Министерство образования Российской Федерации

Иркутский государственный университет

В.А. Срочко

Численные методы

Курс лекций

Иркутск 2003

Печатается по решению редакционно-издательского совета
Иркутского государственного университета

УДК 519.6

Срочко В.А. Численные методы: Курс лекций. – Иркутск: Иркут. ун-т, 2003. - 168с.

Излагается традиционный материал курса "Численные методы"(алгебра, математический анализ, дифференциальные уравнения), адаптированный для студентов математических специальностей вузов. Представлены только теоретические аспекты курса, поэтому изложение необходимо сопровождать практическими и лабораторными занятиями по соответствующим пособиям.
Библиогр. 9 назв.

ОГЛАВЛЕНИЕ

2. Численные методы математического анализа

3. Численные методы решения дифференциальных уравнений

Введение

Вычислительная математика – это раздел прикладной математики, в котором проводится разработка, обоснование и реализация (на базе вычислительной техники) методов приближенного решения разнообразных задач на уровне математических моделей.

Основное содержание вычислительной математики составляют численные методы, представляющие собой упорядоченные схемы (итерационные процедуры, расчетные формулы, алгоритмы) переработки информации с целью нахождения приближенного решения рассматриваемой задачи в числовой форме.

Численные методы являются основным инструментом решения современных прикладных задач. Аналитическое решение той или иной задачи (в виде формульных соотношений) является скорее исключением, нежели правилом в силу *сложного* (вообще говоря, нелинейного) и *приближенного* (погрешности входных данных) характера исследуемых моделей. Вот почему численный анализ математических моделей – метод, алгоритм, программа, вычислительный эксперимент – является в настоящее времы актуальным и наиболее эффективным аппаратом конструктивного исследования прикладных проблем.

Следует также подчеркнуть компьютерно-ориентированный характер численных методов – в конечном итоге их реализация необходимо связана с применением вычислительной техники и программирования. Естественно, что прогресс в области вычислительной математики в немалой степени обусловлен новыми возможностями в развитии компьютерных ресурсов. Однако даже сравнительно высокая производительность современных компьютеров не снимает проблему разработки эффективных и экономичных в плане вычислительных затрат методов решения, специализированных для определенных классов задач. Проблема оптимизации (модификации, модернизации) вычислительных методов по-прежнему сохраняет свою актуальность и определяет перспективу дальнейшего развития численного анализа.

Отметим универсальный, многоплановый характер вычислительной математики, которая в качестве объектов исследования объединяет задачи, возникающие в математических, естественно-научных и гуманитарных дисциплинах. Все эти разнообразные задачи интегрируются с помощью единого общего подхода – конструктивное исследование с целью фактического

получения решения на основе применения компьютерных ресурсов.

Данный курс лекций подготовлен на основе учебных пособий [1]-[6] и предназначен в первую очередь для студентов математических специальностей. Тем не менее, отдельные разделы могут быть использованы при чтении соответствующих курсов в рамках смежных специальностей и направлений подготовки.

Материал лекций отражает теоретические аспекты стандартного курса "Численные методы"и не содержит иллюстрирующих упражнений и заданий. Поэтому в обязательном порядке изложение необходимо сопровождать семинарскими и лабораторными занятиями по изучаемым темам (см., например, учебные пособия [7]-[9]).

В курсе последовательно изучаются следующие вопросы:

1. численные методы алгебры (системы линейных алгебраических уравнений, проблема собственных значений, решение нелинейных уравнений);
2. численные методы математического анализа (приближение функций, численное дифференцирование и интегрирование);
3. численные методы решения дифференциальных уравнений (обыкновенные дифференциальные уравнения, уравнения с частными производными).

1. Численные методы алгебры

Глава 1. Методы решения линейных систем

§1. Вспомогательный материал

1. Основные обозначения, определения и утверждения

Всюду в дальнейшем R^n – векторное пространство размерности n с координатными ортами $e^i, \quad i = \overline{1,n}$. Всякий элемент $x \in R^n$ есть вектор-столбец с координатами $x_1, \ldots, x_n$. Для пары $x, y \in R^n$ обозначим операцию скалярного произведения

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i.$$

Пусть A – $(n \times n)$-матрица. Будем обозначать: A^T – транспонированная матрица, A^{-1} – обратная матрица, $\det A$ – определитель матрицы A. Всюду далее E – $(n \times n)$ - единичная матрица.

Матрица $A = \{a_{ij},\, i,j = \overline{1,n}\}$ называется
1) диагональной, если $a_{ij} = 0\ ,\ i \neq j$;
2) нижней (левой) треугольной, если $a_{ij} = 0\ ,\ i < j$;
3) верхней (правой) треугольной, если $a_{ij} = 0\ ,\ i > j$;
4) симметричной, если $A = A^T$;
5) невырожденной, если $\det A \neq 0$;
6) ортогональной, если $A^T A = E$;
7) матрицей со строгим диагональным преобладанием, если

$$|a_{ii}| > \sum_{j=1 j \neq i}^{n} |a_{ij}|, \quad i = \overline{1,n}.$$

Отметим, что определитель треугольной матрицы равен произведению ее диагональных элементов. Нетрудно проверить, что произведение двух треугольных матриц одного типа есть треугольная матрица того же типа. Матрица, обратная к треугольной, сохраняет это свойство. Если A – ортогональная матрица, то $A^T = A^{-1}$, $AA^T = E$, причем матрица A^T также является ортогональной.

Пусть A – симметричная матрица. Образуем квадратичную форму

$$\langle x, Ax \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

Матрица A называется положительно определенной, если $\langle x, Ax\rangle > 0$, $x \neq 0$, неотрицательно определенной, если $\langle x, Ax\rangle \geq 0$, $x \neq 0$.

Соответствующие обозначения: $A > 0$, $A \geq 0$.

На основе A образуем угловые подматрицы

$$A_m = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mm} \end{pmatrix}, \quad m = \overline{1,n}.$$

Тогда $\det A_m$ – угловой (ведущий) минор порядка m матрицы A.

Сформулируем необходимое и достаточное условие положительной определенности (критерий Сильвестра):

$$A > 0 \Leftrightarrow \det A_m > 0, \quad m = \overline{1,n}.$$

Число λ называется собственным значением (числом) матрицы A, если существует такой вектор $x \neq 0$, что $Ax = \lambda x$. Этот вектор называется собственным вектором матрицы A, соответствующим собственному значению λ. При этом (λ, x) – собственная пара матрицы A.

Согласно определению, собственные числа матрицы A и только они являются корнями характеристического уравнения $\det(A - \lambda E) = 0$. Это алгебраическое уравнение степени n, которое имеет на комплексной плоскости n корней $\lambda_i(A)$, $i = \overline{1,n}$.

Спектр матрицы A есть совокупность ее собственных значений:

$$\sigma(A) = \{\lambda_i(A), \quad i = \overline{1,n}\}.$$

Спектральный радиус есть максимум из модулей собственных значений:

$$\rho(A) = \max_{1 \leq i \leq n} |\lambda_i(A)|.$$

Отметим, что вырожденную матрицу A можно определить включением $0 \in \sigma(A)$.

Образуем матричный многочлен

$$p(A) = \alpha_0 E + \alpha_1 A + \alpha_2 A^2 + \ldots + \alpha_m A^m,$$

где степень матрицы определяется естественным образом

$$A^2 = AA, \ldots, A^m = \underbrace{AA \ldots A}_{m}.$$

Тогда нетрудно проверить, что

$$\lambda(p(A)) = \alpha_0 + \alpha_1\lambda(A) + \alpha_2\lambda^2(A) + \ldots + \alpha_m\lambda^m(A).$$

Отметим ряд утверждений по собственным значениям:
– спектр обратной матрицы представляется в виде

$$\sigma(A^{-1}) = \{1/\lambda_i(A), \quad i = \overline{1,n}\};$$

– собственные числа треугольной матрицы совпадают с ее диагональными элементами;
– собственные значения симметричной матрицы являются вещественными числами;
– собственные числа симметричной, положительно (неотрицательно) определенной матрицы положительны (неотрицательны).

Для $x \neq 0$ образуем выражение с симметричной матрицей A

$$r(x) = \frac{\langle x, Ax\rangle}{\langle x, x\rangle},$$

которое называется отношением Релея.

Пусть $\lambda_{\min}(A)$, $\lambda_{\max}(A)$ – минимальное и максимальное собственные числа матрицы A (границы спектра). Тогда выполняются экстремальные соотношения

$$\lambda_{\min}(A) = \min_{x\neq 0} r(x) = \min_{\langle x,x\rangle\ =1} \langle x, Ax\rangle,$$
$$\lambda_{\max}(A) = \max_{x\neq 0} r(x) = \max_{\langle x,x\rangle\ =1} \langle x, Ax\rangle,$$

т.е. имеет место двусторонняя оценка для квадратичной формы

$$\lambda_{min}(A) \leq \langle x, Ax\rangle \leq \lambda_{\max}(A), \quad \langle x, x\rangle = 1.$$

В дальнейшем постоянно используются простые следствия из правила матричного умножения. Пусть A, B $(n \times n)$ – матрицы, a^i - i-ая вектор-строка матрицы A, $i = \overline{1,n}$, b^j - j-ый вектор-столбец матрицы B, $j = \overline{1,n}$. Тогда

1) $(AB)_{ij} = \langle a^i, b^j\rangle$;

2) i-ая вектор-строка произведения AB имеет вид a^iB;

3) j-ый вектор-столбец произведения AB имеет вид Ab^j;

4) если $x \in R^n$, то

$$Bx = \sum_{j=1}^{n} x_j b^j, \qquad x^T A = \sum_{i=1}^{n} x_i a^i,$$

в частности, $Be^j = b^j$, $(e^i)^T A = a^i$.

2. Элементарные треугольные матрицы

Введем треугольные матрицы специального вида, которые используются как вспомогательные в методе Гаусса. Для каждого $m = \overline{1, n-1}$ $(n \times n)$ – матрицу

$$B_m = \begin{pmatrix} 1 & & & \\ & \dots & 1 & \\ & & b_{m+1,m} & \\ & & \dots & \dots \\ & & b_{nm} & 1 \end{pmatrix}$$

назовем элементарной нижней треугольной матрицей.

Таким образом, матрица B_m отличается от единичной элементами m-го столбца, стоящими ниже главной диагонали (поддиагональные элементы m-го столбца). Понятно, что B_m- невырожденная матрица, $\det B_m = 1$. При этом обратная матрица B_m^{-1} является также элементарной треугольной и имеет следующий вид

$$B_m^{-1} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & \dots & & \\ & & 1 & \\ & -b_{m+1,m} & 1 & \\ & \dots & \dots & \\ & -b_{nm} & & 1 \end{pmatrix}$$

Этот факт нетрудно проверить непосредственно ($B_m B_m^{-1} = E$), используя свойства матричного умножения.

Кроме того, отметим, что матрица $B = B_1 B_2 \dots B_{n-1}$ имеет следующую

структуру

$$B = \begin{pmatrix} 1 & & & & \\ b_{21} & 1 & & & \\ b_{31} & b_{32} & 1 & & \\ & \dots & \dots & & \\ & & & 1 & \\ b_{n1} & b_{n2} & \dots & b_{n,n-1} & 1 \end{pmatrix}$$

Пусть A - произвольная $(n \times n)$ - матрица со строками a^i, $i = \overline{1,n}$. Охарактеризуем структуру матрицы $C = B_m A$ по строкам c^i, $i = \overline{1,n}$. Согласно свойствам матричного умножения первые m строк матрицы A остаются без изменения: $c^i = (e^i)^T A = a^i$, $i = \overline{1,m}$. Последующие строки образуются по правилу:

$$c^{m+1} = a^{m+1} + b_{m+1,m} a^m, \quad c^{m+2} = a^{m+2} + b_{m+2,m} a^m, \quad \dots, c^n = a^n + b_{n,m} a^m.$$

Здесь $b_{m+1,m}, b_{m+2,m}, \dots, b_{n,m}$ - поддиагональные элементы, образующие матрицу B_m.

3. Матрица отражения

Пусть $p \in R^n$ – ненулевой вектор, pp^T – соответствующая ему матрица-диада

$$pp^T = \begin{pmatrix} p_1^2 & p_1 p_2 & \dots & p_1 p_n \\ p_2 p_1 & p_2^2 & \dots & p_2 p_n \\ & & \dots & \\ p_n p_1 & p_n p_2 & \dots & p_n^2 \end{pmatrix} = \{p_i p_j,\ i,j = \overline{1,n}\}.$$

Матрица отражения (вспомогательная в методе отражений) имеет вид

$$H = E - \frac{2}{\langle p,p \rangle} pp^T.$$

Нетрудно видеть, что для любого вектора $x \in R^n$

$$pp^T x = \begin{pmatrix} p_1 \ \langle p,x \rangle \\ p_2 \ \langle p,x \rangle \\ \dots \\ p_n \ \langle p,x \rangle \end{pmatrix} = \langle p,x \rangle p.$$

Следовательно,

$$Hx = x - \frac{2\langle x, p\rangle}{\langle p, p\rangle} p.$$

Отметим основные свойства матрицы H.

1) H - симметричная матрица;

2) $Hp = -p$;

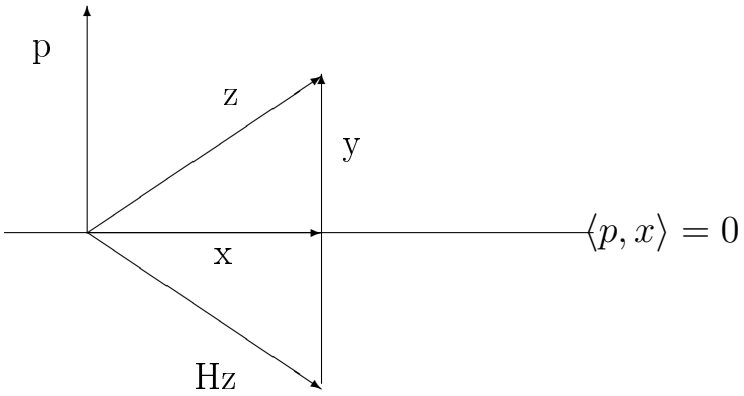
3) если $\langle x, p\rangle = 0$, то $Hx = x$;

4) H- ортогональная матрица: $H^T H = H^2 = E$.

Поясним смысл названия матрицы. Любой вектор $z \in R^n$ можно представить в виде $z = x + y$, где $\langle x, p\rangle = 0$, $y = \alpha p$, $\alpha \in R$. Тогда согласно свойствам 2) и 3)

$$Hz = Hx + Hy = x + \alpha Hp = x - \alpha p = x - y.$$

Таким образом, матрица H осуществляет ортогональное отражение относительно гиперплоскости $\langle p, x\rangle = 0$.



Нормируем вектор p в матрице H по правилу $\langle p, p\rangle = 2$. Тогда матрица $H = E - pp^T$ называется нормализованной. При этом $Hx = x - \langle x, p\rangle p$.

Пусть $m \in \{1, \ldots, n-1\}$, $a = (a_1, \ldots, a_m, a_{m+1}, \ldots, a_n)$, причем

$$\sum_{i=m+1}^{n} a_i^2 > 0.$$

Поставим *вспомогательную задачу*: найти вектор $p \in R^n$, $\langle p, p\rangle = 2$ так, чтобы соответствующая матрица отражения обладала свойством ($\stackrel{\triangle}{=}$ равно

по определению)

$$Ha \stackrel{\triangle}{=} a - \langle a, p\rangle p = \begin{pmatrix} a_1 \\ \cdot \\ a_{m-1} \\ y_m \\ 0 \\ \cdot \\ 0 \end{pmatrix}. \qquad (1)$$

Покажем, что решение данной задачи определяется соотношениями

$$p_i = 0, \quad i = \overline{1, m-1}, \quad p_m = \frac{a_m - y_m}{\beta}, \quad p_i = \frac{a_i}{\beta}, \quad i = \overline{m+1, n}, \qquad (2)$$

где

$$y_m = -sign\ a_m \cdot (\sum_{i=m}^{n} a_i^2)^{1/2}, \quad (sign\ 0 = 1),$$

$$\beta = \sqrt{y_m(y_m - a_m)}.$$

Обозначим $\|a\|_m = (\sum_{i=m}^{n} a_i^2)^{1/2}$. Отметим, что $\|a\|_m^2 = y_m^2$ и проверим, что выражение под знаком корня положительно:

$$y_m(y_m - a_m) = \|a\|_m^2 + |a_m| \cdot \|a\|_m > 0$$

(в силу предположения $\|a\|_{m+1}^2 > 0$).

Проверим условие нормировки

$$\langle p, p\rangle = \frac{1}{\beta^2}((a_m - y_m)^2 + \|a\|_{m+1}^2) = \frac{1}{\beta^2}(\|a\|_m^2 - 2y_m a_m + y_m^2) =$$

$$= \frac{1}{\beta^2}(2y_m^2 - 2y_m a_m) = \frac{2}{\beta^2} y_m(y_m - a_m) = 2$$

Подсчитаем скалярное произведение

$$\langle a, p\rangle = \frac{1}{\beta}(a_m(a_m - y_m) + \|a\|_{m+1}^2) = \frac{1}{\beta}(\|a\|_m^2 - a_m y_m) = \frac{y_m(y_m - a_m)}{\beta} = \beta.$$

Найдем координаты вектора Ha (проверим свойство (1))

$$(Ha)_i = a_i, \quad i = \overline{1, m-1},$$

$$(Ha)_m = a_m - \beta\frac{a_m - y_m}{\beta} = y_m,$$

$$(Ha)_i = a_i - \beta\frac{a_i}{\beta} = 0, \quad i = \overline{m+1, n}.$$

Таким образом, для любого вектора $a \in R^n$ и индекса $m \in \{1, \ldots, n-1\}$ вектор p с координатами (2) обеспечивает матрице H свойство (1). Подчеркнем зависимость этой матрицы от индекса m : $H = H_{(m)}$. В частности,

$$H_{(1)}a = \begin{pmatrix} y_1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix}; \quad H_{(2)}a = \begin{pmatrix} a_1 \\ y_2 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix}; \quad \cdots; \quad H_{(n-1)}a = \begin{pmatrix} a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_{n-2} \\ y_{n-1} \\ 0 \end{pmatrix}.$$

Отметим одно свойство матрицы $H_{(m)}$. Пусть $c = (c_1, \ldots, c_{m-1}, 0, \ldots, 0)$. Тогда, в силу (2) $\langle c, p \rangle = 0$, т.е. с учетом свойства 3) $H_{(m)}c = c$.

4. Нормы векторов и матриц

Для $x \in R^n$ введем семейство норм Гельдера с показателем $p \geq 1$

$$\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}.$$

Выделим наиболее распространенные из этих норм:

1) $\|x\|_1 = \sum_{i=1}^n |x_i|$ – октаэдрическая,

2) $\|x\|_2 = (\sum_{i=1}^n x_i^2)^{1/2}$ – сферическая (евклидова),

3) $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$ – кубическая (чебышевская).

Введем понятие матричной нормы. Пусть M_n – множество $(n \times n)$-матриц.

Определение. *Функция $\|\cdot\| : M_n \to R$ называется матричной нормой, если для всех матриц $A, B \in M_n$ выполнены следующие аксиомы:*

1) $\|A\| \geq 0$, $\|A\| = 0 \Leftrightarrow A = 0$,

2) $\|\alpha A\| = |\alpha| \|A\|$ *для любого числа* α,

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

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

Приведем наиболее употребительные матричные нормы:

1) $\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^n |a_{ij}|$ – максимальная столбцовая норма,

2) $\|A\|_2 = \max_{1\le i\le n} \sqrt{\lambda_i(A^TA)}$ – спектральная норма,

3) $\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}|$ – максимальная строчная норма,

4) $\|A\|_F = (\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2)^{1/2}$ – норма Фробениуса (евклидова).

Отметим, что $\|A\|_1 = \|A^T\|_\infty$. В отношении $\|A\|_2$ заметим следующее. Матрица A^TA симметрична и неотрицательно определена:

$$\langle x, A^TAx\rangle = \langle Ax, Ax\rangle \ge 0, \quad x \in R^n.$$

Поэтому ее собственные значения $\lambda_i(A^TA)$ вещественны и неотрицательны, т.е. выражение для $\|A\|_2$ вполне корректно. Если матрица A симметрична, то

$$\lambda_i(A^TA) = \lambda_i(A^2) = \lambda_i^2(A).$$

Следовательно,

$$\|A\|_2 = \max_{1\le 1\le n} \sqrt{\lambda_i^2(A)} = \max_{1\le i\le n} |\lambda_i(A)| = \rho(A).$$

Таким образом, *спектральная норма симметричной матрицы равна ее спектральному радиусу.*

Установим связь между векторными и матричными нормами.

Определение. *Норма матрицы* $\|A\|$ *называется согласованной с векторной нормой* $\|x\|$, *если для любого* $x \in R^n$ *справедливо неравенство* $\|Ax\| \le \|A\| \cdot \|x\|$.

Определение. *Матричная норма* $\|A\|$ *называется подчиненной векторной норме* $\|x\|$, *если*

$$\|A\| = \max_{x\ne 0} \|Ax\|/\|x\|.$$

Укажем эквивалентное выражение для подчиненной нормы

$$\|A\| = \max_{\|x\|=1} \|Ax\|.$$

Подчиненную норму называют также операторной, индуцированной.

Лемма 1. *Пусть* $\|\cdot\|$ *- заданная векторная норма. Подчиненная матричная норма является согласованной с данной векторной нормой. Среди всех согласованных матричных норм подчиненная норма является наименьшей.*

Доказательство. Согласно определению подчиненной нормы имеем

$$\|A\| \geq \|Ax\|/\|x\|, \quad x \neq 0.$$

Отсюда получаем условие согласования $\|Ax\| \geq \|A\| \cdot \|x\|$, $x \in R^n$.

Допустим, вопреки второму утверждению леммы, что имеется норма $\|A\|_0 < \|A\|$, согласованная с данной векторной нормой

$$\|Ax\| \leq \|A\|_0 \cdot \|x\|, \quad x \in R^n.$$

Отсюда, для $x \neq 0$ имеем $\|Ax\|/\|x\| \leq \|A\|_0$. Следовательно,

$$\max_{x \neq 0} \|Ax\|/\|x\| \leq \|A\|_0 < \|A\|,$$

что противоречит определению подчиненной нормы. □

Следствие. *Пусть норма $\|A\|$ согласована с $\|x\|$ причем существует вектор $\overline{x} \neq 0$, для которого $\|A\overline{x}\| = \|A\| \cdot \|\overline{x}\|$. Тогда $\|A\|$ является подчиненной для $\|x\|$.*

Лемма 2. *Матричная норма $\|A\|_p$ является подчиненной по отношению к векторной норме $\|x\|_p$, $p = 1, 2, \infty$.*

Доказательство проведем для случая $p = 1$.

Пусть $a^i \in R^n$ - i-ый столбец матрицы A. Тогда

$$\|A\|_1 = \max_{1 \leq i \leq n} \|a^i\|_1.$$

Для любого $x \in R^n$ получаем

$$\|Ax\|_1 = \|\sum_{i=1}^{n} x_i a^i\|_1 \leq \sum_{i=1}^{n} |x_i| \cdot \|a^i\|_1 \leq$$

$$\leq \|A\|_1 \sum_{i=1}^{n} |x_i| = \|A\|_1 \cdot \|x\|_1.$$

Итак, $\|A\|_1$ согласована с $\|x\|_1$: $\|Ax\|_1 \leq \|A\|_1 \|x\|_1$. Укажем вектор $\overline{x} \neq 0$, реализующий здесь равенство. Пусть max в выражении для $\|A\|_1$ достигается при $i = k$, т.е. $\|A\|_1 = \|a^k\|_1$. Возьмем $\overline{x} = e^k$, где e^k - координатный орт. Тогда

$$\|e^k\|_1 = 1, \quad \|Ae^k\|_1 = \|a^k\|_1 = \|A\|_1.$$

□

Установим связь между спектральным радиусом и матричной нормой.

Лемма 3. *Модули собственных значений матрицы A не превосходят любую из ее норм:* $|\lambda_i(A)| \le \|A\|, \quad i = \overline{1,n}$.

Доказательство. Пусть (λ, x) - произвольная собственная пара матрицы A. Образуем $(n \times n)$ – матрицу X, у которой первый столбец есть вектор x, а остальные столбцы нулевые. Тогда условие $Ax = \lambda x$ можно представить в виде матричного равенства $AX = \lambda X$. Отсюда, используя аксиомы матричной нормы, получаем

$$|\lambda| \cdot \|X\| = \|AX\| \le \|A\| \cdot \|X\|.$$

Поскольку $x \neq 0$, то $\|X\| > 0$. Тогда $|\lambda| \le \|A\|$. □

Следствие. $\rho(A) \le \|A\|$.

Выясним возможность равенства в последней оценке. Пусть A - симметричная матрица. Тогда, как известно, $\rho(A) = \|A\|_2$. Таким образом, для симметричных матриц

$$\rho(A) = \min_{\|\cdot\|} \|A\|$$

(минимум по всем нормам матрицы A).

В общем случае, для произвольной матрицы A имеет место соотношение

$$\rho(A) = \inf_{\|\cdot\|} \|A\|. \tag{3}$$

5. Матричная геометрическая прогрессия

Рассмотрим последовательность матриц $A_k, \quad k = 1, 2, \ldots$, где $A_k = \{a_{ij}^k, \ i, \ j = \overline{1,n}\}$. *Будем говорить, что последовательность* $\{A_k\}$ *сходится к матрице* $A_* = \{a_{ij}^*, \quad i, \ j = \overline{1,n}\}$, *если имеет место поэлементная сходимость:* $a_{ij}^k \to a_{ij}^*$, $k \to \infty$.

Как и в векторном случае, сходимость $A_k \to A_*$, $k \to \infty$ эквивалентна сходимости по любой норме: $\|A_k - A_*\| \to 0$, $k \to \infty$.

Изучим последовательность степеней $(n \times n)$-матрицы A (матричную геометрическую прогрессию)

$$E, A, A^2, \ldots, A^k, \ldots$$

Заметим, что $A^2 = AA$, $A^k = A^{k-1}A = AA^{k-1}$.

Выясним условия сходимости матричной последовательности $\{A^k\}$ к нулевой матрице O.

Докажем одно простое достаточное условие сходимости.

Лемма 4. *Пусть $\|\cdot\|$ – некоторая матричная норма. Если $\|A\| < 1$, то $A^k \to O$, $k \to \infty$.*

Доказательство. Нетрудно проверить, что для любой нормы имеет место неравенство $\|A^k\| \le \|A\|^k$, $k = 2, 3 \ldots$ Переходя здесь к пределу при $k \to \infty$ получаем требуемый результат. □

Следующая лемма полностью решает вопрос об условиях сходимости $A^k \to O$.

Лемма 5. *Для того, чтобы $A^k \to O$, $k \to \infty$ необходимо и достаточно, чтобы $\rho(A) < 1$.*

Доказательство. Необходимость. Пусть $A^k \to O$, $k \to \infty$, но $\rho(A) \ge 1$. Пусть (λ, x) – собственная пара матрицы A. Тогда

$$\|A^k x\| = \|A^{k-1}Ax\| = |\lambda| \cdot \|A^{k-1}x\| = \ldots = |\lambda|^k \cdot \|x\|.$$

Считая матричную и векторную нормы согласованными, имеем

$$\|A^k x\| \le \|A^k\| \cdot \|x\|.$$

Поскольку $x \ne 0$, то с учетом предыдущего $|\lambda|^k \le \|A\|^k$. Так как $\rho(A) \ge 1$, то найдется собственное число λ_s такое, что $|\lambda_s| \ge 1$. Следовательно, $\|A^k\| \ge |\lambda_s|^k \ge 1$, $k = 1, 2, \ldots$ Получили противоречие с исходным условием $A^k \to O$, $k \to \infty$. Необходимость доказана.

Достаточность. Пусть $\rho(A) < 1$. Используя соотношение (3), заключаем, что найдется матричная норма $\|A\|$ такая, что $\|A\| < 1$. Отсюда, на основании леммы 4 получаем $A^k \to O$, $k \to \infty$. □

Перейдем к рассмотрению матричного ряда

$$\sum_{k=0}^{\infty} A^k = E + A + A^2 + \ldots, \tag{4}$$

составленного из членов геометрической прогрессии $\{A^k\}$.

Лемма 6. *Для сходимости матричного ряда (4) необходимо и достаточно, чтобы $A^k \to O$, $k \to \infty$. При этом сумма ряда равна $(E - A)^{-1}$.*

Доказательство. Необходимость. Пусть ряд (4) сходится. Это значит, что сходится последовательность частичных сумм $\{S_k\}$, где $S_k = E + A + \ldots + A^k$. Следовательно, $S_k - S_{k-1} = A^k \to O$, $k \to \infty$. Необходимость доказана.

Достаточность. Пусть $A^k \to O$, $k \to \infty$. Тогда в силу леммы 5 $|\lambda_i(A)| < 1$, $i = \overline{1,n}$. Следовательно, 1 не является собственным числом матрицы A, т.е.

$$\det(A - E) = (-1)^n \det(E - A) \neq 0.$$

Это значит, что существует обратная матрица $(E - A)^{-1}$.

Рассмотрим матричное равенство

$$(E + A + \ldots + A^k)(E - A) = E - A^{k+1}.$$

Умножим его справа на матрицу $(E - A)^{-1}$

$$S_k = (E - A)^{-1} - A^{k+1}(E - A)^{-1}.$$

Перейдем здесь к пределу при $k \to \infty$. Учитывая, что $A^k \to O$, $k \to \infty$, получаем $A^{k+1}(E - A)^{-1} \to 0$, $k \to \infty$, т.е. предел правой части существует и равен $(E - A)^{-1}$. Следовательно, существует предел левой части

$$\lim_{k\to\infty} S_k = \sum_{k=0}^{\infty} A^k = (E - A)^{-1}.$$

□

§2. Линейная система. Вводные сведения

Рассмотрим систему линейных алгебраических уравнений в следующей записи

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \quad i = \overline{1,n}. \tag{1}$$

Здесь a_{ij} - коэффициенты, b_i – правые части системы (входные данные), x_j – искомые переменные.

Пусть $A = \{a_{ij}\}$ – матрица коэффициентов, $b = (b_1, \ldots, b_n)$ – вектор правых частей, $x = (x_1, \ldots, x_n)$– вектор неизвестных. Тогда система (1) представляется в векторно-матричной форме

$$Ax = b, \tag{1}$$

которая в дальнейшем является основной. Наша цель состоит в построении и изучении методов решения системы (1).

Отметим, что две системы вида (1) называются эквивалентными, если множества их решений совпадают. Наиболее типичное преобразование системы с сохранением эквивалентности связано с умножением на невырожденную матрицу:

$$Ax = b \quad \Leftrightarrow \quad DAx = Db, \quad \det D \neq 0.$$

Система (1) называется невырожденной, если $\det A \neq 0$. Невырожденная система имеет единственное решение $x = A^{-1}b$.

Выделим класс линейных систем, решение которых проводится элементарно. Это системы с треугольными матрицами *(треугольные системы)*.

Пусть A - нижняя треугольная матрица ($a_{ij} = 0, \quad i < j$), т.е. система имеет вид

$$\sum_{j=1}^{i} a_{ij}x_j = b_i, \quad i = \overline{1,n}.$$

В предположении, что $\det A \neq 0$ единственное решение такой системы находится по формулам (прямая подстановка)

$$x_1 = \frac{b_1}{a_{11}}, \quad x_i = \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j), \quad i = \overline{2,n}.$$

Аналогично, пусть A - верхняя треугольная матрица ($a_{ij} = 0, \quad i > j$), т.е. система имеет вид

$$\sum_{j=i}^{n} a_{ij}x_j = b_i, \quad i = \overline{1,n}.$$

Тогда ее решение от x_n до x_1 в случае $\det A \neq 0$ определяется формулами (обратная подстановка)

$$x_n = \frac{b_n}{a_{nn}}, \quad x_i = \frac{1}{a_{ii}}(b_i - \sum_{j=i+1}^{n} a_{ij}x_j), \quad i = \overline{n-1,1}.$$

Одной из базовых идей при построении методов решения общих линейных систем является последовательное преобразование (редукция) исходной системы к эквивалентной системе с треугольной матрицей.

Методы решения линейных систем можно разделить на две группы: *точные* (прямые) и *итерационные*.

Точные методы позволяют найти точное решение системы за конечное число арифметических операций (входные данные заданы точно, все операции выполняются точно, без округлений). Сравнение различных точных методов проводится обычно по числу арифметических действий, необходимых для получения решения в системе из n уравнений (как правило, явно указывается главная часть этого числа относительно n). К примеру, для решения системы из n уравнений с треугольной матрицей по вышеприведенным формулам требуется $n^2/2 + O(n)$ арифметических операций (точное число всех арифметических операций $[\frac{n(n-1)}{2} + n + n - 1]$).

Итерационные методы – это методы последовательных приближений $x^0, x^1, \ldots, x^k, \ldots$ к точному решению x^*. Переход $x^k \Rightarrow x^{k+1}$ называется k-ой итерацией метода, величина $\|x^k - x^*\|$ – погрешностью k-го приближения. Итерационный метод сходится, если $x^k \to x^*$, $k \to \infty$ ($\|x^k - x^*\| \to 0$, $k \to \infty$).

Пусть $\varepsilon > 0$ – заданная погрешность приближенного решения (требуется найти решение x^* с точностью ε). Вектор x^ε называется ε - приближенным решением системы (1), если $\|x^\varepsilon - x^*\| \leq \varepsilon$ (решение с точностью ε).

Сходящийся метод находит ε-приближенное решение за конечное число итераций: для любого $\varepsilon > 0$ найдется номер итерации k_ε такой, что $\|x^k - x^*\| \leq \varepsilon$, $k \geq k_\epsilon$.

Важной характеристикой итерационного метода является оценка погрешности, т.е. неравенство вида $\|x^k - x^*\| \leq \Delta_k$, где величина Δ_k не зависит от x^* (может быть вычислена). Если $\Delta_k \to 0$, $k \to \infty$, то метод является сходящимся. Если $\Delta_k \leq \varepsilon$ (условие остановки), то x^k – ε-приближенное решение.

Сходящиеся итерационные методы сравниваются между собой по скорости сходимости.

Определение 1. *Последовательность $\{x^k\}$ сходится к x^* с линейной скоростью, если*

$$\|x^{k+1} - x^*\| \leq q\|x^k - x^*\|, \quad k = 0, 1, \ldots$$

при условии $q \in (0, 1)$.

С помощью индукции получаем (оценка скорости сходимости)

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

Линейная скорость сходимость $\Leftrightarrow$ сходимость со скоростью геометрической прогрессии (знаменатель q).

Определение 2. *Последовательность* $\{x^k\}$ *сходится к* x^* *с квадратичной скоростью, если*

$$\|x^{k+1} - x^*\| \le q\|x^k - x^*\|^2, \quad k = 0, 1, \ldots$$

при условии $q\|x^0 - x^*\| \in (0, 1)$.

Отсюда по индукции получаем характеризацию квадратичной скорости

$$\|x^{k+1} - x^*\| \le \frac{1}{q}(q\|x^0 - x^*\|)^{2^{k+1}}.$$

§3. Метод Гаусса

1. Схема единственного деления

Классическая идея метода Гаусса – приведение исходной системы к треугольному виду с помощью последовательного исключения неизвестных. Существует несколько вычислительных схем метода.

Представим стандартную версию метода – схему единственного деления.

Рассмотрим невырожденную линейную систему в развернутой форме

$$\begin{array}{lllll} a_{11}x_1 + & a_{12}x_2 + & \ldots & + a_{1n}x_n & = b_1, \\ \ldots & \ldots & \ldots & & \\ a_{n1}x_1 + & a_{n2}x_2 + & \ldots & + a_{nn}x_n & = b_n. \end{array} \tag{1}$$

Шаг 1. Предполагая, что $a_{11} \neq 0$, составим отношения

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

Числа l_{i1} назовем множителями первого шага. Умножим первое уравнение системы (1) на l_{i1} и вычтем из i-го уравнения, $i = \overline{2, n}$. В результате переменная x_1 исключается из i-го уравнения, что приводит к следующей системе

$$\begin{array}{lllll} a_{11}x_1 + & a_{12}x_2 + & \ldots & + a_{1n}x_n & = b_1, \\ & a_{22}^{(1)}x_2 + & \ldots & + a_{2n}^{(1)}x_n & = b_2^{(1)}, \\ & \ldots & \ldots & \ldots & \\ & a_{n2}^{(1)}x_2 + & \ldots & + a_{nn}^{(1)}x_n & = b_n^{(1)}. \end{array} \tag{1.1}$$

Формулы преобразования коэффициентов при переходе от (1) к (1.1) имеют вид

$$a_{ij}^{(1)} = a_{ij} - l_{i1}a_{1j}, \quad b_i^{(1)} = b_i - l_{i1}b_1, \quad i,j = \overline{2,n}.$$

Пусть $A_{(1)}$ матрица коэффициентов, $b_{(1)}$ – вектор правых частей системы (1.1). Тогда после первого шага система (1) имеет вид $A_{(1)}x = b_{(1)}$.

В элементарном плане шаг 1 метода исключения означает, что переменная x_1 находится из первого уравнения и подставляется в последующие уравнения системы (1).

Если в системе (1) $a_{11} = 0$, то в первом столбце матрицы A найдется ненулевой элемент a_{i1} (в силу невырожденности A). В этом случае до начала шага 1 переставляют уравнения с номерами 1 и i после чего исключение x_1 проводится прежним образом.

Шаг 2. Считая, что $a_{22}^{(1)} \neq 0$, вычислим множители второго шага

$$l_{i2} = \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \quad i = \overline{3,n}.$$

Умножим второе уравнение системы (1.1) на l_{i2} и вычтем из i-го уравнения, $i = \overline{3,n}$. В результате получим систему вида

$$\begin{array}{rrrrrl} a_{11}x_1 + & a_{12}x_2 + & a_{13}x_3 + & \dots & + a_{1n}x_n & = b_1, \\ & a_{22}^{(1)}x_2 + & a_{23}^{(1)}x_3 + & \dots & + a_{2n}^{(1)}x_n & = b_2^{(1)}, \\ & & a_{33}^{(2)}x_3 + & \dots & + a_{3n}^{(2)}x_n & = b_3^{(2)}, \\ & \dots & \dots & \dots & & \\ & & a_{n3}^{(2)}x_3 + & \dots & + a_{nn}^{(2)}x_n & = b_n^{(2)}. \end{array} \tag{1.2}$$

Формулы преобразования коэффициентов при переходе от (1.1) к (1.2) имеют вид

$$a_{ij}^{(2)} = a_{ij}^{(1)} - l_{i2}a_{2j}^{(1)}, \quad b_i^{(2)} = b_i^{(1)} - l_{i2}b_2^{(1)}, \quad i,j = \overline{3,n}.$$

В векторно-матричной форме после второго шага преобразований исходная система имеет вид $A_{(2)}x = b_{(2)}$, где $A_{(2)}$ - матрица коэффициентов, $b_{(2)}$ – вектор правых частей системы (1.2).

Если $a_{22}^{(1)} = 0$, то необходимо переставить второе уравнение системы (1.1) с одним из нижеследующих.

Дальнейший ход процесса вполне понятен. После $(n-1)$-го шага перемен-

ная x_{n-1} исключена из n-го уравнения, что приводит к треугольной системе

$$\begin{array}{lllll} a_{11}x_1 + & a_{12}x_2 + & a_{13}x_3 + & \dots & + a_{1n}x_n = b_1, \\ & a_{22}^{(1)}x_2 + & a_{23}^{(1)}x_3 + & \dots & + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ & & a_{33}^{(2)}x_3 + & \dots & + a_{3n}^{(2)}x_n = b_3^{(2)}, \\ & \dots & \dots & \dots & \\ & & & & a_{nn}^{(n-1)}x_n = b_n^{(n-1)}. \end{array} \qquad (1.n-1)$$

Решение полученной системы от x_n к x_1 проводится элементарно, что и завершает реализацию метода Гаусса. Отметим векторно-матричный вариант системы (1.n-1): $A_{(n-1)}x = b_{(n-1)}$.

Подчеркнем, что в силу характера преобразований на каждом шаге метода Гаусса системы (1), (1.1), (1.2), ..., $(1.n-1)$ эквивалентны.

Переход от системы (1) к треугольной системе $(1.n-1)$ *называется прямым ходом метода Гаусса. Процесс последовательного вычисления компонент решения из системы* $(1.n-1)$ *называется обратным ходом метода. Коэффициенты* $a_{11}, a_{22}^{(1)}, \dots, a_{nn}^{(n-1)}$, *на которые производится деление в процессе преобразований системы, называются ведущими элементами метода.*

Нетрудно проверить, что арифметическая трудоемкость метода Гаусса (число арифметических операций) равна $\frac{2}{3}n^3 + O(n^2)$. Таким образом, основная работа при реализации метода Гаусса связана с преобразованием матрицы A к треугольному виду (прямой ход).

Основным ограничением метода является предположение об отличии от нуля ведущих элементов $a_{11}, a_{22}^{(1)}, \dots, a_{nn}^{(n-1)}$. Для невырожденной системы это условие может быть выполнено за счет подходящей перестановки уравнений.

Если какой-то ведущий элемент не равен нулю, но просто близок к нему, то в процессе вычислений может происходить сильное накопление погрешностей. Для предотвращения этой ситуации разработаны модификации стандартного метода Гаусса – схемы с выбором главного элемента, которые обеспечивают, вообще говоря, численную устойчивость процедуры решения.

Рассмотрим, например, *метод Гаусса с выбором главного элемента по столбцу.*

Шаг 1. В качестве ведущего элемента выберем максимальный по модулю (главный) элемент первого столбца матрицы A. Пусть это будет $a_{i_1 1}$. Если

$i_1 > 1$, то переставим уравнения с номерами $1, i_1$ и реализуем первый шаг стандартного метода Гаусса.

Отметим, что в результате такого выбора все множители первого шага по модулю не больше 1.

Шаг 2. В качестве ведущего элемента выберем максимальный по модулю (главный) элемент первого столбца матрицы $\{a_{ij}^{(1)}, \quad i, j = \overline{2, n}\}$.

Пусть это будет $a_{i_2 2}^{(1)}$. Если $i_2 > 2$, то переставим уравнения $2, i_2$ в системе (1.1) и реализуем второй шаг метода Гаусса. В данном случае $|l_{i2}| \leq 1$, $i = \overline{3, n}$.

Дальнейший ход метода вполне очевиден. Отметим, что в данной схеме порядок исключения неизвестных сохраняется: $1, 2, \ldots, n$.

Аналогичным образом выглядит метод Гаусса с выбором главного элемента по строкам. В этом случае переставляются столбцы матриц A, $A_{(1)}$, $\ldots$, т.е. изменяется порядок исключения неизвестных.

Описанные модификации называют схемами с частичным выбором.

Прокомментируем метода Гаусса с выбором главного элемента по всей матрице (полный выбор). В этой схеме на первом шаге в качестве ведущего элемента выбирается максимальный по модулю элемент матрицы A. Пусть это будет $a_{i_1 j_1}$. Переставляют уравнения с номерами $1, i_1$, и столбцы матрицы A с номерами $1, \ j_1$, после чего реализуется первый шаг метода Гаусса.

На втором шаге аналогичным образом поступают с матрицей $\{a_{ij}^{(1)}, i, j = \overline{2, n}\}$ и т.д.

Понятно, что для этой схемы множители l_{ij} на всех шагах ограничены по модулю единицей.

2. Матричное описание метода

Как следует из предыдущего прямой ход метода Гаусса характеризуется последовательностью эквивалентных систем

$$Ax = b, \quad A_{(1)}x = b_{(1)}, \quad A_{(2)}x = b_{(2)}, \ \ldots, \ A_{(n-1)}x = b_{(n-1)},$$

причем в матрице $A_{(1)}$ поддиагональные элементы первого столбца равны нулю, в матрице $A_{(2)}$ поддиагональные элементы первого и второго столб-

цов равны нулю и т.д. В результате итоговая матрица $A_{(n-1)}$ является верхней треугольной.

Опишем прямой ход метода Гаусса на языке матричных умножений.

Шаг 1. Используя множители первого шага $l_{i1}, \quad i = \overline{2,n}$, образуем элементарную треугольную матрицу

$$L_{(1)} = \begin{pmatrix} 1 & & \\ -l_{21} & 1 & \\ & \dots & \\ -l_{n1} & \dots & 1 \end{pmatrix}.$$

Умножим систему (1) на матрицу $L_{(1)}$. В результате, как нетрудно проверить, получаем систему (1.1), т.е. $A_{(1)} = L_{(1)}A$, $b_{(1)} = L_{(1)}b$.

Таким образом, матрица $L_{(1)}$ реализует первый шаг метода Гаусса: в произведении $L_{(1)}A$ поддиагональные элементы первого столбца равны нулю.

Шаг 2. Используя множители второго шага $l_{i2}, \quad i = \overline{3,n}$, образуем элементарную треугольную матрицу

$$L_{(2)} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & -l_{32} & \dots & \\ & \dots & & \\ & -l_{n2} & \dots & 1 \end{pmatrix}.$$

Умножим систему (1.1) на $L_{(2)}$. В результате получаем систему (1.2), т.е. $A_{(2)} = L_{(2)}A_{(1)}$, $b_{(2)} = L_{(2)}b_{(1)}$.

Итак, матрица $L_{(2)}$ осуществляет второй шаг метода Гаусса: в произведении $L_{(2)}A_{(1)}$ поддиагональные элементы первого и второго столбцов равны нулю.

Шаг $(n-1)$. Образуем матрицу

$$L_{(n-1)} = \begin{pmatrix} 1 & & & \\ & \dots & & \\ & & 1 & \\ & & -l_{n,n-1} & 1 \end{pmatrix},$$

где $l_{n,n-1}$ – множитель $(n-1)$-го шага.

Тогда $A_{(n-1)} = L_{(n-1)}A_{(n-2)}$, $b_{(n-1)} = L_{(n-1)}b_{(n-2)}$, т.е. матрица $L_{(n-1)}$ реализует последний шаг прямого хода метода Гаусса: произведение $L_{(n-1)}A_{(n-2)}$ есть верхняя треугольная матрица.

Подведем итог. Согласно полученным формулам

$$A_{(n-1)} = L_{(n-1)}A_{(n-2)} = \ldots = L_{(n-1)}L_{(n-2)}\ldots L_{(1)}\ A,$$

т.е. треугольная матрица $A_{(n-1)}$ есть результат последовательного умножения матрицы A на элементарные треугольные матрицы $L_{(1)}, L_{(2)}, \ldots, L_{(n-1)}$.

Из предыдущего представления найдем выражение для матрицы A

$$A = LA_{(n-1)}, \quad L = L_{(1)}^{-1}L_{(2)}^{-1}\ldots L_{(n-1)}^{-1}.$$

Отметим, что

$$L_{(m)}^{-1} = \begin{pmatrix} 1 & & & & \\ & \ldots & & & \\ & & 1 & & \\ & & l_{m+1,m} & 1 & \\ & & \ldots & & \\ & & l_{nm} & & 1 \end{pmatrix}, \qquad m = \overline{1, n-1}.$$

При этом L - нижняя треугольная матрица с единичной диагональю (матрица множителей $l_{ij}, i > j$)

$$L = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ & \ldots & & \ldots & \\ l_{n1} & l_{n2} & \ldots & l_{n,n-1} & 1 \end{pmatrix}$$

Аналогичные формулы описывают преобразование вектора правых частей системы

$$b_{(n-1)} = L_{(n-1)}L_{(n-2)}\ldots L_{(1)}\ b, \quad b = Lb_{(n-1)}.$$

Таким образом, прямой ход метода Гаусса можно проводить отдельно для матрицы A (независимо от вектора b). В результате, при условии сохранения множителей на всех шагах, для матрицы A получается представление вида $A = LA_{(n-1)}$, где L - нижняя треугольная матрица с единичной диагональю, $A_{(n-1)}$ - верхняя треугольная матрица. Такое представление называют треугольным разложением матрицы A.

3. Компактная схема метода

Рассмотрим линейную систему $Ax = b$. Допустим, что матрица A представлена в виде произведения двух треугольных матриц L и U : $A = LU$, где L - нижняя треугольная $(n \times n)$ - матрица (L - Lower - нижний), U - верхняя треугольная $(n \times n)$- матрица (U - Upper- верхний). Такое представление называют LU - разложением или LU - факторизацией матрицы A.

Если для матрицы A получено LU - разложение, то решение системы $Ax = b$ сводится к последовательному решению двух треугольных систем

1) $Ly = b$ (с нижней треугольной матрицей),

2) $Ux = y$ (с верхней треугольной матрицей).

Компактная схема метода Гаусса связана с непосредственной реализацией LU - разложения матрицы A. Проведем описание этой схемы.

Будем искать для матрицы $A = \{a_{ij}\}$ LU - разложение с матрицами $L = \{l_{ij}\}$, $U = \{u_{ij}\}$, $i, j = \overline{1, n}$ при условии, что диагональные элементы матрицы L равны единице: $l_{ii} = 1, i = \overline{1, n}$. Учитывая правило матричного умножения, запишем равенство $A = LU$ поэлементно

$$a_{ij} = \sum_{k=1}^{n} l_{ik} u_{kj}, \quad i, j = \overline{1, n}.$$

Поскольку $l_{ik} = 0$ при $i < k$, $u_{kj} = 0$, $k > j$, то получаем

$$\sum_{k=1}^{\min\{i,j\}} l_{ik} u_{kj} = a_{ij}, \quad i, j = \overline{1, n}. \tag{2}$$

Отсюда, при $i \leq j$ имеем

$$u_{ij} + \sum_{k=1}^{i-1} l_{ik} u_{kj} = a_{ij}. \tag{2.1}$$

Для $i > j$

$$\sum_{k=1}^{j} l_{ik} u_{kj} = a_{ij}.$$

Переставим в этом равенстве индексы i, j :

$$\sum_{k=1}^{i} l_{jk} u_{ki} = a_{ji}, \quad i < j. \tag{2.2}$$

На основании соотношений (2.1), (2.2) искомые элементы l_{ij}, $i > j$, u_{ij}, $i \le j$ матриц L, U подсчитываются следующим образом.

Из (2.1) при $i = 1$ получаем элементы первой строки матрицы U :

$$u_{1j} = a_{1j}, \quad j = \overline{1,n}. \tag{3}$$

Из (2.2) при $i = 1$ получаются элементы первого столбца матрицы L :

$$l_{j1} = \frac{a_{j1}}{u_{11}}, \quad j = \overline{2,n} \ \ (u_{11} \neq 0). \tag{4}$$

Пусть уже вычислены первые $(i-1)$ строк матрицы U и первые $(i-1)$ столбцов матрицы L, $i = \overline{2,n}$.

Тогда элементы i-ой строки матрицы U подсчитываются согласно формулам (2.1)

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}, \quad j = \overline{i,n}. \tag{5}$$

Элементы i-го столбца матрицы L вычисляются на основе формул (2.2) $(u_{ii} \neq 0)$

$$l_{ji} = \frac{1}{u_{ii}}(a_{ji} - \sum_{k=1}^{i-1} l_{jk}u_{ki}), \quad j = \overline{i+1,n}. \tag{6}$$

Полученные соотношения для подсчета матриц L, U называют компактной схемой метода Гаусса. Фактически она реализует прямой ход метода Гаусса для матрицы A.

Проведем обоснование этой схемы. Выясним условия на матрицу A, при которых вычисления по формулам (3)-(6) осуществимы (при которых $u_{ii} \neq 0$, $i = \overline{1,n}$).

Теорема. *Матрица A допускает единственное LU - разложение (L - нижняя треугольная матрица с единичной диагональю, U - верхняя треугольная матрица) тогда и только тогда, когда все ее угловые миноры отличны от нуля:* $\det A_m \neq 0$, $m = \overline{1,n}$.

Доказательство. Из формул (3)-(6) для элементов матриц L, U следует, что однозначное LU - разложение справедливо в том и только в том случае, когда $u_{ii} \neq 0$, $i = \overline{1,n}$. Выразим эти элементы через угловые миноры матрицы A.

Пусть L_m, U_m - угловые подматрицы матриц L, U соответственно. Из соотношений (2) получаем

$$a_{ij} = \sum_{k=1}^{m} l_{ik}u_{kj}, \quad i,j = \overline{1,m}; \quad (m \ge \min\{i,j\}).$$

В матричной записи это означает, что $A_m = L_m U_m$. Поскольку $\det L_m = 1$, $\det U_m = u_{11} \dots u_{mm}$, то $\det A_m = u_{11} \dots u_{mm}$, $m = \overline{1,n}$. Отсюда

$$u_{11} = \det A_1, \quad u_{mm} = \frac{\det A_m}{\det A_{m-1}}, \quad m = \overline{2,n}.$$

Таким образом, неравенства $u_{mm} \neq 0$ и $\det A_m \neq 0$, $m = \overline{1,n}$ эквивалентны. □

Замечание 1. Отметим, что матрицы $L, A_{(n-1)}$ вычисленные по схеме единственного деления и компактной схеме одинаковы: $A_{(n-1)} = U$. Поэтому условие $\det A_m \neq 0$, $m = \overline{1,n}$ гарантирует осуществимость (применимость) схемы единственного деления. При этом диагональные элементы матрицы U совпадают с ведущими элементами схемы единственного деления:

$$u_{11} = a_{11}, \quad u_{22} = a_{22}^{(1)}, \dots, \quad u_{nn} = a_{nn}^{(n-1)}.$$

Замечание 2. Укажем элементарное преобразование невырожденной матрицы A, для которого имеет место LU-разложение.

Матрица – перестановка P – это матрица, полученная из E произвольной перестановкой строк.

Понятно, что P – ортогональная матрица (её строки образуют ортонормированную систему). Кроме того, матрица PA получается из A аналогичной перестановкой строк.

Справедливо **утверждение** [2, с.65]. *Если* $\det A \neq 0$, *то существует матрица – перестановка* P *такая, что имеет место разложение* $PA = LU$.

Таким образом, невырожденная матрица A допускает LU-разложение после некоторой перестановки строк.

4. Некоторые применения метода

Пусть требуется вычислить определитель матрицы A. Применяя прямой ход метода Гаусса для матрицы A, получаем треугольное разложение $A = LU$. Отсюда $\det A = a_{11} a_{22}^{(1)} \dots a_{nn}^{(n-1)}$. Таким образом, определитель матрицы A равен произведению ведущих элементов прямого хода метода Гаусса (схема единственного деления). В схемах с выбором главного элемента $\det A = (-1)^p a_{11} a_{22}^{(1)} \dots a_{nn}^{(n-1)}$, где p - общее число перестановок строк и

столбцов в процессе реализации прямого хода, $a_{11},\ a_{22}^{(1)}, \ldots, a_{nn}^{(n-1)}$ - ведущие элементы схемы.
Аналогичным образом можно вычислить угловые миноры матрицы A :

$$\det A_m = a_{11}a_{22}^{(1)} \ldots a_{mm}^{(m-1)}, \quad m = \overline{1,n}.$$

Пусть требуется решить серию систем с одной и той же матрицей $A : Ax = b^j,\ j = \overline{1,s}$. Реализуем для матрицы A прямой ход метода Гаусса и найдем треугольное разложение $A = LU$. Далее, для каждого $j = \overline{1,s}$ решаются две треугольные системы $Ly = b^j,\ Ux = y$. В результате вектор x является решением исходной линейной системы.

Рассмотрим, наконец, задачу обращения матрицы A. По определению, обратная матрица A^{-1} является решением матричного уравнения $AX = E$. Распишем это уравнение по столбцам матрицы X. В результате получаем n линейных систем вида $Ax = e^j,\ j = \overline{1,n}$. Решением системы $Ax = e^j$ является j-ый столбец матрицы A^{-1}. Далее действуем по вышеприведенной схеме:

$$A = LU, \quad \forall\ j = \overline{1,n} : \quad Ly = e^j, \quad Ux = y.$$

Арифметическая трудоемкость этой процедуры – $n^3 + O(n^2)$ операций.

Приведем ряд итоговых замечаний по методу Гаусса.

1. Для решения линейной системы порядка n требуется $\frac{2}{3}n^3 + O(n^2)$ арифметических операций.

2. Если после решения системы $Ax = b$ сохранено треугольное разложение матрицы A, то решение каждой новой системы с той же матрицей требует $n^2 + O(n)$ операций.

3. Метод обладает численной устойчивостью, если прямой ход не сопровождается сильным ростом модулей элементов матриц $A_{(1)}, A_{(2)}, \ldots$. С целью ограничить этот рост используют различные схемы выбора главного элемента.

При численной реализации на ЭВМ вследствие возможных погрешностей входных данных и неизбежных ошибок округления метод Гаусса дает, вообще говоря, приближенное решение $\tilde{x}$ системы $Ax = b$, т.е. вектор невязок $\tilde{r} = b - A\tilde{x}$ отличен от нуля. Опишем способ уточнения найденного решения (метод итерационного уточнения).

Введем вектор погрешности $x^* = \tilde{x} + \Delta^*$, где x^* - точное решение системы. Тогда $A\Delta^* = b - A\tilde{x} = \tilde{r}$, т.е. поправка Δ^* является точным решением

системы $Ax = \tilde{r}$. Применяя метод Гаусса, найдем приближенное решение $\tilde{\Delta}$ этой системы и уточним решение $\tilde{x}$, полагая $\tilde{\tilde{x}} = \tilde{x} + \tilde{\Delta}$. Далее процедура может повторяться для приближения $\tilde{\tilde{x}}$.

В результате общий шаг метода итерационного уточнения описывается формулами

$$r^m \simeq b - Ax^m$$

$$A\Delta^m \simeq r^m$$

$$x^{m+1} \simeq x^m + \Delta^m.$$

Здесь $m = 1, 2, \ldots, x^m$ - приближенное решение, знак $\simeq$ означает приближенную реализацию. Отметим, что поправка Δ^m находится как решение исходной системы с измененной правой частью: $Ax = b - Ax^m$. Поскольку для матрицы A известно LU - разложение, то вычисление Δ^m требует решения двух треугольных систем.

§4. Метод квадратного корня

Рассмотрим линейную систему

$$Ax = b \tag{1}$$

при условии, что A – симметричная, положительно определенная матрица

$$A^T = A, \quad \langle x, Ax\rangle > 0, \quad x \neq 0. \tag{2}$$

Систему (1) при условии (2) назовем нормальной линейной системой.

В принципе, условие (2) не является слишком ограничительным. Действительно, систему (1) с произвольной невырожденной матрицей A можно привести к эквивалентной системе вида

$$A^T Ax = A^T b. \tag{1'}$$

Здесь матрица коэффициентов $A^T A$ является симметричной и положительно определенной:

$$(A^T A)^T = A^T A,$$
$$\langle x, A^T Ax\rangle = \langle Ax, Ax\rangle = \|Ax\|^2 > 0, \quad x \neq 0.$$

Переход от (1) к (1′) называется симметризацией системы (1) или трансформацией Гаусса.

Отметим, что согласно критерию Сильвестра свойство (2) эквивалентно условию положительности всех угловых миноров матрицы A : $\det A_m > 0$, $m = \overline{1,n}$.

Для решения задачи (1), (2) возьмем за основу идею треугольного разложения и исследуем возможность представления матрицы A в виде

$$A = U^T U, \tag{3}$$

где $U - (n \times n)$ верхняя треугольная матрица с положительными диагональными элементами ($u_{ii} > 0$, $i = \overline{1,n}$).

Используя представление (3), найдем элементы u_{ij}, $i,j = \overline{1,n}$, $i \leq j$ матрицы U (при $i > j$ $u_{ij} = 0$).

Пусть $u^i = (u_{1i}, \ldots, u_{ii}, 0, \ldots, 0)$ - i-ый столбец матрицы U. Согласно правилу матричного умножения

$$(U^T U)_{ij} = \langle u^i, u^j\rangle = \sum_{k=1}^{n} u_{ki} u_{kj} =$$

$$= \sum_{k=1}^{\min\{i,j\}} u_{ki}u_{kj} = a_{ij}.$$

Тогда при $i = j$

$$\sum_{k=1}^{i} u_{ki}^2 = a_{ii}. \tag{4}$$

Если $i < j$, то

$$\sum_{k=1}^{i} u_{ki}u_{kj} = a_{ij}. \tag{5}$$

Отсюда получаем расчетные формулы метода. Для отыскания элементов первой строки матрицы U положим в (4), (5) $i = 1$. В результате

$$u_{11} = \sqrt{a_{11}}, \quad u_{1j} = \frac{a_{1j}}{u_{11}}, \quad j = \overline{2,n}. \tag{6}$$

Пусть известны элементы первых $(i-1)$ строк матрицы U. Тогда элементы i-ой строки находятся по формулам (разрешаем (4) относительно u_{ii}, (5) относительно u_{ij})

$$u_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2}, \quad u_{ij} = \frac{1}{u_{ii}}(a_{ij} - \sum_{k=1}^{i-1} u_{ki}u_{kj}), \tag{7}$$

$$j = \overline{i+1,n}, \quad i = \overline{2,n}.$$

Обоснуем корректность формул (6), (7) в рассматриваемом случае.

Пусть $A_m, U_m,\ m = \overline{1,n}$ - угловые подматрицы порядка m матриц A и U соответственно. Тогда в силу свойства "треугольности"матрицы U имеем

$$A_m = (U^T)_m U_m = U_m^T U_m,$$

т.е.

$$\det A_m = \det U_m^T \det U_m = (\det U_m)^2 = (u_{11} \dots u_{mm})^2.$$

Следовательно,

$$\det A_1 = u_{11}^2, \quad \det A_m = \det A_{m-1} \cdot u_{mm}^2, \quad m = \overline{2,n}.$$

Отсюда

$$u_{11} = \sqrt{\det A_1}, \quad u_{mm} = \sqrt{\frac{\det A_m}{\det A_{m-1}}}, \quad m = \overline{2,n}.$$

Поскольку $\det A_m > 0$, $m = \overline{1,n}$, то диагональные элементы u_{ii}, $i = \overline{1,n}$ вещественны и положительны, т.е. формулы (6), (7) действуют в области вещественных чисел (представление (3) справедливо).

Нетрудно проверить, что реализация формул (6), (7) требует $\frac{1}{3}n^3+O(n^2)$ арифметических операций (трудоемкость в 2 раза ниже, чем в методе Гаусса – следствие учета специфики системы).

После нахождения матрицы U обратный ход метода состоит в последовательном решении двух треугольных систем

$$U^T y = b, \quad Ux = y.$$

Замечание 1. Представление (3) называется разложением Холецкого для матрицы A, при этом матрица U - множитель (фактор) Холецкого.

Замечание 2. В отличие от метода Гаусса в методе квадратного корня отсутствует проблема роста модулей элементов при подсчете матрицы U. Действительно, с учетом равенства (4) имеет место оценка

$$u_{ji}^2 \le \sum_{k=1}^{i} u_{ki}^2 = a_{ii} \le \max_{1\le i\le n} a_{ii}, \quad j \le i$$

(квадрат любого элемента матрицы U не превосходит максимального из диагональных элементов матрицы A).

Замечание 3. Метод обобщается на системы с симметричной матрицей (без условия $A > 0$). В этом случае за основу берется представление $A = U^T DU$, где D- диагональная матрица: $D = diag\ (d_{11}, \ldots, d_{nn})$ с условием $|d_{ii}| = 1, \quad i = \overline{1,n}$.

§5. Метод отражений

Рассмотрим линейную систему

$$Ax = b. \tag{1}$$

Допустим, что матрица A представлена в виде произведения $A = QR$ (QR - разложение), где Q – ортогональная матрица R – верхняя треугольная матрица. Тогда система (1) легко приводится к треугольному виду

$$QRx = b \Leftrightarrow Rx = Q^T b$$

и решается элементарно.

Обоснуем принципиальную возможность QR - разложения.

Лемма. *Любая невырожденная матрица A допускает QR - разложение.*

Доказательство. Пусть A - невырожденная матрица. Тогда матрица A^TA положительно определена. Следовательно, имеет место разложение Холецкого $A^TA = R^TR$, где R - невырожденная верхняя треугольная матрица.

Введем матрицу $Q = AR^{-1}$. Проверим свойство её ортогональности

$$Q^TQ = (R^{-1})^TA^TAR^{-1} = (R^T)^{-1}R^TRR^{-1} = E.$$

Остается заметить, что $A = QR$. □

Реализовать QR - разложение можно на основе матриц отражения.

Рассмотрим систему (1). Преобразуем её к треугольному виду с помощью матриц отражения $H_{(m)}, m = \overline{1, n-1}$.

Построим матрицу $H_{(1)}$ относительно первого столбца матрицы A. Образуем матрицу $A_{(1)} = H_{(1)}A$ и вектор $b_{(1)} = H_{(1)}b$. Тогда поддиагональные элементы первого столбца матрицы $A_{(1)}$ равны нулю (этот столбец есть результат умножения матрицы $H_{(1)}$ на первый столбец матрицы A).

Построим матрицу $H_{(2)}$ для второго столбца матрицы $A_{(1)}$ и вычислим $A_{(2)} = H_{(2)}A_{(1)}, \quad b_{(2)} = H_{(2)}b_{(1)}$. Тогда поддиагональные элементы первого и второго столбцов матрицы $A_{(2)}$ равны нулю (матрица $H_{(2)}$ сохраняет первый столбец матрицы $A_{(1)}$ и зануляет поддиагональные элементы второго столбца).

Продолжая этот процесс, получим верхнюю треугольную матрицу $A_{(n-1)} = H_{(n-1)}A_{(n-2)}$ и вектор $b_{(n-1)} = H_{(n-1)}b_{(n-2)}$.

Решим треугольную систему $A_{(n-1)}x = b_{(n-1)}$, которая эквивалентна исходной:

$$Ax = b \Leftrightarrow A_{(1)}x = b_{(1)} \Leftrightarrow A_{(2)}x = b_{(2)} \Leftrightarrow \ldots \Leftrightarrow A_{(n-1)}x = b_{(n-1)}.$$

Проведем обсуждение метода. Согласно построению

$$A_{(n-1)} = H_{(n-1)}H_{(n-2)} \ldots H_{(1)}A.$$

Отсюда с учетом свойств симметричности и ортогональности матриц отражения ($H^T_{(m)} = H_{(m)}, \quad H^2_{(m)} = E, \quad m = \overline{1, n-1}$)

$$A = H_{(1)}H_{(2)} \ldots H_{(n-1)}A_{(n-1)}.$$

Введем матрицу $H = H_{(1)}H_{(2)}\dots H_{(n-1)}$. Нетрудно видеть, что она является ортогональной:

$$H^TH = H_{(n-1)}\dots H_{(2)}H_{(1)}H_{(1)}H_{(2)}\dots H_{(n-1)} = E.$$

Следовательно, $A = HA_{(n-1)}$, т.е. в результате реализации метода отражений получается QR - разложение матрицы A.

Нетрудно проверить, что арифметическая трудоемкость метода отражений равна $\frac{4}{3}n^3 + O(n^2)$, что в два раза больше, чем в методе Гаусса.

Преимуществом метода отражений является вычислительная устойчивость (отсутствие существенного роста модулей элементов матриц $A_{(m)}$ в процессе преобразований). Это связано со свойством ортогональности матриц преобразований $H_{(1)}H_{(2)}\dots\ H_{(n-1)}$, поскольку умножение на ортогональную матрицу сохраняет норму Фробениуса матрицы и евклидову норму вектора: $\|QA\|_F = \|A\|_F$, $\|Qb\|_2 = \|b\|_2$, если $Q^TQ = E$. Таким образом, в методе отражений

$$\|A_{(m)}\|_F = \|A\|_F, \quad \|b_{(m)}\|_2 = \|b\|_2, \quad m = \overline{1, n-1}.$$

§6. Обусловленность линейных систем

Рассмотрим линейную систему

$$Ax = b \tag{1}$$

с невырожденной матрицей A. Перейдем к системе с возмущенной правой частью

$$A(x + \delta x) = b + \delta b. \tag{2}$$

Здесь δb – вариация (погрешность) правой части, δx – соответствующая вариация (погрешность) решения. Установим связь между указанными вариациями.

Из (1),(2) получим уравнение для вариаций $A(\delta x) = \delta b$. Отсюда $\delta x = A^{-1}\delta b$ и имеет место оценка (в согласованных нормах)

$$\|\delta x\| \le \|A^{-1}\|\|\delta b\|. \tag{3}$$

В неравенство (3) входят абсолютные погрешности правой части и решения. Более практичной является оценка через относительные погрешности: $\frac{\|\delta b\|}{\|b\|}$, $\frac{\|\delta x\|}{\|x\|}$. На основании уравнения (1)

$$\|b\| \leq \|A\|\|x\|. \tag{4}$$

Перемножим неравенства (3), (4)

$$\|b\|\|\delta x\| \leq \|A\|\|A^{-1}\|\|\delta b\|\|x\|.$$

Отсюда получаем требуемую оценку ($b \neq 0$)

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}.$$

Величину $condA = \|A\|\|A^{-1}\|$ называют *числом обусловленности* матрицы A. Оно связывает относительные погрешности решения и правой части.

Рассмотрим второй случай, когда в системе (1) варьируется матрица A. В результате

$$(A + \delta A)(x + \delta x) = b. \tag{5}$$

Здесь δA – вариация матрицы коэффициентов, δx – соответствующая вариация решения.

Получим оценку δx через δA (в относительных величинах). Поскольку $Ax = b$, то из (5) получаем

$$A\delta x + \delta A(x + \delta x) = 0.$$

Следовательно,

$$\delta x = -A^{-1}\delta A(x + \delta x).$$

Отсюда в согласованных нормах

$$\|\delta x\| \leq \|A^{-1}\|\|\delta A\|\|x + \delta x\|.$$

Переходя к относительным величинам, получаем требуемую оценку

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta A\|}{\|A\|}.$$

Таким образом, величина $condA$ имеет одинаковый смысл как при варьировании правой части, так и при варьировании матрицы коэффициентов системы (1).

Чтобы оценить точность полученного приближенного решения $\tilde{x}$, вычисляют вектор невязок $r = A\tilde{x} - b$ и невязку $\|r\|$. Оценим относительную погрешность приближенного решения через невязку. Рассмотрим вектор погрешности

$$\tilde{x} - x = \tilde{x} - A^{-1}b = A^{-1}(A\tilde{x} - b) = A^{-1}r.$$

Учитывая, что $\|b\| \leq \|A\|\|x\|$, получаем требуемый результат

$$\frac{\|\tilde{x} - x\|}{\|x\|} = \frac{\|A^{-1}r\|}{\|x\|} \leq \frac{1}{\|x\|}\|A^{-1}\|\|r\| \leq \frac{\|A\|}{\|b\|}\|A^{-1}\|\|r\|.$$

Таким образом,

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq condA\frac{\|r\|}{\|b\|}.$$

В полученной оценке вновь фигурирует число обусловленности.

Рассмотрим основные свойства $condA$.

1. $condA \geq \|E\| \geq 1$.

   Действительно,

   $$condA = \|A\|\|A^{-1}\| \geq \|AA^{-1}\| = \|E\| \geq 1.$$

   Если норма является подчиненной, то $condE = 1$. Если A– ортогональная матрица, то в спектральной норме $condA = \|A\|\|A^T\| = 1$.

2. $cond(\alpha A) = condA$, $\forall \alpha \neq 0$, $condA^{-1} = condA$,
   если $\|A^T\| = \|A\|$, то $condA^T = condA$.

3. $condA \geq \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}$,
   где $\lambda_{\max}(A), \lambda_{\min}(A)$ – наибольшее и наименьшее по модулю собственные числа матрицы A.

Проверим это свойство. На основании связи между спектрами матриц A, A^{-1} заключаем, что $\frac{1}{\lambda_{\min}(A)}$ – максимальное по модулю собственное число матрицы A^{-1}. Таким образом, в терминах спектральных радиусов

$$\rho(A) = |\lambda_{\max}(A)|, \quad \rho(A^{-1}) = \frac{1}{|\lambda_{\min}(A)|}.$$

Согласно известному свойству для любой матричной нормы $\rho(B) \leq \|B\|$, поэтому

$$condA = \|A\|\|A^{-1}\| \geq \rho(A)\rho(A^{-1}) = \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}.$$

Выделим из полученной оценки случай равенства.

Пусть A – симметричная матрица и используется спектральная норма: $\|A\| = \rho(A)$. Тогда

$$condA = \rho(A)\rho(A^{-1}) = \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}.$$

Если дополнительно предположить, что A – положительно определенная матрица, то

$$condA = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Определение 1. *Матрицы (системы) с $condA$, близким к единице называются хорошо обусловленными* (погрешности входных данных переносятся на решение без заметного увеличения).

Определение 2. *Матрицы (системы) с большим числом обусловленности ($\sim 10^3$ и выше) называются плохо обусловленными* (при решении плохо обусловленных систем возможно сильное увеличение ошибки в решении по сравнению с погрешностью входных данных).

Классическим примером плохо обусловленной матрицы является матрица Гильберта

$$H_n = \left\{ \frac{1}{i+j-1}, \quad i,j = \overline{1,n} \right\}.$$

Число обусловленности матрицы H_n существенно возрастает при увеличении n. Например, $cond\ H_8 > 10^{10}$.

Другим примером плохой обусловленности может служить следующая двухдиагональная матрица

$$A_n(a) = \begin{pmatrix} 1 & a & & \\ & 1 & a & \\ & \dots & & \\ & & 1 & a \\ & & & 1 \end{pmatrix}$$

с условием $a > 1$.

Нетрудно проверить, что в норме $\|A\|_\infty$

$$condA_n(a) = (1+a)\frac{a^n - 1}{a-1},$$

т.е. при увеличении параметра a и порядка n число обусловленности быстро растет. К примеру, $condA_{20}(50) > 10^{14}$.

Пусть A – симметричная, положительно определенная матрица. Укажем элементарное преобразование матрицы A, улучшающее её обусловленность.

Рассмотрим α-параметрическое семейство матриц $A_\alpha = A + \alpha E$, $\alpha \geq 0$. Понятно, что $A_\alpha > 0$, и в спектральной норме

$$condA_\alpha = \frac{\lambda_{\max}(A_\alpha)}{\lambda_{\min}(A_\alpha)}.$$

Поскольку $\lambda(A_\alpha) = \lambda(A) + \alpha$, то

$$condA_\alpha = \frac{\lambda_{\max}(A) + \alpha}{\lambda_{\min}(A) + \alpha}.$$

Найдем производную

$$\frac{d}{d\alpha} condA_\alpha = \frac{\lambda_{\min}(A) - \lambda_{\max}(A)}{(\lambda_{\min}(A) + \alpha)^2} < 0.$$

Это значит, что функция $condA_\alpha$ монотонно убывает по α, причем

$$condA_\alpha \to 1, \quad \alpha \to \infty$$

(обусловленность матрицы A_α улучшается с ростом α).

§7. Метод простой итерации

Рассмотрим линейную систему $Ax = b$. Преобразуем её к эквивалентному виду (удобному для итераций)

$$x = Bx + c. \tag{1}$$

Организуем итерационный процесс по правилу

$$x^{k+1} = Bx^k + c, \quad k = 0, 1, \ldots \tag{2}$$

Здесь k – номер итерации, x^k– k-ое приближение к решению, x^0 – начальное приближение.

Отметим, что

1) если $x^{k+1} = x^k$, то x^k – решение системы (1),

2) если имеет место сходимость $x^k \to x^*, \quad k \to \infty$ то x^* – решение системы (1).

Выясним условия сходимости метода (2).

Теорема *(необходимое и достаточное условие сходимости). Метод простой итерации сходится для любого начального приближения $x^0 \neq x^*$ тогда и только тогда, когда спектральный радиус матрицы B меньше 1:*

$$\rho(B) < 1, \quad (|\lambda_i(B)| < 1, \quad i = \overline{1,n}).$$

Доказательство.

Необходимость. Пусть $x^k \to x^*, k \to \infty$ для любого $x^0 \neq x^*$. Тогда $x^* = Bx^* + c$, поэтому

$$x^* - x^k = (Bx^* + c) - (Bx^{k-1} + c) = B(x^* - x^{k-1}) = \ldots = B^k(x^* - x^0).$$

Следовательно, $B^k(x^* - x^0) \to 0$, $k \to \infty$. Поскольку $x^0 \neq x^*$ – произвольный вектор, то $B^k \Rightarrow O$, $k \to \infty$. На основании леммы 5 (§ 1, п.5) заключаем, что $\rho(B) < 1$.

Достаточность. Пусть $\rho(B) < 1$. Последовательно применяя формулу (2), получаем

$$x^{k+1} = Bx^k + c = B(Bx^{k-1} + c) + c = B^2x^{k-1} + (B + E)c = \ldots =$$

$$= B^{k+1}x^0 + (B^k + B^{k-1} + \ldots + B + E)c.$$

С учетом лемм 5,6 (§ 1, п.5) имеет место сходимость

$$B^{k+1} \to O, \quad B^k + B^{k-1} + \ldots + B + E \to (E - B)^{-1}, \quad k \to \infty.$$

Отсюда заключаем, что $x^{k+1} \to (E - B)^{-1}c = x^*$, $k \to \infty$. (последовательность $\{x^k\}$ сходится к единственному решению системы (1)). □

Поскольку $\rho(B) \leq \|B\|$ (лемма 3, § 1, п.4), то справедливо утверждение.

Следствие *(достаточное условие сходимости). Если $\|B\| < 1$, то метод (2) сходится для любого начального приближения.*

Рассмотрим вопрос о скорости сходимости метода. Пусть векторная $\|x\|$ и матричная $\|B\|$ нормы согласованы, причем $\|B\| < 1$. Тогда

$$x^{k+1} - x^* = (Bx^k + c) - (Bx^* + c) = B(x^k - x^*).$$

Отсюда

$$\|x^{k+1} - x^*\| \leq \|B\|\|x^k - x^*\|, \quad k = 0, 1, \ldots$$

Следовательно, $\|x^{k+1} - x^*\| \le \|B\|^{k+1}\|x^0 - x^*\|$. Это значит, что *последовательность* $\{x^k\}$ *сходится к решению* x^* *со скоростью геометрической прогрессии со знаменателем* $\|B\|$.

Получим оценку погрешности метода при условии, что $\|B\| < 1$.

Запишем тождество ($p = 1, 2..$)

$$x^{k+p} - x^k = (x^{k+p} - x^{k+p-1}) + (x^{k+p-1} - x^{k+p-2}) + \ldots + (x^{k+1} - x^k).$$

Согласно итерационной формуле (2)

$$x^{k+m} - x^{k+m-1} = B(x^{k+m-1} - x^{k+m-2}) = \ldots = B^m(x^k - x^{k-1}), \quad m = \overline{1, p}.$$

Следовательно,

$$x^{k+p} - x^k = (B^p + B^{p-1} + \ldots + B)(x^k - x^{k-1}).$$

Перейдем к нормам

$$\|x^{k+p} - x^k\| \le \|B^p + B^{p-1} + \ldots + B\|\|x^k - x^{k-1}\| \le$$

$$\le (\|B\| + \|B\|^2 + \ldots + \|B\|^p)\|x^k - x^{k-1}\| =$$

$$= \frac{\|B\| - \|B\|^{p+1}}{1 - \|B\|}\|x^k - x^{k-1}\|.$$

При $p \to \infty$ имеем $\|B\|^{p+1} \to 0$, $x^{k+p} \to x^*$. В результате предельного перехода приходим к оценке погрешности (матричная и векторные нормы согласованы).

$$\|x^k - x^*\| \le \frac{\|B\|}{1 - \|B\|}\|x^k - x^{k-1}\|, \quad k = 1, 2, \ldots$$

Замечание. Если $\|B\| \le \frac{1}{2}$, то из условия $\|x^k - x^{k-1}\| \le \varepsilon$ следует, что $\|x^k - x^*\| \le \varepsilon$. (если $\|B\| \le \frac{1}{2}$, то ε-близость соседних приближений означает получение ε-приближенного решения).

Поскольку $\|x^k - x^{k-1}\| \le \|B\|^{k-1}\|x^1 - x^0\|$, то полученную оценку можно "ослабить"следующим образом

$$\|x^k - x^*\| \le \frac{\|B\|^k}{1 - \|B\|}\|x^1 - x^0\|.$$

В качестве начального приближения, как правило, выбирают $x^0 = c$. При этом $x^1 = Bc+c$, т.е. $x^1-x^0 = Bc$, $\|x^1-x^0\| \le \|B\|\|c\|$. Оценка погрешности для этого случая имеет вид

$$\|x^k - x^*\| \le \frac{\|B\|^{k+1}}{1-\|B\|}\|c\|.$$

Здесь в правой части фигурируют только входные данные системы (1).

В заключение опишем общий прием приведения системы $Ax = b$ к виду $x = Bx + c$, удобному для итераций. Представим исходную систему в виде

$$x = x + D(b - Ax),$$

где D – невырожденная матрица. В результате получаем систему (1), в которой $B = E - DA$, $c = Db$. Проблема состоит в приемлемом выборе матрицы преобразования D.

Пусть диагональные элементы матрицы A отличны от нуля. Положим

$$D = diag(\frac{1}{a_{11}}, \dots, \frac{1}{a_{nn}}).$$

В этом случае матрица B имеет вид

$$B = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots. & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots. & -\frac{a_{2n}}{a_{22}} \\ & \dots & \dots & \dots \\ -\frac{a_{n1}}{a_{nn}} & \dots & -\frac{a_{n,n-1}}{a_{nn}} & 0 \end{pmatrix}.$$

Такое преобразование системы называют преобразованием Якоби к виду, удобному для итераций. Соответствующий вариант метода простой итерации называют методом Якоби. Если A – матрица со строгим диагональным преобладанием, то $\|B\|_\infty < 1$, что обеспечивает сходимость метода Якоби.

Рассмотрим другой способ выбора матрицы D. Пусть система $Ax = b$ является нормальной, т.е. A – симметричная, положительно определенная матрица. Это значит, что $\lambda_i(A) > 0$, $i = \overline{1,n}$.

Положим $D = \alpha E$, где $\alpha \ne 0$ – параметр. Выберем α таким образом, чтобы обеспечить условие сходимости

$$|\lambda_i(B)| = |\lambda_i(E - \alpha A)| < 1, \quad i = \overline{1,n}.$$

Согласно известному свойству собственных чисел

$$\lambda_i(E-\alpha A)=\lambda_i(E)-\alpha\lambda_i(A)=1-\alpha\lambda_i(A).$$

Решая неравенство $|1-\alpha\lambda_i(A)|<1$, получаем $0<\alpha<\frac{2}{\lambda_i(A)}$. Поскольку $\lambda_i(A)\le\|A\|$, то

$$\frac{2}{\|A\|}\le\frac{2}{\lambda_i(A)},\quad i=\overline{1,n}.$$

Следовательно, при $\alpha\in(0,\frac{2}{\|A\|})$ получаем $|\lambda_i(B)|<1,\quad i=\overline{1,n}$, что гарантирует сходимость метода простой итерации для системы $x=x+\alpha(b-Ax)$.

§8. Метод Зейделя

Рассмотрим линейную систему в итерационной форме

$$x=Bx+c\ . \tag{1}$$

Как известно, метод простой итерации в координатной записи имеет вид

$$x_i^{k+1}=\sum_{j=1}^{n}b_{ij}x_j^k+c_i,\quad i=\overline{1,n},\quad k=0,1,\dots$$

С целью ускорения сходимости проведем следующую модификацию данного метода

$$x_i^{k+1}=\sum_{j=1}^{i-1}b_{ij}x_j^{k+1}+\sum_{j=i}^{n}b_{ij}x_j^k+c_i,\quad i=\overline{1,n}. \tag{2}$$

Таким образом, при подсчете x_i^{k+1}, $i=\overline{2,n}$ используются уже известные координаты нового приближения x_j^{k+1}, $j=\overline{1,i-1}$.

Итерационная формула (2) определяет метод Зейделя для линейной системы (1).

Укажем векторно-матричную интерпретацию процедуры (2). Представим матрицу B в виде суммы двух треугольных матриц: $B=B_1+B_2$, где

$$B_1=\begin{pmatrix}0&&&\\b_{21}&0&&\\b_{31}&b_{32}&0&\\&\dots&\dots&\\b_{n1}&\dots&b_{n,n-1}&0\end{pmatrix},\quad B_2=\begin{pmatrix}b_{11}&b_{12}&\dots&\dots&b_{1n}\\&b_{22}&\dots&\dots&b_{2n}\\&&b_{33}&\dots&b_{3n}\\&&\dots&\dots&\\&&&&b_{nn}\end{pmatrix}.$$

Тогда формула (2) записывается следующим образом

$$x^{k+1} = B_1 x^{k+1} + B_2 x^k + c, \quad k = 0, 1, \ldots$$

или

$$(E - B_1)x^{k+1} = B_2 x^k + c, \quad k = 0, 1, \ldots$$

Поскольку матрица $E - B_1$ не вырождена ($\det(E - B_1) = 1$), то процедура представляется в виде

$$x^{k+1} = (E - B_1)^{-1} B_2 x^k + (E - B_1)^{-1} c, \quad k = 0, 1, \ldots$$

Итак, метод Зейделя можно рассматривать как метод простой итерации применительно к системе

$$x = (E - B_1)^{-1} B_2 x + (E - B_1)^{-1} c,$$

которая эквивалентна исходной.

Отсюда получаем достаточное условие сходимости метода Зейделя:

$$\|(E - B_1)^{-1} B_2\| < 1.$$

Докажем явный критерий сходимости, выраженный через элементы матрицы B.

Теорема *(достаточное условие сходимости). Если* $\|B\|_\infty < 1$, *то метод Зейделя (2) сходится при любом начальном приближении.*

Доказательство. Пусть x^* – решение системы (1), т.е.

$$x_i^* = \sum_{j=1}^{n} b_{ij} x_j^* + c, \quad i = \overline{1, n}, \quad k = 0, 1, \ldots.$$

Тогда с учетом (2) получаем

$$x_i^* - x_i^{k+1} = \sum_{j=1}^{i-1} b_{ij}(x_j^* - x_j^{k+1}) + \sum_{j=i}^{n} b_{ij}(x_j^* - x_j^k).$$

Отсюда

$$|x_i^* - x_i^{k+1}| \le \sum_{j=1}^{i-1} |b_{ij}||x_j^* - x_j^{k+1}| + \sum_{j=i}^{n} |b_{ij}||x_j^* - x_j^k|.$$

Поскольку $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$, то

$$|x_i^* - x_i^{k+1}| \le \alpha_i \|x^* - x^{k+1}\|_\infty + \beta_i \|x^* - x^k\|_\infty, \quad i = \overline{1, n}, \tag{3}$$

где

$$\alpha_i = \sum_{j=1}^{i-1} |b_{ij}|, \quad \beta_i = \sum_{j=i}^{n} |b_{ij}|.$$

Пусть $s \in \{1, \ldots, n\}$ – максимизирующий индекс:

$$|x_s^* - x_s^{k+1}| = \max_{1 \le i \le n} |x_i^* - x_i^{k+1}| = \|x^* - x^{k+1}\|_\infty.$$

Тогда неравенство (3) при $i = s$ принимает вид

$$\|x^* - x^{k+1}\|_\infty \le \alpha_s \|x^* - x^{k+1}\|_\infty + \beta_s \|x^* - x^k\|_\infty.$$

Поскольку $\alpha_s + \beta_s \le \|B\|_\infty < 1$, то $1 - \alpha_s > 0$, т.е.

$$\|x^* - x^{k+1}\|_\infty \le \frac{\beta_s}{1 - \alpha_s} \|x^* - x^k\|_\infty \le \mu \|x^* - x^k\|_\infty,$$

$$\mu = \max_{1 \le s \le n} \frac{\beta_s}{1 - \alpha_s}.$$

Покажем, что $\mu \le \|B\|_\infty$. Для любого $s = \overline{1, n}$

$$\alpha_s + \beta_s - \frac{\beta_s}{1 - \alpha_s} = \frac{\alpha_s(1 - \alpha_s - \beta_s)}{1 - \alpha_s} \ge 0.$$

Следовательно,

$$\mu = \max_{1 \le s \le n} \frac{\beta_s}{1 - \alpha_s} = \frac{\beta_{s_1}}{1 - \alpha_{s_1}} \le \alpha_{s_1} + \beta_{s_1} \le \|B\|_\infty.$$

Таким образом,

$$\|x^* - x^{k+1}\|_\infty \le \mu \|x^* - x^k\|_\infty, \;\; \mu \le \|B\|_\infty < 1.$$

Полученное означает, что $x^k \to x^*, k \to \infty$. □

Как следствие отметим, что метод сходится со скоростью геометрической прогрессии со знаменателем μ. Поскольку $\mu \le \|B\|_\infty$, то метод Зейделя по быстроте сходимости "не хуже"метода простой итерации.

Рассмотрим вопрос об оценке погрешности метода в условиях теоремы. Применим преобразования теоремы к векторам x^k, x^{k+1} (вместо x^*, x^{k+1}). Тогда в полной аналогии с предыдущим получим неравенство

$$\|x^{k+1} - x^k\|_\infty \le \mu \|x^k - x^{k-1}\|_\infty, \;\; k = 1, 2, \ldots$$

Следовательно, для $p = 1, 2, \ldots$

$$\|x^{k+p} - x^{k+p-1}\|_\infty \le \mu\|x^{k+p-1} - x^{k+p-2}\|_\infty \le \ldots \le \mu^p\|x^k - x^{k-1}\|_\infty.$$

Тогда

$$\|x^{k+p} - x^k\|_\infty \le \|x^{k+p} - x^{k+p-1}\|_\infty + \|x^{k+p-1} - x^{k+p-2}\|_\infty + \ldots +$$
$$+\|x^{k+1} - x^k\|_\infty \le (\mu^p + \mu^{p-1} + \ldots + \mu)\|x^k - x^{k-1}\|_\infty =$$
$$= \frac{\mu - \mu^{p+1}}{1 - \mu}\|x^k - x^{k-1}\|_\infty.$$

При $p \to \infty$ получаем требуемую оценку

$$\|x^k - x^*\|_\infty \le \frac{\mu}{1 - \mu}\|x^k - x^{k-1}\|_\infty,$$

которая по структуре вполне соответствует аналогичной оценке для метода простой итерации.

§9. Градиентные методы решения линейных систем

1. Редукция системы к экстремальной задаче

Будем рассматривать нормальную линейную систему

$$Ax = b, \tag{1}$$

$$A^T = A, \quad A > 0. \tag{2}$$

По данной системе образуем квадратичную функцию

$$\varphi(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle \tag{3}$$

и поставим задачу на минимум

$$\varphi(x) \to \min, \quad x \in R^n. \tag{4}$$

Это задача отыскания точки $\overline{x} \in R^n$ с условием

$$\varphi(\overline{x}) \le \varphi(x), \quad x \in R^n.$$

В этом случае $\overline{x}$ – точка минимума функции $\varphi(x)$, решение задачи (4), $\varphi(\overline{x})$ – минимальное значение функции $\varphi(x)$, значение задачи (4).

Установим связь между задачами (1),(2) и (3),(4). Пусть x^* – решение системы (1). Представим функцию $\varphi(x)$ в виде

$$\varphi(x) = \frac{1}{2}\langle x - x^*, A(x - x^*)\rangle - \frac{1}{2}\langle x^*, Ax^*\rangle.$$

Поскольку $A > 0$, то x^* является единственным решением задачи (4). Таким образом, задачи (1),(2) и (3),(4) эквивалентны.

Для решения задачи (4) применяют, как правило, методы градиентного типа. Напомним соответствующие понятия.

Рассмотрим приращение функции φ в точке x

$$\varphi(x + \Delta x) - \varphi(x) = \langle Ax - b, \Delta x\rangle + \frac{1}{2}\langle \Delta x, A\Delta x\rangle. \tag{5}$$

Вектор $\nabla\varphi(x) = Ax - b$ называется градиентом функции φ в точке x. Отметим, что это вектор невязок для системы (1). Вектор $-\nabla\varphi(x) = b - Ax$ есть антиградиент функции φ в точке x.

Согласно предыдущему имеет место *утверждение:*

x^* - решение задачи (3),(4) $\Leftrightarrow \nabla\varphi(x^*) = 0 \Leftrightarrow Ax^* = b$.

Пусть $p \in R^n$ – ненулевой вектор, $\alpha > 0$ – числовой параметр. Положим в (5) $\Delta x = \alpha p$

$$\varphi(x + \alpha p) - \varphi(x) = \langle \nabla\varphi(x), p\rangle\ \alpha + \frac{1}{2}\langle p, Ap\rangle\ \alpha^2. \tag{5'}$$

Величина $\varphi_p(x) = \langle \nabla\varphi(x), p\rangle$ называется производной функции φ в точке x по направлению p.

Вектор $p \neq 0$ называется направлением спуска функции φ в точке x, если $\varphi(x + \alpha p) < \varphi(x)$ для достаточно малых $\alpha > 0$.

Согласно $(5')$ имеет место *утверждение:* если $\langle \nabla\varphi(x), p\rangle < 0$, то вектор p есть направление спуска функции φ в точке x.

Пусть $\nabla\varphi(x) \neq 0$. Сформулируем задачу

$$\langle \nabla\varphi(x), p\rangle \to \min, \quad \langle p, p\rangle = 1.$$

Её решение, вектор p^*, называют направлением скорейшего спуска функции φ в точке x. Нетрудно проверить, что (норма евклидова)

$$p^* = -\frac{\nabla\varphi(x)}{\|\nabla\varphi(x)\|},$$

т.е. нормированный антиградиент есть направление скорейшего спуска с производной

$$\langle \nabla\varphi(x), p^* \rangle = -\|\nabla\varphi(x)\|.$$

2. Градиентный метод с постоянным шагом

Рассмотрим задачу (1),(2) в интерпретации (3),(4). Для решения экстремальной задачи (4) организуем следующую итерационную процедуру

$$x^{k+1} = x^k - \alpha\nabla\varphi(x^k), \quad k = 0, 1, \ldots \tag{6}$$

Здесь x^k - k-ое приближение к решению x^*, $\alpha > 0$ - шаг вдоль направления антиградиента. Если $\nabla\varphi(x^k) \neq 0$ ($x^k \neq x^*$), то для малых $\alpha > 0$ выполняется свойство уменьшения: $\varphi(x^{k+1}) < \varphi(x^k)$.

Процедуру (6) для решения задачи (4) назовем градиентным методом с постоянным шагом $\alpha > 0$. Возможна другая интерпретация итерационной формулы (6). Учитывая выражение для градиента $\nabla\varphi$ перепишем (6) в виде

$$x^{k+1} = x^k + \alpha(b - Ax^k), \quad k = 0, 1, \ldots$$

Это метод простой итерации для системы $x = x + \alpha(b - Ax)$ с параметром α.

Рассмотрим вопрос о выборе шага α в методе (6).

Пусть нам известны границы спектра матрицы A, т.е. величины

$$m = \min_{1\leq i\leq n} \lambda_i(A), \quad M = \max_{1\leq i\leq n} \lambda_i(A).$$

Понятно, что $0 < m \leq M$, причем M – спектральный радиус матрицы A. Запишем неравенства для квадратичной формы

$$m \leq \langle x, Ax \rangle \leq M, \quad \langle x, x \rangle = 1. \tag{7}$$

Для последующих оценок будем использовать евклидову векторную норму: $\|x\|^2 = \langle x, x \rangle$ и спектральную норму для симметричной матрицы B :

$$\|B\| = \rho(B) = \max_{1\leq i\leq n} |\lambda_i(B)|.$$

Как известно, это нормы согласованы.

Получим другое выражение для $\|B\|$. Используя неравенство Коши-Шварца для скалярного произведения, имеем

$$|\langle x, Bx\rangle| \le \|x\|\|Bx\| \le \|B\|\|x\|^2.$$

В частности, если $\|x\| = 1$, то $|\langle x, Bx\rangle| \le \|B\|$.

Пусть $\overline{\lambda}$ – максимальное по модулю собственное значение матрицы B, $\overline{x}$ ($\|\overline{x}\| = 1$) – соответствующий собственный вектор. Тогда $\|B\| = |\overline{\lambda}|$, $|\langle \overline{x}, B\overline{x}\rangle| = |\overline{\lambda}|\langle \overline{x}, \overline{x}\rangle = \|B\|$. С учетом предыдущего неравенства заключаем, что

$$\|B\| = \max_{\|x\|=1} |\langle x, Bx\rangle|.$$

Перейдем к анализу итерационного процесса (6).

Теорема 1. *Для градиентного метода (6) имеет место оценка*

$$\|x^{k+1} - x^*\| \le q(\alpha)\|x^k - x^*\|, \quad k = 0, 1, \ldots, \tag{8}$$

где $q(\alpha) = \max\{|1 - \alpha m|, |1 - \alpha M|\}$.

Доказательство. Поскольку $Ax^* = b$, то

$$\nabla\varphi(x^k) = Ax^k - b = A(x^k - x^*).$$

На основании формулы (6)

$$x^{k+1} - x^* = x^k - x^* - \alpha A(x^k - x^*) = (E - \alpha A)(x^k - x^*).$$

Отметим, что $(E - \alpha A)$ – симметричная матрица. Перейдем к нормам

$$\|x^{k+1} - x^*\| \le \|E - \alpha A\|\|x^k - x^*\|.$$

Используя выражение для спектральной нормы и оценку (7), получаем

$$\|E - \alpha A\| = \max_{\|x\|=1} |\langle x, x\rangle - \alpha\langle x, Ax\rangle| = \max_{m \le z \le M} |1 - \alpha z| =$$

$$= \max\{|1 - \alpha m|, |1 - \alpha M|\} = q(\alpha).$$

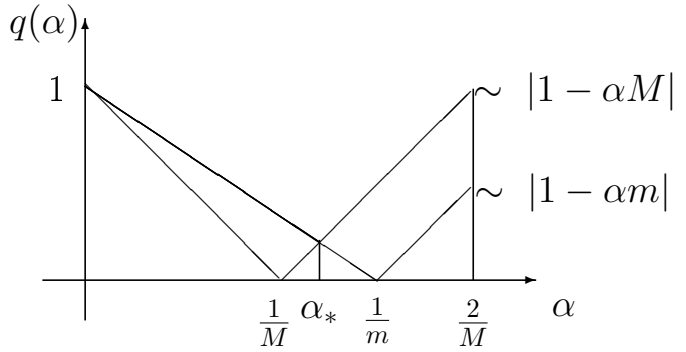
Следовательно, $\|x^{k+1} - x^*\| \le q(\alpha)\|x^k - x^*\|$. □

Приступим к выбору шага α с целью оптимизации метода (6) в рамках оценки (8). В первую очередь, обеспечим условие сходимости $x^k \to x^*$, $k \to \infty$, т.е. потребуем, чтобы $q(\alpha) < 1$. Решим это неравенство относительно α : $q(\alpha) < 1 \Leftrightarrow |1 - \alpha m| < 1$, $|1 - \alpha M| < 1$. Из первого неравенства $0 < \alpha < \frac{2}{m}$, из второго $0 < \alpha < \frac{2}{M}$. Следовательно, при $\alpha \in (0, \frac{2}{M})$ имеем $q(\alpha) < 1 \Rightarrow$

$x^k \to x^*, k \to \infty$ со скоростью геометрической прогрессии. Величина $q(\alpha)$ (знаменатель прогрессии) характеризует скорость сходимости: чем меньше $q(\alpha)$, тем быстрее сходимость. Поэтому естественно поставить задачу поиска оптимального шага α_* в виде

$$q(\alpha) \to \min, \quad \alpha \in (0, \frac{2}{M}).$$

Для решения этой задачи воспользуемся геометрической интерпретацией



В результате

$$\alpha_* = \frac{2}{M+m}, \quad q_* = q(\alpha_*) = \frac{M-m}{M+m}.$$

Таким образом, выбор шага $\alpha = \alpha_*$ в методе (6) является оптимальным с точки зрения скорости сходимости последовательности $\{x^k\}$ к решению x^*.

Полученный результат позволяет охарактеризовать эффективность градиентного метода в зависимости от числа обусловленности матрицы A.

Как известно, в спектральной норме $condA = \frac{M}{m}$. Тогда $q_* = \frac{condA-1}{condA+1}$.

Пусть матрица A хорошо обусловлена, т.е. $condA$ близко к единице. Это означает, что q_* близко к нулю, что гарантирует быструю сходимость $x^k \to x^*, k \to \infty$. (метод (6) при $\alpha = \alpha_*$ эффективен).

Пусть A – плохо обусловленная матрица, т.е. $condA$ много больше единицы. Тогда величина q_* близка к единице. Это плохая ситуация для градиентного метода, когда его эффективность невелика: сходимость $x^k \to x^*$, $k \to \infty$ может быть очень медленной.

Замечание 1. Поскольку $\|E - \alpha A\| = q(\alpha)$, то на основании известного результата для метода простой итерации получаем оценку погрешности

градиентного метода (6)

$$\|x^k - x^*\| \le \frac{q(\alpha)}{1-q(\alpha)}\|x^k - x^{k-1}\|, \quad k = 1, 2, \ldots, \quad \alpha \in (0, \frac{2}{M}).$$

3. Метод скорейшего спуска

Продолжим исследование задачи (3), (4) с позиций градиентного спуска. Рассмотрим следующую модификацию градиентного метода (6), когда шаг α зависит от номера итерации k

$$x^{k+1} = x^k - \alpha_k \nabla\varphi(x^k), \quad \alpha_k > 0, \quad \nabla\varphi(x^k) \neq 0, \quad k = 0, 1, \ldots \tag{9}$$

Учитывая смысл задачи (4), будем выбирать шаг α_k из условия минимума функции φ вдоль направления антиградиента $(-\nabla\varphi(x^k))$. Иными словами, в качестве α_k возьмем решение следующей задачи

$$\alpha_k : \quad \varphi(x^k - \alpha\nabla\varphi(x^k)) \to \min, \quad \alpha > 0. \tag{10}$$

Итерационную процедуру (9), (10) называют методом скорейшего спуска для решения задачи (4).

Найдем явное выражение для α_k, решая задачу (10). Распишем минимизируемую функцию

$$g_k(\alpha) = \varphi(x^k - \alpha\nabla\varphi(x^k)) = \frac{1}{2}\alpha^2\langle\nabla\varphi(x^k), A\nabla\varphi(x^k)\rangle -$$

$$-\alpha\langle\nabla\varphi(x^k), \nabla\varphi(x^k)\rangle + \varphi(x^k).$$

Итак, $g_k(\alpha)$ – выпуклая парабола, поэтому её точка минимума находится из условия стационарности $\frac{dg_k(\alpha)}{d\alpha} = 0$. Корень этого уравнения положителен, т.е. он совпадает с решением задачи (10)

$$\alpha_k = \frac{\langle\nabla\varphi(x^k), \nabla\varphi(x^k)\rangle}{\langle\nabla\varphi(x^k), A\nabla\varphi(x^k)\rangle}. \tag{11}$$

Метод скорейшего спуска полностью определен.

Установим одно свойство метода по части градиентов. Согласно определению и с учетом формулы (9)

$$\nabla\varphi(x^{k+1}) = Ax^{k+1} - b = Ax^k - b - \alpha_k A\nabla\varphi(x^k) =$$

$$= \nabla\varphi(x^k) - \alpha_k A\nabla\varphi(x^k).$$

Отсюда

$$\langle\nabla\varphi(x^k), \nabla\varphi(x^{k+1})\rangle = \langle\nabla\varphi(x^k), \nabla\varphi(x^k)\rangle -$$
$$-\alpha_k\langle\nabla\varphi(x^k), A\nabla\varphi(x^k)\rangle = 0$$

в силу выбора α_k.

Таким образом, $\langle\nabla\varphi(x^k), \nabla\varphi(x^{k+1})\rangle = 0, \quad k = 0, 1, \ldots$ – градиенты соседних приближений ортогональны.

Изучим вопрос о сходимости метода.

Теорема 2. *Для метода (9),(11) в задаче (4) справедлива следующая оценка сходимости*

$$\|x^{k+1} - x^*\| \le q_*^{k+1}\sqrt{\frac{2}{m}(\varphi(x^0) - \varphi(x^*))}, \quad q_* = \frac{M-m}{M+m}.$$

Доказательство. В соответствии с методом

$$\varphi(x^{k+1}) = g_k(\alpha_k) = \varphi(x^k) - \frac{1}{2}\frac{\langle\nabla\varphi(x^k), \nabla\varphi(x^k)\rangle^2}{\langle\nabla\varphi(x^k), A\nabla\varphi(x^k)\rangle}. \tag{12}$$

Согласно формуле приращения

$$\varphi(x^k) - \varphi(x^*) = \langle\nabla\varphi(x^*), x^k - x^*\rangle + \frac{1}{2}\langle x^k - x^*, A(x^k - x^*)\rangle,$$

причем $\nabla\varphi(x^*) = 0$. Кроме того, $\nabla\varphi(x^k) = A(x^k - x^*)$, т.е. $x^k - x^* = A^{-1}\nabla\varphi(x^k)$.

Следовательно,

$$\varphi(x^k) - \varphi(x^*) = \frac{1}{2}\langle\nabla\varphi(x^k), A^{-1}\nabla\varphi(x^k)\rangle.$$

Тогда из (12) получаем (предварительно вычтем из обеих частей $\varphi(x^*)$)

$$\frac{\varphi(x^{k+1}) - \varphi(x^*)}{\varphi(x^k) - \varphi(x^*)} = 1 - \frac{\langle\nabla\varphi(x^k), \nabla\varphi(x^k)\rangle^2}{\langle\nabla\varphi(x^k), A\nabla\varphi(x^k)\rangle\langle\nabla\varphi(x^k), A^{-1}\nabla\varphi(x^k)\rangle}.$$

Далее, используем неравенство Канторовича [2, с.117]

$$\langle x, Ax\rangle\langle x, A^{-1}x\rangle \le \frac{(M+m)^2}{4mM}\langle x, x\rangle^2.$$

В результате

$$\frac{\varphi(x^{k+1})-\varphi(x^*)}{\varphi(x^k)-\varphi(x^*)} \le 1-\frac{4mM}{(M+m)^2} = \frac{(M-m)^2}{(M+m)^2} = q_*^2.$$

Таким образом, получили оценку уменьшения

$$\varphi(x^{k+1})-\varphi(x^*) \le q_*^2\ (\varphi(x^k)-\varphi(x^*)),\ \ k=0,1,\dots$$

Отсюда

$$\varphi(x^{k+1})-\varphi(x^*) \le q_*^{2(k+1)}(\varphi(x^0)-\varphi(x^*)).$$

Заметим, что на основании оценки (7)

$$\varphi(x^{k+1})-\varphi(x^*) = \frac{1}{2}\langle x^{k+1}-x^*, A(x^{k+1}-x^*)\rangle \ge$$

$$\ge \frac{1}{2}m\|x^{k+1}-x^*\|^2.$$

Следовательно,

$$\|x^{k+1}-x^*\|^2 \le \frac{2}{m}(\varphi(x^{k+1})-\varphi(x^*)) \le \frac{2}{m}q_*^{2(k+1)}(\varphi(x^0)-\varphi(x^*)),$$

что равносильно утверждению теоремы.

Замечание 2. Метод скорейшего спуска (9), (11) не требует знания величин m, M, но обеспечивает сходимость $x^k \to x^*, k \to \infty$ со скоростью геометрической прогрессии с минимальным знаменателем q_*.

4. Метод минимальных невязок

Рассмотрим линейную систему $Ax = b$ с условием $A^T = A, A > 0$. Введем вектор невязок $r(x) = Ax - b$. При этом $\|r(x)\|$ – невязка системы (норма евклидова). Отметим, что $r(x) = \nabla\varphi(x)$, $\varphi(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle$.

Построим итерационный метод следующим образом. По известному приближению x^k $(r(x^k) \ne 0)$ введем семейство точек $x^k(\alpha) = x^k - \alpha r(x^k)$ с параметром $\alpha > 0$. Сформулируем вспомогательную задачу на минимум невязки

$$\|r(x^k(\alpha)\| \to \min, \quad \alpha > 0.$$

Пусть α_k – решение этой задачи. Следующее приближение имеет вид $x^{k+1} = x^k(\alpha_k)$.

Найдем явное выражение для α_k, решая вспомогательную задачу. Согласно определению

$$r(x^k(\alpha)) = Ax^k(\alpha) - b = r(x^k) - \alpha Ar(x^k).$$

Тогда

$$\|r(x^k(\alpha))\|^2 = \langle r(x^k(\alpha)), r(x^k(\alpha))\rangle = \langle r(x^k), r(x^k)\rangle - 2\alpha\langle r(x^k), Ar(x^k)\rangle +$$
$$+\alpha^2\langle Ar(x^k), Ar(x^k)\rangle.$$

Таким образом, решение вспомогательной задачи имеет вид (точка минимума выпуклой параболы)

$$\alpha_k = \frac{\langle r(x^k), Ar(x^k)\rangle}{\|Ar(x^k)\|^2}.$$

Отметим, что $\alpha_k > 0$.

Итерационная формула метода минимальных невязок:

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

Обсудим вопрос о сходимости метода. Согласно выбору α_k выполняется неравенство

$$\|r(x^{k+1})\| \le \|r(x^k(\alpha))\|, \quad \alpha > 0.$$

При этом $r(x^k(\alpha)) = Ax^k(\alpha) - b = (E - \alpha A)r(x^k)$. Отсюда, используя спектральную норму матрицы, получаем

$$\|r(x^k(\alpha))\| \le \|E - \alpha A\|\|r(x^k)\| = q(\alpha)\|r(x^k)\|,$$

$$q(\alpha) = \max\{|1 - \alpha m|, |1 - \alpha M|\}.$$

Положим $\alpha = \alpha_* = \frac{2}{M+m}$. Тогда $q(\alpha_*) = q_* = \frac{M-m}{M+m}$. В результате получаем оценку сходимости метода

$$\|r(x^{k+1})\| \le q_*\|r(x^k)\| \le q_*^{k+1}\|r(x^0)\|.$$

Это значит, что имеет место сходимость по невязке: $\|r(x^k)\| \to 0, k \to \infty$ со скоростью геометрической прогрессии со знаменателем q_*.

Получим оценку сходимости по погрешности $\|x^{k+1} - x^*\|$, где $x^* = A^{-1}b$ – решение системы. Отметим, что $\|A^{-1}\| = \frac{1}{m}$.

Поскольку

$$x^{k+1} - x^* = x^{k+1} - A^{-1}b = A^{-1}(Ax^{k+1} - b) = A^{-1}r(x^{k+1}),$$

то

$$\|x^{k+1} - x^*\| \le \|A^{-1}\|\|r(x^{k+1})\| \le q_*^{k+1}\frac{\|r(x^0)\|}{m}, \quad k = 0, 1, \ldots$$

Следовательно, имеет место сходимость $x^k \to x^*, k \to \infty$ со скоростью геометрической прогрессии со знаменателем q_*.

5. Метод сопряженных градиентов

Пусть $(n \times n)$ – матрица A симметрична и положительно определена. Введем понятие сопряженных направлений относительно A.

Определение 1. *Векторы $x, y \in R^n$ называются A - сопряженными (A - ортогональными), если $\langle x, Ay\rangle = 0$.*

Определение 2. *Система векторов $p^i \in R^n$, $i = \overline{0, r}$ называется A - сопряженной, если каждая пара $p^i, p^j, i \ne j$, A – сопряжена.*

Укажем связь A - сопряженности с линейной независимостью.

Лемма. *Если ненулевые векторы $p^0, p^1, \ldots, p^r$, $r+1 \le n$ A - сопряжены, то они линейно независимы.*

Доказательство легко проводится от противного.

Выясним роль сопряженных направлений для решения задачи (4).

Возьмем за основу следующую итерационную процедуру

$$x^{k+1} = x^k + \alpha_k p^k, \quad k = 0, 1, \ldots, \tag{13}$$

где p^k – некоторое направление спуска функции φ в точке x^k. Это значит, что

$$\langle p^k, \nabla\varphi(x^k)\rangle < 0. \tag{14}$$

Шаг $\alpha_k > 0$ будем выбирать по правилу скорейшего спуска

$$\alpha_k : \ \varphi(x^k + \alpha p^k) \to \min, \ \ \alpha > 0. \tag{15}$$

Решая эту задачу, получаем следующее выражение

$$\alpha_k = -\frac{\langle p^k, \nabla\varphi(x^k)\rangle}{\langle p^k, Ap^k\rangle}.$$

Отметим одно свойство процедуры (13)-(15):

$$\langle p^k, \nabla\varphi(x^{k+1})\rangle = 0, \ \ k = 0, 1, \ldots \tag{16}$$

Действительно, распишем левую часть этого равенства

$$\langle p^k, Ax^{k+1} - b\rangle = \langle p^k, Ax^k - b\rangle + \alpha_k \langle p^k, Ap^k\rangle = 0$$

в силу выбора α_k.

Докажем основное свойство процедуры (13)-(15) применительно к задаче (4).

Теорема. *Если направления $p^k, k = \overline{0, n-1}$ A - сопряжены, то метод (13)-(15) решает задачу (4) не более, чем за n итераций.*

Утверждение теоремы означает, что в процессе итераций найдется такой номер $r \le n$, что $\nabla\varphi(x^r) = 0$, т.е. $x^r = x^*$.

Перейдем к *доказательству* теоремы. Если для некоторого $k < n$ окажется, что $\nabla\varphi(x^k) = 0$, то теорема доказана. Рассмотрим общий случай, когда $\nabla\varphi(x^k) \neq 0$, $k = \overline{0, n-1}$. Покажем, что $\nabla\varphi(x^n) = 0$.

Найдем специальное представление для градиента $\nabla\varphi(x^n)$. Применяя последовательно формулу (13), имеем

$$x^n = x^{n-1} + \alpha_{n-1}p^{n-1} = x^{n-2} + \alpha_{n-2}p^{n-2} + \alpha_{n-1}p^{n-1} = \ldots. =$$

$$= x^{j+1} + \alpha_{j+1}p^{j+1} + \alpha_{j+2}p^{j+2} + \ldots + \alpha_{n-1}p^{n-1} =$$

$$= x^{j+1} + \sum_{i=j+1}^{n-1} \alpha_i p^i, \quad j = \overline{0, n-2}.$$

Отсюда

$$\nabla\varphi(x^n) = Ax^n - b = \nabla\varphi(x^{j+1}) + \sum_{i=j+1}^{n-1} \alpha_i A p^i.$$

Учитывая условие ортогональности (16) и свойство A – сопряженности, получаем

$$\langle p^j, \nabla\varphi(x^n)\rangle =$$

$$= \langle p^j, \nabla\varphi(x^{j+1})\rangle + \sum_{i=j+1}^{n-1} \alpha_i \langle p^j, Ap^i\rangle = 0, \quad j = \overline{0, n-2}.$$

Согласно тому же условию (16) $\langle p^{n-1}, \nabla\varphi(x^n)\rangle = 0$. Таким образом, $\langle p^j, \nabla\varphi(x^n)\rangle = 0$, $j = \overline{0, n-1}$. Поскольку векторы p^j – линейно независимы, то должно быть $\nabla\varphi(x^n) = 0$. Это значит, что $x^n = x^*$. □

Итак, в рамках схемы (13)-(15) решение задачи (4) сводится к построению A - сопряженных векторов $p^0, p^1, \ldots, p^{n-1}$. В методе сопряженных градиентов такие векторы строятся следующим образом.

Пусть x^0 – произвольное начальное приближение. Положим $p^0 = -\nabla\varphi(x^0)$. Пусть направления $p^0, p^1, \ldots, p^{k-1}$ и соответствующие точки $x^1, \ldots, x^k$ построены. Если $\nabla\varphi(x^k) = 0$, то $x^k = x^*$. В противном случае полагаем

$$p^k = -\nabla\varphi(x^k) + \beta_k p^{k-1}. \tag{17}$$

Итерационный параметр β_k подберем из условия сопряжения $\langle p^{k-1}, Ap^k\rangle = 0$. Это приводит к следующему выражению

$$\beta_k = \frac{\langle p^{k-1}, A\nabla\varphi(x^k)\rangle}{\langle p^{k-1}, Ap^{k-1}\rangle}. \tag{18}$$

Нетрудно проверить, что p^k – направление спуска функции φ в точке x^k :

$$\langle p^k, \nabla\varphi(x^k)\rangle = \langle -\nabla\varphi(x^k) + \beta_k p^{k-1}, \nabla\varphi(x^k)\rangle =$$

$$= -\|\nabla\varphi(x^k)\|^2 < 0.$$

Отметим также, что из формулы (13)

$$p^k = \frac{1}{\alpha_k}(x^{k+1} - x^k),$$

следовательно,

$$Ap^k = \frac{1}{\alpha_k}(Ax^{k+1} - Ax^k) = \frac{1}{\alpha_k}(\nabla\varphi(x^{k+1}) - \nabla\varphi(x^k)). \tag{19}$$

Теорема. *Для последовательностей $x^k, p^k, k = 0, 1, \ldots$ имеют место соотношения*

$$\langle \nabla\varphi(x^i), p^j\rangle = 0, \quad \langle \nabla\varphi(x^i), \nabla\varphi(x^j)\rangle = 0, \quad \langle p^i, Ap^j\rangle = 0, \tag{20}$$

$$i = 1, 2 \ldots, \quad j = \overline{0, i-1}.$$

Проведем *доказательство* с помощью индукции по индексу i. При $i = 1$ утверждение теоремы справедливо, ибо

$$\langle \nabla\varphi(x^1), p^0\rangle = 0,$$

$$\langle \nabla\varphi(x^1), \nabla\varphi(x^0)\rangle = -\langle \nabla\varphi(x^1), p^0\rangle = 0,$$

$$\langle p^1, Ap^0\rangle = 0.$$

Допустим, что соотношения (20) справедливы для $i = r$, $j = \overline{0, r-1}$. Рассмотрим случай $i = r+1$, $j = \overline{0, r}$. Требуется проверить выполнение условий

$$\langle \nabla\varphi(x^{r+1}), p^j \rangle = 0, \quad \langle \nabla\varphi(x^{r+1}), \nabla\varphi(x^j) \rangle = 0,$$
$$\langle p^{r+1}, Ap^j \rangle = 0, \quad j = \overline{0, r}. \tag{21}$$

Из формулы (19) при $k = r$ имеем

$$\nabla\varphi(x^{r+1}) = \nabla\varphi(x^r) + \alpha_r A p^r.$$

Следовательно, $\langle \nabla\varphi(x^{r+1}), p^j \rangle = 0$, для $j = \overline{0, r-1}$ в силу предположения индукции. При $j = r$ это равенство справедливо согласно (16). Итак, первая группа равенств в (21) доказана.

Далее, поскольку из (17) $\nabla\varphi(x^k) = \beta_k p^{k-1} - p^k$, $k = 1, 2, \ldots$, то

$$\langle \nabla\varphi(x^{r+1}), \nabla\varphi(x^j) \rangle =$$

$$= \langle \nabla\varphi(x^{r+1}), \beta_j p^{j-1} - p^j \rangle = 0, \quad j = \overline{1, r}$$

в силу предыдущего. При $j = 0$

$$\langle \nabla\varphi(x^{r+1}), \nabla\varphi(x^0) \rangle = -\langle \nabla\varphi(x^{r+1}), p^0 \rangle = 0.$$

Следовательно, справедлива вторая группа условий из (21).

Рассмотрим условия сопряжения для $j = \overline{0, r-1}$

$$\langle p^{r+1}, Ap^j \rangle = \langle -\nabla\varphi(x^{r+1}) + \beta_{r+1} p^r, Ap^j \rangle =$$

$$= -\langle \nabla\varphi(x^{r+1}), Ap^j \rangle = -\frac{1}{\alpha_j} \langle \nabla\varphi(x^{r+1}), \nabla\varphi(x^{j+1}) - \nabla\varphi(x^j) \rangle = 0.$$

Здесь использована формула (19). Кроме того, при $j = r$ $\quad \langle p^{r+1}, Ap^r \rangle = 0$ согласно построению. Итак, соотношения (21) справедливы, что и доказывает теорему.

Подведем итог.

По доказанному, векторы p^k, $k = \overline{1, n-1}$, построенные согласно правилу (17), A - сопряжены и отличны от нуля ($\nabla\varphi(x^k) \neq 0$, $k = \overline{0, n-1}$). Следовательно, метод (13), (15), (17), (18) решает задачу (4) не более, чем за n итераций.

В заключение, проведем некоторые упрощения в структуре метода. Исключим из рассмотрения матрицу A при подсчете коэффициентов β_k. Используя формулу (19) и условия (20), получим

$$\beta_k = \frac{\langle \nabla\varphi(x^k), \nabla\varphi(x^k) - \nabla\varphi(x^{k-1})\rangle}{\langle p^{k-1}, \nabla\varphi(x^k) - \nabla\varphi(x^{k-1})\rangle} =$$

$$= \frac{\langle \nabla\varphi(x^k), \nabla\varphi(x^k)\rangle}{\langle \nabla\varphi(x^{k-1}), \nabla\varphi(x^{k-1})\rangle}.$$

Здесь использовано представление

$$p^{k-1} = -\nabla\varphi(x^{k-1}) + \beta_{k-1}p^{k-2}.$$

Заметим также, что согласно той же формулы (19)

$$\nabla\varphi(x^k) = \nabla\varphi(x^{k-1}) + \alpha_{k-1}Ap^{k-1}.$$

Таким образом, вычислительная схема метода сопряженных градиентов имеет вид

$$x^{k+1} = x^k + \alpha_k p^k, \quad k = 0, 1, \ldots;$$

$$\alpha_k = -\frac{\langle p^k, \nabla\varphi(x^k)\rangle}{\langle p^k, Ap^k\rangle};$$

$$p^k = \begin{cases} -\nabla\varphi(x^0), & k = 0 \\ -\nabla\varphi(x^k) + \beta_k p^{k-1}, & k = 1, 2, .. \end{cases}$$

$$\beta_k = \frac{\langle \nabla\varphi(x^k), \nabla\varphi(x^k)\rangle}{\langle \nabla\varphi(x^{k-1}), \nabla\varphi(x^{k-1})\rangle};$$

$$\nabla\varphi(x^k) = \nabla\varphi(x^{k-1}) + \alpha_{k-1}Ap^{k-1}, \quad k = 1, 2, \ldots$$

Отметим, что на каждой итерации $k = 1, 2, \ldots$ требуется только одно умножение матрицы на вектор - при подсчете α_k.

Заметим, что шаг α_k выбран как решение задачи

$$\varphi(x^k + \alpha p^k) \to \min, \quad \alpha > 0.$$

Итерационный параметр β_k получен из условия сопряжения

$$\langle p^{k-1}, Ap^k\rangle = 0.$$

Заключение. *Согласно полученному результату (свойство конечности) метод сопряженных градиентов, итерационный по форме, является точным методом решения линейной системы (1).*

§10. Системы с прямоугольными матрицами

Рассмотрим систему

$$Ax = b, \tag{1}$$

где A – $(m \times n)$ – матрица коэффициентов, $b \in R^m$ – вектор правых частей (система m уравнений с n неизвестными).

Отметим, что система (1) может не иметь решений (в обычном смысле). Расширим понятие решения системы (1). Сформулируем экстремальную задачу на минимум невязки в евклидовой норме

$$\|Ax - b\| \to \min, \quad x \in R^n. \tag{2}$$

Переход от (1) к (2) называется методом наименьших квадратов. Отметим, что в отличие от (1) задача (2) всегда имеет решение.

Действительно, введем новую переменную $y = Ax$ и множество её значений

$$Y = \{y \in R^m : y = Ax, \quad x \in R^n\}.$$

Тогда задача (2) принимает вид

$$\|y - b\| \to \min, \quad y \in Y.$$

Пусть $y^0 \in Y$ – произвольный вектор. Представим y-задачу в эквивалентной форме

$$\|y - b\| \to \min, \quad \|y - b\| \le \|y^0 - b\|, \quad y \in Y.$$

Пусть Y_0 – допустимое множество полученной задачи. Оно является замкнутым и ограниченным. Следовательно, по теореме Вейерштрасса непрерывная функция $\|y - b\|$ достигает на Y_0 своего минимального значения. Это значит, что y-задача имеет решение $y^* \in Y$. Поскольку $y^* = Ax^*$, то задача (2) имеет решение x^* (возможно, неединственное).

Любое решение задачи (2) назовем псевдорешением системы (1). Если система (1) разрешима (совместна), то её псевдорешения совпадают с обычными решениями. В этом случае значение задачи (2) равно нулю.

Нормальным псевдорешением системы (1) называется псевдорешение с наименьшей нормой. Нормальное псевдорешение существует и единственно для любой системы (1). *Если система (1) разрешима, то её нормальное псевдорешение называют нормальным решением.*

По задаче (2) введем в рассмотрение функцию

$$\varphi(x) = \frac{1}{2}\langle Ax - b, Ax - b\rangle = \frac{1}{2}\langle x, A^T Ax\rangle - \langle x, A^T b\rangle + \frac{1}{2}\langle b, b\rangle.$$

Отметим, что

$$\nabla\varphi(x) = A^T Ax - A^T b = A^T(Ax - b).$$

Тогда задача (2) представляется в виде

$$\varphi(x) \to \min, \quad x \in R^n \tag{2'}$$

и эквивалентна линейной системе

$$A^T Ax = A^T b \quad (\nabla\varphi(x) = 0) \tag{3}$$

с $(n \times n)$ – симметричной, неотрицательно определенной матрицей $A^T A$.

Систему (3) называют нормальной системой уравнений. Она всегда совместна, и любое её решение есть псевдорешение системы (1).

Расмотрим случай

$$m \geq n, \quad rankA = n \tag{4}$$

(число уравнений не меньше числа неизвестных, матрица A имеет полный ранг).

Тогда, как известно, матрица $A^T A$ не вырождена (положительно определена), и единственное решение системы (3) определяется по формуле

$$x = (A^T A)^{-1} A^T b. \tag{5}$$

Матрицу $A^+ = (A^T A)^{-1} A^T$ размеров $(n \times m)$ называют псевдообратной матрицей. Она определяет нормальное псевдорешение системы (1): $x = A^+ b$. Если $m = n$, и матрица A не вырождена, то $A^+ = A^{-1}$.

Для численного решения задачи (2) можно использовать градиентные методы. Точные методы могут применяться для решения нормальной системы (3). При этом приходится работать с матрицей $A^T A$, формирование которой требует дополнительных затрат ($\frac{n^3}{2}$ операций).

Рассмотрим вопрос о решении задачи (2) в случае (4) на основе QR-разложения матрицы A.

Пусть получено представление $A = QR$, в котором

Q – $(m\times n)$-матрица с ортонормированной системой столбцов, т.е. $Q^TQ = E$;

R – $(n\times n)$-верхняя треугольная матрица с положительными диагональными элементами.

Тогда согласно формуле (5) для решения нормальной системы (3) (задачи (2)) получаем

$$x = (R^TQ^TQR)^{-1}R^TQ^Tb = R^{-1}(R^T)^{-1}R^TQ^Tb = R^{-1}Q^Tb.$$

Последнее означает, что задача (2) эквивалентна треугольной системе $Rx = Q^Tb$.

Опишем технику получения QR-разложения на основе известной процедуры ортогонализации Грама-Шмидта, примененной к столбцам $a^1, \ldots, a^n$ матрицы A.

Положим $p^1 = a^1$, $q^1 = \frac{p^1}{\|p^1\|}$. Представим общий шаг процедуры.

Пусть уже построены попарно ортогональные векторы $q^1, \ldots, q^{k-1}$ единичной нормы. Введем очередной вектор

$$p^k = a^k - \sum_{i=1}^{k-1} r_{ik}q^i, \tag{6}$$

и коэффициенты r_{ik} определим из условий ортогональности

$$\langle q^j, p^k\rangle = 0, \quad j = \overline{1, k-1}.$$

Отсюда $r_{jk} = \langle q^j, a^k\rangle$. Положим $r_{kk} = \|p^k\|$ и сформируем очередной единичный вектор $q^k = \frac{1}{r_{kk}}p^k$. В результате выстраивается ортонормированная система векторов $q^1, \ldots, q^n$.

Образуем $(m\times n)$-матрицу Q со столбцами $q^1, \ldots, q^n$ и верхнюю треугольную $(n\times n)$-матрицу R с элементами r_{jk}, $j = \overline{1, k}$. Нетрудно видеть, что это объекты искомого разложения матрицы A.

Действительно, представим равенство (6) в виде

$$a^k = \sum_{i=1}^{k-1} r_{ik}q^i + r_{kk}q^k, \quad k = \overline{1, n}.$$

В матричной форме это соотношение имеет вид $A = QR$.

Таким образом, QR-разложение прямоугольной матрицы A с условием (4) может быть получено в результате процесса ортогонализации Грама-Шмидта (с нормировкой) применительно к столбцам $a^1, \ldots, a^n$.

Глава 2. Проблема собственных значений

§1. Введение

1. Вспомогательный материал

Напомним, что собственные числа (значения) λ и собственные векторы x квадратной матрицы A порядка n определяются соотношением

$$Ax = \lambda x, \;\; x \neq 0.$$

Отсюда следует, что все собственные числа матрицы A являются корнями (нулями) характеристического многочлена

$$p(\lambda) = \det(A - \lambda E) = (-1)^n \lambda^n + p_1 \lambda^{n-1} + \ldots + p_n.$$

Уравнение $p(\lambda) = 0$ называется характеристическим уравнением матрицы A.

Итак, матрица A порядка n имеет n собственных значений λ_i, $i = \overline{1,n}$ среди которых могут быть комплексно-сопряженные пары. Всякий собственный вектор $x \neq 0$ соответствует некоторому собственному значению λ и определен с точностью до множителя $\alpha \neq 0 : Ax = \lambda x \Leftrightarrow A(\alpha x) = \lambda(\alpha x)$.

Полная проблема собственных значений - найти все собственные числа и соответствующие собственные векторы матрицы A.

Частичная проблема собственных значений - найти некоторые собственные числа и векторы матрицы A (как правило, минимальное и максимальное по модулю - границы спектра матрицы A).

Симметричная проблема собственных значений - найти собственные значения и векторы симметричной матрицы A.

Отметим известный факт: собственные значения $\lambda_1, \ldots, \lambda_n$ симметричной матрицы являются действительными числами, а соответствующие собственные векторы $x^1, \ldots, x^n$ можно выбрать попарно ортогональными (ортонормированными).

Укажем наиболее простой случай для проблемы собственных значений. Пусть A - треугольная матрица с диагональными элементами a_{ii}, $i = \overline{1,n}$. Тогда

$$p(\lambda) = (a_{11} - \lambda)(a_{22} - \lambda) \cdot \ldots \cdot (a_{nn} - \lambda),$$

и собственные значения совпадают с диагональными элементами: $\lambda_i = a_{ii}$, $i = \overline{1,n}$.

Пусть $A = diag\,(a_{11}, \ldots, a_{nn})$- диагональная матрица. Тогда $Ax = (a_{11}x_1, \ldots, a_{nn}x_n)$ и в качестве собственных векторов можно выбрать единичные орты: $x^i = e^i$, $i = \overline{1,n}$.

Определение. *Квадратные матрицы A, B порядка n называются подобными, если существует невырожденная матрица P такая, что $B = P^{-1}AP$ ($P^{-1}AP$ – преобразование подобия, P – матрица подобия).*

Преобразование подобия играет фундаментальную роль в проблеме собственных значений, что подтверждается следующим утверждением.

Лемма 1. *Если (λ, x) - собственная пара матрицы $B = P^{-1}AP$, то (λ, Px) - собственная пара матрицы A.*

Доказательство. По условию $Bx = \lambda x$, т.е. $P^{-1}APx = \lambda x$. Отсюда $APx = \lambda Px$. □

Таким образом, преобразование подобия сохраняет собственные числа матрицы (подобные матрицы имеют один и тот же спектр). При этом собственные векторы связаны через матрицу подобия: $x \sim Px$.

Если P - ортогональная матрица, то преобразование подобия имеет вид P^TAP. Отметим, что это преобразование сохраняет свойство симметричности:

$$A^T = A, \quad B = P^TAP, \quad B^T = P^TAP = B.$$

Кроме того, ортогональное преобразование подобия сохраняет норму Фробениуса:

$$B = P^TAP \Rightarrow \|B\|_F = \|A\|_F.$$

Действительно, столбцы c^j матрицы $C = P^TA$ связаны со столбцами a^j матрицы A следующим образом: $c^j = P^Ta^j$, $j = \overline{1,n}$. Поскольку P^T-ортогональная матрица, то $\langle c^j, c^j\rangle = \langle a^j, a^j\rangle$. Следовательно,

$$\|C\|_F^2 = \sum_{j=1}^{n}\langle c^j, c^j\rangle = \sum_{j=1}^{n}\langle a^j, a^j\rangle = \|A\|_F^2.$$

Аналогично, строки $\overline{b}^i$ матрицы $B = CP$ связаны со строками $\overline{c}^i$ матрицы C соотношениями

$$\overline{b}^i = \overline{c}^iP, \quad i = \overline{1,n}, \quad \text{т.е. } \langle\overline{b}^i, \overline{b}^i\rangle = \langle\overline{c}^i, \overline{c}^i\rangle.$$

Следовательно,

$$\|B\|_F^2 = \sum_{i=1}^{n} \langle \overline{b}^i, \overline{b}^i \rangle = \sum_{i=1}^{n} \langle \overline{c}^i, \overline{c}^i \rangle = \|C\|_F^2.$$

В результате получаем требуемое свойство.

Особенно важную роль имеет свойство подобия с диагональной матрицей, для которой проблема собственных значений разрешается элементарно. Докажем критерий (необходимое и достаточное условие) такого подобия.

Лемма 2. *Матрица A подобна диагональной матрице тогда и только тогда, когда она имеет n линейно-независимых собственных векторов.*

Доказательство. Достаточность. Пусть $x^1, \ldots, x^n$ линейно-независимые собственные векторы матрицы A, соответствующие собственным значениям $\lambda_1, \ldots, \lambda_n$. Образуем матрицу P из векторов $x^1, \ldots, x^n$, располагая их по столбцам: $P = (x^1 \ldots x^n)$. Учитывая соотношения $Ax^i = \lambda_i x^i$, $i = \overline{1,n}$, получаем

$$AP = A(x^1 \ldots x^n) = (Ax^1 \ldots Ax^n) = (\lambda_1 x^1 \ldots \lambda_n x^n) = PD,$$

где $D = diag\ (\lambda_1, \ \ldots, \ \lambda_n)$ – диагональная матрица. Следовательно, $D = P^{-1}AP$.

Необходимость. Пусть существует невырожденная матрица P со столбцами $x^1 \ldots x^n$ такая, что $D = P^{-1}AP$. Тогда, $AP = PD$. Расписывая это равенство по столбцам, получаем: $Ax^i = \lambda_i x^i$, $i = \overline{1,n}$, где λ_i – диагональные элементы матрицы D. Поскольку P – невырожденная матрица, то $x^1, \ldots, x^n$ – линейно-независимые векторы. □

Следствие 1. *Если все собственные числа матрицы A различны, то она подобна диагональной.*

Доказательство. Пусть $x^1, \ldots, x^n$ – собственные векторы матрицы A, соответствующие $\lambda_1, \ldots, \lambda_n$. Покажем, что они линейно-независимы. Допустим противное. Пусть r – наименьшее число линейно зависимых векторов набора $(x^1, \ldots, x^n)$ таких, что

$$\sum_{i \in I} \alpha_i x^i = 0, \quad \alpha_i \neq 0, \quad I = \{i_1, \ldots, i_r\}. \tag{1}$$

Поскольку $x^i \neq 0$, то $r \geq 2$. Умножим (1) слева на A

$$\sum_{i \in I} \alpha_i A x^i = \sum_{i \in I} \alpha_i \lambda_i x^i = 0. \tag{2}$$

Умножим теперь (1) на λ_{i_1} и вычтем из (2). В результате

$$\sum_{i\in I\setminus\{i_1\}} \alpha_i(\lambda_i - \lambda_{i_1})x^i = 0.$$

Все коэффициенты этой комбинации не равны нулю. Получили противоречие с определением числа r. Итак, собственные векторы $x^1, \ldots, x^n$ линейно-независимы. □

Следствие 2. *Симметричная матрица A подобна диагональной с ортогональной матрицей подобия:*

$$D = P^T AP, \quad P^T P = E, \quad D = diag\ (\lambda_1, \ldots, \lambda_n).$$

Доказательство. Для симметричной матрицы существует ортонормированная система собственных векторов

$$x^1, \ldots, x^n : \quad \langle x^i, x^j\rangle = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}.$$

Составим матрицу P из столбцов $x^1, \ldots, x^n$. Тогда $(P^T P)_{ij} = \langle x^i, x^j\rangle$, т.е. $P^T P = E$. В силу леммы 1 P – матрица подобия, т.е. $P^T AP = D$. □

Рассмотрим вопрос о локализации собственных значений матрицы A.

Теорема (Гершгорин). *Пусть*

$$A = \{a_{ij}\}, \quad i, j = \overline{1, n}, \quad R_i(A) = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i = \overline{1, n}.$$

Тогда все собственные значения матрицы A заключены в объединении n кругов на комплексной плоскости (круги Гершгорина)

$$G(A) = U_{i=1}^n \{z \in C : |z - a_{ii}| \leq R_i(A)\}.$$

Доказательство. Пусть (λ, x) – собственная пара матрицы A, т.е. $Ax = \lambda x, \quad x \neq 0$. Пусть, далее, x_p – координата вектора x с наибольшей абсолютной величиной: $|x_p| = \max_{1\leq i\leq n} |x_i|$. Понятно, что $x_p \neq 0$. Запишем равенство $\lambda x = Ax$ для p-ой координаты

$$\lambda x_p = (Ax)_p = \sum_{j=1}^{n} a_{pj}x_j. \quad \Rightarrow \quad x_p(\lambda - a_{pp}) = \sum_{j=1 j\neq p}^{n} a_{pj}x_j.$$

Следовательно,

$$|x_p||\lambda - a_{pp}| \leq \sum_{j=1 j\neq p}^{n} |a_{pj}x_j| \leq |x_p| \sum_{j=1 j\neq p}^{n} |a_{pj}| = |x_p|R_p(A).$$

Итак, $|\lambda - a_{pp}| \leq R_p(A)$, т.е. λ находится в замкнутом круге с центром a_{pp} и радиусом $R_p(A)$. Поэтому для любого собственного значения имеем $\lambda \in G(A)$. $\square$

Следствие 1. *Пусть A – матрица со строгим диагональным преобладанием, т.е.*

$$|a_{ii}| > \sum_{j=1 j \neq i}^{n} a_{ij}, \quad i = \overline{1,n}.$$

Тогда A – невырожденная матрица.

Действительно, запишем неравенство в виде $|a_{ii}| > R_i(A)$, $i = \overline{1,n}$. Это значит, что действительное число $0 \notin G(A)$, т.е. нуль не является собственным значением матрицы A. Поэтому A – невырожденная матрица.

Следствие 2. *Если A – симметричная матрица, то все ее собственные значения (действительные числа) находятся в объединении n отрезков на вещественной прямой:*

$$|\alpha - a_{ii}| \leq R_i(A), \quad i = \overline{1,n}.$$

2. Матрица вращения

Пусть на плоскости R^2 задан вектор a с координатами u, v. Построим вектор $a' = (u', v')$, повернув вектор a вокруг начала координат на угол φ против часовой стрелки. Известно, что координаты векторов a, a' связаны соотношениями

$$u' = u\cos\varphi - v\sin\varphi,$$
$$v' = u\sin\varphi + v\cos\varphi.$$

Введем матрицу

$$U = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}.$$

Тогда можно записать $a' = Ua$. Матрицу U называют матрицей вращения, соотношение $a' = Ua$ - преобразованием вращения, при этом φ - угол поворота. Нетрудно видеть, что матрица вращения является ортогональной.

В общем, n – мерном случае матрица вращения определяется индексами $i, j \in (1, \dots n)$ с условием $i < j$, параметром $\varphi \in R$ и имеет вид

$$U_{ij}(\varphi) = \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & \dots & & & & & & & \\ & & \cos\varphi & \dots & -\sin\varphi & & & & i \\ & & \dots & \dots & & & & & \\ & & \sin\varphi & \dots & \cos\varphi & & & & j \\ & & & \dots & & & & & \\ & & & & & 1 & & & \\ & & & & & & 1 & & \\ & & & & & & & 1 & \\ & & i & & j & & & & \end{pmatrix}$$

Таким образом, $U_{ij}(\varphi)$ отличается от единичной матрицы четырьмя элементами в позициях $(i,j),(j,i),(i,i),(j,j)$. Нетрудно проверить, что $U_{ij}(\varphi)$ - ортогональная матрица.

Пусть $A = \{a_{ij}\}$ - $(n \times n)$ – симметричная матрица. Рассмотрим преобразование подобия с матрицей вращения

$$C = U_{ij}(\varphi)^T A U_{ij}(\varphi). \tag{3}$$

Понятно, что C - симметричная матрица.

Поставим *задачу:* найти параметр φ так, чтобы $c_{ij} = 0$.

С этой целью найдем выражение для элемента c_{ij} через параметр φ. Положим $B = AU_{ij}(\varphi)$. В силу структуры $U_{ij}(\varphi)$ матрица B отличается от A только столбцами с номерами i, j. В соответствии с правилом матричного умножения элементы этих столбцов определяются по формулам

$$b_{pi} = a_{pi}\cos\varphi + a_{pj}\sin\varphi,$$

$$b_{pj} = -a_{pi}\sin\varphi + a_{pj}\cos\varphi, \quad p = \overline{1,n}. \tag{4}$$

Аналогично, матрица $C = U_{ij}(\varphi)^T B$ отличается от матрицы B только строками с номерами i, j. Элементы этих строк определяются по формулам

$$c_{ip} = b_{ip}\cos\varphi + b_{jp}\sin\varphi,$$

$$c_{jp} = -b_{ip}\sin\varphi + b_{jp}\cos\varphi, \quad p = \overline{1,n} \tag{5}$$

Из соотношений (4), (5) получаем

$$c_{ij} = b_{ij}\cos\varphi + b_{jj}\sin\varphi =$$

$$= (-a_{ii}\sin\varphi + a_{ij}\cos\varphi)\cos\varphi + (-a_{ji}\sin\varphi + a_{jj}\cos\varphi)\sin\varphi.$$

Отсюда с учетом симметричности матрицы A получаем итоговое выражение

$$c_{ij} = a_{ij}\cos 2\varphi + \frac{1}{2}(a_{jj} - a_{ii})\sin 2\varphi.$$

Решим уравнение $c_{ij} = 0$ относительно φ. В результате найдем требуемое значение для параметра φ

$$\varphi(i,j) = \frac{1}{2}\arctan\alpha_{ij}, \quad \alpha_{ij} = \frac{2a_{ij}}{a_{ii} - a_{jj}}. \tag{6}$$

Таким образом, для любой пары индексов i, j $(i < j)$ выбор параметра φ по формуле (6) обращает в нуль элемент c_{ij} матрицы C. С учетом свойства симметричности $c_{ji} = 0$. Подчеркнем, что матрица C отличается от матрицы A только элементами строк и столбцов с номерами i, j.

Выясним одно свойство преобразования (3), (6). Для матрицы A введем величину

$$\Delta(A) = \sum_{i,j=1 \; i\neq j}^{n} a_{ij}^2.$$

Это сумма квадратов внедиагональных элементов матрицы A. Величина $\Delta(A)$ характеризует меру близости матрицы A к диагональной:

$$\Delta(A) \geq 0, \quad \Delta(A) = 0 \Leftrightarrow \quad A = \ diag.$$

Найдем связь между величинами $\Delta(A)$, $\Delta(C)$. Рассмотрим формулы (5) при $p \neq i, j$. С учетом связи между матрицами A, B имеем $b_{ip} = a_{ip}$, $b_{jp} = a_{jp}$, $p \neq i, j$. Следовательно,

$$c_{ip} = a_{ip}\cos\varphi + a_{jp}\sin\varphi, \quad c_{jp} = -a_{ip}\sin\varphi + a_{jp}\cos\varphi.$$

Отсюда получаем $c_{ip}^2 + c_{jp}^2 = a_{ip}^2 + a_{jp}^2$, $p \neq i, j$. Это значит, что при переходе от A к C сумма квадратов внедиагональных элементов, за исключением элементов в позициях (i,j), (j,i), остается постоянной:

$$\Delta(C) - 2c_{ij}^2 = \Delta(A) - 2a_{ij}^2.$$

Поскольку $c_{ij} = 0$, то получаем результат

$$\Delta(C) = \Delta(A) - 2a_{ij}^2.$$

Таким образом, матрица C, полученная с помощью преобразования (3), (6), при условии $a_{ij} \neq 0$ имеет меньшее отклонение от диагональной, нежели матрица A.

3. Спектральные задачи

Пусть A – симметричная матрица с собственными значениями λ_i, $i = \overline{1,n}$. Выделим границы её спектра

$$\lambda_{\min} = \min_{1 \le i \le n} \lambda_i, \quad \lambda_{\max} = \max_{1 \le i \le n} \lambda_i$$

и установим связь $\lambda_{\min}, \lambda_{\max}$ с квадратичной формой $\langle x, Ax \rangle$.

Поскольку A – симметричная матрица, то имеет место представление $P^T A P = \mathcal{D}$, где $P^T P = E$, $\mathcal{D} = diag(\lambda_1, \ldots, \lambda_n)$.

Тогда $A = P\mathcal{D}P^T$, причем $\langle x, Ax \rangle = \langle P^T x, \mathcal{D}P^T x \rangle$. Положим $y = P^T x$. Тогда

$$\langle y, y \rangle = \langle x, x \rangle, \quad \langle x, Ax \rangle = \langle y, \mathcal{D}y \rangle = \sum_{i=1}^{n} \lambda_i y_i^2.$$

Запишем очевидные оценки для суммы

$$\sum_{i=1}^{n} \lambda_i y_i^2 \le \lambda_{\max} \langle y, y \rangle = \lambda_{\max} \langle x, x \rangle,$$

$$\sum_{i=1}^{n} \lambda_i y_i^2 \ge \lambda_{\min} \langle y, y \rangle = \lambda_{\min} \langle x, x \rangle.$$

С учетом предыдущего получаем двустороннюю оценку для квадратичной формы

$$\lambda_{\min} \langle x, x \rangle \le \langle x, Ax \rangle \le \lambda_{\max} \langle x, x \rangle.$$

Перейдем к единичным по норме векторам

$$\lambda_{\min} \le \langle x, Ax \rangle \le \lambda_{\max}, \quad \langle x, x \rangle = 1.$$

Пусть $x^{\min}$ – нормированный собственный вектор, соответствующий $\lambda_{\min}$: $Ax^{\min} = \lambda_{\min} x^{\min}$. Тогда $\lambda_{\min} = \langle x^{\min}, Ax^{\min} \rangle$. Следовательно, имеет место представление

$$\lambda_{\min} = \min_{\langle x,x \rangle = 1} \langle x, Ax \rangle.$$

Аналогично,

$$\lambda_{\max} = \max_{\langle x,x \rangle = 1} \langle x, Ax \rangle.$$

Таким образом, экстремальные собственные пары $(\lambda_{\min}, x^{\min})$, $(\lambda_{\max}, x^{\max})$ симметричной матрицы A можно находить в результате решения спектральных задач

$$\langle x, Ax\rangle \to \min, \max, \qquad \langle x, x\rangle = 1.$$

§2. Степенной метод

Метод носит итерационный характер и предназначен для приближенного решения частичной проблемы собственных значений - нахождения максимального по модулю собственного значения матрицы A вместе с соответствующим собственным вектором. Проведем описание метода для случая симметричной матрицы.

Пусть собственные числа матрицы A упорядочены по модулю:

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|,$$

т.е. искомое значение $\lambda_1 \neq 0$ является единственным. Отметим, что в данном случае

$$1 > \left|\frac{\lambda_2}{\lambda_1}\right| \geq \left|\frac{\lambda_3}{\lambda_1}\right| \geq \ldots \geq \left|\frac{\lambda_n}{\lambda_1}\right|.$$

В силу симметричности A соответствующая система собственных векторов $x^1, x^2, \ldots, x^n$ может быть выбрана ортонормированной:

$$\langle x^i, x^j\rangle = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

Возьмем произвольный вектор $y^0 \in R^n$, $y^0 \neq 0$ и организуем последовательные приближения по правилу

$$y^1 = Ay^0, \;\; y^2 = Ay^1, \ldots, y^k = Ay^{k-1}, \ldots$$

Понятно, что $y^k = A^k y^0, k = 1, 2, \ldots$ Поскольку $\{x^1, x^2, \ldots, x^n\}$ – базис пространства R^n, то имеет место представление

$$y^0 = \alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_n x^n.$$

Предположим, что $\alpha_1 \neq 0$. Согласно определению $Ax^i = \lambda_i x^i, i = \overline{1, n}$. Тогда $A^k x^i = \lambda_i^k x^i$, $k = 2, \ldots$ Следовательно, имеет место разложение

$$y^k = A^k y^0 = \alpha_1 \lambda_1^k x^1 + \alpha_2 \lambda_2^k x^2 + \ldots + \alpha_n \lambda_n^k x^n. \qquad (1)$$

Представим его в виде

$$\lambda_1^{-k} y^k = \alpha_1 x^1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k x^2 + \ldots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^k x^n$$

и перейдем к пределу при $k \to \infty$.

Поскольку

$$\left(\frac{\lambda_2}{\lambda_1}\right)^k \to 0, \ \ldots, \ \left(\frac{\lambda_n}{\lambda_1}\right)^k \to 0, \quad k \to \infty,$$

то $\lambda_1^{-k} y^k \to \alpha_1 x^1$, $k \to \infty$. Это значит, что последовательность $\{y^k\}$ сходится к собственному вектору x^1 по направлению, т.е. вектор y^k можно считать приближением к вектору $\lambda_1^k \alpha_1 x^1$, пропорциональному собственному вектору x^1.

Построим числовую последовательность, сходящуюся к собственному значению λ_1 на основе информации об y^k. Учитывая разложение (1) и свойство ортонормированности системы $\{x^1, \ldots, x^n\}$, образуем скалярные произведения

$$\langle y^k, y^k \rangle = \alpha_1^2 \lambda_1^{2k} + \alpha_2^2 \lambda_2^{2k} + \ldots + \alpha_n^2 \lambda_n^{2k} =$$

$$= \lambda_1^{2k} (\alpha_1^2 + \delta_1(k)), \quad \delta_1(k) = \alpha_2^2 \left(\frac{\lambda_2}{\lambda_1}\right)^{2k} + \ldots + \alpha_n^2 \left(\frac{\lambda_n}{\lambda_1}\right)^{2k};$$

$$\langle y^{k+1}, y^k \rangle = \alpha_1^2 \lambda_1^{2k+1} + \alpha_2^2 \lambda_2^{2k+1} + \ldots + \alpha_n^2 \lambda_n^{2k+1} =$$

$$= \lambda_1^{2k} (\alpha_1^2 \lambda_1 + \delta_2(k)), \quad \delta_2(k) = \alpha_2^2 \lambda_2 \left(\frac{\lambda_2}{\lambda_1}\right)^{2k} + \ldots + \alpha_n^2 \lambda_n (\frac{\lambda_n}{\lambda_1})^{2k}.$$

Отметим, что при $k \to \infty$ $\delta_1(k) \to 0$, $\delta_2(k) \to 0$.

Положим

$$\mu_k = \frac{\langle y^{k+1}, y^k \rangle}{\langle y^k, y^k \rangle}.$$

Согласно предыдущим представлениям

$$\mu_k = \frac{\lambda_1 \alpha_1^2 + \delta_1(k)}{\alpha_1^2 + \delta_2(k)}.$$

Следовательно, имеет место сходимость $\mu_k \to \lambda_1$, $k \to \infty$, т.е. величина μ_k может служить приближением к искомому собственному значению λ_1.

Отметим, что $\rho(A) = |\lambda_1|$. Следовательно, величина $|\mu_k|$ для больших k является приближенным значением для спектрального радиуса матрицы A.

Замечание. Пусть $r(x) = \frac{\langle x,Ax\rangle}{\langle x,x\rangle}$, $x \neq 0$ – отношение Релея. Поскольку $y^{k+1} = Ay^k$, то $\mu_k = \frac{\langle y^k,Ay^k\rangle}{\langle y^k,y^k\rangle} = r(y^k)$.

В практических вычислениях степенной метод сопровождается процедурой евклидовой нормировки последовательных приближений y^k. Тогда модифицированная схема степенного метода имеет вид:

$$y^0 \in R^n, \quad y^k = \frac{Ay^{k-1}}{\|Ay^{k-1}\|}, \quad \mu_k = \langle y^k, Ay^k\rangle, \quad k = 1, 2, \ldots$$

Пусть A – симметричная, невырожденная матрица. Поставим задачу отыскания минимального по модулю собственного значения λ_n в предположении его единственности. Рассмотрим обратную матрицу A^{-1}. Как известно, обратная величина $\frac{1}{\lambda_n}$ есть максимальное по модулю собственное значение матрицы A^{-1} с соответствующим собственным вектором x^n. Поэтому для отыскания собственной пары $(\frac{1}{\lambda_n}, x^n)$ можно использовать степенной метод применительно к матрице A^{-1}:

$$y^0 \in R^n, \quad y^k = A^{-1}y^{k-1}, \quad k = 1, 2, \ldots$$

$$y^k \approx \alpha x^n, \quad \mu_k \approx \frac{1}{\lambda_n}.$$

Отметим, что вектор y^k является решением линейной системы $Ay = y^{k-1}$. Таким образом, степенной метод для отыскания собственной пары (λ_n, x^n) имеет вид (обратные итерации):

$$y^0 \in R^n, \quad Ay^k = y^{k-1}, \quad k = 1, 2, \ldots$$

§3. Метод вращений

Рассмотрим полную проблему собственных значений для симметричной матрицы A. Метод вращений – итерационная процедура диагонализации матрицы A на основе преобразования подобия с матрицей вращения. Обозначим $A_0 = A$ и опишем общий шаг метода.

Пусть построена симметричная матрица A_k, подобная A_0. Найдем наибольший по модулю наддиагональный элемент матрицы A_k. Пусть это будет

$$a_{ls}^{(k)} : \quad |a_{ls}^{(k)}| = \max_{1\le i<j\le n} |a_{ij}^{(k)}|, \quad l < s$$

(ведущий элемент на k-том шаге). Если $a_{ls}^{(k)} = 0$, то A_k диагональная матрица и метод завершается: собственные числа A совпадают с диагональными элементами матрицы A_k.

Рассмотрим общий случай, когда $a_{ls}^{(k)} \neq 0$. Применим к матрице A_k преобразование подобия с матрицей вращения $U_{ls} = U_{ls}(\varphi)$, $\varphi = \varphi(l, s)$. В результате получаем следующее матричное приближение $A_{k+1} = U_{ls}^T A_k U_{ls}$. Отметим свойства нового приближения:

1) матрица A_{k+1} симметрична и подобна A_k,

2) $a_{ls}^{(k+1)} = 0, \ \|A_{k+1}\|_F = \|A_k\|_F$,

3) матрица A_{k+1} "ближе"к диагональной, чем A_k :

$$\Delta(A_{k+1}) = \Delta(A_k) - 2(a_{ls}^{(k)})^2.$$

Изучим вопрос о сходимости метода по невязке $\Delta(A_k)$, $k = 0, 1, \ldots$

Согласно свойству ведущего элемента $a_{ls}^{(k)}$ имеет место оценка

$$\Delta(A_k) \leq (n^2 - n)(a_{ls}^{(k)})^2.$$

Отсюда

$$(a_{ls}^{(k)})^2 \geq \frac{\Delta(A_k)}{n(n-1)}.$$

На основании свойства 3)

$$\Delta(A_{k+1}) \leq \Delta(A_k) - \frac{2\Delta(A_k)}{n(n-1)} = (1 - \frac{2}{n(n-1)})\Delta(A_k).$$

Положим $q = 1 - \frac{2}{n(n-1)}$. Поскольку $n \geq 2$, то $0 \leq q < 1$.

Оценка уменьшения невязки имеет вид

$$\Delta(A_{k+1}) \leq q \ \Delta(A_k), \quad k = 0, 1, \ldots$$

Следовательно, $\Delta(A_k) \to 0$, $k \to \infty$ с линейной скоростью со знаменателем q.

Пусть $\varepsilon > 0$ – заданная точность приближения к диагональной матрице. За конечное число итераций метода вращений придем к неравенству $\Delta(A_{k+1}) \leq \varepsilon$. Тогда $\lambda_i(A) \approx a_{ii}^{(k+1)}$, $i = \overline{1, n}$.

Обсудим вопрос о вычислении собственных векторов матрицы A. Пусть метод остановлен после k-ой итерации. Обозначим через U_k – матрицу вращения на k-ой итерации. Тогда

$$A_{k+1} = U_k^T A_k U_k = U_k^T U_{k-1}^T A_{k-1} U_{k-1} U_k = \ldots =$$
$$= U_k^T \ldots U_0^T A_0 U_0 \ldots U_k.$$

Введем матрицу

$$P_k = U_0 U_1 \ldots U_k.$$

Согласно определению она является ортогональной. Тогда $A_{k+1} = P_k^T A_0 P_k$ или $A_0 P_k = P_k A_{k+1}$, причем $A_{k+1} \approx diag$. Это значит, что столбцы матрицы P_k являются собственными векторами матрицы $A_0 = A$ ($\approx$).

Замечание 1. Если $n = 2$, то $q = 0$, т.е. $\Delta(A_1) = 0 \Leftrightarrow A_1 = diag$. Таким образом, метод вращений диагонализирует матрицу второго порядка за один шаг. При увеличении n q приближается к 1, т.е. скорость диагонализации уменьшается.

Замечание 2. Описанный вариант метода вращений называют методом Якоби.

§4. Метод Данилевского

Укажем структуру матрицы A, которая определяет в явном виде характеристический многочлен. Пусть

$$A = \begin{pmatrix} p_1 & p_2 & \ldots & & p_n \\ 1 & 0 & \ldots & & 0 \\ 0 & 1 & \ldots & & 0 \\ & & \ldots & & \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}. \qquad (1)$$

Представление (1) называется канонической формой Фробениуса для матрицы A.

При условии (1) характеристический многочлен матрицы A имеет вид

$$p(\lambda) = \det(A - \lambda E) = (-1)^n(\lambda^n - p_1\lambda^{n-1} - p_2\lambda^{n-2} - \ldots - p_n).$$

Действительно, подсчитаем определитель, последовательно применяя формулу Лапласа

$$\det(\lambda E - A) = (-1)^n \det(A - \lambda E) =$$

$$= \det \begin{pmatrix} \lambda - p_1 & -p_2 & \dots & -p_n \\ -1 & \lambda & & \\ & & \dots & \\ & & -1 & \lambda \end{pmatrix} =$$

$$= (\lambda - p_1)\lambda^{n-1} + \det \begin{pmatrix} -p_2 & -p_3 & \dots & -p_n \\ -1 & \lambda & & \\ & & \dots & \\ & & -1 & \lambda \end{pmatrix} =$$

$$= (\lambda - p_1)\lambda^{n-1} - p_2\lambda^{n-2} + \det \begin{pmatrix} -p_3 & & \dots & -p_n \\ -1 & \lambda & & \\ & & \dots & \\ & & -1 & \lambda \end{pmatrix} =$$

$$= (\lambda - p_1)\lambda^{n-1} - p_2\lambda^{n-2} - p_3\lambda^{n-3} - \ldots - p_n = (-1)^n p(\lambda).$$

Пусть λ – собственное число матрицы (1) т.е. $p(\lambda) = 0$. Для нахождения соответствующего собственного вектора x получаем систему ($Ax = \lambda x$)

$$p_1x_1 + p_2x_2 + \ldots + p_nx_n = \lambda x_1,$$

$$x_1 = \lambda x_2, \quad x_2 = \lambda x_3, \quad \ldots \quad , x_{n-1} = \lambda x_n.$$

Поскольку собственный вектор определен с точностью до множителя, то можно положить $x_n = 1$. Тогда $x_{n-1} = \lambda$, $x_{n-2} = \lambda^2, \ldots, x_1 = \lambda^{n-1}$. Таким образом, собственный вектор, соответствующий λ, имеет вид

$$x = (\lambda^{n-1}, \lambda^{n-2}, \ldots, \lambda, 1).$$

При этом первое уравнение системы принимает вид

$$p_1\lambda^{n-1} + p_2\lambda^{n-2} + \ldots + p_n = \lambda^n,$$

что равносильно характеристическому уравнению $p(\lambda) = 0$.

Метод Данилевского приводит произвольную матрицу A к канонической форме (1) с помощью $(n-1)$ преобразований подобия (с сохранением собственных значений).

Опишем первый шаг метода. Преобразуем последнюю строку матрицы $A = \{a_{ij}\}$ к канонической форме. С этой целью образуем элементарную

матрицу ($a_{n,n-1} \neq 0$)

$$M_{(n-1)} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & \dots & & \\ -\frac{a_{n1}}{a_{n,n-1}} & -\frac{a_{n2}}{a_{n,n-1}} & \dots & \frac{1}{a_{n,n-1}} & -\frac{a_{nn}}{a_{n,n-1}} \\ & & & & 1 \end{pmatrix}.$$

Понятно, что $\det M_{(n-1)} = \frac{1}{a_{n,n-1}} \neq 0$, причем обратная матрица имеет вид

$$M^{-1}_{(n-1)} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & \dots & & \\ a_{n1} & a_{n2} & \dots & a_{n,n-1} & a_{nn} \\ & & & & 1 \end{pmatrix}.$$

Построим матрицу $A_{(1)} = M^{-1}_{(n-1)} A M_{(n-1)}$. Нетрудно видеть, что она имеет следующую структуру

$$A_{(1)} = \begin{pmatrix} a^{(1)}_{11} & \dots & a^{(1)}_{1,n-1} & a^{(1)}_{1n} \\ & & \dots & \\ a^{(1)}_{n-1,1} & \dots & a^{(1)}_{n-1,n-1} & a^{(1)}_{n-1,n} \\ 0 & \dots & 1 & 0 \end{pmatrix},$$

т.е. последняя строка соответствует форме Фробениуса.

Второй шаг метода имеет вид: $A_{(2)} = M^{-1}_{(n-2)} A^{(1)} M_{(n-2)}$ при условии $a^{(1)}_{n-1,n-2} \neq 0$. Матрица $M_{(n-2)}$ образуется относительно матрицы $A_{(1)}$ аналогично предыдущему. В результате последние две строки матрицы $A_{(2)}$ имеют каноническую форму.

После выполнения $(n-1)$ шагов матрица A будет приведена к форме Фробениуса

$$A_{(n-1)} = \begin{pmatrix} a^{(n-1)}_{11} & a^{(n-1)}_{12} & \dots & a^{(n-1)}_{1n} \\ 1 & 0 & & \\ & & & \\ & 1 & & \\ & \dots & \dots & \\ & & 1 & 0 \end{pmatrix}.$$

Глава 3. Методы решения нелинейных систем

§1. Метод простой итерации

Метод простой итерации (метод итераций, метод повторных подстановок, метод последовательных приближений) является одним из основных в вычислительной математике и применяется для решения широкого класса уравнений. Проведем описание и обоснование метода для системы нелинейных уравнений вида

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

Представим её в виде, удобном для итераций

$$x_i = \varphi_i(x_1, \ldots, x_n), \quad i = \overline{1, n}. \tag{1}$$

Полагая $x = (x_1, \ldots, x_n)$, $\Phi = (\varphi_1, \ldots, \varphi_n)$, запишем систему (1) в векторной форме

$$x = \Phi(x). \tag{2}$$

Отметим, что решение уравнения (2) есть неподвижная точка отображения $\Phi : R^n \to R^n$.

Последовательные приближения $x^1, x^2, \ldots$ в методе простой итерации строятся по правилу

$$x^{k+1} = \Phi(x^k), \quad k = 0, 1, \ldots \tag{3}$$

где x^0 – заданное начальное приближение.

Проведем обоснование метода в некоторой норме $\|x\|$ пространства R^n.

Теорема*(о сходимости). Пусть*

1) вектор-функция $\Phi(x)$ определена в области

$$S = \{x : \|x - x^0\| \le \delta\}, \quad \delta > 0;$$

2) для $\forall x, y \in S$ выполняется условие

$$\|\Phi(x) - \Phi(y)\| \le q\|x - y\|, \quad q \in (0, 1); \tag{4}$$

3) справедливо неравенство

$$\|\Phi(x^0) - x^0\| = r \le (1 - q)\delta. \tag{5}$$

Тогда в методе (3)

1) $x^k \in S, \quad k = 1, 2, \ldots;$

2) $x^k \to x^*, \quad k \to \infty$, *где* $x^* \in S$ *– решение системы (2);*

3) $\|x^k - x^*\| \le \frac{r}{1-q} q^k, \quad k = 1, 2, \ldots$

Замечание 1. Неравенство (4) есть условие Липшица для вектор-функции $\Phi(x)$ в области S с константой $q \in (0, 1)$ (условие сжатия). Оно показывает, что Φ является оператором сжатия в области S, т.е. для уравнения (2) действует принцип сжатых отображений. Утверждения теоремы означают, что уравнение (2) имеет решение x^* в области S, и последовательные приближения x^k сходятся к этому решению со скоростью геометрической прогрессии со знаменателем q.

Доказательство. Поскольку $x^0 \in S$, то для приближения $x^1 = \Phi(x^0)$ в силу предположения (5) имеем $\|x^1 - x^0\| = r \le (1-q)\delta < \delta$. Это значит, что $x^1 \in S$. Покажем, что $x^k \in S$, $k = 2, \ldots$, причем для соседних приближений выполняется неравенство

$$\|x^{k+1} - x^k\| \le rq^k, \quad k = 0, 1, \ldots \tag{6}$$

Будем рассуждать по индукции. При $k = 0$ утверждение справедливо, т.к. $x^1 \in S$ и $\|x^1 - x^0\| = r$. Допустим, что приближения $x^1, \ldots, x^m$ принадлежат S, и неравенство (6) выполнено для $k = \overline{0, m-1}$. Поскольку $x^m \in S$, то для $x^{m+1} = \Phi(x^m)$ с учетом условия (4) имеем

$$\|x^{m+1} - x^m\| = \|\Phi(x^m) - \Phi(x^{m-1})\| \le q\|x^m - x^{m-1}\|.$$

По индуктивному предположению $\|x^m - x^{m-1}\| \le rq^{m-1}$. Следовательно, $\|x^{m+1} - x^m\| \le rq^m$, т.е. неравенство (6) справедливо для $k = m$. Покажем, что $x^{m+1} \in S$. Учитывая свойство (6) при $k = \overline{0, m}$, получаем

$$\|x^{m+1} - x^0\| \le \|x^{m+1} - x^m\| + \|x^m - x^{m-1}\| + \ldots + \|x^1 - x^0\| \le$$

$$\le rq^m + rq^{m-1} + \ldots + r = r\frac{1 - q^{m+1}}{1 - q} \le \frac{r}{1 - q} \le \delta.$$

Итак, $x^{m+1} \in S$, и первое утверждение теоремы доказано.

Покажем, что последовательность x^k, $k = 0, 1, \ldots$ является сходящейся. С этой целью проверим признак сходимости Коши (покажем, что последовательность $\{x^k\}$ является фундаментальной).

По аналогии с предыдущим для любых $p = 1, 2, \ldots$ имеем

$$\|x^{k+p} - x^k\| \le \|x^{k+p} - x^{k+p-1}\| + \|x^{k+p-1} - x^{k+p-2}\| + \ldots + \|x^{k+1} - x^k\| \le$$

$$\le r(q^{k+p-1} + q^{k+p-2} + \ldots + q^k) = r\frac{q^k - q^{k+p}}{1-q} \le \frac{r}{1-q}\, q^k.$$

Поскольку $q \in (0,1)$, то $q^k \to 0$, $k \to \infty$, поэтому для $\forall \varepsilon > 0$ найдется такой номер k_ε, что для $k > k_\varepsilon$ будет

$$\|x^{k+p} - x^k\| < \varepsilon, \quad p = 1, 2, \ldots$$

Это означает выполнение признака Коши, что гарантирует сходимость последовательности $\{x^k\}$. Обозначим $x^* = \lim_{k\to\infty} x^k$. Поскольку $x^k \in S$, и множество S замкнуто, то $x^* \in S$. Переходя к пределу при $k \to \infty$ в соотношении (3) и учитывая непрерывность вектор-функции Φ в области S (следствие условия (4)), получаем $x^* = \Phi(x^*)$. Утверждение 2) теоремы доказано.

Для доказательства последнего утверждения воспользуемся полученным выше неравенством

$$\|x^k - x^{k+p}\| \le \frac{r}{1-q}\, q^k, \quad k = 0, 1, \ldots, \quad p = 1, 2, \ldots$$

Перейдем здесь к пределу при $p \to \infty$. Учитывая непрерывность функции $\|x\|$ и тот факт, что $x^{k+p} \to x^*$, $p \to \infty$, получаем требуемый результат - утверждение 3). □

Замечание 2. В условиях теоремы решение x^* уравнения (2) в области S является единственным.

Действительно, пусть имеются два решения $x, y \in S : x = \Phi(x)$, $y \in \Phi(y)$, причем $x \ne y$. Тогда

$$\|x - y\| = \|\Phi(x) - \Phi(y)\| \le q\|x - y\| < \|x - y\|,$$

что и требовалось показать.

Обсудим условие 2) доказанной теоремы. Рассмотрим уравнение (2) в покомпонентной записи (1) и предположим, что функции $\varphi_i(x)$, $i = \overline{1,n}$ непрерывно-дифференцируемы в области S (т.е. существуют и непрерывны в S частные производные $\frac{\partial \varphi_i(x)}{\partial x_j}$, $i, j = \overline{1,n}$).

Выясним достаточное условие выполнения неравенства (4) в этом случае.

Образуем матрицу Якоби системы функций $\varphi_i(x)$, $i = \overline{1,n}$

$$\Phi'(x) = \begin{pmatrix} \frac{\partial\varphi_1(x)}{\partial x_1} & \dots & \frac{\partial\varphi_1(x)}{\partial x_n} \\ \dots\dots & & \\ \frac{\partial\varphi_n(x)}{\partial x_1} & \dots & \frac{\partial\varphi_n(x)}{\partial x_n} \end{pmatrix}.$$

Далее, будем использовать обобщенную теорему о среднем (обобщение на случай вектор-функции формулы конечных приращений Лагранжа)

$$\|\Phi(x) - \Phi(y)\| \le \max_{\alpha\in[0,1]} \|\Phi'(z^\alpha\|\|x - y\|, \;\; x, y \in S.$$

Здесь матричная норма согласована с векторной, $z^\alpha = x + \alpha(y - x)$, $\alpha \in [0,1]$ – точки отрезка, соединяющего x, y.

Поскольку S – выпуклое множество, то $z^\alpha \in S$. Предположим, что имеет место оценка

$$\|\Phi'(x)\| \le q, \quad x \in S, \tag{7}$$

причем $q \in (0,1)$.

Тогда согласно предыдущему выполняется условие 2) теоремы

$$\|\Phi(x) - \Phi(y)\| \le q\|x - y\|, \;\; x, y \in S.$$

Таким образом, в случае дифференцируемости условие (7) на матрицу Якоби $\Phi'(x)$ гарантирует условие сжатия для вектор-функции $\Phi(x)$.

§2. Метод Ньютона

Рассмотрим систему уравнений

$$f_i(x_1, \dots, x_n) = 0, \;\; i = \overline{1,n}$$

в предположении, что $f_i : R^n \to R$ – непрерывно - дифференцируемые функции.

Полагая $x = (x_1, \dots, x_n)$, $F = (f_1, \dots, f_n)$, перейдем к векторной записи

$$F(x) = 0. \tag{1}$$

Опишем общий шаг метода. Пусть уже получено приближение x^k, $k = 0, 1, \dots$. Проведем линеаризацию вектор-функции $F(x)$ в окрестности точки x^k

$$F(x) \approx F_k(x) = F(x^k) + F'(x^k)(x - x^k).$$

Здесь $F^{'}(x) = \{\frac{\partial f_i(x)}{\partial x_j}\}$ – матрица Якоби для вектор-функции $F(x)$.

Очередное приближение x^{k+1} определяется как решение линейной системы $F_k(x) = 0$, т.е.

$$F^{'}(x^k)(x - x^k) = -F(x^k).$$

Если матрица Якоби $F^{'}(x^k)$ не вырождена, то решение линейной системы можно записать в явном виде, что приводит к стандартной формуле метода Ньютона

$$x^{k+1} = x^k - (F^{'}(x^k))^{-1}F(x^k), \quad k = 0, 1, \ldots \tag{2}$$

Таким образом, в основе метода Ньютона лежит идея линеаризации вектор-функции $F(x)$ в окрестности каждого приближения (на каждой итерации), что позволяет свести решение системы (1) к последовательному решению линейных систем.

Рассмотрим вопрос о сходимости метода. Пусть в пространстве R^n выбрана некоторая векторная норма $\|x\|$ и согласованная с ней матричная норма $\|A\|$.

Теорема *(о сходимости). Пусть*

1) вектор-функция $F(x)$ определена и непрерывно-дифференцируема в области

$$S_\delta = \{x : \|x - x^*\| < \delta\}, \quad \delta > 0,$$

где x^* – решение уравнения (1),

2) для всех $x \in S_\delta$ существует обратная матрица $(F^{'}(x))^{-1}$, причем

$$\|(F^{'}(x))^{-1}\| \le \alpha_1, \quad \alpha_1 > 0,$$

3) для всех $x, y \in S_\delta$

$$\|F(x) - F(y) - F^{'}(y)(x - y)\| \le \alpha_2\|x - y\|^2, \quad \alpha_2 > 0,$$

4) $x^0 \in S_\varepsilon$, $\varepsilon = \min\{\delta, \frac{1}{\alpha}\}$, $\alpha = \alpha_1\alpha_2$.

Тогда для метода Ньютона (2)

1) $x^k \in S_\varepsilon$, $k = 1, 2, \ldots$

2) $x^k \to x^*$, $k \to \infty$,

3) $\|x^k - x^*\| \le \frac{1}{\alpha}(\alpha\|x^0 - x^*\|)^{2^k}$, $k = 1, 2, \ldots$

Доказательство. Докажем первое утверждение теоремы с помощью индукции. По условию $x^0 \in S_\varepsilon$. Допустим, что $x^k \in S_\varepsilon$. Поскольку $\varepsilon \le \delta$, то $S_\varepsilon \subset S_\delta \Rightarrow x^k \in S_\delta$. Рассмотрим условие 3) теоремы для $x = x^*$, $y = x^k$

$$\|F(x^*) - F(x^k) - F'(x^k)(x^* - x^k)\| \le \alpha_2 \|x^* - x^k\|^2.$$

Согласно формуле (2)

$$-F(x^k) = F'(x^k)(x^{k+1} - x^k),$$

кроме того $F(x^*) = 0$. Тогда предыдущее неравенство принимает вид

$$\|F'(x^k)(x^{k+1} - x^*)\| \le \alpha_2 \|x^* - x^k\|^2.$$

Следовательно,

$$\|(x^{k+1} - x^*)\| = \|(F'(x^k))^{-1} F'(x^k)(x^{k+1} - x^*\| \le$$

$$\le \|(F'(x^k))^{-1}\| \|F'(x^k)(x^{k+1} - x^*)\| \le \alpha_1 \alpha_2 \|x^k - x^*\|^2 = \alpha \|x^k - x^*\|^2.$$

Таким образом, имеет место неравенство

$$\|x^{k+1} - x^*\| \le \alpha \|x^k - x^*\|^2. \tag{3}$$

По предположению индукции $x^k \in S_\varepsilon$. Поскольку в силу условия 4) $\varepsilon \le \frac{1}{\alpha}$, то

$$\|x^{k+1} - x^*\| < \alpha \varepsilon^2 \le \varepsilon.$$

Это значит, что $x^{k+1} \in S_\varepsilon$, и шаг индукции реализован. Первое утверждение теоремы доказано.

Продолжим доказательство. Положим $q_k = \alpha \|x^k - x^*\|$, $k = 0, 1, \ldots$. Перепишем оценку (3) после умножения на α в виде $q_{k+1} \le q_k^2$. Покажем, что

$$q_k \le q_0^{2^k}, \quad k = 0, 1, \ldots. \tag{4}$$

Будем рассуждать по индукции. При $k = 0$ неравенство (4) очевидно. Допустим, что оно справедливо для некоторого k. Тогда

$$q_{k+1} \le q_k^2 \le (q_0^{2^k})^2 = q_0^{2^{k+1}}.$$

Переход $k \to k + 1$ завершен, т.е. неравенство (4) справедливо для всех k. Перепишем его исходных обозначениях

$$\|x^k - x^*\| \le \frac{1}{\alpha} (\alpha \|x^0 - x^*\|)^{2^k}.$$

Получили утверждение 3). При этом $q_0 = \alpha\|x^0 - x^*\| < \alpha\varepsilon \le 1$, т.е. $q_0 < 1$. Это значит, что имеет место сходимость: $x^k \to x^*$, $k \to \infty$ $\square$

Замечание 1. Неравенство (3) при условии $\alpha\|x^0 - x^*\| < 1$ означает, что последовательность $\{x^k\}$ сходится к решению x^* с квадратичной скоростью.

Замечание 2. Поскольку $\|x^0 - x^*\| < \varepsilon$, то из утверждения 3) следует оценка погрешности метода Ньютона

$$\|x^k - x^*\| \le \frac{1}{\alpha}(\alpha\varepsilon)^{2^k}, \quad \alpha = \alpha_1\alpha_2.$$

Укажем некоторые *модификации* метода (2).

1. Естественным упрощением метода является следующая процедура (снижение трудоемкости реализации, потеря в скорости сходимости)

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

2. В итерационную формулу метода Ньютона введем параметр $\beta_k > 0$ следующим образом

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

На каждой итерации β_k находится так, чтобы уменьшить невязку уравнения (1), т.е. выполнить неравенство

$$\|F(x^{k+1})\| < \|F(x^k)\|. \tag{5}$$

Проведем обоснование такой процедуры в евклидовой норме.

Введем в рассмотрение функцию-невязку для уравнения (1)

$$\varphi(x) = \frac{1}{2}\langle F(x), F(x)\rangle = \frac{1}{2}\|F(x)\|^2.$$

Найдем градиент $\nabla\varphi(x)$, используя представление

$$F(x + \triangle x) = F(x) + F'(x)\triangle x + o(\|\triangle x\|).$$

С этой целью выделим главный член приращения

$$\varphi(x + \triangle x) - \varphi(x) = \frac{1}{2}(\langle F(x + \triangle x), F(x + \triangle x)\rangle - \langle F(x), F(x)\rangle) =$$

$$= \langle F(x), F'(x)\triangle x\rangle + o_1(\|\triangle x\|) = \langle (F'(x))^T F(x), \triangle x\rangle + o_1(\|\triangle x\|).$$

Следовательно, по определению $\nabla\varphi(x) = (F'(x))^T F(x)$.

Обозначим $p^k = -(F'(x^k))^{-1}F(x^k)$ и найдем производную функции φ в точке x^k по направлению p^k :

$$\langle p^k, \nabla\varphi(x^k)\rangle = -\langle (F'(x^k))^{-1}F(x^k), (F'(x^k))^T F(x^k)\rangle =$$
$$= -\langle F(x^k), F(x^k)\rangle < 0,$$

если $x^k \neq x^*$.

Таким образом, p^k – есть направление спуска для функции $\varphi(x)$ в точке x^k : $\varphi(x^k + \beta p^k) < \varphi(x^k)$ для малых $\beta > 0$. Это значит, что выбор шага β_k согласно условию (5) возможен.

Замечание 3. По части реализации метода (2) отметим следующее. Для каждого $k = 0, 1, ..$ нет необходимости вычислять обратную матрицу $(F'(x^k))^{-1}$. Полагая $y = x - x^k$ найдем решение y^k линейной системы с матрицей Якоби $F'(x^k)y = -F(x^k)$. Тогда $x^{k+1} = x^k + y^k$.

§3. Квазиньютоновский метод

С целью снижения трудоемкости расчетов построим метод линеаризации без использования матрицы частных производных.

Рассмотрим систему уравнений $F(x) = 0$, где $F : R^n \to R^n$ – непрерывная вектор-функция. Пусть имеется приближение x^k и для $F(x)$ построена линейная аппроксимация $F_k(x)$ следующей структуры

$$F_k(x) = F(x^k) + A_k(x - x^k),$$

где A_k – известная матрица.

Очередное приближение x^{k+1} найдем как решение линейной системы $F_k(x) = 0$. Отметим, что если $x^{k+1} = x^k$, то $F_k(x^k) = F(x^k) = 0$, т.е. x^k – решение исходной системы. Предположим, что $x^{k+1} \neq x^k$.

Построим класс линейных аппроксимаций относительно точки x^{k+1}

$$F_{k+1}(x, A) = F(x^{k+1}) + A(x - x^{k+1})$$

и рассмотрим вопрос о выборе матрицы A.

Потребуем, чтобы $F_{k+1}(x^k, A) = F(x^k)$. Это приводит к следующему условию на матрицу A :

$$A(x^k - x^{k+1}) = F(x^k) - F(x^{k+1}).$$

Обозначим: $s^k = x^{k+1} - x^k$, $y^k = F(x^{k+1}) - F(x^k)$. Тогда $As^k = y^k$. Назовем это равенство соотношением секущих.

Используя некоторую матричную норму, сформулируем экстремальную задачу на множестве матриц A

$$\|A - A_k\| \to \min, \quad As^k = y^k. \tag{1}$$

Пусть A_{k+1} – решение этой задачи. Тогда очередная линейная аппроксимация имеет вид $F_{k+1}(x) = F_{k+1}(x, A_{k+1})$, что и завершает k-ую итерацию метода секущих: переход $(x^k, A_k) \Rightarrow (x^{k+1}, A_{k+1})$.

Найдем решение задачи (1) в явном виде, используя матричную норму Фробениуса: $\|A\|_F^2 = \sum_{i,j=1}^n a_{ij}^2$.

Предварительно введем в рассмотрение специальный класс матриц. Пусть $u, v \in R^n$. Образуем $(n \times n)$ матрицу

$$uv^T = \begin{pmatrix} u_1v_1 & u_1v_2 & \dots & u_1v_n \\ u_2v_1 & u_2v_2 & \dots & u_2v_n \\ & \dots & & \\ u_nv_1 & u_nv_2 & \dots & u_nv_n \end{pmatrix} = \{u_iv_j,\ i,j = \overline{1,n}\}.$$

Понятно, что любая пара строк (столбцов) матрицы uv^T линейно зависима. Следовательно, ранг uv^T при $u, v \neq 0$ равен 1, поэтому uv^T называют матрицей ранга 1. При $u = v$ получаем матрицу - диаду $uu^T = \{u_iu_j\}$. Подсчитаем норму Фробениуса матрицы - диады

$$\|uu^T\|_F^2 = \sum_{i,j=1}^n u_i^2u_j^2 = (\sum_{i=1}^n u_i^2)(\sum_{j=1}^n u_j^2) = \langle u, u\rangle^2.$$

Таким образом,

$$\|uu^T\|_F = \langle u, u\rangle = \|u\|_2^2 = u^Tu.$$

Лемма. *Решение задачи (1) представляется по формуле*

$$A_{k+1} = A_k + \frac{1}{\langle s^k, s^k\rangle}(y^k - A_ks^k)(s^k)^T. \tag{2}$$

Доказательство. Проверим, что матрица A_{k+1} удовлетворяет соотношению секущих (ограничению задачи (1))

$$A_{k+1}s^k = A_ks^k + \frac{1}{\langle s^k, s^k\rangle}(y^k - A_ks^k)(s^k)^Ts^k =$$

$$= A_k s^k + y^k - A_k s^k = y^k.$$

Пусть A – произвольная матрица с условием $As^k = y^k$. Тогда

$$\|A_{k+1} - A_k\|_F = \frac{1}{\langle s^k, s^k \rangle} \|(y^k - A_k s^k)(s^k)^T\|_F =$$

$$= \frac{1}{\langle s^k, s^k \rangle} \|(A - A_k) s^k (s^k)^T\|_F \leq$$

$$\leq \frac{1}{\langle s^k, s^k \rangle} \|A - A_k\|_F \|s^k (s^k)^T\|_F = \|A - A_k\|_F.$$

Итак, получено неравенство

$$\|A_{k+1} - A_k\|_F \leq \|A - A_k\|_F, \quad As^k = y^k.$$

Подведем итог. Метод секущих для решения системы $F(x) = 0$ состоит в следующем.

Зададим начальное приближение: вектор x^0 и матрицу A_0 (A_0 – некоторое приближение к матрице Якоби $F'(x^0)$). Опишем переход $(x^k, A_k) \Rightarrow (x^{k+1}, A_{k+1})$.

Найдем решение s^k линейной системы $A_k s = -F(x^k)$. Тогда $x^{k+1} = x^k + s^k$. Положим $y^k = F(x^{k+1}) - F(x^k)$ и подсчитаем матрицу A_{k+1} по формуле (2). Итерация метода завершена.

Формулу (2) называют формулой пересчета Бройдена. Методы описанного типа называют квазиньютоновскими.

2. Численные методы математического анализа

Глава 1. Аппроксимация функций

§1. Задача интерполирования

Задача приближения функций возникает при решении многих проблем вычислительной математики. Интерполирование представляет собой один из стандартных способов аппроксимации. Теория интерполирования является важнейшим аппаратом численного анализа, на его основе строится большое число методов решения других задач вычислительной математики. Роль интерполяционных многочленов в численном анализе аналогична роли разложения Тейлора в классическом анализе.

Пусть функция $f(x)$ определена на $[a,b]$ числовой оси и задана своими значениями в попарно несовпадающих точках $x_0, x_1, \ldots, x_n$, причем подсчет этих значений требует большой вычислительной работы. В этих условиях естественно возникает задача аппроксимации функции $f(x)$ на $[a,b]$ с помощью легко вычисляемой функции $\phi(x)$ некоторого класса на основе информации об известных значениях $f(x_i)$, $i = \overline{0,n}$. В теории интерполирования эта задача решается следующим образом.

Пусть $\phi_0(x), \ldots, \phi_n(x)$ – некоторый набор линейно-независимых функций, определенных на $[a,b]$. Будем искать аппроксимирующую функцию $\phi(x)$ в виде линейной комбинации

$$\phi(x) = \sum_{i=0}^{n} a_i \phi_i(x) \ , \ x \in [a,b] \tag{1}$$

с неопределенными коэффициентами a_i, $i = \overline{0,n}$. Функцию $\phi(x)$ называют обобщенным многочленом по системе $\{\phi_i(x)\}$.

Параметры a_i в (1) выберем из условия совпадения значений функций f и ϕ в точках $x_0, x_1, \ldots, x_n$, т.е. потребуем, чтобы

$$\sum_{i=0}^{n} a_i \phi_i(x_j) = f(x_j) \ , \ j = \overline{0,n}. \tag{2}$$

Получили систему линейных алгебраических уравнений для отыскания коэффициентов a_i, определяющих искомую функцию $\phi(x)$.

Таким образом, в рассматриваемом случае приближение функции $f(x)$, заданной таблицей $\{x_i, f(x_i)\}$, с помощью обобщенных многочленов (1) проводится на основе принципа (2) совпадения значений в точках x_i.

Задача построения функции $\phi(x)$ вида (1) в соответствии с условиями (2) называется задачей интерполирования в классе обобщенных многочленов по данной системе функций $\{\phi_i(x)\}$. При этом используется следующая терминология:

x_i – *узлы интерполирования*;

$\{x_i, f(x_i)\}$ – *исходные данные интерполирования*;

$\phi(x)$ – *обобщенный интерполяционный многочлен для функции* $f(x)$.

Соотношения (2) называют *условиями интерполирования*, разность $R(x) = f(x) - \phi(x)$, $x \in [a, b]$ – *погрешностью интерполирования.*

Рассмотрим вопрос о существовании и единственности обобщенного интерполяционного многочлена. Понятно, что в данном случае этот вопрос сводится к условиям однозначной разрешимости линейной системы (2) относительно переменных a_i . Введем определитель матрицы системы

$$\Delta = \begin{vmatrix} \phi_0(x_0) & \phi_1(x_0) & \dots & \phi_n(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_n(x_1) \\ \dots & \dots & \dots & \dots \\ \phi_0(x_n) & \phi_1(x_n) & \dots & \phi_n(x_n) \end{vmatrix}$$

и сформулируем условие на функции $\phi_i(x)$.

Определение. *Система функций* $\{\phi_i(x),\ x \in [a, b]\}$, $i = \overline{0, n}$ *называется системой Чебышева на* $[a, b]$ *если определитель* Δ *не равен нулю для любого набора попарно различных точек* $x_0, x_1, \dots, x_n$ *отрезка* $[a, b]$.

Предположим, что функции $\phi_i(x)$, $i = \overline{0, n}$ из (1) образуют систему Чебышева на $[a, b]$. Тогда $\Delta \neq 0$, т.е. система (2) имеет единственное решение, которое можно представить по формуле Крамера

$$a_i = \frac{\Delta_i}{\Delta}\ ,\ i = \overline{0, n}\ ,$$

где определитель Δ_i получается из Δ заменой i-го столбца столбцом свободных членов $f(x_j)$, $j = \overline{0, n}$.

Интерполяционный многочлен принимает вид

$$\phi(x) = \sum_{i=0}^{n} \frac{\Delta_i}{\Delta} \phi_i(x)\ . \tag{3}$$

Получим другое представление для $\phi(x)$, в котором явно фигурируют значения $f(x_j)$. С этой целью проведем разложение определителя Δ_i по

элементам i-го столбца

$$\Delta_i = \sum_{j=0}^{n} f(x_j)\Delta_{ij} \ , \ i = \overline{0,n} \ .$$

Здесь Δ_{ij} – алгебраические дополнения элементов i-го столбца. Введем функции

$$\Phi_j(x) = \sum_{i=0}^{n} \frac{\Delta_{ij}}{\Delta}\phi_i(x) \ , \ j = \overline{0,n} \ .$$

Тогда на основании (2) имеет место представление

$$\phi(x) = \sum_{j=0}^{n} f(x_j)\Phi_j(x) \ . \tag{4}$$

В силу условий интерполирования $\phi(x_i) = f(x_i)$, $i = \overline{0,n}$ функции $\Phi_j(x)$ в узлах x_i принимают следующие значения

$$\Phi_j(x_i) = \begin{cases} 0 \ , & i \neq j, \\ 1 \ , & i = j \ . \end{cases} \tag{5}$$

Таким образом, обобщенный интерполяционный многочлен $\phi(x)$ по системе функций Чебышева $\{\phi_i(x)\}$ существует и является единственным для любых исходных данных $\{x_i \ , \ f(x_i)\}$, $i = \overline{0,n}$. При этом имеет место представление (4), где $\Phi_j(x)$ – линейные комбинации базовых функций $\phi_i(x)$, $i = \overline{0,n}$ с условиями (5).

Набор $\{\Phi_j(x)\}$, $j = \overline{0,n}$ называется интерполирующим базисом обобщенного многочлена $\phi(x)$.

Сформулируем *основные проблемы* теории интерполирования:

1) построение интерполяционных многочленов для конкретных систем Чебышева $\{\phi_i(x)\}$;

2) оценка погрешности интерполирования;

3) выбор узлов интерполирования с целью уменьшения погрешности.

Наиболее полно указанные проблемы исследованы для интерполирования в классе алгебраических многочленов. Именно этот случай будет изучаться в дальнейшем.

§2. Интерполяционный многочлен Лагранжа

1. Построение многочлена Лагранжа

Выберем в качестве базовых функций $\phi_i(x)$ набор x^i, $i = \overline{0,n}$ степеней x . Это значит, что задача интерполирования решается в классе алгебраических многочленов степени не выше n :

$$\phi(x) = \sum_{i=0}^{n} a_i x^i \ . \tag{6}$$

В данном случае определитель Δ имеет вид

$$\Delta = \begin{vmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \dots & \dots & \dots & \dots \\ 1 & x_n & \dots & x_n^n \end{vmatrix} .$$

Это определитель Вандермонда, не равный нулю в силу условия $x_i \neq x_j$, $i \neq j$.

Следовательно, набор $\{x^i\}$ образует систему Чебышева, и задача интерполирования в классе полиномов (6) допускает единственное решение. Будем искать функцию $\phi(x)$ в виде (4), где $\Phi_j(x)$ – полиномы степени n . Согласно условиям (5) каждый полином Φ_j обращается в нуль в точках $x_0, \dots, x_{j-1}, x_{j+1}, \dots, x_n$, т.е. имеет следующее представление

$$\Phi_j(x) = C(x - x_0) \dots (x - x_{j-1})(x - x_{j+1}) \dots (x - x_n) \ .$$

Постоянная C определяется из условия $\Phi_j(x_j) = 1$. В результате

$$\Phi_j(x) = \frac{(x - x_0) \dots (x - x_{j-1})(x - x_{j+1}) \dots (x - x_n)}{(x_j - x_0) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_n)} =$$

$$= \prod_{i=0, i\neq j}^{n} \frac{x - x_i}{x_j - x_i} \ , \ j = \overline{0,n} \ .$$

Таким образом, многочлен

$$L_n(x) = \sum_{j=0}^{n} f(x_j) \frac{(x - x_0) \dots (x - x_{j-1})(x - x_{j+1}) \dots (x - x_n)}{(x_j - x_0) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_n)}$$

решает задачу алгебраического интерполирования. Он называется *интерполяционным многочленом Лагранжа*. При этом $\Phi_j(x)$, $j = \overline{0,n}$ – *базисные многочлены Лагранжа.*

Укажем другую запись многочлена $L_n(x)$. По данной системе узлов интерполирования введем в рассмотрение функцию

$$\omega_{n+1}(x) = (x - x_0)(x - x_1)\dots(x - x_n) .$$

Нетрудно видеть, что

$$\omega'_{n+1}(x) = \sum_{k=0}^{n}(x - x_0)\dots(x - x_{k-1})(x - x_{k+1})\dots(x - x_n),$$

причем

$$\omega'_{n+1}(x_j) = (x_j - x_0)\dots(x_j - x_{j-1})(x_j - x_{j+1})\dots(x_j - x_n).$$

Следовательно, интерполяционный многочлен Лагранжа может быть представлен в виде

$$L_n(x) = \sum_{j=0}^{n} f(x_j)\frac{\omega_{n+1}(x)}{\omega'_{n+1}(x_j)(x - x_j)} . \tag{7}$$

2. Оценка погрешности интерполирования

Найдем выражение для погрешности интерполирования функции $f(x)$ с помощью многочлена Лагранжа $L_n(x)$ в произвольной точке $x \in [a, b]$, не совпадающей с узлами интерполирования: $x \neq x_i$, $i = \overline{0, n}$. С этой целью предположим, что функция $f(x)$ непрерывна на $[a, b]$ вместе со своими производными до $(n + 1)$-го порядка включительно.

Составим вспомогательную функцию

$$\psi(z) = f(z) - L_n(z) - K\omega_{n+1}(z) , \ z \in [a, b] ,$$

которая обладает теми же свойствами дифференцируемости, что и $f(x)$. Выберем постоянную K из условия $\psi(x) = 0$. Это значит, что

$$K = \frac{f(x) - L_n(x)}{\omega_{n+1}(x)} . \tag{8}$$

Итак, функция $\psi(z)$ по определению обращается в нуль в $(n+2)$ точках $x_0, x_1, \dots, x_n, x$ отрезка $[a, b]$. На основании теоремы Ролля ее производная $\psi'(z)$ равна нулю по крайней мере в $(n+1)$ точках $[a, b]$. Применяя теорему Ролля к функции $\psi'(z)$, получаем, что производная $\psi''(z)$ обращается в нуль по меньшей мере в n точках $[a, b]$. Продолжая эти заключения,

приходим к выводу, что $(n+1)$-ая производная $\psi^{(n+1)}(z)$ обращается в нуль по крайней мере в одной точке $\xi \in [a,b]$. Согласно определению

$$\psi^{(n+1)}(z) = f^{(n+1)}(z) - K(n+1)! \, .$$

Отсюда, учитывая формулу (8), получаем выражение для погрешности

$$f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)\omega_{n+1}(x)}{(n+1)!} \ , \quad x, \xi \in [a,b] \, . \tag{9}$$

Пусть

$$M_{n+1} = \max_{x \in [a,b]} \mid f^{(n+1)}(x) \mid \ .$$

Тогда справедлива следующая оценка погрешности интерполирования

$$\mid f(x) - L_n(x) \mid \leq \ \frac{M_{n+1} \mid \omega_{n+1}(x) \mid}{(n+1)!} \ , \ x \in [a,b] \, . \tag{10}$$

Здесь $\omega_{n+1}(x) = (x - x_0)(x - x_1) \ldots (x - x_n)$. Снимая зависимость от x в правой части (10), приходим к оценке погрешности вида

$$\mid f(x) - L_n(x) \mid \leq \ \frac{M_{n+1}}{(n+1)!} \max_{a \leq x \leq b} \mid \omega_{n+1}(x) \mid \ . \tag{11}$$

3. Оптимальный выбор узлов интерполирования

Рассмотрим вопрос об оптимальном выборе узлов x_0, x_1, $\ldots$, x_n в смысле минимизации оценки погрешности (11). Это приводит к следующей минимаксной задаче

$$\min_{x_0, \ldots, x_n} \max_{a \leq x \leq b} \mid (x - x_0) \ldots (x - x_n) \mid \ . \tag{12}$$

Поскольку каждый набор $(x_0, \ldots, x_n)$ однозначно определяется полиномом $\omega_{n+1}(x) = (x - x_0)(x - x_1) \ldots (x - x_n)$ степени $(n+1)$ со старшим коэффициентом 1 , то фактически мы получили классическую задачу об отыскании многочлена, наименее уклоняющегося от нуля (в норме пространства C). Как известно, решение такой задачи дают многочлены Чебышева $T_n(x)$, которые определяются на $[-1,1]$ по формуле

$$T_n(x) = \cos(n \arccos x) \ , \ |x| \leq 1 \, .$$

Согласно этому выражению

$$T_1(x) = x \ , \ T_2(x) = 2x^2 - 1.$$

Далее, используя тождество

$$\cos(n+1)\theta = 2\cos\theta\cos n\theta - \cos(n-1)\theta$$

при $\theta = \arccos x$, приходим к рекуррентной формуле

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \ .$$

Таким образом, $T_{n+1}(x)$ есть многочлен степени $(n+1)$ со старшим коэффициентом 2^n .

Решая уравнение $\cos[(n+1)\arccos x] = 0$, определим корни многочлена $T_{n+1}(x)$

$$x_m^* = \cos\frac{(2m+1)\pi}{2(n+1)} \ , \ m = \overline{0,n} \ .$$

Согласно определению $|T_{n+1}(x)| \le 1$, $x \in [-1,1]$, поэтому для отыскания точек максимума и минимума многочлена $T_{n+1}(x)$ решим уравнение $|T_{n+1}(x)| = 1$. В результате получаем

$$x_k = \cos\frac{k\pi}{n+1} \ , \ k = \overline{0,n+1} \ .$$

При этом $T_{n+1}(x_k) = \cos k\pi = (-1)^k$.

Многочлены $T_{n+1}^*(x) = 2^{-n}T_{n+1}(x)$ со старшим коэффициентом 1 называются многочленами, наименее отклоняющимися от нуля. Это название обусловлено следующим свойством.

Лемма. *Если $P_{n+1}(x)$ – многочлен степени $(n+1)$ со старшим коэффициентом 1 , то*

$$\max_{-1\le x\le 1}|P_{n+1}(x)| \ \ge \ \max_{-1\le x\le 1}|T_{n+1}^*(x)| = 2^{-n} \ . \tag{13}$$

Доказательство. Допустим противное, т.е.

$$\max_{-1\le x\le 1}|P_{n+1}(x)| \ < \ 2^{-n} \ . \tag{14}$$

Многочлен $Q_n(x) = T_{n+1}^*(x) - P_{n+1}(x)$ имеет степень не выше n и отличен от тождественного нуля. Кроме того,

$$\mathrm{sign}Q_n(x_k) = \mathrm{sign}[(-1)^k 2^{-n} - P_{n+1}(x_k)] =$$

$$= (-1)^k \ , \ k = \overline{0,n+1} \ ,$$

так как по условию (14) $|P_{n+1}(x)| < 2^{-n}$.

Таким образом, нетривиальный многочлен $Q_n(x)$ степени не выше n имеет $(n+1)$ различных корней, что невозможно. □

Замечание. Нетрудно доказать более сильное утверждение: если $P_{n+1}(x) \neq T^*_{n+1}(x)$, то в (13) имеет место строгое неравенство. Это значит, что для каждого n многочлен $T^*_{n+1}(x)$, наименее отклоняющийся от нуля, является единственным.

Вернемся к задаче (12) об оптимальном выборе узлов интерполирования. Пусть задача решается на $[-1,1]$. Выберем в качестве узлов $x_0, \ldots, x_n$ корни x^*_m , $m = \overline{0,n}$ полинома $T_{n+1}(x)$. Тогда $\omega_{n+1}(x) = T^*_{n+1}(x)$ и на основании леммы приходим к решению задачи (12). При этом оптимальная оценка погрешности имеет вид

$$|f(x) - L_n(x)| \leq \frac{M_{n+1}}{(n+1)!\ 2^n}\ ,\ x \in [-1,1]\ .$$

В общем случае, когда интерполирование производится на $[a,b]$, воспользуемся заменой переменных

$$x = \frac{1}{2}[(b-a)z + (b+a)]\ ,$$

в силу которой $z \in [-1,1]$. Корни многочлена $T_{n+1}(z)$ являются оптимальными узлами на $[-1,1]$. Через исходную переменную они запишутся в виде

$$x_m = \frac{1}{2}[(b-a)\cos\frac{(2m+1)\pi}{2(n+1)} + (b+a)]\ ,\ m = \overline{0,n}\ .$$

Это и есть оптимальные узлы интерполирования для $[a,b]$. Оценка погрешности имеет вид

$$|f(x) - L_n(x)| \leq \frac{M_{n+1}(b-a)^{n+1}}{(n+1)!\ 2^{2n+1}}\ ,\ x \in [a,b]\ .$$

§3. Интерполяционная формула Ньютона

Найдем другую форму представления интерполяционного многочлена Лагранжа. Предварительно введем понятие разделенной разности.

1. Разделенные разности и их свойства

Пусть функция $f(x)$ определена на $[a,b]$ и задана таблицей своих значений $\{x_i,\ f(x_i)\}$ $i=\overline{0,n}$ в попарно не совпадающих точках $x_0,\dots,x_n$.

Разделенные разности нулевого порядка определяются как значения $f(x_i)\ ,\ i=\overline{0,n}\ .$

Далее, образуем отношения

$$f(x_i;x_{i+1})=\frac{f(x_{i+1})-f(x_i)}{x_{i+1}-x_i}\ ,\ i=\overline{0,n-1}\ ,$$

которые назовем разделенными разностями первого порядка.

Разделенные разности второго порядка $f(x_i;x_{i+1};x_{i+2})$ определим соотношениями

$$f(x_i;x_{i+1};x_{i+2})=\frac{f(x_{i+1};x_{i+2})-f(x_i;x_{i+1})}{x_{i+2}-x_i}\ ,\ i=\overline{0,n-2}\ .$$

В общем случае разделенная разность k-го порядка $f(x_i;\ x_{i+1};\ \dots;x_{i+k})$ определяется через разности $(k-1)$-го порядка по формуле

$$f(x_i;x_{i+1};\dots;x_{i+k})=\frac{f(x_{i+1};\dots;x_{i+k})-f(x_i;\dots;x_{i+k-1})}{x_{i+k}-x_i}\ ,$$

$$i=\overline{0,n-k}\ .$$

Разделенные разности удобно располагать в виде таблицы (случай $n=3$)

x_0	$f(x_0)$			
		$f(x_0;x_1)$		
x_1	$f(x_1)$		$f(x_0;x_1;x_2)$	
		$f(x_1;x_2)$		$f(x_0;x_1;x_2;x_3)$
x_2	$f(x_2)$		$f(x_1;x_2;x_3)$	
		$f(x_2;x_3)$		
x_3	$f(x_3)$			

Установим связь между разделенными разностями k-го порядка и значениями функции в точках x_i .

Лемма. *Для каждого* $k=\overline{1,n}$ *справедливо представление*

$$f(x_i;x_{i+1};\dots;x_{i+k})=\sum_{j=0}^{k}\frac{f(x_{i+j})}{\prod_{l=0,l\neq j}^{k}(x_{i+j}-x_{i+l})}\ ,\ i=\overline{0,n-k}\ . \qquad (15)$$

Доказательство проведем по индукции. При $k = 1$ утверждение леммы справедливо, так как

$$f(x_i; x_{i+1}) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{f(x_i)}{x_i - x_{i+1}} + \frac{f(x_{i+1})}{x_{i+1} - x_i} \ .$$

Допустим, что представление (15) имеет место при $k = m-1$ и пусть $k = m$. Тогда

$$f(x_i; x_{i+1}; \ldots; x_{i+m}) = \frac{f(x_{i+1}; \ldots; x_{i+m}) - f(x_i; \ldots; x_{i+m-1})}{x_{i+m} - x_i} =$$

$$= \frac{1}{x_{i+m} - x_i}[\sum_{j=1}^{m} \frac{f(x_{i+j})}{\prod_{l=1, l\neq j}^{m}(x_{i+j} - x_{i+l})} - \sum_{j=0}^{m-1} \frac{f(x_{i+j})}{\prod_{l=0, l\neq j}^{m-1}(x_{i+j} - x_{i+l})}] \ .$$

Подсчитаем коэффициенты при значениях $f(x_{i+j})$, $j = \overline{0, m}$ –

$$f(x_i) \ \ (j = 0) \ : \ \frac{1}{\prod_{l=1}^{m}(x_i - x_{i+l})} \ ;$$

$$f(x_{i+m}) \ \ (j = m) \ : \ \frac{1}{\prod_{l=0}^{m-1}(x_{i+m} - x_{i+l})} \ ;$$

$$f(x_{i+j}) \ \ (0 < j < m) \ :$$

$$\frac{1}{x_{i+m} - x_i}[\frac{1}{\prod_{l=1, l\neq j}^{m}(x_{i+j} - x_{i+l})} - \frac{1}{\prod_{l=0, l\neq j}^{m-1}(x_{i+j} - x_{i+l})}] =$$

$$= \frac{x_{i+j} - x_i - x_{i+j} + x_{i+m}}{(x_{i+m} - x_i)\prod_{l=0, l\neq j}^{m}(x_{i+j} - x_{i+l})} = \frac{1}{\prod_{l=0, l\neq j}^{m}(x_{i+j} - x_{i+l})} \ .$$

В совокупности получаем формулу (15) при $k = m$. □

Выделим частный случай формулы (15), полагая в ней $i = 0$ и используя известный нам многочлен $\omega_{n+1}(x)$

$$f(x_0; x_1; \ldots; x_k) = \sum_{j=0}^{k} \frac{f(x_j)}{\prod_{l=0, l\neq j}^{k}(x_j - x_l)} = \tag{16}$$

$$= \sum_{j=0}^{k} \frac{f(x_j)}{\omega'_{k+1}(x_j)} \ , \ k = 1, 2, \ldots \ .$$

Учитывая представление (15), укажем простейшие свойства разделенной разности:

a) разделенная разность является линейным оператором от функции f : если $f = \alpha_1 f_1 + \alpha_2 f_2$, $\alpha_i = \text{const}$, то

$$f(x_i; \ldots; x_{i+k}) = \alpha_1 f_1(x_i; \ldots; x_{i+k}) + \alpha_2 f_2(x_i; \ldots; x_{i+k}) ;$$

b) разделенная разность есть симметрическая функция своих аргументов $x_i, \ldots, x_{i+k}$, т.е. не меняется при любой их перестановке.

2. Вывод формулы Ньютона с разделенными разностями

Проведем теперь вывод формулы Ньютона. Пусть $L_k(x)$ - интерполяционный многочлен Лагранжа, построенный для функции $f(x)$ по узлам x_0, $\ldots, x_k$. Тогда

$$L_n(x) = L_0(x) + \sum_{k=1}^{n} [L_k(x) - L_{k-1}(x)] . \tag{17}$$

Каждая разность $L_k(x) - L_{k-1}(x)$ есть многочлен степени k , равный нулю в точках $x_0, \ldots, x_{k-1}$, т.е.

$$L_k(x) - L_{k-1}(x) = A(x - x_0) \ldots (x - x_{k-1}) =$$

$$= A\omega_k(x) , \ A = \text{const} .$$

Для определения коэффициента A положим $x = x_k$. Учитывая условие интерполирования $L_k(x_k) = f(x_k)$ и формулу (7) для многочлена Лагранжа, получаем

$$A = \frac{f(x_k)}{\omega_k(x_k)} - \frac{1}{\omega_k(x_k)} \sum_{j=0}^{k-1} f(x_j) \frac{\omega_k(x_k)}{\omega_k'(x_j)(x_k - x_j)} =$$

$$= \frac{f(x_k)}{\omega_{k+1}'(x_k)} + \sum_{j=0}^{k-1} \frac{f(x_j)}{\omega_{k+1}'(x_j)} = \sum_{j=0}^{k} \frac{f(x_j)}{\omega_{k+1}'(x_j)} .$$

Следовательно, на основании выражения (16) заключаем, что

$$A = f(x_0; \ldots; x_k) .$$

Таким образом,

$$L_k(x) - L_{k-1}(x) = f(x_0; \ldots; x_k)\omega_k(x),$$

и согласно представлению (17) интерполяционный многочлен принимает вид

$$L_n(x) = f(x_0) + \sum_{k=1}^{n} f(x_0; \ldots; x_k)\omega_k(x) .$$

В развернутой записи

$$L_n(x) = f(x_0) + f(x_0; x_1)(x - x_0) + f(x_0; x_1; x_2)(x - x_0)(x - x_1) + \ldots + \quad (18)$$
$$+ f(x_0; \ldots; x_n)(x - x_0) \ldots (x - x_{n-1}) \ .$$

Получили *интерполяционный многочлен в форме Ньютона с разделенными разностями*. Представим погрешность интерполирования в произвольной точке $x \neq x_i$ с помощью разделенной разности

$$f(x) - L_n(x) = f(x) - \sum_{j=0}^{n} f(x_j) \frac{\omega_{n+1}(x)}{\omega'_{n+1}(x_j)(x - x_j)} =$$

$$= \omega_{n+1}(x)[\frac{f(x)}{\omega_{n+1}(x)} + \sum_{j=0}^{n} \frac{f(x_j)}{\prod_{i=0, i\neq j}^{n}(x_j - x_i)(x_j - x)}] \ .$$

Согласно формуле (15), выражение в квадратных скобках есть разделенная разность $f(x; x_0; \ldots; x_n)$. Следовательно, погрешность интерполирования допускает следующее представление, согласованное с формулой (18)

$$f(x) - L_n(x) = f(x; x_0; \ldots; x_n)(x - x_0) \ldots (x - x_n) \ , \quad x \neq x_i \ .$$

Сравнивая это представление с выражением (9), полученным ранее для погрешности, легко установить связь между разделенной разностью порядка $(n + 1)$ и производной порядка $(n + 1)$ функции $f(x)$:

$$f(x; x_0; \ldots; x_n) = \frac{f^{(n+1)}(\xi)}{(n + 1)!} \ , \quad \xi, x \in [a, b] \ .$$

§4. Интерполирование с кратными узлами

До сих пор изучалась задача интерполирования по заданной таблице значений функции. Рассмотрим более общий вариант этой задачи, когда в узлах интерполирования известны не только значения функции, но и ее производных до некоторого порядка. В рамках этой информации приходим к задаче интерполирования с кратными узлами.

1. Интерполяционный многочлен Эрмита

Пусть на $[a,b]$ выделены попарно не совпадающие точки $x_0, x_1, \ldots, x_n$ – узлы интерполирования. Предположим, что задана следующая таблица интерполяционных данных

$$\begin{array}{cccc} f(x_0) & f'(x_0) & \ldots & f^{(\alpha_0-1)}(x_0) \\ f(x_1) & f'(x_1) & \ldots & f^{(\alpha_1-1)}(x_1) \\ \ldots & \ldots & \ldots & \ldots \\ f(x_n) & f'(x_n) & \ldots & f^{(\alpha_n-1)}(x_n) \end{array} \tag{19}$$

Числа $\alpha_0, \alpha_1, \ldots, \alpha_n$ назовем кратностями узлов $x_0, x_1, \ldots, x_n$ соответственно. Пусть $\alpha_0 + \alpha_1 + \ldots + \alpha_n = m+1$ – общее число известных данных о функции $f(x)$.

Поставим задачу интерполирования функции $f(x)$ в классе многочленов степени m

$$H_m(x) = a_0 x^m + a_1 x^{m-1} + \ldots + a_m \ , \tag{20}$$

выбирая коэффициенты $a_0 \ldots, a_m$ так, чтобы выполнялись условия кратного интерполирования

$$H_m^{(l)}(x_k) = f^{(l)}(x_k), \quad l = \overline{0, \alpha_k - 1}, \quad k = \overline{0, n}. \tag{21}$$

Полином $H_m(x)$ с условиями (21) называется *интерполяционным многочленом Эрмита*. Выясним вопрос о его существовании и единственности.

Равенства (21) представляют собой, по существу, систему линейных алгебраических уравнений относительно переменных $a_0, \ldots, a_m$. Покажем, что она имеет единственное решение. Для этого достаточно проверить, что соответствующая однородная система

$$H_m^{(l)}(x_k) = 0 \ , \ l = \overline{0, \alpha_k - 1}, \quad k = \overline{0, n}. \tag{22}$$

имеет только нулевое решение.

Соотношения (22) означают, что каждый узел x_k является для полинома $H_m(x)$ корнем с кратностью не меньше α_k . Тогда сумма кратностей всех корней полинома $H_m(x)$ не меньше, чем $\alpha_0+\alpha_1+\ldots+\alpha_n = m+1$. Поскольку m – степень полинома $H_m(x)$, то заключаем, что $H_m(x) \equiv 0$, т.е. все коэффициенты $a_0 \ldots, a_m$ равны нулю. Таким образом, однородная система (22) имеет только нулевое решение. Это значит, что интерполяционный многочлен Эрмита существует и единственен.

Замечание. Отметим некоторые частные случаи многочлена Эрмита. Если $\alpha_i = 1$, $i = \overline{0,n}$, то получаем обычную задачу интерполирования, и многочлен $H_m(x)$ превращается в многочлен Лагранжа. Если положить $n = 0$ (взять один узел x_0), то интерполяционный многочлен Эрмита представляет собой полином Тейлора для функции $f(x)$ относительно точки x_0 .

2. Погрешность кратного интерполирования

Рассмотрим вопрос о представлении погрешности кратного интерполирования.

Теорема. *Пусть функция $f(x)$ $(m+1)$ раз непрерывно-дифференцируема на $[a,b]$. Тогда существует такая точка $\xi \in [a,b]$, что*

$$f(x) - H_m(x) = \frac{f^{(m+1)}(\xi)}{(m+1)!}(x-x_0)^{\alpha_0}\dots(x-x_n)^{\alpha_n} \ , \ x \in [a,b] \ .$$

Доказательство. Обозначим

$$A(x) = (x-x_0)^{\alpha_0}\dots(x-x_n)^{\alpha_n}$$

и зафиксируем точку $x \neq x_k$, $k = \overline{0,n}$. Рассмотрим вспомогательную функцию

$$F(z) = f(z) - H_m(z) - \frac{A(z)}{A(x)}[f(x) - H_m(x)] \ , \ z \in [a,b] \ .$$

По определению функция $F(z)$ нерерывно дифференцируема на $[a,b]$ до порядка $(m+1)$ включительно. При этом точки $x_0,\dots,x_n,x$ являются нулями функции $F(z)$ кратности не ниже, чем $\alpha_0,\dots,\alpha_n,1$ соответственно. Значит, сумма кратностей всех нулей функции $F(z)$ не меньше, чем $\alpha_0 + \dots + \alpha_n + 1 = m+2$. На основании теоремы Ролля производная $F'(z)$ обращается в нуль по крайней мере один раз на каждом интервале (x_0,x_1), (x_1,x_2) , $\dots$, (x_n,x) . Кроме того, точки $x_0,\dots,x_n$ будут для $F'(z)$ нулями кратности не меньше, чем $\alpha_0-1,\dots,\alpha_n-1$. Следовательно, сумма кратностей всех нулей для $F'(z)$ на $[a,b]$ не меньше величины $(\alpha_0-1)+\dots+(\alpha_n-1)+(n+1) = m+1$.

Продолжая эти рассуждения для второй и последующих производных, приходим к заключению, что производная $F^{(m+1)}(z)$ имеет на $[a,b]$ по крайней мере один корень. Таким образом, существует точка $\xi \in [a,b]$, для которой

$$F^{(m+1)}(\xi) = f^{(m+1)}(\xi) - \frac{(m+1)!}{A(x)}[f(x) - H_m(x)] = 0 \ .$$

Отсюда сразу следует утверждение теоремы.

§5. Сплайн – интерполирование

Для повышения качества аппроксимации функции $f(x)$ на $[a,b]$ проведем разбиение этого отрезка на части с помощью точек

$$a = x_0 < x_1 < \ldots < x_{N-1} < x_N = b.$$

Совокупность точек x_j, $j = \overline{0,N}$ называется сеткой Δ на $[a,b]$, при этом x_j – узлы сетки, $\Delta_j = [x_{j-1}, x_j]$, $j = \overline{1,N}$ – ячейки разбиения.

На каждом частичном отрезке Δ_j функция $f(x)$ аппроксимируется некоторым интерполяционным многочленом $P_j(x)$ степени n относительно выбранной системы узлов интерполирования x_{jl}, $l = \overline{1,m}$ с условием

$$x_{j-1} \le x_{j1} < x_{j2} < \ldots < x_{jm} \le x_j, \quad m \le n+1.$$

Условия интерполирования имеют вид

$$P_j(x_{jl}) = f(x_{jl}), \quad l = \overline{1,m}.$$

Кроме того, задаются условия непрерывности во внутренних узлах сетки $x_1, x_2, \ldots x_{N-1}$ (условия состыковки, сопряжения многочленов)

$$P_j^{(k)}(x_j) = P_{j+1}^{(k)}(x_j), \quad k = \overline{0,q}\ , \quad j = \overline{1,N-1}.$$

Для граничных узлов $x_0 = a$, $x_N = b$ задаются краевые условия

$$P_1^{(k)}(a) = \gamma_a^{(k)}, \quad k = \overline{0,q_o}\ ; \quad P_N^{(k)}(b) = \gamma_b^{(k)}, \quad k = \overline{0,q_N}.$$

Функцию $s(x)$, определенную на $[a,b]$ и совпадающую на каждом частичном отрезке Δ_j с многочленом $P_j(x)$, $j = \overline{1,N}$ при выполнении всех указанных условий называют интерполяционным сплайном относительно сетки Δ.

При этом узлы сетки называются узлами сплайна, число n – степенью сплайна.

Таким образом, сплайн – кусочно-полиномиальная функция относительно выбранной сетки с определенным порядком гладкости в её узлах.

Условия интерполирования, непрерывности и краевые условия образуют систему линейных алгебраических уравнений относительно параметров

сплайна (коэффициентов многочленов $P_j(x)$, $j = \overline{1, N}$). Вопрос о существовании и единственности сплайн-функции $s(x)$, $x \in [a, b]$ определяется условиями однозначной разрешимости этой системы. Как минимум, число требуемых условий должно совпадать с числом параметров сплайна.

Опишем наиболее употребительные сплайны для случаев $n = 1, 2, 3$.

1. Линейный сплайн

Пусть функция $f(x)$ определена и непрерывна на $[a, b]$. Введем сетку

$$\Delta : a = x_0 < x_1 < \ldots < x_N = b$$

и вычислим значения функции f в узлах x_j, $j = \overline{0, N}$.

Определение. *Функция $s(x)$, $x \in [a, b]$ называется линейным сплайном относительно сетки Δ, если*

1) *$s(x)$ – линейная функция на каждой ячейке $\Delta_j = [x_{j-1}, x_j]$, $j = \overline{1, N}$;*

2) *$s(x)$ непрерывна на $[a, b]$;*

3) *$s(x_j) = f(x_j)$, $j = \overline{0, N}$.*

Понятно, что сплайн-функция $s(x)$ определяется условиями 1)–3) однозначно – число параметров сплайна $(2N)$ совпадает с числом условий $(N - 1 + N + 1 = 2N)$.

С геометрической точки зрения $s(x)$ – ломаная с угловыми точками $(x_j, f(x_j))$, $j = \overline{1, N-1}$.

В данном случае узлы сплайна совпадают с узлами интерполяции.

2. Параболический сплайн

Пусть функция $f(x)$ определена и непрерывно–дифференцируема на $[a,b]$. Введем на $[a,b]$ две сетки

$$\Delta: \quad a = x_0 < x_1 < \ldots < x_N = b,$$

$$\Delta': \quad a \le z_0 < z_1 < \ldots < z_{N-1} \le b$$

с условием $z_{j-1} < x_j < z_j$, $j = \overline{1, N-1}$.

Вычислим значения функции $f(x)$ в узлах сетки Δ' : $f(z_j)$ $j = \overline{0, N-1}$.

Определение. *Функция $s(x)$, $x \in [a,b]$ называется параболическим сплайном относительно сеток Δ, Δ', если*

1) *$s(x)$ – многочлен второй степени на каждой ячейке $\Delta_j = [x_{j-1}, x_j]$, $j = \overline{1, N}$;*

2) *$s(x)$ непрерывно–дифференцируема на $[a,b]$;*

3) $s(z_j) = f(z_j)$, $j = \overline{0, N-1}$.

В данном случае узлы сетки Δ – узлы сплайна, узлы сетки Δ' – узлы интерполяции. При этом число параметров сплайна: $3N$, число условий: $3N-2$.

Для однозначного определения сплайна необходимо задать два краевых условия.

Сплайн $s(x)$ называется периодическим, если

$$s(a) = s(b), \quad s'(a) = s'(b).$$

В непериодическом случае краевые условия задаются в виде

$$s'(a) = \gamma'_a, \quad s'(b) = \gamma'_b \quad \text{(краевые условия 1-го рода)},$$

либо

$$s''(a) = \gamma''_a, \quad s''(b) = \gamma''_b \quad \text{(краевые условия 2-го рода)}.$$

Если функция $f(x)$ имеет соответствующие производные, то полагают:

$$\text{в условиях 1 рода} \; - \quad \gamma'_a = f'(a), \quad \gamma'_b = f'(b),$$

$$\text{в условиях 2 рода} \; - \quad \gamma''_a = f''(a), \quad \gamma''_b = f''(b).$$

3. Кубический сплайн

Пусть функция $f(x)$ определена и дважды непрерывно-дифференцируема на $[a,b]$. Введем сетку

$$\Delta: \quad a = x_0 < x_1 < \ldots < x_N = b$$

и вычислим значения функции $f(x)$ в узлах сетки Δ : $f(x_j)$, $j = \overline{0,N}$.

Определение. *Функция $s(x)$, $x \in [a,b]$ называется кубическим сплайном на сетке Δ, если*

1) $s(x)$ *– многочлен третьей степени на каждой ячейке* $\Delta_j = [x_{j-1}, x_j]$, $j = \overline{1,N}$;

2) $s(x)$ *дважды непрерывно–дифференцируема на* $[a,b]$;

3) $s(x_j) = f(x_j)$, $j = \overline{0,N}$.

В данном случае узлы сплайна совпадают с узлами интерполяции.

Число параметров: $4N$, общее число условий: $4N-2$. Необходимо дополнительно задать два краевых условия аналогично параболическому случаю.

§6. Наилучшее приближение функций в классе полиномов

1. Наилучшее равномерное приближение (чебышевская аппроксимация)

Пусть функция $y = f(x)$, $x \in [a,b]$ задана таблицей своих значений

$$y_i = f(x_i), \quad i = \overline{0,N} \tag{1}$$

в точках (узлах) $x_0, x_1, \ldots, x_N$, упорядоченных по возрастанию

$$a \le x_0 < x_1 < \ldots < x_N \le b.$$

Проведем аппроксимацию функции $f(x)$ по таблице (1) в классе алгебраических полиномов $P_n(x,\alpha)$ степени не выше n $(n \le N)$ с набором коэффициентов $\alpha = (a_0, a_1, \ldots, a_n)$

$$P_n(x,\alpha) = a_0 + a_1 x + \ldots + a_n x^n.$$

В качестве критерия аппроксимации выберем величину максимального отклонения

$$\varphi_n(\alpha) = \max_{0 \le i \le N} |y_i - P_n(x_i, \alpha)|.$$

Положим

$$\rho_n = \inf_{\{\alpha\}} \varphi_n(\alpha)$$

(точная нижняя грань по всем наборам коэффициентов).

Задача наилучшего приближения функции $f(x)$ по таблице (1) (задача чебышевской аппроксимации) состоит в построении многочлена $P_n(x, \alpha^*)$, для которого

$$\varphi_n(\alpha^*) = \rho_n.$$

При этом $P_n(x, \alpha^*)$ – полином наилучшего приближения для функции $f(x)$ по таблице (1).

Решение задачи наилучшего приближения существенно зависит от соотношения между n и N (при общем условии $n \le N$).

Если $n = N$, то $P_n(x, \alpha^*)$ есть, очевидно, интерполяционный полином функции $f(x)$ по таблице (1), удовлетворяющий условиям

$$P_n(x_i, \alpha^*) = y_i, \quad i = \overline{0, n}.$$

В этом случае $\rho_n = 0$, и задача о наилучшем приближении переходит в обыкновенную задачу интерполирования функции $f(x)$ по таблице (1).

При $n = N - 1$ задача о наилучшем приближении называется задачей чебышевской интерполяции. Сформулируем основной результат для этого случая.

Теорема 1. *В случае $n = N - 1$ полином наилучшего приближения существует и является единственным. Для того, чтобы полином $P_n(x, \alpha)$ был полиномом наилучшего приближения необходимо и достаточно, чтобы для некоторого числа h выполнялись соотношения*

$$y_i - P_n(x_i, \alpha) = (-1)^i h, \quad i = \overline{0, n+1}. \tag{2}$$

На основании этой теоремы задача чебышевской интерполяции эквивалентна решению системы линейных алгебраических уравнений относительно неизвестных h, $a_0, \ldots, a_n$

$$h + a_0 + a_1 x_0 + \ldots + a_n x_0^n = y_0,$$

$$-h + a_0 + a_1 x_1 + \ldots + a_n x_1^n = y_1, \qquad (2')$$

$$\ldots\ldots\ldots$$

$$(-1)^{n+1} h + a_0 + a_1 x_{n+1} + \ldots + a_n x_{n+1}^n = y_{n+1}.$$

При этом $\rho_n = |h|$.

Рассмотрим общий случай, когда $n < N - 1$. Справедливо следующее утверждение.

Теорема 2. *В случае $n < N - 1$ полином наилучшего приближения существует и является единственным. Для того, чтобы полином $P_n(x, \alpha)$ решал задачу о наилучшем приближении необходимо и достаточно, чтобы он осуществлял чебышевскую интерполяцию на некоторой совокупности из $(n + 2)$ узлов $x_{i_0} < x_{i_1} < \ldots < x_{i_{n+1}}$ исходного набора $(x_0, x_1, \ldots, x_N)$.*

Итак, общая задача о наилучшем приближении сводится к последовательности задач чебышевской интерполяции для всевозможных наборов $\{x_{i_0}, \ldots, x_{i_{n+1}}\}$.

Представим непрерывный вариант задачи о наилучшем приближении. Пусть $f(x)$ – непрерывная функция, заданная на $[a, b]$. Проведем её аппроксимацию в классе алгебраических многочленов $P_n(x, \alpha)$ по критерию

$$\varphi_n(\alpha) = \max_{a \le x \le b} |f(x) - P_n(x, \alpha)|.$$

Положим $\rho_n = \inf_{\{\alpha\}} \varphi_n(\alpha)$.

Задача наилучшего равномерного приближения функции $f(x)$ на $[a, b]$ состоит в построении многочлена $P_n(x, \alpha^*)$, для которого $\varphi_n(\alpha^*) = \rho_n$. При этом $P_n(x, \alpha^*)$ – полином наилучшего равномерного приближения для функции $f(x)$ на $[a, b]$.

Решение задачи дается следующим утверждением.

Теорема 3. (П.Л. Чебышев) *Полином наилучшего равномерного приближения для функции $f(x)$ на $[a, b]$ существует и является единственным. Для того, чтобы полином $P_n(x, \alpha)$ был полиномом наилучшего равномерного приближения необходимо и достаточно, чтобы он осуществлял чебышевскую интерполяцию на некотором наборе из $(n + 2)$ точек $x_0, x_1, \ldots, x_{n+1}$ отрезка $[a, b]$.*

Точки $x_0, x_1, \ldots, x_{n+1}$, удовлетворяющие условиям теоремы, называют *точками чебышевского альтернанса.*

2. Наилучшее среднеквадратичное приближение (метод наименьших квадратов)

Пусть для функции $f(x)$, $x \in [a,b]$ вычислены её значения $f(x_i)$, $i = \overline{0,N}$ в точках $x_0, \ldots, x_N$ отрезка $[a,b]$. В рамках этой информации поставим задачу аппроксимации функции $f(x)$, $x \in [a,b]$ в классе алгебраических полиномов $P_n(x,\alpha)$ степени не выше n ($n \le N$) по критерию среднеквадратичного отклонения

$$\varphi_n(\alpha) = \sum_{i=0}^{N} [f(x_i) - P_n(x_i,\alpha)]^2.$$

Задача наилучшего среднеквадратичного приближения состоит в отыскании набора коэффициентов $\alpha^* = (a_0^*, \ldots, a_n^*)$, который минимизирует функцию $\varphi_n(\alpha)$

$$\varphi_n(\alpha^*) = \min_{\{\alpha\}} \varphi_n(\alpha).$$

Соответствующий многочлен $P_n(x,\alpha^*)$ называют многочленом наилучшего среднеквадратичного приближения по таблице значений. Такой способ аппроксимации называют *методом наименьших квадратов.*

Заметим, что при $n = N$ $P_n(x,\alpha^*)$ есть интерполяционный полином для функции $f(x)$ по заданной таблице, причем $\varphi_n(\alpha^*) = 0$.

Теорема 4. *При $n < N$ полином наилучшего среднеквадратичного приближения существует и является единственным. Для того, чтобы полином $P_n(x,\alpha)$ решал задачу о наилучшем среднеквадратичном приближении необходимо и достаточно, чтобы выполнялись соотношения*

$$\sum_{i=0}^{N} [f(x_i) - P_n(x_i,\alpha)] x_i^k = 0, \quad k = \overline{0,n}. \tag{3}$$

Таким образом, поставленная задача эквивалентна решению системы линейных алгебраических уравнений (3) относительно коэффициентов a_0, $\ldots, a_n$ искомого многочлена.

Заметим, что применительно к функции $\varphi_n(\alpha)$ соотношения (3) имеют смысл условий стационарности

$$\frac{1}{2} \frac{\partial \varphi_n(\alpha)}{\partial a_k} = 0, \quad k = \overline{0,n}.$$

Сохраняя терминологию, укажем набор соотношений, характеризующих непрерывную задачу о наилучшем среднеквадратичном приближении:

$$f(x), \quad x \in [a, b]$$

– аппроксимируемая функция;

$$P_n(x, \alpha) = a_0 + a_1 x + \ldots + a_n x^n$$

– аппроксимирующий многочлен;

$$\varphi_n(\alpha) = \int_a^b [f(x) - P_n(x, \alpha)]^2 dx$$

– критерий аппроксимации;

$$\varphi_n(\alpha^*) = \min_{\{\alpha\}} \varphi_n(\alpha)$$

– задача о наилучшем приближении;

$$\int_a^b [f(x) - P_n(x, \alpha)] x^k dx = 0, \quad k = \overline{0, n}$$

– необходимые и достаточные условия минимума.

Глава 2. Численное интегрирование и дифференцирование

§1. Задача численного интегрирования

Пусть функция $f(x)$ определена и интегрируема на $[a,b]$. Задача численного интегрирования состоит в приближенном вычислении $\int_a^b f(x)dx$ по известным значениям $f(x_i)$, $i=\overline{1,n}$ подынтегральной функции в некоторых точках $x_1,\dots,x_n$ из $[a,b]$.

Будем строить формулы численного интегрирования согласно правилу

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} A_i f(x_i). \tag{1}$$

Это соотношение называют *квадратурной формулой*. При этом:

правая часть (1) – *квадратурная сумма*,

$\{A_i, x_i, n\}$ – *параметры* квадратурной формулы,

A_i – *квадратурные* (весовые) *коэффициенты*, x_i – *квадратурные узлы*.

Если пределы интегрирования a, b являются квадратурными узлами, то получаем формулу *замкнутого типа*. В противном случае имеем квадратурную формулу *открытого типа*.

Величина

$$I_n(f) = \int_a^b f(x)dx - \sum_{i=1}^{n} A_i f(x_i)$$

называется *погрешностью* квадратурной формулы (1).

Если для некоторой функции $f(x)$ имеем $I_n(f) = 0$, то квадратурная формула является для этой функции точной.

Определение. *Будем говорить, что квадратурная формула (1) имеет алгебраическую степень точности m, если она является точной при $f(x) = x^k$, $k = \overline{1,m}$ и не точна при $f(x) = x^{m+1}$ ($I_n(x^k) = 0$, $k = \overline{1,m}$, $I_n(x^{m+1}) \neq 0$).*

Из определения следует, что квадратурная формула степени точности m является точной для всех алгебраических многочленов степени не выше m, причем число m – максимальная степень таких многочленов.

Укажем оценку сверху для степени точности формулы (1) при фиксированном n.

Лемма. *Степень точности формулы (1) не может быть больше* $(2n-1)$ *при любом выборе параметров* $\{A_i, x_i\}$, $i = \overline{1,n}$.

Доказательство. Рассмотрим произвольную квадратурную формулу (1). Пусть $f(x) = (x-x_1)^2 \dots (x-x_n)^2$. Это многочлен степени $2n$. Поскольку $f(x) \geq 0$, $x \in [a,b]$, то

$$\int_a^b f(x)dx > 0.$$

С другой стороны, $\sum_{i=1}^n A_i f(x_i) = 0$, т.е. формула (1) в этом случае не является точной. □

Опишем основные подходы к построению квадратурных формул.

1. Интерполяционный метод

Зафиксируем узлы x_i, $i = \overline{1,n}$ и представим функцию $f(x)$ с помощью интерполяционной формулы Лагранжа по заданной таблице $\{x_i, f(x_i)\}$

$$f(x) = L_{n-1}(x) + R_{n-1}(x), \quad x \in [a,b].$$

Здесь

$$L_{n-1}(x) = \sum_{i=1}^n f(x_i)\Phi_i(x), \quad \Phi_i(x) = \prod_{j=1 j\neq i}^n \frac{x-x_j}{x_i-x_j}.$$

($L_{n-1}(x)$ – интерполяционный многочлен Лагранжа, $\Phi_i(x)$ – базисный многочлен Лагранжа)

После интегрирования получаем квадратурную формулу

$$\int_a^b f(x)dx = \sum_{i=1}^n A_i f(x_i) + I_n(f), \tag{2}$$

в которой

$$A_i = \int_a^b \Phi_i(x)dx, \quad i = \overline{1,n},$$

$$I_n(f) = \int_a^b R_{n-1}(x)dx.$$

Эту квадратурную формулу называют *интерполяционной.*

Если квадратурные узлы берутся равноотстоящими: $x_i = x_1 + ih$, $i = \overline{2,n}$, то интерполяционная формула носит название формулы Ньютона-Котеса.

Выясним вопрос о степени точности формулы (2). Пусть $f(x)$ – алгебраический многочлен степени не выше $(n-1)$. Тогда, согласно свойству интерполяции $L_{n-1}(x) \equiv f(x)$, т.е. $I_n(f) = 0$. Следовательно, *интерполяционная квадратурная формула (2) имеет степень точности не ниже, чем* $(n-1)$.

Отсюда следует, что квадратурные коэффициенты A_i формулы (2) являются единственным решением линейной системы

$$\sum_{i=1}^{n} A_i x_i^k = \int_a^b x^k dx, \quad k = \overline{0, n-1},$$

которая получена из (2) при $f(x) = x^k$.

2. Метод неопределенных параметров

Возьмем за основу формулу (1) и будем считать все квадратурные коэффициенты равными: $A_1 = A_2 = \ldots = A_n = A$. Тогда

$$\int_a^b f(x)dx \approx A \sum_{i=1}^{n} f(x_i). \tag{3}$$

Параметры A, x_i $i = \overline{1, n}$ выберем так, чтобы формула (3) была точной для всех полиномов степени не выше n.

При этом достаточно рассмотреть функции $f(x) = x^k$, $k = \overline{0, n}$.

Для $f(x) = 1$ имеем $b - a = An$, т.е. $A = \frac{b-a}{n}$.

Полагая в (3) $f(x) = x^k$, $k = \overline{1, n}$, приходим к системе нелинейных уравнений для определения квадратурных узлов x_i, $i = \overline{1, n}$

$$\frac{b-a}{n} \sum_{i=1}^{n} x_i^k = \frac{b^{k+1} - a^{k+1}}{k+1}, \quad k = \overline{1, n}.$$

Полученная квадратурная формула носит название формулы Чебышева. В частности при $n = 1$ $x_1 = \frac{a+b}{2} \Rightarrow \int_a^b f(x)dx \approx f(\frac{a+b}{2})(b-a)$.

В заключение, укажем общий подход к построению квадратурной формулы вида (1). Он состоит в выборе параметров $\{A_i, x_i\}$, $i = \overline{1, n}$ (n – фиксировано) так, чтобы обеспечить формуле (1) максимально возможную степень точности. Квадратурная формула с таким свойством носит название формулы Гаусса.

§2. Простейшие и составные квадратурные формулы

Введем на $[a, b]$ равномерную сетку с шагом $h > 0$, т. е. множество точек $x_i = a + ih$, $i = \overline{0, n}$, $hn = b - a$. Возьмем за основу следующее представление

$$\int_a^b f(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx.$$

Для каждого частичного интеграла S_i будем использовать простейшие формулы интерполяционного типа. При оценке погрешности функция $f(x)$ предполагается достаточно гладкой.

1. Формулы прямоугольников

Пусть $\overline{x}_i = x_{i-1} + \frac{1}{2}h$ – середина $[x_{i-1}, x_i]$. Рассмотрим приближенную формулу вида

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx f(\overline{x}_i)h. \tag{1}$$

Это простейшая формула прямоугольников для частичного отрезка $[x_{i-1}, x_i]$ (формула средних прямоугольников, $\overline{x}_i$ – середина $[x_{i-1}, x_i]$.) Оценим погрешность r_i формулы (1). Согласно определению

$$r_i = \int_{x_{i-1}}^{x_i} (f(x) - f(\overline{x}_i))dx.$$

На основании формулы Тейлора

$$f(x) = f(\overline{x}_i) + f'(\overline{x}_i)(x - \overline{x}_i) + \frac{1}{2}f''(\xi_i)(x - \overline{x}_i)^2,$$

$$\xi_i = \xi_i(x) \in [x_{i-1}, x_i].$$

Тогда

$$r_i = f'(\overline{x}_i) \int_{x_{i-1}}^{x_i} (x - \overline{x}_i)dx + \frac{1}{2} \int_{x_{i-1}}^{x_i} f''(\xi_i)(x - \overline{x}_i)^2 dx.$$

Обозначим $M_{2,i} = \max_{x_{i-1} \le x \le x_i} |f''(x)|$. Поскольку первый интеграл в правой части равен нулю, то

$$|r_i| \le \frac{1}{2} M_{2,i} \int_{x_{i-1}}^{x_i} (x - \overline{x}_i)^2 dx = \frac{h^3}{24} M_{2,i}.$$

Таким образом, для погрешности простейшей формулы (1) справедлива оценка

$$|r_i| \le \frac{h^3}{24} M_{2,i}. \tag{2}$$

Это значит, что *простейшая формула (1) имеет третий порядок точности (погрешность есть величина порядка* $h^3 : r_i \sim O(h^3)$ *).*

Отметим, что оценка (2) неулучшаема, т.е. существует функция $f(x)$, для которой в (2) имеет место равенство. Действительно, для $f(x) = (x - \overline{x}_i)^2$ имеем $M_{2,i} = 2$, $f(\overline{x}_i) = 0$, $r_i = \frac{h^3}{12} = \frac{h^3}{24} M_{2,i}$.

Составная формула средних прямоугольников получается в результате суммирования в (1) по $i = \overline{1,n}$

$$\int_a^b f(x)dx \approx h\sum_{i=1}^{n} f(\overline{x}_i).$$

Оценим погрешность R этой формулы. Понятно, что

$$R = \sum_{i=1}^{n} r_i.$$

Обозначим $M_2 = \max_{x\in[a,b]} |f''(x)|$. Тогда

$$|R| \le \frac{M_2 n h^3}{24} = \frac{h^2(b-a)}{24} M_2.$$

Это значит, что погрешность составной формулы средних прямоугольников есть величина порядка h^2 (*составная формула имеет второй порядок точности*).

Замечание. Укажем другие варианты формул прямоугольников:

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx f(x_{i-1})h, \quad \int_a^b f(x)dx \approx h\sum_{i=1}^{n} f(x_{i-1})$$

– простейшая и составная формулы левых прямоугольников;

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx f(x_i)h, \quad \int_a^b f(x)dx \approx h\sum_{i=1}^{n} f(x_i)$$

– простейшая и составная формулы правых прямоугольников.

Нетрудно проверить, что простейшие (составные) формулы левых и правых прямоугольников имеют второй (первый) порядок точности.

2. Формулы трапеций

Простейшая формула трапеций имеет вид

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx \frac{f(x_{i-1} + f(x_i)}{2} h \tag{3}$$

и получается в результате замены функции $f(x)$ её интерполяционным многочленом по значениям $f(x_{i-1}), f(x_i)$ (линейная интерполяция). Этот многочлен имеет вид

$$L_{1,i}(x) = \frac{1}{h}[(x - x_{i-1})f(x_i) - (x - x_i)f(x_{i-1})].$$

При этом

$$\int_{x_{i-1}}^{x_i} L_{1,i}(x)dx = \frac{f(x_{i-1}) + f(x_i)}{2}h.$$

Для оценки погрешности r_i формулы (3) будем использовать известное выражение для погрешности интерполирования

$$f(x) - L_{1,i}(x) = \frac{(x - x_{i-1})(x - x_i)}{2}f''(\xi_i(x)).$$

Отсюда

$$r_i = \int_{x_{i-1}}^{x_i}(f(x) - L_{1,i}(x))dx = \frac{1}{2}\int_{x_{i-1}}^{x_i}(x - x_{i-1})(x - x_i)f''(\xi_i(x))dx.$$

Следовательно, после элементарного интегрирования

$$\int_{x_{i-1}}^{x_i}(x - x_{i-1})(x_i - x)dx = -\int_{x_{i-1}}^{x_i}(x - x_{i-1})^2dx + h\int_{x_{i-1}}^{x_i}(x - x_{i-1})dx =$$

$$= -\frac{h^3}{3} + \frac{h^3}{2} = \frac{h^3}{6}$$

получаем оценку

$$|r_i| \le \frac{h^3}{12}M_{2,i}\ .$$

Это значит, что *простейшая формула трапеций имеет третий порядок точности.*

Оценка не улучшаема. Равенство достигается при $f(x) = (x - x_i)^2$.

Составная формула трапеций имеет вид

$$\int_a^b f(x)dx \approx h\sum_{i=1}^{n}\frac{f(x_{i-1}) + f(x_i)}{2} = h[\frac{1}{2}f(x_0) + f(x_1) + \ldots + f(x_{n-1}) + \frac{1}{2}f(x_n)].$$

Оценка погрешности

$$|R| \le \frac{h^2(b-a)}{12}M_2.$$

Составная формула имеет второй порядок точности

3. Формулы парабол (Симпсона)

Для аппроксимации частичного интеграла $\int_{x_{i-1}}^{x_i} f(x)dx$ заменим функцию $f(x)$ параболой, проходящей через точки $(x_{i-1}, f(x_{i-1}))$, $(\overline{x}_i, f(\overline{x}_i))$, $(x_i, f(x_i))$. Это значит, что используется параболическая интерполяция

$$f(x) \approx L_{2,i}(x), \quad x \in [x_{i-1}, x_i],$$

где $L_{2,i}(x)$ – интерполяционный многочлен Лагранжа второй степени

$$L_{2,i}(x) = \frac{2}{h^2}[(x-\overline{x}_i)(x-x_i)f(x_{i-1}) -$$

$$-2(x-x_{i-1})(x-x_i)f(\overline{x}_i) + (x-x_{i-1})(x-\overline{x}_i)f(x_i)].$$

После интегрирования получаем *простейшую квадратурную формулу парабол Симпсона)*

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx \frac{h}{6}(f(x_{i-1}) + 4f(\overline{x}_i) + f(x_i)). \tag{4}$$

Согласно построению данная формула является точной для многочленов второй степени. Нетрудно проверить, что формула (4) является также точной для многочленов третьей степени (достаточно положить в (4) $f(x) = x^3$). Действительно, для интеграла в левой части

$$\int_{x_{i-1}}^{x_i} x^3 dx = \frac{1}{4}(x_i^4 - x_{i-1}^4) = \frac{1}{4}(x_i - x_{i-1})(x_i + x_{i-1})(x_i^2 + x_{i-1}^2) =$$

$$= \frac{h}{4}(x_i^3 + x_i x_{i-1}^2 + x_{i-1}x_i^2 + x_{i-1}^3).$$

С другой стороны, поскольку

$$4\overline{x}_i^3 = 4\frac{(x_i + x_{i-1})^3}{8} = \frac{1}{2}(x_i^3 + 3x_i^2 x_{i-1} + 3x_{i-1}^2 x_i + x_{i-1}^3),$$

то

$$\frac{h}{6}(x_{i-1}^3 + 4\overline{x}_i^3 + x_i^3) = \frac{h}{4}(x_i^3 + x_i^2 x_{i-1} + x_{i-1}^2 x_i + x_{i-1}^3) = \int_{x_{i-1}}^{x_i} x^3 dx.$$

Следовательно, для $f(x) = x^3$ в (4) имеет место точное равенство.

Оценим погрешность формулы (4), используя интерполяционный многочлен Эрмита $H_3(x)$ (полином 3 степени), удовлетворяющий условиям

$$H_3(x_{i-1}) = f(x_{i-1}), \;\; H_3(\overline{x}_i) = f(\overline{x}_i),$$

$$H_3^{'}(\overline{x}_i) = f^{'}(\overline{x}_i), \;\; H_3(x_i) = f(x_i).$$

Как известно, такой многочлен существует и является единственным. При этом погрешность интерполирования $R_i(x) = f(x) - H_3(x)$, $x \in [x_{i-1}, x_i]$ выражается по формуле

$$R_i(x) = \frac{f^{(4)}(\xi_i)}{24}(x - x_{i-1})(x - \overline{x}_i)^2(x - x_i).$$

Поскольку $H_3(x)$ – полином 3-й степени, то формула (4) для него является точной

$$\int_{x_{i-1}}^{x_i} H_3(x)dx = \frac{h}{6}[H_3(x_{i-1}) + 4H_3(\overline{x}_i) + H_3(x_i)].$$

С учетом условий интерполирования получаем

$$\int_{x_{i-1}}^{x_i} H_3(x)dx = \frac{h}{6}[f(x_{i-1}) + 4f(\overline{x}_i) + f(x_i)].$$

Следовательно, погрешность формулы (4) имеет вид

$$r_i = \int_{x_{i-1}}^{x_i} f(x)dx - \int_{x_{i-1}}^{x_i} H_3(x)dx = \int_{x_{i-1}}^{x_i} R_i(x)dx.$$

Отсюда получаем оценку

$$|r_i| \le \frac{M_{4,i}}{24} \int_{x_{i-1}}^{x_i} (x - x_{i-1})(x - \overline{x}_i)^2(x_i - x)dx,$$

где $M_{4,i} = \max_{x\in[x_{i-1},x_i]} |f^{(4)}(x)|$.

Вычисляя интеграл (замена $x - \overline{x}_i = y$), приходим к окончательной оценке погрешности для простейшей формулы

$$|r_i| \le \frac{h^5}{2880} M_{4,i}.$$

Это значит, что *простейшая формула Симпсона имеет пятый порядок точности.*

Составная формула Симпсона имеет вид

$$\int_a^b f(x)dx \approx \frac{h}{6} \sum_{i=1}^{n} (f(x_{i-1}) + 4f(\overline{x}_i) + f(x_i)) =$$

$$= \frac{h}{6}[f(x_0) + f(x_n) + 2(f(x_1) + f(x_2) + \ldots + f(x_{n-1}))+$$

$$+4(f(\overline{x}_1) + f(\overline{x}_2) + \ldots + f(\overline{x}_n))], \quad \overline{x}_i = \frac{x_{i-1} + x_i}{2}, \quad i = \overline{1, n}.$$

Погрешность составной формулы оценивается следующим образом *(четвертый порядок точности)*

$$|R| \le \frac{h^4(b-a)}{2880} M_4, \quad M_4 = \max_{x\in[a,b]} |f^{(4)}(x)|.$$

4. Правило Рунге практической оценки погрешности

Полученные выше оценки погрешности квадратурных формул носят, вообще говоря, теоретический (асимптотический) характер, ибо вычисление (или оценка) величин M_2, M_4, как правило, невозможно. Опишем один способ оценки погрешности, пригодный для практических вычислений.

Рассмотрим задачу приближенного вычисления интеграла

$$I(f) = \int_a^b f(x)dx$$

с заданной погрешностью ε.

Для решения этой задачи будем использовать какую-либо составную формулу порядка точности m с шагом h_1

$$I(f) \approx S_{h_1}(f),$$

где $S_{h_1}(f)$ – квадратурная сумма.

При этом оценка погрешности имеет вид: $|I(f) - S_{h_1}(f)| \leq Ch_1^m$, где C не зависит от h_1 и является неизвестной величиной. Данная оценка позволяет записать приближенное равенство $I(f) - S_{h_1}(f) \approx \overline{C}h_1^m$ с точностью до членов порядка $o(h_1^m)$ с константой $\overline{C}$, не зависящей от h_1.

Повторим расчет по данной квадратурной формуле с шагом $h_2 = \frac{h_1}{2}$. В результате получаем $I(f) - S_{h_2}(f) \approx \overline{C}h_2^m$. Таким образом, имеем два представления

$$I(f) - S_{h_1}(f) \approx \overline{C}h_1^m, \quad I(f) - S_{h_2}(f) \approx \overline{C}h_2^m.$$

Отсюда

$$S_{h_2}(f) - S_{h_1}(f) \approx \overline{C}h_1^m - \overline{C}h_2^m = \overline{C}(2^m - 1)h_2^m,$$

т.е.

$$\overline{C}h_2^m \approx \frac{S_{h_2}(f) - S_{h_1}(f)}{2^m - 1}.$$

Таким образом, приходим к приближенному представлению погрешности

$$I(f) - S_{h_2}(f) \approx \frac{S_{h_2}(f) - S_{h_1}(f)}{2^m - 1},$$

в котором правая часть известна, причем $h_2 = \frac{h_1}{2}$.

Отсюда следует, что требуемую оценку погрешности $|I(f) - S_{h_2}(f)| \leq \varepsilon$ может, вообще говоря, обеспечить неравенство

$$\frac{|S_{h_2}(f) - S_{h_1}(f)|}{2^m - 1} \leq \varepsilon$$

(если неравенство не выполнено, то дробление шага следует продолжить: $h_3 = \frac{h_2}{2}$ и т.д.).

§3. Квадратурная формула Гаусса

Рассмотрим общую квадратурную формулу

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} A_i f(x_i). \tag{1}$$

Считая n фиксированным, выберем параметры A_i, x_i, $i = \overline{1,n}$ так, чтобы обеспечить формуле (1) максимально возможную степень точности. Отметим, что при любом выборе параметров степень точности формулы (1) не может быть больше $(2n-1)$.

Теорема 1. *Для того, чтобы квадратурная формула (1) имела степень точности* $(2n-1)$ *необходимо и достаточно выполнение условий*

$$A_i = \int_a^b \Phi_i(x)dx, \quad i = \overline{1,n}, \tag{2}$$

$$\int_a^b \omega_n(x)p(x)dx = 0, \tag{3}$$

где

$$\Phi_i(x) = \prod_{j=1 j\neq i}^{n} \frac{x - x_j}{x_i - x_j}, \quad \omega_n(x) = \prod_{i=1}^{n}(x - x_i),$$

$p(x)$ *– произвольный многочлен степени не выше* $(n-1)$.

Доказательство. Необходимость. Пусть формула (1) имеет степень точности $(2n-1)$. Поскольку $\Phi_i(x)$ есть многочлен степени $(n-1)$, причем $\Phi_i(x_k) = \begin{cases} 0, & k \neq i \\ 1, & k = i \end{cases}$, то

$$\int_a^b \Phi_i(x)dx = \sum_{k=1}^{n} A_k \Phi_i(x_k) = A_i, \quad i = \overline{1,n}.$$

Получили выражение (2) для квадратурных коэффициентов.

Пусть, далее, $p(x)$ – произвольный многочлен степени не выше $(n-1)$. Положим $f(x) = \omega_n(x)p(x)$. Это многочлен, имеющий степень не выше $(2n-1)$. Следовательно, формула (1) является для него точной. Поскольку $f(x_i) = 0$, $i = \overline{1,n}$, то приходим к условию (3). Необходимость доказана.

Достаточность. Пусть выполнены соотношения (2), (3). Условие (2) означает, что формула (1) является интерполяционной, т.е. имеет степень точности не меньше, чем $(n-1)$.

Пусть теперь $f(x)$ – произвольный многочлен степени не выше $(2n-1)$. Представим его в форме $f(x) = \omega_n(x)p(x) + q(x)$, где степень многочленов $p(x)$, $q(x)$ не выше $(n-1)$ (разделим $f(x)$ на $\omega_n(x)$). Так как $\omega_n(x_i) = 0$, $i = \overline{1,n}$, то $f(x_i) = q(x_i)$. Следовательно,

$$\int_a^b f(x)dx = \int_a^b \omega_n(x)p(x)dx + \int_a^b q(x)dx = \sum_{i=1}^{n} A_i f(x_i).$$

□

Рассмотрим вопрос о существовании многочлена $\omega_n(x)$ (степень n, старший коэффициент 1, корни из $[a,b]$), удовлетворяющего условию (3).

Теорема 2. *Существует единственный многочлен $Q_n(x)$ степени n со старшим коэффициентом* 1, *удовлетворяющий условию*

$$\int_a^b Q_n(x)p(x)dx = 0 \qquad (3')$$

.

Доказательство. Будем искать многочлен $Q_n(x)$ в форме разложения по степеням x, т.е. $Q_n(x) = x^n + a_1x^{n-1} + \cdots + a_n$. Полагая в $(3')$ $p(x) = x^k$, $k = \overline{0,n-1}$, получим линейную систему для определения коэффициентов $a_1, \ldots, a_n$

$$\int_a^b (x^n + a_1x^{n-1} + \ldots + a_n)x^k dx = 0, \quad k = \overline{0,n-1}.$$

Покажем, что данная система имеет единственное решение.

Для этого достаточно убедиться в том, что соответствующая однородная система

$$\int_a^b (a_1x^{n-1} + \cdots + a_n)x^k dx = 0, \quad k = \overline{0,n-1} \qquad (4)$$

имеет только нулевое решение. С этой целью умножим равенство (4) на a_{n-k} и просуммируем по $k = \overline{0,n-1}$. В результате получаем

$$\int_a^b (a_1x^{n-1} + \cdots + a_n)^2 dx = 0.$$

Отсюда следует, что $a_1x^{n-1} + \cdots + a_n = 0$, $x \in [a,b]$, т.е. $a_i = 0$, $i = \overline{1,n}$.
□

Теорема 3. *Корни многочлена $Q_n(x)$ действительны, различны и лежат внутри $[a, b]$.*

Доказательство. Пусть $y_1, \ldots, y_m$ – действительные корни многочлена $Q_n(x)$, которые лежат внутри $[a, b]$ и имеют нечетную кратность. Для доказательства теоремы достаточно показать, что $m = n$. Допустим противное, т.е. $m < n$. Составим многочлен

$$q(x) = (x - y_1) \ldots (x - y_m).$$

Его степень $m < n$, следовательно, согласно условию $(3')$

$$\int_a^b Q_n(x) q(x) dx = 0.$$

С другой стороны, произведение $Q_n(x)q(x) \neq 0$ и сохраняет знак на $[a, b]$, поскольку многочлены $Q_n(x)$, $q(x)$ имеют внутри $[a, b]$ одинаковые точки перемены знака $y_1, \ldots, y_m$. Это значит, что

$$\int_a^b Q_n(x) q(x) dx \neq 0.$$

Полученное противоречие доказывает теорему.

Таким образом, многочлен $Q_n(x)$ совпадает с $\omega_n(x)$, фигурирующим в теореме 1.

Вывод: *для каждого $n = 1, 2, \ldots$ квадратурная формула вида (1) со степенью точности $(2n - 1)$ существует, единственна и определяется условиями (2), (3). Эта формула имеет наивысшую алгебраическую степень точности и носит название квадратурной формулы Гаусса.*

Отметим, что условие (3) фактически эквивалентно следующей системе уравнений для определения квадратурных узлов $x_1, \ldots, x_n$:

$$\int_a^b (x - x_1) \ldots (x - x_n) x^k dx = 0, \quad k = \overline{0, n-1}.$$

Укажем дополнительную характеристику квадратурных узлов формулы Гаусса. С помощью замены переменной осуществим отображение $[a, b] \Leftrightarrow [-1, 1]$ и рассмотрим формулу Гаусса на $[-1, 1]$.

Введем в рассмотрение полиномы Лежандра

$$\mathcal{L}_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n, \;\; n = 1, 2, \ldots, \;\; x \in [-1, 1].$$

Старший коэффициент полинома $\mathcal{L}_n(x)$ равен $a_0 = \frac{(2n)!}{2^n(n!)^2}$. Кроме того, имеет место свойство ортогональности

$$\int_{-1}^{1} \mathcal{L}_n(x)p(x)dx = 0,$$

где $p(x)$ – произвольный многочлен степени не выше $(n-1)$. Отсюда заключаем, что $\omega_n(x) = \frac{1}{a_0}\mathcal{L}_n(x)$. Следовательно, *квадратурные узлы формулы Гаусса для* $[-1,1]$ *являются корнями полинома Лежандра* $\mathcal{L}_n(x)$.

Отметим свойство положительности квадратурных коэффициентов формулы (1).

Действительно, рассмотрим квадраты $\Phi_i^2(x)$ базисных многочленов Лагранжа. Это многочлены степени $(2n-2)$, и формула Гаусса является для них точной

$$\int_a^b \Phi_i^2(x)dx = \sum_{k=1}^{n} A_k\Phi_i^2(x_k) = A_i,$$

т.е. $A_i > 0$, $i = \overline{1,n}$.

Найдем выражение для *погрешности* квадратурной формулы Гаусса. Предположим, что функция $f(x)$ непрерывно-дифференцируема на $[a,b]$ до порядка $2n$ включительно. Пусть $H_{2n-1}(x)$ – полином степени не выше $2n-1$, удовлетворяющий условиям кратного интерполирования

$$H_{2n-1}(x_i) = f(x_i), \quad H'_{2n-1}(x_i) = f'(x_i), \ i = \overline{1,n}.$$

Иными словами, $H_{2n-1}(x)$ – интерполяционный полином Эрмита для функции $f(x)$ по системе узлов $x_1, \ldots, x_n$ кратности 2. Как известно, такой полином существует и является единственным, причем погрешность интерполирования имеет вид

$$f(x) - H_{2n-1}(x) = \frac{f^{(2n)}(\xi)}{(2n)!}\omega_n^2(x), \quad \xi, x \in [a,b].$$

При этом

$$\int_a^b H_{2n-1}(x)dx = \sum_{i=1}^{n} A_i H_{2n-1}(x_i) = \sum_{i=1}^{n} A_i f(x_i),$$

где A_i, x_i – коэффициенты и узлы квадратурной формулы Гаусса. Таким образом, погрешность $I_n(f)$ формулы Гаусса имеет вид

$$I_n(f) = \frac{1}{(2n)!}\int_a^b f^{(2n)}(\xi)\omega_n^2(x)dx.$$

§4. Численное дифференцирование функций

Пусть функция $f(x)$ определена на $[a,b]$ вместе со своими производными до $(n+1)$ порядка включительно. Вычислим её значения в точках $x_0, x_1, \ldots, x_n$ из $[a,b]$. Задача численного дифференцирования состоит в приближенном отыскании производных $f^{(k)}(x)$, $k = \overline{1,n}$ в некоторой точке $x \in [a,b]$ по известным значениям $f(x_i)$, $i = \overline{0,n}$.

Основной метод решения этой задачи связан с применением аппарата интерполирования. Представим функцию $f(x)$ по таблице $\{x_i, f(x_i)\}$ с помощью некоторой интерполяционной формулы

$$f(x) = L_n(x) + R_n(x), \quad x \in [a,b].$$

Здесь $L_n(x)$ – интерполяционный многочлен, $R_n(x)$ – погрешность интерполирования.

В результате последовательного дифференцирования этого соотношения получаем

$$f^{(k)}(x) = L_n^{(k)}(x) + R_n^{(k)}(x), \quad k = \overline{1,n}.$$

Величину $L_n^{(k)}(x)$ примем за приближенное значение k-ой производной в точке x : $f^{(k)}(x) \approx L_n^{(k)}(x)$. Это формула численного дифференцирования в общем виде. При этом $R_n^{(k)}(x)$ – погрешность аппроксимации.

Отметим, что формулы численного дифференцирования, вообще говоря, не являются точными даже в узлах интерполирования x_i, т.е. $R_n^{(k)}(x_i) \neq 0$.

1. Простейшие формулы численного дифференцирования

Введем на $[a,b]$ равномерную сетку $x_i = a + ih$, $i = \overline{0,n}$ с шагом $h = \frac{b-a}{n}$. Положим $f_i = f(x_i)$ и построим разностные аппроксимации для производных $f'(x_i)$, $f''(x_i)$ на основе линейной и параболической интерполяции функции $f(x)$ на частичных отрезках $[x_{i-1}, x_i]$, $[x_i, x_{i+1}]$, $[x_{i-1}, x_{i+1}]$. При этом функция $f(x)$ считается на $[a,b]$ достаточно гладкой.

Проведем линейную интерполяцию функции $f(x)$ на $[x_{i-1}, x_i]$ по значениям f_{i-1}, f_i. Соответствующий полином Лагранжа имеет вид

$$L_{1,i}(x) = \frac{1}{h}[f_i(x - x_{i-1}) - f_{i-1}(x - x_i)].$$

Отсюда

$$f'(x_i) \approx L'_{1,i} = \frac{f_i - f_{i-1}}{h}, \quad i = \overline{1,n}.$$

Это *левая разностная производная функции f в точке x_i.*

Аналогичным образом $([x_i, x_{i+1}], f_i, f_{i+1})$ получаем выражение для *правой разностной производной*

$$f'(x_i) \approx \frac{f_{i+1} - f_i}{h}, \quad i = \overline{0,n-1}.$$

Найдем выражения для погрешности данных аппроксимаций. Рассмотрим левую производную. На основании формулы Тейлора

$$f_i - f_{i-1} = f(x_i) - f(x_i - h) = f(x_i) - f(x_i) + hf'(x_i) - \frac{h^2}{2}f''(\xi_i), \;\; \xi_i \in [x_{i-1}, x_i].$$

Следовательно,

$$f'(x_i) = \frac{f_i - f_{i-1}}{h} + \frac{h}{2}f''(\xi_i)$$

(аппроксимация первого порядка относительно h).

Таким образом, *левая разностная производная имеет первый порядок аппроксимации.*

Аналогично, для правой производной

$$f'(x_i) = \frac{f_{i+1} - f_i}{h} - \frac{h}{2}f''(\xi_i), \;\; \xi_i \in [x_i, x_{i+1}].$$

Рассмотрим, далее, отрезок $[x_{i-1}, x_{i+1}]$ и проведем параболическую интерполяцию функции $f(x)$ по значениям f_{i-1}, f_i, f_{i+1}. Соответствующий интерполяционный многочлен имеет вид

$$L_{2,i}(x) = \frac{1}{2h^2}[f_{i-1}(x - x_i)(x - x_{i+1}) - 2f_i(x - x_i)(x - x_{i+1}) +$$

$$+ f_{i+1}(x - x_i)(x - x_{i+1})].$$

В результате для $i = \overline{1, n-1}$

$$f'(x_i) \approx L'_{2,i}(x_i) = \frac{f_{i+1} - f_{i-1}}{2h} - \textit{ центральная разностная производная,}$$

$$f''(x_i) \approx L''_{2,i} = \frac{f_{i-1} - 2f_i + f_{i+1}}{h^2} - \textit{ вторая разностная производная.}$$

Изучим вопрос о *погрешности* данных формул.

Рассмотрим центральную производную. На основании формулы Тейлора

$$f_{i+1} - f_{i-1} = f(x_i + h) - f(x_i - h) = 2hf'(x_i) + f^{(3)}(\xi_i^{(1)})\frac{h^3}{6} + f^{(3)}(\xi_i^{(2)})\frac{h^3}{6},$$

$$\xi_i^{(1)} \in [x_i, x_{i+1}], \;\; \xi_i^{(2)} \in [x_{i-1}, x_i].$$

Отметим, что согласно теореме о промежуточных значениях непрерывной функции

$$\frac{f^{(3)}(\xi_i^{(1)}) + f^{(3)}(\xi_i^{(2)})}{2} = f^{(3)}(\xi_i), \;\; \xi_i \in [x_{i-1}, x_{i+1}].$$

Следовательно,

$$f'(x_i) = \frac{f_{i+1} - f_{i-1}}{2h} - \frac{h^2}{6}f^{(3)}(\xi_i),$$

т.е. *центральная разностная производная имеет второй порядок аппроксимации.*

Погрешность второй разностной производной, как нетрудно проверить, определяется выражением

$$f''(x_i) = \frac{f_{i-1} - 2f_i + f_{i+1}}{h^2} - \frac{h^2}{12}f^{(4)}(\xi_i) \text{ (второй порядок аппроксимации).}$$

2. Некорректность операции численного дифференцирования

Как правило, значения $f_i = f(x_i),\ i = \overline{0,n}$ вычисляются неточно, с некоторой погрешностью. Оказывается, что погрешность, возникающая при вычислении разностных производных намного превосходит погрешность вычисления значений функции. Поэтому операцию вычисления разностных отношений (операцию численного дифференцирования) называют некорректной.

Поясним сказанное на примере вычисления левой разностной производной

$$f(x_{i-1}; x_i) = \frac{f_i - f_{i-1}}{h}.$$

Пусть вместо точных значений f_i, f_{i-1} вычисляются приближенные значения $\tilde{f}_i = f_i + \delta_i,\ \tilde{f}_{i-1} = f_{i-1} + \delta_{i-1}$. Тогда вместо разностной производной будет вычислена величина

$$\frac{\tilde{f}_i - \tilde{f}_{i-1}}{h} = f(x_{i-1}; x_i) + \frac{\delta_i - \delta_{i-1}}{h}.$$

Следовательно, погрешность вычисления равна $\delta(i) = \frac{\delta_i - \delta_{i-1}}{h}$.

Пусть $|\delta_i| \le \delta$, $|\delta_{i-1}| \le \delta$ (известна оценка погрешности приближенных значений). Тогда $|\delta(i)| \le \frac{2\delta}{h}$. Если δ не зависит от h, то погрешность $\delta(i)$ неограниченно возрастает при $h \to 0$.

Выход из этой ситуации состоит в том, что величины δ, h нужно выбирать согласованно. К примеру, если h задано, то необходимо задать $\delta = Ch^2$. При этом $|\delta(i)| \le 2Ch$, т.е. при $h \to 0$, $|\delta(i)| \to 0$.

С другой стороны, если зафиксирована погрешность δ, то шаг сетки h следует выбрать по правилу: $h = \sqrt{\delta}$. Тогда $|\delta(i)| \le 2\sqrt{\delta}$, т.е. при $\delta \to 0$, $|\delta(i)| \to 0$.

3. Численные методы решения дифференциальных уравнений

Глава 1. Обыкновенные дифференциальные уравнения

§1. Введение

1. Задача Коши. Численный подход

Как известно, задача Коши заключается в отыскании решения $y(x)$, $x_0 \le x \le x_0 + a$ обыкновенного дифференциального уравнения $y' = f(x, y)$, удовлетворяющего заданному начальному условию $y(x_0) = y_0$. Будем считать, как минимум, что функция $f(x, y)$ определена и непрерывна в некотором прямоугольнике

$$\mathcal{D} = \{x, y : x_0 \le x \le x_0 + a, |y - y_0| \le b\}$$

и удовлетворяет в $\mathcal{D}$ условию Липшица по переменной y :

$$|f(x, y + \Delta y) - f(x, y)| \le L|\Delta y|, \;\; (x, y) \in \mathcal{D}, \;\; (x, y + \Delta y) \in \mathcal{D}.$$

В этих условиях, на основании известной теоремы, решение $y(x)$ задача Коши существует и единственно по крайней мере в некоторой окрестности $(x_0, x_0 + d)$ точки x_0. Будем предполагать, что это решение можно однозначно продолжить на весь $[x_0, x_0 + a]$.

Численные методы решения задачи Коши характеризуются тем, что на $[x_0, x_0 + a]$ вводится сетка $\{x_1, \ldots, x_n\}$ *узлов интегрирования*. Для типичного случая равномерной сетки имеем $x_i = x_0 + ih, i = \overline{1, n}$, при этом величина $h = x_{i+1} - x_i$ – *шаг интегрирования*. Далее, последовательно отыскиваются приближенные значения y_i точного решения $y(x)$ в узлах интегрирования: $y_i \approx y(x_i), i = \overline{1, n}$. В результате получается таблица $\{x_i, y_i\}$ (сеточная функция) приближенных значений искомого решения $y(x)$ в узлах интегрирования (функция $y(x)$ аппроксимируется последовательностью точек).

Величина $r_i = y(x_i) - y_i, \;\; i = \overline{1, n}$ характеризует погрешность метода в узле x_i. Понятно, что $r_0 = 0$. Погрешность r_i оценивают в зависимости от шага h. Эта оценка определяет порядок погрешности метода (порядок точности). При этом используются локальная и глобальная оценки погрешности:

1) $r_i = 0$, $|r_{i+1}| \le C_i h^s$ – *локальная* (шаговая) *оценка погрешности*, s – *порядок погрешности на шаге*;

2) $R = \max_{1 \le i \le n} |r_i|$, $R \le Ch^q$ – *глобальная оценка погрешности*, q – *порядок точности метода*.

2. Простейшие методы численного решения задачи Коши

Рассмотрим задачу

$$y' = f(x, y), \;\; y(x_0) = y_0, \;\; x_0 \le x \le x_0 + a. \tag{1}$$

Введем сетку узлов интегрирования $x_i = x_0 + ih, i = \overline{1, n}$ с шагом $h > 0$. Рассмотрим уравнение $y'(x) = f(x, y(x))$ вдоль решения $y(x)$ на частичном отрезке $[x_i, x_{i+1}]$. После интегрирования по $x \in [x_i, x_{i+1}]$ получаем

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(x, y(x))dx, \;\; i = \overline{0, n-1}. \tag{2}$$

Представим интеграл по формуле левых прямоугольников

$$\int_{x_i}^{x_{i+1}} f(x, y(x))dx = hf(x_i, y(x_i)) + O(h^2).$$

Отбрасывая остаток и переходя к приближенным значениям y_i, из (2) получаем следующую расчетную формулу

$$y_{i+1} = y_i + hf(x_i, y_i), \;\; i = \overline{0, n-1}. \tag{3}$$

Процедура (3) носит название *метода Эйлера*.

Рассмотрим другие подходы к построению метода. С помощью представления по формуле Тейлора получаем

$$y(x_{i+1}) = y(x_i + h) = y(x_i) + y'(x_i)h + O(h^2) =$$
$$= y(x_i) + hf(x_i, y(x_i)) + O(h^2).$$

Отбрасывая остаточный член разложения, приходим к процедуре Эйлера (3).

Укажем ещё один способ получения формулы (3). Запишем уравнение (1) в узле x_i

$$y'(x_i) = f(x_i, y(x_i)).$$

Воспользуемся формулой численного дифференцирования (правая разностная производная)

$$y'(x_i) = \frac{y(x_{i+1}) - y(x_i)}{h} + O(h).$$

Это сразу приводит к методу Эйлера.

Получим другой вариант метода. Применяя для интеграла из (2) формулу правых прямоугольников, получаем *неявный метод Эйлера*:

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}), \quad i = \overline{0, n-1}$$

(для нахождения y_{i+1} необходимо решить уравнение, y_{i+1} задается формулой неявно).

Формула (3) определяет *явный метод Эйлера*.

Построим некоторые модификации метода Эйлера. Возьмем за основу выражение (2) и представим интеграл по формуле трапеций

$$\int_{x_i}^{x_{i+1}} f(x, y(x))dx = \frac{h}{2}[f(x_i, y(x_i)) + f(x_{i+1}, y(x_{i+1}))] + O(h^3).$$

В результате получаем расчетную формулу неявного типа (метод трапеций)

$$y_{i+1} = y_i + \frac{h}{2}[f(x_i, y_i) + f(x_{i+1}, y_{i+1})], \quad i = \overline{0, n-1}.$$

Для упрощения этой процедуры значение y_{i+1} в правой части можно считать по явной формуле Эйлера. Это приводит к следующей расчетной схеме (*метод Эйлера с пересчетом*, метод Хьюна)

$$\begin{gathered} \tilde{y}_i = y_i + hf(x_i, y_i), \\ y_{i+1} = y_i + \frac{h}{2}[f(x_i, y_i) + f(x_{i+1}, \tilde{y}_i)], \quad i = \overline{0, n-1}. \end{gathered} \tag{4}$$

Построим вторую модификацию метода Эйлера. С этой целью интеграл в правой части (2) заменим по формуле средних прямоугольников

$$\int_{x_i}^{x_{i+1}} f(x, y(x))dx \approx hf(x_i + \frac{h}{2}, y(x_i + \frac{h}{2})).$$

Пусть $\overline{x}_i = x_i + \frac{h}{2}$. Значение $y(\overline{x}_i)$ приближенно вычислим по явному методу Эйлера

$$y(\overline{x}_i) \approx \overline{y}_i = y_i + \frac{h}{2} f(x_i, y_i).$$

В результате, вторая модифицированная схема имеет вид (*метод средней точки*)

$$\overline{y}_i = y_i + \frac{h}{2} f(x_i, y_i), \quad y_{i+1} = y_i + hf(\overline{x}_i, \overline{y}_i). \tag{5}$$

Получим еще одну модификацию метода Эйлера. Проинтегрируем уравнение $y'(x) = f(x, y(x))$ по $x \in [x_{i-1}, x_{i+1}]$. Тогда

$$y(x_{i+1}) = y(x_{i-1}) + \int_{x_{i-1}}^{x_{i+1}} f(x, y(x))dx.$$

Для приближенного вычисления интеграла применим формулу средних прямоугольников

$$\int_{x_i-h}^{x_i+h} f(x, y(x))dx \approx 2hf(x_i, y(x_i)).$$

В результате получаем расчетную формулу (*уточненный метод Эйлера*)

$$y_{i+1} = y_{i-1} + 2hf(x_i, y_i), \quad i = \overline{1, n-1}. \tag{6}$$

Этот же метод может быть получен на основе равенства $y'(x_i) = f(x_i, y(x_i))$ с помощью формулы численного дифференцирования (центральная разностная производная)

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

Формула (6) определяет явный двушаговый метод: для подсчета y_{i+1} необходимо знать два предыдущих приближения y_{i-1}, y_i.

§2. Методы Рунге - Кутта

Общая схема данного класса методов имеет вид

$$y_{i+1} = y_i + h\sum_{j=1}^{m} p_j k_j, \quad i = \overline{0, n-1}, \tag{7}$$

$$k_1 = f(x_i, y_i),$$

$$k_2 = f(x_i + \alpha_2 h, y_i + \beta_{21} h k_1),$$

$$\dots$$

$$k_m = f(x_i + \alpha_m h, y_i + \beta_{m1} h k_1 + \ldots + \beta_{m,m-1} h k_{m-1}).$$

Здесь $p_1, \ldots, p_m$, $\alpha_2, \ldots \alpha_m$, β_{eq}, $0 < q < l \leq m$ – параметры метода, подлежащие выбору. Формула (7) определяет m – *этапный метод Рунге-Кутта.*

Предположим, что в узле x_i погрешность метода равна нулю: $y(x_i) = y_i$. Исследуем локальную погрешность в узле x_{i+1}

$$r(h) = y(x_{i+1}) - y_{i+1} = y(x_i + h) - y_i - h\sum_{j=1}^{m} p_j k_j.$$

Будем считать, что функция $f(x, y)$ является достаточное число раз дифференцируемой по своим аргументам, так что допустимо следующее разложение погрешности по формуле Тейлора

$$r(h) = \sum_{k=0}^{s} \frac{r^{(k)}(0)}{k!} h^k + \frac{r^{(s+1)}(\theta h)}{(s+1)!} h^{s+1}, \quad 0 < \theta < 1.$$

Пусть параметры метода (7) выбраны так, чтобы $r(0) = r'(0) = \ldots = r^{(s)}(0) = 0$. Тогда $r(h) \sim O(h^{s+1})$, и число s называют *порядком точности метода.*

Основная цель выбора параметров для заданного m состоит в максимально возможном повышении порядка погрешности при сохранении относительной простоты расчетных формул.

Выделим из (7) некоторые конкретные формулы типа Рунге-Кутта, соответствующие специальным выборам параметров.

1) Пусть $m = 1$ (*одноэтапный метод*).

В этом случае

$$r(h) = y(x_i + h) - y_i - hp_1 f(x_i, y_i).$$

Отсюда

$$r(0) = y(x_i) - y_i = 0,$$
$$r'(0) = y'(x_i) - p_1 f(x_i, y_i) = (1 - p_1) f(x_i, y_i),$$
$$r''(h) = y''(x_i + h).$$

Полагая $p_1 = 1$, получаем $r'(0) = 0$. Этому значению p_1 соответствует метод Эйлера

$$y_{i+1} = y_i + hf(x_i, y_i), \quad i = 0, 1, \ldots$$

Погрешность метода на шаге (локальная погрешность) имеет вид

$$r(h) = \frac{1}{2} y''(x_i + \theta h) h^2 \sim O(h^2).$$

Таким образом, метод Эйлера является простейшим случаем методов Рунге- Кутта и имеет *первый порядок точности*

2) Рассмотрим случай $m = 2$ (*двухэтапные методы*).

Тогда

$$r(h) = y(x_i + h) - y_i - h(p_1 f(x_i, y_i) + p_2 f(\overline{x}_i, \overline{y}_i)),$$
$$\overline{x}_i = x_i + \alpha_2 h, \;\; \overline{y_i} = y_i + \beta_{21} h f(x_i, y_i).$$

Для сокращения записи будем обозначать

$$f = f(x_i, y_i), \;\; \overline{f} = f(\overline{x}_i, \overline{y}_i),$$
$$\overline{f}_x = \frac{\partial f}{\partial x}(\overline{x}_i, \overline{y}_i), \;\; \overline{f}_y = \frac{\partial f}{\partial y}(\overline{x}_i, \overline{y}_i).$$

Отметим, что по правилу дифференцирования сложной функции

$$\frac{d}{dh}\overline{f} = \overline{f}_x \alpha_2 + \overline{f}_y f \beta_{21}.$$

Понятно, что $r(0) = 0$. Подсчитаем производные

$$r'(h) = y'(x_i + h) - p_1 f - p_2 \overline{f} - h p_2 [\overline{f}_x \alpha_2 + \overline{f}_y f \beta_{21}],$$
$$r''(h) = y''(x_i + h) - 2p_2 [\overline{f}_x \alpha_2 + \overline{f}_y f \beta_{21}] - p_2 h \frac{d^2}{dh^2}\overline{f}.$$

Согласно уравнению (1)

$$y'(x) = f(x, y(x)), \;\; y''(x) = f_x(x, y(x)) + f_y(x, y(x)) f(x, y(x)).$$

Следовательно

$$r'(0) = (1 - p_1 - p_2) f,$$
$$r''(0) = (1 - 2p_2\alpha_2) f_x + (1 - 2p_2\beta_{21}) f_y f.$$

Чтобы выполнить равенства $r'(0) = r''(0) = 0$ для любой функции f, положим

$$1 - p_1 - p_2 = 0, \;\; 1 - 2p_2\alpha_2 = 0, \;\; 1 - 2p_2\beta_{21} = 0.$$

Получили три уравнения относительно четырех параметров. Произвольно задавая один из них, можно получить различные варианты методов Рунге-Кутта с локальной погрешностью $r(h) = O(h^3)$ (*методы второго порядка точности*).

Положим, например, $p_1 = \frac{1}{2}$. Тогда $p_2 = \frac{1}{2}$, $\alpha_2 = 1$, $\beta_{21} = 1$. Формула (7) принимает вид

$$y_{i+1} = y_i + \frac{h}{2}[f(x_i, y_i) + f(x_i + h, y_i + h f(x_i, y_i))], \;\; i = 0, 1, \ldots$$

Это метод Эйлера с пересчетом (формула (4)).

Пусть теперь $p_1 = 0$. Тогда $p_2 = 1$, $\alpha_2 = \frac{1}{2}$, $\beta_{21} = \frac{1}{2}$. Из (7) получаем процедуру

$$y_{i+1} = y_i + hf(x_i + \frac{h}{2}, y_i + \frac{h}{2}f(x_i, y_i)), \quad i = 0, 1, \ldots$$

Получили вторую модификацию метода Эйлера (метод средней точки(5)).

Таким образом, *модифицированные схемы (4), (5) являются методами Рунге- Кутта с локальной погрешностью порядка h^3 (второго порядка точности).*

Отметим, что увеличить порядок погрешности в случае $m = 2$ невозможно.

3) Пусть $m = 4$ (*четырехэтапный метод*).

В этом случае за счет выбора параметров можно обеспечить локальную погрешность $r(h) \sim O(h^5)$ (четвертый порядок точности).

Приведем наиболее употребительную совокупность формул четвертого порядка точности

$$y_{i+1} = y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

$$k_1 = f(x_i, y_i),$$
$$k_2 = f(x_i + \frac{h}{2}, y_i + \frac{h}{2}k_1),$$
$$k_3 = f(x_i + \frac{h}{2}, y_i + \frac{h}{2}k_2),$$
$$k_4 = f(x_i + h, y_i + hk_3).$$

Замечание. В принципе методы Рунге-Кутта описываются следующей структурой

$$y_{i+1} = y_i + h\varphi(x_i, y_i, h), \quad i = 0, 1, \ldots$$

Эта формула определяет класс явных одношаговых методов.

§3. Метод Эйлера. Анализ глобальной погрешности

Рассмотрим задачу Коши

$$y' = f(x, y), \quad y(x_0) = y_0 \tag{1}$$

на отрезке $[x_0, x_0 + a]$. Предположим, что функция $f(x, y)$ непрерывна в области

$$\mathcal{D} = \{(x, y) : x_0 \le x \le x_0 + a, \quad |y - y_0| \le b\}$$

вместе с производными f_x, f_y. Это значит, что

1) выполняется условие Липшица для функции $f(x, y)$ по переменной y

$$|f(x, y + \Delta y) - f(x, y)| \le L|\Delta y| \quad (L = \max_{x,y \in \mathcal{D}} |f_y(x, y)|),$$

2) решение $y(x), \quad x \in [x_0, x_0 + a]$ имеет непрерывную вторую производную

$$y''(x) = f_x(x, y(x)) + f_y(x, y(x))f(x, y(x))$$

[результат дифференцирования тождества $y'(x) = f(x, y(x))$].

Обозначим $M_2 = \max_{x \in [x_0, x_0 + a]} |y''(x)|$.

Для численного решения задачи (1) введем сетку узлов интегрирования $x_i = x_0 + ih$, $i = \overline{1, n}$, $nh = a$ и будем использовать метод Эйлера

$$y_{i+1} = y_i + hf(x_i, y_i), \quad i = \overline{0, n-1}. \tag{2}$$

Величина $r_i = y(x_i) - y_i$ характеризует локальную погрешность метода в узле $x_i, i = \overline{0, n}$. Понятно, что $r_0 = 0$. Если $r_i = 0$, то $r_{i+1} = O(h^2)$ (локальная погрешность имеет порядок h^2).

Введем глобальную погрешность метода на сетке $R = \max_{1 \le i \le n} |r_i|$ и найдем оценку R через h (выясним порядок точности метода Эйлера).

Рассмотрим погрешность метода (2) в узле x_{i+1}

$$r_{i+1} = y(x_i + h) - y_{i+1} = y(x_i) + hy'(x_i) + \frac{h^2}{2} y''(x_i + \xi h) - y_i - hf(x_i, y_i) =$$

$$= r_i + h[f(x_i, y(x_i)) - f(x_i, y_i)] + \frac{h^2}{2} y''(x_i + \xi h), \quad \xi \in [0, 1].$$

Переходя к оценке по модулю с учетом условий 1), 2), получаем

$$|r_{i+1}| \le |r_i| + hL|r_i| + \frac{h^2}{2} M_2 = (1 + hL)|r_i| + \frac{1}{2} h^2 M_2, \quad i = \overline{0, n-1}.$$

Введем обозначения: $q = 1 + hL$, $\delta = \frac{1}{2}h^2 M_2$. Тогда

$$|r_{i+1}| \le q|r_i| + \delta, \quad i = \overline{0, n-1}, \quad r_0 = 0.$$

Отсюда

$$|r_{i+1}| \le q^2|r_{i-1}| + q\delta + \delta \le$$
$$\le q^3|r_{i-2}| + q^2\delta + q\delta + \delta \le \ldots \le$$
$$\le q^{i+1}|r_0| + q^i\delta + q^{i-1}\delta + \ldots + q\delta + \delta =$$
$$= (1 + q + \ldots + q^i)\delta \le (1 + q + \ldots + q^{n-1})\delta \le nq^n\delta \quad (q^i \le q^n, \quad i = \overline{0, n-1}).$$

Следовательно, имеет место оценка

$$R = \max_{1 \le i \le n} |r_i| \le nq^n\delta.$$

Поскольку $n = \frac{a}{h}$, то в исходных обозначениях

$$R \le \frac{a}{h}(1 + hL)^n \frac{h^2}{2} M_2 = \frac{a}{2} M_2 (1 + hL)^n h.$$

Далее используем известное неравенство $1 + x \le e^x$, $x > 0$. В нашем случае

$$1 + hL \le e^{hL} \Rightarrow (1 + hL)^n \le e^{hLn} = e^{aL}.$$

В результате получаем итоговую оценку глобальной погрешности метода Эйлера

$$R \le Kh, \quad K = \frac{1}{2} a M_2 e^{aL}.$$

Это значит, что *метод Эйлера имеет первый порядок точности.*

§4. Методы Адамса

Задача:

$$y' = f(x, y), \quad y(x_0) = y_0. \tag{1}$$

Точное решение: $y(x), \quad x \in [x_0, x_0 + a]$.

Сетка узлов интегрирования: $x_i = x_0 + ih, \ i = \overline{1, n}, \ hn = a$.

Приближенное решение: $y_i \approx y(x_i), \quad i = \overline{1, n}$.

Обозначение: $f_i = f(x_i, y_i), \quad i = \overline{1, n}$.

Общая схема методов Адамса имеет вид

$$y_i = y_{i-1} + h\sum_{j=0}^{m} b_j f_{i-j}, \quad i = \overline{m,n}, \tag{2}$$

где $b_0, \ldots, b_m$ - параметры метода.

Формула (2) определяет m- *шаговый метод Адамса*, $m = 1, 2, \ldots$.

Расчет по формуле (2) начинается с $i = m$

$$y_m = y_{m-1} + h(b_0 f_m + b_1 f_{m-1} + \ldots + b_m f_0).$$

Следовательно, для подсчета y_m необходимо иметь m начальных значений $y_0, y_1, \ldots, y_{m-1}$. Значение y_0 задано из (1). Величины $y_1, \ldots, y_{m-1}$ необходимо вычислить с помощью другого метода (например, метода Рунге-Кутта).

Таким образом, в процессе счета по формуле (2) используются значения функции $f(x, y)$ только в узлах $(x_i, y_i), i = 0, 1, \ldots$ При этом для подсчета очередного приближенного значения y_i используются m предыдущих значений $y_{i-1}, y_{i-2}, \ldots, y_{i-m}$.

Если $b_0 = 0$, то формула (2) определяет *явный метод Адамса* (искомое значение y_i явно выражается через m предыдущих).

В случае $b_0 \neq 0$ метод Адамса (2) называется *неявным* (для нахождения y_i необходимо решать уравнение).

Процедуры вида (2) относятся к классу разностных многошаговых методов. Формула (2) задает разностное уравнение относительно значений y_i, $i = \overline{m,n}$ (разностную схему).

Величина

$$\Delta_i(h) = \frac{1}{h}[y(x_i) - y(x_{i-1}) - h\sum_{j=0}^{m} b_j f(x_{i-j}, y(x_{i-j}))]$$

называется *погрешностью аппроксимации* дифференциального уравнения (1) разностной схемой (2) или невязкой метода (2) в узле x_i, $i = m, m+1, \ldots$.

Невязка $\Delta_i(h)$ получается в результате подстановки точного решения $y(x)$ в формулу (2).

Величина $\Delta(h) = \max_{i=\overline{m,n}} |\Delta_i(h)|$ характеризует глобальную погрешность аппроксимации или глобальную невязку метода (2).

Если $\Delta(h) = O(h^s)$, то число s называют *порядком аппроксимации* метода (2).

Отметим, как известный факт [2, с. 223], что *порядок аппроксимации метода совпадает с порядком его точности*:

$$\Delta(h) \sim O(h^s) \Leftrightarrow R \sim O(h^s), \quad R = \max_{m \le i \le n} |r_i|.$$

Выделим основные подходы к построению методов Адамса для заданного m :

a) *метод неопределенных коэффициентов* - параметры $b_0, \ldots, b_m$ находятся с целью обеспечить методу (2) определенный (максимально возможный) порядок аппроксимации;

b) *интерполяционный метод* - функция $f(x, y(x))$ заменяется интерполяционным многочленом $L_m(x)$ по таблице $(x_k, f_k),\ k = i, i-1, \ldots, i-m$. Далее используется интегральное представление

$$y_i = y_{i-1} + \int_{x_{i-1}}^{x_i} L_m(x)dx, \quad i = m, m+1 \ldots,$$

где $L_m(x) = \sum_{j=0}^{m} \Phi_j(x) f_{i-j}$.

Для построения явных методов используется интерполяционный полином $L_{m-1}(x)$ по таблице $(x_k, f_k),\ k = i-1, i-2, \ldots, i-m$.

Построим конкретные варианты методов Адамса.

1) $m = 1$ (*одношаговые методы*).

Расчетная формула: $y_i = y_{i-1} + h(b_0 f_i + b_1 f_{i-1}), \quad i = 1, 2, \ldots$

Погрешность аппроксимации представим в виде

$$\Delta_i(h) = \frac{1}{h}[y(x_i) - y(x_i - h) - h(b_0 y'(x_i) + b_1 y'(x_i - h))].$$

Далее используем разложения

$$y(x_i - h) = y(x_i) - hy'(x_i) + \frac{h^2}{2}y''(x_i) - \frac{h^3}{6}y^{(3)}(x_i - \theta h),$$

$$y'(x_i - h) = y'(x_i) - hy''(x_i) + \frac{h^2}{2}y^{(3)}(x_i - \theta h).$$

В результате

$$h\Delta_i(h) = h(1 - b_0 - b_1)y'(x_i) + h^2(-\frac{1}{2} + b_1)y''(x_i) + O_i(h^3). \qquad (3)$$

Определим условия на параметры

$$1 - b_0 - b_1 = 0, \quad -\frac{1}{2} + b_1 = 0 \Rightarrow b_1 = \frac{1}{2}, \quad b_0 = \frac{1}{2}.$$

Тогда погрешность аппроксимации $\Delta_i(h) = O_i(h^2) \Rightarrow \Delta(h) \sim O(h^2)$ (второй порядок аппроксимации, второй порядок точности).

В результате получаем *неявный одношаговый метод Адамса второго порядка*

$$y_i = y_{i-1} + \frac{h}{2}(f_i + f_{i-1}), \quad i = 1, 2, \ldots$$

Он совпадает с методом трапеций.

Для построения явного метода положим в (3) $b_0 = 0$. Тогда $b_1 = 1$ (обращаем в нуль коэффициент при h). При этом $\Delta_i(h) = O_i(h) \Rightarrow \Delta(h) \sim O(h)$ (первый порядок аппроксимации, первый порядок точности).

В результате получаем *явный одношаговый метод Адамса первого порядка*

$$y_i = y_{i-1} + hf_{i-1}, \quad i = 1, 2, \ldots.$$

Он совпадает с явным методом Эйлера.

Метод прогноза и коррекции является комбинацией явного и неявного методов Адамса:

$\tilde{y}_i = y_{i-1} + hf_{i-1}$ – прогноз искомого значения y_i,

$y_i = y_{i-1} + \frac{h}{2}(f(x_i, \tilde{y}_i) + f_{i-1})$ – коррекция значения $\tilde{y}_i$.

Получили метод Эйлера с пересчетом.

2) $m = 2$ (*двушаговые методы*).

Расчетная формула:

$$y_i = y_{i-1} + h(b_0 f_i + b_1 f_{i-1} + b_2 f_{i-2}), \quad i = 2, 3, \ldots$$

Погрешность аппроксимации представим в виде

$$\Delta_i(h) = \frac{1}{h}[y(x_i) - y(x_i - h) - h(b_0 y'(x_i) + b_1 y'(x_i - h) + b_2 y'(x_i - 2h))].$$

Тейлоровские разложения:

$$y(x_i - h) = y(x_i) - hy'(x_i) + \frac{h^2}{2}y''(x_i) - \frac{h^3}{6}y^{(3)}(x_i) + \frac{h^4}{12}y^{(4)}(x_i - \theta_1 h),$$

$$y'(x_i - h) = y'(x_i) - hy''(x_i) + \frac{h^2}{2}y^{(3)}(x_i) - \frac{h^3}{6}y^{(4)}(x_i - \theta_2 h),$$

$$y'(x_i - 2h) = y'(x_i) - 2hy''(x_i) + \frac{(2h)^2}{2}y^{(3)}(x_i) - \frac{(2h)^3}{6}y^{(4)}(x_i - \theta_3 h).$$

В результате

$$h\Delta_i(h) = h(1 - b_0 - b_1 - b_2)y'(x_i) + h^2(-\frac{1}{2} + b_1 + 2b_2)y''(x_i) +$$

$$+h^3(\frac{1}{6} - \frac{1}{2}b_1 - 2b_2)y^{(3)}(x_i) + O_i(h^4).$$

Условия на параметры

$$1 - b_0 - b_1 - b_2 = 0, \;\; -\frac{1}{2} + b_1 + 2b_2 = 0, \;\; \frac{1}{6} - \frac{1}{2}b_1 - 2b_2 = 0 \;\Rightarrow$$

$$b_0 = \frac{5}{12}, \;\; b_1 = \frac{8}{12}, \;\; b_2 = -\frac{1}{12}.$$

Погрешность аппроксимации $\Delta_i(h) = O_i(h^3) \Rightarrow \Delta(h) \sim O(h^3)$.

Неявный двушаговый метод Адамса третьего порядка

$$y_i = y_{i-1} + \frac{h}{12}(5f_i + 8f_{i-1} - f_{i-2}), \;\; i = 2, 3, \ldots$$

Явный двушаговый метод Адамса

$$b_0 = 0, \;\; b_1 + b_2 = 1, \;\; b_1 + 2b_2 = \frac{1}{2} \Rightarrow \;\; b_1 = \frac{3}{2}, \;\; b_2 = -\frac{1}{2},$$

$$y_i = y_{i-1} + \frac{h}{2}(3f_{i-1} - f_{i-2}), \;\; i = 2, 3, \ldots$$

Погрешность аппроксимации: $\Delta_i(h) \sim O_i(h^2)$ (второй порядок).

§5. Линейная многоточечная задача для системы уравнений

1. Постановка задачи

Пусть на $[a, b]$ задана линейная система обыкновенных дифференциальных уравнений

$$y' = A(x)y + f(x). \tag{1}$$

Здесь $y = y(x)$ – n-мерная вектор-функция (искомое решение), $A(x)$ – $(n \times n)$ матричная функция коэффициентов системы, $f(x)$ – $(n \times 1)$ вектор-функция свободных членов.

Предположим, что матричная функция $A(x)$ и вектор-функция $f(x)$ определены и непрерывны на $[a, b]$.

Отметим, что задача Коши для системы (1) ($y(x_0) = y^0$, одноточечная задача) имеет единственное решение $y(x)$, которое определено на $[a, b]$.

Сформулируем линейную многоточечную задачу для системы (1).
Пусть на $[a, b]$ задан набор точек $a \le x_1 < x_2 < \ldots < x_m \le b$, $m \le n$. Присоединим к системе (1) условия следующего вида

$$B_s y(x_s) = \beta^{(s)}, \quad s = \overline{1, m}. \tag{2}$$

Здесь B_s - заданные матрицы размеров $(k_s \times n)$, $\beta^{(s)}$ - известные векторы размеров $(k_s \times 1)$.

Потребуем, чтобы общее число условий (2) было равно размерности системы (1):

$$\sum_{s=1}^{m} k_s = n. \tag{3}$$

Получили линейную многоточечную (m-точечную) задачу (1)-(3): найти решение $y(x)$ системы (1), удовлетворяющее условиям (2). Отметим, что условия (2) с помощью скалярных произведений представляются в виде

$$\langle B_s^{(i)}, y(x_s) \rangle = \beta_i^{(s)}, \quad i = \overline{1, k_s}, \quad s = \overline{1, m}. \tag{2'}$$

Здесь $B_s^{(i)}$ - i-ая строка матрицы B_s, $\beta_i^{(s)}$ – i-ая координата вектора $\beta^{(s)}$.

Выделим частный случай задачи (1)-(3): $m = 2$, $x_1 = a$, $x_2 = b$. В этом случае (1)-(3) – двухточечная краевая задача.

2. Метод прогонки (последовательного переноса условий)

Пусть в некоторой точке $x_0 \in [a,b]$ задано одно линейное условие

$$\langle c^0, y(x_0)\rangle = \alpha_0, \;\; c^0 \in R^n, \; \alpha_0 \in R \tag{4}$$

относительно решения системы (1).

Будем говорить, что условие (4) перенесено из точки x_0 в точку $z \in [a,b]$ вдоль решения $y(x)$ системы (1), если можно определить n-мерную вектор-функцию $c(x)$ и скалярную функцию $\alpha(x), x \in [a,b]$ таким образом, чтобы

$$c(x_0) = c^0, \;\; \alpha(x_0) = \alpha_0, \;\; \langle c(z), y(z)\rangle = \alpha(z)$$

Задача переноса (прогонки) условия решается следующим утверждением.

Лемма. Пусть $c(x), \alpha(x)$, $x \in [a,b]$ – решение задачи Коши

$$\begin{cases} c' = -A^T(x)c, & c(x_0) = c^0, \\ \alpha' = \langle c, f(x)\rangle, & \alpha(x_0) = \alpha_0. \end{cases}$$

Тогда для любой точки $z \in [a,b]$ $\langle c(z), y(z)\rangle = \alpha(z)$.

Доказательство. Проверим утверждение леммы

$$\langle c(z), y(z)\rangle = \langle c(x_0), y(x_0)\rangle + \int_{x_0}^{z} \langle c(x), y(x)\rangle' dx =$$

$$= \alpha(x_0) + \int_{x_0}^{z} [-\langle A^T(x)c(x), y(x)\rangle + \langle c(x), A(x)y(x)\rangle +$$

$$+ \langle c(x), f(x)\rangle] dx = \alpha(x_0) + \int_{x_0}^{z} \alpha'(x) dx = \alpha(z).$$

□

Замечание. Система $c' = -A^T(x)c$ называется сопряженной для исходной системы (1).

Метод прогонки для решения многоточечной задачи (1)-(3) заключается в последовательном переносе условий (2′) в одну точку, например, x_1. С этой целью согласно утверждению леммы решается задача Коши для единой системы

$$c' = -A^T(x)c, \quad \alpha' = \langle c, f(x)\rangle$$

с различными начальными условиями

$$c(x_s) = B_s^{(i)}, \quad \alpha(x_s) = \beta_i^{(s)}, \;\; i = \overline{1, k_s}, \; s = \overline{2, m}.$$

В результате получаем полный набор условий в точке x_1

$$\langle c^j(x_1), y(x_1)\rangle = \alpha_j(x_1), \quad j = \overline{1,n}. \tag{5}$$

Это система n линейных алгебраических уравнений относительно вектора

$$y(x_1) = (y_1(x_1), \ldots, y_n(x_1)).$$

Пусть $y(x_1) = y^1$ – решение системы(5). Тогда решение многоточечной задачи (1)-(3) находится как решение задачи Коши

$$y' = A(x)y + f(x), \quad y(x_1) = y^1.$$

Таким образом, метод прогонки сводит многоточечную задачу (1)-(3) к задаче Коши для системы (1). При этом вопрос о существовании и единственности решения задачи (1)-(3) сводится к аналогичному вопросу для линейной системы (5).

В заключение рассмотрим метод прогонки для двухточечной краевой задачи с условиями

$$y_i(a) = \beta_i^{(1)} \; i = \overline{1,r}, \quad y_i(b) = \beta_i^{(2)} \; i = \overline{r+1,n}.$$

Отметим, что $y_i(b) = \langle e^i, y(b)\rangle, i = \overline{r+1,n}$, где $e^i \in R^n$ – i-ый орт.

Опишем перенос условий из точки b в точку a.

Для формирования линейной алгебраической системы необходимо решить серию задач Коши: $\forall i = \overline{r+1,n}$

$$c' = -A^T(x)c, \quad c(b) = e^i,$$

$$\alpha' = \langle c, f(x)\rangle, \quad \alpha(b) = \beta_i^{(2)}.$$

Пусть $c^i(x), \alpha_i(x)$, $i = \overline{r+1,n}$ – соответствующее решение.

В результате для значений $y_{r+1}(a), \ldots, y_n(a)$ получаем линейную систему

$$\langle c^i(a), y(a)\rangle = \alpha_i(a), \quad i = \overline{r+1,n}.$$

Далее решается задача Коши для системы (1) с полученным начальным условием $y(a)$.

§6. Линейная краевая задача для уравнения второго порядка

1. Постановка задачи. Разностная аппроксимация

Рассмотрим задачу отыскания решения $y(x)$, $x \in [a,b]$ дифференциального уравнения

$$y'' - p(x)y = f(x), \quad p(x) \geq 0, \tag{1}$$

удовлетворяющего краевым (граничным) условиям

$$y(a) = \gamma_0, \quad y(b) = \gamma_1. \tag{2}$$

Предположим, что функции $p(x), f(x)$ дважды непрерывно-дифференцируемы на $[a,b]$. Как известно, в этом случае решение $y(x)$ непрерывно-дифференцируемо до четвертого порядка включительно.

Для приближенного решения задачи (1),(2) применим разностный метод. С этой целью введем на $[a,b]$ равномерную сетку $x_i = a + ih$, $i = \overline{0,n}$, $nh = b - a$ ($x_0 = a, x_n = b$) с шагом $h > 0$ и рассмотрим уравнение (1) на решении $y(x)$ во внутренних узлах сетки

$$y''(x_i) - p(x_i)y(x_i) = f(x_i), \quad i = \overline{1,n-1}. \tag{1'}$$

Для аппроксимации производной используем разностное отношение

$$y''(x_i) = \frac{y(x_{i-1}) - 2y(x_i) + y(x_{i+1})}{h^2} - \frac{h^2}{12}y^{(4)}(\xi_i), \tag{3}$$

$$\xi_i \in [x_{i-1}, x_{i+1}].$$

Обозначим $p_i = p(x_i)$, $f_i = f(x_i)$ и пусть, как обычно, y_i - приближенное значение решения в точке x_i : $y_i \approx y(x_i)$, $i = \overline{1,n-1}$. При этом $y_0 = y(x_0)$, $y_n = y(x_n)$. Тогда разностная аппроксимация для уравнения (1′) имеет вид

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} - p_i y_i = f_i, \quad i = \overline{1,n-1}. \tag{4}$$

Дополним соотношения (4) краевыми условиями

$$y_0 = \gamma_0, \quad y_n = \gamma_1. \tag{5}$$

В результате получили разностную краевую задачу (разностную схему) (4),(5) для приближенного решения исходной непрерывной задачи (1),(2).

Отметим, что задача (4), (5) представляет собой систему $(n-1)$ линейных алгебраических уравнений относительно неизвестных $y_1, \ldots, y_{n-1}$.

Обсудим связь между дифференциальной и разностной задачами. Величина

$$\Delta_i = \frac{y(x_{i-1}) - 2y(x_i) + y(x_{i+1})}{h^2} - p_i y(x_i) - f_i$$

характеризует погрешность аппроксимации в узле x_i, $i = \overline{1, n-1}$, соответствующую разностной схеме (4). Выясним порядок аппроксимации относительно шага сетки h.

Учитывая уравнение $(1')$ и соотношение (3), получаем

$$\Delta_i = \frac{h^2}{12} y^{(4)}(\xi_i), \quad \xi_i \in [x_{i-1}, x_{i+1}].$$

Пусть, как обычно, $M_4 = \max_{[a,b]} |y^{(4)}(x)|$. Тогда глобальная погрешность аппроксимации $\Delta = \max_{1 \le i \le n-1} |\Delta_i|$ оценивается следующим образом: $\Delta \le \frac{h^2}{12} M_4$. Это значит, что разностная схема (4), (5) имеет *второй порядок аппроксимации.*

Введем локальную погрешность приближенного решения $r_i = y(x_i) - y_i, i = \overline{0, n}$. Понятно, что $r_0 = r_n = 0$. Имеет место следующая оценка погрешности [1, с.403]

$$R = \max_{0<i<n} |r_i| \le C\Delta, \quad C = \frac{(b-a)^2}{8}.$$

Это значит, что разностная схема (4), (5) имеет *второй порядок точности.*

Перейдем к вопросу численного решения разностной задачи (4),(5). Это система линейных алгебраических уравнений с трехдиагональной матрицей коэффициентов

$$y_0 = \gamma_0, \;\; y_{i-1} - (2 + h^2 p_i) y_i + y_{i+1} = h^2 f_i, \;\; i = \overline{1, n-1}, \quad y_n = \gamma_1. \qquad (6)$$

Опишем специальный метод решения такого сорта систем.

2. Метод прогонки (для решения разностной задачи)

Рассмотрим следующую систему $(n+1)$ уравнений с неизвестными y_0, $y_1, \ldots y_n$

$$y_0 + c_0 y_1 = \varphi_0, \qquad (7)$$

$$a_i y_{i-1} + b_i y_i + c_i y_{i+1} = \varphi_i, \quad i = \overline{1, n-1}, \tag{8}$$

$$a_n y_{n-1} + y_n = \varphi_n. \tag{9}$$

Матрица коэффициентов системы имеет вид (трехдиагональная матрица)

$$A = \begin{pmatrix} 1 & c_0 & & & \\ a_1 & b_1 & c_1 & & \\ & a_2 & b_2 & c_2 & \\ & \dots & & & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_n & 1 \end{pmatrix}.$$

Будем искать решение системы в виде

$$y_k = \alpha_{k+1} y_{k+1} + \beta_{k+1}, \quad k = \overline{0, n-1}, \tag{10}$$

где $\alpha_{k+1}, \beta_{k+1}$ – неопределенные коэффициенты.

Найдем их из уравнений (7), (8) системы. Полагая в (10) $k = 0$ и сравнивая с (7), получаем

$$\alpha_1 = -c_0, \quad \beta_1 = \varphi_0.$$

Далее, используя (10), выразим переменные y_{i-1}, y_i через $y_{i+1}, i = \overline{1, n-1}$

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad y_{i-1} = \alpha_i(\alpha_{i+1} y_{i+1} + \beta_{i+1}) + \beta_i.$$

После подстановки в (8) получаем

$$y_{i+1}[c_i + \alpha_{i+1}(b_i + \alpha_i a_i)] + \beta_i a_i +$$

$$+\beta_{i+1}(b_i + \alpha_i a_i) = \varphi_i, \quad i = \overline{1, n-1}.$$

Положим

$$c_i + \alpha_{i+1}(b_i + \alpha_i a_i) = 0.$$

В результате приходим к рекуррентным формулам для подсчета коэффициентов α_{k+1}, β_{k+1} из (10)

$$\alpha_1 = -c_0, \quad \alpha_{i+1} = -\frac{c_i}{b_i + \alpha_i a_i}, \tag{11}$$

$$\beta_1 = \varphi_0, \quad \beta_{i+1} = \frac{\varphi_i - \beta_i a_i}{b_i + \alpha_i a_i}, \quad i = \overline{1, n-1},$$

$$(b_i + \alpha_i a_i \neq 0, \quad i = \overline{1, n-1}). \tag{12}$$

Чтобы воспользоваться формулой (10) для отыскания решения системы, необходимо найти y_n. Полагая в (10) $k = n-1$ и подставляя в (9), получаем

$$a_n(\alpha_n y_n + \beta_n) + y_n = \varphi_n.$$

Отсюда

$$y_n = \frac{\varphi_n - \beta_n a_n}{1 + \alpha_n a_n} \tag{13}$$

при условии

$$1 + \alpha_n a_n \neq 0. \tag{14}$$

Таким образом, решение системы (7)-(9) *методом прогонки* проводится следующим образом:

1) по формулам (11) определяются коэффициенты α_i, β_i, $i = \overline{1,n}$ (*прямая прогонка*),

2) решение системы находится от y_n до y_0 по формулам (13), (10) (*обратная прогонка*).

Замечание 1. Нетрудно видеть, что арифметическая трудоемкость метода прогонки равна $(10n)$ операций.

Проведем обоснование метода прогонки, т.е. сформулируем условия на матрицу A, гарантирующие корректность метода.

Теорема. *Пусть выполнены следующие условия на коэффициенты системы (7)-(9)*

$$a_i \neq 0, \;\; c_i \neq 0, \;\; |b_i| \geq |a_i| + |c_i|, \;\; i = \overline{1, n-1}, \tag{15}$$

$$|c_0| \leq 1, \;\; |a_n| \leq 1, \;\; |c_0| + |a_n| < 2. \tag{16}$$

Тогда метод прогонки реализуем, т.е. условия (12), (14) выполнены.

Доказательство. Предварительно напомним одно модульное неравенство:

$$|a + b| \geq |a| - |b|; \;\; a, b \in R.$$

С помощью математической индукции покажем, что $|\alpha_i| \leq 1$, $i = \overline{1,n}$.

При $i = 1$ в силу (16) имеем $|\alpha_1| = |c_0| \leq 1$. Тогда с учетом (15)

$$|b_1 + \alpha_1 a_1| \geq |b_1| - |\alpha_1||a_1| \geq |a_1|(1 - |\alpha_1|) + |c_1| > 0.$$

Это значит, что условие (12) при $i = 1$ выполнено. Пусть $|\alpha_k| \leq 1$, $k \in \{2, \ldots, n-1\}$ и покажем, что $|\alpha_{k+1}| \leq 1$.

Учитывая условие (15) и предположение индукции, получаем

$$|b_k + \alpha_k a_k| \geq |b_k| - |\alpha_k||a_k| \geq |a_k|(1 - |\alpha_k|) + |c_k| > 0,$$

т.е. $b_k + \alpha_k a_k \neq 0$.

Кроме того, $|b_k + \alpha_k a_k| \geq |c_k|$, поэтому

$$|\alpha_{k+1}| = \frac{|c_k|}{|b_k + \alpha_k a_k|} \leq 1.$$

Таким образом, $|\alpha_i| \leq 1$, $i = \overline{1, n}$, при этом выполнены условия (12).

Заметим, что если $|\alpha_1| < 1$, то с помощью тех же рассуждений получаем $|\alpha_i| < 1$, $i = \overline{2, n}$.

Рассмотрим условие (14). Используя оценку для модуля, имеем

$$|1 + \alpha_n a_n| \geq 1 - |\alpha_n||a_n|.$$

Пусть $|\alpha_1| < 1$, тогда $|\alpha_n| < 1$. Поскольку $|a_n| \leq 1$ (условие (16)), то $|1 + \alpha_n a_n| > 0$.

Если $|\alpha_1| = |c_0| = 1$, то $|\alpha_n| \leq 1$, причем из (16) следует, что $|a_n| < 1$. Тогда, как и ранее, $|1 + \alpha_n a_n| > 0$. Итак, условие (14) выполнено. □

Замечание 2. Условия

$$|b_i| \geq |a_i| + |c_i|, \quad i = \overline{1, n-1}, \quad |c_0| \leq 1, \quad |a_n| \leq 1$$

означают, что A - матрица с диагональным преобладанием.

Замечание 3.Условия теоремы гарантируют существование и единственность решения системы (7)-(9).

Проверим выполнение условий теоремы для разностной задачи (4), (5), которая представлена в виде системы (6).

В данном случае

$$c_0 = 0, \quad a_i = 1, \quad b_i = -(2 + h^2 p_i), \quad c_i = 1, \quad i = \overline{1, n-1}, \quad a_n = 0.$$

Поскольку по условию $p_i \geq 0$, $i = \overline{1, n-1}$, то соотношения (15), (16) выполнены ($|b_i| \geq 2$). Следовательно, *система (6) (разностная задача (4), (5)) имеет единственное решение, которое можно найти по методу прогонки.*

§7. Вариационные методы решения краевых задач

1. Редукция к вариационной задаче

Рассмотрим краевую задачу для дифференциального уравнения второго порядка на $[a, b]$

$$y'' - p(x)y = f(x), \quad p(x) \geq 0, \tag{1}$$

$$y(a) = \gamma_a, \quad y(b) = \gamma_b. \tag{2}$$

Предположим, что функции $p(x), f(x)$ непрерывны на $[a, b]$.

Проведем редукцию задачи (1), (2) к задаче вариационного исчисления. С этой целью образуем функционал

$$I(y) = \int_a^b F(x, y(x), y'(x))dx,$$

где $F(x, y, y') = (y')^2 + p(x)y^2 + 2f(x)y$.

Рассмотрим задачу

$$I(y) \to \min, \quad y(a) = \gamma_a, \quad y(b) = \gamma_b \tag{3}$$

на множестве непрерывно - дифференцируемых (гладких) функций $y(x)$, $x \in [a, b]$. Это простейшая задача вариационного исчисления с закрепленными концами.

Установим связь между задачами (1), (2) и (3).

Теорема 1. *Задачи (1),(2) и (3) эквивалентны: решение задачи (1), (2) является решением задачи (3) и наоборот.*

Доказательство. Пусть $y_*(x)$ - решение задачи (1),(2). Возьмем произвольную допустимую функцию $y(x)$ задачи (3) и положим $\delta(x) = y(x) - y_*(x)$. Понятно, что $\delta(a) = \delta(b) = 0$. Подсчитаем значение функционала

$$I(y) = I(y_* + \delta) =$$

$$= \int_a^b [(y_*'(x) + \delta'(x))^2 + p(x)(y_*(x) + \delta(x))^2 + 2f(x)(y_*(x) + \delta(x))]dx =$$

$$= I(y_*) + 2\int_a^b [y_*'(x)\delta'(x) + p(x)y_*(x)\delta(x) + f(x)\delta(x)]dx +$$

$$+ \int_a^b [(\delta'(x))^2 + p(x)\delta^2(x)]dx =$$

$$= I(y_*) + 2I_1 + I_2.$$

Применим формулу интегрирования по частям

$$\int_a^b y_*'(x)\delta'(x)dx = y_*'(x)\delta(x)|_a^b - \int_a^b y_*''(x)\delta(x)dx$$

и учтем, что $y_*''(x) = p(x)y_*(x) + f(x)$ (в силу уравнения (1)).

Тогда $I_1 = 0$. Поскольку $I_2 \geq 0$ ($p(x) \geq 0$), то получаем неравенство $I(y) \geq I(y_*)$. Это значит, что $y_*(x)$ – решение задачи (3).

Пусть теперь $y_*(x)$ – решение задачи (3). Тогда функция $y_*(x)$ является экстремалью функционала $I(y)$ т.е. удовлетворяет уравнению Эйлера

$$F_y - \frac{d}{dx}F_{y'} = 0.$$

В данном случае это уравнение имеет вид $2[p(x)y_*(x)+f(x)-y_*''(x)] = 0$, что эквивалентно (1). Таким образом, $y_*(x)$ – решение краевой задачи (1),(2). □

Замечание. Отметим, что функционал $I(y)$ построен таким образом, чтобы уравнение Эйлера для него (необходимое условие экстремума в задаче (3), условие стационарности) было эквивалентным уравнению (1) краевой задачи (1), (2) (решение уравнения (1) является экстремалью функционала $I(y)$ и наоборот).

Редукция дифференциальной задачи (1)-(2) к вариационной можно элементарно провести с помощью метода наименьших квадратов. В этом случае функционал имеет вид (среднеквадратичная невязка уравнения (1))

$$J(y) = \int_a^b [y''(x) - p(x)y(x) - f(x)]^2 dx.$$

Соответствующая вариационная задача

$$J(y) \to \min, \quad y(a) = \gamma_a, \quad y(b) = \gamma_b \tag{4}$$

рассматривается в классе дважды непрерывно - дифференцируемых функций $y(x)$.

Связь между задачами (1),(2) и (4) вполне очевидна.

2. Метод Ритца

Метод Ритца сводит вариационную (бесконечномерную) задачу на экстремум функционала к последовательности конечномерных задач на экстремум функций конечного числа переменных. Проведем описание метода применительно к задаче (3).

Выберем последовательность базисных (координатных) функций $\varphi_k(x)$, $x \in [a,b]$, $k = 0,1,\ldots$, удовлетворяющих (как минимум) следующим условиям

1) $\varphi_k \in C_1[a,b]$, $k = 0,1,\ldots$ ($C_1[a,b]$ – пространство непрерывно - дифференцируемых на $[a,b]$ функций),

2) $\varphi_0(a) = \gamma_a$, $\varphi_0(b) = \gamma_b$, $\varphi_k(a) = \varphi_k(b) = 0$, $k = 1,2,\ldots$,

3) для любого конечного n функции $\varphi_1(x)$, $\varphi_2(x),\ldots,$ $\varphi_n(x)$ линейно независимы.

Зафиксируем натуральное число n (параметр метода) и будем искать приближенное решение $y_n(x)$ задачи (3) в виде линейной комбинации

$$y_n(x) = \varphi_0(x) + \sum_{k=1}^{n} c_k \varphi_k(x), \quad x \in [a,b]$$

с неопределенными коэффициентами c_k.

Отметим, прежде всего, что согласно условиям 1), 2) функция $y_n(x)$ является допустимой в задаче (3) при любом выборе коэффициентов c_k :

$$y_n \in C_1[a,b], \quad y_n(a) = \gamma_a, \quad y_n(b) = \gamma_b.$$

При этом $I(y_n) = \Phi(c_1,\ldots,c_n)$, и задача (3) переходит в конечномерную задачу

$$\Phi(c_1,\ldots,c_n) \to \min \tag{5}$$

на безусловный минимум функции n переменных.

В нашем случае функция Φ является квадратичной

$$\Phi(c_1,\ldots,c_n) = A_0 + 2\sum_{k=1}^{n} A_k c_k + \sum_{i,k=1}^{n} A_{ik} c_i c_k,$$

где коэффициенты A_0, A_k, A_{ik} легко выписываются. В результате задача (5) сводится к решению системы линейных алгебраических уравнений относительно неизвестных $c_1,\ldots,c_n$

$$\frac{1}{2}\frac{\partial \Phi}{\partial c_k} = A_k + \sum_{i=1}^{n} A_{ik} c_i = 0, \quad k = \overline{1,n}.$$

Нетрудно проверить, что данная система имеет единственное решение $c_1^*, \ldots, c_n^*$, которое определяет приближенное решение $y_n^*(x)$ задачи (3) (задачи (1), (2)).

В заключение укажем возможные способы выбора базисных функций $\varphi_k(x), k = 1, 2, \ldots$

1) $\varphi_k(x) = (x-a)^k(x-b)$ или $\varphi_k(x) = (x-a)(x-b)^k$,

2) $\varphi_k(x) = \sin k\pi \frac{x-a}{b-a}$.

Глава 2. Уравнения с частными производными

Основным аппаратом численного решения уравнений с частными производными являются разностные методы (методы сеток). В этом случае уравнение и граничные условия аппроксимируются некоторыми разностными соотношениями (схемами), и проблема сводится к решению систем алгебраических уравнений.

§1. Основные понятия теории разностных схем

В пусть в области $\mathcal{D}$ изменения переменных x, y определена некоторая дифференциальная задача (дифференциальное уравнение и граничные условия)

$$Lu(x,y) = f(x,y), \quad (x,y) \in \mathcal{D}. \tag{1}$$

Здесь L – линейный дифференциальный оператор, $u(x,y)$ – искомое решение, $f(x,y)$ – заданная функция.

Перейдем к дискретному аналогу задачи (1). Прежде всего введем сетку $\mathcal{D}_h \subset \mathcal{D}$ – конечное множество точек из $\mathcal{D}$, плотность распределения которых характеризуется параметром h – шагом сетки. В общем случае h – векторный параметр (вектор). Определим величину $|h|$ – длина (норма) вектора h. Сетка $\mathcal{D}_h$ строится таким образом, что при $|h| \to 0$ число точек (узлов) сетки увеличивается (множество $\mathcal{D}_h$ "стремится"заполнить всю область $\mathcal{D}$). Численное решение задачи (1) ищется в узлах сетки.

Функция, определенная в узлах сетки $\mathcal{D}_h$, называется сеточной функцией. Введем линейное, нормированное пространство U_h сеточных функций с нормой $\|\cdot\|$.

Пусть $u(x,y)$, $(x,y) \in \mathcal{D}$ – точное решение дифференциальной задачи (1). Обозначим через u_h соответствующую сеточную функцию: $u_h = u(x,y)$, $(x,y) \in \mathcal{D}_h$ – точное решение в узлах сетки. Введем сеточную функцию v_h – приближенное решение в узлах сетки: $v_h \approx u_h$. Пусть далее f_h – сеточная функция, соответствующая функции $f(x,y)$, например, $f_h = f(x,y)$, $(x,y) \in \mathcal{D}_h$.

Для вычисления приближенного решения v_h проведем аппроксимацию задачи (1), заменяя дифференциальный оператор L разностным L_h (заменяя производные в узлах сетки их разностными аппроксимациями). В результате получаем разностную схему (систему алгебраических уравнений

относительно v_h)

$$L_h v_h = f_h. \tag{2}$$

Здесь L_h – разностный оператор. Предположим, что L_h – линейный оператор, т.е. разностная схема (2) – система линейных уравнений относительно сеточной функции v_h.

Установим связь между задачами (1), (2).

Определение 1. *Сеточная функция $\Delta_h = L_h u_h - f_h$ называется погрешностью аппроксимации разностной схемы (2) на решении $u(x, y)$ задачи (1).*

Величина $\|\Delta_h\|$ – невязка разностной схемы (2) на решении $u(x, y)$.

Определение 2. *Говорят, что разностная схема (2) аппроксимирует задачу (1), если $\|\Delta_h\| \to 0$ при $|h| \to 0$.*

Если $\|\Delta_h\| \le c_1 |h|^s$, $c_1 > 0$, то разностная схема имеет s-ый порядок аппроксимации.

Определение 3. *Разностная схема (2) называется корректной, если:*
1) её решение v_h существует и единственно при любой правой части f_h;
2) для любой правой части f_h имеет место оценка

$$\|v_h\| \le c_2 \|f_h\|, \quad c_2 > 0.$$

Свойство 2) называют устойчивостью разностной схемы (2).

Основным вопросом теории разностных схем является вопрос о сходимости.

Введем сеточную функцию $r_h = u_h - v_h$, которая называется *погрешностью разностной схемы (2).*

Определение 4. *Решение разностной задачи (2) сходится к решению дифференциальной задачи (1) (разностная схема (2) сходится) если $\|r_h\| \to 0$ при $|h| \to 0$.*

Говорят, что разностная схема (2) имеет s-ый порядок точности, если

$$\|r_h\| \le c_3 |h|^s, \quad c_3 > 0.$$

Установим связь между введенными понятиями.

Теорема. *Пусть разностная схема (2) является корректной и аппроксимирует исходную задачу (1). Тогда решение разностной задачи (2) сходится к решению дифференциальной задачи (1), причем порядок точности совпадает с порядком аппроксимации.*

Доказательство. По условию теоремы $\|\Delta_h\| \to 0$, $|h| \to 0$, где $\Delta_h = L_h u_h - f_h$ – погрешность аппроксимации. Для погрешности $r_h = u_h - v_h$ с учетом линейности оператора L_h получаем $L_h r_h = L_h u_h - L_h v_h = L_h u_h - f_h = \Delta_h$. Это значит, что r_h – решение разностной схемы (2) с правой частью Δ_h : $L_h r_h = \Delta_h$. В силу свойства устойчивости имеет место оценка $\|r_h\| \le c_2\|\Delta_h\|$. Отсюда получаем свойство сходимости: $\|r_h\| \to 0$, $|h| \to 0$. Если $\|\Delta_h\| \le c_1|h|^s$, то $\|r_h\| \le c_2 c_1 |h|^s = c|h|^s$, т.е. порядок точности совпадает с порядком аппроксимации. $\square$

В заключение укажем основные этапы построения и исследования разностных схем:

1) устанавливается правило выбора сетки в области $\mathcal{D}$;

2) строится одна или несколько разностных схем, выясняется порядок аппроксимации;

3) исследуется корректность (устойчивость) построенных разностных схем;

4) рассматривается вопрос о численном решении разностных схем.

Проиллюстрируем введенные понятия и соотношения на примере классического уравнения теплопроводности.

§2. Разностные схемы для уравнения теплопроводности

1. Явная разностная схема

В области $\mathcal{D} = \{(x,t) : \ 0 \le x \le X, \ 0 \le t \le T\}$ (прямоугольник) требуется найти решение $u(x,t)$ уравнения

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t) \tag{1}$$

с начальным и краевыми (граничными) условиями

$$u(x,0) = u_0(x), \quad u(0,t) = \mu_1(t), \quad u(X,t) = \mu_2(t). \tag{2}$$

Здесь $u_0(x)$, $\mu_1(t)$, $\mu_2(t)$ – заданные функции.

Предположим, что существует достаточно гладкое решение $u(x,t)$ задачи (1), (2).

Построим разностные схемы для поставленной задачи.

Введем сетку $\omega_{h,\tau}$ в прямоугольнике $\mathcal{D}$ как совокупность точек (x_i, t_j), $i = \overline{0,m}$, $j = \overline{0,n}$, $x_i = ih$, $t_j = j\tau$, $mh = X$, $n\tau = T$. Здесь $h > 0$ – шаг сетки по переменной x, $\tau > 0$ – шаг сетки по переменной t.

Узлы $(x_i, 0)$, $(0, t_j)$, (X, t_j), $i = \overline{0,m}$, $j = \overline{0,n}$ назовем граничными узлами сетки. Остальные узлы внутренние.

Совокупность узлов $S_j = \{(x_0, t_j), (x_1, t_j), \ldots, (x_m, t_j)\}$ называется j-ым слоем сетки $\omega_{h,\tau}$, $j = \overline{0,n}$.

Перейдем к разностной аппроксимации дифференциальной задачи в узлах сетки. Обозначим:

$u_i^j = u(x_i, t_j)$ – точное решение в узлах сетки,

$v_i^j \approx u_i^j$ – приближенное решение,

$f_i^j = f(x_i, t_j)$ $i = \overline{0,m}$, $j = \overline{0,n}$.

В граничных узлах известно точное решение, поэтому условия (2) аппроксимируются точно

$$v_i^0 = u_i^0 = u_0(x_i), \quad v_0^j = u_0^j = \mu_1(t_j), \quad v_m^j = u_m^j = \mu_2(t_j), \quad i = \overline{0,m}, \quad j = \overline{0,n}.$$

Для аппроксимации уравнения (1) в узлах сетки используем формулы численного дифференцирования

$$\frac{\partial u(x_i,t_j)}{\partial t} = \frac{u(x_i,t_{j+1}) - u(x_i,t_j)}{\tau} - \frac{\tau}{2}\frac{\partial^2 u(x_i, t_j^{(1)})}{\partial t^2},$$

$$\frac{\partial^2 u(x_i,t_j)}{\partial x^2} = \frac{u(x_{i-1},t_j) - 2u(x_i,t_j) + u(x_{i+1},t_j)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u(x_i^{(1)}, t_j)}{\partial x^4}, \tag{3}$$

$$i = \overline{1,m-1}, \quad j = \overline{0,n-1}.$$

В результате получаем разностную схему относительно сеточной функции v_i^j

$$\frac{v_i^{j+1} - v_i^j}{\tau} = \frac{v_{i-1}^j - 2v_i^j + v_{i+1}^j}{h^2} + f_i^j,$$
$$i = \overline{1,m-1}, \quad j = \overline{0,n-1}; \tag{4}$$

$$v_i^0 = u_0(x_i), \quad i = \overline{0,m},$$
$$v_0^j = \mu_1(t_j), \quad v_m^j = \mu_2(t_j), \quad j = \overline{0,n}.$$

Эта схема представляет собой систему линейных алгебраических уравнений относительно переменных v_i^j.

Решение данной системы находится по слоям S_j, $j = \overline{0,n}$. Решение на нулевом слое известно: $v_i^0 = u_0(x_i)$, $i = \overline{0,m}$. Если решение на слое S_j уже найдено, то решение на слое S_{j+1} находится по явной формуле

$$v_i^{j+1} = v_i^j + \tau\left(\frac{v_{i-1}^j - 2v_i^j + v_{i+1}^j}{h^2} + f_i^j\right), \quad i = \overline{1,m-1},$$

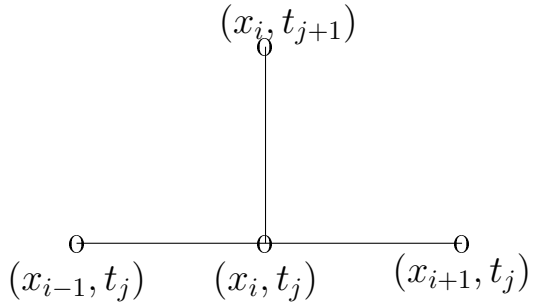
причем

$$v_0^{j+1} = \mu_1(t_{j+1}), \quad v_m^{j+1} = \mu_2(t_{j+1}).$$

Соотношения (4) называют явной разностной схемой.

Определение. *Совокупность узлов сетки, которые используются для разностной аппроксимации дифференциального уравнения (1) в узле (x_i, t_j) называется шаблоном.*

В случае схемы (4) шаблон имеет вид



Это явный двуслойный шаблон.

2. Аппроксимация и устойчивость

Погрешность аппроксимации для разностной схемы (4) во внутренних узлах сетки определяется сеточной функцией

$$\Delta_i^j = \frac{u_i^{j+1} - u_i^j}{\tau} - \frac{u_{i-1}^j - 2u_i^j + u_{i+1}^j}{h^2} - f_i^j, \quad i = \overline{1,m-1}, \quad j = \overline{0,n-1}.$$

Погрешность аппроксимации в граничных узлах сетки равна нулю:

$$\Delta_i^0 = 0, \quad \Delta_0^j = 0, \quad \Delta_m^j = 0.$$

Используя формулы (3) и уравнение (1) в узлах сетки

$$\frac{\partial u(x_i,t_j)}{\partial t}=\frac{\partial^2 u(x_i,t_j)}{\partial x^2}+f(x_i,t_j),$$

получаем выражение

$$\Delta_i^j=\frac{\tau}{2}\,\frac{\partial^2 u(x_i,t_j^{(1)})}{\partial t^2}-\frac{h^2}{12}\,\frac{\partial^4 u(x_i^{(1)},t_j)}{\partial x^4}.$$

Выясним порядок аппроксимации. С этой целью будем использовать сеточную норму пространства C :

$$\text{если}\ \ \Delta=\{\Delta_i^j,\ i=\overline{0,m},\ j=\overline{0,n-1}\},\ \text{то}\ \ \|\Delta\|=\max_{0\le i\le m,\ \ 0\le j\le n-1}|\Delta_i^j|.$$

Предположим, что имеют место оценки

$$\left|\frac{\partial^2 u(x,t)}{\partial t^2}\right|\le M_2,\quad \left|\frac{\partial^4 u(x,t)}{\partial x^4}\right|\le M_4,\quad (x,t)\in\mathcal{D}.$$

Тогда

$$\|\Delta\|\le\frac{\tau}{2}\,M_2+\frac{h^2}{12}\,M_4\le M(\tau+h^2),\quad M=\max\left\{\frac{M_2}{2},\frac{M_4}{12}\right\}.$$

Это значит, что разностная схема (4) аппроксимирует дифференциальную задачу (1), (2) ($\|\Delta\|\to 0$ при $\tau\to 0$, $h\to 0$) и имеет первый порядок аппроксимации относительно величины $(\tau+h^2)$ (первый порядок относительно τ, второй порядок относительно h).

Рассмотрим вопрос об устойчивости разностной схемы (4). С этой целью будем использовать следующие сеточные нормы:

$$\|v^j\|=\max_{0\le i\le m}|v_i^j|,\quad \|v\|=\max_{0\le j\le n}\|v^j\|,\quad \|f\|=\max_{0\le i\le m,\ \ 0\le j\le n}|f_i^j|,$$

$$\|u_0\|=\max_{0\le i\le m}|u_0(x_i)|,\quad \|\mu_k\|=\max_{0\le j\le n}|\mu_k(t_j)|,\quad k=1,2.$$

Будем говорить, что разностная схема (4) устойчива в выбранных нормах, если имеет место оценка

$$\|v\|\le C(\|f\|+\max\{\|u_0\|,\|\mu_1\|,\|\mu_2\|\}),$$

где *const* C не зависит от h,τ.

Теорема 1. *Пусть* $\tau\le\frac{1}{2}h^2$. *Тогда разностная схема (4) устойчива.*

Доказательство. Положим $\rho = \frac{\tau}{h^2}$. По условию теоремы $\rho \le \frac{1}{2}$. Запишем схему (4) в виде

$$v_i^{j+1} = (1-2\rho)v_i^j + \rho v_{i-1}^j + \rho v_{i+1}^j + \tau f_i^j, \quad i = \overline{1, m-1}, \quad j = \overline{0, n-1}.$$

Отсюда $(1 - 2\rho \ge 0)$

$$\max_{0<i<m} |v_i^{j+1}| \le (1-2\rho) \max_{0<i<m} |v_i^j| + \rho \max_{0<i<m} |v_{i-1}^j| +$$

$$\rho \max_{0<i<m} |v_{i+1}^j| + \tau \max_{0<i<m} |f_i^j| \le (1-2\rho)\|v^j\| + 2\rho\|v^j\| +$$

$$+\tau\|f\| = \|v^j\| + \tau\|f\|.$$

Кроме того,

$$|v_0^{j+1}| = |\mu_1(t_{j+1})| \le \|\mu_1\|, \quad |v_m^{j+1}| = |\mu_2(t_{j+1})| \le \|\mu_2\|.$$

В совокупности заключаем, что

$$\|v^{j+1}\| \le \max\{\|\mu_1\|, \|\mu_2\|, \|v^j\| + \tau\|f\|\} \le \tag{5}$$

$$\le \max\{\|\mu_1\|, \|\mu_2\|, \|v^j\|\} + \tau\|f\|, \quad j = \overline{0, n-1}.$$

Покажем, что

$$\|v^k\| \le M + k\tau\|f\|, \quad k = \overline{1, n}, \tag{6}$$

где $M = \max\{\|\mu_1\|, \|\mu_2\|, \|u_0\|\}$.

Будем рассуждать по индукции. При $k = 1$ неравенство (6) справедливо, вследствие (5) при $j = 0$ ($\|v^0\| = \|u_0\|$). Пусть оценка (6) имеет место при $k = j$. Тогда согласно (5)

$$\|v^{j+1}\| \le \max\{\|\mu_1\|, \|\mu_2\|, M + j\tau\|f\|\} + \tau\|f\| \le M + (j+1)\tau\|f\|.$$

Получили оценку (6) при $k = j + 1$.

Таким образом, неравенство (6) доказано. Отметим, что при $k = 0$ оценка (6), очевидно, справедлива. Следовательно, имеет место оценка

$$\|v\| = \max_{0\le k\le n} \|v^k\| \le M + n\tau\|f\| \le$$

$$\le C(\|f\| + M), \quad C = \max\{1, T = n\tau\}.$$

Определение. *Разностная схема, которая обладает устойчивостью при определенных соотношениях между шагами сетки, называется условно устойчивой.*

Если устойчивость имеет место без всяких условий на шаги сетки, то схема называется абсолютно (безусловно) устойчивой.

Таким образом, схема (4) условно устойчива (устойчива при условии $\frac{\tau}{h^2} \leq \frac{1}{2}$).

Вывод. Разностная схема (4) является сходящейся при условии $\tau \leq \frac{1}{2}h^2$, порядок точности – $(\tau + h^2) : \|v - u\| \leq C(\tau + h^2)$.

3. Неявная разностная схема

Построим вторую разностную схему для задачи (1), (2). Сохраняя разностную аппроксимацию для второй производной, будем использовать левую разностную производную

$$\frac{\partial u(x_i, t_j)}{\partial t} = \frac{u(x_i, t_j) - u(x_i, t_{j-1})}{\tau} + \frac{\tau}{2}\,\frac{\partial^2 u(x_i, t_j^{(2)})}{\partial t^2}, \quad t_j^{(2)} \in [t_{j-1}, t_j], \quad j = \overline{1, n}.$$

Далее, проведем сдвиг по индексу: $j \to j + 1$. В результате получаем разностную схему

$$\frac{v_i^{j+1} - v_i^j}{\tau} = \frac{v_{i-1}^{j+1} - 2v_i^{j+1} + v_{i+1}^{j+1}}{h^2} + f_i^{j+1} \quad i = \overline{1, m-1}, \quad j = \overline{0, n-1},$$

$$v_i^0 = u_0(x_i), \quad i = \overline{0, m}, \tag{7}$$

$$v_0^j = \mu_1(t_j), \quad v_m^j = \mu_2(t_j), \quad j = \overline{0, n}.$$

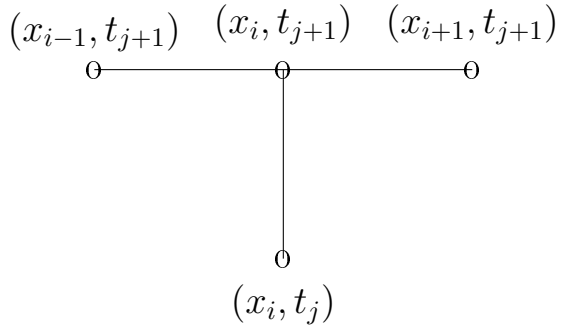
Для подсчета решения на $(j+1)$-ом слое имеем линейную систему

$$-\frac{\tau}{h^2}\, v_{i-1}^{j+1} + (1 + \frac{2\tau}{h^2})\, v_i^{j+1} - \frac{\tau}{h^2}\, v_{i+1}^{j+1} = v_i^j + \tau\, f_i^{j+1}, \quad i = \overline{1, m-1},$$

$$v_0^{j+1} = \mu_1(t_{j+1}), \quad v_m^{j+1} = \mu_2(t_{j+1}).$$

Для каждого $j = \overline{0, n-1}$ это линейная система с трехдиагональной матрицей относительно переменных $v_1^{j+1}, \ldots, v_{m-1}^{j+1}$. Решение существует, единственно и его можно найти с помощью метода прогонки, условия применимости которого в данном случае выполнены (матрица системы со строгим диагональным преобладанием: $1 + \frac{2\tau}{h^2} > \frac{2\tau}{h^2}$).

Таким образом, (7) – неявная (чисто неявная) разностная схема, использующая неявный двуслойный шаблон.



В полной аналогии с предыдущим заключаем, что разностная схема (7) аппроксимирует задачу (1), (2) с порядком $(\tau + h^2)$.

Рассмотрим основной вопрос об устойчивости разностной схемы (7).

Теорема 2. *Разностная схема (7) абсолютно устойчива.*

Доказательство. Представим соотношения (7) в виде

$$v_i^{j+1} + \rho(-v_{i-1}^{j+1} + 2v_i^{j+1} - v_{i+1}^{j+1}) = v_i^j + \tau f_i^{j+1}, \quad i = \overline{1, m-1}, \quad \rho = \frac{\tau}{h^2}.$$

Пусть $k \in \{0, \ldots, m\}$ – наименьший индекс, для которого

$$|v_k^{j+1}| = \max_{0 \le i \le m} |v_i^{j+1}| = \|v^{j+1}\|.$$

Если $k = 0$, то $\|v^{j+1}\| = |v_0^{j+1}| \le \|\mu_1\|$. При $k = m$ получаем $\|v^{j+1}\| = |v_m^{j+1}| \le \|\mu_2\|$.

Пусть $k \in \{1, \ldots, m-1\}$. Тогда

$$|v_k^{j+1}| > |v_{k-1}^{j+1}|, \quad |v_k^{j+1}| \ge |v_{k+1}^{j+1}|.$$

Следовательно,

$$sign\ (2v_k^{j+1} - v_{k-1}^{j+1} - v_{k+1}^{j+1}) = sign\ v_k^{j+1}$$

(если $|a| > |b|$, то $sign\ (a - b) = sign\ a$).

Тогда

$$\|v^{j+1}\| = |v_k^{j+1}| \le |v_k^{j+1} + \rho\ (-v_{k-1}^{j+1} + 2v_k^{j+1} - v_{k+1}^{j+1})| =$$
$$= |v_k^j + \tau f_k^{j+1}| \le \|v^j\| + \tau\|f\|.$$

Таким образом, в любом случае имеет место оценка (см. неравенство (5))

$$\|v^{j+1}\| \le \max\{\|\mu_1\|, \|\mu_2\|, \|v^j\| + \tau\|f\|\}, \quad j = \overline{0, n-1}.$$

Дальнейший вывод полностью повторяет доказательство теоремы 1. □

4. Разностная схема с весами

Продолжим построение разностных схем для задачи (1), (2). Введем в рассмотрение разностный оператор

$$\Lambda_h v_i^j = \frac{v_{i-1}^j - 2v_i^j + v_{i+1}^j}{h^2}, \quad i = \overline{1, m-1}, \quad j = \overline{0, n}.$$

Определим семейство разностных схем с параметром $\sigma \in [0, 1]$

$$\frac{v_i^{j+1} - v_i^j}{\tau} = \sigma \Lambda_h v_i^{j+1} + (1-\sigma)\Lambda_h v_i^j + \varphi_i^j, \quad i = \overline{1, m-1}, \quad j = \overline{0, n-1}. \quad (8)$$

Здесь φ_i^j – сеточная аппроксимация функции $f(x, t)$.

Начальные и граничные условия задаются как и ранее в схемах (4), (7).

Параметр $\sigma \in [0, 1]$ в (8) называют весом. Разностная схема (8) – схема с весами (семейство схем с весами).

При $\sigma = 0$, $\varphi_i^j = f(x_i, t_j)$ получаем явную схему (4).

При $\sigma = 1$, $\varphi_i^j = f(x_i, t_{j+1})$ выделяется чисто неявная схема (7).

Если $\sigma = \frac{1}{2}$, $\varphi_i^j = f(x_i, t_j + \frac{1}{2}\tau)$, то схема (8) называется симметричной.

Введем погрешность аппроксимации для схемы (8)

$$\Delta_i^j = \frac{u_i^{j+1} - u_i^j}{\tau} - \sigma \Lambda_h u_i^{j+1} - (1-\sigma)\Lambda_h u_i^j - \varphi_i^j.$$

Используя разложение функции $u(x, t)$ в ряд Тейлора в точке $(x_i, t_j + \frac{1}{2}\tau)$, нетрудно получить следующее выражение для погрешности

$$\Delta_i^j = f(x_i, t_j + \frac{1}{2}\tau) - \varphi_i^j + \tau(\sigma - \frac{1}{2})\Lambda_h \frac{\partial u(x_i, t_j + \frac{1}{2}\tau)}{\partial t} \; + \; O(\tau^2 + h^2).$$

Для повышения порядка аппроксимации положим $\sigma = \frac{1}{2}$, $\varphi_i^j = f(x_i, t_j + \frac{1}{2}\tau)$. Тогда $\Delta_i^j = O(\tau^2 + h^2)$.

Таким образом, симметричная схема аппроксимирует задачу (1), (2) с порядком $(\tau^2 + h^2)$.

При $\sigma = 0 \vee 1$ порядок аппроксимации – $(\tau + h^2)$.

5. Метод прямых

В заключение опишем метод прямых для решения краевой задачи (1), (2)

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \tag{1}$$

$$u(x,0) = u_0(x), \quad u(0,t) = \mu_1(t), \quad u(X,t) = \mu_2(t), \tag{2}$$

$$\mathcal{D} = \{(x,t) : \ 0 \le x \le X, \quad 0 \le t \le T\}.$$

Рассмотрим *первый вариант* метода прямых. Введем сетку по переменной $x : x_i = ih$, $i = \overline{0,m}$, $mh = X$ и покроем область $\mathcal{D}$ семейством прямых $x = x_i$, параллельных оси t. Обозначим через $v_i(t)$, $t \in [0,T]$ – приближенное решение вдоль прямой $x = x_i$, $i = \overline{0,m} : v_i(t) \approx u(x_i,t)$, $t \in T$.

Рассмотрим уравнение (1) при $x = x_i$, используя разностную аппроксимацию второй производной для $i = \overline{1,m-1}$

$$\frac{\partial^2 u(x_i,t)}{\partial x^2} \approx \frac{u(x_{i-1},t) - 2u(x_i,t) + u(x_{i+1},t)}{h^2}.$$

В результате получаем дифференциально-разностную задачу относительно функций $v_i(t)$

$$\dot{v}_i(t) = \frac{v_{i-1}(t) - 2v_i(t) + v_{i+1}(t)}{h^2} + f(x_i,t), \quad i = \overline{1,m-1},$$

$$v_i(0) = u_0(x_i), \quad i = \overline{0,m},$$

$$v_0(t) = \mu_1(t), \quad v_m(t) = \mu_2(t), \quad t \in [0,T].$$

Получили *задачу Коши* для системы обыкновенных дифференциальных уравнений относительно функций $v_1(t), \ldots, v_{m-1}(t)$, $t \in [0,T]$.

Опишем *второй вариант* метода прямых. Введем сетку по переменной $t : t_j = j\tau$, $j = \overline{0,n}$, $n\tau = T$. Тогда область $\mathcal{D}$ покрывается семейством прямых $t = t_j$, параллельных оси x. Рассмотрим уравнение (1) при $t = t_j$ с использованием разностной аппроксимации производной по t

$$\frac{\partial u(x,t_j)}{\partial t} \approx \frac{u(x,t_j) - u(x,t_{j-1})}{\tau}, \quad j = \overline{1,n}.$$

Пусть $v_j(x) \approx u(x,t_j)$, $j = \overline{0,n}$ – приближенное решение вдоль прямой $t = t_j$.

Дифференциально-разностная задача имеет вид

$$\frac{v_j(x) - v_{j-1}(x)}{\tau} = v_j''(x) + f(x,t_j), \quad j = \overline{1,n},$$

$$v_0(x) = u_0(x), \quad x \in [0, X],$$
$$v_j(0) = \mu_1(t_j), \quad v_j(X) = \mu_2(t_j), \quad j = \overline{0, n}.$$

Для каждого $j = \overline{1, n}$ получаем *краевую задачу* для обыкновенного дифференциального уравнения второго порядка относительно функции $v_j(x)$

$$v_j''(x) - \frac{1}{\tau} v_j(x) = -\frac{1}{\tau} v_{j-1}(x) - f(x, t_j)$$

($v_{j-1}(x)$ – известная функция).

Библиографический список

1. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. *Численные методы.* – М.: Наука, 1987.

2. Самарский А.А., Гулин А.В. *Численные методы.* – М.: Наука, 1989.

3. Крылов В.И., Бобков В.В. Монастырный П.И. *Вычислительные методы, т.I, II.* – М.: Наука, 1976, 1977.

4. Вержбицкий В.М. *Основы численных методов.* – М.: Высшая школа, 2002.

5. Срочко В.А., Захарченко В.С. *Численные методы алгебры.* – Иркутск: РИО ИГУ, 1997.

6. Срочко В.А., Васильева В.Н., Марченко Л.В. *Основы теории интерполирования.* – Иркутск: РИО ИГУ, 1998.

7. Бахвалов Н.С., Лапин А.В., Чижонков Е.В. *Численные методы в задачах и упражнениях.* – М.: Высшая школа, 2000.

8. Самарский А.А., Вабищевич П.Н., Самарская Е.А. *Задачи и упражнения по численным методам.* – М.: Эдиториал УРСС, 2000.

9. Антоник В.Г., Захарченко В.С. *Численные методы. Учебное пособие по практическим занятиям.* – Иркутск: РИО ИГУ, 2000.