выполняется. Здесь, в данном примере, проще привлечь геометрические соображения.

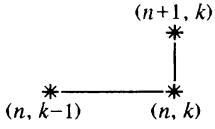


Рис. 10.2.

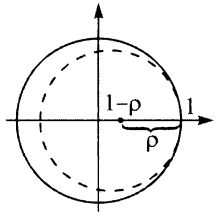


Рис. 10.3.

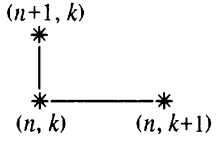


Рис. 10.4.

Учитывая произвольность α, в качестве возможных значений λ имеем круг на комплексной плоскости (рис. 10.3) радиуса ρ с центром в точке $(1-\rho)$ на действительной оси. Очевидно, что $|\lambda|\leqslant 1$ (т.е. при любом α λ не выходит за пределы единичного круга на комплексной плоскости) тогда и только тогда, когда

$$0\leqslant\rho=\frac{a\tau}{h}\leqslant 1 \tag{10.10}$$

типичное условие устойчивости явной схемы для уравнения переноса (его иногда называют условием Куранта, соответственно ρ — число Куранта).

Пример 4. Для той же задачи (10.9) рассмотрим схему с шаблоном, представленным на рис. 10.4:

$$\begin{cases}\dfrac{U_k^{n+1}-U_k^n}{\tau}+a\dfrac{U_{k+1}^n-U_k^n}{h}=f_k^n, & n=0,1,\ldots,N-1,\ k=0,1,\ldots,K-1,\\ U_k^0=\varphi(x_k), & k=0,1,\ldots,K,\\ U_0^n=\psi(t_n), & n=1,2,\ldots,N.\end{cases} \tag{10.9$''$}$$

Привлекая метод гармоник для анализа устойчивости этой схемы, находим

$$\lambda=(1+\rho)-\rho e^{i\alpha}\quad(\rho=a\tau/h).$$

Очевидно, при любом α (кроме $\alpha = 0$) значение λ лежит за пределами единичного круга на комплексной плоскости, т. е. $|\lambda| > 1$. Необходимый критерий устойчивости не выполнен ни при каком отношении между τ и h, схема неустойчива и, стало быть, не годится для проведения расчетов.

Замечание 1. В Лекции 9 мы уже заметили, что эта схема не реализуема с точки зрения организации вычислений. Вспомним о «физическом» смысле решения задачи (10.9), которое представляет собой при $a > 0$ перенос начального возмущения по пространственной переменной слева направо (см. обсуждение этой задачи в Лекции 9). С учетом этого обстоятельства можно сказать, что схема (10.9′) отвечает «физике дела»: решение в каждом узле зависит от данных в узлах, расположенных левее в предыдущие моменты времени. Схема (10.9″), напротив, «переносит» текущее состояние справа налево. ▲

Замечание 2. Если $a < 0$, то в (10.9) краевое условие должно быть поставлено на правой границе области, решение представляет собой перенос возмущения справа налево (по пространственной переменной). Схема (10.9′) будет всегда неустойчивой, схема (10.9″) устойчива при выполнении условия Куранта $|a|\tau/h \leqslant 1$. ▲

ДОПОЛНЕНИЕ К ЛЕКЦИИ 10

Краевая задача для системы гиперболических уравнений. Рассматривается задача

$$\begin{aligned}
&U_t + aV_x = f_1(t,x), \quad \text{при } 0 < t \leqslant T,\ 0 < x < 1,\\
&V_t + bU_x = f_2(t,x), \quad ab > 0,\\
&\left.\begin{aligned} U(0,x) &= \varphi_1(x),\\ V(0,x) &= \varphi_2(x) \end{aligned}\right\} \ \text{— начальные данные, } (0 \leqslant x \leqslant 1)\\
&\left.\begin{aligned} \alpha_1 U(t,0) + \beta_1 V(t,0) &= \psi_1(t),\\ \alpha_2 U(t,1) + \beta_2 V(t,1) &= \psi_2(t) \end{aligned}\right\} \ \text{— краевые условия, } (t > 0)
\end{aligned} \tag{10.11}$$

Замечание. Корректность заданных таким образом граничных условий поясняется ниже. ▲

Условие $ab > 0$ существенно. Оно обеспечивает гиперболичность (эволюционность) задачи. В самом деле, дифференцируя первое уравнение по t, получим

$$U_{tt} + aV_{xt} = (f_1)_t.$$

Дифференцируя второе уравнение по x, имеем

$$V_{tx} + bU_{xx} = (f_2)_x.$$

Умножая последнее соотношение на a и вычитая из предыдущего, получим в качестве следствия исходных уравнений

$$U_{tt} - (ab)U_{xx} = g(t, x) = [(f_1)_t - a(f_2)_x]. \qquad (10.12)$$

(Аналогичное уравнение получается для V.)

Очевидно, уравнение (10.12) является гиперболическим (уравнением малых колебаний струны), только если $ab > 0$. Если $ab < 0$, то (10.12) — эллиптическое уравнение, описывающее равновесное состояние (t нельзя трактовать как время).

З а м е ч а н и е. Система (10.11), как и прежние модельные уравнения, может быть наполнена физическим смыслом. Например, если U — электрический ток, V — напряжение, то это система уравнений электроцепи (a, b при этом выражаются через коэффициенты сопротивления и емкости). ▲

Для построения разностной схемы, как обычно, вводим сеточную область с параметрами τ и h (рис. 10.5). При использовании стандартных приемов для составления разностных уравнений возможностей здесь больше, чем для одного уравнения, так как при аппроксимации каждого уравнения формально можно использовать свой шаблон.

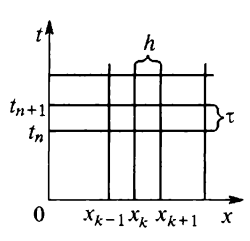


Рис. 10.5.

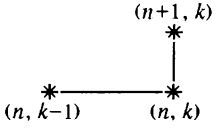


Рис. 10.6.

Построим простейшую схему, аппроксимируя оба уравнения по шаблону, изображенному на рис. 10.6:

$$\begin{cases} \dfrac{U_k^{n+1} - U_k^n}{\tau} + a\dfrac{V_k^n - V_{k-1}^n}{h} = (f_1)_k^n, & n = 0, 1, \ldots, N-1, \\ \dfrac{V_k^{n+1} - V_k^n}{\tau} + b\dfrac{U_k^n - U_{k-1}^n}{h} = (f_2)_k^n, & k = 1, 2, \ldots, K, \\ U_k^0 = \varphi_1(x_k), \quad V_k^0 = \varphi_2(x_k), & k = 0, 1, \ldots, K, \\ \alpha_1 U_0^n + \beta_1 V_0^n = \psi_1(t_n), & n = 1, \ldots, N. \\ \alpha_2 U_0^n + \beta_2 V_0^n = \psi_2(t_n), & \end{cases} \qquad (10.13)$$

Исследуя данную схему на аппроксимацию, мы должны найти невязки всех уравнений (10.13) на решениях $U(t, x)$, $V(t, x)$ исходной задачи. Выбрав в качестве «опорной точки» узел (n, k), из уравнений первой строки (10.13), предполагая, что существуют и ограничены вторые производные от U, V по своим аргументам, получим

$$(\psi_1)_k^n = \left(\frac{\tau}{2} U_{tt} - \frac{ah}{2} V_{xx}\right)_k^n + O(\tau^2, h^2),$$

из уравнений второй строки:

$$(\psi_2)_k^n = \left(\frac{\tau}{2} V_{tt} - \frac{ah}{2} U_{xx}\right)_k^n + O(\tau^2, h^2).$$

Соотношения в последних трех строчках (10.13) аппроксимируют соответствующие начальные и граничные условия из (10.11) точно.

Таким образом, главные члены ошибок аппроксимации суть величины первого порядка малости как по τ, так и по h.

Метод гармоник для анализа устойчивости разностных схем для сеточных функций $U^{(h)}, V^{(h)}$ состоит в том, что в каждом узле компоненты этих функций ищем в виде

$$\begin{pmatrix} U_k^n \\ V_k^n \end{pmatrix} = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \lambda^n e^{ik\alpha}. \tag{10.14}$$

Подставляя (10.14) в разностные уравнения из (10.13) и, как обычно, полагая $f_1 \equiv f_2 \equiv 0$, а $-\infty < k < \infty$, получим после сокращения на $\lambda^n e^{ik\alpha}$:

$$\begin{cases} C_1 \dfrac{\lambda - 1}{\tau} + C_2 a \dfrac{1 - e^{-i\alpha}}{h} = 0, \\ C_2 \dfrac{\lambda - 1}{\tau} + C_1 b \dfrac{1 - e^{-i\alpha}}{h} = 0. \end{cases} \tag{10.15}$$

Нас интересуют нетривиальные решения вида (10.14) (C_1, $C_2 \neq 0$). Рассматривая (10.15) как однородную систему для C_1, C_2, мы, следовательно, должны приравнять нулю ее определитель:

$$\begin{vmatrix} \dfrac{\lambda - 1}{\tau} & a \dfrac{1 - e^{-i\alpha}}{h} \\ b \dfrac{1 - e^{-i\alpha}}{h} & \dfrac{\lambda - 1}{\tau} \end{vmatrix} = 0.$$

Отсюда получаем допустимые значения λ:

$$\lambda = (1 \mp \rho) \pm \rho e^{-i\alpha} \quad \left(\rho = \frac{\tau\sqrt{ab}}{h}\right).$$

Ясно, что нижней последовательности знаков в выражении для λ при любых α соответствуют λ, лежащие за пределами единичного круга комплексной области. Следовательно, необходимое условие устойчивости $|\lambda| \leqslant 1$ не выполнено ни при каких ρ, схема (10.13) *неустойчива*, для расчета непригодна.

Здесь мы столкнулись с принципиальным фактом: при переходе от скалярных уравнений, которые мы рассматривали прежде, к системам (например, вида (10.11)) построение устойчивой явной схемы становится нетривиальной задачей. (Нетрудно построить неявную устойчивую схему, но в этом случае серьезные осложнения возникнут на стадии организации вычислений. См. упражнение 7.)

Чтобы справиться с этой задачей, мы заменим дифференциальные уравнения (10.11) другими, эквивалентными, но, что важно, более простыми. Это оказывается возможным благодаря гиперболичности системы (10.11). Итак, рассмотрим линейную комбинацию дифференциальных уравнений из (10.11) пока с неопределенными коэффициентами α и β:

$$(\alpha U + \beta V)_t + (\alpha a V + \beta b U)_x = \alpha f_1 + \beta f_2. \tag{10.16}$$

Попробуем подобрать параметры α и β так, чтобы и по t, и по x дифференцировалась одна и та же функция. Для этого надо, чтобы при любых U и V имела место пропорциональность

$$\alpha U + \beta V = \varkappa(\alpha a V + \beta b U),$$

т. е. чтобы

$$\alpha = \varkappa \beta b, \quad \beta = \varkappa \alpha a. \tag{10.17}$$

Для трех неопределенных параметров имеем два соотношения. Следствием их (результатом деления одного на другое) является соотношение, связывающее α и β: $(\alpha/\beta)^2 = b/a$, или

$$\frac{\alpha}{\beta} = \pm\sqrt{\frac{b}{a}}. \tag{10.18}$$

Один параметр выбираем по своему усмотрению. Для определенности положим $\beta = 1$. Два значения α определяются из (10.18): $\alpha = \pm\sqrt{b/a}$. Соответственно два значения $\varkappa$ — из (10.17): $\varkappa = \pm(\sqrt{ab})^{-1}$.

Подставляя найденные значения α и $\varkappa$ (по очереди со знаками «+» и «−») вместе с $\beta = 1$ в (10.16), убеждаемся, что существуют две независимые комбинации (10.16) желаемого вида:

$$\left(\sqrt{\frac{b}{a}}U + V\right)_t + \sqrt{ab}\left(\sqrt{\frac{b}{a}}U + V\right)_x = \sqrt{\frac{b}{a}}f_1 + f_2$$

и

$$\left(-\sqrt{\frac{b}{a}}U+V\right)_t-\sqrt{ab}\left(-\sqrt{\frac{b}{a}}U+V\right)_x=-\sqrt{\frac{b}{a}}f_1+f_2.$$

Вводя в рассмотрение новые функции

$$p=\sqrt{\frac{b}{a}}U+V,\quad r=-\sqrt{\frac{b}{a}}U+V, \tag{10.19}$$

однозначно связанные с исходными, так что

$$U=\frac{p-r}{2\sqrt{b/a}},\quad V=\frac{p+r}{2}, \tag{10.20}$$

мы можем перейти к формулировке исходной задачи для новых функций — p и r:

$$\begin{aligned}
&p_t+cp_x=g_1 && \left(c=\sqrt{ab},\quad g_1=\sqrt{\frac{b}{a}}f_1+f_2\right),\\
&r_t-cr_x=g_2 && \left(g_2=-\sqrt{\frac{b}{a}}f_1+f_2\right),\\
&p(0,x)=\widetilde{\varphi}_1(x) && \left(\widetilde{\varphi}_1(x)=\sqrt{\frac{b}{a}}\varphi_1(x)+\varphi_2(x)\right),\\
&r(0,x)=\widetilde{\varphi}_2(x) && \left(\widetilde{\varphi}_2(x)=-\sqrt{\frac{b}{a}}\varphi_1(x)+\varphi_2(x)\right),
\end{aligned} \tag{10.21}$$

$$\widetilde{\alpha}_1p(t,0)+\widetilde{\beta}_1r(t,0)=\psi_1(t)$$
$$\left(\widetilde{\alpha}_1=\frac{1}{2}\left(\frac{\alpha_1}{\sqrt{b/a}}+\beta_1\right),\ \widetilde{\beta}_1=\frac{1}{2}\left(-\frac{\alpha_1}{\sqrt{b/a}}+\beta_1\right)\right),$$

$$\widetilde{\alpha}_2p(t,1)+\widetilde{\beta}_2r(t,1)=\psi_2(t)$$
$$\left(\widetilde{\alpha}_2=\frac{1}{2}\left(\frac{\alpha_2}{\sqrt{b/a}}+\beta_2\right),\ \widetilde{\beta}_2=\frac{1}{2}\left(-\frac{\alpha_2}{\sqrt{b/a}}+\beta_2\right)\right).$$

Очевидно (в силу проделанных выкладок), задача (10.21) эквивалентна исходной задаче (10.11).

Замечание. Функции p и r, для которых уравнения разделяются в том смысле, что в каждом уравнении дифференцируется только одна из них, называются *инвариантами Римана*. Наличие

инвариантов Римана есть следствие гиперболичности исходной системы. Ввиду того, что уравнения для инвариантов являются знакомыми нам уравнениями переноса, мы приходим к выводу, что инвариант p распространяется слева направо вдоль характеристик $dx/dt = c$ ($x = ct + \text{const}$ — так называемые c_+-характеристики). Инвариант r переносится справа налево вдоль c_--характеристик: $dx/dt = -c$ ($x = -ct + \text{const}$). Таким образом, для исходной системы существуют два семейства действительных характеристик, вдоль которых дифференциальные уравнения для инвариантов могут быть представлены в виде обыкновенных дифференциальных уравнений (см. Лекцию 9, соотношение (9.5)), в чем, собственно, и выражается свойство гиперболичности системы (10.11). ▲

Теперь уже не составляет труда построить явную устойчивую схему для решения задачи. Известно (см. примеры 3 и 4 на с. 153–154, что аппроксимация уравнения для p по шаблону, изображенному на рис. 10.7, приводит к устойчивому решению при выполнении условия Куранта $c\tau/h \leqslant 1$. При том же условии устойчива аппроксимация уравнения для r по шаблону, изображенному на рис. 10.8.

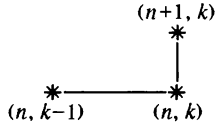


Рис. 10.7.

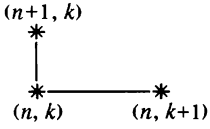


Рис. 10.8.

Уравнения, описывающие переход от n-го слоя по t к $(n+1)$-му, выглядят так:

$$\begin{gathered}
\frac{p_k^{n+1} - p_k^n}{\tau} + c\frac{p_k^n - p_{k-1}^n}{h} = (g_1)_k^n, \quad k = 1, 2, \ldots, K, \\
\frac{r_k^{n+1} - r_k^n}{\tau} - c\frac{r_{k+1}^n - r_k^n}{h} = (g_2)_k^n, \quad k = 0, 1, 2, \ldots, K-1, \\
p_k^0 = \widetilde{\varphi}_1(x_k), \quad r_k^0 = \widetilde{\varphi}_2(x_k), \quad k = 0, 1, \ldots, K,
\end{gathered} \tag{10.22}$$

$$\widetilde{\alpha}_1 p_0^n + \widetilde{\beta}_1 r_0^n = \psi_1(t_n), \qquad \widetilde{\alpha}_2 p_K^n + \widetilde{\beta}_2 r_K^n = \psi_2(t_n) \quad n > 0.$$

Последовательность вычислений при переходе от n-го временного слоя к $(n+1)$-му очевидна. Используя уравнения для p, вычисляем p_k^{n+1} во всех точках, отмеченных крестиком на рис. 10.9. Из уравнений для r находим r_k^{n+1} в точках, отмеченных кружочками. Далее, зная в точке левой границы значение r_0^{n+1}, из предпоследнего уравнения (10.22) находим p_0^{n+1}. Из краевого условия на правой границе

(последнее уравнение (10.22)) вычисляем r_K^{n+1}. На этом расчет очередного слоя заканчивается.

Замечание 1. Описанный способ расчета, очевидно, приводит к единственному решению (если $\widetilde{\alpha}_1^{(k)} \neq 0$ и $\widetilde{\beta}_2 \neq 0$), что свидетельствует в этом случае о корректности граничных условий, заданных в исходной формулировке задачи. Это утверждение справедливо для разностной задачи. Но, вообще говоря, можно установить его справедливость и для исходной (дифференциальной) задачи. ▲

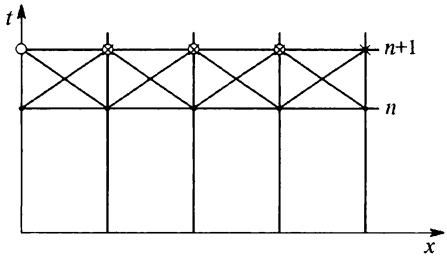


Рис. 10.9.

Замечание 2. Любопытно посмотреть, как выглядит устойчивая явная схема для исходных функций U и V. Для этого заменим в схеме (10.22) p и r на U и V по формулам (10.19). После несложных преобразований (сложений и вычитаний уравнений друг из друга) получим для внутренних точек сетки следующие соотношения:

$$\frac{U_k^{n+1} - U_k^n}{\tau} + a\frac{V_{k+1}^n - V_{k-1}^n}{2h} - \sqrt{ab}\,\frac{h}{2}\,\frac{U_{k+1}^n - 2U_k^n + U_{k-1}^n}{h^2} = (f_1)_k^n,$$

$$\frac{V_k^{n+1} - V_k^n}{\tau} + b\frac{U_{k+1}^n - U_{k-1}^n}{2h} - \sqrt{ab}\,\frac{h}{2}\,\frac{V_{k+1}^n - 2V_k^n + V_{k-1}^n}{h^2} = (f_2)_k^n. \quad ▲$$

Замечание 3. Разумеется, для задачи (10.21) можно предложить и другие (явные и неявные) удобные для расчета схемы. ▲

Замечание 4. В [18] обсуждается вариант явной схемы непосредственно для исходной системы (10.11) на так называемой «шахматной» сеточной области. ▲

В рассмотренном разделе я постарался в максимально простом виде привести иллюстрацию чрезвычайно эффективного метода решения гиперболических систем уравнений в частных производных, называемого сеточно-характеристическим методом (см. [19, с. 386–388], а также монографию [28]).

Об устойчивости нелинейной разностной схемы. Рассмотрим следующую задачу:

$$\begin{cases} U_t + UU_x = 0 \ \text{ при } t > 0 \text{ и } -\infty < x < \infty, \\ U(0, x) = \varphi(x). \end{cases} \tag{10.23}$$

Выпишем какую-либо разностную схему, например,

$$\begin{cases} \dfrac{U_k^{n+1}-U_k^n}{\tau}+U_k^n\dfrac{U_k^n-U_{k-1}^n}{h}=0, \quad U_k^0=\varphi_k, \\ n=0,1,\ldots; \quad k=0,\pm1,\pm2,\ldots \end{cases} \tag{10.24}$$

При анализе устойчивости подобной (нелинейной) схемы мы не можем использовать метод гармоник, так как частные решения $u_k^n = C\lambda^n e^{ik\alpha}$, которые при этом привлекаются, имеют смысл только для линейных разностных уравнений (с постоянными коэффициентами).

Мы прибегнем здесь к следующему эвристическому (подтверждаемому опытом) способу рассуждений.

Пусть $\overline{U}_k^0 = \varphi_k + \delta_k^0$, так что $\|\delta^0\| = \max\limits_k |\delta_k^0| \ll 1$. Соответствующее решение $\overline{U}^{(h)} = \{\overline{U}_k^n\}$; $\delta_k^n = \overline{U}_k^n - U_k^n$ — возмущение решения, вызванное малыми возмущениями входных данных $\{\delta_k^0\}$.

Если схема устойчива, то, очевидно (из определения устойчивости), $\|\delta^{(h)}\| = \max\limits_{k,n} |\delta_k^n| \sim \|\delta^0\| \sim o(1)$.

Будем предполагать, что устойчивость имеет место. Опираясь на это предположение, мы получим далее линеаризованные уравнения для возмущений δ_k^n и, применяя метод гармоник (уже к линейному уравнению!), получим условия, при которых предположение об устойчивости оправдано.

Возмущение решения δ_k^n удовлетворяет разностным уравнениям, которые получаются после подстановки $\overline{U}_k^n = U_k^n + \delta_k^n$ в соотношения, из которых находится возмущенное решение, т.е. в (10.24), записанные для $\overline{U}_k^n$:

$$\frac{\delta_k^{n+1}-\delta_k^n}{\tau}+\frac{U_k^{n+1}-U_k^n}{\tau}+(U_k^n+\delta_k^n)\frac{(U_k^n+\delta_k^n)-(U_{k-1}^n+\delta_{k-1}^n)}{h}=0,$$

или

$$\frac{\delta_k^{n+1}-\delta_k^n}{\tau}+U_k^n\frac{\delta_k^n-\delta_{k-1}^n}{h}+\frac{U_k^n-U_{k-1}^n}{h}\delta_k^n+\delta_k^n\frac{\delta_k^n-\delta_{k-1}^n}{h}$$

$$=-\left(\frac{U_k^{n+1}-U_k^n}{\tau}+U_k^n\frac{U_k^n-U_{k-1}^n}{h}\right).$$

В силу начального предположения об устойчивости δ_k^n — малы, поэтому подчеркнутым штриховой линией слагаемым пренебрегаем как величиной второго порядка малости и получаем линейные уравнения

для величин δ_k^n:

$$\frac{\delta_k^{n+1}-\delta_k^n}{\tau}+U_k^n\frac{\delta_k^n-\delta_{k-1}^n}{h}+\frac{U_k^n-U_{k-1}^n}{h}\delta_k^n=$$
$$=-\left(\frac{U_k^{n+1}-U_k^n}{\tau}+U_k^n\frac{U_k^n-U_{k-1}^n}{h}\right).$$

Заметим, что правые части этих уравнений представляют собой погрешность аппроксимации исходных (нелинейных) уравнений (10.24) во внутренних узлах сетки.

Замораживая коэффициенты ($U_k^n=a$, $(U_k^n-U_{k-1}^n)/h=b$), применяем для анализа устойчивости метод гармоник:

$$\frac{\lambda-1}{\tau}+a\frac{1-e^{-i\alpha}}{h}+b=0$$
$$\lambda=(1-\rho)+\rho e^{-i\alpha}+b\tau\quad\left(\rho=a\frac{\tau}{h}\right).$$

Здесь λ явно зависит от τ, поэтому апеллируем к критерию Неймана (см. условие (10.7′) и Замечание на с. 152). Очевидно, если $0\leqslant\rho\leqslant 1$ и $|b|\leqslant\text{const}$, то $|\lambda|\leqslant 1+|b|\tau$ — *критерий Неймана* выполнен! Далее, «размораживая» коэффициенты в условиях, к которым пришли, и учитывая, что $\dfrac{U_k^n-U_{k-1}^n}{h}\approx\left(\dfrac{\partial U}{\partial t}\right)_k^n$, приходим к окончательным *необходимым условиям устойчивости* схемы для δ_k^n, а вместе с тем (как следует из логики рассуждений) и для *нелинейной* схемы (10.24):

$$0\leqslant\frac{\tau U_k^n}{h}\leqslant 1\ \text{(для всех } n \text{ и } k\text{), или } U_k^n\geqslant 0 \text{ и } \frac{\tau U_k^n}{h}\leqslant 1, \qquad (10.25)$$

$$\left|\frac{U_k^n-U_{k-1}^n}{h}\right|\leqslant\text{const (для всех } n \text{ и } k\text{), или } \left|\frac{\partial U}{\partial x}\right|_k^n\leqslant\text{const}. \qquad (10.26)$$

Замечание 1. Из (10.25) следует, что по схеме (10.24) можно вычислять только такое решение задачи (10.23), которое остается неотрицательным при всех значениях независимых переменных x и t. Далее, при счете шаг τ для перехода с одного временного слоя на другой надо выбирать переменным: $\tau=\tau_n$, так чтобы

$$\tau_n=t_{n+1}-t_n\leqslant\frac{h}{\max\limits_k U_k^n}. \qquad \blacktriangle$$

Замечание 2. Мы могли бы получить условие (10.25), формально применив принцип «замораживания» коэффициентов в нелинейной схеме (10.24), но здесь мы получили еще одно требование —

(10.26), которое, впрочем, носит не конструктивный, т.е. приводящий к какому-то правилу выбора шагов сетки, а ориентировочный характер, т. е. «предупреждает», что в численном решении может появиться неустойчивость, если искомое решение исходной задачи (10.23) характеризуется в какой-то области независимых переменных большими градиентами: $\left|\dfrac{\partial U}{\partial x}\right|$ велико. ▲

Более глубокое и строгое обсуждение рассмотренных здесь вопросов можно найти в следующих источниках: [2, с. 480–526], [4, с. 221–283, 362–418], [5, с. 44–47], [9, с. 311–389, 424–434, 439–451], [12, с. 339–377].

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Установить эквивалентность определений устойчивости: (10.2) в данной лекции и (8.24) в Лекции 8.

2. Доказать достаточность условия $0 \leqslant \rho = a\tau/h \leqslant 1$ для устойчивости разностной схемы (10.9′).

3. Для задачи (10.9) исследовать на устойчивость разностные схемы с шаблонами, показанными на рис. 10.10.

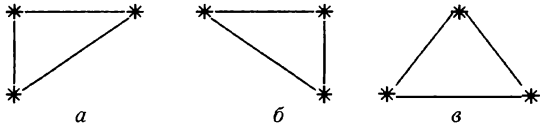


Рис. 10.10.

4. То же задание для схем второго порядка аппроксимации с шаблонами на рис. 10.11.

5. Исследовать на устойчивость неявную схему с шаблоном на рис. 10.12 для задачи (10.3).

6. Сравнить число операций, которые требуется выполнить при расчете величин на одном шаге по времени, при решении задачи (10.3) по явной и неявной четырехточечным схемам.

В чем преимущество неявной схемы для численного решения этой задачи?

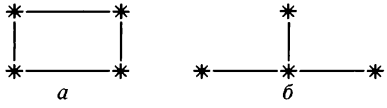


Рис. 10.11.

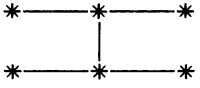


Рис. 10.12.

Рис. 10.13.

7. Для задачи (10.11) построить схему по шаблону рис. 10.13. Исследовать на аппроксимацию и устойчивость. Подумайте, как организовать вычисления.

ЛЕКЦИЯ 11

РАЗНОСТНЫЕ СХЕМЫ ДЛЯ ЭВОЛЮЦИОННЫХ ЗАДАЧ С ДВУМЯ ПРОСТРАНСТВЕННЫМИ ПЕРЕМЕННЫМИ

Формулировка задачи для уравнения теплопроводности. Явная и неявная шеститочечные схемы. Аппроксимация, устойчивость. Организация вычислений, экономичность вычислений. Экономичные разностные схемы. Метод переменных направлений (МПН). Устойчивость, погрешность аппроксимации схемы МПН. Метод покоординатного расщепления.

Рассмотрим следующую задачу для уравнения теплопроводности:

$$
\begin{aligned}
&U_t - \mu_1 U_{xx} - \mu_2 U_{yy} = f(t,x,y),\ (\mu_1,\mu_2>0),\\
&\qquad\qquad 0<t\leqslant T,\quad 0<x,y<1;\\
&U(0,x,y)=\varphi(x,y),\qquad 0\leqslant x,y\leqslant 1,\\
&\left.\begin{aligned}U(t,0,y)&=\psi_1(t,y),\\ U(t,1,y)&=\psi_2(t,y)\end{aligned}\right\}\quad 0<t\leqslant T,\quad 0\leqslant y\leqslant 1,\\
&\left.\begin{aligned}U(t,x,0)&=\chi_1(t,x),\\ U(t,x,1)&=\chi_2(t,x)\end{aligned}\right\}\quad 0<t\leqslant T,\quad 0<x<1.
\end{aligned}
\tag{11.1}
$$

Как следует из (11.1), решение должно быть найдено внутри параллелепипеда области независимых переменных (t,x,y) с единичным квадратом в основании (при $t=0$) (см. рис. 11.1). На основании известны начальные данные, на боковых гранях должны выполняться заданные краевые условия; μ_1, μ_2 — коэффициенты теплопроводности по направлениям x и y соответственно.

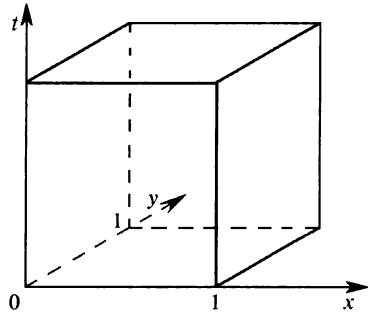


Рис. 11.1.

Замечание. Если говорить конкретно о переносе тепла, то, большей частью, на практике возникают задачи с изотропной теплопроводностью, когда $\mu_1=\mu_2$. Мы сочли возможным рассмотреть более общий, неизотропный случай ($\mu_1\neq\mu_2$), так как это не привносит каких-либо осложнений в последующие обсуждения. ▲

Введем в расчетной области сетку узлов

$$\omega_h = \{(t_n, x_k, y_m):\ t_n = n\tau,\ n = 0, 1, \ldots, N = [T/\tau];\ x_k = kh_x,$$

$$k = 0, 1, \ldots, K = [1/h_x];\ y_m = mh_y,\quad m = 0, 1, \ldots, M = [1/h_y]\}.$$

Как и прежде, мы ограничиваемся рассмотрением сеточных областей простейшей структуры: прямоугольных, с постоянными сеточными параметрами (шагами сетки) τ, h_x, h_y.

В пространстве трех независимых переменных сетка — трехиндексная. Узел $\binom{n}{k,m}$ лежит на пересечении n-го слоя по времени с k-м слоем по $x^{(k)}$ и с m-м слоем по y; τ — шаг по времени, h_x — шаг по x (расстояние между соседними слоями по x), h_y — шаг по y.

Явная и неявная шеститочечные схемы. Ограничим, как обычно, задачу вычислением приближенных значений решения (11.1) в узлах сетки. Искомое значение в узле $\binom{n}{k,m}$ будем обозначать через $U^n_{k,m}$, а всю совокупность искомых величин будем трактовать как сеточную функцию

$$U^{(h)} = \{U^n_{k,m};\ n = 0, 1, \ldots, N;\ k = 0, 1, \ldots, K;\ m = 0, 1, \ldots, M\}.$$

Замечание 1. Временной индекс n принято записывать в верхней позиции для компоненты $U^n_{k,m}$, чтобы подчеркнуть отличие независимой переменной t от пространственных переменных x и y. ▲

Применяя обычный подход, построим для $U^{(h)}$ разностную схему:

$$\frac{U^{n+1}_{k,m} - U^n_{k,m}}{\tau} - \mu_1 \frac{U^n_{k+1,m} - 2U^n_{k,m} + U^n_{k-1,m}}{h_x^2} - \mu_2 \frac{U^n_{k,m+1} - 2U^n_{k,m} + U^n_{k,m-1}}{h_y^2} = f^n_{k,m},$$

$$n = 0, 1, \ldots, N-1,$$

$$k = 1, 2, \ldots, K-1,\quad m = 1, 2, \ldots, M-1,$$

$$U^0_{k,m} = \varphi(x_k, y_m),\qquad k = 0, 1, \ldots, K;\ m = 0, 1, \ldots, M, \tag{11.1'}$$

$$\begin{aligned} &U^n_{0,m} = \psi_1(t_n, y_m),\quad U^n_{K,m} = \psi_2(t_n, y_m), \\ &U^n_{k,0} = \chi_1(t_n, x_k),\quad U^n_{k,M} = \chi_2(t_n, x_k), \end{aligned} \qquad \begin{aligned} &n = 1, 2, \ldots, N, \\ &m = 0, 1, \ldots, M, \\ &k = 1, 2, \ldots, K-1. \end{aligned}$$

Замечание 2. Нетрудно проверить, что это замкнутая система соотношений. Количество их равно числу неизвестных $(N+1)(M+1)(K+1)$ (см. Замечание 1 в Лекции 7). ▲

Схеме (11.1′) отвечает шаблон, изображенный на рис. 11.2. Это *явная* шеститочечная схема для двумерного уравнения теплопроводности. Что касается рассмотренных нами в предыдущих лекциях методов анализа схем на аппроксимацию и устойчивость, то при переходе к многомерным задачам принципиальных сложностей не возникает.

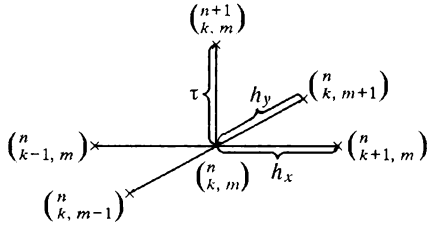


Рис. 11.2.

Исследование схемы (11.1′) на аппроксимацию. Подставляя в (11.1′) проекцию решения задачи (11.1) на сеточную область и вычитая правые части, получаем (для верхних соотношений (11.1′)) невязку, которая представляет собой погрешность аппроксимации в узле $\binom{n}{k,m}$:

$$\psi^n_{k,m} = \frac{\tau}{2}U_{tt} - \frac{\mu_1 h_x^2}{12}U^{(4)}_{x^4} - \frac{\mu_2 h_y^2}{12}U^{(4)}_{y^4}\Big|^n_{k,m} + O(\tau^2, h_x^4, h_y^4).$$

В граничных узлах погрешность аппроксимации нулевая. Таким образом, если существуют и ограничены соответствующие производные от U, то это схема первого порядка аппроксимации по времени и второго порядка — по пространственным переменным.

Исследование на устойчивость методом гармоник. Обобщение метода гармоник на рассматриваемый случай состоит в том, что при упрощающих предположениях, которые обсуждались в Лекции 10, решение разностных уравнений ищется в виде двумерных гармоник:

$$U^n_{k,m} = C\lambda^n e^{i\omega x_k} e^{i\nu y_m} = C\lambda^n e^{i(\omega h_x)k} e^{i(\nu h_y)m} = C\lambda^n e^{i(\alpha k + \beta m)}, \quad (11.2)$$

где $\alpha = \omega h_x$, $\beta = \nu h_y$ — произвольные параметры (ввиду произвольности номеров гармоник ω, ν).

Подставляя (11.2) в верхние соотношения (11.1′) (с $f_k^n = 0$), после сокращения на $C\lambda^n e^{i(\alpha k+\beta m)}$ получим

$$\frac{\lambda-1}{\tau} - \mu_1 \frac{e^{i\alpha} - 2 + e^{-i\alpha}}{h_x^2} - \mu_2 \frac{e^{i\beta} - 2 + e^{-i\beta}}{h_y^2} = 0,$$

$$\lambda = 1 + 2\frac{\mu_1\tau}{h_x^2}(\cos\alpha - 1) + 2\frac{\mu_2\tau}{h_y^2}(\cos\beta - 1) =$$

$$= 1 - 4\frac{\mu_1\tau}{h_x^2}\sin^2\frac{\alpha}{2} - 4\frac{\mu_2\tau}{h_y^2}\sin^2\frac{\beta}{2}.$$

Требуя, чтобы при любых α, β выполнялось неравенство $|\lambda| \leqslant 1$, приходим к необходимому условию устойчивости

$$\tau\left(\frac{\mu_1}{h_x^2} + \frac{\mu_2}{h_y^2}\right) \leqslant \frac{1}{2} \tag{11.3}$$

явной шеститочечной схемы для двумерного уравнения теплопроводности, являющемуся естественным обобщением условия устойчивости, полученного нами для одномерного случая (которое, в свою очередь, формально получается из (11.3), если положить μ_2 или μ_1 равным нулю, что соответствует «блокировке» распространения тепла в одном из направлений).

Замечание. Аналогично тому, как это было сделано в Лекции 10, можно показать достаточность условия (11.3) для устойчивости схемы (11.1′). ▲

Принципиальные трудности при распространении разностных методов на случай двух и более пространственных переменных связаны с проблемой обеспечения *экономичности вычислений*.

Посмотрим на схему (11.1′) с позиций трудоемкости вычислений. Очевидно, число арифметических операций при использовании явной

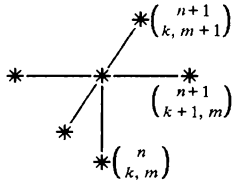


Рис. 11.3.

схемы пропорционально числу узлов в сеточной области. Для наглядности получаемых в этой части оценок будем считать, что $h_x \sim h_y \sim 1/K$. Тогда из условия устойчивости при $\mu_1 \sim \mu_2 \sim 1$ следует, что

$\tau \sim 1/N \sim 1/K^2$. Следовательно, число узлов и соответственно число операций будет $\Omega \sim K^2 N \sim K^4$. При $K \sim 10^2$ $\Omega \sim 10^8$ — это уже много для серийных расчетов и заставляет задуматься о поиске более эффективных алгоритмов.

Обратимся теперь к *неявной* шеститочечной схеме для решения задачи (11.1). Шаблон схемы изображен на рис. 11.3.

Разностные уравнения для решения во внутренних узлах сетки имеют вид

$$\frac{U^{n+1}_{k,m} - U^n_{k,m}}{\tau} - \mu_1 \frac{U^{n+1}_{k+1,m} - 2U^{n+1}_{k,m} + U^{n+1}_{k-1,m}}{h_x^2} - \\ - \mu_2 \frac{U^{n+1}_{k,m+1} - 2U^{n+1}_{k,m} + U^{n+1}_{k,m-1}}{h_y^2} = f^n_{k,m},$$
$$n = 0, 1, \ldots, N-1, \quad k = 1, 2, \ldots, K-1, \quad m = 1, 2, \ldots, M-1. \tag{11.1''}$$

Остальные уравнения схемы выглядят так, как в (11.1′). Легко показать, что эта схема приводит к тем же оценкам (с точностью до знаков) для погрешности аппроксимации, что и схема (11.1′). Метод гармоник приводит к выводу об абсолютной устойчивости данной схемы.

Переходим к обсуждению вопросов, связанных с решением уравнений (11.1″). Как уже отмечалось в Лекции 9, основной вычислительный цикл при использовании разностных схем для эволюционных задач состоит в переходе от n-го слоя по времени к $(n+1)$-му, т. е. в уравнениях (11.1″) неизвестными считаются данные на $(n+1)$-м слое. В каждом уравнении, следовательно, содержится пять неизвестных.

Если каким-то образом перенумеровать неизвестные $U^{n+1}_{k,m}$, чтобы превратить их в элементы одноиндексного массива (с тем, чтобы получить запись линейной системы уравнений (11.1″) в традиционном для линейной алгебры виде), то можно убедиться, что вычисление решения на $(n+1)$-м слое сводится к решению системы с пятидиагональной матрицей. К сожалению, заполненные ненулевыми элементами диагонали не образуют «сплошную ленту» (не равны нулю элементы на главной диагонали, на двух примыкающих к ней сверху и снизу и еще на двух, удаленных (вверх и вниз) примерно на K или M шагов в зависимости от способа нумерации искомых величин $\{U^{n+1}_{k,m}\}$ как элементов одноиндексного массива). Поэтому, несмотря на сильную разреженность матрицы (большая часть ее элементов — нулевые), построить простой экономичный метод решения такой системы (наподобие метода прогонки для систем с трехдиагональной матрицей) не так просто.

Если привлечь для решения общий метод Гаусса, то придем к крайне неблагоприятной оценке эффективности: число требуемых операций пропорционально кубу числа неизвестных, т. е. порядка $(K^2)^3 = K^6$. И это только на одном шаге по времени!

Учитывая ленточную структуру матрицы, можно на два порядка снизить трудоемкость вычислений на одном шаге по времени. В самом деле, количество элементарный операций, необходимых для решения методом исключения системы с ленточной матрицей $\sim s^2 N_A$, где s — полуширина ленты, N_A — порядок системы (см. [9]). В силу абсолютной устойчивости схемы здесь нет ограничений на выбор шага по времени, и если соображения точности позволяют взять $\tau \sim h$ (т. е. $N \sim K$), то (учитывая, что в нашем случае $s \sim K$, $N_A \sim K^2$) общий объем требуемых для решения задачи операций будет $\sim K^5$ (больше, чем для явной схемы).

Сейчас на примере нашей модельной задачи мы рассмотрим некоторые методы конструирования разностных схем для двумерного уравнения теплопроводности, приводящие к высокоэффективным (с точки зрения затрат машинного времени) численным алгоритмам, и, что важно, допускающим обобщение на случай более сложных задач (например, для уравнений с непостоянными коэффициентами, в областях сложной формы).

Замечание. Для гиперболических двумерных задач (например, для двумерного уравнения переноса: $u_t + au_x + bu_y = f$) эта проблема не столь актуальна, так как для явных схем обычное требование к шагу по времени, вытекающее из условия устойчивости: $\tau \sim h_x \sim h_y$, т. е. $N \sim K$ и общий объем арифметических операций для решения задачи порядка $NK^2 \sim K^3$. Это приемлемая оценка. ▲

Метод переменных направлений (МПН). Этот метод иногда называют *методом продольно-поперечных прогонок* (в иностранной литературе — *методом Писмена–Рэкфорда.*)

Введем в рассмотрение промежуточный слой узлов, соответствующий координате $t_{n+1/2} = t + \tau/2$. Компоненты сеточной функции на промежуточном слое будем обозначать через $U^{n+1/2}_{k,m}$. Переход от n-го слоя к $(n+1)$-му предлагается совершать в два этапа. Сначала из уравнений

$$\frac{U^{n+1/2}_{k,m} - U^n_{k,m}}{\tau/2} - \mu_1 \frac{U^{n+1/2}_{k+1,m} - 2U^{n+1/2}_{k,m} + U^{n+1/2}_{k-1,m}}{h_x^2} - \\ - \mu_2 \frac{U^n_{k,m+1} - 2U^n_{k,m} + U^n_{k,m-1}}{h_y^2} = f^{n+1/2}_{k,m}, \quad (11.4)$$

$$k = 1, 2, \ldots, K-1; \quad m = 1, 2, \ldots, M-1,$$

дополненных граничными условиями при $k = 0, K$, находятся величины на промежуточном слое. Затем из уравнений

$$\frac{U_{k,m}^{n+1} - U_{k,m}^{n+1/2}}{\tau/2} - \mu_1 \frac{U_{k+1,m}^{n+1/2} - 2U_{k,m}^{n+1/2} + U_{k-1,m}^{n+1/2}}{h_x^2} - \\ - \mu_2 \frac{U_{k,m+1}^{n+1} - 2U_{k,m}^{n+1} + U_{k,m-1}^{n+1}}{h_y^2} = f_{k,m}^{n+1/2}, \quad (11.5)$$

$$k = 1, 2, \ldots, K-1; \quad m = 1, 2, \ldots, M-1$$

с граничными условиями при $m = 0, M$ отыскивается решение на $(n+1)$-м слое.

Система (11.4) расслаивается на трехточечные (по k) системы алгебраических уравнений при каждом фиксированном m. Последние решаются по формулам прогонки (прогонки по направлению x). Число требуемых операций при этом $\sim MK|_{M=K} \sim K^2$. Аналогично *система (11.5) расслаивается на трехточечные (по m) системы при фиксированных значениях k*, которые решаются прогонками по направлению y. Суммарное число операций на первом и втором этапах $\sim K^2$.

Далее мы покажем, что схема (11.4)–(11.5) абсолютно устойчива, т. е. ограничений на шаг τ нет и, если есть возможность (по соображениям точности) вести счет с шагом $\tau \sim h_x \sim h_y$, то $N \sim K$ и общий объем операций для вычисления решения во всех узлах сетки $NK \sim K^3$, что, как уже отмечалось, приемлемо.

Замечание 1. Можно считать, что схеме МПН соответствует шаблон, изображенный на рис. 11.4. ▲

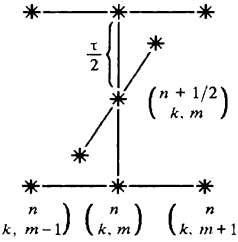


Рис. 11.4.

Замечание 2. Вычислительной неустойчивости при счете по формулам прогонки, к которому сводятся вычисления на каждом этапе, не возникает. В самом деле, как было показано в Лекции 2, для системы уравнений

$$a_k x_{k-1} + b_k x_k + c_k x_{k+1} = g_i$$

диагональное преобладание $|b_i| > |a_i| + |c_i|$ гарантирует устойчивость прогонки. В нашем случае, например, на первом этапе при фиксированном m имеем

$$a_k U_{k-1,m}^{n+1/2} + b_k U_{k,m}^{n+1/2} + c_k U_{k+1,m}^{n+1/2} = g_k$$

с $a_k = -\mu_1\tau/(2h_x^2) = c_k$, $b_k = 1 + \mu_1\tau/h_x^2$. Очевидно, $|a_k| + |c_k| = \mu\tau/h_x^2 < b_k$. ▲

Об устойчивости МПН. Метод гармоник применительно к данной схеме состоит в следующем. При обычных упрощающих предположениях ($f^n_{k,m} = 0$, $k \in (-\infty, \infty)$, $m \in (-\infty, \infty)$) ищем решение разностных уравнений в виде (11.2):

$$U^n_{k,m} = C\lambda^n e^{i(k\alpha + m\beta)},$$

причем $\lambda = \lambda_1\lambda_2$, где λ_1 — коэффициент «усиления» амплитуды гармоники при переходе на промежуточный слой, так что

$$U^{n+1/2}_{k,m} = \lambda_1 U^n_{k,m} = C\lambda_1\lambda^n e^{i(k\alpha + m\beta)}, \tag{11.6}$$

а λ_2 — соответствующий коэффициент при переходе с промежуточного слоя на $(n+1)$-й:

$$U^{n+1}_{k,m} = \lambda_2 U^{n+1/2}_{k,m} = C\lambda_2(\lambda_1\lambda^n e^{(k\alpha + m\beta)}) = C\lambda^{n+1} e^{i(k\alpha + m\beta)}. \tag{11.7}$$

Подставляя (11.6) в разностные уравнения (11.4), находим

$$\lambda_1 = \frac{1 - 2\rho_y \sin^2(\beta/2)}{1 + 2\rho_x \sin^2(\alpha/2)}, \quad \rho_x = \frac{\mu_1\tau}{h_x^2}, \quad \rho_y = \frac{\mu_2\tau}{h_y^2}.$$

Из (11.7) и (11.5)

$$\lambda_2 = \frac{1 - 2\rho_x \sin^2(\alpha/2)}{1 + 2\rho_y \sin^2(\beta/2)},$$

т. е.

$$\lambda = \lambda_1\lambda_2 = \frac{1 - 2\rho_y \sin^2(\beta/2)}{1 + 2\rho_x \sin^2(\beta/2)} \cdot \frac{1 - 2\rho_x \sin^2(\alpha/2)}{1 + 2\rho_y \sin^2(\alpha/2)}.$$

И $|\lambda| \leqslant 1$ при любых τ, h_x, h_y.

О погрешности аппроксимации схемы МПН. Ради компактности выкладок введем обозначения для операторов разностного дифференцирования сеточной функции по пространственным направлениям:

$$\begin{aligned} \Lambda_1 U^n_{k,m} &= \mu_1 \frac{U^n_{k+1,m} - 2U^n_{k,m} + U^n_{k-1,m}}{h_x^2}, \\ \Lambda_2 U^n_{k,m} &= \mu_2 \frac{U^n_{k,m+1} - 2U^n_{k,m} + U^n_{k,m-1}}{h_y^2}. \end{aligned} \tag{11.8}$$

Разностные уравнения из (11.4) и (11.5) можно записать тогда в виде

$$\begin{aligned}&\frac{U_{k,m}^{n+1/2}-U_{k,m}^{n}}{\tau/2}-\Lambda_1 U_{k,m}^{n+1/2}-\Lambda_2 U_{k,m}^{n}=f_{k,m}^{n+1/2},\\&\frac{U_{k,m}^{n+1}-U_{k,m}^{n+1/2}}{\tau/2}-\Lambda_1 U_{k,m}^{n+1/2}-\Lambda_2 U_{k,m}^{n+1}=f_{k,m}^{n+1/2}.\end{aligned}\tag{11.9}$$

Исключим из соотношений (11.9) промежуточный слой, т. е. перейдем к эквивалентной схеме МПН на исходной сеточной области (без промежуточных слоев). Вычитая второе соотношение (11.9) из первого, получим

$$U_{k,m}^{n+1/2}=\frac{U_{k,m}^{n}+U_{k,m}^{n+1}}{2}-\frac{\tau}{4}\Lambda_2(U_{k,m}^{n+1}-U_{k,m}^{n}).\tag{11.10}$$

Складывая соотношения (11.9), имеем

$$\frac{U_{k,m}^{n+1}-U_{k,m}^{n}}{\tau}-\Lambda_1 U_{k,m}^{n+1/2}-\Lambda_2\frac{U_{k,m}^{n}+U_{k,m}^{n+1}}{2}=f_{k,m}^{n+1/2}.$$

Подставляя сюда $U_{k,m}^{n+1/2}$ из (11.10), получаем разностные уравнения МПН для внутренних узлов исходной сеточной области ω^h:

$$\begin{aligned}\frac{U_{k,m}^{n+1}-U_{k,m}^{n}}{\tau}-\Lambda_1\frac{U_{k,m}^{n}+U_{k,m}^{n+1}}{2}-\Lambda_2\frac{U_{k,m}^{(k)}n+U_{k,m}^{n+1}}{2}+\\+\frac{\tau}{4}\Lambda_1\Lambda_2(U_{k,m}^{n+1}-U_{k,m}^{n})=f_{k,m}^{n+1/2}.\end{aligned}\tag{11.11}$$

Используя дальше обычную технику, приходим к выводу, что при достаточной гладкости решения U соотношения (11.11) аппроксимируют исходное дифференциальное уравнение из (11.11) со вторым порядком точности как по x, y, так и по t.

Итак, мы построили эффективный алгоритм численного решения исходной задачи (11.11). МПН широко используется для решения двумерных задач подобного типа.

Замечание. МПН обобщается на случай уравнений с непостоянными коэффициентами и для областей, составленными из прямоугольных подобластей. Для областей с косоугольными и, тем более, с криволинейными границами возникают определенные трудности. ▲

Коснемся коротко других подходов к построению эффективных схем для многомерных задач.

Метод покоординатного расщепления. Этот метод иногда называют *локально-одномерным методом*. Предполагается следующий двухэтапный способ перехода от n-го слоя к $(n+1)$-му слою по времени:

$$\frac{\widehat{U}_{k,m} - U^n_{k,m}}{\tau} - \Lambda_1 \widehat{U}_{k,m} = \frac{1}{2} f^n_{k,m}, \tag{11.12}$$

$$k = 1, 2, \ldots, K-1, \quad m = 1, 2, \ldots, M-1$$

с граничными условиями при $k = 0, K$;

$$\frac{U^{n+1}_{k,m} - \widehat{U}_{k,m}}{\tau} - \Lambda_2 U^{n+1}_{k,m} = \frac{1}{2} f^{n+1}_{k,m} \tag{11.13}$$

$$k = 1, 2, \ldots, K-1, \quad m = 1, 2, \ldots, M-1$$

с граничными условиями при $m = 0, M$. (В записи использованы введенные выше обозначения Λ_1 и Λ_2.)

В отличие от МПН здесь нет промежуточного слоя по t. На первом этапе вычисляются промежуточные значения $\widehat{U}_{k,m}$, которые можно трактовать (апеллируя к физическому смыслу уравнения теплопроводности (11.1)) как изменение U на элементарном временном интервале $\tau = t_{n+1} - t_n$ за счет теплопроводности, «включенной» лишь по направлению x. Затем на втором этапе «включается» теплопроводность по y (выключается по x); значения $\widehat{U}^{(k)}_{k,m}$ используются в качестве новых начальных данных при $t = t_n$, а $U^{n+1}_{k,m}$ рассматриваются как окончательное приближение к решению исходной задачи на слое $t = t_{n+1}$, сформированное за счет действия теплопроводности по тому и другому пространственным направлениям.

На каждом этапе расчет носит экономичный характер. (Так же, как и в схеме МПН, здесь система (11.12) расслаивается на трехточечные системы линейных уравнений при каждом фиксированном m, а (11.13) сводится к аналогичным системам при каждом k.)

Не останавливаясь на подробном анализе, отметим, что на этапе (11.12) при каждом m соответствующая трехточечная система эквивалентна неявной четырехточечной схеме для одномерного уравнения теплопроводности, которая, как мы видели в свое время, абсолютно устойчива (аналогично на этапе (11.13)). В итоге, локально-одномерная схема также будет абсолютно устойчивой. В этой связи выбор шага τ не связан жестким условием типа (11.3), и, если соображения точности позволяют проводить вычисления с шагом $\tau \sim h_x \sim h_y$, то данный алгоритм, как и схема МПН, обеспечивает экономичный расчет не только на одном интервале по времени, но и в целом.

З а м е ч а н и е. Эта схема любопытна тем, что на каждом отдельном этапе разностные соотношения не аппроксимируют дифференциальное уравнение. Однако имеет место так называемая *суммарная аппроксимация*: разностная схема, являющаяся следствием (11.12)–(11.13) (после исключения из последних промежуточных значений $\widehat{U}^{(k)}_{k,m}$), аппроксимирует исходную задачу. ▲

В завершение данной лекции хотелось бы отметить, что здесь весьма схематично представлены два подхода к построению экономичных методов решения многомерных задач. Я имел целью обозначить принципиальную возможность реализации подобных подходов. При необходимости решать конкретную задачу такого типа, естественно, надо разобраться с возможными подходами подробней.

Более подробную информацию по рассмотренным здесь вопросам можно найти в [2, с. 548–561], [4, с. 284–297], [5, с. 96–106], [9, с. 389–400, 435–439], [11, с. 250–259], [12, с. 372–378], [13], [18, с. 133–140], [19, с. 479–493].

ВОПРОСЫ И УПРАЖНЕНИЯ

1. Доказать достаточность условия (11.3) для устойчивости схемы (11.1′).

2. Исследовать на устойчивость неявную схему с шеститочечным шаблоном (рис. 11.3) для задачи (11.1).

3. Построить и исследовать на устойчивость явную разностную схему, аппроксимирующую задачу Коши для уравнения теплопроводности в трехмерном пространстве x, y, z. Каковы главные члены погрешности аппроксимации предложенной схемы?

4. Выписать матрицу системы (11.1″) для $K = 3$, $M = 4$.

5. Изобразить шаблон схемы (11.11), полученной после исключения промежуточного этапа в методе переменных направлений.

ЛЕКЦИЯ 12

ЧИСЛЕННОЕ РЕШЕНИЕ ЭЛЛИПТИЧЕСКИХ УРАВНЕНИЙ

Первая краевая задача для уравнения Пуассона. Простейшая разностная схема, аппроксимация, устойчивость. Проблема решения разностных уравнений. Об итерационных методах решения, связь методов последовательных приближений с решением некоторой эволюционной задачи. О методах решения, основанных на принципе установления. Доказательство устойчивости простейшей разностной схемы.

Мы будем рассматривать здесь первую краевую задачу (задачу Дирихле) для уравнения Пуассона:

$$\begin{cases} U_{xx} + U_{yy} = f(x, y), & (x, y) \in \omega, \\ U|_\Gamma = \varphi(x, y) & (x, y) \in \Gamma; \end{cases} \tag{12.1}$$

ω — множество внутренних точек расчетной области, Γ — граничные точки (рис. 12.1). Объединение множеств ω и Γ ($\overline{\omega} = \omega \bigcup \Gamma$) представляет собой расчетную область $\overline{\omega}$ задачи (12.1).

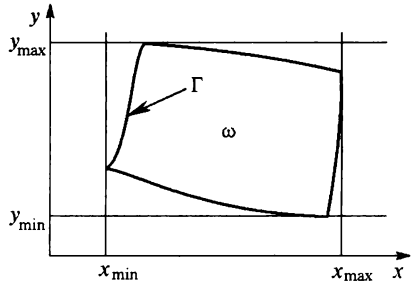


Рис. 12.1.

Замечание. Сопоставляя (12.1) с абстрактной формулировкой исходной задачи $LU = F$, видим, что в данном случае

$$LU = \begin{cases} \dfrac{\partial^2 U}{\partial x^2} + \dfrac{\partial^2 U}{\partial y^2}, & (x, y) \in \omega, \\ U|_\Gamma, & (x, y) \in \Gamma, \end{cases}$$

$$F = \begin{cases} f(x, y), & (x, y) \in \omega, \\ \varphi(x, y), & (x, y) \in \Gamma. \end{cases}$$

▲

Простейшая разностная схема. Построим сеточную область, например, таким образом: промежуток $[x_{\min}, x_{\max}]$ разделим на K слоев: $x_k = x_{\min} + kh_x$, $h_x = (x_{\max} - x_{\min})/K$, $k = 0, 1, \ldots, K$ шириной h_x. Соответственно $[y_{\min}, y_{\max}]$ — на M слоев: $y_m = y_{\min} + mh_y$, $h_y = (y_{\max} - y_{\min})/M$, $m = 0, 1, \ldots, M$ шириной h_y (рис. 12.2).

Внутренними узлами (ω_h) сеточной области будем считать точки с координатами (x_k, y_m), которые принадлежат области $\overline{\omega}$ вместе со своими четырьмя ближайшими соседними точками (на рис. 12.2 отмечены кружочками). Граничные узлы (γ_h) — это совокупность точек $(x_k, m) \in \omega$, для которых по крайней мере одна из четырех соседних точек лежит за пределами $\overline{\omega}$ (крестики на рис. 12.2). Объединение множеств ω_h и γ_h представляет собой сеточную область $\overline{\omega}_h$ для данной задачи ($\overline{\omega}_h = \omega_h \bigcup \gamma_h$). Обозначим через $U_{k,m}$ искомое значение решения в узле (x_k, y_m) сеточной области. Совокупность

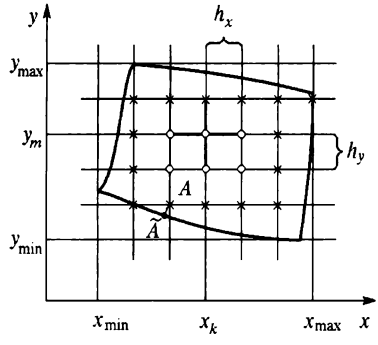


Рис. 12.2.

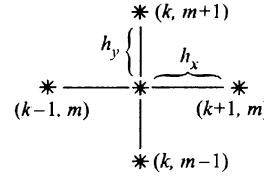


Рис. 12.3.

искомых значений во всех узлах сетки представляет собой сеточную функцию $U^{(h)} = \{U_{k,m}$ для (k, m) таких, что $(x_k, y_m) \in \overline{\omega}_h\}$. Используя обычный подход (замену производных разностными отношениями), аппроксимируем уравнение Пуассона во внутренних узлах сетки разностным уравнением

$$\frac{U_{k+1,m} - 2U_{k,m} + U_{k-1,m}}{h_x^2} + \frac{U_{k,m+1} - 2U_{k,m} + U_{k,m-1}}{h_y^2} = f_{k,m}, \quad (x_k, y_m) \in \omega_h. \quad (12.2)$$

Дополним эту систему соотношений условиями в граничных узлах:

$$U_{k,m} = \widetilde{\varphi}_{k,m}, \quad (x_k, y_m) \in \gamma_h. \tag{12.3}$$

Здесь $\widetilde{\varphi}_{k,m} = \varphi(\widetilde{x}, \widetilde{y})$, где $(\widetilde{x}, \widetilde{y})$ — ближайшая к узлу $(k, m) \in \gamma_h$ точка границы Γ в исходной задаче (12.1). (На рис. 12.2 ближайшей к граничному узлу A является точка $\widetilde{A} \in \Gamma$.)

Соотношения (12.2) и (12.3) образуют замкнутую систему разностных уравнений для компонент сеточной функции $U^{(h)}$, т.е. представляют собой пример возможной разностной схемы для задачи (12.1). (Шаблон этой схемы изображен на рис. 12.3.)

Аппроксимация. Используя обычную технику исследования на аппроксимацию, получаем, что во внутренних узлах сетки (при наличии у решения (12.1) ограниченных производных $U^{(\mathrm{IV})}_{x^4}$ и $U^{(\mathrm{IV})}_{y^4}$) главные члены погрешности аппроксимации имеют вид

$$\psi_{k,m} = U^{(\mathrm{IV})}_{x^4} \frac{h_x^2}{12} + U^{(\mathrm{IV})}_{y^4} \frac{h_y^2}{12}\bigg|_{(k,m)} + O(h_y^4, h_x^4), \quad (x_k, y_m) \in \omega_h.$$

Уместно отметить, что в данном случае определяющей является погрешность аппроксимации граничных условий (раньше, как правило, граничные условия аппроксимировались точно). В рассматриваемой задаче многие из граничных узлов сетки не совпадают с точками исходной границы Γ, а граничные условия (12.1) в этом случае аппроксимируются в (12.3) посредством «сноса» точных условий. Ясно, что при этом привносится погрешность

$$\psi_{k,m} \sim (h_x U_x + h_y U_y)\big|_{(k,m)}, \quad (x_k, y_m) \in \gamma_h.$$

Таким образом, схема (12.2)–(12.3) аппроксимирует исходную задачу с первым порядком по x и по y.

Замечание 1. Применяя для аппроксимации граничных условий не «снос», а более точный способ (например, интерполяцию), можно построить схему второго порядка аппроксимации. ▲

Замечание 2. Если исходная расчетная область представляет собой прямоугольник, то граничные узлы построенной таким образом сетки, очевидно, принадлежат исходной границе Γ, граничные условия (12.3) будут удовлетворяться точно и схема (12.2)–(12.3) будет схемой второго порядка аппроксимации. ▲

Замечание 3. Для области расчета из рис. 12.1 можно было бы совокупности внутренних и граничных узлов определить по-другому, как это было сделано, например, в Лекции 9 (см. рис. 9.2). Тогда

разностные граничные условия удовлетворялись бы точно, но расстояния от внутренних узлов сетки, примыкающих к границе, до ближайших граничных узлов не были бы постоянными. При этом возникает проблема аппроксимации уравнения Пуассона на неравномерной сетке. ▲

Об устойчивости. Напомним сначала, что метод гармоник приспособлен для анализа устойчивости сугубо эволюционных разностных схем. Этот метод существенно использует «слоистый» (по времени) характер решения эволюционных задач и, в конечном счете, сводится к анализу эволюции начального возмущения на множестве возможных (при определенных допущениях) решений разностных уравнений.

Эллиптические уравнения описывают стационарное (установившееся) состояние — здесь нет времени, нет эволюции. Соответственно исследование устойчивости разностных схем, аппроксимирующих задачи для эллиптических уравнений, сводится к непосредственной проверке выполнения неравенства

$$\|U^{(h)}\| \leqslant \text{const} \cdot \|F^{(h)}\|,$$

являющегося определением устойчивости (для линейных задач), если константа не зависит от сеточных параметров (см. Лекцию 10).

Как это можно сделать применительно к схеме (12.2)–(12.3), показывается в дополнении к данной лекции.

Решение разностных уравнений. Переходя к обсуждению вопросов, связанных с вычислением решения задачи (12.2)–(12.3), мы ограничимся рассмотрением прямоугольной расчетной области: $0 \leqslant x \leqslant a$, $0 \leqslant y \leqslant b$. (Как было отмечено выше, схема (12.2)–(12.3) аппроксимирует исходную задачу на прямоугольной области со вторым порядком по x и по y.)

Нетрудно подметить, что в этом случае разностная схема представляет собой систему линейных алгебраических уравнений, подобную которой мы уже встречали в Лекции 11, когда рассматривали неявную схему с шеститочечным шаблоном для двумерного уравнения теплопроводности. И снова на первый план выдвигается вопрос экономичности вычислений.

Метод Гаусса требует выполнения числа операций $\Omega \sim (KM)^3\big|_{K=M} = K^6$. Хотя это и существенно меньше, нежели для решения по упомянутой неявной схеме параболического уравнения (там такой объем работы нужно выполнять для перехода только на один, очередной слой по времени), все равно, это недопустимо много. Мы убедимся далее, что возможны гораздо более рациональные способы расчета.

З а м е ч а н и е. Эффективными методами решения поставленной задачи при определенных условиях являются метод, использующий быстрое преобразование Фурье, и метод редукции ([12, с. 337, 418], [13]). Требуемое количество арифметических операция для того и другого метода характеризуется величиной порядка $K^2 \ln K$. ▲

О методах последовательных приближений. Перепишем разностные уравнения (12.2) в разрешенном относительно $U_{k,m}$ виде и рассмотрим итерационный процесс:

$$\overset{s+1}{U}_{k,m} = \frac{1}{2(1/h_x^2 + 1/h_y^2)} \left[-f_{k,m} + \frac{1}{h_x^2} (\overset{s}{U}_{k+1,m} + \overset{s}{U}_{k-1,m}) + \right.$$
$$\left. + \frac{1}{h_y^2} (\overset{s}{U}_{k,m+1} + \overset{s}{U}_{k,m-1}) \right], \quad (x_k, y_m) \in \omega_h. \qquad (12.4)$$

Здесь s — номер приближения; $\overset{s+1}{U}_{k,m}$ для $(x_k, y_m) \in \gamma_h$ считаются определенными из граничных условий (12.3).

З а м е ч а н и е. Это не что иное, как метод Якоби (см. Лекцию 3), так как мы из (k, m)-го уравнения исключаем $U_{k,m}$, т. е. диагональный элемент. ▲

Не останавливаясь на обосновании, отметим, что итерации (12.4) сходятся. Сходятся довольно медленно; тем не менее, этот метод иногда привлекается для практических вычислений. (Несмотря на медленную сходимость, он намного рациональней метода Гаусса в смысле экономичности. Полное число операций, связанное с реализацией этого метода, $\Omega|_{K=M} \sim K^4 \ln K$ ([4], [12]).)

Для нас сейчас метод (12.4) интересен тем, что оттолкнувшись от него, мы сумеем обозреть способы построения более эффективных методов численного решения систем типа (12.2)–(12.3).

Введем обозначение

$$\tau = \frac{1}{2(1/h_x^2 + 1/h_y^2)} = \frac{1}{2} \frac{h_x^2 h_y^2}{(h_x^2 + h_y^2)}. \qquad (12.5)$$

Тогда (12.4) можно записать в виде

$$\frac{1}{\tau} \overset{s+1}{U}_{k,m} = \left[\frac{\overset{s}{U}_{k+1,m} + \overset{s}{U}_{k-1,m}}{h_x^2} + \frac{\overset{s}{U}_{k,m+1} + \overset{s}{U}_{k,m-1}}{h_y^2} \right] - f_{k,m}.$$

Вычитая из обеих частей последнего равенства $\frac{1}{\tau}U^s_{k,m}$, приведем его к виду

$$\frac{\overset{s+1}{U}{}_{k,m} - \overset{s}{U}_{k,m}}{\tau} = \frac{\overset{s}{U}_{k+1,m} - 2\overset{s}{U}_{k,m} + \overset{s}{U}_{k-1,m}}{h_x^2} +$$

$$+ \frac{\overset{s}{U}_{k,m+1} - 2\overset{s}{U}_{k,m} + \overset{s}{U}_{k,m-1}}{h_y^2} - f_{k,m} = L_h\overset{s}{U}{}^{(h)} - f^{(h)}\big|_{k,m} \qquad (12.6)$$

где L_h — оператор (матрица), определяющий вид уравнений (12.2) с учетом того, что компоненты $U_{k,m}$ в граничных точках определены из (12.3). Далее перепишем соотношения (12.6) в виде

$$\overset{s+1}{U}{}^{(h)} = (E + \tau L_h)\,\overset{s}{U}{}^{(h)} - \tau f^{(h)}. \qquad (12.7)$$

Вспоминая материал Лекции 3 (итерационные методы решения линейных систем уравнений), мы видим, что (12.7) не что иное, как запись однопараметрического итерационного процесса для системы $L_hU^{(h)} = f^{(h)}$, если отвлечься от конкретного значения параметра τ, задаваемого формулой (12.5).

Стало быть, в рамках записи (12.7) мы имеем дело уже с семейством итерационных схем для задачи (12.2)–(12.3). Можно искать диапазон значений параметра τ, при которых процесс (12.7) сходится (значение (12.5) входит в этот диапазон). Можно искать оптимальное τ, при котором сходимость является наиболее быстрой. Дело сводится к оценке минимального и максимального собственных чисел оператора (матрицы) L_h.

Здесь уместно затронуть вопрос о допустимой погрешности при завершении итерационного процесса, используемого для решения рассматриваемой задачи. В силу того, что разностная схема (12.2)–(12.3) аппроксимирует задачу (12.1) со вторым порядком точности по x и по y, точное решение этой системы (12.2)–(12.3) в силу известной теоремы об аппроксимации, устойчивости и сходимости отличается от решения исходной задачи в узлах сетки на величины $O(h_x^2, h_y^2)$. Следовательно, в рамках выбранного итерационного процесса достаточно ограничиться приближением, которое отличается от точного решения системы (12.2)–(12.3) также на величину $\varepsilon \sim O(h_x^2, h_y^2)$. Это и есть характеристика (по порядку величины) для допустимого уровня погрешности при вычислении последовательных приближений. С учетом этого обстоятельства для метода (12.7) получается уже упомянутая оценка полного количества операций: $\Omega|_{K=M} \sim K^4 \ln K$ (см. [4], [12]).

Опираясь на упомянутый анализ собственных чисел оператора L_h, удается строить более эффективные итерационные схемы для реше-

ния систем типа (12.2)–(12.3). (Например, чебышёвский итерационный процесс с $\Omega|_{K=M} \sim K^3 \ln K$ — см. Приложение 3.)

О методах, основанных на принципе установления. Обратимся еще раз к соотношениям (12.6). Согласно (12.5), τ — малый параметр ($\tau \sim O(h^2)$, где $h = \max\{h_x, h_y\}$). Следовательно, (12.6) буквально совпадает с разностными уравнениями явной шеститочечной схемы для двумерного уравнения теплопроводности

$$U_t = U_{xx} + U_{yy} - f(x, y), \tag{12.8}$$

если номер приближения s трактовать как номер слоя по времени, а τ — как шаг сетки по времени.

В Лекции 11 мы получили условие устойчивости для этой схемы, которое применительно к (12.6) запишется в виде

$$\tau\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) \leqslant \frac{1}{2}.$$

Как видно из (12.5), параметр τ, соответствующий простейшей схеме итераций (12.4), удовлетворяет этому условию! Таким образом, метод итераций (12.4) решения системы (12.2)–(12.3) эквивалентен одной из разностных схем решения смешанной (с краевыми и начальными условиями) задачи для нестационарного уравнения (12.8), а решение последней при не зависящих от времени краевых условиях с течением времени устанавливается (перестает зависеть от времени) и в пределе удовлетворяет стационарному уравнению (12.1).

Приведенные рассуждения обобщаются *принципом установления*, согласно которому решение стационарной задачи можно искать в качестве предельного решения эволюционной задачи с не зависящими от времени граничными условиями (согласованная нестационарная задача).

Замечание 1. В соответствии с принципом установления любая схема для согласованной нестационарной задачи может трактоваться как итерационный процесс для решения стационарной системы. ▲

Замечание 2. Из предыдущего замечания вытекает, что для исследования устойчивости итераций можно применять приемы исследования устойчивости разностных схем для эволюционных задач. ▲

Возвращаясь к исходной задаче и опираясь на принцип установления, можно теперь выбрать в качестве способа численного ее решения метод переменных направлений или метод покоординатного расщепления для уравнения (12.8) с краевыми условиями (12.3) и произволь-

ными начальными данными. При этом вычисления следует проводить до тех пор, пока, например, не выполнится условие

$$\frac{1}{\tau}\left|\| \overset{s+1}{U}{}^{(h)}\| - \| \overset{s}{U}{}^{(h)}\|\right| \leqslant \varepsilon,$$

что соответствует, очевидно, требованию установления стационарного решения: $\left|\frac{\partial U}{\partial t}\right| \leqslant \varepsilon$. Анализ показывает, что при этом может быть достигнута столь же высокая эффективность, как и для упомянутого выше чебышёвского итерационного процесса (полное количество требуемых операций $\Omega|_{K=M} \sim K^3 \ln K$).

ДОПОЛНЕНИЕ К ЛЕКЦИИ 12

Доказательство устойчивости схемы (12.2)–(12.3). Предварительно докажем две леммы.

Лемма 1. *Пусть сеточная функция* $V^{(h)} = \{V_{k,m}, (k,m) \in \overline{\omega}_h = \omega_h \cup \gamma_h\}$ *на* ω_h *удовлетворяет условиям*

$$\frac{V_{k+1,m} - 2V_{k,m} + V_{k-1,m}}{h_x^2} + \frac{V_{k,m+1} - 2V_{k,m} + V_{k,m-1}}{h_y^2} \geqslant 0$$

(или в компактной записи $L_h V^{(h)} \geqslant 0$*, если под* L_h *понимать разностный оператор, определенный видом левой части уравнений (12.2)). Тогда наибольшее в* $\overline{\omega}_h$ *значение* $V^{(h)}$ *достигается на границе* γ_h*.*

Доказательство. Допустим противное, т. е. пусть

$$V_{\overline{k},\overline{m}} = \max_{\overline{\omega}_h} V_{k,m} \quad \text{и} \quad (\overline{k},\overline{m}) \in \omega_h.$$

Тогда при $k = \overline{k}$, $m = \overline{m}$ приходим к противоречию с условием леммы:

$$\frac{V_{\overline{k}+1,\overline{m}} - 2V_{\overline{k},\overline{m}} + V_{\overline{k}-1,\overline{m}}}{h_x^2} + \frac{V_{\overline{k},\overline{m}+1} - 2V_{\overline{k},\overline{m}} + V_{\overline{k},\overline{m}-1}}{h_y^2} =$$

$$= \frac{(V_{\overline{k}+1,\overline{m}} - V_{\overline{k},\overline{m}}) + (V_{\overline{k}-1,\overline{m}} - V_{\overline{k},\overline{m}})}{h_x^2} +$$

$$+ \frac{(V_{\overline{k},\overline{m}+1} - V_{\overline{k},\overline{m}}) + (V_{\overline{k},\overline{m}-1} - V_{\overline{k},\overline{m}-1})}{h_y^2} \leqslant 0.$$

Аналогично доказывается следующая лемма.

Лемма 2. *Если* $V^{(h)}$ *на* ω_h *удовлетворяет условию* $L_h V^{(h)} \leqslant 0$, *то* $\min\limits_{\overline{\omega}_h} V_{k,m}$ *достигается на* γ_h.

Из доказанных лемм вытекает теорема.

Теорема. *Если* $L_h V^{(h)} = 0$ *на* ω_h, *то максимальное и минимальное значения* $V_{k,m}$ *достигаются при* $(k, m) \in \gamma_h$.

Из этой теоремы следует вывод о существовании и единственности решения разностной задачи (12.2)–(12.3). В самом деле, если в (12.2)–(12.3) положить $f_{k,m} \equiv 0$ и $\widetilde{\varphi}_{k,m} \equiv 0$ для всех (k, m), то соответствующая однородная система уравнений $L_h U^{(h)} = 0$ согласно теореме будет иметь лишь тривиальное решение.

Перейдем теперь непосредственно к доказательству устойчивости схемы (12.2)–(12.3). Введем в рассмотрение вспомогательную функцию $P^{(h)} = \{P_{k,m}\}$ такую, что в узлах сетки

$$P_{k,m} = \frac{1}{4} M_f [R^2 - (x_k^2 + y_m^2)] + M_\varphi,$$

где $R^2 = \max\limits_{\overline{\omega}} (x^2 + y^2)$, $M_f = \max\limits_{\omega_h} |f_{k,m}|$, $M_\varphi = \max\limits_{\gamma_h} |\widetilde{\varphi}_{k,m}|$.

Очевидно, что результат применения оператора L_h к константе следующий: $L_h\,(\mathrm{const}) = 0$. Далее,

$$L_h\{x_k^2\}\big|_{(k,m)} = \frac{x_{k+1}^2 - 2x_k^2 + x_{k-1}^2}{h_x^2} = \frac{(x_{k+1}^2 - x_k^2) - (x_k^2 - x_{k-1}^2)}{h_x^2} = \\ = \frac{(x_{k+1} + x_k) - (x_k + x_{k-1})}{h_x} = 2.$$

Точно так же $L_h\{y^2\}\big|_{(k,m)} = 2$. Поэтому $L_h P^{(h)} = \frac{1}{4} M_f(-2 - 2) = \\ = -M_f < 0$ во всех внутренних узлах сетки, т. е.

$$L_h(U^{(h)} - P^{(h)})\big|_{(k,m)} = f_{k,m} + M_f \geqslant 0.$$

В силу леммы 1 максимальное значение разности $U^{(h)} - P^{(h)}$ достигается на границе γ_h. Но в точках γ_h имеем

$$U_{k,m} - P_{k,m} = (\widetilde{\varphi}_{k,m} - M_\varphi) - \frac{1}{4} M_f [R^2 - (x_k^2 + y_m^2)] \leqslant 0,$$

т. е. $U_{k,m} \leqslant P_{k,m}$ на всей сетке $\overline{\omega}_h$.

Далее, $L_h(P^{(h)} + U^{(h)})\big|_{(k,m)} = -M_f + f_{k,m} \leqslant 0$. Отсюда с привлечением леммы 2 получаем, что $-P_{k,m} \leqslant U_{k,m}$ на $\overline{\omega}_h$.

Итак, в любом узле сеточной области $\overline{\omega}_h$ имеет место неравенство

$$-P_{k,m} \leqslant U_{k,m} \leqslant P_{k,m},$$

или $|U_{k,m}| \leqslant P_{k,m}$, т. е. $\max\limits_{k,m} |U_{k,m}| \leqslant \max\limits_{k,m} P_{k,m} \leqslant \frac{1}{4} R^2 M_f + M_\varphi$. Привлекая равномерную метрику в пространствах сеточных функций, последнее неравенство можно переписать в виде

$$\|U^{(h)}\| \leqslant \frac{1}{4} R^2 \|f^{(h)}\| + \|\varphi^{(h)}\|.$$

Это означает устойчивость разностной схемы (12.2)–(12.3).

С более строгим и обстоятельным изложением теории разностных схем для эллиптических уравнений можно познакомиться в [2, с. 526–546, 561–570], [4, с. 298–361, 378–425], [5, с. 106–119], [9, с. 401–423], [10, с. 199–289], [11, с. 211–231], [12, с. 291–338, 378–425], [13].

ПРИЛОЖЕНИЕ 1

ПОЛИНОМЫ ЧЕБЫШЁВА 1-ГО РОДА

На отрезке $t \in [-1, 1]$ полиномы, с которыми мы начинаем знакомиться, обычно определяются следующей компактной формулой:

$$T_m(t) = \cos(m \arccos t). \qquad (П1.1)$$

Здесь m — степень полинома.

В том, что формулой (П1.1) задается именно полином m-й степени, мы убедимся чуть ниже, а пока получим полезную для дальнейшего рекуррентную связь между полиномами Чебышёва различных степеней. Привлекая известное тригонометрическое соотношение

$$\cos\alpha + \cos\beta = 2\cos\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right),$$

имеем:

$$\cos[(m+1)\arccos t] + \cos[(m-1)\arccos t] = \\ = 2\cos[m \arccos t] \cdot \cos[\arccos t],$$

что, учитывая введенное для полиномов Чебышёва определение (П1.1), может быть записано в виде

$$T_{m+1}(t) + T_{m-1}(t) = 2tT_m(t)$$

или

$$T_{m+1}(t) = 2tT_m(t) - T_{m-1}(t). \qquad (П1.2)$$

Теперь не составляет труда усмотреть, что формула (П1.1) действительно определяет алгебраический полином степени m. В самом деле, из (П1.1) получаем

$$\begin{aligned} &\text{при} \quad m = 0 \qquad T_0(t) = 1, \\ &\text{при} \quad m = 1 \qquad T_1(t) = t, \end{aligned}$$

Из (П1.2) следует

$$\begin{aligned} T_2(t) &= 2t^2 - 1, \\ T_3(t) &= 4t^3 - 3t, \\ T_4(t) &= 8t^4 - 8t^2 + 1 \end{aligned} \qquad (П1.3)$$

и так далее.

З а м е ч а н и е. Запись полиномов Чебышёва в виде (П1.3) позволяет распространить их область определения за пределы отрезка $[-1, 1]$. ▲

Отметим некоторые характерные особенности полиномов Чебышёва (далее подразумевается, что $m \geqslant 1$).

Как видно из (П1.3), коэффициент при старшей степени t для полинома m-й степени равен 2^{m-1}.

Из представления (П1.1) видно, что корни полинома m-й степени определяются уравнением

$$m \arccos t = \pi/2 + k\pi, \quad k = 0, 1, 2, \ldots \tag{П1.4}$$

Различные (неповторяющиеся) решения этого уравнения:

$$t_k = \cos\left(\frac{(2k+1)\pi}{2m}\right), \quad k = 0, 1, \ldots, m-1. \tag{П1.5}$$

Приведенный в (П1.5) диапазон допустимых значений k определяет совокупность m различных корней полинома m-й степени. Все корни, очевидно, принадлежат отрезку $[-1, 1]$.

Можно отметить следующую геометрическую интерпретацию корней полинома Чебышёва.

Разделим верхнюю полуокружность единичного радиуса (рис. П.1) на m равных частей. Тогда координаты середин элементарных дуг (квадратики на рис. П.1) совпадают с корнями (П1.5).

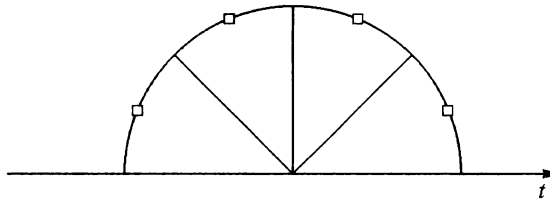


Рис. П1

В свою очередь, координаты точек деления

$$t_k^* = \cos\left(\frac{k\pi}{m}\right), \quad k = 0, 1, \ldots, m \tag{П1.6}$$

представляют собой точки экстремумов полинома Чебышёва степени m.

Соответствующие (экстремальные) значения полинома

$$T_m(t_k^*) = (-1)^k, \quad k = 0, 1, \ldots, m.$$

Рассмотрим теперь полиномы Чебышёва, нормированные таким образом, что коэффициент при старшей степени t равен единице. Для этого, очевидно, надо поделить $T_m(t)$ на 2^{m-1}:

$$\widehat{T}_m(t) = \frac{T_m(t)}{2^{m-1}}. \tag{П1.7}$$

Полиномы, записанные в обычной (алгебраической) форме (П1.3), после указанной нормировки перейдут в

$$\begin{aligned}
&\widehat{T}_1(t) = t,\\
&\widehat{T}_2(t) = t^2 - \frac{1}{2},\\
&\widehat{T}_3(t) = t^3 - \frac{3}{4}t,\\
&\widehat{T}_4(t) = t^4 - t^2 + \frac{1}{8},\\
&\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{П1.8}$$

Замечательным свойством полиномов $\widehat{T}_m(t)$ является то, что они «наименее уклоняются от нуля» на рассматриваемом отрезке среди всех полиномов степени m с единичным коэффициентом при старшей степени t. Последнее утверждение можно выразить в виде неравенства

$$\max_{-1\leqslant t\leqslant 1} |\widehat{T}_m(t)| \leqslant \max_{-1\leqslant t\leqslant 1} |P_m(t)|, \tag{П1.9}$$

в котором под $P_m(t)$ понимается *произвольный* полином степени m с единичным коэффициентом при старшей степени t.

Ясно, что экстремальные значения полиномов $\widehat{T}_m(t)$ достигаются в точках (П1.6) и равны соответственно $\widehat{T}_m(t_k^*) = (-1)^k/2^{m-1}$. Левая часть неравенства (П1.9), стало быть, равна $1/2^{m-1}$.

Чтобы установить справедливость (П1.9), допустим противное. Пусть существует полином $\widetilde{P}_m^{(k)}(t)$ (с единичным коэффициентом при t^m) такой, что

$$\max_{[-1,1]} |\widetilde{P}_m(t)| < \max_{[-1,1]} |\widehat{T}_m(t)| = \frac{1}{2^{m-1}}. \tag{П1.10}$$

Рассмотрим разность двух полиномов

$$R_{m-1}(t) = \widehat{T}_m(t) - \widetilde{P}_m(t), \tag{П1.11}$$

представляющую собой, очевидно, полином $(m-1)$-й степени, так как старшие степени $\widehat{T}_m(t)$ и $\widetilde{P}_m(t)$ при вычитании сокращаются.

В точках (П1.6)

$$R_{m-1}(t_k^*) = \widehat{T}_m(t_k^*) - \widetilde{P}_m(t_k^*) = (-1)^k 2^{1-m} - \widetilde{P}(t_k^*),$$

и в силу предположения (П1.10) знак $R_{m-1}(t_k^*)$ совпадает со знаком $\widehat{T}_m(t_k^*)$, то есть меняется m раз на отрезке $[-1, 1]$ (так как $k = 0, 1, \ldots, m$). Мы пришли к противоречию: полином $(m-1)$-й степени $R_{m-1}(t)$ имеет m корней.

Нетрудно написать полином степени m, наименее уклоняющийся от нуля на произвольном отрезке $x \in [a, b]$. Замечая, что преобразование

$$t = \frac{2x - (b+a)}{b-a} \tag{П1.12}$$

переводит $x \in [a, b]$ в $t \in [-1, 1]$, получаем требуемый полином

$$Q_m(x) = \widehat{T}_m\Big(\frac{2x - (b+a)}{b-a}\Big). \tag{П1.13}$$

Старший коэффициент $Q_m(x)$ равен, очевидно, $(2/(b-a))^m$, то есть, строго говоря, (П1.13) наименее уклоняется от нуля на отрезке $[a, b]$ среди полиномов с указанным старшим коэффициентом. Экстремальные значения при этом достигаются в точках

$$x_k^* = \frac{a+b}{2} + \frac{b-a}{2} t_k^* = \frac{a+b}{2} + \frac{b-a}{2} \cos\frac{k\pi}{m}, \quad k = 0, 1, \ldots, m$$

и равны

$$Q_m(x_k^*) = (-1)^k / 2^{m-1}.$$

Корни (П1.13) расположены на отрезке $[a, b]$ в точках

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} t_k = \frac{a+b}{2} + \frac{b-a}{2} \cos\frac{(2k+1)\pi}{2m}, \quad k = 0, 1, \ldots, m-1. \tag{П1.14}$$

Чтобы получить полином m-й степени с *единичным старшим коэффициентом*, наименее уклоняющийся от нуля на отрезке $[a, b]$, достаточно разделить $Q_m(x)$ на его старший коэффициент:

$$\widehat{Q}_m(x) = \Big(\frac{b-a}{2}\Big)^m Q_m(x). \tag{П1.15}$$

Для $\widehat{Q}_m(x)$ экстремальные значения (максимальные уклонения от нуля), как следует из (П1.15) и (П1.6), равны:

$$\widehat{Q}_m(x_k^*) = \Big(\frac{b-a}{2}\Big)^m \frac{(-1)^k}{2^{m-1}} = (-1)^k \frac{(b-a)^m}{2^{2m-1}}. \tag{П1.16}$$

В дальнейшем нам окажется полезным полином $S_m(x)$, наименее уклоняющийся от нуля на отрезке $[a, b]$ среди полиномов m-й степени со свободным членом, равным единице, то есть среди полиномов, проходящих через единицу при $x = 0$.

$S_m(x)$ очевидным образом связан с полиномами, введенными в рассмотрение выше:

$$S_m(x) = \frac{\widehat{Q}_m(x)}{\widehat{Q}_m(0)} = \frac{Q_m(x)}{Q_m(0)} = \widehat{T}_m\left(\frac{2x - (b+a)}{b-a}\right)\Big/\widehat{T}_m\left(-\frac{(b+a)}{b-a}\right) =$$
$$= T_m\left(\frac{2x - (b+a)}{b-a}\right)\Big/T_m\left(-\frac{(b+a)}{b-a}\right). \qquad (\text{П1.17})$$

В последующих разделах будут приведены примеры использования рассмотренных здесь полиномов применительно к отдельным методам вычислительной математики.

МИНИМИЗАЦИЯ ОШИБКИ ПРИ ПОЛИНОМИАЛЬНОЙ ИНТЕРПОЛЯЦИИ ФУНКЦИЙ

Напомним формулировку задачи об интерполяции. В точках $\{x_i,\ i = 0, 1, \ldots, n\}$ отрезка $[a, b]$ заданы значения функции $f(x)$: $\{f_i,\ i = 0, 1, \ldots, n\}$. По этим табличным данным требуется построить приближенную формулу для $f(x)$.

Если приближение ищется в виде степенного полинома, то, как известно, решением является так называемый интерполяционный полином, который может быть записан в различных формах (Лекция 5). Напомним, например, лагранжеву форму записи интерполяционного полинома:

$$P_n(x) = f_0 \frac{(x - x_1)(x - x_2)\ldots(x - x_n)}{(x_0 - x_1)(x_0 - x_2)\ldots(x_0 - x_n)} + \ldots + \\ + f_k \frac{(x - x_0)\ldots(x - x_{k-1})(x - x_{k+1})\ldots(x - x_n)}{(x_k - x_0)\ldots(x_k - x_{k-1})(x_k - x_{k+1})\ldots(x_k - x_n)} + \ldots + \\ + f_n \frac{(x - x_0)(x - x_1)\ldots(x - x_{n-1})}{(x_n - x_0)(x_n - x_1)\ldots(x_n - x_{n-1})}. \qquad (\text{П2.1})$$

Табличные точки $\{x_i\}$ в контексте рассматриваемой задачи принято называть узлами интерполяции. В общем случае они распределены по отрезку $[a, b]$ произвольным образом. (Предполагаем, что они пронумерованы в порядке возрастания, то есть $x_{i+1} > x_i$). Расстояния между узлами $\{h_i = x_{i+1} - x_i,\ i = 0, 1, \ldots, n - 1\}$ называют шагами интерполирования.

Ошибка интерполяции $R_n(x) = f(x) - P_n(x)$ (или остаточный член интерполяционной формулы $f(x) \approx P_n(x)$) для функции $(n + 1)$ раз непрерывно дифференцируемой на отрезке $[a, b]$, как известно, может быть представлена в виде

$$R_n(x) = \frac{f^{(n+1)}(x')}{(n + 1)!}\omega_{n+1}(x), \qquad (\text{П2.2})$$

где x' — некоторая точка на отрезке $[a, b]$, а $\omega_{n+1}(x)$ — полином $(n + 1)$-й степени вида

$$\omega_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n). \qquad (\text{П2.3})$$

Очевидно, что величина ошибки зависит от расположения узлов. В Лекции 5 рассматривались оценки погрешности интерполяции полиномами различной степени в предположении, что узлы интерполяции

равноотстоят друг от друга: $x_{i+1} - x_i = h = \text{const}$. Здесь мы остановимся на случае, когда возможность «расставлять» узлы на $[a, b]$ находится в нашем распоряжении (например, табличные данные вычисляются на компьютере), и посмотрим, нельзя ли минимизировать ошибку интерполяции за счет этой возможности.

Оказывается, можно. И решение этой задачи становится очевидным, если, во-первых, обратить внимание на то, что $\omega_{n+1}(x)$ представляет собой полином с единичным коэффициентом при старшей степени x, а во-вторых, вспомнить, что в предыдущем Приложении мы выписывали полиномы подобного вида, наименее уклоняющиеся от нуля на рассматриваемом отрезке.

Таковыми среди полиномов степени m являлись полиномы $\widehat{Q}_m(x)$ (см. (П1.15)). В нашем случае $m = n + 1$. Корни этого полинома, согласно (П1.14):

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} t_k = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{(2k+1)\pi}{2(n+1)}, \quad k = 0, 1, \ldots, n. \tag{П2.4}$$

Корни $\omega_{n+1}(x)$, как видно из (П2.3), суть $\{x_i,\ i = 0, 1, \ldots, n\}$.

Если x_i совпадают с (П2.4), то $\omega_{n+1}(x) \equiv \widehat{Q}_{n+1}(x)$.

Говорят, что в таком случае осуществляется *степенная интерполяция по чебышёвским узлам*.

Посмотрим, какое конкретно влияние оказывает интерполяция по чебышёвским узлам на погрешность $R_n(x)$. Согласно (П1.16), максимальное уклонение $\widehat{Q}_{n+1}(x)$ от нуля равно

$$\Delta = \frac{(b-a)^{n+1}}{2^{2n+1}}. \tag{П2.5}$$

Таким образом, можно выписать оценку

$$|R_n(x)| = \left| \frac{f^{(n+1)}(x')}{(n+1)!} \widehat{Q}_{n+1}(x) \right| \leqslant \frac{M_{n+1}}{(n+1)! \cdot 2^n} \left[\frac{b-a}{2} \right]^{n+1} \tag{П2.6}$$

(здесь $M_{n+1} = \max\limits_{[a,b]} |M^{(n+1)}(x)|$ — максимум модуля $(n+1)$-й производной функции $f(x)$ на рассматриваемом отрезке).

В таб. П2.1 для различных значений n приводится величина

$$\eta = \max_{[0,1]} |\widehat{Q}_{n+1}(x)| / \max_{[0,1]} |\widehat{\omega}_{n+1}(x)|,$$

которая характеризует эффективное уменьшение максимально возможной ошибки интерполяции при переходе от равнораспределенных узлов к чебышёвским узлам на отрезке $[a, b]$, если под $\widehat{\omega}_{n+1}(x)$ понимать полином (П2.3), определенный по равноотстоящим узлам.

Таблица П2.1

n	1	2	3	5	10	20	30
η	0.5000	0.6495	0.6328	0.4514	0.1145	0.0041	0.0001

Замечание. Значение η, как можно усмотреть из (П2.2) и (П2.6), не зависит от отрезка $[a, b]$. Поэтому конкретные вычисления проводились на отрезке $[-1, 1]$. ▲

Что касается больших n, то напомним, что интерполяция по равноотстоящим узлам при $n \gtrsim 10$ практически не используется. Во-первых, сходимость интерполяции при увеличении n может отсутствовать. Как показывает известный пример функции Рунге $f(x) = 1/(1+25x^2)$, даже для бесконечно дифференцируемой функции ошибка интерполяции с ростом n может становиться сколь угодно большой. Во-вторых, даже малые погрешности, содержащиеся в табличных данных, приводят к большим (неустранимым) ошибкам интерполяции при использовании полиномов высокой степени.

Все эти неприятности не имеют места, если интерполяция осуществляется по чебышёвским узлам, то есть помимо того, что чебышёвская интерполяция минимизирует теоретическую ошибку, она обладает еще другими положительными свойствами сравнительно с интерполяцией по равномерным узлам. Более подробно эти вопросы еще будут затронуты в Приложении 4, а пока рассмотрим некоторые примеры.

Пример 1. На рис. П.2 показаны результаты интерполирования функции Рунге $f(x) = 1/(1+25x^2)$ полиномами 8-й и 10-й степеней, построенными по равнораспределенным на отрезке $[-1, 1]$ узлам, а также полиномом 8-й степени — $P_8^{\text{Чеб}}(x)$ по чебышёвским узлам.

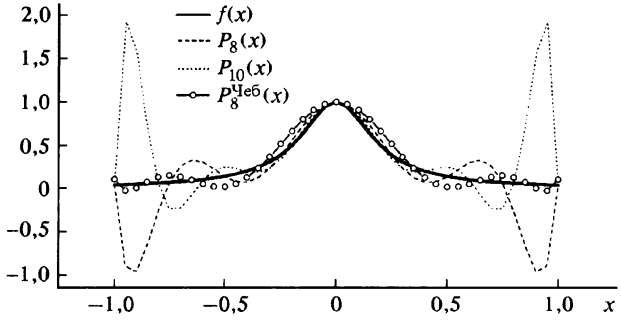


Рис. П2

Более подробные данные приведены в таблице П2.2. В первом столбце таблицы указана степень интерполяционных полиномов, в

последующих приводится максимальная (по модулю) ошибка интерполяции: во втором и третьем столбцах — для полинома по равноотстоящим узлам (во втором столбце приведен максимум ошибки на отрезках $0.7 < |x| \leqslant 1$, где интерполяционный процесс по равноотстоящим узлам для функции Рунге расходится с ростом n; в третьем столбце соответствующая ошибка на отрезке $|x| \leqslant 0.7$); в четвертом столбце показана ошибка интерполяции по чебышёвским узлам.

Таблица П2.2

N	Ошибка интерполяции		
4	0.438	0.375	0.400
5	0.108	0.433	0.550
6	0.608	0.229	0.262
7	0.172	0.274	0.392
8	1.01	0.246	0.167
10	1.88	0.302	0.109
16	12.64	0.245	0.031
20	40.00	0.116	0.014

Как видно, ошибка для этой «нехорошей» (с точки зрения интерполяции) функции при использовании чебышёвских узлов равномерно стремится к нулю с ростом числа узлов.

В следующем примере иллюстрируется влияние неустранимых погрешностей на результат интерполяции.

Пример 2. В таблице П2.3 приведены результаты приближения на отрезке $[-1, 1]$ интерполяционными полиномами различной степени «хорошей» функции $f(x) = \cos \pi x$.

Таблица П2.3

N	Ошибка интерполяции			
2	0.6090	0.6110	0.5503	0.5541
3	0.4844	0.4522	0.3640	0.3620
4	0.0886	0.0939	0.0555	0.0613
5	0.0624	0.0627	0.0320	0.0351
6	0.0071	0.0136	0.0027	0.0090
8	0.0003	0.0067	0.0001	0.0091
10	$1.31 \cdot 10^{-5}$	0.0115	$1.54 \cdot 10^{-6}$	0.0095
14	$6.53 \cdot 10^{-9}$	0.5137	$2.16 \cdot 10^{-10}$	0.0104
20	$2.97 \cdot 10^{-9}$	15.33	$2.55 \cdot 10^{-11}$	0.0111

Как и в предыдущей таблице, здесь для различной степени интерполяции (столбец 1) приводится ошибка интерполяции для полиномов, построенных по равноотстоящим узлам (столбцы 2, 3) и по

чебышёвским (столбцы 4, 5). При этом в столбцах 3, 5 приводятся данные, полученные для случая, когда коэффициенты соответствующих интерполяционных полиномов вычислялись по табличным данным функции $f(x)$, искаженным случайной ошибкой $\leqslant 0.01$.

Видно, что влияние неустранимых погрешностей при использовании чебышёвских узлов существенно меньше сравнительно с интерполяцией по равнораспределенным узлам, когда привлекаются полиномы высокой степени ($n \geqslant 10$).

МЕТОД ИТЕРАЦИЙ С ЧЕБЫШЁВСКИМ НАБОРОМ ПАРАМЕТРОВ ДЛЯ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ

В этом разделе будет рассмотрен еще один пример использования чебышёвских полиномов для построения эффективного численного метода. Речь идет, как следует из названия приложения, об итерационных методах решения линейных систем уравнений.

Итак, пусть задана система

$$A\mathbf{X} = \mathbf{f}. \tag{П3.1}$$

Здесь A — квадратная матрица порядка n с элементами a_{ij}, $\mathbf{X}$ — вектор неизвестных величин $\{x_i,\ i=1,2,\ldots,n\}$, $\mathbf{f}=\{f_i,\ i=1,2,\ldots,n\}$ — правые части уравнений системы.

Приступая к обсуждению метода итераций с чебышёвским набором параметров (или, как иногда говорят, с чебышёвским ускорением), мы оттолкнемся от рассмотренного в Лекции 3 однопараметрического итерационного процесса, который можно использовать в качестве одного из возможных методов вычисления приближенного решения системы (П3.1). Каноническая запись (с. 49) этого метода выглядит следующим образом:

$$\frac{\mathbf{X}^{(m+1)} - \mathbf{X}^m}{\tau} + A\mathbf{X}^{(m)} = \mathbf{f}, \tag{П3.2}$$

m — номер приближения (итерации), τ — некоторый численный параметр.

При заданном начальном (нулевом) приближении — $\mathbf{X}^{(0)}$ каждое следующее, как следует из (П3.2), может быть вычислено по формулам

$$\mathbf{X}^{(m+1)} = \mathbf{X}^{(m)} - \tau A\mathbf{X}^{(m)} + \tau\mathbf{f} = (E - \tau A)\mathbf{X}(m) + \tau\mathbf{f}, \tag{П3.3}$$

(E — единичная матрица).

Если матрица A — симметрична и положительно определена, то, как показывает анализ (см. Лекцию 3), приближения, вычисляемые по формулам (П3.3), сходятся к искомому решению системы (П3.1), когда $0 < \tau < 2/\Lambda$, (Λ — максимальное собственное значение матрицы A), т. е. в этом случае

$$\delta_m = \|\mathbf{X}^{(m)} - \mathbf{X}^*\|_{m\to\infty} \to 0 \tag{П3.4}$$

($\mathbf{X}^*$ — точное решение (П3.1)).

Если при этом используется норма

$$\|\mathbf{X}\| = \|\mathbf{X}\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}, \qquad (\text{П3.5})$$

то удается найти оптимальное значение параметра

$$\tau = \tau_{\text{опт}} = \frac{2}{\bar{\lambda} + \Lambda} \qquad (\text{П3.6})$$

($\bar{\lambda}$ — минимальное собственное значение матрицы A), при котором δ_m убывает с ростом m наиболее быстро:

$$\delta_m = \delta_m^{\text{опт}} = \left(\frac{\Lambda - \bar{\lambda}}{\Lambda + \bar{\lambda}}\right)^m \delta_0 = q^m \delta_0, \qquad (\text{П3.7})$$

где δ_0 — ошибка начального приближения, а q характеризует темп убывания погрешности на одном шаге итераций и определяется через границы спектра матрицы A:

$$q = \frac{\Lambda - \bar{\lambda}}{\Lambda + \bar{\lambda}} = \frac{1 - \xi}{1 + \xi}, \qquad \xi = \frac{\bar{\lambda}}{\Lambda}. \qquad (\text{П3.8})$$

Ниже мы увидим, что если параметр τ менять от итерации к итерации, то можно найти такую последовательность параметров τ_m, при которой ошибка δ_m убывает намного быстрее, чем при $\tau = \text{const}$, даже если $\tau = \tau_{\text{опт}}$.

Последующие выкладки будут опираться на упомянутые выше предположения о матрице системы (П3.1):

она *симметрична* ($A = A^{\mathrm{T}}$, A^{T} — транспонированная матрица) и *положительно определена* ($(A\mathbf{X}, \mathbf{X}) > 0$ для любого вектора $\mathbf{X}$).

Пожалуй, здесь уместно отвлечься от обсуждаемой темы и сделать некоторые пояснения: зачем нужны эти предположения? Не являются ли они слишком ограничительными?

На последний вопрос ответ — нет. Для тех читателей, которые знакомы с методами вычислений в пределах хотя бы вводного курса, понятно, что линейные системы алгебраических уравнений, для решения которых итерационные методы оказываются более эффективными, нежели прямые (типа гауссовского исключения), появляются (в качестве промежуточного этапа) при использовании численных методов решения уравнений математической физики.

Это очень широкий круг задач, как правило, реального физического содержания. В свете же поставленного вопроса важно то, что для возникающих при этом линейных систем предположения о симметричности и положительной определенности матриц очень часто оказываются выполненными.

Читателю, не удовлетворенному подобным неаргументированным ответом на последний вопрос (которым вынужден ограничиться автор), можно порекомендовать разобраться с изучением свойств матриц соответствующих систем, например в [9–13].

Теперь о том, *что дает* предположение о симметричности и положительной определенности матрицы A исходной системы (П3.1). По этому поводу, во-первых, напомним следующие положения из курса линейной алгебры:

а) собственные числа симметричной матрицы действительны;

б) у положительно определенной симметричной матрицы все собственные числа больше нуля;

в) существует ортонормированный базис из собственных векторов симметричной матрицы.

Во-вторых, отметим, что в последующих выкладках мы будем существенно опираться на оговоренные положения.

Возвращаемся к основной теме данного Приложения.

Запишем итерационный процесс с переменным параметром в канонической форме

$$\frac{\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}}{\tau_k} + A\mathbf{X}^{(k)} = \mathbf{f}, \quad k = 0, 1, \ldots \tag{П3.9}$$

Соответствующие формулы для перехода от текущего (k-го) приближения к следующему имеют вид:

$$\mathbf{X}^{(k+1)} = (E - \tau_k A)\mathbf{X}^{(k)} + \tau_k \mathbf{f}. \tag{П3.10}$$

Точное решение системы (П3.1) $\mathbf{X}^*$, очевидно, при любом τ_k удовлетворяет также системе

$$\mathbf{X} = (E - \tau_k A)\mathbf{X} + \tau_k \mathbf{f},$$

т. е. при подстановке $\mathbf{X}^*$ в эту систему получаем тождество

$$\mathbf{X}^* \equiv (E - \tau_k A)\mathbf{X}^* + \tau_k \mathbf{f}. \tag{П3.11}$$

Вычитая (П3.11) из (П3.10), получаем соотношение, которое связывает ошибку $(k+1)$-го приближения с ошибкой k-го:

$$\mathbf{X}^{(k+1)} - \mathbf{X}^* = (E - \tau_k A)(\mathbf{X}^{(k)} - \mathbf{X}^*).$$

Оценим погрешность на m-м шаге.

Вводя обозначение для вектора ошибки $\mathbf{r}^{(m)} = \mathbf{X}^{(m)} - \mathbf{X}^*$ и выражая ошибки последовательно через предыдущие, получаем

$$\begin{aligned} \mathbf{r}^{(m)} &= (E - \tau_{m-1}A)\mathbf{r}^{(m-1)} = \\ &= (E - \tau_{m-1}A)(E - \tau_{m-2}A)\mathbf{r}^{(m-2)} = \ldots = P_m \mathbf{r}^{(0)}, \end{aligned} \tag{П3.12}$$

где матрица $P_m = \prod_{k=0}^{m-1}(E - \tau_k A)$.

Заметим, что матрица P_m симметрична.

В самом деле, для матрицы $B = (E - \alpha A)(E - \beta A)$ с произвольными α, β имеем по правилам перемножения матриц $B = E - (\alpha + \beta)A + \alpha\beta A^2$. Очевидно, если A — симметрична, то симметрична и матрица A^2, а вместе с тем и матрица B.

После этого заключения легко приходим к выводу о симметричности матрицы P_m.

Из (П3.12) следует оценка погрешности m-го приближения:

$$\delta_m = \|\mathbf{r}^{(m)}\| \leqslant \|P_m\|\delta_0. \qquad \text{(П3.13)}$$

Если погрешность оценивается в евклидовой норме (П3.5), то в качестве нормы матриц естественно привлечь спектральную норму (см. Лекцию 2), которая для симметричной матрицы определяется ее спектральным радиусом:

$$\|A\| = \max_i |\lambda_i| = \Lambda.$$

$\{\lambda_i,\ i = 1, 2, \ldots, n\}$ — собственные числа матрицы.

Будем считать их упорядоченными по величине; при этом, в силу положительной определенности матрицы A, $\overline{\lambda} = \lambda_1 > 0$ $(\Lambda = \lambda_n)$.

С учетом сказанного, в формуле (П3.13) $\|P_m\| = \max\limits_i |\mu_i^{(m)}|$, если под $\mu_i^{(m)}$ понимать собственные значения матрицы P_m.

Последние связаны с собственными числами матрицы A следующим образом:

$$\mu_i^{(m)} = \prod_{k=0}^{m-1} (1 - \tau_k\lambda_i), \qquad \text{(П3.14)}$$

что легко устанавливается.

Действительно, пусть i-му собственному числу матрицы A соответствует собственный вектор $\mathbf{e}_i$. Тогда для матрицы $P_1 = (E - \tau_1 A)$ имеем:

$$P_1\mathbf{e}_1 = (E - \tau_1 A)\mathbf{e}_i = \mathbf{e}_i - \tau_1\lambda_i\mathbf{e}_i = (1 - \tau_1\lambda_i)\mathbf{e}_i,$$

т. е. $\mu_i^{(1)} = (1 - \tau_1\lambda_i)$ является собственным значением матрицы P_1. Далее также непосредственной проверкой убеждаемся, что $\mu_i^{(2)} = (1 - \tau_2\lambda_i)\times$ $\times(1 - \tau_1\lambda_i)$ является собственным числом матрицы $P_2 = (E - \tau_2 A)(E - \tau_1 A)$ и т. д.

Настало время вспомнить о полиномах Чебышёва.

Заметим, что формула (П3.14) задает значения полинома m-й степени от λ:

$$\mu^{(m)}(\lambda) = \prod_{k=0}^{m-1} (1 - \tau_k\lambda) \qquad \text{(П3.15)}$$

в точках $\lambda_i,\ i = 1, 2, \ldots, n$ отрезка $[\overline{\lambda},\ \Lambda]$.

Соответственно

$$\|P_m\| = \max_i |\mu_i^{(m)}| \leqslant \max_{[\bar{\lambda}, \Lambda]} |\mu^{(m)}(\lambda)|. \qquad (\text{П3.16})$$

Правая часть последнего неравенства, очевидно, представляет собой максимальное уклонение от нуля полинома $\mu^{(m)}(\lambda)$ на отрезке $[\bar{\lambda}, \Lambda]$, причем $\mu^{(m)}(0) = 1$ при любых значениях параметров τ_k (которые пока не определены).

В Приложении 1 было показано, что среди полиномов подобного вида наименее уклоняются от нуля нормированные должным образом полиномы Чебышёва $S_m(x)$ или применительно к обозначениям данного раздела $S_m(\lambda)$.

Корни полинома $\mu^{(m)}(\lambda)$, как следует из (П3.15) суть

$$\{\lambda_k = 1/\tau_k, \quad k = 0, 1, \ldots, m-1\}.$$

Совмещая λ_k с корнями $S_m(\lambda)$ (см.(П1.14)):

$$\frac{1}{\tau_k} = \frac{\bar{\lambda} + \Lambda}{2} + \frac{\Lambda - \bar{\lambda}}{2} \cos \frac{(2k+1)\pi}{2m}, \quad k = 0, 1, \ldots, m-1, \qquad (\text{П3.17})$$

мы тем самым превращаем $\mu^{(m)}(\lambda)$ в полином, наименее уклоняющийся от нуля на отрезке $[\bar{\lambda}, \Lambda]$ и, как следует из (П3.16) и (П3.13), минимизирующий погрешность приближения на m-м шаге последовательных приближений (П3.10), а формула (П3.17) определяет соответствующую последовательность параметров τ_k метода (П3.10) вплоть до m-го приближения.

Метод (П3.10) с параметрами (П3.17) называют *методом итераций с чебышёвским набором параметров* (или с *чебышёвским ускорением*). В зарубежной литературе используется название — метод Ричардсона.

Какой же выигрыш дает использование чебышёвского набора параметров сравнительно с $\tau = \text{const}$? Соответствующие оценки можно найти в [1, 9–13, 18]. Сошлемся для определенности на выкладки, приведенные в [12, с. 106–107]. Там показано, что максимальное уклонение от нуля интересующих нас сейчас полиномов равно величине q_m, вычисляемой по формулам:

$$q_m = \frac{2\rho_1^m}{1 + \rho_1^{2m}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\bar{\lambda}}{\Lambda}. \qquad (\text{П3.18})$$

Используя (П3.18), можно найти априорную оценку необходимого количества итераций для вычисления решения с требуемой точностью.

На m-м шаге ошибка найденного приближения будет в ε раз меньше начальной ошибки:

$$\|\mathbf{X}^{(m)} - \mathbf{X}^*\| \leqslant \varepsilon \|\mathbf{X}^{(0)} - \mathbf{X}^*\|,$$

если $q_m \leqslant \varepsilon$.

Отсюда в предположении малости $\xi = \underline{\lambda}/\Lambda$ (что, как правило, имеет место для систем, о происхождении которых шла речь выше) получаем оценку для необходимого числа приближений ([12, с.111–112]):

$$m \geqslant m_0 \approx \frac{\ln(2/\varepsilon)}{2\sqrt{\xi}}. \qquad (\text{П3.19})$$

Для метода итераций с постоянным (оптимальным) параметром количество приближений, уменьшающих начальную ошибку в ε раз (Лекция 3):

$$m \geqslant m_0 \approx \frac{\ln(1/\varepsilon)}{2\xi}. \qquad (\text{П3.20})$$

Выигрыш в числе итераций (П3.19) весьма значителен сравнительно с (П3.20). Например, в [11, с.112–113] приводится количество приближений, потребовавшееся при решении конкретной задачи с $\varepsilon = 10^{-4}$: для чебышёвского набора параметров $m_0 \approx 34$, для однопараметрического процесса $m_0 \approx 200$.

Как следует из вышеизложенного, использование данного метода для решения конкретных систем требует знания максимального и минимального собственных чисел матрицы A (по крайней мере их приближенных оценок). При этом априори известна оценка количества приближений, которые необходимо вычислить — (П3.19).

Замечание. Впрочем, иногда придерживаются иной тактики вычислений: задавшись значением m (например, $m \approx 1/\sqrt{\xi} = \sqrt{\Lambda/\underline{\lambda}}$), повторяют m-шаговый вычислительный цикл до тех пор, пока не будет достигнута требуемая точность, которая приближенно контролируется нормой разности двух приближений, полученных на последовательных m-шаговых циклах. ▲

Так или иначе число m достаточно велико для реальных задач (порядка десятков). При этом иногда обнаруживается довольно неприятный эффект. Если привлекать параметры $\{\tau_k\}$ в естественном порядке $k = 0, 1, \ldots, m-1$ (k — номера упорядоченных по возрастанию собственных чисел матрицы A), то расчет завершается вроде бы абсурдным образом: либо наступает «переполнение» (выход чисел за пределы, допустимые при машинном их представлении), либо, если до этого дело не доходит, в качестве результатов выдаются числа экзотических (больших) порядков, вызывая справедливые сомнения в том, что получено приближение к искомому решению.

В действительности ничего абсурдного здесь нет, такой итог обусловлен специфичным характером поведения ошибки приближения при переходе от одной итерации к следующей, а также спецификой машинных вычислений: тем, что числа в ЭВМ всегда представлены с погрешностью округления, и при производстве арифметических операций последняя в неблагоприятном случае может существенно влиять на результаты.

Подробный анализ того, что происходит в изучаемом случае (и как преодолеть возникающие неприятности), можно найти в [9, с.412; 11, с.115–118; 13, с.275–283; 18, с.154–159]. Поэтому мы ограничимся здесь обозначением данной проблемы и кратким (качественным) пояснением «механизма» неограниченного роста погрешности при естественном переборе параметров в обсуждаемом методе.

Представим вектор начальной ошибки $\mathbf{r}^{(0)} = \mathbf{X}^{(0)} - \mathbf{X}^*$ в виде разложения по элементам базиса из собственных векторов матрицы A:

$$\mathbf{r}^{(0)} = \sum_{i=1}^{n} c_i \mathbf{e}_i.$$

Тогда после выполнения первого шага итераций получим, согласно (П3.12):

$$\mathbf{r}^{(1)} = (E - \tau_0 A)\mathbf{r}^{(0)} = \sum_i c_i (E - \tau_0 A)\mathbf{e}_i = \sum_i c_i (1 - \tau_0 \lambda_i)\mathbf{e}_i, \quad (\text{П3.21})$$

где $\tau_0 = 1/\lambda_0^*$ — величина, обратно пропорциональная первому корню полинома $S_m(\lambda)$ (см. (П3.17)). Таким образом, i-я компонета погрешности начального приближения на первом шаге итераций умножается на $(1 - \tau_0 \lambda_i)$.

Так как корень λ_0^* находится вблизи левой границы спектра $\lambda_1 = \overline{\lambda}$), то множитель $(1 - \lambda_1/\lambda_0^*)$ достаточно мал (во всяком случае, меньше единицы). То же можно сказать и о ряде следующих множителей в правой части (П3.21): $(1 - \lambda_2/\lambda_0^*)$, $(1 - \lambda_3/\lambda_0^*), \ldots$

А вот для λ_i, примыкающих к правой границе спектра ($\lambda_i \sim \Lambda$):

$$\left(1 - \frac{\lambda_i}{\lambda_0^*}\right) \sim \left(1 - \frac{\lambda_i}{\overline{\lambda}}\right) \sim \frac{\Lambda}{\overline{\lambda}} > 1.$$

На самом деле, для реальных задач $\Lambda/\overline{\lambda} = 1/\xi \gg 1$. Таким образом, при переходе к первому приближению компоненты ошибки вдоль векторов $\mathbf{e}_i$ с $i \sim n$ значительно возрастают. То же самое на втором шаге и т.д. За полных m шагов метода итераций эти ошибки были бы подавлены (за счет малости множителей $(1 - \lambda_i/\lambda_j^*)$ при $\lambda_i, \lambda_j^* \sim \Lambda$.

Но до такого итога дело не всегда успевает дойти, так как на начальной стадии итераций сильно возрастающие ошибки вдоль упомянутых векторов уже приводят к переполнению либо к гипертрофированному росту неустранимых погрешностей при выполнении арифметических операций над числами чрезмерно большого порядка с ограниченной мантиссой.

Многие другие примеры использования полиномов Чебышёва для построения эффективных итерационных методов решения линейных систем уравнений можно найти в [2, с.277–289; 9, с.408–412; 10, с.176–187; 11, с.110–126; 12, с.109–115, 389–393; 13; 17, с.132–134; 18, с.154–159].

ТРИГОНОМЕТРИЧЕСКАЯ ИНТЕРПОЛЯЦИЯ

В этом приложении речь пойдет о приближении периодических функций.

Пусть $F(x) = F(x + L)$, то есть период функции равен L, и надо найти приближение этой функции по ее значениям, заданным на каком-то отрезке периодичности. Будем считать для определенности, что таковым является отрезок $x \in [0, L]$.

Заметим, что элементарным преобразованием $\varphi = 2\pi x/L$ этот отрезок переводится в $\varphi \in [0, 2\pi]$. Поэтому далее мы будем рассматривать задачу о приближении $f(\varphi)$ по ее значениям на отрезке $[0, 2\pi]$.

Представляется естественным искать приближенное представление периодической функции $f(\varphi)$ в виде

$$Q_n(\varphi) = \sum_{k=0}^{n} a_k \cos k\varphi + \sum_{k=1}^{n} b_k \sin k\varphi. \qquad (\text{П4.1})$$

Замечание. (П4.1) называют *тригонометрическим полиномом n-й степени.* ▲

В записи (П4.1) $N = (2n + 1)$ коэффициентов. Если ставится задача об интерполяции функции $f(\varphi)$, то значения последней должны быть заданы в N точках отрезка $[0, 2\pi]$ (точнее, полуотрезка $[0, 2\pi)$, поскольку в силу периодичности $f(0) = f(2\pi)$ и $Q_n(0) \equiv Q_n(2\pi)$). Пусть это точки $\{\varphi_m,\ m = 0, 1, \ldots, 2n\}$, а соответствующие (заданные) значения функции: $\{f_m = f(\varphi_m),\ m = 0, 1, \ldots, 2n\}$.

Условия интерполяции

$$a_0 + \sum_{k=1}^{n} (a_k \cos k\varphi_m + b_k \sin k\varphi_m) = f_m,\ m = 0, 1, \ldots, 2n, \qquad (\text{П4.2})$$

приводят к замкнутой системе линейных уравнений относительно коэффициентов полинома (П4.1).

Определитель (П4.2)

$$\Delta = \begin{vmatrix} 1 & \cos\varphi_0 & \sin\varphi_0 & \ldots & \cos n\varphi_0 & \sin n\varphi_0 \\ 1 & \cos\varphi_1 & \sin\varphi_1 & \ldots & \cos n\varphi_1 & \sin n\varphi_1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & \cos\varphi_{2n} & \sin\varphi_{2n} & \ldots & \cos n\varphi_{2n} & \sin n\varphi_{2n} \end{vmatrix}. \qquad (\text{П4.3})$$

Вычисления (см. [3, т. 1, с. 156]) приводят к следующему результату:

$$\Delta = 2^{2n^2} \prod_{0 \leqslant l < k \leqslant 2n} \sin \frac{\varphi_k - \varphi_l}{2}, \qquad (\text{П4.4})$$

т. е. при несовпадающих точках $\{\varphi_m\}$ $\Delta \neq 0$ и система (П4.2) имеет единственное решение.

В [3, т. 1, с. 156–163] решение получено в аналитической форме и рассмотрены отдельные примеры.

Это решение — некий аналог полинома Лагранжа для алгебраической интерполяции. Упомянутое аналитическое представление имеет весьма громоздкий вид. Мы более не будем им заниматься, а рассмотрим подробнее иной способ построения тригонометрических полиномов, приближающих функцию $f(x)$.

При этом существенным будет являться предположение о том, что узлы интерполяции равноотстоят друг от друга.

Пусть для определенности координаты узлов задаются формулой

$$\begin{aligned} &\varphi_m = \varphi_0 + m\tau, \quad m = 1, \ldots, N-1 \ (N = 2n+1) \\ &\varphi_0 = \frac{\pi}{N}, \quad \tau = \frac{2\pi}{N}. \end{aligned} \qquad (\text{П4.5})$$

На рис. П.3 крестиками отмечено расположение заданных таким образом узлов на отрезке $[0, 2\pi]$:

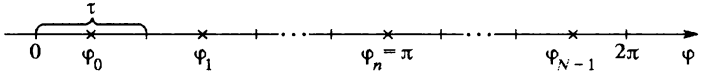


Рис. П3

Рассмотрим различные функции, определенные своими значениями в узлах (П4.5)

$$g^{(\tau)} = \{g_m = g(\varphi_m),\ m = 0, 1, \ldots, N-1\}. \qquad (\text{П4.6})$$

Замечание. Начиная с Лекции 8, объекты вида (П4.6) мы называли сеточными функциями. ▲

Очевидно, что функции $g^{(\tau)}$, заданные множеством своих значений, можно трактовать как N-мерные векторы, которые вкупе образуют N-мерное пространство. Введем в этом пространстве скалярное произведение

$$(g^{(\tau)}, f^{(\tau)}) = \frac{1}{N} \sum_{m=0}^{N-1} g_m f_m \qquad (\text{П4.7})$$

и норму

$$\|g^{(\tau)}\| = (g^{(\tau)}, g^{(\tau)})^{1/2} = \left(\frac{1}{N}\sum_{m=0}^{N-1} g_m^2\right)^{1/2}. \qquad (\text{П4.8})$$

Тригонометрический полином, в виде которого мы ищем приближение к заданной функции $f(\varphi)$, представляет собой, как видно из (П4.1), линейную комбинацию базисных функций

$$\{1, \cos\varphi, \sin\varphi, \cos 2\varphi, \sin 2\varphi, \ldots, \cos n\varphi, \sin n\varphi\}.$$

Последние, если рассматривать их значения в узлах интерполяции, порождают следующие элементы (сеточные функции) в рассматриваемом пространстве:

$$\begin{aligned}
&\psi_{10}^{(\tau)} = \underbrace{1, 1, \ldots, 1}_{N};\\
&\psi_{1r}^{(\tau)} = \{\cos r\varphi_m,\ m = 0, 1, \ldots, N-1\},\quad r = 1, 2, \ldots, n;\\
&\psi_{2r}^{(\tau)} = \{\sin r\varphi_m,\ m = 0, 1, \ldots, N-1\},\quad r = 1, 2, \ldots, n.
\end{aligned} \qquad (\text{П4.9})$$

Оказывается, на равнораспределенных узлах (П4.5) «векторы» (П4.9) взаимно ортогональны в смысле скалярного произведения (П4.7).

Мы ограничимся здесь проверкой этого факта для совокупности элементов, заданных в первой и второй строчках (П4.9).

Итак, покажем, что

$$(\psi_{1r}^{(\tau)},\ \psi_{1s}^{(\tau)}) = \begin{cases} 0, & r \neq s,\\ \sigma \neq 0, & r = s \end{cases} \qquad (\text{П4.10})$$

для $0 \leqslant r, s \leqslant n$.

В самом деле,

$$\begin{aligned}
&\frac{1}{N}\sum_{m=0}^{N-1} \cos r\varphi_m \cos s\varphi_m =\\
&\quad = \frac{1}{4N}\sum (e^{ir(\varphi_0+m\tau)} + e^{-ir(\varphi_0+m\tau)})(e^{is(\varphi_0+m\tau)} + e^{-is(\varphi_0+m\tau)}) =\\
&= \frac{1}{4N}\sum [e^{i(r+s)(\varphi_0+m\tau)} + e^{i(s-r)(\varphi_0+m\tau)} + e^{i(r-s)(\varphi_0+m\tau)} + e^{-i(r+s)(\varphi_0+m\tau)}].
\end{aligned} \qquad (\text{П4.11})$$

В правой части (П4.11) суммирование по m первого слагаемого в квадратных скобках приводит к нулевому результату при любых r и s, не равных нулю одновременно.

Действительно (используя формулу для суммы геометрической прогрессии), получаем:

$$\sum_{m=0}^{N-1} e^{i(r+s)(\varphi_0+m\tau)} = e^{i(r+s)\varphi_0} \sum_{m=0}^{N-1} e^{i(r+s)m\tau} = e^{i(r+s)\varphi_0} \frac{1-e^{i(r+s)N\tau}}{1-e^{i(r+s)\tau}} = 0,$$

так как $N\tau = 2\pi$ и $e^{i(r+s)2\pi} - 1 = 0$, если $r + s \neq 0$.

Точно так же устанавливается, что суммирование последнего слагаемого в (П4.11) всегда дает нулевой результат.

Что касается 2-го и 3-го слагаемых, то суммирование по m для каждого из них приводит к нулю, если $r \neq s$. Если же $r = s$, то

$$e^{i(s-r)(\varphi_0+m\tau)} = e^{i(r-s)(\varphi_0+m\tau)} = 1.$$

В итоге подсчет (П4.11) приводит к следующему результату:

$$\left(\psi_{1r}^{(\tau)},\ \psi_{1r}^{(\tau)}\right) = \frac{1}{N} \sum_{m=0}^{N-1} \cos^2 r\varphi_m = \frac{1}{2}, \qquad \text{(П4.12)}$$

если $r = s \neq 0$ и

$$\left(\psi_{10}^{(\tau)},\ \psi_{10}^{(\tau)}\right) = 1, \quad \text{при} \quad r = s = 0. \qquad \text{(П4.13)}$$

Аналогичным образом проверяется, что

$$\left(\psi_{2r}^{(\tau)},\ \psi_{2s}^{(\tau)}\right) = 0 \text{ при } r \neq s$$

и

$$\left(\psi_{2r}^{(\tau)},\ \psi_{2r}^{(\tau)}\right) = \frac{1}{N} \sum_{m=0}^{N-1} \sin^2 r\varphi_m = \frac{1}{2}. \qquad \text{(П4.14)}$$

Итак, можно считать установленной справедливость соотношений (П4.10) с $\sigma = 1/2$ при $r = s \neq 0$ и $\sigma = 1$ при $r = s = 0$.

В свою очередь эти соотношения означают, что N элементов (векторов) (П4.9) в рассматриваемом пространстве представляют собой ортогональный базис (который с учетом (П4.12)–(П4.14) ничего не стоит сделать ортонормированным).

Возвращаясь теперь к условиям интерполяции (П4.2), мы видим, что их можно трактовать как систему уравнений для коэффициентов разложения «вектора» $f^{(\tau)}$ по элементам указанного базиса:

$$f^{(\tau)} = a_0\psi_0^{(\tau)} + \sum_{k=1}^{n} \left(a_k\psi_{1k}^{(\tau)} + b_k\psi_{2k}^{(\tau)}\right). \qquad \text{(П4.15)}$$

Взаимная ортогональность последних позволяет легко найти эти коэффициенты.

Умножая скалярно (П4.15) последовательно на

$$\psi_{10}^{(\tau)}, \{\psi_{1k}^{(\tau)}, \psi_{2k}^{(\tau)},\ k = 1, 2, \ldots, n\},$$

получаем:

$$a_0 = \frac{1}{N}\sum_{m=0}^{N-1} f_m,$$

$$a_k = \frac{2}{N}\sum_{m=0}^{N-1} f_m \cos k\varphi_m, \qquad \text{(П4.16)}$$

$$b_k = \frac{2}{N}\sum_{m=0}^{N-1} f_m \sin k\varphi_m.$$

Задача о поиске тригонометрического интерполяционного полинома (для случая равноотстоящих узлов) решена. Подробное исследование свойств подобных полиномов в качестве приближения к периодической функции приведено в [17, с. 49–66]. Здесь мы лишь конспективно обозначим некоторые результаты.

1. О погрешности тригонометрической интерполяции. Для погрешности тригонометрической интерполяции полиномами n-й степени в предположении существования r-й производной приближаемой функции $f(\varphi)$ может быть получена следующая оценка ([17]):

$$|R_n(\varphi)| = |f(\varphi) - Q_n(\varphi)| \leqslant \text{const}\,\frac{M_r}{n^{r-2}}, \qquad \text{(П4.17)}$$

где $M_r = \max |f^{(r)}|$ на отрезке периодичности.

Оценка (П4.17) означает, что при наличии у функции третьей непрерывной производной ошибка тригонометрической интерполяции по равноотстоящим узлам *равномерно* стремится к нулю с ростом n (в отличие от алгебраической интерполяции). Скорость убывания погрешности, как видно, зависит от гладкости функции $f(\varphi)$.

Неустранимая погрешность интерполяции это:

$$\Delta_n(\varphi) = \widetilde{Q}_n(\varphi) - Q_n(\varphi),$$

где под $\widetilde{Q}_n(\varphi)$ понимается тригонометрический полином, построенный по «реальным», то есть содержащим погрешность, исходным табличным данным: $\widetilde{f_m} = f_m + \varepsilon_m$.

Пусть погрешность последних ограничена величиной ε :

$$\max_m |\varepsilon_m| \leqslant \varepsilon.$$

Тогда для неустранимой погрешности тригонометрической интерполяции имеет место оценка ([17]):

$$\Delta_n = \max_{\varphi} |\Delta_n(\varphi)| \leqslant L_n \varepsilon, \tag{П4.18}$$

где $L_n \leqslant \dfrac{2}{\pi} \ln n + 1$.

Замечание. Множитель L_n в (П4.18) называют *константой Лебега*. В случае полиномиальной интерполяции по равномерным узлам при больших n ($n \gtrsim 10$) $L_n \gtrsim 2^{n-3}/(n\sqrt{n})$ (см. [17, с. 28], что означает довольно быстрый рост L_n при увеличении n, соответственно неблагоприятное влияние неустранимых погрешностей при полиномиальной интерполяции (см. Пример 2 на с. 194).

Это, как отмечалось в Лекции 5, одна из причин, по которым не рекомендуется использовать для приближения функций интерполяционные алгебраические полиномы высокой степени. ▲

2. Интерполяция четных и нечетных функций. Если функция $f(\varphi)$ на отрезке периодичности является четной, то искомый интерполяционный полином (П4.1), приближающий функцию, принимает вид

$$Q_n(\varphi) = \sum_{k=0}^{n} a_k \cos k\varphi \tag{П4.19}$$

с коэффициентами

$$a_0 = \frac{1}{2n+1}\left(2\sum_{m=0}^{n-1} f_m + f_n\right),$$

$$a_k = \frac{2}{2n+1}\left(2\sum_{m=0}^{n-1} f_m \cos k\varphi_m + f_n \cos k\varphi_n\right), \quad k = 1, 2, \ldots, n. \tag{П4.20}$$

Выписанные выражения для коэффициентов являются следствиями первых двух формул (П4.16), в чем можно убедиться непосредственной проверкой, так же как и в том, что последняя формула (П4.16) дает $b_k = 0$ для всех k.

Таким образом, представляется естественным искать полином, приближающий четную функцию, сразу в виде (П4.19), используя значения функции, заданные на полупериоде ($m = 0, 1, \ldots, n$) либо в виде (П4.26) с узлами (П4.27) (см. ниже).

Если же функция $f(\varphi)$ является нечетной на отрезке периодичности, то (П4.1) превращается в

$$Q_n(\varphi) = \sum_{k=1}^{n} b_k \sin k\varphi, \tag{П4.21}$$

где

$$b_k = \frac{4}{2n+1} \sum_{m=0}^{n-1} f_m \sin k\varphi_m, \quad k = 1, 2, \ldots, n. \tag{П4.22}$$

Замечание. Выражения для коэффициентов (П4.20), (П4.22) отличаются от тех, что приведены в [17, с. 54–55]. Различие связано с тем, что соответствуюшие построения здесь проводились на $N = (2n+1)$ узлах интерполяции, а в [17] на $N = 2(n+1)$. Вообще говоря, в [17] рассмотрены варианты тригонометрической интерполяции и по другим системам узлов. ▲

3. О связи тригонометрической интерполяции с алгебраической интерполяцией по чебышёвским узлам. Существует непосредственная связь полиномиальной интерполяции по чебышёвским узлам, рассмотренной выше, в Приложении 2, с тригонометрической интерполяцией периодических функций.

Памятуя о том, что произвольный отрезок $x \in [a, b]$ всегда может быть преобразован в отрезок $t \in [-1, 1]$ (см. (П1.12)), рассмотрим алгебраическую интерполяцию по чебышёвским узлам для некоторой функции $g(t)$.

Если речь идет о полиноме n-й степени, то требуемые узлы должны совпадать с корнями полинома Чебышёва $(n+1)$-й степени и, согласно (П1.5), распределены по отрезку $[-1, 1]$ следующим образом:

$$t_m = \cos \frac{(2m+1)\pi}{2(n+1)}, \quad m = 0, 1, \ldots, n. \tag{П4.23}$$

Соответствующий интерполяционный полином, в какой бы форме он ни был записан (лагранжевой, ньютоновской и т.д.), очевидно, может быть приведен также к виду

$$P_n(t) = \sum_{m=0}^{n} a_m T_m(t), \tag{П4.24}$$

где $T_m(t)$ — полином Чебышёва m-го порядка. На отрезке $t \in [-1, 1]$ этот полином, как мы видели в Приложении 1, может быть записан в виде $T_m(t) = \cos(m \arccos t)$.

Перепишем с учетом этого обстоятельства (П4.24) в виде

$$P_n(t) = \sum_{m=0}^{n} a_m \cos(m \arccos t). \qquad (П4.25)$$

Перейдем теперь от независимой переменной t к φ по формуле $t = \cos\varphi$. При этом $g(t) \Longrightarrow g(\cos\varphi) = G(\varphi)$, то есть функция $g(t)$, заданная на отрезке $t \in [-1, 1]$, превращается в $G(\varphi)$ — периодическую четную функцию аргумента φ. Соответственно алгебраический интерполяционный полином $P_n(t)$, построенный по чебышёвским узлам, переходит в тригонометрический *четный* полином

$$Q_n(\varphi) = \sum_{m=0}^{n} a_m \cos m\varphi \qquad (П4.26)$$

с равнораспределенными узлами на отрезке $[0, \pi]$, равном половине отрезка периодичности функции $G(\varphi)$. Координаты их, как следует из (П4.23), выражаются формулой

$$\begin{aligned} &\varphi_m = \frac{(2m+1)\pi}{2(n+1)} = \varphi_0 + m\tau, \quad m = 1, \ldots, n, \\ &\varphi_0 = \frac{\pi}{2(n+1)}, \quad \tau = \frac{\pi}{n+1}. \end{aligned} \qquad (П4.27)$$

На рис. П.4 узлы (П4.27) отмечены крестиками.

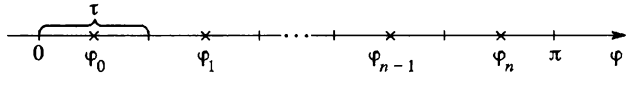


Рис. П4

Коэффициенты полинома (П4.26) определяются по формулам:

$$\begin{aligned} &a_m = \frac{2}{n+1} \sum_{k=0}^{n} G_k \cos m\varphi, \quad m = 1, \ldots, n, \\ &a_0 = \frac{1}{n+1} \sum_{k=0}^{n} G_k. \end{aligned} \qquad (П4.28)$$

Замечание. Отличие формул (П4.28) от (П4.20) объясняется различным расположением узлов интерполяции (сравните рис. П.3 и рис. П.4). Формула (П4.20) была получена для случая, когда узлы распределены по отрезку $[0, 2\pi]$ согласно (П4.5). Всего узловых

точек на этом отрезке $N = 2n + 1$ (нечетное количество), на полупериод ($[0, \pi]$) при этом попадает $(n + 1)$ точка, включая $\varphi_n = \pi$, в то время как текущее распределение (П4.27), будучи распространенным на весь отрезок периодичности, приводит к $2(n + 1)$ узловым точкам на $[0, 2\pi]$ и отвечает расположению узлов, использованному для соответствующих построений в [17, с.50–55] (точка $\varphi = \pi$ не входит в число узловых, на каждый полупериод приходится по $(n + 1)$-му узлу). Соответственно формула (П4.28) выглядит так же, как в [17, с. 54] для аналогичного случая (четной функции). ▲

Таким образом, установлено «родство» алгебраической интерполяции по чебышёвским узлам с тригонометрической интерполяцией некоторой четной периодической функции по равноотстоящим узловым точкам. Опираясь на это обстоятельство, делаем важный вывод о том, что интерполяционные полиномы, построенные по чебышёвским узлам, обладают теми же свойствами (в смысле сходимости, неустранимой погрешности), что и рассмотренные здесь тригонометрические полиномы.

4. О конечных рядах Фурье. Мы поговорим здесь о среднеквадратичном приближении периодических функций тригонометрическими полиномами, некоем аналоге среднеквадратичного полиномиального приближения (см. Лекцию 5).

Снова полагаем, что функция $f(\varphi)$ задана своими значениями f_m в равноотстоящих точках (П4.5) на отрезке периодичности $[0, 2\pi]$.

Будем искать приближенное представление этой функции в виде тригонометрического полинома

$$S_k(\varphi) = \sum_{j=0}^{k} a_j \cos j\varphi + \sum_{j=1}^{k} b_j \sin j\varphi, \qquad \text{(П4.29)}$$

с $k \leqslant n$, используя в качестве критерия близости $S_k(\varphi)$ к $f(\varphi)$ требование, чтобы искомый полином минимизировал сумму квадратов отклонений (функционал)

$$\delta_k(a_0, a_1, b_1, \ldots) = \frac{1}{N} \sum_{m=0}^{N-1} (S_k(\varphi_m) - f_m)^2. \qquad \text{(П4.30)}$$

Замечание. При $k = n$ построенный выше интерполяционный полином $Q_n(\varphi)$, очевидно, обращает величину δ_k в ноль, то есть доставляет абсолютный минимум функционалу (П4.30) и является решением поставленной здесь задачи. Поэтому далее мы будем полагать, что $k < n$. ▲

Запишем условия минимума функционала (П4.30) по не определенным пока коэффициентам $\{a_j, b_j\}$:

$$\frac{\partial \delta_k}{\partial a_j} = \frac{2}{N} \sum_{m=0}^{N-1} (S_k(\varphi_m) - f_m) \cos j\varphi_m = 0, \quad j = 0, 1, \ldots, k,$$

$$\frac{\partial \delta_k}{\partial b_j} = \frac{2}{N} \sum_{m=0}^{N-1} (S_k(\varphi_m) - f_m) \sin j\varphi_m = 0, \quad j = 1, 2, \ldots, k,$$

или в развернутом, с учетом (П4.29), виде (после некоторой перегруппировки слагаемых)

$$a_0 \sum_{m=0}^{N-1} \cos j\varphi_m + \sum_{l=1}^{k} \left(\frac{1}{N} a_l \sum_{m=0}^{N-1} \cos l\varphi_m \cos j\varphi_m \right) + \\ + \sum_{l=1}^{k} \left(\frac{1}{N} b_l \sum_{m=0}^{N-1} \sin l\varphi_m \cos j\varphi_m \right) = \frac{1}{N} \sum_{m=0}^{N-1} f_m \cos j\varphi_m,$$

для $j = 0, 1, \ldots, k$.

$$a_0 \sum_{m=0}^{N-1} \sin j\varphi_m + \sum_{l=1}^{k} \left(\frac{1}{N} a_l \sum_{m=0}^{N-1} \cos l\varphi_m \sin j\varphi_m \right) + \\ + \sum_{l=1}^{k} \left(\frac{1}{N} b_l \sum_{m=0}^{N-1} \sin l\varphi_m \sin j\varphi_m \right) = \frac{1}{N} \sum_{m=0}^{N-1} f_m \sin j\varphi_m,$$

для $j = 1, 2, \ldots, k$.

Учитывая введенное выше определение скалярного произведения (П4.7), перепишем эти соотношения в виде

$$a_0(\psi_{10}^{(\tau)}, \psi_{1j}^{(\tau)}) + \sum_{l=1}^{k} a_l(\psi_{1l}^{(\tau)}, \psi_{1j}^{(\tau)}) + \sum_{l=1}^{k} b_l(\psi_{2l}^{(\tau)}, \psi_{1j}^{(\tau)}) = (f^{(\tau)}, \psi_{1j}^{(\tau)}),$$

для $j = 0, 1, , \ldots, k$. (П4.32)

$$a_0(\psi_{10}^{(\tau)}, \psi_{2j}^{(\tau)}) + \sum_{l=1}^{k} a_l(\psi_{1l}^{(\tau)}, \psi_{2j}^{(\tau)}) + \sum_{l=1}^{k} b_l(\psi_{2l}^{(\tau)}, \psi_{2j}^{(\tau)}) = (f^{(\tau)}, \psi_{2j}^{(\tau)}),$$

для $j = 1, 2, \ldots, k$.

И, наконец, принимая во внимание условия ортогональности (П4.12)–(П4.14), приходим к выводу, что *коэффициенты искомого*

полинома при $k < n$ вычисляются по тем же формулам (П4.16), *что и коэффициенты интерполяционного полинома (при $k = n$).*

5. Ряды Фурье и тригонометрические полиномы. Совершенно естественной представляется связь тригонометрических полиномов (П4.29) с рядом Фурье для периодических функций

$$S(\varphi) = \frac{a_0}{2} + \sum_{l=1}^{\infty} \alpha_l \cos l\varphi + \beta_l \sin l\varphi. \qquad \text{(П4.33)}$$

Коэффициенты $\{\alpha_l, \beta_l\}$ вычисляются по формулам

$$\begin{aligned} \alpha_l &= \frac{1}{\pi} \int_0^{2\pi} f(\varphi) \cos l\varphi \, d\varphi, \quad l = 0, 1, \ldots \\ \beta_l &= \frac{1}{\pi} \int_0^{2\pi} f(\varphi) \sin l\varphi \, d\varphi, \quad l = 1, 2, \ldots \end{aligned} \qquad \text{(П4.34)}$$

(Подразумевается, что $[0, 2\pi]$ является отрезком периодичности функции $f(\varphi)$).

Если функция $f(\varphi)$ задана лишь конечным набором своих значений в N равноостоящих точках отрезка периодичности, то для вычисления интегралов (П4.34) приходится обращаться к формулам численного интегрирования. При этом выясняется любопытное обстоятельство: использование метода прямоугольников превращает формулы (П4.34) в (П4.16)!

Если точки, в которых заданы значения функции, распределены по закону (П4.5), то, как было показано, базисные функции

$$\{1, \cos\varphi, \sin\varphi, \cos 2\varphi, \sin 2\varphi, \ldots\},$$

разложение по которым представляет собой ряд Фурье, переходят в конечный набор линейно независимых элементов (векторов)(П4.9). Соответственно бесконечный ряд (П4.33) переходит в конечную сумму (П4.15), то есть в интерполяционный тригонометрический полином.

В силу отмеченной аналогии тригонометрические полиномы (П4.29) для функции $f(\varphi)$ часто называют *конечными рядами Фурье* (в том числе и при $k < n$).

Замечание. В радиотехнике обработка периодического сигнала $f(\varphi)$ по формулам (П4.34) называется частотным анализом, а представление $f(\varphi)$ в виде ряда Фурье (П4.33) — синтезом сигнала по составляющим его гармоникам.

Возможность разложения $f(\varphi)$ в ряд Фурье означает, что периодическая функция имеет бесконечный дискретный спектр. Если из-

вестен лишь ограниченный набор значений сигнала

$$\{f_m,\ m = 0, 1, \ldots, N\},$$

то частотный анализ сводится к вычислению коэффициентов (амплитуд соответствующих гармоник) по формулам (П4.16). При этом выявляется лишь конечный дискретный спектр (главные частоты) сигнала $f(\varphi)$, который может быть представлен по гармоникам этого спектра в виде конечного ряда Фурье (П4.29). ▲

В заключение приведем пример использования конечных рядов Фурье (тригонометрических интерполяционных и среднеквадратичных полиномов) для приближения конкретной функции.

Пример 1. В табл. П4.1 приведена ошибка (максимальная по модулю) приближения функции $f(\varphi) = \sin^5\varphi + \cos^4\varphi$ с помощью различных тригонометрических полиномов. В первом столбце указана величина n, определяющая количество табличных точек, в которых заданы значения функции ($N = 2n + 1$). Во втором указана степень тригонометрического полинома — k, используемого для приближения данной функции (при $k = n$ имеет место тригонометрическая интерполяция, при $k < n$ — среднеквадратичное приближение).

На рис. П5 показаны результаты, соответствующие случаю $n = 5$, $k = 4$.

Таблица П4.1

n	k	$\Delta_{\text{приближения}}$
5	5	0.0
5	4	0.0625
5	3	0.1875
4	4	0.1219
4	3	0.1875
3	3	0.342

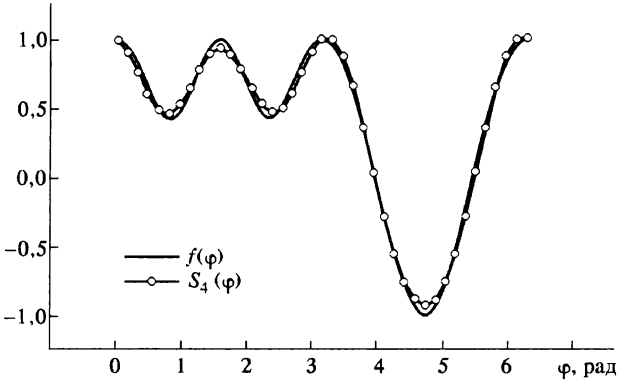


Рис. П5

В свете сказанного в пункте 3 данного Приложения результаты интерполирования по чебышёвским узлам, приведенные в Приложении 2, в таблицах П2.2, П2.3 для полиномов четной степени, можно

рассматривать в качестве примеров тригонометрической интерполяции (по косинусам) для $n = 2, 3, \ldots$ соответственно.

Весьма содержательное и подробное изложение вопросов тригонометрической аппроксимации периодических функции можно найти в книге [37, с. 481–553].

О БЫСТРОМ ПРЕОБРАЗОВАНИИ ФУРЬЕ

Быстрым преобразованием Фурье в литературе по численным методам называют экономичный (рациональный) способ вычисления многих $(0 \leqslant k \leqslant N)$ сумм вида

$$a_k = \frac{1}{N}\sum_{m=0}^{N-1} f_m \cos(k\varphi_m), \quad b_k = \frac{1}{N}\sum_{m=0}^{N-1} f_m \sin(k\varphi_m). \qquad (\text{П5.1})$$

Используя комплексную форму записи, можно объединить определение величин a_k и b_k в виде

$$c_k = \frac{1}{N}\sum_{m=0}^{N-1} f_m \exp(ik\varphi_m), \quad k = 0, 1, \ldots, \qquad (\text{П5.2})$$

где $c_k = a_k + ib_k$ (i — мнимая единица).

Мы сталкивались с подобными суммами, когда обсуждали тригонометрическую интерполяцию. Собственно, и термин (название данного раздела) проистекает оттуда. Вспомним, что тригонометрический интерполяционный полином

$$f(\varphi) \approx a_0 + \sum_{k=1}^{n} (a_k \cos(k\varphi) + b_k \sin(k\varphi))$$

это не что иное, как «конечный ряд Фурье» для периодической функции, заданной своими значениями в равнораспределенных (в пределах периода) точках: $f_m = f(\varphi_m)$ или, другими словами, дискретное преобразование Фурье для этой функции, коэффициенты которого a_k, b_k вычисляются как раз по формулам типа (П5.1).

Проблема в следующем: вычисление коэффициентов ряда непосредственно по формулам (П5.1) требует (по порядку величины) N^2 сложений, умножений и обращений к функциям вычисления sin и cos (по N указанных операций для каждого коэффициента, всего же коэффициентов $\sim 2N$).

Мы обсудим здесь способ существенно более «быстрого» вычисления сумм подобного вида. Будем считать, что область определения функции $f(\varphi)$ — отрезок $[0, 2\pi]$, а точки, в которых заданы значения f_m, распределены по закону

$$\varphi_m = \varphi_0 + m\tau, \; m = 1, \ldots, N-1; \quad \tau = \frac{2\pi}{N}, \quad \varphi_0 = \frac{\pi}{N} = \frac{\tau}{2}. \qquad (\text{П5.3})$$

Последующие выкладки имеют целью пояснить суть дела, показать, за счет чего достигается экономия вычислительных действий.

Учитывая (П5.3), перепишем формулу (П5.2) в виде

$$c_k = \frac{1}{N}\sum_{m=0}^{N-1} f_m e^{ik(\varphi_0+m\tau)} = e^{ik\varphi_0}\frac{1}{N}\sum_m f_m e^{i(km)\tau}. \qquad (П5.4)$$

Внимательный взгляд на правую часть (П5.4) в какой-то степени позволяет понять, что надо бы сделать для оптимизации вычислений (в смысле сокращения количества операций).

В сумме $\sum_m f_m e^{i(km)\tau}$ содержится, очевидно, множество (при больших N) слагаемых с одинаковыми множителями $e^{i(km)\tau}$. За счет периодичности они (множители) совпадают для любых пар (k_1, m_1), (k_2, m_2), для которых $(k_1 m_1)\pi - (k_2 m_2)\pi = 0$ или пропорционально 2π.

Соответственно, обсуждаемый ниже способ «быстрого» суммирования сводится к перегруппировке слагаемых в совокупности сумм (П5.4) для $k = 0, 1, \ldots$, позволяющей не вычислять многократно совпадающие множители и, более того, не вычислять повторно некоторые частичные суммы, общие для (П5.4) при разных k.

Чтобы последующие выкладки, по возможности, носили более компактный характер, введем дополнительные обозначения:

$$s_0 = e^{ik\varphi_0}, \quad w = e^{i\tau} = e^{i\cdot 2\pi/N} \qquad (П5.5)$$

После этого формула (П5.4) перепишется в виде

$$c_k = \frac{s_0}{N}\sum_{m=0}^{N-1} f_m w^{km}. \qquad (П5.6)$$

Далее, предположим, что N представимо в виде произведения двух целых чисел $N = N_1 N_2$.

Разобьем сумму (П5.6) на группы, вводя новые переменные суммирования m_1 и m_2 так, что

$$m = m_1 + N_1 m_2, \quad 0 \leqslant m_1 \leqslant N_1 - 1, \quad 0 \leqslant m_2 \leqslant N_2 - 1. \qquad (П5.7)$$

Индекс k также представим в виде

$$k = k_1 + N_2 k_2, \quad 0 \leqslant k_1 \leqslant N_2 - 1, \quad 0 \leqslant k_2 \leqslant N_1 - 1, \qquad (П5.8)$$

что позволит, как мы сейчас убедимся, внутри сумм (П5.4) выделить с пользой для дела некоторые повторяющиеся частичные суммы.

Формула (П5.6) с использованием m_1, m_2, k_1, k_2 может быть записана так:

$$c_k = \frac{s_0}{N_1} \sum_{m_1=0}^{N_1-1} \left(\frac{1}{N_2} \sum_{m_2=0}^{N_2-1} f_{m_1+N_1 m_2} w^{(k_1+N_2 k_2)(m_1+N_1 m_2)} \right) =$$

$$= \frac{s_0}{N_1} \sum_{m_1=0}^{N_1-1} \left(\frac{1}{N_2} \sum_{m_2=0}^{N_2-1} f_{m_1+N_1 m_2} w^{\overbrace{m_1 k_1 + m_1 N_2 k_2}^{m_1 k} + N_1 k_1 m_2 + N_1 N_2 k_2 m_2} \right). \tag{П5.9}$$

Учитывая, что

$$w^{N_1 N_2 k_2 m_2} = w^{N k_2 m_2} = \left(e^{i \cdot 2\pi/N}\right)^{N k_2 m_2} = e^{i(2\pi k_2 m_2)} \equiv 1$$

(так как $k_2 m_2$ — целое число), перепишем (П5.9) в виде

$$c_k = \frac{s_0}{N_1} \sum_{m_1=0}^{N_1-1} \left(\underbrace{\frac{1}{N_2} \sum_{m_2=0}^{N_2-1} f_{m_1+N_1 m_2} w^{N_1 k_1 m_2}}_{c(k_1, m_1)} \right) w^{m_1 k}. \tag{П5.10}$$

Если теперь использовать обозначение $c(k_1, m_1)$ для внутренних сумм (как показано в (П5.10)), то можно увидеть, что алгоритм вычисления совокупности $\{c_k\}$ сводится к предварительному вычислению величин $c(k_1, m_1)$ по формулам

$$c(k_1, m_1) = \frac{1}{N_2} \sum_{m_2=0}^{N_2-1} f_{m_1+N_1 m_2} w^{N_1 k_1 m_2} \tag{П5.11}$$

для $k_1 = 0, 1, \ldots, N_2 - 1$; $m_1 = 0, 1, \ldots, N_1 - 1$ и окончательному суммированию:

$$c_k = \frac{s_0}{N_1} \sum_{m_1=0}^{N_1-1} c(k_1, m_1) w^{m_1 k} \tag{П5.12}$$

для $k = 0, 1, \ldots, N - 1$ (при выборе номера k_1 для множителей $c(k_1, m_1)$ следует учитывать, что, согласно (П5.8), $k = k_1 + N_2 k_2$, то есть $k_1 = k - N_2 k_2$, причем для текущего номера k при определении k_1 надо взять такое k_2, чтобы значение k_1 попало в интервал $0 \leqslant k_1 \leqslant N_2 - 1$).

Посмотрим, что же мы получили, перейдя от формул (П5.2) непосредственного вычисления сумм c_k к формулам (П5.11), (П5.12). Оценим количество операций, которое требуют последние.

Для вычисления промежуточных величин $c(k_1, m_1)$, с учетом диапазона значений k_1, m_1 и пределов суммирования в (П5.11), нужно выполнить

$$\Omega_2' \sim \underbrace{N_1 N_2}_{\text{кол-во } c(k_1,m_1)} \cdot \underbrace{2N_2}_{\substack{\text{кол-во умножений и сложений} \\ \text{для одного } c(k_1,m_1)}} = 2NN_2$$

операций сложения и умножения.

Соответственно, реализация (П5.12) требует

$$\Omega_2'' \sim \underbrace{N}_{\text{кол-во } c_k} \cdot \underbrace{2N_1}_{\substack{\text{кол-во умножений и сложений} \\ \text{для одного } c_k}} = 2NN_1.$$

Полное количество учитываемых операций

$$\Omega_2 = \Omega_2' + \Omega_2'' = 2N(N_1 + N_2) < 2N^2 = \Omega_1,$$

так как $N_1 + N_2|_{N_1 N_2 > 4} < N_1 N_2 = N$.

В самом деле, имеем $\left(\sqrt{N_1} - \sqrt{N_2}\right)^2 > 0 \Rightarrow N_1 + N_2 < 2\sqrt{N_1 N_2}$. Далее из $2\sqrt{N_1 N_2} < N_1 N_2 \Rightarrow N_1 N_2 > 4$. То есть при $N_1 N_2 > 4$ имеет место $N_1 + N_2 < N_1 N_2$.

Пусть, например, $N_1 = N_2 = \sqrt{N}$. Тогда выигрыш, отношение вычислительных затрат в первом и втором способе вычислений

$$\frac{\Omega_1}{\Omega_2} = \frac{2N^2}{2N(N_1 + N_2)} = \frac{N}{2\sqrt{N}} = \frac{\sqrt{N}}{2}. \qquad \text{(П5.13)}$$

При больших N, как видим, экономия может быть заметной.

Замечание 1. Выше сравнивалось количество элементарных арифметических операций (сложения и умножения). Ясно, что наибольшие затраты процессорного времени связаны с реализацией вычислений синусов и косинусов, «зашифрованных» в формулах (П5.2) и последующих выкладках экспонентой $e^{i\{\dots\}}$. Счет по формулам (П5.2) требует, очевидно, $\sim N^2$ обращений к соответствующим стандартным функциям; формулы: (П5.11) — NN_2, (П5.12) — NN_1. Таким образом, оценка (П5.13) остается в силе. ▲

Замечание 2. Можно попытаться понять, за счет чего достигается экономия.

Во-первых, диапазон различных степеней w

$$\text{в (П5.11)} \quad 0 \leqslant N_1 k_1 m_2 \leqslant N_1 (N_2 - 1)^2 \approx N_1 N_2^2 = N_2 N,$$

$$\text{в (П5.12)} \quad 0 \leqslant m_1 k \leqslant (N_1 - 1)(N - 1) \approx N_1 N$$

меньше, нежели в (П5.2), что влечет уменьшение одинаковых (но каждый раз вычисляемых) множителей вида w^l (l — целое) в разных слагаемых разных сумм.

Во-вторых, можно усмотреть, что величины $c(k_1, m_1)$ представляют собой некие частичные суммы, повторяющиеся в формулах для разных c_k. В последовательности вычислений по формулам (П5.11), (П5.12) они вычисляются по одному разу и далее (в (5.12)) используются столько раз, сколько нужно. ▲

Замечание 3. Возможна следующая трактовка формул (П5.11), (П5.12): (П5.11) представляет собой формулу для коэффициентов ряда Фурье по m_1-му интервалу исходных данных (на котором задано N_2 значений функций $\{f_{m_2}, m_2 = 0, 1, \ldots, N_2 - 1\}$. Если теперь эти коэффициенты рассматривать как значения некоторой функции, заданной в точках, отмеченных номерами m_1, то (П5.12) будет определять коэффициенты ряда Фурье этой функции. ▲

Изложенная процедура модификации вычислений может быть продолжена. Если значение N_2 представимо в виде произведения двух целых чисел, то вычисление совокупности величин $\{c(k_1, m_1)\}$ совершенно аналогичным образом может быть рационализировано и т. д.

Эта идея особенно наглядно «развертывается», если $N = 2^L$. В результате получается последовательность расчетных формул, которую, собственно, и принято в вычислительной математике называть *быстрым преобразованием Фурье* (БПФ).

Можно показать, что количество арифметических операций, требуемых для реализации этого метода,

$$\Omega \approx 4NL = 4N \log_2 N,$$

что в $\Omega/\Omega_1 \approx \dfrac{2 \log_2 N}{N}$ раз меньше, нежели требуется при вычислениях непосредственно по формуле (П5.2).

Более подробное изложение рассмотренных здесь вопросов (в том числе вывод расчетных формул БПФ с указанием последовательности их применения) можно найти в рекомендуемых ниже источниках.

З а м е ч а н и е. Рассмотренный метод рационального суммирования обобщается на двумерный случай: вычисление сумм вида

$$d_{k,l} = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} f_{m,n} e^{i(k\varphi_n + l\psi_m)}$$

и т. д. ▲

С изложенным в данном Приложении материалом можно дополнительно познакомиться по следующим источникам [1, с. 340–342; 2, с. 170–173; 9, с. 416–419; 12, с. 334–338; 13, с. 164–185].

ПРИЛОЖЕНИЕ 6

ВЫЧИСЛЕНИЕ ИНТЕГРАЛОВ ОТ БЫСТРООСЦИЛЛИРУЮЩИХ ФУНКЦИЙ

В этом Приложении мы обсудим способы вычисления интегралов вида

$$I_1 = \int_a^b f(x)\cos\omega x\,dx, \quad I_2 = \int_a^b f(x)\sin\omega x\,dx. \tag{П6.1}$$

Предполагается, что $f(x)$ — не сильно меняющаяся на отрезке $[a, b]$ функция, и частота колебаний подынтегральных функций в (П6.1) определяется величиной ω, относительно которой будем считать, что $\omega \gg 1$.

Заметим, что поставленная задача не абстрактна, в радиотехнических, электротехнических приложениях возникают подобные проблемы. Итак, возвращаемся к обозначенной задаче.

Если воспользоваться комплексным представлением результатов $I = I_1 + iI_2$ (i — мнимая единица), то две формулы (П6.1) могут быть объединены в одну:

$$I = \int_a^b f(x)e^{i\omega x}\,dx. \tag{П6.2}$$

Использование для вычисления интегралов обычных квадратурных формул (трапеций, прямоугольников, Симпсона) в данном случае крайне нерационально, так как для получения результата с приемлемой точностью необходимо брать очень малый шаг интегрирования. В самом деле, погрешность формулы трапеций (прямоугольников) $\sim M_2h^2$, где M_2 — максимум модуля второй производной от подынтегральной функции на отрезке интегрирования. В наших условиях $M_2 \sim \omega^2$ и рассчитывать на малую погрешность можно, если только $\omega h \ll 1$ (для получения всего лишь верного порядка надо, очевидно, чтобы на каждом колебании подынтегральной функции располагалось несколько узлов интегрирования, то есть должно быть $h < 1/\omega$).

Формула Симпсона в данном случае теряет свои преимущества, так как для нее погрешность $\sim M_4h^4$, $M_4 \sim \omega^4$, и мы приходим к тому же ограничению на шаг интегрирования: $\omega h \ll 1$.

Предлагается следующий естественный способ конструирования квадратурных формул, учитывающих специфику подынтегральной функции в (П6.2) (то есть в интегралах (П6.1)).

Как обычно, в рассмотрение вводится сетка узлов интегрирования

$$\{x_0 = a,\ x_{k+1} = x_k + h_k\ (k = 0, 1, \ldots, K-1),$$

так что $x_K = b\}$, которая может быть равномерной, если $h_k = h = \text{const} = (b-a)/K$.

Искомый результат представляется в виде суммы

$$I = \sum_{k=0}^{K-1} \int_{x_k}^{x_{k+1}} f(x) e^{i\omega x} dx = \sum_{k=0}^{K-1} i_k, \qquad (\text{П6.3})$$

где

$$i_k = \int_{x_k}^{x_{k+1}} f(x) e^{i\omega x}\, dx. \qquad (\text{П6.4})$$

Для вычисления же интегралов i_k по элементарным отрезкам $[x_k, x_{k+1}]$ будем использовать полиномиальную интерполяцию не всей подынтегральной функции, а лишь неосциллирующей ее части.

А). Приблизим $f(x)$ на отрезке $[x_k, x_{k+1}]$ полиномом нулевой степени, конкретно ее значением в центре $f(x) \approx f_{k+1/2} = f(x_k + h_k/2)$.

Тогда

$$i_k \approx f_{k+1/2} \int_{x_k}^{x_{k+1}} e^{i\omega x}\, dx.$$

Интеграл от осциллирующей части исходной функции вычисляется элементарно:

$$i_k \approx \frac{2}{\omega} f_{k+1/2} e^{i\omega x_{k+1/2}} \sin\frac{\omega h_k}{2} = \frac{\sin p_k}{p_k} h_k f_{k+1/2} e^{i\omega x_{k+1/2}}, \qquad (\text{П6.5})$$

где $p_k = \dfrac{\omega h_k}{2}$.

Следовательно, сумма (П6.3) может быть записана в виде:

$$I \approx \sum_{k=0}^{K-1} \frac{\sin p_k}{p_k} h_k f_{k+1/2} e^{i\omega x_{k+1/2}}. \qquad (\text{П6.6})$$

Для равноотстоящих узлов интегрирования ($h_k = h = \text{const}$)

$$I \approx h \frac{\sin p}{p} \sum_{k=0}^{K-1} f_{k+1/2} e^{i\omega x_{k+1/2}}. \qquad (\text{П6.7})$$

Очевидно, что построенная квадратурная формула является аналогом формулы прямоугольников с центральной точкой (Лекция 6) и формально переходит в последнюю при $\omega \to 0$ (для $\int_a^b f(x)\,dx$).

Как видно, отличие результатов, вычисленных (для $h = \text{const}$) по методу прямоугольников и по формуле (П6.7), определяется единственным множителем: $(\sin p)/p$. Оказывается, учет этого множителя позволяет заметным образом снизить требования к шагу интегрирования.

Чтобы убедиться в этом, оценим погрешность предложенного метода. Представим функцию $f(x)$ на элементарном отрезке $[x_k, x_{k+1}]$ по формуле Тэйлора относительно центра $x_{k+1/2} = x_k + h_k/2$:

$$f(x) = f_{k+1/2} + f'_{k+1/2}(x - x_{k+1/2}) + \widetilde{f}''_k(x - x_{k+1/2})^2. \qquad (\text{П6.8})$$

Здесь имеется в виду, что вторая производная $\widetilde{f}''_k(x)$ вычислена в некоторой точке $\widetilde{x} \in [x_k, x_{k+1}]$.

Подставляя (П6.8) в формулу (П6.4) для интеграла по элементарному отрезку, получаем

$$i_k = f_{k+1/2} \int_{x_k}^{x_{k+1}} e^{i\omega x}\,dx + f'_{k+1/2} \int_{x_k}^{x_{k+1}} (x - x_{k+1/2}) e^{i\omega x}\,dx +$$

$$+ \widetilde{f}''_k \int_{x_k}^{x_{k+1}} (x - x_{k+1/2})^2 e^{i\omega x}\,dx.$$

Первое слагаемое, очевидно, представляет собой построенную выше приближенную формулу (П6.5). Второе и третье слагаемые характеризуют погрешность формулы (П6.5) для интеграла i_k по локальному отрезку. Разберемся с ними последовательно.

После несложных выкладок можно получить

$$r_k^{(1)} = f'_{k+1/2} \int_{x_k}^{x_{k+1}} (x - x_{k+1/2})\, e^{i\omega x} dx =$$

$$= i\frac{h_k^2}{2} \frac{(\sin p_k - p_k \cos p_k)}{p_k^2} f'_{k+1/2} e^{i\omega x_{k+1/2}}. \qquad (\text{П6.9})$$

$$r_k^{(2)} = \widetilde{f}_k'' \int_{x_k}^{x_{k+1}} (x - x_{k+1/2})^2 e^{i\omega x}\, dx =$$

$$= \frac{h_k^3}{4} \widetilde{f}_k''\, e^{i\omega x_{k+1/2}} \left[\frac{1}{2} \frac{\sin p_k}{p_k} - \frac{(\sin p_k - p_k \cos p_k)}{p_k^3} \right]. \qquad (\text{П6.10})$$

При малых p_k имеем:

$$\frac{(\sin p_k - p_k \cos p_k)}{p_k^2} = \frac{p_k - \dfrac{p_k^3}{6} + \dfrac{p_k^5}{120} + \ldots - p_k \left(1 - \dfrac{p_k^2}{2} + \dfrac{p_k^4}{24} + \ldots\right)}{p_k^2} =$$

$$+ \frac{p_k}{3} - \frac{p_k^3}{30} + \ldots = \frac{p_k}{3} \left(1 - \frac{p_k^2}{10}\right) + O(p_k^5). \qquad (\text{П6.11})$$

Аналогично устанавливается, что при этом

$$\frac{1}{2} \frac{\sin p_k}{p_k} - \frac{(\sin p_k - p_k \cos p_k)}{p_k^3} = \frac{1}{6} - \frac{p_k^2}{20} + O(p_k^4). \qquad (\text{П6.12})$$

Таким образом, формально при $\omega \to 0$ (то есть при $p_k \to 0$) имеет место $r_k^{(1)} \to 0$ и, учитывая (П6.12), мы приходим к известной формуле локальной погрешности метода прямоугольников для интеграла $\int_a^b f(x)\, dx$:

$$r_k = r_k^{(1)} + r_k^{(2)} = \frac{h_k^3}{24} \widetilde{f}_k''.$$

При умеренных значениях $\omega h_k \sim 1$ ($p_k \sim 1/2$) имеем (принимая во внимание (П6.11) и (П6.12)):

$$|r_k^{(1)}| \leqslant \frac{h_k^2}{6} p_k M_1^{(k)} \approx \frac{h_k^2}{12} M_1^{(k)}, \quad \text{где } M_1^{(k)} = \max_{[x_k, x_{k+1}]} |f'|;$$

$$|r_k^{(2)}| \leqslant \frac{h_k^3}{24} M_2^{(k)}, \quad \text{где } M_2^{(k)} = \max_{[x_k, x_{k+1}]} |f''|.$$

Суммируя оценки для локальных погрешностей, получим формулу для глобальной (по всему отрезку интегрирования $[a, b]$) погрешности метода (П6.6):

$$|R| \leqslant \sum_k (|r_k^{(1)}| + |r_k^{(2)}|) \leqslant \frac{h(b-a)}{12} M_1 + \frac{h^2}{24} (b-a) M_2 \qquad (\text{П6.13})$$

с $h = \max\limits_k h_k$, $M_1 = \max\limits_{[a,b]} |f'|$, $M_2 = \max\limits_{[a,b]} |f''|$.

Первое слагаемое в правой части (П6.13) является определяющим по порядку величины (при умеренных значениях M_1, M_2). Таким образом, квадратурная формула (П6.6) для вычисления интегралов вида (П6.2) дает гарантированно малую (порядка $O(h)$) погрешность при существенно менее жестких требованиях к шагу интегрирования сравнительно с обычной формулой прямоугольников ($\omega h \sim 1$ вместо $\omega h \ll 1$).

Замечание. Если функция $f(x)$ слабо меняется на отрезке $[a, b]$, то есть если M_1 мало, то, очевидно, точность формулы (П6.6) повышается. В частности, если $M_1 \sim O(h)$, то погрешность (П6.13) становится величиной второго порядка малости. ▲

Приведем еще некоторые дополнительные соображения относительно свойств предложенного метода. Если бы мы просуммировали выражения для $r_k^{(1)}$, $r_k^{(2)}$ ((П6.9), (П6.10)), а не $|r_k^{(1)}|$, $|r_k^{(2)}|$, то получили бы, в предположении $\omega h_k \sim 1$ ($p_k \sim 1/2$) и с учетом (П6.11), (П6.12), следующее представление для глобальной погрешности метода:

$$R = \sum_k r_k^{(1)} + r_k^{(2)} \approx i\frac{\omega}{12}\sum_{k=0}^{K-1} f'_{k+1/2} h_k^3 e^{i\omega x_{k+1/2}} + \frac{1}{24}\sum_{k=0}^{K-1} \widetilde{f}''_k h_k^3 e^{i\omega x_{k+1/2}}. \tag{П6.14}$$

При $h_k = h = \text{const}$ (П6.14) принимает вид:

$$R \approx i\frac{\omega h^2}{12}\sum_{k=0}^{K-1} f'_{k+1/2} e^{i\omega x_{k+1/2}} h + \frac{h^2}{24}\sum_{k=0}^{K-1} \widetilde{f}''_k e^{i\omega x_{k+1/2}} h \approx$$

$$\approx i\frac{h}{12}\int_a^b f' e^{i\omega x}\,dx + \frac{h^2}{24}\int_a^b f''\, e^{i\omega x}\,dx + O(h^3). \tag{П6.15}$$

Действительно, мы только что (выше) показали, что

$$\int_a^b f e^{i\omega x}\,dx = \sum_{k=0}^{K-1} f_{k+1/2}\, e^{i\omega x_{k+1/2}} h + O(h).$$

Следовательно,

$$\sum_{k=0}^{K-1} f'_{k+1/2}\, e^{i\omega k_{k+1/2}} h = \int_a^b f' e^{i\omega x}\,dx + O(h)$$

и

$$\sum_{k=0}^{K-1} \widetilde{f}_k'' \, e^{i\omega k_{k+1/2}} \, h \approx \int_a^b f'' e^{i\omega x} \, dx + O(h).$$

Из (П6.15) можно извлечь дополнительные выводы. Например, если интеграл вычисляется по отрезку $[-\pi, \pi]$ и $f'(x)$ на этом отрезке — нечетная функция, то погрешность вычисления второго интеграла из (П6.1) по формуле (П6.7) будет величиной второго порядка малости. В самом деле, сопоставляя действительные и мнимые части (П6.7) и (П6.15), убеждаемся, что погрешность для данного интеграла будет определяться выражением

$$R(\sin) \approx \frac{h}{12} \int_{-\pi}^{\pi} f' \cos \omega x \, dx + \frac{h^2}{24} \int_{-\pi}^{\pi} f'' \sin \omega x \, dx.$$

Для нечетной функции $f'(x)$ перый интеграл обращается в ноль.

Если же $f'(x)$ является четной функцией на отрезке интегрирования, то формула (П6.7) обеспечивает второй порядок точности при вычислении первого из интегралов (П6.1).

З а м е ч а н и е. Подход к конструированию методов квадратур для интегралов типа (П6.2) берет начало от работы Филона*). Получаемые на этом пути формулы поэтому принято называть формулами Филона. Следуя этой традиции, будем в дальнейшем называть метод, разобранный в данном пункте, методом Филона типа прямоугольников (или, короче, методом Филона-А). ▲

Б) В этом пункте мы получим и изучим еще одну квадратурную формулу для вычисления интегралов (П6.2) — формулу Филона типа трапеций (Филона-Б).

Приблизим на элементарном отрезке $[x_k, x_{k+1}]$ функцию $f(x)$ интерполяционным полиномом первой степени

$$f(x) \approx f_k + \frac{f_{k+1} - f_k}{h_k} (x - x_k), \qquad \text{(П6.16)}$$

*) *L. N. G. Filon*, Proc. Roy. Soc., Edinb., 1928–1929, p. 49.

тогда

$$i_k = \int\limits_{x_k}^{x_{k+1}} f e^{i\omega x}\,dx \approx \int\limits_{x_k}^{x_{k+1}} \left(f_k + \frac{f_{k+1} - f_k}{h_k}(x - x_k)\right) e^{i\omega x}\,dx =$$

$$= f_k \int\limits_{x_k}^{x_{k+1}} e^{i\omega x}\,dx + \frac{f_{k+1} - f_k}{h_k} \int\limits_{x_k}^{x_{k+1}} (x - x_k) e^{i\omega x}\,dx =$$

после замены $\xi = x - x_{k+1/2}$ $(x = \xi + x_{k+1/2})$

$$= e^{i\omega x_{k+1/2}} \int\limits_{-h_k/2}^{h_k/2} \left[f_k + \frac{f_{k+1} - f_k}{h_k}\left(\xi + \frac{h_k}{2}\right)\right] e^{i\omega\xi}\,d\xi. \qquad \text{(П6.17)}$$

Введем вспомогательные обозначения для промежуточных величин

$$S_0 = \int\limits_{-h_k/2}^{h_k/2} e^{i\omega\xi} d\xi = \frac{1}{i\omega}(e^{i\omega h_k/2} - e^{-i\omega h_k/2}) = \frac{2}{\omega}\sin\frac{\omega h_k}{2} = h_k \frac{\sin p_k}{p_k}$$

(П6.18)

(напомним, что обозначение $p_k = \dfrac{\omega h_k}{2}$ уже было введено прежде),

$$S_1 = \int\limits_{x_k}^{x_{k+1}} \xi e^{i\omega\xi}\,d\xi = i\frac{h_k^2}{2}\frac{\sin p_k - p_k \cos p_k}{p_k^2}. \qquad \text{(П6.19)}$$

Тогда (П6.17) может быть записано в виде

$$i_k \approx e^{i\omega x_{k+1/2}}\left[f_k S_0 + \frac{f_{k+1} - f_k}{h_k}\left(S_1 + \frac{h_k}{2} S_0\right)\right] =$$

$$= e^{i\omega x_{k+1/2}}\left(\frac{f_k + f_{k+1}}{2} S_0 + \frac{f_{k+1} - f_k}{h_k} S_1\right) =$$

$$= e^{i\omega x_{k+1/2}}\left[f_k\left(\frac{S_0}{2} - \frac{S_1}{h_k}\right) + f_{k+1}\left(\frac{S_0}{2} + \frac{S_1}{h_k}\right)\right].$$

(П6.20)

Наконец, привлекая (П6.18), (П6.19), окончательно получаем

$$i_k = \frac{h_k}{2}(A_k f_k + B_k f_{k+1}) e^{i\omega x_{k+1/2}}$$

$$A_k = \frac{\sin p_k}{p_k} - i\frac{\sin p_k - p_k \cos p_k}{p_k^2}, \quad B_k = \frac{\sin p_k}{p_k} + i\frac{\sin p_k - p_k \cos p_k}{p_k^2} = \overline{A}_k. \tag{П6.21}$$

Замечание. При формальном предельном переходе $\omega \to 0$ формула (П6.21) переходит в обычную формулу трапеций $\frac{h_k}{2}(f_k + f_{k+1})$ для интеграла $\int\limits_a^b f(x)\,dx$ по локальному интервалу $[x_k, x_{k+1}]$. ▲

Для вычисления интеграла (П6.2) по конечному отрезку $[a, b]$ надо просуммировать (П6.21) по всем элементарным отрезкам:

$$I = \sum_{k=0}^{K-1} i_k. \tag{П6.22}$$

Если $h_k = h = \text{const}$, ($p_k = p = \text{const}$) то соответствующую квадратурную формулу можно записать в виде

$$I \approx \frac{h}{2}\left(A \sum_{k=0}^{K-1} f_k e^{i\omega x_{k+1/2}} + B \sum_{k=0}^{K-1} f_{k+1} e^{i\omega x_{k+1/2}}\right). \tag{П6.23}$$

Приведем вытекающие из (П6.21) конкретные расчетные формулы для интегралов (П6.1) (порознь) по локальному интервалу $[x_k, x_{k+1}]$:

$$I(\cos) = \int\limits_{x_k}^{x_{k+1}} f(x)\cos\omega x\,dx \approx$$

$$\approx \frac{h_k}{2}\left(\frac{\sin p_k}{p_k}\cos\omega x_{k+1/2} + \frac{\sin p_k - p_k\cos p_k}{p_k^2}\sin\omega x_{k+1/2}\right) f_k +$$

$$+ \frac{h_k}{2}\left(\frac{\sin p_k}{p_k}\cos\omega x_{k+1/2} - \frac{\sin p_k - p_k\cos p_k}{p_k^2}\sin\omega x_{k+1/2}\right) f_{k+1} =$$

$$= h_k\left(\frac{f_k + f_{k+1}}{2}\frac{\sin p_k}{p_k}\cos\omega x_{k+1/2} + \right.$$

$$\left. + \frac{f_k - f_{k+1}}{2}\frac{\sin p_k - p_k\cos p_k}{p_k^2}\sin\omega x_{k+1/2}\right); \tag{П6.24}$$

$$I(\sin)=\int\limits_{x_k}^{x_{k+1}} f(x)\sin\omega x\,dx \approx h_k\left(\frac{f_k+f_{k+1}}{2}\,\frac{\sin p_k}{p_k}\sin\omega x_{k+1/2}-\right.$$

$$\left.-\frac{f_k-f_{k+1}}{2}\,\frac{\sin p_k-p_k\cos p_k}{p_k^2}\cos\omega x_{k+1/2}\right). \qquad (\text{П6.25})$$

Чтобы оценить погрешность метода (П6.22), найдем сначала локальную ошибку на интервале $[x_k, x_{k+1}]$, которая возникает при использовании для вычисления на этом отрезке приближенной формулы (П6.21). Для этого вычислим интеграл от остаточного члена интерполяционного полинома (П6.16), которым мы заменили функцию $f(x)$:

$$r_k=\frac{\widetilde{f}_k''}{2}\int\limits_{x_k}^{x_{k+1}}(x-x_k)(x-x_{k+1})e^{i\omega x}\,dx=$$

$$=\frac{1}{2}\widetilde{f}_k''\,e^{i\omega x_{k+1/2}}\int\limits_{-h_k/2}^{h_k/2}\left(\xi^2-\frac{h_k^2}{4}\right)e^{i\omega\xi}\,d\xi=\frac{1}{2}\left(S_2-\frac{h_k^2}{4}S_0\right)\widetilde{f}_k''\,e^{i\omega x_{k+1/2}}.$$

Здесь используется обозначение S_0, уже введенное ранее. Что касается S_2, то

$$S_2=\int\limits_{-h_k/2}^{h_k/2}\xi^2 e^{i\omega\xi}\,d\xi=\frac{1}{i\omega}\,\xi^2 e^{i\omega\xi}\Big|_{-h_k/2}^{h_k/2}-\frac{2}{i\omega}S_1=$$

$$=\frac{h_k^3}{4}\,\frac{\sin p_k}{p_k}-\frac{h_k^3}{2}\,\frac{(\sin p_k-p_k\cos p_k)}{p_k^3}. \qquad (\text{П6.26})$$

Учитывая (П6.26) и (П6.18), получаем:

$$r_k=\frac{h_k^3}{4}\,\widetilde{f}_k''\,e^{i\omega k_{k+1/2}}\,\frac{\sin p_k-p_k\cos p_k}{p_k^3}.$$

При $\omega h_k\sim 1$ ($p_k\sim 1/2$), как следует из (П6.11),

$$\frac{\sin p_k-p_k\cos p_k}{p_k^3}\sim\frac{1}{3}.$$

Поэтому локальная ошибка в этом случае оценивается сверху точно так же, как ошибка формулы трапеций для $\int\limits_{x_k}^{x_{k+1}} f\,dx$:

$$|r_k| \lesssim \frac{h_k^3}{12} M_2^{(k)}, \quad M_2^{(k)} = \max_{[x_k, x_{k+1}]} |f''|.$$

Соответственно погрешность квадратурной формулы (П6.22) на конечном интервале $[a, b]$

$$|R| \lesssim \frac{h^2}{12} M_2 (b-a), \quad M_2 = \max_{[a,b]} |f''|, \quad h = \max_k h_k. \tag{П6.27}$$

Таким образом, использование метода (П6.22) для вычисления интегралов (П6.2) при $\omega h_k \sim 1$ обеспечивает второй порядок точности.

В) В заключение без выкладок приведем для интегралов (П6.2) квадратурную формулу Филона типа Симпсона и соответствующую оценку погрешности. После замены функции $f(x)$ на отрезке $[x_k, x_{k+1}]$ интерполяционным полиномом второй степени (см. Лекцию 5) получаем следующую приближенную формулу для интеграла по этому отрезку:

$$i_k \approx \frac{h_k}{2} (a_k f_k + 4 b_k f_{k+1/2} + c_k f_{k+1}) e^{i\omega x_{k+1/2}}, \tag{П6.28}$$

где

$$a_k = \frac{\sin p_k}{p_k} - 2\frac{\sin p_k - p_k \cos p_k}{p_k^3} - i\frac{\sin p_k - p_k \cos p_k}{p_k^2}, \; b_k = \frac{\sin p_k - p_k \cos p_k}{p_k^3},$$

$c_k = \overline{a}_k$ — комплексно-сопряженное к a_k.

Соответственно для интеграла по исходному отрезку $[a, b]$ следует локальные результаты (П6.28) просуммировать по всем отрезкам, как в (П6.22).

При формальном переходе $\omega \to 0$ имеем $a_k, b_k, c_k \to 1/3$ и (П6.28) переходит в формулу Симпсона $\frac{h_k}{6}(f_k + 4f_{k+1/2} + f_{k+1})$ для обычного интеграла $\int\limits_a^b f(x)\,dx$ на интервале $[x_k, x_{k+1}]$.

Локальная погрешность формулы (П6.28) получается интегрированием по $[x_k, x_{k+1}]$ остаточного члена интерполяционного полинома

второй степени:

$$r_k = \frac{1}{6}\widetilde{f}''' \int\limits_{x_k}^{x_{k+1}} (x - x_k)(x - x_{k+1/2})(x - x_{k+1})e^{i\omega x} =$$

$$= \frac{1}{6}e^{i\omega x_{k+1/2}}\widetilde{f}''' \int\limits_{-h_k/2}^{h_k/2} \left(\xi^2 - \frac{h_k^2}{4}\right)\xi e^{i\omega\xi}d\xi = \frac{1}{6}e^{i\omega x_{k+1/2}}\widetilde{f}'''\left(S_3 - \frac{h_k^2}{4}S_1\right). \tag{П6.29}$$

Здесь

$$S_3 = \int\limits_{-h_k/2}^{h_k/2} \xi^3 e^{i\omega\xi}d\xi = -i\frac{h_k^4}{8}\frac{p_k^3\cos p_k - 3p_k^2\sin p_k + 6\sin p_k - 6p_k\cos p_k}{p_k^4}. \tag{П6.30}$$

Учитывая (П6.19), в конечном счете получаем:

$$r_k = \frac{i}{24}h_k^4\widetilde{f}'''\rho e^{i\omega x_{k+1/2}},$$

где

$$\rho = \frac{-p_k^2\sin p_k + 3\sin p_k - 3p_k\cos p_k}{p_k^4}.$$

Оценка показывает, что при $p_k \lesssim 1$ $\rho \lesssim 0.1$. Поэтому при $p_k \lesssim 1$ $(\omega h_k \lesssim 2)$

$$|r_k| \leqslant \frac{h_k^4}{240}M_3^{(k)},$$

с $M_3^{(k)} = \max\limits_{[x_k, x_{k+1}]} |f'''|$.

Суммируя левые и правые части последнего неравенства, приходим к оценке глобальной погрешности метода Филона–Симпсона:

$$|R| \leqslant \sum_k |r_k| \leqslant \frac{h^3 M_3}{240}(b - a),$$

где $M_3 = \max\limits_{[a,b]} |f'''|$, а $h = \max\limits_k h_k$.

Замечание. При формальном переходе $\omega \to 0$ имеем: $\rho \to 0$, в чем можно убедиться последовательным применением правила Лопиталя. Следовательно, локальная погрешность становится величиной

пятого порядка малости, как и положено для обычного метода Симпсона применительно к интегралу $\int_a^b f(x)\,dx$. ▲

Пример. В таблице П6.1 приведены результаты вычисления интегралов

$$I_1 = \int_0^1 \cos x \cos \omega x\, dx = 7.730739 \cdot 10^{-4},$$

$$I_2 = \int_0^1 \cos x \sin \omega x\, dx = 9.537752 \cdot 10^{-3}$$

с $\omega = 160$.

Таблица 6.1.

		Метод прямоуг-ков		Филона-А		Филона-Б	
K	ωh	Δ_1	Δ_2	Δ_1	Δ_2	Δ_1	Δ_2
5	32					$5.856 \cdot 10^{-6}$	$2.118 \cdot 10^{-5}$
						0.75%	0.22%
10	16					$3.582 \cdot 10^{-8}$	$8.184 \cdot 10^{-7}$
						0.0046%	0.0086%
20	8					$9.154 \cdot 10^{-8}$	$9.118 \cdot 10^{-7}$
						0.018%	0.096%
40	4			$6.138 \cdot 10^{-5}$	$1.450 \cdot 10^{-5}$	$-5.397 \cdot 10^{-8}$	$7.145 \cdot 10^{-7}$
				7.94%	0.15%	0.0070%	0.0075%
				$7.365 \cdot 10^{-3}$	$1.740 \cdot 10^{-3}$	$1.036 \cdot 10^{-3}$	$1.372 \cdot 10^{-2}$
80	2	$1.32 \cdot 10^{-4}$	$1.80 \cdot 10^{-3}$	$1.148 \cdot 10^{-5}$	$2.549 \cdot 10^{-6}$	$-1.067 \cdot 10^{-8}$	$1.334 \cdot 10^{-7}$
		17.1%	18.9%	1.48%	0.027%	0.0014%	0.0014%
				$5.510 \cdot 10^{-3}$	$1.223 \cdot 10^{-3}$	$8.194 \cdot 10^{-4}$	$1.024 \cdot 10^{-2}$
160	1	$3.034 \cdot 10^{-5}$	$4.099 \cdot 10^{-4}$	$2.719 \cdot 10^{-6}$	$5.974 \cdot 10^{-7}$	$-2.565 \cdot 10^{-9}$	$3.158 \cdot 10^{-8}$
		3.92%	4.29%	0.35%	0.0063%	0.00033%	0.00033%
				$5.221 \cdot 10^{-3}$	$1.147 \cdot 10^{-3}$	$8.017 \cdot 10^{-4}$	$9.70 \cdot 10^{-3}$
320	0.5	$7.434 \cdot 10^{-6}$	$1.002 \cdot 10^{-4}$	$6.712 \cdot 10^{-7}$	$1.471 \cdot 10^{-7}$	$6.524 \cdot 10^{-10}$	$7.790 \cdot 10^{-9}$
		0.96%	1.05%	0.087%	0.0015%	0.00008%	0.00008%
				$5.155 \cdot 10^{-3}$	$1.130 \cdot 10^{-3}$	$8.017 \cdot 10^{-4}$	$9.573 \cdot 10^{-3}$
640	0.25	$1.849 \cdot 10^{-6}$	$2.492 \cdot 10^{-5}$	$1.672 \cdot 10^{-7}$	$3.663 \cdot 10^{-5}$		
		0.24%	0.26%	0.022%	0.00038%		
				$5.137 \cdot 10^{-3}$	$1.125 \cdot 10^{-3}$		

В левом столбце таблицы указано количество элементарных шагов интегрирования, для которых получены результаты, приведенные

в данной строке; во втором столбце — параметр ωh. Далее, в 3-м–4-м столбцах — результаты для первого и второго интегралов, полученные обычным методом прямоугольников, в 5-м–6-м столбцах — методом Филона-А; в 7-м–8-м столбцах — методом Филона-Б. В качестве результатов приведены: абсолютная погрешность, относительная (в процентах). Третье число, если оно присутствует, означает: для метода Филона-А — отношение модуля абсолютной погрешности к величине $\omega h^2/12$ (см. формулу (П6.15)), для метода Филона-Б — соответствующее отношение к величине $h^2/12$ (см. (П6.27)).

Дадим некоторые комментарии к приведенным результатам. Во-первых, очевидно, что формулы Филона обеспечивают существенно лучшую точность при довольно грубом шаге интегрирования сравнительно с обычными квадратурами (которые здесь представлены методом прямоугольников). Подтверждаются выводы о порядке точности использованных формул Филона. При фиксированном значении ω ошибка метода Филона-А пропорциональна h^2 (см. формулу (П6.15)), что проявляется в табличных данных; фактически же, с учетом того, что $\omega h \sim 1$, ошибка для этого метода на порядок больше, нежели для метода Филона-Б. Обращает на себя внимание высокая точность результатов, полученных по последнему методу при очень грубом шаге ($\omega h = 32$).

Наконец, подтверждается вывод о том, что ошибка метода Филона-А пропорциональна интегралу $i\int\limits_a^b f' e^{i\omega x}\,dx$. В нашем случае

$$\Delta_1 \sim I_4 = \int\limits_0^1 \sin x \sin \omega x\,dx = 5.135855 \cdot 10^{-3},$$

$$\Delta_2 \sim I_3 = \int\limits_0^1 \sin x \cos \omega x\,dx = 1.094389 \cdot 10^{-3},$$

что и подтверждается (асимптотически, с уменьшением ωh) третьим числом, приведенным в клетках столбцов таблицы, относящихся к методу Филона-А. (Третье число в клетках, относящихся к методу Филона-Б, свидетельствует, что погрешность этого метода асимптотически пропорциональна $\int\limits_a^b f'' e^{i\omega x}\,dx$. В данном случае это означает, что $\Delta_1 \sim I_1 \cdot h^2$ и $\Delta_2 \sim I_2 \cdot h^2$).

Формулы Филона для интегралов от быстроосциллирующих функций рассматриваются также в [1, с. 405–406; 2, с. 113–115; 9, с. 103–105].

СПИСОК ЛИТЕРАТУРЫ *)

1. *Амосов А. А., Дубинский Ю. А., Копченова Н. А.* Вычислительные методы для инженеров. — М.: Высшая школа, 1994.

2. *Бахвалов Н. С., Жидков Н. П., Кобельков Г. М.* Численные методы. — М.: Наука, 1987.

3. *Березин Н. С., Жидков Н. П.* Методы вычислений. Т. 1, 2. — М.: Физматгиз, 1960.

4. *Годунов С. К., Рябенький В. С.* Разностные схемы. — М.: Наука, 1977.

5. *Дьяченко В. Ф.* Основные понятия вычислительной математики. — М.: Наука, 1977.

6. *Воеводин В. В.* Вычислительные основы линейной алгебры. — М.: Наука, 1977.

7. *Волков Е. А.* Численные методы. — М.: Наука, 1987.

8. *Икрамов Х. Д.* Численные методы линейной алгебры. — М.: Знание (сер. «Математика и кибернетика»), 1987. №4.

9. *Калиткин Н. Н.* Численные методы. — М.: Наука, 1978.

10. *Марчук Г. И.* Методы вычислительной математики. — М.: Наука, 1977.

11. *Самарский А. А.* Введение в численные методы. — М.: Наука, 1982.

12. *Самарский А. А., Гулин А. В.* Численные методы. — М.: Наука, 1989.

13. *Самарский А. А., Николаев Е. С.* Методы решения сеточных уравнений. — М.: Наука, 1978.

14. *Тихонов А. Н., Костомаров Д. П.* Вводные лекции по прикладной математике. — М.: Наука, 1984.

15. *Турчак Л. И.* Основы численных методов. — М.: Наука, 1987.

16. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений. — М.: Мир, 1980.

17. *Рябенький В. С.* Введение в вычислительную математику. — М.: Наука, 1994.

18. *Федоренко Р. П.* Введение в вычислительную физику. — М.: Издательство МФТИ, 1994.

19. *Петров И. Б., Лобанов А. И.* Лекции по вычислительной математике. — М.: Интернет-университет информационных технологий, 2006.

20. *Вержбицкий В. М.* Основы численных методов. — М.: Высшая школа, 2002.

21. *Бахвалов Н. С., Лапин А. В., Чижонков Е. В.* Численные методы в задачах и упражнениях. — М.: Высшая школа, 2000.

22. *Костомаров Д. П., Фаворский А. П.* Вводные лекции по численным методам. — М.: Логос, 2004.

---

*) За время, прошедшее со дня выхода в свет первого издания книги, некоторые учебные пособия из основного списка литературы были переизданы. Однако я счел возможным оставить здесь ссылки на старые издания, принимая во внимание, что они выходили существенно бóльшими тиражами и с бóльшей вероятностью доступны в фондах институтских библиотек и читальных залов. Впрочем, в порядке информации, список переизданий некоторых из пособий приведен после списка литературы.

СПИСОК ДОПОЛНИТЕЛЬНОЙ ЛИТЕРАТУРЫ

23. *Ортега Д., Рейнболдт В.* Итерационные методы решения нелинейных систем уравнений со многими неизвестными. — М.: Мир, 1975.

24. *Уилкинсон Дж. Х.* Алгебраическая проблема собственных значений. — М.: Наука, 1970.

25. *Тихонов А. Н., Гончарский А. В., Степанов В. В.* Численные методы решения некорректных задач. — М.: Наука, 1990.

26. *Белоцерковский О. М.* Численное моделирование в механике сплошных сред. — М.: Наука, 1984.

27. Численные методы в динамике жидкости /Пер. с англ., под ред. О. М. Белоцерковского и В. П. Шидловского. — М.: Мир, 1981.

28. *Магомедов К. М., Холодов А. С.* Сеточно-характеристические численные методы. — М.: Наука, 1988.

29. *Пасконов В. М., Полежаев В. И., Чудов Л. А.* Численное моделирование процессов тепло- и массообмена. — М.: Наука, 1984.

30. *Попов Ю. П., Самарский А. А.* Разностные схемы газовой динамики. — М.: 1980.

31. *Днестровский Ю. Н., Костомаров Д. П.* Математическое моделирование плазмы. — М.: Наука, 1982.

32. Численное моделирование в физике плазмы. — М.: Наука, 1977.

33. Современные проблемы вычислительной физики и вычислительной математики. — М.: Наука, 1982.

34. *Поттер Д.* Вычислительные методы в физике. — М.: Мир, 1975.

35. *Ильин В. П.* Численные методы решения задач электрофизики. — М.: Наука, 1985.

36. *Хайрер Э., Нерсетт С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. — М.: Мир, 1990.

37. *Каханер Д., Моулер К., Нэш С.* Численные методы и програмное обеспечение — М.: Мир, 1998.

СПИСОК ПЕРЕИЗДАНИЙ

1. *Федоренко Р. П.* Введение в вычислительную физику. — М.: Интеллект, 2008.

2. *Калиткин Н. Н.* Численные методы. — М.: БХВ, 2011.

3. *Рябенький В. С.* Введение в вычислительную математику. — М.: Физматлит, 2000.

4. *Бахвалов Н. С., Жидков Н. П., Кобельков Г. М.* Численные методы. — М.: Физматлит, Невский диалект, Лаборатория базовых знаний, 2001.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Учебное издание

Косарев Виталий Иванович

12 лекций по вычислительной математике (вводный курс)

Издание третье, исправленное и дополненное

Редактор *А. К. Розанов*

Набор и верстка выполнены в издательстве «Физматкнига»
Операторы верстки *И. А. Розанов, К. В. Чувилин*
Художник *М. В. Ивановский*

Издательство «Физматкнига».
141700, Московская область, г. Долгопрудный, ул. Первомайская, д. 11-а.
Тел. (495) 971-26-04.

Подписано в печать 23.09.2013. Формат 60×88/16. Бумага офсетная.
Печать офсетная. Усл. печ. л. 14,7. Уч.-изд. л. 15,0. Тираж 1000 экз. Заказ № 4009.

Отпечатано в ППП «Типография «Наука».
121099, Москва, Шубинский пер., 6

12 ЛЕКЦИЙ ПО ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКЕ

9 785891 552142

www.fizmatkniga.ru