Министерство образования
Российской Федерации

Московский государственный университет леса

В.И. Мышенков, Е.В. Мышенков

# ЧИСЛЕННЫЕ МЕТОДЫ

Часть первая

Учебное пособие
для студентов специальности 0101.07

Издательство Московского государственного университета леса
Москва – 2001

УДК 519.6
6Л2 Мышенков В.И., Мышенков Е.В.
Численные методы. Часть первая: Учебное пособие для студентов специальности 0101.07. – М.:МГУЛ,2001. – 120 с.: ил.

Учебное пособие содержит изложение основных понятий и методов теории погрешностей, аппроксимации, численного дифференцирования, вычисления определенных интегралов, решения нелинейных уравнений, систем линейных и нелинейных уравнений, методов решения задач на собственные значения.

Одобрено и рекомендовано к изданию в качестве учебного пособия редакционно-издательским советом университета.

Рецензенты: профессор И.М. Степанов,
Кафедра вычислительной техники,
доцент А.Г. Королев,
Кафедра прикладной математики.

Авторы: - Виталий Иванович Мышенков, профессор;
Евгений Витальевич Мышенков, доцент

Редактор Е.Г. Петрова

Компьютерный набор и верстка В.И. Мышенкова, Е.В. Мышенкова

По тематическому плану внутривузовских изданий учебной литературы на 2001г. поз. 53.

ЛР № 020718 от 02.02.1998 г.

Подписано к печати Тираж 100 экз.
Объем 7,50 п.л. Заказ №

Издательство Московского государственного университета леса.
141005. Мытищи-5. Московская обл., 1-я Институтская, 1, МГУЛ.
Телефон: (095) 588-57-62.
E-mail:izdat@mgul.ac.ru

## Введение

Создание электронных вычислительных машин, ЭВМ, в середине XX века явилось выдающимся техническим достижением – революцией в истории человечества. Если предыдущие технические революции расширяли физические возможности трудовой деятельности человека, то создание ЭВМ расширило его интеллектуальные возможности. Стало возможным более эффективное познание законов реального мира, значительное увеличение производительности труда, развитие производства, совершенствование управления и т. д.

Поскольку для реализации этих потенциальных возможностей использования ЭВМ необходимо наличие квалифицированных специалистов, во многих вузах страны созданы кафедры прикладной математики и вычислительной техники, и читаются студентам соответствующие курсы.

Настоящее пособие представляет собой первую часть курса численных методов, читаемого студентам кафедры прикладной математики МГУЛ. Курс может быть полезен также аспирантам при начальном ознакомлении с численными методами решения различных задач.

Данное издание содержит первую часть курса численных методов, включающую изложение основных понятий и методов теории погрешностей, аппроксимации, численного дифференцирования, вычисления определенных интегралов, решения нелинейных уравнений, систем линейных и нелинейных уравнений, методов решения задач на собственные значения. Для желающих углубить и расширить свои знания методов численного моделирования и решения задач в конце книги приводится список основной и дополнительной литературы по данной тематике.

## 1. МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ИХ СОЗДАНИЕ И СОВЕРШЕНСТВОВАНИЕ

Особенностью настоящего времени является широкое применение математических методов и ЭВМ в различных областях человеческой деятельности: в науке, технике, экономике, медицине и даже в лингвистике. Такое широкое внедрение математики в сферу общественно-политической, произ

водственной и других областей жизни вызвано необходимостью анализа и прогнозирования явлений и процессов, происходящих в обществе и природе. Для осуществления указанных целей прежде всего необходимо разработать математическую модель рассматриваемого явления, процесса или объекта. Математическая модель – это описание наиболее существенных свойств и особенностей явления на языке математических понятий и уравнений.

Математическая модель, основанная на упрощении, идеализации, не тождественна реальному явлению, объекту, а является его приближенным описанием. Однако благодаря замене реального объекта приближенной моделью становится возможным его математическое описание и применение математического аппарата для его анализа. Математика позволяет провести детальный анализ рассматриваемого явления, предсказать его поведение в различных условиях и в будущем.

Сложность математической модели и ее исследования зависит от сложности исследуемого объекта. Если раньше математические методы и модели применялись лишь в механике, физике, астрономии, изучающих простейшие формы движения, то с появлением ЭВМ и развитием вычислительной математики математические методы находят применение и в других областях деятельности человека.

Построение модели объекта, явления начинается с выделения его наиболее существенных черт и свойств и описания их при помощи математических соотношений. Затем, после создания математической модели, ее исследуют математическими методами, то есть решают сформулированную математическую задачу.

В качестве примера рассмотрим задачу определения площади поверхности стола. Моделью этой поверхности, на первый взгляд, может служить прямоугольник со сторонами, равными сторонам стола. Если же длины противоположных сторон стола и его диагоналей окажутся не равными, в качестве модели нужно принять четырехугольник. Для более точного определения площади стола необходимо учесть еще скругления его угловых кромок. Таким образом, с повышением требований к точности определения площади стола его математические модели постоянно уточняются. Следовательно, математическая модель не определяется однозначно исследуемым объектом. Выбор конкретной модели определяется требованиями ее точности.

Построение математической модели является одним из наиболее сложных и ответственных этапов исследования объекта. Математическая модель никогда не бывает тождественна рассматриваемому объекту, не передает всех его свойств, так как основывается на упрощении и идеализации объекта. Поэтому результаты, получаемые на основе этой модели, имеют всегда приближенный характер. Их точность определяется степенью соответствия, адекватности модели и объекта. Вопрос о точности является важнейшим в прикладной математике. Однако он не является чисто математическим во

просом и не может быть решен математическими методами. Основным критерием истины является эксперимент, то есть сопоставление результатов, получаемых на основе математической модели, с рассматриваемым объектом. Только практика позволяет сравнить различные гипотетические модели и выбрать из них наиболее простую и достоверную, указать области применимости различных моделей и направление их совершенствования.

Рассмотрим развитие модели на примере известной задачи баллистики об определении траектории тела, выпущенного с начальной скоростью $v_0$ под углом $\alpha_0$ к горизонту. Для начала предположим, что скорость $v$ и дальность полета тела небольшие. Тогда для данной задачи будет справедлива математическая модель Галилея, основанная на следующих допущениях:

1) Земля – инерциальная система;
2) ускорение свободного падения $g$ постоянно;
3) Земля – плоское тело;
4) сопротивление воздуха отсутствует.

В этом случае составляющие скорости движения тела по осям $x$ и $y$ равны $v_x = v_0 \cos\alpha_0$; $v_y = v_0 \sin\alpha_0 - gt$, а их пути

$$x = tv_0 \cos\alpha_0; \quad y = tv_0 \sin\alpha_0 - gt^2/2,$$

где $t$ – время движения. Определяя $t$ из первого уравнения $t = x/(v_0 \cos\alpha_0)$ и подставляя его во второе, получаем уравнение траектории тела, представляющее собой параболу:

$$y = x \operatorname{tg}\alpha_0 - x^2 g/(2v_0^2 \cos^2\alpha_0).$$

Из условия $y = 0$ получаем дальность полета тела

$$l = \left(v_0^2/g\right)\sin 2\alpha_0. \qquad (1.1)$$

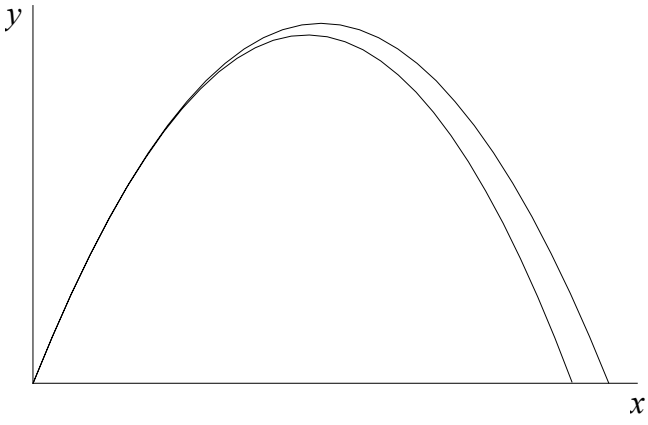


Рис. 1

Однако, как показывает практика, результаты, получаемые на основе этой модели, оказываются справедливыми лишь при малых начальных скоростях движения тела $v_0 < 30$ м/с. С увеличением скорости $v_0$ дальность полета становится меньше величины, даваемой формулой (1.1). На рис. 1 пунктирной линией изображена траектория тела, полученная на основе модели Галилея для $v_0 = 80$ м/с, а сплошной линией – реальная траектория. Такое

расхождение эксперимента с расчетной формулой (1.1) говорит о неточности модели Галилея, не учитывающей сопротивление воздуха.

Дальнейшее уточнение модели баллистической задачи в части учета сопротивления воздуха было сделано Ньютоном. Это позволило с достаточной точностью рассчитывать траектории движения пушечных ядер, выстреливаемых со значительными начальными скоростями.

Переход от гладкоствольного к нарезному оружию позволил увеличить скорость, дальность и высоту полета снарядов, что потребовало дальнейшего уточнения математической модели задачи. В новой математической модели были пересмотрены все допущения, принятые в модели Галилея, то есть Земля уже не считалась плоской и инерциальной системой, и сила земного притяжения не принималась постоянной.

Последующее совершенствование математической модели задачи связано с использованием методов теории вероятности. Это было вызвано тем, что параметры снарядов, орудий, зарядов и окружающей среды в силу допусков обработки деталей и других причин не остаются неизменными, а подчиняются случайным колебаниям.

В результате последовательных уточнений и усовершенствований была создана математическая модель, наиболее полно и точно описывающая задачу внешней баллистики. Сопоставление ее данных с результатами стрельб показало хорошее их совпадение.

На этом примере показаны этапы создания, развития и уточнения математической модели объекта, которые сопровождаются постоянно сопоставлением и проверкой практикой, то есть с самим реальным объектом или явлением. Именно недостаточно хорошее совпадение результатов, предоставляемых моделью, с объектом вызывает дальнейшее совершенствование модели.

Наконец отметим, что выбор конкретной математической модели объекта для его анализа необходимо производить из условия обеспечения достаточной точности получаемых результатов и простоты модели. При этом всегда следует помнить, что нельзя использовать очень точную и сложную модель объекта, когда требуется небольшая точность результатов.

# 2. ПОГРЕШНОСТИ ВЫЧИСЛЕНИЙ

## 2.1. Источники погрешностей. Классификация погрешностей

Анализ погрешности результатов вычислений должен являться непременной частью при математическом моделировании объекта, поскольку знание о нем является приближенным. Каждый инженер должен знать погрешность получаемых результатов эксперимента или расчета и правильно представлять их в отчетах, статьях и пр.

Причины возникновения погрешностей:

1) математическая модель объекта не является точным его образом, то есть не является точным математическое его описание. Не точно задаются исходные данные.

2) применяемый математический метод дает не точное решение задачи, а приближенное.

3) при вводе данных в машину и выполнении арифметических операций в ЭВМ производится округление чисел.

В результате получаются соответствующие погрешности:

1) неустранимая погрешность;

2) погрешность метода;

3) вычислительная погрешность.

В качестве примера можно привести задачу о маятнике. Математическая модель этой задачи представляется уравнением

$$l\varphi''(t) + \mu\varphi'(t) + g\sin\varphi(t) = 0, \qquad (2.1)$$

где $l$ – длина маятника; $\mu$ – коэффициент трения; $g$ – ускорение силы тяжести; $t$ – время; $\varphi(t)$ – угол отклонения маятника. В данной модели (2.1) уже существует неустранимая погрешность, поскольку она соответствует объекту лишь приближенно.

Дифференциальное уравнение (2.1) не решается в явном виде. При применении численного метода для его решения возникает погрешность метода, а при выполнении арифметических операций – вычислительная погрешность.

Пусть $I$ – точное значение отыскиваемого параметра; $\widetilde{I}$ – значение этого параметра, соответствующее принятому математическому описанию; $\widetilde{I}_h$ – решение задачи, получаемое при реализации численного метода при отсутствии округлений; $\widetilde{I}_h^*$ – приближение к решению, получаемое при реальных вычислениях.

Тогда получаем:

$\rho_1 = \left|I - \widetilde{I}\right|$ – неустранимая погрешность;

$\rho_2 = \left|\widetilde{I} - \widetilde{I}_h\right|$ – погрешность метода;

$\rho_3 = \left|\widetilde{I}_h - \widetilde{I}_h^*\right|$ – вычислительная погрешность.

Полная погрешность равна

$$\rho_0 = \left|I - \widetilde{I}_h^*\right| = \rho_1 + \rho_2 + \rho_3.$$

Важно знать неустранимую погрешность. Так как никакой процесс в природе нельзя описать точно, то, зная требования на точность конечного ответа, можно в пределах разумного производить необходимые упрощения ма

тематической модели. С другой стороны, нет никакой необходимости применять метод решения задачи с погрешностью, существенно меньшей, чем величина неустранимой погрешности. Таким образом, зная величину неустранимой погрешности, можно понизить требования к точности применяемых алгоритмов.

Пусть $A$ – точное значение некоторого параметра; $a$ – приближенное его значение, тогда абсолютной погрешностью приближения $a$ называют величину $\Delta a = |A - a|$.

Относительной погрешностью называют некоторую величину $\delta a$:

$$\delta a = \left|\frac{A - a}{A}\right| = \frac{\Delta a}{|A|}.$$

Если $\Delta a << a$, то $\delta a$ можно определить как

$$\delta a = \frac{\Delta a}{|a|}.$$

Значащими цифрами числа $a$ называются все цифры с первой ненулевой слева. Например, в числе $a = 0{,}00\underline{6380}$ все подчеркнутые цифры являются значащими.

Значащую цифру называют верной, если абсолютная погрешность числа меньше или равна половине единицы разряда, соответствующего этой цифре

*Пример*. $a_1 = 0{,}0\underline{345}1\underline{2}4$, $\Delta a_1 = 0{,}0000007$; $a_2 = 0{,}\underline{8362}$, $\Delta a_2 = 0{,}00004$. Подчеркнутые цифры являются верными.

Цифра считается сомнительной, если абсолютная погрешность результата не превышает двух единиц разряда, соответствующего этой цифре.

Точность приближенного числа зависит не от количества значащих цифр, а от количества верных значащих цифр. Поэтому, если приближенное число содержит излишнее количество значащих цифр, прибегают к его округлению. Окончательный результат должен содержать лишь на одну значащую цифру больше числа верных значащих цифр.

Информацию о том, что $a$ является приближенным значением числа $A$ с абсолютной погрешностью $\Delta a$, можно записать как

$$A = a \pm \Delta a.$$

Это эквивалентно определению

$$a - \Delta a < A < a + \Delta a.$$

Например, для приближенного значения числа $\pi$, заданного в пределах $3{,}140 < \pi < 3{,}142$, $\Delta\pi = |\pi - a| < 0{,}01$.

*Задача*. Вес $1\,дм^3$ воды при $t = 0^\circ\text{C}$ равен $p = 1000 \pm 0{,}05$ г. Определить абсолютную и относительную погрешности взвешивания.

Решение: $\Delta p = 0{,}05$ г, $\delta p = \dfrac{\Delta p}{p} = \dfrac{0{,}05}{1000} = 5 \cdot 10^{-5}$.

*Задача.* Экспериментально определили газовую постоянную воздуха $R = 29{,}32$. Относительная погрешность равна 2 %. Найти пределы, в которых заключается *R*.

Решение: $\delta R = 0{,}02$;

$$\Delta R = R\delta R = 29{,}32 \cdot 0{,}02 \cong 0{,}59;$$

$$29{,}32 - \Delta R \le R \le 29{,}32 + \Delta R;$$

$$28{,}73 \le R \le 29{,}91.$$

## 2.2. Связь числа верных знаков с относительной погрешностью

*Теорема*. Если положительное приближенное число *a* имеет *n* верных знаков, то относительная погрешность его $\delta a$ не превосходит величину $\delta a \le 0{,}1^{n-1}/(2\alpha_m)$, где $\alpha_m$ – первая значащая цифра числа *a*.

Представим *a* в виде

$$a = \alpha_m \cdot 10^m + \alpha_{m-1} \cdot 10^{m-1} + \ldots + \alpha_{m-n+1} \cdot 10^{m-n+1},$$

тогда $\Delta a = |A - a| \le 0{,}5 \cdot 10^{m-n+1}$. Отсюда

$$A \ge a - \Delta a = a - 0{,}5 \cdot 10^{m-n+1} \ge \alpha_m \cdot 10^m.$$

$$\delta a = \Delta a / |A| \le 0{,}5 \cdot 10^{m-n+1} / (\alpha_m \cdot 10^m) = 0{,}1^{n-1} / (2\alpha_m).$$

*Пример*. Определить относительную погрешность приближенного значения $a = 3{,}142$ числа $\pi$.

В нашем случае $a_m = 3$, $n = 4$. Отсюда

$$\delta a \le 0{,}1^{4-1} / (2 \cdot 3) = 1/6000 = 0{,}01666\,\%.$$

Для определения количества верных цифр числа *a*, если известна его относительная погрешность $\delta a$, пользуются приближенной формулой $\delta a = \Delta a / |a|$, откуда $\Delta a = |a| \delta a$.

Например, если $\delta a \le 10^{-n}$, то $\Delta a \le (\alpha_m + 1) \cdot 10^m \cdot 10^{-n} \le 10^{m-n+1}$, то есть число *a* заведомо имеет *n* верных знаков.

*Пример*. $a = 57384$, $\delta a = 1\,\%$. Сколько в числе *a* верных знаков?

$$\Delta a = a\delta a = 57384 \cdot 0{,}01 \cong 5{,}738 \cdot 10^2 .$$

Следовательно, число $a$ имеет верными лишь первые две цифры $(n = 2)$, и его правильная запись есть $a = 5{,}74 \cdot 10^4$.

### 2.3. Распространение ошибок в арифметических операциях

Пусть даны два приближенных числа $x = \bar{x} + \Delta x$ и $y = \bar{y} + \Delta y$, где $\bar{x}$ и $\bar{y}$ – их точные значения, и $\Delta x << x$, $\Delta y << y$. Эта запись означает, что $x$ и $y$ находятся в интервалах $(\bar{x} - \Delta x, \bar{x} + \Delta x)$ и $(\bar{y} - \Delta y, \bar{y} + \Delta y)$ соответственно. Найдем абсолютные ошибки суммы, разности, произведения и частного этих чисел.

Сумма чисел $x$ и $y$ находится в интервале $(a, b)$ с нижней и верхней границами:

$$a = \min(\bar{x} - \Delta x, \bar{x} + \Delta x) + \min(\bar{y} - \Delta y, \bar{y} + \Delta y) = \bar{x} + \bar{y} - (\Delta x + \Delta y);$$

$$b = \max(\bar{x} - \Delta x, \bar{x} + \Delta x) + \max(\bar{y} - \Delta y, \bar{y} + \Delta y) = \bar{x} + \bar{y} + (\Delta x + \Delta y).$$

Таким образом, ошибка суммы есть $\Delta(x + y) = \Delta x + \Delta y$.

Разность чисел $x$ и $y$ находится в интервале $(a, b)$ с границами:

$$a = \min(\bar{x} - \Delta x, \bar{x} + \Delta x) - \max(\bar{y} - \Delta y, \bar{y} + \Delta y) = \bar{x} - \bar{y} - (\Delta x + \Delta y);$$

$$b = \max(\bar{x} - \Delta x, \bar{x} + \Delta x) - \min(\bar{y} - \Delta y, \bar{y} + \Delta y) = \bar{x} - \bar{y} + (\Delta x + \Delta y).$$

Отсюда получаем ошибку разности как $\Delta(x - y) = \Delta x + \Delta y$.

Для произведения двух чисел $xy$ имеем интервал $(a, b)$ с границами:

$$a = \min(\bar{x} - \Delta x, \bar{x} + \Delta x) \cdot \min(\bar{y} - \Delta y, \bar{y} + \Delta y) = \bar{x}\bar{y} - (y\Delta x + x\Delta y) + \Delta x \Delta y;$$

$$b = \max(\bar{x} - \Delta x, \bar{x} + \Delta x) \cdot \max(\bar{y} - \Delta y, \bar{y} + \Delta y) = \bar{x}\bar{y} + (y\Delta x + x\Delta y) + \Delta x \Delta y,$$

если $x, y > 0$. Пренебрегая малым квадратичным членом $\Delta x \Delta y$, ошибку произведения записываем в виде

$$\Delta(xy) \approx |\bar{x}|\Delta y + |\bar{y}|\Delta x.$$

Частное $x / y$ попадает в интервал $(a, b)$ с границами:

$$a = \frac{\min(\bar{x} - \Delta x, \bar{x} + \Delta x)}{\max(\bar{y} - \Delta y, \bar{y} + \Delta y)} = \frac{\bar{x} - \Delta x}{\bar{y} + \Delta y} \approx \frac{\bar{x} - \Delta x}{\bar{y}}\left(1 - \frac{\Delta y}{\bar{y}}\right) \approx \frac{\bar{x}}{\bar{y}} - \frac{\bar{y}\Delta x + \bar{x}\Delta y}{\bar{y}};$$

$$b=\frac{\max(\overline{x}-\Delta x,\overline{x}+\Delta x)}{\min(\overline{y}-\Delta y,\overline{y}+\Delta y)}=\frac{\overline{x}+\Delta x}{\overline{y}-\Delta y}\approx\frac{\overline{x}+\Delta x}{\overline{y}}\left(1+\frac{\Delta y}{\overline{y}}\right)\approx\frac{\overline{x}}{\overline{y}}+\frac{\overline{y}\Delta x+\overline{x}\Delta y}{\overline{y}}$$

при том же условии $x, y > 0$ (квадратичными членами пренебрегаем). Отсюда получаем ошибку

$$\Delta(x/y)=\frac{|\overline{y}|\Delta x+|\overline{x}|\Delta y}{\overline{y}^2}. \qquad (2.2)$$

Относительные погрешности суммы и разности $x \pm y$ равны

$$\delta(x\pm y)=\frac{\Delta(x\pm y)}{|\overline{x}\pm\overline{y}|}=\frac{\Delta x}{|\overline{x}\pm\overline{y}|}+\frac{\Delta y}{|\overline{x}\pm\overline{y}|}=\left|\frac{\overline{x}}{\overline{x}\pm\overline{y}}\right|\delta x+\left|\frac{\overline{y}}{\overline{x}\pm\overline{y}}\right|\delta y,$$

а произведения $xy$ и частного $x/y$ равны соответственно:

$$\delta(xy)=\frac{|\overline{x}|\Delta y+|\overline{y}|\Delta x}{|\overline{xy}|}=\delta y+\delta x;$$

$$\delta(x/y)=\left(\frac{\overline{y}\Delta x+\overline{x}\Delta y}{\overline{y}^2}\right)\Bigg/\left(\frac{\overline{x}}{\overline{y}}\right)=\delta x+\delta y.$$

*Задача*. Найти абсолютную и относительную погрешности объема параллелепипеда, имеющего стороны $a=6\pm0{,}03$ см; $b=8\pm0{,}03$ см; $c=3\pm0{,}01$ см.

Решение. $V=abc=6\times8\times3=144$ см$^3$, $\ln V=\ln a+\ln b+\ln c\Rightarrow$

$$\Delta V/V=\Delta a/a+\Delta b/b+\Delta c/c\Rightarrow\quad \delta V=\delta a+\delta b+\delta c=$$

$$=0{,}03/6+0{,}03/8+0{,}01/3\approx0{,}012;$$

$$\Delta V=V\delta V=144\cdot0{,}012\approx1{,}73\text{ см}^3.$$

## 2.4. Общая формула для погрешности функции

Пусть задана дифференцируемая функция $f(x_1,x_2,\ldots,x_n)$, и $|\Delta x_i|$, $i=1,2,\ldots,n$ – абсолютные погрешности аргументов функции. Тогда абсолютная погрешность функции будет

$$\Delta f=\left|f(x_1+\Delta x_1,x_2+\Delta x_2,\ldots,x_n+\Delta x_n)-f(x_1,x_2,\ldots,x_n)\right|.$$

Так как $|\Delta x_i|$ малы, то произведениями и квадратами их можно пренебречь. Поэтому можно положить

$$\Delta f \approx \left|df(x_1, x_2, \ldots, x_n)\right| = \left|\sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i\right| \le \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right| \left|\Delta x_i\right| \quad \Rightarrow \quad \Delta f = \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right| \left|\Delta x_i\right|,$$

$$\delta u \le \sum_{i=1}^{n} \left|\frac{(\partial f / \partial x_i)}{f}\right| \left|\Delta x_i\right| = \sum_{i=1}^{n} \left|\frac{\partial}{\partial x_i} \ln f(x_1, x_2, \ldots, x_{n)}\right| \left|\Delta x_i\right|,$$

то есть относительная погрешность равна

$$\delta f = \sum_{i=1}^{n} \left|\frac{\partial}{\partial x_i} \ln f\right| \left|\Delta x_i\right|.$$

При помощи этих формул можно определить абсолютную и относительную погрешности наиболее употребительных элементарных функций: логарифма, степенной функции и степени числа:

$$\Delta \ln x = \frac{d}{dx}(\ln x)\Delta x = \frac{\Delta x}{\bar{x}} = \delta x, \quad \delta \ln x = \frac{\delta x,}{\bar{x}};$$

$$\Delta e^x = \frac{d}{dx} e^x \Delta x = e^{\bar{x}} \Delta x, \quad \delta e^x = \Delta x;$$

$$\Delta x^a = \frac{d}{dx}(x^a)\Delta x = a x^{a-1} \Delta x = a x^a \delta x, \quad \delta x^a = a \delta x.$$

*Задача*. Найти абсолютную и относительную погрешности объема шара $R = 3 \pm 0{,}05$ см, $\pi = 3{,}14 + 0{,}002$. Объем шара равен $V = 4\pi R^3 / 3$.

Решение: $\partial V / \partial \pi = 4R^3 / 3 = 36{,}0; \quad \partial V / \partial R = 4\pi R^2 = 113{,}0;$

$$\Delta V = \left|\partial V / \partial \pi\right| \left|\Delta \pi\right| + \left|\partial V / \partial \pi\right| \left|\Delta R\right| = 36 \cdot 0{,}002 + 113 \cdot 0{,}05 \approx 5{,}7 \text{ см}^3.$$

Поэтому $V = 4 \cdot 3{,}14 \cdot 27 / 3 = 113 \pm 5{,}7$ см$^3$, $\quad \delta V = 5{,}7 / 113 \approx 0{,}05 \approx 5\ \%$.

## 2.5. Обратная задача теории погрешностей

Обратная задача теории погрешностей решает вопрос о том, каковы должны быть погрешности аргументов функции, чтобы абсолютная погрешность функции не превышала заданной величины. Эта задача математически не определена, поскольку решение можно обеспечить, по-разному устанавливая погрешности аргументов.

Простейшее решение обратной задачи дается принципом равных влияний. Согласно этому принципу предполагается, что дифференциалы

$(\partial f / \partial x_i)\Delta x_i$ функции $u = f(x_1, x_2, \ldots, x_n)$ одинаково влияют на образование общей погрешности функции $\Delta u$. Предполагая, что

$$\left|\frac{\partial u}{\partial x_1}\right|\left|\Delta x_i\right| = \left|\frac{\partial u}{\partial x_2}\right|\left|\Delta x_2\right| = \ldots = \left|\frac{\partial u}{\partial x_n}\right|\left|\Delta x_n\right|,$$

получаем

$$\Delta u = \sum_{i=1}^{n}\left|\frac{\partial u}{\partial x_i}\right|\left|\Delta x_i\right| = n\left|\frac{\partial u}{\partial x_i}\right|\left|\Delta x_i\right| \quad \Rightarrow \quad \left|\Delta x_i\right| = \frac{\Delta u}{n\left|\partial u / \partial x_i\right|}, \quad i = 1, 2, \ldots, n.$$

Эту задачу можно также решить, считая одинаковыми абсолютные или относительные погрешности всех аргументов. В первом случае получаем:

$$\Delta x_1 = \Delta x_2 = \ldots = \Delta x_n \quad \Rightarrow \quad \Delta f = \sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right|\left|\Delta x_i\right| = \left|\Delta x_i\right|\sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right|;$$

$$\left|\Delta x_i\right| = \frac{\Delta u}{\sum\limits_{i=1}^{n}\left|\dfrac{\partial u}{\partial x_i}\right|}.$$

Во втором случае $\delta x_1 = \delta x_2 = \ldots = \delta x_n$

$$\Delta u = \sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right|\left|\Delta x_i\right| = \sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right|\frac{\left|\Delta x_i\right|}{x_i}x_i = \sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right|\delta x_i x_i \quad \Rightarrow$$

$$\delta x_i = \frac{\Delta f}{\sum\limits_{i}\left|\partial u / \partial x_i\right|\left|x_i\right|} \quad \Rightarrow \quad \Delta x_i = \frac{\Delta f\left|x_i\right|}{\sum\limits_{i}\left|\partial u / \partial x_i\right|\left|x_i\right|}.$$

*Задача*. Радиус основания конуса $R \approx 2$ м, его высота $H \approx 3$ м. С какими абсолютными погрешностями нужно определить *R*, *H*, чтобы его объем можно было вычислить с точностью до $0{,}1\ \text{м}^3$?

Решение: Объем конуса равен $V = \pi R^2 H / 3$, $\Delta V = 0{,}1\ \text{м}^3$. Объем конуса зависит от трех аргументов: $\pi$, *R*, *H*. Вычисляем частные производные:

$$\partial V / \partial \pi = R^2 H / 3 = 4;\ \partial V / \partial R = 2\pi RH / 3 = 12{,}6;\ \partial V / \partial H = \pi R^2 / 3 = 4{,}19,$$

отсюда при помощи формулы $\Delta x_i = \dfrac{\Delta u}{n\left|\partial u / \partial x_i\right|}$ получаем:

$$\Delta\pi = \frac{0{,}1}{3 \cdot 4} \le 0{,}0083; \quad \Delta R = \frac{0{,}1}{3 \cdot 12{,}6} \le 0{,}0026\ \text{м}; \quad \Delta H = \frac{0{,}1}{3 \cdot 4{,}19} \le 0{,}008\ \text{м}.$$

# 3. КОНЕЧНЫЕ РАЗНОСТИ

## 3.1. Формулы вычисления *n*-й конечной разности функции

Пусть задана функция $y = f(x)$ и $\Delta x = h$ – фиксированный шаг аргумента, тогда приращением или первой конечной разностью функции $y$ называется выражение

$$\Delta y = \Delta f(x) = f(x + \Delta x) - f(x).$$

$n$-й конечной разностью функции $y$ будет

$$\Delta^n y = \Delta(\Delta^{n-1} y).$$

Например, вторая конечная разность имеет вид

$$\Delta^2 y = \Delta(\Delta y) = \Delta(f(x + \Delta x) - f(x)) =$$

$$= f(x + 2\Delta x) - 2f(x + \Delta x) + f(x).$$

Рассматривая символ $\Delta$ как оператор, можно указать следующие его свойства:

1) $\Delta(u + v) = \Delta u + \Delta v$;
2) $\Delta(cu) = c\Delta u$;
3) $\Delta^m(\Delta^n u) = \Delta^{m+n} u, \quad \Delta^0 u = u$.

Здесь $u$, $v$ – функции; $c$ – константа; $\Delta x$ – фиксированное приращение аргумента. Из выражения для приращения функции

$$\Delta f(x) = f(x + \Delta x) - f(x)$$

получаем значение функции в точке $x + \Delta x$:

$$f(x + \Delta x) = f(x) + \Delta f(x),$$

или, если считать $\Delta$ символьным множителем,

$$f(x + \Delta x) = (1 + \Delta) f(x).$$

Для значений функции в последующих точках имеем:

$$f(x + 2\Delta x) = (1 + \Delta) f(x + \Delta x) = (1 + \Delta)^2 f(x);$$

$$f(x + 3\Delta x) = (1 + \Delta) f(x + 2\Delta x) = (1 + \Delta)^3 f(x);$$

. . . . . .;

$$f(x + n\Delta x) = (1 + \Delta) f(x + (n-1)\Delta x) = (1 + \Delta)^n f(x).$$

Используя формулу бинома Ньютона

$$(a+b)^n = \sum_{m=0}^{n} C_n^m a^{n-m} b^m, \quad n = 1, 2, \ldots,$$

где $C_n^m = \dfrac{n!}{m!(n-m)!}$ $(m = 0, 1, \ldots \le n = 0, 1, \ldots)$, можно получить выражение функции в точке $x + n\Delta x$ через конечные разности в виде

$$f(x + n\Delta x) = (1+\Delta)^n f(x) = \sum_{m=0}^{n} C_n^m \Delta^m f(x). \qquad (3.1)$$

Аналогично можно получить выражение $n$-й конечной разности через значения функции. Полагая $\Delta = (1+\Delta) - 1$, можно записать

$$\Delta^n f(x) = ((1+\Delta) - 1)^n f(x) =$$

$$= (1+\Delta)^n f(x) - C_n^1 (1+\Delta)^{n-1} f(x) + \ldots + (-1)^n f(x).$$

Откуда, используя формулу (3.1), имеем выражение $n$-й конечной разности:

$$\Delta^n f(x) = f(x + n\Delta x) - C_n^1 f(x + (n-1)\Delta x) + \ldots + (-1)^n f(x).$$

### 3.2. Обобщение теоремы Лагранжа о конечном приращении

Пусть функция $f(x)$ имеет $n$ непрерывных производных на отрезке $[x, x + \Delta x]$, то есть принадлежит классу $C^n(x, x + n\Delta x)$, тогда справедливо следующее выражение для $n$-й конечной разности функции:

$$\Delta^n f(x) = (\Delta x)^n f^{(n)}(x + \theta n \Delta x), \qquad (3.2)$$

где $0 < \theta < 1$.

Доказательство (3.2) будет проведено методом индукции. При $n = 1$ следует теорема Лагранжа о конечном приращении функции:

$$\Delta f(x) = (\Delta x) f^{(1)}(x + \theta \Delta x).$$

Пусть для $k < n$ формула (3.2) верна

$$\Delta^k f(x) = (\Delta x)^k f^{(k)}(x + \theta k \Delta x), \quad 0 < \theta < 1.$$

Покажем, что она справедлива и для $(k+1)$:

$$\Delta^{k+1} f(x) = \Delta(\Delta^k f(x)) = \Delta^k(\Delta f(x)) = \Delta^k(f(x + \Delta x) - f(x)) =$$

$$= (\Delta x)^k [f^{(k)}(x + \Delta x + \theta' k \Delta x) - f^{(k)}(x + \theta' k \Delta x)].$$

Согласно теореме Лагранжа отсюда следует

$$\Delta^{k+1} f(x) = (\Delta x)^k \, \Delta x f^{(k+1)}(x + \theta'' \Delta x + \theta'' k \Delta x), \quad 0 < \theta'' < 1.$$

Полагая $\theta = (\theta' k + \theta'')/(k+1)$, получаем требуемое выражение

$$\Delta^{k+1} f(x) = (\Delta x)^{k+1} f^{(k+1)}(x + \theta(k+1)\Delta x).$$

Так как $k$ произвольно, следовательно, формула (3.2) верна и при $k = n$. Отсюда следует

$$f^{(n)}(x + \theta n \Delta x) = \frac{\Delta^n f(x)}{(\Delta x)^n},$$

а при малых $\Delta x$

$$f^{(n)}(x) \approx \frac{\Delta^n f}{(\Delta x)^n}. \tag{3.3}$$

Формулу (3.3) обычно используют для определения производных функций, заданных таблично.

## 4. АППРОКСИМАЦИЯ И ИНТЕРПОЛИРОВАНИЕ ФУНКЦИЙ

Аппроксимацией функции называется приближенное представление сложной (имеющей громоздкое математическое представление) или заданной в виде таблицы функции $f(x)$ более простой функцией $\varphi(x)$, имеющей минимальные отклонения от исходной функции в заданной области $x$. По сути аппроксимация – это моделирование сложной функции более простой с вычислительной точки зрения функцией.

Частным случаем аппроксимации является интерполяция – точечная аппроксимация, когда приближенная функция $\varphi(x)$ строится на заданном множестве точек $x_i$, $i = 0, 1, 2, \ldots, n$, причем в узловых точках $x_i$ значения аппроксимирующей функции $\varphi(x)$ и исходной функции $f(x)$ равны: $\varphi(x_i) = f(x_i)$.

Если приближение строится на заданном дискретном множестве точек $\{x_i\}$, то аппроксимация называется точечной. При построении приближения на непрерывном множестве точек (например, на отрезке $[a,b]$) аппроксимация называется непрерывной (или интегральной).

Интерполяция на всем участке $[a,b]$ называется глобальной, а на отдельных участках отрезка $[a,b]$ – кусочной или локальной.

Погрешность аппроксимации функции $f(x)$ полиномом $\varphi(x)$ можно оценивать по величине среднеквадратичного отклонения $S_a$ или по значению

максимального отклонения $\delta_i(\varphi) = |\varphi(x_i) - f(x_i)|$, $i = 0, 1, 2, \ldots, n$, которые должны быть меньше заданной погрешности $\varepsilon > 0$.

Среднеквадратичное отклонение $\varphi(x)$ от $f(x)$ определяется выражением

$$S_a = \frac{1}{n+1}\sum_{i=0}^{n}[\varphi(x) - f(x)]^2 .$$

При этом коэффициенты полинома $\varphi(x)$ $a_i$, $i = 0, 1, 2, \ldots, n$ подбираются из условия минимальности $S_a$.

Приближение называется равномерным, если

$$\delta_i(\varphi) = |f(x_i) - \varphi(x_i)| < \varepsilon, \quad i = 0, 1, 2, \ldots, n \quad на \quad [a,b].$$

Возможность построения аппроксимирующего многочлена следует из *теоремы* Вейерштрасса: если функция $f(x)$ непрерывна на отрезке $[a,b]$, то для любого $\varepsilon > 0$ существует многочлен $\varphi(x)$ степени $n = n(\varepsilon)$, абсолютное отклонение которого от функции $f(x)$ на $[a,b]$ меньше $\varepsilon$.

На практике для аппроксимации $f(x)$ чаще всего используются степенные многочлены $\varphi(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$, коэффициенты которых $a_i$, $i = 0, 1, 2, \ldots, n$ определяются из условия наименьшего отклонения $\varphi(x)$ от $f(x)$. Так же применяются дробно-рациональные выражения, многочлены Чебышева, ряды Фурье и др.

## 4.1. Обобщенная *n*-я степень числа *x*

Обобщенной *n*-й степенью числа *x* называется выражение

$$x^{[n]} = \prod_{i=0}^{n-1}(x - ih) = x(x-h)(x-2h)\ldots(x-(n-1)h),$$

где *h* – фиксированное постоянное число. По определению $x^{[0]} = 1$. При $h \to 0$ $x^{[n]}$ стремится к *n*-й степени числа *x*: $x^{[n]} \to x^n$.

Вычислим конечные разности $x^{[n]}$:

$$\Delta x^{[n]} = (x+h)^{[n]} - x^{[n]} = (x+h)\prod_{i=0}^{n-2}(x - ih) - (x-(n-1)h)\prod_{i=0}^{n-2}(x - ih) =$$

$$= nh\prod_{i=0}^{n-2}(x - ih) = nhx^{[n-1]}.$$

Продолжая далее аналогичным образом, получим

$$\Delta^2 x^{[n]} = \Delta(\Delta x^{[n]}) = \Delta(nhx^{[n-1]}) = n(n-1)h^2 x^{[n-2]};$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots;$$

$$\Delta^k x^{[n]} = n(n-1)(n-2)\ldots\{n-(k-1)\}h^k x^{[n-k]}.$$

## 4.2. Точечная аппроксимация. Понятие интерполирования

Пусть на отрезке $[a,b]$ заданы точки $x_0, x_1, \ldots, x_n$ и значения некоторой функции $y = f(x)$ в этих точках: $f(x_0) = y_0$; $f(x_1) = y_1$; …; $f(x_n) = y_n$. Требуется построить такую функцию $F(x)$, которая принимает в точках $x_i$ значения, равные значениям функции $f(x_i)$: $F(x_0) = y_0$; $F(x_1) = y_1$; …; $F(x_n) = y_n$. Такая функция $F(x)$ называется интерполирующей, а точки $x_0, x_1, \ldots, x_n$ – узлами интерполяции (рис. 2).

Интерполяционную функцию $F(x)$ используют для вычисления значений функции $f(x)$ в промежутках между точками $x_i, x_{i-1}$. Процесс вычисления функции $f(x)$ в промежуточных точках между $x_0, x_n$ называется интерполяцией, а за пределами отрезка $[a,b]$ – экстраполяцией.

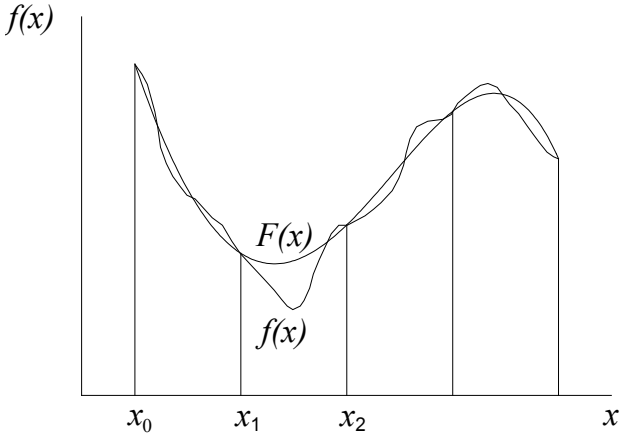


Рис. 2

Наиболее часто встречается интерполяция многочленами

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n.$$

Получим другие записи для многочлена $P_n(x)$.

## 4.3. Первая интерполяционная формула Ньютона

Пусть заданы узлы интерполяции $x_0, x_1, \ldots, x_n$, причем расстояния между узлами одинаковы: $x_i - x_{i-1} = h = \text{const}$, $h$ – шаг интерполяции. Требуется найти для функции $y = f(x)$ такой многочлен $P_n(x)$, что $P_n(x_i) = y_i$ и $\Delta^k P(x_0) = \Delta^k y_0$, для $k, i = 0, 1, 2, \ldots, n$.

$P_n(x)$ будем искать в виде

$$P_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots + a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}).$$

Используя понятие обобщенной степени числа, запишем его как

$$P_n(x) = a_0 + a_1(x - x_0)^{[1]} + a_2(x - x_0)^{[2]} + \ldots + a_n(x - x_0)^{[n]}.$$

Для определения коэффициентов $a_i$, $i = 0, 1, 2, \ldots, n$ вычислим $k$-е конечные разности полинома $P_n(x)$ в точке $x_0$, $k = 0, 1, 2, \ldots, n$ и приравняем их значения $k$-м конечным разностям самой функции $f(x_0)$:

$$\Delta^k P_n(x_0) = \Delta^k y_0, \quad (k = 0,1,\ldots,n);$$

$$\Delta P_n(x) = a_1 h + a_2 2h(x - x_0)^{[1]} + a_3 3h(x - x_0)^{[2]} + \ldots + a_n nh(x - x_0)^{[n-1]};$$

$$\Delta^2 P_n(x) = a_2 2h^2 + a_3 6h^2(x - x_0)^{[1]} + \ldots + a_n n(n-1)h^2(x - x_0)^{[n-2]};$$

. . . . . .

$$\Delta^k P_n(x) = a_k k!h^k + a_{k-1}(k+1)k(k-1)\ldots 2h^k(x - x_0)^{[1]} +$$
$$+ a_n n(n-1)\ldots(n-k+1)h^k(x - x_0)^{[n-k]}.$$

Так как при $x = x_0$ все члены в $\Delta^i P_n(x_0)$, кроме первого, равны нулю, получаем при $i = 0, 1, 2, \ldots, n$:

$$P_n(x_0) = y_0 \quad \Rightarrow \quad a_0 = y_0;$$

$$\Delta P_n(x_0) = a_1 h = \Delta y_0 \quad \Rightarrow \quad a_1 = \Delta y_0 / h;$$

$$\Delta^2 P_n(x_0) = 2a_2 h^2 = \Delta^2 y_0 \quad \Rightarrow \quad a_2 = \Delta^2 y_0 / (2h^2);$$

. . . . . .

$$\Delta^k P_n(x_0) = k! a_k h^k = \Delta^k y_0 \quad \Rightarrow \quad a_k = \Delta^k y_0 / (k! h^k),$$

$$k = 0, 1, 2, \ldots, n.$$

Подставляя найденные значения $a_k$ в полином Ньютона $P_n(x)$, имеем

$$P_n(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0)^{[1]} + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)^{[2]} + \ldots + \frac{\Delta^n y_0}{n!h^n}(x - x_0)^{[n]}.$$

Введя переменную $q = (x - x_0)/h$ и заменив множители $(x - x_0)^{[i]} / h^i$ выражениями $q(q-1)(q-2)\ldots(q-i+1)$, поскольку

$$\frac{(x - x_0)^{[i]}}{h^i} = \prod_{k=0}^{i-1} \frac{x - x_k}{h} = \prod_{k=0}^{i-1} \frac{x - x_0 - kh}{h} = \prod_{k=0}^{i-1} (q - k),$$

где $x_k = x_0 + kh, \quad k = 0, 1, 2, \ldots, n$, первый полином Ньютона представляем следующим образом:

$$P_n^1(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \ldots + \frac{q(q-1)\ldots(q-n+1)}{n!}\Delta^n y_0 . \quad (4.1)$$

При $n = 1$ имеем линейную интерполяцию

$$P_1(x) = y_0 + q\Delta y_0 ;$$

при $n = 2$ – квадратичную

$$P_2(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 .$$

Первая интерполяционная формула Ньютона (4.1) используется для интерполирования у левой границы отрезка $[a,b]$. Чаще всего на практике используются полиномы 1-й, 2-й и 3-й степени.

Погрешность интерполяции первой формулы Ньютона для функции $f(x) \in C^{n+1}[a,b]$, то есть при условии, что $f(x)$ на отрезке $[a,b]$ имеет непрерывные производные до $(n+1)$-го порядка включительно, есть

$$R_n^1(x) = f(x) - P_n^1(x) = \frac{h^{n+1} q(q-1)(q-2)\ldots(q-n)}{(n+1)!} f^{(n+1)}(\xi), \quad a \le \xi \le b .$$

Для интерполирования в конце отрезка $[a,b]$ используется вторая формула Ньютона.

### 4.4. Вторая интерполяционная формула Ньютона

Второй полином Ньютона строится следующим образом:

$$P_n^2(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + \ldots + a_n(x - x_n)\ldots(x - x_1) =$$
$$= a_0 + a_1(x - x_n)^{[1]} + a_2(x - x_{n-1})^{[2]} + \ldots + a_n(x - x_1)^{[n]}.$$

Предполагается $x_i - x_{i-1} = h = \text{const}$. Чтобы выполнялось условие $P_n^2(x_i) = y_i$, необходимо и достаточно, чтобы

$$\Delta^i P_n^2(x_{n-i}) = \Delta^i y_{n-i}, \; i = 0, 1, 2, \ldots, n .$$

Значения коэффициентов $a_i$ находим так же, как и в случае первого полинома Ньютона:

$$P_n(x_n) = y_n \quad \Rightarrow \quad a_0 = y_n ;$$

$$\Delta P_n(x) = a_1 h + a_2 2h(x - x_{n-1})^{[1]} + a_3 3h(x - x_{n-2})^{[2]} + \ldots + a_n nh(x - x_1)^{[n-1]}$$

$$\Rightarrow \quad \Delta P_n(x_{n-1}) = a_1 h = \Delta y_{n-1} \quad \Rightarrow \quad a_1 = \Delta y_{n-1}/h;$$

$$\Delta^2 P_n(x) = a_2 2h^2 + a_3 6h^2 (x - x_{n-2})^{[1]} + \ldots + a_n n(n-1)h^2 (x - x_1)^{[n-2]} \quad \Rightarrow$$

$$\Rightarrow \quad \Delta^2 P_n(x_{n-2}) = 2a_2 h^2 = \Delta^2 y_{n-2} \quad \Rightarrow \quad a_2 = \Delta^2 y_{n-2}/(2h^2);$$

. . . . . .;

$$\Delta^k P_n(x_{n-k}) = k! a_k h^k = \Delta^k y_{n-k} \quad \Rightarrow \quad a_k = \Delta^k y_{n-k}/(k! h^k).$$

Отсюда получаем рекуррентную формулу для коэффициентов

$$a_k = \Delta^k y_{n-k}/(k! h^k), \; k = 0, 1, 2, \ldots, n.$$

Подставляя значения $a_k$ во вторую формулу Ньютона, получаем

$$P_n^2(x) = y_n + \frac{\Delta y_{n-1}}{h}(x - x_n)^{[1]} + \frac{\Delta^2 y_{n-2}}{2! h^2}(x - x_{n-1})^{[2]} + \ldots + \frac{\Delta^n y_0}{n! h^n}(x - x_1)^{[n]}.$$

Положим $q = (x - x_n)/h$, тогда $(x - x_{n-1})/h = (x - x_n + x_n - x_{n-1})/h = q + 1$, $(x - x_{n-2})/h = (x - x_n + x_n - x_{n-2})/h = q + 2$ и т. д. В результате вторая интерполяционная формула Ньютона принимает вид

$$P_n^2(x) = y_n + q\Delta y_{n-1} + \frac{q(q+1)}{2!}\Delta^2 y_{n-2} + \ldots + \frac{q(q+1)\ldots(q+n-1)}{n!}\Delta^n y_0. \quad (4.2)$$

Она используется для интерполяции у правой границы отрезка $[a,b]$.

Погрешность интерполяции второй формулы Ньютона для $x \in [a,b]$ и $f(x) \in C^{n+1}[a,b]$ равна

$$R_n^2(x) = f(x) - P_n^2(x) = \frac{h^{n+1} q(q+1)(q+2)\ldots(q+n)}{(n+1)!} f^{(n+1)}(\xi), \quad \xi \in [a,b].$$

Для интерполяции в середине отрезка $[a,b]$ применяются формулы Бесселя, Гаусса и другие [1].

## 4.5. Формула Лагранжа

Пусть заданы узлы интерполяции $x_0, x_1, \ldots, x_n$, требуется построить полином $L_n(x)$ степени $n$, принимающий в точках $x_i$ значения функции $y_i = f(x_i)$:

$$L_n(x) = y_i, \quad i = 0, 1, \ldots, n.$$

Построим вначале такой полином $P_i(x)$, что

$$P_i(x_j) = \begin{cases} 1 & \text{при} \quad j = i; \\ 0 & \text{при} \quad j \neq i. \end{cases}$$

Этому условию удовлетворяет полином

$$P_i(x) = C_i(x - x_0)(x - x_1)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n) =$$

$$= C_i \prod_{k=0, k\neq i}^{n}(x - x_k), \quad \text{где} \quad C_i = \left( \prod_{k=0, k\neq i}^{n}(x_i - x_k) \right)^{-1}.$$

Используя многочлен $P_i(x)$, запишем полином Лагранжа следующим образом: $L_n(x) = \sum_{i=0}^{n} P_i(x) y_i(x)$, или в развернутом виде

$$L_n(x) = \sum_{i=0}^{n} \frac{(x - x_0)(x - x_1)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)} y_i \,.$$

На практике чаще всего используются полиномы Лагранжа первой, второй и третьей степеней. Например, при $n = 1$ получаем линейную формулу

$$L_1(x) = \frac{x - x_1}{x_0 - x_1} y_0 + \frac{x - x_0}{x_1 - x_0} y_1,$$

где $x_0, x_1$ – узлы интерполяции.

При $n = 2$ получаем уравнение параболы, проходящей через 3 точки $(x_0, y_0)$, $(x_1, y_1), (x_2, y_2)$:

$$L_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} y_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} y_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} y_2 \,.$$

Погрешность формулы Лагранжа есть $R_n(x) = f(x) - L_n(x)$. Пусть $f(x) \in C^{n+1}[a,b]$, то есть функция $f(x)$ имеет непрерывные производные до $(n+1)$-й включительно на $[a,b]$. Введем функцию

$$u(x) = f(x) - L_n(x) - k\Pi_{n+1}(x), \text{ где } \Pi_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n),$$

$k$ – постоянный коэффициент. Фиксируем точку $\bar{x} \in [a,b]$ и подбираем $k$ такое, что $u(\bar{x}) = 0$:

$$k = \frac{f(\bar{x}) - L_n(\bar{x})}{\Pi_{n+1}(\bar{x})}. \tag{4.3}$$

Таким образом, $u(x)$ имеет $(n+2)$ нулей на $[a,b]$ (в точках $x_0, x_1, \ldots, x_n$ и $\bar{x}$). Следовательно, существует такая точка $x = \xi$ на отрезке $[a,b]$, в которой

$u^{(n+1)}(\xi)=0$. Поскольку $L_n^{(n+1)}(x)=0$, и $\Pi_{n+1}^{(n+1)}(x)=(n+1)!$, $u^{(n+1)}(x)=f^{(n+1)}(x)-k(n+1)!$. Отсюда $f^{(n+1)}(\xi)-k(n+1)!=0$, и значение коэффициента есть $k=\dfrac{f^{(n+1)}(\xi)}{(n+1)!}$. Подставляя в (4.3) найденное значение *k*, получаем

$$\frac{f(\bar{x})-L_n(\bar{x})}{\Pi_{n+1}(\bar{x})}=\frac{f^{(n+1)}(\xi)}{(n+1)!},\quad\Rightarrow\quad f(\bar{x})-L_n(\bar{x})=\frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(\bar{x}).$$

Так как $\bar{x}$ произвольно, то $R_n(x)=f(x)-L_n(x)=\dfrac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x)$. Отсюда получаем выражение для погрешности формулы Лагранжа:

$$\left|R_n(x)\right|\le\frac{1}{(n+1)!}\left|\Pi_{n+1}(x)\right|\max_{x\in[a,b]}\left|f^{(n+1)}(x)\right|. \qquad (4.4)$$

Оценку погрешности для первой формулы Ньютона можно получить из формулы (4.4), положив расстояния между узлами интерполяции одинаковыми, $x_i-x_{i-1}=h=\text{const}$, и введя обозначение $q=(x-x_0)/h$:

$$R_n(x)=h^{n+1}\frac{q(q-1)\ldots(q-n)}{(n+1)!}f^{(n+1)}(\xi).$$

Погрешность второй формулы Ньютона имеет аналогичный вид:

$$R_n(x)=h^{n+1}\frac{q(q+1)\ldots(q+n)}{(n+1)!}f^{(n+1)}(\xi),\ \text{где}\ q=(x-x_n)/h:$$

Предполагая, что $\Delta^{n+1}y$ почти постоянно для функции $y=f(x)$ и *h* достаточно мало, и учитывая, что

$$f^{(n+1)}(\xi)\approx\frac{\Delta^{n+1}y_0}{h^{n+1}},$$

для формул Ньютона можно получить:

$$R_n^1(x)\approx\frac{q(q-1)\ldots(q-n)}{(n+1)!}\Delta^{n+1}y_0;\quad R_n^2(x)\approx\frac{q(q+1)\ldots(q+n)}{(n+1)!}\Delta^{n+1}y_n.$$

## 4.6. Практическое интерполирование

Чаще всего при обработке экспериментальных данных применяются линейная и квадратичная интерполяции. Рассмотрим их применение.

*а) Линейная интерполяция*

Для *i*-го интервала интерполирования уравнение прямой, проходящей через точки $(x_{i-1}, y_{i-1})$ и $(x_i, y_i)$, имеет вид

$$y = a_i x + b_i \text{ для } x \in [x_{i-1}, x_i].$$

Для определения коэффициентов $a_i$ и $b_i$ составим систему уравнений:

$$y_{i-1} = a_i x_{i-1} + b_i;$$

$$y_i = a_i x_i + b_i,$$

откуда получаем $a_i = (y_{i-1} - y_i)/(x_{i-1} - x_i)$; $b_i = y_{i-1} - a_i x_{i-1}$.

*б) Квадратичная интерполяция (параболическая)*

В этом случае уравнение имеет вид $y = a_i x^2 + b_i x + c_i$ для $x \in [x_{i-1}, x_{i+1}]$. Неизвестные коэффициенты $a_i$, $b_i$, $c_i$ определяются из условия прохождения параболы через три точки: $(x_{i-1}, y_{i-1})$, $(x_i, y_i)$, $(x_{i+1}, y_{i+1})$, путем решения системы трех линейных алгебраических уравнений:

$$y_{i-1} = a_i x_{i-1}^2 + b_i x_{i-1} + c_i;$$

$$y_i = a_i x_i^2 + b_i x_i + c_i;$$

$$y_{i+1} = a_i x_{i+1}^2 + b_i x_{i+1} + c_i.$$

## 4.7. Интерполяция и приближение сплайнами

Интерполирование функций многочленами Ньютона или Лагранжа на всем отрезке $[a,b]$ с использованием большого числа узлов интерполяции часто приводит к неудовлетворительному приближению из-за накопления погрешностей при вычислениях. Кроме того, увеличение числа узлов не всегда обеспечивает повышение точности интерполяции вследствие расходимости самого процесса интерполяции. Чтобы избежать увеличения погрешностей, обычно применяют кусочно-полиномиальную интерполяцию, то есть разбивают отрезок $[a,b]$ на частичные отрезки, на каждом из которых функцию $f(x)$ заменяют многочленом невысокой степени.

Одним из способов такой интерполяции является интерполирование при помощи сплайн-функций. Сплайн-функцией (сплайном) называют кусочно-полиномиальную функцию, определенную на отрезке $[a,b]$ и имеющую на нем некоторое число непрерывных производных.

Достоинством сплайнов в сравнении с обычной интерполяцией является их сходимость и устойчивость процесса вычислений.

Рассмотрим распространенный на практике случай кубического сплайна. Пусть на отрезке $[a,b]$ задана непрерывная функция $f(x)$. Зададим сетку точек $a = x_0 < x_1 < x_2 < \ldots < x_{N-1} < x_N = b$, перенумерованных в порядке возрастания, и обозначим $f_n = f(x_n), \quad n = 0, 1, \ldots, N$.

Сплайном, соответствующим функции $f(x)$ и узлам $x_n$, называется функция $s(x)$, удовлетворяющая условиям:

1) на каждом частичном отрезке $[x_{n-1}, x_n]$ функция $s(x)$ является многочленом третьей степени;

2) значения функции $s(x)$ в узлах интерполяции равны значениям функции $f(x)$: $s(x_n) = f(x_n), \quad n = 0, 1, \ldots, N$;

3) функция $s(x)$ и ее первая и вторая производные непрерывны на $[a,b]$.

*Построение сплайна.* На каждом отрезке $[x_{n-1}, x_n], \quad n = 1, 2, \ldots, N$ функцию $s(x) = s_n(x)$ будем искать в виде многочлена третьей степени

$$s_n(x) = a_n + b_n(x - x_n) + \frac{c_n}{2}(x - x_n)^2 + \frac{d_n}{6}(x - x_n)^3,$$

$$x_{n-1} \le x \le x_n, \quad n = 1, 2, \ldots, N,$$

где $a_n, b_n, c_n, d_n$ – неизвестные коэффициенты, подлежащие определению. Вычисляя первую, вторую и третью производные функции $s(x)$ –

$$s_n'(x) = b_n + c_n(x - x_n) + \frac{d_n}{2}(x - x_n)^2;$$

$$s_n''(x) = c_n + d_n(x - x_n), \qquad s_n''' = d_n,$$

получаем, что коэффициенты $a_n, b_n, c_n, d_n$ равны:

$$a_n = s_n(x_n); \quad b_n = s_n'(x_n); \quad c_n = s_n''(x_n); \quad d_n = s_n'''(x_n).$$

В узлах интерполирования $x = x_n$ по условию 2) $s_n(x_n) = f(x_n)$, при $n = 1, 2, \ldots, N$, следовательно,

$$a_n = s_n(x_n) = f(x_n) = f_n, \quad n = 1, 2, \ldots, N.$$

Дополнительно положим $a_0 = f(x_0)$.

Используя условия непрерывности функции $s_n(x)$ и ее производных на отрезке $[a,b]$ –

$$s_{n-1}(x_{n-1}) = s_n(x_{n-1}); \quad s'_{n-1}(x_{n-1}) = s'_n(x_{n-1});$$

$$s''_{n-1}(x_{n-1}) = s''_n(x_{n-1}), \quad n = 2, 3, \ldots, N,$$

и вводя обозначение $h_n = x_n - x_{n-1}$, получаем следующие уравнения для определения коэффициентов $b_n$, $c_n$, $d_n$, $n = 2, 3, \ldots, N$:

$$h_n b_n - \frac{h_n^2}{2} c_n + \frac{h_n^3}{6} d_n = f_n - f_{n-1}; \tag{4.5}$$

$$h_n c_n - \frac{h_n^2}{2} d_n = b_n - b_{n-1}; \tag{4.6}$$

$$h_n d_n = c_n - c_{n-1}. \tag{4.7}$$

В уравнениях (4.5), кроме условия непрерывности сплайна $s_{n-1}(x_{n-1}) = s_n(x_{n-1})$, для $n = 2, 3, \ldots, N$ используется условие $s_1(x_0) = f_0$, и нумерация в уравнениях (4.5) начинается с $n = 1$. Полученные уравнения (4.5)–(4.7), образуют систему $3N - 2$ уравнений относительно $3N$ неизвестных $b_n$, $c_n$, $d_n$, $n = 1, 2, \ldots, N$.

Два недостающих уравнения получаем из граничных условий для функции $s(x)$. Например, при свободном закреплении концов стержня кривизна линии на его концах равна нулю, то есть удовлетворяет условиям $f''(a) = f''(b) = 0$. Отсюда следует, что $s''_1(a) = 0, s''_N(b) = 0$, то есть $c_1 - h_1 d_1 = 0, \quad c_N = 0$.

Условие $c_1 - h_1 d_1 = 0$ совпадает с уравнением (4.7), при $n = 1$, если положить $c_0 = 0$. В результате получается замкнутая система уравнений для определения коэффициентов $b_n, c_n, d_n$:

$$h_n d_n = c_n - c_{n-1}, \quad n = 1, 2, \ldots, N, \quad c_0 = c_N = 0; \tag{4.8}$$

$$h_n c_n - \frac{h_n^2}{2} d_n = b_n - b_{n-1}, \quad n = 2, 3, \ldots, N; \tag{4.9}$$

$$h_n b_n - \frac{h_n^2}{2} c_n + \frac{h_n^3}{6} d_n = f_n - f_{n-1}, \quad n = 1, 2, \ldots, N. \tag{4.10}$$

Разрешая уравнение (4.10) относительно $b_n$, найдем разность $(b_n - b_{n-1})$:

$$b_n - b_{n-1} = \frac{1}{2}(h_n c_n - h_{n-1} c_{n-1}) - \frac{1}{6}(h_n^2 d_n - h_{n-1}^2 d_{n-1}) + \frac{f_n - f_{n-1}}{h_n} - \frac{f_{n-1} - f_{n-2}}{h_{n-1}}.$$

Подставляя затем ее в (4.9), получим уравнения

$$h_n c_n + h_{n-1} c_{n-1} - \frac{2}{3} h_n^2 d_n - \frac{1}{3} h_{n-1}^2 d_{n-1} = 2\left( \frac{f_n - f_{n-1}}{h_n} - \frac{f_{n-1} - f_{n-2}}{h_{n-1}} \right), \quad (4.11)$$

после чего при помощи (4.8) исключаем из уравнений (4.11) неизвестные $d_n$:

$$h_{n-1} c_{n-2} + 2(h_{n-1} + h_n) c_{n-1} + h_n c_n = 6\left( \frac{f_n - f_{n-1}}{h_n} - \frac{f_{n-1} - f_{n-2}}{h_{n-1}} \right).$$

Таким образом, для определения коэффициентов $c_n$ окончательно получаем систему уравнений

$$h_n c_{n-1} + 2(h_n + h_{n+1}) c_n + h_{n+1} c_{n+1} = 6\left( \frac{f_{n+1} - f_n}{h_{n+1}} - \frac{f_n - f_{n-1}}{h_n} \right),$$

$$n = 1, 2, \ldots, (N-1), \quad c_0 = c_N = 0,$$

матрица которой трехдиагональна и имеет диагональное преобладание. Данная система решается методом прогонки. По найденным коэффициентам $c_n$ определяются коэффициенты $d_n$ и $b_n$ из уравнений (4.8), (4.10):

$$d_n = \frac{c_n - c_{n-1}}{h_n}; \quad b_n = \frac{h_n}{2} c_n - \frac{h_n^2}{6} d_n + \frac{f_n - f_{n-1}}{h_n}, \; n = 1, 2, \ldots, N.$$

Погрешность интерполирования кубическим сплайном функции $f(x)$ и ее производных при условии, что $f(x) \in C^{(4)}[a,b]$, то есть имеет непрерывные производные до четвертой производной включительно, оценивается по формулам:

$$\left\| f(x) - s_n(x) \right\| \le M_4 h^4;$$

$$\left\| f'(x) - s_n'(x) \right\| \le M_4 h^3;$$

$$\left\| f''(x) - s_n''(x) \right\| \le M_4 h^2,$$

где $M_4 = \max\limits_{x \in [a,b]} \left| f^{(4)}(x) \right|$.

## 4.8. Подбор эмпирических формул

При обработке экспериментальных данных часто приходится представлять их в виде некоторой приближенной зависимости типа $y = f(x)$. Задача формулируется следующим образом.

Пусть в результате измерений получена таблица данных $x_i$, $y_i$. Требуется построить зависимость $y = f(x)$, называемую эмпирической формулой, которая бы приближенно отображала эти данные.

Если характер зависимости неизвестен, то вид эмпирической формулы может быть произвольным. В этом случае предпочтение отдается наиболее простым формулам, обладающим достаточной точностью. Их первоначальный вид можно выбрать из геометрических соображений.

Более строгий выбор эмпирической формулы производится на основе анализа *i*-х конечных разностей функции по данным таблицы. Например, если расстояние между узлами $\Delta x_i = x_i - x_{i-1} = \text{const}$, то:

1) при условии $\Delta y_i \approx \text{const}$ следует в качестве эмпирической формулы использовать линейную зависимость $y = ax + b$;

2) при $\Delta^2 y_i \approx \text{const}$ – квадратичную $y = ax^2 + bx + c$;

3) при $\Delta^3 y_i \approx \text{const}$ – кубическую $y = ax^3 + bx^2 + cx + d$ и т. д.

В случае, когда расстояния между точками $x_i$ различны, $\Delta x_i \neq \text{const}$, вместо конечных разностей функции $\Delta^i y$ необходимо рассматривать конечно-разностные представления производных, а в остальном процедура сохраняется. Например:

$$\Delta y_i \Rightarrow \frac{\Delta y_i}{\Delta x_i};$$

$$\Delta^2 y_i \Rightarrow \frac{(y_{i+1} - y_i)/\Delta x_i - (y_i - y_{i-1})/\Delta x_{i-1}}{(\Delta x_i + \Delta x_{i-1})/2}.$$

## 4.9. Определение параметров эмпирической формулы методом наименьших квадратов

Этот метод находит широкое применение при обработке экспериментальных данных. Предполагается, что каждое данное $y_i = f(x_i)$ измерено с некоторой неизвестной ошибкой $\varepsilon_i = y_i - \bar{y}_i$, где $\bar{y}_i$ – истинное значение. Вероятность же измерения величины с заданной ошибкой $\varepsilon_i$ определяется распределением Гаусса $(\sqrt{2\pi}\sigma)^{-1} \exp(-(y_i - \bar{y}_i)^2/(2\sigma^2))$, где $\sigma$ – дисперсия, которая для всех $i$ предполагается одинаковой. Поскольку измерения независимы, вероятность получения совокупности данных является произведением вероятностей всех независимых измерений:

$$(\sqrt{2\pi}\sigma)^{-m} \exp(-\sum_{i=0}^{m} (y_i - \bar{y}_i)^2/(2\sigma^2)).$$

Если данные предположения (гипотеза Гаусса) выполняются, среди семейства параметрических кривых $\bar{y} = \psi(x, a_0, a_1, \ldots, a_m)$ наибольшую вероятность для данной совокупности данных $y_i = f(x_i)$ обеспечивает кривая, для которой минимальна сумма квадратов отклонений во всех точках $x_0, x_1, \ldots, x_n$:

$$S = \sum_{i=0}^{n} \varepsilon_i^2 = \sum_{i=0}^{n} [\psi(x_i, a_0, a_1, \ldots, a_m) - y_i]^2 ,$$

причем погрешность эмпирической формулы оценивается по величине среднеквадратичного отклонения $R = S/(n+1)$.

Параметры эмпирической формулы $a_0, a_1, \ldots, a_m$ находятся из условия минимума функции $S = S(a_0, a_1, \ldots, a_m)$, который определяется путем приравнивания нулю частных производных этой функции по всем ее переменным $a_0, a_1, \ldots, a_m$:

$$\partial S / \partial a_0 = 0, \quad \partial S / \partial a_1 = 0, \quad \ldots, \quad \partial S / \partial a_m = 0 .$$

Полученные соотношения – система уравнений для определения параметров $a_0, a_1, \ldots, a_m$.

Очень часто в качестве эмпирической функции используется многочлен степени *m*:

$$\psi(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m .$$

Тогда

$$S = \sum_{i=0}^{n} (a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m - y_i)^2 ;$$

$$\frac{\partial S}{\partial a_k} = 2 \sum_{i=0}^{n} (a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m - y_i) x_i^k , \quad k = 0, 1, \ldots, m .$$

Приравнивая эти выражения нулю и собирая коэффициенты при неизвестных $a_0, a_1, \ldots, a_m$, получаем следующую систему уравнений:

$$a_0 \sum_{i=0}^{n} x_i^k + a_1 \sum_{i=0}^{n} x_i^{k+1} + a_2 \sum_{i=0}^{n} x_i^{k+2} + \ldots + a_m \sum_{i=0}^{n} x_i^{2m} = \sum_{i=0}^{n} x_i^k y_i , \qquad (4.12)$$

где $k = 0, 1, \ldots, m$. Решая эту систему линейных уравнений, получаем коэффициенты $a_0, a_1, \ldots, a_m$, которые являются искомыми параметрами эмпирической формулы.

*Пример*. Получить эмпирическую формулу для функции $f(x)$, заданной таблицей, используя метод наименьших квадратов.

| $x$ | 0,75 | 1,50 | 2,25 | 3,10 | 3,75 |
|---|---|---|---|---|---|
| $f(x)$ | 2,3 | 1,3 | 1,0 | 2,2 | 4,2 |

*Решение*. Табличные данные показаны на рис. 3 крестами. Из графика видно, что в качестве эмпирической функции $\psi(x)$ можно принять параболу

$$\psi(x) = a_0 + a_1 x + a_2 x^2 .$$

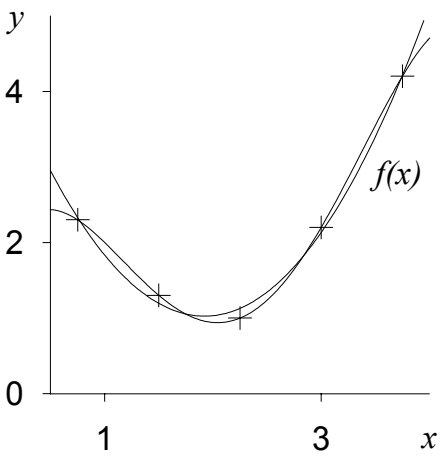


Рис. 3.

В данном случае $m = 2,\ \ n = 4$, и система уравнений (4.12) принимает следующий вид

$$b_{00}a_0 + b_{01}a_1 + b_{02}a_2 = c_0 ;$$

$$b_{10}a_0 + b_{11}a_1 + b_{12}a_2 = c_1 ;$$

$$b_{20}a_0 + b_{21}a_1 + b_{22}a_2 = c_2 .$$

Коэффициенты этой системы вычисляются по формулам

$$b_{jk} = \sum_{i=0}^{n} x_i^{j+k} ,\ c_j = \sum_{i=n}^{n} x_i^j y_i ,$$

откуда получаем:

$$5a_0 + 11{,}25a_1 + 30{,}94a_2 = 11{,}0 ;$$

$$11{,}25a_0 + 30{,}94a_1 + 94{,}92a_2 = 28{,}27 ;$$

$$30{,}94a_0 + 94{,}92a_1 + 309{,}76a_2 = 88{,}14 .$$

Решение этой системы есть $a_0 = 4{,}54;\ \ a_1 = -3{,}66;\ \ a_2 = 0{,}95$, и парабола, приближающая табличную зависимость и показанная на рис. 3 сплошной кривой, описывается формулой

$$\psi(x) = 4{,}54 - 3{,}66x + 0{,}95x^2 .$$

Парабола не проходит ни через одну табличную точку. Для сравнения на том же рисунке пунктиром показана интерполирующая кривая – полином 4-го порядка $P(x) = 1{,}7 + 3{,}39x - 4{,}69x^2 + 1{,}79x^3 - 0{,}198x^4$, которая имеет более осциллирующий характер. Это свойство точных интерполяций усиливается с ростом числа точек в таблице.

# 5. ПРИБЛИЖЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

## 5.1. Использование конечных разностей для дифференцирования

При проведении расчетов часто приходится вычислять значения производных функций. В случае задания функции в виде формулы их значения определяются по известным правилам дифференцирования. Если же функция задана в виде таблицы, то в предположении о существовании производных этой функции можно вычислить только приближенные значения производных, для чего используются их конечно-разностные аппроксимации (приближения).

Так, первая производная функции $y = f(x)$ по определению есть

$$y' = f'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x},$$

где $\Delta x$ – приращение $x$; $\Delta y = f(x + \Delta x) - f(x)$ – приращение $y$. Поскольку при табличном задании функции $f(x)$ $\Delta y$ и $\Delta x$ имеют конечные значения, в качестве приближения производной функции принимается выражение $y' \approx \Delta y / \Delta x$, называемое ее конечно-разностной аппроксимацией.

Пусть функция $y = f(x)$ задана таблично: $y_0, y_1, y_2, \ldots$ при $x = x_0, x_1, x_2, \ldots$, тогда

$$\Delta y_i = y_i - y_{i-1}, \quad \Delta x_i = x_i - x_{i-1}, \quad i = 1, 2, \ldots,$$

и первая производная функции в точке $x_i$ может быть приближенно представлена в виде

$$y'_i \approx (y_i - y_{i-1})/(x_i - x_{i-1}) = \Delta y_i / \Delta x_i$$

при помощи левых разностей, или в виде

$$y'_i \approx (y_{i+1} - y_i)/(x_{i+1} - x_i)$$

при помощи правых разностей, либо с помощью центральных разностей

$$y'_i \approx (y_{i+1} - y_{i-1})/(x_{i+1} - x_{i-1}).$$

Приближенные значения старших производных можно найти аналогичным образом, например, для второй производной:

$$y''_i \approx \frac{(y'_{i+1} - y'_i)}{x_{i+1} - x_i} = \frac{(y_{i+1} - y_i)/h_{i+1} - (y_i - y_{i-1})/h_i}{h_{i+1}},$$

где $h_{i+1} = x_{i+1} - x_i$. Если $h_i = \text{const}$, это выражение принимает вид

$$y''_i \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{h_i^2}.$$

Погрешность аппроксимации функции $f(x)$ некоторой функцией $\varphi(x)$, то есть $R(x) = f(x) - \varphi(x)$, обычно вычисляется с использованием разложения $\varphi(x)$ в ряд Тейлора и определения остаточных членов в выражении $R(x) = f(x) - \varphi(x)$. Пусть функция $f(x)$ имеет производные до *n*-й включительно на отрезке $[a,b]$. Определим погрешность аппроксимации первой производной при помощи правых конечных разностей

$$y_i' \approx \varphi'(x_i) = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, \qquad (5.1)$$

которая определяется как $R'(x) = f'(x) - \varphi'(x)$. Если $x_{i+1} - x_i = h = \text{const}$, то, разлагая функцию $y_{i+1} = f(x_{i+1})$ в ряд Тейлора в окрестности точки $x_i$,

$$y_{i+1} = y(x_i + h) = y(x_i) + y'(x_i)h + y''(x_i)h^2/2 + O(h^3),$$

и подставляя $y_i' = f'(x_i)$ и $\varphi'(x_i)$ в (5.1), имеем

$$R'(x_i) = y_i' - \frac{y_{i+1} - y_i}{h_i} = y'(x_i) - \frac{y_i + y_i'h + y_i''h^2/2 - y_i}{h} + O(h^2) =$$

$$= y'(x_i) - y'(x_i) - y''(x_i)h/2 + O(h^2) = O(h).$$

Степень *k* в члене $O(h^k)$ называется порядком погрешности аппроксимации. Говорят, что выражение (5.1) аппроксимирует производную в точке $x_i$ с первым порядком аппроксимации.

Для симметричной разности погрешность аппроксимации первой производной функции имеет второй порядок:

$$R'(x_i) = y_i' - \frac{y_{i+1} - y_{i-1}}{h_{i+1} - h_{i-1}} = y'(x_i) - \frac{y_i + y_i'h + y_i''h^2/2 + y_i'''h^3/6 - y_i}{2h} + O(h^3) =$$

$$= y'(x_i) - y'(x_i) + y''(x_i)h^2/6 + O(h^3) = O(h^2).$$

Аналогичную оценку погрешности аппроксимации можно получить и для второй производной при $h = \text{const}$:

$$y''(x_i) \approx (y_{i+1} - 2y_i + y_{i-1})/h^2;$$

$$R''(x_i) = y''(x_i) - (y_{i+1} - 2y_i + y_{i-1})/h^2 = y''(x_i) - (y_i + y_i'h + y_i''h^2/2 +$$

$$+ y_i'''h^3/6 - 2y_i + y_i - y_i'h + y_i''h^2/2 - y_i'''h^3/6 + O(h^4))/h^2 =$$

$$= y''(x_i) - y''(x_i) + O(h^2) = O(h^2).$$

Таким образом, погрешность этой аппроксимации – второго порядка.

## 5.2. Использование интерполяционных полиномов

Более точные значения производных можно получить при использовании интерполяционных полиномов. Пусть функция $y = f(x)$ интерполируется полиномом $P_n(x) \approx f(x)$. Приближенные значения производных определяем как $f^{(k)}(x) \approx P_n^{(k)}(x)$, $k = 1, 2, \ldots$. Если известна погрешность интерполяционного полинома $R(x) = f(x) - P_n(x)$, то погрешности приближений к производным будут таковы:

$$R^{(k)}(x) = f^{(k)}(x) - P_n^{(k)}(x).$$

1) *Формулы дифференцирования, основанные на первой формуле Ньютона*. Первая формула Ньютона задает полином, построенный по значениям функции в равноотстоящих узлах интерполирования $x_i$, $i = 0, 1, 2, \ldots, n$:

$$P_n(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2}\Delta^2 y_0 + \ldots + \frac{q(q-1)\ldots(q-n+1)}{n!}\Delta^n y_0,$$

где $q = (x - x_0)/h$, $h = x_i - x_{i-1}$. Считая этот полином приближением к искомой функции $y(x)$, найдем приближенное выражение для первой производной $y'(x)$. Учитывая $\frac{dP_n(x)}{dx} = \frac{dP_n(x)}{dq}\frac{dq}{dx} = \frac{1}{h}\frac{dP_n(x)}{dx}$, получим

$$y'(x) \approx \frac{1}{h}\left[\Delta y_0 + \frac{2q-1}{2}\Delta^2 y_0 + \frac{3q^2 - 6q + 2}{6}\Delta^3 y_0 + \ldots\right].$$

Вторая производная имеет вид $y'' = \frac{1}{h}\frac{d(y')}{dq}$,

$$y''(x) \approx \frac{1}{h^2}\left[\Delta^2 y_0 + (q-1)\Delta^3 y_0 + \frac{6q^2 - 18q + 11}{12}\Delta^4 y_0 + \ldots\right].$$

Для производных в узлах таблицы $x = x_0$; $q = 0$ получаем следующие формулы:

$$y'(x_0) \approx \frac{1}{h}\left[\Delta y_0 - \frac{\Delta^2 y_0}{2} + \frac{\Delta^3 y_0}{3} - \frac{\Delta^4 y_0}{4} + \ldots\right];$$

$$y''(x_0) \approx \frac{1}{h^2}\left[\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 - \frac{5}{6}\Delta^5 y_0 + \ldots\right].$$

Оценим погрешность вычисления производных. Ранее была приведена формула погрешности интерполяционной формулы Ньютона

$$R_k(x) = \frac{h^{k+1} q(q-1)(q-2)\ldots(q-k)}{(k+1)!} y^{(k+1)}(\xi), \quad \xi \in [x_0, x_n], \quad y(x) \in C^{(k+2)}.$$

Тогда, дифференцируя $R_k(x)$ и полагая $x = x_0$, $q = 0$, получим выражение для погрешности вычисления первой производной функции $y'(x)$ в точке $x = x_0$:

$$R_k'(x) = f'(x) - P_k'(x) = \frac{1}{h}\frac{dP_k}{dq} = \frac{h^k}{(k+1)!}\Big[y^{(k+1)}(\xi) \cdot$$

$$\cdot \frac{d}{dq}\{q(q-1)\ldots(q-k)\} + q(q-1)\ldots(q-k)\frac{d}{dq} y^{(k+1)}(\xi)\Big] \quad \Rightarrow$$

$$R_k'(x_0) = \frac{(-1)^k h^k y^{(k+1)}(\xi)}{k+1},$$

так как $\left.\frac{d}{dq}\{q(q-1)\ldots(q-k)\}\right|_{q=o} = (-1)^k k!$.

Аналогично можно получить погрешности производных более высокого порядка.

2) *Формулы дифференцирования, основанные на многочлене Лагранжа* для случая равноотстоящих узлов интерполяции $x_i - x_{i-1} = h = \text{const}$:

$$L_n(x) = \sum_{i=0}^{n} \frac{y_i \Pi_{n+1}(x)}{(x - x_i)\Pi_{n+1}'(x_i)}, \quad \Pi_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n).$$

Полагая $q = (x - x_0)/h$, преобразуем выражения в формуле Лагранжа и саму формулу к виду

$$\Pi_{n+1}(x) = h^{n+1} q(q-1)\ldots(q-n) = h^{n+1} q^{[n+1]};$$

$$\Pi_{n+1}'(x_i) = (x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n) =$$

$$= h^n i(i-1)\ldots 1 \cdot (-1)\ldots(i-n) = (-1)^{n-i} h^n \cdot i!(n-i)! \quad \Rightarrow$$

$$L_n(x) = \sum_{i=0}^{n} \frac{(-1)^{n-i}}{i!(n-i)!} \frac{q^{[n+1]}}{q-i} y_i. \qquad (5.2)$$

Отсюда получаем аппроксимацию первой производной:

$$y'(x) \approx L_n'(x) = \frac{1}{h}\sum_{i=0}^{n} y_i \frac{(-1)^{n-i}}{i!(n-i)!}\frac{d}{dq}\left\{\frac{q^{(n+1)}}{q-i}\right\}.$$

Аналогично получаются приближенные выражения старших производных.

Для оценки погрешности аппроксимации производных воспользуемся формулой

$$R_n(x) = y(x) - L_n(x) = \frac{y^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x), \quad \xi \in [x_0, x_n].$$

Предполагая $y(x) \in C^{n+2}$, имеем

$$R_n'(x) = \frac{1}{(n+1)!}\left\{y^{(n+1)}(\xi)\Pi_{n+1}'(x) + \Pi_{n+1}(x)\frac{d}{dx}y^{(n+1)}(\xi)\right\}.$$

Учитывая выражения для $\Pi_{n+1}'(x_i)$ и предполагая $\frac{d}{dx}y^{(n+1)}(\xi)$ ограниченной, получаем погрешность аппроксимации первой производной в узле $x_i$:

$$R_n'(x_i) = (-1)^{n-i}h^n \frac{i!(n-i)!}{(n+1)!}y^{(n+1)}(\xi).$$

Получим выражения для аппроксимаций первых производных на основе полинома Лагранжа второй степени $n = 2$ (три узловые равноотстоящие точки: $x_0, x_1, x_2$)

$$L_2(x) = 0.5y_0(q-1)(q-2) - y_1 q(q-2) + 0.5y_2 q(q-1).$$

Так как $q = (x - x_0)/h$ и $\frac{d}{dx} = \frac{d}{dq}\frac{dq}{dx} = \frac{1}{h}\frac{d}{dq}$,

$$y'(x) = \frac{1}{h}\left[\frac{1}{2}y_0(2q-3) - 2y_1(q-1) + \frac{1}{2}y_2(2q-1)\right].$$

Аппроксимации производных в отдельных точках $y_i' = y'(x_i)$ будут таковы:

$$y_0' \approx \frac{1}{2h}(-3y_0 + 4y_1 - y_2), \quad R_0' = \frac{1}{3}h^2 y'''(\xi_0), \quad (q_0 = 0);$$

$$y_1' \approx \frac{1}{2h}(-y_0 + y_2), \quad R_1' = -\frac{1}{6}h^2 y'''(\xi_1), \quad (q_1 = 1);$$

$$y_2' \approx \frac{1}{2h}(y_0 - 4y_1 - 3y_2), \quad R_2' = \frac{1}{3}h^2 y'''(\xi_2), \quad (q_2 = 2),$$

где $\xi_i \in [x_0, x_2], \quad i = 0, 1, 2.$

Аналогичным образом получаются формулы аппроксимаций производных для большего числа узлов $n = 4, 5, \ldots$.

## 6. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

К вычислению определенного интеграла сводятся многие практические задачи, такие, как вычисление площадей фигур, объемов тел, работы некоторой силы и др.

Пусть для непрерывной функции $f(x)$, определенной на отрезке $[a,b]$, требуется вычислить определенный интеграл $S = \int_a^b f(x)dx$. Геометрически это означает, что необходимо вычислить площадь фигуры, заключенной между осью $x$ и кривой $y = f(x)$ и ограниченной слева и справа прямыми, проходящими через точки $x = a$ и $x = b$, иначе называемой криволинейной трапецией (рис. 4).

Обычно понятие определенного интеграла вводится как предел интегральной суммы при неограниченном увеличении числа точек разбиения отрезка $[a,b]$. Интегральной суммой называется сумма площадей элементарных прямоугольников, которые получаются в результате разбиения отрезка $[a,b]$ на $n$ элементарных отрезков и построения боковых сторон прямоугольников, проходящих через узловые точки $x_i$, $i = 0, 1, 2, \ldots, n$:

$$S = \sum_{i=1}^{n} s_i = \sum_{i=1}^{n} f(x_i)\Delta x_i, \quad \Delta x_i = x_i - x_{i-1}.$$

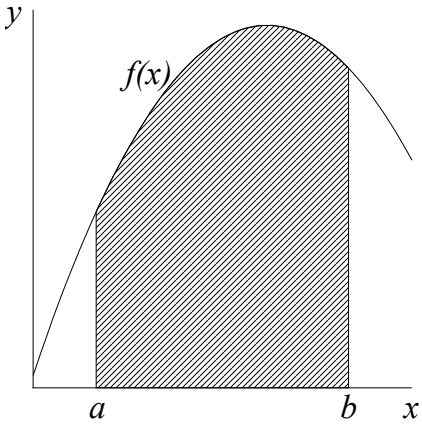


Рис. 4

В результате этой процедуры искомая площадь криволинейной трапеции заменяется площадью ступенчатой фигуры, состоящей из суммы площадей отдельных прямоугольников, интегральной суммой. Площадь этой ступенчатой фигуры при $\Delta x \to 0$ стремится к площади криволинейной трапеции.

В случаях, когда подынтегральная функция $f(x)$ задана в аналитическом виде, определенный интеграл можно вычислить по формуле Ньютона-Лейбница, то есть через значение первообразной $F(x)$

$$\int_a^b f(x)dx = F(x)\Big|_a^b = F(b) - F(a).$$

Однако на практике этот способ вычисления определенного интеграла используется редко, поскольку не каждая функция $f(x)$ имеет первообразную, которая выражается через элементарные функции, когда же $f(x)$ задана таблицей, этот метод вообще не применим. В таких случаях применяются методы численного интегрирования.

Вычислительный алгоритм строится следующим образом. Отрезок интегрирования $[a,b]$ разбивается на $n$ равных частичных отрезков $[x_{i-1}, x_i]$, $i = 1, 2, ..., n$ длиной $h = (b-a)/h$, а интеграл $\int_a^b f(x)dx$ заменяется суммой частичных интегралов

$$\int_a^b f(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx .$$

Затем подынтегральная функция $f(x)$ на частичном отрезке $[x_{i-1}, x_i]$ заменяется некоторым интерполяционным полиномом невысокой степени $m$ $L_{m,i}(x)$, и вычисляется интеграл $\int_{x_{i-1}}^{x_i} L_{m,i}(x)dx$. В результате получается приближенное значение интеграла

$$\int_a^b f(x)dx \approx \sum_{k=0}^{n} c_k f(x_k) .$$

Эта формула называется квадратурной, точки $x_k$ – узлами, а числа $c_k$ – коэффициентами этой формулы. Погрешность квадратурной формулы определяется из выражения

$$R_n = \int_a^b f(x)dx - \sum_{k=0}^{n} c_k f(x_k) .$$

В зависимости от выбора интерполяционного полинома $L_{m,i}(x)$ получаются различные квадратурные формулы. Рассмотрим простейшие из них.

## 6.1. Формула прямоугольников

В этом методе функция $f(x)$ на отрезке $[x_{i-1}, x_i]$ заменяется полиномом нулевой степени $L_{0,i}(x) = f(\xi_i)$, $\xi_i \in [x_{i-1}, x_i]$. В результате получается приближенное значение интеграла на частичном отрезке

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx \int_{x_{i-1}}^{x_i} f(\xi_i)dx = f(\xi_i)h , \qquad (6.1)$$

так как $f(\xi_i) = \text{const}$; $h = x_i - x_{i-1}$.

Полное значение интеграла вычисляется посредством интегральной суммы

$$\int_a^b f(x)dx \approx \sum_{k=1}^{n} f(x_k)h. \qquad (6.2)$$

В зависимости от выбора точки $\xi_i$ получаются различные формулы прямоугольников. Если выбрать в качестве $\xi_i$ координату левой стороны прямоугольника на отрезке $[x_{i-1}, x_i]$, то есть $\xi_i = x_{i-1}$, получается следующая формула для интеграла (6.1) (рис. 5а):

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx f(x_{i-1})h.$$

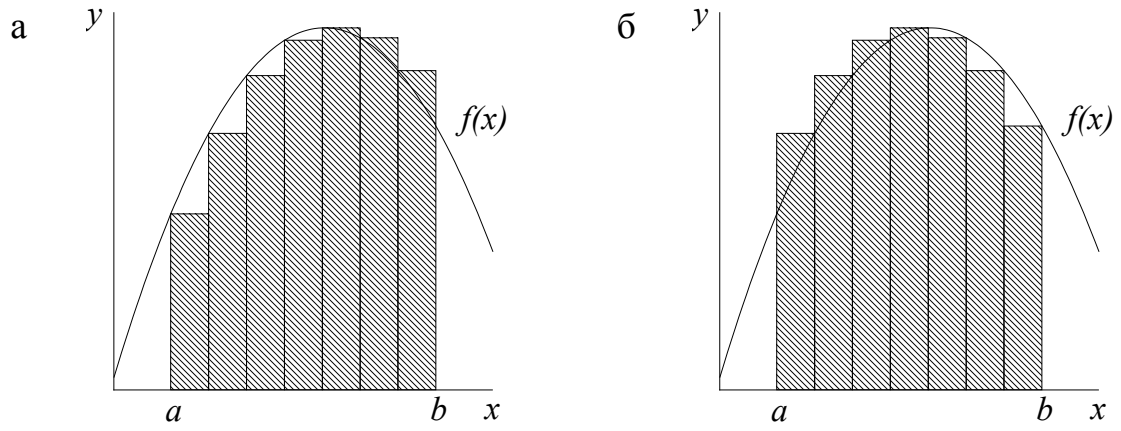


Рис. 5а, б

Подставляя ее в формулу (6.2) и заменяя для простоты $f(x_i)$ через $y_i$, получим общую формулу прямоугольников:

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} f(x_{i-1})h = (y_0 + y_1 + y_2 + \ldots + y_{n-1})h. \qquad (6.3)$$

При использовании значения $\xi_i = x_i$, то есть равное координате правой стороны прямоугольника, получаем следующую величину частичного интеграла (рис. 5б):

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx f(x_i)h.$$

Значение интеграла для всего отрезка $[a,b]$ принимает вид

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} f(x_i)h = (y_1 + y_2 + y_3 + \ldots + y_n)h. \qquad (6.4)$$

Если принять $\xi_i = x_{i-1/2}$, то есть равной координате середины отрезка $[x_{i-1}, x_i]$, то получается более точная квадратурная формула на частичном отрезке (рис. 6)–

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx f(x_{i-1/2})h,$$

и на полном отрезке $[a,b]$–

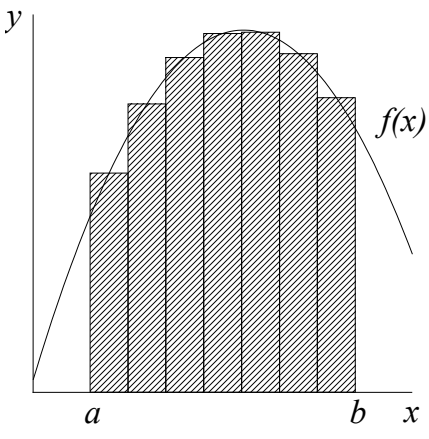


Рис. 6

$$\int_a^b f(x)dx = \sum_{i=0}^{n-1} f(x_{i+1/2})\Delta x =$$

$$= (y_{1/2} + y_{3/2} + y_{5/2} + \ldots + y_{n-1/2})h, \qquad (6.5)$$

где $y_i = f(x_i)$. Эта формула обычно называется формулой метода средних.

Оценим погрешность полученных формул приближенного вычисления интеграла. Погрешность формулы (6.3) на частичном отрезке $[x_{i-1}, x_i]$ определяется величиной

$$r_i = \int_{x_{i-1}}^{x_i} f(x)dx - f(x_{i-1})h = \int_{x_{i-1}}^{x_i} (f(x) - f(x_{i-1}))dx.$$

Заменяя функцию $f(x)$ формулой Тейлора с остаточным членом в форме Лагранжа $f(x) = f(x_{i-1}) + (x - x_{i-1})f'(\xi_i)$, $\xi_i \in [x_{i-1}, x_i]$, имеем

$$|r_i| = \left| \int_{x_{i-1}}^{x_i} (x - x_{i-1}) f'(\xi_i)dx \right| \le \frac{1}{2} h^2 M_i,$$

где $M_i = \max\limits_{x \in [x_{i-1}, x_i]} |f'(x)|$.

Суммируя частичные погрешности на элементарных отрезках, получаем общую погрешность для формулы (6.3) :

$$|R| \le \sum_{i=1}^{n} |r_i| \le \sum_{i=1}^{n} \frac{1}{2} h^2 M_i \le \frac{1}{2} nh^2 M = \frac{b-a}{2} hM,$$

то есть формула (6.3) является формулой первого порядка точности. Здесь $M = \max\limits_i M_i$. Аналогичная оценка погрешности получается и для формулы (6.4).

Погрешность формулы средних (6.5) на частичном отрезке $[x_{i-1}, x_i]$ определяется величиной

$$r_i = \int\limits_{x_{i-1}}^{x_i} f(x)dx - f(x_{i-1/2})h = \int\limits_{x_{i-1}}^{x_i} (f(x) - f(x_{i-1/2}))dx .$$

Используя, как и прежде, формулу Тейлора

$$f(x) = f(x_{i-1/2}) + (x - x_{i-1/2}) f'(x_{i-1/2}) + \frac{(x - x_{i-1/2})^2}{2} f''(\xi_i),$$

$\xi_i \in [x_{i-1}, x_i]$, получаем

$$| r_i | = \left| \int\limits_{x_{i-1}}^{x_i} (f(x) - f(x_{i-1/2}))dx \right| = \left| \int\limits_{x_{i-1}}^{x_i} \frac{(x - x_{i-1/2})^2}{2} f''(\xi_i)dx \right| \le \frac{h^3}{24} M_{2,i},$$

где $M_{2,i} = \max\limits_{x \in [x_{i-1}, x_i]} |f''(x)|$. Полная погрешность формулы (6.5) на отрезке $[a, b]$ равна

$$| R | = \left| \int\limits_a^b f(x)dx - \sum_{i=1}^n f(x_{i-1/2})h \right| \le \sum_{i=1}^n | r_i | \le \sum_{i=1}^n M_{2,i} \frac{h^3}{24} \le \frac{b-a}{24} h^2 M_2,$$

где $M_2 = \max\limits_{x \in [a,b]} |f''(x)|$, таким образом, погрешность формулы средних на $[a, b]$ равна $O(h^2)$.

## 6.2. Формула трапеций

Заменяя в частичном интеграле $\int_{x_{i-1}}^{x_i} f(x)dx$ функцию $f(x)$ линейным полиномом

$$L_{1,i} = -\frac{x - x_i}{h} f(x_{i-1}) + \frac{x - x_{i-1}}{h} f(x_i) = -\frac{x - x_i}{h} y_{i-1} + \frac{x - x_{i-1}}{h} y_i,$$

получаем формулу трапеций на частичном отрезке (рис. 7)

$$\int\limits_{x_{i-1}}^{x_i} f(x)dx \approx \int\limits_{x_{i-1}}^{x_i} \left[ \frac{x - x_{i-1}}{h} y_i - \frac{x - x_i}{h} y_{i-1} \right] dx = \frac{h}{2} (y_{i-1} + y_i). \qquad (6.6)$$

Общая формула трапеций получается суммированием частичных интегралов

$$\int_a^b f(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx \approx \sum_{i=1}^{n} \frac{h}{2}(y_{i-1} - y_i) = \frac{h}{2}(y_0 + 2y_1 + 2y_1 + \ldots + y_n). \quad (6.7)$$

Погрешность формулы (6.6) определяется выражением

$$r_i = \int_{x_{i-1}}^{x_i} f(x)dx - \frac{h}{2}(f(x_{i-1}) + f(x_i)) = \int_{x_{i-1}}^{x_i} (f(x) - L_{1,i}(x))dx.$$

Используя оценку погрешности аппроксимации функции $f(x)$ полиномом Лагранжа

$$f(x) - L_{1,i}(x) = \frac{(x - x_{i-1})(x - x_i)}{2} f''(\xi_i), \quad \xi_i \in [x_{i-1}, x_i],$$

окончательно получаем

$$r_i = \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1})(x - x_i)}{2} f''(\xi_i)dx = -\frac{h^3}{12} f''(\xi_i), \quad \xi_i \in [x_{i-1}, x_i],$$

$$| r_i | \le \frac{h^3}{12} M_{2,i}, \quad M_{2,i} = \max_{x \in [x_{i-1}, x_i]} |f''(x)|.$$

Погрешность общей формулы трапеций (6.7) оценим как сумму погрешностей на отдельных отрезках:

$$| R | \le \sum_{i=1}^{n} | r_i | \le \frac{(b-a)h^2}{12} M_2, \quad M_2 = \max_{x \in [a,b]} |f''(x)|.$$

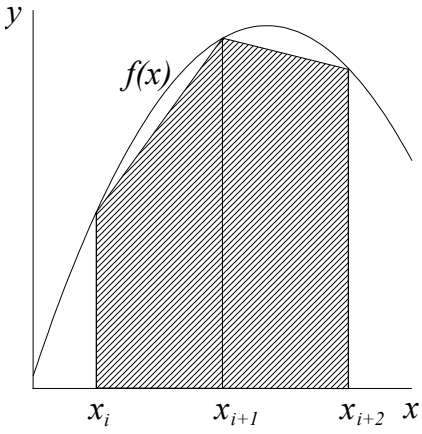


Рис. 7

Формулу трапеций можно получить также из геометрических соображений. В этом случае точки ординат $y_0, y_1, \ldots, y_n$ соединяем хордами, заменяя на каждом элементарном отрезке подынтегральную функцию $f(x)$ линейным полиномом $y = kx + d$ (рис. 7). В результате непрерывная кривая $y = f(x)$ на отрезке $[a,b]$ заменяется ломаной линией, состоящей из отдельных хорд, а определенный интеграл $\int_a^b f(x)dx$ заменяется суммой площадей получившихся трапеций.

Площадь отдельной трапеции равна произведению полусуммы оснований на высоту

$\Delta s_i = (y_{i-1} + y_i)h/2$, где $i = 1, 2, \ldots, n$, $h = x_i - x_{i-1}$, а определенный интеграл будет равен

$$\int_a^b f(x)dx = \sum_{i=1}^{n} \Delta s_i = \sum_{i=1}^{n} \frac{y_{i-1} + y_i}{2} h = \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \ldots + 2y_{n-1} + y_n) \ .$$

## 6.3. Формула Симпсона

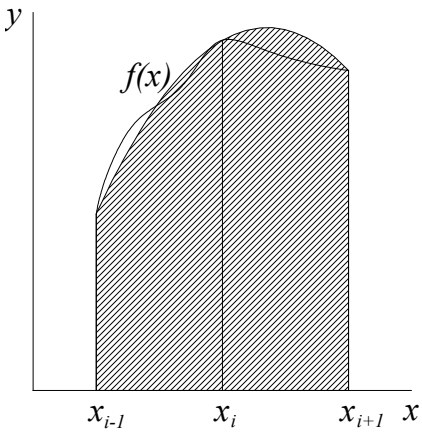


Рис. 8

Аппроксимируя в частичном интеграле функцию $f(x)$ квадратичным полиномом $L_{2,i}(x)$, получаем так называемую формулу Симпсона для частичного интервала. Поскольку квадратичный полином однозначно определяется координатами трех точек $(x_k, y_k), (k = i-1, i, i+1)$, то частичный отрезок интерполирования и интегрирования должен состоять из двух элементарных отрезков $[x_{i-1}, x_i]$ и $[x_i, x_{i+1}]$ (рис. 8), то есть быть двойным, а число $n = (b-a)/h$ четным ($h = x_i - x_{i-1}$). (Либо на каждом отрезке $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, n$ должна быть определена еще одна точка $(x_{i-1/2}, y_{i-1/2})$, где $x_{i-1/2} = (x_{i-1} + x_i)/2$). Тогда полином Лагранжа второй степени имеет вид

$$L_{2,i}(x) = \frac{1}{2h^2}\big[(x - x_i)(x - x_{i+1})f(x_{i-1}) - 2(x - x_{i-1})(x - x_{i+1})f(x_i) +$$
$$+ (x - x_i)(x - x_{i+1})f(x_{i-1})\big],$$

а частичный интеграл –

$$\int_{x_{i-1}}^{x_{i+1}} f(x)dx \approx \int_{x_{i-1}}^{x_{i+1}} L_{2,i}(x)dx = \frac{h}{3}(y_{i-1} + 4y_i + y_{i+1}) . \qquad (6.8)$$

Это формула Симпсона или формула парабол для двойного отрезка $[x_{i-1}, x_{i+1}]$.

Общая формула Симпсона для интегрирования по всему отрезку $[a, b]$ имеет вид

$$\int_a^b f(x)dx \approx \sum_{i=1,3,5,}^{n-1} \frac{h}{3}(y_{i-1} + 4y_i + y_{i+1}) =$$

$$= \frac{h}{3}\left[y_0 + 4(y_1 + y_3 + \ldots + y_{n-1}) + 2(y_2 + y_4 + \ldots + y_{n-2}) + y_n\right]. \qquad (6.9)$$

Следует отметить, что формула Симпсона (6.8) дает точное значение интеграла не только для любых полиномов второй, но и третьей степени: $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$,

$$\int_{x_{i-1}}^{x_{i+1}} y dx = \frac{h}{3}(y_{i-1} + 4y_i + y_{i+1}),$$

в чем можно убедиться непосредственной проверкой этого равенства.

Поэтому для оценки погрешности формулы Симпсона на отрезке $[x_{i-1}, x_{i+1}]$ надо воспользоваться каким-либо полиномом третьей степени, для которого формула (6.8) точна. Удобнее всего использовать для этого полином Эрмита $H_3(x)$, который, как и квадратичный полином, проводится только через три точки и удовлетворяет условиям:

$$H_3(x_{i-1}) = f(x_{i-1}); \quad H_3(x_i) = f(x_i); \quad H_3'(x_i) = f(x_i); \quad H_3(x_{i+1}) = f(x_{i+1}).$$

Погрешность аппроксимации функции $f(x)$ полиномом Эрмита на отрезке $[x_{i-1}, x_{i+1}]$ равна [3]

$$\rho_i(x) = f(x) - H_3(x) = (x - x_{i-1})(x - x_i)^2 (x - x_{i+1}) \frac{f^{(4)}(\xi_i)}{24}, \quad x \in [x_{i-1}, x_{i+1}].$$

Так как формула Симпсона точна для любого полинома третьей степени

$$\int_{x_{i-1}}^{x_{i+1}} H_3(x) dx = \frac{h}{3}\left[H_3(x_{i-1}) + 4H_3(x_i) + H_3(x_{i+1})\right] = \frac{h}{3}(y_{i-1} + 4y_i + y_{i+1}),$$

погрешность формулы Симпсона на частичном отрезке будет равна

$$r_i = \int_{x_{i-1}}^{x_{i+1}} f(x) dx - \frac{h}{3}(y_{i-1} + 4y_i + y_{i+1}) = \int_{x_{i-1}}^{x_{i+1}} \rho(x) dx =$$

$$= \int_{x_{i-1}}^{x_{i+1}} (x - x_{i-1})(x - x_i)^2 (x - x_{i+1}) \frac{f^{(4)}(\xi_i)}{24} dx.$$

Откуда получаем оценку погрешности формулы (6.8):

$$|r_i| \le \frac{h^5}{90} y^{(4)}(\xi), \quad \xi \in [x_{i-1}, x_{i+1}].$$

Погрешность общей формулы Симпсона (6.9) равна сумме погрешностей на частичных отрезках:

$$|R| \le \frac{(b-a)h^4}{180} M_4, \quad M_4 = \max_{x\in[a,b]} \left| f^{(4)}(x) \right|,$$

и имеет четвертый порядок малости по $h$, что на два порядка меньше, чем у погрешностей формул прямоугольников и трапеций.

## 6.4. Формулы интерполяционного типа

Пусть функция $f(x)$ имеет на отрезке $[a,b]$ интерполяционный полином $\varphi(x)$. Заменяя подынтегральную функцию этим полиномом $\varphi(x)$, получаем приближенное значение интеграла:

$$\int_a^b f(x)dx \approx \int_a^b \varphi(x)dx .$$

Если для $\varphi(x)$ интеграл вычисляется непосредственно, то его значение можно получить сразу.

В противном случае строится квадратурная формула интерполяционного типа. Пусть на отрезке $[a,b]$ заданы узлы интерполяции функции $f(x)$: $x_i$ и ее значения $y_i = f(x_i)$, $i = 0, 1, 2, \ldots, n$. По этим данным строим полином Лагранжа для функции $f(x)$:

$$L_n(x) = \sum_{i=0}^{n} \frac{\Pi_{n+1}(x)}{(x-x_i)\Pi'_{n+1}(x_i)} y_i ,$$

$$\Pi_{n+1}(x) = (x-x_0)(x-x_1)\ldots(x-x_n);$$

$$\Pi'_{n+1}(x_i) = (x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n).$$

Подставив в интеграл вместо функции $f(x)$ полином Лагранжа, получим

$$\int_a^b f(x)dx = \int_a^b L_n(x)dx + R_n(f),$$

где $R_n(f)$ – остаточный член (ошибка вычисления интеграла). Пренебрегая остаточным членом, получаем приближенную квадратурную формулу интерполяционного типа

$$\int_a^b ydx \approx \int_a^b \sum_{i=0}^{n} \frac{\Pi_{n+1}(x)}{(x-x_i)\Pi'_{n+1}(x_i)} y_i dx = \sum_{i=0}^{n} y_i \int_a^b \frac{\Pi_{n+1}(x)}{(x-x_i)\Pi'_{n+1}(x_i)} dx = \sum_{i=0}^{n} y_i A_i , \quad (6.10)$$

где $A_i = \int_a^b \frac{\Pi_{n+1}(x)}{(x-x_i)\Pi'_{n+1}(x_i)}dx, \quad i = 0, 1, \ldots, n$.

Если $a$ и $b$ являются узлами интерполяции, то квадратурные формулы называются формулами замкнутого типа, в противном случае – открытого.

При определении коэффициентов $A_i$ следует учитывать, что:

1) коэффициенты $A_i$ при фиксированном расположении узлов $x_i$ не зависят от вида $f(x)$;

2) для полинома степени $n$ формула точная, так как в этом случае $L_n(x) \equiv f(x)$; следовательно, формула (6.10) точная для $y(x) = x^k$, при $k = 0, 1, \ldots, n$, то есть $R_n(x^n) = 0$.

Задавая $y(x) = x^k$, $k = 0, 1, \ldots, n$, из (6.10) получим систему $(n+1)$ линейных уравнений для определения коэффициентов $A_i$:

$$\sum_{i=0}^{k} A_i x_i^k = I_k,$$

где $I_k = \int_a^b x^k dx = \frac{b^{k+1} - a^{k+1}}{k+1}$.

Для оценки погрешности квадратурной формулы интерполяционного типа (6.10) представим $f(x)$ в виде

$$f(x) = L_n(x) + \rho_{n+1}(x),$$

где $L_n(x)$ – интерполяционный полином; $\rho_{n+1}(x)$ – погрешность интерполяции $f(x)$. Подставляя это в выражение интеграла, получим

$$\int_a^b f(x)dx = \int_a^b (L_n(x) + \rho_{n+1}(x))dx = \sum_{i=0}^{n} y_i A_i + \int_a^b \rho_{n+1}(x)dx.$$

Откуда получаем погрешность квадратурной формулы

$$R(f) = \int_a^b \rho_{n+1}(x)dx.$$

Подставляя сюда выражение для погрешности интерполирования функции $f(x)$

$$\rho_{n+1}(x) = \frac{\Pi_{n+1}(x)}{(n+1)!} f^{(n+1)}(\xi), \quad \xi \in [a, b],$$

получаем

$$\left|R(f)\right| \le \frac{M_{n+1}}{(n+1)!}\int_a^b \left|\Pi_{n+1}(x)\right|dx\,, \qquad (6.11)$$

где $M_{n+1} = \max\limits_{x\in[a,b]}\left|f^{(n+1)}(x)\right|$.

Из оценки (6.11) следует, что квадратурная формула интерполяционного типа (6.10) является точной для любого многочлена степени $n$.

Рассмотренные ранее формулы прямоугольников, трапеций и Симпсона являются частными случаями формул интерполяционного типа.

## 6.5. Формулы Ньютона–Котеса

Рассмотрим частный случай равноотстоящих узлов интерполяции $x_i$, $i = 0, 1, \ldots, n$: $x_0 = a$, $x_i = x_0 + ih$, $x_n = b$, $y_i = f(x_i)$, $h = (b-a)/n$ – шаг между узлами интерполяции. В этом случае полином Лагранжа имеет вид (5.2)

$$L_n(x) = \sum_{i=0}^{n} \frac{(-1)^{n-i}}{i!(n-i)!}\frac{q^{[n+1]}}{q-i}y_i\,.$$

Заменяя подынтегральную функцию $y(x)$ данным полиномом Лагранжа, получаем приближенное значение определенного интеграла

$$\int_{x_0}^{x_n} y dx \approx \int_{x_0}^{x_n} L_n(x)dx = \sum_{i=0}^{n} A_i y_i\,, \qquad (6.12)$$

где $A_i = \int\limits_{x_0}^{x_n} \frac{(-1)^{n-1}}{i!(n-1)!}\frac{q^{[n+1]}}{q-i}dx$. Так как $q = \frac{x-x_0}{h}$, $dx = hdq$, для $A_i$ получаем

$$A_i = \frac{h(-1)^{n-i}}{i!(n-i)!}\int_0^n \frac{q^{[n+1]}}{q-i}dq, \quad i = 0,1,\ldots,n\,.$$

Поскольку $h = (b-a)/n$, обычно полагают $A_i = (b-a)H_i$, где постоянные $H_i$, равные

$$H_i = \frac{1}{n}\frac{(-1)^{n-i}}{i!(n-i)!}\int_0^n \frac{q^{[n+1]}}{q-i}dq\,, \qquad (6.13)$$

называются коэффициентами Ньютона–Котеса. Коэффициенты Ньютона–Котеса удовлетворяют следующим условиям: 1) $\sum_{i=0}^{n} H_i = 1$; 2) $H_i = H_{n-i}$.

В результате квадратурная формула (6.12) принимает вид

$$\int_a^b y dx \approx (b-a)\sum_{i=0}^{n} H_i y_i \,, \qquad (6.14)$$

где $y_i = f(a+ih)$, $i = 0, 1, \ldots, n$.

Вычисляя коэффициенты Ньютона–Котеса для *n*=1, 2, получаем квадратурные формулы трапеций и Симпсона соответственно, описанные ранее, а для случая $n = 3$ получим квадратурную формулу Ньютона (правило трех восьмых):

$$\int_{x_0}^{x_3} y dx = \frac{3h}{8}(y_0 + 3y_1 + 3y_2 + y_3),$$

остаточный член которой на частичном отрезке оказывается несколько большим, чем у формулы Симпсона:

$$r_i = -\frac{3h^5}{80} y^{(4)}(\xi).$$

Квадратурные формулы Ньютона–Котеса для больших *n* практически не применяются.

## 6.6. Квадратурная формула Гаусса

При получении квадратурной формулы Гаусса используются полиномы Лежандра, поэтому приведем их определение и основные свойства.

Полиномами Лежандра называются полиномы вида

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left[(x^2-1)^n\right], \quad n = 0, 1, 2, \ldots,$$

определенные на отрезке $[-1, 1]$. Важнейшие их свойства таковы:

1) $P_n(1) = 1, \quad P_n(-1) = (-1)^n$;

2) $\int_{-1}^{1} P_n(x) Q_k(x) dx = 0$, где $Q_k(x)$ – любой полином степени $k < n$;

3) $P_n(x)$ имеет *n* различных действительных корней, все они расположены в интервале $(-1, 1)$.

Первые полиномы Лежандра имеют следующий вид:

$$P_0(x) = 1; \quad P_1(x) = x; \quad P_2(x) = \frac{1}{2}(3x^2 - 1); \quad P_3(x) = \frac{1}{2}(5x^3 - 3x);$$

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3); \ldots.$$

Для вывода квадратурной формулы Гаусса рассмотрим функцию $y = f(t)$, определенную на отрезке $[-1, 1]$.

Нужно решить задачу: выбрать точки $t_1, t_2, \ldots, t_n$ и коэффициенты $A_1, A_2, \ldots, A_n$ так, чтобы квадратурная формула

$$\int_{-1}^{1} f(t)dt = \sum_{i=1}^{n} A_i f(t_i) \tag{6.15}$$

была точной для всех полиномов $f(t)$ наивысшей возможной степени *N*.

Так как в нашем распоряжении имеется $2n$ постоянных $t_i, A_i, i = 1, 2, \ldots, n$, наивысшая степень полинома, определяемого $2n$ коэффициентами, будет равна $N = 2n - 1$.

Для выполнения равенства (6.15) необходимо и достаточно, чтобы оно было верным при $f(t) = 1, t, t^2, \ldots, t^{2n-1}$.

Полагая

$$\int_{-1}^{1} t^k dt = \sum_{i=1}^{n} A_i t_i^k, \quad k = 0, 1, 2, \ldots, (2n-1)$$

и учитывая, что

$$\int_{-1}^{1} t^k dt = \frac{1-(-1)^{k+1}}{k+1} = \begin{cases} \dfrac{2}{k+1} & \text{при четном } k; \\ 0 & \text{при нечетном } k, \end{cases}$$

получаем систему для определения $t_i$ и $A_i$:

$$\sum_{i=1}^{n} A_i = 2; \quad \sum_{i=1}^{n} A_i t_i = 0; \ldots, \quad \sum_{i=1}^{n} A_i t_i^{2n-2} = \frac{2}{2n-1}; \quad \sum_{i=1}^{n} A_i t_i^{2n-1} = 0. \tag{6.16}$$

Для решения полученной системы применим следующий прием.

Рассмотрим полиномы $f(t) = t^k P_n(t), \quad k = 0, 1, \ldots, (n-1)$, где $P_n(t)$ – полином Лежандра. Так как степень $f(t)$ меньше $(2n-1)$, то для него верна формула (6.15):

$$\int_{-1}^{1} t^k P_n(t)dt = \sum_{i=1}^{n} t_i^k P_n(t_i) \cdot A_i.$$

В силу свойства 2) $\int_{-1}^{1} t^k P_n(t)dt = 0$. Отсюда получаем при $k = 0, 1, \ldots, (n-1)$

$$\sum_{i=1}^{n} t_i^k P_n(t_i) \cdot A_i = 0.$$

Эти равенства справедливы для любых $A_i$, если положить $P_n(t_i)=0$, $i=1, 2, \ldots, n$. Следовательно, $t_i$ – корни полинома Лежандра. Зная $t_i$, из первых *n* уравнений линейной системы (6.16) можно найти коэффициенты $A_i$.

Формула (6.15), в которой $t_i$ – нули полинома Лежандра $P_n(t_i)$, коэффициенты $A_i$ находятся из системы (6.16), называется квадратурной формулой Гаусса. Она обладает высокой точностью при небольшом числе ординат.

Для вычисления интеграла при произвольных пределах интегрирования $\int_a^b f(x)dx$, необходимо сделать замену переменных $x=(b+a)/2+t(b-a)/2$, чтобы свести его к стандартной форме

$$\int_a^b f(x)dx=\frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b+a}{2}+\frac{b-a}{2}t\right)dt.$$

Применяя затем квадратурную формулу Гаусса, получим

$$\int_a^b f(x)dx=\frac{b-a}{2}\sum_{i=1}^{n} A_i f(x_i),\quad x_i=\frac{b+a}{2}+\frac{b-a}{2}t_i,\quad i=1, 2, \ldots, n,$$

где $t_i$ – корни полинома Лежандра $P_n(t_i)=0$.

Остаточный член формулы Гаусса с *n* узлами:

$$R_n=\frac{(b-a)^{2n+1}(n!)^4 f^{(2n)}(\xi)}{((2n)!)^3(2n+1)},\quad \xi\in[a,b],$$

откуда для различных *n* получаем:

$$R_2=\frac{1}{135}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi);$$

$$R_3=\frac{1}{15750}\left(\frac{b-a}{2}\right)^7 f^{(6)}(\xi);$$

$$R_4=\frac{1}{3472875}\left(\frac{b-a}{2}\right)^9 f^{(8)}(\xi),\ \text{и т. д.}$$

## 6.7. Экстраполяция по Ричардсону

Пусть точное значение интеграла есть приближенное плюс остаток:

$$I=I_h+R(h).$$

Предположим, что порядок остатка известен: $R(h) = Mh^m$. Например, для метода трапеций $m = 2$, а для метода Симпсона $m = 4$.

Выберем два различных шага интегрирования: $h_1 = (b-a)/n_1$, $h_2 = (b-a)/n_2$, где $n_1, n_2$ – число отрезков разбиения $[a,b]$. Тогда точное значение интеграла в этих случаях будет равно

$$I = I_{h_1} + Mh_1^m, \quad I = I_{h_2} + Mh_2^m.$$

Вычитая первое значение интеграла из второго, получим

$$I_{h_2} - I_{h_1} = M(h_1^m - h_2^m) \quad \Rightarrow \quad M = \frac{I_{h_2} - I_{h_1}}{h_1^m - h_2^m}.$$

Тогда уточненное значение интеграла можно записать в виде

$$I = I_{h_2} + Mh_2^m + O(h^{m+1}) = I_{h_2} + \frac{I_{h_2} - I_{h_1}}{h_1^m - h_2^m} h_2^m + O(h^{m+1}).$$

Предполагая, что для погрешности вычисления интеграла справедливо выражение

$$I - I_h = M_1 h_1^{m_1} + M_2 h_2^{m_2} + \ldots M_k h^{m_k} + O(h^{k+1}),$$

где $k = 1, 2, \ldots$, $0 < m_1 < m_2 < \ldots < m_k$, можно получить алгоритм уточнения интеграла, который также называют методом экстраполяции Ричардсона. Пусть шаг $h$ при каждом последующем вычислении уменьшается вдвое: $h_i = h_{i-1}/2, \quad i = 1, 2, \ldots, k$. Полагаем вначале $I_h^{(0)} = I_h$. Последующие приближения интеграла вычисляем по рекуррентной формуле

$$I_h^{(i)} = I_{h/2}^{(i-1)} + \frac{I_{h/2}^{(i-1)} - I_h^{(i-1)}}{h^{k_i} - (h/2)^{k_i}} (h/2)^{k_i} = I_{h/2}^{(i-1)} + \frac{I_{h/2}^{(i-1)} - I_h^{(i-1)}}{2^k - 1}, \quad i = 1, 2, \ldots.$$

При применении рассмотренного алгоритма экстраполяции Ричардсона к формуле трапеций, получается так называемый метод Ромберга, который часто используется в пакетах прикладных программ.

## 7. СИСТЕМЫ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Системы линейных алгебраических уравнений находят самое широкое распространение в науке и технике. К ним сводятся многие практические задачи. Поэтому решение линейных алгебраических уравнений является одной из важнейших задач вычислительной математики.

Система *n* уравнений с *n* неизвестными называется линейной, если неизвестные входят в нее только в первой степени, например:

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1;$$

$$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2;$$

. . . . . .

$$a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \qquad (7.1)$$

Здесь $a_{ij}$ – коэффициенты уравнений; $x_i$ – неизвестные; $b_i$ – свободные члены уравнений, $i, j = 1, 2, \ldots, n$. Записывая коэффициенты этой системы $a_{ij}$ в виде прямоугольной таблицы, а неизвестные $x_i$ и свободные члены $b_i$ – в виде столбцов

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}; \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}; \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

систему уравнений (7.1) можно представить в матричном виде:

$$Ax = b. \qquad (7.2)$$

Здесь $A$ – матрица коэффициентов системы; $x$ – вектор-столбец неизвестных; $b$ – вектор-столбец свободных членов.

Так как алгебра матриц широко используется в теории линейных уравнений, рассмотрим ее основные понятия.

## 7.1. Основные понятия алгебры матриц и линейной алгебры

Матрицей называется совокупность $m \times n$ чисел, расположенных в виде прямоугольной таблицы из $m$ строк и $n$ столбцов

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}.$$

Матрица $A$, имеющая $m$ строк и $n$ столбцов, называется $(m \times n)$-матрицей, или матрицей размера $m \times n$. Если $m = n$, то матрицу называют квадратной матрицей порядка $n$. Прямоугольную $(m \times n)$-матрицу кратко обозначают как $A = [a_{ij}]_{mn}$.

Матрица размера $1\times n$ называется вектором-строкой, а матрица размера $m\times 1$ – вектором-столбцом. Квадратная матрица вида

$$A=\begin{bmatrix}\alpha_1 & 0 & \dots & 0\\ 0 & \alpha_2 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & \alpha_n\end{bmatrix}$$

называется диагональной и обозначается кратко: $A=\operatorname{diag}(\alpha_1,\alpha_2,\dots,\alpha_n)$. Если $\alpha_i=1,\ i=1,2,\dots,n$, то матрица называется единичной и обычно обозначается буквой $E$:

$$E=\begin{bmatrix}1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & 1\end{bmatrix}.$$

Матрица называется нулевой и обозначается 0 или $0_{mn}$, если все ее элементы равны нулю. Матрица называется верхней треугольной, если все элементы ее ниже главной диагонали равны нулю, и нижней треугольной, если ее элементы выше главной диагонали нулевые. Главной диагональю квадратной матрицы называется линия, проведенная от левого верхнего ее угла к нижнему правому, а линия, проведенная из правого верхнего угла к нижнему левому, называется побочной диагональю.

Матрица называется матрицей ленточного типа, если ее ненулевые элементы располагаются только на нескольких диагоналях, смежных с главной диагональю.

Каждой квадратной матрице $A$ соответствует определитель (детерминант) $\det A$, вычисляющийся по формуле

$$\det A=\sum_{(\alpha_1,\alpha_2,\dots,\alpha_n)}(-1)^{\kappa}a_{1\alpha_1}a_{2\alpha_2}\dots a_{n\alpha_n},\qquad(7.3)$$

где суммирование распространено на все возможные перестановки вторых индексов $\alpha_i$ элементов $1,2,\dots,n$ и содержит $n!$ слагаемых, причем $\kappa=0$, если перестановка четная, и $\kappa=1$, если перестановка нечетная. В общем случае определители вычисляются не по формуле (7.3), а каким-либо другим способом, например, методом Гаусса.

В развернутой форме определитель записывается путем замены скобок в развернутой записи соответствующей матрицы:

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

### 7.1.1. Действия с матрицами

Матрицы $A$ и $B$ считаются равными $A = B$, если они одного размера и соответствующие элементы их равны, то есть $a_{ij} = b_{ij}$. В этом случае имеет смысл и операция сравнения $A \geq B$, которая означает $a_{ij} \geq b_{ij}$ для $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n$. Если матрицы $A$ и $B$ одного размера, то матрицы можно складывать и вычитать:

$$C = A \pm B = \begin{bmatrix} a_{11} \pm b_{11} & a_{12} \pm b_{12} & \dots & a_{1n} \pm b_{1n} \\ \dots & \dots & \dots & \dots \\ a_{m1} \pm b_{m1} & a_{m2} \pm b_{m2} & \dots & a_{mn} \pm b_{mn} \end{bmatrix}.$$

Сумма матриц удовлетворяет следующим свойствам:

1) $A + (B + C) = (A + B) + C$;
2) $A + B = B + A$;
3) $A + 0 = A$.

Матрицу $A$ можно умножить на число $\alpha$. Элементы результирующей матрицы $B$ есть результат умножения соответствующих элементов матрицы $A$ на число $\alpha$:

$$B = A\alpha = \alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \dots & \alpha a_{1n} \\ \dots & \dots & \dots & \dots \\ \alpha a_{m1} & \alpha a_{m2} & \dots & \alpha a_{mn} \end{bmatrix}.$$

Произведение матрицы на число обладает свойствами:

1) $1A = A$;
2) $0A = 0$;
3) $\alpha(\beta A) = (\alpha\beta)A$;
4) $(\alpha + \beta)A = \alpha A + \beta A$;
5) $\alpha(A + B) = \alpha A + \alpha B$.

Матрица $-A = (-1)A$ называется противоположной матрице $A$. Если $A$ – квадратная матрица порядка $n$, то определитель матрицы $C = \alpha A$ равен

$$\det C = \det \alpha A = \alpha^n \det A .$$

Для матриц определена операция умножения. Пусть $A$ и $B$ – матрицы размеров $m \times n$ и $p \times q$ соответственно. Если число столбцов $n$ матрицы $A$ равно числу строк $p$ матрицы $B$, то для этих матриц существует матрица $C$ размера $m \times q$, являющаяся их произведением:

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1q} \\ c_{21} & c_{22} & \dots & c_{2q} \\ \dots & \dots & \dots & \dots \\ c_{m1} & c_{m2} & \dots & c_{mq} \end{bmatrix}, \text{ где } c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \; i = 1, 2, \dots, m; \; j = 1, 2, \dots, q,$$

то есть элемент матрицы $c_{ij}$ равен сумме произведений элементов $i$-й строки матрицы $A$ на соответствующие элементы $j$-го столбца матрицы $B$.

Свойства произведения:

1) $A(BC) = (AB)C$;
2) $\alpha(AB) = (\alpha A)B$;
3) $(A + B)C = AC + BC$;
4) $C(A + B) = CA + CB$.

В общем случае $AB \neq BA$ (свойство некоммутативности). Только квадратные матрицы одного порядка могут быть коммутативными, но даже среди них коммутативностью обладают только немногие пары.

*Пример*.

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}; \; B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}; \; AB = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}; \; BA = \begin{bmatrix} 23 & 34 \\ 31 & 46 \end{bmatrix},$$

то есть $AB \neq BA$.

Единичная матрица $E$ порядка $n$ коммутирует со всеми квадратными матрицами порядка $n$: $AE = EA = A$, и таким образом играет роль единицы для операции умножения.

Если $A$ и $B$–квадратные матрицы одного порядка, то

$$\det(AB) = \det(BA) = \det A \times \det B .$$

Это следует из правила умножения матриц.

Если в матрице $A$ поменять строки на столбцы, то есть зеркально отразить матрицу относительно главной диагонали, то получим так называемую транспонированную матрицу, которую обозначают как $A^T$:

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}; \quad A^T=\begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix}.$$

Если транспонировать вектор-строку $a=[a_1,a_2,\dots,a_n]$, то получим вектор-столбец $a^T=[a_1,a_2,\dots,a_n]^T=\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$, и наоборот.

Свойства $A^T$:

1) $A^{TT}=(A^T)^T=A$;

2) $(A+B)^T=A^T+B^T$;

3) $(AB)^T=B^TA^T$.

Если $A$ – квадратная матрица, то $\det A^T=\det A$.

Квадратная матрица $A=[a_{ij}]$ называется симметричной, если совпадает со своей транспонированной: $A=A^T$, то есть ее элементы, симметричные относительно главной диагонали, равны: $a_{ij}=a_{ji}$. Произведение матрицы на транспонированную $C=AA^T$ – симметричная матрица:

$$C^T=(AA^T)^T=(A^T)^TA^T=AA^T=C.$$

Матрица $A^{-1}$ называется обратной матрице $A$, если $A^{-1}A=AA^{-1}=E$. Вычисление обратной матрицы называется обращением матрицы. Квадратная матрица называется неособенной, если ее определитель не равен нулю. В противном случае матрица особенная, или сингулярная. Всякая неособенная матрица имеет обратную матрицу.

Основные свойства обратной матрицы:

1) $(AB)^{-1}=B^{-1}A^{-1}$;

2) определитель матрицы, обратной к $A$, равен величине, обратной к определителю матрицы $A$:

$$\det A^{-1}=1/\det A;$$

3) транспонированная обратная матрица равна обратной к транспонированной:

$$(A^{-1})^T=(A^T)^{-1};$$

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

Степенью матрицы $A^n$ называется $n$-кратное произведение матрицы $A$ на себя: $A^n = AA\ldots A$; если $\det A \neq 0$, то

$$A^{-n} = (A^{-1})^n; \quad A^0 = E; \quad A^p \cdot A^q = A^{p+q}; \quad (A^p)^q = A^{p \cdot q}.$$

Определитель матрицы, состоящий из $k$ строк и $k$ столбцов, называется минором $k$-го порядка матрицы $A$. Рангом матрицы называется максимальный порядок минора матрицы, отличного от нуля. Матрица $A$ имет ранг $r$, если найдется хотя бы один ее минор $r$-го порядка, отличный от нуля, а миноры порядка $(r+1)$ и выше все равны нулю.

Дефектом матрицы $A$ размера $m \times n$ называется разность между наименьшим из чисел $m$ и $n$ и рангом матрицы.

### 7.1.2. Нормы матриц и векторов

Для оценок сходимости различных методов решения уравнений используются понятия норм матрицы и вектора. Нормой матрицы $A$ называется положительное число $\|A\|$, удовлетворяющее условиям:

1) $\|A\| \geq 0$, причем $\|A\| = 0$ при условии $A = 0$;

2) $\|\alpha A\| = \alpha \|A\|$, где $\alpha$ – положительное число;

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

4) $\|A - B\| \geq \|A\| - \|B\|$;

5) $\|AB\| \leq \|A\| \cdot \|B\|$. (7.4)

В качестве норм используются три основные нормы:

1) $\|A\|_m = \max\limits_i \sum\limits_{j=1}^{n} |a_{ij}|$ – максимальная сумма модулей элементов одной строки);

2) $\|A\|_l = \max\limits_j \sum\limits_{i=1}^{m} |a_{ij}|$ – максимальная сумма модулей элементов одного столбца);

3) $\|A\|_k = \left(\sum\limits_{i,j} a_{ij}^2\right)^{1/2}$.

Норма вектора $x$ есть положительное число $\|x\|$, удовлетворяющее условиям, аналогичным (7.4). Поскольку векторы используются совместно с матрицами, норма вектора должна быть согласована с нормой матрицы. Применяются следующие нормы векторов $x = [x_1, x_2, \ldots, x_n]^T$:

1) $\|x\|_m = \max\limits_i |x_i|$;

2) $\|x\|_l = \sum\limits_{i=1}^{n} |x_i|$;

3) $\|x\|_k = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$.

Пределом матрицы $A_k$ называется матрица $A$, равная $A = \lim\limits_{k\to\infty} A_k$, при этом $\lim\limits_{k\to\infty}\|A - A_k\| = 0$ для любой нормы. Для этого необходимо и достаточно, чтобы $\lim\limits_{k\to\infty} a_{ij}^{(k)} = a_{ij}$, $i, j = 1, 2, \ldots, n$.

Пользуясь понятием предела матрицы, можно определить предельную сумму матричного ряда:

$$\sum_{k=1}^{\infty} A_k = \lim_{N\to\infty} \sum_{k=1}^{N} A_k ,$$

для сходимости суммы матричного ряда достаточно, чтобы сходилась сумма ряда некоторой матричной нормы

$$\sum_{k=1}^{\infty} \|A_k\|.$$

## 7.2. Решение систем линейных уравнений

Система линейных алгебраических уравнений (7.2) $Ax = b$ называется однородной, если вектор $b$ равен нулю, в противном случае – неоднородной. Решением системы называют такое значение вектора $x$, которое обращает ее в тождество. Система неоднородных уравнений (7.2) называется совместной, если существует хотя бы одно решение $x$, и несовместной, если ни одного решения не существует.

*Признак совместности*. Система совместна, если ранг матрицы $A$ равен рангу расширенной матрицы $B$:

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}; \qquad B = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & b_n \end{bmatrix}.$$

Совместная система называется определенной, если она имеет единственное решение, и неопределенной, если решений бесконечное множество.

Необходимым и достаточным условием существования единственного решения системы (7.2) является условие неравенства нулю ее определителя $\det A \neq 0$. В случае $\det A = 0$ система линейных уравнений (7.2) либо не имеет решения, либо имеет их бесконечное множество. Если $\det A \approx 0$, система называется плохо обусловленной, поскольку погрешности определения ко

эффициентов матрицы $A$ приводят к значительным изменениям решения системы.

Система однородных уравнений

$$Ax = 0 \qquad (7.5)$$

всегда имеет нулевое решение $x_i = 0, \quad i = 1, 2, \ldots, n$. Для существования ненулевого решения необходимо, чтобы определитель системы был равен нулю $\det A = 0$.

Если система (7.5) имеет хотя бы одно ненулевое решение $x$, то она имеет бесконечное число таких решений, поскольку все векторы, отличающиеся от $x$ коэффициентом $\alpha$, также являются решениями. Совокупность линейно независимых решений $x^{(i)}$ системы называется фундаментальной системой. Любое решение системы записывается как линейная комбинация фундаментальной системы решений: $x = \alpha_i x^{(i)}$. Число $k$ решений системы, образующих фундаментальную систему, равно разности порядка матрицы системы $n$ и ее ранга $r$: $k = n - r$.

### 7.2.1. Методы решения линейных систем

Если известна обратная матрица $A^{-1}$ матрицы системы, то решение системы уравнений получается умножением слева системы на обратную матрицу:

$$A^{-1}Ax = A^{-1}b \quad \Rightarrow \quad x = A^{-1}b.$$

Однако обратные матрицы бывают известны очень редко, а их вычисление требует усилий, больших, чем решение самой системы. Поэтому этот способ решения имеет лишь теоретическое значение.

Методы решения систем линейных уравнений в основном делятся на две группы: 1) прямые методы, использующие конечные алгоритмы (формулы), и 2) итерационные методы, требующие для получения решения применения сходящихся бесконечных вычислительных процессов.

К прямым методам относятся: правило Крамера, методы Гаусса, главных элементов, квадратных корней и др. Достоинствами этих методов являются простота и универсальность, то есть применимость их для решения многих классов линейных систем.

В то же время эти методы имеют недостатки: необходимость большого объема памяти ЭВМ для хранения матриц, накапливание погрешностей в процессе решения, большое число арифметических операций. Поэтому эти методы применяются обычно для небольших хорошо обусловленных систем уравнений с числом уравнений $n < 1000$.

К итерационным методам относятся методы простой итерации (Якоби), Гаусса–Зейделя, релаксации и др. Эти методы последовательных приближений требуют задания некоторого начального приближения. Затем посредством некоторого алгоритма решение последовательно уточняется. Итерации продолжаются до получения решения с заранее заданной точностью. Итерационые методы иногда используются для уточнения решения, полученного прямыми методами.

### 7.2.2. Правило Крамера

Решение системы уравнений $Ax = b$ осуществляется по формуле

$$x_j = \frac{|A_j|}{|A|}, \quad j = 1, 2, \ldots, n .$$

Здесь $|A| = \det A$, $|A_j| = \det A_j$, $A_j$ – это матрица, полученная из матрицы $A$, в которой $j$-й столбец коэффициентов заменен столбцом свободных членов – вектором $b$, например:

$$A_2 = \begin{bmatrix} a_{11} & b_1 & a_{13} & \ldots & a_{1n} \\ a_{21} & b_2 & a_{23} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & b_n & a_{n3} & \ldots & a_{nn} \end{bmatrix}.$$

Правило Крамера имеет исключительно теоретическое значение, поскольку требует вычисления $(n+1)$ определителей, то есть огромного числа арифметических операций $N$:

$$N = (n+1)(n \cdot n! - 1) + n ,$$

где $n$ – порядок системы.

### 7.2.3. Метод исключения Гаусса

Метод исключения Гаусса чаще других прямых методов применяется для решения систем уравнений. Метод основан на сведении матрицы системы уравнений к треугольному виду посредством преобразований ее строк, а затем решении полученной системы уравнений, и, таким образом, состоит из двух этапов.

Первый этап, называемый *прямым ходом*, для системы (7.2)

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 ;$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2;$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_{3n}x_n = b_3;$$

. . . . . .

$$a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n,$$

состоит из $(n-1)$ шагов исключения.

На первом шаге первое уравнение системы используется для исключения неизвестного $x_1$ из всех последующих уравнений. Для этого каждое уравнение с номером $i = 2, 3, \ldots, n$ складывается с первым уравнением, умноженным на коэффициент $-a_{i1}/a_{11}$, при этом в $i$-м уравнении коэффициент при $x_1$ сокращается, а само уравнение приобретает следующий вид:

$$a_{i2}^{(1)}x_2 + a_{i3}^{(1)}x_3 + \ldots + a_{in}^{(1)}x_n = b_i^{(1)};$$

$$b_i^{(1)} = b_i - b_1\frac{a_{i1}}{a_{11}}, \quad a_{im}^{(1)} = a_{im} - a_{1m}\frac{a_{i1}}{a_{11}}, \quad m = 2, 3, \ldots, n. \qquad (7.6)$$

Коэффициент $a_{11}$ при неизвестном $x_1$ в первом уравнении в данном контексте называется главным, или ведущим, элементом первого шага исключения. Главный элемент не должен быть равен 0, иначе при делении на него произойдет авост.

В результате система приводится к эквивалентному виду:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1;$$

$$a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)};$$

$$a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n = b_3^{(1)};$$

. . . . . .

$$a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)}.$$

Теперь уравнения с номерами $i = 2, 3, \ldots, n$ образуют систему $(n-1)$ уравнений для $(n-1)$ неизвестных, и можно повторить шаг исключения уже для этой системы меньшего размера.

На втором шаге исключается неизвестное $x_2$ из уравнений с номерами $i = 3, \ldots, n$ путем сложения этих уравнений со вторым уравнением, умноженным на коэффициент $-a_{i2}^{(1)}/a_{22}^{(1)}$, в результате чего система приобретает следующий вид:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1;$$

$$a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)};$$

$$a_{33}^{(2)}x_3 + \ldots + a_{3n}^{(2)}x_n = b_3^{(2)};$$

. . . . . .

$$a_{n3}^{(2)}x_3 + \ldots + a_{nn}^{(2)}x_n = b_n^{(2)}.$$

Аналогично, на *k*-м шаге исключается неизвестное $x_k$ из уравнений с номерами $i = (k+1), \ldots, n$ сложением этих уравнений с *k*-м уравнением, умноженным на коэффициент $-a_{ik}^{(k-1)} / a_{kk}^{(k-1)}$, где $a_{kk}^{(k-1)}$ – ненулевой главный элемент на *k*-м шаге исключения. В результате исключения *i*-е уравнение получает следующий вид:

$$a_{ik+1}^{(k)}x_{k+1} + a_{ik+2}^{(k)}x_{k+2} + \ldots + a_{in}^{(k)}x_n = b_i^{(k)};$$

$$b_i^{(k)} = b_i^{(k-1)} - b_k^{(k-1)}\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}},\ a_{im}^{(k)} = a_{im}^{(k-1)} - a_{km}^{(k-1)}\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}},\ m = (k+1), \ldots, n. \quad (7.7)$$

После $(n-1)$, последнего шага исключения, матрица системы уравнений становится верхней треугольной:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1;$$

$$a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)};$$

$$a_{33}^{(2)}x_3 + \ldots + a_{3n}^{(2)}x_n = b_3^{(2)};$$

. . . . . .

$$a_{nn}^{(n-1)}x_n = b_n^{(n-1)}. \quad (7.8)$$

На втором этапе, этапе *обратного хода*, вычисляются неизвестные в порядке, обратном порядку их исключения. Из последнего уравнения системы (7.8) вычисляется значение $x_n = b_n^{(n-1)} / a_{nn}^{(n-1)}$. Подстановкой найденного значения в $(n-1)$-е уравнение определяется $x_{n-1}$, и далее, подстановкой в *k*-е уравнение вычисленных неизвестных $x_{k+1}, \ldots, x_n$ рассчитывается $x_k$:

$$x_k = (b_k^{k-1} - a_{k,k+1}^{(k-1)}x_{k+1} - \ldots - a_{k,n}^{(k-1)}x_n) / a_{kk}^{(k-1)}, \quad k = (n-1), \ldots, 1.$$

Так как при исключении неизвестных из системы выполняются операции деления на главные элементы $a_{11}$, $a_{22}^{(1)}$, $a_{33}^{(2)}$, …, $a_{nn}^{(n-1)}$, необходимо, чтобы в процессе исключения они не становились нулевыми. Если какой-либо главный элемент окажется равным нулю, систему следует преобразовать таким образом, чтобы на месте этого коэффициента оказался другой, не равный нулю коэффициент. Этого можно достичь перестановкой уравнений (строк матрицы) или изменением порядка исключения неизвестных, что эквивалентно перестановке столбцов матрицы. Такие преобразования не оказывают влияния на вектор $x$ – решение системы.

Число арифметических операций в методе Гаусса для системы $n$ уравнений составляет $N \approx 2n^3/3$.

### 7.2.4. Метод Гаусса с выбором главного элемента

Деление коэффициентов $k$-го уравнения на главный элемент $a_{kk}^{(k-1)}$ $k$-го шага исключения метода Гаусса в формулах (7.7) приводит к большим вычислительным погрешностям, если главный элемент близок к нулю. Действительно, пусть на $k$-м шаге исключения все коэффициенты матрицы вычислены с абсолютными погрешностями одного порядка. В этом случае погрешность формул (7.7) можно оценить по формуле для погрешности частного (2.2), в которой следует оставить только зависимость от погрешности главного элемента $\Delta a_{kk}^{(k-1)}$:

$$\Delta b_i^{(k)} \approx \left| b_k^{(k-1)} \frac{a_{ik}^{(k-1)}}{(a_{kk}^{(k-1)})^2} \right| \Delta a_{kk}^{(k-1)}; \quad \Delta a_{im}^{(k)} \approx \left| a_{km}^{(k-1)} \frac{a_{ik}^{(k-1)}}{(a_{kk}^{(k-1)})^2} \right| \Delta a_{kk}^{(k-1)}.$$

Это выражение показывает, что малый главный элемент $a_{kk}^{(k-1)}$ в знаменателе ($a_{kk}^{(k-1)} << a_{ik}^{(k-1)}, a_{km}^{(k-1)}, b_i^{(k-1)}$) приводит к большому множителю при $\Delta a_{kk}^{(k-1)}$ и, как следствие, к катастрофическому росту ошибки всех коэффициентов системы.

Во избежание роста ошибок, обусловленных малостью главного элемента, на каждом шаге исключения главный элемент следует выбирать достаточно большим по абсолютной величине. При этом, чем больше главный элемент, тем медленнее рост ошибки. Существуют различные стратегии выбора главного элемента. Например, в качестве главного элемента можно выбрать максимальный по модулю элемент текущего столбца, что равносильно перестановке уравнений в системе, и этот способ называется методом с выбором главного элемента по столбцу. В другом способе выбора главного элемента ищется максимальный по модулю элемент текущей строки, что

приводит к изменению порядка исключения неизвестных. Такой метод называется методом с выбором главного элемента по строке.

Рассмотрим метод исключения Гаусса с полным выбором главного элемента, в котором поиск производится по всем элементам матрицы. Поскольку выбор производится по большему числу элементов, ошибка этого метода оказывается наименьшей.

Пусть задана система уравнений $Ax = b$ порядка $n$ с матрицей $A$:

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

Найдем в ней максимальный по модулю элемент $a_{ij}$, находящийся на пересечении $i$-й строки и $j$-го столбца, и переставим местами 1-е уравнение системы с $i$-м уравнением, а также 1-й столбец неизвестных с $j$-м столбцом, то есть поменяем нумерацию неизвестных $x_j$ на $x_1$, а $x_1$ на $x_j$. В результате получим эквивалентную систему уравнений с новой матрицей $A'$, в которой элемент $a'_{11}$ будет максимальным.

Далее обычным методом Гаусса при помощи первого уравнения исключается неизвестное $x_1$ из второго, третьего и последующих уравнений системы, и получается система, в которой матрица коэффициентов будет иметь нулевые значения в первом столбце везде, кроме первой строки. Отбрасывая первое уравнение, получаем систему, содержащую на 1 уравнение меньше, с матрицей $A^{(1)}$ порядка $(n-1)$.

Повторяя с этой системой те же операции, что и с исходной, то есть выбирая главный элемент, переставляя на первое место уравнение, его содержащее, и соответствующий столбец с неизвестными, переобозначая неизвестные и исключая из последующих уравнений неизвестную $x_2$ при помощи первого уравнения, получаем новую матрицу $A^{(2)}$, в которой первый столбец содержит нули везде, кроме первой строки.

Таким образом, в алгоритм решения Гаусса на каждом шаге исключения добавляется еще процесс выбора главного элемента и соответствующей трансформации системы.

Продолжая этот процесс далее, получаем в результате систему с верхней треугольной матрицей (7.8), решение которой производится обычным обратным ходом метода Гаусса, изложенным выше. После этого в полученном решении восстанавливается исходный порядок нумерации неизвестных.

### 7.2.5. Метод прогонки

Рассмотрим применение метода исключения Гаусса для решения линейной системы *n* уравнений с трехдиагональной матрицей:

$$b_1 x_1 + c_1 x_2 = d_1;$$

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, \quad i = 2, 3, \ldots, (n-1);$$

$$a_n x_{n-1} + b_n x_n = d_n.$$

Здесь $a_i, b_i, c_i$ – коэффициенты уравнений; $x_i$ – неизвестные; $d_i$ – свободные члены.

Разделим для этого первое уравнение системы на $b_1$ и представим его в виде

$$x_1 = K_1 x_2 + L_1, \quad K_1 = -\frac{c_1}{b_1}, \quad L_1 = \frac{d_1}{b_1}.$$

Затем при его помощи исключим $x_1$ из второго уравнения:

$$x_2 = K_2 x_3 + L_2, \quad K_2 = -\frac{c_2}{b_2 + a_2 K_1}, \quad L_1 = \frac{d_2 - a_2 L_1}{b_2 + a_2 K_1}.$$

Продолжая далее процесс исключения, сведем исходную систему к виду с двухдиагональной матрицей:

$$x_1 - K_1 x_2 = L_1;$$

$$x_2 - K_2 x_3 = L_3;$$

$$x_i - K_i x_{i+1} = L_i, \quad i = 3, 4, \ldots, (n-2);$$

$$x_{n-1} - K_{n-1} x_n = L_{n-1};$$

$$x_n = L_n,$$

где коэффициенты $K_i, L_i$ вычисляются последовательно от $i = 1$ до $i = n$ по рекуррентным формулам:

$$K_1 = -\frac{c_1}{b_1}; \quad L_1 = \frac{d_1}{b_1};$$

$$K_i = -\frac{c_i}{b_i + a_i K_{i-1}}; \quad L_i = \frac{d_i - a_i L_{i-1}}{b_i + a_i K_{i-1}}; \quad i = 2, 3, \ldots, (n-1);$$

$$L_n = \frac{d_n - a_n L_{n-1}}{b_n + a_n K_{n-1}}.$$

На этом заканчивается прямой ход метода Гаусса.

Обратный ход начинается с вычисления значения $x_n$ из последнего уравнения системы с двухдиагональной матрицей

$$x_n = L_n.$$

Последующие компоненты вектора $x$ вычисляются по рекуррентным формулам

$$x_i = K_i x_{i+1} + L_i, \quad i = (n-1), (n-2), \ldots, 1.$$

В литературе изложенное решение системы уравнений с трехдиагональной матрицей получило название метода прогонки. Более того, шел даже спор о приоритете его предложения, хотя из изложенного видно, что это просто частный случай метода исключения Гаусса без выбора главного элемента.

Условием устойчивости метода прогонки являются общие требования, предъявляемые к методу Гаусса без выбора главного элемента, то есть требование диагонального преобладания матрицы коэффициентов $|b_i| > |a_i| + |c_i|$.

### 7.3. Вычисление определителя методом Гаусса

Вычисление определителя матрицы по формуле его определения (7.3) – очень трудоемкая операция. Проще всего вычисляется определитель треугольной матрицы. Он равен произведению ее диагональных элементов, так как все другие входящие в определение определителя произведения элементов матрицы, расположенных в разных строках и столбцах, равны нулю, поскольку включают хотя бы один элемент матрицы, равный нулю:

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{nn} \end{vmatrix} = \sum_{(\alpha_1,\ldots,\alpha_n)} a_{1\alpha_1} a_{2\alpha_2} \ldots a_{n\alpha_n} = a_{11} a_{22} \ldots a_{nn}.$$

Для приведения матрицы $A$ к треугольному виду можно использовать метод исключения Гаусса. В процессе исключения элементов величина определителя не меняется, а потому его величина равна

$$\det A = (-1)^{\gamma} a_{11} a_{22}^{(1)} \ldots a_{nn}^{(n-1)},$$

где $\gamma$ – число перестановок строк и столбцов, сделаных при прямом ходе исключения. Если число перестановок строк и столбцов четно, то знак определителя совпадает со знаком произведения главных элементов исключения, в противном случае знак меняется на противоположный.

## 7.4. Вычисление обратной матрицы методом Гаусса

Матрица $A^{-1}$, обратная к матрице $A$, удовлетворяет условию

$$AA^{-1} = E, \tag{7.9}$$

где $E$ – единичная матрица.

Рассматривая столбцы матриц как векторы и используя правило умножения матриц (строка на столбец), равенство (7.9) можно записать в виде совокупности систем линейных уравнений. Пусть

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}; \quad A^{-1} = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1n} \\ z_{21} & z_{22} & \dots & z_{2n} \\ \dots & \dots & \dots & \dots \\ z_{n1} & z_{n2} & \dots & z_{nn} \end{bmatrix}; \quad E = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix},$$

тогда получаем системы уравнений

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} z_{11} \\ z_{21} \\ \dots \\ z_{n1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}; \qquad \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} z_{12} \\ z_{22} \\ \dots \\ z_{n2} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \dots \\ 0 \end{bmatrix};$$

. . . . . .,

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} z_{1n} \\ z_{2n} \\ \dots \\ z_{nn} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 1 \end{bmatrix},$$

или в более компактном виде

$$Az^{(1)} = \overline{\delta}_{i1}; \quad Az^{(2)} = \overline{\delta}_{i2}; \quad \dots; \quad Az^{(n)} = \overline{\delta}_{in}, \tag{7.10}$$

где $z^{(j)} = \begin{bmatrix} z_{1j} \\ z_{2j} \\ \vdots \\ z_{nj} \end{bmatrix}$; $\overline{\delta}_{ij} = \begin{bmatrix} \delta_{1j} \\ \delta_{2j} \\ \vdots \\ \delta_{nj} \end{bmatrix}$; $\delta_{ij} = \begin{cases} 1 & \text{при } i = j; \\ 0 & \text{при } i \neq j; \end{cases}$ $i, j = 1, 2, \ldots, n.$

Таким образом, решая систему уравнений

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} z_{11} \\ z_{21} \\ \ldots \\ z_{n1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \ldots \\ 0 \end{bmatrix}$$

методом Гаусса, можно определить элементы первого столбца обратной матрицы $A^{-1}$. Для определения всех ее элементов необходимо решить $n$ систем такого вида, то есть все системы уравнений (7.10). Задача облегчается тем, что прямой ход для всех систем делается один раз, поскольку все системы имеют одну и ту же матрицу. Для каждой отдельной системы делается только обратный ход после некоторых преобразований ее свободных членов (7.10).

Этот метод обращения матрицы экономичен. Он требует примерно в три раза больше арифметических операций, чем при решении одной системы уравнений.

## 7.5. Метод Гаусса и *LU*-разложение матрицы

Рассмотрим, какими именно преобразованиями матрицы $A$ сопровождается метод Гаусса решения системы (7.2) $Ax = b$.

На первом шаге прямого хода метода Гаусса первая строка матрицы $A$ и элемент $b_1$ вектора правой части (7.2), умноженные на соответствующие коэффициенты, вычитаются из всех остальных строк и элементов вектора правой части. Это действие можно осуществить умножением $A$ и $b$ на матрицу $L_1$:

$$L_1 = \begin{bmatrix} l_{11} & 0 & 0 & \ldots & 0 \\ l_{21} & 1 & 0 & \ldots & 0 \\ l_{31} & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ l_{n1} & 0 & 0 & \ldots & 1 \end{bmatrix},$$

коэффициенты 1-го столбца которой имеют вид $l_{i1} = -a_{i1} / a_{11},\ i = 2, 3, \ldots, n$. Диагональный коэффициент $l_{11} = 1/a_{11}$ выбирается таким, чтобы диагональный коэффициент преобразованной матрицы был единичным

$$A^{(1)} = L_1 A = \begin{bmatrix} 1 & a_{12}/a_{11} & a_{13}/a_{11} & \ldots & a_{1n}/a_{11} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \ldots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \ldots & a_{3n}^{(1)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \ldots & a_{nn}^{(1)} \end{bmatrix}, \quad b^{(1)} = L_1 b = \begin{bmatrix} b_1/a_{11} \\ b_2^{(1)} \\ b_3^{(1)} \\ \ldots \\ b_n^{(1)} \end{bmatrix},$$

где коэффициенты $a_{ik}^{(1)}$, $b_i^{(1)}$ определяются формулами (7.6).

Аналогично, *k*-й шаг прямого хода исключения производится умножением на матрицу, называемую элементарной нижней треугольной матрицей

$$L_k = \begin{bmatrix} 1 & \ldots & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & l_{kk} & 0 & \ldots & 0 \\ 0 & \ldots & l_{k+1,k} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & l_{nk} & 0 & \ldots & 1 \end{bmatrix}.$$

В матрице такого типа все элементы главной диагонали, кроме $l_{kk}$, равны единице, недиагональные элементы, за исключением элементов *k*-го столбца ниже диагонали, равны нулю. В матрице $L_k$ коэффициенты *k*-го столбца выбираются из аналогичных соображений, что и в матрице $L_1$: матрица $L_k$ должна исключать неизвестное $x_k$ из уравнений с номерами $(k+1), (k+2), \ldots, n$: $l_{kk} = 1/a_{kk}^{(k-1)}$, $l_{ki} = -a_{ki}^{(k-1)} / a_{kk}^{(k-1)}$.

После *n*-го шага исключения, где матрица $L_n$ имеет вид $L_n = \mathrm{diag}(1,1,\ldots,1/a_{nn}^{(n-1)})$, матрица $A^{(n)}$ принимает вид верхней треугольной матрицы с единичной диагональю, которую обычно обозначают через *U*:

$$A^{(n)} = U = L_n \ldots L_2 L_1 A = \begin{bmatrix} 1 & a_{12}/a_{11} & a_{12}/a_{11} & \ldots & a_{12}/a_{11} \\ 0 & 1 & a_{23}^{(1)}/a_{22}^{(1)} & \ldots & a_{2n}^{(1)}/a_{22}^{(1)} \\ 0 & 0 & 1 & \ldots & a_{3n}^{(2)}/a_{33}^{(2)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}.$$

а сама система (7.2) записывается в виде

$$Ux = b^{(n)}, \quad b^{(n)} = L_n \dots L_2 L_1 b, \tag{7.11}$$

с верхней треугольной матрицей с единичной главной диагональю. Если систему (7.11) умножить слева на произведение обратных матриц $L = L_1^{-1} L_2^{-1} \dots L_n^{-1}$, которое есть тоже нижняя треугольная матрица, то система приводится к исходному виду:

$$LUx = b,$$

где $LU = A$ – представление матрицы $A$ в виде произведения нижней треугольной и верхней треугольной матриц называется *LU-разложением*.

Таким образом, *LU*-разложение матрицы $A$ можно получить при помощи элементарных треугольных матриц. Сначала строятся матрицы $L_1, L_2, \dots L_n$, и вычисляется матрица $U = L_n \dots L_2 L_1 A$, а затем находится $L = L_1^{-1} L_2^{-1} \dots L_n^{-1}$. Учитывая конкретный вид обратных матриц $L_k^{-1}$:

$$L_k^{-1} = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & l_{kk}^{-1} & 0 & \dots & 0 \\ 0 & \dots & -l_{k+1,k} / l_{kk} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & -l_{nk} / l_{kk} & 0 & \dots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & a_{kk}^{(k-1)} & 0 & \dots & 0 \\ 0 & \dots & a_{k+1,k}^{(k-1)} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & a_{nk}^{(k-1)} & 0 & \dots & 1 \end{bmatrix},$$

получаем матрицу $L$:

$$L = \begin{bmatrix} a_{11} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22}^{(1)} & 0 & \dots & 0 \\ a_{31} & a_{32}^{(1)} & a_{33}^{(2)} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2}^{(1)} & a_{n3}^{(2)} & \dots & a_{nn}^{(n-1)} \end{bmatrix},$$

При известном *LU*-разложении матрицы $A$ решение системы $Ax = b$ производится в два этапа: 1) находится решение $y$ системы с нижней треугольной матрицей $Ly = b$; 2) находится решение $x$ системы с верхней треугольной матрицей $Ux = y$. Вследствие специального вида матриц $L$ и $U$ решение этих систем не представляет трудности.

В методе Гаусса приведение матрицы $A$ к виду с верхней треугольной формой и решение системы $Ly = b$ выполняется одновременно при прямом ходе. Затем обратным ходом решается система $Ux = y$, и находится вектор $x$.

## 7.6. Теорема об *LU*-разложении

Пусть $\Delta_j$ – угловой минор порядка $j$ матрицы $A$, определенный как

$$\Delta_j = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1j} \\ a_{21} & a_{22} & \dots & a_{2j} \\ \dots & \dots & \dots & \dots \\ a_{j1} & a_{j2} & \dots & a_{jj} \end{vmatrix}.$$

Справедлива *теорема*. Если все угловые миноры матрицы $A$ отличны от нуля, $\Delta_j \neq 0, \quad j = 1, 2, \dots, m$, то матрицу $A$ можно представить единственным образом в виде произведения $A = LU$, где $L$ – нижняя треугольная матрица с ненулевыми диагональными элементами, а $U$ – верхняя треугольная матрица с единичной диагональю.

Доказательство проведем методом индукции. Покажем, что теорема справедлива для матриц второго порядка

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

Представим $A$ в виде произведения $A = LU$:

$$A = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} 1 & u_{12} \\ 0 & 1 \end{bmatrix},$$

где $l_{11}, l_{21}, l_{22}, u_{12}$ – неизвестные числа. После перемножения матриц получим систему уравнений для определения этих неизвестных:

$$l_{11} = a_{11}; \quad l_{11}u_{12} = a_{12}; \quad l_{21} = a_{21}; \quad l_{21}u_{12} + l_{22} = a_{22}.$$

Эта система имеет единственное решение:

$$l_{11} = a_{11}; \quad u_{12} = a_{12} / a_{11}; \quad l_{21} = a_{21}; \quad l_{22} = \frac{a_{11}a_{22} - a_{21}a_{12}}{a_{11}},$$

поскольку из условия теоремы $a_{11} \neq 0$; $a_{11}a_{22} \neq a_{21}a_{12}$. Предположим теперь, что теорема справедлива для матриц порядка $(k-1)$. Докажем, что она справедлива и для матриц порядка $k$. Представим матрицу $A$ порядка $k$ в клеточном виде:

$$A = \left[\begin{array}{ccc|c} a_{11} & \dots & a_{1,k-1} & a_{1k} \\ \dots & \dots & \dots & \dots \\ a_{k-1,1} & \dots & a_{k-1,k-1} & a_{k-1,k} \\ \hline a_{k,1} & \dots & a_{k,k-1} & a_{k,k} \end{array}\right]. \qquad (7.12)$$

Введем обозначения:

$$A_{k-1} = \begin{bmatrix} a_{11} & \dots & a_{1,k-1} \\ \dots & \dots & \dots \\ a_{k-1,1} & \dots & a_{k-1,k-1} \end{bmatrix}; \quad a_{k-1} = \begin{bmatrix} a_{1k} \\ \dots \\ a_{k-1,k} \end{bmatrix};$$

$$r_{k-1} = \left(a_{k,1}, \quad \dots, \quad a_{k,k-1}\right).$$

По предположению индукции существует *LU*-разложение матрицы $A_{k-1}$: $A_{k-1} = L_{k-1}U_{k-1}$, где $L_{k-1}, U_{k-1}$ – указанные треугольные матрицы. Разложение матрицы (7.12) ищем в виде

$$A = \begin{bmatrix} L_{k-1} & 0 \\ l_{k-1} & l_{kk} \end{bmatrix} \begin{bmatrix} U_{k-1} & u_{k-1} \\ 0 & 1 \end{bmatrix}, \tag{7.13}$$

где $l_{k-1}, u_{k-1}$ – неизвестные векторы:

$$l_{k-1} = (l_{k1}, l_{k2}, \dots, l_{k,k-1}); \quad u_{k-1} = (u_{1k}, u_{2k}, \dots, u_{k-1,k})^T.$$

Перемножая матрицы уравнения (7.13) и учитывая (7.12), получаем систему уравнений:

$$L_{k-1}u_{k-1} = a_{k-1};$$

$$l_{k-1}U_{k-1} = r_{k-1};$$

$$l_{k-1}u_{k-1} + l_{kk} = a_{kk}.$$

Отсюда получаем значения неизвестных векторов, так как $L_{k-1}^{-1}$ и $U_{k-1}^{-1}$ по предположению индукции существуют:

$$u_{k-1} = L_{k-1}^{-1}a_{k-1}; \quad l_{k-1} = r_{k-1}U_{k-1}^{-1}; \quad l_{kk} = a_{kk} - l_{k-1}u_{k-1}.$$

Таким образом, *LU*-разложение матрицы *A* порядка *k* существует, так как

$$\det A = (\det L_{k-1})l_{kk}(\det U_{k-1}) = (\det L_{k-1})l_{kk},$$

а по условию $\det A \neq 0$, следовательно, $l_{kk} \neq 0$.

Более того, это разложение единственно. Действительно, предположим, что существует два вида *LU*-разложения матрицы *A*: $A = L_1U_1 = L_2U_2$. Тогда, умножив слева это уравнение на $L_1^{-1}$, а справа на $U_2^{-1}$, получим уравнение

$$U_1U_2^{-1} = L_1^{-1}L_2. \tag{7.14}$$

Матрица в левой части уравнения (7.14) – верхняя треугольная, а в правой части – нижняя треугольная. Такое равенство возможно, если $U_1U_2^{-1}$ и $L_1^{-1}L_2$ диагональны. Но на диагонали $U_1U_2^{-1}$ стоят единицы, следовательно, эти матрицы единичны и $U_1 = U_2$, $L_1 = L_2$, то есть разложение $LU$ единственно.

## 7.7. Метод Холецкого (метод квадратного корня)

Если матрица $A$ системы линейных алгебраических уравнений (7.2) симметрична ($A^T = A$) и положительно определена, то ее можно представить в виде произведения двух транспонированных между собой матриц:

$$A = T^T T,$$

$$T = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ 0 & t_{22} & \dots & t_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & t_{nn} \end{bmatrix}, \quad T^T = \begin{bmatrix} t_{11} & 0 & \dots & 0 \\ t_{12} & t_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ t_{1n} & t_{2n} & \dots & t_{nn} \end{bmatrix},$$

и для ее решения применить более экономный метод Холецкого, учитывающий специфику матрицы $A$.

Матрица $A$ называется положительно определенной, если квадратичная форма

$$u = (Ax, x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j > 0$$

является положительно определенной для любого вектора $x \neq 0$.

Если собственные значения симметричной матрицы положительны, то матрица положительно определена.

Действительная симметричная матрица $A$ является положительно определенной, если все главные диагональные миноры ее больше нуля.

Линейные системы с симметричными матрицами часто встречаются при решении технических задач, например: в задачах механики твердого тела, теории упругости, колебаний, оптимизации и др. Поэтому рассмотрение метода решения таких задач является актуальным.

Построение матриц $T^T$ и $T$ производится следующим образом. Перемножив матрицы $T^T$ и $T$ и приравняв получившиеся выражения элементов матрицы $T^T T$ к значениям элементов матрицы $A$, получим следующие уравнения для определения элементов матрицы $T$:

$$t_{1i}t_{1j} + t_{2i}t_{2j} + \dots + t_{ii}t_{ij} = a_{ij} \quad (i < j);$$

$$t_{1i}^2 + t_{2i}^2 + \ldots + t_{ii}^2 = a_{ii}\,.$$

Разрешая эту систему, находим:

$$t_{11} = \sqrt{a_{11}}\,, \quad t_{1j} = \frac{a_{1j}}{t_{11}} \quad (j > 1);$$

$$t_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} t_{ki}^2} \qquad (1 < i \le n);$$

$$t_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} t_{ki} t_{kj}}{t_{ii}} \qquad (i < j);$$

$$t_{ij} = 0 \text{ при } i > j\,.$$

Этот метод получил название метода квадратного корня потому, что при вычислении диагональных элементов используются операции извлечения корня.

Система (7.2) имеет единственное решение, так как в этом случае $\det A \neq 0$.

Если разложение $A = T^T T$ получено, то решение системы (7.2) сводится к решению двух систем с треугольными матрицами $T^T y = b$ и $Tx = y$, которые решаются очень просто и требуют $N \approx 2n^2$ арифметических операций (здесь $n$ – порядок матрицы $A$).

Общее число арифметических операций при решении системы $Ax = b$ методом Холецкого почти в два раза меньше, чем при применении метода Гаусса, и составляет $N \approx n^3/3 + 2n^2$, тогда как для метода Гаусса требуется $N \approx 2n^3/3$ операций. Кроме того, симметричность матрицы $A$ позволяет хранить в памяти ЭВМ только половину матрицы.

## 7.8. *QR*-разложение матрицы

В настоящее время, кроме метода Гаусса, известно много прямых методов решения систем линейных алгебраических уравнений. Большинство этих методов основано на переходе от исходной системы $Ax = b$ к новой системе $Bx = d$, решаемой проще исходной. Этот переход производится путем умножения исходной системы на некоторую матрицу $C$, которая выбирается из условий, чтобы эта матрица вычислялась не слишком сложно, а само умножение не сильно портило систему, то есть не сильно изменяло ее число обусловленности.

Этим условиям удовлетворяют методы вращений и отражений. Оба метода позволяют получить представление матрицы $A$ в виде произведения ортогональной матрицы $Q$ на верхнюю треугольную матрицу $R$:

$$A = QR,$$

и считаются одними из наиболее устойчивых к вычислительной погрешности.

*Определение*: Матрица $Q$ называется ортогональной, если для нее выполняется условие $Q^T = Q^{-1}$, то есть $QQ^T = E$.

### 7.8.1. Метод вращений

В этом методе матрица $C$, приводящая исходную систему $n$ уравнений $Ax = b$ к системе с верхней треугольной матрицей $CAx = Cb \Rightarrow Bx = d$, где $B = CA;\ d = Cb$, получается последовательным обнулением элементов, лежащих ниже главной диагонали с помощью матриц элементарных вращений $T_{ij}$ Гивенса:

$$T_{ij} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & t_{ii} & 0 & \dots & t_{ij} & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & t_{ji} & 0 & \dots & t_{jj} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 \end{bmatrix}.$$

Они отличаются от единичной матрицы $E$ только четырьмя элементами: $t_{ii} = t_{jj} = \cos\varphi_{ij};\ t_{ij} = -t_{ji} = \sin\varphi_{ij}$.

Умножение вектора $x$ на матрицу $T_{ij}$ геометрически можно интерпретировать как поворот в плоскости $Ox_i x_j$ $n$-мерного пространства, что и дало название методу. Матрица $T_{ij}$ удовлетворяет условию ортогональности $T_{ij}^T = T_{ij}^{-1}$.

На первом шаге прямого хода, получения матрицы $B$, исключается переменная $x_1$ из второго и последующих уравнений исходной системы. Это делается посредством умножения слева системы $Ax = b$ вначале на матрицу вращения

$$T_{12} = \begin{bmatrix} t_{11} & t_{12} & 0 & \dots & 0 \\ t_{21} & t_{22} & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix},$$

элементы которой вычисляются по формулам:

$$t_{11} = \frac{a_{11}}{\sqrt{a_{11}^2 + a_{21}^2}}; \quad t_{12} = \frac{a_{21}}{\sqrt{a_{11}^2 + a_{21}^2}}; \quad t_{21} = -t_{12}; \quad t_{22} = t_{11}, \tag{7.15}$$

где $a_{ij}$ – коэффициенты матрицы $A$. Коэффициенты $t_{ij}$ удовлетворяют условиям:

$$t_{11}^2 + t_{12}^2 = 1; \quad t_{21}a_{11} + t_{11}a_{21} = 0. \tag{7.16}$$

В результате получается система, второе уравнение которой не содержит неизвестное $x_1$:

$$\begin{gathered} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \dots + a_{1n}^{(1)}x_n = b_1^{(1)}; \\ a_{22}^{(1)}x_2 + \dots + a_{2n}^{(1)}x_n = b_2^{(1)}; \\ a_{31}x_1 + a_{32}x_2 + \dots + a_{3n}x_n = b_3; \\ \dots\dots\dots\dots\dots\dots\dots\dots; \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = b_n. \end{gathered} \tag{7.17}$$

Новые коэффициенты первых двух уравнений вычисляются по правилам:

$$a_{1j}^{(1)} = t_{11}a_{1j} + t_{12}a_{2j}; \quad a_{2j}^{(1)} = t_{21}a_{1j} + t_{11}a_{2j} \quad (1 \le j \le n);$$

$$b_1^{(1)} = t_{11}b_1 + t_{12}b_2; \quad b_2^{(1)} = t_{21}b_1 + t_{11}b_2; \quad a_{21}^{(1)} = 0$$

согласно (7.16).

Если в исходной системе $a_{21} = 0$, считается $t_{11} = 1, \quad t_{12} = 0$, и матрица вращений становится равной единичной $T_{12} = E$.

Для исключения $x_1$ из третьего уравнения система (7.17) умножается слева на матрицу вращения $T_{13}$

$$T_{13} = \begin{bmatrix} t_{11} & 0 & t_{13} & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ t_{31} & 0 & t_{33} & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

Коэффициенты $T_{13}$ вычисляются по формулам:

$$t_{11} = \frac{a_{11}^{(1)}}{\sqrt{(a_{11}^{(1)})^2 + a_{31}^2}}; \quad t_{12} = \frac{a_{31}}{\sqrt{(a_{11}^{(1)})^2 + a_{31}^2}}; \quad t_{31} = -t_{13}; \quad t_{33} = t_{11}, \qquad (7.18)$$

и удовлетворяют условиям $t_{11}^2 + t_{13}^2 = 1; \quad t_{31}a_{11}^{(1)} + t_{11}a_{31} = 0$.

Аналогичным образом исключается $x_1$ из всех последующих уравнений. В итоге получается система следующего вида:

$$a_{11}^{(n-1)}x_1 + a_{12}^{(n-1)}x_2 + a_{13}^{(n-1)}x_3 + \dots + a_{1n}^{(n-1)}x_n = b_1^{(n-1)};$$

$$a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \dots + a_{2n}^{(1)}x_n = b_2^{(1)};$$

$$\dots\dots\dots\dots\dots\dots\dots\dots;$$

$$a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \dots + a_{nn}^{(1)}x_n = b_n^{(1)},$$

или в матричной форме

$$A^{(1)}x = b^{(1)},$$

где $A^{(1)} = T_{1n}T_{1n-1}\dots T_{13}T_{12}A; \quad b^{(1)} = T_{1n}T_{1n-1}\dots T_{13}T_{12}b$. Здесь $T_{ij}$ – матрицы элементарных преобразований вращения, элементы которых вычисляются по правилам типа (7.15), (7.18) и удовлетворяют условиям $t_{ii}^2 + t_{ij}^2 = 1$.

На втором шаге метода, состоящем из $(n-2)$ “малых” шагов, аналогичным образом исключается $x_2$ из третьего и последующих уравнений системы $A^{(1)}x = b^{(1)}$. В матричной форме получается

$$A^{(2)}x = b^{(2)},$$

где $A^{(2)} = T_{2n}\dots T_{24}T_{23}A^{(1)}; \quad b^{(2)} = T_{2n}\dots T_{24}T_{23}b^{(1)}$.

После завершения $(n-1)$-го шага система принимает вид

$$a_{11}^{(n-1)}x_1 + a_{12}^{(n-1)}x_2 + a_{13}^{(n-1)}x_3 + \ldots + a_{1n}^{(n-1)}x_n = b_1^{(n-1)};$$

$$a_{22}^{(n-1)}x_2 + a_{23}^{(n-1)}x_3 + \ldots + a_{2n}^{(n-1)}x_n = b_2^{(n-1)};$$

$$a_{33}^{(n-1)}x_3 + \ldots + a_{3n}^{(n-1)}x_n = b_3^{(n-1)};$$

$$\ldots\ldots\ldots\ldots\ldots\ldots;$$

$$a_{nn}^{(n-1)}x_n = b_n^{(n-1)},$$

или в матричной форме $A^{(n-1)}x = b^{(n-1)}$, где $A^{(n-1)} = T_{n-1,n}A^{n-2}$; $b^{(n-1)} = T_{n-1,n}b^{n-2}$.

Верхняя треугольная матрица $A^{(n-1)}$, которая обычно обозначается $R$, связана с исходной матрицей равенством $R = TA$, где $T$ – матрица результирующего вращения,

$$T = T_{n-1,n}\ldots T_{3n}\ldots T_{34}\ldots T_{2n}\ldots T_{23}\ldots T_{1n}\ldots T_{12}.$$

Матрица $T$ ортогональна, так как является произведением ортогональных матриц $T_{ij}$. Обозначая $Q = T^{-1} = T^T$, получаем

$$T^{-1}R = T^{-1}TA \quad \Rightarrow \quad A = QR,$$

то есть $QR$-разложение матрицы $A$.

Обратный ход метода вращения проводится так же, как и в методе Гаусса, то есть решается вначале система $Qy = b$, а затем система $Rx = y$.

Этот метод обладает существенной численной устойчивостью, однако более трудоемок в сравнении с методом Гаусса. Получение матриц $QR$-разложения для квадратной матрицы $A$ порядка $n$ общего вида требует около $2n^3$ арифметических операций.

### 7.8.2. Метод отражений

В этом методе $QR$-разложение матрицы $A$ системы уравнений $Ax = b$ производится при помощи матриц отражения, которые имеют вид

$$U = E - 2ww^T.$$

Здесь $E$ – единичная матрица; $w$ – $n$-мерный вектор единичной длины, $(w, w) = 1$, а $ww^T$ – квадратная симметричная матрица:

$$ww^T = \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_n \end{bmatrix} [w_1, w_2, \dots, w_n] = \begin{bmatrix} w_1^2 & w_1 w_2 & \dots & w_1 w_n \\ w_1 w_2 & w_2^2 & \dots & w_2 w_n \\ \dots & \dots & \dots & \dots \\ w_1 w_n & w_2 w_n & \dots & w_n^2 \end{bmatrix}.$$

Следовательно, и матрица отражения $U$ является симметричной, $U = U^T$. Более того, матрица $U$ ортогональна, то есть $U^T = U^{-1}$, действительно:

$$UU^T = (E - 2ww^T)(E - 2ww^T)^T = E - 2ww^T - 2ww^T + 4ww^T ww^T = E,$$

поскольку $w^T w = (w, w) = 1$.

Так как $U$ симметрична и ортогональна, $U^2 = UU^T = E$, собственные числа матрицы $U$ удовлетворяют условию $\lambda_U^2 = 1$, то есть $\lambda_U = \pm 1$, причем отрицательному собственному значению $\lambda = -1$ соответствует собственный вектор $w$:

$$Uw = \lambda w = Ew - 2ww^T w = w - 2w = -w = (-1)w \quad \rightarrow \quad \lambda = -1.$$

Положительному собственному значению $\lambda_U = 1$ соответствуют все векторы, ортогональные вектору $w$. Если произвольный вектор $v$ ортогонален вектору $w$, то есть $(v, w) = 0$, то

$$Uv = Ev - 2ww^T v = v - 2w(w, v) = v.$$

Рассмотрим действие матрицы $U$ на произвольный вектор $y$, который представим в виде суммы двух ортогональных компонент: $y = z + v$. Компонента $z$ направлена вдоль вектора $w$ и является проекцией $y$ на $w$: $z = dw$, $d = (y, w)$, а компонента $v$ ортогональна этому вектору: $(v, w) = 0$, и равна $v = y - (y, w)w$. Тогда

$$Uy = U(z + v) = E(z + v) - 2ww^T (z + v) = z + v - 2ww^T z - 2ww^T v =$$

$$= z + v - 2ww^T dw = z + v - 2z = -z + v.$$

Таким образом, вектор $Uy$ является зеркальным отражением вектора $y$ относительно плоскости, ортогональной вектору $w$.

Используя это свойство матрицы отражения, можно подобрать вектор $w$ таким, чтобы исходный вектор $y \neq 0$ в результате отражения $Uy$ получил направление некоторого заданного единичного вектора $e$. В результате отражения получается вектор

$$Uy = \alpha e \text{ или } Uy = -\alpha e, \quad \alpha = \sqrt{(y, y)}, \tag{7.19}$$

поскольку при ортогональных преобразованиях длины векторов сохраняются ($U$ – ортогональная матрица). Направление, перпендикулярное к плоскости отражения, будет определяться вектором ($y - \alpha e$) или вектором ($y + \alpha e$).

Таким образом, векторы

$$w_1 = \pm\frac{y - \alpha e}{\rho_1} \text{ или } w_1 = \pm\frac{y + \alpha e}{\rho_2}, \qquad (7.20)$$

где $\rho_1 = \sqrt{(y - \alpha e, y - \alpha e)}$; $\rho_2 = \sqrt{(y + \alpha e, y + \alpha e)}$, являются требуемыми компонентами матрицы отражения. Если векторы $y$ и $e$ параллельны, то отражения делать не надо (при этом $\rho_1$ или $\rho_2$ будут равны нулю).

Покажем, что произвольную квадратную матрицу можно представить в виде произведения ортогональной и правой (верхней) треугольной матриц.

Пусть $A$ – квадратная матрица порядка $n$. Приведем ее к правой треугольной форме путем последовательного умножения слева на ортогональные матрицы отражения. На первом шаге приведения в качестве вектора $y$ рассмотрим первый столбец матрицы $A$:

$$y_1 = [a_{11}, a_{21}, \ldots, a_{n1}]^T.$$

Если $a_{i1} = 0 \quad (i = 2, 3, \ldots n)$, следует перейти к следующему шагу приведения, положив $A^{(1)} = A$; $U_1 = E$ и обозначив $a_{ij}^{(1)} = a_{ij}$. В противном случае умножим матрицу $A$ слева на матрицу отражения $U_1 = E_n - 2w_1 w_1^T$, где $w_1$ подбирается таким образом, чтобы вектор $U_1 y_1$ стал параллелен вектору $e_1 = [1, 0, 0, \ldots, 0]^T$, то есть в соответствии с формулами (7.19), (7.20). Здесь $E_n$ – единичная матрица порядка $n$, а $e_1$, $y_1$ – $n$–мерные векторы. В результате такого преобразования в первом столбце матрицы $A^{(1)}$ все элементы, кроме первого, станут равными нулю.

На втором шаге приведения преобразуемым вектором является второй столбец матрицы $A^{(1)}$ без первого члена

$$y_2 = [a_{22}^{(1)}, a_{32}^{(1)}, \ldots, a_{n2}^{(1)}]^T.$$

Преобразование отражения выполняется умножением матрицу $A^{(1)}$ слева на матрицу

$$U_2 = \begin{bmatrix} 1 & 0 \\ 0 & S_{n-1} \end{bmatrix},$$

где $S_{n-1} = E_{n-1} - 2w_{n-1}w_{n-1}^T$, а $w_{n-1}$ – $(n-1)$-мерный вектор, вычисляющийся по формулам (7.19), (7.20). Тем самым обнуляются элементы второго столбца, расположенные ниже главной диагонали матрицы $A^{(2)} = U_2 A^{(1)}$.

Последующие шаги процесса приведения матрицы $A$ проводятся аналогично. После выполнения $k$-го шага получается матрица $A^{(k)}$, все элементы которой, находящиеся ниже главной диагонали вплоть до $k$-го столбца матрицы, равны нулю: $a_{ij}^{(k)} = 0$ при $i > j$, $j = 1, 2, \ldots, k$.

Для выполнения $(k+1)$-го шага приведения преобразуем вектор

$$y_{k+1} = [a_{k+1,k+1}^{(k)}, a_{k+2,k+1}^{(k)}, \ldots, a_{n,k+1}^{(k)}]^T.$$

Если компоненты вектора $y_{k+1}$ $a_{i,k+1}^{(k)} = 0$ (для $i = (k+2), (k+3), \ldots, n)$), получаем $A^{(k+1)} = A^{(k)}$; $U_{k+1} = E_n$ и переходим к следующему шагу. В противном случае строим матрицу отражения

$$S_{k+1} = E_{n-k} - 2w_{k+1}w_{k+1}^T$$

(вектор $w_{k+1}$ и матрица $S_{k+1}$ порядка $(n-k)$) для преобразования вектора $y_{k+1}$ в вектор, параллельный вектору $e_{k+1} = [1,0,0,\ldots,0]^T$ (длины $(n-k)$), и переходим от матрицы $A^{(k)}$ к матрице $A^{(k+1)}$:

$$A^{(k+1)} = U_{k+1}A^{(k)}, \text{ где } U_{k+1} = \begin{bmatrix} E_k & 0 \\ 0 & S_{k+1} \end{bmatrix}.$$

Процесс этот всегда осуществим, и после $(n-1)$-го шага приходим к матрице

$$A^{(n-1)} = U_{n-1}U_{n-2}\ldots U_1 A,$$

имеющей правую треугольную форму. Обозначив через $U$ произведение матриц вращения $U = U_{n-1}U_{n-2}\ldots U_1$, это выражение можно записать в виде $A^{(n-1)} = UA$, или $A = QR$, где $Q = U^T$ – ортогональная матрица, а $R = A^{(n-1)}$ – правая треугольная матрица.

Решение системы $Ax = b$ посредством метода отражения выполняется следующим образом. Умножив систему слева на последовательность матриц отражения, сводим ее к виду с верхней треугольной матрицей

$$UAx = Ub \quad \Rightarrow \quad Rx = Ub, \text{ или } A^{(n-1)}x = Ub.$$

Если все диагональные элементы матрицы $A^{(n-1)}$ отличны от нуля, то неизвестные $x_i$ для $i = n, (n-1), \ldots, 1$ находятся, как и в методе Гаусса, обычным обратным ходом. Если же хотя бы один из диагональных элементов матрицы равен нулю, то система $A^{(n-1)}x = Ub$ вырождена, и в силу эквивалентности вырождена и исходная система.

Этот метод в настоящее время считается одним из наиболее устойчивых к вычислительной погрешности, но более трудоемок в сравнении с мето

дом Гаусса. Для получения *QR*-разложения квадратной матрицы *A* порядка *n* общего вида требуется около $(4/3)n^3$ арифметических операций.

## 7.9. Обусловленность систем линейных алгебраических уравнений

### 7.9.1. Устойчивость системы линейных алгебраических уравнений

Математическая задача называется корректной, если ее решение существует и единственно, и если оно непрерывно зависит от входных данных.

Корректность исходной математической задачи еще не гарантирует хороших свойств численного метода ее решения. Поэтому свойства используемых методов решения корректных задач должны изучаться отдельно.

Рассмотрим вопросы корректности исходной задачи и численных методов ее решения на примере системы линейных алгебраических уравнений

$$Ax = b, \tag{7.2}$$

где *A* – квадратная матрица порядка *n*, *x* – вектор неизвестных; *b* – вектор свободных членов (порядка *n*). Для существования единственного решения уравнений (7.2) необходимо, чтобы $\det A \neq 0$, тогда существует обратная матрица $A^{-1}$ и решение можно записать в виде

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

Для установления корректности задачи (7.2) необходимо установить еще непрерывную зависимость решения от входных данных.

Входными данными здесь являются правая часть *b* и элементы матрицы *A*.

Различают устойчивость по правой части, когда возмущается вектор *b*, а матрица *A* остается неизменной, и коэффициентную устойчивость, когда возмущается только *A*, а вектор *b* остается неизменным.

Для выполнения этих оценок будем использовать понятие нормы. Предполагая, что решение и правая часть системы (7.2) принадлежат линейному *n*-мерному пространству *H,* будем считать нормой матрицы *A*, подчиненной данной норме вектора, число

$$\|A\| = \sup \frac{\|Ax\|}{\|x\|}, \quad (0 \neq x \in H).$$

Рассмотрим ”возмущенную систему”, отличающуюся от (7.2) правой частью

$$A\tilde{x} = \tilde{b}. \tag{7.21}$$

Полагаем, что в $A$ возмущения не вносятся. Введем обозначения для погрешностей решения $x$ и правой части $b$

$$\delta x = \widetilde{x} - x, \quad \delta b = \widetilde{b} - b,$$

где $x,\ b$ – точные, а $\widetilde{x}, \widetilde{b}$ – возмущенные вектора. Считается, что система (7.2) устойчива по правой части, если при любых $b$, $\widetilde{b}$ справедлива оценка

$$\|\delta x\| \le M_1 \|\delta b\|, \tag{7.22}$$

где $M_1 = \text{const} > 0$, $M_1 \ne \varphi(b, \widetilde{b})$. Условие (7.22) показывает, что при уменьшении погрешности правой части погрешность решения тоже стремится к 0: $\|\delta x\| \to 0$ при $\|\delta b\| \to 0$. Наличие устойчивости важно для решения задачи, поскольку нельзя задать точно правую часть уравнения $b$. Погрешность $\delta b = \widetilde{b} - b$ возникает также в процессе округления.

Для устойчивости по правой части необходимо, чтобы $\det A \ne 0$. Вычитая из уравнения (7.21) уравнение (7.2), получаем уравнение для погрешности

$$A(\widetilde{x} - x) = \widetilde{b} - b \quad \Rightarrow \quad A(\delta x) = \delta b,$$

откуда следует

$$\delta x = A^{-1}(\delta b) \ \text{ и } \ \|\delta x\| \le \left\|A^{-1}\right\| \|\delta b\|, \tag{7.23}$$

то есть неравенство (7.22) выполняется с $M_1 = \left\|A^{-1}\right\|$. Отсюда следует, что чем меньше $\det A$, тем больше $M_1$ и хуже устойчивость по правой части.

### 7.9.2. Число обусловленности

Получим выражение относительной погрешности решения через относительную погрешность правой части. Для этого, умножив (7.23) $\|\delta x\| \le \left\|A^{-1}\right\| \|\delta b\|$ на неравенство $\|b\| \le \|A\| \|x\|$, получим

$$\frac{\|\delta x\|}{\|x\|} \le M_A \frac{\|\delta b\|}{\|b\|},$$

где $M_A = \left\|A^{-1}\right\| \|A\|$ и называется числом обусловленности матрицы $A$. Число $M_A$ характеризует степень зависимости относительной погрешности решения от относительной погрешности правой части. Матрицы с большим числом $M_A$ называются плохо обусловленными матрицами.

Свойства числа обусловленности:

1) $M_A \geq 1$;

2) 2) $M_A \geq |\lambda_{max}(A)|/|\lambda_{min}(A)|$, где $\lambda$ – собственные числа матрицы $A$;

3) 3) $M_{AB} \leq M_A M_B$.

Число $\rho(A) = |\lambda_{max}(A)|$ называется спектральным радиусом матрицы $A$.

Покажем, что для любой нормы матрицы $\rho(A)$ удовлетворяет неравенству

$$\rho(A) \leq \|A\|.$$

Пусть $y$ – собственный вектор матрицы $A$, соответствующий $\lambda_{max}$, то есть $Ay = \lambda_{max} y$, тогда $\|Ay\| = |\lambda_{max}| \|y\|$.

Используя свойство нормы $\|Ay\| \leq \|A\| \|y\|$, получаем

$$|\lambda_{max}(A)| \|y\| \leq \|A\| \|y\| \quad \Rightarrow \quad |\lambda_{max}(A)| \leq \|A\|. \tag{7.24}$$

Поскольку $\lambda_{min}^{-1}(A)$ является максимальным по модулю собственным значением матрицы $A^{-1}$, для него выполняется неравенство

$$|\lambda_{min}(A)|^{-1} \leq \|A^{-1}\|. \tag{7.25}$$

Перемножая неравенства (7.24) и (7.25), получаем требуемое неравенство 2): $M_A \geq |\lambda_{max}(A)|/|\lambda_{min}(A)|$.

Для некоторых матриц и норм условие 2) выполняется со знаком равенства. Например, норма симметричной матрицы совпадает со спектральным радиусом:

$$\|A\| = \rho(A).$$

Аналогично $\|A^{-1}\| = \rho(A^{-1}) = |\lambda_{min}(A)|^{-1}$, и, следовательно,

$$M_A = |\lambda_{max}(A)|/|\lambda_{min}(A)|,$$

для симметричной матрицы $A$ и среднеквадратичной нормы вектора $\|x\| = \sqrt{(x,x)}$.

Свойство 1) следует из 2), а свойство 3) – из свойства матричных норм $\|AB\| \leq \|A\| \|B\|$.

В общем случае, когда имеются возмущения вектора $b$ и матрицы $A$ и решается возмущенная система

$$\widetilde{A}\widetilde{x} = \widetilde{b}, \tag{7.26}$$

где $\delta A = \widetilde{A} - A$; $\delta x = \widetilde{x} - x$; $\delta b = \widetilde{b} - b$, справедлива теорема о полной относительной погрешности, приводимая здесь без доказательства.

Пусть матрица $A$ имеет обратную, и выполнено условие

$$\|\delta A\| < \left\|A^{-1}\right\|^{-1}.$$

Тогда матрица $\widetilde{A} = A + \delta A$ имеет обратную матрицу, и справедлива оценка относительной погрешности:

$$\frac{\|\delta x\|}{\|x\|} \le \frac{M_A}{1 - M_A\left(\dfrac{\|\delta A\|}{\|A\|}\right)}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right). \qquad (7.27)$$

### 7.9.3. Влияние погрешностей округления при решении систем линейных алгебраических уравнений методом Гаусса

Метод Гаусса – прямой и точный метод решения системы уравнений (7.2). Однако из-за ограниченности разрядной сетки ЭВМ вносятся погрешности округления при задании вектора $b$ и матрицы $A$, а также в процессе вычислений. Все это приводит к получению не точного, а приближенного решения уравнения $Ax = b$. Поскольку требуется не абсолютно точное решение, а решение с некоторой погрешностью, то необходимо, чтобы погрешность была в пределах заданной точности. Для этого требуется проводить анализ влияния погрешности округления на точность алгоритма.

Для большинства вычислительных алгоритмов влияние погрешности округления оценивается путем рассмотрения возмущенной системы (7.26)

$$\widetilde{A}\widetilde{x} = \widetilde{b}\,.$$

Предполагается, что решение системы $Ax = b$, искаженное погрешностями округления, совпадает с точным решением некоторой возмущенной системы

$$\widetilde{A}\widetilde{x} = \widetilde{b}\,.$$

Считая, что вектор $b$ задан точно, рассмотрим уравнение

$$\widetilde{A}\widetilde{x} = b\,.$$

Матрицу $\delta A = \widetilde{A} - A$ назовем матрицей эквивалентных возмущений.

Если известна оценка нормы $\delta A$, то согласно (7.27) погрешность, возникающую при расчетах в результате округлений, можно оценить как

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \le \frac{M_A}{1 - M_A\dfrac{\|\delta A\|}{\|A\|}}\frac{\|\delta A\|}{\|A\|}.$$

Отсюда следует, что на точность решения влияют число обусловленности матрицы $A$ и эквивалентное возмущение $\|\delta A\|/\|A\|$. Число обусловленности характеризует свойство исходной матрицы и не связано с используемым алгоритмом, а величина эквивалентного возмущения определяется численным алгоритмом.

Например, при решении методом Гаусса система $Ax = b$ в результате факторизации сводится к системе $LUx = b$ и к решению двух систем

$$Ly = b \quad u \quad Ux = y$$

с треугольными матрицами. Погрешности округления приводят к получению вместо решения $x$ системы $Ux = y$ решения возмущенной системы $\widetilde{U}\widetilde{x} = \widetilde{y}$ и вместо решения $y$ системы $Ly = b$ решения $\widetilde{y}$ системы $\widetilde{L}\widetilde{y} = b$. Следовательно, предполагается, что точно решается возмущенная система $\widetilde{A}\widetilde{x} = b$, где $\widetilde{A} = \widetilde{L}\widetilde{U}$, вместо исходной системы $Ax = b$. Чтобы найти матрицу $\widetilde{A}$, надо выписать все формулы метода Гаусса, внести в них погрешности и получить матрицы $\widetilde{L}$ и $\widetilde{U}$, а затем оценить норму матрицы $\delta A = LU - \widetilde{L}\widetilde{U}$.

Анализ этого алгоритма, приведенный в [15], показывает, что решение $x$ системы вычисляется с относительной погрешностью

$$\frac{\|\delta x\|}{\|x\|} = O(M_A \cdot n \cdot 2^{-t}),$$

где $n$ – порядок матрицы $A$; $t$ – число двоичных разрядов мантиссы числа на используемой ЭВМ. Например, для персональных ЭВМ $2^{-t} \approx 10^{-7}$.

## 7.10. Итерационные методы

Итерационные методы широко используются в линейной алгебре как для непосредственного получения решений систем уравнений, так и для уточнения решений, полученных прямыми методами, которые из-за ограниченной длины мантиссы вещественных чисел в памяти ЭВМ в результате округлений приобретают определенные погрешности.

### 7.10.1. Метод простой итерации (Якоби)

Пусть задана система уравнений

$$Ax = b.$$

Представим ее в виде

$$x = Gx + \beta, \tag{7.28}$$

где $G$ – матрица; $x$, $\beta$ – векторы. Зададим некоторое начальное приближеное значение вектора $x = x^{(0)}$ и подставим его в правую часть уравнения (7.28). В результате получим первое приближение вектора $x^{(1)}$. Его снова подставим в правую часть уравнения (7.28), и получим второе приближение вектора $x^{(2)}$. Продолжая далее этот процесс, получим последовательность значений вектора $x^{(k)}$:

$$x^{(1)} = Gx^{(0)} + \beta;$$

$$x^{(2)} = Gx^{(1)} + \beta;$$

. . . . . .

$$x^{(k+1)} = Gx^{(k)} + \beta,\ k = 0, 1, 2, \ldots. \qquad (7.29)$$

Если последовательность векторов $\{x^{(k)}\}$ сходится, то она сходится к решению системы уравнений. Покажем это. Из последовательностей (7.29) находим:

$$x^{(2)} = Gx^{(1)} + \beta = G(Gx^{(0)} + \beta) + \beta = G^2x^{(0)} + G\beta + \beta;$$

$$x^{(3)} = Gx^{(2)} + \beta = G^3x^{(0)} + G^2\beta + G\beta + \beta;$$

. . . . . .

$$x^{(k)} = G^kx^{(0)} + (G^{k-1} + G^{k-2} + \ldots + E)\beta.$$

Используя понятие нормы, величину вектора $x$ можно оценить так:

$$\left\|x^{(k}\right\| \le \|G\|^k \left\|x^{(0)}\right\| + \left(\|G\|^{k-1} + \|G\|^{k-2} + \ldots + \|G\| + \|E\|\right)\|\beta\|. \qquad (7.30)$$

Если норма матрицы $G$ меньше единицы $\|G\| < 1$, то $\lim\limits_{k\to\infty}\|G\|^k = 0$, и первый член этого неравенства пропадает при любом начальном значении вектора $x$. Второй член неравенства при $k \to \infty$ равен

$$\lim_{k\to\infty}(G^{k-1} + G^{k-2} + \ldots + E)\beta = (E - G)^{-1}\beta.$$

Для доказательства этого утверждения умножим $(G^{k-1} + G^{k-2} + \ldots + E)$ на $(E - G)$. В результате получим

$$(G^{k-1} + G^{k-2} + \ldots + E)(E - G) = G^{k-1} + G^{k-2} + \ldots + E - G^k - G^{k-1} - \ldots - G =$$

откуда $= E - G^k$,

$$\|E\| - \|G\|^k \le \|E - G^k\| \le \|E\| + \|G\|^k .$$

Так как $\|G\| < 1$; $\lim\limits_{k\to\infty}\|G\|^k = 0$, и $\lim\limits_{k\to\infty}\|E - G^k\| = 1$, что и требовалось доказать.

Таким образом, неравенство (7.30) преобразуется к виду

$$\|x^{(k)}\| \le \|(E-G)^{-1}\|\,\|\beta\|. \qquad (7.31)$$

С другой стороны, перенеся неизвестные в левую часть, уравнение $x = Gx + \beta$ можно записать в виде $(E-G)x = b$. Умножив его слева на обратную матрицу $(E-G)^{-1}$, получим решение системы

$$x = (E-G)^{-1}\beta ,$$

что совпадает с оценкой решения методом итераций (7.31) при $k \to \infty$.

Таким образом, для сходимости процесса итераций необходимо, чтобы $\|G\| < 1$. Причем сходимость в этом случае имеет место при произвольном начальном значении $x^{(0)}$.

Итерации продолжаются до выполнения условия

$$\|x^{(k)} - x^{(k-1)}\| < \varepsilon , \qquad (7.32)$$

где $\varepsilon$ – заданная погрешность расчета

Оценка сходимости алгоритма

$$\|x - x^{(k)}\| \le \frac{\|G\|}{1-\|G\|}\|x^{(k)} - x^{(k-1)}\|,$$

откуда следует оценка погрешности решения

$$\|x - x^{(k)}\| \le \frac{\|G\|^k}{1-\|G\|}\|x^{(1)} - x^{(0)}\|;$$

здесь $x$ – точное решение.

Число операций для получения решения $N \approx 2n^2k$, где $n$ – порядок системы; $k$ – число итераций.

### 7.10.2. Метод Гаусса–Зейделя

Метод Гаусса–Зейделя отличается от метода простой итерации тем, что при расчете последующих компонент вектора $x$ в нем используются значения ранее уже вычисленных компонент.

Пусть система $Ax = b$ приведена к виду (7.28) $x = Gx + \beta$. Представим матрицу $G$ в виде суммы трех матриц $G = L + D + R$, где $D$ – диагональная

матрица; $L$ – нижняя треугольная матрица; $R$ – верхняя треугольная матрица (без диагональных элементов). Тогда алгоритм метода Гаусса–Зейделя можно записать в виде

$$x^{(k+1)} = Lx^{(k+1)} + Dx^{(k+1)} + Rx^{(k)} + \beta$$

или в виде

$$(E - D)x^{(k+1)} = Lx^{(k+1)} + Rx^{(k)} + \beta,$$

где $x^{(k)}$, $x^{(k+1)}$– два последовательных приближения к точному решению, $k = 1, 2, \ldots$ – номер итерации.Уравнения этой системы решаем последовательно друг за другом, используя ранее вычисленные компоненты вектора $x^{(k+1)}$. Итерации продолжаются до выполнения условия сходимости (7.32).

Рассмотрим методы Якоби и Гаусса–Зейделя на примере решения системы трех уравнений:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1;$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2;$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3.$$

Алгоритм процесса простой итерации (Якоби) имеет вид следующий:

$$x_1^{(k+1)} = -a_{12}x_2^{(k)} / a_{11} - a_{13}x_3^{(k)} / a_{11} + b_1 / a_{11};$$

$$x_2^{(k+1)} = -a_{21}x_1^{(k)} / a_{22} - a_{23}x_3^{(k)} / a_{22} + b_2 / a_{22};$$

$$x_3^{(k+1)} = -a_{31}x_1^{(k)} / a_{33} - a_{32}x_2^{(k)} / a_{33} + b_3 / a_{33},\ k = 0, 1, 2, \ldots. \quad (7.33)$$

Зададим произвольные начальные значения $x_1^{(0)}$, $x_2^{(0)}$, $x_3^{(0)}$ и подставим их в правую часть системы (7.33). В результате получим следующие приближения неизвестных: $x_1^{(1)}$, $x_2^{(1)}$, $x_3^{(1)}$ и т. д. до выполнения условия сходимости (7.32).

Алгоритм метода Гаусса–Зейделя также использует начальные значения вектора $x$ и имеет следующий вид:

$$x_1^{(k+1)} = -a_{12}x_2^{(k)} / a_{11} - a_{13}x_3^{(k)} / a_{11} + b_1 / a_{11};$$

$$x_2^{(k+1)} = -a_{21}x_1^{(k+1)} / a_{22} - a_{23}x_3^{(k)} / a_{22} + b_2 / a_{22};$$

$$x_3^{(k+1)} = -a_{31}x_1^{(k+1)} / a_{33} - a_{32}x_2^{(k+1)} / a_{33} + b_3 / a_{33},$$

то есть отличается от метода простой итерации тем, что при расчете компонент вектора $x$ на $(k+1)$-й итерации используются все ранее вычисленные компоненты $(k+1)$-й итерации.

Условие сходимости метода Гаусса–Зейделя такое же, $\|G\| < 1$, однако сходимость в общем случае более быстрая, чем у метода Якоби.

Для сходимости итерационного процесса достаточно, чтобы модули диагональных коэффициентов для каждого уравнения системы были не меньше суммы модулей всех остальных коэффициентов:

$$|a_{ii}| \geq \sum_{i \neq j} |a_{ij}|, \quad i, j = 1, 2, \ldots, n,$$

то есть матрица $A$ системы $Ax = b$ должна иметь диагональное преобладание.

Приведем рассмотренные итерационные методы в матричном виде. Для этого представим матрицу $A$ как сумму нижней треугольной, диагональной и верхней треугольной матриц $A = L + D + R$:

$$L = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ a_{21} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & 0 \end{bmatrix}; \; D = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{nn} \end{bmatrix};$$

$$R = \begin{bmatrix} 0 & a_{12} & \ldots & a_{1n} \\ 0 & 0 & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 \end{bmatrix}. \qquad (7.34)$$

Заменяя в уравнении $Ax = b$ матрицу A суммой матриц $A = L + D + R$, представим алгоритмы Якоби и Гаусса–Зейделя соответственно в виде:

$$Dx^{(k+1)} = -(L + R)x^{(k)} + b;$$

$$Dx^{(k+1)} = -Lx^{(k+1)} - Rx^{(k)} + b.$$

В правых частях этих уравнений записаны уже известные компоненты вектора $x$.

### 7.10.3. Метод релаксации

Метод релаксации является эффективным и широко используемым итерационным методом для решения систем линейных алгебраических уравнений $Ax = b$ с симметричными положительно определенными матрицами.

Метод представляет собой некоторую модернизацию итерационного метода Гаусса–Зейделя и относится к методам типа предиктор–корректор. Пусть получено значение вектора $x$ на $k$-й итерации. Последующие итерационные шаги выполняются следующим образом. На этапе предиктор методом Гаусса–Зейделя вычисляется предварительное $(k+1)$-е приближение $i$-й компоненты вектора $x$, то есть $\tilde{x}_i^{(k+1)}$, по формуле

$$\tilde{x}_i^{(k+1)} = -\frac{a_{i1}}{a_{ii}} x_i^{(k+1)} - \frac{a_{i2}}{a_{ii}} x_2^{(k+1)} - \ldots - \frac{a_{i,j-1}}{a_{ii}} x_{i-1}^{(k+1)} - \frac{a_{i,j+1}}{a_{ii}} x_{i+1}^{(k)} - \ldots - \frac{a_{in}}{a_{ii}} x_n^{(k)} + \frac{b_i}{a_{ii}},$$

где $a_{ij}$ – коэффициенты матрицы $A$; $b_i$ – компонента вектора $b$. Затем на этапе корректор эта величина уточняется добавлением к ней смещения компоненты $x_i$, умноженного на коэффициент $(\omega - 1)$, то есть добавлением величины $(\omega - 1)(\tilde{x}_i^{(k+1)} - x_i^{(k)})$, где $\omega$ – параметр релаксации.

Таким образом, уточненное значение $i$-й компоненты $(k+1)$-го приближения вычисляется по формуле

$$x_i^{(k+1)} = \tilde{x}_i^{(k+1)} + (\omega - 1)(\tilde{x}_i^{(k+1)} - x_i^{(k)}) = \omega \tilde{x}_i^{(k+1)} + (1 - \omega) x_i^{(k)}.$$

В матричной виде метод решения можно записать следующим образом:

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

где $D$, $L$, $R$ – матрицы, вид которых описан формулами (7.34).

При $\omega = 1$ метод релаксации совпадает с методом Гаусса–Зейделя. При $\omega > 1$ метод обычно называют методом последовательной верхней релаксации, а при $\omega < 1$ – методом последовательной нижней релаксации. В последнее время при любых $\omega$ этот метод называют методом последовательной верхней релаксации.

Для положительно определенных симметричных матриц $A$ метод релаксации сходится для $0 < \omega < 2$. Можно подобрать $\omega > 1$ так, чтобы метод релаксации сходился значительно быстрее метода Гаусса–Зейделя. Однако теоретически этот выбор сделать чрезвычайно трудно, поэтому обычно он осуществляется экспериментальным путем.

На практике встречается много вариантов метода верхней релаксации, отличающихся использованием различных коэффициентов $\omega_i$ при уточнении отдельных компонент вектора $x$. Эти методы подробно описаны в специальной монографии [24].

# 8. ПРИБЛИЖЕННОЕ РЕШЕНИЕ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

Задачи вычисления корней нелинейных уравнений часто встречаются при научных исследованиях. Корнем, или решением нелинейного уравнения

$$f(x) = 0, \tag{8.1}$$

называется такое значение $x = \xi$, которое превращает уравнение (8.1) в тождество: $f(\xi) \equiv 0$. Нелинейные уравнения обычно подразделяют на алгебраические и трансцендентные. Алгебраическими называются уравнения, содержащие алгебраические функции. Уравнения, содержащие тригонометрические, показательные, логарифмические и др. функции, называются трансцендентными.

Методы решения нелинейных уравнений бывают прямыми и итерационными. Прямые методы дают решение в виде конечной формулы и применимы лишь к узкому классу уравнений. Для решения большинства нелинейных уравнений применяются итерационные методы, то есть методы последовательных приближений.

Приближенное нахождение изолированных корней уравнения обычно состоит из двух этапов:

1) отделения корней, то есть определения интервалов, содержащих отдельные корни;

2) уточнения приближенных корней, т. е. доведения их до заданной степени точности.

## 8.1. Отделение корней уравнения

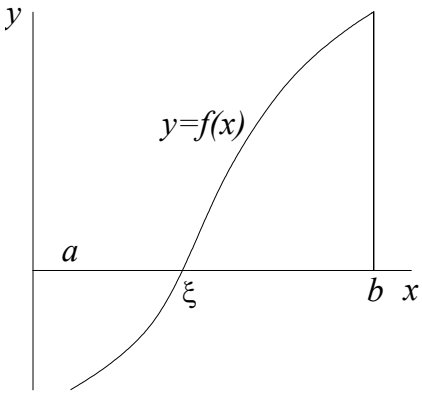


Рис. 9

При отделении корней пользуются теоремой: если непрерывная функция $f(x)$ принимает значения разных знаков на концах отрезка $[a,b]$, то есть $f(a)f(b) < 0$, то внутри этого отрезка содержится по меньшей мере один корень уравнения $f(x) = 0$ (рис. 9). Корень будет единственным, если первая производная функции $f'(x)$ существует и сохраняет постоянный знак внутри интервала $[a,b]$.

Отделение корней начинаем с определения знаков $f(x)$ в крайних точках $x = a$, $x = b$ и в ряде промежуточных точек $x = x_1, x_2, \ldots$, которые выбираются из анализа особенностей функции. Если на некотором интервале окажется $f(x_i)f(x_{i+1}) < 0$, то согласно теореме на этом интервале

$[x_i, x_{i+1}]$ имеется корень уравнения $f(x)=0$. Чтобы убедиться в существовании единственного корня на отрезке, нужно провести процесс половинного деления, определяя знаки в точках деления. При отделении корней алгебраического уравнения

$$a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n = 0$$

следует помнить, что оно имеет *n* корней (включая, возможно, и комплексные корни). Если для такого уравнения получаем *n* перемен знаков, то все его корни вещественные и отделены.

*Пример*. Отделить корни уравнения $f(x) = 2x^3 - 5x + 1 = 0$.

Это уравнение имеет не более 3 действительных корней. Составим таблицу знаков функции в различных точках:

| $X$ | –3 | –1 | 0 | 1 | 3 | 5 |
|---|---|---|---|---|---|---|
| $f(x)$ | – | + | + | – | + | + |

Из таблицы видно, что функция имеет 3 действительных корня, лежащих в интервалах (–3, –1), (0,1), (1,3).

## 8.2. Погрешность приближенного значения корня

Если $\xi$ – точное значение корня уравнения $f(x)=0$, а $\bar{x}$ – его приближенное значение на отрезке $[a,b]$, причем $|f'(x)| \geq m_1 > 0$ на $[a,b]$, то погрешность приближенного значения корня будет равна

$$|\bar{x} - \xi| \leq |f(\bar{x})| / m_1 .$$

Эта оценка следует непосредственно из теоремы Лагранжа о среднем значении производной на отрезке

$$\frac{f(\bar{x}) - f(\xi)}{\bar{x} - \xi} = f'(c),$$

где $c$ – некоторая промежуточная точка отрезка $[\bar{x}, \xi]$. Так как $f(\xi) = 0$ и $|f'(c)| \geq m_1$, получаем

$$|f(\bar{x}) - f(\xi)| = |f(\bar{x})| \geq m_1 |\bar{x} - \xi| \quad \Rightarrow \quad |\bar{x} - \xi| \leq \frac{|f(\bar{x})|}{m_1}.$$

## 8.3. Метод половинного деления

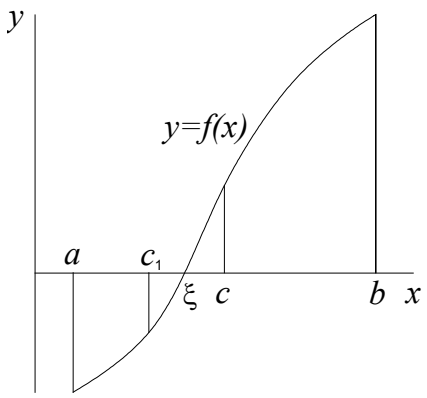


Рис. 10

Пусть требуется найти корень уравнения $f(x)=0$ на отрезке $[a,b]$, который был задан заранее, либо получен методом отделения корней.

Решение задачи выполняется следующим образом. Проверяется условие существования корня на отрезке $[a,b]$: $f(a)f(b)<0$. Если это условие выполнено, приступаем к вычислению корня. Отрезок делится пополам точкой $c=(a+b)/2$ (рис. 10) и вычисляется значение функции $f(c)$ в этой точке. Проверяется, на каком из двух получившихся отрезков $[a,c]$ или $[c,b]$ располагается корень. Для этого следует определить знак произведения $f(a)f(c)$ или $f(c)f(b)$. Если $f(a)f(c)<0$, то $f(c)f(b)>0$, и корень располагается на отрезке $[a,c]$. Следовательно, отрезок $[c,b]$ можно отбросить и искать корень на отрезке $[a,c]$, который обозначается как $[a_1,b_1]$. В противном случае $f(a)f(c)>0$, корень располагается на отрезке $[c,b]$, и этот отрезок обозначается $[a_1,b_1]$. С отрезком $[a_1,b_1]$ производятся точно такие же действия, как и с предыдущим, в результате чего получается отрезок $[a_2,b_2]$ вдвое меньшей длины, содержащий корень. В ходе повторных операций половинного деления получается последовательность вложенных друг в друга отрезков $[a_1,b_1]$, $[a_2,b_2]$,…, $[a_n,b_n]$, таких, что $f(a_n)f(b_n)<0$, и, следовательно, содержащих корень. Длина их уменьшается по закону

$$b_n-a_n=(b-a)/2^n.$$

Процесс половинного деления продолжается до тех пор, пока длина отрезка $[a_n,b_n]$ не станет меньше заданной погрешности $\varepsilon$:

$$|b_n-a_n|<\varepsilon. \qquad (8.2)$$

Среднюю точку отрезка $[a_n,b_n]$ можно принять за приближенное значение корня $\xi=(a_n+b_n)/2$. Из неравенства (8.2) можно определить число *n* операций половинного деления, необходимых для получения заданной точности решения:

$$b_n-a_n=(b-a)/2^n<\varepsilon \quad\Rightarrow\quad n\ge\log_2[(b-a)/\varepsilon].$$

Погрешность приближенного решения можно оценить по формуле:

$$|x_n - \xi| \le \frac{|f(x_n)|}{m_1}, \quad m_1 = \min_{x\in[a,b]} |f'(x)|.$$

Дополнительно погрешность вычисления корня можно контролировать выполнением условия $|f(\xi)| \le \varepsilon$.

### 8.4. Метод хорд или пропорциональных частей

Пусть задан отрезок $[a,b]$, на концах которого функция $f(x)$ имеет значения разных знаков – $f(a)f(b) < 0$. Пусть $f(a) < 0$; $f(b) > 0$. Соединим хордой точки $f(a)$ и $f(b)$. Точку пересечения ее с осью $x$ обозначим $x_1$ (рис 11а). Из подобия треугольников $ABC$ и $Ax_1D$ имеем

$$\frac{f(b)-f(a)}{b-a} = \frac{-f(a)}{x_1-a}, \quad \Rightarrow \quad x_1 = a - \frac{f(a)}{f(b)-f(a)}(b-a).$$

Это будет первым приближением к корню уравнения $f(x) = 0$. Сравнивая знаки величин $f(x_1)$ и $f(b)$, видим, что корень функции $f(x)$ находится на отрезке $[x_1,b]$, поскольку $f(x_1)f(b) < 0$.

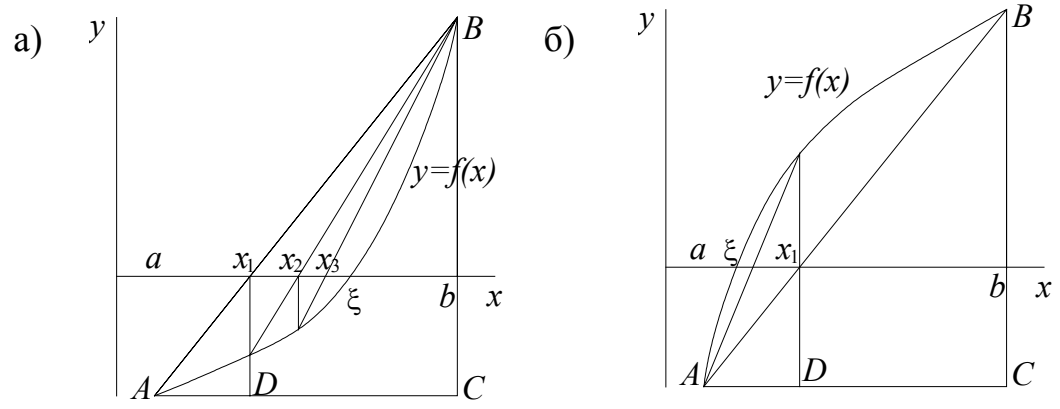


Рис. 11

Следующее приближение к корню $x_2$ получаем в точке пересечения с осью $x$ хорды, соединяющей точки $f(b)$ и $f(x_1)$. Ее координаты определяются аналогично $x_1$ по формуле

$$x_2 = x_1 - \frac{f(x_1)}{f(b)-f(x_1)}(b-x_1).$$

Последующие итерации выполняются по формуле

$$x_{n+1} = x_n - \frac{f(x_n)}{f(b) - f(x_n)}(b - x_n), \quad n = 0, 1, 2, \ldots. \tag{8.3}$$

Итерации продолжаются до тех пор, пока значение $f(x)$ и разность двух последовательных приближений к корню по модулю не станут меньше наперед заданных погрешностей расчета $\varepsilon_1$ и $\varepsilon$:

$$\left|f(x_n)\right| \le \varepsilon_1 \ \text{ и } \ \left|x_n - x_{n-1}\right| \le \varepsilon.$$

Для случая, изображенного на рис.11б, рекуррентная формула для расчета корня функции $f(x)$ имеет вид

$$x_{n+1} = x_n - \frac{f(x_n)}{f(x_n) - f(a)}(x_n - a), \quad n = 0, 1, 2, \ldots. \tag{8.4}$$

В формуле (8.3) неподвижным концом отрезка остается точка *b*, а в формуле (8.4) – точка *a*.

Для произвольной функции $f(x)$ за неподвижный конец отрезка $[a,b]$ выбирается тот, для которого знаки $f(x)$ и $f''(x)$ совпадают, то есть $f(x)f''(x) > 0$. Таким образом, если $f(b)f''(b) > 0$, то для расчетов нужно использовать формулу (8.3), а если $f(a)f''(a) > 0$, то формулу (8.3).

Погрешность приближенного решения можно оценить по формулам:

$$\left|x_n - \xi\right| \le \frac{\left|f(x_n)\right|}{m_1}, \quad m_1 = \min_{x\in[a,b]}\left|f'(x)\right|;$$

$$\left|x_n - \xi\right| \le \frac{M_1 - m_1}{m_1}\left|x_n - x_{n-1}\right|, \quad M_1 = \max_{x\in[a,b]}\left|f'(x)\right|,$$

где $\xi$ – точное значение корня.

## 8.5. Метод Ньютона (касательных)

Этот метод является также методом последовательных приближений. Пусть $\xi$ – корень уравнения $f(x) = 0$, а производные $f'(x)$, $f''(x)$ на отрезке $[a,b]$ непрерывны и сохраняют знак. Тогда *n*-е приближение к корню $x_n$ можно представить в виде

$$\xi = x_n + h_n, \tag{8.5}$$

где $h_n$ – малая величина.

Используя формулу Тейлора, можно записать

$$f(\xi) = f(x_n + h_n) \approx f(x_n) + h_n f'(x_n) \approx 0.$$

Отсюда следует

$$h_n = -f(x_n)/f'(x_n). \qquad (8.6)$$

Принимая $x_{n+1} \approx \xi$ и подставляя (8.6) в (8.5), получаем формулу для вычисления приближенных значений корня:

$$x_{n+1} = x_n - f(x_n)/f'(x_n), \quad n = 0, 1, 2, \ldots. \qquad (8.7)$$

Геометрически этот метод означает следующее (рис.12). В точке $c_0 = f(b)$ проводим касательную к кривой $f(x)$. Точка пересечения ее с осью $x$ дает первое приближение корня $x_1$. Затем проводим перпендикуляр к оси $x$ в точке $x_1$ до пересечения его с кривой $f(x)$ и получаем точку $c_1$. Далее снова проводим касательную к $f(x)$ в точке $c_1$. Ее пересечение с осью $x$ дает следующее приближение корня $x_2$ и т. д. до тех пор, пока не будут выполнены условия $|f(x_n)| \le \varepsilon_1$ и $|x_n - x_{n-1}| \le \varepsilon$, где $\varepsilon_1$, $\varepsilon$ – малые величины, погрешности расчета.

Расчетную формулу (8.7) можно получить также из геометрических соображений. Из прямоугольного треугольника $x_1 b c_0$ на рис. 12 следует

$$\operatorname{tg}\alpha = \frac{bc_0}{x_1 b} = \frac{f(b)}{b - x_1} \quad \Rightarrow$$

$$(b - x_1)\operatorname{tg}\alpha = f(b) \quad \Rightarrow \quad x_1 = b - \frac{f(b)}{\operatorname{tg}\alpha}.$$

Так как $\operatorname{tg}\alpha = f'(b)$, а $b = x_0$, отсюда получаем $x_1 = x_0 - f(x_0)/f'(x_0)$, и для последующих приближений имеем

$$x_{n+1} = x_n - f(x_n)/f'(x_n), \quad n = 0, 1, 2, \ldots.$$

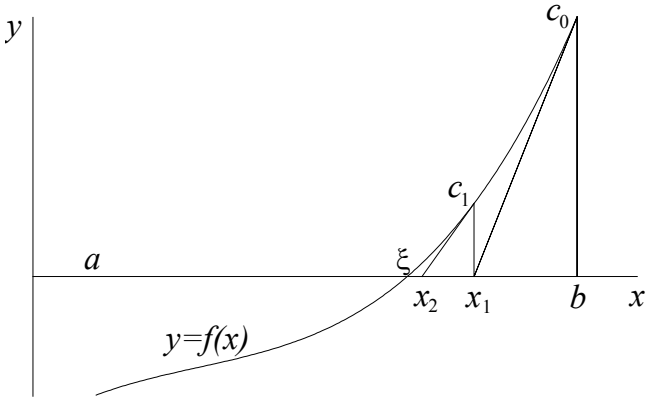


Рис. 12

Для рассматриваемого метода доказана следующая *теорема*. Если $f(a)f(b) < 0$, а $f'(x)$ и $f''(x)$ сохраняют определенные знаки на отрезке $[a,b]$, то для любой начальной точки $x_0 \in [a,b]$, для которой $f(x_0)f''(x_0) > 0$, можно вычислить корень с любой степенью точности методом Ньютона.

Следовательно, за начальную точку в методе Ньютона следует

брать точку $x_0$, в которой $f(x_0)f''(x_0) > 0$.

Оценку погрешности решения можно производить по формулам:

$$|\xi - x_n| \le |f(x_n)| / m_1;$$

$$|\xi - x_n| \le \frac{M_2}{2m_1}(x_n - x_{n-1})^2.$$

Здесь $m_1 = \min\limits_{x\in[a,b]}|f'(x)|$, а $M_2 = \max\limits_{x\in[a,b]}|f''(x)|$. Этот метод – самый эффективный и быстро сходящийся из рассмотренных.

## 8.6. Метод простой итерации

Пусть дано уравнение $f(x) = 0$. Необходимо определить корень $\xi$ этого уравнения на отрезке $[a,b]$. Причем функция $f(x)$ имеет на концах отрезка значения разных знаков $f(a)f(b) < 0$. Заменим уравнение $f(x) = 0$ равносильным уравнением $x = \varphi(x)$. Зададим некоторое начальное приближение корня $x_0 \in [a,b]$ и, подставляя его в правую часть уравнения, получим первое приближение корня $x_1$. Затем, подставляя найденное значение $x_1$ в правую часть уравнения, получим следующее приближение корня $x_2$, и так далее:

$$x_1 = \varphi(x_0);$$

$$x_2 = \varphi(x_1);$$

$$\ldots\ldots\ldots;$$

$$x_n = \varphi(x_{n-1});$$

$$(n \to \infty).$$

Если последовательность $x_0, x_1, \ldots, x_n$ при $n \to \infty$ сходится, то она сходится к точному решению уравнения $f(\xi) = 0$. Процесс итераций прекращаем, если результаты двух последовательных приближений близки:

$$|f(x_n)| \le \varepsilon_1, \quad |x_n - x_{n-1}| \le \varepsilon,$$

где $\varepsilon_1$, $\varepsilon$ – малые величины, погрешности расчета.

Достаточным условием сходимости метода простой итерации является условие

$$|\varphi'(x)| \le q < 1, \quad x \in [a,b].$$

Покажем это: пусть $x_{n+1} = \varphi(x_n)$; $x_n = \varphi(x_{n-1})$. Пользуясь теоремой Лагранжа о среднем, запишем разность

$$x_{n+1} - x_n = \varphi(x_n) - \varphi(x_{n-1}) = (x_n - x_{n-1})\varphi'(c_n),$$

учитывая $|\varphi'(c_n)| \le q$, где $x_{n-1} < c_n < x_n$, получаем

$$|x_{n+1} - x_n| \le q|x_n - x_{n-1}|;$$

$$|x_{n+1} - x_n| \le q^2|x_{n-1} - x_{n-2}|$$

. . . . . .

$$|x_{n+1} - x_n| \le q^n|x_0 - x_1|.$$

$$\lim_{n\to\infty}|x_{n+1} - x_n| = 0, \text{ если } |q| < 1, \text{ т.е. } |\varphi'(x_n)| < 1.$$

Оценка погрешности метода

$$|\xi - x_n| \le \frac{q^n}{1-q}|x_1 - x_0|; \quad |\xi - x_n| \le \frac{q}{1-q}|x_n - x_{n-1}|.$$

Процесс решения для различных типов уравнений приведен на рис. 13а,б.

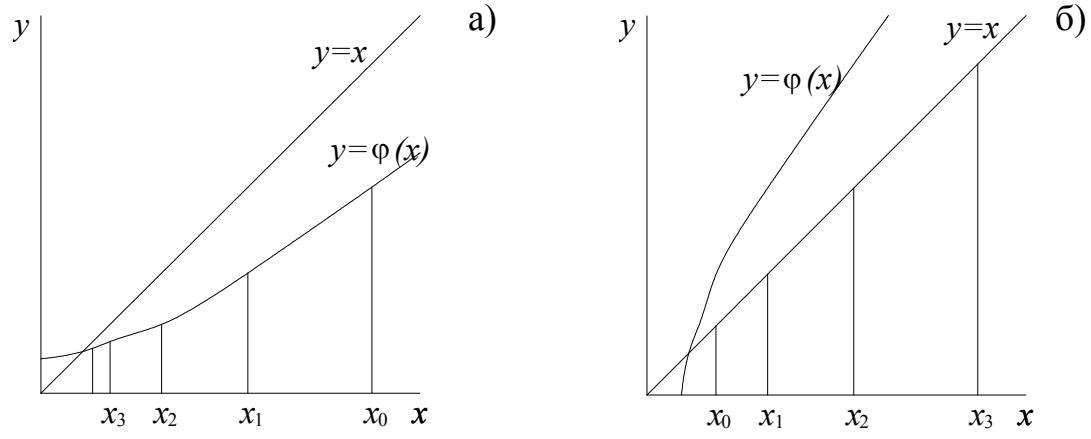


Рис. 13

*Приведение уравнения к виду, удобному для итераций.* Представим уравнение $f(x) = 0$ в виде

$$x = x - \lambda f(x), \quad (\lambda > 0),$$

тогда $\varphi(x) = x - \lambda f(x)$. Подберем $\lambda$ таким, чтобы $\varphi'(x) \le q < 1$:

$$\varphi'(x) = 1 - \lambda f'(x), \quad 0 \le 1 - \lambda f'(x) \le q < 1;$$

$$0 \le 1 - \lambda M_1 \le 1 - \lambda m_1 \le q, \quad \Rightarrow \quad \lambda = 1/M_1;$$

$$q = 1 - m_1 / M_1 \le 1; \;\; m_1 = \min|f'(x)|; \quad M_1 = \max|f'(x)|; \quad x \in [a,b].$$

Отсюда получаем $\varphi(x) = x - f(x)/M_1$ .

*Пример*. Найти наибольший положительный корень $\xi$ уравнения $x^4 + 2x = 10000$ с точностью до $10^{-4}$.

*Решение*. Первое приближение $x_0 = 10$, причем $\xi < x_0$. За основной интервал $[a,b]$ возьмем интервал [9,10]. Исходное уравнение можно переписать относительно *x* тремя способами:

1) $x = 5000 - x^4/2$,

2) $x = \dfrac{10000}{x^3} - \dfrac{2}{x^2}$,

3) $x = \sqrt[4]{10000 - 2x}$.

Уравнение 1) не годится для итераций, поскольку

$$\varphi(x) = 5000 - x^4/2 \quad \Rightarrow \quad |\varphi'(x)| = |-2x^2| > 1.$$

Уравнение 2) также не пригодно для итераций, так как

$$\varphi(x) = \frac{10000}{x^3} - \frac{2}{x^2} \quad \Rightarrow \quad |\varphi'(x)| > 3.$$

Только уравнение 3) пригодно для решения:

$$\varphi'(x) = -\frac{1}{2(10000 - 2x)^{3/4}} \quad \Rightarrow \quad |\varphi'(x)| \le \frac{1}{1000} << 1.$$

Вычисляя последовательные приближения *x* по формуле 3), получаем:

$$x_0 = 10; \; x_1 = 9{,}994996; \; x_2 = 9{,}994999; \; x_3 = 9{,}994999.$$

## 9. РЕШЕНИЕ СИСТЕМ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

Многие задачи оптимизации производственных, экономических, технических и других процессов сводятся к отысканию корней систем нелинейных уравнений. В общем случае систему нелинейных уравнений можно записать в виде

$$f_i(x_1, x_2, \ldots, x_n) = 0, \quad i = 1, 2, \ldots, n. \tag{9.1}$$

Если неизвестные $x_1, x_2, \ldots, x_n$ и функции $f_1, f_2, \ldots, f_n$ рассматривать как *n*–мерные векторы $x = (x_1, x_2, \ldots, x_n)^T$, $F = (f_1, f_2, \ldots, f_n)^T$, то систему (9.1) можно записать кратко в векторном виде $F(x) = 0$.

Решением системы (9.1) будет вектор $x = \xi$, превращающий систему в тождество. Для систем нелинейных уравнений в отличие от линейных не существует прямых методов решения. Лишь в отдельных простейших случаях нелинейные уравнения можно решить непосредственно. Обычно для решения нелинейных систем применяются методы последовательных приближений (итерационные методы) Ньютона и простой итерации.

## 9.1. Метод Ньютона

Формулы метода Ньютона для систем нелинейных уравнений, как и в случае одного нелинейного уравнения, получаются посредством применения формулы Тейлора для функции $F(x)$ в окрестности решения $\xi$. Пусть нам известно некоторое *k*-е приближение $x^{(k)}$ к решению $\xi$ системы (9.1). Поэтому решение можно представить как

$$\xi = x^{(k)} + \Delta x, \tag{9.2}$$

где $\Delta x$ – приращение (поправка) к приближенному решению. В развернутом виде это уравнение записывается как

$$\xi_1 = x_1^{(k)} + \Delta x_1; \quad \xi_2 = x_2^{(k)} + \Delta x_2; \quad \ldots; \quad \xi_n = x_n^{(k)} + \Delta x_n.$$

Разложим функцию $F(x)$ в ряд Тейлора по малому параметру $\Delta x$, оставив только два первых члена разложения в силу малости параметра:

$$F(\xi) = F(x^{(k)} + \Delta x) = F(x^{(k)}) + F'(x^{(k)})\Delta x \approx 0. \tag{9.3}$$

Здесь $F'(x^{(k)})$ – матрица Якоби для системы уравнений

$$F'(x^{(k)}) = W(x^{(k)}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \ldots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \ldots & \dfrac{\partial f_2}{\partial x_n} \\ \ldots & \ldots & \ldots & \ldots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \ldots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}.$$

Полагая, что матрица Якоби $W(x^{(k)})$ неособенная, разрешим уравнение (9.3) относительно вектора $\Delta x$:

$$\Delta x = -W^{-1}(x^{(k)})F(x^{(k)}).$$

Здесь $W^{-1}(x^{(k)})$ – обратная матрица матрицы Якоби.

Подставив значение приращения $\Delta x$ в уравнение (9.2), получаем алгоритм метода Ньютона

$$x^{(k+1)} = x^{(k)} - W^{-1}(x^{(k)})F(x^{(k)}). \qquad (9.4)$$

Здесь вместо точного решения $\xi$ системы (9.1) в левой части алгоритма (9.4) поставлено последующее приближение $x^{(k+1)}$ к решению, так как значение приращения $\Delta x$ получено из приближенного уравнения (9.3). При расчете по формуле (9.4) на каждом шаге итерации необходимо вычислять обратную матрицу $W^{-1}(x^{(k)})$ при новых значениях $x^{(k)}$. Расчеты продолжаются до выполнения условия сходимости решения, т.е. близости двух последовательных приближений

$$\left\| x^{(k+1)} - x^{(k)} \right\| \le \varepsilon, \quad k = 1, 2, \ldots, \qquad (9.5)$$

где $\varepsilon$ – малая величина, погрешность решения.

Этот метод обладает значительно большей скоростью сходимости, чем метод простой итерации. Для его применения необходимо, чтобы матрица Якоби была неособенной.

*Пример*. Пусть требуется решить систему двух нелинейных уравнений:

$$f(x, y) = 0;\ \varphi(x, y) = 0.$$

При этом известно приближенное значение решения: $x = a;\ y = b$.

Введем векторы:

$$z = \begin{bmatrix} x \\ y \end{bmatrix}; \quad F(z) = \begin{bmatrix} f(x, y) \\ \varphi(x, y) \end{bmatrix}; \quad z^{(0)} = \begin{bmatrix} a \\ b \end{bmatrix}$$

и составим матрицу Якоби для системы $F(z) = 0$:

$$W(z) = \begin{bmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \\ \dfrac{\partial \varphi}{\partial x} & \dfrac{\partial \varphi}{\partial y} \end{bmatrix}.$$

Полагая, что $\det W \ne 0$ при $x = a,\ y = b$, вычисляем обратную матрицу $W^{-1}$:

$$WW^{-1} = E = \begin{bmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \\ \dfrac{\partial \varphi}{\partial x} & \dfrac{\partial \varphi}{\partial y} \end{bmatrix} \begin{bmatrix} \eta_{11} & \eta_{12} \\ \eta_{21} & \eta_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Rightarrow$$

$$\frac{\partial f}{\partial x}\eta_{11}+\frac{\partial f}{\partial y}\eta_{21}=1,\quad \frac{\partial f}{\partial x}\eta_{12}+\frac{\partial f}{\partial y}\eta_{22}=0;$$

$$\frac{\partial \varphi}{\partial x}\eta_{11}+\frac{\partial \varphi}{\partial y}\eta_{21}=0,\quad \frac{\partial \varphi}{\partial x}\eta_{12}+\frac{\partial \varphi}{\partial y}\eta_{22}=1;$$

$$W^{-1}=\frac{1}{\det W}\begin{bmatrix}\dfrac{\partial \varphi}{\partial y} & -\dfrac{\partial f}{\partial y}\\ -\dfrac{\partial \varphi}{\partial x} & \dfrac{\partial f}{\partial x}\end{bmatrix}.$$

Для системы двух уравнений удалось получить обратную матрицу Якоби в аналитическом виде, что в общем случае $n$ уравнений не удается.

Далее, применяя формулу (9.4), получаем

$$z^{(k+1)}=z^{(k)}-W^{-1}(z^{(k)})F(z^{(k)}),$$

или в развернутом виде

$$x^{(k+1)}=x^{(k)}-\frac{1}{\det W^{(k)}}\left(f^{(k)}\frac{\partial \varphi^{(k)}}{\partial y}-\varphi^{(k)}\frac{\partial f^{(k)}}{\partial y}\right);$$

$$y^{(k+1)}=y^{(k)}-\frac{1}{\det W^{(k)}}\left(\varphi^{(k)}\frac{\partial f^{(k)}}{\partial x}-f^{(k)}\frac{\partial \varphi^{(k)}}{\partial x}\right).$$

Выражения справа сначала вычисляются при $x=x^{(0)}=a;\ y=y^{(0)}=b$. Дальнейшие уточнения решения продолжаются для $k=1,2,\ldots$ до выполнения условия (9.5).

## 9.2. Метод простой итерации

Систему исходных уравнений (9.1) представим в эквивалентном виде, удобном для проведения итераций:

$$x_1=\varphi_1(x_1,x_2,\ldots,x_n);$$

$$x_2=\varphi_2(x_1,x_2,\ldots,x_n);$$

$$\ldots\ldots\ldots\ldots\ldots;$$

$$x_n=\varphi_n(x_1,x_2,\ldots,x_n). \qquad (9.6)$$

Если известно какое-то приближенное значение решения системы уравнений: $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$, то последующие приближения решения можно вычислить методом простой итерации по алгоритму:

$$x_1^{(k+1)} = \varphi_1(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)});$$

$$x_2^{(k+1)} = \varphi_2(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)});$$

$$\ldots\ldots\ldots\ldots\ldots\ldots;$$

$$x_n^{(k+1)} = \varphi_n(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}), \quad k = 1, 2, \ldots, \tag{9.7}$$

либо методом, похожим на метод Гаусса–Зейделя для линейных систем алгебраических уравнений, с использованием в текущей итерации ранее вычисленных компонент вектора:

$$x_1^{(k+1)} = \varphi_1(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)});$$

$$x_2^{(k+1)} = \varphi_2(x_1^{(k+1)}, x_2^{(k)}, \ldots, x_n^{(k)});$$

$$x_3^{(k+1)} = \varphi_3(x_1^{(k+1)}, x_2^{(k+1)}, x_3^{(k)}, \ldots, x_n^{(k)});$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots;$$

$$x_n^{(k+1)} = \varphi_n(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, x_n^{(k)}), \quad k = 1, 2, \ldots.$$

Итерационный процесс продолжается до выполнения условия сходимости итераций (9.5).

Введя векторы $x = (x_1, x_2, \ldots, x_n)^T$ и $\Phi(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x))^T$, систему уравнений (9.6) можно записать кратко в виде

$$x = \Phi(x),$$

а итерационный алгоритм (9.7) в виде

$$x^{(k+1)} = \Phi(x^{(k)}), \quad k = 1, 2, \ldots.$$

Если данный итерационный процесс сходится, то он сходится к решению системы уравнений.

Для сходимости итерационного процесса необходимо, чтобы норма производной вектор-функции $\Phi(x)$ была меньше некоторого положительного числа $q < 1$ в некоторой окрестности $Q$ области решения системы, из которой не выходят приближенные значения $x^{(k)}$ при итерации

$$\left\|\Phi'(x)\right\| \le q < 1, \quad x \in Q. \tag{9.8}$$

Метод простой итерации может быть применен и к общей системе уравнений

$$F(x) = 0\,.$$

Для этого перепишем ее в эквивалентном виде:

$$x = x + \Lambda F(x)\,, \tag{9.9}$$

где $\Lambda$ – неособенная матрица. Обозначив $\Phi(x) = x + \Lambda F(x)$, получим

$$x = \Phi(x)\,.$$

Значение $\Lambda$ определим из условия (9.8) сходимости итераций, то есть из условия малости нормы $\|\Phi'(x)\|$:

$$\Phi'(x) = E + \Lambda F'(x);$$

$$\Phi'(x^{(0)}) = E + \Lambda F'(x^{(0)}) = 0\,,$$

где $x^{(0)}$ – начальное приближение решения. Если матрица Якоби $F'(x^{(0)}) = W(x^{(0)})$ неособенная, то

$$\Lambda = -W^{-1}(x^{(0)})\,,$$

и итерационная формула (9.9) принимает вид

$$x^{(k+1)} = x^{(k)} - W^{-1}(x^{(0)})F(x^{(k)})\,.$$

Эта формула соответствует алгоритму модифицированного метода Ньютона, когда матрица Якоби $W(x)$ обращается только один раз при $x = x^{(0)}$.

В заключение отметим, что скорость сходимости итерационных методов во многом зависит от близости выбора начальных данных к решению задачи. При неудачном выборе начальных данных итерационный процесс может не сойтись.

# 10. СОБСТВЕННЫЕ ЗНАЧЕНИЯ И СОБСТВЕННЫЕ ВЕКТОРЫ МАТРИЦ

## 10.1. Основные понятия

При решении многих теоретических и прикладных задач возникает необходимость определения собственных значений и собственных векторов матриц.

Собственным значением матрицы $A$, называется такое число $\lambda$, которое удовлетворяет уравнению

$$Ax = \lambda x, \tag{10.1}$$

а вектор $x$, соответствующий данному собственному значению $\lambda$ и удовлетворяющий уравнению (10.1), называется собственным вектором матрицы $A$.

Перенеся неизвестные уравнения (10.1) в левую часть, получим

$$(A - \lambda E)x = 0. \tag{10.2}$$

Это однородная система линейных уравнений. Она имеет ненулевое решение $x$ лишь при условии

$$\det(A - \lambda E) = 0. \tag{10.3}$$

Матрица $(A - \lambda E)$ называется характеристической и имеет вид

$$A - \lambda E = \begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{bmatrix}.$$

Определитель $\det(a - \lambda E)$ называется характеристическим определителем. В развернутом виде этот определитель есть многочлен $n$-й степени от $\lambda$ и называется характеристическим многочленом. Корни этого многочлена являются собственными значениями матрицы $A$, иначе называемые характеристическими числами многочлена

$$\det(A - \lambda E) = (-1)^n [\lambda^n - p_1 \lambda^{n-1} + p_2 \lambda^{n-2} - \dots + (-1)^n p_n]. \tag{10.4}$$

Числа $p_1, p_2, \dots, p_n$ называются коэффициентами характеристического многочлена.

Таким образом, собственные значения матрицы $A$ определяются из уравнения (10.3) или (10.4), а соответствующие им собственные векторы $x$ – из уравнения (10.2), которое можно записать в виде

$$\begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Эта система уравнений называется характеристической. Ее решение неединственно, так как обычно одно уравнение является линейной комбинацией других. Решения получаются с точностью до множителя. Чтобы избавиться от многозначности, собственные векторы нормируют – делят компоненты вектора на какую-либо его норму – или одну какую-то его компоненту при

нимают равной единице. Тогда остальные компоненты определяются однозначно.

*Пример*. Найти собственные числа и векторы матрицы

$$A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}.$$

*Решение*. Составим характеристический многочлен

$$\begin{vmatrix} 1-\lambda & 2 \\ 4 & 3-\lambda \end{vmatrix} = (1-\lambda)(3-\lambda) - 8 = \lambda^2 - 4\lambda - 5 = 0,$$

$$\lambda = (4 \pm \sqrt{16+20}\,)/2 = (4 \pm 6)/2;$$

$$\lambda_1 = 5; \quad \lambda_2 = -1.$$

Вычислим собственные векторы $X_1$ и $X_2$, соответствующие собственным числам $\lambda_1$ и $\lambda_2$. При $\lambda_1 = 5$ получаем систему $Ax = \lambda x$:

$$\begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5x_1 \\ 5x_2 \end{bmatrix} \Rightarrow \begin{matrix} x_1 + 2x_2 = 5x_1; \\ 4x_1 + 3x_2 = 5x_2; \end{matrix} \Rightarrow \begin{matrix} 2x_1 - x_2 = 0; \\ 2x_1 - x_2 = 0. \end{matrix}$$

Получаются линейно зависимые уравнения, из которых следует оставить только одно. Полагая $x_1 = 1$, получаем для второй компоненты $x_2 = 2x_1 = 2$. Собственный вектор $X_1$ имеет вид $X_1 = [1,2]^T$ или $X_1 = e_1 + 2e_2$, где $e_1, e_2$ – единичные базисные векторы некоторого выбранного базиса.

Аналогично находим собственный вектор $X_2$, соответствующий $\lambda_2 = -1$, и решаем систему уравнений:

$$\begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -x_1 \\ -x_2 \end{bmatrix} \Rightarrow \begin{matrix} x_1 + 2x_2 = -x_1; \\ 4x_1 + 3x_2 = -x_2; \end{matrix} \Rightarrow \begin{matrix} x_1 + x_2 = 0; \\ x_1 + x_2 = 0. \end{matrix}$$

Откуда получаем при $x_1 = 1$ $x_2 = -1$ и собственный вектор $X_2 = [1,-1]^T$, или $X_2 = e_1 - e_2$.

Вектор $X_2$ нормирован. Можно нормировать и вектор $X_1$, разделив его компоненты на наибольшую из них: $X_1 = e_1/2 + e_2$. Можно также привести векторы к единичной длине, разделив их компоненты на длину вектора $|X| = \sqrt{x_1^2 + x_2^2}$:

$$X_1 = \frac{1}{\sqrt{5}}(e_1 + 2e_2); \quad X_2 = \frac{1}{\sqrt{2}}(e_1 - e_2).$$

Аналогичные решения можно получить для матриц 3-го порядка. Для матриц более высокого порядка эта задача существенно осложняется.

Задачи определения собственных значений и собственных векторов матриц обычно подразделяется на две:

1) задачу определения всех собственных значений и принадлежащих им собственных векторов матриц, называемую полной проблемой собственных значений;

2) задачу определения одного или нескольких собственных значений и принадлежащих им собственных векторов, называемую частичной проблемой собственных значений.

Задача отыскания собственных значений осложняется трудностями вычисления коэфициентов характеристического полинома, необходимостью вычисления корней нелинейного уравнения высокого порядка, частым наличием среди собственных значений кратных.

Некоторые свойства собственных значений матрицы:

1) все собственные значения симметричной матрицы действительны;

2) если собственные значения матрицы действительны и различны, то соответствующие им собственные векторы ортогональны и образуют базис рассматриваемого пространства;

3) если две матрицы $A$ и $B$ подобны, то есть если они связаны соотношением подобия $B = P^{-1}AP$, то их собственные значения совпадают, $P$ – некоторая матрица.

Преобразованием подобия можно воспользоваться для упрощения исходной матрицы, чтобы свести задачу о вычислении ее собственных значений к аналогичной задаче для более простой матрицы.

Самым лучшим упрощением матрицы было бы приведение ее к треугольному виду

$$C = \begin{bmatrix} a'_{11} & a'_{12} & \dots & a'_{1n} \\ 0 & a'_{22} & \dots & a'_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a'_{nn} \end{bmatrix}.$$

Тогда определитель этой матрицы равен произведению диагональных элементов. В этом случае характеристический полином имеет вид

$$\det C = (a'_{11} - \lambda)(a'_{22} - \lambda)\dots(a'_{nn} - \lambda),$$

и собственные значения матрицы определяются непосредственно:

$$\lambda_1 = a'_{11}; \quad \lambda_2 = a'_{22}; \quad \dots; \quad \lambda_n = a'_{nn}.$$

Таким образом, собственные значения треугольной матрицы равны ее диагональным элементам. Однако лишь некоторые типы матриц можно привести к треугольному виду с помощью преобразования подобия.

Рассмотрим простейшие методы решения полной проблемы собственных значений.

## 10.2. Метод непосредственного развертывания

Рассмотрим вначале метод на примере матрицы 3-го порядка:

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad \det(A-\lambda E)=\begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{vmatrix}.$$

Вычислим ее характеристический определитель:

$$\det(A-\lambda E)=\begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{vmatrix}=(-1)^3\left[\lambda^3-\lambda^2(a_{11}+a_{22}+a_{33})+\right.$$

$$\left.+\lambda\left\{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}+\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}+\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}\right\}-\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}\right],$$

или $\det(A-\lambda E)=(-1)^3(\lambda^3-p_1\lambda^2+p_2\lambda-p_3)=0$. Здесь коэффициент $p_1$ – сумма диагональных элементов матрицы $A$, называемый следом матрицы $\mathrm{Sp}\,A$:

$$p_1=\mathrm{Sp}\,A=a_{11}+a_{22}+a_{33},$$

коэффициент $p_2$ – сумма всех диагональных миноров второго порядка матрицы $A$:

$$p_2=\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}+\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}+\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}$$

(диагональными минорами второго, третьего и т. д. порядка называются миноры, элементы главных диагоналей которых являются элементами главной диагонали определителя $\det A$).

Чтобы развернуть характеристический определитель $\det(A-\lambda E)$ для матрицы порядка $n$ в многочлен $n$-й степени

$$D(\lambda) = (-1)^n [\lambda^n - p_1 \lambda^{n-1} + p_2 \lambda^{n-2} - \ldots + (-1)^n p_n],$$

необходимо коэффициенты $p_1, p_2, \ldots, p_n$ вычислить по формулам:

$p_1 = \sum_{i=1}^{n} a_{ii} = \mathrm{Sp}\, A$ – сумма всех диагональных элементов матрицы $A$;

$p_2 = \sum_{\alpha<\beta} \begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} \\ a_{\beta\alpha} & a_{\beta\beta} \end{vmatrix}$ – сумма всех диагональных миноров второго порядка;

$p_3 = \sum_{\alpha<\beta<\gamma} \begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} & a_{\alpha\gamma} \\ a_{\beta\alpha} & a_{\beta\beta} & a_{\beta\gamma} \\ a_{\gamma\alpha} & a_{\gamma\beta} & a_{\gamma\gamma} \end{vmatrix}$ – сумма всех диагональных миноров третьего порядка, и т. д.;

$p_n = \det A$ – определитель матрицы $A$.

Число диагональных миноров $k$-го порядка матрицы $A$ равно

$$C_n^k = \frac{n(n-1)\ldots(n-k+1)}{k!}, \quad k = 1, 2, \ldots, n.$$

## 10.3. Метод вращений Якоби

Наиболее просто полная проблема собственных значений решается для симметричных матриц $A$, так как все их сообственные значения действительны, а размерность собственного подпространства, принадлежащего собственному значению $\lambda$, совпадает с кратностью $\lambda$.

Для симметричных матриц всегда можно найти такую ортогональную матрицу $P$, что в результате преобразования подобия матрица $A$ сведется к подобной матрице $B$ диагонального вида

$$B = P^{-1} A P = P^T A P,$$

имеющей одинаковые с матрицей $A$ собственные значения $\lambda$. Причем собственными векторами матрицы $A$ будут столбцы матрицы $P$, собственными числами – диагональные элементы матрицы $B$.

Однако вид матрицы $P$ заранее неизвестен. Поэтому диагональную матрицу $B$ можно получить последовательным обнулением недиагональных элементов матрицы $A$ посредством линейных преобразований при помощи матриц элементарных вращений Якоби $T_{pq}$:

$$T_{pq} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & t_{pp} & 0 & \dots & t_{pq} & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & t_{qp} & 0 & \dots & t_{qq} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 \end{bmatrix},$$

где $t_{pp} = t_{qq} = \cos\varphi$; $t_{pq} = \sin\varphi$; $t_{qp} = -\sin\varphi$. Матрица элементарных вращений $T_{pq}$ ортогональна: $T_{pq}^T = T_{pq}^{-1}$, и ее норма равна единице: $\|T_{pq}\| = 1$.

Ортогональной матрицей называется такая матрица $P$, для которой транспонированная матрица $P^T$ равна обратной матрице $P^{-1}$: $P^T = P^{-1}$. Все собственные значения ортогональной матрицы по модулю равны единице.

Всегда можно подобрать угол $\varphi$ таким, чтобы в преобразованной матрице $A' = T_{pq}^T A T_{pq}$ элементы $a'_{pq} = a'_{qp}$ стали равными нулю. Действительно, в матрице $A'' = T_{pq}^T A$ меняются лишь элементы $p$-й и $q$-й строк:

$$a''_{pj} = a_{pj}\cos\varphi - a_{qj}\sin\varphi;$$

$$a''_{qj} = a_{pj}\sin\varphi + a_{qj}\cos\varphi; \quad a''_{ij} = a_{ij} \quad \text{для} \quad i \neq p, q.$$

В матрице $A' = A''T_{pq}$ меняются в сравнении с $A''$ только $p$-й и $q$-й столбцы:

$$a'_{ip} = a''_{ip}\cos\varphi - a''_{iq}\sin\varphi;$$

$$a'_{iq} = a''_{ip}\sin\varphi + a''_{iq}\cos\varphi; \quad a'_{ij} = a''_{ij} \quad \text{для} \quad j \neq p, q.$$

Объединяя обе части процесса преобразования, для элементов матрицы $A'$ на пересечении $p$-х и $q$-х строк и столбцов, получаем:

$$a'_{pp} = a_{pp}\cos^2\varphi - 2a_{pq}\cos\varphi\sin\varphi + a_{qq}\sin^2\varphi;$$

$$a'_{qq} = a_{pp}\sin^2\varphi + 2a_{pq}\cos\varphi\sin\varphi + a_{qq}\cos^2\varphi;$$

$$a'_{pq} = a'_{qp} = (a_{pp} - a_{qq})\cos\varphi\sin\varphi + a_{pq}(\cos^2\varphi - \sin^2\varphi).$$

Чтобы элементы $a'_{pq}$ и $a'_{qp}$ стали равными нулю, необходимо использовать угол $\varphi$, равный

$$\varphi = 0{,}5\text{arctg}[2a_{pq}/(a_{qq} - a_{pp})].$$

Повторяя в определенной последовательности операции вращения, можно все недиагональные элементы матрицы $A$ сделать пренебрежимо малыми, превратив ее в диагональную. Диагональные элементы полученной матрицы будут собственными значениями также исходной матрицы.

Однако уничтожить все недиагональные элементы за конечное число поворотов нельзя, поскольку при последующих поворотах ранее уничтоженные элементы могут снова стать ненулевыми. Поэтому важное значение для эффективности метода имеет выбор последовательности обнуляемых элементов. Наиболее выгодным оказался способ уничтожения оптимального элемента, то есть наибольшего по модулю из элементов строки с максимальной суммой квадратов элементов

$$s_i = \sum_{j \neq i}^{n} a_{ij}^2, \quad i = 1, 2, \ldots, n.$$

В матричной форме итерационный метод элементарных вращений Якоби вычисления собственных значений и векторов матрицы $A$ может быть записан в следующем виде

$$A^{(1)} = T_1^T A T_1, \quad P_1 = T_1;$$

$$A^{(2)} = T_2^T A^{(1)} T_2, \quad P_2 = P_1 T_2;$$

$$A^{(3)} = T_3^T A^{(2)} T_3, \quad P_3 = P_2 T_3;$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots;$$

$$A^{(k)} = T_k^T A^{(k-1)} T_k, \quad P_k = P_{k-1} T_k.$$

Здесь $T_i$ – матрицы элементарных вращений; $P_i = \prod_j^i T_j$ – матрицы преобразования. Последовательности $A^{(k)}$ и $P_k$ сходятся, и пределами их являются диагональная матрица $B$, содержащая собственные значения матрицы $A$, и ортогональная матрица $P$, столбцы которой являются соответствующими собственными векторами матрицы $A$. Процесс итераций прекращается при выполнении условия

$$\sum \sum_{j \neq i} a_{ij}^2 \leq \varepsilon,$$

где $\varepsilon > 0$. Этот метод используется, когда важны точность, надежность и простота расчета и не существен объем вычислений.

## 10.4. Частичная проблема собственных значений

Для решения частичной проблемы собственных значений, когда определяют одно или несколько собственных значений и собственных векторов, обычно используют итерационные методы. Рассмотрим некоторые из них.

### 10.4.1. Метод простой итерации

Этот метод используется для вычисления наибольшего собственного значения матрицы *A* и соответствующего ему собственного вектора. Это следует из следующей *теоремы*. Если матрица *A* обладает простым превалирующим собственным значением $\lambda_1$, таким, что $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|$, то векторная последовательность

$$X^{(k)} = AX^{(k-1)} = A^k X^{(0)}, \quad k = 1, 2, \ldots,$$

где $X^{(0)}$ – задано, асимптотически сходится к собственному вектору $Y_1$, принадлежащему $\lambda_1$, $X^{(k)} \sim \lambda_1 c Y_1$ при условии, что начальный вектор $X^{(0)}$ обладает компонентой в направлении собственного вектора $Y_1$ ($c = \text{const}$). Скорость сходимости итерационного процесса пропорциональна отношению $|\lambda_1 / \lambda_2|$ в случае произвольных матриц и $|\lambda_1 / \lambda_2|^2$ – для симметричных.

Для обеспечения устойчивости метода и уменьшения погрешности расчетов необходимо нормировать итерируемый вектор $X^{(k)}$ на каждом шаге в процессе расчетов:

$$X_N^{(k)} = X^{(k)} / \left\| X^{(k)} \right\|,$$

где $\left\| X^{(k)} \right\| = \sqrt{\sum_i^n x_i^2}$ – норма вектора $X^{(k)}$, $x_i$ – компонента вектора $X^{(k)}$; $X_N^{(k)}$ – нормированный вектор, $i = 1, 2, \ldots, n$. В противном случае компоненты вектора $X^{(k)}$ увеличиваются примерно в $\lambda_1$ раз после каждой итерации, что в конечном счете может привести к переполнению разрядной сетки ЭВМ и увеличению погрешности расчета.

Алгоритм расчета в векторном виде записывается следующим образом:

$$X^{(1)} = AX_N^{(0)}; \quad X_N^{(1)} = X^{(1)} / \left\| X^{(1)} \right\|;$$

$$X^{(2)} = AX_N^{(1)}; \quad X_N^{(2)} = X^{(2)} / \left\| X^{(2)} \right\|;$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots;$$

$$X^{(k)} = AX_N^{(k-1)}; \quad X_N^{(k)} = X^{(k)} / \left\| X^{(k)} \right\|, \qquad (10.5)$$

$k = 1, 2, \ldots;$ $X_N^{(k)}$ – нормированный вектор.

Процесс итераций прекращается при выполнении условия

$$\left\| X_N^{(k)} - X_N^{(k-1)} \right\| \le \varepsilon,$$

где $\varepsilon$ – заданная погрешность вычисления.

После определения собственного вектора соответствующее ему собственное значение вычисляется по формуле Релея

$$\lambda_1 = X^T AX / (X^T X), \quad \text{где} \quad X = \lim_{k \to \infty} X^{(k)},$$

$$X^T X = [x_1, \ldots, x_n] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_i^n x_i^2;$$

$$X^T AX = [x_1, \ldots, x_n] \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_i^n \sum_j^n (x_i a_{ij} x_j).$$

*Вычисление минимального собственного значения матрицы.* В некоторых задачах нужно искать не наибольшее, а наименьшее собственное значение матрицы $A$.

Для этого умножим систему $\lambda x = Ax$ на обратную матрицу $A^{-1}$:

$$\lambda A^{-1} x = A^{-1} Ax,$$

а затем, разделив обе части уравнения на $\lambda$ и учитывая, что $A^{-1} A = E$, получаем

$$A^{-1} x = \lambda^{-1} x.$$

Эта задача отличается от рассмотренной выше тем, что здесь вычисляется наибольшее собственное значение $1/\lambda$ обратной матрицы $A^{-1}$, что достигается при наименьшем $\lambda$ для матрицы $A$, так как собственные значения матриц $A$ и $A^{-1}$ обратны друг другу. Следовательно, рассмотренный выше итерационный процесс (10.5) может быть использован также для нахождения наименьшего собственного значения обратной матрицы $A^{-1}$.

### 10.4.2. Метод одновременных итераций

Этот метод является обобщением изложенного выше метода простой итерации и применяется к группе ортогональных векторов. Метод позволяет определить последовательную группу собственных векторов $X_1, X_2, \ldots, X_m$ и соответствующие им собственные значения $\lambda_1, \lambda_2, \ldots, \lambda_m$ матрицы $A$, где $m < n$, $n$ – порядок матрицы $A$.

В качестве начального приближения векторов $X_i = X_i^{(0)}$ задается произвольная система $m$ ортогональных и нормированных векторов. Векторы $a$ и $b$ называются ортогональными, если их скалярное произведение равно нулю $(a,b) = 0$. Система векторов называется ортогональной, если все векторы системы попарно ортогональны между собой. Например, в качестве начальных для итерирования могут быть заданы векторы:

$$X_{1N}^{(0)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \quad X_{2N}^{(0)} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \quad \ldots; \quad X_{mN}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

Процесс одновременных итераций формируем в следующем виде:

$$X_1^{(k)} = AX_{1N}^{(k-1)};$$

$$X_2^{(k)} = AX_{2N}^{(k-1)};$$

$$\ldots\ldots\ldots\ldots;$$

$$X_m^{(k)} = AX_{mN}^{(k-1)};$$

$$(X_{1N}^{(k)}, X_{2N}^{(k)}, \ldots, X_{mN}^{(k)}) = \mathrm{ort}(X_1^{(k)}, X_2^{(k)}, \ldots, X_m^{(k)}).$$

Здесь $k = 1, 2, \ldots$ – номер итерации; $\mathrm{ort}(X_1, X_2, \ldots)$ – означает ортонормирование векторов. После каждой итерации векторы $X_i^{(k)}$, $i = 1, 2, \ldots, m$ нормируются и ортогонализируются. Если не ортогонализировать систему векторов, то процесс итераций приведет к одному и тому же собственному вектору для всей системы начальных векторов, соответствующему максимальному собственному значению.

Для ортогонализации и нормирования системы векторов $X_i^{(k)}$ обычно применяется алгоритм Грама–Шмидта. Первый вектор ортонормированной системы векторов получается нормированием первого вектора $X_1^{(k)}$:

$$X_{1N}^{(k)} = X_1^{(k)} \Big/ \left\| X_1^{(k)} \right\|,$$

где норма вектора вычисляется по формуле

$$\left\| X^{(k)} \right\| = \left( \sum_{i=1}^{n} x_i^2 \right)^{1/2}.$$

Затем вычисляется скалярное произведение векторов $X_2^{(k)}$ и $X_1^{(k)}$ и ортогонализируется второй вектор относительно первого, после чего полученный вектор нормируется:

$$c_{21} = \left( X_2^{(k)}, X_1^{(k)} \right) = \left( X_2^{(k)} \right)^T X_1^{(k)};$$

$$X_{2O}^{(k)} = X_2^{(k)} - c_{21} X_{1N}^{(k)};$$

$$X_{2N}^{(k)} = X_{2O}^{(k)} \Big/ \left\| X_{2O}^{(k)} \right\|.$$

Далее процесс ортонормирования повторяется для последующих векторов:

$$c_{31} = \left( X_3^{(k)} \right)^T X_1^{(k)}; \quad c_{32} = \left( X_3^{(k)} \right)^T X_2^{(k)};$$

$$X_{3O}^{(k)} = X_3^{(k)} - c_{31} X_{1N}^{(k)} - c_{32} X_{2N}^{(k)};$$

$$X_{3N}^{(k)} = X_{3O}^{(k)} \Big/ \left\| X_{3O}^{(k)} \right\|;$$

$$c_{m1} = \left( X_m^{(k)} \right)^T X_1^{(k)}; \quad c_{m2} = \left( X_m^{(k)} \right)^T X_2^{(k)}; \quad \ldots; \quad c_{m,m-1} = \left( X_m^{(k)} \right)^T X_{m-1}^{(k)};$$

$$X_{mO}^{(k)} = X_m^{(k)} - c_{m1} X_{1N}^{(k)} - c_{m2} X_{2N}^{(k)} - \ldots - c_{m,m-1} X_{(m-1)N}^{(k)};$$

$$X_{mN}^{(k)} = X_{mO}^{(k)} \Big/ \left\| X_{mO}^{(k)} \right\|.$$

Процесс итераций заканчивается при выполнении условия

$$\max_{i=1,2,\ldots,m} \left\| X_{iN}^{(k)} - X_{iN}^{(k-1)} \right\| \le \varepsilon,$$

где $\varepsilon$ – заданная погрешность вычисления собственного вектора. Скорость сходимости метода одновременных итераций определяется отношением $\left| \lambda_i / \lambda_{i+1} \right|$.

После вычисления собственных векторов определяются соответствующие им собственные значения матрицы $A$ из соотношения Релея

$$\lambda_i = X_i^T A X_i / (X_i^T X_i) .$$

## Литература


1. Бахвалов Н.С. Численные методы. – М.: Наука, 1973. – 632 с
2. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. – Численные методы. – М.: Наука, 1987. – 600 с.
3. Самарский А.А., Гулин А.В. Численные методы. – М.: Наука, 1989. – 432 с.
4. Амосов А.А., Дубинский Ю.А., Копченова Н.В. Вычислительные методы для инженеров. – М.: Высшая школа, 1994. – 544 с.
5. Тихонов А.Н., Костомаров Д.П. Вводные лекции по прикладной математике. – М.: Наука, 1984. – 190 с.
6. Калиткин Н.Н. Численные методы. – М.: Наука, 1978. – 512 с.
7. Волков Е.А. Численные методы. – М.: Наука, 1982. – 248 с.
8. Турчак Л.И. Основы численных методов. – М.: Наука, 1987. – 320 с.
9. Демидович Е.П., Марон И.А. Основы вычислительной математики. – М.: Наука, 1966. – 664 с.
10. Мак-Кракен Д., Дорн У. Численные методы и программирование на Фортране. – М.: Мир, 1977. – 584 с.
11. Кузнецов Ю.Н., Кузубов В.И., Волощенко А.Б. Математическое программирование. – М.: Высшая школа, 1980. – 350 с.
12. Боглаев Ю.П. Вычислительная математика и программирование. – М.: Высшая школа, 1990. – 544 с.


## Дополнительная литература


13. Годунов С.К., Рябенький В.С. Разностные схемы. – М.: Наука, 1977.
14. Хорн Р., Джонсон Ч. Матричный анализ. – М.: Мир, 1989.
15. Воеводин В.В. Вычислительные основы линейной алгебры. – М.: Наука, 1977.
16. Гантмахер Ф.Р. Теория матриц. – М.: Наука, 1988.
17. Де Бор К. Практическое руководство по сплайнам. – М.: Радио и связь, 1985.
18. Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов. – М.: Наука, 1986.-232 с.
19. Краснощеков П.С., Петров А.А. Принципы построения моделей. – М.: Изд-во МГУ, 1983.
20. Крылов В.И., Бобков В.В., Монастырский П.И. Вычислительные методы. – Т.1,2. – М.: Наука, 1976–1977.
21. Парлетт Б. Симметричная проблема собственных значений. – М.: Мир, 1983.
22. Писсанецки С. Технология разреженных матриц. – М.: Мир, 1988.

23. Икрамов Ч.Д. Вычислительные методы линейной алгебры. (Решение больших разреженных систем уравнений прямыми методами.) – М.: Знание, 1989.
24. Хейгеман Л., Янг Д. Прикладные итерационные методы. – М.: Мир, 1986.
25. Пустыльник Е.И. Статистические методы анализа и обработки наблюдений. – М.: Наука, 1975.
26. Стечкин С.Б., Субботин Ю.Н. Сплайны в вычислительной математике. – М.: Наука, 1976.
27. Уилкинсон Дж.Х. Алгебраическая проблема собственных значений. – М.: Наука, 1970.
28. Уилкинсон Дж.Х., Райнш К. Справочник алгоритмов на языке Алгол. Линейная алгебра. – М.: Машиностроение, 1976.
29. Фаддеев Д.К. Фаддеева В.Н. Вычислительные методы линейной алгебры. – М.: Физматгиз, 1963.
30. Фаддеев Д.К. Лекции по алгебре. – М.: Наука, 1984.
31. Хемминг Р.И. Численные методы для научных работников и инженеров. – М.: Мир, 1977.

## ОГЛАВЛЕНИЕ