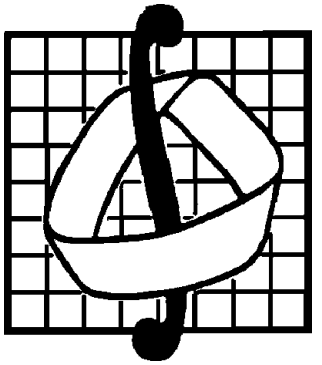
МГУ имени М.В. Ломоносова
Механико-математический факультет

А.А. Корнев

Лекции по курсу
"ЧИСЛЕННЫЕ МЕТОДЫ"

Издательство попечительского совета механико-математического факультета МГУ

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА



Механико-математический факультет

Лекции по курсу
"Численные методы"

А.А. Корнев

Москва 2016 год

УДК 519.6(075.8)

А.А. Корнев

Лекции по курсу "Численные методы"

Издание второе, переработанное

Предлагаемые лекции по численным методам читаются в качестве обязательного годового курса для студентов экономического отделения механико-математического факультета МГУ имени М.В. Ломоносова. Рассматриваются вычислительные алгоритмы решения конечно-разностных, обыкновенных дифференциальных и нелинейных алгебраических уравнений, методы приближенного интегрирования. Особое внимание уделяется численным задачам линейной алгебры и интерполяции функций.

Для студентов и аспирантов, изучающих и применяющих методы вычислительной математики, а также для преподавателей, читающих лекции и проводящих семинарские занятия.

Рецензент — профессор Г.М. Кобельков.

Допущено УМО по классическому университетскому образованию в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению подготовки 010100 Математика.

ISBN 978-5-4294-0010-05

Оглавление

1 Математический аппарат **7**
Лекция 1. Погрешность. Разностные уравнения 7
Лекция 2. Частные решения и собственные функции 13
Лекция 3. Многочлены Чебышёва 18

2 Дифференциальные уравнения **26**
Лекция 4. Теория разностных схем 26
Лекция 5. Численное дифференцирование и задача Коши . . 31
Лекция 6. Методы решения задачи Коши 35
Лекция 7. Оценка глобальной погрешности 40
Лекция 8. Уравнения второго порядка 46
Лекция 9. Методы решения уравнений второго порядка 52

3 Линейная алгебра **59**
Лекция 10 Векторные и матричные нормы 59
Лекция 11. Точные методы решения слау 63
Лекция 12. Линейная задача наименьших квадратов 69
Лекция 13. Линейная знк с ограничениями 76
Лекция 14. Итерационные методы решения слау 80
Лекция 15. Вариационные методы решения слау 86
Лекция 16. Проекционные методы решения слау 93
Лекция 17. Задачи на собственные значения 98

4 Приближение функций **107**
Лекция 18. Полиномиальная интерполяция 107
Лекция 19. Задачи наилучшего приближения 110
Лекция 20. Сплайн-интерполяция 114
Лекция 21. Интерполяция Чебышёва, Фурье, Паде 119

5 Численное интегрирование **126**
Лекция 22. Численное интегрирование 126
Лекция 23. Квадратуры Гаусса. Ортогональные многочлены . 131
Лекция 24. Оптимизация квадратур 138
Лекция 25. Интегрирование функций с особенностями 142

6 Нелинейные уравнения **148**
Лекция 26. Методы решения нелинейных уравнений 148
Лекция 27. Решение систем нелинейных уравнений 156

1. Математический аппарат

Лекция 1. Вычислительная погрешность. Линейные разностные уравнения

Покажем, что между математически точными вычислениями и вычислениями с произвольно высокой, но конечной точностью имеется принципиальное отличие. Это приводит к тому, что алгоритмы, традиционно применяемые в точной арифметике, могут некорректно работать при расчетах на ЭВМ. Как следствие, к методам и постановкам задач вычислительной математики предъявляют дополнительные требования.

Первое, решаемая численно задача должна быть устойчива, т.е. малое изменение входных параметров не должно значительно менять результат;

второе, должен быть численно устойчив выбранный алгоритм, т.е. ошибки округления в промежуточных вычислениях не должны искажать окончательный ответ;

третье, имеющиеся вычислительные ресурсы (память, быстродействие, программное обеспечение) должны позволить реализовать алгоритм и получить ответ за требуемое время.

Пример 1.1. Матрица Уилкинсона

$$A = \begin{pmatrix} 20 & 20 & 0 & 0 & \dots & 0 & 0 \\ 0 & 19 & 20 & 0 & \dots & 0 & 0 \\ 0 & 0 & 18 & 20 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 2 & 20 \\ \varepsilon & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

при $\varepsilon = 0$ имеет наименьшее по модулю собственное значение, равное 1. Как оно изменится при $\varepsilon = 20^{-19} \cdot 20! \approx 5 \cdot 10^{-7}$?

Характеристическое уравнение для возмущенной матрицы Уилкинсона имеет вид

$$\det(A - \lambda I) = (20 - \lambda)(19 - \lambda) \cdots (1 - \lambda) - 20^{19} \cdot \varepsilon = 0.$$

Свободный член в этом уравнении равен 0, следовательно, наименьшее собственное значение также равно 0. Таким образом, задачи вычисления собственных чисел и определителя для данной матрицы являются численно неустойчивыми: незначительная погрешность в элементах матрицы может существенно исказить ответ. В данном случае изменение одного элемента

матрицы 20×20 на величину порядка $5 \cdot 10^{-7}$ привело к изменению величины определителя с $20! \approx 2.4 \cdot 10^{18}$ до нуля.

Для сравнения: $5 \cdot 10^{-7}$ км $= 0.5$ мм, 10^{18} км примерно равно диаметру нашей галактики Млечный Путь.

Рассмотрим примеры возможного влияния вычислительной погрешности в известных алгоритмах. Для этого нам потребуются математические основы машинной арифметики.

Наиболее распространенная форма представления действительных чисел в компьютерах — *числа с плавающей точкой*. Множество F чисел с плавающей точкой характеризуется четырьмя параметрами: основанием системы счисления p, разрядностью t и интервалом показателей $[L, U]$. Каждое число x, принадлежащее F, представимо в виде

$$x = \pm \left(\frac{d_1}{p} + \frac{d_2}{p^2} + \ldots + \frac{d_t}{p^t} \right) p^\alpha ,$$

где целые числа $p, \alpha, d_1, \ldots, d_t$ удовлетворяют неравенствам

$$0 \le d_i \le p - 1 , \quad i = 1, \ldots, t ; \qquad L \le \alpha \le U .$$

Часто d_i называют *разрядами*, t — *длиной мантиссы*, α — *порядком числа*. *Мантиссой* (дробной частью) x называют число в скобках. Множество F называют *нормализованным*, если для каждого $x \neq 0$ справедливо $d_1 \neq 0$.

Округление чисел при работе на компьютере с точностью ε — это некоторое отображение fl действительных чисел $\mathbf{R}$ на множество F чисел с плавающей точкой, в том числе удовлетворяющее условию:

для произвольного $y \in \mathbf{R}$ такого, что результат отображения $fl(y) \in F$, $fl(y) \neq 0$, имеет место равенство

$$fl(y) = y \cdot (1 + \eta), \quad |\eta| \le \varepsilon.$$

Отсюда следует, что относительная погрешность не превосходит точности округления ε.

Если результат округления не принадлежит F, то его обычно называют *переполнением* и обозначают ∞.

Если ε — точная верхняя грань для $|\eta|$, то при традиционном способе округления чисел имеем $\varepsilon = \frac{1}{2} p^{1-t}$, при округлении отбрасыванием разрядов $\varepsilon = p^{1-t}$. Величину ε часто называют *машинной точностью*. Машинную точность можно оценить следующим образом: $double\, e = 1.; \; while(1. + e > 1.)\; e = e/2$. При этом следует помнить, что представление чисел в регистрах повышенной точности может иметь большее число разрядов, чем в оперативной памяти, поэтому точность арифметики на таких регистрах может оказаться выше. На распространенных на данный момент моделях

компьютеров число типа *double* занимает 8 байт: 1 бит для знака, 11 бит для показателя, 52 для мантиссы (старший бит мантиссы всегда равен 1 и в памяти не запоминается); диапазон значений: $\sim 2^{\pm 2^{10}} \sim 10^{\pm 308}$, машинная точность: $\sim 10^{-18}$.

Пример 1.2. Пусть отыскивается наименьший корень x_1 уравнения

$$y(x) = (x - 10^{-3})(x - 10^3) = 0 \Leftrightarrow x^2 - 1000.001x + 1 = 0\,.$$

Какая из формул

$$x_1^{(1)} = a - \sqrt{a^2 - 1},\ x_1^{(2)} = \frac{1}{a + \sqrt{a^2 - 1}}, a = 1000.001/2$$

дает более точный результат?

Вторая формула представляет собой результат избавления от иррациональности в числителе первой формулы, во втором случае точность результата значительно выше: при вычислениях в классе *double* для $x_1^{(1)}$ имеем 10 верных цифр, невязка $|y(x_1^{(1)})| \sim 10^{-11}$; для $x_1^{(2)}$ имеем 17 верных цифр, невязка $|y(x_1^{(2)})| \sim 10^{-18}$. В первом случае приходится вычитать близкие числа, что приводит к *эффекту пропадания значащих цифр*, часто существенно искажающему конечный результат вычислений. Абсолютная погрешность также увеличивается, если имеется *деление на малое (умножение на большое)* число. Еще одна опасность — *выход за диапазон допустимых значений* в промежуточных вычислениях (что произойдет, например, после умножения исходного уравнения на достаточно большое число).

Пример 1.3. Пусть вычисляется сумма $\sum_{j=1}^{10^3} j^{-2}$ Какой алгоритм

$$S_0 = 0\,, \quad S_n = S_{n-1} + \frac{1}{n^2}, \quad n = 1, \ldots, 10^3\,,$$

или

$$R_{10^3+1} = 0\,, \quad R_{n-1} = R_n + \frac{1}{n^2}\,, \quad n = 10^3, \ldots, 1\,, \quad \tilde{S}_{10^3} = R_0,$$

следует использовать, чтобы суммарная вычислительная погрешность была меньше?

Следует воспользоваться вторым способом (полученные значения совпадают только в 14-и знаках). При вычислении первым способом происходит потеря точности в результате сложения чисел S_{n-1} и $1/n^2$, существенно отличающихся по величине.

Пример 1.4. Можно ли непосредственными вычислениями проверить, что ряд $\sum_{j=1}^{\infty} \frac{1}{j}$ расходится? Нет.

Задача 1.1. Известно, что $e^x = 1 + x + \frac{x^2}{2!} + \ldots$. Найти e^1, e^{-15}, e^{-20}. Объяснить результат и улучшить алгоритм: $e^{-x} = 1/e^x$.

Пример 1.5. Метод Крамера решения систем линейных алгебраических уравнений с невырожденной матрицей $n \times n$ позволяет найти точное решение, вычислив $(n+1)$ определитель матриц размерности $n \times n$. Оценим

время $T(n)$ работы такого алгоритма (напомним, что нахождение определителя в общем случае является неустойчивой задачей, и уже поэтому метод применять не стоит).

Вычисление одного определителя методом миноров реализуется за $\sim nn!$ арифметических действий, для вычисления $(n+1)$ определителя потребуется $N \sim n(n+1)!$ арифметических действий. Например, для $n = 20,\ 100$ имеем:

$20! \sim 2.4 \cdot 10^{18}$, $N(20) \sim 10^{21}$; $100! \sim 10^{158}$, $N(100) \sim 10^{162}$.

Для сравнения: число Авогадро $\sim 6 \cdot 10^{23}$, поэтому успешное выполнение такого числа арифметических действий маловероятно.

Производительность Fs современных ЭВМ (число арифметических действий в секунду, flops) весьма высока. Например, для персональных компьютеров на базе Intel Core i7 и суперкомпьютера "Ломоносов" (4-я позиция в Top500 за 2010 г.) имеем:

$Fs(\text{"}i7\text{"}) \sim 50 \cdot 10^9$ flops, $Fs(\text{"}L\text{"}) \sim 10^{15}$ flops.

И так как в году порядка $3 \cdot 10^7$ секунд, то можно оценить требуемое для вычислений время:

$T(20, \text{"}i7\text{"}) \sim 6.5 \cdot 10^2$ лет, $T(100, \text{"}L\text{"}) \sim 3 \cdot 10^{139}$ лет.

Для сравнения: на данный момент считается, что возраст вселенной порядка $13 \cdot 10^9$ лет.

В заключение раздела приведем слова классиков вычислительной математики Р.В. Хемминга и Н.С. Бахвалова: "Цель расчетов — понимание, а не числа", "Не следует думать, что совершенное знание математики, численных методов и навыки работы с ЭВМ позволяют сразу решить любую прикладную математическую задачу. Во многих случаях требуется *доводка* методов, приспособление их к решению конкретных задач".

Линейные разностные уравнения n-го порядка. Теория разностных уравнений является важным математическим аппаратом численных методов. В частности, разностные уравнения возникают при численном решении дифференциальных уравнений.

Определение 1.1 *Пусть неизвестная функция y, заданные функции a_i, f являются функциями одного целочисленного аргумента k. Тогда уравнение*

$$a_0(k)y(k) + a_1(k)y(k+1) + \ \ldots\ + a_n(k)y(k+n) = f(k) \ \Leftrightarrow\ Ly = f,$$

при $a_0(k), a_n(k) \neq 0$ называется линейным разностным уравнением n-го порядка. Уравнение $Ly = 0$ называется однородным.

Для однозначного построения решения $y(k)$ обычно достаточно задать значения $y(k_i)$ в некоторых n точках.

Пример 1.6. Найти $S(k) = \sum_{i=1}^{k}(i^2 + i + 1)$. Это приводит к уравнению $S(k) - S(k-1) = k^2 + k + 1$, $S(1) = 3$.

Пример 1.7. Найти $y(k)$ из условий

$$a(k)y(k-1) - c(k)y(k) + b(k)y(k+1) = f(k), \quad 0 < k < N,$$
$$y(0) = y_0,\ y(N) = y_N,$$

если $a(k), b(k), c(k), f(k), \quad 0 < k < N$ заданы.

Теорема 1.1 *Пусть $y^{(1)}$, …, $y^{(n)}$ — произвольные линейно независимые решения однородного разностного уравнения n-го порядка $Ly = 0$. Тогда общее решение данного уравнения может быть записано в виде*

$$y^0(k) = \sum_{i=1}^{n} c_i y^{(i)}(k).$$

Доказательство. Перепишем исходное уравнение в виде

$$y(k+n) = -\sum_{i=0}^{n-1} \frac{a_i(k)}{a_n(k)} y(k+i).$$

Отсюда следует, что если известны $y(k_0)$, …, $y(k_0+n-1)$, то искомая функция $y(k)$ однозначно восстанавливается для $k \geq k_0 + n$. Аналогично, из равенства

$$y(k) = -\sum_{i=1}^{n} \frac{a_i(k)}{a_0(k)} y(k+i)$$

следует, что $y(k)$ однозначно восстанавливается для $k \leq k_0$. Таким образом, если два решения совпадают в n последовательных точках, то они тождественно равны. Пусть $\left\{y^{(i)}(k)\right\}_{i=1}^{n}$ — некоторая линейно независимая система частных решений, и $y(k)$ — произвольное решение однородного уравнения. Тогда n-мерные вектора $\left\{y^{(i)}(k)\right\}_{i=1}^{n}$ при $0 \leq k \leq n-1$ образуют базис в $\mathbf{R}^n$. Следовательно, при $0 \leq k \leq n-1$ мы можем выразить $y(k) = \sum_{i=1}^{n} c_i y^{(i)}(k)$. А так как $y(x)$ и $\sum_{i=1}^{n} c_i y^{(i)}(k)$ являются решениями $Ly = 0$, совпадающими при $0 \leq k \leq n-1$, то $y(k) \equiv \sum_{i=1}^{n} c_i y^{(i)}(k)$ для всех k.

Теорема 1.2 *Пусть $y^0(k)$ — общее решение линейного однородного уравнения $Ly = 0$, а $y^1(k)$ — частное решение неоднородного уравнения $Ly = f$. Тогда общее решение уравнения $Ly = f$ можно представить в виде суммы*

$$y(k) = y^0(k) + y^1(k)\,.$$

Доказательство. Действительно, разность произвольного решения $y(k)$ неоднородного уравнения и некоторого частного решения неоднородного уравнения $y^1(k)$ является решением однородного уравнения и, как следует из теоремы 1, может быть представлено в виде суммы линейно независимых частных решений однородной задачи: $y(k) - y^1(k) = \sum_{i=1}^{n} c_i y^{(i)}(k) =: y^0(k)$.

Линейные уравнения n-го порядка с постоянными коэффициентами. Для многих важных приложений класс рассматриваемых линейных уравнений можно сузить.

Определение 1.2 *Линейное уравнение*

$$a_0 y(k) + a_1 y(k+1) + \cdots + a_n y(k+n) = f(k)$$

где a_i $(i = 0, 1, \ldots, n)$ — постоянные коэффициенты и $a_0 \neq 0$, $a_n \neq 0$, называется линейным разностным уравнением n-го порядка с постоянными коэффициентами.

Если в этом уравнении положить $y(k+i) = y_{k+i}$ и $f(k) = f_k$, то уравнение примет вид

$$a_0 y_k + a_1 y_{k+1} + \cdots + a_n y_{k+n} = f_k.$$

Для однозначного определения решения требуется задать n условий, например, $y_i = b_i$, $i = 0, \ldots, n-1$. Имеется глубокая аналогия между рассмотренным разностным уравнением и обыкновенным дифференциальным уравнением с постоянными коэффициентами

$$\tilde{a}_0 y(x) + \tilde{a}_1 y'(x) + \cdots + \tilde{a}_n y^{(n)}(x) = \tilde{f}(x)$$

как в постановках задач, так и в методах их решения.

Будем искать решение однородного разностного уравнения в виде $y_k = \mu^k$. После подстановки этого выражения в разностное уравнение и сокращения на μ^k получим характеристическое уравнение $p(\mu) = \sum_{j=0}^{n} a_j \mu^j = 0$.

Утверждение 1.1 *Пусть $\mu_1, \ldots, \mu_r$ — различные корни характеристического уравнения, а $\sigma_1, \ldots, \sigma_r$ — их кратности. Тогда общее решение однородного разностного уравнения представляется в виде*

$$y_k = c_{11}\mu_1^k + c_{12}k\mu_1^k + \cdots + c_{1\sigma_1}k^{\sigma_1-1}\mu_1^k + \cdots +$$
$$+c_{r1}\mu_r^k + c_{r2}k\mu_r^k + \cdots + c_{r\sigma_r}k^{\sigma_r-1}\mu_r^k\,,$$

где c_{ij} — произвольные постоянные.

Таким образом, каждому корню μ кратности σ соответствует набор частных решений вида

$$\mu^k, k\mu^k, \ldots, k^{\sigma-1}\mu^k\,.$$

Пример 1.8. Найти общее решение уравнения

$$by_{k+1} - cy_k + ay_{k-1} = 0\,.$$

Найдем корни характеристического уравнения $b\mu^2 - c\mu + a = 0$:

$$\mu_{1,2} = \frac{c \pm \sqrt{D}}{2b}\,, \quad D = c^2 - 4ab$$

и рассмотрим три случая для дискриминанта (в последующих формулах C_1, C_2 — произвольные постоянные).

а) $D > 0$, $\mu_1 \neq \mu_2$ — вещественные: $y_k = C_1\mu_1^k + C_2\mu_2^k$;

б) $D < 0$, $\mu_{1,2} = \rho e^{\pm \mathbf{i}\varphi}$ — комплексно сопряженные. Здесь

$$\rho = \sqrt{\frac{a}{b}}, \quad \varphi = \begin{cases} \operatorname{arctg}\dfrac{\sqrt{|D|}}{c} & c/b > 0, \\ \pi - \operatorname{arctg}\dfrac{\sqrt{|D|}}{c} & c/b < 0, \\ \dfrac{\pi}{2} & c = 0. \end{cases}$$

Тогда $\mu_{1,2}^k = \rho^k(\cos k\varphi \pm \mathbf{i}\sin k\varphi)$. Следовательно, $y_k = \rho^k(C_1\cos k\varphi + C_2\sin k\varphi)$. Это форма записи действительного решения, для комплексного можно использовать предыдущий вид, т.е. $y_k = C_1\mu_1^k + C_2\mu_2^k$.

в) $D = 0$, $\mu_1 = \mu_2 = \mu$ — кратные: $y_k = C_1\mu^k + C_2 k\mu^k$.

Лекция 2. Частное решение неоднородного уравнения. Фундаментальное решение. Задачи на собственные значения

Как и в случае дифференциальных уравнений, частное решение разностного уравнения для правой части специального вида может быть найдено методом неопределенных коэффициентов.

Утверждение 2.1 *Пусть*

$$f_k = \alpha^k\left(P_{m_1}(k)\,\cos\beta k + \tilde{P}_{m_2}(k)\,\sin\beta k\right),$$

где $P_{m_1}(k)$, $\tilde{P}_{m_2}(k)$ — многочлены степени m_1 и m_2 соответственно. Тогда частное решение может быть найдено в виде

$$y_k^1 = k^s\,\alpha^k\left(Q_n(k)\,\cos\beta k + \tilde{Q}_n(k)\,\sin\beta k\right), \qquad (1)$$

где $s = 0$, если $\alpha e^{\pm\mathbf{i}\beta}$ не являются корнями характеристического уравнения, и s равно кратности корня в противном случае; $n = \max(m_1, m_2)$ — степень многочленов $Q_n(k)$ и $\tilde{Q}_n(k)$.

Чтобы найти коэффициенты многочленов $Q_n(k)$ и $\tilde{Q}_n(k)$, надо подставить выражение (1) в неоднородное уравнение и приравнять коэффициенты при подобных членах.

Пример 2.1. Найти вид частного решения уравнения

$$y_{k+2} + y_k = \cos\frac{\pi}{2}k.$$

Найдем корни характеристического уравнения $\mu = \pm\mathbf{i}$. Так как $(\alpha, \beta) = (1, \pi/2)$ является корнем кратности 1, следовательно, $s = 1$, и решение ищется в виде: $y_k^1 = k\left(c_1\cos\frac{\pi}{2}k + c_2\sin\frac{\pi}{2}k\right)$.

Фундаментальное решение. Рассмотрим метод построения частного решения с произвольной правой частью.

Определение 2.1 *Фундаментальным решением G_k называется решение разностного уравнения*

$$a_0y_k + a_1y_{k+1} + \cdots + a_ny_{k+n} = f_k$$

с правой частью $f_k = \delta_k^0$, где $\delta_k^n = \begin{cases} 0, & k \neq n \\ 1, & k = n. \end{cases}$

Пример 2.2. Построим ограниченное фундаментальное решение уравнения первого порядка

$$a\,y_k + b\,y_{k+1} = \delta_k^0\,.$$

Для этого найдем общее решение по известной схеме: $y_k = y_k^0 + y_k^1$, $y_k^0 = C\left(-\frac{a}{b}\right)^k$. Для определения y_k^1 имеем три группы уравнений:

$$\begin{cases} a\,y_k + b\,y_{k+1} = 0 & \text{при} \quad k \leq -1, \quad \text{т.е. } y_k^1 = c_-\left(-\frac{a}{b}\right)^k, \\ a\,y_0 + b\,y_1 = 1 & \text{при} \quad k = 0, \\ a\,y_k + b\,y_{k+1} = 0 & \text{при} \quad k \geq 1, \quad \text{т.е. } y_k^1 = c_+\left(-\frac{a}{b}\right)^k. \end{cases}$$

Для $k \leq -1$ возьмем $c_- = 0$. Тогда все уравнения первой группы выполнены и $y_0 = 0$. Из второго уравнения следует, что $y_1 = 1/b$. Отсюда и из третьей группы уравнений имеем $c_+ = -1/a$. Таким образом, получаем частное решение неоднородного уравнения

$$y_k^1 = \begin{cases} 0, & k \leq 0, \\ -\dfrac{1}{a}\left(-\dfrac{a}{b}\right)^k, & k \geq 1. \end{cases}$$

После прибавления к нему общего решения однородного уравнения находим выражение для общего решения:

$$y_k = \begin{cases} C\Big(-\dfrac{a}{b}\Big)^k, & k \le 0, \\ \Big(C - \dfrac{1}{a}\Big)\Big(-\dfrac{a}{b}\Big)^k, & k \ge 1. \end{cases}$$

Обозначим построенное фундаментальное решение через G_k^0. Его ограниченность выражается в виде зависимости постоянной C от величины $|a/b|$:

$$\begin{array}{ll} C = 0, & \text{при } |a/b| < 1, \\ \forall\, C, & \text{при } |a/b| = 1, \\ C = 1/a, & \text{при } |a/b| > 1. \end{array}$$

Теорема 2.1 *Пусть $|a/b| \ne 1$, $|f_k| \le F$, а G_k^n — ограниченное фундаментальное решение уравнения $a\,y_k + b\,y_{k+1} = \delta_k^n$. Тогда ряд*

$$y_k^1 = \sum_{n=-\infty}^{\infty} G_k^n f_n$$

абсолютно сходится и является частным решением уравнения

$$a\,y_k + b\,y_{k+1} = f_k\,.$$

Доказательство. Рассмотрим случай $|a/b| > 1$. Из предыдущего примера следует, что $C = \dfrac{1}{a}$, т.е.

$$G_k^n = \begin{cases} \dfrac{1}{a}\Big(-\dfrac{a}{b}\Big)^{k-n}, & k - n \le 0, \\ 0, & k - n \ge 1. \end{cases}$$

Следовательно,

$$y_k^1 = \sum_{(k-n)\le 0} \frac{1}{a}(-\frac{a}{b})^{k-n} f_n,$$

и имеет место следующая оценка

$$|y_k^1| \le \frac{F}{|a|} \sum_{(n-k)\ge 0} \left|\frac{b}{a}\right|^{n-k} = \frac{F}{|a|}\frac{1}{1 - |b|/|a|} = \frac{F}{|a| - |b|}\,,$$

т.е. ряд для функции y_k^1 является абсолютно сходящимся. Кроме того, ряд является частным решением исходного уравнения:

$$a\,y_k + b\,y_{k+1} = a\sum_{n=-\infty}^{\infty} G_k^n f_n + b\sum_{n=-\infty}^{\infty} G_{k+1}^n f_n =$$

$$= \sum_{n=-\infty}^{\infty} \big(a\,G_k^n + b\,G_{k+1}^n\big)\,f_n = \sum_{n=-\infty}^{\infty} \delta_k^n\,f_n = f_k\,.$$

Случай $|a/b| < 1$ рассматривается аналогичным образом. Теорема доказана.

Отметим, что изложенная техника применима для построения фундаментального решения для уравнения n-го порядка.

Задачи на собственные значения. В некоторых важных для приложений случаях задачу нахождения собственных векторов и собственных чисел заданной матрицы удается решить методами разностных уравнений.

Пример 2.3. Найдем все λ, для которых разностная задача

$$\frac{y_{k+1} - y_{k-1}}{2h} = -\lambda\,y_k, \quad y_0 = y_N = 0, \quad h = 1/N$$

имеет нетривиальные решения.

Перепишем разностное уравнение следующим образом

$$y_{k+1} + 2h\lambda\,y_k - y_{k-1} = 0,$$

и найдем характеристическое уравнение $mu^2 + 2h\lambda\,\mu - 1 = 0$. Его корни: $\mu_{1,2} = -h\lambda \pm \sqrt{1 + h^2\lambda^2}$. Если $\mu_1 = \mu_2 = \mu$, то $y_k = C_1\mu^k + kC_2\mu^k$. И из краевых условий следует, что $C_1 = C_2 = 0$. Если $\mu_1 \ne \mu_2$, то общее решение разностного уравнения имеет вид

$$y_k = C_1\mu_1^k + C_2\mu_2^k,$$

а константы C_1 и C_2 определяются из системы

$$C_1 + C_2 = 0, \quad C_1\mu_1^N + C_2\mu_2^N = 0.$$

Отсюда получаем, что $C_2 = -C_1$ и $C_1(\mu_1^N - \mu_2^N) = 0$, т.е. нетривиальное решение разностной задачи существует тогда и только тогда, когда $\mu_1^N = \mu_2^N$. Следовательно,

$$\frac{\mu_1}{\mu_2} = \exp\left(\mathbf{i}\frac{2\pi n}{N}\right), \quad n = 0, \ldots, N-1\,.$$

Так как $\mu_1\mu_2 = -1$, то $\mu_1^2 = -\exp\left(\mathbf{i}\dfrac{2\pi n}{N}\right)$, т.е.

$$\mu_1 = \mathbf{i}\exp\left(\mathbf{i}\frac{\pi n}{N}\right), \quad \mu_2 = \mathbf{i}\exp\left(-\mathbf{i}\frac{\pi n}{N}\right).$$

Нетривиальные решения исходной задачи соответствуют $n = 1, \ldots, N-1$ и имеют вид

$$y_k^{(n)} = C_1\left(\mu_1^k - \mu_2^k\right) = C_1\mathbf{i}^k\left(\exp\left(\mathbf{i}\frac{\pi kn}{N}\right) - \exp\left(-\mathbf{i}\frac{\pi kn}{N}\right)\right) = \\ = C_1\mathbf{i}^k 2\mathbf{i}\sin\frac{\pi kn}{N} = C\,\mathbf{i}^{k+1}\sin\frac{\pi kn}{N}.$$

Поскольку

$$\mu_1 + \mu_2 = -2h\lambda = \mathbf{i}\left(\exp\left(\mathbf{i}\frac{\pi n}{N}\right) + \exp\left(-\mathbf{i}\frac{\pi n}{N}\right)\right) = 2\mathbf{i}\cos\frac{\pi n}{N},$$

имеем

$$\lambda^{(n)} = -\frac{\mathbf{i}}{h}\cos\frac{\pi n}{N}, \quad n = 1, \ldots, N-1\,.$$

Отметим, что количество различных ненулевых собственных значений равно $N-1$.

Задача 2.1. Провести аналогию с дифференциальной задачей

$$y' = -\lambda\, y\,.$$

Пример 2.4. Найти все λ, для которых разностная задача

$$\frac{y_{k+1} - 2\,y_k + y_{k-1}}{h^2} = -\lambda\, y_k, \quad y_0 = y_N = 0, \quad h = 1/N$$

имеет нетривиальные решения. Найдем характеристическое уравнение разностной задачи:

$$\mu^2 - 2p\mu + 1 = 0, \quad \text{где} \quad p = \big(1 - \frac{h^2}{2}\lambda\big).$$

Его корни: $\mu_{1,2} = p \pm \sqrt{p^2 - 1}$. Если $\mu_1 = \mu_2 = \mu$, то $y_k = C_1\mu^k + kC_2\mu^k$, а из краевых условий следует, что $C_1 = C_2 = 0$. Если $\mu_1 \neq \mu_2$, то общее решение разностного уравнения имеет вид

$$y_k = C_1\mu_1^k + C_2\mu_2^k,$$

а константы C_1 и C_2 определяются из системы

$$C_1 + C_2 = 0,$$
$$C_1\mu_1^N + C_2\mu_2^N = 0.$$

Отсюда получаем, что $C_2 = -C_1$ и $C_1(\mu_1^N - \mu_2^N) = 0$, т.е. нетривиальное решение разностной задачи существует тогда и только тогда, когда $\mu_1^N = \mu_2^N$. Следовательно,

$$\frac{\mu_1}{\mu_2} = \exp\left(\mathbf{i}\frac{2\pi n}{N}\right), \quad n = 0, \ldots, N-1\,.$$

Так как $\mu_1\mu_2 = 1$, то $\mu_1^2 = \exp\left(\mathbf{i}\frac{2\pi n}{N}\right)$, откуда

$$\mu_{1,2} = \exp\left(\pm\mathbf{i}\frac{\pi n}{N}\right).$$

Нетривиальные решения исходной задачи соответствуют $n = 1, \ldots, N-1$ и имеют вид

$$y_k^{(n)} = C_1\left(\mu_1^k - \mu_2^k\right) = C_1\left(\exp\left(\mathbf{i}\frac{\pi kn}{N}\right) - \exp\left(-\mathbf{i}\frac{\pi kn}{N}\right)\right) = \\ = C_1 2\mathbf{i}\sin\frac{\pi kn}{N} = C\,\sin\frac{\pi kn}{N}.$$

Поскольку

$$2p = 2\big(1 - \frac{h^2}{2}\lambda\big) = \mu_1 + \mu_2 = \left(\exp\left(\mathbf{i}\frac{\pi n}{N}\right) + \exp\left(-\mathbf{i}\frac{\pi n}{N}\right)\right) = 2\cos\frac{\pi n}{N},$$

имеем

$$\lambda^{(n)} = \frac{2}{h^2}\left(1 - \cos\frac{\pi n}{N}\right) = \frac{4}{h^2}\sin^2\frac{\pi n}{2N}, \quad n = 1, 2, \ldots, N-1.$$

Отметим, что количество различных ненулевых собственных значений равно $N-1$.

Задача 2.2. Провести аналогию с дифференциальной задачей:

$$y'' = -\lambda\, y,\ y(0) = y(1) = 0,$$

$$y_{(n)}(x) = C\sin(\pi n x),\ \lambda_{(n)} = (\pi n)^2,\ n = 1, \ldots\,.$$

Лекция 3. Многочлены Чебышёва

Найдем общее решение однородного разностного уравнения с параметром:

$$y_{n+1}(x) = 2x\,y_n(x) - y_{n-1}(x), \quad x \in \mathbf{R}.$$

Имеем:

$$\mu^2 - 2x\mu + 1 = 0,\ \mu_{1,2} = x \pm \sqrt{x^2-1}\,.$$

Следовательно, при $x \neq \pm 1$ верна формула

$$y_n(x) = C_1(x)(x + \sqrt{x^2-1})^n + C_2(x)(x - \sqrt{x^2-1})^n.$$

Случай $x = \pm 1$ далее рассматривается отдельно.

Если $|x| < 1$, то удобно перейти к действительной форме записи решения $y_n(x)$, представив $\mu_{1,2}$ в тригонометрическом виде. Пусть $x = \cos\varphi$, тогда:

$$y_n(x) = \hat{C}_1(x)\cos(n\arccos x) + \hat{C}_2(x)\sin(n\arccos x).$$

Значения соответствующих констант определяются начальными условиями, например, для $y_0(x), y_1(x)$.

Многочлены Чебышёва первого рода $T_n(x)$.

Определение 3.1 *Многочленами Чебышёва первого рода $T_n(x)$ называется последовательность многочленов, удовлетворяющих рекуррентному соотношению*

$$T_{n+1}(x) = 2x\,T_n(x) - T_{n-1}(x), \quad T_0(x) = 1, \quad T_1(x) = x\,.$$

Теорема 3.1 *Для многочленов $T_n(x)$ имеет место иррациональное представление*

$$T_n(x) = \frac{\left(x + \sqrt{x^2-1}\right)^n + \left(x - \sqrt{x^2-1}\right)^n}{2}.$$

Доказательство. Пусть $x \neq \pm 1$. Тогда из начальных условий получаем $C_1(x) = C_2(x) = \dfrac{1}{2}$, что приводит к указанной формуле. Для $x = \pm 1$ имеем $\mu_{1,2} = x$, т.е. $T_n(x) = C_1(x)x^n + C_2(x)nx^n$. Из начальных условий получаем $C_1(x) = 1$, $C_2(x) = 0$, т.е $T_n(x)|_{\pm 1} = (\pm 1)^n$. Это совпадает со значениями, получаемыми при формальной подстановке $x = \pm 1$ в иррациональное представление, поэтому, найденная формула верна при всех x. Корректность указанного выражения для $x = \pm 1$ также следует из непрерывности рекуррентного соотношения и иррационального представления для всех $x \in \mathbf{R}$ и их тождественного равенства для $x \neq \pm 1$.

Теорема 3.2 *Для многочленов $T_n(x)$ имеет место тригонометрическая форма*

$$T_n(x) = \cos(n\arccos x),\ |x| \leq 1.$$

Доказательство. Либо из начальных условий, либо подставив $x = \cos\varphi$ в иррациональное представление, получаем $\hat{C}_1(x) = 1$, $\hat{C}_2(x) = 0$. Это приводит к указанной формуле.

Отметим, что тригонометрическую форму можно взять за определение многочлена Чебышёва первого рода. Действительно, при любом η имеем

$$\cos\big((n+1)\eta\big) + \cos\big((n-1)\eta\big) = 2\cos\eta\cos(n\eta).$$

Полагая $\eta = \arccos x$, получаем, что тригонометрическая формула удовлетворяет рекуррентному соотношению и указанным начальным условиям. Отсюда следует, что при $|x| \leq 1$ рекуррентное соотношение и тригонометрическая формула задают один и тот же многочлен.

Можно показать, что тригонометрическая форма верна для произвольного комплексного x. Для этого достаточно воспользоваться формулой Муавра $\cos(n\theta) = \frac{1}{2}((\cos\theta + \mathbf{i}\sin\theta)^n + (\cos\theta - \mathbf{i}\sin\theta)^n)$, где $\sin\theta = \sqrt{1-x^2}$, и выбрать для всех функций подходящие ветви. Отсюда также можно обосновать иррациональное представление для комплексных x.

Многочлены Чебышёва второго рода $U_n(x)$.

Определение 3.2 *Многочленами Чебышёва второго рода $U_n(x)$ называется последовательность многочленов, удовлетворяющих рекуррентному соотношению*

$$U_{n+1}(x) = 2x\,U_n(x) - U_{n-1}(x)\,, \quad U_0(x) = 1\,,\ U_1(x) = 2\,x\,.$$

Теорема 3.3 *Для многочленов $U_n(x)$ имеет место иррациональное представление*

$$U_n(x) = \frac{\left(x + \sqrt{x^2-1}\right)^{n+1} - \left(x - \sqrt{x^2-1}\right)^{n+1}}{2\sqrt{x^2-1}}.$$

Доказательство. Выписав решение разностного уравнения с учетом начальных условий, находим указанную формулу для $x \neq \pm 1$. Из непрерывности и тождественного совпадения рекуррентной последовательности и иррационального представления получаем, что формула также верна для $x = \pm 1$.

Теорема 3.4 *Для многочленов $U_n(x)$ имеет место тригонометрическая форма*

$$U_n(x) = \frac{\sin((n+1)\arccos x)}{\sin(\arccos x)},\ |x| \leq 1.$$

Доказательство. Либо из начальных условий, либо подставив $x = \cos\varphi$ в иррациональное представление, получаем указанный вид.

Свойства многочленов Чебышёва первого рода $T_n(x)$.

Теорема 3.5 *Все многочлены $T_{2n}(x)$ — четные, $T_{2n+1}(x)$ — нечетные; коэффициент при старшем члене x^n равен 2^{n-1}. Для $|x| \le 1$ имеем $\left|T_n(x)\right| \le 1$.*

Доказательство первого утверждения можно получить методом индукции из рекуррентной формы записи a). Второе утверждение следует из тригонометрической формы записи b).

Теорема 3.6

$$I_{mn} = \int\limits_{-1}^{1} \frac{T_n(x)\,T_m(x)}{\sqrt{1-x^2}}\,dx = \begin{cases} 0 & \text{при } n \ne m, \\ \pi/2 & \text{при } n = m \ne 0, \\ \pi & \text{при } n = m = 0. \end{cases}$$

Доказательство. Положим $x = \cos\varphi$, тогда $dx = -\sin\varphi\,d\varphi$ и

$$I_{mn} = \int\limits_{0}^{\pi} \cos(n\varphi)\cos(m\varphi)d\varphi = \frac{\pi}{2}(\delta^0_{n-m} + \delta^0_{n+m})\,.$$

Таким образом, система многочленов Чебышёва ортогональна, а система $\{\frac{1}{\sqrt{\pi}}T_0(x), \frac{\sqrt{2}}{\sqrt{\pi}}T_1(x), \ldots, \frac{\sqrt{2}}{\sqrt{\pi}}T_n(x), \ldots\}$ ортонормальна на отрезке $[-1,1]$ с весом $p(x) = \dfrac{1}{\sqrt{1-x^2}}$.

Теорема 3.7 *Все нули многочленов Чебышёва $T_n(x)$ имеют вид: $x_m = \cos\dfrac{\pi\,(2m-1)}{2n}$, где $m = 1, \ldots, n$ (все нули лежат внутри отрезка $[-1,1]$, их ровно n).*

Доказательство. Искомые нули являются корнями уравнения

$$\cos(n \arccos x_m) = 0.$$

Теорема 3.8 *Все экстремумы многочлена Чебышёва $T_n(x)$ на отрезке $[-1,1]$ имеют вид: $x_{(m)} = \cos\dfrac{\pi m}{n}$, $m = 0, \ldots, n$ (на $[-1,1]$ имеется $n+1$ экстремум и $T_n(x_{(m)}) = (-1)^m$).*

Доказательство. Искомые экстремумы являются корнями уравнения $\cos(n \arccos x_{(m)}) = \pm 1$.

Теорема 3.9 *Для произвольных положительных целых m, n верны тождества*

$$T_m(x)T_n(x) = \frac{1}{2}(T_{m+n}(x) + T_{m-n}(x)),\ m \ge n;$$
$$T_m(T_n(x)) = T_{mn}(x).$$

Теорема 3.10 (Лебедев В.И.) *Для произвольных положительных целых m, n верны тождества*

$$T_{3n}(x) = T_3(T_n(x)) = T_n(x)(2T_2(T_n(x)) - 1) = $$
$$= 4T_n(x)(T_n(x) - \frac{\sqrt{3}}{2})(T_n(x) + \frac{\sqrt{3}}{2});$$
$$U_{mn-1}(x) = U_{m-1}(T_n(x))U_{n-1}(x).$$

Доказательство может быть получено из вида многочленов $T_n(x)$ и $U_n(x)$.

Из теоремы Лебедева в том числе следует, что треть корней $T_{3n}(x)$ являются корнями $T_n(x)$, а корни $U_{mn-1}(x)$ состоят из корней $U_{m-1}(T_n(x))$ и $U_{n-1}(x)$.

Теорема 3.11 *Среди всех многочленов $P_n(x) = x^n + \ldots$ со старшим коэффициентом 1 приведенный многочлен Чебышёва $\overline{T}_n(x) = 2^{1-n}T_n(x)$ наименее уклоняется от нуля на отрезке $[-1,1]$, т.е.*

$$\max_{[-1,1]} \left|P_n(x)\right| \ge \max_{[-1,1]} \left|\overline{T}_n(x)\right| = 2^{1-n}\,.$$

Доказательство. Пусть $\|P_n(x)\|_{C[-1,1]} < 2^{1-n}$. Тогда многочлен $Q_{n-1}(x) = \overline{T}_n(x) - P_n(x)$ имеет степень $n-1$ и отличен от нулевого. Однако при этом в точках $x_{(m)}$ экстремума многочлена Чебышёва знак разности $Q_{n-1}(x)$ определяется знаком $\overline{T}_n(x)$:

$$\text{sign}\,\left(\overline{T}_n(x_{(m)}) - P_n(x_{(m)})\right) = (-1)^m\,.$$

Следовательно, $Q_{n-1}(x)$ является отличным от нуля многочленом степени $n-1$, но имеет n нулей, поскольку $n+1$ раз меняет знак в точках экстремума. Полученное противоречие завершает доказательство.

Таким образом, приведенный многочлен Чебышёва является решением следующей минимаксной задачи:

$$\arg\{\inf_{P_n(x)=x^n+\ldots}\ \max_{x\in[-1,1]} |P_n(x)|\ \} = \overline{T}_n(x).$$

Также можно доказать единственность такого решения, т.е. для произвольного $P_n(x)$ из указанного класса, отличного от $\overline{T}_n(x)$, выполняется оценка: $\|\overline{T}_n(x)\|_{C[-1,1]} < \|P_n\|_{C[-1,1]}$.

Теорема 3.12 *Среди всех многочленов $P_n(x) = x^n + \ldots$ со старшим коэффициентом 1 приведенный многочлен Чебышёва*

$$\overline{T}_n^{[a,b]}(x) = \left(\frac{b-a}{2}\right)^n 2^{1-n} T_n\left(\frac{2x-(b+a)}{b-a}\right)$$

наименее уклоняется от нуля на отрезке $[a,b]$ *, т.е.*

$$\max_{[a,b]}\left|P_n(x)\right| \geq \max_{x\in[a,b]}|\overline{T}_n^{[a,b]}(x)| = (b-a)^n 2^{1-2n}.$$

Доказательство. Сделаем линейную замену переменных $x = \dfrac{a+b}{2} + \dfrac{b-a}{2}x'$ для отображения отрезка $[-1,1]$ в заданный отрезок $[a,b]$. Многочлен $T_n(x)$ при этом преобразуется в многочлен $T_n\left(\dfrac{2x-(b+a)}{b-a}\right)$ со старшим коэффициентом $2^{n-1}(2/(b-a))^n$ и экстремумами в точках $x_{(m)}^{[a,b]} = \dfrac{a+b}{2} + \dfrac{b-a}{2}x_{(m)}$. После перенормировки по старшему коэффициенту и использования предыдущей схемы доказательства имеем, что многочлен

$$\overline{T}_n^{[a,b]}(x) = (b-a)^n 2^{1-2n} T_n\left(\frac{2x-(b+a)}{b-a}\right)$$

является наименее уклоняющимся от нуля на отрезке $[a,b]$ многочленом со старшим коэффициентом 1. Искомая оценка для $\|\overline{T}_n^{[a,b]}\|_{C[a,b]}$ следует из равенства $\max\limits_{x\in[a,b]}|T_n\left(\dfrac{2x-(b+a)}{b-a}\right)| = 1$.

Теорема 3.13 *Среди всех многочленов* $P_n(x)$ *равных* 1 *при* $x_0 = 0$, $x_0 \notin [a,b]$, *приведенный многочлен Чебышёва*

$$\tilde{T}_n^{[a,b]}(x) = \frac{T_n\left(\frac{2x-b-a}{b-a}\right)}{T_n(-\frac{b+a}{b-a})}$$

наименее уклоняется от нуля на отрезке $[a,b]$. *При этом*

$$\max_{x\in[a,b]}|\tilde{T}_n^{[a,b]}(x)| = \frac{2}{q_1^{-n}+q_1^n}, \quad q_1 = \frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}.$$

Доказательство. Оптимальность приведенного многочлена Чебышёва может быть доказана по рассмотренной ранее схеме от противного, т.к. разность $\tilde{T}_n^{[a,b]}(x) - P_n(x)$ отлична от тождественного нуля, является многочленом степени не выше n, имеет n нулей на отрезке $[a,b]$ и ноль в точке $x_0 = 0 \notin [a,b]$.

Найдем величину $\|\tilde{T}_n^{[a,b]}\|_{C[a,b]} = \dfrac{1}{|T_n(-\frac{b+a}{b-a})|}$. Для этого вычислим значение

$$T_n(x) = \frac{(x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n}{2}$$

при $x = -\dfrac{b+a}{b-a}$. Так как

$$-\frac{b+a}{b-a} \pm \sqrt{\left(\frac{b+a}{b-a}\right)^2 - 1} = -\frac{b+a}{b-a} \pm \frac{2\sqrt{ba}}{b-a} =$$

$$= -\frac{(\sqrt{b}\mp\sqrt{a})^2}{(\sqrt{b}-\sqrt{a})(\sqrt{b}+\sqrt{a})} = -q_1^{\pm 1}, \quad q_1 = \frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}},$$

то

$$\|\tilde{T}_n^{[a,b]}\|_{C[a,b]} = \frac{2}{q_1^n + q_1^{-n}} = \frac{2q_1^n}{1+q_1^{2n}} \leq 2q_1^n.$$

Теорема доказана.

Теорема 3.14 *Среди всех многочленов* $P_n(x)$, *удовлетворяющих условию* $\max\limits_{x\in[a,b]}\left|P_n(x)\right| = M$, *приведенный многочлен Чебышёва* $M\,T_n\left(\dfrac{2x-(a+b)}{b-a}\right)$ *принимает наибольшее по модулю значение для всех* $\xi \notin (a,b)$, *т.е.*

$$\left|P_n(\xi)\right| \leq M\left|T_n\left(\frac{2\xi-(a+b)}{b-a}\right)\right|, \ \xi \notin (a,b).$$

Доказательство. Предположим противное, т.е. пусть существует такое $\xi \notin (a,b)$, что $\left|P_n(\xi)\right| > M\left|T_n\left(\dfrac{2\xi-(a+b)}{b-a}\right)\right|$. Тогда у ненулевого полинома

$$Q_n(x) = \frac{P_n(\xi)}{T_n\left(\frac{2\xi-(a+b)}{b-a}\right)} T_n\left(\frac{2x-(a+b)}{b-a}\right) - P_n(x)$$

как минимум $(n+1)$ перемена знака на $[a,b]$ в точках $x_{(m)}^{[a,b]}$, т.е. n нулей на (a,b) и ноль в точке ξ. Противоречие.

Теорема 3.15 (Марков А.А.) *(Обобщенная теорема). Среди всех многочленов* $P_n(x)$, *удовлетворяющих условию* $\max\limits_{x\in[a,b]}\left|P_n(x)\right| = M$, *приведенный многочлен Чебышёва* $M\,T_n\left(\dfrac{2x-(a+b)}{b-a}\right)$ *принимает наибольшее по модулю значение производной для всех* $\xi \notin (a,b)$, *т.е.*

$$\left|P_n'(x)\right|_{x=\xi} \leq M\left|T_n'\left(\frac{2x-(a+b)}{b-a}\right)\right|_{x=\xi}.$$

Равенство достигается только для указанного полинома.

Доказательство. Предположим противное, т.е. пусть существует такое $\xi \notin (a,b)$, что $\left|P_n'(x)\right|_{x=\xi} > M \left|T_n'\left(\frac{2x-(a+b)}{b-a}\right)\right|_{x=\xi}$. Тогда у ненулевого полинома

$$Q_n(x) = \frac{\left|P_n'(x)\right|_{x=\xi}}{\left|(T_n'(\frac{2x-(a+b)}{b-a})\right|_{x=\xi}} T_n\left(\frac{2x-(a+b)}{b-a}\right) - P_n(x)$$

как минимум $(n+1)$ перемена знака на $[a,b]$ в точках $x_{(m)}^{[a,b]}$, т.е. n нулей на (a,b). Но тогда у полинома $Q_n'(x)$ как минимум $n-1$ нуль на (a,b) и нуль в точке ξ. Отсюда следует, что $Q_n'(x) \equiv 0$, т.е. $Q_n(x) \equiv \text{const}$ и имеет n нулей, т.е. $Q_n(x) \equiv 0$. Противоречие. (Единственность экстремального полинома без доказательства).

2. Дифференциальные уравнения

Лекция 4. Теория разностных схем

Пусть в области D с границей Γ задана дифференциальная задача с граничным условием

$$L\,u = f \quad \text{в} \quad D\,, \tag{1}$$

$$l\,u = \varphi \quad \text{на} \quad \Gamma\,. \tag{2}$$

Здесь L и l — дифференциальные операторы; f и φ — заданные элементы, а u — искомый элемент некоторых линейных нормированных пространств F, Φ и U соответственно.

Конечно-разностный метод.

Для применения разностного метода задают некоторую *сетку* — конечное множество точек (узлов) $\overline{D}_h = D_h \bigcup \Gamma_h$, принадлежащее области $\overline{D} = D \bigcup \Gamma$ (как правило, $\Gamma_h \subset \Gamma$), определяют сеточные пространства F_h, Φ_h и U_h и задают операторы проектирования $(\cdot)_h$ элементов исходных пространств на элементы сеточных пространств (часто пространства F_h, Φ_h и U_h определяют как пространства следов функций из F, Φ и U на D_h, Γ_h и $\overline{D}_h$ соответственно, в этом случае каждая непрерывная функция v совпадает с функцией $(v)_h$ в узлах сетки). При этом в пространствах задаются согласованные нормы.

Определение 4.1 *Нормы $\|\cdot\|, \|\cdot\|_h$ называются согласованными, если для произвольной достаточно гладкой функции v выполняется соотношение*

$$\lim_{h\to 0} \|\,(v)_h\,\|_h = \|\,v\,\|\,.$$

Если нормы не являются согласованными, то из условия $\lim_{h\to 0} \|(u)_h\|_h = 0$ не следует $\|u\| = 0$, т.е. что $u \equiv 0$ в исходном пространстве U. Однако, это требование необходимо для обоснования сходимости сеточной функций u_h к непрерывной u. Отметим, что пространства U_h и F_h могут различаться, т.е. различаться узлы, где ищется решение и задана правая часть.

Все производные, входящие в уравнение и краевые условия, заменяются *разностными аппроксимациями*. В результате дифференциальные операторы L и l заменяются разностными L_h и l_h.

Для нахождения приближенного решения задачи (1), (2) определим *разностную схему* — семейство разностных задач, зависящих от параметра h:

$$L_h u_h = f_h \quad \text{в} \quad D_h\,, \tag{3}$$

$$l_h u_h = \varphi_h \quad \text{на} \quad \Gamma_h\,. \tag{4}$$

Решение u_h, называемое *разностным*, принимается в качестве приближенного решения дифференциальной задачи.

Пример 4.1.

$$u'(x) = f(x), u(0) = u^0; u(x), f(x) \in C^m[0,1]; \|v\| = \max_{x\in[0,1]} |v(x)|;$$
$$h = 1/N, x_i = ih, (u)_{U_h} = u(ih), (f)_{F_h} = f(ih + h/2);$$
$$\frac{u_{i+1} - u_i}{h} = f_i,\ f_i = f((i + 1/2)h), i = 0, \ldots, N-1,\ u_0 = u^0;$$
$$u_h = [u_0, \ldots, u_N] \in \mathbf{R}^{N+1},\ f_h = [f_0, \ldots, f_{N-1}] \in \mathbf{R}^N;$$
$$\|u_h\|_h = \max_{0\le i\le N} |u_i|,\ \|f_h\|_h = \max_{0\le i\le N-1} |f_i|.$$

Отметим, что евклидова норма $\|v_h\|_{**} = (\sum_{i=0}^{N} |v_i|^2)^{1/2}$ и норма $\|v_h\|_* = h\|v_h\|_h$ не согласованны с нормой исходной задачи $\|v\|$, т.к. при $h \to 0$ имеем $\|(v)_h\|_* \to 0$, $\|(v)_h\|_{**} \to \infty$ для произвольной ненулевой функции $v \in C[0,1]$.

Сходимость.

Определение 4.2 *Решение u_h разностной схемы (3), (4) сходится к решению u дифференциальной задачи (1), (2), если существуют такие постоянные h_0, c и p, что для всех $h \le h_0$ выполнено неравенство*

$$\|(u)_{U_h} - u_h\|_{U_h} \le c\,h^p,$$

причем c и p не зависят от h. Число p называют порядком сходимости разностной схемы; при этом говорят, что разностное решение u_h имеет порядок точности p.

Наибольший интерес представляет именно сходимость как условие близости полученного численно и точного решений. Проверку сходимости удобно свести к проверке аппроксимации и устойчивости.

Аппроксимация.

Определение 4.3 *Разностная задача (3), (4) аппроксимирует с порядком аппроксимации $p = \min(p_1, p_2)$ дифференциальную задачу (1), (2), если для любых достаточно гладких функций u, f, φ из соответствующих пространств существуют такие постоянные h_0, c_1, p_1, c_2 и p_2, что для всех $h \le h_0$ выполняются неравенства*

$$\|L_h(u)_{U_h} - (Lu)_{F_h}\|_{F_h} + \|(f)_{F_h} - f_h\|_{F_h} \le c_1 h^{p_1},$$
$$\|l_h(u)_{U_h} - (lu)_{\Phi_h}\|_{\Phi_h} + \|(\varphi)_{\Phi_h} - \varphi_h\|_{\Phi_h} \le c_2 h^{p_2},$$

причем c_1, p_1, c_2 и p_2 не зависят от h.

Выражения, стоящие под знаком норм, называют погрешностями аппроксимации.

Определение 4.4 *Оператор L_h из (3) локально аппроксимирует в точке x_i дифференциальный оператор L из (1), если для достаточно гладкой функции $u \in U$ существуют такие положительные постоянные h_0, c и p, не зависящие от h, что при всех $h \le h_0$ справедливо неравенство*

$$\left|\left(\, L_h(u)_{U_h} - (Lu)_{F_h}\,\right)_{x=x_i}\right| \le c\,h^p\,.$$

Число p при этом называется порядком аппроксимации. Аналогично определяется порядок локальной аппроксимации оператора l_h.

Также используется понятие аппроксимации на решении, позволяющее строить схемы более высокого порядка точности на фиксированном шаблоне.

Определение 4.5 *Говорят, что разностная схема (3), (4) аппроксимирует на решении u с порядком аппроксимации $p = \min(p_1, p_2)$ дифференциальную задачу (1), (2), если существуют такие постоянные h_0, c_1, p_1, c_2 и p_2, что для всех $h \le h_0$ выполняются неравенства*

$$\|L_h(u)_{U_h} - f_h\|_{F_h} \le c_1 h^{p_1}, \quad \|l_h(u)_{U_h} - \varphi_h\|_{\Phi_h} \le c_2 h^{p_2},$$

причем c_1, p_1, c_2 и p_2 не зависят от h и выполнены условия нормировки

$$\lim_{h\to 0} \|\, f_h - (f)_{F_h}\,\|_{F_h} = 0\,, \quad \lim_{h\to 0} \|\, \varphi_h - (\varphi)_{\varphi_h}\,\|_{\Phi_h} = 0\,.$$

Лемма 4.1 *Из аппроксимации задач следует аппроксимация на решении с порядком не ниже p.*

Доказательство. Действительно, так как на решении u имеем $(Lu)_{F_h} = (f)_{F_h}$, $(lu)_{\Phi_h} = (\varphi)_{\Phi_h}$, то

$$\|L_h(u)_{U_h} \pm (Lu)_{F_h} - f_h\|_{F_h} \le \|L_h(u)_{U_h} - (Lu)_{F_h}\|_{F_h} + \|(f)_{F_h} - f_h\|_{F_h} \le c_1 h^{p_1},$$
$$\|l_h(u)_{U_h} \pm (lu)_{\Phi_h} - \varphi_h\|_{\Phi_h} \le \|l_h(u)_{U_h} - (lu)_{\Phi_h}\|_{\Phi_h} + \|(\varphi)_{\Phi_h} - \varphi_h\|_{\Phi_h} \le c_2 h^{p_2}.$$

В обратную сторону не верно, т.е. аппроксимация на решении может быть выше аппроксимации задач, т.к. имеется связь $Lu = f$, $lu = \varphi$ между решением u и функциями f, φ.

Устойчивость.

Определение 4.6 *Разностная схема (3), (4) устойчива, если существует такое $h_0 > 0$, что для любого $\varepsilon > 0$ найдется такое $\delta = \delta(\varepsilon)$, что для произвольных функций $u_h^{(i)}$, $i = 1, 2$, являющихся решениями (3), (4), из неравенств*

$$h \le h_0,\ \left\|f_h^{(1)} - f_h^{(2)}\right\|_{F_h} \le \delta,\ \left\|\varphi_h^{(1)} - \varphi_h^{(2)}\right\|_{\Phi_h} \le \delta$$

следует оценка $\left\|u_h^{(1)} - u_h^{(2)}\right\|_{U_h} \le \varepsilon$.

Определение 4.7 *Линейная схема устойчива, если существуют такое h_0 и такие постоянные c_1 и c_2, что для произвольных решений $u_h^{(i)}$, $i = 1, 2$ задачи (3), (4) при всех $h \le h_0$ верна оценка*

$$\left\| u_h^{(1)} - u_h^{(2)} \right\|_{U_h} \le c_1 \left\| f_h^{(1)} - f_h^{(2)} \right\|_{F_h} + c_2 \left\| \varphi_h^{(1)} - \varphi_h^{(2)} \right\|_{\Phi_h}.$$

Для случая линейных задач разностная схема представляет собой систему линейных алгебраических уравнений. И если $f_h^{(1)} = f_h^{(2)}$, $\varphi_h^{(1)} = \varphi_h^{(2)}$, то из условия устойчивости имеем $\|u_h^{(1)} - u_h^{(2)}\|_h \le 0$. Т.е. с нулевой правой частью задача имеет единственное решение, следовательно, по теореме Кронекера–Капелли задача однозначно разрешима при произвольных правых частях. Таким образом, для линейных уравнений устойчивость влечет корректность — задача имеет единственное решение при всех правых частях, и малые изменения в f_h, φ_h приводят к малым изменениям в решении.

Непрерывную зависимость по f_h (равномерную относительно h) называют устойчивостью *по правой части*, а непрерывную зависимость по φ_h называют устойчивостью *по граничным условиям*.

Найдем константу C в условии устойчивости. Представим задачу (3), (4) в виде системы линейных уравнений $A_h u_h = b_h$:

$$\begin{pmatrix} L_h \\ l_h \end{pmatrix} u_h = \begin{pmatrix} f_h \\ \varphi_h \end{pmatrix} \Rightarrow \|u_h^{(1)} - u_h^{(2)}\|_h \le \|(A_h)^{-1}\|_h \|b_h^{(1)} - b_h^{(2)}\|_h,$$

т.е. можно взять $C \ge \|(A_h)^{-1}\|_h$. Задача устойчива, если норма оператора $\|(A_h)^{-1}\|_h$ остается ограниченной при $h \to 0$.

Пример 4.2. "Разностная схема" $h \cdot u_h = f_h$ неустойчива.

Теорема 4.1 (Филиппов А.Ф.) *(О связи аппроксимации, устойчивости и сходимости). Пусть выполнены следующие условия:*

1) *операторы L, l и L^h, l^h — линейные;*

2) *решение и дифференциальной задачи* (1), (2) *существует и единственно;*

3) *разностная схема* (3), (4) *аппроксимирует на решении дифференциальную задачу* (1), (2) *с порядком p;*

4) *разностная схема* (3), (4) *устойчива.*

Тогда решение разностной схемы u_h сходится к решению и дифференциальной задачи с порядком не ниже p.

Доказательство. Рассмотрим следующие задачи

$$\begin{cases} L_h u_h = f_h, \\ l_h u_h = \varphi_h; \end{cases} \qquad \begin{cases} L_h(u)_{U_h} = f_h + (L_h(u)_{U_h} - f_h), \\ l_h(u)_{U_h} = \varphi_h + (l_h(u)_{U_h} - \varphi_h). \end{cases}$$

Так как разностная схема устойчива, то по определению для линейных задач имеем

$$\|u_h - (u)_{U_h}\|_{U_h} \le c_1 \|L_h(u)_{U_h} - f_h\|_{F_h} + c_2 \|(l_h(u)_{U_h} - \varphi_h)\|_{\Phi_h}.$$

И так как разностная схема аппроксимирует на решении дифференциальную задачу, то

$$\|u_h - (u)_{U_h}\|_{U_h} \le ch^p.$$

Это неравенство по определению означает сходимость с порядком p. Теорема доказана.

Отметим, что в классе задач с решениями конечной гладкости требование устойчивости является необходимым условием сходимости.

Метод неопределенных коэффициентов построения разностных схем. Построим для уравнения $u^{(m)}(x) = f(x)$ разностную схему $L_h u_h = f_h$ на сетке с узлами $\{x_i\}$. Пусть $f_h = (f)_{F_h}$. Тогда правая часть аппроксимируется точно. Оператор $L_h u_h$ будем искать в виде $\frac{1}{h^m} \sum_{i=0}^{n} c_i u_i$ на шаблоне x_i, а коэффициенты определим из условия $|(Lu)_{F_h} - L_h(u)_{U_h}| \le ch^p$, т.е.

$$|u^{(m)}(x_k) - \frac{1}{h^m} \sum_{i=0}^{n} c_i u(x_i)| \le ch^p$$

с наивысшим p. Представим $x_i = x_k + \alpha_i h$ и разложим каждое $u(x_i)$ в ряд Тейлора относительно узла сетки x_k, будем иметь $u(x_k + \alpha_i h) = u(x_k) + u'(x_k)\alpha_i h + \ldots + O(h^{p+m})$. Подставив данные разложения в условие аппроксимации и приравняв коэффициенты при соответствующих степенях h, получим систему линейных уравнений относительно неизвестных c_i. При этом константа c в оценке будет зависеть от величины $\|u^{(s)}\|_C$ для некоторого s.

Пример 4.3. Найти коэффициенты разностного оператора при наибольшем p:

$$u'(x_k) = \frac{1}{h}(c_1 u(x_k - h) + c_2 u(x_k) + c_3 u(x_k + h)) + O(h^p).$$

Подставим ряд Тейлора $u(x_k \pm h) = u(x_k) \pm h u'(x_k) + \frac{h^2}{2} u''(x_k) \pm \frac{h^3}{6} u'''(\xi_\pm)$ в схему и приведем подобные слагаемые. Это приводит к системе уравнений

$$\frac{1}{h}(c_1 + c_2 + c_3) = 0,\ \frac{1}{h}(c_3 h - c_1 h) = 1,\ \frac{1}{h}(\frac{h^2}{2} c_1 + \frac{h^2}{2} c_3) = 0.$$

Ее решение имеет вид $c_3 = -c_1 = \frac{1}{2}$, $c_2 = 0$, следовательно, $L_h u_h|_{x_k} = \frac{u_{k+1} - u_{k-1}}{2h}$. При этом константа в оценке погрешности находится из условия:

$$\left| \frac{1}{h} \left(\frac{h^3}{6} c_3 u'''(\xi_+) - \frac{h^3}{6} c_1 u'''(\xi_-) \right) \right| \le ch^2.$$

Отсюда в том числе следует, что найденная схема *точна* для произвольного многочлена второй степени. Таким образом, систему уравнений на коэффициенты можно найти из условия точности формулы разностного дифференцирования для многочленов наиболее высокой степени. Для этого подставляем последовательно $u(x) = 1, x, x^2, \dots$ в разностную формулу и приравниваем к точному значению производной $u^{(m)}(x)$. Решение полученной линейной системы определяет те же коэффициенты схемы.

Лекция 5. Погрешность формул численного дифференцирования. Задача Коши

Рассмотрим задачу численного нахождения значения производной в условиях приближенных вычислений:

$$u^{(m)}(x) = \frac{1}{h^m}\sum_{i=0}^{n} c_i u(x_i) + ch^p = D^m u(x) + E_1.$$

Пусть известны только приближенные значения $\tilde{u}(x_i)$ и $u(x_i) - \tilde{u}(x_i) = \varepsilon_i$. Тогда

$$u^{(m)}(x) = \frac{1}{h^m}\sum_{i=0}^{n} c_i \tilde{u}(x_i) + \frac{1}{h^m}\sum_{i=0}^{n} c_i\varepsilon_i + E_1 = D^m\tilde{u}(x) + E_2 + E_1.$$

Величина $D^m\tilde{u}(x)$ равна вычисленному приближению к $u^{(m)}(x)$, величина $E_2 + E_1 = \sum\limits_{i=0}^{n} \frac{c_i\varepsilon_i}{h^m} + ch^p$ дает погрешность. Пусть

$$|E_1 + E_2| \le \frac{A\varepsilon}{h^m} + ch^p = E(h),$$

где $\sum_{i=0}^{n} |c_i| \le A$, $|\varepsilon_i| \le \varepsilon$. Нас интересует минимизация погрешности $E(h)$ за счет выбора h. В точке минимума h_0 имеем $E'(h_0) = 0$. Отсюда следует, что

$$-m\frac{\varepsilon A}{h_0^{m+1}} + cph_0^{p-1} = 0,\ h_0^{p+m} = \frac{\varepsilon Am}{cp},\ h_0 = \left(\frac{\varepsilon Am}{cp}\right)^{\frac{1}{m+p}}.$$

Таким образом, $h_0 \sim \varepsilon^{\frac{1}{m+p}}, E_0 \sim \left(\varepsilon^{1-\frac{m}{m+p}} + \varepsilon^{\frac{p}{m+p}}\right) \sim \varepsilon^{\frac{p}{m+p}}$.

Пример 5.1. Известно, что

$$u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + ch^2.$$

Пусть погрешность ε вычисления функции u не превосходит 10^{-4}. Тогда при $h \sim 10^{-4}$ имеем $E \sim 10^4$. Оптимальный шаг $h_0 \sim 10^{-1}$ дает $E_0 \sim 10^{-2}$.

Конечно-разностный метод решения задачи Коши. Для задачи

$$y'(x) = f(x, y(x)),\ y(x_0) = y^0,$$
$$y \in C^{(m)}[x_0, x_0 + X],\ \|y\| = \max\nolimits_{x\in[x_0,x_0+X]} |y(x)|$$

построим следующую разностную схему

$$\frac{1}{h}\sum_{i=0}^{n} a_{-i}y_{k-i} = \sum_{i=0}^{n} b_{-i}f(x_{k-i}, y_{k-i}),$$
$$y_h = \{y_i\}_0^N, x_i = x_0 + ih, h = \tfrac{X}{N},\ \|y_h\|_{U_h} = \max_{0\le i\le N} |y_i|$$

с некоторыми начальными условиями $y_0, y_1, \dots, y_{n-1}$.

Если $a_0 \ne 0$, $b_0 \ne 0$, то схема называется неявной; если $a_0 \ne 0$, $b_0 = 0$, то явной; если $a_0 = 0$, $b_0 \ne 0$, то с забеганием вперед.

Напомним, что разностная задача аппроксимирует дифференциальную на решении на отрезке $[x_0, x_0 + X]$ с порядком p, если для функции погрешности

$$r_h = \frac{1}{h}\sum_{i=0}^{n} a_{-i}y(x_{k-i}) - \sum_{i=0}^{n} b_{-i}f\big(x_{k-i}, y(x_{k-i})\big)$$

выполняется оценка $\|r_h\|_{F_h} \le ch^p$ и имеется нормировка по правым частям, т.е. $\|f_h - (f)_{F_h}\|_{F_h} \to 0$ при $h \to 0$. Коэффициенты a_{-i}, b_{-i} формально определяются с точностью до множителя, указанная нормировка устраняет такой произвол и означает, что правая часть разностного уравнения аппроксимирует правую часть дифференциального уравнения.

В общем случае поиск коэффициентов a_{-i}, b_{-i} сводится к нетривиальной нелинейной системе, поэтому аппроксимация левой и правой частей дифференциального уравнения может рассматриваться отдельно. В этом случае коэффициенты a_{-i}, b_{-i} определяются из условий аппроксимации задачи

$\|L_h(y)_{U_h} - (Ly)_{F_h}\|_{F_h} \le c_1h^p$, $\|f_h - (f)_{F_h}\|_{F_h} \le c_2h^p$.

В данном разделе мы не рассматриваем методы построения краевых условий $y_0, \dots, y_{n-1}$ с требуемым порядком точности. Отметим также, что решение задачи Коши по сути сводится к интегрированию функции f, поэтому результаты о погрешности формул численного дифференцирования не применимы.

Условия аппроксимации p-го порядка. Рассмотрим задачу нахождения аппроксимации высокого порядка для уравнения $y'(x) = f(x)$ следующей схемой:

$$\frac{1}{h}\sum_{i=0}^{n} a_{-i}y_{k-i} = \sum_{i=0}^{n} b_{-i}f_{k-i}.$$

Выпишем ряд Тейлора для решения задачи y и правой части $f = y'$:

$$y(x_k - ih) = \sum_{s=0}^{p} y^{(s)}(x_k)\frac{(-ih)^s}{s!} + O(h^{p+1}),$$

$$f(x_k - ih) = y'(x_k - ih) = \sum_{s=1}^{p} y^{(s)}(x_k)\frac{(-ih)^{s-1}}{(s-1)!} + O(h^p).$$

Тогда для функции погрешности имеем: $r_h = \frac{1}{h}y(x_k)E_0 + h^0 y'(x_k)E_1 + h^1 y''(x_k)E_2 + \ldots + h^{p-1}y^{(p)}(x_k)E_p + O(h^p)$, где

$$E_0 = \sum_{i=0}^{n} a_{-i}, \quad E_1 = -\sum_{i=0}^{n} a_{-i}i - \sum_{i=0}^{n} b_{-i},$$

$$E_s = \sum_{i=0}^{n}\left(a_{-i}\frac{(-i)^s}{s!} - b_{-i}\frac{(-i)^{s-1}}{(s-1)!}\right),\ s = 2, \ldots, p.$$

Данные соотношения и условие нормировки $\sum\limits_{i=0}^{n} b_{-i} = 1$ для правой части, означающее, что $\|(f)_h - f_h\|_{F_h} \to 0$ при $h \to 0$, позволяют сформулировать следующее утверждение.

Теорема 5.1 *Для задачи $y'(x) = f(x)$ условия аппроксимации на решении p-го порядка имеют вид:*

$$\sum_{i=0}^{n} a_{-i} = 0,\ \sum_{i=0}^{n} b_{-i} = 1, \sum_{i=0}^{n} a_{-i}i = -1,$$
$$\sum_{i=0}^{n}(a_{-i}i + b_{-i}s)i^{s-1} = 0,\ s = 2, \ldots, p.$$

Первая группа (три уравнения) образует необходимые и достаточные условия аппроксимации (т.е. $p = 1$). Всего имеется $p + 2$ уравнения и $2n + 2$ неизвестных, для корректности $p = 2n$, при этом возможная точность $O(h^{2n})$.

Однако, методы с наивысшим порядком аппроксимации обычно неустойчивы, поэтому практически не применимы для расчетов. В настоящее время используют алгоритмы при $a_{-i} = 0$, $i \geq 2$ (в этом случае производная аппроксимируется по двум точкам). Для остальных схем обычно отсутствует сходимость даже для класса бесконечно гладких правых частей и при отсутствии погрешности округления, а если сходимость и наблюдается, то погрешность растет как e^{CX}. При этом аппроксимация производной по большому числу точек влечет проблему начальных условий.

Пример 5.2. Схемы Адамса с порядком аппроксимации на решении $O(h^2)$,

явная схема: $\dfrac{y_k - y_{k-1}}{h} = \dfrac{3}{2}f_{k-1} - \dfrac{1}{2}f_{k-2}$,

неявная схема: $\dfrac{y_k - y_{k-1}}{h} = \dfrac{1}{2}(f_k + f_{k-1})$.

Устойчивость разностных схем. Для задачи Коши принято проверять более слабое условие α-устойчивости.

Определение 5.1 *Разностная схема $\dfrac{1}{h}\sum\limits_{i=0}^{n} a_{-i}y_{k-i} = f_k$ для задачи $y'(x) = f(x)$ называется α-устойчивой, если все корни соответствующего характеристического многочлена однородного разностного уравнения принадлежат единичному кругу и на границе нет кратных корней.*

Можно показать, что для любой разностной схемы, не удовлетворяющей условию α-устойчивости, существует дифференциальное уравнение с бесконечно дифференцируемой правой частью, для которого даже при отсутствии округлений и погрешности в начальных данных решение разностной задачи не стремится к решению дифференциальной при измельчении шага. Условие α-устойчивости для широкого класса задач обеспечивает сходимость, но константа в оценке близости решений может увеличиваться при увеличении длины отрезка X. Поэтому всегда полезно проверить свойства выбранной разностной схемы для модельного уравнения.

Пример 5.3. Для задачи $y' + y = 0$, $y(0) = 1$ с точным решением $y(x) = e^{-x}$ рассмотрим различные схемы, имеющие второй порядок аппроксимации и являющиеся α-устойчивыми. Отметим, что равенства $y_0 = 1$, $y_1 = 1 - h$ аппроксимируют решение $y(x)$ исходной задачи в точках $x = 0, h$ со вторым порядком. Нас интересует качественная близость решений разностной и дифференциальной задач на большом отрезке $[0, X]$, $X = Nh$.

Первая схема:

$$\begin{cases} \dfrac{y_{k+1} - y_k}{h} + \dfrac{y_{k+1} + y_k}{2} = 0; \quad \mu = \dfrac{1 - h/2}{1 + h/2}, \quad |\mu| < 1, \\ y_0 = 1. \end{cases}$$

При малых h решение разностной задачи $y_k = \mu^k$ качественно похоже на решение дифференциальной задачи: y_k монотонно убывает и $\lim\limits_{h\to 0} y_N = \lim\limits_{h\to 0} \mu^{X/h} = e^{-X} = y(x_N)$.

Вторая схема:

$$\begin{cases} \dfrac{y_{k+1} - y_{k-1}}{2h} + y_k = 0; \quad \mu_{1,2} = -h \pm \sqrt{1 + h^2}, \quad \max\limits_{i=1,2}|\mu_i| > 1, \\ y_0 = 1, y_1 = 1 - h. \end{cases}$$

В данном случае у задачи имеется растущее по модулю (но ограниченное на $[0, X]$ при $h \to 0$) знакопеременное решение. Схему не стоит применять.

Если, например, погрешность вычислений $\sim 10^{-18}$, то растущее решение весьма заметно для $X \sim 10$, а для $X \sim 20$ сходимость не наблюдается.

Третья схема:

$$\begin{cases} \dfrac{y_{k+1}-y_{k-1}}{2h} + \dfrac{y_{k+1}+y_{k-1}}{2} = 0;\ \mu = \pm\sqrt{\dfrac{1-h}{1+h}},\ \max\limits_{i=1,2}|\mu_i| < 1, \\ y_0 = 1, y_1 = 1-h. \end{cases}$$

Данная схема по своим свойствам похожа на первую схему, но имеет знакопеременное решение.

Еще раз отметим, что рассмотренное α-условие характеризует *устойчивость схемы* и при наличии аппроксимации обычно влечет сходимость решения разностной схемы к решению дифференциальной задачи. Однако, сходимость может существенно зависеть от длины отрезка интегрирования и отсутствовать для больших X.

Лекция 6. Методы решения задачи Коши

Кроме конечно-разностного метода для решения задачи Коши применяются следующие алгоритмы.

Метод Тейлора. Пусть на отрезке $[x_0, x_0+X]$ требуется найти решение дифференциального уравнения $y' = f(x,y)$ при начальном условии $y(x_0)$. Дифференцируя уравнение по x, имеем следующую цепочку соотношений:

$$y'' = f_x + f_y y',\ y''' = f_{xx} + 2f_{xy}y' + f_{yy}(y')^2 + f_y y'', \ldots .$$

Последовательно подставляя начальные данные в полученные соотношения, получаем значения производных искомой функции: $y'(x_0)$, $y''(x_0)$, ... Это позволяет написать приближенное равенство: $y(x) \approx \sum\limits_{i=0}^{n} \frac{y^{(i)}(x_0)}{i!}(x-x_0)^i$. Если величина $|x-x_0|$ больше радиуса сходимости ряда, то формально метод неприменим, но можно разбить $[x_0, x_0+X]$ на подотрезки и построить решение в точке x за несколько шагов. Данный алгоритм может быть полезен, когда требуется решить большое количество задач вполне определенного вида с различными начальными данными. В этом случае требуемые производные можно найти аналитически и сохранить для многократного применения.

Методы типа Адамса. Рассмотрим следующее интегральное тождество, верное на решении задачи $y'(x) = f(x,y)$:

$$y(x+h) = y(x) + \int\limits_x^{x+h} y'(\tau)d\tau = y(x) + \int\limits_x^{x+h} f(\tau, y(\tau))d\tau.$$

Заменим точное значение интеграла на приближенное значение, применив следующую лемму.

Лемма 6.1 *Пусть $g \in C^2[a,b]$. Тогда для интеграла $I(g) = \int\limits_a^b g(x)dx$ верны следующие оценки:*

1) $|I(g) - g(a)(b-a)| \le \|g'\|\frac{(b-a)^2}{2}$,
2) $|I(g) - g(b)(b-a)| \le \|g'\|\frac{(b-a)^2}{2}$,
3) $|I(g) - g(\frac{a+b}{2})(b-a)| \le \|g''\|\frac{(b-a)^3}{24}$,
4) $|I(g) - \frac{1}{2}(g(a)+g(b))(b-a)| \le \|g''\|\frac{(b-a)^3}{12}$,

где $\|g\| = \max_{x\in[a,b]}|g(x)|$.

Доказательство. Первые три оценки могут быть получены из разложений Тейлора:

$$g(x) = g(a) + g'(\xi_1(x))(x-a),$$
$$g(x) = g(b) + g'(\xi_2(x))(x-b),$$
$$g(x) = g(\tfrac{a+b}{2}) + g'(\tfrac{a+b}{2})(x-\tfrac{a+b}{2}) + \tfrac{1}{2}g''(\xi_3(x))(x-\tfrac{a+b}{2})^2,$$

четвертое — из тождества:

$$\int\limits_a^b g(x)\,dx - \frac{b-a}{2}\,(g(a)+g(b)) = \frac{1}{2}\int\limits_a^b (a-x)(b-x)g''(x)\,dx,$$

которое можно проверить, проинтегрировав правую часть равенства по частям два раза. Лемма доказана.

Явный метод Эйлера. Из оценки 1) леммы следует, что

$$y(x+h) = y(x) + hf(x, y(x)) + O(h^2).$$

Это приводит к расчетным формулам

$$y_{k+1} = y_k + hf(x_k, y_k), \quad y_0 = y(x_0),$$

имеющим локальную погрешность $O(h^2)$ и сходимость с порядком $O(h)$, т.е. схема имеет *первый порядок точности.*

Неявный метод Эйлера. Из оценки 2) леммы следует, что

$$y(x+h) = y(x) + hf(x+h, y(x+h)) + O(h^2).$$

Это приводит к расчетным формулам

$$y_{k+1} = y_k + hf(x_{k+1}, y_{k+1}), \quad y_0 = y(x_0),$$

также имеющим локальную погрешность $O(h^2)$, сходимость $O(h)$, но устойчивым для более широкого класса уравнений. На каждом шаге неявного метода Эйлера требуется решать нелинейное уравнение относительно y_{k+1}. При фиксированном k решение можно найти следующим внутренним итерационным процессом по $j = 0, 1, \ldots$:

$$y_{k+1}^{j+1} = y_k + hf(x_{k+1}, y_{k+1}^j), \quad y_{k+1}^0 = y_k.$$

Можно показать, что при достаточно малых h и гладкой функции f метод сходится, т.к. отображение $y_{k+1}^{j+1} = Y(y_{k+1}^j)$ является сжимающим.

Формулы Адамса второго порядка точности. Для формулы трапеций оценка 4) из леммы дает

$$y(x+h) = y(x) + \frac{h}{2}\big(f(x, y(x)) + f(x+h, y(x+h))\big) + O(h^3).$$

Соответствующая расчетная формула имеет вид

$$y_{k+1} = y_k + \frac{h}{2}\big(f(x_k, y_k) + f(x_{k+1}, y_{k+1})\big), \quad y_0 = y(x_0).$$

Решение y_{k+1} для каждого k часто удается найти внутренним итерационным методом:

$$y_{k+1}^{j+1} = y_k + \frac{h}{2}\big(f(x_k, y_k) + f(x_{k+1}, y_{k+1}^j)\big),\ j = 0, 1, \ldots.$$

При малых значениях h соответствующее отображение обычно является сжимающим, и итерационный процесс сходится к искомому y_{k+1}.

Упростим расчетные формулы неявного метода Адамса без понижения порядка точности. Добавим в правую часть слагаемые $\pm\frac{h}{2}f(x+h, y^*(x+h))$, где $y(x+h) - y^*(x+h) = O(h^2)$. Так как

$$\begin{gathered}\tfrac{h}{2}\left(f(x+h, y(x+h)) - f(x+h, y^*(x+h))\right) = \\ \tfrac{h}{2} f_y(x+h, \tilde{y})(y(x+h) - y^*(x+h)) = O(h^3),\end{gathered}$$

где $\tilde{y}$ лежит между $y(x+h)$ и $y^*(x+h)$, то

$$y(x+h) = y(x) + \frac{h}{2}\big(f(x, y(x)) + f(x+h, y^*(x+h))\big) + O(h^3).$$

И если выбрать $y^*(x+h) = y(x) + hf(x, y(x))$, то требуемое на y^* условие точности будет выполнено. Соответствующие расчетные формулы имеют вид:

$$\begin{gathered}y_{k+1}^* = y_k + hf(x_k, y_k), \quad y_0 = y(x_0), \\ y_{k+1} = y_k + \tfrac{h}{2}\big(f(x_k, y_k) + f(x_{k+1}, y_{k+1}^*)\big).\end{gathered}$$

Замечание. Данный подход по сути соответствует первому шагу в приближенном решении имеющегося нелинейного уравнения. Отметим, что дальнейшие итерации формально не повышают порядок точности, но обычно уменьшают главный член погрешности.

Замечание. Применим идею замены $y(x+h)$ на некоторое приближение y^* в случае неявного метода Эйлера. Получим:

$$\begin{gathered}y_{k+1}^* = y_k + hf(x_k, y_k), \\ y_{k+1} = y_k + hf(x_{k+1}, y_{k+1}^*).\end{gathered}$$

Формально локальная погрешность остается $O(h^2)$, но устойчивость схемы повышается.

Построим другую пару расчетных формул. Для этого заменим интеграл по формуле прямоугольников по центральной точке:

$$y(x+h) = y(x) + hf(x + \frac{h}{2}, y(x + \frac{h}{2})) + O(h^3).$$

Если $y(x+\frac{h}{2}) = y^*(x+\frac{h}{2}) + O(h^2)$, то как и в предшествующем случае имеем

$$y(x+h) = y(x) + hf(x + \frac{h}{2}, y^*(x + \frac{h}{2})) + O(h^3).$$

Соответствующие расчетные формулы принимают вид:

$$\begin{gathered}y_{k+\frac{1}{2}}^* = y_k + \tfrac{h}{2} f(x_k, y_k), \\ y_{k+1} = y_k + hf(x_{k+\frac{1}{2}}, y_{k+\frac{1}{2}}^*).\end{gathered}$$

Методы Рунге–Кутта. Один из наиболее популярных подходов к решению задачи Коши для уравнений первого порядка $y' = f(x, y),\ y(x_0) = y_0$ заключается в следующем. Зафиксируем некоторые числа

$$\alpha_2, \ldots, \alpha_q, \quad p_1, \ldots, p_q, \quad \beta_{i,j},\ 0 < j < i \le q$$

и последовательно вычислим

$$\begin{aligned}&k_1(h) = hf(x,\, y), \\ &k_2(h) = hf(x + \alpha_2 h,\, y + \beta_{2,1} k_1(h)), \\ &\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ &k_q(h) = hf(x + \alpha_q h,\, y + \beta_{q,1} k_1(h) + \ldots + \beta_{q,q-1} k_{q-1}(h))\,.\end{aligned}$$

В этом случае расчетная формула будет иметь вид:

$$y(x+h) \approx z(x+h) = y(x) + \sum_{i=1}^{q} p_i k_i(h).$$

Обозначим погрешность метода на шаге через $\varphi(h) = y(x+h) - z(x+h)$. Если $f(x,\, y)$ — достаточно гладкая функция своих аргументов, то справедлива формула Тейлора:

$$\varphi(h) = \sum_{i=0}^{s} \frac{\varphi^{(i)}(0)}{i!} h^i + \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1},$$

где $0 \le \theta \le 1$. Выберем параметры метода $\alpha_i,\ p_i,\ \beta_{i,j}$ так, что $\varphi'(0) = \ldots = \varphi^{(s)}(0) = 0$. Тогда локальная погрешность метода равна $O(h^{s+1})$, а величина s называется порядком метода.

Построим метод при $q = 1$ и выпишем формулу погрешности. Имеем:
$\varphi(h) = y(x+h) - y(x) - p_1 h f(x,\, y), \quad \varphi(0) = 0,$
$\varphi'(0) = (y'(x+h) - p_1 f(x,\, y))|_{h=0} = f(x,\, y)(1 - p_1),$
$\varphi''(h) = y''(x+h).$
Равенство $\varphi'(0) = 0$ выполняется для всех гладких функций $f(x,\, y)$ лишь в случае $p_1 = 1$. Для погрешности этого метода на шаге получаем выражение $\varphi(h) = y''(x+\theta h)\frac{h^2}{2}$.

Построим все методы при $q = 2$. Для погрешности имеем:

$$\varphi(h) = y(x+h) - y(x) - p_1 h f(x,\, y) - p_2 h f(\bar{x},\, \bar{y}),$$

где $\bar{x} = x + \alpha_2 h,\ \bar{y} = y + \beta_{21} h f(x,\, y)$. Вычислим производные функции $\varphi(h)$:

$$\varphi'(h) = y'(x+h) - p_1 f(x,\, y) - p_2 f(\bar{x},\, \bar{y}) - p_2 h\big(\alpha_2 f_x(\bar{x},\, \bar{y}) + \beta_{21} f_y(\bar{x},\, \bar{y}) f(x,\, y)\big),$$

$$\varphi''(h) = y''(x+h) - 2p_2\big(\alpha_2 f_x(\bar{x},\, \bar{y}) + \beta_{21} f_y(\bar{x},\, \bar{y}) f(x,\, y)\big) - p_2 h\big(\alpha_2^2 f_{xx}(\bar{x},\, \bar{y}) + 2\alpha_2\beta_{21} f_{xy}(\bar{x},\, \bar{y}) f(x,\, y) + \beta_{21}^2 f_{yy}(\bar{x},\, \bar{y})(f(x,\, y))^2\big),$$

$$\varphi'''(h) = y'''(x+h) - 3p_2\big(\alpha_2^2 f_{xx}(\bar{x},\, \bar{y}) + 2\alpha_2\beta_{21} f_{xy}(\bar{x},\, \bar{y}) f(x,\, y)) + \beta_{21}^2 f_{yy}(\bar{x},\, \bar{y})(f(x,\, y))^2\big) + O(h).$$

Согласно исходному дифференциальному уравнению имеем:

$$y' = f, \quad y'' = f_x + f_y f, \quad y''' = f_{xx} + 2f_{xy} f + f_{yy} f^2 + f_y y''.$$

Подставим в выражения $\varphi(h),\ \varphi'(h),\ \varphi''(h),\ \varphi'''(h)$ значение $h = 0$ и, воспользовавшись этими соотношениями, получим:

$$\begin{aligned}
&\varphi(0) = y(x) - y(x) = 0,\\
&\varphi'(0) = (1 - p_1 - p_2) f(x,\, y),\\
&\varphi''(0) = (1 - 2p_2\alpha_2) f_x(x,\, y) + (1 - 2p_2\beta_{21}) f_y(x,\, y) f(x,\, y),\\
&\varphi'''(0) = (1 - 3p_2\alpha_2^2) f_{xx}(x,\, y) + (2 - 6p_2\beta_{21}) f_{xy}(x,\, y) f(x,\, y) +\\
&\qquad + (1 - 3p_2\beta_{21}^2) f_{yy}(x,\, y)(f(x,\, y))^2 + f_y(x,\, y) y''(x). \qquad (1)
\end{aligned}$$

Соотношение $\varphi'(0) = 0$ выполняется при всех $f(x,\, y)$, если

$$1 - p_1 - p_2 = 0; \qquad (2)$$

соотношение $\varphi''(0) = 0$ выполняется, если

$$1 - 2p_2\alpha_2 = 0 \quad \text{и} \quad 1 - 2p_2\beta_{21} = 0. \qquad (3)$$

Таким образом, $\varphi(0) = \varphi'(0) = \varphi''(0) = 0$ при всех $f(x,\, y)$, если выполнены три указанных выше соотношения (2), (3) относительно четырех параметров. Задавая произвольно один из параметров, получим различные методы Рунге–Кутта второго порядка. Например, при $p_1 = 1/2$ получаем $p_2 = 1/2,\ \alpha_2 = 1,\ \beta_{21} = 1$. При $p_1 = 0$ получаем $p_2 = 1,\ \alpha_2 = 1/2,\ \beta_{21} = 1/2$. В случае уравнения $y' = y$, согласно (1), имеем $\varphi'''(0) = y$ независимо от значений $p_1,\ p_2,\ \alpha_2,\ \beta_{21}$. Отсюда следует, что нельзя построить формул Рунге–Кутта со значениями $q = 2$ и $s = 3$.

Лекция 7. Оценка глобальной погрешности

Главный член погрешности. Для задачи

$$y'(x) = y, \quad y(0) = 1$$

рассмотрим схему

$$\frac{y_{k+1} - y_k}{h} = y_k, \quad y_0 = 1, \quad k \ge 0\,.$$

Найдем в разложении ошибки $y(x_n) - y_n = c_0 + c_1 h + c_2 h^2 + \cdots$ постоянные c_0, c_1 для точки $x_n = 1$.

Точное решение разностной задачи имеет вид:

$$y_n = (1+h) y_{n-1} = (1+h)^n y_0 = (1+h)^n\,,$$

а точное решение дифференциальной задачи при $x = x_n$ равно $y(x_n) = \exp(x_n)$. Так как $x_n = n\,h = 1$, следовательно,

$$y(x_n) - y_n = e - (1+h)^{1/h} = e - \exp\left[\tfrac{1}{h}\ln(1+h)\right] =$$

$$= e\,\left(1 - \exp\left[-\tfrac{h}{2} + O(h^2)\right]\right) = \tfrac{e}{2}\,h + O(h^2)\,.$$

Таким образом, имеем $c_0 = 0, c_1 = \dfrac{e}{2}$.

Оценка погрешности явного одношагового метода. Для уравнения $y' = f(x,y)$ с начальным условием $y(x_0)$ рассмотрим явную одношаговую разностную схему общего вида

$y_0,\ y_{k+1} = F(x_k, y_k, x_{k+1} - x_k),\ k = 0, 1, \ldots, n-1,\ x_n = x_0 + X.$

Нас интересует оценка погрешности в n-ой точке интегрирования, т.е. величина $|y(x_n) - y_n|$. Докажем вспомогательное утверждение.

Лемма 7.1 *Пусть $y_1(x)$ и $y_2(x)$ — два решения дифференциального уравнения $y' = f(x,y)$, $x \in [a,b]$, где f — непрерывная и непрерывно дифференцируемая по y функция. Тогда*

$$y_2(b) - y_1(b) = (y_2(a) - y_1(a))e^{\int\limits_a^b f_y(x,\tilde y(x))dx},$$

где функция $\tilde y(x)$ заключена между $y_1(x)$ и $y_2(x)$.

Доказательство. Рассмотрим два решения с различными начальными данными

$$y_2'(x) = f(x, y_2(x)),\ y_1'(x) = f(x, y_1(x)).$$

Для их разности имеем тождество

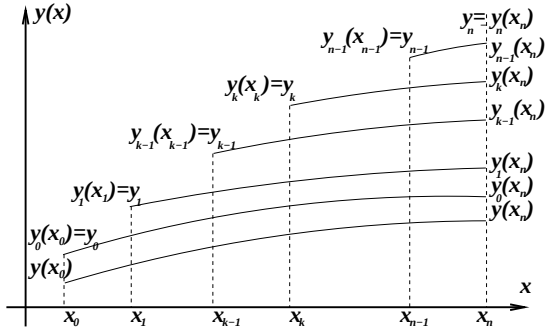
$$(y_2(x) - y_1(x))' = f_y(x,\tilde y)(y_2(x) - y_1(x))$$

с некоторой функцией $\tilde y(x)$. Так как $\tilde y(x)$ фиксирована, и $f_y(x,\tilde y)$ зависит только от x, то можно определить функцию $g(x) = f_y(x,\tilde y)$. Тогда

$$e^{-\int\limits_a^x g(\tau)d\tau}(y_2(x) - y_1(x))' = e^{-\int\limits_a^x g(\tau)d\tau} g(x)(y_2(x) - y_1(x)).$$

Следовательно, $\dfrac{d}{dx}\left(e^{-\int\limits_a^x g(\tau)d\tau}(y_2(x) - y_1(x))\right) = 0$. Интегрируя по $[a,b]$, имеем $(y_2(b) - y_1(b))e^{-\int\limits_a^b g(\tau)d\tau} = y_2(a) - y_1(a)$. Отсюда, умножая на экспоненту, получаем требуемое равенство. Лемма доказана.

Перейдем к получению оценки глобальной погрешности для одношагового явного метода, см. рис.:



Здесь $y(x)$ — точное решение исходного уравнения, y_k — найденное приближенное решение, а $y_k(x)$ — решение уравнения $y_k'(x) = f(x, y_k(x))$ с начальными условиями $y_k(x_k) = y_k$.

Пусть $y_1 = F(x_0, y_0, h_1)$. При этом $y_0 - y(x_0) = r_0$ — погрешность начальных данных, $y_1 - y_0(x_1) = c_1 h_1^{s+1} + \delta_1$ — погрешность метода плюс вычислительная погрешность. И так далее, $y_k - y_{k-1}(x_k) = c_k h_k^{s+1} + \delta_k$ для всех $k = 2, \ldots, n$.

Рассмотрим следующее тождество:

$$y_n - y(x_n) = y_n(x_n) - y(x_n) = \sum_{k=1}^{n}(y_k(x_n) - y_{k-1}(x_n)) + y_0(x_n) - y(x_n).$$

Из определения функции $y_k(x)$, (см. рис.) и леммы имеем:

$$y_k(x_n) - y_{k-1}(x_n) = (y_k(x_k) - y_{k-1}(x_k))e^{\int\limits_{x_k}^{x_n} f_y(\tau,\tilde y_k(\tau))d\tau}.$$

Так как $y_k(x_k) = y_k$, а оценка локальной погрешности имеет вид $y_k - y_{k-1}(x_k) = c_k h_k^{s+1} + \delta_k$, то

$$y_n - y(x_n) = \sum_{k=1}^{n}(c_k h_k^{s+1} + \delta_k)e^{\int\limits_{x_k}^{x_n} f_y(\tau,\tilde y_k(\tau))d\tau} + r_0 e^{\int\limits_{x_0}^{x_n} f_y(\tau,\tilde y_0(\tau))d\tau}.$$

Пусть $|c_k| \le c, |\delta_k| \le \delta, h_k \le h$. Рассмотрим два случая.

1. $|f_y(\tau, \tilde y_k(\tau))| \le L$. Тогда

$$|y_n - y(x_n)| \le e^{L(x_n - x_0)}(ch^s hn + \delta n) + r_0 e^{L(x_n - x_0)} = e^{LX}(cXh^s + \delta n + r_0).$$

2. $f_y(\tau, \tilde{y}_k(\tau)) \le -L$. Тогда, т.к. $\sum_{k=1}^{n} e^{-L\int_{x_k}^{x_n} d\tau} \le \frac{1}{1-e^{-Lh}}$, имеем

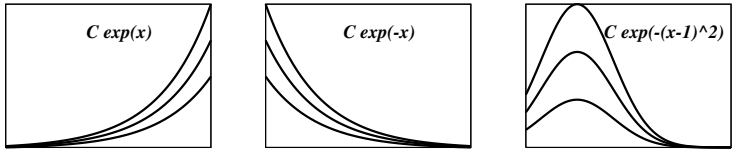
$$|y_n - y(x_n)| \le (ch^{s+1} + \delta)\frac{1}{1-e^{-Lh}} + r_0 e^{-LX} \le \frac{ch^{s+1}+\delta}{Lh} + r_0 e^{-LX}.$$

Формально данная оценка не ухудшается при увеличении отрезка интегрирования X, но с ростом X константа c может существенно расти.

Устойчивые и неустойчивые задачи. Рассмотрим качественное поведение решений следующих уравнений:

$$y_1' = y_1; \quad y_2' = -y_2; \quad y_3' = -2(x-1)y_3;$$
$$y_1(x) = ce^x; \quad y_2(x) = ce^{-x}; \quad y_3(x) = ce^{-(x-1)^2}.$$

Интегральные кривые первого семейства расходятся с увеличением x, второго — сближаются, а третьего — сначала расходятся, а затем сближаются. Так как на k-ом шаге найденное приближенное значение y_k смещается с интегральной кривой $y_{k-1}(x)$, то внесенная в результате этого погрешность может в зависимости от поведения семейства решений либо возрастать, либо уменьшаться, см. рис.:



В связи с этим имеется существенная разница в численном интегрировании двух на первый взгляд эквивалентных задач

$$y_1' = y_1,\ y_1(0) = 1; \quad y_2' = e^x,\ y_2(0) = 1,$$

так как соответствующие им семейства интегральных кривых существенно отличаются: $y_1(x) = ce^x$, $y_2(x) = c + e^x$.

Задача 7.1. Выяснить, какое из рассмотренных уравнение явный метод Эйлера проинтегрирует точнее.

Жесткие системы. Рассмотрим задачу интегрирования уравнения $y' = -ay$, $a = \text{const} > 0$, $x \in [0, X]$. Для решения данной задачи запишем явную разностную схему

$$\frac{y_{k+1} - y_k}{h} = -ay_k \Rightarrow y_{k+1} = (1-ha)y_k.$$

Для качественного совпадения поведения решений разностной и дифференциальной задач необходимо, чтобы $|1 - ha| \le 1$. Соответствующее множество шагов h называется областью устойчивости разностной схемы, а максимально допустимый шаг $h_{cou} = 2/a$ — числом Куранта. Если $h > h_{cou}$, то

мы не только теряем качественное совпадение решений разностной и дифференциальной задач, но и столкнемся с катастрофическим ростом нормы приближенного решения и вычислительной погрешности.

Так как $y(x+h) = y(x) + hy'(x) + \frac{h^2}{2}y''(\xi)$, где $x \le \xi \le x+h$, то погрешность аппроксимации имеет вид $|\frac{h}{2}y''(\xi)|$. Величина $y''(\xi) = a^2e^{-a\xi} \ll 1$ при $a\xi \gg 1$. Поэтому при $a\xi \gg 1$ для достижения требуемой точности локальной аппроксимации $|\frac{h}{2}y''(\xi)| \le \varepsilon$ не требуются мелкие шаги, но $h \le h_{cou}$ для всех x из условия качественного совпадения решений (условия устойчивости).

Данный пример показывает, что разумно выделить класс т.н. жестких задач. Будем считать задачу $y' = f$ жесткой, если характерное время изменения решения много меньше отрезка интегрирования (в данном случае это означает, что $aX \gg 1$). Система уравнений $\mathbf{y}' = A\mathbf{y}$, $\mathbf{y} = (y_1, \ldots, y_n)^T$ считается жесткой, если

$$1)\, Re(\lambda_i(A)) < 0, \quad 2) s = \frac{\max_i |Re(\lambda_i(A))|}{\min_i |Re(\lambda_i(A))|} \gg 1.$$

Число s принято называть числом жесткости. Данное определение естественно обобщается на случай матриц $A(x)$, а также (на основе метода замороженных коэффициентов) на случай правых частей общего вида $F(y, x)$.

В случае одного уравнения задача будет жесткой, если решение содержит несколько компонент с существенно отличающимися характерными временами изменения. Например, пусть $y' = -a(y - \sin x) + \cos x, y(0) = 1$, $a \gg 1$. Тогда точное решение имеет вид $y(x) = e^{-ax} + \sin x$. Здесь можно выделить пограничный слой быстрого изменения решения, далее решение мало отличается от плавной функции $\sin x$. Однако при всех x для успешного интегрирования необходимо выполнение условия $h \le h_{cou}$.

Еще раз отметим, что устойчивость дифференциальной задачи определяется типом уравнения и значениями параметров, устойчивость разностной задачи — типом разностной схемы, значениями параметров и величиной шага. Так явные схемы с постоянным шагом $y_{k+1} = y_k + hf(x_k, y_k)$ обычно требуют мелких шагов, но просты в реализации. Неявные схемы $y_{k+1} = y_k + hf(x_k, y_{k+1})$ обычно имеют менее жесткие условия на h, но в общем случае каждый шаг требует решения нелинейного уравнения. Для рассмотренного модельного уравнения $y' = -ay$ условие устойчивости неявной схемы принимает вид $1/(1+ha) < 1$, т.е. выполнено при всех h, поэтому величина h выбирается только с учетом точности аппроксимации.

Метод Лебедева решения жестких систем. Для решения системы обыкновенных дифференциальных уравнений $\mathbf{y}'(x) = -A\mathbf{y}$, $\mathbf{y} \in \mathbf{R}^n$ рассмотрим явную схему с переменными шагами. Тогда за N шагов будем

иметь

$$\mathbf{y}_N = P_N(A)\mathbf{y}_0, \quad P_N(A) = \prod_{i=1}^{N}(I - h_i A) := S, \quad \mathbf{y}_{kN} = S^k \mathbf{y}_0.$$

Лемма 7.2 *Пусть собственные функции оператора S образуют базис. Тогда*

$$\overline{\lim_{k\to\infty}} \|S^k\| \le \text{const} \Leftrightarrow |\lambda(S)| \le 1.$$

Доказательство. Напомним, что $\|S^k\| = \sup\limits_{\|x\|=1} \|S^k x\|$. Отсюда, разложив по базису из собственных векторов $\{e_i\}$ вектор $x = \sum\limits_{i=1}^{n} \alpha_i e_i$, получим требуемую оценку. Лемма доказана.

Пусть $sp(A) \in [0, M]$, и собственные функции оператора A образуют базис в $\mathbf{R}^n$. Найдем набор таких шагов $\{h_i\}_{i=1}^N$, что схема, оставаясь устойчивой, позволяет за N шагов проинтегрировать задачу по отрезку наибольшей длины. Так как $\lambda(S) = P_N(\lambda(A)) = \prod_{i=1}^N \big(1 - h_i\lambda(A)\big)$, то, с учетом леммы, искомые значения $\{h_i\}_{i=1}^N$ равны обратным величинам корней такого полинома $P_N(x) = \prod_{i=1}^N (1 - h_i x)$, что

$$\max_{x\in[0,M]} |P_N(x)| \le 1, \quad \textstyle\sum_{i=1}^N h_i = -P_N'(0) \to \sup.$$

Решением является многочлен Чебышёва, т.к. $P_N(x) = T_N\left(\frac{2x-M}{M}\right)$ имеет требуемый вид, удовлетворяет условию устойчивости и имеет, согласно теореме Маркова, наибольшую производную в нуле. Следовательно, формула для чебышёвских шагов найдена: $\{h_i = (\frac{M}{2} + \frac{M}{2}\cos\frac{(2i-1)}{2N}\pi)^{-1}\}_{i=1}^N$. Вычислим соответствующую длину отрезка интегрирования, т.е. значение $P_N'(0)$:

$$T_N(x) = \cos N\theta,\ \cos\theta = x;$$

$$T_N'(x) = \frac{N\sin(N\arccos x)}{\sin\arccos x} = NU_{N-1}(x);$$

$$U_{N-1}(x) = \frac{(x+\sqrt{x^2-1})^N - (x-\sqrt{x^2-1})^N}{2\sqrt{x^2-1}} = Nx^{N-1} + O(\sqrt{x^2-1});$$

отсюда находим $|\frac{d}{dx}T_N\left(\frac{2x-M}{M}\right)|_{x=0} = |\frac{2}{M}T_N'(x)|_{x=-1} = \frac{2N^2}{M}$.

Таким образом, суммарная длина отрезка интегрирования с N чебышёвскими шагами равна $\frac{2N^2}{M}$. Максимально допустимый постоянный шаг явной схемы равен $h_{cou} = \frac{2}{M}$, поэтому суммарная длина отрезка интегрирования за N постоянных шагов равна $\frac{2N}{M}$, что в N раз меньше, чем с чебышёвскими шагами.

Важной проблемой в случае чебышёвских шагов является устойчивость метода по отношению к ошибкам округления. Дело в том, что условие устойчивости $\|\prod_{i=1}^N (I - h_i A)\| \le 1$ гарантированно выполняется за N шагов, хотя для больших i величина шага h_i и, следовательно, норма $\|I - h_i A\|$,

будут существенно больше единицы. Действительно, пусть e — нормированный собственный вектор матрицы A, отвечающий наибольшему собственному числу $\lambda(A) = M$. Найдем норму вектора $y = (I - h_i A)e$ для $i = N$. Имеем

$$h_N = \left(\frac{M}{2} + \frac{M}{2}\cos\pi(1 - \frac{1}{2N})\right)^{-1} \sim \left(\frac{M}{2}\frac{1}{2}\frac{\pi^2}{4N^2}\right)^{-1} \sim \frac{N^2}{M}.$$

Поэтому $y = (I - h_N A)e \sim (1 - \frac{N^2}{M}M)e$, т.е. $\|y\| \sim N^2$. Следовательно, для больших значений N применение метода может привести как катастрофическому росту погрешности вычислений, так и к переполнению машинных разрядов. Для устранения данной проблемы необходимо переупорядочить последовательность шагов $\{h_1, h_2, \ldots, h_N\}$. На данный момент известны и строго обоснованы алгоритмы для $N = 2^p 3^q$. Приведем без доказательства один из таких результатов для $N = 2^p$, гарантирующий, что норма оператора перехода $P_n(A)$ по модулю не превосходит единицы для каждого $n = 1, 2, \ldots, N$.

Пусть $h_{N+1-i} = (\frac{M}{2} + \frac{M}{2}\cos\frac{(2i-1)}{2N}\pi)^{-1}, \quad i = 1, \ldots, N$, т.е. нумерация шагов ведется "от большего к меньшему".

Для $N = 2$ устойчивая последовательность имеет вид $(2, 1)$. Пусть для $N = 2^{p-1}$ последовательность построена: $(i_1, i_2, \ldots, i_{2^{p-1}})$, т.е. шаги берутся в следующем порядке: $\{h_{i_1}, h_{i_2}, \ldots, h_{i_{2^{p-1}}}\}$.

Тогда для $N = 2^p$ последовательность определяется формулой:
$(2^p + 1 - i_1, i_1, 2^p + 1 - i_2, i_2, \ldots, 2^p + 1 - i_{2^{p-1}}, i_{2^{p-1}})$.
Например: $(3, 2, 4, 1)$, $(6, 3, 7, 2, 5, 4, 8, 1)$. Таким образом, делается серия шагов по правилу "малый — большой", самый малый шаг выполняется предпоследним, а самым большой — последним.

Лекция 8. Уравнения второго порядка

Рассмотрим следующую задачу:

$$y'' = f(x, y, y'), \quad y(a) = \xi_a, \ \ y'(b) = \xi_b.$$

Введением новой неизвестной функции $v(x) = y'(x)$ она может быть сведена к системе уравнений первого порядка:

$$\begin{aligned} v' &= f(x, y, v), \quad v(b) = \xi_b, \\ y' &= v, \quad y(a) = \xi_a. \end{aligned}$$

И если $b = a$, то для ее решения можно применить методы, рассмотренные выше. В некоторых случаях данный прием оказывается полезным. Однако

в общем случае методы, приспособленные к решению конкретного класса задач, оказываются более эффективны. Далее будем рассматривать задачи с правыми частями, не зависящими от y': $f(x,y,y') = f(x,y)$, т.е. уравнения вида

$$y'' = f(x,y) \tag{1}$$

с краевыми условиями, заданными на границе отрезка $[0,1]$. В данном случае (по аналогии с задачей Коши для уравнения первого порядка) разностной схемой на равномерной сетке $x_k = x_0 + kh$, $k \geq 0$ называется система разностных уравнений

$$\frac{1}{h^2}\sum_{i=0}^{n} a_{-i} y_{k-i} = \sum_{i=0}^{n} b_{-i} f_{k-i}\,, \quad k = n, n+1, \ldots \tag{2}$$

с известными начальными условиями $y_0, y_1, \ldots, y_{n-1}$, где a_{-i}, b_{-i} не зависят от h, $a_0, a_n \neq 0$, и $f_{k-i} = f(x_{k-i}, y_{k-i})$.

Теорема 8.1 *Необходимые и достаточные условия аппроксимации на решении уравнения (1) разностной схемой (2) имеют вид*

$$\frac{1}{h^2}\sum_{i=0}^{n} a_{-i} = 0, \quad -\frac{1}{h}\sum_{i=0}^{n} i a_{-i} = 0, \quad \frac{1}{2}\sum_{i=0}^{n} i^2 a_{-i} = 1, \quad \sum_{i=0}^{n} b_{-i} = 1\,.$$

Доказательство. Подставим решение $y(x)$ в формулу разностной схемы, для каждого слагаемого выпишем разложение в ряд Тейлора в точке x_k, приравняем нулю коэффициенты при h^{-2}, h^{-1}, h^0 и, добавив условие нормировки правой части, получим требуемые соотношения. Теорема доказана.

Рассмотрим разностные схемы для уранения $y''(x) = f(x)$.

Пример 8.1. Естественная аппроксимация:

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = f_k\,.$$

Главный член погрешности на решении равен $r_h = L_h(y)_h - f_h = \frac{h^2}{12} y^{(4)}(x_k) + O(h^4)$.

Пример 8.2. Схема

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = f_k + \frac{h^2}{12} f''(x_k)$$

аппроксимирует уравнение на решении порядком $O(h^4)$.

Пример 8.3. Схема Нумерова:

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = \frac{f_{k+1} + 10 f_k + f_{k-1}}{12}\,.$$

Главный член погрешности на решении имеет вид: $r_h = L_h(y)_h - f_h = \frac{h^2}{12} y^{(4)}(x_k) + \frac{h^4}{360} y^{(6)}(x_k) - \frac{h^2}{12} f^{(2)}(x_k) - \frac{h^4}{144} f^{(4)}(x_k) = -\frac{h^4}{240} y^{(6)}(x_k) + O(h^6)$.

Пример 8.4. Рассмотрим следующее уравнение

$$-\big(k(x)\,y'\big)' + p(x)\,y = f(x)\,,$$

гладкие коэффициенты которого удовлетворяют условиям $0 < k_0 \leq k(x) \leq k_1$, $\quad 0 \leq p(x) \leq p_1$. Тогда схема

$$-\frac{1}{h}\left[k(x_{i+1/2})\frac{y_{i+1} - y_i}{h} - k(x_{i-1/2})\frac{y_i - y_{i-1}}{h}\right] + p(x_i) y_i = f_i$$

имеет порядок аппроксимации $O(h^2)$.

Краевое условие для задачи (1) может быть задано на любом из концов отрезка в виде линейной комбинации функции и производной, т.н. краевое условие третьего рода:

$$a\,y + b\,y' = c\,.$$

В этом случае следует обратить внимание на способ его аппроксимации.

Аппроксимация граничных условий третьего рода. Пусть для уравнения

$$-y'' + p(x)\,y = f(x)$$

задано граничное условие $a\,y(0) + b\,y'(0) = c$. Построим для этого условия конечно-разностную аппроксимацию второго порядка точности на решении, используя значения функции y в точках $x_0 = 0$ и $x_1 = h$. Из формулы Тейлора имеем

$$y(h) = y(0) + h\,y'(0) + \frac{h^2}{2}\,y''(0) + O\big(h^3\big)\,,$$

откуда следует

$$y'(0) = \frac{y(h) - y(0)}{h} - \frac{h}{2}\,y''(0) + O\big(h^2\big)\,.$$

Из исходного уравнения получаем, что $-y''(0) = f(0) - p(0)\,y(0)$. Поэтому $a\,y(0) + b\left(\frac{y(h) - y(0)}{h} + \frac{h}{2}\big(f(0) - p(0)\,y(0)\big)\right) = c + O\big(h^2\big)$. Искомая аппроксимация имеет вид:

$$\left(a - \frac{h}{2}\,b\,p(0)\right) y_0 + b\,\frac{y_1 - y_0}{h} = c - \frac{h}{2}\,b\,f(0)\,.$$

Устойчивость разностных схем второго порядка. В случае задачи Коши принято проверять более слабое определение α-устойчивости.

Определение 8.1 *Для задачи Коши $y''(x) = f(x)$ схема называется α - устойчивой, если все корни соответствующего характеристического многочлена $\sum\limits_{i=0}^{n} a_{-i}\mu^{k-i}$ однородного уравнения принадлежат единичному кругу*

и на границе круга нет кратных корней, за исключением $\mu = 1$ кратности 2.

Отличие данного условия устойчивости от условия устойчивости для уравнения первого порядка обусловлено более высокой степенью h в правой части: $\sum_{i=0}^{n} a_{-i} y_{k-i} = h^2 \sum_{i=0}^{n} b_{-i} f_{k-i}$.

Устойчивость краевой задачи. Пусть уравнение $y''(x) = f(x)$ доопределено краевыми условиями на разных концах отрезка. Напомним определение устойчивости для линейных задач. Разностная схема $A_h y_h = f_h$ линейной задачи устойчива, если существуют C, h_0 такие, что для произвольных $A_h y_h^{(1,2)} = f_h^{(1,2)}$ выполняется оценка $\|y_h^{(1)} - y_h^{(2)}\|_h \le C \|f_h^{(1)} - f_h^{(2)}\|_h$ при всех $h \le h_0$ с константой C, не зависящей от h.

Если матрица A_h не вырождена, то $y_h = A_h^{-1} f_h$. Отсюда получаем неравенство для нормы векторов:

$$\|y_h^{(1)} - y_h^{(2)}\|_h \le \|A_h^{-1}\|_h \, \|f_h^{(1)} - f_h^{(2)}\|_h.$$

Следовательно, можно выбрать $C \ge \|A_h^{-1}\|_h$.

Метод собственных функций. Исследуем устойчивость разностной схемы

$$-\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = f_k, \quad y_0 = y_N = 0 \quad \Leftrightarrow \quad A_h y_h = f_h$$

в сеточной интегральной норме $\|y_h\|_h^2 = (y_h, y_h)_h = \sum_{k=1}^{N-1} y_k^2 h$, эквивалентной норме $\|y(x)\|_{L_2(0,1)}^2 = \int\limits_0^1 y^2(x) dx$ исходной задачи. Для этого оценим норму оператора $\|A_h^{-1}\|_h$, где

$$\|A_h^{-1}\|_h^2 = \sup_{y_h \ne 0} \frac{(A_h^{-1} y_h, A_h^{-1} y_h)_h}{(y_h, y_h)_h}.$$

Пусть известны собственные числа λ_n матрицы A_h, а собственные вектора $y_k^{(n)}$ ортонормальны: $(y^{(n)}, y^{(m)})_h = \delta_m^n$. Тогда собственные вектора образуют базис, и произвольный вектор y_k можно представить в виде $y_k = \sum_{n=1}^{N-1} c_n y_k^{(n)}$. С учетом $A_h^{-1} y_k = \sum_{n=1}^{N-1} \lambda_n^{-1} c_n y_k^{(n)}$ имеем $\|A^{-1}\|_h = \max_n |\lambda_n^{-1}|$.

Ранее было получено, что решение разностной задачи на собственные значения

$$\frac{y_{k+1} - 2\,y_k + y_{k-1}}{h^2} = -\lambda\, y_k, \quad y_0 = y_N = 0, \quad h = 1/N$$

имеет вид

$$y_k^{(n)} = \sin \pi n k h, \quad \lambda_n = \frac{4}{h^2} \sin^2 \frac{\pi n h}{2}, \quad n = 1, \ldots, N-1\,.$$

Проверим, что $(y^{(n)}, y^{(m)})_h = 0$ при $m \ne n$. Рассмотрим оператор A_h как оператор, действующий на $y_1, \ldots, y_{N-1}$. В этом случае матрица A_h симметрична $A_h = A_h^T$, следовательно,

$$0 = (A_h \varphi^{(n)}, \varphi^{(m)})_h - (\varphi^{(n)}, A_h \varphi^{(m)})_h = (\lambda^{(n)} - \lambda^{(m)})(\varphi^{(m)}, \varphi^{(n)})_h,$$

т.е. $(\varphi^{(m)}, \varphi^{(n)})_h = 0$ для $\lambda^{(m)} \ne \lambda^{(n)}$. Ортогональность собственных векторов доказана, поэтому $\|A^{-1}\|_h = \lambda_{\min}^{-1}$.

Получим оценку для $\lambda_{\min} = \lambda_1$. Из неравенства $\sin|\beta| \ge \frac{2}{\pi}|\beta|$ при $|\beta| \le \frac{\pi}{2}$ имеем

$$\lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2} \ge 4, \quad \lambda_{\max} = \lambda_{N-1} \le \frac{4}{h^2}\,.$$

Таким образом, верна не зависящая от h оценка для нормы $\|A^{-1}\|_h \le \frac{1}{4}$. По определению это означает устойчивость схемы.

Энергетический метод исследования устойчивости. Рассмотрим энергетический метод исследования устойчивости на примере дифференциальной задачи

$$-y'' + p(x)\, y = f(x), \quad y(0) = y(1) = 0, \quad p(x) \ge 0\,.$$

Умножим уравнение на $y(x)$, и результат проинтегрируем по отрезку $[0, 1]$:

$$\int\limits_0^1 (-y'') y \, dx + \int\limits_0^1 p y^2 dx = \int\limits_0^1 f y \, dx\,.$$

После интегрирования по частям получим интегральное тождество

$$\int\limits_0^1 (y')^2 dx + \int\limits_0^1 p y^2 dx = \int\limits_0^1 f y \, dx\,.$$

Далее нам потребуется неравенство, связывающее интегралы от квадратов функции и ее производной. Из равенства

$$y(x_0) = \int\limits_0^{x_0} y'(x)\, dx$$

следует, что

$$\big|y(x_0)\big|^2 \le \left(\int\limits_0^{x_0} 1^2 dx\right) \left(\int\limits_0^{x_0} (y')^2 dx\right) \le \int\limits_0^{x_0} (y')^2 dx \le \int\limits_0^1 (y')^2 dx\,.$$

После интегрирования по x_0 по отрезку $[0,1]$ обеих частей получим искомое неравенство

$$\int_0^1 \big|y(x_0)\big|^2 dx_0 \le \int_0^1 (y')^2 dx \int_0^1 dx_0 \quad \text{или} \quad \int_0^1 y^2 dx \le \int_0^1 (y')^2 dx\,.$$

В результате имеем

$$\int_0^1 y^2 dx \le \int_0^1 (y')^2 dx + \int_0^1 p y^2 dx = \int_0^1 f y dx \le \frac{1}{2}\left(\int_0^1 f^2 dx + \int_0^1 y^2 dx\right),$$

откуда $$\|y\|_{L_2} \le \|f\|_{L_2}, \quad \text{где} \quad \|y\|^2_{L_2} = \int_0^1 y^2 dx\,.$$

Это — априорная оценка для решения, означающая устойчивость дифференциальной задачи по правой части.

Докажем энергетическим методом устойчивость стандартной разностной схемы

$$-\frac{y_{k+1}-2y_k+y_{k-1}}{h^2} + p_k\, y_k = f_k,\ k = 1,\dots,N-1, \quad y_0 = y_N = 0.$$

Умножим на y_k и просуммируем от 1 до $(N-1)$. Так как $y_0 = y_N = 0$, то:

$$-\frac{1}{h^2}\sum_{k=1}^{N-1}(y_{k+1}-y_k)y_k + \frac{1}{h^2}\sum_{k=1}^{N-1}(y_k - y_{k-1})y_k =$$
$$= -\frac{1}{h^2}\sum_{k=2}^{N}(y_k-y_{k-1})y_{k-1} + \frac{1}{h^2}\sum_{k=1}^{N-1}(y_k-y_{k-1})y_k = \frac{1}{h^2}\sum_{k=1}^{N}(y_k-y_{k-1})^2.$$

Отсюда следует конечномерный аналог интегрального тождества:

$$\frac{1}{h^2}\sum_{k=1}^{N}(y_k-y_{k-1})^2 + \sum_{k=1}^{N-1} p_k y_k^2 = \sum_{k=1}^{N-1} f_k y_k.$$

Введем следующие обозначения:

$$\nabla y_k = y_k - y_{k-1},\ (u,v] = \sum_{i=1}^{N} u_i v_i,\ (u,v) = \sum_{i=1}^{N-1} u_i v_i.$$

Тогда конечно-разностное интегральное тождество принимает вид:

$$\frac{1}{h^2}(\nabla y, \nabla y] + (p\,y, y) = (f, y).$$

Докажем сеточный аналог неравенства для функции и ее производной в точках $k = 1,\dots,N-1$.

Из тождества $y_k = \sum_{i=1}^{k}(y_i - y_{i-1})$ следует

$$y_k^2 \le \left(\sum_{i=1}^{k} 1^2\right)\cdot\left(\sum_{i=1}^{k}(y_i-y_{i-1})^2\right) \le N\sum_{i=1}^{N}(y_i-y_{i-1})^2.$$

Отсюда получается требуемая оценка

$$\sum_{k=1}^{N-1} y_k^2 \le N^2 \sum_{i=1}^{N}(y_i-y_{i-1})^2 \le \frac{1}{h^2}\sum_{i=1}^{N}(y_i-y_{i-1})^2 + \sum_{i=1}^{N-1} p_i y_i^2 =$$
$$= \sum_{i=1}^{N-1} f_i y_i \le \frac{1}{2}\left(\sum_{i=1}^{N-1} f_i^2 + \sum_{i=1}^{N-1} y_i^2\right).$$

Таким образом,

$$\sum_{i=1}^{N-1} y_i^2 \le \sum_{i=1}^{N-1} f_i^2, \text{ т.е. } \sum_{i=1}^{N-1} y_i^2 h \le \sum_{i=1}^{N-1} f_i^2 h.$$

Отсюда следует, что априорная оценка для решения разностной задачи в норме $\|y\|_{L_{2,h}} = (\sum_{i=1}^{N-1} y_i^2 h)^{1/2}$, согласованной с $L_2(0,1)$, имеет вид $\|y\|_{L_{2,h}} \le \|f\|_{L_{2,h}}$. Устойчивость доказана. И т.к. схема имеет второй порядок аппроксимации, то согласно теореме Филиппова верна

Теорема 8.2 *Решение рассмотренной разностной схемы сходится к решению дифференциальной задачи с порядком $O(h^2)$ в норме $L_{2,h}$, т.е.*

$$\|(y)_h - y_h\|_{L_{2,h}} \le ch^2.$$

Лекция 9. Методы решения уравнений второго порядка

Рассмотрим эффективные численные алгоритмы решения систем линейных алгебраических уравнений, возникающих, например, при аппроксимации обыкновенных дифференциальных уравнений второго порядка.

Метод прогонки. Пусть требуется найти решение системы уравнений $A\,\mathbf{y} = \mathbf{f}$, где $\mathbf{y} = (y_0, y_1, \dots, y_N)^T$ — вектор неизвестных, $\mathbf{f} = (f_0, f_1, \dots, f_N)^T$

— заданный вектор правых частей, A — квадратная $(N+1)\times(N+1)$ матрица:

$$\begin{pmatrix} c_0 & -b_0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ -a_1 & c_1 & -b_1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & -a_2 & c_2 & -b_2 & \dots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -a_{N-2} & c_{N-2} & -b_{N-2} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -a_{N-1} & c_{N-1} & -b_{N-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -a_N & c_N \end{pmatrix}.$$

Перепишем задачу следующим образом:

$$\begin{array}{cc} c_0y_0 - b_0y_1 = f_0\,, & k=0\,, \\ -a_ky_{k-1} + c_ky_k - b_ky_{k+1} = f_k\,, & 1\le k\le N-1\,, \\ -a_Ny_{N-1} + c_Ny_N = f_N\,, & k=N\,. \end{array} \qquad (1)$$

Основная идея метода состоит в представлении решения в виде

$$y_k = \alpha_{k+1}y_{k+1} + \beta_{k+1}\,, \quad k = N-1, N-2, \dots, 0\,, \qquad (2)$$

для которого значения α_k, β_k и y_k вычисляются по коэффициентам исходной системы и правой части.

Из первого уравнения (1) следует:

$$y_0 = \alpha_1y_1 + \beta_1\,, \quad \alpha_1 = b_0/c_0\,, \ \ \beta_1 = f_0/c_0\,.$$

Найдем последующие коэффициенты в равенстве $y_k = \alpha_{k+1}y_{k+1}+\beta_{k+1}$. Для этого подставим уже найденную формулу $y_{k-1} = \alpha_ky_k + \beta_k$ в уравнение из (1):

$$\begin{array}{c} y_{k-1} = \alpha_ky_k + \beta_k\,, \\ -a_ky_{k-1} + c_ky_k - b_ky_{k+1} = f_k\,. \end{array}$$

Отсюда имеем

$$\begin{array}{c} (-a_k\alpha_k + c_k)y_k - \alpha_k\beta_k - b_ky_{k+1} = f_k\,, \\ \alpha_{k+1} = \dfrac{b_k}{c_k - a_k\alpha_k}\,,\ \beta_{k+1} = \dfrac{f_k + a_k\beta_k}{c_k - a_k\alpha_k}\,. \end{array}$$

Этот процесс закончится, когда мы придем к последнему уравнению системы (1), содержащему только два значения неизвестных:

$$\begin{array}{c} y_{N-1} = \alpha_Ny_N + \beta_N\,, \\ -a_Ny_{N-1} + c_Ny_N = f_N\,. \end{array}$$

Исключение из этой системы y_{N-1} приводит к формуле для y_N:

$$y_N = \frac{f_N + a_N\beta_N}{c_N - a_N\alpha_N}\,.$$

Формально отметим, что $y_N = \beta_{N+1}$. Вывод формул закончен. Сформулируем алгоритм в целом.

Для решения системы (1) сначала рекуррентно вычисляются прогоночные коэффициенты α_k, β_k:

$$\begin{array}{c} \alpha_1 = b_0/c_0\,, \quad \alpha_{k+1} = \dfrac{b_k}{c_k - a_k\alpha_k}\,, \\ \beta_1 = f_0/c_0\,, \quad \beta_{k+1} = \dfrac{f_k + a_k\beta_k}{c_k - a_k\alpha_k}\,, \end{array}$$

где k последовательно принимает значения $1, 2, \dots, N-1$.

Затем вычисляется y_N:

$$y_N = \frac{f_N + a_N\beta_N}{c_N - a_N\alpha_N}\,.$$

И, наконец, рекуррентно определяются остальные компоненты вектора неизвестных:

$$y_k = \alpha_{k+1}y_{k+1} + \beta_{k+1}\,, \quad k = N-1, N-2, \dots, 0\,.$$

Полученные соотношения называют формулами *правой прогонки.*

Теорема 9.1 *(Достаточные условия корректности и устойчивости метода прогонки). Пусть коэффициенты системы* (1) *действительны и удовлетворяют условиям:* c_0, c_N, a_k, c_k, b_k *при* $k = 1, 2, \dots, N-1$ *отличны от нуля и*

$$\begin{array}{l} |c_k| \ge |a_k| + |b_k|\,, \quad k = 1, 2, \dots, N-1\,, \\ |c_0| \ge |b_0|\,, \quad |c_N| \ge |a_N|\,, \end{array}$$

причем хотя бы одно из неравенств является строгим. Тогда для формул метода прогонки справедливы неравенства:

$$c_k - a_k\alpha_k \ne 0\,, \quad |\alpha_k| \le 1\,, \quad k = 1, 2, \dots, N\,,$$

гарантирующие корректность и устойчивость метода.

Доказательство проведем методом индукции.
База: $|\alpha_1| \le 1$. Шаг: Пусть $|\alpha_k| \le 1$.
Покажем, что $|c_k - \alpha_ka_k| > 0$ и $|\alpha_{k+1}| \le 1$. Действительно,

$$|c_k - \alpha_ka_k| \ge |c_k| - |\alpha_k||a_k| \ge |c_k| - |a_k| \ge |b_k| > 0.$$

Отсюда следует, что

$$|\alpha_{k+1}| = \frac{|b_k|}{|c_k - \alpha_k a_k|} \le 1.$$

Отметим, что если при некотором k_0 выполняется $|\alpha_{k_0}| < 1$, тогда $|\alpha_j| < 1$ при всех $j > k_0$.

Рассмотрим $|c_N - \alpha_N a_N| \ge |c_N| - |\alpha_N||a_N|$. По условию:
либо $|\alpha_N| < 1$, если $|c_0| > |b_0|$ или $|c_{k_0}| > |a_{k_0}| + |b_{k_0}|$,
либо $|c_N| > |a_N|$,
следовательно, $|c_N - \alpha_N a_N| > 0$. Теорема доказана.

Пример 9.1. Для задачи $-y'' = f$, $y(0) = a, y(1) = b$ разностная аппроксимация имеет вид

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & \\ 0 & -1 & 2 & -1 & 0 & \dots & 0 \\ . & . & . & . & . & . & . \\ 0 & & \dots & & -1 & 2 & -1 \\ 0 & & \dots & & 0 & -1 & 2 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_{N-2} \\ y_{N-1} \end{pmatrix} = \mathbf{f},$$

где $\mathbf{f} = (f_1 + \frac{a}{h^2}, f_2, \dots, f_{N-2}, f_{N-1} + \frac{b}{h^2})^T$.

В данном случае мы исключили известные значения y_0, y_N. Метод прогонки устойчив.

Пример 9.2. Для задачи $-y'' = f$, $y'(0) = a, y'(1) = b$ разностная аппроксимация имеет вид

$$\frac{1}{h^2}\begin{pmatrix} 2 & -2 & 0 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & \\ 0 & -1 & 2 & -1 & 0 & \dots & 0 \\ . & . & . & . & . & . & . \\ 0 & & \dots & & -1 & 2 & -1 \\ 0 & & \dots & & 0 & -2 & 2 \end{pmatrix}\begin{pmatrix} y_0 \\ y_1 \\ . \\ . \\ . \\ y_{N-1} \\ y_N \end{pmatrix} = \mathbf{f},$$

где $\mathbf{f} = (f_0 - \frac{2a}{h}, f_1, \dots, f_{N-1}, f_N - \frac{2b}{h})^T$.

Аппроксимация, например, левого краевого условия получена следующим образом: $y(h) = y(0) + hy'(0) + \frac{h^2}{2}y''(0) + O(h^3)$, и из уравнения находим $-y''(0) = f(0)$.

В данном случае из теоремы не следует корректность метода прогонки. Но и задача некорректна, т.к. если $\mathbf{y}$ решение системы, то $\mathbf{y} + \text{const}$ также решение.

Пример 9.3. Для задачи $-y'' = f$, $y'(0) = a, y(1) = b$ при рассмотренных аппроксимациях метод прогонки применим.

Метод стрельбы. Идея этого подхода выглядит следующим образом. Для решения системы уравнений

$$-\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = f_k, \big(\Leftrightarrow L_h y_h = f_h\big), \quad y_0 = a, y_N = b$$

рассмотрим две вспомогательные задачи

$$L_h u_h = f_h,\ u_0 = a,\ u_1 = \bar{u}_1;$$
$$L_h v_h = 0,\ v_0 = 0,\ v_1 = \bar{v}_1,$$

и построим решение $y_h = u_h + Cv_h$, определив коэффициент C из условия $y_N = b$. Формально значения u_1, v_1 могут быть произвольными, однако, для повышения устойчивости схемы разумно выбрать $u_1 - u_0 = O(h)$, $v_1 - v_0 = O(h)$.

Метод Фурье (базисных функций). Пусть требуется найти решение системы линейных уравнений $A\mathbf{y} = \mathbf{b}$, $\mathbf{y} \in \mathbf{R}^M$ при условии, что известны все собственные вектора и собственные числа матрицы A:

$$A\mathbf{y}^{(n)} = \lambda_n \mathbf{y}^{(n)},\ n = 1, ..., M,$$

и $\mathbf{y}^{(n)}$ образуют ортонормированный базис в пространстве $\mathbf{R}^M$.

Основная идея метода состоит в формальном разложении решения $\mathbf{y}$ по собственным векторам $\mathbf{y} = \sum_{n=1}^{M} c_n \mathbf{y}^{(n)}$, определении коэффициентов c_n в данном представлении и последующем восстановлении вектора $\mathbf{y}$.

Так как проблема нахождения собственных векторов и собственных значений существенно сложнее решения системы линейных уравнений, то данный метод позволяет эффективно находить вектор $\mathbf{y}$, если все собственные вектора и собственные числа заданы аналитически, и система собственных векторов образует в пространстве решений ортонормированный базис относительно некоторого скалярного произведения: $(\mathbf{y}^{(n)}, \mathbf{y}^{(m)})_h = \delta_m^n$. В этом случае коэффициенты c_m могут быть найдены по явной формуле. Действительно,

$$A(\sum_{n=1}^{M} c_n \mathbf{y}^{(n)}) = \mathbf{b} \quad \Rightarrow \quad \sum_{n=1}^{M} c_n \lambda_n \mathbf{y}^{(n)} = \mathbf{b}.$$

Умножим данное равенство скалярно на $\mathbf{y}^{(m)}$ для $m = 1, \dots, M$. С учетом ортонормированности базиса получим

$$(\sum_{n=1}^{M} \lambda_n c_n \mathbf{y}^{(n)}, \mathbf{y}^{(m)})_h = (\mathbf{b}, \mathbf{y}^{(m)})_h \quad \Rightarrow \quad c_m \lambda_m = (\mathbf{b}, \mathbf{y}^{(m)})_h.$$

Отсюда имеем $c_m = d_m/\lambda_m$, где величины $d_m = (\mathbf{b}, \mathbf{y}^{(m)})_h$ являются коэффициентами в разложении вектора $\mathbf{b} = \sum\limits_{m=1}^{M} d_m \mathbf{y}^{(m)}$.

Пример 9.4. Выписать алгоритм решения системы

$$-\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = b_k,\ k = 1, ..., N-1,\ y_0 = y_N = 0,\ h = 1/N$$

методом Фурье. Для данной задачи собственные векторы и собственные числа могут быть найдены аналитически

$$y_k^{(n)} = \sqrt{2}\sin(\pi nkh), \quad \lambda_n = \frac{4}{h^2}\sin^2\big(\frac{\pi nh}{2}\big),\ n = 1, \ldots, N-1.$$

При этом $y^{(n)}$ ортонормированны относительно скалярного произведения $(y^{(n)}, y^{(m)})_h = \sum\limits_{k=1}^{N-1} y_k^{(n)} y_k^{(m)} h$.

Ортогональность следует из симметричности оператора A, действующего на $y_1, \ldots, y_{N-1}$. Проверим условие ортонормальности:

$$(\mathbf{y}^{(n)}, \mathbf{y}^{(n)})_h = h\sum_{k=1}^{N-1}(y_k^{(n)})^2 = 2h\sum_{k=1}^{N-1}\sin^2(\pi nkh) =$$

$$= h\sum_{k=1}^{N-1}(1 - \cos(2\pi nkh)) = (N-1)h - h\sum_{k=1}^{N-1}\cos(2\pi nkh),$$

$$\sum_{k=1}^{N-1}\cos(2\pi nkh) = \mathrm{Re}\sum_{k=1}^{N-1} e^{(\mathbf{i}2\pi nhk)} =$$

$$= \mathrm{Re}(q + \ldots + q^{N-1})(1-q)/(1-q) = \mathrm{Re}\frac{q - q^N}{1-q}.$$

Так как $q^N = \cos 2\pi Nhn + \mathbf{i}\sin 2\pi Nhn = 1$, $hN = 1$, то $(\mathbf{y}^{(n)}, \mathbf{y}^{(n)})_h = Nh = 1$.

Пример 9.5. Для задачи

$$-\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = b_k,\ k = 1, ..., N-1,$$
$$-\frac{2}{h^2}(y_1 - y_0) = b_0,\ \frac{2}{h^2}(y_N - y_{N-1}) = b_N,\ h = 1/N$$

собственные векторы, собственные числа и соответствующее скалярное произведение имеют вид

$y_k^{(0)} = 1, y_k^{(n)} = \sqrt{2}\cos(\pi nkh),\ n = 1, \ldots, N-1, y_k^{(N)} = \cos(\pi Nkh)$;

$\lambda^{(0)} = 0,\ \lambda^{(n)} = \frac{4}{h^2}\sin(\frac{\pi nh}{2}), n = 1, \ldots, N$;

$(y^{(m)}, y^{(n)})_h = \sum\limits_{k=1}^{N-1} y_k^{(m)} y_k^{(n)} h + \frac{h}{2} y_0^{(m)} y_0^{(n)} + \frac{h}{2} y_N^{(m)} y_N^{(n)}$.

Для доказательства ортогональности собственных векторов покажем, что оператор A, действующий на векторах $(y_0, y_1, \ldots, y_{N-1}, y_N)^T$, симметричен относительно выбранного скалярного произведения. Действительно,

$$(Au, v)_h = h\sum_{i=1}^{N-1}\sum_{j=0}^{N} a_{ij}u_j v_i + \frac{h}{2}\sum_{j=0}^{N} a_{0j}u_j v_0 + \frac{h}{2}\sum_{j=0}^{N} a_{Nj}u_j v_N = h(\tilde{A}u, v),$$

где $\frac{1}{2}a_{0j} = \tilde{a}_{0j}$, $\frac{1}{2}a_{Nj} = \tilde{a}_{Nj}$ для $\forall j$; $a_{ij} = \tilde{a}_{ij}$ иначе.

Но матрица $\tilde{A} = \tilde{A}^T$ симметрична, следовательно,

$(Au, v)_h = h(\tilde{A}u, v) = h(u, \tilde{A}v) = h(\tilde{A}v, u) = (Av, u)_h = (u, Av)_h$.

Ортонормальность собственных векторов проверяется аналогично предыдущему примеру. Так как $\lambda^{(0)} = 0$, $y^{(0)} \equiv 1$, то необходимым и достаточным условием корректности исходной системы уравнений является условие $(\mathbf{b}, 1)_h = 0$.

Отметим, что для рассмотренных примеров нахождение коэффициентов d_m и восстановление решения y_k можно существенно ускорить за счет арифметических свойств $y^{(n)}$ при помощи так называемого быстрого преобразования Фурье.

3. Линейная алгебра

Лекция 10. Векторные и матричные нормы

Определение 10.1 *Нормой вектора* $\mathbf{x} = (x_1, \ldots, x_n)^T$ *называется функционал, обозначаемый* $\|\mathbf{x}\|$ *и удовлетворяющий следующим условиям:*

$$\|\mathbf{x}\| > 0\,, \quad \text{если} \quad \mathbf{x} \neq 0\,, \quad \|0\| = 0,$$
$$\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|,$$
$$\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$$

Наиболее употребительны следующие нормы:

$$\|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} |x_i|,\ \|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|,\ \|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{(\mathbf{x},\mathbf{x})}.$$

Определение 10.2 *Нормы* $\|\cdot\|_{\mathrm{I}}$ *и* $\|\cdot\|_{\mathrm{II}}$ *называются эквивалентными, если для всех* $\mathbf{x} \in \mathbf{R}^n$ *с одними и теми же положительными постоянными* c_1 *и* c_2 *справедливы неравенства*

$$c_1\,\|\mathbf{x}\|_{\mathrm{II}} \leq \|\mathbf{x}\|_{\mathrm{I}} \leq c_2\,\|\mathbf{x}\|_{\mathrm{II}}.$$

Пример 10.1. $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$.

Теорема 10.1 *Пусть* C *— симметричная положительно определенная матрица. Тогда функционал* $\sqrt{(C\mathbf{x},\mathbf{x})} = \|x\|_C$ *задает норму вектора* $\mathbf{x}$*, и верна следующая оценка*

$$\min_i \sqrt{\lambda_i}\|x\|_2 \leq \|x\|_C \leq \max_i \sqrt{\lambda_i}\|x\|_2,$$

где $\{\lambda_i\}$ *— собственные числа матрицы* C*.*

Доказательство. Первая часть теоремы следует из представления $C = Q^T D^{1/2} Q Q^T D^{1/2} Q$, здесь $Q^T Q = I$, $D^{1/2} = \mathrm{diag}\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\}$. Найдем константы эквивалентности с нормой $\|\cdot\|_2$. Пусть $\mathbf{e}_1, \ldots, \mathbf{e}_n$ — ортонормированная система собственных векторов матрицы C (т.е. $(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij}$), а $\lambda_1, \ldots, \lambda_n$ — соответствующие собственные значения. Любой вектор $\mathbf{x}$ представим в виде $\mathbf{x} = \sum\limits_{i=1}^n \alpha_i \mathbf{e}_i$. Поэтому,

$$(C\mathbf{x},\mathbf{x}) = \left(\sum_{i=1}^n \lambda_i\alpha_i\mathbf{e}_i, \sum_{i=1}^n \alpha_i\mathbf{e}_i\right) = \sum_{i=1}^n \lambda_i\alpha_i^2.$$

Отсюда для произвольного вектора $\mathbf{x}$ получаем

$$\min_i \lambda_i \cdot (\mathbf{x},\mathbf{x}) \leq (C\mathbf{x},\mathbf{x}) \leq \max_i \lambda_i \cdot (\mathbf{x},\mathbf{x}), \quad (\mathbf{x},\mathbf{x}) = \sum_i \alpha_i^2.$$

Поскольку все $\lambda_i > 0$, полученное неравенство означает эквивалентность евклидовой норме $\|\mathbf{x}\|_2$ с постоянными

$$c_1 = \sqrt{\min_i \lambda_i}, \quad c_2 = \sqrt{\max_i \lambda_i}.$$

Определение 10.3 *Нормой матрицы* A *называется функционал, обозначаемый* $\|A\|$ *и удовлетворяющий следующим условиям:*

$$\|A\| > 0\,, \quad \text{если} \quad A \neq 0\,, \quad \|0\| = 0,$$
$$\|\alpha\,A\| = |\alpha|\,\|A\|,$$
$$\|A+B\| \leq \|A\| + \|B\|,$$
$$\|A\,C\| \leq \|A\|\,\|C\|.$$

Пример 10.2. Функционал $N(A) = \left(\sum\limits_{i,j=1}^n a_{ij}^2\right)^{1/2}$ является матричной нормой. Норму $N(A)$ называют *нормой Фробениуса (нормой Шура, евклидовой матричной нормой)* и обозначают $\|A\|_F$.

Пример 10.3. Функционал $\eta(A) = \max_{i,j} |a_{ij}|$ не является нормой в пространстве матриц. Свойства 1-3 определения матричной нормы очевидно выполнены. Рассмотрим матрицы $A = B : a_{ij} = b_{ij} = 1$, для которых имеют место соотношения $\eta(AB) = n$, $\eta(A) = \eta(B) = 1$, противоречащие четвертому свойству матричной нормы: $\eta(AB) \leq \eta(A)\eta(B)$. Отметим, что функционал $M(A) = n\,\eta(A)$ является матричной нормой.

Лемма 10.1 *Пусть задана некоторая векторная норма* $\|\cdot\|_v$*. Тогда матричную норму можно определить как операторную*

$$\|A\|_v = \sup_{\|\mathbf{x}\|_v \neq 0} \frac{\|A\,\mathbf{x}\|_v}{\|\mathbf{x}\|_v} = \sup_{\|\mathbf{x}\|_v = 1} \|A\,\mathbf{x}\|_v.$$

Доказательство сводится к элементарной проверке свойств 1-4 матричной нормы.

Построенная матричная норма называется *подчиненной* соответствующей векторной норме $\|\cdot\|_v$. Отметим, что для произвольной подчиненной матричной нормы $\|\cdot\|_v$ и единичной матрицы I имеем $\|I\|_v = 1$.

Теорема 10.2 *Матричные нормы, подчиненные векторным нормам* $\|\cdot\|_\infty$, $\|\cdot\|_1$ *и* $\|\cdot\|_2$, *имеют вид*

$$\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|,\ \|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|,\ \|A\|_2 = \sqrt{\max_i \lambda_i(A^T A)}\,.$$

Доказательство. Получим оценку сверху для величины $\|A\mathbf{x}\|_\infty$:

$$\|A\mathbf{x}\|_\infty = \max_i \left|\sum_j a_{ij}x_j\right| \le \max_i \left(\sum_j |a_{ij}|\ \max_j |x_j|\right) \le$$
$$\le \max_i \left(\sum_j |a_{ij}|\right) \|\mathbf{x}\|_\infty.$$

Покажем, что эта оценка достигается. Пусть максимум по i имеет место при $i = l$; тогда возьмем $\mathbf{x} = \big(\mathrm{sign}\,(a_{l1}), \mathrm{sign}\,(a_{l2}),\ \ldots, \mathrm{sign}\,(a_{ln})\big)$. Имеем $\|\mathbf{x}\|_\infty = 1$ и точные равенства во всей цепочке выше, т.е. $\|A\|_\infty = \max\limits_i \sum\limits_{j=1}^n |a_{ij}|$. Аналогично показывается, что $\|A\|_1 = \max\limits_j \sum\limits_{i=1}^n |a_{ij}|$.

Найдем $\|A\|_2$. По определению матричной нормы, подчиненной евклидовой векторной норме, имеем

$$\|A\|_2 = \sup_{\mathbf{x}\ne 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x}\ne 0} \sqrt{\frac{(A\mathbf{x}, A\mathbf{x})}{(\mathbf{x},\mathbf{x})}} = \sup_{\mathbf{x}\ne 0} \sqrt{\frac{(A^T A\mathbf{x}, \mathbf{x})}{(\mathbf{x},\mathbf{x})}}.$$

Матрица $C = A^T A$ — симметричная, и $(A^T A\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) \ge 0$, следовательно, все $\lambda(C) \ge 0$. Рассуждая далее как и в предыдущей теореме, получим $\sup\limits_{x\ne 0} \dfrac{(C\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})} \le \max\limits_i \lambda_i(C)$, а равенство достигается на соответствующем собственном векторе. Поэтому $\|A\|_2 = \sqrt{\max\limits_i \lambda_i(A^T A)}\,.$

Следует отметить важный частный случай симметричной матрицы: если $A = A^T$, то $\|A\|_2 = \max\limits_i |\lambda_i(A)|$.

Теорема 10.3 *Модуль любого собственного значения матрицы не больше любой ее нормы:* $|\lambda(A)| \le \|A\|$.

Доказательство. Зафиксируем произвольный собственный вектор $\mathbf{x}$ матрицы A и построим матрицу X, столбцами которой являются вектора $\mathbf{x}$. Получим равенство $\lambda X = A X$. Отсюда следует $|\lambda|\,\|X\| \le \|A\|\,\|X\|$, что порождает искомый ответ.

Следствие 10.1 *Для любого собственного значения* $\lambda(A)$ *невырожденной матрицы* A *справедлива оценка* $1/\|A^{-1}\| \le |\lambda(A)|$.

Элементы теории возмущений. Рассмотрим систему линейных алгебраических уравнений $A\,\mathbf{x} = \mathbf{b}$ с квадратной невырожденной матрицей A и точным решением $\mathbf{x}$. В результате численного решения с конечной разрядностью вместо $\mathbf{x}$ получается *приближенное* решение $\tilde{\mathbf{x}}$: $A\,\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$. При этом вектор $\mathbf{z} = \mathbf{x} - \tilde{\mathbf{x}}$ называется *вектором ошибки*, а вектор $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$ называется *вектором невязки*.

Найдем насколько приближенное решение отличается от точного. Из неравенств

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \le \|A^{-1}\|\|\mathbf{b} - \tilde{\mathbf{b}}\|, \quad \|A\|\|\mathbf{x}\| \ge \|\mathbf{b}\|$$

получаем, что для относительной ошибки верна оценка

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \|A\|\,\|A^{-1}\|\,\frac{\|\mathbf{b} - \tilde{\mathbf{b}}\|}{\|\mathbf{b}\|} = \|A\|\,\|A^{-1}\|\,\frac{\|\mathbf{b} - A\,\tilde{\mathbf{x}}\|}{\|\mathbf{b}\|}\,.$$

Величина $\|A\|\,\|A^{-1}\|$ называется *числом обусловленности* матрицы A и часто обозначается $\mathrm{cond}(A)$. Для вырожденных матриц $\mathrm{cond}(A) = \infty$. Конкретное значение $\mathrm{cond}(A)$ зависит от выбора матричной нормы, однако, в силу их эквивалентности при практических оценках этим различием обычно можно пренебречь. Если $\mathrm{cond}(A)$ велико, то матрицу называют *плохо обусловленной*.

Пример 10.4. Покажем, что если матрица A плохо обусловлена, то малая невязка *не может* гарантировать малость относительной ошибки. Более того, может оказаться так, что достаточно точное решение будет иметь большую невязку. Пусть

$$\begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix}, \quad \varepsilon \ll 1.$$

Вектор $\tilde{\mathbf{x}} = (1+1, \varepsilon)^T$, который не является близким к $\mathbf{x}$, дает маленькую невязку $\mathbf{r} = (-\varepsilon, 0)^T$. Вектор $\tilde{\mathbf{x}} = (1, \varepsilon + \sqrt{\varepsilon})^T$ достаточно близок к $\mathbf{x}$ в смысле относительной погрешности, однако $\tilde{\mathbf{x}}$ дает большую невязку $\mathbf{r} = (0, -1/\sqrt{\varepsilon})^T$.

Лемма 10.2 *Для любой матрицы* A *имеем* $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$, $\mathrm{cond}(A) \ge 1$ *и* $\mathrm{cond}_2(Q) = 1$ *для ортогональной матрицы* Q.

Доказательство. Первое утверждение следует из свойств обратной матрицы и определения матричной нормы. Далее, так как $I = I \cdot I$, $\|I\| \le \|I\|\|I\|$, $I = A^{-1}A$, то

$$1 \le \|I\| = \|A^{-1}\,A\| \le \|A^{-1}\|\,\|A\| = \mathrm{cond}(A)\,.$$

Далее, имеем:

$$\|Q\|_2 = \sup_{\|x\|_2\neq 0} \frac{(Qx,Qx)^{1/2}}{\|x\|_2} = \sup_{\|x\|_2\neq 0} \frac{(Q^TQx,x)^{1/2}}{\|x\|_2} = 1.$$

Аналогично находим, что $\|Q^{-1}\|_2 = \|Q^T\|_2 = 1$.

Пример 10.5. Покажем, что если определитель матрицы мал, то матрица не обязательно плохо обусловлена, а определитель плохо обусловленной матрицы может равняться 1.

Пусть дана диагональная матрица $A = \varepsilon I$, где $\varepsilon > 0$ — малое число и I — единичная матрица. Определитель $\det(A) = \varepsilon^n$ весьма мал, тогда как матрица A хорошо обусловлена, поскольку

$$\mathrm{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = \varepsilon\|I\|_\infty\, \varepsilon^{-1}\|I^{-1}\|_\infty = 1\,.$$

$$\text{Пусть } A = \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix}, \text{ тогда } \det(A) = 1, \quad \mathrm{cond}_2(A) = \varepsilon^{-2}.$$

Пример 10.6. Пусть

$$A = \begin{pmatrix} 1 & -1 & -1 & \dots & -1 \\ 0 & 1 & -1 & \dots & -1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \; A^{-1} = \begin{pmatrix} 1 & 1 & 2 & \dots & 2^{n-2} \\ 0 & 1 & 1 & \dots & 2^{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Здесь матрица A^{-1} найдена как решение системы $A\mathbf{x} = \mathbf{b}$ при помощи обратного хода метода Гаусса. Определитель A равен 1, вычислим ее число обусловленности. Получим: $\|A\|_\infty = n$, $\|A^{-1}\|_\infty = 1 + 1 + 2 + 2^2 + \dots + 2^{n-2} = 2^{n-1}$, $\mathrm{cond}_\infty(A) = n\,2^{n-1}$, т.е. матрица A плохо обусловлена, хотя $\det(A) = 1$. Отметим также, что в данном случае $\lambda_i(A) = 1$, но матрица A несимметрична, поэтому за обусловленность отвечают собственные числа A^TA.

Лекция 11. Точные методы решения систем линейных алгебраических уравнений

К точным методам решения системы $A\mathbf{x} = \mathbf{b}$ линейных алгебраических уравнений (*слау*) относятся алгоритмы, которые при отсутствии ошибок округления позволяют точно вычислить искомый вектор $\mathbf{x}$. Если число ненулевых элементов матрицы имеет порядок n^2, то большинство такого рода алгоритмов позволяет найти решение за $O(n^3)$ арифметических

действий. Данная оценка, а также необходимость хранения всех элементов матрицы в памяти машины, накладывают существенное ограничение на область применимости точных методов. Однако, для решения задач не очень большой размерности ($n \sim 10^{3\div 4}$) в большинстве случаев разумно применение точных алгоритмов. Отметим, что при численном решении задач математической физики часто требуется обращать матрицы блочно-диагонального вида. В этом случае удается построить точные методы с меньшим по порядку числом арифметических действий. К таким алгоритмам относятся метод прогонки, стрельбы, Фурье (базисных функций).

Метод исключения Гаусса является наиболее известным из точных методов, применяемых для задач с матрицами общего вида. В предположении, что коэффициент $a_{11} \neq 0$, уравнения исходной системы заменяются на следующие

$$\begin{cases} x_1 + \sum\limits_{j=2}^{n} \dfrac{a_{1j}}{a_{11}} x_j = \dfrac{b_1}{a_{11}}, \\ \sum\limits_{j=2}^{n} \left(a_{ij}x_j - \dfrac{a_{1j}}{a_{11}} a_{i1}x_j \right) = b_i - \dfrac{b_1}{a_{11}} a_{i1}, \quad i = 2,\dots,n, \end{cases}$$

т.е. первое уравнение делится на a_{11}, а затем, умноженное на соответствующий коэффициент a_{i1}, вычитается из последующих уравнений. В полученной системе $A^{(1)}x = b^{(1)}$ неизвестное x_1 оказывается исключенным из всех уравнений, кроме первого. Далее, при условии, что коэффициент $a_{22}^{(1)}$ матрицы $A^{(1)}$ отличен от нуля, исключаем x_2 из всех уравнений кроме первого и второго, и т.д. В итоге получим систему $A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$ с верхнетреугольной матрицей. Данная последовательность вычислений называется прямым ходом метода Гаусса. Из последнего уравнения приведенной системы определяем компоненту решения x_n. Далее подставляем x_n в $(n-1)$-е уравнение, находим x_{n-1} и т.д. Соответствующая последовательность вычислений называется обратным ходом Гаусса. Если на k-м шаге прямого хода коэффициент $a_{kk}^{(k-1)}$ равен нулю, тогда k-я строка уравнения переставляется с произвольной l-й строкой, $l > k$ с ненулевым коэффициентом $a_{lk}^{(k-1)}$ при x_k. Такая строка всегда найдется, если $\det(A) \neq 0$.

Если на k-м шаге прямого хода диагональный элемент $a_{kk}^{(k-1)}$ отличен от нуля, но имеет малое абсолютное значение, то коэффициенты очередной матрицы $A^{(k)}$ будут вычислены с большой абсолютной погрешностью. Это может существенно исказить найденный ответ. Поэтому при практической реализации метода Гаусса следует на каждом шаге прямого хода переставлять на k-е место строку с максимальным по модулю элементом $a_{lk}^{(k-1)}$ среди всех $l \geq k$. Такая *необходимая* при расчетах модификация названа методом Гаусса с частичным выбором главного элемента. Данный алгоритм позволяет гарантированно найти приближенное решение $\tilde{\mathbf{x}}$ с малой нормой

невязки $\|\mathbf{b} - A\tilde{\mathbf{x}}\|/\|\mathbf{b}\|$ но, возможно, с большой ошибкой $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$.

Лемма 11.1 *Реализация прямого и обратного хода метода Гаусса требует порядка* $\frac{2}{3}n^3$ *и* $\frac{1}{2}n^2$ *арифметических действий соответственно.*

Действительно, число умножений прямого хода равно $n^2 + \ldots + 1 \approx \int_0^n x^2 dx = n^3/3$, столько же сложений; для обратного хода количесто арифметических действий находится аналогично.

Лемма 11.2 *Прямой ход метода Гаусса соответствует последовательному умножению исходной системы на диагональные матрицы* C_k *и нижнетреугольные матрицы* C'_k*: матрица* C_k *получается из матрицы* I *заменой диагонального элемента с индексом* $i_{k,k}$ *на элемент* $\left(a_{k,k}^{(k-1)}\right)^{-1}$*, матрица* C'_k *получается из матрицы* I *заменой* k*-го столбца на столбец*

$$\left(0, \ldots, 1, -a_{k+1,k}^{(k-1)}, -a_{k+2,k}^{(k-1)}, \ldots, -a_{n,k}^{(k-1)}\right)^T.$$

Метод Гаусса по сути соответствует разложению исходной матрицы A на произведение нижнетреугольной L и верхнетреугольной R. Действительно, $CA = R$, где $C = C_n C'_{n-1} \ldots C'_1 C_1$ — нижнетреугольная матрица, а R — верхнетреугольная матрица с единичной диагональю. Поэтому $A = LR$, где $L = C^{-1}$.

Прямой ход метода Гаусса с частичным выбором главного элемента соответствует последовательному умножению исходной системы на некоторые диагональные матрицы C_k, нижнетреугольные матрицы C'_k и матрицы перестановок P_k. При этом матрицы C_k, C'_k совпадают с матрицами метода Гаусса, матрицы P_k получаются из единичной матрицы I некоторой перестановкой строк.

Такой подход гарантирует, что метод находит приближенное решение с малой относительной невязкой (но, возможно, большой ошибкой). Метод Гаусса с выбором на каждом шаге наибольшего элемента по всей подматрице на практике почти не применяется, хотя имеются немногочисленные примеры, когда он дает качественное улучшение.

Отметим, что для невырожденной матрицы A существуют матрицы перестановок P_1 и P_2, нижняя треугольная матрица L и верхнетреугольная R такие, что $P_1 A P_2 = LR$. При этом достаточно одной из матриц P_i. Умножение A на матрицу P_1 слева переставляет строки исходной матрицы, а умножение на P_2 справа — столбцы. Для того чтобы матрица имела LR-разложение, необходимо и достаточно, чтобы все ее ведущие подматрицы (в том числе и A) были невырожденные.

Среди точных методов, требующих для реализации порядка $O(n^3)$ действий, одним из наиболее устойчивых к вычислительной погрешности является *метод отражений*.

Пусть имеется некоторый единичный вектор $\mathbf{w} \in \mathbf{R}^n$, $\|\mathbf{w}\|_2 = 1$. Построим по нему следующую матрицу $U = I - 2\mathbf{w}\,\mathbf{w}^T$, называемую матрицей Хаусхолдера. Здесь I — единичный оператор, $\mathbf{w}\,\mathbf{w}^T = \Omega$ — матрица с элементами $\omega_{ij} = w_i w_j$, являющаяся результатом произведения вектор-столбца $\mathbf{w}$ на вектор-строку $\mathbf{w}^T$.

Теорема 11.1 *(Свойства матрицы Хаусхолдера).*

1. Матрица U *является симметричной и ортогональной матрицей, т.е.* $U = U^T$ *и* $U^T U = I$*, и все ее собственные значения равны* ± 1*.*

2. $U\mathbf{w} = -\mathbf{w}$*; если вектор* $\mathbf{v}$ *ортогонален* $\mathbf{w}$*, тогда* $U\mathbf{v} = \mathbf{v}$*.*

3. Образ $U\mathbf{y}$ *произвольного вектора* $\mathbf{y}$ *является зеркальным отражением относительно гиперплоскости, ортогональной вектору* $\mathbf{w}$*.*

Доказательство. 1. Симметричность U следует из явного вида U. Так как $(\mathbf{w}, \mathbf{w}) = 1$, следовательно,

$\Omega\Omega|_{ij} = \sum_{k=1}^{n} w_i w_k w_k w_j = \Omega|_{ij}$ и $UU = I - 4\Omega + 4\Omega\Omega = I$, т.е. $U^2 = U^T U = I$.

2. Так как $(\Omega\mathbf{w})_i = \sum_{j=1}^n w_i w_j w_j = w_i$, следовательно $(I - 2\Omega)\mathbf{w} = \mathbf{w} - 2\mathbf{w} = -\mathbf{w}$. Аналогично $\Omega\mathbf{v} = 0$, если $\sum_{j=1}^n w_j v_j = (\mathbf{w}, \mathbf{v}) = 0$.

3. Представим $\mathbf{y}$ в виде $\mathbf{y} = (\mathbf{y}, \mathbf{w})\mathbf{w} + \mathbf{v}$. Тогда из п. 2 следует $U\mathbf{y} = -(\mathbf{y}, \mathbf{w})\mathbf{w} + \mathbf{v}$, где $\mathbf{v} = \mathbf{y} - (\mathbf{y}, \mathbf{w})\mathbf{w}$, $\mathbf{v} \perp \mathbf{w}$. Теорема доказана.

Преобразование Хаусхолдера. Для единичных векторов $\mathbf{y}$ и $\mathbf{e}$ найдем единичный вектор $\mathbf{w}$ такой, что $U\mathbf{y} = \mathbf{e}$, где $U = I - 2\mathbf{w}\,\mathbf{w}^T$. Из свойства зеркально отражения 3 предыдущей теоремы несложно заметить, что $\mathbf{w} = \pm(\mathbf{y} - \mathbf{e})/\sqrt{(\mathbf{y} - \mathbf{e}, \mathbf{y} - \mathbf{e})}$. Действительно, $\left(I - 2\mathbf{w}\,\mathbf{w}^T\right)\mathbf{y} = \mathbf{y} - \xi = \mathbf{e}$, так как

$$\xi_i = \frac{2\sum\limits_{k=1}^{n}(y_i - e_i)(y_k - e_k)y_k}{(\mathbf{y} - \mathbf{e}, \mathbf{y} - \mathbf{e})} = \frac{2(y_i - e_i)(1 - (\mathbf{y}, \mathbf{e}))}{2 - 2(\mathbf{y}, \mathbf{e})} = (y_i - e_i).$$

Отметим, что преобразование U не меняет длины вектора, следовательно, для неединичного вектора $\mathbf{y}$ имеем:

$$U\mathbf{y} = \alpha\mathbf{e},\ \alpha = \|\mathbf{y}\|_2,\ \mathbf{w} = \pm\frac{(\mathbf{y} - \alpha\mathbf{e})}{\|\mathbf{y} - \alpha\mathbf{e}\|_2} = \pm\frac{(\alpha^{-1}\mathbf{y} - \mathbf{e})}{\|\alpha^{-1}\mathbf{y} - \mathbf{e}\|_2}.$$

Метод отражений. Произвольная квадратная матрица A может быть приведена к верхнетреугольному виду в результате последовательного умно-

жения слева на ортогональные матрицы отражений. Действительно, по векторам $\mathbf{y}_1 = (a_{1,1}, \ldots, a_{n,1})^T$ и $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ можно построить вектор $\mathbf{w}_1$ и соответствующую матрицу U_1 так, чтобы первый столбец матрицы $A^{(1)} = U_1 A$ был пропорционален вектору $\mathbf{e}_1 \in \mathbf{R}^n$, т.е. $U_1\mathbf{y_1} = \pm\alpha_1\mathbf{e}_1$. Вычислим $\alpha_1 = (a_{1,1}^2 + a_{2,1}^2 + \ldots + a_{n,1}^2)^{1/2}$ и определим

$$\tilde{\mathbf{w}}_1 = (a_{1,1}/\alpha_1 + \mathrm{sign}(a_{1,1}), a_{2,1}/\alpha_1, \ldots, a_{n,1}/\alpha_1)^T,\ \mathbf{w}_1 = \tilde{\mathbf{w}}_1/\|\tilde{\mathbf{w}}_1\|_2.$$

Такой выбор знака и предварительная нормировка на α_1 гарантируют малость вычислительной погрешности и устойчивость алгоритма.

Далее в пространстве $\mathbf{R}^{n-1}$ по вектору $\mathbf{y}_2 = (a_{22}^{(1)}, \ldots, a_{2n}^{(1)})^T$ строится матрица U_2', отображающая его в вектор, коллинеарный $\mathbf{e}_2 = (1, 0, \ldots, 0)^T \in \mathbf{R}^{n-1}$. Затем определяется $U_2 = \begin{pmatrix} 1 & 0 \\ 0 & U_2' \end{pmatrix}$ и рассматривается матрица $A^{(2)} = U_2U_1A$, и так далее. На k-м шаге имеем $U_k = \begin{pmatrix} I_{k-1} & 0 \\ 0 & U_k' \end{pmatrix}$. Таким образом, матрица отражений U_k строится по вектору $\mathbf{w}_k = \tilde{\mathbf{w}}_k/\|\tilde{\mathbf{w}}_k\|_2$, $\mathbf{w}_k \in \mathbf{R}^n$, где

$$\tilde{\mathbf{w}}_k = (0, \ldots, 0, a_{k,k}^{(k-1)}/\alpha_k + \mathrm{sign}(a_{k,k}^{(k-1)}), a_{k+1,k}^{(k-1)}/\alpha_k, \ldots, a_{n,k}^{(k-1)}/\alpha_k)^T,$$

$$\alpha_k = ((a_{k,k}^{(k-1)})^2 + (a_{k+1,k}^{(k-1)})^2 + \ldots + (a_{n,k}^{(k-1)})^2)^{1/2}.$$

В результате преобразований получится верхняя треугольная матрица $R = UA$, где $U = U_{n-1}\ldots U_1$. При практической реализации явное вычисление U_k не требуется, так как $U_kA^{(k-1)} = A^{(k-1)} - 2\mathbf{w}_k\left(\mathbf{w}_k^TA^{(k-1)}\right)$. При этом изменяются только элементы $a_{ij}^{(k-1)}$, $k \le i, j \le n$ матрицы $A^{(k-1)}$.

Из условия $UU = I$ имеем $A = UR$. Таким образом, произвольная квадратная матрица A может быть представлена в виде произведения симметричной ортогональной матрицы U и верхней треугольной матрицы R.

Рассмотренный алгоритм позволяет свести систему линейных уравнений $A\mathbf{x} = \mathbf{b}$ к виду $R\mathbf{x} = U_{n-1}\ldots U_1\mathbf{b}$, а затем найти ее решение обратным ходом метода Гаусса. Пусть решается задача с возмущенной правой частью $A\tilde{\mathbf{x}} = \mathbf{b} + \delta$ и $\|\delta\| \ll \|\mathbf{b}\|$. Так как ортогональные преобразования не меняют евклидову норму векторов, то для приведенной системы $R\tilde{\mathbf{x}} = U\mathbf{b} + U\delta$ имеем $\|U\delta\| = \|\delta\| \ll \|\mathbf{b}\| = \|U\mathbf{b}\|$, и относительная погрешность правой части не увеличилась. В методе Гаусса матрица преобразования C не ортогональна, и в общем случае может оказаться, что $\|C\delta\|$ станет сравнимым с $\|C\mathbf{b}\|$. То есть малая начальная погрешность может существенно исказить ответ.

QR-разложение. Рассмотрим еще один метод (в дополнение к методу отражений) построения разложения $A = QR$.

Утверждение 11.1 *Пусть A — $m \times n$ матрица, $m \ge n$, вектор-столбцы которой линейно независимы. Тогда существуют и единственны матрица $Q \in \mathbf{R}^{m\times n}$, $Q^TQ = I_n$, и $R \in \mathbf{R}^{n\times n}$, верхнетреугольная с положительными диагональными элементами такие, что $A = QR$.*

Классический алгоритм Грама–Шмидта ортогонализации набора линейно-независимых векторов $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$:

$$\begin{array}{ll} \mathbf{q}_1 = \mathbf{a}_1, & \mathbf{q}_1 := \mathbf{q}_1/\|\mathbf{q}_1\|_2; \\ \mathbf{q}_2 = \mathbf{a}_2 - (\mathbf{a}_2, \mathbf{q}_1)\mathbf{q}_1, & \mathbf{q}_2 := \mathbf{q}_2/\|\mathbf{q}_2\|_2; \\ \ldots & \ldots \\ \mathbf{q}_n = \mathbf{a}_n - \sum_{j=1}^{n-1}(\mathbf{a}_n, \mathbf{q}_j)\mathbf{q}_j, & \mathbf{q}_n := \mathbf{q}_n/\|\mathbf{q}_n\|_2. \end{array}$$

Теорема 11.2 *Вектора $\{\mathbf{q}_i\}_1^n$ ортонормальны: $(\mathbf{q}_i, \mathbf{q}_j) = \delta_{ij}$.*

Доказательство несложно провести методом индукции, используя явный вид формул и условие

$$\mathrm{span} < \mathbf{a}_1, \ldots, \mathbf{a}_k >= \mathrm{span} < \mathbf{q}_1, \ldots, \mathbf{q}_k > \perp \mathbf{q}_m, \quad m > k.$$

Модифицированный алгоритм Грама–Шмидта математически эквивалентен предыдущему, но более устойчив при наличии вычислительной погрешности:

$$\begin{array}{ll} \mathbf{q}_1 = \mathbf{a}_1, & \mathbf{q}_1 := \mathbf{q}_1/\|\mathbf{q}_1\|_2; \\ \mathbf{q}_2 = \mathbf{a}_2, & \\ \mathbf{q}_2 = \mathbf{q}_2 - (\mathbf{q}_2, \mathbf{q}_1)\mathbf{q}_1, & \mathbf{q}_2 := \mathbf{q}_2/\|\mathbf{q}_2\|_2; \\ \ldots & \ldots \\ \mathbf{q}_n = \mathbf{a}_n, & \\ \mathbf{q}_n = \mathbf{q}_n - (\mathbf{q}_n, \mathbf{q}_1)q_1, & \mathbf{q}_n = \mathbf{q}_n - (\mathbf{q}_n, \mathbf{q}_2)\mathbf{q}_2, \\ \ldots & \ldots \\ \mathbf{q}_n = \mathbf{q}_n - (\mathbf{q}_n, \mathbf{q}_{n-1})\mathbf{q}_{n-1}, & \mathbf{q}_n := \mathbf{q}_n/\|\mathbf{q}_n\|_2. \end{array}$$

Вычислительная устойчивость достигается за счет вычитания уже найденных $\mathbf{q}_i$ из $\mathbf{a}_n$. В этом случае накопление погрешности при вычислении $\mathbf{q}_n$ из-за неточной ортогональности $\{\mathbf{q}_i\}$, $i < n$ происходит существенно медленнее.

Построим по найденным векторам матрицу $Q = (\mathbf{q}_1 \ldots \mathbf{q}_n)$ и рассмотрим матрицу $R = Q^TA$:

$$Q^TA = \begin{pmatrix} \mathbf{q}_1^T \\ \ldots \\ \mathbf{q}_n^T \end{pmatrix}(\mathbf{a}_1 \ldots \mathbf{a}_n) = \begin{pmatrix} (\mathbf{q}_1, \mathbf{a}_1) & (\mathbf{q}_1, \mathbf{a}_2) & . & (\mathbf{q}_1, \mathbf{a}_n) \\ 0 & (\mathbf{q}_2, \mathbf{a}_2) & . & (\mathbf{q}_2, \mathbf{a}_n) \\ . & . & . & . \\ 0 & . & 0 & (\mathbf{q}_n, \mathbf{a}_n) \end{pmatrix}$$

с элементами $r_{ij} = (\mathbf{q}_i, \mathbf{a}_j)$. Отметим, что по условию

$$\text{span} < \mathbf{a}_1, \ldots, \mathbf{a}_k >= \text{span} < \mathbf{q}_1, \ldots, \mathbf{q}_k > \perp \mathbf{q}_m,\ m > k.$$

Отсюда следует, что $(\mathbf{q}_i, \mathbf{a}_j) = 0$ при $i > j$. Построенная матрица Q ортогональна $Q^TQ = I_n$, т.к. $(\mathbf{q}_i, \mathbf{q}_j) = \delta_{ij}$. Отсюда и из $Q^TA = R$ имеем $A = QR$.

Задача 11.1. Построить подобное разложение $A = QR$ с квадратной матрицей Q.

Если модифицированный алгоритм Грама–Шмидта необходимо применить для решения задачи $A\mathbf{x} = \mathbf{b}$ (хотя для больших n происходит значительное накопление вычислительной погрешности, и поэтому лучше воспользоваться методом отражений), то задачу сводят к системе $R\mathbf{x} = \mathbf{d}$. При этом матрица R находится указанным способом, а компоненты вектора d_i для сохранения вычислительной устойчивости определяются по следующему алгоритму: $\mathbf{r}_0 = \mathbf{b},\ d_i = (\mathbf{r}_{i-1}, \mathbf{q}_i),\ \mathbf{r}_i = \mathbf{r}_{i-1} - d_i\mathbf{q}_i,\ i = 1, \ldots, n$.

Отметим, что если приближенное решение $\tilde{\mathbf{x}}$ системы $A\mathbf{x} = \mathbf{b}$ получено каким-либо методом, основанном на QR-разложении, то можно выполнить следующий процесс уточнения. Найдем вектор невязки $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$ с удвоенным количеством значащих цифр и решим систему $A\mathbf{z} = \mathbf{r}/\|\mathbf{r}\|$. Положим $\tilde{\mathbf{x}} := \tilde{\mathbf{x}} + \|\mathbf{r}\|\mathbf{z}$. Процесс уточнения значительно экономичнее, чем решение исходного уравнения, так как разложение матрицы A уже имеется. Уточнение можно повторять до тех пор, пока убывает норма вектора невязки.

Лекция 12. Линейная задача наименьших квадратов

Пусть требуется решить слау $A\mathbf{x} = \mathbf{b}$ с матрицей A размерности $m \times n$, правой частью $\mathbf{b} \in \mathbf{R}^m$ и вектором неизвестных $\mathbf{x} \in \mathbf{R}^n$.

Рассмотрим три случая: 1) $m = n$, $\det(A) \neq 0$; 2) $m < n$, $\text{rank}(A) = m$; 3) $m > n$, $\text{rank}(A) = n$.

В случае 1) задача невырождена и вектор $\mathbf{x} = A^{-1}\mathbf{b}$ является точным решением. Для вектора невязки $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ имеем $\|\mathbf{r}\| = 0$. В случае 2) задача недоопределена. Исходная система имеет подпространство решений размерности $(n - m)$. Для каждого решения имеем $\|\mathbf{r}\| = \mathbf{0}$. В случае 3) система переопределена и, если она несовместна, то точного решения не существует, т.е. для произвольного $\mathbf{x} \in \mathbf{R}^n$ имеем $\|\mathbf{b} - A\mathbf{x}\| = \|\mathbf{r}\| > 0$.

Представляют интерес методы решения переопределенных задач, поэтому далее, если не оговаривается иное, считаем, что $m > n$ и $\text{rank}(A) = n$. Для задач такого рода Гаусс предложил считать решением вектор $\mathbf{x}$, минимизирующий евклидову норму вектора невязки $\inf\limits_{\mathbf{y}} \|\mathbf{b} - A\mathbf{y}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2$.

Рассмотрим некоторые методы решения данной минимизационной задачи, называемой *задачей наименьших квадратов* (знк).

Метод нормального уравнения Гаусса для знк. Рассмотрим следующую, называемую *нормальной*, систему уравнений $A^TA\mathbf{x} = A^T\mathbf{b}$ с квадратной матрицей A^TA размерности $n \times n$. Отсюда найдем вектор $\mathbf{x}$.

Теорема 12.1 (Гаусса К.Ф.) *Пусть $m \geq n$ и $\text{rank}(A) = n$. Тогда нормальное уравнение имеет единственное решение.*

Доказательство. Если $A\mathbf{x} \neq 0$, то $(A^TA\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) > 0$. Но $A\mathbf{x} \neq 0$ для всякого $\mathbf{x} \neq 0$, так как $\text{rank}(A) = n$. Следовательно, матрица A^TA невырождена и нормальное уравнение имеет единственное решение. Отметим, что $A^TA = (A^TA)^T$, т.е. полученная матрица симметрична.

Теорема 12.2 *Пусть $m \geq n$ и $\text{rank}(A) = n$. Вектор $\mathbf{x}$ — решение задачи наименьших квадратов $\min\limits_{\mathbf{y}} \|\mathbf{b} - A\mathbf{y}\|_2$ тогда и только тогда, когда $\mathbf{x}$ — решение системы $A^TA\mathbf{x} = A^T\mathbf{b}$.*

Доказательство. Из предыдущей теоремы следует существование и единственность такого вектора $\mathbf{x}$, что $A^T(\mathbf{b} - A\mathbf{x}) = 0$. Рассмотрим $\mathbf{y} = \mathbf{x} + \Delta$. В этом случае $\|\mathbf{b} - A\mathbf{y}\|_2^2 = (\mathbf{b} - A(\mathbf{x} + \Delta), \mathbf{b} - A(\mathbf{x} + \Delta)) = (\mathbf{b} - A\mathbf{x}, \mathbf{b} - A\mathbf{x}) + (A\Delta, A\Delta) + 2(A\Delta, \mathbf{b} - A\mathbf{x}) = \|\mathbf{b} - A\mathbf{x}\|_2^2 + (A\Delta, A\Delta) + 2(\Delta, A^T(\mathbf{b} - A\mathbf{x}))$. Следовательно, минимум достигается при $\Delta = 0$, т.е. на векторе $\mathbf{y} = \mathbf{x}$.

Метод QR-разложения. Метод нормального уравнения прост в реализации, однако в приближенной арифметике для почти вырожденных задач большой размерности может давать плохой результат. Например, в случае квадратной матрицы $A = A^T$ имеем $\text{cond}_2(A^TA) = \text{cond}_2^2(A)$. Таким образом, обусловленность исходной задачи возводится в квадрат, поэтому полученное численно решение может сильно отличаться от точного, если $\text{cond}(A) \gg 1$. Метод, основанный на QR-разложении матрицы A, более устойчив к вычислительной погрешности. Соответствующее разложение $A = QR$ при $Q^TQ = I$, $\det R \neq 0$ можно построить методом отражений или методом вращений.

Теорема 12.3 *Пусть $m > n$ и $\text{rank}(A) = n$. Тогда матрицу A можно привести к виду*

$$A = QR = (Q_1Q_2)\begin{pmatrix} R_1 \\ 0 \end{pmatrix} = Q_1R_1;$$

$$Q \in \mathbf{R}^{m\times m}, R \in \mathbf{R}^{m\times n},\ Q^TQ = I, \det R_1 \neq 0$$

с верхней треугольной $R_1 \in \mathbf{R}^{n\times n}$. В этом случае решение $\mathbf{x}$ задачи наименьших квадратов является решением системы $R_1\mathbf{x} = Q_1^T\mathbf{b}$.

Доказательство. Разложение с указанными Q и R можно построить методом отражений. Из метода нормального уравнения следует, что $\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b}$. Поэтому $\mathbf{x} = (R_1^TQ_1^TQ_1R_1)^{-1}R_1^TQ_1^T\mathbf{b} = R_1^{-1}Q_1^T\mathbf{b}$. Таким образом, искомый вектор $\mathbf{x}$ является решением системы $R_1\mathbf{x} = Q_1^T\mathbf{b}$.

Формально метод более трудоемкий, но построив однажды QR-разложение, можно быстро решать задачи с различными правыми частями.

Вырожденные задачи. Задача наименьших квадратов называется *вырожденной*, если $\mathrm{rank}(A) < n$. При численном решении вырожденных и почти вырожденных систем требуется изменить постановку задачи и соответственно применять иные методы. Рассмотрим следующий пример

$$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Для данного уравнения $m = n = 2$ и задача имеет семейство решений $\mathbf{x} = (1 - x_2, x_2)^T$, $\mathbf{r} = \mathbf{b} - A\mathbf{x} = (0, 1)^T$. Можно выбрать решения как с нормами порядка единицы, так и со сколь угодно большими. Однако, для возмущенной задачи

$$\begin{pmatrix} 1 & 1 \\ 0 & \varepsilon \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

формально близкой к исходной при малых ε, имеется единственное решение $\left(1 - \varepsilon^{-1}, \varepsilon^{-1}\right)^T$ с большой нормой порядка ε^{-1} и нулевой невязкой. Это означает, что сколь угодно малое возмущение элементов матрицы может существенно изменить структуру и норму решения.

Теорема 12.4 *Пусть* $\mathrm{rank}(A) = k$, $A \in \mathbf{R}^{m\times n}$, $m \geq n$ *и* $k < n$. *Тогда множество векторов* $\mathbf{x}$, *минимизирующих* $\|\mathbf{b} - A\mathbf{x}\|_2$, *образует* $(n-k)$*-мерное линейное подпространство.*

Доказательство. Пусть вектор $\mathbf{z} \in \ker(A)$ и $\dim(\ker(A)) = n-k$. Тогда $A\mathbf{z} = 0$, и если $\mathbf{x}$ минимизирует $\|\mathbf{b} - A\mathbf{x}\|_2$, то $\mathbf{x} + \mathbf{z}$ также минимизирует невязку, так как $\|\mathbf{b} - A(\mathbf{x} + \mathbf{z})\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2$.

Теорема 12.5 *Пусть ранг матрицы* $A \in \mathbf{R}^{m\times n}$, $m \geq n$, *в точной арифметике равен* $k < n$, *и первые* k *столбцов линейно независимы. Тогда матрицу можно привести к виду*

$$A = QR = (Q_1\, Q_2)\begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix},$$

$$Q \in \mathbf{R}^{m\times m}, Q^TQ = I,$$
$$R_{11} \in \mathbf{R}^{k\times k}, \quad \det R_{11} \neq 0, \quad R_{12} \in \mathbf{R}^{k\times(n-k)},$$

и для задачи наименьших квадратов с матрицей A *имеется семейство решений*

$$\mathbf{x} = \left(R_{11}^{-1}(Q_1^Tb - R_{12}\mathbf{x}_2), \mathbf{x}_2\right)^T,$$

где $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)^T, \mathbf{x}_1 \in \mathbf{R}^k, \mathbf{x}_2 \in \mathbf{R}^{n-k}$.

Доказательство. Искомое разложение $A = QR$ можно построить методом отражений. Решим задачу наименьших квадратов:

$$\|\mathbf{b} - A\mathbf{x}\|_2^2 = \|Q^T(\mathbf{b} - QR\mathbf{x})\|_2^2 = \|Q^T\mathbf{b} - R\mathbf{x}\|_2^2 =$$

$$= \left\|\begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix}\mathbf{b} - \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix}\mathbf{x}\right\|_2^2 = \left\|\begin{array}{c} Q_1^T\mathbf{b} - R_{11}\mathbf{x}_1 - R_{12}\mathbf{x}_2 \\ Q_2^T\mathbf{b} \end{array}\right\|_2^2 =$$

$$= \|Q_1^T\mathbf{b} - R_{11}\mathbf{x}_1 - R_{12}\mathbf{x}_2\|_2^2 + \|Q_2^T\mathbf{b}\|_2^2.$$

В данном случае умножение на Q^T корректно, т.к. Q ортогональная матрица, не меняющая при умножении евклидовой длины вектора. Теорема доказана.

Замечание. В общем случае метод QR-разложения не позволяет из всего множества решений выделить решение с минимальной евклидовой нормой, хотя выбор $\mathbf{x}_2 = 0$ может давать неплохое приближение.

Метод QR-разложения с выбором главного столбца. Преобразуем исходную задачу так, чтобы первые k столбцов полученной матрицы $\tilde{A}$ были линейно независимы, т.е. $\tilde{A} = AP = QR$, где P — некоторая матрица перестановок. Отсюда имеем $\tilde{A}\tilde{\mathbf{x}} = \mathbf{b}$, и далее найдем решение задачи наименьших квадратов. Перестановки столбцов в матрице A удобно проводить в процессе вычислений. Цель соответствующих перестановок — получить в матрице $R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \\ 0 & 0 \end{pmatrix}$ как можно лучше обусловленный блок R_{11} и как можно меньшие элементы в R_{22}. В приближенных вычислениях блок R_{22} отличен от нуля, хотя исходная задача могла быть неполного ранга.

Соответствующие вычисления проводятся на основе стандартного QR-разложения, например, методом отражений. На k-м шаге ($k = 1, \ldots, n-1$) в матрице $A^{(k)}$ выбирают столбец с номером j_k, $k \leq j_k \leq n$, с наибольшей величиной $\max\limits_{k\leq j\leq n}\left(\sum\limits_{i=k}^{m}\left(a_{ij}^{(k)}\right)^2\right)^{1/2}$. Если таких столбцов несколько, то берут произвольный из них. В матрице $A^{(k)}$ найденный столбец j_k переставляют с k-м столбцом. Далее реализуют очередной шаг QR-разложения.

Задача 12.1. Оценить величину элемента r_{nn} в методе QR-разложения с выбором главного столбца для

$$A = \begin{pmatrix} 1 & -1 & \dots & -1 \\ 0 & 1 & -1 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 \end{pmatrix}, \quad A \in \mathbf{R}^{n\times n}.$$

Метод сингулярного (SVD) **разложения.** Метод применяют для решения гарантированно наилучшим образом плохо обусловленных и вырожденных задач.

Утверждение 12.1 *Пусть A — матрица размерности $m \times n$, $m \geq n$. Тогда справедливо сингулярное разложение*

$$A = UDV^T = (U_1 U_2) \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T = U_1 \Sigma V^T,$$

где

U — ортогональная матрица размерности $m \times m$;

V — ортогональная матрица размерности $n \times n$;

Σ — диагональная матрица размерности $n \times n$ с элементами $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$.

Столбцы $\mathbf{u}_1, \dots, \mathbf{u}_m$ матрицы U называют *левыми сингулярными* векторами матрицы A, столбцы $\mathbf{v}_1, \dots, \mathbf{v}_n$ матрицы V — *правыми сингулярными* векторами, величины σ_i — *сингулярными числами*. Если $m = n$, то нулевой блок в матрице D отсутствует.

Построив SVD-разложение можно установить, является ли задача вырожденной ($\sigma_n = 0$), невырожденной ($\sigma_n \neq 0$), "хорошей" (σ_1/σ_n не слишком велико).

Если $m < n$, то сингулярное разложение строят для матрицы A^T. Если $m = n$ и $A = A^T$, то сингулярные числа $\sigma_i = |\lambda_i|$, т.е. с точностью до знака совпадают с собственными числами, сингулярные векторы $\mathbf{v}_i$ являются соответствующими собственными векторами.

Геометрическая интерпретация SVD-разложения. Рассмотрим оператор A, переводящий элемент $\mathbf{x} \in \mathbf{R}^n$ в элемент $\mathbf{y} \in \mathbf{R}^m$. Единичная сфера из $\mathbf{R}^n$ под действием A переходит в эллипсоид в подпространстве $\mathrm{span} < \mathbf{u_1}, \dots, \mathbf{u_n} > \subset \mathbf{R}^m$. Вектора $\mathbf{u}_i$, $i = 1, \dots, n$, задают полуоси эллипсоида, $\mathbf{v}_i \in \mathbf{R}^n$ — их прообразы, σ_i — коэффициенты удлинения векторов $\mathbf{v}_i$.

Алгебраическая интерпретация SVD-разложения. Рассмотрим оператор A, переводящий элемент $\mathbf{x} \in \mathbf{R}^n$ в элемент $\mathbf{y} \in \mathbf{R}^m$. В этом случае в пространстве $\mathbf{R}^n$ существует ортонормальный базис $\mathbf{v}_1, \dots, \mathbf{v}_n$, а в пространстве $\mathbf{R}^m$ — ортонормальный базис $\mathbf{u}_1, \dots, \mathbf{u}_m$ такие, что матрица оператора

A имеет диагональный вид, т.е. для произвольного вектора $\mathbf{x} = \sum_{i=1}^n \beta_i \mathbf{v}_i$ имеем $\mathbf{y} = A\mathbf{x} = \sum_{i=1}^n \sigma_i \beta_i \mathbf{u}_i$. Иначе говоря, всякая матрица A становится диагональной, если в области определения и в области значений подходящим образом выбраны ортогональные системы координат.

Теорема 12.6 *Пусть $A = U_1 \Sigma V^T$, $\mathrm{rank}(A) = n$. Тогда решение $\mathbf{x}$ задачи наименьших квадратов $\min\limits_{\mathbf{y}} \|\mathbf{b} - A\mathbf{y}\|_2$ имеет вид $\mathbf{x} = V\Sigma^{-1}U_1^T\mathbf{b}$.*

Доказательство. Из метода нормального уравнения следует, что решение $\mathbf{x}$ можно представить в виде

$\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b} = (V\Sigma U_1^T U_1 \Sigma V^T)^{-1} V\Sigma U_1^T \mathbf{b} = V\Sigma^{-1}U_1^T\mathbf{b}$.

Теорема 12.7 *Пусть матрица $A = (U_1\, U_2) \begin{pmatrix} \Sigma_k & 0 \\ 0 & 0 \end{pmatrix} (V_1\, V_2)^T$ имеет ранг $k < n$, здесь $U_1 \in \mathbf{R}^{m\times k}$, $\Sigma_k \in \mathbf{R}^{k\times k}$, $V_1 \in \mathbf{R}^{n\times k}$.*

Тогда пространство решений $\mathbf{x}$ задачи наименьших квадратов имеет вид $\mathbf{x} = V_1\Sigma_k^{-1}U_1^T\mathbf{b} + V_2\mathbf{z}$ с произвольным $\mathbf{z} \in \mathbf{R}^{n-k}$. Норма $\mathbf{x}$ минимальна при $\mathbf{z} = 0$.

Доказательство. $\|\mathbf{b} - A\mathbf{x}\|_2^2 = \|U^T(\mathbf{b} - UDV^T\mathbf{x})\|_2^2 =$

$$= \left\| \begin{matrix} U_1^T\mathbf{b} - (\Sigma_k\, 0) \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} \mathbf{x} \\ U_2^T\mathbf{b} \end{matrix} \right\|_2^2 = \|U_1^T\mathbf{b} - \Sigma_k V_1^T\mathbf{x}\|_2^2 + \|U_2^T\mathbf{b}\|_2^2.$$

Отсюда следует, что вектор $V_1\Sigma_k^{-1}U_1^T\mathbf{b}$ является решением знк, а все решения имеют вид $\mathbf{x} = V_1\Sigma_k^{-1}U_1^T\mathbf{b} + V_2\mathbf{z}$ для вех $\mathbf{z} \in \mathbf{R}^{n-k}$. Действительно, $\dim \ker A = n - k$; т.к. $V_1^TV_2 = 0$, то $AV_2 = UDV^TV_2 = 0$; следовательно, $\ker A = V_2\mathbf{z}$, $\mathbf{z} \in \mathbf{R}^{n-k}$.

Норма $\mathbf{x}$ минимальна при $\mathbf{z} = 0$, т.к. вектора $V_1\Sigma_k^{-1}U_1^T\mathbf{b}$ и $V_2\mathbf{z}$ ортогональны.

Теорема 12.8 *Пусть $A = UDV^T$, $\mathrm{rank}(A) = n$. Тогда*

$$\inf_{\substack{B\in\mathbf{R}^{m\times n},\\ \mathrm{rank}(B)=k\leq n}} \|A - B\|_2 = \sigma_{k+1}$$

и достигается на матрице

$$A_k = U \begin{pmatrix} \Sigma_k & 0 \\ 0 & 0 \end{pmatrix} V^T.$$

Доказательство. По построению (доказать это) матрица A_k имеет ранг k. Покажем, что $\|A - A_k\|_2 = \sigma_{k+1}$. Действительно,

$$\|A - A_k\|_2 = \left\| U \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{n-k} \\ 0 & 0 \end{pmatrix} V^T \right\|_2 \leq$$

$$\leq \|U\|_2 \left\| \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{n-k} \\ 0 & 0 \end{pmatrix} \right\|_2 \|V^T\|_2 = \sigma_{k+1}.$$

При этом

$$\|A - A_k\|_2 \geq \|(A - A_k)\mathbf{v}_{k+1}\|_2 = \sigma_{k+1}\|U\mathbf{e}_{k+1}\| = \sigma_{k+1},$$

для $\|\mathbf{v}_{k+1}\|_2 = 1$, где $\mathbf{e}_{k+1} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$. Следовательно, $\|A - A_k\|_2 = \sigma_{k+1}$. Покажем, что не существует матрицы B ранга k более близкой к A. Так как

$$\dim(\ker(B)) + \dim(\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}) = n + 1 > n,$$

то существует ненулевое пересечение соответствующих подпространств. Рассмотрим вектор $\mathbf{h} = \sum_{i=1}^{k+1} c_i \mathbf{v}_i$ из этого пересечения с нормой $\|\mathbf{h}\|_2^2 = \sum_{i=1}^{k+1} c_i^2 = 1$. Имеем

$$\|A - B\|_2 \geq \|(A - B)\mathbf{h}\|_2 = \|A\mathbf{h}\|_2 = \|UDV^T\mathbf{h}\|_2 = \|U_1\Sigma V^T\mathbf{h}\|_2.$$

Так как $V^TV = I$, то $V^T\mathbf{v}_i = \mathbf{e}_i$, следовательно,

$$V^T\mathbf{h} = \sum_{i=1}^{k+1} c_i\mathbf{e}_i, \quad \Sigma V^T\mathbf{h} = \sum_{i=1}^{k+1} \sigma_i c_i \mathbf{e}_i, \quad U_1\Sigma V^T\mathbf{h} = \sum_{i=1}^{k+1} \sigma_i c_i \mathbf{u}_i.$$

В результате

$$\|U_1\Sigma V^T\mathbf{h}\|_2 = (\sum_{i=1}^{k+1} \sigma_i^2 c_i^2)^{1/2} \geq \sigma_{k+1}(\sum_{i=1}^{k+1} c_i^2)^{1/2} = \sigma_{k+1}\|\mathbf{h}\|_2 = \sigma_{k+1}.$$

Теорема доказана.

Сформулируем правило решения задачи наименьших квадратов в приближенной арифметике. В реальных вычислениях все σ_i получатся (с учётом машинной точности) отличными от нуля, поэтому зафиксируем некоторое значение ε_0. Будем считать, что величины $\sigma_k < \varepsilon_0$ при $k = k_0 + 1, \ldots, n$ соответствуют погрешности вычислений, следовательно, можно заменить исходную задачу на задачу с матрицей A_k. Такой способ усечения матрицы A является оптимальным в том смысле, что полученная матрица A_k наиболее близка к A в норме $\|\cdot\|_2$.

Лекция 13. Линейная знк с ограничениями

Задача наименьших квадратов с линейными ограничениями–равенствами. Среди векторов $\mathbf{x}$, удовлетворяющих системе $B\mathbf{x} = \mathbf{d}$, требуется найти вектор $\mathbf{x}$, минимизирующий евклидову норму вектора невязки $\inf_{\tilde{\mathbf{x}}}\|\mathbf{b} - A\tilde{\mathbf{x}}\|_2$, т.е.

$$\inf_{B\tilde{\mathbf{x}}=\mathbf{d}} \|\mathbf{b} - A\tilde{\mathbf{x}}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2.$$

Здесь $A \in \mathbf{R}^{m\times n}$, $m \geq n$, $B \in \mathbf{R}^{p\times n}$. Если $\operatorname{rank}(B) = n$, то задача сводится к решению $B\mathbf{x} = \mathbf{d}$.

Метод исключения. Пусть $\operatorname{rank}(B) = p < n$. Построим QR-разложение следующего вида: $B^T = Q \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$, где $Q \in \mathbf{R}^{n\times n}$, $R_1 \in \mathbf{R}^{p\times p}$, $\det(R_1) \neq 0$. Тогда

$$B\mathbf{x} = \begin{pmatrix} R_1^T & 0 \end{pmatrix} Q^T\mathbf{x} = \mathbf{d},$$

можно сделать замену $Q^T\mathbf{x} = \mathbf{y}$ и переписать исходную задачу в виде

$$\begin{pmatrix} R_1^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \mathbf{d},$$

где $\mathbf{y}_1 \in \mathbf{R}^p$, $\mathbf{y}_2 \in \mathbf{R}^{n-p}$ — новые переменные. С учетом $AQ\,Q^T\mathbf{x} \equiv (A_1A_2)\mathbf{y}$, $A_1 \in \mathbf{R}^{m\times p}$, $A_2 \in \mathbf{R}^{m\times(n-p)}$, исходная задача минимизации принимает вид

$$\inf_{B\tilde{\mathbf{x}}=\mathbf{d}} \|\mathbf{b} - A\tilde{\mathbf{x}}\|_2 = \inf_{\substack{R_1^T\tilde{\mathbf{y}}_1=\mathbf{d};\\ \tilde{\mathbf{y}}_2}} \|\mathbf{b} - A_1\tilde{\mathbf{y}}_1 - A_2\tilde{\mathbf{y}}_2\|_2 = \inf_{\tilde{\mathbf{y}}_2} \|(\mathbf{b} - A_1\mathbf{y}_1) - A_2\tilde{\mathbf{y}}_2\|_2,$$

т.е. сводится к стандартной знк относительно $\tilde{\mathbf{y}}_2$ при вычисленном из системы $R_1^T\mathbf{y}_1 = \mathbf{d}$ векторе $\mathbf{y}_1$. Найденные таким образом компоненты $(\mathbf{y}_1, \mathbf{y}_2)^T$ позволяют получить искомый ответ $\mathbf{x} = Q\mathbf{y}$.

Замечание. Пусть $\operatorname{rank}(B) = r < p < n$, и первые r столбцов линейно независимы. Тогда

$$B^T = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix}, \; B\mathbf{x} = \begin{pmatrix} R_{11}^T & 0 \\ R_{12}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{pmatrix},$$

где $\mathbf{y}_1 \in \mathbf{R}^r$, $\mathbf{y}_2 \in \mathbf{R}^{n-r}$ — новые переменные (для вычисления матриц можно применить QR-разложение с выбором главного столбца). Если полученная система окажется несовместной, то решение исходной задачи не существует. Иначе исходная задача с ограничениями сводится к стандартной знк относительно $\tilde{\mathbf{y}}_2$ при вычисленном из системы $R_{11}^T\mathbf{y}_1 = \mathbf{d}_1$ векторе $\mathbf{y}_1$.

Метод обобщенного сингулярного разложения. Решение задачи наименьших квадратов с линейными ограничениями–равенствами будем строить на основании следующего утверждения.

Утверждение 13.1 *(Обобщенное сингулярное разложение — GSVD). Пусть* $A \in \mathbf{R}^{m\times n}$, $m \geq n$, $B \in \mathbf{R}^{p\times n}$, $p \leq n$. *Тогда справедливо разложение*

$$A = UD_AX^{-1}, \quad B = VD_BX^{-1},$$

где

U — *ортогональная матрица размерности* $m \times m$;
V — *ортогональная матрица размерности* $p \times p$;
X — *обратимая матрица размерности* $n \times n$;
$D_A = \mathrm{diag}(\alpha_1, \ldots, \alpha_n), \alpha_1 \geq \ldots \geq \alpha_n \geq 0, D_A \in \mathbf{R}^{m\times n}$;
$D_B = \mathrm{diag}(\beta_1, \ldots, \beta_p), \beta_1 \geq \ldots \geq \beta_p \geq 0, D_B \in \mathbf{R}^{p\times n}$.

Данное утверждение можно также обобщить на случай $m < n$ и $p > n$.

Пусть обобщенное сингулярное разложение построено:

$$U = (\mathbf{u}_1 \ldots \mathbf{u}_m), V = (\mathbf{v}_1 \ldots \mathbf{v}_p), X = (\mathbf{x}_1 \ldots \mathbf{x}_n).$$

Тогда исходная задача эквивалента следующей:

$$\begin{cases} D_B\mathbf{y} = V^T\mathbf{d}, \quad \mathbf{y} = X^{-1}\mathbf{x}, \\ \|D_A\mathbf{y} - U^T\mathbf{b}\|_2 \to \inf. \end{cases}$$

Рассмотрим случай $p \leq n$, $\mathrm{rank}B = p$, $\mathrm{rank}(A) = n$. Решая первое уравнение для $\mathbf{y}_1, \ldots, \mathbf{y}_p$ точно и минимизируя по оставшимся переменным $\mathbf{y}_{p+1}, \ldots, \mathbf{y}_n$ второе уравнение, получаем, что искомое решение имеет вид:

$$y_i = \frac{(\mathbf{v}_i, \mathbf{d})}{\beta_i}, i = 1, \ldots, p, \quad y_i = \frac{(\mathbf{u}_i, \mathbf{b})}{\alpha_i}, i = p+1, \ldots, n.$$

Отсюда находим:

$$\mathbf{x} = X\mathbf{y} = \sum_{i=1}^{p} \frac{(\mathbf{v}_i, \mathbf{d})}{\beta_i}\mathbf{x}_i + \sum_{i=p+1}^{n} \frac{(\mathbf{u}_i, \mathbf{b})}{\alpha_i}\mathbf{x}_i.$$

Отметим, что построение GSVD — процедура дорогостоящая.

Для случая $\mathrm{rank}(B) < p$, $\mathrm{rank}(A) < n$ решение строится аналогично.

Метод взвешиванием. Исходная система с ограничением заменяется на следующую задачу наименьших квадратов:

$$\begin{pmatrix} \lambda B \\ A \end{pmatrix}\mathbf{x} = \begin{pmatrix} \lambda\mathbf{d} \\ \mathbf{b} \end{pmatrix} \Leftrightarrow \left\| \begin{matrix} \lambda B\mathbf{x} - \lambda\mathbf{d}, \\ A\mathbf{x} - \mathbf{b} \end{matrix} \right\|_2 \to \inf$$

с некоторым $\lambda \gg 1$. И так как для ее решения можно применить один из стандартных алгоритмов, то данный подход прост в реализации. Покажем,

что полученное решение сходится к решению исходной задачи при увеличении λ (однако, при увеличении λ обусловленность системы стремится к бесконечности).

Пусть имеется GSVD для матриц A, B. Тогда исходная задача принимает вид:

$$\|\lambda D_B\mathbf{y} - \lambda V^T\mathbf{d}\|_2^2 + \|D_A\mathbf{y} - U^T\mathbf{b}\|_2^2 \to \inf, \quad \mathbf{x} = X\mathbf{y},$$

или покомпонентно,

$$\sum_{i=1}^{p}(\beta_i y_i - (\mathbf{v}_i, \mathbf{d}))^2\lambda^2 + \sum_{i=1}^{n}(\alpha_i y_i - (\mathbf{u}_i, \mathbf{b}))^2 + \sum_{i=n+1}^{m}(\mathbf{u}_i, \mathbf{b})^2 \to \inf.$$

Дифференцируя по y_i и приравнивая результат к нулю, находим оптимальные значения y_i. В результате имеем

$$\mathbf{x} = \sum_{i=1}^{p} \frac{\alpha_i(\mathbf{u}_i, \mathbf{b}) + \lambda^2\beta_i(\mathbf{v}_i, \mathbf{d})}{\alpha_i^2 + \lambda^2\beta_i^2}\mathbf{x}_i + \sum_{i=p+1}^{n} \frac{(\mathbf{u}_i, \mathbf{b})}{\alpha_i}\mathbf{x}_i.$$

Отсюда следует оценка скорости сходимости $\mathbf{x}(\lambda) \to \mathbf{x}$ при $\lambda \to \infty$.

Задача наименьших квадратов с ограничениями типа квадратных неравенств. Пусть $A \in \mathbf{R}^{m\times n}$, $B \in \mathbf{R}^{p\times n}$, $n \leq m$, $p \leq n$ и требуется найти такой вектор $\mathbf{x}$, что

$$\begin{cases} \|A\mathbf{x} - \mathbf{b}\|_2 \to \inf, \\ \|Bx - \mathbf{d}\|_2 \leq E. \end{cases}$$

Дополнительное ограничение выделяет в $\mathbf{R}^n$ эллипсоид — область допустимых векторов $\mathbf{x}$.

Пусть GSVD для матриц A, B найдено. Тогда исходная задача принимает вид

$$\begin{cases} \|D_A\mathbf{y} - U^T\mathbf{b}\|_2 \to \inf, \\ \|D_B\mathbf{y} - V^T\mathbf{d}\|_2 \leq E, \end{cases}$$

т.е.

$$\begin{cases} \sum\limits_{i=1}^{k_A}(\alpha_i y_i - (\mathbf{u}_i, \mathbf{b}))^2 + \sum\limits_{i=k_A+1}^{m}(\mathbf{u}_i, \mathbf{b})^2 \to \inf, \\ \sum\limits_{i=1}^{k_B}(\beta_i y_i - (\mathbf{v}_i, \mathbf{d}))^2 + \sum\limits_{i=k_B+1}^{p}(\mathbf{v}_i, \mathbf{d})^2 \leq E^2, \end{cases}$$

где $k_A = \mathrm{rank(A)}$, т.е. $\alpha_j = 0$ при $j = k_A + 1, \ldots, n$; $k_B = \mathrm{rank(B)}$, т.е. $\beta_j = 0$ при $j = k_B + 1, \ldots, p$. Найдем решение $(y_1, \ldots, y_n)$ полученной задачи.

Если $\sum\limits_{i=k_B+1}^{p}(\mathbf{v}_i, \mathbf{d})^2 > E^2$, то решения не существует.

Если $\sum\limits_{i=k_B+1}^{p}(\mathbf{v}_i,\mathbf{d})^2 = E^2$, то решение имеет вид:

$$y_i = \begin{cases} (\mathbf{v}_i,\mathbf{d})/\beta_i, & 1 \le i \le k_B, \\ (\mathbf{u}_i,\mathbf{b})/\alpha_i, & k_B+1 \le i \le k_A, \\ \text{произвольные} & \text{для} \quad k_A+1 \le i \le n. \end{cases}$$

Выбор произвольных компонент нулевыми дает решение минимальной нормы. Здесь и далее считается, что если указан диапазон $i_1 \le i \le i_2$ для $i_1 > i_2$, то соотвествующих i не существует.

Пусть $\sum\limits_{i=k_B+1}^{p}(\mathbf{v}_i,\mathbf{d})^2 < E^2$. Тогда выберем первые компоненты вектора $\mathbf{y}$ из условия $\|D_A\mathbf{y} - U^T\mathbf{b}\|_2 \to \inf$, а оставшиеся — из условия минимизации величины $\|D_B\mathbf{y} - V^T\mathbf{d}\|_2$

$$y_i = \begin{cases} (\mathbf{u}_i,\mathbf{b})/\alpha_i, & 1 \le i \le k_A, \\ (\mathbf{v}_i,\mathbf{d})/\beta_i, & k_{A+1} \le i \le k_B, \\ \text{произвольные} & \text{для остальных } i. \end{cases} \qquad (1)$$

Если для него будет выполнено требуемое ограничение $\|D_B\mathbf{y} - V^T\mathbf{d}\|_2 \le E$, то найденный вектор является решением.

Иначе искомое решение лежит на границе эллипсоида, т.к. ищется минимум квадратичной функции на эллипсоиде, при этом ее абсолютный минимум определяемый по формуле (1), не удовлетворяет ограничению $\|D_B\mathbf{y} - V^T\mathbf{d}\|_2 \le E$, т.е. эллипсоиду не принадлежит. В этом случае решение может быть найдено из условий

$$\begin{cases} \sum\limits_{i=1}^{k_A}(\alpha_i y_i - (\mathbf{u}_i,\mathbf{b}))^2 + \sum\limits_{i=k_A+1}^{m}(\mathbf{u}_i,\mathbf{b})^2 = f(\mathbf{y}) \to \inf, \\ \sum\limits_{i=1}^{k_B}(\beta_i y_i - (\mathbf{v}_i,\mathbf{d}))^2 + \sum\limits_{i=k_B+1}^{p}(\mathbf{v}_i,\mathbf{d})^2 - E^2 = \varphi(\mathbf{y}) = 0. \end{cases}$$

Для решения данной задачи применим принцип множителей Лагранжа. Для функции Лагранжа

$$L(\lambda,\mathbf{y}) = f(\mathbf{y}) + \lambda\varphi(\mathbf{y})$$

условия $\partial L/\partial y_i = 0, \quad 1 \le i \le n$ имеют вид

$$2\alpha_i(\alpha_i y_i - (\mathbf{u}_i,\mathbf{b})) + \lambda 2\beta_i(\beta_i y_i - (\mathbf{v}_i,\mathbf{d})) = 0,$$

где $\alpha_i = 0, k_A < i \le n$; $\beta_i = 0, k_B < i \le n$. Отсюда

$$y_i(\lambda) = \begin{cases} \dfrac{\alpha_i(\mathbf{u}_i,\mathbf{b}) + \lambda\beta_i(\mathbf{v}_i,\mathbf{d})}{\alpha_i^2 + \lambda\beta_i^2}, & 1 \le i \le \max\{k_A,k_B\}, \\ \text{произвольные} & \text{для остальных } i. \end{cases}$$

При этом множитель Лагранжа λ определяется из *секулярного уравнения*: $\varphi\big(\mathbf{y}(\lambda)\big) = 0$. Так как в рассматриваемом случае $\varphi\big(\mathbf{y}(0)\big) > 0$, а функция $\varphi\big(\mathbf{y}(\lambda)\big)$ является монотонно убывающей при $\lambda > 0$, следовательно, существует единственное λ^*: $\varphi\big(\mathbf{y}(\lambda^*)\big) = 0$. Его можно найти, например, методом Ньютона (начальное приближение определяется методом бисекции). Далее вычисляем $\mathbf{y}(\lambda^*)$ и получаем искомое решение $\mathbf{x} = X\mathbf{y}(\lambda^*)$.

Отметим, что если $B = I, \mathbf{d} = 0, E = 1$, то исходная задача соответствует задаче минимизации на единичном шаре.

Лекция 14. Итерационные методы решения слау

Рассмотрим класс итерационных методов решения систем линейных алгебраических уравнений, основанный на сжимающем свойстве оператора перехода. Различные постановки задачи минимизации нормы оператора перехода приводят к различным алгоритмам расчета.

Метод простой итерации. Преобразуем систему линейных алгебраических уравнений

$$A\,\mathbf{x} = \mathbf{b} \qquad (1)$$

с невырожденной матрицей A к виду

$$\mathbf{x} = B\,\mathbf{x} + \mathbf{c}. \qquad (2)$$

Если решение системы (2) находится как предел последовательности

$$\mathbf{x}^{k+1} = B\,\mathbf{x}^k + \mathbf{c}, \qquad (3)$$

то такой процесс называется *двухслойным итерационным методом*, или *методом простой итерации*. При этом B называется *оператором перехода*. Рассмотрим общий способ перехода от задачи (1) к задаче (2). Всякая система

$$\mathbf{x} = \mathbf{x} - D\,(A\,\mathbf{x} - \mathbf{b}) \qquad (4)$$

имеет вид (2) и при $\det(D) \ne 0$ равносильна (1). В то же время всякая система (2), равносильная (1), записывается в виде (4) с матрицей $D = (I - B)\,A^{-1}$.

Справедливы следующие теоремы о сходимости метода.

Теорема 14.1 *Пусть $\|B\| = q < 1$. Тогда система уравнений (2) имеет единственное решение, и итерационный процесс (3) сходится с произвольного начального приближения к решению задачи со скоростью геометрической прогрессии: $\|\mathbf{x} - \mathbf{x}^k\| \le q^k\|\mathbf{x} - \mathbf{x}^0\|$.*

Доказательство. 1. Известно, что система линейных уравнений имеет единственное решение, если однородная задача имеет единственное решение. Для задачи (2) получаем

$$\mathbf{x} = B\mathbf{x} + \mathbf{c} \Rightarrow \|\mathbf{x}\| \le \|B\|\|\mathbf{x}\| + \|\mathbf{c}\| \Rightarrow \|\mathbf{x}\| \le \frac{\|\mathbf{c}\|}{1 - \|B\|},$$

т.к. $\|B\| < 1$. Отсюда следует, что при $\|\mathbf{c}\| = 0$ имеем $\|\mathbf{x}\| = 0$.

2. Пусть $\mathbf{x}$ — точное решение, а $\mathbf{x}^k$ — k-е приближение, тогда

$$\mathbf{x} = B\mathbf{x} + c, \quad \mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c},$$

и для вектора ошибки $\mathbf{z}^k = \mathbf{x} - \mathbf{x}^k$ верна оценка:

$$\mathbf{z}^k = B\mathbf{z}^{k-1}, \ \|\mathbf{z}^k\| \le \|B\|^k\|\mathbf{z}^0\| \le q^k\|\mathbf{z}^0\| \to 0$$

при $k \to \infty$.

Теорема 14.2 *Пусть система уравнений* (2) *имеет единственное решение. Итерационный процесс* (3) *сходится к решению системы* (2) *при любом начальном приближении тогда и только тогда, когда все собственные значения матрицы B по модулю меньше* 1.

Доказательство. Напомним, что для произвольной матрицы B существуют S и Λ, такие, что

$$S^{-1}BS = \Lambda = \begin{pmatrix} \lambda_1 & \alpha_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \alpha_2 & 0 & \dots \\ . & . & . & . & . \\ 0 & \dots & 0 & \lambda_{n-1} & \alpha_{n-1} \\ 0 & \dots & & & \lambda_n \end{pmatrix}.$$

Данное разложение называется Жордановой нормальной формой. Здесь λ_i — собственные числа матрицы B, а α_i=0 или 1. Пусть $\eta > 0$. Тогда для матрицы $\frac{1}{\eta}B$ с собственными числами $\frac{\lambda_i}{\eta}$ существует $\tilde{S}$:

$$\tilde{S}^{-1}\frac{1}{\eta}B\tilde{S} = \begin{pmatrix} \frac{\lambda_1}{\eta} & \alpha_1 & 0 & \dots & 0 \\ 0 & \frac{\lambda_2}{\eta} & \alpha_2 & 0 & \dots \\ . & . & . & . & . \\ 0 & \dots & 0 & \frac{\lambda_{n-1}}{\eta} & \alpha_{n-1} \\ 0 & \dots & & & \frac{\lambda_n}{\eta} \end{pmatrix} \Rightarrow$$

$$\tilde{S}^{-1}B\tilde{S} = \begin{pmatrix} \lambda_1 & \eta\alpha_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \eta\alpha_2 & 0 & \dots \\ . & . & . & . & . \\ 0 & \dots & 0 & \lambda_{n-1} & \eta\alpha_{n-1} \\ 0 & \dots & & & \lambda_n \end{pmatrix} = \tilde{\Lambda}.$$

Пусть $\max_i|\lambda_i| < q < 1$. Покажем, что метод сходится. Для $\eta = q - \max_i|\lambda_i| > 0$ имеем

$$\|\tilde{\Lambda}\|_\infty = \max_{i,\alpha_n=0}(|\lambda_i| + \eta\alpha_i) \le \max_i|\lambda_i| + \eta = q.$$

И так как $\mathbf{z}^k = B\mathbf{z}^{k-1} = B^k\mathbf{z}^0 = \tilde{S}\tilde{\Lambda}^k\tilde{S}^{-1}\mathbf{z}^0$, то

$$\|\mathbf{z}^k\|_\infty \le \|\tilde{S}\|_\infty\|\tilde{S}^{-1}\|_\infty q^k\|\mathbf{z}^0\|_\infty = \mathrm{cond}_\infty(\tilde{S})q^k\|\mathbf{z}^0\|_\infty,$$

т.е. метод сходится.

Пусть теперь для некоторого i_0 выполняется $|\lambda_{i_0}| \ge 1$, $B\mathbf{e}_{i_0} = \lambda_{i_0}\mathbf{e}_{i_0}$. Тогда для начального вектора $\mathbf{x}^0 = \mathbf{x} + c\,\mathbf{e}_{i_0}$ находим $\|\mathbf{z}^k\| = \|B^k\mathbf{z}^0\| = \|cB^k\mathbf{e}_{i_0}\| = |c\lambda_{i_0}^k|\,\|\mathbf{e}_{i_0}\|$, следовательно, норма ошибки не стремится к нулю при $k \to \infty$.

Оптимальный линейный одношаговый метод. Для систем со знакоопределёнными матрицами метод (3) обычно строится в виде

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\,\mathbf{x}^k = \mathbf{b}, \quad \text{т.е.} \quad B = I - \tau\,A, \ \ \mathbf{c} = \tau\,\mathbf{b}. \tag{5}$$

Здесь τ — итерационный параметр.

Так как точное решение $\mathbf{x}$ удовлетворяет уравнению (5), то имеет место следующий закон изменения вектора ошибки $\mathbf{z}^k = \mathbf{x} - \mathbf{x}^k$:

$$\mathbf{z}^{k+1} = (I - \tau A)\,\mathbf{z}^k, \quad \|\mathbf{z}^{k+1}\| \le \|I - \tau A\|\,\|\mathbf{z}^k\|, \ k = 0, 1, 2, \dots .$$

Оптимальный итерационный параметр τ_0 ищется из условия минимума оператора перехода: $\min\limits_\tau \|I - \tau A\|$. Данная минимизационная задача решается явно при $A = A^T > 0$. В этом случае в качестве нормы $\|\cdot\|$ удобно взять евклидову $\|\cdot\|_2$ норму. Тогда подчиненная ей матричная норма имеет вид

$$\|A\|_2 = \sup_{x\ne0}\frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\max|\lambda(A^TA)|} = \max\lambda(A),$$

а соответствующая оптимизационная задача сводится к следующей:

$$\min_\tau \|I - \tau A\|_2 = \min_\tau\Big(\max_{\lambda(A)}|1 - \tau\lambda(A)|\Big) = q_0.$$

Теорема 14.3 *При условии $A = A^T > 0$, $\lambda(A) \in [m, M]$ и $0 < m \le M < \infty$ оптимальное значение параметра равно $\tau_0 = \frac{2}{m+M}$ и верна оценка*

$$\|\mathbf{x} - \mathbf{x}^k\|_2 \le q_0^k \|\mathbf{x} - \mathbf{x}^0\|_2, \quad q_0 = \frac{M-m}{M+m} < 1.$$

Доказательство. Пусть $\mathbf{z}^k = \mathbf{x} - \mathbf{x}^k$. Тогда уравнение для вектора ошибки имеет вид $\mathbf{z}^k = (I - \tau A)\mathbf{z}^{k-1}$. Отсюда получаем $\|\mathbf{z}^k\|_2 \le \|(I - \tau A)\|_2 \|\mathbf{z}^{k-1}\|_2$. Будем минимизировать по параметру τ евклидову норму оператора перехода:

$$\min_\tau \|I - \tau A\|_2 = \min_\tau \max_{\lambda(A)\in[m,M]} |1 - \tau\lambda(A)|.$$

Найдем решение данной минимизационной задачи:

$$|1 - \tau_0 m| = |1 - \tau_0 M|; \tau_0 = \frac{2}{m+M},\ q_0 = \|I - \tau_0 A\|_2 = \frac{M-m}{M+m}.$$

При этом $\|\mathbf{x} - \mathbf{x}^k\|_2 \le q_0^k \|\mathbf{x} - \mathbf{x}^0\|_2$.

Замечание. Минимизационная задача по сути означает, что требуется найти полином первой степени, наименее уклоняющийся от нуля на отрезке $[m, M]$, имеющий ровно один корень и равный единице свободный член. Ее решением является нормированный многочлен Чебышева $\tilde{T}_1^{[m,M]}(\lambda) = \frac{T_1\left(\frac{2\lambda - M - m}{M-m}\right)}{T_1\left(-\frac{M+m}{M-m}\right)}$.

Оптимальный линейный N-шаговый метод. Будем считать, что в итерационном алгоритме

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau_{k+1}} + A\,\mathbf{x}^k = \mathbf{b}$$

допускается циклическое изменение (с периодом N) параметра τ в зависимости от номера итерации

$$\tau_1, \tau_2, \ldots, \tau_N, \tau_1, \tau_2, \ldots .$$

В этом случае после N итераций для вектора ошибки имеем:

$$\mathbf{z}^{k+N} = \prod_{j=1}^{N} (I - \tau_j A)\,\mathbf{z}^k, \quad \|\mathbf{z}^{k+N}\|_2 \le \left\| \prod_{j=1}^{N} (I - \tau_j A) \right\|_2 \|\mathbf{z}^k\|_2.$$

Будем искать набор τ_j, $j = 1, \ldots, N$, из условия минимума нормы оператора перехода.

Теорема 14.4 *При условии $A = A^T > 0$, $\lambda(A) \in [m, M]$ и $0 < m \le M < \infty$ оптимальные значения параметров равны обратным величинам корней многочлена Чебышёва степени N на отрезке $[m, M]$: $\tau_j^{-1} = \frac{M+m}{2} + \frac{M-m}{2}\cos\frac{\pi(2j-1)}{2N}$, и после N итераций для погрешности справедлива оценка:*

$$\|\mathbf{x} - \mathbf{x}^N\|_2 \le \frac{2}{q_1^{-N} + q_1^N} \|\mathbf{x} - \mathbf{x}^0\|_2 \le 2q_1^N \|\mathbf{x} - \mathbf{x}^0\|_2, \quad q_1 = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}.$$

Доказательство. Для вектора ошибки имеем:

$$\mathbf{z}^{k+N} = \prod_{j=1}^{N} (I - \tau_j A)\mathbf{z}^k \Rightarrow \|\mathbf{z}^{k+N}\|_2 \le \| \prod_{j=1}^{N} (I - \tau_j A)\|_2 \|\mathbf{z}^k\|_2.$$

Будем минимизировать по параметрам τ_j евклидову норму оператора перехода. Так как для $A = A^T$ имеем $\lambda(P_N(A)) = P_N(\lambda(A))$, то минимизационную задачу можно переписать следующим образом:

$$\min_{\tau_j}\Big(\max_{\lambda\in[m,M]} | \prod_{j=1}^{N} (1 - \tau_j \lambda)| \Big) = q_N.$$

Данная задача сводится к нахождению многочлена $P_N(\lambda)$ степени N, наименее уклоняющегося от нуля на $[m, M]$ в классе $P(0) = 1$. Решением является нормированный многочлен Чебышёва $\tilde{T}_N^{[m,M]}(\lambda) = \frac{T_N\left(\frac{2\lambda - M - m}{M-m}\right)}{T_N\left(-\frac{M+m}{M-m}\right)}$, где $T_N(x) = \cos(N \arccos x)$. Отсюда следует, что τ_j^{-1} равны корням многочлена Чебышёва на отрезке $[m, M]$, т.е. $\{\tau_j^{-1}\}_{j=1}^N = \{\frac{m+M}{2} + \frac{M-m}{2}\cos\frac{\pi(2j-1)}{2N}\}_{j=1}^N$, а скорость сходимости алгоритма характеризуется величиной

$$q_N = \frac{2}{q_1^{-N} + q_1^N} = \frac{2q_1^N}{1 + q_1^{2N}} \le 2q_1^N,\ q_1 = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}.$$

Теорема доказана.

При численной реализации N-шагового процесса для устойчивости необходимо специальным образом перемешивать значения параметров τ_j. Напомним алгоритм для $N = 2^p$.

Пусть $\tau_{N+1-j} = (\frac{m+M}{2} + \frac{M-m}{2}\cos\frac{\pi(2j-1)}{2N})^{-1}$, т.е. нумерация τ_j ведется "от большего к меньшему" . Тогда

для $N = 2$ устойчивая последовательность имеет вид $(2, 1)$;

пусть для $N = 2^{p-1}$ имеем: $(j_1, j_2, \ldots, j_{2^{p-1}})$;

тогда для $N = 2^p$: $(2^p + 1 - j_1, j_1, \ldots, 2^p + 1 - j_{2^{p-1}}, j_{2^{p-1}})$.

Например: $(3, 2, 4, 1)$, $(6, 3, 7, 2, 5, 4, 8, 1)$.

Отметим, что мы построили неулучшаемый по скорости сходимости метод в данном классе. Однако указанное уменьшение погрешности будет получено только через N итераций, а реализация метода требует знания границ спектра m, M.

Для оценки границ спектра часто используется следующее утверждение.

Утверждение 14.1 *(Теорема о кругах Гершгорина). Все собственные значения матрицы A принадлежат объединению кругов*

$$|z - a_{ii}| \le \sum_{j \ne i} |a_{ij}|, \quad i = 1, \ldots, n\,.$$

Если указанное объединение кругов распадается на несколько связных частей, то каждая такая часть содержит столько собственных значений, сколько кругов ее составляют.

Отметим, что первая часть теоремы следует из покоординатной записи определения собственного вектора $(A\mathbf{x})_i = \lambda x_i$, $i = 1, \ldots, n$:

$$\sum_{j=1}^{n} a_{ij} x_j = \lambda x_i, \quad (\lambda - a_{ii}) = \sum_{j \ne i} a_{ij} x_j / x_i.$$

Так как равенство выполняется для всех i, то искомая оценка для фиксированного λ верна для $i = i(\lambda) : |x_j| \le |x_i|$.

Полная информация о собственных числах позволяет формально построить следующий точный метод.

Лемма 14.1 *Пусть для невырожденной матрицы простой структуры A порядка $n \times n$ известны все собственные значения λ. Тогда в точной арифметике итерационный метод с переменными параметрами $\tau_k = \lambda_k^{-1}$, $k = 1, 2, \ldots, n$, не более чем за n шагов приведет к точному решению системы $A\mathbf{x} = \mathbf{b}$.*

Доказательство следует из разложения ошибки $\mathbf{x} - \mathbf{x}^k$ при $k = 0, 1, \ldots, n$ по базису из собственных векторов матрицы A.

Отметим, что метод имеет теоретическое значение, так как требует n умножений на матицу A и весьма чувствителен к вычислительной погрешности. При этом задача нахождения собственных чисел в общем случае существенно сложнее задачи решения системы линейных уравнений.

Лекция 15. Вариационные методы решения слау

Данный класс методов строится как методы минимизации некоторого функционала, минимум которого достигается на решении исходной системы линейных уравнений. Конкретный вид функционала и алгоритм минимизации определяют параметры итерационного процесса.

Метод наискорейшего градиентного спуска. Основная цель — построить оптимальный одношаговый метод, не требующий для реализации информации о границах спектра.

Лемма 15.1 *Пусть $A = A^T > 0$ и $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$ — квадратичная функция. Равенство $F(\mathbf{x}^*) = \min\limits_{\mathbf{x}} F(\mathbf{x})$ выполнено тогда и только тогда, когда $\mathbf{x}^*$ — решение системы $A\mathbf{x} = \mathbf{b}$.*

Доказательство. Пусть $\mathbf{x} = \mathbf{x}^* + \Delta$. Преобразуем выражение для $F(x)$. С учетом $A\mathbf{x}^* = \mathbf{b}$ имеем

$$(A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x}) = (A\mathbf{x}^*, \mathbf{x}^*) - 2(A\mathbf{x}^* - b, \Delta) + (A\Delta, \Delta) - 2(\mathbf{b}, \mathbf{x}^*) = \\ = (A\mathbf{x}^*, \mathbf{x}^*) + (A\Delta, \Delta) - 2(\mathbf{b}, \mathbf{x}^*).$$

Если $A > 0$, то $(A\Delta, \Delta) > 0$ при $\Delta \ne 0$, поэтому функция $F(\mathbf{x})$ имеет единственный минимум при $\Delta = 0$.

Замечание. Справедлива формула $\operatorname{grad} F(\mathbf{x}) = 2\,(A\mathbf{x} - \mathbf{b})$, проверяемая покомпонентным дифференцированием $\partial F(\mathbf{x})/\partial x_i$.

Замечание. Вдоль вектора $\operatorname{grad} F(x)$ происходит наискорейшее изменение функции $F(x)$. Действительно, пусть $\mathbf{e}$ — некоторый единичный вектор. Разложим в ряд $F(\mathbf{x} + \delta\,\mathbf{e})$ при условии малости величины $\delta \in \mathbf{R}^1$, имеем

$$F(x_1 + \delta e_1, \ldots, x_n + \delta e_n) = F(\mathbf{x}) + \delta \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(\mathbf{x}) e_i + O(\delta^2) \sim$$

$$\sim F(\mathbf{x}) + \delta(\operatorname{grad} F(\mathbf{x}), \mathbf{e}) = F(\mathbf{x}) + \delta \cos(\alpha) \|\operatorname{grad} F(\mathbf{x})\| \|\mathbf{e}\|.$$

Отсюда следует, что в точке $\mathbf{x}$ наискорейшему убыванию соответствует $\alpha = -\pi$, т.е. $\mathbf{e} = -\operatorname{grad} F(\mathbf{x}) / \|\operatorname{grad} F(\mathbf{x})\|$.

Лемма 15.2 *Пусть $A = A^T > 0$ и решение системы $A\mathbf{x}^* = \mathbf{b}$ ищется как точка минимума функционала $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$ по следующему алгоритму:*

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \delta_{k+1} \operatorname{grad} F(\mathbf{x}^k),$$

где параметр δ_{k+1} выбирается из условия минимума величины $F(\mathbf{x}^{k+1})$. Тогда $2\delta_{k+1} = (\mathbf{r}^k, \mathbf{r}^k)/(A\mathbf{r}^k, \mathbf{r}^k)$.

Доказательство. Подставляя $\operatorname{grad} F(\mathbf{x}) = 2(A\mathbf{x} - \mathbf{b})$ в выражение для $\mathbf{x}^{k+1}$, получаем

$$\mathbf{x}^{k+1} = \mathbf{x}^k - 2\delta_{k+1}(A\mathbf{x}^k - \mathbf{b}) = \mathbf{x}^k + \tau_{k+1}\mathbf{r}^k,$$

где $\tau_{k+1} = 2\delta_{k+1}, \mathbf{r}^k = \mathbf{b} - A\mathbf{x}^k$. Далее

$$F(\mathbf{x}^{k+1}) = (A(\mathbf{x}^k + \tau_{k+1}\mathbf{r}^k), (\mathbf{x}^k + \tau_{k+1}\mathbf{r}^k)) - 2(\mathbf{b}, \mathbf{x}^k + \tau_{k+1}\mathbf{r}_k) =$$
$$= (A\mathbf{x}^k, \mathbf{x}^k) + 2\tau_{k+1}(A\mathbf{x}^k, \mathbf{r}^k) - 2\tau_{k+1}(\mathbf{b}, \mathbf{r}^k) + \tau_{k+1}^2(A\mathbf{r}^k, \mathbf{r}^k) - 2(\mathbf{b}, \mathbf{x}^k).$$

Из условия $F'_{\tau_{k+1}}(\mathbf{x}^{k+1}) = 0$ находим $\tau_{k+1} = (\mathbf{r}^k, \mathbf{r}^k)/(A\mathbf{r}^k, \mathbf{r}^k)$.

Замечание. Найденные расчетные формулы можно переписать в виде

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau_{k+1}} + A\mathbf{x}^k = \mathbf{b}, \quad \tau_{k+1} = \frac{(\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, \mathbf{r}^k)}.$$

Соответствующий метод называется методом наискорейшего градиентного спуска.

Расчетные формулы данного метода также могут быть получены следующим образом. Будем минимизировать на каждом шаге A-норму вектора ошибки $\mathbf{z}^{k+1} = \mathbf{x} - \mathbf{x}^{k+1}$. Напомним, что в оптимальном линейном одношаговом методе минимизировалась норма оператора перехода $\|I - \tau A\|_2$.

Лемма 15.3 *Пусть $A = A^T > 0$. Тогда на k-м шаге метода наискорейшего градиентного спуска минимизируется норма $\|\mathbf{z}^k\|_A = \sqrt{(A\mathbf{z}^k, \mathbf{z}^k)}$ вектора ошибки $\mathbf{z}^k = \mathbf{x} - \mathbf{x}^k$, где $\mathbf{x}$ — точное решение.*

Действительно, так как $\mathbf{z}^{k+1} = (I - \tau_{k+1}A)\mathbf{z}^k$, то

$$\|\mathbf{z}^{k+1}\|_A^2 = (A(I - \tau_{k+1}A)\mathbf{z}^k, (I - \tau_{k+1}A)\mathbf{z}^k) =$$
$$= (A\mathbf{z}^k, \mathbf{z}^k) - 2\tau_{k+1}(A\mathbf{z}^k, A\mathbf{z}^k) + \tau_{k+1}^2(AA\mathbf{z}^k, A\mathbf{z}^k).$$

Отсюда, дифференцируя по τ_{k+1}, находим, что на $(k+1)$-м шаге минимум $\|\mathbf{z}^{k+1}\|_A^2$ достигается при $\tau_{k+1} = (A\mathbf{z}^k, A\mathbf{z}^k)/(AA\mathbf{z}^k, A\mathbf{z}^k)$. С учетом $A\mathbf{z}^k = \mathbf{b} - A\mathbf{x}^k = \mathbf{r}^k$ имеем $\tau_{k+1} = (\mathbf{r}^k, \mathbf{r}^k)/(A\mathbf{r}^k, \mathbf{r}^k)$.

Замечание. Минимизация евклидовой нормы $\|\mathbf{z}^{k+1}\|_2 = \sqrt{(\mathbf{z}^{k+1}, \mathbf{z}^{k+1})}$ вектора ошибки приводит к неконструктивным формулам для параметра $\tau_{k+1} = (\mathbf{r}^k, \mathbf{z}^k)/(\mathbf{r}^k, \mathbf{r}^k)$.

Теорема 15.1 *Пусть $A = A^T > 0$ и $\lambda(A) \in [m, M]$. Тогда для системы $A\mathbf{x} = \mathbf{b}$ метод наискорейшего градиентного спуска сходится с любого начального приближения со скоростью геометрической прогрессии, и для вектора ошибки имеет место следующая оценка:*

$$\|\mathbf{z}^k\|_A \le \left(\frac{M-m}{M+m}\right)^k \|\mathbf{z}^0\|_A, \quad \text{где } \mathbf{z}^k = \mathbf{x} - \mathbf{x}^k,\ \|\mathbf{z}\|_A^2 = (A\mathbf{z}, \mathbf{z}).$$

Доказательство. Параметр τ_{k+1} минимизирует на $(k+1)$-м шаге норму $\|\mathbf{z}^{k+1}\|_A$, следовательно, с параметром τ_0 оптимального линейного одношагового метода оценка не лучше:

$$\min_{\tau'_{k+1}} \|\mathbf{z}^{k+1}\|_A = \|(I - \tau_{k+1}A)\mathbf{z}^k\|_A \le \|(I - \tau_0 A)\mathbf{z}^k\|_A \le$$
$$\le \|I - \tau_0 A\|_A \|\mathbf{z}^k\|_A = \left(\tfrac{M-m}{M+m}\right)\|\mathbf{z}^k\|_A.$$

Здесь A-норма оператора перехода вычислена следующим образом:

$$\|I - \tau_0 A\|_A^2 = \sup_{x \ne 0} \tfrac{(A(I-\tau_0 A)x, (I-\tau_0 A)x)}{(Ax, x)} = \sup_{x \ne 0} \tfrac{((I-\tau_0 A)A^{1/2}x, (I-\tau_0 A)A^{1/2}x)}{(A^{1/2}x, A^{1/2}x)} =$$
$$= \sup_{y \ne 0} \tfrac{((I-\tau_0 A)y, (I-\tau_0 A)y)}{(y, y)} = \|I - \tau_0 A\|_2^2 = (\tfrac{M-m}{M+m})^2.$$

При этом мы применили следующее

Утверждение 15.1 *Для произвольной матрицы $A = A^T > 0$ найдется такая матрица $\tilde{A}$, что $\tilde{A} = \tilde{A}^T > 0$, $\tilde{A}\tilde{A} = A$. Матрицу $\tilde{A}$ называют квадратным корнем из A и обозначают $A^{1/2}$.*

Метод минимальных невязок. Пусть $A = A^T > 0$. Будем минимизировать на $(k+1)$-м шаге A^2-норму вектора ошибки $\mathbf{z}^{k+1} = \mathbf{x} - \mathbf{x}^{k+1}$.

Так как $\mathbf{z}^{k+1} = (I - \tau_{k+1}A)\mathbf{z}^k$, то

$$\|\mathbf{z}^{k+1}\|_{A^2}^2 = (A(I - \tau_{k+1}A)\mathbf{z}^k, A(I - \tau_{k+1}A)\mathbf{z}^k) =$$
$$= \|\mathbf{z}^k\|_{A^2}^2 - 2\tau_{k+1}(A^2\mathbf{z}^k, A\mathbf{z}^k) + \tau_{k+1}^2(A^2\mathbf{z}^k, A^2\mathbf{z}^k).$$

Отсюда, дифференцируя по τ_{k+1} с учетом $A\mathbf{z}^k = \mathbf{b} - A\mathbf{x}^k = \mathbf{r}^k$ находим, что минимум $\|\mathbf{z}^{k+1}\|_{A^2}^2$ достигается при $\tau_{k+1} = (A\mathbf{r}^k, \mathbf{r}^k)/(A\mathbf{r}^k, A\mathbf{r}^k)$. Итерационный процесс с таким набором параметров называется методом минимальных невязок, так как $\|\mathbf{z}^{k+1}\|_{A^2}^2 = (A\mathbf{z}^{k+1}, A\mathbf{z}^{k+1}) = \|\mathbf{r}^{k+1}\|_2^2$. Нами доказана

Теорема 15.2 *Пусть $A = A^T > 0$. Тогда на k-м шаге метода минимальных невязок минимизируется норма $\|\mathbf{z}^k\|_{A^2} = \sqrt{(A\mathbf{z}^k, A\mathbf{z}^k)}$ вектора ошибки $\mathbf{z}^k = \mathbf{x} - \mathbf{x}^k$.*

Имеет место следующая теорема об оценке скорости сходимости метода.

Теорема 15.3 *Пусть $A = A^T > 0$ и $\lambda(A) \in [m, M]$. Тогда для системы $A\mathbf{x} = \mathbf{b}$ метод минимальных невязок сходится с любого начального приближения со скоростью геометрической прогрессии, и для вектора ошибки имеет место следующая оценка:*

$$\|\mathbf{z}^k\|_{A^2} \le \left(\frac{M-m}{M+m}\right)^k \|\mathbf{z}^0\|_{A^2}, \quad \text{где } \mathbf{z}^k = \mathbf{x} - \mathbf{x}^k,\ \|\mathbf{z}\|_{A^2}^2 = (A\mathbf{z}, A\mathbf{z}).$$

Доказательство совпадает с доказательством теоремы о скорости сходимости наискорейшего градиентного спуска, т.к. $\|I-\tau_0 A\|_{A^2} = \|I-\tau_0 A\|_2$.

Скорость сходимости рассмотренных вариационных методов по порядку не хуже, чем у линейного одношагового метода, но в A- и A^2-нормах соответственно. При этом для практической реализации данных методов не требуется знание спектральных границ m, M матрицы A.

Метод спектрально-эквивалентных операторов. Скорость сходимости рассмотренных итерационных процессов зависела от отношения m/M границ спектра матрицы $A = A^T > 0$, то есть от свойств конкретной задачи. Для "улучшения исходной задачи" можно перейти к некоторой эквивалентной системе $B^{-1}A\mathbf{x} = B^{-1}\mathbf{b}$ при условии невырожденности B:

$$\frac{\mathbf{x}^{k+1}-\mathbf{x}^k}{\tau_{k+1}} + B^{-1}A\mathbf{x}^k = B^{-1}\mathbf{b}. \qquad (1)$$

Если явный вид матрицы B^{-1} не известен, то (1) удобно переписать в канонической (по Самарскому) форме:

$$B\frac{\mathbf{x}^{k+1}-\mathbf{x}^k}{\tau_{k+1}} + A\mathbf{x}^k = \mathbf{b}. \qquad (2)$$

Данный метод также называют *методом с предобусловливателем* B. Неявный двухслойный итерационный процесс (2) требует на каждом шаге решения задач вида $By = f$ и совпадает с рассмотренными методами при $B = E$.

Теорема 15.4 *Пусть $B = B^T > 0$, $A = A^T > 0$ и $m_1 \le \frac{(Ax,x)}{(Bx,x)} \le M_1$. Тогда для $\tau = \frac{2}{m_1+M_1}$ метод сходится со скоростью геометрической прогрессии с показателем $q = \frac{M_1-m_1}{M_1+m_1}$, и для вектора ошибки имеет место следующая оценка:*

$$\|B^{\frac12}\mathbf{z}^{k+1}\|_2 \le q\|B^{\frac12}\mathbf{z}^k\|_2.$$

Доказательство. Выпишем уравнение для вектора ошибки:

$$B\mathbf{z}^{k+1} = (B-\tau A)\mathbf{z}^k \Rightarrow B^{\frac12}\mathbf{z}^{k+1} = B^{\frac12}\mathbf{z}^k - \tau B^{-\frac12}AB^{-\frac12}B^{\frac12}\mathbf{z}^k \Rightarrow$$

$$\mathbf{y}^{k+1} = (I-\tau B^{-\frac12}AB^{-\frac12})\mathbf{y}^k, \quad \text{где } \mathbf{y} = B^{\frac12}\mathbf{z}.$$

Таким образом, для вектора $\mathbf{y}^k$ данный метод соответствует линейному одношаговому методу с матрицей $C = B^{-\frac12}AB^{-\frac12} = C^T$. Оценим границы спектра C. Имеем

$$\frac{(C\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})} = \frac{(B^{-\frac12}AB^{-\frac12}\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})} = \frac{(A\tilde{\mathbf{x}},\tilde{\mathbf{x}})}{(B\tilde{\mathbf{x}},\tilde{\mathbf{x}})},$$

где $\tilde{\mathbf{x}} = B^{-\frac12}\mathbf{x}$. Отсюда и из условия теоремы имеем оценку $m_1 \le \frac{(C\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})} \le M_1$, следовательно, $\lambda(C) \in [m_1, M_1]$. Поэтому можно применить теорему о скорости сходимости оптимального одношагового метода и получить требуемую оценку для нормы вектора

$$\|\mathbf{y}^k\|_2^2 = (B^{\frac12}\mathbf{z}^k, B^{\frac12}\mathbf{z}^k) = (B\mathbf{z}^k, \mathbf{z}^k).$$

Теорема доказана.

Аналогично строятся неявные методы типа минимальных невязок и скорейшего градиентного спуска.

При удачном выборе оператора B можно принципиально улучшить скорость сходимости соответствующих итерационных процессов, однако необходимо учитывать трудоемкость нахождения $y = B^{-1}f$. Например, при $B = A$, $\tau = 1$ метод (2) сойдется за одну итерацию, но потребует решения исходной задачи $A\mathbf{x} = \mathbf{b}$.

Докажем общий результат о сходимости методов с предобусловливателем.

Теорема 15.5 *Пусть*

$$A = A^T > 0, \quad \tau > 0, \quad B - \frac{\tau}{2}A > 0.$$

Тогда обобщенный метод простой итерации

$$B\frac{\mathbf{x}^{k+1}-\mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}$$

сходится для произвольного $\mathbf{x}^0$.

Доказательство. Отметим, что $\det(B) \neq 0$. Докажем сходимость метода. Перепишем уравнение для вектора ошибки

$$B\frac{\mathbf{z}^{k+1}-\mathbf{z}^k}{\tau} + A\mathbf{z}^k = 0$$

в виде

$$(B-\frac{\tau}{2}A)(\mathbf{z}^{k+1}-\mathbf{z}^k) + \frac{\tau}{2}A(\mathbf{z}^{k+1}+\mathbf{z}^k) = 0,$$

умножим скалярно на $\frac{2}{\tau}(\mathbf{z}^{k+1}-\mathbf{z}^k)$ и, используя симметрию A, получим энергетическое тождество:

$$\|\mathbf{z}^{k+1}\|_A^2 - \|\mathbf{z}^k\|_A^2 + \frac{2}{\tau}((B-\frac{\tau}{2}A)(\mathbf{z}^{k+1}-\mathbf{z}^k), (\mathbf{z}^{k+1}-\mathbf{z}^k)) = 0 \quad \forall k \ge 0.$$

Если $\mathbf{z}^{k+1} = \mathbf{z}^k$, то из уравнения для вектора ошибки имеем $\mathbf{z}^k = 0$, т.е. решение найдено. Если $\mathbf{z}^{k+1} \neq \mathbf{z}^k$, то из энергетического тождества и условия $B - \frac{\tau}{2}A > 0$ следует монотонное ограниченное убывание $0 \le \|\mathbf{z}^{k+1}\|_A < \|\mathbf{z}^k\|_A$ и существование предельного значения для $\|\mathbf{z}^k\|_A$ при $k \to \infty$. Покажем, что $\|\mathbf{z}^k\|_A \to 0$ при $k \to \infty$. Перейдем к пределу в энергетическом тождестве. Получим

$$\Big((B - \frac{\tau}{2}A)(\mathbf{z}^{k+1} - \mathbf{z}^k), (\mathbf{z}^{k+1} - \mathbf{z}^k)\Big) \to 0.$$

Отсюда и из условия $B - \frac{\tau}{2}A > 0$ следует, что $\mathbf{z}^{k+1} - \mathbf{z}^k$ сходится к нулевому вектору, т.е. $\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_A \to 0$. Далее $\mathbf{z}^k = -\frac{1}{\tau}A^{-1}B(\mathbf{z}^{k+1} - \mathbf{z}^k)$, поэтому $\|\mathbf{z}^k\|_A \le \frac{1}{\tau}\|A^{-1}\|_A\|B\|_A\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_A \to 0$. Таким образом $\|\mathbf{z}^k\|_A \to 0$, и метод сходится. Теорема доказана.

Метод Зейделя. Представим матрицу системы $A\,\mathbf{x} = \mathbf{b}$ в виде $A = L + D + R$, где D — диагональная матрица, L и R — соответственно левая нижняя и правая верхняя треугольные матрицы с нулевыми диагоналями (строго нижняя и строго верхняя треугольные матрицы). Итерационный метод (2) при $B = D + L$, $\tau = 1$ называется методом Зейделя (Гаусса–Зейделя)

$$(D + L)(\mathbf{x}^{k+1} - \mathbf{x}^k) + A\mathbf{x}^k = \mathbf{b}, \quad \text{т.е.}$$
$$(D + L)\mathbf{x}^{k+1} + R\mathbf{x}^k = \mathbf{b}.$$

Теорема 15.6 *Пусть $A = A^T > 0$. Тогда метод Гаусса–Зейделя сходится с любого начального приближения.*

Доказательство. Проверим, что $B - \frac{\tau}{2}A > 0$. С учетом $L^T = R$ имеем:

$$((D + L - \frac{1}{2}A)\mathbf{x}, \mathbf{x}) = \frac{1}{2}(D\mathbf{x}, \mathbf{x}) + \frac{1}{2}(L\mathbf{x}, \mathbf{x}) - \frac{1}{2}(R\mathbf{x}, \mathbf{x}) = \frac{1}{2}(D\mathbf{x}, \mathbf{x}).$$

Так как $(A\mathbf{x}, \mathbf{x}) > 0$, то $(A\mathbf{e}_i, \mathbf{e}_i) = a_{ii} > 0$ и $(D\mathbf{x}, \mathbf{x}) > 0$.

Теорема 15.7 *Пусть матрица A обладает свойством строгого диагонального преобладания, т.е. справедливо*

$$\sum_{i \neq j} |a_{ij}| \le q|a_{ii}|,\ i = 1, \ldots, n, \quad q < 1\,.$$

Тогда метод Гаусса–Зейделя сходится с любого начального приближения, и для вектора ошибки верна оценка:

$$\|\mathbf{x} - \mathbf{x}^k\|_\infty \le q^k\,\|\mathbf{x} - \mathbf{x}^0\|_\infty\,.$$

Доказательство. Отметим, что условие строгого диагонального преобладания влечет (см. теорему Гершгорина) невырожденность матрицы A. Запишем уравнение для вектора ошибки $(D + L)\,\mathbf{z}^{k+1} + R\,\mathbf{z}^k = 0$. Выпишем i-е уравнение

$$\sum_{j=1}^{i-1} a_{ij}\,z_j^{k+1} + a_{ii}\,z_i^{k+1} + \sum_{j=i+1}^{n} a_{ij}\,z_j^k = 0$$

и разрешим его относительно z_i^{k+1} :

$$z_i^{k+1} = -\sum_{j=1}^{i-1} \tfrac{a_{ij}}{a_{ii}}\,z_j^{k+1} - \sum_{j=i+1}^{n} \tfrac{a_{ij}}{a_{ii}}\,z_j^k\,.$$

Отсюда получим

$$|z_i^{k+1}| \le \sum_{j=1}^{i-1} |\tfrac{a_{ij}}{a_{ii}}|\,|z_j^{k+1}| + \sum_{j=i+1}^{n} |\tfrac{a_{ij}}{a_{ii}}|\,|z_j^k| \le$$
$$\le \|z^{k+1}\|_\infty \sum_{j=1}^{i-1} |\tfrac{a_{ij}}{a_{ii}}| + \|z^k\|_\infty \sum_{j=i+1}^{n} |\tfrac{a_{ij}}{a_{ii}}|.$$

Так как соотношение верно для всех i, следовательно, оценка верна и для нормы $\|\mathbf{z}^{k+1}\|_\infty$. В результате имеем

$$(1 - \sum_{j<i} |\tfrac{a_{ij}}{a_{ii}}|)\|\mathbf{z}^{k+1}\|_\infty \le \|\mathbf{z}^k\|_\infty \sum_{j>i} |\tfrac{a_{ij}}{a_{ii}}|.$$

По условию теоремы

$$\sum_{j>i} |\tfrac{a_{ij}}{a_{ii}}| \le q - \sum_{j<i} |\tfrac{a_{ij}}{a_{ii}}| \le q(1 - \sum_{j<i} |\tfrac{a_{ij}}{a_{ii}}|),$$

следовательно, $\|\mathbf{z}^{k+1}\|_\infty \le q\|\mathbf{z}^k\|_\infty$ и $\|\mathbf{z}^k\|_\infty \le q\|\mathbf{z}^{k-1}\|_\infty \le \ldots \le q^k\|\mathbf{z}^0\|_\infty$. Теорема доказана.

Метод Якоби. Пусть $A = D + L + R$. Рассмотрим итерационный процесс (2) с матрицей $B = D$ и $\tau = 1$:

$$D(\mathbf{x}^{k+1} - \mathbf{x}^k) + A\mathbf{x}^k = \mathbf{b}, \quad \text{т.е.}$$
$$D\mathbf{x}^{k+1} + (L + R)\mathbf{x}^k = \mathbf{b}.$$

Теорема 15.8 *Пусть матрица A обладает свойством диагонального преобладания с коэффициентом q. Тогда метод Якоби сходится с любого начального приближения, и для вектора ошибки верна оценка*

$$\|\mathbf{x} - \mathbf{x}^k\|_\infty \le q^k\,\|\mathbf{x} - \mathbf{x}^0\|_\infty\,.$$

Доказательство. Из уравнения $D\,\mathbf{z}^{k+1} + (R+L)\,\mathbf{z}^k = 0$ для вектора ошибки имеем $\mathbf{z}^{k+1} = S\mathbf{z}^k$, где

$$S = -\begin{pmatrix} 0 & a_{12}/a_{11} & a_{13}/a_{11} & \dots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & a_{23}/a_{22} & \dots & a_{2n}/a_{22} \\ . & . & . & . & . \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & \dots & a_{nn-1}/a_{nn} & 0 \end{pmatrix}.$$

Отсюда и условия теоремы находим оценку для нормы оператора перехода: $\|S\|_\infty \le q < 1$.

Метод верхней релаксации. Методом верхней релаксации называется итерационный процесс (2) с матрицей $B = (D + \omega L)$:

$$(D + \omega\, L)\,\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}.$$

Здесь итерационный параметр ω называется *параметром релаксации* (обычно $\omega = \tau$). Методы Якоби ($\omega = 0, \tau = 1$) и Гаусса–Зейделя ($\omega = \tau = 1$) также являются методами релаксации.

Теорема 15.9 *Пусть $A = A^T > 0$. Тогда метод верхней релаксации при $0 < \omega = \tau < 2$ сходится с любого начального приближения.*

Доказательство следует из оценки $B - \frac{\tau}{2}A > 0$.

Лекция 16. Проекционные методы решения слау

Эффективными методами решения системы линейных алгебраических уравнений $A\mathbf{x}^* = \mathbf{b}$ большой размерности являются итерационные методы проекционного типа. На каждом шаге такого метода реализуется *проекционный алгоритм*:

в зависимости от текущего приближения $\mathbf{x} \in \mathbf{R}^n$ и номера итерации выбирают два m-мерных ($m \le n$) подпространства $\mathcal{K}$ и $\mathcal{L}$; следующее приближение $\hat{\mathbf{x}}$ к точному решению $\mathbf{x}^*$ ищут в виде $\hat{\mathbf{x}} = \mathbf{x} + \mathbf{k}$, $\mathbf{k} \in \mathcal{K}$ из условия $\mathbf{r} \perp \mathcal{L}$, где $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$.

Таким образом, основная идея данного подхода заключается в построении вектора поправки $\mathbf{k}$ из подпространства $\mathcal{K}$, обеспечивающего ортогональность вектора невязки $\mathbf{r}$ подпространству $\mathcal{L}$. Различные правила выбора подпространств $\mathcal{K}$ и $\mathcal{L}$ приводят к различным расчетным формулам.

Пример 16.1. Покажем, что метод Гаусса–Зейделя решения систем линейных уравнений является проекционным методом.

Определим $\mathcal{K} = \mathcal{L} = \{\mathbf{e}_i\}$ для $i = 1, ..., n$, где $\mathbf{e}_i$ — естественный i-й базисный вектор пространства $\mathbf{R}^n$, и выполним n шагов. Последовательно имеем $\hat{\mathbf{x}} = \mathbf{x} + c_i\mathbf{e}_i$, $(\mathbf{b} - A(\mathbf{x} + c_i\mathbf{e}_i), \mathbf{e}_i) = 0$, т.е. $b_i - \sum\limits_{j \ne i} a_{ij}x_j - \hat{x}_i a_{ii} = 0$. Отсюда находим $\hat{x}_i$ при известных компонентах x_j, $j = i, i+1, \dots, n$ и найденных x_j, $j = 1, 2, \dots, i-1$. Таким образом, за n шагов проекционного алгоритма имеем $\mathbf{x}^{k+1} = \mathbf{x}^k + \sum\limits_{i=1}^{n} c_i\mathbf{e}_i$, что соответствует шагу метода Гаусса–Зейделя: $(D + L)\mathbf{x}^{k+1} + R\mathbf{x}^k = \mathbf{b}$.

Экстремальное свойство проекционного алгоритма. Нам потребуется следующая проекционная теорема.

Теорема 16.1 *Для произвольного вектора $\mathbf{z}$ вектор $\hat{\mathbf{k}}$ является решением следующей задачи минимизации*

$$\min_{\mathbf{k} \in \mathcal{K}} (\mathbf{z} - \mathbf{k}, \mathbf{z} - \mathbf{k})$$

тогда и только тогда, когда $(\mathbf{z} - \hat{\mathbf{k}}, \mathbf{v}) = 0$ для $\forall \mathbf{v} \in \mathcal{K}$.

Доказательство. Рассмотрим разложение $\mathbf{z} = P\mathbf{z} + (\mathbf{z} - P\mathbf{z})$, где $P\mathbf{z} \in \mathcal{K}$, $\mathbf{z} - P\mathbf{z} \in \mathcal{K}^\perp$ (в этом случае P называют *оператором ортогонального проектирования на $\mathcal{K}$*). Тогда

$$(\mathbf{z} - \mathbf{k}, \mathbf{z} - \mathbf{k}) = \|\mathbf{z} - \mathbf{k}\|^2 = \|\mathbf{z} - P\mathbf{z} + P\mathbf{z} - \mathbf{k}\|^2 = \|\mathbf{z} - P\mathbf{z}\|^2 + \|P\mathbf{z} - \mathbf{k}\|^2,$$

т. е. $\|\mathbf{z} - \mathbf{k}\|^2 \ge \|\mathbf{z} - P\mathbf{z}\|^2$, равенство возможно лишь при $\mathbf{k} = P\mathbf{z}$.

Теорема 16.2 *Пусть $A = A^T > 0$ и $\mathcal{L} = \mathcal{K}$. Тогда вектор $\hat{\mathbf{x}}$ является результатом проекционного алгоритма тогда и только тогда, когда*

$$E(\hat{\mathbf{x}}) = \min_{\tilde{\mathbf{x}} \in \mathbf{x} + \mathcal{K}} E(\tilde{\mathbf{x}}), \quad \text{где } E(\tilde{\mathbf{x}}) = (A(\mathbf{x}^* - \tilde{\mathbf{x}}), \mathbf{x}^* - \tilde{\mathbf{x}})\ ,\ A\mathbf{x}^* = \mathbf{b}.$$

Доказательство. Из проекционной теоремы следует, что решение задачи $\min\limits_{\mathbf{k} \in \mathcal{K}} (\mathbf{z} - \mathbf{k}, \mathbf{z} - \mathbf{k})_A$ для $\mathbf{z} = \mathbf{x}^* - \mathbf{x}$, где $(\mathbf{u}, \mathbf{v})_A = (A\mathbf{u}, \mathbf{v})$ эквивалентно нахождению вектора $\mathbf{k}$ из условия $(\mathbf{z} - \mathbf{k}, \mathbf{v})_A = 0 \ \ \forall \mathbf{v} \in \mathcal{K}$, что соответствует определению проекционного алгоритма:

$$(\mathbf{z} - \mathbf{k}, \mathbf{v})_A = (A(\mathbf{x}^* - (\mathbf{x} + \mathbf{k})), \mathbf{v}) = (\mathbf{b} - A\hat{\mathbf{x}}, \mathbf{v}) = 0 \quad \forall \mathbf{v} \in \mathcal{K}.$$

Такой подход к аппроксимации вектора $\mathbf{x}^*$ вектором $\hat{\mathbf{x}}$ называется *методом Галеркина*: $(\mathbf{b} - A\hat{\mathbf{x}}, \mathbf{v}) = 0 \ \ \forall \mathbf{v} \in \mathcal{K}$.

Теорема 16.3 *Пусть A невырождена и $\mathcal{L} = A\mathcal{K}$. Тогда вектор $\hat{\mathbf{x}}$ является результатом проекционного алгоритма тогда и только тогда, когда*

$$E(\hat{\mathbf{x}}) = \min_{\tilde{\mathbf{x}} \in \mathbf{x}+\mathcal{K}} E(\tilde{\mathbf{x}}), \quad \text{где } E(\tilde{\mathbf{x}}) = (A(\mathbf{x}^* - \tilde{\mathbf{x}}), A(\mathbf{x}^* - \tilde{\mathbf{x}}))\ ,\ A\mathbf{x}^* = \mathbf{b}.$$

Доказательство совпадает с доказательством предыдущей теоремы с точностью до замены скалярного произведения на $(\mathbf{u}, \mathbf{v})_{A^TA} = (A^TA\mathbf{u}, \mathbf{v}) = (A\mathbf{u}, A\mathbf{v})$. При этом по условию $\mathbf{v} \in \mathcal{K}$, $A\mathbf{v} \in \mathcal{L}$. Такой подход к аппроксимации вектора $\mathbf{x}^*$ вектором $\hat{\mathbf{x}}$ называется *методом Петрова–Галеркина*: $(\mathbf{b} - A\hat{\mathbf{x}}, \mathbf{v}) = 0 \ \ \forall \mathbf{v} \in A\mathcal{K}$.

Одномерные проекционные методы. В простейшем случае в качестве базовых пространств $\mathcal{K}$ и $\mathcal{L}$ выбирают одномерные подпространства.

Лемма 16.1 *Проекционный алгоритм при $\mathcal{K} = \mathcal{L} = \{\mathbf{r}\}$, где $\mathbf{r} = \mathbf{b} - A\mathbf{x}$, соответствует методу наискорейшего градиентного спуска.*

Доказательство. Пространства $\mathcal{K}$ и $\mathcal{L}$ одномерны, следовательно, $\hat{\mathbf{x}} = \mathbf{x} + \tau\mathbf{r}$, и τ определяется из условия ортогональности $(\mathbf{b} - A(\mathbf{x} + \tau\mathbf{r}), \mathbf{r}) = 0$. Отсюда имеем $(\mathbf{r} - \tau A\mathbf{r}, \mathbf{r}) = 0$ и $\tau = (\mathbf{r}, \mathbf{r})/(A\mathbf{r}, \mathbf{r})$.

Лемма 16.2 *Проекционный алгоритм при $\mathcal{K} = \{\mathbf{r}\}$ и $\mathcal{L} = \{A\mathbf{r}\}$, где $\mathbf{r} = \mathbf{b} - A\mathbf{x}$, соответствует методу минимальных невязок.*

Доказательство. В терминах предыдущей леммы находим, что $\tau = (A\mathbf{r}, \mathbf{r})/(A\mathbf{r}, A\mathbf{r})$.

Проекционные методы в пространствах Крылова. Пусть пространства $\mathcal{L}$ зависят от номера итерации и $\mathcal{L}^1 \subset \mathcal{L}^2 \subset \ldots \subset \mathcal{L}^m \subset \ldots \subset \mathcal{L}^n = \mathbf{R}^n$. Тогда точное решение системы будет получено не позже, чем за n шагов. Если же цепочка $\mathcal{L}^m$ задается некоторым оптимальным образом, то можно расчитывать, что требуемая точность $\|\mathbf{x}^* - \mathbf{x}^m\| \le \varepsilon$, где $A\mathbf{x}^* = \mathbf{b}$, будет достигнута значительно раньше.

Эффективные алгоритмы удается построить, если в качестве пространства $\mathcal{K}^m$ выбрать пространство Крылова $\mathcal{K}^m = \mathrm{span}\{\mathbf{r}, A\mathbf{r}, ..., A^{m-1}\mathbf{r}\}$ порядка m для $\mathbf{r} = \mathbf{b} - A\mathbf{x}^0$. При этом пространство $\mathcal{L}^m$ определяется либо как $\mathcal{L}^m = \mathcal{K}^m$, либо $\mathcal{L}^m = A\mathcal{K}^m$.

Метод сопряженных градиентов. Пусть $A = A^T > 0$. Построим проекционный метод для пары пространств

$$\mathcal{K}^m = \mathrm{span}\left\{\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{m-1}\mathbf{r}^0\right\}, \quad \mathcal{L}^m = \mathcal{K}^m, \quad \mathbf{r}^0 = \mathbf{b} - A\mathbf{x}^0,$$

очередное приближение найдем в виде $\mathbf{x}^m = \mathbf{x}^0 + \sum\limits_{i=1}^{m} c_i A^{i-1}\mathbf{r}^0$, а коэффициенты c_i определим из условия $\mathbf{r}^m = (\mathbf{b} - A\mathbf{x}^m) \perp \mathcal{L}^m$. Такая форма алгоритма для нахождения $\{c_i\}$ требует решения системы линейных уравнений

(при этом формально $\{c_i = c_i^{(m)}\}$, т.е. коэффициенты могут изменяться при увеличении m). Рассмотрим эквивалентную, но более удобную с практической точки зрения реализацию этого алгоритма.

Пусть в пространстве $\mathcal{K}^m = \mathrm{span}\{\mathbf{k}_1, \ldots, \mathbf{k}_m\}$ известен A-ортогональный базис, т.е. $(A\mathbf{k}_i, \mathbf{k}_j) = 0$ при $i \ne j$ и $\mathbf{k}_1 = \mathbf{r}^0$. Тогда $\mathbf{x}^m = \mathbf{x}^0 + \sum\limits_{i=1}^{m} \alpha_i \mathbf{k}_i$ и $\mathbf{r}^m = \mathbf{b} - A\mathbf{x}^m = \mathbf{b} - A\left(\mathbf{x}^0 + \sum\limits_{i=1}^{m} \alpha_i \mathbf{k}_i\right)$. В этом случае из условия $\mathbf{r}^m \perp \mathcal{L}^m$ имеем формулы для определения коэффициентов

$$(\mathbf{r}^m, \mathbf{k}_j) = (\mathbf{r}^0, \mathbf{k}_j) - \alpha_j (A\mathbf{k}_j, \mathbf{k}_j) = 0,\ \alpha_j = \frac{(\mathbf{r}^0, \mathbf{k}_j)}{(A\mathbf{k}_j, \mathbf{k}_j)},\ j = 1, \ldots, m.$$

Таким образом, коэффициенты $\alpha_1, \alpha_2, \ldots$ не зависят от выбора m, поэтому могут быть найдены последовательно из условия $\mathbf{x}^m = \mathbf{x}^{m-1} + \alpha_m \mathbf{k}_m$. Отсюда получаем $\mathbf{r}^m = \mathbf{r}^{m-1} - \alpha_m A\mathbf{k}_m$ и $\alpha_m = (\mathbf{r}^{m-1}, \mathbf{k}_m)/(A\mathbf{k}_m, \mathbf{k}_m)$. Для вычислений такая рекуррентная форма записи предпочтительнее.

Построим соответствующий рекуррентный алгоритм для определения $\{\mathbf{k}_i\}$, т.к. стандартная процедура типа Грама–Шмидта, требующая хранения всех элементов базиса $\{\mathbf{k}_i\}_{i=1}^m$, в данном случае оказывается существенно менее эффективна. Имеем $\mathcal{K}^{m+1} = \mathcal{L}^{m+1} =$

$$= \mathrm{span}\Big\{\mathrm{span}\left\{\mathbf{r}^0, \ldots, A^{m-1}\mathbf{r}^0\right\}, A^m\mathbf{r}^0\Big\} = \mathrm{span}\Big\{\mathrm{span}\left\{\mathbf{k}_1, \ldots, \mathbf{k}_m\right\}, \mathbf{k}_{m+1}\Big\}.$$

Отсюда следует, что $\mathbf{k}_{m+1} = A^m\mathbf{r}^0 + \sum\limits_{i=1}^{m} \tilde{\beta}_i \mathbf{k}_i$. Заметим, что

$$\mathbf{r}^m = \mathbf{r}^0 - A\sum_{i=1}^{m} c_i A^{i-1}\mathbf{r}^0 = \mathbf{r}^0 - \sum_{i=1}^{m} c_i A^i \mathbf{r}^0,$$

следовательно, при $\mathbf{r}^m \ne 0$ и $c_m \ne 0$ вектор $\mathbf{k}_{m+1}$ можно искать в виде $\mathbf{k}_{m+1} = \mathbf{r}^m + \sum\limits_{i=1}^{m} \beta_i \mathbf{k}_i$. Умножим данное соотношение на $A\mathbf{k}_j$. С учетом $(\mathbf{k}_i, A\mathbf{k}_j) = 0$ при $i \ne j$, получим $0 = (\mathbf{r}^m, A\mathbf{k}_j) + \beta_j(\mathbf{k}_j, A\mathbf{k}_j)$. Покажем, что $(\mathbf{r}^m, A\mathbf{k}_i) = 0$ для $i < m$. По построению пространства $\mathcal{L}^m$ имеем

$$\mathrm{span}\left\{\mathbf{k}_1, \ldots, \mathbf{k}_{m-1}\right\} = \mathrm{span}\left\{\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{m-2}\mathbf{r}^0\right\},$$
$$A\left(\mathrm{span}\left\{\mathbf{k}_1, \ldots, \mathbf{k}_{m-1}\right\}\right) = A\left(\mathrm{span}\left\{\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{m-2}\mathbf{r}^0\right\}\right) =$$
$$= \mathrm{span}\left\{A\mathbf{r}^0, A^2\mathbf{r}^0, \ldots, A^{m-1}\mathbf{r}^0\right\} \subset \mathcal{L}^m.$$

Следовательно, $A\mathbf{k}_i \in \mathcal{L}^m$ при $i = 1, \ldots, m-1$. Отсюда и из условия $\mathbf{r}^m \perp \mathcal{L}^m$ имеем $(\mathbf{r}^m, A\mathbf{k}_i) = 0$ для $i < m$, т.е. $\beta_i = 0$ при $i < m$ и $\mathbf{k}_{m+1} = \mathbf{r}^m + \beta_m \mathbf{k}_m$. Из данного представления находим $\beta_m = -\dfrac{(\mathbf{r}^m, A\mathbf{k}_m)}{(A\mathbf{k}_m, \mathbf{k}_m)}$.

Таким образом расчетные формулы имеют вид:

$$\mathbf{x}^m = \mathbf{x}^{m-1} + \alpha_m \mathbf{k}_m, \quad \alpha_m = \frac{(\mathbf{r}^{m-1}, \mathbf{k}_m)}{(A\mathbf{k}_m, \mathbf{k}_m)},$$

$$\mathbf{k}_{m+1} = \mathbf{r}^m + \beta_m \mathbf{k}_m, \quad \beta_m = -\frac{(\mathbf{r}^m, A\mathbf{k}_m)}{(A\mathbf{k}_m, \mathbf{k}_m)}, \ \mathbf{k}_1 = \mathbf{r}^0.$$

Замечание. На шаге m данного метода минимизируется A-норма вектора ошибки на подпространствах Крылова $\mathcal{K}^m$, поэтому с точки зрения проекционных методов метод сопряженных градиентов является обобщением метода наискорейшего градиентного спуска. Метод сопряженных градиентов минимизирует значение функционала $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$ на векторах вида $\mathbf{x}^m = \mathbf{x}^0 + \sum\limits_{i=1}^{m} c_i A^{i-1}\mathbf{r}^0$ относительно c_i.

Теорема 16.4 *Пусть $A = A^T > 0$. Тогда метод сопряженных градиентов сходится с любого начального приближения, и имеет место следующая оценка скорости сходимости*

$$\|\mathbf{z}^N\|_A \le \frac{2}{q_1^{-N} + q_1^N}\|\mathbf{z}^0\|_A, \quad q_1 = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}},$$

где $\mathbf{z}^N = \mathbf{x}^* - \mathbf{x}^N$, $A\mathbf{x}^* = \mathbf{b}$ и $\lambda(A) \in [m, M]$.

Доказательство. Сравним задачи минимизации ошибки, соответствующие методу сопряженных градиентов и оптимальному линейному N-шаговому методу, принимая во внимание, что оптимальный N-шаговый процесс представляет собой метод простой итерации с чебышёвским набором параметров. Для нахождения приближения $\mathbf{x}_{ch}^k$ имеем

$$\frac{\mathbf{x}_{ch}^k - \mathbf{x}_{ch}^{k-1}}{\tau_k} + A\mathbf{x}_{ch}^{k-1} = \mathbf{b}, \quad k = 1, \ldots, N.$$

Отсюда следует

$$\mathbf{x}_{ch}^k = \mathbf{x}_{ch}^{k-1} + \tau_k \mathbf{r}_{ch}^{k-1}, \quad \mathbf{r}_{ch}^k = \mathbf{r}_{ch}^{k-1} - \tau_k A\mathbf{r}_{ch}^{k-1}$$

при $k = 1, 2, \ldots, N$. По индукции получаем, что $\mathbf{x}_{ch}^N = \mathbf{x}^0 + \sum_{i=1}^N \hat{c}_i A^{i-1}\mathbf{r}^0$. Приближение $\mathbf{x}_{sg}^N$ метода сопряженных градиентов по определению имеет вид $\mathbf{x}_{sg}^N = \mathbf{x}^0 + \sum_{i=1}^N c_i A^{i-1}\mathbf{r}^0$. Отсюда следует, что вектора $\mathbf{x}_{ch}^N$ и $\mathbf{x}_{sg}^N$ принадлежат одному и тому же подпространству Крылова и могут отличаться только коэффициентами. Так как вектор $\mathbf{x}_{sg}^N$ является решением задачи минимизации $\min\limits_{x^N \in x^0 + \mathcal{K}} \big(A\mathbf{z}^N, \mathbf{z}^N\big)$, где $\mathbf{z}^N = x^* - \mathbf{x}^N$, то справедливо неравенство

$$\big(A\mathbf{z}_{sg}^N, \mathbf{z}_{sg}^N\big) = \min_{\mathbf{x}^N \in x^0 + \mathcal{K}} \big(A\mathbf{z}^N, \mathbf{z}^N\big) \le \big(A\mathbf{z}_{ch}^N, \mathbf{z}_{ch}^N\big).$$

Для $\mathbf{z}_{ch}^N$ имеем $\mathbf{z}_{ch}^N = \prod_{i=1}^N (1 - \tau_i A)\mathbf{z}^0$ и следующую оценку:

$$\|\mathbf{z}_{ch}^N\|_A \le \|\prod_{i=1}^N (1 - \tau_i A)\|_A \|\mathbf{z}^0\|_A = \|\prod_{i=1}^N (1 - \tau_i A)\|_2 \|\mathbf{z}^0\|_A \le$$

$$\le \frac{2}{q_1^{-N} + q_1^N}\|\mathbf{z}^0\|_A \le 2q_1^N \|\mathbf{z}^0\|_A, \quad q_1 = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}.$$

Теорема доказана.

Утверждение 16.1 *В методе сопряженных градиентов необходимыми и достаточными условиями минимума функционала $F(\mathbf{x}^m) = (A\mathbf{x}^m, \mathbf{x}^m) - 2(\mathbf{b}, \mathbf{x}^m)$ для любого $m \ge 1$ являются равенства $(\mathbf{r}^m, \mathbf{r}^j) = 0$ при $j = 0, 1, \ldots, m-1$.*

Утверждение 16.2 *В методе сопряженных градиентов для любого $m \ge 2$ имеют место соотношения ортогональности $(A\mathbf{r}^m, \mathbf{r}^j) = 0$ при $j = 0, 1, \ldots, m-2$.*

Лекция 17. Задачи на собственные значения

Напомним, что ненулевой вектор $\mathbf{e}$, удовлетворяющий условию $A\mathbf{e} = \lambda\mathbf{e}$ называется *собственным вектором*, при этом λ называется *собственным числом*. Многочлен $P(\lambda) = \det(A - \lambda I)$ называется *характеристическим многочленом* матрицы A, его корни суть собственные числа. Матрица A называется матрицей *простой структуры*, если ее собственные вектора $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ образуют базис в $\mathbf{C}^n$.

Пусть S — произвольная невырожденная матрица. Говорят, что матрицы одинаковой размерности A и $B = S^{-1}AS$ *подобны*, а матрица S осуществляет подобие.

Утверждение 17.1 *Для произвольной вещественной матрицы $A \in \mathbf{R}^{n\times n}$ найдется вещественная ортогональная матрица $Q \in \mathbf{R}^{n\times n}$ такая, что*

$$Q^T A Q = R = \begin{pmatrix} R_{11} & R_{12} & \ldots & R_{1m} \\ 0 & R_{22} & \ldots & R_{2m} \\ . & . & \ldots & . \\ 0 & 0 & \ldots & R_{mm} \end{pmatrix},$$

где каждый диагональный блок R_{ii} $(m \le n)$ представляет собой либо вещественное собственное значение, либо 2×2-матрицу, отвечающую сопряженной паре комплексных собственных значений.

Матрицу R называют *действительной формой Шура*, из которой можно легко определить собственные векторы и собственные числа исходной матрицы.

Лемма 17.1 *Пусть $B = S^{-1}AS$, т.е. матрицы подобны. Тогда $\lambda_B = \lambda_A$, $\mathbf{e}_B = S^{-1}\mathbf{e}_A$.*

Доказательство. Собственные значения B находим из условия

$$0 = \det(B - \lambda I) = \det(S^{-1}AS - \lambda S^{-1}S) = \det(S^{-1}(A - \lambda I)S) =$$
$$= \det(S^{-1})\det(A - \lambda I)\det(S) = \det(A - \lambda I).$$

Собственные значения матриц A и B совпадают, так как равны характеристические многочлены. Из условия $A\mathbf{e} = \lambda\,\mathbf{e}$ имеем

$$S^{-1}ASS^{-1}\mathbf{e} = \lambda S^{-1}\mathbf{e}, \quad BS^{-1}\mathbf{e} = \lambda S^{-1}\mathbf{e},$$

т.е. собственные векторы $\mathbf{e}_A$ и $\mathbf{e}_B$ связаны соотношением $\mathbf{e}_B = S^{-1}\mathbf{e}_A$.

Лемма 17.2 *Пусть A — симметричная матрица размерности $n \times n$, $\lambda \in \mathbf{R}^1$, $\mathbf{x} \in \mathbf{R}^n$ — соответственно произвольное число и вектор, причем $||\mathbf{x}||_2 = 1$. Тогда существует собственное значение λ_k матрицы A, для которого $|\lambda_k - \lambda| \le ||A\mathbf{x} - \lambda\mathbf{x}||_2$.*

Доказательство. Пусть $\{\mathbf{e}_k\}$, $k = 1, \ldots, n$ — полная ортонормированная система собственных векторов матрицы A, $\mathbf{x} = \sum\limits_{k=1}^{n} c_k\mathbf{e}_k$. Тогда

$$||A\mathbf{x} - \lambda\mathbf{x}||_2^2 = \sum_{k=1}^{n}(\lambda_k - \lambda)^2 c_k^2 \ge \min_k(\lambda_k - \lambda)^2\|\mathbf{x}\|_2^2.$$

Так как малые изменения элементов матрицы могут приводить к существенным изменениям собственных значений и определителя (см. пример для матрицы Уилкинсона), то задача нахождения собственных значений относится к числу неустойчивых, а алгоритмы ее решения, основанные на приближенном построения характеристического многочлена и вычислении его корней, на практике обычно не применяются.

Степенной метод. Алгоритм вычисления максимального по модулю собственного значения λ матрицы A имеет вид

$$\mathbf{x}^{k+1} = A\mathbf{x}^k, \;\; \lambda^k = \frac{(\mathbf{x}^{k+1}, \mathbf{x}^k)}{(\mathbf{x}^k, \mathbf{x}^k)}, \;\; \mathbf{x}^k \ne 0; \;\; k = 0, 1, 2, \ldots.$$

При его практической реализации на каждом шаге нормируют текущий вектор: $\mathbf{x}^k := \mathbf{x}^k/\|\mathbf{x}^k\|_2$.

Теорема 17.1 *Пусть A — матрица простой структуры. Пусть $|\lambda_1| > |\lambda_2| \ge |\lambda_3| \ge \ldots \ge |\lambda_n|$ и L — линейная оболочка $\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_n$. Тогда для степенного метода при условии $\mathbf{x}^0 \notin L$ справедлива оценка $\lambda^k = \lambda_1 + O(|\lambda_2/\lambda_1|^k)$.*

Доказательство. Пусть $\mathbf{x}^0 = \sum_{i=1}^{n} c_i\mathbf{e}_i$. Тогда

$$\mathbf{x}^k = c_1\lambda_1^k\mathbf{e}_1 + \sum_{i=2}^{n} c_i\lambda_i^k\mathbf{e}_i, \;\; \mathbf{x}^{k+1} = c_1\lambda_1^{k+1}\mathbf{e}_1 + \sum_{i=2}^{n} c_i\lambda_i^{k+1}\mathbf{e}_i, \;\; c_1 \ne 0.$$

Отсюда следует, что

$$(\mathbf{x}^k, \mathbf{x}^k) = \sum_{i,j=1}^{n} c_ic_j\lambda_i^k\lambda_j^k(\mathbf{e}_i, \mathbf{e}_j) = \lambda_1^{2k}c_1^2\Big(1 + O\left(\left|\tfrac{\lambda_2}{\lambda_1}\right|^k\right)\Big),$$
$$(\mathbf{x}^{k+1}, \mathbf{x}^k) = \sum_{i,j=1}^{n} c_ic_j\lambda_i^{k+1}\lambda_j^k(\mathbf{e}_i, \mathbf{e}_j) = \lambda_1^{2k+1}c_1^2\Big(1 + O\left(\left|\tfrac{\lambda_2}{\lambda_1}\right|^k\right)\Big).$$

Поэтому $\dfrac{(x^{k+1}, x^k)}{(x^k, x^k)} = \lambda_1 + O\left(|\lambda_2/\lambda_1|^k\right)$. Теорема доказана.

Следствие 17.1 *Если в условии Теоремы 1 матрица A является симметричной, то для степенного метода справедлива оценка $\lambda^k = \lambda_1 + O(|\lambda_2/\lambda_1|^{2k})$.*

Доказательство в точности повторяет доказательство предыдущей теоремы. Более сильная скорость сходимости получается из-за ортонормированности $(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij}$ базиса $\{\mathbf{e}_i\}$.

Следствие 17.2 *Пусть матрица A размерности $n \times n$ имеет n различных по модулю собственных значений. Предположим, что $\mathbf{x}^0$ принадлежит линейной оболочке некоторых собственных векторов $\mathbf{e}_{i_1}$, $\mathbf{e}_{i_2}$, …, $\mathbf{e}_{i_t}$, но не принадлежит никакой их линейной подоболочке. Тогда итерации степенного метода в точной арифметике сходятся к максимальному по модулю λ из λ_{i_k}, $1 \le k \le t$.*

При численных расчетах как правило (конечно, возможно исключение: $A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$, $\mathbf{x}^0 = (0, 1)^T$) имеется ненулевой (возможно близкий к машинной погрешности) коэффициент c_1 в разложении $\mathbf{x}^0$ по векторам $\mathbf{e}_i$, поэтому метод сходится к λ_1 — к максимальному по модулю собственному значению. Сходимость к следующему по абсолютной величине λ_2 можно обеспечить только постоянным исключением каким-либо способом вектора $\mathbf{e}_1$ из очередного приближения $\mathbf{x}^k$. Например, можно вычислить $\tilde{\mathbf{e}}_1$ — собственный

вектор сопряженной матрицы A^T, соответствующий λ_1, и ортогонализовывать $\mathbf{x}^k$ к $\tilde{\mathbf{e}}_1$.

Метод обратной итерации. Этот метод, по сути соответствующий степенному методу для матрицы A^{-1}, применяют для вычисления наименьшего по модулю собственного значения λ:

$$\mathbf{x}^k := \mathbf{x}^k/\|\mathbf{x}^k\|_2,\ A\mathbf{x}^{k+1} = \mathbf{x}^k,\ \lambda^k = \frac{(\mathbf{x}^k, \mathbf{x}^{k+1})}{(\mathbf{x}^{k+1}, \mathbf{x}^{k+1})}.$$

При этом на каждом шаге алгоритма требуется решать систему $A\mathbf{x}^{k+1} = \mathbf{x}^k$.

Теорема 17.2 *Пусть A — матрица простой структуры. Пусть $|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_{n-1}| > |\lambda_n|$ и L — линейная оболочка $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, ..., \mathbf{e}_{n-1}$. Тогда для метода обратной итерации при условии $\mathbf{x}^0 \notin L$ справедлива оценка $\lambda^k = \lambda_n + O(|\lambda_n/\lambda_{n-1}|^k)$.*

Степенной метод и метод обратной итерации можно также применять к матрице $A - cI$, что позволяет влиять на сходимость. Например, если с высокой точностью известно приближение $\tilde{\lambda}$ к некоторому собственному значению λ, то метод обратной итерации с параметром $c = \tilde{\lambda}$ обычно сходится за 1-2 итерации. Скорость сходимости существенно замедляется при вычислении одного из группы близких собственных значений.

Теорема 17.3 *Пусть собственные значения симметричной матрицы A удовлетворяют цепочке неравенств $\lambda_1 < \lambda_2 < \ldots < \lambda_n$. Тогда итерационный процесс*

$$\mathbf{x}^{k+1} = (A - cI)\mathbf{x}^k,\ \lambda^k = c + \frac{(\mathbf{x}^{k+1}, \mathbf{x}^k)}{(\mathbf{x}^k, \mathbf{x}^k)}.$$

в зависимости от параметра c сходится к такому λ_s, для которого справедливо $|\lambda_s - c| = \max\limits_i |\lambda_i - c|$. Процесс сходится к λ_1 при $c > (\lambda_1 + \lambda_n)/2$ или к λ_n при $c < (\lambda_1 + \lambda_n)/2$. Скорость сходимости равна $O(q^{2n})$, где $q = \max\limits_{i \neq s} |\lambda_i - c|/|\lambda_s - c|$.

Теорема 17.4 *В условиях теоремы итерационный процесс сходится к λ_1 с наилучшей скоростью при $c_1 = (\lambda_2 + \lambda_n)/2$, сходится к λ_n с наилучшей скоростью при $c_n = (\lambda_1 + \lambda_{n-1})/2$.*

Доказательство. Из теоремы следует, что оптимальное значение c_s для $s = 1, n$ является решением следующей минимаксной задачи: $\min\limits_c \max\limits_{i \neq s} |\lambda_i - c|/|\lambda_s - c|$. Рассматриваемая функция линейна по λ_i, поэтому принимает максимальное значение в граничных точках. Например, при $s = 1$ это соответствует $\min\limits_c \max\{|\lambda_2 - c|/|\lambda_1 - c|, |\lambda_n - c|/|\lambda_1 - c|\}$. Можно показать

(например, графически), что оптимальное значение $c_1 = (\lambda_2 + \lambda_n)/2$. Аналогично $c_n = (\lambda_1 + \lambda_{n-1})/2$. Скорость сходимости степенного метода при оптимальном сдвиге зависит от $\lambda_1, \ldots, \lambda_n$ и не может стать сколь угодно высокой за счет параметра сдвига.

Теорема 17.5 *Пусть собственные значения симметричной матрицы A удовлетворяют цепочке неравенств $\lambda_1 < \lambda_2 < \ldots < \lambda_n$. Тогда метод обратной итерации со сдвигом*

$$(A - cI)\mathbf{x}^{k+1} = \mathbf{x}^k,\ \lambda^k = c + \frac{(\mathbf{x}^k, \mathbf{x}^{k+1})}{(\mathbf{x}^{k+1}, \mathbf{x}^{k+1})}$$

в зависимости от параметра c сходится к λ_t, для которого справедливо $|\lambda_t - c| = \min\limits_i |\lambda_i - c|$. Скорость сходимости равна $O(q^{2n})$, где $q = |\lambda_t - c|/\min\limits_{i \neq t} |\lambda_i - c|$. Процесс в зависимости от значения c может сходиться к любому λ_t. При этом $q \to 0$, если $c \to \lambda_t$.

Скорость сходимости метода обратной итерации со сдвигом можно значительно повысить, если изменять значение сдвига от шага к шагу. Рассмотрим функцию $R_A(\mathbf{x}) = \dfrac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}$, называемую *отношением Рэлея*.

Теорема 17.6 *Пусть $A = A^T$. Тогда*

$$\lambda_{\max}(A) = \max_{\mathbf{x} \neq 0} R_A(\mathbf{x}), \quad \lambda_{\min}(A) = \min_{\mathbf{x} \neq 0} R_A(\mathbf{x}).$$

Доказательство. Пусть λ_i — i-е собственное значение и $A\mathbf{e}_i = \lambda_i \mathbf{e}_i$. Из условия $A = A^T$ следует, что собственные векторы образуют ортонормальный базис. При этом $\lambda_i = \dfrac{(A\mathbf{e}_i, \mathbf{e}_i)}{(\mathbf{e}_i, \mathbf{e}_i)} = (A\mathbf{e}_i, \mathbf{e}_i)$. Представив произвольный вектор $\mathbf{x}$ в виде разложения по собственным векторам, получим требуемый результат.

Лемма 17.3 *Пусть $A = A^T$. Тогда $\forall \mathbf{x} \in \mathbf{R}^n$ и $\forall \mu \in \mathbf{R}^1$ имеет место свойство минимальности невязки*

$$\|\big(A - R_A(\mathbf{x})\, I\big)\mathbf{x}\|_2 \leq \|\big(A - \mu\, I\big)\mathbf{x}\|_2.$$

Доказательство. Минимум квадратичной по μ функции $\|(A - \mu I)\mathbf{x}\|_2^2 = (A\mathbf{x}, A\mathbf{x}) - 2\mu(A\mathbf{x}, \mathbf{x}) + \mu^2(\mathbf{x}, \mathbf{x})$ достигается при $\mu = R_A(\mathbf{x})$.

Неравенство из леммы показывает, что наилучший сдвиг для метода обратной итерации, который можно получить из найденного приближения $\mathbf{x}^k$ к собственному вектору, есть отношение Рэлея $R_A(\mathbf{x}^k)$. При таком выборе

сдвига сходимость к собственному вектору, если она есть, является кубической: $\lim\limits_{k\to\infty} |\varphi_{k+1}/\varphi_k^3| \le 1$, где φ_k — угол между собственным вектором $\mathbf{e}$ и его приближением $\mathbf{x}^k$.

Теорема 17.7 *Пусть метод обратной итерации со сдвигом сходится к собственному значению λ_c матрицы $A = A^T$. Тогда*

$$|\lambda_c - \lambda^k| \le \frac{1}{\|\mathbf{x}^{k+1}\|_2},$$

т.е. величина $\|\mathbf{x}^{k+1}\|_2^{-1}$ характеризует скорость сходимости итерационного процесса.

Доказательство. Для приближений $\mathbf{x}^k$, $\mathbf{x}^{k+1}$ метода обратной итерации справедливо выражение

$$R_A(\mathbf{x}^{k+1}) = \frac{(A\mathbf{x}^{k+1}, \mathbf{x}^{k+1})}{(\mathbf{x}^{k+1}, \mathbf{x}^{k+1})} = c + \frac{(\mathbf{x}^k, \mathbf{x}^{k+1})}{(\mathbf{x}^{k+1}, \mathbf{x}^{k+1})} = \lambda^k.$$

Отсюда и из леммы получаем, что

$$1 = \|\mathbf{x}^k\|_2 = \|(A - cI)\mathbf{x}^{k+1}\|_2 \ge \|(A - R_A(\mathbf{x}^{k+1})I)\mathbf{x}^{k+1}\|_2 =$$

$$= \|(A - \lambda^k I)\mathbf{x}^{k+1}\|_2 \ge \min_i |\lambda_i - \lambda^k| \|\mathbf{x}^{k+1}\|_2.$$

И так как метод сходится к λ_c, то $\min\limits_i |\lambda_i - \lambda^k| = |\lambda_c - \lambda^k|$. Отсюда следует искомая оценка. Теорема доказана.

Рассмотрим методы нахождения нескольких (всех) собственных значений и поиска инвариантных подпространств. Подпространство $\mathbf{H} \subset \mathbf{R}^n$ называется *инвариантным подпространством* матрицы A, если $A\mathbf{H} \subset \mathbf{H}$. В качестве $\mathbf{H}$ можно взять, например, подпространство $\operatorname{span}\{\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_m}\}$, являющееся линейной оболочкой собственных векторов $\mathbf{e}_{i_j}$.

Теорема 17.8 *Пусть A — матрица размерности $n \times n$ и $\mathbf{X} = \operatorname{span} < \mathbf{x}_1, \ldots, \mathbf{x}_m >$, $\mathbf{x}_i \in \mathbf{R}^n$ — подпространство, задаваемое линейно независимыми столбцами, $X = (\mathbf{x}_1 \ldots \mathbf{x}_m)$ — соответствующая матрица. Подпространство $\mathbf{X}$ тогда и только тогда является инвариантным относительно A, когда найдется такая матрица B размерности $m \times m$, что $AX = XB$. В случае $m = n$ собственные значения матрицы B будут собственными значениями матрицы A.*

Действительно, инвариантность подпространства означает, что существуют константы c_j такие, что $A\mathbf{x}_i = \sum\limits_{j=1}^{m} c_j \mathbf{x}_j$. В данном случае $A\mathbf{x}_i = \sum\limits_{j=1}^{n} b_{ji}\mathbf{x}_j$. Если же $m = n$, то $B = X^{-1}AX$ и матрицы подобны. Если $A\mathbf{x}_i = \lambda_i \mathbf{x}_i$, то $B = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$.

Ортогональная итерация (Метод итерирования подпространств). Рассмотрим следующий итерационный алгоритм нахождения m-мерного инвариантного подпространства матрицы A, образованного линейной комбинацией собственных векторов, отвечающих m наибольшим по модулю собственным значениям. Возьмем произвольную матрицу $V_0 \in \mathbf{R}^{n\times m}$ с m ортонормированными столбцами;
для $k = 0, 1, \ldots$ вычислим
$W_{k+1} = AV_k$;
$W_{k+1} = V_{k+1}R_{k+1}$, т. е. проведем ортонормировку столбцов W_{k+1};
пока V_k и V_{k+1} значительно отличаются.

Замечание. При $|\lambda_m| > |\lambda_{m+1}|$ и естественных условиях на V_0 метод сходится, и столбцы матрицы V_∞ задают базис в искомом инвариантном подпространстве. Скорость сходимости характеризуется величиной $|\lambda_{m+1}/\lambda_m|$.

QR-алгоритм. Модифицируем метод ортогональной итерации так, чтобы он (как и метод обратной итерации) допускал сдвиги и обращения матрицы. Это приводит к QR-алгоритму — наиболее популярному методу вычисления всех собственных значений и векторов матрицы $A \in \mathbf{R}^{n\times n}$ не слишком большой размерности. Положим $A_0 = A$;
для $k = 0, 1, 2, \ldots$ вычислим
$A_k = Q_k R_k$, где Q_k — ортогональная (в комплексном случае — унитарная) матрица, R_k — верхняя треугольная матрица;
определим $A_{k+1} = R_k Q_k$.
пока A_{k+1} значительно отличается от верхней треугольной матрицы.

Лемма 17.4 *Собственные значения матриц A и A_k из QR-алгоритма при $k = 1, 2, \ldots$ совпадают.*

Доказательство. Так как $A_{k+1} = R_k Q_k = Q_k^T (Q_k R_k) Q_k = Q_k^T A_k Q_k$, то из $Q_k^{-1} = Q_k^T$ следует, что матрицы A_{k+1} и A_k ортогонально подобны и имеют одинаковые собственные значения.

Отметим, что в случае невырожденной матрицы QR-разложение с положительными элементами r_{ii} треугольной матрицы R единственно, поэтому в дальнейшем будем полагать $r_{ii} \ge 0$ для каждой верхней треугольной матрицы.

Лемма 17.5 *Пусть A_k — матрица из QR-алгоритма, а V_k — из метода ортогональной итерации для $m = n$, $V_0 = I$. Тогда $A_k = V_k^T A V_k$.*

Доказательство. Проверим равенство $A_k = V_k^T A V_k$ по индукции. Для QR-алгоритма имеем: $A = A_0 = Q_0 R_0$, $A_1 = R_0 Q_0 = Q_0^T A Q_0$. Для метода ортогональной итерации имеем: $W_1 = AI = A = V_1 R_1$. Отсюда $V_1 = Q_0$, $A_1 = V_1^T A V_1$. База индукции проверена. Шаг индукции: пусть $A_k = V_k^T A V_k$. Тогда для QR-алгоритма имеем

$$A_{k+1} = Q_k^T A_k Q_k, \quad A_{k+1} = Q_k^T V_k^T A V_k Q_k.$$

Т.е. для доказательства леммы необходимо показать, что $V_{k+1} = V_k Q_k$.

Согласно шагу индукции $AV_k = V_k A_k$. Так как для QR-алгоритма имеем, что $A_k = Q_k R_k$, то $AV_k = V_k Q_k R_k$. Но для метода ортогональной итерации $W_{k+1} = AV_k = V_{k+1}\tilde{R}_{k+1}$. В силу единственности QR-разложения с положительными элементами r_{ii} получим $V_k Q_k = V_{k+1}$, $R_k = \tilde{R}_{k+1}$. Лемма доказана.

Утверждение 17.2 *Пусть все собственные числа матрицы A по модулю различны. Тогда QR-алгоритм сходится к действительной форме Шура — "почти верхней треугольной" матрице A_∞, для которой собственные значения вычисляются явно.*

Пример 17.1. Для матрицы $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ применение QR-алгоритма дает: $A_0 = A_1 = \ldots = A_k \quad \forall k$.

QR-алгоритм со сдвигом. Скорость сходимости QR-алгоритма можно существенно повысить, если применять разложение к матрицам вида $A_k - c_k I$. Положим $A_0 = A$;
для $k = 0, 1, 2, \ldots$ вычислим
$A_k - c_k I = Q_k R_k$;
$A_{k+1} = R_k Q_k + c_k I$.
пока A_{k+1} значительно отличается от верхней треугольной матрицы.

Если c_k выбрать равным некоторому собственному значению, то у матрицы R_k на диагонали появится нуль, поэтому можно найти соответствующий собственный вектор, и задачу для A_{k+1} сформулировать как задачу на единицу меньшей размерности.

Лемма 17.6 *Собственные значения матриц A и A_k из QR-алгоритма со сдвигом при $k = 1, 2, \ldots$ совпадают.*

Доказательство. Покажем, что матрицы A_k и A_{k+1} подобны:

$$A_{k+1} = R_k Q_k + c_k I = Q_k^T (Q_k R_k + c_k I) Q_k = Q_k^T A_k Q_k.$$

Лемма доказана.

При практическом использовании метода сначала проводят масштабирование (уравновешивание) матрицы A, сближающее ее норму со спектральным радиусом, а затем приводят к верхней форме Хессенберга H ($h_{ij} = 0$ при $i > j+1$), которая инвариантна относительно QR-итераций. Само же разложение применяют к матрицам вида $A_k - c_k I$. Полное обоснование QR-алгоритма на данный момент отсутствует, однако для практических задач метод работает стабильно.

4. Приближение функций

Лекция 18. Полиномиальная интерполяция

Пусть $a = x_1 < x_2 < \ldots < x_n = b$ — набор различных точек (узлов) на отрезке $[a,b]$, в которых заданы значения достаточно гладкой функции $f(x)$ так, что $f_i = f(x_i)$, $i = 1,\ldots,n$. Требуется построить многочлен $L_n(x)$, принимающий в точках x_i значения f_i, и оценить погрешность приближения на всем отрезке $[a,b]$ в равномерной норме.

Пример 18.1. Для функции Рунге $f(x) = \dfrac{1}{25x^2+1}$ на отрезке $[-1,1]$ на равномерной сетке $x_i = x_{i-1} + h$ будем иметь $\max\limits_{0.726\ldots\le|x|<1} |f(x) - L_n(x)| \to \infty$ при $n \to \infty$.

Пример 18.2. На равномерной сетке также нет сходимости в C-норме для функции $|x|$ на $[-1,1]$.

Теорема 18.1 *Для любой таблицы узлов интерполяции $(x_1^n,\ldots,x_n^n)$, заданной на отрезке $[a,b]$, существует такая непрерывная на этом отрезке функция $f(x)$, что погрешность $||f(x) - L_n(x)||$ в равномерной норме не стремится к нулю при $n \to \infty$.*

Замечание. Построение $L_n(x)$ формально эквивалентно задаче нахождения коэффициентов c_i из системы уравнений $\sum\limits_{i=0}^{n-1} c_i x_j^i = f_j$ при $j = 1,\ldots,n$. Поэтому существование и единственность многочлена $L_n(x)$ степени $n-1$, принимающего в n различных точках заданные значения, следует из невырожденности данной системы — определитель задачи является отличным от нуля определителем Вандермонда. Однако при больших n система близка к вырожденной. Действительно, можно считать, что узлы интерполяции принадлежат отрезку $[0,1]$. Тогда сеточные функции x_j^{n-2}, x_j^{n-1} при больших n почти неотличимы, т.е. базис из данных многочленов почти линейно зависим, а соответствующие столбцы матрицы почти равны.

Поэтому при численном построении $L_n(x)$ обычно не вычисляют коэффициенты c_i, а поступают следующим образом.

Интерполяционный многочлен Лагранжа. Приведем в явном виде вспомогательные многочлены $\Phi_i(x)$ степени $n-1$, удовлетворяющие условиям $\Phi_i(x_i) = 1$, $\Phi_i(x_j) = 0$ при $j \ne i$, и далее с их помощью запишем формулу для искомого интерполяционного многочлена:

$$\Phi_i(x) = \prod_{\substack{j=1\\ j\ne i}}^{n} \frac{x - x_j}{x_i - x_j}\,, \quad L_n(x) = \sum_{i=1}^{n} f_i\,\Phi_i(x)\,.$$

Многочлен $L_n(x)$ называется интерполяционным многочленом в форме Лагранжа. Так как интерполяционный многочлен единственен, то $L_n(x)$ есть решение поставленной задачи. Если в точной арифметике привести подобные слагаемые, то будут получены те же значения коэффициентов, что и при решении системы. При численных расчетах этого делать не рекомендуется, так как вычислительная погрешность в случае больших n может существенно исказить ответ.

Отметим, что если коэффициенты многочлена известны, то можно применить следующую схему вычисления $L_n(x)$:

$$L_n(x) = (\ldots((c_{n-1}x + c_{n-2})\,x + c_{n-3})\,x + \ldots)\,x + c_0.$$

Такая форма записи имеет неулучшаемое по порядку число арифметических действий $O(n)$, необходимое для нахождения $L_n(x)$.

Теорема 18.2 *Пусть $f(x) \in C^n[a,b]$. Тогда для любой точки $x \in [a,b]$ существует такая точка $\xi(x) \in [a,b]$, что справедливо равенство*

$$f(x) - L_n(x) = \frac{f^{(n)}(\xi(x))}{n!}\,\omega_n(x), \quad \text{где} \;\; \omega_n(x) = \prod_{i=1}^{n}(x - x_i)\,.$$

Доказательство. Если $x = x_i$, то равенство очевидно. Для фиксированного $x \in [a,b]$, $x \ne x_i$ построим функцию $\varphi(t) = f(t) - L_n(t) - K\omega_n(t)$, где $K = \frac{f(x)-L_n(x)}{\omega(x)}$. Тогда $\varphi(t)$ обращается в нуль в точках $x, x_1, \ldots, x_n$; по теореме Ролля $\varphi'(t)$ равна нулю в некоторых n точках, и т.д., $\varphi^{(n)}(t)$ — в некоторой $\xi(x) \in [a,b]$. Из условия $\varphi^{(n)}(\xi(x)) = 0$ и выбранного для K представления находим $K = \frac{f^{(n)}(\xi(x))}{n!} = \frac{f(x)-L_n(x)}{\omega(x)}$.

Следствие 18.1 *Имеет место следующая оценка погрешности в равномерной норме*

$$\|f(x) - L_n(x)\| \le \frac{\|f^{(n)}(x)\|}{n!}\,\|\omega_n(x)\|, \quad \text{где} \quad \|f(x)\| = \sup_{x\in[a,b]} |f(x)|\,.$$

Величина $\lambda_n = \max\limits_{x\in[a,b]} \sum\limits_{i=1}^{n} |\Phi_i(x)|$ называется *константой Лебега интерполяционного процесса*. Значение λ_n зависит от взаимного расположения узлов, а скорость ее роста при увеличении n влияет на скорость сходимости многочлена $L_n(x)$ к функции $f(x)$ в равномерной норме и на оценку для вычислительной погрешности интерполяции.

Например, для системы равноотстоящих узлов λ_n растет экспоненциально, т.е. $C_1 2^n/n^{3/2} \le \lambda_n \le C_2 2^n$, а построенный на такой *равномерной сетке*

интерполяционный полином $L_n(x)$ при большом числе узлов может принципиально отличаться от приближаемой функции. Например, для функции Рунге на $[-1,1]$ имеем $\|f(x)-L_n(x)\| \geq \varepsilon p^n$, $\varepsilon>0, p>1$. В этом случае принято говорить, что алгоритм имеет насыщение.

Для чебышёвских узлов верна оценка $\lambda_n \leq C_3 \ln n$. И если f, например, функция Рунге или произвольная аналитическая в каждой точке отрезка функция, то для чебышёвских узлов имеет место оценка $\|f-L_n\| \leq Cq^n$, $q<1$, где величины q, C не зависят от n. Следовательно, погрешность убывает с ростом n, и алгоритм является алгоритмом без насыщения.

Минимизация остаточного члена погрешности. Запишем полученное ранее представление погрешности

$$f(x)-L_n(x)=\frac{f^{(n)}(\xi(x))}{n!}\omega_n(x).$$

Рассмотрим класс функций $\mathcal{F}=\{f:\ f\in C^{(n)}[a,b], \|f^{(n)}\|_C \leq A_n\}$. Для набора узлов $\{x_i\}_1^n$ определим соответственно погрешность интерполяции для функции f и для класса $\mathcal{F}$:

$$l(f,\{x_i\})=\|f-L_n\|, \quad l(F,\{x_i\})=\sup_{f\in F} l(f,\{x_i\}).$$

Нас интересует оптимальный набор узлов $\{\bar{x}_i\}$:

$$\inf_{\{x_i\}} l(\mathcal{F},\{x_i\}) = l(\mathcal{F},\{\bar{x}_i\}) = l(\mathcal{F}).$$

Возьмем некоторый набор $\{x_i\}$. Тогда

$$l(\mathcal{F},\{x_i\})=\sup_{f\in\mathcal{F}}\|\frac{f^{(n)}(\xi(x))}{n!}\omega_n(x)\| \leq \frac{A_n}{n!}\|\omega_n\|.$$

Отсюда следует, что $\inf\limits_{\{x_i\}} l(\mathcal{F},\{x_i\}) \leq \dfrac{A_n}{n!}\inf\limits_{\{x_i\}}\|\omega_n\|$. Решением задачи $\inf\limits_{\{x_i\}}\|\omega_n\| = \inf\limits_{\{x_i\}}\max\limits_{x\in[a,b]}|(x-x_1)\dots(x-x_n)|$ является нормированный многочлен Чебышёва, x_i — его корни: $x_i=\frac{a+b}{2}+\frac{b-a}{2}\cos\frac{\pi(2i-1)}{2n}$. При этом

$$\omega_n(x)=(b-a)^n 2^{1-2n} T_n\left(\frac{2x-(a+b)}{b-a}\right), \quad \|\omega_n\|=(b-a)^n 2^{1-2n}.$$

Это приводит к оценке

$$\inf_{\{x_i\}} l(\mathcal{F},\{x_i\}) \leq \frac{A_n}{n!}(b-a)^n 2^{1-2n}.$$

Результат оптимизации на классе $\mathcal{F}$ всегда не лучше, чем результат оптимизации на конкретном элементе $f_0\in\mathcal{F}$. Поэтому для $f_0(x)=\frac{A_n}{n!}x^n$ имеем

$$\inf_{\{x_i\}} l(\mathcal{F},\{x_i\}) \geq \inf_{\{x_i\}} l(f_0,\{x_i\}) = \frac{A_n}{n!}(b-a)^n 2^{1-2n},$$

т.е. найденная для класса оценка сверху достигается на функции $f_0(x)$, т.е. является точной.

Таким образом, интерполяция по узлам Чебышёва оптимальна на классе $\mathcal{F}$. Напомним, что для произвольной аналитической функции f интерполяционный полином, построенный по чебышёвским узлам, сходится к f в непрерывной норме.

Лекция 19. Задачи наилучшего приближения

Наилучшее приближение в линейном нормированном пространстве. Пусть $\mathcal{L}$ — линейное (векторное) нормированное пространство. Требуется найти наилучшее приближение для $f\in\mathcal{L}$ в виде линейной комбинации элементов линейно независимой системы $g^{(1)},\dots,g^{(n)}\in\mathcal{L}$:

$$\inf_{c_1,\dots,c_n}\|f-\sum_{i=1}^n c_i g^{(i)}\|.$$

Соответствующее решение $\sum\limits_{i=1}^n c_i^f g^{(i)}$ (если существует) называется элементом наилучшего приближения.

Пример 19.1. Пусть $\mathcal{L}=R^N$, $f=(f_1,\dots,f_N)$, $g^{(i)}=(x_1^{i-1},\dots,x_N^{i-1})$, $\|f\|=\max\limits_{i=1,\dots,N}|f_i|$.

Теорема 19.1 *Элемент наилучшего приближения (т.е. соответствующий набор $\{c_i^f\}$) в линейном нормированном пространстве существует.*

Доказательство. Рассмотрим функцию $F_f(c_1,\dots,c_n)=\|f-\sum\limits_{i=1}^n c_i g^{(i)}\|$. По неравенству треугольника имеем:

$$|F_f(\mathbf{c})-F_f(\tilde{\mathbf{c}})| = \left|\|f-\sum_{i=1}^n c_i g^{(i)}\| - \|f-\sum_{i=1}^n \tilde{c}_i g^{(i)}\|\right| \leq$$
$$\leq \|\sum_{i=1}^n (c_i-\tilde{c}_i)g^{(i)}\| \leq \sum_{i=1}^n |c_i-\tilde{c}_i|\,\|g^{(i)}\|.$$

Таким образом $F_f(\mathbf{c})$ непрерывна по $\mathbf{c}$ для любого $f \in \mathcal{L}$ (в том числе для $f \equiv 0$). Рассмотрим в n-мерном пространстве коэффициентов шар $D_R = \{\mathbf{c} : |\mathbf{c}| = \sqrt{\sum_{i=1}^n c_i^2} \le R\}$. Функция $F_f(\mathbf{c})$ непрерывна, следовательно, в некоторой точке $\mathbf{c}^f \in D_R$ достигает точной нижней грани. При этом $F_f(\mathbf{c}^f) \le F_f(0) = \|f\|$. Покажем, что если R достаточно велико, то вне шара D_R оценка заведомо хуже. Вне шара имеем $|\mathbf{c}| > R$, следовательно,

$$F_f(\mathbf{c}) \ge \|\sum_{i=1}^n c_i g_i\| - \|f\| = |\mathbf{c}| F_0\left(\frac{\mathbf{c}}{|\mathbf{c}|}\right) - \|f\| > R F_0\left(\frac{\mathbf{c}}{|\mathbf{c}|}\right) - \|f\|.$$

Так как непрерывная функция $F_0(\tilde{\mathbf{c}}) = \|\tilde{c}_1 g^{(1)} + \ldots + \tilde{c}_n g^{(n)}\|$ при $\tilde{\mathbf{c}} = \dfrac{\mathbf{c}}{|\mathbf{c}|}$ рассматривается на единичной сфере $S_1 = \{\tilde{\mathbf{c}} : |\tilde{\mathbf{c}}| = \sqrt{\sum_{i=1}^n \tilde{c}_i^2} = 1\}$, то на S_1 функция достигает своей точной нижней грани: $\inf\limits_{\tilde{\mathbf{c}} \in S_1} F_0(\tilde{\mathbf{c}}) = F_0(\mathbf{c}^0)$. При этом $F_0(\mathbf{c}^0) > 0$, так как $0 \notin S_1$, а условие $F_0(\mathbf{c}^0) = 0$, $\mathbf{c}^0 \ne 0$ означает, что $g^{(i)}$ линейно зависимы.

Выберем $R = 2\frac{\|f\|}{F_0(\mathbf{c}^0)}$. Тогда вне рассматриваемого шара D_R имеем $F_f(\mathbf{c}) > 2\frac{\|f\| F_0(\mathbf{c}^0)}{F_0(\mathbf{c}^0)} - \|f\| = \|f\|$, следовательно, найденный на шаре минимум $F_f(\mathbf{c}^f)$ доставляет минимум по всему пространству $\mathbf{R}^n$.

Наилучшее приближение в гильбертовом пространстве. Пусть в пространстве H с нормой $\|\cdot\| = (\cdot,\cdot)^{1/2}$ задан элемент f и линейно независимая система $\{g^{(i)}\}_{i=1}^n \in H$. Требуется найти коэффициенты $\{c_i\}_{i=1}^n$ из условия минимума функционала $\Phi(\mathbf{c}) = \|f - \sum_{i=1}^n c_i g^{(i)}\|^2$, т.е. требуется найти элемент $g \in H$ вида $g = \sum_{i=1}^n c_i g^{(i)}$, наилучшим образом приближающий f:

$$\inf_{\tilde{\mathbf{c}}} \Phi(\tilde{\mathbf{c}}) = (f - \sum_{i=1}^n c_i g^{(i)}, f - \sum_{i=1}^n c_i g^{(i)}).$$

В точке минимума выполняются условия $\dfrac{\partial \Phi}{\partial c_j} = 0$, что приводит к линейной системе уравнений

$$-2(f - \sum_{i=1}^n c_i g^{(i)}, g^{(j)}) = 0 \Rightarrow \sum_{i=1}^n c_i (g^{(i)}, g^{(j)}) = (f, g^{(j)}),\ j = 1, \ldots, n$$

с матрицей Грама $G_n = \left[(g^{(i)}, g^{(j)})\right]$, $G_n = G_n^T$.

Теорема 19.2 *Если $g^{(1)}, \ldots, g^{(n)}$ линейно независимы, то G_n положительно определена, и задача $G_n \mathbf{c} = \mathbf{b}$ имеет единственное решение при всех $\mathbf{b}$.*

Доказательство. Для линейно независимой системы имеем

$$(G_n \mathbf{c}, \mathbf{c}) = \sum_{i,j=1}^n (g^{(i)}, g^{(j)}) c_j c_i = (\sum_{i=1}^n c_i g^{(i)}, \sum_{j=1}^n c_j g^{(j)}) = \|\sum_{i=1}^n c_i g^{(i)}\|^2,$$

т.е. $(G_n \mathbf{c}, \mathbf{c}) > 0$ при $c \ne 0$.

Пример 19.2. Пусть в пространстве $L_2(0,1)$ выбрана система $\{g^{(i)}\}_{i=1}^n = \{1, x, \ldots, x^{n-1}\}$ и $(x^{i-1}, x^{j-1}) = \int\limits_0^1 x^{i+j-2} dx = \frac{1}{i+j-1}$.
Требуется решить задачу наилучшего приближения

$$\inf_{\mathbf{c}} \int_0^1 (f - \sum_{i=1}^n c_i x^{i-1})^2 dx.$$

Задача при всех n имеет единственное решение, а набор коэффициентов c_i можно найти из указанной системы уравнений. Однако матрица G_n в данном случае является матрицей Гильберта, обусловленность которой растет экспоненциально: $\|G_n^{-1}\|_\infty \sim \frac{1}{\sqrt{n}}(1+\sqrt{2})^{4n}$. Следовательно, при $n \gtrsim 20$ коэффициенты c_i могут быть найдены с большой погрешностью.

Отметим, что при неудачном выборе базиса (как в данном примере) вычислительная погрешность для больших n может достигать катастрофических размеров, и при добавлении новых функций $g^{(n+1)}, g^{(n+2)}, \ldots$ качество приближения будет только ухудшаться.

Если выбранная система ортонормальна $(g^{(i)}, g^{(j)}) = \delta_i^j$, то $g = \sum\limits_{i=1}^n c_i g^{(i)}$, где $c_i = (f, g^{(i)})$. При этом

$$\|f - g\|^2 = (f - \sum_{i=1}^n c_i g^{(i)}, f - \sum_{i=1}^n c_i g^{(i)}) = (f,f) - \sum_{i=1}^n (f, g^{(i)})^2.$$

Отсюда в том числе следует неравенство Бесселя: $(f,f) \ge \sum_{i=1}^n (f, g^{(i)})^2$.

Многочлен наилучшего равномерного приближения. Пусть $\mathcal{F}$ — пространство ограниченных вещественных функций, определенных на отрезке $[a,b]$ вещественной оси с нормой $\|f(x)\| = \sup\limits_{x \in [a,b]} \big|f(x)\big|$. Для элемента $f \in \mathcal{F}$ отыскивается наилучшее приближение вида $Q_n^0(x) = \sum_{j=0}^n a_j\, x^j$, являющееся решением следующей задачи:

$$\inf_{a_j} \max_{x \in [a,b]} |f(x) - \sum_{j=0}^n a_j\, x^j| = \inf_{a_j} \|f - Q_n\| = \|f - Q_n^0(x)\|.$$

Определение 19.1 *Многочлен $Q_n^0(x)$ называется многочленом наилучшего равномерного приближения для функции $f(x)$, если для любого многочлена $Q_n(x)$ степени n справедливо неравенство $\|f - Q_n^0\| \le \|f - Q_n\|$.*

Такой многочлен существует всегда (по теореме об элементе наилучшего приближения в линейном нормированном пространстве), а его единственность (см. далее) имеет место для непрерывных функций $f(x)$.

Теорема 19.3 *(Валле–Пуссен). Пусть существует $(n+2)$ точки $a \le x_0 < \ldots < x_{n+1} \le b$ такие, что*

$$\operatorname{sign}\big(f(x_i) - Q_n(x_i)\big) \cdot (-1)^i = \text{const},$$

т. е. при переходе от точки к точке разность $f(x_i) - Q_n(x_i)$ меняет знак. Тогда

$$\|f - Q_n^0\| \ge \mu = \min_{i=0,\ldots n+1} |f(x_i) - Q_n(x_i)|.$$

Доказательство. Если $\mu = 0$, то утверждение очевидно. Пусть $\mu > 0$ и $\|Q_n^0 - f\| < \mu$. Тогда полином $Q_n(x) - Q_n^0(x)$ отличен от нулевого. При этом $\operatorname{sign}\big(Q_n(x_i) - Q_n^0(x_i)\big) = \operatorname{sign}\big((Q_n(x_i) - f(x_i)) - (Q_n^0(x_i) - f(x_i))\big) = \operatorname{sign}\big(Q_n(x_i) - f(x_i)\big)$, так как модуль первой разности всегда больше μ, второй не превосходит $\|Q_n^0 - f\|$ и, по предположению, $\mu > \|Q_n^0 - f\|$. Таким образом, ненулевой многочлен $Q_n(x) - Q_n^0(x)$ степени n меняет знак в $(n+2)$ точках, что невозможно. Теорема доказана.

Теорема 19.4 (Чебышёв П.Л.) *Чтобы многочлен $Q_n(x)$ был многочленом наилучшего равномерного приближения непрерывной функции $f(x)$, необходимо и достаточно существования на $[a, b]$ по крайней мере $(n+2)$ точек $x_0 < \ldots < x_{n+1}$ таких, что*

$$f(x_i) - Q_n(x_i) = \alpha(-1)^i \|f - Q_n\|,$$

где $i = 0, \ldots, n+1$; $\alpha = 1$ или $\alpha = -1$ одновременно для всех i.

Точки $x_0, \ldots, x_{n+1}$, удовлетворяющие условию теоремы, называются *точками чебышёвского альтернанса.*

Доказательство достаточности. По условию теоремы, а затем по теореме Валле–Пуссена имеем:

$$\|f - Q_n^0\| \ge \min_{i=0,\ldots,n+1} |f(x_i) - Q_n(x_i)| = \|f - Q_n\|.$$

Таким образом, имеющийся многочлен Q_n приближает f не хуже, чем оптимальный Q_n^0. По определению Q_n^0 это означает, что Q_n является многочленом наилучшего равномерного приближения.

Пример 19.3. Для функции $f(x) = \sin 100x$ многочленом наилучшего равномерного приближения степени $n = 98$ на отрезке $[0, \pi]$ является $Q_{98}^0(x) \equiv 0$. В данном случае имеем точки чебышёвского альтернанса $f(x_k) = (-1)^k$ для $100x_k = \pi/2 + \pi k$, $k = 0, \ldots, 99$.

Теорема 19.5 *Многочлен наилучшего равномерного приближения непрерывной функции единственен.*

Доказательство. Пусть $Q_n^1(x) \ne Q_n^2(x)$, $\|f - Q_n^1\| = \|f - Q_n^2\| = \|f - Q_n^0\| = E$. Тогда

$$\|f - \frac{Q_n^1 + Q_n^2}{2}\| \le \|\frac{f - Q_n^1}{2}\| + \|\frac{f - Q_n^2}{2}\| = E.$$

Отсюда следует, что $Q_n = \frac{Q_n^1 + Q_n^2}{2}$ является многочленом наилучшего равномерного приближения, и для Q_n существуют точки чебышёвского альтернанса $x_0 < \ldots < x_{n+1}$:

$$\big|\frac{Q_n^1(x_i) + Q_n^2(x_i)}{2} - f(x_i)\big| = E, \quad i = 0, \ldots, n+1.$$

Но в этом случае

$$|Q_n^1(x_i) - f(x_i) + Q_n^2(x_i) - f(x_i)| = 2E.$$

И так как $|Q_n^j(x_i) - f(x_i)| \le \max_x |Q_n^j(x) - f(x)| = E$, $j = 1, 2$, то равенство возможно только при $Q_n^1(x_i) - f(x_i) = Q_n^2(x_i) - f(x_i) = \pm E$. Это означает, что $Q_n^1(x_i) = Q_n^2(x_i)$ в $(n+2)$ точках, то есть многочлены $Q_n^{1,2}$ тождественно равны.

Пример 19.4. В классе разрывных функций при $n \ge 1$ теоремы Чебышёва и единственности могут нарушаться:

$$f(x) = \operatorname{sign} x \text{ на } [-1, 1], \;\; Q_1(x) = \alpha\, x, \;\; \alpha \in [0, 2].$$

Следствие 19.1 *Если $f(x)$ — непрерывная симметричная (кососимметричная) относительно $(a+b)/2$ функция, то $Q_n^0(x)$ симметричный (кососимметричный) относительно $(a+b)/2$ многочлен.*

Доказательство может быть получено с учетом единственности Q_n^0 методом от противного.

Лекция 20. Сплайн-интерполяция

Пусть на отрезке $[a, b]$ вещественной оси задана сетка: $a = x_0 < x_1 < \ldots < x_n = b$, $\mathcal{P}_m$ — множество многочленов степени не выше m $(m \ge 1)$, $C^{(r)}[a, b]$

— множество функций, имеющих на $[a,b]$ непрерывные производные до r-го порядка включительно ($r \geq 0$).

Определение 20.1 *Функцию $S_{m,k}(x)$ называют полиномиальным сплайном степени m дефекта k ($1 \leq k \leq m$) с узлами $\{x_i\}$, $i = 0,1,\ldots,n$ для функции $f(x) \in C[a,b]$, если выполнены следующие условия:*

1) на каждом из отрезков $[x_{i-1},x_i]$, $i = 1,\ldots,n$ она является многочленом, т.е. $S_{m,k}(x) \in \mathcal{P}_m[x_{i-1},x_i]$;

2) на всем отрезке $[a,b]$ обладает непрерывностью производных, т.е. $S_{m,k}(x) \in C^{(m-k)}[a,b]$.

Сплайн называется *интерполяционным*, если в узлах $\{x_i\}$ справедливы равенства $S_{m,k}(x_i) = f(x_i)$, $i = 0,1,\ldots,n$; иначе — *аппроксимационным*.

Далее рассматривается только случай $k = 1$, поэтому термин "дефекта k" будем опускать. Также далее задача интерполяции рассматривается на единичном отрезке $[0,1]$ на сетке $x_0 = 0$, $x_i = x_{i-1} + h_i$, $i = 1,\ldots,n$, $x_n = 1$.

Линейный интерполяционный сплайн на каждом отрезке $[x_{i-1},x_i]$ является полиномом первой степени $S_1(x) = a_i x + b_i$ и проходит через точки $(x_i, f(x_i))$, $i = 0,\ldots,n$ (как следствие непрерывен на всем отрезке интерполирования $[x_0,x_n]$). Его коэффициенты a_i, b_i однозначно находятся из условий $S_1(x_i) = f(x_i)$.

Теорема 20.1 *Пусть $f(x) \in C^{(2)}[0,1]$. Тогда линейный интерполяционный сплайн на отрезке $[x_{i-1},x_i]$ имеет вид*

$$S_1(x) = f(x_{i-1})\frac{x - x_i}{x_{i-1} - x_i} + f(x_i)\frac{x - x_{i-1}}{x_i - x_{i-1}}\,,$$

и верна оценка

$$\|f(x) - S_1(x)\|_{C[0,1]} \leq \|f''(x)\|_{C[0,1]}\,\frac{1}{8}\max_{i=1,\ldots,n} h_i^2\,.$$

Доказательство. Из оценки погрешности для интерполяционного многочлена Лагранжа первой степени следует неравенство для каждого отрезка $[x_{i-1},x_i]$ и, следовательно, для $[0,1]$ в целом.

Утверждение 20.1 *Пусть $\mathcal{S}_1$ — множество непрерывных, кусочно непрерывно-дифференцируемых функций $s(x)$, удовлетворяющих условиям*

$$s(x_i) = f(x_i), \quad i = 0,\ldots,n; \quad I_1(s) = \int_0^1 (s'(x))^2 dx < \infty.$$

Тогда решением задачи

$$\inf_{s\in\mathcal{S}_1} I_1(s)$$

является линейный сплайн $S_1(x)$.

Кубический интерполяционный сплайн на каждом из отрезков $[x_{i-1},x_i]$ является полиномом третьей степени $S_3(x) = a_i x^3 + b_i x^2 + c_i x + d_i$, проходит через точки $(x_i, f(x_i))$, $i = 0,\ldots,n$ (два условия на каждом отрезке) и имеет непрерывные первую и вторую производные в точках x_i, $i = 1,\ldots,n-1$ (два условия для каждой внутренней точки). Формально из этих условий для нахождения $4n$ неизвестных коэффициентов имеем $4n-2$ уравнений, еще два уравнения необходимо дополнительно задать.

Теорема 20.2 *Пусть $M_i = S_3''(x_i)$, $i = 0,1,\ldots,n$. Тогда M_i удовлетворяют системе линейных уравнений $C\,M = d$ с прямоугольной матрицей C размерности $(n-1)\times(n+1)$, где C и d соответственно имеют вид:*

$$\begin{pmatrix}
\frac{h_1}{6} & \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & . & . & . & . & 0 \\
. & . & . & . & . & . & . & . & . \\
0 & . & 0 & \frac{h_i}{6} & \frac{h_i+h_{i+1}}{3} & \frac{h_{i+1}}{6} & 0 & . & 0 \\
. & . & . & . & . & . & . & . & . \\
0 & 0 & . & . & . & 0 & \frac{h_{n-1}}{6} & \frac{h_{n-1}+h_n}{3} & \frac{h_n}{6}
\end{pmatrix},$$

$$\left(\frac{f_2-f_1}{h_2} - \frac{f_1-f_0}{h_1}, \ldots, \frac{f_{i+1}-f_i}{h_{i+1}} - \frac{f_i-f_{i-1}}{h_i}, \ldots, \frac{f_n-f_{n-1}}{h_n} - \frac{f_{n-1}-f_{n-2}}{h_{n-1}}\right)^T.$$

Доказательство. По определению $S_3''(x)$ — линейная на каждом отрезке $[x_{i-1},x_i]$ функция. В силу ее непрерывности в точках x_i имеем представление:

$$S_3''(x) = M_{i-1}\frac{x_i - x}{h_i} + M_i\frac{x - x_{i-1}}{h_i}\,.$$

Двукратно интегрируя и учитывая условия $S_3(x_i) = f_i\,, S_3(x_{i-1}) = f_{i-1}$, получим аналитическое представление кубического интерполяционного сплайна на отрезке $[x_{i-1},x_i]$:

$$S_3(x) = M_{i-1}\frac{(x_i - x)^3}{6\,h_i} + M_i\frac{(x - x_{i-1})^3}{6\,h_i} + \left(f_{i-1} - \frac{M_{i-1}h_i^2}{6}\right)\frac{x_i - x}{h_i} + \left(f_i - \frac{M_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i}\,.$$

Отсюда вычислим производную сплайна $S_3'(x)$ слева в точке x_i:

$$S_3'(x_i - 0) = M_{i-1}\frac{h_i}{6} + M_i\frac{h_i}{3} + \frac{f_i - f_{i-1}}{h_i}\,.$$

Аналогично найдем представление $S_3(x)$ на $[x_i,x_{i+1}]$ и вычислим производную сплайна $S_3'(x)$ справа в точке x_i:

$$S_3'(x_i + 0) = -M_i\frac{h_{i+1}}{3} - M_{i+1}\frac{h_{i+1}}{6} + \frac{f_{i+1} - f_i}{h_{i+1}}\,.$$

Непрерывность $S_3'(x)$ в точках $x_i\,, i = 1, \ldots, n-1$, т.е. $S_3'(x_i - 0) = S_3'(x_i + 0)$, порождает искомую систему из $(n-1)$ уравнения относительно $(n+1)$ неизвестных. Теорема доказана.

Имеются три стандартных способа доопределения системы.

1-й способ: $S_3''(x_0) = M_0 = S_3''(x_n) = M_n = 0$ — естественный сплайн.

2-й способ: $S_3'(x_0 + 0) = f'(x_0)$, $S_3'(x_n - 0) = f'(x_n)$.

3-й способ: строим параболу $L_3^0(x)$ по точкам x_0, x_1, x_2, два раза дифференцируя, находим $M_0 = (L_3^0(x))''$; строим параболу $L_3^n(x)$ по точкам x_n, x_{n-1}, x_{n-2}, два раза дифференцируя, находим $M_n = (L_3^n(x))''$.

Утверждение 20.2 *Пусть $\mathcal{S}_2$ — множество непрерывных, непрерывно-дифференцируемых и дважды кусочно непрерывно-дифференцируемых функций $s(x)$, удовлетворяющих условиям*

$$s(x_i) = f(x_i), \quad i = 0, \ldots, n; \quad I_2(s) = \int_0^1 (s''(x))^2 dx < \infty.$$

Тогда решением задачи

$$\inf_{s \in \mathcal{S}_2} I_2(s)$$

является кубический сплайн $S_3(x)$ с условиями $S_3''(x_0) = S_3''(x_n) = 0$.

Теорема 20.3 *Пусть $M_0 = M_n = 0$. Тогда решение системы $C\,M = d$ удовлетворяет неравенству*

$$\max_{1 \le i \le n-1} |M_i| \le 3 \frac{\max\limits_{1 \le i \le n-1} |d_i|}{\min\limits_{1 \le i \le n} h_i}\,.$$

Доказательство. Пусть $\max\limits_i |M_i| = |M_j|\,, 1 \le j \le n-1$. Рассмотрим j-е уравнение системы

$$d_j = M_{j-1}\frac{h_j}{6} + M_j \frac{h_j + h_{j+1}}{3} + M_{j+1}\frac{h_{j+1}}{6}\,,$$

из которого следует неравенство:

$$|d_j| \ge |M_j|\frac{h_j + h_{j+1}}{3} - \left(|M_{j-1}|\frac{h_j}{6} + |M_{j+1}|\frac{h_{j+1}}{6}\right) \ge |M_j|\frac{h_j + h_{j+1}}{6}\,,$$

так как $|M_{j\pm1}| \le |M_j|$. Оценивая левую часть неравенства сверху через $\max\limits_i |d_i|$ и множитель в его правой части снизу как

$$\min_i \frac{h_i + h_{i+1}}{6} \ge \frac{1}{3}\,\min_i h_i\,,$$

приходим к искомой оценке.

Утверждение 20.3 *Пусть $f(x) \in C^{(4)}[0,1]$, $\|f^{(4)}(x)\|_{C[0,1]} \le A_4$, задана сетка с постоянным шагом $h_i = h$, и дополнительные условия для определения кубического интерполяционного сплайна имеют следующий вид*

$$S_3'(x_0 + 0) = f'(x_0), \quad S_3'(x_n - 0) = f'(x_n)\,.$$

Тогда справедлива оценка погрешности

$$\|S_3^{(l)}(x) - f^{(l)}(x)\|_{C[0,1]} \le C_l A_4 h^{4-l}\,, \quad l = 0, 1, 2, 3\,.$$

Используют также **локальные (аппроксимационные) сплайны**, значения которых в узлах, как правило, не совпадают со значениями $f(x)$. Это обстоятельство не принципиально, так как значения $f(x)$ обычно известны приближенно. Рассмотрим построение локального сплайна третьей степени на сетке с постоянным шагом $h = x_{i+1} - x_i$, $i = 0, 1, \ldots, n-1$, для отрезка $[0,1]$. Возьмем *стандартный* сплайн $B(x)$, определяемый соотношениями

$$B(x) = \begin{cases} \dfrac{2}{3} - x^2 + \dfrac{1}{2}|x|^3 & \text{при} \quad |x| \le 1\,, \\ \dfrac{1}{6}\,(2 - |x|)^3 & \text{при} \quad 1 \le |x| \le 2\,, \\ 0 & \text{при} \quad 2 \le |x|\,. \end{cases}$$

Представление $B(x) = C(2-|x|)^3$ при $1 \le |x| \le 2$ следует из непрерывности функции и первых двух ее производных в точках $x = \pm 2$. Если значение C задано, то представление $B(x)$ при $|x| \le 1$ следует из непрерывности функции, двух ее производных в точках $x = \pm 1$ и условия $B(-x) = B(x)$. Величина C определяется из условия нормировки $B(0) + B(1) + B(-1) = 1$.

Локальные сплайны третьей степени $B_2^{(1)}(x)$ и $B_2^{(2)}(x)$ записываются в виде

$$B_2^{(k)}(x) = \sum_{i=-1}^{n+1} \alpha_i^{(k)} B\left(\frac{x - ih}{h}\right), \quad k = 1, 2$$

и отличаются выбором коэффициентов.

При $k = 1$ доопределяют значения f_{-1} и f_{n+1} линейной интерполяцией по значениям f_0, f_1 и f_n, f_{n-1} соответственно и полагают $\alpha_i^{(1)} = f_i$, где $f_i = f(x_i)$ при $0 \le i \le n$.

При $k = 2$ доопределяют значения f_{-2}, f_{-1} и f_{n+1}, f_{n+2} кубической интерполяцией по значениям f_0, f_1, f_2, f_3 и f_n, f_{n-1}, f_{n-2}, f_{n-3} соответственно, и полагают $\alpha_i^{(2)} = (8f_i - f_{i+1} - f_{i-1})/6$.

При любых $\alpha_i^{(k)}$, $k = 1, 2$, функции $B_2^{(k)}(x)$ являются сплайнами третьей степени, причем они тождественно равны нулю вне отрезка $[-3h, 1 + 3h]$. Значения полученных сплайнов в узлах сетки равны некоторому среднему значений функции в ближайших узлах.

Утверждение 20.4 *Пусть* $f(x) \in C^{(2)}[0,1]$, $\|f^{(2)}\|_{C[0,1]} \le A_2$ *и задана сетка с постоянным шагом* $h_i = h$. *Тогда*

$$\| \left(B_2^{(1)}\right)^{(l)} - f^{(l)}\|_{C[0,1]} \le C_l A_2 h^{2-l}, \quad l = 0, 1\,.$$

Утверждение 20.5 *Пусть* $f(x) \in C^{(4)}[0,1]$, $\|f^{(4)}\|_{C[0,1]} \le A_4$ *и задана сетка с постоянным шагом* $h_i = h$. *Тогда*

$$\| \left(B_2^{(2)}\right)^{(l)} - f^{(l)}\|_{C[0,1]} \le C_l A_4 h^{4-l}, \quad l = 0, 1, 2, 3\,.$$

Лекция 21. Интерполяция Чебышёва, Фурье, Паде

Пусть требуется приблизить функцию $f(x)$ на отрезке $[-1,1]$ многочленами. Так как система $\{1, x, \ldots, x^n, \ldots\}$ при больших n близка к линейно зависимой, то в качестве базисных функций возьмем многочлены Чебышёва $\{T_n(x)\}$.

Пример 21.1. Рассмотрим задачу приближения $f(x)$ в гильбертовом пространстве $L_2(-1,1)$:

$$\|f\|_p^2 = (f,f)_p, \quad (f,g)_p = \int\limits_{-1}^{1} p(x) f(x) g(x) dx.$$

В качестве весовой функции удобно взять $p(x) = \frac{1}{\sqrt{1-x^2}}$, так как в этом случае система $\{\frac{1}{\sqrt{\pi}}T_0(x), \frac{\sqrt{2}}{\sqrt{\pi}}T_1(x), \ldots, \frac{\sqrt{2}}{\sqrt{\pi}}T_n(x), \ldots\} = \{\tilde{T}_n(x)\}$ ортонормальна, и искомые коэффициенты в разложении $f(x) \approx \sum\limits_{m=0}^{n} c_m \tilde{T}_m(x)$ определяются явно $c_m = (f, \tilde{T}_m)$. Согласно доказанным ранее результатам построенное приближение является наилучшим для выбранного пространства.

Пример 21.2. Рассмотрим задачу приближения $f(x)$ в пространстве $C[-1,1]$, $\|f\|_C = \max\limits_{x\in[-1,1]} |f(x)|$. Пусть

$$f(x) \approx Q_n(x) = \sum_{m=0}^{n} c_m T_m(x) = Q_{n-1}(x) + c_n T_n(x).$$

Тогда $Q_n(x) - Q_{n-1}(x) = c_n T_n(x)$. Так как $c_n T_n(x)$ является многочленом наименее уклоняющимся от нуля в классе $2^{n-1}c_n x^n + \ldots$, то частичная сумма $Q_{n-1}(x)$ является многочленом наилучшего равномерного приближения

к $Q_n(x)$, поэтому переход от $Q_n(x)$ к $Q_{n-1}(x)$ осуществляется оптимальным образом.

На данной идее основан следующий *телескопический метод*. Пусть получено приближение $f(x) \approx \sum\limits_{m=0}^{N} c_m x^m = P_N(x)$, например, на основе ряда Тейлора, но степень N излишне велика. Построим разложение $P_N(x) = \sum\limits_{m=0}^{N} c_m T_m(x)$, определив коэффициенты c_m рекуррентно, и заменим $P_N(x)$ на частичную сумму $P_N(x) \approx Q_n(x) = \sum\limits_{m=0}^{n} c_m T_m(x)$. Так как на каждом шаге цепочки $P_N(x) \approx Q_{N-1}(x) \approx \ldots \approx Q_n(x)$ имеем наилучшее приближение в равномерной норме, то при такой замене обычно удается значительно понизить степень N с малой погрешностью.

Сделаем следующее замечание. Задача приближения функции по многочленам Чебышёва $f(x) \approx \sum\limits_{m=0}^{n} c_m T_m(x)$, $x \in [-1,1]$ заменой переменных $x = \cos t$ сводится к задаче тригонометрической интерполяции $f(\cos t) \approx \sum\limits_{m=0}^{n} c_m T_m(\cos t) = \sum\limits_{m=0}^{n} c_m \cos(mt)$. Данная замена также сводит задачу приближения $f(x)$ по узлам Чебышёва $x_j = \cos(\frac{\pi j}{n})$ к задаче интерполяции $f(\cos t) \approx \sum\limits_{m=0}^{n} c_m \cos(mt)$ по равноотстоящим узлам $t_j = \frac{\pi j}{n}$.

Таким образом тригонометрическая интерполяция в некотором смысле оптимальна для системы равноотстоящих узлов.

Дискретное преобразование Фурье. Если на $[0,1]$ известна 1-периодическая функция $f(x)$, $f(0) = f(1)$, то в качестве базиса можно взять систему

$$\{1, \cos 2\pi x, \sin 2\pi x, \ldots, \cos 2\pi m x, \sin 2\pi m x, \ldots\}$$

и построить ряд Фурье для $f(x)$. Рассмотрим конечномерные аналоги такого разложения.

Пример 21.3. Пусть на $[0,1]$ известна функция $f(x)$, $f(0) = f(1) = 0$. Рассмотрим $f_j = f(jh)$, $h = 1/N$, $0 \le j \le N$. На пространстве дискретных функций f_j, $f_0 = f_N = 0$ определим скалярное произведение $(f_j, g_j) = \frac{1}{N}\sum\limits_{j=1}^{N-1} f_j g_j$. Тогда система $\{\varphi_j^{(m)} = \sqrt{2}\sin \pi m j h, m = 1, \ldots, N-1\}$ полна и ортонормальна. Следовательно, $f_j = \sum\limits_{m=1}^{N-1} c_m \varphi_j^{(m)}$, $c_m = (f_j, \varphi_j^{(m)})$ и можно предположить, что $f(x) \approx \sum\limits_{m=1}^{N-1} c_m \varphi^{(m)}(x)$.

Отметим, что формально любую последовательность $f_j = f(x_j)$, $j = 1, \ldots, N-1$ можно дополнить нулями $f_0 = f_N = 0$ и построить разложение $f_j = \sum\limits_{m=1}^{N-1} b_m \sin(\pi m j h)$. Однако, если $f(0), f(1) \neq 0$, то полученная непрерывная функция $\tilde{f}(x) = \sum\limits_{m=1}^{N-1} b_m \sin(\pi m x)$ может плохо приближать $f(x)$ вне точек x_j, $j = 1, \ldots, N-1$.

Пример 21.4. Пусть на $[0,1]$ известна функция $f(x)$, $f'(0) = f'(1) = 0$. На пространстве дискретных функций f_j определим скалярное произведение

$$(f_j, g_j) = \tfrac{1}{N}\big(\sum_{j=1}^{N-1} f_j g_j + \tfrac{1}{2}(f_0 g_0 + f_N g_N)\big).$$

Тогда система

$$\{\psi_j^{(m)}\} = \{1, \cos \pi N j h, \sqrt{2}\cos \pi m j h, m = 1, \ldots, N-1\}$$

полна и ортонормальна. Следовательно,

$$f_j = \sum_{m=0}^{N} c_m \psi_j^{(m)},\ c_m = (f_j, \psi_j^{(m)}) \text{ и } f(x) \approx \sum_{m=0}^{N} c_m \psi^{(m)}(x).$$

Пример 21.5. Пусть $f(0) = f(1)$, $f_j = f(jh)$, $h = 1/N$, $0 \le j < N$. На дискретном пространстве N-периодических функций ($f_0 = f_N$) определим скалярное произведение

$$(f, g) = \tfrac{1}{N} \sum_{j=0}^{N-1} f_j \bar{g}_j$$

и выберем следующую базисную систему:

$$\big\{g_j^{(m)} = e^{2\pi \mathbf{i} \frac{mj}{N}}\big\}, m = 0, \ldots N-1.$$

Теорема 21.1 *Система функций $\big\{g_j^{(m)}\big\}, m = 0, \ldots, N-1$, ортонормальна, т.е. $(g^{(m)}, g^{(n)}) = \delta_{mn}$.*

Доказательство.

$$m = n, \quad (g^{(m)}, g^{(m)}) = 1.$$

$$m \neq n, \quad (g^{(m)}, g^{(n)}) = \tfrac{1}{N} \sum_{j=0}^{N-1} e^{2\pi \mathbf{i} \frac{(m-n)j}{N}} =$$

$$= \tfrac{1}{N}(1 + q + \ldots + q^{N-1}) = \tfrac{1}{N}\tfrac{1-q^N}{1-q} = 0, \quad q = e^{2\pi \mathbf{i} \frac{(m-n)}{N}}.$$

Следствие 21.1 *Каждая дискретная функция f_j, $j = 0, \ldots, N-1$ может быть представлена в виде*

$$f_j = \sum_{m=0}^{N-1} A_m e^{2\pi \mathbf{i} \frac{mj}{N}}, \quad A_m = \frac{1}{N} \sum_{j=0}^{N-1} f_j e^{-2\pi \mathbf{i} \frac{mj}{N}},$$

и имеется взаимно однозначное соответствие между значениями $\{f_j\}_{j=0}^{N-1}$ и коэффициентами $\{A_m\}_{m=0}^{N-1}$.

В общем случае преобразование Фурье $\{f_j\} \leftrightarrow \{A_m\}$ невозможно вычислить менее, чем за $O(N)$, и не сложно реализовать за $O(N^2)$ арифметических действий.

Быстрое преобразование Фурье (БПФ). Метод требует при $N = n^2$ порядка $O(N^{3/2})$, а при $N = 2^n$ порядка $O(N \log_2 N)$ арифметических действий. Это дает следующий выигрыш:

$N = 32,\ N^2 = 1024,\ N\log_2 N = 160;$

$N = 128,\ N^2 = 16384,\ N\log_2 N = 896;$

$N = 1024,\ N^2 = 1048576,\ N\log_2 N = 10240.$

Отметим, что алгоритм БПФ принято включать в десятку важнейших численных алгоритмов двадцатого века (современный подход основан на алгоритме Кули–Тьюки, хотя основная идея была известна Гауссу).

Построим расчетные формулы для случая $N = n_1 n_2$. Для m-го коэффициента имеем:

$$A_m = \tfrac{1}{N} \sum_{j=0}^{N-1} f_j e^{-2\pi \mathbf{i} m \frac{j}{N}}.$$

Запишем результат деления с остатком для m, j:

$$m = m_1 + m_2 n_1, \quad j = j_1 + j_2 n_2.$$

Тогда

$$A_m = \tfrac{1}{n_1 n_2} \sum_{j_1=0}^{n_2-1} \sum_{j_2=0}^{n_1-1} f_{j_1 + j_2 n_2} e^{-2\pi \mathbf{i} \frac{(m_1 + m_2 n_1)(j_1 + j_2 n_2)}{N}}.$$

Далее $\frac{(m_1+m_2n_1)(j_1+j_2n_2)}{n_1 n_2} = \frac{mj_1}{N} + \frac{m_1 j_2}{n_1} + m_2 j_2$, т.е. $e^{-2\pi \mathbf{i} \frac{mj}{N}} = e^{-2\pi \mathbf{i} \frac{mj_1}{N}} e^{-2\pi \mathbf{i} \frac{m_1 j_2}{n_1}}$. Следовательно, можно переписать выражение для A_m в виде

$$A_m = \tfrac{1}{n_2} \sum_{j_1=0}^{n_2-1} \left(\tfrac{1}{n_1} \sum_{j_2=0}^{n_1-1} f_{j_1 + j_2 n_2} e^{-2\pi \mathbf{i} \frac{m_1 j_2}{n_1}} \right) e^{-2\pi \mathbf{i} \frac{mj_1}{N}} =$$

$$= \tfrac{1}{n_2} \sum_{j_1=0}^{n_2-1} A(j_1, m_1) e^{-2\pi \mathbf{i} \frac{mj_1}{N}}.$$

Такая форма записи позволяет выделить слагаемые $A(j_1, m_1)$, $j_1 = 0, \ldots, n_2 - 1$, $m_1 = 0, \ldots, n_1 - 1$. Для их подсчета требуется порядка $n_1(n_1 n_2)$ арифметических действий; для вычисления A_m при найденных $A(j_1, m_1)$ потребуется порядка $n_2(n_1 n_2)$ ар. действ.; в результате имеем $O(n_1(n_1 n_2) + n_2(n_1 n_2))$ арифметических действий.

Рекурсивная реализация для $N = 2^n$. Рассмотрим полином $P_N(x) = f_0 + f_1 x + f_2 x^2 + \ldots + f_{N-1} x^{N-1}$. Вычисление коэффициентов A_m по сути

соответствует вычислению значений полинома $P_N(x)$ в корнях N-ой степени из единицы, т.е. в точках $\{x_m = e^{-2\pi \mathbf{i} m/N},\ m = 0, \ldots, N-1\}$. Так как корни x_m обладают симметрией $x_m = -x_{m+N/2}$, то $x^2_{m+N/2} = x^2_m$, т.е. $\{x_m\} \supset \{x^2_m\} \supset \{x^4_m\} \supset \ldots \supset \{x^{N/2}_m\} \supset \{x^N_m\} = \{1\}$, $m = 0, \ldots N-1$. При этом на каждом шаге число различных элементов в множестве уменьшается в два раза. Представим $P_N(x_m) = P^e_{N/2}(x^2_m) + x_m P^o_{N/2}(x^2_m)$, где

$P^e_{N/2}(x) = f_0 + f_2 x + \ldots + f_{N-2} x^{N/2-1}$,

$P^o_{N/2}(x) = f_1 + f_3 x + \ldots + f_{N-1} x^{N/2-1}$.

Таким образом, вычисление полинома $P_N(x)$ степени $(N-1)$ в N точках сведено к вычислению двух полиномов $P^{e,o}_{N/2}(x)$ степени $(N/2-1)$ в $N/2$ точках и последующему сложению $P^e_{N/2}(x^2_m) + x_m P^o_{N/2}(x^2_m)$. Так как $P^e_{N/2}(x)$ и $P^o_{N/2}(x)$ имеют ту же структуру, что и $P_N(x)$, следовательно, к каждому из них можно рекурсивно применить аналогичную процедуру разложения. Найдем вычислительную сложность алгоритма. Пусть $D(N)$ — число арифметических действий, требуемых для вычисления $P_N(x_m)$ для N точек. Тогда вычисляем $P^{o,e}_{N/2}(x^2_m)$ за $2D(N/2)$ арифметических действий; вычисляем $P_N(x_m)$ с учетом $x_m = -x_{m+N/2}$ в виде $P^e_{N/2}(x^2_m) \pm x_m P^o_{N/2}(x^2_m)$, $m = 0, \ldots, \frac{N}{2} - 1$, за $N/2 + (N/2 + N/2)$ арифметических действий; в итоге имеем:

$$D(N) = 2D(N/2) + 3N/2.$$

При этом $D(2) = 3$, $D(1) = 0$. Покажем по индукции, что отсюда следует аналитическая формула $D(N) = \frac{3N}{2}\log_2 N$. Действительно, для $N = 2$ формула верна. Пусть верна для $N/2$, т.е. $D(N/2) = \frac{3N}{4}\log_2(N/2)$. Тогда

$$D(N) = 2(\tfrac{3N}{4}\log_2 N - \tfrac{3N}{4}) + 3N/2 = \tfrac{3N}{2}\log_2 N.$$

При практической реализации для ускорения работы программы рекурсия заменяется на вложенные циклы, а коэффициенты $\{f_j\}$ полинома $P_N(x)$ специальным образом упорядочиваются. Отметим, что имеются эффективные обобщения данного алгоритма для $N = 2^{n_1} \cdot 3^{n_2} \cdot 5^{n_3}$..., если число простых множителей в разложении не слишком велико.

Замечание. Пусть $m_1 = m_2 + mN$, тогда $e^{\frac{2\pi\mathbf{i}}{N} m_1 j} = e^{\frac{2\pi\mathbf{i}}{N} m_2 j}$, т.е. гармоники m_1 и m_2 на сетке с шагом $1/N$ совпадают (по сути это соответствует работе стробоскопа — кратные гармоники видны одинаково). Отсюда имеем, что низкочастотная функция $f(x) = e^{-2\pi\mathbf{i}x}$ на сетке видна как высокочастотная $f(x) = e^{2\pi\mathbf{i}(N-1)x}$. В связи с этим разумно удается восстановить коэффициенты только при гармониках до частоты Найквиста $m \le \frac{N}{2}$. Поэтому для повышения качества численной интерполяции непрерывной функции $f(x)$ по дискретному набору значений $\{f(x_j), x_j = \frac{j}{N}, j = 0, 1, \ldots, N-1\}$ лучше выбирать дискретный базис $\{g^{(m)}_j = e^{2\pi\mathbf{i}m x_j}, m = 0, \pm 1, \ldots, \pm(N/2-1), N/2\}$. В этом случае в узлах сетки дискретная функция восстанавливается без изменений, однако вне узлов интерполяции ее непрерывный аналог обычно обеспечивает существенно лучшее качество приближения для $f(x)$.

Интерполяция Паде–Якоби. Для решения многих прикладных задач интерполяции успешно применяется следующий подход. Будем искать представление функции $f(x)$ в окрестности точки x_0 в виде отношения двух полиномов. Для этого построим отрезок ряда Тейлора в точке x_0: $f(x) = \sum_{i=0}^{\infty} c_i z^i$, $z = (x - x_0)$ и найдем коэффициенты искомого приближения из условия

$$\textstyle\sum_{i=0}^{\infty} c_i z^i = \frac{a_0 + a_1 z + \ldots + a_p z^p}{b_0 + b_1 z + \ldots + b_q z^q} + O(z^{p+q+1}),\ b_0 = 1.$$

Это означает, что первые $p + q + 1$ слагаемые ряда Тейлора для построенной дробно-рациональной функции совпадают с первыми слагаемыми ряда для $f(x)$. Умножим данное равенство на знаменатель и приравняем коэффициенты при $1, z, \ldots, z^{p+q}$. В результате получим систему линейных уравнений для $\{b_0 = 1, b_1, \ldots, b_q\}$, решив которую можно вычислить коэффициенты $\{a_i\}$.

Пример. Для $e^x \approx \frac{12+6x+x^2}{12-6x+x^2}$, $p = q = 2$, при $|x| \le 1$ точность приближения порядка $1/6! \approx 0.0014$.

Многоточечная интерполяция Паде (Рациональная интерполяция). Рассмотрим задачу построения

$$R(x) = \tfrac{a_0 + a_1 x + \ldots + a_p x^p}{b_0 + b_1 x + \ldots + b_q x^q}$$

из условий $R(x_i) = f(x_i)$, $b_0 = 1$, $i = 1, \ldots, n$ при $n = p+q+1$. Данные равенства образуют систему линейных алгебраических уравнений относительно неизвестных коэффициентов:

$$b_0 = 1,\ \sum_{j=0}^{p} a_j x_i^j - f(x_i) \sum_{j=0}^{q} b_j x_i^j = 0, \quad i = 1, \ldots, n.$$

Ее решение может быть получено численно стандартными методами.

Если же $p = q$ (т.е. $n = 2q+1$) или $p = q+1$ (т.е. $n = 2q+2$), то $R(x)$ удается выписать в явном виде методом *обратных разностей Тиле*. Определим

$$f^-(x_1) = f(x_1),\ f^-(x_1; x_2) = \frac{x_2 - x_1}{f^-(x_2) - f^-(x_1)},$$

и далее для $k > 2$ рекуррентно:

$$f^-(x_1; \ldots; x_k) = \frac{x_k - x_1}{f^-(x_2; \ldots; x_k) - f^-(x_1; \ldots; x_{k-1})} + f^-(x_2; \ldots; x_{k-1}).$$

Докажем, что $R(x) =$

$$f(x_1) + \cfrac{x - x_1}{f^-(x_1; x_2) + \cfrac{x - x_2}{f^-(x_1; x_2; x_3) - f^-(x_2) + \ldots + \cfrac{x - x_{n-1}}{f^-(x_1; \ldots; x_n) - f^-(x_2; \ldots; x_{n-1})}}}.$$

Действительно, явно проверяется, что

$$f(x) = f(x_1) + \cfrac{x - x_1}{f^-(x_1; x_2) + \cfrac{x - x_2}{f^-(x; x_1; x_2) - f^-(x_1)}}.$$

Далее для $k \geq 2$ имеем
$f^-(x; x_1; \ldots; x_{k+1}) = \dfrac{x_{k+1} - x}{f^-(x_1; \ldots; x_{k+1}) - f^-(x; x_1; \ldots; x_k)} + f^-(x_1; \ldots; x_k).$
Отсюда выражаем $f^-(x; x_1; \ldots; x_k)$, подставляем в предыдущее тождество и по индукции получаем

$$f(x) = f(x_1) + \cfrac{x - x_1}{f^-(x_1; x_2) + \cfrac{x - x_2}{f^-(x_1; x_2; x_3) - f^-(x_2) + \ldots}}$$

$$\cfrac{}{\ldots + \cfrac{x - x_{n-1}}{f^-(x_1; \ldots; x_n) - f^-(x_2; \ldots; x_{n-1}) + \frac{x - x_n}{f^-(x; x_1; \ldots; x_n) - f^-(x_1; \ldots; x_{n-1})}}}.$$

Из данного представления для функции $f(x)$ следует, что $R(x_i) = f(x_i)$ при $i = n, n-1, \ldots, 2, 1$, т.к. в указанных точках x_i обнуляются последние слагаемые цепной дроби как для $f(x)$, так и для $R(x)$. Утверждение доказано.

Применение дробно-рациональной интерполяции по подходящим образом выбранным узлам для функций с нерегулярным характером поведения часто оказывается целесообразнее тригонометрической интерполяции и интерполяции многочленами.

5. Численное интегрирование

Лекция 22. Численное интегрирование

Рассмотрим интеграл вида $I(f) = \int\limits_a^b p(x)\, f(x)\, dx$, где $[a, b]$ — конечный или бесконечный промежуток числовой оси, и $f(x)$ — произвольная функция из некоторого класса F. Заданная функция $p(x)$ называется *весовой*. Будем предполагать, что на $[a, b]$ функция $p(x)$ измерима, тождественно не равна нулю, и ее произведение на любую $f(x) \in F$ суммируемо.

Для приближенного вычисления интеграла $I(f)$ строятся линейные квадратурные формулы (*квадратуры*) следующего вида: $S_n(f) = \sum\limits_{i=1}^{n} c_i\, f(x_i)$. Постоянные c_i называются *коэффициентами* (*весами*) квадратуры, а x_i — ее *узлами*.

Для каждой функции $f(x) \in F$ погрешность квадратурной формулы $S_n(f)$ определяется как $R_n(f) = I(f) - S_n(f)$. При этом оценкой погрешности на классе F называют величину $R_n(F) = \sup_{f \in F} \big|R_n(f)\big|$. На практике часто используют оценки сверху для $|R_n(f)|$, которые будем обозначать через R_n.

Метод неопределенных коэффициентов.
Дано: $n, \{x_1, \ldots, x_n\}, p(x), [a, b]$.
Найти: набор $\{c_i\}$ из условия точности квадратуры для многочленов наиболее высокой степени, т.е. $I(P_k) = S_n(P_k)$, $P_k(x) = x^k$, $k = 0, 1, \ldots n-1$.

Постановка задачи корректна, т.к. система

$$\begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \\ . & \ldots & . \\ x_1^{n-1} & \ldots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ . \\ c_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ . \\ b_n \end{pmatrix}, \quad b_i = \int\limits_a^b p(x) x^{i-1} dx$$

имеет и при том единственное решение $c_1, \ldots, c_n$ в случае различных узлов.

Пример 22.1. Пусть $p(x) \equiv 1$. Тогда можно получить следующие квадратуры:

$$\int\limits_a^b f(x)\, dx \approx \begin{cases} S_1(f) = (b-a) f(a), \\ S_1(f) = (b-a) f(b), \\ S_1(f) = (b-a)\, f(\frac{a+b}{2}), \\ S_2(f) = \frac{b-a}{2}(f(a) + f(b)), \\ S_3(f) = \frac{b-a}{6}\big(f(a) + 4f(\frac{a+b}{2}) + f(b)\big) \end{cases}$$

(*формулы прямоугольников (по крайним и центральной точкам), формула трапеций, формула Симпсона (парабол)*).

Оценка погрешности. Нас интересует величина $|R_n(f)| = |I(f) - S_n(f)|$. Пусть квадратура точна для многочленов некоторой степени m. Тогда $R_n(f) = R_n(f - P_m)$ для всякого многочлена P_m. Обозначим $g(x) = f(x) - P_m(x)$, будем иметь

$$|R_n(g)| = \Big|\int_a^b g(x)p(x)dx - \sum_{i=1}^n c_i g(x_i)\Big| \le \|g\|\Big(\int_a^b |p(x)|dx + \sum_{i=1}^n |c_i|\Big),$$

где $\|g\| = \max\limits_{[a,b]} |g(x)|$. Далее также используется равномерная норма.

Данная оценка верна для произвольного $P_m(x)$. Возьмем в качестве $P_m(x)$ интерполяционный многочлен $L_{m+1}(x)$, построенный по нулям многочлена Чебышёва, тогда

$$\|g\| = \|f - P_m\| \le \frac{\|f^{(m+1)}\|}{(m+1)!}\frac{(b-a)^{m+1}}{2^{m+1}}2^{1-m-1}.$$

Таким образом,

$$|R_n(f)| = |R_n(g)| \le \frac{\|f^{(m+1)}\|}{(m+1)!}\frac{(b-a)^{m+1}}{2^{2m+1}}\Big(\int_a^b |p(x)|dx + \sum_{i=1}^n |c_i|\Big).$$

Пусть $p(x) > 0$ п.в., а все $c_i > 0$. Тогда

$$\sum_{i=1}^n |c_i| = \sum_{i=1}^n c_i = S(1) = I(1) = \int_a^b p(x)dx.$$

В этом случае

$$R_n = \frac{\|f^{(m+1)}\|}{(m+1)!}\frac{(b-a)^{m+1}}{2^{2m+1}}2\int_a^b p(x)dx.$$

Если к тому же $p(x) \equiv 1$, то $R_n = \dfrac{\|f^{(m+1)}\|}{(m+1)!}\dfrac{(b-a)^{m+2}}{2^{2m}}$.

Пример 22.2. Для формул трапеций и прямоугольников по центральной точке $m = 1$, следовательно,

$$R_1 = \frac{(b-a)^3}{8}\|f''\|, \quad R_2 = \frac{(b-a)^3}{8}\|f''\|.$$

Для формулы Симпсона $m = 3$, т.е. $R_3 = \dfrac{(b-a)^5}{1536}\|f^{(4)}\|$.

Интерполяционные квадратуры. Имеется большая группа квадратурных формул, построенных на основе замены $f(x)$ алгебраическим интерполяционным многочленом.

Дано: $n, \{x_1, \dots, x_n\}, p(x), [a, b]$.

Найти: набор $\{\tilde{c}_i\}$, заменив подынтегральную функцию $f(x)$ многочленом Лагранжа $L_n(x)$ степени $n-1$, построенным по точкам x_i

$$L_n(x) = \sum_{i=1}^n f(x_i)\,\Phi_i(x), \quad \Phi_i(x) = \prod_{\substack{j=1\\ j\ne i}}^n \frac{x - x_j}{x_i - x_j}\,.$$

Положим $\tilde{S}_n(f) = \int\limits_a^b p(x)\,L_n(x)\,dx$. Отсюда получаем явные формулы для набора коэффициентов:

$$\tilde{c}_i = \int_a^b p(x)\,\Phi_i(x)\,dx,\ i = 1, \dots, n.$$

Квадратурные формулы интерполяционного типа, построенные в случае весовой функции $p(x) \equiv 1$ для системы равноотстоящих узлов, называются *формулами Ньютона–Котеса.*

Лемма 22.1 *Набор $\{c_i\}$, найденный методом неопределенных коэффициентов, совпадает с $\{\tilde{c}_i\}$.*

Доказательство. Пусть квадратура $S_n(f)$ с набором $\{c_i\}$ точна для многочленов степени $n-1$. Тогда по определению $\tilde{c}_i = \int\limits_a^b p(x)\,\Phi_i(x) = I(\Phi_i) =$ (т.к. $\deg \Phi_i(x) \le n-1$) $= S_n(\Phi_i) = \sum\limits_{j=1}^n c_j \Phi_i(x_j) = c_i$. Утверждение доказано.

При вычислении коэффициентов квадратуры на отрезке $[a, b]$ полезно сделать замену переменных $x = \dfrac{a+b}{2} + \dfrac{b-a}{2}t$ и перейти к $[-1, 1]$:

$$c_i = \int_a^b p(x) \prod_{\substack{j=1\\ j\ne i}}^n \frac{x - x_j}{x_i - x_j}\,dx = \left(\frac{b-a}{2}\right)\int_{-1}^1 p^0(t) \prod_{\substack{j=1\\ j\ne i}}^n \frac{t - t_j}{t_i - t_j}\,dt.$$

Здесь

$$p^0(t) = p\left(\frac{a+b}{2} + \frac{b-a}{2}t\right),\ x_i = \frac{a+b}{2} + \frac{b-a}{2}t_i,\ i = 1, \dots, n.$$

Оценка погрешности интерполяционных квадратур. Так как выполняется тождество $S_n(f) = I(L_n)$, то

$$R_n(f) = I(f) - S_n(f) = I(f) - I(L_n) = \int\limits_a^b p(x)\big(f(x) - L_n(x)\big)dx.$$

Далее, для произвольной точки x имеем оценку

$|f(x) - L_n(x)| \le \frac{\|f^{(n)}\|}{n!}|\omega_n(x)|$, следовательно,

$$R_n(f) \le R_n = \frac{\|f^{(n)}\|}{n!}\int\limits_a^b |p(x)\omega_n(x)|dx \le \frac{\|f^{(n)}\|}{n!}\left(\frac{b-a}{2}\right)^{n+1}\int\limits_{-1}^1 |p^0(t)\omega_n^0(t)|dt,$$

где $\omega^0(t) = (t-t_1)\dots(t-t_n)$.

Пример 22.3. Для формулы прямоугольников по центральной точке имеем $n = 1, p(x) \equiv 1$, следовательно,

$$R_1 = \|f'\|\frac{(b-a)^2}{4}\int\limits_{-1}^1 |t|dt = \|f'\|\frac{(b-a)^2}{4}.$$

Данную оценку можно улучшить. Действительно,

$$\int\limits_a^b f(x)dx = \int\limits_a^b \left(f(\frac{a+b}{2}) + f'(\frac{a+b}{2})(x - \frac{a+b}{2}) + \frac{f''(\xi)}{2}(x - \frac{a+b}{2})^2\right)dx\,.$$

Отсюда следует, что $R_1 = \|f''\|\frac{(b-a)^3}{24}$.

Формула трапеций: $n = 2, p(x) \equiv 1$, следовательно,

$$R_2 = \|f''\|\frac{(b-a)^3}{8}\frac{1}{2}\int\limits_{-1}^1 |t^2 - 1|dt = \|f''\|\frac{(b-a)^3}{12}\,.$$

Формула парабол (Симпсона): $n = 3, p(x) \equiv 1$, следовательно,

$$R_3 = \left\|f^{(3)}(x)\right\|\frac{(b-a)^4}{192}\,.$$

Теорема 22.1 *Пусть $p(x)$ — четная относительно середины отрезка $[a,b]$ функция, и узлы x_j расположены симметрично относительно $\frac{a+b}{2}$, т.е. $\frac{a+b}{2} - x_j = x_{n+1-j} - \frac{a+b}{2}$. Тогда $c_i = c_{n+1-i}$.*

Доказательство.

$$c_i = (\frac{b-a}{2})\int\limits_{-1}^1 p^0(t)\prod_{\substack{j=1\\ j\ne i}}^n \frac{t-t_j}{t_i-t_j}\,dt,$$

$$c_{n+1-i} = (\frac{b-a}{2})\int\limits_{-1}^1 p^0(t)\prod_{\substack{j=1\\ j\ne n+1-i}}^n \frac{t-t_j}{t_{n+1-i}-t_j}\,dt =$$

(так как $t_{n+1-i} = -t_i$, $t_j = -t_{n+1-j}$)

$$= (\frac{b-a}{2})\int\limits_{-1}^1 p^0(t)\prod_{\substack{j=1\\ j\ne n+1-i}}^n \frac{t+t_{n+1-j}}{-(t_i-t_{n+1-j})}\,dt =$$

(заменим $j \ne n+1-i$ на $n+1-j \ne i$ и переобозначим $n+1-j := j$)

$$= (\frac{b-a}{2})\int\limits_{-1}^1 p^0(t)\prod_{\substack{j=1\\ j\ne i}}^n \frac{-t-t_j}{(t_i-t_j)}\,dt = c_i,$$

так как $\int\limits_{-1}^1 g(t)dt = \int\limits_{-1}^1 g(-t)dt$. Теорема доказана.

Следствие 22.1 *Пусть узлы квадратуры и вес $p(x)$ симметричны относительно середины отрезка. Тогда интерполяционная квадратура с числом узлов $2q-1$ точна для многочленов степени $2q-1$.*

Доказательство. Для произвольного полинома имеем разложение $P_{2q-1}(x) = a_{2q-1}(x - \frac{a+b}{2})^{2q-1} + Q_{2n-2}(x)$. Интеграл и квадратура для первого слагаемого равны нулю, а для второго совпадают по построению.

Составные квадратурные формулы. Пусть $h = (b-a)/N$ и $x_k = a + kh$, $k = 0, 1, \dots, N$. Введем следующие обозначения:

$$I^{(k)}(f) = \int\limits_{x_k}^{x_{k+1}} p(x)f(x)\,dx,\ S_n^{(k)}(f) = S_n^{[x_k, x_{k+1}]}(f),\ k = 0, \dots, N-1.$$

Тогда $I(f) = \sum\limits_{k=0}^{N-1} I^{(k)}(f)$, и для его вычисления можно применить *составную* квадратурную формулу $S_n^N(f) = \sum\limits_{k=0}^{N-1} S_n^{(k)}(f)$ с оценкой погрешности $|R_n^N(f)| \le \sum\limits_{k=0}^{N-1} |R_n^{(k)}(f)|$.

Например, для составной формулы прямоугольников

$$S_1^N(f) = \frac{b-a}{N} \sum_{k=0}^{N-1} f\left(x_k + \frac{h}{2}\right)$$

для погрешности на отрезке $[x_k, x_{k+1}]$ имеем неравенство

$$\left|R_1^{(k)}(f)\right| \le \|f''\|_{[x_k,x_{k+1}]} \frac{(x_{k+1}-x_k)^3}{24} = \|f''\|_{[x_k,x_{k+1}]} \frac{(b-a)^3}{24\,N^3},$$

следовательно, для всего отрезка $[a,b]$: $R_1^N = \|f''\|_{[a,b]} \dfrac{(b-a)^3}{24\,N^2}$.

Лекция 23. Квадратуры Гаусса. Ортогональные многочлены

Для приближенного вычисления интеграла

$$I(f) = \int_a^b p(x)\, f(x)\, dx\,, \quad p(x) > 0 \text{ п.в.}$$

построим при заданном n квадратурную формулу вида

$$S_n(f) = \sum_{i=1}^{n} c_i\, f(x_i),$$

точную для многочленов наиболее высокой степени. Такие квадратуры называются квадратурами Гаусса. Таким образом, имеем следующую задачу.

Дано: $p(x), [a,b], n$.

Найти: набор $\{x_i, c_i\}_{i=1}^n$ из условия точности квадратуры для многочленов наиболее высокой степени, т.е. $I(P_k) = S_n(P_k)$, $P_k(x) = x^k$, $k = 0, 1, \ldots, n-1, \ldots, m$.

Так как число свободных параметров квадратуры равно $2n$, то можно надеяться, что $m = 2n-1$. Однако, соответствующая нелинейная система

весьма нетривиальна для $n \ge 3$. Поэтому, следуя Гауссу, поступают следующим образом.

Ортогональные многочлены. Важную роль при построении квадратурных формул Гаусса играют ортогональные многочлены на отрезке $[a,b]$ с весом $p(x) > 0$ почти всюду.

Пусть H — пространство комплекснозначных функций, определенных на $[a,b]$, с ограниченным интегралом (нормой):

$$(f,g) = \int_a^b f(x)\bar{g}(x)p(x)dx,\ p(x) > 0 \text{ п.в.}, \quad \|f\|^2 = (f,f).$$

Напомним, что система функций $\{\psi_j\}_{j=0}^n$ называется ортогональной, если $(\psi_i, \psi_j) = 0$; называется линейно независимой, если $\sum_{j=0}^n c_j\psi_j = 0$ возможно только при $c_j = 0$, $j = 0, 1, \ldots, n$.

Построение.

1-й метод. Система ортогональных многочленов может быть получена из произвольной линейно независимой системы многочленов $\{\varphi_j\}_{j=0}^n$ в результате стандартной процедуры ортогонализации типа Грама–Шмидта. Например, в качестве исходной можно выбрать систему $\{1, x, \ldots, x^k, \ldots\}$.

2-ой метод. Не сложно доказать, что многочлен $\psi_n(x)$ является ортогональным многочленом тогда и только тогда, когда $\psi_n(x)$ ортогонален произвольному многочлену $P_{n-1}(x)$. Таким образом, коэффициенты $\psi_n(x) = a_0 + a_1x + \ldots + a_{n-1}x^{n-1} + x^n$ можно найти из системы уравнений $(\psi_n, x^j) = 0$, $j = 0, \ldots, n-1$. В данном случае коэффициент $a_n = 1$ выбирается из соображений удобства нормировки.

В практических расчетах наиболее употребительны следующие ортогональные многочлены:

Лежандра $\big([-1,1],\ p(x) = 1\big)$,

Чебышёва первого рода $\left([-1,1],\ p(x) = \dfrac{1}{\sqrt{1-x^2}}\right)$,

Лагерра $\big([0,\infty),\ p(x) = e^{-x}\big)$,

Эрмита $\left((-\infty,\infty),\ p(x) = e^{-x^2}\right)$.

Замечание. Для стандартных систем $\{\psi_n(x)\}$ удается построить явные формулы. Например,

для многочленов Лежандра:

$$\psi_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} \left((x^2-1)^n\right),$$

$$\psi_{n+1}(x) = \frac{2n+1}{n+1} x\psi_n(x) - \frac{n}{n+1}\psi_{n-1}(x);$$

для многочленов Чебышёва:

$$\psi_n(x) = \cos(n \arccos x),$$

$$\psi_{n+1}(x) = 2x\psi_n(x) - \psi_{n-1}(x).$$

Теорема 23.1 *Ортогональный многочлен степени n имеет ровно n различных корней на отрезке $[a, b]$.*

Доказательство. Если $\psi_n(x)$ имеет на $[a, b]$ только $r < n$ нулей нечетной кратности, то многочлен $Q_{n+r}(x) = \psi_n(x) \prod_{l=1}^{r} (x - x_l)$ не меняет знака на этом отрезке, что противоречит свойству $\int_a^b p(x)\psi_n(x) \prod_{l=1}^{r} (x - x_l)\, dx = 0$ ортогональности $\psi_n(x)$ всем многочленам низшей степени.

Теорема 23.2 *Среди многочленов вида $P_n(x) = x^n + \ldots$ минимальную норму $\|P_n\|^2 = \int_a^b p(x)P_n^2(x)\, dx$ имеет ортогональный многочлен $\psi_n(x)$ со старшим коэффициентом* 1.

Доказательство. Пусть $P_n(x)$ — произвольный многочлен степени n со старшим коэффициентом 1. Тогда $P_n(x) = \psi_n(x) + r_{n-1}(x)$, и из ортогональности $\psi_n(x)$ любому многочлену низшей степени следует

$$\big\|P_n(x)\big\|^2 = \big\|\psi_n(x)\big\|^2 + \big\|r_{n-1}(x)\big\|^2 .$$

Теорема 23.3 *Для ортогональных многочленов вида $\psi_n(x) = x^n + \ldots$ справедливы рекуррентные соотношения*

$$\psi_n(x) = (x - b_n)\psi_{n-1}(x) - c_n\psi_{n-2}(x),\ n = 2, 3, \ldots ,$$

$$b_n = \frac{(x\psi_{n-1}, \psi_{n-1})}{(\psi_{n-1}, \psi_{n-1})}, \quad c_n = \frac{(\psi_{n-1}, \psi_{n-1})}{(\psi_{n-2}, \psi_{n-2})} > 0.$$

Доказательство. Представим многочлен $x\psi_{n-1}(x)$ в виде: $x\psi_{n-1} = \sum_{k=0}^{n} \alpha_k\psi_k$. Определим коэффициенты α_j из условий ортогональности

$$(x\psi_{n-1}, \psi_j) = \alpha_j(\psi_j, \psi_j)\,.$$

При $j < n - 2$ имеем

$$(x\psi_{n-1}, \psi_j) = (\psi_{n-1}, x\psi_j) = \big(\psi_{n-1}, Q_{j+1}\big) = 0\,,$$

т.е. все $\alpha_j = 0$ при $j < n - 2$ (здесь $Q_{j+1} = x\psi_j(x)$ обозначает многочлен степени $j + 1$). Таким образом,

$$x\psi_{n-1} = \alpha_n\psi_n + \alpha_{n-1}\psi_{n-1} + \alpha_{n-2}\psi_{n-2}\,,$$

при этом $\alpha_n = 1$ в силу равенства единице коэффициентов при старшей степени x. Отсюда следует, что

$$\psi_n(x) = (x - \alpha_{n-1})\psi_{n-1} - \alpha_{n-2}\psi_{n-2}.$$

И так как $(\psi_n, \psi_{n-1}) = (\psi_{n-2}, \psi_{n-1}) = 0$, то

$$b_n := \alpha_{n-1} = \frac{(x\psi_{n-1}, \psi_{n-1})}{(\psi_{n-1}, \psi_{n-1})}.$$

Далее $(\psi_n, \psi_{n-2}) = 0$, т.е.

$$0 = (x\psi_{n-1}, \psi_{n-2}) - \alpha_{n-1}(\psi_{n-1}, \psi_{n-2}) - \alpha_{n-2}(\psi_{n-2}, \psi_{n-2}).$$

Поскольку $x\psi_{n-2} = \psi_{n-1} + \sum_{i=0}^{n-2} \beta_i\psi_i$, то $(x\psi_{n-1}, \psi_{n-2}) = (\psi_{n-1}, x\psi_{n-2}) = (\psi_{n-1}, \psi_{n-1})$. Таким образом,

$$c_n := \alpha_{n-2} = \frac{(\psi_{n-1}, \psi_{n-1})}{(\psi_{n-2}, \psi_{n-2})} > 0\,.$$

Лемма 23.1 *Ортогональные многочлены на симметричном относительно нуля отрезке с четным весом $p(x)$ обладают свойством $\psi_n(-x) = (-1)^n\psi_n(x)$. При этом $b_n = 0$ для всех n.*

Доказательство. Проверим базу индукции. Так как $\psi_0(x) = 1$, $\psi_1(x) = x$, то из четности $p(x)$ и полученной формулы для b_n следует, что $b_2 = 0$. Шаг индукции: пусть утверждение леммы верно для полиномов степени до $(n - 1)$ и коэффициентов до b_n включительно. Тогда из рекуррентной формулы для ψ_n получаем четность/нечетность $\psi_n(x)$ и, как следствие, из формулы для коэффициентов равенство $b_{n+1} = 0$.

Теорема 23.4 *Нули ортогональных многочленов с фиксированным на отрезке $[a, b]$ весом $p(x) > 0$ перемежаются, т.е.*

$$a < x_1^{(n)} < x_1^{(n-1)} < \ldots < x_{n-1}^{(n-1)} < x_n^{(n)} < b.$$

Доказательство. Будем строить доказательство по индукции. База индукции. Проверим для $\psi_0(x), \psi_1(x), \psi_2(x)$. Имеем: $\psi_0 = 1$, $\psi_1 = x - x_1^{(1)}$. Так

как $\psi_2(x) = x^2 + \ldots$ и $a < x_1^{(1)} < b$, то $\operatorname{sign}\psi_2(b) = 1$. Рассмотрим рекуррентное соотношение для ψ_2 в точке $x = x_1^{(1)}$:

$$\psi_2(x_1^{(1)}) = -c_2\psi_0(x_1^{(1)}) < 0\,,$$

т.к. $c_2 > 0$. Все корни ψ_2 принадлежат отрезку $[a, b]$, поэтому $\operatorname{sign}\psi_2(a) = 1$. Таким образом, имеем перемены знака $\psi_2(x)$ в последовательно расположенных точках $a < x_1^{(1)} < b$, следовательно,

$$a < x_1^{(2)} < x_1^{(1)} < x_2^{(2)} < b.$$

Шаг индукции. Пусть нули перемежаются для ψ_{n-1}, ψ_{n-2}:

$$a < x_1^{(n-1)} < x_1^{(n-2)} < \ldots < x_{n-2}^{(n-2)} < x_{n-1}^{(n-1)} < b.$$

Так как $\psi_n(x) = x^n + \ldots$ и $x_n^{(n)} < b$, то $\operatorname{sign}\psi_n(b) = 1$. Подставим $x = x_{n-1}^{(n-1)}$ в рекуррентное соотношение

$$\psi_n(x) = (x - b_n)\psi_{n-1}(x) - c_n\psi_{n-2}(x).$$

Имеем:

$$\psi_n\left(x_{n-1}^{(n-1)}\right) = -c_n\psi_{n-2}\left(x_{n-1}^{(n-1)}\right).$$

Напомним, что здесь $c_n > 0$. Из условия $x_{n-2}^{(n-2)} < x_{n-1}^{(n-1)} < b$ следует, что $\psi_{n-2}\left(x_{n-1}^{(n-1)}\right) > 0$, таким образом, $\psi_n\left(x_{n-1}^{(n-1)}\right) < 0$. Аналогично показывается, что $\psi_n\left(x_{n-2}^{(n-1)}\right) > 0$. И так далее. Так как многочлен имеет вид $\psi_n(x) = x^n + \ldots$ и $a < x_1^{(n)}$, и все его корни принадлежат отрезку $[a, b]$, то $\operatorname{sign}\psi_n(a) = (-1)^n$. Таким образом, имеем перемены знака $\psi_n(x)$ в последовательно расположенных точках $a, x_1^{(n-1)}, \ldots, x_n^{(n-1)}, b$, что и завершает доказательство.

Замечание. Полиномы $\{x^m\}_{m=0}^{\infty}$ не могут быть ортогональны ни на каком отрезке $[a, b]$ ни с каким весом $p(x) > 0$.

Квадратурные формулы Гаусса. Из полученных ранее оценок следует, что точность квадратуры повышается с повышением степени многочленов для которых квадратура точна. Рассмотрим задачу построения квадратур Гаусса: при заданном числе узлов n построить квадратурную формулу $S_n(f) = \sum_{i=1}^{n} c_i f(x_i)$ для вычисления интегралов вида $I(f) = \int_a^b p(x)f(x)\,dx\,,$ точную для многочленов максимально высокой степени. Весовая функция $p(x)$ здесь предполагается почти всюду положительной.

В этой постановке имеется $2n$ свободных параметров (узлы x_i и коэффициенты c_i неизвестны), поэтому можно попытаться построить квадратуру, точную для многочленов степени $2n - 1$.

Лемма 23.2 *Не существует квадратуры с n узлами, точной для всех многочленов степени $2n$.*

Доказательство. Возьмем $P_{2n}(x) = (x - x_1)^2 \cdots (x - x_n)^2$. Тогда $0 = S_n(P_{2n}) \neq I(P_{2n}) > 0$.

Отметим, что не рекомендуется строить квадратуры Гаусса методом "неопределенных коэффициентов и узлов" из условия точности квадратуры для многочленов наиболее высокой степени. Полученная таким образом система нелинейных уравнений весьма нетривиальна уже при $n = 3$.

При построении квадратурных формул Гаусса ключевым является следующее утверждение.

Теорема 23.5 *Пусть узлы $x_1, \ldots, x_n$ — нули ортогонального многочлена $\psi_n(x)$ степени n на $[a, b]$ с весом $p(x)$, и $S_n(f)$ — квадратура, точная для многочленов степени $n - 1$. Тогда квадратура точна для многочленов степени $2n - 1$.*

Доказательство. Для произвольного многочлена имеем $P_{2n-1}(x) = \psi_n(x)q_{n-1}(x) + r_{n-1}(x) = F(x) + G(x)$. Поэтому

$$I(P_{2n-1}) = \int_a^b P_{2n-1}(x)p(x)dx = \int_a^b G(x)p(x)dx.$$

Равенство нулю интеграла $\int_a^b F(x)p(x)dx$ следует из определения ортогонального многочлена $\psi_n(x)$. Для квадратуры имеем

$$S_n(P_{2n-1}) = S_n(F + G) = S_n(G) = \int_a^b G(x)p(x)dx.$$

Равенство $S_n(F) = 0$ следует из условий $\psi(x_j) = 0$, т.к. $F(x_j) = 0$. В результате имеем равенство $I(P_{2n-1}) = S_n(P_{2n-1})$. Теорема доказана.

Данное утверждение позволяет разбить сам процесс построения квадратуры Гаусса на два последовательных этапа:
— нахождение нулей ортогонального многочлена ψ_n,
— нахождение весов методом неопределенных коэффициентов.

Докажем следующее полезное утверждение.

Лемма 23.3 *Все коэффициенты квадратуры Гаусса положительны.*

Доказательство. Рассмотрим многочлен степени $k = 2n-2$ вида $P_k(x) = \prod\limits_{\substack{i=1\\ i\neq k}}^{n} (x - x_i)^2$. Для интеграла от этого многочлена формула Гаусса дает точный результат:

$$\int\limits_a^b p(x)P_k(x)\,dx = \sum_{j=1}^{n} c_j P_k(x_j) = \sum_{\substack{j=1\\ j\neq k}}^{n} c_j P_k(x_j) + c_k P_k(x_k)\,.$$

Поскольку справедливо $P_k(x_j) = 0$ при $j \neq k$, следовательно, имеем $c_k = \int\limits_a^b p(x)P_k(x)\,dx / P_k(x_k) > 0$. Лемма доказана.

Так как квадратуры Гаусса точны для многочленов степени $m = 2n-1$ и $c_k > 0$, следовательно, верна оценка

$$R_n \le \frac{\|f^{(m+1)}\|}{(m+1)!}\frac{(b-a)^{m+1}}{2^{m+1+m}} 2 \int\limits_a^b p(x)dx.$$

Имеет место следующее полезное утверждение.

Лемма 23.4 *Пусть весовая функция $p(x)$ — четная относительно середины отрезка интегрирования $(a+b)/2$. Тогда узлы квадратуры Гаусса для вычисления $I(f) = \int\limits_a^b p(x)f(x)\,dx$ расположены симметрично относительно $(a+b)/2$, а соответствующие симметричным узлам коэффициенты квадратуры равны.*

Действительно, симметрия узлов квадратуры следует из симметрии (кососимметрии) ортогонального многочлена, а симметрия коэффициентов есть следствие симметрии узлов.

Отметим, что на данный момент имеются готовые таблицы квадратур до 2^{12} с 20-ю знаками для стандартных весовых функций.

Пример 23.1. Для интеграла $\int\limits_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x)dx$ квадратура Гаусса имеет вид

$$S_n(f) = \frac{\pi}{n}\sum_{j=1}^{n} f(x_j), \quad x_j = \cos\frac{2j-1}{2n}\pi.$$

Данную квадратуру также называют квадратурой Эрмита. Соответствующими ортогональными многочленами являются многочлены Чебышёва, поэтому корни $T_n(x)$ задают узлы квадратуры n-го порядка. Чтобы для данного набора коэффициентов квадратура была точна для произвольного многочлена $P_{2n-1}(x)$ достаточно проверить, что квадратура точна для произвольного $P_{n-1}(x)$. Представим $P_{n-1}(x) = \sum\limits_{m=0}^{n-1} a_m T_m(x)$ и установим, что квадратура точна для T_m при $m < n$.

Лекция 24. Оптимизация квадратур

Рассмотрим задачу вычисления интеграла с оптимальными в некотором смысле затратами. Пусть заданы область интегрирования $[a, b]$, вес $p(x)$, класс подынтегральных функций F.

Погрешностью квадратуры $S_N(f) = \sum\limits_{i=1}^{N} c_i f(x_i)$ на классе функций F называют величину $R_N(F) = \sup\limits_{f\in F} |R_N(f)|$, где $R_N(f)$ — погрешность на функции, т.е. $R_N(f) = I(f) - S_N(f)$.

Нижняя грань $W_N(F) = \inf\limits_{c_i, x_i} |R_N(F)|$ называется оптимальной оценкой погрешности квадратуры на классе. Если существует квадратура, для которой $R_N(F) = W_N(F)$, то такую квадратуру называют оптимальной на данном классе.

Пример 24.1. Пусть F: $|f'(x)| \le A,\ x \in [0, 1]$. Можно доказать, что оптимальная квадратура — составная формула прямоугольников по центральной точке:

$$S_N(f) = \frac{1}{N}\sum_{j=1}^{N} f\left(\frac{j-1/2}{N}\right),\ R_N(F) \le \frac{A}{4N}.$$

Если требуется обеспечить точность интегрирования ε, а соответствующее $N \ge \frac{A}{4\varepsilon}$ для нас не достижимо, то можно изменить класс F и либо попытаться построить оптимальную квадратуру (что часто нетривиально), либо для выбранного класса найти асимптотически оптимальную квадратуру, и для нее решить задачу об оптимальном распределении узлов.

Пусть $F : \|f''\|_{C[a_i,b_i]} \le A_i$, $i = 1, \ldots, q$. Представим

$$I(f) = \int\limits_a^b f(x)dx = \sum_{i=1}^{q} \int\limits_{a_i}^{b_i} f(x)dx.$$

На каждом отрезке $[a_i, b_i]$ длины H_i будем применять составную формулу трапеций с шагом $h_i = H_i/N_i$:

$$\int\limits_{a_i}^{b_i} f(x)dx \approx h_i \left(\frac{f(a_i)}{2} + f(a_i + h_i) + \ldots + f(b_i - h_i) + \frac{f(b_i)}{2} \right).$$

При этом

$$|R^{(i)}(f)| \le \frac{\|f''\|_{[a_i,b_i]}}{12} \frac{H_i^3}{N_i^2} \le \frac{A_i H_i^3}{12 N_i^2},\ |R_N(f)| \le R_N = \sum_{i=1}^{q} |R^{(i)}(f)| = \sum_{i=1}^{q} \frac{A_i H_i^3}{12 N_i^2}.$$

Первая постановка оптимизационной задачи.
Дано: q, N.
Найти $\{N_i\}$, $\sum_{i=1}^{q} N_i = N$, из условия минимума погрешности R_N, т.е.

$$\begin{cases} R_N \to \min, \\ \Psi = N_1 + \ldots + N_q - N = 0. \end{cases}$$

Далее будем считать, что N_i — действительные числа (хотя интересуют целые). Тогда для функции Лагранжа $L = R_N + \lambda\Psi$ имеем:

$$\frac{\partial L}{\partial N_i} = -\frac{A_i H_i^3}{6 N_i^3} + \lambda = 0 \Rightarrow N_i = H_i \left(\frac{A_i}{6\lambda}\right)^{1/3},\ i = 1, \ldots, q.$$

Подставляем N_i в условие $\sum_{i=1}^{q} N_i = N$, имеем $\sum\limits_{i=1}^{q} H_i \left(\frac{A_i}{6\lambda}\right)^{1/3} = N$. Из данного выражения находим λ и подставляем в формулы для N_i. Полученные таким образом N_i не целые, поэтому результат округляем. Найденное в итоге распределение узлов несколько отлично от оптимального, но дальнейшее уточнение вряд ли разумно.

Вторая постановка оптимизационной задачи.
Дано: $q, R_N = \varepsilon$.
Найти $\{N_i\}$ из условия минимума числа узлов $N = \sum_{i=1}^{q} N_i$, т.е.

$$\begin{cases} N = N_1 + \ldots + N_q \to \min, \\ \Psi = R_N - \varepsilon = 0. \end{cases}$$

Для функции Лагранжа $L = (N_1 + \ldots + N_q) + \frac{1}{\mu}(R_N - \varepsilon)$, выписанной для удобства в терминах $\mu = \frac{1}{\lambda}$ при действительных N_i, имеем

$$\frac{\partial L}{\partial N_i} = 1 + \frac{1}{\mu}\left(-\frac{A_i H_i^3}{6 N_i^3}\right) = 0 \Rightarrow N_i = H_i \left(\frac{A_i}{6\mu}\right)^{1/3}.$$

Подставим N_i в условие на R_N, получим $\sum\limits_{i=1}^{q} \frac{A_i H_i^3}{12 H_i^2} \left(\frac{6\mu}{A_i}\right)^{\frac{2}{3}} = \varepsilon$. Отсюда следует, что $\mu^{\frac{2}{3}} \sum\limits_{i=1}^{q} H_i \left(\frac{A_i}{48}\right)^{\frac{1}{3}} = \varepsilon$. Из данного выражения находим μ, подставим в формулы для N_i, и результат округлим.

Полученные формулы позволяют сформулировать работу разумного алгоритма: строим набор подотрезков $[a_i, b_i]$ плавного изменения $f''(x)$, на каждом отрезке приближенно оцениваем $A_i = \|f''\|_{[a_i,b_i]}$, находим N_i по указанным формулам и далее вычисляем по составной формуле интеграл. Отметим, что большинство алгоритмов базируется на качественной стороне полученных результатов, а не на количественных соотношениях. В данном случае найденная формула $\frac{1}{12} A_i \frac{H_i^3}{N_i^3} = \frac{\lambda}{2}$ означает, что погрешность по каждому элементарному подотрезку интегрирования одинакова в случае оптимального распределения узлов.

Правило Рунге оценки погрешности. Пусть для подсчета интеграла $I(f) = \int\limits_a^b p(x)\, f(x)\, dx$ применяется некоторая составная формула $S_n^N(f) = \sum\limits_{i=1}^{N} S_n^{(i)}(f)$. Можно доказать, что в случае достаточно гладкой f имеет место следующее представление

$$I(f) = S_n^{(N)}(f) + Ch^m + O\left(h^{m+1}\right)$$

с не зависящей от h константой C. По сути, это соответствует разложению функции ошибки по степеням h.

Обозначим через $S_n^{(2N)}(f)$ составную формулу, полученную применением формулы $S_n^{(N)}(f)$ с вдвое большим числом узлов (половинным шагом). Тогда с той же C находим:

$$I(f) = S_n^{(2N)}(f) + C\left(\frac{h}{2}\right)^m + O\left(h^{m+1}\right).$$

Следовательно, с точностью до членов $O\left(h^{m+1}\right)$ справедливо следующее *правило Рунге*:

$$C \approx \frac{S_n^{(2N)}(f) - S_n^{(N)}(f)}{1 - 2^{-m}},$$
$$I(f) - S_n^{(2N)}(f) \approx \frac{S_n^{(2N)}(f) - S_n^{(N)}(f)}{2^m - 1}.$$

Отсюда следует, что используя значения $S_n^{(N)}$ и $S_n^{(2N)}$ квадратуры с главным членом погрешности Ch^m, можно построить квадратурную формулу

$S_n^{(N,2N)} = S_n^{(2N)} + \dfrac{S_n^{(2N)} - S_n^{(N)}}{2^m - 1}$ порядка точности не ниже $(n+1)$.

При применении данного правила к формуле трапеций получается формула Симпсона, т.к. в этом случае $m = 2$, и на отрезке $[a_i, b_i]$ имеем:

$$S_n^{(N)} = \frac{b_i - a_i}{2}\left(f(a_i) + f(b_i)\right),$$
$$S_n^{(2N)} = \frac{b_i - a_i}{4}\left(f(a_i) + f(\frac{a_i + b_i}{2})\right) + \frac{b_i - a_i}{4}\left(f(\frac{a_i + b_i}{2}) + f(b_i)\right).$$

Формула Симпсона имеет четвертый порядок точности, поэтому в данном случае порядок главного члена погрешности увеличится на 2.

Процедура построения формулы $S_n^{(N,2N)} = S_n^{(2N)} + \dfrac{S_n^{(2N)} - S_n^{(N)}}{2^m - 1}$ является экстраполяционной, т.е. при $S_n^{(N)} \neq S_n^{(2N)}$ величина $S_n^{(N,2N)}$ всегда лежит вне отрезка с концами $S_n^{(N)}$ и $S_n^{(2N)}$. Действительно, если $S_n^{(2N)} > S_n^{(N)}$, то $S_n^{(N,2N)} > S_n^{(2N)} > S_n^{(N)}$. Если $S_n^{(2N)} < S_n^{(N)}$, то $S_n^{(N,2N)} < S_n^{(2N)} < S_n^{(N)}$.

Пусть для вычисления интеграла $I(f)$ от некоторой функции используется квадратурная формула $S_n^{(N)}$, фактический порядок точности m которой неизвестен для данной функции. Рассмотрим *процесс Эйткена* численной оценки значения порядка m, являющийся обобщением правила Рунге. Пусть $I(f)$ — точное значение интеграла. Вычислим значения $S_n^{(N)}$, $S_n^{(2N)}$, $S_n^{(4N)}$. Если учитывать только главный член погрешности, то получаем систему трех уравнений

$$I(f) \approx S_n^{(N)}(f) + Ch^m,$$
$$I(f) \approx S_n^{(2N)}(f) + C\left(\tfrac{h}{2}\right)^m,$$
$$I(f) \approx S_n^{(4N)}(f) + C\left(\tfrac{h}{4}\right)^m$$

с неизвестными $I(f), C$ и m. Формальное решение позволяет найти $I(f)$, C, m. Однако для конкретного N, отброшенные слагаемые $O(h^{m+1})$ могут оказаться порядка Ch^m, что существенно исказит ответ. Поэтому более разумным является следующий подход. Так как $S_n^{(2N)}(f) - S_n^{(N)}(f) \approx C\left(1 - 2^{-m}\right)h^m$, следовательно,

$$F(\ln h^{-1}) = \ln|S_n^{(2N)}(f) - S_n^{(N)}(f)|^{-1} \approx \text{const} + m\ln h^{-1}.$$

Дифференцирование асимптотических равенств вообще говоря незаконно, поэтому величину m можно попытаться найти в результате линейной интерполяции $F(\ln h^{-1})$ по нескольким значениям $h/2^k$ (например, в смысле наименьших квадратов).

Правило Рунге позволяет сформулировать принцип построения программ интегрирования с автоматическим выбором шага, учитывающих поведение интегрируемых функций. Их суть заключается в следующем. Пусть для интегрирования применяется некоторая квадратура $S(f)$. Для приближенного вычисления интеграла на отрезке $[a_i, b_i]$ длины h_i, находятся два значения $S_1 = S^{[a_i,b_i]}$ и $S_2 = S^{[a_i,a_i+h_i/2]} + S^{[a_i+h_i/2,b_i]}$. С их помощью по правилу Рунге оценивается главный член погрешности. Если погрешность мала, тогда значение интеграла считается найденным. В противном случае требуется уменьшить длину отрезка интегрирования h_i.

При нахождении интеграла на большом отрезке $[a, b]$ естественно выделяются две процедуры, которые принято назвать вертикальной и горизонтальной.

В случае горизонтальной процедуры фиксируются два числа $\varepsilon_1 < \varepsilon_2$ (точность) и начальный шаг h_1. На отрезке $[a_1, b_1]$, $a_1 = a$, $b_1 = a_1 + h_1$, находятся значения S_1, S_2, $R = |S_1 - S_2|$.

Если $R \leq \varepsilon_2$, то точность интегрирования на данном отрезке считается удовлетворительной, поэтому переходим к следующему отрезку $[a_2, b_2]$ длины h_2. Если $\varepsilon_1 \leq R \leq \varepsilon_2$, то $h_2 = h_1$, т.е. шаг признается оптимальным, если $R < \varepsilon_1$, то шаг признается излишне мелким и $h_2 = 2h_1$.

Если же $R > \varepsilon_2$, то точность считается недостаточной, поэтому переопределяем $h_1 = h_1/2$, и метод заново применяется для $[a_1, a_1 + h_1]$.

Значение интеграла по всему отрезку $[a, b]$ находится как сумма найденных по подотрезкам интегралов.

В случае вертикальной процедуры также фиксируется точность ε. Для всего отрезка $[a, b]$ находятся значения S_1, S_2, R. Если $R \leq \varepsilon$, то считается, что значение интеграла по всему отрезку найдено. Иначе исходный отрезок делится на два $[a, a + \frac{b-a}{2}]$, $[a + \frac{b-a}{2}, b]$, и метод применяется рекурсивно к каждому из подотрезков.

Несложно привести примеры функций, для которых каждый из методов дает плохой результат (например, исходная функция неотрицательна, а все узлы попадают в нули).

Лекция 25. Метод Монте–Карло. Интегрирование функций с особенностями

Рассмотрим задачу вычисления интеграла в многомерном случае:

$$I(f) = \int\limits_{\Omega} f(\omega)d\omega, \quad \Omega \subset R^d,\ d > 1.$$

Пусть $f \in C^{1,\dots,1}(A,\dots,A)|_\Omega$. Можно доказать, что не существует методов интегрирования на данном классе с оценкой погрешности лучше, чем $|R(f)| \geq \dfrac{cA}{N^{\frac{1}{d}}}$, где d — кратность вычисляемого интеграла. И если необходимо обеспечить точность $|R(f)| \leq \varepsilon = cA10^{-3}$, то $N \geq 10^{3d}$. Поскольку нахождение каждого значения подынтегральной функции обычно требует значительного числа арифметических действий, то уже при $d = 6$ такое условие на N может оказаться невыполнимым. В этом случае для достижения требуемой точности можно попытаться детальнее описать класс F и применить более точные методы.

Рассмотрим иной приближенный способ вычисления значения интеграла, оценка погрешности которого гарантируется только с некоторой степенью достоверности.

Пусть область $\Omega = [0,1]^d$. Рассмотрим случайную величину p, равномерно распределенную в Ω. Построим с.в. $s = f(p)$. Тогда

$$M(s) = \int\limits_\Omega f(p)dp = I(f) = M(f),$$
$$D(s) = M(s^2) - M^2(s) = I(f^2) - I^2(f) = D(f).$$

Пусть выбрано N случайных величин p_j, независимых в совокупности.

Определим $s_j = f(p_j)$ и построим с.в. $S_N = \dfrac{1}{N}\sum\limits_{j=1}^{N} s_j$. Тогда:

$$M(S_N) = \frac{1}{N}\sum_{j=1}^{N} M(s_j) = M(s),$$
$$D(S_N) = \frac{1}{N^2}\Big(\sum_{j=1}^{N} D(s_j) + 2\sum_{i<j} cov(s_i, s_j)\Big) = \frac{1}{N}D(s).$$

Рассмотрим с.в. $Z_N = \dfrac{S_N - M(S_N)}{\sqrt{D(S_N)}}$. Согласно центральной предельной теореме случайная величина Z_N распределена асимптотически нормально с функцией распределения

$$\Phi(y) = \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{y} e^{-\frac{t^2}{2}}dt.$$

При этом

$$P(|Z_N| \leq y) \sim p_0(y) = 1 - \frac{2}{\sqrt{2\pi}}\int\limits_{y}^{\infty} e^{-\frac{t^2}{2}}dt,$$

т.е. вероятность того, что случайная величина с такой функцией распределения не превосходит по модулю значения $y > 0$ асимптотически равна $p_0(y)$.

Таким образом, при больших N с вероятностью близкой к $p_0(y)$ выполняется неравенство:

$$|S_N - I(f)| = |S_N(f) - M(S_N)| \leq y\sqrt{D(S_N)} = y\sqrt{\frac{D(s)}{N}}.$$

Данная оценка позволяет сформулировать правила 3-х сигм: $y = 3$, $p_0(3) \approx 0.997$; и 5-и сигм: $y = 5$, $p_0(5) \approx 0.99999$. В правой части оценки стоит неизвестная величина $D(s) = D(f) = I(f^2) - I^2(f)$, которую можно оценить в процессе вычислений.

Приведем практическую схему применения метода Монте–Карло. Пусть мера Ω равна единице. Пусть ε — требуемая точность, а $1-\eta$ — требуемая мера надежности. Из условия

$$\eta \leq \frac{2}{\sqrt{2\pi}}\int\limits_{y}^{\infty} e^{-\frac{t^2}{2}}dt$$

определим y.

Далее последовательно

для $n = 1, 2, \dots$ строятся случайные точки p_n, находятся

$t_n = t_{n-1} + f(p_n)$, $d_n = d_{n-1} + f^2(p_n)$,

$S_n(f) = t_n/n$, $D_n(f) = d_n/n - S_n^2(f)$,

$\varepsilon_n = y\sqrt{D_n(f)/n}$;

пока $\varepsilon_n > \varepsilon$.

По окончании расчета полагаем, что $|I(f) - S_N(f)| \leq \varepsilon$ с вероятностью $1 - \eta$.

Отметим, что наибольшие проблемы данного метода заключены не в вероятностном подходе вычисления интеграла, а в отсутствии датчиков случайных чисел с требуемыми характеристиками. Формально трудоемкость метода не зависит от размерности d. Однако, типичным для практики является требование малости $|S_N - I(f)|/I(f)$, т.е. малости $\sqrt{D(f)/N}/|I(f)|$. При этом величина $\sqrt{D(f)}/|I(f)|$ быстро растет с увеличением кратности d интеграла, что приводит к возрастанию вычислительных затрат.

Для ускорения сходимости метода Монте–Карло можно попытаться представить подынтегральную функцию в виде суммы $f(\omega) = F(\omega) + G(\omega)$, где F интегрируется явно и содержит все выстро меняющиеся компоненты, а G имеет малую дисперсию. Существенное ускорение можно получить за счет разбиения исходной области на подобласти с малым изменением функции f (например, задача об "Опросе" или интегрирование $\operatorname{sign} x$ на $[-1,1]$). При

наличии информации о поведении подынтегральной функции также можно изменить плотность распределения узлов.

Быстро осциллирующие функции. Пусть требуется вычислить интеграл

$$I(f) = \int_a^b \exp\{\mathbf{i}\omega x\} f(x)dx, \quad \omega(b-a) \gg 1,$$

где $f(x)$ — гладкая функция.

Функции $\mathrm{Re}\,(\exp\{\mathbf{i}\omega x\}f(x))$, $\mathrm{Im}\,(\exp\{\mathbf{i}\omega x\}f(x))$ имеют на рассматриваемом отрезке примерно $\omega(b-a)/\pi$ нулей. Такие функции могут быть хорошо приближены многочленами степени $n \geq \omega(b-a)/\pi$, следовательно, для непосредственного вычисления интеграла $I(f)$ потребуется применение квадратур, точных для многочленов очень высокой степени. Поэтому более выгодным может оказаться использование $\exp\{\mathbf{i}\omega x\}$ в качестве весовой функции.

Зададимся узлами интерполирования $x_j, j = 1, 2, \ldots, n$, построим многочлен Лагранжа $L_n(x)$ и рассмотрим квадратурную формулу, сделав замену переменных $x = \dfrac{b+a}{2} + \dfrac{b-a}{2}t$. Будем иметь:

$$S_n(f) = \int_a^b e^{\mathbf{i}\omega x}\left(\sum_{j=1}^n f(x_j)\Phi_j(x)\right)dx = \frac{b-a}{2}e^{\mathbf{i}\frac{\omega(a+b)}{2}}\sum_{j=1}^n c_j f(x_j),$$

где $c_j = \displaystyle\int_{-1}^1 e^{\mathbf{i}pt}\left(\prod_{k\neq j}\frac{t-t_k}{t_j-t_k}\right)dt, \quad p = \frac{b-a}{2}\omega.$

При этом оценка погрешности

$$|R_n(f)| = |\int_a^b \exp\{\mathbf{i}\omega x\} f^{(n)}(\xi(x))(x-x_1)\ldots(x-x_n)dx| \leq$$

$$\leq C\|f^{(n)}\|_{C[a,b]}(b-a)^{n+1}$$

не зависит от ω.

Приведем расчетную формулу для $n = 2,\ d_1 = -1,\ d_2 = 1$:

$$c_{1,2} = \int_{-1}^1 \frac{1 \mp t}{2}e^{\mathbf{i}pt}dt = \frac{\sin p}{p} \pm \frac{p\cos p - \sin p}{p^2}\mathbf{i}.$$

Отметим, что при малых ω данные формулы могут иметь большую вычислительную погрешность. Действительно, при малых ω величина p мала,

и функции $\cos p$ и $\sin p$ вычисляются с погрешностями $O(2^{-t})$ и $O(p2^{-t})$ соответственно, где t — длина мантиссы. Вследствие этого коэффициенты c_1 и c_2 приобретают погрешность $O(2^{-t})/p$. Таким образом, при малых p лучше построить квадратурную формулу, не выделяя вес.

Вычисление интегралов от функций с особенностями. Существенную часть реально встречающихся подынтегральных функций составляют функции с особенностями, причем особенность может содержаться либо в функции, либо в ее производных. Если нерегулярность функции не вызвана колебательным характером ее поведения, то для вычисления больших серий интегралов такого типа применяется ряд специальных приемов: выделение особенности в весовую функцию, разбиение интеграла на части, аддитивное представление подынтегральной функции, замена переменных и т.д.

1. Выделение весовой функции. Пусть требуется вычислить интеграл $I(f) = \displaystyle\int_a^b f(x)dx$ с нерегулярной функцией $f(x)$. В этом случае применение стандартных методов интегрирования недопустимо, так как оценка погрешности выражается через $f^{(m)}(\xi)$. Пусть $f(x)$ представима в виде $f(x) = p(x)g(x)$, где $p(x)$ содержит особенность, но имеет простой вид, а $g(x)$ — гладкая функция. В этом случае разумно применение квадратурных формул для интегралов с весом.

Пример 25.1. Для интеграла $\displaystyle\int_{-1}^1 \frac{dx}{\sqrt{1-x^4}} = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}}\frac{dx}{\sqrt{1+x^2}}$ с весовой функцией $\dfrac{1}{\sqrt{1-x^2}}$ рассмотренный метод приводит к т.н. квадратуре Эрмита:

$$S_n(f) = \frac{\pi}{n}\sum_{j=1}^n g(x_j); \quad x_j = \cos\frac{2j-1}{2n}\pi,\ j = 1, \ldots, n.$$

Обратим внимание, что соответствующими ортогональными многочленами являются многочлены Чебышёва, и корни $T_n(x)$ задают узлы квадратуры n-го порядка.

2. Явное выделение особенности. Пусть вычисляется интеграл $I(f) = \displaystyle\int_0^1 g(x)x^\alpha dx$, где $\alpha \in (-1, 1)$, $g(x)$ — гладкая функция, $g(0) \neq 0$. Можно выделить x^α в качестве весовой. Можно разбить интеграл на сумму двух по отрезкам $[0, \varepsilon]$ и $[\varepsilon, 1]$. Первый отрезок поделить на k частей и на каждом

заменить $g(x)$ на постоянную функцию $g(x_i)$: $\int\limits_0^{\varepsilon} x^\alpha g(x)\approx\sum\limits_{i=0}^{k} g(x_i)\int\limits_{a_k}^{b_k} x^\alpha dx$.

Погрешность такой замены на каждом подотрезке можно оценить, например, на основе метода Тейлора: $g(x)=g(x_i)+g'(\xi_i)(x-x_i)$. Для $x\in[\varepsilon,1]$ возможно применение стандартных методов интегрирования.

При вычислении интеграла $\int\limits_0^1 \dfrac{\ln x}{1+x^2}dx$ можно представить подынтегральную функцию в виде $f(x)=G(x)+g(x)$, где

$$G(x)=\ln x,\quad g(x)=-\frac{x^2\ln x}{1+x^2},$$

и вычислить $\int\limits_0^1 G(x)dx$ в явном виде, а для $\int\limits_0^1 g(x)dx$ построить квадратуру.

Можно выделить особенность более точно, рассмотрев $G(x)=(1-x^2)\ln x$.

3. Замена переменных. Для рассматриваемой подынтегральной функции $f(x)$ особенность также можно устранить, сделав замену переменных $x=t^k$:

$$\int\limits_0^1 \frac{\ln x}{1+x^2}dx=\int\limits_0^1 \frac{k^2t^{k-1}\ln t}{1+t^{2k}}dt,\quad k>2.$$

В этом случае, увеличивая степень k, можно повысить гладкость интегрируемой функции, следовательно, степень квадратуры. Однако, при увеличении k также растет норма производных полученной функции, что приводит к ухудшению оценки погрешности (узлы при этом сгущаются к нулевой точке).

В некоторых случаях приходится совмещать несколько приемов. Построим квадратурную формулу для вычисления интеграла

$\int\limits_0^\infty f(x)e^x x^\alpha\sin(\omega x)dx$, где $\alpha\in(-1,1)$, $\omega\gg1$, $|f^{(k)}(x)|\le A_k$, $k=0,1,2,\ldots$.

Разобьем отрезок интегрирования на подотрезки $[0,\varepsilon]$, $[\varepsilon,A]$, $[A,\infty[$, где $\varepsilon\approx1/\omega$, $A\gg1$. На первом отрезке в качестве весовой можно взять x^α, на втором — функцию $\sin\omega x$. На третьем отрезке можно, например, сделать замену переменных типа $x=A/t$ и перейти к интегрированию функции с особенностью на конечном отрезке. Либо за счет выбора A обеспечить выполнение условия $|\int\limits_A^\infty g(x)dx|\le\delta$ с требуемой точностью δ и пренебречь значением данного интеграла.

6. Нелинейные уравнения

Лекция 26. Методы решения нелинейных уравнений

Итерационные методы вычисления изолированного (отделенного от других) корня z уравнения $f(x)=0$, как правило, требуют указания какой-либо области D, содержащей этот единственный корень, и алгоритма нахождения очередного приближения x_{n+1} по уже имеющимся x_n, ..., x_{n-k}.

Широко используемые способы отделения корней — графический и табличный — базируются на свойствах гладкости функции; в случае, когда $f(x)$ является алгебраическим полиномом степени n, имеются аналитические подходы.

Если $f(x)$ — непрерывная, то вещественный корень z принадлежит любому отрезку, на концах которого функция имеет значения разных знаков. Деля отрезок пополам, получаем универсальный метод вычисления корня (метод бисекции). Этот подход не требует знания хорошего начального приближения. Если оно имеется, то для гладких функций используются более эффективные методы.

Пусть отыскивается единственный на отрезке $[a,b]$ корень z уравнения $f(x)=0$ в предположении непрерывности функции $f(x)$. Если в его окрестности функция представляется в виде $f(x)=(x-z)^pg(x)$, где p — натуральное число, а $g(x)$ — такая ограниченная функция, что $g(z)\ne0$, то число p называют *кратностью* корня. Если $p=1$, то корень называют *простым*. При нечетном p функция $f(x)$ меняет знак на $[a,b]$, т.е. $f(a)f(b)<0$, а при четном p — нет.

Итерационный метод решения порождает последовательность приближений x_n, которая сходится к корню: $\lim\limits_{n\to\infty}|x_n-z|=0$. Величину $e_n=|x_n-z|$ называют *абсолютной ошибкой* на n-й итерации. Итерационный метод имеет *порядок* m (или *скорость сходимости* m), если m — наибольшее положительное число, для которого существует такая конечная постоянная $q>0$, что

$$\lim_{n\to\infty}\sup\frac{e_{n+1}}{e_n^m}\le q<\infty.$$

Постоянную q называют *константой асимптотической ошибки*, её обычно оценивают через производные функции $f(x)$ в точке $x=z$. При $m=1$ ($q\in(0,1)$) сходимость называется *линейной* (иногда говорят, что в этом случае метод сходится со скоростью геометрической прогрессии, знаменатель которой q), при $1<m<2$ — *сверхлинейной*, при $m=2$ — *квадратичной* и т.д. Из сходимости с порядком $m>1$ следует оценка $e_{n+1}\le q_ne_n$,

$q_n \to 0$ при $n \to \infty$. При этом $e_{n+1} \le e_0 \prod_{i=0}^{n} q_i$. Иногда скорость сходимости может замедляться при приближении к искомому решению, что соответствует $q_n \to 1$, но $e_n \to 0$ при $n \to \infty$. Таким свойством обладают методы с *полиномиальной* скоростью сходимости $e_{n+1} \le (1 - \alpha e_n^p) e_n$ с некоторыми $p \ge 1$ и $0 < \alpha e_0^p < 1$.

Особое внимание в теории решения нелинейных уравнений уделяется методам со сверхлинейной скоростью сходимости. При практических расчетах традиционно применяют методы с квадратичной сходимостью, так как итерационные процессы более высокого порядка ($m > 2$) обычно требуют серьезного увеличения вычислительных затрат.

Отметим, что при нахождении кратных корней ($p > 1$) для большинства алгоритмов характерно замедление скорости сходимости.

Метод простой итерации. Исходное уравнение $f(x) = 0$ заменяют эквивалентным ему уравнением

$$x = \varphi(x).$$

Эту замену можно сделать положив, например,

$$\varphi(x) = x + \psi(x) f(x),$$

где $\psi(x)$ — произвольная непрерывная знакопостоянная функция.

Выберем некоторое начальное приближение $x_0 \in [a, b]$ к корню z, дальнейшие приближения будем вычислять по формуле

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, 2, \ldots .$$

Определение 26.1 *Отображение $y = \varphi(x)$ метрического пространства H называется сжимающим (экспоненциально), если $\varphi(H) \subset H$ и при некотором $0 < q < 1$ выполнено условие*

$$\rho(\varphi(x_1), \varphi(x_2)) \le q\, \rho(x_1, x_2) \quad \forall x_{1,2}.$$

Здесь $\rho(x_1, x_2)$ — расстояние между точками x_1 и x_2.

Теорема 26.1 *Пусть отображение $y = \varphi(x)$ является сжимающим на полном метрическом пространстве H. Тогда существует единственная точка z такая, что $z = \varphi(z)$. При этом для произвольной $x_0 \in H$ верна оценка: $\rho(\varphi^n(x_0), z) \le q^n \rho(x_0, z)$.*

Доказательство. Покажем, что последовательность $x_n = \varphi^n(x_0)$ является фундаментальной. Действительно,

$$\rho(\varphi^n(x_0), \varphi^{n+m}(x_0)) \le q^n \rho(x_0, \varphi^m(x_0)) \le$$
(с учетом неравенства треугольников для точек $x_0, \varphi(x_0), \ldots, \varphi^m(x_0)$)
$$\le q^n \Big(\rho(x_0, \varphi(x_0)) + \rho(\varphi(x_0), \varphi^2(x_0)) + \ldots + \rho(\varphi^{m-1}(x_0), \varphi^m(x_0)) \Big) \le$$
$$\le q^n \rho(x_0, \varphi(x_0)) \left(1 + q + \ldots + q^{m-1}\right) \le q^n \rho(x_0, \varphi(x_0)) \frac{1}{1-q}.$$

Если взять достаточно большое n, то величина $\rho(\varphi^n(x_0), \varphi^{n+m}(x_0))$ будет сколь угодно малой независимо от m. Это означает фундаментальность последовательности x_n и существование предела z. При этом из непрерывности φ имеем:

$$z = \lim_{n\to\infty} x_n = \lim_{n\to\infty} \varphi(x_{n-1}) = \varphi(\lim_{n\to\infty} x_{n-1}) = \varphi(z).$$

Если z_1, z_2 — две неподвижные точки, то
$\rho(z_1, z_2) = \rho(\varphi(z_1), \varphi(z_2)) \le q\rho(z_1, z_2) \Rightarrow \rho(z_1, z_2) = 0.$
Теорема доказана.

В некоторых случаях скорость сходимости может замедляться при приближении к корню, например:

$$\rho(\varphi(x_1), \varphi(x_2)) \le \rho(x_1, x_2)\left(1 - \alpha\rho^p(x_1, x_2)\right) \le$$
$$\le \frac{\rho(x_1, x_2)}{(1 + \alpha\rho^p(x_1, x_2)} \le \frac{\rho(x_1, x_2)}{(1 + \alpha\rho^p(x_1, x_2))^{1/p}},\ \alpha > 0,\ p \ge 1.$$

Определение 26.2 *Отображение $y = \varphi(x)$ метрического пространства H называется слабо (полиномиально) сжимающим , если $\varphi(H) \subset H$, и при некоторых $\alpha > 0, p \ge 1$ выполнено условие*

$$\rho(\varphi(x_1), \varphi(x_2)) \le \frac{\rho(x_1, x_2)}{(1 + \alpha\rho^p(x_1, x_2))^{1/p}} \quad \forall x_{1,2}.$$

Теорема 26.2 *Пусть отображение $y = \varphi(x)$ является слабо (полиномиально) сжимающим на полном метрическом пространстве H. Тогда существует единственная точка z такая, что $z = \varphi(z)$. При этом для произвольной $x_0 \in H$ верна оценка: $\rho(\varphi^n(x_0), z) \le \rho(x_0, z)\,(1+\alpha n\rho^p(x_0, z))^{-1/p}$.*

Доказательство данного утверждения аналогично доказательству принципа сжимающих отображений. Явно проверяется, что для полиномиально сжимающих отображений имеет место оценка:

$$\rho(\varphi^n(x_1), \varphi^n(x_2)) \le \frac{\rho(x_1, x_2)}{(1 + n\alpha\rho^p(x_1, x_2))^{1/p}}.$$

Отсюда следует, что произвольная последовательность $\varphi^n(x_0)$ для $x_0 \in H$ является фундаментальной. Действительно, имеем $\rho(\varphi^{n+m}(x_0), \varphi^n(x_0)) \le$

$$\frac{\rho\left(\varphi^m(x_0), x_0\right)}{\left(1+n\alpha\rho^p\left(\varphi^m(x_0), x_0\right)\right)^{1/p}} \le \frac{\rho\left(\varphi^m(x_0), x_0\right)}{\left(n\alpha\rho^p\left(\varphi^m(x_0), x_0\right)\right)^{1/p}} \le \frac{1}{(n\alpha)^{1/p}} \le \varepsilon$$

с произвольно малым ε и всех m, если n достаточно велико.

Из фундаментальности последовательности следует, что существует предельная точка $z = \lim\limits_{n\to\infty} \varphi^n(x_0)$. Аналогично случаю экспоненциально сжимающего отображения доказывается, что точка z является единственной неподвижной точкой отображения φ: $\varphi(z) = \varphi(\lim\limits_{n\to\infty} x_n) = \lim\limits_{n\to\infty} x_{n+1} = z$, при этом каждая точка x_0 притягивается к ней с полиномиальной скоростью. Теорема доказана.

Для методов с полиномиальной скоростью сходимости число итераций n, необходимое для достижения ошибки порядка ε, имеет асимптотику $n \approx \varepsilon^{-p}$, что существенно ограничивает их применение для расчетов с высокой точностью.

Отметим, что в случае $H = \mathbf{R^1}$ гарантировать сходимость метода можно при условии, что соответствующие оценки сжимаемости выполняются либо для всех точек $x_{1,2} \in \mathbf{R^1}$, либо для точек $x_{1,2} \in [a, b]$, но дополнительно требуется $\varphi(x_0) \in [a, b]$ для $\forall x_0 \in [a, b]$, т.е. все приближения $\varphi^n(x_0)$ принадлежат отрезку $[a, b]$.

Теорема 26.3 *Пусть уравнение $f(x) \equiv x - \varphi(x) = 0$ имеет корень z, и для его вычисления применяется метод простой итерации $x_{n+1} = \varphi(x_n)$. Пусть функция $\varphi(x) \in C^1$ в открытой окрестности z, и $\left|\varphi'(z)\right| \le q < 1$. Тогда найдется такая окрестность $Q_\delta = [z-\delta, z+\delta]$, что для произвольного $x_0 \in Q_\delta$ метод сходится к z с линейной скоростью.*

Доказательство. Функция $\varphi'(x)$ непрерывна и $|\varphi'(z)| < 1$, следовательно, $|\varphi'(x)| \le \tilde{q} < 1$ в некоторой окрестности Q_δ. Возьмем произвольную точку $x_0 \in Q_\delta$. По теореме Лагранжа имеем $\varphi(x_0) = \varphi(z) + \varphi'(\xi)(x_0 - z)$, где точка $\xi \in Q_\delta$. Так как $z = \varphi(z)$, то $|\varphi(x_0) - z| = |\varphi'(\xi)(x_0 - z)| \le \tilde{q}|x_0 - z| < |x_0 - z|$, т.е. φ отображает отрезок Q_δ в себя и является сжимающим. Теорема доказана.

Следствие 26.1 *Пусть в условиях теоремы $\varphi'(z) = 0$, $\varphi(x) \in C^2$. Тогда для метода $x_{n+1} = \varphi(x_n)$ в окрестности корня z верна квадратичная оценка сходимости*

$$|z - x_{n+1}| \le C(z - x_n)^2.$$

Доказательство. Сходимость метода в окрестности $x_n \in Q_\delta$ корня z следует из теоремы. Оценим скорость сходимости:

$$\begin{aligned} x_{n+1} - z &= \varphi(x_n) - \varphi(z) = \varphi\big(z + (x_n - z)\big) - \varphi(z) = \\ &= \varphi(z) + (x_n - z)\varphi'(z) + \frac{1}{2}(x_n - z)^2\varphi''(\xi_n) - \varphi(z) = \\ &= \frac{\varphi''(\xi_n)}{2}(x_n - z)^2,\ \xi_n \in [x_n,\, z]. \end{aligned}$$

Теорема 26.4 *Пусть на некотором отрезке $Q_\delta = [a-\delta, a+\delta]$ функция $\varphi(x)$ удовлетворяет условию Липшица $|\varphi(x_1) - \varphi(x_2)| \le q|x_1 - x_2|$ с константой $q < 1$, и в точке a выполняется неравенство*

$$|a - \varphi(a)| \le (1-q)\delta.$$

Тогда на отрезке Q_δ уравнение $f(x) \equiv x - \varphi(x) = 0$ имеет единственный корень z, и последовательность $x_n = \varphi(x_{n-1})$ сходится к корню z для произвольного $x_0 \in Q_\delta$.

Доказательство. Пусть $x_0 \in Q_\delta$, т.е. $|a - x_0| \le \delta$. Тогда

$$\begin{aligned} &|\varphi(x_0) - a| = |\varphi(x_0) - \varphi(a) + \varphi(a) - a| \le \\ &\le |\varphi(x_0) - \varphi(a)| + |\varphi(a) - a| \le q|x_0 - a| + (1-q)\delta \le \delta. \end{aligned}$$

Таким образом, функция $\varphi(x)$ отображает Q_δ в себя и является сжимающей с константой q. Применение принципа сжимающих отображений завершает решение.

Теорема 26.5 *Пусть функция $f'(y)$ непрерывна и $|f(x_n)/f'(y)| \le \varepsilon$ для всех $y \in [x_n - \varepsilon, x_n + \varepsilon]$. Тогда для некоторого $z \in [x_n - \varepsilon, x_n + \varepsilon]$ справедливо равенство $f(z) = 0$.*

Доказательство. По теореме Лагранжа имеем $f(x_n + t) = f(x_n) + f'(y)t$, отсюда следует $f(x_n + t)/f'(y) = f(x_n)/f'(y) + t$. Выражение в правой части равенства неотрицательно при $t = \varepsilon$ и неположительно при $t = -\varepsilon$. Из условия следует, что производная функции не меняет знак при $t \in [-\varepsilon, \varepsilon]$, поэтому если оба значения $f(x_n - \varepsilon)$ и $f(x_n + \varepsilon)$ отличны от нуля, то они имеют разные знаки. Теорема доказана.

Замечание. Таким образом в качестве критерия остановки итерационного метода для нахождения простых корней уравнения $f(x) = 0$ условие $|f(x_n)/f'(x_n)| \le \varepsilon$ предпочтительнее, чем условие $|f(x_n)| \le \varepsilon$.

Пример 26.1. Пусть уравнение $f(x) = 0$ имеет корень на интервале (a, b), причем $f(x)$ дифференцируема, а $f'(x)$ знакопостоянна на этом отрезке.

Построим равносильное уравнение вида $x = \varphi(x)$, для которого на $[a,b]$ выполнено условие $\left|\varphi'(x)\right| \le q < 1$.

Для определенности будем считать, что $f'(x) > 0$. Пусть $0 < m \le f'(x) \le M$. Заменим исходное уравнение $f(x) = 0$ равносильным

$$x = \varphi(x), \quad \varphi(x) = x - \lambda f(x), \quad \lambda > 0.$$

Подберем параметр λ так, чтобы на $[a,b]$ выполнялось неравенство

$$0 \le \varphi'(x) = 1 - \lambda f'(x) \le q < 1.$$

При $\lambda = \dfrac{1}{M}$ получаем $q = 1 - \dfrac{m}{M} < 1$.

Метод хорд. Пусть $f(a)f(b) < 0$. Идея метода (его еще называют методом *ложного положения*) состоит в замене кривой $y = f(x)$ хордами, проходящими через концы отрезков, в которых $f(x)$ имеет противоположные знаки. Метод хорд требует, чтобы один конец отрезка, на котором ищется корень, был неподвижен. В качестве неподвижного конца x_0 выбирают конец отрезка, для которого знак $f(x)$ совпадает со знаком второй производной $f''(x)$. Расчетная формула имеет вид

$$x_{n+1} = x_n - \frac{x_n - x_0}{f(x_n) - f(x_0)}\, f(x_n)$$

и может быть получена из уравнения прямой $y = (f(x_n) - f(x_0))/(x_n - x_0)(x - x_n) + f(x_n)$, имеющей нужный угол наклона и проходящей через точку $(x_n, f(x_n))$.

Метод секущих. Пусть x_{n-1} и x_n — последовательные приближения к корню. Заменим функцию $y = f(x)$ на линейную, проходящую через точки $\big(x_{n-1}, f(x_{n-1})\big)$ и $\big(x_n, f(x_n)\big)$. В качестве следующего приближения к корню возьмем точку пересечения этой прямой с осью абсцисс. Расчетная формула принимает вид

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\, f(x_n).$$

Метод парабол. Пусть x_{n-2}, x_{n-1} и x_n — три последовательных приближения к корню. Заменим кривую $y = f(x)$ параболой, проходящей через точки $\big(x_{n-2}, f(x_{n-2})\big)$, $\big(x_{n-1}, f(x_{n-1})\big)$ и $\big(x_n, f(x_n)\big)$. В качестве следующего приближения к корню возьмем ближайшую к x_n точку пересечения этой параболы с осью абсцисс. Этот подход исключительно эффективен для нахождения корней многочлена как с действительными, так и с комплексными коэффициентами.

Теорема 26.6 *Пусть $f(x) \in C^2$, z — простой корень. Тогда найдется такая окрестность $Q_\delta(z)$, что для произвольных $x_0, x_1 \in Q_\delta(z)$ метод хорд сходится с линейной скоростью.*

Доказательство. Представим метод хорд как частный случай метода простой итерации:

$$x = \varphi(x), \quad \varphi(x) = x - \frac{f(x)}{f(x) - f(x_0)}\,(x - x_0).$$

Тогда

$$\varphi'(x) = 1 - \frac{(f'(x)(x - x_0) + f(x))(f(x) - f(x_0)) + f(x)(x - x_0)f'(x)}{(f(x) - f(x_0))^2},$$

$$\varphi'(z) = 1 + \frac{f'(z)}{f(x_0)}\,(z - x_0) =$$

$$= \frac{f(z) + f'(z)(x_0 - z) + \dfrac{f''(\eta)}{2}(x_0 - z)^2 + f'(z)(z - x_0)}{f(z) + f'(\xi)(x_0 - z)} =$$

$$= \frac{(x_0 - z)^2}{2}\,\frac{f''(\eta)}{f'(\xi)(x_0 - z)} = (x_0 - z)\,\frac{f''(\eta)}{2f'(\xi)}, \quad \eta, \xi \in [x_0, z].$$

Для простого корня имеем $f'(z) \ne 0$, поэтому начальное приближение x_0 можно взять в такой окрестности корня, что $\left|\varphi'(z)\right| \le q < 1$. Согласно доказанным ранее результатам отсюда следует, что метод хорд имеет линейную скорость сходимости для произвольных x_0, x_1 из некоторой подокрестности $Q_\delta(z)$.

Утверждение 26.1 *Пусть $f(x) \in C^3$, z — простой корень. Тогда найдется такая окрестность $Q_\delta(z)$, что для произвольных $x_0, x_1 \in Q_\delta(z)$ метод секущих сходится и верна оценка*

$$|x_{n+1} - z| \le A|x_n - z|^m,$$

где $m = \dfrac{(1+\sqrt{5})}{2} \approx 1.618$, $A \sim \left(\dfrac{1}{2}\,\dfrac{f''(z)}{f'(z)}\right)^{1/m}$, т.е. сходится сверхлинейно.

Метод Ньютона. В случае одномерного уравнения формула метода Ньютона имеет вид

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

В данном случае на каждом шаге дуга кривой $y = f(x)$ заменяется на касательную к ней, проведенную в точке $\big(x_n, f(x_n)\big)$: $y - f(x_n) = f'(x_n)(x - x_n)$. Отсюда находим формулу итерационного процесса, положив $y = 0$ и $x = x_{n+1}$. Метод Ньютона соответствует методу простой итерации $\dfrac{x_{n+1} - x_n}{\tau} +$

$f(x_n) = 0$ с оптимальным в некотором смысле параметром τ. Действительно, пусть z — изолированный простой корень и все $x_n \in [a, b]$. Тогда

$$z = z - \tau f(z),$$
$$x_{n+1} = x_n - \tau f(x_n),$$

следовательно, для некоторого ξ имеем

$$z - x_{n+1} = (1 - \tau f'(\xi))(z - x_n).$$

Если $\tau = 1/f'(\xi)$, то $z - x_{n+1} = 0$, т.е. метод сойдется за один шаг. Но ξ не известно, поэтому выбираем $\tau = 1/f'(x_n)$.

Теорема 26.7 *Пусть $f(x) \in C^3$, z — простой корень. Тогда найдется такая окрестность $Q_\delta(z)$, что для произвольного $x_0 \in Q_\delta(z)$ метод Ньютона сходится и верна оценка*

$$|x_{n+1} - z| \le C|x_n - z|^2,$$

т.е. сходится квадратично.

Доказательство. Рассмотрим метод Ньютона как частный случай метода простой итерации, для которого

$$\varphi'(x) = 1 - \frac{\big(f'(x)\big)^2 - f(x)f''(x)}{\big(f'(x)\big)^2} = \frac{f(x)\, f''(x)}{\big(f'(x)\big)^2}, \quad \text{т.е.} \quad \varphi'(z) = 0.$$

Согласно доказанным ранее результатам отсюда следует, что метод Ньютона в некоторой подокрестности $Q_\delta(z)$ имеет квадратичную скорость сходимости.

Теорема 26.8 *Пусть $f(x) \in C^2$, z — корень кратности p. Тогда найдется такая окрестность $Q_\delta(z)$, что для произвольного $x_0 \in Q_\delta(z)$ метод Ньютона сходится и верна оценка*

$$|x_{n+1} - z| \le q|x_n - z|\,, q \approx (p-1)/p,$$

т.е. сходится линейно.

Доказательство. Поступая так же, как и в случае простого корня, получим $x_{n+1} - z = (x_n - z)\,\varphi'(z) + \dfrac{1}{2}\,(x_n - z)^2\,\varphi''(\xi)$, где $\xi \in [x_n,\, z]$. Однако, в случае $p > 1$ в выражении $\varphi'(x) = \dfrac{f(x)\, f''(x)}{\big(f'(x)\big)^2}$ содержится неопределенность "ноль на ноль", так как z является также корнем уравнения $f'(x) = 0$. Оценим $\varphi'(x)$.

Представим $f(x)$ в виде $f(x) = g(x)\,(x - z)^p$, где $g(z) = a \ne 0$. Тогда, вычисляя $f'(x)$ и $f''(x)$, получаем представление для $\varphi'(x)$ в малой окрестности корня

$$\varphi'(x) = \frac{f(x)\, f''(x)}{\big(f'(x)\big)^2} = \frac{a\,(x-z)^p\, ap(p-1)\,(x-z)^{p-2}}{a^2p^2\,(x-z)^{2p-2}} + O(|x-z|).$$

Отсюда следует, что $\varphi'(z) \approx \dfrac{p-1}{p} < 1$, и чем выше кратность корня, тем медленнее сходимость. Теорема доказана.

Пример 26.2. Пусть уравнение $f(x) = 0$ имеет на отрезке $[a, b]$ корень z кратности p, причем $f(x) \in C^3$. Построим модификацию метода Ньютона, имеющую квадратичную скорость сходимости. Требуемую модификацию будем искать в виде

$$x_{n+1} = x_n - \alpha\,\frac{f(x_n)}{f'(x_n)}$$

и подберем параметр α так, чтобы $\varphi'(z) = 0$. Вблизи корня имеем $\varphi'(x) = 1 - \alpha + \alpha\,\dfrac{f(x)f''(x)}{\big(f'(x)\big)^2} = 1 - \alpha + \alpha\,\dfrac{p-1}{p} + O(|x-z|) = \dfrac{p-\alpha}{p} + O(|x-z|)$. Отсюда находим $\alpha = p$.

Лекция 27. Решение систем нелинейных уравнений

Будем считать, что исходная система имеет вид:

$$\begin{cases} f_1(x^1, x^2, ..., x^m) = 0, \\ f_2(x^1, x^2, ..., x^m) = 0, \\ \dots \quad \dots \quad \dots \quad \dots \\ f_m(x^1, x^2, ..., x^m) = 0; \end{cases} \Leftrightarrow \mathbf{F}(\mathbf{x}) = 0,\ \mathbf{x} = (x^1, ..., x^m)^T.$$

Для решения задачи можно попытаться построить некоторое сжимающее отображение $\mathbf{G}(\mathbf{x})$ и применить *метод простой итерации* $\mathbf{x}_{n+1} = \mathbf{G}(\mathbf{x}_n)$. В многомерном случае также возможна следующая модификация метода:

$$\begin{cases} x^1_{n+1} = g_1(x^1_n, x^2_n, ..., x^m_n), \\ x^2_{n+1} = g_2(x^1_{n+1}, x^2_n, ..., x^m_n), \\ \dots \quad \dots \quad \dots \quad \dots \quad \dots \\ x^m_{n+1} = g_m(x^1_{n+1}, ..., x^{m-1}_{n+1}, x^m_n). \end{cases}$$

Если из функции $\mathbf{F}$ удается выделить главную линейную часть, т.е. получить представление $\mathbf{F}(\mathbf{x}) = \mathbf{B}\mathbf{x} + \mathbf{H}(\mathbf{x})$, то можно использовать либо *метод Пикара* $\mathbf{B}\mathbf{x}_{n+1} + \mathbf{H}(\mathbf{x}_n) = 0$, либо его модификацию

$$\mathbf{B}\frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{\tau} + \mathbf{F}(\mathbf{x}_n) = 0.$$

По сути это соответствует некоторому обобщению метода Ньютона.

Обобщением метода Гаусса–Зейделя и метода Якоби на нелинейный случай являются *методы покоординатного спуска*, для которых компоненты очередного приближения $\mathbf{x}_{n+1}$ соответственно определяются при решении одномерных нелинейных уравнений одной из следующих систем:

$$\begin{cases} f_1(x^1_{n+1}, x^2_n, ..., x^m_n) = 0, \\ f_2(x^1_{n+1}, x^2_{n+1}, ..., x^m_n) = 0, \\ \dots \quad \dots \quad \dots \quad \dots \quad \dots \\ f_m(x^1_{n+1}, x^2_{n+1}, ..., x^m_{n+1}) = 0; \end{cases} \qquad \begin{cases} f_1(x^1_{n+1}, x^2_n, ..., x^m_n) = 0, \\ f_2(x^1_n, x^2_{n+1}, ..., x^m_n) = 0, \\ \dots \quad \dots \quad \dots \quad \dots \quad \dots \\ f_m(x^1_n, ..., x^{m-1}_n, x^m_{n+1}) = 0. \end{cases}$$

Нахождение корней системы уравнений бывает удобно свести к нахождению минимума некоторого функционала $\mathbf{\Phi}(\mathbf{x})$, например, для $\mathbf{\Phi}(\mathbf{x}) = \|\mathbf{F}(\mathbf{x})\|^2$. Напомним, что подобным образом строились методы решения линейных систем. Для решения задачи минимизации $\mathbf{\Phi}(\mathbf{x})$ можно рассмотреть следующий нестационарный процесс, называемый *методом установления*:

$$\frac{d}{dt}\mathbf{x}(t) + \operatorname{grad}\mathbf{\Phi}\big(\mathbf{x}(t)\big) = 0.$$

Так как при $\operatorname{grad}\mathbf{\Phi}(\mathbf{x}) \neq 0$ имеем

$$\frac{d}{dt}\mathbf{\Phi}(\mathbf{x}(t)) = \Big(\operatorname{grad}\mathbf{\Phi}(\mathbf{x}), \frac{d\mathbf{x}}{dt}\Big) = -\Big(\operatorname{grad}\mathbf{\Phi}(\mathbf{x}), \operatorname{grad}\mathbf{\Phi}(\mathbf{x})\Big) < 0\,,$$

то для произвольного $x(0)$ вдоль траектории $\mathbf{x}(t)$ значение $\mathbf{\Phi}(\mathbf{x}(t))$ не возрастает, поэтому $\mathbf{x}(t)$ сходится к некоторой точке x вида $\operatorname{grad}\mathbf{\Phi}(\mathbf{x}) = 0$.

Другой возможный нестационарный процесс, решение которого в точке локального минимума также устанавливается, имеет вид:

$$\frac{d^2\mathbf{x}}{dt^2} + \gamma\frac{d\mathbf{x}}{dt} + \operatorname{grad}\mathbf{\Phi}(\mathbf{x}) = 0,\ \gamma > 0.$$

Можно проверить, что в данном случае вдоль траектории не возрастает значение функционала $\dfrac{1}{2}\left(\dfrac{d\mathbf{x}}{dt}, \dfrac{d\mathbf{x}}{dt}\right) + \mathbf{\Phi}(\mathbf{x})$. При численной реализации методов установления операторы производных заменяют соответствующими разностными аналогами.

Рассмотрим случай системы m нелинейных уравнений $\mathbf{F}(\mathbf{x}) = 0$, где $\mathbf{x} = (x^1, \dots, x^m)^T$, $\mathbf{F} = (f_1, \dots, f_m)^T$. Будем предполагать, что отображение $\mathbf{F} : \mathbf{R}^m \to \mathbf{R}^m$ непрерывно дифференцируемым в некоторой окрестности решения $\mathbf{z}$, так что $\mathbf{F}'(\mathbf{x}) = \left[\frac{\partial f_i}{\partial x^j}\right]$. В предположении обратимости этого оператора метод Ньютона можно записать в виде $\mathbf{x}_{n+1} = \mathbf{x}_n - (\mathbf{F}'(\mathbf{x}_n))^{-1}\,\mathbf{F}(\mathbf{x}_n)$.

Теорема 27.1 *(О сходимости метода Ньютона). Обозначим* $\Omega_a = \{\mathbf{x} : \|\mathbf{x} - \mathbf{z}\| < a\}$, *где* $\|\cdot\|$ *— норма в* $\mathbf{R}^m$. *Пусть при некоторых* $a, a_1, a_2 : 0 < a$, $0 \le a_1, a_2 < \infty$, *выполнены условия:*

1) $\|(\mathbf{F}'(\mathbf{x}))^{-1}y\| \le a_1\|y\|$ *при* $\mathbf{x} \in \Omega_a\,,\quad \forall y$;
2) $\|\mathbf{F}(\tilde{\mathbf{x}}) - \mathbf{F}(\mathbf{x}) - \mathbf{F}'(\mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})\| \le a_2\|\tilde{\mathbf{x}} - \mathbf{x}\|^2$ *при* $\tilde{\mathbf{x}}, \mathbf{x} \in \Omega_a\,.$

Обозначим также $c = a_1a_2$, $b = \min\{a, c^{-1}\}$.

При условиях 1), 2) *и* $\mathbf{x}_0 \in \Omega_b$ *итерационный процесс Ньютона сходится с оценкой погрешности*

$$\|\mathbf{x}_n - \mathbf{z}\| \le c^{-1}\left(c\|\mathbf{x}_0 - \mathbf{z}\|\right)^{2^n}\,,$$

т. е. имеет квадратичную скорость сходимости.

Доказательство. Пусть $\mathbf{x}_n \in \Omega_b$. Докажем по индукции, что $\mathbf{x}_{n+1} \in \Omega_b$. Так как $b \le a$, следовательно, точки $\mathbf{z}, \mathbf{x}_n \in \Omega_a$ и по условию 2) имеем

$$\|F(\mathbf{z}) - F(\mathbf{x}_n) - F'(\mathbf{x}_n)(\mathbf{z} - \mathbf{x}_n)\| \le a_2\|\mathbf{z} - \mathbf{x}_n\|^2.$$

Далее $F(\mathbf{z}) = 0$, а из расчетных формул находим

$$F(\mathbf{x}_n) = -F'(\mathbf{x}_n)(\mathbf{x}_{n+1} - \mathbf{x}_n).$$

Отсюда следует, что

$$\|F'(\mathbf{x}_n)(\mathbf{z} - \mathbf{x}_{n+1})\| \le a_2\|\mathbf{z} - \mathbf{x}_n\|^2.$$

Таким образом,

$$\begin{aligned} &\|\mathbf{x}_{n+1} - \mathbf{z}\| = \|(F'(x_n))^{-1}F'(x_n)(\mathbf{x}_{n+1} - \mathbf{z})\| \le \\ &\le \|(F'(\mathbf{x}_n))^{-1}\|\|F'(\mathbf{x}_n)(\mathbf{x}_{n+1} - \mathbf{z})\| \le \\ &\le (a_1a_2)\|\mathbf{x}_n - \mathbf{z}\|^2 = c\|\mathbf{x}_n - \mathbf{z}\|^2 < cb^2 = (cb)b \le b. \end{aligned}$$

Вложение $\mathbf{x}_n \in \Omega_b$, $n \ge 0$ доказано. Докажем сходимость. Пусть $q_n = c\|\mathbf{x}_n - \mathbf{z}\|$, тогда $\|\mathbf{x}_{n+1} - \mathbf{z}\| \le q_n\|\mathbf{x}_n - \mathbf{z}\|$. Умножим данное неравенство на c, получим $q_{n+1} \le q_n^2$. Покажем по индукции, что отсюда следует оценка $q_n \le$

$q_0^{2^n}$. Имеем: $n=0$, $q_0 \le q_0$; пусть верно для n; тогда $q_{n+1} \le q_n^2 \le (q_0^{2^n})^2 = q_0^{2^{n+1}}$. Таким образом, верна следующая оценка скорости сходимости:

$$c\|\mathbf{x}_n - \mathbf{z}\| \le (c\|\mathbf{x}_0 - \mathbf{z}\|)^{2^n}.$$

И так как $c\|\mathbf{x}_0 - \mathbf{z}\| < cb \le 1$, следовательно, $\mathbf{x}_n \to \mathbf{z}$. Теорема доказана.

Условия теоремы гарантируют, что корень $\mathbf{z}$ простой. Напомним, что в случае кратных корней метод Ньютона сходится с линейной скоростью, и скорость замедляется при повышении кратности.

Построение итераций высшего порядка. Как отмечалось, особый интерес представляют методы со сверхлинейной оценкой сходимости, т.к. в этом случае скорость сходимости увеличивается при приближении к корню.

Интерполяционный метод. Пусть имеется набор $x_n, ..., x_{n-m+1}$ из m приближений к корню z одномерной функции $f(x)$. Тогда в качестве очередного приближения целесообразно выбрать нуль интерполяционного многочлена $L_m(x)$, построенного по узлам $x_n, ..., x_{n-m+1}$. Это требует нахождения корней многочлена L_m. Как следствие широкое применение имеют только алгоритмы при $m = 2, 3$ — метод секущих и метод парабол.

Чтобы избежать проблем, связанных с решением алгебраического уравнения $L_m(x) = 0$, естественно интерполировать обратную к $y = f(x)$ функцию $x = F(y)$ по узлам $y_{n-i} = f(x_{n-i})$, $i = 0, ..., m-1$, и в качестве очередного приближения взять значение полученного интерполяционного многочлена в нуле. Линейная обратная интерполяция ($m = 2$) соответствует методу секущих, но уже при $m = 3$ прямая и обратная интерполяция приводят к различным алгоритмам.

Метод Чебышёва. Пусть z — корень уравнения $f(x) = 0$ и $F(y)$ — обратная к $f(x)$ функция. Тогда $x \equiv F(f(x))$ и $z = F(0)$. Разложим $F(0)$ в ряд Тейлора в окрестности некоторой точки y:

$$F(0) = F(y) + \sum_{k=1}^{m-1} F^{(k)}(y)\frac{(-y)^k}{k!} + \dots .$$

Приблизим значение $F(0)$ значением частичной суммы в точке $y = f(x)$:

$$z = F(0) \approx \varphi_m(x) = x + \sum_{k=1}^{m-1} (-1)^k F^{(k)}(f(x))\frac{(f(x))^k}{k!},$$

что соответствует замене исходной функции $F(y)$ на многочлен $\varphi_m(y)$ степени $m-1$, производные которого совпадают с соответствующими производными F в точке $y = f(x)$. Соответствующий итерационный процесс $x_{n+1} = \varphi_m(x_n)$ имеет порядок сходимости m.

Пример 27.1. Запишем формулы метода Чебышёва для функции $f(x) = x^p - a$. Обратная к f функция имеет вид $F(y) = (a+y)^{1/p}$, а производные F выражаются (с учетом $y = x^p - a$) формулой

$$F^{(k)}(y) = (a+y)^{1/p-k}\prod_{j=0}^{k-1}(\frac{1}{p} - j) = x^{1-kp}\prod_{j=0}^{k-1}\left(\frac{1}{p} - j\right).$$

Таким образом,

$$\varphi_m(x) = x + x\sum_{k=1}^{m-1}\frac{1}{k!}\left(\frac{a-x^p}{px^p}\right)^k\prod_{j=0}^{k-1}(1-jp).$$

В частности, $\varphi_2(x) = (x/p)(p - 1 + a/x^p)$.

При $p = 2$ получаем формулу Ньютона–Херона $x_{n+1} = \varphi_2(x_n)$ для приближенного вычисления квадратных корней.

Если $p = -1$, то $\varphi_m(x) = x\sum\limits_{k=0}^{m-1}(1-ax)^k$. В этом случае итерационный процесс $x_{n+1} = \varphi_m(x_n)$ при $|1 - ax| < 1$ сходится к решению уравнения $x - 1/a = 0$. Данный метод позволяет находить значение $1/a$ с произвольной точностью без использования операции деления.

δ^2-процесс Эйткена. Вычислим по имеющемуся приближению x_n значения $x_{n+1} = \varphi(x_n)$ и $x_{n+2} = \varphi(x_{n+1})$. Так как в малой окрестности z имеются представления

$$x_{n+1} - z \approx \varphi'(z)(x_n - z),$$
$$x_{n+2} - z \approx \varphi'(z)(x_{n+1} - z),$$

то из данных соотношений получаем $\varphi'(z) \approx \dfrac{x_{n+2} - x_{n+1}}{x_{n+1} - x_n}$,

$$z \approx \frac{x_{n+2} - \varphi'(z)x_{n+1}}{1 - \varphi'(z)} \approx \frac{x_{n+2}x_n - x_{n+1}^2}{x_{n+2} - 2x_{n+1} + x_n}.$$

Таким образом, за следующее после x_n приближение правильно принять

$$x_{n+1} = \frac{x_n\varphi(\varphi(x_n)) - \varphi(x_n)\varphi(x_n)}{\varphi(\varphi(x_n)) - 2\varphi(x_n) + x_n} =$$
$$= \varphi(\varphi(x_n)) - \frac{(\varphi(\varphi(x_n)) - \varphi(x_n))^2}{\varphi(\varphi(x_n)) - 2\varphi(x_n) + x_n}.$$

Известно, что если исходный процесс имел линейную скорость сходимости, то данная модификация имеет скорость сходимости более высокого порядка, но, возможно, только сверхлинейную. Применение рассмотренной модификации, например, к квадратично сходящейся последовательности формально не приводит к повышению порядка.

Данное преобразование является частным случаем (при $\varphi_1 = \varphi_2 = \varphi$) метода Стеффенсона–Хаусхолдера–Островского построения итерационной функции φ_3 более высокого порядка по известным φ_1 и φ_2:

$$\varphi_3(x) = \frac{x\varphi_1\left(\varphi_2(x)\right) - \varphi_1(x)\varphi_2(x)}{x - \varphi_1(x) - \varphi_2(x) + \varphi_1\left(\varphi_2(x)\right)}.$$

В данном случае функции φ_1 и φ_2 выбираются так, чтобы обеспечить эффективное подавление вектора ошибки $z - x_n$.

Литература

1. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. — М.: БИНОМ. Лаборатория знаний, 2011.

2. *Бахвалов Н.С., Корнев А.А., Чижонков Е.В.* Численные методы. Решения задач и упражнения. — М.: Лаборатория знаний, 2016.

3. *Деммель Дж.* Вычислительная линейная алгебра. — М.: Мир, 2001.

4. *Лебедев В.И.* Функциональный анализ и вычислительная математика. — М.: Физматлит, 2005.

Программа курса.

Вычислительная погрешность.

1. Вычислительная погрешность. Устойчивость задачи и численного алгоритма.

Разностные уравнения.

2. Линейные разностные уравнения n-го порядка. Теоремы о представлении общего решения однородного уравнения и общего решения неоднородного уравнения.

3. Линейные разностные уравнения n-го порядка с постоянными коэффициентами. Формулировка теорем о представлении общего решения однородного уравнения и частного решения неоднородного уравнения с квазимногочленом в правой части. Форма записи действительного решения.

4. Фундаментальное решение разностного уравнения. Теорема о представлении частного решения неоднородного уравнения первого порядка с постоянными коэффициентами.

5. Решение задач на собственные значения для разностных уравнений, сравнение с дифференциальным случаем.

6. Построение многочленов Чебышёва первого и второго рода.

7. Свойства многочленов Чебышёва первого рода: симметричность, нули, экстремумы. Теоремы о композиции.

8. Экстремальные свойства многочленов Чебышёва первого рода на отрезке $[a, b]$.

9. Экстремальные свойства многочленов Чебышёва первого рода вне (a, b).

Решение дифференциальных уравнений.

10. Конечно-разностный метод. Аппроксимация, устойчивость, сходимость, теорема Филиппова.

11. Метод неопределенных коэффициентов построения разностных схем. Погрешность формул численного дифференцирования, оценка для оптимального шага.

12. Задача Коши, условия аппроксимации p-го порядка на решении, α-устойчивость. Модельные схемы.

13. Численные методы решения задачи Коши: метод Тейлора, методы Адамса.

14. Методы Рунге–Кутта для решения задачи Коши.

15. Вычисление главного члена погрешности для простейших схем для задачи Коши. Оценка глобальной погрешности явного одношагового метода.

16. Устойчивые и неустойчивые задачи. Жесткие системы.

17. Метод Лебедева решения жестких систем.

18. Обыкновенные дифференциальные уравнения второго порядка, аппроксимация, α-устойчивость. Аппроксимация краевых условий третьего рода.

19. Устойчивость краевой задачи для уравнения второго порядка: метод собственных функций.

20. Устойчивость краевой задачи для уравнения второго порядка: энергетический метод.

21. Метод прогонки.

22. Метод стрельбы и метод Фурье.

Численные методы линейной алгебры.

23. Нормы векторов, линейных операторов, обусловленность матрицы. Оценка возмущения решения системы линейных алгебраических уравнений при возмущении правой части.

24. Метод Гаусса решения систем линейных алгебраических уравнений. Алгоритм ортогонализации Грама–Шмидта.

25. Метод отражений.

26. Невырожденная задача наименьших квадратов: метод нормального уравнения, метод QR-разложения.

27. Задача наименьших квадратов неполного ранга: методы QR-разложения и QR-разложения с выбором главного столбца

28. Сингулярное разложение. Теорема о наилучшем приближении матрицы малоранговыми матрицами в норме, подчиненной евклидовой.

29. Решение задачи наименьших квадратов полного и неполного рангов методом сингулярного разложения.

30. Задача наименьших квадратов с линейными ограничениями–равенствами: методы исключения, обобщенного сингулярного разложения, взвешиванием

31. Задача наименьших квадратов с ограничениями типа квадратных неравенств: метод обобщенного сингулярного разложения.

32. Метод простой итерации $\mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c}$ для решения систем линейных алгебраических уравнений.

33. Линейный оптимальный одношаговый метод и линейный оптимальный N-шаговый метод.

34. Метод наискорейшего градиентного спуска и метод минимальных невязок.

35. Итерационные методы с предобусловливателем

36. Методы Гаусса–Зейделя, Якоби и верхней релаксации

37. Проекционный алгоритм решения систем линейных алгебраических уравнений. Проекционная теорема, экстремальное свойство. Одномерные алгоритмы

38. Метод сопряженных градиентов.

39. Степенной метод со сдвигом для задач на собственные значения.
40. Метод обратной итерации со сдвигом для задач на собственные значения. Отношение Рэлея.
41. Инвариантные подпространства. Метод итерирования подпространств. QR-алгоритм.

Приближение функций одной переменной.

42. Интерполяционный многочлен Лагранжа. Минимизация остаточного члена погрешности.
43. Наилучшее приближение в линейном нормированном и гильбертовом пространствах.
44. Многочлен наилучшего равномерного приближения. Теорема Валле–Пуссена. Теорема Чебышёва.
45. Примеры построения многочлена наилучшего равномерного приближения. Теорема единственности.
46. Сплайн–интерполяция. Линейный интерполяционный сплайн.
47. Кубический интерполяционный сплайн.
48. Локальный (аппроксимационный) сплайн.
49. Приближение по многочленам Чебышёва.
50. Дискретное преобразование Фурье. Свойства, примеры.
51. Быстрое преобразование Фурье для $N = n_1 n_2$, $N = 2^k$ (рекурсивная форма).
52. Интерполяция Паде–Якоби. Многоточечная интерполяция Паде.

Численное интегрирование.

53. Метод неопределенных коэффициентов построения квадратур
54. Интерполяционные квадратуры.
55. Составные квадратуры.
56. Ортогональные многочлены.
57. Квадратурные формулы Гаусса.
58. Задачи оптимизации квадратур.
59. Правило Рунге оценки погрешности. Построение программ с автоматическим выбором шага.
60. Метод Монте–Карло вычисления интегралов.
61. Вычисление интегралов в нерегулярном случае.

Решение нелинейных уравнений.

62. Метод простой итерации: сжимающие и слабо сжимающие отображения.
63. Конструктивные теоремы о сходимости метода простой итерации и существовании корней уравнения.
64. Метод хорд, метод секущих: расчетные формулы и теоремы сходимости. Метод парабол.

65. Метод Ньютона в $\mathbf{R}^1$.
66. Методы типа простой итерации для решения систем нелинейных уравнений. Методы установления.
67. Метод Ньютона в $\mathbf{R}^m$.
68. Интерполяционные методы построения итераций высшего порядка: метод Чебышёва, δ^2-процесс Эйткена, метод Стефенсона–Хаусхолдера–Островского.

Список дополнительных задач.

Задача 27.1. *Найти общее решение в действительной форме для следующего уравнения:* $y_{k+2} + y_k = \cos\frac{\pi}{2}k$.

Задача 27.2. *Найти ограниченное фундаментальное решение уравнения* $y_{k+1} - y_k - 12y_{k-1} = \delta_k^0$.

Задача 27.3. *Найти все решения задачи на собственные значения*

$$\frac{y_{k+1} - 2\,y_k + y_{k-1}}{h^2} = -\lambda\, y_k\,, \quad 1 \le k \le N-1\,,$$
$$y_0 = 0\,, \quad -\frac{2}{h^2}(y_N - y_{N-1}) = -\lambda\, y_N\,, \quad h = \frac{1}{N}\,.$$

Задача 27.4. *Для задачи* $y' + y = \exp 2x,\ y(0) = 1$ *построить двухточечную разностную схему второго порядка сходимости.*

Задача 27.5. *Для уравнения* $y'(x) = f(x)$ *построить схему Адамса второго порядка аппроксимации* $\dfrac{y_k - y_{k-1}}{h} = \dfrac{c_1 f_{k-1} + c_2 f_{k-2}}{2}$. *Найти главный член погрешности и исследовать устойчивость.*

Задача 27.6. *Для задачи* $-u'' + p(x)\,u = f(x)$, $u'(0) = 1$, $u(1) = 0$, *где* $p(x) \ge 0$, *построить разностную схему второго порядка аппроксимации на сетке* $x_i = (i-1/2)h$, $i = 0, \ldots, N$, $h = 1/(N-1/2)$. *Исследовать устойчивость и сходимость.*

Задача 27.7. *Для задачи* $-u'' + p(x)\,u = f(x)$, $u'(0) = 1$, $u(1) = 0$, *где* $p(x) \ge 0$, *построить разностную схему второго порядка аппроксимации на сетке* $x_i = ih$, $i = 0, \ldots, N$, $h = 1/N$. *Исследовать устойчивость и сходимость.*

Задача 27.8. *Функция* $f(x) = e^x$ *приближается на* $[-1, 1]$ *многочленом Лагранжа по четырем равноотстоящим узлам. Найти наибольшее целое* p *в оценке погрешности следующего вида:* $\|f - L_n\|_{C[-1,1]} \le 10^{-p}$.

Задача 27.9. *Функция* $f(x) = \cos x$ *приближается многочленом Лагранжа на* $[-1, 1]$ *по* n *чебышёвским узлам. При каком значении* n *величина погрешности приближения в непрерывной норме не превосходит* 10^{-5}.

Задача 27.10. *Среди всех многочленов вида* $a_3x^3 + 5x^2 + a_1x + a_0$ *найти наименее уклоняющийся от нуля на отрезке* $[1, 2]$.

Задача 27.11. *Построить многочлен наилучшего равномерного приближения степени* $n = 3$ *для функции* $f(x) = 3\sin^2 10x + |x^2 - 7x + 10|$ *на отрезке* $[3, 4]$.

Задача 27.12. *Оценить число разбиений отрезка* N *для вычисления интеграла* $\int_0^1 \sin(x^2)\,dx$ *по составной квадратурной формуле трапеций, обеспечивающее точность* 10^{-4}.

Задача 27.13. *Построить квадратуру Гаусса с тремя узлами для вычисления интеграла* $I(f) = \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}dx$. *Указать алгебраический порядок точности построенной квадратуры.*

Задача 27.14. *Предложить способ вычисления интеграла* $\int_0^1 \frac{\ln x}{1+x^2}dx$ *по составной квадратурной формуле с постоянным шагом* h *и погрешностью* $O\left(h^2\right)$.

Задача 27.15. *Пусть* A *— матрица простой структуры, т.е. подобна диагональной, и все* $\lambda(A) \in [m, M]$, $m > 0$. *Доказать, что метод* $\dfrac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}$ *сходится при* $0 < \tau < \dfrac{2}{M}$.

Задача 27.16. *Пусть у задачи* $A\mathbf{x} = \mathbf{b}$ *с матрицей простой структуры имеется одно отрицательное собственное значение* $\lambda_1 \in [-2.01, -1.99]$, *а остальные — положительны:* $\lambda_i \in [1, 3]$, $i = 2, \ldots, n$. *Построить сходящийся итерационный метод для решения такой системы.*

Задача 27.17. *Для решения системы* $\begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}\mathbf{x} = \mathbf{b}$ *применяется метод Гаусса–Зейделя. Найти все значения параметров* α, β, *для которых метод сходится с произвольного начального приближения.*

Задача 27.18. *Уравнение* $x = 2^{x-1}$, *имеющее два корня* $z_1 = 1$ *и* $z_2 = 2$, *решают методом простой итерации* $x_{n+1} = 2^{x_n - 1}$. *Исследовать сходимость метода в зависимости от выбора начального приближения* x_0.

Задача 27.19. *Для решения системы* $\begin{cases} x^3 - y^2 = 4, \\ xy^3 - y = 14 \end{cases}$ *применяется метод Ньютона. Указать ненулевую окрестность* Q_δ *корня* $\mathbf{z} = (2, 2)$ *и оценить число итераций* n, *необходимое для достижения точности* $\|\mathbf{z} - \mathbf{x}_n\|_2 \le 10^{-3}$ *для произвольного* $\mathbf{x}_0 \in Q_\delta$.

Задача 27.20. *Для вычисления* $a^{1/p}$ *применяется следующий алгоритм:*

$$x_{n+1} = \varphi(x_n),\ \varphi(x) = x\frac{(p-1)x^p + (p+1)a}{(p+1)x^p + (p-1)a}.$$

Найти порядок сходимости метода.

Учебное издание

Корнев Андрей Алексеевич
Лекции по курсу
"Численные методы"

Учебное пособие для вузов

М., Издательство попечительского совета
механико-математического факультета МГУ, 167 стр.

Подписано в печать 29.08.2011 г.
Формат 60×90 1/16. Объем 10,5 п.л.
Заказ 5 Доп. тираж.

Издательство попечительского совета механико-математического факультета МГУ
г. Москва, Ленинские горы.

Отпечатано на типографском оборудовании механико-математического факультета