Большинство вычислительных задач современной физики, как правило, на том или ином этапе приводит к необходимости решения систем алгебраических уравнений

$$\begin{aligned} a_{11}x_1 + ... + a_{1n}x_n &= b_1 \\ &..... \\ a_{n1}x_1 + ... + a_{nn}x_n &= b_n, \end{aligned} \tag{1}$$

или в матричной форме

$$Ax = b,$$

где $A = \{a_{ij}\}$ – невырожденная $(n \times n)$-матрица, $x = \{x_i\}$ – n-вектор (столбец) неизвестных, $b = \{b_i\}$ – заданный n-вектор.

В множестве разнообразных численных методов решения системы (1) можно выделить две большие (и наиболее широко используемые) группы методов: прямые и итерационные. Им и посвящена данная лабораторная работа. Прежде, чем приступать к описанию численных методов, напомним необходимые сведения из линейной алгебры.

§1. Нормы векторов и матриц ([1], стр. 120-123)

В численном анализе наиболее часто используются следующие три нормы n-мерного вектора $x = \{x_1, ..., x_n\}$:

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

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2},$$

$$\|x\|_\infty = \max_{1 \le i \le n} |x_i|.$$

Нормой квадратной $(n \times n)$-матрицы A называется функционал, обозначаемый $\|A\|$, удовлетворяющий условиям:

1) $\|A\| \ge 0; \quad \|A\| = 0 \leftrightarrow A$ – нулевая матрица;

2)$\|\alpha A\| = |\alpha| \|A\|, \quad \alpha$ – произвольное число;

3)$||A + B|| \le ||A|| + ||B||$;

4)$||AB|| \le ||A||\ ||B||$.

Норма $(n \times n)$-матрицы $||A||_\alpha$ называется согласованной с нормой n-мерного вектора $||x||_\beta$, если для любых x и A выполняется неравенство

$$||Ax||_\beta \leq ||A||_\alpha \, ||x||_\beta.$$

Нормы матриц

$$||A||_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|,$$

$$||A||_* = (\sum_{i,j=1}^{n} |a_{ij}|^2)^{1/2},$$

$$||A||_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$$

согласованы соответственно с нормами вектора $||x||_1$, $||x||_2$, $||x||_\infty$.

Норма $(n \times n)$-матрицы $||A||$ называется подчиненной норме n-вектора $||x||_\beta$, если для любой матрицы A

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

Норма матрицы $||A||$, подчиненная норме вектора $||x||_\beta$, является наименьшей среди всех норм матрицы, согласованных с этой нормой вектора. Нормы $||A||_1$ и $||A||_\infty$ подчинены нормам $||x||_1$ и $||x||_\infty$ соответственно. Норме $||x||_2$ подчинена норма

$$||A||_2 = (\lambda_{max})^{1/2},$$

где λ_{max} – максимальное собственное число матрицы $A^T A$, A^T – транспонированная матрица к матрице A. Норму матрицы $||A||$, подчиненную норме вектора $||x||_\beta$, будем обозначать $||A||_\beta$.

§ 2. Прямые методы

Прямыми методами решения алгебраических систем называют методы, использующие конечное, известное заранее число арифметических операций. При этом точность решения определяется лишь точностью арифметических вычислений. Основной идеей прямых методов, в конечном итоге, является идея исключения неизвестных.

1. Метод Гаусса([1], § 16)

Одним из прямых методов для решения линейных систем общего вида (1) является хорошо известный метод последовательного исключения, связываемый с именем Гаусса. Метод заключается в приведении системы (1) к системе уравнений с треугольной матрицей (прямой ход метода)

$$\begin{aligned} c_{11}x_1 + c_{12}x_2 + ... + c_{1n}x_n &= y_1 \\ c_{22}x_2 + ... + c_{2n}x_n &= y_2 \\ ..... & \\ c_{nn}x_n &= y_n \end{aligned} \tag{2}$$

с последующим решением системы(2), начиная с последнего уравнения (обратный ход метода). Различные модификации метода отличаются друг от друга путями приведения системы (1) к виду (2).

Опишем одну из модификаций, которая называется компактной схемой Гаусса. Эта модификация использует представление матрицы A как произведения нижней треугольной матрицы $B = \{b_{ij}\}$ и верхней треугольной матрицы $C = \{c_{ij}\}$ вида:

$$A = BC = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ b_{11} & 1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ b_{n1} & b_{n2} & b_{n3} & ... & 1 \end{bmatrix} \times \begin{bmatrix} c_{11} & c_{12} & c_{13} & ... & c_{1n} \\ 0 & c_{22} & c_{23} & ... & c_{2n} \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & c_{nn} \end{bmatrix} \tag{3}$$

(представление (3) возможно, если, например, все главные миноры матрицы A отличны от нуля. Решение исходной системы (1) сводится к последовательному решению систем $By = b$ и $Cx = y$ с треугольными матрицами. Нахождение матриц B, C и вектора y (из системы $By = b$) является прямым ходом метода, а нахождение вектора x из системы $Cx = y$ обратным ходом. Формулы для

элементов матриц B и C имеют вид:

$$
\begin{aligned}
&c_{11} = a_{11}; \qquad c_{1j} = a_{1j}, \qquad b_{j1} = a_{j1}/c_{11}, \qquad j = 2, 3, ..., n; \\
&c_{ii} = a_{ii} - \sum_{k=1}^{i-1} b_{ik}c_{ki}, \qquad i = 2, 3, ..., n; \\
&c_{ij} = a_{ij} - \sum_{k=1}^{i-1} b_{ik}c_{kj}, \\
&b_{ji} = (a_{ji} - \sum_{k=1}^{j-1} b_{jk}c_{ki})/c_{ii}, \\
&j = i+1, i+2, ..., n, \qquad i = 2, 3, ..., n.
\end{aligned}
$$

2. Метод прогонки ([2], §5)

Важным классом систем (1) являются системы вида:

$$
\begin{aligned}
&d_1x_1 + c_1x_2 = b_1\,, \\
&a_ix_{i-1} + d_ix_i + c_ix_{i+1} = b_i, \qquad i = 2, 3, ..., n-1, \\
&a_nx_{n-1} + d_nx_n = b_n
\end{aligned}
\tag{4}
$$

с трехдиагональной матрицей

$$
A = \begin{bmatrix}
d_1 & c_1 & 0 & 0 & ... & ... & 0 & 0 \\
a_2 & d_2 & c_2 & 0 & ... & ... & 0 & 0 \\
... & ... & ... & ... & ... & ... & ... & ... \\
0 & 0 & ... & ... & 0 & a_{n-1} & d_{n-1} & c_{n-1} \\
0 & 0 & ... & ... & 0 & 0 & a_n & d_n
\end{bmatrix}.
$$

Такие системы возникают при решении краевых задач для дифференциальных уравнений разностными методами.

Специальный вид матрицы A позволяет применить идею исключения неизвестных в системе (4) следующим простым способом, который носит название метода прогонки.

Первое уравнение системы (4) дает соотношение между x_1 и x_2, в силу которого второе уравнение дает соотношение между x_2 и x_3. Следовательно, третье уравнение дает соотношение между x_3 и x_4 и т.д.. Запишем связь между неизвестными x_{i-1} и x_i в виде:

$$
x_{i-1} = L_ix_i + M_i, \qquad i = 2, 3, ..., n. \tag{5}
$$

Из первого уравнения системы (4) следует, что

$$
L_2 = -c_1/d_1, \qquad M_2 = b_1/d_1. \tag{6}
$$

Подставляя (5) в i-ое уравнение системы (4), получим

$$x_i = -\frac{c_i}{a_i L_i + d_i} x_{i+1} + \frac{b_i - M_i a_i}{a_i L_i + d_i}. \qquad (7)$$

Сравнивая (5) и (7), находим рекуррентные соотношения

$$L_{i+1} = -\frac{c_i}{a_i L_i + d_i}, \qquad M_{i+1} = \frac{b_i - M_i a_i}{a_i L_i + d_i}, \qquad i = 2, ..., n-1,$$

которые вместе с формулами (6) позволяют последовательно найти все прогоночные коэффициенты $L_i, M_i, i = 2, ..., n$. Процесс нахождения этих коэффициентов называется прямым ходом метода прогонки. Из последнего уравнения системы (4) и соотношения (5) для $i = n$ находим

$$x_n = \frac{b_n - M_n a_n}{L_n a_n + d_n},$$

что позволяет по формулам (5) последовательно найти все остальные неизвестные $x_{n-1}, x_{n-2}, ..., x_1$ (обратный ход метода прогонки).

Мы рассмотрели два характерных для прямых методов алгоритма. С другими алгоритмами можно ознакомится, например, по [1], [2].

§ 3. Итерационные методы

Итерационными называются приближенные методы, в которых решение системы (1) получается как предел последовательности векторов $\{x^k\}_{k=1}^{\infty}$, каждый последующий элемент которой вычисляется по некоторому единому правилу. Начальный элемент x^1 выбирается произвольно. Последовательность $\{x^k\}_{k=1}^{\infty}$ называется итерационной, а ее элементы последовательными итерациями (приближениями).

Важной характеристикой итерационного процесса является скорость сходимости итерационной последовательности. Говорят, что итерация x^k является с точностью ϵ (в смысле некоторой нормы $||x||$) приближенным решением системы (1), если

$$||x^k - x^0|| \leq \epsilon, \tag{8}$$

где x^0 -точное решение системы (1).

Как правило, для итерационного метода решения системы (1) существует такая последовательность невырожденных матриц H_k, $k = 1, 2, ...$, что правило построения элементов итерационной последовательности записывается в виде (см. [1], стр. 204-207)

$$x^{k+1} = x^k - H_k(Ax^k - b). \tag{9}$$

Запишем (9) в виде

$$x^{k+1} = T_k x^k + H_k b, \tag{10}$$

где $T_k = E - H_k E$, E – единичная $(n \times n)$-матрица. Вектор $\varphi^k = x^k - x^0$ называется вектором ошибки, а вектор $r^k = Ax^k - b$ – вектором невязки.

Итерационный метод называется стационарным, если матрица H_k не зависит от номера шага k. В противном случае метод называется нестационарным. Для того, чтобы стационарный итерационный процесс

$$x^{k+1} = Tx^k + Hb, \tag{11}$$

сходился, достаточно, чтобы для какой-либо одной нормы матрицы T выполнялось неравенство

$$||T|| < 1. \tag{12}$$

Критерием окончания итерационного процесса при заданной точности ϵ (см.(8)) может служить неравенство

$$||x^k - x^{k-1}|| \le \epsilon \frac{1 - ||T||}{||T||}. \tag{13}$$

Рассмотрим некоторые итерационные методы.

1. Метод последовательных приближений ([1], §30)

Для этого метода

$$H^k = E, \quad T_k = E - A,$$

т.е. (10) имеет вид

$$x^{k+1} = (E - A)x^k + b,$$

или в покоординатной форме

$$\begin{aligned}
x_1^{k+1} &= -(a_{11} - 1)x_1^k - a_{12}x_2^k - ... - a_{1n}x_n^k + b_1 \\
x_2^{k+1} &= -a_{21}x_1^k - (a_{22} - 1)x_2^k - ... - a_{2n}x_n^k + b_2 \\
&..... \\
x_n^{k+1} &= -a_{n1}x_1^k - a_{n2}x_2^k - ... - (a_{nn} - 1)x_n^k + b_n.
\end{aligned}$$

2. Одношаговый циклический метод Зейделя ([1] , §32)

Здесь

$$H_k = (E - M)^{-1}, \quad T_k = (E - M)^{-1}N,$$

где $M + N = E - A$, M–нижняя треугольная матрица с нулевыми диагональными элементами , N–верхняя треугольная матрица

$$M = \begin{bmatrix}
0 & 0 & 0 & 0 & ... & 0 & 0 \\
-a_{21} & 0 & 0 & 0 & ... & 0 & 0 \\
-a_{31} & -a_{32} & 0 & 0 & ... & 0 & 0 \\
... & ... & ... & ... & ... & ... & \\
-a_{n1} & -a_{n2} & ... & ... & ... & -a_{nn-1} & 0
\end{bmatrix}, \quad N = \begin{bmatrix}
1 - a_{11} & -a_{12} & ... & -a_{1n} \\
0 & 1 - a_{22} & ... & -a_{2n} \\
... & ... & ... & ... \\
... & ... & ... & ... \\
0 & 0 & ... & 1 - a_{nn}
\end{bmatrix},$$

т.е. (9) имеет вид

$$x^{k+1} = (E - M)^{-1}Nx^k + (E - M)^{-1}b,$$

или в покоординатной форме

$$x_1^{k+1} = -(a_{11} - 1)x_1^k - a_{12}x_2^k - ... - a_{1n}x_n^k + b_1$$
$$x_2^{k+1} = -a_{21}x_1^{k+1} - (a_{22} - 1)x_2^k - ... - a_{2n}x_n^k + b_2$$
$$.....$$
$$x_n^{k+1} = -a_{n1}x_1^{k+1} - a_{n2}x_2^{k+1} - ... - (a_{nn} - 1)x_n^k + b_n.$$

3. Итерационный метод Гаусса

Для этого метода

$$H_k = D^{-1}, \qquad T_k = -D^{-1}(L + R),$$

где $L + D + R = A$, L–нижняя треугольная матрица с нулевыми диагональными элементами, D–диагональная матрица, R–верхняя треугольная матрица с нулевыми диагональными элементами.

Итерационный процесс (10) имеет вид

$$x^{k+1} = -D^{-1}(L + R)x^k + D^{-1}b,$$

или в покоординатной форме

$$x_1^{k+1} = (b_1 - a_{12}x_2^k - ... - a_{1n}x_n^k)/a_{11}$$
$$x_2^{k+1} = (b_2 - a_{21}x_1^k - ... - a_{2n}x_n^k)/a_{22}$$
$$.....$$
$$x_n^{k+1} = (b_n - a_{n1}x_1^k - ... - a_{nn-1}x_{n-1}^k)/a_{nn}$$

4. Метод Гаусса-Зейделя ([1], §33)

Здесь

$$H_k = (D + L)^{-1}, \qquad T_k = -(D + L)^{-1}R,$$

L, D и R – те же, что и в итерационном методе Гаусса. Таким образом, процесс (10) имеет вид

$$x^{k+1} = -(D+L)^{-1}Rx^k + (D+L)^{-1}b,$$

или в покоординатной форме

$$x_1^{k+1} = (b_1 - a_{12}x_2^k - ... - a_{1n}x_n^k)/a_{11}$$
$$x_2^{k+1} = (b_2 - a_{21}x_1^{k+1} - ... - a_{2n}x_n^k)/a_{22}$$
$$.....$$
$$x_n^{k+1} = (b_n - a_{n1}x_1^{k+1} - ... - a_{nn-1}x_{n-1}^{k+1})/a_{nn} \quad .$$

Данный метод отличается от предыдущего, как и метод 2 от метода 1, тем, что для вычисления каждой координаты очередного вектора x^{k+1} используется (уже вычисленные) все предыдущие его координаты.

Как правило, перечисленные методы не применяются без ускорения процесса сходимости. Ускорение достигается, например, введением так называемых ускоряющих множителей. В простейшем случае введение ускоряющего множителя τ заключается в незначительной линейной экстраполяции с коэффициентом τ, применяемой к каждой неизвестной x_i на каждой итерации по формуле

$$\bar{x}_i^{k+1} = x_i^k + \tau(x_i^{k+1} - x_i^k), \tag{14}$$

где x_i^{k+1}- последнее значение x_i полученное без ускорения, а $\bar{x}_i^{k+1}$ – ускоренное значение x_i. Оптимальное значение параметра τ зависит от вида матрицы A и обычно для пололожительных симметрических матриц заключено в пределах от 0 до 2. Приведем пример итерационного процесса с ускоряющим множителем.

5. Метод последовательной релаксации([1], §35)

Здесь

$$H_k = \tau(D+\tau L)^{-1}, T_k = (D+\tau L)^{-1}((1-\tau)D - \tau R),$$

где D, L, R те же, что и в пункте 3, $0 < \tau < 2$. Итерационный процесс (10) имеет вид

$$x^{k+1} = (D + \tau L)^{-1}((1 - \tau)D - \tau R)x^k + \tau(D + \tau L)^{-1}b,$$

или в покоординатной форме

$$x_1^{k+1} = x_1^k + (\tau/a_{11})(b_1 - a_{11}x_1^k - a_{12}x_2^k - ... - a_{1n}x_n^k)$$

$$x_2^{k+1} = x_2^k + (\tau/a_{22})(b_2 - a_{21}x_1^{k+1} - a_{22}x_2^k - ... - a_{2n}x_n^k)$$

.....

$$x_n^{k+1} = x_n^k + (\tau/a_{nn})(b_n - a_{n1}x_1^{k+1} - a_{n2}x_2^{k+1} - ... - a_{nn}x_n^k).$$

При $\tau > 1$ метод называется методом последовательности верхней релаксации, при $\tau < 1$ методом последовательной нижней релаксации, при $\tau = 1$ совпадает с методом Гаусса-Зейделя.

Оценить теоретически оптимальное значение параметра τ представляется затруднительным. На практике для определения наилучшего значения τ обычно пользуются методом подбора. Строят зависимость от τ числа итерации N, необходимых для получения решения с некоторой наперед заданной невысокой точностью. В качестве ускоряющего множителя выбирают то значение τ, при котором функция $N(\tau)$ достигает минимума.

Рассмотренные итерационные методы являются стационарными. Рассмотрим теперь два нестационарных метода.

6. Методы наискорейшего спуска и минимальных невязок ([1], §70)

Здесь

$$H_k = \tau_k E, \qquad T_k = E - \tau_k A.$$

Параметр τ_k в методе наискорейшего спуска выбирается из условия ортогональности невязок на соседних шагах метода (невязка на $(k+1)$-ом шаге равна $r^k = Ax^k - f$) и имеет вид

$$\tau_k = \frac{(r^k, r^k)}{(Ar^k, r^k)}. \tag{15}$$

В методе минимальных невязок итерационный параметр τ_k выбирается из условия минимума нормы невязки на $(k+1)$-ом шаге и имеет вид

$$\tau_k = \frac{(Ar^k, r^k)}{(Ar^k, Ar^k)}. \tag{16}$$

Процесс (10) имеет вид

$$x^{k+1} = (E - \tau_k A)x^k + \tau_k Eb, \tag{17}$$

или в покоординатной форме

$$\begin{aligned}
x_1^{k+1} &= -(\tau_k a_{11} - 1)x_1^k - \tau_k a_{12}x_2^k - \ \ldots\ - \tau_k a_{1n}x_n^k + \tau_k b_1 \\
x_2^{k+1} &= -\tau_k a_{21}x_1^k - (\tau_k a_{22} - 1)x_2^k - \ \ldots\ - \tau_k a_{2n}x_n^k + \tau_k b_2 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
x_n^{k+1} &= -\tau_k a_{n1}x_1^k - \tau_k a_{n2}x_2^k - \ \ldots\ - (\tau_k a_{nn} - 1)x_n^k + \tau_k b_n.
\end{aligned}$$

§ 4. Обусловленность линейной системы

Обычно элементы матрицы A и вектора b системы (1) заданы лишь с некоторой степенью точности. Посмотрим, как такая неопределенность сказывается на решении системы (1). Пусть, например, матрица A известна точно, а вместо значения вектора b задано его приближение (b+Δb). Пусть x^0 точное решение системы (1) т.е.

$$Ax^0 = b, \tag{18}$$

а вектор Δx удовлетворяет системе

$$A(x^0 + \Delta x) = b + \Delta b. \tag{19}$$

Выбрав некоторую норму вектора $||x||_\beta$ оценим относительную погрешность

$$\frac{||\Delta x||_\beta}{||x^0||_\beta}.$$

Из (18) и (19) находим

$$\Delta x = A^{-1}\Delta b$$

и, следовательно,

$$||x||_\beta \le ||A^{-1}||_\beta \, ||\Delta b||_\beta. \tag{20}$$

Кроме того, из (19) получаем

$$||b||_\beta \le ||A||_\beta \, ||x^0||_\beta. \tag{21}$$

Из неравенств (20) и (21) следует, что

$$||b||_\beta \, ||\Delta x||_\beta \le ||A||_\beta \, ||A^{-1}||_\beta \, ||\Delta b||_\beta \, ||x^0||_\beta.$$

Предположив , что $b \neq 0$, находим

$$\frac{||\Delta x||_\beta}{||x^0||_\beta} \le ||A||_\beta ||A^{-1}||_\beta \frac{||\Delta b||_\beta}{||b||_\beta}.$$

Величина

$$cond_\beta(A) = ||A||_\beta \, ||A^{-1}||_\beta$$

называется числом обусловленности матрицы A (это число зависит от используемой нормы матрицы).Таким образом, относительная погрешность решения $\frac{||\Delta x||_\beta}{||x^0||_\beta}$ не превышает относительной погрешности $\frac{||\Delta b||_\beta}{||b||_\beta}$, умноженной на число обусловленности:

$$\frac{\|\Delta x\|_\beta}{\|x^0\|_\beta} \le cond_\beta(A) \frac{\|\Delta b\|_\beta}{\|b\|_\beta}. \tag{22}$$

Неравенство (22) является точным в том смысле, что при заданной погрешности $\|\Delta b\|_\beta / \|b\|_\beta$ существует вектор $\Delta b \neq 0$ и соответствующий ему вектор Δx такие, что (22) обращается в равенство.

Аналогично, если вектор b известен точно, а вместо матрицы A задано ее приближение $A + \Delta A$, и вектор Δx удовлетворяет системе

$$(A + \Delta A)(x^0 + \Delta x) = b,$$

то

$$\frac{\|\Delta x\|_\beta}{\|x^0 + \Delta x\|_\beta} \le cond_\beta(A) \frac{\|\Delta A\|_\beta}{\|A\|_\beta}. \tag{23}$$

Неравенство (23) точное.

Если число $cond_\beta(A)$ относительно мало, то система (1) называется хорошо обусловленной. Если же число $cond_\beta(A)$ относительно велико, то система называется плохо обусловленной. Неравенства (22),(23) показывают, что к решению плохо обусловленных систем надо относится очень внимательно, т.к. даже небольшие ошибки округления могут привести к неверным результатам.

Пример ([3],стр.37): точным решением системы

$$x_1 + 0.99x_2 = 1.99$$

$$0.99x_1 + 0.98x_2 = 1.97$$

является вектор $x^0 = \{1, 1\}$. Точным решением системы

$$x_1 + 0.99x_2 = 1.989903$$

$$0.99x_1 + 0.98x_2 = 1.970106$$

является вектор $\{3.0000, -1.0203\}$. Таким образом, изменение $\Delta b = \{-0.000097, -0.000106\}$ привело к изменению решения $\Delta x = \{-2.0000, -2.0203\}$. В то время , как $\|\Delta b\|_2/\|b\|_2 \cong 0.510^{-4}$, имеем $\|\Delta x\|_2/\|x\|_2 \cong 2$. Большая относительная ошибка решения здесь обусловлена большим значением $cond_2 = 39600$.

Оценка числа обусловленности обычно довольно трудоемка. При практических вычислениях плохую обусловленность определяют часто по некоторым внешним признакам матрицы A. Например, если определитель матрицы близок к нулю, или некоторые диагональные элементы матрицы A малы по сравнению с недиагональными, то может оказаться, что система (1) плохо обусловлена.

Для важного в приложении частного случая систем вида (3) с трехдиагональной матрицей A существует легко проверяемый достаточный признак хорошей обусловленности:

$$|d_i| \geq |a_i| + |b_i|, \qquad i = 1, 2, ..., n,$$
$$|c_1/d_1| < 1, \qquad |a_n/d_n| < 1.$$

§5. Итерационные методы при неточных входных данных ([6], стр.259)

До сих пор при рассмотрении методов решения линейных систем мы исходили из того, что матрица и правые части нашей системы заданы точно. Однако, в практических задачах входные данные, т.е.

матрица и правая часть линейной системы, обычно заданы приближенно и вместо системы

$$Ax = b \tag{24}$$

мы имеем дело с системой

$$A_h x = b_h, \tag{25}$$

где A_h и b_h - известные нам приближения матрицы A и вектора b соответственно (входные данные зависят либо от случайных ошибок или статистических погрешностей, либо от погрешностей, появляющихся при вычислениях).

Будем предполагать, что погрешности входных данных нам известны в следующем виде (x – точное решение системы (24)):

$$||(A - A_h)x|| \le \varepsilon(h), \quad ||\Delta b|| = ||b - b_h|| \le \eta(h). \tag{26}$$

Следовательно, нам нужно приближенно решить систему (24)), имея в распоряжении систему (25) и априорную информацию (26). Для решения системы (25) используем метод последовательных приближений (x_h^j – j-ое приближение; сказанное ниже без труда распространяется и на другие изложенные выше методы)

$$x_h^{j+1} = (E - A_h)x_h^j + b_h \quad (x_h^0 = 0), \quad j = 0, 1, 2, ..., \tag{27}$$

считая выполненным условие

$$q = ||E - A_h|| < 1, \tag{28}$$

обеспечивающее сходимость.

Естественно предположить, что при неточно известных входных данных последовательные приближения (27) следует вычислять до тех пор, пока ошибка итерационного процесса (27) не станет сравнимой с заведомой (неустранимой) ошибкой решения системы (24), появляющейся из-за неточности входных данных. Более того, оказывается, если матрица A плохо обусловлена и, следовательно, обратная матрица A_h^{-1} может отличаться от обратной матрицы A^{-1} значительно, то продолжение итерационного процесса (27) может привести не к улучшению, а наоборот, к существенному ухудшению окончательного результата.

Для нахождения оптимального числа итераций в итерационном процессе (27) проведем следующий анализ. Пусть x и x_h – точные

решения систем (24) и (25) соответственно, т.е.

$$x = A^{-1}b, \quad x_h = A_h^{-1}b_h. \tag{29}$$

Вычитая из второго уравнения (29) первое и проводя тождественные преобразования, получим

$$x_h - x = A_h^{-1}(b_h - b + Ax - A_hA^{-1}b) = A_h^{-1}(b_h - b + (A - A_h)x). \tag{30}$$

Из последнего следует

$$||x_h - x|| \le ||A_h^{-1}||(||b_h - b|| + ||(A - A_h)x||), \tag{31}$$

или с учетом априорных оценок (26)

$$||x - x_h|| \le ||A_h^{-1}||(\varepsilon(h) + \eta(h)). \tag{32}$$

Представим уравнение (25) в виде

$$x_h = (E - A_h)x_h + b_h. \tag{33}$$

Из (33) и (27) получим

$$(x_h - x_h^j) = (E - A_h)(x_h - x_h^j), \tag{34}$$

откуда

$$x_h - x_h^j = (E - A_h)^j A_h^{-1} b_h. \tag{35}$$

Т.к.

$$||x - x_h^j|| = ||x + x_h - x_h - x_h^j|| \le ||x_h^j - x_h|| + ||x_h - x||, \tag{36}$$

то из (32) и (35) получаем

$$||x_h - x_h^j|| \le q^j ||A_h^{-1}||||b_h|| + ||A_h^{-1}||(\varepsilon(h) + \eta(h)). \tag{37}$$

Первое слагаемое в последнем неравенстве дает оценку погрешности итерационного процесса (27), а второе слагаемое оценивает "неустранимую"погрешность решения системы ((24), возникающую за счет неточностей входных данных. Требуя равенства этих величин, получим формулу для номера итерации j_0, на которой следует закончить итерационный процесс ((27):

$$j_0 = \frac{1}{\ln q} \ln \frac{(\varepsilon(h) + \eta(h))}{||b_h||}. \tag{38}$$

Следует отметить, что в формуле (38) отсутствует норма обратной матрицы, что существенно упрощает вычисление оптимального числа итераций.

Вопросы для самопроверки

1) Оценить возможное возмущение решения системы

$$\begin{aligned} x_1 - 2x_2 &= -1 \\ 2x_1 + 4.01x_2 &= 2 \end{aligned}$$

при изменении компонент правой части на 0.01. Найти решения указанной системы и системы с той же матрицей и правой частью $b + \Delta b = \{-1, 2.01\}$

2) Найти число обусловленности $cond_\infty(A)$ матрицы системы

$$\begin{aligned} 5x_1 - 3.331x_2 &= 1.69 \\ 6x_1 - 3.97x_2 &= 2.03 \end{aligned}$$

Указать изменение решения этой системы при переходе к системе с той же матрицей и правой частью $b + \Delta b = \{1.7, 2\}$

3) Пусть $||x||_\alpha$ - некоторая норма вектора, $||A||_\alpha$ - подчиненная ему норма матрицы. Показать, что при изменении правой части системы (1) на вектор с нормой равной числу $\varepsilon > 0$, решение системы (1) может измениться на вектор, имеющий норму $\varepsilon||A^{-1}||_\alpha$.

4) Доказать, что норма матрицы $||A||_1$ (соотв. $||A||_\infty$) согласована с нормой вектора $||x||_1$ (соотв. $||x||_\infty$).

5) Показать, что для системы (1) с матрицей

$$A = \begin{bmatrix} 2 & 0.3 & 0.5 \\ 0.1 & 3 & 0.4 \\ 0.1 & 0.1 & 1.8 \end{bmatrix}$$

итерационный процесс

$$x^{k+1} = (E - \tau A)x^k + \tau B$$

будет сходиться при $0 < \tau < 0.4$.

6) Написать формулы метода прогонки для системы

$$\begin{aligned} &b_1 x_1 = f_1 \\ &a_i x_{i-1} - b_i x_i + c_i x_{i+1} = f_i, \qquad i = 2, 3, ..., n-1, \\ &b_n x_n = f_n. \end{aligned} \tag{39}$$

7) Доказать, что достаточным условием сходимости процесса (11) является выполнение неравенства (12).

8) Доказать, что достаточным признаком хорошей обусловленности системы (39) является условие:

$$|b_i| \geq |a_i| + |c_i|, \qquad i = 1, 2, 3, ..., n-1.$$

9) Доказать,что итерационный процесс (11) при условии(12) сходится со скоростью геометрической прогрессии и выполняется неравенство:

$$||x^k - x^0|| \leq \frac{||T||^{k-1}}{1 - ||T||} ||x^2 - x^1||.$$

10) Доказать,что для итерационного процесса (11) при условии (12) из неравенства (13) следует неравенство (8).

11) Решению системы двух линейных уравнений с двумя неизвестными

$$a_{11}x_1 + a_{12}x_2 = b_1$$

$$a_{21}x_1 + a_{22}x_2 = b_2,$$

матрица которой действительная и невырожденная, соответствует геометрическая задача отыскания на плоскости x_1, x_2 точки пересечения двух прямых, заданных уравнениями системы.

Доказать,что угол между этими прямыми удовлетворяет неравенству

$$|ctg\alpha| \leq 0.5 cond_*(A).$$

12) Показать, что число обусловленности не меняется при умножении матрицы на нулевое число.

13) Найти градиент функции $h(x) = (Ax, x), x \in R^2$.

14) Показать, что если матрица A положительно определена, то задача решения системы (1) эквивалентна задаче отыскания минимума функции

$$H(x) = (Ax, x) - 2(b, x). \tag{40}$$

15) Пользуясь тем, что в случае положительно определенной матрицы A задача решения системы (1) эквивалентна задаче нахождения минимума функции (39); получить формулы (15),(17) метода наискорейшего спуска.

16) Доказать, что для диагональной матрицы $A = \{a_{ij}\}, a_{ij} = 0, i \neq j$, число обусловленности $cond_1(A) = max_i(a_{ii})/min_i(a_{ii})$.

17) Привести примеры матриц, показывающие, что число обусловленности может быть равным единице при сколь угодно малом определителе. (Указание: рассмотреть случай диагональной матрицы.)

Варианты заданий

В лабораторной работе требуется решить заданную линейную систему с неточно заданной правой частью методом исключения Гаусса и указанным итерационным методом. В каждом из приводимых ниже вариантов заданий необходимо:

а) Оценить число обусловленности матрицы системы.

б) Проверить выполнение достаточных условий разрешимости заданной системы указанным методом (в случае необходимости преобразовать систему).

в) Решить заданную систему на ЭВМ с помощью подпрограммы, описанной в приложении и заданным итерационным методом. Сравнить полученные результаты в графической форме.

г) Оценить погрешность найденного решения.

Приводимые ниже линейные алгебраические системы возникают при численном решении двухточечных граничных задач для обыкновенных дифференциальных уравнений (системы 1,2), интегральных уравнений Фредгольма второго рода (системы 3,4), первой краевой задачи для уравнения Пуассона (система 5).

1) $$x_{i-1} - 2x_i + x_{i+1} = 0, \qquad i = 2, 3, ..., n-1,$$

$$x_1 = 1, x_n = 5.$$

$$x_{i-1} - 4x_i + x_{i+1} = 0, \qquad i = 2, 3, ..., n-1,$$

$$2x_1 = x_2 + 1,$$

$$x_{n-1} = 3x_n + 2.$$

$$x_i - \frac{i}{2n^2}\sum_{j=1}^{n} x_j = 1, \qquad i = 1, 2, ..., n.$$

$$x_i - \frac{i}{2n^3}\sum_{j=1}^{n} jx_j = 1, \qquad i = 1, 2, ..., n.$$

$$x_m = 0,$$

$$m = 1, 2, ..., k+1, 2k, 2k+1,$$

$$3k, 3k+1,$$

$$.....,$$

$$(k-1)k, (k-1)k+1, ..., k^2.$$

$$4x_m - x_{m+1} - x_{m-1} - x_{m+k} - x_{m-k} = 1/(k-1)^2, \qquad (41)$$

$$m = k+2, k+3, ..., 2k-1$$

$$2k+2, 2k+3, ..., 3k-1,$$

$$.....,$$

$$(k-2)k+2, (k-2)k+3, ..., (k-1)k-1.$$

(число неизвестных $n = k^2$).

Таблица I

N варианта	N системы	численный метод	число неизвестных	точность ϵ
1	1	§2,п.2	10	
2	1	§2,п.2	20	
3	2	§2,п.2	10	
4	2	§2,п.2	20	
5	2	§3,п.1	10	0.01
6	3	§3,п.1	10	0.01
7	3	§3,п.3	10	0.01
8	3	§3,п.4	10	0.01
9	4	§3,п.1	10	0.01
10	4	§3,п.3	10	0.01
11	4	§3,п.4	10	0.01
12	5	§3,п.2	25	0.01
13	5	§3,п.4	25	0.01
14	5	§3,п.5	25	0.01

Приложение

В файле

i:\users\studendts\libr\gauss.for

содержится программа [4] для анализа и решения систем с невырожденными матрицами методом Гаусса.
Программа состоит из основной программы и двух подпрограмм DECOMP и SOLVE. Подпрограмма DECOMP осуществляет прямой ход метода Гаусса и одновременно проводит оценку cond исходной матрицы. Подпрограмма SOLVE использует результаты подпрограммы DECOMP для получения решения с произвольной правой частью.
Заметим, что значения правых частей устанавливаются лишь после того, как выяснилось, что исходная матрица хорошо обусловлена.
Обращение к подпрограмме DECOMP осуществляется оператором
CALL DECOMP (NDIN,N,A,COND,IPVT,WORK)
Здесь входными параметрами являются:
NDIN - заявленная строчная размерность матрицы A в основной программной единице.
N - порядок матрицы A.
A(NDIM,N) - матрица системы размера NDIM*N.
Выходными параметрами являются:
WORK(N) - рабочий массив вещественного типа.
Обращение к подпрограмме SOLVE осуществляется оператором
CALL SOLVE (NDIM,N,A,B,IPVT)
Здесь параметры NDIM,N,F,IPVT имеют тот же смысл, что и в программе DECOMP.
Массив B(N) является вектором правых частей исходной системы.
После выполнения подпрограммы в массиве B(N) содержится вектор решения x(N).
Описание входной и выходной информации приводится также в тексте программы, более подробное описание содержится в [4].

ЛИТЕРАТУРА:

1. Фаддеев Д.К.,Фаддеева В.Н. Вычислительные методы линейной алгебры. М.–Л.: Физматгиз, 1960.

2. Годунов С.К., Рябенький В.С. Разностные схемы. Введение в теорию.-М., Наука, 1977, стр.38–41,51–53.
3. Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений.-М.: Мир. 1960.
4. Форсайт Дж., Малькольм М., Молер К. Машинные методы математических вычислений.-М.: Мир. 1980.
5. Воеводин В.В. Вычислительные основы линейной алгебры. М.: Наука. 1977.
6. Марчук Г.И. Методы вычислительной математики. М.: Наука. 1989.

C Иллюстрирующая программа для подпрограм DECOMP и SOLVE

```
 REAL A(10,10),B(10),WORK(10),COND,CONDP1
 INTEGER IPVT(10), I, J, N, NDIM
 NDIM=10
 N=3
 A(1,1)=10
 A(2,1)=-3
 A(3,1)=5
 A(1,2)=-7
 A(2,2)=2
 A(3,2)=-1
 A(1,3)=0
 A(2,3)=6
 A(3,3)=5
 DO 1 I=1,N
WRITE(6,2) (A(I,J), J=1,N)
1    CONTINUE
2    FORMAT (1H,10F5.0)
 WRITE (6,8)
 CALL DECOMP(NDIM,N,A,COND,IPVT,WORK)
 WRITE (6,3) COND
3    FORMAT (6H COND=, E15.5)
 WRITE (6,8)
 CONDP1=COND+1
 IF (CONDP1.EQ.COND) WRITE (6,4)
4    FORMAT (40H MATRIX IS SINGULAR TO WORKING PRECISION)
 IF (CONDP1 .EQ. COND) STOP
 B(1)=7
 B(2)=4
 B(3)=6
 DO 5 I=1,N
WRITE(6,2) B(I)
5    CONTINUE
```

```
 WRITE(6,8)
 CALL SOLVE(NDIM,N,A,B,IPVT)
 DO 6 I=1,N
WRITE (6,7) B(I)
6    CONTINUE
7    FORMAT (1H, F10.5)
 STOP
8    FORMAT(1H )
 END
C
SUBROUTINE DECOMP (NDIM,N,A,COND,IPVT,WORK)
C
 INTEGER NDIM, N
 REAL A(NDIM,N),COND,WORK(N)
 INTEGER IPVT(N)
C
C     Программа вычисляет разложение  вещественной матрицы
C     посредством гауссова исключения и оценивает обусловленность
C     матрицы
C
C     она используется для вычисления решений линейных систем
C
C     входная информация..
C
C    NDIM=заявленная строчная размерность массива, содержащего A
C
C    N=порядок матриц.
C
C    A=матрица, которую нужно разложить.
C
C     выходная информация..
C
C    A содержит верхнюю треугольную матрицу U и учитывающую
C     перестановки версию нижней треугольной матрицы I-L, такие,
C     что (матрица перестановок) *A=L*U
C
C    COND=оценка обусловленности A.
C      для линейной системы A*X=B изменения в A и B могут вызвать
C     изменения в X, большие в COND раз. Если COND+1.0.EQ.COND, то A
C     в пределах машинной точности является вырожденной
C     матрицей. COND полагается равным 1.0F+32, если обнаружена точная
C     вырожденность.
C
C     IPVT=вектор ведущих элементов.
C    IPVT(K)=индекс k-го ведущей строки
C    IPVT(N)=(-1)**(число перестановок)
C
```

```
C     рабочее поле.. вектор WORK должен быть описан и включен в вызов.
C           его входное содержимое обычно не дает важной ин-
C           формации.
C
C     определитель матрицы A может быть получен на выходе по формуле
C       DET(A)=IPVT(N)*A(1,1)*A(2,2)*...*A(N,N).
C
  REAL EK,T, ANORM,YNORM,ZNORM
  INTEGER NM1, I, J, K, KP1, KB, KM1, M
C
  IPVT(N)=1
  IF (N.EQ.1) GO TO 80
  NM1=N-1
C
C     вычислить 1-норму матрицы A
C
   ANORM=0.0
   DO 10 J=1,N
 T=0.0
 DO 5 I=1,N
T=T+ABS(A(I,J))
 5   CONTINUE
 IF (T.GT.ANORM) ANORM=T
 10   CONTINUE
C
C     гауссово исключение с частичным выбором ведущего элемента
C
  DO 35 K=1,NM1
 KP1=K+1
C
C     найти ведущий элемент
C
   M=K
DO 15 I=KP1,N
  IF(ABS(A(I,K)).GT.ABS(A(M,K))) M=I
 15  CONTINUE
 IPVT(K)=M
IF (M.NE.K) IPVT(N)=-IPVT(N)
T=A(M,K)
A(M,K)=A(K,K)
A(K,K)=T
C
C      пропустить этот шаг, если ведущий элемент равен нулю
C
   IF (T.EQ.0.0) GO TO 35
C
C      вычислить множители
```

```
C
   DO 20 I=KP1,N
   A(I,K)=-A(I,K)/T
 20    CONTINUE
C
C     переставлять  и исключать по столбцам
C
  DO 30 J=KP1,N
 T=A(M,J)
 A(M,J)=A(K,J)
 A(K,J)=T
 IF (T.EQ.0.0) GO TO 30
 DO 25 I=KP1,N
A(I,J)=A(I,J)+A(I,K)*T
 25      CONTINUE
 30    CONTINUE
 35  CONTINUE
C
C     COND=(1-норма матрицы A)*(оценка для 1-нормы  матрицы, обратной
C     к A)
C     оценка получается посредством одного шага метода обратных
C     итераций для наименьшего сингулярного вектора. Это требует
C     решения двух систем уравнений  (транспонированная для A)
C     *Y=E i A*Z=Y,gde E- вектор из +1 i =1, выбранный так, чтобы
C      максимизировать величину Y.
C     ESTIMATE=(1-норма Z)/(1-норма Y)
C
C     решить систему (транспонированная для A)*Y=E
C
  DO 50 K=1,N
 T=0.0
 IF(K.EQ.1) GO TO 45
 KM1=K-1
 DO 40 I=1,KM1
T=T+A(I,K)*WORK(I)
 40  CONTINUE
 45  EK=1.0
 IF (T.LT.0.0) EK=-1.0
 IF (A(K, K) .EQ. 0.0) GO TO 90
 WORK(K)=-(EK+T)/A(K,K)
 50    CONTINUE
  DO 60 KB=1,NM1
 K=N-KB
 T=0.0
 KP1=K+1
 DO 55 I=KP1,N
 T=T+A(I,K)*WORK(K)
```

```
 55  CONTINUE
 WORK(K)=T
 M=IPVT(K)
 IF(M.EQ.K) GO TO 60
 T=WORK(M)
 WORK(M)=WORK(K)
 WORK(K)=T
 60  CONTINUE
C
  YNORM=0.0
  DO 65 I=1,N
 YNORM=YNORM+ABS(WORK(I))
 65   CONTINUE
C
C     решить систему A*Z=Y
C
  CALL SOLVE(NDIM,N,A,WORK,IPVT)
C
  ZNORM=0.0
  DO 70 I=1,N
 ZNORM=ZNORM+ABS(WORK(I))
 70   CONTINUE
C
C     оценить обусловленность
C
  COND=ANORM*ZNORM/YNORM
  IF (COND.LT.1.0) COND=1.0
  RETURN
C
C     случай матрицы 1*1
C
 80   COND=1.0
  IF(A(1,1).NE.0.0) RETURN
C
C     точная вырожденность
C
 90   COND=1.0E+32
  RETURN
  END
SUBROUTINE SOLVE(NDIM,N,A,B,IPVT)
C
   INTEGER NDIM,N,IPVT(N)
   REAL A(NDIM,N),B(N)
C
C     решение линейной системы A*X=B
C     подпрограмму  не следует использовать, если DECOMP обнаружила
C     вырожденность
```

```
C
C     входная информация..
C
C        NDIM=заявленная строчная размерность массива, содержащего A.
C
C   A=факторизованная матрица, полученная из DECOMP
C
C      B=вектор правых частей.
C
C      IPVT=вектор ведущих элементов, полученный из DECOMP
C
C      выходная информация
C
C      B=вектор решения X.
C
  INTEGER KB,KM1,NM1,KP1,I,K,M
  REAL T
C
C     прямой ход
C
  IF(N.EQ.1)GO TO 50
  NM1=N-1
  DO 20 K=1,NM1
 KP1=K+1
 M=IPVT(K)
 T=B(M)
 B(M)=B(K)
 B(K)=T
 DO 10 I=KP1,N
B(I)=B(I)+A(I,K)*T
 10    CONTINUE
 20   CONTINUE
C
C     обратная подстановка
C
  DO 40 KB=1,NM1
 KM1=N-KB
 K=KM1+1
 B(K)=B(K)/A(K,K)
 T=-B(K)
 DO 30 I=1,KM1
B(I)=B(I)+A(I,K)*T
 30  CONTINUE
 40   CONTINUE
 50   B(1)=B(1)/A(1,1)
  RETURN
  END
```