Глава 1

Решение систем линейных алгебраических уравнений

В данной главе мы будем рассматривать следующую задачу — требуется решить систему линейных алгебраических уравнений (в дальнейшем СЛАУ), задаваемых в виде матричного уравнения:

$$Ax = F,$$

где $A = (a_{ij})$ — матрица размерности $n \times n$, причем $\det A \neq 0$; $x = (x_1,\ x_2,\ \ldots,\ x_n)^T$ — вектор-столбец неизвестных, $F = (f_1,\ f_2,\ \ldots,\ f_n)^T$ — заранее заданный вектор-столбец правой части.

Далее мы будем предполагать, что у всех рассматриваемых задач решение существует и единственно (в данном случае это следует из того, что определитель матрицы A не равен нулю).

Рассмотрим несколько методов решения СЛАУ. Методы решения делятся на **прямые**, где мы получаем x за конечное число действий, и **итерационные** — получаем x как предел некоторой сходящейся последовательности $\{x^k\} :\ x = \lim\limits_{k\to\infty} x^k$.

В случае прямых методов получается точное (до погрешности аппаратуры) решение. В итерационных методах вводится точность решения $\varepsilon > 0$. Если $|x^k - x| < \varepsilon$ (x — точное решение) — заканчиваем процесс вычислений, иначе — вычисляем очередное x^{k+1} и т. д.

К прямым методам относятся формулы Крамера, метод Гаусса. К итерационным методам — метод Зейделя, метод верхней релаксации и т.д.

Основным показателем при оценке эффективности метода является его **сложность**. В прямых методах это чаще всего количество арифметических операций, необходимых для вычисления x, а в итерационных — количество итераций, необходимых для достижения заданной точности ε (его записывают как $k_0(\varepsilon)$). К примеру, метод Крамера (модифицированный) имеет сложность $O(n^4)$, а метод Гаусса — $O(n^3)$.

Выбор метода, в основном, зависит от размерности матрицы. При решении систем большой размерности чаще используют итерационные методы.

1.1 Прямые методы решения СЛАУ. Метод квадратного корня

В этой главе мы рассмотрим несколько прямых методов для решения системы

$$Ax = F. \tag{1.1}$$

В самом общем случае используется **метод Гаусса**[1], в котором мы будем выделять **прямой ход** — приведение матрицы коэффициентов к верхнетреугольному виду (мы думаем, что все читатели знают,

[1] Впервые описан К. Гауссом в 1849г., правда, без обратного хода. Этот же метод часто называют методом Гаусса-Остроградского.

как это делается), и **обратный ход** — приведение верхнетреугольной матрицы к диагональной и нахождение самого решения.

Однако существует целый класс методов, использующих специфику конкретной матрицы — например, **метод прогонки**.

Метод прогонки имеет сложность $O(n)$, но он применим только к трехдиагональным матрицам.

Вычисление элементов обратной матрицы

Чтобы решить систему (1.1), можно вычислить A^{-1}, тогда решение будет представимо в виде

$$x = A^{-1}F.$$

Для определения элементов матрицы A^{-1} рассмотрим уравнение $AA^{-1} = E$, где матрица A — задана, а требуется найти A^{-1}.

Обозначим за a_{ij} элементы матрицы A, а за z_{ij} — элементы матрицы A^{-1}. Умножая строки матрицы A на столбцы A^{-1} (по определению произведения матриц), получим n^2 алгебраических уравнений:

$$\sum_{l=1}^{n} a_{il} z_{lj} = \delta_{ij}, \quad i,\ j = \overline{1,\ n}.$$

Эти уравнения можно объединить в группы с фиксированным индексом j, то есть у нас получится по n уравнений для определения каждого столбца матрицы A^{-1}.

Это будут уравнения вида $Az^j = e^j$, где z^j — j-й столбец матрицы A^{-1}, а e^j — j-й столбец матрицы E. Получаем n систем по n уравнений в каждой. Далее используем метод Гаусса решения систем уравнений, при этом сложность данного метода будет равна $O(n^4)$.

Кроме того, можно применить прямой ход метода Гаусса сразу ко всем системам (т. к. матрица A — общая для всех систем), и сложность метода станет $\approx \frac{n^3}{3}$.

Наложим ограничения на матрицу A из уравнения (1.1).

Определение. Матрица A называется **положительно определенной**, если выполнено одно из условий (они эквивалентны друг другу):

1. все главные миноры больше нуля;
2. все собственные числа матрицы A (далее обозначаются как $\lambda(\mathbf{A})$) положительны;
3. $\forall\, x \neq 0 \quad \langle Ax,\, x\rangle > 0.$

Примечание. *Из пункта 1 (а также из 3) определения следует необходимое условие положительной определенности — все диагональные элементы положительны.*

Достаточным условием положительной определенности является **условие диагонального преобладания**: если элементы $a_{ii} > \sum\limits_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$, то матрица $A = (a_{ij})$ положительно определена.

Метод квадратного корня

Будем рассматривать симметричные (т. е. $A = A^T$) положительно определенные матрицы. Для таких матриц справедливо представление (разложение Холецкого):

$$A = LL^T, \tag{1.2}$$

где L — нижнетреугольная матрица:

$$L = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ & \ddots & & \vdots \\ & l_{ij} & \ddots & 0 \\ & & & l_{nn} \end{pmatrix}.$$

Рассмотрим задачу (1.1). Подставив разложение (1.2) для A в уравнение (1.1), получим:

$$LL^T x = F.$$

Обозначим $L^T x = y$, тогда получим систему уравнений относительно новой переменной y :

$$Ly = F.$$

Учитывая то, что матрица L нижнетреугольная, для решения этой системы требуется только обратный ход метода Гаусса (решение уравнений построчно сверху вниз). Аналогично, так как $y = L^T x$, то нужно решать только СЛАУ с верхней треугольной матрицей. Поэтому основной вклад в сложность даст вычисление элементов матрицы L — и общая сложность метода будет приближенно равна $\frac{n^3}{6}$.

Теперь покажем, что матрицу L можно найти. Обозначим $L = (l_{ij})$, $L^T = (\overline{l_{ij}})$ — из определения транспонированной матрицы вытекает, что $\overline{l_{ij}} = l_{ji}$. Найдем элементы этих матриц, для чего распишем разложение (1.2):

$$LL^T = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ & \ddots & & \vdots \\ & l_{ij} & \ddots & 0 \\ & & & l_{nn} \end{pmatrix} \begin{pmatrix} \overline{l_{11}} & & & \\ 0 & \ddots & \overline{l_{ij}} & \\ \vdots & & \ddots & \\ 0 & \dots & 0 & \overline{l_{nn}} \end{pmatrix} = (a_{ij}).$$

По правилу умножения двух матриц получаем:

$$a_{ij} = \sum_{l=1}^{n} l_{il}\overline{l_{lj}} = \{\text{учитывая треугольный вид матриц}\} = \sum_{l=1}^{\min(i,\,j)} l_{il}\overline{l_{lj}}.$$

Найдем элементы l_{ij} матрицы L. Сначала вычислим l_{11} :

$$a_{11} = l_{11}\overline{l_{11}} \quad \Longrightarrow \quad a_{11} = l_{11}^2 \quad \Longrightarrow \quad l_{11} = \sqrt{a_{11}} \qquad (a_{11} > 0).$$

Получив l_{11}, найдем элементы первого столбца матрицы L:

$$a_{i1} = l_{i1}\overline{l_{11}} = l_{i1}l_{11} \quad \Longrightarrow \quad l_{i1} = \frac{a_{i1}}{l_{11}}, \qquad i = \overline{2, n}.$$

Рассмотрим второй столбец матрицы A :

$$a_{i2} = l_{i1}\overline{l_{12}} + l_{i2}\overline{l_{22}} \quad \Longrightarrow \quad l_{i1}l_{21} + l_{i2}l_{22} = a_{i2}, \qquad i = \overline{2, n}. \tag{1.3}$$

Пусть $i = 2$, тогда $l_{21}^2 + l_{22}^2 = a_{22} \quad \Longrightarrow \quad l_{22}^2 = a_{22} - l_{21}^2$ — а l_{21} уже найдено.
Извлекая корень, получим:

$$l_{22} = \sqrt{a_{22} - l_{21}^2} = \sqrt{a_{22} - \frac{a_{21}^2}{l_{11}^2}} \quad \Longrightarrow \quad l_{22} = \sqrt{a_{22} - \frac{a_{21}^2}{a_{11}}} = \sqrt{\frac{a_{11}a_{22} - a_{21}^2}{a_{11}}}.$$

Корректность вытекает из того, что $a_{11} > 0$, $\quad a_{11}a_{22} - a_{12}^2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0$ по условию положительной определенности матрицы A.

Подставив в (1.3) только что вычисленное значение l_{22}, получим представление для остальных элементов второго столбца:

$$l_{i2} = \frac{a_{i2} - l_{i1}l_{21}}{l_{22}}, \qquad i = \overline{3, n}.$$

Рассмотрим произвольный (k-й) столбец матрицы A:

$$l_{i1}\overline{l_{1k}} + l_{i2}\overline{l_{2k}} + \ldots + l_{i(k-1)}\overline{l_{(k-1)k}} + l_{ik}\overline{l_{kk}} = a_{ik} \iff$$

$$\iff l_{i1}l_{k1} + l_{i2}l_{k2} + \ldots + l_{i(k-1)}l_{k(k-1)} + l_{ik}l_{kk} = a_{ik}, \qquad i = \overline{k, n}. \tag{1.4}$$

Пусть $i = k$, тогда из (1.4) получаем:

$$l_{k1}^2 + l_{k2}^2 + \ldots + l_{k(k-1)}^2 + l_{kk}^2 = a_{kk}.$$

Переносим все слагаемые, кроме l_{kk}^2, в правую часть и извлекаем корень:

$$l_{kk} = \sqrt{a_{kk} - l_{k1}^2 - \ldots - l_{k(k-1)}^2}. \tag{1.5}$$

Можно показать, что под корнем получается отношение положительного минора k-го порядка, к положительной сумме:

$$l_{kk} = \sqrt{\frac{|\ldots| > 0}{(\ldots) > 0}}.$$

Подставив (1.5) в (1.4), получим:

$$l_{ik} = \frac{a_{ik} - l_{i1}l_{k1} - \ldots - l_{i(k-1)}l_{k(k-1)}}{l_{kk}}, \qquad i = \overline{k+1,\, n}.$$

Мы полностью определили элементы матрицы L, доказав тем самым работоспособность нашего метода. Заметим, что мы использовали симметричность матрицы A, когда брали в уравнениях для нахождения l_{ij} только те ее элементы, которые лежат на главной диагонали или ниже ее.

Модифицированный метод квадратного корня

Теперь несколько модифицируем метод квадратного корня, а точнее, распространим его на более широкий класс задач. Итак, мы снова рассматриваем уравнение

$$Ax = f \quad \text{при условии, что } \det A \neq 0. \tag{1.6}$$

Предположим, что матрица A симметрична ($A = A^T$) — положительной определенности требовать не будем. Покажем, что для таких матриц справедливо представление

$$A = LDL^T, \tag{1.7}$$

где L — нижнетреугольная матрица с единицами на главной диагонали, а D — диагональная матрица:

$$L = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ & \ddots & & \vdots \\ & l_{ij} & \ddots & 0 \\ & & & 1 \end{pmatrix}; \qquad D = \begin{pmatrix} d_{11} & 0 & \ldots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & d_{nn} \end{pmatrix}.$$

Для наглядности (1.7) можно представить так ($L^T = (\overline{l_{ij}})$):

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ & \ddots & & \vdots \\ & l_{ij} & \ddots & 0 \\ & & & 1 \end{pmatrix} \begin{pmatrix} d_{11} & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & d_{nn} \end{pmatrix} \begin{pmatrix} 1 & & & \\ 0 & \ddots & \overline{l_{ij}} & \\ \vdots & & \ddots & \\ 0 & \dots & 0 & 1 \end{pmatrix} = A.$$

Обозначим произведение матриц DL^T за новую матрицу $B = (b_{ij})$ и заметим, что $\overline{l_{ij}} = l_{ji}$.

Теперь найдем элементы матриц L и D. Согласно определению произведения матриц, элемент a_{ik} при $i \geqslant k$ определяется так:

$$a_{ik} = l_{i1}b_{1k} + l_{i2}b_{2k} + \ldots + l_{i(k-1)}b_{(k-1)k} + l_{ik}b_{kk}$$

— мы отбросили нулевые последние слагаемые, обусловленные видом матрицы L.

А элемент a_{ik} при $i = k$ определяется так:

$$a_{kk} = l_{k1}b_{1k} + l_{k2}b_{2k} + \ldots + l_{k(k-1)}b_{(k-1)k} + l_{kk}b_{kk}.$$

Теперь воспользуемся тем, что

$$b_{jk} = d_{jj}\overline{l_{jk}} = d_{jj}l_{kj},$$

и перепишем два последних уравнения:

$$a_{ik} = l_{i1}(d_{11}l_{k1}) + l_{i2}(d_{22}l_{k2}) + \ldots + l_{i(k-1)}(d_{(k-1)(k-1)}l_{k(k-1)}) + l_{ik}b_{kk}; \tag{1.8}$$

$$a_{kk} = l_{k1}(d_{11}l_{k1}) + l_{k2}(d_{22}l_{k2}) + \ldots + l_{k(k-1)}(d_{(k-1)(k-1)}l_{k(k-1)}) + l_{kk}b_{kk}. \tag{1.9}$$

При $k = 1$ (1.9) выглядит так:

$$l_{11}^2 d_{11} = a_{11}.$$

Зная, что $l_{11} = 1$, получим формулу для d_{11} :

$$d_{11} = a_{11}.$$

Теперь рассмотрим случай $k = 1,\ i = \overline{2,\ n}$. Из (1.8) получим:

$$l_{i1}(d_{11}l_{11}) = a_{i1} \quad \Longrightarrow \quad \{\mathsf{l_{11} = 1}\} \quad \Longrightarrow \quad l_{i1} = \frac{a_{i1}}{d_{11}}.$$

Мы получили уравнения, определяющие первый столбец матрицы L.

Аналогично, при $k = 2$ (1.9) и (1.8) перепишутся так:

$$l_{21}(d_{11}l_{21}) + l_{22}(d_{22}l_{22}) = a_{22} \quad \Longrightarrow \quad \{\mathsf{l_{22} = 1}\} \quad \Longrightarrow \quad d_{22} = a_{22} - l_{21}^2 d_{11};$$

$$l_{i1}(d_{11}l_{21}) + l_{i2}(d_{22}l_{22}) = a_{i2} \quad \Longrightarrow \quad l_{i2} = \frac{a_{i2} - l_{i1}d_{11}l_{21}}{d_{22}}.$$

И так далее. Вот как будут выглядеть общие формулы для элементов матриц D и L :

$$d_{kk} = a_{kk} - (l_{k1}^2 d_{11} + \ldots + l_{k(k-1)}^2 d_{(k-1)(k-1)});$$

$$l_{ik} = \frac{(a_{ik} - l_{i1}d_{11}l_{k1} - \ldots - l_{i(k-1)}d_{(k-1)(k-1)}l_{k(k-1)})}{d_{kk}}.$$

Заметим, что формулы корректны — элементы d_{kk} не могут быть равны нулю, так как в этом случае определитель матрицы A обращался бы в ноль.

Итак, разложение (1.7) получено. Подставим его в исходное уравнение (1.6):

$$LDL^Tx = f \quad \Longrightarrow \quad \{\text{обозначим } \mathsf{DL^Tx = y}\} \quad \Longrightarrow \quad Ly = f.$$

Мы получили уравнение относительно вектора y. В левой части стоит нижнетреугольная матрица L, что позволяет нам сильно сократить вычисления, применяя только обратный ход метода Гаусса.

После нахождения y вспомним, как мы его определяли:

$$DL^Tx = y \quad \Longrightarrow \quad \{\text{обозначим } \mathsf{L^Tx = z}\} \quad \Longrightarrow \quad Dz = y$$

— это система уравнений, определяемая диагональной матрицей D. Решая ее, получим z.

Теперь у нас есть простая система для нахождения x :

$$L^Tx = z.$$

Наличие в левой части верхнетреугольной матрицы L, опять же, сильно упрощает вычисления.

Сложив количество операций, необходимых для получения разложения (1.7) и для поиска x, получим сложность данного метода, равную $\approx \frac{n^3}{6}$.

Итерационные методы решения СЛАУ

Это следующая группа методов поиска решений систем линейных уравнений, особенно эффективная при решении больших систем (с числом неизвестных порядка тысячи и более).

В общем случае сначала задаются некоторым вектором x^0, называемым **начальным приближением**. От него строится последовательность $x^1,\ x^2,\ \ldots,\ x^k$ и так далее, где число k называют номером итерации. В общем случае $(k+1)$-е приближение зависит от всех предыдущих:

$$x^{k+1} = F_{k+1}(x^0,\ x^1,\ \ldots,\ x^k).$$

От последовательности, естественно, ожидается сходимость к вектору x, который будет являться решением исходной системы.

Определение. Итерационный метод называется $m-$**шаговым**, если каждое последующее итерационное приближение строится лишь по m предыдущим:

$$x^{k+1} = F_{k+1}(x^{k-m+1},\ \ldots,\ x^{k-1},\ x^k)$$

— на практике наиболее часто используется $m = 1$ и $m = 2$.

Определение. Если F_{k+1} — линейная функция, то соответствующий итерационный метод называется **линейным**.

1.2 Линейные одношаговые итерационные методы

Согласно определению, общее выражение для x^{k+1} в линейных одношаговых итерационных методах таково:

$$x^{k+1} = S^{k+1}x^k + \psi_{k+1}, \tag{1.10}$$

где S^{k+1} — некоторая матрица, а ψ_{k+1} — некоторый вектор.

Логично потребовать, чтобы вектор $x = A^{-1}f$ (то есть искомое решение) при подстановке вместо x^{k+1} и x^k обращал (1.10) в тождество:

$$A^{-1}f = S^{k+1}A^{-1}f + \psi_{k+1},$$

верное для всех k.

Преобразовав, получим:

$$(A^{-1} - S^{k+1}A^{-1})f = \psi_{k+1} \quad \Longrightarrow \quad \psi_{k+1} = Q^{k+1}f,$$

где $Q^{k+1} = A^{-1} - S^{k+1}A^{-1}$. Потребуем, чтобы эта матрица была обратима, и формулу для Q^{k+1} домножим на A :

$$S^{k+1} = E - Q^{k+1}A.$$

Это выражение подставим в (1.10):

$$x^{k+1} = x^k - Q^{k+1}Ax^k + Q^{k+1}f.$$

Домножим это выражение на $(Q^{k+1})^{-1}$ и перенесем слагаемые:

$$\left(Q^{k+1}\right)^{-1}(x^{k+1} - x^k) + Ax^k = f.$$

Теперь домножим и разделим первое слагаемое на некоторую константу τ_{k+1} :

$$\left(Q^{k+1}\right)^{-1}\tau_{k+1}\frac{x^{k+1} - x^k}{\tau_{k+1}} + Ax^k = f.$$

Обозначив $B_{k+1} = \left(Q^{k+1}\right)^{-1}\tau_{k+1}$, получим так называемую **каноническую форму** записи одношагового итерационного метода:

$$B_{k+1}\frac{x^{k+1} - x^k}{\tau_{k+1}} + Ax^k = f. \tag{1.11}$$

Примечание. *Мы могли бы и не вводить константу τ_{k+1}, однако впоследствии будет видно, что с матрицей B_{k+1} и константой τ_{k+1} удобнее работать по отдельности.*

Примечание. *Далее мы иногда будем отождествлять итерационный метод и его каноническую форму — так что не пугайтесь.*

Примечание. *Матрица B обязана быть обратимой — иначе вывод канонической формы записи итерационного метода будет некорректен.*

Определение. Если $B_{k+1} = E$, то соответствующий метод называется **явным** (находим x^{k+1}, не решая системы уравнений), в противном случае — **неявным**.

Определение. Если $B_{k+1} = B$ и $\tau_{k+1} = \tau$, то метод называется **стационарным**, в противном случае — **нестационарным**.

Определение. Вектор $z^k = x^k - x$ (отклонение от точного решения) называется **погрешностью на k-й итерации**.

Определение. Итерационный метод называется **сходящимся**, если $\|z^k\| \overset{k\to\infty}{\longrightarrow} 0$ для некоторой выбранной нормы $||\cdot||$.

Понятно, что предел может не быть достигнут за конечное число шагов. В таком случае мы задаемся некоторой точностью ε — достаточно малым числом ($\sim 10^{-3},\ 10^{-5}$) — и будем останавливать итерационный процесс при выполнении неравенства

$$||x^k - x|| \leqslant \frac{||x^0 - x||}{\frac{1}{\varepsilon}}, \quad \text{или} \quad ||x^k - x|| \leqslant \varepsilon||x^0 - x||$$

для некоторого k. В этом случае мы будем говорить, что решение получено с точностью ε.

Кроме того, естественно потребовать, что если это неравенство выполняется для $k_0 = k_0(\varepsilon)$, то оно должно выполняться и для любого $k > k_0(\varepsilon)$.

Число $k_0(\varepsilon)$ называется **минимальным числом итераций, необходимых для достижении заданной точности** ε. Понятно, что чем меньше $k_0(\varepsilon)$, тем лучше метод.

Примеры одношаговых итерационных методов

Для иллюстрации нижеприведенных методов сначала более детально распишем уравнение $Ax = f$:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= f_1; \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= f_2; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= f_n. \end{cases}$$

Расставляя итерационные индексы над компонентами вектора X, можно получить различные итерационные процессы.

Метод Якоби

Расставим индекс следующего приближения $(k+1)$ над диагональными элементами, а индекс k — над остальными:

$$\begin{cases} a_{11}x_1^{k+1} + a_{12}x_2^k \quad + \ldots + a_{1n}x_n^k \quad = f_1; \\ a_{21}x_1^k \quad + a_{22}x_2^{k+1} + \ldots + a_{2n}x_n^k \quad = f_2; \\ \qquad\qquad \ldots\ldots\ldots \\ a_{n1}x_1^k \quad + a_{n2}x_2^k \quad + \ldots + a_{nn}x_n^{k+1} = f_n. \end{cases}$$

Из этой системы легко получить общее уравнение для x_i^{k+1} :

$$x_i^{k+1} = \frac{f_i - \sum\limits_{j=1}^{i-1} a_{ij}x_j^k - \sum\limits_{j=i+1}^{n} a_{ij}x_j^k}{a_{ii}}, \qquad i = \overline{1,\, n}. \tag{1.12}$$

Чтобы от этой расчетной формулы перейти к канонической форме записи метода, представим матрицу A в виде суммы трех матриц — нижнетреугольной, диагональной, и верхнетреугольной:

$$A = A_1 + D + A_2,$$

$$A_1 = \begin{pmatrix} 0 & & \\ \ddots & 0 & \\ a_{ij} & \ddots & \\ & & 0 \end{pmatrix}; \quad D = \begin{pmatrix} a_{11} & & & \\ & \ddots & 0 & \\ & 0 & \ddots & \\ & & & a_{nn} \end{pmatrix}; \quad A_2 = \begin{pmatrix} 0 & & & \\ & \ddots & a_{ij} & \\ & 0 & \ddots & \\ & & & 0 \end{pmatrix}. \tag{1.13}$$

Из (1.12) легко вывести, что

$$A_1x^k + Dx^{k+1} + A_2x^k = f.$$

Для приведения к канонической форме (1.11) прибавим и вычтем Dx^k :

$$D(x^{k+1} - x^k) + Ax^k = f.$$

Отсюда получаем формальное определение метода Якоби через матрицу B_{k+1} и константу τ_{k+1} :

$$\begin{cases} B_{k+1} &= D; \\ \tau_{k+1} &= 1. \end{cases}$$

Легко видеть, что он неявный и стационарный.

Метод Зейделя[2]

Пусть теперь итерационные индексы $(k+1)$ стоят не только на диагонали, но и под ней:

$$\begin{cases} a_{11}x_1^{k+1} + a_{12}x_2^k + \ldots + a_{1n}x_n^k = f_1; \\ a_{21}x_1^{k+1} + a_{22}x_2^{k+1} + \ldots + a_{2n}x_n^k = f_2; \\ \ldots\ldots\ldots\ldots \\ a_{n1}x_1^{k+1} + a_{n2}x_2^{k+1} + \ldots + a_{nn}x_n^{k+1} = f_n. \end{cases}$$

Относительно x^{k+1} эта система решается так: сначала из первого уравнения находится x_1^{k+1}, потом из второго — x_2^{k+1} и т.д.

Для записи метода в канонической форме представим матрицу A в виде суммы матриц A_1, D, A_2 (как и в методе Якоби):

$$A = A_1 + D + A_2,$$

где A_1 — нижнетреугольная, D — диагональная, а A_2 — верхнетреугольная матрица. Тогда написанная выше итерационная схема будет выглядеть так:

$$A_1x^{k+1} + Dx^{k+1} + A_2x^k = f.$$

Преобразуя это уравнение, приведем его к каноническому виду:

$$\begin{aligned} & (A_1 + D)x^{k+1} + A_2x^k = f \quad \Longleftrightarrow \\ \Longleftrightarrow \quad & (A_1 + D)(x^{k+1} - x^k) + Ax^k = f. \end{aligned}$$

Отсюда получаем формальное определение **метода Зейделя**:

$$\begin{cases} B_{k+1} = A_1 + D; \\ \tau_{k+1} = 1. \end{cases}$$

Как несложно заметить, он является неявным и стационарным.

Метод релаксации

Формально определим метод следующим образом:

$$\begin{cases} B_{k+1} = D + \omega A_1; \\ \tau_{k+1} = \omega, \end{cases}$$

где ω — некий числовой параметр. Нетрудно заметить, что метод Зейделя является частным случаем этого метода при $\omega = 1$.

Кроме того, метод релаксации является стационарным и неявным.

Метод простой итерации

И здесь никакой подробной системы не будет:

$$\begin{cases} B_{k+1} = E; \\ \tau_{k+1} = \tau. \end{cases}$$

Подставив эти равенства в (1.11), получим каноническую форму метода простой итерации:

$$\frac{x^{k+1} - x^k}{\tau} + Ax^k = f.$$

На этот раз метод является явным и стационарным.

[2]Предложен Л. Зейделем в 1874 г.

Метод Ричардсона

В данном случае матрица B_{k+1} — единичная, а константа τ_{k+1} определяется по некоторым расчетным формулам. Выражение для τ_{k+1} приводить не будем, запишем только каноническую форму записи метода:

$$\frac{x^{k+1}-x^k}{\tau_{k+1}} + Ax^k = f.$$

1.3 Сходимость одношаговых стационарных методов

Теорема 1.1 (Достаточное условие сходимости одношагового стационарного итерационного метода)**.** *Пусть матрица A — симметрическая положительно определенная, B — положительно определена, и $\tau > 0$, тогда итерационный процесс $B\dfrac{(x^{k+1}-x^k)}{\tau} + Ax^k = f$ сходится для любого начального приближения x^0, если $B - \dfrac{\tau}{2}A > 0$.*

Доказательство этой теоремы было дано в курсе «Введение в численные методы», кроме того, его можно найти в [1].

Утверждение 1.1. *Пусть матрица A — симметрическая положительно определенная, и выполнено условие диагонального преобладания: $a_{ii} > \sum\limits_{\substack{j=1\\ i\neq j}}^{n} |a_{ij}|$. Тогда метод Якоби является сходящимся для любого начального приближения.*

Доказательство. Так как матрица A — положительно определена, $\forall\, x \neq 0 \quad \langle Ax,\, x\rangle > 0$.

Распишем скалярное произведение поподробнее:

$$\begin{aligned}
\langle Ax,\, x\rangle &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j \leqslant \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}||x_ix_j| \leqslant \left\{|x_ix_j| \leqslant \frac{x_i^2+x_j^2}{2}\right\} \leqslant \\
&\leqslant \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|x_i^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ji}|x_j^2 = \{i \leftrightarrow j\} = \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|x_i^2 = \sum_{i=1}^{n}x_i^2\Big(|a_{ii}| + \sum_{\substack{j=1\\ i\neq j}}^{n}|a_{ij}|\Big) < \\
&< \{\text{так как матрица } A \text{ — матрица с диагональным преобладанием}\} < \\
&< 2\sum_{i=1}^{n}x_i^2|a_{ii}| = 2\sum_{i=1}^{n}|a_{ii}|x_ix_i = 2\,\langle Dx,\, x\rangle\,.
\end{aligned}$$

Таким образом, получаем:

$$0 < \langle Ax,\, x\rangle < 2\,\langle Dx,\, x\rangle\,.$$

Откуда следует, что

$$2\,\langle Dx,\, x\rangle - \langle Ax,\, x\rangle > 0 \quad \Longrightarrow \quad 2D - A > 0.$$

То есть матрица $(2D - A)$ положительно определена. В этом случае выполняются все условия теоремы 1.1 при $B = D$ и $\tau = 1$ — это и дает сходимость метода Якоби. □

Утверждение 1.2 (Сходимость метода релаксации)**.** *Пусть $B = D + \omega A_1$, $\tau = \omega$, A — симметрическая положительно определенная матрица и $\omega \in [0;\, 2]$, тогда метод релаксации является сходящимся для любого начального приближения.*

Доказательство. Проверим выполнение достаточного условия сходимости стационарного одношагового итерационного метода (теорема 1.1):

$$B - \frac{\tau}{2}A > 0,$$

подставив формулы для B и τ из условия теоремы. Таким образом, для доказательства теоремы достаточно показать, что

$$2D + 2\omega A_1 - \omega A > 0 \tag{1.14}$$

Представим матрицу A в виде (1.13), тогда:

$$\begin{aligned}\langle Ax,\, x\rangle &= \langle A_1 x,\, x\rangle + \langle Dx,\, x\rangle + \left\langle x,\, A_2^T x\right\rangle = \ \{\left\langle \mathsf{x},\, \mathsf{A_2^T x}\right\rangle = \langle \mathsf{A_1 x},\, \mathsf{x}\rangle\} = \\ &= \langle Dx,\, x\rangle + 2\left\langle A_1 x,\, x\right\rangle.\end{aligned}$$

Теперь перепишем (1.14):

$$\langle (2D + 2\omega A_1 - \omega A)x,\, x\rangle > 0.$$

Подставим сюда разложение матрицы A и раскроем скобки:

$$2\left\langle Dx,\, x\right\rangle + 2\omega\left\langle A_1 x,\, x\right\rangle - 2\omega\left\langle A_1 x,\, x\right\rangle - \omega\left\langle Dx,\, x\right\rangle > 0.$$

В результате получаем:

$$(2-\omega)\left\langle Dx,\, x\right\rangle > 0,$$

где $\langle Dx,\, x\rangle > 0$ в силу того, что $\langle Ax,\, x\rangle > 0$ и примечания к определению положительно определенной матрицы, а множитель $(2-\omega)$ больше нуля в силу условия теоремы. □

Утверждение 1.3 (необходимое и достаточное условие сходимости метода простой итерации)**.** *Пусть матрица A — симметрическая положительно определенная, а параметр τ больше нуля. Тогда для сходимости метода простой итерации необходимо и достаточно, чтобы $\tau < \dfrac{2}{\lambda_{max}(A)}$, где $\lambda(A)$ — некоторое собственное значение матрицы A.*[3]

Доказательство. **Достаточность.** Пусть, согласно требованию теоремы,

$$\tau < \frac{2}{\lambda_{max}(A)}.$$

Тогда это верно для любого собственного значения матрицы A :

$$\tau < \frac{2}{\lambda(A)} \quad \Longleftrightarrow \quad 1 - \frac{\tau}{2}\lambda(A) > 0 \quad \Longleftrightarrow \quad \lambda(E - \frac{\tau}{2}A) > 0.$$

Это значит, что все собственные значения матрицы $E - \frac{\tau}{2}A$ положительны, то есть она положительно определена. Так как $B = E$, то по теореме 1.1 метод простой итерации является сходящимся.

Необходимость. Возьмем в качестве начального приближения $x^0 = x + \mu$, где μ — собственный вектор A :

$$A\mu = \lambda_{max}(A)\mu.$$

Рассмотрим наш итерационный процесс:

$$\frac{x^{k+1} - x^k}{\tau} + Ax^k = f. \tag{1.15}$$

[3]Далее такое обозначение будет использовано для собственных значений других матриц; иногда (например, в неравенствах) оно может обозначать весь спектр.

Заменим x^k на погрешность решения z^k : $z^k = x^k - x$, где x — точное решение. Тогда для погрешности получим такое выражение:

$$\frac{z^{k+1} - z^k}{\tau} + Az^k = 0 \quad \Longrightarrow \quad z^{k+1} = (E - \tau A)z^k \quad \forall k.$$

Выразим z^k через предыдущие значения:

$$z^k = (E - \tau A)z^{k-1} = \ldots = (E - \tau A)^k z^0.$$

Заметим, что $z^0 = \mu$:

$$(E - \tau A)^k \mu = (E - \tau A)^{k-1}(\mu - \tau A\mu) = (E - \tau A)^{k-1}(1 - \tau\lambda_{max}(A))\mu = \ldots = (1 - \tau\lambda_{max}(A))^k \mu.$$

Таким образом:

$$z^k = (1 - \tau\lambda_{max}(A))^k \mu.$$

Посчитаем норму z^k :

$$||z^k|| = |1 - \tau\lambda_{max}(A)|^k ||\mu||.$$

Так как итерационный процесс сходится ($||z^k|| \overset{k\to\infty}{\longrightarrow} 0$), то $|1-\tau\lambda_{max}(A)| < 1$. Раскрыв знак модуля, получим:

$$\begin{cases} 1 - \tau\lambda_{max}(A) & < & 1; \\ 1 - \tau\lambda_{max}(A) & > & -1. \end{cases} \quad \Longrightarrow \quad 0 < \tau < \frac{2}{\lambda_{max}(A)}.$$

Утверждение доказано. □

Сходимость стационарных методов

Теперь установим необходимое и достаточное условие сходимости для стационарных одношаговых итерационных методов. Для этого перейдем в итерационном процессе

$$B\frac{x^{k+1} - x^k}{\tau} + Ax^k = f \tag{1.16}$$

к погрешности $z^k = x^k - x$:

$$B\frac{z^{k+1} - z^k}{\tau} + Az^k = 0.$$

Отсюда получим формулу $z^{k+1} = (E - \tau B^{-1}A)z^k$, и обозначим $S = E - \tau B^{-1}A$. Матрица S называется **матрицей перехода**. Теперь выражение для z^{k+1} выглядит так:

$$z^{k+1} = Sz^k. \tag{1.17}$$

Теорема 1.2 (Критерий сходимости одношагового стационарного итерационного метода)**.** *Итерационный метод (1.16) сходится для любого начального приближения x^0 тогда и только тогда, когда для всех собственных значений $\lambda(S)$ выполнено $|\lambda(S)| < 1$.*

Доказательство. **Необходимость.** Пусть μ — собственный вектор S, соответствующий собственному значению $\lambda(S)$. Рассмотрим вектор начального приближения $x^0 = \mu + x$, тогда $z^0 = x^0 - x = \mu$.

Все погрешности связаны соотношением (1.17). Выразим из него z^k

$$z^k = Sz^{k-1} = S^2 z^{k-2} = \ldots = S^k z^0,$$

где $z^0 = \mu$, тогда

$$z^k = S^k\mu = S^{k-1}\lambda(S)\mu = \ldots = \lambda^k(S)\mu.$$

В силу сходимости верно, что

$$||z^k|| \overset{k\to\infty}{\longrightarrow} 0 \quad \Longrightarrow \quad ||\lambda^k(S)\mu|| \overset{k\to\infty}{\longrightarrow} 0 \quad \Longrightarrow \quad |\lambda(S)|^k||\mu|| \overset{k\to\infty}{\longrightarrow} 0.$$

Отсюда следует, что $|\lambda(S)| < 1$.

Достаточность. Пусть матрица S — матрица простой структуры, причем $|\lambda(S)| < 1$. Из этого следует, что существуют собственные вектора, образующие ортонормированный базис $\{\xi_i\}_{i=1}^n$.

Пусть $z^0 = \sum\limits_{i=1}^{n} c_i\xi_i$ — разложение z^0 по базису. Тогда получим следующее выражение для погрешности:

$$z^k = S^k z^0 = S^k \sum_{i=1}^{n} c_i\xi_i = S^{k-1} \sum_{i=1}^{n} c_i S\xi_i = S^{k-1} \sum_{i=1}^{n} c_i\lambda_i\xi_i = \ldots = \sum_{i=1}^{n} c_i\lambda_i^k\xi_i.$$

Рассмотрим норму погрешности:

$$||z^k|| = \left|\left|\sum_{i=1}^{n} c_i\lambda_i^k\xi_i\right|\right| \leqslant \sum_{i=1}^{n} |c_i| \cdot |\lambda_i|^k||\xi_i|| \leqslant \{\rho = \max_{\mathsf{i}} |\lambda_{\mathsf{i}}(\mathsf{S})| < 1\} \leqslant \rho^k \sum_{i=1}^{n} |c_i| \cdot ||\xi_i|| \overset{k\to\infty}{\longrightarrow} 0,$$

в силу того что $\sum\limits_{i=1}^{n} |c_i| \cdot ||\xi_i||$ ограничено. □

Примечание. *1. При доказательстве достаточности предполагалась, что матрица S — матрица простой структуры. Теорема верна и без этого предположения, которое сделали для упрощения доказательства.*

2. На первый взгляд кажется, что область применения этого утверждения широка, но на практике найти весь спектр матрицы перехода непросто (иногда сложнее, чем решить систему прямым методом).

Пусть для погрешности итерационного метода справедливо неравенство:

$$||x^k - x|| \leqslant q^k||x^0 - x||, \quad \text{где} \quad q \in (0,1). \tag{1.18}$$

Определение. Итерационный метод, погрешность итерационного приближения которого удовлетворяет (1.18), **сходится со скоростью геометрической прогрессии со знаменателем** q.

Потребуем, взяв $\varepsilon > 0$, чтобы $q^k < \varepsilon$. Тогда для погрешности итерационного метода будет выполнена оценка:

$$||x^k - x|| < \varepsilon||x^0 - x||.$$

А это означает, что к k-й итерации погрешность начального приближения уменьшится в $\dfrac{1}{\varepsilon}$ раз. Таким образом, число итераций, необходимых для достижения требуемой точности ε, будет

$$k > k_0(\varepsilon) = \left\lceil \frac{\ln\frac{1}{\varepsilon}}{\ln\frac{1}{q}} \right\rceil.$$

Определение. Число $\ln\frac{1}{q}$ называется **скоростью сходимости** итерационного метода.

Примечание. *Вообще говоря, число q из вышеприведенных неравенств определяется неоднозначно. Для формальности можно считать, что это минимальное из всех q, удовлетворяющих (1.18).*

1.4 Оценка погрешности одношаговых стационарных методов

Далее мы будем активно использовать матричные неравенства, например $A \geqslant B$. Оно означает, что для всех x $\langle Ax, x\rangle \geqslant \langle Bx, x\rangle$.

Введем норму вектора x, порожденную симметричной положительно определенной матрицей A:

$$||x||_A = \sqrt{\langle Ax, x\rangle}.$$

Теорема 1.3 (Оценка погрешности стационарных одношаговых итерационных методов)**.** *Пусть матрицы A, B симметричны и положительно определены, и существуют такие положительные константы γ_1, γ_2, что*

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B. \tag{1.19}$$

Тогда итерационный метод, задаваемый уравнением

$$B\frac{x^{k+1}-x^k}{\tau} + Ax^k = f, \quad \text{где } \tau = \frac{2}{\gamma_1+\gamma_2}, \tag{1.20}$$

сходится для любого начального приближения x^0 со скоростью геометрической прогрессии:

$$||x^k - x||_A \leqslant q^k ||x^0 - x||_A, \quad \text{где } q = \frac{1-\xi}{1+\xi},\ \xi = \frac{\gamma_1}{\gamma_2}.$$

Доказательство. Перейдем от приближений x^k к погрешности на k-й итерации: $z^k = x^k - x$. Как уже было показано, уравнение (1.20) можно переписать так:

$$B\frac{z^{k+1}-z^k}{\tau} + Az^k = 0,$$

и из него следует, что

$$z^{k+1} = Sz^k, \quad S = E - \tau B^{-1}A. \tag{1.21}$$

Теперь установим справедливость такого неравенства:

$$||z^{k+1}||_A \leqslant q||z^k||_A. \tag{1.22}$$

Известно, что для матрицы $A = A^T > 0$ существует такая матрица, обозначаемая $A^{\frac{1}{2}}$, что $(A^{\frac{1}{2}})^2 = A$, причем $A^{\frac{1}{2}} = (A^{\frac{1}{2}})^T > 0$.

Домножим (1.21) слева на эту матрицу:

$$A^{\frac{1}{2}}z^{k+1} = A^{\frac{1}{2}}Sz^k \quad \Longleftrightarrow \quad A^{\frac{1}{2}}z^{k+1} = A^{\frac{1}{2}}SA^{-\frac{1}{2}}A^{\frac{1}{2}}z^k.$$

Обозначив $\omega^k = A^{\frac{1}{2}}z^k$ и $\overline{S} = A^{\frac{1}{2}}SA^{-\frac{1}{2}}$, получим, что

$$\omega^{k+1} = \overline{S}\omega^k.$$

Заметим, что $\overline{S} = A^{\frac{1}{2}}SA^{-\frac{1}{2}} = A^{\frac{1}{2}}(E - \tau B^{-1}A)A^{-\frac{1}{2}} = E - \tau A^{\frac{1}{2}}B^{-1}A^{\frac{1}{2}}$.

Обозначив $C = A^{\frac{1}{2}}B^{-1}A^{\frac{1}{2}}$, получим, что

$$\overline{S} = E - \tau C. \tag{1.23}$$

Легко проверить, что матрицы C и $\overline{S}$ будут симметричны, а C — еще и положительно определена (это понадобится нам чуть позже).

Теперь преобразуем $||z^k||_A$:

$$||z^k||_A = \sqrt{\langle Az^k, z^k\rangle} = \sqrt{\left\langle A^{\frac{1}{2}}A^{\frac{1}{2}}z^k, z^k\right\rangle} = \sqrt{\left\langle A^{\frac{1}{2}}z^k, A^{\frac{1}{2}}z^k\right\rangle} = \sqrt{\langle \omega^k, \omega^k\rangle} = ||\omega^k||.$$

Таким образом, от доказательства неравенства (1.22) можно перейти к доказательству того, что

$$||\omega^{k+1}|| \leqslant q||\omega^k||. \tag{1.24}$$

Преобразуем ω^{k+1} :

$$||\omega^{k+1}||^2 = \left\langle \omega^{k+1}, \omega^{k+1} \right\rangle = \left\langle \overline{S}\omega^k, \overline{S}\omega^k \right\rangle = \left\langle \overline{S}^2\omega^k, \omega^k \right\rangle.$$

Теперь будем искать такие q, что $\overline{S}^2 \leqslant q^2 E$ — как видно из последних преобразований, этого будет достаточно для доказательства (1.24).

Итак, предположим, что $\overline{S}^2 \leqslant q^2 E$. Тогда, по свойству симметричных матриц, и в силу того, что C положительно определена, это эквивалентно тому, что

$$-qE \leqslant \overline{S} \leqslant qE \iff \{\text{подставим (1.23)}\} \iff -qE \leqslant E - \tau C \leqslant qE \iff$$

$$\iff \begin{cases} \tau C & \leqslant & (1+q)E; \\ \tau C & \geqslant & (1-q)E. \end{cases} \iff \{\text{домножим на } \mathsf{C}^{-1}\} \iff \begin{cases} \tau E & \leqslant & (1+q)C^{-1}; \\ \tau E & \geqslant & (1-q)C^{-1}. \end{cases}$$

Из этой системы вытекает такое двойное неравенство:

$$\frac{1-q}{\tau}C^{-1} \leqslant E \leqslant \frac{1+q}{\tau}C^{-1}.$$

Из того, что $C = A^{\frac{1}{2}}B^{-1}A^{\frac{1}{2}}$, следует, что $C^{-1} = A^{-\frac{1}{2}}BA^{-\frac{1}{2}}$. Таким образом, предыдущее двойное неравенство эквивалентно такому:

$$\frac{1-q}{\tau}A^{-\frac{1}{2}}BA^{-\frac{1}{2}} \leqslant E \leqslant \frac{1+q}{\tau}A^{-\frac{1}{2}}BA^{-\frac{1}{2}}.$$

Умножив его слева и справа на $A^{\frac{1}{2}}$, получим

$$\frac{1-q}{\tau}B \leqslant A \leqslant \frac{1+q}{\tau}B. \tag{1.25}$$

Очень похоже на одно из условий теоремы. Покажем, что

$$\frac{1-q}{\tau} = \gamma_1, \quad \frac{1+q}{\tau} = \gamma_2. \tag{1.26}$$

Действительно:

$$\frac{1-q}{\tau} = \frac{1 - \frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1}}{\frac{2}{\gamma_2+\gamma_1}} = \frac{(\gamma_2+\gamma_1)-(\gamma_2-\gamma_1)}{2} = \gamma_1;$$

$$\frac{1+q}{\tau} = \frac{1 + \frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1}}{\frac{2}{\gamma_2+\gamma_1}} = \frac{(\gamma_2+\gamma_1)+(\gamma_2-\gamma_1)}{2} = \gamma_2.$$

Итак, мы показали, что неравенство $||z^{k+1}||_A \leqslant q||z^k||_A$ эквивалентно матричному неравенству $\gamma_1 B \leqslant A \leqslant \gamma_2 B$ из условия теоремы.

Таким образом, показано, что $||z^{k+1}||_A \leqslant q||z^k||_A$ для любого k. Тогда, переходя к более ранним членам последовательности, получим:

$$||z^k||_A \leqslant q||z^{k-1}||_A \leqslant \ldots \leqslant q^k||z^0||.$$

Из этого напрямую следует, что

$$||x^k - x||_A \leqslant q^k||x^0 - x||_A.$$

Теорема доказана. □

Замечание 1. В случае, когда ξ мало, можно получить такую оценку для скорости сходимости:

$$\ln \frac{1}{q} = \ln \left(\frac{1+\xi}{1-\xi} \right) = \ln \left(1 + \frac{2\xi}{1-\xi} \right) \approx 2\xi.$$

Замечание 2. Из теоремы 1.3 следует, что выбор чисел γ_1, γ_2 напрямую влияет на скорость сходимости. Для выяснения их возможных значений рассмотрим произвольный собственный вектор μ матрицы $B^{-1}A$:

$$B^{-1}A\mu = \lambda(B^{-1}A)\mu.$$

Это эквивалентно тому, что $A\mu = \lambda(B^{-1}A)B\mu$ (мы просто домножили на B). Как известно, неравенство

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B \tag{1.27}$$

означает, что $\gamma_1 \langle Bx, x\rangle \leqslant \langle Ax, x\rangle \leqslant \gamma_2 \langle Bx, x\rangle$ для любого x. Положим $x = \mu$ и преобразуем это двойное неравенство:

$$\begin{aligned}
&\gamma_1 \langle B\mu, \mu\rangle \leqslant \langle A\mu, \mu\rangle \leqslant \gamma_2 \langle B\mu, \mu\rangle \iff \\
&\iff \gamma_1 \langle B\mu, \mu\rangle \leqslant \lambda(B^{-1}A) \langle B\mu, \mu\rangle \leqslant \gamma_2 \langle B\mu, \mu\rangle \implies \\
&\implies \gamma_1 \leqslant \lambda(B^{-1}A) \leqslant \gamma_2.
\end{aligned}$$

Так как собственный вектор мы выбирали произвольно, то получаем, что

$$\gamma_1 \leqslant \lambda_{min}(B^{-1}A), \quad \gamma_2 \geqslant \lambda_{max}(B^{-1}A).$$

Таким образом, наиболее точными константами, с которыми выполняется неравенство (1.27), являются константы

$$\gamma_1 = \lambda_{min}(B^{-1}A), \quad \gamma_2 = \lambda_{max}(B^{-1}A).$$

В этом случае параметр

$$\tau_{\text{опт}} = \frac{2}{\lambda_{min}(B^{-1}A) + \lambda_{max}(B^{-1}A)}$$

называется **оптимальным итерационным параметром**. Кроме того, так как мы берем $q = \dfrac{1 - \frac{\gamma_1}{\gamma_2}}{1 + \frac{\gamma_1}{\gamma_2}}$, то наилучшим вариантом будет как раз $\lambda_{min}(B^{-1}A) = \lambda_{max}(B^{-1}A)$.

Все, мы закончили с главной теоремой этого параграфа, пора построить пример использования наших методов.

Примечание. *Выполнение неравенства (1.27) при $\gamma_1 = \lambda_{min}(B^{-1}A)$, $\gamma_2 = \lambda_{max}(B^{-1}A)$ следует из возможности построения в данном линейном пространстве базиса из собственных векторов.*

Модельная задача. Сравнение скорости сходимости различных итерационных методов

Рассмотрим краевую задачу:

$$\begin{cases} -u''(x) = f(x), \ 0 < x < 1; \\ u(0) = u(1) = 0. \end{cases}$$

Найдем ее решение, используя численные методы. Для этого сначала разделим отрезок $[0; 1]$ на N равных промежутков длины $h = \frac{1}{N}$, обозначив границы отрезков как x_i :

$$0 = x_0 < x_1 < x_2 < \ldots < x_N = 1; \quad h = x_{i+1} - x_i.$$

Воспользуемся тем, что $u''(x_i)$ можно приблизить 2-й разностной производной:

$$u''(x_i) \approx \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2}$$

Пусть $y(x)$ — приближение нашей функции. Потребуем, чтобы $y_i \stackrel{\text{def}}{=} y(x_i) = u(x_i)$. Тогда, зная, что $u''(x_i) = f(x_i)$, и используя разностное приближение, получим такую систему уравнений относительно значений y в узлах сетки:

$$-\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} = f(x_i), \quad i = \overline{1,\, N-1}.$$

Обозначим $f_i = f(x_i)$. Из краевых условий можно получить, что $y_0 = y_N = 0$. Тогда получим такую систему:

$$\begin{cases} -\dfrac{y_{i+1} - 2y_i + y_{i-1}}{h^2} &= f_i, \quad i = \overline{1,\, N-1}; \\ y_0 = y_N &= 0. \end{cases} \tag{1.28}$$

Понятно, что решив ее, мы получим приближенные значения $u(x_i) = y_i$. Теперь обозначим

$$Y = (y_1,\, y_2,\, \ldots,\, y_{N-1})^T; \quad F = (f_1,\, f_2,\, \ldots,\, f_{N-1})^T.$$

Тогда систему (1.28), использовав значения y_0 и y_N, можно переписать в виде матричного уравнения

$$AY = F, \tag{1.29}$$

где $A = \begin{pmatrix} \frac{2}{h^2} & -\frac{1}{h^2} & & & \\ -\frac{1}{h^2} & \frac{2}{h^2} & \ddots & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \ddots & -\frac{1}{h^2} \\ & & & -\frac{1}{h^2} & \frac{2}{h^2} \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & \ddots & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix}.$

Легко видеть, что $A = A^T$. Покажем, что $A > 0$ — для этого обоснуем положительность всех собственных значений.

Утверждается, что уравнение для поиска собственных значений $A\xi = \lambda\xi$ в каком-то смысле эквивалентно задаче Штурма-Лиувилля:

$$\begin{cases} X'' + \lambda X = 0, \quad 0 < x < 1; \\ X(0) = X(1) = 0, \end{cases}$$

решением которой будут собственные функции $X_m(x) = \sin \pi m x$. Отсюда делается вывод, что для собственных значений матрицы A справедлива формула:

$$\lambda_m = \frac{4}{h^2} \left(\sin \frac{\pi m h}{2} \right)^2, \quad m = \overline{1,\, N-1}.$$

Таким образом, все λ_m положительны и матрица A положительно определена. Легко проверить, что справедлива следующая цепочка неравенств:

$$0 < \lambda_1 < \lambda_2 < \ldots < \lambda_{N-1},$$

причем $\lambda_1 = \lambda_{min} = \frac{4}{h^2}\left(\sin\frac{\pi h}{2}\right)^2$ и $\lambda_{N-1} = \lambda_{max} = \frac{4}{h^2}\left(\cos\frac{\pi h}{2}\right)^2$. ч Будем решать уравнение (1.29) методом простой итерации:

$$\frac{x^{k+1} - x^k}{\tau} + Ax^k = f.$$

Для него справедлива теорема 1.3 о скорости сходимости, так как $A = A^T > 0$ (по доказанному) и $B = B^{-1} = E$ — симметричная и положительно определенная матрица. Тогда, согласно замечанию 2 к теореме 1.3 мы можем взять $\gamma_1 = \lambda_{min}(A)$, $\gamma_2 = \lambda_{max}(A)$ и положить

$$\tau = \frac{2}{\lambda_{min}(A) + \lambda_{max}(A)} = \frac{2}{\frac{4}{h^2}(\sin^2(\frac{\pi h}{2}) + \cos^2(\frac{\pi h}{2}))} = \frac{h^2}{2}.$$

При этом для погрешности верна оценка

$$||x^k - x||_A \leqslant q^k ||x^0 - x||_A, \quad \text{где } q = \frac{1-\xi}{1+\xi},\ \xi = \frac{\lambda_{min}}{\lambda_{max}}.$$

Тогда, если в качестве условия выхода из итерационного процесса использовать неравенство

$$||x^k - x||_A \leqslant \varepsilon ||x^0 - x||_A,$$

то число итераций, необходимых для достижения точности ε, равно

$$k_0(\varepsilon) = \left\lceil \frac{\ln\frac{1}{\varepsilon}}{\ln\frac{1}{q}} \right\rceil.$$

В нашем случае ξ достаточно мало, поэтому верна такая оценка для $\ln\frac{1}{q}$:

$$\ln\frac{1}{q} \approx 2\xi = 2\frac{\lambda_{min}(A)}{\lambda_{max}(A)} = 2\,\mathrm{tg}^2\left(\frac{\pi h}{2}\right) \approx \frac{2\pi^2 h^2}{4} = \{\mathsf{h} = \frac{1}{\mathsf{N}}\} = \frac{\pi^2}{2N^2}.$$

Отсюда следует, что

$$k_0(\varepsilon) = \frac{2N^2}{\pi^2}\ln\frac{1}{\varepsilon} \approx \frac{N^2}{5}\ln\frac{1}{\varepsilon}.$$

К примеру, для $\varepsilon = 0,5 \cdot 10^{-4} \approx e^{-10}$ имеем:

$$k_0(\varepsilon) = 2N^2.$$

Таким образом, например, если разбить отрезок на 10 частей, получив систему уравнений с числом уравнений $N = 10$, требуется выполнить 200 итераций, в случае разбиения отрезка на 100 частей (соответственно $N = 100$ уравнений) уже требуется выполнить 20000 итераций для получения решения с заданной точностью ε.

Впоследствии мы увидим, что по сравнению с другими методами метод простой итерации является очень медленно сходящимся.

Впоследствии мы увидим, что по сравнению с другими методами метод простой итерации является очень медленно сходящимся (хотя это понятно и так, 20000 итераций — это же убить себя можно...).

1.5 Попеременно-треугольный итерационный метод

В этом пункте мы рассмотрим еще один итерационный метод решения СЛАУ. Этот метод работает быстрее метода простой итерации, но и параметры B и τ в нем выбираются не так тривиально.

Рассмотрим одношаговый итерационный метод:

$$B\frac{x^{k+1}-x^k}{\tau}+Ax^k=f.$$

Возьмем матрицу B как произведение верхнетреугольной и нижнетреугольной матриц особого вида:

$$B=(E+\omega R_1)(E+\omega R_2), \tag{1.30}$$

где матрицы

$$R_1=\begin{pmatrix} r_{11} & & & \\ & \ddots & 0 & \\ & r_{ij} & \ddots & \\ & & & r_{nn}\end{pmatrix}, \qquad R_2=\begin{pmatrix} r_{11} & & & \\ & \ddots & r_{ij} & \\ & 0 & \ddots & \\ & & & r_{nn}\end{pmatrix}$$

такие, что их сумма $R_1+R_2=A$, а их элементы на их диагонали равны половине соответствующих элементов A.

Теорема 1.4 (Достаточное условие сходимости ПТИМ)**.** *Пусть A — симметрическая положительно определенная матрица, а матрица B задается формулой (1.30), где $\omega>\frac{\tau}{4}$ (просто согласование ω с τ). Тогда соответствующий попеременно-треугольный итерационный метод сходится.*

Доказательство. Преобразуем матрицу B к виду, показывающему выполнение достаточного условия сходимости. Раскроем скобки в представлении (1.30)

$$\begin{aligned}B=(E+\omega R_1)(E+\omega R_2)=E+\omega(R_1+R_2)+\omega^2R_1R_2=\\ =\{\text{так как } \mathsf{R_1+R_2=A}\}=E+\omega A+\omega^2R_1R_2.\end{aligned}$$

Заметим, что в силу симметричности матрицы A, матрицы R_1 и R_2 связаны следующим образом: $R_1=R_2^T$. Вычтем и добавим ωA :

$$B=E-\omega A+\omega^2R_1R_2+2\omega A=(E-\omega R_1)(E-\omega R_2)+2\omega A.$$

Покажем, что выполняется достаточное условие сходимости из теоремы 1.1. Для этого распишем такое скалярное произведение:

$$\langle Bx,\,x\rangle=\langle(E-\omega R_1)(E-\omega R_2)x,\,x\rangle+2\omega\,\langle Ax,\,x\rangle\,.$$

Перенесем скобку $(E-\omega R_1)$ в правую часть скалярного произведения, а потом воспользуемся тем, что $(E-\omega R_1)^*=E-\omega R_1^T=E-\omega R_2$, получим:

$$\langle Bx,\,x\rangle=\langle(E-\omega R_2)x,\,(E-\omega R_2)x\rangle+2\omega\,\langle Ax,\,x\rangle=||(E-\omega R_2)x||^2+2\omega\,\langle Ax,\,x\rangle\,.$$

В силу неотрицательности нормы $\langle Bx,\,x\rangle\geqslant 2\omega\,\langle Ax,\,x\rangle$, что равносильно $B\geqslant 2\omega A$, а, так как в условии теоремы мы взяли $\omega>\frac{\tau}{4}$, то выполняется достаточное условие сходимости: $B>\frac{\tau}{2}A$. Теорема доказана. □

Теорема 1.5. *Пусть матрица A — симметрическая положительно определенная, и пусть существуют такие константы $\delta>0$ и $\Delta>0$, что*

$$A\geqslant\delta E,\quad \frac{\Delta}{4}A\geqslant R_1R_2\quad (R_1 \text{ и } R_2 \text{ — такие же, как в теореме 1.4}).$$

Тогда ПТИМ сходится при любом начальном приближении со скоростью геометрической прогрессии, причем при значениях параметров

$$\begin{cases} \tau = \dfrac{2}{\gamma_1 + \gamma_2}, \text{ где } \gamma_1 = \dfrac{\sqrt{\delta}}{2} \cdot \dfrac{\sqrt{\Delta\delta}}{\sqrt{\Delta} + \sqrt{\delta}}, \quad \gamma_2 = \dfrac{\sqrt{\delta\Delta}}{4}; \\ \omega = \dfrac{2}{\sqrt{\Delta\delta}}. \end{cases}$$

скорость сходимости наилучшая: $||x^{k+1} - x||_A \leqslant q||x^k - x||_A$, *где* $q = \dfrac{1 - \sqrt{\xi}}{1 + 3\sqrt{\xi}}, \quad \xi = \dfrac{\delta}{\Delta}$.

Доказательство. **(1)** Для корректности последующих формул сначала убедимся, что $\dfrac{\delta}{\Delta} \leqslant 1$.

Из условий теоремы, так как $A \geqslant \delta E$ и $\dfrac{\Delta}{4}A \geqslant R_1R_2$, следует выполнение неравенств:

$$\delta||x||^2 = \delta \langle x, x\rangle \leqslant \langle Ax, x\rangle, \qquad \langle R_1R_2x, x\rangle \leqslant \frac{\Delta}{4}\langle Ax, x\rangle. \tag{1.31}$$

Левую часть второго неравенства можно переписать в виде:

$$\langle R_1R_2x, x\rangle = \langle R_2x, R_2x\rangle = ||R_2x||^2. \tag{1.32}$$

Из (1.31) очевидно следует, что

$$\delta||x||^2 \leqslant \langle Ax, x\rangle = \frac{\langle Ax, x\rangle^2}{\langle Ax, x\rangle}.$$

Кроме того, в силу представления матрицы A

$$\langle Ax, x\rangle = \langle R_1x, x\rangle + \langle R_2x, x\rangle = \langle R_2x, x\rangle + \langle x, R_1^Tx\rangle = 2\langle R_2x, x\rangle,$$

поэтому

$$\frac{\langle Ax, x\rangle^2}{\langle Ax, x\rangle} = \frac{4\langle R_2x, x\rangle^2}{\langle Ax, x\rangle} \leqslant \{|\langle \mathsf{u}, \mathsf{v}\rangle| \leqslant ||\mathsf{u}|| \cdot ||\mathsf{v}||\} \leqslant \frac{4||R_2x||^2 \cdot ||x||^2}{\langle Ax, x\rangle}.$$

Распишем это с помощью (1.31) и (1.32):

$$\frac{4||R_2x||^2 \cdot ||x||^2}{\langle Ax, x\rangle} \leqslant \frac{4\Delta\langle Ax, x\rangle ||x||^2}{4\langle Ax, x\rangle} = \Delta||x||^2.$$

В итоге получаем, что $\delta||x||^2 \leqslant \Delta||x||^2$ для любого $x \neq 0$, или, что то же самое,

$$\delta \leqslant \Delta \implies \xi = \frac{\delta}{\Delta} \leqslant 1.$$

(2) Теперь фиксируем некоторое ω и посмотрим, что происходит. Из доказательства теоремы 1.4 следует, что $B \geqslant 2\omega A$, откуда

$$A \leqslant \frac{1}{2\omega}B.$$

Обозначим $\dfrac{1}{2\omega}$ некоторой константой γ_2, и перейдем к доказательству симметричного неравенства для γ_1.

Преобразуем условия нашей теоремы:

$$\begin{cases} A \geqslant \delta E; \\ \dfrac{\Delta}{4}A \geqslant R_1R_2. \end{cases} \implies \begin{cases} E \leqslant \dfrac{1}{\delta}A; \\ R_1R_2 \leqslant \dfrac{\Delta}{4}A. \end{cases}$$

Использовав эти неравенства в очевидном (хотя и ранее доказанном) равенстве

$$B = E + \omega A + \omega^2 R_1 R_2,$$

получим, что

$$B \leqslant \frac{1}{\delta} A + \omega A + \omega^2 \frac{\Delta}{4} A.$$

Отсюда легко получается выражение для A :

$$A \geqslant \left(\frac{1}{\delta} + \omega + \frac{\omega^2 \Delta}{4} \right)^{-1} B.$$

Обозначив $\gamma_1 = \left(\frac{1}{\delta} + \omega + \frac{\omega^2 \Delta}{4} \right)^{-1}$, получим, что выполняется условие теоремы 1.3, то есть

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B.$$

Кроме того, мы можем выбирать различные γ_1 и γ_2, варьируя параметр ω.

Теперь воспользуемся теоремой 1.3, и получим:

$$||x^k - x||_A \leqslant q ||x^{k-1} - x||_A,$$

где $q = \dfrac{1 - \frac{\gamma_1}{\gamma_2}}{1 + \frac{\gamma_1}{\gamma_2}} = 1 - \dfrac{2}{\frac{\gamma_2}{\gamma_1} + 1}$. Здесь γ_1 и γ_2 (а, значит, и q) неявно зависят от ω.

При уменьшении q скорость сходимости возрастает. Выберем ω так, чтобы q было минимально. Очевидно, для этого надо минимизировать отношение $\dfrac{\gamma_2}{\gamma_1}$. Обозначим $g = \dfrac{\gamma_2(\omega)}{\gamma_1(\omega)} \geqslant 1$, и получим для него такую формулу:

$$g = \frac{\frac{1}{\delta} + \omega + \omega^2 \frac{\Delta}{4}}{2\omega} = \frac{1}{2} + \frac{\omega \Delta}{8} + \frac{1}{2\omega\delta}.$$

Чтобы найти экстремум, возьмем производную от g по ω :

$$g' = \frac{\Delta}{8} - \frac{1}{2\delta\omega^2}.$$

Приравняв производную к нулю, получим точку минимума $\omega = \dfrac{2}{\sqrt{\delta\Delta}}$ (заметим, что вторая производная будет больше нуля). Теперь найдем значения γ_1 и γ_2 в этой точке.

$$\begin{cases} \gamma_1 &= \left. \dfrac{1}{\frac{1}{\delta} + \omega + \frac{\omega^2 \Delta}{4}} \right|_{\omega = \frac{2}{\sqrt{\delta\Delta}}} = \dfrac{1}{\frac{2}{\delta} + \frac{2}{\sqrt{\delta\Delta}}} = \dfrac{\sqrt{\delta}}{2} \cdot \dfrac{\sqrt{\delta\Delta}}{\sqrt{\Delta} + \sqrt{\delta}}; \\ \gamma_2 &= \dfrac{1}{2\omega} = \dfrac{\sqrt{\delta\Delta}}{4}. \end{cases}$$

Воспользовавшись найденными значениями γ_1 и γ_2, подсчитаем q :

$$q = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} = \frac{\frac{\sqrt{\delta\Delta}}{4} - \frac{\sqrt{\delta}}{2} \cdot \frac{\sqrt{\delta\Delta}}{\sqrt{\Delta} + \sqrt{\delta}}}{\frac{\sqrt{\delta\Delta}}{4} + \frac{\sqrt{\delta}}{2} \cdot \frac{\sqrt{\delta\Delta}}{\sqrt{\Delta} + \sqrt{\delta}}} =$$

$$= \frac{\sqrt{\Delta}\sqrt{\delta\Delta} + \sqrt{\delta}\sqrt{\delta\Delta} - 2\sqrt{\delta}\sqrt{\delta\Delta}}{\sqrt{\Delta}\sqrt{\delta\Delta} + \sqrt{\delta}\sqrt{\delta\Delta} + 2\sqrt{\delta}\sqrt{\delta\Delta}} = \frac{\sqrt{\Delta} - \sqrt{\delta}}{\sqrt{\Delta} + 3\sqrt{\delta}} = \frac{1 - \sqrt{\frac{\delta}{\Delta}}}{1 + 3\sqrt{\frac{\delta}{\Delta}}}.$$

Итак, мы получили, что наибольшая скорость сходимости достигается при параметрах, указанных в условии теоремы. Это, в общем-то, и требовалось доказать. □

Оценим скорость сходимости этого метода при $q \neq 0$. По определению

$$\ln \frac{1}{q} = \ln \frac{1 + 3\sqrt{\frac{\delta}{\Delta}}}{1 - \sqrt{\frac{\delta}{\Delta}}} = \ln \left(1 + 4\sqrt{\frac{\delta}{\Delta}} + \frac{4\frac{\delta}{\Delta}}{1 - \sqrt{\frac{\delta}{\Delta}}} \right) \approx 4\sqrt{\frac{\delta}{\Delta}}.$$

Применив попеременно-треугольный итерационный метод к решению системы уравнений (1.29) с трехдиагональной матрицей A, получим оценку на $k_0(\varepsilon)$:

$$k_0(\varepsilon) = \frac{\ln \frac{1}{\varepsilon}}{4\sqrt{\frac{\delta}{\Delta}}}.$$

Как уже показывалось для метода простой итерации, $\frac{\delta}{\Delta} = \frac{\lambda_{min}(B^{-1}A)}{\lambda_{max}(B^{-1}A)} = \operatorname{tg}^2\left(\frac{\pi h}{2}\right) \approx \frac{\pi^2 h^2}{4}$, откуда получаем оценку на $k_0(\varepsilon)$:

$$k_0(\varepsilon) \approx \frac{\ln \frac{1}{\varepsilon}}{4\frac{\pi h}{2}} = \frac{\ln \frac{1}{\varepsilon}}{2\pi} N = \{\varepsilon = \mathsf{5} \cdot \mathsf{10}^{-\mathsf{5}} \approx \mathsf{e}^{-\mathsf{10}}\} \approx 1,6N.$$

Это уже неплохой результат (гораздо меньше, чем в случае метода простой итерации), например, взяв $N = 100$, мы должны будем сделать 160 итераций вместо 20000 в МПИ.

1.6 Чебышевский набор итерационных параметров

В данном разделе мы будем применять для уже порядком заколебавшего нас уравнения $Ax = f$ такую итерационную схему:

$$B\frac{x^l - x^{l-1}}{\tau_l} + Ax^{l-1} = f. \tag{1.33}$$

Теперь фиксируем число итераций (скажем, k, а то m — маловато...), и постараемся выбрать τ_l так, чтобы погрешность на k-й итерации была минимальной:

$$||z^k|| \longrightarrow \inf_{\tau_l}.$$

Чтобы получить оценку на погрешность, перейдем от l-го приближения к погрешности $z^l = x^l - x$, как это уже делали раньше сто раз. Тогда итерационный процесс будет выглядеть просто прекрасно:

$$B\frac{z^l - z^{l-1}}{\tau_l} + Az^{l-1} = 0.$$

Отсюда без проблем получается выражение для z^l :

$$z^l = (E - \tau_l B^{-1}A)z^{l-1}, \quad l = \overline{1,\, k}.$$

Рекурсивно применяя эту формулу для z^k, получим очень-очень длинную формулу:

$$z^k = (E - \tau_k B^{-1}A)(E - \tau_{k-1}B^{-1}A)\dots(E - \tau_1 B^{-1}A)z^0. \tag{1.34}$$

Для тех, кто в танке, повторяем: необходимо так подобрать τ_l, чтобы минимизировать z^k. Оказывается, что эта задача не только имеет решение, но нам даже разрешили его опубликовать (*тоже мне, ценность великая...*). Это решение, а, точнее, набор τ_l называется **чебышевским набором итерационных параметров**. Прямо корпускулярно-волновая теория света, мдаа...

А теперь приколитесь — следующая теорема пойдет без доказательства (кому неймется, может посмотреть в [1]).

Теорема 1.6. *Пусть матрицы A и B симметричны и положительно определены, τ_l считаются по следующей формуле (k — фиксировано):*

$$\tau_l = \frac{\tau_0}{1+\rho_0 t_l},\ \text{где}\ \begin{cases} \tau_0 &= \dfrac{2}{\lambda_{min}(B^{-1}A)+\lambda_{max}(B^{-1}A)}; \\ \rho_0 &= \dfrac{1-\xi}{1+\xi}, \quad \xi = \dfrac{\lambda_{min}(B^{-1}A)}{\lambda_{max}(B^{-1}A)}; \\ t_l &= \cos\dfrac{(2l-1)\pi}{2k},\ l=\overline{1,\,k}, \end{cases}$$

а x^l вычисляется по формуле (1.33).

Тогда погрешность $||x^k - x||_A$ будет минимальной среди всех возможных, и для нее справедливо неравенство:

$$||x^k - x||_A \leqslant q_k ||x^0 - x||_A,$$

где $q_k = \dfrac{\rho_1^k}{1+\rho_1^{2k}},\ \rho_1 = \dfrac{1-\sqrt{\xi}}{1+\sqrt{\xi}}$.

Замечание 1. Подробную инструкцию о том, что надо выкурить, чтобы догадаться до таких формул, можно найти в [2].

Замечание 2. Найдем, при каких k выполняется условие остановки итерационного процесса:

$$||x^k - x||_A \leqslant \varepsilon ||x^0 - x||_A. \tag{1.35}$$

Из теоремы 1.6 следует, что оно заведомо (т.е. всегда (т.е. даже ночью)) выполняется, если $q_k \leqslant \varepsilon$. Это, в свою очередь, эквивалентно тому, что

$$\frac{\rho_1^k}{1+\rho_1^{2k}} \leqslant \varepsilon \iff \varepsilon(\rho_1^k)^2 - \rho_1^k + \varepsilon \geqslant 0.$$

Корнями этого квадратного трехчлена будут

$$\rho_1^k = \frac{1 \pm \sqrt{1-4\varepsilon^2}}{2\varepsilon} \approx \frac{1 \pm (1-2\varepsilon^2)}{2\varepsilon} \implies \left[\begin{array}{l} \rho_1^k = \varepsilon; \\ \rho_1^k = \dfrac{1}{\varepsilon} + \varepsilon. \end{array}\right.$$

Отсюда следует, что $q_k \leqslant \varepsilon$ при

$$\left[\begin{array}{l} \rho_1^k \leqslant \varepsilon; \\ \rho_1^k \geqslant \dfrac{1}{\varepsilon} + \varepsilon. \end{array}\right.$$

В теореме 1.6 мы выбирали ρ_1^k так, что оно было меньше единицы. Поэтому второй случай нам не подойдет ($\frac{1}{\varepsilon} \gg 1$), из чего следует, что

$$q_k \leqslant \varepsilon \iff \rho_1^k \leqslant \varepsilon.$$

Таким образом, (1.35) верно при

$$k \geqslant k_0(\varepsilon) = \left\lceil \frac{\ln\frac{1}{\varepsilon}}{\ln\frac{1}{\rho_1}} \right\rceil. \tag{1.36}$$

В данном случае за скорость сходимости принимается $\ln\frac{1}{\rho_1}$. Оценим этот параметр:

$$\ln\frac{1}{\rho_1} = \ln\frac{1+\sqrt{\xi}}{1-\sqrt{\xi}} = \ln\left(1 + 2\sqrt{\xi} + O(\xi)\right) \approx 2\sqrt{\xi}.$$

Посмотрим, как работает этот метод (1.33) (для упрощения сделаем матрицу B единичной) применительно к нашей модельной задаче (1.28). Напомним, матрица системы в примере была вида:

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & \ddots & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix}.$$

Собственные значения для нее выглядят так:

$$\lambda_{min} = \frac{4}{h^2}\sin^2(\frac{\pi h}{2}), \quad \lambda_{max} = \frac{4}{h^2}\cos^2(\frac{\pi h}{2}),$$

поэтому $\xi = \frac{\lambda_{min}}{\lambda_{max}} = \mathrm{tg}^2(\frac{\pi h}{2})$.

И снова, в который раз (и не в последний, по-моему) зададимся $\varepsilon = 0,5\cdot 10^{-4} \approx e^{-10}$. В этом случае из (1.36) можно поиметь следующую формулу:

$$k_0(\varepsilon) = \frac{10}{2\,\mathrm{tg}(\frac{\pi h}{2})} \approx \frac{10}{\pi h} = \ \{\mathsf{h} = \frac{1}{\mathsf{N}}\} \ = \frac{10}{\pi}N \approx 3,5N.$$

Тогда для $N = 10$ нам понадобится 35 итераций, а для $N = 100$ — 350 итераций, то есть уже можно себя не убивать... Откровенно говоря, этот метод немеряно круче (на порядок) метода простой итерации, да и то, что схема явная, не может не радовать.

С другой стороны, этот метод медленнее, чем попеременно-треугольный, но там-то надо было матрицы обращать, а здесь этого иногда можно и не делать...

Упорядоченный набор чебышевских параметров

Вообще говоря, в описанном в предыдущем разделе процессе не предполагалось монотонности убывания погрешности. Обычно ее и нет, что на практике при реализации метода с чебышевским набором параметров часто приводит к возникновению неприятных ситуаций (переполнение регистров и прочий отстой). Для предотвращения таких дел набор параметров можно особым образом упорядочить, чтобы как-то скомпенсировать увеличение погрешности за счет ее уменьшения на соседних итерациях.

К примеру, в нашем методе отличия на разных итерациях заключались в изменяющемся параметре τ_l, который в свою очередь зависит от $t_l = \cos(\frac{(2l-1)\pi}{2k})$. По-другому это можно записать так:

$$t_l = \cos\left(\frac{(2l-1)\pi}{2k}\right) = \cos\left(\frac{\pi}{2k}\,\theta_l^k\right), \quad \theta_l^k = \{1, 3, \ldots, 2k-1\}.$$

Одним из способов упорядочивания параметров τ_l является простая перестановка значений θ_l^k. Мы рассмотрим пример такой перестановки, когда k является степенью двойки: $k = 2^p$. Тогда значения будут считаться по рекуррентным формулам:

$$\left\{\begin{array}{rcll} \theta_1^1 & = & 1; & \\ \theta_{2i-1}^{2m} & = & \theta_i^m, & i = \overline{1,\,m}; \\ \theta_{2i}^{2m} & = & 4m - \theta_{2i-1}^{2m}, & i = \overline{1,\,m}. \end{array}\right. \tag{1.37}$$

— последние две формулы необходимо применять для всех $m = 1, 2, 4, \ldots, 2^{p-1}$. Проще это будет понять на примере:

$$k = 2^4 = 16:$$

θ_1^1		1;
$\theta_{1,2}^2$	$m=1:$	1, 4-1=3;
$\theta_{1,\dots,4}^4$	$m=2:$	1, 8-1=7, 3, 8-3=5;
$\theta_{1,\dots,8}^8$	$m=4:$	1, 15, 7, 9, 3, 13, 5, 11;

— потом по θ_l^k строятся t_l и τ_l, при этом получается более устойчивый метод.

В следующей теореме, которую мы опять примем без доказательства, будет показано, как можно соединить попеременно-треугольный итерационный метод с упорядоченным набором чебышевских параметров.

Теорема 1.7. *Пусть уравнение $Ax=f$ решается по следующей итерационной схеме (число шагов — k — фиксировано):*

$$B\frac{x^l-x^{l-1}}{\tau_l}+Ax^{l-1}=f,\quad l=\overline{1,k},$$

где $B=(E+\omega R_1)(E+\omega R_2)$, $R_1+R_2=A$ (матрицы R_1, R_2 имеют тот же смысл, что и в теореме 1.4).

Пусть также существуют такие положительные константы δ и Δ, что

$$A\geqslant\delta E,\ \frac{\Delta}{4}A\geqslant R_1R_2.$$

Тогда, если выбирать итерационные параметры τ_l так:

$$\begin{cases} \tau_0 = \dfrac{2}{\gamma_1+\gamma_2}; \\ \tau_l = \dfrac{\tau_0}{1+\rho_0 t_l},\quad \rho_0=\dfrac{1-\xi}{1+\xi},\quad \xi=\dfrac{\gamma_1}{\gamma_2}, \quad \begin{cases} \gamma_1 = \dfrac{\delta\sqrt{\Delta}}{2(\sqrt{\Delta}+\sqrt{\delta})}; \\ \gamma_2 = \dfrac{\sqrt{\delta\Delta}}{4}, \end{cases} \\ t_l = \cos\left(\dfrac{\pi}{2k}\theta_l^k\right),\quad \theta_l^k \text{ считаются по формулам вида (1.37).} \end{cases}$$

то для достижения точности

$$||A(x^k-x)||_{B^{-1}}\leqslant\varepsilon||A(x^0-x)||_{B^{-1}}$$

достаточно выполнить $k_0(\varepsilon)=\dfrac{\ln\frac{2}{\varepsilon}}{2\sqrt{2}\sqrt[4]{\frac{\delta}{\Delta}}}$ итераций.

Доказательство. Доказательство можно найти в [1]. □

Теперь посмотрим, какие улучшения даст этот метод в модельной задаче (1.28). Напомним, что константы δ, Δ можно взять такими:

$$\begin{cases} \delta = \lambda_{min}(A)=\dfrac{4}{h^2}\sin^2(\dfrac{\pi h}{2}); \\ \Delta = \lambda_{max}(A)=\dfrac{4}{h^2}\cos^2(\dfrac{\pi h}{2}). \end{cases}$$

Тогда при $\varepsilon=0,5\cdot10^{-4}\approx e^{-10}$ получим, что

$$k_0(\varepsilon)=\frac{\ln2+10}{2\sqrt{2}\sqrt[4]{\operatorname{tg}^2(\frac{\pi h}{2})}}\approx\frac{\ln2+10}{2\sqrt{2}\sqrt{\frac{\pi h}{2}}}=\ \{\mathsf{h}=\frac{1}{\mathsf{N}}\}\ =\frac{(\ln2+10)\sqrt{N}}{2\sqrt{\pi}}\approx3\sqrt{N}.$$

Таким образом, для $N = 10$ нам понадобится 10 итераций, а для $N = 100$ — 30 итераций. Превосходство в скорости над ранее рассмотренными методами очевидно, однако недостатки тоже налицо: необходимо уметь обращать матрицу B и много знать о спектре матрицы A. Заметим, что это общие требования у быстрых итерационных методов — но, как мы увидим далее, они не являются необходимыми.

1.7 Одношаговые итерационные методы вариационного типа

В данном разделе мы вновь будем работать с одношаговыми методами. Запишем итерационную схему метода:

$$B\frac{x^k - x^{k-1}}{\tau_k} + Ax^{k-1} = f.$$

Как и раньше, перейдем от x^k к погрешности $z^k = x^k - x$:

$$B\frac{z^k - z^{k-1}}{\tau_k} + Az^{k-1} = 0.$$

Отсюда выразим $(k+1)$-ю погрешность:

$$z^{k+1} = (E - \tau_{k+1}B^{-1}A)z^k. \tag{1.38}$$

Теперь зафиксируем некоторую матрицу D : $D = D^T > 0$. Как уже говорилось, в этом случае существует такая матрица $D^{\frac{1}{2}}$, что

$$D^{\frac{1}{2}}D^{\frac{1}{2}} = D, \quad D^{\frac{1}{2}} = (D^{\frac{1}{2}})^T.$$

Далее мы будем подбирать τ_{k+1} так, чтобы минимизировать $||z^{k+1}||_D$ (считаем, что z^k уже задана) — такой способ построения итерационного процесса называется **локальной минимизацией**.

Согласно определению $||z^{k+1}||_D$, имеем:

$$||z^{k+1}||_D = \sqrt{\langle Dz^{k+1},\, z^{k+1}\rangle} = \sqrt{\left\langle D^{\frac{1}{2}}z^{k+1},\, D^{\frac{1}{2}}z^{k+1}\right\rangle} = ||y^{k+1}||,$$

где $y^{k+1} = D^{\frac{1}{2}}z^{k+1}$.

То есть, минимизация $||z^{k+1}||_D$ эквивалентна минимизации $||y^{k+1}||$. Преобразуем немного выражение для z^{k+1} в (1.38):

$$z^{k+1} = (E - \tau_{k+1}B^{-1}A)D^{-\frac{1}{2}}D^{\frac{1}{2}}z^k.$$

Домножим обе части равенства на $D^{\frac{1}{2}}$:

$$\begin{aligned} D^{\frac{1}{2}}z^{k+1} &= D^{\frac{1}{2}}(E - \tau_{k+1}B^{-1}A)D^{-\frac{1}{2}}D^{\frac{1}{2}}z^k \iff \\ \iff D^{\frac{1}{2}}z^{k+1} &= (E - \tau_{k+1}D^{\frac{1}{2}}B^{-1}AD^{-\frac{1}{2}})D^{\frac{1}{2}}z^k. \end{aligned}$$

Согласно принятым обозначениям, имеем:

$$y^{k+1} = (E - \tau_{k+1}D^{\frac{1}{2}}B^{-1}AD^{-\frac{1}{2}})y^k.$$

Переобозначив $C = D^{\frac{1}{2}}B^{-1}AD^{-\frac{1}{2}}$, получим более короткую запись:

$$y^{k+1} = (E - \tau_{k+1}C)y^k.$$

Вспомним, что нам надо уменьшать $||y^{k+1}||$. Более удобно работать с квадратом этого выражения:

$$||y^{k+1}||^2 = \left\langle y^{k+1},\, y^{k+1}\right\rangle = \left\langle (E - \tau_{k+1}C)y^k,\, (E - \tau_{k+1}C)y^k\right\rangle.$$

Из линейности скалярного произведения получаем:

$$||y^{k+1}||^2 = ||y^k||^2 + \tau_{k+1}^2||Cy^k||^2 - 2\tau_{k+1}\left\langle y^k,\, Cy^k\right\rangle.$$

Потребуем дополнительное условие: положительную определенность C. В этом случае квадрат нормы $||y^{k+1}||$ примет удобный для нас вид:

$$||y^{k+1}||^2 = ||y^k||^2 + \left\langle Cy^k,\, Cy^k\right\rangle\left[\tau_{k+1}^2 - 2\tau_{k+1}\frac{\left\langle Cy^k,\, y^k\right\rangle}{\left\langle Cy^k,\, Cy^k\right\rangle}\right] =$$

$$= ||y^k||^2 + \left\langle Cy^k,\, Cy^k\right\rangle\left[\tau_{k+1} - \frac{\left\langle Cy^k,\, y^k\right\rangle}{\left\langle Cy^k,\, Cy^k\right\rangle}\right]^2 - \frac{\left\langle Cy^k,\, y^k\right\rangle^2}{\left\langle Cy^k,\, Cy^k\right\rangle}.$$

Очевидно, что наименьшее значение $||y^{k+1}||$ достигается, когда второе слагаемое обращается в нуль, то есть, когда:

$$\tau_{k+1} = \frac{\left\langle Cy^k,\, y^k\right\rangle}{\left\langle Cy^k,\, Cy^k\right\rangle}$$

— это выражение больше нуля, так как $C > 0$. Теперь преобразуем эту формулу, используя ранние обозначения ($y^k = D^{\frac{1}{2}}z^k,\ C = D^{\frac{1}{2}}B^{-1}AD^{-\frac{1}{2}}$):

$$\tau_{k+1} = \frac{\left\langle D^{\frac{1}{2}}B^{-1}AD^{-\frac{1}{2}}D^{\frac{1}{2}}z^k,\, D^{\frac{1}{2}}z^k\right\rangle}{\left\langle D^{\frac{1}{2}}B^{-1}AD^{-\frac{1}{2}}D^{\frac{1}{2}}z^k,\, D^{\frac{1}{2}}B^{-1}AD^{-\frac{1}{2}}D^{\frac{1}{2}}z^k\right\rangle} = \frac{\left\langle DB^{-1}Az^k,\, z^k\right\rangle}{\left\langle DB^{-1}Az^k,\, B^{-1}Az^k\right\rangle}. \tag{1.39}$$

Получилось, что для получения оптимального выражения для τ_{k+1} необходимо знать z^k. Далее с помощью варьирования матрицы D сведем это к вычислению известных нам величин. Для этого введем некоторые дополнительные понятия, связанные с погрешностью итерационного метода.

Рассмотрим произведение $Az^k = A(x^k - x) = Ax^k - Ax = Ax^k - f = r^k$ — полученная разность вычислима на каждом шаге итерационного процесса и называется **невязкой**[4].

Запишем схему итерационного процесса

$$B\frac{x^{k+1} - x^k}{\tau_{k+1}} + Ax^k = f,$$

и выразим x^{k+1} :

$$x^{k+1} = x^k - \tau_{k+1}B^{-1}(Ax^k - f) = x^k - \tau_{k+1}B^{-1}r^k = x^k - \tau_{k+1}\omega^k.$$

Число $\omega^k = B^{-1}r^k$ называется **поправкой**.

Перепишем (1.39), используя новые определения:

$$\tau_{k+1} = \frac{\left\langle D\omega^k,\, z^k\right\rangle}{\left\langle D\omega^k,\, \omega^k\right\rangle}. \tag{1.40}$$

Проще ли нам от этого стало? Да пока не очень, потому что величина z^k по-прежнему неизвестна, и вычислить ее мы не можем...

Воспользуемся возможностью варьировать D.

[4] **невязка**(от греч. $\eta\varepsilon\vartheta\alpha\varsigma\upsilon\varsigma$) — дисбаланс.

1.8 Примеры итерационных методов вариационного типа

Метод скорейшего спуска

Пусть матрица A, задающая систему уравнений $Ax = f$, симметрична и положительно определена ($A = A^T > 0$), и выберем матрицу $D = A$, тогда

$$\tau_{k+1} = \frac{\langle A\omega^k,\, z^k\rangle}{\langle A\omega^k,\, \omega^k\rangle} = \frac{\langle \omega^k,\, Az^k\rangle}{\langle A\omega^k,\, \omega^k\rangle} = \frac{\langle \omega^k,\, r^k\rangle}{\langle A\omega^k,\, \omega^k\rangle}$$

— теперь, так как нам известно r^k на каждом шаге, то τ_{k+1} становится вычислимым.

Получим необходимое условие на матрицу B для применимости итерационных методов вариационного типа. Используя то, что

$$C = D^{\frac12}B^{-1}AD^{-\frac12} = \{\mathsf{D} = \mathsf{A} \implies \mathsf{D}^{\frac12} = \mathsf{A}^{\frac12}\} = A^{\frac12}B^{-1}AA^{-\frac12} = A^{\frac12}B^{-1}A^{\frac12},$$

и положительную определенность матрицы C, получим

$$0 < \langle Cx,\, x\rangle = \left\langle A^{\frac12}B^{-1}A^{\frac12}x,\, x\right\rangle = \left\langle B^{-1}A^{\frac12}x,\, A^{\frac12}x\right\rangle.$$

Обозначив $A^{\frac12}x = y$, получим $\left\langle B^{-1}y,\, y\right\rangle > 0$, или, еще раз переобозначив $B^{-1}y = u$, $\langle u,\, Bu\rangle > 0$, то есть матрица B должна быть положительно определена — это и есть необходимое условие применимости метода вариационного типа.

Возьмем $B = E$:

$$x^{k+1} = x^k - \tau_{k+1}r^k, \quad \text{где} \quad \tau_{k+1} = \frac{\langle r^k,\, r^k\rangle}{\langle Ar^k,\, r^k\rangle}.$$

Этот набор $(\widetilde{\tau})$ соответствует **методу скорейшего спуска**.

Пояснение названия метода. Введем функцию $F(x)$:

$$F(x) = \langle Ax,\, x\rangle - 2\langle f,\, x\rangle = \sum_{i=1}^{n}(Ax)_i x_i - 2\sum_{i=1}^{n} f_i x_i = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_j x_i - 2\sum_{i=1}^{n} f_i x_i, \quad \text{где } x \in \mathbb{R}^n.$$

Пусть $A = A^T$, найдем градиент функции $F(x)$, чтобы определить направление максимального убывания:

$$\operatorname{grad} F(x) = \left(\frac{\partial F}{\partial x_1},\, \ldots,\, \frac{\partial F}{\partial x_n}\right)^T,$$

где $\dfrac{\partial F}{\partial x_l} = \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} a_{ij}\dfrac{\partial}{\partial x_l}(x_j,\, x_i) - 2f_l = \sum\limits_{i=1}^{n} a_{il}x_i + \sum\limits_{j=1}^{n} a_{lj}x_j - 2f_l.$

Используя симметричность матрицы A, получим:

$$\frac{\partial F}{\partial x_l} = 2\sum_{i=1}^{n} a_{li}x_i - 2f_l = 2((Ax)_l - f_l) = 2r_l.$$

Таким образом, мы показали, что градиент функции $F(x)$ равен удвоенному вектору невязки:

$$\operatorname{grad} F(x) = 2r.$$

То есть, учитывая формулу $x^{k+1} = x^k - \tau_{k+1}r^k$, и представление для невязки, получается, что каждый раз мы строим x^{k+1} от x^k в направлении антиградиента функции $F(x)$.

При этом остается свободным параметр τ_{k+1}. Будем его выбирать так, чтобы минимизировать значение функции $F(x^{k+1})$ на $k+1$ приближении.

$$\begin{aligned}\min_{\tau_{k+1}} F(x^{k+1}) &= \min_{\tau_{k+1}}\big(\langle Ax^{k+1},\, x^{k+1}\rangle - 2\langle f,\, x^{k+1}\rangle\big) = \\ &= \min_{\tau_{k+1}}\big(\langle A(x^k-\tau_{k+1}r^k),\, x^k-\tau_{k+1}r^k\rangle - 2\langle f,\, x^k-\tau_{k+1}r^k\rangle\big).\end{aligned}$$

Раскроем скобки, и, в силу свойств скалярного произведения, получим

$$\min_{\tau_{k+1}} F(x^{k+1}) = \min_{\tau_{k+1}}\Big(\langle Ax^k,\, x^k\rangle - 2\tau_{k+1}\langle r^k,\, Ax^k\rangle + \tau_{k+1}^2\langle Ar^k,\, r^k\rangle - 2\langle f,\, x^k\rangle + 2\tau_{k+1}\langle f,\, r^k\rangle\Big).$$

Чтобы найти минимум по τ_{k+1}, возьмем производную по этой переменной от функции $F(x)$ и приравняем ее к 0.

$$\frac{\partial F(x^{k+1})}{\partial \tau_{k+1}} = -2\langle r^k,\, Ax^k\rangle + 2\tau_{k+1}\langle Ar^k,\, r^k\rangle + 2\langle f,\, r^k\rangle = 0.$$

Откуда получаем выражение для τ_{k+1} :

$$\tau_{k+1} = \frac{\langle r^k,\, Ax^k\rangle - \langle r^k,\, f\rangle}{\langle Ar^k,\, r^k\rangle} = \frac{\langle r^k,\, r^k\rangle}{\langle Ar^k,\, r^k\rangle}.$$

Как нетрудно заметить оно совпадает с записанным ранее τ_{k+1} в определении метода скорейшего спуска. Таким образом, этот метод гарантирует построение итерационной последовательности, которая строится в направлении максимального уменьшения функции $F(x)$. Из курса вариационного исчисления известно, что минимизация функции $F(x)$ приводит к наиболее быстрой сходимости x^k к точному решению системы.

Рассмотрим скорость сходимости этого метода. Будем подбирать τ_{k+1} так, чтобы минимизировать z^{k+1} в методе скорейшего спуска. Понятно, что при выборе τ_{k+1} отличным от оптимального, гарантирующего минимум, будут получены величины не меньше. Таким образом, получаем оценку сверху:

$$||z^{k+1}||_A \leqslant ||(E-\tau_0 B^{-1}A)z^k||_A.$$

В параграфе об оценке сходимости одношаговых стационарных методов была получена верхняя оценка $||z^{k+1}||_A \leqslant q||z^k||_A$. Очевидно, метод скорейшего спуска имеет оценку не хуже, чем ОСИМ, и, соответственно, итерационные методы вариационного типа имеют погрешность не хуже, чем соответствующие (по матрице B) ОСИМ с оптимальным выбором параметра τ_0. Отсюда можно сделать заключение, что вариационные методы сходятся не медленнее стационарных методов.

Кроме того, для метода вариационного типа следует обратить внимание на то, что данный результат (оценку погрешности) можно получить, практически не исследуя матрицу $B^{-1}A$.

Метод минимальных невязок

В качестве матрицы D возьмем $D = A^TA$. Из этого следует, что $D = D^T > 0$. Тогда, учитывая представление (1.40), параметр τ_{k+1} будет считаться так:

$$\tau_{k+1} = \frac{\langle D\omega^k,\, z^k\rangle}{\langle D\omega^k,\, \omega^k\rangle}.$$

Перепишем τ_{k+1} так, чтобы его можно было вычислить. Распишем D :

$$\tau_{k+1} = \frac{\langle A^TA\omega^k,\, z^k\rangle}{\langle A^TA\omega^k,\, \omega^k\rangle} = \frac{\langle A\omega^k,\, r^k\rangle}{\langle A\omega^k,\, A\omega^k\rangle} = \frac{\langle A\omega^k,\, r^k\rangle}{||A\omega^k||^2}$$

— теперь параметр τ_{k+1} вычислим, так как невязка и поправка нам известны на каждом шаге.

Для сходимости итерационных методов вариационного типа мы требовали положительной определенности матрицы C. Посмотрим к чему это условие приведет в методе минимальных невязок. Итак, найдем ограничения на исходную систему.

$$0 < \langle Cx,\, x\rangle = \left\langle D^{\frac{1}{2}} B^{-1} A D^{-\frac{1}{2}} x,\, x \right\rangle =$$
$$= \ \{\text{обозначим } \mathsf{D^{-\frac{1}{2}}x = y}\} \ = \left\langle D^{\frac{1}{2}} B^{-1} A y,\, D^{\frac{1}{2}} y \right\rangle = \left\langle DB^{-1}Ay,\, y\right\rangle = \left\langle A^T A B^{-1} A y,\, y\right\rangle.$$

Сделаем еще две замены: $Ay = u$ и $u = Bv$, тогда получим

$$\langle Cx,\, x\rangle = \left\langle AB^{-1}u,\, u\right\rangle = \langle Av,\, Bv\rangle = \left\langle B^T Av,\, v\right\rangle.$$

То есть, надо следить за тем, чтобы матрица $B^T A$ была положительно определена. Если в качестве B возьмем единичную матрицу, то метод минимальных невязок будет применим к системам уравнений с положительно определенными матрицами.

Объяснение названия метода. Параметр τ_{k+1} мы подбираем таким образом, чтобы минимизировать норму погрешности:

$$\min_{\tau_{k+1}} ||z^{k+1}||_D = \min_{\tau_{k+1}} \sqrt{\langle Dz^{k+1},\, z^{k+1}\rangle} = \min_{\tau_{k+1}} \sqrt{\langle A^T A z^{k+1},\, z^{k+1}\rangle} = \min_{\tau_{k+1}} \sqrt{\langle r^{k+1},\, r^{k+1}\rangle} = \min_{\tau_{k+1}} ||r^{k+1}||$$

— таким образом, минимизируя погрешность, мы минимизируем невязки.

Метод минимальных поправок

В качестве матрицы D возьмем $D = A^T B^{-1} A$, и наложим ограничения на матрицу: $D = D^T > 0$, т.е. матрица B должна быть симметрической положительно определенной.

Возьмем $B = E \implies r^k = \omega^k$. При таком выборе матриц B и D получим следующее выражение для параметра τ_{k+1} :

$$\tau_{k+1} = \frac{\left\langle D\omega^k,\, z^k\right\rangle}{\left\langle D\omega^k,\, \omega^k\right\rangle} = \frac{\left\langle A^T B^{-1} A\omega^k,\, z^k\right\rangle}{\left\langle A^T B^{-1} A\omega^k,\, \omega^k\right\rangle} = \frac{\left\langle B^{-1} A\omega^k,\, r^k\right\rangle}{\left\langle B^{-1} A\omega^k,\, A\omega^k\right\rangle} = \frac{\left\langle A\omega^k,\, \omega^k\right\rangle}{\left\langle A\omega^k,\, A\omega^k\right\rangle}.$$

Итак, формула для итерационного параметра становится реализуемой. Исходя из ограничений наложенных на матрицы C выше ($C > 0$) опишем класс систем, которые можно решать с помощью метода минимальных поправок. Раскроем условие положительной определенности:

$$0 < \langle Cx,\, x\rangle = \left\langle D^{\frac{1}{2}} B^{-1} A D^{-\frac{1}{2}} x,\, x\right\rangle.$$

Как и в методе минимальных невязок, обозначим $D^{-\frac{1}{2}} x = y,\ Ay = u,\ B^{-1}u = v$, тогда получим:

$$0 < \left\langle D^{\frac{1}{2}} B^{-1} A y,\, D^{\frac{1}{2}} y\right\rangle = \left\langle DB^{-1}Ay,\, y\right\rangle = \left\langle A^T B^{-1} A B^{-1} A y,\, y\right\rangle =$$
$$= \left\langle B^{-1} A B^{-1} A y,\, Ay\right\rangle = \left\langle B^{-1} A B^{-1} u,\, u\right\rangle = \left\langle B^{-1} Av,\, Bv\right\rangle = \langle Av,\, v\rangle$$

— то есть, метод минимальных поправок применим для систем с положительно определенной матрицей.

Объяснение названия метода. Как и в методе минимальных невязок, будем минимизировать норму погрешности

$$||z^{k+1}||_D^2 = \left\langle Dz^{k+1},\, z^{k+1}\right\rangle = \left\langle A^T B^{-1} A z^{k+1},\, z^{k+1}\right\rangle =$$
$$= \left\langle B^{-1} A z^{k+1},\, Az^{k+1}\right\rangle = \left\langle B^{-1} r^{k+1},\, r^{k+1}\right\rangle = \left\langle \omega^{k+1},\, B\omega^{k+1}\right\rangle = ||\omega^{k+1}||_B^2.$$

Таким образом, минимизируя норму погрешности, мы минимизируем норму поправки.

Метод минимальных погрешностей

В этом примере матрица D берется равной некоторой матрице B_0 : $B_0 = B_0^T > 0$, которая должна быть к тому же легко обратимой. Матрица B определяется так:

$$B = (A^T)^{-1} B_0.$$

Как и ранее, покажем корректность формулы для τ_{k+1} :

$$\tau_{k+1} = \frac{\langle D\omega^k,\, z^k\rangle}{\langle D\omega^k,\, \omega^k\rangle} = \frac{\langle B_0 B^{-1} r^k,\, z^k\rangle}{\langle B_0 B^{-1} r^k,\, \omega^k\rangle} =$$
$$= \{\mathsf{B^{-1} = B_0^{-1} A^T}\} = \frac{\langle B_0 B_0^{-1} A^T r^k,\, z^k\rangle}{\langle B_0 B_0^{-1} A^T r^k,\, \omega^k\rangle} = \frac{\langle r^k,\, Az^k\rangle}{\langle r^k,\, A\omega^k\rangle} = \frac{\langle r^k,\, r^k\rangle}{\langle r^k,\, A\omega^k\rangle}.$$

— как уже показывалось, r^k (невязку) и w^k (поправку) мы умеем вычислять на каждом шаге.

Другим ограничением была положительная определенность матрицы C. Проверим это:

$$\langle Cx,\, x\rangle = \left\langle D^{\frac12} B^{-1} A D^{-\frac12} x,\, x\right\rangle = \{\text{обозначим } \mathsf{y = D^{-\frac12} x}\} = \left\langle D^{\frac12} B^{-1} Ay,\, D^{\frac12} y\right\rangle =$$
$$= \left\langle DB^{-1} Ay,\, y\right\rangle = \{\mathsf{D = B_0},\ \mathsf{B = (A^T)^{-1} B_0}\} = \left\langle B_0 B_0^{-1} A^T Ay,\, y\right\rangle = \langle Ay,\, Ay\rangle .$$

Очевидно, $\langle Ax,\, Ax\rangle$ всегда больше нуля, если матрица A — невырожденная. Таким образом, метод минимальных погрешностей верен для любых невырожденных матриц A.

Название метода произошло от требования минимизации погрешности $||z^{k+1}||_D = ||z^{k+1}||_{B_0}$ на $(k+1)$-м шаге.

Замечание. Все эти методы возникли из стратегии локальной минимизации погрешности от шага к шагу. При таких требованиях достигается высокая скорость сходимости, причем нам не требуется дополнительная информация, к примеру, о спектре матрицы A. Однако нет предела совершенству...

1.9 Двухшаговые итерационные методы вариационного типа

Данные методы используются при решении систем большой размерности (тысячи, десятки тысяч уравнений). Решаем мы по-прежнему систему $Ax = f$. Действуя так же, как и при работе с одношаговыми итерационными методами, можно показать, что каноническая форма двухшаговых методов такова:

$$B\frac{x^{k+1} - x^k + (1-\alpha_{k+1})(x^k - x^{k-1})}{\alpha_{k+1}\tau_{k+1}} + Ax^k = f, \quad k = 1, 2, \ldots \tag{1.41}$$

Для того, чтобы задать итерационный процесс, мы определяем два начальных приближения: x^0 и x^1. Остальные вычисляются по этой формуле, где B, α_{k+1}, τ_{k+1} — параметры, определяющие метод.

Также неявным параметром будет матрица D. Она определяет норму $||z^{k+1}||_D$, которую мы, как и раньше, будем минимизировать на каждой итерации. Для этого необходимо брать α_{k+1} и τ_{k+1} такими:

$$\begin{cases} \tau_{k+1} &= \dfrac{\langle D\omega^k,\, z^k\rangle}{\langle D\omega^k,\, \omega^k\rangle}, \quad k = 1, 2, \ldots; \\ \alpha_1 &= 1; \\ \alpha_{k+1} &= \left(1 - \dfrac{\tau_{k+1}}{\tau_k \alpha_k} \cdot \dfrac{\langle D\omega^k,\, z^k\rangle}{\langle D\omega^{k-1},\, z^{k-1}\rangle}\right)^{-1}, \quad k = 1, 2, \ldots \end{cases}$$

— подробный вывод этих формул можно найти в [2].

Как и раньше, при надлежащем выборе мы можем избавиться от z^k в скалярном произведении и корректно применять эти формулы при подсчете приближений. Поясним это на примерах.

Метод сопряженных градиентов

В данном методе берется $D = A$, поэтому матрица A должна быть симметричной и положительно определенной. Как уже показывалось ранее, в этом случае для τ_{k+1} можно получить такую формулу:

$$\tau_{k+1} = \frac{\langle \omega^k,\, r^k \rangle}{\langle A\omega^k,\, \omega^k \rangle}.$$

В свою очередь, для α_{k+1} будет справедливо такое выражение:

$$\alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k \alpha_k} \cdot \frac{\langle \omega^k,\, r^k \rangle}{\langle \omega^{k-1},\, r^{k-1} \rangle}\right)^{-1}$$

— в состав этих формул входит уже не погрешность, а невязка и поправка, которые мы вычислять умеем.

Метод сопряженных невязок

В данном случае берется $D = A^T A$, и несложно проверить, что $D = D^T > 0$. Получающиеся формулы для τ_{k+1} и α_{k+1} мы опускаем (вывод предоставляется читателю).

Кроме того, в данном методе и во всех последующих вводится дополнительное требование:

$$DB^{-1}A = (DB^{-1}A)^T > 0. \tag{1.42}$$

В данном случае оно эквивалентно тому, что $B^T A > 0$. Это небольшое ограничение на матрицу B.

Метод сопряженных поправок

Здесь мы берем $D = A^T B^{-1} A$. Легко проверить, что $D = D^T > 0$, если $B = B^T > 0$. Ограничение (1.42) приводит к требованию положительной определенности матрицы A. Формулы для τ_{k+1} и α_{k+1} также становятся корректными.

Метод сопряженных погрешностей

Здесь все тоже просто: $D = B_0$, где $B_0 = B_0^T > 0$, причем матрица B_0 должна быть легко обратимой.

Матрица B задается строго:

$$B = (A^T)^{-1} B_0.$$

Ограничение (1.42) приводит к требованию невырожденности матрицы A.

Общие замечания

(1) Можно показать, что любой из выше перечисленных методов сходится не медленнее, чем соответствующий ему одношаговый итерационный метод с чебышевским набором параметров.

(2) Кроме того, если количество шагов итерации превысит размерность системы, то последовательность итерационных приближений выйдет на точное решение — то есть вышеперечисленные методы фактически являются прямыми. Другое дело, что нужная точность будет достигнута намного раньше.

(3) Учитывая, что $\alpha_1 = 1$, можно не угадывать x^1 (первое приближение), а просто подставить в (1.41) $k = 0$:

$$B\frac{x^1 - x^0}{\alpha_1 \tau_1} + Ax^0 = f$$

— это простая одношаговая схема для поиска x^1.

Глава 2

Задачи на собственные значения

Сначала о терминологии. Пусть $A = (a_{ij})$ — матрица размера $n \times n$, а $x = (x_1, x_2, \ldots, x_n)^T$ — вектор неизвестных. Тогда поиск таких констант λ и векторов x, что

$$Ax = \lambda x,$$

называется **задачей на собственные значения**. Нетрудно заметить, что эта задача эквивалентна поиску таких x и λ, для которых

$$(A - \lambda E)x = 0.$$

При этом λ называются **собственными значениями**, а соответствующие им вектора x — **собственными векторами**.

Из курса линейной алгебры известно, что если $\det(A - \lambda E) = 0$, то решение существует. Для его нахождения надо найти такие λ, чтобы определитель $|A - \lambda E|$ обращался в ноль. Этот определитель является полиномом от λ с коэффициентами из A — поэтому, если n мало, то корни этого полинома легко найти. Это также просто сделать, если матрица A является диагональной или верхнетреугольной: в этом случае определитель будет равен произведению диагональных элементов.

Далее мы будем рассматривать различные методы нахождения собственных значений.

Определение. Метод решает **полную проблему собственных значений**, если он находит их все. В противном случае (например, надо найти только границы спектра матрицы A) говорят, что решается **частичная проблема**.

2.1 Поиск собственных значений методом вращений

Данный метод, предложенный К. Якоби, позволяет найти все собственные значения (решает полную проблему поиска). Для этого требуется симметричность матрицы A: $A = A^T$. Известно, что в этом случае для A справедливо такое представление (оно единственно):

$$A = Q^T \Lambda Q, \tag{2.1}$$

где Q — некая ортогональная матрица ($Q^T = Q^{-1}$), а Λ — диагональная матрица, причем на диагонали у нее стоят именно собственные значения матрицы A: $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.

Если домножить равенство (2.1) на $(Q^T)^{-1} = Q$ слева, а на $Q^{-1} = Q^T$ — справа, то получим такую формулу:

$$\Lambda = QAQ^T \tag{2.2}$$

— то есть для нахождения собственных значений матрицы A нам необходимо найти соответствующую матрицу Q и провести два матричных умножения.

Матрицу Q будем находить с помощью ортогональных преобразований матрицы A, постепенно уменьшая абсолютные значения ее внедиагональных элементов:

$$\begin{aligned} A_1 &= V_{ij}^1\, A\, (V_{ij}^1)^T; \\ A_2 &= V_{ij}^2\, A_1\, (V_{ij}^2)^T; \\ &\dots \\ A_k &= V_{ij}^k\, A_{k-1}\, (V_{ij}^k)^T, \end{aligned}$$

и так далее. Здесь V_{ij}^k — некие ортогональные матрицы, а индексы в них говорят о номере преобразования — (k) и об индексе уменьшаемого элемента из A_{k-1} — (ij). Произведение ортогональных матриц дает ортогональную матрицу, поэтому, если мы на некотором шаге придем к диагональной матрице, то это будет означать получение преобразования (2.2).

Матрицы V_{ij}^k будут задаваться так (φ — пока свободный параметр):

$$V_{ij}^k = \begin{pmatrix} 1 & & & & 0 & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & \cos\varphi & & -\sin\varphi & & \\ & & & & 1 & & 0 & \\ 0 & & & & & \ddots & & 0 \\ & & & & 0 & & 1 & \\ & & & \sin\varphi & & \cos\varphi & & \\ & & & & & & & 1 \\ & & & & & & \ddots & \\ & & & & 0 & & & 1 \end{pmatrix}$$

— здесь на диагонали стоят единицы везде, кроме v_{ii} и v_{jj}, где стоят косинусы, а в позициях v_{ji} и v_{ij} стоят синусы. Все остальные элементы — нулевые. Легко проверить, что все эти матрицы ортогональны.

Индексы i и j задаются на каждом шаге заново. Они обозначают индекс максимального по модулю внедиагонального элемента, то есть

$$|a_{ij}| = \max_{\substack{l,m \\ l \neq m}} |a_{lm}^k|,$$

где a_{lm}^k — элементы матрицы A_k (k-е приближение к Λ). Если i и j можно выбрать несколькими способами, то берется произвольная пара. Если же все внедиагональные элементы — нулевые, то процесс прекращается.

Назовем **количественной характеристикой диагональности** матрицы A_k такое число:

$$t(A_k) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} (a_{ij}^k)^2$$

Понятно, что если $t(A_k) \overset{k\to\infty}{\longrightarrow} 0$, то A_k сходится к диагональному виду. Таким образом, через $t(A_k)$ можно наглядно оценивать скорость сходимости данного метода.

Теперь фиксируем максимальный элемент a_{ij} и установим соотношение между матрицами A_{k+1} и A_k. A_{k+1} мы задавали так:

$$A_{k+1} = V_{ij}^{k+1}\, A_k\, (V_{ij}^{k+1})^T.$$

Обозначим $B_k = A_k\, (V_{ij}^{k+1})^T = (b_{ms}^k)$ и $(V_{ij}^{k+1})^T = (\overline{v}_{lm}^{k+1})$ — обозначение для элементов матрицы $(V_{ij}^{k+1})^T$. Тогда, согласно определению матричного произведения,

$$b_{ms}^k = \sum_{p=1}^{n} a_{mp}^k\, \overline{v}_{ps}^{k+1} = \begin{cases} a_{ms}^k, & s \neq i, j; \\ a_{mi}^k \cos\varphi - a_{mj}^k \sin\varphi, & s = i; \\ a_{mi}^k \sin\varphi + a_{mj}^k \cos\varphi, & s = j. \end{cases} \tag{2.3}$$

— эта сложная формула вытекает из того, что среди элементов $\overline{v}_{ps}^{k+1}$ один ненулевой при $p \neq i,\ j$, и два ненулевых при $p = i,\ j$.

Теперь, согласно принятым обозначениям, $A_{k+1} = V_{ij}^{k+1} B_k$. Выведем отсюда формулу для элементов матрицы A_{k+1} :

$$a_{ms}^{k+1} = \sum_{p=1}^{n} v_{mp}^{k+1}\, b_{ps}^{k} = \begin{cases} b_{ms}^{k}, & m \neq i, j; \\ b_{is}^{k}\cos\varphi - b_{js}^{k}\sin\varphi, & m = i; \\ b_{is}^{k}\sin\varphi + b_{js}^{k}\cos\varphi, & m = j. \end{cases} \tag{2.4}$$

— идея построения системы та же (v_{mp}^{k+1} — элементы матрицы V_{ij}^{k+1}).

Напомним, что a_{ij} — максимальный по модулю элемент из A. Положив в (2.4) $m = i$ и $s = j$, получим такую формулу для него:

$$a_{ij}^{k+1} = b_{ij}^{k}\cos\varphi - b_{jj}^{k}\sin\varphi.$$

Теперь распишем b_{ij}^{k} и b_{jj}^{k} через (2.3), взяв $m = i,\ j$ и $s = j$:

$$\begin{aligned} a_{ij}^{k+1} &= (a_{ii}^{k}\sin\varphi + a_{ij}^{k}\cos\varphi)\cos\varphi - (a_{ji}^{k}\sin\varphi + a_{jj}^{k}\cos\varphi)\sin\varphi = \quad \{\mathsf{A} = \mathsf{A}^{\mathsf{T}} \Rightarrow \mathsf{A_k} = \mathsf{A_k^{\mathsf{T}}}\} = \\ &= (a_{ii}^{k} - a_{jj}^{k})\sin\varphi\cos\varphi + a_{ij}^{k}(\cos^2\varphi - \sin^2\varphi) = \frac{(a_{ii}^{k} - a_{jj}^{k})\sin 2\varphi}{2} + a_{ij}^{k}\cos 2\varphi. \end{aligned}$$

Напомним, мы пытаемся уменьшить внедиагональные элементы матрицы A при ортогональном преобразовании. Потребуем равенство нулю для a_{ij}^{k+1}. Таким образом, написанное выше преобразование превращается в уравнение относительно φ. Решая его, получим:

$$\varphi = \frac{1}{2}\operatorname{arctg}\frac{2a_{ij}^{k}}{a_{ii}^{k} - a_{jj}^{k}}.$$

Вычислим количественную характеристику диагональности получившейся матрицы A^{k+1} :

$$t(A^{k+1}) = \sum_{m=1}^{n}\sum_{\substack{s=1 \\ s\neq m}}^{n} (a_{ms}^{k+1})^2.$$

Согласно формулам, элементы в A^k при умножении на $(V_{ij}^{k+1})^T$ изменяются только в i-м и j-м столбцах. Аналогично, элементы в A^{k+1} изменяются относительно B^k только в i-й и j-й строках. То есть A^{k+1} отлична от A^k только в i-х и j-х строках и столбцах.

Выделим в сумме неменяющиеся элементы. Распишем сумму, раскроем скобки, и проведем все преобразования:

$$\begin{aligned} t(A^{k+1}) &= \sum_{\substack{m=1 \\ m\neq i,j}}^{n}\sum_{\substack{s=1 \\ s\neq i,j,m}}^{n} (a_{ms}^{k})^2 + \sum_{\substack{m=1 \\ m\neq i,j}}^{n}\left[(b_{mi}^{k})^2 + (b_{mj}^{k})^2\right] + \sum_{\substack{s=1 \\ s\neq i,j}}^{n}\left[(a_{is}^{k+1})^2 + (a_{js}^{k+1})^2\right] = \\ &= \sum_{\substack{m=1 \\ m\neq i,j}}^{n}\sum_{\substack{s=1 \\ s\neq i,j,m}}^{n} (a_{ms}^{k})^2 + \sum_{\substack{m=1 \\ m\neq i,j}}^{n}\left[(a_{mi}^{k})^2\cos^2\varphi + (a_{mj}^{k})^2\sin^2\varphi - 2a_{mi}^{k}a_{mj}^{k}\sin\varphi\cos\varphi + \right. \\ &\quad \left. + (a_{mi}^{k})^2\sin^2\varphi + (a_{mj}^{k})^2\cos^2\varphi + 2a_{mi}^{k}a_{mj}^{k}\sin\varphi\cos\varphi\right] + \sum_{\substack{s=1 \\ s\neq i,j}}^{n}\left[(b_{is}^{k})^2\cos^2\varphi + (b_{js}^{k})^2\sin^2\varphi - \right. \\ &\quad \left. - 2b_{is}^{k}b_{js}^{k}\sin\varphi\cos\varphi + (b_{is}^{k})^2\sin^2\varphi + (b_{js}^{k})^2\cos^2\varphi + 2b_{is}^{k}b_{js}^{k}\sin\varphi\cos\varphi\right] = \end{aligned}$$

$$= \sum_{\substack{m=1\\ m\neq i,j}}^{n} \sum_{\substack{s=1\\ s\neq i,j,m}}^{n} (a_{ms}^k)^2 + \sum_{\substack{m=1\\ m\neq i,j}}^{n} \left[(a_{mi}^k)^2 + (a_{mj}^k)^2\right] + \sum_{\substack{s=1\\ s\neq i,j}}^{n} \left[(a_{is}^k)^2 + (a_{js}^k)^2\right] + 2(a_{ij}^k)^2 - 2(a_{ij}^k)^2 =$$

$$= t(A^k) - 2(a_{ij}^k)^2.$$

Из всей этой последовательности формул следует, что $t(A^{k+1}) < t(A^k)$ — как видно, характеристика диагональности монотонно убывает с ростом индекса, причем уменьшается каждый раз на $2(a_{ij}^k)^2$ — удвоенный квадрат максимального элемента. Из этого следует, что описанный итерационный процесс приводит к диагональной матрице.

Запишем оценку на скорость сходимости процесса. Пусть a_{ij}^k — максимальный внедиагональный элемент. Простым подсчетом элементов матрицы можно получить такое неравенство:

$$t(A^k) \leqslant n(n-1)(a_{ij}^k)^2.$$

Отсюда следует, что $(a_{ij}^k)^2 \geqslant \dfrac{t(A^k)}{n(n-1)}$ для $n \geqslant 2$. Подставим это неравенство в ранее полученное соотношение между $t(A^k)$ и $t(A^{k+1})$:

$$t(A^{k+1}) = t(A^k) - 2(a_{ij}^k)^2 \leqslant t(A^k) - \frac{2}{n(n-1)} t(A^k) = qt(A^k), \quad \text{где } q = 1 - \frac{2}{n(n-1)} < 1.$$

Применив эту операцию k раз, получим:

$$t(A^k) \leqslant q^k t(A).$$

Из последнего неравенства видно, что процесс нахождения матрицы Q сходится к диагональной матрице Λ со скоростью геометрической прогрессии со знаменателем q.

Можно немного оптимизировать способ выбора «плохого» элемента: сначала выбираем «плохую» строку (например, выбираем строку s, где $\sum_{i=1}^{n}(a_{si})^2$ имеет максимальное значение), а потом из этой строчки выбираем максимальный по модулю элемент. При этом, немного неоптимальный выбор a_{sk} компенсируется скоростью нахождения «плохого» элемента ($2n$ вместо n^2 сравнений).

Теперь будем решать частичные проблемы. Как пример частичной проблемы, можно привести задачу нахождения границ спектра в итерационном методе.

2.2 Степенной метод поиска собственных значений

Будем рассматривать задачу нахождения максимального по модулю собственного значения симметрической матрицы A.

Алгоритм поиска

Возьмем любой вектор x^0, отличный от нуля, и построим последовательность векторов x^k такую, что

$$x^{k+1} = Ax^k, \quad k = 0, 1, \ldots \tag{2.5}$$

По этой последовательности построим числовую последовательность $\{\Lambda_1^k\}$ по следующему правилу:

$$\Lambda_1^k = \frac{\langle x^{k+1}, x^k\rangle}{\langle x^k, x^k\rangle}.$$

Пусть собственные числа занумерованы так, что первым стоит максимальное по модулю — искомое: $|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|$.

Примечание. *Обратите внимание, что максимальное по модулю собственное значение единственно — мы будем этим пользоваться.*

Покажем, что в этом случае последовательность $\{\Lambda_1^k\}$ будет сходиться к λ_1.

Так как матрица A симметрическая, то для нее существует ортонормированный базис $\{\xi_i\}$ из собственных векторов. Разложим начальное приближение x^0 по этому базису:

$$x^0 = \sum_{i=1}^{n} \alpha_i \xi_i. \tag{2.6}$$

Из (2.5) получаем выражение для x^k через x^0 :

$$x^k = Ax^{k-1} = \ldots = A^k x^0.$$

Используя разложение x^0 по базису, запишем следующую оценку на x^k :

$$x^k = A^k \sum_{i=1}^{n} \alpha_i \xi_i = A^{k-1} \sum_{i=1}^{n} \alpha_i A\xi_i = A^{k-1} \sum_{i=1}^{n} \alpha_i \lambda_i \xi_i = \ldots = \sum_{i=1}^{n} \alpha_i \lambda_i^k \xi_i = \alpha_1 \lambda_1^k \xi_1 + \sum_{i=2}^{n} \alpha_i \lambda_i^k \xi_i$$

В силу того, что $|\lambda_2| \geqslant |\lambda_i|$, где $i = \overline{3, n}$, получим, что $\sum\limits_{i=2}^{n} \alpha_i \lambda_i^k \xi_i = O(|\lambda_2|^k)$. Подсчитаем два скалярных произведения для нахождения Λ_1^k.

$$\begin{aligned} \left\langle x^k, x^k \right\rangle &= \left\langle \alpha_1 \lambda_1^k \xi_1 + O(|\lambda_2|^k), \alpha_1 \lambda_1^k \xi_1 + O(|\lambda_2|^k) \right\rangle = \alpha_1^2 \lambda_1^{2k} + O(|\lambda_1|^k \cdot |\lambda_2|^k); \\ \left\langle x^{k+1}, x^k \right\rangle &= \left\langle \alpha_1 \lambda_1^{k+1} \xi_1 + O(|\lambda_2|^{k+1}), \alpha_1 \lambda_1^k \xi_1 + O(|\lambda_2|^k) \right\rangle = \alpha_1^2 \lambda_1^{2k+1} + O(|\lambda_1|^{k+1} \cdot |\lambda_2|^k). \end{aligned}$$

Зная скалярные произведения, вычислим члены последовательности Λ_1^k :

$$\Lambda_1^k = \frac{\alpha_1^2 \lambda_1^{2k+1} + O(|\lambda_1|^{k+1} \cdot |\lambda_2|^k)}{\alpha_1^2 \lambda_1^{2k} + O(|\lambda_1|^k \cdot |\lambda_2|^k)} = \frac{\alpha_1^2 \lambda_1^{2k+1} \left(1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right)}{\alpha_1^2 \lambda_1^{2k} \left(1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right)} = \lambda_1 \left(1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right) \longrightarrow \lambda_1$$

при $k \to \infty$.

То есть видно, что последовательность Λ_1^k сходится к λ_1 — искомому максимальному собственному значению. Определим, к чему же сходится последовательность x^k. Для этого рассмотрим

$$\frac{x^k}{||x^k||} = \frac{\alpha_1 \lambda_1^k \xi_1 + O(|\lambda_2|^k)}{\sqrt{\langle x^k, x^k \rangle}} = \frac{\alpha_1 \lambda_1^k \xi_1 + O(|\lambda_2|^k)}{|\alpha_1||\lambda_1|^k \left(1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right)} = \pm\xi_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$$

Откуда видно, что x^k сходится к собственному вектору ξ_1 по направлению (чем больше k, тем ближе $\frac{x^k}{||x^k||}$ по направлению к ξ_1).

Примечание. *В разложении (2.6) в общем случае коэффициент α_1 может быть равен нулю, и итерационный процесс может не сходиться. Для устранения неопределенности обычно прогоняют алгоритм для нескольких случайно выбранных начальных приближений.*

Примеры

Поиск максимального собственного значения. В общем случае максимальное собственное значение может не быть максимальным по модулю. В этом случае переходят от матрицы A к матрице $D = A + cE$, где c — некоторая положительная константа. Легко заметить, что $\lambda(D) = \lambda(A) + c$ и $\lambda_{max}(D) = \lambda_{max}(A) + c$. Очевидно, что можно подобрать такое c, что все собственные значения D будут положительны. Теперь мы пользуемся вышеприведенным алгоритмом, находим $\lambda_{max}(D)$, а потом и $\lambda_{max}(A)$.

Для поиска минимального собственного значения можно брать константу $c < 0$ и проводить аналогичные рассуждения.

Поиск собственного значения, близкого к заданному числу Λ. Это задача о поиске такого λ, что

$$|\lambda - \Lambda| = \min_i |\lambda_i - \Lambda|.$$

Аналогично предыдущему примеру, переходим от матрицы A к матрице $D = E - c(A - \Lambda E)^2$, $c > 0$. В этом случае $\lambda(D) = 1 - c(\lambda(A) - \Lambda)^2$, и максимальное собственное значение D будет соответствовать собственному значению A, ближайшему к Λ.

2.3 Метод обратной итерации

Пусть нам дана матрица A. Обозначим соответствующие ее собственному значению λ_i собственные вектора как $x_i : Ax_i = \lambda_i x_i$. Тогда

$$(A - \lambda_i E)x_i = 0. \tag{2.7}$$

Задача состоит в том, чтобы найти один из собственных векторов x_i. Сначала для поиска собственного значения λ_i необходимо решать уравнение

$$\det(A - \lambda_i E) = 0.$$

Если мы нашли λ_i точно, то вычисление x_i не составит труда. Однако, если мы получили неточное собственное значение $\overline{\lambda_i} \approx \lambda_i$, то определитель $\det(A - \overline{\lambda_i}E)$ будет отличен от нуля. Так как должна выполняться формула (2.7), то единственным подходящим x_i будет $x_i = 0$.

Покажем, что даже при неточно вычисленном собственном значении можно вычислить собственный вектор. Фиксируем произвольный вектор b и решим систему:

$$(A - \overline{\lambda_i}E)x = b. \tag{2.8}$$

Утверждается, что решение этой системы будет приближенно равняться искомому собственному вектору: $x \approx x_i$. Для решения исходной системы построим последовательность векторов по следующему правилу:

$$(A - \overline{\lambda_i}E)x^{k+1} = x^k, \quad k = 0, 1, \ldots, \tag{2.9}$$

где $x^0 = b$.

При таком задании итерационного процесса достаточно примерно 5 итераций, чтобы с хорошей точностью определить собственный вектор x_i, соответствующий собственному значению λ_i.

Покажем, что алгоритм работает правильно. Пусть матрица A такова, что существует базис $\{\xi_i\}$ из собственных векторов этой матрицы ($x_i = C \cdot \xi_i$). Пусть $b = \sum_j \beta_j \xi_j$ и $x = \sum_j \alpha_j \xi_j$ — разложения фиксированного вектора b и искомого вектора x по этому базису. Подставим эти разложения в (2.8)

$$(A - \overline{\lambda_i}E)\sum_j \alpha_j \xi_j = \sum_j \beta_j \xi_j,$$

$$\sum_j (\alpha_j \lambda_j - \alpha_j \overline{\lambda_i})\xi_j = \sum_j \beta_j \xi_j,$$

$$\sum_j [\alpha_j(\lambda_j - \overline{\lambda_i}) - \beta_j]\xi_j = 0.$$

Только коэффициенты, равные нулю, могут обратить линейную комбинацию базисных векторов в нуль. Следовательно, получаем такое выражение для α_j :

$$\alpha_j = \frac{\beta_j}{\lambda_j - \overline{\lambda_i}}.$$

Отсюда видно, что в разложении x по базису из собственных векторов коэффициент α_i при базисном векторе ξ_i будет велик по сравнению с другими (если λ_i близко к $\overline{\lambda_i}$). Это означает, что каждый шаг итерационного процесса (2.9) приводит нас к вектору, который все больше похож на искомый собственный вектор ξ_i. Можно говорить, что итерационный процесс сходится к ξ_i по направлению.

Глава 3

Численные методы решения нелинейных уравнений

Пусть $f(x)$ — некоторая непрерывная функция, заданная на отрезке $[a;\, b]$. Нашей задачей будет поиск корней уравнения $f(x) = 0$ на этом отрезке.

Обычно это проходит в два этапа: сначала проводится **локализация корней**: выделение небольших отрезков, содержащих только один корень, а потом на этих отрезках проводится уточнение его значения.

3.1 Методы разделения корней

Как легко можно убедиться, локализация корней даже для функции одной переменной представляет собой достаточно трудоемкую задачу. Одним из способов является разбиение отрезка $[a;\, b]$ произвольным образом на N подотрезков $[x_k;\, x_{k+1}]$:

$$x_0 = a < x_1 < x_2 < \ldots < x_N = b.$$

Теперь мы считаем значение функции на границах этих маленьких отрезков: $f(x_k) = f_k$. Обратим свое внимание на те из них, для которых выполняется неравенство:

$$f_k \cdot f_{k+1} < 0. \tag{3.1}$$

Оно означает, что значения функции на краях отрезка различны по знаку, а из непрерывности $f(x)$ следует, что она где-то внутри обращается в ноль. Далее повторяем вышеописанную процедуру для всех отрезков, для которых верно (3.1).

Упрощенным вариантом предыдущего метода является **метод бисекции**. Пусть $f(a) \cdot f(b) < 0$ — это означает, что внутри $[a;\, b]$ точно есть корень. Теперь обозначим $x_0 = \frac{a+b}{2}$ — середина отрезка $[a;\, b]$. Если он не является корнем, то либо $f(a) \cdot f(x_0) < 0$, либо $f(x_0) \cdot f(b) < 0$. Соответственно, делим пополам $[a;\, x_0]$ или $[x_0;\, b]$ и так до достижения нужной точности — очевидно, это всегда можно сделать.

3.2 Примеры численных методов

Метод простой итерации

Пусть x^* — корень уравнения $f(x) = 0$, лежащий на отрезке $[a;\, b]$. Зададим $\tau(x)$ — некоторую функцию-параметр, не обращающуюся в ноль на $[a;\, b]$. Тогда, очевидно,

$$f(x) = 0 \iff -\tau(x)f(x) = 0 \iff x - \tau(x)f(x) = x.$$

Обозначим $S(x) = x - \tau(x)f(x)$, тогда получим, что

$$f(x) = 0 \iff S(x) = x. \tag{3.2}$$

Теперь будем строить последовательность приближений к x^* по следующему правилу:

$$x^{k+1} = S(x^k), \quad k = 0, 1, \ldots,$$

задаваясь при этом некоторым начальным приближением x^0.

Если предел последовательности $\{x^k\}$ существует: $\lim\limits_{k\to\infty} x^k = x^*$, то $x^* = S(x^*)$, и, согласно (3.2), он будет являться корнем исходного уравнения. Очевидно, эта последовательность не всегда будет сходиться, а зависит это от функции $S(x)$, которая в свою очередь определяется параметром $\tau(x)$.

Из геометрических соображений (нарисовав соответствующие графики) можно сделать предположение, что условие $|S'(x)| < 1$ на некотором отрезке $[c; d]$ внутри $[a; b]$ будет достаточным для существования предела $\{x^k\}$, если начальное приближение тоже взять из $[c; d]$. Докажем это позже, а пока рассмотрим один из вариантов данного метода — метод Ньютона.

Метод Ньютона[1]

Итак, пусть x^* — нуль функции $f(x)$:

$$f(x^*) = 0, \tag{3.3}$$

а $f'(x)$ существует, непрерывна и отлична от нуля на $[a; b]$. Это означает, в частности, что других нулей у $f(x)$ на этом отрезке нет. Теперь перепишем (3.3) следующим образом:

$$f(x^k + (x^* - x^k)) = 0.$$

Применим теперь к этому выражению формулу Лагранжа:

$$f(x^k) + f'(\overline{x})(x^* - x^k) = 0, \ \overline{x} \in [a; b].$$

Для получения формулы итерационного процесса заменим в этом равенстве $\overline{x}$ на x^k, а x^* — на x^{k+1}. Равенство превратится в уравнение:

$$f(x^k) + f'(x^k)(x^{k+1} - x^k) = 0$$

Выразим отсюда x^{k+1} :

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}. \tag{3.4}$$

Проведенное преобразование называется **линеаризацией** уравнения (3.3).

Метод Ньютона называется также **методом касательных**, так как новое приближение является абсциссой точки пересечения касательной к графику функции $f(x)$, проведенной в точке $M(x^k,\, f(x^k))$, с осью OX.

Замечание. Метод Ньютона имеет (когда сходится) квадратичную скорость сходимости:

$$x^{k+1} - x^* = O\left((x^k - x^*)^2\right).$$

Это хорошее свойство, однако метод ведь может и не сходиться. Возвращаясь к ранее введенным обозначениям, получим, что $S(x)$ в методе Ньютона считается так:

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

[1] Разработан И. Ньютоном (1669).

то есть $\tau(x) = \dfrac{1}{f'(x)}$.

Мы предположили, что метод простой итерации (т.е. и метод Ньютона) является сходящимся, если $|S'(x)| < 1$. Рассмотрим это неравенство поподробнее:

$$S'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} \implies |S'(x)| < 1 \iff$$

$$\iff \left| \frac{f(x)f''(x)}{(f'(x))^2} \right| < 1. \tag{3.5}$$

Таким образом, метод Ньютона будет сходится, если неравенство (3.5) будет выполняться на некотором отрезке, содержащем начальное приближение x^0 и нужный нам корень x^*. Часто такое требование приводит к необходимости брать x^0 очень близко к x^*, что не всегда выполнимо.

Модифицированный метод Ньютона

Если у нас есть проблемы с подсчетом производной на каждом шаге, то можно воспользоваться **модифицированным методом Ньютона**, где берется $\tau(x) = \frac{1}{f'(x^0)}$, то есть

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^0)}, \quad k = 0, 1, \ldots,$$

где x^0 — начальное приближение.

Однако, в этом методе, как мы можем подсчитать, уже не квадратичная, а линейная скорость сходимости:

$$|x^{k+1} - x^*| = O(x^k - x^*).$$

Замечание. Модифицированный метод Ньютона можно назвать **методом одной касательной**, так как новые приближения являются абсциссами точек, получающихся при пересечении с осью OX прямых, параллельных касательной к графику $f(x)$ в точке $M(x^0,\, f(x^0))$.

Метод секущих

Когда нет возможности считать производную, просто заменим ее в формуле (3.4) на разностное приближение:

$$f'(x^k) \approx \frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}}.$$

Тогда получим такую формулу:

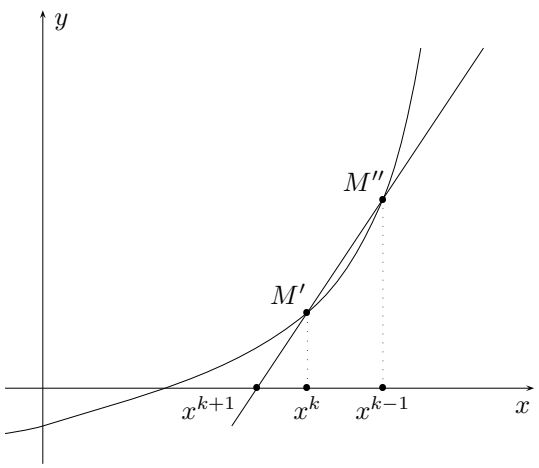
$$x^{k+1} = x^k - \frac{x^k - x^{k-1}}{f(x^k) - f(x^{k-1})} f(x^k), \quad k = 1, 2, \ldots \tag{3.6}$$

— здесь уже требуется задать два начальных приближения: x^0 и x^1. Скорость сходимости будет линейной.

Этот метод называется **методом секущих**. Для пояснения названия заметим, что уравнение для секущей, проходящей через точки $M'(x^{k-1},\, f(x^{k-1}))$ и $M''(x^k,\, f(x^k))$ (точки предыдущих приближений), будет выглядеть так:

$$\frac{y - f(x^k)}{x - x^k} = \frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}}.$$

Положив $y = 0$ и $x = x^{k+1}$, можно получить формулу (3.6). Это означает, что x^{k+1} — это абсцисса точки пересечения нашей секущей с осью OX :



3.3 Сходимость метода простой итерации

Как было сказано ранее, в методе простой итерации уравнение для нового приближения x^{k+1} таково:

$$x^{k+1} = S(x^k), \quad k = 0, 1, \ldots, \tag{3.7}$$

а x^0 — заданное начальное приближение.

Кроме того, введем такое обозначение для r-окрестности точки a :

$$U_r(a) = \{x : \ |x - a| \leqslant r\}.$$

Теперь мы готовы к формулировке теоремы о сходимости этого метода.

Теорема 3.1. *Пусть для некоторых r и a функция $S(x)$ удовлетворяет на множестве $U_r(a)$ условию Липшица с константой $q \in (0; 1)$:*

$$|S(x') - S(x'')| \leqslant q|x' - x''| \quad \forall\, x',\, x'' \in U_r(a),$$

причем $|S(a) - a| \leqslant (1 - q)r$.

Тогда уравнение $x = S(x)$ имеет на множестве $U_r(a)$ единственное решение, а метод простой итерации (3.7) сходится к этому решению при любом начальном приближении из $U_r(a)$, причем для погрешности на k-й итерации справедлива оценка:

$$|x^k - x^*| \leqslant q^k |x^0 - x^*|$$

— то есть скорость сходимости линейная.

Доказательство. Возьмем начальное приближение x^0 из множества $U_r(a)$. Индукцией покажем, что остальные x^j тоже принадлежат множеству $U_r(a)$.

Пусть $x^j \in U_r(a)$, докажем, что следующий член последовательности $x^{j+1} \in U_r(a)$.

$$|x^{j+1} - a| = |S(x^j) - a| = |S(x^j) - S(a) + S(a) - a| \leqslant |S(x^j) - S(a)| + |S(a) - a|$$

из того, что функция $S(x)$ Липшиц-непрерывна и условия теоремы получаем:

$$|x^{j+1} - a| \leqslant q|x^j - a| + (1 - q)r \leqslant qr + (1 - q)r = r.$$

Теперь покажем, что последовательность $\{x^k\}$ имеет предел, являющийся решением уравнения $x^{k+1} = S(x^k)$, причем это решение единственно. Сначала установим сходимость, для этого оценим разность двух соседних элементов:

$$|x^{j+1} - x^j| = |S(x^j) - S(x^{j-1})| \leqslant q|x^j - x^{j-1}| \leqslant \ldots \leqslant q^j|x^1 - x^0|.$$

Покажем, что выполняется критерий Коши сходимости числовой последовательности:

$$|x^{k+p} - x^k| = \left|\sum_{j=1}^{p}\left(x^{k+j} - x^{k+j-1}\right)\right| \leqslant \sum_{j=1}^{p}\left|x^{k+j} - x^{k+j-1}\right| \leqslant$$

$$\leqslant \sum_{j=1}^{p} q^{k+j-1}|x^1 - x^0| = q^k|x^1 - x^0| \cdot \sum_{j=1}^{p} q^{j-1} < q^k|x^1 - x^0| \cdot \sum_{j=1}^{\infty} q^{j-1} = \frac{q^k}{1-q}|x^1 - x^0|.$$

Последнее выражение не зависит от p, и его можно сделать меньше любого $\varepsilon > 0$, и мы можем вычислить, с какого $k(\varepsilon)$ будет выполнена эта оценка.

Таким образом, числовая последовательность $\{x^k\}_{k=0}^{\infty}$ сходится при $k \to \infty$ к некоторому $x^* \in U_r(a)$. Покажем, что этот предел является корнем уравнения $x = S(x)$.

Запишем итерационную форму нашего уравнения $x^{k+1} = S(x^k)$, и перейдем к пределу при $k \to \infty$. Левая часть x^{k+1}, как было уже показано, сходится к $x^* \in U_r(a)$, а правая часть $S(x^k)$ в силу Липшиц-непрерывности $S(x)$ сходится к $S(x^*)$. Таким образом, решение существует.

Покажем единственность найденного корня. Пусть существуют два решения: x^* и $\overline{x^*}$. Рассмотрим модуль их разности:

$$|x^* - \overline{x^*}| = |S(x^*) - S(\overline{x^*})| \leqslant q|x^* - \overline{x^*}|,$$

откуда вытекает, что $x^* = \overline{x^*}$, иначе возникает противоречие, так как $q < 1$.

В завершение докажем оценку на погрешность

$$|x^k - x^*| = |S(x^{k-1}) - S(x^*)| \leqslant q|x^{k-1} - x^*| \leqslant \cdots \leqslant q^k|x^0 - x^*|.$$

Теперь теорема полностью доказана. □

Замечание 1. В теореме получена оценка

$$\forall p\ |x^{k+p} - x^k| \leqslant \frac{q^k}{1-q}|x^1 - x^0|,$$

то есть, переходя к пределу при $p \to \infty$, $|x^* - x^k| \leqslant \dfrac{q^k}{1-q}|S(x^0) - x^0|$.

Потребуем, чтобы x^k отличалось от x^* не более, чем на ε. Так как

$$\frac{q^k}{1-q}|S(x^0) - x^0| \leqslant \varepsilon \implies |x^* - x^k| \leqslant \varepsilon,$$

то число итераций, необходимых для достижения такой точности, можно подсчитать так:

$$k(\varepsilon) = \left\lceil \frac{\ln \frac{(1-q)\varepsilon}{|S(x^0) - x^0|}}{\ln q} \right\rceil.$$

Замечание 2. В условиях теоремы сделано предположение, что $S(x)$ Липшиц-непрерывна:

$$|S(x') - S(x'')| \leqslant q|x' - x''| \quad \forall\, x',\, x'' \in U_r(a).$$

Это достаточно слабое ограничение, но его сложно проверять. Тем не менее, если функция дифференцируема, а ее производная ограничена той самой константой q, то условие Липшица будет выполнено:

$$|S(x') - S(x'')| = \{\text{применяя формулу Лагранжа}\} = |S'(\xi)| \cdot |x' - x''| \leqslant q|x' - x''|$$

— это и есть условие Липшиц-непрерывности.

3.4 Метод Эйткена

Пусть метод имеет линейную скорость сходимости, то есть для погрешности выполнена следующая оценка:

$$|x^k - x^*| \leqslant q^k |x^0 - x^*| \implies x^k - x^* \approx aq^k \tag{3.8}$$

— мы выразили погрешность через некоторую константу a и параметр q.

Допустим, что данное соотношение выполняется точно. Рассмотрим погрешность на трех соседних итерациях.

$$\begin{aligned} x^k - x^* &= aq^k; \\ x^{k+1} - x^* &= aq^{k+1}; \\ x^{k+2} - x^* &= aq^{k+2}. \end{aligned}$$

Из этих условий легко найти x^*. Для этого вычтем из второго уравнения первое, а из третьего второе, тогда получим:

$$\begin{aligned} x^{k+1} - x^k &= aq^k(q-1); \\ x^{k+2} - x^{k+1} &= aq^{k+1}(q-1). \end{aligned}$$

Вычтем одно уравнение из другого:

$$x^{k+2} - 2x^{k+1} + x^k = aq^k(q-1)^2.$$

Свяжем это уравнение с предыдущим следующим соотношением:

$$\frac{(x^{k+2} - x^{k+1})^2}{x^{k+2} - 2x^{k+1} + x^k} = \frac{a^2 q^{2k+2}(q-1)^2}{aq^k(q-1)^2} = aq^{k+2} = x^{k+2} - x^*.$$

Если бы равенство (3.8) выполнялось точно, то можно было бы получить x^* через три последних приближения:

$$x^* = x^{k+2} - \frac{(x^{k+2} - x^{k+1})^2}{x^{k+2} - 2x^{k+1} + x^k}.$$

Тем не менее, выражение, стоящее в правой части, приближает x^* намного лучше, чем x^{k+2}. Обозначим его

$$\overline{x^{k+2}} = x^{k+2} - \frac{(x^{k+2} - x^{k+1})^2}{x^{k+2} - 2x^{k+1} + x^k},$$

и будем считать это равенство алгоритмом построения подправленной итерационной последовательности с элементами $\overline{x^k}$.

Если эту операцию (вычисление подправленных значений) проводить достаточно часто, то новая последовательность из $\overline{x^k}$ сходится к точному решению значительно быстрее, чем исходный метод.

Построение подправленных значений последовательности называется **методом Эйткена**.

3.5 Сходимость метода Ньютона

Запишем итерационный процесс:

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}.$$

Он является модификацией метода простой итерации, тогда условием сходимости этого итерационного процесса будет неравенство

$$|S'(x)| \leqslant q < 1,$$

где $S(x) = x - \dfrac{f(x)}{f'(x)}$. $S'(x) = 1 - \dfrac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2}$, поэтому

$$|S'(x)| \leqslant q < 1 \iff \left| \frac{f(x)f''(x)}{(f'(x))^2} \right| \leqslant q < 1.$$

Но в силу того, что мы ищем корень уравнения $f(x) = 0$, существует такая окрестность, где $S'(x) \leqslant q < 1$, но в общем случае эта область будет мала, то есть нужно подбирать начальное приближение достаточно близко расположенным к корню.

Теорема 3.2 (о сходимости метода Ньютона)**.** *Пусть x^* — простой вещественный корень уравнения $f(x) = 0$, а функция $f(x)$ — дважды дифференцируема в некоторой окрестности $U_r(x^*)$, причем первая производная нигде не обращается в нуль.*

Тогда, следуя обозначениям

$$0 < m_1 = \inf_{x \in U_r(x^*)} |f'(x)|, \quad M_2 = \sup_{x \in U_r(x^*)} |f''(x)|,$$

при выборе начального приближения x^0 из той же окрестности $U_r(x^*)$ такого, что

$$\frac{M_2 |x^0 - x^*|}{2m_1} = q < 1,$$

итерационная последовательность

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)},\ k = 0,\ 1,\ \ldots$$

будет сходиться к x^*, причем для погрешности на k-м шаге будет справедлива оценка:

$$|x^k - x^*| \leqslant q^{2^k - 1} |x^0 - x^*|. \tag{3.9}$$

Доказательство. Основным этапом нашего доказательства будет получение x^{k+1} из x^k. Получим оценку погрешности на $(k+1)$-й итерации из определения итерационного процесса:

$$x^{k+1} - x^* = x^k - \frac{f(x^k)}{f'(x^k)} - x^* = \frac{(x^k - x^*) f'(x^k) - f(x^k)}{f'(x^k)} =$$
$$= \ \{\text{обозначим } \mathsf{F(x) = (x - x^*) f'(x) - f(x)}\} \ = \frac{F(x^k)}{f'(x^k)}.$$

Преобразуем $F(x^k)$ по формуле Ньютона-Лейбница:

$$F(x^k) = F(x^*) + \int_{x^*}^{x^k} F'(\xi) d\xi = \ \{\mathsf{F(x^*) = f(x^*) = 0}\} \ = \int_{x^*}^{x^k} F'(\xi) d\xi = \int_{x^*}^{x^k} (\xi - x^*) f''(\xi) d\xi.$$

Применив к последнему интегралу формулу среднего значения ($\xi_k \in [x^*;\, x^k]$), получим

$$F(x^k) = f''(\xi_k) \int_{x^*}^{x^k} (\xi - x^*) d\xi = f''(\xi_k) \frac{(x^k - x^*)^2}{2}.$$

Подставив запись для $F(x^k)$ в выражение для погрешности, получим

$$x^{k+1} - x^* = f''(\xi_k) \frac{(x^k - x^*)^2}{2 f'(x^k)}.$$

Так как вторая производная по модулю ограничена сверху, а первая — снизу, то из последнего равенства следует, что метод Ньютона имеет квадратичную скорость сходимости.

Докажем оценку на погрешность по индукции.

База индукции. Рассмотрим погрешность при $k = 1$:

$$x^1 - x^* = f''(\xi_0)\frac{|x^0 - x^*|^2}{2f'(x^0)} \leqslant \{\xi_0 \in \mathsf{U_r(x^*)}\} \leqslant \frac{M_2|x^0 - x^*|}{2m_1}|x^0 - x^*| = q|x^0 - x^*|.$$

Таким образом, база индукции верна.

Предположение индукции. Пусть оценка (3.9) выполняется для некоторого k :

$$|x^k - x^*| \leqslant q^{2^k-1}|x^0 - x^*|.$$

Индуктивный переход. Докажем, что она выполняется для $k + 1$. Согласно показанному ранее,

$$|x^{k+1} - x^*| = |f''(\xi^k)|\frac{|x^k - x^*|^2}{2|f'(x^k)|}.$$

$\xi_k \in U_r(x^*)$, так как ξ_k выбиралась в соответствии с теоремой о среднем, и поэтому принадлежит отрезку $U_r(x^*)$.

Получим верхнюю оценку, используя предположение индукции:

$$|f''(\xi^k)|\frac{|x^k - x^*|^2}{2|f'(x^k)|} \leqslant \frac{M_2(x^k - x^*)^2}{2m_1} \leqslant \frac{M_2}{2m_1}(q^{2^k-1})^2 \cdot |x^0 - x^*|^2.$$

Вспомним, что мы обозначали $\dfrac{M_2}{2m_1}|x^0 - x^*| = q$, тогда получим:

$$|x^{k+1} - x^*| \leqslant q^{2^{k+1}-1}|x^0 - x^*|.$$

Откуда следует, что оценка верна, и, следовательно, теорема доказана ($q < 1$ по предположению, и при $k \longrightarrow \infty$ правая часть стремится к нулю, а это значит, что последовательность сходится к x^*). □

Замечание.

1. В условии теоремы мы требуем, чтобы $\dfrac{M_2|x^0 - x^*|}{2m_1} < 1$. Но как это проверить, ведь мы не знаем точного решения x^*? Можно поступить так.

   Рассмотрим условие

   $$\frac{M_2|x^0 - x^*|}{2m_1} < 1. \tag{3.10}$$

   Распишем $f(x^0) = f(x^0) - f(x^*) = f'(\overline{x})(x^0 - x^*)$, тогда

   $$|x^0 - x^*| = \frac{|f(x^0)|}{|f'(\overline{x})|} \leqslant \frac{|f(x^0)|}{m_1}.$$

   Подставим эту запись для $|x^0 - x^*|$ в (3.10), тогда из неравенства

   $$\frac{M_2|f(x^0)|}{2m_1^2} < 1$$

   следует, что выполняется условие (3.10).

   Таким образом, зная m_1 и M_2, можно подбирать x^0, исходя из этого неравенства (подбираем достаточно малую окрестность на этапе локализации корней, и дальше работаем с ней; если окрестность велика, уточняем расположение корня).

2. В условии требовалось, чтобы x^* был простым вещественным корнем; если же x^* — корень кратности p, то метод Ньютона будет иметь квадратичную скорость сходимости, если некоторым образом подправить итерационную последовательность.

3.6 Решение систем нелинейных уравнений

Перейдем к поиску численных методов для решения систем нелинейных уравнений. Пусть имеется n уравнений следующего вида:

$$f_i(x_1, x_2, \ldots, x_n) = 0, \quad i = \overline{1, n}. \tag{3.11}$$

Поиск решений данной системы обычно проводится в два этапа: сначала проходит **разделение решений** (термин имеет тот же смысл, что и локализация корней при решении одиночных уравнений), а затем на полученных участках производится их уточнение. Как мы увидим, для поиска решений систем будут применяться те же методы, что и для поиска корней одиночных уравнений, но с небольшими модификациями.

Метод простой итерации

Итак, мы решаем систему вида (3.11). Построим итерационный процесс для нахождения точного решения (обозначим его x^*) на отрезке $[a; b]$. Для этого потребуем, чтобы оно являлось решением такой системы уравнений:

$$x_i = S_i(x_1, x_2, \ldots, x_n), \quad i = \overline{1, n}, \tag{3.12}$$

где S_i — некоторая функция. Будем задавать $S_i(x)$ в виде:

$$S_i(x_1, x_2, \ldots, x_n) = x_i - \tau_i(x_1, x_2, \ldots, x_n) f_i(x_1, x_2, \ldots, x_n).$$

где τ_i — функция-параметр, не обращающаяся в нуль в некоторой окрестности x^*. Легко проверить, что в данном случае x^* будет решением (3.12). Теперь зададим итерационный процесс следующим образом (x_l^k — координаты вектора x^k, который должен будет сходиться к решению):

$$x_l^{k+1} = S_l(x_1^k, x_2^k, \ldots, x_n^k), \quad k = 0, 1, \ldots$$

При этом задается вектор x^0 — начальное приближение.

Итак, мы ожидаем сходимость:

$$\left\{x^k\right\} = \left\{(x_1^k, x_2^k, \ldots, x_n^k)^T\right\} \overset{k\to\infty}{\longrightarrow} x^*.$$

Для одиночных уравнений достаточным условием сходимости было выполнение неравенства

$$|S'(x)| < 1$$

всюду на рассматриваемом отрезке. Посмотрим, что нам потребуется в данном случае для уменьшения погрешности на каждом шаге:

$$z_i^{k+1} = x_i^{k+1} - x_i^* = S_i(x_1^k, x_2^k, \ldots, x_n^k) - S_i(x_1^*, x_2^*, \ldots, x_n^*).$$

Пусть у функций S_i существуют первые производные. Тогда мы можем применить обобщенную формулу Лагранжа и получим:

$$S_i(x_1^k, x_2^k, \ldots, x_n^k) - S_i(x_1^*, x_2^*, \ldots, x_n^*) = \sum_{l=1}^{n} \frac{\partial S_i(\xi_1^k, \xi_2^k, \ldots, \xi_n^k)}{\partial x_l}(x_l^k - x_l^*).$$

Соберем все производные в одну матрицу: $A^k = (a_{ij}^k) = \left(\dfrac{\partial S_i(\xi_1^k, \xi_2^k, \ldots, \xi_n^k)}{\partial x_j}\right)$. Тогда предыдущая формула может быть переписана так:

$$z^{k+1} = A^k z^k.$$

Из этого следует, что $||z^{k+1}|| = ||A^k z^k|| \leqslant ||A^k|| \cdot ||z^k||$. Таким образом, если для всех k на нашем отрезке $[a;\, b]$ выполняется неравенство

$$||A^k|| \leqslant q < 1, \tag{3.13}$$

то последовательность $||z^k||$ будет сходится по норме. Этого нам будет достаточно.

Заметим, что условие (3.13) будет выполнено, если матрица $A = \left(\max\limits_{[a;\, b]} \left|\dfrac{\partial S_i}{\partial x_j}\right|\right)$ будет иметь норму, меньшую единицы. Тогда, очевидно, $||A^k|| \leqslant ||A||$, и процесс будет сходиться. Такое требование можно удовлетворить, подбирая параметры τ_i. В частности, связывая их с производными функций $f_i(x)$, можно получить метод Ньютона.

Метод Ньютона

Здесь мы требуем от функций $f_i(x_1,\, x_2,\, \ldots,\, x_n)$ существование первых производных. Согласно определению x^*,

$$f_i(x_1^*,\, x_2^*,\, \ldots,\, x_n^*) = 0, \quad i = \overline{1,\, n}. \tag{3.14}$$

Теперь с помощью преобразований этого тождества получим формулу для итерационного процесса. Для этого зададимся k-м приближением $x^k = (x_1^k,\, x_2^k,\, \ldots,\, x_n^k)^T$ и зафиксируем некоторое $i \in [1;\, n]$. Заметим, что (3.14) можно переписать так:

$$f_i(x_1^k + (x_1^* - x_1^k),\, x_2^k + (x_2^* - x_2^k),\, \ldots,\, x_n^k + (x_n^* - x_n^k)) = 0.$$

Применив обобщенную формулу Лагранжа, получим:

$$f_i(x_1^k,\, x_2^k,\, \ldots,\, x_n^k) + \sum_{l=1}^{n} \frac{\partial f_i(\xi_1^k,\, \xi_2^k,\, \ldots,\, \xi_n^k)}{\partial x_l}(x_l^* - x_l^k) = 0.$$

Теперь, заменив ξ^k на x^k, а x_l^* — на новое приближение x_l^{k+1}, получим такую формулу для поиска x_l^{k+1} :

$$f_i(x_1^k,\, x_2^k,\, \ldots,\, x_n^k) + \sum_{l=1}^{n} \frac{\partial f_i(x_1^k,\, x_2^k,\, \ldots,\, x_n^k)}{\partial x_l}(x_l^{k+1} - x_l^k) = 0. \tag{3.15}$$

Обозначив $\Delta x_l^k = x_l^{k+1} - x_l^k$, перепишем ее так:

$$\sum_{l=1}^{n} \frac{\partial f_i(x_1^k,\, x_2^k,\, \ldots,\, x_n^k)}{\partial x_l}\Delta x_l^k = -f_i(x_1^k,\, x_2^k,\, \ldots,\, x_n^k).$$

Это система линейных уравнений для поиска Δx^k вида $A\Delta x^k = -f$. Решая ее, мы находим Δx^k, а затем и $x_l^{k+1} = x_l^k + \Delta x_l^k$ для $l = \overline{1,\, n}$.

Теорему о сходимости данного метода мы напишем неформально и примем без доказательства.

Теорема 3.3. *В достаточно малой окрестности искомого корня итерационный процесс по методу Ньютона (задаваемый формулой (3.15)) сходится, если определитель матрицы A не обращается в этой окрестности в нуль. При этом скорость сходимости — квадратичная.*

Замечание. Данный итерационный процесс сходится быстро: для достижения неравенства (обычный для метода Ньютона критерий завершения ИП)

$$||x^{k+1} - x^k|| < \varepsilon$$

при $\varepsilon = 10^{-5}$ достаточно всего 3-5 итераций. Однако он требует большого объема вычислений — ведь на каждом шаге нам приходится решать систему уравнений.

Примеры решения нелинейных уравнений

Пример 3.1. (Метод простой итерации)

Пусть $f(x) = x^3 - x - 1$. Найдем корень уравнения $x^3 - x - 1 = 0$ на отрезке $[-2; 3]$. Для начала проведем локализацию корней: введем на отрезке сетку и посчитаем значение функции в ее узлах:

x_i	-2	-1	0	1	2	3
$f(x_i)$	-7	-1	-1	-1	5	23

Так как $f(1)f(2) < 0$, а наша функция непрерывна, то на отрезке $[1; 2]$ обязательно есть корень уравнения. Будем искать его методом простой итерации.

1. Как уже говорилось, мы ставим в соответствие уравнению $f(x) = 0$ уравнение $x = S(x)$, где $S(x)$ наиболее часто берется в виде $S(x) = x - \tau(x)f(x)$. Выберем $\tau(x) = \tau > 0$. Тогда

$$S(x) = x - \tau(x^3 - x - 1).$$

Соответственно, итерационный процесс задается так:

$$x^{k+1} = S(x^k) = x^k - \tau((x^k)^3 - x^k - 1).$$

Для его сходимости, согласно замечанию к теореме 3.1, нам было достаточно выполнения неравенства $|S'(x)| < 1$ на $[1; 2]$. Посмотрим, какие ограничения это условие даст на τ :

$$|S'(x)| < 1 \iff |1 - \tau(3x^2 - 1)| < 1 \iff$$

$$-1 < 1 - \tau(3x^2 - 1) < 1. \tag{3.16}$$

При $x \in [1; 2]$ и для любого положительного τ правое неравенство в (3.16) верно всегда. Левое неравенство перепишется так:

$$\tau < \frac{2}{3x^2 - 1}.$$

При $x \in [1; 2]$ знаменатель дроби достигает минимума при $x = 2$. Отсюда итоговое ограничение на τ таково:

$$\tau < \frac{2}{3 \cdot 2^2 - 1} = \frac{2}{11}.$$

Возьмем $\tau = \dfrac{1}{11}$. В этом случае $S(x) = x - \dfrac{1}{11}(x^3 - x - 1)$, $\max\limits_{x \in [1; 2]} |S'(x)| = \dfrac{9}{11}$. Это число — знаменатель геометрической прогрессии, характеризующей скорость сходимости процесса. Как видно, оно не очень мало, и сходится все медленно. Насколько медленно, можно понять из таблицы первых приближений:

k	x^k	$x^{k+1} = S(x^k)$
0	1.1	1.16991
1	1.16991	1.22161
2	1.22161	1.25784
3	1.25784	1.28218
4	1.28218	1.29802
5	1.29802	1.30812

— в данном примере $x^* \approx 1.32472$.

2. Можно по-разному выбирать функцию $S(x)$:

$$x^3 - x - 1 = 0 \implies x = \sqrt[3]{x+1}.$$

— и можно взять $S(x) = \sqrt[3]{x+1}$. При этом скорость сходимости будет выше, так как

$$S'(x) = \frac{1}{3(x+1)^{\frac{2}{3}}} < \frac{1}{3(1+1)^{\frac{2}{3}}} \approx 0.2 = q.$$

3. С другой стороны,

$$x^3 - x - 1 = 0 \iff x = x^3 - 1.$$

Однако, если взять $S(x) = x^3 - 1$, то $S'(x) = 3x^2 > 1$ на $[1;2]$, что противоречит достаточному условию сходимости.

Пример 3.2. (Метод простой итерации и метод Ньютона)
Будем искать корень уравнения $x^3 - 1 = 0$. Зная, что $x^* = 1$ подходит, не будем утруждать себя локализацией корней, а просто покажем, как проходят итерационные процессы различных методов.

Метод простой итерации. Возьмем отрезок $[\frac{2}{3}, \frac{4}{3}]$ и зададим на нем итерационный процесс:

$$x^{k+1} = S(x^k),$$

где зададим $S(x) = x - \tau(x^3 - 1)$. Рассмотрим, какие ограничения накладываются на τ в этом примере:

$$S'(x) = 1 - 3\tau x^2.$$

Потребуем выполнение достаточного условия сходимости:

$$|S'(x)| < 1 \iff -1 < 1 - 3\tau x^2 < 1.$$

Отсюда ограничение на τ таково:

$$0 < \tau < \left.\frac{2}{3x^2}\right|_{x=\frac{4}{3}} = \frac{3}{8}.$$

Возьмем $\tau = 0.25$, тогда $S(x) = x - \dfrac{x^3 - 1}{4}$. Распишем первые несколько шагов итерационного процесса, взяв $x^0 = 1.1$:

k	x^k	$x^k - x^*$
0	1.1	0.1
1	1.01725	0.01725
2	1.00409	0.00409
3	1.00101	0.00101
4	1.00025	0.00025
5	1.00006	0.00006

Видно, что скорость сходимости не очень высокая. Однако на третьем и четвертом шаге уже можно применить коррекцию Эйткена:

$$\overline{x^{k+2}} = x^{k+2} - \frac{(x^{k+2} - x^{k+1})^2}{x^{k+2} - 2x^{k+1} + x^k}.$$

Тогда получим такую таблицу:

k	x^k	$x^k - x^*$
0	1.1	0.1
1	1.01725	0.01725
2	1.00409	0.00409
3	1.000068	0.000068
4	1.000001	0.000001

Видно, что точность существенно возросла.

Метод Ньютона. Согласно канонической форме метода Ньютона,

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}.$$

Зная, что $f(x) = x^3 - 1$, а $f'(x) = 3x^2$, получим такую формулу:

$$x^{k+1} = x^k - \frac{(x^k)^3 - 1}{3(x^k)^2}.$$

Таблица приближений такова:

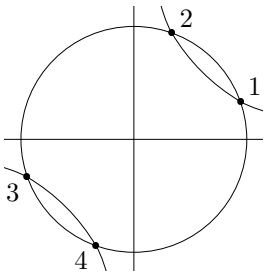
k	x^k	$x^k - x^*$
0	1.1	0.1
1	1.00882	0.00882
2	1.00008	0.00008
3	1.000000006	0.000000006

Здесь уже заметно, что скорость сходимости — квадратичная.

Пример 3.3. (Метод Ньютона для систем уравнений)
Рассмотрим следующую систему нелинейных уравнений:

$$\begin{cases} F(x,\, y) &= x^2 + y^2 - 4 = 0; \\ G(x,\, y) &= xy - 1 = 0. \end{cases} \tag{3.17}$$

На первом этапе (разделение корней) получаем 4 корня. Это можно показать, нарисовав графики соответствующих кривых.



Возьмем начальное приближение для поиска одного из корней таким:

$$x^0 = 2,\ y^0 = 0.$$

Решать систему будем методом Ньютона. Линеаризованные уравнения имеют вид:

$$\begin{aligned} \frac{\partial F}{\partial x}(x^k,\, y^k)\Delta x^k + \frac{\partial F}{\partial y}(x^k,\, y^k)\Delta y^k &= -F(x^k,\, y^k); \\ \frac{\partial G}{\partial x}(x^k,\, y^k)\Delta x^k + \frac{\partial G}{\partial y}(x^k,\, y^k)\Delta y^k &= -G(x^k,\, y^k), \end{aligned} \qquad k = 0,\, 1,\, \ldots$$

Учитывая вид системы (3.17), можно найти частные производные:

$$\frac{\partial F}{\partial x}=2x, \quad \frac{\partial F}{\partial y}=2y;$$
$$\frac{\partial G}{\partial x}=y, \quad \frac{\partial G}{\partial y}=x.$$

Подставив их в (3.17), можно найти приращения, решив систему:

$$\left\{ \begin{array}{rcl} 2x^k\Delta x^k + 2y^k\Delta y^k & = & 4-(x^k)^2-(y^k)^2; \\ y^k\Delta x^k + x^k\Delta y^k & = & 1-x^k y^k. \end{array} \right.$$

и подставить в определение итерационного процесса:

$$\left\{ \begin{array}{rcl} x^{k+1} & = & x^k+\Delta x^k; \\ y^{k+1} & = & y^k+\Delta y^k. \end{array} \right.$$

Получим следующую последовательность итерационных приближений:

N	x^k	y^k	F	G	Δx^k	Δy^k
1	2	0	0	-1	0	0.5
2	2	0.5	0.25	0	$-\frac{1}{15}$	$\frac{1}{60}$
3	1.93	0.517	-0.0077	-0.0022	$\ldots$	$\ldots$

Как видно из приведенной таблицы, процесс очень быстро сходится. Об этом можно судить по величине функций $F(x,\,y)$ и $G(x,\,y)$, так как мы решаем уравнения $F(x,\,y)=0$ и $G(x,\,y)=0$.

Как же работает метод Ньютона? Дадим геометрическую интерпретацию этого метода. В каждой текущей точке $(x^k,\,y^k)$ к поверхностям $z=F(x,\,y)$ и $z=G(x,\,y)$ строятся касательные плоскости. Потом рассматриваем линии пересечения этих плоскостей с плоскостью $z=0$ — это две прямые. И в заключение, определяем точку пересечения полученных прямых как новое, $k+1$ значение.

Глава 4

Интерполяция и приближение функций

В этой главе мы будем восстанавливать функцию по значениям в некоторых заданных точках.

Итак, пусть на отрезке $[a; b]$ задан набор точек ($n + 1$ точка). Эти точки называют **узлами интерполяции**. Занумеруем их в следующем порядке $a = x_0 < x_1 < \ldots < x_n = b$, и пусть в каждой из этих точек известно $f(x_k) = f_k,\ k = \overline{0, n}$. Наша задача заключается в том, чтобы вычислить приближенные (с некоторой точностью) значения функции между значениями в узлах интерполяции. Опираясь на эти сведения, построим на том же отрезке функцию $\Phi(x)$, которую назовем **интерполянтой** — она и будет служить приближением исходной функции. Не любая $\Phi(x)$ нам подойдет — потребуем, чтобы она была легко вычислима, и совпадала с исходной функцией в узлах интерполяции: $\Phi(x_k) = f(x_k),\ k = \overline{0, n}$. Иногда, правда, от последнего условия отказываются — тогда говорят о построении **наилучшего приближения**. Впрочем, вся терминология будет объясняться по ходу дела.

Построим интерполянту. Введем **базисные функции** $\varphi_i(x),\ i = \overline{0, m}$ — линейно независимые элементы в нашем пространстве функций. $\Phi(x)$ будем строить как линейную комбинацию базисных функций $\varphi_i(x)$:

$$\Phi(x) = \sum_{i=0}^{m} a_i \varphi_i(x).$$

В качестве базисных функций мы можем выбрать следующие:

1. Степенные функции: $\varphi_i(x) = x^i$;
2. Тригонометрические функции: в случае интерполяции периодических функций с периодом, к примеру, $2l$ лучше взять $\sin \frac{\pi i x}{l}$ и $\cos \frac{\pi i x}{l}$;
3. Дробно-полиномиальные функции.

Интерполяция алгебраическими многочленами

Одним из наших первых требований было совпадение значений интерполянты и исходной функции в узлах сетки. Записав интерполянту через базисные функции (пусть их будет $n + 1$ — столько же, сколько узлов), получим такие уравнения:

$$\sum_{i=0}^{n} a_i \varphi_i(x_k) = f(x_k),\ k = \overline{0, n}. \tag{4.1}$$

Относительно a_i мы получили СЛАУ с $n+1$ неизвестными, так как $\varphi_i(x_k)$ и $f(x_k)$ заданы. Условием разрешимости в данном случае является то, что определитель системы (4.1) отличен от нуля. При этом решение будет существовать и оно будет единственным. Определитель матрицы системы будет выглядеть так:

$$\begin{vmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \dots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \dots & \varphi_n(x_1) \\ & \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot & & \\ \varphi_0(x_n) & \varphi_1(x_n) & \dots & \varphi_n(x_n) \end{vmatrix} \tag{4.2}$$

Будем брать в качестве базисных функций степенные. Тогда (4.2) — это определитель Вандермонда:

$$\begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ & \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot & & \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} = (x_1 - x_0)(x_2 - x_0)\dots(x_n - x_0)(x_2 - x_1)(x_3 - x_1)\dots(x_n - x_{n-1}) =$$

$$= \prod_{0 \leqslant k < m \leqslant n} (x_m - x_k) \neq 0, \text{ если } x_i \neq x_j \quad \forall\, i \neq j.$$

Откуда следует необходимое и достаточное условие, накладываемое на систему (4.1): все узлы интерполяции должны быть различными, тогда решение существует и единственно.

Оно даст коэффициенты a_i, и мы получим выражение для $\Phi(x)$.

4.2 Наилучшее приближение табличной функции

В предыдущих разделах были рассмотрены примеры интерполяции функции $f(x)$ многочленами Лагранжа и сплайнами. Интерполянта

$$\Phi(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + \dots + a_m\varphi_m(x).$$

обычно содержала $m+1$ неизвестный коэффициент, которые определялись из $n+1$ условия совпадения с табличными значениями ($m = n$ при интерполяции многочленами и $m > n$ — сплайнами).

Теперь обсудим случай, когда $m < n$ (узлов, в которых известно значение функции, больше, чем неизвестных коэффициентов). В этом случае возникает понятие **наилучшего приближения**: мы отказываемся от требования совпадения значения интерполянты и табличной функции, а требуем минимизировать некоторый функционал от, к примеру, вектора погрешностей в узлах сетки:

$$\vec{r} = \left(\Phi(x_0) - f(x_0),\, \Phi(x_1) - f(x_1),\, \dots,\, \Phi(x_n) - f(x_n)\right),$$

где $x_0, \dots, x_n$ — точки, в которых нам задана функция $f(x)$. В качестве функционала рассмотрим норму этого вектора, которую можно задавать по-разному:

$$\begin{aligned} ||\vec{r}||_{(1)} &= \left(\sum_{i=0}^{n} (\Phi(x_i) - f(x_i))^2\right)^{\frac{1}{2}}; \\ ||\vec{r}||_{(2)} &= \max_i |\Phi(x_i) - f(x_i)|. \end{aligned}$$

В каждом случае мы пытаемся подобрать коэффициенты a_i в задании $\Phi(x)$ так, чтобы минимизировать эту норму. Эту задачу называют **поиском наилучшего среднеквадратичного приближения** (а метод ее решения — **методом наименьших квадратов**) или **поиском наименьшего равномерного приближения** соответственно. Приведем пример решения такой задачи.

Наилучшее приближение в гильбертовом пространстве

Пусть $\mathbb{H}$ — евклидово пространство функций с нормой $||f|| = \sqrt{\langle f,\, f\rangle}$, а $\varphi_i\,(i = \overline{0,\, n})$ — его линейно независимые элементы. Нашей задачей будет поиск наилучшего приближения

$$\varphi = c_0\varphi_0 + \ldots + c_n\varphi_n, \quad c_i \in \mathbb{R},$$

для элемента $f \in \mathbb{H}$. Оценкой точности будет служить величина погрешности $||f - \varphi||$.

Определение. Элемент $\widetilde{\varphi}$, доставляющий минимум этой норме (для которого верно равенство $\min\limits_{\varphi} ||f - \varphi|| = ||f - \widetilde{\varphi}||$), называется **элементом наилучшего приближения**.

Покажем, что такой элемент существует и единствен. В качестве примера можно взять пространство $L_2[a;\, b]$ (пространство функций, интегрируемых со своими квадратами) и в нем некоторую функцию f. В качестве скалярного произведения берут обычное в L_2 :

$$\langle g,\, f\rangle_{L_2} = \int\limits_a^b f(x)g(x)\,dx, \quad \text{соответственно } ||f||_{L_2} = \left(\int\limits_a^b f^2(x)\,dx\right)^{\frac{1}{2}}.$$

Продолжим рассуждения в общем виде. Наша задача — это минимизировать $||f - \varphi||$, подобрав соответствующую функцию φ. Приведем эту норму (будем работать с ее квадратом) к более удобному виду:

$$\begin{aligned}
||f-\varphi||^2 &= \langle f-\varphi,\, f-\varphi\rangle = \left\langle f - \sum_{l=0}^{n} c_l\varphi_l,\, f - \sum_{k=0}^{n} c_k\varphi_k \right\rangle = \\
&= \langle f,\, f\rangle - \sum_{l=0}^{n} c_l\,\langle \varphi_l,\, f\rangle - \sum_{k=0}^{n} c_k\,\langle \varphi_k,\, f\rangle + \sum_{l=0}^{n}\sum_{k=0}^{n} c_l c_k\,\langle \varphi_l,\, \varphi_k\rangle = \\
&= ||f||^2 - 2\sum_{l=0}^{n} c_l\,\langle \varphi_l,\, f\rangle + \sum_{l=0}^{n}\sum_{k=0}^{n} c_l c_k\,\langle \varphi_l,\, \varphi_k\rangle = \\
&= \{\text{ Введем обозначения } \begin{cases} f_l &= \langle \varphi_l,\, f\rangle = \int\limits_a^b f(x)\varphi_l(x)\,dx; \\ a_{kl} &= \langle \varphi_k,\, \varphi_l\rangle = \int\limits_a^b \varphi_k(x)\varphi_l(x)\,dx. \end{cases} \} = \\
&= ||f||^2 - 2\sum_{l=0}^{n} c_l f_l + \sum_{l=0}^{n}\sum_{k=0}^{n} c_k c_l a_{kl} = \\
&= \{\text{ Обозначим } \begin{cases} \overline{c} &= (c_0,\, c_1,\, \ldots,\, c_n)^T; \\ \overline{f} &= (f_0,\, f_1,\, \ldots,\, f_n)^T; \\ A &= (a_{kl}). \end{cases} \} = \\
&= ||f||^2 - 2\,\langle \overline{c},\, \overline{f}\rangle + \langle A\overline{c},\, \overline{c}\rangle = ||f||^2 + F(\overline{c}).
\end{aligned}$$

Соединив первое и последнее равенство, получим:

$$||f - \varphi||^2 = ||f||^2 + F(\overline{c}). \tag{4.18}$$

Таким образом, задача о минимизации $||f - \varphi||$ свелась к задаче минимизации функции $F(\overline{c})$ от вектора переменных $\overline{c}$.

Из определения матрицы A следует, что она симметрична. Покажем, что она положительно определена, то есть

$$\forall \overline{c} \neq 0 \quad \langle A\overline{c},\, \overline{c}\rangle > 0.$$

Если взять в равенстве (4.18) $f \equiv 0$, то получим, что $\langle A\overline{c}, \overline{c}\rangle = ||\varphi||^2 > 0$. Предположим, что существует вектор $\overline{y} \neq 0$ такой, что $\langle A\overline{y}, \overline{y}\rangle = 0$. Но это будет означать, что $||\varphi|| = 0$. Так как φ — линейная комбинация линейно независимых элементов φ_i, то это возможно тогда и только тогда, когда эта комбинация тривиальна — то есть $y_i = 0$ для всех i. Отсюда делаем вывод, что

$$\overline{y} = 0 \implies \forall \overline{c} \neq 0 \quad \langle A\overline{c}, \overline{c}\rangle > 0.$$

Таким образом, матрица A положительно определена. Это позволяет воспользоваться следующей теоремой.

Теорема 4.4. *Пусть даны матрица A такая, что $A = A^T > 0$, и $\overline{f}$ — некоторый вектор (соответствующей размерности). Тогда у функции*

$$F(\overline{c}) = \langle A\overline{c}, \overline{c}\rangle - 2\left\langle \overline{f}, \overline{c}\right\rangle$$

векторного переменного $\overline{c}$ точка минимума существует и единственна, причем элемент $\overline{c}$ реализует этот минимум тогда и только тогда, когда он является решением системы:

$$A\overline{c} = \overline{f}. \tag{4.19}$$

Доказательство. Сначала докажем утверждение об эквивалентности.

Достаточность. Пусть $\overline{c}$ — решение системы (4.19). Покажем, что для любого ненулевого вектора $\overline{v}$ $F(\overline{c} + \overline{v}) > F(\overline{c})$:

$$F(\overline{c}+\overline{v}) = \langle A(\overline{c}+\overline{v}), \overline{c}+\overline{v}\rangle - 2\left\langle \overline{f}, \overline{c}+\overline{v}\right\rangle = \langle A\overline{c}, \overline{c}\rangle + \langle A\overline{c}, \overline{v}\rangle + \langle A\overline{v}, \overline{c}\rangle + \langle A\overline{v}, \overline{v}\rangle - 2\left\langle \overline{f}, \overline{c}\right\rangle - 2\left\langle \overline{f}, \overline{v}\right\rangle =$$
$$= \{\mathsf{A} = \mathsf{A}^{\mathsf{T}} \Longrightarrow \langle \mathsf{A}\overline{\mathsf{c}}, \overline{\mathsf{v}}\rangle = \langle \mathsf{A}\overline{\mathsf{v}}, \overline{\mathsf{c}}\rangle\} = F(\overline{c}) + \langle A\overline{v}, \overline{v}\rangle + 2\left\langle \overline{v}, A\overline{c} - \overline{f}\right\rangle = F(\overline{c}) + \langle A\overline{v}, \overline{v}\rangle .$$

Так как $A > 0$, то $\langle A\overline{v}, \overline{v}\rangle > 0$. Это означает, что $\overline{c}$ — точка минимума . Достаточность доказана.

Необходимость. Пусть $\overline{c}$ — точка минимума $F(\overline{c})$. Фиксируем произвольный ненулевой вектор $\overline{y}$ и положим $\overline{v} = \lambda\overline{y}$ (λ — параметр). Тогда, согласно проведенным в доказательстве достаточности преобразованиям,

$$F(\overline{c}+\overline{v}) = F(\overline{c}) + \langle A\overline{v}, \overline{v}\rangle + 2\left\langle \overline{v}, A\overline{c} - \overline{f}\right\rangle = F(\overline{c}) + \lambda^2\langle A\overline{y}, \overline{y}\rangle + 2\lambda\left\langle \overline{y}, A\overline{c} - \overline{f}\right\rangle . \tag{4.20}$$

Обозначим выражение, стоящее в правой части равенства, за $g(\lambda)$. Из принятых условий следует, что в точке $\lambda = 0$ функция $g(\lambda)$ достигает минимума. Она, очевидно, дифференцируема, поэтому $g'(0) = 0$. Продифференцировав (4.20) по λ, получим, что

$$\left(2\lambda\langle A\overline{y}, \overline{y}\rangle + 2\left\langle \overline{y}, A\overline{c} - \overline{f}\right\rangle\right)\Big|_{\lambda=0} = 0 \quad \Longleftrightarrow \quad \left\langle A\overline{c} - \overline{f}, \overline{y}\right\rangle = 0.$$

Из произвольности выбора $\overline{y}$ следует, что $A\overline{c} - \overline{f} = 0$. Это означает, что $\overline{c}$ — решение системы (4.19). Необходимость доказана.

Осталось заметить, что из доказанной эквивалентности следует существование и единственность точки минимума функции $F(\overline{c})$. Это вытекает из того, что матрица A положительно определена, поэтому система (4.19) имеет единственное решение. Теорема полностью доказана. □

Алгоритм построения наилучшего приближения

Опираясь на доказанную теорему, можно построить алгоритм нахождения наилучшего приближения для функции $f \in L_2[a;\, b]$. Он будет выглядеть так:

1. Выбираем $(n+1)$ линейно независимый элемент $\varphi_k,\ k = \overline{0,\, n}$ из $L_2[a;\, b]$.
2. Строим матрицу $A = (a_{kl})$ скалярных произведений:
$$a_{kl} = \int\limits_a^b \varphi_k(x)\varphi_l(x)\,dx.$$
3. Готовим вектор скалярных произведений $\overline{f} = (f_0,\, f_1,\, \ldots,\, f_n)$, где f_i находятся так:
$$f_i = \int\limits_a^b f(x)\varphi_i(x)\,dx, \quad i = \overline{0,\, n}.$$
4. Ищем вектор коэффициентов $\overline{c} = (\widetilde{c}_0,\, \widetilde{c}_1,\, \ldots,\, \widetilde{c}_n)$, решая систему уравнений $A\overline{c} = \overline{f}$.
5. Строим элемент $\widetilde{\varphi}$:
$$\widetilde{\varphi} = \widetilde{c}_0\varphi_0 + \widetilde{c}_1\varphi_1 + \ldots + \widetilde{c}_n\varphi_n.$$
Он будет являться наилучшим приближением согласно доказанной теореме.

Теперь посмотрим, насколько точно $\widetilde{\varphi}$ приближает f, то есть оценим $||f - \widetilde{\varphi}||$. Для этого нам понадобится лемма.

Лемма. *Пусть $\widetilde{\varphi}$ — элемент наилучшего приближения для f. Тогда*
$$\langle f - \widetilde{\varphi},\, \widetilde{\varphi}\rangle = 0$$
— то есть $\widetilde{\varphi}$ ортогонален $(f - \widetilde{\varphi})$.

Доказательство. Подставим представление $\widetilde{\varphi}$ через φ_k в искомое скалярное произведение:
$$\langle f - \widetilde{\varphi},\, \widetilde{\varphi}\rangle = \left\langle f - \sum_{k=0}^{n} \widetilde{c}_k\varphi_k,\, \sum_{l=0}^{n} \widetilde{c}_l\varphi_l \right\rangle = \sum_{l=0}^{n} \widetilde{c}_l \langle f,\, \varphi_l\rangle - \sum_{k=0}^{n}\sum_{l=0}^{n} \widetilde{c}_k\widetilde{c}_l \langle \varphi_k,\, \varphi_l\rangle =$$
$$= \sum_{l=0}^{n} \widetilde{c}_l f_l - \sum_{k=0}^{n}\sum_{l=0}^{n} \widetilde{c}_k\widetilde{c}_l a_{kl} = \{\text{согласно старым обозначениям }\} = \left\langle \overline{c},\, \overline{f}\right\rangle - \langle A\overline{c},\, \overline{c}\rangle = \left\langle \overline{f} - A\overline{c},\, \overline{c}\right\rangle.$$

Согласно построению $\widetilde{\varphi},\ \overline{c}$ находился из условия $A\overline{c} = \overline{f}$, поэтому получаем, что
$$\langle f - \widetilde{\varphi},\, \widetilde{\varphi}\rangle = 0.$$

Лемма доказана. □

Теперь можно оценить отклонение:

$$||f-\widetilde{\varphi}||^2 = \langle f-\widetilde{\varphi},\, f-\widetilde{\varphi}\rangle = \langle f-\widetilde{\varphi},\, f\rangle - \langle f-\widetilde{\varphi},\, \widetilde{\varphi}\rangle = \langle f-\widetilde{\varphi},\, f\rangle = \langle f,\, f\rangle - \langle \widetilde{\varphi},\, f\rangle\,.$$

Согласно лемме, $\langle f,\, \widetilde{\varphi}\rangle = \langle \widetilde{\varphi},\, \widetilde{\varphi}\rangle$, поэтому

$$||f-\widetilde{\varphi}||^2 = \langle f,\, f\rangle - \langle \widetilde{\varphi},\, \widetilde{\varphi}\rangle = ||f||^2 - ||\widetilde{\varphi}||^2.$$

Замечание. Если $\{\varphi_k\}$ — ортонормированная система, то есть $\langle \widetilde{\varphi}_k,\, \widetilde{\varphi}_l\rangle = \delta_{kl}$, тогда $A = E$. Отсюда следует, что $c_k = \langle f,\, \varphi_k\rangle = f_k$, и для наилучшего приближения получается простая формула:

$$\widetilde{\varphi} = \sum_{i=0}^{n} f_k\varphi_k.$$

В этом случае коэффициенты c_k называются **коэффициентами Фурье**, а построенный элемент $\widetilde{\varphi}$ — **многочленом Фурье**.

Пример 4.3.

Построим для функции из предыдущего примера наилучшее приближение. Итак, $f(x)$ задана табличо в точках x_0, $x_1 = x_0 + h$, $x_2 = x_0 + 2h$. Обозначим $F_0 = f(x_0)$, $F_1 = f(x_1)$, $F_2 = f(x_2)$.

Приближать будем снова многочленами:

$$\varphi_0(x) = 1, \quad \varphi_1(x) = x - x_1.$$

Напомним, что, применяя метод наименьших квадратов, мы нашли такое приближение:

$$\Phi(x) = \frac{F_0+F_1+F_2}{3} + \frac{F_2-F_0}{2h}(x-x_1)$$

Для поиска наилучшего приближения для f последовательно пройдем по построенному нами алгоритму:

1. $\varphi_0(x) = 1, \quad \varphi_1(x) = x - x_1.$

2. Подсчитаем $a_{kl} = \int\limits_{x_1-h}^{x_1+h} \varphi_k(x)\varphi_l(x)\,dx$:

$$\begin{aligned} a_{00} &= 2h; \\ a_{10} &= a_{01} = \int\limits_{x_1-h}^{x_1+h} (x-x_1)\,dx = \frac{(x-x_1)^2}{2}\Big|_{x_1-h}^{x_1+h} = 0; \\ a_{11} &= \int\limits_{x_1-h}^{x_1+h} (x-x_1)^2\,dx = \frac{(x-x_1)^3}{3}\Big|_{x_1-h}^{x_1+h} = \frac{h^3}{3} + \frac{h^3}{3} = \frac{2h^3}{3}. \end{aligned}$$

$$\Longrightarrow A = \begin{pmatrix} 2h & 0 \\ 0 & \dfrac{2h^3}{3} \end{pmatrix}.$$

3. На данном этапе возникают сложности, так как, не зная $f(x)$, мы не можем точно вычислить f_0 и f_1. Будем вычислять их приближенно: f_0 через формулу среднего значения, а f_1 — по

формуле Симпсона (как известно, она дает маленькую погрешность):

$$\begin{aligned} f_0 &= \int\limits_{x_1-h}^{x_1+h} f(x)\,dx = f(\overline{x})\cdot 2h \approx 2h\frac{F_0+F_1+F_2}{3}; \\ f_1 &= \int\limits_{x_1-h}^{x_1+h} f(x)(x-x_1)\,dx \approx \{\int\limits_{\mathsf{a}}^{\mathsf{b}} \mathsf{G(x)\,dx} \approx \frac{\mathsf{b-a}}{\mathsf{6}}(\mathsf{G_0+4G_1+G_2})\} \approx \\ &\approx \frac{2h}{6}[F_0(-h)+4F_1\cdot 0+F_2\cdot h] = \frac{h^2}{3}(F_2-F_0) \end{aligned}$$

4. Теперь решаем систему $A\overline{c}=\overline{f}$. Матрица A — диагональная, поэтому ее решение запишется просто:

$$\begin{cases} 2h\widetilde{c}_0 = 2h\dfrac{F_0+F_1+F_2}{3}; \\ \dfrac{2h^3}{3}\widetilde{c}_1 = \dfrac{h^2}{3}(F_2-F_0). \end{cases} \Longleftrightarrow \begin{cases} \widetilde{c}_0 = \dfrac{F_0+F_1+F_2}{3}; \\ \widetilde{c}_1 = \dfrac{F_2-F_0}{2h}. \end{cases}$$

5. Вычислив коэффициенты $\widetilde{c}_0$ и $\widetilde{c}_1$, можем записать построенное приближение:

$$\widetilde{\varphi}(x) = \frac{F_0+F_1+F_2}{3} + \frac{F_2-F_0}{2h}(x-x_1).$$

Оно совпало с построенной ранее функцией $\Phi(x)$. Совпадение это не случайно: мы очень неточно вычислили f_0, хотя могли бы этого избежать. Вычислим f_0, применяя формулу Симпсона:

$$\int\limits_{x_1-h}^{x_1+h} f(x)\,dx \approx \frac{h}{3}[F_0+4F_1+F_2].$$

Заново решив систему $A\overline{c}=\overline{f}$, получим такие выражения для $\widetilde{c}_0$ и $\widetilde{c}_1$:

$$\begin{cases} \widetilde{c}_0 = \dfrac{F_0+4F_1+F_2}{6}; \\ \widetilde{c}_1 = \dfrac{F_2-F_0}{2h}. \end{cases}$$

Эти коэффициенты дадут более точное приближение:

$$\overline{\widetilde{\varphi}}(x) = \frac{F_0+4F_1+F_2}{6} + \frac{F_2-F_0}{2h}(x-x_1).$$

Естественно задаться вопросом: «А насколько оно точнее?». Легко показать, что $\widetilde{\varphi}$ отличается от $\overline{\widetilde{\varphi}}$ на константу:

$$\widetilde{\varphi}(x) - \overline{\widetilde{\varphi}}(x) = \frac{F_0-2F_1+F_2}{6} = \frac{h^2}{6}\cdot\frac{F_0-2F_1+F_2}{h^2} = \frac{h^2}{6}f_{\overline{x}x,1},$$

где $f_{\overline{x}x,1}$ — вторая разностная производная $f(x)$ в точке x_1. Ранее было показано, что такое же значение принимает квадрат нормы погрешности между $\widetilde{\varphi}$ и f — то есть $||\widetilde{\varphi}(x)-f(x)||^2$:

$$||\widetilde{\varphi}(x)-f(x)||^2 = \frac{1}{6}(F_0-2F_1+F_2)^2.$$

Неформально можно сказать, что

$$f-\widetilde{\varphi} \approx \widetilde{\varphi} - \overline{\widetilde{\varphi}}.$$

Столь большие отклонения возникают из-за неточности при вычислении $\overline{f}$. Отсюда следует вывод, что алгоритм построения наилучшего приближения слишком зависит от методов приближенных вычислений, и лучше использовать единые формулы, например, для подсчета интегралов — ту же формулу Симпсона. При этом получаются неплохие результаты.

Глава 5

Численные методы решения краевых задач

Нашей первой задачей будет поиск численных методов решения задачи Коши:

$$\begin{cases} \dfrac{du}{dt} &= f(t,\, u(t)), \quad 0 < t < T; \\ u\Big|_{t=0} &= u_0. \end{cases}$$

Для начала необходимо построить соответствующую дискретную модель. Для этого разобьем весь отрезок $[0; T]$ на точки $\omega_\tau = \{t_n = n\tau,\ n = 0,\, 1,\, \ldots\ \frac{T}{\tau}\}$, где τ — диаметр дискретной сетки. Обозначим значения искомой функции — $u_n = u(t_n)$, приближенное решение — y_n, и погрешность на n-й итерации как $z_n = y_n - u_n$ в узлах сетки ($z_n,\ y_n,\ u_n$ — сеточные функции).

От будущего алгоритма потребуем как можно более точного воспроизведения функции u — для этого нужно, чтобы погрешность z_n была мала.

Обсудим понятие сходимости приближенного решения к точному. Фиксируем точку t_n и построим последовательность сеток ω_τ такую, что точка t_n является узлом для сеток с номерами $m \geqslant k$, то есть при сгущении сетки только добавляются новые узлы. На каждой из этих сеток строится сеточная функция y_n.

Определение. Сеточная функция y_n **сходится к решению** u_n **в узле** t_n, если $|z_n| \overset{\tau \to 0}{\longrightarrow} 0$.

Определение. Сходимость на отрезке означает сходимость в каждой точке этого отрезка.

Определение. Пусть погрешность по порядку роста ведет себя как $|z_n| = O(\tau^p)$, тогда приближенное решение имеет p**-й порядок точности**.

Если мы будем приближать производную в узлах сетки ее разностным аналогом:

$$u'(t_n) = \frac{y_{n+1} - y_n}{\tau},$$

то исходное уравнение примет такой вид:

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n,\, y_n), \quad n = 0,\, 1,\, \ldots \tag{5.1}$$

Получившиеся уравнения относительно y_i называются **разностной схемой**.

Из этого аналога нашего непрерывного уравнения можно определить значения y_n во всех точках сетки, если стартовать с фигурирующего в условии задачи y_0.

$$\begin{cases} y_{n+1} &= y_n + \tau f(t_n,\, y_n); \quad n = 0,\, 1,\, \ldots \\ y_0 &= u_0. \end{cases}$$

Получившаяся схема называется **явной**, так как $(n+1)$-е приближение выражается через n-е явно, непосредственно (не требуется решать уравнение).

Рассмотрим другой вариант, в котором мы заменяем обыкновенную производную на разностную производную назад. Тогда функция $f(t,\, y)$ будет браться в точке $(t_{n+1},\, y_{n+1})$:

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+1}, y_{n+1}).$$

При этом относительно y_{n+1} возникнет нелинейное уравнение:

$$\begin{cases} y_{n+1} &= \tau f(t_{n+1}, y_{n+1}) + y_n, \quad y = 0,\, 1,\, \ldots; \\ y_0 &= u_0. \end{cases}$$

То есть на каждом шаге требуется решать нелинейное, вообще говоря, уравнение, и формула для y соответственно является неявной. Отсюда и название — **неявные разностные схемы**.

Определение. Значение $\psi_n = -\dfrac{u_{n+1} - u_n}{\tau} + f(t_n,\, u_n)$ называется **невязкой** (**погрешностью аппроксимации**) **расчетной схемы (5.1)** в узле t_n.

Примечание. *Вообще говоря, далее под невязкой мы будем понимать разность левой и правой частей расчетной схемы при подстановке в нее точного решения. Вышеприведенная формула выражает невязку для методов Рунге-Кутта с параметром* $m = 1$.

Определение. Разностная схема **аппроксимирует** исходное ОДУ в узле t_n, если $\psi_n \overset{\tau \to 0}{\longrightarrow} 0$ (этот предельный переход связан с последовательностью сеток).

Определение. Разностная схема имеет p-й **порядок аппроксимации** в точке t_n, если выполнено равенство: $|\psi_n| = O(\tau^p)$.

Данную терминологию будем использовать для сравнения методов.

5.1 Сходимость методов Рунге-Кутта

Методами Рунге-Кутта[1] называется семейство методов, общий вид разностных схем которых задается так:

$$\frac{y_{n+1} - y_n}{\tau} = \sum_{i=1}^{m} \sigma_i K_i(y), \tag{5.2}$$

где величины K_i вычисляются по следующим формулам:

$$\begin{aligned} K_1(y) &= f(t_n,\, y_n); \\ K_2(y) &= f(t_n + a_2\tau,\, y_n + b_{21}\tau K_1(y)); \\ K_3(y) &= f(t_n + a_3\tau,\, y_n + b_{31}\tau K_1(y) + b_{32}\tau K_2(y)); \\ &\ldots \\ K_m(y) &= f(t_n + a_m\tau,\, y_n + \sum_{i=1}^{m-1} b_{mi}\tau K_i(y)). \end{aligned}$$

Параметры a_i, b_{ij}, σ_i выделяют конкретный метод Рунге-Кутта.

В этих методах нужно вычислять значения функции f в промежуточных точках сетки. Их количество определяется параметром m, а соответствующие методы называются m**-этапными** методами. Схемы с $m \geqslant 5$ используются крайне редко (чаще всего используют 4-этапные методы).

[1] Основная идея — К. Рунге (1885).

Какие ограничения накладываются на параметры методов Рунге-Кутта для того, чтобы обеспечить сходимость? Попробуем ответить на этот вопрос.

Потребуем, чтобы разностная схема (5.2) аппроксимировала исходное ОДУ в соответствии с введенным определением ($|\psi_n| \overset{\tau\to 0}{\longrightarrow} 0$). Невязка в данном случае будет иметь вид:

$$\psi_n = -\frac{u_{n+1} - u_n}{\tau} + \sum_{i=1}^{m} \sigma_i K_i(u).$$

Разложим u_{n+1} в ряд Тейлора с центром в точке t_n :

$$u_{n+1} = u_n + u'_n\tau + O(\tau^2).$$

Рассмотрим выражение для K_i (аналогично разложив его в ряд Тейлора):

$$K_i(u) = f(t_n,\, u_n) + O(\tau).$$

Подставим эти формулы в выражение для невязки:

$$\psi_n = -u'_n + f(t_n,\, u_n)\sum_{i=1}^{m}\sigma_i + O(\tau) = f(t_n,\, u_n)\left(-1 + \sum_{i=1}^{m}\sigma_i\right) + O(\tau).$$

Очевидно, для того, чтобы разностная схема аппроксимировала исходное ОДУ (с порядком аппроксимации $p = 1$, то есть $\psi_n = O(\tau)$), достаточно, чтобы $\sum\limits_{i=1}^{m}\sigma_i = 1$; тогда невязка будет равна $O(\tau)$.

Таким образом, можно ввести первое ограничение на параметры метода (5.2):

$$\sum_{i=1}^{m}\sigma_i = 1.$$

Теорема 5.1 (О сходимости методов Рунге-Кутта)**.** *Пусть метод Рунге-Кутта аппроксимирует исходное уравнение, тогда приближенное решение y_n сходится к точному u_n, и порядок точности приближенного решения совпадает с порядком аппроксимации разностной схемы исходного ОДУ.*

Доказательство. Будем предполагать, что функция $f(t,\, u)$ Липшиц-непрерывна по второму аргументу, то есть $\forall\, u_1,\, u_2 \quad |f(t,\, u_1) - f(t,\, u_2)| \leqslant L|u_1 - u_2|$, где L — некоторая константа.

Рассмотрим функцию погрешности: $z_n = y_n - u_n$. Выразим из нее $y_n = z_n + u_n$ и подставим в левую часть (5.2). Тогда получим:

$$\frac{z_{n+1} - z_n}{\tau} = -\frac{u_{n+1} - u_n}{\tau} + \sum_{i=1}^{m}\sigma_i K_i(u) + \sum_{i=1}^{m}\sigma_i(K_i(y) - K_i(u)).$$

Заметим, что $-\dfrac{u_{n+1} - u_n}{\tau} + \sum\limits_{i=1}^{m}\sigma_i K_i(u) = \psi_n$, и обозначим $\overline{\psi_n} = \sum\limits_{i=1}^{m}\sigma_i(K_i(y) - K_i(u))$.

Оценим для разных i выражение $|K_i(y) - K_i(u)|$:

i = 1 : $|K_1(y) - K_1(u)| = |f(t_n,\, y_n) - f(t_n,\, u_n)| \leqslant L|y_n - u_n| = L|z_n|.$

i = 2 : $|K_2(y) - K_2(u)| = |f(t_n + a_2\tau,\, y_n + b_{21}\tau K_1(y)) - f(t_n + a_2\tau,\, u_n + b_{21}\tau K_1(u))| \leqslant$
$\leqslant L|y_n - u_n + b_{21}\tau(K_1(y) - K_1(u))| \leqslant L(|z_n| + \tau|b_{21}|\cdot|K_1(y) - K_1(u)|) \leqslant$
$\leqslant L(|z_n| + \tau|b_{21}|L|z_n|) \leqslant \{$обозначим $\mathsf{b} = \max\limits_{\substack{i=\overline{2,\,n}\\ j=\overline{1,\,m-1}}} \mathsf{b_{ij}}\} \leqslant L|z_n|(1 + \tau bL).$

i=3: Если тоже самое проделать для $i = 3$, то получим такую оценку:

$$|K_3(y) - K_3(u)| \leqslant L|z_n|(1 + \tau bL)^2.$$

Теперь перейдем к общей оценке. Докажем, что

$$|K_l(y) - K_l(u)| \leqslant L|z_n|(1 + \tau bL)^{l-1}, \quad l = \overline{1,\, m}. \tag{5.3}$$

Пусть эта оценка верна для некоторого i :

$$|K_i(y) - K_i(u)| \leqslant L|z_n|(1 + \tau bL)^{i-1},$$

докажем ее для $i + 1$:

$$|K_{i+1}(y) - K_{i+1}(u)| = \left| f\left(t_n + a_{i+1}\tau,\, y_n + \tau \sum_{j=1}^{i} b_{(i+1)j} K_j(y)\right) - \right.$$

$$\left. - f\left(t_n + a_{i+1}\tau,\, u_n + \tau \sum_{j=1}^{i} b_{(i+1)j} K_j(u)\right) \right| \leqslant \left| L \left[y_n - u_n + \tau \left(\sum_{j=1}^{i} b_{(i+1)j}(K_j(y) - K_j(u)) \right) \right] \right| \leqslant$$

$$\leqslant \{\text{воспользуемся неравенством } b_{ij} \leqslant b \text{ и подсчитаем сумму геометрической прогрессии}\} \leqslant$$

$$\leqslant L\left(|y_n - u_n| + \tau b \sum_{j=1}^{i} |K_j(y) - K_j(u)| \right) \leqslant L\left(|z_n| + \tau bL|z_n| \sum_{j=1}^{i} (1 + \tau bL)^{j-1} \right) =$$

$$= L|z_n|\left(1 + \tau bL \frac{1-(1+\tau bL)^i}{1-(1+\tau bL)} \right) = L|z_n|(1 + \tau bL)^i.$$

Таким образом, оценка (5.3) действительно имеет место. Получим теперь оценку на $|\overline{\psi_n}|$:

$$|\overline{\psi_n}| \leqslant \sum_{i=1}^{m} |\sigma_i| \cdot |K_i(y) - K_i(u)| \leqslant \{\text{обозначим } \sigma = \max_{i=\overline{1,\,m}} |\sigma_i|\} \leqslant$$

$$\leqslant \sigma \sum_{i=1}^{m} L|z_n|(1 + \tau bL)^{i-1} = \sigma L|z_n|(1 + \tau bL)^{m-1} m.$$

Оценим два последних множителя:

$m(1 + \tau bL)^{m-1} \leqslant \{(1 + y)^\alpha \leqslant e^{\alpha y}\} \leqslant me^{\tau bL(m-1)} \leqslant \{\tau \leqslant T\} \leqslant me^{TbL(m-1)}$ — обозначим за $\overline{\alpha}$,

тогда $|\overline{\psi_n}| \leqslant |z_n|\sigma L\overline{\alpha}$.

Откуда, учитывая, что $\dfrac{z_{n+1} - z_n}{\tau} = \psi_n + \overline{\psi_n}$, получаем оценку на погрешность:

$$|z_{n+1}| \leqslant |z_n| + \tau|\psi_n| + \tau|\overline{\psi_n}| \leqslant |z_n| + \tau|\psi_n| + \tau|z_n|\sigma L\overline{\alpha} =$$
$$= |z_n|(1 + \tau\sigma L\overline{\alpha}) + \tau|\psi_n| \leqslant |z_{n-1}|(1 + \tau\sigma L\overline{\alpha})^2 + \tau|\psi_{n-1}|(1 + \tau\sigma L\overline{\alpha}) + \tau|\psi_n|.$$

Применив эту же операцию $n - 1$ раз, получим:

$$|z_{n+1}| \leqslant |z_0|(1 + \tau\sigma L\overline{\alpha})^n + \tau \sum_{j=0}^{n} |\psi_j|(1 + \tau\sigma L\overline{\alpha})^{n-j}.$$

Если мы обозначим $\psi = \max\limits_{j=\overline{0,\,n}} |\psi_j|$ и учтем, что $z_0 = y_0 - u_0 = 0$, то получим:

$$|z_{n+1}| \leqslant \psi\tau \sum_{j=0}^{n} (1 + \tau\sigma L\overline{\alpha})^j \leqslant \psi\tau \max_{j=\overline{0,\,n}} (1 + \tau\sigma L\overline{\alpha})^j (n + 1).$$

В силу того, что $\tau n = T$,

$$|z_{n+1}| \leqslant \psi T e^{T\sigma L\overline{\alpha}},$$

откуда следует, что если наша схема аппроксимирует на всей сетке ОДУ, то есть $|\psi_i| \to 0$, то имеет место сходимость ($|z_n| \to 0$). Кроме того, если наша разностная схема аппроксимирует исходное ОДУ с p-м порядком аппроксимации ($\psi_n = O(\tau^p)$), то погрешность имеет соответственно p-й порядок: $|z_n| = O(\tau^p)$.

Таким образом, теорема полностью доказана. □

Теперь свяжем требование на порядок аппроксимации с количеством промежуточных точек, в которых требуется вычислять значение $f(t, y)$.

5.2 Методы Рунге-Кутта второго порядка аппроксимации

Рассмотрим семейство методов Рунге-Кутта при $m = 2$. Схема для вычисления приближенного значения (y_{n+1}) будет выглядеть так:

$$\begin{cases} \dfrac{y_{n+1} - y_n}{\tau} &= \sigma_1 K_1(y) + \sigma_2 K_2(y); \\ K_1(y) &= f(t_n,\, y_n); \\ K_2(y) &= f(t_n + a_2\tau,\, y_n + \tau b_{21}\, K_1(y)). \end{cases}$$

Найдем, что является достаточным условием для достижения 2-го порядка точности. Для этого рассмотрим выражение для невязки:

$$\psi_n = -\frac{u_{n+1} - u_n}{\tau} + \sigma_1 K_1(u) + \sigma_2 K_2(u), \tag{5.4}$$

и потребуем, чтобы $\psi_n = O(\tau^2)$.

Для начала распишем дробь в правой части (5.4), применив формулу Тейлора:

$$\frac{u_{n+1} - u_n}{\tau} = \frac{1}{\tau}(u(t_n + \tau) - u_n) = \frac{1}{\tau}(u_n + u'_n\tau + u''_n\frac{\tau^2}{2} + O(\tau^3) - u_n) = u'_n + u''_n\frac{\tau}{2} + O(\tau^2).$$

Согласно постановке, $K_1(u) = f(t_n,\, y_n)$. Обозначим это число за f_n. Используя это обозначение, разложим выражение для $K_2(u)$ по формуле Тейлора для функции двух переменных:

$$K_2(u) = f(t_n + a_2\tau,\, u_n + \tau b_{21} f_n) = f_n + f'_t(t_n,\, y_n)a_2\tau + f'_u(t_n, u_n)b_{21}f_n\tau + O(\tau^2).$$

Подставив это выражение в (5.4), получим такое выражение для невязки:

$$\psi_n = -u'_n - u''_n\frac{\tau}{2} + \sigma_1 f_n + \sigma_2(f_n + f'_t(t_n,\, y_n)a_2\tau + f'_u(t_n,\, u_n)b_{21}f_n\tau) + O(\tau^2).$$

Так как u — точное решение, то для него справедливы формулы:

$$\begin{cases} u' &= f(t, u); \\ u'' &= f'_t + f'_u u' = f'_t + f'_u f. \end{cases}$$

С их использованием выражение для невязки перепишется так:

$$\psi_n = f_n(-1 + \sigma_1 + \sigma_2) + f'_t(t_n,\, y_n)(-\frac{\tau}{2} + \sigma_2 a_2\tau) + f'_u(t_n,\, y_n)f_n(-\frac{\tau}{2} + \sigma_2 b_{21}\tau) + O(\tau^2).$$

Нетрудно подобрать такие σ_1, σ_2, a_2, b_{21}, чтобы первые три слагаемых обратились в ноль — это и даст требуемую оценку для невязки. Коэффициенты σ_1, σ_2, a_2, b_{21} должны быть такими, что

$$\begin{cases} \sigma_1 + \sigma_2 = 1; \\ \sigma_2 a_2 = \dfrac{1}{2}; \\ \sigma_2 b_{21} = \dfrac{1}{2}. \end{cases} \tag{5.5}$$

Из последних двух уравнений следует, что a_2 и b_{21} равны друг другу. Обозначим это число за a. Теперь обозначим $\sigma_2 = \sigma$, и тогда из первого уравнения будет следовать, что $\sigma_1 = 1 - \sigma$. Это позволяет переписать систему (5.5) так:

$$\begin{cases} \sigma_1 = 1 - \sigma; \\ \sigma a = \dfrac{1}{2}. \end{cases}$$

Таким образом, методы с расчетной схемой следующего вида:

$$y_{n+1} - y_n = \tau((1-\sigma) f(t_n,\, y_n) + \sigma f(t_n + a\tau,\, y_n + a f(t_n,\, y_n)\tau))$$

имеют 2-й порядок аппроксимации, если выполняется условие $\sigma a = \dfrac{1}{2}$. Согласно доказанной теореме, построенное решение будет иметь 2-й порядок точности.

Можно привести примеры таких методов. При $\sigma = 1$, $a = \dfrac{1}{2}$ получается такая схема:

$$y_{n+1} - y_n = \tau f(t_n + \frac{\tau}{2},\, y_n + \frac{\tau f(t_n,\, y_n)}{2}).$$

А при $\sigma = \dfrac{1}{2}$, $a = 1$ — вот такая:

$$y_{n+1} - y_n = \frac{\tau}{2}\left[f(t_n,\, y_n) + f(t_{n+1},\, y_n + \tau f(t_n,\, y_n)\right].$$

Несмотря на бесконечное количество схем заданного порядка точности, нам может не подойти ни одна. Это связано с тем, что методы Рунге-Кутта не являются устойчивыми, и при их использовании накапливается машинная погрешность, в конце вычислений сравнимая с полученными величинами. Подробнее о вычислительной устойчивости речь пойдет в следующих разделах.

Методы Рунге-Кутта четвертого порядка точности

Приведем без вывода расчетную схему 4-го порядка точности в методе Рунге-Кутта с параметром $m = 4$:

$$\begin{cases} \dfrac{y_{n+1} - y_n}{\tau} &= \dfrac{1}{6}(K_1(y) + 2K_2(y) + 2K_3(y) + K_4(y)); \\ K_1(y) &= f(t_n,\, y_n); \\ K_2(y) &= f(t_n + \dfrac{\tau}{2},\, y_n + \dfrac{\tau}{2} K_1(y)); \\ K_3(y) &= f(t_n + \dfrac{\tau}{2},\, y_n + \dfrac{\tau}{2} K_2(y)); \\ K_4(y) &= f(t_n + \tau,\, y_n + \tau K_3(y)). \end{cases}$$

Все методы Рунге-Кутта требуют возможность вычислять значение функции в произвольной точке. Теперь рассмотрим методы, в которых этого делать не надо.

5.3 Описание многошаговых методов

Определение. Линейным m-шаговым методом называется метод с расчетной схемой следующего вида:

$$\frac{a_0 y_n + a_1 y_{n-1} + a_2 y_{n-2} + \ldots + a_m y_{n-m}}{\tau} = b_0 f_n + b_1 f_{n-1} + \ldots + b_m f_{n-m}. \tag{5.6}$$

где a_i, b_j — параметры метода, а y_{n-k} и f_{n-i} означают следующее:

$$\begin{cases} y_{n-k} = y(t_{n-k}); \\ f_{n-i} = f(t_{n-i},\, y_{n-i}). \end{cases}$$

Таким образом, для реализации m-шагового метода на первом шаге требуется знать значения $y_0, y_1, \ldots, y_{m-1}$. Значение y_0 можно взять равным u_0 — начальному условию, а вот для вычисления $y_1, \ldots, y_{m-1}$ применяют методы Рунге-Кутта соответствующего порядка точности.

Заметим также, что в методе используются только табличные данные о $f(x)$ — то есть уметь вычислять функцию f на промежуточных точках в общем случае не требуется, а может понадобиться только для получения $y_1, \ldots, y_{m-1}$.

Если в схеме (5.6) коэффициент b_0 равен нулю, то в правой части f_n не присутствует, и соответствующий метод называется **явным** (по тем же причинам, что и раньше). Если же $b_0 \neq 0$, то метод называется **неявным** (как ни странно, тоже по тем же причинам, что и раньше); возникает нелинейное уравнение относительно y_n, которое в общем случае решается методом Ньютона.

Очевидно, что метод не изменится, если выражение (5.6) домножить на какую-нибудь ненулевую константу. Поэтому устраним неоднозначность, введя условие нормировки: $\sum_{i=0}^{m} b_i = 1$. Покажем, что в этом случае правая часть уравнения (5.6) будет аппроксимировать правую часть дифференциального уравнения исходной задачи:

$$f(t_n, u_n) - \sum_{i=0}^{m} b_i f(t_n - i\tau,\, u(t_n - i\tau)) = \{\text{разлагая слагаемые в сумме в ряд Тейлора}\} =$$

$$= f(t_n, u_n) - \sum_{i=0}^{m} b_i [f(t_n, u_n) + O(\tau)]gg = f_n(1 - \sum_{i=0}^{m} b_i) + O(\tau) = O(\tau)$$

— это и означает аппроксимацию.

Теперь выведем достаточные условия для достижения k-го порядка аппроксимации исходной функции. Для этого рассмотрим выражение для невязки:

$$\psi_n = -\frac{\sum_{i=0}^{m} a_i u(t_n - i\tau)}{\tau} + \sum_{i=0}^{m} b_i f(t_n - i\tau,\, u(t_n - i\tau)) = -\frac{\sum_{i=0}^{m} a_i u_{n-i}}{\tau} + \sum_{i=0}^{m} b_i f(t_{n-i},\, u_{n-i}).$$

Теперь разложим составляющие равенства в ряд Тейлора:

$$u(t_n - i\tau) = \sum_{j=0}^{k} \frac{u_n^{(j)}}{j!}(-i\tau)^j + O(\tau^{k+1});$$

$$f(t_{n-i},\, u_{n-i}) = \{\mathsf{u' = f(t, u)}\} = u'_{n-i} = u'(t_n - i\tau) = \sum_{j=0}^{k-1} \frac{u_n^{(j+1)}}{j!}(-i\tau)^j + O(\tau^k).$$

Подставив эти формулы в выражение для невязки, получим:

$$\psi_n = -\frac{1}{\tau} \sum_{i=0}^{m} a_i \left[\sum_{j=0}^{k} \frac{u_n^{(j)}}{j!}(-i\tau)^j \right] + \sum_{i=0}^{m} b_i \sum_{j=0}^{k-1} \frac{u_n^{(j+1)}}{j!}(-i\tau)^j + O(\tau^k).$$

Поменяем порядки суммирования в двойных суммах, при этом из первой суммы вынесем отдельно составляющую при $j = 0$, а во второй сделаем замену $j = j + 1$. Тогда выражение для невязки перепишется так:

$$\begin{aligned}\psi_n &= -\frac{1}{\tau}\sum_{i=0}^{m} u_n a_i - \frac{1}{\tau}\sum_{j=1}^{k}\frac{u_n^{(j)}}{j!}\sum_{i=0}^{m} a_i(-i\tau)^j + \sum_{j=1}^{k}\frac{u_n^{(j)}}{(j-1)!}\sum_{i=0}^{m} b_i(-i\tau)^{j-1} + O(\tau^k) = \\ &= -\frac{u_n}{\tau}\sum_{i=0}^{m} a_i + \sum_{j=1}^{k}\frac{u_n^{(j)}}{j!}\tau^{j-1}\sum_{i=0}^{m}(-i)^{j-1}(ia_i + jb_i) + O(\tau^k).\end{aligned}$$

Заметим, что если все суммы подбором коэффициентов a_i, b_j обратить в нуль, то для невязки будет справедлива оценка:

$$\psi_n = O(\tau^k).$$

Таким образом, достаточным условием k-го порядка аппроксимации будет выполнение системы равенств:

$$\begin{cases} \sum\limits_{i=0}^{m} a_i = 0; \\ \sum\limits_{i=0}^{m}(-i)^{j-1}(ia_i + jb_i) = 0, \quad j = \overline{1,\,k}. \end{cases} \tag{5.7}$$

Рассмотрим отдельно последнее условие при $j = 1$:

$$\sum_{i=0}^{m} ia_i + \sum_{i=0}^{m} b_i = 0. \tag{5.8}$$

Согласно условию нормировки, $\sum\limits_{i=0}^{m} b_i = 1$, поэтому (5.8) перепишется так:

$$\sum_{i=0}^{m} ia_i = -1 \iff \sum_{i=1}^{m} ia_i = -1.$$

Добавив это уравнение и условие нормировки в систему (5.7), получим окончательный вариант достаточного условия k-го порядка аппроксимации:

$$\begin{cases} \sum\limits_{i=0}^{m} b_i = 1; \\ \sum\limits_{i=0}^{m} a_i = 0; \\ \sum\limits_{i=1}^{m} ia_i = -1; \\ \sum\limits_{i=1}^{m} i^{j-1}(ia_i + jb_i) = 0, \quad j = \overline{2,\,k}. \end{cases} \tag{5.9}$$

Мы получили систему из $k + 2$ линейных уравнений, решив которую, мы получим параметры, определяющие метод k-го порядка аппроксимации. Система содержит $2m+2$ неизвестных. Чтобы она не была переопределенной, потребуем, чтобы $k + 2 \leqslant 2m + 2$.

Таким образом, порядок аппроксимации m-шагового линейного метода не может превышать $2m$ — для неявного метода. Если же метод явный, то одним неизвестным в системе становится меньше, и максимально возможный порядок аппроксимации будет равен $2m - 1$.

Перейдем к практическим примерам.

5.4 Методы Адамса и Гира

Определение. Методы Адамса[2] — семейство m-шаговых линейных методов решения задачи Коши, в которых берется

$$a_0 = 1,\ a_1 = -1,\ a_2 = a_3 = \ldots = a_m = 0.$$

Таким образом, общая формула для нахождения приближения y_n выглядит так:

$$\frac{y_n - y_{n-1}}{\tau} = b_0 f_n + b_1 f_{n-1} + \ldots + b_m f_{n-m}. \tag{5.10}$$

Посмотрим, к чему приведет требование m-го порядка аппроксимации для данного метода. Ранее было показано, что достаточным условием будет выполнение системы равенств:

$$\begin{cases} \sum\limits_{i=0}^{m} b_i = 1; \\ \sum\limits_{i=0}^{m} a_i = 0; \\ \sum\limits_{i=1}^{m} i a_i = -1; \\ \sum\limits_{i=1}^{m} i^{j-1}(i a_i + j b_i) = 0, \quad j = \overline{2,\, k}. \end{cases}$$

Второе и третье равенство следуют из определения методов Адамса. Подставив значения a_i в последнее условие, получим такую систему:

$$\begin{cases} \sum\limits_{i=0}^{m} b_i = 1; \\ j \sum\limits_{i=1}^{m} i^{j-1} b_i = 1, \quad j = \overline{2,\, k}. \end{cases} \tag{5.11}$$

Эти уравнения на коэффициенты b_i — достаточные условия k-го порядка аппроксимации. Их k штук, а должны они определять $m+1$ неизвестное. Очевидно, чтобы система была разрешима, необходимо выполнение неравенства:

$$k \leqslant m + 1.$$

Таким образом, максимально возможный порядок аппроксимации не может превышать $m+1$. Если мы потребуем, чтобы $k = m+1$, то система (5.11) даст единственное решение — то есть схема, отвечающая максимально возможному порядку аппроксимации, одна.

Взглянув на формулу (5.10), легко заметить, что при $b_0 = 0$ она становится явной относительно y_n. Очевидно, что в этом случае система (5.11) содержит меньше неизвестных, и разрешима она будет уже при $k \leqslant m$.

Если же $b_0 \neq 0$, то схема является неявной, и для нахождения y_n приходится решать такое, в общем случае нелинейное, уравнение:

$$y_n - \tau b_0 f(t_n,\, y_n) = y_{n-1} + \tau \sum_{i=1}^{m} b_m f_{n-i}.$$

В общем случае оно решается методом Ньютона.

[2] Впервые предложены Дж. Адамсом (1855).

Пример 5.1. Положим $b_0 = 0,\ m = 1$. Это означает, что схема для вычисления y_n будет явная, поэтому максимально возможный порядок аппроксимации будет равен m, то есть 1. Пусть будет так.

От системы (5.11) останется только первое уравнение — на b_1, и из него просто получается, что $b_1 = 1$. Общая схема метода будет такова:

$$\begin{cases} \dfrac{y_n - y_{n-1}}{\tau} = f(t_{n-1},\, y_{n-1}), \quad n = 1, 2, \dots \\ y_0 = u_0 \text{ — начальное условие.} \end{cases}$$

Пример 5.2. Здесь возьмем $b_0 = 0,\ m = 3$. Максимально возможный порядок аппроксимации будет равен 3 — его и потребуем. При этом система (5.11) для нахождения b_i будет записана так:

$$\begin{cases} b_1 + b_2 + b_3 = 1; \\ 2(b_1 + 2b_2 + 3b_3) = 1; \\ 3(b_1 + 4b_2 + 9b_3) = 1. \end{cases} \iff \begin{cases} b_1 = \dfrac{23}{12}; \\ b_2 = -\dfrac{4}{3}; \\ b_3 = \dfrac{5}{12}. \end{cases}$$

Общая схема такова:

$$\frac{y_n - y_{n-1}}{\tau} = \frac{23f_{n-1} - 16f_{n-2} + 5f_{n-3}}{12},\ n \geqslant 3.$$

Значение y_0, как и раньше, берется равным u_0, а y_1 и y_2 обычно ищутся методами Рунге-Кутта. В данном случае они должны быть не менее, чем третьего порядка точности (так как k у нас равно трем).

Пример 5.3. Рассмотрим пример неявного метода Адамса при $m = 1$. Тогда можно взять k, равное 2 — максимально возможному порядку аппроксимации. Точно так же решаем систему:

$$\begin{cases} b_0 + b_1 = 1; \\ 2b_1 = 1. \end{cases} \iff \begin{cases} b_0 = \dfrac{1}{2}; \\ b_1 = \dfrac{1}{2}. \end{cases}$$

и получаем такую схему:

$$\begin{cases} \dfrac{y_n - y_{n-1}}{\tau} = \dfrac{f_n + f_{n-1}}{2}, \quad n \geqslant 1 \\ y_0 = u_0. \end{cases}$$

В данном случае y_n приходится находить из, вообще говоря, нелинейного уравнения. По-другому его можно записать так:

$$y_n - \frac{\tau}{2} f(t_n,\, y_n) = y_{n-1} + \frac{\tau}{2} f_{n-1}.$$

Обычно его решают методом Ньютона, где в качестве начального приближения берут y_{n-1}.

Пример 5.4. Последним примером на метод Адамса будет неявный метод, с $m = 2$ и $k = 3$. Получаем для нахождения b_i такую систему:

$$\begin{cases} b_0 + b_1 + b_2 = 1; \\ 2(b_1 + 2b_2) = 1; \\ 3(b_1 + 4b_2) = 1. \end{cases} \iff \begin{cases} b_0 = \dfrac{5}{12}; \\ b_1 = \dfrac{2}{3}; \\ b_2 = -\dfrac{1}{12}. \end{cases}$$

Отсюда получаем общую схему:

$$\begin{cases} \dfrac{y_n - y_{n-1}}{\tau} = \dfrac{5f_n + 8f_{n-1} - f_{n-2}}{12}, \quad n \geqslant 2 \\ y_0 = u_0; \\ y_1 \textit{ ищется методом Рунге-Кутта 3-го порядка точности.} \end{cases}$$

y_n ищется из неявной формулы по методу Ньютона.

Методы Гира

Определение. Методами Гира называется семейство линейных m-шаговых методов, в которых заранее определяется $b_0 = 1,\ b_1 = b_2 = \ldots = b_m = 0$. Общая расчетная формула будет такова:

$$\frac{a_0 y_n + a_1 y_{n-1} + \ldots + a_m y_{n-m}}{\tau} = f_n.$$

Заметим, что все эти методы являются неявными, так как y_n приходится находить из нелинейного, вообще говоря, уравнения:

$$a_0 y_n - \tau f(t_n,\, y_n) = -\sum_{i=1}^{m} a_i y_{n-i}.$$

В системе достаточных условий для k-го порядка аппроксимации

$$\begin{cases} \sum\limits_{i=0}^{m} b_i = 1; \\ \sum\limits_{i=0}^{m} a_i = 0; \\ \sum\limits_{i=1}^{m} i a_i = -1; \\ \sum\limits_{i=1}^{m} i^{j-1}(i a_i + j b_i) = 0, \quad j = \overline{2,\, k}. \end{cases}$$

первое равенство будет выполняться всегда. Подставив в последнее условие значения из определения методов Гира, получим такую систему:

$$\begin{cases} \sum\limits_{i=0}^{m} a_i = 0; \\ \sum\limits_{i=1}^{m} i a_i = -1; \\ \sum\limits_{i=1}^{m} i^{j} a_i = 0, \quad j = \overline{2,\, k}. \end{cases} \tag{5.12}$$

Она содержит $m + 1$ неизвестное и состоит из $k + 1$ уравнения. Таким образом, чтобы обеспечить разрешимость системы и требование k-го порядка аппроксимации, приходится ограничивать k числом m. При $k = m$ порядок аппроксимации будет максимально возможным, а схема, им определяемая, тоже будет одна.

Пример 5.5. В первом простейшем примере берем $m = 1,\ k = 1$. Система для нахождения a_i будет такова:

$$\begin{cases} a_0 + a_1 = 0; \\ a_1 = -1. \end{cases} \iff \begin{cases} a_0 = 1; \\ a_1 = -1. \end{cases}$$

Отсюда получаем простейшую разностную схему:

$$\begin{cases} \dfrac{y_n - y_{n-1}}{\tau} = f(t_n,\, y_n), \quad n \geqslant 1; \\ y_0 = u_0. \end{cases}$$

Пример 5.6. Взяв в этом примере $m = 2$ и $k = 2$, мы получим такую систему уравнений:

$$\begin{cases} a_0 + a_1 + a_2 = 0; \\ a_1 + 2a_2 = -1; \\ a_1 + 4a_2 = 0. \end{cases} \iff \begin{cases} a_0 = \dfrac{3}{2}; \\ a_1 = -2; \\ a_2 = \dfrac{1}{2}. \end{cases}$$

Соответствующая схема будет выглядеть так:

$$\begin{cases} \dfrac{3y_n - 4y_{n-1} + y_{n-2}}{2} = \tau f(t_n,\, y_n), \quad n \geqslant 2 \\ y_0 = u_0; \\ y_1 \textit{ ищется методом Рунге-Кутта 2-го порядка точности.} \end{cases}$$

То есть, для определения y_n мы имеем такое нелинейное уравнение:

$$\frac{3}{2} y_n - \tau f(t_n\, y_n) = 2y_{n-1} - \frac{1}{2} y_{n-2}.$$

Пример 5.7. В этом примере мы возьмем $m = 3,\ k = 3$. Тогда последнее условие в системе (5.12) разобьется на два уравнения, и мы получим такую систему:

$$\begin{cases} a_0 + a_1 + a_2 + a_3 = 0; \\ a_1 + 2a_2 + 3a_3 = -1; \\ a_1 + 4a_2 + 9a_3 = 0; \\ a_1 + 8a_2 + 27a_3 = 0. \end{cases} \iff \begin{cases} a_0 = \dfrac{11}{6}; \\ a_1 = -3; \\ a_2 = \dfrac{3}{2}; \\ a_3 = -\dfrac{1}{3}. \end{cases}$$

Этому будет соответствовать такие расчетные формулы:

$$\begin{cases} \dfrac{11y_n - 18y_{n-1} + 9y_{n-2} - 2y_{n-3}}{6} = \tau f(t_n,\, y_n), \quad n \geqslant 2 \\ y_0 = u_0; \\ y_1,\ y_2 \textit{ ищутся методом Рунге-Кутта 3-го порядка точности.} \end{cases}$$

Замечание. На практике используются методы Гира вплоть до десятого порядка аппроксимации (и соответствующей точности). Это связано с тем, что эти методы обладают свойством вычислительной устойчивости, о котором мы поговорим в следующем разделе.

5.5 Устойчивость численных методов решения задачи Коши

Мы будем рассматривать численные методы для поиска функции, являющейся решением такой задачи Коши:

$$\begin{cases} \dfrac{du}{dt} &= f(t,u), \quad t>0; \\ u\big|_{t=0} &= u_0. \end{cases}$$

Для обоснования дальнейших действий сначала проведем теоретические рассуждения. Будем считать, что функция $\overline{u}$ является решением задачи Коши:

$$\begin{cases} \dfrac{d\overline{u}}{dt} &= f(t,\overline{u}), \quad t>0; \\ \overline{u}\big|_{t=0} &= \overline{u}_0. \end{cases} \tag{5.13}$$

Пусть функция U — решение аналогичной задачи, но с «возмущенными» начальными данными:

$$\begin{cases} \dfrac{dU}{dt} &= f(t,U), \quad t>0; \\ U\big|_{t=0} &= \overline{u}_0+u_0. \end{cases} \tag{5.14}$$

Представим ее следующим образом: $U=\overline{u}+u$, где u — функция-погрешность. Исследуем u, подставив представление для U в (5.14) и разложив $f(t,u)$ в ряд Тейлора по второй переменной:

$$\begin{cases} \dfrac{d\overline{u}}{dt}+\dfrac{du}{dt} &= f(t,\overline{u}+u)=f(t,\overline{u})+f'_u(t,\overline{u})u+R_f, \quad t>0; \\ \overline{u}\big|_{t=0}+u\big|_{t=0} &= \overline{u}_0+u_0. \end{cases}$$

Используя (5.13), получим:

$$\begin{cases} \dfrac{du}{dt} &= f'_u(t,\overline{u})u+R_f, \quad t>0; \\ u\big|_{t=0} &= u_0. \end{cases} \tag{5.15}$$

Из курса «Дифференциальные уравнения» известно, что если $f'_u(t,\overline{u})<0$, то функция $u(t)$ монотонно стремится к нулю на бесконечности. В этом случае говорят, что задача (5.13) **устойчива по начальным данным**.

Теперь вернемся к численным методам. Если их применять к возмущенной задаче (при условии $f'_u<0$), то естественно требовать, чтобы функция u получалась убывающей (или, что тоже самое, функция U приближалась к $\overline{u}$, то есть погрешность не накапливалась). После проведенных выкладок понятно, что это то же самое, что требовать получение убывающей функции при решении системы (5.15). Поэтому методы обычно тестируют на такой модельной задаче:

$$\begin{cases} \dfrac{du}{dt} &= \lambda u, \quad t>0; \quad \lambda=\text{const}<0; \\ u\big|_{t=0} &= u_0, \end{cases}$$

справедливо ожидая, что если при ее решении будет получаться убывающая на бесконечности функция, то метод, примененный к системе (5.14), даст функцию, не сильно отличающуюся от $\overline{u}$. Такие методы называются **устойчивыми**.

Примечание. *Мы требуем убывания от функции, если u_0, задающее начальное условие, положительно. Если же оно отрицательно, то логично требовать возрастание, или, точнее, стремление к нулю слева.*

Будем называть метод **условно устойчивым**, если он устойчив при некоторых значениях своих параметров.

Теперь рассмотрим несколько связанных с этой модельной задачей примеров.

Пример 5.8. Явный одношаговый метод Адамса первого порядка аппроксимации. Разностная схема имеет вид:

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n).$$

Правая часть, согласно дифференциальному уравнению, выражается как $f(t_n, y_n) = \lambda y_n$, тогда схема будет иметь следующий вид:

$$y_{n+1} = (1 + \tau\lambda)y_n.$$

Как уже говорилось, метод называется устойчивым, если сеточная функция y_n не возрастает по n. В нашем примере $|y_{n+1}| = |1 + \tau\lambda| \cdot |y_n|$ не будет возрастать, если $|1 + \tau\lambda| \leqslant 1$, то есть при

$$-2 \leqslant \tau\lambda \leqslant 0.$$

Правая часть неравенства выполняется всегда, так как $\tau > 0$, а $\lambda < 0$. Из этого следует, что явный одношаговый метод Адамса удовлетворяет условию устойчивости лишь при

$$\tau \leqslant -\frac{2}{\lambda} = \frac{2}{|\lambda|}.$$

Этот метод условно устойчивый, и это не очень хорошо, так как, чтобы не набирать погрешность, надо учитывать ограничение на τ, где величина λ зависит от поведения функции $f(t, u)$ (λ — разностный аналог производной функции $f(t, u)$). То есть надо выбирать очень мелкий шаг интегрирования в соответствии с поведением функции f.

Пример 5.9. Одношаговый метод Гира первого порядка аппроксимации.

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+1},\, y_{n+1}).$$

Согласно модельной задаче, $f(t_{n+1},\, y_{n+1}) = \lambda y_{n+1}$, откуда

$$y_{n+1}(1 - \tau\lambda) = y_n.$$

Для устойчивости метода мы требуем, чтобы $|y_{n+1}| \leqslant |y_n|$. Это неравенство выполнено всегда, так как $\tau > 0$ и $\lambda < 0$, то есть сеточная функция не возрастает при любых шагах интегрирования. Мы получили, что данный метод Гира является абсолютно устойчивым. Неоспоримым достоинством этого метода по сравнению с методами Адамса является то, что шаг интегрирования мы выбираем с оглядкой лишь на требуемую точность, а к недостаткам можно отнести то, что на каждом шаге приходится решать нелинейную систему уравнений.

При рассмотрении модельной задачи у нас возникают так называемые сеточные уравнения. В общем виде их можно записать так:

$$y_{n+1} = a_n y_n + a_{n-1} y_{n-1} + \ldots + a_0 y_0.$$

В правой части стоит линейная комбинация значений сеточной функции. По сути дела, это некоторое рекуррентное соотношение. В общем виде оно решается с помощью характеристического многочлена, но мы ограничимся случаем, когда решение можно задать в виде $y_i = q^i$.

На основе такого представления решений аппарат исследования на устойчивость таков. Если удается показать, что $|q| \leqslant 1$ для любых параметров метода (от них будет зависеть рекуррентное соотношение), то общее решение будет монотонно убывать. Если же существует хотя бы один набор параметров такой, что $|q| > 1$, то условие устойчивости будет нарушено для данного метода.

Рассмотрим пример. Проверим на устойчивость явные одношаговые методы Адамса.

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n).$$

Решение сеточного уравнения будем искать в виде $y_n = q^n$. Так как схема сводится к виду $y_{n+1} = (1 + \tau\lambda)y_n$, то q будет равным $1 + \tau\lambda$.

Для устойчивости метода необходимо, чтобы $|1 + \tau\lambda| \leqslant 1$, но это условие мы уже получали выше. Таким образом, пока наш метод проверки на устойчивость ничего нового не дал, но мы увидим в полученных формулах закономерности.

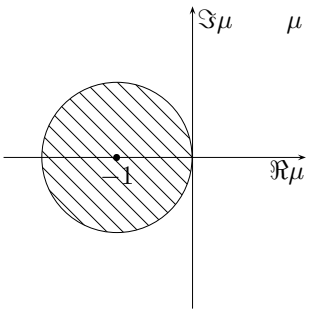
Если применить этот аппарат к методам Гира, опять получаем знакомую формулу:

$$|q| = \frac{1}{|1 - \tau\lambda|}.$$

В общем случае для сеточных уравнений, в которых задано более двух узлов, основание решения q может быть, вообще говоря, комплексным числом. Таким образом, проведем некоторое обобщение аппарата исследования устойчивости — будем считать, что в модельной задаче $\lambda \in \mathbb{C}$.

Еще раз применим метод Адамса к решению модельной задачи, но уже с комплексным λ. Обозначив в получившихся выкладках $\tau\lambda = \mu \in \mathbb{C}$, получим:

$$|1 + \mu| \leqslant 1.$$



Для всех точек, лежащих внутри области, соответствующие методы Адамса устойчивы.

Аналогично для методов Гира:

$$|1 - \mu| \geqslant 1.$$

Как нетрудно заметить, метод устойчив во внешней области.

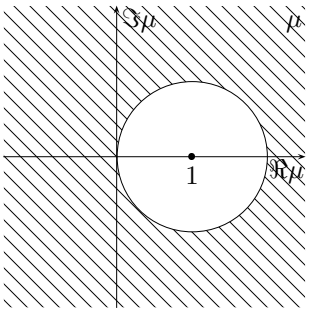
Определение. Назовем **областью устойчивости** численного метода решения задачи Коши для ОДУ ту область значений μ, в каждой точке которой $|q| \leqslant 1$.

Определение. Численный метод называется **A-устойчивым**[3], если его область устойчивости содержит отрицательную полуплоскость μ.

В соответствии с этим определением, метод Гира — A-устойчив, а метод Адамса не является A-устойчивым.

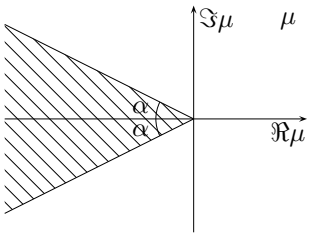
Переход с оси на комплексную полуплоскость привел к тому, что можно показать, что явных A-устойчивых методов не существует, а среди неявных нет A-устойчивых методов выше второго порядка аппроксимации.

[3]от абсолютно устойчивый.



Все это вытекает из «плохого» определения устойчивости. Попробуем немного исправить эту ситуацию.

Определение. Линейный многошаговый метод называется **A(α)-устойчивым**, если область его устойчивости содержит угол $|arg(-\mu)| < \alpha$.



Заметим, что

$$\mu = re^{i\varphi} \quad \Longrightarrow \quad -\mu = re^{i(\pi-\varphi)} \quad \Longrightarrow \quad \arg(-\mu) = \pi - \varphi.$$

Можно доказать, что для введенного определения устойчивости выполняется следующее утверждение.

Утверждение 5.1. *Среди явных линейных m-шаговых методов нет A(α)-устойчивых.*

Это следует учитывать при решении химических, биологических задач, где часто решение имеет экспоненциальный вид.

Среди неявных методов существуют A(α)-устойчивые методы. Рассмотрим пару примеров таких методов.

Пример 5.10. Двухшаговый метод Гира второго порядка аппроксимации.

$$\frac{3}{2}y_n - 2y_{n-1} + \frac{1}{2}y_{n-2} = \tau f(t_n, y_n).$$

Пример 5.11. Четырехшаговый метод Гира четвертого порядка аппроксимации.

$$\frac{1}{12}(25y_n - 48y_{n-1} + 36y_{n-2} - 16y_{n-3} + 3y_{n-4}) = \tau f(t_n, y_n).$$

Найдем область устойчивости для рассмотренного выше двухшагового метода Гира второго порядка аппроксимации. Как уже делали ранее, зададим $f(t_n, y_n) = \lambda y_n$ и, обозначим, $\tau\lambda = \mu$, тогда

сеточное уравнение будет выглядеть так:

$$\frac{3}{2}y_n - 2y_{n-1} + \frac{1}{2}y_{n-2} = \mu y_n. \tag{5.16}$$

Как мы уже договаривались, нас интересуют только решения вида $y_n = q^n$. Подставив это значение в (5.16), получим:

$$\frac{3}{2}q^2 - 2q + \frac{1}{2} = \mu q^2.$$

или:

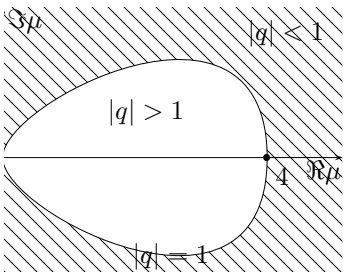
$$\mu = \frac{3}{2} - \frac{2}{q} + \frac{1}{2q^2}.$$

Найдем те μ, для которых $|q| < 1$. При этом комплексная плоскость разбивается на две области — ту, где метод устойчив, и — где неустойчив. При этом, так как на границе $|q| = 1$, то $q = e^{i\varphi}$ и уравнение границы примет вид:

$$\begin{aligned}\mu(\varphi) &= \frac{3}{2} - 2e^{-i\varphi} + \frac{1}{2}e^{-i2\varphi} = \frac{3}{2} - 2\cos\varphi + 2i\sin\varphi + \frac{1}{2}\cos 2\varphi - i\frac{1}{2}\sin 2\varphi = \\ &= \frac{3}{2} - 2\cos\varphi + \frac{1}{2}(2\cos^2\varphi - 1) + i(2\sin\varphi - \sin\varphi\cos\varphi).\end{aligned}$$

Сделаем замену переменной $x = \cos\varphi$ ($x \in [-1;\, 1]$), получим:

$$\mu = 1 - 2x + x^2 \pm i(2 - x)\sqrt{1 - x^2}.$$



Вообще то неплохо было бы отметить здесь еще область определения, а то, например, при $\mu = \frac{3}{2}$ итерационный процесс неопределен. Проверить область определения мы предоставим читателю. Подставим точку $\mu = -\frac{3}{2}$, она лежит во внешней части плоскости, а $q = \frac{1 \pm i\sqrt{1/2}}{3}$, $|q| < 1$ и, таким образом, мы получили, что внешняя область является областью устойчивости данного метода Гира.

Хотя при реализации неявных методов требуется больше времени на просчет одного шага, но их плюс в том, что мы можем выбирать произвольный шаг — тот, который нам нужен.

Численные методы решения задачи Коши для систем ОДУ

Исследуем методы решения задачи Коши для систем ОДУ. Будем искать приближенное решение для такой системы:

$$\begin{cases} \overline{u}_t &= \overline{f}(t, \overline{u}), \quad t > 0; \\ \overline{u}\big|_{t=0} &= \overline{u_0}, \end{cases}$$

где $\overline{u} = (u_1,\, u_2,\, \ldots,\, u_n)$ и $\overline{f} = (f_1,\, f_2,\, \ldots,\, f_n)$.

Методы Рунге-Кутта и линейные одношаговые методы легко переносятся на системы уравнений. Чтобы стало яснее, приведем пример задачи:

$$\begin{cases} u_t &= f(t,u,v); \\ v_t &= g(t,u,v); \\ u(0) &= u_0; \\ v(0) &= v_0. \end{cases}$$

Покажем, как к этой задаче применяется метод Рунге-Кутта четвертого порядка аппроксимации (и точности соответственно). Обозначим $u(t_n) = y_n$, а $v(t_n) = z_n$, тогда формулы для подсчета следующих сеточных значений будут таковы:

$$\begin{cases} \dfrac{y_{n+1} - y_n}{\tau} = \dfrac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4); \\ \dfrac{z_{n+1} - z_n}{\tau} = \dfrac{1}{6}(M_1 + 2M_2 + 2M_3 + M_4). \end{cases}$$

Параметры $K_i,\ M_i,\ i = 1,2,3,4$ вычисляются по следующим формулам:

$$\begin{array}{ll} K_1 = f(t_n, y_n, z_n), & M_1 = g(t_n, y_n, z_n); \\ K_2 = f(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}K_1, z_n + \frac{\tau}{2}M_1), & M_2 = g(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}K_1, z_n + \frac{\tau}{2}M_1); \\ K_3 = f(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}K_2, z_n + \frac{\tau}{2}M_2), & M_3 = g(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}K_2, z_n + \frac{\tau}{2}M_2); \\ K_4 = f(t_n + \tau, y_n + \tau K_3, z_n + \tau M_3), & M_4 = g(t_n + \tau, y_n + \tau K_3, z_n + \tau M_3). \end{array}$$

Мы долго и упорно говорим, что условно устойчивые методы хуже устойчивых, но не привели не одного примера, объясняющего разницу между ними. Исправим ситуацию. Представим, что мы моделируем два процесса, причем один из них протекает существенно быстрее другого. Система уравнений для такой задачи будет иметь вид:

$$\begin{cases} u_1' + a_1 u_1 &= 0, \quad a_1 > 0; \\ u_2' + a_2 u_2 &= 0, \quad a_2 > 0; \\ u_1(0) &= u_0; \\ u_2(0) &= u_0, \end{cases}$$

а точные решения соответственно:

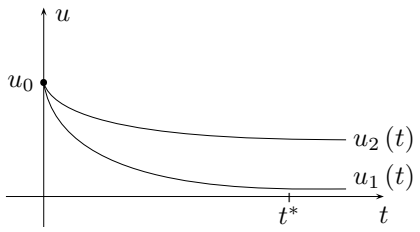
$$\begin{cases} u_1 &= u_0 e^{-a_1 t}; \\ u_2 &= u_0 e^{-a_2 t}. \end{cases}$$

Как мы уже говорили, $a_1 \gg a_2$ — константы суть характерное время протекания первого и второго процессов. Нас интересует описание системы при достаточно больших $t \geqslant t^*$. Пусть мы используем для решения этой задачи условно устойчивый метод (например, какой-нибудь из методов Адамса). При этом на шаг τ мы должны сделать ограничение $\tau\lambda \geqslant -2$, где роль параметра λ играют a_1 и a_2, то есть $\lambda = -a_1$ или $\lambda = -a_2$. Неравенства должны выполняться одновременно, то есть $\tau \leqslant \frac{2}{a_1}$ (в силу того, что $a_1 \gg a_2$).

Теперь представим, что нам важно исследовать первый процесс в то время, когда второй уже не влияет на систему. Тем не менее, от поведения второго процесса зависит шаг аппроксимации (он будет уменьшаться), поэтому весьма вероятно, что мы будем выполнять «лишнюю» работу. Отсюда можно сделать вывод, что в случае решения систем ОДУ, описывающих разномасштабные процессы, надо использовать абсолютно устойчивые методы (например, методы Гира).

Опишем еще одно свойство систем уравнений. В качестве примера системы возьмем такую:

$$\begin{cases} \dfrac{\partial \overline{u}}{\partial t} &= f(\overline{t}, \overline{u}), \quad 0 < t \leqslant T; \\ \overline{u}\big|_{t=0} &= \overline{u}_0. \end{cases} \tag{5.17}$$



Поставим ей в соответствие матрицу из производных (якобиан):

$$A(t, \overline{u}) = \left(\frac{\partial f_j(t, \overline{u})}{\partial u_i} \right)$$

и обозначим ее собственные значения за $\lambda_k(t),\ t \in [0;\, T]$.

Определение. Система (5.17) называется **жесткой**, если выполнены два условия:

$$\operatorname{Re} \lambda_k(t) < 0 \quad \forall\, k,\ t \in [0;\, T];$$

$$\sup_{t \in [0;\, T]} \frac{\max\limits_k |\operatorname{Re} \lambda_k(t)|}{\min\limits_k |\operatorname{Re} \lambda_k(t)|} \gg 1 \text{ — много больше единицы.}$$

Такие системы чаще всего решают неявным абсолютно устойчивым методом.

5.6 Интегро-интерполяционный метод построения разностных схем

Перейдем к рассмотрению более сложных краевых задач. Для начала исследуем применение численных методов для решения такой модельной задачи:

$$\begin{cases} (k(x)u'(x))' - q(x)u(x) + f(x) = 0, \quad 0 < x < l; \\ -k(0)u'(0) + \beta u(0) = \mu_1; \\ u(l) = \mu_2. \end{cases} \tag{5.18}$$

Это краевая задача для обыкновенного дифференциального уравнения второго порядка с заданными функциями k, q, f и неизвестной функцией u. Известно, что если выполняются условия

$$\begin{cases} k(x) \geqslant k_0 > 0; \\ q(x) \geqslant 0; \\ \beta \geqslant 0, \end{cases}$$

то решение задачи (5.18) существует и единственно.

Задача (5.18) содержит уравнение параболического типа. Обычно такие уравнения возникают при исследовании распределения температуры в тонком стержне или в диффузионных процессах.

Решение системы типа (5.18) проходит в несколько этапов. Сначала ей сопоставляется дискретная модель, а на ее основе строится разностная схема. Существует несколько методов построения таких схем, и первым мы рассмотрим **интегро-интерполяционный метод**. Название его происходит от того, что в процессе построения соответствующей разностной схемы мы переходим от интегральных соотношений к интерполяционным уравнениям.

Построение разностной схемы

Перейдем к построению схемы. Для начала введем на отрезке $[0;\, l]$ равномерную сетку:

$$\omega_h = \{x_i = ih,\ i = \overline{0,\, N},\ h = \frac{l}{N}\}$$

— очевидно, $x_0 = 0$, $x_N = l$. Теперь введем такое обозначение для средних точек между узлами сетки:

$$x_{i\pm\frac{1}{2}} = x_i \pm \frac{h}{2}.$$

Также обозначим $u_i = u(x_i)$ — значение искомой функции в узлах сетки, и $W(x) = k(x)u'(x)$. Применяя эти обозначения, фиксируем произвольное $i \in [1;\, N-1]$ и проинтегрируем первое уравнение системы (5.18) по отрезку $[x_{i-\frac{1}{2}};\, x_{i+\frac{1}{2}}]$:

$$\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} ((k(x)u'(x))' - q(x)u(x) + f(x))\,dx = 0 \iff$$

$$W_{i+\frac{1}{2}} - W_{i-\frac{1}{2}} - \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)u(x)\,dx + \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx = 0, \tag{5.19}$$

где $W_{i\pm\frac{1}{2}} = W(x_{i\pm\frac{1}{2}})$.

Первый интеграл можно приблизить значением $u_i \cdot \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)\,dx$. Тогда (5.19) можно переписать как приближенное равенство:

$$W_{i+\frac{1}{2}} - W_{i-\frac{1}{2}} - u_i \cdot \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)\,dx + \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx \approx 0. \tag{5.20}$$

Перейдем от интегральных выражений к линейным. Для этого введем новые обозначения:

$$\varphi_i = \frac{1}{h}\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx; \quad d_i = \frac{1}{h}\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)\,dx. \tag{5.21}$$

Теперь заметим, что $u'(x) = \dfrac{W(x)}{k(x)}$. Проинтегрировав это равенство на отрезке $[x_i;\, x_{i+1}]$, получим:

$$u_{i+1} - u_i = \int\limits_{x_i}^{x_{i+1}} \frac{W(x)}{k(x)}\,dx.$$

Заменим это равенство приближенным:

$$u_{i+1} - u_i \approx W_{i+\frac{1}{2}} \int\limits_{x_i}^{x_{i+1}} \frac{dx}{k(x)}.$$

Отсюда следует, что если обозначить

$$a_{i+1} = \left[\frac{1}{h} \int\limits_{x_i}^{x_{i+1}} \frac{dx}{k(x)} \right]^{-1},$$

то будут справедливы приближенные равенства:

$$\begin{cases} W_{i+\frac{1}{2}} \approx a_{i+1} \dfrac{u_{i+1} - u_i}{h}; \\ W_{i-\frac{1}{2}} \approx a_i \dfrac{u_i - u_{i-1}}{h}. \end{cases} \tag{5.22}$$

Воспользовавшись обозначениями (5.21) и (5.22), приближенное равенство (5.20) можно переписать так:

$$a_{i+1} \frac{u_{i+1} - u_i}{h} - a_i \frac{u_i - u_{i-1}}{h} - h d_i u_i + \varphi_i h \approx 0. \tag{5.23}$$

Обозначим за y_i такие числа, которые при подстановке в (5.23) вместо u_i дают точное равенство:

$$a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} - h d_i y_i + \varphi_i h = 0. \tag{5.24}$$

Найденные из таких уравнений значения y_i и будут считаться приближениями к u_i. Полученное равенство и будет искомой разностной схемой, однако ее можно переписать и в более компактном виде, заметив, что первые две дроби — не что иное, как разностные производные назад. Обозначив их $y_{\overline{x},i+1}$ и $y_{\overline{x},i}$, перепишем (5.24) в следующем виде:

$$a_{i+1} y_{\overline{x},i+1} - a_i y_{\overline{x},i} - h d_i y_i + \varphi_i h = 0.$$

Разделив это равенство на h, мы сможем объединить две разностные производные во вторую разностную производную вперед:

$$\frac{(a y_{\overline{x}})_{i+1} - (a y_{\overline{x}})_i}{h} \equiv (a y_{\overline{x}})_{x,i}.$$

В итоге мы получим такой вид разностной схемы:

$$(a y_{\overline{x}})_{x,i} - d_i y_i + \varphi_i = 0. \tag{5.25}$$

Мы имеем право написать такие равенства для $i = \overline{1,\, N-1}$. Их можно объединить в систему линейных (по построению) уравнений относительно y_i. Она будет содержать $N-1$ уравнение и $N+1$ неизвестное. Необходимые для однозначной разрешимости системы 2 уравнения добавим из краевых условий. Мы можем заменить в последнем равенстве в (5.18) $u(l)$ на y_N, тогда получим, что

$$y_N = \mu_2. \tag{5.26}$$

Для получения последнего уравнения мы выполним те же самые действия, что и при выводе равенства (5.19), но интегрирование будем проводить на отрезке $[0;\, \frac{h}{2}]$. Тогда можно получить такое равенство:

$$W_{\frac{1}{2}} - W_0 - u_0 \int\limits_0^{\frac{h}{2}} q(x)\, dx + \int\limits_0^{\frac{h}{2}} f(x)\, dx \approx 0. \tag{5.27}$$

$W_{\frac{1}{2}}$ и W_0 мы найдем, заменив в приближенных равенствах

$$W_{\frac{1}{2}} \approx a_1 \frac{u_1 - u_0}{h};$$
$$W_0 \approx \beta u_0 - \mu_1.$$

u_k на y_k и получив уравнения:

$$W_{\frac{1}{2}} = a_1 \frac{y_1 - y_0}{h};$$
$$W_0 = \beta y_0 - \mu_1.$$

После этого, воспользовавшись обозначениями

$$\varphi_0 = \frac{1}{\frac{h}{2}} \int_0^{\frac{h}{2}} f(x)\,dx, \quad d_0 = \frac{1}{\frac{h}{2}} \int_0^{\frac{h}{2}} q(x)\,dx,$$

мы приведем (5.27) к такому виду:

$$a_1 y_{x,0} - \beta y_0 + \mu_1 - d_0 \frac{h}{2} y_0 + \frac{h}{2}\varphi_0 = 0 \iff$$

$$a_1 y_{x,0} - \overline{\beta} y_0 = \overline{\mu}, \tag{5.28}$$

где $\overline{\beta} = \beta + d_0 \frac{h}{2}$, $\overline{\mu} = \mu_1 + \varphi_0 \frac{h}{2}$.

Уравнения (5.25), (5.26) и (5.28) представляют собой окончательный вариант разностной схемы, полученной с использованием интегро-интерполяционного метода.

Решение разностной схемы

Вторым шагом на пути решения краевой задачи численно становится выбор метода решения построенной схемы. В нашем случае мы получили систему линейных уравнений, для которой метод решения и выбирается. Заметим, что (5.25) можно переписать так:

$$\frac{1}{h}\left(a_{i+1}\frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h}\right) - d_i y_i + \varphi_i = 0.$$

Собрав коэффициенты при y_i, получим:

$$A_i y_{i+1} - C_i y_i + B_i y_{i-1} = -F_i, \quad i = \overline{1,\, N-1},$$

где $A_i = a_{i+1}$, $B_i = a_i$, $C_i = a_i + a_{i+1} + d_i h^2$.

Добавив к этим уравнениям уравнения (5.26) и (5.28), получим систему из $N+1$ уравнения. Матрица, задающая эту систему уравнений, будет являться трехдиагональной, а такие системы обычно решаются методом прогонки. Он применим, так как выполнены условия:

$$A_i,\ B_i > 0,\ C_i > B_i + A_i$$

— они дают существование и единственность y_i, отвечающих уравнениям (5.25), (5.26) и (5.28).

5.7 Метод аппроксимации квадратичного функционала

Это другой метод построения разностных схем. Будем рассматривать его для задачи, схожей с (5.18), но с более простыми краевыми условиями:

$$\begin{cases} (k(x)u'(x))' - q(x)u(x) + f(x) = 0, \quad 0 < x < 1; \\ u(0) = u(1) = 0. \end{cases}$$

Известно, что решение такой задачи эквивалентно поиску u, минимизирующих функционал

$$J[u] = \int_0^1 [k(x)(u'(x))^2 + q(x)u^2(x) - 2f(x)u(x)]\,dx.$$

Задав на отрезке [0; 1] равномерную сетку, разобьем интеграл по всему отрезку на сумму по подотрезкам:

$$J[u] = \sum_{i=1}^{N} \int\limits_{x_{i-1}}^{x_i} \left[k(x)(u'(x))^2 + q(x)u^2(x) - 2f(x)u(x)\right] dx.$$

Для упрощения поиска экстремума заменим обыкновенную производную на ее разностный аналог:

$$J[u] \approx \sum_{i=1}^{N} \int\limits_{x_{i-1}}^{x_i} \left[k(x)u_{\overline{x},i}^2 + q(x)u^2(x) - 2f(x)u(x)\right] dx.$$

Воспользовавшись обозначением $a_i = \dfrac{1}{h} \int\limits_{x_{i+1}}^{x_i} k(x)\, dx$, получим:

$$J[u] \approx \sum_{i=1}^{N} \left[a_i u_{\overline{x},i}^2 h + \int\limits_{x_{i-1}}^{x_i} \left(q(x)u^2(x) - 2f(x)u(x)\right)\, dx \right].$$

Интегралы посчитаем по формуле трапеций, заменив всюду u_i на y_i — приближенные значения:

$$J[u] \approx J[y] = J_h(y_0, y_1, \ldots, y_N) = \sum_{i=1}^{N} \left[a_i y_{\overline{x},i}^2 h + \frac{h}{2}(q_i y_i^2 - 2f_i y_i + q_{i-1} y_{i-1}^2 - 2f_{i-1} y_{i-1}) \right] =$$

$$= \{\mathsf{y_0} = \mathsf{y_N} = 0\} = \sum_{i=1}^{N} a_i y_{\overline{x},i}^2 h + \sum_{i=1}^{N-1} (q_i y_i^2 - 2f_i y_i)h.$$

Мы свели задачу о поиске элемента, минимизирующего функционал, к поиску чисел y_i, доставляющих минимум функции многих переменных — при этом, правда, мы потеряли в точности.

Необходимым условием экстремума будет равенство нулю всех частных производных:

$$\begin{aligned} \frac{\partial J_h}{\partial y_i} = 0 \quad &\Longleftrightarrow \quad 2a_{i+1} y_{\overline{x},i+1}(-\frac{1}{h} \cdot h) + 2a_i y_{\overline{x},i}(\frac{1}{h} \cdot h) + (2q_i y_i - 2f_i)h = 0 \iff \\ &\Longleftrightarrow \quad \frac{a_{i+1} y_{\overline{x},i+1} - a_i y_{\overline{x},i}}{h} - q_i y_i + f_i = 0 \\ &\Longleftrightarrow \quad (a y_{\overline{x}})_{x,i} - q_i y_i + f_i = 0. \end{aligned}$$

Последнее уравнение — это уже часть итоговой разностной схемы (осталось добавить только краевые условия). Можно заметить, что она похожа на схему, возникающую в интегро-интерполяционном методе, однако коэффициенты различны, да и свойства схем тоже довольно сильно различаются.

5.8 Корректность разностной схемы

Напомним несколько определений.

Определение. Разностная схема **аппроксимирует** исходную дифференциальную задачу в точке x_i [на всей сетке], если погрешность аппроксимации в этой точке [соответственно, норма погрешности аппроксимации] стремится к нулю [соответственно, тоже к нулю] с уменьшением h :

$$\psi(x_i) \xrightarrow{h \to 0} 0 \quad \left[||\psi||_h = ||\psi||_{C(\omega_n)} = \max_i |\psi_i| \xrightarrow{h \to 0} 0 \right].$$

Определение. Разностная схема **аппроксимирует** исходную дифференциальную задачу с p-м порядком аппроксимации, если $\psi_i = O(h^p)$ в точках x_i, или в целом на сетке, если $||\psi_i||_h = O(h^p)$.

Будем поступать так же, как и в случае выяснения порядка аппроксимации для задачи Коши в разностной задаче аппроксимации.

Можно показать, что в целом на сетке схема, построенная интегро-интерполяционным методом, будет иметь второй порядок аппроксимации. Вычисления громоздки и мы их опускаем, заметив, что невязка имеет порядок не хуже $O(h^2)$ даже в крайних узлах:

$$\psi_0 = O(h^2), \quad \psi_N = 0.$$

Если мы требуем такой порядок аппроксимации, то можно сэкономить на вычислении параметров, вычисляя $a_i,\ d_i,\ \varphi_i$ по формуле прямоугольников, при этом получим следующие значения параметров:

$$\begin{cases} a_i &= k\left(x_{i-\frac{1}{2}}\right); \\ d_i &= q(x_i); \\ \varphi_i &= f(x_i). \end{cases}$$

Если считать параметры по квадратурной формуле трапеций, то получим следующие выражения:

$$\begin{cases} \dfrac{1}{a_i} &= \dfrac{1}{k(x_{i-1})} + \dfrac{1}{k(x_i)}; \\ d_i &= \dfrac{1}{2}\big(q\big(x_{i-\frac{1}{2}}\big) + q\big(x_{i+\frac{1}{2}}\big)\big); \\ \varphi_i &= \dfrac{1}{2}\big(f\big(x_{i-\frac{1}{2}}\big) + f\big(x_{i+\frac{1}{2}}\big)\big). \end{cases}$$

Рассмотрим вопрос о сходимости приближенного решения к точному. Как обычно, обозначим $z_i = y_i - u_i$ — погрешность и напомним несколько определений.

Определение. Приближенное решение y_i **сходится к точному в точке** x_i, если $z_i \xrightarrow{h\to 0} 0$.

Определение. Приближенное решение y_i **сходится к точному на всей сетке**, если $||z_i||_h \xrightarrow{h\to 0} 0$.

Определение. Если величина погрешности z_i в каждой точке (или на всей сетке) есть $O(h^p)$, то метод имеет **p-й порядок точности**.

Можно установить, что сеточная норма $||z||_{C(\omega_i)} = O(h^2)$. Доказательство этого утверждения можно посмотреть в книге [1].

Подставим в расчетную схему $y_i = u_i + z_i$:

$$(az_{\overline{x}})_{x,i} - d_i z_i = -(au_{\overline{x}})_{x,i} + d_i u_i - \varphi_i.$$

Аналогичную операцию проведем для граничных условий:

$$\begin{cases} (az_x)_{x,i} - d_i z_i &= -\psi_i; \\ -a_1 z_{x,0} + \overline{\beta} z_0 &= -\psi; \\ z_N &= 0. \end{cases}$$

Как нетрудно заметить, задача для погрешности имеет ту же структуру, что и исходная разностная схема, с заменой правой части на невязку.

После преобразований системы, которые мы снова опускаем, можно получить, что

$$||z||_h \leqslant M_1 ||\psi||_h.$$

Разностная схема имеет 2-й порядок аппроксимации ($||\psi||_h = O(h^2)$), а, следовательно, и второй порядок точности $||z||_h = O(h^2)$. Подробные указания на то, как это получить, можно найти в [1].

Проделав те же самые действия, можем получить оценку на приближенное решение в сеточной норме (по аналогии — уравнения очень похожи).

$$||y||_h \leqslant M_1(||\varphi||_h + |\mu_1| + |\mu_2|), \tag{5.29}$$

— такой оценки и следовало ожидать.

Теперь мы можем дать несколько определений.

Определение. Если для решения разностной задачи выполняется оценка (5.29), то решение называется **устойчивым по правой части**.

Это определение устойчивости разностной задачи является непосредственным следствием общего определения устойчивости.

Определение. Задача называется **корректно поставленной по Адамару**[4], если:
1) решение существует;
2) решение единственно;
3) решение непрерывно зависит от входных данных (устойчиво по правой части).

По аналогии запишем определение для разностной схемы.

Определение. Разностная схема называется **корректной**, если:
1) решение существует;
2) решение единственно;
3) решение устойчиво по правой части.

Теорема 5.2. *Пусть дифференциальная задача корректно поставлена, и разностная схема, соответствующая этой дифференциальной задаче, корректна. Тогда, если разностная схема аппроксимирует исходную задачу, то решение разностной схемы сходится к решению исходной дифференциальной задачи и порядок аппроксимации совпадает с порядком точности.*

Доказательство. Как уже записывали ранее, оценка на приближенное решение такова

$$||y||_h \leqslant M_1||\varphi||_h.$$

В силу линейности разностной схемы, оценка на погрешность будет:

$$||z||_h \leqslant M_1||\psi||_h \overset{h\to 0}{\longrightarrow} 0.$$

Таким образом, так как $||\psi||_h = O(h^p)$, решение разностной схемы сходится и имеет p-й порядок точности. □

Далее работать с теоремой мы будем по следующему плану. Сначала исследуем разностную схему на аппроксимацию (аппроксимирует ли она исходное ОДУ). Затем проверяем схему на устойчивость, и потом уже можно пользоваться теоремой, что мы и будем с успехом делать.

5.9 Явная разностная схема для уравнения теплопроводности

Мы будем рассматривать достаточно простые задачи, в которых решение можно построить и аналитическими методами, но наша задача — изучать численные методы. Рассмотренная ниже техника решения краевых задач легко обобщается на более сложные случаи, не имеющие аналитического решения.

[4]Понятие впервые предложено Ж. Адамаром (1923).

Запишем краевую задачу для уравнения теплопроводности (УТ):

$$\begin{cases} u_t(x,t) &= u_{xx}(x,t) + f(x,t), & 0 < x < 1,\ 0 < t \leqslant T; \\ u(x,0) &= u^0(x), & 0 \leqslant x \leqslant 1; \\ u(0,t) &= \mu_1(t), & 0 \leqslant t \leqslant T; \\ u(1,t) &= \mu_2(t), & 0 \leqslant t \leqslant T. \end{cases}$$

Проведем дискретизацию области изменения независимого переменного:

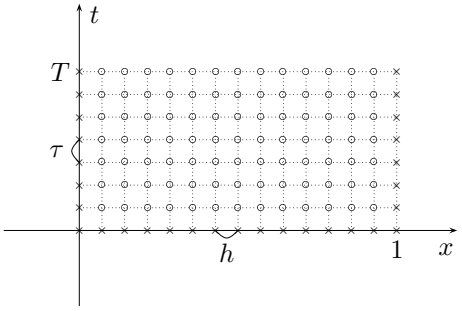
$$\begin{aligned} \omega_h &= \left\{ x_j = jh,\ j = \overline{0,N},\ h = \tfrac{1}{N} \right\}; \\ \omega_\tau &= \left\{ t_n = n\tau,\ n = \overline{0,K},\ \tau = \tfrac{T}{K} \right\}. \end{aligned}$$

Параметры N и K характеризуют «густоту» сетки.

Теперь построим семейство линий

$$\begin{aligned} x &= x_j, \quad j = \overline{0,N}; \\ t &= t_n, \quad n = \overline{0,K}. \end{aligned}$$

Будем рассматривать точки пересечения этих линий. Разделим узлы на две группы — граничные узлы (в них заданы дополнительные условия — краевые и граничные условия) и внутренние узлы.



После дискретизации строим некоторый аналог исходного уравнения — разностную схему. Сначала введем обозначения:

$$\begin{aligned} u(x_j,\, t_n) &= u_j^n; \\ f(x_j,\, t_n) &= \varphi_j^n, \end{aligned}$$

а для частных производных возьмем такие приближения:

$$\begin{aligned} u_{xx}\big|_{(x_j,\, t_n)} &\approx \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}; \\ u_t\big|_{(x_j,\, t_n)} &\approx \frac{u_j^{n+1} - u_j^n}{\tau}. \end{aligned}$$

— эти конструкции возникают при применении интегро-интерполяционного метода с формулой прямоугольников. Подставив эти формулы в краевую задачу, получим ее алгебраический аналог:

$$\begin{cases} \dfrac{y_j^{n+1} - y_j^n}{\tau} &= \dfrac{y_{j+1}^n - 2y_j^n + y_{j-1}^n}{h^2} + \varphi_j^n, \quad j = \overline{1,\, N-1},\ n = \overline{0,\, K-1}; \\ y_j^0 &= u_0(x_j), \quad j = \overline{0,\, N}; \\ y_0^n &= \mu_1(t_n), \quad n = \overline{0,K}; \\ y_N^n &= \mu_2(t_n), \quad n = \overline{0,K}. \end{cases} \tag{5.30}$$

Будем исследовать класс схем для решения задачи (5.30). Нас будут интересовать следующие вопросы:

1) существование и единственность решения;

2) методы получения решения разностной схемы;

3) как соотносятся разностная схема и исходная дифференциальная задача (т. е. аппроксимация);

4) есть ли сходимость приближенного решения к точному.

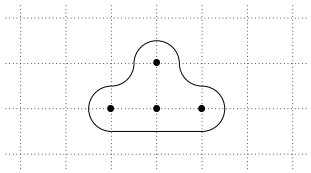
Будем рассматривать (пытаться разрешить) уравнение относительно y_j^{n+1}. Множество узлов дискретной сетки с одинаковым $t = \text{const}$ назовем **временным слоем**. Первое уравнение из (5.30) можно переписать так:

$$y_j^{n+1} = \left(1 - \frac{2\tau}{h^2}\right) y_j^n + \frac{\tau}{h^2}\left(y_{j+1}^n + y_{j-1}^n\right) + \tau\varphi_j^n, \quad j = \overline{1,\, N-1},\ n = \overline{0,\, K-1}.$$

При $n = 0$ получим:

$$y_j^1 = \left(1 - \frac{2\tau}{h^2}\right) y_j^0 + \frac{\tau}{h^2}\left(y_{j+1}^0 + y_{j-1}^0\right) + \tau\varphi_j^0, \quad j = \overline{1,\, N-1}.$$

В правой части все значения известны из краевых условий. Поэтому мы можем получить искомую сеточную функцию на всем первом временном слое. Аналогично можно рассчитать второй и последующие слои. Таким образом, данная разностная схема решается по слоям, и понятно, что решение существует и единственно. Так как все формулы явные, то вся разностная схема является явной разностной схемой. Эта схема построена по четырем узлам дискретной сетки, что было необходимо для представления первой производной по t и второй по x. Четырех узлов оказалось достаточно, и схема получилась простой.



Определение. Совокупность узлов дискретной сетки на базе которых построено разностное уравнение, называется **шаблоном разностной схемы**.

Далее мы будем изучать более сложные схемы с большим числом узлов в шаблоне, а пока вернемся к исследованию только что полученного разностного уравнения.

Изучим поведение невязки:

$$\psi_j^n = -\frac{u_j^{n+1} - u_j^n}{\tau} + \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} + \varphi_j^n.$$

Для этого разложим $u_{j\pm1}^n$ и u_j^{n+1} в ряд Тейлора в точке u_j^n :

$$\begin{aligned} u_{j\pm1}^n &= u_j^n \pm u_{x,j}^n h + u_{xx,j}^n \tfrac{h^2}{2} \pm u_{xxx,j}^n \tfrac{h^3}{6} + O(h^4); \\ u_j^{n+1} &= u_j^n + u_{t,j}^n \tau + O(\tau^2). \end{aligned}$$

Тогда получим выражение для невязки:

$$\psi_j^n = (-u_{t,j}^n + u_{xx,j}^n + f_j^n) + \varphi_j^n - f_j^n + O(\tau + h^2).$$

Выражение в скобках равно нулю в силу того, что u_j^n — точное решение уравнения теплопроводности.

Если взять функции φ_j^n так, что $\varphi_j^n = f_j^n + O(\tau + h^2)$ (мы допускаем некоторый произвол при выборе φ_j^n), то для невязки будет справедлива оценка:

$$||\psi||_h = O(\tau + h^2)$$

— то есть наша разностная схема аппроксимирует исходное ДУ со вторым порядком аппроксимации по h и с первым порядком аппроксимации по τ.

Рассмотрим теперь вопрос о сходимости приближенного решения к точному. Как обычно, выразим искомое решение y_j^n через точное решение и погрешность: $y_j^n = u_j^n + z_j^n$. Тогда разностная схема примет вид:

$$\begin{cases} \dfrac{z_j^{n+1} - z_j^n}{\tau} = \dfrac{z_{j+1}^n - 2z_j^n + z_{j-1}^n}{h^2} + \psi_j^n, & j = \overline{1,\, N-1}, \quad n = \overline{1,\, K-1}; \\ z_j^0 = 0, & j = \overline{0,\, N}; \\ z_0^n = 0, & n = \overline{0,\, K}; \\ z_N^n = 0, & n = \overline{0,\, K}. \end{cases}$$

Выразим из первого уравнения погрешность на $(n+1)$-м временном слое:

$$z_j^{n+1} = (1 - \frac{2\tau}{h^2})z_j^n + \frac{\tau}{h^2}(z_{j+1}^n + z_{j-1}^n) + \tau\psi_j^n.$$

Получим оценку на погрешность:

$$|z_j^{n+1}| \leqslant \left|1 - 2\frac{\tau}{h^2}\right| |z_j^n| + \frac{\tau}{h^2}\left(|z_{j+1}^n| + |z_{j-1}^n|\right) + |\tau\psi_j^n| \leqslant \left(\left|1 - 2\frac{\tau}{h^2}\right| + 2\frac{\tau}{h^2}\right) \max_j |z_j^n| + \tau \max_j |\psi_j^n|$$

— это неравенство выполняется для любого j, поэтому:

$$\max_j |z_j^{n+1}| \leqslant \left(\left|1 - 2\frac{\tau}{h^2}\right| + 2\frac{\tau}{h^2}\right) \max_j |z_j^n| + \tau \max_j |\psi_j^n| .$$

Норма погрешности на n-м временном слое считается так: $\left\|z_j^n\right\|_{C(\omega_n)} = \max_j |z_j^n|$. Тогда на $(n+1)$-м временном слое норма погрешности оценивается как

$$\left\|z_j^{n+1}\right\|_{C(\omega_{n+1})} \leqslant \left(\left|1 - 2\frac{\tau}{h^2}\right| + 2\frac{\tau}{h^2}\right) \left\|z_j^n\right\|_{C(\omega_n)} + \tau \left\|\psi_j^n\right\|_{C(\omega_n)} .$$

Если предположить, что $1 - 2\dfrac{\tau}{h^2} \geqslant 0$ (накладываем ограничения на шаг), то

$$\left\|z_j^{n+1}\right\|_{C(\omega_{n+1})} \leqslant \left\|z_j^n\right\|_{C(\omega_n)} + \tau \left\|\psi_j^n\right\|_{C(\omega_n)} ,$$

причем эта оценка выполняется для любого n. Применив оценку рекурсивно n раз, получим:

$$\left\|z_j^n\right\|_{C(\omega_n)} \leqslant \left\|z_j^0\right\|_{C(\omega_0)} + \tau \sum_{k=0}^{n-1} \left\|\psi_j^k\right\|_{C(\omega_k)} ,$$

но $\left\|z_j^0\right\|_{C(\omega_0)} = 0$ согласно постановке задачи, поэтому:

$$\left\|z_j^n\right\|_{C(\omega_n)} \leqslant \tau \sum_{k=0}^{n-1} \left\|\psi_j^k\right\|_{C(\omega_k)} .$$

Напомним, что у нас получена следующая оценка на невязку: $\psi_j^n = O(\tau + h^2)$. Обозначим $\left\|\psi_j^k\right\|_{C(\omega_k)} = M_k(\tau + h^2)$, $\max\limits_{k=\overline{0,\,n-1}} M_k = \overline{M}$, тогда получим:

$$\left\|z_j^n\right\|_{C(\omega_n)} \leqslant \tau(\tau + h^2) \sum_{k=0}^{n-1} M_k \leqslant \tau n \overline{M}(\tau + h^2).$$

Напомним, что $\tau n = t_n$ и не зависит от τ, тогда обозначим $\tau n \overline{M} = \overline{\overline{M}}$ — не зависящая от τ константа, и получим:

$$\left\|z_j^n\right\|_{C(\omega_n)} \leqslant \overline{\overline{M}}(\tau + h^2)$$

— то есть полученная разностная схема имеет первый порядок точности по τ и второй — по h.

Структура разностной схемы и задачи для погрешности одинакова, разница только в том, что в разностной схеме правая часть равна f, а в задаче для погрешности — ψ. Значит, по аналогии (то есть проведя те же действия) получим, что $\left\|y_j^n\right\|_{C(\omega_n)} \leqslant \left\|y_i^0\right\|_{C(\omega_0)} + \tau \sum\limits_{k=0}^{n-1} \left\|\varphi_j^k\right\|_{C(\omega_k)}$. Это дискретный аналог принципа максимума для уравнения теплопроводности (подробнее он описан в курсе «Уравнения математической физики»). Он говорит о том, что решение краевой задачи для уравнения теплопроводности устойчиво по начальным данным и по правой части. Но вспомним, что эта оценка получена при ограничении шагов дискретизации $\frac{\tau}{h^2} \leqslant \frac{1}{2}$, то есть разностная схема, которую мы исследуем, скорее всего, является условно устойчивой.

Далее мы рассмотрим некоторый метод исследования разностной схемы на устойчивость, называемый **методом гармоник**.

Сопоставим рассматриваемой разностной схеме однородное разностное уравнение:

$$\frac{y_j^{n+1} - y_j^n}{\tau} = \frac{y_{j+1}^n - 2y_j^n + y_{j-1}^n}{h^2}. \tag{5.31}$$

Исследуем полученное уравнение на решения вида:

$$y_j^n = q^n e^{ijh\varphi}, \tag{5.32}$$

где q, φ — некоторые параметры.

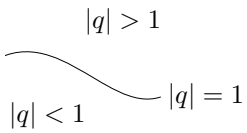
Подставим такое y_j^n в (5.31) и сократим:

$$\begin{aligned}
\frac{q-1}{\tau} &= \frac{1}{h^2}\left(e^{ih\varphi} - 2 + e^{-ih\varphi}\right) \quad \Longrightarrow \\
\frac{q-1}{\tau} &= \frac{1}{h^2}(2\cos(h\varphi) - 2) \quad \Longrightarrow \\
\frac{q-1}{\tau} &= -\frac{2}{h^2} \cdot 2\sin^2\tfrac{h\varphi}{2} \quad \Longrightarrow \\
q &= 1 - 4\frac{\tau}{h^2}\sin^2\tfrac{h\varphi}{2}.
\end{aligned}$$

Из формулы (5.32) нетрудно заметить, что если $|q| \leqslant 1$, то ограниченность начального условия влечет ограниченность в любой момент времени n. То есть сеточная функция будет устойчива. Если же $|q| > 1$ при каких-то τ и h, то решения разностной схемы y_j^n будут расти с ростом n. Отсюда следует необходимое условие сходимости разностной схемы — $|q| \leqslant 1$.

В нашем случае оно перепишется так:

$$\left|1 - 4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2}\right| \leqslant 1.$$



Раскрыв модуль, получим:

$$-1 \leqslant 1 - 4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2} \leqslant 1.$$

Правое неравенство выполнено всегда, перепишем второе неравенство:

$$\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2} \leqslant \frac{1}{2},$$

или

$$\frac{\tau}{h^2} \leqslant \frac{1}{2\sin^2\frac{h\varphi}{2}}.$$

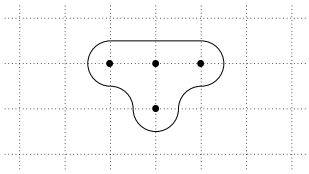
Оно должно выполняться для всех φ. Взяв минимум по правой части, получим окончательное ограничение на параметры схемы:

$$\frac{\tau}{h^2} \leqslant \frac{1}{2}.$$

Это не очень хорошо. Для примера рассмотрим $h = \frac{1}{100}$, тогда τ должно быть меньше $\frac{1}{2}10^{-4}$. Если верхнюю границу отрезка, на котором мы ищем функцию, взять $t^* = 1$, то количество шагов по времени будет $N = \frac{1}{\tau} \approx 20000$, а это, понятно, немало. Как исправить этот глобальный дефект? Можно ли, меняя шаблон, на котором происходит аппроксимация исходного дифференциального уравнения, менять соответствующий разностный аналог?

5.10 Неявная разностная схема для уравнения теплопроводности

Изменим шаблон:



Уравнения разностной схемы перепишутся следующим образом:

$$\begin{cases} \dfrac{y_j^{n+1} - y_j^n}{\tau} = \dfrac{y_{j+1}^{n+1} - 2y_j^{n+1} + y_{j-1}^{n+1}}{h^2} + \varphi_j^{n+1}, \quad j = \overline{1,\, N-1},\ n = \overline{0,\, K-1}; \\ y_j^0 = u_0(x_j), \quad j = \overline{0,\, N}; \\ y_0^{n+1} = \mu_1(t_{n+1}), \quad n = \overline{0,\, K-1}; \\ y_N^{n+1} = \mu_2(t_{n+1}), \quad n = \overline{0,\, K-1}. \end{cases} \tag{5.33}$$

Тогда невязка будет иметь вид:

$$\psi_j^{n+1} = -\frac{u_j^{n+1} - u_j^n}{\tau} + \frac{1}{h^2}\left(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}\right) + \varphi_j^{n+1}. \tag{5.34}$$

Как и в предыдущем случае, разложим все $u_{j\pm1}^{n+1}$ и u_j^n в ряд Тейлора:

$$\begin{aligned} u_{j\pm1}^{n+1} &= u_j^{n+1} \pm u_{x,j}^{n+1}h + u_{xx,j}^{n+1}\frac{h^2}{2} \pm u_{xxx,j}^{n+1}\frac{h^3}{6} + O(h^4); \\ u_j^n &= u_j^{n+1} - u_{t,j}^{n+1}\tau + O(\tau^2). \end{aligned}$$

Теперь подставим эти разложения в формулу (5.34) и в правой части добавим и вычтем f_j^{n+1}. Тогда невязка посчитается так:

$$\psi_j^{n+1} = (-u_{t,j}^{n+1} + u_{xx,j}^{n+1} + f_j^{n+1}) + \varphi_j^{n+1} - f_j^{n+1} + O(\tau + h^2).$$

Если функцию φ_j^{n+1} взять равной f_j^{n+1} с точностью $O(\tau + h^2)$, то выражение для невязки сильно сократится и примет вид:

$$\psi_j^{n+1} = O(\tau + h^2).$$

Как видно, эта разностная схема имеет первый порядок аппроксимации по τ и второй по h.

Для исследования на устойчивость воспользуемся методом гармоник. Сопоставим нашему разностному уравнению однородное уравнение:

$$\frac{y_j^{n+1} - y_j^n}{\tau} = \frac{y_{j+1}^{n+1} - 2y_j^{n+1} + y_{j-1}^{n+1}}{h^2}.$$

Проделаем те же действия, что и в предыдущем случае. Подставив в качестве решения $y_j^n = q^n e^{ijh\varphi}$ и сократив множители, получим такое уравнение относительно параметров этого решения:

$$\frac{q-1}{\tau} = q\frac{e^{ih\varphi} - 2 + e^{-ih\varphi}}{h^2} \quad \Longrightarrow \quad \frac{q-1}{\tau} = -q\frac{4}{h^2}\sin^2\frac{h\varphi}{2}.$$

Выразим отсюда q :

$$q\left(1 + \frac{4\tau}{h^2}\sin^2\frac{h\varphi}{2}\right) = 1 \quad \Longrightarrow \quad q = \frac{1}{\left(1 + \frac{4\tau}{h^2}\sin^2\frac{h\varphi}{2}\right)}.$$

Очевидно (так как знаменатель всегда больше или равен единице), что эта неявная разностная схема абсолютно устойчива (устойчива при любых значениях τ и h).

Посмотрим, как можно получить решение разностного уравнения из системы (5.33). Перепишем его так:

$$\frac{\tau}{h^2}y_{j-1}^{n+1} - \left(1 + \frac{2\tau}{h^2}\right)y_j^{n+1} + \frac{\tau}{h^2}y_{j+1}^{n+1} = -y_j^n + \tau\varphi_j^{n+1}. \tag{5.35}$$

Обозначим для удобства

$$A_j = \frac{\tau}{h^2}, \quad B_j = \frac{\tau}{h^2}, \quad C_j = \left(1 + \frac{2\tau}{h^2}\right), \quad F_j^n = y_j^n - \tau\varphi_j^{n+1}.$$

Тогда (5.35) будет выглядеть постройнее:

$$A_j y_{j-1}^{n+1} - C_j y_j^{n+1} + B_j y_{j+1}^{n+1} = -F_j^n.$$

Рассмотрим эту систему при $n = 0$:

$$A_j y_{j-1}^1 - C_j y_j^1 + B_j y_{j+1}^1 = -F_j^0, \quad j = \overline{1,\, N-1}$$

— это система линейных алгебраических уравнений относительно $y^1 = (y_0^1,\, y_1^1,\, \ldots,\, y_N^1)$. Перепишем ее в компактном виде:

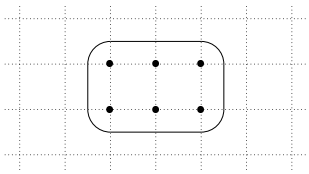
$$My^1 = F^0,$$

где матрица F^0 состоит из элементов F_j^0 (которые полностью определяются начальными условиями), а M — из коэффициентов $A_j,\ B_j,\ C_j$, расположенных на трех диагоналях (то есть матрица имеет трехдиагональный вид). Следовательно, применим метод прогонки, и мы можем найти сеточную функцию на первом временном слое.

Поступая так дальше, мы сможем определить искомую сеточную функцию на всех временных слоях («послойно» применяя метод прогонки).

5.11 Разностная схема с весами для уравнения теплопроводности

Рассмотрим теперь не минимально возможный шаблон, а «избыточный». Будем аппроксимировать производные в шести узлах. «Избыточность» схемы скомпенсируем введением некоторого параметра — весового множителя.



Соответствующая этому шаблону разностная схема такова:

$$\begin{cases} \dfrac{y_j^{n+1}-y_j^n}{\tau} = \sigma\dfrac{y_{j+1}^{n+1}-2y_j^{n+1}+y_{j-1}^{n+1}}{h^2}+(1-\sigma)\dfrac{y_{j+1}^n-2y_j^n+y_{j-1}^n}{h^2}+\varphi_j^n, & j=\overline{1,\,N-1}, \\ & n=\overline{0,\,K-1}; \\ y_j^0 = u^0(x_j), \quad j=\overline{0,\,N}; \\ y_0^{n+1} = \mu_1(t_{n+1}), \quad n=\overline{0,\,K-1}; \\ y_N^{n+1} = \mu_2(t_{n+1}), \quad n=\overline{0,\,K-1}. \end{cases} \tag{5.36}$$

— это так называемая **разностная схема с весами** для уравнения теплопроводности.

Список вопросов остается тем же.

Рассмотрим вопрос об аппроксимации. Выясним, как ведет себя невязка:

$$\psi_j^{n+\frac12} \stackrel{def}{=} -\frac{u_j^{n+1}-u_j^n}{\tau}+\sigma u_{\overline{x}x,j}^{n+1}+(1-\sigma)u_{\overline{x}x,j}^n+\varphi_j^n.$$

Как мы уже поступали, разложим функции u_j^n и u_j^{n+1} в ряд Тейлора:

$$\begin{aligned} u_j^{n+1} &= u_j^{n+\frac12}+u_{t,j}^{n+\frac12}\frac{\tau}{2}+\frac12 u_{tt,j}^{n+\frac12}\left(\frac{\tau}{2}\right)^2+O(\tau^3); \\ u_j^{n} &= u_j^{n+\frac12}-u_{t,j}^{n+\frac12}\frac{\tau}{2}+\frac12 u_{tt,j}^{n+\frac12}\left(\frac{\tau}{2}\right)^2+O(\tau^3). \end{aligned}$$

Тогда получим:

$$\psi_j^{n+\frac12} = -u_{t,j}^{n+\frac12}+O(\tau^2)+\sigma u_{\overline{x}x,j}^{n+1}+(1-\sigma)u_{\overline{x}x,j}^n+\varphi_j^n. \tag{5.37}$$

Теперь в представлении второй разностной производной разложим все вхождения функции в ряд Тейлора с членами до пятого порядка включительно:

$$u_{\overline{x}x}(x_j,t) = \frac{1}{h^2}(u(x_{j+1},t) - 2u(x_j,t) + u(x_{j-1},t)) =$$
$$= \{\mathsf{u}(\mathsf{x_j}\pm 1,\mathsf{t}) = \mathsf{u}(\mathsf{x_j},\mathsf{t}) \pm \mathsf{u_x}(\mathsf{x_j},\mathsf{t})\mathsf{h} + \mathsf{u_{xx}}(\mathsf{x_j},\mathsf{t})\frac{\mathsf{h}^2}{2} \pm \mathsf{u_{xxx}}(\mathsf{x_j},\mathsf{t})\frac{\mathsf{h}^3}{6} + \mathsf{u_{xxxx}}(\mathsf{x_j},\mathsf{t})\frac{\mathsf{h}^4}{4!} \pm \mathsf{u_{xxxxx}}(\mathsf{x_j},\mathsf{t})\frac{\mathsf{h}^5}{5!} + \mathsf{O}(\mathsf{h}^6)\} =$$
$$= u_{xx}(x_j,t) + u_{xxxx}(x_j,t)\frac{h^2}{12} + O(h^4).$$

Воспользуемся этим разложением для слагаемых в выражении (5.37):

$$u^{n+1}_{\overline{x}x,j} = u^{n+1}_{xx,j} + u^{n+1}_{xxxx,j}\cdot\frac{h^2}{12} + O(h^4) = \{\text{Разложение в ряд Тейлора с центром в точке } \mathsf{x_j^{n+\frac{1}{2}}}\} =$$
$$= u^{n+\frac{1}{2}}_{xx,j} + u^{n+\frac{1}{2}}_{xxt,j}\cdot\frac{\tau}{2} + u^{n+\frac{1}{2}}_{xxxx,j}\cdot\frac{h^2}{12} + u^{n+\frac{1}{2}}_{xxxxt,j}\cdot\frac{h^2}{12}\cdot\frac{\tau}{2} + O(\tau^2+h^4);$$
$$u^{n}_{\overline{x}x,j} = u^{n}_{xx,j} + u^{n}_{xxxx,j}\cdot\frac{h^2}{12} + O(h^4) = \{\text{Разложение в ряд Тейлора с центром в точке } \mathsf{x_j^{n+\frac{1}{2}}}\} =$$
$$= u^{n+\frac{1}{2}}_{xx,j} - u^{n+\frac{1}{2}}_{xxt,j}\cdot\frac{\tau}{2} + u^{n+\frac{1}{2}}_{xxxx,j}\cdot\frac{h^2}{12} - u^{n+\frac{1}{2}}_{xxxxt,j}\cdot\frac{h^2}{12}\cdot\frac{\tau}{2} + O(\tau^2+h^4).$$

Таким образом, выражение для невязки принимает вид:

$$\psi_j^{n+\frac{1}{2}} = -u^{n+\frac{1}{2}}_{t,j} + u^{n+\frac{1}{2}}_{xx,j} + u^{n+\frac{1}{2}}_{xxxx,j}\frac{h^2}{12} + \varphi_j^n + (\sigma-\frac{1}{2})u^{n+\frac{1}{2}}_{xxt,j}\tau + (\sigma-\frac{1}{2})u^{n+\frac{1}{2}}_{xxxxt,j}\frac{h^2}{12}\tau + O(\tau^2+h^4).$$

Добавляя и вычитая $f_j^{n+\frac{1}{2}}$, получим эквивалентное выражение:

$$\psi_j^{n+\frac{1}{2}} = -u^{n+\frac{1}{2}}_{t,j} + u^{n+\frac{1}{2}}_{xx,j} + f_j^{n+\frac{1}{2}} + u^{n+\frac{1}{2}}_{xxxx,j}\frac{h^2}{12} + \varphi_j^n - f_j^{n+\frac{1}{2}} + (\sigma-\frac{1}{2})u^{n+\frac{1}{2}}_{xxt,j}\tau + (\sigma-\frac{1}{2})u^{n+\frac{1}{2}}_{xxxxt,j}\frac{h^2}{12}\tau + O(\tau^2+h^4).$$

Согласно уравнению теплопроводности, первые три слагаемых обращаются в нуль:

$$\psi_j^{n+\frac{1}{2}} = u^{n+\frac{1}{2}}_{xxxx,j}\frac{h^2}{12} + \varphi_j^n - f_j^{n+\frac{1}{2}} + (\sigma-\frac{1}{2})u^{n+\frac{1}{2}}_{xxt,j}\tau + (\sigma-\frac{1}{2})u^{n+\frac{1}{2}}_{xxxxt,j}\frac{h^2}{12}\tau + O(\tau^2+h^4).$$

При $\sigma = \frac{1}{2}$ схема (5.36) называется симметричной. Тогда в последнем равенстве последние слагаемые обнулятся, и с помощью условия на параметр φ_j^n

$$\varphi_j^n = f_j^{n+\frac{1}{2}} + O(\tau^2+h^2)$$

мы можем достичь такого порядка аппроксимации:

$$\psi_j^{n+\frac{1}{2}} = O(\tau^2+h^2).$$

Теперь вернемся на шаг назад и воспользуемся тем, что

$$u_t = u_{xx} + f \implies u_{xxt} = u_{xxxx} + f_{xx}.$$

Тогда формула для невязки будет несколько иной:

$$\psi_j^{n+\frac{1}{2}} = \varphi_j^n - f_j^{n+\frac{1}{2}} - f^{n+\frac{1}{2}}_{xx,j}\cdot\frac{h^2}{12} + \left[(\sigma-\frac{1}{2})\tau + \frac{h^2}{12}\right]u^{n+\frac{1}{2}}_{xxt,j} + (\sigma-\frac{1}{2})u^{n+\frac{1}{2}}_{xxxxt,j}\cdot\frac{h^2}{12}\tau + O(\tau^2+h^4).$$

Взяв $\sigma = \dfrac{1}{2} - \dfrac{h^2}{12\tau}$, мы обнулим четвертое слагаемое, а коэффициент при пятом оценится как $O(h^4)$. Осталось потребовать, чтобы

$$\varphi_j^n = f_j^{n+\frac{1}{2}} + f^{n+\frac{1}{2}}_{xx,j}\cdot\frac{h^2}{12} + O(\tau^2+h^4),$$

тогда порядок аппроксимации будет таков:

$$\psi_j^{n+\frac{1}{2}} = O(\tau^2 + h^4).$$

При данном σ выражение (5.36) называется разностной схемой повышенного порядка аппроксимации. При всех остальных σ порядок аппроксимации будет меньше:

$$\psi_j^{n+\frac{1}{2}} = O(\tau + h^2)$$

при условии, что

$$\varphi_j^n = f_j^{n+\frac{1}{2}} + O(\tau + h^2).$$

Теперь исследуем схему на **устойчивость** методом гармоник. Для начала запишем однородное уравнение:

$$\frac{y_j^{n+1} - y_j^n}{\tau} = \sigma \frac{y_{j+1}^{n+1} - 2y_j^{n+1} + y_{j-1}^{n+1}}{h^2} + (1-\sigma)\frac{y_{j+1}^n - 2y_j^n + y_{j-1}^n}{h^2}.$$

Подставим в качестве решения $y_j^n = q^n e^{ijh\varphi}$. Сокращая степени q, получим:

$$\frac{q-1}{\tau} = \sigma \frac{q}{h^2}(e^{ih\varphi} - 2 + e^{-ih\varphi}) + (1-\sigma)\frac{1}{h^2}(e^{ih\varphi} - 2 + e^{-ih\varphi}) \iff$$
$$\frac{q-1}{\tau} = -4\sigma \frac{q}{h^2}\sin^2\frac{h\varphi}{2} - 4(1-\sigma)\frac{1}{h^2}\sin^2\frac{h\varphi}{2}.$$

Отсюда получаем выражение для q :

$$q(1 + 4\sigma\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2}) = 1 - (1-\sigma)4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2} \implies$$
$$q = \frac{1 - (1-\sigma)4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2}}{1 + 4\sigma\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2}}.$$

Для получения условий на устойчивость мы требуем, чтобы $|q| \leqslant 1$. В данном случае это дает два неравенства:

$$\begin{cases} 1 - (1-\sigma)4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2} & \leqslant & 1 + 4\sigma\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2}; \\ -1 - 4\sigma\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2} & \leqslant & 1 - (1-\sigma)4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2}. \end{cases}$$

Первое выполнено всегда, так как $\tau > 0$. Второе перепишется так:

$$-8\sigma\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2} \leqslant -4\frac{\tau}{h^2}\sin^2\frac{h\varphi}{2} + 2 \iff \sigma \geqslant \frac{1}{2} - \frac{h^2}{4\tau\sin^2\frac{h\varphi}{2}}$$

— оно должно быть выполнено при любом φ. Взяв максимум по правой части, приходим к окончательному условию для σ :

$$\sigma \geqslant \frac{1}{2} - \frac{h^2}{4\tau}.$$

Значение $\sigma = \dfrac{1}{2}$ удовлетворяет этому неравенству. Это означает, что соответствующий метод является абсолютно устойчивым.

Последним нашим долгом будет обосновать возможность вычисления приближения по этой схеме. Для этого перепишем (5.36) в таком виде:

$$\sigma\frac{\tau}{h^2}y_{j+1}^{n+1} - (1 + 2\sigma\frac{\tau}{h^2})y_j^{n+1} + \sigma\frac{\tau}{h^2}y_{j-1}^{n+1} = -y_j^n - (1-\sigma)\tau y_{\overline{x}x,j}^n - \varphi_j^n.$$

Теперь почти очевидно, что мы можем получить все y_j^k ($j = \overline{0,\, N}$, $k = \overline{0,\, K}$). Действительно, фиксируем $n = 0$. Тогда мы получим систему линейных уравнений с трехдиагональной матрицей. Правые части уравнений мы можем найти, используя начальные условия (заметим, что φ_j^n мы задали при исследовании порядка аппроксимации). После этого применяется метод прогонки, после которого становятся известны все y_j^1. Теперь можно увеличить n и снова получить СЛАУ, подставив в правую часть уравнений только что найденные y_j^1. Так действуем, пока не найдем все y_j^k.

Мы рассмотрели схему для нахождения численного решения простейшей краевой задачи. Для нее существует более простое аналитическое решение, но в общем случае его может и не быть. В то же время вся методика построения разностных схем и их решение достаточно легко переносятся на более сложные задачи, к которым мы и перейдем.

5.12 Разностные схемы для уравнения теплопроводности особого типа

Разностная схема для уравнения теплопроводности с переменными коэффициентами

Рассмотрим такую краевую задачу:

$$\begin{cases} \rho(x,t)\dfrac{\partial u}{\partial t} &= \dfrac{\partial}{\partial x}\left(k(x,t)\dfrac{\partial u}{\partial x}\right) + f(x,t), \quad 0 < x < 1,\ 0 < t \leqslant T; \\ u(0,t) &= \mu_1(t), \quad 0 \leqslant t \leqslant T; \\ u(1,t) &= \mu_2(t), \quad 0 \leqslant t \leqslant T; \\ u(x,0) &= u^0(x), \quad 0 \leqslant x \leqslant 1. \end{cases}$$

Аналитического выражения для решения нет. Тем не менее, известно, что если всюду верны неравенства

$$0 < c_1 \leqslant \rho(x,t);$$
$$0 < c_2 \leqslant k(x,t),$$

то оно существует и единственно.

Теперь будем приближать u_t соответствующей разностной производной:

$$\frac{\partial u}{\partial t} \approx u_{t,j}^n,$$

а производные по x с использованием интегро-интерполяционного метода можно представить следующим образом:

$$\frac{\partial}{\partial x}\left(k(x,t)\frac{\partial u}{\partial x}\right) \approx \left(a u_{\overline{x}}\right)_{x,j} = \frac{1}{h}\left(a_{j+1}\frac{u_{j+1}-u_j}{h} - a_j\frac{u_j - u_{j-1}}{h}\right),$$

где a_j вычисляются по такой формуле:

$$a_j = \left[\frac{1}{h}\int\limits_{x_{j-1}}^{x_j}\frac{dx}{k(x,t)}\right]^{-1} \approx k(x_{j-\frac{1}{2}}, t).$$

Используя шаблон из шести точек, мы построим разностную схему. Выкладки аналогичны предыдущим схемам, поэтому опустим их и приведем только окончательный результат:

$$\begin{cases} \rho(x_j,t)\dfrac{y_j^{n+1}-y_j^n}{\tau} &= \sigma\left(a y_{\overline{x}}\right)_{x,j}^{n+1} + (1-\sigma)\left(a y_{\overline{x}}\right)_{x,j}^{n} + \varphi_j^n; \quad j = \overline{1,\, N-1},\ n = \overline{0,\, K-1} \\ y_j^0 &= u^0(x_j), \quad j = \overline{0,\, N}; \\ y_0^n &= \mu_1(t_n), \quad n = \overline{1,\, K}; \\ y_N^n &= \mu_2(t_n), \quad n = \overline{1,\, K}. \end{cases} \tag{5.38}$$

Здесь остаются нефиксированными момент времени t в первом уравнении (от него зависят a_j и ρ) и параметр метода σ. Взяв $\sigma = \dfrac{1}{2}$ и $t = t_{n+\frac{1}{2}}$, мы можем получить такую оценку на порядок аппроксимации:

$$\psi_j^{n+\frac{1}{2}} = O(\tau^2 + h^2).$$

В противном случае оценка будет похуже:

$$\psi_j^n = O(\tau + h^2).$$

Наконец, общая формула для получения y будет такова:

$$A_j y_{j+1}^{n+1} - C_j y_j^{n+1} + B_j y_{j-1}^{n+1} \;=\; -F_j.$$

При этом правая часть уравнения зависит только от y_n. Это означает, что, решая данное уравнение послойно (с $n = 0$) и используя начальные условия, мы можем найти все y_j^k. Все эти умозаключения абсолютно идентичны тем, что были в конце предыдущего параграфа.

Разностная схема для нелинейного уравнения теплопроводности

Исследуем случай, когда уравнение теплопроводности в краевой задаче имеет такой вид:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k(u,x,t)\frac{\partial u}{\partial x}\right) + f(u,x,t).$$

— это означает, что коэффициенты при производных зависят еще и от искомой функции u, причем, вообще говоря, нелинейно.

В этом случае рекомендуется использовать неявные разностные схемы, так как они чаще всего абсолютно устойчивы. Приведем пример:

$$\frac{y_j^{n+1} - y_j^n}{\tau} = \left(a y_{\overline{x}}\right)_{x,j}^{n+1} + f(y_j^{n+1}),$$

при этом коэффициенты a зависят еще и от y. Расписав разностные производные, получим такую формулу:

$$\frac{y_j^{n+1} - y_j^n}{\tau} = \frac{1}{h}\left[a_{j+1}(y_j^{n+1}) \cdot \frac{y_{j+1}^{n+1} - y_j^{n+1}}{h} - a_j(y_j^{n+1}) \cdot \frac{y_j^{n+1} - y_{j-1}^{n+1}}{h}\right] + f(y_j^{n+1}).$$

— для каждого слоя это система нелинейных уравнений относительно y_j^{n+1}. Решается она итерационным методом следующего вида. Если обозначить за $\overset{(k)}{y_j}$ приближение для y_j^{n+1}, то формула для получения следующего приближения будет такова:

$$\frac{\overset{(k+1)}{y_j} - y_j^n}{\tau} = \frac{1}{h}\left[a_{j+1}(\overset{(k)}{y_j}) \cdot \frac{\overset{(k+1)}{y_{j+1}} - \overset{(k+1)}{y_j}}{h} - a_j(\overset{(k)}{y_j}) \cdot \frac{\overset{(k+1)}{y_j} - \overset{(k+1)}{y_{j-1}}}{h}\right] + f(\overset{(k)}{y_j})$$

— это уже система линейных уравнений с трехдиагональной матрицей. Она, в свою очередь, решается методом прогонки, и мы получаем $(k+1)$-е приближение к y_j^{n+1}. Обычно ограничиваются пятью приближениями:

$$y_j^{n+1} = \overset{(5)}{y_j}.$$

5.13 Разностная схема для уравнения колебаний

Рассмотрим стандартную краевую задачу на уравнение колебаний:

$$\left\{\begin{array}{rcll} u_{tt} & = & u_{xx} + f(x,t), & 0 < x < 1,\ 0 < t \leqslant T; \\ u(0,t) & = & \mu_1(t), & 0 \leqslant t \leqslant T; \\ u(1,t) & = & \mu_2(t), & 0 \leqslant t \leqslant T; \\ u(x,0) & = & u^0(x), & 0 \leqslant x \leqslant 1; \\ u_t(x,0) & = & \psi(x), & 0 \leqslant x \leqslant 1. \end{array}\right. \tag{5.39}$$

Решение такой задачи существует и единственно.

Введем дискретную сетку на рассматриваемой области:

$$\begin{aligned} \omega_k &= \left\{x_j = jh,\ h = \tfrac{1}{N},\ j = \overline{0,\ N}\right\}; \\ \omega_\tau &= \left\{t_n = n\tau,\ \tau = \tfrac{T}{K},\ n = \overline{0,\ K}\right\}. \end{aligned}$$

Заметим, что начальные условия дают нам значения искомой функции на границе прямоугольника.

Будем использовать уже привычные нам обозначения: $u(x_j, t_n) = u_j^n$ — точное решение в узлах сетки, y_j^n — искомое приближение.

Выпишем аналоги краевых и начальных условий:

$$\left\{\begin{array}{rcl} y_0^n & = & \mu_1(t_n); \\ y_N^n & = & \mu_2(t_n); \\ y_j^0 & = & u^0(x_j). \end{array}\right.$$

Сопоставим уравнению дискретный аналог в каждом узле. Будем использовать шаблон из пяти точек (это минимально необходимый шаблон для аппроксимации вторых производных по t и x). Приближения построим так:

$$\begin{aligned} u_{tt} &\approx u^n_{\overline{t}t,j} = \frac{1}{\tau^2}\left(u_j^{n+1} - 2u_j^n + u_j^{n-1}\right); \\ u_{xx} &\approx u^n_{\overline{x}x,j} = \frac{1}{h^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right). \end{aligned}$$

Тогда получим следующий дискретный аналог исходной задачи:

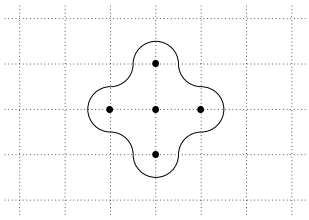
$$\left\{\begin{array}{rcll} y^n_{\overline{t}t,j} & = & y^n_{\overline{x}x,j} + \varphi_j^n, & j = \overline{1,\ N-1},\ n = \overline{1,\ K-1}; \\ y_0^n & = & \mu_1(t_n), & n = \overline{1,\ K}; \\ y_N^n & = & \mu_2(t_n), & n = \overline{1,\ K}; \\ y_j^0 & = & u^0(x_j), & j = \overline{0,\ N}; \\ \dfrac{y_j^1 - y_j^0}{\tau} & = & \psi(x_j), & j = \overline{0,\ N}. \end{array}\right. \tag{5.40}$$

Шаблон будет выглядеть следующим образом:

Стоит отметить, что для того, чтобы разностная схема была сбалансированной (то есть, чтобы не делать лишних вычислений в одном месте, а потом загрублять их в другом), необходимо, чтобы порядки аппроксимации в уравнении и в краевом условии второго типа были согласованы. Иначе, если использовать низкий (первый) порядок аппроксимации в краевом условии, то вся схема будет аппроксимировать исходную задачу с первым порядком аппроксимации.

Постараемся все данные в задаче приблизить со вторым порядком точности. Для этого разложим u_j^1 в ряд Тейлора в точке $(x_j,\ 0)$ с остаточным членом второго порядка малости:

$$\frac{u_j^1 - u_j^0}{\tau} = u^0_{t,j} + u^0_{tt,j}\frac{\tau}{2} + O(\tau^2).$$



Постараемся избавиться от $u^0_{tt,j}$. Пусть на границе также выполняется уравнение колебаний, тогда $u_{tt} = u_{xx} + f(x,0)$, и краевое условие примет вид:

$$\frac{u^1_j - u^0_j}{\tau} = u^0_{t,j} + \left(u^0_{xx,j} + f(x_j,0)\right)\frac{\tau}{2} + O(\tau^2),$$

где выражение $u^0_{xx,j} + f(x_j,0)$ точно вычислимо из начальных условий.

Тогда разностную схему можно переписать в виде:

$$\begin{cases} y^n_{\overline{t}t,j} &= y^n_{\overline{x}x,j} + \varphi^n_j, & j = \overline{1,\, N-1},\ n = \overline{1,\, K-1};\\ y^n_0 &= \mu_1(t_n), & n = \overline{1,\, K};\\ y^n_N &= \mu_2(t_n), & n = \overline{1,\, K};\\ y^0_j &= u^0(x_j), & j = \overline{0,\, N};\\ \dfrac{y^1_j - y^0_j}{\tau} &= \psi(x_j) + \dfrac{\tau}{2}\left(u^0_{xx}(x_j) + f^0_j\right), & j = \overline{0,\, N}. \end{cases}$$

Порядок аппроксимации краевого условия будет $O(\tau^2)$. Потребуем в задаче второй порядок аппроксимации и в первом уравнении. Для этого достаточно взять $\varphi^n_j = f(x_j,t_n) + O(h^2+\tau^2)$.

Таким образом, разностный аналог исходного ДУ будет его аппроксимировать со вторым порядком по τ и по h.

Рассмотрим вопрос о поиске решения. Найдем из построенной разностной схемы выражение для y^{n+1}_j :

$$y^{n+1}_j = 2y^n_j - y^{n-1}_j + \tau^2 y^n_{\overline{x}x,j} + \tau^2\varphi^n_j. \tag{5.41}$$

Покажем,что значение сеточной функции вычислимо на всей сетке. Рассмотрим формулу (5.41) при $n = 1$:

$$y^2_j = 2y^1_j - y^0_j + \frac{\tau^2}{h^2}\left(y^1_{j+1} - 2y^1_j + y^1_{j-1}\right) + \tau^2\varphi^1_j.$$

Так как y^0_j известно из начальных условий, а y^1_j можем найти из второго краевого условия, то y^2_j находится явно на всем втором временном слое. Аналогично можно найти y^3_j и так далее, то есть определить значение искомой сеточной функции на всей сетке (как и раньше, делая это послойно).

Как видно из процесса поиска решения, эта схема — явная, и построенное решение будет единственным.

Рассмотрим вопрос устойчивости. Будем использовать привычный нам метод гармоник. Запишем однородное уравнение:

$$y^{n+1}_j - 2y^n_j + y^{n-1}_j = \frac{\tau^2}{h^2}\left(y^n_{j+1} - 2y^n_j + y^n_{j-1}\right).$$

Подставим в него решение вида $y^n_j = q^n e^{ijh\varphi}$, тогда получим:

$$q^2 - 2q + 1 = \frac{\tau^2}{h^2}q(e^{ih\varphi} - 2 + e^{-ih\varphi}) \quad\Longleftrightarrow\quad q^2 - \left(2 - 4\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)q + 1 = 0.$$

Решения этого квадратного уравнения будут такими:

$$q_{1,2} = \left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right) \pm \sqrt{\left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)^2 - 1}.$$

Выделим два случая на дискриминант (подкоренное выражение):

1. $D > 0$:

   Так как свободный член равен единице, то $|q_1||q_2| = 1$. В этом случае (можно показать, что корни не могут быть противоположными по знаку) $|q_1|$ или $|q_2|$ больше единицы, поэтому устойчивости не будет при данных значениях параметров.

2. $D \leqslant 0$:

   Аналогично, $|q_1||q_2| = 1$. Найдем значение абсолютной величины корня:

$$|q_1|^2 = \left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)^2 + 1 - \left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)^2 = 1.$$

   Тогда $|q_1| = |q_2| = 1$, и, следовательно, решение будет устойчивым.

Распишем условие неположительности дискриминанта:

$$|1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}| \leqslant 1,$$

или, расписав:

$$-1 \leqslant 1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2} \leqslant 1.$$

Правое неравенство выполнено всегда, перепишем левое:

$$\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2} \leqslant 1.$$

В худшем случае получим $\frac{\tau^2}{h^2} \leqslant 1$, что эквивалентно $\frac{\tau}{h} \leqslant 1$ (напомним, что в уравнении теплопроводности мы получали ограничение вида $\frac{\tau}{h^2} \leqslant \frac{1}{2}$). Получается, что построенная схема является условно устойчивой.

Итак, можно выделить несколько особенностей для уравнения колебаний:
1) дополнительное условие на аппроксимацию (из начального условия);
2) решение строится послойно (впрочем, как и для многих других задач);
3) условная устойчивость с простыми ограничениями.

5.14 Разностная аппроксимация задачи Дирихле

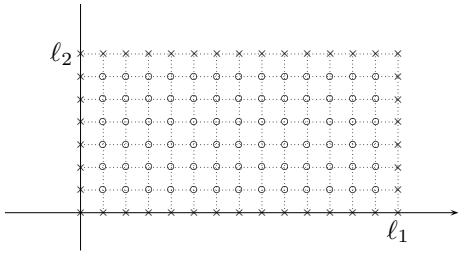
Кратко опишем численное решение задачи Дирихле на уравнение Пуассона. Пусть область, для которой задано уравнение — прямоугольник в E^2:

$$\begin{cases} u_{x_1x_1} + u_{x_2x_2} = f(x_1, x_2), & 0 < x_1 < l_1,\ 0 < x_2 < l_2; \\ u(0, x_2) = \mu_1(x_2), & 0 \leqslant x_2 \leqslant l_2; \\ u(l_1, x_2) = \mu_2(x_2), & 0 \leqslant x_2 \leqslant l_2; \\ u(x_1, 0) = \overline{\mu_1}(x_1), & 0 \leqslant x_1 \leqslant l_1; \\ u(x_1, l_2) = \overline{\mu_2}(x_1), & 0 \leqslant x_1 \leqslant l_1. \end{cases}$$

Дискретизируем область:

$$\omega_{h_{x_1}} = \left\{ x_{1i} = ih_1,\ i = \overline{0,\ N_1},\ h_1 = \frac{l_1}{N_1} \right\};$$
$$\omega_{h_{x_2}} = \left\{ x_{2j} = jh_2,\ j = \overline{0,\ N_2},\ h_2 = \frac{l_2}{N_2} \right\}.$$

Выглядеть это будет следующим образом (обозначения аналогичны предыдущим рисункам):



Запишем дискретные аналоги граничных условий:

$$\begin{cases} y_{0,j} = \mu_1(x_{2j}), & j = \overline{0,\ N_2}; \\ y_{N_1,j} = \mu_2(x_{2j}), & j = \overline{0,\ N_2}; \\ y_{i,0} = \overline{\mu_1}(x_{1i}), & i = \overline{0,\ N_1}; \\ y_{i,N_2} = \overline{\mu_2}(x_{1i}), & i = \overline{0,\ N_1}. \end{cases} \tag{5.42}$$

Производные будем приближать разностными производными со вторым порядком аппроксимации:

$$y_{\overline{x_1}x_1,ij} + y_{\overline{x_2}x_2,ij} = \varphi_{ij}, \quad i = \overline{1,\ N_1 - 1},\ j = \overline{1,\ N_2 - 1}, \tag{5.43}$$

где для достижения второго порядка аппроксимации во всем уравнении φ_{ij} считается по формуле:

$$\varphi_{ij} = f(x_{1i}, x_{2j}) + O(h_1^2 + h_2^2).$$

Найдем формулу для поиска приближенного решения. Явно получить выражение для приближений на следующем слое из (5.43) довольно сложно, поэтому сделаем хитрее.

Введем вектор всех неизвестных (заметим, **всех**, то есть получим огромный вектор), используя самый простой, регулярный способ нумерации:

$$Y = \left(y_{11},\ y_{21},\ \ldots,\ y_{(N_1-1)1},\ y_{12},\ y_{22},\ \ldots,\ y_{(N_1-1)2},\ \ldots,\ y_{1(N_2-1)},\ \ldots,\ y_{(N_1-1)(N_2-1)}\right)^T.$$

Аналогично введем вектор правой части:

$$F = \left(\varphi_{11},\ \varphi_{21},\ \ldots,\ \varphi_{(N_1-1)1},\ \varphi_{12},\ \varphi_{22},\ \ldots,\ \varphi_{(N_1-1)2},\ \ldots,\ \varphi_{1(N_2-1)},\ \ldots,\ \varphi_{(N_1-1)(N_2-1)}\right)^T.$$

Объединив, используя эти обозначения, уравнения (5.42) и (5.43), получим уравнение:

$$AY = F.$$

Из структуры шаблона следует такой вид матрицы A (она будет иметь пять ненулевых «диагона-

лей»; i, j, s — некоторые параметры, зависящие от N_1, N_2):

$$A = \begin{pmatrix}
a_{11} & a_{12} & 0 & 0 & \dots & 0 & a_{1j} & 0 & 0 & \dots & 0 \\
a_{21} & a_{22} & a_{23} & 0 & \dots & 0 & 0 & a_{2j} & 0 & \dots & 0 \\
0 & a_{32} & a_{33} & a_{34} & \dots & 0 & 0 & 0 & a_{3j} & \dots & 0 \\
 & & & \dots & & & & & & & \\
0 & \dots & 0 & a_{j(j-1)} & a_{jj} & a_{j(j+1)} & 0 & \dots & 0 & 0 & a_{j1} \\
 & & & \dots & & & & & & & \\
a_{i1} & 0 & \dots & 0 & a_{i(i-1)} & a_{ii} & a_{i(i+1)} & 0 & \dots & 0 & 0 \\
 & & & \dots & & & & & & & \\
0 & \dots & a_{(s-2)(i-2)} & 0 & 0 & \dots & 0 & a_{(s-2)(s-3)} & a_{(s-2)(s-2)} & a_{(s-2)(s-1)} & 0 \\
0 & \dots & 0 & a_{(s-1)(i-1)} & 0 & 0 & \dots & 0 & a_{(s-1)(s-2)} & a_{(s-1)(s-1)} & a_{(s-1)s} \\
0 & 0 & \dots & 0 & a_{si} & 0 & 0 & \dots & 0 & a_{s(s-1)} & a_{ss}
\end{pmatrix}$$

— так называемая матрица ленточной структуры[5].

Мы, кстати, опять получили сеточное уравнение (напомним, подробные указания о том, как их решать, можно найти в [2]).

[5] названа в честь ленточного червя (тоже большая гадость).

Оглавление

*Литература

[1] А. А. Самарский, А. В. Гулин. *"Численные методы"*.

[2] А. А. Самарский, Е. С. Николаев. *"Методы решения сеточных уравнений"*.