Лекции по вычислительной математике (черновик)

Першин А.Ю.

Кафедра РК6 (Системы автоматизированного проектирования)
МГТУ им. Н.Э. Баумана

9 июня 2020 г.

Оглавление

Введение

1.1 Введение в курс

Исторически, математика появилась в первую очередь как ответ на необходимость измерять те или иные величины окружающего мира и производить над ними вычисления. Однако оказалось, что некоторые, даже сравнительно простые вычислительные задачи, не так легко решить. Например, вавилонская глиняная табличка YBC 7289 (1800-1600 до н.э.) содержит в себе сравнительно точное приближение иррационального числа $\sqrt{2} \approx 1.414222$, которое является длиной диагонали квадрата с единичной стороной. Для нахождения этой аппроксимации вавилоняне использовали одну из вариаций численного метода решения алгебраического нелинейного уравнения $x^2 = 2$, который сейчас известен как метод простой итерации (он будет пройдет в курсе). Таким образом уже древние вавилоняне столкнулись с фундаментальной проблемой, которая породила всё направление вычислительной математики – сформулировать математическую задачу еще не значит ее решить. Проходя сквозь века математическая наука совершенствовалась и научилась формулировать свои модели не только в виде алгебраических уравнений, но и в виде обыкновенных дифференциальных уравнений, уравнений в частных производных, интегральных уравнений и проч. Развивались и методы нахождения аналитического решения математических задач – например, Эйлер изобрел метод решения однородных линейных дифференциальных уравнений, а Фурье обнаружил, что некоторые линейные уравнений в частных производных могут быть решены с помощью бесконечных рядов косинусов и синусов. За вычетом нескольких исключительных случаев (например, корни многочленов до 4 степени включительно или уравнения Бернулли), математикам удавалось находить точные аналитические решения только для линейных уравнений. Однако, как показала практика, многие интересные физические явления описываются нелинейными уравнениями. Так, например, динамика вязкой несжимаемой жидкости описывается уравнением Навье–Стокса, которое за счет члена $\boldsymbol{u} \cdot \nabla \boldsymbol{u}$, где $\boldsymbol{u}$ – поле скоростей потока, является нелинейным уравнением в частных производных. Для этого уравнения, дополненного уравнением непрерывности, не только не найдено аналитического решения в общем виде, но даже не доказано, что оно в принципе существует и является гладким для гладких начальных условий (это доказательство является одной из «задач тысячелетия», сформулированных институтом Клэя). В таком случае ученым приходится полагаться исключительно на приближенные решения, полученные либо аналитически с помощью упрощения оригинального уравнения (такой подход используется, например, в асимптотических методах или слабонелинейном анализе), либо с помощью численных методов. Очевидно, что упрощение оригинального уравнения сильно

ограничивает валидность полученного решения, что и мотивирует активное использование численных методов при решении современных задач математической физики.

Величайшие ученые своего времени были озадачены вопросами численного решения математических задач – по мере прохождения курса мы встретим фамилии Ньютона, Лагранжа, Эйлера, Гаусса и многих других известных математиков. Нельзя не сказать о вкладе русских ученых в развитие численных методов – полиномы Чебышёва, подпространство Крылова, метод Галеркина, метод Годунова хорошо известны современным ученым по всему миру.

Глобально вычислительные методы можно разбить на три больших группы (по крайней мере в рамках нашего курса):

1. методы приближения (аппроксимации) дискретных данных, принадлежащих некоторой сложной, часто недоступной, функции, к сравнительно простым аналитическим функциям;

2. численные методы линейной и нелинейной алгебры.

3. вероятностные численные методы;

Первая группа исторически развивалась раньше двух других и была нацелена на нахождение решений нелинейных уравнений с помощью аппроксимации решения набором простых функций, что сводило нелинейную задачу к более простой линейной (под «сложной» функцией подразумевается сильно нелинейная или зашумленная функция). Линейная задача чаще всего выражена с помощью СЛАУ, и подавляющая часть методов второй группы посвящена точным и приближенным методам решения СЛАУ. Приближенные методы решения СЛАУ появились вследствие того, что с каждым днем размерность вычислительных задач N увеличивалась, и решение СЛАУ, например, методом Гаусса не всегда было реализуемо ввиду его большой сложности $O(N^3)$. Увеличение размерности задач так же мотивировало создание многих методов из третьей группы. Например, решение задачи многих тел в квантовой механике сводится к вычислению значения интеграла, кратность которого может достигать нескольких тысяч. В таких случаях вместо классических квадратур численного интегрирования используется метод Монте-Карло, где подынтегральная функция вычисляется в M случайных точках, сгенерированных с помощью заданного закона распределения, после чего их значения суммируются. Это позволяет уменьшить зависимость сложности от размерности задачи.

В данном курсе будут изучаться базовые численные методы, знание которых необходимо, чтобы перейти к более сложным методам, так как последние почти всегда содержат в себе базовые в той или иной форме. К примеру:

- в методе конечного элемента (МКЭ), используемого для численного решения уравнений в частных производных, функции формы аппроксимируют решение уравнения в элементе с помощью полиномов Лагранжа или Эрмита (классический МКЭ) или полиномов Чебышёва (метод спектральных элементов);

- при обучении нейронных сетей, локальная оптимизация сильно нелинейной целевой функции ошибки происходит с помощью вариации метода градиентного спуска (стохастический градиентный спуск).

Для успешного прохождения курса представляется важным освежить воспоминания из прошлых курсов:

- математический анализ:
  - понятия предела, непрерывности, производной и интеграла Римана;
  - теорема о среднем значении;
  - теорема о промежуточном значении;
  - теорема Ньютона-Лейбница;
  - ряды Тейлора;
- линейная алгебра:
  - понятия линейного нормированного пространства, матрицы, определителя, обратной матрицы;
  - собственные числа и собственные вектора матрицы;
  - свойства положительно определенных и симметричных матриц;
  - простейшие методы решения систем линейных алгебраических уравнений (СЛАУ);

Несмотря на то, что многие вопросы вычислительной математики удобно изъяснять в рамках функционального анализа, его знание не является обязательным, и понятия теории функциональных пространств будут даны в лекциях по необходимости.

Тем, кто заинтересован в более глубоком и всестороннем изучении численных методов, к прочтению рекомендуются следующие книги:

1. Численные методы. Н.Н. Калиткин. Главная редакция физико-математической литературы изд-ва «Наука», М., 1978.
   - *комментарий: оптимальный учебник для первого знакомства с численными методами (многие примеры и описания в курсе взяты из него).*
2. Численные методы / Н.С. Бахвалов, Н.П. Жидков, Г.М. Кобельков. – 7-е изд. – М. : БИНОМ. Лаборатория знаний, 2017. – 636 с. : ил. – (Классический университетский учебник).
   - *комментарий: полезен для тех случаев, когда необходимо разобраться со строгим теоретическим обоснованием численных методов.*
3. Элементы теории функций и функционального анализа. А.Н. Колмогоров, С.В. Фомин. Главная редакция физико-математической литературы изд-ва «Наука», М., 1976.
   - *комментарий: удобен для подробного теоретического знакомства с функциональными пространствами, интегралом Лебега и тригонометрическими рядами.*

За вычетом приведенных выше книг, при подготовке лекций активно использовались следующие источники на английском языке, которые также рекомендуются к ознакомлению:

1. Numerical Analysis, Ninth Edition, Richard L. Burden and J. Douglas Faires.

   - *комментарий: полный, подробный и вместе с тем доступный курс численного анализа.*

2. Analysis of Numerical Methods, Eugene Isaacson and Herbert Bishop Keller (1994).

   - *комментарий: курс численного анализа, ориентированный на строгое доказательное обоснование свойств численных методов.*

3. Introductory Functional Analysis with Applications, Erwin Kreyszig.

   - *комментарий: курс функционального анализа с подробным разбором примеров использования функциональных пространств в различных прикладных задачах.*

4. The Elements of Real Analysis, Robert G Bartle.

   - *комментарий: углубленный курс математического анализа функций вещественного переменного.*

5. Essential Topology, Martin D Crossley.

   - *комментарий: книга доступно излагает базовые понятия топологии, которая, будучи наукой о непрерывных преобразованиях, представляется полезной для освоения в контексте курса численных методов.*

6. Private lecture notes by Prof Mark Kelmanson and Dr Kevin Gouder.

   - *комментарий: эти лекции отсутствуют в открытом доступе, однако соответствующие конспекты представляются ценными и могут быть предоставлены по запросу.*

1.2 Моделирование объектов исследования

Конечной целью прикладной математики, а следовательно и вычислительной математики как ее подраздела, является моделирование некоторых объектов и процессов окружающего мира (далее – объектов исследования). В математике под моделированием подразумевается описание самых важных и существенных свойств объекта исследования в математических понятиях. Допустим, например, что объектом исследования является футбольный мяч. Если из всех его свойств нам больше всего интересна геометрическая форма, то его моделью будет сфера как математический объект. Однако, если нас интересует его деформация при сжатии руками (в этом случае объектом исследования формально является процесс сжатия мяча, а не сам мяч), то модель будет сформулирована как система дифференциальных и алгебраических уравнений статической задачи линейной теории упругости. В обоих случаях полученная модель называется *математической моделью* объекта исследования.

Математические модели зачастую настолько сложны, что для них не существует способа нахождения аналитического (точного) решения (“сложными” могут быть как разрешающие уравнения, так и граничные условия). В таком случае может ли сама математическая

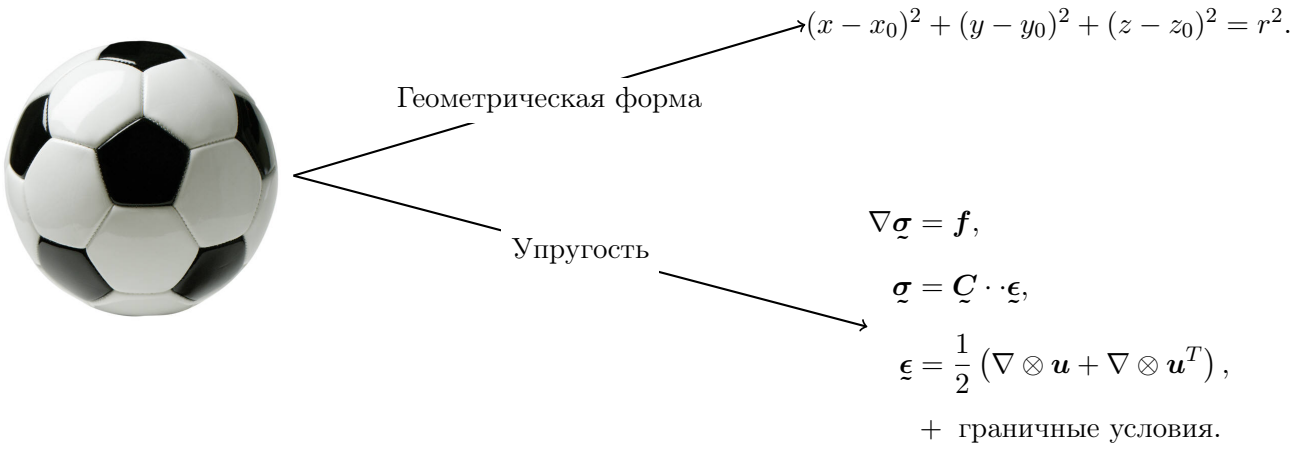

Рисунок 1.1 – Моделирование различных свойств мяча.

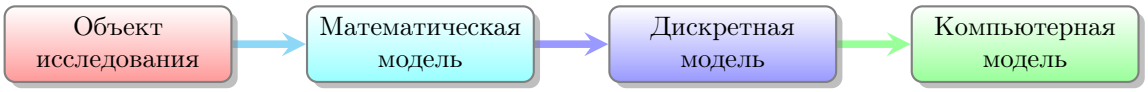

Рисунок 1.2 – Классические этапы моделирования некоторого объекта исследования в научных и инженерных задачах.

модель быть объектом исследования? Безусловно, так как она в свою очередь тоже может быть смоделирована. К примеру, система уравнений упругости, приведенная на рисунке **??**, может быть дискретизирована с помощью метода конечных элементов. Полученная модель будет представлять собой алгоритм по сборке СЛАУ и саму СЛАУ. Мы будем называть подобную модель *дискретной моделью*. Принципиально, точное решение такой модели может быть найдено, однако для этого потребуется неоправданно большое количество времени и людей (**[TODO: ссылка на Льюиса Ричардсона]**). В современном мире для нахождения решений дискретных моделей используются компьютеры, что вынуждает трансформировать дискретную модель в компьютерную программу, называемую также *компьютерной моделью*. Результирующая диаграмма подобного моделирования показана на рисунке **??**. Заметим, что решения всех трех моделей в общем случае не равны и могут лишь сходиться друг к другу при определенных обстоятельствах.

1.3 Приближенный анализ

1.3.1 Источники погрешностей

Очевидно, что упрощения, осуществляемые на каждом шаге моделирования, вносят определенную неточность в решение относительно изначального объекта исследования, которую называют *погрешностью*. Принято выделять четыре вида погрешностей: неустра-

нимая погрешность, погрешность математической модели, погрешность метода и вычислительная погрешность. Дадим им определения.

Определение 1.3.1. *Неустранимой погрешностью называют неточность при задании исходных данных.*

Пример 1.3.1. *Допустим, вы бросаете шар с Пизанской башни. В независимости от того, как вы моделируете законы природы, для того, чтобы рассчитать скорость шара в момент касания поверхности Земли, вам необходимо знать высоту, с которой вы отпускаете шар в свободное падение. В случае, если вы измеряете высоту рулеткой с миллиметровой шкалой, погрешность измерения составляет $O(10^{-3})$ метра. Тогда, даже если вы способны идеально точно предсказывать скорость шара, погрешность в $O(10^{-3})$ метра устранить не удастся.*

Определение 1.3.2. *Погрешностью математической модели называют неточность при описании реального объекта математическими понятиями.*

Пример 1.3.2. *Математическая модель, описывающая движение тела в свободном падении, выражается простым уравнением Ньютона*

$$m\frac{d^2x}{dt^2} = -mg \tag{1.1}$$

где m – масса тела, g – ускорение свободного падения, $x(t)$ – высота тела в момент времени t. Применительно к земной атмосфере, погрешность этой математической модели возникает в результате пренебрежения сопротивлением воздуха. Более точной моделью была бы:

$$m\frac{d^2x}{dt^2} = -mg - Asign\left(\frac{dx}{dt}\right)\frac{dx}{dt}^2 \tag{1.2}$$

Очевидно, что погрешность растет пропорционально квадрату скорости, что делает уравнение (1.1) неприменимым, скажем, для моделирования движения спускаемого аппарата сквозь атмосферу.

Определение 1.3.3. *Погрешностью метода называют неточность при замене математической модели приближенной.*

Пример 1.3.3. *Уравнения 1.1 и 1.2 имеют аналитические решения, однако в случае более сложной правой части нам необходимо будет использовать один из численных методов решения обыкновенных дифференциальных уравнений (ОДУ), например, метод Рунге-Кутта, который мы будем проходить в курсе. В таком случае погрешностью метода будет разница между точным решением и решением уравнения, дискретизированного методом Рунге-Кутта.*

Определение 1.3.4. *Вычислительной погрешностью называют погрешность математических операций, производимых компьютером.*

С математической точки зрения вычислительная погрешность появляется в результате того, что при программировании численного метода, алгебраические структуры которого определены для поля вещественных чисел $\mathbb{R}$, мы неявно заменяем их на алгебраические структуры, определенные для поля рациональных чисел $\mathbb{Q}$. Это связано с формой представления (формально, приближения) вещественных чисел в памяти компьютера. Стандарт IEEE 754-2008 гласит, что 64-битное представление вещественного числа состоит из последовательно расположенных бита знака s, 11 бит экспоненты (порядка) c и 52 бит мантиссы f, а само значение числа вычисляется как

$$(-1)^s 2^{c-1023}(1+f), \tag{1.3}$$

где $f = \sum_{i=1}^{52} 2^{-b_i}$ и $b_i \in \{0,1\}$ – i-ый бит мантиссы.

Возникающую погрешность при вычислительных операциях часто называют погрешностью округления.

Так как процесс моделирования реального объекта является последовательным (сначала составляется математическая модель, затем дискретная модель, а после компьютерная), логично заключить, что погрешность, вносимая на очередном этапе моделирования не должна быть меньше, чем погрешность предыдущих этапов.

Пример 1.3.4. *Вернемся к примеру с бросанием тела с Пизанской башни. Допустим, мы используем такой численный метод для интегрирования уравнений движения тела, что погрешность вычисляемой высоты тела получилась $O(10^{-5})$ метра. Очевидно, что при погрешности линейки, с помощью которой мы измеряли начальную высоту, равной $O(10^{-3})$ метра эти усилия окажутся напрасными.*

1.3.2 Абсолютная и относительная погрешности

Определим, что мы формально называем погрешностями.

Определение 1.3.5. *Абсолютной погрешностью приближенного значения a^* называют величину $\Delta(a^*)$, которая определена как*

$$\Delta(a^*) = |a - a^*|, \tag{1.4}$$

где a – точное значение.

Число a записывают с учетом абсолютной погрешности в следующей форме:

$$a = a^* \pm \Delta(a^*). \tag{1.5}$$

Определение 1.3.6. *Относительной погрешностью приближенного значения a^* называют величину $\delta(a^*)$, которая определена как*

$$\delta(a^*) = \left|\frac{a - a^*}{a}\right|, \tag{1.6}$$

где a – точное значение.

Число a записывают с учетом относительной погрешности в следующей форме:

$$a = a^*(1 \pm \delta(a^*)). \tag{1.7}$$

1.3.3 Некоторые понятия функционального анализа

Понятия абсолютной и относительной погрешности определены для некоторых приближенных величин, которые неявно предполагаются принадлежащими одному из числовых множеств (чаще всего $\mathbb{R}$ или $\mathbb{Q}$). Однако понятие близости распространяется на куда более широкий класс множеств и, в частности, на множества функций, что особенно важно при доказательстве сходимости численных методов. Соответствующей обобщенной мерой "близости" называется метрика.

Определение 1.3.7. *Множество X называется метрическим пространством, если на нем определена функция $\rho : X \times X \longrightarrow \mathbb{R}$, называемая метрикой или расстоянием, для которой выполняются следующие аксиомы:*

1. $\rho(x_1, x_2) \geq 0$,
2. $\rho(x_1, x_2) = 0 \Longleftrightarrow x_1 = x_2$,
3. $\rho(x_1, x_2) = \rho(x_2, x_1)$,
4. $\rho(x_1, x_3) \leq \rho(x_1, x_2) + \rho(x_2, x_3)$ *(аксиома треугольника),*

Ремарка 1.3.1. *Метрическим пространством чаще называют пару (X, ρ).*

Определение 1.3.8. *Последовательность $\{x_n\}_{n=1}^{\infty}$ метрического пространства (X, ρ) называется сходящейся к элементу $x \in X$, если для нее верно*

$$n \to \infty \Longrightarrow \rho(x_n, x) \to 0.$$

Определение 1.3.9. *Последовательность $\{x_n\}_{n=1}^{\infty}$ метрического пространства (X, ρ) называется фундаментальной, если $\forall n > 1, \epsilon > 0 \ \exists k(\epsilon) : \rho(x_n, x_m) < \epsilon, m > k$*

Ремарка 1.3.2. *Фундаментальную последовательность также называют сходящейся в себе последовательностью или последовательностью Коши.*

Определение 1.3.10. *Метрическое пространство называется полным, если любая последовательность его элементов сходится к элементу того же пространства.*

Ремарка 1.3.3. *Сложность при анализе вычислительной погрешности в частности возникает из-за того, что множество рациональных чисел $\mathbb{Q}$ является неполным. К примеру, последовательность*

$$x_n = \left(1 + \frac{1}{k}\right)^k$$

сходится к $e \in \mathbb{R}$.

На протяжении всего курса мы практически всегда будем иметь дело с элементами линейных (векторных) пространств, т.е. множеств с определенными для них операциями сложения и умножения на число. Для линейных пространств в качестве метрики часто выбирают норму, что делает его линейным нормированным пространством.

Определение 1.3.11. *Линейным нормированным пространством называется пара* $(X, ||\cdot||)$*, где* X *– линейное пространство, а* $||\cdot|| : X \longrightarrow \mathbb{R}$ *– норма, удовлятворяющая следующим аксиомам:*

1. $||x|| \geq 0$,
2. $||x|| = 0 \Longleftrightarrow x = \mathbf{0}$,
3. $||\lambda x|| = |\lambda| \cdot ||x||$,
4. $||x_1 + x_2|| \leq ||x_1|| + ||x_2||$,

где $x \in X$.

Легко показать, что линейное нормированное пространство является метрическим пространством с метрикой $\rho(x_1, x_2) = ||x_1 - x_2||$.

Определение 1.3.12. *Банаховым пространством называется полное линейное нормированное пространство.*

Самым очевидным примером банахова пространства является $(\mathbb{R}, |\cdot|)$, где нормой является модуль числа. Так как банаховы пространства и соответствующие им нормы играют важную роль в численных методах, мы рассмотрим внимательно несколько самых важных банаховых пространств.

Конечномерные нормированные пространства

Определение 1.3.13. *Конечномерным нормированным пространством* $l_p^{(n)}$ *называется пара* $(X, ||\cdot||_p)$*, где* X *– множество векторов* $\boldsymbol{x} = (x_1, ..., x_n)$ *в* n*-мерном линейном пространстве и норма определена функцией*

$$||\boldsymbol{x}||_p = \left(\frac{1}{n} \sum_{i=1}^{n} |x_i|^p \right)^{1/p}. \tag{1.8}$$

Для простоты будем предполагать, что n-мерное линейное пространство определенно над полем вещественных или комплексных чисел. Эвклидово пространство $\mathbb{R}^n$ является частным случаем конечномерного нормированного пространства: $l_2^{(n)}$. Если индекс нормы опущен, то предполагается, что используется классическая эвклидова норма, т.е. $||\cdot|| = ||\cdot||_2$.

Важно отметить, что сходимость в одной из норм $||\cdot||_p$ гарантирует сходимость во всех остальных нормах этого типа. Если последовательность векторов $\{\boldsymbol{x_m}\}_{n=1}^{\infty}$ не сходится, но сходится последовательность

$$\left\{ \frac{\boldsymbol{x_m}}{||\boldsymbol{x_m}||} \right\}_{n=1}^{\infty},$$

то говорят о сходимости по направлению.

Бесконечномерные нормированные пространства

Логичным расширением случая конечномерных нормированных пространств является аналогичное пространство с бесконечной размерностью. Заметим, что в таком случае множества, образуемые векторами, должны оставаться счетными.

Определение 1.3.14. *Бесконечномерным нормированным пространством l_p называется пара $(X, ||\cdot||_p)$, где X – множество векторов $\boldsymbol{x} = (x_1, x_2, ...)$, каждый из которых является в свою очередь счетным множеством, и норма определена функцией*

$$||\boldsymbol{x}||_p = \lim_{nto\infty} \left(\frac{1}{n} \sum_{i=1}^{n} |x_i|^p \right)^{1/p}. \tag{1.9}$$

Лебегово пространство

Следующий шаг состоит в переходе от счетных векторов к несчетным. Такой шаг не является излишним теоретизированием, так как функции, с которыми мы чаще всего имеет дело, являются бесконечномерными векторами с несчетным числом элементов. Это утверждение звучит контринтуитивно из-за терминологической путаницы, которая вносится понятием мерности пространства. Предположим, что у нас есть функция $d(x, y, z)$, вычисляющая расстояние от начала координат до точки в трехмерном пространстве:

$$d(x, y, z) = \sqrt{x^2 + y^2 + z^2}. \tag{1.10}$$

С точки зрения теории множеств, эта функция задана на трехмерном пространстве и отображает его на положительную вещественную ось: $f : \mathbb{R}^3 \longrightarrow \mathbb{R}_+$. Однако если мы рассматриваем саму функцию как элемент некоторого множества, то выясняется, что она является бесконечномерным вектором с несчетным числом элементов. Действительно, рассмотрим следующий $(n + 1)$-мерный вектор:

$$\boldsymbol{f} = [f_0, f_1, \ldots, f_n]^T, \tag{1.11}$$

где $f_j = \sin \frac{2\pi j}{n}$ и $j = 0, \ldots, n$. При $n \to \infty$ вектор $\boldsymbol{f}$ стремится к функции $\sin x$ на отрезке $[0; 2\pi]$, что демонстрируется на рисунке 1.3. Учитывая, что отрезок $[0; 2\pi]$ является несчетным множеством, соответствующий бесконечномерный вектор, восстанавливающий функцию $\sin x$, так же будет несчетным множеством.

В таком случае естественное обобщение предела суммы в норме (1.9) до интеграла приводит к лебеговым пространствам L_p.

Определение 1.3.15. *Лебеговым пространством L_p называется пара $(F, ||\cdot||_p)$, где F – множество функций $x(t)$, p-я степень которых интегрируема на отрезке $[a, b]$, и норма определена функционалом*

$$||x(t)||_p = \left(\int_a^b |x(t)|^p dt \right)^{1/p}. \tag{1.12}$$

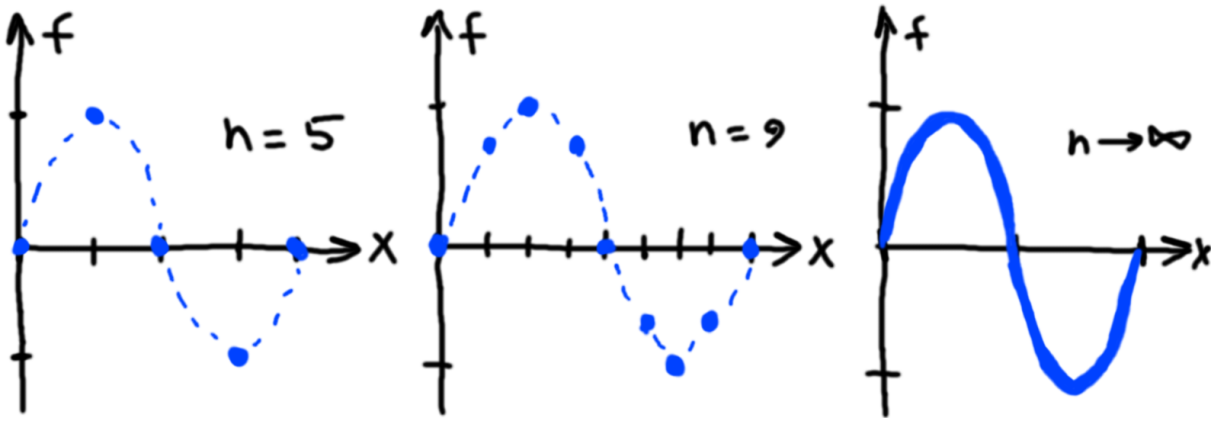

Рисунок 1.3 – Иллюстрация функции $f(x) = \sin x$ как бесконечномерного вектора с несчетным числом элементов.

Пространство L_2 называют гильбертовым, а его норму $||\cdot||_2$ среднеквадратичной. Норму $||x(t)||_\infty = \max_{t\in[a;b]} |x(t)|$ называют равномерной или чебышевской. Несложно доказать, что для норм верны следующие соотношения:

$$||x(t)||_1 \leq ||x(t)||_2 \leq ... \leq ||x(t)||_\infty, \tag{1.13}$$

что означает, что из равномерной сходимости (т.е. сходимости в норме $||x(t)||_\infty$) следует сходимость в среднем (т.е. сходимость в норме $||x(t)||_p, p < \infty$), однако обратное не является в общем случае верным.

На рисунке 1.4 демонстрируется сходимость последовательности функций $\{f_1(x), f_2(x), f_3(x), \dots\}$ к некоторой функции $f(x)$. На левом рисунке последовательность функций сходится равномерно и, следовательно, в среднем, в то время как на правом рисунке последовательность функций сходится только в среднем: «пик», формируемый $f_i(x)$ при $i \to \infty$, не окажет влияния на значение интеграла в норме $||\cdot||_2$, но при этом приведет к ненулевому значению $||f_i(x)||_\infty = \max_{x\in[a;b]} |f_i(x)|$.

Вопрос равномерной и средней сходимости имеет прикладное значение, так как если некоторый численный метод сходится в среднем, но не сходится равномерно, это означает, что численное решение может включать в себя паразитное решение, например, в форме паразитных осцилляций. Сходимость в среднем гарантирует, что паразитные осцилляции имеют меру нуль (т.е. определены в одной точке и нигде больше) при бесконечно малом размере сетки. Однако, так как размер сетки всегда конечен, они появятся в численном решении задачи и могут сделать его неудовлетворительным. В курсе мы встретимся с примерами паразитных осцилляций.

Пространство непрерывных функций

Так как класс непрерывных функций играет важную роль в уравнениях математической физики, рассмотрим их отдельно.

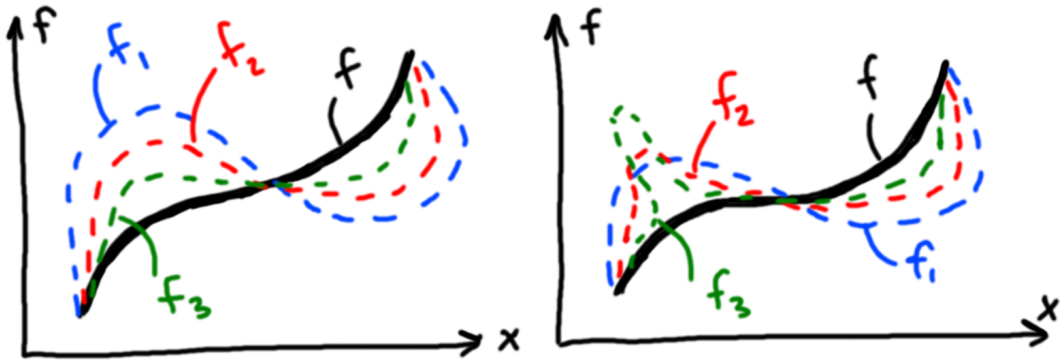

Рисунок 1.4 – Иллюстрация равномерной сходимости (левый рисунок) и сходимости только в среднем (правый рисунок).

Определение 1.3.16. *Пространством непрерывных функций $C[a,b]$ называется пара $(F, ||\cdot||_C)$, где F – множество функций, непрерывных на отрезке $[a,b]$, а $||x||_C = \max_{t\in[a,b]} |x(t)|$ – равномерная норма.*

Пространство непрерывных функций также обозначают $C^0[a,b]$. В свою очередь пространство $C^p[a,b]$ называют пространство функций, p-я производная которых непрерывна. Гладкой называют функцию, имеющую непрерывные производные (порядок последней непрерывной производной часто опускают, подразумевая, что функция достаточно гладкая для решения данной задачи). Бесконечно гладкие функции называют аналитическими – такие функции можно представить в виде бесконечной суммы ряда Тейлора.

Определение 1.3.17. *Функция $x(t)$ называется равномерно-непрерывной на заданном отрезке, если $\forall\epsilon > 0\ \exists\delta = \delta(\epsilon) : |x(t_1) - x(t_2)| \le \epsilon, |t_1 - t_2| \le \delta$.*

Определение 1.3.18. *Функция $x(t)$ называется липшиц-непрерывной, если $\epsilon \le K\delta$, что эквивалентно $|x(t_1) - x(t_2)| \le K|t_1 - t_2|$, где K называется константой Липшица.*

Если функция имеет ограниченную производную, т.е. $|x'(t)| \le K$, то она липщиц-непрерывна, причем точная верхняя грань модуля производной равна константе Липшица K. Липшиц-непрерывные функции играют важную роль в теории дифференциальных уравнений, так как если функция f в уравнении $x'(t) = f(t,x)$ является липшиц-непрерывной по переменной x (т.е. имеет ограниченную частную производную по x), то из этого следует существование и единственность решения уравнения.

Интерполяция

2.1 Основные понятия теории приближений

Предположим, что существует некоторая функция $f(x)$, определенная на отрезке $x \in [a; b]$, при том, что ее аналитическое выражение нам неизвестно. Например, $f(x)$ может быть как гладким решением какого-то нелинейного дифференциального уравнения, так и быть кривой вариации биржевого курса. Мы также предполагаем, что нам известны значения $f(x)$ для некоторых x:

- в случае дифференциальных уравнений, мы можем разбить $[a; b]$ на $n - 1$ отрезков и предположить, что нам известны значения $f(x)$ в узлах;
- в случае кривой вариации биржевого курса, этими значениями является временной ряд, предоставляемый биржей.

Задача *приближения* или *аппроксимации* (эти термины синонимичны) состоит в представлении недоступной функции $f(x)$ в виде более простой аналитической функции $\widetilde{f}(x)$ по n значениям $f(x_i)$, где $x_i \in [a; b]$ называются *узлами*. Для решения этой задачи мы задаемся некоторой параметризированной формой для $\widetilde{f}(x; \boldsymbol{c})$, где $\boldsymbol{c}$ – вектор параметров, а затем подбираем $\boldsymbol{c}$ так, что отклонение $\rho[f(x) - \widetilde{f}(x; \boldsymbol{c})]$ минимизировано, где ρ – некоторая метрика, определяющая оценку погрешности приближения. Частными случаями аппроксимации являются интерполяция и экстраполяция.

Определение 2.1.1. *Интерполяцией называется приближение, при котором требуется, чтобы $\widetilde{f}(x)$ проходила через заданные узлы $(x_i, f(x_i))$ внутри отрезка $x \in [a; b]$.*

Определение 2.1.2. *Экстраполяцией называется приближение, при котором требуется, чтобы $\widetilde{f}(x)$ по заданным узлы $(x_i, f(x_i))$ предсказывала значение $f(x)$ вне отрезка $[a; b]$.*

Рисунок 2.1 демонстрирует разницу между интерполяцией и экстраполяцией одних и тех же данных (центральный и правый графики), а также показывает приближение данных линейным полиномом (так называемая *линейная регрессия*). Интерполяция требует, чтобы аппроксимирующая функция $\widetilde{f}(x)$ проходила через заданные узлы, что необходимо, например, при аппроксимации решения дифференциального уравнения (решение, очевидно, должно проходить через собственные узлы). Это мотивирует использование интерполяционных многочленов в самых разных численных методах решения ДУ – например, в методе конечного элемента, где решение в элементе интерполируется между узлами элемента. Отметим, что вне отрезка $[a; b]$ поведение интерполирующей кривой уже не должно

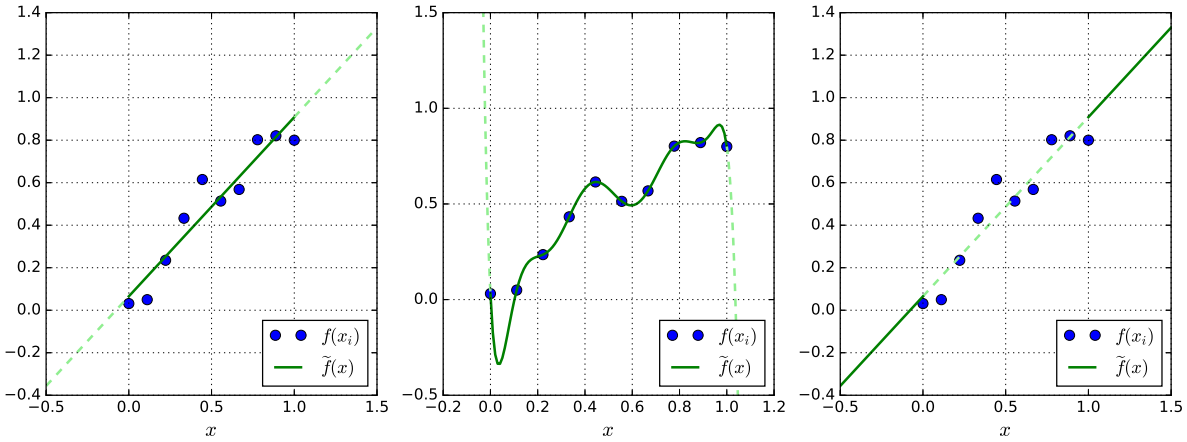

Рисунок 2.1 – Пример приближения данных линейным полиномом (левый график), их интерполяции (центральный график) и экстраполяции по линейному полиному (правый график). Сплошные линии обозначают часть кривой, представляющей главный интерес при том или ином виде приближения данных.

представлять интереса и чаще всего является неудовлетворительным для экстраполяции (штриховая линия на центральном графике рисунка 2.1).

С другой стороны для целей экстраполяции основной интерес представляет поведение аппроксимирующей кривой вне заданного отрезка. Анализ временных рядов биржевых курсов, очевидно, нацелен на экстраполяцию имеющихся данных временного ряда с целью предсказания их эволюции в будущем.

2.2 Приближение в линейном пространстве

Наиболее распространенным случаем приближения является приближение в линейном пространстве, где аппроксимирующая функция $\widetilde{f}(x)$ представляется в виде линейной комбинации *базисных функций* $\phi_i(x) \in F$, где F – некоторое линейное функциональное пространство:

$$\widetilde{f}(x) = \sum_{i=1}^{n} c_i \phi_i(x), \tag{2.1}$$

где $c_i \in \mathbb{R}$ и n может быть бесконечностью.

Задача приближения в таком случае формулируется следующим образом: необходимо найти такие c_i, что аппроксимирующая функция $\widetilde{f}(x)$ "приближается" к $f(x)$ в том или ином смысле. Частным случаем является линейная интерполяция, где под приближением понимается совпадение значений $f(x)$ и $\widetilde{f}(x)$ в интерполяционных узлах $(x_1, x_2, ..., x_n)$:

$$f(x_j) = \widetilde{f}(x_j), \quad j = 1, \ldots, n. \tag{2.2}$$

Подставив (2.1) в уравнение (2.2), мы получаем систему линейных алгебраических уравне-

ний:

$$\begin{bmatrix} \phi_1(x_1) & \dots & \phi_n(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_n) & \dots & \phi_n(x_n) \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} f(x_1) \\ \vdots \\ f_(x_n) \end{bmatrix} \tag{2.3}$$

Прямое определение коэффициентов c_i решением системы (2.3) называется *методом неопределенных коэффициентов*. Обратим внимание на то, что количество узлов, равное n, не случайно совпадает с количеством базисных функций в сумме в (2.1) – это позволяет системе уравнений иметь единственное и нетривиальное решение (при условии $\forall i \neq j : \; x_i \neq x_j$).

Как видно из постановки задачи линейной интерполяции, основным вопросом является выбор базисных функций ϕ_i. Для начала мы рассмотрим самый разработанный случай базисных функций, а именно случай интерполяции многочленами, где $\phi_i(x) = x^{i-1}$. Тогда система (2.3) принимает вид:

$$\begin{bmatrix} 1 & x_1 & \dots & x_1^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^{n-1} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} f(x_1) \\ \vdots \\ f_(x_n) \end{bmatrix}, \tag{2.4}$$

где определитель системы является *определителем Вандермонда*, который при условии несовпадающих узлов всегда отличен от нуля:

$$\begin{vmatrix} 1 & x_1 & \dots & x_1^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^{n-1} \end{vmatrix} = \prod_{1 \leq j < i \leq n} (x_i - x_j) \tag{2.5}$$

Естественным является вопрос о том, какие функции $f(x)$ можно аппроксимировать подобной линейной композицией многочленов (или, что эквивалентно, полиномов). Ответ на него дает знаменитая аппроксимационная теорема Вейерштрасса.

2.3 Аппроксимационная теорема Вейерштрасса

Мы приводим теорему для случая аппроксимации непрерывных функций полиномами, но необходимо помнить, что теорема Вейерштрасса имеет ряд обобщений (например, теорема Вейештрасса-Стоуна). Рассмотренное ниже доказательство этой теоремы, не являющееся обязательным для усвоения и приведенное только для иллюстрации техники теории приближений, было выполнено выдающимся российским и советским математиком Сергеем Бернштейном, который ввел в математический оборот *полиномы Бернштейна*:

$$B_n(x; f) = \sum_{k=0}^{n} f(k/n) \binom{n}{k} x^k (1-x)^{n-k}, \tag{2.6}$$

где $f : [0; 1] \longrightarrow \mathbb{R}$ – некоторая функция, $\binom{n}{k}$ – биномиальный коэффициент:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2.7}$$

Выведем некоторые равенства, которые будут полезны при доказательстве теоремы. Биномиальный коэффициент (2.7) обладает следующим рекурсивным свойством:

$$\binom{n-1}{k-1} = \frac{(n-1)!}{(k-1)!(n-k)!} = \frac{k}{n}\binom{n}{k} \tag{2.8}$$

Можно увидеть, что при $f(\frac{k}{n}) = 1$, сумма в (2.6) представляет собой бином Ньютона, равный единице:

$$\begin{aligned} \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} &= ((1-x)+x)^n \\ &= 1. \end{aligned} \tag{2.9}$$

Так как (2.9) верно для любого n, заменив n на $n-1$ и k на $\tilde{k}$, мы получаем

$$\begin{aligned} 1 &= \sum_{\tilde{k}=0}^{n-1} \binom{n-1}{\tilde{k}} x^{\tilde{k}} (1-x)^{n-(\tilde{k}-1)} \\ \xRightarrow{\cdot x} \quad x &= \sum_{\tilde{k}=0}^{n-1} \binom{n-1}{\tilde{k}} x^{\tilde{k}+1} (1-x)^{n-(\tilde{k}-1)} \\ \xRightarrow{(2.8)} \quad x &= \sum_{\tilde{k}=0}^{n-1} \frac{\tilde{k}+1}{n} \binom{n}{\tilde{k}+1} x^{\tilde{k}+1} (1-x)^{n-(\tilde{k}-1)} \\ \xRightarrow{k=\tilde{k}+1} \quad x &= \sum_{k=1}^{n} \frac{k}{n} \binom{n}{k} x^k (1-x)^{n-k}, \end{aligned} \tag{2.10}$$

откуда, помня, что при $k = 0$ член серии обращается в ноль, следует равенство:

$$x = \sum_{k=0}^{n} \frac{k}{n} \binom{n}{k} x^k (1-x)^{n-k}, \tag{2.11}$$

Аналогичным образом при замене n на $n-2$ из (2.9) имеем:

$$\begin{aligned} (n^2-n)x^2 &= \sum_{k=0}^{n} (k^2-k) \binom{n}{k} x^k (1-x)^{n-k} \\ \xRightarrow{:n^2 \text{ и } (2.11)} \quad (1-\frac{1}{n})x^2 + \frac{x}{n} &= \sum_{k=0}^{n} \left(\frac{k}{n}\right)^2 \binom{n}{k} x^k (1-x)^{n-k} \\ \xRightarrow{+x^2\cdot(2.9)-2x\cdot(2.11)} \quad \frac{1}{n}(x-x^2) &= \sum_{k=0}^{n} \left(x-\frac{k}{n}\right)^2 \binom{n}{k} x^k (1-x)^{n-k}. \end{aligned} \tag{2.12}$$

Закончив подготовительную работу, перейдем к формулировке и доказательству теоремы.

Теорема 2.3.1. *(аппроксимационная теорема Вейерштрасса) Пусть $g(t)$ – функция, непрерывная на замкнутом отрезке $[a;b]$. Тогда существует такая последовательность многочленов $\{P_n(t)\}_{n=1}^{\infty}$, что она равномерно сходится к $g(t)$ при $n \to \infty$.*

Доказательство. Без потери общности рассмотрим доказательство для непрерывной функции $f(x) = g\left((b-a)\,x + a\right)$, где $x \in [0;1]$. Нам достаточно показать, что последовательность полиномов Берштейна сходится к $f(x)$. Домножив (2.9) на $f(x)$ имеем:

$$\begin{aligned} & f(x) = \sum_{k=0}^{n} f(x) \binom{n}{k} x^k (1-x)^{n-k} \\ \implies \quad & f(x) - B_n(x;f) = \sum_{k=0}^{n} \left[f(x) - f(k/n)\right] \binom{n}{k} x^k (1-x)^{n-k} \qquad (2.13) \\ \xRightarrow{\text{нерав-во треугольника}} \quad & |f(x) - B_n(x;f)| = \sum_{k=0}^{n} |f(x) - f(k/n)| \binom{n}{k} x^k (1-x)^{n-k} \end{aligned}$$

По теореме Вейерештрасса об ограниченности непрерывной на отрезке функции справедливо неравенство $|f(x)| \leq M$, где M – некоторая конечная константа. Доказательство строится на разделении суммы в (2.13) на те члены, для которых $x \approx k/n$ (они естественным образом малы) и остальные. Пусть $\epsilon > 0$ и $\delta(\epsilon)$ являются величинами, взятыми в соотстветствии с определением равномерной непрерывности 1.3.18. Выберем такое n, что

$$n \geq \sup\left\{\delta^{-4}(\epsilon), \frac{M^2}{\epsilon^2}\right\} \qquad (2.14)$$

Рассмотрим те k-ые члены в сумме (2.13), для которых верно $|x - \frac{k}{n}| < n^{-1/4} \geq \delta(\epsilon)$. Используя определение ϵ, получаем для них неравенство:

$$\begin{aligned} \sum_{k} |f(x) - f(k/n)| \binom{n}{k} x^k (1-x)^{n-k} & \leq \epsilon \sum_{k} \binom{n}{k} x^k (1-x)^{n-k} \\ & \leq \epsilon \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} \qquad (2.15) \\ (2.9) \implies \quad & = \epsilon \end{aligned}$$

Теперь рассмотрим те члены в сумме (2.13), для которых $|x - \frac{k}{n}| \geq n^{-1/4}$. Для них,

используя свойство ограниченности $f(x)$, имеем:

$$\begin{aligned}
\sum_k |f(x) - f(k/n)| \binom{n}{k} x^k (1-x)^{n-k} &\leq \sum_k 2M \binom{n}{k} x^k (1-x)^{n-k} \\
&= \sum_k 2M \frac{\left(x - \frac{k}{n}\right)^2}{\left(x - \frac{k}{n}\right)^2} \binom{n}{k} x^k (1-x)^{n-k} \\
\left|x - \frac{k}{n}\right| \geq n^{-1/4} \Longrightarrow \quad &\leq \sqrt{n} \sum_k 2M \left(x - \frac{k}{n}\right)^2 \binom{n}{k} x^k (1-x)^{n-k} \qquad (2.16) \\
(2.12) \Longrightarrow \quad &\leq 2M\sqrt{n} \frac{1}{n} (x - x^2) \\
x - x^2 \leq 1/4 \text{ на } [0;1] \Longrightarrow \quad &\leq \frac{M}{2\sqrt{n}} \\
(2.14) \Longrightarrow \quad &\leq \frac{\epsilon}{2}
\end{aligned}$$

Таким образом обе части суммы в (2.13) ограничены ϵ для любого $x \in [0;1]$:

$$|f(x) - B_n(x;f)| < \frac{3}{2}\epsilon, \tag{2.17}$$

что по определению означает равномерную сходимость последовательности полиномов Бернштейна к функции $f(x)$. □

2.4 Интерполяционный многочлен Лагранжа

Непосредственное решение системы (2.4) с целью вычисления коэффициентов c_i не является оптимальным для одномерного случая. Вместо этого мы получил явное представление интерполяционного многочлена. Можно заметить, что условие равенства значений заданной и интерполируемой функций в узлах (2.2) выполняется, если нам удастся построить такие $\phi_i(x)$, что

$$\phi_i(x_j) = \delta_{ij}, \tag{2.18}$$

где δ_{ij} называется *символом Кронекера* и определяется как

$$\delta_{ij} = \begin{cases} 1, & \text{если } i = j, \\ 0, & \text{если } i \neq j. \end{cases} \tag{2.19}$$

Иными словами мы хотим построить такую функцию $\phi_i(x)$, что она равна 1 только в узле x_i и обращается в ноль во всех остальных узлах. Тогда искомый интерполяционный многочлен будет вычислятся как

$$\widetilde{f}_n(x) = \sum_{i=1}^{n} f(x_i)\phi_i(x). \tag{2.20}$$

Действительно, легко убедиться, что в интерполяционных узлах значения функций совпадают:

$$\widetilde{f}_n(x_j) = \sum_{i=1}^{n} f(x_i)\phi_i(x_j) = \sum_{i=1}^{n} f(x_i)\delta_{ij} = f(x_i). \tag{2.21}$$

Зная, что $\phi_i(x_j) = 0$ при $i \neq j$, следующий полином удовлетворяет указанному требованию:

$$\phi_i(x) = C_i \prod_{i \neq j} (x - x_j), \tag{2.22}$$

где C_i – некоторая константа. Определить неизвестную C_i можно из условия

$$\begin{aligned} & \phi_i(x_i) = 1 \\ \Longrightarrow \quad & C_i \prod_{i \neq j} (x_i - x_j) = 1 \\ \Longrightarrow \quad & C_i = \frac{1}{\prod_{i \neq j}(x_i - x_j)}, \end{aligned} \tag{2.23}$$

что в результате дает многочлен, известный как *интеполяционный многочлен Лагранжа*:

$$\widetilde{f}_n(x_j) = L_{n-1}(x) = \sum_{i=1}^{n} f(x_i) \prod_{i \neq j} \frac{x - x_j}{x_i - x_j}. \tag{2.24}$$

Определение 2.4.1. *Пусть функция $f(x)$ задана в n интерполяционных узлах $x_1, x_2, ..., x_n$ на отрезке $[a; b]$, т.е. $x_1 = a$ и $x_n = b$. Тогда интерполяционным многочленом для функции $f(x)$ и соответствующих узлов интерполяции называется функция*

$$L_{n-1}(x) = \sum_{i=1}^{n} f(x_i) l_i(x) = \sum_{i=1}^{n} f(x_i) \prod_{i \neq j} \frac{x - x_j}{x_i - x_j}, \tag{2.25}$$

где $l_i(x) = \prod_{i \neq j} \frac{x - x_j}{x_i - x_j}$ является базисным многочленом $(n-1)$-й степени.

К примеру, квадратичная интерполяция между узлами x_1, x_2 и x_3:

$$L_2(x) = f(x_1)\frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} + \ + f(x_2)\frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} + f(x_3)\frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)}. \tag{2.26}$$

Форма базисных многочленов (также называемые базисными полиномами Лагранжа) позволяет многое сказать о поведении интерполирующей функции. Графики на рисунке 2.2 показывают базисные полиномы для двух, трех и четырех равномерно распределенных узлов. Можно заметить, что базисные полиномы при увеличении их степени (т.е. при увеличении количества узлов) имеют тенденцию к росту амплитуды ближе к граничным узлам отрезка. Чем выше степень базисного полинома, тем более заметным становится этот эффект. Подобное поведение может привести к появлению нежелательных, паразитных осцилляций у граничных узлов. Интуитивно можно заключить, что использование неравномерно распределенных узлов, концентрирующихся у границ отрезка, потенциально могло бы решить проблему. Как мы увидим в обсуждении многочленов Чебышева, такой выбор узлов действительно является оптимальным с точки зрения минимизации ошибки интерполирования.

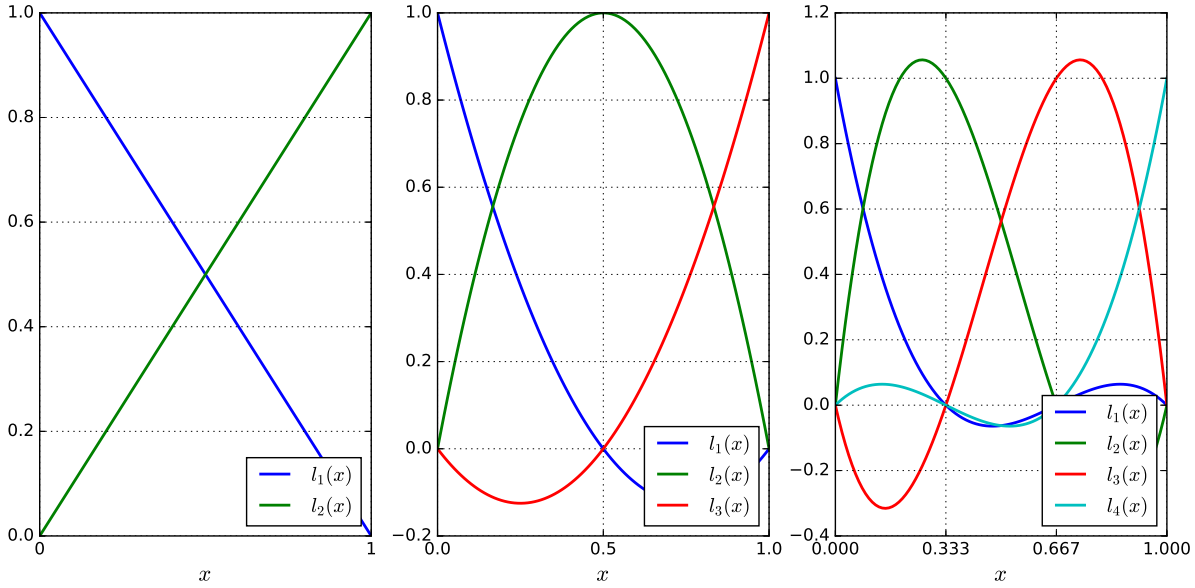

Рисунок 2.2 – Базисные многочлены Лагранжа первой (левый график), второй (центральный график) и третьей степени (правый график), определенные на равномерно распределенных узлах на отрезке $[0;1]$ (сетка x-координаты соответствует интерполяционным узлам).

2.5 Оценка остаточного члена многочлена Лагранжа

Определим ошибку, возникающую при аппроксимации функции $f(x)$ интерполяционным многочленом Лагранжа. Для этого нам необходимо найти значение остаточного члена $f(x) - L_n(x)$.

Теорема 2.5.1. *Пусть $x_1, ..., x_n \in [a;b]$ – интерполяционные узлы и $f(x) \in C^n[a;b]$. Тогда $\forall x \in [a;b] \quad \exists \xi \in (a;b)$ такое, что*

$$f(x) - L_{n-1}(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{i=1}^{n} (x - x_i) \tag{2.27}$$

Доказательство. Случай $x = x_i$ тривиален, так что рассмотрим $x \neq x_i, i = 1, \ldots, n$. Введем функцию $g(t)$:

$$g(t) = f(t) - L_{n-1}(t) - [f(x) - L_{n-1}(x)] \prod_{i=1}^{n} \frac{t - x_i}{x - x_i}. \tag{2.28}$$

Несложно проверить, что $g(x_i) = 0, i = 1, \ldots, n$ и $g(x) = 0$. Таким образом, функция $g(t)$ имеет $n + 1$ корней. Теорема Ролля гласит, что в этом случае $g'(t)$ имеет как минимум n корней. Обобщая теорему Ролля на производные высшего порядка, получаем, что $g^{(n)}(t)$ имеет как минимум один корень в точке $\xi \in (a;b)$. Тогда, учитывая, что $L_n(t)$ является

полиномом $n-1$ степени, имеем

$$\begin{aligned} g^{(n)}(\xi) = 0 &= f^{(n)}(\xi) - L_{n-1}^{(n)}(\xi) - [f(x) - L_{n-1}(x)] \frac{d^n}{dt^n} \prod_{i=1}^{n} \left[\frac{t-x_i}{x-x_i}\right]_{t=\xi} \\ &= f^{(n)}(t) - [f(x) - L_{n-1}(x)] \frac{d^n}{dt^n} \prod_{i=1}^{n} \left[\frac{t-x_i}{x-x_i}\right]_{t=\xi}. \end{aligned} \tag{2.29}$$

Заметим, что

$$\prod_{i=1}^{n} \frac{t-x_i}{x-x_i} = t^n \prod_{i=1}^{n} \frac{1}{x-x_i} + O(t^{n-1}), \tag{2.30}$$

и тогда мы получаем

$$0 = f^{(n)}(\xi) - [f(x) - L_{n-1}(x)] \frac{n!}{\prod_{i=1}^{n}(x-x_i)}, \tag{2.31}$$

из чего следуем искомое

$$f(x) - L_{n-1}(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{i=1}^{n}(x-x_i). \tag{2.32}$$

□

2.6 Интерполяция Эрмита

Зачастую кроме значений функции $f(x)$ в нескольких узлам, нам также известны ее производные в них же. В таком случае возникает желание построить полином $\tilde{f}(x)$, который был бы одновременно согласован с $f(x)$ и M ее производными $f^{(m)}(x), 1 \le m \le M$ в n узлах:

$$f(x_i) = \tilde{f}(x_i), \tag{2.33}$$

$$\frac{d^m}{dx^m} f(x)\Big|_{x=x_i} = \frac{d^m}{dx^m} \tilde{f}(x)\Big|_{x=x_i}. \tag{2.34}$$

Это формирует n равенств для $f(x_i)$ и nM для ее производных. Таким образом, максимальная степень полинома, удовлетворяющего этим условиям, равна $n(M+1)-1$.

В случае $M=1$ интерполирование полиномами производится с помощью *многочленов Эрмита*. Они позволяют сформировать полином $2n-1$ степени, согласующийся с $f(x)$ и ее первой производной в n точках. Согласованность с первой производной эквивалентна согласованности с касательной к функции $f(x)$, что продемонстрировано рисунком 2.3.

Сформируем теорему о многочлене Эрмита, в которой он выражен с помощью уже знакомых нам базисных полиномов Лагранжа.

Теорема 2.6.1. *Пусть $x_1, ..., x_n \in [a;b]$ – интерполяционные узлы и $f(x) \in C^1[a;b]$. Тогда единственный многочлен наименьшей степени согласующийся с $f(x_i)$ и $f'(x_i)$, $i = 1, \dots, n$ является многочленом Эрмита степени (максимум) $2n-1$, заданный выражением*

$$H_{2n-1}(x) = \sum_{i=1}^{n} f(x_i) h_i(x) + \sum_{i=1}^{n} f'(x_i) \hat{h}_i(x), \tag{2.35}$$

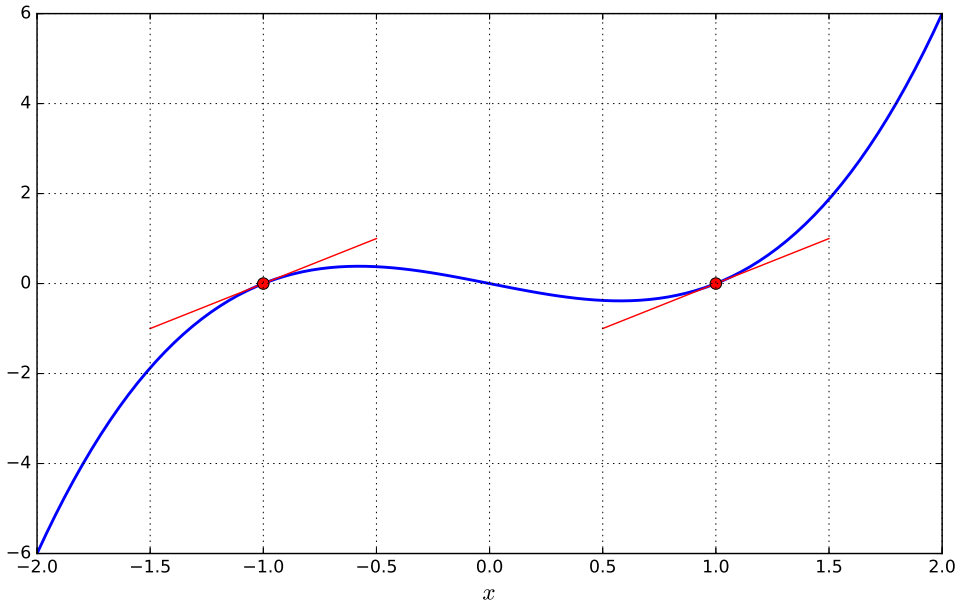

Рисунок 2.3 – Часть функции, интерполированной по узлам $x_1 = -1$ и $x_2 = 1$ (красные точки) с учетом равенства первых производных в узлах. Согласованность с первой производной удобно визуализировать касательными прямыми в узлах интерполяции (красные линии).

где $h_i(x)$ *и* $\hat{h}_i(x)$ *заданы как*

$$h_i(x) = \left[1 - 2(x - x_i)l_i'(x_i)\right] l_i^2(x), \tag{2.36}$$

$$\hat{h}_i(x) = (x - x_i)l_i^2(x), \tag{2.37}$$

где l_i *– базисные полинома Лагранжа* $n - 1$ *степени.*

Если $f \in C^{2n}[a; b]$*, тогда остаточный член интерполяции выражается формулой:*

$$f(x) - H_{2n-1}(x) = \frac{\prod_{i=1}^{n}(x - x_i)^2}{(2n)!} f^{(2n)}(\xi) \tag{2.38}$$

для некоторого $\xi \in (a; b)$.

Оставив доказательство единственности и вывод формулы для остаточного члена (он аналогичен выводу для интерполяционного многочлена Лагранжа) в стороне, убедимся, что построенный многочлен действительно согласуется с $f(x)$ и его первой производной в интерполяционных узлах x_i. Так как $l_i(x_j) = \delta_{ij}$, мы имеем $h_i(x_j) = \delta_{ij}$ и $\hat{h}_i(x_j) = 0$, что автоматически дает согласование $H_{2n-1}(x)$ с $f(x)$ в узлах интерполяции:

$$H_{2n-1}(x_i) = f(x_i). \tag{2.39}$$

Рассмотрим производные от $h_i(x)$ и $\hat{h}_i(x)$:

$$\begin{aligned} h_i'(x) &= 2l_i(x)l_i'(x)\left[1 - 2(x - x_i)l_i'(x_i)\right] - 2l_i'(x_i)l_i^2(x) \\ \hat{h}_i'(x) &= 2l_i(x)l_i'(x)(x - x_i) + l_i^2(x). \end{aligned} \tag{2.40}$$

Легко заметить, что $h_i'(x_j) = 0$ для любых i и j, в то время как $\hat{h}_i'(x_j) = \delta_{ij}$, что результирует в

$$H_{2n-1}'(x_i) = f'(x_i) \tag{2.41}$$

и таким образом доказывает согласованность составленного многочлена Эрмита с первой производной $f(x)$ в интерполяционных узлах.

2.7 Оптимальное распределение узлов интерполяции

До текущего момента мы не задавались вопросом о том, каким образом должны быть распределены узлы $x_1, \ldots, x_n$ в отрезке $[a; b]$. Самый очевидный случай, а именно случай равномерно распределенных узлов, где $x_{i+1} - x_i = h$ для любых $i = 1, \ldots n-1$, приводит к паразитным осцилляциям у границ отрезка интерполирования, что связанно с соответствующей формой базисных полиномов Лагранжа, пример которых показан на рисунке 2.2. Логично предположить, что минимизация остаточного члена, формула для которого была выведена в теореме 2.5.1, относительно значений $x_1, \ldots, x_n$ могла бы дать оптимальные значения $x_1, \ldots, x_n$, которые следует использовать для интерполяции. Решение этой оптимизационной задачи связанно с многочленами Чебышева, которые теперь необходимо детально рассмотреть.

2.7.1 Ортогональные функции

Перед рассмотрением многочленов Чебышева, нам предварительно необходимо ввести несколько новых понятий.

Определение 2.7.1. *Интегрируемая функция $\omega(x)$ называется весовой функцией, определенной на интервале $[-1; 1]$, если $\omega(x) \geq 0$ для любых $x \in [-1; 1]$, но при этом $\omega(x) \neq 0$ на любом подинтервале $[-1; 1]$.*

Задача весовой функции состоит в том, чтобы сделать одни части интервала $[-1; 1]$ более "важными", чем другие. В частности, весовая функция

$$\omega(x) = \frac{1}{\sqrt{1 - x^2}}, \tag{2.42}$$

изображенная на рисунке 2.4, будучи домноженной на некоторую другую функцию $f(x)$, будет увеличивать вклад $f(x)$ ближе к границам отрезка $(-1; 1)$.

Весовые функции необходимы для определения ортогональности функций. В конечномерных векторных пространствах, которые известны из классического курса линейной алгебры, ортогональность векторов определяется через обнуление скалярного произведения:

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i=1}^{n} a_i b_i = 0, \tag{2.43}$$

где $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ – ортогональные вектора. Как уже отмечалось во введении, функции являются суть бесконечномерными векторами, что позволяет сформулировать ортогональность

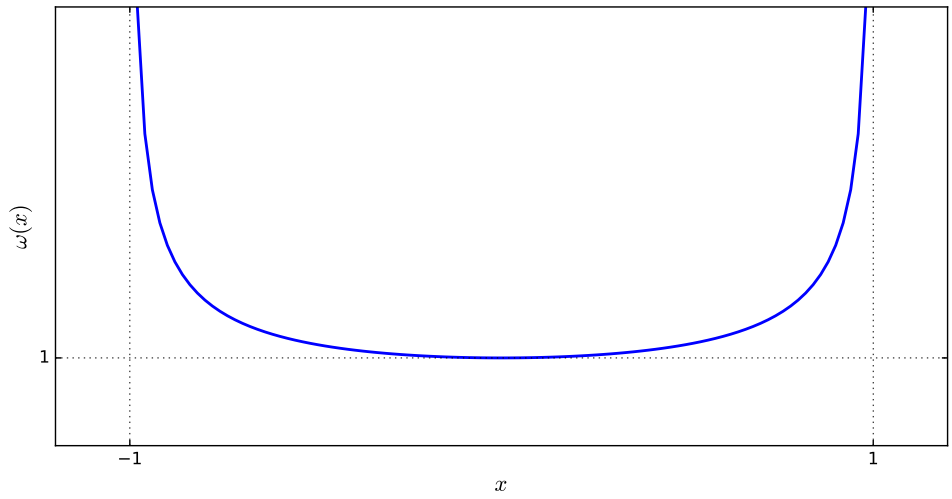

Рисунок 2.4 – Весовая функция $\omega(x) = \frac{1}{\sqrt{1-x^2}}$, определенная на интервале $[-1; 1]$.

функций через *скалярное произведение функций*:

$$\langle f(x), g(x) \rangle = \int_{-1}^{1} f(x)g(x)dx, \tag{2.44}$$

которое порождает среднеквадратическую норму и, опуская вопросы полноты, автоматически обязывает функции $f(x), g(x)$ принадлежать гильбертову пространству $L_2[-1; 1]$. В таком случае мы говорим, что функции $f(x)$ и $g(x)$ ортогональны, если $\langle f(x), g(x) \rangle = 0$. Однако на практике удобным оказался несколько модифицированный вид ортогональности, называемый *ортогональностью с весом*.

Определение 2.7.2. *Множество функций $\{\phi_1, \dots, \phi_n\}$ называется ортогональной системой функций с весом $\omega(x)$ на интервале $[a; b]$, если*

$$\langle \phi_i(x), \phi_j(x) \rangle_\omega = \int_a^b \omega(x)\phi_i(x)\phi_j(x) = \alpha_i \delta_{ij}, \tag{2.45}$$

где $\alpha_i > 0$. Если $\langle \phi_i(x), \phi_j(x) \rangle_\omega = \delta_{ij}$, то система называется ортонормальной.

Пример 2.7.1. *Тригонометрическая система функций $\{1, \cos x, \sin x, \cos 2x, \sin 2x, \dots\}$ является ортогональной на отрезке $[-\pi; \pi]$. Действительно, для $\phi_1(x) = 1$ и $k \in \mathbb{N}$ мы имеем:*

$$\int_{-\pi}^{\pi} 1 \cdot \cos kx dx = \left. \frac{\sin kx}{k} \right|_{-\pi}^{\pi} = 0, \tag{2.46}$$

$$\int_{-\pi}^{\pi} 1 \cdot \sin kx dx = \left. -\frac{\cos kx}{k} \right|_{-\pi}^{\pi} = 0. \tag{2.47}$$

Затем, для $k \in \mathbb{N}$ и $m \in \mathbb{N}$, где $k \neq m$ получаем:

$$\int_{-\pi}^{\pi} \cos kx \sin mx dx = \int_{-\pi}^{\pi} \frac{1}{2} \left[\sin(k+m)x - \sin(k-m)x\right] dx = 0, \tag{2.48}$$

$$\int_{-\pi}^{\pi} \cos kx \cos mx dx = \int_{-\pi}^{\pi} \frac{1}{2} \left[\cos(k-m)x + \cos(k+m)x\right] dx = 0, \tag{2.49}$$

$$\int_{-\pi}^{\pi} \sin kx \sin mx dx = \int_{-\pi}^{\pi} \frac{1}{2} \left[\cos(k-m)x - \cos(k+m)x\right] dx = 0. \tag{2.50}$$

И наконец для случая $k = m$:

$$\int_{-\pi}^{\pi} \cos kx \sin kx dx = \int_{-\pi}^{\pi} \frac{1}{2} \sin 2kx dx = 0, \tag{2.51}$$

$$\int_{-\pi}^{\pi} \cos^2 kx dx = \int_{-\pi}^{\pi} \frac{1}{2}(1 + \cos 2kx) dx = \pi, \tag{2.52}$$

$$\int_{-\pi}^{\pi} \sin^2 kx dx = \int_{-\pi}^{\pi} \frac{1}{2}(1 - \cos 2kx) dx = \pi. \tag{2.53}$$

Резюмируя, для скалярного произведения тригонометрических функций имеет место равенство $\langle \phi_k, \phi_m \rangle = \pi \delta_{km}$, в то время как для функции $\phi_1(x) = 1$ скалярное произведение имеет вид $\langle \phi_1, \phi_k \rangle = 2\pi\delta_{1k}$. Таким образом тригонометрическая система функций является ортогональной на $[-\pi; \pi]$.

Теперь мы можем перейти к обсуждению многочленов Чебышева, которые, как будет показано ниже, составляют ортогональную систему функций.

2.7.2 Многочлены Чебышева

Многочлены Чебышева компактнее всего выражаются с помощью тригонометрических функций:

$$T_k(x) = \cos[k \arccos x], \tag{2.54}$$

где $k \geq 0$ и $x \in I$. Чтобы показать, что выражение (2.54) действительно генерирует ряд полиномов, в первую очередь заметим, что

$$T_0(x) = 1, \tag{2.55}$$

$$T_1(x) = x. \tag{2.56}$$

Для $k > 1$ произведем замену:

$$\begin{aligned} & \theta = \arccos x, \\ \Longrightarrow \quad & T_k(\theta) = \cos(k\theta), \end{aligned} \tag{2.57}$$

где $\theta \in [0; \pi]$. Подобная тригонометрическая форма представления многочленов Чебышева задает их как ряд косинусов, что позволяет обращаться с ними так же, как и с классическими тригонометрическими системами функций (в частности, для них можно определить

полное и дискретное преобразование Фурье). Используя тригонометрическую форму, выражения для T_{k+1} и T_{k-1} будут иметь вид:

$$T_{k+1}(\theta) = \cos\left[(k+1)\theta\right] = \cos(k\theta)\cos(\theta) - \sin(k\theta)\sin(\theta), \tag{2.58}$$

$$T_{k-1}(\theta) = \cos\left[(k-1)\theta\right] = \cos(k\theta)\cos(\theta) + \sin(k\theta)\sin(\theta). \tag{2.59}$$

Сложив два уравнения, получаем:

$$\begin{aligned} & T_{k+1}(\theta) = 2\cos(k\theta)\cos(\theta) - T_{k-1}(\theta). \\ \Longrightarrow \quad & T_{k+1}(\theta) = 2T_k(\theta)\cos(\theta) - T_{k-1}(\theta). \end{aligned} \tag{2.60}$$

Проведя обратную замену, мы получаем рекуррентное соотношение для многочленов Чебышева:

$$\begin{aligned} & x = \cos(\theta) \\ \Longrightarrow \quad & T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x). \end{aligned} \tag{2.61}$$

Так, для $k = 2$ и $k = 3$ мы имеем:

$$T_2(x) = 2xT_1(x) - T_0(x) = 2x^2 - 1, \tag{2.62}$$

$$T_3(x) = 2xT_2(x) - T_1(x) = 2x(2x^2 - 1) - x = 4x^3 - 3x. \tag{2.63}$$

Заметим, что рекуррентное соотношение (2.61) предполагает, что $T_k(x)$ является полиномом k-й степени с ведущим членом $2^{k-1}x^k$.

Из рисунка 2.5, иллюстрирующего полиномы Чебышева для $1 \leq n \leq 4$, можно заметить, что на отрезке $[-1; 1]$ их амплитуда ограничена единицей, т.е. $\forall x \in [-1; 1] : |T_k(x)| \leq 1$, что отличает их от базисных полиномов Лагранжа, изображенных на рисунке 2.2. Более того, расположение экстремумов полиномов Чебышева стремится к границам отрезка $[-1; 1]$ при увеличении k. В дальнейшем мы строго докажем теорему об экстремумах полиномов Чебышева и воспользуемся ими для минимизации ошибки интерполяции.

Теперь продемонстрируем, что многочлены Чебышева составляют ортогональную систему функций с весом $\omega(x) = 1/\sqrt{1-x^2}$ на отрезке $[-1; 1]$. Рассмотрим скалярное произведение многочленов:

$$\langle T_k(x), T_m(x)\rangle_\omega = \int_{-1}^{1} \frac{T_k(x)T_m(x)}{\sqrt{1-x^2}}dx = \int_{-1}^{1} \frac{\cos(k\arccos x)\cos(m\arccos x)}{\sqrt{1-x^2}}dx, \tag{2.64}$$

и произведем уже известную замену

$$\begin{aligned} & \theta = \arccos x \\ \Longrightarrow \quad & d\theta = -\frac{1}{\sqrt{1-x^2}}dx \\ \Longrightarrow \quad & \langle T_k(x), T_m(x)\rangle_\omega = -\int_{\pi}^{0} \cos k\theta \cos m\theta d\theta = \int_{0}^{\pi} \cos k\theta \cos m\theta d\theta. \end{aligned} \tag{2.65}$$

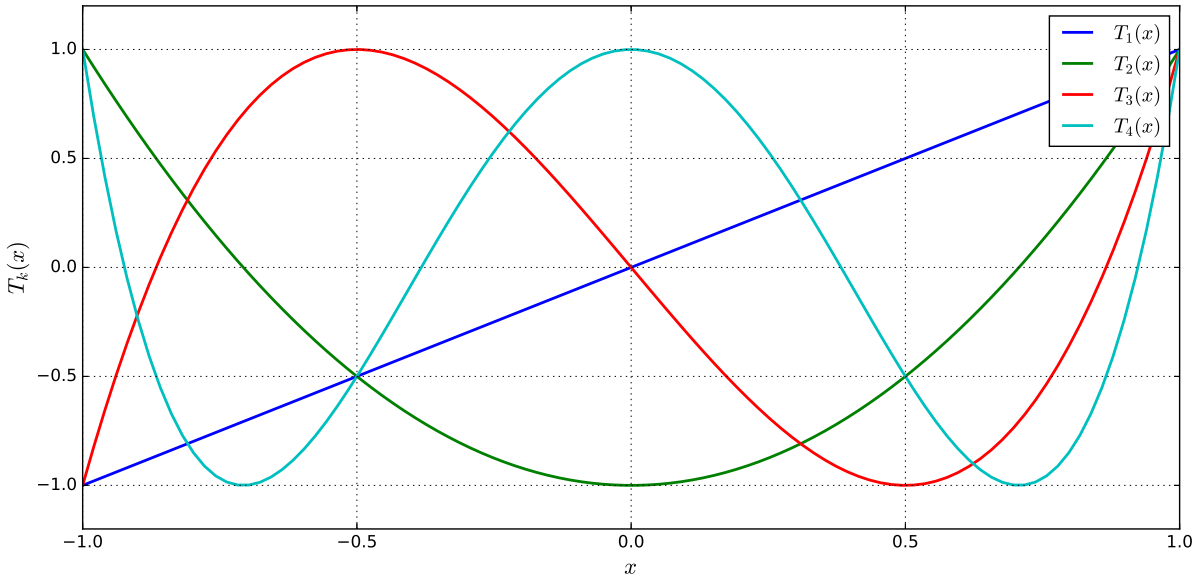

Рисунок 2.5 – Многочлены Чебышева $T_1(x), T_2(x), T_3(x)$ и $T_4(x)$, заданные на отрезке $[-1; 1]$.

Тогда для случаев $k \neq m$ и $k = m$ имеем:

$$\begin{aligned} k \neq m \quad &\Longrightarrow \quad \int_0^\pi \cos k\theta \cos m\theta d\theta = \int_0^\pi \frac{1}{2}\left[\cos(k-m)\theta + \cos(k+m)\theta\right] d\theta = 0, \\ k = m \quad &\Longrightarrow \quad \int_0^\pi \cos^2 k\theta d\theta = \frac{1}{2}\int_0^\pi (1 + \cos 2k\theta) d\theta = \frac{\pi}{2}, \end{aligned} \tag{2.66}$$

что результирует в следующем равенстве для скалярного произведения:

$$\langle T_k(x), T_m(x)\rangle_\omega = \frac{\pi}{2}\delta_{km}, \tag{2.67}$$

которое подтверждает, что система, сгенерированная выражением для полиномов Чебышева (2.54), действительно является ортогональной системой функций с весом $\omega(x) = 1/\sqrt{1-x^2}$ на отрезке $[-1; 1]$.

Установив ортогональность многочленов Чебышева, перейдем к теореме об их корнях и экстремумах, которая формализует некоторые интуитивные выводы, которые мы сделали из рисунка 2.5.

Теорема 2.7.1. *Пусть $T_k(x)$ является многочленом Чебышева и $k \geq 1$. Тогда $T_k(x)$ имеет k корней в замкнутом интервале $[-1; 1]$ в точках*

$$\bar{x}_m = \cos\left(\frac{2m-1}{2k}\pi\right), \quad m = 1, \ldots, k. \tag{2.68}$$

Более того, глобальные экстремумы $T_k(x)$ расположены в точках

$$\hat{x}_m = \cos\left(\frac{m\pi}{k}\right), \quad m = 0, \ldots, k. \tag{2.69}$$

и имеют соответствующие значения $T_k(\hat{x}_m) = (-1)^m$.

Доказательство. Удостоверимся, что $\bar{x}_m$ действительно являются корнями $T_k(x)$:

$$\begin{aligned} T_k(\bar{x}_m) &= \cos(k \arccos \bar{x}_m) \\ &= \cos\left(k \arccos\cos\left(\frac{2m-1}{2k}\pi\right)\right) \\ &= \cos\left(\frac{2m-1}{2}\pi\right) \\ &= 0. \end{aligned} \tag{2.70}$$

Так как $T_k(x)$ является полиномом k-й степени, и все $\bar{x}_m$ отличны друг от друга, других дополнительных корней $T_k(x)$ иметь не может.

Для доказательства утверждения о глобальных экстремумах необходимо рассмотреть первую производную $T_k(x)$:

$$\begin{aligned} T_k'(x) &= \frac{d}{dx}\left[\cos(k \arccos x)\right] \\ &= \frac{k \sin(k \arccos x)}{\sqrt{1-x^2}}. \end{aligned} \tag{2.71}$$

Тогда, подставляя вражение для $\hat{x}_m$, имеем:

$$\begin{aligned} T_k'(\hat{x}_m) &= \frac{k \sin\left(k \arccos\cos\left(\frac{m\pi}{k}\right)\right)}{\sqrt{1-\cos^2\left(\frac{m\pi}{k}\right)}} \\ &= \frac{k \sin m\pi}{\sin\left(\frac{m\pi}{k}\right)} \\ &= 0. \end{aligned} \tag{2.72}$$

Случаи $m = 0$ и $m = k$ соответствуют неопределенности $0/0$, так что для начала рассмотрим $m = 1, \ldots, k-1$. Так как $T_k'(x)$ является полиномом $(k-1)$-й степени, и все $\hat{x}_m$ отличны друг от друга, они соответствуют всем возможным корнях $T_k'(x)$. Другие глобальные экстремумы могут существовать только на границах отрезка $\hat{x} = -1$ и $\hat{x} = 1$. Рассмотрим экстремальные значения функции $T_k'(x)$:

$$\begin{aligned} T_k(\hat{x}_m) &= \cos(k \arccos \hat{x}_m) \\ &= \cos\left(k \arccos\cos\left(\frac{m\pi}{k}\right)\right) \\ &= \cos(m\pi) \\ &= (-1)^m, \end{aligned} \tag{2.73}$$

что включает в себя в том числе и $m = 0$ и $m = n$, соответствующие границам отрезкам. Таким образом $\hat{x}_m$ для $m = 0, \ldots, k$ действительно являются точками глобального экстремума функции $T_k(x)$ на отрезке $[-1; 1]$. □

Важнейшим свойством полиномов Чебышева является тот факт, что, будучи приведенными к нормированной форме, где коэффициент при члене с наибольшей степенью

равен единице, они имеют наименьшие по модулю экстремумы среди всех нормированных полиномов той же степени. Так как это свойство позволит нам в скором будущем минимизировать ошибку интерполяции, докажем его строго.

Теорема 2.7.2. *Пусть* $\widetilde{T}_k(x) = \frac{1}{2^{k-1}} T_k(x)$, *где* $k \geq 1$ *и* $T_k(x)$ *является полиномом Чебышева* k*-й степени, и пусть* $\widetilde{\Pi}_k$ *– множество всех нормированных полиномов степени* k*. Тогда верным является следующее утверждение:*

$$\frac{1}{2^{k-1}} = \max_{x\in[-1;1]} |\widetilde{T}_k(x)| \leq \max_{x\in[-1;1]} |P_k(x)| \quad \forall P_k(x) \in \widetilde{\Pi}_k. \tag{2.74}$$

Более того, равенство верно только при $P_k(x) = \widetilde{T}_k(x)$.

Доказательство. Рассмотрим доказательство от обратного. Пусть верно

$$\max_{x\in[-1;1]} |P_k(x)| \leq \max_{x\in[-1;1]} |\widetilde{T}_k(x)| = \frac{1}{2^{k-1}}. \tag{2.75}$$

Введем полином $(k-1)$-й степени $Q(x) = \widetilde{T}_k(x) - P_k(x)$. Для $k+1$ точки $\hat{x}_m$ мы имеем:

$$Q(\hat{x}_m) = \frac{(-1)^m}{2^{k-1}} - P_k(\hat{x}_m). \tag{2.76}$$

Однако из (2.75) следует $|P_k(\hat{x}_m)| \leq \frac{1}{2^{k-1}}$. Тогда мы получаем:

$$\begin{aligned} Q(\hat{x}_m) &\geq 0 \quad \text{для } m = 0, 2, \dots \\ Q(\hat{x}_m) &\leq 0 \quad \text{для } m = 1, 3, \dots \end{aligned} \tag{2.77}$$

Таким образом, вследствие непрерывности $Q(x)$ полином имеет корень в каждом подотрезке $[x_m; x_{m+1}]$, где $m = 0, \dots, k-1$, что суммарно дает k различных корней. Однако $Q(x)$ является полиномом $(k-1)$-й степени. Это возможно только в случае $Q(x) = 0$, что дает $P_k(x) = \widetilde{T}_k(x)$. □

2.7.3 Минимизация ошибки интерполяции Лагранжа

Вспомним выражение для остаточного члена интерполяции Лагранжа, доказанное в теореме 2.5.1, на отрезке $[-1; 1]$:

$$f(x) - L_{n-1}(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{i=1}^{n} (x - x_i), \tag{2.78}$$

где $\xi \in (-1; 1)$ и $x_1, \dots, x_n$ – узлы интерполяции. Оптимальная интерполяция соответствует случаю, при котором значение остаточного члена минимизировано. Так как мы не имеем контроля над $f(x)$, минимизация возможна только относительно значений интерполяционных узлов. Таким образом, мы получаем оптимизационную задачу следующего вида:

$$\begin{gathered} \min_{x_1,\dots,x_n} \prod_{i=1}^{n} (x - x_i), \\ \text{при условии } x_i \in [-1; 1], \quad i = 1, \dots, n. \end{gathered} \tag{2.79}$$

Заметим, что целевая функция является нормированным полиномом степени n с n различными корнями, равными $x_1, \ldots, x_n$. По теореме 2.7.2 ее максимальное значение на $[-1;1]$ минимально тогда, когда x_i соответствуют корням нормированного многочлена Чебышева $\widetilde{T}_n(x)$, т.е. когда

$$x_i = \bar{x}_i = \cos\left(\frac{2i-1}{2n}\pi\right), \quad i = 1, \ldots, n. \tag{2.80}$$

Более того, теорема 2.7.2 позволяет оценить остаточный член подобной оптимальной интерполяции:

$$\max|f(x) - L_{n-1}(x)| \leq \frac{1}{2^n n!}\max|f^{(n)}(x)|, \tag{2.81}$$

для любой функции $f(x) \in C^n[-1;1]$. Обобщение до произвольного отрезка $[a;b]$ реализуется с помощью непрерывной замены переменной

$$\tilde{x} = \frac{1}{2}\left[(b-a)x + a + b\right], \tag{2.82}$$

где $\tilde{x} \in [a;b]$ и $x \in [-1;1]$. С помощью подобной замены значения $\bar{x}_i$ пересчитываются на отрезке $[a;b]$.

Резюмируя, интерполяция Лагранжа для любой достаточно гладкой функции $f(x)$ является оптимальной тогда, когда узлы интерполяции распределены в соответствии с корнями многочлена Чебышева $\bar{x}_i$, имеющего степень равную количеству узлов интерполяции. Подобные узлы принято называть *чебышевскими узлами*. На примере функции Рунге $f(x) = \dfrac{1}{1+25x^2}$ рисунок 2.6 показывает, как использование чебышевских узлов позволяет снизить ошибку интерполяции, которая в случае равномерно распределенных узлов сильно возрастает ближе к границам отрезка.

2.8 Локальная интерполяция

До этого момента мы рассматривали глобальную интерполяцию, т.е. интерполяцию одной аппроксимирующей функцией по всему отрезку $[a;b]$. При обсуждении многочленов Лагранжа мы выяснили, что увеличение степени интерполирующего полинома может приводить к паразитным осцилляциям рядом с границами отрезка. Увеличение степени в свою очередь является результатом увеличения количества узлов интерполяции. Одним из выходов из этого положения является локальная интерполяция, где мы делим отрезок $[a;b]$ на маленькие подотрезки и используем интерполяцию полиномом малой степени на каждом из этих подотрезков, после чего "склеиваем" полученное множество полиномов в единую функцию, заданную на всем отрезке $[a;b]$. Такую интерполяцию называют *кусочной интерполяцией*.

Самым простым случаем локальной интерполяции является *кусочно-линейная интерполяция*, где между каждой парой точек строится линейная функция, соединяющая их. Пример подобной интерполяции показан на рисунке 2.7. Очевидным недостатком кусочно-линейной интерполяции является отсутствие гладкости, так как результирующая интерполирующая функция принадлежит $C^0[a;b]$ и, соответственно, имеет прерывную первую

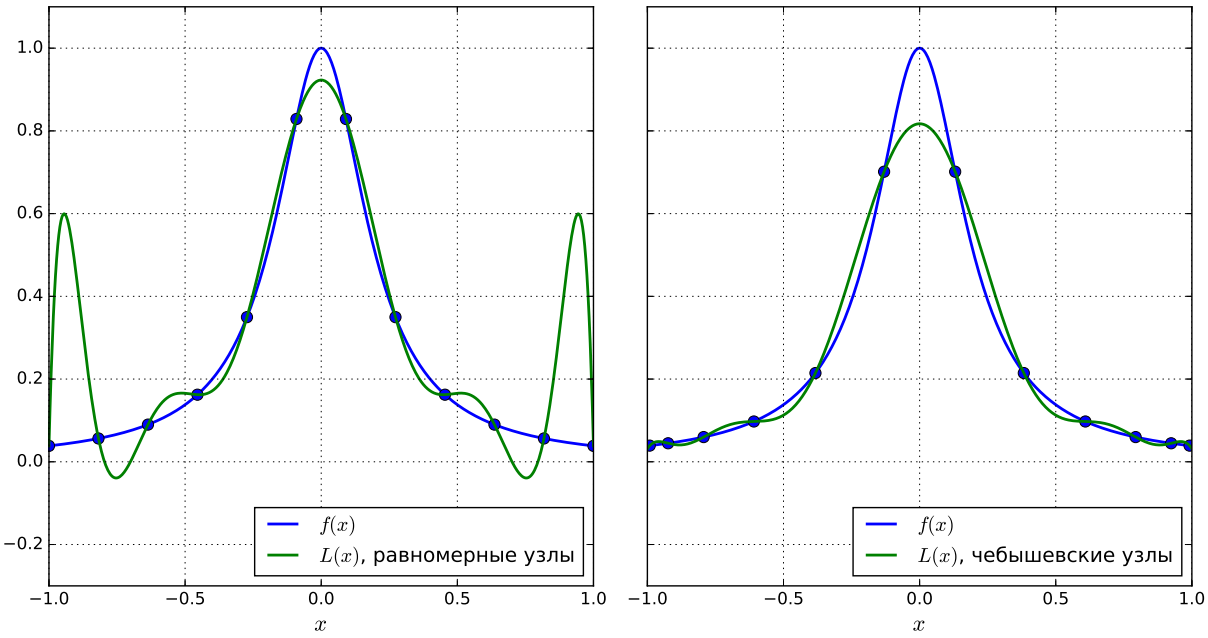

Рисунок 2.6 – Интерполяция функции Рунге $f(x) = \dfrac{1}{1+25x^2}$ с помощью равномерно распределенных (левый график) и чебышевских (правый график) узлов (синие точки).

производному. Более того, увеличение степени интерполяционных многочленов (что приводит,например, к кусочно-квадратичной интерполяции, кусочно-кубической интерполяции и т.д.) не исправляет эту ситуацию. Проблема гладкости в кусочной интерполяции решается с помощью введения дополнительных условий на значения интерполяционных многочленов в узлах, а именно, условий равенства их производных. На практике самым распространенным случаем является равенство первых и вторых производных в узлах, что требует использования кубических интерполяционных многочленом между парами узлов. Подобная интерполяция называется *интерполяцией кубическими сплайнами.*

2.8.1 Интерполяция кубическими сплайнами

Определение 2.8.1. *Пусть функция $f(x)$ задана в n интерполяционных узлах $a = x_1, x_2, \ldots, x_n = b$ на отрезке $[a; b]$. Тогда кубическим сплайном для функции $f(x)$ называется функция $S(x)$, для которой верно:*

1. *$S(x)$ кусочно задана кубическими многочленами $S_i(x)$ на каждом отрезке $[x_i; x_{i+1}]$, $i = 1, \ldots, n-1$;*
2. *$S_i(x_i) = f(x_i)$ и $S_i(x_{i+1}) = f(x_{i+1})$, $i = 1, \ldots, n-1$;*
3. *значения смежных многочленов совпадают в общих узлах: $S_i(x_{i+1}) = S_{i+1}(x_{i+1})$, $i = 1, \ldots, n-2$;*

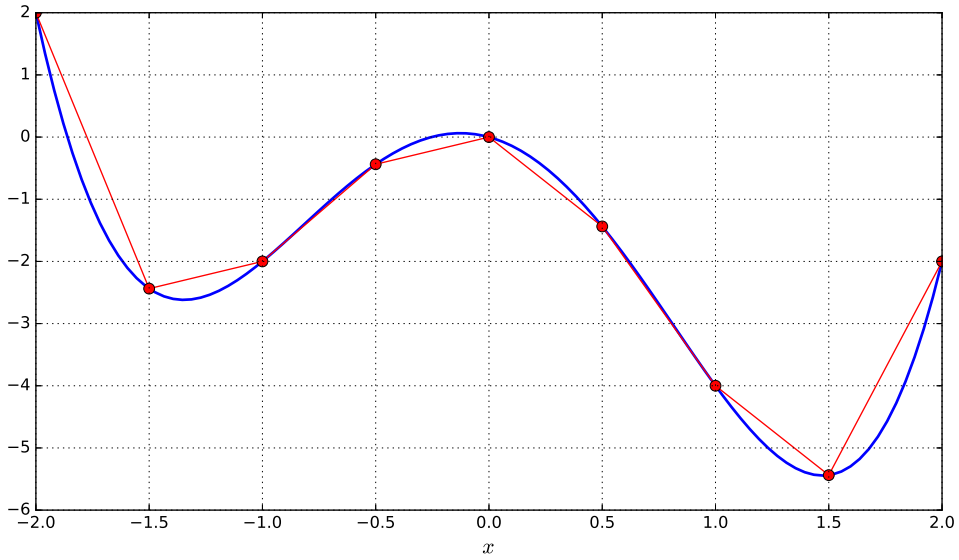

Рисунок 2.7 – Пример кусочно-линейной интерполяции функции $f(x) = x^4 - 4x^2 - x$.

4. *значения первых производных смежных многочленов совпадают в общих узлах:* $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$, $i = 1, \ldots, n-2$;
5. *значения вторых производных смежных многочленов совпадают в общих узлах:* $S_i''(x_{i+1}) = S_{i+1}''(x_{i+1})$, $i = 1, \ldots, n-2$;
6. *заданы граничные условия:*
   - *естественные граничные условия:* $S''(x_1) = S''(x_n) = 0$;
   - *граничные условия на касательную:* $S'(x_1) = f'(x_1)$ *и* $S'(x_n) = f'(x_n)$;

Так как кубический многочлен задается 4 константами, для задания кубического сплайна нам необходимо определить $4(n-1)$ констант. Сделаем это в общем виде. Запишем кубический многочлен $S_i(x)$ на отрезке $[x_i, x_{i+1}]$ в форме

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \tag{2.83}$$

что автоматически дает

$$S_i(x_i) = a_i = f(x_i). \tag{2.84}$$

Тогда из условия равенства значений смежных многочленов в общих узлах имеем:

$$a_{i+1} = S_{i+1}(x_{i+1}) = S_i(x_{i+1}) = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3, \tag{2.85}$$

где $h_i = x_{i+1} - x_i$. Так как $S_i'(x_i) = b_i$, из условия равенства значений первых производных в общих узлах смежных многочленов получаем:

$$b_{i+1} = b_i + 2c_i h_i + 3d_i h_i^2. \tag{2.86}$$

Наконец, из условия для второй производной имеем:

$$c_{i+1} = c_i + 3d_i h_i, \tag{2.87}$$

где $c_i = \frac{S_i''(x_i)}{2}$. Из последнего уравнения выразим d_i через c_i:

$$d_i = \frac{c_{i+1} - c_i}{3h_i}. \tag{2.88}$$

Подставив его в выражение для a_{i+1}, имеем:

$$a_{i+1} = a_i + b_i h_i + \frac{h_i^2}{3}(c_{i+1} + 2c_i). \tag{2.89}$$

Аналогично получаем для b_{i+1}:

$$b_{i+1} = b_i + h_i(c_{i+1} + c_i). \tag{2.90}$$

Чтобы получить замыкание для c_i, выраженное через $a_i = f(x_i)$, мы подставляем b_i и b_{i+1}, выраженное из (2.89)

$$\begin{aligned} & b_i = \frac{1}{h_i}(a_{i+1} - a_i) - \frac{h_i}{3}(c_{i+1} + 2c_i), \\ \Longrightarrow \quad & b_{i-1} = \frac{1}{h_{i-1}}(a_i - a_{i-1}) - \frac{h_{i-1}}{3}(c_i + 2c_{i-1}), \end{aligned} \tag{2.91}$$

в уравнение (2.90), записанное для b_i, что дает

$$h_{i-1}c_{i-1} + 2(h_i + h_{i-1})c_i + h_i c_{i+1} = \frac{3}{h_i}(a_{i+1} - a_i) - \frac{3}{h_{i-1}}(a_i - a_{i-1}) \tag{2.92}$$

Теперь мы можем составить систему уравнений относительно c_i, решив которую можно вычислить недостающие коэффициенты d_i и b_i по формулам (2.88) и (2.86) соответственно. Система уравнений и доказательство единственности ее решения для случая естественных граничных условий рассматриваются в следующей теореме.

Теорема 2.8.1. *Пусть функция $f(x)$ задана в n интерполяционных узлах $a = x_1, x_2, ..., x_n = b$ на отрезке $[a; b]$. Тогда функция $f(x)$ имеет уникальный естественный кубический сплайн $S(x)$, т.е. удовлетворяющий граничным условиям $S''(a) = 0$ и $S''(b) = 0$.*

Доказательство. Граничное условие $S''(a) = 0$ соответствует следующему условию на кубический многочлен $S_1(x)$:

$$\begin{aligned} S_1''(x)\big|_{x=a} &= [2c_1 + 6d_1(x - x_1)]\big|_{x=a} \\ &= 2c_1 \\ &= 0, \end{aligned} \tag{2.93}$$

т.е. $c_1 = 0$. Аналогично граничное условие $S''(b) = 0$ трансформируется в $c_n = 0$. Исходя из полученных ограничений на c_1 и c_n и уравнения (2.92), запишем матричное уравнение $\boldsymbol{Ac} = \boldsymbol{b}$, где $\boldsymbol{c} = [c_1, \ldots, c_n]^T$:

$$\begin{bmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ h_1 & 2(h_2+h_1) & h_2 & 0 & \ldots & 0 \\ 0 & h_2 & 2(h_3+h_2) & h_3 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & \ldots & \ldots & \ldots & \ldots & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_{n-1} \\ c_n \end{bmatrix} = \\ = \begin{bmatrix} 0 \\ \frac{3}{h_2}(a_3 - a_2) - \frac{3}{h_1}(a_2 - a_1) \\ \frac{3}{h_3}(a_4 - a_3) - \frac{3}{h_2}(a_3 - a_2) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 0 \end{bmatrix} \tag{2.94}$$

Можно заметить, что матрица $\boldsymbol{A}$ является матрицей со строгим диагональным преобладанием, т.е. $|A_{ii}| > \sum_{j \neq i} |A_{ij}|, \quad i = 1, \ldots, n$. Действительно:

$$2(h_i + h_{i-1}) > h_i + h_{i-1}. \tag{2.95}$$

Теорема о СЛАУ с матрицами со строгим диагональным преобладанием гласит, что такая СЛАУ имеет единственное решение для $\boldsymbol{c}$ (она будет рассмотрена и доказана в дальнейших лекциях **[TODO: добавить ссылку на секцию]**) □

Пример кубического сплайна, построенного после решения системы (2.94) и определения всех коэфициентов a_i, b_i, c_i и d_i, показан на рисунке 2.8.

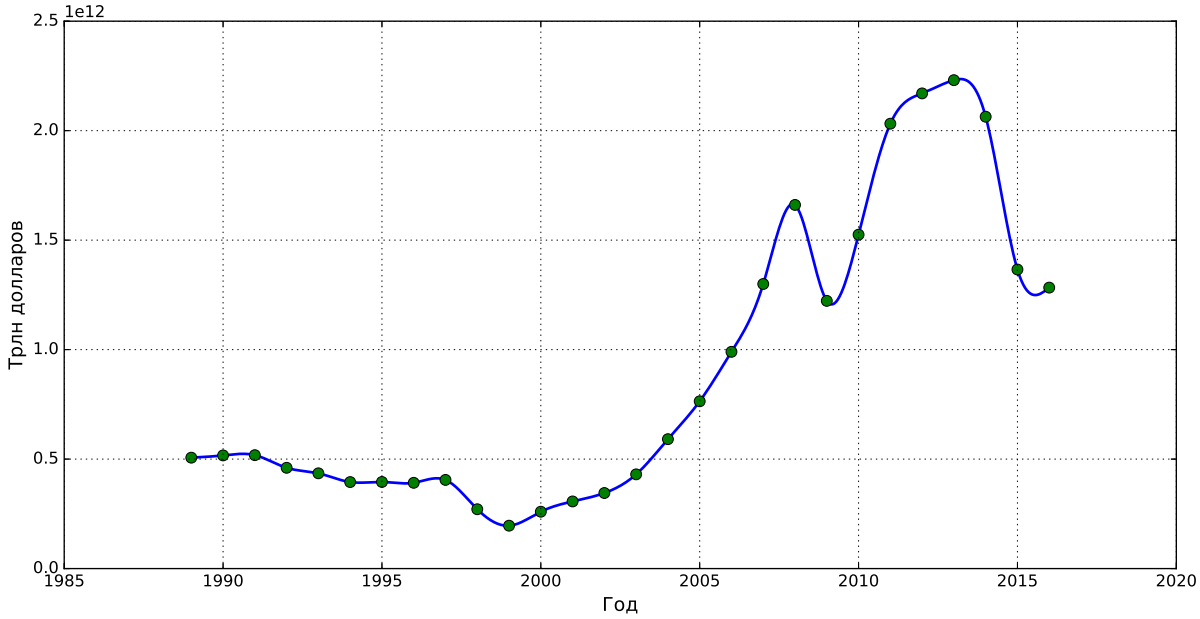

Рисунок 2.8 – Интерполяция ВВП России с помощью кубических сплайнов.

Численное дифференцирование и интегрирование

3.1 Численное дифференцирование

3.1.1 Метод дифференцирования многочлена Лагранжа

Простейшая формула для численного дифференцирования функции $f(x)$ в точке x^* может быть построена исходя из математического определения производной:

$$f'(x^*) = \lim_{h \to 0} \frac{f(x^* + h) - f(x^*)}{h}. \tag{3.1}$$

Предполагая h достаточно малым, мы всегда можем записать:

$$f'(x^*) \approx \frac{f(x^* + h) - f(x^*)}{h}. \tag{3.2}$$

В такой аппроксимации неявным образом была использована интерполяция линейным полиномом Лагранжа. Действительно, рассмотрим функцию $f(x) \in C^2[a; b]$ и такие два узла $x_1, x_2 \in [a; b]$, что $x_2 = x_1 + h$. Тогда функция $f(x)$ может быть выражена следующим образом:

$$\begin{aligned} f(x) &= L_1(x) + \frac{(x - x_1)(x - x_2)}{2!} f''(\xi(x)) \\ &= f(x_1)\frac{x - x_1 - h}{-h} + f(x_1 + h)\frac{x - x_1}{h} + \frac{(x - x_1)(x - x_1 - h)}{2!} f''(\xi(x)), \end{aligned} \tag{3.3}$$

где $\xi(x) \in (x_1; x_1 + h)$. Дифференцирование $f(x)$ приводит к следующему выражению:

$$f'(x) = \frac{f(x_1 + h) - f(x_1)}{h} + \frac{2x - 2x_1 - h}{2} f''(\xi(x)) + \frac{(x - x_1)(x - x_1 - h)}{2!} \frac{d}{dx} \left[f''(\xi(x)) \right], \tag{3.4}$$

из которого видно, что при исключении членов, зависящих от $\xi(x)$, производная аппроксимируется как:

$$f'(x) \approx \frac{f(x_1 + h) - f(x_1)}{h}. \tag{3.5}$$

В силу наличия производной от неизвестной функции в (3.4), остаточный член подобного численного дифференцирования мы можем определить только в точке x_1 (или, эквивалентно, в точке x_2):

$$f'(x_1) = \frac{f(x_1+h) - f(x_1)}{h} - \frac{h}{2}f''(\xi(x)). \tag{3.6}$$

Легко заметить, что при $h \to 0$ погрешность численного дифференцирования будет стремиться к нулю пропорционально $O(h)$. Таким образом мы имеем формулу для численного дифференцирования *первого порядка точности*.

Очевидно, что данный подход можно распространить на полиномы Лагранжа произвольной степени. Рассмотрим функцию $f(x) \in C^n[a;b]$ и n различных узлов $x_1, \ldots, x_n$. Тогда разложение $f(x)$ в базисные полиномы Лагранжа имеет вид:

$$f(x) = \sum_{i=1}^{n} f(x_i)l_i(x) + \frac{\prod_{i=1}^{n}(x-x_i)}{n!}f^{(n)}(\xi(x)), \tag{3.7}$$

что при дифференцировании дает:

$$f'(x) = \sum_{i=1}^{n} f(x_i)l_i'(x) + \frac{d}{dx}\left[\frac{\prod_{i=1}^{n}(x-x_i)}{n!}\right]f^{(n)}(\xi(x)) + \frac{\prod_{i=1}^{n}(x-x_i)}{n!}\frac{d}{dx}\left[f^{(n)}(\xi(x))\right]. \tag{3.8}$$

При дифференцировании в точке $x = x_j$ мы имеем:

$$\begin{aligned} f'(x_j) &= \sum_{i=1}^{n} f(x_i)l_i'(x_j) + \frac{d}{dx}\left[\frac{\prod_{i=1}^{n}(x-x_i)}{n!}\right]_{x=x_j} f^{(n)}(\xi(x_j)) \\ &= \sum_{i=1}^{n} f(x_i)l_i'(x_j) + \frac{1}{n!}\sum_{k=1}^{n}\prod_{i=1,i\neq k}^{n}(x_j - x_i)f^{(n)}(\xi(x_j)) \\ &= \sum_{i=1}^{n} f(x_i)l_i'(x_j) + \frac{\prod_{i=1,i\neq j}^{n}(x_j-x_i)}{n!}f^{(n)}(\xi(x_j)), \end{aligned} \tag{3.9}$$

что является общим выражением для численного дифференцирования по n точкам. Форма остаточного члена дает понять, что точность дифференцирования будет повышаться при использовании все большего числа точек, однако это будет происходить за счет увеличения числа арифметических операций, что всегда необходимо помнить, так как большое количество арифметических операций не только влияет на производительность численного метода, но и может приводить к вычислительным неустойчивостям, связанным с погрешностью округления.

В качестве примера использования (3.9) рассмотрим вывод формулы численного дифференцирования второго порядка точности. Для этого нам необходимо использовать интерполяцию $f(x)$ в трех точках x_1, x_2, x_3:

$$\begin{aligned} f'(x_j) = f(x_1)\frac{d}{dx}\left[\frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)}\right]_{x=x_j} + f(x_2)\frac{d}{dx}\left[\frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)}\right]_{x=x_j} + \\ + f(x_3)\frac{d}{dx}\left[\frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)}\right]_{x=x_j} + \frac{1}{6}\prod_{i=1,i\neq j}^{3}(x_j-x_i)f^{(3)}(\xi(x_j)), \end{aligned} \tag{3.10}$$

из чего следует:

$$f'(x_j) = f(x_1)\frac{2x_j - x_2 - x_3}{(x_1 - x_2)(x_1 - x_3)} + f(x_2)\frac{2x_j - x_1 - x_3}{(x_2 - x_1)(x_2 - x_3)} + +f(x_3)\frac{2x_j - x_1 - x_2}{(x_3 - x_1)(x_3 - x_2)} + \\ + \frac{1}{6}\prod_{i=1, i\neq j}^{3}(x_j - x_i)f^{(3)}(\xi(x_j)). \quad (3.11)$$

Теперь предположим, что узлы распределены равномерно, т.е. $x_2 = x_1 + h$ и $x_3 = x_1 + 2h$. Тогда выражение (3.11) принимает вид:

$$f'(x_j) = f(x_1)\frac{2x_j - 2x_1 - 3h}{-h \cdot (-2h)} + f(x_1 + h)\frac{2x_j - 2x_1 - 2h}{h \cdot (-h)} + f(x_1 + 2h)\frac{2x_j - 2x_1 - h}{2h \cdot h} + \\ + \frac{1}{6}\prod_{i=1, i\neq j}^{3}(x_j - x_i)f^{(3)}(\xi(x_j)). \quad (3.12)$$

Записав последнее выражение для $x_1, x_1 + h$ и $x_1 + 2h$, мы получаем три формулы для численного дифференцирования второго порядка точности:

$$f'(x_1) = \frac{-3f(x_1) + 4f(x_1 + h) - f(x_1 + 2h)}{2h} + \frac{h^2}{3}f^{(3)}(\xi), \quad (3.13)$$

$$f'(x_1 + h) = \frac{f(x_1 + 2h) - f(x_1)}{2h} - \frac{h^2}{6}f^{(3)}(\xi), \quad (3.14)$$

$$f'(x_1 + 2h) = \frac{f(x_1) - 2f(x_1 + h) + 3f(x_1 + 2h)}{2h} + \frac{h^2}{3}f^{(3)}(\xi). \quad (3.15)$$

Несмотря на то, что все три формулы имеют точность $O(h^2)$, можно заметить, что остаточный член численного дифференцирования в случае граничных узлов x_1 и $x_1 + 2h$ в два раза больше, чем в случае центрального узла $x_1 + h$. Это связано с тем, что центральная формула использует значения функции в точках, расположенных по обе стороны узла, что повышает точность дифференцирования. Несложно показать, что формулы (3.13), (3.14) и (3.15) можно вывести с помощью разложения функции $f(x)$ в ряд Тейлора относительно точек x_1, $x_1 + h$ и $x_1 + 2h$ соответственно и последующего вычисления значений ряда Тейлора в оставшихся узлах. Тогда разложение в ряд Тейлора в центральном узле $x_1 + h$ дейтвительно имеет преимущество, так как остаточный член ряда Тейлора пропорционален разности от точки разложения до точки, в которой требуется вычислить значение ряда.

3.1.2 Метод разложения функции в ряд Тейлора

Метод разложения в ряд Тейлора становится удобен в случае, когда требуется построить формулу для вычисления высших производных. В качестве примера построим формулу для второй производной функции $f(x)$. Для этого допустим, что нам известны ее значения в точках $x_1 - h$, x_1 и $x_1 + h$, и разложим ее в ряд Тейлора в точке x_1:

$$f(x) = f(x_1) + f'(x_1)(x - x_1) + \frac{f''(x_1)}{2}(x - x_1)^2 + \frac{f'''(x_1)}{6}(x - x_1)^3 + \frac{f^{(4)}(\xi)}{24}(x - x_1)^4. \quad (3.16)$$

где $\xi \in (x_1; x)$. Тогда значения ряда в точках $x_1 - h$ и $x_1 + h$ будут равны:

$$f(x_1 - h) = f(x_1) - hf'(x_1) + \frac{h^2}{2}f''(x_1) - \frac{h^3}{6}f'''(x_1) + \frac{h^4}{24}f^{(4)}(\xi_1), \tag{3.17}$$

$$f(x_1 + h) = f(x_1) + hf'(x_1) + \frac{h^2}{2}f''(x_1) + \frac{h^3}{6}f'''(x_1) + \frac{h^4}{24}f^{(4)}(\xi_2). \tag{3.18}$$

где $\xi_1 \in (x_1 - h; x_1)$ и $\xi_2 \in (x; x_1 + h)$. Сложив два равенства, получаем:

$$\begin{aligned} & f(x_1 - h) - 2f(x_1) + f(x_1 + h) = h^2 f''(x_1) + \frac{h^4}{24}\left[f^{(4)}(\xi_1) + f^{(4)}(\xi_2)\right] \\ \Longrightarrow \quad & f''(x_1) = \frac{f(x_1 - h) - 2f(x_1) + f(x_1 + h)}{h^2} - \frac{h^2}{24}\left[f^{(4)}(\xi_1) + f^{(4)}(\xi_2)\right]. \end{aligned} \tag{3.19}$$

Предположим, что $f(x) \in C^4[x_1 - h; x_1 + h]$. Тогда по теореме о промежуточном значении существует такое $\xi \in (x_1 - h; x_1 + h)$, что

$$f^{(4)}(\xi) = \frac{1}{2}\left[f^{(4)}(\xi_1) + f^{(4)}(\xi_2)\right], \tag{3.20}$$

что в результате дает формулу численного дифференцирования второго порядка для второй производной:

$$f''(x_1) = \frac{f(x_1 - h) - 2f(x_1) + f(x_1 + h)}{h^2} - \frac{h^2}{12}f^{(4)}(\xi), \tag{3.21}$$

где $\xi \in (x_1 - h; x_1 + h)$.

3.1.3 Вычислительная неустойчивость операции дифференцирования

Как отмечалось во введении, погрешность округления может становиться в особенности заметной тогда, когда происходит вычитание двух близких значений. Именно это происходит во всех формулах численного дифференцирования, что делает операцию дифференцирования неустойчивой с вычислительной точки зрения. В качестве примера рассмотрим формулы численного дифференцирования второго порядка:

$$f'(x_1 + h) = \frac{f(x_1 + 2h) - f(x_1)}{2h} - \frac{h^2}{6}f^{(3)}(\xi), \tag{3.22}$$

и предположим, что при округлении значений $f(x_1 + 2h)$ и $f(x_1)$ вычислительная погрешность равна $e(x_1 + 2h)$ и $e(x_1)$, то есть

$$f(x_1 + 2h) = \tilde{f}(x_1 + 2h) + e(x_1 + 2h), \tag{3.23}$$

$$f(x_1) = \tilde{f}(x_1) + e(x_1). \tag{3.24}$$

Тогда полная погрешность, включающая погрешность метода и вычислительную погрешность, вычисляется следующим образом:

$$f'(x_1 + h) - \frac{\tilde{f}(x_1 + 2h) - \tilde{f}(x_1)}{2h} = \frac{e(x_1 + 2h) - e(x_1)}{2h} - \frac{h^2}{6}f^{(3)}(\xi). \tag{3.25}$$

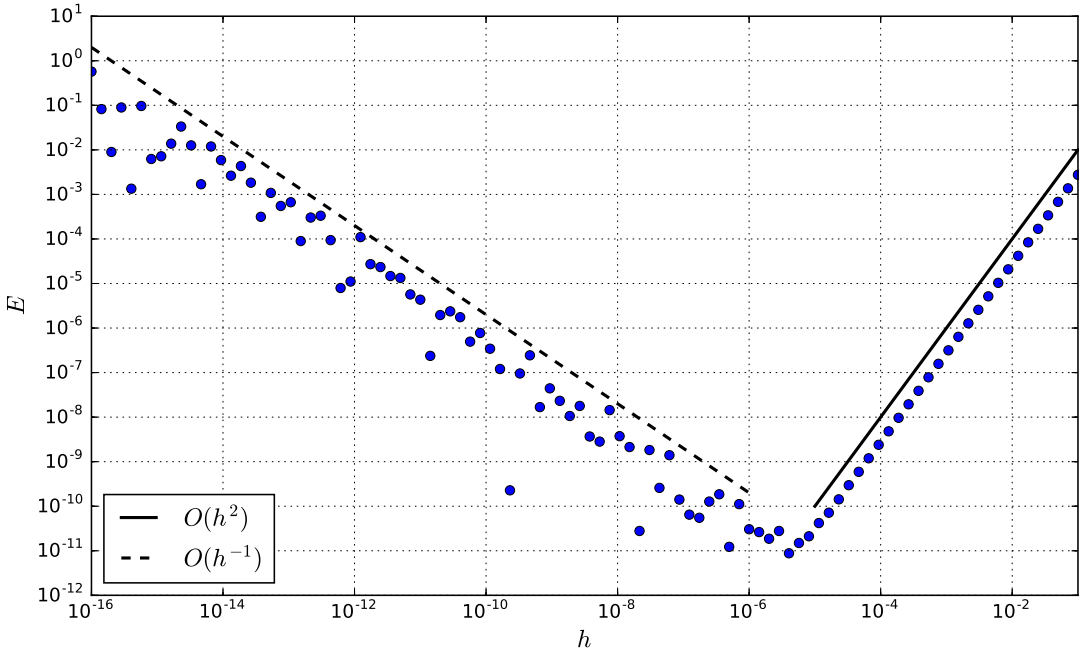

Рисунок 3.1 – Зависимость полной погрешности E, выражение для которой дано в (3.28), от шага дифференцирования h в случае формулы дифференцирования второго порядка для функции $f(x)$ и значения производной $f'(1/2)$.

Пусть вычислительная погрешность ограничена ϵ (например, машинным эпсилон) и пусть $f^{(3)}$ ограничена M. Тогда верным является следующее неравенство:

$$\left| f'(x_1 + h) - \frac{\tilde{f}(x_1 + 2h) - \tilde{f}(x_1)}{2h} \right| \leq \frac{\epsilon}{h} + \frac{h^2}{6} M. \tag{3.26}$$

Можно заметить, что при $h \to 0$ погрешность будет стремиться к бесконечности, что и обусловливаниет вычислительную неустойчивость численного дифференцирования. Несложно проверить, что выражение справа имеет минимум в точке

$$h_{opt} = \left(\frac{3\epsilon}{M} \right)^{\frac{1}{3}}, \tag{3.27}$$

что задает оптимальный шаг дифференцирования для данной формулы.

Рисунок 3.1 демонстрирует, как полная погрешность, определенная как

$$E = \left| f'(x_1 + h) - \frac{\tilde{f}(x_1 + 2h) - \tilde{f}(x_1)}{2h} \right|, \tag{3.28}$$

меняется в зависимости от шага дифференцирования h в случае использования формулы дифференцирования второго порядка для $f(x) = e^x$ и $f'(1/2)$. Обе оси графика имеют

логарифмический масштаб (так называемый log-log график), что позволяет легко выявить зависимости вида $E(h) = \alpha h^\gamma$. Можно заметить, что для $h \gtrsim 7 \cdot 10^{-6}$ полная погрешность уменьшается пропорционально h^2 при уменьшении h, что и предсказывается формулой (3.14). Однако для $h \lesssim 7 \cdot 10^{-6}$ вычислительная погрешность начинает доминировать в соответствии с правой частью неравенства (3.26) и полная погрешность начинает расти пропорционально $1/h$ при уменьшении h. В соответствии с формулой (3.27) оптимальный шаг дифференцирования равен

$$h_{opt} = \left(\frac{3 \cdot 2.2 \cdot 10^{-16}}{e^{1/2}} \right)^{\frac{1}{3}} \approx 7.4 \cdot 10^{-6}, \tag{3.29}$$

что действительно совпадает с оптимальным значением, определяемым численно из графика на рисунке 3.1.

3.2 Численное интегрирование

Базовая идея численного интегрирования состоит в том, чтобы аппроксимировать интеграл следующим образом:

$$\int_a^b f(x)dx \approx \sum_{i=1}^n a_i f(x_i), \quad a_i \in \mathbb{R}, \tag{3.30}$$

где выражение справа называется *квадратурой*. Как и в случае численного дифференцирования, простейшие квадратуры можно вывести, аппроксимируя $f(x)$ интерполяционным многочленом Лагранжа.

3.2.1 Формулы Ньютона-Котеса

Рассмотрим функцию $f(x) \in C^n[a;b]$ и n различных узлов $x_1, \ldots, x_n$. Тогда, раскладывая $f(x)$ в базисные многочлены Лагранжа, мы получаем

$$f(x) = \sum_{i=1}^n f(x_i) l_i(x) + \frac{\prod_{i=1}^n (x - x_i)}{n!} f^{(n)}(\xi(x)), \tag{3.31}$$

где $\xi(x) \in (a;b)$. Тогда интегрирование $f(x)$ на интервале $[a;b]$ дает:

$$\begin{aligned} \int_a^b f(x)dx &= \int_a^b \sum_{i=1}^n \left[f(x_i) l_i(x) \right] dx + \int_a^b \frac{\prod_{i=1}^n (x - x_i)}{n!} f^{(n)}(\xi(x))dx \\ &= \sum_{i=1}^n \left[f(x_i) \int_a^b l_i(x)dx \right] + \int_a^b \frac{\prod_{i=1}^n (x - x_i)}{n!} f^{(n)}(\xi(x))dx, \end{aligned} \tag{3.32}$$

что дает выражение для коэффициентов квадратуры a_i:

$$a_i = \int_a^b l_i(x)dx. \tag{3.33}$$

В случае, когда интерполяционные узлы $x_1, \ldots, x_n$ распределены равномерно, мы получаем *формулы Ньютона-Котеса для численного интегрирования*. Оценим остаточный член формул Ньютона-Котеса для n узлов $x_i = x_1 + (i-1)h$, где $i = 1, \ldots, n$ и $x_1 = a, x_n = b$:

$$\begin{aligned} R &= \frac{1}{n!} \int_a^b \prod_{i=1}^{n} (x - x_i) f^{(n)}(\xi(x)) dx \\ &= \frac{1}{n!} \int_a^b \prod_{i=1}^{n} \left[x - x_1 - (i-1)h\right] f^{(n)}(\xi(x)) dx. \end{aligned} \tag{3.34}$$

Представим переменную x как $x = x_1 + (t-1)h$, что дает $dx = hdt$ и $t = \dfrac{x - x_1}{h} + 1$. Тогда, произведя замену в выражении выше, имеем:

$$R = \frac{h^n}{n!} \int_1^n \prod_{i=1}^{n} (t - i)\, f^{(n)}(\xi(t)) dt. \tag{3.35}$$

Таким образом, порядок точности можно оценить как $O(h^n)$. Чтобы получить более точное выражение для остаточного члена, необходимо заметить, что полином в подинтегральном выражении меняет знак только в узлах x_i. Тогда, пользуя первой теоремой о среднем значении, мы получаем:

$$R = \frac{h^n}{n!} \sum_{j=1}^{n-1} f^{(n)}(\xi_j) \int_j^{j+1} \prod_{i=1}^{n} (t - i)\, dt. \tag{3.36}$$

где $\xi_j \in [x_j; x_{j+1}]$. Эта форма R приводит к теореме об остаточном члене формул Ньютона-Котеса, которую мы рассмотрим без доказательства.

Теорема 3.2.1. *Пусть $a_i = \int_a^b l_i(x)dx$ и заданы узлы $x_i = x_1 + (i-1)h$, где $i = 1, \ldots, n$ и $x_1 = a, x_n = b$. Тогда для четных n и $f(x) \in C^n[a; b]$ существует $\xi \in (a; b)$ такое, что:*

$$\int_a^b f(x)dx = \sum_{i=1}^{n} a_i f(x_i) + \frac{h^{n+1}}{n!} f^{(n)}(\xi) \int_1^n \prod_{i=1}^{n} (t - i)\, dt, \tag{3.37}$$

в то время как для нечетных n и $f(x) \in C^{n+1}[a; b]$ существует $\xi \in (a; b)$ такое, что:

$$\int_a^b f(x)dx = \sum_{i=1}^{n} a_i f(x_i) + \frac{h^{n+2}}{(n+1)!} f^{(n+1)}(\xi) \int_1^n (t-1) \prod_{i=1}^{n} (t - i)\, dt, \tag{3.38}$$

Основным следствием теоремы является тот факт, что предпочтительным является нечетное число узлов. Это связано с формой полинома в подинтегральном выражении в (3.35), который в случае нечетного числа узлов имеет равное количество отрезков, на которых функция всюду положительна и всюду отрицальна, что снижает погрешность метода.

3.2.2 Формулы трапеций и Симпсона

Частными случаями формул Ньютона-Котеса являются формула трапеций ($n = 2$) и формула Симпсона ($n = 3$). В связи с их распространенностью рассмотрим вывод этих формул с явными выражением для остаточного члена. Так, для $n = 2$ выражение (3.32) принимает вид:

$$\int_a^b f(x)dx = f(x_1)\int_{x_1}^{x_2}\frac{(x-x_2)}{(x_1-x_2)}dx + f(x_2)\int_{x_1}^{x_2}\frac{(x-x_1)}{(x_2-x_1)}dx + \\ + \int_{x_1}^{x_2}\frac{(x-x_1)(x-x_2)}{2}f''(\xi(x))dx, \quad (3.39)$$

что после интегрирования линейных полиномов дает:

$$\int_a^b f(x)dx = \frac{h}{2}\left[f(x_1)+f(x_2)\right] + \int_{x_1}^{x_2}\frac{(x-x_1)(x-x_2)}{2}f''(\xi(x))dx. \quad (3.40)$$

Так как полином внутри подинтегрального выражения не меняет знак для $x \in [a;b]$, первая теорема о среднем значении позволяет записать выражение

$$\int_a^b f(x)dx = \frac{h}{2}\left[f(x_1)+f(x_2)\right] + f''(\xi)\int_{x_1}^{x_2}\frac{(x-x_1)(x-x_2)}{2}dx, \quad (3.41)$$

где $\xi \in (a;b)$. Таким образом мы получаем выражение для *формулы трапеций*:

$$\int_a^b f(x)dx = \frac{h}{2}\left[f(x_1)+f(x_2)\right] - \frac{h^3}{12}f''(\xi), \quad (3.42)$$

где, как мы помним, $x_1 = a$, $x_2 = b$ и $h = b - a$.

Для вывода формулы Симпсона с $n = 3$ мы воспользуемся альтернативным подходом, который сделает вывод остаточного члена проще, а также в очередной раз продемонстрирует связь между интерполяционным многочленом Лагранжа и рядом Тейлора. Рассмотрим разложение функции $f(x)$ в ряд Тейлора в промежуточном узле x_2:

$$f(x) = f(x_2) + f'(x_2)(x-x_2) + \frac{f''(x_2)}{2}(x-x_2)^2 + \frac{f'''(x_2)}{6}(x-x_2)^3 + \frac{f^{(4)}(\xi(x))}{24}(x-x_2)^4. \quad (3.43)$$

Тогда интеграл от $f(x)$ принимает вид:

$$\int_a^b f(x)dx = \left[f(x_2)(x-x_2) + \frac{f'(x_2)}{2}(x-x_2)^2 + \frac{f''(x_2)}{6}(x-x_2)^3 + \frac{f'''(x_2)}{24}(x-x_2)^4\right]_{x_1}^{x_3} + \\ + \int_{x_1}^{x_3}\frac{f^{(4)}(\xi(x))}{24}(x-x_2)^4dx. \quad (3.44)$$

В силу четности степеней члены, включающие в себя нечетные производные, обращаются в ноль. Более того, так как $(x-x_2)^4 \geq 0$, мы можем применить первую теорему о среднем значении, что в результате дает:

$$\int_a^b f(x)dx = 2hf(x_2) + \left[\frac{f''(x_2)}{6}(x-x_2)^3\right]_{x_1}^{x_3} + \frac{f^{(4)}(\xi)}{24}\int_{x_1}^{x_3}(x-x_2)^4dx, \quad (3.45)$$

где $\xi \in (x_1; x_3)$. Для того, чтобы избавиться от второй производной, воспользуемся формулой для численного дифференцирования с остаточным членом (3.21). После подстановки равенство выше принимает вид:

$$\begin{aligned}\int_a^b f(x)dx &= 2hf(x_2) + \frac{h^3}{3}\left[\frac{f(x_1) - 2f(x_2) + f(x_3)}{h^2} - \frac{h^2}{12}f^{(4)}(\xi_1)\right] + \frac{f^{(4)}(\xi_2)}{24}\int_{x_1}^{x_3}(x - x_2)^4 dx \\ &= \frac{h}{3}\left[f(x_1) + 4f(x_2) + f(x_3)\right] - \frac{h^5}{36}f^{(4)}(\xi_1) + \frac{h^5}{60}f^{(4)}(\xi_2),\end{aligned} \tag{3.46}$$

где $\xi_1, \xi_2 \in (x_1; x_3)$. Так как в общем случае $\xi_1 \neq \xi_2$ нам необходимо каким-то образом скомбинировать два последних члена. Для этого рассмотрим остаточный член в общем случае и предположим, что существует такое $\xi \in (x_1; x_3)$, что:

$$\int_a^b f(x)dx = \frac{h}{3}\left[f(x_1) + 4f(x_2) + f(x_3)\right] + \alpha f^{(4)}(\xi), \tag{3.47}$$

где $\alpha \in \mathbb{R}$ – неопределенный коэффициент. Для того чтобы найти его, заметим, что $f^{(4)}(\xi) = 24$ для любого нормированного многочлена 4-й степени. Тогда рассмотрим в качестве $f(x)$ многочлен $(x - x_2)^4$:

$$\begin{aligned}&\int_{x_1}^{x_3}(x - x_2)^4 dx = \frac{2h^5}{5} \\ \implies\quad &\frac{2h^5}{5} = \frac{h}{3}\left[(x_1 - x_2)^4 + 4(x_2 - x_2)^4 + (x_3 - x_2)^4\right] - 24\alpha \\ \implies\quad &\alpha = -\frac{h^5}{90}.\end{aligned} \tag{3.48}$$

Таким образом мы получаем *формулу Симпсона* с явным выражением для остаточного члена:

$$\int_a^b f(x)dx = \frac{h}{3}\left[f(x_1) + 4f(x_2) + f(x_3)\right] - \frac{h^5}{90}f^{(4)}(\xi), \tag{3.49}$$

где $x_1 = a$, $x_2 = x_1 + h$ и $x_3 = x_1 + 2h = b$. Малость погрешности формулы Симпсона, $O(h^5)$, при использовании всего лишь трех узлов объясняет ее частое использование в реальных приложениях.

Необходимо отметить, что такая малая погрешность сохраняется только при достаточной гладкости функции $f(x)$. Для того, чтобы численно исследовать зависимость остаточного члена формулы Симпсона от гладкости интегрируемой функции, мы проведет следующий вычислительный эксперимент. Рассмотрим две функции $f_1(x) = e^x$ и $f_2(x) = |x|$, первая из которых является бесконечно гладкой, а вторая лишь непрерывной. Построим на основе последовательности шагов $\{h_i\}_{i=1}^n$ соответствующую последовательность интегралов:

$$\{I_i^{(1)}\}_{i=1}^n, \quad \text{где } I_i^{(1)} = \int_{\frac{1}{2}-h_i}^{\frac{1}{2}+h_i} e^x dx = e^{\frac{1}{2}+h_i} - e^{\frac{1}{2}-h_i}, \tag{3.50}$$

$$\{I_i^{(2)}\}_{i=1}^n, \quad \text{где } I_i^{(2)} = \int_{-h_i}^{h_i} |x| dx = h_i^2, \tag{3.51}$$

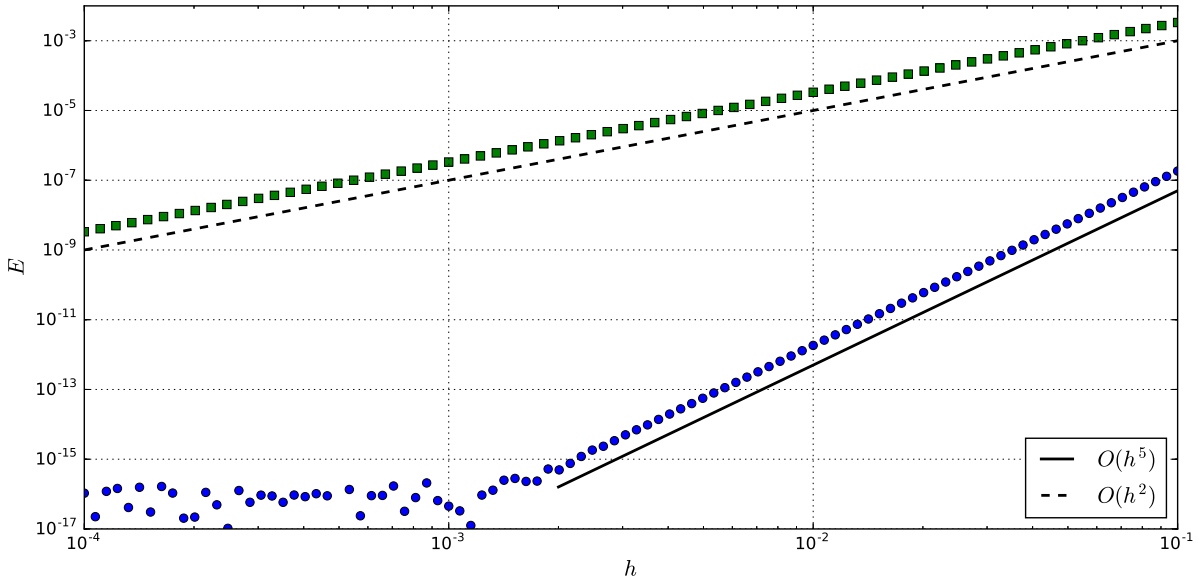

Рисунок 3.2 – Зависимость полной погрешности E численного интегрирования с помощью формулы Симпсона от шага интегрирования h для интегралов $\{I_i^{(1)}\}_{i=1}^n$ (синие круги) и $\{I_i^{(2)}\}_{i=1}^n$ (зеленые квадраты).

где $x = 1/2$ и $x = 0$ являются центральными узлами численного интегрирования для $f_1(x)$ и $f_2(x)$ соответственно. Рассчитав те же интегралы с помощью формулы Симпсона и найдя полную погрешность E, включающую в себя как остаточный член, так и вычислительную погрешность, для каждого случая, мы получаем рисунок 3.2. Для начала заметим, что численное интегрирование устойчиво с вычислительной точки зрения, и полная погрешность полностью соответствует остаточному члену вплоть до тех пор, пока E не достигнет машинного эпсилон. Сам остаточный член, как мы видим, пропорционален $O(h^5)$ в случае бесконечно гладкой функции $f_1(x)$, что и предполагается равенством (3.49), в то время как в случае функции $f_2(x)$ погрешность метода пропорциональна лишь $O(h^2)$, что связано с тем, что уже первая производная от $f_2(x)$ имеет разрыв, что делает формулу (3.49) нерабочей.

3.2.3 Формула средних

Отдельно рассмотрим случай, когда на отрезке $[a; b]$ мы имеем только один узел, расположенный в центре отрезка, т.е. $x_1 = \frac{a+b}{2}$. Разложим функцию $f(x) \in C^2[a; b]$ в ряд Тейлора в этом узле:

$$f(x) = f(x_1) + f'(x_1)(x - x_1) + \frac{f''(\xi(x))}{2}(x - x_1)^2. \tag{3.52}$$

Тогда интеграл от $f(x)$ вычисляется следующим образом:

$$\int_a^b f(x)dx = \left[f(x_1)(x-x_1) + \frac{1}{2}f'(x_1)(x-x_1)^2\right]_a^b + \int_a^b \frac{f''(\xi(x))}{2}(x-x_1)^2 dx. \tag{3.53}$$

Так как выражение $(x-x_1)^2$ всюду неотрицательно, мы можем воспользоваться первой теоремой о среднем значении, что дает:

$$\int_a^b f(x)dx = f(x_1)(b-a) + \frac{f''(\xi)}{2}\int_a^b (x-x_1)^2 dx, \tag{3.54}$$

что после интегрирования приводит к *формуле средних*:

$$\int_a^b f(x)dx = f(x_1)(b-a) + \frac{f''(\xi)}{24}(b-a)^3. \tag{3.55}$$

Заметим, что остаточный член в формуле средних пропорционален $O(h^3)$, где $h = b-a$, т.е. имеет тот же порядок погрешности, что и формула трапеций. Однако формула средних использует только один узел и, соответственно, только одно значение функции $f(x)$, что делает ее предпочтительной с вычислительной точки зрения.

3.2.4 Степень точности численного интегрирования

Несмотря на то, что оценка остаточного члена с точки зрения его зависимости от шага интегрирования h является удобной для анализа и сравнения различных формул интегрирования, на практике часто пользуются другой оценкой, напрямую следующей из формы остаточного члена. Этой оценкой является степень точности численного интегрирования.

Определение 3.2.1. *Степенью точности квадратуры называется такое наибольшее* $n \in \mathbb{N}$*, что формула квадратуры дает точный результат для всех* $x^i, i = 0, \dots n$.

Легко проверить, что вследствие линейности операции интегрирования, это определение автоматически приводит к следующей теореме:

Теорема 3.2.2. *Степень точности квадратуры равна* $n \in \mathbb{N}$ *тогда и только тогда, когда остаточный член равен нулю для всех многочленов степеней* $i = 0, \dots n$ *и не равен нулю для хотя бы одного многочлена степени* $n+1$.

Форма остаточного члена в формулах трапеций и Симпсона позволяет заключить, что они имеют степени точности 1 и 3 соответственно. Например, интегрируя любой полином, имеющий степень $0, 1, 2$ или 3, формула Симпсона будет давать точный результат. В дальнейшем мы рассмотрим способ построения квадратур, имеющих наибольшую степень точности при наименьшем числе используемых узлов.

3.2.5 Составные формулы численного интегрирования

Как и в случае с интерполяцией, использование формул Ньютона-Котеса большого порядка для увелечения точности интегрирования на больших отрезках приводит к паразитным решениям. В таком случае предпочтительным является кусочный подход к интегрированию. Мы построим составные формулы из формул Симпсона, трапеций и средних, хотя подобный подход, очевидно, можно применить к любым другим квадратурам.

Рассмотрим интеграл $\int_a^b f(x)dx$. Разделим отрезок $[a;b]$ на четное число подотрезков n и применим формулу Симпсона на каждом из них. Тогда для $f(x) \in C^4[a;b]$, $h = \frac{b-a}{n}$ и $x_i = a + (i-1)h$, где $i = 1, \ldots, n+1$ мы имеем:

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{i=1}^{n/2} \int_{x_{2i-1}}^{x_{2i+1}} f(x)dx \\ &= \frac{h}{3} \sum_{i=1}^{n/2} [f(x_{2i-1}) + 4f(x_{2i}) + f(x_{2i+1})] - \frac{h^5}{90} \sum_{i=1}^{n/2} f^{(4)}(\xi_i),\end{aligned} \tag{3.56}$$

где $\xi_i \in (x_{2i-1}; x_{2i+1})$. Заметим, что все нечетные узлы за исключением x_1 и x_{n+1} дважды повторяются в сумме, что позволяет упростить выражение:

$$\int_a^b f(x)dx = \frac{h}{3} \left[f(x_1) + 2 \sum_{i=1}^{n/2-1} f(x_{2i+1}) + 4 \sum_{i=1}^{n/2} f(x_{2i}) + f(x_{n+1}) \right] - \frac{h^5}{90} \sum_{i=1}^{n/2} f^{(4)}(\xi_i). \tag{3.57}$$

Для упрощения выражения для остаточного члена заметим, что так как $f(x) \in C^4[a;b]$, мы имеем:

$$\begin{aligned}&\min_{x\in[a;b]} f^{(4)}(x) \le f^{(4)}(\xi_i) \le \max_{x\in[a;b]} f^{(4)}(x) \\ \Longrightarrow &\frac{n}{2} \min_{x\in[a;b]} f^{(4)}(x) \le \sum_{i=1}^{n/2} f^{(4)}(\xi_i) \le \frac{n}{2} \max_{x\in[a;b]} f^{(4)}(x) \\ \Longrightarrow &\min_{x\in[a;b]} f^{(4)}(x) \le \frac{2}{n} \sum_{i=1}^{n/2} f^{(4)}(\xi_i) \le \max_{x\in[a;b]} f^{(4)}(x).\end{aligned} \tag{3.58}$$

Тогда по теореме о промежуточном значении мы получаем:

$$f^{(4)}(\xi) = \frac{2}{n} \sum_{i=1}^{n/2} f^{(4)}(\xi_i), \tag{3.59}$$

где $\xi \in (a;b)$. Остаточный член в таком случае принимает следующую форму:

$$\begin{aligned}R &= -\frac{h^5}{90} \sum_{i=1}^{n/2} f^{(4)}(\xi_i) \\ &= -\frac{nh^5}{180} f^{(4)}(\xi) \\ &= -\frac{(b-a)h^4}{180} f^{(4)}(\xi).\end{aligned} \tag{3.60}$$

Заметим, что по сравнению со стандартной формулой Симпсона, остаточный член составной формулы Симпсона пропорционален $O(h^4)$. Результирующее выражение для составной формулы представлено в следующей теореме.

Теорема 3.2.3. *Пусть $x_i = a + (i-1)h$, $h = \frac{b-a}{n}$ и $i = 1, \ldots, n+1$, где n – четное число. Тогда существует такое $\xi \in (a;b)$ для $f(x) \in C^4[a;b]$, что составная формула Симпсона имеет вид:*

$$\int_a^b f(x)dx = \frac{h}{3}\left[f(x_1) + 2\sum_{i=1}^{n/2-1} f(x_{2i+1}) + 4\sum_{i=1}^{n/2} f(x_{2i}) + f(x_{n+1})\right] - \frac{(b-a)h^4}{180} f^{(4)}(\xi). \quad (3.61)$$

Аналогично можно вывести составные формулы трапеций и средних, в которых n может быть как четным, так и нечетным. Мы приведет соответствующие теоремы без доказательства.

Теорема 3.2.4. *Пусть $x_i = a + (i-1)h$, $h = \frac{b-a}{n}$ и $i = 1, \ldots, n+1$, где $n \in N$. Тогда существует такое $\xi \in (a;b)$ для $f(x) \in C^2[a;b]$, что составная формула трапеций имеет вид:*

$$\int_a^b f(x)dx = \frac{h}{2}\left[f(x_1) + 2\sum_{i=2}^{n} f(x_i) + f(x_{n+1})\right] - \frac{(b-a)h^2}{12} f''(\xi). \quad (3.62)$$

Теорема 3.2.5. *Пусть $x_i = a + \frac{(2i-1)h}{2}$, $h = \frac{b-a}{n}$ и $i = 1, \ldots, n$, где $n \in N$. Тогда существует такое $\xi \in (a;b)$ для $f(x) \in C^2[a;b]$, что составная формула средних имеет вид:*

$$\int_a^b f(x)dx = h\sum_{i=1}^{n} f(x_i) + \frac{(b-a)h^2}{6} f''(\xi). \quad (3.63)$$

3.2.6 Вычислительная устойчивость операции интегрирования

По сравнению с дифференцированием, операция интегрирования, как мы вскоре покажем, способна к стабилизации вычислительной погрешности. Интуитивное объяснение этого эффект заключается в том, что интегрирование предполагает суммирование близких значений, в то время как дифференцирование вычисляет их разность.

Рассмотрим составную формулу Симпсона и предположим, что значение $f(x)$ в точке x_i вычисляется с погрешностью округления e_i:

$$f(x_i) = \tilde{f}(x_i) + e_i, \quad i = 1, \ldots, n+1. \quad (3.64)$$

Тогда полная погрешность округления, накапливаемая составной формулой Симпсона, может быть оценена следующим образом:

$$\begin{aligned} e(h) &= \frac{h}{3}\left[e_1 + 2\sum_{i=1}^{n/2-1} e_{2i+1} + 4\sum_{i=1}^{n/2} e_{2i} + e_{n+1}\right] \\ &\leq \frac{h}{3}\left[|e_1| + 2\sum_{i=1}^{n/2-1} |e_{2i+1}| + 4\sum_{i=1}^{n/2} |e_{2i}| + |e_{n+1}|\right]. \end{aligned} \quad (3.65)$$

Предположим, что погрешность округления ограничена, например, машинным эпсилон: $|e_i| \leq \epsilon$, $i = 1, \ldots, n+1$. Тогда полная погрешность погрешность оценивается как:

$$\begin{aligned} e(h) &\leq \frac{h\epsilon}{3}\left[1 + 2\left(\frac{n}{2} - 1\right) + 4\frac{n}{2} + 1\right] \\ &= nh\epsilon \\ &= (b-a)\epsilon. \end{aligned} \tag{3.66}$$

Этот результат показывает, что верхняя грань для накопленной погрешности округления не зависит от n или h, что означает, что увеличение числа подотрезков не приводит к дестабилизации полной погрешности. Действительно, из рисунка 3.2 можно заметить, что полная погрешность интегрирования падает до тех пор, пока она не достигнет значения, сравнимого с машинным эпсилон, после чего уменьшение погрешности становится невозможным, и она стабилизируется на уровне машинного эпсилон.

3.2.7 Квадратуры Гаусса

Формулы Ньютона-Котеса были построены, исходя из равномерного распределения узлов. Как и в случае интерполяции, логично предположить, что существует иное, неравномерное распределение узлов, которое позволило бы максимизировать точность. Как уже было сказано ранее, на практике часто возникает необходимость максимизировать степень точности, т.е. максимизировать степень полинома, интегрирование которого с помощью данной квадратуры дает точный результат при заданном количестве узлов. Эту задачу решают квадратуры Гаусса, суть которых сводится к нахождению таких $x_1, \ldots, x_n$ и $c_1, \ldots, c_n$, что приближение

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} c_i f(x_i) \tag{3.67}$$

максимизирует степень точности. Так как всего мы имеем $2n$ оптимизируемых параметров, логично предположить, что полином $2n-1$ степени, имеющий так же $2n$ параметров, может быть интегрирован точно при правильно подобранных параметрах. Очевидно, что если квадратура дает точный результат для любого полинома этой степени, то все полиномы низших степеней автоматически интегрируются точно как частные случаи.

Рассмотрим случай $n = 2$ и интервал интегрирования $[-1; 1]$. Тогда квадратура принимает вид:

$$\int_a^b f(x)dx \approx c_1 f(x_1) + c_2 f(x_2). \tag{3.68}$$

Мы ожидаем, что эта квадратура дает точный результат при интегрировании полинома третьей степени:

$$\begin{aligned} & f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \\ \implies \quad & \int_{-1}^{1} f(x)dx = a_0 \int_{-1}^{1} dx + a_1 \int_{-1}^{1} x dx + a_2 \int_{-1}^{1} x^2 dx + a_3 \int_{-1}^{1} x^3 dx, \end{aligned} \tag{3.69}$$

где $a_0, a_1, a_2, a_3 \in \mathbb{R}$ – произвольные константы. Можно заметить, что квадратура будет точно вычислять интеграл этого полинома тогда, когда точно будут вычисляться интегралы

от функций $f(x) = 1$, $f(x) = x$, $f(x) = x^2$, $f(x) = x^3$. В таком случае мы получаем систему уравнений:

$$\begin{cases} c_1 + c_2 = \int_{-1}^{1} dx = 2, \\ c_1 x_1 + c_2 x_2 = \int_{-1}^{1} x dx = 0, \\ c_1 x_1^2 + c_2 x_2^2 = \int_{-1}^{1} x^2 dx = \frac{2}{3}, \\ c_1 x_1^3 + c_2 x_2^3 = \int_{-1}^{1} x^3 dx = 0. \end{cases} \tag{3.70}$$

Из первых двух уравнений мы получаем:

$$c_1 = \frac{2x_2}{x_2 - x_1} \tag{3.71}$$

$$c_2 = -\frac{2x_1}{x_2 - x_1}, \tag{3.72}$$

что при подстановке в третье дает:

$$x_1 x_2 = -\frac{1}{3}. \tag{3.73}$$

Тогда четвертое уравнение становится:

$$x_1 + x_2 = 0, \tag{3.74}$$

что в результате дает следующее решение:

$$c_1 = 1, \tag{3.75}$$

$$c_2 = 1, \tag{3.76}$$

$$x_1 = -\frac{1}{\sqrt{3}}, \tag{3.77}$$

$$x_2 = \frac{1}{\sqrt{3}}. \tag{3.78}$$

Таким образом квадратура Гаусса со степенью точности 3 имеет следующий вид:

$$\int_{-1}^{1} f(x)dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right). \tag{3.79}$$

Подобным образом мы можем построить квадратуры Гаусса произвольной степени точности, однако из-за нелинейности решаемой системы уравнений это становится алгебраически утомительным. Для рассмотрения более удобного подхода нам необходимо ознакомиться с понятием линейно-независимых функций и многочленами Лежандра.

3.2.8 Ортогональные полиномы и многочлены Лежандра

По аналогии с векторами мы введем понятие линейно независимых функций.

Определение 3.2.2. *Система функций $\{\phi_1, \ldots, \phi_n\}$ называется линейно независимой на $[a; b]$, если*

$$\sum_{i=1}^{n} c_i \phi_1(x) = 0 \quad \Longleftrightarrow \quad c_1 = \cdots = c_n = 0. \tag{3.80}$$

В противном случае система функций называется линейно зависимой.

Важное свойство полиномов состоит в том, что любые полиномы различных степеней являются линейно независимыми. Это свойство доказывается в следующей теореме.

Теорема 3.2.6. *Пусть $\{\phi_i\}_{i=0}^n$ – система полиномов, где полином ϕ_i имеет i-ую степень. Тогда $\{\phi_i\}_{i=0}^n$ является линейно независимой системой функций на $[a;b]$.*

Доказательство. Рассмотрим линейную комбинацию:

$$P(x) = \sum_{i=0}^{n} c_i \phi_i(x). \tag{3.81}$$

Пусть коэффициенты c_i имеют такие значения, что $P(x) = 0$. В свою очередь $P(x)$ является полиномом степени n, что означает, что для удовлетворения условия $P(x) = 0$ коэффициенты при всех степенях x^i, $i = 0, \ldots, n$ должны быть равны нулю. Это возможно только тогда, когда $c_n = 0$. В результате мы получаем:

$$P(x) = \sum_{i=0}^{n-1} c_i \phi_i(x). \tag{3.82}$$

Продолжая эту логику, мы получаем $c_0 = \cdots = c_n = 0$, что означает, что система $\{\phi_i\}_{i=0}^n$ является линейно независимой. □

Также добавим, что если система полиномов $\{\phi_i\}_{i=0}^n$ является линейно независимой, то любой полином степени меньшей или равной n можно представить в виде единственной линейной комбинации полиномов $\{\phi_i\}_{i=0}^n$. Более того, если такая система полиномов является к тому же ортогональной, то верно следующее равенство:

$$\langle \phi_n(x), P_k(x) \rangle_\omega = \int_a^b \omega(x) \phi_n(x) P_k(x) dx = 0, \tag{3.83}$$

где $P_k(x)$ – полиномом любой степени $k < n$. Действительно, мы может разложить $P_k(x)$ в линейную комбинацию полиномов $\phi(x)$:

$$P_k(x) = \sum_{i=0}^{k} c_i \phi_i(x), \tag{3.84}$$

что дает

$$\int_a^b \omega(x) \phi_n(x) P_k(x) dx = \sum_{i=0}^{k} c_i \int_a^b \omega(x) \phi_n(x) \phi_i(x) dx = 0. \tag{3.85}$$

Линейно независимая система функций необязательно является ортогональной. Для случая полиномов мы можем построить ортогональную систему полиномов, адаптировав процесс Грама-Шмидта, хорошо известный из курса линейной алгебры.

Теорема 3.2.7. *Пусть система полиномов* $\{\phi_i\}_{i=0}^n$ *определена следующим образом на* $[a;b]$ *относительно весовой функции* $\omega(x)$*:*

$$\phi_0(x) = 1 \tag{3.86}$$

$$\phi_1(x) = x - \frac{\langle x\phi_0, \phi_0\rangle_\omega}{\langle \phi_0, \phi_0\rangle_\omega} \tag{3.87}$$

$$\phi_k(x) = x\phi_{k-1}(x) - \frac{\langle x\phi_{k-1}, \phi_{k-1}\rangle_\omega}{\langle \phi_{k-1}, \phi_{k-1}\rangle_\omega}\phi_{k-1}(x) - \frac{\langle x\phi_{k-2}, \phi_{k-1}\rangle_\omega}{\langle \phi_{k-2}, \phi_{k-2}\rangle_\omega}\phi_{k-2}(x), \tag{3.88}$$

где $x \in [a;b]$ *и* $k \geq 2$*. Тогда эта система является ортогональной на* $[a;b]$ *с весом* $\omega(x)$*.*

Доказательство. Рассмотрим систему полиномов $\{\chi_k(x)\}_{k=0}^n = \{1, x, x^2, \dots\}$, которая по теореме 3.2.6 является линейно независимой. Пусть $\phi_0(x) = \chi_0(x) = 1$. Процесс Грама-Шмидта предполагает, что $\phi_1(x)$ ищется в следующей форме:

$$\phi_1 = \chi_1 + c_{01}\phi_0, \tag{3.89}$$

где c_{01} находится из условия ортогональности ϕ_0 и ϕ_1:

$$\begin{aligned} 0 = \langle \phi_0, \phi_1\rangle_\omega &= \langle \phi_0, \chi_1\rangle_\omega + c_{01}\langle \phi_0\phi_0\rangle_\omega \\ \implies \quad c_{01} &= -\frac{\langle \phi_0, \chi_1\rangle_\omega}{\langle \phi_0\phi_0\rangle_\omega}. \end{aligned} \tag{3.90}$$

Тогда $\phi_1(x)$ имеет вид:

$$\phi_1 = x - \frac{\langle \phi_0, x\rangle_\omega}{\langle \phi_0, \phi_0\rangle_\omega} = x - \frac{\langle x\phi_0, \phi_0\rangle_\omega}{\langle \phi_0, \phi_0\rangle_\omega}. \tag{3.91}$$

Для $\phi_2(x)$ имеем:

$$\phi_2 = \chi_2 + c_{20}\phi_0 + c_{21}\phi_1. \tag{3.92}$$

Функция $\phi_2(x)$ должна быть ортогональна $\phi_0(x)$ и $\phi_1(x)$:

$$0 = \langle \phi_0, \phi_2\rangle_\omega = \langle \phi_0, \chi_2\rangle_\omega + c_{20}\langle \phi_0, \phi_0\rangle_\omega + c_{21}\langle \phi_0, \phi_1\rangle_\omega, \tag{3.93}$$

$$0 = \langle \phi_1, \phi_2\rangle_\omega = \langle \phi_1, \chi_2\rangle_\omega + c_{20}\langle \phi_1, \phi_0\rangle_\omega + c_{21}\langle \phi_1, \phi_1\rangle_\omega. \tag{3.94}$$

$$\tag{3.95}$$

Тогда, помня о взаимной ортогональности ϕ_0 и ϕ_1, находим c_{20} и c_{21}:

$$c_{20} = -\frac{\langle \phi_0, \chi_2\rangle_\omega}{\langle \phi_0, \phi_0\rangle_\omega}, \tag{3.96}$$

$$c_{21} = -\frac{\langle \phi_1, \chi_2\rangle_\omega}{\langle \phi_1, \phi_1\rangle_\omega}, \tag{3.97}$$

$$\tag{3.98}$$

что дает следующее выражение для функции ϕ_2:

$$\phi_2 = \chi_2 - \frac{\langle \phi_0, \chi_2\rangle_\omega}{\langle \phi_0, \phi_0\rangle_\omega}\phi_0 - \frac{\langle \phi_1, \chi_2\rangle_\omega}{\langle \phi_1, \phi_1\rangle_\omega}\phi_1. \tag{3.99}$$

Продолжая тем же образом вплоть до ϕ_k, мы получаем:

$$\phi_k = \chi_k - \sum_{j=0}^{k-1} \frac{\langle \phi_j, \chi_k \rangle_\omega}{\langle \phi_j, \phi_j \rangle_\omega} \phi_j. \tag{3.100}$$

Это выражение можно упростить, если заметить, что полином степени k может быть построен как $\chi_k = x\phi_{k-1}$. Тогда ϕ_k принимает форму:

$$\begin{aligned} \phi_k &= x\phi_{k-1} - \sum_{j=0}^{k-1} \frac{\langle \phi_j, x\phi_{k-1} \rangle_\omega}{\langle \phi_j, \phi_j \rangle_\omega} \phi_j \\ &= x\phi_{k-1} - \sum_{j=0}^{k-1} \frac{\langle x\phi_j, \phi_{k-1} \rangle_\omega}{\langle \phi_j, \phi_j \rangle_\omega} \phi_j. \end{aligned} \tag{3.101}$$

Так как $x\phi_j$ является полиномом степени $j+1$, мы имеем

$$\langle x\phi_j, \phi_{k-1} \rangle_\omega = 0 \tag{3.102}$$

для $j < k-2$. Тогда для ϕ_k получаем:

$$\phi_k = x\phi_{k-1} - \frac{\langle x\phi_{k-1}, \phi_{k-1} \rangle_\omega}{\langle \phi_{k-1}, \phi_{k-1} \rangle_\omega} \phi_{k-1} - \frac{\langle x\phi_{k-2}, \phi_{k-1} \rangle_\omega}{\langle \phi_{k-2}, \phi_{k-2} \rangle_\omega} \phi_{k-2}. \tag{3.103}$$

□

Многочлены Лежандра могут быть получены различными способами, однако мы рассмотрим их как ортогональную систему полиномов, построенную с помощью процесса Грама-Шмидта. Действительно, для случая $\omega(x) = 1$ и интервала $[-1; 1]$ мы имеем:

$$\phi_0(x) = 1 \tag{3.104}$$

$$\phi_1(x) = x - \frac{\int_{-1}^{1} x dx}{\int_{-1}^{1} dx} = x \tag{3.105}$$

$$\phi_2(x) = \left(x - \frac{\int_{-1}^{1} x^3 dx}{\int_{-1}^{1} x^2 dx} \right) x - \frac{\int_{-1}^{1} x^2 dx}{\int_{-1}^{1} dx} = x^2 - \frac{1}{3}, \tag{3.106}$$

$$\phi_3(x) = \left(x - \frac{\int_{-1}^{1} x \left(x^2 - \frac{1}{3}\right)^2 dx}{\int_{-1}^{1} \left(x^2 - \frac{1}{3}\right)^2 dx} \right) \left(x^2 - \frac{1}{3} \right) - \frac{\int_{-1}^{1} x^2 \left(x^2 - \frac{1}{3}\right) dx}{\int_{-1}^{1} x^2 dx} x = x^3 - \frac{3}{5}x. \tag{3.107}$$

Остальные многочлены Лежандра могут быть получены аналогичным образом.

Наконец, мы имеем возможность доказать следующую теорему, которая постулирует, что квадратуры Гаусса могут быть построены, если в качестве узлов выбраны корни соответствующего многочлена Лежандра.

Теорема 3.2.8. *Пусть $x_1, \ldots, x_n$ являются корнями полинома Лежандра n-ой степени $\phi_n(x)$, и пусть коэффициенты $c_1, \ldots, c_n$ определены следующим образом:*

$$c_i = \int_{-1}^{1} l_i(x)dx = \int_{-1}^{1} \prod_{j=1, i\neq j}^{n} \frac{x - x_j}{x_i - x_j} dx. \tag{3.108}$$

Тогда, если $P_m(x)$ является полиномом степени $m < 2n$, то верным является следующее равенство:

$$\int_{-1}^{1} P_m(x)dx = \sum_{i=1}^{n} c_i P_m(x_i). \tag{3.109}$$

Доказательство. Для начала рассмотрим случай $m < n$. Тогда полином $P_m(x)$ может быть переписан в виде многочлена Лагранжа с нулевым остаточным членом, так как n-ая производная от $P_m(x)$ будет равна нулю. Тогда имеем:

$$P_m(x) = \sum_{i=1}^{n} P_m(x_i) l_i(x) = \sum_{i=1}^{n} P_m(x_i) \prod_{j=1, i\neq j}^{n} \frac{x - x_j}{x_i - x_j}. \tag{3.110}$$

В таком случае интеграл от $P_m(x)$ имеет вид:

$$\begin{aligned} \int_{-1}^{1} P_m(x)dx &= \sum_{i=1}^{n} P_m(x_i) \left[\int_{-1}^{1} \prod_{j=1, i\neq j}^{n} \frac{x - x_j}{x_i - x_j} dx \right] \\ &= \sum_{i=1}^{n} c_i P_m(x_i). \end{aligned} \tag{3.111}$$

Теперь рассмотрим случай $n \leq m < 2n$. Разделим многочлен $P_m(x)$ на многочлен Лежандра $\phi_n(x)$. Тогда классическое деление многочленов столбиком дает:

$$\begin{aligned} & \frac{P_m(x)}{\phi_n(x)} = Q(x) + \frac{R(x)}{\phi_n(x)} \\ \implies \quad & P_m(x) = Q(x)\phi_n(x) + R(x), \end{aligned} \tag{3.112}$$

где $Q(x)$ – полином степени $m - n$ и $R(x)$ – полином степени обязательно меньшей, чем $\phi_n(x)$. Тогда, учитывая, что $m < 2n$ и соответственно $m - n < n$ мы имеем (см. равенство (3.83) и прилегающие параграфы):

$$\int_{-1}^{1} Q(x)\phi_n(x)dx = 0. \tag{3.113}$$

С другой стороны, так как степень многочлена $R(x)$ меньше n, интеграл от него попадает под первый случай этой теоремы, из чего следует:

$$\int_{-1}^{1} R(x)dx = \sum_{i=1}^{n} c_i R(x_i). \tag{3.114}$$

В таком случае интеграл от $P_m(x)$ принимает форму:

$$\int_{-1}^{1} P_m(x)dx = \sum_{i=1}^{n} c_i R(x_i). \tag{3.115}$$

Однако вследствие того, что $\phi_n(x_i) = 0$, мы получаем $P_m(x_i) = R(x_i)$. Тогда финальная форма интеграла имеет вид:

$$\int_{-1}^{1} P_m(x)dx = \sum_{i=1}^{n} c_i P_m(x_i). \tag{3.116}$$

□

Интегрирование на произвольном интервале $[a; b]$ с помощью квадратуры Гаусса реализуется с помощью замены переменных:

$$x = \frac{1}{2}\left[(b-a)t + a + b\right], \tag{3.117}$$

где $x \in [a; b]$ и $t \in [-1; 1]$. Тогда квадратура Гаусса вычисляется следующим образом:

$$\int_{a}^{b} f(x)dx = \int_{-1}^{1} f\left(\frac{(b-a)t + a + b}{2}\right)\frac{b-a}{2}dt. \tag{3.118}$$

Наилучшее приближение

До текущего момента мы апроксимировали неизвестные функции или дискретные данные с помощью интерполяции, т.е. требовали, чтобы аппроксимирующая функция проходила через заранее заданные узлы. В случае с интерполяцией полиномами, степень полинома при этом всегда жестко зависит от количества используемых узлов. Увеличение степени, как мы выяснили, часто негативно влияет на качество интерполяции. Более того, зачастую дискретные данные формируются из численного или реального эксперимента и являются зашумленными. В таком случае логичнее было бы найти полином малой степени, который бы наилучшим образом приближался к дискретным данным. В математической статистике эта операция называется *регрессией*, но мы будем использовать его без статистического контекста, подразумевая исключительно приближение дискретных данных некоторыми кривыми. Очевидно, что подобная идея легко распространяется, например, на класс тригонометрических полиномов.

Для примера рассмотрим набор дискретных данных $D = \{(x_i, y_i)\}_{i=1}^n$. В случае линейной регрессии нам необходимо найти такую функцию $f(x) = a_0 + a_1 x$, что она будет приближаться к данным D наилучшим образом. Самый очевидный подход состоит в решении следующей оптимизационной задачи:

$$\min_{a_0,a_1} E_\infty(a_0, a_1) = \min_{a_0,a_1} \max_{1\le i\le n} |y_i - f(x_i)|, \tag{4.1}$$

известной под названием *минимакс* (минимизация максимальной ошибки), и результатом которой является наилучший аппроксимационный полином. Несмотря на свою простоту, эту задачу тяжело решить даже для линейной функции $f(x)$, что делает ее применение нерациональным **[TODO: Демьянов и Малоземов]**. Другой подход состоит в минимизации суммы абсолютных отклонений:

$$\min_{a_0,a_1} E_1(a_0, a_1) = \min_{a_0,a_1} \sum_{i=1}^{n} |y_i - f(x_i)|. \tag{4.2}$$

Минимизация функции $E_1(a_0, a_1)$ сводится к взятию производных от нее по a_0 и a_1 и приравнивание их к нулю. Однако функция $E_1(a_0, a_1)$ не дифференцируема в нуле, что означает, что мы не всегда сможем найти решение.

Проблемы, сформулированные выше, обходятся с помощью использования суммы квадратов отклонений, что приводит нас к методу наименьших квадратов.

4.1 Метод наименьших квадратов

4.1.1 Линейная регрессия в одномерном пространстве

Проиллюстрируем метод наименьших квадратов на примере, изложенном выше. Для этого рассмотрим следующую задачу минимизации:

$$\min_{a_0,a_1} E_2(a_0,a_1) = \min_{a_0,a_1} \sum_{i=1}^{n} \left[y_i - f(x_i)\right]^2. \tag{4.3}$$

Выражение для $E_2(a_0,a_1)$ представляет собой суммы квадратом отклонений. Название «метод наименьших квадратов», очевидно, мотивировано именно этой задачей минимизации. Подстановка функции $f(x) = a_0 + a_1 x$ в $E_2(a_0,a_1)$ дает:

$$E_2(a_0,a_1) = \sum_{i=1}^{n} \left[y_i - a_0 - a_1 x_i\right]^2. \tag{4.4}$$

Функция $E_2(a_0,a_1)$ принимает экстремальное значение при таких a_0, a_1, что:

$$\frac{\partial E_2}{\partial a_0} = 0 \quad \Longrightarrow \quad -2\sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right) = 0, \tag{4.5}$$

$$\frac{\partial E_2}{\partial a_1} = 0 \quad \Longrightarrow \quad -2\sum_{i=1}^{n} x_i \left(y_i - a_0 - a_1 x_i\right) = 0, \tag{4.6}$$

что дает систему уравнений относительно a_0 и a_1:

$$\begin{cases} \sum\limits_{i=1}^{n} y_i - n a_0 - a_1 \sum\limits_{i=1}^{n} x_i = 0, \\ \sum\limits_{i=1}^{n} (x_i y_i) - a_0 \sum\limits_{i=1}^{n} x_i - a_1 \sum\limits_{i=1}^{n} x_i^2 = 0. \end{cases} \tag{4.7}$$

Решением системы являются следующие выражения для a_0 и a_1:

$$a_0^{(opt)} = \frac{\sum\limits_{i=1}^{n} x_i^2 \sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} (x_i y_i)}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}, \tag{4.8}$$

$$a_1^{(opt)} = \frac{n \sum\limits_{i=1}^{n} (x_i y_i) - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}. \tag{4.9}$$

Тогда прямая $f(x) = a_0^{(opt)} + a_1^{(opt)} x$ наилучшим образом приближается к дискретным данным $\{(x_i, y_i)\}_{i=1}^{n}$ в среднеквадратичном смысле. Однако это не означает, что такое приближение является наилучшим в принципе. Действительно, если мы сравним решение, полученное методом наименьших квадратов и решение, полученное минимаксом, то обнаружим небольшую разницу между ними.

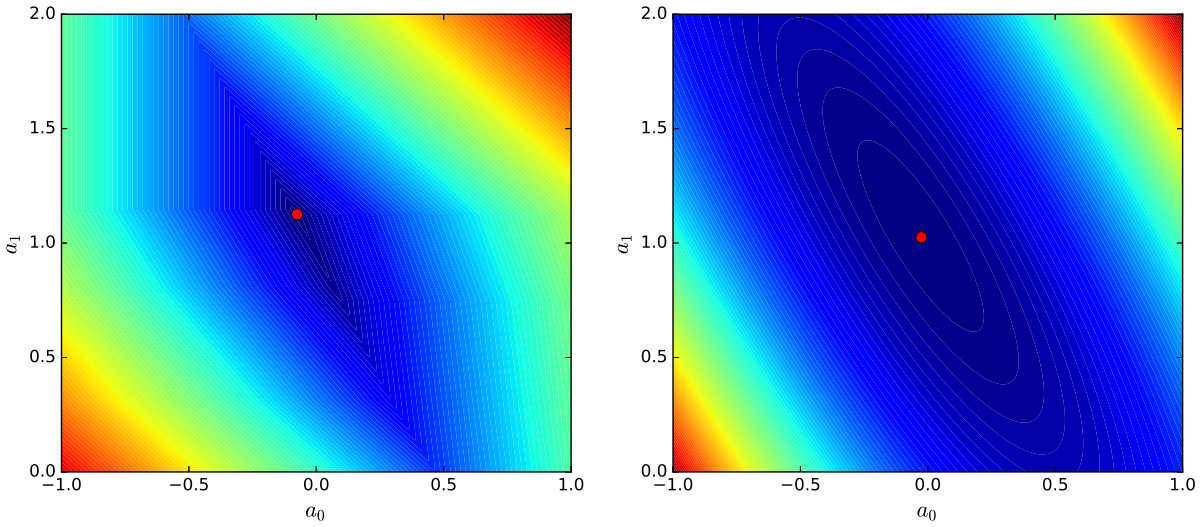

Рисунок 4.1 – Контуры поверхностей $E_\infty(a_0, a_1)$ и $E_2(a_0, a_1)$. Чем цвет контура ближе к красному, тем больше значение ошибки. Чем цвет контура ближе к синему, тем меньше значение ошибки. Красные точки обозначают минимальные значения соответствующих ошибок.

Продемонстрируем это на примере. Рисунок 4.1 показывает контуры поверхностей $E_\infty(a_0, a_1)$ и $E_2(a_0, a_1)$, сформированных при решении задачи приближения функции $f(x) = a_0 + a_1 x$ к данным, сгенерированным функцией $\widetilde{f}(x, \omega_g) = x + \omega_g$, где $\omega_g \sim \mathcal{N}(0, \sigma^2)$ – гауссовский белый шум. Легко заметить, что оптимальные значения соответствуют несовпадающим a_0, a_1. Это несовпадение легко увидеть, если изобразить на графике оптимальные прямые $f(x)$ для обоих случаев, как сделано на рисунке 4.2.

4.1.2 Линейная регрессия в общем случае

В многомерном случае линейная регрессия производится с помощью гиперплоскостей. Для перехода к многомерному случаю нам необходимо воспользоваться векторной формой записи задачи оптимизации, что позволит избежать лишних алгебраических выкладок. Итак, мы имеем:

$$\min_{\boldsymbol{a}} E_2(\boldsymbol{a}) = \min_{\boldsymbol{a}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}), \tag{4.10}$$

где $\boldsymbol{X} \in \mathbb{R}^{m \times (n+1)}$, $\boldsymbol{a} \in \mathbb{R}^{n+1}$ и $\boldsymbol{y} \in \mathbb{R}^m$. Здесь m обозначает число дискретных данных, а n – размерность пространства данных. Строка матрицы $\boldsymbol{X}$ хранит в себе одну многомерную точку данных, при этом первый столбец является столбцом единиц, что и дает размерность матрицы $m \times (n + 1)$. В общем виде матрица $\boldsymbol{X}$ может быть представлена следующим

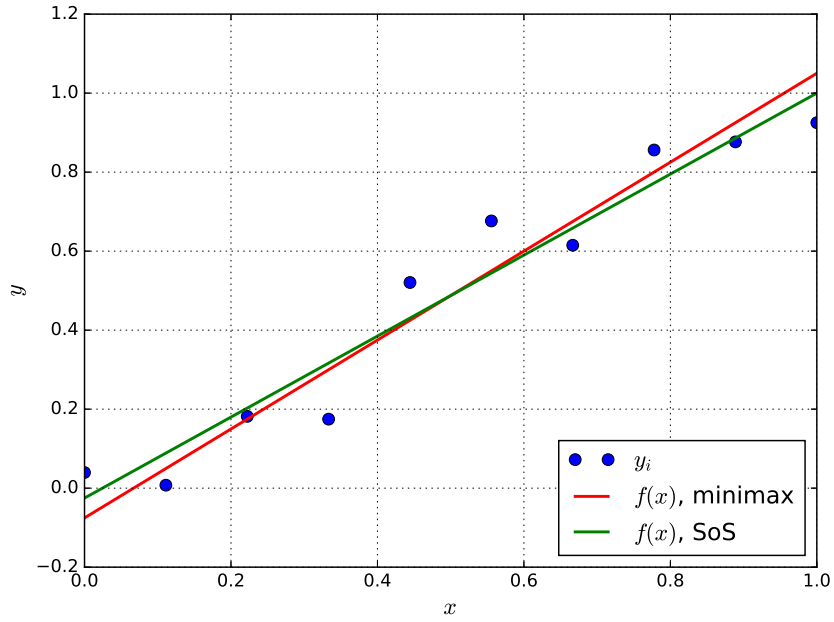

Рисунок 4.2 – Оптимальные прямые $f(x) = a_0^{(opt)} + a_1^{(opt)} x$, найденные при решении минимакса (красная прямая) и метода наименьших квадратов (зеленая прямая). Дискретные данные, приближение к которым осуществлялось, обозначены синими точками.

образом:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_n^{(1)} \\ 1 & x_1^{(2)} & \dots & x_n^{(2)} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_1^{(m)} & \dots & x_n^{(m)} \end{bmatrix}, \tag{4.11}$$

где нижние индексы обозначают координаты многомерной точки данных, а верхние индексы номера точек. Продифференцируем функцию $E_2(\boldsymbol{a})$ относительно $\boldsymbol{a}$:

$$\begin{aligned} & E_2(\boldsymbol{a}) = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{a} + \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a} = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a} \\ \implies & \frac{\partial E_2}{\partial \boldsymbol{a}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a} \end{aligned} \tag{4.12}$$

В справедливости подобного «векторного» дифференцирования можно убедиться, перейдя от матричной формы к суммированию:

$$E_2(\boldsymbol{a}) = \sum_{j=1}^{m} y_j^2 - 2\sum_{j=1}^{m}\sum_{i=0}^{n} a_i X_{ji} y_j + \sum_{j=1}^{m}\left(\sum_{i=0}^{n} a_i X_{ji}\right)^2. \tag{4.13}$$

Тогда дифференцирование относительно произвольного a_k дает:

$$\frac{\partial E_2}{\partial a_k} = -2\sum_{j=1}^{m} X_{jk} y_j + 2\sum_{j=1}^{m} X_{jk}\left(\sum_{i=0}^{n} a_i X_{ji}\right), \tag{4.14}$$

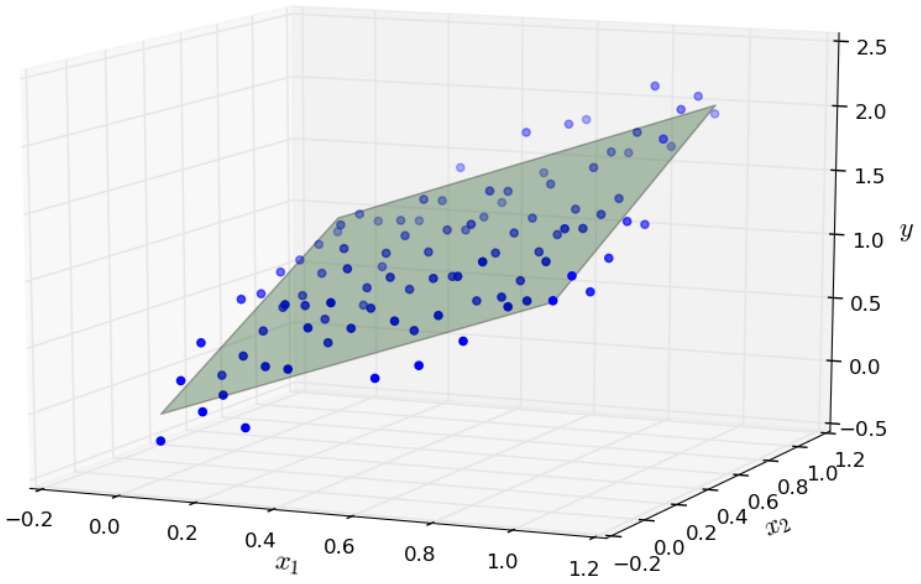

Рисунок 4.3 – Поверхность $f(x) = a_0^{(opt)} + a_1^{(opt)} x_1 + a_2^{(opt)} x_2$ (зеленая поверхность), приближенная к дискретным данным $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ (синие точки) с помощью метода наименьших квадратов.

что при записи в матричном виде дает (4.12). Оптимальное значение вектора $\boldsymbol{a}$ находится с помощью приравнивания производной (4.12) к нулю:

$$\begin{aligned} & -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a} = 0 \\ \Longrightarrow \quad & \boldsymbol{a} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \end{aligned} \tag{4.15}$$

Последнее уравнение известно как *нормальное уравнение* и широко используется как для нахождения аналитических решений задач, сформулированных с помощью метода наименьших квадратов, так и численных решений (в том случае, когда размерность матрицы $\boldsymbol{X}$ сравнительно мала).

Рисунок 4.3 показывает пример приближения двумерной плоскости к дискретным данным, сгенерированным функцией $\widetilde{f}(x, \omega_g) = x_1 + x_2 + \omega_g$, где $\omega_g \sim \mathcal{N}(0, \sigma^2)$ – гауссовский белый шум.

4.1.3 Нелинейная регрессия

Метод наименьших квадратов и нормальное уравнение могут быть с тем же успехом использованы для аппроксимации нелинейной функцией многих переменных. Рассмотрим

аппроксимирующую функцию $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, заданную следующим образом:

$$f(\boldsymbol{a}; \boldsymbol{x}) = \sum_{i=0}^{n} a_i \phi_i(\boldsymbol{x}), \tag{4.16}$$

где $\boldsymbol{x} \in \mathbb{R}^l$, $\phi_0(\boldsymbol{x}) = 1$ и $\phi_i(\boldsymbol{x}), i = 1, \ldots, n$ в общем случае нелинейные функции. Несложно заметить, что в таком случае задача минимизации (4.3) может быть записана в матричном виде (4.10), где матрица $\boldsymbol{X} \in \mathbb{R}^{m \times (n+1)}$ определена как:

$$\boldsymbol{X} = \begin{bmatrix} 1 & \phi_1(\boldsymbol{x}^{(1)}) & \ldots & \phi_n(\boldsymbol{x}^{(1)}) \\ 1 & \phi_1(\boldsymbol{x}^{(2)}) & \ldots & \phi_n(\boldsymbol{x}^{(2)}) \\ \vdots & \vdots & \ldots & \vdots \\ 1 & \phi_1(\boldsymbol{x}^{(m)}) & \ldots & \phi_n(\boldsymbol{x}^{(m)}) \end{bmatrix}, \tag{4.17}$$

где $\boldsymbol{x}^{(i)}, i = 1, \ldots, m$ – дискретные многомерные данные, к которым приближается функция $f(\boldsymbol{a}; \boldsymbol{x})$. Тогда оптимальные значения параметров $\boldsymbol{a}$ определяются с помощью нормального уравнения:

$$\boldsymbol{a} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{4.18}$$

Подобный метод может также интерпретироваться как линейная многомерная регрессия в пространстве с нелинейно трансформированными координатами.

Заметим, что повторное использование нормального уравнения возможно благодаря линейности $f(\boldsymbol{a}; \boldsymbol{x})$ относительно вектора параметров $\boldsymbol{a}$, несмотря на то, что функция нелинейна относительно $\boldsymbol{x}$. В случае, когда $f(\boldsymbol{a}; \boldsymbol{x})$ нелинейна относительно $\boldsymbol{a}$, требуется пересчет соответствующих производных, что в общем случае будет приводить к более сложным и, вероятно, нелинейным уравнениям для определения оптимальных $\boldsymbol{a}$.

В качестве примера рассмотрим полиномиальную регрессию в одномерном пространстве, где $f(\boldsymbol{a}; x) = \sum_{i=0}^{n} a_i x^i$. Тогда матрица $\boldsymbol{X}$ принимает вид:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{(1)} & \ldots & x_{(1)}^n \\ 1 & x_{(2)} & \ldots & x_{(2)}^n \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{(m)} & \ldots & x_{(m)}^n \end{bmatrix}, \tag{4.19}$$

Так как $\boldsymbol{X}$ является матрицей Вандермонда, ее определитель всегда отличен от нуля, если среди дискретных данных нет повторяющихся. Это доказывает, что при удовлетворении этого условия нормальное уравнение (4.18) всегда будет иметь единственное и нетривиальное решение.

Рисунок 4.4 демонстрирует пример приближения кубического полинома к дискретным данным, сгенерированным функцией $\widetilde{f}(x, \omega_g) = x^3 + x^2 + x + \omega_g$, где $\omega_g \sim \mathcal{N}(0, \sigma^2)$ – гауссовский белый шум.

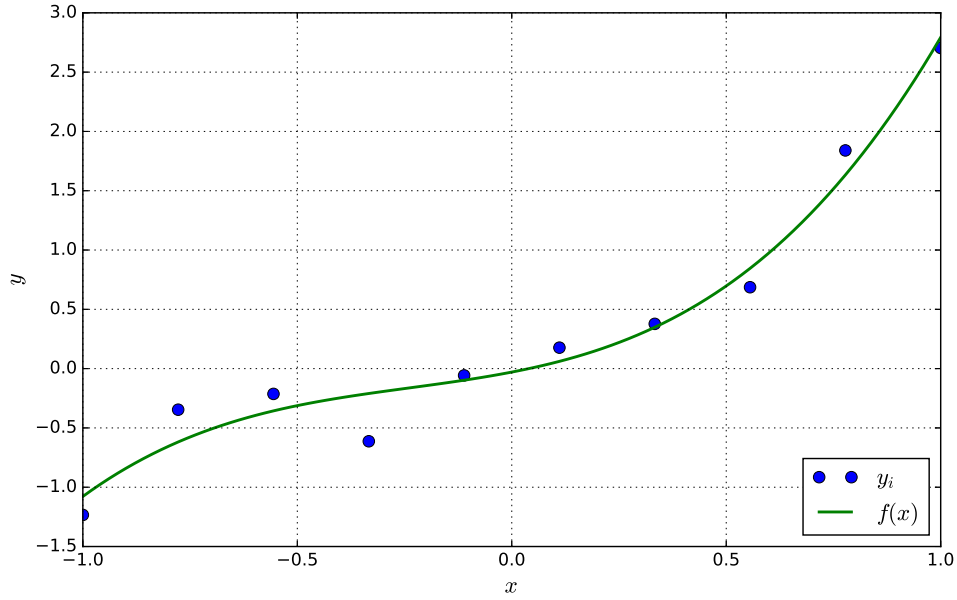

Рисунок 4.4 – Полином $f(x) = a_0^{(opt)} + a_1^{(opt)}x + a_2^{(opt)}x^2 + a_3^{(opt)}x^3$ (зеленая кривая), приближенный к дискретным данным $\{(x_i, y_i)\}_{i=1}^n$ (синие точки) с помощью метода наименьших квадратов.

4.1.4 Метод наименьших квадратов для приближения к непрерывной функции

До текущего момента мы рассматривали метод наименьших квадратов для для нахождения оптимального приближения к дискретных данных. Однако тот же метод можно применить и для приближения к некоторой непрерывной функции $y(x)$ на отрезке $[a; b]$. Для его демонстрации мы рассмотрим одномерный случай. Пусть аппроксимирующая функция задана так же, как и в случае нелинейной регрессии:

$$f(\boldsymbol{a}; x) = \sum_{i=0}^{n} a_i \phi_i(x). \tag{4.20}$$

Метод наименьших квадратов, сформулированный для дискретного случая в (4.3), в случае приближения к непрерывной функции трансформируется в интеграл от квадрата отклонения:

$$\begin{aligned} \min_{\boldsymbol{a}} E_2(\boldsymbol{a}) &= \min_{\boldsymbol{a}} \int_a^b [y(x) - f(x)]^2 \, dx \\ &= \min_{\boldsymbol{a}} \int_a^b \left[y(x) - \sum_{i=0}^{n} a_i \phi_i(x) \right]^2 dx. \end{aligned} \tag{4.21}$$

Как мы скоро увидим, в качестве системы функций $\{\phi_i(x)\}_{i=0}^n$ удобно выбрать ортогональную систему. Так как многие системы функций являются ортогональными только с весом,

активно используется модифицированный метод, называемый *взвешенным методом наименьших квадратов*:

$$\min_{\boldsymbol{a}} E_2(\boldsymbol{a}) = \min_{\boldsymbol{a}} \int_a^b \omega(x) \left[y(x) - \sum_{i=0}^{n} a_i \phi_i(x) \right]^2 dx. \tag{4.22}$$

Предположим, что $\{\phi_i(x)\}_{i=0}^n$ является ортогональной системой на $[a;b]$ с весом $\omega(x)$. Для нахождения наименьшего значения функции $E_2(\boldsymbol{a})$ необходимо найти нули производной:

$$\begin{aligned} &\frac{\partial E_2}{\partial a_k} = -2 \int_a^b \omega(x)\phi_k(x) \left[y(x) - \sum_{i=0}^{n} a_i \phi_i(x) \right] dx \\ \Longrightarrow \quad &\int_a^b \omega(x)\phi_k(x) \left[y(x) - \sum_{i=0}^{n} a_i \phi_i(x) \right] dx = 0. \end{aligned} \tag{4.23}$$

Пользуясь свойством оргональности функций, мы получаем:

$$\int_a^b \left[\omega(x) y(x) \phi_k(x) - a_k \omega(x) \phi_k^2(x) \right] dx = 0. \tag{4.24}$$

Тогда оптимальные значения a_k находятся с помощью выражения:

$$a_k = \frac{\langle y(x), \phi_k(x) \rangle_\omega}{\langle \phi_k(x), \phi_k(x) \rangle_\omega}. \tag{4.25}$$

4.2 Приближение тригонометрическими полиномами

Рассмотрим приближение к непрерывной функции $y : \mathbb{R} \longrightarrow \mathbb{R}$ с помощью тригонометрического ряда:

$$\begin{aligned} f(x) = S_n(x) &= a_0 + \sum_{k=1}^{n} \left(a_k \cos kx + b_k \sin kx \right) \\ &= \sum_{k=0}^{n} \left(a_k \cos kx + b_k \sin kx \right), \end{aligned} \tag{4.26}$$

где $a_k, b_k \in \mathbb{R}$ для $k = 0, \ldots, n$ и b_0 может быть произвольным, так как $\sin 0 = 0$. Оптимизационная задача метода наименьших квадратов в таком случае имеет вид:

$$\min_{\boldsymbol{a}} E_2(\boldsymbol{a}) = \min_{\boldsymbol{a}} \int_{-\pi}^{\pi} \omega(x) \left[y(x) - \sum_{k=0}^{n} \left(a_k \cos kx + b_k \sin kx \right) \right]^2 dx. \tag{4.27}$$

Как мы уже доказали в примере 2.7.1, система тригонометрических функций $\{\cos kx, \sin kx\}_{i=0}^n$ является ортогональной на $[-\pi;\pi]$ с весом $\omega(x) = 1$. Тогда пользуясь выражением (4.25)

для оптимальных коэффициентов метода наименьших квадратов, можно легко вывести следующие выражения для оптимальных a_k и b_k:

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} y(x)dx, \tag{4.28}$$

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} y(x)\cos kx dx, \tag{4.29}$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} y(x)\sin kx dx. \tag{4.30}$$

Тригонометрические ряды вида (4.26) часто называют *тригонометрическими полиномами*. Это название мотивировано экспоненциальной формой тригонометрического ряда. Действительно, рассмотрим формулу Эйлера:

$$e^{ikx} = \cos kx + i\sin kx. \tag{4.31}$$

Тогда домножение на $\hat{a}_k \in \mathbb{C}$ дает:

$$\begin{aligned}\hat{a}_k e^{ikx} &= \hat{a}_k \cos kx + i\hat{a}_k \sin kx \\ &= \Re(\hat{a}_k)\cos kx + i\Re(\hat{a}_k)\sin kx + i\Im(\hat{a}_k)\cos kx - \Im(\hat{a}_k)\sin kx,\end{aligned} \tag{4.32}$$

где $\Re(\hat{a}_k)$ и $\Im(\hat{a}_k)$ – вещественная и мнимая части $\hat{a}_k$ соответственно. Добавим к обоим частям равенства комплексно сопряженное:

$$\hat{a}_k e^{ikx} + \hat{a}_k^* e^{-ikx} = 2\Re(\hat{a}_k)\cos kx - 2\Im(\hat{a}_k)\sin kx, \tag{4.33}$$

где комлексно сопряженное число обозначено через $*$. Тогда суммирование обоих частей равенства по $k = 0, \ldots, n$ приводит к выражению:

$$\sum_{k=0}^{n}\left(\hat{a}_k e^{ikx} + \hat{a}_k^* e^{-ikx}\right) = \sum_{k=-n}^{n}\hat{a}_k e^{ikx} = \sum_{k=0}^{n}\left[2\Re(\hat{a}_k)\cos kx - 2\Im(\hat{a}_k)\sin kx\right], \tag{4.34}$$

где $\hat{a}_{-k} = \hat{a}_k^*$. Таким образом мы получили экспоненциальную форму тригонометрического ряда (4.26), где коэффициенты связаны друг с другом следующим образом:

$$a_k = 2\Re(\hat{a}_k), \tag{4.35}$$

$$b_k = -2\Im(\hat{a}_k), \tag{4.36}$$

где $k = 0, \ldots, n$. Для того, чтобы перейти к форме полинома, обозначим $z = e^{ix}$. Тогда имеем:

$$\sum_{k=-n}^{n}\hat{a}_k e^{ikx} = \sum_{k=-n}^{n}\hat{a}_k z^k, \tag{4.37}$$

что при домножении на z^n дает комплекснозначный полином степени $2n$, по конвенции называемый тригонометрическим полиномом n-й степени:

$$Q_n(z) = z^n\sum_{k=-n}^{n}\hat{a}_k z^k. \tag{4.38}$$

4.2.1 Дискретное приближение тригонометрическими полиномами

Тригонометрические полиномы могут быть использованы для приближения к дискретным данным. Пусть дано $2m$ равномерно распределенных узлов $x_j = -\pi + \frac{j}{m}\pi$ и соответствующие дискретные значения $\{y_j\}_{j=0}^{2m-1}$. Тогда метод наименьших квадратов формулируется следующим образом:

$$\min_{\boldsymbol{a}} \sum_{j=0}^{2m-1} \left[y_j - \sum_{k=0}^{n} (a_k \cos kx_j + b_k \sin kx_j) \right]^2. \tag{4.39}$$

Чтобы найти решение этой оптимизационной задачи, нам необходимо доказать ортогональность тригонометрической системы функций относительно операции суммирования $\sum_{j=0}^{2m-1}$. Для этого докажем следующую лемму.

Лемма 4.2.1. *Пусть* $r \in \mathbb{Z}$ *и* $r \bmod 2m \neq 0$. *Тогда*

$$\sum_{j=0}^{2m-1} \cos rx_j = 0, \tag{4.40}$$

$$\sum_{j=0}^{2m-1} \sin rx_j = 0. \tag{4.41}$$

Более того, если $r \bmod m \neq 0$, *то*

$$\sum_{j=0}^{2m-1} \cos^2 rx_j = m, \tag{4.42}$$

$$\sum_{j=0}^{2m-1} \sin^2 rx_j = m. \tag{4.43}$$

Доказательство. Для доказательства первых двух равенств достаточно доказать, что в ноль обращается следующее выражение:

$$\sum_{j=0}^{2m-1} e^{irx_j} = \sum_{j=0}^{2m-1} \cos rx_j + i \sum_{j=0}^{2m-1} \sin rx_j. \tag{4.44}$$

Тогда имеем:

$$\begin{aligned} \sum_{j=0}^{2m-1} e^{irx_j} &= \sum_{j=0}^{2m-1} e^{ir\left(-\pi + \frac{j}{m}\pi\right)} \\ &= e^{-ir\pi} \sum_{j=0}^{2m-1} e^{ir\frac{j}{m}\pi}. \end{aligned} \tag{4.45}$$

Можно заметить, что данная сумма представляет собой сумму членов геометрической прогрессии. Тогда получаем:

$$\begin{aligned}\sum_{j=0}^{2m-1} e^{ir\frac{j}{m}\pi} &= \frac{1-e^{i\frac{r}{m}(2m)\pi}}{1-e^{i\frac{r}{m}\pi}} \\ &= \frac{1-e^{2ir\pi}}{1-e^{i\frac{r}{m}\pi}}\end{aligned} \tag{4.46}$$

Легко проверить, что $e^{2ir\pi} = 1$ и числитель, следователь, обращается в ноль. Однако для $r \bmod 2m = 0$ в ноль обращается и знаменатель, что приводит к неопределенности $0/0$. Таким образом мы получаем

$$\begin{aligned}&\sum_{j=0}^{2m-1} e^{irx_j} = 0 \\ \Longrightarrow \quad &\sum_{j=0}^{2m-1} \cos rx_j = 0, \\ \Longrightarrow \quad &\sum_{j=0}^{2m-1} \sin rx_j = 0,\end{aligned} \tag{4.47}$$

для $r \bmod 2m \neq 0$.

Теперь докажем справедливость второй пары неравенств:

$$\begin{aligned}\sum_{j=0}^{2m-1} \cos^2 rx_j &= \frac{1}{2}\sum_{j=0}^{2m-1}(1+\cos 2rx_j) \\ &= m + \frac{1}{2}\sum_{j=0}^{2m-1}\cos 2rx_j \\ &= m,\end{aligned} \tag{4.48}$$

где $r \bmod m \neq 0$. Аналогично доказывается и второе равенство:

$$\begin{aligned}\sum_{j=0}^{2m-1} \sin^2 rx_j &= \frac{1}{2}\sum_{j=0}^{2m-1}(1-\cos 2rx_j) \\ &= m,\end{aligned} \tag{4.49}$$

где $r \bmod m \neq 0$. □

Рассмотренная лемма позволяет легко доказать теорему об ортогональности.

Теорема 4.2.1. *Пусть* $\phi_k(x_j) = \cos kx_j$ *и* $\psi_l(x_j) = \sin lx_j$, *где* $k = 0, \ldots, n$ *и* $l = 1, \ldots, n$. *Тогда верны следующие равенства:*

$$\sum_{j=0}^{2m-1} \phi_k(x_j)\psi_l(x_j) = 0, \quad \textit{при } \forall k, l, \tag{4.50}$$

$$\sum_{j=0}^{2m-1} \phi_k(x_j)\phi_l(x_j) = \sum_{j=0}^{2m-1} \psi_k(x_j)\psi_l(x_j) = 0, \quad \textit{при } k \neq l, \tag{4.51}$$

$$\sum_{j=0}^{2m-1} \phi_k^2(x_j) = \sum_{j=0}^{2m-1} \psi_k^2(x_j) = m, \quad \textit{при любом } k \neq 0, \tag{4.52}$$

$$\sum_{j=0}^{2m-1} \phi_k^2(x_j) = 2m, \quad \textit{при } k = 0. \tag{4.53}$$

Доказательство. Первые два равенства доказываются, используя тригонометрические тождества

$$\cos kx_j \sin lx_j = \frac{1}{2}\left[\sin(k+l)x_j + \sin(k-l)x_j\right] \tag{4.54}$$

$$\cos kx_j \cos lx_j = \frac{1}{2}\left[\cos(k+l)x_j + \cos(k-l)x_j\right] \tag{4.55}$$

$$\sin kx_j \sin lx_j = \frac{1}{2}\left[\cos(k-l)x_j - \sin(k+l)x_j\right] \tag{4.56}$$

$$\tag{4.57}$$

и лемму 4.2.1. Третье равенство напрямую следует из леммы 4.2.1. Четвертое равенство очевидно, так как:

$$\sum_{j=0}^{2m-1} \phi_0^2(x_j) = \sum_{j=0}^{2m-1} 1 = 2m. \tag{4.58}$$

□

Теперь мы можем найти оптимальные значения коэффициентов тригонометрического ряда в оптимизационной задаче (4.39). Производная суммы квадратов отклонений по $a_p, p = 0, \ldots, n$ имеет вид:

$$\frac{\partial E_2}{\partial a_p} = -2 \sum_{j=0}^{2m-1} \cos px_j \left[y_j - \sum_{k=0}^{n} (a_k \cos kx_j + b_k \sin kx_j) \right]. \tag{4.59}$$

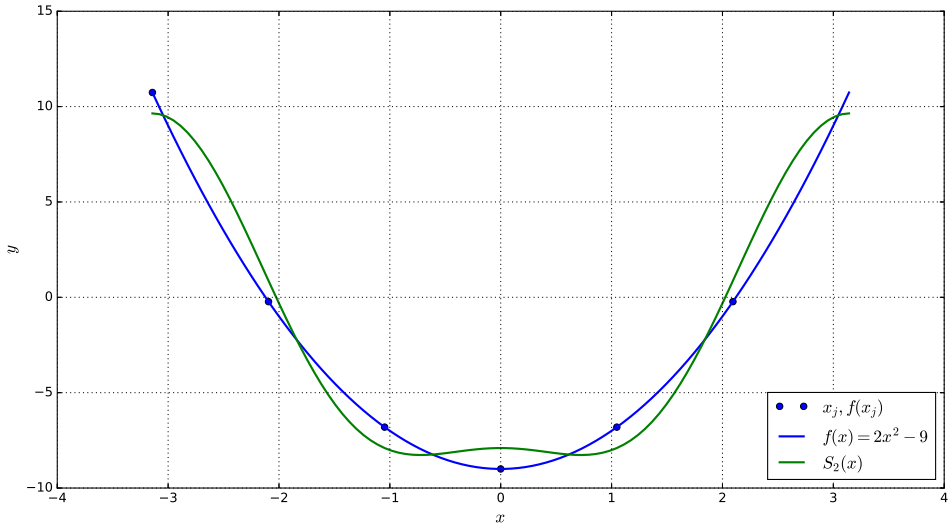

Рисунок 4.5 – Приближение тригонометрическим полиномом второй степени (зеленая линия) к дискретным данных (синие точки), сгенерированными функцией $f(x) = 2x^2 - 9$ (синяя линия) для шести узлов.

Приравнивая к нулю и используя теорему 4.2.1, получаем:

$$\sum_{j=0}^{2m-1} \cos px_j \left[y_j - \sum_{k=0}^{n} (a_k \cos kx_j + b_k \sin kx_j) \right] = 0$$

$$\Longrightarrow \quad a_0 = \frac{1}{2m} \sum_{j=0}^{2m-1} y_j \cos px_j, \tag{4.60}$$

$$\Longrightarrow \quad a_p = \frac{1}{m} \sum_{j=0}^{2m-1} y_j \cos px_j, \quad p = 1, \ldots, n.$$

Аналогично получаем оптимальные значения для b_k:

$$b_p = \frac{1}{m} \sum_{j=0}^{2m-1} y_j \sin px_j, \quad p = 1, \ldots, n. \tag{4.61}$$

Пример дискретного приближения тригонометрическим полиномом второй степени с помощью МНК можно наблюдать на рисунке **??**. В данном случае узлы $\{(x_j, y_j)\}_{j=0}^{5}$ были сгенерированы функцией $f(x) = 2x^2 - 9$, обладающей четностью: $x \longrightarrow -x$.

4.2.2 Дискретное преобразование Фурье

До текущего момента мы рассматривали наилучшее (в среднеквадратическом смысле) приближение тригометрического полинома к некоторым непрерывным или дискрет-

ным данным. Однако более широкое распространение получила *тригонометрическая интерполяция* по равномерно распределенным узлам, т.е. интерполяции дискретных данных тригонометрическими полиномами. Как мы уже знаем из (4.37), тригонометрический ряд представляет из себя полином степени n с $2n+1$ неизвестными вещественными коэффициентами. При $2m$ дискретных узлов $\{(x_j, y_j)\}_{j=0}^{2m-1}$, где $x_j = -\pi + \frac{j}{m}\pi$, мы всегда можем найти единственный интерполяционный тригонометрический полином степени m с некоторыми ограничениями, которые будут обсуждаться позже.

Действительно, рассмотрим следующий тригонометрический полином в экспоненциальной форме:

$$\sum_{k=-m}^{m-1} \hat{a}_k e^{ikx} \tag{4.62}$$

и потребуем, чтобы он согласовывался с узлами (x_j, y_j):

$$\begin{aligned} &\sum_{k=-m}^{m-1} \hat{a}_k e^{ikx_j} = y_j \\ \Longrightarrow \quad &\sum_{k=-m}^{m-1} \hat{a}_k e^{ikx_j} = y_j. \end{aligned} \tag{4.63}$$

Домножим обе стороны равенства на e^{-ilx_j} и просуммируем результат по индексуj:

$$\sum_{k=-m}^{m-1} \hat{a}_k \sum_{j=0}^{2m-1} e^{i(k-l)x_j} = \sum_{j=0}^{2m-1} y_j e^{-ilx_j}. \tag{4.64}$$

Благодаря лемме 4.2.1 мы знаем, что:

$$\sum_{j=0}^{2m-1} e^{i(k-l)x_j} = \begin{cases} 0, & k \neq l \\ 2m, & k = l \end{cases} \tag{4.65}$$

Тогда мы сразу получаем выражение для коэффициентов $\hat{a}_k$, составляющее *дискретное преобразование Фурье*:

$$\hat{a}_k = \frac{1}{2m} \sum_{j=0}^{2m-1} y_j e^{-ikx_j}. \tag{4.66}$$

Заметим, что в тригонометрическом интерполянте вы выбрали несимметричное суммирование относительно k:

$$\sum_{k=-m}^{m-1} \hat{a}_k e^{ikx}, \tag{4.67}$$

что приводит к несбалансированности комплексного члена $\hat{a}_{-m} e^{-imx}$. Можно заметить, что

для $\hat{a}_{-m}$ мы имеем:

$$\begin{aligned} \hat{a}_{-m} &= \frac{(-1)^{-m}}{2m} \sum_{j=0}^{2m-1} y_j e^{ij\pi} = \frac{(-1)^m}{2m} \sum_{j=0}^{2m-1} (-1)^j y_j \\ \implies & \hat{a}_{-m} \in \mathbb{R}, \\ \implies & \hat{a}_{-m} = \hat{a}_m. \end{aligned} \tag{4.68}$$

Так как $y_j \in \mathbb{R}$, тригонометрический полином должен отображать $\mathbb{R}$ в $\mathbb{R}$. Это означает, что в члене $\hat{a}_{-m}e^{-imx}$ требуется обнулить мнимую часть. В таком случае вследствие вещественности $\hat{a}_{-m}$ получаем окончательную форму тригонометрического интерполянта:

$$\sum_{k=-m}^{m-1} \hat{a}_k e^{ikx} = \sum_{k=-m+1}^{m-1} \hat{a}_k e^{ikx} + \hat{a}_{-m}\Re(e^{-imx}) = \sum_{k=-m+1}^{m-1} \hat{a}_k e^{ikx} + \hat{a}_m \cos mx.$$

4.2.3 Быстрое преобразование Фурье

Дискретное преобразование Фурье в представленной форме требует $2m$ комплексных умножений и $2m-1$ комплексных сложений для вычисления одного коэффициента a_k. Так как всего имеется $2m$ коэффициентов, алгоритмическая сложность дискретного преобразование Фурье имеет форму $O(m^2)$. Подобная сложность ограничивала масштабное использование дискретного преобразования Фурье вплоть до тех пор, пока не был открыт алгоритм *быстрого преобразования Фурье* (БПФ), имеющий сложность $O(m \log_2 m)$. Мы рассмотрим его в той форме, в которой он был представлен Кули и Тьюки.

Для начала запишем коэффициент $\hat{a}_k$ в следующей форме:

$$\hat{a}_k = \frac{1}{2m} \sum_{j=0}^{2m-1} y_j e^{ik\pi} e^{-i\frac{kj}{m}\pi} = \frac{(-1)^k}{2m} \sum_{j=0}^{2m-1} y_j e^{-\frac{ikj\pi}{m}}. \tag{4.69}$$

Алгоритм Кули–Тьюки вычисляет значение суммы, которую мы обозначим как:

$$A_k = \sum_{j=0}^{2m-1} y_j e^{-\frac{ikj\pi}{m}}, \quad k = 0, \dots 2m-1. \tag{4.70}$$

Разделим A_k на две части с четными и нечетными индексами соответственно:

$$\begin{aligned} A_k &= \sum_{j=0}^{m-1} y_{2j} e^{-\frac{2jik\pi}{m}} + \sum_{j=0}^{m-1} y_{2j+1} e^{-\frac{(2j+1)ik\pi}{m}} \\ &= \sum_{j=0}^{m-1} y_{2j} e^{-\frac{2jik\pi}{m}} + e^{-\frac{ik\pi}{m}} \sum_{j=0}^{m-1} y_{2j+1} e^{-\frac{2jik\pi}{m}}. \end{aligned} \tag{4.71}$$

Обозначим полученные суммы E_k и O_k:

$$E_k = \sum_{j=0}^{m-1} y_{2j} e^{-\frac{2jik\pi}{m}}, \tag{4.72}$$

$$O_k = \sum_{j=0}^{m-1} y_{2j+1} e^{-\frac{2jik\pi}{m}}. \tag{4.73}$$

Заметим, что E_k и O_k являются периодическими относительно k и периода m:

$$E_{k\pm m} = \sum_{j=0}^{m-1} y_{2j} e^{-\frac{2ji(k\pm m)\pi}{m}} = e^{\mp 2ji\pi} \sum_{j=0}^{m-1} y_{2j} e^{-\frac{2jik\pi}{m}} = E_k, \tag{4.74}$$

$$O_{k\pm m} = \sum_{j=0}^{m-1} y_{2j+1} e^{-\frac{2ji(k\pm m)\pi}{m}} = e^{\mp 2ji\pi} \sum_{j=0}^{m-1} y_{2j+1} e^{-\frac{2jik\pi}{m}} = O_k. \tag{4.75}$$

Это свойство позволяет посчитать E_k и O_k только для $k = 0, \ldots, m$, так как E_k и O_k для $k = m, \ldots, 2m-1$ будут иметь те же значения. Тогда A_k для $k = 0, \ldots, 2m$ высчитывается следующим образом:

$$A_k = E_k + e^{-\frac{ik\pi}{m}} O_k, \quad k = 0, \ldots, m \tag{4.76}$$

$$A_{k+m} = E_k + e^{-\frac{i(k+m)\pi}{m}} O_k = E_k - e^{-\frac{ik\pi}{m}} O_k, \quad k = 0, \ldots, m. \tag{4.77}$$

Так как E_k и O_k так же могут быть восприняты как БПФ на более грубых сетках, этот алгоритм рекурсивно применяется и к ним, что и замыкает в конечном итоге алгоритм Кули–Тьюки. Ясно, что в приведенном изложении алгоритма предполагается, что $m = 2^{\tilde{m}}, \tilde{m} \in \mathbb{N}$.

Рассмотрим в качестве примера развернутый рекурсивный алгоритм для $2m = 8$. Тогда имеем:

$$\begin{aligned} A_k &= \sum_{j=0}^{7} y_j e^{-\frac{ijk\pi}{4}} \\ &= \sum_{j=0}^{3} y_{2j} e^{-\frac{ijk\pi}{2}} + e^{-\frac{ik\pi}{4}} \sum_{j=0}^{3} y_{2j} e^{-\frac{ijk\pi}{2}} \\ &= E_k^{(1)} + e^{-\frac{ik\pi}{4}} O_k^{(1)}. \end{aligned} \tag{4.78}$$

На этом этапе нам необходимо рассчитать $E_k^{(1)}$ и $O_k^{(1)}$ только для $k = 0, \ldots, 4$. Для удобства введем индекс $k^{(1)} = 0, \ldots, 4$ и рекурсивно применим алгоритм разделения на четные и нечетные индексы для $E_k^{(1)}$ и $O_k^{(1)}$:

$$E_{k^{(1)}}^{(1)} = \sum_{j=0}^{3} y_{2j} e^{-\frac{ijk^{(1)}\pi}{2}} = \sum_{j=0}^{1} y_{4j} e^{-ijk^{(1)}\pi} + e^{-\frac{ik^{(1)}\pi}{2}} \sum_{j=0}^{1} y_{4j+2} e^{-ijk^{(1)}\pi} = E_{k^{(1)}}^{(E,2)} + e^{-\frac{ik^{(1)}\pi}{2}} O_{k^{(1)}}^{(E,2)},$$

$$O_{k^{(1)}}^{(1)} = \sum_{j=0}^{3} y_{2j+1} e^{-\frac{ijk^{(1)}\pi}{2}} = \sum_{j=0}^{1} y_{4j+1} e^{-ijk^{(1)}\pi} + e^{-\frac{ik^{(1)}\pi}{2}} \sum_{j=0}^{1} y_{4j+3} e^{-ijk^{(1)}\pi} = E_{k^{(1)}}^{(O,2)} + e^{-\frac{ik^{(1)}\pi}{2}} O_{k^{(1)}}^{(O,2)}.$$

Наконец, для полученных выражений $E^{(E,2)}_{k^{(1)}}, O^{(E,2)}_{k^{(1)}}, E^{(O,2)}_{k^{(1)}}$ и $O^{(O,2)}_{k^{(1)}}$ нам достаточно рассмотреть индексы $k^{(1)} = 0, 1, 2$. Тогда, введя новый индекс $k^{(2)} = 0, 1, 2$, получаем:

$$E^{(E,2)}_{k^{(2)}} = \sum_{j=0}^{1} y_{4j} e^{-ijk^{(2)}\pi} = y_0 + (-1)^{k^{(2)}} y_4, \tag{4.79}$$

$$O^{(E,2)}_{k^{(2)}} = \sum_{j=0}^{1} y_{4j+2} e^{-ijk^{(2)}\pi} = y_2 + (-1)^{k^{(2)}} y_6, \tag{4.80}$$

$$E^{(O,2)}_{k^{(2)}} = \sum_{j=0}^{1} y_{4j+1} e^{-ijk^{(2)}\pi} = y_1 + (-1)^{k^{(2)}} y_5, \tag{4.81}$$

$$O^{(O,2)}_{k^{(2)}} = \sum_{j=0}^{1} y_{4j+3} e^{-ijk^{(2)}\pi} = y_3 + (-1)^{k^{(2)}} y_7. \tag{4.82}$$

$$\tag{4.83}$$

Рассчитав все значения по рекурсии вверх, мы получаем искомый результат.

Численные методы линейной алгебры

Зачастую конечным результатом применения тех или иных сложных численных методов решения дифференциальных или иных уравнений является система линейных алгебраических уравнений (СЛАУ), матричная форма которой имеет вид:

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \tag{5.1}$$

или кратко:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{5.2}$$

где $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, $\boldsymbol{x} \in \mathbb{R}^n$ и $\boldsymbol{b} \in \mathbb{R}^n$

Так как в сложных численных методах решение СЛАУ после некоторых простых манипуляций дает полное численное решение задачи, очевидна необходимость нахождения оптимальных методов решения СЛАУ. Глобально большинство методов решения СЛАУ можно разделить на две группы:

- прямые;
- итерационные.

Прямые методы позволяют найти точное решение СЛАУ (5.2), в то время как итерационные методы рассчитывают такое $\boldsymbol{x}^{(k)}$, что $\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b} \to \boldsymbol{0}$ при $k \to \infty$. Прямые методы обычно используются для СЛАУ малой размерности, в то время как для СЛАУ большой размерности используют итерационные методы.

Разнообразие как прямых, так и итерационных методов связано с разнообразием особенностей матриц, составляющих СЛАУ. Так, например, если матрица является положительно определенной, то из класса прямых методов логично выбрать разложение Холецкого.

5.1 Прямые методы

5.1.1 Метод Гаусса (метод последовательного исключения)

Метод Гаусса является базовым методом решения СЛАУ и состоит в приведении матрицы $\boldsymbol{A}$ к треугольному виду с помощью элементарных преобразований, после чего решение

$\boldsymbol{x}$ находится сравнительно легко. Рассмотрим следующую расширенную матрицу:

$$\tilde{\boldsymbol{A}} = [\boldsymbol{A}, \boldsymbol{b}] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{array}\right]. \tag{5.3}$$

Для того, чтобы привести матрицу $\boldsymbol{A}$ к треугольному виду, необходимо последовательно обнулять элементы, находящиеся под главной диагональю. Так, если из строк $i = 2, 3, \dots, n$ отнять первую строку, домноженную на $\frac{a_{i1}}{a_{11}}$, то все элементы, расположенные под a_{11} будут равны нулю. Мы обозначим это элементарное преобразование следующим образом:

$$(i) \longrightarrow (i) - \frac{a_{i1}}{a_{11}}(1), \quad i = 2, \dots, n \tag{5.4}$$

где (i) обозначает i-ю строку. Тогда, применив операцию к расширенной матрицу, имеем:

$$\tilde{\boldsymbol{A}} = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right], \tag{5.5}$$

где $a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}$ и $b_i^{(1)} = b_i - \frac{a_{i1}}{a_{11}} b_1$. Аналогично, для обнуления элементов, находящихся под диагональным элементом $a_{22}^{(1)}$, нам необходимо применить следующее элементарное преобразование:

$$(i) \longrightarrow (i) - \frac{a_{i2}}{a_{22}^{(1)}}(2), \quad i = 3, \dots, n \tag{5.6}$$

Применяя подобные преобразования каскадно вплоть до последнего диагольного элемента, мы получаем верхнюю треугольную матрицу для $\boldsymbol{A}$:

$$\tilde{\boldsymbol{A}} = \left[\begin{array}{ccccc|c} a_{11} & a_{12} & \dots & \dots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & \dots & \dots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & a_{n-1,n-1}^{(n-2)} & a_{n-1,n}^{(n-2)} & b_n^{(n-2)} \\ 0 & 0 & \dots & 0 & a_{nn}^{(n-1)} & b_n^{(n-1)} \end{array}\right]. \tag{5.7}$$

Этот этап метода Гаусса называется *прямым ходом*. Очевидно, что тот же метод можно использовать для получения нижней треугольной матрицы. Можно заметить, что решение СЛАУ, в которой матрица имеет треугольную форму, легко находится с помощью так

называемого *обратного хода* метода Гаусса:

$$\begin{aligned} x_n &= \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}, \\ x_{n-1} &= \frac{b_{n-1}^{(n-2)} - a_{n-1,n}^{(n-2)} x_n}{a_{n-1,n-1}^{(n-2)}}, \\ &\dots \\ x_1 &= \frac{b_1 - \sum_{i=2}^{n} a_{1i} x_i}{a_{11}}. \end{aligned}$$

Рассчитаем количество операций, необходимых для нахождения решения СЛАУ методом Гаусса. Так как вычислительное время, потребное для операций умножения и деления, сильно больше, чем вычислительное время сложения и вычитания, мы сконцентрируемся на расчете числа умножений и делений. В первую очередь заметим, что на k-й итерации из всего $(n-1)$ итераций прямого хода требуется $(n-k)$ делений для вычисления множителя $m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$. Затем требуется для $(n-k)$ строк требуется произвести $(n-k+1)$ перемножений, т.е. суммарно $(n-k)(n-k+1)$. Таким образом, на k-й итерации мы имеем $n-k+(n-k)(n-k+1) = (n-k)(n-k+2)$ умножений и делений. Суммируя по $k = 1, \dots, n-1$, получаем общее число умножений и делений в прямом ходе метода Гаусса:

$$\begin{aligned} \sum_{k=1}^{n-1}(n-k)(n-k+2) &= \sum_{k=1}^{n-1}(n-k)^2 + 2\sum_{k=1}^{n-1}(n-k) \\ &= \sum_{k=1}^{n-1} k^2 + 2\sum_{k=1}^{n-1} k \\ &= \frac{(n-1)n(2n-1)}{6} + 2\frac{(n-1)n}{2} \\ &= \frac{2n^3+3n^2-5n}{6}. \end{aligned} \tag{5.8}$$

В обратном ходе метода Гаусса для нахождения x_k требуется $n-k$ перемножений и одно деление, что в сумме дает:

$$\begin{aligned} \sum_{k=1}^{n}(n-k+1) &= \sum_{k=1}^{n} k \\ &= \frac{n(n+1)}{2} \\ &= \frac{n^2+n}{2}. \end{aligned} \tag{5.9}$$

Наконец, суммарно требуется следующее число операций для нахождения решения методом Гаусса:

$$\frac{2n^3+3n^2-5n}{6} + \frac{n^2+n}{2} = O(n^3). \tag{5.10}$$

5.1.2 Метод Гаусса с выбором главного элемента

В случае, когда на k-й итерации алгоритма диагональный элемент оказывается равен нулю, мы не имеем возможности рассчитать множитель $m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$. Эта проблема решается соответствующим переставлением строк. Если элемент $a_{kk}^{(k)}$ мал по сравнению с $a_{ik}^{(k)}$, множитель m_{ik} будет иметь большое значение, что приводит к усилению погрешности округления в результате ряда последующих умножений на m_{ik}. Более того, деление на малый элемент $a_{kk}^{(k)}$ происходит также во время обратного года Гаусса и усиливает накопленные погрешности в числителе. Простейший способ избежать подобного поведения состоит в перестановке строк матрицы так, что диагональным элементом становится наибольший по модулю элемент k-го столбца:

$$|a_{kk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|. \tag{5.11}$$

Такой подход называется *частичным выбором главного элемента*. Логичным развитием является перестановка строк и столбцов так, что диагональным элементом становится наибольший по модулю элемент $|a_{ij}^{(k)}|$, где $k \le i \le n$ и $k \le j \le n$:

$$|a_{kk}^{(k)}| = \max_{k \le i,j \le n} |a_{ij}^{(k)}|. \tag{5.12}$$

Такой подход именуется *полным выбором главного элемента*. Легко убедиться, что полный выбор главного элемента требует суммарно $O(n^3)$ сравнений и таким образом рекомендуется к использованию только тогда, когда решаемая задача особенно чувствительна к погрешности округления.

5.1.3 LU-разложение

Предположим, что матрица $\boldsymbol{A}$ может быть разложена в нижнюю и верхнюю треугольные матрицы:

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}. \tag{5.13}$$

Такое разложение называется *LU-разложение*. В таком случае система линейных уравнений $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ решается в два шага. Обозначив вектор $\boldsymbol{y}$ как $\boldsymbol{y} = \boldsymbol{U}\boldsymbol{x}$, мы находим его как решение уравнения:

$$\boldsymbol{L}\boldsymbol{y} = \boldsymbol{b}. \tag{5.14}$$

Вектор $\boldsymbol{x}$ тогда является решением уравнения:

$$\boldsymbol{U}\boldsymbol{x} = \boldsymbol{y}. \tag{5.15}$$

Каждый из шагов требует $O(n^2)$ операций. Таким образом, один раз разложив матрицу $\boldsymbol{A}$ в нижнюю и верхнюю треугольные матрицы, мы можем находить решение $\boldsymbol{x}$ для различных $\boldsymbol{b}$ в $O(n^2)$ шагов. LU-разложение связано с прямым ходом метода Гаусса и соответственно требует $O(n^3)$ операций.

Получим явные выражения для матриц $\boldsymbol{L}$ и $\boldsymbol{U}$. Пусть перед k-й итерацией прямого хода метода Гаусса матрица $\boldsymbol{A}$ находится в форме $\boldsymbol{A}^{(k)}$. Тогда k-я итерация эквивалентна домножению матрицы $\boldsymbol{A}^{(k)}$ и вектора $\boldsymbol{b}^{(k)}$ слева на матрицу $\boldsymbol{M}^{(k)}$, имеющую вид:

$$\boldsymbol{M}^{(k)} = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & \ddots & \ddots & & & \vdots \\ \vdots & & \vdots & -m_{k+1,k} & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -m_{nk} & 0 & \dots & 0 & 1 \end{bmatrix}. \tag{5.16}$$

Верхняя треугольная матрица, получающаяся в результате обратного хода метода Гаусса, тогда выражается следующим образом:

$$\boldsymbol{A}^{(n)} = \boldsymbol{M}^{(n-1)} \cdots \boldsymbol{M}^{(1)} \boldsymbol{A}. \tag{5.17}$$

В контексте LU-разложения мы таким образом получаем матрицу $\boldsymbol{U} = \boldsymbol{A}^{(n)}$:

$$\boldsymbol{U} = \boldsymbol{M}^{(n-1)} \cdots \boldsymbol{M}^{(1)} \boldsymbol{A}. \tag{5.18}$$

Матрица $\boldsymbol{L}$ в таком случае может быть получена следующим образом:

$$\begin{aligned} & \boldsymbol{A} = \left(\boldsymbol{M}^{(n-1)} \cdots \boldsymbol{M}^{(1)}\right)^{-1} \boldsymbol{U} \\ \Longrightarrow \quad & \boldsymbol{A} = \left(\boldsymbol{M}^{(1)}\right)^{-1} \cdots \left(\boldsymbol{M}^{(n-1)}\right)^{-1} \boldsymbol{U} \\ \Longrightarrow \quad & \boldsymbol{L} = \left(\boldsymbol{M}^{(1)}\right)^{-1} \cdots \left(\boldsymbol{M}^{(n-1)}\right)^{-1}. \end{aligned} \tag{5.19}$$

Явные выражения для обратных матриц могут быть получены, если мы заметим, что операция, обратная $(i) \longrightarrow (i) - \frac{a_{ik}}{a^{(1)}_{kk}}(k), \quad i = k+1, \dots, n$, есть операция $(i) \longrightarrow (i) + \frac{a_{ik}}{a^{(1)}_{kk}}(k), \quad i = k+1, \dots, n$. Тогда обратная матрица $\left(\boldsymbol{M}^{(k)}\right)^{-1}$ имеет вид:

$$\left(\boldsymbol{M}^{(k)}\right)^{-1} = \boldsymbol{L}^{(k)} = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & \ddots & \ddots & & & \vdots \\ \vdots & & \vdots & m_{k+1,k} & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & m_{nk} & 0 & \dots & 0 & 1 \end{bmatrix}. \tag{5.20}$$

Легко убедиться, что перемножение таких матриц дает:

$$\boldsymbol{L} = \boldsymbol{L}^{(1)} \cdots \boldsymbol{L}^{(n-1)} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ m_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ m_{n1} & \dots & m_{n,n-1} & 1 \end{bmatrix}. \tag{5.21}$$

В данном выводе LU-разложения мы не предполагали перестановок строк, которые зачастую являются необходимыми. Подобные перестановки можно учесть, используя *матрицу перестановки*. Матрица перестановки $\boldsymbol{P}$ образуется с помощью перестановки строк единичной матрицы $\boldsymbol{E}$. Несложно проверить, что при умножении такой матрицы слева на матрицу $\boldsymbol{A}$ происходит та же перестановка строк, что использовалась при построении матрицы $\boldsymbol{P}$ из единичной матрицы. При умножении матрицы перестановки справа на матрицу $\boldsymbol{A}$ происходит перестановка столбцов. Пусть мы имеем матрицу необходимых перестановок $\boldsymbol{P}$. Тогда мы можем найти LU-разложение для матрицы:

$$\boldsymbol{PA} = \boldsymbol{LU}, \tag{5.22}$$

и найти решение матричного уравнения:

$$\boldsymbol{PAx} = \boldsymbol{Pb}. \tag{5.23}$$

5.1.4 Матрицы с диагональным преобладанием

Матрица $\boldsymbol{A}$ обладает свойством диагольного преобладания, если

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{ij}|, \quad i = 1, \dots, n. \tag{5.24}$$

Для матрицы со строгим диагональным преобладанием верны соответствующие строгие неравенства:

$$|a_{ii}| > \sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{ij}|, \quad i = 1, \dots, n. \tag{5.25}$$

Матрицы со строгим диагональным преобладанием примечательны тем, что они всегда являются невырожденными. Более того, метод Гаусса может быть использован для них без перестановок строк и столбцов и является вычислительно устойчивым.

Теорема 5.1.1. *Матрица $\boldsymbol{A}$ со строгим диагональным преобладанием является невырожденной.*

Доказательство. Рассмотрим доказательство от обратного. Пусть матрица $\boldsymbol{A}$ является вырожденной. Это означает, что уравнение

$$\boldsymbol{Ax} = \boldsymbol{0} \tag{5.26}$$

имеет нетривиальное решение. Пусть k – индекс, для которого верно

$$x_k = \max_{1 \le j \le n} |x_j|. \tag{5.27}$$

Тогда имеем:

$$\begin{aligned} &\sum_{1 \le j \le n} a_{kj} x_j = 0 \\ \implies \quad &a_{kk} x_k = - \sum_{\substack{1 \le j \le n \\ k \ne j}} a_{kj} x_j. \end{aligned} \tag{5.28}$$

В соответствии с неравенством треугольника имеем:

$$|a_{kk}||x_k| \le \sum_{\substack{1 \le j \le n \\ k \ne j}} |a_{kj}||x_j|, \tag{5.29}$$

что дает

$$\begin{aligned} |a_{kk}| &\le \sum_{\substack{1 \le j \le n \\ k \ne j}} |a_{kj}| \frac{|x_j|}{|x_k|}, \\ &\le \sum_{\substack{1 \le j \le n \\ k \ne j}} |a_{kj}|, \end{aligned} \tag{5.30}$$

что противоречит свойству строгого диагонального преобладания. □

5.1.5 Положительно определенные матрицы

Матрица $\boldsymbol{A}$ называется *положительно определенной*, если она симметричная, и верным является неравенство $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$ для любого вектора $\boldsymbol{x} \ne \boldsymbol{0}$ подходящей размерности. Легко проверить, что выражение $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$ имеет вид:

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j. \tag{5.31}$$

Положительно определенные матрицы рядом полезных свойств, которые мы докажем в следующей теореме.

Теорема 5.1.2. *Пусть $\boldsymbol{A}$ – положительно определенная матрица. Тогда верно следующее:*

1. *существует $\boldsymbol{A}^{-1}$;*
2. *$a_{ii} > 0, i = 1, \ldots, n$;*
3. *$\max_{1 \le i \le n} |a_{ii}| \ge \max_{1 \le k,j \le n} |a_{kj}|$;*

4. $a_{ii}a_{jj} > a_{ij}^2, i \neq j$.

Доказательство. Первый пункт доказывается через определение положительно определенной матрицы. Если $\boldsymbol{Ax} = \boldsymbol{0}$, то верно и $\boldsymbol{x}^T\boldsymbol{Ax} = 0$. По определению это возможно только в том случае, когда $\boldsymbol{x} = \boldsymbol{0}$. Соответственно $\boldsymbol{Ax} = \boldsymbol{0}$ имеет только тривиальное решение, что эквивалентно существованию обратной матрицы для $\boldsymbol{A}$.

Второй пункт доказывается через введение ненулевого вектора $\boldsymbol{x^{(i)}}$ для $i = 1, \ldots, n$:

$$\begin{cases} x_i^{(i)} = 1, \\ x_j^{(i)} = 0, \text{ при } j \neq i. \end{cases} \tag{5.32}$$

По определению имеем:

$$\begin{gathered} \left(\boldsymbol{x^{(i)}}\right)^T \boldsymbol{Ax^{(i)}} > 0 \\ \Longrightarrow \left(\boldsymbol{x^{(i)}}\right)^T \boldsymbol{Ax^{(i)}} = \left(\boldsymbol{x^{(i)}}\right)^T \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ni} \end{bmatrix} = a_{ii} > 0. \end{gathered} \tag{5.33}$$

Схожим образом доказывается и третий пункт. Определим $\boldsymbol{z^{(j,k)}}$ как:

$$\begin{cases} z_j^{(j,k)} = 1, \\ z_k^{(j,k)} = -1, \\ z_k^{(j,k)} = 0, \text{ при } i \neq j, i \neq k. \end{cases} \tag{5.34}$$

Тогда имеем:

$$\left(\boldsymbol{z^{(j,k)}}\right)^T \boldsymbol{Az^{(j,k)}} = \left(\boldsymbol{z^{(j,k)}}\right)^T \begin{bmatrix} a_{1j} - a_{1k} \\ a_{2j} - a_{2k} \\ \vdots \\ a_{nj} - a_{nk} \end{bmatrix} = a_{jj} - a_{jk} - (a_{kj} - a_{kk}) > 0. \tag{5.35}$$

В силу симметричности матрицы получаем:

$$a_{kj} < \frac{a_{jj} + a_{kk}}{2}. \tag{5.36}$$

Теперь определим $\boldsymbol{x^{(j,k)}}$ как:

$$\begin{cases} x_j^{(j,k)} = 1, \\ x_k^{(j,k)} = 1, \\ x_k^{(j,k)} = 0, \text{ при } i \neq j, i \neq k. \end{cases} \tag{5.37}$$

Аналогичным образом получаем:

$$\left(\boldsymbol{x}^{(\boldsymbol{j},\boldsymbol{k})}\right)^T \boldsymbol{A}\boldsymbol{x}^{(\boldsymbol{j},\boldsymbol{k})} = \left(\boldsymbol{x}^{(\boldsymbol{j},\boldsymbol{k})}\right)^T \begin{bmatrix} a_{1j} + a_{1k} \\ a_{2j} + a_{2k} \\ \vdots \\ a_{nj} + a_{nk} \end{bmatrix} = a_{jj} + a_{jk} + a_{kj} + a_{kk} > 0, \tag{5.38}$$

из чего следует:

$$-a_{kj} < \frac{a_{jj} + a_{kk}}{2}. \tag{5.39}$$

Иными словами, мы имеем:

$$|a_{kj}| < \frac{a_{jj} + a_{kk}}{2}. \tag{5.40}$$

Несложно заметить, что:

$$\frac{a_{jj} + a_{kk}}{2} \leq \max_{1 \leq i \leq n} |a_{ii}|, \tag{5.41}$$

что в результате дает:

$$\max_{1 \leq i \leq n} |a_{ii}| \geq \max_{1 \leq k,j \leq n} |a_{kj}|. \tag{5.42}$$

Последний пункт доказывается через модификацию вектора $\boldsymbol{x}^{(\boldsymbol{j},\boldsymbol{k})}$:

$$\begin{cases} x_j^{(j,k)} = \alpha, \\ x_k^{(j,k)} = 1, \\ x_k^{(j,k)} = 0, \text{ при } i \neq j, i \neq k. \end{cases} \tag{5.43}$$

Тогда получаем:

$$\left(\boldsymbol{x}^{(\boldsymbol{j},\boldsymbol{k})}\right)^T \boldsymbol{A}\boldsymbol{x}^{(\boldsymbol{j},\boldsymbol{k})} = \left(\boldsymbol{x}^{(\boldsymbol{j},\boldsymbol{k})}\right)^T \begin{bmatrix} \alpha a_{1j} + a_{1k} \\ \alpha a_{2j} + a_{2k} \\ \vdots \\ \alpha a_{nj} + a_{nk} \end{bmatrix} = \alpha(\alpha a_{jj} + a_{jk}) + \alpha a_{kj} + a_{kk} > 0, \tag{5.44}$$

После упрощений имеем:

$$a_{jj}\alpha^2 + 2a_{kj}\alpha + a_{kk} > 0. \tag{5.45}$$

Это неравенство выполняется, когда дискриминант меньше нуля:

$$D = 4a_{kj}^2 - 4a_{kk}a_{jj} < 0, \tag{5.46}$$

из чего следует:

$$a_{kk}a_{jj} > a_{kj}^2. \tag{5.47}$$

□

Особенностью положительно определенных матриц является тот факт, что метод Гаусса для них является устойчивым с вычислительной точки зрения и не требует перестановки строк. Более того, для положительно определенных матриц существует два специальных вида разложений – LDL^T и LL^T разложения. Разложение вида LDL^T существует также для симметричных матриц. Для случая положительно определенных матриц мы сформулируем эти разложения в виде двух теорем без доказательства.

Теорема 5.1.3. *Матрица $\boldsymbol{A}$ является положительно определенной тогда и только тогда, когда существует разложение $\boldsymbol{A} = \boldsymbol{LDL}^T$, где $\boldsymbol{L}$ – нижняя треугольная матрица с единицами на диагонали, $\boldsymbol{D}$ – диагональная матрица с положительными элементами.*

В качестве примера LDL^T-разложения рассмотрим положительно определенную матрицу размерности 3×3:

$$\begin{aligned}
\begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix} \begin{bmatrix} 1 & l_{21} & l_{31} \\ 0 & 1 & l_{32} \\ 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} d_1 & d_1 l_{21} & d_1 l_{31} \\ 0 & d_2 & d_2 l_{32} \\ 0 & 0 & d_3 \end{bmatrix} \\
&= \begin{bmatrix} d_1 & d_1 l_{21} & d_1 l_{31} \\ d_1 l_{21} & d_1 l_{21}^2 + d_2 & d_1 l_{21} l_{31} + d_2 l_{32} \\ d_1 l_{31} & d_1 l_{21} l_{31} + d_2 l_{32} & d_1 l_{31}^2 + d_2 l_{32}^2 + d_3 \end{bmatrix}
\end{aligned} . \tag{5.48}$$

Тогда неизвестные элемента матриц вычисляются следующим образом:

$$d_1 = a_{11}, \tag{5.49}$$

$$l_{21} = \frac{a_{21}}{d_1}, \tag{5.50}$$

$$l_{31} = \frac{a_{31}}{d_1}, \tag{5.51}$$

$$d_2 = a_{22} - d_1 l_{21}^2, \tag{5.52}$$

$$l_{32} = \frac{1}{d_2}\left(a_{32} - d_1 l_{21} l_{31}\right), \tag{5.53}$$

$$d_3 = a_{33} - d_1 l_{31}^2 - d_2 l_{32}^2. \tag{5.54}$$

Теорема 5.1.4. *Матрица $\boldsymbol{A}$ является положительно определенной тогда и только тогда, когда существует разложение $\boldsymbol{A} = \boldsymbol{LL}^T$, называемое разложением Холецкого, где $\boldsymbol{L}$ – нижняя треугольная матрица с ненулевыми элементами на диагонали.*

Из LDL^T-разложения можно легко получить разложение Холецкого, если разложить $\boldsymbol{D}$ как $\boldsymbol{D} = \sqrt{\boldsymbol{D}}\sqrt{\boldsymbol{D}}$. Тогда имеем:

$$\boldsymbol{LDL}^T = \boldsymbol{L}\sqrt{\boldsymbol{D}}\sqrt{\boldsymbol{D}}\boldsymbol{L}^T = \left(\boldsymbol{L}\sqrt{\boldsymbol{D}}\right)\left(\boldsymbol{L}\sqrt{\boldsymbol{D}}\right)^T = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^T. \tag{5.55}$$

В очередной раз воспользуемся положительно определенной матрицей размерности 3×3 для демонстрации разложения Холецкого:

$$\begin{aligned}\begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} &= \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix} \\ &= \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{11}l_{31} & l_{21}l_{31} + l_{22}l_{32} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix}.\end{aligned} \tag{5.56}$$

Неизвестные элементы матрицы $\boldsymbol{L}$ тогда вычисляются следующим образом:

$$l_{11} = \sqrt{a_{11}}, \tag{5.57}$$

$$l_{21} = \frac{a_{21}}{l_{11}}, \tag{5.58}$$

$$l_{31} = \frac{a_{31}}{l_{11}}, \tag{5.59}$$

$$l_{22} = \sqrt{a_{22} - l_{21}^2}, \tag{5.60}$$

$$l_{32} = \frac{1}{l_{22}} \left(a_{32} - l_{21}l_{31} \right), \tag{5.61}$$

$$l_{33} = \sqrt{a_{33} - l_{31}^2 - l_{32}^2}. \tag{5.62}$$

Несложно убедиться, что справедливы следующие общие формулы для матрицы произвольной размерности $n \times n$:

$$l_{ii} = \sqrt{a_{ii} - \sum_{j=1}^{i-1} l_{ij}^2}, \quad i = 1, \ldots, n \tag{5.63}$$

$$l_{ij} = \frac{1}{l_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk} \right), \quad j < i. \tag{5.64}$$

5.1.6 Ленточные матрицы

В ряде численных методов результирующая матрица часто является разреженной, то есть обладает преимущественно нулевыми элементами. Одним из важнейших частных случаев разреженных матриц являются ленточные матрицы.

Определение 5.1.1. *Квадратная матрица $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ называется ленточной, если существуют такие $p, q \in \mathbb{Z}$, где $1 < p, q < n$, что $j - i \geq p \Longrightarrow a_{ij} = 0$ и $i - j \geq q \Longrightarrow a_{ij} = 0$. Ширина ленты при этом определяется как $w = p + q - 1$.*

В приведенном определении p обозначает число ненулевых диагоналей над главной диагональю (включая ее), в то время как q обозначает число ненулевых диагоналей под

главной диагональю (включая ее). Частным случаем ленточной матрицы является трехдиагональная матрица, возникающая, например, при выводе разрещающих уравнений кубического сплайна:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & 0 & \dots & \dots & \dots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & & & \vdots \\ 0 & a_{32} & a_{23} & a_{34} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & \dots & \dots & \dots & 0 & a_{n,n-1} & a_{nn} \end{bmatrix}. \tag{5.65}$$

Рассмотрим специальный метод решения СЛАУ, имеющих трехдиагональную матрицу – *метод прогонки*. Пусть мы имеем следующее матричное уравнение:

$$\begin{bmatrix} a_{11} & a_{12} & 0 & \dots & \dots & \dots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & & & \vdots \\ 0 & a_{32} & a_{23} & a_{34} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & \dots & \dots & \dots & 0 & a_{n,n-1} & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}, \tag{5.66}$$

где каждая строка эквивалентна рекуррентному соотношению вида:

$$a_{i,i-1}x_{i-1} + a_{ii}x_i + a_{i,i+1}x_{i+1} = b_i. \tag{5.67}$$

Очевидно, что методом последовательного исключения можно привести такую СЛАУ к верхней треугольной форме:

$$\begin{bmatrix} \tilde{a}_{11} & \tilde{a}_{12} & 0 & \dots & \dots & 0 \\ 0 & \tilde{a}_{22} & \tilde{a}_{23} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \tilde{a}_{n-1,n-1} & \tilde{a}_{n-1,n} \\ 0 & \dots & \dots & \dots & 0 & \tilde{a}_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} \tilde{b}_1 \\ \tilde{b}_2 \\ \vdots \\ \vdots \\ \vdots \\ \tilde{b}_{n-1} \\ \tilde{b}_n \end{bmatrix}. \tag{5.68}$$

Тогда решение для x_{i-1} может быть выражено через x_i:

$$x_{i-1} = \frac{\tilde{b}_{i-1} - \tilde{a}_{i-1,i}x_i}{\tilde{a}_{i-1,i-1}}, \tag{5.69}$$

что задает выражение для обратного хода метода Гаусса. Вместо явного вывода неизвестных коэффициентов, мы построим рекуррентное соотношение, позволяющее рекурсивно найти нужные коэффициенты. Для упрощения записи переопределим коэффициенты:

$$x_{i-1} = \gamma_i x_i + \beta_i. \tag{5.70}$$

Тогда подстановка в (5.67) дает:

$$\begin{aligned} & a_{i,i-1}(\gamma_i x_i + \beta_i) + a_{ii} x_i + a_{i,i+1} x_{i+1} = b_i \\ \Longrightarrow \quad & x_i = \frac{-a_{i,i+1}}{a_{i,i-1}\gamma_i + a_{ii}} x_{i+1} + \frac{b_i - a_{i,i-1}\beta_i}{a_{i,i-1}\gamma_i + a_{ii}}. \end{aligned} \tag{5.71}$$

Сравнив полученное выражение с (5.70), получаем рекуррентные соотношения для определения γ_i и β_i:

$$\gamma_{i+1} = \frac{-a_{i,i+1}}{a_{i,i-1}\gamma_i + a_{ii}}, \tag{5.72}$$

$$\beta_{i+1} = \frac{b_i - a_{i,i-1}\beta_i}{a_{i,i-1}\gamma_i + a_{ii}}. \tag{5.73}$$

Последовательно вычислив все коэффициенты γ_i и β_i, решение СЛАУ находится с помощью обратного хода метода Гаусса по формуле (5.70). Для завершения построения метода прогонки достаточно найти выражения для коэффициентов γ_1, β_1 и неизвестной x_n. Рассмотрим уравнение для x_1:

$$\begin{aligned} & a_{11} x_1 + a_{12} x_2 = b_1 \\ \Longrightarrow \quad & x_1 = -\frac{a_{12}}{a_{11}} x_2 + \frac{b_1}{a_{11}}. \end{aligned} \tag{5.74}$$

Тогда сравнение с (5.71) дает:

$$\frac{-a_{12}}{a_{10}\gamma_1 + a_{11}} = -\frac{a_{12}}{a_{11}}, \tag{5.75}$$

$$\frac{b_1 - a_{10}\beta_1}{a_{10}\gamma_1 + a_{11}} = \frac{b_1}{a_{11}}, \tag{5.76}$$

из чего следует $\gamma_1 = \beta_1 = 0$ (или, эквивалентно, $a_{10} = 0$). Для определения x_n рассмотрим уравнение, соответствующее последней строке матрицы:

$$\begin{aligned} & a_{n,n-1} x_{n-1} + a_{nn} x_n = b_n \\ \Longrightarrow \quad & x_n = -\frac{a_{n,n-1}}{a_{nn}} x_{n-1} + \frac{b_n}{a_{nn}}. \end{aligned} \tag{5.77}$$

Подстановка в уравнение (5.70), записанное для $i = n$ дает:

$$\begin{aligned} & x_{n-1} = \gamma_n x_n + \beta_n \\ \Longrightarrow \quad & x_n = \frac{b_n - a_{n,n-1}\beta_n}{a_{nn} + a_{n,n-1}\gamma_n}. \end{aligned} \tag{5.78}$$

Таким образом, используя рекуррентные соотношения (5.70), (5.72) и (5.73), можно найти решение СЛАУ с помощью $O(n)$ операций, что делает метод прогонки предпочтительным перед методом последовательного исключения, требующим $O(n^3)$ операций. Более того, если исходная матрица обладает свойством строгого диагонального преобладания, то метод прогонки является вычислительно устойчивым.

5.2 Итерационные методы

5.2.1 Нормы векторов и матриц

Для обсуждения итерационных методов нам необходимо определить несколько норм векторов и матриц и доказать важнейшие их свойства. Будучи снабженными подходящими нормами, вектора и матрицы формируют соответствующие линейные нормированными пространства (определение 1.3.11).

Рассмотрим в первую очередь векторные нормы l_2 и l_∞, также обозначаемые как $||\cdot||_2$ и $||\cdot||_\infty$. Они определяются следующим образом:

$$||\boldsymbol{x}||_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}, \tag{5.79}$$

$$||\boldsymbol{x}||_\infty = \max_{i\in[1;n]} |x_i|. \tag{5.80}$$

где $\boldsymbol{x} = [x_1, \ldots, x_n] \in \mathbb{R}^n$. Заметим, что геометрическим местом точек $\boldsymbol{x}$, таких, что $||\boldsymbol{x}||_2 = 1$ и $||\boldsymbol{x}||_\infty = 1$, являются $(n-1)$-мерная сфера и $(n-1)$-мерный куб соответственно. Одним из главных свойств l_2-нормы является *неравенство Коши-Буняковского.*

Теорема 5.2.1 (неравенство Коши-Буняковского)**.** *Пусть $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Тогда верно следующее неравенство:*

$$\boldsymbol{x}^T\boldsymbol{y} \le ||\boldsymbol{x}||_2||\boldsymbol{y}||_2 \tag{5.81}$$

Доказательство. Доказательство для $\boldsymbol{x} = \boldsymbol{0}$ и $\boldsymbol{y} = \boldsymbol{0}$ очевидно, так что рассмотрим случай $\boldsymbol{x}, \boldsymbol{y} \neq \boldsymbol{0}$. Заметим, что для любого $\lambda \in \mathbb{R}$ верно:

$$\begin{aligned} & ||\boldsymbol{x} - \lambda\boldsymbol{y}||_2^2 \ge 0 \\ \Longrightarrow \quad & \sum_{i=1}^{n} x_i^2 - 2\lambda\sum_{i=1}^{n} x_i y_i + \lambda^2\sum_{i=1}^{n} y_i^2 \ge 0 \\ \Longrightarrow \quad & 2\lambda\boldsymbol{x}^T\boldsymbol{y} \le ||\boldsymbol{x}||_2^2 + \lambda^2||\boldsymbol{y}||_2^2. \end{aligned} \tag{5.82}$$

Выберем $\lambda = \frac{||\boldsymbol{x}||_2}{||\boldsymbol{y}||_2}$. Тогда имеем:

$$\begin{aligned} & 2\frac{||\boldsymbol{x}||_2}{||\boldsymbol{y}||_2}\boldsymbol{x}^T\boldsymbol{y} \le 2||\boldsymbol{x}||_2^2 \\ \Longrightarrow \quad & \boldsymbol{x}^T\boldsymbol{y} \le ||\boldsymbol{x}||_2||\boldsymbol{y}||_2. \end{aligned} \tag{5.83}$$

□

Из этого неравенства следует неравенство треугольника, которому по определению должна удовлетворять норма:

$$||\boldsymbol{x} + \boldsymbol{y}||_2 \leq ||\boldsymbol{x}||_2 + ||\boldsymbol{y}||_2. \tag{5.84}$$

Расстояние между двумя векторами, индуцированное нормой $||\cdot||$, имеет вид $||\boldsymbol{x} - \boldsymbol{y}||$.

Перейдем к рассмотрению норм матриц. Кроме стандартных аксиом норм линейных нормированных пространств, нормы квадартных матриц также обязаны удовлетворяют аксиоме $||\boldsymbol{AB}|| \leq ||\boldsymbol{A}||||\boldsymbol{B}||$. Здесь и далее мы будет использовать матричные нормы, построенные на основе векторных норм. Такие нормы матриц называются *индуцированными*. Для вывода индуцированных матричных норм из векторных норм l_2 и l_∞, нам необходимо сформулировать следующую теорему без доказательства:

Теорема 5.2.2. *Пусть $||\cdot||$ – векторная норма в пространстве $\mathbb{R}^n$. Тогда следующий функционал является нормой матрицы:*

$$||\boldsymbol{A}|| = \max_{||\boldsymbol{x}||=1} ||\boldsymbol{Ax}||. \tag{5.85}$$

Так как $\boldsymbol{x} = \frac{\boldsymbol{z}}{||\boldsymbol{z}||}$, где $\boldsymbol{z} \neq \boldsymbol{0}$, удовлетворяет требованию $||\boldsymbol{x}|| = 1$, мы можем переписать утверждение теоремы 5.2.2 следующим образом:

$$||\boldsymbol{A}|| = \max_{||\boldsymbol{z}||\neq 0} \frac{||\boldsymbol{Az}||}{||\boldsymbol{z}||}. \tag{5.86}$$

Очевидным является также следующее следствие:

Следствие 5.2.1. *Для любого вектора $\boldsymbol{x} \neq \boldsymbol{0}$, матрицы $\boldsymbol{A}$ и индуцированной матричной нормы $||\cdot||$ верно следующее неравенство:*

$$||\boldsymbol{Ax}|| \leq ||\boldsymbol{A}|| \cdot ||\boldsymbol{x}||. \tag{5.87}$$

Исходя из теоремы 5.2.2, мы можем записать индуцированные матричные нормы $||\cdot||_2$ и $||\cdot||_\infty$ как:

$$||\boldsymbol{A}||_2 = \max_{||\boldsymbol{x}||_2=1} ||\boldsymbol{Ax}||_2, \tag{5.88}$$

$$||\boldsymbol{A}||_\infty = \max_{||\boldsymbol{x}||_\infty=1} ||\boldsymbol{Ax}||_\infty. \tag{5.89}$$

$$\tag{5.90}$$

Более того, можно доказать, что эти нормы имеют следующий вид:

$$||\boldsymbol{A}||_2 = \sqrt{\rho(\boldsymbol{A}^T\boldsymbol{A})}, \tag{5.91}$$

$$||\boldsymbol{A}||_\infty = \max_{i\in[1;n]} \sum_{j=1}^{n} |a_{ij}|, \tag{5.92}$$

где $\rho(\boldsymbol{A}^T\boldsymbol{A})$ – спектральный радиус матрицы $\boldsymbol{A}^T\boldsymbol{A}$, определение которому будет дано ниже.

5.2.2 Собственные числа и вектора

Проблема о собственных числах и векторах матрицы появляется в самых разных контекстах математической физики. В численных методах они играют важную роль в определении сходимости итерационного процесса, так что нам необходимо ознакомиться с некоторыми важными свойствами собственных чисел и собственных векторов. В первую очередь рассмотрим ряд определений.

Определение 5.2.1. *Пусть* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$*. Тогда собственным вектором матрицы* $\boldsymbol{A}$ *называется такой вектор* $\boldsymbol{x} \neq \boldsymbol{0}$*, что:*

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}, \tag{5.93}$$

где $\lambda \in \mathbb{R}$ *называется собственным числом матрицы* $\boldsymbol{A}$*, ассоциированным с собственным вектором* $\boldsymbol{x}$*.*

Несложно заметить, что собственный вектор $\boldsymbol{x}$ является нетривиальным решением однородной СЛАУ:

$$(\boldsymbol{A} - \lambda \boldsymbol{E})\boldsymbol{x} = \boldsymbol{0}, \tag{5.94}$$

которая будет иметь нетривиальное решение только в том случае, когда:

$$\det(\boldsymbol{A} - \lambda \boldsymbol{E}) = 0. \tag{5.95}$$

Определение 5.2.2. *Пусть* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$*. Тогда характеристическим многочленом матрицы* $\boldsymbol{A}$ *является многочлен:*

$$P_n(\lambda) = \det(\boldsymbol{A} - \lambda \boldsymbol{E}) = \sum_{k=0}^{n} a_k \lambda^k \tag{5.96}$$

Таким образом, собственные числа матрицы являются корнями ее характеристического многочлена.

Заметим, что собственные вектора матрицы определены вплоть до произвольного множителя и являются такими направлениями в пространстве $\mathbb{R}^n$, вдоль которых действие оператора $\boldsymbol{A}$ эквивалентно коллинеарному преобразованию (т.е. растяжению или сжатию вектора в зависимости от значения $|\lambda|$), что продемонстрировано на рисунке 5.1.

Важнейшим свойством собственных векторов является тот факт, что в случае, когда они все являются линейно независимыми, они формируют базис данного пространства. Это позволяет разложить любой вектор, на который действует оператор $\boldsymbol{A}$, в линейную комбинацию собственных векторов. Тогда, например, если все собственные числа по модулю меньше 1, то можно утверждать, что для любого вектора $\boldsymbol{x_1}$ вектор $\boldsymbol{x_2} = \boldsymbol{A}\boldsymbol{x_1}$ будет удовлетворять неравенству $||x_2|| \leq ||x_1||$. Эта простая мысль приводит нас к рассмотрению спектрального радиуса, сходящихся матриц и связи между ними.

Определение 5.2.3. *Спектральным радиусом матрицы* $\boldsymbol{A}$ *называется число* $\rho(\boldsymbol{A}) \in \mathbb{R}$ *такое, что:*

$$\rho(\boldsymbol{A}) = \max_{i\in[1,m]} |\lambda_i|, \tag{5.97}$$

где λ_i *– одно из* m *собственных чисел матрицы* $\boldsymbol{A}$*.*

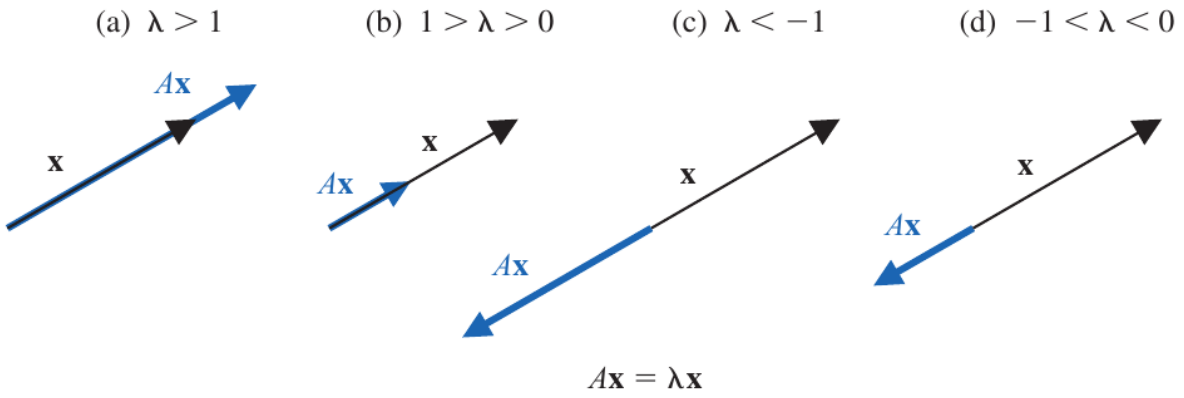

Рисунок 5.1 – Коллинеарное преобразование собственного вектора $\boldsymbol{x}$ под действием оператора $\boldsymbol{A}$. В случаях (a) и (c) ассоциированное собственное число λ больше единицы по модулю, что приводит к растяжению собственного вектора. В случаях (b) и (d) мы имеем $|\lambda| < 1$, что приводит к сжатию собственного вектора.

Мы уже ознакомились с одним из случаев использования спектрального радиуса:

$$||\boldsymbol{A}||_2 = \sqrt{\rho(\boldsymbol{A}^T\boldsymbol{A})}. \tag{5.98}$$

Теперь докажем другое важное свойство спектрального радиуса.

Теорема 5.2.3. *Пусть $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Тогда $\rho(\boldsymbol{A}) \leq ||\boldsymbol{A}||$ для любой индуцированной матричной нормы $||\cdot||$.*

Доказательство. Рассмотрим собственное число λ с собственным вектором $\boldsymbol{x}$. Тогда имеем:

$$\begin{aligned} & \boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \\ \implies \quad & ||\boldsymbol{A}\boldsymbol{x}|| = ||\lambda\boldsymbol{x}|| \\ \implies \quad & ||\boldsymbol{A}\boldsymbol{x}|| = |\lambda| \cdot ||\boldsymbol{x}||. \end{aligned} \tag{5.99}$$

Так как $||\boldsymbol{A}\boldsymbol{x}|| \leq ||\boldsymbol{A}|| \cdot ||\boldsymbol{x}||$, мы получаем:

$$\begin{aligned} & |\lambda| \cdot ||\boldsymbol{x}|| \leq ||\boldsymbol{A}|| \cdot ||\boldsymbol{x}|| \\ \implies \quad & |\lambda| \leq ||\boldsymbol{A}||, \end{aligned} \tag{5.100}$$

что приводит к $\rho(A) = \max|\lambda| \leq ||\boldsymbol{A}||$. □

Расширением этой теоремы является следующая теорема, которую мы рассмотрим без доказательства.

Теорема 5.2.4. *Для любой матрицы $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ и любого $\epsilon > 0$ можно найти такую индуцированную норму, что:*

$$\rho(\boldsymbol{A}) \leq ||\boldsymbol{A}|| \leq \rho(\boldsymbol{A}) + \epsilon. \tag{5.101}$$

5.2.3 Сходящиеся матрицы

Так как итерационные методы построены на большом количестве перемножений матриц на самих себя, т.е. возведении их в степень, важно рассмотреть такие матрицы, которые при такой операции сходятся к нулевой матрице.

Определение 5.2.4. *Матрица $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ называется сходящейся, если для нее верно:*

$$\lim_{k\to\infty} \boldsymbol{A}^k = \boldsymbol{O}, \tag{5.102}$$

где $\boldsymbol{O}$ – нулевая матрица.

Несложно доказать, что если матрица является сходящейся, то к нулю сходится любая матричная норма степеней матрицы:

$$\lim_{k\to\infty} ||\boldsymbol{A}^k|| = 0. \tag{5.103}$$

Более того, в случае индуцированных норм сходимость в одной из норм эквивалентна сходимости в любой другой норме.

Важнейшим свойством сходящихся матриц является тот факт, что их спектральный радиус всегда строго меньше единицы.

Теорема 5.2.5. *Матриц $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ является сходящейся тогда и только тогда, когда $\rho(\boldsymbol{A}) < 1$.*

Доказательство. Докажем утверждение в обе стороны относительно схождения норм:

$$\lim_{k\to\infty} ||\boldsymbol{A}^k|| = 0. \tag{5.104}$$

В первую очередь докажем $\lim_{k\to\infty} ||\boldsymbol{A}^k|| = 0 \implies \rho(\boldsymbol{A}) < 1$. Для начала заметим, что $\rho(\boldsymbol{A}^k) = \rho^k(\boldsymbol{A})$. Действительно, пусть λ собственное число матрицы $\boldsymbol{A}$, ассоциированное с собственным вектором $\boldsymbol{x}$. Предположим, что $\boldsymbol{A}^k$ имеет собственное число $\tilde{\lambda}$, ассоциированное с тем же собственным вектором. Тогда имеем:

$$\begin{aligned} & \boldsymbol{A}^k\boldsymbol{x} = \boldsymbol{A}\cdots\boldsymbol{A}\boldsymbol{A}\boldsymbol{x} = \tilde{\lambda}\boldsymbol{x} \\ \implies\quad & \boldsymbol{A}\cdots\boldsymbol{A}\lambda\boldsymbol{x} = \tilde{\lambda}\boldsymbol{x} \\ \implies\quad & \lambda^k\boldsymbol{x} = \tilde{\lambda}\boldsymbol{x} \\ \implies\quad & \tilde{\lambda} = \lambda^k. \end{aligned} \tag{5.105}$$

Очевидным следствием из этого является равенство $\rho(\boldsymbol{A}^k) = \rho^k(\boldsymbol{A})$. Теперь, воспользовавшись теоремой 5.2.3, получаем:

$$\begin{aligned} & \rho(\boldsymbol{A}^k) \le ||\boldsymbol{A}^k|| \\ \implies\quad & \rho^k(\boldsymbol{A}) \le ||\boldsymbol{A}^k||, \end{aligned} \tag{5.106}$$

что приводит нас к утверждению:

$$\lim_{k\to\infty} ||\boldsymbol{A}^k|| = 0 \implies \lim_{k\to\infty} \rho^k(\boldsymbol{A}) = 0 \iff \rho(\boldsymbol{A}) < 1. \tag{5.107}$$

Теперь докажем утверждение в обратную сторону: $\rho(\boldsymbol{A}) < 1 \implies \lim_{k\to\infty} ||\boldsymbol{A}^k|| = 0$. По теореме 5.2.4 мы можем найти такую норму $||\cdot||$ и $\epsilon > 0$, что:

$$||\boldsymbol{A}|| \leq \rho(\boldsymbol{A}) + \epsilon < 1. \tag{5.108}$$

Тогда пользуясь одной из аксиом матричных норм, имеем:

$$||\boldsymbol{A}^k|| \leq ||\boldsymbol{A}||^k \leq (\rho(\boldsymbol{A}) + \epsilon)^k \implies \lim_{k\to\infty} ||\boldsymbol{A}^k|| = 0. \tag{5.109}$$

□

Последнее свойство сходящихся матриц, которое мы рассмотрим, очевидно следует из предыдущих рассуждений:

$$\lim_{k\to\infty} \boldsymbol{A}^k \boldsymbol{x} = \boldsymbol{0}. \tag{5.110}$$

5.2.4 Методы простой итерации

Как мы уже упомянули в начале раздела, итерационные методы ищут такое $\boldsymbol{x}^{(k)}$, что $\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b} \to \boldsymbol{0}$ при $k \to \infty$. Зачастую такие методы могут быть записаны в виде следующей итерации:

$$\boldsymbol{x}^{(k)} = \boldsymbol{T}\boldsymbol{x}^{(k-1)} + \boldsymbol{c}, \quad k = 1, 2, \ldots, \tag{5.111}$$

где $\boldsymbol{T}$ и $\boldsymbol{c}$ некоторые матрица и вектор, получаемые в результате разложения $\boldsymbol{A}$ и $\boldsymbol{b}$ тем или иным образом. Начальное приближение $\boldsymbol{x}^{(0)}$ предполагается случайным. Решением такого рекуррентного соотношения является неподвижная точка $\boldsymbol{x}$:

$$\boldsymbol{x} = \boldsymbol{T}\boldsymbol{x} + \boldsymbol{c}. \tag{5.112}$$

Таким образом, нам необходимо определить, какие условия следует наложить на матрицу $\boldsymbol{T}$, чтобы итерации сходились к неподвижной точке, т.е. $\boldsymbol{x}^{(k)} \to \boldsymbol{x}$ при $k \to \infty$. Прежде чем перейти к соответствующей теореме, рассмотрим вспомогательную лемму.

Лемма 5.2.1. *Пусть $\rho(\boldsymbol{T}) < 1$. Тогда матрица $(\boldsymbol{E} - \boldsymbol{T})$ имеет обратную, и при этом:*

$$(\boldsymbol{E} - \boldsymbol{T})^{-1} = \boldsymbol{E} + \boldsymbol{T} + \boldsymbol{T}^2 + \cdots = \sum_{j=0}^{\infty} \boldsymbol{T}^j \tag{5.113}$$

Доказательство. Докажем существование обратной матрицы, доказав существования только ненулевых собственных чисел. Действительно, рассмотрим произвольное собственное число λ матрицы $\boldsymbol{T}$:

$$\begin{aligned} & \boldsymbol{T}\boldsymbol{x} = \lambda\boldsymbol{x} \\ \implies \quad & (\boldsymbol{E} - \boldsymbol{T})\boldsymbol{x} = (1 - \lambda)\boldsymbol{x}. \end{aligned} \tag{5.114}$$

Иными словами, собственные числа матрицы $\boldsymbol{E} - \boldsymbol{T}$ равны $1 - \lambda$. Мы знаем, что $|\lambda| \leq \rho(\boldsymbol{T}) < 1$, из чего следует:

$$0 < 1 - \lambda < 2, \tag{5.115}$$

что означает, что матрица $\boldsymbol{E}-\boldsymbol{T}$ невырожденная. Для доказательства второй части леммы рассмотрим следующую матрицу:

$$\boldsymbol{S}_m = \boldsymbol{E} + \boldsymbol{T} + \boldsymbol{T}^2 + \cdots + \boldsymbol{T}^m. \tag{5.116}$$

Домножив ее слева на $\boldsymbol{E} - \boldsymbol{T}$, получаем:

$$\begin{aligned} & (\boldsymbol{E} - \boldsymbol{T})\boldsymbol{S}_m = (\boldsymbol{E} + \boldsymbol{T} + \boldsymbol{T}^2 + \cdots + \boldsymbol{T}^m) - (\boldsymbol{T} + \boldsymbol{T}^2 + \cdots + \boldsymbol{T}^{m+1}) \\ \Longrightarrow \quad & (\boldsymbol{E} - \boldsymbol{T})\boldsymbol{S}_m = \boldsymbol{E} - \boldsymbol{T}^{m+1}. \end{aligned} \tag{5.117}$$

Рассмотрим предел этого выражения:

$$\lim_{m\to\infty} (\boldsymbol{E} - \boldsymbol{T})\boldsymbol{S}_m = \lim_{m\to\infty} (\boldsymbol{E} - \boldsymbol{T}^{m+1}) = \boldsymbol{E}, \tag{5.118}$$

где последнее равенство справедливо, так как $\rho(\boldsymbol{T}) < 1$. Таким образом, мы получаем:

$$(\boldsymbol{E} - \boldsymbol{T})^{-1} = \boldsymbol{S}_\infty = \sum_{j=0}^{\infty} \boldsymbol{T}^j. \tag{5.119}$$

□

Теперь мы можем рассмотреть фундаментальную теорему о сходимости метода простой итерации.

Теорема 5.2.6. *Последовательность $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$, сгенерированная итерацией $\boldsymbol{x}^{(k)} = \boldsymbol{T}\boldsymbol{x}^{(k-1)} + \boldsymbol{c}$, сходится к единственному решению уравнения $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{x} + \boldsymbol{c}$, т.е. неподвижной точке $\boldsymbol{x}$, для любого $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ тогда и только тогда, когда $\rho(\boldsymbol{T}) < 1$.*

Доказательство. В первую очередь докажем обратную часть теоремы, т.е. предположим, что $\rho(\boldsymbol{T}) < 1$. Тогда, развернув все итерации, имеем:

$$\begin{aligned} \boldsymbol{x}^{(k)} &= \boldsymbol{T}\boldsymbol{x}^{(k-1)} + \boldsymbol{c} \\ &= \boldsymbol{T}(\boldsymbol{T}\boldsymbol{x}^{(k-2)} + \boldsymbol{c}) + \boldsymbol{c} \\ &= \boldsymbol{T}^2\boldsymbol{x}^{(k-2)} + (\boldsymbol{T} + \boldsymbol{E})\boldsymbol{c} \\ &= \dots \\ &= \boldsymbol{T}^k\boldsymbol{x}^{(0)} + (\boldsymbol{T}^{k-1} + \cdots + \boldsymbol{T} + \boldsymbol{E})\boldsymbol{c}. \end{aligned} \tag{5.120}$$

Рассмотрим предел $\lim_{k\to\infty} \boldsymbol{x}^{(k)}$. Тогда, воспользовавшись фактом $\rho(\boldsymbol{T}) < 1$ и леммой 5.2.1, имеем:

$$\begin{aligned} \boldsymbol{x} = \lim_{k\to\infty} \boldsymbol{x}^{(k)} &= \lim_{k\to\infty} \left[\boldsymbol{T}^k\boldsymbol{x}^{(0)} + (\boldsymbol{T}^{k-1} + \cdots + \boldsymbol{T} + \boldsymbol{E})\boldsymbol{c}\right] \\ &= \lim_{k\to\infty} (\boldsymbol{T}^{k-1} + \cdots + \boldsymbol{T} + \boldsymbol{E})\boldsymbol{c} \\ &= (\boldsymbol{E} - \boldsymbol{T})^{-1}\boldsymbol{c}, \end{aligned} \tag{5.121}$$

из чего следует $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{x} + \boldsymbol{c}$. Теперь рассмотрим прямой случай, т.е. предположим, что $\boldsymbol{x} = \lim_{k\to\infty} \boldsymbol{x}^{(k)}$ есть единственное решение уравнения $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{x} + \boldsymbol{c}$. Так как $\rho(\boldsymbol{T}) < 1 \iff \lim_{k\to\infty} \boldsymbol{T}^k\boldsymbol{z} = 0$ для любого $\boldsymbol{z}$, докажем вторую часть утверждения. Пусть $\boldsymbol{x}^{(0)} = \boldsymbol{x} - \boldsymbol{z}$. Заметим, что:

$$\begin{aligned} \boldsymbol{x} - \boldsymbol{x}^{(k)} &= \boldsymbol{T}\boldsymbol{x} + \boldsymbol{c} - (\boldsymbol{T}\boldsymbol{x}^{(k-1)} + \boldsymbol{c}) \\ &= \boldsymbol{T}(\boldsymbol{x} - \boldsymbol{x}^{(k-1)}) \\ &= \boldsymbol{T}^2(\boldsymbol{x} - \boldsymbol{x}^{(k-2)}) \\ &= \dots \\ &= \boldsymbol{T}^k(\boldsymbol{x} - \boldsymbol{x}^{(0)}) \\ &= \boldsymbol{T}^k\boldsymbol{z}. \end{aligned} \tag{5.122}$$

Однако по условию теоремы $\lim_{k\to\infty}(\boldsymbol{x} - \boldsymbol{x}^{(k)}) = \boldsymbol{0}$, из чего следует:

$$\lim_{k\to\infty} \boldsymbol{T}^k\boldsymbol{z} = 0. \tag{5.123}$$

Так как $\boldsymbol{z}$ был выбран произвольно, это эквивалентно утверждению $\rho(\boldsymbol{T}) < 1$. □

Из доказательства теоремы можно вывести две полезные оценки погрешности. Рассмотрим одно из уравнений из 5.122:

$$\boldsymbol{x} - \boldsymbol{x}^{(k)} = \boldsymbol{T}^k(\boldsymbol{x} - \boldsymbol{x}^{(0)}), \tag{5.124}$$

и оценим его:

$$\begin{aligned} & ||\boldsymbol{x} - \boldsymbol{x}^{(k)}|| = ||\boldsymbol{T}^k(\boldsymbol{x} - \boldsymbol{x}^{(0)})|| \\ \implies \quad & ||\boldsymbol{x} - \boldsymbol{x}^{(k)}|| \le ||\boldsymbol{T}^k|| \cdot ||\boldsymbol{x} - \boldsymbol{x}^{(0)}||, \end{aligned} \tag{5.125}$$

из чего следует оценка погрешности метода:

$$||\boldsymbol{x} - \boldsymbol{x}^{(k)}|| \le ||\boldsymbol{T}||^k ||\boldsymbol{x} - \boldsymbol{x}^{(0)}||. \tag{5.126}$$

Для вывода второй оценки рассмотрим следующее выражение:

$$\begin{aligned} \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} &= \boldsymbol{T}\boldsymbol{x}^{(k)} + \boldsymbol{c} - \boldsymbol{T}\boldsymbol{x}^{(k-1)} - \boldsymbol{c} \\ &= \boldsymbol{T}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}) \\ &= \dots \\ &= \boldsymbol{T}^k(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}). \end{aligned} \tag{5.127}$$

Тогда для $m > k \ge 1$ справедливо следующее:

$$\begin{aligned} ||\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(k)}|| &= ||\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m-1)} + \boldsymbol{x}^{(m-1)} - \dots + \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}|| \\ &\le ||\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m-1)}|| + ||\boldsymbol{x}^{(m-1)} - \boldsymbol{x}^{(m-2)}|| + \dots + ||\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}|| \\ &\le (||\boldsymbol{T}||^{m-1} + ||\boldsymbol{T}||^{m-2} + \dots ||\boldsymbol{T}||^k)||\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}|| \\ &= ||\boldsymbol{T}||^k(1 + \dots + ||\boldsymbol{T}||^{m-k-1} + ||\boldsymbol{T}||^{m-k-2})||\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}||. \end{aligned} \tag{5.128}$$

В пределе $m \to \infty$ мы имеем:

$$\begin{aligned} ||\boldsymbol{x} - \boldsymbol{x}^{(k)}|| &= \lim_{m\to\infty} ||\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(k)}|| \\ &= ||\boldsymbol{T}||^k \sum_{j=0}^{\infty} ||\boldsymbol{T}||^j \cdot ||\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}||. \end{aligned} \tag{5.129}$$

Сумма представляет собой сумму геометрической прогрессии, и так как $\rho(\boldsymbol{T}) < 1 \Longrightarrow ||\boldsymbol{T}|| < 1$ мы имеем:

$$||\boldsymbol{x} - \boldsymbol{x}^{(k)}|| = \frac{||\boldsymbol{T}||^k}{1 - ||\boldsymbol{T}||} ||\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}||. \tag{5.130}$$

Таким образом скорость сходимости методов простой итерации зависит от $||\boldsymbol{T}||$, и вследствие $\rho(\boldsymbol{T}) \approx ||\boldsymbol{T}||$ она, в сущности, зависит от спектрального радиуса матрицы $\boldsymbol{T}$:

$$||\boldsymbol{x} - \boldsymbol{x}^{(k)}|| \approx \rho(\boldsymbol{T})^k ||\boldsymbol{x} - \boldsymbol{x}^{(0)}||. \tag{5.131}$$

Очевидно, что мы хотели бы найти итерационный метод с минимально возможным $\rho(\boldsymbol{T})$.

В конце отметим, что в качестве условия окончания итерационного процесса логично выбрать ограничение на относительное улучшение решения:

$$\frac{||x^{(k)} - x^{(k-1)}||}{||x^{(k)}||} < \epsilon, \tag{5.132}$$

где $\epsilon \in \mathbb{R}$ – выбранная точность решения.

Рассмотрев метод простой итерации в общем виде, мы можем перейти к его частным случаям, а именно к методам Якоби и Гаусса–Зейделя.

5.2.5 Метод Якоби

Метод Якоби строит итерационную процедуру, напрямую решая i-ое уравнение СЛАУ:

$$x_i = \frac{1}{a_{ii}} \left(- \sum_{\substack{j=1 \\ i \neq j}}^{n} a_{ij} x_j + b_i \right), \tag{5.133}$$

и находя решение на шаге k из решения на шаге $k-1$:

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(- \sum_{\substack{j=1 \\ i \neq j}}^{n} a_{ij} x_j^{(k-1)} + b_i \right), \tag{5.134}$$

Легко убедиться, что в матричном виде метод Якоби предполагает разложение матрицы $\boldsymbol{A}$ на диагональную, нижнюю треугольную и верхнюю треугольную составляющие:

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & a_{nn} \end{bmatrix} - \begin{bmatrix} 0 & \dots & \dots & 0 \\ -a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -a_{n1} & \dots & -a_{n,n-1} & 0 \end{bmatrix} - \begin{bmatrix} 0 & -a_{12} & \dots & -a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -a_{n-1,n} \\ 0 & \dots & \dots & 0 \end{bmatrix} \tag{5.135}$$

или кратко:

$$\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{L} - \boldsymbol{U}. \tag{5.136}$$

Подстановка в СЛАУ дает:

$$\begin{aligned} & (\boldsymbol{D} - \boldsymbol{L} - \boldsymbol{U})\boldsymbol{x} = \boldsymbol{b} \\ \Longrightarrow \quad & \boldsymbol{D}\boldsymbol{x} = (\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x} + \boldsymbol{b}. \end{aligned} \tag{5.137}$$

При этом, если $\boldsymbol{D}$ имеет обратную (т.е. все диагональные элементы ненулевые), то можно записать:

$$\boldsymbol{x} = \boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x} + \boldsymbol{D}^{-1}\boldsymbol{b}, \tag{5.138}$$

что легко превращается в следующее рекуррентное соотношение:

$$\boldsymbol{x}^{(k)} = \boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x}^{(k-1)} + \boldsymbol{D}^{-1}\boldsymbol{b}, \tag{5.139}$$

из которого мы получаем матрицу $\boldsymbol{T}$ и вектор $\boldsymbol{c}$ в контексте метода простой итерации:

$$\boldsymbol{T_J} = \boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{U}), \tag{5.140}$$

$$\boldsymbol{c_J} = \boldsymbol{D}^{-1}\boldsymbol{b}. \tag{5.141}$$

5.2.6 Метод Гаусса–Зейделя

Метод Якоби можно улучшить, если заметить, что в уравнении (5.134) вместо $x_j^{(k-1)}, j < i$ можно использовать уже посчитанные $x_j^{(k)}, j < i$, которые априори будут "ближе" к точному решению. Это наблюдение приводит нас к методу Гаусса–Зейделя:

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} + b_i \right), \tag{5.142}$$

В матричном виде метод Гаусса–Зейделя можно записать в следующем виде:

$$\begin{aligned} & (\boldsymbol{D} - \boldsymbol{L})\boldsymbol{x}^{(k)} = \boldsymbol{U}\boldsymbol{x}^{(k-1)} + \boldsymbol{b} \\ \Longrightarrow \quad & \boldsymbol{x}^{(k)} = (\boldsymbol{D} - \boldsymbol{L})^{-1}\boldsymbol{U}\boldsymbol{x}^{(k-1)} + (\boldsymbol{D} - \boldsymbol{L})^{-1}\boldsymbol{b}, \end{aligned} \tag{5.143}$$

где матрица $(\boldsymbol{D} - \boldsymbol{L})^{-1}$ существует тогда и только тогда, когда все диагональные элементы являются ненулевыми. В свою очередь в контексте метода простой итерации мы получаем следующие выражения для $\boldsymbol{T}$ и $\boldsymbol{c}$:

$$\boldsymbol{T_G} = (\boldsymbol{D} - \boldsymbol{L})^{-1}\boldsymbol{U}, \tag{5.144}$$

$$\boldsymbol{c_G} = (\boldsymbol{D} - \boldsymbol{L})^{-1}\boldsymbol{b}. \tag{5.145}$$

Методы Якоби и Гаусса–Зейделя являются безусловно сходящимися для матриц со строго диагональным преобладанием. Докажем это утверждение для метода Якоби.

Теорема 5.2.7. *Пусть $\boldsymbol{A}$ – матрица со строгим диагональным преобладанием. Тогда метод Якоби формирует последовательность $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$, которая сходится к единственному решению $\boldsymbol{x}$.*

Доказательство. Найдем верхнюю границу для $||\boldsymbol{T_J}||_\infty$. По определению мы имеем:

$$\begin{aligned} ||\boldsymbol{T_J}||_\infty &= \max_{i\in[1;n]} \sum_{j=1}^{n} \left| \frac{1}{a_{ii}}(l_{ij}+u_{ij}) \right| \\ &\leq \max_{i\in[1;n]} \frac{\sum\limits_{j=1}^{n} |l_{ij}+u_{ij}|}{|a_{ii}|}. \end{aligned} \tag{5.146}$$

Заметим, что $l_{ij}+u_{ij}=a_{ij}$ для $i\neq j$ и $l_{ij}+u_{ij}=0$ для $i=j$. Тогда достаточно записать:

$$||\boldsymbol{T_J}||_\infty \leq \max_{i\neq j} \frac{|a_{ij}|}{|a_{ii}|}. \tag{5.147}$$

По определению матриц со строгим диагональным преобладанием мы имеем:

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i=1,\ldots,n, \tag{5.148}$$

из чего следует:

$$\frac{\sum\limits_{j=1, j\neq i}^{n} |a_{ij}|}{|a_{ii}|} < 1. \tag{5.149}$$

что немедленно дает:

$$||\boldsymbol{T_J}||_\infty < 1, \tag{5.150}$$

что равносильно $\rho(\boldsymbol{T_J}) < 1$. □

5.2.7 Методы релаксации

Как показывает выражение (5.131), скорость сходимости итерационного метода напрямую зависит от спектрального радиуса матрицы $\boldsymbol{T}$ – чем он меньше, тем больше скорость сходимости к точному результату. Методы релаксации, как мы скоро покажем, позволяют модифицировать матрицу $\boldsymbol{T}$ так, что ее спектральный радиус уменьшается.

Для начала необходимо ввести новый показатель близости решения СЛАУ на определенной итерации к ее точному решению.

Определение 5.2.5. *Пусть $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ является приближением к точному решению СЛАУ $\boldsymbol{Ax}=\boldsymbol{b}$. Тогда вектором невязки называется $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\tilde{\boldsymbol{x}}$.*

На примере метода Гаусса–Зейделя рассмотрим связь между итерационными уравнениями и вектором невязки. Пусть $\boldsymbol{r}_i^{(k)} = [r_{1i}^{(k)}, r_{2i}^{(k)}, \ldots, r_{ni}^{(k)}]^T$ обозначает вектор невязки,

соответствующий k-й итерации метода при расчете $x_i^{(k)}$ и, следовательно, вектору приближения к решению $\boldsymbol{x}_i^{(k)} = [x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k)}, \ldots, x_n^{(k)}]^T$. Элемент вектора невязки $r_{mi}^{(k)}$ тогда имеет вид:

$$\begin{aligned} r_{mi}^{(k)} &= b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i}^{n} a_{mj} x_j^{(k-1)} \\ &= b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i+1}^{n} a_{mj} x_j^{(k-1)} - a_{mi} x_i^{(k-1)}. \end{aligned} \tag{5.151}$$

Тогда i-й компонент вектора невязки будет иметь вид:

$$\begin{aligned} r_{ii}^{(k)} &= b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} - a_{ii} x_i^{(k-1)} \\ \implies \quad r_{ii}^{(k)} + a_{ii} x_i^{(k-1)} &= b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)}. \end{aligned} \tag{5.152}$$

Подставив выражение для $x_i^{(k)}$ метода Гаусса-Зейделя из уравнения (5.142), получаем:

$$\begin{aligned} r_{ii}^{(k)} + a_{ii} x_i^{(k-1)} &= a_{ii} x_i^{(k)} \\ \implies \quad x_i^{(k)} &= x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}, \end{aligned} \tag{5.153}$$

что задает связь между итерацией метода Гаусса-Зейделя и вектором невязки. *Методы релаксации* основаны на модификации уравнения (5.153):

$$x_i^{(k)} = x_i^{(k-1)} + \omega \frac{r_{ii}^{(k)}}{a_{ii}}, \tag{5.154}$$

где $\omega > 0$. При $0 < \omega < 1$ мы имеем *методы нижней релаксации*, а при $\omega > 1$ *методы верхней релаксации*. Последние оказались подходящими для большинства практических случаев, так что позже мы рассмотрим их подробнее. Подставив выражение для $r_{ii}^{(k)}$ из (5.152), получаем формулу, которая может использоваться для реализации итерационного метода верхней релаксации:

$$x_i^{(k)} = (1-\omega) x_i^{(k-1)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right). \tag{5.155}$$

Для вывода матричной формы переформулируем выражение выше следующим образом:

$$a_{ii} x_i^{(k)} + \omega \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} = (1-\omega) a_{ii} x_i^{(k-1)} - \omega \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} + \omega b_i, \tag{5.156}$$

что в матричной форме дает:

$$(\boldsymbol{D} - \omega \boldsymbol{L})\boldsymbol{x}^{(k)} = \left[(1-\omega)\boldsymbol{D} + \omega \boldsymbol{U}\right]\boldsymbol{x}^{(k-1)} + \omega \boldsymbol{b}. \tag{5.157}$$

В форме простой итерации мы имеем:

$$\boldsymbol{x}^{(k)} = \boldsymbol{T}_{\boldsymbol{\omega}}\boldsymbol{x}^{(k-1)} + \boldsymbol{c}_{\boldsymbol{\omega}}. \tag{5.158}$$

где $\boldsymbol{T}_{\boldsymbol{\omega}} = (\boldsymbol{D} - \omega \boldsymbol{L})^{-1}\left[(1-\omega)\boldsymbol{D} + \omega \boldsymbol{U}\right]$ и $\boldsymbol{c}_{\boldsymbol{\omega}} = \omega(\boldsymbol{D} - \omega \boldsymbol{L})^{-1}\boldsymbol{b}$.

Выбор значения ω не всегда является тривиальной задачей, и очевидного ответа для общего случая СЛАУ нет. Однако для некоторых частных случаев можно получить оценки интервала, которому должен принадлежать ω для сходимости метода. Рассмотрим эти случаи в виде ряда теорем без доказательств.

Теорема 5.2.8. *Если $a_{ii} \neq 0$, $i = 1, \ldots, n$, то $\rho(\boldsymbol{T}_{\boldsymbol{\omega}}) \geq |\omega - 1|$. Следовательно, для этого случая метод верхней релаксации может сойтись только при $0 \leq \omega \leq 2$.*

Теорема 5.2.9. *Если $\boldsymbol{A}$ – положительная определенная матрица и $0 \leq \omega \leq 2$, то метод верхней релаксации сходится для любого $\boldsymbol{x}^{(0)}$.*

Теорема 5.2.10. *Если $\boldsymbol{A}$ – положительная определенная и трехдиагональная матрица, то $\rho(\boldsymbol{T}_{\boldsymbol{G}}) = \rho^2(\boldsymbol{T}_{\boldsymbol{J}}) < 1$ и оптимальное значение ω вычисляется как:*

$$\omega^{(opt)} = \frac{2}{1 + \sqrt{1 - \rho^2(\boldsymbol{T}_{\boldsymbol{J}})}}, \tag{5.159}$$

и при этом $\rho(\boldsymbol{T}_{\boldsymbol{\omega}^{(opt)}}) = \omega^{(opt)} - 1$.

5.2.8 Обусловленность матриц

Несмотря на кажущуюся разумность предположения, что малая относительная погрешность вычислений $\frac{||\boldsymbol{x} - \tilde{\boldsymbol{x}}||}{||\boldsymbol{x}||}$ соответствует малой невязке $||\boldsymbol{r}||$, в реальности соотношение между этими величинами регулируется *числом обусловленности*, которое, как мы скоро убедимся, является важнейшей показателем устойчивости решения СЛАУ к ее малым изменениям. Для начала найдем верхнюю грань для относительной погрешности вычислений через норму невязки.

Теорема 5.2.11. *Пусть $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ является приближением к точному решению СЛАУ $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{A}$ – невырожденная матрица и $\boldsymbol{r}$ – вектор невязки. Тогда для любой индуцированной матричной нормы верно:*

$$||\boldsymbol{x} - \tilde{\boldsymbol{x}}|| \leq ||\boldsymbol{r}|| \cdot ||\boldsymbol{A}^{-1}||, \tag{5.160}$$

и при $\boldsymbol{x}, \boldsymbol{b} \neq \boldsymbol{0}$:

$$\frac{||\boldsymbol{x} - \tilde{\boldsymbol{x}}||}{||\boldsymbol{x}||} \leq K(\boldsymbol{A})\frac{||\boldsymbol{r}||}{||\boldsymbol{b}||}, \tag{5.161}$$

где $K(\boldsymbol{A}) = ||\boldsymbol{A}|| \cdot ||\boldsymbol{A}^{-1}||$ называется числом обусловленности.

Доказательство. Первое неравенство следует из:

$$\begin{aligned} & \boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{A}(\boldsymbol{x} - \tilde{\boldsymbol{x}}) \\ \implies \quad & \boldsymbol{x} - \tilde{\boldsymbol{x}} = \boldsymbol{A}^{-1}\boldsymbol{r} \\ \implies \quad & ||\boldsymbol{x} - \tilde{\boldsymbol{x}}|| \leq ||\boldsymbol{A}^{-1}|| \cdot ||\boldsymbol{r}||. \end{aligned} \tag{5.162}$$

Второе неравенство следует из рассмотрения верхней границы для $\frac{1}{||\boldsymbol{x}||}$:

$$\begin{aligned} & ||\boldsymbol{b}|| = ||\boldsymbol{A}\boldsymbol{x}|| \\ \implies \quad & ||\boldsymbol{b}|| \leq ||\boldsymbol{A}|| \cdot ||\boldsymbol{x}|| \\ \implies \quad & \frac{1}{||\boldsymbol{x}||} \leq \frac{||\boldsymbol{A}||}{||\boldsymbol{b}||}. \end{aligned} \tag{5.163}$$

Тогда домножая первое неравенство на $\frac{1}{||\boldsymbol{x}||}$, получаем:

$$\frac{||\boldsymbol{x} - \tilde{\boldsymbol{x}}||}{||\boldsymbol{x}||} \leq ||\boldsymbol{A}|| \cdot ||\boldsymbol{A}^{-1}|| \frac{||\boldsymbol{r}||}{||\boldsymbol{b}||}. \tag{5.164}$$

□

Для того, чтобы увидеть связь между число обусловленности и устойчивостью решения СЛАУ к малым изменениям ее коэффициентов, представим, что приближенное решение $\tilde{\boldsymbol{x}}$ является точным решением модифицированной СЛАУ, полученной с помощью малых возмущений матрицы $\boldsymbol{A}$:

$$(\boldsymbol{A} + \boldsymbol{\delta A})\tilde{\boldsymbol{x}} = \boldsymbol{b}. \tag{5.165}$$

Матрица возмущений $\boldsymbol{\delta A}$, например, может имитировать погрешность округления, ассоциированную с каждым элементом матрицы $\boldsymbol{A}$. Вектор невязки тогда вычисляется следующим образом:

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}} \\ &= \boldsymbol{\delta A}\tilde{\boldsymbol{x}}, \end{aligned} \tag{5.166}$$

что при подстановке во второе неравенство теоремы 5.2.11 дает:

$$\begin{aligned} \frac{||\boldsymbol{x} - \tilde{\boldsymbol{x}}||}{||\boldsymbol{x}||} &\leq K(\boldsymbol{A}) \frac{||\boldsymbol{\delta A}\tilde{\boldsymbol{x}}||}{||\boldsymbol{b}||} \\ &\leq K(\boldsymbol{A}) ||\boldsymbol{\delta A}|| \frac{||\tilde{\boldsymbol{x}}||}{||\boldsymbol{b}||}. \end{aligned} \tag{5.167}$$

Даже такой грубый пример позволяет заметить, что относительная погрешность вычислений будет тем меньше, чем меньше будет число обусловленности $K(\boldsymbol{A})$ (при этом легко увидеть, что $K(\boldsymbol{A}) \geq 1$). Слишком большое число обусловленности, в свою очередь, приводит к дестабилизации вычислительных погрешностей. Матрицы с малым числом обусловленности называют *хорошо обусловленными*, а матрицы с большим числом обусловленности *плохо обусловленными*.

В свете этих рассуждений попробуем оценить число обусловленности без явного вычисления обратной матрицы, что само по себе было бы источником погрешностей, и найти связь между числом обусловленности и точностью полученного решения в контексте арифметики с t значащими цифрами. Можно доказать, что если приближенное решение $\tilde{\boldsymbol{x}}$ было получено с помощью метода Гаусса и арифметики с t значащими цифрами, то верно следующее приближение для нормы вектора невязки: **[TODO: Forsythe, G. E. and C. B. Moler, Computer solution of linear algebraic systems]**

$$||\boldsymbol{r}|| \approx 10^{-t}||\boldsymbol{A}|| \cdot ||\tilde{\boldsymbol{x}}||. \tag{5.168}$$

Приближение для числа обусловленности теперь можно получить, рассмотрев следующую СЛАУ:

$$\boldsymbol{A}\boldsymbol{y} = \boldsymbol{r}. \tag{5.169}$$

Так как предполагалось, что решение $\tilde{\boldsymbol{x}}$ было получено с помощью метода Гаусса, в наличии гарантированно имеются множители, необходимые для LU-разложения матрицы $\boldsymbol{A}$, что означает, что решение $\boldsymbol{y}$ может быть быстро найдено. Приближенное решение $\boldsymbol{y}$ удовлетворяет:

$$\begin{aligned} \tilde{\boldsymbol{y}} \approx \boldsymbol{A}^{-1}\boldsymbol{r} &= \boldsymbol{A}^{-1}(\boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}}) \\ &= \boldsymbol{A}^{-1}\boldsymbol{b} - \tilde{\boldsymbol{x}} \\ &= \boldsymbol{x} - \tilde{\boldsymbol{x}}. \end{aligned} \tag{5.170}$$

Таким образом вектор $\tilde{\boldsymbol{y}}$ является оценкой вычислительной погрешности, возникшей при нахождении решения $\tilde{\boldsymbol{x}}$. Оценим его норму:

$$\begin{aligned} ||\tilde{\boldsymbol{y}}|| &\approx ||\boldsymbol{A}^{-1}\boldsymbol{r}|| \\ &\leq ||\boldsymbol{A}^{-1}|| \cdot ||\boldsymbol{r}|| \\ &\approx ||\boldsymbol{A}^{-1}||(10^{-t}||\boldsymbol{A}|| \cdot ||\tilde{\boldsymbol{x}}||) \\ &= 10^{-t}K(\boldsymbol{A})||\tilde{\boldsymbol{x}}||. \end{aligned} \tag{5.171}$$

Из полученного результата можно сделать следующий вывод. Если число обусловленности имеет порядок $O(1)$, то вычислительная погрешность приближенного решения $\tilde{\boldsymbol{x}}$ имеет оценочный порядок $O(10^{-t})$, в том время как при $K(\boldsymbol{A}) \propto O(10^t)$ ее оценочный порядок $O(1)$. Иными словами, число обусловленности позволяет оценить, сколько значимых цифр теряется в процессе вычислений:

$$\begin{aligned} & K(\boldsymbol{A}) \approx \frac{||\tilde{\boldsymbol{y}}||}{||\tilde{\boldsymbol{x}}||}10^t \approx \frac{||\boldsymbol{x} - \tilde{\boldsymbol{x}}||}{||\tilde{\boldsymbol{x}}||}10^t \\ \implies \quad & \frac{||\boldsymbol{x} - \tilde{\boldsymbol{x}}||}{||\tilde{\boldsymbol{x}}||} \approx 10^{-t}K(\boldsymbol{A}). \end{aligned} \tag{5.172}$$

На основании этого результата может быть построена техника итерационного уточнения решения $\tilde{\boldsymbol{x}}$. Действительно, если $\tilde{\boldsymbol{y}} \approx \boldsymbol{x} - \tilde{\boldsymbol{x}}$, то логично предположить, что для достаточно обусловленных матриц (а именно $K(\boldsymbol{A}) < 10^t$) вектор $\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{y}}$ будет давать более точное

решение. Обозначив $\tilde{\boldsymbol{x}}^{(1)} = \tilde{\boldsymbol{x}}^{(0)} + \tilde{\boldsymbol{y}}^{(0)}$ и продолжив итерацию уже относительно нового приближенного значения $\tilde{\boldsymbol{x}}^{(1)}$, мы получаем метода итерационного уточнения на основе метода Гаусса. Из (5.172) очевидно, что при $K(\boldsymbol{A}) \propto 10^q$ после k итераций ожидается $k(t-q)$ корректных значащих цифр в решении. Для хорошо обусловленных матриц будет достаточно одной-двух итераций.

5.2.9 Метод сопряженных градиентов

Еще один класс итерационных методов может быть построен при рассмотрении задачи минимизации вектора невязки. Действительно, пусть $\boldsymbol{x}^*$ является решением СЛАУ $\boldsymbol{Ax} = \boldsymbol{b}$, где $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Тогда верным является утверждение:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}}\, \boldsymbol{r}^T\boldsymbol{r} = \underset{\boldsymbol{x}}{\operatorname{argmin}} (\boldsymbol{b} - \boldsymbol{Ax})^T(\boldsymbol{b} - \boldsymbol{Ax}). \tag{5.173}$$

Рассмотрим случай положительно определенной и, следовательно, симметричной матрицы $\boldsymbol{A}$. Легко убедиться, что для симметричной матрицы $\boldsymbol{A}$ верным является утверждение:

$$\langle \boldsymbol{Ax}, \boldsymbol{y}\rangle = \langle \boldsymbol{x}, \boldsymbol{Ay}\rangle, \tag{5.174}$$

где $\langle\cdot,\cdot\rangle$ – скалярное произведение векторов, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ – произвольные вектора. Тогда задача минимизации (5.173) может быть записана следующим образом:

$$\begin{aligned}\underset{\boldsymbol{x}}{\operatorname{argmin}} \left[\langle \boldsymbol{b}, \boldsymbol{b}\rangle - 2\langle \boldsymbol{Ax}, \boldsymbol{b}\rangle + \langle \boldsymbol{Ax}, \boldsymbol{Ax}\rangle\right] &= \underset{\boldsymbol{x}}{\operatorname{argmin}} \left[\langle \boldsymbol{Ax}, \boldsymbol{Ax}\rangle - 2\langle \boldsymbol{Ax}, \boldsymbol{b}\rangle\right] \\ &= \underset{\boldsymbol{x}}{\operatorname{argmin}} \left[\langle \boldsymbol{x}, \boldsymbol{Ax}\rangle - 2\langle \boldsymbol{x}, \boldsymbol{b}\rangle\right].\end{aligned} \tag{5.175}$$

Обозначив целевую функцию как $g(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{Ax}\rangle - 2\langle \boldsymbol{x}, \boldsymbol{b}\rangle$, мы имеем конечную форму задачи минимизации:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}}\, g(\boldsymbol{x}). \tag{5.176}$$

Для нахождения минимума, начав с некоторого приближения $\boldsymbol{x}$, необходимо определить направление поиска $\boldsymbol{v}$ и шаг t. Рассмотрим подобный поиск в контексте функции $g(\boldsymbol{x})$:

$$\begin{aligned} g(\boldsymbol{x} + t\boldsymbol{v}) &= \langle \boldsymbol{x} + t\boldsymbol{v}, \boldsymbol{A}(\boldsymbol{x} + t\boldsymbol{v})\rangle - 2\langle \boldsymbol{x} + t\boldsymbol{v}, \boldsymbol{b}\rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{Ax}\rangle + t\langle \boldsymbol{x}, \boldsymbol{Av}\rangle + t\langle \boldsymbol{v}, \boldsymbol{Ax}\rangle + t^2\langle \boldsymbol{v}, \boldsymbol{Av}\rangle - 2\langle \boldsymbol{x}, \boldsymbol{b}\rangle - 2t\langle \boldsymbol{v}, \boldsymbol{b}\rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{Ax}\rangle - 2\langle \boldsymbol{x}, \boldsymbol{b}\rangle + 2t\langle \boldsymbol{v}, \boldsymbol{Ax}\rangle + t^2\langle \boldsymbol{v}, \boldsymbol{Av}\rangle - 2t\langle \boldsymbol{v}, \boldsymbol{b}\rangle \\ &= g(\boldsymbol{x}) + t^2\langle \boldsymbol{v}, \boldsymbol{Av}\rangle - 2t\langle \boldsymbol{v}, \boldsymbol{b} - \boldsymbol{Ax}\rangle. \end{aligned} \tag{5.177}$$

В первую очередь найдем оптимальный шаг $t^{(opt)}$, минимизирующий функцию $g(\boldsymbol{x} + t\boldsymbol{v})$:

$$\begin{aligned} &\frac{\partial g}{\partial t} = 2t\langle \boldsymbol{v}, \boldsymbol{Av}\rangle - 2\langle \boldsymbol{v}, \boldsymbol{b} - \boldsymbol{Ax}\rangle = 0 \\ \Longrightarrow\quad & t^{(opt)} = \frac{\langle \boldsymbol{v}, \boldsymbol{b} - \boldsymbol{Ax}\rangle}{\langle \boldsymbol{v}, \boldsymbol{Av}\rangle} \\ \Longrightarrow\quad & g(\boldsymbol{x} + t^{(opt)}\boldsymbol{v}) = g(\boldsymbol{x}) - \frac{\langle \boldsymbol{v}, \boldsymbol{b} - \boldsymbol{Ax}\rangle^2}{\langle \boldsymbol{v}, \boldsymbol{Av}\rangle}. \end{aligned} \tag{5.178}$$

Из последнего выражения очевидно, что решение $\boldsymbol{x}^*$ минимизирует $g(\boldsymbol{x})$ для любого направления $\boldsymbol{v} \neq \boldsymbol{0}$. Теперь, задавшись начальным приближением $\boldsymbol{x}^{(0)}$ и начальным направлением поиска $\boldsymbol{v}^{(1)}$, мы можем построить следующий итерационный алгоритм:

$$t_k = \frac{\langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle}, \tag{5.179}$$

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t_k \boldsymbol{v}^{(k)}. \tag{5.180}$$

Следующим шагом является генерация таких направлений поиска $\boldsymbol{v}^{(k)}$, что метод сходится достаточно быстро. Очевидным выбором является направление наискорейшего спуска:

$$-\frac{\partial g}{\partial \boldsymbol{x}} = -2(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = 2\boldsymbol{r}, \tag{5.181}$$

то есть направлением наискорейшего спуска является вектор невязки. Тогда направление поиска определяется как:

$$\boldsymbol{v}^{(k+1)} = \boldsymbol{r}^{(k)}, \tag{5.182}$$

что задает *метод градиентного спуска* для решения СЛАУ. Так как этот метод известен относительно медленной сходимостью, мы воспользуется альтернативным подходом, который базируется на построении таких $\boldsymbol{v}^{(i)}$, что

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{A}\boldsymbol{v}^{(j)} \rangle = 0, i \neq j. \tag{5.183}$$

Такой вид ортогональности векторов мы будем называть A-ортогональностью, а полученную систему векторов $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\}$ A-ортогональной. Эта система задает базис пространства, так как является линейно независимой. Это легко проверить с помощью доказательства от обратного. Действительно, пусть можно выразить вектор $\boldsymbol{v}^{(i)}$ через линейную комбинацию остальных векторов:

$$\begin{aligned} & \boldsymbol{v}^{(i)} = \sum_{\substack{j=1 \\ j \neq i}} \alpha_j \boldsymbol{v}^{(j)} \\ \Longrightarrow \quad & \langle \boldsymbol{v}^{(i)}, \boldsymbol{A}\boldsymbol{v}^{(i)} \rangle = \sum_{\substack{j=1 \\ j \neq i}} \alpha_j \langle \boldsymbol{v}^{(j)}, \boldsymbol{A}\boldsymbol{v}^{(i)} \rangle \\ \Longrightarrow \quad & \langle \boldsymbol{v}^{(i)}, \boldsymbol{A}\boldsymbol{v}^{(i)} \rangle = 0, \end{aligned} \tag{5.184}$$

что нарушает условие A-ортогональности, из чего следует, что система векторов действительно является линейно независимой. Результирующий итерационный метод, построенный на основе уравнений (5.179) и (5.180), называется *методом сопряженных направлений*. Следующая теорема доказывает, что такой метод дает точное решение в контексте арифметики с бесконечной точностью за n итераций.

Теорема 5.2.12. *Пусть система векторов $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\}$ является A-ортогональной, а матрица $\boldsymbol{A}$ положительно определенной. Тогда $\boldsymbol{A}\boldsymbol{x}^{(n)} = \boldsymbol{b}$, где $\boldsymbol{x}^{(n)}$ определяется с помо-*

щью итерационного алгоритма ниже:

$$t_k = \frac{\langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle}, \tag{5.185}$$

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t_k \boldsymbol{v}^{(k)}. \tag{5.186}$$

Доказательство. Раскроем итерации в выражении $\boldsymbol{A}\boldsymbol{x}^{(n)}$ вплоть до начального приближения $\boldsymbol{x}^{(0)}$:

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x}^{(n)} &= \boldsymbol{A}\left(\boldsymbol{x}^{(n-1)} + t_n \boldsymbol{v}^{(n)}\right), \\ &= \boldsymbol{A}\left(\boldsymbol{x}^{(n-2)} + t_{n-1}\boldsymbol{v}^{(n-1)} + t_n \boldsymbol{v}^{(n)}\right), \\ &= \dots \\ &= \boldsymbol{A}\left(\boldsymbol{x}^{(0)} + t_1 \boldsymbol{v}^{(1)} + \dots + t_n \boldsymbol{v}^{(n)}\right), \end{aligned} \tag{5.187}$$

из чего следует:

$$\boldsymbol{A}\boldsymbol{x}^{(n)} - \boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^{(0)} - \boldsymbol{b} + \boldsymbol{A}\left(t_1 \boldsymbol{v}^{(1)} + \dots + t_n \boldsymbol{v}^{(n)}\right). \tag{5.188}$$

Посчитаем скалярное произведение обоих сторон уравнения, домножив его на $\boldsymbol{v}^{(k)}$:

$$\langle \boldsymbol{A}\boldsymbol{x}^{(n)} - \boldsymbol{b}, \boldsymbol{v}^{(k)} \rangle = \langle \boldsymbol{A}\boldsymbol{x}^{(0)} - \boldsymbol{b}, \boldsymbol{v}^{(k)} \rangle + t_k \langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle. \tag{5.189}$$

Из определения t_k можно получить выражение:

$$\begin{aligned} t_k \langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle &= \langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k-1)} \rangle \\ &= \langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)} + \boldsymbol{A}\boldsymbol{x}^{(0)} - \boldsymbol{A}\boldsymbol{x}^{(1)} + \boldsymbol{A}\boldsymbol{x}^{(1)} - \dots - \boldsymbol{A}\boldsymbol{x}^{(k-2)} + \boldsymbol{A}\boldsymbol{x}^{(k-2)} - \boldsymbol{A}\boldsymbol{x}^{(k-1)} \rangle \\ &= \langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)} \rangle + \langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{x}^{(0)} - \boldsymbol{A}\boldsymbol{x}^{(1)} \rangle + \dots + \langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{x}^{(k-2)} - \boldsymbol{A}\boldsymbol{x}^{(k-1)} \rangle. \end{aligned} \tag{5.190}$$

Заметим, что из $\boldsymbol{x}^{(i)} = \boldsymbol{x}^{(i-1)} + t_i \boldsymbol{v}^{(i)}$ следует $\boldsymbol{A}\boldsymbol{x}^{(i-1)} - \boldsymbol{A}\boldsymbol{x}^{(i)} = -t_i \boldsymbol{A}\boldsymbol{v}^{(i)}$. Тогда подстановка в выражение выше дает:

$$t_k \langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle = \langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)} \rangle, \tag{5.191}$$

из чего после подстановки в (5.189) мы получаем:

$$\begin{aligned} & \langle \boldsymbol{A}\boldsymbol{x}^{(n)} - \boldsymbol{b}, \boldsymbol{v}^{(k)} \rangle = \langle \boldsymbol{A}\boldsymbol{x}^{(0)} - \boldsymbol{b}, \boldsymbol{v}^{(k)} \rangle + \langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)} \rangle \\ \implies \quad & \langle \boldsymbol{A}\boldsymbol{x}^{(n)} - \boldsymbol{b}, \boldsymbol{v}^{(k)} \rangle = 0 \quad \forall k = 1, \dots, n. \end{aligned} \tag{5.192}$$

Последнее выражение говорит о том, что вектор невязки на n-й итерации ортогонален всем векторам $\{\boldsymbol{v}^{(1)}, \dots, \boldsymbol{v}^{(n)}\}$, которые, будучи линейно независимыми, задают базис пространства. Это возможно тогда и только тогда, когда $\boldsymbol{A}\boldsymbol{x}^{(n)} - \boldsymbol{b} = 0$, что означает, что $\boldsymbol{x}^{(n)}$ является решением исходной СЛАУ. □

Также можно доказать, что получаемые векторы невязки $\boldsymbol{r}^{(k)}$ будут ортогональны векторам $\boldsymbol{v}^{(i)}$, $i = 1, \ldots, k$. *Метод сопряженных градиентов* выбирает такие $\{\boldsymbol{v}^{(i)}\}$, что система, состоящая из векторов невязки $\{\boldsymbol{r}^{(i)}\}$, является ортогональной, т.е. $\langle \boldsymbol{r}^{(i)}, \boldsymbol{r}^{(j)} \rangle$ для $i \neq j$. Соответствующие формулы для $\{\boldsymbol{v}^{(i)}\}$ можно построить, адаптировав процесс Грама-Шмидта для A-ортогональности. Действительно, пусть начальное направление является направлением наискорейшего спуска, т.е. $\boldsymbol{v}^{(1)} = \boldsymbol{r}^{(0)}$. Предположим, что мы уже рассчитали приближения $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k-1)}\}$ и направления $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(k-1)}\}$. Применим процесс Грама-Шмидт для нахождения $\boldsymbol{v}^{(k)}$:

$$\boldsymbol{v}^{(k)} = \boldsymbol{r}^{k-1} + s_{k-1}\boldsymbol{v}^{(k-1)}, \tag{5.193}$$

где коэффициент s_{k-1} мы находим исходя из условия A-ортогональности $\langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k-1)} \rangle = 0$:

$$\begin{aligned} & 0 = \langle \boldsymbol{r}^{k-1}, \boldsymbol{A}\boldsymbol{v}^{(k-1)} \rangle + s_{k-1} \langle \boldsymbol{v}^{(k-1)}, \boldsymbol{A}\boldsymbol{v}^{(k-1)} \rangle \\ \implies \quad & s_{k-1} = -\frac{\langle \boldsymbol{r}^{(k-1)}, \boldsymbol{A}\boldsymbol{v}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k-1)}, \boldsymbol{A}\boldsymbol{v}^{(k-1)} \rangle}. \end{aligned} \tag{5.194}$$

Имея выражение для $\boldsymbol{v}^{(k)}$, переформулируем t_k:

$$\begin{aligned} t_k &= \frac{\langle \boldsymbol{v}^{(k)}, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle} \\ &= \frac{\langle \boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle} + s_{k-1} \frac{\langle \boldsymbol{v}^{(k-1)}, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle} \\ &= \frac{\langle \boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle}, \end{aligned} \tag{5.195}$$

где мы использовали свойство $\langle \boldsymbol{v}^{(k-1)}, \boldsymbol{r}^{(k-1)} \rangle = 0$. Мы также можем записать рекурсивное выражение для $\boldsymbol{r}^{(k)}$:

$$\begin{aligned} & \boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t_k \boldsymbol{v}^{(k)} \\ \implies \quad & \boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^{(k-1)} - \boldsymbol{b} + t_k \boldsymbol{A}\boldsymbol{v}^{(k)} \\ \implies \quad & \boldsymbol{r}^{(k)} = \boldsymbol{r}^{(k-1)} - t_k \boldsymbol{A}\boldsymbol{v}^{(k)}. \end{aligned} \tag{5.196}$$

Требование ортогональности $\boldsymbol{r}^{(k)}$ и $\boldsymbol{r}^{(k-1)}$ при этом дает

$$\langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle = \frac{1}{t_k} \langle \boldsymbol{r}^{(k-1)} \boldsymbol{r}^{(k-1)} \rangle. \tag{5.197}$$

Это позволяет упросить выражение для s_k, если мы заметим что:

$$\langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{r}^{(k)} \rangle = \langle \boldsymbol{A}\boldsymbol{v}^{(k)}, \boldsymbol{r}^{(k)} \rangle = \frac{1}{t_k} \langle \boldsymbol{r}^{(k-1)} - \boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)} \rangle = -\frac{1}{t_k} \langle \boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)} \rangle, \tag{5.198}$$

Тогда выражение для s_k принимает вид:

$$s_k = \frac{\langle \boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)} \rangle}{\langle \boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)} \rangle}. \tag{5.199}$$

Резюмируем метод сопряженных градиентов в виде следующих рекурсивных уравнений:

$$t_k = \frac{\langle \boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)}, \boldsymbol{A}\boldsymbol{v}^{(k)} \rangle}, \tag{5.200}$$

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t_k \boldsymbol{v}^{(k)}, \tag{5.201}$$

$$\boldsymbol{r}^{(k)} = \boldsymbol{r}^{(k-1)} - t_k \boldsymbol{A}\boldsymbol{v}^{(k)}, \tag{5.202}$$

$$s_k = \frac{\langle \boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)} \rangle}{\langle \boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)} \rangle}, \tag{5.203}$$

$$\boldsymbol{v}^{(k+1)} = \boldsymbol{r}^{(k)} + s_k \boldsymbol{v}^{(k)}, \tag{5.204}$$

где $\boldsymbol{x}^{(0)}$, $\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$ и $\boldsymbol{v}^{(1)} = \boldsymbol{r}^{(0)}$ являются начальными условиями.

Важно отметить, что сопряженные направления напрямую связаны с подпространством Крылова.

Определение 5.2.6. *Подпространством Крылова размерности n называется линейное пространство $\mathcal{K}_m(\boldsymbol{z}, \boldsymbol{A})$, порожденное множеством векторов $\{\boldsymbol{z}, \boldsymbol{A}\boldsymbol{z}, \boldsymbol{A}^2\boldsymbol{z}, \ldots, \boldsymbol{A}^{m-1}\boldsymbol{z}\}$, где $\boldsymbol{z} \in \mathbb{R}^n$ и $\boldsymbol{A} \in \mathbb{R}^{n \times n}$.*

Рассмотрим выражения для сопряженных направлений:

$$\begin{aligned}
&\boldsymbol{v}^{(1)} = \boldsymbol{r}^{(0)}, \\
&\boldsymbol{v}^{(2)} = -t_1 \boldsymbol{A}\boldsymbol{r}^{(0)} + (1 + s_1)\boldsymbol{r}^{(0)}, \\
&\boldsymbol{v}^{(3)} = \boldsymbol{r}^{(0)} - [t_1 + t_2(1 + s_1)]\, \boldsymbol{A}\boldsymbol{r}^{(0)} + t_1 t_2 \boldsymbol{A}^2 \boldsymbol{r}^{(0)}, \\
&\quad \ldots
\end{aligned}$$

Можно заметить, что сопряженные направления $\boldsymbol{v}^k$ принадлежат подпространству Крылова $\mathcal{K}_n(\boldsymbol{r}^{(0)}, \boldsymbol{A})$. Более того, несложно продемонстрировать, что они составляют ортонормальный базис подпространства Крылова $\mathcal{K}_n(\boldsymbol{r}^{(0)}, \boldsymbol{A})$.

5.2.10 Предобуславливание матриц

В случае, когда матрица $\boldsymbol{A}$ является плохо обусловленной (т.е. имеет большое число обусловленности $K(\boldsymbol{A})$), вычислительная погрешность, как мы выяснили, становится достаточно большой и значимой. Более того, плохо обусловленные матрицы обладают медленной сходимостью. Выходом из этой ситуации является модификация исходной матрицы $\boldsymbol{A}$ так, что полученная в результате матрица обладает значительно меньшим числом обусловленности. Матрица, используемая для модификации исходной матрицы, называется *матрицей предобуславливания*. Для положительно определенных матрицы такая модификация должна иметь вид:

$$\tilde{\boldsymbol{A}} = \boldsymbol{C}^{-1} \boldsymbol{A} \left(\boldsymbol{C}^{-1}\right)^T, \tag{5.205}$$

где обозначение $\boldsymbol{C}^{-1}$ говорит о том, что *матрица предобуславливания* $\boldsymbol{C}^{-1}$ должна быть невырожденной. При подобном предобуславливании полученная матрица $\tilde{\boldsymbol{A}}$ сохраняет свойство положительной определенности. Одним из вариантов матриц $\boldsymbol{C}^{-1}$ является следующее

предобуславливание:

$$\tilde{\boldsymbol{A}} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\left(\boldsymbol{D}^{-1/2}\right)^T, \tag{5.206}$$

где $\boldsymbol{D}^{-1/2}$ состоит из корней от обратных диагональных элементов матрицы $\boldsymbol{A}$. Метод сопряженных градиентов, сформулированный для модифицированной матрицы, называется *методом сопряженных градиентов с предобуславливанием*.

Численные методы нелинейной алгебры

Несмотря на широкую распространенность линейных систем уравнений и соответствующих методов их решения, мир по своей природе остается нелинейным. Этот факт остроумно резюмирован польским математиком Станиславом Уламом в афоризме «использование термина “нелинейная динамика” эквивалентно отношению к зоологии как к науке о неслонах». Действительно, большинство прикладных задач, решаемых в математической физике и машинном обучении, являются изначально нелинейными и становятся линейными лишь в процессе тех или иных упрощений. Когда эти упрощения перестают быть валидными, нам необходимо решать полностью нелинейную задачу. Первым шагом на пути к решению нелинейных задач является изучение методов нахождения корней нелинейных систем алгебраических уравнений. В отличие от линейных систем, нелинейные системы в общем случае не могут быть решены прямыми методами, поэтому целью этой главы является рассмотрение исключительно итерационных методов. Также отметим, что доказательства для многомерных пространств могут быть технически сложны, так что мы будем демонстрировать доказательства теорем для одномерным пространств (т.е. для системы из одного нелинейного алгебраического уравнения), после чего обобщим их утверждения до многомерного случая без доказательства.

В общем виде система нелинейных алгебраических уравнений имеет вид:

$$\begin{aligned} &f_1(x_1, \dots, x_n) = 0, \\ &f_2(x_1, \dots, x_n) = 0, \\ &\dots \\ &f_n(x_1, \dots, x_n) = 0. \end{aligned} \tag{6.1}$$

Такую систему удобно представить в виде векторной функции $\boldsymbol{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$:

$$\boldsymbol{f}(\boldsymbol{x}) : \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \mapsto \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_n(x_1, \dots, x_n) \end{bmatrix}. \tag{6.2}$$

Тогда система принимает вид:

$$\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0} \tag{6.3}$$

6.1 Метод простой итерации

Мы уже рассмотрели метод простой итерации для линейных систем. Формулировка этого метода для нелинейных систем строится вокруг понятия неподвижной точки.

Определение 6.1.1. *Вектор $\boldsymbol{x^*} \in \mathbb{R}^n$ называется неподвижной точкой векторной функции $\boldsymbol{g}(\boldsymbol{x})$, если $\boldsymbol{g}(\boldsymbol{x^*}) = \boldsymbol{x^*}$.*

Систему вида (6.3) можно легко свести к задаче о поиске неподвижной точки, рассмотрев следующую функцию:

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{\Omega}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}), \tag{6.4}$$

где невырожденная матрица $\boldsymbol{\Omega}(\boldsymbol{x}) : \mathbb{R}^n \in \mathbb{R}^{n\times n}$ выступает в качестве аналога параметра релаксации в линейных системах. В таком случае множество неподвижных точек функции $\boldsymbol{g}(\boldsymbol{x})$ совпадает с множеством корней функции $\boldsymbol{f}(\boldsymbol{x})$.

Докажем теорему о достаточном условии существования и единственности неподвижной точки для одномерного случая.

Теорема 6.1.1. *Пусть $D = [a; b]$. Если $g \in C(D)$ и $g(D) \subset D$, то g имеет хотя бы одну неподвижную точку в D. Если при этом существует производная $g'(x)$ в $(a; b)$ и $|g'(x)| < 1$ для любого $x \in (a; b)$, то эта неподвижная точка является единственной.*

Доказательство. Случаи $g(a) = a, g(b) = b$ очевидны, так что, с учетом $g(D) \subset D$, мы рассмотрим $g(a) > a, g(b) < b$. Определим вспомогательную функцию $h(x)$ следующим образом: $h(x) = g(x) - x$. Для нее является справедливым следующее:

$$h(a) = g(a) - a > 0, \tag{6.5}$$
$$h(b) = g(b) - b < 0. \tag{6.6}$$

Следовательно, по теореме о промежуточном значении можно найти такое $x^* \in D$, что $h(x^*) = 0$, то есть $g(x^*) = x^*$.

Доказажем вторую часть теоремы от обратного. Предположим, что существуют две неподвижных точки $x^* \neq y^*$. Тогда по теореме Лагранжа о среднем значении существует такое $\xi \in [x^*; y^*] \subset D$, что:

$$\begin{aligned} g'(\xi) &= \frac{g(x^*) - g(y^*)}{x^* - y^*} \\ &= \frac{x^* - y^*}{x^* - y^*} \\ &= 1, \end{aligned} \tag{6.7}$$

что противоречит установке $|g'(x)| < 1$. Следовательно, существующая неподвижная точка является единственной. □

Как и в случае с линейными системами, неподвижная точка находится с помощью метода простых итераций:

$$\boldsymbol{x}^{(k)} = \boldsymbol{g}(\boldsymbol{x}^{(k-1)}), \tag{6.8}$$

который дает последовательность точек $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$. Следующая теорема задает условие, при котором эта последовательность будет сходиться к неподвижной точке $\boldsymbol{x}$. Рассмотрим ее для одномерного случая.

Теорема 6.1.2. *Пусть $D = [a; b]$, $g \in C(D)$, $g(D) \subset D$ и существует производная $g'(x)$ для $x \in (a; b)$ с таким $\gamma \in (0; 1)$, что $|g'(x)| \leq \gamma$ для любого $x \in (a; b)$. Тогда для любого $x^{(0)} \in D$ последовательность $\{x^{(k)}\}_{k=0}^{\infty}$, сгенерированная итерацией $x^{(k)} = g(x^{(k-1)})$, сходится к единственной неподвижной точке $x^* \in [a; b]$.*

Доказательство. Так как $|g'(x)| < 1$ для любого $x \in (a; b)$, по теореме 6.1.1 существует единственная неподвижная точка в D. По теореме Лагранжа о среднем значении для любого $k > 1$ существует такое $\xi^{(k)} \in D$, что:

$$\begin{aligned} & g'(\xi^{(k)}) = \frac{g(x^{(k-1)}) - g(x^*)}{x^{(k-1)} - x^*} = \frac{x^{(k)} - x^*}{x^{(k-1)} - x^*} \\ \Longrightarrow \quad & |x^{(k)} - x^*| = |g'(\xi^{(k)})| \cdot |x^{(k-1)} - x^*| \\ \Longrightarrow \quad & |x^{(k)} - x^*| \leq \gamma |x^{(k-1)} - x^*|. \end{aligned} \tag{6.9}$$

Применив тот же подход для нахождения верхней границы для $|x^{(k-1)} - x^*|$, $|x^{(k-2)} - x^*|$ и т.д., получаем:

$$\begin{aligned} |x^{(k)} - x^*| &\leq \gamma |x^{(k-1)} - x^*| \\ &\leq \gamma^2 |x^{(k-2)} - x^*| \\ &\dots \\ &\leq \gamma^k |x^{(0)} - x^*|. \end{aligned} \tag{6.10}$$

Тогда предел $\lim_{k\to\infty} |x^{(k)} - x^*|$ принимает вид:

$$\lim_{k\to\infty} |x^{(k)} - x^*| \leq \lim_{k\to\infty} \gamma^k |x^{(0)} - x^*| = 0, \tag{6.11}$$

что эквивалентно сходимости метода простой итерации. □

Следующее следствие дает оценки погрешности метода простой итерации. Мы оставляем его без доказательства, так как оно абсолютно аналогично выводу оценок погрешностей метода простой итерации для случая СЛАУ.

Следствие 6.1.1. *Пусть функция $g(x)$ удовлетворяет требованиями теоремы 6.1.2. Тогда верными являются следующие неравенства:*

$$|x^{(k)} - x^*| \leq \gamma^k \max\{x^{(0)} - a, x^{(0)} - b\}, \tag{6.12}$$

$$|x^{(k)} - x^*| \leq \frac{\gamma^k}{1 - \gamma} |x^{(1)} - x^{(0)}|, \tag{6.13}$$

для любых $k \geq 1$.

Теперь приведем утверждения для этих же теорем для n-мерного случая без доказательства.

Теорема 6.1.3. *Пусть* $D = \{(x_1, \ldots, x_n) | x_i \in [a_i; b_i]$ *для* $i = 1, \ldots, n\}$. *Если* $\boldsymbol{g} \in C(D)$ *и* $\boldsymbol{g}(D) \subset D$, *то* $\boldsymbol{g}$ *имеет хотя бы одну неподвижную точку в* D. *Если при этом существует матрица Якоби* $\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}$ *в* $\widetilde{D} = \{(x_1, \ldots, x_n) | x_i \in (a_i; b_i)$ *для* $i = 1, \ldots, n\}$ *и такое число* $\gamma < 1$, *что для любого* $\boldsymbol{x} \in \widetilde{D}$ *верны следующие неравенства:*

$$\left| \frac{\partial g_i(\boldsymbol{x})}{\partial x_j} \right| \leq \frac{\gamma}{n}, \quad i, j = 1, \ldots, n, \tag{6.14}$$

то эта неподвижная точка является единственной.

Теорема 6.1.4. *Пусть верны условия теоремы 6.1.3. Тогда для любого* $\boldsymbol{x}^{(0)} \in D$ *последовательность* $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$, *сгенерированная итерацией* $\boldsymbol{x}^{(k)} = \boldsymbol{g}(\boldsymbol{x}^{(k-1)})$, *сходится к единственной неподвижной точке* $\boldsymbol{x}^* \in D$.

Следствие 6.1.2. *Пусть верны условия теоремы 6.1.4. Тогда верным является следующее неравенство:*

$$||\boldsymbol{x}^{(k)} - \boldsymbol{x}^*||_\infty \leq \frac{\gamma^k}{1 - \gamma} ||\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}||_\infty, \tag{6.15}$$

для любых $k \geq 1$.

Сформулированные теоремы позволяют однозначно установить сходимость метода простой итерации и оценить скорость сходимости. Таким образом, для решения практических задач достаточно сформировать функцию $\boldsymbol{g}(\boldsymbol{x})$ по формуле (6.4) и убедиться, что ее производные удовлетворяют заданному критерию сходимости. Скорость сходимости можно увеличить, если использовать подход, аналогичный подходу в методе Гаусса-Зейделя: при расчете x_i^k вместо $x_1^{k-1}, \ldots, x_{i-1}^{k-1}$ использовать уже посчитанные $x_1^k, \ldots, x_{i-1}^k$.

6.2 Метод Ньютона

Метод Ньютона является одним из самых распространенных методов решения систем нелинейных алгебраических уравнений. Для начала мы рассмотрим классический вывод метода Ньютона через разложение в ряд Тейлора, после чего взглянем на метод Ньютона как на частный случай метода простой итерации. Как и раньше, мы в деталях рассмотрим одномерный случай, после чего обобщим результаты для многомерного случая.

Пусть нам дана функция $f(x) \in C^2[a; b]$ и точка $x^{(0)} \in [a; b]$ является приближением к корню x^* функции $f(x)$, т.е. $f(x^*) = 0$. Разложим функцию $f(x)$ в ряд Тейлора в точке $x^{(0)}$:

$$f(x) = f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)}) + \frac{f''(\xi)}{2}(x - x^{(0)})^2, \tag{6.16}$$

где $\xi \in (x; x^{(0)})$. Метод Ньютона основан на предположении, что дистанция $|x^* - x^{(0)}|$ от приближения до корня функции достаточно мала. Тогда, вычислив ряд в точке x^*, мы можем отбросить квадратичный член:

$$f(x^*) \approx f(x^{(0)}) + f'(x^{(0)})(x^* - x^{(0)}). \tag{6.17}$$

Так как $f(x^*) = 0$, мы получаем оценку для x^*, т.е. более точное приближение:

$$\begin{aligned} & 0 \approx f(x^{(0)}) + f'(x^{(0)})(x^* - x^{(0)}) \\ \Longrightarrow \quad & x^* \approx x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})}. \end{aligned} \tag{6.18}$$

Таким образом, можно сформулировать следующий итерационный метод, называемый *методом Ньютона*:

$$x^{(k)} = x^{(k-1)} - \frac{f(x^{(k-1)})}{f'(x^{(k-1)})}. \tag{6.19}$$

Несложно заметить, что полученный итерационный метод является формой метода простой итерации:

$$x^{(k)} = g(x^{(k-1)}), \tag{6.20}$$

где $g(x) = x - \frac{f(x)}{f'(x)}$. Сравнение полученной функции $g(x)$ с общим выражением (6.4) показывает, что метода Ньютона выбирает в качестве функции $\Omega(x) = -\frac{1}{f'(x)}$.

Отметим две важных особенности метода Ньютона:

1. начальное приближение $x^{(0)}$ должно быть достаточно близким к корню x^*;
2. градиент $f'(x^{(k)})$ должен быть отличен от нуля для любого k. Более того, метод Ньютона тем эффективнее, чем дальше $f'(x^*)$ от нуля.

Первое условие накладывает серьезные ограничения на сходимость метода Ньютона для произвольных $x^{(0)}$. Следующая теорема доказывает существование некоторой окрестности x^*, внутри которой сходимость метода Ньютона является безусловной.

Теорема 6.2.1. *Пусть $f(x) \in C^2[a;b]$ и существует такое $x^* \in (a;b)$, что $f(x^*) = 0$ и $f'(x^*) \neq 0$. Тогда существует такое $\delta > 0$, что последовательность $\{x^{(k)}\}_{k=0}^{\infty}$, генерируемая метода Ньютона, сходится к x^* для любого $x^{(0)} \in [x^* - \delta; x^* + \delta]$.*

Доказательство. Рассмотрим метода Ньютона как метод простой итерации в соответствии с (6.20). Тогда, следуя теореме 6.1.2, необходимо найти интервал $D = [x^* - \delta; x^* + \delta]$, который функция $g(x)$ отображает в себя, и в котором $|g'(x)| \leq \gamma$, где $\gamma \in (0;1)$.

Так как f' является непрерывной и $f'(x^*) \neq 0$, существует такое $\delta_1 > 0$, что $f'(x) \neq 0$ в замкнутой δ_1-окрестности точки x^*, т.е. для $x \in [x^* - \delta_1; x^* + \delta_1] \subset [a;b]$. Тогда непрерывная производная $g'(x) \in C^1[a;b]$ также будет существовать в этом интервале:

$$\begin{aligned} g'(x) &= 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} \\ &= \frac{f(x)f''(x)}{(f'(x))^2}. \end{aligned} \tag{6.21}$$

Найдем значение $g'(x)$ в точке x^*:

$$\begin{aligned} g'(x^*) &= \frac{f(x^*)f''(x^*)}{(f'(x^*))^2} \\ &= 0. \end{aligned} \tag{6.22}$$

Тогда существует такое $\delta \in (0;\delta_1)$, что $|g'(x)| \leq \gamma$ для $\gamma \in (0;1)$ в замкнутой δ-окрестности точки x^*, т.е. для $x \in [x^* - \delta; x^* + \delta] \subset [a;b]$.

Теперь докажем, что полученный интервал $D = [x^* - \delta; x^* + \delta]$ отображается функцией $g(x)$ сам в себя. По теореме Лагранжа о среднем значении для $x \in D$ существует такое $\xi \in (x;x^*)$ или $\xi \in (x^*;x)$, что:

$$\begin{aligned} & |g'(\xi)| = \frac{|g(x) - g(x^*)|}{|x - x^*|} = \frac{|g(x) - x^*|}{|x - x^*|} \\ \Longrightarrow \quad & |g(x) - x^*| = |g'(\xi)| \cdot |x - x^*| \leq \gamma |x - x^*| < |x - x^*|. \end{aligned} \tag{6.23}$$

Так как $|x - x^*| \leq \delta$, мы получаем $|g(x) - x^*| \leq \delta$, из чего следует, что $g(x)$ отображает D в себя. Тогда по теореме 6.1.2 последовательность $\{x^{(k)}\}_{k=0}^{\infty}$ сходится к x^* для любого $x^{(0)} \in [x^* - \delta; x^* + \delta]$. □

Прежде чем перейти к рассмотрению скорости сходимости метода Ньютона, дадим строгое определение этому термину.

Определение 6.2.1. *Пусть последовательность $\{x^{(k)}\}_{k=0}^{\infty}$ сходится к x^*. Тогда если существуют такие $\lambda, \alpha \in \mathbb{R}$, что:*

$$\lim_{k \to \infty} \frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|^{\alpha}} = \lambda, \tag{6.24}$$

то метод, генерирующий данную последовательность, обладает сходимостью степени α. При $\alpha = 1, \lambda \in (0;1)$ сходимость называется линейной, при $\alpha = 1, \lambda = 0$ сверхлинейной и при $\alpha = 2$ квадратичной.

Следующая теорема доказывает, что сходящийся метод простой итерации в общем случае сходится только линейно.

Теорема 6.2.2. *Пусть $D = [a;b]$, $g \in C(D)$, $g(D) \subset D$ и существует производная $g'(x)$ для $x \in (a;b)$ с таким $\gamma \in (0;1)$, что $|g'(x)| \leq \gamma$ для любого $x \in (a;b)$. Если $g'(x^*) \neq 0$, то для любого $x^{(0)} \in D$ последовательность $\{x^{(k)}\}_{k=0}^{\infty}$, сгенерированная итерацией $x^{(k)} = g(x^{(k-1)})$, сходится линейно к единственной неподвижной точке $x^* \in [a;b]$.*

Доказательство. По теореме 6.1.2 данная последовательность действительно сходится к x^*, так что необходимо доказать только факт линейной сходимости. В силу непрерывности $g'(x)$ по теореме Лагранжа о среднем значении для любого $k > 0$ существует такое $\xi \in (x^{(k)};x^*)$ или $\xi_k \in (x^*;x^{(k)})$, что:

$$\begin{aligned} g'(\xi_k) &= \frac{g(x^{(k)}) - g(x^*)}{x^{(k)} - x^*} \\ &= \frac{x^{(k+1)} - x^*}{x^{(k)} - x^*}. \end{aligned} \tag{6.25}$$

Так как $\{x^{(k)}\}_{k=0}^{\infty}$ сходится к x^*, к x^* сходится и $\{\xi^{(k)}\}_{k=0}^{\infty}$. Более того, благодаря непрерывности $g'(x)$, мы имеем

$$\begin{aligned} & \lim_{k\to\infty} g'(\xi_k) = g'(x^*) \\ \Longrightarrow \quad & \lim_{k\to\infty} \frac{x^{(k+1)} - x^*}{x^{(k)} - x^*} = g'(x^*) \\ \Longrightarrow \quad & \lim_{k\to\infty} \frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|} = |g'(x^*)|. \end{aligned} \tag{6.26}$$

Тогда, так как $0 < |g'(x^*)| < 1$, последовательность $\{x^{(k)}\}_{k=0}^{\infty}$ сходится к x^* линейно. □

Рассмотренная теорема предполагает, что сходимость быстрее линейной возможна только при $g'(x^*) = 0$. Действительно, как доказывает следующая теорема, сходимость в таком случае становится как минимум квадратичной.

Теорема 6.2.3. *Пусть $x^* \in (a; b)$ является неподвижной точкой функции $g(x)$, т.е. $x^* = g(x^*)$. Пусть также $g'(x^*) = 0$ и $|g''(x)| < M$ для $x \in (a; b)$. Тогда существует такое $\delta > 0$, что последовательность $\{x^{(k)}\}_{k=0}^{\infty}$, генерируемая итерацией $x^{(k)} = g(x^{(k-1)})$, $k > 1$, сходится к x^* в области $[x^* - \delta; x^* + \delta]$ как минимум квадратически. Более того, для асимптотически больших k, верно следующее неравенство:*

$$|x^{(k+1)} - x^*| < \frac{M}{2}|x^{(k)} - x^*|^2. \tag{6.27}$$

Доказательство. По аналогии с доказательством теоремы 6.2.1 выберем такое $\delta > 0$ и $\gamma \in (0; 1)$, что $|g'(x)| \le \gamma$ для $x \in [x^* - \delta; x^* + \delta] \subset [a; b]$. В том же доказательстве мы установили, что последовательность $\{x^{(k)}\}_{k=0}^{\infty}$ для $x^{(0)} \in [x^* - \delta; x^* + \delta]$ будем так же содержаться в $[x^* - \delta; x^* + \delta]$. Разложим $g(x)$ в ряд Тейлора в точке x^*:

$$g(x) = g(x^*) + g'(x^*)(x - x^*) + \frac{g''(\xi)}{2}(x - x^*)^2 \tag{6.28}$$

где $x \in [x^* - \delta; x^* + \delta]$ и, следовательно, $\xi \in [x^*; x]$ или $\xi \in [x; x^*]$. Используя установки теоремы, а именно $g(x^*) = x^*$ и $g'(x^*) = 0$, мы имеем:

$$g(x) = x^* + \frac{g''(\xi)}{2}(x - x^*)^2. \tag{6.29}$$

Тогда для $x = x^{(k)}$ мы получаем:

$$\begin{aligned} & g(x^{(k)}) = x^* + \frac{g''(\xi_k)}{2}(x^{(k)} - x^*)^2 \\ \Longrightarrow \quad & x^{(k+1)} - x^* = \frac{g''(\xi_k)}{2}(x^{(k)} - x^*)^2. \end{aligned} \tag{6.30}$$

Так как $|g'(x)| \leq \gamma$ для $x \in [x^* - \delta; x^* + \delta]$, последовательность $\{x^{(k)}\}_{k=0}^{\infty}$ сходится к $x^* \Longrightarrow$ $\{\xi^{(k)}\}_{k=0}^{\infty}$ сходится к $x^* \Longrightarrow \lim_{k\to\infty} g''(\xi_k) = g''(x^*)$. Тогда имеем:

$$\begin{aligned} & \lim_{k\to\infty} g''(\xi_k) = g''(x^*) \\ \Longrightarrow \quad & \lim_{k\to\infty} \frac{x^{(k+1)} - x^*}{(x^{(k)} - x^*)^2} = \frac{g''(x^*)}{2} \\ \Longrightarrow \quad & \lim_{k\to\infty} \frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|^2} = \frac{|g''(x^*)|}{2}, \end{aligned} \tag{6.31}$$

что доказывает квадратичную сходимость. Так как $|g''(x^*)| < M$, мы также получаем следующую оценку:

$$|x^{(k+1)} - x^*| < \frac{M}{2}|x^{(k)} - x^*|^2. \tag{6.32}$$

□

Теорема 6.2.3 дает нам возможность построить такую функцию $g(x)$ в соответствии с определением (6.4) для нахождения корня функции $f(x)$, что метод простой итерации будет сходиться квадратично. Как мы убедились, для этого необходимо, чтобы $g'(x^*) = 0$. Тогда по формуле (6.4) для одномерного случая имеем:

$$\begin{aligned} & g'(x^*) = 0 \\ \Longrightarrow \quad & 1 - \Omega'(x^*)f(x^*) - \Omega(x^*)f'(x^*) = 0 \\ \Longrightarrow \quad & 1 - \Omega(x^*)f'(x^*) = 0 \\ \Longrightarrow \quad & \Omega(x^*) = \frac{1}{f'(x^*)}. \end{aligned} \tag{6.33}$$

Тогда, выбрав $\Omega(x) = \frac{1}{f'(x)}$, мы получаем уже рассмотренный метод Ньютона:

$$x^{(k)} = g(x^{(k)}) = x^{(k-1)} - \frac{f(x^{(k-1)})}{f'(x^{(k-1)})}. \tag{6.34}$$

Таким образом, метод Ньютона является "оптимальным" методом простой итерации и имеет квадратичную сходимость.

Рассмотрев одномерный случай, мы можем обобщить результаты на многомерный случай. Мы сразу обратимся к методу Ньютона в форме метода простой итерации и сформулируем без доказательства теорему о методе простой итерации с квадратичной сходимостью.

Теорема 6.2.4. *Пусть $\boldsymbol{x^*}$ является неподвижной точкой функции $\boldsymbol{g}(\boldsymbol{x})$, т.е. $\boldsymbol{g}(\boldsymbol{x^*}) = \boldsymbol{x}^*$, и пусть функция $\boldsymbol{g}(\boldsymbol{x})$ удовлетворяет следующим условиям:*

- *существует такое δ, что производные $\dfrac{\partial g_i}{\partial x_j}$ непрерывны в $D = \{\boldsymbol{x} \quad | \quad ||\boldsymbol{x} - \boldsymbol{x^*}|| < \delta\}$;*
- *вторые производные $\dfrac{\partial^2 g_i}{\partial x_j \partial x_k}$ непрерывны и ограничены, т.е. $\left|\dfrac{\partial^2 g_i}{\partial x_j \partial x_k}\right| \leq M$, в $D = \{\boldsymbol{x} \quad | \quad ||\boldsymbol{x} - \boldsymbol{x^*}|| < \delta\}$;*

- $\dfrac{\partial g_i(\boldsymbol{x}^*)}{\partial x_j} = 0,$

для $i = 1, \ldots, n$, $j = 1, \ldots, n$ *и* $k = 1, \ldots, n$. *Тогда существует такое* $\hat{\delta} \leq \delta$, *что последовательность* $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$, *генерируемая итерацией* $\boldsymbol{x}^{(k)} = \boldsymbol{g}(\boldsymbol{x}^{(k-1)})$, *сходится квадратично к* $\boldsymbol{x}^*$ *для любого* $\boldsymbol{x}^{(0)}$, *удовлетворяющего* $||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*|| < \hat{\delta}$, *так, что верно следующее неравенство:*

$$||\boldsymbol{x}^{(k)} - \boldsymbol{x}^*||_\infty \leq \frac{n^2 M}{2} ||\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^*||_\infty^2, \quad k \geq 1. \tag{6.35}$$

Для вывода метода Ньютона для многомерного случая рассмотрим формулу (6.4) покоординатно:

$$g_i(\boldsymbol{x}) = x_i - \sum_{j=1}^{n} \omega_{ij}(\boldsymbol{x}) f_j(\boldsymbol{x}). \tag{6.36}$$

По условиям теоремы 6.2.4 для построения метода простой итерации с квадратичной сходимостью нам необходимо обнулить производные функций $g_i(\boldsymbol{x})$ в точке $\boldsymbol{x}^*$. Найдем для начала соответствующие производные:

$$\frac{\partial g_i(\boldsymbol{x})}{\partial x_j} = \begin{cases} 1 - \sum\limits_{j=1}^{n} \dfrac{\partial \omega_{ij}(\boldsymbol{x})}{\partial x_j} f_j(\boldsymbol{x}) - \sum\limits_{j=1}^{n} \omega_{ij}(\boldsymbol{x}) \dfrac{\partial f_j(\boldsymbol{x})}{\partial x_j}, & i = j, \\ - \sum\limits_{j=1}^{n} \dfrac{\partial \omega_{ij}(\boldsymbol{x})}{\partial x_j} f_j(\boldsymbol{x}) - \sum\limits_{j=1}^{n} \omega_{ij}(\boldsymbol{x}) \dfrac{\partial f_j(\boldsymbol{x})}{\partial x_j}, & i \neq j. \end{cases} \tag{6.37}$$

Тогда обнуление производных дает уравнения:

$$\sum_{j=1}^{n} \omega_{ij}(\boldsymbol{x}^*) \frac{\partial f_j(\boldsymbol{x}^*)}{\partial x_j} = 1, \quad i = j, \tag{6.38}$$

$$\sum_{j=1}^{n} \omega_{ij}(\boldsymbol{x}^*) \frac{\partial f_j(\boldsymbol{x}^*)}{\partial x_j} = 0, \quad i \neq j. \tag{6.39}$$

где мы использовали факт $f_j(\boldsymbol{x}^*) = 0$. Легко убедиться, что в матричном виде это эквивалентно следующему выражению:

$$\Omega(\boldsymbol{x}^*) \boldsymbol{J}(\boldsymbol{x}^*) = \boldsymbol{E} \tag{6.40}$$

где матрица $\boldsymbol{J}$ называется *матрицей Якоби* и имеет следующую форму:

$$\boldsymbol{J}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_n(\boldsymbol{x})}{\partial x_n} \end{bmatrix}. \tag{6.41}$$

Таким образом, для квадратичной сходимости в качестве $\Omega(\boldsymbol{x})$ мы можем выбрать обратную матрицу Якоби:

$$\Omega(\boldsymbol{x}) = \boldsymbol{J}^{-1}(\boldsymbol{x}), \tag{6.42}$$

что дает формулировку *метода Ньютона для систем нелинейных алгебраических уравнений*:

$$\boldsymbol{x}^{(k)} = \boldsymbol{g}(\boldsymbol{x}^{(k)}) = \boldsymbol{x}^{(k-1)} - \boldsymbol{J}^{-1}(\boldsymbol{x}^{(k-1)})\boldsymbol{f}(\boldsymbol{x}^{(k-1)}). \tag{6.43}$$

На практике обратная от матрицы Якоби не вычисляется, вместо чего используется следующая двухшаговая процедура:

$$\boldsymbol{J}(\boldsymbol{x}^{(k-1)})\boldsymbol{y}^{(k-1)} = \boldsymbol{f}(\boldsymbol{x}^{(k-1)}), \tag{6.44}$$

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - \boldsymbol{y}^{(k-1)}, \tag{6.45}$$

где $\boldsymbol{y}^{(k-1)}$ находится через решение первого уравнения.

Заметим, что в такой форме метод Ньютона требует на каждой итерации n вычислений значений векторной функции $\boldsymbol{f}(\boldsymbol{x}^{(k-1)})$, n^2 вычислений значений матрицы Якоби $\boldsymbol{J}(\boldsymbol{x}^{(k-1)})$ и $O(n^3)$ операций для нахождения $\boldsymbol{y}^{(k-1)}$. Подобная алгоритмическая сложность не позволяет использовать метод Ньютона в его стандартной форме для систем большой размерности.

6.3 Квазиньютоновские методы

Квазиньютоновские методы представляют собой класс методов, уменьшающих алгоритмическую сложность метода Ньютона за счет аппроксимации матрицы Якоби тем или иных способом. Уменьшение алгоритмической сложности, однако, приводит к замене квадратичной сходимости на сверхлинейную. Более того, так как аппроксимация матрицы Якоби в сущности является заменой точных производных скалярных функций на аппроксимации путем численного дифференцирования, квазиньютоновские методы являются вычислительно неустойчивыми и могут накапливать погрешность округления с каждой итерацией.

В этом разделе мы рассмотрим один из квазиньютоновских методов, а именно метод Бройдена, который обобщает метод хорд на многомерный случай. Предположим, что $\boldsymbol{x}^{(0)}$ является начальным приближением и следующее приближение $\boldsymbol{x}^{(1)}$ было рассчитано по методу Ньютона. Это также означает, что мы уже рассчитали матрицу Якоби в точке $\boldsymbol{x}^{(0)}$, т.е. $\boldsymbol{J}(\boldsymbol{x}^{(0)})$. Для нахождения аппроксимации матрицы Якоби в точке $\boldsymbol{x}^{(1)}$ нам необходимо сначала взглянуть на одномерный случай. Рассмотрим формулу для численного дифференцирования первого порядка:

$$f'(x^{(1)}) \approx \frac{f(x^{(1)}) - f(x^{(0)})}{x^{(1)} - x^{(0)}}. \tag{6.46}$$

Обобщить это выражение на многомерный случай можно, если переписать его в следующей форме:

$$f'(x^{(1)}) \left(x^{(1)} - x^{(0)}\right) \approx f(x^{(1)}) - f(x^{(0)}), \tag{6.47}$$

из чего следует:

$$\boldsymbol{J}(\boldsymbol{x}^{(1)}) \left(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\right) \approx \boldsymbol{f}(\boldsymbol{x}^{(1)}) - \boldsymbol{f}(\boldsymbol{x}^{(0)}). \tag{6.48}$$

Введем обозначение для аппроксимации матрицы Якоби: $\boldsymbol{A}_1 \approx \boldsymbol{J}(\boldsymbol{x}^{(1)})$. Тогда мы получаем следующее равенство:

$$\boldsymbol{A}_1 \left(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \right) = \boldsymbol{f}(\boldsymbol{x}^{(1)}) - \boldsymbol{f}(\boldsymbol{x}^{(0)}). \tag{6.49}$$

Это выражение не позволяет однозначно определить матрицу $\boldsymbol{A}_1$, так как оно определяет поведение матрицы лишь в направлении одного вектора $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}$ в n-мерном пространстве. Бройден логично предположил, что действие матрицы $\boldsymbol{A}_1$ на все вектора $\boldsymbol{z}$, ортогональные вектору $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}$, должны быть аналогичны действию на них матрицы $\boldsymbol{J}(\boldsymbol{x}^{(0)})$:

$$\boldsymbol{A}_1 \boldsymbol{z} = \boldsymbol{J}(\boldsymbol{x}^{(0)}) \boldsymbol{z}, \quad \langle \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}, \boldsymbol{z} \rangle = 0. \tag{6.50}$$

Эти два условия однозначно задают матрицу $\boldsymbol{A}_1$, выражение для которой будет тогда иметь вид:

$$\boldsymbol{A}_1 = \boldsymbol{J}(\boldsymbol{x}^{(0)}) + \frac{\left[\boldsymbol{f}(\boldsymbol{x}^{(1)}) - \boldsymbol{f}(\boldsymbol{x}^{(0)}) - \boldsymbol{J}(\boldsymbol{x}^{(0)})(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}) \right] \left[\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \right]^T}{\langle \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \rangle}. \tag{6.51}$$

Действительно, легко убедиться, что такая матрица удовлетворяет заданным условиям. Например, для вектора $\boldsymbol{z}$, удовлетворяющего $\langle \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}, \boldsymbol{z} \rangle = 0$, мы имеем:

$$\begin{aligned} \boldsymbol{A}_1 \boldsymbol{z} &= \boldsymbol{J}(\boldsymbol{x}^{(0)}) \boldsymbol{z} + \frac{\left[\boldsymbol{f}(\boldsymbol{x}^{(1)}) - \boldsymbol{f}(\boldsymbol{x}^{(0)}) - \boldsymbol{J}(\boldsymbol{x}^{(0)})(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}) \right] \langle \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}, \boldsymbol{z} \rangle}{\langle \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \rangle} \\ &= \boldsymbol{J}(\boldsymbol{x}^{(0)}) \boldsymbol{z}, \end{aligned} \tag{6.52}$$

в то время как для $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}$ перемножение дает:

$$\begin{aligned} \boldsymbol{A}_1 \left(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \right) &= \boldsymbol{J}(\boldsymbol{x}^{(0)}) \left(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \right) + \\ &\quad + \frac{\left[\boldsymbol{f}(\boldsymbol{x}^{(1)}) - \boldsymbol{f}(\boldsymbol{x}^{(0)}) - \boldsymbol{J}(\boldsymbol{x}^{(0)})(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}) \right] \langle \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \rangle}{\langle \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \rangle} \\ &= \boldsymbol{J}(\boldsymbol{x}^{(0)}) \left(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \right) + \left[\boldsymbol{f}(\boldsymbol{x}^{(1)}) - \boldsymbol{f}(\boldsymbol{x}^{(0)}) - \boldsymbol{J}(\boldsymbol{x}^{(0)})(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}) \right] \\ &= \boldsymbol{f}(\boldsymbol{x}^{(1)}) - \boldsymbol{f}(\boldsymbol{x}^{(0)}). \end{aligned} \tag{6.53}$$

Используя подобную аппроксимацию для матрицы Якоби, модифицированный метод Ньютона имеет вид:

$$\boldsymbol{A}_k = \boldsymbol{A}_{k-1} + \frac{(\boldsymbol{y}_k - \boldsymbol{A}_{k-1} \boldsymbol{s}_k) \, \boldsymbol{s}_k^T}{||\boldsymbol{s}_k||_2^2}, \tag{6.54}$$

$$\boldsymbol{A}_k \boldsymbol{w}^{(k)} = \boldsymbol{f}(\boldsymbol{x}^{(k)}), \tag{6.55}$$

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \boldsymbol{w}^{(k)}, \tag{6.56}$$

где $\boldsymbol{y}_k = \boldsymbol{f}(\boldsymbol{x}^{(k)}) - \boldsymbol{f}(\boldsymbol{x}^{(k-1)})$, $\boldsymbol{s}_k = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}$. Даже после такого упрощения нам все еще необходимо $O(n^3)$ операций для нахождения $\boldsymbol{w}^{(k)}$, так что полученный метод в его текущей форме не имеет преимуществ перед методом Ньютона. Его можно улучшить, если

сформулировать рекуррентное соотношение для обратной матрицы $\boldsymbol{A}_k^{-1}$. В этом нам поможет *формула Шермана–Моррисона*, которая гласит, что для невырожденной матрицы $\boldsymbol{A}$ и векторов $\boldsymbol{x}, \boldsymbol{y}$ верно следующее выражение:

$$\left(\boldsymbol{A} + \boldsymbol{x}\boldsymbol{y}^T\right)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{x}\boldsymbol{y}^T\boldsymbol{A}^{-1}}{\sigma}, \tag{6.57}$$

где $\sigma = 1 + \boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{x}$. Убедимся, что формула действительно является тождественно верной. Домножение слева на $\boldsymbol{A} + \boldsymbol{x}\boldsymbol{y}^T$ дает:

$$\begin{aligned} \boldsymbol{E} &= \left(\boldsymbol{A} + \boldsymbol{x}\boldsymbol{y}^T\right)\boldsymbol{A}^{-1} - \frac{\left(\boldsymbol{A} + \boldsymbol{x}\boldsymbol{y}^T\right)\boldsymbol{A}^{-1}\boldsymbol{x}\boldsymbol{y}^T\boldsymbol{A}^{-1}}{\sigma} \\ \implies \quad \boldsymbol{x}\boldsymbol{y}^T\boldsymbol{A}^{-1} &- \frac{\boldsymbol{x}\boldsymbol{y}^T\boldsymbol{A}^{-1} + \boldsymbol{x}\boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{x}\boldsymbol{y}^T\boldsymbol{A}^{-1}}{\sigma} = \boldsymbol{O}. \end{aligned} \tag{6.58}$$

Тогда домножение справа на $\boldsymbol{x}$ и подстановка $\boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{x} = \sigma - 1$ дает:

$$\begin{aligned} \boldsymbol{x}&\left(\boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{x} - \frac{\boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{x} + \boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{x}\boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{x}}{\sigma}\right) = \boldsymbol{0}, \\ \implies \quad \boldsymbol{x}&\left(\sigma - 1 - \frac{\sigma - 1 + (\sigma - 1)^2}{\sigma}\right) = \boldsymbol{0}, \end{aligned} \tag{6.59}$$

где, как несложно убедиться, выражение в скобках тождественно обращается в ноль.

Найдем выражение для $\boldsymbol{A}_k^{-1}$, используя формулу Шермана–Моррисона:

$$\begin{aligned} \boldsymbol{A}_k^{-1} &= \left(\boldsymbol{A}_{k-1} + \frac{(\boldsymbol{y}_k - \boldsymbol{A}_{k-1}\boldsymbol{s}_k)\,\boldsymbol{s}_k^T}{||\boldsymbol{s}_k||_2^2}\right)^{-1} \\ &= \boldsymbol{A}_{k-1}^{-1} - \frac{\boldsymbol{A}_{k-1}^{-1}\,(\boldsymbol{y}_k - \boldsymbol{A}_{k-1}\boldsymbol{s}_k)\,\boldsymbol{s}_k^T\boldsymbol{A}_{k-1}^{-1}}{||\boldsymbol{s}_k||_2^2\left(1 + \boldsymbol{s}_k^T\boldsymbol{A}_{k-1}^{-1}\dfrac{\boldsymbol{y}_k - \boldsymbol{A}_{k-1}\boldsymbol{s}_k}{||\boldsymbol{s}_k||_2^2}\right)} \\ &= \boldsymbol{A}_{k-1}^{-1} - \frac{\boldsymbol{A}_{k-1}^{-1}\,(\boldsymbol{y}_k - \boldsymbol{A}_{k-1}\boldsymbol{s}_k)\,\boldsymbol{s}_k^T\boldsymbol{A}_{k-1}^{-1}}{||\boldsymbol{s}_k||_2^2 + \boldsymbol{s}_k^T\boldsymbol{A}_{k-1}^{-1}(\boldsymbol{y}_k - \boldsymbol{A}_{k-1}\boldsymbol{s}_k)} \\ &= \boldsymbol{A}_{k-1}^{-1} - \frac{\left(\boldsymbol{A}_{k-1}^{-1}\boldsymbol{y}_k - \boldsymbol{s}_k\right)\boldsymbol{s}_k^T\boldsymbol{A}_{k-1}^{-1}}{\boldsymbol{s}_k^T\boldsymbol{A}_{k-1}^{-1}\boldsymbol{y}_k}, \end{aligned} \tag{6.60}$$

что позволяет рассчитать $\boldsymbol{A}_k^{-1}$, используя $\boldsymbol{A}_{k-1}^{-1}$, за $O(n^2)$ операций. Таким образом, мы получаем следующие рекуррентные соотношения для *метода Бройдена*:

$$\boldsymbol{A}_k^{-1} = \boldsymbol{A}_{k-1}^{-1} - \frac{\left(\boldsymbol{A}_{k-1}^{-1}\boldsymbol{y}_k - \boldsymbol{s}_k\right)\boldsymbol{s}_k^T\boldsymbol{A}_{k-1}^{-1}}{\boldsymbol{s}_k^T\boldsymbol{A}_{k-1}^{-1}\boldsymbol{y}_k}, \tag{6.61}$$

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \boldsymbol{A}_k^{-1}\boldsymbol{f}(\boldsymbol{x}^{(k)}). \tag{6.62}$$

6.4 Метод градиентного спуска

Метод градиентного спуска, используемый для нахождения минимума нелинейной целевой функции, может быть использован и для нахождения корней нелинейной системы алгебраических уравнений так же, как и в случае со СЛАУ (см. вывод метода градиентого спуска в разделе 5.2.9). Он обладает линейной сходимостью, вследствие чего редко используется для нахождения окончательного значения корня. Чаще его роль заключается в том, чтобы найти подходящее начальное приближение для метода Ньютона. Подобная страгегия работает благодаря тому, что метод градиентного спуска накладывает гораздо меньше ограничений на бассейн притяжения (т.е. область, которой должно принадлежать начальное приближение).

Так как в случае нелинейной системы вида 6.3 вектор невязки формально равен $\boldsymbol{f}(\boldsymbol{x})$, корень нелинейной системы алгебраических уравнений совпадает со следующим аргументом минимизации:

$$\underset{\boldsymbol{x}}{\operatorname{argmin}}\, g(\boldsymbol{x}) = \underset{\boldsymbol{x}}{\operatorname{argmin}} \langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}) \rangle. \tag{6.63}$$

Как мы уже обсуждали в разделе 5.2.9, в подобных задачах минимизации необходимо найти шаг поиска t и направление поиска $\boldsymbol{v}$ для формирования итерации вида:

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t^{(k)} \boldsymbol{v}^{(k)}. \tag{6.64}$$

Оптимальным направлением поиска является направление, обратное вектору градиента функции $g(\boldsymbol{x})$, также называемое направлением наискорейшего спуска:

$$\begin{aligned} \boldsymbol{v}^{(opt)} &= -\nabla g(\boldsymbol{x}) \\ &= -\left[\frac{\partial g}{\partial x_1}, \frac{\partial g}{\partial x_2}, \ldots, \frac{\partial g}{\partial x_n}\right]^T \\ &= -\left[\sum_{i=1}^{n} 2 f_i(\boldsymbol{x}) \frac{\partial f_i(\boldsymbol{x})}{\partial x_1}, \sum_{i=1}^{n} 2 f_i(\boldsymbol{x}) \frac{\partial f_i(\boldsymbol{x})}{\partial x_2}, \ldots, \sum_{i=1}^{n} 2 f_i(\boldsymbol{x}) \frac{\partial f_i(\boldsymbol{x})}{\partial x_n}\right]^T \\ &= -2 \boldsymbol{J}^T(\boldsymbol{x}) \boldsymbol{f}(\boldsymbol{x}). \end{aligned} \tag{6.65}$$

Выбрав в качестве направления поиска направление наискорейшего спуска, мы получаем *метод градиентного спуска*:

$$\boldsymbol{z}^{(k)} = \boldsymbol{J}^T(\boldsymbol{x}^{(k-1)}) \boldsymbol{f}(\boldsymbol{x}^{(k-1)}), \boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - t^{(k)} \frac{\boldsymbol{z}^{(k)}}{||\boldsymbol{z}^{(k)}||_2}. \tag{6.66}$$

Нахождение оптимального значения для $t^{(k)}$ не всегда является тривиальной задачей, так как это требует дифференцирования функции $h(t) = g(\boldsymbol{x}^{(k-1)} - t\frac{\boldsymbol{z}^{(k)}}{||\boldsymbol{z}^{(k)}||_2})$ относительно t. Для упрощения задачи, мы саппроксимируем функцию $h(t)$ квадратичным полиномом $P_2(t)$ и найдем ее минимум. Для этого нам необходимо найти три значения $t_1^{(k)}, t_2^{(k)}, t_3^{(k)}$. Мы сразу можем выбрать $t_1^{(k)} = 0$, так как оно является минимальным значением t и гарантированно не является оптимальным. Затем мы выбираем такое $t_3^{(k)}$, что $h(t_1^{(k)}) > h(t_3^{(k)})$, что можно

сделать, начав подбирать с $t_3^{(k)} = 1$, и после делить $t_3^{(k)}$ пополам или удваивать до тех пор, пока не выполнится указанное условие. Когда $t_3^{(k)}$, мы автоматически можем найти $t_2^{(k)} = \frac{t_3^{(k)}}{2}$. Тогда, в соответствии с интерполяцией Лагранжа (см. 2.26), мы получаем следующее выражение для $P_2(t)$:

$$P_2(t) = h(t_1^{(k)})\frac{(t-t_2^{(k)})(t-t_3^{(k)})}{(t_1^{(k)}-t_2^{(k)})(t_1^{(k)}-t_3^{(k)})} + h(t_2^{(k)})\frac{(t-t_1^{(k)})(t-t_3^{(k)})}{(t_2^{(k)}-t_1^{(k)})(t_2^{(k)}-t_3^{(k)})} + h(t_3^{(k)})\frac{(t-t_1^{(k)})(t-t_2^{(k)})}{(t_3^{(k)}-t_1^{(k)})(t_3^{(k)}-t_2^{(k)})}. \tag{6.67}$$

Производная $P_2(t)$ имеет вид:

$$\frac{dP_2}{dt} = h(t_1^{(k)})\frac{2t-t_2^{(k)}-t_3^{(k)}}{(t_1^{(k)}-t_2^{(k)})(t_1^{(k)}-t_3^{(k)})} + h(t_2^{(k)})\frac{2t-t_1^{(k)}-t_3^{(k)}}{(t_2^{(k)}-t_1^{(k)})(t_2^{(k)}-t_3^{(k)})} + h(t_3^{(k)})\frac{2t-t_1^{(k)}-t_2^{(k)}}{(t_3^{(k)}-t_1^{(k)})(t_3^{(k)}-t_2^{(k)})}. \tag{6.68}$$

Для упрощения записи введем следующие обозначения:

$$a^{(k)} = \frac{h(t_1^{(k)})}{(t_1^{(k)}-t_2^{(k)})(t_1^{(k)}-t_3^{(k)})}, \tag{6.69}$$

$$b^{(k)} = \frac{h(t_2^{(k)})}{(t_2^{(k)}-t_1^{(k)})(t_2^{(k)}-t_3^{(k)})}, \tag{6.70}$$

$$c^{(k)} = \frac{h(t_3^{(k)})}{(t_3^{(k)}-t_1^{(k)})(t_3^{(k)}-t_2^{(k)})}. \tag{6.71}$$

Тогда путем обнуления производной мы находим следующее квазиоптимальное значение для $t^{(k)}$:

$$t^{(k)} = \frac{a^{(k)}\left(t_2^{(k)}+t_3^{(k)}\right) + b^{(k)}\left(t_1^{(k)}+t_3^{(k)}\right) + c^{(k)}\left(t_1^{(k)}+t_2^{(k)}\right)}{2\left(a^{(k)}+b^{(k)}+c^{(k)}\right)}. \tag{6.72}$$

Численное решение задачи Коши для систем ОДУ

Дифференциальные уравнения являются каркасом большинства математических моделей, формулируемых во всех направлениях науки, где производится попытка математически формализовать то или иное явление или процесс. Это объясняется тем, что в большинстве случаев описание рассматриваемого явления или процесса сводится к заданию взаимосвязи между изменениями значений переменных рассматриваемой системы (например, характеризуемыми производной $\frac{dy(t)}{dt}$) и самими значениями (например, значением переменной $y(t)$) при некотором значении параметра системы (например, в момент времени t). В том случае, когда переменные рассматриваемой системы изменяются непрерывно, дифференциальные уравнения являются очевидным способом описания соответствующего явления или процесса. Для контраста отметим, что если переменные изменяются прерывно (это может происходить, например, в случае дискретного времени), операция дифференцирования становится невозможной, и используют другие способы описания: конечные автоматы, итерации отображений и проч.

Среди дифференциальных уравнений принято выделять две основные группы – *обыкновенные дифференциальные уравнения (ОДУ)* и *уравнения в частных производных*. Мы будем рассматривать *нормальные ОДУ* n-го порядка, т.е. разрешенные относительно производной:

$$y^{(n)}(t) = f(t, y, y', \ldots, y^{(n-1)}), \quad t \in [a; b]. \tag{7.1}$$

В частности, мы рассмотрим нахождение решения ОДУ с помощью задания начальных условий для переменных, что известно как *задача Коши*:

$$y(a) = y_0, \tag{7.2}$$
$$y'(a) = y_{01}, \tag{7.3}$$
$$\ldots \tag{7.4}$$
$$y^{(n-1)}(a) = y_{0,n-1}. \tag{7.5}$$

Несложно убедиться, что подобное ОДУ n-го порядка можно представить в виде системы

ОДУ 1-го порядка:

$$\frac{d}{dt}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_n \\ f(t, y_1, \ldots, y_n) \end{bmatrix}, \tag{7.6}$$

где $y_1 = y, y_2 = y', \ldots, y_n = y^{(n-1)}$. В таком случае рассмотрение численных методов решения задачи Коши для систем ОДУ 1-го порядка позволяет закрыть большой класс практических задач, связанных с нахождением решения различных систем ОДУ. Более того, те же численные методы используются при решении задачи Коши для уравнений в частных производных. Рассмотрим, например, уравнение теплопроводности:

$$\frac{\partial u}{\partial t} = \alpha \nabla^2 u + q(\boldsymbol{x}, t), \tag{7.7}$$

где $u = u(\boldsymbol{x}, t)$, $\alpha \in \mathbb{R}$ и $q(\boldsymbol{x}, t)$ – некоторая функция. Дискретизация пространства приводит к тому, что функция u, изначально являющаяся бесконечномерным вектором, заменяется на конечномерный вектор $\boldsymbol{u} = [u_1(t), u_2(t), \ldots, u_n(t)]^T$, каждый элемент которого все еще является функцией от времени t. Такая форма дискретизации называется *полудискретизацией*. Аналогичная метаморфоза происходит в функцией $q(\boldsymbol{x}, t)$, которая заменяется на конечномерный вектор $\boldsymbol{q} = [q_1(t), q_2(t), \ldots, q_n(t)]^T$. Оператор Лапласа ∇^2 также преобразуется в конечномерную форму, а именно в матрицу $\boldsymbol{A}$. В результате мы получаем следующую систему ОДУ:

$$\frac{d\boldsymbol{u}}{dt} = \alpha \boldsymbol{A}\boldsymbol{u} + \boldsymbol{q} \tag{7.8}$$

с задачей Коши для $t = t_0$. Для ее решения мы можем использовать те же численные методы, которые мы используем для любых других систем ОДУ 1-го порядка.

Убедившись в важности задачи Коши для систем ОДУ 1-го порядка, рассмотрим n-мерную систему 1-го порядка в общем виде:

$$\frac{d}{dt}\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{bmatrix} = \begin{bmatrix} f_1(t, y_1, \ldots, y_n) \\ f_2(t, y_1, \ldots, y_n) \\ \vdots \\ f_n(t, y_1, \ldots, y_n) \end{bmatrix} \tag{7.9}$$

или кратко в векторном виде:

$$\frac{d\boldsymbol{y}}{dt} = \boldsymbol{f}(t, \boldsymbol{y}). \tag{7.10}$$

с начальными условиями $y_1(a) = \alpha_1, \ldots, y_n(a) = \alpha_n$, где $t \in [a; b]$.

Важным моментом в задаче Коши является существование и единственность решений данной системы ОДУ 1-го порядка. Соответствующая теорема полагается на понятие липшиц-непрерывных функций, так что нам необходимо обобщить определение 1.3.18, данное во введении.

Определение 7.0.1. *Пусть дана область* $D = \{(t, y_1, \ldots, y_n) \mid t \in [a; b], y_1, \ldots, y_n \in \mathbb{R}\}$. *Тогда функция* $f(t, y_1, \ldots, y_n)$, *определенная на* D, *называется липшиц-непрерывной по переменным* $y_1, \ldots, y_n$, *если существует такое* $K > 0$, *что*

$$|f(t, \tilde{y}_1, \ldots, \tilde{y}_n) - f(t, \tilde{\tilde{y}}_1, \ldots, \tilde{\tilde{y}}_n)| \leq K \sum_{i=1}^{n} |\tilde{y}_i - \tilde{\tilde{y}}_i| \tag{7.11}$$

для любых $(t, \tilde{y}_1, \ldots, \tilde{y}_n) \in D$ *и* $(t, \tilde{\tilde{y}}_1, \ldots, \tilde{\tilde{y}}_n) \in D$.

Теперь можем рассмотреть формулировку теоремы о существовании и единственности решения задачи Коши без доказательства.

Теорема 7.0.1. *Пусть дана область* $D = \{(t, y_1, \ldots, y_n) \mid t \in [a; b], y_1, \ldots, y_n \in \mathbb{R}\}$. *Пусть функции* $f_i(t, y_1, \ldots, y_n), i = 1, \ldots, n$, *липшиц-непрерывны по переменным* $y_1, \ldots, y_n$ *в* D. *Тогда система ОДУ 1-го порядка вида*

$$\frac{d}{dt} \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{bmatrix} = \begin{bmatrix} f_1(t, y_1, \ldots, y_n) \\ f_2(t, y_1, \ldots, y_n) \\ \vdots \\ f_n(t, y_1, \ldots, y_n) \end{bmatrix} \tag{7.12}$$

с начальными условиями $y_1(a) = \alpha_1, \ldots, y_n(a) = \alpha_n$ *имеет единственное решение для* $t \in [a; b]$.

Здесь и далее мы будем рассматривать детальные выводы относительно одного ОДУ, а после презентовать обобщение для систем ОДУ.

7.1 Метод Эйлера

Метод Эйлера является стартовой точкой в разговоре о численных методах решения задачи Коши в силу простоты своего вывода и возможности легко аналитически оценить как глобальные, так и локальные погрешности метода.

Рассмотрим следующее ОДУ:

$$\frac{dy}{dt} = f(t, y), \tag{7.13}$$

где $t \in [a; b]$ и $y(a) = \alpha$. Все методы, которые мы будем рассматривать, предполагают дискретизацию координаты t в сетку вида $t_i = a + ih, i = 1, \ldots, m$, где $h = \frac{b-a}{m} = t_{i+1} - t_i$ называют шагом. Это автоматически дает дискретизацию решения $y(t)$ в виде $y_i = y(t_i)$. Предположим, что $y(t) \in C^2[a; b]$ и разложим функцию $y(t)$ в ряд Тейлора в точке t_i:

$$y(t) = y(t_i) + y'(t_i)(t - t_i) + \frac{y''(\xi)}{2}(t - t_i)^2, \tag{7.14}$$

где $\xi \in (t_i; t)$ для $t > t_i$. Вычислим значение ряда в точке t_{i+1}:

$$\begin{aligned} & y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2 y''(\xi_i)}{2} \\ \Longrightarrow \quad & y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2 y''(\xi_i)}{2}, \end{aligned} \tag{7.15}$$

где $\xi_i \in (t_; t_{i+1})$. Теперь предположив, что h мало, мы можем отбросить член порядка $O(h^2)$, что дает нам формулировку *метода Эйлера*:

$$w_0 = \alpha, \tag{7.16}$$

$$w_{i+1} = w_i + hf(t_i, w_i), \quad i = 0, 1, \ldots, m-1, \tag{7.17}$$

где мы ожидаем, что $w_i \approx y(t_i)$.

Рисунок 7.1 демонстрирует работу метода Эйлера на пример ОДУ $\frac{dy}{dt} = y - t^2 + 1$ для $y(0) = \frac{1}{2}$ и $t \in [0; \frac{5}{2}]$. Это ОДУ первого порядка является линейным, что позволяет легко найти точное решение: $y(t) = (t+1)^2 + (y(0) - 1)e^t$. Этот рисунок также позволяет продемонстрировать два важных вида погрешности метода, характерных для численных методов решения задачи Коши: *локальную погрешность метода* и *глобальную погрешность метода*. Локальной погрешностью (погрешностью на шаге) называют абсолютное отклонение приближенного решения от точного в пределах одного шага. На рисунке 7.1 локальную погрешность проще всего наблюдать для $t = \frac{1}{2}$. Глобальной погрешностью в свою очередь называют погрешность, накопленную на всем интервале $[a; b]$. В случае решения, изображенного на рисунке 7.1, глобальную погрешность можно оценить как максимальное абсолютное отклонение, наблюдаемое при $t = \frac{5}{2}$.

Так как при выводе метода Эйлера предполагалось, что член порядка $O(h^2)$ мал и может быть отброшен, мы можем утверждать, что локальная погрешность метода Эйлера имеет порядок $O(h^2)$. В таком случае глобальную погрешность можно оценить как наполенную локальную погрешность, что дает порядок $m \cdot O(h^2) = \frac{b-a}{h} \cdot O(h^2) \sim O(h)$. Более точный вывод верхней границы и, соответственно, порядка для глобальной погрешности гораздо сложнее, так что нам необходимо рассмотреть две вспомогательных леммы.

Лемма 7.1.1. *Для любого $x \geq -1$ и $k > 0$ справедливы следующие неравенства:*

$$0 \leq (1+x)^k \leq e^{kx} \tag{7.18}$$

Доказательство. Ряд Маклорена для $f(x) = e^x$ имеет вид:

$$e^x = 1 + x + \frac{1}{2}x^2 e^{\xi}, \tag{7.19}$$

где ξ расположен между x и 0. Тогда для $x \geq -1$ справедливо неравенство:

$$\begin{aligned} & 0 \leq 1 + x \leq 1 + x + \frac{1}{2}x^2 e^{\xi} = e^x \\ \Longrightarrow \quad & 0 \leq (1+x)^k \leq e^{kx}. \end{aligned} \tag{7.20}$$

□

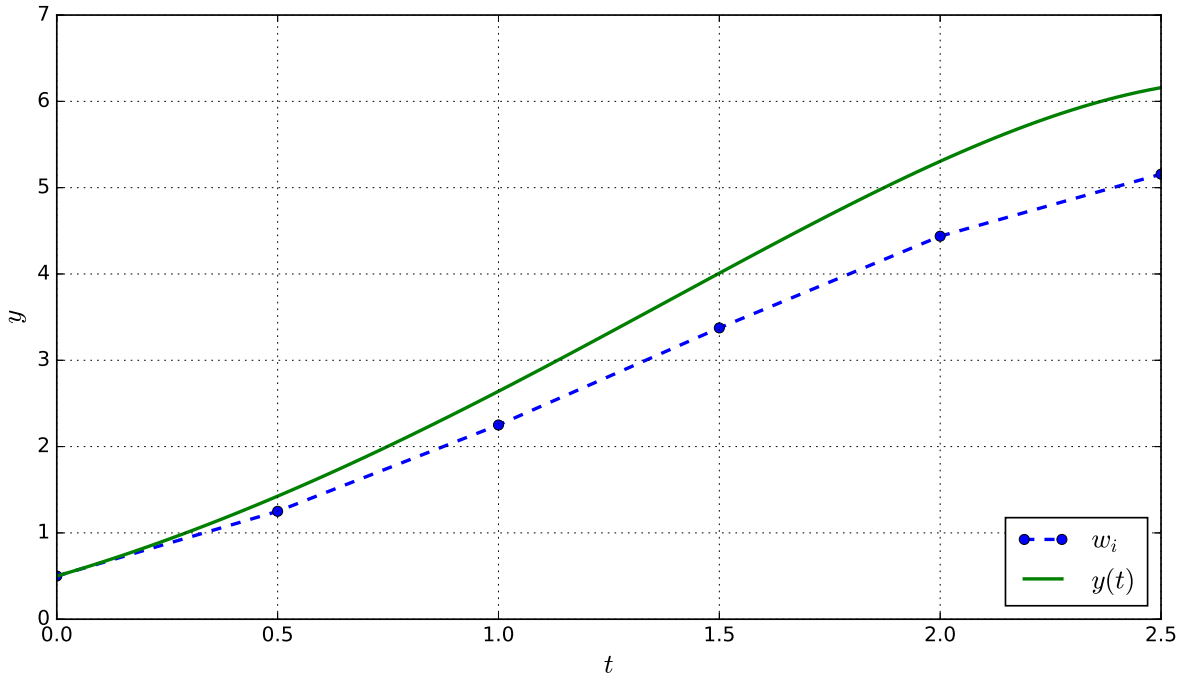

Рисунок 7.1 – Синие точки обозначают решение ОДУ $\frac{dy}{dt} = y - t^2 + 1$ с начальным условием $y(0) = \frac{1}{2}$, полученное методом Эйлера. Зеленая кривая обозначает точное решение задачи Коши.

Лемма 7.1.2. *Пусть $s, t > 0$ и последовательность $\{a_i\}_{i=0}^{k}$ задана так, что:*

$$a_0 \geq -\frac{t}{s}, \tag{7.21}$$

$$a_{i+1} \leq (1+s)a_i + t, \quad i = 0, 1, \ldots, k-1. \tag{7.22}$$

Тогда верным является следующее неравенство:

$$a_{i+1} \leq e^{(i+1)s}\left(a_0 + \frac{t}{s}\right)a_i - \frac{t}{s}. \tag{7.23}$$

Доказательство. Раскроем рекуррентное неравенство, заданное в условии леммы:

$$\begin{aligned} a_{i+1} &\leq (1+s)a_i + t \\ &\leq (1+s)\left[(1+s)a_{i-1} + t\right] + t \\ &= (1+s)^2 a_{i-1} + (1+1+s)t \\ &\leq (1+s)^3 a_{i-2} + \left[1+1+s+(1+s)^2\right]t \\ &\vdots \\ &\leq (1+s)^{i+1}a_0 + \left[1+1+s+(1+s)^2+\cdots+(1+s)^i\right]t \\ &= (1+s)^{i+1}a_0 + t\sum_{j=0}^{i}(1+s)^j. \end{aligned} \tag{7.24}$$

Несложно убедиться, что сумма по j является суммой членов геометрической прогрессии. Тогда имеем:

$$\begin{aligned} a_{i+1} &\leq (1+s)^{i+1}a_0 + t\frac{1-(1+s)^{i+1}}{1-(1+s)} \\ &= (1+s)^{i+1}a_0 + \frac{t}{s}\left[(1+s)^{i+1} - 1\right] \\ &= (1+s)^{i+1}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}, \end{aligned} \tag{7.25}$$

где, используя лемму 7.1.1, мы получаем:

$$a_{i+1} \leq e^{(i+1)s}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}. \tag{7.26}$$

□

Теперь мы можем приступить к рассмотрению теоремы о верхней границе глобальной погрешности метода Эйлера.

Теорема 7.1.1. *Пусть функция $f(t, y)$ является липшиц-непрерывной в $D = \{(t, y) \mid t \in [a; b], y \in \mathbb{R}\}$ с константой Липшица L. Пусть существует такое $M > 0$, что $|y''(t)| < M$ для любого $t \in [a; b]$, где $y(t)$ является единственным решением задачи Коши $y' = f(t, y)$,*

$y(a) = \alpha$. Тогда для последовательности $\{w_i\}_{i=0}^m$, сгенерированной методом Эйлера, верно следующее неравенство:

$$|y(t_i) - w_i| \leq \frac{hM}{2L}\left(e^{L(t_i-a)} - 1\right), \quad i = 0, 1, \ldots, m, \tag{7.27}$$

где $h = \frac{b-a}{m}$.

Доказательство. Из вывода метода Эйлера мы имеем:

$$y_{i+1} = y_i + hf(t_i, y_i) + \frac{h^2}{2}y''(\xi_i), \tag{7.28}$$

где $y_i = y(t_i)$. Тогда выражение для $y_{i+1} - w_{i+1}$ имеет вид:

$$y_{i+1} - w_{i+1} = y_i - w_i + h\left[f(t_i, y_i) - f(t_i, w_i)\right] + \frac{h^2}{2}y''(\xi_i), \tag{7.29}$$

из чего следует неравенство:

$$|y_{i+1} - w_{i+1}| \leq |y_i - w_i| + h|f(t_i, y_i) - f(t_i, w_i)| + \frac{h^2}{2}|y''(\xi_i)|. \tag{7.30}$$

Так как функция $f(t, y)$ является липщиц-непрерывной, мы имеем:

$$|y_{i+1} - w_{i+1}| \leq (1 + hL)|y_i - w_i| + \frac{h^2}{2}|y''(\xi_i)|. \tag{7.31}$$

Более того, в силу ограниченности $y''(t)$ мы получаем:

$$|y_{i+1} - w_{i+1}| \leq (1 + hL)|y_i - w_i| + \frac{h^2 M}{2}. \tag{7.32}$$

Тогда применение леммы 7.1.2 вместе с $|y_0 - w_0| = 0$ и $(i+1)h = t_{i+1} - a$ дает:

$$\begin{aligned} |y_{i+1} - w_{i+1}| &\leq e^{(i+1)hL}\left(|y_0 - w_0| + \frac{h^2 M}{2hL}\right) - \frac{h^2 M}{2hL} \\ &= \frac{hM}{2L}\left(e^{(i+1)hL} - 1\right) \\ &= \frac{hM}{2L}\left(e^{(t_{i+1}-a)L} - 1\right). \end{aligned} \tag{7.33}$$

□

7.1.1 Методы решения задачи Коши, основанные на разложении в ряд Тейлора

Очевидно, что подход, основанный на разложении в ряд Тейлора, который мы использовали при выводе метода Эйлера, может быть обобщен на разложения более высокого порядка:

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2}y''(t_i) + \cdots + \frac{h^n}{n!}y^{(n)}(t_i) + \frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\xi_i), \tag{7.34}$$

где $\xi \in (t_i, t_{i+1})$. Так как $y^{(k)} = f^{(k-1)}(t, y(t))$, мы имеем:

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2} f'(t_i, y(t_i)) + \cdots + \frac{h^n}{n!} f^{(n-1)}(t_i, y(t_i)) + \frac{h^{n+1}}{(n+1)!} f^{(n)}(\xi_i, y(\xi_i)). \tag{7.35}$$

Это позволяет сформировать метод решения задачи Коши порядка n относительно глобальной погрешности:

$$w_0 = \alpha, \tag{7.36}$$

$$w_{i+1} = w_i + hT^{(n)}(t_i, w_i), \tag{7.37}$$

где функция $T^{(n)}(t_i, w_i)$ определена как:

$$T^{(n)}(t_i, w_i) = f(t_i, y(t_i)) + \frac{h}{2} f'(t_i, y(t_i)) + \cdots + \frac{h^{n-1}}{n!} f^{(n-1)}(t_i, y(t_i)). \tag{7.38}$$

Метод Эйлера является частным случаем подобного обобщенного метода и имеет первый порядок точности.

Подобные методы не используются на практике, так как требуют дополнительных вычислений для нахождения производных $f^{(k)}(t, y(t))$. Однако они позволяют создавать методы произвольного порядка точности и являются хорошей базой для вывода более эффективных методов, таких, как метод Рунге–Кутты.

7.2 Методы Рунге–Кутты

Прежде чем перейти к выводу методов Рунге-Кутты, нам необходимо рассмотреть теорему о формуле Тейлора для функции двух переменных.

Теорема 7.2.1. *Пусть $f(t, y) \in C^{n+1}$ в области $D = \{(t, y) \mid t \in [a; b], t \in [c; d]\}$ и пусть $(t_0, y_0) \in D$. Тогда для любого $(t, y) \in D$ существуют такие ξ между t и t_0 и μ между y и y_0, что:*

$$f(t, y) = P_n(t, y) + R_n(t, y), \tag{7.39}$$

где n-й многочлен Тейлора $P_n(t, y)$ имеет вид:

$$\begin{aligned} P_n(t, y) = f(t_0, y_0) + \left[(t - t_0) \frac{\partial f}{\partial t}(t_0, y_0) + (y - y_0) \frac{\partial f}{\partial y}(t_0, y_0) \right] + \\ + \left[\frac{(t - t_0)^2}{2} \frac{\partial^2 f}{\partial t^2}(t_0, y_0) + (t - t_0)(y - y_0) \frac{\partial f}{\partial t \partial y}(t_0, y_0) + \frac{(y - y_0)^2}{2} \frac{\partial^2 f}{\partial y^2}(t_0, y_0) \right] + \cdots + \\ + \frac{1}{n!} \sum_{j=0}^{n} \binom{n}{j} (t - t_0)^{n-j} (y - y_0)^j \frac{\partial^n f}{\partial t^{n-j} \partial y^j}(t_0, y_0), \end{aligned} \tag{7.40}$$

и остаточный член $R_n(t, y)$:

$$R_n(t, y) = \frac{1}{(n+1)!} \sum_{j=0}^{n+1} \binom{n+1}{j} (t - t_0)^{n+1-j} (y - y_0)^j \frac{\partial^{n+1} f}{\partial t^{n+1-j} \partial y^j}(\xi, \mu). \tag{7.41}$$

Методы Рунге–Кутты основаны на аппроксимации $T^{(n)}(t,y)$, определенного формулой (7.38). Так как формула для $T^{(n)}(t,y)$ была построена путем отбрасывания члена порядка $O(h^n)$, нам достаточно найти аппроксимацию $\phi(t,y)$ точную вплоть до члена порядка $O(h^n)$:

$$T^{(n)}(t,y) = \phi(t,y) + O(h^n). \tag{7.42}$$

Подобная аппроксимация позволяет получить метод n-го порядка с такой функцией $\phi(t,y)$, что в ней присутствуют только вычисления значений функции $f(t,y)$ без каких-либо ее производных. Это достигается за счет использования некоторой обобщенной формы с неопределенными коэффициентами, которые затем находятся методом неопределенных коэффициентов.

Рассмотрим, например, вывод метода Рунге–Кутты 2-го порядка. Это означает, что нам необходимо найти аппроксимацию для $T^{(2)}(t,y)$ точную до члена порядка $O(h^2)$. Рассмотрим следующую аппроксимирующую функцию:

$$\phi(t,y) = a_1 f(t+\alpha_1, y+\beta_1). \tag{7.43}$$

с неопределенными коэффициентами a_1, α_1, β_1. Раскроем выражение для $T^{(2)}(t,y)$:

$$\begin{aligned} T^{(2)}(t,y) &= f(t,y) + \frac{h}{2} f'(t,y) \\ &= f(t,y) + \frac{h}{2}\left[\frac{\partial f}{\partial t}(t,y) + \frac{\partial f}{\partial y}(t,y) y'(t)\right] \\ &= f(t,y) + \frac{h}{2}\frac{\partial f}{\partial t}(t,y) + \frac{h}{2}\frac{\partial f}{\partial y}(t,y) f(t,y). \end{aligned} \tag{7.44}$$

Теперь разложим функцию $f(t,y)$ в ряд Тейлора в точке (t,y) и вычислим значение ряда, соответствующее $\phi(t,y) = a_1 f(t+\alpha_1, y+\beta_1)$:

$$a_1 f(t+\alpha_1, y+\beta_1) = a_1 f(t,y) + a_1\alpha_1 \frac{\partial f}{\partial t}(t,y) + a_1\beta_1 \frac{\partial f}{\partial y}(t,y) + a_1 R_1(t+\alpha_1, y+\beta_1). \tag{7.45}$$

Приравнивая коэффициенты при $f(t,y)$, $\frac{\partial f}{\partial t}(t,y)$ и $\frac{\partial f}{\partial y}(t,y)$ в выражениях (7.44) и (7.45), получаем:

$$a_1 = 1, \tag{7.46}$$

$$\alpha_1 = \frac{h}{2}, \tag{7.47}$$

$$\alpha_2 = \frac{h}{2} f(t,y), \tag{7.48}$$

из чего следует:

$$T^{(2)}(t,y) = f(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)) - R_1(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)). \tag{7.49}$$

Мы ожидаем, что функция R_1 будет иметь порядок $O(h^2)$. Убедимся в этом:

$$\begin{aligned} R_1(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)) &= \frac{h^2}{8}\frac{\partial^2 f}{\partial t^2}(\xi,\mu) + \frac{h^2}{4} f(t,y)\frac{\partial^2 f}{\partial t \partial y}(\xi,\mu) + \frac{h^2}{8} f^2(t,y)\frac{\partial^2 f}{\partial y^2}(\xi,\mu) \\ &= O(h^2). \end{aligned} \tag{7.50}$$

Таким образом, мы получаем формулировку *метода Рунге–Кутты 2-го порядка*, также называемого *модифицированным методом Эйлера*:

$$w_0 = \alpha, \tag{7.51}$$

$$w_{i+1} = w_i + hf(t_i + \frac{h}{2}, w_i + \frac{h}{2} f(t_i, w_i)), \quad i = 0, 1, \ldots, m-1. \tag{7.52}$$

Функция $\phi(t, y)$ была подобрана таким образом, что соответствующее разложение в ряд Тейлора включало в себя составляющие функции $T^{(2)}(t, y)$, т.е. $f(t, y)$, $\frac{\partial f}{\partial t}(t, y)$ и $\frac{\partial f}{\partial y}(t, y)$. Рассмотрим теперь вывод метода Рунге–Кутты 3-го порядка. Для этого нам необходимо найти аппроксимацию для выражения $T^{(3)}(t, y)$:

$$\begin{aligned}
T^{(3)}(t, y) &= f + \frac{h}{2} f' + \frac{h^2}{6} f'' \\
&= f + \frac{h}{2} \left[\partial_t f + f \partial_y f\right] + \frac{h^2}{6} \frac{d}{dt} \left[f \partial_y f + \partial_t f\right] \\
&= f + \frac{h}{2} \left[\partial_t f + f \partial_y f\right] + \frac{h^2}{6} \left[f' \partial_y f + f \frac{d}{dt} \left(\partial_y f\right) + \frac{d}{dt} \left(\partial_t f\right)\right] \\
&= f + \frac{h}{2} \left[\partial_t f + f \partial_y f\right] + \frac{h^2}{6} \left[\left(\partial_t f + f \partial_y f\right) \partial_y f + f \left(\partial_{yt} f + f \partial_{yy} f\right) + \left(\partial_{tt} f + f \partial_{ty} f\right)\right], \\
&= f + \frac{h}{2} \left[\partial_t f + f \partial_y f\right] + \frac{h^2}{6} \left[\partial_t f \partial_y f + f \left(\partial_y f\right)^2 + 2 f \partial_{yt} f + f^2 \partial_{yy} f + \partial_{tt} f\right],
\end{aligned} \tag{7.53}$$

где для компактности была опущена зависимость $f(t, y)$ от t и y и использовались сокращенные записи производных: например, $\partial_t f = \frac{\partial f}{\partial t}$ и $\partial_{ty} f = \frac{\partial^2 f}{\partial t \partial y}$. Заметим, что в выражении для $T^{(3)}(t, y)$ присутствуют $\partial_t f \partial_y f$ и $\left(\partial_y f\right)^2$, нехарактерные для обычного разложения в ряд Тейлора. Чтобы те же члены присутствовали в аппроксимации $\phi(t, y)$, нам необходимо рассмотреть следующую форму для этой функции:

$$\phi(t, y) = a_1 f(t, y) + a_2 f\left(t + \alpha_1, y + \beta_1 f\left(t + \alpha_2, y + \beta_2 f(t, y)\right)\right). \tag{7.54}$$

Прежде чем рассмотреть соответствующее разложение в ряд Тейлора, заметим, что вполне логично предположить следующие оценки для коэффициентов (см. (7.52)):

$$a_1 \sim O(1), \tag{7.55}$$

$$a_2 \sim O(1), \tag{7.56}$$

$$\alpha_1 \sim O(h), \tag{7.57}$$

$$\alpha_2 \sim O(h), \tag{7.58}$$

$$\beta_1 \sim O(h), \tag{7.59}$$

$$\beta_2 \sim O(h). \tag{7.60}$$

Подобные оценки помогут сразу записать получившийся порядок и опустить некоторые выкладки. Теперь рассмотрим выражения для $\phi(t, y)$ с учетом разложения $f(t, y)$ в ряд

Тейлора:

$$\phi(t,y) = a_1 f + a_2 \left[f + \alpha_1 \partial_t f + \beta_1 f(t+\alpha_2, y+\beta_2 f)\partial_y f + \frac{1}{2}\alpha_1^2 \partial_{tt} f + \right.$$
$$\left. + \alpha_1\beta_1 f(t+\alpha_2, y+\beta_2 f)\partial_{ty} f + \frac{1}{2}\beta_1^2 f^2 (t+\alpha_2, y+\beta_2 f)\partial_{yy} f \right] + O(h^3), \quad (7.61)$$

где $f = f(t,y)$. Заметим при этом, что:

$$f(t+\alpha_2, y+\beta_2 f) = f + \alpha_2 \partial_t f + \beta_2 \partial_y f + O(h^2). \quad (7.62)$$

Тогда мы имеем:

$$\phi(t,y) = a_1 f + a_2 \left[f + \alpha_1 \partial_t f + \beta_1 f \partial_y f + \beta_1\alpha_2 \partial_t f \partial_y f + \beta_1\beta_2 \left(\partial_y f\right)^2 + \frac{1}{2}\alpha_1^2 \partial_{tt} f + \right.$$
$$\left. + \alpha_1\beta_1 f \partial_{ty} f + \frac{1}{2}\beta_1^2 f^2 \partial_{yy} f \right] + O(h^3). \quad (7.63)$$

Затем, как и в предыдущем случае, составляем систему уравнений, приравнивая коэффициенты при подобных членах в выражениях (7.53) и (7.63):

$$f \Longrightarrow a_1 + a_2 = 1, \quad (7.64)$$
$$\partial_y f \Longrightarrow \frac{h}{2} f = a_2\beta_1 f, \quad (7.65)$$
$$\partial_t f \Longrightarrow \frac{h}{2} = a_2\alpha_1, \quad (7.66)$$
$$\partial_y f \partial_t f \Longrightarrow \frac{h^2}{6} = a_2\beta_1\alpha_2, \quad (7.67)$$
$$(\partial_y f)^2 \Longrightarrow \frac{h^2}{6} f = a_2\beta_1\beta_2 f, \quad (7.68)$$
$$\partial_{tt} f \Longrightarrow \frac{h^2}{6} = \frac{1}{2} a_2 \alpha_1^2, \quad (7.69)$$
$$\partial_{ty} f \Longrightarrow 2\frac{h^2}{6} f = \frac{1}{2} a_2\alpha_1\beta_1 f, \quad (7.70)$$
$$\partial_{yy} f \Longrightarrow \frac{h^2}{6} f^2 = \frac{1}{2} a_2 \beta_1^2 f^2, \quad (7.71)$$

решением которой являются следующие коэффициенты:

$$a_1 = \frac{1}{4},\ a_2 = \frac{3}{4},\ \alpha_1 = \beta_1 = \frac{2h}{3},\ \alpha_2 = \beta_2 = \frac{h}{3}. \quad (7.72)$$

Таким образом, мы вывели формулу для *метода Рунге–Кутты 3-го порядка*:

$$w_0 = \alpha, \quad (7.73)$$
$$w_{i+1} = w_i + \frac{h}{4}\left[f(t_i, w_i) + 3f\left(t_i + \frac{2h}{3}, w_i + \frac{2h}{3} f\left(t_i + \frac{h}{3}, w_i + \frac{h}{3} f(t_i, w_i) \right) \right) \right], \quad i = 0, 1, \ldots, m-1, \quad (7.74)$$

который также можно записать в следующем виде:

$$w_0 = \alpha, \tag{7.75}$$
$$k_1 = hf(t_i, w_i), \tag{7.76}$$
$$k_2 = hf\left(t_i + \frac{h}{3}, w_i + \frac{1}{3}k_1\right), \tag{7.77}$$
$$k_3 = hf\left(t_i + \frac{2h}{3}, w_i + \frac{2}{3}k_2\right), \tag{7.78}$$
$$w_{i+1} = w_i + \frac{1}{4}\left(k_1 + 3k_3\right), \quad i = 0, 1, \ldots, m-1. \tag{7.79}$$

Аналогично можно вывести и метод Рунге–Кутты 4-го порядка, что, правда, становится алгебраически более трудоемким (для дальнейших выводов рекомендуется пользоваться обобщенная схемой формирования выражений для методов Рунге–Кутты произвольного порядка **[TODO: Butcher]**). Приведем формулировку метода Рунге–Кутты 4-го порядка:

$$w_0 = \alpha, \tag{7.80}$$
$$k_1 = hf(t_i, w_i), \tag{7.81}$$
$$k_2 = hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_1\right), \tag{7.82}$$
$$k_3 = hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_2\right), \tag{7.83}$$
$$k_4 = hf\left(t_i + h, w_i + k_3\right), \tag{7.84}$$
$$w_{i+1} = w_i + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right), \quad i = 0, 1, \ldots, m-1. \tag{7.85}$$

Также отметим, что в случае системы ОДУ 1-го порядка выражения, полученные для методов Рунге–Кутта не претерпевают никаких существенных изменений. Например, формулировка метода Рунге–Кутты 4-го порядка для систем ОДУ имеет вид:

$$\boldsymbol{w}_0 = \boldsymbol{\alpha}, \tag{7.86}$$
$$\boldsymbol{k}_1 = h\boldsymbol{f}(t_i, \boldsymbol{w_i}), \tag{7.87}$$
$$\boldsymbol{k}_2 = h\boldsymbol{f}\left(t_i + \frac{h}{2}, \boldsymbol{w}_i + \frac{1}{2}\boldsymbol{k}_1\right), \tag{7.88}$$
$$\boldsymbol{k}_3 = h\boldsymbol{f}\left(t_i + \frac{h}{2}, \boldsymbol{w}_i + \frac{1}{2}\boldsymbol{k}_2\right), \tag{7.89}$$
$$\boldsymbol{k}_4 = h\boldsymbol{f}\left(t_i + h, \boldsymbol{w}_i + \boldsymbol{k}_3\right), \tag{7.90}$$
$$\boldsymbol{w}_{i+1} = \boldsymbol{w}_i + \frac{1}{6}\left(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4\right), \quad i = 0, 1, \ldots, m-1. \tag{7.91}$$

Заметим, что увеличение порядка формулы приводит к увеличению количества арифметических операций. В частности, формулы метода Рунге–Кутты 3-го и 4-го порядка требуют 3 и 4 вычислений функции $f(t, y)$ соответственно, в то время как формула 5-го порядка

будет требовать уже 6 вычислений функции $f(t, y)$ **[TODO: Butcher]**. Тренд на непропорциональное увеличение числа вычислений сохраняется и для формул большего порядка. Таким образом, оптимальным выбором среди прочих методов Рунге–Кутты является метод Рунге–Кутты 4-го порядка.

7.3 Многошаговые методы

Все методы, которые мы рассматривали до этого момента, были построены таким образом, что они использовали информацию только в пределах одного шага $t \in [t_i; t_{i+1}]$. Логично предположить, что мы можем увеличить порядок точности за счет использования уже посчитанных значений функции $y(t)$ на предыдущих шагах t_{i-1}, t_{i-2} и так далее. В общем случае соответствующая p-шаговая формула численного решения задачи Коши будет иметь вид:

$$w_{i+1} = a_{p-1}w_i + a_{p-2}w_{i-1} + \cdots + a_0 w_{i+1-p} + \\ + h\left[b_p f(t_{i+1}, w_{i+1}) + b_{p-1}f(t_i, w_i) + \cdots + b_0 f(t_{i+1-p}, w_{i+1-p})\right], \quad (7.92)$$

где при $b_p \neq 0$ формула называется *неявной*, так как рекуррентное соотношение в таком случае не разрешено относительно w_{i+1}. Заметим, что $i \geq p-1$ и, следовательно, для расчета значений $w_{i+1}, i < p-1$ требуется использовать одношаговый метод (например, метод Рунге–Кутты 4-го порядка).

Читатель, ознакомившийся с главой о численном дифференцировании, уже может догадаться, что подобные формулы могут быть выведены с помощью разложения в ряд Тейлора или интерполяции Лагранжа. Мы воспользуемся альтернативным подходом, дающим, однако, тот же результат и так же включающий в себя интерполяцию. Как и раньше, рассмотрим следующее ОДУ:

$$\frac{dy}{dt} = f(t, y), \quad (7.93)$$

где $t \in [a; b]$ и $y(a) = \alpha$. Проинтегрируем обе части уравнения на интервале $[t_i; t_{i+1}]$:

$$\begin{aligned} \int_{t_i}^{t_{i+1}} \frac{dy}{dt} dt &= \int_{t_i}^{t_{i+1}} f(t, y(t)) dt \\ \implies \quad y(t_{i+1}) &= y(t_i) + \int_{t_i}^{t_{i+1}} f(t, y(t)) dt. \end{aligned} \quad (7.94)$$

Заметим, что функция $f(t, y(t))$ в сущности является функцией одной переменной t. Тогда для того, чтобы проинтегрировать $f(t, y(t))$, нам необходимо построить для нее соответствующую аппроксимацию вида $f(t, y(t)) \approx P(t)$. Простейшим решением является полиномиальная интерполяция, т.е. построение полинома Лагранжа, проходящего через точки, значения которых уже известны:

$$(t_i, f(t_i, y(t_i))), (t_{i-1}, f(t_{i-1}, y(t_{i-1}))), \ldots, (t_{i+1-p}, f(t_{i+1-p}, y(t_{i+1-p}))). \quad (7.95)$$

Если при этом требуется построить неявную формулу, то интерполянт также должен проходить через точку $(t_{i+1}, f(t_{i+1}, y(t_{i+1})))$. В случае явной формулы, используя теорему 2.5.1,

мы получаем:

$$\begin{aligned} f(t,y(t)) &= L_{p-1}(t) + \frac{f^{(p)}(\xi_i, y(\xi_i))}{p!}\prod_{j=1}^{p}(t-t_{i-j+1}) \\ &= \sum_{j=1}^{p} f(t_{i-j+1}, y(t_{i-j+1}))\prod_{j\neq k}\frac{t-t_{i-k+1}}{t_{i-j+1}-t_{i-k+1}} + \frac{f^{(p)}(\xi_i, y(\xi_i))}{p!}\prod_{j=1}^{p}(t-t_{i-j+1}) \quad (7.96) \\ &= \sum_{j=1}^{p} f(t_{i-j+1}, y(t_{i-j+1}))\prod_{j\neq k}\frac{t-t_{i-k+1}}{h(k-j)} + \frac{f^{(p)}(\xi_i, y(\xi_i))}{p!}\prod_{j=1}^{p}(t-t_{i-j+1}), \end{aligned}$$

где $\xi_i = \xi_i(t) \in (t_{i-p+1}; t_i)$. Тогда интеграл от $f(t, y(t))$ принимает вид:

$$\begin{aligned} \int_{t_i}^{t_{i+1}} f(t,y(t))dt = \sum_{j=1}^{p} f(t_{i-j+1}, y(t_{i-j+1}))\int_{t_i}^{t_{i+1}}\prod_{j\neq k}\frac{t-t_{i-k+1}}{h(k-j)}dt + \\ + \int_{t_i}^{t_{i+1}}\frac{f^{(p)}(\xi_i, y(\xi_i))}{p!}\prod_{j=1}^{p}(t-t_{i-j+1})dt. \quad (7.97) \end{aligned}$$

Для вывода зависимости остаточного члена от h, проведем замену $t = t_i + sh$ во втором интеграле:

$$\begin{aligned} \int_{t_i}^{t_{i+1}}\frac{f^{(p)}(\xi_i, y(\xi_i))}{p!}\prod_{j=1}^{p}(t-t_{i-j+1})dt &= \frac{h}{p!}\int_0^1 f^{(p)}(\xi_i, y(\xi_i))\prod_{j=1}^{p}(t_i+sh-t_{i-j+1})ds \\ &= \frac{h^{p+1}}{p!}\int_0^1 f^{(p)}(\xi_i, y(\xi_i))\prod_{j=1}^{p}(s+j-1)ds \quad (7.98) \\ &= \frac{h^{p+1}}{p!}f^{(p)}(\mu_i, y(\mu_i))\int_0^1\prod_{j=1}^{p}(s+j-1)ds, \end{aligned}$$

где в последнем шаге мы воспользовались теоремой о среднем значении ($\prod_{j=1}^{p}(s+j-1)$ не меняет знак для $s \in [0;1]$) и $\mu_i \in (t_i; t_{i+1})$. Полученное выражение говорит о том, что локальная погрешность метода будет иметь порядок $O(h^{p+1})$. Проведем аналогичную замену для первого интеграла:

$$\int_{t_i}^{t_{i+1}} f(t,y(t))dt = h\sum_{j=1}^{p} f(t_{i-j+1}, y(t_{i-j+1}))\int_0^1\prod_{j\neq k}\frac{s+k-1}{k-j}ds. \quad (7.99)$$

Таким образом мы получаем следующую обобщенную формулировку явного p-шагового метода, известного как *метод Адамса–Башфорта* p-го порядка:

$$w_0 = \alpha_0,\ w_1 = \alpha_1,\ \ldots,\ w_p = \alpha_p, \quad (7.100)$$

$$w_{i+1} = w_i + h\sum_{j=1}^{p} a_j f(t_{i-j+1}, w_{i-j+1}), \quad i = p, p+1, \ldots, m-1, \quad (7.101)$$

где коэффициенты a_j имеют вид:

$$a_j = \int_0^1 \prod_{j \neq k} \frac{s+k-1}{k-j} ds, \quad j = 1, \ldots, p. \tag{7.102}$$

Как и в случае с методом Рунге–Кутты, порядок метода Адамса–Башфорта определяется порядком глобальной погрешности, полученной в результате накопления локальной погрешности.

В качестве примера рассмотрим двухшаговый метод. Коэффициенты a_j для него имеют вид:

$$a_1 = \int_0^1 \prod_{j \neq k} \frac{s+2-1}{2-1} ds = \int_0^1 (s+1) ds = \frac{3}{2}, \tag{7.103}$$

$$a_2 = \int_0^1 \prod_{j \neq k} \frac{s+1-1}{1-2} ds = -\int_0^1 s ds = -\frac{1}{2}. \tag{7.104}$$

Тогда метод Адамса–Башфорта второго порядка принимает вид:

$$w_0 = \alpha_0, \ w_1 = \alpha_1, \tag{7.105}$$

$$w_{i+1} = w_i + \frac{h}{2} \left[3f(t_i, w_i) - f(t_{i-1}, w_{i-1}) \right], \quad i = 1, 2, \ldots, m-1, \tag{7.106}$$

Аналогичным образом можно получить формулировку для неявного метода. В таком случае мы имеем следующие p точек интерполяции:

$$(t_{i+1}, f(t_{i+1}, y(t_{i+1}))), (t_i, f(t_i, y(t_i))), \ldots, (t_{i+2-p}, f(t_{i+2-p}, y(t_{i+2-p}))). \tag{7.107}$$

Дальнейший вывод будет идентичен выводу явного метода. В результате мы получаем неявный $(p-1)$-шаговый метод, известный как *метод Адамса–Моултона*:

$$w_0 = \alpha_0, \ w_1 = \alpha_1, \ \ldots, \ w_{p-1} = \alpha_{p-1}, \tag{7.108}$$

$$w_{i+1} = w_i + h \sum_{j=1}^{p} a_j f(t_{i-j+2}, w_{i-j+2}), \quad i = p-1, p, \ldots, m-1, \tag{7.109}$$

где коэффициенты a_j имеют вид:

$$a_j = \int_0^1 \prod_{j \neq k} \frac{s+k-2}{k-j} ds, \quad j = 1, \ldots, p, \tag{7.110}$$

и остаточный член, формирующийся из-за интерполяции функции $f(t, y)$, имеет форму:

$$\frac{h^{p+1}}{p!} f^{(p)}(\mu_i, y(\mu_i)) \int_0^1 \prod_{j=1}^{p} (s+j-2) ds, \tag{7.111}$$

Таким образом, $(p-1)$-шаговый метод Адамса–Моултона имеет порядок $O(h^p)$, что делает его на один порядок более точным, чем $(p-1)$-шаговый метод Адамса–Башфорта.

Неявные методы в общем случае всегда дают более точный результат, чем явные методы, для одного и того же количества шагов или вычислений функции $f(t,y)$. В случае методов Адамса–Башфорта и Адамса–Моултона это является следствием того, что методы Адамса–Башфорта для вычисления интеграла фактически экстраполируют функцию $f(t,y(t))$ на интервале $[t_i;t_{i+1}]$ на основе интерполянта, построенного на интервале $[t_{i-p+1};t_i]$, в то время как методы Адамса–Моултона интегрируют сам интерполянт. Очевидно, что интегрирование интерполянта дает более точный результат, чем интегрирование экстраполянта.

Более того, неявные методы в общем случае оказываются абсолютно устойчивыми, тогда явные методы почти всегда являются лишь условно устойчивыми.

7.4 Методы, построенные по схеме предиктор-корректор

Несмотря на рассмотренные преимущества неявных методов, основным их недостатком является необходимость искать корни системы нелинейных алгебраических уравнений. Выходом из этого положения является использование двух стадий в расчете w_{i+1}. На первой стадии с помощью явного метода, называемого *предиктором*, находится приближение $\tilde{w}_{i+1} \approx w_{i+1}$. Затем, на второй стадии, с помощью неявного метода, называемого *корректором*, находится само значение w_{i+1}, при этом в правой части неявного метода вместо w_{i+1} используется $\tilde{w}_{i+1}$. Вторую стадию можно затем повторять, имитируя метод простой итерации, что зачастую позволяет улучшить устойчивость численного решения. Рассмотрим несколько конкретных примеров подобных методов.

7.4.1 Метод Хойна

Простейший метод такого рода может быть построен путем комбинации одношаговых методов Адамса–Башфорта (т.е. метода Эйлера) и Адамса–Моултона. Одношаговый метод Адамса–Моултона также известен как *метод трапеций*, названный в честь аналогичной формулы численного интегрирования (3.42):

$$w_0 = \alpha, \tag{7.112}$$

$$w_{i+1} = w_i + \frac{h}{2}\left[f(t_{i+1}, w_{i+1}) + f(t_i, w_i)\right], \quad i = 0, \ldots, m-1. \tag{7.113}$$

Метод Хойна в таком случае может быть сформулирован следующим образом:

$$w_0 = \alpha, \tag{7.114}$$

$$\tilde{w}_{i+1} = w_i + hf(t_i, w_i), \tag{7.115}$$

$$w_{i+1} = w_i + \frac{h}{2}\left[f(t_{i+1}, \tilde{w}_{i+1}) + f(t_i, w_i)\right], \quad i = 0, \ldots, m-1. \tag{7.116}$$

Использование схемы предиктор-корректор позволило улучшить точность численного решения (метод Эйлера имеет порядок $O(h)$, в то время как метод трапеций $O(h^2)$) и одно-

временно повысить устойчивость численной схемы (метод Эйлера является условно устойчивым методом, в то время как метод трапеций абсолютно устойчив).

7.4.2 Метод Адамса–Башфорта–Моултона

Метод Адамса–Башфорта–Моултона строится за счет комбинации многошаговых методов Адамса–Башфорта и Адамса–Моултона одного порядка точности, что повышает устойчивость численной схемы. Например, для методов 4-го порядка мы получаем Метод Адамса–Башфорта–Моултона 4-го порядка следующего вида:

$$w_0 = \alpha_0, w_1 = \alpha_1, w_2 = \alpha_2, w_3 = \alpha_3 \tag{7.117}$$

$$\tilde{w}_{i+1} = w_i + \frac{h}{24}\left[55f(t_i, w_i) - 59f(t_{i-1}, w_{i-1}) + 37f(t_{i-2}, w_{i-2}) - 9f(t_{i-3}, w_{i-3})\right], \tag{7.118}$$

$$w_{i+1} = w_i + \frac{h}{24}\left[9f(t_{i+1}, \tilde{w}_{i+1}) + 19f(t_i, w_i) - 5f(t_{i-1}, w_{i-1}) + f(t_{i-2}, w_{i-2})\right], \quad i = 3, \ldots, m-1. \tag{7.119}$$

7.4.3 Метод Милна–Симпсона

Методы Адамса – не единственные многошаговые методы, построенные на основе интегрирования ОДУ и дальнейшей полиномиальной интерполяции. Другой класс многошаговых методов можно построить, если проинтегрировать ОДУ $\frac{dy}{dt} = f(t, y)$ на интервале $[t_j; t_{i+1}]$ для $j < i$ вместо $[t_i; t_{i+1}]$, как в методах Адамса:

$$\begin{aligned} &\int_{t_j}^{t_{i+1}} \frac{dy}{dt} dt = \int_{t_j}^{t_{i+1}} f(t, y(t)) dt \\ \Longrightarrow \quad &y(t_{i+1}) = y(t_j) + \int_{t_j}^{t_{i+1}} f(t, y(t)) dt. \end{aligned} \tag{7.120}$$

Так, например, для интервала $[t_{i-3}; t_{i+1}]$ и квадратичного интерполянта для $f(t, y(t))$ мы получаем явный метод, называемый *методом Милна*:

$$w_0 = \alpha_0, w_1 = \alpha_1, w_2 = \alpha_2, w_3 = \alpha_3 \tag{7.121}$$

$$w_{i+1} = w_{i-3} + \frac{4h}{3}\left[2f(t_i, w_i) - f(t_{i-1}, w_{i-1}) + 2f(t_{i-2}, w_{i-2})\right], \quad i = 3, \ldots, m-1. \tag{7.122}$$

Подобным же образом для интервала $[t_{i-1}; t_{i+1}]$ и квадратичного интерполянта для $f(t, y(t))$ мы получаем неявный метод, называемый *методом Симпсона* по аналогии с известной формулой численного интегрирования 3.49:

$$w_0 = \alpha_0, w_1 = \alpha_1, \tag{7.123}$$

$$w_{i+1} = w_{i-1} + \frac{h}{3}\left[f(t_{i+1}, w_{i+1}) + 4f(t_i, w_i) + f(t_{i-1}, w_{i-1})\right], \quad i = 1, \ldots, m-1. \tag{7.124}$$

Рассмотренные методы часто используются в комбинации по схеме предиктор-корректор:

$$w_0 = \alpha_0, w_1 = \alpha_1, w_2 = \alpha_2, w_3 = \alpha_3 \tag{7.125}$$

$$\tilde{w}_{i+1} = w_{i-3} + \frac{4h}{3}\left[2f(t_i, w_i) - f(t_{i-1}, w_{i-1}) + 2f(t_{i-2}, w_{i-2})\right], \tag{7.126}$$

$$w_{i+1} = w_{i-1} + \frac{h}{3}\left[f(t_{i+1}, \tilde{w}_{i+1}) + 4f(t_i, w_i) + f(t_{i-1}, w_{i-1})\right], \quad i = 3, \ldots, m-1. \tag{7.127}$$

7.5 Устойчивость численных схем

Мы уже несколько раз упоминали факт устойчивости или неустойчивости тех или иных численных методов решения задачи Коши. Чаще всего под понятием "устойчивости" в этом контексте понимается *A-устойчивость*, являющейся, в сущности, формой линейной устойчивости. Прежде чем перейти к устойчивости численных схем, рассмотрим линейную устойчивость самого ОДУ. Как и раньше, мы рассматриваем следующее ОДУ:

$$y' = f(t, y). \tag{7.128}$$

Предположим, что нам известно решение $y(t_i) = y_i$ на шаге t_i. Рассмотрим разложение $f(t, y)$ в ряд Тейлора в точке (t_i, y_i):

$$f(t, y) = f(t_i, y_i) + (t - t_i)\frac{\partial f}{\partial t}(t_i, y_i) + (y - y_i)\frac{\partial f}{\partial y}(t_i, y_i) + O\left((t - t_i)^2, (y - y_i)^2\right). \tag{7.129}$$

Линеаризация исходного ОДУ подразумевает отбрасывание квадратичного члена в этом разложении, что приводит к линейному дифференциальному уравнению первого порядка:

$$y' - y\frac{\partial f}{\partial y}(t_i, y_i) = f(t_i, y_i) + (t - t_i)\frac{\partial f}{\partial t}(t_i, y_i) - y_i\frac{\partial f}{\partial y}(t_i, y_i). \tag{7.130}$$

Для упрощения записи введем следующие обозначения:

$$f_i = f(t_i, y_i), \tag{7.131}$$

$$k_t = \frac{\partial f}{\partial t}(t_i, y_i), \tag{7.132}$$

$$k_y = \frac{\partial f}{\partial y}(t_i, y_i). \tag{7.133}$$

Тогда мы имеем:

$$y' - k_y y = f_i + (t - t_i)k_t - k_y y_i. \tag{7.134}$$

Это ОДУ можно решить, используя хорошо известный метод интегрирующего множителя. Разберем его детально. Домножим (7.134) на множитель $\mu(t)$:

$$\mu(t)y'(t) - k_y\mu(t)y(t) = \mu(t)\left[f_i + (t - t_i)k_t - k_y y_i\right]. \tag{7.135}$$

Тогда при $\mu'(t) = -k_y\mu(t)$, мы имеем:

$$(\mu(t)y(t))' = \mu(t)\left[f_i + (t - t_i)k_t - k_y y_i\right], \tag{7.136}$$

из чего следует решение:

$$y(t) = \frac{\int \mu(t) \left[f_i + (t - t_i)k_t - k_y y_i \right] dt + c_1}{\mu(t)}. \tag{7.137}$$

Для того, чтобы найти явное выражение для $y(t)$, нам необходимо рассчитать $\mu(t)$:

$$\begin{aligned} & \mu'(t) = -k_y \mu(t) \\ \implies \quad & \mu(t) = c_2 e^{-k_y t}. \end{aligned} \tag{7.138}$$

Тогда выражение для $y(t)$ принимает вид:

$$y(t) = ce^{k_y t} + e^{k_y t} \int_{t_i}^{t} e^{-k_y \tau} \left[f_i + k_t(\tau - t_i) - k_y y_i \right] d\tau. \tag{7.139}$$

Коэффициент c находится из условия $y(t_i) = y_i$:

$$\begin{aligned} & y(t_i) = ce^{k_y t_i} \\ \implies \quad & c = y_i e^{-k_y t_i}, \end{aligned} \tag{7.140}$$

что дает:

$$y(t) = y_i e^{k_y (t - t_i)} + e^{k_y t} \int_{t_i}^{t} e^{-k_y \tau} \left[f_i + k_t(\tau - t_i) - k_y y_i \right] d\tau. \tag{7.141}$$

После интегрирования мы имеем:

$$\begin{aligned} y(t) &= y_i e^{k_y (t - t_i)} + e^{k_y t} \left(e^{-k_y t} - e^{-k_y t_i} \right) (f_i - k_t t_i - k_y y_i) + k_t e^{k_y t} \int_{t_i}^{t} e^{-k_y \tau} \tau d\tau \\ &= \left[y_i - f_i + k_t t_i + k_y y_i + \frac{k_t t_i}{k_y} - \frac{k_t}{k_y^2} \right] e^{k_y (t - t_i)} - \frac{k_t}{k_y} t + \left[f_i - k_t t_i - k_y y_i + \frac{k_t}{k_y^2} \right]. \end{aligned} \tag{7.142}$$

Из данного выражения можно заметить, что при $k_y > 0$ будет наблюдаться бесконечный экспоненциальный рост, что говорит о линейной неустойчивости исходного ОДУ. Необходимо однако отметить, что это не говорит о глобальной неустойчивости, но вопрос глобальной неустойчивости в общем случае является трудно разрешимым. Кроме того, мы рассматриваем поведение функции на некотором шаге t_i, и нас интересует поведение ОДУ в некоторой окрестности t_i. Таким образом, необходимым условием для линейной устойчивости как ОДУ, так и численной схемы, является $k_y < 0$ для любого $t_i \in [a; b]$. Более того, так как правая часть линеаризованного уравнения (7.134) не влияет на линейную устойчивость решения, нам достаточно рассмотреть следующее ОДУ:

$$y' - ky = 0, \tag{7.143}$$

где мы опустили индекс y, т.е. $k = k_y$. Применение численной схемы к этому уравнению позволяет установить линейную устойчивость численной схемы, также называемую A-устойчивостью. Для того, чтобы также проанализировать колебательные неустойчивости схемы, необходимо кроме того предположить, что $k \in \mathbb{C}$.

Рассмотрим анализ A-устойчивости на примере явного и неявного методов Эйлера. Рекуррентное соотношение для явного метода Эйлера имеет вид:

$$w_{i+1} = w_i + hf(t_i, w_i). \tag{7.144}$$

Применение метода к линейному ОДУ (7.143) дает:

$$\begin{aligned} & w_{i+1} = w_i + hkw_i \\ \Longrightarrow \quad & w_{i+1} = (1 + hk)w_i. \end{aligned} \tag{7.145}$$

Подобное рекуррентное соотношение стремиться к неподвижной точке (т.е. является устойчивым) тогда, когда $|1 + hk| < 1$ (это очевидно, если вы проинтерпетируете это соотношение в контексте метода простой итерации). Так как k является комплексным числом, область устойчивости будет кругом в координатах $\Re(hk)$, $\Im(hk)$, которую принято изображать в виде соответствующей hk-диаграммы, представленной на левом графике рисунка 7.2 (штриховка обозначает область устойчивости). Подобная устойчивость называется *условной устойчивостью*, так как при фиксированном k величина шага h строго ограничена областью устойчивости. Заметим, что $\Re(k) < 0$ (левая половина графика), является необходимым условием для устойчивости любой схемы и независит от самой схемы.

Теперь рассмотрим неявный метод Эйлера:

$$\begin{aligned} & w_{i+1} = w_i + hf(t_i, w_{i+1}) \\ \Longrightarrow \quad & w_{i+1} = w_i + hkw_{i+1} \\ \Longrightarrow \quad & w_{i+1} = \frac{1}{1 - hk} w_i. \end{aligned} \tag{7.146}$$

Область устойчивости в этом случае определяется неравенством $|1-hk| > 1$. Соответствующая область устойчивости с учетом необходимого условия $\Re(k) < 0$ изображена на правом графике рисунка 7.2. Легко заменить, что неявный метод Эйлера будет устойчивым для любых h. Подобная устойчивость называется *абсолютной устойчивостью*. Иными словами, всегда, когда само исходное ОДУ является линейно устойчивым, неявный метод Эйлера, примененный к этому ОДУ, будет так же устойчивым в независимости от шага h.

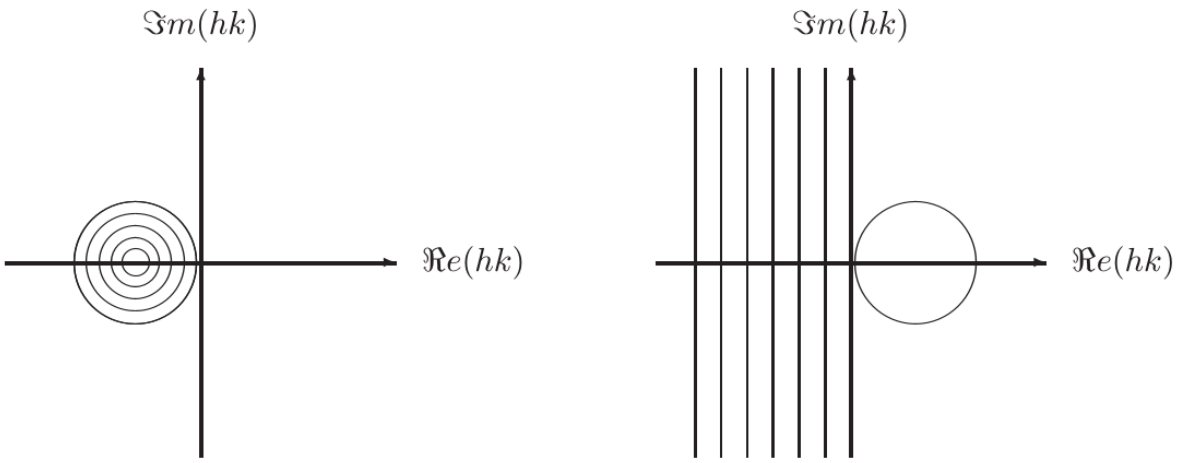

Рисунок 7.2 – Левый график демонстрирует hk-диаграмму для явного метода Эйлера, где окружность соответствует равенству $|1 + hk| = 1$. Правый график демонстрирует hk-диаграмму для неявного метода Эйлера, где окружность соответствует равенству $|1 - hk| = 1$. На обоих графиках область устойчивости численного метода обозначена штриховкой.