ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

§1. Необходимые сведения из линейной алгебры

Предварительно вспомним материал линейной алгебры, относящийся к понятиям нормы векторов и матриц, а также некоторые свойства матриц. Далее будем полагать, что компоненты векторов и элементы матриц являются действительными числами.

Нормой вектора $x=\left(x_1,x_2,...,x_n\right)^T$ называется число, обозначаемое $\|x\|$ и удовлетворяющее следующим условиям:

1. $\|x\|\geq 0, \|x\|=0 \Leftrightarrow x=0$;
2. $\|\alpha x\|=|\alpha|\cdot\|x\|$, α – действительное число;
3. $\|x+y\|\leq\|x\|+\|y\|$ (неравенство треугольника).

Примерами норм вектора являются:

$\|x\|_1=\sum_{i=1}^{n}|x_i|$;

$\|x\|_2=\left(\sum_{i=1}^{n}x_i^2\right)^{\frac{1}{2}}$ – евклидова норма;

$\|x\|_\infty=\|x\|_c=\max_{i=1,n}|x_i|$ – равномерная норма.

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

$$\rho(x,y)=\|x-y\|.$$

Поэтому далее, как равнозначными, будем пользоваться, например, терминами: пространство с евклидовой нормой, пространство с евклидовой метрикой.

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

1. $\|A\|\geq 0, \|A\|=0 \Leftrightarrow A=O$;
2. $\|\alpha A\|=|\alpha|\cdot\|A\|$, α – действительное число;
3. $\|A+B\|\leq\|A\|+\|B\|$ (неравенство треугольника);
4. $\|A\cdot B\|\leq\|A\|\cdot\|B\|$,

Норма матрицы $\|A\|$ ***согласована*** с нормой вектора $\|x\|$, если $\|A\cdot x\|\leq\|A\|\cdot\|x\|$.

Использование согласованных норм позволяет получать требуемые оценки для погрешности методов последовательных приближений. Норма матрицы A называется ***подчиненной норме вектора x,*** если она вводится следующим образом

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|.$$

Нетрудно видеть, что подчиненная норма согласована с соответствующей метрикой векторного пространства, так как

$$\frac{\|Ax\|}{\|x\|} \leq \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \|A\|,$$

отсюда $\|A \cdot x\| \leq \|A\| \cdot \|x\|$. Чтобы получить конкретное выражение подчиненной нормы матрицы через ее элементы, надо найти $\sup\limits_{\|x\|=1} \|Ax\|$. Найдем, например, $\|A\|_c$ – норму матрицы, подчиненную равномерной метрике векторного пространства. Пусть $y = Ax$. Тогда

$$\|Ax\|_c = \|y\|_c = \max_{i=1,n} |y_i| = \max_{i=1,n} \left| \sum_{j=1}^{n} A_{ij} x_j \right|.$$

Так как

$$\left| \sum_{j=1}^{n} A_{ij} x_j \right| \leq \sum_{j=1}^{n} |A_{ij}| \cdot |x_j| \leq \max_{j=1,n} |x_j| \cdot \sum_{j=1}^{n} |A_{ij}| = \|x\|_c \sum_{j=1}^{n} |A_{ij}|,$$

то для векторов с $\|x\|_c = 1$

$$\|Ax\|_c \leq \max_{i=1,n} \sum_{j=1}^{n} |A_{ij}|.$$

Можно доказать, что существует вектор, на котором найденная верхняя оценка достигается, то есть что она является точной верхней гранью для оцениваемых величин. Это означает, что

$$\|A\|_c = \max_{i=1,n} \sum_{j=1}^{n} |A_{ij}|.$$

Аналогичным образом устанавливается, что $\|A\|_1 = \max\limits_{j=1,n} \sum\limits_{i=1}^{n} |A_{ij}|$ – норма, подчиненная метрике $\|x\|_1$ векторного пространства. Можно показать также, что норма матрицы, подчиненная евклидовой метрике векторного пространства, определяется следующим образом

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

где $\rho\left(A^T A\right) = \max\left(\lambda_{A^T A}\right)$ – спектральный радиус матрицы $A^T A$ (наибольшее абсолютное значение собственных значений $A^T A$, A^T – транспонированная к матрице A), $\lambda_{A^T A}$ – собственные значения матрицы $A^T A$. Определенная таким образом норма называется ***спектральной.*** Если при этом матрица A является симметричной, то есть

$$A^T = A, \text{ то } A^T A = A^2,\ \max\left(\lambda_{A^T A}\right) = \max\left(\lambda_{A^2}\right) = \left[\max\left(\lambda_A\right)^2\right)$$

и, следовательно, в этом случае $\|A\|_{sp} = \|A\|_2 = \max|\lambda_A|$ (λ_A – собственные значения самой матрицы A).

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

§2. Прямые методы решения линейных систем уравнений

Задачи линейной алгебры являются самыми распространенными с прикладной точки зрения , и поэтому вызывают неизменно возрастающий интерес и со стороны математиков. Задачи линейной алгебры с точки зрения математиков-вычислителей – это

- ❖ решение систем линейных уравнений;
- ❖ вычисление определителей и обращение матриц;
- ❖ вычисление собственных значений и собственных векторов матриц.

В практических приложениях наиболее часто приходится иметь дело с первой задачей. Проанализировав численные методы, можно заметить, что необходимость решать линейные системы уравнений возникает при реализации большинства методов, с которыми сталкивается инженер в своей прикладной деятельности. Решение второй задачи, в конечном счете, сводится к первой. Последняя же связана с проблемой поиска корней алгебраического полинома n-й степени (n – порядок матрицы). Хотя многие современные методы не требуют выписывания характеристического многочлена.

Постановка задачи: найти решение система n линейных уравнений с n неизвестными

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n, \end{cases}$$

Обозначая через

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & & \cdots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

матрицу коэффициентов системы, через $b = \left(b_1, b_2, \ldots, b_n\right)^T$ столбец (вектор) свободных членов и через $X = \left(x_1, x_2, \ldots, x_n\right)^T$ столбец (вектор) неизвестных,

представим систему линейных уравнений в более компактной матричной форме

$$AX = b.$$

Будем считать, что определитель матрицы системы $\Delta = \det A \neq 0$, то есть решение системы существует и единственно. Известны формулы, дающие в явной форме решение этой СЛАУ. Это формулы Крамера

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

где Δ_i - определитель матрицы, которая получается из матрицы A заменой столбца с номером i столбцом свободных членов. Определители при этом приходится вычислять по формулам, рассматриваемым в курсах линейной алгебры. Например, по определению

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

где последовательности $p_1, p_2, \ldots, p_n$ представляют собой различные перестановки натуральных чисел от 1 до n, $P_n = n!$ – число возможных перестановок, а $[p_1, p_2, \ldots, p_n]$ – число так называемых «беспорядков» в перестановке. Однако в качестве конкретного метода решения исходной системы данные формулы совершенно неприменимы, так как при подсчете каждого определителя по приведенной выше формуле надо вычислить $n!$ слагаемых, что нереально при весьма умеренных n. Например, уже при n=100 имеем $100! \gg 10^{90}$. Если одно слагаемое вычисляется, скажем, за 10^{-6}, то время расчета T составит совершенно фантастическую цифру

$$T \geq 10^{90} \cdot 10^{-6} \approx 3 \cdot 10^{76} \text{ лет.}$$

Фактически же в настоящее время с использованием подходящих методов решаются системы гораздо более высокого порядка (до $n \approx 10^4$). Они разделяются на две группы: ***прямые и итерационные***.

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

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

С помощью точных методов, проделав конечное число операций, можно точно получить искомые значения неизвестных. При этом предполагается, что коэффициенты и правые части системы известны точно, а все вычисления проводятся без округлений. В основном решение таких задач осуществляется поэтапно. На первом этапе систему преобразуют к более простому виду, на втором – решают упрощенную систему и получают значения неизвестных.

Метод Гаусса является почти оптимальным по быстродействию и почти универсальным по отношению к свойствам матрицы A. Этим объясняется его широкое применение. Наибольшее распространение имеют схемы Гаусса с выбором главного элемента: по строке, по столбцу, по всей матрице. Если нет необходимости в выборе главных элементов каким-либо специальным образом, целесообразно применять схему единственного деления, которая уступает всем трем схемам в точности, но выигрывает в простоте вычислений.

Очевидно, что сложность исходной системы определяется структурой матрицы A. Если A - диагональная матрица

$$A = D = diag\left[d_1, d_2, ..., d_n\right] = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & d_n \end{bmatrix},$$

то система распадается на n линейных уравнений, каждое из которых содержит одну неизвестную величину, и проблем с вычислениями не возникает. Просто решается задача и в случае, когда матрица A является треугольной. Пусть, например,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}.$$

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

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

далее,

$$x_m = \frac{1}{a_{mm}}\left(b_m - a_{mn}x_n - a_{mn-1}x_{n-1} - \ldots - a_{mm+1}x_{m+1}\right) \ (m = n-1, n-2, ..., 2, 1).$$

Ввиду большого многообразия методов решения этой задачи в дальнейшем важную роль будет играть такой критерий отбора, как трудоемкость метода, выраженная в количестве требуемых арифметических операций. Оценим объем вычислений, связанный с решением системы с треугольной матрицей. Чтобы вычислить x_n, надо выполнить одну операцию, для вычисления x_{n-1} – три, x_{n-2} – пять и так далее. Нетрудно получить, что общее число операций равно $\Omega = \sum_{m=1}^{n}(2m-1) = n^2.$

Перейдем теперь к обсуждению вариантов метода исключения для системы с матрицей общего вида. Типичная схема реализации метода

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

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

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ \qquad a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)}, \end{cases}$$

где $a_{ij}^{(1)} = a_{ij} + l_i^{(1)}a_{1j}$, $b_i^{(1)} = b_i + l_i^{(1)}b_1$. для i= 2, 3, ... , n, j = 2,3, ... , n.

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

Пример. Пусть требуется решить систему линейных уравнений четвертого порядка вида:

$$\begin{cases} 2x_1 + 2x_2 + 4x_3 - 2x_4 = 10 \\ x_1 + 3x_2 + 2x_3 + x_4 = 17 \\ 3x_1 + x_2 + 3x_3 + x_4 = 18 \\ x_1 + 3x_2 + 4x_3 + 2x_4 = 27. \end{cases}$$

Разделим первое уравнение системы на 2. Из второго уравнения вычитаем первое. Получаем новое второе уравнение. Из третьего уравнения вычитаем первое уравнение, умноженное на 3. Получим новое третье уравнение. Из четвертого уравнения вычитаем первое. Получаем новое четвертое уравнение.
В результате этих преобразований придем к системе уравнений вида:

$$\begin{cases} x_1 + x_2 + 2x_3 - x_4 = 5 \\ \quad 2x_2 \quad\quad + 2x_4 = 12 \\ -2x_2 - 3x_3 + 4x_4 = 3 \\ \quad 2x_2 + 2x_3 + 3x_4 = 22. \end{cases}$$

Далее первое уравнение оставляем без изменений, второе сокращаем на 2 и проводим с ним такие же вычисления, как и с первым уравнением. Преобразованная система примет вид

$$\begin{cases} x_1 + x_2 + 2x_3 - x_4 = 5 \\ \quad x_2 \quad\quad + x_4 = 6 \\ \quad\quad x_3 - 2x_4 = -5 \\ \quad\quad 2x_3 + x_4 = 10. \end{cases}$$

Наконец, оставляя без изменения первые два уравнения, работаем с третьим. Получим систему уравнений диагонального вида:

$$\begin{cases} x_1 + x_2 + 2x_3 - x_4 = 5 \\ \quad x_2 \quad\quad + x_4 = 6 \\ \quad\quad x_3 - 2x_4 = -5 \\ \quad\quad\quad 5x_4 = 20. \end{cases}$$

Теперь из четвертого уравнения находим $x_4 = 4$. Подставляя его в третье уравнение, выражаем $x_3 = 3$. Продолжая подставлять найденные неизвестные в оставшиеся уравнения, получим $x_2 = 2$ и $x_1 = 1$.

Итак, все искомые неизвестные найдены. Проверкой устанавливается правильность полученных результатов.

Ответ: $x_1 = 1;\ \ x_2 = 2;\ \ x_3 = 3;\ \ x_4 = 4$.

В данном алгоритме сначала обеспечивается преобразование матрицы *A* к треугольному виду (прямой ход метода), а затем определяются корни системы линейных уравнений (обратный ход метода).

Естественно возникает вопрос какова трудоемкость этого метода? Обратный ход, как мы видели, требует выполнения n^2 арифметических операций. В то же время из общих соображений ясно, что полное число операций должно быть пропорционально n^3 (мы ищем *n* неизвестных, каждая из которых должна зависеть от n^2 элементов матрицы). На самом деле, аккуратный подсчет операций, выполняемых на первом этапе исключения, приводит к следующей оценке объема вычислительной работы в рамках прямого хода

$$\Omega = \frac{1}{6} n(n-1)(4n+7) \ .$$

При больших n количество операций $\Omega \sim \frac{2}{3} n^3$. Это вполне приемлемая величина (при $n \sim 10^3$, быстродействии ЭВМ порядка 10^6 операций в секунду требуемое для расчета время порядка одного часа). Мы рассмотрели простейшую схему исключения (этот метод называют также методом Гаусса) и далеко не лучшую. Рассмотрим систему

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

Первый метод. Исключая x_1 из первого уравнения: $x_1 = 10^7 x_2 - 10^7$, и подставляя это выражение во второе уравнение, получаем $x_2 = \frac{10^7 + 4}{10^7 + 2}$. Проведя вычисления с семью значащими цифрами (при работе в режиме с ординарной точностью), получаем $x_2 = 1.000000$, $x_1 = 0.000000$, что неверно, как видно из второго уравнения.

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

Видим, что в первом варианте метода исключения результаты получились совершенно неверными. «Механизм» возникновения больших погрешностей: деление на малые числа, появление больших (по величине) промежуточных результатов, потеря точности при вычитании больших (близких друг к другу) чисел.

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

Выбор главного элемента по столбцам. Перед исключением x_1 отыскивается $\max\limits_{i=1,n} |a_{i1}|$. Допустим, максимум соответствует $i = i_0$. Тогда первое уравнение в исходной системе меняем местами с i_0-м уравнением (для ЭВМ эта процедура связана с перестановкой двух строк расширенной матрицы). После этого осуществляется первый шаг исключения. Затем перед исключением

x_2 из оставшихся уравнений отыскивается $\max\limits_{i=2,n}\left|a_{i2}^{(1)}\right|$ и осуществляется соответствующая перестановка уравнений и так далее.

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

Метод Гаусса с выбором главного элемента состоит в следующем.

1. В системе сначала выбирают уравнение, в котором содержится наибольший по абсолютной величине коэффициент системы (это и будет главный элемент).
2. Делят данное уравнение на этот коэффициент.
3. Так же, как и в простейшей схеме метода Гаусса, исключают из остальных уравнений то неизвестное, при котором был наибольший коэффициент в выбранном уравнении (для удобства наибольший коэффициент помещают в первую строку и первый столбец матрицы, над которой производятся соответствующие преобразования).
4. Уравнение с главным элементом оставляем неизменным и ищем наибольший по абсолютной величине коэффициент в остальных уравнениях (новый главный элемент).
5. На новый главный элемент делим уравнение, в котором он находится, и исключаем из остальных уравнений соответствующие неизвестные.
6. Продолжаем этот процесс до тех пор, пока не останется одно уравнение с одним неизвестным, то есть пока система не будет приведена к диагональному виду.

Рассмотренные модификации метода Гаусса позволяют, как правило, существенно уменьшить неблагоприятное влияние погрешностей округления на результаты расчета. Впрочем, в прикладных задачах довольно часто приходится сталкиваться с линейными системами, при решении которых можно не заботиться о «вредном» воздействии неустранимых погрешностей на решение, спокойно применяя простейшую схему гауссова исключения (без выбора главного элемента). Это системы, для матриц которых выполнено ***условие диагонального преобладания***

$$\left|a_{ii}\right| > \sum_{j\neq i}^{n}\left|a_{ij}\right|, i = 1,2,...,n.$$

Можно показать, что условие диагонального преобладания остается справедливым после каждого шага исключений в процессе приведения матрицы к треугольному виду, то есть

$$\left|a_{ii}^{(k)}\right| > \sum_{\substack{j=k,\\ j\neq i}}^{n} \left|a_{ij}^{(k)}\right|, i = k, k+1, k+2, \ldots, n$$

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

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

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

Необходимость решать подобные системы довольно часто возникает в качестве составного элемента при реализации различных численных методов решения дифференциальных уравнений. Применим для решения системы простейшую схему метода Гаусса (без выбора главного элемента). Поделив первое уравнение на b_1, перепишем его в виде

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

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

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

Переписываем последнее уравнение в виде

$$x_2 + p_2x_3 = q_2, \left(p_2 = \frac{c_2}{b_2 - p_1a_2}, q_2 = \frac{f_2 - q_1a_2}{b_2 - p_1a_2} \right).$$

Исключаем с помощью этого соотношения x_2 из третьего уравнения системы и так далее. В итоге приходим к системе уравнений с двухдиагональной матрицей, состоящей из элементов

$$x_i + p_ix_{i+1} = q_i \quad \left(i = 1, 2, \ldots, n-1\right)$$

$$x_n = q_n,$$

Коэффициенты и правые части которой вычисляются по формулам

$$p_1 = \frac{c_1}{b_1}, \quad p_i = \frac{c_i}{b_i - p_{i-1}a_i}, \; i = 2,3,\ldots,n-1,$$

$$q_1 = \frac{f_1}{b_1}, \quad q_i = \frac{f_i - q_{i-1}a_i}{b_i - p_{i-1}a_i}, \; i = 2,3,\ldots,n.$$

На основании проделанных выкладок можно сформулировать алгоритм решения системы с трехдиагональной матрицей, состоящий из двух этапов:

1) по вышеприведенным формулам вычисляются массивы p_i и q_i;

2) по формулам

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

вычисляется искомое решение.

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

Ввиду опасности обращения знаменателя в формулах в ноль возникает вопрос об условиях применимости метода прогонки. Таковыми являются, в частности, условия диагонального преобладания в исходной матрице: $|b_i| > |a_i| + |c_i|$ для всех i. В этом случае $|p_1| < 1$. Используя метод математической индукции, несложно показать, что $|p_i| < 1$ для всех i = 2, 3, ... , n – 1, что гарантирует отличный от нуля знаменатель.

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

§3. О неустранимой погрешности при решении линейных систем методом исключения

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

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

$$AX = b$$

решается задача

$$\left(A+\delta A\right)\left(X+\delta X\right)=\left(b+\delta b\right).$$

Здесь δA – матрица возмущений, моделирующих ошибки коэффициентов исходных уравнений, δb – соответственно возмущения правых частей, δX – обусловленный этими возмущениями вектор «ошибок.

Переписывая последнее матричное уравнение в виде

$$A\cdot X + A\cdot\delta X + \delta A\cdot X + \delta A\cdot\delta X = b + \delta b$$

и вычитая из последнего соотношение $AX = b$, приходим к системе уравнений

$$A\cdot\delta X + \delta A\cdot\delta X = \delta b - \delta A\cdot X$$

которая описывает зависимость δX от возмущений (ошибок) исходных данных. Далее будем полагать, что возмущения коэффициентов уравнений δA и погрешности решения δX в достаточной мере малы, так что в последнем уравнении можно пренебречь квадратичными членами $\delta A\cdot\delta X$. Тогда интересующую нас ошибку δX можно представить в виде

$$\delta X \simeq A^{-1}\left(\delta b - \delta A\cdot X\right).$$

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

$$\left\|\delta X\right\| \simeq \left\|\left(\delta b - \delta A\cdot X\right)\right\| \le \left\|A^{-1}\right\|\left(\left\|\delta b\right\| + \left\|\delta A\right\|\cdot\left\|X\right\|\right) =$$

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

Учитывая, что $\left\|b\right\| = \left\|AX\right\| \le \left\|A\right\|\cdot\left\|X\right\|$, получаем далее

$$\left\|\delta X\right\| \le \left\|A^{-1}\right\|\left(\left\|A\right\|\cdot\left\|X\right\|\frac{\left\|\delta b\right\|}{\left\|b\right\|} + \left\|A\right\|\cdot\left\|X\right\|\frac{\left\|\delta A\right\|}{\left\|A\right\|}\right) =$$

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

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

$$\frac{\left\|\delta X\right\|}{\left\|X\right\|} \le \mu_A\left(\frac{\left\|\delta b\right\|}{\left\|b\right\|} + \frac{\left\|\delta A\right\|}{\left\|A\right\|}\right),$$

где $\mu_A = \left\|A^{-1}\right\| \cdot \|A\|$. Значение μ_A называется ***числом обусловленности матрицы*** A. Именно эта величина определяет, насколько сильно погрешности входных данных могут повлиять на решение системы. Так ка всегда имеем $E = A^{-1} \cdot A,$ то $1 = \|E\| = \left\|A^{-1}A\right\| \le \left\|A^{-1}\right\| \cdot \|A\| = \mu_A$ и всегда $\mu_A \ge 1$. Если значение μ_A является умеренным ($\mu_A \sim 1 \div 10$), ошибки входных данных слабо сказываются на решении и система в этом случае называется ***хорошо обусловленной***. Если μ_A велико ($\mu_A \ge 10^3$), система ***плохо обусловлена***, решение ее сильно зависит от ошибок в правых частях и коэффициентах. Вообще говоря, более точное представление о хорошей или плохой обусловленности системы должно опираться на требования, предъявляемые к решению. Если, к примеру, погрешность входных данный $\sim 10^{-6}$, а допустимая погрешность решения $\sim 10^{-2}$, то даже при $\mu_A \sim 10^4$ систему можно считать хорошо обусловленной.

Хотелось бы подчеркнуть, что данное свойство (обусловленность), выражаемое неравенством

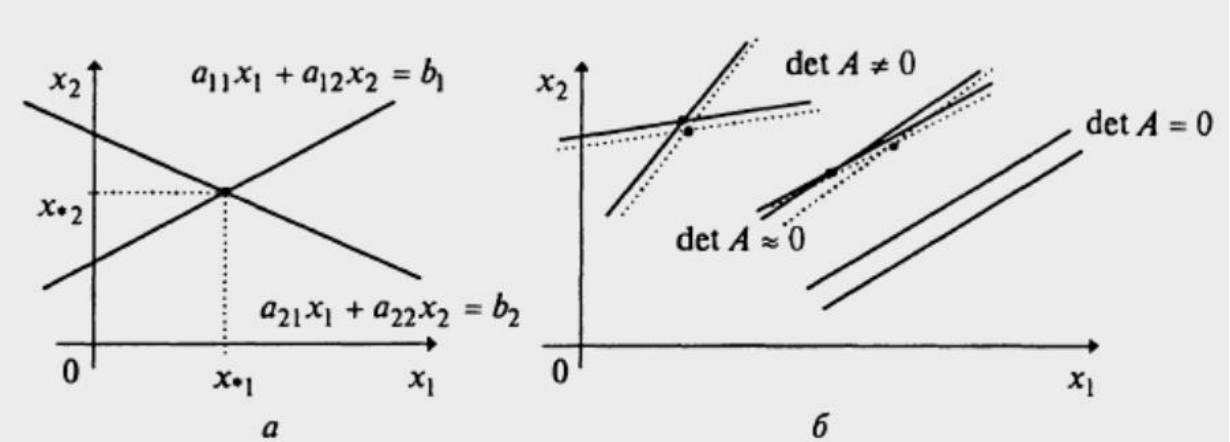
$$\frac{\|\delta X\|}{\|X\|} \le \mu_A \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right),$$

никак не связано с предполагаемым методом решения системы, а является изначальной характеристикой решаемой задачи.

Характер задачи и точность получаемого решения в большой степени зависят от ее обусловленности, являющейся важнейшим математическим понятием, влияющим на выбор метода ее решения. Поясним это понятие обусловленности на примере двумерной задачи:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1, \\ a_{21}x_1 + a_{22}x_2 = b_2. \end{cases}$$

Точным решением этой задачи является вектор $x_* = \left(x_{*1}, x_{*2}\right)^T$, компоненты которого определяются координатами точки пересечения двух прямых, соответствующих уравнениям $a_{11}x_1 + a_{12}x_2 = b_1$, $a_{21}x_1 + a_{22}x_2 = b_2$ (рис. *а*).



На рисунке *б* применительно к трем наборам входных данных, заданных с некоторыми погрешностями и соответствующих различным системам линейных уравнений, иллюстрируется характер обусловленности системы. Если определитель системы A существенно отличен от нуля, то точка пересечения пунктирных прямых, смещенных относительно сплошных прямых из-за погрешностей задания A и b, сдвигается несильно. Это свидетельствует о хорошей обусловленности системы. При $\det A \approx 0$ небольшие погрешности в коэффициентах могут привести к большим погрешностям в решении (плохо обусловленная матрица), поскольку прямые близки к параллельным. При $\det A=0$ прямые параллельны или они совпадают, и тогда решение задачи не существует или оно не единственно.

Следует иметь в виду, что реализация хорошей или плохой обусловленности в корректной и некорректной задачах напрямую связана с вытекающей отсюда численной устойчивостью или неустойчивостью. При этом для решения некорректных задач обычно применяются специальные методы или математические преобразования этих задач к корректным.

Пример. Рассмотрим систему

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

Ее решение $x_1 = x_2 = 1.$

Исказим теперь слегка ее правые части

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

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

Чтобы сопоставить полученные результаты с оценкой

$$\frac{\|\delta X\|}{\|X\|} \le \mu_A \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right),$$

будем пользоваться следующими согласованными нормами для векторов и матриц

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

Для рассмотренного примера имеем

$$b = \begin{pmatrix} 199 \\ 199 \end{pmatrix},\ \delta b = \begin{pmatrix} -0.01 \\ 0.01 \end{pmatrix},\ \text{то есть } \|b\| = 199,\ \|\delta b\| = 0.01.$$

Относительная погрешность $\dfrac{\|\delta b\|}{\|b\|} \approx \dfrac{1}{2} \cdot 10^{-4}.$ Это очень малая величина.

Далее, вычислим число обусловленности. Так как

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

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

Согласно оценке

$$\frac{\|\delta X\|}{\|X\|} \le \mu_A \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right) \approx 4 \cdot 10^4 \cdot \frac{10^{-4}}{2} = 2,$$

что, как видно, согласуется с результатами решения рассмотренных систем.

Плохо обусловленные системы вызывают определенные трудности при решении. Из оценки $\frac{\|\delta X\|}{\|X\|}$ следует, что решение их сильно зависит от ошибок входных данных, и даже при отсутствии ошибок во входных величинах может произойти значительная (если не полная) потеря точности на стадии вычислений по методу Гаусса за счет погрешностей округлений.

§4. Метод LU-разложения для решения СЛАУ

Рассмотрим ещё один метод решения системы линейных алгебраических уравнений $AX = b$. Метод опирается на возможность представления квадратной матрицы A системы в виде произведения двух треугольных матриц $A = L \cdot U$, где L – нижняя, а U – верхняя треугольные матрицы,

$$L = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix},\ U = \begin{pmatrix} 1 & u_{12} & \cdots & u_{1n} \\ 0 & 1 & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

С учётом этого система $AX=b$ представляется в форме $L \cdot U \cdot X = b$ и её решение сводится к последовательному решению двух простых систем с треугольными матрицами. В итоге процедура решения состоит из двух этапов.

Прямой ход. Произведение UX обозначим через Y. В результате решения системы $LY=b$ находится вектор Y.

Обратный ход. В результате решения системы $UX=Y$ находится решение задачи – столбец X.

В силу треугольности матриц L и U решения обеих систем находятся рекуррентно (как в обратном ходе метода Гаусса).

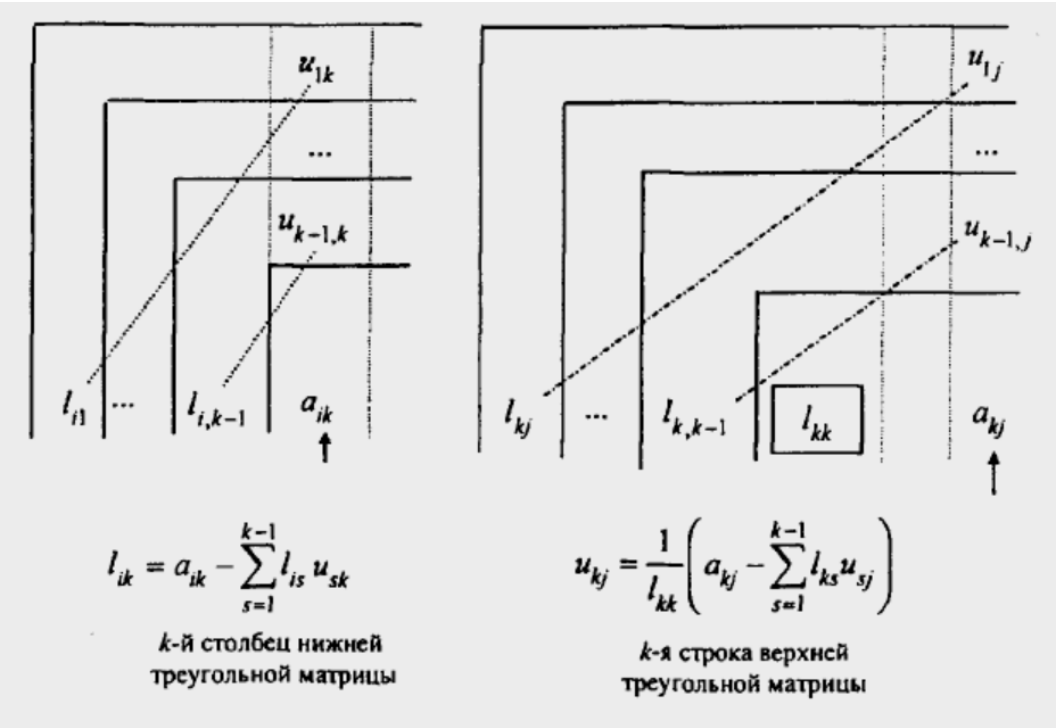
Из общего вида элемента произведения $A=LU$, а также структуры матриц L и U следуют формулы для определения элементов этих матриц

$$l_{ij} = a_{ij} - \sum_{s=1}^{j-1} l_{is} u_{sj},\ i \ge j,\ u_{ij} = \frac{1}{l_{ii}} \left(a_{ij} - \sum_{s=1}^{i-1} l_{is} u_{sj} \right),\ i < j.$$

Результат представления матрицы A в виде произведения двух треугольных матриц (операции факторизации) удобно хранить в одной матрице следующей структуры

$$\begin{pmatrix} l_{11} & u_{12} & \cdots & u_{1n} \\ l_{21} & l_{22} & \cdots & u_{2n} \\ l_{31} & l_{32} & \cdots & u_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix}.$$

Вычисления на k-м шаге метода LU-разложения удобно производить, пользуясь двумя схемами, изображенными на рисунке



Замечание. Всякую квадратную матрицу A имеющую отличные от нуля главные миноры

$$\Delta_1 = a_{11} \neq 0, \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0, \cdots, \Delta_n = |A| \neq 0$$

можно представить в виде LU-разложения, причем это разложение будет единственным. Это условие выполняется для матриц с преобладанием диагональных элементов, у которых

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

Алгоритм метода LU-разложения:

1. Выполнить операцию факторизации исходной матрицы A, и получить матрицы L и U.

2. Решить систему $L \cdot Y = b$.

3. Решить систему $U \cdot X = Y$.

Пример. Решить систему линейных алгебраических уравнений методом *LU*-разложения

$$\begin{cases} 2x_1 + x_2 + 4x_3 = 16, \\ 3x_1 + 2x_2 + x_3 = 10, \\ x_1 + 3x_2 + 3x_3 = 16. \end{cases}$$

Решение. 1. Выполним операцию факторизации:

$$\begin{pmatrix} 2 & 1 & 4 \\ 3 & 2 & 1 \\ 1 & 3 & 3 \end{pmatrix} \xrightarrow{k=1} \begin{pmatrix} 2 & 0{,}5 & 2 \\ 3 & 2 & 1 \\ 1 & 3 & 3 \end{pmatrix} \xrightarrow{k=2} \begin{pmatrix} 2 & 0{,}5 & 2 \\ 3 & 0{,}5 & -10 \\ 1 & 2{,}5 & 3 \end{pmatrix} \xrightarrow{k=3} \begin{pmatrix} 2 & 0{,}5 & 2 \\ 3 & 0{,}5 & -10 \\ 1 & 2{,}5 & 26 \end{pmatrix}.$$

При k=1: $l_{11} = a_{11} = 2; l_{21} = a_{21} = 3; l_{31} = a_{31} = 1; u_{12} = \frac{1}{2} a_{12} = 0{,}5; u_{13} = \frac{1}{2} a_{13} = 2.$

При k=2: $l_{22} = a_{22} - l_{21} \cdot u_{12} = 2 - 3 \cdot 0{,}5 = 0{,}5; l_{32} = a_{32} - l_{31} \cdot u_{12} = 3 - 1 \cdot 0{,}5 = 2{,}5;$

$u_{23} = \frac{1}{l_{22}} (a_{23} - l_{21} \cdot u_{13}) = \frac{1}{0{,}5} (1 - 3 \cdot 2) = -10.$

При k=2: $l_{33} = a_{33} - l_{31} \cdot u_{13} - l_{32} \cdot u_{23} = 3 - 1 \cdot 2 - 2{,}5 \cdot (-10) = 26$

В результате получены две треугольные матрицы:

$$L = \begin{pmatrix} 2 & 0 & 0 \\ 3 & 0{,}5 & 0 \\ 1 & 2{,}5 & 26 \end{pmatrix}, \quad U = \begin{pmatrix} 1 & 0{,}5 & 2 \\ 0 & 1 & -10 \\ 0 & 0 & 1 \end{pmatrix}.$$

Определитель матрицы A находится в результате перемножения диагональных элементов матрицы L: $\det A = 2 \cdot 0{,}5 \cdot 26 = 26$.

2. Решим систему $L \cdot Y = b$

$$\underbrace{\begin{pmatrix} 2 & 0 & 0 \\ 3 & 0{,}5 & 0 \\ 1 & 2{,}5 & 26 \end{pmatrix}}_{L} \cdot \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 16 \\ 10 \\ 16 \end{pmatrix}}_{b} \Rightarrow \begin{cases} 2y_1 = 16, \\ 3y_1 + 0{,}5y_2 = 10, \\ y_1 + 2{,}5y_2 + 26y_3 = 16 \end{cases} \Rightarrow \begin{cases} y_1 = 8, \\ y_2 = -28, \\ y_3 = 3. \end{cases}$$

3. Решим систему $U \cdot X = Y$:

$$\underbrace{\begin{pmatrix} 1 & 0{,}5 & 2 \\ 0 & 1 & -10 \\ 0 & 0 & 1 \end{pmatrix}}_{U} \cdot \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{X} = \underbrace{\begin{pmatrix} 8 \\ -28 \\ 3 \end{pmatrix}}_{Y} \Rightarrow \begin{cases} x_1 + 0{,}5x_2 + 2x_3 = 8, \\ x_2 - 10x_3 = -28, \\ x_3 = 3 \end{cases} \Rightarrow \begin{cases} x_1 = 1, \\ x_2 = 2, \\ x_3 = 3. \end{cases}$$

Ответ: $X_* = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^T$.

Недостатки прямых методов:

– необходимость хранения в оперативной памяти компьютера сразу всей матрицы (при большой размерности матрицы требуется много памяти);

– накопление погрешностей в процессе решения, что особенно опасно для больших систем, а также для плохо обусловленных систем, весьма чувствительных к погрешностям. В связи с этим прямые методы используют обычно для не слишком больших ($n \le 1000$) систем с плотно заполненной матрицей и не близким к нулю определителем.

§4. Итерационные методы решения линейных систем

§4.1. Метод простых итераций

Переходим к обсуждению итерационных методов (то есть методов последовательных приближений) решения линейных систем уравнений $AX = b.$ при этом, как и ранее считаем, что решение этой системы существует и единственно. Различные варианты метода простых итераций связаны с переходом от исходной системы к эквивалентной системе

$$X=PX+g.$$

Итерационный процесс, опираясь на это равенство, строим очевидным образом

$$X^{(k)} = PX^{(k-1)} + g \ ,$$

Здесь k – номер приближения, $X^{(0)}$ задано.

Условия сходимости метода последовательных приближений формулируются в следующих теоремах.

Теорема. Для сходимости итераций к решению исходной системы достаточно, чтобы в какой-либо норме выполнялось условие $\|P\| \le q < 1$. Тогда независимо от выбора $X^{(0)}$

$$\left\|X^{(k)} - X^*\right\| \le q^k \left\|X^{(0)} - X^*\right\|,$$

где X^* – точное решение исходной системы $AX = b.$

Доказательство. Подстановка точного решения X^* в систему $X=PX+g$ обращает последнюю в тождество:

$$X^* = PX^* + g.$$

Вычитая его из $X^{(k)} = PX^{(k-1)} + g$, получим

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

где $X^{(k)} - X^*$ – вектор погрешности (или просто погрешность) k-го приближения. Оценивая величину погрешности по какой-либо норме (с которой согласована норма матрицы, фигурирующая в условии теоремы), получаем

$$\left\|X^{(k)} - X^*\right\| \le \|P\| \cdot \left\|X^{(k-1)} - X^*\right\| \le q\left\|X^{(k-1)} - X^*\right\| \le$$

$$\le q^2\left\|X^{(k-2)} - X^*\right\| \le \ldots \le q^n\left\|X^{(0)} - X^*\right\|.$$

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

Теорема. Для сходимости итераций $X^{(k)} = PX^{(k-1)} + g$ к решению системы $AX = b.$ необходимо и достаточно, чтобы все собственные значения матрицы P по абсолютной величине были меньше единицы.

Хотя вторая теорема дает более общие условия сходимости метода простых итераций, чем первая, однако ею воспользоваться сложнее, так как нужно предварительно вычислить границы собственных значений матрицы P или сами собственные значения.

Преобразование системы $AX = b$ к виду $X=PX+g$ с матрицей P, удовлетворяющей условиям сходимости, может быть выполнено несколькими способами. Приведем способы, используемые наиболее часто.

1. Уравнения, входящие в систему $AX = b$, переставляются так, чтобы выполнялось условие преобладания диагональных элементов (для той же цели можно использовать другие элементарные преобразования). Затем первое уравнение разрешается относительно x_1, второе — относительно x_2 и так далее. При этом получается матрица P с нулевыми диагональными элементами. Например, система

$$\begin{cases} -2,8x_1 + x_2 + 4x_3 = 60, \\ 10x_1 - x_2 + 8x_3 = 10, \\ -x_1 + 2x_2 - 0,6x_3 = 20 \end{cases}$$

с помощью перестановки уравнений приводится к виду

$$\begin{cases} 10x_1 - x_2 + 8x_3 = 10, \\ -x_1 + 2x_2 - 0,6x_3 = 20, \\ -2,8x_1 + x_2 + 4x_3 = 60, \end{cases}$$

где $|10| > |1| + |8|$, $|2| > |-1| + |-0,6|$, $|4| > |-2,8| + |1|$, то есть диагональные элементы преобладают. Выражая x_1 из первого уравнения, x_2 – из второго, а x_3 – из третьего, получаем систему вида $X=PX+g$:

$$\begin{cases} x_1 = 0x_1 + 0,1x_2 - 0,8x_3 + 1, \\ x_2 = 0,5x_1 + 0x_2 + 0,3x_3 + 10, \\ x_3 = 0,7x_1 - 0,25x_2 + 0x_3 + 15, \end{cases} \text{ где } P = \begin{pmatrix} 0 & 0,1 & -0,8 \\ 0,5 & 0 & 0,3 \\ 0,7 & -0.25 & 0 \end{pmatrix}, \; g = \begin{pmatrix} 1 \\ 10 \\ 15 \end{pmatrix}.$$

Заметим, что $\|P\|_1 = \max\{0,9; 0,8; 0,95\} < 1$, то есть условие теоремы выполнено. Существуют и более сложные методы приведения матрицы системы к матрице с диагональным преобладанием.

Алгоритм метода простых итераций

1. Преобразовать систему $AX = b$ к виду $X=PX+g$.

2. Задать начальное приближение решения $X^{(0)}$ произвольно (на основании каких-либо предположений) или положить $X^{(0)} = g$, а также малое положительное число ε (точность). Положить k=1.

3. Вычислить следующее приближение $X^{(k)} = PX^{(k-1)} + g$.

4. Если выполнено условие $\left\| X^{(k+1)} - X^{(k)} \right\| < \varepsilon$, процесс завершить и в качестве приближенного решения задачи принять $X^* \approx X^{(k+1)}$. Иначе положить k=k+1 и перейти к пункту 3 алгоритма.

Опираясь на оценку первой теоремы $\left\| X^{(k)} - X^* \right\| \le q^k \left\| X^{(0)} - X^* \right\|$, можно получить априорную (доопытную, предполагаемую до начала процесса вычислений) оценку числа приближений, гарантирующую, что вычисленное решение будет отличаться от точного не более, чем на малое число ε:

$$\left\| X^{(k)} - X^* \right\| \le q^k \left\| X^{(0)} - X^* \right\| \le \varepsilon .$$

Разрешая последнее неравенство относительно k, получаем

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

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

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

Метод Гаусса (в простейшей интерпретации), как мы видели, при количестве неизвестных n существенно больших единицы требует выполнения приблизительно $\frac{2}{3}n^3$ арифметических операций. Метод итераций реализуется приблизительно за $\left(2n^2\right)K$ операций ($2n^2$ умножений и сложений связано с умножением матрицы P на вектор $X^{(k-1)}$, K – число приближений). Если допустимая погрешность достигается при $K < \frac{n}{3}$, то метод итераций становится предпочтительней. В задачах, с которыми практически приходится иметь

дело, зачастую $K \ll n$. Кроме того, имеют место и следующие преимущества итерационных методов:

– методы итераций могут оказаться предпочтительней с точки зрения устойчивости вычислений, в смысле влияния вычислительных погрешностей на результаты расчетов. Это происходит за счет того, что погрешности окончательных результатов при использовании итерационных методов не накапливаются, а точность вычислений в каждой итерации определяется результатами предыдущей итерации и практически не зависти от ранее выполненных вычислений;

- требуют хранения в памяти машины не всей матрицы системы, а лишь нескольких векторов с n компонентами.

§4.2. Метод Якоби.

Запишем каждое уравнение системы $AX = b$ в виде, разрешенном относительно неизвестного с коэффициентом на главной диагонали матрицы A

$$x_m = \frac{1}{a_{mm}}\left(b_m - a_{m1}x_1 - \ldots - a_{m\,m-1}x_{m-1} - a_{m\,m+1}x_{m+1} - \ldots - a_{m\,n}x_n\right),\ \ m = 1,2,\ldots,n.$$

То есть переписали систему $AX = b$ в виде $X=PX+g$ с матрицей

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

на главной диагонали которой находятся нули. Если ввести в рассмотрение диагональную матрицу

$$D = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix},$$

то $P = -D^{-1}\left(A - D\right),\ \ b = D^{-1}b$. Итерационный процесс $X^{(k)} = PX^{(k-1)} + g$ с определенной таким образом матрицей P называется ***методом Якоби***. Фактически вычисления проводятся по формулам

$$x_m^{(k)} = \frac{1}{a_{mm}}\left(b_m - a_{m1}x_1^{(k-1)} - \ldots - a_{m\,m-1}x_{m-1}^{(k-1)} - a_{m\,m+1}x_{m+1}^{(k+1)} - \ldots - a_{m\,n}x_n^{(k-1)}\right),$$

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

Для сходимости метода Якоби достаточно, чтобы для исходной матрицы A имело место диагональное преобладание, то есть чтобы коэффициенты исходных уравнений удовлетворяли условиям

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

Так как условие первой теоремы выполнено для нормы

$$\|P\|_C = \max_i \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} = \max_i \frac{\sum_{j \neq i} |a_{ij}|}{|a_{ii}|} < 1.$$

Пример. Методом Якоби с точностью ε=0,01 решить систему линейных алгебраических уравнений

$$\begin{cases} 2x_1 + 2x_2 + 10x_3 = 14, \\ 10x_1 + x_2 + x_3 = 12, \\ 2x_1 + 10x_2 + x_3 = 13 \end{cases}$$

Решение. **1.** Переставим уравнения так, чтобы выполнялось условие преобладания диагональных элементов

$$\begin{cases} 10x_1 + x_2 + x_3 = 12, \\ 2x_1 + 10x_2 + x_3 = 13, \\ 2x_1 + 2x_2 + 10x_3 = 14. \end{cases}$$

Выразим из первого уравнения x_1, из второго x_2, из третьего x_3

$$\begin{cases} x_1 = -0,1x_2 - 0,1x_3 + 1,2 \\ x_2 = -0,2x_1 - 0,1x_3 + 1,3 \\ x_3 = -0,2x_1 - 0,2x_2 + 1,4 \end{cases}, \; P = \begin{pmatrix} 0 & -0,1 & -0,1 \\ -0,2 & 0 & -0,1 \\ -0,2 & -0,2 & 0 \end{pmatrix}, \; g = \begin{pmatrix} 1,2 \\ 1,3 \\ 1,4 \end{pmatrix}.$$

Заметим, что $\|P\|_1 = \max\{0,2; 0,3; 0,4\} = 0,4 < 1$, следовательно, условие сходимости выполнено.

2. Зададим $X^{(0)} = g = \begin{pmatrix} 1,2 \\ 1,3 \\ 1,4 \end{pmatrix}$.

3. Выполним расчеты по формуле

$$X^{(k+1)} = \begin{pmatrix} 0 & -0,1 & -0,1 \\ -0,2 & 0 & -0,1 \\ -0,2 & -0,2 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{pmatrix} + \begin{pmatrix} 1,2 \\ 1,3 \\ 1,4 \end{pmatrix} \text{ или } \begin{cases} x_1^{(k+1)} = -0,1x_2^{(k)} - 0,1x_3^{(k)} + 1,2 \\ x_2^{(k+1)} = -0,2x_1^{(k)} - 0,1x_3^{(k)} + 1,3 \\ x_3^{(k+1)} = -0,2x_1^{(k)} - 0,2x_2^{(k)} + 1,4 \end{cases}$$

до выполнения условия окончания $\left\| X^{(k+1)} - X^{(k)} \right\| < \varepsilon = 0,01$ и результаты занесем в таблицу

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\left\|x^{(k)} - x^{(k-1)}\right\|_1$
0	1,2000	1,3000	1,4000	–
1	0,9300	0,9200	0,900	0,5
2	1,0180	1,0240	1,0300	0,13
3	0,9946	0,9934	0,9916	0,0384
4	1,0015	1,0020	1,0024	0,0108
5	0,9996	0,9995	0,9993	$0,0027 < \varepsilon$

4. Расчет закончен, поскольку выполнено условие окончания

$$\left\|X^{(k+1)} - X^{(k)}\right\| = 0,0027 < \varepsilon = 0,01.$$

Приближенное решение задачи: $X^* \approx (0,99960;99950;9993)^T$. Очевидно, точное решение: $X^* = (1;1;1)^T$.

Приведем результаты расчетов для другого начального приближения $X^{(0)} = (1,2 \quad 0 \quad 0)^T$ и ε=0,001:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\left\|x^{(k)} - x^{(k-1)}\right\|_1$
0	1,2000	0	0	–
1	1,2000	1,0600	1,1600	1,1600
2	0,9780	0,9440	0,9480	0,2220
3	1,0108	1,0096	1,0156	0,0676
4	0,9975	0,9963	0,9959	0,0133
5	1,0008	0,0009	1,0012	0,0053
6	0,9998	0,9997	0,9997	0,0015
7	1,0001	1,0001	1,0001	$0,0004 < \varepsilon$

Приближенное решение задачи $X^* = (1,0001 \quad 1,0001 \quad 1,0001)^T$

§4.2. Метод Зейделя

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

$$x_1^{(k)} = \frac{1}{a_{11}}\left(b_1 - a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)} - \ldots - a_{1n}x_n^{(k-1)}\right),$$

$$x_2^{(k)} = \frac{1}{a_{22}}\left(b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k-1)} - \ldots - a_{2n}x_n^{(k-1)}\right),$$

--- ---

$$x_m^{(k)} = \frac{1}{a_{mm}}\left(b_m - a_{m1}x_1^{(k)} - a_{m2}x_2^{(k)} - \cdots - a_{m\,m-1}x_{m-1}^{(k)} - a_{m\,m+1}x_{m+1}^{(k)} \ldots - a_{mn}x_n^{(k-1)}\right),$$

--- ---

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

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

Можно доказать, что метод Зейделя гарантировано сходится, если выполнено условие диагонального преобладания матрицы A или матрица A является симметричной и положительно определенной.

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

Пример. Методом Зейделя с точностью ε=0,001 решить систему линейных алгебраических уравнений:

$$\begin{cases} 2x_1 + 2x_2 + 10x_3 = 14, \\ 10x_1 + x_2 + x_3 = 12, \\ 2x_1 + 10x_2 + x_3 = 13 \end{cases}$$

Решение. Приведем исходную систему $AX = b$ к виду $X=PX+g$ (смотри предыдущий пример)

$$\begin{cases} x_1 = -0,1x_2 - 0,1x_3 + 1,2 \\ x_2 = -0,2x_1 - 0,1x_3 + 1,3 \\ x_3 = -0,2x_1 - 0,2x_2 + 1,4 \end{cases}, \ P = \begin{pmatrix} 0 & -0,1 & -0,1 \\ -0,2 & 0 & -0,1 \\ -0,2 & -0,2 & 0 \end{pmatrix}, \ g = \begin{pmatrix} 1,2 \\ 1,3 \\ 1,4 \end{pmatrix}.$$

Так как $\|P\|_1 = \max\{0,2;0,3;0,4\} = 0,4 < 1$, условие сходимости выполнено.

Зададим $X^{(0)} = \begin{pmatrix} 1,2 & 0 & 0 \end{pmatrix}^T$. В поставленной задаче ε=0,001.

Выполним расчеты по формуле

$$\begin{cases} x_1^{(k+1)} = -0,1x_2^{(k)} - 0,1x_3^{(k)} + 1,2 \\ x_2^{(k+1)} = -0,2x_1^{(k+1)} - 0,1x_3^{(k)} + 1,3 \\ x_3^{(k+1)} = -0,2x_1^{(k+1)} - 0,2x_2^{(k+1)} + 1,4 \end{cases}$$

k=0,1,…) и результаты занесем в таблицу

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\left\|x^{(k)} - x^{(k-1)}\right\|_1$
0	1,2000	0	0	—
1	1,2000	1,0600	0,9480	1,0600
2	0,9992	1,0054	0,9991	0,1008
3	0,9996	1,0002	1,0000	0,0052
4	1,0000	1,0000	1,0000	$0{,}0004 < \varepsilon$

Очевидно, найденное решение $X^* = (1{,}0001 \quad 1{,}0001 \quad 1{,}0001)^T$ является точным. Расчет завершен, поскольку выполнено условие окончания

$$\left\|X^{(k+1)} - X^{(k)}\right\| = 0{,}0004 < \varepsilon = 0{,}001.$$

Замечания:

1. Для обеспечения сходимости метода Зейделя требуется преобразовать систему $AX = b$ к виду $X=PX+g$ с преобладанием диагональных элементов в матрице P (см. метод простых итераций).
2. Процесс, реализованный в методе Зейделя, называется ***последовательным итерированием***, так как на каждой итерации полученные из предыдущих уравнений значения подставляются в последующие. Как правило, метод Зейделя обеспечивает лучшую сходимость, чем метод простых итераций (за счет накопления информации, полученной при решении предыдущих уравнений). Метод Зейделя может сходиться, если расходится метод простых итераций.
3. Преимуществом метода Зейделя, как и метода простых итераций, является его "самоисправляемость".
4. Метод Зейделя имеет преимущества перед методом простых итераций, так как он всегда сходится для нормальных систем линейных алгебраических уравнений, то есть таких систем, в которых матрица A является симметрической и положительно определенной. Систему линейных алгебраических уравнений с невырожденной матрицей A всегда можно преобразовать к нормальной, если ее умножить слева на матрицу A^T (матрица $A^T A$ – симметрическая). Система $A^T AX = A^T b$ является нормальной.

§4.2. Однопараметрический метод итераций.

Перепишем $AX = b$ в виде $X = X - \tau(AX - b) = (E - \tau A)X + \tau b$, где τ – пока неопределенный параметр. Мы фактически привели $AX = b$ к форме $X=PX+g$ с матрицей $P = (E - \tau A)$ и $g = \tau b$. Чтобы обеспечить сходимость итераций

$$X^{(k)} = PX^{(k-1)} + g,$$

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

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

Итак, замечая, что

$$X^* = PX^* + g,$$

X^* – точное решение $AX = b$, и вводя в рассмотрение вектор ошибки k-го приближения $r^{(k)} = X^{(k)} - X^*$, получим, вычитая из равенства $X^{(k)} = PX^{(k-1)} + g$ равенство $X^* = PX^* + g$:

$$r^{(k)} = Pr^{(k-1)}.$$

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

Пусть $r^{(k-1)} = \sum\limits_{i=1}^{n} c_i e_i$ – разложение по элементам базиса вектора ошибки (k–1)-го приближения. Пользуясь евклидовой метрикой, имеем $\left\|r^{(k-1)}\right\|_2^2 = \left(r^{(k-1)}, r^{(k-1)}\right) = \sum c_i^2$. Подставляя это разложение в $r^{(k)} = Pr^{(k-1)}$, получим

$$r^{(k-1)} = P\left(\sum_{i=1}^{n} c_i e_i\right) = \sum_{i=1}^{n} c_i (Pe_i) = \sum_{i=1}^{n} c_i (E - \tau A) e_i = \sum_{i=1}^{n} c_i (1 - \tau\lambda_i) e_i = \sum_{i=1}^{n} c_i \mu_i e_i.$$

Как следует из цепочки соотношений, μ_i – i-е собственное значение матрицы P, которое связано с i-м собственным значением матрицы A равенством

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

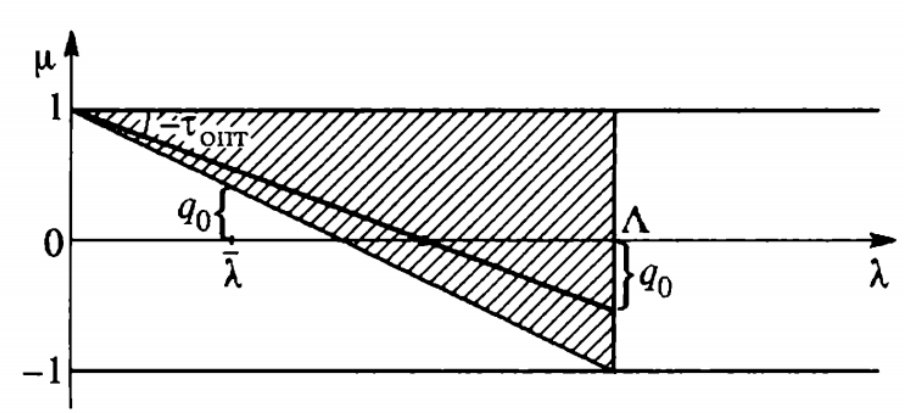
Далее отсюда следует, что

$$\left\|r^{(k)}\right\|_2^2 = \sum_i c_i^2 \mu_i^2 \le \max_i \mu_i^2 \left\|r^{(k-1)}\right\|_2^2.$$

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

Мы могли бы и сразу, без выкладок, связанных с привлечением разложения вектора ошибок по базису, сформулировать этот вывод, заметив, что P – симметричная матрица (следовательно для нее $P^T = P$), и вспомнив, что норма матрицы P, подчиненная евклидовой норме вектора, определяется в этом случае как $\|P\|_2 = \max_i |\mu_i|$. Проделанные выкладки представляют собой доказательство теоремы о сходимости итерационного процесса для частного случая $A^T = A > 0$.

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



Очевидно, собственные числа $\mu_i = 1 - \tau\lambda_i$ лежат на прямых $\mu = 1 - \tau\lambda$ с тангенсом угла наклона $-\tau$. Из рисунка видно, что условие $|\mu_i| < 1$ выполняется при всех i (то есть при всех $\lambda \le \lambda_i \le \Lambda$), когда соответствующие прямые лежат внутри заштрихованного сектора, то есть когда

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

Из разложения $r^{(k-1)} = \sum_{i=1}^{n} c_i \mu_i e_i$. следует, что величину $|\mu_i|$ можно. трактовать как коэффициент «подавления» составляющей вектора ошибки вдоль орта e_i при переходе от одного приближения к следующему. Из рисунка видно, что $\max_i |\mu_i|$ достигается либо при $\lambda_i = \lambda$, либо при $\lambda_i = \Lambda$; соответствующая компонента ошибки при увеличении числа итераций убывает наиболее медленно. Оптимальным среди всех возможных значений является $\tau = \tau_{опт}$ при котором достигается

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

Это вариант так называемой задачи о минимаксе, одной из основных задач теории оптимального управления. Из геометрических соображений (см. рис.)

ясно, что решение этой задачи доставляет прямая, которая наименее уклоняется от нуля на отрезке $[\lambda, \Lambda]$, то есть прямая, проходящая через середину этого отрезка. Соответствующее значение $\tau_{опт} = \dfrac{2}{\lambda + \Lambda}$, а минимальный по модулю коэффициент подавления погрешности достигается одновременно в конечных точках отрезка $[\lambda, \Lambda]$ и равен

$$q_0 = \min_{0<\tau<\frac{2}{\Lambda}} \left\{ \max_{\lambda<\lambda_i<\Lambda} \left|1 - \tau\lambda_i\right| \right\} = \left|1 - \tau_{опт}\lambda\right| = \left|1 - \tau_{опт}\Lambda\right| = 1 - \frac{2\lambda}{\lambda + \Lambda} = \frac{\Lambda - \lambda}{\Lambda + \lambda}.$$

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

$$\left\| r^{(k)} \right\| \le q_0 \left\| r^{(0)} \right\| .$$

Если обозначить $\eta = \dfrac{\lambda}{\Lambda}$, то $q_0 = \dfrac{1-\eta}{1+\eta}$. В данном случае величина η обратно пропорциональна числу обусловленности матрицы *A*. В самом деле, $\mu_A = \left\| A^{-1} \right\| \left\| A \right\|$. l. При сделанных предположениях относительно матрицы *A* $(A^T = A > 0)$ подчиненная евклидовой метрике векторного пространства спектральная норма для матриц приводит к $\left\| A \right\|_2 = \Lambda$ и $\left\| A^{-1} \right\|_2 = \lambda$, то есть $\mu_A = \dfrac{\Lambda}{\lambda}$.

Используя оценку невязки, получаем априорную оценку числа приближений, гарантирующих достижение заданной точности ε

$$k \ge k_0 = \frac{\ln \dfrac{\varepsilon}{\left\| r^{(0)} \right\|}}{\ln q_0}.$$

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

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

и

$$k_0 \approx \frac{1}{2\eta} \ln \frac{\left\| r^{(0)} \right\|}{\varepsilon} .$$

Здесь приводятся некоторые оценки для систем с большим числом обусловленности. Это не случайно. Дело в том, что с такими системами приходится иметь дело довольно часто при численном решении уравнений с частными производными.

Например, пусть надо решить систему из $n = 10^4$ линейных уравнений с симметричной положительно определенной матрицей. Метод Гаусса требует в этом случае выполнения порядка $n^3 \sim 10^{12}$ элементарных операций. В итерационных методах для перехода от одного приближения к следующему

необходимо выполнить порядка n^2 операций. Таким образом, объем вычислений, который сопряжен с методом итераций с одним оптимальным параметром $(k_0 n^2)$, согласно вышеприведенным рассуждениям, порядка $\frac{1}{\eta}\ln\frac{1}{\varepsilon}\cdot 10^8$, и при не слишком больших числах обусловленности этот метод эффективнее метода Гаусса.

Подробный анализ однопараметрического метода итераций открывает (в методическом плане) пути построения более эффективных итерационных процессов. Например, если при переходе от одного приближения к другому использовать различные итерационные параметры, то, оказывается, можно найти такую их последовательность, что скорость сходимости существенно возрастает сравнительно с однопараметрическим методом. Например, в *m*-параметрическом (или, как иногда говорят, *m*-шаговом) итерационном процессе последовательные приближения вычисляются по формулам

$$X^{(k)} = \left(E - \tau_k A\right) X^{(k-1)} + \tau_k b,\ k = 1, 2, \ldots, m$$

Затем *m*-е приближение принимается за нулевое, цикл повторяется и так далее, до сходимости с требуемой точностью. Последовательные ошибки (в пределах одного цикла) $r^{(k)} = X^{(k)} - X^*$ (X^* – точное решение исходной системы) в этом случае связаны между собой соотношением

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

В предположении, что $A^T = A > 0$, имеет место равенство

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

где λ_i – собственные значения матрицы *A*.