2

Решение нелинейных уравнений $f(x) = 0$

Рассмотрим следующую задачу из физики. Шар радиуса r погружен в воду на глубину d (рис. 2.1). Предположим, что шар имеет радиус $r = 10$ см и сделан из старой сосны, имеющей плотность $\rho = 0{,}638$. Какая часть шара будет находиться в воде?

Масса вытесненной воды M_w, когда шар погружен в воду на глубину d, равна

$$M_w = \int_0^d \pi(r^2 - (x - r)^2)\, dx = \frac{\pi d^2(3r - d)}{3},$$

и масса шара равна $M_b = 4\pi r^3\rho/3$. Применяя закон Архимеда $M_w = M_b$, получим следующее уравнение, которое необходимо решить:

$$\frac{\pi(d^3 - 3d^2r + 4r^3\rho)}{3} = 0.$$

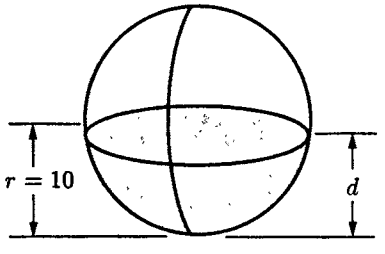


Рис. 2.1. Часть шара радиуса r, погруженного на глубину d

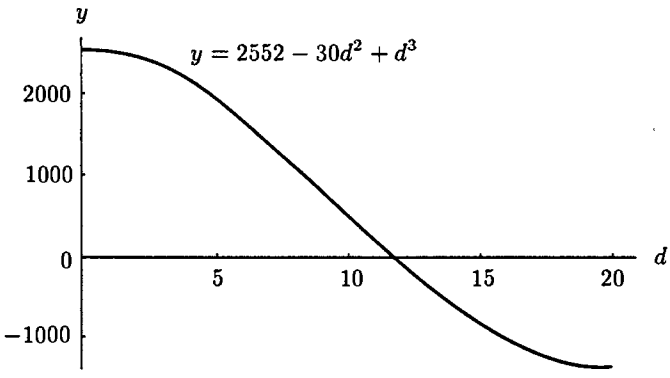


Рис. 2.2. Кубическое уравнение $y = 2552 - 30d^2 + d^3$

В нашем случае (при $r = 10$ и $\rho = 0{,}638$) это уравнение становится таким:

$$\frac{\pi(2552 - 30d^2 + d^3)}{3} = 0.$$

График кубического полинома $y = 2552 - 30d^2 + d^3$ показан на рис. 2.2, и можно увидеть, что решение находится около значения $d = 12$. В этом разделе излагается ряд методов приближенного нахождения корней уравнения. Например, метод деления пополам дает три корня: $d_1 = -8{,}17607212$, $d_2 = 11{,}86150151$ и $d_3 = 26{,}31457061$. Первый корень, d_1, является недопустимым решением этой задачи, так как d не может быть отрицательным. Третий корень, d_3, больше, чем диаметр сферы, и это также недопустимо. Корень $d_2 = 11{,}86150151$ лежит на интервале $[0; 20]$ и является правильным решением. Его величина приемлема, потому что немного больше половины шара должно быть погружено.

2.1. Использование итерации для решения уравнения $x = g(x)$

Фундаментальным принципом компьютерной науки является *итерация*. Как подсказывает название, процесс повторяется до тех пор, пока не будет получен ответ. Техника итераций используется для нахождения корней уравнений, решения систем линейных и нелинейных уравнений и решения дифференциальных уравнений. В этом разделе рассматривается процесс итераций, использующий повторные подстановки.

Для последовательного вычисления членов итерации необходимы правило или функция $g(x)$ и начальное значение p_0. Тогда получаем последовательность значе-

ний $\{p_k\}$, используя правило $p_{k+1} = g(p_k)$. Последовательность имеет такой вид:

$$
\begin{aligned}
& p_0 && \text{(начальное значение)} \\
& p_1 = g(p_0) \\
& p_2 = g(p_1) \\
& \vdots \\
& p_k = g(p_{k-1}) \\
& p_{k+1} = g(p_k) \\
& \vdots
\end{aligned}
\qquad (1)
$$

Что можно сказать о бесконечной последовательности чисел? Если последовательность чисел стремится к пределу, мы чувствуем, что он в некотором роде достигается. Но что происходит, если последовательность чисел расходится или является периодической? В следующем примере описана такая ситуация.

Пример 2.1. Правило итерации $p_0 = 7$ и $p_{k+1} = 1{,}001p_k$ для $k = 0, 2, \ldots$ порождает расходящуюся последовательность. Первые ее 200 членов выглядят следующим образом:

$$
\begin{aligned}
p_1 = 1{,}001p_0 &= (1{,}001)(1{,}000000) = 1{,}001000, \\
p_2 = 1{,}001p_1 &= (1{,}001)(1{,}001000) = 1{,}002001, \\
p_3 = 1{,}001p_2 &= (1{,}001)(1{,}002001) = 1{,}003003, \\
\vdots \qquad &\quad \vdots \qquad\qquad\quad \vdots \\
p_{100} = 1{,}001p_{99} &= (1{,}001)(1{,}104012) = 1{,}105116.
\end{aligned}
$$

Процесс можно продолжать бесконечно, и легко показать, что $\lim_{n\to\infty} p_n = +\infty$. В разделе 9 показано, что последовательность $\{p_k\}$ является численным решением дифференциального уравнения $y' = 0{,}001y$. Как известно, решением уравнения является $y(x) = e^{5{,}001x}$. Действительно, если сравнить сотый член последовательности с $y(100)$, то получится, что $p_{100} = 1{,}105116 \approx 1{,}185171 = e^{0{,}1} = y(100)$. ■

В этом разделе будут рассмотрены типы функций $g(x)$, которые производят сходящиеся последовательности p_k.

Нахождение неподвижных точек

Определение 2.1 (неподвижная точка). ***Неподвижной точкой*** функции $g(x)$ называется такое действительное число P, что $P = g(P)$. ▲

Геометрически неподвижные точки функции $y = g(x)$ — это точки пересечения $y = g(x)$ и $y = x$.

Определение 2.2 (итерация неподвижной точки). Итерация $p_{n+1} = g(p_n)$ для $n = 0, 1, \ldots$ называется ***итерацией неподвижной точки.*** ▲

Теорема 2.1. Предположим, что g — непрерывная функция и что $\{p_n\}_{n=0}^{\infty}$ — последовательность, сгенерированная с помощью итерации неподвижной точки. Если $\lim_{n\to\infty} p_n = P$, то P является неподвижной точкой $g(x)$.

Доказательство. Если $\lim_{n\to\infty} p_n = P$, то $\lim_{n\to\infty} p_{n+1} = P$. Из непрерывности g и соотношения $p_{n+1} = g(p_n)$ следует, что

$$g(P) = g\left(\lim_{n\to\infty} p_n\right) = \lim_{n\to\infty} g(p_n) = \lim_{n\to\infty} p_{n+1} = P. \tag{2}$$

Таким образом, P является неподвижной точкой $g(x)$. •

Пример 2.2. Рассмотрим сходящуюся итерацию

$$p_0 = 0{,}5 \quad \text{и} \quad p_{k+1} = e^{-p_k} \quad \text{для } k = 0, 1, \ldots.$$

Первые 10 членов получаем с помощью следующих вычислений:

$$\begin{aligned}
p_1 &= e^{-0{,}500000} = 0{,}606531,\\
p_2 &= e^{-0{,}606531} = 0{,}545239,\\
p_3 &= e^{-0{,}545239} = 0{,}579703,\\
&\vdots \qquad\qquad \vdots\\
p_9 &= e^{-0{,}566409} = 0{,}567560,\\
p_{10} &= e^{-0{,}567560} = 0{,}566907.
\end{aligned}$$

Последовательность сходится, и дальнейшие вычисления показывают, что

$$\lim_{n\to\infty} p_n = 0{,}567143\ldots.$$

Так мы нашли приближение для неподвижной точки функции $y = e^{-x}$. ■

Условия существования неподвижной точки и сходимость итерационного процесса неподвижной точки к неподвижной точке устанавливаются в следующих двух теоремах.

Теорема 2.2. Предположим, что $g \in C[a; b]$.

(3) Если область отображения $y = g(x)$ удовлетворяет условию $y \in [a; b]$ для всех $x \in [a; b]$, то g имеет неподвижную точку на $[a; b]$.

(4) Кроме того, предположим, что $g'(x)$ определена на $(a; b)$ и что существует положительная константа $K < 1$, такая, что $|g'(x)| \leq K < 1$ для всех $x \in$ $\in (a; b)$. Тогда g имеет единственную неподвижную точку P на $[a; b]$.

Доказательство (3). Если $g(a) = a$ или $g(b) = b$, то утверждение справедливо. Иначе значения $g(a)$ и $g(b)$ должны удовлетворять условиям $g(a) \in (a; b]$ и $g(b) \in$ $\in [a; b)$. Функция $f(x) = x - g(x)$ обладает следующим свойством:

$$f(a) = a - g(a) < 0 \quad \text{и} \quad f(b) = b - g(b) > 0.$$

Теперь применяем теорему 1.2 и теорему о промежуточном значении к $f(x)$, когда константа $L = 0$, и заключаем, что существует такое число P, $P \in (a; b)$, при котором $f(P) = 0$. Следовательно, $P = g(P)$ и P является требуемой неподвижной точкой $g(x)$.

Доказательство (4). Сейчас нужно показать, что это решение единственное. Доказываем от противного, предположив, что существуют две неподвижные точки P_1 и P_2. Теперь применяем теорему 1.6 и теорему о среднем значении и приходим к заключению, что существует такое число $d \in (a; b)$, что

$$g'(d) = \frac{g(P_2) - g(P_1)}{P_2 - P_1}. \tag{5}$$

Далее, учитывая тот факт, что $g(P_1) = P_1$ и $g(P_2) = P_2$, упростим правую часть выражения (5) и получим

$$g'(d) = \frac{P_2 - P_1}{P_2 - P_1} = 1.$$

Но это противоречит гипотезе (4), что $|g'(x)| < 1$ на $(a; b)$, поэтому существование двух неподвижных точек невозможно. Следовательно, при условиях, заданных в (4), $g(x)$ имеет единственную неподвижную точку P на $(a; b)$. •

Пример 2.3. Применим теорему 2.2, чтобы строго доказать, что $g(x) = \cos(x)$ имеет единственную неподвижную точку.

Очевидно, что $g \in C[0; 1]$. Кроме того, $g(x) = \cos(x)$ на интервале $[0; 1]$ — убывающая функция. Таким образом, выполняется условие (3) теоремы 2.2 и g имеет неподвижную точку на $[0; 1]$. Наконец, если $x \in (0; 1)$, то $|g'(x)| =$ $= |-\sin(x)| = \sin(x) \leq \sin(1) < 0{,}8415 < 1$. Поэтому $K = \sin(1) < 1$, и выполняется условие (4) теоремы 2.2. Значит, g имеет единственную неподвижную точку на $[0; 1]$. ■

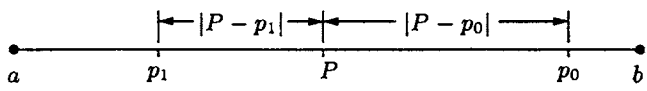


Рис. 2.3. Соотношения между P, p_0, p_1, $|P - p_0|$ и $|P - p_1|$

Теорема 2.3 (теорема о неподвижной точке). Предположим, что (i) g', $g \in C[a; b]$, (ii) K — положительная константа, (iii) $p_0 \in (a; b)$, (iv) $g(x) \in [a; b]$ для всех $x \in [a; b]$.

(6) Если $|g'(x)| \leq K < 1$ для всех $x \in [a; b]$, то итерация $p_n = g(p_{n-1})$ будет сходиться к единственной неподвижной точке $P \in [a; b]$. В этом случае говорят, что P — неподвижная точка притяжения.

(7) Если $|g'(x)| > 1$ для всех $x \in [a; b]$, то итерация $p_n = g(p_{n-1})$ не сходится к P. В этом случае говорят, что P — неподвижная точка отталкивания и итерация проявляет локальную расходимость.

Замечание 1. В утверждении (7) предполагается, что $p_0 \neq P$.
Замечание 2. Так как g непрерывна на интервале, содержащем P, допустимо использование более простых критериев $|g'(P)| \leq K<1$ и $|g'(P)| > 1$ в (6) и (7) соответственно.

Доказательство. Прежде всего покажем, что все точки $\{p_n\}_{n=0}^{\infty}$ лежат на интервале $(a; b)$. Начиная с p_0, применим теорему 1.6 и теорему о среднем значении. Существует такое значение $c_0 \in (a; b)$, что

$$\begin{aligned} |P - p_1| &= |g(P) - g(p_0)| = |g'(c_0)(P - p_0)| = \\ &= |g'(c_0)||P - p_0| \leq K|P - p_0| < |P - p_0|. \end{aligned} \tag{8}$$

Поэтому p_1 не более удалено от P, чем p_0, из чего следует, что $p_1 \in (a; b)$ (рис. 2.3). В общем случае предположим, что $p_{n-1} \in (a; b)$. Тогда

$$\begin{aligned} |P - p_n| &= |g(P) - g(p_{n-1})| = |g'(c_{n-1})(P - p_{n-1})| = \\ &= |g'(c_{n-1})||P - p_{n-1}| \leq K|P - p_{n-1}| < |P - p_{n-1}|. \end{aligned} \tag{9}$$

Таким образом, $p_{n-1} \in (a; b)$ и по индукции все точки $\{p_n\}_{n=0}^{\infty}$ лежат на интервале $(a; b)$.

Чтобы завершить доказательство (6), осталось показать, что

$$\lim_{n \to \infty} |P - p_n| = 0. \tag{10}$$

Прежде всего докажем по индукции неравенство

$$|P - p_n| \leq K^n |P - p_0|. \tag{11}$$

Случай, когда $n = 1$, следует из соотношения (8). Используя предположение индукции $|P - p_{n-1}| \leq K^{n-1}|P - p_0|$ и соотношение (9), получим

$$|P - p_n| \leq K|P - p_{n-1}| \leq KK^{n-1}|P - p_0| = K^n|P - p_0|.$$

Следовательно, по индукции неравенство (11) выполняется для всех n. Так как $0 < K < 1$, член K^n стремится к нулю, когда n стремится к бесконечности. Отсюда

$$0 \leq \lim_{n\to\infty} |P - p_n| \leq \lim_{n\to\infty} K^n|P - p_0|. \tag{12}$$

Предел $|P - p_n|$ заключен между нулем слева и нулем справа, поэтому можно сделать вывод, что $\lim_{n\to\infty} |P - p_n| = 0$. Таким образом, $\lim_{n\to\infty} p_n = P$ и согласно теореме 2.1 итерация $p_n = g(p_{n-1})$ сходится к неподвижной точке P. Следовательно, утверждение (6) теоремы 2.3 доказано. Оставляем читателю доказательство правильности утверждения (7). •

Следствие 2.1. Предположим, что g удовлетворяет условиям (6) теоремы 2.3. Грани ошибки, возникающей при использовании приближения p_n для P, задаются формулами

$$|P - p_n| \leq K^n|P - p_0| \quad \text{для всех} \quad n \geq 1 \tag{13}$$

и

$$|P - p_n| \leq \frac{K^n|p_1 - p_0|}{1 - K} \quad \text{для всех} \quad n \geq 1. \tag{14}$$

Графическая интерпретация итерации неподвижной точки

Для существования неподвижной точки P кривой $g(x)$ необходимо, чтобы график кривой $y = g(x)$ и прямая $y = x$ пересекались в точке (P, P). Два простых типа сходящейся итерации, монотонной и колеблющейся, показаны на рис. 2.4(a) и 2.4(b) соответственно.

Чтобы проследить за процессом, начинаем с точки p_0 на оси x и двигаемся вертикально к точке $(p_0; p_1) = (p_0; g(p_0))$ на кривой $y = g(x)$. Затем двигаемся горизонтально от $(p_0; p_1)$ к точке $(p_1; p_1)$ на прямой $y = x$. Наконец, двигаемся вертикально вниз к p_1 на оси x-в. Рекуррентная формула $p_{n+1} = g(p_n)$ используется для построения точки $(p_n; p_{n+1})$ на графике. Затем, двигаясь по прямой, определяем положение точки $(p_{n+1}; p_{n+1})$ на прямой $y = x$. Вертикальное движение оканчивается в точке p_{n+1} на оси x. Эта ситуация показана на рис. 2.4. Если $|g'(P)| > 1$, то итерация $p_{n+1} = g(p_n)$ порождает последовательность, которая расходится от P. Два простых типа расходящейся итерации, монотонной и колеблющейся, показаны на рис. 2.5(a) и (b) соответственно.

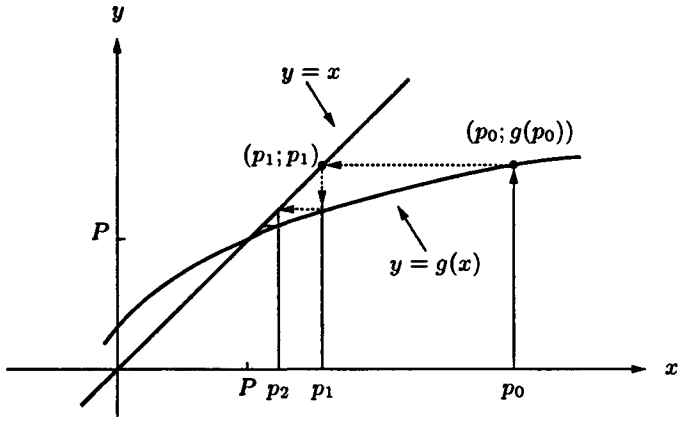


Рис. 2.4. (a) Монотонная сходимость, когда $0 < g'(P) < 1$

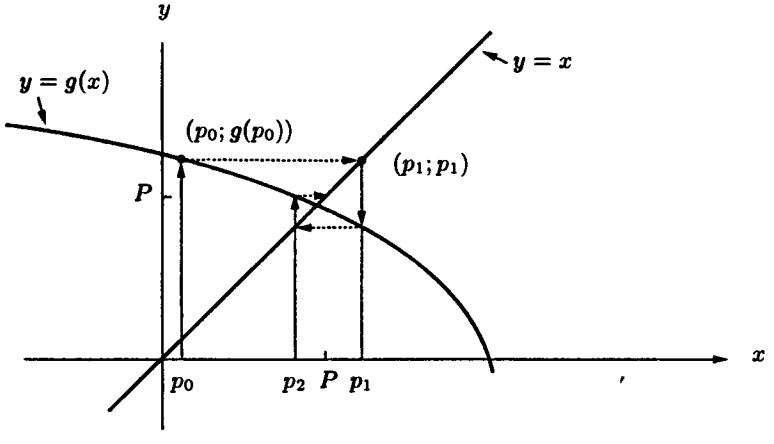


Рис. 2.4. (b) Колеблющаяся сходимость, когда $-1 < g'(P) < 0$

Пример 2.4. Рассмотрим итерацию $p_{n+1} = g(p_n)$, в которой используется функция $g(x) = 1 + x - x^2/4$. Неподвижные точки можно найти, решив уравнение $x = g(x)$. Решениями являются (неподвижные точки g) $x = -2$ и $x = 2$, производная функции равна $g'(x) = 1 - x/2$, и необходимо рассмотреть только два случая.

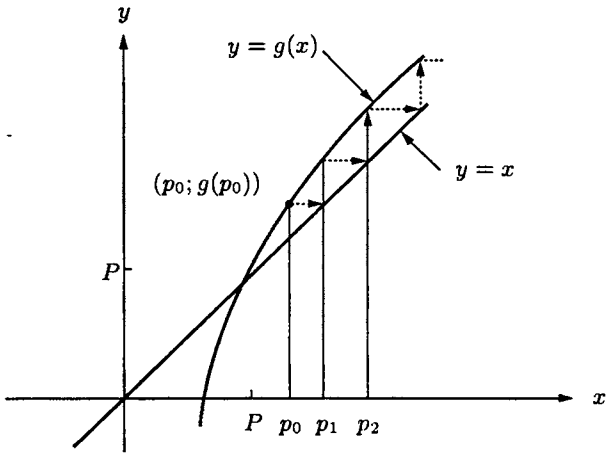


Рис. 2.5. (а) Монотонное расхождение, когда $1 < g'(P)$

Случай (i)	$P = -2$
Начинаем с	$p_0 = -2{,}05$
Тогда получаем	$p_1 = -2{,}100625$
	$p_2 = -2{,}20378135$
	$p_3 = -2{,}41794441$
	$\vdots$
	$\lim_{n\to\infty} p_n = -\infty$

Так как $|g'(x)| > \frac{3}{2}$ на $[-3; l-1]$, то согласно теореме 2.3 последовательность не будет сходиться к $P = -2$.

Случай (ii)	$P = 2$
Начинаем с	$p_0 = 1{,}6$
Тогда получаем	$p_1 = 1{,}96$
	$p_2 = 1{,}9996$
	$p_3 = 1{,}99999996$
	$\vdots$
	$\lim_{n\to\infty} p_n = 2$

Так как $|g'(x)| < \frac{1}{2}$ на $[1; 3]$, то согласно теореме 2.3 последовательность не будет сходиться к $P = 2$. ■

Теорема 2.3 не говорит, что произойдет, когда $g'(P) = 1$. Следующий пример специально построен так, что последовательность $\{p_n\}$ сходится, как только $p_0 > P$, и расходится, если выбрать $p_0 < P$.

Пример 2.5. Рассмотрим итерацию $p_{n+1} = g(n)$, когда функция $g(x) = 2(x - -1)^{1/2}$ для $x \geq 1$. Существует только одна неподвижная точка $P = 2$. Производная функции равна $g'(x) = 1/(x-1)^{1/2}$ и $g'(2) = 1$, поэтому теорема 2.3 не применяется. Рассмотрим два случая, когда начальное значение лежит слева или справа от точки $P = 2$.

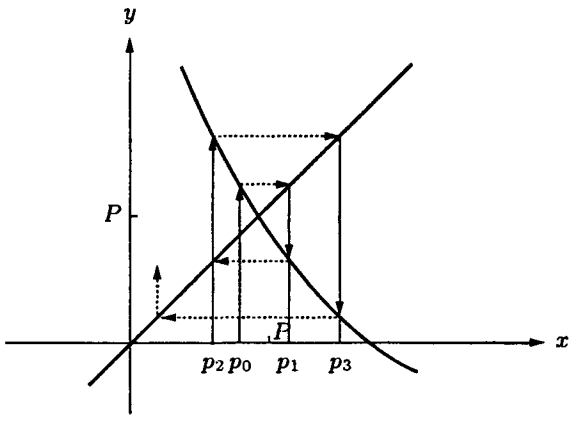


Рис. 2.5. (b) Расходящиеся колебания, когда $g'(P) < -1$

Случай (i) Начинаем с $p_0 = 1{,}5$.
Тогда получаем
$$p_1 = 1{,}41421356$$
$$p_2 = 1{,}28718851$$
$$p_3 = 1{,}07179943$$
$$p_4 = 0{,}53590832$$
$$\vdots$$
$$p_5 = 2(-0{,}46409168)^{1/2}$$

Поскольку p_4 лежит вне области определения $g(x)$, член p_5 не вычисляем.

Случай (ii) Начинаем с $p_0 = 2{,}5$.
Тогда получаем
$$p_1 = 2{,}44948974$$
$$p_2 = 2{,}40789513$$
$$p_3 = 2{,}37309514$$
$$p_4 = 2{,}34358284$$
$$\vdots$$
$$\lim_{n\to\infty} p_n = 2$$

Эта последовательность сходится также медленно к значению $P = 2$; действительно, $P_{1000} = 2{,}00398714$. ■

Абсолютная и относительная ошибки

В примере 2.5, случай (ii), последовательность сходится медленно, и после 1000 итераций три последовательных члена равны

$$p_{1000} = 2{,}00398714; \quad p_{1001} = 2{,}00398317 \quad \text{и} \quad p_{1002} = 2{,}00397921.$$

Но не следует возмущаться; затем можно вычислить еще несколько тысяч членов и найти лучшее приближение! Но что станет критерием для остановки итерации? Заметим, что можно использовать разницу между последовательными членами

$$|p_{1001} - p_{1002}| = |2{,}00398317 - 2{,}00397921| = 0{,}00000396.$$

Кроме того, известно, что абсолютная ошибка приближения p_{1000} равна

$$|P - p_{1000}| = |2{,}00000000 - 2{,}00398714| = 0{,}00398714.$$

Существует около 1000, членов больших, чем $|p_{1001} - p_{1002}|$, и это показывает, что близость последующих членов не дает гарантии, что точность достигнута. Но обычно это самый доступный критерий, и его часто используют, чтобы остановить итерационный процесс.

Программа 2.1 (итерация неподвижной точки). Получить приближенное решение уравнения $x = g(x)$, начав с предположительно угаданного значения p_0 и итерационного правила $p_{n+1} = g(p_n)$.

```
function [k,p,err,P]=fixpt(g,p0,tol,max1)
%  Вход - g - итерационная функция, вводимая как строка 'g'
%       - p0 - начальное угаданное значение неподвижной точки
%       - tol - допустимое отклонение
%       - max1 - максимальный член итерации
%Выход  - k - число произведенных итераций
%       - p - приближение для неподвижной точки
%       - err - ошибка приближения
%       - P - содержит последовательность {pn}
P(1)= p0;
for k=2:max1
    P(k)=feval(g,P(k-1));
    err=abs(P(k)-P(k-1));
    relerr=err/(abs(P(k))+eps);
    p=P(k);
    if (err<tol) | (relerr<tol),break;end
end
if k == max1
    disp('максимально допустимое число итераций')
end
P=P';
```

Замечание. Если использовать строго определенную функцию `fixpt`, то необходимо вводить M-файл g.m в виде строки: 'g' (см. MATLAB Дополнение).

Упражнения к разделу 2.1

1. Определите точно, имеет ли функция единственную неподвижную точку на заданном интервале (следуя примеру 2.3).
   (a) $g(x) = 1 - x^2/4$ на $[0; 1]$
   (b) $g(x) = 2^{-x}$ на $[0; 1]$

(c) $g(x) = 1/x$ на $[0{,}5; 5{,}2]$

2. Исследуйте природу итерации неподвижной точки, когда

$$g(x) = -4 + 4x - \frac{1}{2}x^2.$$

(a) Решите уравнение $g(x) = x$ и покажите, что $P = 2$ и $P = 4$ — неподвижные точки.

(b) Используйте в качестве начального значения $p_0 = 1{,}9$ и определите p_1, p_2 и p_3.

(c) Используйте $p_0 = 3{,}8$ в качестве начального значения и определите p_1, p_2 и p_3.

(d) Найдите ошибку E_k и относительную ошибку R_k значения p_k в пп. (b) и (c).

(e) Какой вывод можно сделать из теоремы 2.3?

3. В одной и той же системе координат заданы график $g(x)$, прямая $y = x$ и неподвижная точка P. Используя начальное значение p_0, вычислите p_1 и p_2. Постройте рисунок, подобный рис. 2.4 и 2.5. Исходя из своего графика, определите геометрически, когда итерация неподвижной точки сойдется.

(a) $g(x) = (6 + x)^{1/2}$, $P = 3$ и $p_0 = 7$

(b) $g(x) = 1 + 2/x$, $P = 2$ и $p_0 = 4$

(c) $g(x) = x^2/3$, $P = 3$ и $p_0 = 3{,}5$

(d) $g(x) = -x^2 + 2x + 2$, $P = 2$ и $p_0 = 2{,}5$

4. Пусть $g(x) = x^2 + x - 4$. Можно ли использовать итерацию неподвижной точки для нахождения корня (или корней) уравнения $x = g(x)$? Если можно, то объясните, почему?

5. Пусть $g(x) = x\cos(x)$. Решите уравнение $x = g(x)$ и определите все неподвижные точки g (их бесконечно много). Можно ли использовать итерацию неподвижной точки для нахождения корня (корней) уравнения $x = g(x)$? Объясните, почему?

6. Предположим, что $g(x)$ и $g'(x)$ определены и непрерывны на $(a; b)$; $p_0, p_1, p_2 \in (a; b)$, $p_1 = g(p_0)$ и $p_2 = g(p_1)$. Предположим также, что существует такая константа K, что $|g'(x)| < K$. Покажите, что $|p_2 - p_1| < K|p_1 - p_0|$. *Указание.* Воспользуйтесь теоремой о среднем значении.

7. Предположим, что $g(x)$ и $g'(x)$ непрерывны на интервале $(a; b)$ и что на этом интервале $|g'(x)| > 1$. Если неподвижная точка P и начальные приближения p_0 и p_1 лежат на интервале $(a; b)$, то покажите, что из $p_1 = g(p_0)$ следует, что $|E_1| = |P - p_1| > |P - p_0| = |E_0|$. Таким образом будет доказано утверждение (7) (локальная несходимость) теоремы 2.3.

8. Пусть $g(x) = -0{,}0001x^2 + x$ и $p_0 = 1$. Рассмотрите итерацию неподвижной точки.
   (a) Покажите, что $p_0 > p_1 > \cdots > p_n > p_{n+1} > \cdots$.
   (b) Покажите, что $p_n > 0$ для всех n.
   (c) Так как последовательность $\{p_n\}$ убывающая и ограничена снизу, то она имеет предел. Чему равен этот предел?
9. Пусть $g(x) = 0{,}5x + 1{,}5$ и $p_0 = 4$. Рассмотрите итерацию неподвижной точки.
   (a) Покажите, что неподвижной точкой является $P = 3$.
   (b) Покажите, что $|P - p_n| = |P - p_{n-1}|/2$ для $n = 1, 2, 3, \ldots$.
   (c) Покажите, что $|P - p_n| = |P - p_0|/2^n$ для $n = 1, 2, 3, \ldots$.
10. Пусть $g(x) = x/2$. Рассмотрите итерацию неподвижной точки.
   (a) Найдите величину $|p_{k+1} - p_k|/|p_{k+1}|$.
   (b) Что случится, если только относительная ошибка явится критерием остановки, как в программе 2.1?
11. Почему для итерации неподвижной точки благоприятно, когда $g'(P) \approx 0$?

Алгоритмы и программы

1. Используйте программу 2.1, чтобы приблизить неподвижные точки (если они есть) следующих функций. Должна ли точность иметь 12 десятичных знаков? Постройте график каждой функции и прямой $y = x$, где ясно показана любая из неподвижных точек.
   (a) $g(x) = x^5 - 3x^3 - 2x^2 + 2$
   (b) $g(x) = \cos(\sin(x))$
   (c) $g(x) = x^2 - \sin(x + 0{,}15)$
   (d) $g(x) = x^{x - \cos(x)}$

2.2. Методы интервалов локализации корня

Рассмотрим для интереса знакомый пример. Предположим, что вы ежемесячно вкладываете в банк сумму, равную P, и ежегодный процент дохода равен I. Тогда общая сумма A после N вкладов составляет

$$A = P + P\left(1 + \frac{I}{12}\right) + P\left(1 + \frac{I}{12}\right)^2 + \cdots + P\left(1 + \frac{I}{12}\right)^{N-1}. \tag{1}$$

Первый член справа равенства (1) равен последнему вкладу. Вклад, который благодаря процентам, выплаченным за один период, равен $P\left(1 + \frac{I}{12}\right)$. Еще более

ранний вклад теперь составляет $P\left(1+\frac{I}{12}\right)^2$ и т. д. Наконец, первый взнос, сделанный $N-1$ месяц тому назад, составляет $P\left(1+\frac{I}{12}\right)^{N-1}$. Напомним, что сумма N членов геометрического ряда равна

$$1+r+r^2+r^3+\cdots+r^{N-1}=\frac{1-r^N}{1-r}. \tag{2}$$

Выражение (1) можно записать в виде

$$A=P\left(1+\left(1+\frac{I}{12}\right)+\left(1+\frac{I}{12}\right)^2+\cdots+\left(1+\frac{I}{12}\right)^{N-1}\right),$$

и, выполнив замену $r=(1+I/12)$ в (2), получить

$$A=P\frac{1-(1+\frac{I}{12})^N}{1-(1+\frac{I}{12})}.$$

Это выражение можно упростить, чтобы получить формулу для ежегодно причитающейся ренты:

$$A=\frac{P}{I/12}\left(\left(1+\frac{I}{12}\right)^N-1\right). \tag{3}$$

В следующем примере равенство используется для вычисления ежегодно причитающейся ренты и рекуррентная последовательность вычислений — для нахождения ответа.

Пример 2.6. Будем платить \$250 в месяц в течение 20 лет и потребуем, чтобы общая сумма всех вкладов и процентов через 20 лет была равна \$250 000. Какой должна быть процентная ставка I для достижения этой цели? Если зафиксировать $N=240$, то A — функция только от I, т. е. $A=A(I)$. Начнем с двух предположений, что $I_0=0{,}12$ и $I_1=0{,}13$, и, выполнив последовательность вычислений, придем к окончательному ответу. Начиная с $I_0=0{,}12$, получим

$$A(0{,}12)=\frac{250}{0{,}12/12}\left(\left(1+\frac{0{,}12}{12}\right)^{240}-1\right)=247\,314.$$

Поскольку это значение является несколько меньшим, чем требуется в задаче, следующей попыткой будет присвоение $I_1=0{,}13$:

$$A(0{,}13)=\frac{250}{0{,}13/12}\left(\left(1+\frac{0{,}13}{12}\right)^{240}-1\right)=282\,311.$$

Это значение несколько больше, поэтому попытаемся взять среднее значение $I_2 = 0{,}125$:

$$A(0{,}125) = \frac{250}{0{,}125/12}\left(\left(1 + \frac{0{,}125}{12}\right)^{240} - 1\right) = 264\,623.$$

Полученное значение снова больше, и можно заключить, что требуемая процентная ставка лежит на интервале $[0{,}12; 0{,}125]$. Следующее предположение — средняя точка $I_3 = 0{,}1225$:

$$A(0{,}1225) = \frac{250}{0{,}1225/12}\left(\left(1 + \frac{0{,}1225}{12}\right)^{240} - 1\right) = 255\,803.$$

Это снова больше, и сейчас интервал сузился до $[0{,}12; 0{,}1225]$. В последнем вычислении опять используется средняя точка $I_4 = 0{,}12125$:

$$A(0{,}12125) = \frac{250}{0{,}12125/12}\left(\left(1 + \frac{0{,}12125}{12}\right)^{240} - 1\right) = 251\,518.$$

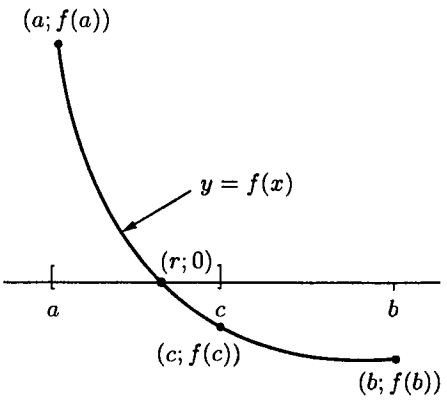
Последующая итерация дает возможность получить столько значащих цифр, сколько требуется. Назначение этого примера — найти значение I, которое обеспечивает точно установленное значение функции L (оно является решением уравнения $A(I) = L$). Обычно константа L стоит слева и решается уравнение $A(I) - L = 0$. ■

Определение 2.3 (корень уравнения, нуль функции). Предположим, что $f(x)$ — непрерывная функция. Любое число r, для которого $f(r) = 0$, называется ***корнем уравнения*** $f(x) = 0$. Говорят также, что r является ***нулем функции*** $f(x)$. ▲

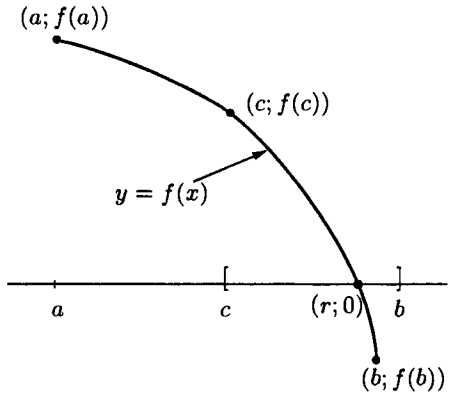
Например, уравнение $2x^2 + 5x - 3 = 0$ имеет два действительных нуля $r_1 = 0{,}5$ и $r_2 = -3$, в то время как соответствующая функция $f(x) = 2x^2 + 5x - 3 = (2x - 1)(x + 3)$ имеет два действительных корня: $r_1 = 0{,}5$ и $r_2 = -3$.

Метод Больцано деления пополам (метод бисекции)

В этом разделе рассматривается первый метод интервалов для нахождения нулей непрерывной функции. Начнем с исходного интервала $[a; b]$, на котором $f(a)$ и $f(b)$ имеют противоположные знаки. Так как график непрерывной функции $y = f(x)$ является непрерывным, он пересекает ось x в точке $x = r$, которая лежит где-то на интервале (рис. 2.6). Метод деления пополам сдвигает крайние точки все ближе и ближе, пока на интервале не получится произвольно малый отрезок,



(a) Если $f(a)$ и $f(c)$ имеют противоположные знаки, то выполнять сжатие справа.



(b) Если $f(c)$ и $f(b)$ имеют противоположные знаки, то выполнять сжатие слева.

Рис. 2.6. Решение методом деления пополам

содержащий нуль функции. Решающим шагом процесса деления интервала пополам является выбор средней точки $c = (a + b)/2$ и анализ трех возможностей, которые могут возникнуть.

(4) Если $f(a)$ и $f(c)$ имеют разные знаки, нуль лежит на интервале $[a; c]$.

(5) Если $f(c)$ и $f(b)$ имеют разные знаки, нуль лежит на интервале $[c; b]$.

(6) Если $f(c) = 0$, значит, нулем является c.

В любом из двух случаев, (4) или (5), мы рассматриваем половину интервала как начальный интервал, который содержит корень, и "сжимаем его" (см. рис. 2.6). Продолжаем процесс, обозначая новый меньший интервал $[a; b]$ и повторяя процесс до тех пор, пока интервал не станет настолько малым, насколько требуется. Таким образом, процесс деления пополам включает последовательность вложенных интервалов и их средних точек. Используем следующее обозначение, чтобы отслеживать детали процесса.

(7)

$[a_0; b_0]$ — начальный интервал и $c_0 = \frac{a_0+b_0}{2}$ — средняя точка.

$[a_1; b_1]$ — второй интервал, который содержит нуль r, и c_1 — его средняя точка; длина интервала $[a_1; b_1]$ равна половине длины $[a_0; b_0]$.

После достижения n-го интервала $[a_n; b_n]$, который включает r и имеет среднюю точку c_n, строим интервал $[a_{n+1}; b_{n+1}]$, который также содержит r и длина которого равна половине длины $[a_n; b_n]$.

Оставим в качестве упражнения для читателя доказательство того, что последовательность левых конечных точек убывающая, а последовательность правых крайних точек — возрастающая, т. е.

$$a_0 \le a_1 \le \cdots \le a_n \le \cdots \le r \le \cdots \le b_n \le \cdots \le b_1 \le b_0, \tag{8}$$

где $c_n = \frac{a_n + b_n}{2}$, и если $f(a_{n+1}) f(b_{n+1}) < 0$, то

$$[a_{n+1}; b_{n+1}] = [a_n; c_n] \quad \text{или} \quad [a_{n+1}; b_{n+1}] = [c_n; b_n] \quad \text{для всех } n. \tag{9}$$

Теорема 2.4 (теорема о делении отрезка пополам). Предположим, что $f \in C[a; b]$ и существует такое число $r \in [a; b]$, что $f(r) = 0$. Если $f(a)$ и $f(b)$ имеют различные знаки и $\{c_n\}_{n=0}^{\infty}$ представляет последовательность средних точек, полученных в результате деления пополам (8) и (9), то

$$|r - c_n| \le \frac{b - a}{2^{n+1}} \quad \text{для } n = 0, 1, \ldots, \tag{10}$$

и, значит, последовательность $\{c_n\}_{n=0}^{\infty}$ сходится к нулю функции $x = r$, т. е.

$$\lim_{n \to \infty} c_n = r. \tag{11}$$

Доказательство. Так как и нуль r и средняя точка c_n лежат на интервале $[a_n; b_n]$, расстояние между c_n и r не может быть больше половины длины этого интервала (рис. 2.7). Поэтому

$$|r - c_n| \le \frac{b_n - a_n}{2} \quad \text{для всех } n. \tag{12}$$

Заметим, что следующие один за другим интервалы образуют последовательность

$$b_1 - a_1 = \frac{b_0 - a_0}{2^1},$$

$$b_2 - a_2 = \frac{b_1 - a_1}{2} = \frac{b_0 - a_0}{2^2}.$$

Оставим читателю в качестве упражнения доказать с помощью математической индукции, что

$$b_n - a_n = \frac{b_0 - a_0}{2^n}. \tag{13}$$

Объединив результаты (12) и (13), получим, что

$$|r - c_n| \le \frac{b_0 - a_0}{2^{n+1}} \quad \text{для всех } n. \tag{14}$$

А сейчас можно воспользоваться рассуждениями, сходными с приведенными в теореме 2.3, для доказательсва, что из (14) вытекает, что последовательность $\{c_n\}_{n=0}^{\infty}$ сходится к r, и теорема доказана. •

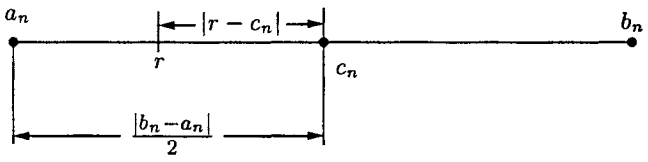


Рис. 2.7. Корень r и средняя точка c_n интервала $[a_n; b_n]$ для метода деления пополам

Пример 2.7. Функция $h(x) = x\sin(x)$ возникает при изучении недемпфированных форсированных колебаний. Найти значение x, которое лежит на интервале $[0; 2]$, на котором функция принимает значение $h(x) = 1$ (функция $\sin(x)$ вычисляется в радианах).

Воспользуемся методом деления пополам для нахождения нуля функции $f(x) = x\sin(x) - 1$. Начав с $a_0 = 0$ и $b_0 = 2$, вычислим

$$f(0) = -1{,}000000 \quad \text{и} \quad f(2) = 0{,}818595.$$

Таким образом, корень $f(x) = 0$ лежит на интервале $[0; 2]$. Средняя точка — $c_0 = 1$ и $f(1) = -0{,}158529$. Следовательно, функция меняет знак на интервале $[c_0; b_0] = [1; 2]$.

Далее выполняем сжатие влево и присваиваем $a_1 = c_0$ и $b_1 = b_0$. Средняя точка равна $c_1 = 1{,}5$ и $f(c_1) = 0{,}496242$. Из $f(1) = -0{,}158529$ и $f(1{,}5) = 0{,}496242$ вытекает, что корень лежит на интервале $[a_1; c_1] = [1{,}0; 1{,}5]$. Следующим шагом является сжатие вправо и присвоение $a_2 = a_1$ и $b_2 = c_1$. Таким образом получаем последовательность $\{c_k\}$, которая сходится к $r \approx 1{,}114157141$. Пример вычислений приведен в табл. 2.1. ■

Достоинством метода деления пополам является то, что формула (10) дает предопределенную оценку точности вычисляемого решения. В примере 2.7 начальная длина интервала равнялась $b_0 - a_0 = 2$. Предположим, что табл. 2.1 продлена до 31-й итерации. Тогда в силу (10) ошибка была бы ограничена значением $|E_{31}| \leq (2-0)/2^{32} \approx 4{,}656613 \times 10^{-10}$. Поэтому c_{31} было бы приближением к r с 9-ю десятичными знаками точности. N повторяемых делений пополам, необходимых для гарантии того, что N-я средняя точка c_N является приближением к нулю функции и ошибка приближения меньше, чем наперед заданное значение δ, равно

$$N = \left[\left(\frac{\ln(b-a) - \ln(\delta)}{\ln(2)}\right)\right]. \tag{15}$$

Доказательство этой формулы оставлено читателю в качестве упражнения.

Другим популярным алгоритмом является ***метод ложного положения***, или ***метод regula falsi***. Его стали использовать потому, что метод деления пополам

Таблица 2.1. Решение уравнения $x\sin(x) - 1 = 0$ методом деления пополам

k	Крайняя слева точка, a_k	Средняя точка, c_k	Крайняя справа точка, b_k	Значение функции, $f(c_k)$
0	0	1,	2,	−0,158529
1	1,0	1,5	2,0	0,496242
2	1,00	1,25	1,50	0,186231
3	1,000	1,125	1,250	0,015051
4	1,0000	1,0625	1,1250	−0,071827
5	1,06250	1,09375	1,12500	−0,028362
6	1,093750	1,109375	1,125000	−0,006643
7	1,1093750	1,1171875	1,1250000	0,004208
8	1,10937500	1,11328125	1,11718750	−0,001216
⋮	⋮	⋮	⋮	⋮

сходится весьма медленно. Как и раньше, предполагаем, что $f(a)$ и $f(b)$ имеют противоположные знаки. В методе деления пополам средняя точка интервала $[a; b]$ используется как следующая итерация. Приближение получается лучше, если найти точку $(c; 0)$, в которой секущая L, соединяющая точки $(a; f(a))$ и $(b; f(b))$, пересекает ось x (рис. 2.8). Чтобы найти значение c, запишем ниже два варианта для тангенса угла наклона m прямой L:

$$m = \frac{f(b) - f(a)}{b - a}, \tag{16}$$

с использованием точек $(a; f(a))$ и $(b; f(b))$ и

$$m = \frac{0 - f(b)}{c - b}, \tag{17}$$

где используются точки $(c; 0)$ и $(b; f(b))$.

Приравнивая тангенсы угла наклона прямой в (16) и (17), получим уравнение

$$\frac{f(b) - f(a)}{b - a} = \frac{0 - f(b)}{c - b},$$

которое легко решаем относительно c, и

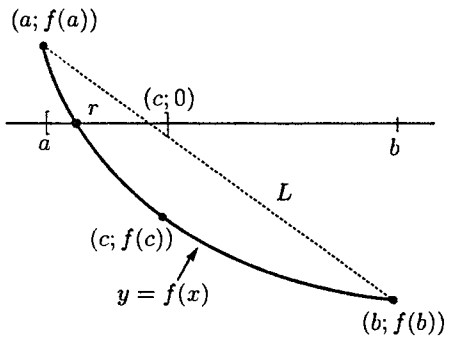
$$c = b - \frac{f(b)(b - a)}{f(b) - f(a)}. \tag{18}$$

Здесь, как и ранее, также существуют три возможности.

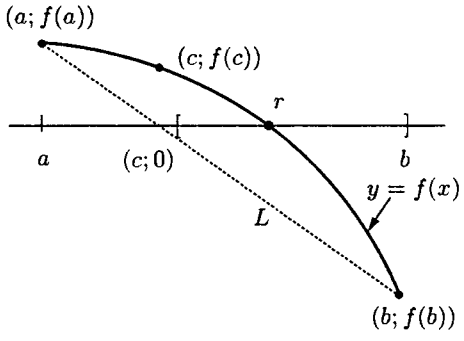
(19) Если $f(a)$ и $f(c)$ имеют различные знаки, нуль лежит на интервале $[a; c]$.

(20) Если$f(c)$ и $f(b)$ имеют различные знаки, нуль лежит на интервале $[c; b]$.

(21) Если $f(c) = 0$, значит, нулем является c.



(a) Если $f(a)$ и $f(c)$ имеют противоположные знаки, то выполняем сжатие справа.



(b) Если $f(c)$ и $f(b)$ имеют противоположные знаки, то выполняем сжатие слева.

Рис. 2.8. Процесс решения для метода ложного положения

Сходимость метода ложного положения

Процесс решения, включающий (19) и (20) вместе с (18), используется для построения последовательности интервалов $\{[a_n; b_n]\}$, каждый из которых содержит нуль. Каждый шаг приближения нуля r равен

$$c_n = b_n - \frac{f(b_n)(b_n - a_n)}{f(b_n) - f(a_n)}, \tag{22}$$

и можно доказать, что последовательность $\{c_n\}$ будет сходиться к r. Но необходима осторожность: хотя интервалы длины $b_n - a_n$ становятся все меньше, может случиться, что они не сходятся к нулю. Если график $y = f(x)$ вогнутый вблизи точки $(r; 0)$, одна крайняя точка остается фиксированной, а другая сходится к решению (см. рис. 2.9).

Сейчас снова решим уравнение $x\sin(x) - 1 = 0$ с помощью метода ложного положения и отметим, что он сходится быстрее, чем метод деления пополам. Отметим также, что $\{b_n - a_n\}_{n=0}^{\infty}$ не приводит к нулю.

Пример 2.8. Воспользуемся методом ложного положения для нахождения корня уравнения $x\sin(x) - 1 = 0$, который находится на интервале $[0, 2]$ (функция $\sin(x)$ вычисляется в радианах).

Начиная с $a_0 = 0$ и $b_0 = 2$, получим, что $f(0) = -1{,}00000000$ и $f(2) = 0{,}81859485$, поэтому корень лежит на интервале $[0; 2]$. Воспользуемся формулой (22) и получим

$$c_0 = 2 - \frac{0{,}81859485(2 - 0)}{0{,}81859485 - (-1)} = 1{,}09975017 \quad \text{и} \quad f(c_0) = -0{,}02001921.$$

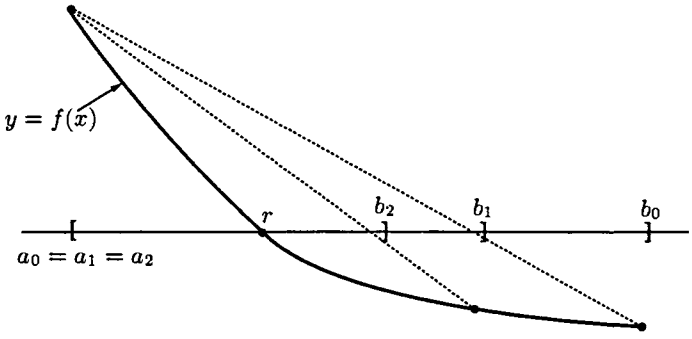


Рис. 2.9. Стационарные крайние точки для метода ложного положения

Функция изменяет знак на интервале $[c_0; b_0] = [1{,}09975017; 2]$, поэтому сдвигаемся влево и присваиваем $a_1 = c_0$ и $b_1 = b_0$. Формула (22) приводит к приближению

$$c_1 = 2 - \frac{0{,}81859485(2 - 1{,}09975017)}{0{,}81859485 - (-0{,}02001921)} = 1{,}12124074$$

и

$$f(c_1) = 0{,}00983461.$$

Следующее изменение знака функции $f(x)$ происходит на интервале $[a_1; c_1] = [1{,}09975017; 1{,}12124074]$, и следующим решением будет сдвиг вправо и присвоение $a_2 = a_1$ и $b_2 = c_1$. Результаты вычислений приведены в табл. 2.2. ■

Таблица 2.2. Решение уравнения $x\sin(x) - 1 = 0$ методом ложного положения

k	Левая крайняя точка, a_k	Средняя точка, c_k	Правая крайняя точка, b_k	Значение функции, $f(c_k)$
0	0,00000000	1,09975017	2,00000000	−0,02001921
1	1,09975017	1,12124074	2,00000000	0,00983461
2	1,09975017	1,11416120	1,12124074	0,00000563
3	1,09975017	1,11415714	1,11416120	0,00000000

Критерий окончания итерационного процесса, который используется в методе деления пополам, не пригоден для метода ложного положения и может привести к бесконечному циклу. Близость последовательных итераций и величина $|f(c_n)|$ используются в критерии окончания итерации в программе 2.3. В разделе 2.3 обсуждается причина такого выбора.

Программа 2.2 (метод деления пополам). Приближенное нахождение корня уравнения $f(x) = 0$ на интервале $[a; b]$. Применять только для случая, когда функция $f(x)$ непрерывна и $f(a)$ и $f(b)$ имеют различные знаки.

```
function [c,err,yc]=bisect(f,a,b,delta)
%Ввод  - f - функция вводится как строка 'f'
%      - a и b - левая и правая крайние точки
%      - дельта - допустимое отклонение
%Выход - c - нуль
%      - yc=f(c)
%      - err - ошибка вычисления c
ya=feval(f,a);
yb=feval(f,b);
if ya*yb>0,break,end
max1=1+round((log(b-a)-log(delta))/log(2));
for k=1:max1
    c=(a+b)/2;
    yc=feval(f,c);
    if yc==0
       a=c;
       b=c;
elseif yb*yc>0
      b=c;
      yb=yc;
   else
      a=c;
      ya=yc;
   end
   if b-a < delta, break,end
end
c=(a+b)/2;
err=abs(b-a);
yc=feval(f,c);
```

Программа 2.3 (метод ложного положения или метод regula falsi). Приближенное нахождение корня уравнения $f(x) = 0$ на интервале $[a; b]$. Только для случая, когда функция $f(x)$ непрерывна и $f(a)$ и $f(b)$ имеют различные знаки.

```
function [c,err,yc]=regula(f,a,b,delta,epsilon,max1)
%Ввод  - f - функция вводится как строка 'f'
%      - a и b - левая и правая крайние точки
%      - delta - допустимое отклонение для нуля
```

```
%       - epsilon - допустимое отклонение для значения f в нуле
%       - max1 - максимальное число итераций
%Вывод - c - нуль
%       - yc=f(c)
%       - err - ошибка вычисления для c
ya=feval(f,a);
yb=feval(f,b);
if ya*yb>0
   disp('Замечание: f(a)*f(b)>0'),
   break,
end
for k=1:max1
   dx=yb*(b-a)/(yb-ya);
   c=b-dx;
   ac=c-a;
   yc=feval(f,c);
   if yc==0,break;
   elseif yb*yc>0
      b=c;
      yb=yc;
   else
      a=c;
      ya=yc;
   end
   dx=min(abs(dx),ac);
   if abs(dx)<delta,break,end
   if abs(yc)<epsilon,break,end
end
c;
err=abs(b-a)/2;
yc=feval(f,c);
```

Упражнения к разделу 2.2

В упр. 1 и 2 найдите приближение для процентной ставки I, которая даст общую сумму ренты, равную A, если 240 месяцев делать ежемесячный вклад, равный P. Используйте два начальных значения для I и вычислите три последующих приближения с помощью метода деления пополам.

1. $P = \$275$; $A = \$250\,000$; $I_0 = 0{,}11$; $I_1 = 0{,}12$

2. $P = \$325$; $A = \$400\,000$; $I_0 = 0{,}13$; $I_1 = 0{,}14$

3. Для каждой из функций найдите такой интервал $[a; b]$, что $f(a)$ и $f(b)$ имеют различные знаки.
   (a) $f(x) = e^x - 2 - x$
   (b) $f(x) = \cos(x) + 1 - x$
   (c) $f(x) = \ln(x) - 5 + x$
   (d) $f(x) = x^2 - 10x + 23$

В упр. 4-7 начните с интервала $[a_0; b_0]$ и используйте метод ложного положения для вычисления c_0, c_1, c_2 и c_3.

4. $e^x - 2 - x = 0, [a_0; b_0] = [-2{,}4; -1{,}6]$

5. $\cos(x) + 1 - x = 0, [a_0; b_0] = [0{,}8; 1{,}6]$

6. $\ln(x) - 5 + x = 0, [a_0; b_0] = [3{,}2; 4{,}0]$

7. $x^2 - 10x + 23 = 0, [a_0; b_0] = [6{,}0; 6{,}8]$

8. Обозначим интервалы, которые появятся в методе деления пополам, как $[a_0; b_0]$, $[a_1; b_1]$, ..., $[a_n; b_n]$.
   (a) Покажите, что $a_0 \le a_1 \le \cdots \le a_n \le \cdots$ и $\cdots \le b_n \le \cdots \le b_1 \le b_0$.
   (b) Покажите, что $b_n - a_n = (b_0 - a_0)/2^n$.
   (c) Пусть $c_n = (a_n + b_n)/2$ — средняя точка каждого интервала. Покажите, что
   $$\lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n = \lim_{n\to\infty} b_n.$$
   Указание. Найдите в учебниках теоремы о сходимости монотонных последовательностей.

9. Что произойдет, если метод деления пополам использовать для функции $f(x) = 1/(x - 2)$ и
   (a) интервала $[3; 7]$; (b) интервала $[1; 7]$?

10. Что произойдет, если метод деления пополам использовать для функции $f(x) = \tan(x)$ и
    (a) интервала $[3; 4]$; (b) интервала $[1; 3]$?

11. Предположим, что метод деления пополам используется для нахождения нуля функции $f(x)$ на интервале $[2; 7]$. Сколько раз нужно делить этот интервал, чтобы иметь гарантию, что точность приближения c_N будет равна 5×10^{-9}?

12. Покажите, что формула (22) для метода ложного положения является алгебраическим эквивалентом выражения
    $$c_n = \frac{a_n f(b_n) - b_n f(a_n)}{f(b_n) - f(a_n)}.$$

13. Докажите формулу (15) для определения числа итераций, которые необходимо выполнить в методе деления пополам. *Указание.* Воспользуйтесь неравенством $|b - a|/2^{n+1} < \delta$ и постройте алгоритм.

14. Полином $f(x) = (x-1)^3(x-2)(x-3)$ имеет три нуля: $x = 1$ кратности 3 и $x = 2$ и $x = 3$ каждый кратности 1. Если a_0 и b_0 — любые два действительных числа, таких, что $a_0 < 1$ и $b_0 > 3$, то $f(a_0)f(b_0) < 0$. Значит, на интервале $[a_0; b_0]$ метод деления пополам сходится к одному из трех нулей. Если $a_0 < 1$ и $b_0 > 3$ выбрать так, что $c_n = \frac{a_n+b_n}{2}$ не равно 1, 2 или 3 для любого $n \geq 1$, то метод деления пополам никогда не сойдется к одному из нулей. Объясните, почему?

15. Если полином $f(x)$ имеет нечетное число нулей на интервале $[a_0; b_0]$ и каждый нуль имеет нечетную кратность, то $f(a_0)f(b_0) < 0$ и метод деления пополам сойдется к одному из нулей. Если $a_0 < 1$ и $b_0 > 3$ выбрать так, что $c_n = \frac{a_n+b_n}{2}$ не равно какому-либо из нулей функции $f(x)$ для любого $n \geq 1$, то метод деления пополам никогда не сойдется к одному из нулей. Объясните, почему?

Алгоритмы и программы

1. Найдите приближение (с точностью до 10 десятичных знаков) для процентной ставки I, которая даст общую сумму годовой ренты, равную \$500 000, если 240 месяцев постоянно вносить по \$300.

2. Рассмотрите шар радиуса $r = 15$ см, который сделан из вида белого дуба, имеющего плотность $\rho = 0{,}710$. Какая часть шара (с точностью до 8 десятичных знаков) погружена в воду, если он находится в воде?

3. Преобразуйте программы 2.2 и 2.3, чтобы получить на выходе матрицу, подобную матрице, представленной в табл. 2.1 и 2.2 (т. е. первая строка матрицы должна иметь вид $\begin{bmatrix} 0 & a_0 & c_0 & b_0 & f(c_0) \end{bmatrix}$).

4. Используйте свою программу для решения упр. 3, чтобы найти приближенно три наименьших положительных корня уравнения $x = \tan(x)$ (с точностью до 8 десятичных знаков).

5. Шар радиусом, равным единице, разрезан плоскостью на два сегмента. Один сегмент содержит две части объема другого. Определите расстояние x плоскости от центра шара (с точностью до 10 десятичных знаков).

2.3. Начальное приближение и критерий сходимости

Метод интервалов зависит от того, можно ли найти интервал $[a; b]$ таким образом, чтобы $f(a)$ и $f(b)$ имели различные знаки. Как только будет найден такой

интервал, величина которого не имеет значения, итерация будет длиться до тех пор, пока не будет найден корень. Поэтому данные методы называют ***глобально сходящимися***. Однако, если $f(x) = 0$ имеет несколько корней на интервале $[a; b]$, следует использовать различные начальные интервалы для нахождения каждого корня. Но определять эти малые интервалы, на которых $f(x)$ меняет знак, нелегко.

В разделе 2.4 излагается метод Ньютона–Рафсона и метод секущих для решения уравнения $f(x) = 0$. Оба метода требуют, чтобы сначала было найдено такое приближение к корню, которое гарантировало бы сходимость. Поэтому данные методы называются ***локально сходящимися***. Обычно они сходятся более быстро, чем глобально сходящиеся методы. Некоторые смешанные алгоритмы начинают с глобально сходящихся методов и переходят к локально сходящемуся методу, когда итерация близко подходит к корню.

Если вычисление корней является частью большего проекта, то удобно сначала построить график функции. Увидев график функции $y = f(x)$, можно принять решение на основании того, как выглядит график (вогнутая функция, наклон, колебательное поведение, локальный экстремум, точки перегиба и т. д.). Но более важно следующее: когда координаты точек на графике приблизительно известны, тогда их можно проанализировать и приближенно определить расположение корней. Эти приближения затем можно использовать в качестве начальных значений в алгоритмах нахождения корней уравнения.

Поступать следует осторожно. В программном обеспечении компьютера графики используются в несколько искаженном виде. Предположим, что компьютер использует график функции $y = f(x)$ на $[a; b]$. Обычно интервал делят на $N + 1$ равных отрезков точками, $a = x_0 < x_1 < \cdots < x_N = b$, и вычисляют значения функции $y_k = f(x_k)$. Затем каждый отрезок прямой или "подогнанной кривой" вычерчивают между последовательными точками $(x_{k-1}; y_{k-1})$ и $(x_k; y_k)$ для $k = 1, 2, \ldots, N$. Точек должно быть настолько много, чтобы не пропустить корень в той части кривой, где функция изменяется медленно. Если $f(x)$ непрерывна и две смежные точки, $(x_{k-1}; y_{k-1})$ и $(x_k; y_k)$, лежат по различные стороны оси x, то согласно теореме о промежуточном значении по крайней мере один корень лежит на интервале $[x_{k-1}; x_k]$. Но если существует корень или даже несколько близко расположенных корней на интервале $[x_{k-1}; x_k]$ и две соседние точки $(x_{k-1}; y_{k-1})$ и $(x_k; y_k)$ лежат на одной и той же стороне оси x, то построенный компьютером график не отразит ситуацию, в которой применима теорема о промежуточном значении. График не будет соответствовать действительному графику функции f. Это не редкость, когда функция имеет "сливающиеся" корни, т.е. корни, в которых график соприкасается с осью x, но не пересекает ее, или корни, "сливающиеся" с вертикальной асимптотой. Такие ситуации необходимо рассматривать, когда применяется любой численный алгоритм для нахождения корня.

Таким образом, вблизи двух соседних корней или около сливающихся корней у построенной компьютером кривой между точкой $(x_{k-1}; y_{k-1})$ и $(x_k; y_k)$ может

Таблица 2.3. Приближенное нахождение места расположения корней

x_k	Значения функции		Разность y-ов		Изменение знака у $f(x)$ или $f'(x)$
	y_{k-1}	y_k	$y_k - y_{k-1}$	$y_{k+1} - y_k$	
−1,2	−3,125	−0,968	2,157	1,329	
−0,9	−0,968	0,361	1,329	0,663	f изменяет знак на $[x_{k-1}, x_k]$
−0,6	0,361	1,024	0,663	0,159	
−0,3	1,024	1,183	0,159	−0,183	f' изменяет знак около x_k
0,0	1,183	1,000	−0,183	−0,363	
0,3	1,000	0,637	−0,363	−0,381	
0,6	0,637	0,256	−0,381	−0,237	
0,9	0,256	0,019	−0,237	0,069	f' изменяет знак около x_k
1,2	0,019	0,088	0,069	0,537	

недоставать пересечения или соприкосновения с осью x. Если $|f(x_k)|$ меньше наперед заданного значения ϵ (т. е. $f(x_k) \approx 0$), то x_k является предварительным приближением корня. Однако график может быть близок к нулю на широком диапазоне значений около точки x_k. Тогда точка x_k, возможно не близка к истинному корню. Поэтому добавляется требование, чтобы тангенс угла наклона графика изменял знак вблизи точки $(x_k; y_k)$, т.е. $m_{k-1} = \frac{y_k - y_{k-1}}{x_k - x_{k-1}}$ и $m_k = \frac{y_{k+1} - y_k}{x_{k+1} - x_k}$ должны иметь различные знаки. Так как $x_k - x_{k-1} > 0$ и $x_{k+1} - x_k > 0$, нет необходимости использовать отношение разностей, и достаточно для проверки увидеть, что разности $y_k - y_{k-1}$ и $y_{k+1} - y_k$ меняют знак. В этом случае x_k является приближением корня. К несчастью, нет гарантии, что такое начальное значение приведет к сходящейся последовательности. Если график $y = f(x)$ имеет локальный минимум (или максимум), чрезвычайно близкий к нулю, то, возможно, что x_k будет рассматриваться как приближенное значение корня, когда $f(x_k) \approx 0$, несмотря на то, что x_k, возможно, и не близка к корню.

Пример 2.9. Найдем приближенное расположение корней уравнения $x^3 - x^2 - x + 1 = 0$ на интервале $[-1.2; 1.2]$. Для определенности выберем $N = 8$ и посмотрим табл. 2.3.

Рассмотрим три абсциссы: $-1{,}05$; $-0{,}3$ и $0{,}9$. Так как $f(x)$ меняет знак на интервале $[-1{,}2; -0{,}9]$, значение $-1{,}05$ является приближенным корнем. Действительно, $f(-1{,}05) = -0{,}210$.

Несмотря на то что тангенс угла наклона меняет знак около точки $-0{,}3$, $f(-0{,}3) = 1{,}183$; следовательно, $-0{,}3$ не близко к корню. И наконец, тангенс угла наклона меняет знак около $0{,}9$ и $f(0{,}9) = 0{,}019$, поэтому $0{,}9$ является приближенным корнем (рис. 2.10). ■

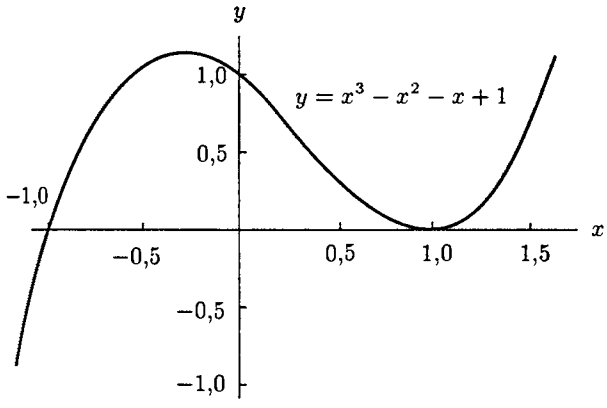


Рис. 2.10. График кубического полинома $y = x^3 - x^2 - x + 1$

Проверка сходимости

График можно использовать для того, чтобы увидеть приблизительно место нахождения корня, но алгоритм должен использовать для вычисления значения p_n, т. е. приемлемое компьютерное решение. Часто применяют итерацию для получения последовательности $\{p_k\}$, сходящейся к корню p, и критерий (или стратегия) останова должен быть составлен заранее так, чтобы компьютер прекратил вычисления, когда достигнет необходимого приближения. Поскольку нужно решить уравнение $f(x) = 0$, окончательное значение p_n должно удовлетворять неравенству $|f(p_n)| < \epsilon$.

Пользователь может задать допустимое значение ϵ для величины $|f(p_n)|$, и тогда итеративный процесс будет порождать точки $P_k = (p_k, f(p_k))$ до тех пор, пока последняя точка P_n лежит в горизонтальной полосе, ограниченной прямыми $y = +\epsilon$ и $y = -\epsilon$, как показано на рис. 2.11(a). Этот критерий пригоден, если пытаться решить уравнение $h(x) = L$, применив алгоритм для нахождения корня к функции $f(x) = h(x) - L$.

Другой критерий останова использует абсциссу, и можно попытаться остановиться, если последовательность $\{p_k\}$ сходящаяся. Если нарисовать вертикальные линии $x = p + \delta$ и $x = p - \delta$ с каждой стороны от $x = p$, то остановить итерацию можно, когда точка P_n лежит между этими двумя вертикальными линиями, как показано на рис. 2.11(b).

Последний критерий часто удобен, но его сложно выполнять, потому что он включает неизвестное значение p. Используем эту идею и в дальнейшем будем останавливать вычисления, когда последовательные итерации p_{n-1} и p_n достаточно близки или если у них совпадают M значащих цифр.

Иногда достаточно выполнения соотношения $p_n \approx p_{n-1}$ или $f(p_n) \approx 0$. Необходимо правильно логически объяснить, что это обозначает. Если потребовать, чтобы было $|p_n - p| < \delta$ и $|f(p_n)| < \epsilon$, точка P_n будет определена в прямоуголь-

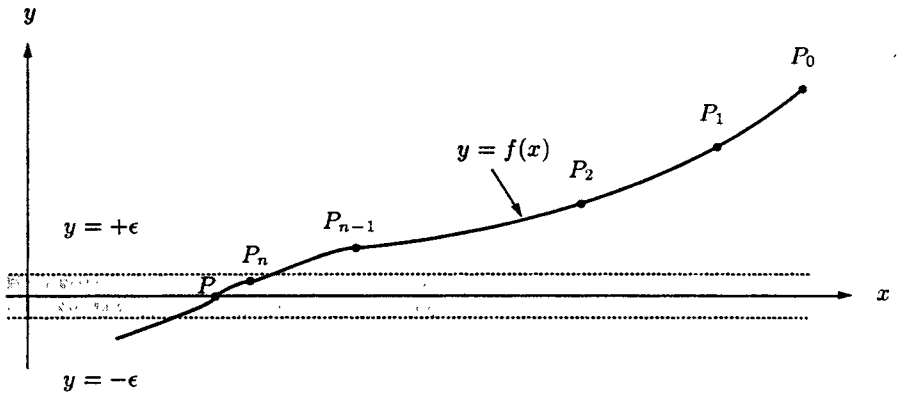


Рис. 2.11. (a) Горизонтальная полоса сходимости для нахождения решения уравнения $f(x) = 0$

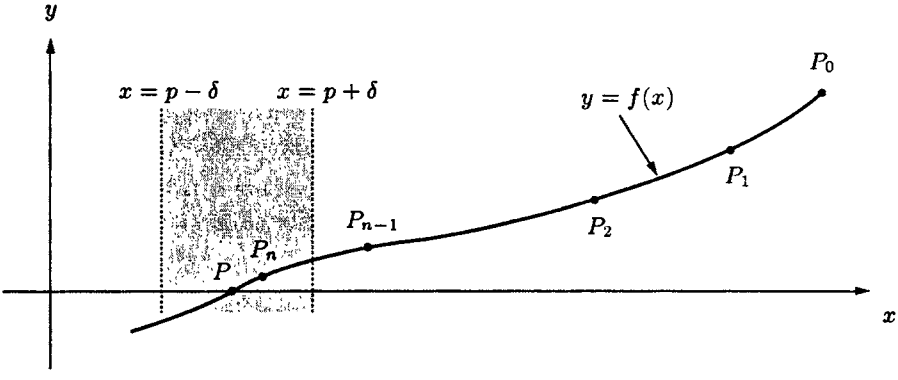


Рис. 2.11. (b) Вертикальная полоса сходимости для нахождения решения уравнения $f(x) = 0$

ной области около решения $(p, 0)$, как показано на рис. 2.12(a). Если потребовать, чтобы было $|p_n - p| < \delta$ **или** $|f(p_n)| < \epsilon$, точка P_n может быть определена где-нибудь в области, образованной объединением горизонтальной и вертикальной полос, как показано на рис. 2.12(b). Область допустимых отклонений δ и ϵ является крестообразной. Если допустимые отклонения выбраны слишком малыми, то итерация может продолжаться вечно. Их следует выбирать приблизительно в 100 раз больше, чем 10^{-M}, где M — число десятичных знаков компьютера для чисел с плавающей точкой. Близость к абсциссе проверяется одним из критериев:

$$|p_n - p_{n-1}| < \delta \quad \text{(оценка абсолютной ошибки)}$$

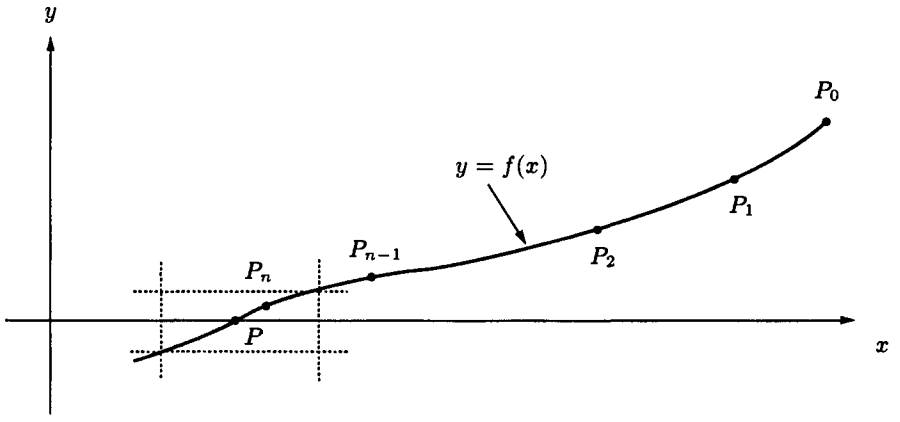


Рис. 2.12. (a) Прямоугольная область, определенная неравенствами $|x - p| < \delta$ И $|y| < \epsilon$

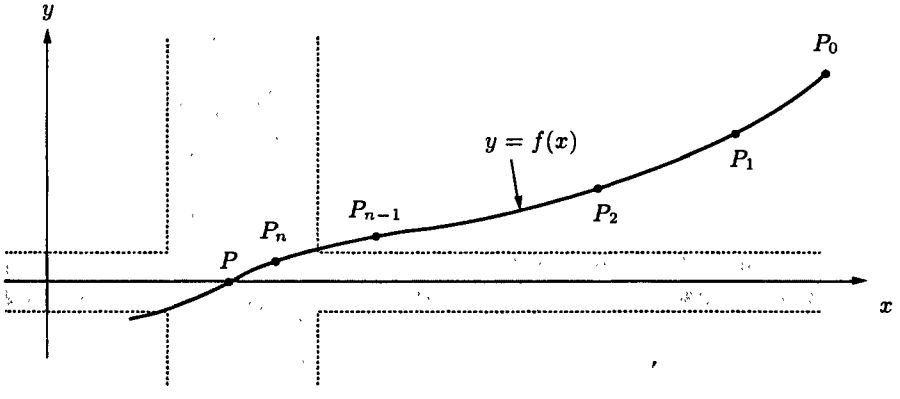


Рис. 2.12. (b) Безграничная область, определенная неравенством $|x - p| < \delta$ ИЛИ $|y| < \epsilon$

или

$$\frac{2|p_n - p_{n-1}|}{|p_n| + |p_{n-1}|} < \delta \quad \text{(вычисление относительной ошибки).}$$

Близость ординаты обычно проверяется неравенством $|f(p_n)| < \epsilon$.

Неудобные функции

Компьютерное решение уравнения $f(x) = 0$ почти всегда имеет ошибку, обусловленную округлением и/или неустойчивостью в вычислениях. Если график $y = f(x)$ крутой около корня $(p; 0)$, то задача нахождения корня хорошо обусловлена (т. е. можно легко получить решение с несколькими значащими цифрами).

Если график $y = f(x)$ пологий около $(p; 0)$, то задача нахождения корня плохо обусловлена (т. е. вычисленный корень может иметь лишь несколько значащих цифр). Это случается, когда $f(x)$ имеет кратный корень в p. Более детальное обсуждение будет приведено в следующем разделе.

Программа 2.4 (приближенное нахождение места расположения корней). Найти приблизительное место расположения корней уравнения $f(x) = 0$ на интервале $[a; b]$, используя точки $(x_k; f(x_k))$ и следующий критерий:

(i) $(y_{k-1})(y_k) < 0$, или

(ii) $|y_k| < \epsilon$ и $(y_k - y_{k-1})(y_{k+1} - y_k) < 0$.

Иначе говоря, либо $f(x_{k-1})$ и $f(x_k)$ имеют различные знаки, либо $|f(x_k)|$ мал и тангенс угла наклона кривой $y = f(x)$ меняет знак около $(x_k, f(x_k))$.

```
function R = approot (X,epsilon)
%  Вход - f - функция записанная как M-файл под именем f.m
%       - X - вектор абсцисс
%       - epsilon - допустимое отклонение
% Выход - R - вектор приближений корней
Y=f(X);
yrange = max(Y)-min(Y);
epsilon2 = yrange*epsilon;
n=length(X);
m=0;
X(n+1)=X(n);
Y(n+1)=Y(n);
for k=2:n,
   if Y(k-1)*Y(k)<=0,
      m=m+1;
      R(m)=(X(k-1)+X(k))/2;
   end
   s=(Y(k)-Y(k-1))*(Y(k+1)-Y(k));
   if (abs(Y(k)) < epsilon2) & (s<=0),
      m=m+1;
      R(m)=X(k);
   end
end
```

Пример 2.10. Используем программу approot для того, чтобы найти место приближенного расположения корней уравнения $f(x) = \sin(\cos(x^3))$ на интервале $[-2; 2]$. Сначала запишем f в виде M-файла под именем f.m. Затем используем

результаты как начальное приближение для алгоритма нахождения корня и построим вектор X так, чтобы приближение имело точность 4 десятичных знака.

```
>>X=-2:.001:2;
>>approot (X,0.00001)
ans=
-1.9875 -1.6765 -1.1625 1.1625 1.6765 1.9875
```

Сравнивая результаты с графиком функции f, видим, что получено хорошее приближение для одного из алгоритмов для нахождения корня. ■

Упражнения к разделу 2.3

В упр. 1–6 используйте компьютер или графическое вычисление, чтобы графически определить приближенное место расположения корней уравнения $f(x) = 0$ на заданном интервале. В каждом случае определите интервал $[a; b]$ так, чтобы найти корни, используя программы 2.2 и 2.3 (т. е. $f(a)f(b) < 0$).

1. $f(x) = x^2 - e^x$ для $-2 \le x \le 2$
2. $f(x) = x - \cos(x)$ для $-2 \le x \le 2$
3. $f(x) = \sin(x) - 2\cos(x)$ для $-2 \le x \le 2$
4. $f(x) = \cos(x) + (1 + x^2)^{-1}$ для $-2 \le x \le 2$
5. $f(x) = (x - 2)^2 - \ln(x)$ для $0{,}5 \le x \le 4{,}5$
6. $f(x) = 2x - \tan(x)$ для $-1{,}4 \le x \le 1{,}4$

Алгоритмы и программы

В задачах 1 и 2 используйте компьютер или графическое вычисление и программу 2.4, чтобы приближенно вычислить действительные корни с 4 десятичными знаками каждой заданой функции на заданом интервале. Затем воспользуйтесь программой 2.2 или 2.3, чтобы найти каждый корень с 12 десятичными знаками.

1. $f(x) = 1\,000\,000x^3 - 111\,000x^2 + 1110x - 1$ для $-2 \le x \le 2$
2. $f(x) = 5x^{10} - 38x^9 + 21x^8 - 5\pi x^6 - 3\pi x^5 - 5x^2 + 8x - 3$ для $-15 \le x \le 15$.
3. Программа построения графика $y = f(x)$ на интервале $[a; b]$ использует точки $(x_0; y_0), (x_1; y_1), \ldots$ и $(x_N; y_N)$ и обычно определяет масштаб вертикально по

высоте графика, поэтому должна быть записана процедура для определения минимального и максимального значений функции f на интервале.

(a) Постройте алгоритм, который найдет значения $Y_{\max} = \max_k\{y_k\}$ и $Y_{\min} = \min_k\{y_k\}$.

(b) Напишите MATLAB-программу для нахождения приближенного расположения максимального (минимального) значения функции $f(x)$ и ее величину на интервале $[a; b]$.

(c) Используйте свою программу из п. (b), чтобы найти приближенное место расположения и значение максимума (минимума) функций из задач 1 и 2. Сравните свои приближения с действительными значениями.

2.4. Метод Ньютона–Рафсона и метод секущих

Метод касательных для нахождения корней

Если $f(x)$, $f'(x)$ и $f''(x)$ непрерывны в окрестности корня p, эту дополнительную информацию о свойствах функции $f(x)$ можно использовать для построения алгоритмов, которые порождают последовательности $\{p_k\}$, сходящиеся к p быстрее, чем при методе деления пополам или методе ложного положения. Метод Ньютона–Рафсона (или просто Ньютона) является одним из наиболее полезных и самым известным алгоритмом, в котором используется непрерывность $f'(x)$ и $f''(x)$. Введем его графически и затем дадим более строгую трактовку на основании полиномов Тейлора.

Предположим, что начальное приближение p_0 близко к корню p. Тогда график $y = f(x)$ пересекает ось x в точке $(p; 0)$ и точка $(p_0; f(p_0))$ лежит на кривой около точки $(p; 0)$ (рис. 2.13). Определим точку p_1, как точку пересечения оси x и

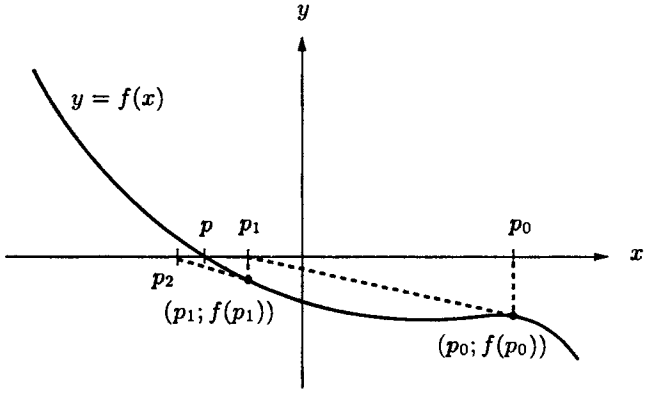


Рис. 2.13. Геометрическое построение p_1 и p_2 для метода Ньютона–Рафсона

касательной к кривой в точке $(p_0; f(p_0))$. На рис. 2.13 показано, что в этом случае p_1 ближе к p, чем p_0. Соотношение, связывающее p_1 и p_0, можно найти, если записать два варианта для тангенса угла наклона касательной L:

$$m = \frac{0 - f(p_0)}{p_1 - p_0}, \tag{1}$$

где m равно тангенсу угла наклона линии, проходящей через точки $(p_1; 0)$ и $(p_0; f(p_0))$, и

$$m = f'(p_0), \tag{2}$$

где m равно тангенсу угла наклона к кривой в точке $(p_0; f(p_0))$. Приравняем значения тангенса угла наклона m в выражениях (1) и (2) и результатом решения относительно p_1 будет

$$p_1 = p_0 - \frac{f(p_0)}{f'(p_0)}. \tag{3}$$

Приведенный выше процесс можно повторять, чтобы получить последовательностью $\{p_k\}$, сходящуюся к p. А сейчас изложим это более строго.

Теорема 2.5 (теорема Ньютона–Рафсона). Предположим, что $f \in C^2[a; b]$ и существует такое число $p \in [a; b]$, что $f(p) = 0$. Если $f'(p) \neq 0$, то существует такое $\delta > 0$, что последовательность $\{p_k\}_{k=0}^{\infty}$, определенная итерацией

$$p_k = g(p_{k-1}) = p_{k-1} - \frac{f(p_{k-1})}{f'(p_{k-1})} \quad \text{для } k = 1, 2, \ldots \tag{4}$$

будет сходиться к p для любого начального приближения $p_0 \in [p - \delta, p + \delta]$.

Замечание. Функция $g(x)$, определенная формулой

$$g(x) = x - \frac{f(x)}{f'(x)}, \tag{5}$$

называется ***интерполяционной функцией Ньютона–Рафсона***. Так как $f(p) = 0$, легко видеть, что $g(p) = p$. Таким образом, итерация Ньютона–Рафсона для нахождения корней уравнения $f(x) = 0$ осуществляется путем нахождения неподвижной точки функции $g(x)$.

Доказательство. Геометрическое построение точки p_1, показанное на рис. 2.13, не помогает понять, почему необходимо, чтобы p_0 была близко расположена к p, или почему важна непрерывность $f''(x)$. Начнем анализ с полинома Тейлора степени $n = 1$ и его остаточного члена:

$$f(x) = f(p_0) + f'(p_0)(x - p_0) + \frac{f''(c)(x - p_0)^2}{2!}, \tag{6}$$

где c лежит где-то между p_0 и x. Подставляя $x = p$ в равенство (6) и учитывая то, что $f(p) = 0$, получим

$$0 = f(p_0) + f'(p_0)(p - p_0) + \frac{f''(c)(p - p_0)^2}{2!}. \tag{7}$$

Если p_0 достаточно близко к p, то последний член правой части выражения (7) будет существенно меньше суммы двух первых членов. Следовательно, им можно пренебречь и использовать приближение

$$0 \approx f(p_0) + f'(p_0)(p - p_0). \tag{8}$$

Решив уравнение (8) относительно p, получаем $p \approx p_0 - f(p_0)/f'(p_0)$. Это приближенное равенство используем, чтобы определить p_1 — следующее приближение к корню:

$$p_1 = p_0 - \frac{f(p_0)}{f'(p_0)}. \tag{9}$$

Общее правило (4) будет установленным, когда место p_0 в уравнении (9) займет p_{k-1}. Это все, что необходимо понимать для большинства случаев. Однако для полного понимания того, что происходит, необходимо рассмотреть итерации неподвижной точки функции и применить в нашей ситуации теорему 2.2. Ключом является анализ $g'(x)$:

$$g'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}.$$

По предположению $f(p) = 0$, поэтому $g'(p) = 0$. Так как $g'(p) = 0$ и $g(x)$ непрерывна, можно найти такое $\delta > 0$, что предположение $|g'(x)| < 1$ теоремы 2.2 выполняется на интервале $(p - \delta, p + \delta)$. Тогда достаточным условием для того, чтобы p_0 была начальной точкой сходящейся последовательности $\{p_k\}_{k=0}^{\infty}$, которая сходится к корню $f(x) = 0$, является выбор $p_0 \in (p - \delta, p + \delta)$, где δ такое, что

$$\frac{|f(x)f''(x)|}{|f'(x)|^2} < 1 \quad \text{для всех } x \in (p - \delta, p + \delta). \tag{10}$$

•

Следствие 2.2 (итерация Ньютона для нахождения квадратных корней). Предположим, что $A > 0$ — действительное число, и пусть $p_0 > 0$ — начальное приближение к $\sqrt{A}$. Найдем последовательность $\{p_k\}_{k=0}^{\infty}$, используя рекуррентное правило

$$p_k = \frac{p_{k-1} + \dfrac{A}{p_{k-1}}}{2} \quad \text{для } k = 1, 2, \ldots. \tag{11}$$

Тогда последовательность $\{p_k\}_{k=0}^{\infty}$ сходится к $\sqrt{A}$, т. е. $\lim_{k\to\infty} p_k = \sqrt{A}$.

Схема доказательства. Начнем с функции $f(x) = x^2 - A$ и заметим, что корнями уравнения $x^2 - A = 0$ являются $\pm\sqrt{A}$. Используем функцию $f(x)$ и ее производную $f'(x)$ в формуле (5) и запишем ниже итерационную формулу Ньютона–Рафсона:

$$g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^2 - A}{2x}. \tag{12}$$

После упрощения формулы получим

$$g(x) = \frac{x + \frac{A}{x}}{2}. \tag{13}$$

Затем $g(x)$ в (13) используем, чтобы определить рекуррентную итерацию в (4); в результате получим формулу (11). Можно доказать, что последовательность, генерируемая в (11), будет сходиться для любого начального значения $p_0 > 0$. Подробное доказательство оставим для упражнений. •

Важным моментом следствия 2.2 является то, что итерационная функция $g(x)$ включает только арифметические операции $+, -, \times$ и $/$. Если бы $g(x)$ включала вычисление квадратного корня, то мы вычисляли бы квадратные корни без построения рекуррентной последовательности, которая сходится к $\sqrt{A}$. Из этих соображений было выбрано выражение $f(x) = x^2 - A$, так как оно приводит только к арифметическим операциям.

Пример 2.11. Воспользуемся алгоритмом Ньютона нахождения квадратного корня для вычисления $\sqrt{5}$.

Начав с $p_0 = 2$ и использовав формулы (11), вычислим

$$p_1 = \frac{2 + 5/2}{2} = 2{,}25$$

$$p_2 = \frac{2{,}25 + 5/2{,}25}{2} = 2{,}236111111$$

$$p_3 = \frac{2{,}236111111 + 5/2{,}236111111}{2} = 2{,}236067978$$

$$p_4 = \frac{2{,}36067978 + 5/2{,}236067978}{2} = 2{,}236067978.$$

Дальнейшие итерации дают $p_k \approx 2{,}236067978$ для $k > 4$, и мы видим, что достигается точность сходимости 9 десятичных знаков. ■

А сейчас вернемся к знакомой задаче из элементарной физики и выясним, почему определение места расположения корня является важной задачей. Предположим, что был выпущен снаряд под углом b_0 с начальной скоростью v_0. Из элементарного курса известно, что сопротивлением воздуха можно пренебречь и

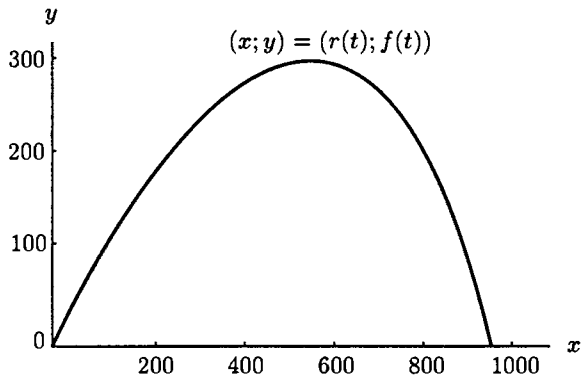


Рис. 2.14. Траектория движения снаряда с учетом сопротивления воздуха

что высота $y = y(t)$ и дальность полета $x = x(t)$, измеренные в футах (1 фут = = 30,5 см), удовлетворяют уравнениям

$$y = v_y t - 16t^2 \quad \text{и} \quad x = v_x t, \tag{14}$$

где горизонтальная и вертикальная составляющие начальной скорости соответственно равны $v_x = v_0 \cos(b_0)$ и $v_y = v_0 \sin(b_0)$. Математическая модель, выраженная уравнением (14), проста для вычислений, но дает слишком большую высоту и слишком большую величину пути. Если сделать дополнительное предположение о том, что сопротивление воздуха пропорционально скорости, то получим уравнения движения

$$y = f(t) = (Cv_y + 32C^2)\left(1 - e^{-t/C}\right) - 32Ct \tag{15}$$

и

$$x = r(t) = Cv_x\left(1 - e^{-t/C}\right), \tag{16}$$

где $C = m/k$ и k равно коэффициенту сопротивления воздуха, а m — масса снаряда. Большее значение C приводит к большей максимальной высоте и большей дальности полета снаряда. График пути полета снаряда, когда учитывается сопротивление воздуха, показан на рис. 2.14. Эта уточненная модель более реалистична, но требует использования алгоритма нахождения корня для решения уравнения $f(t) = 0$, чтобы найти время полета снаряда до падения. Элементарная модель (14) не требует сложной процедуры нахождения времени полета.

Пример 2.12. Снаряд выпущен под углом $b_0 = 45°$, $v_y = v_x = 160$ фут/с и $C = 10$. Найти время полета до падения и дальность полета.

Используя формулы (15) и (16), получим уравнения движения $y = f(t) = = 4800(1 - e^{-t/10}) - 320t$ и $x = r(t) = 1600(1 - e^{-t/10})$. Так как $f(8) = 83{,}220972$

Таблица 2.4. Определение времени полета (с момента, когда высота $f(t)$ равна нулю)

k	Время, p_k	$p_{k+1} - p_k$	Высота, $f(p_k)$
0	8,00000000	0,79773101	83,22097200
1	8,79773101	−0,05530160	−6,68369700
2	8,74242941	−0,00025475	−0,03050700
3	8,74217467	−0,00000001	−0,00000100
4	8,74217466	0,00000000	0,00000000

и $f(9) = -31{,}534367$, используем начальное приближение $p_0 = 8$. Производная $f(t)$ равна $f'(t) = 480e^{-t/10} - 320$; ее значение в точке $p_0 - f'(p_0) = f'(8) = -104{,}3220972$, тогда из формулы (4) получаем

$$p_1 = 8 - \frac{83{,}22097200}{-104{,}3220972} = 8{,}797731010.$$

Результаты вычислений приведены в табл. 2.4.

Величина p_4 имеет 9 десятичных знаков точности, и время до падения равно $t \approx 8{,}74217466$ с. Дальность полета можно вычислить, используя $r(t)$:

$$r(8{,}74217466) = 1600\left(1 - e^{-0{,}874217466}\right) = 932{,}4986302 \text{ футов.} \quad \blacksquare$$

Ошибка деления на нуль

Одной очевидной ловушкой в методе Ньютона–Рафсона является возможность деления на нуль в формуле (4), которая возникает, если $f'(p_{k-1}) = 0$. Программа 2.5 имеет процедуру проверки такой ситуации, но как в таком случае использовать последнее вычисленное приближение p_{k-1}? Вполне вероятно, что $f(p_{k-1})$ достаточно близко к нулю и p_{k-1} — приемлемое приближение к корню. Изучим сейчас эту ситуацию и откроем интересный факт, т. е. определим скорость сходимости итерации.

Определение 2.4 (порядок корня). Предположим, что функция $f(x)$ и ее производные $f'(x), \ldots, f^{(M)}(x)$ определены и непрерывны на интервале в окрестности точки $x = p$. Говорят, что $f(x) = 0$ имеет корень порядка M в точке $x = p$ тогда и только тогда, когда

$$(17) \qquad f(p) = 0, \quad f'(p) = 0, \quad \ldots, \quad f^{(M-1)}(p) = 0 \quad \text{и} \quad f^{(M)}(p) \neq 0.$$

Корень порядка $M = 1$ часто называют ***простым корнем***, а если $M > 1$, его называют ***кратным корнем***. Корень порядка $M = 2$ иногда называют ***двойным корнем*** и т. д. Следующий результат пояснит эти понятия. ▲

Лемма 2.1. Если уравнение $f(x) = 0$ имеет корень порядка M при $x = p$, то существует такая непрерывная функция $h(x)$, что $f(x)$ можно представить в виде произведения

$$f(x) = (x - p)^M h(x), \quad \text{где } h(p) \neq 0. \tag{18}$$

Пример 2.13. Функция $f(x) = x^3 - 3x + 2$ имеет простой корень в $p = -2$ и двойной — в $p = 1$. Это можно проверить, рассмотрев производные $f'(x) = 3x^2 - 3$ и $f''(x) = 6x$. При значении $p = -2$ получим $f(-2) = 0$ и $f'(-2) = 9$, так что $M = 1$ в определении 2.4, поэтому $p = -2$ — простой корень. Для значения $p = 1$ получаем, что $f(1) = 0, f'(1) = 0$ и $f''(1) = 6$, так что $M = 2$ в определении 2.4, поэтому $p = 1$ — двойной корень. Заметим также, что разложение на множители функции $f(x)$ имеет вид $f(x) = (x + 2)(x - 1)^2$. ■

Скорость сходимости

Рассмотрим следующие свойства сходимости. Если p — простой корень уравнения $f(x) = 0$, то метод Ньютона сходится быстро и количество десятичных знаков точности приблизительно удваивается с каждой итерацией. С другой стороны, если p является кратным корнем, то ошибка в каждом последующем приближении равна части предыдущей ошибки. Чтобы уточнить это, определим понятие ***порядка сходимости***. Он является мерой скорости сходимости последовательности.

Определение 2.5 (порядок сходимости). Предположим, что $\{p_n\}_{n=0}^{\infty}$ сходится к p и положим $E_n = p - p_n$ для $n \geq 0$. Если существуют такие две положительные константы $A \neq 0$ и $R > 0$, что

$$\lim_{n\to\infty} \frac{|p - p_{n+1}|}{|p - p_n|^R} = \lim_{n\to\infty} \frac{|E_{n+1}|}{|E_n|^R} = A, \tag{19}$$

то говорят, что последовательность сходится к p с порядком сходимости R. Число A называют постоянной асимптотической ошибкой. Случаи, когда $R = 1, 2$, рассматриваются особо.

(20) Если $R = 1$, то сходимость $\{p_n\}_{n=0}^{\infty}$ называется ***линейной***.

(21) Если $R = 2$, то сходимость $\{p_n\}_{n=0}^{\infty}$ называется ***квадратичной***. ▲

Если R большое, последовательность $\{p_n\}$ сходится быстро к p, т. е. из соотношения (19) следует, что для больших значений n справедливо приближенное равенство $|E_{n+1}| \approx A|E_n|^R$. Например, предположим, что $R = 2$ и $|E_n| \approx 10^{-2}$. Тогда можно ожидать, что $|E_{n+1}| \approx A \times 10^{-4}$.

Некоторые последовательности сходятся с порядком, не являющимся целым числом, и мы увидим, что порядок сходимости в методе секущих равен $R = (1 + \sqrt{5})/2 \approx 1{,}618033989$.

Таблица 2.5. Квадратичная сходимость к простому корню в методе Ньютона

k	p_k	$p_{k+1}-p_k$	$E_k = p - p_k$	$\frac{\lvert E_{k+1}\rvert}{\lvert E_k\rvert^2}$
0	−2,400000000	0,323809524	0,400000000	0,476190475
1	−2,076190476	0,072594465	0,076190476	0,619469086
2	−2,003596011	0,003587422	0,003596011	0,664202613
3	−2,000008589	0,000008589	0,000008589	
4	−2,000000000	0,000000000	0,000000000	

Пример 2.14 (квадратичная сходимость к простому корню). Начнем с $p_0 = -2{,}4$ и воспользуемся итерацией Ньютона–Рафсона, чтобы найти корень $p = -2$ полинома $f(x) = x^3 - 3x + 2$. Итерационная формула для вычисления $\{p_k\}$ имеет вид

$$p_k = g(p_{k-1}) = \frac{2p_{k-1}^3 - 2}{3p_{k-1}^2 - 3}. \tag{22}$$

Используем формулу (21) для проверки квадратичной сходимости. Полученные значения приведены в табл. 2.5. ■

Детальное рассмотрение скорости сходимости в примере 2.14 показывает, что ошибка в каждой последующей итерации пропорциональна квадрату ошибки предыдущей итерации, т. е.

$$|p - p_{k+1}| \approx A|p - p_k|^2,$$

где $A \approx 2/3$. Для проверки воспользуемся тем, что

$$|p - p_3| = 0{,}000008589 \quad \text{и} \quad |p - p_2|^2 = |0{,}003596011|^2 = 0{,}000012931.$$

Легко видеть, что

$$|p - p_3| = 0{,}000008589 \approx 0{,}000008621 = \frac{2}{3}|p - p_2|^2.$$

Пример 2.15 (линейная сходимость к двойному корню). Начнем с $p_0 = 1{,}2$ и воспользуемся итерацией Ньютона–Рафсона, чтобы найти двойной корень $p = 1$ полинома $f(x) = x^3 - 3x + 2$.

Для проверки линейной сходимости используем формулу (20). Полученные значения приведены в табл. 2.6. ■

Таблица 2.6. Линейная сходимость метода Ньютона к двойному корню

k	p_k	$p_{k+1} - p_k$	$E_k = p - p_k$	$\frac{\lvert E_{k+1}\rvert}{\lvert E_k\rvert}$
0	1,200000000	−0,096969697	−0,200000000	0,515151515
1	1,103030303	−0,050673883	−0,103030303	0,508165253
2	1,052356420	−0,025955609	−0,052356420	0,496751115
3	1,026400811	−0,013143081	−0,026400811	0,509753688
4	1,013257730	−0,006614311	−0,013257730	0,501097775
5	1,006643419	−0,003318055	−0,006643419	0,500550093
⋮	⋮	⋮	⋮	⋮

Отметим, что метод Ньютона–Рафсона медленно сходится к двойному корню. Значения $f(p_k)$ в примере 2.15 сходятся к нулю быстрее, чем значения $f'(p_k)$, так что отношение $f(p_k)/f'(p_k)$ в формуле (4) определено, когда $p_k \neq p$. Последовательность сходится линейно и ошибка убывает с каждой последующей итерацией с коэффициентом приближения $1/2$. В следующей теореме показано действие метода Ньютона на простой и двойной корни.

Теорема 2.6 (скорость сходимости итерации Ньютона–Рафсона). Предположим, что итерация Ньютона–Рафсона производит последовательность $\{p_n\}_{n=0}^{\infty}$, которая сходится к корню p функции $f(x)$. Если p — простой корень, то сходимость является квадратичной и

$$|E_{n+1}| \approx \frac{|f''(p)|}{2|f'(p)|}|E_n|^2 \quad \text{для достаточно больших } n. \tag{23}$$

Если p — кратный корень порядка M, то сходимость линейная и

$$|E_{n+1}| \approx \frac{M-1}{M}|E_n| \quad \text{для достаточно больших } n. \tag{24}$$

Ловушки

Ошибку деления на нуль легко предупредить, но существуют другие сложности, которые не так легко выявить. Предположим, что задана функция $f(x) = x^2 - 4x + 5$, тогда последовательность действительных чисел $\{p_k\}$, полученная с помощью формулы (4), будет блуждать назад и вперед, слева направо и не сойдется. Простой анализ ситуации покажет, что функция $f(x) > 0$ и не имеет действительных корней.

Иногда начальное приближение p_0 слишком далеко от требуемого корня и последовательность $\{p_k\}$ сходится к некоторому другому корню. Это обычно происходит, когда тангенс наклона $f'(p_0)$ мал и касательная к кривой $y = f(x)$ близка

к горизонтали. Например, если $f(x) = \cos(x)$ и ищем корень $p = \pi/2$, начиная с $p_0 = 3$, вычисления покажут, что $p_1 = -4{,}01525255, p_2 = -4{,}85265757, \ldots$ и $\{p_k\}$ будут сходиться к другому корню: $-3\pi/2 \approx -4{,}71238898$.

Предположим, что $f(x)$ — положительная и монотонно убывающая функция на неограниченном интервале $[a, \infty)$ и $p_0 > a$. Тогда последовательность $\{p_k\}$ должна расходиться к $+\infty$. Например, если $f(x) = xe^{-x}$ и $p_0 = 2{,}0$, то

$$p_1 = 4{,}0; \quad p_2 = 5{,}333333333; \quad \ldots; \quad p_{15} = 19{,}723549434; \quad \ldots,$$

и $\{p_k\}$ медленно расходится к $+\infty$ (рис. 2.15(a)). Эта особенная функция ставит другую неожиданную проблему. Значение $f(x)$ стремится к нулю так же быстро, как увеличивается x, например $f(p_{15}) = 0{,}0000000536$, и вероятно, p_{15} можно ошибочно принять за корень. Из этих соображений в программе 2.5 построен критерий останова, основанный на относительной ошибке $2|p_{k+1} - p_k|/(|p_k| + 10^{-6})$, и, когда $k = 15$, его значение равно 0,106817, так что допустимое отклонение $\delta = 10^{-6}$ защищает от ложного корня.

Другой феномен, ***циклический***, встречается, когда члены последовательности $\{p_k\}$ стремятся повторяться или почти повторяться. Например, если $f(x) = x^3 - x - 3$ и начальное приближение равно $p_0 = 0$, то последовательность будет такой:

$$p_1 = -3{,}000000; \quad p_2 = -1{,}961538; \quad p_3 = -1{,}147176; \quad p_4 = -0{,}006579;$$
$$p_5 = -3{,}000389; \quad p_6 = -1{,}961818; \quad p_7 = -1{,}147430; \quad \ldots$$

Здесь мы приходим к циклу, где $p_{k+4} \approx p_k$ для $k = 0, 1, \ldots$ (рис. 2.15(b)). Но если начальное значение p_0 достаточно близко к корню $p \approx 1{,}671699881$, то $\{p_k\}$

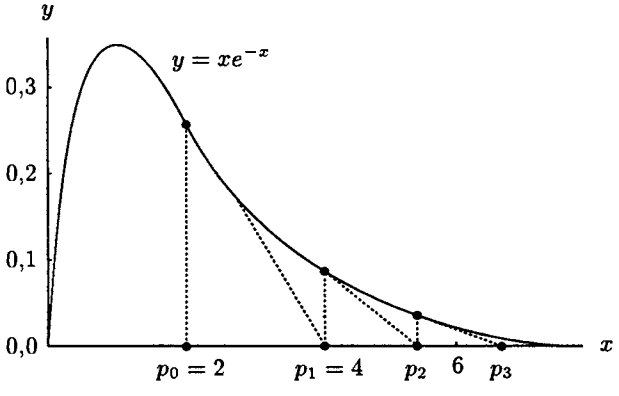


Рис. 2.15. (a) Итерация Ньютона–Рафсона для $f(x) = xe^{-x}$ может дать расходящуюся последовательность

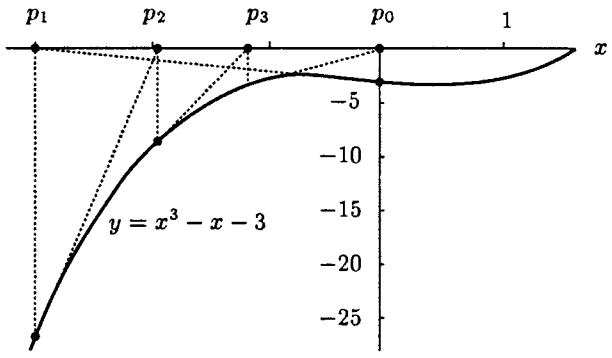


Рис. 2.15. (b) Итерация Ньютона–Рафсона для $f(x) = x^3 - x - 3$ может дать циклическую последовательность

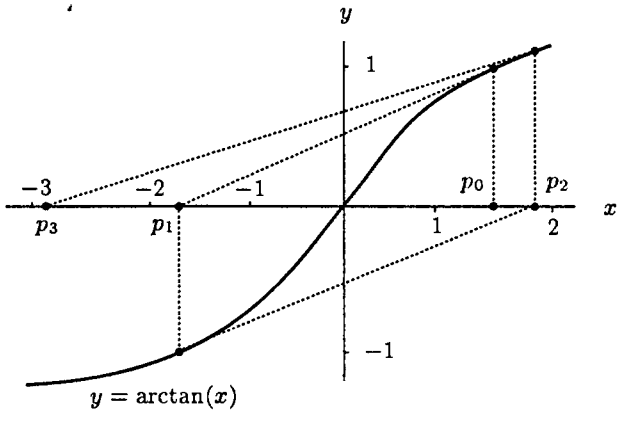


Рис. 2.15. (c) Итерация Ньютона–Рафсона для $f(x) = \arctan(x)$ может дать расходящуюся осциллирующую последовательность

сходится. Если $p_0 = 2$, последовательность сходится: $p_1 = 1{,}72727272$; $p_2 = 1{,}67369173$; $p_3 = 1{,}671702570$ и $p_4 = 1{,}671699881$.

Когда $|g'(x)| \geq 1$ на интервале, содержащем корень p, существует вероятность появления расходящихся колебаний. Например, пусть $f(x) = \arctan(x)$. Тогда функция итерации Ньютона–Рафсона равна $g(x) = x - (1+x^2)\arctan(x)$ и $g'(x) = -2x\arctan(x)$. Если выбрать начальное значение равным $p_0 = 1{,}45$, то

$$p_1 = -1{,}550263297; \qquad p_2 = 1{,}845931751; \qquad p_3 = -2{,}889109054$$

и т. д. (рис. 2.15(c)). Но если начальное значение достаточно близко к корню $p = 0$, в результате получим сходящуюся последовательность. Если $p_0 = 0{,}5$, то

$$p_1 = -0{,}079559511; \qquad p_2 = 0{,}000335302; \qquad p_3 = 0{,}000000000.$$

Ситуация, рассмотренная выше, указывает на тот факт, что следует быть осторожным при сообщении ответа. Иногда последовательность не сходится. Не всегда бывает так, что решение найдено после N итераций. Тот, кто использует алгоритм нахождения корня, должен быть предупрежден о ситуации, когда нельзя найти корень. Если существует дополнительная информация относительно смысла задачи, то менее вероятно, чем именно найденный корень ошибочен. Иногда известно, что корень $f(x)$ находится на определенном интервале. Если доступна информация о поведении функции или известен "точный" график, то легче выбрать p_0.

Метод секущих

В алгоритме Ньютона–Рафсона требуется вычислить две функции для каждой итерации — $f(p_{k-1})$ и $f'(p_{k-1})$. Традиционно вычисление производных элементарных функций требует значительных усилий. Но при наличии современного пакета программного обеспечения компьютеров эти сложности уменьшаются. Все еще много функций имеют непростую форму (интегралы, суммы и т. д.), и желательно иметь метод, который сходится почти так же быстро, как метод Ньютона, и включает только вычисление функции $f(x)$, но не вычисление $f'(x)$. Метод секущих потребует только одного вычисления функции $f(x)$ при одной итерации, и простой корень имеет порядок сходимости $R \approx 1{,}618033989$. Этот метод почти так же быстр, как и метод Ньютона, который имеет порядок сходимости 2.

В методе секущих используется такая же формула, как и в методе regula falsi, но существуют различные логические решения относительно способа поиска каждого последующего члена. Необходимо иметь две начальные точки, $(p_0; f(p_0))$ и $(p_1; f(p_1))$, около точки $(p; 0)$ как показано на рис. 2.16. Определим p_2 как абсциссу точки пересечения линии, проходящей через эти две точки, и оси x. Тогда на рис. 2.16 показано, что p_2 будет ближе к p, чем любая из точек p_0 или p_1. Уравнение, связывающее p_2, p_1 и p_0, находим, рассматривая тангенс угла наклона

$$m = \frac{f(p_1) - f(p_0)}{p_1 - p_0} \quad \text{и} \quad m = \frac{0 - f(p_1)}{p_2 - p_1}. \tag{25}$$

Значения m в (25) равны тангенсу угла наклона секущей, которая проходит через два первых приближения к тангенсу угла наклона прямой, проходящей через точки $(p_1; f(p_1))$ и $(p_2; 0)$ соответственно. Приравняем правые части в (25), решим относительно $p_2 = g(p_1; p_0)$ и получим

$$p_2 = g(p_1, p_0) = p_1 - \frac{f(p_1)(p_1 - p_0)}{f(p_1) - f(p_0)}. \tag{26}$$

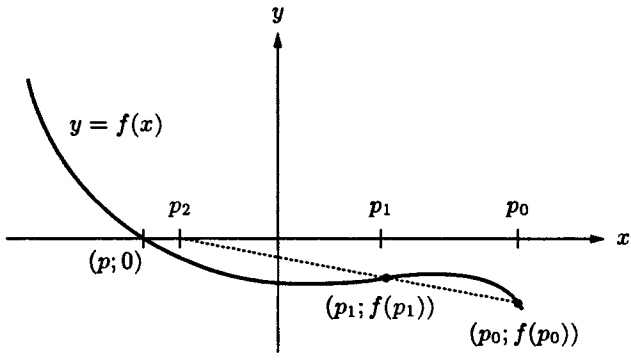


Рис. 2.16. Геометрическое построение p_2 для метода секущих

Общий член, определенный согласно двухточечной итерационной формуле, равен

$$p_{k+1} = g(p_k, p_{k-1}) = p_k - \frac{f(p_k)(p_k - p_{k-1})}{f(p_k) - f(p_{k-1})}. \tag{27}$$

Пример 2.16 (метод секущих при простом корне). Начнем с $p_0 = -2{,}6$ и $p_1 = -2{,}4$ и воспользуемся методом секущих, чтобы найти корень $p = -2$ полиномиальной функции $f(x) = x^3 - 3x + 2$.

В этом случае итерационная формула (27) имеет вид

$$p_{k+1} = g(p_k, p_{k-1}) = p_k - \frac{(p_k^3 - 3p_k + 2)(p_k - p_{k-1})}{p_k^3 - p_{k-1}^3 - 3p_k + 3p_{k-1}}. \tag{28}$$

Преобразовав эту формулу, получим

$$p_{k+1} = g(p_k, p_{k-1}) = \frac{p_k^2 p_{k-1} + p_k p_{k-1}^2 - 2}{p_k^2 + p_k p_{k-1} + p_{k-1}^2 - 3}. \tag{29}$$

Последовательность итераций приведена в табл. 2.7. ■

Существует связь между методом секущих и методом Ньютона. Для полиномиальной функции $f(x)$ двухточечная формула $p_{k+1} = g(p_k; p_{k-1})$ метода секущих может быть приведена к одноточечной формуле метода Ньютона $p_{k+1} = g(p_k)$, если p_{k-1} заменить на p_k. Действительно, если заменить p_{k-1} на p_k в (29), то правая часть становится такой же, как правая часть формулы (22) примера 2.14.

Результаты относительно скорости сходимости метода секущих можно найти в специальных книгах по численному анализу. Заметим только, что члены последовательности ошибок удовлетворяет соотношению

$$|E_{k+1}| \approx |E_k|^{1{,}618} \left| \frac{f''(p)}{2f'(p)} \right|^{0{,}618}, \tag{30}$$

Таблица 2.7. Сходимость метода секущих к простому корню

k	p_k	$p_{k+1}-p_k$	$E_k = p - p_k$	$\frac{\lvert E_{k+1}\rvert}{\lvert E_k\rvert^{1,618}}$
0	−2,600000000	0,200000000	0,600000000	0,914152831
1	−2,400000000	0,293401015	0,400000000	0,469497765
2	−2,106598985	0,083957573	0,106598985	0,847290012
3	−2,022641412	0,021130314	0,022641412	0,693608922
4	−2,001511098	0,001488561	0,001511098	0,825841116
5	−2,000022537	0,000022515	0,000022537	0,727100987
6	−2,000000022	0,000000022	0,000000022	
7	−2,000000000	0,000000000	0,000000000	

где порядок сходимости равен $R = (1+\sqrt{5})/2 \approx 1{,}618$, и что соотношение (30) выполняется только при простых корнях.

Чтобы проверить это, воспользуемся примером 2.16, специально подбирая следующие значения:

$$|p - p_5| = 0{,}000022537$$
$$|p - p_4|^{1{,}618} = 0{,}001511098^{1{,}618} = 0{,}000027296,$$

и

$$A = |f''(-2)/2f'(-2)|^{0{,}618} = (2/3)^{0{,}618} = 0{,}778351205.$$

Теперь легко видеть, что

$$|p - p_5| = 0{,}000022537 \approx 0{,}000021246 = A|p - p_4|^{1{,}618}.$$

Ускоренная сходимость

Можно надеяться, что существуют технические приемы нахождения корня, которые сходятся быстрее, чем линейно, когда p — корень порядка M. Следующий результат показывает, что можно так модифицировать метод Ньютона, что для кратного корня сходимость станет квадратичной.

Теорема 2.7 (ускорение итераций Ньютона–Рафсона). Предположим, что алгоритм Ньютона–Рафсона порождает последовательность, которая линейно сходится к корню $x = p$ порядка $M > 1$. Тогда интерполяционная формула Ньютона–Рафсона

$$p_k = p_{k-1} - \frac{Mf(p_{k-1})}{f'(p_{k-1})} \tag{31}$$

производит последовательность $\{p_k\}_{k=0}^{\infty}$, которая квадратично сходится к p.

Пример 2.17 (ускоренная сходимость к двойному корню). Начнем с $p_0 = 1{,}2$ и воспользуемся итерацией ускоренного метода Ньютона–Рафсона, чтобы найти двойной корень $p = 1$ функции $f(x) = x^3 - 3x + 2$.

При $M = 2$ формула ускоренного метода (31) принимает вид

$$p_k = p_{k-1} - 2\frac{f(p_{k-1})}{f'(p_{k-1})} = \frac{p_{k-1}^3 + 3p_{k-1} - 4}{3p_{k-1}^2 - 3},$$

и мы получаем значения, показанные в табл. 2.8. ■

Таблица 2.8. Ускоренная сходимость к двойному корню

k	p_k	$p_{k+1} - p_k$	$E_k = p - p_k$	$\frac{\lvert E_{k+1}\rvert}{\lvert E_k\rvert^2}$
0	1,200000000	−0,193939394	−0,200000000	0,151515150
1	1,006060606	−0,006054519	−0,006060606	0,165718578
2	1,000006087	−0,000006087	−0,000006087	
3	1,000000000	0,000000000	0,000000000	

В табл. 2.9 сравниваются скорости сходимостей различных методов нахождения корня, которые мы изучили. Значения константы A различны для каждого метода.

Таблица 2.9. Сравнение скорости сходимости

Метод	Кратность корня	Соотношение между членами последовательности ошибок
Деления пополам		$E_{k+1} \approx \frac{1}{2}\lvert E_k\rvert$
Regula falsi		$E_{k+1} \approx A\lvert E_k\rvert$
Метод секущих	Кратный корень	$E_{k+1} \approx A\lvert E_k\rvert$
Ньютона–Рафсона	Кратный корень	$E_{k+1} \approx A\lvert E_k\rvert$
Метод секущих	Простой корень	$E_{k+1} \approx A\lvert E_k\rvert^{1{,}618}$
Ньютона–Рафсона	Простой корень	$E_{k+1} \approx A\lvert E_k\rvert^2$
Ускоренный метод Ньютона–Рафсона	Кратный корень	$E_{k+1} \approx A\lvert E_k\rvert^2$

Программа 2.5 (итерация Ньютона–Рафсона). Найти корень функции $f(x) = 0$ с одним заданным начальным приближением p_0, используя итерацию

$$p_k = p_{k-1} - \frac{f(p_{k-1})}{f'(p_{k-1})} \quad \text{для } k = 1,\ 2,\ \ldots.$$

```
function [p0,err,k,y]=newton(f,df,p0,delta,epsilon,max1)
% Вход - f - функция, вводимая как строка  'f'
%      - df - производная f, вводимая как строка 'df'
```

```
%       - p0 - начальное приближение функции f к нулю
%       - delta - допустимое отклонение для p0
%       - epsilon - допустимое отклонение для значений функции y
%       - max1 - максимальное число итераций
%Выход - p0 - приближение Ньютона-Рафсона к нулю
%       - err - ошибка вычисления для p0
%       - k - число итераций
%       - y - значение функции f(p0)
for k=1:max1
   p1=p0-feval(f,p0)/feval(df,p0);
   err=abs(p1-p0);
   relerr=2*err/(abs(p1)+delta);
   p0=p1;
   y=feval(f,p0);
   if (err<delta)|(relerr<delta)|(abs(y)<epsilon),break,end
end
```

Программа 2.6 (метод секущих). Найти корень уравнения $f(x) = 0$ с двумя заданными начальными приближениями p_0, p_1 и используя итерацию

$$p_{k+1} = p_k - \frac{f(p_k)(p_k - p_{k-1})}{f(p_k) - f(p_{k-1})} \quad \text{для } k = 1,\ 2,\ \ldots.$$

```
function [p1,err,k,y]=secant(f,p0,p1,delta,epsilon,max1)
% Ввод - f - функция, вводимая как строка 'f'
%      - p0 и p1 - начальные приближения к нулю
%      - delta - допустимое отклонение для p1
%      - epsilon - допустимое отклонение для значений функции y
%      - max1 - максимальное число итераций
%Выход - p1 - приближение к нулю для метода секущих
%      - err - ошибка вычисления для p1
%      - k - число итераций
%      - y - значение функции f(p1)
for k=1:max1
   p2=p1-feval(f,p1)*(p1-p0)/(feval(f,p1)-feval(f,p0));
   err=abs(p2-p1);
   relerr=2*err/(abs(p2)+delta);
   p0=p1;
   p1=p2;
   y=feval(f,p1);
   if (err<delta)|(relerr<delta)|(abs(y)<epsilon),break,end
end
```

Упражнения к разделу 2.4

Для задач, требующих вычисления, можете использовать либо калькулятор, либо компьютер.

1. Пусть задана функция $f(x) = x^2 - x + 2$.
   (a) Найдите формулу Ньютона–Рафсона для $p_k = g(p_{k-1})$.
   (b) Начните с $p_0 = -1{,}5$ и найдите p_1, p_2 и p_3.
2. Пусть функция $f(x) = x^2 - x - 3$.
   (a) Найдите формулу Ньютона–Рафсона для $p_k = g(p_{k-1})$.
   (b) Начните с $p_0 = 1{,}6$ и найдите p_1, p_2 и p_3.
   (c) Начните с $p_0 = 0{,}0$ и найдите p_1, p_2, p_3 и p_4. Что можно сказать об этой последовательности?
3. Пусть дана функция $f(x) = (x-2)^4$.
   (a) Найдите формулу Ньютона–Рафсона для $p_k = g(p_{k-1})$.
   (b) Начните с $p_0 = 2{,}1$ и найдите p_1, p_2, p_3 и p_4.
   (c) Эта последовательность сходится квадратично или линейно?
4. Пусть дана функция $f(x) = x^3 - 3x - 2$.
   (a) Найдите формулу Ньютона–Рафсона для $p_k = g(p_{k-1})$.
   (b) Начните с $p_0 = 2{,}1$ и найдите p_1, p_2, p_3 и p_4.
   (c) Эта последовательность сходится квадратично или линейно?
5. Рассмотрите функцию $f(x) = \cos(x)$.
   (a) Найдите формулу Ньютона–Рафсона для $p_k = g(p_{k-1})$.
   (b) Найдите корень $p = 3\pi/2$. Можно ли использовать начальное приближение $p_0 = 3$? Почему?
   (c) Найдите корень $p = 3\pi/2$. Можно ли использовать начальное приближение $p_0 = 5$? Почему?
6. Рассмотрите функцию $f(x) = \arctan(x)$.
   (a) Найдите формулу Ньютона–Рафсона для $p_k = g(p_{k-1})$.
   (b) Если $p_0 = 1{,}0$, найдите p_1, p_2, p_3 и p_4. Чему равен $\lim_{n\to\infty} p_k$?
   (c) Если $p_0 = 2{,}0$, найдите p_1, p_2, p_3 и p_4. Чему равен $\lim_{n\to\infty} p_k$?
7. Рассмотрите функцию $f(x) = xe^{-x}$.
   (a) Найдите формулу Ньютона–Рафсона для $p_k = g(p_{k-1})$.
   (b) Если $p_0 = 0{,}2$, найдите p_1, p_2, p_3 и p_4. Чему равен $\lim_{n\to\infty} p_k$?
   (c) Если $p_0 = 20$, найдите p_1, p_2, p_3 и p_4. Чему равен $\lim_{n\to\infty} p_k$?
   (d) Чему равно значение $f(p_4)$ в п. (c)?

В упр. 8–10, используя метод секущих и формулу (27), вычислите следующие две итерации p_2 и p_3.

8. Пусть $f(x) = x^2 - 2x - 1$. Начните с $p_0 = 2{,}6$ и $p_1 = 2{,}5$.
9. Пусть $f(x) = x^2 - x - 3$. Начните с $p_0 = 1{,}7$ и $p_1 = 1{,}67$.
10. Пусть $f(x) = x^3 - x + 2$. Начните с $p_0 = -1{,}5$ и $p_1 = -1{,}52$.
11. *Алгоритм для нахождения корня кубического уравнения.* Начните с функции $f(x) = x^3 - A$, где A — любое действительное число, и используйте рекуррентную формулу

$$p_k = \frac{2p_{k-1} + A/p_{k-1}^2}{3} \quad \text{для } k = 1,\ 2,\ \ldots.$$

12. Рассмотрите функцию $f(x) = x^N - A$, где N — положительное целое число.
    (a) Чему равны действительные решения уравнения $f(x) = 0$ для различных N и A?
    (b) Используйте рекуррентную формулу

$$p_k = \frac{(N-1)p_{k-1} + A/p_{k-1}^{N-1}}{N} \quad \text{для } k = 1,\ 2,\ \ldots$$

    для нахождения N-го корня A.
13. Можно ли использовать итерацию Ньютона–Рафсона, чтобы решить уравнение $f(x) = 0$, если $f(x) = x^2 - 14x + 50$? Объясните, почему?
14. Можно ли использовать итерацию Ньютона–Рафсона, чтобы решить уравнение $f(x) = 0$, если $f(x) = x^{1/3}$? Почему?
15. Можно ли использовать итерацию Ньютона–Рафсона, чтобы решить уравнение $f(x) = 0$, если $f(x) = (x-3)^{1/2}$ и начальное значение равно $p_0 = 4$? Почему?
16. Найдите предел последовательности в (11).
17. Докажите, что последовательность $\{p_k\}$ в итерации (4) теоремы 2.5 сходится к p. Используйте следующие шаги.
    (a) Покажите, что если p — неподвижная точка $g(x)$ в формуле (5), то p является нулем функции $f(x)$.
    (b) Если p — нуль функции $f(x)$ и $f'(p) \neq 0$, то покажите, что $g'(p) = 0$. Воспользуйтесь п. (b) и теоремой 2.3, чтобы показать, что последовательность $\{p_k\}$ в итерации (4) сходится к p.
18. Докажите приближенное равенство (23) теоремы 2.6. Ипользуйте приведенную ниже последовательность действий. Согласно теореме 1.11 можно разложить функцию $f(x)$ в окрестности точки $x = p_k$, чтобы получить

$$f(x) = f(p_k) + f'(p_k)(x - p_k) + \frac{1}{2}f''(c_k)(x - p_k)^2.$$

Так как p — нуль функции $f(x)$, положим $x = p$ и получим

$$0 = f(p_k) + f'(p_k)(p - p_k) + \frac{1}{2}f''(c_k)(p - p_k)^2.$$

(a) Сейчас предположим, что $f'(x) \neq 0$ для всех x около корня p. Используем приведенные выше факты и то, что $f'(p_k) \neq 0$, чтобы показать следующее:

$$p - p_k + \frac{f(p_k)}{f'(p_k)} = \frac{-f''(c_k)}{2f'(p_k)}(p - p_k)^2.$$

(b) Предположим, что $f'(x)$ и $f''(x)$ не настолько быстро изменяются, чтобы можно было использовать приближения $f'(p_k) \approx f'(p)$ и $f''(c_k) \approx f''(p)$. Тогда используем п. (a), чтобы получить

$$E_{k+1} \approx \frac{-f''(p)}{2f'(p)}E_k^2.$$

19. Предположим, что A — действительное положительное число.

(a) Покажите, что A можно представить в виде $A = q \times 2^{2m}$, где $1/4 \leq q < 1$ и m — целое число.

(b) Используйте п. (a), чтобы показать, что квадратный корень равен $A^{1/2} = q^{1/2} \times 2^m$. *Примечание.* Пусть $p_0 = (2q + 1)/3$, где $1/4 \leq q < 1$; воспользуйтесь формулой Ньютона (11). После трех итераций p_3 будет приближением к $q^{1/2}$ с точностью, равной 24 двоичным знакам. Этот алгоритм часто используется во встроенных прикладных программах на компьютере для вычисления квадратного корня.

20. **(a)** Покажите, что формула (27) метода секущих является алгебраическим эквивалентом выражения

$$p_{k+1} = \frac{p_{k-1}f(p_k) - p_k f(p_{k-1})}{f(p_k) - f(p_{k-1})}.$$

(b) Объясните, почему потеря значащих разрядов в вычитании делает эту формулу худшей для вычислений, чем формула (27).

21. Предположим, что p — корень порядка $M = 2$ уравнения $f(x) = 0$. Докажите, что ускоренная итерация Ньютона–Рафсона

$$p_k = p_{k-1} - \frac{2f(p_{k-1})}{f'(p_{k-1})}$$

сходится квадратично (см. упр. 18).

22. ***Метод Хейли*** является еще одним из способов улучшения скорости сходимости метода Ньютона. Итерационная формула Хейли имеет вид

$$g(x) = x - \frac{f(x)}{f'(x)}\left(1 - \frac{f(x)f''(x)}{2(f'(x))^2}\right)^{-1}.$$

Член в скобках — это модификация формулы Ньютона–Рафсона. Метод Хейли дает кубическую сходимость ($R = 3$) к простому нулю функции $f(x)$.

(a) Начните с функции $f(x) = x^2 - A$ и найдите итерационную формулу Хейли $g(x)$ для нахождения $\sqrt{A}$. Используйте $p_0 = 2$ в качестве приближения к $\sqrt{5}$ и вычислите p_1, p_2 и p_3.

(b) Начните с функции $f(x) = x^3 - 3x + 2$ и выведите итерационную формулу Хейли $g(x)$. Используйте $p_0 = -2{,}4$ и вычислите p_1, p_2 и p_3.

23. *Модифицированный метод Ньютона–Рафсона для кратных корней.* Если p — корень кратности M, то $f(x) = (x - p)^M q(x)$, где $q(p) \neq 0$.

(a) Покажите, что $h(x) = f(x)/f'(x)$ имеет простой корень в p.

(b) Покажите, что, когда применяется метод Ньютона–Рафсона для нахождения простого корня p функции $h(x)$, получаем $g(x) = x - h(x)/h'(x)$, которая приводится к виду

$$g(x) = x - \frac{f(x)f'(x)}{(f'(x))^2 - f(x)f''(x)}.$$

(c) Итерация, использующая $g(x)$ в п. (b), сходится квадратично к p. Объясните, почему так происходит.

(d) Нуль — корень кратности 3 для функции $f(x) = \sin(x^3)$. Начните с $p_0 = 1$ и вычислите p_1, p_2 и p_3, используя модифицированный метод Ньютона–Рафсона.

24. Предположим, что итеративный метод решения уравнения $f(x) = 0$ производит следующие четыре члена последовательности ошибок (см. пример 2.11): $E_0 = 0{,}400000$; $E_1 = 0{,}043797$; $E_2 = 0{,}000062$ и $E_3 = 0{,}000000$. Оцените асимптотическую ошибку A и порядок сходимости R последовательности, генерируемой итеративным методом.

Алгоритмы и программы

1. Преобразуйте программы 2.5 и 2.6, чтобы появлялось сообщение о соответствующей ошибке, когда (i) возникает деление на нуль в итерационных формулах (4) и (27) соответственно или (ii) превышено максимальное число итераций `max1`.

2. Часто поучительно показывать члены последовательностей, генерируемых итерационными формулами (4) и (27) (вторая колонка табл. 2.4). Преобразуйте программы 2.5 и 2.6 так, чтобы можно было показывать последовательности, генерируемые итерационными формулами (4) и (27) соответственно.

3. Преобразуйте программу 2.5 так, чтобы можно было использовать алгоритм Ньютона нахождения квадратного корня для нахождения каждого из следующих квадратных корней с 10 десятичными знаками.
   (a) Начните с $p_0 = 3$ и вычислите $\sqrt{8}$.
   (b) Начните с $p_0 = 10$ и вычислите $\sqrt{91}$.
   (c) Начните с $p_0 = -3$ и вычислите $-\sqrt{8}$.

4. Измените программу 2.5 так, чтобы можно было использовать алгоритм для нахождения кубического корня из упр. 11, чтобы найти кубические корни с 10 десятичными знаками.
   (a) Начните с $p_0 = 2$ и вычислите $7^{1/3}$.
   (b) Начните с $p_0 = 6$ и вычислите $200^{1/3}$.
   (c) Начните с $p_0 = -2$ и вычислите $(-7)^{1/3}$.

5. Измените программу 2.5 так, чтобы, используя алгоритм ускоренного метода теоремы 2.7, можно было найти корень p порядка M следующих функций.
   (a) $f(x) = (x-2)^5$, $M = 5$, $p = 2$; начать с $p_0 = 1$.
   (b) $f(x) = \sin(x^3)$, $M = 3$, $p = 0$; начать с $p_0 = 1$.
   (c) $f(x) = (x-1)\ln(x)$, $M = 2$, $p = 1$; начать с $p_0 = 2$.

6. Преобразуйте программу 2.5 так, чтобы, используя метод Хелли из упр. 22, можно было найти простой нуль функции $f(x) = x^3 - 3x + 2$, полагая $p_0 = -2{,}4$.

7. Предположим, что уравнение движения снаряда имеет вид

$$y = f(t) = 9600(1 - e^{-t/15}) - 480t,$$
$$x = r(t) = 2400(1 - e^{-t/15}).$$

   (a) Найдите время полета до падения с точностью до 10 десятичных знаков.
   (b) Найдите дальность полета с точностью до 10 десятичных знаков.

8. (a) Найдите точку на параболе $y = x^2$, примыкающую к точке $(3; 1)$, с точностью до 10 десятичных знаков.
   (b) Найдите точку на графике $y = \sin(x - \sin(x))$, примыкающую к точке $(2{,}1; 0{,}5)$, с точностью до 10 десятичных знаков.
   (c) Найдите значение x, равное минимальному расстоянию по вертикали между графиками функций $f(x) = x^2 + 2$ и $g(x) = (x/5) - \sin(x)$, с точностью до 10 десятичных знаков.

9. Открытый сверху ящик построен из прямоугольного куска листа металла размером 10×16 дюймов (1 дюйм $= 2{,}54$ см). Квадраты какого размера (точность до 0,000000001 дюймов) следует срезать по углам, если объем ящика должен быть равен 100 кубическим дюймам?

10. Цепная линия — это кривая, сделанная из подвесного каната. Предположим, что самая низкая точка — $(0;0)$. Тогда цепная линия описывается формулой $y = C\cosh(x/C) - C$. Чтобы найти цепную линию, которая проходит через точки $(\pm a; b)$, следует решить уравнение $b = C\cosh(a/C) - C$ для C.
   (a) Покажите, что цепная линия, которая проходит через точку $(\pm 10; 6)$, имеет вид $y = 9{,}1889\cosh(x/9{,}1889) - 9{,}1889$.
   (b) Найдите цепную линию, которая проходит через точки $(\pm 12; 5)$.

2.5. Процесс Эйткена и методы Стеффенсена и Мюллера (оптимальные)

В разделе 2.4 было показано, что метод Ньютона медленно сходится к кратному корню и порядок последовательности итераций $\{p_k\}$ — линейный. В теореме 2.7 показано, как ускорить сходимость, но это возможно, если заранее известен порядок корня.

Процесс Эйткена

Технику, называемую процесс ***Эйткена*** $\boldsymbol{\Delta^2}$, можно использовать для ускорения сходимости любой последовательности, которая сходится линейно. Для ее описания необходимо следующее определение.

Определение 2.6. Задана последовательность $\{p_n\}_{n=0}^{\infty}$. Определим сначала разности Δp_n

$$\Delta p_n = p_{n+1} - p_n \quad \text{для } n \geq 0. \tag{1}$$

Разности высшего порядка $\Delta^k p_n$ определяются рекуррентно по формуле

$$\Delta^k p_n = \Delta^{k-1}(\Delta p_n) \quad \text{для } k \geq 2. \tag{2}$$

▲

Теорема 2.8 (ускорение Эйткена). Предположим, что последовательность $\{p_n\}_{n=0}^{\infty}$ линейно сходится к пределу p и что $p - p_n \neq 0$ для всех $n \geq 0$. Если существует такое действительное число A, $|A| < 1$, что

$$\lim_{n\to\infty} \frac{p - p_{n+1}}{p - p_n} = A, \tag{3}$$

то последовательность $\{q_n\}_{n=0}^{\infty}$, определяемая формулой

$$q_n = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n} = p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}, \tag{4}$$

Таблица 2.10. Линейно сходящаяся последовательность $\{p_n\}$

n	p_n	$E_n = p_n - p$	$A_n = \dfrac{E_n}{E_{n-1}}$
1	0,606530660	0,039387369	−0,586616609
2	0,545239212	−0,021904079	−0,556119357
3	0,579703095	0,012559805	−0,573400269
4	0,560064628	−0,007078663	−0,563596551
5	0,571172149	0,004028859	−0,569155345
6	0,564862947	−0,002280343	−0,566002341

сходится к p быстрее, чем $\{p_n\}_{n=0}^{\infty}$, в том смысле, что

$$\lim_{n\to\infty} \left| \frac{p - q_n}{p - p_n} \right| = 0. \tag{5}$$

Доказательство. Покажем, как вывести формулу (4), и оставим читателю доказательство (5) в качестве упражнения. Так как члены в (3) стремятся к пределу, можно записать

$$\frac{p - p_{n+1}}{p - p_n} \approx A \quad \text{и} \quad \frac{p - p_{n+2}}{p - p_{n+1}} \approx A \quad \text{когда } n \text{ большое.} \tag{6}$$

Из соотношений (6) следует, что

$$(p - p_{n+1})^2 \approx (p - p_{n+2})(p - p_n). \tag{7}$$

После преобразования обеих частей (7) члены с p^2 пропадут, и в результате получим

$$p \approx \frac{p_{n+2}p_n - p_{n+1}^2}{p_{n+2} - 2p_{n+1} + p_n} = q_n \quad \text{для } n = 0,\ 1,\ \ldots. \tag{8}$$

Формулу (8) используют для определения члена q_n. Формулу (4) можно получить из (8) путем простых алгебраических преобразований. При ее использовании ошибка увеличивается меньше при вычислениях на компьютере. •

Пример 2.18. Покажем, что последовательность $\{p_n\}$ из примера 2.2 имеет линейный порядок сходимости и что последовательность $\{q_n\}$, полученная в ходе процесса Эйткена Δ^2, сходится быстрее.

Последовательность $\{p_n\}$ получена методом итерации неподвижной точки для функции $g(x) = e^{-x}$ с начальным приближением $p_0 = 0{,}5$. Последовательность

Таблица 2.11. Получение последовательности $\{q_n\}$ с использованием процесса Эйткена

n	q_n	$q_n - p$
1	0,567298989	0,000155699
2	0,567193142	0,000049852
3	0,567159364	0,000016074
4	0,567148453	0,000005163
5	0,567144952	0,000001662
6	0,567143825	0,000000534

итераций сходится к пределу $P \approx 0{,}567143290$. Значения p_n и q_n приведены в табл. 2.10 и 2.11. Для наглядности покажем, как вычисляется q_1:

$$q_1 = p_1 - \frac{(p_2 - p_1)^2}{p_3 - 2p_2 + p_1} =$$

$$= 0{,}606530660 - \frac{(-0{,}061291448)^2}{0{,}095755331} = 0{,}567298989. \quad \blacksquare$$

Хотя последовательность $\{q_n\}$ в табл. 2.11 сходится линейно, она сходится быстрее, чем последовательность $\{p_n\}$, в смысле теоремы 2.8. Обычно метод Эйткена дает большее улучшение, чем полученное. Если объединить процесс Эйткена и итерацию неподвижной точки, то в результате получим метод, называемый ***ускорением Стеффенсена***. Детали приведены в программе 2.7 и упражнениях.

Метод Мюллера

Метод Мюллера является обобщением метода секущих в том смысле, что в нем не требуется производная функции. Это итерационный метод, для которого необходимы три начальные точки: $(p_0, f(p_0))$, $(p_1, f(p_1))$ и $(p_2, f(p_2))$. Далее строится парабола, проходящая через эти три точки, и следующее приближение находится, как корень квадратного уравнения. Докажем, что в окрестности простого корня метод Мюллера сходится быстрее, чем метод секущих, и почти так же быстро, как метод Ньютона. Этот метод можно использовать, чтобы найти действительные или комплексные нули функции. Его можно запрограммировать с использованием комплексной арифметики.

Без потери общности предположим, что p_2 — наилучшее приближение к корню, и рассмотрим параболу, которая проходит через три начальных значения, как показано на рис. 2.17. Сделаем замену переменной

$$t = x - p_2, \tag{9}$$

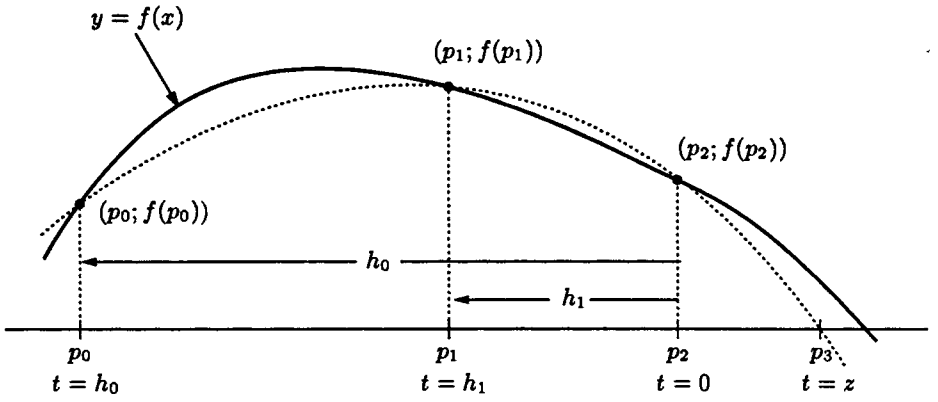


Рис. 2.17. Начальные приближения p_0, p_1 и p_2 для метода Мюллера и разности h_0 и h_1

и используем разности

$$h_0 = p_0 - p_2 \quad \text{и} \quad h_1 = p_1 - p_2. \tag{10}$$

Рассмотрим квадратный полином от переменной t:

$$y = at^2 + bt + c. \tag{11}$$

Будем использовать каждую точку, чтобы получить уравнение для нахождения a, b и c:

$$\begin{aligned} &\text{при } t = h_0: && ah_0^2 + bh_0 + c = f_0, \\ &\text{при } t = h_1: && ah_1^2 + bh_1 + c = f_1, \\ &\text{при } t = 0: && a0^2 + b0 \ + c = f_2. \end{aligned} \tag{12}$$

Из третьего уравнения в (12) видно, что

$$c = f_2. \tag{13}$$

Подставляя (13) в первые два уравнения из (12) и используя определение $e_0 = f_0 - c$ и $e_1 = f_1 - c$, получим в результате систему линейных уравнений

$$\begin{aligned} ah_0^2 + bh_0 &= f_0 - c = e_0, \\ ah_1^2 + bh_1 &= f_1 - c = e_1. \end{aligned} \tag{14}$$

Решаем ее относительно a и b и получаем в результате

$$\begin{aligned} a &= \frac{e_0 h_1 - e_1 h_0}{h_1 h_0^2 - h_0 h_1^2} \\ b &= \frac{e_1 h_0^2 - e_0 h_1^2}{h_1 h_0^2 - h_0 h_1^2}. \end{aligned} \tag{15}$$

Квадратичная формула используется для нахождения корней $t = z_1, z_2$ уравнения (11):

$$z = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}. \tag{16}$$

Формула (16) эквивалентна обычной формуле для вычисления корней квадратного уравнения, но удобнее для данного случая, так как известно, что $c = f_2$.

Чтобы гарантировать устойчивость метода, выберем в (16) корень, наименьший по абсолютной величине. Если $b > 0$, используем положительный знак квадратного корня, и, если $b < 0$, используем отрицательный знак. Тогда p_3 показано на рис. 2.17 и задается равенством

$$p_3 = p_2 + z. \tag{17}$$

Чтобы выбрать итерации, возьмем два значения, p_0 и p_1, среди $\{p_0, p_1, p_3\}$, которые лежат ближе к p_3 (т. е. отбросим то, которое находится дальше остальных). Затем поменяем p_2 с p_3. Несмотря на то что в методе Мюллера проделано множество дополнительных вычислений, он требует вычисления только одного значения функции на итерацию.

Если использовать метод Мюллера, чтобы находить действительные корни уравнения $f(x) = 0$, то, вероятно, может встретиться комплексное приближение, так как квадратные корни в (16) могут быть комплексными (не равная нулю мнимая часть). В этих случаях мнимая часть будет иметь малую величину и может быть установлена равной нулю, так что вычисления продолжатся с действительными числами.

Сравнение методов

Метод Стеффенсена можно использовать вместе с методом Ньютона–Рафсона для нахождения неподвижной точки функции $g(x) = x - f(x)/f'(x)$. В следующих двух примерах рассмотрим корни полинома $f(x) = x^3 - 3x + 2$. Функцией Ньютона–Рафсона является $g(x) = (2x^3 - 2)/(3x^2 - 3)$. Как только эта функция будет использована в программе 2.7, приведем вычисления в колонке для методов Стеффенсена и Ньютона табл. 2.12 и 2.13. Например, начиная с $p_0 = -2{,}4$, можно вычислить

$$p_1 = g(p_0) = -2{,}076190476, \tag{18}$$

и

$$p_2 = g(p_1) = -2{,}003596011. \tag{19}$$

Тогда улучшение Эйткена даст значение $p_3 = -1{,}982618143$.

Таблица 2.12. Сравнение сходимости около простого корня

k	Метод секущих	Метод Мюллера	Метод Ньютона	Метод Стеффенсена и Ньютона
0	−2,600000000	−2,600000000	−2,400000000	−2,400000000
1	−2,400000000	−2,500000000	−2,076190476	−2,076190476
2	−2,106598985	−2,400000000	−2,003596011	−2,003596011
3	−2,022641412	−1,985275287	−2,000008589	−1,982618143
4	−2,001511098	−2,000334062	−2,000000000	−2,000204982
5	−2,000022537	−2,000000218		−2,000000028
6	−2,000000022	−2,000000000		−2,000002389
7	−2,000000000			−2,000000000

Таблица 2.13. Сравнение сходимости около двойного корня

k	Метод секущих	Метод Мюллера	Метод Ньютона	Метод Стеффенсена и Ньютона
0	1,400000000	1,400000000	1,200000000	1,200000000
1	1,200000000	1,300000000	1,103030303	1,103030303
2	1,138461538	1,200000000	1,052356417	1,052356417
3	1,083873738	1,003076923	1,026400814	0,996890433
4	1,053093854	1,003838922	1,013257734	0,998446023
5	1,032853156	1,000027140	1,006643418	0,999223213
6	1,020429426	0,999997914	1,003325375	0,999999193
7	1,012648627	0,999999747	1,001663607	0,999999597
8	1,007832124	1,000000000	1,000832034	0,999999798
9	1,004844757		1,000416075	0,999999999
	⋮		⋮	

Пример 2.19 (сходимость около простого корня). Это сравнение методов для функции $f(x) = x^3 - 3x + 2$ около простого корня $p = -2$.

Метод Ньютона и метод секущих для этой функции приведены в примерах 2.14 и 2.16 соответственно. В таб. 2.12 приведены результаты вычислений для этих методов. ■

Пример 2.20 (сходимость около двойного корня). Сравнение методов для функции $f(x) = x^3 - 3x + 2$ около двойного корня $p = 1$. В таб. 2.13 приведены результаты вычислений. ■

Метод Ньютона является наилучшим для нахождения простого корня (см. табл. 2.12). Для двойного корня хорошо выбирать либо метод Мюллера, либо

метод Стеффенсена с формулой Ньютона–Рафсона (см. табл. 2.13). Отметим, что в формуле (4) ускоренного метода Эйткена может появиться деление на нуль, тогда как последовательность $\{p_k\}$ сходится. В этом случае последнее вычисленное приближение к нулю следует использовать в качестве приближения к нулю функции f.

В приведенной ниже программе последовательность $\{p_k\}$, которая генерируется методом Стеффенсена с формулой Ньютона–Рафсона, хранится в матрице $\boldsymbol{Q}$, состоящей из max1 строк и 3 столбцов. В первом столбце $\boldsymbol{Q}$ содержатся начальное приближение к корню, p_0, и члены $p_3, p_6, \ldots, p_{3k}, \ldots$, генерируемые ускоренным методом Эйткена (4). Во втором и третьем столбцах матрицы $\boldsymbol{Q}$ содержатся члены, генерируемые методом Ньютона. Критерий останова в программе основан на разнице между последовательными членами первого столбца матрицы $\boldsymbol{Q}$.

Программа 2.7 (ускоренный метод Стеффенсена). Быстрое нахождение решения уравнения неподвижной точки $x = g(x)$ с заданным начальным приближением p_0. Предполагается, что и $g(x)$, и $g'(x)$ непрерывны, $|g'(x)| < 1$ и обычная итерация неподвижной точки медленно (линейно) сходится к p.

```
function [p,Q]=steff(f,df,p0,delta,epsilon,max1)
% Вход - f - функция, вводимая как строка 'f'
%      - df - производная, вводимая как строка 'df'
%      - p0 - начальное приближение к нулю функции f
%      - delta - допустимое отклонение для p0
%      - epsilon - допустимое отклонение для значений функции y
%      - max1 - максимальное число итераций
%Выход - p - приближение Стеффенсена к нулю
%      - Q - матрица, содержащая последовательность Стеффенсена
%Инициализация матрицы R
R=zeros(max1,3);
R(1,1)=p0;
for k=1:max1
   for j=2:3
      %Вычисление знаменателя в формуле метода Ньютона-Рафсона
      nrdenom=feval(df,R(k,j-1));
      %Вычисление приближений Ньютона-Рафсона
      if nrdenom==0
         'деление на нуль в методе Ньютона-Рафсона'
         break
      else
         R(k,j)=R(k,j-1)-feval(f,R(k,j-1))/nrdenom;
      end
```

```
      %Вычисление знаменателя в ускоренном процессе Эйткена
      aadenom=R(k,3)-2*R(k,2)+R(k,1);
      %Вычисление приближений ускоренного процесса Эйткена
      if aadenom==0
         'деление на нуль в ускоренном процессе Эйткена'
         break
      else
         R(k+1,1)=R(k,1)-(R(k,2)-R(k,1))^2/aadenom;
      end
   end
   %Конец программы при появлении деления на нуль
   if (nrdenom==0)|(aadenom==0)
      break
   end
   %Вычисление p для критерия останова и матрицы Q
   err=abs(R(k,1)-R(k+1,1));
   relerr=err/(abs(R(k+1,1))+delta);
   y=feval(f,R(k+1,1));
   if (err<delta)|(relerr<delta)|(y<epsilon)
      % вычисляем p и матрицу Q
      p=R(k+1,1);
      Q=R(1:k+1,:);
      break
   end
end
```

Программа 2.8 (метод Мюллера). Нахождение корня уравнения $f(x) = 0$ с тремя заданными различными начальными приближениями p_0, p_1 и p_2.

```
function [p,y,err]=muller(f,p0,p1,p2,delta epsilon,max1)
% Вход - f - функция, вводимая как строка 'f'
%      - p0, p1 и p2 - начальные приближения
%      - delta - допустимое отклонение для p0, p1 и p2
%      - epsilon - допустимое отклонение для значений функции y
%      - max1 - максимальное число итераций
%Выход - p - приближение Мюллера к нулю функции f
%      - y - значение функции  y = f(p)
%      - err - ошибка приближения к p
%Инициализация матриц P и Y
P=[p0 p1 p2];
Y=feval(f,P);
```

```
%Вычисление a и b в формуле (15)
for k=1:max1
   h0=P(1)-P(3);h1=P(2)-P(3);e0=Y(1)-Y(3);e1=Y(2)-Y(3);c=Y(3);
   denom=h1*h0^2-h0*h1^2;
   a=(e0*h1-e1*h0)/denom;
   b=(e1*h0^2-e0*h1^2)/denom;
   %Подавление любых комплексных корней
   if b^2-4*a*c > 0
      disc=sqrt(b^2-4*a*c);
   else
      disc=0;
   end
   %Нахождение наименьшего корня уравнения (17)
   if b < 0
      disc=-disc;
   end
   z=-2*c/(b+disc);
   p=P(3)+z;
   %Сортировка входных P для поиска двух ближайших к p
   if abs(p-P(2))<abs(p-P(1))
      Q=[P(2) P(1) P(3)];
      P=Q;
      Y=feval(f,P);
   end
   if abs(p-P(3))<abs(p-P(2))
      R=[P(1) P(3) P(2)];
      P=R;
      Y=feval(f,P);
   end
   %Замена входного P ближайшим от p pn
   P(3)=p;
   Y(3) = feval(f,P(3));
   y=Y(3);
   %Вычисление критерия останова
   err=abs(z);
   relerr=err/(abs(p)+delta);
   if (err<delta)|(relerr<delta)|(abs(y)<epsilon)
      break
   end
end
```

Упражнения к разделу 2.5

1. Найдите Δp_n, где

   (a) $p_n = 5$ (b) $p_n = 6n + 2$ (c) $p_n = n(n + 1)$

2. Пусть $p_n = 2n^2 + 1$. Найдите $\Delta^k p_n$, где

   (a) $k = 2$ (b) $k = 3$ (c) $k = 4$

3. Пусть $p_n = 1/2^n$. Покажите, что $q_n = 0$ для всех n, где q_n задано формулой (4).

4. Пусть $p_n = 1/n$. Покажите, что $q_n = 1/(2n + 2)$ для всех n. Следовательно, существует небольшое ускорение сходимости. Сходится ли $\{p_n\}$ к 0 линейно? Почему?

5. Пусть $p_n = 1/(2^n - 1)$. Покажите, что $q_n = 1/\left(4^{n+1} - 1\right)$ для всех n.

6. Последовательность $p_n = 1/(4^n + 4^{-n})$ сходится линейно к 0. Воспользуйтесь формулой Эйткена (4), чтобы найти q_1, q_2 и q_3, и таким образом ускорить сходимость.

n	p_n	q_n
0	0,5	−0,26437542
1	0,23529412	
2	0,06225681	
3	0,01562119	
4	0,00390619	
5	0,00097656	

7. Последовательность $\{p_n\}$, генерируемая итерацией неподвижной точки, начинается с $p_0 = 2{,}5$, использует функцию $g(x) = (6 + x)^{1/2}$ и сходится линейно к $p = 3$. Воспользуйтесь формулой Эйткена (4), чтобы найти q_1, q_2 и q_3 и таким образом ускорить сходимость.

8. Последовательность $\{p_n\}$ генерируется итерацией неподвижной точки, начиная с точки $p_0 = 3{,}14$, и, используя функцию $g(x) = \ln(x) + 2$, сходится линейно к $p \approx 3{,}1419322$. Воспользуйтесь формулой (4) Эйткена, чтобы найти q_1, q_2 и q_3 и таким образом ускорить сходимость.

9. Для уравнения $\cos(x) - 1 = 0$ функция метода Ньютона–Рафсона равна $g(x) = x - (1 - \cos(x))/\sin(x) = x - \tan(x/2)$. Воспользуйтесь алгоритмом Стеффенсена с функцией $g(x)$, начните с $p_0 = 0{,}5$ и найдите сначала p_1, p_2 и p_3, а затем — p_4, p_5 и p_6.

10. *Сходимость рядов.* Метод Эйткена можно использовать для ускорения сходимости рядов. Если n-я частичная сумма ряда равна

$$S_n = \sum_{k=1}^{n} A_k,$$

покажите, что ряд, полученный с помощью метода Эйткена, имеет вид

$$T_n = S_n + \frac{A_{n+1}^2}{A_{n+1} - A_{n+2}}.$$

В упр. 11–14 для ускорения сходимости ряда применяются метод Эйткена и результаты упр. 10.

11. $S_n = \sum_{k=1}^{n} (0{,}99)^k$

12. $S_n = \sum_{k=1}^{n} \frac{1}{4^k + 4^{-k}}$

13. $S_n = \sum_{k=1}^{n} \frac{k}{2^{k-1}}$

14. $S_n = \sum_{k=1}^{n} \frac{1}{2^k k}$

15. Воспользуйтесь методом Мюллера, чтобы найти корень уравнения $f(x) = x^3 - x - 2$. Начните с $p_0 = 1{,}0$, $p_1 = 1{,}2$ и $p_2 = 1{,}4$ и найдите p_3, p_4 и p_5.

16. Воспользуйтесь методом Мюллера, чтобы найти корень уравнения $f(x) = 4x^2 - e^x$. Начните с $p_0 = 4{,}0$, $p_1 = 4{,}1$ и $p_2 = 4{,}2$ и найдите p_3, p_4 и p_5.

17. Пусть $\{p_n\}$ и $\{q_n\}$ — любые две последовательности действительных чисел. Покажите, что

(a) $\Delta(p_n + q_n) = \Delta p_n + \Delta q_n$

(b) $\Delta(p_n q_n) = p_{n+1}\Delta q_n + q_n \Delta p_n$

18. Начните с формулы (8), прибавьте члены p_{n+2} и $-p_{n+2}$ к правой части и покажите, что эквивалентная формула имеет вид

$$p \approx p_{n+2} - \frac{(p_{n+2} - p_{n+1})^2}{p_{n+2} - 2p_{n+1} + p_n} = q_n.$$

19. Предположим, что ошибка итерационного процесса удовлетворяет соотношению $E_{n+1} = KE_n$ для некоторой константы K и $|K| < 1$.

(a) Найдите выражение для E_n, в которую входят E_0, K и n.

(b) Найдите такое наименьшее целое число N, чтобы $|E_N| < 10^{-8}$.

Алгоритмы и программы

1. Воспользуйтесь методом Стеффенсена с начальным приближением $p_0 = 0{,}5$, чтобы найти приближение к нулю функции $f(x) = x - \sin(x)$ с точностью до 10 десятичных знаков.
2. Воспользуйтесь методом Стеффенсена с начальным приближением $p_0 = 0{,}5$, чтобы найти приближение к нулю функции $f(x) = \sin(x^3)$, ближайшее к 0,5, с точностью до 10 десятичных знаков.
3. Воспользуйтесь методом Мюллера с начальными приближениями $p_0 = 1{,}5$; $p_1 = 1{,}4$ и $p_2 = 1{,}3$, чтобы найти нуль функции $f(x) = 1 + 2x - \tan(x)$ с точностью до 12 десятичных знаков.
4. В программе 2.8 (метод Мюллера) матрица $\boldsymbol{P}$ размера 1×3 инициализирована с p_0, p_1 и p_2. Затем в конце цикла одно из значений p_0, p_1 и p_2 заменяется новым приближением к нулю. Этот процесс продолжается до тех пор, пока не выполнится критерий останова, скажем, при $k = K$. Преобразуйте программу 2.8 так, чтобы в дополнение к p и *err* матрица $\boldsymbol{Q}$ размера $(K+1) \times 3$ получалась такой, чтобы ее первая строка включала матрицу $\boldsymbol{P}$ размера 1×3 с начальными приближениями к нулю и k-я строка матрицы $\boldsymbol{Q}$ содержала k-й набор трех начальных приближений к нулю.

   Используйте это преобразование программы 2.8 с начальными приближениями $p_0 = 2{,}4$; $p_1 = 2{,}3$ и $p_2 = 2{,}2$, чтобы найти нули функции $f(x) = 3\cos(x) + 2\sin(x)$ с точностью до 8 десятичных знаков.

3

Решение систем линейных уравнений $AX = B$

На рис. 3.1 изображены три плоскости, которые ограничивают тело в первом октанте. Предположим, что уравнения этих плоскостей записываются так:

$$5x + y + z = 5,$$
$$x + 4y + z = 4,$$
$$x + y + 3z = 3.$$

Чему равны координаты точки пересечения трех плоскостей? Для решения системы линейных уравнений использовался метод приведения к одному неизвестному (метод Гаусса):

$$x = 0{,}76, \quad y = 0{,}68 \quad \text{и} \quad z = 0{,}52.$$

В этой главе будут рассмотрены численные методы решения систем линейных уравнений.

3.1. Введение в теорию векторов и матриц

Действительный N-мерный вектор $\boldsymbol{X}$ — это упорядоченный набор из N действительных чисел, который обычно в координатной форме записывается как

(1) $$\boldsymbol{X} = (x_1, x_2, \ldots, x_N).$$

Числа x_1, x_2, ... и x_N называются ***компонентами*** $\boldsymbol{X}$. Множество всех N-мерных векторов называется *N**мерным пространством***. Когда вектор используется

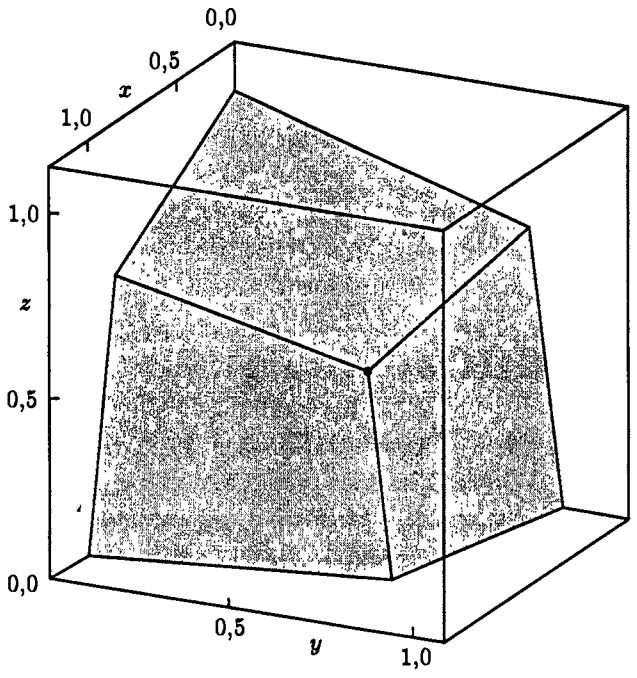


Рис. 3.1. Пересечение трех плоскостей

для обозначения точки или положения в пространстве, его называют ***радиус-вектором***. Когда вектор используется для обозначения перемещения между двумя точками в пространстве, его называют ***вектором переноса***.

Пусть $\boldsymbol{Y} = (y_1, y_2, \ldots, y_N)$ — произвольный вектор. Говорят, что вектор $\boldsymbol{X}$ равен вектору $\boldsymbol{Y}$ тогда и только тогда, когда каждая координата одного вектора равна такой же координате другого, т. е.

(2) $\boldsymbol{X} = \boldsymbol{Y}$ тогда и только тогда, когда $x_j = y_j$ для $j = 1, 2, \ldots, N$.

Сумма векторов $\boldsymbol{X}$ и $\boldsymbol{Y}$ вычисляется покомпонентно согласно определению

$$\boldsymbol{X} + \boldsymbol{Y} = (x_1 + y_1, x_2 + y_2, \ldots, x_N + y_N). \tag{3}$$

Отрицание вектора $\boldsymbol{X}$ получаем, заменяя каждую координату ее отрицанием:

$$-\boldsymbol{X} = (-x_1, -x_2, \ldots, -x_N). \tag{4}$$

Разность $\boldsymbol{Y} - \boldsymbol{X}$ является вектором, координаты которого равны разностям координат соответствующих векторов:

$$\boldsymbol{Y} - \boldsymbol{X} = (y_1 - x_1, y_2 - x_2, \ldots, y_N - x_N). \tag{5}$$

Для векторов в N-мерном пространстве выполняется алгебраическое свойство

$$\boldsymbol{Y} - \boldsymbol{X} = \boldsymbol{Y} + (-\boldsymbol{X}). \tag{6}$$

Если c — действительное число (скаляр), то определим ***умножение на скаляр*** $c\boldsymbol{X}$ следующим образом:

$$c\boldsymbol{X} = (cx_1, cx_2, \ldots, cx_N). \tag{7}$$

Если c и d — скаляры, то взвешенную сумму $c\boldsymbol{X} + d\boldsymbol{Y}$ называют ***линейной комбинацией*** векторов $\boldsymbol{X}$ и $\boldsymbol{Y}$ и записывают как

$$c\boldsymbol{X} + d\boldsymbol{Y} = (cx_1 + dy_1, cx_2 + dy_2, \ldots, cx_N + dy_N). \tag{8}$$

Скалярное произведение двух векторов $\boldsymbol{X}$ и $\boldsymbol{Y}$ — это скалярная величина (действительное число), определенная выражением

$$\boldsymbol{X} \cdot \boldsymbol{Y} = x_1y_1 + x_2y_2 + \cdots + x_Ny_N. \tag{9}$$

Норма (или ***длина***) вектора $\boldsymbol{X}$ определяется как

$$\|\boldsymbol{X}\| = (x_1^2 + x_2^2 + \cdots + x_N^2)^{1/2}. \tag{10}$$

Выражение (10) еще называют ***нормой Евклида*** (или ***длиной***) вектора $\boldsymbol{X}$.

Умножение на скаляр $c\boldsymbol{X}$ удлиняет вектор $\boldsymbol{X}$, когда $|c| > 1$, и сокращает, когда $|c| < 1$. Это можно показать, используя формулу (10):

$$\begin{aligned} \|c\boldsymbol{X}\| &= (c^2x_1^2 + c^2x_2^2 + \cdots + c^2x_N^2)^{1/2} = \\ &= |c|(x_1^2 + x_2^2 + \cdots + x_N^2)^{1/2} = |c|\|\boldsymbol{X}\|. \end{aligned} \tag{11}$$

Существует важное соотношение между скалярным произведением и нормой вектора. Если обе части равенства (10) возведем в квадрат и используем равенство (9), заменив $\boldsymbol{Y}$ на $\boldsymbol{X}$, то получим

$$\|\boldsymbol{X}\|^2 = x_1^2 + x_2^2 + \cdots + x_N^2 = \boldsymbol{X} \cdot \boldsymbol{X}. \tag{12}$$

Если $\boldsymbol{X}$ и $\boldsymbol{Y}$ — радиус-векторы, которые определяют место расположения двух точек $(x_1, x_2, \ldots, x_N)$ и $(y_1, y_2, \ldots, y_N)$ в N-мерном пространстве, то ***вектор переноса*** от $\boldsymbol{X}$ к $\boldsymbol{Y}$ задается разностью

$$\boldsymbol{Y} - \boldsymbol{X} \quad \text{(перенос из положения } \boldsymbol{X} \text{ в положение } \boldsymbol{Y}). \tag{13}$$

Заметим, что если частица выходит из положения $\boldsymbol{X}$ и двигается по вектору $\boldsymbol{Y} - \boldsymbol{X}$, то ее новой позицией будет $\boldsymbol{Y}$. Это можно получить с помощью следующей суммы векторов:

$$\boldsymbol{Y} = \boldsymbol{X} + (\boldsymbol{Y} - \boldsymbol{X}). \tag{14}$$

Используя равенства (10) и (13), запишем формулу для расстояния между двумя точками в N-мерном пространстве:

$$\|\boldsymbol{Y} - \boldsymbol{X}\| = \left((y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_N - x_N)^2\right)^{1/2}. \tag{15}$$

Когда расстояние между точками вычисляется согласно формуле (15), говорят, что точки лежат в ***N-мерном Евклидовом пространстве.***

Пример 3.1. Пусть $\boldsymbol{X} = (2, -3, 5, -1)$ и $\boldsymbol{Y} = (6, 1, 2, -4)$. Проиллюстрируем для векторов в 4-мерном пространстве упомянутые выше понятия.

Сумма	$\boldsymbol{X} + \boldsymbol{Y} = (8, -2, 7, -5)$
Разность	$\boldsymbol{X} - \boldsymbol{Y} = (-4, -4, 3, 3)$
Умножение на скаляр	$3\boldsymbol{X} = (6, -9, 15, -3)$
Длина	$\|\boldsymbol{X}\| = (4 + 9 + 25 + 1)^{1/2} = 39^{1/2}$
Скалярное произведение	$\boldsymbol{X} \cdot \boldsymbol{Y} = 12 - 3 + 10 + 4 = 23$
Перенос из $\boldsymbol{X}$ к $\boldsymbol{Y}$	$\boldsymbol{Y} - \boldsymbol{X} = (4, 4, -3, -3)$
Расстояние от $\boldsymbol{X}$ до $\boldsymbol{Y}$	$\|\boldsymbol{Y} - \boldsymbol{X}\| = (16 + 16 + 9 + 9)^{1/2} = 50^{1/2}$ ■

Иногда векторы записывают в виде столбцов, а не строк, например

$$\boldsymbol{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \text{и} \quad \boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}. \tag{16}$$

Тогда линейная комбинация $c\boldsymbol{X} + d\boldsymbol{Y}$ имеет вид

$$c\boldsymbol{X} + d\boldsymbol{Y} = \begin{bmatrix} cx_1 + dy_1 \\ cx_2 + dy_2 \\ \vdots \\ cx_N + dy_N \end{bmatrix}. \tag{17}$$

Выбрав соответственно c и d в равенстве (17), можно получить сумму $1\boldsymbol{X} + 1\boldsymbol{Y}$, разность $1\boldsymbol{X} - 1\boldsymbol{Y}$ и умножение на скаляр $c\boldsymbol{X} + 0\boldsymbol{Y}$. В дальнейшем будем использовать верхний индекс "$'$" для обозначения транспонирования, чтобы обозначить вектор-строку, который преобразуется в вектор-столбец и наоборот:

$$(x_1, x_2, \ldots, x_N)' = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \text{и} \quad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}' = (x_1, x_2, \ldots, x_N). \tag{18}$$

Множество векторов имеет нулевой элемент $\mathbf{0}$, который определен как

$$\mathbf{0} = (0, 0, \ldots, 0). \tag{19}$$

Теорема 3.1 (векторная алгебра). Предположим, что $\boldsymbol{X}$, $\boldsymbol{Y}$ и $\boldsymbol{Z}$ — N-мерные векторы и a и b — скаляры (действительные числа). Тогда выполняются следующие свойства сложения векторов и умножения вектора на скаляр.

(20)	$\boldsymbol{Y} + \boldsymbol{X} = \boldsymbol{X} + \boldsymbol{Y}$	Свойство коммутативности
(21)	$\mathbf{0} + \boldsymbol{X} = \boldsymbol{X} + \mathbf{0}$	Аддитивное тождество
(22)	$\boldsymbol{X} - \boldsymbol{X} = \boldsymbol{X} + (-\boldsymbol{X}) = \mathbf{0}$	Аддитивное обращение
(23)	$(\boldsymbol{X} + \boldsymbol{Y}) + \boldsymbol{Z} = \boldsymbol{X} + (\boldsymbol{Y} + \boldsymbol{Z})$	Свойство ассоциативноси
(24)	$(a + b)\boldsymbol{X} = a\boldsymbol{X} + b\boldsymbol{X}$	Свойство дистрибутивности для скаляров
(25)	$a(\boldsymbol{X} + \boldsymbol{Y}) = a\boldsymbol{X} + a\boldsymbol{Y}$	Свойство дистрибутивности для вектора
(26)	$a(b\boldsymbol{X}) = (ab)\boldsymbol{X}$	Свойство ассоциативности для скаляров

Матрицы и двухмерные массивы

Матрица — это прямоугольный массив чисел, который упорядочен по строкам и столбцам. Матрица, состоящая из M строк и N столбцов, называется матрицей размера $M \times N$ (читается "M на N"). Обозначается матрица большой буквой $\boldsymbol{A}$; буквой a_{ij} с нижними индексами обозначается одно из чисел, образующих матрицу. Запишем

$$\boldsymbol{A} = [a_{ij}]_{M \times N} \quad \text{для } 1 \le i \le M, 1 \le j \le N, \tag{27}$$

где a_{ij} — число на месте (i, j) (т. е. число, которое хранится в i-й строке и j-м столбце матрицы). Мы обращаемся к a_{ij}, как к элементу, занимающему место (i, j). В развернутом виде матрица записывается так:

$$\text{строка } i \rightarrow \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{iN} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{Mj} & \cdots & a_{MN} \end{bmatrix} = A. \tag{28}$$

$$\text{столбец } j \uparrow$$

Строки матрицы $\boldsymbol{A}$ размера $M \times N$ — это N-мерные векторы:

$$\boldsymbol{V}_i = (a_{i1}, a_{i2}, \ldots, a_{iN}) \quad \text{для } i = 1,\ 2,\ \ldots,\ M. \tag{29}$$

Вектор-строку в (29) также можно представить в виде матрицы размера $1 \times N$, т. е. разделить матрицу $\boldsymbol{A}$ размера $M \times N$ на M частей (подматриц), каждая из которых является матрицей размера $1 \times N$.

В этом случае можно представить матрицу $\boldsymbol{A}$ в виде матрицы размера $M \times 1$, содержащей $1 \times N$ матриц в виде $\boldsymbol{V}_i$, т. е.

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{V}_1 \\ \boldsymbol{V}_2 \\ \vdots \\ \boldsymbol{V}_i \\ \vdots \\ \boldsymbol{V}_M \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 & \cdots & \boldsymbol{V}_i & \cdots & \boldsymbol{V}_M \end{bmatrix}'. \tag{30}$$

Аналогично столбцы матрицы $\boldsymbol{A}$ размера $M \times N$ являются матрицами размера $M \times 1$:

$$\boldsymbol{C}_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{M1} \end{bmatrix}, \quad \ldots, \quad \boldsymbol{C}_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{ij} \\ \vdots \\ a_{Mj} \end{bmatrix}, \quad \ldots, \quad \boldsymbol{C}_N = \begin{bmatrix} a_{1N} \\ a_{2N} \\ \vdots \\ a_{iN} \\ \vdots \\ a_{MN} \end{bmatrix}. \tag{31}$$

В этом случае можно представить матрицу $\boldsymbol{A}$ в виде матрицы размера $1 \times N$, содержащую $M \times 1$ матриц в виде столбцов $\boldsymbol{C}_j$:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{C}_1 & \boldsymbol{C}_2 & \cdots & \boldsymbol{C}_j & \cdots & \boldsymbol{C}_N \end{bmatrix}. \tag{32}$$

Пример 3.2. Идентифицируем строки и столбцы матрицы, ассоциированной с матрицей размера 4×3:

$$\boldsymbol{A} = \begin{bmatrix} -2 & 4 & 9 \\ 5 & -7 & 1 \\ 0 & -3 & 8 \\ -4 & 6 & -5 \end{bmatrix}.$$

Четыре матрицы строки равны $\boldsymbol{V}_1 = \begin{bmatrix} -2 & 4 & 9 \end{bmatrix}$, $\boldsymbol{V}_2 = \begin{bmatrix} 5 & -7 & 1 \end{bmatrix}$, $\boldsymbol{V}_3 = \begin{bmatrix} 0 & -3 & 8 \end{bmatrix}$ и $\boldsymbol{V}_4 = \begin{bmatrix} -4 & 6 & -5 \end{bmatrix}$. Три матрицы столбца равны

$$\boldsymbol{C}_1 = \begin{bmatrix} -2 \\ 5 \\ 0 \\ -4 \end{bmatrix}, \quad \boldsymbol{C}_2 = \begin{bmatrix} 4 \\ -7 \\ -3 \\ 6 \end{bmatrix}, \quad \text{и} \quad \boldsymbol{C}_3 = \begin{bmatrix} 9 \\ 1 \\ 8 \\ -5 \end{bmatrix}.$$

Отметим, как матрицу $\boldsymbol{A}$ можно представить с помощью этих матриц:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{V}_1 \\ \boldsymbol{V}_2 \\ \boldsymbol{V}_3 \\ \boldsymbol{V}_4 \end{bmatrix} = [\boldsymbol{C}_1 \quad \boldsymbol{C}_2 \quad \boldsymbol{C}_3]. \qquad \blacksquare$$

Пусть $\boldsymbol{A} = [a_{ij}]_{M\times N}$ и $\boldsymbol{B} = [b_{ij}]_{M\times N}$ — две матрицы одинакового размера. Говорят, что две матрицы $\boldsymbol{A}$ и $\boldsymbol{B}$ равны тогда и только тогда, когда равны все их соответствующие элементы, т. е.

(33) $\boldsymbol{A} = \boldsymbol{B}$ только тогда, когда $a_{ij} = b_{ij}$ для $1 \le i \le M,\ 1 \le j \le N.$

Сумма двух матриц $\boldsymbol{A}$ и $\boldsymbol{B}$ размера $M \times N$ вычисляется поэлементно согласно определению

(34) $$\boldsymbol{A} + \boldsymbol{B} = [a_{ij} + b_{ij}]_{M\times N} \quad \text{для } 1 \le i \le M,\ 1 \le j \le N.$$

Отрицание матрицы $\boldsymbol{A}$ получается в результате замены каждого элемента его отрицанием:

(35) $$-\boldsymbol{A} = [-a_{ij}]_{M\times N} \quad \text{для } 1 \le i \le M,\ 1 \le j \le N.$$

Разность матриц $\boldsymbol{A} - \boldsymbol{B}$ — это матрица, элементы которой равны разности соответствующих элементов матриц:

(36) $$\boldsymbol{A} - \boldsymbol{B} = [a_{ij} - b_{ij}]_{M\times N} \quad \text{для } 1 \le i \le M,\ 1 \le j \le N.$$

Если c — действительное число (скаляр), то можно определить умножение матрицы на скаляр $c\boldsymbol{A}$ следующим образом:

(37) $$c\boldsymbol{A} = [ca_{ij}]_{M\times N} \quad \text{для } 1 \le i \le M,\ 1 \le j \le N.$$

Если p и q — скаляры, то взвешенная сумма $p\boldsymbol{A} + q\boldsymbol{B}$ называется линейной комбинацией матриц $\boldsymbol{A}$ и $\boldsymbol{B}$ и записывается в форме

(38) $$p\boldsymbol{A} + q\boldsymbol{B} = [pa_{ij} + qb_{ij}]_{M\times N} \quad \text{для } 1 \le i \le M,\ 1 \le j \le N.$$

Нулевая матрица размера $M \times N$ состоит из всех нулей:

(39) $$\boldsymbol{0} = [0]_{M\times N}.$$

Пример 3.3. Найдем произведение на скаляр $2\boldsymbol{A}$ и $3\boldsymbol{B}$ и линейную комбинацию $2\boldsymbol{A} - 3\boldsymbol{B}$ матриц

$$\boldsymbol{A} = \begin{bmatrix} -1 & 2 \\ 7 & 5 \\ 3 & -4 \end{bmatrix} \quad \text{и} \quad \boldsymbol{B} = \begin{bmatrix} -2 & 3 \\ 1 & -4 \\ -9 & 7 \end{bmatrix}.$$

Используем формулу (37) и получим

$$2\boldsymbol{A} = \begin{bmatrix} -2 & 4 \\ 14 & 10 \\ 6 & -8 \end{bmatrix} \quad \text{и} \quad 3\boldsymbol{B} = \begin{bmatrix} -6 & 9 \\ 3 & -12 \\ -27 & 21 \end{bmatrix}.$$

Найдем линейную комбинацию $2\boldsymbol{A} - 3\boldsymbol{B}$:

$$2\boldsymbol{A} - 3\boldsymbol{B} = \begin{bmatrix} -2+6 & 4-9 \\ 14-3 & 10+12 \\ 6+27 & -8-21 \end{bmatrix} = \begin{bmatrix} 4 & -5 \\ 11 & 22 \\ 33 & -29 \end{bmatrix}. \quad \blacksquare$$

Теорема 3.2 (сложение матриц). Предположим, что $\boldsymbol{A}$, $\boldsymbol{B}$ и $\boldsymbol{C}$ — матрицы размера $M \times N$ и p и q — скаляры. Тогда операции сложения матриц и умножения матрицы на скаляр имеют следующие свойства.

(40)	$\boldsymbol{B} + \boldsymbol{A} = \boldsymbol{A} + \boldsymbol{B}$	Свойство коммутативности
(41)	$\mathbf{0} + \boldsymbol{A} = \boldsymbol{A} + \mathbf{0}$	Аддитивное тождество
(42)	$\boldsymbol{A} - \boldsymbol{A} = \boldsymbol{A} + (-\boldsymbol{A}) = \mathbf{0}$	Аддитивное обращение
(43)	$(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C})$	Свойство ассоциативности
(44)	$(p+q)\boldsymbol{A} = p\boldsymbol{A} + q\boldsymbol{A}$	Свойство дистрибутивности для скаляров
(45)	$p(\boldsymbol{A} + \boldsymbol{B}) = p\boldsymbol{A} + p\boldsymbol{B}$	Свойство дистрибутивности для матриц
(46)	$p(q\boldsymbol{A}) = (pq)\boldsymbol{A}$	Свойство ассоциативности для скаляров

Упражнения к разделу 3.1

Следующие упражнения читатель может выполнять вручную и с помощью MATLAB.

1. Заданы векторы $\boldsymbol{X}$ и $\boldsymbol{Y}$. Найдите (a) $\boldsymbol{X} + \boldsymbol{Y}$, (b) $\boldsymbol{X} - \boldsymbol{Y}$, (c) $3\boldsymbol{X}$, (d) $\|\boldsymbol{X}\|$, (e) $7\boldsymbol{Y} - 4\boldsymbol{X}$, (f) $\boldsymbol{X} \cdot \boldsymbol{Y}$, и (g) $\|7\boldsymbol{Y} - 4\boldsymbol{X}\|$.
   (i) $\boldsymbol{X} = (3, -4)$ и $\boldsymbol{Y} = (-2, 8)$
   (ii) $\boldsymbol{X} = (-6, 3, 2)$ и $\boldsymbol{Y} = (-8, 5, 1)$
   (iii) $\boldsymbol{X} = (4, -8, 1)$ и $\boldsymbol{Y} = (1, -12, -11)$
   (iv) $\boldsymbol{X} = (1, -2, 4, 2)$ и $\boldsymbol{Y} = (3, -5, -4, 0)$

2. Используя закон косинуса, можно показать, что угол θ между векторами $\boldsymbol{X}$ и $\boldsymbol{Y}$ задается отношением

$$\cos(\theta) = \frac{\boldsymbol{X} \cdot \boldsymbol{Y}}{\|\boldsymbol{X}\| \, \|\boldsymbol{Y}\|}.$$

   Найдите угол (в радианах) между следующими векторами.
   (a) $\boldsymbol{X} = (-6, 3, 2)$ и $\boldsymbol{Y} = (2, -2, 1)$
   (b) $\boldsymbol{X} = (4, -8, 1)$ и $\boldsymbol{Y} = (3, 4, 12)$

3. Говорят, что два вектора $\boldsymbol{X}$ и $\boldsymbol{Y}$ ортогональны (перпендикулярны), если угол между ними равен $\pi/2$.

   (a) Докажите, что $\boldsymbol{X}$ и $\boldsymbol{Y}$ ортогональны тогда и только тогда, когда $\boldsymbol{X} \cdot \boldsymbol{Y} = 0$.

   Используя п. (a), определите, ортогональны ли следующие векторы.

   (b) $\boldsymbol{X} = (-6, 4, 2)$ и $\boldsymbol{Y} = (6, 5, 8)$

   (c) $\boldsymbol{X} = (-4, 8, 3)$ и $\boldsymbol{Y} = (2, 5, 16)$

   (d) $\boldsymbol{X} = (-5, 7, 2)$ и $\boldsymbol{Y} = (4, 1, 6)$

   (e) Найдите два различных вектора, ортогональных вектору $\boldsymbol{X} = (1, 2, -5)$.

4. Найдите (a) $\boldsymbol{A} + \boldsymbol{B}$, (b) $\boldsymbol{A} - \boldsymbol{B}$ и (c) $3\boldsymbol{A} - 2\boldsymbol{B}$ для матриц

$$\boldsymbol{A} = \begin{bmatrix} -1 & 9 & 4 \\ 2 & -3 & -6 \\ 0 & 5 & 7 \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} -4 & 9 & 2 \\ 3 & -5 & 7 \\ 8 & 1 & -6 \end{bmatrix}.$$

5. ***Транспонированная*** матрица $\boldsymbol{A}$ размера $M \times N$, обозначаемая $\boldsymbol{A}'$, — это матрица размера $N \times M$, которую получают из $\boldsymbol{A}$ посредством обращения строк матрицы $\boldsymbol{A}$ в столбцы $\boldsymbol{A}'$, т. е. если $\boldsymbol{A} = [a_{ij}]_{M \times N}$ и $\boldsymbol{A}' = [b_{ij}]_{N \times M}$, то элементы удовлетворяют отношению

$$b_{ji} = a_{ij} \quad \text{for} \quad 1 \le i \le M, 1 \le j \le N.$$

   Найдите транспонированные матрицы для следующих матриц.

   (a) $\begin{bmatrix} -2 & 5 & 12 \\ 1 & 4 & -1 \\ 7 & 0 & 6 \\ 11 & -3 & 8 \end{bmatrix}$ (b) $\begin{bmatrix} 4 & 9 & 2 \\ 3 & 5 & 7 \\ 8 & 1 & 6 \end{bmatrix}'$

6. Говорят, что квадратная матрица $\boldsymbol{A}$ размера $N \times N$ симметрична, если $\boldsymbol{A} = \boldsymbol{A}'$ (см. определение $\boldsymbol{A}'$ в упр. 5). Определите, будут ли симметричными следующие квадратные матрицы.

   (a) $\begin{bmatrix} 1 & -7 & 4 \\ -7 & 2 & 0 \\ 4 & 0 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 4 & -7 & 1 \\ 0 & 2 & -7 \\ 3 & 0 & 4 \end{bmatrix}$

   (c) $\boldsymbol{A} = [a_{ij}]_{N \times N}$, где $a_{ij} = \begin{cases} ij, & i = j \\ i - ij + j, & i \ne j \end{cases}$

   (d) $\boldsymbol{A} = [a_{ij}]_{N \times N}$, где $a_{ij} = \begin{cases} \cos(ij), & i = j \\ i - ij - j, & i \ne j \end{cases}$

7. Докажите утверждения (20), (24) и (25) теоремы 3.1.

3.2. Свойства векторов и матриц

Линейная комбинация величин $x_1, x_2, \ldots, x_N$ — это сумма

$$a_1x_1 + a_2x_2 + \cdots + a_Nx_N, \tag{1}$$

где a_k — коэффициенты при x_k для $k = 1, 2, \ldots, N$.

Линейное уравнение относительно $x_1, x_2, \ldots, x_N$ получаем, приравнивая линейную комбинацию (1) к значению b, т. е.

$$a_1x_1 + a_2x_2 + \cdots + a_Nx_N = b. \tag{2}$$

Систему линейных уравнений, которая часто возникает, когда заданы M уравнений от N неизвестных, можно записать в виде

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1N}x_N = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2N}x_N = b_2 \\ \quad\vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\quad \vdots \\ a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kN}x_N = b_k \\ \quad\vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\quad \vdots \\ a_{M1}x_1 + a_{M2}x_2 + \cdots + a_{MN}x_N = b_M. \end{array} \tag{3}$$

Для того чтобы различать коэффициенты каждого уравнения, необходимо использовать два нижних индекса (k, j). Первый индекс определяет уравнение k, которому принадлежат коэффициенты, и второй индекс определяет переменную x_j, перед которой стоит коэффициент.

Решением системы уравнений (3) является набор численных значений x_1, x_2, ..., x_N, которые одновременно удовлетворяют всем уравнениям в (3). Следовательно, решение может иметь вид N-мерного вектора:

$$X = (x_1, x_2, \ldots, x_N). \tag{4}$$

Пример 3.4. Бетон (строительный материал) представляет собой смесь портланд-цемента, песка и гравия. Распределитель имеет в распоряжении три замеса бетона для подрядчика. Первый замес бетона содержит цемент, песок и гравий, смешанные в пропорции $1/8, 3/8, 4/8$; второй замес бетона имеет пропорции $2/10, 5/10, 3/10$ и третий — $2/5, 3/5, 0/5$.

Пусть x_1, x_2, и x_3 определяют количество бетона в каждом замесе (в кубических ярдах, 1 ярд = 3 фута = 91,44 см), общий объем бетона равен 10 кубическим ярдам. Также предположим, что смесь содержит $b_1 = 2{,}3$; $b_2 = 4{,}8$ и $b_3 = 2{,}9$ кубических ярдов портланд-цемента, песка и гравия соответственно. Тогда система

линейных уравнений для ингредиентов имеет следующий вид:

$$\begin{aligned} 0{,}125x_1 + 0{,}200x_2 + 0{,}400x_3 &= 2{,}3 \quad \text{(цемент)} \\ 0{,}375x_1 + 0{,}500x_2 + 0{,}600x_3 &= 4{,}8 \quad \text{(песок)} \\ 0{,}500x_1 + 0{,}300x_2 + 0{,}000x_3 &= 2{,}9 \quad \text{(гравий)} \end{aligned} \tag{5}$$

Решениями системы линейных уравнений (5) являются $x_1 = 4, x_2 = 3$ и $x_3 = 3$, что можно проверить, подставив эти значения в уравнения:

$$\begin{aligned} (0{,}125)(4) + (0{,}200)(3) + (0{,}400)(3) &= 2{,}3; \\ (0{,}375)(4) + (0{,}500)(3) + (0{,}600)(3) &= 4{,}8; \\ (0{,}500)(4) + (0{,}300)(3) + (0{,}000)(3) &= 2{,}9. \end{aligned}$$

■

Умножение матриц

Определение 3.1. Если $\boldsymbol{A} = [a_{ik}]_{M\times N}$ и $\boldsymbol{B} = [b_{kj}]_{N\times P}$ — такие две матрицы, что $\boldsymbol{A}$ состоит из стольких же столбцов, из скольких строк состоит матрица $\boldsymbol{B}$, произведение матриц $\boldsymbol{AB}$ — это матрица $\boldsymbol{C}$ размера $M \times P$:

$$\boldsymbol{AB} = \boldsymbol{C} = [c_{ij}]_{M\times P}, \tag{6}$$

где элемент c_{ij} матрицы $\boldsymbol{C}$ равен скалярному произведению i-й строки матрицы $\boldsymbol{A}$ и j-го столбца $\boldsymbol{B}$:

$$c_{ij} = \sum_{k=1}^{N} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iN}b_{Nj} \tag{7}$$

для $i = 1, 2, \ldots, M$ и $j = 1, 2, \ldots, P$. ▲

Пример 3.5. Найдем произведение $\boldsymbol{C} = \boldsymbol{AB}$ для следующих матриц и объясним, почему матрица $\boldsymbol{BA}$ не определена:

$$\boldsymbol{A} = \begin{bmatrix} 2 & 3 \\ -1 & 4 \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} 5 & -2 & 1 \\ 3 & 8 & -6 \end{bmatrix}.$$

Матрица $\boldsymbol{A}$ состоит из двух столбцов, матрица $\boldsymbol{B}$ — из двух строк, поэтому произведение матриц $\boldsymbol{AB}$ определено. Произведение матриц размера 2×2 и 2×3 равно матрице размера 2×3. Вычисления показывают, что

$$\begin{aligned} \boldsymbol{AB} &= \begin{bmatrix} 2 & 3 \\ -1 & 4 \end{bmatrix} \begin{bmatrix} 5 & -2 & 1 \\ 3 & 8 & -6 \end{bmatrix} = \\ &= \begin{bmatrix} 10+9 & -4+24 & 2-18 \\ -5+12 & 2+32 & -1-24 \end{bmatrix} = \begin{bmatrix} 19 & 20 & -16 \\ 7 & 34 & -25 \end{bmatrix} = \boldsymbol{C}. \end{aligned}$$

Когда мы попытались получить произведение матриц $\boldsymbol{BA}$, то обнаружили, что размерности не совместимы по порядку, так как строки матрицы $\boldsymbol{B}$ — это трехмерные векторы, а столбцы матрицы $\boldsymbol{A}$ двухмерные векторы. Поэтому скалярное произведение j-й строки матрицы $\boldsymbol{B}$ и k-го столбца $\boldsymbol{A}$ не определено. ■

В случае, когда $\boldsymbol{AB} = \boldsymbol{BA}$, говорят, что матрицы $\boldsymbol{A}$ и $\boldsymbol{B}$ коммутативны. Очень часто, даже когда обе матрицы, $\boldsymbol{AB}$ и $\boldsymbol{BA}$, определены, эти произведения не обязательно совпадают.

Рассмотрим, как можно использовать матрицы для представления системы линейных уравнений. Систему линейных уравнений (3) можно записать как произведение матриц. Коэффициенты a_{kj} содержатся в матрице $\boldsymbol{A}$ размера $M \times N$ (называемой матрицей коэффициентов), а неизвестные x_j содержатся в матрице $\boldsymbol{X}$ размера $N \times 1$. Константы b_k содержатся в матрице $\boldsymbol{B}$ размера $M \times 1$. Общепринято использовать матрицы-столбцы и для $\boldsymbol{X}$, и для $\boldsymbol{B}$. Тогда запишем

$$\boldsymbol{AX} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kj} & \cdots & a_{kN} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{Mj} & \cdots & a_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_j \\ \vdots \\ b_M \end{bmatrix} = \boldsymbol{B}. \tag{8}$$

Умножение матриц $\boldsymbol{AX} = \boldsymbol{B}$ в (8) напоминает скалярное произведение для обычных векторов, потому что каждый элемент b_k в матрице $\boldsymbol{B}$ равен скалярному произведению строки k матрицы $\boldsymbol{A}$ и столбца матрицы $\boldsymbol{X}$.

Пример 3.6. Выразим систему линейных уравнений (5) из примера 3.4 в виде произведения матриц. Используем умножение матриц, чтобы проверить, что $[4 \quad 3 \quad 3]'$ — это решение (5):

$$\begin{bmatrix} 0{,}125 & 0{,}200 & 0{,}400 \\ 0{,}375 & 0{,}500 & 0{,}600 \\ 0{,}500 & 0{,}300 & 0{,}000 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2{,}3 \\ 4{,}8 \\ 2{,}9 \end{bmatrix}. \tag{9}$$

Чтобы убедиться, что $[4 \quad 3 \quad 3]'$ является решением (5), нужно показать, что $\boldsymbol{A}[4 \quad 3 \quad 3]' = [2{,}3 \quad 4{,}8 \quad 2{,}9]'$:

$$\begin{bmatrix} 0{,}125 & 0{,}200 & 0{,}400 \\ 0{,}375 & 0{,}500 & 0{,}600 \\ 0{,}500 & 0{,}300 & 0{,}000 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 0{,}5 + 0{,}6 + 1{,}2 \\ 1{,}5 + 1{,}5 + 1{,}8 \\ 2{,}0 + 0{,}9 + 0{,}0 \end{bmatrix} = \begin{bmatrix} 2{,}3 \\ 4{,}8 \\ 2{,}9 \end{bmatrix}. \qquad \blacksquare$$

Некоторые специальные матрицы

Матрица размера $M \times N$, все элементы которой равны нулю, называется ***нулевой матрицей размера*** $M \times N$ и записывается как

$$\mathbf{0} = [0]_{M \times N}. \tag{10}$$

Когда размерность известна, для записи нулевой матрицы используется $\mathbf{0}$.

Единичной матрицей порядка N называется квадратная матрица, заданная в виде

$$\boldsymbol{I}_N = [\delta_{ij}]_{N\times N} \quad \text{где} \quad \delta_{ij} = \begin{cases} 1 & \text{где } i = j, \\ 0 & \text{где } i \neq j. \end{cases} \tag{11}$$

Как будет показано в следующем примере, при умножении эта матрица осуществляет тождественное преобразование.

Пример 3.7. Пусть $\boldsymbol{A}$ — матрица размера 2×3. Тогда $\boldsymbol{I_2 A} = \boldsymbol{A I_4} = \boldsymbol{A}$. Умножив матрицу $\boldsymbol{A}$ слева на матрицу $\boldsymbol{I_2}$, в результате получим

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} = \begin{bmatrix} a_{11}+0 & a_{12}+0 & a_{13}+0 \\ a_{21}+0 & a_{22}+0 & a_{23}+0 \end{bmatrix} = \boldsymbol{A}.$$

Умножаем матрицу $\boldsymbol{A}$ справа на матрицу $\boldsymbol{I_3}$ и получаем в результате

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11}+0+0 & 0+a_{12}+0 & 0+0+a_{13} \\ a_{21}+0+0 & 0+a_{22}+0 & 0+0+a_{23} \end{bmatrix} = \boldsymbol{A}. \quad \blacksquare$$

Некоторые свойства умножения матриц приведены в следующей теореме.

Теорема 3.3 (умножение матриц). Предположим, что c — скаляр и что $\boldsymbol{A}$, $\boldsymbol{B}$ и $\boldsymbol{C}$ такие матрицы, что определены указанные ниже их суммы и произведения. Тогда получаем следующее.

(12)	$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$	Ассоциативность умножения матриц
(13)	$\boldsymbol{IA} = \boldsymbol{AI} = \boldsymbol{A}$	Тождественная матрица
(14)	$\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$	Свойство левой дистрибутивности
(15)	$(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$	Свойство правой дистрибутивности
(16)	$c(\boldsymbol{AB}) = (c\boldsymbol{A})\boldsymbol{B} = \boldsymbol{A}(c\boldsymbol{B})$	Свойство ассоциативности скаляра

Обращение невырожденных матриц

Понятие "обращение" применяется к матрицам, но требует специального рассмотрения. Матрица $\boldsymbol{A}$ размера $N \times N$ называется ***невырожденной (несингулярной)*** или обратимой (неособенной), если существует такая матрица $\boldsymbol{B}$ размера $N \times N$, что

$$\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}. \tag{17}$$

Если такая матрица $\boldsymbol{B}$ не существует, говорят, что матрица $\boldsymbol{A}$ — ***вырожденная (сингулярная)***. Когда существует матрица $\boldsymbol{B}$ и равенство (17) выполняется, то обычно записывают $\boldsymbol{B} = \boldsymbol{A}^{-1}$ и используют хорошо известное отношение

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} \quad \text{если } \boldsymbol{A} \text{ не вырожденная.} \tag{18}$$

Легко показать, что можно найти самое большее одну матрицу $\boldsymbol{B}$, которая может удовлетворять отношению (17). Предположим, что матрица $\boldsymbol{C}$ есть также обращение матрицы $\boldsymbol{A}$ (т. е. $\boldsymbol{AC} = \boldsymbol{CA} = \boldsymbol{I}$). Тогда можно использовать свойства (12) и (13), чтобы получить

$$\boldsymbol{C} = \boldsymbol{IC} = (\boldsymbol{BA})\boldsymbol{C} = \boldsymbol{B}(\boldsymbol{AC}) = \boldsymbol{BI} = \boldsymbol{B}.$$

Определители

Определитель квадратной матрицы $\boldsymbol{A}$ — это скалярная величина (действительное число); обозначается как $\det(\boldsymbol{A})$ или $|\boldsymbol{A}|$. Если $\boldsymbol{A}$ — матрица размера $N \times N$

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix},$$

то обычно для определителя используется запись

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{vmatrix}.$$

Хотя по форме определитель, возможно, выглядит, как матрица, их свойства совершенно отличаются. Во-первых, определитель — это скалярная величина (действительное число). Во-вторых, определитель $\det(\boldsymbol{A})$, определение которого можно найти почти во всех книгах по линейной алгебре, нелегко поддается вычислению, когда $N > 3$. Сейчас рассмотрим, как использовать метод алгебраического дополнения для вычисления определителя. Для вычисления определителей высокого порядка можно использовать метод исключения неизвестных Гаусса, о чем упоминается в описании программы 3.3.

Если $\boldsymbol{A} = [a_{ij}]$ — матрица размера 1×1, то определитель равен $\det(\boldsymbol{A}) = a_{11}$. Если $\boldsymbol{A} = [a_{ij}]_{N \times N}$, где $N \geq 2$, то пусть $\boldsymbol{M}_{ij}$ — определитель подматрицы $\boldsymbol{A}$ размера $N-1 \times N-1$, полученной путем удаления i-й строки и j-го столбца в матрице $\boldsymbol{A}$. Говорят, что определитель $\boldsymbol{M}_{ij}$ называется ***минором*** a_{ij}. ***Алгебраическое дополнение*** $\boldsymbol{A}_{ij}$ к a_{ij} определяется как $\boldsymbol{A}_{ij} = (-1)^{i+j}\boldsymbol{M}_{ij}$. Тогда определитель матрицы $\boldsymbol{A}$ размера $N \times N$ задается формулой

$$\det(\boldsymbol{A}) = \sum_{j=1}^{N} a_{ij}A_{ij} \quad (i\text{-я строка разложения}) \tag{19}$$

или

$$\det(\boldsymbol{A}) = \sum_{i=1}^{N} a_{ij} A_{ij} \quad (j\text{-й столбец разложения}). \tag{20}$$

Применяя формулу (19) при $i = 1$ к матрице размера 2×2

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

видим, что $\det \boldsymbol{A} = a_{11}a_{22} - a_{12}a_{21}$. В следующем примере иллюстрируется применение формул (19) и (20) для рекуррентного вычисления определителя матрицы размера $N \times N$ путем вычисления определителей матриц размера 2×2.

Пример 3.8. Используем формулу (19) с $i = 1$ и формулу (20) с $j = 2$, чтобы вычислить определитель матрицы

$$\boldsymbol{A} = \begin{bmatrix} 2 & 3 & 8 \\ -4 & 5 & -1 \\ 7 & -6 & 9 \end{bmatrix}.$$

С помощью формулы (19) с $i = 1$ получим

$$\det(\boldsymbol{A}) = (2)\begin{vmatrix} 5 & -1 \\ -6 & 9 \end{vmatrix} - (3)\begin{vmatrix} -4 & -1 \\ 7 & 9 \end{vmatrix} + (8)\begin{vmatrix} -4 & 5 \\ 7 & -6 \end{vmatrix} =$$
$$= (2)(45 - 6) - (3)(-36 + 7) + (8)(24 - 35) = 77.$$

При использовании формулы (20) с $j = 2$ получим

$$\det(\boldsymbol{A}) = -(3)\begin{vmatrix} -4 & -1 \\ 7 & 9 \end{vmatrix} + (5)\begin{vmatrix} 2 & 8 \\ 7 & 9 \end{vmatrix} - (-6)\begin{vmatrix} 2 & 8 \\ -4 & -1 \end{vmatrix} = 77. \qquad \blacksquare$$

Следующая теорема дает достаточные условия существования и единственности решения системы линейных уравнений $\boldsymbol{AX} = \boldsymbol{B}$ для квадратных матриц.

Теорема 3.4. Предположим, что $\boldsymbol{A}$ — матрица размера $N \times N$. Следующие утверждения эквивалентны.

(21) Для любой заданной матрицы $\boldsymbol{B}$ размера $N \times 1$ система линейных уравнений $\boldsymbol{AX} = \boldsymbol{B}$ имеет единственное решение.

(22) Матрица $\boldsymbol{A}$ невырождена (т. е. существует $\boldsymbol{A}^{-1}$).

(23) Система уравнений $\boldsymbol{AX} = \boldsymbol{0}$ имеет единственное решение $\boldsymbol{X} = \boldsymbol{0}$.

(24) $\det(\boldsymbol{A}) \neq 0$.

Теоремы 3.3 и 3.4 связывают алгебру матриц с обычной алгеброй. Если справедливо утверждение (21), то утверждение (22) вместе со свойствами (12) и (13) приводит к следующему простому заключению:

(25) $\boldsymbol{AX} = \boldsymbol{B}$ влечет $\boldsymbol{A}^{-1}\boldsymbol{AX} = \boldsymbol{A}^{-1}\boldsymbol{B}$, которое влечет $\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{B}$.

Пример 3.9. Используем обращение матрицы

$$\boldsymbol{A}^{-1} = \frac{1}{5}\begin{bmatrix} 4 & -1 \\ -7 & 3 \end{bmatrix}$$

и рассуждения из (25), чтобы решить систему линейных уравнений $\boldsymbol{AX} = \boldsymbol{B}$:

$$\boldsymbol{AX} = \begin{bmatrix} 3 & 1 \\ 7 & 4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \boldsymbol{B}.$$

Используем (25) и получим

$$\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{B} = \frac{1}{5}\begin{bmatrix} 4 & -1 \\ -7 & 3 \end{bmatrix}\begin{bmatrix} 2 \\ 5 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 0{,}6 \\ 0{,}2 \end{bmatrix}. \quad \blacksquare$$

Примечание. Практически мы никогда не используем численные методы для обращения невырожденной матрицы или определителя квадратной матрицы. Эти понятия употребляются как теоретический "инструмент", чтобы установить существование и единственность решений, или как средство алгебраического выражения решения системы линейных уравнений (см. пример 3.9).

Вращение плоскости

Предположим, что $\boldsymbol{A}$ — матрица размера 3×3 и $\boldsymbol{U} = \begin{bmatrix} x & y & z \end{bmatrix}'$ — матрица размера 3×1, тогда их произведение $\boldsymbol{V} = \boldsymbol{AU}$ — матрица размера 3×1. Это пример линейного преобразования, которое находит применение в компьютерной графике. Матрица $\boldsymbol{U}$ эквивалентна радиус-вектору $\boldsymbol{U} = (x, y, z)$, который обозначает координаты точки в трехмерном пространстве. Рассмотрим три специально подобранные матрицы:

(26) $$\boldsymbol{R}_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{bmatrix},$$

(27) $$\boldsymbol{R}_y(\beta) = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{bmatrix},$$

(28) $$\boldsymbol{R}_z(\gamma) = \begin{bmatrix} \cos(\gamma) & -\sin(\gamma) & 0 \\ \sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Матрицы $\boldsymbol{R}_x(\alpha)$, $\boldsymbol{R}_y(\beta)$ и $\boldsymbol{R}_z(\gamma)$ используются для вращения точки вокруг осей x, y и z под углами α, β и γ соответственно. Обратные матрицы $\boldsymbol{R}_x(-\alpha)$, $\boldsymbol{R}_y(-\beta)$ и $\boldsymbol{R}_z(-\gamma)$ вращают плоскость вокруг осей x, y и z под углами $-\alpha, -\beta$ и $-\gamma$ соответственно. Следующий пример иллюстрирует этот случай, но дальнейшие исследования оставлены читателю.

Пример 3.10. Единичный куб расположен в первом октанте с одной из вершин в начале координат. Сначала поворачиваем куб на угол $\pi/4$ вокруг оси z, затем поворачиваем этот образ на угол $\pi/6$ вокруг оси y. Найдем образ всех восьми вершин куба.

Первый поворот задан отображением

$$\boldsymbol{V} = \boldsymbol{R}_z\left(\frac{\pi}{4}\right)\boldsymbol{U} = \begin{bmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) & 0 \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}707107 & -0{,}707107 & 0{,}000000 \\ 0{,}707107 & 0{,}707107 & 0{,}000000 \\ 0{,}000000 & 0{,}000000 & 1{,}000000 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

Затем задан второй поворот:

$$\boldsymbol{W} = \boldsymbol{R}_y\left(\frac{\pi}{6}\right)\boldsymbol{V} = \begin{bmatrix} \cos(\frac{\pi}{6}) & 0 & \sin(\frac{\pi}{6}) \\ 0 & 1 & 0 \\ -\sin(\frac{\pi}{6}) & 0 & \cos(\frac{\pi}{6}) \end{bmatrix} \boldsymbol{V} =$$

$$= \begin{bmatrix} 0{,}866025 & 0{,}000000 & 0{,}500000 \\ 0{,}000000 & 1{,}000000 & 0{,}000000 \\ -0{,}500000 & 0{,}000000 & 0{,}866025 \end{bmatrix} \boldsymbol{V}.$$

Композиция обоих поворотов имеет вид

$$\boldsymbol{W} = \boldsymbol{R}_y\left(\frac{\pi}{6}\right)\boldsymbol{R}_z\left(\frac{\pi}{4}\right)\boldsymbol{U} = \begin{bmatrix} 0{,}612372 & -0{,}612372 & 0{,}500000 \\ 0{,}707107 & 0{,}707107 & 0{,}000000 \\ -0{,}353553 & 0{,}353553 & 0{,}866025 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

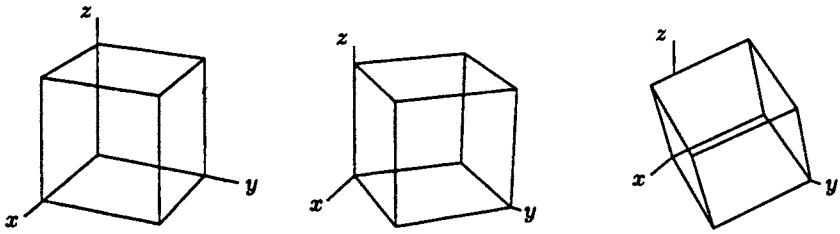
Вычисленные значения координат углов куба заданы в табл. 3.1 (как радиус-векторы); образы куба показаны на рис. 3.2. ■

MATLAB

MATLAB-функции `det(A)` и `inv(A)` вычисляют определитель и обратную матрицу (если матрица $\boldsymbol{A}$ обратима) соответственно квадратной матрицы $\boldsymbol{A}$.

Таблица 3.1. Координаты вершин куба при последовательных вращениях

U	$V = R_z\left(\frac{\pi}{4}\right)U$	$W = R_y\left(\frac{\pi}{6}\right)R_z\left(\frac{\pi}{4}\right)U$
$(0;0;0)'$	$(0{,}000000;\ 0{,}000000;\ 0)'$	$(0{,}000000;\ 0{,}000000;\ 0{,}000000)'$
$(1;0;0)'$	$(0{,}707107;\ 0{,}707107;\ 0)'$	$(0{,}612372;\ 0{,}707107;\ -0{,}353553)'$
$(0;1;0)'$	$(-0{,}707107;\ 0{,}707107;\ 0)'$	$(-0{,}612372;\ 0{,}707107;\ 0{,}353553)'$
$(0;0;1)'$	$(0{,}000000;\ 0{,}000000;\ 1)'$	$(0{,}500000;\ 0{,}000000;\ 0{,}866025)'$
$(1;1;0)'$	$(0{,}000000;\ 1{,}414214;\ 0)'$	$(0{,}000000;\ 1{,}414214;\ 0{,}000000)'$
$(1;0;1)'$	$(0{,}707107;\ 0{,}707107;\ 1)'$	$(1{,}112372;\ 0{,}707107;\ 0{,}512472)'$
$(0;1;1)'$	$(-0{,}707107;\ 0{,}707107;\ 1)'$	$(-0{,}112372;\ 0{,}707107;\ 1{,}219579)'$
$(1;1;1)'$	$(0{,}000000;\ 1{,}414214;\ 1)'$	$(0{,}500000;\ 1{,}414214;\ 0{,}866025)'$



(a) (b) (c)

Рис. 3.2. (a) Начальное положение куба. (b) $V = R_z(\pi/4)U$. Поворот вокруг оси z. (c) $W = R_y(\pi/6)V$. Поворот вокруг оси y

Пример 3.11. Воспользуемся MATLAB, чтобы решить систему линейных уравнений из примера 3.6. Используем для обращения матрицы, метод описаный в (25).

Сначала убедимся, что матрица A — невырожденная, показав, что $\det(A) \neq 0$ (см. теорему 3.4).

```
>>A=[0.125 0.200 0.400;0.375 0.500 0.600;0.500 0.300 0.000];
>>det(A)
ans=
   -0.0175
```

Следуя рассуждениям в (25), решение системы $AX = B$ представим в виде $X = A^{-1}B$.

```
>>X=inv(A)*[2.3 4.8 2.9]'
X=
   4.0000
   3.0000
   3.0000
```

Проверим наше решение, убедившись, что $\boldsymbol{AX} = \boldsymbol{B}$.

```
>>B=A*X
B=
   2.3000
   4.8000
   2.9000
```

■

Упражнения к разделу 3.2

Следующие упражнения читатель может выполнять как вручную так и с помощью MATLAB.

1. Найдите $\boldsymbol{AB}$ и $\boldsymbol{BA}$ для следующих матриц:

$$\boldsymbol{A} = \begin{bmatrix} -3 & 2 \\ 1 & 4 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 5 & 0 \\ 2 & -6 \end{bmatrix}.$$

2. Найдите $\boldsymbol{AB}$ и $\boldsymbol{BA}$ для следующих матриц:

$$\boldsymbol{A} = \begin{bmatrix} 1 & -2 & 3 \\ 2 & 0 & 5 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 3 & 0 \\ -1 & 5 \\ 3 & -2 \end{bmatrix}.$$

3. Пусть $\boldsymbol{A}, \boldsymbol{B}$ и $\boldsymbol{C}$ заданные матрицы:

$$\boldsymbol{A} = \begin{bmatrix} 3 & 1 \\ 0 & 4 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 1 & 2 \\ -2 & -6 \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} 2 & -5 \\ 3 & 4 \end{bmatrix}.$$

   (a) Найдите $(\boldsymbol{AB})\boldsymbol{C}$ и $\boldsymbol{A}(\boldsymbol{BC})$.
   (b) Найдите $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C})$ и $\boldsymbol{AB} + \boldsymbol{AC}$.
   (c) Найдите $(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C}$ и $\boldsymbol{AC} + \boldsymbol{BC}$.
   (d) Найдите $(\boldsymbol{AB})'$ и $\boldsymbol{B}'\boldsymbol{A}'$.

4. Воспользуемся обозначением $\boldsymbol{A}^2 = \boldsymbol{AA}$. Найдите $\boldsymbol{A}^2$ и $\boldsymbol{B}^2$ для следующих матриц:

$$\boldsymbol{A} = \begin{bmatrix} -1 & -7 \\ 5 & 2 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 2 & 0 & 6 \\ -1 & 5 & -4 \\ 3 & -5 & 2 \end{bmatrix}$$

5. Найдите определитель следующих матриц, если он существует.

   (a) $\begin{bmatrix} -1 & -7 \\ 5 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 0 & 6 \\ -1 & 5 & -4 \\ 3 & -5 & 2 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 0 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 4 & 6 \\ 0 & 0 & 5 & 4 \\ 0 & 0 & 0 & 7 \end{bmatrix}$

6. Покажите, что $\boldsymbol{R}_x(\alpha)\boldsymbol{R}_x(-\alpha) = \boldsymbol{I}$ согласно обычному умножению матриц $\boldsymbol{R}_x(\alpha)$ и $\boldsymbol{R}_x(-\alpha)$ (см. формулу (26)).

7. (a) Покажите, что $\boldsymbol{R}_x(\alpha)\boldsymbol{R}_y(\beta) =$

$$\begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ \sin(\beta)\sin(\alpha) & \cos(\alpha) & -\cos(\beta)\sin(\alpha) \\ -\cos(\alpha)\sin(\beta) & \sin(\alpha) & \cos(\beta)\cos(\alpha) \end{bmatrix}$$

(см. формулы (26) и (27)).

(b) Покажите, что $\boldsymbol{R}_y(\beta)\boldsymbol{R}_x(\alpha) =$

$$\begin{bmatrix} \cos(\beta) & \sin(\beta)\sin(\alpha) & \cos(\alpha)\sin(\beta) \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ -\sin(\alpha) & \cos(\beta)\sin(\alpha) & \cos(\beta)\cos(\alpha) \end{bmatrix}.$$

8. Если $\boldsymbol{A}$ и $\boldsymbol{B}$ — невырожденные матрицы размера $N \times N$ и $\boldsymbol{C} = \boldsymbol{AB}$, покажите, что $\boldsymbol{C}^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$. *Указание.* Используйте свойство ассоциативности умножения матриц.

9. Докажите утверждения (13) и (16) теоремы 3.3.

10. Пусть $\boldsymbol{A}$ — матрица размера $M \times N$ и $\boldsymbol{X}$ — матрица размера $N \times 1$.
    (a) Сколько операций умножения необходимо для вычисления $\boldsymbol{AX}$?
    (b) Сколько операций сложения необходимо для вычисления $\boldsymbol{AX}$?

11. Пусть $\boldsymbol{A}$ — матрица размера $M \times N$ и $\boldsymbol{B}$ и $\boldsymbol{C}$ — матрицы размера $N \times P$. Докажите левый дистрибутивный закон для умножения матриц: $\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$.

12. Пусть $\boldsymbol{A}$ и $\boldsymbol{B}$ — матрицы размера $M \times N$ и $\boldsymbol{C}$ — матрица размера $N \times P$. Докажите правый дистрибутивный закон для умножения матриц: $(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$.

13. Найдите $\boldsymbol{XX}'$ и $\boldsymbol{X}'\boldsymbol{X}$, где $\boldsymbol{X} = \begin{bmatrix} 1 & -1 & 2 \end{bmatrix}$. *Заметка.* Матрица $\boldsymbol{X}'$ равна транспонированной матрице $\boldsymbol{X}$.

14. Пусть $\boldsymbol{A}$ — матрица размера $M \times N$ и $\boldsymbol{B}$ — матрица размера $N \times P$. Докажите, что $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$. *Указание.* Положите $\boldsymbol{C} = \boldsymbol{AB}$ и покажите, используя определение умножения матриц, что (i, j)-й элемент матрицы $\boldsymbol{C}'$ равен (i, j)-му элементу матрицы $\boldsymbol{B}'\boldsymbol{A}'$.

15. Воспользуйтесь результатом упр. 14 и свойством ассоциативности умножения матриц, чтобы показать, что $(\boldsymbol{ABC})' = \boldsymbol{C}'\boldsymbol{B}'\boldsymbol{A}'$.

Алгоритмы и программы

В первом столбце табл. 3.1 содержатся координаты вершин единичного куба, расположенного в первом октанте с одной вершиной в начале координат. Заметим, что координаты всех восьми вершин можно записать в виде матрицы $\boldsymbol{U}$ размера 8×3, где каждая строка представляет собой координату одной вершины. Из упр. 14 следует, что матрица, равная произведению $\boldsymbol{U}$ и транспонированной матрицы $\boldsymbol{R}_z(\pi/4)$, имеет размер 8×3 (см. второй столбец табл. 3.1, где каждая строка представляет собой преобразование соответствующей строки матрицы $\boldsymbol{U}$). Из вышесказанного и упр. 15 следует, что координаты вершин куба при любом количестве последовательных вращений можно записать, как произведение матриц.

1. Единичный куб расположен в первом октанте с одной вершиной в начале координат. Сначала поворачиваем куб на угол $\pi/6$ вокруг оси y, затем поворачиваем этот образ на угол $\pi/4$ вокруг оси z. Найдите образы всех восьми вершин начального положения куба. Сравните результаты с результатами упр. 3.10.

   Каковы различия? Объясните свой ответ, учитывая тот факт, что, вообще, умножение матриц не коммутативно (рис. 3.3). Используйте программу `plot3`, чтобы построить каждый из трех кубов.

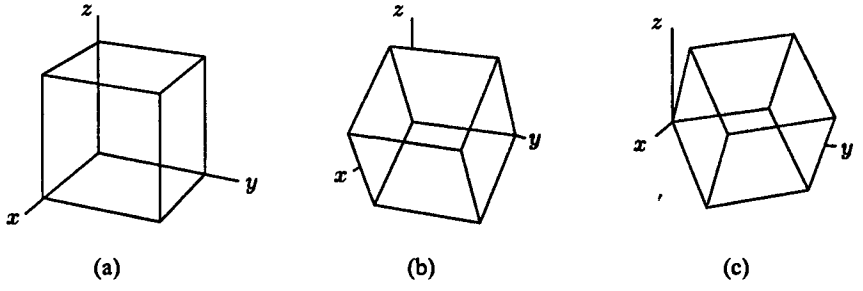


Рис. 3.3. (a) Начальное положение куба. (b) $\boldsymbol{V} = \boldsymbol{R}_y(\pi/6)\boldsymbol{U}$. Вращение вокруг оси y. (c) $\boldsymbol{W} = \boldsymbol{R}_z(\pi/4)\boldsymbol{V}$. Вращение вокруг оси z

2. Единичный куб расположен в первом октанте с одной вершиной в начале координат. Сначала поворачиваем куб на угол $\pi/12$ вокруг оси x, затем поворачиваем этот образ на угол $\pi/6$ вокруг оси z. Найдите образы всех восьми вершин начального положения куба. Используйте программу `plot3`, чтобы построить каждый из трех кубов.

3. Тетраэдр с вершинами в точках $(0;0;0), (1;0;0), (0;1;0)$ и $(0;0;1)$ сначала поворачивают на угол 0,15 радиан вокруг оси y, затем - на угол $-1{,}5$ радиан вокруг оси z и наконец на угол 2,7 радиан вокруг оси x. Найдите координаты образа всех четырех вершин. Используйте программу `plot3` для построения каждого из четырех образов.

3.3. Верхняя треугольная система линейных уравнений

Рассмотрим ***алгоритм обратной подстановки***, который полезен для решения систем линейных уравнений, имеющих верхнюю треугольную матрицу коэффициентов. В разделе 3.4 этот алгоритм войдет в алгоритм для решения общих систем линейных уравнений.

Определение 3.2. Матрица $\boldsymbol{A} = [a_{ij}]$ размера $N \times N$ называется ***верхней треугольной*** в том случае, если элементы $a_{ij} = 0$, как только $i > j$. Матрица $\boldsymbol{A} = [a_{ij}]$ размера $N \times N$ называется ***нижней треугольной*** в том случае, если элементы матрицы $a_{ij} = 0$ всякий раз, как только $i < j$. ▲

Рассмотрим метод получения решения верхней треугольной системы линейных уравнений и оставим читателю исследование нижней треугольной системы уравнений. Если $\boldsymbol{A}$ — верхняя треугольная матрица, то говорят, что $\boldsymbol{AX} = \boldsymbol{B}$ — это ***верхняя треугольная система*** линейных уравнений и она имеет вид

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + {} & a_{13}x_3 + \cdots + a_{1N-1}x_{N-1} + a_{1N}x_N = b_1 \\
& a_{22}x_2 + a_{23}x_3 + \cdots + a_{2N-1}x_{N-1} + a_{2N}x_N = b_2 \\
& a_{33}x_3 + \cdots + a_{3N-1}x_{N-1} + a_{3N}x_N = b_3 \\
& \qquad\qquad \vdots \qquad\qquad\qquad\quad \vdots \\
& a_{N-1N-1}x_{N-1} + a_{N-1N}x_N = b_{N-1} \\
& \qquad\qquad\qquad\qquad a_{NN}x_N = b_N.
\end{aligned} \tag{1}
$$

Теорема 3.5 (обратная подстановка). Предположим, что $\boldsymbol{AX} = \boldsymbol{B}$ — верхняя треугольная система линейных уравнений, заданная в виде (1). Если

$$
a_{kk} \neq 0 \quad \text{для } k = 1,\ 2,\ \ldots,\ N, \tag{2}
$$

то существует единственное решение системы (1).

Конструктивное доказательство. Решение сожно легко найти. Последнее уравнение включает только одно неизвестное x_N, поэтому вычисляем его первым:

$$
x_N = \frac{b_N}{a_{NN}}. \tag{3}
$$

Сейчас x_N известно и его можно использовать в следующем уравнении:

$$
x_{N-1} = \frac{b_{N-1} - a_{N-1N}x_N}{a_{N-1N-1}}. \tag{4}
$$

Затем x_N и x_{N-1} используем для нахождения x_{N-2}:

$$
x_{N-2} = \frac{b_{N-2} - a_{N-2N-1}x_{N-1} - a_{N-2N}x_N}{a_{N-2N-2}}. \tag{5}
$$

Когда значения $x_N, x_{N-1}, \ldots, x_{k+1}$ известны, можно записать общую формулу:

$$x_k = \frac{b_k - \sum_{j=k+1}^{N} a_{kj}x_j}{a_{kk}} \quad \text{для } k = N-1,\ N-2,\ \ldots,\ 1. \tag{6}$$

Легко видеть, что решение единственно. Из N-го уравнения следует, что b_N/a_{NN} — единственное возможное значение x_N. Наконец используем индукцию, чтобы установить единственность решений $x_{N-1}, x_{N-2}, \ldots, x_1$. •

Пример 3.12. Используем обратную подстановку для решения следующей системы линейных уравнений.

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ -2x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Решение для x_4 дано в последнем уравнении:

$$x_4 = \frac{6}{3} = 2.$$

Используя $x_4 = 2$ в третьем уравнении, получим

$$x_3 = \frac{4 - 5(2)}{6} = -1.$$

Сейчас используем $x_3 = -1$ и $x_4 = 2$, чтобы найти x_2 из следующего уравнения:

$$x_2 = \frac{-7 - 7(-1) + 4(2)}{-2} = -4.$$

И окончательно из первого уравнения получаем x_1:

$$x_1 = \frac{20 + 1(-4) - 2(-1) - 3(2)}{4} = 3. \qquad \blacksquare$$

Условие, что $a_{kk} \neq 0$, существенно, так как уравнение (6) включает деление на a_{kk}. Если это требование не выполняется, то либо не существует решения, либо существует бесконечно много решений.

Пример 3.13. Покажем, что система линейных уравнений (7) не имеет решения.

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ 0x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned} \tag{7}$$

Из последнего уравнения в (7) получим $x_4 = 2$, которое подставим во второе и третье уравнения, и получим

$$\begin{aligned} 7x_3 - \ 8 &= -7 \\ 6x_3 + 10 = \ \ 4. \end{aligned} \tag{8}$$

Из первого уравнения (8) следует, что $x_3 = 1/7$, а из второго — $x_3 = -1$. Это противоречие приводит к заключению, что система линейных уравнений (7) не имеет решения. ■

Пример 3.14. Покажем, что система уравнений (9) имеет бесконечно много решений.

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= \ 20 \\ 0x_2 + 7x_3 + 0x_4 &= -7 \\ 6x_3 + 5x_4 &= \ \ 4 \\ 3x_4 &= \ \ 6. \end{aligned} \tag{9}$$

Из последнего уравнения в (9) получим $x_4 = 2$, подставим его значение во второе и третье уравнения, чтобы получить $x_3 = -1$, которое удовлетворяет обоим уравнениям. Но только два значения, x_3 и x_4, можно получить из уравнений второго по четвертое. Если подставить их в первое уравнение системы уравнений (9), в результате получится уравнение

$$x_2 = 4x_1 - 16, \tag{10}$$

которое имеет бесконечно много решений. Следовательно, (9) имеет бесконечно много решений. Если выбрать значение x_1 в (10), то значение x_2 определено единственным образом. Например, если включить равенство $x_1 = 2$ в систему (9), из (10) получим, что $x_2 = -8$. ■

Теорема 3.4 утверждает, что система линейных уравнений $\boldsymbol{AX} = \boldsymbol{B}$, где $\boldsymbol{A}$ — матрица размера $N \times N$, имеет единственное решение тогда и только тогда, когда $\det(\boldsymbol{A}) \neq 0$. Следующая теорема утверждает, что если элемент на главной диагонали верхней или нижней треугольной матрицы равен нулю, то $\det(\boldsymbol{A}) = 0$. Таким образом, при внимательном рассмотрении коэффициентов матриц предыдущих уравнений становится ясно, что система линейных уравнений примера 3.12 имеет единственное значение, а системы линейных уравнений примеров 3.13 и 3.14 не имеют единственного решения. Доказательство теоремы 3.6 можно найти в большинстве учебников по линейной алгебре.

Теорема 3.6. Если $\boldsymbol{A} = [a_{ij}]$ размера $N \times N$ — либо верхняя, либо нижняя треугольная матрица, то

$$\det(\boldsymbol{A}) = a_{11}a_{22}\cdots a_{NN} = \prod_{i=1}^{N} a_{ii}. \tag{11}$$

Значение определителя матрицы коэффициентов в примере 3.12 равно $\det \boldsymbol{A} = 4(-2)(6)(3) = -144$. Значение определителя матрицы коэффициентов в примерах 3.13 и 3.14 равно $4(0)(6)(3) = 0$.

Следующая программа будет решать верхнюю треугольную систему линейных уравнений (1) методом обратной подстановки при условии, что $a_{kk} \neq 0$ для $k = 1, 2, \ldots, N$.

Программа 3.1 (обратная подстановка). Решение верхней треугольной системы линейных уравнений $\boldsymbol{AX} = \boldsymbol{B}$ методом обратной подстановки. Обращаться к методу только в том случае, когда диагональные элементы не равны нулю. Сначала вычисляем $x_N = b_N/a_{NN}$, а затем используем правило

$$x_k = \frac{b_k - \sum_{j=k+1}^{N} a_{kj}x_j}{a_{kk}} \quad \text{для } k = N-1,\ N-2,\ \ldots,\ 1.$$

```
function X=backsub(A,B)
%Вход  - A - верхняя треугольная матрица размера n x n
%      - B - матрица размера n x 1
%Выход - X - решение системы линейных уравнений AX = B
%Находим размер матрицы B и инициализируем X
n=length(B);
X=zeros(n,1);
X(n)=B(n)/A(n,n);
for k=n-1:-1:1
   X(k)=(B(k)-A(k,k+1:n)*X(k+1:n))/A(k,k);
end
```

Упражнения к разделу 3.3

В упр. 1–3 решите верхнюю треугольную систему линейных уравнений и найдите значение определителя матрицы коэффициентов.

1.
$$\begin{aligned} 3x_1 - 2x_2 + x_3 - x_4 &= 8 \\ 4x_2 - x_3 + 2x_4 &= -3 \\ 2x_3 + 3x_4 &= 11 \\ 5x_4 &= 15 \end{aligned}$$

2.
$$\begin{aligned} 5x_1 - 3x_2 - 7x_3 + x_4 &= -14 \\ 11x_2 + 9x_3 + 5x_4 &= 22 \\ 3x_3 - 13x_4 &= -11 \\ 7x_4 &= 14 \end{aligned}$$

3.
$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 2x_4 - x_5 &= 4 \\ -2x_2 + 6x_3 + 2x_4 + 7x_5 &= 0 \\ x_3 - x_4 - 2x_5 &= 3 \\ -2x_4 - x_5 &= 10 \\ 3x_5 &= 6 \end{aligned}$$

4. (a) Рассмотрите две верхние треугольные матрицы

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} \quad \text{и} \quad \boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ 0 & b_{22} & b_{23} \\ 0 & 0 & b_{33} \end{bmatrix}.$$

Покажите, что их произведение $\boldsymbol{C} = \boldsymbol{AB}$ также является верхней треугольной матрицей.

(b) Пусть $\boldsymbol{A}$ и $\boldsymbol{B}$ — две верхние треугольные матрицы размера $N \times N$. Покажите, что их произведение также является верхней треугольной матрицей.

5. Решите нижнюю треугольную систему линейных уравнений $\boldsymbol{AX} = \boldsymbol{B}$ и найдите определитель $\det(\boldsymbol{A})$.

$$\begin{aligned} 2x_1 &= 6 \\ -x_1 + 4x_2 &= 5 \\ 3x_1 - 2x_2 - x_3 &= 4 \\ x_1 - 2x_2 + 6x_3 + 3x_4 &= 2 \end{aligned}$$

6. Решите нижнюю треугольную систему линейных уравнений $\boldsymbol{AX} = \boldsymbol{B}$ и найдите определитель $\det(\boldsymbol{A})$.

$$\begin{aligned} 5x_1 &= -10 \\ x_1 + 3x_2 &= 4 \\ 3x_1 + 4x_2 + 2x_3 &= 2 \\ -x_1 + 3x_2 - 6x_3 - x_4 &= 5 \end{aligned}$$

7. Покажите, что для обратной подстановки требуется N операций деления, $(N^2 - N)/2$ операций умножения и $(N^2 - N)/2$ операций сложения или вычитания. *Указание.* Можно воспользоваться формулой

$$\sum_{k=1}^{M} k = M(M+1)/2.$$

Алгоритмы и программы

1. Воспользуйтесь программой 3.1, чтобы решить систему линейных уравнений $\boldsymbol{UX} = \boldsymbol{B}$, где

$$\boldsymbol{U} = [u_{ij}]_{10\times 10} \quad \text{и} \quad u_{ij} = \begin{cases} \cos(ij), & i \le j, \\ 0, & i > j. \end{cases}$$

   и $\boldsymbol{B} = [b_{i1}]_{10\times 1}$ и $b_{i1} = \tan(i)$.

2. *Алгоритм прямой подстановки.* Система линейных уравнений $\boldsymbol{AX} = \boldsymbol{B}$ называется нижней треугольной при условии, что $a_{ij} = 0$, когда $i < j$. Постройте программу `forsub`, аналогичную программе 3.1, для решения следующей системы нижней треугольной системы линейных уравнений. *Примечание.* Эта программа найдет применение в разделе 3.5.

$$\begin{aligned}
a_{11}x_1 & & & & & & & = b_1 \\
a_{21}x_1 + & a_{22}x_2 & & & & & & = b_2 \\
a_{31}x_1 + & a_{32}x_2 + & a_{33}x_3 & & & & & = b_3 \\
\vdots & \vdots & \vdots & & & & & \vdots \\
a_{N-1\,1}x_1 + & a_{N-1\,2}x_2 + & a_{N-1\,3}x_3 + & \cdots + & a_{N-1\,N-1}x_{N-1} & & & = b_{N-1} \\
a_{N1}x_1 + & a_{N2}x_2 + & a_{N3}x_3 + & \cdots + & a_{N\,N-1}x_{N-1} + & a_{NN}x_N & & = b_N
\end{aligned}$$

3. Воспользуйтесь программой `forsub`, чтобы решить систему линейных уравнений $\boldsymbol{LX} = \boldsymbol{B}$, где

$$\boldsymbol{L} = [l_{ij}]_{20\times 20} \quad \text{и} \quad l_{ij} = \begin{cases} i+j, & i \ge j, \\ 0, & i < j, \end{cases} \quad \text{и} \quad \boldsymbol{B} = [b_{i1}]_{20\times 1} \quad \text{и} \quad b_{i1} = i.$$

3.4. Метод исключения Гаусса и выбор главного элемента

В этом разделе будет рассмотрена схема решения системы $\boldsymbol{AX} = \boldsymbol{B}$ N уравнений с N неизвестными. Нашей целью является построение эквивалентной верхней треугольной системы $\boldsymbol{UX} = \boldsymbol{Y}$, которую можно решить методами, рассмотренными в разделе 3.3. (Далее, как это принято в литературе по вычислительным методам, под термином "система" (или "линейная система") будем понимать систему линейных уравнений. — *Прим. ред.*)

Говорят, что две линейные системы размера $N \times N$ ***эквивалентны***, если они имеют одно и то же множество решений. Теоремы из линейной алгебры показывают, что применение определенных преобразований к заданной системе не изменяет множества решений.

Теорема 3.7 (элементарные преобразования). Следующие операции, примененные к линейной системе, приводят к эквивалентной системе.

(1)	Перестановки	Порядок двух уравнений может быть изменен
(2)	Масштабирование	Умножение уравнения на не равную нулю константу
(3)	Замещение	Уравнение можно заменить суммой этого же уравнения и любого другого уравнения, умноженного на не равную нулю константу

Обычно используют (3), чтобы заменить уравнение разностью между этим уравнением и кратным другому уравнению. Эти понятия проиллюстрируем на следующем примере.

Пример 3.15. Найдем параболу $y = A + Bx + Cx^2$, которая проходит через три точки: $(1;1)$, $(2;-1)$ и $(3;1)$.

Для каждой точки получим уравнение, связывающее значения x со значением y. Результатом является система линейных уравнений.

$$\begin{aligned} A + B + C &= 1 \quad \text{в } (1;1) \\ A + 2B + 4C &= -1 \quad \text{в } (2;-1) \\ A + 3B + 9C &= 1 \quad \text{в } (3;1). \end{aligned} \tag{4}$$

Если вычесть первое уравнение из второго и третьего, то исключается переменная A. Это применение преобразования замещения (3). В результате получаем эквивалентную линейную систему.

$$\begin{aligned} A + B + C &= 1 \\ B + 3C &= -2 \\ 2B + 8C &= 0. \end{aligned} \tag{5}$$

Переменная B исключается из третьего уравнения (5) путем двукратного вычитания из него второго уравнения. Мы пришли к эквивалентной верхней треугольной системе линейных уравнений.

$$\begin{aligned} A + B + C &= 1 \\ B + 3C &= -2 \\ 2C &= 4. \end{aligned} \tag{6}$$

А сейчас воспользуемся алгоритмом обратной подстановки для нахождения коэффициентов $C = 4/2 = 2$, $B = -2 - 3(2) = -8$ и $A = 1 - (-8) - 2 = 7$. Следовательно, уравнение параболы имеет вид $y = 7 - 8x + 2x^2$. ∎

Эффективнее всего можно хранить коэффициенты линейной системы $\boldsymbol{AX} = \boldsymbol{B}$, как матрицу размера $N \times (N+1)$. Коэффициенты $\boldsymbol{B}$ хранятся в $(N+1)$-м столбце матрицы (т. е., $a_{kN+1} = b_k$). В каждой строке содержатся все необходимые для уравнения линейной системы коэффициенты. ***Расширенная матрица*** обозначается как $[\boldsymbol{A}|\boldsymbol{B}]$, и систему линейных уравнений представляют в следующем виде:

$$[\boldsymbol{A}|\boldsymbol{B}] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1N} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2N} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} & b_N \end{array}\right]. \tag{7}$$

Систему $\boldsymbol{AX} = \boldsymbol{B}$ с заданной в (7) расширенной матрицей можно решить, выполнив ряд операций над расширенной матрицей $[\boldsymbol{A}|\boldsymbol{B}]$. Переменные x_k определяются положением коэффициентов и могут быть опущены до конца вычислений.

Теорема 3.8 (элементарный ряд операций). Следующие операции, примененные к расширенной матрице (7), приводят к эквивалентной линейной системе.

(8) Перестановки — Порядок двух строк может быть изменен

(9) Масштабирование — Умножение строки на не равную нулю константу

(10) Замещение — Строку можно заменить суммой этой же строки и любой другой строки, умноженной на не равную нулю константу (ненулевое кратное другой строки), т. е. $\text{строка}_r = \text{строка}_r - m_{rp} \times \text{строка}_p$.

Обычно, чтобы воспользоваться (10), заменяют строку разностью между этой строкой и кратным другой строки.

Определение 3.3 (главный элемент). Коэффициент a_{rr} матрицы $\boldsymbol{A}$, который используется, чтобы исключить элементы a_{kr}, где $k = r+1, r+2, \ldots, N$, называется r-м ***главным элементом*** и r-я строка — ***главной строкой***. ▲

Следующий пример иллюстрирует, как использовать операции теоремы 3.8, чтобы получить эквивалентную верхнюю треугольную линейную систему $\boldsymbol{UX} = \boldsymbol{Y}$ для линейной системы $\boldsymbol{AX} = \boldsymbol{B}$, где $\boldsymbol{A}$ — матрица размера $N \times N$.

Пример 3.16. Выразим следующую систему в форме расширенной матрицы, найдем эквивалентную ей верхнюю треугольную систему линейных уравнений и ее решение.

$$\begin{aligned} x_1 + 2x_2 + x_3 + 4x_4 &= 13 \\ 2x_1 + 0x_2 + 4x_3 + 3x_4 &= 28 \\ 4x_1 + 2x_2 + 2x_3 + x_4 &= 20 \\ -3x_1 + x_2 + 3x_3 + 2x_4 &= 6. \end{aligned}$$

Расширенная матрица имеет вид

$$\begin{array}{l} \text{гл. эл.} \rightarrow \\ m_{21} = 2 \\ m_{31} = 4 \\ m_{41} = -3 \end{array} \left[\begin{array}{cccc|c} \underline{1} & 2 & 1 & 4 & 13 \\ 2 & 0 & 4 & 3 & 28 \\ 4 & 2 & 2 & 1 & 20 \\ -3 & 1 & 3 & 2 & 6 \end{array}\right].$$

Первая строка используется, чтобы исключить элементы под диагональю в первом столбце. Мы обращаемся к первой строке, как к главной, и называем элемент $a_{11} = 1$ главным. Значение m_{k1} является множителем строки 1, которую вычитаем из k строк, $k = 2, 3, 4$. Результатом первого исключения будет

$$\begin{array}{l} \\ \text{гл. эл.} \rightarrow \\ m_{32} = 1{,}5 \\ m_{42} = -1{,}75 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & \underline{-4} & 2 & -5 & 2 \\ 0 & -6 & -2 & -15 & -32 \\ 0 & 7 & 6 & 14 & 45 \end{array}\right].$$

Вторая строка используется, чтобы исключить элементы под диагональю во втором столбце. Эта строка является главной, и значение m_{k1} является множителем строки 2, которую вычитаем из k строк, $k = 3, 4$. Результатом исключения будет

$$\begin{array}{l} \\ \text{гл. эл.} \rightarrow \\ m_{43} = -1{,}9 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & \underline{-5} & -7{,}5 & -35 \\ 0 & 0 & 9{,}5 & 5{,}25 & 48{,}5 \end{array}\right].$$

Наконец, умножаем $m_{43} = -1{,}9$ на третью строку, вычитаем из четвертой строки и в результате получаем верхнюю треугольную систему линейных уравнений

$$\left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & -5 & -7{,}5 & -35 \\ 0 & 0 & 0 & -9 & -18 \end{array}\right]. \tag{11}$$

Для решения системы линейных уравнений (11) воспользуемся алгоритмом обратной подстановки и получим

$$x_4 = 2, \quad x_3 = 4, \quad x_2 = -1, \quad x_1 = 3. \qquad \blacksquare$$

Описанный выше процесс называется ***методом исключения Гаусса*** и должен быть модифицирован так, чтобы его можно было использовать в большинстве случаев. Если $a_{kk} = 0$, то строку k нельзя использовать для исключения элементов столбца k и строку k следует заменить такой же строкой под диагональю, чтобы получить не равный нулю главный элемент. Если этого сделать нельзя, значит, матрица коэффициентов системы линейных уравнений является вырожденной и система не имеет единственного решения.

Теорема 3.9 (метод исключения Гаусса с обратной подстановкой[1]). Если матрица $\boldsymbol{A}$ размера $N \times N$ невырождена, то существует система $\boldsymbol{UX} = \boldsymbol{Y}$, эквивалентная $\boldsymbol{AX} = \boldsymbol{B}$, где $\boldsymbol{U}$ — верхняя треугольная матрица с элементом $u_{kk} \neq 0$. После построения матриц $\boldsymbol{U}$ и $\boldsymbol{Y}$ можно использовать обратную подстановку, чтобы решить систему $\boldsymbol{UX} = \boldsymbol{Y}$ для $\boldsymbol{X}$.

Доказательство. Используем расширенную матрицу с элементами матрицы столбца $\boldsymbol{B}$, записанными в $(N+1)$-й столбец:

$$\boldsymbol{AX} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2N}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3N}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{N1}^{(1)} & a_{N2}^{(1)} & a_{N3}^{(1)} & \cdots & a_{NN}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} a_{1\,N+1}^{(1)} \\ a_{2\,N+1}^{(1)} \\ a_{3\,N+1}^{(1)} \\ \vdots \\ a_{N\,N+1}^{(1)} \end{bmatrix} = \boldsymbol{B}.$$

Затем построим эквивалентную верхнюю треугольную систему $\boldsymbol{UX} = \boldsymbol{Y}$:

$$\boldsymbol{UX} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN}^{(N)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} a_{1\,N+1}^{(1)} \\ a_{2\,N+1}^{(2)} \\ a_{3\,N+1}^{(3)} \\ \vdots \\ a_{N\,N+1}^{(N)} \end{bmatrix} = \boldsymbol{Y}.$$

Шаг 1. Запишем коэффициенты в расширенной матрице. Верхний индекс $a_{rc}^{(1)}$ означает, что сначала число занимает место (r, c), т. е. находится на пересечении r-й строки и столбца c:

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1\,N+1}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2N}^{(1)} & a_{2\,N+1}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3N}^{(1)} & a_{3\,N+1}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_{N1}^{(1)} & a_{N2}^{(1)} & a_{N3}^{(1)} & \cdots & a_{NN}^{(1)} & a_{N\,N+1}^{(1)} \end{array}\right].$$

[1] Иногда говорят — метод Гаусса с обратным ходом. — Прим. перев.

Шаг 2. Если есть необходимость, переставляем строки так, чтобы $a_{11}^{(1)} \neq 0$, затем исключаем x_1 в строках 2–N. При этом m_{r1} — множитель строки 1 при ее вычитании из строки r.

```
for r = 2 : N
  m_r1 = a_r1^(1) / a_11^(1);
  a_r1^(2) = 0;
  for c = 2 : N + 1
    a_rc^(2) = a_rc^(1) - m_r1 * a_1c^(1);
  end
end
```

Новые элементы записаны как $a_{rc}^{(2)}$, чтобы показать, что на втором шаге число хранится в матрице на том же месте (r, c). После второго шага получаем

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1\,N+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2\,N+1}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3N}^{(2)} & a_{3\,N+1}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{N2}^{(2)} & a_{N3}^{(2)} & \cdots & a_{NN}^{(2)} & a_{N\,N+1}^{(2)} \end{array}\right].$$

Шаг 3. Если необходимо, так переставляем вторую строку со строкой, стоящей ниже, чтобы $a_{22}^{(2)} \neq 0$, затем исключаем x_2 из строк 3–N. В этом случае m_{r2} — множитель строки 2 при ее вычитании из строки r.

```
for r = 3 : N
  m_r2 = a_r2^(2) / a_22^(2);
  a_r2^(3) = 0;
  for c = 3 : N + 1
    a_rc^(3) = a_rc^(2) - m_r2 * a_2c^(2);
  end
end
```

Новые элементы записаны как $a_{rc}^{(3)}$, чтобы показать, что на третьем шаге число хранится в матрице на том же месте (r, c). После третьего шага получаем

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1\,N+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2\,N+1}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3\,N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & a_{N3}^{(3)} & \cdots & a_{NN}^{(3)} & a_{N\,N+1}^{(3)} \end{array}\right].$$

Шаг $p+1$. Это шаг общего вида. Если необходимо, так переставляем строку p со строкой, стоящей ниже, чтобы $a_{pp}^{(p)} \neq 0$, затем исключаем x_p в строках $p+1$–N. Здесь m_{rp} — множитель строки p при вычитании из строки r.

```
for r = p + 1 : N
  m_rp = a_rp^(p) / a_pp^(p);
  a_rp^(p+1) = 0;
  for c = p + 1 : N + 1
    a_rc^(p+1) = a_rc^(p) − m_rp * a_pc^(p);
  end
end
```

Окончательно после исключения x_{N-1} из строки N получаем

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1\,N+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2\,N+1}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3\,N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN}^{(N)} & a_{N\,N+1}^{(N)} \end{array}\right].$$

На этом процесс построения верхней треугольной матрицы окончен.

Так как матрица $\boldsymbol{A}$ невырождена, после осуществления этих операций получим также невырожденную матрицу. Это является гарантией, что $a_{kk}^{(k)} \neq 0$ для всех k в процессе построения. Таким образом, можно использовать обратную подстановку, чтобы решить $\boldsymbol{UX} = \boldsymbol{Y}$ для $\boldsymbol{X}$. Теорема доказана. •

Выбор главных элементов во избежание $a_{pp}^{(p)} = 0$

Если $a_{pp}^{(p)} = 0$, то строку p нельзя использовать, чтобы исключать элементы в столбце p ниже главной диагонали. Необходимо найти строку k, где $a_{kp}^{(p)} \neq 0$ и $k > p$, и затем поменять местами строку p и строку k так, чтобы получить главный элемент, не равный нулю. Этот процесс называется выбором главного элемента, и критерий для решения, какую выбрать строку, называется стратегией выбора главного элемента. Стратегия ***тривиального выбора главного элемента*** следующая. Если $a_{pp}^{(p)} \neq 0$, то не переставляем строки. Если $a_{pp}^{(p)} = 0$, найдем первую строку ниже строки p, в которой $a_{kp}^{(p)} \neq 0$, и поменяем местами строки k и p. В результате получим новый элемент $a_{pp}^{(p)} \neq 0$, который и будет главным элементом, не равным нулю.

Выбор главного элемента для уменьшения ошибки

Поскольку компьютер использует арифметику с фиксированной точностью, то, возможно, при выполнении арифметической операции каждый раз вводится небольшая ошибка. Проиллюстрируем на следующем примере, как использование стратегии тривиального выбора главного элемента в методе исключения Гаусса может привести к значительной ошибке в решении системы линейных уравнений.

Пример 3.17. Значения $x_1 = x_2 = 1{,}000$ являются решениями системы

$$\begin{aligned} 1{,}133x_1 + 5{,}281x_2 &= 6{,}414 \\ 24{,}14x_1 - 1{,}210x_2 &= 22{,}93. \end{aligned} \tag{12}$$

Используем арифметику с четырьмя знаками точности (см. упр. 6 и 7 раздела 1.3) и метод исключения Гаусса с тривиальным выбором главного элемента, чтобы найти приближенное решение системы.

Чтобы получить верхнюю треугольную систему линейных уравнений, строку 1, умноженную на $m_{21} = 24{,}14/1{,}133 = 21{,}31$, вычитаем из строки 2. Используя четыре знака в вычислениях, получим новые коэффициенты:

$$\begin{aligned} a_{22}^{(2)} &= -1{,}210 - 21{,}31(5{,}281) = -1{,}210 - 112{,}5 = -113{,}7 \\ a_{23}^{(2)} &= \quad 22{,}93 - 21{,}31(6{,}414) = \quad 22{,}93 - 136{,}7 = -113{,}8. \end{aligned}$$

Вычисленная верхняя треугольная система линейных уравнений имеет вид

$$\begin{aligned} 1{,}133x_1 + 5{,}281x_2 &= \quad 6{,}414 \\ -113{,}7x_2 &= -113{,}8. \end{aligned}$$

С помощью обратной подстановки вычислим $x_2 = -113{,}8/(-113{,}7) = 1{,}001$ и $x_1 = (6{,}414 - 5{,}281(1{,}001))/(1{,}133) = (6{,}414 - 5{,}286)/(1{,}133) = 0{,}9956$. ∎

Ошибка в решении системы линейных уравнений (12) обусловлена значением множителя $m_{21} = 21{,}31$. В следующем примере значение множителя m_{21} уменьшается после первой замены первого и второго уравнений линейной системы (12). Затем, чтобы получить решение системы, в методе исключения Гаусса используется стратегия тривиального выбора главного элемента.

Пример 3.18. Чтобы решить следующую линейную систему, используем арифметику с четырьмя знаками и метод исключения Гаусса с тривиальным выбором главного элемента:

$$\begin{aligned} 24{,}14x_1 - 1{,}210x_2 &= 22{,}93 \\ 1{,}133x_1 + 5{,}281x_2 &= 6{,}414. \end{aligned}$$

В этот раз множитель строки (1) равен $m_{21} = 1{,}133/24{,}14 = 0{,}04693$. Новые коэффициенты равны

$$\begin{aligned} a_{22}^{(2)} &= 5{,}281 - 0{,}04693(-1{,}210) = 5{,}281 + 0{,}05679 = 5{,}338 \\ a_{23}^{(2)} &= 6{,}414 - 0{,}04693(22{,}93) \quad = 6{,}414 - 1{,}076 \quad = 5{,}338. \end{aligned}$$

Получим верхнюю треугольную систему

$$\begin{aligned} 24{,}14x_1 - 1{,}210x_2 &= 22{,}93 \\ 5{,}338x_2 &= 5{,}338. \end{aligned}$$

Используем обратную подстановку, чтобы вычислить $x_2 = 5{,}338/5{,}338 = 1{,}000$ и $x_1 = (22{,}93 + 1{,}210(1{,}000))/(24{,}14) = 1{,}000$. ■

Назначение стратегии выбора главного элемента — сдвиг наибольшего по величине элемента на главную диагональ и его использование для исключения оставшихся в столбце элементов. Если существует больше одного, не равного нулю, элемента в столбце p, который лежит на главной диагонали или ниже нее, то выбор определяет, какую из строк заменить. Стратегию ***частного выбора главного элемента*** иллюстрирует пример 3.18; она является наиболее общей и используется в программе 3.2. Чтобы уменьшить распространение ошибки, советуем проверить значения всех элементов в столбце p, которые лежат на главной диагонали или ниже нее. Определим строку k, в которой находится наибольший по абсолютной величине элемент, т. е.

$$|a_{kp}| = \max\{|a_{pp}|, |a_{p+1p}|, \ldots, |a_{N-1p}|, |a_{Np}|\},$$

и затем поменяем строку p со строкой k, если $k > p$. Сейчас каждый из множителей m_{rp} для $r = p + 1, \ldots, N$ меньше или равен 1 по абсолютной величине. Этот процесс обычно сохраняет относительные величины элементов матрицы U в теореме 3.9 такими же, как и элементов исходной матрицы коэффициентов A.

Обычно выбор наибольшего главного элемента приводит в результате к меньшему распространению ошибки.

В разделе 3.5 указывается, что для решения системы размера $N \times N$ достаточно всего $(4N^3 + 9N^2 - 7N)/6$ арифметических операций. Когда $N = 20$, общее число арифметических операций, которые нужно выполнить, равно 5910 и распространение ошибки в вычислениях может привести к ошибочному ответу. Технику ***определения масштаба частного выбора главного элемента*** или уравновешивания можно использовать для дальнейшего уменьшения влияния распространения ошибки. При масштабном частном выборе главного элемента мы выбираем из всех элементов в столбце p, которые лежат на главной диагонали или ниже нее, элемент, являющийся наибольшим относительно элементов в его строке. Сначала рассмотрим строки p–N, чтобы найти наибольший по величине элемент в каждой строке, например s_r:

$$s_r = \max\{|a_{rp}|, |a_{rp+1}|, \ldots, |a_{rN}|\} \quad \text{для } r = p,\ p+1,\ \ldots,\ N. \tag{13}$$

Главную строку k находим, определяя

$$\frac{|a_{kp}|}{s_k} = \max\left\{\frac{|a_{pp}|}{s_p}, \frac{|a_{p+1p}|}{s_{p+1}}, \ldots, \frac{|a_{Np}|}{s_N}\right\}. \tag{14}$$

Заменим строку p строкой k, исключая случай, когда $p = k$. Этот процесс выбора главного элемента составлен так, чтобы сохранить отношение величин элементов матрицы $\boldsymbol{U}$ в теореме 3.9 такими же, как в исходной матрице коэффициентов $\boldsymbol{A}$.

Плохая обусловленность

Матрица $\boldsymbol{A}$ называется ***плохо обусловленной***, если существует такая матрица $\boldsymbol{B}$, что при небольших возмущениях коэффициентов матриц $\boldsymbol{A}$ или $\boldsymbol{B}$ произойдут большие изменения в $\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{B}$. Говорят, что система $\boldsymbol{AX} = \boldsymbol{B}$ плохо обусловлена, когда матрица $\boldsymbol{A}$ плохо обусловлена. В этом случае численные методы приближенного вычисления могут привести к большим ошибкам.

Плохая обусловленность возникает, когда матрица $\boldsymbol{A}$ "почти вырождена" и определитель $\boldsymbol{A}$ близок к нулю. Плохая обусловленность также имеет место в системах двух уравнений, когда две линии почти параллельны (или в системе из трех уравнений, когда три плоскости почти параллелльны). Следствием плохой обусловленности является то, что может произойти замена истинного решения ошибочным значением. Рассмотрим, например, два уравнения:

$$\begin{aligned} x + 2y - 2{,}00 &= 0 \\ 2x + 3y - 3{,}40 &= 0. \end{aligned} \tag{15}$$

Подстановка $x_0 = 1{,}00$ и $y_0 = 0{,}48$ в эти два уравнения "почти приводит к нулям":

$$\begin{aligned} 1 + 2(0{,}48) - 2{,}00 &= 1{,}96 - 2{,}00 = -0{,}04 \approx 0 \\ 2 + 3(0{,}48) - 3{,}40 &= 3{,}44 - 3{,}40 = \ \ 0{,}04 \approx 0. \end{aligned}$$

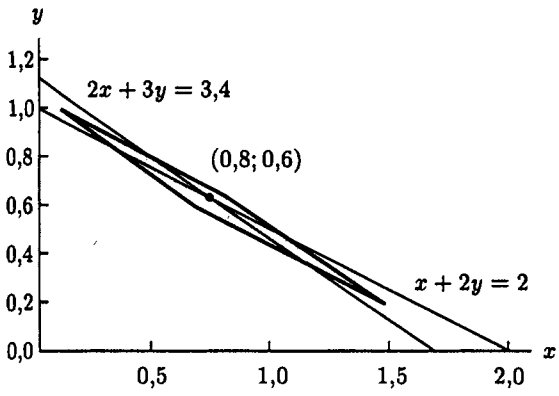


Рис. 3.4. Область, точки которой "почти удовлетворяют" обоим уравнениям

Здесь расхождение с 0 равно только $\pm 0{,}04$. Тем не менее истинное решение этой линейной системы равно $x = 0{,}8$ и $y = 0{,}6$, поэтому ошибки приближенного решения равны $x - x_0 = 0{,}80 - 1{,}00 = -0{,}20$ и $y - y_0 = 0{,}60 - 0{,}48 = 0{,}12$. Таким образом, просто подстановка значений в систему уравнений не является надежным критерием точности. Ромбообразная область R на рис. 3.4 представляет множество, точки которого "почти удовлетворяет" обоим уравнениям в (15):

$$R = \{(x; y) : |x + 2y - 2{,}00| < 0{,}1 \quad \text{и} \quad |2x + 3y - 3{,}40| < 0{,}2\}.$$

В области R существуют точки, которые достаточно далеки от точки, которая является решением, (0,8; 0,6); все-таки они дают малые значения, если подставить их в уравнения в (15). Если окажется, что линейная система плохо обусловлена, вычисления нужно будет выполнять с арифметикой многократной точности. Заинтересованный читатель может познакомиться с понятием чисел обусловленности матрицы и их свойствами, чтобы получить больше информации об этом феномене.

Плохая обусловленность имеет более сильные последствия, когда нужно решить несколько уравнений. Рассмотрим задачу нахождения кубического полинома $y = c_1x^3 + c_2x^2 + c_3x + c_4$, который проходит через четыре точки: (2; 8), (3; 27), (4; 64) и (5; 125) (очевидно, что $y = x^3$ — требуемый кубический полином). В разделе 3.5 будет введен метод наименьших квадратов. Применим метод наименьших квадратов, чтобы найти требуемые коэффициенты. Для этого решим следующую линейную систему:

$$\begin{bmatrix} 20\,514 & 4\,424 & 978 & 224 \\ 4\,424 & 978 & 224 & 54 \\ 978 & 224 & 54 & 14 \\ 224 & 54 & 14 & 4 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ lc_4 \end{bmatrix} = \begin{bmatrix} 20\,514 \\ 4\,424 \\ 978 \\ 224 \end{bmatrix}.$$

Для вычисления коэффициентов использовался компьютер с 9-ю знаками точности. Были получены значения

$$c_1 = 1{,}000004; \quad c_2 = -0{,}000038; \quad c_3 = 0{,}000126 \quad \text{и} \quad c_4 = -0{,}000131.$$

Хотя они близки к истинному решению $c_1 = 1$ и $c_2 = c_3 = c_4 = 0$, этот пример показывает, как легко ошибка может прокрасться в решение. Более того, предположим, что коэффициент $a_{11} = 20\,514$ в верхнем левом углу матрицы коэффициентов заменен значением $20\,515$ и решается возмущенная система. Значения, полученные на таком же компьютере и равные

$$c_1 = 0{,}642857; \quad c_2 = 3{,}75000; \quad c_3 = -12{,}3928 \quad \text{и} \quad c_4 = 12{,}7500,$$

являются ничего не стоящим ответом. Плохую обусловленность нелегко обнаружить. Если после повторного решения системы со слабо возмущенными коэффициентами обнаружится, что ответ существенно отличается от предыдущего, значит, присутствует плохая обусловленность. Раздел, посвященный анализу чувствительности методов, обычно входит в книги повышенной сложности по численному анализу.

MATLAB

В программе 3.2 MATLAB обозначение `[A B]` используется для построения расширенной матрицы линейной системы $\boldsymbol{AX} = \boldsymbol{B}$; команда `max` используется для определения главного элемента в частном выборе главного элемента. Как только получена эквивалентная треугольная матрица $[\boldsymbol{U}|\boldsymbol{Y}]$, она разделяется на матрицы $\boldsymbol{U}$ и $\boldsymbol{Y}$; для выполнения обратной подстановки используется программа 3.1 (`backsub(U,Y)`). Использование этих команд и процессов иллюстрируется в следующем примере.

Пример 3.19. (a) Используем MATLAB, чтобы построить расширенную матрицу для линейной системы из примера 3.16; (b) используем команду `max`, чтобы найти наибольший по величине элемент в первом столбце матрицы коэффициентов $\boldsymbol{A}$; (c) разъединим расширенную матрицу (11) на матрицу коэффициентов $\boldsymbol{U}$ и матрицу постоянных членов $\boldsymbol{Y}$ верхней треугольной системы $\boldsymbol{UX} = \boldsymbol{Y}$.

(a)

```
>> A=[1 2 1 4;2 0 4 3;4 2 2 1;-3 1 3 2];
>> B=[13 28 20 6]';
>> Aug=[A B]
Aug=
    1 2 1 4 13
    2 0 4 3 28
    4 2 2 1 20
   -3 1 3 2 6
```

(b) В следующей программе MATLAB показано, что a — наибольший элемент в первом столбце матрицы A и j — номер строки.

```
>> [a,j]=max{abs(A(1:4,1))}
a=
   4
j=
   3
```

(c) Пусть Augup $= [U|Y]$ — верхняя треугольная матрица для (11).

```
>> Augup=[1 2 1 4 13;0 -4 2 -5 2;0 0 -5 -7.5 -35;0 0 0 -9 -18];
>> U=Augup(1:4,1:4)
U=
   1.0000 2.0000 1.0000  4.0000
    0    -4.0000 2.0000 -5.0000
    0     0     -5.0000 -7.5000
    0     0      0      -9.0000
>> Y=Augup(1:4,5)
Y=
    13
     2
   -35
   -18
```

■

Программа 3.2 (построение верхней треугольной матрицы для применения метода обратной подстановки). Чтобы найти решение системы $AX = B$, сначала приводим расширенную матрицу $[A|B]$ к верхней треугольной форме, а затем выполняем обратную подстановку.

```
function X = uptrbk(A,B)
%Вход  - A - невырожденная матрица размера N x N
%      - B - матрица размера N x 1
%Выход - X - матрица размера N x 1, содержащая решение AX=B
%Инициализация X и временное сохранение матрицы C
    [N N]=size(A);
    X=zeros(N,1);
    C=zeros(1,N+1);
%Вид расширенной матрицы:Aug=[A|B]
    Aug=[A B];
for p=1:N-1
   %Частный выбор главного элемента для столбца p
   [Y,j]=max(abs(Aug(p:N,p)));
```

```
   %Меняем местами строки p и j
   C=Aug(p,:);
   Aug(p,:)=Aug(j+p-1,:);
   Aug(j+p-1,:)=C;
if Aug(p,p)==0
   'A вырождена. Нет единственного решения'
   break
end
%Процесс исключения для столбца p
   for k=p+1:N
      m=Aug(k,p)/Aug(p,p);
      Aug(k,p:N+1)=Aug(k,p:N+1)-m*Aug(p,p:N+1);
   end
end
%Обратная подстановка в [U|Y] с использованием программы 3.1
X=backsub(Aug(1:N,1:N),Aug(1:N,N+1));
```

Упражнения к разделу 3.4

В упр. 1–4 покажите, что $\boldsymbol{AX} = \boldsymbol{B}$ эквивалентна верхней треугольной системе линейных уравнений $\boldsymbol{UX} = \boldsymbol{Y}$, и найдите ее решение.

1.
$$\begin{aligned} 2x_1 + 4x_2 - 6x_3 &= -4 \\ x_1 + 5x_2 + 3x_3 &= 10 \\ x_1 + 3x_2 + 2x_3 &= 5 \end{aligned} \qquad \begin{aligned} 2x_1 + 4x_2 - 6x_3 &= -4 \\ 3x_2 + 6x_3 &= 12 \\ 3x_3 &= 3 \end{aligned}$$

2.
$$\begin{aligned} x_1 + x_2 + 6x_3 &= 7 \\ -x_1 + 2x_2 + 9x_3 &= 2 \\ x_1 - 2x_2 + 3x_3 &= 10 \end{aligned} \qquad \begin{aligned} x_1 + x_2 + 6x_3 &= 7 \\ 3x_2 + 15x_3 &= 9 \\ 12x_3 &= 12 \end{aligned}$$

3.
$$\begin{aligned} 2x_1 - 2x_2 + 5x_3 &= 6 \\ 2x_1 + 3x_2 + x_3 &= 13 \\ -x_1 + 4x_2 - 4x_3 &= 3 \end{aligned} \qquad \begin{aligned} 2x_1 - 2x_2 + 5x_3 &= 6 \\ 5x_2 - 4x_3 &= 7 \\ 0{,}9x_3 &= 1{,}8 \end{aligned}$$

4.
$$\begin{aligned} -5x_1 + 2x_2 - x_3 &= -1 \\ x_1 + 0x_2 + 3x_3 &= 5 \\ 3x_1 + x_2 + 6x_3 &= 17 \end{aligned} \qquad \begin{aligned} -5x_1 + 2x_2 - x_3 &= -1 \\ 0{,}4x_2 + 2{,}8x_3 &= 4{,}8 \\ -10x_3 &= -10 \end{aligned}$$

5. Найдите коэффициенты параболы $y = A + Bx + Cx^4$, которая проходит через точки $(0;4)$, $(2;7)$ и $(3;14)$.

6. Найдите коэффициенты параболы $y = A + Bx + Cx^2$, которая проходит через точки $(1;6)$, $(2;5)$ и $(3;2)$.

7. Найдите коэффициенты кривой третьего порядка $y = A + Bx + Cx^3 + Dx^3$, которая проходит через точки $(0;4), (1;1), (2;2)$ и $(3;2)$.

В упр. 8–10 покажите, что $\boldsymbol{AX} = \boldsymbol{B}$ эквивалентна верхней треугольной системе линейных уравнений $\boldsymbol{UX} = \boldsymbol{Y}$, и найдите решение.

8.
$$\begin{aligned} 4x_1 + 8x_2 + 4x_3 + 0x_4 &= 8 \\ x_1 + 5x_2 + 4x_3 - 3x_4 &= -4 \\ x_1 + 4x_2 + 7x_3 + 2x_4 &= 10 \\ x_1 + 3x_2 + 0x_3 - 2x_4 &= -4 \end{aligned} \qquad \begin{aligned} 4x_1 + 8x_2 + 4x_3 + 0x_4 &= 8 \\ 3x_2 + 3x_3 - 3x_4 &= -6 \\ 4x_3 + 4x_4 &= 12 \\ x_4 &= 2 \end{aligned}$$

9.
$$\begin{aligned} 2x_1 + 4x_2 - 4x_3 + 0x_4 &= 12 \\ x_1 + 5x_2 - 5x_3 - 3x_4 &= 18 \\ 2x_1 + 3x_2 + x_3 + 3x_4 &= 8 \\ x_1 + 4x_2 - 2x_3 + 2x_4 &= 8 \end{aligned} \qquad \begin{aligned} 2x_1 + 4x_2 - 4x_3 + 0x_4 &= 12 \\ 3x_2 - 3x_3 - 3x_4 &= 12 \\ 4x_3 + 2x_4 &= 0 \\ 3x_4 &= -6 \end{aligned}$$

10.
$$\begin{aligned} x_1 + 2x_2 + 0x_3 - x_4 &= 9 \\ 2x_1 + 3x_2 - x_3 + 0x_4 &= 9 \\ 0x_1 + 4x_2 + 2x_3 - 5x_4 &= 26 \\ 5x_1 + 5x_2 + 2x_3 - 4x_4 &= 32 \end{aligned} \qquad \begin{aligned} x_1 + 2x_2 + 0x_3 - x_4 &= 9 \\ -x_2 - x_3 + 2x_4 &= -9 \\ -2x_3 + 3x_4 &= -10 \\ 1{,}5x_4 &= -3 \end{aligned}$$

11. Найдите решение линейной системы.

$$\begin{aligned} x_1 + 2x_2 &= 7 \\ 2x_1 + 3x_2 - x_3 &= 9 \\ 4x_2 + 2x_3 + 3x_4 &= 10 \\ 2x_3 - 4x_4 &= 12 \end{aligned}$$

12. Найдите решение линейной системы.

$$\begin{aligned} x_1 + x_2 &= 5 \\ 2x_1 - x_2 + 5x_3 &= -9 \\ 3x_2 - 4x_3 + 2x_4 &= 19 \\ 2x_3 + 6x_4 &= 2 \end{aligned}$$

13. Компании Rockmore Corp. необходим новый компьютер. Пока она остановила свой выбор на DoGood 174 и MightDo 11 и проверяют возможности обоих компьютеров на решении системы линейных уравнений

$$\begin{aligned} 34x + 55y - 21 &= 0 \\ 55x + 89y - 34 &= 0. \end{aligned}$$

Компьютер DoGood 174 дал значения $x = -0{,}11$ и $y = 0{,}45$; его точность проверяется подстановкой:

$$34(-0{,}11) + 55(0{,}45) - 21 = 0{,}01$$
$$55(-0{,}11) + 89(0{,}45) - 34 = 0{,}00.$$

Компьютер MightDo 11 получил значения $x = -0{,}99$ и $y = 1{,}01$; его точность проверяется подстановкой:

$$34(-0{,}99) + 55(1{,}01) - 21 = 0{,}89$$
$$55(-0{,}99) + 89(1{,}01) - 34 = 1{,}44.$$

Ответ какого компьютера лучше? Объясните, почему?

14. Решите следующую систему линейных уравнений с помощью (i) метода исключения Гаусса с частным выбором главных элементов и (ii) метода исключения Гаусса с определением масштаба частного выбора главных элементов.

(a)
$$\begin{aligned} 2x_1 - 3x_2 + 100x_3 &= 1 \\ x_1 + 10x_2 - 0{,}001x_3 &= 0 \\ 3x_1 - 100x_2 + 0{,}01x_3 &= 0 \end{aligned}$$

(b)
$$\begin{aligned} x_1 + 20x_2 - x_3 + 0{,}001x_4 &= 0 \\ 2x_1 - 5x_2 + 30x_3 - 0{,}1x_4 &= 1 \\ 5x_1 + x_2 - 100x_3 - 10x_4 &= 0 \\ 2x_1 - 100x_2 - x_3 + x_4 &= 0 \end{aligned}$$

15. Матрица Гильберта — это пример классической плохо обусловленной матрицы, и малые изменения ее коэффициентов приводят к большим изменениям в решении возмущенной системы.

(a) Найдите точное решение $\boldsymbol{AX} = \boldsymbol{B}$ (оставьте все числа в виде дробей и точно выполняйте арифметические операции), используя матрицу Гильберта размера 4×4:

$$\boldsymbol{A} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

(b) Решите $\boldsymbol{AX} = \boldsymbol{B}$, используя арифметику с округлением до четырех знаков:

$$\boldsymbol{A} = \begin{bmatrix} 1{,}0000 & 0{,}5000 & 0{,}3333 & 0{,}2500 \\ 0{,}5000 & 0{,}3333 & 0{,}2500 & 0{,}2000 \\ 0{,}3333 & 0{,}2500 & 0{,}2000 & 0{,}1667 \\ 0{,}2500 & 0{,}2000 & 0{,}1667 & 0{,}1429 \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Заметка. Матрица коэффициентов п. (b) является приближением для матрицы коэффициентов п. (a).

Алгоритмы и программы

1. Многие приложения включают матрицы с большим количеством нулей. Практическое значение имеют ***трехдиагональные системы*** (см. упр. 11 и 12) вида

$$\begin{aligned}
d_1x_1 + c_1x_2 &= b_1\\
a_1x_1 + d_2x_2 + c_2x_3 &= b_2\\
a_2x_2 + d_3x_3 + c_3x_4 &= b_3\\
&\cdots\\
a_{N-2}x_{N-2} + d_{N-1}x_{N-1} + c_{N-1}x_N &= b_{N-1}\\
a_{N-1}x_{N-1} + d_Nx_N &= b_N.
\end{aligned}$$

   Постройте программу, которая решит трехдиагональную систему. Можете предположить, что нет необходимости переставлять строки и что строку k можно использовать для исключения элемента x_k в строке $k+1$.

2. Используйте программу 3.2, чтобы найти кривую шестого порядка $y = a_1 + a_2x + a_3x^2 + a_4x^3 + a_5x^4 + a_6x^5 + a_7x^6$, которая проходит через точки $(0;1)$, $(1;3)$, $(2;2)$, $(3;1)$, $(4;3)$, $(5;2)$ и $(6;1)$. Используйте команду `plot`, чтобы построить график кривой с заданными точками на нем. Объясните любые расхождения в вашем графике.

3. Используйте программу 3.2, чтобы решить линейную систему $\boldsymbol{AX} = \boldsymbol{B}$, где $\boldsymbol{A} = [a_{ij}]_{N\times N}$, $a_{ij} = i^{j-1}$, $\boldsymbol{B} = [b_{ij}]_{N\times 1}$, где $b_{11} = N$ и $b_{i1} = i^{N-2}/(i-1)$ для $i \geq 2$. Используйте $N = 3, 7$ и 11. Точное решение равно $\boldsymbol{X} = [1 \;\; 1 \;\; \ldots \;\; 1 \;\; 1]'$. Объясните любые отклонения от точного решения.

4. Постройте программу, которая заменит стратегию выбора главного элемента программы 3.2 методом определения масштаба частного выбора главного элемента.

5. Используйте свою программу определения масштаба частного выбора главного элемента в задаче 4, чтобы решить систему, заданную в задаче 3 для $N = 11$. Объясните любые отклонения от точного решения.

6. Модифицируйте программу 3.2 таким образом, чтобы она стала эффективной для решения M линейных систем с одинаковыми матрицами коэффициентов $\boldsymbol{A}$, но различными матрицами-столбцами $\boldsymbol{B}$. M линейных систем выглядят так:

$$\boldsymbol{AX}_1 = \boldsymbol{B}_1, \quad \boldsymbol{AX}_2 = \boldsymbol{B}_2, \quad \ldots, \quad \boldsymbol{AX}_M = \boldsymbol{B}_M.$$

7. Следующие результаты обсуждаются для матриц размера 3×3, но их можно применять к матрицам размера $N \times N$. Если $\boldsymbol{A}$ невырождена, то существуют

A^{-1} и $AA^{-1} = I$. Пусть C_1, C_2 и C_3 — столбцы матрицы A^{-1} и E_1, E_2 и E_3 — столбцы матрицы I. Уравнение $AA^{-1} = I$ можно записать в виде

$$A\begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix} = \begin{bmatrix} E_1 & E_2 & E_3 \end{bmatrix}.$$

Произведение матриц эквивалентно трем линейным системам:

$$AC_1 = E_1, \quad AC_2 = E_2, \quad \text{и} \quad AC_3 = E_3.$$

Поэтому нахождение матрицы A^{-1} эквивалентно решению трех линейных систем.

Воспользуйтесь программой 3.2 или своей программой из задачи 6 и найдите обратную к каждой из следующих матриц. Проверьте ответ с помощью вычисления произведения AA^{-1} и команды `inv(A)`. Объясните любые различия.

(a) $\begin{bmatrix} 2 & 0 & 1 \\ 3 & 2 & 5 \\ 1 & -1 & 0 \end{bmatrix}$

(b) $\begin{bmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{bmatrix}$

3.5. Разложение на треугольные матрицы

Из раздела 3.3 видно, как легко решить верхнюю треугольную систему. Сейчас введем понятие разложения (также используется термин "факторизация". — *Прим. ред.*) данной матрицы A на произведение нижней треугольной матрицы L, на главной диагонали которой стоят 1, и верхней треугольной матрицы U с не равными нулю диагональными элементами. Для простоты проиллюстрируем эти понятия на матрице размера 4×4, но они применимы к произвольной системе размера $N \times N$.

Определение 3.4. Невырожденную матрицу A можно ***разложить на треугольные матрицы***, если ее можно представить как произведение нижней треугольной матрицы L и верхней треугольной матрицы U:

$$A = LU. \tag{1}$$

В матричном виде это записывается как

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & m_{32} & 1 & 0 \\ m_{41} & m_{42} & m_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

(Такое разложение автор называет LU разложением. — *Прим. ред.*) ▲

Из условия, что $\boldsymbol{A}$ не вырождена, следует, что $u_{kk} \neq 0$ для всех k. Обозначим элементы матрицы $\boldsymbol{L}$ через m_{ij} и дадим обоснование, почему выбрано обозначение m_{ij} вместо l_{ij}.

Решение линейной системы

Предположим, что матрицу коэффициентов $\boldsymbol{A}$ линейной системы $\boldsymbol{AX} = \boldsymbol{B}$ можно разложить на треугольные матрицы (1). Тогда решение

$$\boldsymbol{LUX} = \boldsymbol{B} \tag{2}$$

можно получить, полагая $\boldsymbol{Y} = \boldsymbol{UX}$, и затем решить две системы:

$$\boldsymbol{LY} = \boldsymbol{B} \text{ для } \boldsymbol{Y}, \quad \text{чтобы получить } \boldsymbol{UX} = \boldsymbol{Y} \text{ для } \boldsymbol{X}. \tag{3}$$

Сначала решаем нижнюю треугольную систему

$$\begin{aligned} y_1 &= b_1 \\ m_{21}y_1 + y_2 &= b_2 \\ m_{31}y_1 + m_{32}y_2 + y_3 &= b_3 \\ m_{41}y_1 + m_{42}y_2 + m_{43}y_3 + y_4 &= b_4 \end{aligned} \tag{4}$$

чтобы получить y_1, y_2, y_3 и y_4. Используем их для решения верхней треугольной системы.

$$\begin{aligned} u_{11}x_1 + u_{12}x_2 + u_{13}x_3 + u_{14}x_4 &= y_1 \\ u_{22}x_2 + u_{23}x_3 + u_{24}x_4 &= y_2 \\ u_{33}x_3 + u_{34}x_4 &= y_3 \\ u_{44}x_4 &= y_4. \end{aligned} \tag{5}$$

Пример 3.20. Решить следующую систему

$$\begin{aligned} x_1 + 2x_2 + 4x_3 + x_4 &= 21 \\ 2x_1 + 8x_2 + 6x_3 + 4x_4 &= 52 \\ 3x_1 + 10x_2 + 8x_3 + 8x_4 &= 79 \\ 4x_1 + 12x_2 + 10x_3 + 6x_4 &= 82. \end{aligned}$$

Воспользуемся методом разложения на треугольные матрицы и тем фактом, что

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 4 & 1 \\ 2 & 8 & 6 & 4 \\ 3 & 10 & 8 & 8 \\ 4 & 12 & 10 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 1 & 1 & 0 \\ 4 & 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 & 1 \\ 0 & 4 & -2 & 2 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 0 & -6 \end{bmatrix} = \boldsymbol{LU}.$$

Чтобы решить систему $\boldsymbol{LY} = \boldsymbol{B}$, используем метод прямой подстановки.

$$
\begin{aligned}
y_1 &= 21 \\
2y_1 + y_2 &= 52 \\
3y_1 + y_2 + y_3 &= 79 \\
4y_1 + y_2 + 2y_3 + y_4 &= 82.
\end{aligned} \tag{6}
$$

Получаем значения $y_1 = 21$, $y_2 = 52 - 2(21) = 10$, $y_3 = 79 - 3(21) - 10 = 6$ и $y_4 = 82 - 4(21) - 10 - 2(6) = -24$ или $\boldsymbol{Y} = \begin{bmatrix} 21 & 10 & 6 & -24 \end{bmatrix}'$. Затем записываем систему $\boldsymbol{UX} = \boldsymbol{Y}$.

$$
\begin{aligned}
x_1 + 2x_2 + 4x_3 + x_4 &= 21 \\
4x_2 - 2x_3 + 2x_4 &= 10 \\
-2x_3 + 3x_4 &= 6 \\
-6x_4 &= -24.
\end{aligned} \tag{7}
$$

Воспользуемся обратной подстановкой и вычислим $x_4 = -24/(-6) = 4$, $x_3 = (6 - -3(4))/(-2) = 3$, $x_2 = (10 - 2(4) + 2(3))/4 = 2$ и $x_1 = 21 - 4 - 4(3) - 2(2) = 1$ или $\boldsymbol{X} = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}'$. ■

Разложение на треугольные матрицы

А сейчас обсудим, как получить разложение на треугольные матрицы. Если нет необходимости менять местами строки, то, когда используется метод исключения Гаусса, коэффициенты m_{ij} являются элементами, расположенными под диагональю матрицы $\boldsymbol{L}$.

Пример 3.21. Используем метод исключения Гаусса, чтобы построить разложение на треугольные матрицы

$$
\boldsymbol{A} = \begin{bmatrix} 4 & 3 & -1 \\ -2 & -4 & 5 \\ 1 & 2 & 6 \end{bmatrix}.
$$

Построим матрицу $\boldsymbol{L}$ из единичной матрицы, расположенной слева. В результате каждой операции со строкой, используемой для построения верхней треугольной матриц?, множители m_{ij} будут помещены на их собственные места слева. Начнем с

$$
\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ -2 & -4 & 5 \\ 1 & 2 & 6 \end{bmatrix}.
$$

Строку 1 используем, чтобы исключить элементы матрицы $\boldsymbol{A}$ в столбце 1 под a_{11}. Элементы $m_{21} = -0{,}5$ и $m_{31} = 0{,}25$ умножаем на строку 1 и вычитаем из

строк 2 и 3 соответственно. Эту же операцию производим с матрицей слева и в в результате получаем

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ -0{,}5 & 1 & 0 \\ 0{,}25 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ 0 & -2{,}5 & 4{,}5 \\ 0 & 1{,}25 & 6{,}25 \end{bmatrix}.$$

Строку 2 используем, чтобы исключить элементы матрицы $\boldsymbol{A}$ в столбце 2 ниже a_{22}. Множитель $m_{32} = -0{,}5$ второй строки вычитаем из строки 3, коэффициент помещаем в матрице слева и получаем требуемое разложение матрицы $\boldsymbol{A}$ на треугольные матрицы.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ -0{,}5 & 1 & 0 \\ 0{,}25 & -0{,}5 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ 0 & -2{,}5 & 4{,}5 \\ 0 & 0 & 8{,}5 \end{bmatrix}. \tag{8}$$

■

Теорема 3.10 (прямое разложение $\boldsymbol{A} = \boldsymbol{LU}$; без перестановки строк). Предположим, что методом исключения Гаусса, без перестановок строк, можно успешно решить линейную систему $\boldsymbol{AX} = \boldsymbol{B}$. Тогда матрицу $\boldsymbol{A}$ можно разложить на множители в виде произведения нижней треугольной матрицы $\boldsymbol{L}$ и верхней треугольной матрицы $\boldsymbol{U}$:

$$\boldsymbol{A} = \boldsymbol{LU}.$$

Более того, матрицу $\boldsymbol{L}$ можно построить так, чтобы на диагонали стояли 1 и матрица $\boldsymbol{U}$ на диагонали имела не равные нулю элементы. После того как построены матрицы $\boldsymbol{L}$ и $\boldsymbol{U}$, решение $\boldsymbol{X}$ получаем за два шага.

1. Решаем $\boldsymbol{LU} = \boldsymbol{B}$ для $\boldsymbol{Y}$, используя прямую подстановку.
2. Решаем $\boldsymbol{UX} = \boldsymbol{Y}$ для $\boldsymbol{X}$, используя обратную подстановку.

Доказательство. Покажем, что, когда применен метод исключения Гаусса и матрица $\boldsymbol{B}$ хранится как $(N+1)$-й столбец расширенной матрицы, в результате шага, на котором матрица приводится к виду верхней треугольной матрицы, получаем эквивалентную верхнюю треугольную систему $\boldsymbol{UX} = \boldsymbol{Y}$. Матрицы $\boldsymbol{L}, \boldsymbol{U}, \boldsymbol{B}$ и $\boldsymbol{Y}$ будут иметь следующий вид:

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 & & 0 \\ m_{21} & 1 & 0 & \cdots & 0 \\ m_{31} & m_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ m_{N1} & m_{N2} & m_{N3} & \cdots & 1 \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} a^{(1)}_{1\,N+1} \\ a^{(2)}_{2\,N+1} \\ a^{(3)}_{3\,N+1} \\ \vdots \\ a^{(N)}_{N\,N+1} \end{bmatrix},$$

$$U = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN}^{(N)} \end{bmatrix}, \qquad Y = \begin{bmatrix} a_{1\,N+1}^{(1)} \\ a_{2\,N+1}^{(2)} \\ a_{3\,N+1}^{(3)} \\ \vdots \\ a_{N\,N+1}^{(N)} \end{bmatrix}.$$

Примечание. Чтобы найти именно L и U, в $(N+1)$-м столбце нет необходимости.

Шаг 1. Запишем коэффициенты в расширенную матрицу. Верхний индекс $a_{rc}^{(1)}$ означает, что на первом шаге число занимает место (r, c).

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1\,N+1}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2N}^{(1)} & a_{2\,N+1}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3N}^{(1)} & a_{3\,N+1}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_{N1}^{(1)} & a_{N2}^{(1)} & a_{N3}^{(1)} & \cdots & a_{NN}^{(1)} & a_{N\,N+1}^{(1)} \end{array}\right]$$

Шаг 2. Исключим элемент x_1 в строках 2–N и сохраним множитель m_{r1}, используемый для исключения x_1 из ряда r. В матрице он занимает место $(r, 1)$.

```
for r = 2 : N
  m_r1 = a_r1^(1) / a_11^(1);
  a_r1 = m_r1;
  for c = 2 : N + 1
    a_rc^(2) = a_rc^(1) − m_r1 * a_1c^(1);
  end
end
```

Новые элементы записаны как $a_{rc}^{(2)}$, чтобы показать, что это второй шаг и число будет храниться в матрице на месте (r, c). В результате после второго шага получаем

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1\,N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2\,N+1}^{(2)} \\ m_{31} & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3N}^{(2)} & a_{3\,N+1}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & a_{N2}^{(2)} & a_{N3}^{(2)} & \cdots & a_{NN}^{(2)} & a_{N\,N+1}^{(2)} \end{array}\right].$$

Шаг 3. Исключим элемент x_2 в строках 3–N и сохраним множитель m_{r2}, который использовался для исключения элемента x_2 в строке r, в матрице на месте $(r, 2)$.

```
for r = 3 : N
   m_{r2} = a_{r2}^{(2)} / a_{22}^{(2)};
   a_{r2} = m_{r2};
   for c = 3 : N + 1
      a_{rc}^{(3)} = a_{rc}^{(2)} - m_{r2} * a_{2c}^{(2)};
   end
end
```

Новые элементы записаны как $a_{rc}^{(3)}$, чтобы показать, что это третий шаг и что число занимает в матрице место $(r, 2)$.

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1\,N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2\,N+1}^{(2)} \\ m_{31} & m_{32} & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3\,N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & m_{N2} & a_{N3}^{(3)} & \cdots & a_{NN}^{(3)} & a_{N\,N+1}^{(3)} \end{array}\right]$$

Шаг $p+1$. Это шаг общего вида. Исключаем x_p в строках $p+1$–N и сохраняем множитель на месте (r, p).

```
for r = p + 1 : N
   m_{rp} = a_{rp}^{(p)} / a_{pp}^{(p)};
   a_{rp} = m_{rp};
   for c = p + 1 : N + 1
      a_{rc}^{(p+1)} = a_{rc}^{(p)} - m_{rp} * a_{pc}^{(p)};
   end
end
```

Окончательным результатом после исключения элемента x_{N-1} в N-й строке будет

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1\,N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2\,N+1}^{(2)} \\ m_{31} & m_{32} & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3\,N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & m_{N2} & m_{N3} & \cdots & a_{NN}^{(N)} & a_{N\,N+1}^{(N)} \end{array}\right].$$

Построение верхней треугольной матрицы завершено. Отметим, что одна матрица используется для записи элементов обеих матриц $\boldsymbol{L}$ и $\boldsymbol{U}$. Ни единицы матрицы $\boldsymbol{L}$, ни нули матриц $\boldsymbol{L}$ и $\boldsymbol{U}$, которые лежат выше и ниже диагонали, соответственно не записаны. Необходимы только существенные коэффициенты, чтобы воссоздать перестроенные матрицы $\boldsymbol{L}$ и $\boldsymbol{U}$!

Убедимся, что произведение $\boldsymbol{LU} = \boldsymbol{A}$. Предположим, что $\boldsymbol{D} = \boldsymbol{LU}$, и рассмотрим случай, когда $r \leq c$. Тогда d_{rc} равно

$$d_{rc} = m_{r1}a_{1c}^{(1)} + m_{r2}a_{2c}^{(2)} + \cdots + m_{rr-1}a_{r-1c}^{(r-1)} + a_{rc}^{(r)}. \tag{9}$$

Используя замену уравнений на шагах с 1-го по $p + 1 = r$, получим следующие подстановки:

$$\begin{aligned} m_{r1}a_{1c}^{(1)} &= a_{rc}^{(1)} - a_{rc}^{(2)}, \\ m_{r2}a_{2c}^{(2)} &= a_{rc}^{(2)} - a_{rc}^{(3)}, \\ &\vdots \\ m_{rr-1}a_{r-1c}^{(r-1)} &= a_{rc}^{(r-1)} - a_{rc}^{(r)}. \end{aligned} \tag{10}$$

Если подстановки в (10) использовать в (9), то в результате получим

$$d_{rc} = a_{rc}^{(1)} - a_{rc}^{(2)} + a_{rc}^{(2)} - a_{rc}^{(3)} + \cdots + a_{rc}^{(r-1)} - a_{rc}^{(r)} + a_{rc}^{(r)} = a_{rc}^{(1)}.$$

Другой случай, $r > c$, доказывается аналогично. •

Вычислительные трудности

Процесс приведения к треугольному виду одинаков как для метода исключения Гаусса, так и для метода разложения на треугольные матрицы. Можно подсчитать число операций, если посмотреть на первые N столбцов расширенной матрицы теоремы 3.10. Внешний цикл шага $p+1$ требует $N - p = N - (p+1) + 1$ делений для вычисления множителей m_{rp}. Внутренние циклы, но только для первых N

столбцов, требуют для вычисления новых строк элементов $a_{rc}^{(p+1)}$ всего $(N - p)(N - p)$ умножений и такое же число вычитаний. Этот процесс выполняется для $p = 1, 2, \ldots, N - 1$. Таким образом, часть вычислений, которую занимает разложение на треугольные матрицы $\boldsymbol{A} = \boldsymbol{LU}$, требует

$$\sum_{p=1}^{N-1}(N-p)(N-p+1) = \frac{N^3 - N}{3} \quad \text{умножений и делений} \tag{11}$$

и

$$\sum_{p=1}^{N-1}(N-p)(N-p) = \frac{2N^3 - 3N^2 + N}{6} \quad \text{вычитаний.} \tag{12}$$

Чтобы получить (11), используем формулы суммирования

$$\sum_{k=1}^{M} k = \frac{M(M+1)}{2} \quad \text{и} \quad \sum_{k=1}^{M} k^2 = \frac{M(M+1)(2M+1)}{6}.$$

Произведя замену переменных $k = N - p$, перепишем (11) в виде

$$\begin{aligned}\sum_{p=1}^{N-1}(N-p)(N-p+1) &= \sum_{p=1}^{N-1}(N-p) + \sum_{p=1}^{N-1}(N-p)^2 = \\ &= \sum_{k=1}^{N-1} k + \sum_{k=1}^{N-1} k^2 = \\ &= \frac{(N-1)N}{2} + \frac{(N-1)(N)(2N-1)}{6} = \\ &= \frac{N^3 - N}{3}.\end{aligned}$$

Как только разложение $\boldsymbol{A} = \boldsymbol{LU}$ на треугольные матрицы будет получено, для решения нижней треугольной системы $\boldsymbol{LY} = \boldsymbol{B}$ потребуется $0+1+\cdots+N-1 = (N^2 - N)/2$ умножений и вычитаний, но операции деления не потребуются, потому что диагональные элементы матрицы $\boldsymbol{L}$ равны 1. Решение верхней треугольной системы $\boldsymbol{UX} = \boldsymbol{Y}$ потребует $1+2+\cdots+N = (N^2+N)/2$ умножений и делений и $(N^2-N)/2$ вычитаний. Следовательно, для решения системы $\boldsymbol{LUX} = \boldsymbol{B}$ потребуется

$$N^2 \text{ умножений и делений и } N^2 - N \text{ вычитаний.}$$

Мы видим, что большая часть вычислений приходится на разложение на треугольные матрицы. Если линейную систему решать много раз с той же матрицей

коэффициентов $\boldsymbol{A}$, но с различными матрицами-столбцами $\boldsymbol{B}$, нет необходимости в разложении матрицы каждый раз, если запомнить множители. Поэтому в таком случае метод разложения на треугольные матрицы предпочитают методу исключения Гаусса. Но, если решается только одна линейная система, оба метода одинаковы, за исключением того, что метод разложения на треугольные матрицы позволяет сохранить множители.

Перестановка матриц

Разложение матрицы $\boldsymbol{A} = \boldsymbol{LU}$ в теореме 3.10 предполагает, что строки не переставляют. Но случается, что невырожденную матрицу $\boldsymbol{A}$ нельзя непосредственно представить в виде $\boldsymbol{A} = \boldsymbol{LU}$.

Пример 3.22. Покажем, что следующую матрицу нельзя непосредственно представить в виде $\boldsymbol{A} = \boldsymbol{LU}$:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 6 \\ 4 & 8 & -1 \\ -2 & 3 & 5 \end{bmatrix}.$$

Предположим, матрицу $\boldsymbol{A}$ можно представить в виде $\boldsymbol{LU}$, тогда

$$\begin{bmatrix} 1 & 2 & 6 \\ 4 & 8 & -1 \\ -2 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}. \tag{13}$$

Можно перемножить матрицы $\boldsymbol{L}$ и $\boldsymbol{U}$, расположенные справа в (13), и каждый элемент произведения сравнить с соответствующим элементом матрицы $\boldsymbol{A}$. В первом столбце $1 = 1u_{11}$, затем $4 = m_{21}u_{11} = m_{21}$ и наконец $-2 = m_{31}u_{11} = m_{31}$. Во втором столбце $2 = 1u_{12}$, затем $8 = m_{21}u_{12} = (4)(2) + u_{22}$ (откуда следует, что $u_{22} = 0$) и наконец $3 = m_{31}u_{12} + m_{32}u_{22} = (-2)(2) + m_{32}(0) = -4$, что приводит к противоречию. Следовательно, $\boldsymbol{A}$ нельзя разложить на произведение матриц $\boldsymbol{LU}$. ■

Перестановка первых N положительных чисел 1, 2, ..., N является размещением k_1, k_2, ..., k_N этих целых чисел в определенном порядке. Например, $1, 4, 2, 3, 5$ является перестановкой пяти целых чисел $1, 2, 3, 4, 5$. В следующем определении используем стандартные обозначения для векторов, образующих базис, $\boldsymbol{E}_i = [0\ 0\ \cdots\ 0\ 1_i\ 0\ \cdots\ 0]$ для $i = 1, 2, \ldots, N$.

Определение 3.5. ***Матрицей перестановок*** $\boldsymbol{P}$ размера $N \times N$ называется матрица с точно одним элементом, равным 1 в каждом столбце и строке, и всеми остальными элементами, равными 0. Строки матрицы $\boldsymbol{P}$ являются перестановками строк единичной матрицы и могут быть записаны как

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{E}'_{k_1} & \boldsymbol{E}'_{k_2} & \cdots & \boldsymbol{E}'_{k_N} \end{bmatrix}'. \tag{14}$$

Элементы матрицы $\boldsymbol{P} = [p_{ij}]$ имеют вид

$$p_{ij} = \begin{cases} 1 & j = k_i, \\ 0 & \text{в остальных случаях.} \end{cases}$$

Например, следующая матрица размера 4×4 является матрицей перестановок:

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{E}_2' & \boldsymbol{E}_1' & \boldsymbol{E}_4' & \boldsymbol{E}_3' \end{bmatrix}'. \tag{15}$$

▲

Теорема 3.11. Предположим, что $\boldsymbol{P} = \begin{bmatrix} \boldsymbol{E}_{k_1}' & \boldsymbol{E}_{k_2}' & \dots & \boldsymbol{E}_{k_N}' \end{bmatrix}'$ — матрица перестановок. Произведение $\boldsymbol{PA}$ — это новая матрица, строки которой состоят из строк матрицы $\mathbf{A}$, размещенных в таком порядке: строка$_{k_1}$ $\boldsymbol{A}$, строка$_{k_2}$ $\boldsymbol{A}$, ..., строка$_{k_N}$ $\boldsymbol{A}$.

Пример 3.23. Пусть $\boldsymbol{A}$ — матрица размера 4×4 и $\boldsymbol{P}$ — матрица перестановок, заданная в (15). Тогда $\boldsymbol{PA}$ — это матрица, строки которой содержат строки матрицы $\boldsymbol{A}$, размещенные в таком порядке: строка$_2$ $\boldsymbol{A}$, строка$_1$ $\boldsymbol{A}$, строка$_4$ $\boldsymbol{A}$, строка$_3$ $\boldsymbol{A}$.

Вычислим произведение и получим

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{21} & a_{22} & a_{23} & a_{24} \\ a_{11} & a_{12} & a_{13} & a_{14} \\ a_{41} & a_{42} & a_{43} & a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}. \tag{16}$$

■

Теорема 3.12. Если $\boldsymbol{P}$ — матрица перестановок, то она невырождена и $\boldsymbol{P}^{-1} = \boldsymbol{P}'$.

Теорема 3.13. Если $\boldsymbol{A}$ — невырожденная матрица, то существует такая матрица перестановок $\boldsymbol{P}$, что $\boldsymbol{PA}$ можно разложить на треугольные матрицы

$$\boldsymbol{PA} = \boldsymbol{LU}. \tag{17}$$

Доказательство можно найти в книге повышенной сложности по линейной алгебре.

Пример 3.24. Если строки 2 и 3 матрицы из примера 3.22 поменять местами, то в результате получим матрицу $\boldsymbol{PA}$, которую можно разложить на треугольные матрицы.

Матрица перестановок, меняющая местами строки 2 и 3, имеет вид $\boldsymbol{P} = \begin{bmatrix} \boldsymbol{E}_1' & \boldsymbol{E}_3' & \boldsymbol{E}_2' \end{bmatrix}'$. Вычислим произведение $\boldsymbol{PA}$ и получим

$$\boldsymbol{PA} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 6 \\ 4 & 8 & -1 \\ -2 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 6 \\ -2 & 3 & 5 \\ 4 & 8 & -1 \end{bmatrix}.$$

А сейчас можно использовать метод исключения Гаусса без перестановки строк:

$$\begin{matrix}\text{гл. эл.} \rightarrow \\ m_{21} = -2 \\ m_{31} = 4\end{matrix} \begin{bmatrix} \underline{1} & 2 & 6 \\ -2 & 3 & 5 \\ 4 & 8 & -1 \end{bmatrix}.$$

После исключения элемента, стоящего на пересечении столбца 2 и строки 3, получим

$$\begin{matrix}\text{гл. эл.} \rightarrow \\ m_{32} = 0\end{matrix} \begin{bmatrix} 1 & 2 & 6 \\ 0 & \underline{7} & 17 \\ 0 & 0 & -25 \end{bmatrix} = U. \tag{18}$$

■

Расширение метода исключения Гаусса

Следующая теорема является обобщением теоремы 3.10, поскольку включает случай, требующий перестановки строк. Такое разложение на треугольные матрицы может быть использовано для нахождения решения любой линейной системы $AX = B$, где матрица A не вырождена.

Теорема 3.14 (непрямое разложение: $PA = LU$). Пусть задана матрица A размера $N \times N$. Предположим, что методом исключения Гаусса можно решить линейную систему общего вида $AX = B$, но требуются перестановки строк. Тогда существует такая матрица перестановок P, что произведение PA можно разложить на произведение нижней треугольной матрицы L и верхней треугольной матрицы U:

$$PA = LU.$$

Более того, матрицу L можно построить так, чтобы на главной диагонали стояли единицы, а на диагонали матрицы U были элементы, не равные нулю. Решение X находим за четыре шага.

1. Строим матрицы L, U и P.
2. Вычисляем вектор-столбец PB.
3. Используя прямую подстановку, решаем $LY = PB$ для Y.
4. Используя обратную подстановку, решаем $UX = Y$ для X.

Примечание. Предполагаем, что линейная система $AX = B$ имеет решение для фиксированной матрицы A и нескольких различных матриц столбцов B. Тогда шаг 1 выполняем только один раз и шаги 2–4 используем, чтобы найти решение X, соответствующее столбцу B. Шаги 2–4 для нахождения решения X вычисляются эффективным методом и требуют $O(N^2)$ операций вместо $O(N^3)$, необходимых для метода исключения Гаусса.

MATLAB

Команда MATLAB `[L,U,P]=lu(A)` строит нижнюю треугольную матрицу L, верхнюю треугольную матрицу U (для разложения A на треугольные матрицы) и матрицу перестановок P из теоремы 3.14.

Пример 3.25. Используем команду MATLAB `[L,U,P]=lu(A)` для матрицы A из примера 3.22. Убедимся, что $A = P^{-1}AU$ (эквивалентно тому, что $PA = LU$).

```
>>A=[1 2 6 ;4 8 -1;-2 3 -5];
>>[L,U,P]=lu(A)
L=
    1.0000  0       0
    -0.5000 1.0000  0
    0.2500  0       1.0000
U=
    4.0000 8.0000 -1.0000
    0      7.0000 4.5000
    0      0      6.2500
P=
   0 1 0
   0 0 1
   1 0 0
>>inv(P)*L*U
   1  2 6
   4  8 -1
   -2 3 5
```

■

Как отмечалось ранее, метод разложения на треугольные матрицы часто предпочитают методу исключения. К тому же его используют в программах MATLAB `inv(A)` и `det(A)`. Например, из курса линейной алгебры известно, что определитель невырожденной матрицы A равен $(-1)^q \det U$, где U — верхняя треугольная матрица разложения A и q — число переставленных строк, которые требуются для получения матрицы P из единичной матрицы I. Так как U — верхняя треугольная матрица, известно, что определитель U точно равен произведению элементов ее главной диагонали (см. теорему 3.6). Читателю следует убедиться, что в примере 3.25 $\det(A) = 175 = (-1)^2(175) = (-1)^2 \det(U)$.

В следующей программе осуществлен процесс, описанный в доказательстве теоремы 3.10. Это расширение программы 3.2, использующее частный выбор главного элемента. Перестановка строк, обусловленная частным выбором главного элемента, записана в матрицу R. Затем матрица R используется на шаге прямой подстановки для нахождения матрицы Y.

Программа 3.3 ($PA = LU$: разложение с выбором главного элемента). Построение решения линейной системы $\boldsymbol{AX} = \boldsymbol{B}$, где $\boldsymbol{A}$ — невырожденная матрица.

```
function X = lufact(A,B)
%Вход  - A - матрица размера N x N
%      - B - матрица размера N x 1
%Выход - X - матрица размера N x 1, содержащая решение AX = B
%Инициализация  X, Y, временное сохранение матрицы C и строк
%заданной матрицы перестановок R
   [N,N]=size(A);
   X=zeros(N,1);
   Y=zeros(N,1);
   C=zeros(1,N);
   R=1:N;
for p=1:N-1
%Находим главный элемент строки для столбца p
   [max1,j]=max(abs(A(p:N,p)));
%Меняем местами строки p и j
   C=A(p,:);
   A(p,:)=A(j+p-1,:);
   A(j+p-1,:)=C;
   d=R(p);
   R(p)=R(j+p-1);
   R(j+p-1)=d;
if A(p,p)==0
   'A вырождена. Нет единственного решения'
   break
end
%Вычисление множителя и размещение под диагональю матрицы A
   for k=p+1:N
      mult=A(k,p)/A(p,p);
      A(k,p) = mult;
      A(k,p+1:N)=A(k,p+1:N)-mult*A(p,p+1:N);
   end
end
%Решение для Y
Y(1) = B(R(1));
for k=2:N
   Y(k)= B(R(k))-A(k,1:k-1)*Y(1:k-1);
```

```
end
%Решение для X
X(N)=Y(N)/A(N,N);
for k=N-1:-1:1
   X(k)=(Y(k)-A(k,k+1:N)*X(k+1:N))/A(k,k);
end
```

Упражнения к разделу 3.5

1. Решите $\boldsymbol{LY} = \boldsymbol{B}, \boldsymbol{UX} = \boldsymbol{Y}$ и убедитесь, что $\boldsymbol{B} = \boldsymbol{AX}$ для **(a)** $\boldsymbol{B} = [-4 \quad 10 \quad 5]'$ и **(b)** $\boldsymbol{B} = [20 \quad 49 \quad 32]'$, где $\boldsymbol{A} = \boldsymbol{LU}$ имеет вид

$$\begin{bmatrix} 2 & 4 & -6 \\ 1 & 5 & 3 \\ 1 & 3 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 1/3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 & -6 \\ 0 & 3 & 6 \\ 0 & 0 & 3 \end{bmatrix}.$$

2. Решите $\boldsymbol{LY} = \boldsymbol{B}, \boldsymbol{UX} = \boldsymbol{Y}$ и проверьте, что $\boldsymbol{B} = \boldsymbol{AX}$ для **(a)** $\boldsymbol{B} = [7 \quad 2 \quad 10]'$ и **(b)** $\boldsymbol{B} = [23 \quad 35 \quad 7]'$, где $\boldsymbol{A} = \boldsymbol{LU}$ имеет вид

$$\begin{bmatrix} 1 & 1 & 6 \\ -1 & 2 & 9 \\ 1 & -2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 6 \\ 0 & 3 & 15 \\ 0 & 0 & 12 \end{bmatrix}.$$

3. Найдите разложение на треугольные матрицы $\boldsymbol{A} = \boldsymbol{LU}$ следующих матриц.

(a) $\begin{bmatrix} -5 & 2 & -1 \\ 1 & 0 & 3 \\ 3 & 1 & 6 \end{bmatrix}$ **(b)** $\begin{bmatrix} 1 & 0 & 3 \\ 3 & 1 & 6 \\ -5 & 2 & -1 \end{bmatrix}$

4. Найдите разложение $\boldsymbol{A} = \boldsymbol{LU}$ следующих матриц.

(a) $\begin{bmatrix} 4 & 2 & 1 \\ 2 & 5 & -2 \\ 1 & -2 & 7 \end{bmatrix}$ **(b)** $\begin{bmatrix} 1 & -2 & 7 \\ 4 & 2 & 1 \\ 2 & 5 & -2 \end{bmatrix}$

5. Найдите решение $\boldsymbol{LY} = \boldsymbol{B}, \boldsymbol{UX} = \boldsymbol{Y}$ и подтвердите, что $\boldsymbol{B} = \boldsymbol{AX}$ для **(a)** $\boldsymbol{B} = [8 \quad -4 \quad 10 \quad -4]'$ и **(b)** $\boldsymbol{B} = [28 \quad 13 \quad 23 \quad 4]'$, где $\boldsymbol{A} = \boldsymbol{LU}$ имеет вид

$$\begin{bmatrix} 4 & 8 & 4 & 0 \\ 1 & 5 & 4 & -3 \\ 1 & 4 & 7 & 2 \\ 1 & 3 & 0 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{4} & 1 & 0 & 0 \\ \frac{1}{4} & \frac{2}{3} & 1 & 0 \\ \frac{1}{4} & \frac{1}{3} & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 4 & 8 & 4 & 0 \\ 0 & 3 & 3 & -3 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

6. Найдите разложение на треугольные матрицы $\boldsymbol{A} = \boldsymbol{LU}$ для матрицы

$$\begin{bmatrix} 1 & 1 & 0 & 4 \\ 2 & -1 & 5 & 0 \\ 5 & 2 & 1 & 2 \\ -3 & 0 & 2 & 6 \end{bmatrix}.$$

7. Получите формулу (12) из этого раздела.
8. Покажите, что разложение на треугольные матрицы единственно в следующем смысле: если $\boldsymbol{A}$ не вырождена и $\boldsymbol{L}_1\boldsymbol{U}_1 = \boldsymbol{A} = \boldsymbol{L}_2\boldsymbol{U}_2$, то $\boldsymbol{L}_1 = \boldsymbol{L}_2$ и $\boldsymbol{U}_1 = \boldsymbol{U}_2$.
9. Докажите теорему 3.10 в случае, если $r > c$.
10. **(a)** Проверьте утверждение теоремы 3.12, показав, что $\boldsymbol{PP}' = \boldsymbol{I} = \boldsymbol{P}'\boldsymbol{P}$ для матрицы перестановок

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

(b) Докажите теорему 3.12. *Указание.* Воспользуйтесь определением умножения матриц и тем фактом, что каждая строка и столбец матриц $\boldsymbol{P}$ и $\boldsymbol{P}'$ содержат точно одну единицу.

11. Докажите, что матрица, обратная к невырожденной верхней треугольной матрице размера $N \times N$, является верхней треугольной матрицей.

Алгоритмы и программы

1. Используя программу 3.3, решите систему $\boldsymbol{AX} = \boldsymbol{B}$, где

$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & -1 & 3 & 5 \\ 0 & 0 & 2 & 5 \\ -2 & -6 & -3 & 1 \end{bmatrix} \quad \text{и} \quad \boldsymbol{B} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}.$$

Используйте команду MATLAB `[L,U,P]=lu(A)`, чтобы проверить свой ответ.

2. Воспользуйтесь программой 3.3, чтобы решить линейную систему $\boldsymbol{AX} = \boldsymbol{B}$, где $\boldsymbol{A} = [a_{ij}]_{N\times N}$, $a_{ij} = i^{j-1}$, и $\boldsymbol{B} = [b_{ij}]_{N\times 1}$, где $b_{11} = N$ и $b_{i1} = i^{N-2}/(i-1)$ для $i \geq 2$. Возьмите $N = 3, 7$ и 11. Точным решением является $\boldsymbol{X} = [1\ 1\ \ldots\ 1\ 1]'$. Объясните любые отклонения от точного решения.

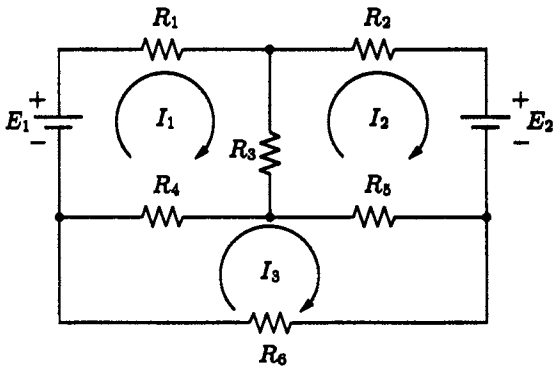


Рис. 3.5. Электрическая цепь для упр. 4

3. Модифицируйте программу 3.3 таким образом, чтобы она вычисляла A^{-1}, повторяя вычисления N линейных систем

$$AC_J = E_J \quad \text{для } J = 1,\ 2,\ \ldots,\ N.$$

Тогда

$$A\,[C_1\ \ C_2\ \ \ldots\ \ C_N] = [E_1\ \ E_2\ \ \ldots\ \ E_N]$$

и

$$A^{-1} = [C_1\ \ C_2\ \ \ldots\ \ C_N].$$

Убедитесь, что вы вычисляете разложение LU только один раз!

4. Закон Кирхгофа гласит, что сумма падения напряжения по любому замкнутому пути в сети в заданном направлении равна нулю. Когда этот закон применяется к цепи (рис. 3.5), мы получаем следующую систему линейных уравнений:

$$\begin{aligned} (R_1 + R_3 + R_4)I_1 + \quad & R_3 I_2 + \quad & R_4 I_3 &= E_1 \\ R_3 I_1 + (R_2 + R_3 + R_5)I_2 - \quad & & R_5 I_3 &= E_2 \\ R_4 I_1 - \quad & R_5 I_2 + (R_4 + R_5 + R_6)I_3 &= 0. \end{aligned} \tag{19}$$

Используйте программу 3.3, чтобы найти решение для токов I_1, I_2 и I_3, если

(a) $R_1 = 1$, $R_2 = 1$, $R_3 = 2$, $R_4 = 1$, $R_5 = 2$, $R_6 = 4$ и $E_1 = 23$, $E_2 = 29$;

(b) $R_1 = 1$, $R_2 = 0{,}75$, $R_3 = 1$, $R_4 = 2$, $R_5 = 1$, $R_6 = 4$ и $E_1 = 12$, $E_2 = 21{,}5$;

(c) $R_1 = 1$, $R_2 = 2$, $R_3 = 4$, $R_4 = 3$, $R_5 = 1$, $R_6 = 5$ и $E_1 = 41$, $E_2 = 38$.

5. Для вычисления следующего интеграла следует применить метод разложения на элементарные дроби:

$$\int \frac{x^2 + x + 1}{(x-1)(x-2)(x-3)^2(x^2+1)}\,dx.$$

Требуется найти коэффициенты A_i для $i = 1, 2, \ldots, 6$ в выражении

$$\frac{x^2 + x + 1}{(x-1)(x-2)(x-3)^2(x^2+1)} = \\ = \frac{A_1}{(x-1)} + \frac{A_2}{(x-2)} + \frac{A_3}{(x-3)^2} + \frac{A_4}{(x-3)} + \frac{A_5 x + A_6}{(x^2+1)}.$$

Воспользуйтесь программой 3.3, чтобы найти коэффициенты элементарных дробей.

6. Используя программу 3.3, решите линейную систему $\boldsymbol{AX} = \boldsymbol{B}$, где матрица $\boldsymbol{A}$ сгенерирована командами MATLAB `A=rand(10,10)` и `B=[1 2 3 ... 10]'`. Не забудьте проверить, не вырождена ли матрица $\boldsymbol{A}$ (`det(A)`$\neq 0$), прежде чем использовать программу 3.3. Проверьте точность своего ответа, формируя разницу матриц $\boldsymbol{AX} - \boldsymbol{B}$ и проверяя, насколько элементы этой матрицы близки к нулю (точный ответ дает $\boldsymbol{AX} - \boldsymbol{B} = \boldsymbol{0}$). Повторите этот процесс с матрицей $\boldsymbol{A}$, коэффициенты которой сгенерированы командами `A=rand(20,20)` и `B=[1 2 3 ... 20]'`. Объясните любые видимые расхождения в точности программы 3.3 на этих двух системах.

7. В выражении (8) раздела 3.1 определено понятие линейной комбинации в N-мерном пространстве. Например, вектор $(4, -3)$, который эквивалентен матрице $\begin{bmatrix} 4 & -3 \end{bmatrix}'$, можно записать как линейную комбинацию столбцов $\begin{bmatrix} 1 & 0 \end{bmatrix}'$ и $\begin{bmatrix} 0 & 1 \end{bmatrix}'$:

$$\begin{bmatrix} 4 \\ -3 \end{bmatrix} = 4 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + (-3) \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Используйте программу 3.3, чтобы показать, что матрицу $\begin{bmatrix} 1 & 3 & 5 & 7 & 9 \end{bmatrix}'$ можно записать как линейную комбинацию матриц столбцов

$$\begin{bmatrix} 0 \\ 4 \\ -2 \\ 3 \\ -1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \\ 4 \\ 4 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \\ 0 \\ 5 \\ 1 \end{bmatrix}, \begin{bmatrix} 5 \\ 6 \\ -3 \\ 0 \\ 2 \end{bmatrix} \text{ и } \begin{bmatrix} 1 \\ 4 \\ -2 \\ 7 \\ 0 \end{bmatrix}.$$

Объясните, почему любую матрицу $\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{bmatrix}'$ можно записать в виде линейной комбинации этих матриц.

3.6. Итеративные методы для линейных систем

В этом разделе речь пойдет о распространении некоторых итеративных методов, введенных в разделе 2, на системы больших размерностей. Рассмотрим расширение метода итерации неподвижной точки, которую применим к системам линейных уравнений.

Итерация Якоби

Пример 3.26. Рассмотрим систему линейных уравнений

$$\begin{aligned} 4x - y + z &= 7 \\ 4x - 8y + z &= -21 \\ -2x + y + 5z &= 15. \end{aligned} \tag{1}$$

Уравнения можно записать в виде

$$\begin{aligned} x &= \frac{7 + y - z}{4} \\ y &= \frac{21 + 4x + z}{8} \\ z &= \frac{15 + 2x - y}{5}. \end{aligned} \tag{2}$$

Это позволяет предложить следующий итеративный процесс Якоби:

$$\begin{aligned} x_{k+1} &= \frac{7 + y_k - z_k}{4} \\ y_{k+1} &= \frac{21 + 4x_k + z_k}{8} \\ z_{k+1} &= \frac{15 + 2x_k - y_k}{5}. \end{aligned} \tag{3}$$

Покажем, что если начать с точки $\boldsymbol{P}_0 = (x_0; y_0; z_0) = (1; 2; 2)$, то итерация (3) сходится к решению $(2; 4; 3)$.

Подставим $x_0 = 1, y_0 = 2$ и $z_0 = 2$ в правую часть каждого уравнения в (3), чтобы получить новые значения:

$$\begin{aligned} x_1 &= \frac{7 + 2 - 2}{4} = 1{,}75 \\ y_1 &= \frac{21 + 4 + 2}{8} = 3{,}375 \\ z_1 &= \frac{15 + 2 - 2}{5} = 3{,}00. \end{aligned}$$

Новая точка $\boldsymbol{P}_1 = (1{,}75; 3{,}375; 3{,}00)$ ближе к $(2; 4; 3)$, чем $\boldsymbol{P}_0$. Итерация, использующая (3), генерирует последовательность точек $\{\boldsymbol{P}_k\}$, которая сходится к решению $(2; 4; 3)$ (табл. 3.2). ■

Этот процесс называется ***итерацией Якоби*** и может использоваться для решения определенных типов линейных систем. После 19-и шагов итерация сходится к приближению с девятью значащими цифрами (2,00000000; 4,00000000; 3,00000000).

Таблица 3.2. Сходимость итерации Якоби для систем линейных уравнений (1)

k	x_k	y_k	z_k
0	1,0	2,0	2,0
1	1,75	3,375	3,0
2	1,84375	3,875	3,025
3	1,9625	3,925	2,9625
4	1,99062500	3,97656250	3,00000000
5	1,99414063	3,99531250	3,00093750
⋮	⋮	⋮	⋮
15	1,99999993	3,99999985	2,99999993
⋮	⋮	⋮	⋮
19	2,00000000	4,00000000	3,00000000

Линейные системы с таким большим количеством переменных, как 100 000, часто возникают в решениях дифференциальных уравнений в частных производных. Матрицы коэффициентов этих систем полупустые, т. е. большой процент элементов матрицы коэффициентов равен нулю. Если существует модель с не равными нулю элементами (т. е. трехдиагональная система), в таком случае итеративный процесс является эффективным методом решения этих больших систем.

Иногда метод Якоби не работает. Проведем эксперимент и увидим, что перестановка начальной линейной системы приведет к системе итерационных уравнений, которые дадут расходящуюся последовательность точек.

Пример 3.27. Пусть линейная система (1) переставлена следующим образом:

$$\begin{aligned} -2x + y + 5z &= 15 \\ 4x - 8y + z &= -21 \\ 4x - y + z &= 7. \end{aligned} \tag{4}$$

Полученные уравнения можно записать в виде

$$\begin{aligned} x &= \frac{-15 + y + 5z}{3} \\ y &= \frac{21 + 4x + z}{8} \\ z &= 7 - 4x + y. \end{aligned} \tag{5}$$

Это подразумевает следующий итеративный процесс Якоби:

$$
\begin{aligned}
x_{k+1} &= \frac{-15 + y_k + 5z_k}{3} \\
y_{k+1} &= \frac{21 + 4x_k + z_k}{8} \\
z_{k+1} &= 7 - 4x_k + y_k.
\end{aligned} \tag{6}
$$

Очевидно, что если начать с точки $\boldsymbol{P}_0 = (x_0; y_0; z_0) = (1; 2; 2)$, то итерация (6) отклоняется от решения $(2; 4; 3)$.

Подставим $x_0 = 1, y_0 = 2$ и $z_0 = 2$ в правую часть каждого уравнения (6), чтобы получить новые значения x_1, y_1 и z_1:

$$
\begin{aligned}
x_1 &= \frac{-15 + 2 + 10}{2} = -1{,}5 \\
y_1 &= \frac{21 + 4 + 2}{8} = 3{,}375 \\
z_1 &= 7 - 4 + 2 = 5{,}00.
\end{aligned}
$$

Новая точка $\boldsymbol{P}_1 = (-1{,}5; 3{,}375; 5{,}00)$ дальше от решения $(2; 4; 3)$, чем $\boldsymbol{P}_0$. Используемые итерационные уравнения (6) порождают расходящуюся последовательность (табл. 3.3). ■

Таблица 3.3. Расходящаяся итерация Якоби для линейной системы (4)

k	x_k	y_k	z_k
0	1,0	2,0	2,0
1	−1,5	3,375	5,0
2	6,6875	2,5	16,375
3	34,6875	8,015625	−17,25
4	−46,617188	17,8125	−123,73438
5	−307,929688	−36,150391	211,28125
6	502,62793	−124,929688	1202,56836
⋮	⋮	⋮	⋮

Итерация Гаусса–Зейделя

Иногда сходимость можно ускорить. Отметим, что итеративный процесс Якоби (3) производит три последовательности, $\{x_k\}, \{y_k\}$ и $\{z_k\}$, которые сходятся соответственно к 2, 4 и 3 (см. таблицу 3.2). Кажется разумным, что x_{k+1} может быть использовано вместо x_k в вычислении y_{k+1}. Аналогично x_{k+1} и y_{k+1} можно использовать в вычислении z_{k+1}. В следующем примере показано, что произойдет, когда применить это к уравнениям из примера 3.26.

Пример 3.28. Рассмотрим систему уравнений, заданную в (1), и итеративный процесс Гаусса–Зейделя, использующий (2):

$$(7)\qquad \begin{aligned} x_{k+1} &= \frac{7 + y_k - z_k}{4} \\ y_{k+1} &= \frac{21 + 4x_{k+1} + z_k}{8} \\ z_{k+1} &= \frac{15 + 2x_{k+1} - y_{k+1}}{5}. \end{aligned}$$

Ясно, что если начать с точки $\boldsymbol{P}_0 = (x_0; y_0; z_0) = (1; 2; 2)$, то итерация, заданная (7), сходится к решению $(2; 4; 3)$.

Подставим $y_0 = 2$ и $z_0 = 2$ в первое уравнение системы (7) и получим

$$x_1 = \frac{7 + 2 - 2}{4} = 1{,}75.$$

Затем подставим $x_1 = 1{,}75$ и $z_0 = 2$ во второе уравнение и получим

$$y_1 = \frac{21 + 4(1{,}75) + 2}{8} = 3{,}75.$$

Наконец подстановка $x_1 = 1{,}75$ и $y_1 = 3{,}75$ в третье уравнение даст

$$z_1 = \frac{15 + 2(1{,}75) - 3{,}75}{5} = 2{,}95.$$

Новая точка $\boldsymbol{P}_1 = (1{,}75; 3{,}75; 2{,}95)$ ближе к точке $(2; 4; 3)$, чем $\boldsymbol{P}_0$, и лучше, чем значение, заданное в примере 3.26. Итерация, использующая уравнения (7), генерирует последовательность $\{\mathbf{P}_k\}$, которая сходится к $(2; 4; 3)$ (см. табл. 3.4). ■

Таблица 3.4. Сходимость итерации Гаусса–Зейделя для системы (1)

k	x_k	y_k	z_k
0	1,0	2,0	2,0
1	1,75	3,75	2,95
2	1,95	3,96875	2,98625
3	1,995625	3,99609375	2,99903125
⋮	⋮	⋮	⋮
8	1,99999983	3,99999988	2,99999996
9	1,99999998	3,99999999	3,00000000
10	2,00000000	4,00000000	3,00000000

Принимая во внимание примеры 3.26 и 3.27, необходимо иметь некоторый критерий, определяющий, будет ли сходиться итерация Якоби. Поэтому дадим следующее определение.

Определение 3.6. Говорят, что матрица $\boldsymbol{A}$ размера $N \times N$ называется ***строго диагонально доминирующей***, если выполняется условие, что

$$|a_{kk}| > \sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{kj}| \quad \text{для } k = 1,\ 2,\ \ldots,\ N. \tag{8}$$

▲

Это означает, что в каждой строке матрицы величина элемента на главной диагонали должна превышать сумму величин всех остальных элементов в строке. Матрица коэффициентов линейной системы (1) из примера 3.26 строго диагонально доминирующая, потому что

$$\begin{aligned} &\text{В строке 1} && |4| > |-1| + |1|, \\ &\text{В строке 2} && |-8| > |4| + |1|, \\ &\text{В строке 3} && |5| > |-2| + |1|. \end{aligned}$$

Все строки удовлетворяют соотношению (8) определения 3.6; следовательно, матрица коэффициентов $\boldsymbol{A}$ линейной системы (1) строго диагонально доминирующая.

Матрица коэффициентов $\boldsymbol{A}$ линейной системы (4) из примера 3.27 не строго диагонально доминирующая, так как

$$\begin{aligned} &\text{В строке 1} && |-2| < |1| + |5|, \\ &\text{В строке 2} && |-8| > |4| + |1|, \\ &\text{В строке 3} && |1| < |4| + |-1|. \end{aligned}$$

Строки 1 и 3 не удовлетворяют соотношению (8) в определении 3.6, поэтому матрица коэффициентов $\boldsymbol{A}$ для линейной системы (4) не строго диагонально доминирующая.

А сейчас обобщим итерационные процессы Якоби и Гаусса–Зейделя. Предположим, что задана линейная система

$$\begin{array}{ccccccc} a_{11}x_1 + a_{12}x_2 & + \cdots + & a_{1j}x_j & + \cdots + & a_{1N}x_N & = b_1 \\ a_{21}x_1 + a_{22}x_2 & + \cdots + & a_{2j}x_j & + \cdots + & a_{2N}x_N & = b_2 \\ \vdots \quad \vdots & & \vdots & & \vdots & \vdots \\ a_{j1}x_1 + a_{j2}x_2 & + \cdots + & a_{jj}x_j & + \cdots + & a_{jN}x_N & = b_j \\ \vdots \quad \vdots & & \vdots & & \vdots & \vdots \\ a_{N1}x_1 + a_{N2}x_2 & + \cdots + & a_{Nj}x_j & + \cdots + & a_{NN}x_N & = b_N. \end{array} \tag{9}$$

Пусть k-я точка имеет вид $\boldsymbol{P}_k = (x_1^{(k)}, x_2^{(k)}, \ldots, x_j^{(k)}, \ldots, x_N^{(k)})$. Следующая точка, $(k+1)$-я, — $\boldsymbol{P}_{k+1} = (x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_j^{(k+1)}, \ldots, x_N^{(k+1)})$. Верхний индекс (k) координат точки $\boldsymbol{P}_k$ предоставляет возможность устанавливать координаты, которые относятся к этой точке. Итерационная формула использует j-ю строку (9), чтобы выразить $x_j^{(k+1)}$ в терминах линейной комбинации предыдущих значений $x_1^{(k)}, x_2^{(k)}, \ldots, x_j^{(k)}, \ldots, x_N^{(k)}$.

Итерация Якоби:

$$x_j^{(k+1)} = \frac{b_j - a_{j1}x_1^{(k)} - \cdots - a_{jj-1}x_{j-1}^{(k)} - a_{jj+1}x_{j+1}^{(k)} - \cdots - a_{jN}x_N^{(k)}}{a_{jj}} \tag{10}$$

для $j = 1, 2, \ldots, N$.

Итерация Якоби использует все старые координаты, чтобы получить все новые координаты, тогда как итерация Гаусса–Зейделя использует новые координаты по мере их появления.

Итерация Гаусса–Зейделя:

$$x_j^{(k+1)} = \frac{b_j - a_{j1}x_1^{(k+1)} - \cdots - a_{jj-1}x_{j-1}^{(k+1)} - a_{jj+1}x_{j+1}^{(k)} - \cdots - a_{jN}x_N^{(k)}}{a_{jj}} \tag{11}$$

для $j = 1, 2, \ldots, N$.

В следующей теореме приводятся эффективные условия для сходимости итерации Якоби.

Теорема 3.15 (итерация Якоби). Предположим, что $\boldsymbol{A}$ — строго диагонально доминирующая матрица. Тогда $\boldsymbol{AX} = \boldsymbol{B}$ имеет единственное решение $\boldsymbol{X} = \boldsymbol{P}$. Итерационная формула (10) порождает последовательность векторов $\{\boldsymbol{P}_k\}$, которые сходятся к $\boldsymbol{P}$ для любого выбора начального вектора $\boldsymbol{P}_0$.

Доказательство. Доказательство можно найти в книгах повышенной сложности по численному анализу. •

Можно доказать, что метод Гаусса–Зейделя также будет сходиться, когда $\boldsymbol{A}$ — строго диагонально доминирующая матрица. Во многих случаях метод Гаусса–Зейделя сходится быстрее, чем метод Якоби, поэтому предпочитают обычно его (ср. примеры 3.26 и 3.28). Важно понять, что достаточно небольшой модификации формулы (10), чтобы получить формулу (11). В некоторых случаях метод Якоби будет сходиться даже несмотря на то, что метод Гаусса–Зейделя не сходится.

Сходимость

Следует дать определение близости двух векторов так, чтобы можно было ввести понятие сходимости $\{\boldsymbol{P}_k\}$ к $\boldsymbol{P}$. Расстояние Евклида (см. раздел 3.1) между $\boldsymbol{P} = (x_1, x_2, \ldots, x_N)$ и $\boldsymbol{Q} = (y_1, y_2, \ldots, y_N)$ равно

$$\|\boldsymbol{P} - \boldsymbol{Q}\| = \left(\sum_{j=1}^{N} (x_j - y_j)^2 \right)^{1/2}. \tag{12}$$

Это неудобное определение, так как оно требует значительных усилий для вычислений. Поэтому введем другую норму $\|\boldsymbol{X}\|_1$:

$$\|\boldsymbol{X}\|_1 = \sum_{j=1}^{N} |x_j|. \tag{13}$$

Следующий результат гарантирует, что норма $\|\boldsymbol{X}\|_1$ имеет математическую структуру метрик и, следовательно, подходит для использования в качестве обобщения "формулы расстояния". Из курса линейной алгебры известно, что в векторном пространстве ограниченной размерности все нормы эквивалентны, т. е. если два вектора близки в норме $\|*\|_1$, они также близки в норме Евклида $\|*\|$.

Теорема 3.16. Пусть $\boldsymbol{X}$ и $\boldsymbol{Y}$ — векторы размера N и c — скаляр. Тогда функция $\|\boldsymbol{X}\|_1$ имеет следующие свойства:

$$\|\boldsymbol{X}\|_1 \geq 0, \tag{14}$$

$$\|\boldsymbol{X}\|_1 = 0 \quad \text{тогда и только тогда, когда} \quad \boldsymbol{X} = \boldsymbol{0}, \tag{15}$$

$$\|c\boldsymbol{X}\|_1 = |c| \, \|\boldsymbol{X}\|_1, \tag{16}$$

$$\|\boldsymbol{X} + \boldsymbol{Y}\|_1 \leq \|\boldsymbol{X}\|_1 + \|\boldsymbol{Y}\|_1. \tag{17}$$

Доказательство. Докажем (17) и оставим доказательство остальных свойств в качестве упражнений. Для каждого j неравенство треугольника для действительных чисел гласит, что $|x_j + y_j| \leq |x_j| + |y_j|$. Суммируя их, получаем неравенство (17):

$$\|\boldsymbol{X} + \boldsymbol{Y}\|_1 = \sum_{j=1}^{N} |x_j + y_j| \leq \sum_{j=1}^{N} |x_j| + \sum_{j=1}^{N} |y_j| = \|\boldsymbol{X}\|_1 + \|\boldsymbol{Y}\|_1.$$

Норму, заданную формулой (13), можно использовать, чтобы определять расстояние между точками.

Определение 3.7. Предположим, что $\boldsymbol{X}$ и $\boldsymbol{Y}$ — две точки в N-мерном пространстве. Определим расстояние между $\boldsymbol{X}$ и $\boldsymbol{Y}$ в норме $\|*\|_1$ как

$$\|\boldsymbol{X} - \boldsymbol{Y}\|_1 = \sum_{j=1}^{N} |x_j - y_j|. \qquad \blacktriangle$$

Пример 3.29. Определим расстояние Евклида и расстояние $\|*\|_1$ между точками $\boldsymbol{P} = (2; 4; 3)$ и $\boldsymbol{Q} = (1{,}75; 3{,}75; 2{,}95)$.

Расстояние Евклида равно

$$\|\boldsymbol{P} - \boldsymbol{Q}\| = ((2 - 1{,}75)^2 + (4 - 3{,}75)^2 + (3 - 2{,}95)^2)^{1/2} = 0{,}3570.$$

Расстояние $\|*\|_1$ равно

$$\|\boldsymbol{P} - \boldsymbol{Q}\|_1 = |2 - 1{,}75| + |4 - 3{,}75| + |3 - 2{,}95| = 0{,}55.$$

Расстояние $\|*\|_1$ легче вычисляется и используется для определения сходимости в N-мерном пространстве. ■

Команда MATLAB `A(j,[1:j-1,j+1:N])` используется в программе 3.4. Она эффективно выбирает все элементы в j-й строке матрицы $\boldsymbol{A}$, кроме элемента в j-м столбце (т. е. `A(j,j)`). Это замечание используется, чтобы упростить шаг (10) итерации Якоби в программе 3.4.

В обеих программах, 3.4 и 3.5, используется команда MATLAB `norm`, которая является нормой Евклида. Норму $\|*\|_1$ также можно использовать и читателю предлагаем проверить Help menu в MATLAB или одну из работ, указанных в списке литературы, чтобы получить информацию о команде `norm`.

Программа 3.4 (итерация Якоби). Решение линейной системы $\boldsymbol{AX} = \boldsymbol{B}$, начиная с исходного значения $\boldsymbol{X} = \boldsymbol{P}_0$, и генерирования последовательности $\{\boldsymbol{P}_k\}$, которая сходится к решению. Эффективным условием для применения метода является то, что $\boldsymbol{A}$ — строго диагонально доминирующая матрица.

```
function X=jacobi(A,B,P,delta, max1)
%Вход  - A - невырожденная матрица размера N x N
%      - B - матрица размера N x 1
%      - P - матрица размера N x 1, начальное предположение
%      - delta - допустимое отклонение для P
%      - max1 - максимальное число итераций
%Выход - X - матрица размера N x 1: приближение Якоби к решению
%        AX = B
N = length(B);
```

```
for k=1:max1
   for j=1:N
      X(j)=(B(j)-A(j,[1:j-1,j+1:N])*P([1:j-1,j+1:N]))/A(j,j);
   end
   err=abs(norm(X'-P));
   relerr=err/(norm(X)+eps);
   P=X';
      if(err<delta)|(relerr<delta)
      break
   end
end
X=X';
```

Программа 3.5 (итерация Гаусса–Зейделя). Решение линейной системы $\boldsymbol{AX} = \boldsymbol{B}$, начиная с исходного значения $\boldsymbol{X} = \boldsymbol{P}_0$ методом генерирования последовательности $\{\boldsymbol{P}_k\}$, которая сходится к решению. Эффективным условием для применения метода является то, что $\boldsymbol{A}$ — строго диагонально доминирующая матрица.

```
function X=gseid(A,B,P,delta, max1)
%Вход  - A - невырожденная матрица размера N x N
%      - B - матрица размера N x 1
%      - P - матрица размера N x 1, начальное предположение
%      - delta - допустимое отклонение для P
%      - max1 - максимльное число итераций
%Выход - X матрица размера N x 1: приближение Гаусса-Зейделя к
%        решению  AX = B
N = length(B);
for k=1:max1
   for j=1:N
      if j==1
         X(1)=(B(1)-A(1,2:N)*P(2:N))/A(1,1);
      elseif j==N
         X(N)=(B(N)-A(N,1:N-1)*(X(1:N-1))')/A(N,N);
      else
         %X содержит k-е приближение и P_{k-1}
         X(j)=(B(j)-A(j,1:j-1)*X(1:j-1)
               -A(j,j+1:N)*P(j+1:N))/A(j,j);
      end
   end
   err=abs(norm(X'-P));
```

```
   relerr=err/(norm(X)+eps);
   P=X';
      if(err<delta)|(relerr<delta)
         break
      end
end
X=X';
```

Упражнения к разделу 3.6

В упр. 1–8 поступайте следующим образом.

(a) Начните с $\boldsymbol{P}_0 = \mathbf{0}$ и используйте итерацию Якоби, чтобы найти $\boldsymbol{P}_k$ для $k = 1, 2, 3$. Будет ли итерация Якоби сходиться к решению?

(b) Начните с $\boldsymbol{P}_0 = \mathbf{0}$ и, используя итерацию Гаусса–Зейделя, найдите $\boldsymbol{P}_k$ для $k = 1, 2, 3$. Сходится ли итерация Гаусса–Зейделя к решению?

1.
$$\begin{aligned} 4x - y &= 15 \\ x + 5y &= 9 \end{aligned}$$

2.
$$\begin{aligned} 8x - 3y &= 10 \\ -x + 4y &= 6 \end{aligned}$$

3.
$$\begin{aligned} -x + 3y &= 1 \\ 6x - 2y &= 2 \end{aligned}$$

4.
$$\begin{aligned} 2x + 3y &= 1 \\ 7x - 2y &= 1 \end{aligned}$$

5.
$$\begin{aligned} 5x - y + z &= 10 \\ 2x + 8y - z &= 11 \\ -x + y + 4z &= 3 \end{aligned}$$

6.
$$\begin{aligned} 2x + 8y - z &= 11 \\ 5x - y + z &= 10 \\ -x + y + 4z &= 3 \end{aligned}$$

7.
$$\begin{aligned} x - 5y - z &= -8 \\ 4x + y - z &= 13 \\ 2x - y - 6z &= -2 \end{aligned}$$

8.
$$\begin{aligned} 4x + y - z &= 13 \\ x - 5y - z &= -8 \\ 2x - y - 6z &= -2 \end{aligned}$$

9. Пусть $\boldsymbol{X} = (x_1, x_2, \ldots, x_N)$. Докажите, что норма $\|*\|_1$

$$\|\boldsymbol{X}\|_1 = \sum_{k=1}^{N} |x_k|$$

удовлетворяет свойствам (14)–(16) теоремы 3.16.

10. Пусть $\boldsymbol{X} = (x_1, x_2, \ldots, x_N)$. Докажите, что норма Евклида

$$\|\boldsymbol{X}\| = \left(\sum_{k=1}^{N} (x_k)^2 \right)^{1/2}$$

удовлетворяет свойствам (14)–(17) теоремы 3.16.

11. Пусть $\boldsymbol{X} = (x_1, x_2, \ldots, x_N)$. Докажите, что норма $\|*\|_\infty$

$$\|\boldsymbol{X}\|_\infty = \max_{1 \le k \le N} |x_k|$$

удовлетворяет свойствам (14)–(17) теоремы 3.16.

Алгоритмы и программы

1. Используя программы 3.4 и 3.5, решите линейные системы упр. 1–8. Используйте команды `format long` и `delta` $= 10^{-9}$.

2. В теореме 3.14 условие, что $\boldsymbol{A}$ — строго диагональная доминантная матрица, достаточное, но не необходимое. Используйте обе программы, 3.4 и 3.5, и несколько различных начальных предположений для $\boldsymbol{P}_0$ в следующей линейной системе. *Примечание.* Итерация Якоби сходится, в то время как итерация Гаусса–Зейделя расходится.

$$\begin{aligned} x \quad\quad + z &= 2 \\ -x + y \quad\quad &= 0 \\ x + 2y - 3z &= 0 \end{aligned}$$

3. Рассмотрим следующую трехдиагональную систему и предположим, что матрица коэффициентов — строго диагональная доминантная матрица.

$$\begin{aligned} d_1x_1 + c_1x_2 &= b_1 \\ a_1x_1 + d_2x_2 + c_2x_3 &= b_2 \\ a_2x_2 + d_3x_3 + c_3x_4 &= b_3 \\ \vdots \quad\quad \vdots \quad\quad \vdots &\quad\quad \vdots \\ a_{N-2}x_{N-2} + d_{N-1}x_{N-1} + c_{N-1}x_N &= b_{N-1} \\ a_{N-1}x_{N-1} + d_Nx_N &= b_N. \end{aligned}$$

(i) Напишите алгоритм итерации, следуя (9)–(11), который решит эту систему. Ваш алгоритм должен эффективно использовать то, что матрица коэффициентов "полупустая".

(ii) Постройте MATLAB-программу, основанную на вашем алгоритме, и решите следующие трехдиагональные системы.

(a)
$$\begin{aligned}
4m_1 + m_2 &= 3\\
m_1 + 4m_2 + m_3 &= 3\\
m_2 + 4m_3 + m_4 &= 3\\
m_3 + 4m_4 + m_5 &= 3\\
&\vdots\\
m_{48} + 4m_{49} + m_{50} &= 3\\
m_{49} + 4m_{50} &= 3
\end{aligned}$$

(b)
$$\begin{aligned}
4m_1 + m_2 &= 1\\
m_1 + 4m_2 + m_3 &= 2\\
m_2 + 4m_3 + m_4 &= 1\\
m_3 + 4m_4 + m_5 &= 2\\
&\vdots\\
m_{48} + 4m_{49} + m_{50} &= 1\\
m_{49} + 4m_{50} &= 2
\end{aligned}$$

4. Используйте итерацию Гаусса–Зейделя, чтобы решить следующую полосовую систему.

$$\begin{aligned}
12x_1 - 2x_2 + x_3 &= 5\\
-2x_1 + 12x_2 - 2x_3 + x_4 &= 5\\
x_1 - 2x_2 + 12x_3 - 2x_4 + x_5 &= 5\\
x_2 - 2x_3 + 12x_4 - 2x_5 + x_6 &= 5\\
&\vdots\\
x_{46} - 2x_{47} + 12x_{48} - 2x_{49} + x_{50} &= 5\\
x_{47} - 2x_{48} + 12x_{49} - 2x_{50} &= 5\\
x_{48} - 2x_{49} + 12x_{50} &= 5
\end{aligned}$$

5. В программах 3.4 и 3.5 в качестве критерия останова используется относительная ошибка между двумя последовательными итерациями. Задачи с использованием исключительно этого критерия обсуждались в разделе 2.3. Линейную систему $\boldsymbol{AX} = \boldsymbol{B}$ можно записать как $\boldsymbol{AX} - \boldsymbol{B} = \boldsymbol{0}$. Если $\boldsymbol{X}_k$ – это k-я итерация для итерационного процесса Якоби или Гаусса–Зейделя, то норма ***разности*** $\boldsymbol{AX}_k - \boldsymbol{B}$ является, в общем, более подходящим критерием останова.

   Модифицируйте программы 3.4 и 3.5, чтобы они использовали разность в качестве критерия останова. Затем используйте модифицированные программы для решения полосовой системы из задачи 4.

3.7. Итерация для нелинейных систем: методы Ньютона и Зейделя (оптимальные)

В этом разделе обсуждаются способы обобщения методов, описанных в главе 2 и в разделах 3.6, на случай систем нелинейных функций. Рассмотрим функции

$$\begin{aligned}
f_1(x; y) &= x^2 - 2x - y + 0{,}5\\
f_2(x; y) &= x^2 + 4y^2 - 4.
\end{aligned} \tag{1}$$

Мы ищем метод решения системы нелинейных уравнений

$$f_1(x;y)=0 \quad \text{и} \quad f_2(x;y)=0. \tag{2}$$

Уравнения $f_1(x;y)=0$ и $f_2(x;y)=0$ неявно определяют кривые на плоскости. Следовательно, решением системы (2) является точка $(p;q)$, в которой пересекаются две кривые (т. е. и $f_1(p;q)=0$, и $f_2(p;q)=0$). Кривые системы (1) хорошо известны:

$$\begin{aligned} x^2-2x+0{,}5&=0 \quad &&\text{— график параболы,}\\ x^2+4y^2-4&=0 \quad &&\text{— график эллипса.}\end{aligned} \tag{3}$$

Графики на рис. 3.6 показывают, что существуют две точки решения и что они находятся в окрестности точек $(-0{,}2;1{,}0)$ и $(1{,}9;0{,}3)$.

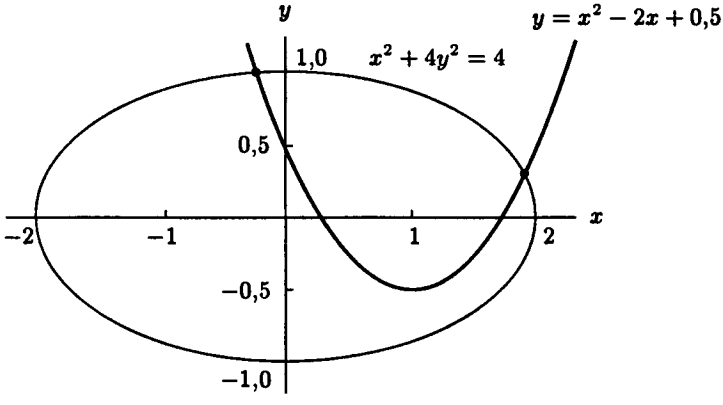


Рис. 3.6. Графики для нелинейной системы $y \doteq x^2 - 2x + 0{,}5$ и $x^2+4y^2=4$

Первый метод — это итерация неподвижной точки. Метод изобретен для генерирования последовательности $\{(p_k;q_k)\}$, которая сходится к решению $(p;q)$. Первое уравнение в (3) можно решить непосредственно относительно x. Однако к каждой части второго уравнения можно прибавить число, кратное y, чтобы получить $x^2+4y^2-8y-4=-8y$. Выбор величины $-8y$ является решающим и будет объяснен позже. А сейчас мы получили эквивалентную систему уравнений:

$$\begin{aligned} x&=\frac{x^2-y+0{,}5}{2}\\ y&=\frac{-x^2-4y^2+8y+4}{8}.\end{aligned} \tag{4}$$

Эти два уравнения используются для получения рекуррентных формул. В качестве начальной точки берем $(p_0;q_0)$ и затем вычисляем последовательность

Таблица 3.5. Итерация неподвижной точки, использующая формулы (5)

Случай (i). Начинаем с (0; 1)			Случай (ii). Начинаем с (2; 0)		
k	p_k	q_k	k	p_k	q_k
0	0,00	1,00	0	2,00	0,00
1	−0,25	1,00	1	2,25	0,00
2	−0,21875	0,9921875	2	2,78125	−0,1328125
3	−0,2221680	0,9939880	3	4,184082	−0,6085510
4	−0,2223147	0,9938121	4	9,307547	−2,4820360
5	−0,2221941	0,9938029	5	44,80623	−15,891091
6	−0,2222163	0,9938095	6	1 011,995	−392,60426
7	−0,2222147	0,9938083	7	512 263,2	−205 477,82
8	−0,2222145	0,9938084	Эта последовательность расходится.		
9	−0,2222146	0,9938084			

$\{(p_{k+1}; q_{k+1})\}$, используя соотношения

$$(5) \qquad \begin{aligned} p_{k+1} &= g_1(p_k; q_k) = \frac{p_k^2 - q_k + 0{,}5}{2} \\ q_{k+1} &= g_2(p_k; q_k) = \frac{-p_k^2 - 4q_k^2 + 8q_k + 4}{8}. \end{aligned}$$

Случай (i). Если в качестве начального значения выбрать $(p_0; q_0) = (0; 1)$, то

$$p_1 = \frac{0^2 - 1 + 0{,}5}{2} = -0{,}25 \quad \text{и} \quad q_1 = \frac{-0^2 - 4(1)^2 + 8(1) + 4}{8} = 1{,}0.$$

В табл. 3.5 приведена последовательность, генерируемая итерацией для случая (i). В этом случае последовательность сходится к решению, которое лежит около начального значения $(0; 1)$.

Случай (ii). Если в качестве начального значения выбрать $(p_0; q_0) = (2; 0)$, то

$$p_1 = \frac{2^2 - 0 + 0{,}5}{2} = 2{,}25 \quad \text{и} \quad q_1 = \frac{-2^2 - 4(0)^2 + 8(0) + 4}{8} = 0{,}0.$$

В табл. 3.5 приведена последовательность, генерируемая итерацией для случая (ii). В этом случае последовательность расходится.

Итерационные формулы (5) не могут быть использованы для нахождения второго решения $(1{,}900677; 0{,}3112186)$. Чтобы найти эту точку, необходима отличная от (5) пара итерационных формул. Возьмем уравнения (3), добавим $-2x$ к первому уравнению и $-11y$ ко второму уравнению и получим

$$x^2 - 4x - y + 0{,}5 = -2x \quad \text{и} \quad x^2 + 4y^2 - 11y - 4 = -11y.$$

Таблица 3.6. Итерация неподвижной точки, использующая формулы (6)

k	p_k	q_k
0	2,00	0,00
1	1,75	0,0
2	1,71875	0,0852273
3	1,753063	0,1776676
4	1,808345	0,2504410
8	1,903595	0,3160782
12	1,900924	0,3112267
16	1,900652	0,3111994
20	1,900677	0,3112196
24	1,900677	0,3112186

Затем используем эти уравнения, чтобы получить итерационные формулы

$$
\begin{aligned}
p_{k+1} &= g_1(p_k; q_k) = \frac{-p_k^2 + 4p_k + q_k - 0{,}5}{2} \\
q_{k+1} &= g_2(p_k; q_k) = \frac{-p_k^2 - 4q_k^2 + 11q_k + 4}{11}.
\end{aligned} \tag{6}
$$

В табл. 3.6 показан процесс использования (6) для нахождения второго решения.

Теория

Выясним, почему уравнения (6) подходят для того, чтобы найти поиска решения вблизи точки (1,9; 0,3), а уравнения (5) не подходят. В разделе 2.1 величина производной в неподвижной точке определяла, будет ли итерация сходиться. Когда применяется функция от нескольких переменных, тогда должны использоваться частные производные. Обобщением "производной" для систем функций от нескольких переменных является матрица Якоби. Рассмотрим лишь несколько предварительных идей, имеющих отношение к рассматриваемым задачам. Более подробное изложение можно найти в любой книге повышенной сложности по численным методам.

Определение 3.8 (матрица Якоби). Предположим, что $f_1(x; y)$ и $f_2(x; y)$ — функции от независимых переменных x и y, тогда их матрица Якоби $J(x; y)$ имеет вид

$$
\begin{bmatrix}
\dfrac{\partial f_1}{\partial x} & \dfrac{\partial f_1}{\partial y} \\
\dfrac{\partial f_2}{\partial x} & \dfrac{\partial f_2}{\partial y}
\end{bmatrix}. \tag{7}
$$

Аналогично, если $f_1(x;y;z)$, $f_2(x;y;z)$ и $f_3(x;y;z)$ — функции от независимых переменных $x;y$ и z, то их матрица Якоби $J(x;y;z)$ размера 3×3 определяется следующим образом:

$$\begin{bmatrix} \dfrac{\partial f_1}{\partial x} & \dfrac{\partial f_1}{\partial y} & \dfrac{\partial f_1}{\partial z} \\ \dfrac{\partial f_2}{\partial x} & \dfrac{\partial f_2}{\partial y} & \dfrac{\partial f_2}{\partial z} \\ \dfrac{\partial f_3}{\partial x} & \dfrac{\partial f_3}{\partial y} & \dfrac{\partial f_3}{\partial z} \end{bmatrix}. \tag{8}$$

▲

Пример 3.30. Найдем матрицу Якоби $J(x;y;z)$ размера 3×3 в точке $(1;3;2)$ для трех функций:

$$f_1(x;y;z) = x^3 - y^2 + y - z^4 + z^2$$
$$f_2(x;y;z) = xy + yz + xz$$
$$f_3(x;y;z) = \frac{y}{xz}.$$

Матрица Якоби имеет вид

$$J(x;y;z) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x} & \dfrac{\partial f_1}{\partial y} & \dfrac{\partial f_1}{\partial z} \\ \dfrac{\partial f_2}{\partial x} & \dfrac{\partial f_2}{\partial y} & \dfrac{\partial f_2}{\partial z} \\ \dfrac{\partial f_3}{\partial x} & \dfrac{\partial f_3}{\partial y} & \dfrac{\partial f_3}{\partial z} \end{bmatrix} = \begin{bmatrix} 3x^2 & -2y+1 & -4z^3+2z \\ y+z & x+z & y+x \\ \dfrac{-y}{x^2z} & \dfrac{1}{xz} & \dfrac{-y}{xz^2} \end{bmatrix}.$$

Таким образом, матрица Якоби в точке $(1;3;2)$ — это матрица размера 3×3:

$$J(1;3;2) = \begin{bmatrix} 3 & -5 & -28 \\ 5 & 3 & 4 \\ -\frac{3}{2} & \frac{1}{2} & -\frac{3}{4} \end{bmatrix}.$$

■

Дифференциал функции от нескольких переменных

Для функции от нескольких переменных дифференциал используется, чтобы показать, как повлияет изменение независимых переменных на зависимые переменные. Предположим, что есть

$$u = f_1(x;y;z), \quad v = f_2(x;y;z), \quad \text{и} \quad w = f_3(x;y;z). \tag{9}$$

Предположим также, что значения функций в (9) известны в точке $(x_0;y_0;z_0)$ и необходимо предугадать их значения в близкой точке $(x;y;z)$. Пусть du, dv и

dw обозначают дифференциалы зависимых переменных, а dx, dy и dz — дифференциалы независимых переменных. Справедливы соотношения

$$
\begin{aligned}
du &= \frac{\partial f_1}{\partial x}(x_0; y_0; z_0)\,dx + \frac{\partial f_1}{\partial y}(x_0; y_0; z_0)\,dy + \frac{\partial f_1}{\partial z}(x_0; y_0; z_0)\,dz,\\
dv &= \frac{\partial f_2}{\partial x}(x_0; y_0; z_0)\,dx + \frac{\partial f_2}{\partial y}(x_0; y_0; z_0)\,dy + \frac{\partial f_2}{\partial z}(x_0; y_0; z_0)\,dz,\\
dw &= \frac{\partial f_3}{\partial x}(x_0; y_0; z_0)\,dx + \frac{\partial f_3}{\partial y}(x_0; y_0; z_0)\,dy + \frac{\partial f_3}{\partial z}(x_0; y_0; z_0)\,dz.
\end{aligned}
\tag{10}
$$

Если использовать векторное обозначение, то с помощью матриц Якоби (10) можно записать сокращенно. Изменения функций обозначим, как $d\boldsymbol{F}$, а изменения переменных как — $d\boldsymbol{X}$:

$$
d\boldsymbol{F} = \begin{bmatrix} du \\ dv \\ dw \end{bmatrix} = \boldsymbol{J}(x_0; y_0; z_0) \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = \boldsymbol{J}(x_0; y_0; z_0)\; d\boldsymbol{X}. \tag{11}
$$

Пример 3.31. Используем матрицу Якоби, чтобы найти дифференциал $(du; dv; dw)$, когда независимые переменные из точки $(1; 3; 2)$ перемещаются в точку $(1{,}02; 2{,}97; 2{,}01)$ для системы функций

$$
\begin{aligned}
u &= f_1(x; y; z) = x^3 - y^2 + y - z^4 + z^2\\
v &= f_2(x; y; z) = xy + yz + xz\\
w &= f_3(x; y; z) = \frac{y}{xz}.
\end{aligned}
$$

Используем соотношение (11) с матрицей Якоби $\boldsymbol{J}(1; 3; 2)$ из примера 3.30 и дифференциал $(dx; dy; dz) = (0{,}02; -0{,}03; 0{,}01)$, чтобы получить

$$
\begin{bmatrix} du \\ dv \\ dw \end{bmatrix} = \begin{bmatrix} 3 & -5 & -28 \\ 5 & 3 & 4 \\ -\frac{3}{2} & \frac{1}{2} & -\frac{3}{4} \end{bmatrix} \begin{bmatrix} 0{,}02 \\ -0{,}03 \\ 0{,}01 \end{bmatrix} = \begin{bmatrix} -0{,}07 \\ 0{,}05 \\ -0{,}0525 \end{bmatrix}.
$$

Заметим, что значение функции в $(1{,}02; 2{,}97; 2{,}01)$ близко к линейному приближению, полученному посредством добавления дифференциалов $du = -0{,}07$; $dv = 0{,}05$ и $dw = -0{,}0525$ к соответствующим значениями функций $f_1(1; 3; 2) = -17$, $f_2(1; 3; 2) = 11$ и $f_3(1; 3; 2) = 1{,}5$, т. е.

$$
\begin{aligned}
f_1(1{,}02; 2{,}97; 2{,}01) &= -17{,}072 \approx -17{,}01 = f_1(1; 3; 2) + du\\
f_2(1{,}02; 2{,}97; 2{,}01) &= 11{,}0493 \approx 11{,}05 = f_2(1; 3; 2) + dv\\
f_3(1{,}02; 2{,}97; 2{,}01) &= 1{,}44864 \approx 1{,}4475 = f_3(1; 3; 2) + dw.
\end{aligned}
$$

■

Сходимость вблизи неподвижных точек

Расширим определения и теоремы из раздела 2.1 для функций от двух и трех переменных. Функции от N переменных рассматриваться не будут. Читатель легко найдет обобщения для этого случая во многих книгах по численному анализу.

Определение 3.9. ***Неподвижной точкой*** для системы двух уравнений от двух переменных

$$x = g_1(x; y) \quad \text{и} \quad y = g_2(x; y) \tag{12}$$

называется такая точка $(p; q)$, что $p = g_1(p; q)$ и $q = g_2(p; q)$. Аналогично для трех переменных неподвижной точкой для системы

$$x = g_1(x; y; z), \quad y = g_2(x; y; z) \quad \text{и} \quad z = g_3(x; y; z) \tag{13}$$

называется такая точка $(p; q; r)$, что $p = g_1(p; q; r)$, $q = g_2(p; q; r)$ и $r = g_3(p; q; r)$.▲

Определение 3.10. Для функций (12) ***итерация неподвижной точки*** определяется следующим образом:

$$p_{k+1} = g_1(p_k; q_k) \quad \text{и} \quad q_{k+1} = g_2(p_k; q_k) \tag{14}$$

для $k = 0, 1, \ldots$. Аналогично для функций (13) ***итерация неподвижной точки*** определяется следующим образом:

$$\begin{aligned} p_{k+1} &= g_1(p_k; q_k; r_k) \\ q_{k+1} &= g_2(p_k; q_k; r_k) \\ r_{k+1} &= g_3(p_k; q_k; r_k) \end{aligned} \tag{15}$$

для $k = 0, 1, \ldots$. ▲

Теорема 3.17 (итерация неподвижной точки). Предположим, что функции в (12) и (13) и их первые частные производные непрерывны в области, содержащей неподвижную точку $(p; q)$ или $(p; q; r)$ соответственно. Если выбрать начальную точку достаточно близко к неподвижной точке, то будет справедливым одно из следующих утверждений.

Случай (i). Размерность — два. Если точка $(p_0; q_0)$ достаточно близка к $(p; q)$ и если

$$\begin{aligned} \left|\frac{\partial g_1}{\partial x}(p; q)\right| + \left|\frac{\partial g_1}{\partial y}(p; q)\right| &< 1, \\ \left|\frac{\partial g_2}{\partial x}(p; q)\right| + \left|\frac{\partial g_2}{\partial y}(p; q)\right| &< 1, \end{aligned} \tag{16}$$

то итерация, заданная в (14), сходится к неподвижной точке $(p; q)$.

Случай (ii). Трехмерный. Если $(p_0; q_0; r_0)$ достаточно близка к точке $(p; q; r)$ и если

$$(17)\qquad \begin{aligned} \left|\frac{\partial g_1}{\partial x}(p;q;r)\right| + \left|\frac{\partial g_1}{\partial y}(p;q;r)\right| + \left|\frac{\partial g_1}{\partial z}(p;q;r)\right| &< 1,\\ \left|\frac{\partial g_2}{\partial x}(p;q;r)\right| + \left|\frac{\partial g_2}{\partial y}(p;q;r)\right| + \left|\frac{\partial g_2}{\partial z}(p;q;r)\right| &< 1,\\ \left|\frac{\partial g_3}{\partial x}(p;q;r)\right| + \left|\frac{\partial g_3}{\partial y}(p;q;r)\right| + \left|\frac{\partial g_3}{\partial z}(p;q;r)\right| &< 1, \end{aligned}$$

то итерация, заданная в (15), сходится к неподвижной точке $(p; q; r)$.

Если условия (16) или (17) не выполняются, итерация может расходиться. Обычно это происходит, когда сумма значений частных производных намного больше 1. Теорему 3.17 можно использовать, чтобы показать, почему итерация (5) сходится к неподвижной точке вблизи точки $(-0{,}2; 1{,}0)$. Частные производные равны

$$\begin{aligned} \frac{\partial}{\partial x} g_1(x;y) &= x, & \frac{\partial}{\partial y} g_1(x;y) &= -\frac{1}{2},\\ \frac{\partial}{\partial x} g_2(x;y) &= -\frac{x}{4}, & \frac{\partial}{\partial y} g_2(x;y) &= -y + 1. \end{aligned}$$

Поэтому для всех (x, y), удовлетворяющих неравенствам $-0{,}5 < x < 0{,}5$ и $0{,}5 < y < 1{,}5$, частные производные удовлетворяют неравенствам

$$\begin{aligned} \left|\frac{\partial}{\partial x} g_1(x;y)\right| + \left|\frac{\partial}{\partial y} g_1(x;y)\right| &= |x| + |-0{,}5| < 1,\\ \left|\frac{\partial}{\partial x} g_2(x;y)\right| + \left|\frac{\partial}{\partial y} g_2(x;y)\right| &= \frac{|-x|}{4} + |-y+1| < 0{,}625 < 1. \end{aligned}$$

Таким образом, условия (16) для частных производных выполняются и из теоремы 3.17 следует, что итерация неподвижной точки сходится к $(p; q) \approx$ $\approx(-0{,}2222146; 0{,}9938084)$. Отметим, что вблизи другой неподвижной точки $(1{,}90068; 0{,}31122)$ частные производные не удовлетворяют условиям (16), поэтому сходимость не гарантируется, т. е.

$$\begin{aligned} \left|\frac{\partial}{\partial x} g_1(1{,}90068; 0{,}31122)\right| + \left|\frac{\partial}{\partial y} g_1(1{,}90068; 0{,}31122)\right| &= 2{,}40068 > 1,\\ \left|\frac{\partial}{\partial x} g_2(1{,}90068; 0{,}31122)\right| + \left|\frac{\partial}{\partial y} g_2(1{,}90068; 0{,}31122)\right| &= 1{,}16395 > 1. \end{aligned}$$

Итерация Зейделя

Можно получить итерацию неподвижной точки, подобную методу Гаусса–Зейделя, для линейных систем. Предположим, что p_{k+1} используется для вычисления q_{k+1} (в трехмерном случае и p_{k+1} и q_{k+1} используются для вычисления r_{k+1}). Когда эти модификации включены в формулы (14) и (15), метод называется ***итерацией Зейделя***:

$$p_{k+1} = g_1(p_k; q_k) \quad \text{и} \quad q_{k+1} = g_2(p_{k+1}; q_k), \tag{18}$$

и

$$\begin{aligned} p_{k+1} &= g_1(p_k; q_k; r_k) \\ q_{k+1} &= g_2(p_{k+1}; q_k; r_k) \\ r_{k+1} &= g_3(p_{k+1}; q_{k+1}; r_k). \end{aligned} \tag{19}$$

Программа 3.6 выполняет итерацию Зейделя для нелинейных систем. Выполнение итерации неподвижной точки оставляем читателю.

Метод Ньютона для нелинейных систем

Изложим основные принципы метода Ньютона для двух переменных. Метод Ньютона можно легко расширить на большие размерности.

Систему

$$\begin{aligned} u &= f_1(x; y) \\ v &= f_2(x; y), \end{aligned} \tag{20}$$

можно рассматривать как преобразование плоскости xy в плоскость uv. Нас интересует поведение этого преобразования около точки $(x_0; y_0)$, образом которой является точка $(u_0; v_0)$. Если две функции имеют непрерывные частные производные, то можно использовать дифференциал, чтобы записать систему линейных приближений, которая справедлива в точке (x_0, y_0):

$$\begin{aligned} u - u_0 &= \frac{\partial}{\partial x} f_1(x_0; y_0)(x - x_0) + \frac{\partial}{\partial y} f_1(x_0; y_0)(y - y_0), \\ v - v_0 &= \frac{\partial}{\partial x} f_2(x_0; y_0)(x - x_0) + \frac{\partial}{\partial y} f_2(x_0; y_0)(y - y_0). \end{aligned} \tag{21}$$

Система (21) — это локальное линейное преобразование, которое связывает небольшие изменения в независимых переменных с небольшими изменениями в зависимых переменных. Если использовать матрицу Якоби $\boldsymbol{J}(x_0; y_0)$, то эти соотношения можно легко представить в виде

$$\begin{bmatrix} u - u_0 \\ v - v_0 \end{bmatrix} = \begin{bmatrix} \dfrac{\partial}{\partial x} f_1(x_0; y_0) & \dfrac{\partial}{\partial y} f_1(x_0; y_0) \\ \dfrac{\partial}{\partial x} f_2(x_0; y_0) & \dfrac{\partial}{\partial y} f_2(x_0; y_0) \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}. \tag{22}$$

Если система (20) записана как вектор-функция $\boldsymbol{V} = \boldsymbol{F}(\boldsymbol{X})$, матрица Якоби $\boldsymbol{J}(x; y)$ — это двухмерный аналог производной, потому что (22) можно записать как

$$\Delta \boldsymbol{F} \approx \boldsymbol{J}(x_0; y_0)\, \Delta \boldsymbol{X}. \tag{23}$$

А сейчас для описания метода Ньютона в случае двух переменных воспользуемся выражением (23).

Рассмотрим систему (20), приравняв u и v к нулю:

$$\begin{aligned} 0 &= f_1(x; y) \\ 0 &= f_2(x; y). \end{aligned} \tag{24}$$

Предположим, что $(p; q)$ — решение (24), т. е.

$$\begin{aligned} 0 &= f_1(p; q) \\ 0 &= f_2(p; q). \end{aligned} \tag{25}$$

Чтобы обобщить метод Ньютона для решения системы (24), необходимо рассмотреть небольшие изменения в функциях около точки $(p_0; q_0)$:

$$\begin{aligned} \Delta u &= u - u_0, \qquad \Delta p = x - p_0. \\ \Delta v &= v - v_0, \qquad \Delta q = y - q_0. \end{aligned} \tag{26}$$

Присвоим $(x; y) = (p; q)$ в (20) и используем (25), чтобы убедиться, что $(u; v) = (0; 0)$. Следовательно, изменения зависимых переменных равны

$$\begin{aligned} u - u_0 &= f_1(p, q) - f_1(p_0, q_0) = 0 - f_1(p_0, q_0) \\ v - v_0 &= f_2(p, q) - f_2(p_0, q_0) = 0 - f_2(p_0, q_0). \end{aligned} \tag{27}$$

Используем результаты (27) в (22), чтобы получить линейное преобразование

$$\begin{bmatrix} \dfrac{\partial}{\partial x} f_1(p_0; q_0) & \dfrac{\partial}{\partial y} f_1(p_0; q_0) \\ \dfrac{\partial}{\partial x} f_2(p_0; q_0) & \dfrac{\partial}{\partial y} f_2(p_0; q_0) \end{bmatrix} \begin{bmatrix} \Delta p \\ \Delta q \end{bmatrix} \approx - \begin{bmatrix} f_1(p_0; q_0) \\ f_2(p_0; q_0) \end{bmatrix}. \tag{28}$$

Если матрица Якоби $\boldsymbol{J}(p_0; q_0)$ в (28) не вырождена, то решение для системы $\Delta \boldsymbol{P} = \begin{bmatrix} \Delta p & \Delta q \end{bmatrix}' = \begin{bmatrix} p & q \end{bmatrix}' - \begin{bmatrix} p_0 & q_0 \end{bmatrix}'$ можно записать в виде

$$\Delta \boldsymbol{P} \approx -\boldsymbol{J}(p_0; q_0)^{-1} \boldsymbol{F}(p_0; q_0). \tag{29}$$

Тогда следующее приближение $\boldsymbol{P}_1$ к решению $\boldsymbol{P}$ имеет вид

$$\boldsymbol{P}_1 = \boldsymbol{P}_0 + \Delta \boldsymbol{P} = \boldsymbol{P}_0 - \boldsymbol{J}(p_0; q_0)^{-1} \boldsymbol{F}(p_0; q_0). \tag{30}$$

Отметим, что (30) является обобщением метода Ньютона для одной переменной, т. е. $p_1 = p_0 - f(p_0)/f'(p_0)$.

Схема метода Ньютона

Предположим, что $\boldsymbol{P}_k$ получена.

Шаг 1. Вычисление функции

$$\boldsymbol{F}(\boldsymbol{P}_k) = \begin{bmatrix} f_1(p_k; q_k) \\ f_2(p_k; q_k) \end{bmatrix}.$$

Шаг 2. Вычисление матрицы Якоби

$$\boldsymbol{J}(\boldsymbol{P}_k) = \begin{bmatrix} \dfrac{\partial}{\partial x} f_1(p_k; q_k) & \dfrac{\partial}{\partial y} f_1(p_k; q_k) \\ \dfrac{\partial}{\partial x} f_2(p_k; q_k) & \dfrac{\partial}{\partial y} f_2(p_k; q_k) \end{bmatrix}.$$

Шаг 3. Решение линейной системы

$$\boldsymbol{J}(\boldsymbol{P}_k)\Delta\boldsymbol{P} = -\boldsymbol{F}(\boldsymbol{P}_k) \quad \text{для} \quad \Delta\boldsymbol{P}.$$

Шаг 4. Вычисление следующей точки:

$$\boldsymbol{P}_{k+1} = \boldsymbol{P}_k + \Delta\boldsymbol{P}.$$

Повтор процесса.

Пример 3.32. Рассмотрим нелинейную систему

$$\begin{aligned} 0 &= x^2 - 2x - y + 0{,}5 \\ 0 &= x^2 + 4y^2 - 4. \end{aligned}$$

Воспользуемся методом Ньютона с начальным значением $(p_0; q_0) = (2{,}00; 0{,}25)$ и вычислим $(p_1; q_1)$, $(p_2; q_2)$ и $(p_3; q_3)$.

Вектор-функция и матрица Якоби имеют вид

$$\boldsymbol{F}(x; y) = \begin{bmatrix} x^2 - 2x - y + 0{,}5 \\ x^2 + 4y^2 - 4 \end{bmatrix}, \quad \boldsymbol{J}(x; y) = \begin{bmatrix} 2x - 2 & -1 \\ 2x & 8y \end{bmatrix}.$$

В точке $(2{,}00; 0{,}25)$ они принимают значения

$$\boldsymbol{F}(2{,}00; 0{,}25) = \begin{bmatrix} 0{,}25 \\ 0{,}25 \end{bmatrix}, \quad \boldsymbol{J}(2{,}00; 0{,}25) = \begin{bmatrix} 2{,}0 & -1{,}0 \\ 4{,}0 & 2{,}0 \end{bmatrix}.$$

Дифференциалы Δp и Δq являются решениями линейной системы

$$\begin{bmatrix} 2{,}0 & -1{,}0 \\ 4{,}0 & 2{,}0 \end{bmatrix} \begin{bmatrix} \Delta p \\ \Delta q \end{bmatrix} = - \begin{bmatrix} 0{,}25 \\ 0{,}25 \end{bmatrix}.$$

Таблица 3.7. Значения функции, матрицы Якоби и дифференциалы, необходимые на каждой итерации метода Ньютона при решении примера 3.32

$\boldsymbol{P}_k$	Решение линейной системы $J(\boldsymbol{P}_k)\Delta\boldsymbol{P} = -\boldsymbol{F}(\boldsymbol{P}_k)$	$\boldsymbol{P}_k + \Delta\boldsymbol{P}$
$\begin{bmatrix}2{,}00\\0{,}25\end{bmatrix}$	$\begin{bmatrix}2{,}0 & -1{,}0\\4{,}0 & 2{,}0\end{bmatrix}\begin{bmatrix}-0{,}09375\\0{,}0625\end{bmatrix} = -\begin{bmatrix}0{,}25\\0{,}25\end{bmatrix}$	$\begin{bmatrix}1{,}90625\\0{,}3125\end{bmatrix}$
$\begin{bmatrix}1{,}90625\\0{,}3125\end{bmatrix}$	$\begin{bmatrix}1{,}8125 & -1{,}0\\3{,}8125 & 2{,}5\end{bmatrix}\begin{bmatrix}-0{,}005559\\-0{,}001287\end{bmatrix} = -\begin{bmatrix}0{,}008789\\0{,}024414\end{bmatrix}$	$\begin{bmatrix}1{,}900691\\0{,}311213\end{bmatrix}$
$\begin{bmatrix}1{,}900691\\0{,}311213\end{bmatrix}$	$\begin{bmatrix}1{,}801381 & -1{,}000000\\3{,}801381 & 2{,}489700\end{bmatrix}\begin{bmatrix}-0{,}000014\\0{,}000006\end{bmatrix} = -\begin{bmatrix}0{,}000031\\0{,}000038\end{bmatrix}$	$\begin{bmatrix}1{,}900677\\0{,}311219\end{bmatrix}$

Простое вычисление показывает, что

$$\Delta\boldsymbol{P} = \begin{bmatrix}\Delta p\\\Delta q\end{bmatrix} = \begin{bmatrix}-0{,}09375\\0{,}0625\end{bmatrix}.$$

Следующая точка итерации равна

$$\boldsymbol{P}_1 = \boldsymbol{P}_0 + \Delta\boldsymbol{P} = \begin{bmatrix}2{,}00\\0{,}25\end{bmatrix} + \begin{bmatrix}-0{,}09375\\0{,}0625\end{bmatrix} = \begin{bmatrix}1{,}90625\\0{,}3125\end{bmatrix}.$$

Аналогично следующими двумя точками будут

$$\boldsymbol{P}_2 = \begin{bmatrix}1{,}900691\\0{,}311213\end{bmatrix} \quad \text{и} \quad \boldsymbol{P}_3 = \begin{bmatrix}1{,}900677\\0{,}311219\end{bmatrix}.$$

Координаты $\boldsymbol{P}_3$ вычислены с точностью до шести десятичных знаков. Вычисления для $\boldsymbol{P}_2$ и $\boldsymbol{P}_3$ приведены в табл. 3.7. ■

При выполнении метода Ньютона может потребоваться вычислить несколько частных производных. Допускается использование численных приближений для значений этих частных производных, но нужна осторожность при выборе следующего шага. При больших размерностях это необходимо, чтобы использовать методы для решения линейных систем, включенные ранее в этот раздел для решения $\Delta\boldsymbol{P}$.

MATLAB

Программы 3.6 (нелинейная итерация Зейделя) и 3.7 (метод Ньютона–Рафсона) требуют записи нелинейных систем $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{X})$ и $\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{0}$ и их матриц Якоби

JF соответственно в форме М-файлов. Например, рассмотрим запись нелинейной системы из примера 3.32 и соответствующей матрицы Якоби в форме М-файлов F.m и JF.m соответственно.

```
function Z=F(X)
x=X(1);y=X(2);
Z=zeros(1,2);
Z(1)=x^2-2*x-y+0.5;
Z(2)=x^2+4y^2-4;
```

```
function W=JF(X)
x=X(1);y=X(2);
W=[2*x-2 -1;2*x 8*y];
```

Функцию следует вычислять, используя стандартные команды MATLAB.

```
>>A=feval('F',[2.00 0.25])
A=
   0.2500 0.2500
>>V=JF([2.00 0.25])
B=
   2 -1
   4 2
```

Программа 3.6 (нелинейная итерация Зейделя). Решение нелинейной системы неподвижной точки $X = G(X)$ с заданным начальным приближением P_0 и генерированием последовательности $\{P_k\}$, которая сходится к решению P.

```
function [P,iter] = seidel(G,P,delta, max1)
%Вход  - G - нелинейная система, записанная в форме М-файла G.m
%      - P - начальное приближение к решению
%      - delta - грань ошибки
%      - max1 - число итераций
%Выход - P - приближение Зейделя к решению
%      - iter - число потребовавшихся итераций
N=length(P);
for k=1:max1
   X=P;
   % X k-oe приближение к решению
   for j=1:N
      A=feval('G',X);
      % Вывод членов X по мере их вычисления
      X(j)=A(j);
   end
   err=abs(norm(X-P));
   relerr=err/(norm(X)+eps);
   P=X;
```

```
   iter=k;
   if(err<delta)|(relerr<delta)
      break
   end
end
```

В следующей программе команда MATLAB command A\B используется для решения линейной системы $\boldsymbol{AX} = \boldsymbol{B}$ (см. Q=P-(J\Y')'). Вместо нее можно использовать приведенную ранее в этом разделе программу. Выбор соответствующей программы для решения линейной системы зависит от размера и характеристики матрицы Якоби.

Программа 3.7 (метод Ньютона–Рафсона). Решение нелинейной системы $\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{0}$ с одним заданным приближением $\boldsymbol{P}_0$ и генерированием последовательности $\{\boldsymbol{P}_k\}$, которая сходится к решению $\boldsymbol{P}$.

```
function [P,iter,err]=newdim(F,JF,P,delta,epsilon,max1)

%Вход  - F - система, записанная в форме M-файла F.m
%      - JF - матрица Якоби F, записанная в форме M-файла JF.M
%      - P - начальное приближение к решению
%      - delta - допустимое отклонение для P
%      - epsilon - допустимое отклонение для F(P)
%      - max1 - максимальное число итераций
%Выход - P - приближенное решение
%      - iter - число потребовавшихся итераций
%      - err - ошибка вычисления P

Y=feval(F,P);

for k=1:max1
   J=feval(JF,P);
   Q=P-(J\Y')';
   Z=feval(F,Q);
   err=norm(Q-P);
   relerr=err/(norm(Q)+eps);
   P=Q;
   Y=Z;
   iter=k;
   if (err<delta)|(relerr<delta)|(abs(Y)<epsilon)
      break
   end
end
```

Упражнения к разделу 3.7

1. Найдите аналитически неподвижную точку (или точки) для каждой из следующих систем.
   (a) $x = g_1(x; y) = x - y^2$
   $y = g_2(x; y) = -x + 6y$
   (b) $x = g_1(x; y) = (x^2 - y^2 - x - 3)/3$
   $y = g_2(x; y) = (-x + y - 1)/3$
   (c) $x = g_1(x; y) = \sin(y)$
   $y = g_2(x; y) = -6x + y$
   (d) $x = g_1(x; y; z) = 9 - 3y - 2z$
   $y = g_2(x; y; z) = 2 - x + z$
   $z = g_3(x; y; z) = -9 + 3x + 4y - z$
2. Найдите аналитически нуль (или нули) для каждой из следующих систем. Вычислите матрицу Якоби каждой системы в каждом нуле.
   (a) $0 = f_1(x; y) = 2x + y - 6$
   $0 = f_2(x; y) = x + 2y$
   (b) $0 = f_1(x; y) = 3x^2 + 2y - 4$
   $0 = f_2(x; y) = 2x + 2y - 3$
   (c) $0 = f_1(x; y) = 2x - 4\cos(y)$
   $0 = f_2(x; y) = 4x \sin(y)$
   (d) $0 = f_1(x; y; z) = x^2 + y^2 - z$
   $0 = f_2(x; y; z) = x^2 + y^2 + z^2 - 1$
   $0 = f_3(x; y; z) = x + y$
3. Найдите в плоскости xy такую область, в которой при условии, что $(p_0; q_0)$ находится в этой области, итерация неподвижной точки гарантированно сходится (используйте обоснование, подобное тому, которое следует из теоремы 3.17) для системы:

$$x = g_1(x; y) = (x^2 - y^2 - x - 3)/3$$
$$y = g_2(x; y) = (x + y + 1)/3.$$

4. Перепишите следующую линейную систему в форме системы неподвижной точки. Найдите такие ограничения для x, y и z, что итерация неподвижной точки гарантированно сойдется для любого начального значения $(p_0; q_0; r_0)$, которое удовлетворяет граничным условиям.

$$\begin{aligned} 6x + \ y + \ z &= 1 \\ x + 4y + \ z &= 2 \\ x + \ y + 5z &= 0 \end{aligned}$$

5. Для заданной нелинейной системы используйте начальное приближение $(p_0; q_0) = (1{,}1; 2{,}0)$ и вычислите следующие три приближения к неподвижной точке, используя **(a)** итерацию неподвижной точки и формулы (14), и **(b)** итерацию Зейделя, используя формулы (18).

$$x = g_1(x; y) = \frac{8x - 4x^2 + y^2 + 1}{8} \qquad \text{(гипербола)}$$

$$y = g_2(x; y) = \frac{2x - x^2 + 4y - y^2 + 3}{4} \qquad \text{(окружность)}.$$

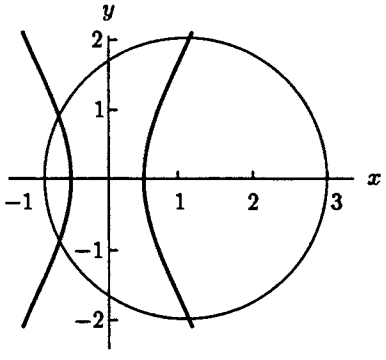


Рис. 3.7. Гипербола и окружность для упр. 5

6. Вычислите следующие три приближения к неподвижной точке, используя начальное приближение $(p_0; q_0) = (-0{,}3; -1{,}3)$, с помощью **(a)** итерации неподвижной точки и формул (14) и **(b)** итерации Зейделя и формул (18) для следующей нелинейной системы.

$$x = g_1(x; y) = \frac{y - x^3 + 3x^2 + 3x}{7} \qquad \text{(кривая третьего порядка)}$$

$$y = g_2(x; y) = \frac{y^2 + 2y - x - 2}{2} \qquad \text{(парабола)}.$$

7. Рассмотрим нелинейную систему

$$\begin{aligned} 0 &= f_1(x; y) = x^2 - y - 0{,}2 \\ 0 &= f_2(x; y) = y^2 - x - 0{,}3. \end{aligned}$$

Как показано на рис. 3.9, эти параболы пересекаются в двух точках.

(a) Начните с точки $(p_0; q_0) = (1{,}2; 1{,}2)$ и примените метод Ньютона для нахождения точек $(p_1; q_1)$ и $(p_2; q_2)$.

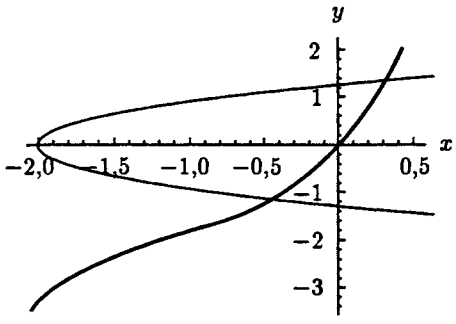


Рис. 3.8. Кубическая кривая и парабола для упр. 6

(b) Начните с точки $(p_0; q_0) = (-0{,}2; -0{,}2)$ и, применяя метод Ньютона, найдите $(p_1; q_1)$ и $(p_2; q_2)$.

8. Рассмотрим нелинейную систему, изображенную на рис. 3.10.

$$0 = f_1(x; y) = x^2 + y^2 - 2$$
$$0 = f_2(x; y) = xy - 1.$$

(a) Убедитесь, что решениями являются точки $(1; 1)$ и $(-1; -1)$.

(b) Какие могут возникнуть сложности, если попытаться найти решение методом Ньютона?

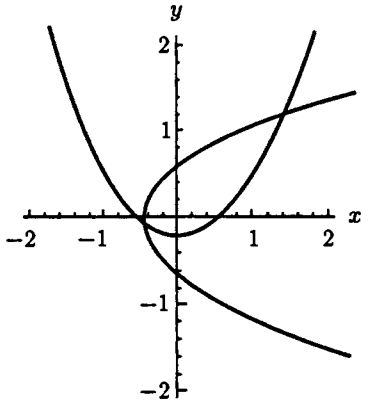


Рис. 3.9. Параболы для упр. 7

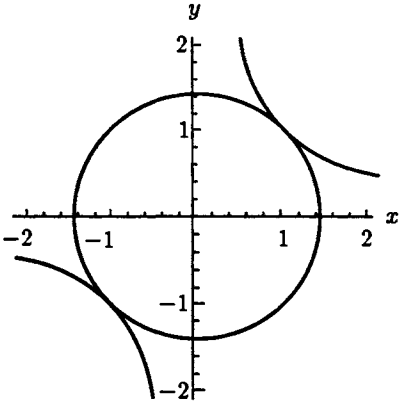


Рис. 3.10. Окружность и гипербола для упр. 8

9. Покажите, что итерация Якоби для линейной системы размера 3×6 является частным случаем итерации неподвижной точки (15). Кроме того, проверьте, что если матрица коэффициентов линейной системы размера 3×3 строго диагонально доминирующая, то выполняются условия (17) теоремы 3.17.

10. Покажите, что метод Ньютона для двух уравнений можно записать в форме итерации неподвижной точки

$$x = g_1(x;y), \quad y = g_2(x;y),$$

где $g_1(x;y)$ и $g_2(x;y)$ заданы выражениями

$$g_1(x;y) = x - \frac{f_1(x;y)\frac{\partial}{\partial y}f_2(x;y) - f_2(x;y)\frac{\partial}{\partial y}f_1(x;y)}{\det(\boldsymbol{J}(x;y))}$$

$$g_2(x;y) = y - \frac{f_2(x;y)\frac{\partial}{\partial x}f_1(x;y) - f_1(x;y)\frac{\partial}{\partial x}f_2(x;y)}{\det(\boldsymbol{J}(x;y))}.$$

11. Для решения нелинейной системы (12) использовалась итерация неподвижной точки. Используйте приведенные ниже шаги, чтобы доказать следующее: условий (16) достаточно для гарантии того, что $\{(p_k;q_k)\}$ сходится к $(p;q)$. Предположим, что существует такая константа K, $0 < K < 1$, что

$$\left|\frac{\partial}{\partial x}g_1(x;y)\right| + \left|\frac{\partial}{\partial y}g_1(x;y)\right| < K$$

и

$$\left|\frac{\partial}{\partial x}g_2(x;y)\right| + \left|\frac{\partial}{\partial y}g_2(x;y)\right| < K$$

для всех $(x;y)$, лежащих в прямоугольнике $R = \{(x;y) : a < x < b, c < y < d\}$. Предположим также, что $a < p_0 < b$ и $c < q_0 < d$. Определим

$$e_k = p - p_k, \qquad E_k = q - q_k, \qquad \text{и} \qquad r_k = \max\{|e_k|, |E_k|\}.$$

Используем следующую форму теоремы о среднем значении, примененную к функции двух переменных:

$$e_{k+1} = \frac{\partial}{\partial x}g_1(a_k^*;q_k)e_k + \frac{\partial}{\partial y}g_1(p;c_k^*)E_k,$$

$$E_{k+1} = \frac{\partial}{\partial x}g_2(b_k^*;q_k)e_k + \frac{\partial}{\partial y}g_2(p;d_k^*)E_k,$$

где a_k^* и b_k^* принадлежат $[a,b]$ и c_k^* и d_k^* лежат в интервале $[c;d]$. Докажите следующее.

(a) $|e_1| \le Kr_0$ и $|E_1| \le Kr_0$

(b) $|e_2| \le Kr_1 \le K^2r_0$ и $|E_2| \le Kr_1 \le K^2r_0$

(c) $|e_k| \le Kr_{k-1} \le K^kr_0$ и $|E_k| \le Kr_{k-1} \le K^kr_0$

(d) $\lim_{n\to\infty} p_k = p$ и $\lim_{n\to\infty} q_k = q$

12. Как отмечалось ранее, матрица Якоби системы (20) является двухмерным аналогом производной. Запишем систему (20) как вектор-функцию $\boldsymbol{V} = \boldsymbol{F}(\boldsymbol{X})$, и пусть $\boldsymbol{J}(\boldsymbol{F})$ — матрица Якоби этой системы. Заданы две нелинейные системы, $\boldsymbol{V} = \boldsymbol{F}(\boldsymbol{X})$, и $\boldsymbol{V} = \boldsymbol{G}(\boldsymbol{X})$ и действительное число c. Докажите следующее.
    (a) $\boldsymbol{J}(c\boldsymbol{F}(\boldsymbol{X})) = c\boldsymbol{J}(\boldsymbol{F}(\boldsymbol{X}))$
    (b) $\boldsymbol{J}(\boldsymbol{F}(\boldsymbol{X}) + \boldsymbol{G}(\boldsymbol{X})) = \boldsymbol{J}(\boldsymbol{F}(\boldsymbol{X})) + \boldsymbol{J}(\boldsymbol{G}(\boldsymbol{X}))$

Алгоритмы и программы

1. Воспользуйтесь программой 3.6, чтобы найти приближения к неподвижной точке систем из упр. 5 и 6. Ответ должен иметь точность 10 десятичных знаков.
2. Воспользуйтесь программой 3.7, чтобы найти приближения к нулям систем из упр. 7 и 8. Ответ должен иметь точность 10 десятичных знаков.
3. Постройте программу для нахождения неподвижной точки системы методом итерации неподвижной точки. Воспользуйтесь программой, чтобы найти приближения к неподвижной точке систем из упр. 5 и 6. Ответ должен иметь точность 8 десятичных знаков.
4. Воспользуйтесь программой 3.7, чтобы найти приближения к нулям следующих систем. Ответ должен иметь точность 10 десятичных знаков.

    (a) $0 = x^2 - x + y^2 + z^2 - 5$
    $0 = x^2 + y^2 - y + z^2 - 4$
    $0 = x^2 + y^2 + z^2 + z - 6$

    (b) $0 = x^2 - x + 2y^2 + yz - 10$
    $0 = 5x - 6y + z$
    $0 = z - x^2 - y^2$

    (c) $0 = (x+1)^2 + (y+1)^2 - z$
    $0 = (x-1)^2 + y^2 - z$
    $0 = 4x^2 + 2y^2 + z^2 - 16$

    (d) $0 = 9x^2 + 36y^2 + 4z^2 - 36$
    $0 = x^2 - 2y^2 - 20z$
    $0 = 16x - x^3 - 2y^2 - 16z^2$

5. Решите нелинейную систему

$$0 = 7x^3 - 10x - y - 1,$$
$$0 = 8y^3 - 11y + x - 1.$$

С помощью MATLAB нарисуйте графики обеих кривых в той же системе координат. Используйте график для проверки того, что существует 9 точек, в которых графики пересекаются. Используйте график, чтобы оценить точки пересечения. Используйте эти оценки и программу 3.7 для приближения точек пересечения с 9 десятичными знаками.

6. Систему из задачи 5 можно переписать в форме системы неподвижной точки:

$$x = \frac{7x^3 - y - 1}{10},$$
$$y = \frac{8y^3 + x - 1}{11}.$$

Проведем несколько экспериментов на компьютере и обнаружим, что безразлично, каким было начальное значение. Только одно из девяти решений можно найти, используя итерацию неподвижной точки (при этой особой форме метода неподвижной точки). Существует ли другая форма итерации неподвижной точки системы в 5, которую можно использовать для нахождения других решений системы?

8

Численная оптимизация

Двумерное волновое уравнение находит применение в механике при моделировании вибрирующей прямоугольной пластинки. Если закреплены все четыре края пластинки, то синусоидальные колебания описываются двухкратным рядом Фурье. Предположим, что в определенный момент высота $z = f(x, y)$ в точке

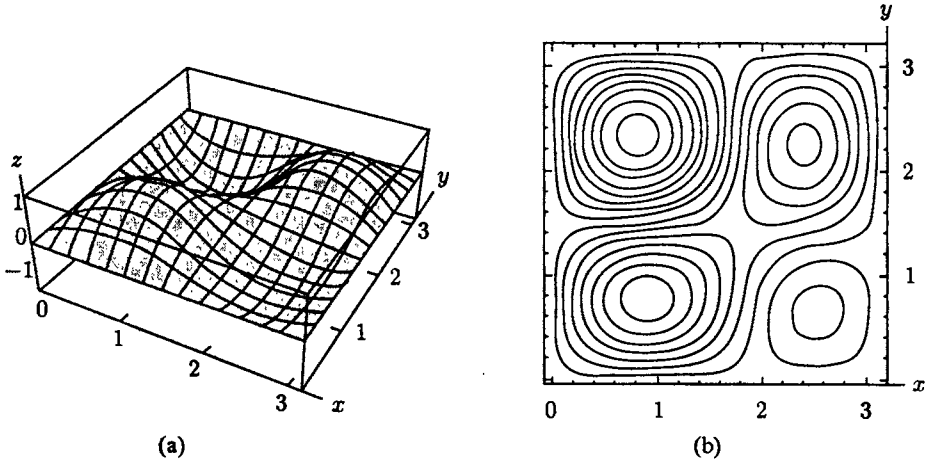


Рис. 8.1. (a) Смещение $z = f(x, y)$ вибрирующей пластинки. (b) Контурное изображение $f(x, y) = C$ вибрирующей пластинки

(x, y) задана функцией

$$z = f(x,y) = 0{,}02\sin(x)\sin(y) - 0{,}03\sin(2x)\sin(y) + \\ + 0{,}04\sin(x)\sin(2y) + 0{,}08\sin(2x)\sin(2y).$$

Где расположены точки с максимальным отклонением? Если рассматривать трехмерный график и соседнее с ним контурное изображение на рис. 8.1, можно увидеть, что существуют два локальных минимума и два локальных максимума на квадрате $0 \le x \le \pi$, $0 \le y \le \pi$. Численные методы используются для приближенного определения их места расположения:

$$f(0{,}8278; 2{,}3322) = -0{,}1200 \quad \text{и} \quad f(2{,}5351; 0{,}6298) = -0{,}0264$$

представляют собой локальные минимумы и

$$f(0{,}9241; 0{,}7640) = 0{,}0998 \quad \text{и} \quad f(2{,}3979; 2{,}2287) = 0{,}0853$$

являются локальные максимумами.

В этой главе кратко излагаются некоторые основные методы нахождения экстремума функций от одной или нескольких переменных.

8.1. Минимизация функции

Определение 8.1 (локальный экстремум). Говорят, что функция f имеет ***локальный минимум*** в точке $x = p$, если существует такой открытый интервал I, содержащий p, что $f(p) \le f(x)$ для всех $x \in I$. Аналогично говорят, что функция f имеет ***локальный максимум*** в точке $x = p$, если $f(x) \le f(p)$ для всех $x \in I$. Если функция f имеет или локальный минимум, или локальный максимум в точке $x = p$, то говорят, что она имеет ***локальный экстремум*** в точке $x = p$. ▲

Определение 8.2 (возрастание и убывание функции). Предположим, что функция $f(x)$ определена на интервале I.

(i) Если из неравенства $x_1 < x_2$ следует, что $f(x_1) < f(x_2)$ для всех $x_1, x_2 \in I$, то говорят, что функция f ***возрастает*** на интервале I.

(ii) Если из неравенства $x_1 < x_2$ вытекает, что $f(x_1) > f(x_2)$ для всех $x_1, x_2 \in I$, то говорят, что функция f ***убывает*** на интервале I. ▲

Теорема 8.1. Предположим, что функция $f(x)$ непрерывна на интервале $I = [a; b]$ и дифференцируема на (a, b).

(i) Если $f'(x) > 0$ для всех $x \in (a, b)$, то функция $f(x)$ является возрастающей на интервале I.

(ii) Если $f'(x) < 0$ для всех $x \in (a, b)$, то функция $f(x)$ является убывающей на интервале I.

Теорема 8.2. Предположим, что функция $f(x)$ определена на отрезке $I = [a; b]$ и имеет локальный экстремум во внутренней точке $p \in (a, b)$. Если $f(x)$ дифференцируема в точке $x = p$, то $f'(p) = 0$.

Теорема 8.3 (критерий первой производной). Предположим, что функция $f(x)$ непрерывна на отрезке $I = [a; b]$. Кроме того, предположим, что $f'(x)$ определена для всех $x \in (a, b)$ за исключением, возможно, точки $x = p$.

(i) Если $f'(x) < 0$ на интервале (a, p) и $f'(x) > 0$ на (p, b), то $f(p)$ — локальный минимум.

(ii) Если $f'(x) > 0$ на интервале (a, p) и $f'(x) < 0$ на (p, b), то $f(p)$ — локальный максимум.

Теорема 8.4 (критерий второй производной). Предположим, что функция f непрерывна на отрезке $[a; b]$ и f' и f'' определены на (a, b). Также предположим, что $p \in (a, b)$ — критическая точка, в которой $f'(p) = 0$.

(i) Если $f''(p) > 0$, то значение $f(p)$ является локальным минимумом f.

(ii) Если $f''(p) < 0$, значение $f(p)$ является локальным максимумом f.

(iii) Если $f''(p) = 0$, то этот критерий не является окончательным.

Пример 8.1. Используем критерий второй производной, чтобы классифицировать локальный экстремум функции $f(x) = x^3 + x^2 - x + 1$ на интервале $[-2; 2]$.

Первая производная равна $f'(x) = 3x^2 + 2x - 1 = (3x - 1)(x + 1)$, а вторая — $f''(x) = 6x + 2$. Существуют две точки, в которых $f'(x) = 0$ (т. е. $x = 1/3, -1$).

Случай (i). Находим, что в точке $x = 1/3$ $f'(1/3) = 0$ и $f''(1/3) = 4 > 0$; таким образом, функция $f(x)$ имеет локальный минимум в точке $x = 1/3$.

Случай (ii). Находим, что в точке $x = -1$ $f'(-1) = 0$ и $f''(-1) = -4 < 0$; таким образом, функция $f(x)$ имеет в точке $x = -1$ локальный максимум. ■

Метод поиска

Другой метод нахождения минимума функции $f(x)$ состоит в многократном вычислении функции и поиске локального минимума. Для уменьшения количества вычислений функции важно иметь хорошую стратегию, чтобы определить, где вычислять функцию $f(x)$. Одним из наиболее эффективных методов является *поиск методом золотого сечения*, который так назван из-за отношения, используемого при выборе точек.

Золотое сечение

Пусть $[0; 1]$ — начальный интервал. Если $0{,}5 < r < 1$, то $0 < 1 - r < 0{,}5$ и интервал делится на три подынтервала: $[0; 1 - r]$, $[1 - r; r]$ и $[r; 1]$. В процессе решения используется либо сжатие вправо и получение нового интервала $[0; r]$, либо сжатие влево и получение интервала $[1 - r; 1]$. Затем эти новые подинтервалы делятся на три подынтервала в таком же соотношении, как и интервал $[0; 1]$.

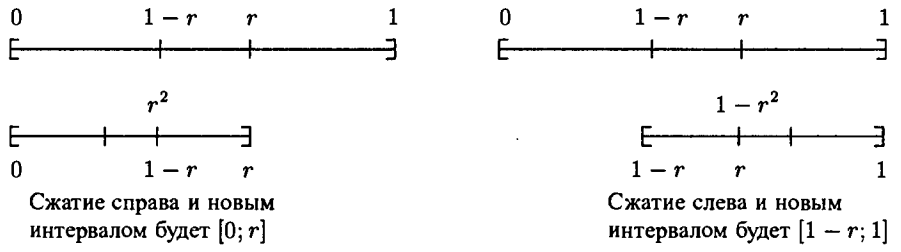


Рис. 8.2. Интервалы, которые используются при поиске методом золотого сечения

Требуется так выбрать r, чтобы одна из старых точек была в правильном положении относительно нового интервала, как показано на рис. 8.2. Из этого следует, что отношение $(1-r) : r$ такое же, как и $r : 1$. Следовательно, r удовлетворяет уравнению $1 - r = r^2$, которое можно записать в виде квадратного уравнения $r^2 + r - 1 = 0$. Решение r, удовлетворяющее неравенству $0{,}5 < r < 1$, равно $r = \left(\sqrt{5} - 1\right)/2$.

Функция $f(x)$ должна удовлетворять особым условиям, которые гарантируют существование истинного минимума на интервале, чтобы можно было использовать поиск минимума функции $f(x)$ методом золотого сечения.

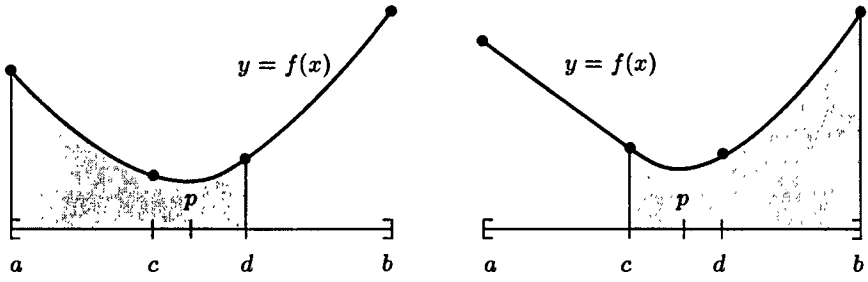
Определение 8.3 (унимодальная функция). Функция $f(x)$ является унимодальной на интервале $I = [a; b]$, если существует такое единственное число $p \in I$, что

(1) $f(x)$ убывает на $[a, p]$

(2) $f(x)$ возрастает на $[p, b]$. ▲

Если известно, что функция $f(x)$ унимодальна на интервале $[a; b]$, то можно заменить интервал подынтервалом, на котором функция $f(x)$ принимает минимальное значение. Для поиска методом золотого сечения требуется, чтобы использовались две внутренние точки, $c = a + (1 - r)(b - a)$ и $d = a + r(b - a)$, где r является золотым сечением, о котором упоминалось выше. Эти точки удовлетворяют неравенству $a < c < d < b$. Условие, что функция $f(x)$ унимодальна, гарантирует, что значения функции $f(c)$ и $f(d)$ меньше, чем $\max\{f(a), f(b)\}$. Рассмотрим два случая (рис. 8.3).

Если $f(c) \leq f(d)$, то минимум должен находиться на подынтервале $[a; d]$. Заменяем b на d и продолжаем поиск на новом подынтервале. Если $f(d) < f(c)$, то минимум должен находиться на подинтервале $[c; b]$. Заменяем a на c и продолжаем поиск. В следующем примере метод нахождения корня сравнивается с поиском методом золотого сечения.



Если $f(c) \le f(d)$, то сжимаем справа и используем $[a; d]$

Если $f(d) < f(c)$, то сжимаем слева и используем $[c; b]$

Рис. 8.3. Процесс решения для поиска методом золотого сечения

Пример 8.2. Найдем минимум унимодальной функции $f(x) = x^2 - \sin(x)$ на интервале $[0; 1]$.

Решение с помощью корня уравнения $f'(x) = 0$. Метод нахождения корня можно использовать для определения, где производная $f'(x) = 2x - \cos(x)$ равна нулю. Так как $f'(0) = -1$ и $f'(1) = 1{,}4596977$, корень $f'(x)$ находится на интервале $[0; 1]$. Начиная с точек $p_0 = 0$ и $p_1 = 1$, в табл. 8.1 приведены все итерации.

Если применить метод секущих, то можно получить $f'(0{,}4501836) = 0$. Вторая производная равна $f''(x) = 2 + \sin(x)$, и можно вычислить $f''(0{,}4501836) = 2{,}435131 > > 0$. Таким образом, минимальное значение равно $f(0{,}4501836) = -0{,}2324656$.

Решение, найденное методом золотого сечения. Сравниваем на каждом шаге значения функции $f(c)$ и $f(d)$ и принимаем решение продолжать поиск на интервале $[a; d]$ или $[c; b]$. Некоторые вычисления приведены в табл. 8.2.

На 23-й итерации интервал сужается до $[a_{23}; b_{23}] = [0{,}4501827; 0{,}4501983]$ длиной 0,0000156. Тем не менее полезно вычислять значения функции в крайних точках с девятью десятичными знаками (т. е. $f(a_{23}) \approx -0{,}23246558 \approx f(b_{23})$); по-

Таблица 8.1. Метод секущих для решения $f'(x) = 2x - \cos(x) = 0$

k	p_k	$2p_k - \cos(p_k)$
0	0,0000000	−1,00000000
1	1,0000000	1,45969769
2	0,4065540	−0,10538092
3	0,4465123	−0,00893398
4	0,4502137	0,00007329
5	0,4501836	−0,00000005

Таблица 8.2. Поиск минимума функции $f(x) = x^2 - \sin(x)$ методом золотого сечения

k	a_k	c_k	d_k	b_k	$f(c_k)$	$f(d_k)$
0	0,0000000	0,3819660	0,6180340	1	−0,22684748	−0,19746793
1	0,0000000	0,2360680	0,3819660	0,6180340	−0,17815339	−0,22684748
2	0,2360680	0,3819660	0,4721360	0,6180340	−0,22684748	−0,23187724
3	0,3819660	0,4721360	0,5278640	0,6180340	−0,23187724	−0,22504882
4	0,3819660	0,4376941	0,4721360	0,5278640	−0,23227594	−0,23187724
5	0,3819660	0,4164079	0,4376941	0,4721360	−0,23108238	−0,23227594
6	0,4164079	0,4376941	0,4508497	0,4721360	−0,23227594	−0,23246503
⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	0,4501574	0,4501730	0,4501827	0,4501983	−0,23246558	−0,23246558
22	0,4501730	0,4501827	0,4501886	0,4501983	−0,23246558	−0,23246558
23	0,4501827	0,4501886	0,4501923	0,4501983	−0,23246558	−0,23246558

этому происходит останов алгоритма. Проблемой в использовании метода поиска является то, что функция может быть плоской около минимума. Это ограничивает точность, которую можно получить. Метод секущих позволяет найти более точный ответ: $p_5 = 0{,}4501836$.

Хотя метод золотого сечения действует медленнее в этом примере, он имеет хорошее, будущее поскольку его можно применять в тех случаях, когда функция $f(x)$ не дифференцируема. ■

Нахождение экстремальных значений функции $f(x, y)$

Определение 8.1 легко обобщить для функций от нескольких переменных. Предположим, что $f(x, y)$ определена в области

$$R = \{(x, y) : (x - p)^2 + (y - q)^2 < r^2\}. \tag{3}$$

Функция $f(x, y)$ имеет локальный минимум в точке (p, q), если

$$f(p, q) \leq f(x, y) \quad \text{для каждой точки } (x, y) \in R. \tag{4}$$

Функция $f(x, y)$ имеет локальный максимум в точке (p, q), если

$$f(x, y) \leq f(p, q) \quad \text{для каждой точки } (x, y) \in R. \tag{5}$$

Критерий второй производной для экстремального значения является обобщением теоремы 8.4.

Теорема 8.5 (критерий второй производной). Предположим также, что функция $f(x,y)$ и ее первая и вторая частные производные непрерывны в области R. Предположим, что $(p,q) \in R$ — критическая точка, в которой и $f_x(p,q) = 0$, и $f_y(p,q) = 0$. Частные производные высшего порядка используются для определения природы критической точки.

(i) Если $f_{xx}(p,q)f_{yy}(p,q) - f_{xy}^2(p,q) > 0$ и $f_{xx}(p,q) > 0$, то $f(p,q)$ — локальный минимум функции f.

(ii) Если $f_{xx}(p,q)f_{yy}(p,q) - f_{xy}^2(p,q) > 0$ и $f_{xx}(p,q) < 0$, то $f(p,q)$ — локальный максимум функции f.

(iii) Если $f_{xx}(p,q)f_{yy}(p,q) - f_{xy}^2(p,q) < 0$, то функция $f(x,y)$ не имеет локального экстремума в точке (p,q).

(iv) Если $f_{xx}(p,q)f_{yy}(p,q) - f_{xy}^2(p,q) = 0$, этот критерий не является окончательным.

Пример 8.3. Найдем минимум функции $f(x,y) = x^2 - 4x + y^2 - y - xy$.

Частные производные первого порядка равны

$$f_x(x,y) = 2x - 4 - y \quad \text{и} \quad f_y(x,y) = 2y - 1 - x. \tag{6}$$

Приравняем эти частные производные к нулю и получим систему линейных уравнений

$$\begin{aligned} 2x - y &= 4, \\ -x + 2y &= 1. \end{aligned} \tag{7}$$

Решением системы (7) будет $(x,y) = (3,2)$. Частные производные второго порядка функции $f(x,y)$ имеют вид

$$f_{xx}(x,y) = 2, \quad f_{yy}(x,y) = 2 \quad \text{и} \quad f_{xy}(x,y) = -1.$$

Легко видеть, что это случай (i) теоремы 8.5, т. е.

$$f_{xx}(3,2)f_{yy}(3,2) - f_{xy}^2(3,2) = 3 > 0 \quad \text{и} \quad f_{xx}(3,2) = 2 > 0.$$

Следовательно, $f(x,y)$ имеет локальный минимум $f(3,2) = -7$ в точке $(3,2)$. ∎

Метод Нелдера–Мида

Симплекс-метод нахождения локального минимума функции от нескольких переменных изобретен Нелдером и Мидом. Для двух переменных симплексом является треугольник, и метод — это схема поиска, который сравнивает значения функции в трех вершинах треугольника. Наихудшая вершина, в которой функция $f(x,y)$ принимает наибольшее значение, отбрасывается и заменяется новой вершиной. Формируется новый треугольник, и поиск продолжается. При этом

строится последовательность треугольников (они могут иметь различную форму), значения функции в вершинах которой становятся все меньше и меньше. Уменьшается размер треугольника, и координаты точки минимума найдены.

В формулировке алгоритма используется термин "симплекс" (обобщенный N-мерный треугольник). С его помощью находим минимум функции от N переменных. Он эффективен и компактен при вычислении.

Исходный треугольник BGW

Предположим, что нужно минимизировать функцию $f(x, y)$. Для начала зададим три вершины треугольника: $V_k = (x_k, y_k)$, $k = 1, 2, 3$. Вычислим значения функции $f(x, y)$ в каждой из трёх точек $z_k = f(x_k, y_k)$, $k = 1, 2, 3$. Упорядочим индексы таким образом, чтобы $z_1 \le z_2 \le z_3$. Чтобы запомнить, что $\boldsymbol{B}$ — наилучшая вершина, $\boldsymbol{G}$ — хорошая (следует за наилучшей) и $\boldsymbol{W}$ — нахудшая вершина, введем обозначения

$$\boldsymbol{B} = (x_1, y_1), \quad \boldsymbol{G} = (x_2, y_2) \quad \text{и} \quad \boldsymbol{W} = (x_3, y_3) \tag{8}$$

Средняя точка хорошей стороны

При построении используется средняя точка отрезка, соединяющего вершины $\boldsymbol{B}$ и $\boldsymbol{G}$. Находим ее посредством усреднения координат:

$$\boldsymbol{M} = \frac{\boldsymbol{B} + \boldsymbol{G}}{2} = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right). \tag{9}$$

Отражение, использующее точку R

Функция убывает при движении вдоль стороны треугольника от вершины $\boldsymbol{W}$ к вершине $\boldsymbol{B}$ так же, как при движении вдоль стороны от вершины $\boldsymbol{W}$ к $\boldsymbol{G}$. Следовательно, существует возможность, что функция $f(x, y)$ принимает наименьшие значения в точках, которые лежат вдали от вершины $\boldsymbol{W}$ на противоположной стороне между вершинами $\boldsymbol{B}$ и $\boldsymbol{G}$. Выберем для проверки точку $\boldsymbol{R}$, т. е. точку, полученную путем "отражения" треугольника относительно стороны $\overline{BG}$. Чтобы найти $\boldsymbol{R}$, сначала определяем среднюю точку $\boldsymbol{M}$ стороны $\overline{BG}$. Затем проводим линию от вершины $\boldsymbol{W}$ к $\boldsymbol{M}$ и обозначаем длину полученного отрезка через d. Этот отрезок продолжается через точку $\boldsymbol{M}$ на длину d до точки $\boldsymbol{R}$ (рис. 8.4). Формула для вектора $\boldsymbol{R}$ имеет вид

$$\boldsymbol{R} = \boldsymbol{M} + (\boldsymbol{M} - \boldsymbol{W}) = 2\boldsymbol{M} - \boldsymbol{W}. \tag{10}$$

Расширение, использующее точку E

Если значение функции в вершине $\boldsymbol{R}$ меньше значения функции в вершине $\boldsymbol{W}$, то выбрано правильное направление в сторону минимума. Возможно, минимум находится несколько дальше, чем точка $\boldsymbol{R}$. Поэтому продлим отрезок через

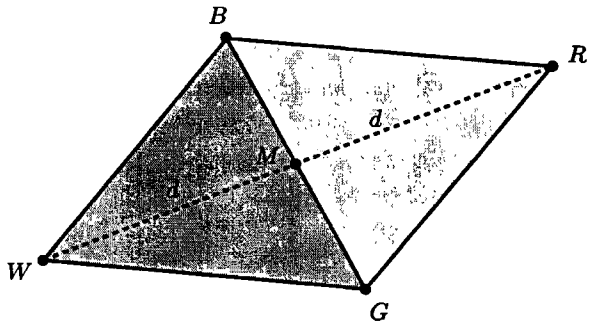


Рис. 8.4. Треугольник $\triangle BGW$, средняя точка M и отраженная точка R для метода Нелдера–Мида

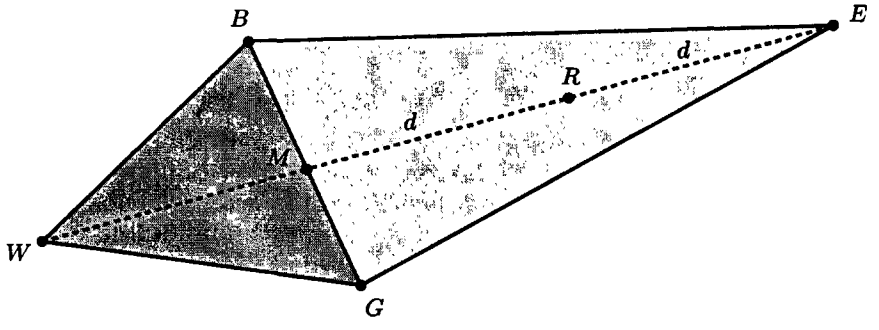


Рис. 8.5. Треугольник $\triangle BGW$, точка R и продолженная точка E

вершины M и R к точке E. Получится вытянутый треугольник BGE. Точку E находим, двигаясь на расстояние d вдоль линии, соединяющей вершины M и R (рис. 8.5). Если значение функции в вершине E меньше значения функции в вершине R, значит, найдена лучшая вершина, чем R. Формула для вектора E имеет вид

$$E = R + (R - M) = 2R - M. \tag{11}$$

Сжатие, использующее точку C

Если значения функции в точках R и W одинаковы, следует проверить другую точку. Возможно, значение функции меньше в точке M, но нельзя заменять вершину W точкой M, так как три точки должны составлять треугольник. Рассмотрим две средние точки C_1 и C_2, которые лежат на отрезках $\overline{WM}$ и $\overline{MR}$ соответственно (рис. 8.6). Точку, в которой функция принимает наименьшее значение, обозначаем через C и получаем новый треугольник BGC.

Замечание. Выбор между точками C_1 и C_2 в двумерном случае может показаться неуместным, однако он очень важен для больших размерностей.

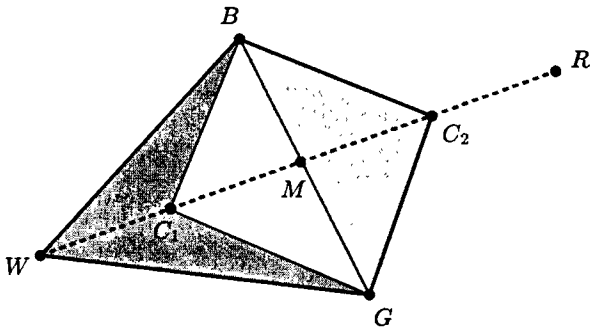


Рис. 8.6. Сжатие точки C_1 или C_2 для метода Нелдера–Мида

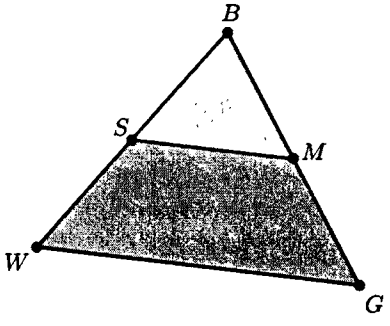


Рис. 8.7. Сокращение треугольника к точке B

Сокращение по направлению к B

Если значение функции в точке C не меньше, чем значение в W, точки G и W следует "стянуть" к точке B (рис. 8.7). Точку G заменяем точкой M, а W — точкой S, которая является средней точкой на отрезке, соединяющем точки B и W.

Логическое решение на каждом шаге

Численно эффективный алгоритм будет вычислять функцию только при необходимости. На каждом шаге находим новую вершину, которой заменяем вершину W. После ее нахождения в дальнейшем исследовании не будет необходимости и шаг итерации будет завершен. Логические детали для двумерного случая объяснены в табл. 8.3.

Пример 8.4. Используем алгоритм Нелдера–Мида, чтобы найти минимум функции $f(x,y) = x^2 - 4x + y^2 - y - xy$. Начнем с трех вершин:

$$V_1 = (0;0), \qquad V_2 = (1{,}2;0{,}0), \qquad V_3 = (0{,}0;0{,}8).$$

Вычисляем значения функции $f(x,y)$ в вершинах

$$f(0;0) = 0{,}0, \qquad f(1{,}2;0{,}0) = -3{,}36, \qquad f(0{,}0;0{,}8) = -0{,}16.$$

Таблица 8.3. Логическое решение для алгоритма Нелдера–Мида

Если $f(R) < f(G)$, ТО выполняем случай (i) {либо отражение, либо растягивание} ИНАЧЕ выполняем случай (ii) {либо сжатие, либо стягивание}	

BEGIN {Случай (i).}
IF $f(B) < f(R)$ THEN
  замена W на R
ELSE
 	
  Вычисление E и $f(E)$
  IF $f(E) < f(B)$ THEN
  замена W на E
  ELSE
  замена W на R
  ENDIF
ENDIF
END {Случай (i).}
	BEGIN {Случай (ii).}
	IF $f(R) < f(W)$ THEN
	  замена W на R
	Вычисление $C = (W + M)/2$
	или $C = (M + R)/2$ и $f(C)$
	IF $f(C) < f(W)$ THEN
	  замена W на C
	ELSE
	  Вычисление S и $f(S)$
	  замена W на S
	  замена G на M
	ENDIF
	END {Случай (ii).}

Следует сравнить значения функции, чтобы определить $\boldsymbol{B}$, $\boldsymbol{G}$ и $\boldsymbol{W}$:

$$\boldsymbol{B} = (1{,}2; 0{,}0), \qquad \boldsymbol{G} = (0{,}0; 0{,}8), \qquad \boldsymbol{W} = (0; 0).$$

Вершину $\boldsymbol{W} = (0; 0)$ следует заменить. Точки $\boldsymbol{M}$ и $\boldsymbol{R}$ имеют координаты

$$\boldsymbol{M} = \frac{\boldsymbol{B} + \boldsymbol{G}}{2} = (0{,}6; 0{,}4) \quad \text{и} \quad \boldsymbol{R} = 2\boldsymbol{M} - \boldsymbol{W} = (1{,}2; 0{,}8).$$

Значение функции $f(\boldsymbol{R}) = f(1{,}2; 0{,}8) = -4{,}48$ меньше значения $f(\boldsymbol{G})$. Таким образом, приходим к случаю (i). Так как $f(\boldsymbol{R}) \leq f(\boldsymbol{B})$, движемся вправо, и нужно построить вершину $\boldsymbol{E}$:

$$\boldsymbol{E} = 2\boldsymbol{R} - \boldsymbol{M} = 2(1{,}2; 0{,}8) - (0{,}6; 0{,}4) = (1{,}8; 1{,}2).$$

Значение функции $f(\boldsymbol{E}) = f(1{,}8; 1{,}2) = -5{,}88$ меньше значения $f(\boldsymbol{B})$, и получаем новый треугольник с вершинами

$$\boldsymbol{V}_1 = (1{,}8; 1{,}2), \qquad \boldsymbol{V}_2 = (1{,}2; 0{,}0), \qquad \boldsymbol{V}_3 = (0{,}0; 0{,}8).$$

Продолжаем процесс и генерируем последовательность треугольников, которая сходится к решению в точке $(3; 2)$ (рис. 8.8). В табл. 8.4 приведены значения функции в вершинах треугольника на различных шагах итерации. Вычисление алгоритма на компьютере продолжается до 33-го шага, когда наилучшей вершиной будут $\boldsymbol{B} = (2{,}99996456; 1{,}99983839)$, причем $f(\boldsymbol{B}) = -6{,}99999998$. Эти значения найдены в примере 8.3 и являются приближениями к $f(3; 2) = -7$. Основанием для прекращения итерации до вычисления точки $(3; 2)$ является то, что функция плоская около минимума. Значения функции $f(\boldsymbol{B})$, $f(\boldsymbol{G})$ и $f(\boldsymbol{W})$ проверены, и найдено, что они все одинаковы (это пример ошибки округления). Алгоритм завершен. ■

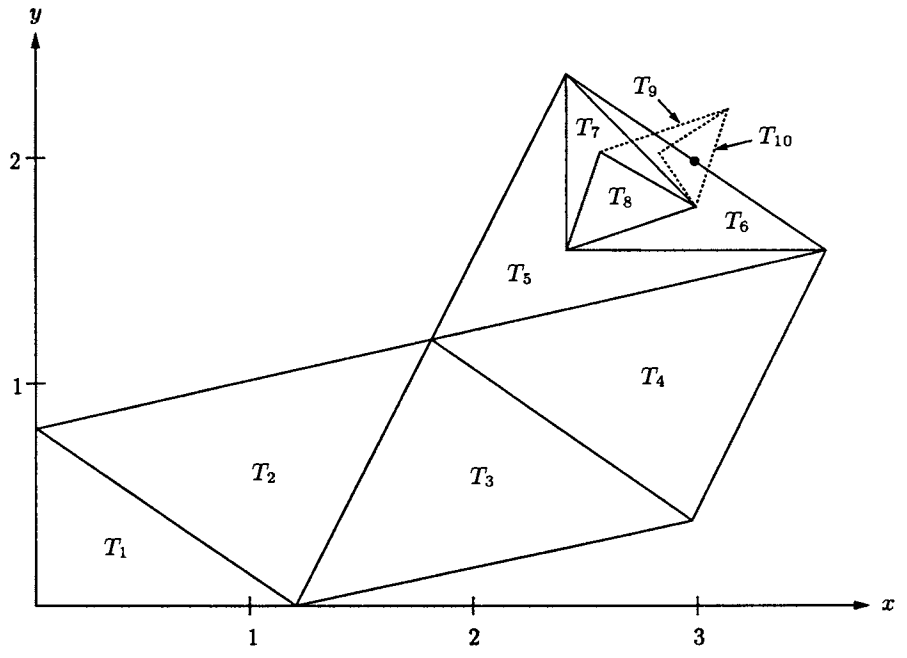


Рис. 8.8. Последовательность треугольников $\{T_k\}$, сходящаяся к точке $(3; 2)$ для метода Нелдера–Мида

Таблица 8.4. Значения функции в различных треугольниках из примера 8.4

k	Наилучшая точка	Хорошая точка	Наихудшая точка
1	$f(1{,}2; 0{,}0) = -3{,}36$	$f(0{,}0; 0{,}8) = -0{,}16$	$f(0{,}0; 0{,}0) = 0{,}00$
2	$f(1{,}8; 1{,}2) = -5{,}88$	$f(1{,}2; 0{,}0) = -3{,}36$	$f(0{,}0; 0{,}8) = -0{,}16$
3	$f(1{,}8; 1{,}2) = -5{,}88$	$f(3{,}0; 0{,}4) = -4{,}44$	$f(1{,}2; 0{,}0) = -3{,}36$
4	$f(3{,}6; 1{,}6) = -6{,}24$	$f(1{,}8; 1{,}2) = -5{,}88$	$f(3{,}0; 0{,}4) = -4{,}44$
5	$f(3{,}6; 1{,}6) = -6{,}24$	$f(2{,}4; 2{,}4) = -6{,}24$	$f(1{,}8; 1{,}2) = -5{,}88$
6	$f(2{,}4; 1{,}6) = -6{,}72$	$f(3{,}6; 1{,}6) = -6{,}24$	$f(2{,}4; 2{,}4) = -6{,}24$
7	$f(3{,}0; 1{,}8) = -6{,}96$	$f(2{,}4; 1{,}6) = -6{,}72$	$f(2{,}4; 2{,}4) = -6{,}24$
8	$f(3{,}0; 1{,}8) = -6{,}96$	$f(2{,}55; 2{,}05) = -6{,}7725$	$f(2{,}4; 1{,}6) = -6{,}72$
9	$f(3{,}0; 1{,}8) = -6{,}96$	$f(3{,}15; 2{,}25) = -6{,}9525$	$f(2{,}55; 2{,}05) = -6{,}7725$
10	$f(3{,}0; 1{,}8) = -6{,}96$	$f(2{,}8125; 2{,}0375) = -6{,}95640625$	$f(3{,}15; 2{,}25) = -6{,}9525$

Минимизация с использованием производных

Предположим, что функция $f(x)$ унимодальна на интервале $[a; b]$ и имеет единственный минимум в точке $x = p$. Предположим также, что $f'(x)$ определена во всех точках интервала $(a; b)$. Начнем с точки p_0, лежащей на интервале $(a; b)$. Если $f'(p_0) < 0$, точка минимума p лежит справа от точки p_0. Если $f'(p_0) > 0$, точка p лежит слева от p_0 (рис. 8.9).

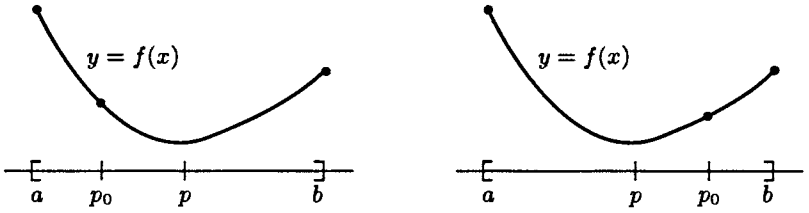


Рис. 8.9. Использование $f'(x)$ для нахождения минимума унимодальной функции $f(x)$ на интервале $[a, b]$

Нахождение минимума методом интервалов

Первая задача — получить такие три значения для проверки

$$p_0, \quad p_1 = p_0 + h \quad \text{и} \quad p_2 = p_0 + 2h, \tag{12}$$

что

$$f(p_0) > f(p_1) \quad \text{и} \quad f(p_1) < f(p_2). \tag{13}$$

Предположим, что $f'(p_0) < 0$. Тогда $p_0 < p$ и следует выбрать длину шага h положительной. Легко найти такое значение h, чтобы три точки в (12) удовлетворяли неравенствам (13). Начнем с $h = 1$ в формуле (12) (при условии, что $a + 1 < b$).

Случай (i). Если выполняются неравенства (13), то шаг определен.

Случай (ii). Если $f(p_0) > f(p_1)$ и $f(p_1) > f(p_2)$, то $p_2 < p$. Необходимо проверить точки, которые лежат дальше справа. Удваиваем шаг и повторяем процесс.

Случай (iii). Если $f(p_0) \leq f(p_1)$, значит, мы "перепрыгнули" через точку p и h слишком велико. Необходимо проверить значения, ближайшие к точке p_0. Уменьшаем шаг в 2 раза и повторяем процесс.

Когда $f'(p_0) > 0$, то длину шага h следует выбирать отрицательной и затем рассматривать случаи, подобные (i)–(iii).

Нахождение p квадратическим приближением

Наконец, есть три точки (12), которые удовлетворяют неравенствам (13). Используем квадратическое приближение, чтобы найти $p_{\min}$, которое является приближением к p. Полином Лагранжа, построенный на узлах (12), имеет вид

$$Q(x) = \frac{y_0(x - p_1)(x - p_2)}{2h^2} - \frac{y_1(x - p_0)(x - p_2)}{h^2} + \frac{y_2(x - p_0)(x - p_1)}{2h^2}. \tag{14}$$

Производная $Q(x)$ равна

$$Q'(x) = \frac{y_0(2x - p_1 - p_2)}{2h^2} - \frac{y_1(2x - p_0 - p_2)}{h^2} + \frac{y_2(2x - p_0 - p_1)}{2h^2}. \tag{15}$$

Запишем $Q'(x) = 0$ в виде $Q'(p_0 + h_{\min}) = 0$:

$$0 = \frac{y_0(2(p_0 + h_{\min}) - p_1 - p_2)}{2h^2} - \frac{y_1(4(p_0 + h_{\min}) - 2p_0 - 2p_2)}{2h^2} + \frac{y_2(2(p_0 + h_{\min}) - p_0 - p_1)}{2h^2}. \tag{16}$$

Умножим каждый член в (16) на $2h^2$ и объединим члены, содержащие $h_{\min}$:

$$\begin{aligned} -h_{min}(2y_0 - 4y_1 + 2y_2) &= y_0(2p_0 - p_1 - p_2) - \\ &\quad - y_1(4p_0 - 2p_0 - 2p_2) + y_2(2p_0 - p_0 - p_1) = \\ &= y_0(-3h) - y_1(-4h) + y_2(-h). \end{aligned}$$

Последнее уравнение легко решить относительно $h_{\min}$:

$$h_{\min} = \frac{h(4y_1 - 3y_0 - y_2)}{4y_1 - 2y_0 - 2y_2}. \tag{17}$$

Значение $p_{\min} = p_0 + h_{\min}$ является лучшим приближением к p, чем p_0. Поэтому можно заменить p_0 на $p_{\min}$ и повторить схему двух описанных выше процессов, чтобы определить новую длину шага h и новое $h_{\min}$. Продолжаем итерацию до тех пор, пока не достигнем требуемой точности. Подробная схема реализована в программе 8.3.

Метод наискорейшего спуска, или градиентный метод

Обратимся к минимизации функции $f(\boldsymbol{X})$ от N переменных, где $\boldsymbol{X} = (x_1, x_2, \ldots, x_N)$. Градиент $f(\boldsymbol{X})$ — это вектор (векторная функция), определенный следующим образом:

$$\operatorname{grad} f(\boldsymbol{X}) = (f_1, f_2, \ldots, f_N), \tag{18}$$

где частные производные $f_k = \partial f / \partial x_k$ вычисляются в точке $\boldsymbol{X}$.

Напомним, что градиент (18) указывает направление наибольшей скорости возрастания функции $f(\boldsymbol{X})$. Следовательно, $-\operatorname{grad} f(\boldsymbol{X})$ указывает направление наибольшего убывания. Начнем поиск из точки $\boldsymbol{P}_0$ вдоль линии, проходящей через $\boldsymbol{P}_0$ в направлении $\boldsymbol{S}_0 = -\boldsymbol{G}/\|\boldsymbol{G}\|$, где $\boldsymbol{G} = \operatorname{grad} f(\boldsymbol{P}_0)$. И тогда придем в точку $\boldsymbol{P}_1$, где находится локальный минимум, когда точка $\boldsymbol{X}$ вынуждена будет попасть на линию $\boldsymbol{X} = \boldsymbol{P}_0 + t\boldsymbol{S}_0$.

Затем можно вычислить $\boldsymbol{G} = \operatorname{grad} f(\boldsymbol{P}_1)$ и двигаться в направлении $\boldsymbol{S}_1 = -\boldsymbol{G}/\|\boldsymbol{G}\|$. Придем в точку $\boldsymbol{P}_2$, где находится локальный минимум, когда точка $\boldsymbol{X}$ вынуждена будет попасть на линию $\boldsymbol{X} = \boldsymbol{P}_1 + t\boldsymbol{S}_1$. Итерация порождает

последовательность точек $\{\boldsymbol{P}_k\}$, обладающих свойством $f(\boldsymbol{P}_0) > f(\boldsymbol{P}_1) > \cdots > f(\boldsymbol{P}_k) > \cdots$. Если $\lim_{k\to\infty} \boldsymbol{P}_k = \boldsymbol{P}$, то $f(\boldsymbol{P})$ будет локальным минимумом для $f(\boldsymbol{X})$.

Схема метода градиента

Предположим, что последовательность точек $\boldsymbol{P}_k$ получена.

Шаг 1. Вычислим градиент $\boldsymbol{G} = \operatorname{grad} f(\boldsymbol{P}_k)$.

Шаг 2. Вычислим направление поиска $\boldsymbol{S} = -\boldsymbol{G}/\|\boldsymbol{G}\|$.

Шаг 3. Определим единственный параметр минимизации $\Phi(t) = f(\boldsymbol{P}_k + t\boldsymbol{S})$ на интервале $[0, b]$, где b большое. Это даст значение $t = h_{\min}$, где для $\Phi(t)$ находится локальный минимум. Соотношение $\Phi(h_{\min}) = f(\boldsymbol{P}_k + h_{\min}\boldsymbol{S})$ показывает, что это минимум для $f(\boldsymbol{X})$ вдоль выбранной линии $\boldsymbol{X} = \boldsymbol{P}_k + h_{\min}\boldsymbol{S}$.

Шаг 4. Построим следующую точку $\boldsymbol{P}_{k+1} = \boldsymbol{P}_k + h_{\min}\boldsymbol{S}$.

Шаг 5. Определим критерий останова для минимизации, т. е. достаточно ли близки значения функции $f(\boldsymbol{P}_k)$ и $f(\boldsymbol{P}_{k+1})$ и достаточно ли мало расстояние $\|\boldsymbol{P}_{k+1} - \boldsymbol{P}_k\|$?

Повторение процесса.

Программа 8.1 (поиск минимума методом золотого сечения). Программа предназначена для поиска численного приближения минимума функции $f(x)$ на интервале $[a; b]$ методом золотого сечения. Метод применяется, только если функция $f(x)$ унимодальна на интервале $[a; b]$.

```
function[S,E,G]=golden(f,a,b,delta,epsilon)
%Вход  - f - функция, вводимая как строка 'f'
%      - a и b - крайние точки интервала
%      - delta - допустимое отклонение для абсцисс
%      - epsilon - допустимое отклонение для ординат
%Выход - S=(p,yp) - содержит абсциссу p и ординату yp минимума
%      - E=(dp,dy) - содержит грани ошибки для p и yp
%      - G - матрица размера n x 4: k-я строка содержит
%        [ak ck dk bk]; значения a, c, d и b на k-й итерации
r1=(sqrt(5)-1)/2;
r2=r1^2;
h=b-a;
ya=feval(f,a);
yb=feval(f,b);
c=a+r2*h;
d=a+r1*h;
yc=feval(f,c);
```

```
yd=feval(f,d);
k=1;
A(k)=a;B(k)=b;C(k)=c;D(k)=d;
while(abs(yb-ya)>epsilon)|(h>delta)
   k=k+1;
   if(yc<yd)
      b=d;
      yb=yd;
      d=c;
      yd=yc;
      h=b-a;
      c=a+r2*h;
      yc=feval(f,c);
   else
      a=c;
      ya=yc;
      c=d;
      yc=yd;
      h=b-a;
      d=a+r1*h;
      yd=feval(f,d);
   end
   A(k)=a;B(k)=b;C(k)=c;D(k)=d;
end
dp=abs(b-a);
dy=abs(yb-ya);
p=a;
yp=ya;
if(yb<ya)
   p=b;
   yp=yb;
end
G=[A' C' D' B'];
S=[p yp];
E=[dp dy];
```

В программах 8.2 и 8.4 требуется, чтобы функция f записывалась, как М-файл. Аргумент функции f должен быть матрицей размера $1 \times n$. Для иллюстрации рассмотрим запись функции из примера 8.3 в виде М-файла.

```
function z=f(V)
z=0; x=V(1); y=V(2);
z=x.^2-4x+y.^2-y-x.*y;
```

Программа 8.2 (метод минимизации Нелдера–Мида). Программа предназначена для приближенного нахождения локального минимума функции $f(x_1, x_2, \ldots, x_N)$, где f — непрерывная функция от N действительных переменных, и задана $N+1$ начальная точка $V_k = (v_{k,1}, \ldots, v_{k,N})$ для $k = 0, 1, \ldots, N$.

```
function[V0,y0,dV,dy]=nelder(F,V,min1,max1,epsilon,show)
%Вход  - F - функция, вводимая как строка 'F'
%      - V - матрица размера 3 x n, содержащая исходный симплекс
%      - min1 & max1 - минимальное и максимальное количество
%        итераций
%      - epsilon - допустимое отклонение
%      - show == 1 - показывает число итераций (P и Q)
%Выход - V0 - вершина для минимума
%      - y0 - значение функции F(V0)
%      - dV - размер окончательного симплекса
%      - dy - грань ошибки для минимума
%      - P - матрица, содержащая итерации вершин
%      - Q - массив, содержащий итерации для F(P)
if nargin==5,
   show=0;
end
[mm n]=size(V);
% Последовательность вершин
for j=1:n+1
   Z=V(j,1:n);
   Y(j)=feval(F,Z);
end
[mm lo]=min(Y);
[mm hi]=max(Y);
li=hi;
ho=lo;
for j=1:n+1
   if(j~=lo&j~=hi&Y(j)<=Y(li))
      li=j;
   end
   if(j~=hi&j~=lo&Y(j)>=Y(ho))
      ho=j;
   end
end
cnt=0;
% Начало алгоритма Нелдера-Мида
```

```
while(Y(hi)>Y(lo)+epsilon&cnt<max1)|cnt<min1
   S=zeros(1,1:n);
   for j=1:n+1
      S=S+V(j,1:n);
   end
   M=(S-V(hi,1:n))/n;
   R=2*M-V(hi,1:n);
   yR=feval(F,R);
   if(yR<Y(ho))
      if(Y(li)<yR)
         V(hi,1:n)=R;
         Y(hi)=yR;
      else
         E=2*R-M;
         yE=feval(F,E);
         if(yE<Y(li))
            V(hi,1:n)=E;
            Y(hi)=yE;
         else
            V(hi,1:n)=R;
            Y(hi)=yR;
         end
      end
   else
      if(yR<Y(hi))
         V(hi,1:n)=R;
         Y(hi)=yR;
      end
      C=(V(hi,1:n)+M)/2;
      yC=feval(F,C);
      C2=(M+R)/2;
      yC2=feval(F,C2);
      if(yC2<yC)
         C=C2;
         yC=yC2;
      end
      if(yC<Y(hi))
         V(hi,1:n)=C;
         Y(hi)=yC;
      else
         for j=1:n+1
            if(j~=lo)
               V(j,1:n)=(V(j,1:n)+V(lo,1:n))/2;
```

```
                Z=V(j,1:n);
                Y(j)=feval(F,Z);
             end
          end
       end
    end
    [mm lo]=min(Y);
    [mm hi]=max(Y);
    li=hi;
    ho=lo;
    for j=1:n+1
       if(j~=lo&j~=hi&Y(j)<=Y(li))
          li=j;
       end
       if(j~=hi&j~=lo&Y(j)>=Y(ho))
          ho=j;
       end
    end
    cnt=cnt+1;
    P(cnt,:)=V(lo,:);
    Q(cnt)=Y(lo);
end
% Конец алгоритма Нелдера-Мида
%Определение размера симплекса
snorm=0;
for j=1:n+1
    s=norm(V(j)-V(lo));
    if(s>=snorm)
       snorm=s;
    end
end
Q=Q';
V0=V(lo,1:n);
y0=Y(lo);
dV=snorm;
dy=abs(Y(hi)-Y(lo));
if (show==1)
    disp(P);
    disp(Q);
end
```

Программа 8.3 (поиск локального минимума, использующий квадратичное интерполирование). Программа предназначена для нахождения локального минимума функции $f(x)$ на интервале $[a;b]$. Выбирается начальное приближение p_0 и затем происходит поиск на интервалах $[a;p_0]$ и $[p_0;b]$.

```
function[p,yp,dp,dy,P]=quadmin(f,a,b,delta,epsilon)
%Вход  - f - функция, вводимая как строка 'f'
%      - a и b - крайние точки интервала
%      - delta - допустимое значение для абсцисс
%      - epsilon - допустимое значение для ординат
%Выход - p - абсцисса минимума
%      - yp - ордината минимума
%      - dp - грань ошибки для p
%      - dy - грань ошибки для yp
%      - P - вектор итераций
p0=a;
maxj=20;
maxk=30;
big=1e6;
err=1;
k=1;
P(k)=p0;
cond=0;
h=1;
if (abs(p0)>1e4),h=abs(p0)/1e4;end
while(k<maxk&err>epsilon&cond~=5)
   f1=(feval(f,p0+0.00001)-feval(f,p0-0.00001))/0,00002;
   if(f1>0),h=-abs(h);end
   p1=p0+h;
   p2=p0+2*h;
   pmin=p0;
   y0=feval(f,p0);
   y1=feval(f,p1);
   y2=feval(f,p2);
   ymin=y0;
   cond=0;
   j=0;
   %Определение такого h, что y1<y0&y1<y2
   while(j<maxj&abs(h)>delta&cond==0)
      if (y0<=y1),
         p2=p1;
         y2=y1;
         h=h/2;
```

```
      p1=p0+h;
      y1=feval(f,p1);
   else
      if(y2<y1),
         p1=p2;
         y1=y2;
         h=2*h;
         p2=p0+2*h;
         y2=feval(f,p2);
      else
         cond=-1;
      end
   end
j=j+1;
if(abs(h)>big|abs(p0)>big),cond=5;end
end
if(cond==5),
   pmin=p1;
   ymin=feval(f,p1);
else
   %Квадратичное интерполирование для нахождения yp
   d=4*y1-2*y0-2*y2;
   if(d<0),
      hmin=h*(4*y1-3*y0-y2)/d;
   else
      hmin=h/3;
      cond=4;
   end
   pmin=p0+hmin;
   ymin=feval(f,pmin);
   h=abs(h);
   h0=abs(hmin);
   h1=abs(hmin-h);
   h2=abs(hmin-2*h);
   %Определение величины следующего h
   if(h0<h),h=h0;end
   if(h1<h),h=h1;end
   if(h2<h),h=h2;end
   if(h==0),h=hmin;end
   if(h<delta),cond=1;end
   if (abs(h)>big|abs(pmin)>big),cond=5;end
   %Критерий останова для минимизации
   e0=abs(y0-ymin);
```

```
      e1=abs(y1-ymin);
      e2=abs(y2-ymin);
      if(e0~=0 & e0<err),err=e0;end
      if(e1~=0 & e1<err),err=e1;end
      if(e2~=0 & 2<err),err=e2;end
      if(e0~=0 & e1==0 & e2==0),error=0;end
      if(err<epsilon),cond=2;end
      p0=pmin;
      k=k+1;
      P(k)=p0;
   end
   if(cond==2&h<delta),cond=3;end
end
p=p0;
dp=h;
yp=feval(f,p);
dy=err;
```

В программе 8.4 требуется, чтобы функция f записывалась, как М-файл. Дополнительно поиск направления $-\operatorname{grad} f / \|\operatorname{grad} f\|$ необходимо записывать, как М-файл. Для иллюстрации рассмотрим функцию f из примера 8.3, где градиент f равен $(2x-4-y, 2y-1-x)$. Соответствующий М-файл для этой функции f имеет следующий вид:

```
function z=G(V)
z=zeros(1,2);
x=V(1);y=V(2);
g=[2x-4-y 2*y-1-x];
z=-(1/norm(g))*g;
```

Программа 8.4 (метод наискорейшего спуска или градиентный метод). Программа презназначена для нахождения численного приближения локального минимума $f(\boldsymbol{X})$, где f — непрерывная функция от N действительных переменных и $\boldsymbol{X} = (x_1, x_2, \ldots, x_N)$, начиная с одной точки $\boldsymbol{P}_0$ и применяя градиентный метод.

```
function[P0,y0,err]=grads(F,G,P0,max1,delta,epsilon,show)
%Вход  - F - функция, вводимая, как строка 'F'
%      - G = -(1/norm(grad F)) * grad F; выбор направления
%        вводится, как строка 'G'
%      - P0 - начальная точка
%      - max1 - максимальное число итераций
%      - delta - допустимое отклонение для hmin в единственном
%        параметре минимизации в выбранном направлении
```

```
%       - epsilon - допустимое отклонение для ошибки в y0
%       - show; если show==1, итерации выводятся на дисплей
%Выход - P0 - точка минимума
%       - y0 - значение функции F(P0)
%       - err - грань ошибки для y0
%       - P - вектор, содержащий итерации
if nargin==5,show=0;end
[mm n]=]size(P0);
maxj=10; big=1e8; h=1;
P=zeros(maxj,n+1);
len=norm(P0);
y0=feval(F,P0);
if (len>e4),h=len/1e4;end
err=1;cnt=0;cond=0;
P(cnt+1,:)=[P0 y0];
while(cnt<max1&cond~=5&(h>delta|err>epsilon))
   %Вычисление направления поиска
   S=feval(G,P0);
   %Начало выбора параметра квадратичной минимизации
   P1=P0+h*S;
   P2=P0+2*h*S;
   y1=feval(F,P1);
   y2=feval(F,P2);
   cond=0;j=0;
   while(j<maxj&cond==0)
      len=norm(P0);
      if (y0<y1)
      P2=P1;
      y2=y1;
      h=h/2;
      P1=P0+h*S;
      y1=feval(F,P1);
   else
      if(y2<y1)
         P1=P2;
         y1=y2;
         h=2*h;
         P2=P0+2*h*S;
         y2=feval(F,P2);
      else
         cond=-1;
      end
```

```
   end
   j=j+1;
   if(h<delta),cond=1;end
   if(abs(h)>big|len>big),cond=5;end
end
if(cond==5)
   Pmin=P1;
   ymin=y1;
else
   d=4*y1-2*y0-2*y2;
   if(d<0)
      hmin=h*(4*y1-3*y0-y2)/d;
   else
      cond=4;
      hmin=h/3;
   end
   %Построение следующей точки
   Pmin=P0+hmin*S;
   ymin=feval(F,Pmin);
   %Определение величины следующего h
   h0=abs(hmin);
   h1=abs(hmin-h);
   h2=abs(hmin-2*h);
   if(h0<h),h=h0;end
   if(h1<h),h=h1;end
   if(h2<h),h=h2;end
   if(h==0),h=hmin;end
   if(h<delta),cond=1;end
   %Критерий останова для минимизации
   e0=abs(y0-ymin);
   e1=abs(y1-ymin);
   e2=abs(y2-ymin);
   if(e0~=0&e0<err),err=e0;end
   if(e1~=0&e1<err),err=e1;end
   if(e2~=0&e2<err),err=e2;end
   if(e0==0&e1==0&e2==0),err=0;end
   if(err<epsilon),cond=2;end
   if(cond==2&h<delta),cond=3;end
end
cnt=cnt+1;
P(cnt+1,:)=[Pmin ymin];
P0=Pmin;
```

```
y0=ymin;
end
if(show==1)
   disp(P);
end
```

Упражнения к разделу 8.1

1. Используйте теорему 8.1, чтобы определить, где каждая из следующих функций возрастает и где убывает.
   (a) $f(x) = 2x^3 - 9x^2 + 12x - 5$
   (b) $f(x) = x/(x+1)$
   (c) $f(x) = (x+1)/x$
   (d) $f(x) = x^x$
2. Используйте определение 8.3, чтобы показать, что следующие функции унимодальны на интервале.
   (a) $f(x) = x^2 - 2x + 1; [0; 4]$
   (b) $f(x) = \cos(x); [0; 3]$
   (c) $f(x) = x^x; [1; 10]$
   (d) $f(x) = -x(3-x)^{5/3}; [0; 3]$
3. Воспользуйтесь теоремами 8.3 и 8.4 по возможности, чтобы найти все локальные минимумы и максимумы каждой из следующих функций на заданном интервале.
   (a) $f(x) = 4x^3 - 8x^2 - 11x + 5;\ [0; 2]$
   (b) $f(x) = x + 3/x^2;\ [0{,}5; 3]$
   (c) $f(x) = (x + 2{,}5)/(4 - x^2);\ [-1{,}9; 1{,}9]$
   (d) $f(x) = e^x/x^2;\ [0{,}5; 3]$
   (e) $f(x) = -\sin(x) - \sin(3x)/3;\ [0; 2]$
   (f) $f(x) = -2\sin(x) + \sin(2x) - 2\sin(3x)/3;\ [1; 3]$
4. Найдите на параболе $y = x^2$ точку, которая ближе всех к точке $(3; 1)$.
5. Найдите на кривой $y = \sin(x)$ точку, которая ближе всех к точке $(2; 1)$.
6. Найдите на окружности $x^2 + y^2 = 25$ самую дальнюю от хорды AB точку (точки), если $A = (3; 4)$ и $B = (-1, \sqrt{24})$.
7. Воспользуйтесь теоремой 8.5, чтобы найти локальный минимум каждой из следующих функций.
   (a) $f(x, y) = x^3 + y^3 - 3x - 3y + 5$
   (b) $f(x, y) = x^2 + y^2 + x - 2y - xy + 1$

(c) $f(x,y) = x^2y + xy^2 - 3xy$

(d) $f(x,y) = (x-y)/(x^2+y^2+2)$

(e) $f(x,y) = 100(y-x^2)^2 + (1-x)^2$

(Параболический желоб Розенброка; ок. 1960 г.)

8. Пусть $\boldsymbol{B} = (2;-3)$, $\boldsymbol{G} = (1;1)$ и $\boldsymbol{W} = (5;2)$. Найдите точки $\boldsymbol{M}$, $\boldsymbol{R}$ и $\boldsymbol{E}$ и сделайте наброски треугольников, которые они включают.

9. Пусть $\boldsymbol{B} = (-1;2)$, $\boldsymbol{G} = (-2;-5)$ и $\boldsymbol{W} = (3;1)$. Найдите точки $\boldsymbol{M}$, $\boldsymbol{R}$ и $\boldsymbol{E}$ и сделайте наброски треугольников, которые они включают.

10. Приведите схему доказательства, что $\boldsymbol{M} = (\boldsymbol{B} + \boldsymbol{G})/2$ — средняя точка отрезка, соединяющего точки $\boldsymbol{B}$ и $\boldsymbol{G}$.

11. Дайте схему доказательства формулы (10).

12. Дайте схему доказательства формулы (11).

13. Дайте схему доказательства, что медианы любого треугольника пересекаются в точке, которая делит расстояние от каждой вершины до средней точки противоположной стороны в отношении $\frac{1}{3}$ и $\frac{2}{3}$.

14. Пусть $\boldsymbol{B} = (0;0;0)$, $\boldsymbol{G} = (1;1;0)$, $\boldsymbol{P} = (0;0;1)$ и $\boldsymbol{W} = (1;0;0)$.

(a) Сделайте набросок тетраэдра $BGPW$.

(b) Найдите $\boldsymbol{M} = (\boldsymbol{B} + \boldsymbol{G} + \boldsymbol{P})/3$.

(c) Найдите $\boldsymbol{R} = 2\boldsymbol{M} - \boldsymbol{W}$ и сделайте набросок тетраэдра $BGPR$.

(d) Найдите $\boldsymbol{E} = 2\boldsymbol{R} - \boldsymbol{M}$ и сделайте набросок тетраэдра $BGPE$.

15. Пусть $\boldsymbol{B} = (0;0;0)$, $\boldsymbol{G} = (0;2;0)$, $\boldsymbol{P} = (0;1;1)$ и $\boldsymbol{W} = (2;1;0)$. Выполните пункты упр. 14.

Алгоритмы и программы

1. Воспользуйтесь программой 8.1, чтобы найти локальный минимум каждой функции из упр. 3 с точностью до восьми десятичных знаков.

2. Воспользуйтесь программой 8.3, чтобы найти локальный минимум каждой функции из упр. 3 с точностью до восьми десятичных знаков. Начните со средней точки заданного интервала.

3. Воспользуйтесь программой 8.2, чтобы найти минимум каждой функции из упр. 7 с точностью до восьми десятичных знаков. Возьмите в качестве начальных следующие вершины.

(a) $(1;2)$, $(2;0)$ и $(2;2)$

(b) $(0;0)$, $(2;0)$ и $(2;1)$

(c) $(0;0)$, $(2;0)$ и $(2;1)$

(d) $(0;0)$, $(0;1)$ и $(1;1)$

(e) $(0;0)$, $(1;0)$ и $(0;2)$

4. Воспользуйтесь программой 8.4, чтобы найти минимум каждой функции из упр. 7 с точностью до восьми десятичных знаков. Используйте исходные вершины.
   (a) (1; 2) **(b)** (0; 0,3) **(c)** (0,1; 0,1)
   (d) (0,5; 0,11) **(e)** (0; 0)
5. В программе 8.4 координаты итераций x и y хранятся соответственно в первых двух столбцах матрицы P. Модифицируйте программу 8.4 таким образом, чтобы она могла построить графики координат x и y E итераций в одной и той же системе координат. *Указание.* Включите в свою программу команду `plot(P(:,1),P(:,2),'.')`. Проверьте работу программы на функциях из упр. 7.
6. Используйте программу 8.2, чтобы найти локальный минимум каждой из следующих функций с точностью до восьми десятичных знаков.
   (a) $f(x;y;z) = 2x^2 + 2y^2 + z^2 - 2xy + yz - 7y - 4z$
   Начните с (1; 1; 1), (0; 1; 0), (1; 0; 1) и (0; 0; 1).
   (b) $f(x;y;z;u) = 2(x^2+y^2+z^2+u^2) - x(y+z-u) + yz - 3x - 8y - 5z - 9u$
   Начните поиск вблизи точки (1; 1; 1; 1).
   (c) $f(x;y;z;u) = xyzu + \frac{1}{x} + \frac{1}{y} + \frac{1}{z} + \frac{1}{u}$
   Начните поиск вблизи точки (0,7; 0,7; 0,7; 0,7).
7. Воспользуйтесь программой 8.4, чтобы найти локальный минимум каждой из функций задачи 6. Используйте начальные значения около одной из заданных вершин.
8. Воспользуйтесь программой 8.1 и/или 8.3, чтобы найти все максимумы и минимумы указанной функции на интервале [0; 2].
$$f(x) = \frac{x^3 + x^2 - 12x - 12}{2x^6 - 3x^5 - 4x^4 + 9x^2 + 12x - 18}$$
9. Найдите точку на поверхности $z = x^2 + y^2$, ближайшую к точке (2; 3; 1).
10. У компании есть пять фабрик, A, B, C, D и E, расположенных в точках (10; 10), (30; 50), (16,667; 29), (0,555; 29,888) и (22,2221; 49,988) соответственно на площади xy. Предположим, что расстояние между двумя точками задает расстояние в милях между фабриками. Компания планирует построить склад в некоторой точке. Ожидается, что в среднем в течение недели должно быть 10, 18, 20, 14 и 25 поставок на фабрики A, B, C, D и E соответственно. Покажите, как минимизировать еженедельное число пройденных миль для поставок, где следует расположить склад?
11. Где в задаче 10 следует расположить склад, если соответствующий район ограничен? Следует ли расположить его в точке на кривой $y = x^2$?

11

Собственные значения и собственные векторы

Проектирование определенных инженерных систем включает ***теорию повреждений при максимальных нагрузках***. Она основана на предположении, что максимальная основная нагрузка, действуя на тело, разрушает его. Математическим результатом, связанным с этой проблемой, является основная теорема о линейном преобразовании $\boldsymbol{Y} = \boldsymbol{AX}$. В двумерном случае существуют такие базисные векторы $\boldsymbol{U}_1$ и $\boldsymbol{U}_2$, что действие такого преобразования выражается в растяжении пространства в направлениях, параллельных $\boldsymbol{U}_1$ и $\boldsymbol{U}_2$, на величины λ_1 и λ_2 соответственно. Рассмотрим симметричную матрицу

$$\begin{bmatrix} 3{,}8 & 0{,}6 \\ 0{,}6 & 2{,}2 \end{bmatrix};$$

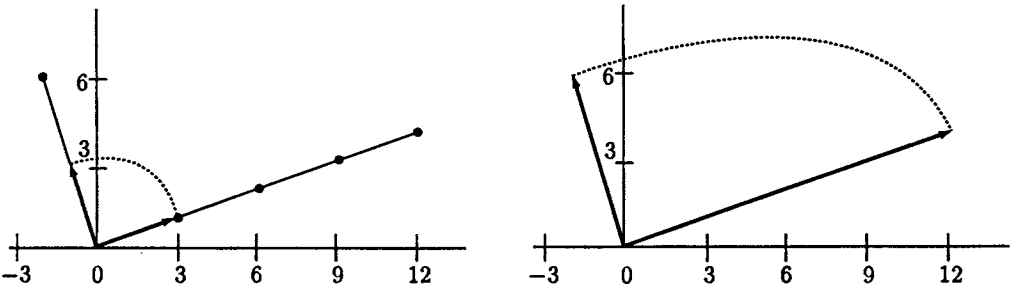


Рис. 11.1. (a) Прообразы $\boldsymbol{U}_1 = [3 \quad 1]'$ и $\boldsymbol{U}_2 = [-1 \quad 3]'$ для преобразования $\boldsymbol{Y} = \boldsymbol{AX}$. (b) Образ векторов $\boldsymbol{V}_1 = \boldsymbol{AU}_1 = [12 \quad 4]'$ и $\boldsymbol{V}_2 = \boldsymbol{AU}_2 = [-2 \quad 6]'$

главные направления — $\boldsymbol{U}_1 = [3 \quad 1]'$ и $\boldsymbol{U}_2 = [-1 \quad 3]'$ с соответствующими собственными значениями $\lambda_1 = 4$ и $\lambda_2 = 2$. Образы этих векторов — $\boldsymbol{V}_1 = \boldsymbol{A}\boldsymbol{U}_1 = [12 \quad 4]' = 4[3 \quad 1]'$ и $\boldsymbol{V}_2 = \boldsymbol{A}\boldsymbol{U}_2 = [-2 \quad 6]' = 2[-1 \quad 3]'$. В результате данного преобразования четверть круга растягивается в четверть эллипса, как показано на рис. 11.1.

11.1. Однородные системы: задача о собственных значениях

Предпосылка

Напомним некоторые понятия из линейной алгебры. Доказательства теорем оставлены читателю в качестве упражнений; их можно найти в любом стандартном учебнике по линейной алгебре (см. [132]).

В главе 3 показано, как решить n линейных уравнений с n неизвестными. Предполагалось, что определитель матрицы не равен нулю и, следовательно, решение единственное. Для случая однородных систем $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}$, если $\det(\boldsymbol{A}) \neq 0$, то единственное решение тривиально: $\boldsymbol{X} = \boldsymbol{0}$. Если $\det(\boldsymbol{A}) = 0$, то существуют нетривиальные решения системы $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}$. Предположим, что $\det(\boldsymbol{A}) = 0$, и рассмотрим решения однородной линейной системы

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\
\vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\quad &\vdots \\
a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= 0.
\end{aligned} \tag{1}
$$

Система уравнений (1) всегда имеет тривиальное решение $x_1 = 0$, $x_2 = 0$, ..., $x_n = 0$. Чтобы получить решение, образуя совокупность соотношений между переменными, можно использовать метод исключения Гаусса.

Пример 11.1. Найдем нетривиальные решения однородной системы

$$
\begin{aligned}
x_1 + 2x_2 - x_3 &= 0, \\
2x_1 + \ x_2 + x_3 &= 0, \\
5x_1 + 4x_2 + x_3 &= 0.
\end{aligned}
$$

Используя метод исключения Гаусса, исключим x_1 и в результате получим

$$
\begin{aligned}
x_1 + 2x_2 - \ x_3 &= 0, \\
-3x_2 + 3x_3 &= 0, \\
-6x_2 + 6x_3 &= 0.
\end{aligned}
$$

Поскольку третье уравнение кратно второму, система приводится к двум уравнениям с тремя неизвестными:

$$\begin{aligned} x_1 + x_2 \quad &= 0 \\ -x_2 + x_3 &= 0. \end{aligned}$$

Выберем одно неизвестное и используем его как параметр. Например, положим $x_3 = t$. Тогда из второго уравнения следует, что $x_2 = t$, и из первого уравнения вычисляем, что $x_1 = -t$. Значит, решение можно выразить через совокупность соотношений:

$$\begin{aligned} x_1 &= -t \\ x_2 &= \ \ t \\ x_3 &= \ \ t \end{aligned} \quad \text{или} \quad \boldsymbol{X} = \begin{bmatrix} -t \\ t \\ t \end{bmatrix} = t \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix},$$

где t — любое действительное число. ■

Определение 11.1 (линейная независимость). Говорят, что векторы $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, ..., $\boldsymbol{U}_n$ ***линейно независимы***, если из уравнения

$$c_1\boldsymbol{U}_1 + c_2\boldsymbol{U}_2 + \cdots + c_n\boldsymbol{U}_n = \boldsymbol{0} \tag{2}$$

следует, что $c_1 = 0$, $c_2 = 0, \ldots, c_n = 0$. Если векторы не являются линейно независимыми, то говорят, что они линейно зависимы. Другими словами, векторы ***линейно зависимы***, если существует множество не равных нулю чисел $\{c_1, c_2, \ldots, c_n\}$, таких, что выполняется равенство (2). ▲

Два вектора в $\Re^2$ линейно независимы тогда и только тогда, когда они не параллельны. Три вектора в $\Re^3$ линейно независимы тогда и только тогда, когда они не лежат в одной плоскости.

Теорема 11.1. Векторы $\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_n$ линейно зависимы тогда и только тогда, когда по крайней мере один из них является линейной комбинацией других.

Хорошим свойством векторного пространства является возможность выражения каждого вектора в виде линейной комбинации векторов, выбранных из малого подмножества векторов. Это служит поводом для следующего определения.

Определение 11.2 (базис). Предположим, что $S = \{\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_m\}$ — множество m векторов в пространстве $\Re^n$. Множество S называется базисом для $\Re^n$, если для каждого вектора $\boldsymbol{X}$ в $\Re^n$ существует единственное множество скаляров $\{c_1, c_2, \ldots, c_m\}$, таких, что $\boldsymbol{X}$ можно выразить в виде линейной комбинации

$$\boldsymbol{X} = c_1\boldsymbol{U}_1 + c_2\boldsymbol{U}_2 + \cdots + c_m\boldsymbol{U}_m. \tag{3}$$

▲

Теорема 11.2. Любое множество из n линейно независимых векторов образует базис пространства $\Re^n$. Каждый вектор $\boldsymbol{X}$ в $\Re^n$ может быть единственным образом выражен в виде линейной комбинации базисных векторов, как показано в равенстве (3).

Теорема 11.3. Пусть $\boldsymbol{K}_1, \boldsymbol{K}_2, \ldots, \boldsymbol{K}_m$ — векторы в $\Re^n$.

(4) Если $m > n$, то векторы линейно зависимы.

(5) Если $m = n$, то векторы линейно зависимы тогда и только тогда, когда $\det(\boldsymbol{K}) = 0$, где $\boldsymbol{K} = \begin{bmatrix}\boldsymbol{K}_1 & \boldsymbol{K}_2 & \ldots & \boldsymbol{K}_m\end{bmatrix}$.

Собственные значения

Иногда в математических приложениях возникают следующие вопросы. "Каковы особенности выражения $\boldsymbol{A} - \lambda\boldsymbol{I}$, где λ — параметр?" "Каково поведение последовательности векторов $\{\boldsymbol{A}^j\boldsymbol{X}_0\}_{j=0}^{\infty}$?" "Каковы геометрические свойства линейного преобразования?" Решения задач для различных областей приложения математики, таких как экономика, техника и физика, могут включать понятия, имеющие отношение к этим вопросам. Теория собственных значений и собственных векторов достаточно мощна, чтобы решить эти или другие подобные задачи.

Пусть $\boldsymbol{A}$ — квадратная матрица размера $n \times n$, и пусть $\boldsymbol{X}$ — вектор размера n. Произведение $\boldsymbol{Y} = \boldsymbol{AX}$ можно рассматривать как линейное преобразование n-мерного пространства самого в себя. Требуется найти скаляры λ, для которых существует такой не равный нулю вектор $\boldsymbol{X}$, что

$$\boldsymbol{AX} = \lambda\boldsymbol{X}; \tag{6}$$

т. е. линейное преобразование $T(\boldsymbol{X}) = \boldsymbol{AX}$ отображает $\boldsymbol{X}$ на кратное $\lambda\boldsymbol{X}$. Когда это происходит, $\boldsymbol{X}$ называют собственным вектором, который соответствует собственному значению λ, а вместе они образуют собственную пару λ, $\boldsymbol{X}$ для матрицы $\boldsymbol{A}$. Вообще, скаляр λ и вектор $\boldsymbol{X}$ могут быть комплексными. Для простоты в большинстве примеров будут использоваться действительные числа. Тем не менее техника вычислений легко переносится на случай с комплексными числами. Единичную матрицу $\boldsymbol{I}$ можно использовать, чтобы выразить уравнение (6) как $\boldsymbol{AX} = \lambda\boldsymbol{IX}$, которое затем можно переписать в обычной форме для линейных систем, как-то

$$(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{X} = \boldsymbol{0}. \tag{7}$$

Значение системы (7) состоит в том, что произведение матрицы $(\boldsymbol{A} - \lambda\boldsymbol{I})$ и не равного нулю вектора $\boldsymbol{X}$ — нулевой вектор! Согласно теореме 3.5 эта линейная

система имеет нетривиальные решения тогда и только тогда, когда матрица $\boldsymbol{A}-\lambda\boldsymbol{I}$ вырождена, т. е.

$$\det(\boldsymbol{A}-\lambda\boldsymbol{I})=0. \tag{8}$$

Этот определитель можно записать в виде

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn}-\lambda \end{vmatrix}=0. \tag{9}$$

Когда определитель в (9) записан в виде полинома степени n, его называют характеристическим полиномом

$$\begin{aligned} p(\lambda)&=\det(\boldsymbol{A}-\lambda\boldsymbol{I})= \\ &=(-1)^n(\lambda^n+c_1\lambda^{n-1}+c_2\lambda^{n-2}+\cdots+c_{n-1}\lambda+c_n). \end{aligned} \tag{10}$$

Полином степени n имеет точно n корней (не обязательно различных). Каждый из корней λ можно подставить в уравнение (7), чтобы получить конкретную систему уравнений, которая имеет соответственный нетривиальный вектор решения $\boldsymbol{X}$. Если λ — действительное число, то можно построить действительный собственный вектор $\boldsymbol{X}$. Чтобы подчеркнуть это, дадим следующие определения.

Определение 11.3 (собственное значение). Если $\boldsymbol{A}$ — действительная матрица размера $n\times n$, то ее n собственных значений $\lambda_1, \lambda_2, \ldots, \lambda_n$ — это действительные и комплексные корни характеристического полинома

$$p(\lambda)=\det(\boldsymbol{A}-\lambda\boldsymbol{I}). \tag{11}$$

▲

Определение 11.4 (собственный вектор). Если λ — собственное значение матрицы $\boldsymbol{A}$ и не равный нулю вектор $\boldsymbol{V}$ удовлетворяет равенству

$$\boldsymbol{AV}=\lambda\boldsymbol{V}, \tag{12}$$

то $\boldsymbol{V}$ называется собственным вектором матрицы $\boldsymbol{A}$, соответствующим собственному значению λ. ▲

Характеристический полином (11) можно разложить на множители следующим образом:

$$p(\lambda)=(-1)^n(\lambda-\lambda_1)^{m_1}(\lambda-\lambda_2)^{m_2}\cdots(\lambda-\lambda_k)^{m_k}, \tag{13}$$

Здесь m_j называются кратностью собственных значений λ_j. Сумма кратностей всех собственных значений равна n, т. е.

$$n=m_1+m_2+\cdots+m_k.$$

Следующие три результата позволяют прояснить вопрос о существовании собственных векторов.

Теорема 11.4. (a) Для каждого собственного значения λ существует по крайней мере один собственный вектор $\boldsymbol{V}$, соответствующий λ.

(b) Если λ имеет кратность r, то существует самое большее r линейно независимых векторов $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_r$, которые соответствуют λ.

Теорема 11.5. Предположим, что $\boldsymbol{A}$ — квадратная матрица и $\lambda_1, \lambda_2, \ldots, \lambda_k$ — различные собственные значения матрицы $\boldsymbol{A}$ с соответствующими собственными векторами $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_k$. Тогда $\{\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_k\}$ — множество линейно независимых векторов.

Теорема 11.6. Если собственные значения матрицы $\boldsymbol{A}$ размера $n \times n$ различны, то существует n собственных векторов $\boldsymbol{V}_j$, $j = 1, 2, \ldots, n$.

Теорема 11.4 обычно применяется следующим образом при вычислениях вручную. Собственное значение λ кратности $r \geq 1$ подставляют в уравнение

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{V} = \boldsymbol{0}. \tag{14}$$

Затем можно применить метод исключения Гаусса, чтобы получить уменьшенную форму Гаусса, которая содержит $n - k$ уравнений с n неизвестными, где $1 \leq k \leq r$. Следовательно, существует k свободных величин, которые можно выбрать. Свободные переменные можно выбрать, в некотором смысле исходя из рассуждений, как получить k линейно независимых векторов $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_k$, соответствующих λ.

Пример 11.2. Найдем собственные пары $\lambda_j, \boldsymbol{V}_j$ для матрицы

$$\boldsymbol{A} = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix}.$$

Кроме того, покажем, что собственные векторы линейно независимы.

Характеристическое уравнение $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$, которое имеет вид

$$\begin{vmatrix} 3-\lambda & -1 & 0 \\ -1 & 2-\lambda & -1 \\ 0 & -1 & 3-\lambda \end{vmatrix} = -\lambda^3 + 8\lambda^2 - 19\lambda + 12 = 0, \tag{15}$$

можно записать как $-(\lambda - 1)(\lambda - 3)(\lambda - 4) = 0$. Таким образом, три собственные значения равны $\lambda_1 = 1$, $\lambda_2 = 3$ и $\lambda_3 = 4$.

Случай (i). Подставим $\lambda_1 = 1$ в уравнение (14) и получим

$$\begin{aligned} 2x_1 - x_2 &= 0 \\ -x_1 + x_2 - x_3 &= 0 \\ -x_2 + 2x_3 &= 0. \end{aligned}$$

Так как сумма первого уравнения, удвоенного второго уравнения и третьего уравнения тождественно равна нулю, систему можно свести к двум уравнениям с тремя неизвестными:

$$\begin{aligned} 2x_1 - x_2 &= 0 \\ -x_2 + 2x_3 &= 0. \end{aligned}$$

Выберем $x_2 = 2a$, где a — произвольная постоянная, затем используем первое и второе уравнения для вычисления $x_1 = a$ и $x_3 = a$ соответственно. Таким образом, первая собственная пара равна $\lambda_1 = 1$, $\boldsymbol{V}_1 = \begin{bmatrix} a & 2a & a \end{bmatrix}' = a\begin{bmatrix} 1 & 2 & 1 \end{bmatrix}'$.

Случай (ii). Подставим $\lambda_2 = 3$ в уравнение (14) и получим

$$\begin{aligned} -x_2 &= 0 \\ -x_1 - x_2 - x_3 &= 0 \\ -x_2 &= 0. \end{aligned}$$

Это эквивалентно системе из двух уравнений:

$$\begin{aligned} x_1 + x_3 &= 0 \\ x_2 &= 0. \end{aligned}$$

Выберем $x_1 = b$, где b — произвольная постоянная, и вычислим $x_3 = -b$. Таким образом, вторая собственная пара равна $\lambda_2 = 3$, $\boldsymbol{V}_2 = \begin{bmatrix} b & 0 & -b \end{bmatrix}' = b\begin{bmatrix} 1 & 0 & -1 \end{bmatrix}'$.

Случай (iii). Подставим $\lambda_3 = 4$ в (14) и в результате получим

$$\begin{aligned} -x_1 - x_2 &= 0 \\ -x_1 - 2x_2 - x_3 &= 0 \\ -x_2 - x_3 &= 0. \end{aligned}$$

Это эквивалентно двум уравнениям:

$$\begin{aligned} x_1 + x_2 &= 0 \\ x_2 + x_3 &= 0. \end{aligned}$$

Выберем $x_3 = c$, где c — постоянная, затем используем второе уравнение для вычисления $x_2 = -c$. Затем из первого уравнения получим $x_1 = c$. Итак, третья собственная пара есть $\lambda_3 = 4$, $\boldsymbol{V}_3 = \begin{bmatrix} c & -c & c \end{bmatrix}' = c\begin{bmatrix} 1 & -1 & 1 \end{bmatrix}'$.

Для доказательства того, что векторы линейно независимы, достаточно применить теорему 11.5. Однако полезно вспомнить методы из линейной алгебры и использовать теорему 11.3. Запишем определитель

$$\det\left(\begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 & \boldsymbol{V}_3 \end{bmatrix}\right) = \begin{vmatrix} a & b & c \\ 2a & 0 & -c \\ a & -b & c \end{vmatrix} = -6abc.$$

Так как $\det\left(\begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 & \boldsymbol{V}_3 \end{bmatrix}\right) \neq 0$, из теоремы 11.3 следует, что векторы $\boldsymbol{V}_1$, $\boldsymbol{V}_2$ и $\boldsymbol{V}_3$ линейно независимы. ■

В примере 11.2 показано, как выполняя вычисления вручную, найти собственные значения, когда размерность n мала: (1) найдем коэффициенты характеристического полинома; (2) найдем его корни; (3) найдем не равные нулю решения однородной линейной системы $(\boldsymbol{A}-\lambda\boldsymbol{I})\boldsymbol{V}=\boldsymbol{0}$. Выберем распространенный подход изучения методов степеней, Якоби и QR-алгоритма. QR-алгоритм и его варианты используются во всех специализированных пакетах прикладных программ, таких как EISPACK и MATLAB [178].

Поскольку вектор $\boldsymbol{V}$ в (12) умножается справа на матрицу $\boldsymbol{A}$, его называют ***правым собственным вектором***, соответствующим λ. Существует также левый собственный вектор $\boldsymbol{Y}$, такой, что

$$\boldsymbol{Y}'\boldsymbol{A}=\lambda\boldsymbol{Y}'. \tag{16}$$

Вообще, левый собственный вектор $\boldsymbol{Y}$ не равен правому собственному вектору $\boldsymbol{V}$. Однако, если матрица $\boldsymbol{A}$ — действительная и симметричная ($\boldsymbol{A}'=\boldsymbol{A}$), то

$$\begin{aligned}(\boldsymbol{A}\boldsymbol{V})'&=\boldsymbol{V}'\boldsymbol{A}'=\boldsymbol{V}'\boldsymbol{A},\\(\lambda\boldsymbol{V})'&=\lambda\boldsymbol{V}'\end{aligned} \tag{17}$$

Поэтому правый собственный вектор $\boldsymbol{V}$ равен левому собственному вектору, когда матрица $\boldsymbol{A}$ симметрична. В последних разделах книги рассматриваются только правые собственные векторы.

Собственный вектор $\boldsymbol{V}$ — единственный с точностью до кратности постоянной. Предположим, что c — скаляр; тогда следующие вычисления показывают, что $c\boldsymbol{V}$ — собственный вектор:

$$\boldsymbol{A}(c\boldsymbol{V})=c(\boldsymbol{A}\boldsymbol{V})=c(\lambda\boldsymbol{V})=\lambda(c\boldsymbol{V}). \tag{18}$$

Чтобы получить некоторое представление о единственности, нормируем собственный вектор одним из следующих способов. Используем одну из норм вектора

$$\|\boldsymbol{X}\|_\infty=\max_{1\le k\le n}\{|x_k|\} \tag{19}$$

или

$$\|\boldsymbol{X}\|_2=\left(\sum_{k=1}^{n}|x_k|^2\right)^{1/2} \tag{20}$$

и потребуем, чтобы выполнялось либо равенство $\|\boldsymbol{X}\|_\infty=1$ либо $\|\boldsymbol{X}\|_2=1$.

Приведение матрицы к диагональному виду

Ситуация с собственным значением прояснится, если рассмотреть пример диагональной матрицы $\boldsymbol{D}$, которая имеет вид

$$\boldsymbol{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}. \tag{21}$$

Пусть $\boldsymbol{E}_j = [0\ 0\ \cdots\ 0\ 1\ 0\ \cdots\ 0]'$ — стандартный базисный вектор, где j-я компонента равна 1, а все остальные — равны 0. Тогда из

$$\boldsymbol{D}\boldsymbol{E}_j = [0\ 0\ \cdots\ 0\ \lambda_j\ 0\ \cdots\ 0]' = \lambda_j \boldsymbol{E}_j, \tag{22}$$

вытекает, что собственными парами матрицы $\boldsymbol{D}$ являются λ_j, $\boldsymbol{E}_j$, $j = 1, 2, \ldots, n$. Следовательно, требуется придумать простой способ приведения матрицы $\boldsymbol{A}$ к такому диагональному виду, чтобы собственные значения были инвариантны слева. Это приводит к следующему определению.

Определение 11.5. Говорят, что две матрицы $\boldsymbol{A}$ и $\boldsymbol{B}$ размера $n \times n$ подобны, если существует такая невырожденная матрица $\boldsymbol{K}$, что

$$\boldsymbol{B} = \boldsymbol{K}^{-1}\boldsymbol{A}\boldsymbol{K}. \tag{23}$$ ▲

Теорема 11.7. Предположим, что $\boldsymbol{A}$ и $\boldsymbol{B}$ — подобные матрицы и что λ — собственное значение матрицы $\boldsymbol{A}$ с соответствующим собственным вектором $\boldsymbol{V}$. Тогда λ — также собственное значение матрицы $\boldsymbol{B}$. Если $\boldsymbol{K}^{-1}\boldsymbol{A}\boldsymbol{K} = \boldsymbol{B}$, то $\boldsymbol{Y} = \boldsymbol{K}^{-1}\boldsymbol{V}$ — собственный вектор матрицы $\boldsymbol{B}$, соответствующий собственному значению λ.

Матрица $\boldsymbol{A}$ размера $n \times n$ называется ***приводимой к диагональному виду***, если она подобна диагональной матрице. Следующая теорема показывает роль собственых векторов в этом процессе.

Теорема 11.8 (приведение матрицы к диагональному виду). Матрица $\boldsymbol{A}$ подобна диагональной матрице $\boldsymbol{D}$ тогда и только тогда, когда она имеет n линейно независимых собственных векторов. Если матрица $\boldsymbol{A}$ подобна матрице $\boldsymbol{D}$, то

$$\begin{aligned} \boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V} &= \boldsymbol{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \\ \boldsymbol{V} &= [\boldsymbol{V}_1\ \boldsymbol{V}_2\ \ldots\ \boldsymbol{V}_n], \end{aligned} \tag{24}$$

где n собственных пар — λ_j, $\boldsymbol{V}_j$, $j = 1, 2, \ldots, n$.

Из теоремы 11.8 вытекает, что каждую матрицу $\boldsymbol{A}$, имеющую n различных собственных значений, можно привести к диагональному виду.

Пример 11.3. Покажем, что следующую матрицу можно привести к диагональному виду:

$$\boldsymbol{A} = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix}.$$

В примере 11.2 найдены собственные значения $\lambda_1 = 1$, $\lambda_2 = 3$ и $\lambda_3 = 4$ и матрица собственных векторов

$$\boldsymbol{V} = [\boldsymbol{V}_1 \ \boldsymbol{V}_2 \ \boldsymbol{V}_3] = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -1 \\ 1 & -1 & 1 \end{bmatrix}.$$

Обратная матрица $\boldsymbol{V}^{-1}$ имеет вид

$$\boldsymbol{V}^{-1} = \begin{bmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Оставляем читателю проверку деталей вычисления произведения в (24):

$$\begin{bmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -1 \\ 1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}.$$

Таким образом, показано, что матрицу $\boldsymbol{A}$ можно привести к диагональному виду, т. е. $\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{D} = \operatorname{diag}(1, 3, 4)$. ■

В следующей теореме приведен более общий результат, связывающий построение матрицы с ее собственными значениями.

Теорема 11.9 (Шур). Предположим, что $\boldsymbol{A}$ — произвольная матрица размера $n \times n$. Существует невырожденная матрица $\boldsymbol{P}$, для которой выполняется $\boldsymbol{T} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$, где $\boldsymbol{T}$ — верхняя треугольная матрица, диагональные элементы которой состоят из собственных значений матрицы $\boldsymbol{A}$.

Для определенного типа структурного анализа в технике требуется выбирать базис в $\Re^n$ таким, чтобы он содержал собственные векторы матрицы $\boldsymbol{A}$. Такой выбор упрощает отчетливое представление о том, каким образом трансформируется пространство при отбражении $\boldsymbol{Y} = \boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{X}$. Напомним, что собственная пара λ_j, $\boldsymbol{V}_j$ обладает таким свойством, что $\boldsymbol{T}$ отображает вектор $\boldsymbol{V}_j$ в кратный вектор $\lambda_j \boldsymbol{V}_j$. Это свойство используется в следующей теореме.

Теорема 11.10. Предположим, что $\boldsymbol{A}$ — матрица размера $n \times n$, имеющая n линейно независимых собственных пар λ_j, $\boldsymbol{V}_j$, $j = 1, 2, \ldots, n$; тогда любой вектор $\boldsymbol{X}$ из пространства $\Re^n$ имеет единственное представление в виде линейной комбинации собственных векторов:

$$\boldsymbol{X} = c_1\boldsymbol{V}_1 + c_2\boldsymbol{V}_2 + \cdots + c_n\boldsymbol{V}_n. \tag{25}$$

Линейное преобразование $T(\boldsymbol{X}) = \boldsymbol{AX}$ отображает вектор $\boldsymbol{X}$ в вектор

$$\boldsymbol{Y} = T(\boldsymbol{X}) = c_1\lambda_1\boldsymbol{V}_1 + c_2\lambda_2\boldsymbol{V}_2 + \cdots + c_n\lambda_n\boldsymbol{V}_n. \tag{26}$$

Пример 11.4. Предположим, что $\lambda_1 = 2$, $\lambda_2 = -1$ и $\lambda_3 = 4$ — собственные значения матрицы $\boldsymbol{A}$ размера 3×3, которым соответствуют собственные векторы $\boldsymbol{V}_1 = [1\ \ 2\ \ -2]'$, $\boldsymbol{V}_2 = [-2\ \ 1\ \ 1]'$ и $\boldsymbol{V}_3 = [1\ \ 3\ \ -4]'$. Для $\boldsymbol{X} = [-1\ \ 2\ \ 1]'$ найдем образ $\boldsymbol{X}$ при отображении $T(\boldsymbol{X}) = \boldsymbol{AX}$.

Сначала следует выразить $\boldsymbol{X}$ как линейную комбинацию собственных векторов. Получаем ее, решая уравнения

$$[-1\ \ 2\ \ 1]' = c_1[1\ \ 2\ \ -2]' + c_2[-2\ \ 1\ \ 1]' + c_3[1\ \ 3\ \ -4]'$$

относительно c_1, c_2 и c_3. Заметим, что это эквивалентно решению линейной системы

$$\begin{aligned} c_1 - 2c_2 + \ c_3 &= -1 \\ 2c_1 + \ c_2 + 3c_3 &= \ 2 \\ -2c_1 + \ c_2 - 4c_3 &= \ 1. \end{aligned}$$

Полученные решения равны $c_1 = 2$, $c_2 = 1$ и $c_3 = -1$. Используя определение 11.4 для собственных векторов, найдем $T(\boldsymbol{X})$ из вычислений

$$\begin{aligned} T(\boldsymbol{X}) &= \boldsymbol{A}(2\boldsymbol{V}_1 + \boldsymbol{V}_2 - \boldsymbol{V}_3) = \\ &= 2\boldsymbol{AV}_1 + \boldsymbol{AV}_2 - \boldsymbol{AV}_3 = \\ &= 2(2\boldsymbol{V}_1) - \boldsymbol{V}_2 - 4\boldsymbol{V}_3 = \\ &= [2\ \ -5\ \ 7]'. \end{aligned}$$

■

Достоинства симметрии

Не существует простого способа для определения, насколько много линейно независимых собственных векторов имеет матрица, без того, чтобы не обратиться за помощью к более эффективным алгоритмам специализированных пакетов программ, таких как EISPACK и MATLAB. Однако известно, что действительная симметричная матрица имеет n действительных собственных векторов и что каждому собственному значению кратности m_j соответствует m_j линейно независимых векторов. Следовательно, каждую действительную симметричную матрицу можно привести к диагональному виду.

Определение 11.6 (ортогональность). Говорят, что множество векторов $\{V_1, V_2, \ldots, V_n\}$ ортогонально, если выполняется условие

$$V_j' V_k = 0 \quad \text{всякий раз, когда} \quad j \neq k. \tag{27}$$

▲

Определение 11.7 (ортонормированность). Предположим, что $\{V_1, V_2, \ldots, V_n\}$ — множество ортогональных векторов; говорят, что они ортонормированные, если все они имеют единичную норму, т. е.

$$V_j' V_k = 0 \quad \text{всякий раз, когда} \quad j \neq k. \tag{28}$$

$$V_j' V_j = 1 \quad \text{для всех} \quad j = 1, 2, \ldots, n.$$

▲

Теорема 11.11. Множество ортонормированных векторов линейно независимо.

Замечание. Нулевой вектор не принадлежит множеству ортонормированных векторов.

Определение 11.8 (ортогональная матрица). Говорят, что матрица A размера $n \times n$ ортогональна, если матрица A' является обратной к матрице A, т. е.

$$A'A = I, \tag{29}$$

что эквивалентно

$$A^{-1} = A'. \tag{30}$$

Также, матрица A ортогональна тогда и только тогда, когда столбцы (строки) A образуют множество ортонормальных векторов. ▲

Теорема 11.12. Если A — действительная симметричная матрица, то существует такая ортогональная матрица K, что

$$K'AK = K^{-1}AK = D, \tag{31}$$

где D — диагональная матрица, состоящая из собственных значений матрицы A.

Следствие 11.1. Если A — действительная симметричная матрица размера $n \times n$, то существует n линейно независимых собственных векторов матрицы A и они образуют ортогональное множество.

Следствие 11.2. Все собственные значения действительной симметричной матрицы — действительные числа.

Теорема 11.13. Собственные векторы, соответствующие различным собственным значениям симметричной матрицы, ортогональны.

Теорема 11.14. Симметричная матрица A положительно определена тогда и только тогда, когда все собственные значения A положительны.

Оценка величин собственных значений

Такая оценка полезна для нахождения грани величин собственных значений матрицы A. Следующие результаты дают об этом некоторое представление.

Определение 11.9 (норма матрицы). Пусть $\|X\|$ — норма вектора. Тогда естественная норма матрицы задается так:

$$\|A\| = \max_{\|X\|=1} \left\{ \frac{\|AX\|}{\|X\|} \right\}. \tag{32}$$

Для нормы $\|A\|_\infty$ имеет место следующая формула:

$$\|A\|_\infty = \max_{1 \le i \le n} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\}. \tag{33}$$

▲

Теорема 11.15. Если λ — любое собственное значение матрицы A, то

$$|\lambda| \le \|A\|, \tag{34}$$

для любой естественной нормы матрицы $\|A\|$.

Теорема 11.16 (круговая теорема Гершгорина). Предположим, что A — матрица размера $n \times n$, и обозначим через C_j круг в комплексной плоскости с центром в точке a_{jj} и радиусом

$$r_j = \sum_{k=1, k \ne j}^{n} |a_{jk}| \quad \text{для каждого } j = 1, 2, \ldots, n; \tag{35}$$

т. е. C_j содержит все такие комплексные числа $z = x + iy$, что

$$C_j = \{z : |z - a_{jj}| \le r_j\}. \tag{36}$$

Если $S = \bigcup_{i=1}^{n} C_i$, то все собственные значения матрицы A принадлежат к множеству S. Кроме того, объединение любых k из этих кругов, которые не пересекают оставшиеся $n - k$ кругов, должны содержать точно k (считая кратные) собственных значений.

Теорема 11.17 (теорема о спектральном радиусе). Пусть A — симметричная матрица. Спектральный радиус матрицы A равен $\|A\|_2$ и удовлетворяет соотношению

$$\|A\|_2 = \max\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|\}. \tag{37}$$

Обзор методов

Для задач, использующих умеренного размера симметричные матрицы, лучше использовать метод Якоби. Для задач, включающих симметричные матрицы больших размеров (для n больше нескольких сотен), лучше использовать метод Хаусхольдера (Householder) для получения трехдиагональной матрицы, следуя QR-алгоритму. В отличие от действительных симметричных матриц действительные несимметричные матрицы могут иметь комплексные собственные значения и векторы.

Для матриц, обладающих мажорирующими собственными значениями, можно использовать метод степеней для нахождения мажорирующих собственных векторов. Технику уменьшения размера матрицы можно использовать с этого момента, чтобы найти несколько первых близких к мажорирующим собственных векторов. По отношению к действительным несимметричным матрицам метод Хаусхольдера используется для получения матрицы Гессенберга (Hessenberg), следуя LR- или QR-алгоритму.

Упражнения к разделу 11.1

1. Для каждой из следующих матриц найдите (i) характеристический полином $p(\lambda)$, (ii) собственные значения и (iii) соответствующие им собственные векторы.

   (a) $\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}$ (b) $\boldsymbol{A} = \begin{bmatrix} 1 & 6 \\ 9 & 2 \end{bmatrix}$ (c) $\boldsymbol{A} = \begin{bmatrix} -2 & 3 \\ 3 & -2 \end{bmatrix}$

   (d) $\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ -1 & 3 & 2 \end{bmatrix}$ (e) $\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 3 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 4 \end{bmatrix}$

2. Определите спектральный радиус каждой матрицы из упр. 1.
3. Определите $\|\boldsymbol{A}\|_2$ и $\|\boldsymbol{A}\|_\infty$ — нормы каждой матрицы из упр. 1.
4. Определите, какую матрицу, если это возможно, из упр. 1 можно привести к диагональному виду. Для каждой приведенной к диагональному виду матрицы из упр. 1 найдите матрицы $\boldsymbol{V}$ и $\boldsymbol{D}$ из теоремы 11.8 и вычислите произведение матриц, как в (24).
5. (a) Для каждого фиксированного θ покажите, что

$$\boldsymbol{R} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

   есть ортогональная матрица.

   Замечание. Матрица $\boldsymbol{R}$ называется матрицей вращения.

   (b) Определите все числа θ, для которых все собственные значения матрицы $\boldsymbol{R}$ действительные.

6. В разделе 3.2 введены плоскости вращений $\boldsymbol{R}_x(\alpha)$, $\boldsymbol{R}_y(\beta)$ и $\boldsymbol{R}_z(\gamma)$.
   (a) Покажите, что для любого фиксированного α, β и γ, $\boldsymbol{R}_x(\alpha)$, $\boldsymbol{R}_y(\beta)$ и $\boldsymbol{R}_z(\gamma)$ соответственно являются отртогональными матрицами.
   (b) Определите все значения α, β и γ, для которых все собственные значения матриц $\boldsymbol{R}_x(\alpha)$, $\boldsymbol{R}_y(\beta)$ и $\boldsymbol{R}_z(\gamma)$ соответственно будут действительными.
7. Пусть $\boldsymbol{A} = \begin{bmatrix} a+3 & 2 \\ 2 & a \end{bmatrix}$.
   (a) Покажите, что ее характеристический полином равен $p(\lambda) = \lambda^2 - (3 + 2a)\lambda + a^2 - 3a - 4$.
   (b) Покажите, что $\lambda_1 = a + 4$ и $\lambda_2 = a - 1$ — собственные значения матрицы $\boldsymbol{A}$.
   (c) Покажите, что собственные векторы матрицы $\boldsymbol{A}$ равны $\boldsymbol{V}_1 = \begin{bmatrix} 2 & 1 \end{bmatrix}'$ и $\boldsymbol{V}_2 = \begin{bmatrix} -1 & 2 \end{bmatrix}'$.
8. Предположим, что λ, $\boldsymbol{V}$ образуют собственную пару матрицы $\boldsymbol{A}$. Докажите, что если k — положительное целое число, то λ^k, $\boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A}^k$.
9. Предположим, что $\boldsymbol{V}$ — собственный вектор матрицы $\boldsymbol{A}$, соответствующий собственному значению $\lambda = 3$. Докажите, что $\lambda = 9$ — собственное значение матрицы $\boldsymbol{A}^2$, соответствующее $\boldsymbol{V}$.
10. Предположим, что $\boldsymbol{V}$ — собственный вектор матрицы $\boldsymbol{A}$ соответствующий собственному значению $\lambda = 2$. Докажите, что $\lambda = \frac{1}{2}$ — собственное значение матрицы $\boldsymbol{A}^{-1}$, соответствующее $\boldsymbol{V}$.
11. Предположим, что $\boldsymbol{V}$ — собственный вектор матрицы $\boldsymbol{A}$, соответствующий собственному значению $\lambda = 5$. Докажите, что $\lambda = 4$ — собственное значение матрицы $\boldsymbol{A} - \boldsymbol{I}$, соответствующее $\boldsymbol{V}$.
12. Пусть $\boldsymbol{A}$ — квадратная матрица размера $n \times n$ с заданным характеристическим полиномом $p(\lambda)$

$$p(\lambda) = \det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \\ = (-1)^n(\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_{n-1}\lambda + c_n).$$

   (a) Покажите, что постоянный член полинома $p(\lambda)$ равен $c_n = (-1)^n \det(\boldsymbol{A})$.
   (b) Покажите, что коэффициент при λ^{n-1} равен $c_1 = -(a_{11} + a_{22} + \cdots + a_{nn})$.
13. Предположим, что матрица $\boldsymbol{A}$ подобна диагональной матрице, т. е.

$$\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n).$$

Докажите, что если k — положительное целое число, то

$$\boldsymbol{A}^k = \boldsymbol{V} \operatorname{diag}(\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k)\boldsymbol{V}^{-1}.$$

11.2. Метод степеней

Опишем метод степеней для вычисления мажорирующей собственной пары. Его обобщение на обратный метод степеней можно использовать для нахождения любого собственного значения при условии, что известно хорошее начальное приближение. В некоторых схемах нахождения собственных значений используются другие методы, которые сходятся быстрее, однако имеют ограниченную точность. В этом случае обращаются к обратному методу степеней, чтобы уточнить численные значения и получить нужную точность. Для обсуждения этой ситуации понадобятся следующие определения.

Определение 11.10. Если λ_1 — собственное значение матрицы $\boldsymbol{A}$, большее по абсолютной величине, чем другие собственные значения, то оно называется ***мажорирующим собственным значением***. Собственный вектор $\boldsymbol{V}_1$, соответствующий λ_1, называется ***мажорирующим собственным вектором***. ▲

Определение 11.11. Говорят, что собственный вектор $\boldsymbol{V}$ нормализован, если координата, имеющая наибольшее значение по абсолютной величине, равна единице (т. е. наибольшая координата вектора $\boldsymbol{V}$ равна 1). ▲

Можно легко нормализовать собственный вектор $[v_1 \ v_2 \ \dots \ v_n]'$, образуя новый вектор $\boldsymbol{V} = (1/c)[v_1 \ v_2 \ \dots \ v_n]'$, где $c = v_j$ и $|v_j| = \max_{1 \le i \le n}\{|v_i|\}$.

Предположим, что матрица $\boldsymbol{A}$ имеет мажорирующее собственное значение λ и что существует единственный нормализованный собственный вектор $\boldsymbol{V}$, соответствующий λ. Эту собственную пару λ, $\boldsymbol{V}$ можно найти, выполнив следующую итеративную процедуру, которая называется ***методом степеней***. Начнем с вектора

$$\boldsymbol{X}_0 = [1 \ 1 \ \dots \ 1]'. \tag{1}$$

Рекуррентно генерируем последовательность $\{\boldsymbol{X}_k\}$, используя соотношение

$$\begin{aligned} \boldsymbol{Y}_k &= \boldsymbol{A}\boldsymbol{X}_k, \\ \boldsymbol{X}_{k+1} &= \frac{1}{c_{k+1}}\boldsymbol{Y}_k, \end{aligned} \tag{2}$$

где c_{k+1} — наибольшая по величине координата вектора $\boldsymbol{Y}_k$ (в случае совпадения выбирают первую координату). Последовательности $\{\boldsymbol{X}_k\}$ и $\{c_k\}$ сходятся соответственно к $\boldsymbol{V}$ и λ:

$$\lim_{k\to\infty} \boldsymbol{X}_k = \boldsymbol{V} \quad \text{и} \quad \lim_{k\to\infty} c_k = \lambda. \tag{3}$$

Замечание. Если $\boldsymbol{X}_0$ — собственный вектор и $\boldsymbol{X}_0 \neq \boldsymbol{V}$, то следует выбрать другой начальный вектор.

Пример 11.5. Используем метод степеней, чтобы найти мажорирующие собственные значение и вектор матрицы

$$A = \begin{bmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{bmatrix}.$$

Начнем с $X_0 = [1\ 1\ 1]'$ и используем для генерирования последовательности векторов $\{X_k\}$ и постоянных $\{c_k\}$ формулу (2). После первой итерации получаем

$$\begin{bmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \\ 12 \end{bmatrix} = 12 \begin{bmatrix} \frac{1}{2} \\ \frac{2}{3} \\ 1 \end{bmatrix} = c_1 X_1.$$

Вторая итерация дает

$$\begin{bmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ \frac{2}{3} \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{7}{3} \\ \frac{10}{3} \\ \frac{16}{3} \end{bmatrix} = \frac{16}{3} \begin{bmatrix} \frac{7}{16} \\ \frac{5}{8} \\ 1 \end{bmatrix} = c_2 X_2.$$

Итерация генерирует последовательность $\{X_k\}$ (где X_k — нормализованный вектор):

$$12 \begin{bmatrix} \frac{1}{2} \\ \frac{2}{3} \\ 1 \end{bmatrix}, \frac{16}{3} \begin{bmatrix} \frac{7}{16} \\ \frac{5}{8} \\ 1 \end{bmatrix}, \frac{9}{2} \begin{bmatrix} \frac{5}{12} \\ \frac{11}{18} \\ 1 \end{bmatrix}, \frac{38}{9} \begin{bmatrix} \frac{31}{76} \\ \frac{23}{38} \\ 1 \end{bmatrix}, \frac{78}{19} \begin{bmatrix} \frac{21}{52} \\ \frac{47}{78} \\ 1 \end{bmatrix}, \frac{158}{39} \begin{bmatrix} \frac{127}{316} \\ \frac{95}{158} \\ 1 \end{bmatrix}, \ldots.$$

Последовательность векторов сходится к $V = [\frac{2}{3}\ \frac{3}{5}\ 1]'$, а последовательность постоянных — к $\lambda = 4$ (табл. 11.1). Можно доказать, что скорость сходимости линейна. ■

Теорема 11.18 (метод степеней). Предположим, что матрица A размера $n \times n$ имеет n различных собственных значений $\lambda_1, \lambda_2, \ldots, \lambda_n$ и что они расположены в порядке убывания, т. е.

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|. \tag{4}$$

Если вектор X_0 выбран подходящим образом, то последовательности $\{X_k = [x_1^{(k)}\ x_2^{(k)}\ \ldots\ x_n^{(k)}]'\}$ и $\{c_k\}$ рекуррентно генерируются согласно формулам

$$Y_k = AX_k \tag{5}$$

и

$$X_{k+1} = \frac{1}{c_{k+1}} Y_k, \tag{6}$$

Таблица 11.1. Метод степеней, используемый в примере 11.5 для нахождения нормализованного мажорирующего собственного вектора $V = \left[\frac{2}{5} \quad \frac{3}{5} \quad 1\right]'$ и соответствующего собственного значения $\lambda = 4$

AX_k =	Y_k	=	$c_{k+1}X_{k+1}$
AX_0 =	[6,000000 8,000000 12,00000]′	=	12,00000[0,500000 0,666667 1]′ = c_1X_1
AX_1 =	[2,333333 3,333333 5,333333]′	=	5,333333[0,437500 0,625000 1]′ = c_2X_2
AX_2 =	[1,875000 2,750000 4,500000]′	=	4,500000[0,416667 0,611111 1]′ = c_3X_3
AX_3 =	[1,722222 2,555556 4,222222]′	=	4,222222[0,407895 0,605263 1]′ = c_4X_4
AX_4 =	[1,657895 2,473684 4,105263]′	=	4,105263[0,403846 0,602564 1]′ = c_5X_5
AX_5 =	[1,628205 2,435897 4,051282]′	=	4,051282[0,401899 0,601266 1]′ = c_6X_6
AX_6 =	[1,613924 2,417722 4,025316]′	=	4,025316[0,400943 0,600629 1]′ = c_7X_7
AX_7 =	[1,606918 2,408805 4,012579]′	=	4,012579[0,400470 0,600313 1]′ = c_8X_8
AX_8 =	[1,603448 2,404389 4,006270]′	=	4,006270[0,400235 0,600156 1]′ = c_9X_9
AX_9 =	[1,601721 2,402191 4,003130]′	=	4,003130[0,400117 0,600078 1]′ = $c_{10}X_{10}$
AX_{10} =	[1,600860 2,401095 4,001564]′	=	4,001564[0,400059 0,600039 1]′ = $c_{11}X_{11}$

где

$$c_{k+1} = x_j^{(k)} \quad \text{и} \quad x_j^{(k)} = \max_{1 \le i \le n} \{|x_i^{(k)}|\}, \tag{7}$$

соответственно будут сходиться к мажорирующему собственному вектору V_1 и собственному значению λ_1, т. е.

$$\lim_{k \to \infty} X_k = V_1 \quad \text{и} \quad \lim_{k \to \infty} c_k = \lambda_1. \tag{8}$$

Доказательство. Так как матрица A имеет n собственных значений, существует n соответствующих собственных векторов V_j, $j = 1, 2, \ldots, n$, которые линейно независимы, нормализованы и образуют базис n-мерного пространства. Следовательно, начальный вектор X_0 можно выразить как линейную комбинацию

$$X_0 = b_1V_1 + b_2V_2 + \cdots + b_nV_n. \tag{9}$$

Предположим, что вектор $X_0 = [x_1 \; x_2 \; \ldots \; x_n]'$ выбран таким образом, что $b_1 \neq 0$. Предположим также, что координаты вектора X_0 выбраны так, что $\max_{1 \le j \le n}\{|x_j|\} = 1$. Поскольку $\{V_j\}_{j=1}^n$ — собственные векторы матрицы A, умножение AX_0 и последующая нормализация дают

$$\begin{aligned} Y_0 = AX_0 &= A(b_1V_1 + b_2V_2 + \cdots + b_nV_n) = \\ &= b_1AV_1 + b_2AV_2 + \cdots + b_nAV_n = \\ &= b_1\lambda_1V_1 + b_2\lambda_2V_2 + \cdots + b_n\lambda_nV_n = \\ &= \lambda_1\left(b_1V_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)V_2 + \cdots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)V_n\right), \end{aligned} \tag{10}$$

и

$$\boldsymbol{X}_1 = \frac{\lambda_1}{c_1}\left(b_1\boldsymbol{V}_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)\boldsymbol{V}_2 + \cdots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)\boldsymbol{V}_n\right).$$

После k итераций приходим к соотношениям

(11)

$$\begin{aligned}
\boldsymbol{Y}_{k-1} &= \boldsymbol{A}\boldsymbol{X}_{k-1} = \\
&= \boldsymbol{A}\frac{\lambda_1^{k-1}}{c_1c_2\cdots c_{k-1}}\left(b_1\boldsymbol{V}_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k-1}\boldsymbol{V}_2 + \cdots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k-1}\boldsymbol{V}_n\right) = \\
&= \frac{\lambda_1^{k-1}}{c_1c_2\cdots c_{k-1}}\left(b_1\boldsymbol{A}\boldsymbol{V}_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k-1}\boldsymbol{A}\boldsymbol{V}_2 + \cdots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k-1}\boldsymbol{A}\boldsymbol{V}_n\right) = \\
&= \frac{\lambda_1^{k-1}}{c_1c_2\cdots c_{k-1}}\left(b_1\lambda_1\boldsymbol{V}_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k-1}\lambda_2\boldsymbol{V}_2 + \cdots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k-1}\lambda_n\boldsymbol{V}_n\right) = \\
&= \frac{\lambda_1^{k}}{c_1c_2\cdots c_{k-1}}\left(b_1\boldsymbol{V}_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k}\boldsymbol{V}_2 + \cdots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k}\boldsymbol{V}_n\right),
\end{aligned}$$

и

$$\boldsymbol{X}_k = \frac{\lambda_1^k}{c_1c_2\cdots c_k}\left(b_1\boldsymbol{V}_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k-1}\boldsymbol{V}_2 + \cdots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k-1}\boldsymbol{V}_n\right).$$

Предполагалось, что $|\lambda_j|/|\lambda_1| < 1$ для каждого $j = 2, 3, \ldots, n$, поэтому имеем

(12) $$\lim_{k\to\infty} b_j\left(\frac{\lambda_j}{\lambda_1}\right)^k \boldsymbol{V}_j = \boldsymbol{0} \quad \text{для каждого } j = 2,\ 3,\ \ldots,\ n.$$

Значит,

(13) $$\lim_{k\to\infty} \boldsymbol{X}_k = \lim_{k\to\infty} \frac{b_1\lambda_1^k}{c_1c_2\cdots c_k}\boldsymbol{V}_1.$$

Предполагалось, что оба вектора (и $\boldsymbol{X}_k$, и $\boldsymbol{V}_1$) были нормализованы и их наибольшая компонента равнялась 1. Поэтому предельный вектор в правой части (13) будет нормализован и его наибольшая компонента равна 1. Следовательно, предел скалярного множителя вектора $\boldsymbol{V}_1$ в правой части (13) существует и его значение должно быть равно 1, т. е.

(14) $$\lim_{k\to\infty} \frac{b_1\lambda_1^k}{c_1c_2\cdots c_k} = 1.$$

Поэтому последовательность векторов $\{\boldsymbol{X}_k\}$ сходится к мажорирующему собственному вектору:

$$\lim_{k\to\infty} \boldsymbol{X}_k = \boldsymbol{V}_1. \tag{15}$$

Замена k на $k-1$ в членах последовательности в формуле (14) даст

$$\lim_{k\to\infty} \frac{b_1\lambda_1^{k-1}}{c_1c_2\cdots c_{k-1}} = 1,$$

и, если разделить обе части формулы (14) до и после замены одна на другую, можно получить

$$\lim_{k\to\infty} \frac{\lambda_1}{c_k} = \lim_{k\to\infty} \frac{b_1\lambda_1^k/(c_1c_2\cdots c_k)}{b_1\lambda_1^{k-1}/(c_1c_2\cdots c_{k-1})} = \frac{1}{1} = 1.$$

Значит, последовательность постоянных $\{c_k\}$ сходится к мажорирующему собственному значению:

$$\lim_{k\to\infty} c_k = \lambda_1. \tag{16}$$

Доказательство теоремы закончено. •

Скорость сходимости

Из соотношения (12) видно, что коэффициент $\boldsymbol{V}_j$ в $\boldsymbol{X}_k$ стремится к нулю пропорционально $(\lambda_j/\lambda_1)^k$ и что скорость сходимости $\{\boldsymbol{X}_k\}$ к $\boldsymbol{V}_1$ определяется членами $(\lambda_2/\lambda_1)^k$. Следовательно, скорость сходимость линейная. Аналогично сходимость постоянных $\{c_k\}$ к λ_1 линейная. Метод Δ^2 Эйткена можно использовать для любой линейно сходящейся последовательности $\{p_k\}$, чтобы сформировать новую последовательность

$$\left\{\widehat{p}_k = \frac{(p_{k+1}-p_k)^2}{p_{k+2}-2p_{k+1}+p_k}\right\},$$

которая сходится быстрее. В примере 11.4 можно применить метод Δ^2 Эйткена, чтобы ускорить сходимость последовательности постоянных $\{c_k\}$ так же, как сходимость первых двух компонент последовательности векторов $\{\boldsymbol{X}_k\}$. Полученные таким способом результаты сравниваются с исходными последовательностями в табл. 11.2.

Таблица 11.2. Сравнение скорости сходимости метода степеней и ускорения метода степеней с использованием техники метода Δ^2 Эйткена

$c_k Y_k$	$\widehat{c}_k \widehat{X}_k$
$c_1 X_1 = 12{,}000000[0{,}5000000\ \ 0{,}6666667\ \ 1]'$;	$4{,}3809524[0{,}4062500\ \ 0{,}6041667\ \ 1]' = \widehat{c}_1 \widehat{X}_1$
$c_2 X_2 = 5{,}3333333[0{,}4375000\ \ 0{,}6250000\ \ 1]'$;	$4{,}0833333[0{,}4015152\ \ 0{,}6010101\ \ 1]' = \widehat{c}_2 \widehat{X}_2$
$c_3 X_3 = 4{,}5000000[0{,}4166667\ \ 0{,}6111111\ \ 1]'$;	$4{,}0202020[0{,}4003759\ \ 0{,}6002506\ \ 1]' = \widehat{c}_3 \widehat{X}_3$
$c_4 X_4 = 4{,}2222222[0{,}4078947\ \ 0{,}6052632\ \ 1]'$;	$4{,}0050125[0{,}4000938\ \ 0{,}6000625\ \ 1]' = \widehat{c}_4 \widehat{X}_4$
$c_5 X_5 = 4{,}1052632[0{,}4038462\ \ 0{,}6025641\ \ 1]'$;	$4{,}0012508[0{,}4000234\ \ 0{,}6000156\ \ 1]' = \widehat{c}_5 \widehat{X}_5$
$c_6 X_6 = 4{,}0512821[0{,}4018987\ \ 0{,}6012658\ \ 1]'$;	$4{,}0003125[0{,}4000059\ \ 0{,}6000039\ \ 1]' = \widehat{c}_6 \widehat{X}_6$
$c_7 X_7 = 4{,}0253165[0{,}4009434\ \ 0{,}6006289\ \ 1]'$;	$4{,}0000781[0{,}4000015\ \ 0{,}6000010\ \ 1]' = \widehat{c}_7 \widehat{X}_7$
$c_8 X_8 = 4{,}0125786[0{,}4004702\ \ 0{,}6003135\ \ 1]'$;	$4{,}0000195[0{,}4000004\ \ 0{,}6000002\ \ 1]' = \widehat{c}_8 \widehat{X}_8$
$c_9 X_9 = 4{,}0062696[0{,}4002347\ \ 0{,}6001565\ \ 1]'$;	$4{,}0000049[0{,}4000001\ \ 0{,}6000001\ \ 1]' = \widehat{c}_9 \widehat{X}_9$
$c_{10} X_{10} = 4{,}0031299[0{,}4001173\ \ 0{,}6000782\ \ 1]'$;	$4{,}0000012[0{,}4000000\ \ 0{,}6000000\ \ 1]' = \widehat{c}_{10} \widehat{X}_{10}$

Метод степеней с обратным сдвигом

Обсудим метод степеней с обратным сдвигом. Он требует хорошего начального приближения для собственного значения и затем использует итерацию для получения точного решения. Другие процедуры, такие как QM и метод Гивена, используются для получения начальных приближений. Случаи, когда собственные значения комплексные либо кратные или когда существует два собственных значения, которые имеют одинаковую величину или одинаковое по величине приближение, вызывают трудности при вычислениях и требуют более совершенных методов. Остановимся на иллюстрации тех случаев, когда собственные значения различны. Метод степеней с обратным сдвигом основан на следующих трех результатах (доказательства оставляем читателям в качестве упражнений).

Теорема 11.19 (сдвиг собственных значений). Предположим, что λ, $\boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A}$. Если α — любая постоянная, то $\lambda - \alpha$, $\boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A} - \alpha \boldsymbol{I}$.

Теорема 11.20 (обратные собственные значения). Предположим, что λ, $\boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A}$. Если $\lambda \neq 0$, то $1/\lambda$, $\boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A}^{-1}$.

Теорема 11.21. Предположим, что λ, $\boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A}$. Если $\alpha \neq \lambda$, то $1/(\lambda - \alpha)$, $\boldsymbol{V}$ — собственная пара матрицы $(\boldsymbol{A} - \alpha \boldsymbol{I})^{-1}$.

Теорема 11.22 (метод степеней с обратным сдвигом). Предположим, что матрица $\boldsymbol{A}$ размера $n \times n$ имеет различные собственные значения $\lambda_1, \lambda_2, \ldots, \lambda_n$ и рассмотрим собственное значение λ_j. Тогда можно так выбрать постоянную α, что $\mu_1 = 1/(\lambda_j - \alpha)$ будет мажорирующим собственным значением матрицы

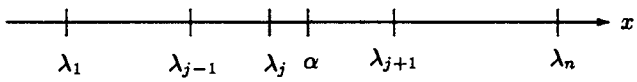


Рис. 11.2. Определение положения α для метода степеней с обратным сдвигом

$(A-\alpha I)^{-1}$. Кроме того, если соответствующим образом выбрано X_0, последовательности $\{X_k = [x_1^{(k)}\ x_2^{(k)}\ \dots\ x_n^{(k)}]'\}$ и $\{c_k\}$, которые рекуррентно генерируются согласно формулам

$$Y_k = (A-\alpha I)^{-1}X_k \tag{17}$$

и

$$X_{k+1} = \frac{1}{c_{k+1}}Y_k, \tag{18}$$

где

$$c_{k+1} = x_j^{(k)} \quad \text{и} \quad x_j^{(k)} = \max_{1\le j\le n}\{|x_i^{(k)}|\} \tag{19}$$

будут сходиться к мажорирующей собственной паре μ_1, V_j матрицы $(A-\alpha I)^{-1}$. И наконец, соответствующее собственное значение матрицы A задается следующей формулой:

$$\lambda_j = \frac{1}{\mu_1} + \alpha. \tag{20}$$

Замечание. Применяя на практике теорему 11.22 для вычисления Y_k, на каждом шаге необходимо решить линейную систему $(A-\alpha I)Y_k = X_k$.

Доказательство. Можно предположить без потери общности, что $\lambda_1 < \lambda_2 < \dots < \lambda_n$. Выберем число α ($\alpha \neq \lambda_j$), которое ближе к λ_j, чем любое другое собственное значение (рис. 11.2), т. е.

$$|\lambda_j - \alpha| < |\lambda_i - \alpha| \quad \text{для каждого } i = 1,\ 2,\ \dots,\ j-1,\ j+1,\ \dots,\ n. \tag{21}$$

Согласно теореме 11.21 $1/(\lambda_j-\alpha)$, V — собственная пара матрицы $(A-\alpha I)^{-1}$. Из соотношения (21) вытекает, что $1/|\lambda_i-\alpha| < 1/|\lambda_j-\alpha|$ для каждого $i \neq j$. Таким образом, $\mu_1 = 1/(\lambda_j-\alpha)$ — мажорирующее собственное значение матрицы $(A-\alpha I)^{-1}$. Метод степеней с обратным сдвигом использует модификацию метода степеней для определения собственной пары μ_1, V_j. Затем, вычисляя $\lambda_j = 1/\mu_1 + \alpha$, получаем требуемое собственное значение матрицы A. •

Таблица 11.3. Метод степеней с обратным сдвигом для матрицы $(A - 4{,}2I)^{-1}$ из примера 11.6: сходимость к собственному вектору $V = \left[\frac{2}{5}\ \ \frac{3}{5}\ \ 1\right]'$ и $\mu_1 = -5$

$(A-\alpha I)^{-1}X_k =$	$c_{k+1}X_{k+1}$		
$(A-\alpha I)^{-1}X_0 = -23{,}18181818$	[0,4117647059	0,6078431373	$1]' = c_1X_1$
$(A-\alpha I)^{-1}X_1 = -5{,}356506239$	[0,4009983361	0,6006655574	$1]' = c_2X_2$
$(A-\alpha I)^{-1}X_2 = -5{,}030252609$	[0,4000902120	0,6000601413	$1]' = c_3X_3$
$(A-\alpha I)^{-1}X_3 = -5{,}002733697$	[0,4000081966	0,6000054644	$1]' = c_4X_4$
$(A-\alpha I)^{-1}X_4 = -5{,}000248382$	[0,4000007451	0,6000004967	$1]' = c_5X_5$
$(A-\alpha I)^{-1}X_5 = -5{,}000022579$	[0,4000000677	0,6000000452	$1]' = c_6X_6$
$(A-\alpha I)^{-1}X_6 = -5{,}000002053$	[0,4000000062	0,6000000041	$1]' = c_7X_7$
$(A-\alpha I)^{-1}X_7 = -5{,}000000187$	[0,4000000006	0,6000000004	$1]' = c_8X_8$
$(A-\alpha I)^{-1}X_8 = -5{,}000000017$	[0,4000000001	0,6000000000	$1]' = c_9X_9$

Пример 11.6. Применим метод степеней с обратным сдвигом, чтобы найти собственные пары матрицы

$$A = \begin{bmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{bmatrix}.$$

Воспользуемся тем фактом, что собственные значения матрицы A равны $\lambda_1 = 4$, $\lambda_2 = 2$ и $\lambda_3 = 1$, и выберем для каждого случая соответствующие α и начальный вектор.

Случай (i). Для собственного значения $\lambda_1 = 4$ выберем $\alpha = 4{,}2$ и начальный вектор $X_0 = [1\ \ 1\ \ 1]'$. Сначала образуем матрицу $A - 4{,}2I$ и решим уравнение

$$\begin{bmatrix} -4{,}2 & 11 & -5 \\ -2 & 12{,}8 & -7 \\ -4 & 26 & -14{,}2 \end{bmatrix} Y_0 = X_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

а затем получим вектор $Y_0 = [-9{,}545454545\ \ -14{,}09090909\ \ -23{,}18181818]'$. Затем вычислим $c_1 = -23{,}18181818$ и $X_1 = [0{,}4117647059\ \ 0{,}6078431373\ \ 1]'$. Значения, генерируемые итерацией, приведены в табл. 11.3. Последовательность $\{c_k\}$ сходится к $\mu_1 = -5$, которое является мажорирующим собственным значением матрицы $(A - 4{,}2I)^{-1}$, и $\{X_k\}$ сходится к $V_1 = \left[\frac{2}{5}\ \ \frac{3}{5}\ \ 1\right]'$. Собственное значение λ_1 матрицы A получаем следующим образом: $\lambda_1 = 1/\mu_1 + \alpha = 1/(-5) + 4{,}2 = -0{,}2 + 4{,}2 = 4$.

Таблица 11.4. Метод степеней с обратным сдвигом для матрицы $(\boldsymbol{A} - 2{,}1\boldsymbol{I})^{-1}$ из примера 11.6: сходимость к мажорирующему собственному вектору $\boldsymbol{V} = \left[\frac{1}{4}\ \ \frac{1}{2}\ \ 1\right]$ и $\mu_1 = -10$

$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_k =$	$c_{k+1}\boldsymbol{X}_{k+1}$		
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_0 =$ 42,63157895	[0,2592592593	0,5061728395	$1]' = c_1\boldsymbol{X}_1$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_1 =$ −9,350227420	[0,2494788047	0,4996525365	$1]' = c_2\boldsymbol{X}_2$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_2 =$ −10,03657511	[0,2500273314	0,5000182209	$1]' = c_3\boldsymbol{X}_3$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_3 =$ −9,998082009	[0,2499985612	0,4999990408	$1]' = c_4\boldsymbol{X}_4$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_4 =$ −10,00010097	[0,2500000757	0,5000000505	$1]' = c_5\boldsymbol{X}_5$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_5 =$ −9,999994686	[0,2499999960	0,4999999973	$1]' = c_6\boldsymbol{X}_6$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_6 =$ −10,00000028	[0,2500000002	0,5000000001	$1]' = c_7\boldsymbol{X}_7$

Таблица 11.5. Метод степеней с обратным сдвигом для матрицы $(\boldsymbol{A} - 0{,}875\boldsymbol{I})^{-1}$ из примера 11.6: сходимость к мажорирующему собственному вектору $\boldsymbol{V} = \left[\frac{1}{2}\ \ \frac{1}{2}\ \ 1\right]'$ и $\mu_1 = 8$

$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_k =$	$c_{k+1}\boldsymbol{X}_{k+1}$		
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_0 =$ −30,40000000	[0,5052631579	0,4947368421	$1]' = c_1\boldsymbol{X}_1$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_1 =$ 8,404210526	[0,5002004008	0,4997995992	$1]' = c_2\boldsymbol{X}_2$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_2 =$ 8,015390782	[0,5000080006	0,4999919994	$1]' = c_3\boldsymbol{X}_3$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_3 =$ 8,000614449	[0,5000003200	0,4999996800	$1]' = c_4\boldsymbol{X}_4$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_4 =$ 8,000024576	[0,5000000128	0,4999999872	$1]' = c_5\boldsymbol{X}_5$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_5 =$ 8,000000983	[0,5000000005	0,4999999995	$1]' = c_6\boldsymbol{X}_6$
$(\boldsymbol{A}-\alpha\boldsymbol{I})^{-1}\boldsymbol{X}_6 =$ 8,000000039	[0,5000000000	0,5000000000	$1]' = c_7\boldsymbol{X}_7$

Случай (ii). Для собственного значения $\lambda_2 = 2$ выберем $\alpha = 2{,}1$ и начальный вектор $\boldsymbol{X}_0 = [1\ \ 1\ \ 1]'$. Сформируем матрицу $\boldsymbol{A} - 2{,}1\boldsymbol{I}$, решим уравнение

$$\begin{bmatrix} -2{,}1 & 11 & -5 \\ -2 & 14{,}9 & -7 \\ -4 & 26 & -12{,}1 \end{bmatrix} \boldsymbol{Y}_0 = \boldsymbol{X}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

и получим вектор $\boldsymbol{Y}_0 = [11{,}05263158\ \ 21{,}57894737\ \ 42{,}63157895]'$. Затем $c_1 = 42{,}63157895$ и вектор $\boldsymbol{X}_1 = [0{,}2592592593\ \ 0{,}5061728395\ \ 1]'$. Итерационная процедура получения значений приведена в табл. 11.4. Мажорирующее собственное значение матрицы $(\boldsymbol{A} - 2{,}1\boldsymbol{I})^{-1}$ равно $\mu_1 = -10$ и собственная пара матрицы $\boldsymbol{A}$ равна $\lambda_2 = 1/(-10) + 2{,}1 = -0{,}1 + 2{,}1 = 2$ и $\boldsymbol{V}_2 = \left[\frac{1}{4}\ \ \frac{1}{2}\ \ 1\right]'$.

Случай (iii). Для собственного значения $\lambda_3 = 1$ выберем $\alpha = 0{,}875$ и начальный вектор $\boldsymbol{X}_0 = [0\ \ 1\ \ 1]'$. Итерационная процедура получения значений приведена в табл. 11.5. Мажорирующее собственное значение матрицы $(\boldsymbol{A} - 0{,}875\boldsymbol{I})^{-1}$ равно $\mu_1 = 8$ и собственная пара матрицы $\boldsymbol{A}$ равна $\lambda_3 = 1/8 + 0{,}875 = 0{,}125 +$

$+ 0{,}875 = 1$ и $\boldsymbol{V}_3 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}'$. Последовательность векторов $\{\boldsymbol{X}_k\}$ при начальном векторе $\begin{bmatrix} 0 & 1 & 1 \end{bmatrix}'$ сходится на седьмой итерации. (Вычислительные трудности возникли, когда использовался $\boldsymbol{X}_0 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}'$, и процедура сошлась значительно позже.) ■

Программа 11.1 (метод степеней). Программа предназначена для вычисления мажорирующего собственного значения λ_1 и соответствующего ему вектора $\boldsymbol{V}_1$ для матрицы $\boldsymbol{A}$ размера $n \times n$. Предполагается, что n собственных значений имеют мажорирующее свойство $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| > 0$.

```
function [lambda,V]=power1(A,X,epsilon,max1)
%Вход  - A - матрица размера n x n
%      - X - начальный вектор размера n x 1
%      - epsilon - допустимое значение
%      - max1 - максимальное число итераций
%Выход - lambda - мажорирующее собственное значение
%      - V - мажорирующий собственный вектор
%Инициализация параметров
lambda=0;
cnt=0;
err=1;
state=1;
while ((cnt<=max1)&(state==1))
   Y=A*X;
   %Нормализация Y
   [m j]=max(abs(Y));
   c1=m;
   dc=abs(lambda-c1);
   Y=(1/c1)*Y;
   %Обновление X и lambda и проверка сходимости
   dv=norm(X-Y);
   err=max(dc,dv);
   X=Y;
   lambda=c1;
   state=0;
   if(err>epsilon)
      state=1;
   end
   cnt=cnt+1;
end
V=X;
```

Программа 11.2 (метод степеней с обратным сдвигом). Программа предназначена для вычисления мажорирующего собственного значения λ_j и соответствующего ему собственного вектора V_j для матрицы A размера $n \times n$. Предполагается, что n собственных значений обладают свойством $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ и что α — такое действительное число, что $|\lambda_j - \alpha| < |\lambda_i - \alpha|$ для каждого $i = 1, 2, \ldots, j-1, j+1, \ldots, n$.

```
function [lambda,V]=invpow(A,X,alpha,epsilon,max1)
%Вход  - A - матрица размера n x n
%      - X - начальный вектор размера n x 1
%      - alpha - заданый сдвиг
%      - epsilon - допустимое значение
%      - max1 - максимальное число итераций
%Выход - lambda - мажорирующее собственное значение
%      - V - мажорирующий собственный вектор
%Инициализация матрицы A-alphaI и параметров
[n n]=size(A);
A=A-alpha*eye(n);
lambda=0;
cnt=0;
err=1;
state=1;
while ((cnt<=max1)&(state==1))
   %Решение системы AY=X
   Y=A\X;
   %Нормализация Y
   [m j]=max(abs(Y));
   c1=m;
   dc=abs(lambda-c1);
   Y=(1/c1)*Y;
   %Обновление X и lambda и проверка сходимости
   dv=norm(X-Y);
   err=max(dc,dv);
   X=Y;
   lambda=c1;
   state=0;
   if (err>epsilon)
      state=1;
   end
   cnt=cnt+1;
end
lambda=alpha+1/c1;
V=X;
```

Упражнения к разделу 11.2

1. Пусть λ, $\boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A}$. Покажите, что, если α — любая постоянная, $\lambda - \alpha, \boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A} - \alpha\boldsymbol{I}$.

2. Пусть λ, $\boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A}$. Покажите, что, если $\lambda \neq 0$, $1/\lambda$, $\boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A}^{-1}$.

3. Пусть λ, $\boldsymbol{V}$ — собственная пара матрицы $\boldsymbol{A}$. Если $\alpha \neq \lambda$, покажите, что $1/(\lambda - \alpha)$, $\boldsymbol{V}$ — собственная пара матрицы $(\boldsymbol{A} - \alpha\boldsymbol{I})^{-1}$.

4. *Техника сокращения.* Предположим, что $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ — собственные значения матрицы $\boldsymbol{A}$ с соответствующими собственными векторами $\boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{V}_3, \ldots, \boldsymbol{V}_n$ и что λ_1 имеет кратность 1. Если $\boldsymbol{X}$ — любой вектор, обладающий свойством $\boldsymbol{X}'\boldsymbol{V}_1 = 1$, докажите, что матрица

$$\boldsymbol{B} = \boldsymbol{A} - \lambda_1 \boldsymbol{V}_1 \boldsymbol{X}'$$

имеет собственные значения $0, \lambda_2, \lambda_3, \ldots, \lambda_n$ с соответствующими им собственными векторами $\boldsymbol{V}_1, \boldsymbol{W}_2, \boldsymbol{W}_3, \ldots, \boldsymbol{W}_n$, где $\boldsymbol{V}_j$ и $\boldsymbol{W}_j$ связаны соотношением

$$\boldsymbol{V}_j = (\lambda - \lambda_1)\boldsymbol{W}_j + \lambda_1(\boldsymbol{X}'\boldsymbol{W}_j)\boldsymbol{V}_1 \quad \text{для каждого } j = 2, 3, \ldots, n.$$

5. *Марковский процесс и собственные значения.* Марковский процесс можно описать квадратной матрицей $\boldsymbol{A}$, все элементы которой положительны и суммы столбцов все равны 1. Например, пусть в некоторой местности люди покупают товары сортов X и Y соответственно. $\boldsymbol{P}_0 = \begin{bmatrix} x^{(0)} & y^{(0)} \end{bmatrix}'$ — такой вектор, что $x^{(0)}$ — отношение числа людей, покупающих товар сорта X, к общему числу жителей и $y^{(0)} = 1 - x^{(0)}$. Каждый месяц они решают, какой вектор покупать: тот же или другой. Вероятность того, что покупатель товара сорта X станет покупать товар сорта Y, равна 0,3. Вероятность того, что покупатель товара сорта Y станет покупать товар сорта X, равна 0,2. Переходная матрица для этого процесса такова:

$$\boldsymbol{P}_{k+1} = \boldsymbol{A}\boldsymbol{P}_k = \begin{bmatrix} 0{,}8 & 0{,}3 \\ 0{,}2 & 0{,}7 \end{bmatrix} \begin{bmatrix} x^{(k)} \\ y^{(k)} \end{bmatrix}.$$

Если $\boldsymbol{A}\boldsymbol{P}_j = \boldsymbol{P}_j$ для некоторого j, то говорят, что $\boldsymbol{P}_j = \boldsymbol{V}$ — стационарное распределение для марковского процесса. Таким образом, если существует стационарное распределение, то $\lambda = 1$ должно быть собственным значением матрицы $\boldsymbol{A}$. Дополнительно стационарно распределенный вектор $\boldsymbol{V}$ является собственным вектором, соответствующим $\lambda = 1$, (т. е. решением $(\boldsymbol{A} - \boldsymbol{I})\boldsymbol{V} = \boldsymbol{0}$).

(а) Для приведенного выше примера убедитесь, что $\lambda = 1$ — собственное значение переходной матрицы $\boldsymbol{A}$.

(b) Убедитесь, что совокупность собственных векторов, соответствующих $\lambda = 1$, равно $\{t[3/2 \;\; 1]' : t \in \Re, t \neq 0\}$.

(c) Предположим, что население города составляет 50 000. Используя результаты выполнения п. (b), проверьте, что стационарное распределение имеет вид $[\frac{3}{5} \;\; \frac{2}{5}]'$.

Алгоритмы и программы

В задачах 1–4 используйте следующее.

(a) Программу 11.1 для нахождения мажорирующей собственной пары заданной матрицы.

(b) Программу 11.2 для нахождения остальных собственных пар.

1. $\boldsymbol{A} = \begin{bmatrix} 7 & 6 & -3 \\ -12 & -20 & 24 \\ -6 & -12 & 16 \end{bmatrix}$ **2.** $\boldsymbol{A} = \begin{bmatrix} -14 & -30 & 42 \\ 24 & 49 & -66 \\ 12 & 24 & -32 \end{bmatrix}$

3. $\boldsymbol{A} = \begin{bmatrix} 2{,}5 & -2{,}5 & 3{,}0 & 0{,}5 \\ 0{,}0 & 5{,}0 & -2{,}0 & 2{,}0 \\ -0{,}5 & -0{,}5 & 4{,}0 & 2{,}5 \\ -2{,}5 & -2{,}5 & 5{,}0 & 3{,}5 \end{bmatrix}$ **4.** $\boldsymbol{A} = \begin{bmatrix} 2{,}5 & -2{,}0 & 2{,}5 & 0{,}5 \\ 0{,}5 & 5{,}0 & -2{,}5 & -0{,}5 \\ -1{,}5 & 1{,}0 & 3{,}5 & -2{,}5 \\ 2{,}0 & 3{,}0 & -5{,}0 & 3{,}0 \end{bmatrix}$

5. Предположим, что вероятность того, что покупатель товара X будет покупать товар Y или Z, соответственно равна 0,4 и 0,2. Вероятность того, что покупатель товара Y будет покупать товар X или Z, соответственно равна 0,2 и 0,2. Вероятность того, что покупатель товара Z будет покупать товар X или Y, соответственно равна 0,1 и 0,1. Переходная матрица этого процесса имеет вид

$$\boldsymbol{P}_{k+1} = \boldsymbol{A}\boldsymbol{P}_k = \begin{bmatrix} 0{,}4 & 0{,}2 & 0{,}1 \\ 0{,}4 & 0{,}6 & 0{,}1 \\ 0{,}2 & 0{,}2 & 0{,}8 \end{bmatrix} \begin{bmatrix} x^{(k)} \\ y^{(k)} \\ z^{(k)} \end{bmatrix}.$$

(a) Убедитесь, что $\lambda = 1$ является собственным значением матрицы $\boldsymbol{A}$.

(b) Определите стационарное распределение для населения в 80 000.

6. Предположим, что выпускается пять марок кофе: B_1, B_2, B_3, B_4 и B_5. Предположим, что каждый клиент покупает трехфунтовую (1 фунт = 453,6 г) банку кофе каждый месяц и 60 млн фунтов кофе продается каждый месяц. Вне зависимости от марки каждый фунт кофе приносит доход в один доллар. Кофейная промышленность эмпирически определила следующую переходную матрицу $\boldsymbol{A}$ для ежемесячной продажи кофе, где a_{ij} — вероятность того, что

клиент будет покупать марку B_i при том, что предыдущий клиент покупал кофе марки B_j.

$$A = \begin{bmatrix} 0{,}1 & 0{,}2 & 0{,}2 & 0{,}6 & 0{,}2 \\ 0{,}1 & 0{,}1 & 0{,}1 & 0{,}1 & 0{,}2 \\ 0{,}1 & 0{,}3 & 0{,}4 & 0{,}1 & 0{,}2 \\ 0{,}3 & 0{,}3 & 0{,}1 & 0{,}1 & 0{,}2 \\ 0{,}4 & 0{,}1 & 0{,}2 & 0{,}1 & 0{,}2 \end{bmatrix}$$

Рекламное агентство производителей марки B_1 гарантирует, что за \$40 млн в год они могут изменить первый столбец матрицы A на $[0{,}3 \;\; 0{,}1 \;\; 0{,}1 \; 0{,}2 \;\; 0{,}3]'$. Стоит ли производителю марки B_1 платить рекламному агентству?

7. Напишите программу, основанную на технике сокращения из упр. 4, для нахождения всех собственных значений данных матриц. Ваша программа должна обращаться к программе 11.1, как к подпрограмме для определения мажорирующего собственного значения и собственного вектора на каждой итерации.

8. Воспользуйтесь своей программой из задачи 7, чтобы найти все собственные значения следующих матриц.

(a) $A = \begin{bmatrix} 1 & 2 & -1 \\ 1 & 0 & 1 \\ 4 & -4 & 5 \end{bmatrix}$

(b) $A = [a_{ij}]$, где $a_{ij} = \begin{cases} i + j & i = j \\ ij & i \neq j \end{cases}$ и $i, j = 1, 2, \ldots, 15$.

11.3. Метод Якоби

Метод Якоби представляет собой простой для понимания алгоритм, который находит все собственные пары симметричной матрицы. Этот надежный метод дает одинаковую точность для всех собственных пар. Для матриц размера выше 10 алгоритм конкурирует с более сложными алгоритмами. Если скорость не является основным критерием, то он вполне допустим для матриц размера выше 20.

Если использовать метод Якоби, то решение будет гарантировано для всех действительных симметричных матриц. Это несущественное ограничение, так как в большинстве практических задач, которые встречаются в математике и технике, содержатся симметричные матрицы. С теоретической точки зрения метод объединяет техники, присущие более сложным алгоритмам. В поучительных целях имеет смысл подробно рассмотреть метод Якоби.

Вращение плоскости

Сначала напомним некоторые факты о преобразовании координат. Пусть $\boldsymbol{X}$ — вектор в n-мерном пространстве. Рассмотрим линейное преобразование $\boldsymbol{Y} = \boldsymbol{RX}$, где $\boldsymbol{R}$ — матрица размера $n \times n$:

$$\boldsymbol{R} = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & \cos\phi & \cdots & \sin\phi & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & -\sin\phi & \cdots & \cos\phi & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{matrix} \\ \\ \leftarrow \text{row } p \\ \\ \leftarrow \text{row } q \\ \\ \\ \end{matrix}$$

$$\begin{matrix} & & \uparrow & & \uparrow & & \\ & & \text{col } p & & \text{col } q & & \end{matrix}$$

Здесь все недиагональные элементы матрицы $\boldsymbol{R}$ равны нулю, за исключением значения $\pm\sin\phi$, и все диагональные элементы равны 1, кроме $\cos\phi$. Действие преобразования $\boldsymbol{Y} = \boldsymbol{RX}$ легко понять:

$$\begin{aligned} y_j &= x_j \qquad \text{когда } j \neq p \text{ и } j \neq q, \\ y_p &= x_p \cos\phi + x_q \sin\phi, \\ y_q &= -x_p \sin\phi + x_q \cos\phi. \end{aligned}$$

Преобразование понимается как вращение n-мерного пространства параллельно $x_p x_q$-плоскости на угол ϕ. Выбирая подходящим образом угол ϕ, можно достичь, чтобы либо $y_p = 0$, либо $y_q = 0$. Обратное преобразование $\boldsymbol{X} = \boldsymbol{R}^{-1}\boldsymbol{Y}$ вращает пространство параллельно той же плоскости $x_p x_q$ на угол $-\phi$. Отметим, что $\boldsymbol{R}$ — ортогональная матрица, т. е.

$$\boldsymbol{R}^{-1} = \boldsymbol{R}' \quad \text{или} \quad \boldsymbol{R}'\boldsymbol{R} = \boldsymbol{I}.$$

Подобные и ортогональные преобразования

Рассмотрим задачу о собственных значениях:

$$\boldsymbol{AX} = \lambda \boldsymbol{X}. \tag{1}$$

Предположим, что $\boldsymbol{K}$ — невырожденная матрица и что $\boldsymbol{B}$ определяется следующим образом:

$$\boldsymbol{B} = \boldsymbol{K}^{-1}\boldsymbol{AK}. \tag{2}$$

Умножим обе части выражения (2) справа на величину $\boldsymbol{K}^{-1}\boldsymbol{X}$ и получим

$$\begin{aligned}\boldsymbol{B}\boldsymbol{K}^{-1}\boldsymbol{X} &= \boldsymbol{K}^{-1}\boldsymbol{A}\boldsymbol{K}\boldsymbol{K}^{-1}\boldsymbol{X} = \boldsymbol{K}^{-1}\boldsymbol{A}\boldsymbol{X} = \\ &= \boldsymbol{K}^{-1}\lambda\boldsymbol{X} = \lambda\boldsymbol{K}^{-1}\boldsymbol{X}.\end{aligned} \tag{3}$$

Произведем замену переменных:

$$\boldsymbol{Y} = \boldsymbol{K}^{-1}\boldsymbol{X} \quad \text{или} \quad \boldsymbol{X} = \boldsymbol{K}\boldsymbol{Y}. \tag{4}$$

Если (4) подставить в (3), то новая задача о собственных будет значениях иметь вид

$$\boldsymbol{B}\boldsymbol{Y} = \lambda\boldsymbol{Y}. \tag{5}$$

Сравнивая выражения (1) и (5), видим, что подобное преобразование (2) сохраняет собственное значение λ. Собственные векторы другие, однако связаны с прежними как следует из (4).

Предположим, что матрица $\boldsymbol{R}$ ортогональная (т. е. $\boldsymbol{R}^{-1} = \boldsymbol{R}'$) и что $\boldsymbol{D}$ определено как

$$\boldsymbol{D} = \boldsymbol{R}'\boldsymbol{A}\boldsymbol{R}. \tag{6}$$

Умножим справа обе части выражения (6) на $\boldsymbol{R}'\boldsymbol{X}$ и получим

$$\boldsymbol{D}\boldsymbol{R}'\boldsymbol{X} = \boldsymbol{R}'\boldsymbol{A}\boldsymbol{R}\boldsymbol{R}'\boldsymbol{X} = \boldsymbol{R}'\boldsymbol{A}\boldsymbol{X} = \boldsymbol{R}'\lambda\boldsymbol{X} = \lambda\boldsymbol{R}'\boldsymbol{X}. \tag{7}$$

Произведем замену переменных:

$$\boldsymbol{Y} = \boldsymbol{R}'\boldsymbol{X} \quad \text{или} \quad \boldsymbol{X} = \boldsymbol{R}\boldsymbol{Y}. \tag{8}$$

Подставим (8) в (7) и получим новую задачу о собственных значениях

$$\boldsymbol{D}\boldsymbol{Y} = \lambda\boldsymbol{Y}. \tag{9}$$

Как и раньше, собственные значения такие же, как в (1), так и в (9). Тем не менее, если в выражении (9) выполнить замену переменной (8), то это упростит преобразование $\boldsymbol{X}$ в $\boldsymbol{Y}$ и $\boldsymbol{Y}$ обратно в $\boldsymbol{X}$, так как $\boldsymbol{R}^{-1} = \boldsymbol{R}'$.

Кроме того, предположим, что $\boldsymbol{A}$ — симметричная матрица (т. е. $\boldsymbol{A} = \boldsymbol{A}'$). Тогда получаем

$$\boldsymbol{D}' = (\boldsymbol{R}'\boldsymbol{A}\boldsymbol{R})' = \boldsymbol{R}'\boldsymbol{A}(\boldsymbol{R}')' = \boldsymbol{R}'\boldsymbol{A}\boldsymbol{R} = \boldsymbol{D}. \tag{10}$$

Следовательно, $\boldsymbol{D}$ — симметричная матрица. Приходим к заключению, что, если $\boldsymbol{A}$ — симметричная матрица и $\boldsymbol{R}$ — ортогональная матрица, преобразование матрицы $\boldsymbol{A}$ в $\boldsymbol{D}$, задаваемое выражением (6), сохраняет симметрию так же, как собственные значения. Соотношения между их собственными векторами задаются посредством замены переменной (8).

Последовательность преобразований Якоби

Начнем с действительной симметричной матрицы $\boldsymbol{A}$. Затем построим последовательность ортогональных матриц $\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_n$:

$$\begin{aligned} &\boldsymbol{D}_0 = \boldsymbol{A}, \\ &\boldsymbol{D}_j = \boldsymbol{R}_j' \boldsymbol{D}_{j-1} \boldsymbol{R}_j \quad \text{для } j = 1,\ 2,\ \ldots. \end{aligned} \tag{11}$$

Покажем, как построить последовательность $\{\boldsymbol{R}_j\}$, чтобы были справедливы следующие соотношения:

$$\lim_{j \to \infty} \boldsymbol{D}_j = \boldsymbol{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n). \tag{12}$$

В действительности останавливаемся, когда принадлежащие диагонали элементы близки к нулю. Тогда получаем

$$\boldsymbol{D}_n \approx \boldsymbol{D}. \tag{13}$$

Легко видеть, что

$$\boldsymbol{D}_n = \boldsymbol{R}_n' \boldsymbol{R}_{n-1}' \cdots \boldsymbol{R}_1' \boldsymbol{A} \boldsymbol{R}_1 \boldsymbol{R}_2 \cdots \boldsymbol{R}_{n-1} \boldsymbol{R}_n. \tag{14}$$

Если определить

$$\boldsymbol{R} = \boldsymbol{R}_1 \boldsymbol{R}_2 \cdots \boldsymbol{R}_{n-1} \boldsymbol{R}_n, \tag{15}$$

то $\boldsymbol{R}^{-1} \boldsymbol{A} \boldsymbol{R} = \boldsymbol{D}$, откуда вытекает, что

$$\boldsymbol{A}\boldsymbol{R} = \boldsymbol{R}\boldsymbol{D} = \boldsymbol{R} \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n). \tag{16}$$

Пусть столбцы матрицы $\boldsymbol{R}$ обозначены, как векторы $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$. Тогда матрицу $\boldsymbol{R}$ можно записать как вектор-строку вектор-столбцов:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 & \ldots & \boldsymbol{X}_n \end{bmatrix}. \tag{17}$$

Столбцы произведения (16) принимают следующий вид:

$$\begin{bmatrix} \boldsymbol{A}\boldsymbol{X}_1 & \boldsymbol{A}\boldsymbol{X}_2 & \ldots & \boldsymbol{A}\boldsymbol{X}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 \boldsymbol{X}_1 & \lambda_2 \boldsymbol{X}_2 & \ldots & \lambda_n \boldsymbol{X}_n \end{bmatrix}. \tag{18}$$

Из выражений (17) и (18) видно, что вектор $\boldsymbol{X}_j$, который является j-м столбцом матрицы $\boldsymbol{R}$, — это собственный вектор, соответствующий собственному значению λ_j.

Общий шаг

Каждый шаг итерации Якоби будет определяться стремлением свести два не лежащих на диагонали элемента a_{pq} и a_{qp} к нулю. Пусть $\boldsymbol{R}_1$ — первая из используемых ортогональных матриц. Предположим, что

$$\boldsymbol{D}_1 = \boldsymbol{R}_1' \boldsymbol{A} \boldsymbol{R}_1 \tag{19}$$

сводит элементы a_{pq} и a_{qp} к нулю, где $\boldsymbol{R}_1$ имеет форму

$$\boldsymbol{R}_1 = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{matrix} \\ \\ \leftarrow \text{row } p \\ \\ \leftarrow \text{row } q \\ \\ \\ \end{matrix} \tag{20}$$

$$\begin{matrix} & & \uparrow & & \uparrow & & \\ & & \text{col } p & & \text{col } q & & \end{matrix}$$

Здесь все не стоящие на диагонали элементы матрицы $\boldsymbol{R}_1$ равны нулю, за исключением элемента s, стоящего в p-й строке и q-м столбце, и элемента $-s$, стоящего в q-й строке и p-м столбце. Заметим также, что все диагональные элементы равны 1, за исключением элемента c, который появляется два раза — в p-й строке, p-м столбце и в q-й строке, q-м столбце. Матрица задает вращения плоскости, если используются обозначения $c = \cos\phi$ и $s = \sin\phi$.

Следует убедиться, что преобразование (19) приведет к замене только строк p и q и столбцов p и q. Рассмотрим результат умножения матрицы $\boldsymbol{A}$ на $\boldsymbol{R}_1$ и произведение $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{R}_1$.

$$\boldsymbol{B} = \begin{bmatrix} a_{11} & \cdots & a_{1p} & \cdots & a_{1q} & \cdots & a_{1n} \\ a_{p1} & \cdots & a_{pp} & \cdots & a_{pq} & \cdots & a_{pn} \\ a_{q1} & \cdots & a_{qp} & \cdots & a_{qq} & \cdots & a_{qn} \\ a_{n1} & \cdots & a_{np} & \cdots & a_{nq} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \tag{21}$$

Применим правило умножения строки на столбец и заметим, что не заменяются столбцы 1-й на $(p-1)$-й, $(p+1)$-й на $(q-1)$-й и $(q+1)$-й на n-й. Следовательно, только столбцы p и q заменяются.

$$\begin{aligned} b_{jk} &= a_{jk}, && \text{когда } k \neq p \text{ и } k \neq q, \\ b_{jp} &= ca_{jp} - sa_{jq} && \text{для } j = 1, 2, \ldots, n, \\ b_{jq} &= sa_{jp} + ca_{jq} && \text{для } j = 1, 2, \ldots, n. \end{aligned} \tag{22}$$

Подобные рассуждения показывают, что умножение начиная слева матрицы A на матрицу R_1' приведет к изменению только строк p и q. Поэтому в результате преобразования

$$D_1 = R_1' A R_1 \tag{23}$$

изменятся только столбцы p и q и строки p и q матрицы A. Элементы d_{jk} матрицы D_1 вычисляются по формулам

$$\begin{aligned}
d_{jp} &= ca_{jp} - sa_{jq}, \qquad \text{когда } j \neq p \text{ и } j \neq q,\\
d_{jq} &= sa_{jp} + ca_{jq}, \qquad \text{когда } j \neq p \text{ и } j \neq q,\\
d_{pp} &= c^2 a_{pp} + s^2 a_{qq} - 2csa_{pq},\\
d_{qq} &= s^2 a_{pp} + c^2 a_{qq} + 2csa_{pq},\\
d_{pq} &= (c^2 - s^2)a_{pq} + cs(a_{pp} - a_{qq}).
\end{aligned} \tag{24}$$

Остальные элементы матрицы D_1 находятся из соображений симметрии.

Обнуление d_{pq} и d_{qp}

Назначение каждого шага итерации Якоби — приведение двух не стоящих на диагонали элементов d_{pq} и d_{qp} к нулю. Следует отметить, что очевидная стратегия — это присвоение

$$c = \cos\phi \quad \text{и} \quad s = \sin\phi, \tag{25}$$

где ϕ — угол вращения, который приводит к необходимому результату. Однако сейчас потребуются некоторые остроумные манипуляции тригонометрическими тождествами. Тождество для $\cot\phi$ используется с (25) для определения

$$\theta = \cot 2\phi = \frac{c^2 - s^2}{2cs}. \tag{26}$$

Предположим, что $a_{pq} \neq 0$ и необходимо получить $d_{pq} = 0$. Тогда, используя последнее равенство в (24), получим

$$0 = (c^2 - s^2)a_{pq} + cs(a_{pp} - a_{qq}). \tag{27}$$

Его можно перегруппировать и получить равенство $(c^2 - s^2)/(cs) = (a_{qq} - a_{pp})/a_{pq}$, которое используется в (26) для нахождения решения относительно θ:

$$\theta = \frac{a_{qq} - a_{pp}}{2a_{pq}}. \tag{28}$$

Несмотря на то что для вычисления c и s можно использовать (28) с формулами (25) и (26), менее всего распространится ошибка округления, если вычислять $\tan\phi$ и использовать его в дальнейших вычислениях. Тогда определим

$$t = \tan\phi = \frac{s}{c}. \tag{29}$$

Разделим числитель и знаменатель в (26) на c^2 и получим

$$\theta = \frac{1 - s^2/c^2}{2s/c} = \frac{1 - t^2}{2t},$$

которое приводит к уравнению

$$t^2 + 2t\theta - 1 = 0. \tag{30}$$

Так как $t = \tan\phi$, наименьший корень уравнения (30) соответствует наименьшему углу вращения при $|\phi| \leq \pi/4$. Запишем частную форму квадратичной формулы для нахождения этого корня

$$t = -\theta \pm (\theta^2 + 1)^{1/2} = \frac{\operatorname{sign}(\theta)}{|\theta| + (\theta^2 + 1)^{1/2}}, \tag{31}$$

где $\operatorname{sign}(\theta) = 1$, когда $\theta \geq 0$, и $\operatorname{sign}(\theta) = -1$, когда $\theta < 0$. Затем c и s вычислим по формулам

$$\begin{aligned} c &= \frac{1}{(t^2 + 1)^{1/2}} \\ s &= ct. \end{aligned} \tag{32}$$

Краткое описание общего шага

В этом разделе предлагается схема вычислений для приведения к нулю элемента d_{pq}. Вначале отметим строку p и столбец q, для которых $a_{pq} \neq 0$. Затем предварительно сформируем величины

$$\begin{aligned} \theta &= \frac{a_{qq} - a_{pp}}{2a_{pq}}, \\ t &= \frac{\operatorname{sign}(\theta)}{|\theta| + (\theta^2 + 1)^{1/2}}, \\ c &= \frac{1}{(t^2 + 1)^{1/2}}, \\ s &= ct. \end{aligned} \tag{33}$$

После этого для построения матрицы $\boldsymbol{D} = \boldsymbol{D}_1$ используем

$$
\begin{aligned}
&d_{pq} = 0;\\
&d_{qp} = 0;\\
&d_{pp} = c^2 a_{pp} + s^2 a_{qq} - 2csa_{pq};\\
&d_{qq} = s^2 a_{pp} + c^2 a_{qq} + 2csa_{pq};
\end{aligned}
$$

(34)

```
for   j = 1 : N
   if   (j~ = p)   and   (j~ = q)
      d_jp = c a_jp − s a_jq;
      d_pj = d_jp;
      d_jq = c a_jq + s a_jp;
      d_qj = d_jq;
   end
end
```

Обновление матрицы собственных векторов

Отследим произведение матриц $\boldsymbol{R}_1\boldsymbol{R}_2 \cdots \boldsymbol{R}_n$. Остановившись на n-й итерации, получим

$$\boldsymbol{V}_n = \boldsymbol{R}_1\boldsymbol{R}_2 \cdots \boldsymbol{R}_n, \tag{35}$$

где $\boldsymbol{V}_n$ — ортогональная матрица. Необходимо следить только за текущей матрицей $\boldsymbol{V}_j$, $j = 1, 2, \ldots, n$. Начнем с инициализации матрицы $\boldsymbol{V} = \boldsymbol{I}$. Используем вектор переменных **XP** и **XQ** соответственно для хранения столбцов p и q матрицы $\boldsymbol{A}$. Затем на каждом шаге выполним вычисления

(36)

```
for   j = 1 : N
   XP_j = v_jp;
   XQ_j = v_jq;
end
for   j = 1 : N
   v_jp = c XP_j − s XQ_j;
   v_jq = s XP_j + c XQ_j;
end
```

Стратегия исключения a_{pq}

Скорость сходимости метода Якоби можно определить, если рассмотреть суммы квадратов, не стоящих на диагонали элементов матрицы:

$$S_1 = \sum_{\substack{j,k=1 \\ k \neq j}}^{n} |a_{jk}|^2 \tag{37}$$

$$S_2 = \sum_{\substack{j,k=1 \\ k \neq j}}^{n} |d_{jk}|^2, \quad \text{где} \quad \boldsymbol{D}_1 = \boldsymbol{R}'\boldsymbol{A}\boldsymbol{R}. \tag{38}$$

Читатель может убедиться, что соотношения, заданные в (34), можно использовать для доказательства того, что

$$S_2 = S_1 - 2|a_{pq}|^2. \tag{39}$$

Обозначим на каждом шаге через S_j сумму квадратов, не стоящих на диагонали элементов матрицы $\boldsymbol{D}_j$. Тогда последовательность $\{S_j\}$ монотонно уменьшается и имеет нижнюю грань, равную нулю. Первоначальный алгоритм Якоби, предложенный в 1846 году, на каждом шаге обнулял элемент с наибольшим значением a_{pq}, не стоящим на диагонали, и включал вычисление значений

$$\max\{\boldsymbol{A}\} = \max_{p<q}\{|a_{pq}|\}. \tag{40}$$

Такой выбор гарантировал, что последовательность $\{S_j\}$ сойдется к нулю. Как следствие это доказывает, что $\{\boldsymbol{D}_j\}$ сходится к $\boldsymbol{D}$ и $\{\boldsymbol{V}_j\}$ сходится к матрице собственных векторов $\boldsymbol{V}$ (см. [68]).

На выполнение этого алгоритма Якоби уходит много времени, так как требуется порядка $(n^2 - n)/2$ сравнений за цикл, что делает его непригодным для больших значений n. Лучшей стратегией является циклический метод Якоби, когда он обнуляет элементы, пересекая строки в определенном порядке. Выберем допустимое значение ϵ, затем исследуем всю матрицу и, если найденный элемент a_{pq} будет больше, чем ϵ, то он обнулится. За один проход по матрице проверяются элементы в строке 1, $a_{12}, a_{13}, \ldots, a_{1n}$, затем — в строке 2, $a_{23}, a_{24}, \ldots, a_{2n}$ и т. д. Это доказывает, что скорость сходимости квадратична для первоначального и циклического методов Якоби. Выполнение циклического метода Якоби начнем с замечания, что сумма квадратов диагональных элементов увеличивается с каждой итерацией, т. е. если

$$T_0 = \sum_{j=1}^{n} |a_{jj}|^2 \tag{41}$$

и

$$T_1 = \sum_{j=1}^{n} |d_{jj}|^2,$$

то

$$T_1 = T_0 + 2|a_{pq}|^2.$$

Значит, последовательность $\{\boldsymbol{D}_j\}$ сходится к диагональной матрице $\boldsymbol{D}$. Заметим, что среднее значение диагональных элементов можно вычислить по формуле $(T_0/n)^{1/2}$. Величины не лежащих на диагонали элементов сравнимы с $\epsilon(T_0/n)^{1/2}$, где ϵ — наперед заданное допустимое отклонение. Кроме того, элемент a_{pq} становится равным нулю, если

$$|a_{pq}| > \epsilon \left(\frac{T_0}{n}\right)^{1/2}. \tag{42}$$

Другой вариант метода, называемый пороговым методом Якоби, оставляем для исследования читателю (см. [178]).

Пример 11.7. Используем итерацию Якоби для приведения следующей симметричной матрицы к диагональному виду.

$$\begin{bmatrix} 8 & -1 & 3 & -1 \\ -1 & 6 & 2 & 0 \\ 3 & 2 & 9 & 1 \\ -1 & 0 & 1 & 7 \end{bmatrix}.$$

Детали вычисления оставляем читателю. Первое вращение матрицы для обнуления элемента $a_{13} = 3$ имеет вид

$$\boldsymbol{R}_1 = \begin{bmatrix} 0{,}763020 & 0{,}000000 & 0{,}646375 & 0{,}000000 \\ 0{,}000000 & 0{,}000000 & 0{,}000000 & 0{,}000000 \\ -0{,}646375 & 0{,}000000 & 0{,}763020 & 0{,}000000 \\ 0{,}000000 & 0{,}000000 & 0{,}000000 & 0{,}000000 \end{bmatrix}.$$

Вычисления показывают, что матрица $\boldsymbol{A}_2 = \boldsymbol{R}_1\boldsymbol{A}_1\boldsymbol{R}_1$ равна

$$\boldsymbol{A}_2 = \begin{bmatrix} 5{,}458619 & -2{,}055770 & 0{,}000000 & -1{,}409395 \\ -2{,}055770 & 6{,}000000 & 0{,}879665 & 0{,}000000 \\ 0{,}000000 & 0{,}879665 & 11{,}541381 & 0{,}116645 \\ -1{,}409395 & 0{,}000000 & 0{,}116645 & 7{,}000000 \end{bmatrix}.$$

Затем обнуляем элемент $a_{12} = -2{,}055770$ и получаем

$$\boldsymbol{A}_3 = \begin{bmatrix} 3{,}655795 & 0{,}000000 & 0{,}579997 & -1{,}059649 \\ 0{,}000000 & 7{,}802824 & 0{,}661373 & 0{,}929268 \\ 0{,}579997 & 0{,}661373 & 11{,}541381 & 0{,}116645 \\ -1{,}059649 & 0{,}929268 & 0{,}116645 & 7{,}000000 \end{bmatrix}.$$

После десяти итераций получаем

$$A_{10} = \begin{bmatrix} 3{,}295870 & 0{,}002521 & 0{,}037859 & 0{,}000000 \\ 0{,}002521 & 8{,}405210 & -0{,}004957 & 0{,}066758 \\ 0{,}037859 & -0{,}004957 & 11{,}704123 & -0{,}001430 \\ 0{,}000000 & 0{,}066758 & -0{,}001430 & 6{,}594797 \end{bmatrix}.$$

Чтобы получить близкую к диагональной матрицу

$$D = \operatorname{diag}(3{,}295699; 8{,}407662; 11{,}704301; 6{,}592338).$$

выполняем для диагональных элементов на шесть итераций больше.

Тем не менее элементы, не лежащие на диагонали, недостаточно малы и потребовалось еще три итерации, чтобы они получились меньше по величине, чем 10^{-6}. Тогда собственные векторы — это столбцы матрицы $V = R_1 R_2 \cdots R_{18}$

$$V = \begin{bmatrix} 0{,}528779 & -0{,}573042 & 0{,}582298 & 0{,}230097 \\ 0{,}591967 & 0{,}472301 & 0{,}175776 & -0{,}628975 \\ -0{,}536039 & 0{,}282050 & 0{,}792487 & -0{,}071235 \\ 0{,}287454 & 0{,}607455 & 0{,}044680 & 0{,}739169 \end{bmatrix}. \quad \blacksquare$$

Программа 11.3 (итерация Якоби для собственных значений и векторов). Программа предназначена для вычисления полной совокупности собственных пар $\{\lambda_j, V_j\}_{j=1}^n$ действительной симметричной матрицы A размера $n \times n$. Итерация Якоби используется для нахождения собственных пар.

```
function [V,D]=jacobi1(A,epsilon)
%Вход  - A - матрица размера n x n
%      - epsilon - допустимое значение
%Выход - V - матрица собственных векторов размера nxn
%      - D - диагональная матрица собственных значений размера nxn
%Инициализация V,D и параметров
D=A;
[n,n]=size(A);
V=eye(n);
state=1;
%Вычисление строки p и столбца q наибольшего по абсолютной
%величине элемента матрицы A, не лежащего на диагонали
[m1 p]=max(abs(D-diag(diag(D))));
[m2 q]=max(m1);
```

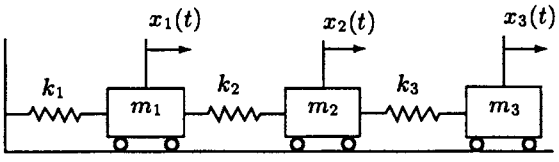


Рис. 11.3. Система недемпфированных масс, связанных пружинами

```
p=p(q);
while(state==1)
   %Обнуление Dpq и Dqp
   t=D(p,q)/(D(q,q)-D(p,p));
   c=1/sqrt(t^2+1);
   s=c*t;
   R=[c s;-s c];
   D([p q],:)=R'*D([p q],:);
   D(:,[p q])=D(:,[p q])*R;
   V(:,[p q])=V(:,[p q])*R;
   [m1 p]=max(abs(D-diag(diag(D))));
   [m2 q]=max(m1);
   p=p(q);
   if (abs(D(p,q))<epsilon*sqrt(sum(diag(D).^2)/n))
      state=0;
   end
end
D=diag(diag(D));
```

Упражнения к разделу 11.3

1. *Системы масс, связанных пружинами.* Рассмотрим недемпфированную систему масс, связанных пружинами, которая изображена на рис. 11.3. Математическая модель, описывающая положение статического равновесия, имеет следующий вид.

$$\begin{bmatrix} k_1+k_2 & -k_2 & 0 \\ -k_2 & k_2+k_3 & -k_3 \\ 0 & -k_3 & k_3 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix} \begin{bmatrix} x_1''(t) \\ x_2''(t) \\ x_3''(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

   (a) Используйте подстановки $x_j(t) = v_j \sin(\omega t + \theta)$ для $j = 1, 2, 3$, где θ — постоянная, и покажите, что решение математической модели можно

переписать в следующем виде:

$$\begin{bmatrix} \frac{k_1+k_2}{m_1} & \frac{-k_2}{m_1} & 0 \\ \frac{-k_2}{m_2} & \frac{k_2+k_3}{m_2} & \frac{-k_3}{m_2} \\ 0 & \frac{-k_3}{m_3} & \frac{k_3}{m_3} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \omega^2 \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

(b) Положим $\lambda = \omega^2$. Тогда три решения из п. (а) являются собственными парами λ_j, $\boldsymbol{V}_j = \begin{bmatrix} v_1^{(j)} & v_2^{(j)} & v_3^{(j)} \end{bmatrix}'$ для $j = 1, 2, 3$. Покажите, что для их построения используются три фундаментальные решения:

$$\boldsymbol{X}_j(t) = \begin{bmatrix} v_1^{(j)} \sin(\omega_j t + \theta) \\ v_2^{(j)} \sin(\omega_j t + \theta) \\ v_3^{(j)} \sin(\omega_j t + \theta) \end{bmatrix} = \sin(\omega_j t + \theta) \begin{bmatrix} v_1^{(j)} \\ v_2^{(j)} \\ v_3^{(j)} \end{bmatrix},$$

где $\omega_j = \sqrt{\lambda_j}$ для $j = 1, 2, 3$.

Замечание. Эти три решения рассматриваются как ***три основных вида колебания***.

2. ***Однородную линейную систему*** дифференциальных уравнений

$$\begin{aligned} x_1'(t) &= x_1(t) + x_2(t), \\ x_2'(t) &= -2x_1(t) + 4x_2(t), \end{aligned}$$

можно записать в матричном виде:

$$\boldsymbol{X}'(t) = \begin{bmatrix} x_1'(t) \\ x_2'(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = {}'\boldsymbol{A}\boldsymbol{X}(t).$$

(a) Проверьте, что 2, $\begin{bmatrix} 1 & 1 \end{bmatrix}'$ и 3, $\begin{bmatrix} 1 & 2 \end{bmatrix}'$ — собственные пары матрицы $\boldsymbol{A}$.

(b) Выполнив прямую подстановку в систему, записанную в матричном виде, убедитесь, что и $\boldsymbol{X}(t) = e^{2t}\begin{bmatrix} 1 & 1 \end{bmatrix}'$, и $\boldsymbol{X}(t) = e^{3t}\begin{bmatrix} 1 & 2 \end{bmatrix}'$ — решения системы дифференциальных уравнений.

(c) Выполнив прямую подстановку в систему, записанную в матричном виде, убедитесь, что $\boldsymbol{X}(t) = c_1 e^{2t}\begin{bmatrix} 1 & 1 \end{bmatrix}' + c_2 e^{3t}\begin{bmatrix} 1 & 2 \end{bmatrix}'$ является общим решением системы дифференциальных уравнений.

Замечание. Если матрица $\boldsymbol{A}$ имеет n различных собственных значений, то она будет иметь n линейно независимых собственных векторов. В этом случае общее решение однородной системы дифференциальных уравнений можно записать как линейную комбинацию: $\boldsymbol{X}(t) = c_1 e^{\lambda_1 t}\boldsymbol{V}_1 + c_2 e^{\lambda_2 t}\boldsymbol{V}_2 + \cdots + c_n e^{\lambda_n t}\boldsymbol{V}_n$.

3. Воспользуйтесь техникой (вручную), схема которой описана в упр. 2, для решения каждой из следующих задач Коши.

(a) $\begin{aligned} x_1' &= 4x_1 + 2x_2 \\ x_2' &= 3x_1 - x_2 \end{aligned}$ с $\begin{cases} x_1(0) = 1 \\ x_2(0) = 2 \end{cases}$

(b) $\begin{aligned} x_1' &= 2x_1 - 12x_2 \\ x_2' &= x_1 - 5x_2 \end{aligned}$ с $\begin{cases} x_1(0) = 2 \\ x_2(0) = 2 \end{cases}$

(c) $\begin{aligned} x_1' &= x_2 \\ x_2' &= x_3 \\ x_3' &= 8x_1 - 14x_2 + 7x_3 \end{aligned}$ с $\begin{cases} x_1(0) = 1 \\ x_2(0) = 2 \\ x_3(0) = 3 \end{cases}$

Алгоритмы и программы

1. Воспользуйтесь программой 11.3 для нахождения собственных пар заданной матрицы с допустимым отклонением $\epsilon = 10^{-7}$. Сравните свои результаты с результатами, которые получены с помощью команды MATLAB `eig` при вводе `[eig(A) diag(D)]` в командном окне MATLAB.

(a) $A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$.

(b) $A = \begin{bmatrix} 2{,}25 & -0{,}25 & -1{,}25 & 2{,}75 \\ -0{,}25 & 2{,}25 & 2{,}75 & 1{,}25 \\ -1{,}25 & 2{,}75 & 2{,}25 & -0{,}25 \\ 2{,}75 & 1{,}25 & -0{,}25 & 2{,}25 \end{bmatrix}$.

(c) $A = [a_{ij}]$, где $a_{ij} = \begin{cases} i + j & i = j \\ ij & i \neq j \end{cases}$ и $i, j = 1, 2, \ldots, 30$.

(d) $A = [a_{ij}]$, где $a_{ij} = \begin{cases} \cos(\sin(i + j)) & i = j \\ i + ij + j & i \neq j \end{cases}$ и $i, j = 1, 2, \ldots, 40$.

2. Воспользуйтесь техникой, схема которой описана в упр. 1, и программой 11.3, чтобы найти собственные пары и три основных вида колебания для недемпированной системы масс, связанных пружинами, со следующими коэффициентами.

(a) $k_1 = 3$; $k_2 = 2$; $k_3 = 1$; $m_1 = 1$; $m_2 = 1$; $m_3 = 1$.

(b) $k_1 = \frac{1}{2}$; $k_2 = \frac{1}{4}$; $k_3 = \frac{1}{4}$; $m_1 = 4$; $m_2 = 4$; $m_3 = 4$.

(c) $k_1 = 0{,}2$; $k_2 = 0{,}4$; $k_3 = 0{,}3$; $m_1 = 2{,}5$; $m_2 = 2{,}5$; $m_3 = 2{,}5$.

3. Воспользуйтесь техникой, схема которой описана в упр. 2, и программой 11.3, чтобы найти общее решение заданной однородной системы дифференциальных уравнений.

(a) $$\begin{aligned} x_1' &= 4x_1 + 3x_2 + 2x_3 + x_4 \\ x_2' &= 3x_1 + 4x_2 + 3x_3 + 2x_4 \\ x_3' &= 2x_1 + 3x_2 + 4x_3 + 3x_4 \\ x_4' &= x_1 + 2x_2 + 3x_3 + 4x_4 \end{aligned}$$

(b) $$\begin{aligned} x_1' &= 5x_1 + 4x_2 + 3x_3 + 2x_4 + x_5 \\ x_2' &= 4x_1 + 5x_2 + 4x_3 + 3x_4 + 2x_5 \\ x_3' &= 3x_1 + 4x_2 + 5x_3 + 4x_4 + 3x_5 \\ x_4' &= 2x_1 + 3x_2 + 4x_3 + 5x_4 + 4x_5 \\ x_5' &= x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 \end{aligned}$$

4. Модифицируйте программу 11.3 таким образом, чтобы реализовать "циклический" метод Якоби.

5. Используйте свою программу из задачи 4 для симметричных матриц из задачи 1. В частности, сравните число итераций, необходимых вашей циклической программе и программе 11.3 при заданном допустимом отклонении.

11.4. Собственные значения симметричных матриц

Метод Хаусхольдера

Каждое преобразование метода Якоби дает два равных нулю элемента, не стоящих на диагонали, но последующие итерации могут сделать их не равными нулю. Значит, потребуется много итераций, чтобы элементы, не лежащие на диагонали, стали достаточно близки к нулю. Сейчас изложим метод, который в каждой итерации дает несколько равных нулю не лежащих на диагонали элементов и они остаются нулями в последующих итерациях. Начнем изложение с важного шага процесса.

Теорема 11.23 (отражение Хаусхольдера). Если $\boldsymbol{X}$ и $\boldsymbol{Y}$ — векторы с одной и той же нормой, то существует такая ортогональная симметричная матрица $\boldsymbol{P}$, что

(1) $$\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{X},$$

где

(2) $$\boldsymbol{P} = \boldsymbol{I} - 2\boldsymbol{W}\boldsymbol{W}'$$

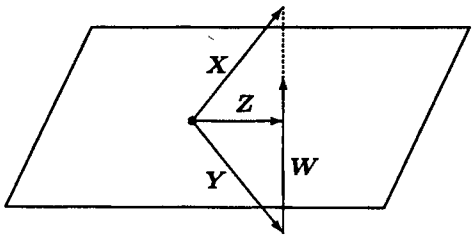


Рис. 11.4. Векторы W, X, Y и Z, используемые при отражении Хаусхольдера

и

$$W = \frac{X - Y}{\|X - Y\|_2}. \tag{3}$$

Из того, что матрица P одновременно и ортогональная, и симметричная, следует

$$P^{-1} = P. \tag{4}$$

Доказательство. Используем соотношение (3) и определим вектор W как единичный вектор в направлении $X - Y$. Следовательно,

$$W'W = 1 \tag{5}$$

и

$$Y = X + cW, \tag{6}$$

где $c = -\|X - Y\|_2$. Так как векторы X и Y имеют одну и ту же норму, можно использовать правило параллелограмма для сложения векторов и увидеть, что вектор $Z = (X + Y)/2 = X + (c/2)W$ ортогонален к вектору W (рис. 11.4). Из этого следует, что

$$W'(X + \frac{c}{2}W) = 0.$$

Используем (5), чтобы преобразовать предыдущее уравнение и получить

$$W'X + \frac{c}{2}W'W = W'X + \frac{c}{2} = 0. \tag{7}$$

Решающим шагом является использование выражения (7) и представление c в виде

$$c = -2(W'X). \tag{8}$$

Тогда (8) можно подставить в (6) и увидеть, что

$$Y = X + cW = X - 2W'XW.$$

Поскольку величина $\boldsymbol{W}'\boldsymbol{X}$ — скаляр, последнее уравнение можно переписать в виде

$$\boldsymbol{Y} = \boldsymbol{X} - 2\boldsymbol{W}\boldsymbol{W}'\boldsymbol{X} = (\boldsymbol{I} - 2\boldsymbol{W}\boldsymbol{W}')\boldsymbol{X}. \tag{9}$$

Если посмотреть на уравнение (9), то можно увидеть, что $\boldsymbol{P} = \boldsymbol{I} - 2\boldsymbol{W}\boldsymbol{W}'$. Матрица $\boldsymbol{P}$ симметрична потому, что

$$\begin{aligned}\boldsymbol{P}' &= (\boldsymbol{I} - 2\boldsymbol{W}\boldsymbol{W}')' = \boldsymbol{I} - 2(\boldsymbol{W}\boldsymbol{W}')' = \\ &= \boldsymbol{I} - 2\boldsymbol{W}\boldsymbol{W}' = \boldsymbol{P}.\end{aligned}$$

Следующие вычисления показывают, что матрица $\boldsymbol{P}$ ортогональна:

$$\begin{aligned}\boldsymbol{P}'\boldsymbol{P} &= (\boldsymbol{I} - 2\boldsymbol{W}\boldsymbol{W}')(\boldsymbol{I} - 2\boldsymbol{W}\boldsymbol{W}') = \\ &= \boldsymbol{I} - 4\boldsymbol{W}\boldsymbol{W}' + 4\boldsymbol{W}\boldsymbol{W}'\boldsymbol{W}\boldsymbol{W}' = \\ &= \boldsymbol{I} - 4\boldsymbol{W}\boldsymbol{W}' + 4\boldsymbol{W}\boldsymbol{W}' = \boldsymbol{I},\end{aligned}$$

и доказательство завершено. •

Отметим, что следствием отображения $\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{X}$ является отражение вектора $\boldsymbol{X}$ относительно линии, имеющей направление $\boldsymbol{Z}$. Отсюда и название — ***отражение Хаусхольдера.***

Следствие 11.3 (k-я матрица Хаусхольдера). Пусть $\boldsymbol{A}$ — матрица размера $n \times$ $\times n$ и $\boldsymbol{X}$ — любой вектор. Если k — целое число, $1 \le k \le n-2$, то можно так построить вектор $\boldsymbol{W}_k$ и матрицу $\boldsymbol{P}_k = \boldsymbol{I} - 2\boldsymbol{W}_k\boldsymbol{W}_k'$, что

$$\boldsymbol{P}_k\boldsymbol{X} = \boldsymbol{P}_k \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ x_{k+2} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ -S \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{Y}. \tag{10}$$

Доказательство. Ключом доказательства является определение значения S таким образом, что $\|\boldsymbol{X}\|_2 = \|\boldsymbol{Y}\|_2$, и последующее применение теоремы 11.23. Подходящее значение для S должно удовлетворять равенству

$$S^2 = x_{k+1}^2 + x_{k+2}^2 + \cdots + x_n^2, \tag{11}$$

которое без труда проверяется путем вычисления норм векторов $\boldsymbol{X}$ и $\boldsymbol{Y}$:

$$\begin{aligned}\|\boldsymbol{X}\|_2 &= x_1^2 + x_2^2 + \cdots + x_n^2 = \\ &= x_1^2 + x_2^2 + \cdots + x_k^2 + S^2 = \\ &= \|\boldsymbol{Y}\|_2 .\end{aligned} \tag{12}$$

Вектор $\boldsymbol{W}$ находим, используя выражение (3) теоремы 11.23:

$$\begin{aligned} \boldsymbol{W} &= \frac{1}{R}(\boldsymbol{X}-\boldsymbol{Y}) = \\ &= \frac{1}{R}[0 \ \dots \ 0 \ \ (x_{k+1}+S) \ \ x_{k+2} \ \dots \ x_n]'. \end{aligned} \tag{13}$$

В наименьшей степени ошибка округления распространяется тогда, когда знак S выбираем таким же, как и знак x_{k+1}. Следовательно, вычисляем

$$S = \operatorname{sign}(x_{k+1})(x_{k+1}^2 + x_{k+2}^2 + \dots + x_n^2)^{1/2}. \tag{14}$$

Число R в (13) выбрано таким образом, что $\|\boldsymbol{W}\|_2 = 1$, и должно удовлетворять равенству

$$\begin{aligned} R^2 &= (x_{k+1}+S)^2 + x_{k+2}^2 + \dots + x_n^2 = \\ &= 2x_{k+1}S + S^2 + x_{k+1}^2 + x_{k+2}^2 + \dots + x_n^2 = \\ &= 2x_{k+1}S + 2S^2. \end{aligned} \tag{15}$$

Поэтому матрица $\boldsymbol{P}_k$ задается формулой

$$\boldsymbol{P}_k = \boldsymbol{I} - 2\boldsymbol{W}\boldsymbol{W}'. \tag{16}$$

На этом доказательство завершено. •

Преобразование Хаусхольдера

Предположим, что $\boldsymbol{A}$ — симметричная матрица размера $n \times n$. Тогда последовательность $n-2$ преобразований вида $\boldsymbol{PAP}$ приводит $\boldsymbol{A}$ к симметричной трехдиагональной матрице. Представим себе процесс, когда $n = 5$. Первое преобразование определяет $\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{P}_1$, где $\boldsymbol{P}_1$ построено согласно следствию 11.3 с вектором $\boldsymbol{X}$, который является первым столбцом матрицы $\boldsymbol{A}$. Общий вид матрицы $\boldsymbol{P}_1$ таков:

$$\boldsymbol{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & p & p & p & p \\ 0 & p & p & p & p \\ 0 & p & p & p & p \\ 0 & p & p & p & p \end{bmatrix}, \tag{17}$$

где буква p установлена вместо некоторых элементов матрицы $\boldsymbol{P}_1$. В результате преобразование $\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{P}_1$ не затрагивает элемент a_{11} матрицы $\boldsymbol{A}$:

$$\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{P}_1 = \begin{bmatrix} a_{11} & v_1 & 0 & 0 & 0 \\ u_1 & w_1 & w & w & w \\ 0 & w & w & w & w \\ 0 & w & w & w & w \\ 0 & w & w & w & w \end{bmatrix} = \boldsymbol{A}_1. \tag{18}$$

Элемент, обозначенный через u_1, изменяется, так как выполняется умножение слева на матрицу $\boldsymbol{P}_1$, а v_1 изменяется, так как выполняется умножение справа на $\boldsymbol{P}_1$. Следовательно, матрица $\boldsymbol{A}_1$ симметрична и имеем $u_1 = v_1$. Изменение элементов, обозначенных через w, является следствием как умножения слева, так и умножения справа. Также, поскольку вектор $\boldsymbol{X}$ является первым столбцом матрицы $\boldsymbol{A}$, из (10) следует, что $u_1 = -S$.

Второе преобразование Хаусхольдера, примененное к матрице $\boldsymbol{A}_1$, определенной в (18), обозначается через $\boldsymbol{P}_2\boldsymbol{A}\boldsymbol{P}_2$, где матрица $\boldsymbol{P}_2$ строится согласно следствию 11.3 с вектором $\boldsymbol{X}$, который, в свою очередь, является вторым столбцом матрицы $\boldsymbol{A}_1$. Матрица $\boldsymbol{P}_2$ имеет вид

$$\boldsymbol{P}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & p & p & p \\ 0 & 0 & p & p & p \\ 0 & 0 & p & p & p \end{bmatrix}, \tag{19}$$

где p установлено вместо некоторых элементов в матрице $\boldsymbol{P}_2$. Благодаря единичному блоку размера 2×2 в левом верхнем углу частичная трехдиагональность, достигаемая на первом шаге, не будет изменяться вторым преобразованием $\boldsymbol{P}_2\boldsymbol{A}_1\boldsymbol{P}_2$. В результате этого преобразования получим

$$\boldsymbol{P}_2\boldsymbol{A}_1\boldsymbol{P}_2 = \begin{bmatrix} a_{11} & v_1 & 0 & 0 & 0 \\ u_1 & w_1 & v_2 & 0 & 0 \\ 0 & u_2 & w_2 & w & w \\ 0 & 0 & w & w & w \\ 0 & 0 & w & w & w \end{bmatrix} = \boldsymbol{A}_2. \tag{20}$$

Элементы u_2 и v_2 изменяются в результате умножения слева и умножения справа на $\boldsymbol{P}_2$. Дополнительные изменения при преобразовании относятся к другим элементам w.

Третье преобразование Хаусхольдера, $\boldsymbol{P}_3\boldsymbol{A}_2\boldsymbol{P}_3$, применяется к матрице $\boldsymbol{A}_2$, определенной в (20), где следствие используется вектором $\boldsymbol{X}$, который, в свою очередь, является третьим столбцом матрицы $\boldsymbol{A}_2$. Матрица $\boldsymbol{P}_3$ имеет вид

$$\boldsymbol{P}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & p & p \\ 0 & 0 & 0 & p & p \end{bmatrix}. \tag{21}$$

И снова благодаря единичному блоку размера 3×3 матрица $\boldsymbol{P}_3\boldsymbol{A}_2\boldsymbol{P}_3$ не изменяет элементы матрицы $\boldsymbol{A}_2$, которые лежат в верхнем углу размера 3×3.

Получаем

$$\boldsymbol{P}_3\boldsymbol{A}_2\boldsymbol{P}_3 = \begin{bmatrix} a_{11} & v_1 & 0 & 0 & 0 \\ u_1 & w_1 & v_2 & 0 & 0 \\ 0 & u_2 & w_2 & v_3 & 0 \\ 0 & 0 & u_3 & w & w \\ 0 & 0 & 0 & w & w \end{bmatrix} = \boldsymbol{A}_3. \tag{22}$$

Таким образом, понадобилось три преобразования, чтобы привести матрицу $\boldsymbol{A}$ к трехдиагональному виду.

В действительности преобразование $\boldsymbol{PAP}$ не выполняется в матричном виде. Следующий результат показывает, что это эффективнее выполнять некоторыми искусными манипуляциями векторами.

Теорема 11.24 (вычисление одного преобразования Хаусхольдера). Если $\boldsymbol{P}$ – матрица Хаусхольдера, то преобразование $\boldsymbol{PAP}$ выполняется следующим образом. Пусть

$$\boldsymbol{V} = \boldsymbol{AW}. \tag{23}$$

Вычислим

$$c = \boldsymbol{W}'\boldsymbol{V} \tag{24}$$

и

$$\boldsymbol{Q} = \boldsymbol{V} - c\boldsymbol{W}. \tag{25}$$

Тогда

$$\boldsymbol{PAP} = \boldsymbol{A} - 2\boldsymbol{WQ}' - 2\boldsymbol{QW}'. \tag{26}$$

Доказательство. Сначала образуем произведение

$$\boldsymbol{AP} = \boldsymbol{A}(\boldsymbol{I} - 2\boldsymbol{WW}') = \boldsymbol{A} - 2\boldsymbol{AWW}'.$$

Используя (23), запишем его в виде

$$\boldsymbol{AP} = \boldsymbol{A} - 2\boldsymbol{VW}'. \tag{27}$$

Теперь используем (27) и запишем

$$\boldsymbol{PAP} = (\boldsymbol{I} - 2\boldsymbol{WW}')(\boldsymbol{A} - 2\boldsymbol{VW}'). \tag{28}$$

Если преобразовать эту величину, член $2(2\boldsymbol{WW}'\boldsymbol{VW}')$ можно разделить на две части и затем (28) переписать в виде

$$\boldsymbol{PAP} = \boldsymbol{A} - 2\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{A}) + 2\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{VW}') - 2\boldsymbol{VW}' + 2\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{V})\boldsymbol{W}'. \tag{29}$$

Если предположить, что матрица $\boldsymbol{A}$ симметрична, можно использовать тождество $(\boldsymbol{W}'\boldsymbol{A}) = (\boldsymbol{W}'\boldsymbol{A}') = \boldsymbol{V}'$. Частично фокус состоит в том, чтобы вспомнить, что $(\boldsymbol{W}'\boldsymbol{V})$ — скалярная величина. Следовательно, она может коммутировать с любым членом. Другое скалярное тождество $\boldsymbol{W}'\boldsymbol{V} = (\boldsymbol{W}'\boldsymbol{V})'$ используется для получения соотношения $\boldsymbol{W}'\boldsymbol{V}\boldsymbol{W}' = (\boldsymbol{W}'\boldsymbol{V})\boldsymbol{W}' = \boldsymbol{W}'(\boldsymbol{W}'\boldsymbol{V}) = \boldsymbol{W}'(\boldsymbol{W}'\boldsymbol{V})' = (((\boldsymbol{W}'\boldsymbol{V})\boldsymbol{W})' = (\boldsymbol{W}'\boldsymbol{V}\boldsymbol{W})'$. Эти результаты применяются в стоящих в скобках членах выражения (29) для получения

$$\boldsymbol{P}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{A} - 2\boldsymbol{W}\boldsymbol{V}' + 2\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{V}\boldsymbol{W})' - 2\boldsymbol{V}\boldsymbol{W}' + 2\boldsymbol{W}'\boldsymbol{V}\boldsymbol{W}\boldsymbol{W}'. \tag{30}$$

Воспользуемся в (30) законом дистрибутивности и получим

$$\boldsymbol{P}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{A} - 2\boldsymbol{W}(\boldsymbol{V}' - (\boldsymbol{W}'\boldsymbol{V}\boldsymbol{W})') - 2(\boldsymbol{V} - \boldsymbol{W}'\boldsymbol{V}\boldsymbol{W})\boldsymbol{W}'. \tag{31}$$

Наконец используем в (31) определение $\boldsymbol{Q}$, данное в (25), и в результате получим выражение (26). На этом доказательство закончено. •

Приведение к трехдиагональному виду

Предположим, что $\boldsymbol{A}$ — симметричная матрица размера $n \times n$. Начнем с

$$\boldsymbol{A}_0 = \boldsymbol{A}. \tag{32}$$

Построим последовательность матриц Хаусхольдера $\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_{n-1}$ таким образом, что

$$\boldsymbol{A}_k = \boldsymbol{P}_k\boldsymbol{A}_{k-1}\boldsymbol{P}_k \quad \text{для } k = 1, 2, \ldots, n-2, \tag{33}$$

где матрица $\boldsymbol{A}_k$ имеет нули под диагональю в столбцах $1, 2, \ldots, k$. Тогда $\boldsymbol{A}_{n-2}$ является симметричной трехдиагональной матрицей, подобной матрице $\boldsymbol{A}$. Этот метод называется ***методом Хаусхольдера***.

Пример 11.8. Используем метод Хаусхольдера, чтобы привести следующую матрицу к симметричному трехдиагональному виду:

$$\boldsymbol{A}_0 = \begin{bmatrix} 4 & 2 & 2 & 1 \\ 2 & -3 & 1 & 1 \\ 2 & 1 & 3 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}.$$

Подробности оставляем читателю. Используем для построения вектора постоянные $S = 3$ и $R = 30^{1/2} = 5{,}477226$:

$$\boldsymbol{W}' = \frac{1}{\sqrt{30}}[0 \;\; 5 \;\; 2 \;\; 1] = [0{,}000000 \;\; 0{,}912871 \;\; 0{,}365148 \;\; 0{,}182574].$$

Затем используем умножение матриц $\boldsymbol{V} = \boldsymbol{AW}$, чтобы сформировать

$$\boldsymbol{V}' = \frac{1}{\sqrt{30}}[0 \ \ -12 \ \ 12 \ \ 9] =$$
$$= [0{,}000000 \ \ -2{,}190890 \ \ 2{,}190890 \ \ 1{,}643168].$$

Находим, что постоянная $c = \boldsymbol{W}'\boldsymbol{V}$ равна

$$c = -0{,}9.$$

После этого формируем вектор $\boldsymbol{Q} = \boldsymbol{V} - c\boldsymbol{W} = \boldsymbol{V} + 0{,}9\boldsymbol{W}$:

$$\boldsymbol{Q}' = \frac{1}{\sqrt{30}}[0{,}000000 \ \ -7{,}500000 \ \ 13{,}800000 \ \ 9{,}900000] =$$
$$= [0{,}000000 \ \ -1{,}369306 \ \ 2{,}519524 \ \ 1{,}807484].$$

Вычисляя $\boldsymbol{A}_1 = \boldsymbol{A}_0 - 2\boldsymbol{WQ}' - 2\boldsymbol{QW}'$, получаем

$$\boldsymbol{A}_1 = \begin{bmatrix} 4{,}0 & -3{,}0 & 0{,}0 & 0{,}0 \\ -3{,}0 & 2{,}0 & -2{,}6 & -1{,}8 \\ 0{,}0 & -2{,}6 & -0{,}68 & -1{,}24 \\ 0{,}0 & -1{,}8 & -1{,}24 & 0{,}68 \end{bmatrix}.$$

И на последнем шаге используем постоянные $S = -3{,}1622777$, $R = 6{,}0368737$, $c = -1{,}2649111$ и векторы

$$\boldsymbol{W}' = [0{,}000000 \ \ 0{,}000000 \ \ -0{,}954514 \ \ -0{,}298168],$$
$$\boldsymbol{V}' = [0{,}000000 \ \ 0{,}000000 \ \ 1{,}018797 \ \ 0{,}980843],$$
$$\boldsymbol{Q}' = [0{,}000000 \ \ 0{,}000000 \ \ -0{,}188578 \ \ 0{,}603687].$$

Тогда трехдиагональная матрица $\boldsymbol{A}_2 = \boldsymbol{A}_1 - 2\boldsymbol{WQ}' - 2\boldsymbol{QW}'$ имеет вид

$$\boldsymbol{A}_2 = \begin{bmatrix} 4{,}0 & -3{,}0 & 0{,}0 & 0{,}0 \\ -3{,}0 & 2{,}0 & 3{,}162278 & 0{,}0 \\ 0{,}0 & 3{,}162278 & -1{,}4 & -0{,}2 \\ 0{,}0 & 0{,}0 & -0{,}2 & 1{,}4 \end{bmatrix}. \quad \blacksquare$$

Программа 11.4 (приведение к трехдиагональному виду). Программа предназначена для приведения симметричной матрицы $\boldsymbol{A}$ размера $n \times n$ к трехдиагональному виду с использованием $n - 2$ преобразований Хаусхольдера.

```
function T=house (A)
%Вход  - A - симметричная матрица размера n x n
%Выход - T - трехдиагональная матрица
```

```
[n,n]=size(A);
for k=1:n-2
   %Построение W
   s=norm(A(k+1:n,k));
   if (A(k+1,k)<0)
      s=-s;
   end
   r=sqrt(2*s*(A(k+1,k)+s));
   W(1:k)=zeros(1,k);
   W(k+1)=(A(k+1,k)+s)/r;
   W(k+2:n)=A(k+2:n,k)'/r;
   %Построение V
   V(1:k)=zeros(1,k);
   V(k+1:n)=A(k+1:n,k+1:n)*W(k+1:n)';
   %Построение Q
   c=W(k+1:n)*V(k+1:n)';
   Q(1:k)=zeros(1,k);
   Q(k+1:n)=V(k+1:n)-c*W(k+1:n);
   %Построение Ak
   A(k+2:n,k)=zeros(n-k-1,1);
   A(k,k+2:n)=zeros(1,n-k-1);
   A(k+1,k)=-s;
   A(k,k+1)=-s;
   A(k+1:n,k+1:n)=A(k+1:n,k+1:n) ...
   -2*W(k+1:n)'*Q(k+1:n)-2*Q(k+1:n)'*W(k+1:n);
end
T=A;
```

QR-метод

Предположим, что A — действительная симметричная матрица. В предыдущем разделе было показано, метод Хаусхольдера используется для построения подобной трехдиагональной матрицы. QR-метод применяется для нахождения всех собственных значений трехдиагональной матрицы. Подобное этому вращение плоскости введено в методе Якоби, чтобы построить ортогональную матрицу $Q_1 = Q$ и верхнюю треугольную матрицу $U_1 = U$ таким образом, чтобы $A_1 = A$ можно было разложить на множители

$$A_1 = Q_1 U_1. \tag{34}$$

Затем образуем произведение

$$A_2 = U_1 Q_1. \tag{35}$$

Так как Q_1 — ортогональная матрица, можно использовать (34) и увидеть, что

$$Q_1' A_1 = Q_1' Q_1 U_1 = U_1. \tag{36}$$

Следовательно, A_2 можно вычислить по формуле

$$A_2 = Q_1' A_1 Q_1. \tag{37}$$

А поскольку $Q_1' = Q_1^{-1}$, A_2 подобна матрице A_1 и имеет такие же собственные значения. В общем, строим ортогональную матрицу Q_k и верхнюю треугольную матрицу U_k таким образом, что

$$A_k = Q_k U_k. \tag{38}$$

Затем определим

$$A_k + 1 = U_k Q_k = Q_k' A_k Q_k. \tag{39}$$

И снова получаем $Q_k' = Q_k^{-1}$, из чего следует, что матрицы A_{k+1} и A_k подобны. Важным следствием является то, что матрица A_k подобна A и, значит, имеет такую же структуру. Определенно, можно заключить, что если A — трехдиагональная матрица, то A_k — также трехдиагональная матрица для всех k. Теперь предположим, что матрица A записана в виде

$$A = \begin{bmatrix} d_1 & e_1 & & & & & \\ e_1 & d_2 & e_2 & & & & \\ & e_2 & d_3 & \cdots & & & \\ & & \vdots & d_{n-2} & e_{n-2} & & \\ & & & e_{n-2} & d_{n-1} & e_{n-1} & \\ & & & & e_{n-1} & d_n & \end{bmatrix}. \tag{40}$$

Можно найти вращение плоскости P_{n-1}, которое сводит к нулю элемент матрицы A, занимающий место $(n, n-1)$, т. е.

$$P_{n-1}A = \begin{bmatrix} d_1 & e_1 & & & & & \\ e_1 & d_2 & e_2 & & & & \\ & e_2 & d_3 & \cdots & & & \\ & & \vdots & d_{n-2} & q_{n-2} & r_{n-2} & \\ & & & e_{n-2} & p_{n-1} & q_{n-1} & \\ & & & & 0 & p_n & \end{bmatrix}. \tag{41}$$

Аналогично можно построить вращение плоскости P_{n-2}, которое уменьшит до нуля элемент матрицы $P_{n-1}A$, расположенный на месте $(n-1, n-2)$. После

$n-1$ шага получим

$$\boldsymbol{P}_1 \cdots \boldsymbol{P}_{n-1}\boldsymbol{A} = \begin{bmatrix} p_1 & q_1 & r_1 & \cdots & & & \\ 0 & p_2 & q_2 & \ddots & & & \\ 0 & 0 & p_3 & \ddots & r_{n-4} & & \\ & & \vdots & \ddots & q_{n-3} & r_{n-3} & \\ & & & & p_{n-2} & q_{n-2} & r_{n-2} \\ & & & & 0 & p_{n-1} & q_{n-1} \\ & & & & 0 & 0 & p_n \end{bmatrix} = \boldsymbol{U}. \tag{42}$$

Поскольку каждое вращение плоскости описывается ортогональной матрицей, из уравнения (42) следует, что

$$\boldsymbol{Q} = \boldsymbol{P}'_{n-1}\boldsymbol{P}'_{n-2}\cdots\boldsymbol{P}'_1. \tag{43}$$

Умножив матрицу $\boldsymbol{U}$ на матрицу $\boldsymbol{Q}$, получим, что все элементы, лежащие ниже второй нижней диагонали, равны нулю. Из трехдиагонального вида матрицы $\boldsymbol{A}_2$ следует, что она также имеет нули выше второй верхней диагонали. Исследования показывают, что члены r_j используются только для вычисления этих нулевых элементов. Значит, нет необходимости хранить или использовать в компьютере числа $\{r_j\}$.

Для каждого вращения плоскости $\boldsymbol{P}_j$ предполагается, что сохраняются коэффициенты c_j и s_j, которые его определяют. Тогда нет необходимости вычислять и хранить именно матрицу $\boldsymbol{Q}$; можно использовать последовательности $\{c_j\}$ и $\{s_j\}$ вместе с точными формулами, чтобы получить произведение

$$\boldsymbol{A}_2 = \boldsymbol{U}\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{P}'_{n-1}\boldsymbol{P}'_{n-2}\cdots\boldsymbol{P}'_1. \tag{44}$$

Ускоренные сдвиги

Выше приводилась схема работы QR метода, однако сходимость его медленна даже для матриц малого размера. Можно дополнить его техникой сдвига, что повысит скорость сходимости. Напомним, что если λ_j — собственное значение матрицы $\boldsymbol{A}$, то $\lambda_j - s_i$ — собственное значение матрицы $B = \boldsymbol{A} - s_i\boldsymbol{I}$. Этот метод предусматривает выполнение модификации

$$\boldsymbol{A}_i - s_i\boldsymbol{I} = \boldsymbol{U}_i\boldsymbol{L}_i \tag{45}$$

и формирования матрицы

$$\boldsymbol{A}_{i+1} = \boldsymbol{U}_i\boldsymbol{Q}_i \quad \text{для } i = 1,\ 2,\ \ldots,\ k_j, \tag{46}$$

где $\{s_i\}$ — последовательность, сумма которой равна λ_j, т. е. $\lambda_j = s_1 + s_2 + \cdots + s_{k_j}$.

На каждом этапе правильное число сдвигов находят, используя четыре элемента в нижнем правом углу матрицы. Начнем с определения λ_1 и вычисления собственных значений матрицы размера 2×2:

$$\begin{bmatrix} d_{n-1} & e_{n-1} \\ e_{n-1} & d_n \end{bmatrix}. \tag{47}$$

Они равны x_1 и x_2 и являются корнями квадратного уравнения

$$x^2 - (d_{n-1} + d_n)x + d_{n-1}d_n - e_{n-1}e_{n-1} = 0. \tag{48}$$

Значение s_i в соотношении (45) выбирается таким образом, чтобы оно было корнем уравнения (48), т. е. ближайшим к d_n.

Затем итерация со сдвигом QR метода повторяется до тех пор, пока не будет получено $e_{n-1} \approx 0$. Это дает первое собственное значение $\lambda_1 = s_1 + s_2 + \cdots + s_{k_1}$. Подобный процесс повторяется с верхними $n-1$ рядами, чтобы получить $e_{n-2} \approx 0$, и следующее собственное значение равно λ_2. Последовательно итерациии применяются к меньшим подматрицам, пока не будет получено $e_2 \approx 0$ и собственное значение λ_{n-2}. Наконец, квадратичная формула используется, чтобы найти два последних собственных значения. Детали можно уточнить, если проанализировать программу.

Пример 11.9. Найдем собственные значения матрицы

$$M = \begin{bmatrix} 4 & 2 & 2 & 1 \\ 2 & -3 & 1 & 1 \\ 2 & 1 & 3 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}.$$

В примере 11.8 трехдиагональная матрица A_1 была построена так, что она подобна матрице M. Начнем процесс приведения к диагональному виду с этих матриц:

$$A_1 = \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & 2 & 3{,}16228 & 0 \\ 0 & 3{,}16228 & -1{,}4 & -0{,}2 \\ 0 & 0 & -0{,}2 & 1{,}4 \end{bmatrix}.$$

Четыре элемента в нижнем правом углу равны $d_3 = -1{,}4$, $d_4 = 1{,}4$ и $e_3 = -0{,}2$, используем их для формирования квадратного уравнения

$$x^2 - (-1{,}4 + 1{,}4)x + (-1{,}4)(1{,}4) - (-0{,}2)(-0{,}2) = x^2 - 2 = 0.$$

Вычисляем корни $x_1 = -1{,}41421$ и $x_2 = 1{,}41421$. Корень, ближайший к d_4, выбираем в качестве первого сдвига $s_1 = 1{,}41421$, и после первого сдвига матрица

имеет вид

$$\boldsymbol{A}_1 - s_1\boldsymbol{I} = \begin{bmatrix} 2{,}58579 & -3 & 0 & 0 \\ -3 & 0{,}58579 & 1{,}74806 & 0 \\ 0 & 1{,}74806 & -2{,}81421 & -1{,}61421 \\ 0 & 0 & -1{,}61421 & -0{,}01421 \end{bmatrix}.$$

Затем вычисляем разложение на множители $\boldsymbol{A}_1 - s_1\boldsymbol{I} = \boldsymbol{Q}_1\boldsymbol{U}_1$:

$$\boldsymbol{Q}_1\boldsymbol{U}_1 = \begin{bmatrix} -0{,}65288 & -0{,}38859 & -0{,}55535 & 0{,}33814 \\ 0{,}75746 & -0{,}33494 & -0{,}47867 & 0{,}29145 \\ 0 & 0{,}85838 & -0{,}43818 & 0{,}26610 \\ 0 & 0 & 0{,}52006 & 0{,}85413 \end{bmatrix} \times$$

$$\times \begin{bmatrix} -3{,}96059 & 2{,}40235 & 2{,}39531 & 0 \\ 0 & 3{,}68400 & -3{,}47483 & -0{,}17168 \\ 0 & 0 & -0{,}38457 & 0{,}08024 \\ 0 & 0 & 0 & -0{,}06550 \end{bmatrix}.$$

Теперь вычисляем в обратном порядке произведение матриц и получаем

$$\boldsymbol{A}_2 = \boldsymbol{U}_1\boldsymbol{Q}_1 = \begin{bmatrix} 4{,}40547 & 2{,}79049 & 0 & 0 \\ 2{,}79049 & -4{,}21663 & -0{,}33011 & 0 \\ 0 & -0{,}33011 & 0{,}21024 & -0{,}03406 \\ 0 & 0 & -0{,}03406 & -0{,}05595 \end{bmatrix}.$$

Второй сдвиг равен $s_2 = -0{,}06024$. После второго сдвига матрица имеет вид $\boldsymbol{A}_2 - s_2\boldsymbol{I} = \boldsymbol{Q}_2\boldsymbol{U}_2$ и

$$\boldsymbol{A}_3 = \boldsymbol{U}_2\boldsymbol{Q}_2 = \begin{bmatrix} 4{,}55257 & -2{,}65725 & 0 & 0 \\ -2{,}65725 & -4{,}26047 & 0{,}01911 & 0 \\ 0 & 0{,}01911 & 0{,}29171 & 0{,}00003 \\ 0 & 0 & 0{,}00003 & 0{,}00027 \end{bmatrix}.$$

Третий сдвиг равен $s_3 = 0{,}00027$. Матрица после третьего сдвига имеет вид $\boldsymbol{A}_3 - s_3\boldsymbol{I} = \boldsymbol{Q}_3\boldsymbol{U}_3$ и

$$\boldsymbol{A}_4 = \boldsymbol{U}_3\boldsymbol{Q}_3 = \begin{bmatrix} 4{,}62640 & 2{,}53033 & 0 & 0 \\ 2{,}53033 & -4{,}33489 & -0{,}00111 & 0 \\ 0 & -0{,}00111 & 0{,}29150 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Первое собственное значение, округленное до пяти десятичных знаков, получаем в результате вычислений:

$$\lambda_1 = s_1 + s_2 + s_3 = 1{,}41421 - 0{,}06023 + 0{,}00027 = 1{,}35425.$$

Следующее λ_1 расположено на последней позиции диагонали матрицы A_4, и процесс повторяется. Однако изменения происходят только в верхнем углу размера 3×3 матрицы

$$A_4 = \begin{bmatrix} 4{,}62640 & 2{,}53033 & 0 & 0 \\ 2{,}53033 & -4{,}33489 & -0{,}00111 & 0 \\ 0 & -0{,}00111 & 0{,}29150 & 0 \\ 0 & 0 & 0 & 1{,}35425 \end{bmatrix}.$$

Аналогично один дополнительный сдвиг приводит элемент во второй строке и третьем столбце к нулю (определяется до десяти десятичных знаков):

$$s_4 = 0{,}29150, \qquad A_4 - s_4 I = Q_4 U_4, \qquad A_5 = U_4 Q_4.$$

Поэтому второе собственное значение равно

$$\lambda_2 = \lambda_1 + s_4 = 1{,}35425 + 0{,}29150 = 1{,}64575.$$

Наконец λ_2 располагаем на диагонали матрицы A_5 в третьей строке и столбце и получаем

$$A_5 = \begin{bmatrix} 4{,}26081 & -2{,}65724 & 0 & 0 \\ -2{,}65724 & -4{,}55232 & 0 & 0 \\ 0 & 0 & 1{,}64575 & 0 \\ 0 & 0 & 0 & 1{,}35425 \end{bmatrix}.$$

Для завершения вычисления необходимо найти собственные значения матрицы размера 2×2 в верхнем левом углу матрицы A_5. Характеристическое уравнение

$$x^2 - (-4{,}26081 + 4{,}55232)x + (4{,}26081)(-4{,}55232) - (2{,}65724)(2{,}65724) = 0$$

приводим к виду

$$x^2 + 0{,}29151x - 26{,}45749 = 0.$$

Корни его равны $x_1 = 5{,}00000$ и $x_2 = -5{,}29150$, и два последних собственных значения получаем в результате вычислений

$$\lambda_3 = \lambda_2 + x_1 = 1{,}64575 + 5{,}0000 = 6{,}64575$$

и

$$\lambda_4 = \lambda_2 + x_2 = 1{,}64575 - 5{,}29150 = -3{,}64575. \quad \blacksquare$$

Программу 11.5 можно использовать для получения всех приближенных собственных значений симметричной трехдиагональной матрицы. Алгоритм программы следует из предыдущих рассуждений, но с двумя значительными исключениями. Первое: команда MATLAB `eig` используется для нахождения корней характеристического уравнения (48) каждой подматрицы (47) размера 2×2. Второе:

QR-разложение матрицы $A_i - s_i I$ (45) выполняется с использованием команды MATLAB `[Q,R]=qr(B)`, которая получает такие ортогональную матрицу `Q` и верхнюю треугольную матрицу `R`, что `B=Q*R` (читатели могут написать собственную программу QR-разложения).

Программа 11.5 (QR-метод со сдвигами). Программа предназначена для приближенного вычисления собственных значений симметричной трехдиагональной матрицы A с использованием QR-метода со сдвигами.

```
function D=qr2(A,epsilon)
%Вход  - A - симметричная трехдиагональная матрица размера n x n
%      - epsilon - допустимое отклонение
%Выход - D - вектор собственных значений размера n x 1
%Инициализация параметров
[n,n]=size(A);
m=n;
D=zeros(n,1);
B=A;
while (m>1)
   while (abs(B(m,m-1))>=epsilon)
      %Вычисление сдвига
      S=eig(B(m-1:m,m-1:m));
      [j,k]=min([abs(B(m,m)*[1 1]'-S)]);
      %QR разложение B
      [Q,U]=qr(B-S(k)*eye(m));
      %Вычисление следующего B
      B=U*Q+S(k)*eye(m);
   end
   %Расположение m-го собственного значения в матрице A(m,m)
   A(1:m,1:m)=B;
   %Повторение процесса на подматрице A размера m-1 x m-1
   m=m-1;
   B=A(1:m,1:m);
end
D=diag(A);
```

Упражнения к разделу 11.4

1. Подробно объясните, почему в доказательстве теоремы 11.23 Z перпендикулярно W.

2. Если $\boldsymbol{X}$ — произвольный вектор, и $\boldsymbol{P} = \boldsymbol{I} - 2\boldsymbol{X}\boldsymbol{X}'$, то покажите, что $\boldsymbol{P}$ — симметричная матрица.
3. Пусть $\boldsymbol{X}$ — произвольный вектор и пусть $\boldsymbol{P} = \boldsymbol{I} - 2\boldsymbol{X}\boldsymbol{X}'$.
   (a) Найдите величину $\boldsymbol{P}'\boldsymbol{P}$.
   (b) Какое дополнительное условие необходимо для того, чтобы $\boldsymbol{P}$ была ортогональной матрицей?

Алгоритмы и программы

В задачах 1–6 используйте следующее.
(a) Программу 11.4, чтобы привести данную матрицу к трехдиагональному виду.
(b) Программу 11.5, чтобы найти собственные значения данной матрицы.

1. $\begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix}$.

2. $\begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$.

3. $\begin{bmatrix} 2{,}75 & -0{,}25 & -0{,}75 & 1{,}25 \\ -0{,}25 & 2{,}75 & 1{,}25 & -0{,}75 \\ -0{,}75 & 1{,}25 & 2{,}75 & -0{,}25 \\ 1{,}25 & -0{,}75 & -0{,}25 & 2{,}75 \end{bmatrix}$.

4. $\begin{bmatrix} 3{,}6 & 4{,}4 & 0{,}8 & -1{,}6 & -2{,}8 \\ 4{,}4 & 2{,}6 & 1{,}2 & -0{,}4 & 0{,}8 \\ 0{,}8 & 1{,}2 & 0{,}8 & -4{,}0 & -2{,}8 \\ -1{,}6 & -0{,}4 & -4{,}0 & 1{,}2 & 2{,}0 \\ -2{,}8 & 0{,}8 & -2{,}8 & 2{,}0 & 1{,}8 \end{bmatrix}$.

5. $\boldsymbol{A} = [a_{ij}]$, где $a_{ij} = \begin{cases} i+j & i = j \\ ij & i \neq j \end{cases}$ и $i, j = 1, 2, \ldots, 30$.

6. $\boldsymbol{A} = [a_{ij}]$, где $a_{ij} = \begin{cases} \cos(\sin(i+j)) & i = j \\ i + ij + j & i \neq j \end{cases}$ и $i, j = 1, 2, \ldots, 40$.

7. Напишите программу, которая реализует QR-метод для симметричной матрицы.
8. Модифицируйте программу 11.5 таким образом, чтобы она обращалась к вашей программе из задачи 7, как к подпрограмме. Используйте эту модификацию программы, чтобы найти собственные значения матриц из задач 1–6.