Учебно–методические материалы
по курсу "Численные методы"
4 курс, II поток

2019–2020 уч.г.

Вопросы по курсу "Численные методы"
4 курс, II поток

1. Погрешность метода и вычислительная погрешность. Пример неустойчивого алгоритма.
2. Алгебраическая интерполяция. Многочлен Лагранжа.
3. Константа Лебега интерполяционного процесса для равноотстоящих узлов.
4. Многочлены Чебышева и их свойства.
5. Интерполяционные сплайны. Конструкция и обоснование кубического сплайна.
6. Понятие об аппроксимационных сплайнах.
7. Наилучшее приближение в линейном нормированном пространстве.
8. Наилучшее приближение в гильбертовом пространстве.
9. Дискретное преобразование Фурье. Идея быстрого дискретного преобразования Фурье.
10. Наилучшее равномерное приближение многочленами.
11. Квадратурные формулы интерполяционного типа.
12. Ортогональные многочлены и квадратуры Гаусса.
13. Составные квадратурные формулы. Правило Рунге для оценки погрешности.
14. Основные приемы для вычисления нерегулярных интегралов.
15. Метод прогонки для решения трехдиагональных систем. Корректность и устойчивость метода прогонки.
16. Прямые методы решения систем линейных уравнений. Методы Гаусса и Холецкого.
17. Прямые методы решения систем линейных уравнений. Методы отражений и вращений.
18. Число обусловленности. Неравенства для ошибки и невязки.
19. Метод простой итерации решения систем линейных уравнений.
20. Оптимальный одношаговый итерационный метод.
21. Оптимальный циклический итерационный метод.
22. Обобщенный метод простой итерации.
23. Методы Якоби и Гаусса – Зейделя.
24. Метод верхней релаксации.
25. Метод наискорейшего градиентного спуска.
26. Линейная задача наименьших квадратов. Метод нормального уравнения.
27. Линейная задача наименьших квадратов. Методы QR-разложения и сингулярного разложения.

28. Общая идея и примеры проекционных методов.
29. Пространства Крылова. Понятие о методе сопряженных градиентов.
30. Частичная проблема собственных значений.
31. Полная проблема собственных значений. QR-алгоритм.
32. Метод простой итерации для нелинейных уравнений.
33. Метод Ньютона.
34. Явный метод Эйлера для обыкновенных дифференциальных уравнений (ОДУ). Устойчивость. Локальная и глобальная ошибки.
35. Явные методы Рунге – Кутты.
36. Неявные одношаговые методы решения ОДУ.
37. Многошаговые методы решения ОДУ.
38. Основы метода конечных элементов: вариационная постановка задачи, метод Ритца, базисные функции.
39. Оценка точности приближения кусочно – линейными функциями.
40. Проекционная теорема в методе конечных элементов.
41. Система уравнений в методе конечных элементов.
42. Решение модельной задачи методом Фурье.
43. Исследование устойчивости модельной задачи методом Фурье.
44. Метод стрельбы для решения трехдиагональных систем.
45. Пример аппроксимации уравнения и краевых условий.
46. Определения аппроксимации и устойчивости.
47. Определение сходимости. Теорема А.Ф.Филиппова.
48. Интегро – интерполяционный метод.
49. Исследование устойчивости методом априорных оценок.
50. Метод конечных разностей для уравнения Пуассона.
51. Спектральный признак устойчивости и примеры его применения для аппроксимаций гиперболического уравнения.
52. Принцип замороженных коэффициентов.
53. Исследование устойчивости простейших схем для уравнения теплопроводности в равномерной метрике.
54. Исследование устойчивости схемы с весами для уравнения теплопроводности в интегральной метрике.

Задачи к билетам по курсу "Численные методы"
4 курс, II поток

1. Найти $\sum\limits_{i=1}^{n} x_i^n \, \Phi_i(x)$, где $\Phi_i(x) = \prod\limits_{\substack{j=1 \\ j \neq i}}^{n} \dfrac{x - x_j}{x_i - x_j}$, $x_1 < x_2 < \cdots < x_n$.

2. Доказать, что если узлы интерполяции расположены симметрично относительно некоторой точки c, а значения интерполируемой функции в симметричных узлах равны, то интерполяционный многочлен Лагранжа — функция, четная относительно точки c.

3. Пусть функция $f(x) = \sin x$ задана на отрезке $[0, b]$. При каком b многочлен Лагранжа третьей степени, построенный по равноотстоящим узлам, приближает эту функцию с погрешностью $\varepsilon \leq 10^{-3}$?

4. Доказать следующее свойство многочленов Чебышева:

$$T_{2n}(x) = 2\,T_n^2(x) - 1\,;$$

5. Пусть $x^2 + y^2 = 1$. Доказать следующее свойство многочленов Чебышева: $T_{2n}(y) = (-1)^n T_{2n}(x)$.

6. Найти многочлен, наименее уклоняющийся от нуля на отрезке $[a, b]$ среди всех многочленов со старшим коэффициентом 1.

7. Пусть $\omega_n(x) = \prod\limits_{i=1}^{n}(x - x_i)$. Показать, что при любом выборе узлов $x_i \in [a, b]$ имеет место неравенство $\max\limits_{[a,b]} |\omega_n(x)| \geq (b-a)^n \, 2^{1-2n}$.

8. Среди всех многочленов вида $a_2x^2 + x + a_0$ найти наименее уклоняющийся от нуля на отрезке $[-1, 1]$.

9. Построить многочлен наилучшего равномерного приближения степени $n = 2$ для функции $f(x) = x^3$ на отрезке $[-1, 1]$.

10. Построить многочлен наилучшего равномерного приближения степени $n = 3$ для функции $f(x) = \exp(x^2)$ на отрезке $[-1, 1]$.

11. Построить многочлен наилучшего равномерного приближения степени $n = 3$ для функции $f(x) = |x^2 - 7x + 10|$ на отрезке $[3, 4]$.

12. Найти наилучшее приближение в $L_2(-1, 1)$, где $\|f\|^2_{L_2(-1,1)} = \int\limits_{-1}^{1} |f(x)|^2 \mathrm{d}x$, для функции $f(x) = x^2$ алгебраическими многочленами $Q_1(x)$.

13. Найти для функции $\exp(x)$ наилучшее приближение многочленом нулевой степени в норме $L_1(0,1)$, где $\|f\|_{L_1(0,1)} = \int\limits_0^1 |f(x)|\mathrm{d}x$.

14. Пусть P_2 — пространство алгебраических многочленов второй степени на отрезке $[-1,1]$ с нормой $||p(x)|| = |p(-1)| + |p(0)| + |p(1)|$. Найти наилучшее приближение константой для функции $p(x) = x^2 \in P_2$.

15. Рассмотреть формулы Ньютона – Котеса при $n = 1$ (прямоугольников) и $n = 2$ (трапеций) и сравнить оценки их погрешностей в случае гладких подынтегральных функций.

16. Доказать, что для погрешности квадратурной формулы трапеций справедливо представление

$$R_2(f) = \int\limits_a^b f(x)\,\mathrm{d}x - \frac{b-a}{2}\,(f(a)+f(b)) = \frac{1}{2}\int\limits_a^b (a-\xi)(b-\xi)f''(\xi)\,\mathrm{d}\xi.$$

17. Оценить минимальное число разбиений N отрезка $[0,1]$ для вычисления интеграла $\int\limits_0^1 \exp(x^2)\,\mathrm{d}x$, по составной квадратурной формуле прямоугольников, обеспечивающее погрешность не более 10^{-4}.

18. Оценить минимальное число разбиений N отрезка $[0,1]$ для вычисления интеграла $\int\limits_0^1 \exp(x^2)\,\mathrm{d}x$, по составной квадратурной формуле трапеций, обеспечивающее погрешность не более 10^{-4}.

19. Пусть T — треугольник на плоскости, $s(T)$ — его площадь, A, B, C — середины сторон. Показать, что кубатурная формула

$$S(f) = \frac{1}{3}\,s(T)\big(f(A)+f(B)+f(C)\big) \approx \iint\limits_T f(x)\,\mathrm{d}x,$$

где $x = (x_1, x_2)$, $\mathrm{d}x = \mathrm{d}x_1\mathrm{d}x_2$, точна для всех многочленов второй степени вида $a_0 + a_1x_1 + a_2x_2 + a_{11}x_1^2 + a_{12}x_1x_2 + a_{22}x_2^2$.

20. Построить квадратуру Гаусса с двумя узлами для вычисления интегралов $I(f) = \int\limits_{-1}^1 x^2 f(x)\,\mathrm{d}x$.

21. Пусть задан отрезок $[a,b]$. Доказать, что при $b > a \geq 0$ все коэффициенты ортогонального многочлена отличны от нуля.

22. Доказать, что все коэффициенты квадратуры Гаусса положительны.

23. Найти матричную норму, подчиненную векторной норме $||\cdot||_1$.

24. Найти матричную норму, подчиненную векторной норме $||\cdot||_\infty$.

25. Доказать, что модуль любого собственного значения матрицы не больше любой ее нормы.

26. Показать, что $\mathrm{cond}(A) \geq 1$ для любой матрицы A и любой матричной нормы.

27. Доказать неравенство $n^{-1} \leq \mathrm{cond}_1(A)/\mathrm{cond}_2(A) \leq n$ для квадратных невырожденных матриц размерности $n \times n$.

28. Получить неравенство $\mathrm{cond}(A) \geq |\lambda_{\max}(A)/\lambda_{\min}(A)|$ для произвольной невырожденной матрицы A и любой матричной нормы.

29. Оценить $\mathrm{cond}_2(A)$ трехдиагональной $n \times n$ матрицы A с элементами $a_{ij} = \{2$ для $i = j;\ -1$ для $|i-j| = 1;\ 0$ для остальных индексов$\}$.

30. Пусть элементы матрицы B имеют вид $b_{ij} = \frac{1}{2} \cdot 3^{-|i-j|}$. Доказать, что система $\mathbf{x} = B\mathbf{x} + \mathbf{c}$ имеет единственное решение и метод простой итерации сходится при любом начальном приближении.

31. Пусть матрица B имеет вид

$$B = \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

Найти все α, β, при которых метод простой итерации $\mathbf{x}^{k+1} = B\,\mathbf{x}^k + \mathbf{c}$ сходится с произвольного начального приближения.

32. Пусть матрица B в методе $\mathbf{x}^{k+1} = B\,\mathbf{x}^k + \mathbf{c}$ имеет вид

$$B = \begin{pmatrix} \alpha & 4 \\ 0 & \beta \end{pmatrix} \quad 0 < \alpha, \beta < 1\,.$$

Показать, что величина ошибки $\mathbf{z}^k = \mathbf{x} - \mathbf{x}^k$ в норме $\|\cdot\|_\infty$ начинает монотонно убывать лишь с некоторого номера итерации N. Оценить N при $\alpha = \beta \approx 1$.

33. Пусть все собственные значения матрицы A вещественные и положительные. Доказать сходимость метода $\mathbf{x}^{k+1} = (I - \tau A)\mathbf{x}^k + \tau\mathbf{b}$ при $\tau = \|A\|^{-1}$ с любой матричной нормой.

34. Пусть спектр матрицы A удовлетворяет условиям: $|\mathrm{Im}(\lambda(A))| \leq 1$, $0 < \delta \leq \mathrm{Re}(\lambda(A)) \leq 1$. Найти область значений вещественного параметра τ, при которых итерационный метод $\mathbf{x}^{k+1} =$

$(I - \tau A)\mathbf{x}^k + \tau\,\mathbf{b}$ для системы $A\,\mathbf{x} = \mathbf{b}$ сходится с произвольного начального приближения.

35. При каких условиях на спектр матрицы B итерационный метод $\mathbf{x}^{k+1} = (2B^2 - I)\mathbf{x}^k + 2(B + I)\mathbf{c}$ сходится быстрее метода простой итерации $\mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c}$?

36. Доказать, что для систем линейных уравнений второго порядка методы Якоби и Гаусса–Зейделя сходятся и расходятся одновременно.

37. Исследовать сходимость метода Гаусса – Зейделя, если матрица размерности $n \times n$ системы $A\,\mathbf{x} = \mathbf{b}$ имеет элементы: $a_{ij} = 3^{-|i-j|}$.

38. Пусть симметричная матрица A имеет собственные значения $\lambda(A) \in [m, M],\ m > 0$. Доказать, что при любом положительном значении итерационного параметра τ сходится метод

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\left(\frac{\mathbf{x}^{k+1} + \mathbf{x}^k}{2}\right) = \mathbf{b}\,.$$

Определить оптимальное значение τ_{opt}.

39. Пусть симметричная матрица A имеет собственные значения $\lambda(A) \in [m, M],\ m > 0$. При каких $\alpha \in [0, 1]$ метод

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\left(\alpha\mathbf{x}^{k+1} + (1 - \alpha)\mathbf{x}^k\right) = \mathbf{b}$$

сходится при любом $\tau > 0$?

40. Найти все $\alpha,\ \beta$, при которых метод Гаусса – Зейделя является сходящимся для системы уравнений $A\,\mathbf{x} = \mathbf{b}$ с матрицей

$$A = \begin{pmatrix} \alpha & \alpha & 0 \\ \alpha & \beta & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

41. Пусть $x_{n+1} = \sqrt{x_n + 2}$. Доказать, что $\lim\limits_{n\to\infty} x_n = 2$ для любого $x_0 \geq -2$.

42. Доказать, что итерационный процесс $x_{n+1} = \cos x_n$ сходится для любого начального приближения $x_0 \in \mathbf{R}^1$.

43. Исследовать сходимость метода простой итерации $x_{n+1} = x_n^2 - 2x_n + 2$ в зависимости от выбора начального приближения x_0.

44. Пусть уравнение $f(x) = 0$ имеет на отрезке $[a, b]$ корень z кратности p, причем $f(x)$ — дважды непрерывно дифференцируемая функция. Построить модификацию метода Ньютона, имеющую квадратичную скорость сходимости.

45. Проверить, что $\mathbf{z} = (1,1,1)^T$ — одно из решений системы уравнений $\mathbf{F}(\mathbf{x}) = 0$, где $\mathbf{F} : \mathbf{R}^3 \to \mathbf{R}^3$ имеет вид

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 x_2^3 + x_2 x_3 - x_1^4 - 1 \\ x_2 + x_2^2 + x_3 - 3 \\ x_2 x_3 - 1 \end{bmatrix}.$$

Сходится ли метод Ньютона к $\mathbf{z}$ при достаточно близких начальных приближениях?

46. Для дифференциальной задачи

$$u'' = f(x)\,, \quad x \in [0,1]\,, \quad u(0) = u(1) = 0$$

построить разностную схему интегро – интерполяционным методом на равномерной сетке.

47. Дана дифференциальная задача

$$-u'' + c\,u = f(x)\,,\ x \in [0,1]\,, \quad u(0) = u(1) = 0\,, \quad c = \text{const}\,.$$

При каких c для решения этой задачи можно применить метод конечных элементов?

48. Проверить, аппроксимирует ли разностная схема уравнение $y' = f(x,y)$ на равномерной сетке $x_k = x_0 + kh$, $k \geq 0$:

$$\frac{1}{8h}(y_k - 3y_{k-2} + 2y_{k-3}) = \frac{1}{2}(f_{k-1} + f_{k-2}), \quad \text{где} \quad f_k = f(x_k, y_k)\,.$$

49. Для задачи $y' + y = x + 1$, $y(0) = 0$ рассматривается схема

$$\frac{y_{k+1} - y_{k-1}}{2h} + y_k = kh + 1, \quad y_0 = 0, \quad y_1 = 0\,.$$

Каков порядок аппроксимации на решении данной схемы? Можно ли его улучшить?

50. Для уравнения $y' = f(x,y)$ построить разностную схему

$$\frac{y_k - y_{k-2}}{2h} = a_1 f_k + a_0 f_{k-1} + a_{-1} f_{k-2}, \quad \text{где} \quad f_k = f(x_k, y_k),$$

с наивысшим порядком аппроксимации p на решении.

51. Для задачи $y' = y$, $y(0) = 1$ рассмотреть схему

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

и в разложении ошибки $y(x_N) - y_N = c_1 h + c_2 h^2 + \cdots$ найти постоянную c_1 для $x_N = Nh = 1$.

52. Для задачи $y' = y$, $y(0) = 1$ рассмотреть схему

$$4\,\frac{y_{k+1} - y_{k-1}}{2h} - 3\,\frac{y_{k+1} - y_k}{h} = y_k, \quad y_0 = 1, \quad y_1 = e^h\,, \quad k \geq 1,$$

и в разложении ошибки $y(x_N) - y_N = c_1 h + c_2 h^2 + \cdots$ найти постоянную c_1 для $x_N = Nh = 1$.

53. Имеется краевая задача $u'' - 2u = \sin x - 1$, $u'(0) - u(0) = 0$, $u(1) = 0$. На сетке с шагом h построить разностную схему с аппроксимацией второго порядка на решении.

54. Исследовать устойчивость разностной схемы с постоянным коэффициентом a с помощью спектрального признака

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\,\frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2h} = 0\,.$$

55. Исследовать устойчивость разностной схемы с постоянным коэффициентом a с помощью спектрального признака

$$\frac{u_m^{n+1} - \dfrac{u_{m+1}^n + u_{m-1}^n}{2}}{\tau} + a\,\frac{u_{m+1}^n - u_{m-1}^n}{2h} = 0\,.$$

56. Исследовать устойчивость разностной схемы с постоянным коэффициентом a с помощью спектрального признака

$$\frac{u_m^{n+1} - \dfrac{u_{m+1}^n + u_{m-1}^n}{2}}{\tau} + a\,\frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2h} = 0\,.$$

57. При каком соотношении τ и h разностная схема

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}$$

имеет на решении порядок аппроксимации $O\left(\tau^2 + h^4\right)$?

58. Первая краевая задача для однородного уравнения теплопроводности $u_t = u_{xx}$ аппроксимируется явной разностной схемой ($Mh = 1$)

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}\,,\ 1 \le m \le M-1,$$
$$u_m^0 = \varphi(mh), \quad u_0^n = u_M^n = 0\ \forall n \ge 0\,.$$

Определить порядок сходимости решения разностной схемы к решению дифференциальной задачи при различных $\rho = \tau/h^2$.

59. Исследовать устойчивость разностной схемы по начальным данным в интегральной метрике ($Mh = 1$)

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} = \frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2},\ 1 \le m \le M-1,\ u_0^n = u_M^n = 0\ \forall n \ge 1.$$

60. Исследовать устойчивость разностной схемы по начальным данным в интегральной метрике ($Mh = 1$)

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} = \frac{u_{m-1}^{n-1} - 2u_m^{n-1} + u_{m+1}^{n-1}}{h^2},\ 1 \le m \le M-1,\ u_0^n = u_M^n = 0\ \forall n \ge 1.$$

61. Построить аппроксимацию оператора Лапласа и оценить ее погрешность на шаблоне "крест":

$$\Delta^h u(x_1, x_2) = h^{-2}[a_{0,0}u(x_1, x_2) + a_{1,0}u(x_1 + h, x_2) + a_{-1,0}u(x_1 - h, x_2) + a_{0,1}u(x_1, x_2 + h) + a_{0,-1}u(x_1, x_2 - h)],$$

где коэффициенты $a_{i,j}$ не зависят от h.

62. Построить аппроксимацию оператора Лапласа и оценить ее погрешность на шаблоне "косой крест":

$$\Delta^h u(x_1, x_2) = h^{-2}[a_{0,0}u(x_1, x_2) + a_{1,1}u(x_1 + h, x_2 + h) + a_{1,-1}u(x_1 + h, x_2 - h) + a_{-1,1}u(x_1 - h, x_2 + h) + a_{-1,-1}u(x_1 - h, x_2 - h)],$$

где коэффициенты $a_{i,j}$ не зависят от h.

63. Для уравнения $\Delta u = f$ на равномерной сетке с шагом h построить аппроксимацию на решении с порядком $O(h^2)$ граничного условия $\partial u/\partial x_1 - \alpha u = 0$ при $x_1 = 0$, используя минимальное количество узлов вдоль оси x_1.

Литература по курсу "Численные методы" 4 курс, II поток

1. Бахвалов Н.С., Корнев А.А., Чижонков Е.В. „Численные методы. Решения задач и упражнения“. – 2-е изд. – М.: Лаборатория Знаний, 2016.

2. Локуциевский О.В., Гавриков М.Б. „Начала численного анализа“. – М.: ТОО „Янус“, 1995.

3. Марчук Г.И. „Методы вычислительной математики“. – М.: Наука, 1980.

4. Самарский А.А. „Теория разностных схем“. – М.: Наука, 1977.

5. Стренг Г., Фикс Дж. „Теория метода конечных элементов“. – М.: Мир, 1977.

6. Тыртышников Е.Е. „Методы численного анализа“. – М.: Издательский центр „Академия“, 2007.

7. Чижонков Е.В. „Численные методы. Конспект лекций. “ – Эл. версия.

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ М.В ЛОМОНОСОВА
МЕХАНИКО – МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ

ЧИСЛЕННЫЕ МЕТОДЫ.
Конспект лекций.

Лектор — проф. ЧИЖОНКОВ Е.В.

Лекция 1

Погрешность метода и вычислительная погрешность. — Пример неустойчивого алгоритма. — Алгебраическая интерполяция. Многочлен Лагранжа.

Погрешность метода и вычислительная погрешность

Чтобы получить простейшее представление о предмете „Численные методы“, рассмотрим следующий пример. Пусть имеется возможность вычисления значений некоторой функции $f(x)$ при произвольном $x_1, x_2, \ldots$, а требуется приближенно определить значение производной $f'(x_0)$ в фиксированной точке x_0. Справочные пособия предлагают различные формулы. Расмотрим одну из них:

$$f'(x_0) \approx \frac{f(x_0+h) - f(x_0-h)}{2h}\,.$$

В этой формуле имеется параметр $h > 0$, который рекомендуется выбирать „достаточно малым“, но сами эти слова при этом не конкретизируются.

Так как всегда следует учитывать возможность опечаток, проверим эту формулу на тестовом примере $f(x) = e^x$ в точке $x_0 = 1$. Для этого составим элементарную программу для вычисления погрешности

$$err = \left| f'(x_0) - \frac{f(x_0+h) - f(x_0-h)}{2h} \right|.$$

Если использовать переменные типа float, занимающие по 4 байта памяти, то с помощью компьютера можно получить следующие данные:

h	err
10^{-1}	$4.54 \cdot 10^{-3}$
10^{-2}	$4.16 \cdot 10^{-5}$
10^{-3}	$4.11 \cdot 10^{-5}$

h	err
10^{-4}	$5.57 \cdot 10^{-4}$
10^{-5}	$4.28 \cdot 10^{-3}$
10^{-6}	$3.01 \cdot 10^{-2}$

Они нам не очень нравятся, так как в них отсутствует стремление погрешности к нулю при $h \to 0$. Проверка аналогичных формул приведет к такому же качественному результату: сначала погрешность будет убывать, а затем — монотонно возрастать. Это означает, что проблема не в опечатках. Тогда — в чем? Попробуем разобраться. Сначала применим формулу Тейлора для трижды непрерывно дифференцируемой функции $f(x) \in C^{(3)}[x_0-h, x_0+h]$

$$f(x_0 \pm h) = f(x_0) \pm hf'(x_0) + \frac{h^2}{2}f''(x_0) \pm \frac{h^3}{6}f^{(3)}(x_\pm).$$

Подстановка этих выражений в формулу для вычисления производной дает

$$\frac{f(x_0+h) - f(x_0-h)}{2h} - f'(x_0) = \frac{h^3\left[f^{(3)}(x_+) + f^{(3)}(x_-)\right]}{6 \cdot 2h}\,.$$

Теперь применим теорему о среднем для непрерывной функции $f^{(3)}(x)$

$$\frac{h^3\left[f^{(3)}(x_+) + f^{(3)}(x_-)\right]}{6 \cdot 2h} = \frac{h^2 f^{(3)}(\xi)}{6}, \quad \xi \in [x_0-h, x_0+h].$$

Величина в правой части оценивается следующим образом:

$$\left|\frac{h^2}{6}f^{(3)}(\xi)\right| \le \frac{h^2}{6}M_3, \quad M_3 = \max_{[x_0-h,x_0+h]}|f^{(3)}(x)|\,.$$

Полученная оценка означает, что $err \le \dfrac{h^2}{6}M_3$, или что уменьшение h в 10 раз должно приводить к уменьшению погрешности по крайней мере в 100 раз соответственно. Но такая зависимость видна в таблице только при переходе от первой строки ко второй. Что же происходит дальше? А дальше начинает оказывать влияние конечная разрядность представления чисел в компьютере. Для анализа эффектов такого рода необходимо привлечение других понятий, в первую очередь, *машинной точности* ε. Это такая наибольшая положительная величина, что $1 \oplus \varepsilon = 1$, где символом $\oplus$ обозначена операция сложения чисел в компьютере. Величина ε по смыслу близка к относительной погрешности представления чисел в компьютере. Напомним, что если a^* — приближенное значение величины $a \neq 0$, то относительной погрешностью называют величину $\delta(a)$, про которую известно, что

$$\left|\frac{a^* - a}{a}\right| \le \delta(a)\,.$$

Для проведения анализа влияния машинной точности на погрешность вычислений удобно считать, что $\delta(a) \le \varepsilon$, что в действительности и имеет место при использовании так называемых *чисел с плавающей точкой (запятой)*.

Чтобы получить информацию о порядке величины ε в конкретном компьютере, достаточно знать основание системы счисления β и длину мантиссы t (количество разрядов для представления дробной части в этой системе). При обычном способе округления, как правило, ориентируются на формулу

$$\varepsilon = \frac{1}{2}\,\beta^{-t}.$$

В нашем случае переменных типа float имеем $\beta = 2, t = 23$, что дает $\varepsilon = 2^{-24} \approx 6{\cdot}10^{-8}$. Однако заметим, что для вычислительной практики наиболее характерно использование переменных типа double ($\beta = 2$, $t = 52$), что соответствует машинной точности — $\varepsilon \approx 10^{-16}$.

Обозначим через $f^*(x)$ приближенное значение функции $f(x)$ в точке x и, имея в виду неравенство

$$\left|\frac{f^*(x) - f(x)}{f(x)}\right| \le \varepsilon \quad \forall x \in [x_0 - h, x_0 + h],$$

попытаемся заново проанализировать погрешность исходной формулы.

$$\left|\frac{f^*(x_0+h) - f^*(x_0-h)}{2h} - f'(x_0)\right| =$$

$$= \left|\frac{f^*(x_0+h) - f^*(x_0-h) \pm f(x_0+h) \pm f(x_0-h)}{2h} - f'(x_0)\right| =$$

$$= \left|\left(\frac{f(x_0+h) - f(x_0-h)}{2h} - f'(x_0)\right) + \frac{f^*(x_0+h) - f(x_0+h)}{2h} -\right.$$

$$- \frac{f^*(x_0-h) - f(x_0-h)}{2h}\Bigg| \le \frac{h^2}{6}M_3 + \frac{\varepsilon}{h}M_0\,,$$

где $M_0 = \max\limits_{[x_0-h,x_0+h]} |f(x)|$. В нашем случае $M_0 = M_3\,(\approx e)\, < 3$, и численные эксперименты хорошо согласуются с оценкой

$$err \le 3\left(\frac{h^2}{6} + \frac{\varepsilon}{h}\right), \quad \varepsilon \approx 6\cdot 10^{-8}.$$

Далее можно было бы поговорить об уменьшении погрешности при использовании других формул, переменных типа double и т.п., но для нас более важной является иллюстрация на этом примере следующей схемы: **постановка задачи → приближенный метод решения → оценка погрешности (*погрешность метода*) → оценка погрешности с учетом округлений (*влияние вычислительной погрешности*)**. Такой структуры мы будем стараться придерживаться в течение всего курса. Следует отметить, что последний этап зачастую определяет применимость метода вообще. Это хорошо иллюстрирует следующий пример.

Пример неустойчивого алгоритма

Пусть требуется вычислить последовательность интегралов:

$$I_n = \int\limits_0^1 x^n e^{x-1} dx, \quad n = 1,2,3,\ldots,N\,.$$

Для построения численного алгоритма проведем интегрирование по частям

$$I_n = \int\limits_0^1 x^n d(e^{x-1}) = x^n e^{x-1}\Big|_0^1 - \int\limits_0^1 n x^{n-1} e^{x-1} dx = 1 - nI_{n-1}\,.$$

В результате мы получили рекуррентное соотношение, к которому необходимо добавить недостающее начальное условие: $I_1 = 1/e$. Легко заметить, что при отсутствии ошибок округления погрешность метода равна нулю (точный метод!). Что будет при реальных вычислениях? Рассмотрим модельную ситуацию, когда погрешность возникает только вследствие определения величины I_1. Введем обозначение для ошибки $z_n = I_n - I_n^*$, где величины со звездочкой по – прежнему обозначают приближенные значения. Так как мы предположили справедливость равенств $I_n^* = 1 - nI_{n-1}^*$, то для ошибок имеем соотношения $z_n = -nz_{n-1} = n!(-1)^{n+1}z_1$. Это означает, что погрешность очень быстро (факториально!) растет, чередуя при этом знак. Очень скоро это приведет к сильному искажению искомого результата, вплоть до потери смысла, так как сами величины I_n положительны и монотонно убывают, что следует из неравенств

$$0 < I_n < \int\limits_0^1 x^n dx = \frac{1}{n+1}\,.$$

Алгоритмы такого рода называют *неустойчивыми*, пользоваться ими нельзя.

Можно ли исправить ситуацию? Если переписать рекуррентное соотношение в виде

$$I_{n-1} = \frac{1 - I_n}{n}\,,$$

то при его использовании вычислительная погрешность будет наоборот убывать, но где взять начальное значение? Его можно извлечь из полученной выше оценки. Пусть, например, требуется вычислить серию интегралов для $n \le 500$. Положим $I^*_{600} = 0$, в этом случае абсолютная ошибка не будет превышать величины $1/601$. Далее будем считать по формуле в сторону уменьшения n. Когда мы дойдем до $n \le 500$, начальная погрешность исчезнет практически полностью, и полученные приближения к интегралам будут находиться в пределах машинной точности.

Алгебраическая интерполяция. Многочлен Лагранжа

Пусть $a = x_1 < x_2 < \ldots < x_n = b$ — набор различных точек (узлов) на отрезке $[a, b]$, в которых заданы значения функции $f(x)$, так что $f_i = f(x_i)$, $i = 1, \ldots, n$. Требуется построить многочлен $L_{n-1}(x)$, принимающий в точках x_i значения f_i и оценить погрешность приближения достаточно гладкой функции этим многочленом на всем отрезке $[a, b]$.

Сначала построим в явном виде вспомогательные многочлены $\Phi_i(x)$ степени $n-1$, удовлетворяющие условиям $\Phi_i(x_i) = 1$, $\Phi_i(x_j) = 0$ при $j \ne i$:

$$\Phi_i(x) = \prod_{\substack{j=1 \\ j \ne i}}^{n} \frac{x - x_j}{x_i - x_j}\,.$$

Далее с их помощью запишем формулу для многочлена Лагранжа

$$L_{n-1}(x) = \sum_{i=1}^{n} f_i\, \Phi_i(x)\,.$$

Существование и единственность многочлена степени $n-1$, принимающего в n различных точках заданные значения, следует из отличия от нуля соответствующего определителя Вандермонда, поэтому указанный многочлен $L_{n-1}(x)$ есть единственное решение поставленной задачи.

Оценим погрешность замены $f(x)$ на $L_{n-1}(x)$, считая $f(x)$ достаточно гладкой функцией. Рассмотрим для этого функцию

$$\varphi(t) = f(t) - L_{n-1}(t) - K\,\omega_n(t)\,,$$

где K — некоторая постоянная, $\omega_n(t) = (t - x_1)\ldots(t - x_n)$. Выберем K из условия $\varphi(x) = 0$, где $x \in [a, b]$ — точка, в которой оценивается погрешность. Величина K при этом определяется однозначно

$$K = \frac{f(x) - L_{n-1}(x)}{\omega_n(x)}\,.$$

Заметим, что $x \ne x_i$, $i = 1, \ldots, n$, так как $f(x_i) = L_{n-1}(x_i)$ для любого значения K. При указанном выборе K функция $\varphi(t)$ обращается в нуль на $[a, b]$ в $(n+1)$-ой точке: $x, x_1, \ldots, x_n$. Поэтому по теореме Ролля $\varphi'(t)$ обращается в нуль в n точках. Продолжая рассуждения, получим, что $\varphi^{(n)}(t)$

имеет по крайней мере один нуль на $[a,b]$ в некоторой точке ξ. Так как $\varphi^{(n)}(t) = f^{(n)}(t) - K\,n!$, то из условия $\varphi^{(n)}(\xi) = 0$ будем иметь $K = \dfrac{f^{(n)}(\xi)}{n!}$. Сформулируем окончательное

Утверждение. *Пусть n-я производная функции $f(x)$ непрерывна на отрезке $[a,b]$. Тогда для любой точки $x \in [a,b]$ существует точка $\xi \in [a,b]$ такая, что справедливо равенство*

$$f(x) - L_{n-1}(x) = \frac{f^{(n)}(\xi)}{n!}\,\omega_n(x),\ \text{где}\ \ \omega_n(x) = \prod_{i=1}^{n}(x - x_i)\,.$$

Следствием этого представления является оценка погрешности в равномерной норме

$$\|f(x) - L_{n-1}(x)\| \le \frac{\|f^{(n)}(x)\|}{n!}\|\omega_n(x)\|,\ \ \text{где}\ \|g(x)\| = \sup_{x\in[a,b]}|g(x)|.$$

Лекция 2

Константа Лебега интерполяционного процесса для равноотстоящих узлов. — Многочлены Чебышева и их свойства.

Константа Лебега интерполяционного процесса для равноотстоящих узлов.

Если значения интерполируемой функции $f(x)$ известны с некоторой относительной погрешностью ε, т.е.

$$\frac{|f(x_i) - f^*(x_i)|}{|f(x_i)|} \le \varepsilon,$$

то для равномерной нормы разности точного $L_{n-1}(x)$ и приближенного $L^*_{n-1}(x)$многочленов Лагранжа справедлива оценка

$$\|L_{n-1}(x) - L^*_{n-1}(x)\| \le \varepsilon \|f(x)\| \lambda_n.$$

Величина $\lambda_n = \max\limits_{x\in[a,b]} \sum\limits_{i=1}^{n} |\Phi_i(x)|$ называется *константой Лебега интерполяционного процесса.* Скорость ее роста в зависимости от величины n существенно влияет как на сходимость $L_n(x)$ к $f(x)$, так и на оценку вычислительной погрешности интерполяции. Для равномерных сеток λ_n растет экспоненциально. Это приводит к тому, что построенный на *равномерной сетке* интерполяционный полином $L_n(x)$ при большом числе узлов может сильно отличаться от приближаемой функции. Так, например, для функции Рунге $f(x) = \dfrac{1}{25x^2+1}$ на отрезке $[-1,1]$ известно, что при использовании равномерной сетки узлов $\max\limits_{x\in[-1,1]} |L_n(x) - f(x)| \to \infty$ при $n \to \infty$.

Утверждение 1. *Пусть $a \le x \le b$ и $-1 \le y \le 1$, и, соответственно, узлы интерполяции x_i и y_i, $i = 1, \dots, n$, связаны линейным соотношением $x_i = x(y_i) = \frac{a+b}{2} + \frac{b-a}{2} y_i$. Тогда что константы Лебега интерполяционного процесса $\lambda_n^{[a,b]}$ и $\lambda_n^{[-1,1]}$, соответствующие этим отрезкам, совпадают.*

Доказательство. По определению вспомогательные многочлены $(n-1)$–й степени $\Phi_i(y)$, $i = 1, \dots, n$ обладают свойством $\Phi_i(y_k) = \delta_i^k$. Положим в формуле для $\Phi_i(x)$, обладающей теми же свойствами, $x = x(y) = \frac{a+b}{2} + \frac{b-a}{2} y$. Линейное преобразование не меняет степени многочлена. Кроме того, $\Phi_i(x_k) = \Phi_i(x(y_k)) = \Phi_i(y_k) = \delta_i^k$, т.е. два многочлена $(n-1)$–й степени совпадают в n точках. Отсюда следует их тождественное совпадение, и, следовательно, равенство констант Лебега $\lambda_n^{[a,b]}$ и $\lambda_n^{[-1,1]}$.

Таким образом, величина λ_n не зависит от длины и расположения отрезка интерполяции $[a,b]$, а определяется только взаимным расположением узлов.

Утверждение 2. *Для системы равноотстоящих узлов $\{x_i = i,\ i = 1, \dots, n\}$ при $n \ge 2$ справедлива оценка снизу для константы Лебега $\lambda_n \ge K\, 2^n / n^{3/2}$ с постоянной K, не зависящей от n.*

Доказательство. По определению λ_n на отрезке $[1,n]$ имеем

$$\lambda_n = \max_{x\in[1,n]} \sum_{i=1}^{n} \left| \prod_{\substack{j=1 \\ j\ne i}}^{n} \frac{x-j}{i-j} \right|.$$

Отметим справедливость соотношений

$$\prod_{\substack{j=1\\ j\neq i}}^{n} |i-j| = (i-1)!\,(n-i)!\,, \qquad \prod_{j=1}^{n}(j-1/2) \geq \frac{n!}{2\sqrt{n}},\ \ n \geq 1\,,$$

первое из которых очевидно, а второе показывается по индукции. Теперь с их помощью проведем оценку снизу для λ_n:

$$\lambda_n = \max_{x\in[1,n]} \sum_{i=1}^{n} \frac{1}{(i-1)!\,(n-i)!} \prod_{\substack{j=1\\ j\neq i}}^{n} |x-j| \ \geq \ \sum_{i=1}^{n} \frac{1}{(i-1)!\,(n-i)!} \prod_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{3}{2}-j\right|$$

(использовано неравенство $\max\limits_{x\in[1,n]} |f(x)| \geq |f(3/2)|$). Для оценки произведения в правой части проделаем преобразования:

$$\prod_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{3}{2}-j\right| = \frac{1}{|i-\frac{3}{2}|} \prod_{j=1}^{n} \left|\frac{3}{2}-j\right| = \frac{1}{2|i-\frac{3}{2}|} \prod_{j=1}^{n-1} \left|\frac{1}{2}-j\right| \geq \frac{(n-1)!}{4(n-\frac{3}{2})\sqrt{n-1}}\,.$$

И наконец, получим искомое неравенство $\left(K = \dfrac{1}{8}\right)$:

$$\lambda_n \geq \frac{1}{4(n-\frac{3}{2})\sqrt{n-1}} \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!\,(n-i)!} \geq \frac{1}{4n^{3/2}} \sum_{i=1}^{n} C_{n-1}^{i-1} = \frac{1}{8}\,\frac{2^n}{n^{3/2}}\,.$$

Утверждение доказано.

В качестве самостоятельного упражнения предлагается доказать следующее утверждение: *Для системы равноотстоящих узлов* $\{x_i = i,\ i = 1,\dots,n\}$ *при* $n \geq 2$ *справедлива оценка сверху для константы Лебега* $\lambda_n \leq K\,2^n$ *с постоянной* K*, не зависящей от* n*.*

Многочлены Чебышева и их свойства

Имеется несколько способов определения последовательности многочленов Чебышева первого рода $T_n(x)$. Рассмотрим некоторые из них.

1) Рекуррентное соотношение:

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

2) Тригонометрическая форма. При любом η имеем

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

Полагая $\eta = \arccos x$, получим

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

Эта форма удобна для применения на отрезке $[-1,1]$. Заметим, что $\big|T_n(x)\big| \leq 1$ при $|x| \leq 1$.

3) Разностное уравнение. Рекуррентное соотношение является разностным уравнением по переменной n. Ему соответствует характеристическое уравнение

$$\mu^2 - 2x\mu + 1 = 0\,.$$

Следовательно, $\mu_{1,2} = x \pm \sqrt{x^2-1}$. При $x \neq \pm 1$ справедливо $T_n(x) = C_1\mu_1^n + C_2\mu_2^n$. Из начальных условий получаем $C_1 = C_2 = 1/2$, что приводит к формуле

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

В силу непрерывности многочлена, формула верна и при $x = \pm 1$.

Отметим, что все многочлены $T_{2n}(x)$ — четные, а $T_{2n+1}(x)$ — нечетные, т.е. $T_n(-x) = (-1)^n T_n(x)$. При этом коэффициент при старшем члене равен 2^{n-1}.

Рассмотрим некоторые полезные свойства этих многочленов.

Нули. Найдем на $[-1,1]$ все решения уравнения $T_n(x) = 0$, используя тригонометрическую форму:

$$x_k = \cos\frac{\pi\,(2k-1)}{2n} \quad k = 1,\ldots,n.$$

Легко проверить, что они все различны, поэтому вне отрезка $[-1,1]$ других нулей многочлена $T_n(x)$ не существует.

Экстремумы на отрезке $[-1,1]$. Внутри отрезка имеется ровно $(n-1)$ экстремум (между двумя последовательными нулями). Для их определения можно воспользоваться уравнением

$$0 = T_n'(x) = \frac{n\,\sin(n\arccos x)}{\sqrt{1-x^2}}\,.$$

Это дает $x_m = \cos(\pi m/n)$, $m = 1,\ldots,n-1$. В экстремальных точках справедливы равенства $T_n(x_m) = (-1)^m$. Кроме этого, в концевых точках $x = \pm 1$ имеем: $T_n(1) = 1$, $T_n(-1) = (-1)^n$. Поэтому общее количество экстремумов на отрезке $[-1,1]$ равно $(n+1)$:

$$x_m = \cos\frac{\pi m}{n} \quad m = 0,\ldots,n\,.$$

Наименьшее уклонение от нуля. Рассмотрим приведенные многочлены Чебышева:

$$\bar{T}_n(x) = 2^{1-n}T_n(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0\,,$$

т.е. многочлены с коэффициентом 1 при старшей степени. Для них справедливо

Утверждение 3. *Приведенный многочлен Чебышева степени n наименее уклоняется от нуля среди всех многочленов $P_n(x) = x^n + b_{n-1}x^{n-1} + \ldots + b_1x + b_0$ со старшим коэффициентом 1 на отрезке $[-1,1]$, т.е.*

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

Доказательство. Пусть $\max\limits_{[-1,1]}\left|P_n(x)\right| < 2^{1-n}$. Тогда в точках экстремума многочлена Чебышева знак разности $\bar{T}_n(x) - P_n(x)$ определяется знаком $\bar{T}_n(x)$

$$\operatorname{sign}\left(\bar{T}_n(x_m) - P_n(x_m)\right) = \operatorname{sign}\left((-1)^m 2^{1-n} - P_n(x_m)\right) = (-1)^m.$$

При этом указанная разность является отличным от нуля многочленом степени $n-1$, но имеет n нулей, поскольку $n+1$ раз меняет знак в точках экстремума. Полученное противоречие и дает искомый результат.

Отображение на отрезок $[a,b]$. Пусть требуется найти многочлен, наименее уклоняющийся от нуля среди всех многочленов со старшим коэффициентом 1 на отрезке $[a,b]$. Тогда 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 $\bar{T}_n(x')$ при этом преобразуется в многочлен $\bar{T}_n\left(\frac{2x-(b+a)}{b-a}\right)$ со старшим коэффициентом $(2/(b-a))^n$. После перенормировки и использования схемы доказательства из предыдущего пункта имеем

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

Иногда требуется иметь многочлен вида: $P_n(x) = b_n x^n + b_{n-1}x^{n-1} + \ldots + b_1 x + 1$, т.е. с коэффициентом 1 при младшей степени, наименее отклоняющийся от нуля на отрезке $[a,b]$, $a \geq 0$. Его можно получить перенормировкой $P_n^*(x) = \dfrac{\bar{T}_n^{[a,b]}(x)}{\bar{T}_n^{[a,b]}(0)}$, так как многочлен $\bar{T}_n^{[a,b]}(x)$ имеет все нули внутри отрезка $[a,b]$. Покажем, что построенный многочлен имеет наименьшее отклонение от нуля в своем классе. Если применить уже использованную выше схему рассуждений, то получим, что разность $P_n^*(x) - P_n(x)$ является многочленом n-ой степени, имеет n нулей внутри $[a,b]$ и дополнительный нуль при $x=0$. Следовательно эта разность должна тождественно равняться нулю, что приводит к противоречию.

Минимизация оценки погрешности алгебраической интерполяции. Напомним, что при приближении достаточно гладкой функции $f(x)$ многочленом Лагранжа справедливо представление погрешности

$$f(x) - L_{n-1}(x) = \frac{f^{(n)}(\xi)}{n!}\omega_n(x),$$

откуда следует оценка погрешности в равномерной норме

$$\|f(x) - L_{n-1}(x)\| \leq \frac{\|f^{(n)}(x)\|}{n!}\|\omega_n(x)\|,$$

где $\omega_n(x) = (x-x_1)\ldots(x-x_n) = x^n + \ldots$ — многочлен с коэффициентом 1 при старшей степени. Из свойства наименьшего уклонения от нуля получаем

$$\|\omega_n(x)\| \geq \|\bar{T}_n^{[a,b]}(x)\| = (b-a)^n 2^{1-2n},$$

т.е. выбор в качестве узлов интерполяции нулей многочлена Чебышева n-й степени на отрезке $[a,b]$ приводит к минимизации оценки погрешности сверху. Заметим, что эту наименьшую оценку качественно улучшить нельзя. Пусть $f(x) = a_n x^n$, тогда $f^{(n)}(\xi) = a_n n!$ и

$$\|f(x) - L_{n-1}(x)\| = |a_n|\|\omega_n(x)\| \geq |a_n|(b-a)^n 2^{1-2n},$$

т.е. при фиксированном n выбор в качестве узлов интерполяции нулей многочлена Чебышева является оптимальным.

Следует отметить, что для системы узлов интерполяции $x_i = \cos\frac{2i-1}{2n}\pi$, $i = 1, \ldots, n$ (нули многочлена Чебышева $T_n(x)$), справедлива асимптотическая оценка сверху для константы Лебега $\lambda_n \leq K \ln n$ с постоянной K, не зависящей от n.

В качестве самостоятельного упражнения для многочлена $\omega_n(x)$ с равноотстоящими корнями на отрезке $[a, b]$ при $n \geq 2$, т.е.

$$x_i = a + (i-1)\frac{b-a}{n-1}, \quad i = 1, 2, \ldots, n,$$

предлагается получить оценку

$$\|\omega_n(x)\| \leq \frac{(b-a)^n (n-1)!}{4(n-1)^n}.$$

Лекция 3

Интерполяционные сплайны. Конструкция и обоснование кубического сплайна. — Понятие об аппроксимационных сплайнах.

Интерполяционные сплайны. Конструкция и обоснование кубического сплайна

Пусть на отрезке $[a,b]$ вещественной оси задана сетка: $a = x_0 < x_1 < \cdots < x_n = b$, $P_m(x)$ — множество многочленов степени не выше $m\,(m \geq 1)$, $C^{(r)}[a,b]$ — множество функций, имеющих на $[a,b]$ непрерывные производные до r–го порядка включительно $(r \geq 0)$.

Функцию $S_m(x) \equiv S_{m,k}(x)$ называют полиномиальным сплайном степени m дефекта k $(1 \leq k \leq m)$ с узлами $\{x_i\}$, $i = \overline{0,n}$ для функции $f(x) \in C[a,b]$, если выполнены следующие условия:

1) на каждом из отрезков $[x_i, x_{i+1}]$, $i = \overline{0,n-1}$ она является многочленом — $S_m(x) \in P_m(x)$;

2) на всем отрезке $[a,b]$ обладает непрерывностью производных — $S_m(x) \in C^{(m-k)}[a,b]$.

В дальнейшем термин "дефекта k" будет опускаться, так как далее рассматривается только случай $k = 1$.

Сплайн называется *интерполяционным*, если в узлах $\{x_i\}$ справедливы равенства $S_m(x_i) = f(x_i)$, $i = \overline{0,n}$.

Займемся построением **кубического** интерполяционного сплайна. Обозначим через M_i значения второй производной $S_3''(x)$ кубического интерполяционного сплайна в узлах $\{x_i\}$, $i = \overline{0,n}$, и будем для простоты считать, что все расстояния между узлами одинаковы, т.е. $x_i - x_{i-1} = h$.

Утверждение 1. *Величины $M_0, M_1, \ldots, M_n$ удовлетворяют системе линейных уравнений $C\,M = d$, где*

$$c_{ij} = \begin{cases} 1/6 & \text{при } j = i-1\,, j = i+1\,, \\ 2/3 & \text{при } j = i\,, \\ 0 & \text{при } |j-i| > 1\,; \end{cases} \qquad d_i = \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2}\,,$$

$$i, j = 1, 2, \ldots, n-1\,.$$

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: $S_3''(x) = M_{i-1}\dfrac{x_i - x}{h} + M_i\dfrac{x - x_{i-1}}{h}$. После двухкратного интегрирования и учета условий $S_3(x_i) = f_i\,, S_3(x_{i-1}) = f_{i-1}$ получим аналитическое представление кубического интерполяционного сплайна на отрезке $[x_{i-1}, x_i]$:

$$S_3(x) = M_{i-1}\frac{(x_i - x)^3}{6\,h} + M_i\frac{(x - x_{i-1})^3}{6\,h} +$$
$$+ \left(f_{i-1} - \frac{M_{i-1}h^2}{6}\right)\frac{x_i - x}{h} + \left(f_i - \frac{M_i h^2}{6}\right)\frac{x - x_{i-1}}{h}\,.$$

Теперь вычислим производную сплайна $S_3'(x)$ слева в точке x_i, воспользовавшись представлением на $[x_{i-1}, x_i]$:

$$S_3'(x_i - 0) = M_{i-1}\frac{h}{6} + M_i\frac{h}{3} + \frac{f_i - f_{i-1}}{h}\,,$$

и аналогичным образом — производную сплайна $S_3'(x)$ справа, в точке x_i, воспользовавшись представлением на $[x_i, x_{i+1}]$:

$$S_3'(x_i + 0) = -M_i \frac{h}{3} - M_{i+1} \frac{h}{6} + \frac{f_{i+1} - f_i}{h}\,.$$

Непрерывность $S_3'(x)$ в точках $x_i\,, i = \overline{0, n-1}$, т.е. $S_3'(x_i - 0) = S_3'(x_i + 0)$, после деления обеих частей на h порождает искомую систему уравнений. Утверждение доказано.

В рассмотренной системе число уравнений на два меньше числа неизвестных. Если наложены ограничения $M_0 = M_n = 0$, то сплайн называется *естественным*. Познакомимся с экстремальным свойством естественного сплайна. Рассмотрим множество функций

$$\Phi = \left\{ \varphi \in C^{(2)}[a, b] \,:\, \varphi(x_i) = f(x_i),\, i = 0, 1, \ldots, n \right\}$$

и *функционал энергии* на нем

$$E(\varphi) = \int\limits_a^b \left(\varphi''(x)\right)^2 dx\,.$$

Утверждение 2. *Естественный сплайн $S_3(x)$ доставляет минимум функционалу энергии на множестве Φ, т.е.*

$$E(\varphi) \geq E(S_3) \quad \forall \varphi \in \Phi,$$

причем неравенство строгое, если $\varphi \neq S_3$.

Доказательство. Рассмотрим равенство

$$y^2 - z^2 = (y-z)^2 + 2z(y-z) \quad \rightarrow \quad \left(\varphi''\right)^2 - \left(S_3''\right)^2 = \left(\varphi'' - S_3''\right)^2 + 2S_3''\left(\varphi'' - S_3''\right)\,.$$

Отсюда следует

$$E(\varphi) - E(S_3) = E(\varphi - S_3) + 2\int\limits_a^b S_3''\left(\varphi'' - S_3''\right) dx.$$

Интегрируя по частям, получаем

$$\int\limits_a^b S_3''\left(\varphi'' - S_3''\right) dx = S_3''\left(\varphi' - S_3'\right)\Big|_a^b - \int\limits_a^b S_3'''\left(\varphi' - S_3'\right) dx =$$

$$= -\sum_{i=1}^{n} \int\limits_{x_{i-1}}^{x_i} S_3'''\left(\varphi' - S_3'\right) dx = -\sum_{i=1}^{n} \left[S_3'''\left(\varphi - S_3\right)\Big|_{x_{i-1}}^{x_i} - \int\limits_{x_{i-1}}^{x_i} S_3^{(4)}\left(\varphi - S_3\right) dx \right] = 0\,.$$

Первое слагаемое обращается в нуль, так как $\varphi(x_i) = S_3(x_i)$, $i = 0, 1, \ldots, n$, а второе — из-за $S_3^{(4)} \equiv 0$ почти всюду. Таким образом,

$$E(\varphi) - E(S_3) = E(\varphi - S_3) \geq 0.$$

Утверждение доказано.

Утверждение 3. *В случае естественного сплайна решение системы* $C\,M = d$ *удовлетворяет неравенству*

$$\max_{1\le i\le n-1} |M_i| \le 3 \max_{1\le i\le n-1} |d_i|\,.$$

Доказательство. Пусть $\max\limits_i |M_i| = |M_j|\,, 1 \le j \le n-1$. Тогда рассмотрим j–е уравнение системы

$$d_j = M_{j-1}\frac{1}{6} + M_j\frac{2}{3} + M_{j+1}\frac{1}{6}\,,$$

из которого следует неравенство:

$$|d_j| \ge |M_j|\frac{2}{3} - \frac{1}{6}(|M_{j-1}| + |M_{j+1}|) \ge |M_j|\frac{1}{3}\,,$$

так как $|M_{j\pm1}| \le |M_j|$. Теперь, оценивая левую часть сверху через $\max\limits_i |d_i|$ приходим к искомому неравенству. Утверждение доказано.

Отметим, что из аналитической формулы и из полученной оценки следует существование и единственность естественного кубического интерполяционного сплайна, так как однородная система линейных уравнений имеет только тривиальное решение.

Рассмотрим вопрос о вычислительной устойчивости естественного кубического интерполяционного сплайна.

Утверждение 4. *Пусть на сетке с постоянным шагом* h *построены естественные сплайны* $S_3(x)$ *и* $S_3^*(x)$ *при использовании точных* f_i *и приближенных* f_i^* *значений функции:* $|f_i - f_i^*| \le \varepsilon$. *Покажем справедливость оценки*

$$\max_x |S_3(x) - S_3^*(x)| \le K\,\varepsilon\,, \quad K = 10\,.$$

Доказательство. Пусть $x \in [x_{i-1}, x_i]$, тогда, используя аналитическое представление сплайна, получим:

$$\max_{[x_{i-1},x_i]} |S_3(x) - S_3^*(x)| \le$$
$$\le |M_{i-1} - M_{i-1}^*|\frac{h^2}{3} + |M_i - M_i^*|\frac{h^2}{3} + |f_{i-1} - f_{i-1}^*| + |f_i - f_i^*|\,.$$

Для вектора разности $M - M^*$ справедливо уравнение с номером i:

$$C(M - M^*)|_i = d_i - d_i^* = \frac{1}{h^2}\left[f_{i+1} - 2\,f_i + f_{i-1} - (f_{i+1}^* - 2\,f_i^* + f_{i-1}^*)\right]\,.$$

Отсюда на основании полученного выше неравенства для коэффициентов M_i имеем оценку $\max\limits_i |M_i - M_i^*| \le \dfrac{12}{h^2}\,\varepsilon$. Поэтому справедливо

$$\max_{[x_{i-1},x_i]} |S_3(x) - S_3^*(x)| \le \frac{12}{h^2}\varepsilon\,\frac{2\,h^2}{3} + 2\,\varepsilon = 10\,\varepsilon\,.$$

Правая часть неравенства не зависит от рассматриваемого отрезка $[x_{i-1}, x_i]$, значит оно справедливо для $x_0 \le x \le x_n$. Утверждение доказано.

Полученная оценка означает, что возмущение сплайна пропорционально возмущению исходных данных с некоторой абсолютной постоянной, т.е. сплайн вычислительно устойчив.

Приведем без доказательства оценку погрешности для одного „неестественного“ сплайна. Пусть $f(x) \in C^{(4)}[a,b]$, $\max\limits_{[a,b]} |f^{(4)}(x)| \le A_4$, задана сетка с постоянным шагом $h_i = h$, и дополнительные условия для определения кубического интерполяционного сплайна имеют следующий вид:

$$S_3'(x_0 + 0) = f'(x_0), \quad S_3'(x_n - 0) = f'(x_n)\,.$$

Можно показать, что справедлива оценка погрешности $\forall x \in [x_0, x_n]$

$$|S_3^{(l)}(x) - f^{(l)}(x)| \le C_l A_4 h^{4-l}\,, \quad l = 0,1,2,3\,.$$

Понятие об аппроксимационных сплайнах.

Интерполяционные сплайны имеют значимый недостаток: при изменении даже одного значения f_i требуется пересчитывать заново все коэффициенты M_i. Чтобы избавиться от этого недостатка были введены аппроксимационные сплайны, коэффициенты которых строятся локально, т.е. на отрезке $[x_{i-1}, x_i]$ зависят только от нескольких ближайших значений f_i. В свою очередь, значения в узлах *локальных (аппроксимационных)* сплайнов, как правило, не совпадают со значениями $f(x)$. Однако это обстоятельство не носит принципиального характера, так как при вычислениях обычно используются приближенные значения функций.

Приведем построение локального сплайна третьей степени на сетке с постоянным шагом $h = x_{i+1} - x_i,\ i = \overline{0, n-1}$ для отрезка $[0,1]$. Рассмотрим *стандартный* сплайн $B(x)$, определяемый соотношениями

$$B(x) = \begin{cases} \dfrac{2}{3} - x^2 + \dfrac{1}{2}|x|^3 & \text{при } |x| \le 1\,, \\ \dfrac{1}{6}\,(2 - |x|)^3 & \text{при } 1 \le |x| \le 2\,, \\ 0 & \text{при } 2 \le |x|\,. \end{cases}$$

Отметим его полезные для дальнейшего изложения свойства:

$$\begin{aligned} &B(0) = \frac{2}{3}, && B(\pm 1) = \frac{1}{6}, \\ &B'(0) = 0, && B'(\pm 1) = \mp\frac{1}{2}, \\ &B''(0) = -2, && B''(\pm 1) = 1. \end{aligned}$$

Локальные сплайны третьей степени $B_2^{(1)}(x)$ и $B_2^{(2)}(x)$ записываются в виде

$$B_2^{(m)}(x) = \sum_{i=-1}^{n+1} \alpha_i^{(m)} B\left(\frac{x - ih}{h}\right), \quad m = 1,2$$

и отличаются выбором коэффициентов $\alpha_i^{(m)}$.

При $m = 1$ 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 $\alpha_i^{(1)} = f_i$ ($f_i = f(x_i)$) при $-1 \le i \le n+1$.

При $m = 2$ доопределяют значения f_{-2}, f_{-1} и f_{n+1}, f_{n+2} с помощью кубического интерполяционного многочлена 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 $\alpha_i = (8f_i - f_{i+1} - f_{i-1})/6$.

Значения полученных сплайнов в узлах сетки равны некоторому среднему значений функции в ближайших узлах.

Для погрешности приближения аппроксимационным сплайном $B_2^{(2)}(x)$ справедлива оценка, асимптотически такая же, как для интерполяционного сплайна: если $f(x) \in C^{(4)}[0,1]$, $|f^{(4)}(x)| \le A_4$, то $\forall x \in [0,1]$

$$\left| \left(B_2^{(2)}(x) \right)^{(l)} - f^{(l)}(x) \right| \le C_l A_4 h^{4-l}\,, \quad l = 0,1,2,3\,.$$

Ниже мы рассмотрим получение более простой оценки, справедливой для функций меньшей гладкости: пусть $f(x) \in C^{(2)}[0,1]$, $|f^{(2)}(x)| \le A_2$, тогда $\forall x \in [0,1]$ справедливо

$$\left| \left(B_2^{(1)}(x) \right)^{(l)} - f^{(l)}(x) \right| \le C_l A_2 h^{2-l}\,, \quad l = 0,1,2\,.$$

Будем использовать обозначение $B_2(x) = B_2^{(1)}(x)$, чтобы избежать недоразумений с символами производных.

Поточечное неравенство для второй производной сплайна. Рассмотрим выражение $B_2''(x)$ в одном из узлов $x_k = k\,h$:

$$B_2''(x_k) = \frac{1}{h^2}\left(\alpha_{k-1}B''(1) + \alpha_k B''(0) + \alpha_{k+1}B''(-1)\right) = \\ = \frac{\alpha_{k-1} - 2\,\alpha_k + \alpha_{k+1}}{h^2} = \frac{f_{k-1} - 2\,f_k + f_{k+1}}{h^2}\,.$$

Используя разложения в ряд Тейлора для величин $f_{k\pm1}$ в точке $x = x_k$ в предположении непрерывности $f''(x)$, получим $|B_2''(x_k)| \le A_2$. Более точно: $B_2''(x_k)$ совпадает со значением $f''(\xi_k)$, где $\xi_k \in [x_k - h, x_k + h]$.

Оценка для второй производной следует из поточечного неравенства. По определению $B_2''(x)$ на каждом отрезке $[x_{k-1}, x_k]$ является линейной функцией, а линейная функция принимает на концах отрезка экстремальные значения. Отсюда имеем $|B_2''(x)| \le A_2$. Так как вторые производные функции и сплайна ограничены одной постоянной A_2, то для их разности справедлива тривиальная оценка с $C_2 = 2$.

Оценка для первой производной. Рассмотрим на отрезке $[x_{k-1}, x_k]$ функцию $g(x)$, про которую известно:
1) $g'(x)$ — непрерывна, 2)$|g'(x)| \le K$. Тогда

$$g(x) = \int\limits_{x_{k-1}}^{x} g'(\xi)d\xi + g(x_{k-1}) \quad \text{и} \quad |g(x)| \le K\,h + |g(x_{k-1})|\,.$$

Рассмотрим сначала $g(x) = B_2'(x) - f'(x)$. Имеется оценка $|g'(x)| = |B_2''(x) - f''(x)| \le K = 2A_2$. Для нахождения недостающей величины $|g(x_{k-1})|$ рассмотрим значение $B_2'(x)$ в одном из узлов $x_k = k\,h$:

$$B_2'(x_k) = \frac{1}{h}\left(\alpha_{k-1}B'(1) + \alpha_k B'(0) + \alpha_{k+1}B'(-1)\right) = \frac{\alpha_{k+1} - \alpha_{k-1}}{2\,h} = \frac{f_{k+1} - f_{k-1}}{2\,h}.$$

Используя разложения в ряд Тейлора для величин $f_{k\pm1}$ в точке $x = x_k$ в предположении непрерывности $f''(x)$, получим $|B_2'(x_k) - f'(x_k)| \le \dfrac{h}{2} A_2$.

Теперь имеем искомую оценку:

$$\max_x |f'(x) - B_2'(x)| \le C_1 h\, A_2,\ C_1 = 5/2\,.$$

Оценка для разности получается таким же способом: $g(x) = B_2(x) - f(x)$. В данном случае $K = C_1 A_2 h$, а

$$B_2(x_k) = \alpha_{k-1}B(1) + \alpha_k B(0) + \alpha_{k+1}B(-1) = \frac{\alpha_{k+1} + 4\,\alpha_k + \alpha_{k-1}}{6} = \\ = \frac{f_{k+1} + 4\,f_k + f_{k-1}}{6} = f(x_k) + \tilde{C}_0 f''(\xi_k)\,h^2\,,$$

и это приводит к завершающей оценке для $l = 0$.

Оценку вычислительной устойчивости аппроксимационного сплайна предлагается получить самостоятельно.

Лекция 4

Наилучшее приближение в линейном нормированном пространстве. — Наилучшее приближение в гильбертовом пространстве.

Наилучшее приближение в линейном нормированном пространстве.

Пусть задан элемент f линейного нормированного пространства $\mathcal{L}$. Требуется найти его наилучшее приближение линейной комбинацией данных линейно независимых элементов $g_1, \ldots, g_n \in \mathcal{L}$. Это означает: найти элемент $x = \sum\limits_{j=1}^{n} c_j^0 g_j$ такой, что

$$\|f - x\| = \|f - \sum_{j=1}^{n} c_j^0 g_j\| = \inf_{c_1,\ldots,c_n} \|f - \sum_{j=1}^{n} c_j g_j\|.$$

Если x существует, то он называется *элементом наилучшего приближения.*

Утверждение 1. *Элемент наилучшего приближения существует.*

Доказательство. Рассмотрим следствие неравенства треугольника (модуль разности длин двух сторон меньше длины третьей, т.е. $|a - b| < c$):

$$\left| \|f - \sum_{j=1}^{n} c_j g_j\| - \|f - \sum_{j=1}^{n} \hat{c}_j g_j\| \right| \le \|\sum_{j=1}^{n} (c_j - \hat{c}_j) g_j\| \le \sum_{j=1}^{n} |c_j - \hat{c}_j| \, \|g_j\|.$$

Отсюда следует, что функция $F_f(c_1, \ldots, c_n) = \|f - \sum\limits_{j=1}^{n} c_j g_j\|$ является непрерывной функцией аргументов $c_1, \ldots, c_n$ при любой $f \in \mathcal{L}$ (в том числе, для $f \equiv 0$).

Пусть $|\mathbf{c}| = \sqrt{\sum\limits_{j=1}^{n} c_j^2}$ — евклидова норма вектора $\mathbf{c} = (c_1, \ldots, c_n)$. Функция $F_0(c_1, \ldots, c_n) = \|c_1 g_1 + \ldots + c_n g_n\|$ непрерывна на единичной сфере $|\mathbf{c}| = 1$ и, следовательно, в некоторой ее точке $(\tilde{c}_1, \ldots, \tilde{c}_n)$ достигает своей нижней грани $\tilde{F}$ по сфере, причем $\tilde{F} > 0$, так как равенство $\tilde{F} = \|\tilde{c}_1 g_1 + \ldots + \tilde{c}_n g_n\| = 0$ противоречит линейной независимости элементов $g_1, \ldots, g_n \in \mathcal{L}$. Для любого ненулевого вектора коэффициентов $\mathbf{c} = (c_1, \ldots, c_n)$ справедлива оценка

$$\|c_1 g_1 + \ldots + c_n g_n\| = F_0(c_1, \ldots, c_n) = |\mathbf{c}| F_0\left(\frac{c_1}{|\mathbf{c}|}, \ldots, \frac{c_n}{|\mathbf{c}|}\right) \ge |\mathbf{c}| \tilde{F}.$$

Рассмотрим в n-мерном пространстве коэффициентов шар $|\mathbf{c}| \le \gamma$ при $\gamma > 2\|f\|/\tilde{F}$. Функция $F_f(c_1, \ldots, c_n)$ непрерывна в этом шаре, следовательно, в некоторой точке шара $(c_1^0, \ldots, c_n^0)$ достигает своей нижней грани F^0 по шару. Имеем $F^0 \le F_f(0, \ldots, 0) = \|f\|$. Вне рассматриваемого шара выполняются соотношения

$$F_f(c_1, \ldots, c_n) \ge \|c_1 g_1 + \ldots + c_n g_n\| - \|f\| > \frac{2\|f\|}{\tilde{F}} \tilde{F} - \|f\| = \|f\| \ge F^0.$$

Таким образом, вне шара радиуса γ имеем

$$F_f(c_1,\dots,c_n) \geq F^0 = F_f(c_1^0,\dots,c_n^0)$$

при всех $c_1,\dots,c_n$. Утверждение доказано.

Элементов наилучшего приближения в общем случае может быть несколько. Единственность наилучшего приближения определяется либо свойствами приближаемого элемента, либо свойствами метрики.

Пространство $\mathcal{L}$ называется *строго нормированным*, если из условия

$$\|f+g\| = \|f\| + \|g\|, \quad \|f\|, \|g\| \neq 0$$

следует $f = \alpha g,\ \alpha \neq 0$.

Утверждение 2. *В случае строго нормированного пространства элемент наилучшего приближения определяется единственным образом.*

Доказательство. Введем обозначение $\Delta = \inf\limits_{c_1,\dots,c_n} \|f - \sum\limits_{j=1}^{n} c_j g_j\|$ и допустим, что существует два элемента x^1 и x^2 такие, что

$$x^1 \neq x^2, \quad x^i = \sum_{j=1}^{n} c_j^i g_j, \quad \|f - x^1\| = \|f - x^2\| = \Delta\,.$$

Ясно, что $\Delta > 0$, так как иначе $x^1 = f = x^2$. Кроме того, справедлива цепочка неравенств

$$\Delta \leq \left\| f - \frac{x^1+x^2}{2} \right\| = \left\| \frac{f-x^1}{2} + \frac{f-x^2}{2} \right\| \leq \left\| \frac{f-x^1}{2} \right\| + \left\| \frac{f-x^2}{2} \right\| = \Delta\,.$$

Отсюда следует, что элемент $\dfrac{x^1+x^2}{2}$ также является элементом наилучшего приближения. Вследствие предположения о строгой нормированности пространства имеем равенство $\dfrac{f-x^1}{2} = \alpha \dfrac{f-x^2}{2}$. Отсюда следует, что $|\alpha| = 1$. Если $\alpha = 1$, то $x^1 = f = x^2$ (противоречие). Если $\alpha \neq 1$, то имеем равенство

$$f = \frac{x^1 - \alpha x^2}{1-\alpha} = \sum_{j=1}^{n} \frac{c_j^1 - \alpha c_j^2}{1-\alpha} g_j,$$

которое означает, что f является линейной комбинацией $g_1,\dots,g_n$. т.е. $\Delta = 0$ (противоречие). Утверждение доказано.

Отметим, что пространство непрерывных функций на отрезке $[a,b]$ с нормой $\|f\| = \max\limits_{x\in[a,b]} |f(x)|$ не является строго нормированным.

Наилучшее приближение в гильбертовом пространстве

Если линейное нормированное пространство $\mathcal{L}$ является гильбертовым, т.е. на нем задано скалярное произведение $(x,y)\ \forall x,y \in \mathcal{L}$, а норма определена как $\|x\| = (x,x)^{1/2}$, то нахождение элемента наилучшего приближения весьма удобно.

Во-первых, такой элемент — единственный.

Утверждение 3. *Гильбертово пространство является строго нормированным.*

Доказательство. Если $x \neq 0$, $y \neq 0$, то возводя в квадрат равенство $\|x+y\| = \|x\| + \|y\|$, получим $(x,y) = \|x\|\|y\|$. Но в неравенстве Коши – Буняковского $(x,y) \leq \|x\|\,\|y\|$ равенство достигается только тогда, когда элементы x и y — линейно зависимы, т.е. $x = \alpha y$. Подставим это соотношение в полученное выше равенство, получим $\alpha = |\alpha| > 0$, т.е. $\alpha \neq 0$. Утверждение доказано.

Во-вторых, определение коэффициентов наилучшего приближения $c_1^0, \ldots, c_n^0$ сводится к решению системы линейных алгебраических уравнений. Функция

$$F_f^2(c_1, \ldots, c_n) = \left\| f - \sum_{j=1}^n c_j g_j \right\|^2 = \left(f - \sum_{j=1}^n c_j g_j, f - \sum_{j=1}^n c_j g_j \right)$$

является квадратичной относительно искомых коэффициентов, поэтому условия ее минимума $\dfrac{\partial F_f^2}{\partial c_i} = 0, \quad i = 1, \ldots, n$:

$$-2 \left(f - \sum_{j=1}^n c_j g_j, g_i \right) = 0, \quad i = 1, \ldots, n$$

удобно переписать в виде системы линейных алгебраических уравнений вида $A\,\mathbf{c} = \mathbf{b}$ —

$$\sum_{j=1}^n c_j (g_i, g_j) = (f, g_i), \quad i = 1, \ldots, n,$$

где $a_{ij} = (g_i, g_j), \quad b_i = (f, g_i)$. При этом симметричную вещественную матрицу A традиционно называют матрицей Грама.

Утверждение 4. *Если функции $g_1, \ldots, g_n$ линейно независимы, то матрица Грама A симметрична и положительно определена (т.е. $(A\mathbf{c}, \mathbf{c}) > 0 \quad \forall \mathbf{c} \neq 0$), и система $A\,\mathbf{c} = \mathbf{b}$ имеет единственное решение при любом векторе $\mathbf{b}$.*

Доказательство. Для линейно независимой системы имеем

$$(A\mathbf{c}, \mathbf{c}) = \sum_{i,j=1}^n (g_i, g_j) c_j c_i = \left(\sum_{i=1}^n c_i g_i, \sum_{j=1}^n c_j g_j \right) = \left\| \sum_{j=1}^n c_j g_j \right\|^2 > 0\,.$$

Пусть $\mathbf{x} \neq 0$ — произвольный собственный вектор матрицы A и λ — соответствующее ему собственное значение, т.е. $A\,\mathbf{x} = \lambda\,\mathbf{x}$. Имеем $0 < (A\mathbf{x}, \mathbf{x}) = \lambda \|\mathbf{x}\|^2$, т.е. $\lambda > 0$. Таким образом, все собственные значения матрицы A положительны, следовательно, $\det(A) \neq 0$, и система $A\,\mathbf{c} = \mathbf{b}$ имеет единственное решение при любом векторе $\mathbf{b}$. Утверждение доказано.

Если выбранная система ортонормальна $(g_i, g_j) = \delta_i^j$, то матрица A становится единичной и элемент наилучшего приближения имеет вид $x = \sum_{j=1}^n c_j g_j$, где $c_j = (f, g_j)$. При этом для вычислений удобно использовать равенство

$$\|f - x\|^2 = \left(f - \sum_{j=1}^n c_j g_j, f - \sum_{j=1}^n c_j g_j \right) = (f,f) - \sum_{j=1}^n |(f, g_j)|^2.$$

Так как $\|f - x\| \geq 0$, то из приведенного равенства, в том числе, следует известное неравенство Бесселя.

Рассмотрим полезный пример, приводящий к матрице Гильберта. Пусть в пространстве $\mathcal{L}$ вещественных функций, имеющих ограниченный интеграл $\int\limits_0^1 f^2(x)dx < \infty$ задано скалярное произведение $(f,g) = \int\limits_0^1 f(x)g(x)dx$ и порождаемая им норма. Выберем в качестве $g_1, \ldots, g_n$ систему многочленов $1, x, \ldots, x^{n-1}$ и попытаемся определить для некоторой $f \in \mathcal{L}$ элемент наилучшего приближения. В результате придем к системе линейных уравнений, матричные элементы которой имеют вид $a_{ij} = (x^{i-1}, x^{j-1}) = \int\limits_0^1 x^{i+j-2}dx = \frac{1}{i+j-1}$. Система $A\,\mathbf{c} = \mathbf{b}$ при любом n имеет единственное решение, однако ее решение сильно осложняется влиянием вычислительной погрешности (об этом будет позже — при решении линейных систем). В результате 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}, ... качество приближения будет только ухудшаться.

Лекция 5

Дискретное преобразование Фурье. — Идея быстрого дискретного преобразования Фурье.

Дискретное преобразование Фурье.

Пусть $f(x)$ — периодическая функция с периодом 1, т.е. $f(x) = f(x+1)\ \forall x$, разложена в ряд Фурье

$$f(x) = \sum_{q=-\infty}^{\infty} a_q \exp\{2\pi \mathbf{i} q x\}, \quad \text{причем} \quad \sum_{q=-\infty}^{\infty} |a_q| < \infty\,. \tag{1}$$

Здесь $\mathbf{i}$ — мнимая единица.

Рассмотрим значения этой функции на сетке из точек $x_l = l/N$, где $0 \le l \le N-1, N$ — целые, N фиксировано, и обозначим $f(x_l) = f_l$. Если $q_2 - q_1 = kN$, где k — целое, то $q_2 x_l - q_1 x_l = kN x_l = kl$ — также целое. Следовательно

$$\exp\{2\pi \mathbf{i} q_1 x_l\} = \exp\{2\pi \mathbf{i} q_2 x_l\}\,.$$

Поэтому если функция $f(x)$ рассматривается лишь в узлах сетки x_l, то в (1) можно привести подобные члены

$$f_l = \sum_{q=0}^{N-1} A_q \exp\{2\pi \mathbf{i} q x_l\}, \quad \text{где} \quad A_q = \sum_{s=-\infty}^{\infty} a_{q+sN}\,. \tag{2}$$

В отличие от традиционного преобразования Фурье (1) (отображения $f(x)$ в общем случае в бесконечное множество коэффициентов a_q) *дискретным преобразованием Фурье* (2) называется отображение вектора значений функции f_l, $l = 0, \ldots, N-1$ в вектор коэффициентов A_q, $q = 0, \ldots, N-1$ разложения по базисным векторам (функциям) $g_q(x_l) = \exp\{2\pi \mathbf{i} q x_l\}$.

Определим скалярное произведение функций на сетке следующим образом:

$$(f, g) = \frac{1}{N} \sum_{l=0}^{N-1} f_l \bar{g}_l$$

(множитель $1/N$ введен для аналогии с непрерывным случаем). Функции $g_q(x_l)$, $q = 0, \ldots, N-1$ образуют ортонормированную систему относительно введенного таким образом скалярного произведения.

Следующие выкладки для проверки ортонормированности рекомендуется провести самостоятельно. Действительно,

$$(g_q, g_j) = \frac{1}{N} \sum_{l=0}^{N-1} \exp\left\{2\pi \mathbf{i} \frac{q-j}{N} l\right\}\,.$$

При $q \ne j$, суммируя геометрическую прогрессию, имеем

$$(g_q, g_j) = \frac{1}{N} \frac{\exp\{2\pi \mathbf{i}(q-j)\} - 1}{\exp\left\{2\pi \mathbf{i} \frac{q-j}{N}\right\} - 1} = 0$$

(при $0 \le q, j \le N-1$, $q \ne j$ знаменатель отличен от нуля). Поскольку $(g_q, g_q) = 1$, то окончательно имеем $(g_q, g_j) = \delta_j^q$ при $0 \le q, j \le N-1$.

Вспомним построение наилучшего приближения в гильбертовом пространстве, когда базисные функции ортонормальны. В нашем частном случае это позволяет для коэффициентов дискретного преобразования Фурье после скалярного умножения первой суммы в (2) получить удобную формулу:

$$A_q = (f, g_q) = \frac{1}{N} \sum_{l=0}^{N-1} f_l \exp\{-2\pi \mathbf{i} q x_l\} .$$

Несложно заметить, что равенства не изменятся, если пределы суммирования $[0, N-1]$ в первой сумме (2) заменить на $[m, m+N-1$, где m — любое целое. Поэтому на практике используют представление следующего вида

$$f_l = \sum_{-N/2 < q \le N/2} A_q \exp\{2\pi \mathbf{i} q x_l\} . \tag{3}$$

Оно аргументируется так: если $f(x)$ — достаточно гладкая функция, то величины $|a_q|$ с ростом $|q|$ убывают быстро, поэтому $A_q \approx a_q$ при малых $|q|$; кроме того, при гладкой $f(x)$ величины A_q и a_q малы при больших $|q|$.

Однако следует иметь в виду, что следующее соотношение:

$$f(x) \approx \sum_{-N/2 < q \le N/2} A_q \exp\{2\pi \mathbf{i} q x\} \tag{4}$$

в общем случае, т.е. для непрерывного аргумента x, не имеет места. Например, пусть $f(x) = a_0 + a_{N-1} \exp\{2\pi \mathbf{i}(N-1)x\}$, тогда по формуле для A_q имеем: $A_0 = a_0,\ A_{-1} = a_{N-1}$. Сумма в правой части совпадает с $f(x)$ только в узлах сетки x_l, но далека от нее вне этих точек.

Тем не менее, способ аппроксимации (4) носит название *тригонометрической* интерполяции, а соотношение (3) (или (2)) называют *конечным* или *дискретным рядом Фурье*.

Упражнение. Пусть $f(x)$ непрерывно дифференцируема на $[0,1]$. Доказать, что

$$\max_{[0,1]} \left| f(x) - \sum_{-N/2 < q \le N/2} A_q \exp\{2\pi \mathbf{i} q x\} \right| \to 0$$

при $N \to \infty$.

Идея быстрого дискретного преобразования Фурье

Сами формулы быстрого преобразования достаточно громоздки, поэтому мы проследим только идею, лежащую в основе этого подхода.

Рассмотрим на отрезке $[0,1]$ сетку $x_l = l/N,\ 0 \le l \le N-1$. Пусть $f(x)$ — периодическая функция с периодом 1, заданная в узлах. Разложим ее в дискретный ряд Фурье (в силу конечномерности пространства он превращается в конечную сумму):

$$f(x_l) = \sum_{q=0}^{N-1} A_q g_q(x_l) , \quad g_q(x) = \exp\{2\pi \mathbf{i} q x\} ,$$

где коэффициенты вычисляются по формуле

$$A_q = (f, g_q) = \frac{1}{N} \sum_{l=0}^{N-1} f(x_l) \exp\{-2\pi \mathbf{i} q x_l\} \ .$$

Следует заметить, что вычисление каждого A_q требует по порядку $O(N)$ арифметических операций, $q \in [0, N-1]$, поэтому вычисление всех коэффициентов требует по порядку $O(N^2)$ арифметических операций.

Пусть $N = p_1 p_2$ $(p_1, p_2 \neq 1)$, представим индексы суммирования q и l в виде деления с остатком на p_1 и p_2 соответственно $q = q_1 + p_1 q_2, \quad l = l_2 + p_2 l_1$, где $0 \leq q_1, l_1 < p_1$, $\quad 0 \leq q_2, l_2 < p_2$. Используем эти выражения для преобразования суммы A_q :

$$A_q \equiv A(q_1, q_2) = \frac{1}{N} \sum_{l_1=0}^{p_1-1} \sum_{l_2=0}^{p_2-1} f_{l_2+p_2 l_1} \exp\left\{-2\pi \mathbf{i} \frac{l\, q}{N}\right\} \ .$$

Рассмотрим множитель lq/N у агумента функции exp

$$\frac{lq}{N} = \frac{(l_2 + p_2 l_1)(q_1 + p_1 q_2)}{p_1 p_2} = l_1 q_2 + \frac{q l_2}{N} + \frac{q_1 l_1}{p_1} \ .$$

На основании такого представления имеем равенство

$$\exp\left\{-2\pi \mathbf{i} \frac{lq}{N}\right\} = \exp\left\{-2\pi \mathbf{i} \frac{q l_2}{N}\right\} \exp\left\{-2\pi \mathbf{i} \frac{q_1 l_1}{p_1}\right\} \ .$$

Перепишем теперь выражение для $A(q_1, q_2)$ в виде

$$A(q_1, q_2) = \frac{1}{p_2} \sum_{l_2=0}^{p_2-1} B(q_1, l_2) \exp\left\{-2\pi \mathbf{i} \frac{q l_2}{N}\right\} \ ,$$

где использовано обозначение

$$B(q_1, l_2) = \frac{1}{p_1} \sum_{l_1=0}^{p_1-1} f_{l_2+l_1 p_2} \exp\left\{-2\pi \mathbf{i} \frac{q_1 l_1}{p_1}\right\} \ .$$

Двух последних формул уже достаточно для получения требуемой асимптотики. Для вычисления **всех** величин $B(q_1, l_2)$ требуется по порядку $O(p_2 p_1^2)$ действий, так как их всего $O(p_2 p_1)$ штук и для вычисления каждого необходимо провести $O(p_1)$ операций. Далее, при известных $B(q_1, l_2)$ для вычисления **всех** $A(q_1, q_2)$ аналогичным образом требуется по порядку $O(p_2^2 p_1)$ действий.

Если положить $p_i \approx \sqrt{N}$, $i = 1, 2$, то общее количество требуемых действий будет $O(N^{\frac{3}{2}})$, т.е. уже меньше, чем $O(N^2)$. Предположим, что $N = 2^p$, тогда сумму для A_q можно разбить не на две, а на p сумм, и проведение аналогичных преобразований приводит к асимптотике $O(Np) = O(N \log_2 N)$.

Для практических приложений выведены формулы быстрого преобразования Фурье для составных N вида $N = 2^p 3^q 5^r$. Этого оказывается вполне достаточно для большинства задач.

Лекция 6

Наилучшее равномерное приближение многочленами.

Наилучшее равномерное приближение многочленами

Пусть $\mathcal{L}$ — 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 $f \in \mathcal{L}$ отыскивается наилучшее приближение вида

$$Q_n^0(x) = \sum_{j=0}^{n} a_j^0\, x^j,$$

являющееся решением следующей задачи:

$$\inf_{a_j} \sup_{x\in[a,b]} |f(x) - \sum_{j=0}^{n} a_j\, x^j| \equiv \inf_{a_j} \|f - Q_n\| = \|f - Q_n^0(x)\|.$$

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 $\Delta \equiv \|f - Q_n^0\| \le \|f - Q_n\|$. 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 Валле–Пуссена. *Пусть существуют $(n+2)-$е точки $a \le x_0 < \ldots < x_{n+1} \le b$ и многочлен $Q_n(x)$ такие, что*

$$\operatorname{sign}\big(f(x_i) - Q_n(x_i)\big)\cdot(-1)^i = \text{const},$$

то есть при переходе от точки к точке разность $f(x_i) - Q_n(x_i)$ меняет знак. Тогда

$$\Delta = \|f - Q_n^0\| \ge \mu = \min_{i=0,\ldots n+1} |f(x_i) - Q_n(x_i)|.$$

Доказательство. Если $\mu = 0$, то утверждение очевидно. Пусть $\mu > 0$ и предположим противное: $\Delta = \|Q_n^0 - f\| < \mu$. Тогда в рассматриваемых точках x_i имеем

$$\operatorname{sign}\big(Q_n(x_i) - Q_n^0(x_i)\big) = \operatorname{sign}\big((Q_n(x_i) - f(x_i)) - (Q_n^0(x_i) - f(x_i))\big) = \operatorname{sign}\big(Q_n(x_i) - f(x_i)\big),$$

так как модуль первой разности не меньше μ, модуль второй — не превосходит Δ и, по предположению, $\mu > \Delta$. Таким образом, ненулевой многочлен $Q_n(x) - Q_n^0(x)$ степени n меняет знак в $n+2$ точках, что невозможно. Теорема доказана.

Теорема Чебышева. *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 чебышевского альтернанса.*

Доказательство достаточности. Обозначим $\|f - Q_n\|$ через L. Имеем по теореме Валле – Пуссена

$$L = \mu = \min_{i=0,\ldots,n+1} |f(x_i) - Q_n(x_i)| \le \Delta\,.$$

Однако $\Delta \le \|f - Q_n\| = L$. Отсюда $\Delta = L$.

Таким образом, имеющийся многочлен Q_n приближает f не хуже, чем наилучший Q_n^0. По определению Q_n^0 это означает, что Q_n является многочленом наилучшего равномерного приближения. Конец доказательства достаточности.

Пример. Для функции $f(x) = \sin 100x$ многочленом наилучшего равномерного приближения степени $n = 90$ на отрезке $[0, \pi]$ является $Q_{90}^0(x) \equiv 0$. В данном случае имеем точки чебышевского альтернанса $f(x_i) = (-1)^i$ для $x_i = (\pi/2 + \pi i)/100$, $i = 0, \ldots, 99$.

Утверждение. *Многочлен наилучшего равномерного приближения непрерывной функции единственен.*

Доказательство. Пусть $Q_n^1(x) \ne Q_n^2(x)$, $\|f - Q_n^1\| = \|f - Q_n^2\| = \Delta$. Тогда

$$\left\| f - \frac{Q_n^1 + Q_n^2}{2} \right\| \le \left\| \frac{f - Q_n^1}{2} \right\| + \left\| \frac{f - Q_n^2}{2} \right\| = \Delta.$$

Отсюда следует, что $Q_n = (Q_n^1 + Q_n^2)/2$ является многочленом наилучшего равномерного приближения, и для Q_n существуют точки чебышевского альтернанса $x_0 < \ldots < x_{n+1}$:

$$\left| \frac{Q_n^1(x_i) + Q_n^2(x_i)}{2} - f(x_i) \right| = \Delta, \quad i = 0, \ldots, n+1.$$

Но в этом случае

$$|Q_n^1(x_i) - f(x_i) + Q_n^2(x_i) - f(x_i)| = 2\Delta.$$

И так как $|Q_n^j(x_i) - f(x_i)| \le \max\limits_x |Q_n^j(x) - f(x)| = \Delta, \quad j = 1, 2$, то равенство возможно только при $Q_n^1(x_i) - f(x_i) = Q_n^2(x_i) - f(x_i) = \pm\Delta$. Это означает, что $Q_n^1(x_i) = Q_n^2(x_i)$ в $(n+2)$ точках, то есть многочлены Q_n^1 и Q_n^2 тождественно равны. Утверждение доказано.

Следствие. *Если $f(x)$ — непрерывная четная (нечетная) относительно $(a+b)/2$ функция, то $Q_n^0(x)$ четный (нечетный) относительно $(a+b)/2$ многочлен.*

Доказательство может быть получено с учетом единственности Q_n^0 методом от противного.

Пример. В классе разрывных функций при $n \ge 1$ теоремы Чебышева и единственности могут нарушаться:

$$f(x) = \operatorname{sign} x \text{ на } [-1, 1], \;\; Q_1(x) = \alpha\, x, \;\; \alpha \in [0, 2].$$

Лекция 7

Квадратурные формулы интерполяционного типа. — Ортогональные многочлены и квадратуры Гаусса.

Рассмотрим интеграл вида

$$I(f) = \int\limits_{\Omega} p(x)\, f(x)\, dx\,,$$

где Ω — конечный или бесконечный промежуток числовой оси и $f(x)$ — произвольная функция из некоторого класса F. В первую очередь под Ω будем иметь в виду отрезок $[a, b]$ и, если не оговаривается противное, то считать, что $f(x)$ непрерывна на нем. Заданная функция $p(x)$ называется *весовой*. Будем предполагать, что на $[a, b]$ она измерима, тождественно не равна нулю (как правило, почти всюду положительна) и ее произведение на любую $f(x) \in F$ суммируемо.

Для приближенного вычисления интеграла $I(f)$ строят линейные квадратурные формулы (*квадратуры*) следующего вида:

$$S_n(f) = \sum_{i=1}^{n} c_i\, f(x_i)\,.$$

Постоянные c_i называются *коэффициентами* (*весами*) квадратуры, x_i — ее *узлами*. Будем считать, что выполнено условие

$$I(1) = S_n(1), \quad \text{т.е.} \quad \sum_{i=1}^{n} c_i = \int\limits_{a}^{b} p(x)dx > 0\,.$$

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

$$R_n(F) = \sup_{f \in F} |R_n(f)|\,.$$

На практике часто используют оценки сверху для $|R_n(f)|$, которые будем обозначать через R_n.

Содержательной стороной численного интегрирования является построение квадратурных формул для заданного класса функций, вывод оценок погрешностей и анализ устойчивости формул при большом количестве слагаемых.

Квадратурные формулы интерполяционного типа

Имеется большая группа квадратурных формул, основанных на замене $f(x)$ алгебраическим интерполяционным многочленом. Пусть на отрезке $[a, b]$ по заданному набору различных узлов x_i, $1 \le i \le n$, функция $f(x)$ приближается интерполяционным многочленом Лагранжа $L_{n-1}(x)$ степени $n-1$

$$L_{n-1}(x) = \sum_{i=1}^{n} f(x_i)\Phi_i(x)\,, \quad \Phi_i(x) = \prod_{\substack{j=1 \\ j \ne i}}^{n} \frac{x - x_j}{x_i - x_j}\,.$$

Положим

$$S_n(f) = \int_a^b p(x)\, L_{n-1}(x)\, dx\,.$$

Отсюда получаем явные формулы для набора коэффициентов c_i, $1 \le i \le n$, и оценку погрешности R_n:

$$c_i = \int_a^b p(x)\, \Phi_i(x) dx, \quad R_n = \frac{\left\| f^{(n)}(x) \right\|}{n!} \int_a^b p(x)\, |\, \omega_n(x)|\, dx\,,$$

где

$$\left\| f^{(n)}(x) \right\| = \max_{[a,b]} \left| f^{(n)}(x) \right|\,, \quad \omega_n(x) = \prod_{i=1}^{n} (x - x_i)\,.$$

В оценках, приводимых ниже, также используется равномерная норма.

Определение. Квадратурная формула $S_n(f)$ имеет алгебраический порядок точности m, если справедливо равенство $S_n(P_m(x)) = I(P_m(x))$ для произвольного многочлена $P_m(x)$ степени не выше m, и найдется хотя бы один многочлен $Q_{m+1}(x)$ степени $m+1$, для которого выполнено неравенство

$$S_n(Q_{m+1}(x)) \ne I(Q_{m+1}(x))\,.$$

Утверждение 1. *Квадратурная формула $S_n(f)$ имеет алгебраический порядок точности $m \ge n-1$ тогда и только тогда, когда она является интерполяционной квадратурной формулой.*

Доказательство. Алгебраический порядок точности интерполяционной квадратурной формулы $S_n(f)$ не меньше, чем $n-1$, так как при ее построении мы использовали многочлен Лагранжа и, в силу оценки погрешности, справедливо тождество $\forall P_{n-1}(x) \equiv L_{n-1}(x)$. В другую сторону, если формула $S_n(f)$ точна для многочленов степени $n-1$, то используя равенства $S_n(\Phi_i(x)) = I(\Phi_i(x))$, $1 \le i \le n$, имеем $c_i = \int_a^b p(x)\Phi_i(x)dx$, $i = 1, 2, \ldots, n$. Последнее означает, что квадратурная формула является интерпроляционной. Утверждение доказано.

Квадратурные формулы интерполяционного типа, построенные в случае весовой функции $p(x) \equiv 1$ и системы равноотстоящих узлов, называются *формулами Ньютона – Котеса*.

Под системой из n равноотстоящих узлов на отрезке $[a, b]$ понимают следующее:

$$\begin{cases} \text{при} \quad n = 1: \quad x_1 = \dfrac{a+b}{2}\,, \\ \text{при} \quad n \ge 2: \quad x_i = a + (i-1)h,\ h = \dfrac{b-a}{n-1},\ i = 1, 2, \ldots, n. \end{cases}$$

Приведем наиболее употребительные формулы и оценки погрешности для них.

1. Формула прямоугольников: $n = 1$,

$$S_1(f) = (b-a) f\left(\frac{a+b}{2}\right), \quad R_1 = \|f'(x)\| \frac{(b-a)^2}{4}.$$

2. Формула трапеций: $n = 2$,

$$S_2(f) = \frac{b-a}{2}[f(a) + f(b)], \quad R_2 = \|f''(x)\|\frac{(b-a)^3}{12}.$$

3. Формула парабол (Симпсона): $n = 3$,

$$S_3(f) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{b+a}{2}\right) + f(b)\right], \quad R_3 = \|f^{(3)}(x)\|\frac{(b-a)^4}{192}.$$

Утверждение 2. *Для формул Ньютона – Котеса справедливо свойство симметрии коэффициентов:* $c_k = c_{n+1-k}\,\forall k = 1, 2, \ldots, n$.

Доказательство. Это свойство следует из явного вида фундаментальных многочленов

$$\Phi_k(x) = \prod_{\substack{j=1\\ j\neq k}}^{n} \frac{x - x_j}{x_k - x_j} \equiv \frac{\omega_n(x)}{(x - x_k)\omega_n'(x_k)}, \quad \omega_n(x) = \prod_{i=1}^{n}(x - x_i)\,.$$

Дело в том, что полином $\omega_n(x)$ и его производная $\omega_n'(x)$ имеют различную четность относительно середины отрезка $(a + b)/2$, поэтому функция под интегралом в правой части равенства

$$c_k - c_{n+1-k} = \int_a^b \omega_n(x)\left[\frac{1}{(x - x_k)\omega_n'(x_k)} - \frac{1}{(x - x_{n+1-k})\omega_n'(x_{n+1-k})}\right]dx$$

всегда является нечетной относительно середины отрезка — $(a + b)/2$, следовательно интеграл от нее равен нулю. Утверждение доказано.

Кроме того, симметрия узлов при **нечетных** n приводит к увеличению алгебраического порядка точности квадратурной формулы на единицу. На этом основан любопытный факт, что формула прямоугольников точнее, чем формула трапеций. Поскольку сравнение точности можно проводить только для функций из одного класса, необходимо получить для формулы прямоугольников несколько другую оценку погрешности. Для этого воспользуемся в качестве приближения к функции $f(x)$ отрезком ряда Тейлора в точке $(a + b)/2$:

$$f(x) = f\left(\frac{a+b}{2}\right) + f'\left(\frac{a+b}{2}\right)\left(x - \frac{a+b}{2}\right) + \frac{f''(\xi)}{2}\left(x - \frac{a+b}{2}\right)^2.$$

Тогда для квадратурной формулы $\widetilde{S}_1(f)$, полученной с помощью интегрирования двух первых слагаемых, справедливо равенство

$$\widetilde{S}_1(f) = \int_a^b \left[f\left(\frac{a+b}{2}\right) + f'\left(\frac{a+b}{2}\right)\left(x - \frac{a+b}{2}\right)\right]dx = S_1(f)\,,$$

при этом оценка погрешности принимает вид

$$\widetilde{R}_1 = \frac{\|f''(x)\|}{2}\int_a^b \left(x - \frac{a+b}{2}\right)^2 dx = \|f''(x)\|\frac{(b-a)^3}{24}\,.$$

Следовательно, на классе функций с непрерывной второй производной оценка погрешности формулы прямоугольников в два раза меньше, чем оценка погрешности формулы трапеций.

Отметим, что известны оценки погрешности через величины $\|f^{(n+1)}(x)\|$ и при других нечетных n. В частности, справедлива оценка для формулы Симпсона

$$\tilde{R}_3 = \left\|f^{(4)}(x)\right\| \frac{(b-a)^5}{2880}\,.$$

Основным недостатком формул Ньютона – Котеса является вычислительная неустойчивость, проявляющаяся в появлении отрицательных коэффициентов c_i. Сравним две величины $S_n(f)$ и $S_n(f^*)$, где $|f(x_i) - f^*(x_i)| \le \varepsilon$, $1 \le i \le n$. Будем иметь $|S_n(f) - S_n(f^*)| \le \varepsilon \sum\limits_{i=1}^{n} |c_i|$. Напомним, что $\sum\limits_{i=1}^{n} c_i = b-a$, т.е. для положительных коэффициентов квадратурной формулы сумма их модулей всегда ограничена длиной отрезка. С другой стороны, для формул Ньютона – Котеса известен факт

$$\lim_{n\to\infty} \left(\sum_{i=1}^{n} |c_i^{(n)}| \right) = \infty\,.$$

По этой причине на практике используют эти формулы только при малых $n\,(\le 7)$.

Ортогональные многочлены и квадратуры Гаусса

Рассмотрим следующую задачу: при заданном числе узлов $n \ge 1$ для вычисления интегралов вида

$$I(f) = \int\limits_a^b p(x) f(x)\,dx$$

построить квадратурную формулу

$$S_n(f) = \sum_{i=1}^{n} c_i f(x_i), \tag{1}$$

точную для многочленов максимально высокой степени. Весовая функция $p(x)$ предполагается почти всюду положительной без дополнительных оговорок.

В этой постановке имеется $2n$ свободных параметров (узлы x_i и коэффициенты c_i неизвестны), поэтому можно попытаться построить квадратуру, точную для многочленов степени $2n-1$.

Важную роль при построении формул Гаусса играют ортогональные многочлены на отрезке $[a,b]$ с весом $p(x) > 0$. Они могут быть получены, например, в результате стандартной процедуры ортогонализации примененной к системе $\{1, x, \dots, x^k, \dots\}$, при скалярном произведении

$$(f,g) = \int\limits_a^b p(x) f(x) g(x)\,dx\,.$$

Пусть на отрезке $[a,b]$ имеется система ортогональных многочленов с весом $p(x)$

$$\left\{\, 1, \psi_1(x), \psi_2(x), \,\ldots\,, \psi_k(x), \ldots \,\right\}.$$

Тогда многочлен k-й степени $\psi_k(x)$ ортогонален произвольному многочлену $P_l(x)$ при $l = 0, \ldots, k-1$. Действительно, многочлен $P_l(x)$ представим в виде $P_l(x) = \sum\limits_{j=0}^{l} \alpha_j \psi_j(x)$, и при $k \neq l$ имеют место равенства

$$\int\limits_a^b p(x)\psi_k(x)\psi_l(x)\,dx = 0\,.$$

На практике наиболее употребительны следующие ортогональные многочлены: Лежандра $\big([-1,1],\ p(x)=1\big)$, Чебышева $\left([-1,1],\ p(x)=\dfrac{1}{\sqrt{1-x^2}}\right)$, Лагерра $\big([0,\infty),\ p(x)=e^{-x}\big)$, Эрмита $\left((-\infty,\infty),\ p(x)=e^{-x^2}\right)$.

Рассмотрим пример построения многочленов Лежандра. Так как условия ортогональности определяют многочлен с точностью до постоянного множителя, будем считать, что коэффициент при старшем члене всегда равен единице. Поэтому $\psi_0(x) = 1$. Следующий многочлен будем искать в виде $\psi_1(x) = x + A$, а неизвестный коэффициент A определим из условия ортогональности

$$0 = (\psi_1, \psi_0) = \int\limits_{-1}^{1} (x + A) \cdot 1 dx = 2\,A\,,$$

т.е. $\psi_1(x) = x$. Для многочлена второй степени $\psi_2(x) = x^2 + Ax + B$ соотношения ортогональности имеют вид

$$\int\limits_{-1}^{1} (x^2 + Ax + B) \cdot 1 dx = 0\,, \quad \int\limits_{-1}^{1} (x^2 + Ax + B) x dx = 0\,.$$

Отсюда следует, что $\psi_2(x) = x^2 - 1/3$. Аналогично вычисляется $\psi_3(x) = x^3 + Ax^2 + Bx + C = x\left(x^2 - 3/5\right)$.

Рассмотрим общее свойство ортогональных многочленов.

Утверждение 3. *Ортогональный многочлен $\psi_n(x)$ имеет на интервале (a,b) ровно n различных корней.*

Доказательство. Пусть $\psi_n(x)$ имеет на (a,b) только $l < n$ нулей нечетной кратности. Обозначим их x_j и построим по ним многочлен $P_l(x) = \prod\limits_{j=1}^{l}(x - x_j)$. Тогда, в силу предположения, будет справедливо представление $\psi_n(x) = Q_{n-l}(x)\,P_l(x)$, где многочлен $Q_{n-l}(x)$ не меняет знака на (a,b). Так как $\psi_n(x)$ ортогонален произвольному многочлену меньшей степени, то справедливы равенства

$$0 = \int\limits_a^b p(x)\psi_n(x)P_l(x)dx = \int\limits_a^b p(x)P_l^2(x)Q_{n-l}(x)dx\,.$$

Но интеграл от знакопостоянной функции не может быть равен нулю, поэтому полученное противоречие завершает доказательство.

Перейдем к построению квадратур Гаусса.

Утверждение 4. *Для любого $n \geq 1$ существует единственная квадратурная формула Гаусса $S_n(f)$, которая имеет алгебраический порядок точности $m = 2n - 1$.*

Доказательство. Сначала покажем, что $m \leq 2n-1$. В этом легко убедиться от противного, взяв многочлен степени $2n$ следующего вида $P_{2n}(x) = (x - x_1)^2 \cdots (x - x_n)^2$, где x_i — узлы квадратурной формулы. Для такого многочлена имеем противоречие: $0 = S_n(P_{2n}) \neq I(P_{2n}) > 0$.

Проанализируем сначала необходимое свойство квадратурной формулы Гаусса. Пусть квадратурная формула $S_n(f)$ с узлами x_i, $i = 1, 2, \ldots n$, имеет порядок $m = 2n - 1$, тогда для многочлена $P_{2n-1}(x) = \psi_n(x)Q_{n-1}(x)$, где $\psi_n(x) = \prod_{i=1}^{n}(x - x_i)$, а $Q_{n-1}(x)$ — произвольный многочлен степени не выше $n - 1$, справедливо равенство

$$I(\psi_n(x)Q_{n-1}(x)) = S_n(\psi_n(x)Q_{n-1}(x)) = 0\,.$$

Следовательно, $\psi_n(x)$ — ортогональный многочлен на отрезке $[a, b]$ с весом $p(x)$. Мы знаем, что на (a, b) он имеет n различных корней, поэтому будем использовать их в качестве узлов квадратуры x_i. В этом случае коэффициенты c_i определяются по известным узлам однозначно (как в интерполяционных квадратурах: $c_i = \int_a^b p(x)\,\Phi_i(x)dx$).

Докажем теперь, что построенная таким образом квадратура действительно имеет алгебраический порядок точности $m = 2n - 1$. Возьмем произвольный многочлен

$$P_{2n-1}(x) = \psi_n(x)r_{n-1}(x) + q_{n-1}(x)\,.$$

Для него имеем

$$\begin{gathered} I(P_{2n-1}) = I(\psi_n r_{n-1} + q_{n-1}) = I(\psi_n r_{n-1}) + I(q_{n-1}) = \\ = I(q_{n-1}) = S_n(q_{n-1}) = S_n(q_{n-1}) + S_n(\psi_n r_{n-1}) = \\ = S_n(\psi_n r_{n-1} + q_{n-1}) = S_n(P_{2n-1})\,. \end{gathered}$$

В этой цепочке равенств мы последовательно использовали аддитивность интеграла, ортогональность $\psi_n(x)$ к $r_{n-1}(x)$, способ построения коэффициентов квадратуры (гарантирует точность $\geq n - 1$), выбор в качестве узлов квадратуры нулей $\psi_n(x)$ и аддитивность квадратурной формулы. Отметим, что существование формулы следует из явной конструкции, а ее единственность — это единственность ортогонального многочлена.

Утверждение доказано.

Приведем пример построения квадратуры Гаусса $S_3(f)$ на отрезке $[-1, 1]$ с весом $p(x) \equiv 1$. Ранее мы построили соответствующий ортогональный многочлен третьей степени $\psi_3(x) = x\left(x^2 - 3/5\right)$, его корни равны: $x_1 = -\sqrt{3/5}$, $x_2 = 0$, $x_3 = \sqrt{3/5}$. Далее методом неопределенных коэффициентов найдем недостающие веса, в результате получим формулу

$$S_3(f) = \frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right].$$

Несложно проверить, что ее алгебраический порядок точности равен $m = 2n - 1 = 5$.

Для сравнения приведем на том же отрезке формулу Симпсона $\bar{S}_3(f) = [f(-1) + 4f(0) + f(1)]/3$, порядок m которой при одинаковом количестве узлов равен 3.

Приведем без доказательства оценку погрешности для квадратур Гаусса

$$R_n = \|f^{(2n)}\| \int_a^b p(x) \frac{\psi_n^2(x)}{(2n)!} dx \,.$$

Лекция 8

Составные квадратурные формулы. — Правило Рунге для оценки погрешности. — Основные приемы для вычисления нерегулярных интегралов.

Составные квадратурные формулы

Рассмотрим задачи на построение составных квадратурных формул и вывод оценок их погрешностей. Пусть $h = (b-a)/N$ и $x_k = a + kh$, $k = 0, 1, \ldots, N$. Введем следующие обозначения: $I^{(k)}(f) = \int\limits_{x_k}^{x_{k+1}} p(x)f(x)dx$, $S_n^{(k)}(f) := S_n(f)$ для отрезка $[x_k, x_{k+1}]$, $k = 0, \ldots, N-1$. Исходный интеграл $I(f)$ равен $I(f) = \sum\limits_{k=0}^{N-1} I^{(k)}(f)$, соответствующая составная квадратурная формула принимает вид $S_n^N(f) = \sum\limits_{k=0}^{N-1} S_n^{(k)}(f)$, а для ее погрешности справедливо неравенство $|R_n^N(f)| \le \sum\limits_{k=0}^{N-1} |R_n^{(k)}(f)|$. Например, в случае составной формулы трапеций

$$S_2^N(f) = \frac{b-a}{N}\left[\frac{f(a)+f(b)}{2} + \sum_{k=1}^{N-1} f\left(a + \frac{b-a}{N}k\right)\right]$$

для погрешности на отрезке $[x_k, x_{k+1}]$ имеем неравенство

$$|R_2^{(k)}(f)| \le \|f''(x)\|_{[x_k,x_{k+1}]} \frac{(x_{k+1}-x_k)^3}{12} = \|f''(x)\|_{[x_k,x_{k+1}]} \frac{(b-a)^3}{12\,N^3}.$$

Следовательно, для всего отрезка $[a, b]$ оценка погрешности получается суммированием по всем $[x_k, x_{k+1}]$

$$R_2^N = \|f''(x)\|_{[a,b]} \frac{(b-a)^3}{12\,N^2}\,.$$

Заметим, что объем вычислительной работы при использовании составных квадратур растет линейно по N, а оценка погрешности убывает существенно быстрее. Причем скорость убывания погрешности напрямую зависит от алгебраического порядка точности формулы, поэтому применение составных квадратурных формул оправдано. Но такая стратегия разумна для функций с регулярным поведением производных, в противном случае исходный отрезок дробят так, чтобы величины погрешностей, приходящихся на каждый отрезок, были примерно одинаковы.

Правило Рунге для оценки погрешности

При разбиении отрезка на элементарные части важно учитывать поведение интегрируемой функции. Если же о ней заранее ничего не известно, то можно проводить разбиение постепенно шаг за шагом, двигаясь, например, слева направо. Для очередного элементарного отрезка длины h мы должны уметь оценивать погрешность и принимать решение об уменьшении или увеличении шага интегрирования.

Пусть на отрезке длины h используется некоторая квадратурная формула $S_h(f)$, точная для многочленов степени не выше $m-1$. Разлагая $f(x)$ в ряд Тейлора в середине отрезка (точке c), получим

$$I(f) - S_h(f) = D\, f^{(m)}(c) h^{m+1} + O(h^{m+2}), \quad D \neq 0.$$

Обозначим через $S_{h/2}(f)$ составную формулу, полученную с помощью формулы $S_h(f)$ для двух половинок отрезка длины h. Тогда при том же D находим

$$I(f) - S_{h/2}(f) = D\, f^{(m)}(c) \frac{h^{m+1}}{2^m} + O(h^{m+2}).$$

Следовательно, с точностью до членов порядка $O(h^{m+2})$ справедливо следующее правило Рунге:

$$I(f) - S_{h/2}(f) \approx \frac{S_{h/2}(f) - S_h(f)}{2^m - 1}.$$

Поэтому, если мы хотим найти $I(f)$ с абсолютной погрешностью ε на всем отрезке $[a, b]$, то каждый шаг h следует выбирать из условия

$$\left| \frac{S_{h/2}(f) - S_h(f)}{2^m - 1} \right| \leq \frac{h}{b-a}\, \varepsilon.$$

Основные приемы для вычисления нерегулярных интегралов

Пусть для вычисления интеграла $I(f) = \int\limits_a^b p(x) f(x) dx$ имеется некоторая квадратурная формула $S_n(f)$. Рассмотрим оценку погрешности

$$|I(f) - S_n(f)| \leq D \max_{[a,b]} |f^{(m)}(x)| (b-a)^{m+1}, \quad D \neq 0.$$

Она теряет всякий смысл в двух случаях: если по крайней мере один из пределов интегрирования равен бесконечности или $m-$я производная функции $f(x)$ не ограничена (не существует) на $[a, b]$. Это (формальная неприменимость оценки погрешности квадратурной формулы) и заложено в понятие нерегулярных интегралов. Познакомимся с наиболее типичными приемами избавления от нерегулярности на следующем примере (пример А.Н. Крылова)

$$I = \int\limits_0^\infty \frac{dx}{(x+1)\sqrt{x}}.$$

Первый прием — выделение бесконечности:

$$\int\limits_0^\infty \frac{dx}{(x+1)\sqrt{x}} = \int\limits_0^1 \frac{dx}{(x+1)\sqrt{x}} + \int\limits_1^\infty \frac{dx}{(x+1)\sqrt{x}}.$$

Второй прием — замена переменных: $x = \dfrac{1}{z}$, $dx = -\dfrac{dz}{z^2}$. Это дает

$$\int\limits_1^\infty \frac{dx}{(x+1)\sqrt{x}} = -\int\limits_1^0 \frac{dz}{z^2 \left(\dfrac{1}{z} + 1\right) \sqrt{\dfrac{1}{z}}} = \int\limits_0^1 \frac{dz}{(z+1)\sqrt{z}},$$

то есть

$$I = 2 \int\limits_0^1 \frac{dx}{(x+1)\sqrt{x}}\,.$$

Третий прием — интегрирование по частям (устранение особенности):

$$\int\limits_0^1 \frac{dx}{(x+1)\sqrt{x}} = 2\left[\frac{\sqrt{x}}{x+1}\Big|_0^1 + \int\limits_0^1 \frac{\sqrt{x}}{(x+1)^2}dx\right] = 1 + 2\int\limits_0^1 \frac{\sqrt{x}}{(x+1)^2}dx.$$

Окончательно имеем

$$I = 2 + 4\int\limits_0^1 \frac{\sqrt{x}}{(x+1)^2}dx\,.$$

Четвертый прием — выделение весовой функции. Примем

$$p(x) = \sqrt{x}, \quad f(x) = \frac{1}{(x+1)^2}\,.$$

Теперь отрезок интегрирования конечен и функция $f(x)$ имеет на этом отрезке ограниченные производные любого порядка, т.е. для приближенного вычисления интеграла можно использовать произвольную квадратурную формулу. Например, формулу Гаусса с весовой функцией $p(x) = \sqrt{x}$.

Заметим также, что величину этого интеграла можно найти точно: $I = \pi$, но это не умаляет достоинств приведенного примера.

Лекция 9

Интегрирование быстроосциллирующих функций. — Метод прогонки для решения трехдиагональных систем. — Корректность и устойчивость метода прогонки.

Интегрирование быстроосциллирующих функций

Пусть требуется вычислить интеграл

$$\int_a^b \exp\{\mathbf{i}\,\theta x\} f(x)dx, \qquad \text{где} \quad \theta(b-a) \gg 1,$$

а $f(x)$ — достаточно гладкая функция. Вещественная и мнимая компоненты подынтегральной функции — $\cos(\theta x)f(x)$ и $\sin(\theta x)f(x)$ — имеют на рассматриваемом отрезке примерно $\theta(b-a)/\pi$ нулей. Поскольку многочлен степени n имеет не более n нулей на этом отрезке, такие функции могут быть хорошо приближены многочленами степени n лишь при $n > \theta(b-a)/\pi$. Поэтому для непосредственного вычисления интегралов от таких функций потребуется применение квадратур, точных для многочленов очень высокой степени.

Более выгодным может оказаться использование $\exp\{\mathbf{i}\,\theta x\}$ в качестве **весовой** функции. Зададимся узлами интерполирования

$$x_j = \frac{b+a}{2} + \frac{b-a}{2} d_j\,, \quad j = 1, 2, \ldots, n\,, \quad d_j \in [-1, 1]\,,$$

построим многочлен Лагранжа $L_{n-1}(x)$ для $f(x)$, т.е.

$$f(x) = L_{n-1}(x) + \frac{f^{(n)}(\eta(x))}{n!}\omega_n(x),$$

и рассмотрим квадратурную формулу

$$S_n(f) \equiv \int_a^b \exp\{\mathbf{i}\,\theta x\} L_{n-1}(x)dx = \frac{b-a}{2} \exp\left\{\mathbf{i}\,\theta \frac{a+b}{2}\right\} \sum_{j=1}^{n} c_j\,(p)\, f(x_j),$$

где

$$p = \theta\frac{b-a}{2}, \quad c_j(p) = \int_{-1}^{1} \left(\prod_{k \neq j} \frac{\xi - d_k}{d_j - d_k} \right) \exp\{\mathbf{i}\,p\xi\} d\xi.$$

При этом оценка погрешности

$$R_n = D\,(d_1, \ldots, d_n) \max_{[a,b]} \left| f^{(n)}(x) \right| \left(\frac{b-a}{2} \right)^{n+1}$$

не зависит от θ.

Рассмотрим пример такой формулы для $n = 2$ с узлами $d_1 = -1,\ d_2 = 1$. После несложных вычислений получим

$$c_1(p) = \int_{-1}^{1} \frac{1-\xi}{2} \exp\{\mathbf{i}\,p\xi\} d\xi = \frac{\sin p}{p} + \frac{p\cos p - \sin p}{p^2}\mathbf{i}\,,$$

$$c_2(p) = \int\limits_{-1}^{1} \frac{1+\xi}{2} \exp\{\mathbf{i}\, p\xi\} d\xi = \frac{\sin p}{p} - \frac{p\cos p - \sin p}{p^2}\mathbf{i}\,.$$

Отметим, что при малых θ на фиксированном отрезке $[a, b]$ величина p мала. Функции $\cos p$ и $\sin p$ вычисляются с абсолютными погрешностями порядка ε и $p\,\varepsilon$ соответственно, где ε – машинная точность. Вследствие этого величины $c_1(p)$ и $c_2(p)$ приобретают погрешность порядка ε/p, т.е. такие формулы хороши при больших значениях p, а при малых — лучше пользоваться традиционными.

Метод прогонки для решения трехдиагональных систем

Пусть требуется найти решение системы уравнений:

$$\begin{array}{cc} c_0 y_0 - b_0 y_1 = f_0\,, & i = 0\,, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i\,, & 1 \le i \le N-1\,, \\ -a_N y_{N-1} + c_N y_N = f_N\,, & i = N\,, \end{array} \tag{1}$$

или в векторном виде

$$A\, y_h = f_h\,,$$

где $y_h = (y_0, y_1, \ldots, y_N)^T$ — искомый вектор неизвестных, $f_h = (f_0, f_1, \ldots, f_N)^T$ — заданный вектор правых частей, A — квадратная $(N+1)\times(N+1)$ матрица:

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

Основная идея метода состоит в представлении решения в виде

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}\,, \quad i = N-1, N-2, \ldots, 0\,, \tag{2}$$

для которого значения α_i, β_i и y_N необходимо вычислять по коэффициентам исходной системы и компонентам вектора правой части.

Перепишем первое уравнение (1) в виде (2):

$$y_0 = \alpha_1 y_1 + \beta_1\,, \quad \alpha_1 = b_0/c_0\,, \quad \beta_1 = f_0/c_0\,.$$

Затем к полученному соотношению добавим уравнение из (1) при $i = 1$:

$$\begin{array}{c} y_0 = \alpha_1 y_1 + \beta_1\,, \\ -a_1 y_0 + c_1 y_1 - b_1 y_2 = f_1\,. \end{array} \tag{3}$$

Из этой системы исключим переменную y_0:

$$(c_1 - a_1\alpha_1) y_1 - b_1 y_2 = f_1 + a_1\beta_1\,,$$

и перепишем полученное соотношение в виде (2):

$$y_1 = \alpha_2 y_2 + \beta_2\,, \quad \alpha_2 = \frac{b_1}{c_1 - a_1\alpha_1}\,, \quad \beta_2 = \frac{f_1 + a_1\beta_1}{c_1 - a_1\alpha_1}\,.$$

Следующий шаг подобен предыдущему: возьмем последнее соотношение и добавим к нему уравнение из (1) при $i = 2$:

$$\begin{gathered} y_1 = \alpha_2 y_2 + \beta_2 \,, \\ -a_2 y_1 + c_2 y_2 - b_2 y_3 = f_2 \,. \end{gathered}$$

Отличие этой пары уравнений от (3) состоит только в сдвиге индексов на единицу, поэтому результат шага можно написать сразу:

$$y_2 = \alpha_3 y_3 + \beta_3 \,, \quad \alpha_3 = \frac{b_2}{c_2 - a_2\alpha_2} \,, \quad \beta_3 = \frac{f_2 + a_2\beta_2}{c_2 - a_2\alpha_2} \,.$$

Таким образом, каждый раз добавляя к последнему полученному соотношению вида (2) следующее уравнение из системы (1), найдем формулы для вычисления α_i, β_i:

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i\alpha_i} \,, \quad \beta_{i+1} = \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i} \,.$$

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

$$\begin{gathered} y_{N-1} = \alpha_N y_N + \beta_N \,, \\ -a_N y_{N-1} + c_N y_N = f_N \,. \end{gathered}$$

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

$$y_N = \frac{f_N + a_N\beta_N}{c_N - a_N\alpha_N} \,(= \beta_{N+1}) \,.$$

Вывод формул закончен. Опишем алгоритм в целом.

Для решения системы (1) сначала рекуррентно вычисляются прогоночные коэффициенты α_i, β_i:

$$\begin{aligned} \alpha_1 &= b_0/c_0 \,, \quad \alpha_{i+1} = \frac{b_i}{c_i - a_i\alpha_i} \,, \\ \beta_1 &= f_0/c_0 \,, \quad \beta_{i+1} = \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i} \,, \end{aligned}$$

где i последовательно принимает значения $1, 2, \ldots, N-1$.

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:

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1} \,, \quad i = N-1, N-2, \ldots, 0 \,.$$

Полученные соотношения называют формулами *правой* прогонки. Формулы *левой* прогонки можно получить, проводя исключение неизвестных в обратном порядке. Рекомендуется это проделать в качестве самостоятельного упражнения.

Корректность и устойчивость метода прогонки

В приведенных выше формулах имеются два "узких"места: возможность обращения в нуль знаменателя (корректность) и вычислительная устойчивость. Поясним второй момент подробнее. Пусть единственная погрешность была допущена при вычислении y_N, а коэффициенты α_i, β_i найдены точно. Обозначим эту погрешность через ε_N, тогда погрешность в определении y_0 будет выражаться так:

$$\varepsilon_0 = (\alpha_1 \cdot \alpha_2 \cdot \ldots \cdot \alpha_N)\, \varepsilon_N \,.$$

Эта величина может быть очень большой ввиду большого количества сомножителей ($N \gg 1$). Поэтому говорят, что метод прогонки устойчив, если $\forall\, i \quad |\alpha_i| \le 1$. Справедливость этих неравенств зависит только от коэффициентов исходной системы. Существуют различные достаточные условия для корректности и устойчивости метода. Рассмотрим одно из них.

Утверждение 1. *Пусть коэффициенты системы уравнений* (1) *вещественны и удовлетворяют условиям:* c_0, c_N, *все* a_i, b_i *при* $i = 1, 2, \ldots, N-1$ *отличны от нуля и*

$$|c_i| \ge |a_i| + |b_i| \,, \quad i = 1, 2, \ldots, N-1 \,, \qquad |c_0| \ge |b_0| \,, \quad |c_N| \ge |a_N| \,,$$

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

$$c_i - a_i \alpha_i \ne 0 \,, \quad |\alpha_i| \le 1 \,, \quad i = 1, 2, \ldots, N \,,$$

гарантирующие корректность и устойчивость метода.

Доказательство. Рассуждения будем проводить по индукции. Из условия $|c_0| \ge |b_0|$ следует $|\alpha_1| \le 1$.

Пусть $|\alpha_i| \le 1$ для $i \le k$, тогда по неравенству треугольника имеем

$$|c_k - \alpha_k a_k| \ge |c_k| - |\alpha_k||a_k| \ge |a_k|(1 - |\alpha_k|) + |b_k| \ge |b_k| > 0 \,,$$

т.е. знаменатель не обращается в нуль. Кроме того, из этой оценки следует, что в формуле для α_{k+1} знаменатель не меньше чем числитель, следовательно $|\alpha_{k+1}| \le 1$. Требование хотя бы одного строгого неравенства объясняется следующим. Если все неравенства нестрогие, то знаменатель может обратиться в нуль. В примере

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

все $\alpha_i = 1$. Это приводит к равенству нулю знаменателя в определении последней компоненты решения y_4.

Условие $|a_i|$, $|b_i| \ne 0$ говорит о невозможности разложения исходной матрицы на две (или больше) трехдиагональные матрицы меньшей размерности, для каждой из которых применимо приведенное доказательство.

Утверждение доказано.

Отметим экономичность рассмотренного алгоритма. Для нахождения решения требуется проделать по порядку такое же количество арифметических операций, какова размерность вектора неизвестных, т.е. $O(N)$, причем постоянная в главном члене асимптотики не превышает 8.

Лекция 10

Прямые методы решения систем линейных уравнений. — Метод Гаусса. — Метод Холецкого.

Одной из наиболее часто решаемых задач на компьютере является система линейных алгебраических уравнений (СЛАУ): $A\,\mathbf{x} = \mathbf{b}$, $\det(A) \neq 0$, где A — заданная $n \times n$ – матрица, $\mathbf{b}$ — заданный вектор размерности n. Сначала мы рассмотрим некоторую группу методов, применяемых в случае матриц общего вида.

Определение. *Метод называется прямым (или точным), если в предположении отсутствия округления чисел он дает точное решение после конечного числа арифметических и логических операций.*

Мы, как и ранее, будем придерживаться схемы: задача → метод → оценка погрешности → учет влияния вычислительной погрешности. В данном случае (для всех прямых методов) оценка погрешности равна нулю, поэтому удобно рассмотреть сразу идеи всех методов, а затем изучить влияние вычислительной погрешности на точность получаемого решения.

Каждый из прямых методов базируется на разложении исходной матрицы в произведение более простых для обращения матриц и характеризуется асимптотикой количества арифметических действий относительно параметра n (размерности системы). Далее мы расмотрим наиболее часто используемые методы: Гаусса, Холецкого, вращений, отражений. Но сначала подчеркнем основную идею, присутствующую всюду: треугольная матрица — быстро и легко обратима. Напомним, что матрица R называется верхней (нижней) треугольной, если $r_{ij} = 0$ при $i > j$ $(i < j)$. Действительно, пусть имеется невырожденная система $R\,\mathbf{x} = \mathbf{c}$:

$$\begin{array}{rcccl} r_{11}x_1 + r_{12}x_2 & + \ldots + & r_{1n}x_n & = & c_1 \\ r_{22}x_2 & + \ldots + & r_{2n}x_n & = & c_2 \\ & & \ldots & \ldots & \ldots \\ & & r_{nn}x_n & = & c_n \end{array}$$

Из последнего уравнения находим x_n за одно действие. Подставляем полученное значение в предыдущее уравнение (1 умножение и 1 вычитание из правой части). В результате опять имеем уравнение с одним неизвестным x_{n-1}. Находим его за одно деление. Далее процедура использования уже найденных компонент вектора $\mathbf{x}$ повторяется. Если требуется решить систему с нижней треугольной матрицей, то используется та же процедура, начиная с определения компоненты x_1. Посчитаем количество арифметических действий

$$1 + (1 + 2) + (1 + 4) + \cdots + (1 + 2(n - 1)) = 2n \cdot n/2 = n^2.$$

Отметим, что процедура умножения квадратной матрицы $n-$го порядка на вектор требует $2n^2 - n$ арифметических действий, поэтому в прямых методах затраты порядка $O(n^2)$ считаются незначительными. Также следует иметь в виду, что построение используемых в прямых методах матричных разложений требует по порядку $O(n^3)$ действий. При этом вычислительные затраты в методах отличаются только постоянными в главном члене асимптотики. Перейдем к последовательному изложению алгоритмов.

Метод Гаусса

Сначала рассмотрим хорошо известный метод Гаусса решения линейных невырожденных систем. Он основан на разложении $A = L\,R$, где L — нижняя треугольная матрица с $l_{ii} = 1$, а R — верхняя треугольная матрица. При этом матрицу L, как правило, в явном виде не вычисляют (она оказывается полезной только при многократном решении одной и той же системы с различными правыми частями). Требуемое число арифметических действий: $2n^3/3 + O(n^2)$ (здесь и далее будем приводить константу главного члена асимптотики). Будем предполагать исходную матрицу *строго регулярной*, т.е. имеющей все ведущие подматрицы невырожденными. Тогда процедуру последовательного исключения неизвестных, т.е. приведение исходной системы к треугольному виду, можно описать следующим образом. Обозначим исходную матрицу $A = A^{(0)}$. Первый шаг. Первое уравнение системы

$$\sum_{j=1}^{n} a_{1j} x_j = b_1$$

не трогаем. Зато из каждого последующего вычитаем первое, умноженное соответственно на коэффициент $m_{i1} = a_{i1}/a_{11},\ i = 2, \ldots, n$. Это эквивалентно умножению системы на невырожденную матрицу $M^{(1)}$, имеющую начиная со второй строки только по два ненулевых элемента:

$$M^{(1)} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -m_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -m_{n1} & 0 & \ldots & 1 \end{pmatrix}.$$

Теперь в новой системе $A^{(1)}\mathbf{x} \equiv M^{(1)}A^{(0)}\mathbf{x} = M^{(1)}\mathbf{b}$ первая компонента вектора неизвестных осталась только в первом уравнении. На втором шаге оставляем без изменения два первых уравнения системы с матрицей $A^{(1)}$, на третьем — три с $A^{(2)}$ и т.д. Этот процесс запишем в виде: $M^{(r)}A^{(r-1)} = A^{(r)}$ для $2 \le r \le n-1$, где

$$M^{(r)} = \begin{pmatrix} I_{r-1,r-1} & & 0 & & \\ & & 1 & \ldots & 0 \\ & 0 & \vdots & \ddots & \vdots \\ & & -m_{nr} & \ldots & 1 \end{pmatrix},\ m_{ir} = \frac{a_{ir}^{(r-1)}}{a_{rr}^{(r-1)}},\quad i \ge r+1.$$

Здесь через $I_{r-1,r-1}$ обозначена единичная матрица $r-1$-го порядка. В результате мы получаем

$$M^{(n-1)}\ldots M^{(1)} A\,\mathbf{x} = M^{(n-1)}\ldots M^{(1)}\,\mathbf{b}, \qquad \text{или } R\,\mathbf{x} = \mathbf{c},$$

и задача сводится к обращению верхней треугольной матрицы R. Отметим, что $\mathbf{c} = L^{-1}\mathbf{b}$, $L^{-1} = \prod_{r=1}^{n-1} M^{(r)}$, так как произведение нижних треугольных матриц есть нижняя треугольная матрица и обратная к нижней треугольной также является нижней треугольной матрицей.

При практической реализации алгоритма для вычислительной устойчивости необходимо переставлять стороки и/или столбцы матрицы так, чтобы

ведущий элемент (на который происходит деление) был максимальным по модулю. В противном случае влияние вычислительной погрешности может существенно исказить решение задачи.

Метод Холецкого

Если матрица системы A симметрична и положительно определена ($A = A^T > 0$), то используется метод Холецкого, также называемый методом квадратного корня, который требует асимптотически в два раза меньше действий, чем метод Гаусса: $n^3/3+O(n^2)$. Несложно выписать формулы для решения системы $A\mathbf{x} = \mathbf{b}$, основанные на разложении $A = R^T R$ с верхней треугольной матрицей R.

Сначала определим элементы матрицы R. В силу формулы умножения матриц

$$a_{ij} = \sum_{k=1}^{n} \left(R^T\right)_{ik} (R)_{kj} = \sum_{k=1}^{n} r_{ki} r_{kj},$$

имеем

$$\begin{gathered} a_{ij} = r_{1i} r_{1j} + r_{2i} r_{2j} + \ldots + r_{ii} r_{ij} \quad \text{при} \quad i < j\,, \\ a_{ii} = r_{1i}^2 + r_{2i}^2 + \ldots + r_{ii}^2 \quad \text{при} \quad i = j\,. \end{gathered}$$

Отсюда следуют формулы для определения r_{ij} :

$$\begin{gathered} r_{11} = \sqrt{a_{11}}, \quad r_{1j} = \frac{a_{1j}}{r_{11}} \quad (j > 1), \\ r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2} \quad (i > 1), \quad r_{ij} = \frac{a_{ij} - \sum\limits_{k=1}^{i-1} r_{ki} r_{kj}}{r_{ii}} \quad (j > i), \\ r_{ij} = 0 \quad (i > j). \end{gathered}$$

Далее решение исходной системы сводится к последовательному решению двух систем с треугольными матрицами:

$$R^T \mathbf{y} = \mathbf{b} \quad \text{и} \quad R\mathbf{x} = \mathbf{y}\,.$$

Лекция 11

Метод отражений. — Метод вращений. — Число обусловленности. Неравенства для ошибки и невязки.

Метод отражений

Среди точных методов, требующих для реализации порядка $O(n^3)$ действий, одним из наиболее устойчивых к вычислительной погрешности является метод отражений (у него константа при старшем члене асимптотики равна 4/3), основанный на матричном разложении $A = VR$, где V — некоторая ортогональная матрица ($V^{-1} = V^T$).

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

Легко проверить, что матрица U является симметричной и ортогональной матрицей, т.е. $U = U^T$ и $UU^T = I$, и все ее собственные значения равны ± 1.

Симметричность следует из явного вида элементов U: $u_{ij} = \delta_i^j - 2w_i w_j$. Так как $(\mathbf{w}, \mathbf{w}) = 1$, следовательно $\Omega\Omega|_{ij} = \sum\limits_{k=1}^{n} w_i w_k w_k w_j = w_i w_j \sum\limits_{k=1}^{n} w_k^2 = \Omega|_{ij}$ и $UU = I - 4\Omega + 4\Omega\Omega = I$, т.е. $U^2 = UU^T = I$.

Дополнительно матрица U обладает свойствами: $U\mathbf{w} = -\mathbf{w}$; а если вектор $\mathbf{v}$ ортогонален $\mathbf{w}$, тогда $U\mathbf{v} = \mathbf{v}$. Отсюда следует, что образ $U\mathbf{y}$ произвольного вектора $\mathbf{y}$ является зеркальным отражением относительно гиперплоскости, ортогональной вектору $\mathbf{w}$. Действительно, представим $\mathbf{y}$ в виде $\mathbf{y} = (\mathbf{y}, \mathbf{w})\mathbf{w} + \mathbf{v}$. Тогда из указанных свойств U следует $U\mathbf{y} = -(\mathbf{y}, \mathbf{w})\mathbf{w} + \mathbf{v}$.

Это дает возможность решить вспомогательную задачу: для векторов единичной длины $\mathbf{y}$ и $\mathbf{e}$ найти вектор единичной длины $\mathbf{w}$ такой, что $U\mathbf{y} = \mathbf{e}$, где $U = I - 2\mathbf{w}\,\mathbf{w}^T$.

Приведем формулу искомого решения — $\mathbf{w} = \pm(\mathbf{y} - \mathbf{e})/\sqrt{(\mathbf{y} - \mathbf{e}, \mathbf{y} - \mathbf{e})}$. Действительно, $U\mathbf{y} = \left(I - 2\mathbf{w}\,\mathbf{w}^T\right)\mathbf{y} = \mathbf{y} - \xi = \mathbf{e}$, 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 так как преобразование U не меняет длины вектора, следовательно для вектора произвольной ненулевой длины $\mathbf{y}$ имеем $U\mathbf{y} = \alpha\mathbf{e}$, $\alpha = \|\mathbf{y}\|_2$ и решением тогда является вектор $\mathbf{w} = \pm\dfrac{(\mathbf{y} - \alpha\mathbf{e})}{\|\mathbf{y} - \alpha\mathbf{e}\|_2}$.

Теперь изложим собственно метод отражений, т.е. покажем, что произвольная квадратная матрица A может быть приведена к верхнетреугольному виду в результате последовательного умножения слева на ортогональные матрицы.

По векторам $\mathbf{y}_1 = (a_{1,1}, \ldots, a_{n,1})^T$ и $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ можно построить матрицу U_1 (см. вспомогательную задачу), так чтобы первый столбец матрицы $A^{(1)} = U_1 A$ был пропорционален вектору $\mathbf{e}_1$, т.е. $U_1\mathbf{y}_1 = \pm\alpha_1\mathbf{e}_1$, где знак плюс или минус определяется величиной $\operatorname{sign}(a_{1,1})$.

Вектор $\mathbf{w}_1$ определяется так: сначала конструируется верктор $\tilde{\mathbf{w}}_1 = (a_{1,1} + \mathrm{sign}(a_{1,1})\alpha_1, a_{2,1}, \ldots, a_{n,1})$, где $\alpha_1 = \|\mathbf{y}_1\|_2$, а затем производится его нормировка — $\mathbf{w}_1 = \tilde{\mathbf{w}}_1/\|\tilde{\mathbf{w}}_1\|_2$. Такой выбор знака гарантирует корректность алгоритма и малость вычислительной погрешности. В этом легко убедиться выбирая для проверки в качестве $\mathbf{y}_1$ векторы: $(1, 0, \ldots, 0)^T$ и $(-1, 0, \ldots, 0)^T$.

Далее, в пространстве 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 $\mathbf{e}_2 = (1, 0, \ldots, 0)^T \in R^{n-1}$. Затем определяется $U_2 = \begin{pmatrix} 1 & 0 \\ 0 & U_2' \end{pmatrix}$ и рассматривается матрица $A^{(2)} = U_2 U_1 A$.

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 $\mathbf{w}_k = \tilde{\mathbf{w}}_k/\|\tilde{\mathbf{w}}_k\|_2$, $\mathbf{w}_k \in R^n$, где $\alpha_k = (a_{k,k}^2 + a_{k+1,k}^2 + \ldots + a_{n,k}^2)^{1/2}$ — параметр для $\tilde{\mathbf{w}}_k = (0, \ldots, 0, a_{k,k} + \mathrm{sign}(a_{k,k})\alpha_k, a_{k+1,k}, \ldots, a_{n,k})^T$.

В результате преобразований получится верхнетреугольная матрица $R = \hat{U}A$, где $\hat{U} = U_{n-1} \ldots U_1$. Для завершения изложения остается положить $V = \hat{U}^T$.

Отметим, что при практической реализации явное вычисление U_k не требуется, так как $U_k A^{(k-1)} = A^{(k-1)} - 2\mathbf{w}_k \left(\mathbf{w}_k^T A^{(k-1)}\right)$. При этом изменяются только элементы $a_{ij}^{(k-1)}$, $k \le i, j \le n$ матрицы $A^{(k-1)}$.

Метод вращений

Имеется другой, также очень популярный, способ приведения матрицы A к виду $A = QR$, где $Q^{-1} = Q^T$ — ортогональная матрица. Он называется методом вращений и требует $2\,n^3 + O(n^2)$ арифметических операций, зато обладает большей простотой.

Рассмотрим элементарную матрицу вращения второго порядка (матрицу Гивенса):

$$G(\varphi) = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix},$$

зависящую от некоторого параметра — угла φ. Если двумерный вектор $(a_1, a_2)^T$ ненулевой, то выбор

$$\cos\varphi = \frac{a_1}{\sqrt{a_1^2 + a_2^2}}, \quad \sin\varphi = \frac{-a_2}{\sqrt{a_1^2 + a_2^2}}$$

позволяет обнулить его вторую компоненту

$$G(\varphi)\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sqrt{a_1^2 + a_2^2} \\ 0 \end{pmatrix}.$$

Отсюда следует, что при умножении матрицы A слева на матрицу G_{kl}

$$G_{kl}(\varphi) = \begin{pmatrix} 1 & 0 & \ldots & 0 & \ldots \\ 0 & \cos\varphi & 0 & -\sin\varphi & 0 \\ \vdots & 0 & 1 & 0 & \vdots \\ 0 & \sin\varphi & 0 & \cos\varphi & 0 \\ 0 & & \ldots & 0 & 1 \end{pmatrix},$$

($g_{kl} = \sin\varphi$, т.е. синусы и косинусы находятся на пересечении строк и столбцов с номерами k и l) можно получить нуль на пересечении k-й строки и l-го столбца. Следовательно, матрицу A можно привести к верхнему треугольному виду R с помощью умножений слева на последовательность матриц вращения:

$$G_{n\,n-1}\dots G_{3\,2}G_{n\,1}\dots G_{3\,1}G_{2\,1}A = R\,,$$

или, что то же самое —

$$Q^T = G_{n\,n-1}\dots G_{3\,2}G_{n\,1}\dots G_{3\,1}G_{2\,1}\,,\quad A = QR\,.$$

Отметим, что алгоритмы, использующие ортогональные матрицы более устойчивы к ошибкам округлений по сравнению с методом Гаусса, так как умножение на них не меняет евклидову длину вектора.

Число обусловленности. Неравенства для ошибки и невязки

Напомним распространенные векторные нормы и подчиненные им матричные нормы ($\|A\| = \sup\limits_{\mathbf{x}\neq 0}\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$):

$$\|\mathbf{x}\|_\infty = \max_i |x_i|,\quad \|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|,\quad \|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|,$$
$$\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|,\quad \|\mathbf{x}\|_2^2 = \sum_{i=1}^n x_i^2,\quad \|A\|_2 = \sqrt{\max_i \lambda_i(A^TA)}.$$

Отметим, что таким образом введенные нормы удовлетворяют неравенствам:

$$\|AB\|_* \le \|A\|_*\|B\|_*,\quad \|A\mathbf{x}\|_* \le \|A\|_*\|\mathbf{x}\|_*.$$

В дальнейшем, если это специально не оговаривается, в лекциях всюду будут использоваться только подчиненные матричные нормы.

Удобной характеристикой для оценки влияния вычислительной погрешности является *число обусловленности*:

$$\mathrm{cond}(A) = \|A\|\cdot\|A^{-1}\|\quad (\det(A)\neq 0)\,.$$

Перейдем к получению двух основных неравенств (Уилкинсона).

Неравенство для ошибки. Приближенное решение $\mathbf{x}^*$ системы уравнений $A\mathbf{x} = \mathbf{b}$ можно трактовать как точное для задачи $A\mathbf{x}^* = \mathbf{b}^*$, где $\mathbf{x}^* = \mathbf{x} + \Delta\mathbf{x}$, $\mathbf{b}^* = \mathbf{b} + \Delta\mathbf{b}$. Отсюда получим:

$$\|\mathbf{b}\| \le \|A\|\|\mathbf{x}\|,\qquad \|\Delta\mathbf{x}\| \le \|A^{-1}\|\|\Delta\mathbf{b}\|.$$

Из этих неравенств следует искомое

$$\frac{\|\mathbf{x}^* - \mathbf{x}\|}{\|\mathbf{x}\|} \le \mathrm{cond}(A)\frac{\|\mathbf{b}^* - \mathbf{b}\|}{\|\mathbf{b}\|}\,.$$

Множитель при числе обусловленности есть возмущение правой части исходной системы, которое может быть вызвана просто машинным представлением компонент вектора $\mathbf{b}$. В этом случае оно определяется машинной точностью (см. лекцию 1), и устранить этот эффект невозможно. Относительная же погрешность полученного решения зависит от свойств матрицы

решаемой системы и может быть сколь угодно велика. Отметим, что здесь совершенно не важен метод нахождения $\mathbf{x}^*$, так как неравенство носит общий характер.

Ярким примером "плохой"матрицы является матрица Гильберта H_n с элементами $h_{ij} = (i + j - 1)^{-1}$, $1 \le i, j \le n$. Ее число обусловленности растет экспоненциально $\mathrm{cond}_2(H_n) = \exp(æ_n \cdot n)$, с показателем $æ_n \approx 3$. Это означает, что уже при $n = 15$ и вычислениях с двойной точностью ($\varepsilon \approx 10^{-16}$) относительная погрешность решения, полученного, например, методом Гаусса, становится больше единицы. Другими словами, приближенное решение не имеет ничего общего с точным.

Неравенство для невязки. Приближенное решение системы уравнений $A\mathbf{x} = \mathbf{b}$ допускает и другую интерпретацию: его можно трактовать как точное для задачи с возмущенной матрицей $A_*\mathbf{x}_* = \mathbf{b}$, где $A_* = A + \Delta A$. Определим вектор невязки как $\mathbf{r} = \mathbf{b} - A\mathbf{x}_*$, тогда $\Delta A\mathbf{x}_* = \mathbf{r}$. Отсюда следует неравенство $\|\mathbf{r}\| \le \|\Delta A\|\|\mathbf{x}_*\|$. Поделим обе его части для нормировки на величину $\|A\|\|\mathbf{x}_*\|$, в результате получим

$$\frac{\|A\mathbf{x}_* - \mathbf{b}\|}{\|A\|\|\mathbf{x}_*\|} \le \frac{\|A_* - A\|}{\|A\|}.$$

Величина в правой части по порядку совпадает с машинной точностью, значит, при решении прямым методом (любым!) нормированная указанным образом невязка всегда мала.

Хорошие программы, реализующие прямые методы, отличаются тем, что одновременно с нахождением решения они оценивают число обусловленности матрицы.

Лекция 12

Метод простой итерации решения систем линейных уравнений. — Оптимальный одношаговый итерационный метод. — Оптимальный циклический итерационный метод.

Метод простой итерации решения систем линейных уравнений

Преобразуем систему линейных алгебраических уравнений

$$A\,\mathbf{x} = \mathbf{b} \tag{1}$$

с невырожденной матрицей A к виду

$$\mathbf{x} = G\,\mathbf{x} + \mathbf{c}. \tag{2}$$

Если решение системы (2) находится как предел последовательности

$$\mathbf{x}^{k+1} = G\,\mathbf{x}^k + \mathbf{c}, \tag{3}$$

то такой процесс называется *двухслойным итерационным методом*, или *методом простой итерации*. При этом G называется *оператором перехода*.

Рассмотрим общий способ перехода от системы (1) к системе (2). Всякая система

$$\mathbf{x} = \mathbf{x} - D\,(A\,\mathbf{x} - \mathbf{b}) \tag{4}$$

имеет вид (2) и при $\det(D) \neq 0$ равносильна системе (1). В то же время всякая система (2), равносильная (1), записывается в виде (4) с матрицей $D = (I - G)\,A^{-1}$. Для систем со знакоопределенными матрицами метод (3) обычно строится в виде

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\,\mathbf{x}^k = \mathbf{b}, \qquad \text{т.е.} \quad G = I - \tau\,A, \;\; \mathbf{c} = \tau\,\mathbf{b}. \tag{5}$$

Здесь τ — итерационный параметр. Отметим, что произвольную невырожденную систему (1) можно свести к задаче с симметричной положительно определенной матрицей путем умножения обеих частей равенства на A^T слева. При этом число обусловленности возрастает квадратично, что не всегда является приемлемым.

Справедливы следующие утверждения о сходимости метода простой итерации.

Утверждение 1 (достаточное условие сходимости). *Если $\|G\| < 1$, то система уравнений (2) имеет единственное решение и итерационный процесс (3) сходится к решению со скоростью геометрической прогрессии.*

Доказательство. Для любого решения системы (2) имеем оценку $\|\mathbf{x}\| \leq \|G\| \cdot \|\mathbf{x}\| + \|\mathbf{c}\|$. Поэтому справедливо неравенство $\|\mathbf{x}\| \leq (1 - \|G\|)^{-1}\|\mathbf{c}\|$. Отсюда следует существование и единственность решения однородной системы, а следовательно, и неоднородной.

Пусть $\mathbf{e}^k = \mathbf{x}^k - \mathbf{x}$ — погрешность решения на k-й итерации, получим для нее уравнение. Так как $\mathbf{x} = G\mathbf{x} + \mathbf{c}$ и $\mathbf{x}^{k+1} = G\mathbf{x}^k + \mathbf{c}$, то вычитая из второго уравнения первое, получим $\mathbf{e}^{k+1} = G\mathbf{e}^k$. Отсюда следует сходимость со скоростью геометрической прогрессии: так как $\mathbf{e}^k = G^k\mathbf{e}^0$, то справедлива оценка $\|\mathbf{e}^k\| \leq \|G\|^k\|\mathbf{e}^0\|$ с показателем $\|G\| < 1$. Утверждение доказано.

Для выяснения вопроса о необходимых и достаточных условиях сходимости метода простой итерации нам потребуется

Лемма. *Пусть все собственные значения λ_i матрицы G лежат в круге $|\lambda| \le q$, причем собственным значениям λ, по модулю равным q, соответствуют жордановы клетки размерности 1. Тогда существует матрица $\Lambda : \Lambda = D^{-1}GD$ с нормой $\|\Lambda\|_\infty \le q$.*

Доказательство. Напомним, что $\|A\|_\infty = \max\limits_i \sum\limits_{j=1}^{n} |a_{ij}|$. Положим $\eta = q - \max\limits_{|\lambda|<q} |\lambda| > 0$. Собственными значениями матрицы $\eta^{-1}G$ будут величины $\eta^{-1}\lambda_i(G)$. Преобразуем матрицу $\eta^{-1}G$ к нормальной жордановой форме

$$D^{-1}(\eta^{-1}G)D = \begin{pmatrix} \eta^{-1}\lambda_1 & \alpha_{1,2} & 0 & \dots \\ 0 & \ddots & \ddots & 0 \\ \dots & 0 & \eta^{-1}\lambda_{n-1} & \alpha_{n-1,n} \\ \dots & 0 & \dots & \eta^{-1}\lambda_n \end{pmatrix},$$

где величины $\alpha_{i,i+1}$ принимают значения 0 или 1. После умножения на η получим

$$\Lambda = D^{-1}GD = \begin{pmatrix} \lambda_1 & \eta\alpha_{1,2} & 0 & \dots \\ 0 & \ddots & \ddots & 0 \\ \dots & 0 & \lambda_{n-1} & \eta\alpha_{n-1,n} \\ \dots & 0 & \dots & \lambda_n \end{pmatrix}.$$

Займемся оценкой нормы матрицы Λ. Если $|\lambda_i| = q$, то по условию $\alpha_{i,i+1} = 0$ и $|\lambda_i| + \eta\alpha_{i,i+1} = q$. Если $|\lambda_i| < q$, то

$$|\lambda_i| + \eta\alpha_{i,i+1} \le \max_{|\lambda|<q} |\lambda| + \eta = q.$$

Таким образом, в любом случае $\|\Lambda\|_\infty \le q$. Лемма доказана.

Утверждение 2 (необходимое и достаточное условие сходимости). *Пусть система* (2) *имеет единственное решение. Итерационный процесс* (3) *сходится к решению системы* (2) *при любом начальном приближении $\mathbf{x}^0$ тогда и только тогда, когда все собственные значения матрицы G по модулю меньше* 1.

Доказательство. Достаточность. Возьмем произвольное q в пределах $\max\limits_i |\lambda_i(G)| < q < 1$. Условие леммы выполнено для матрицы G по отношению к этому q. Поэтому существует матрица D такая, что $\|\Lambda\|_\infty \le q$ при $\Lambda = D^{-1}GD$. Так как $G = D\Lambda D^{-1}$, то $G^k = D\Lambda D^{-1}D\Lambda D^{-1}\dots D\Lambda D^{-1} = D\Lambda^k D^{-1}$. Поэтому $\|G^k\|_\infty \le \|D\|_\infty \|D^{-1}\|_\infty q^k \to 0$. Следовательно $\|\mathbf{x}^k - \mathbf{x}\|_\infty \le \mathrm{cond}_\infty(D) q^k \|\mathbf{x}^0 - \mathbf{x}\|_\infty \to 0$. Так как в конечномерных пространствах все нормы эквивалентны, то отсюда следует сходимость в произвольной норме. Достаточность доказана.

Необходимость. Доказывается от противного. Пусть для некоторого l найдется собственное значение $|\lambda_l(G)| \ge 1$, обозначим через $\mathbf{y}_l$ соответствующий ему собственный вектор $G : G\mathbf{y}_l = \lambda_l \mathbf{y}_l$. Тогда при начальном приближении $\mathbf{x}^0 = \mathbf{x} + \gamma\mathbf{y}_l$, $\gamma \ne 0$ имеем для ошибки $\mathbf{e}^0 = \gamma\mathbf{y}_l$, и $\mathbf{e}^k = \gamma\lambda_l^k \mathbf{y}_l$, т.е. ее норма не стремится к нулю при $k \to \infty$. Полученное противоречие завершает доказательство.

Пусть A — симметричная положительно определенная матрица, т.е. подобна диагональной ($A = QDQ^{-1}$, где столбцы $\mathbf{q}_i$ матрицы Q есть собственные векторы матрицы A, а элементы диагональной матрицы D — соответствующие им собственные значения, т.е. $d_{ii} = \lambda_i > 0$). Отметим, что Q является ортогональной матрицей. Без ограничения общности можно считать, что $\lambda(A) \in [m, M],\ m > 0$.

Покажем, что метод $\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right)/\tau + A\mathbf{x}^k = \mathbf{b}$ сходится при $0 < \tau < \dfrac{2}{M}$. Его оператор перехода имеет вид $G = I - \tau A$, значит собственные значения G можно представить как $\lambda_i(G) = 1 - \tau\lambda_i(A)$. Поэтому по теореме о необходимом и достаточном условии сходимости метода простой итерации имеем

$$|1 - \tau\,\lambda(A)| < 1 \qquad \forall\lambda(A) \in [m, M]\,,$$

откуда и следует искомое неравенство для τ.

Оптимальный одношаговый итерационный метод

Выясним значение параметра τ, при котором сходимость метода

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}$$

будет наилучшей.

Утверждение 3. *При условии $A = A^T > 0$, $\lambda(A) \in [m, M]$ оптимальное значение равно $\tau_0 = \frac{2}{m+M}$. При этом имеет место геометрическая скорость убывания ошибки $\|\mathbf{x} - \mathbf{x}^k\|_2 \le q_0^k\|\mathbf{x} - \mathbf{x}^0\|_2$ с $q_0 = \frac{M-m}{M+m} < 1$.*

Доказательство. Напомним, что справедлив следующий закон изменения вектора ошибки $\mathbf{e}^k = \mathbf{x} - \mathbf{x}^k$:

$$\mathbf{e}^{k+1} = G\,\mathbf{e}^k, \quad G = I - \tau A\,.$$

Отсюда следует, $\mathbf{e}^k = G^k\mathbf{e}^0$ и $\|\mathbf{e}^k\|_2 \le \|G\|_2^k\|\mathbf{e}^0\|_2$. Так как матрица G является симметричной, то евклидова матричная норма имеет вид $\|G\|_2 = \sqrt{\max|\lambda(G^TG)|} = \max|\lambda(G)|$, и соответствующая оптимизационная задача сводится к следующей

$$q_0 = \min_\tau\Big(\max_{\lambda\in[m,M]}|1 - \tau\lambda|\Big).$$

Искомое значение τ_0 находится из равенства

$$1 - \tau_0 m = -(1 - \tau_0 M)\,.$$

Это следует из того, что функция $1 - \tau\lambda$ является линейной по λ, и ее максимум модуля на $[m, M]$ достигается на одном из концов отрезка. Приведенное выше равенство означает, что на обоих концах минимизируемая функция принимает одинаковые значения. Если оно не выполнено, то в силу непрерывности по τ, максимальное значение можно уменьшить, что противоречит оптимальности τ_0.

Вычислим величину q_0. При оптимальном значении τ она равна

$$q_0 = \frac{M - m}{M + m} = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{m}{M} < 1\,.$$

Утверждение доказано.

Отметим, что величина ξ совпадает с $1/\operatorname{cond}_2(A)$, т.е. от числа обусловленности исходной матрицы зависит не только точность получаемого решения (как это следует из неравенства Уилкинсона), но и скорость его получения.

Оптимальный циклический итерационный метод

Рассмотрим следующий алгоритм с переменным итерационным параметром

$$\frac{\mathbf{x}^{k+1}-\mathbf{x}^k}{\tau_{k+1}}+A\mathbf{x}^k=\mathbf{b}.$$

Будем считать, что допускается изменение параметра τ в зависимости от номера итерации следующим (циклическим с периодом N) образом:

$$\tau_1,\tau_2,\ldots,\tau_N,\tau_1,\tau_2,\ldots\,.$$

Утверждение 4. *При условии $A=A^T>0$, $\lambda(A)\in[m,M]$ оптимальные значения параметров τ_k равны обратным величинам корней многочлена Чебышева степени N на отрезке $[m,M]$: $\tau_k^{-1}=\frac{M+m}{2}+\frac{M-m}{2}\cos\frac{\pi(2k-1)}{2N}$, $k=1,\ldots,N$. При этом имеет место следующая оценка скорости сходимости за N шагов:*

$$\|\mathbf{x}-\mathbf{x}^N\|_2\le\frac{2q_1^N}{1+q_1^{2N}}\|\mathbf{x}-\mathbf{x}^0\|_2,\quad q_1=\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}.$$

Доказательство. Для рассматриваемого метода за N шагов имеем следующий закон изменения вектора ошибки $\mathbf{e}^k=\mathbf{x}-\mathbf{x}^k$

$$\mathbf{e}^{l+N}=G_N\,\mathbf{e}^l,\quad G_N=\prod_{k=1}^{N}(I-\tau_kA)\,.$$

Будем искать набор итерационных параметров τ_k, $k=1,...,N$ из условия минимума евклидовой нормы оператора перехода G_N. Если $A=A^T>0$, тогда справедливо $A=QDQ^{-1}$ и $G_N=Q\left[\prod_{k=1}^{N}(I-\tau_kD)\right]Q^{-1}$. Это означает, что G_N является симметричной матрицей. Поэтому

$$\min_{\tau_k}\|G_N\|_2=\min_{\tau_k}\Big(\max_{\lambda\in[m,M]}\left|\prod_{k=1}^{N}(1-\tau_k\lambda)\right|\Big).$$

Заметим, что выражение $\prod_{k=1}^{N}(1-\tau_k\lambda)$ является многочленом N-ой степени с младшим коэффициентом, равным единице. В лекции 3 было показано, что многочлен, наименее отклоняющийся от нуля на $[m,M]$ в этом классе, имеет вид $P_N^*(\lambda)=\dfrac{\bar{T}_N^{[m,M]}(\lambda)}{\bar{T}_N^{[m,M]}(0)}$, где $\bar{T}_n^{[a,b]}(x)=(b-a)^n2^{1-2n}T_n\left(\frac{2x-(b+a)}{b-a}\right)$ — приведенный многочлен Чебышева на этом отрезке. Его корни выписываются аналитически, а искомые τ_k есть величины, обратные этим корням.

Найдем величину нормы $\|G_N\|_2 = \|P_N^*(\lambda)\|_C = 1/|T_N^{[m,M]}(0)|$. Для этого воспользуемся следующим представлением для многочлена Чебышева

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

Так как $T_N^{[m,M]}(0) = T_N\left(-\dfrac{M+m}{M-m}\right)$, требуется вычислить величины $x \pm \sqrt{x^2-1}$ для $x = -\dfrac{M+m}{M-m}$. Элементарные преобразования приводят к равенствам

$$x \pm \sqrt{x^2-1} = -\frac{\left(\sqrt{M} \pm \sqrt{m}\right)^2}{(\sqrt{M}-\sqrt{m})(\sqrt{M}+\sqrt{m})}.$$

Используя обозначение $q_1 = \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}$, можно записать

$$T_N\left(-\frac{M+m}{M-m}\right) = \frac{(-q_1)^N + (-q_1)^{-N}}{2},$$

откуда и следует искомое выражение для оценки погрешности. Утверждение доказано.

Следует отметить, что в среднем за каждую из N итераций при выборе чебышевских шагов ошибка убывает примерно как $\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}$, что при малых ξ существенно меньше, чем аналогичная величина $\frac{1-\xi}{1+\xi}$ для оптимального одношагового метода.

Улучшить полученную скорость сходимости нельзя, так как многочлен Чебышева обладает свойством минимальности нормы, — это безусловное достоинство. Недостатком метода является требование информации о границах спектра матрицы A.

Однако здесь появляется важный аспект **упорядочивания шагов**. В каком порядке брать τ_k? Практические вычисления показали, что на некоторых шагах (когда $\max\limits_{[m,M]} |1-\tau\lambda| > 1$) или их последовательностях вычислительная погрешность может сильно расти и за это время полностью исказить решение задачи. Поэтому используется специальная процедура упорядочивания шагов, гарантирующая вычислительную устойчивость:

$$\begin{gathered}
\max_\lambda |1-\tau_1\lambda| \le 1, \\
\max_\lambda |(1-\tau_1\lambda)(1-\tau_2\lambda)| \le 1, \\
\dots \quad \dots \quad \dots \quad \dots \quad \dots \\
\max_\lambda |(1-\tau_1\lambda)\dots(1-\tau_N\lambda)| \le 1.
\end{gathered}$$

Приведем пример для $N = 2^l$. В формуле для τ_k используется перестановка множества $\{1, 2, \dots, N\}$. Ее можно построить рекуррентным образом:

$$l = 1 \to \{1,2\}, \;\; l = 2 \to \{1,4,2,3\}, \;\; l = 3 \to \{1,8,4,5,2,7,3,6\}.$$

Здесь раздвигаются элементы для предыдущего значения l и на свободные места (после имеющихся) записываются величины 2^l+1-"предыдущий элемент". Подобные процедуры упорядочивания шагов производят их такую попарную группировку, чтобы каждая полусумма не сильно отличалась от среднего.

Лекция 13

Обобщенный метод простой итерации. — Методы Якоби и Гаусса – Зейделя.

Обобщенный метод простой итерации

Скорость сходимости рассмотренных итерационных процессов зависела от отношения M/m границ спектра матрицы $A = A^T > 0$, то есть от числа обусловленности задачи. Если $M/m \gg 1$, то для "улучшения"исходной задачи можно перейти к некоторой равносильной системе $B^{-1}A\mathbf{x} = B^{-1}\mathbf{b}$ при условии невырожденности $B : \det(B) \neq 0$. Этот процедура называется *предобусловливанием* (precondition). Новую систему можно решать ранее рассмотренными алгоритмами, например

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + B^{-1}A\mathbf{x}^k = B^{-1}\mathbf{b}. \tag{1}$$

Пусть $B = B^T > 0$, определим величины

$$\tilde{M} = \sup_{\mathbf{y}\neq 0} \frac{(A\mathbf{y},\mathbf{y})}{(B\mathbf{y},\mathbf{y})}, \quad \tilde{m} = \inf_{\mathbf{y}\neq 0} \frac{(A\mathbf{y},\mathbf{y})}{(B\mathbf{y},\mathbf{y})}.$$

Теперь сходимость итерационных методов к решению предобусловленной системы определяется уже отношением $\tilde{M}/\tilde{m}$. При удачном выборе оператора B, т.е. если $\tilde{M}/\tilde{m} \ll M/m$, скорость сходимости улучшается принципиально, однако необходимо учитывать трудоемкость процедуры $\mathbf{y} = B^{-1}\mathbf{f}$. Например, при $B = A$, $\tau = 1$ метод (1) сойдется за одну итерацию, но будет эквивалентен решению исходной задачи $A\mathbf{x} = \mathbf{b}$. Другими словами, может оказаться так, что применение предобусловливателя B потребует настолько большого количества арифметических действий, что отказ от метода простой итерации окажется нецелесообразным.

При $B \neq I$ формулу (1) часто называют обобщенным методом простой итерации и записывают в виде

$$B\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}\,.$$

Требование $B = B^T > 0$ для сходимости в общем случае не обязательно.

Утверждение 1. *Пусть $A = A^T > 0$, $\tau > 0$, тогда обобщенный метод простой итерации сходится $\forall \mathbf{x}^0$ при условии $B - \frac{\tau}{2}A > 0$ (напомним, что это означает справедливость неравенства $(B\mathbf{x},\mathbf{x}) > \frac{\tau}{2}(A\mathbf{x},\mathbf{x})$ $\forall \mathbf{x} \neq 0$).*

Доказательство. Из уравнения для ошибки $\mathbf{e}^k$

$$B\frac{\mathbf{e}^{k+1} - \mathbf{e}^k}{\tau} + A\mathbf{e}^k = 0$$

следует

$$\mathbf{e}^{k+1} = (I - \tau B^{-1}A)\mathbf{e}^k\,, \quad A\mathbf{e}^{k+1} = (A - \tau AB^{-1}A)\mathbf{e}^k\,. \tag{2}$$

Введем обозначение $C = B - \frac{\tau}{2}A$ и вычислим скалярное произведение, используя симметрию A,

$$(A\mathbf{e}^{k+1},\mathbf{e}^{k+1}) = (A\mathbf{e}^k,\mathbf{e}^k) - 2\tau\left(CB^{-1}A\mathbf{e}^k, B^{-1}A\mathbf{e}^k\right)\,. \tag{3}$$

Из первого соотношения (2) имеем $B^{-1}A\mathbf{e}^k = -(\mathbf{e}^{k+1} - \mathbf{e}^k)/\tau$, это позволяет переписать (3) в виде

$$||\mathbf{e}^{k+1}||_A^2 - ||\mathbf{e}^k||_A^2 + \frac{2}{\tau}\left(C(\mathbf{e}^{k+1} - \mathbf{e}^k), \mathbf{e}^{k+1} - \mathbf{e}^k\right) = 0\,,$$

где $||\mathbf{u}||_A = (A\mathbf{u}, \mathbf{u})^{1/2}$.

Условие $C > 0$ равносильно условию $C \geq \varepsilon I$ с некоторым $\varepsilon > 0$ в силу конечномерности векторного пространства (здесь через I обозначена единичная матрица). Это дает

$$||\mathbf{e}^{k+1}||_A^2 - ||\mathbf{e}^k||_A^2 + 2\varepsilon\tau^{-1}||\mathbf{e}^{k+1} - \mathbf{e}^k||_2^2 \leq 0 \quad \forall k \geq 0\,.$$

Из этого неравенства следует монотонное убывание последовательности $\{||\mathbf{e}^k||_A^2\}$. Учитывая ее ограниченность снизу нулем, имеем сходимость к некоторому пределу $||\mathbf{e}^\infty||_A^2$. Отсюда сразу следует $\lim\limits_{k\to\infty} ||\mathbf{e}^{k+1} - \mathbf{e}^k||_2^2 = \lim\limits_{k\to\infty} ||\mathbf{x}^{k+1} - \mathbf{x}^k||_2^2 = 0$. Переходя к пределу при $k \to \infty$ в уравнении

$$A^{-1}B\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + \mathbf{x}^k = A^{-1}\mathbf{b}\ (\equiv \mathbf{x}),$$

убеждаемся (т.к. $\det(A^{-1}B) \neq 0$), что $\mathbf{x}^\infty$ является решением уравнения $A\mathbf{x} = \mathbf{b}$, т.е. последовательность приближений $\{\mathbf{x}^k\}$ сходится к $\mathbf{x}$. Утверждение доказано.

Рассмотрим обобщенный метод простой итерации с диагональной либо треугольной матрицей B:

$$B\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}\,.$$

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). Будем предполагать, что все диагональные элементы a_{ii} отличны от нуля, и, следовательно, любая матрица вида $D + \tau\,L$ с произвольным параметром τ обратима.

К **методам релаксации** обычно относят методы Якоби, Гаусса – Зейделя, верхней релаксации (SOR) и симметричной верхней релаксации (SSOR). Мы познакомимся с первыми тремя из них:
метод Якоби ($B = D$, $\tau = 1$):

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

метод Гаусса–Зейделя ($B = D + L$, $\tau = 1$):

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

метод SOR ($B = D + \omega L$, $\tau = \omega$):

$$(D + \omega\,L)\,\mathbf{x}^{k+1} + [\omega\,R + (\omega - 1)\,D]\,\mathbf{x}^k = \omega\,\mathbf{b}.$$

Здесь итерационный параметр ω традиционно называется *параметром релаксации.*

Для применимости этих методов симметрия и положительная определенность исходной матрицы A не обязательна. Рассмотрим следующий пример. Будем говорить, что **невырожденная** матрица A размерности $n \times n$ обладает свойством *диагонального преобладания*, если для всех i справедливо

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| \leq q|a_{ii}|, \quad 0 \leq q < 1\,. \tag{4}$$

Методы Якоби и Гаусса – Зейделя

Утверждение 2. *При выполнении условия (4) метод Якоби сходится с произвольного начального приближения* $\mathbf{x}^0$ *с оценкой* $\|\mathbf{x}^k - \mathbf{x}\|_\infty \leq q^k \|\mathbf{x}^0 - \mathbf{x}\|_\infty$.

Доказательство. Перепишем метод Якоби в виде, эквивалентном методу простой итерации $\mathbf{x}^{k+1} = G\mathbf{x}^k + \mathbf{c}$. Оператор перехода G здесь имеет вид $G = -D^{-1}(L+R)$, откуда для ошибки $\mathbf{e}^k = \mathbf{x}^k - \mathbf{x}$ следует равенство $\mathbf{e}^{k+1} = G\mathbf{e}^k$, или $\mathbf{e}^k = G^k\mathbf{e}^0$. Оценим ее норму $\|\mathbf{e}^k\|_\infty = \max\limits_{1\leq i\leq n} |e_i^k| \leq \|G\|_\infty^k \|\mathbf{e}^0\|_\infty$. Для вычисления величины $\|G\|_\infty = \max\limits_{1\leq i\leq n} \sum\limits_{j=1}^{n} |g_{ij}|$ воспользуемся явным видом матрицы G

$$G = -\begin{pmatrix} 0 & a_{12}/a_{11} & \dots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & \dots & a_{2n}/a_{22} \\ \dots & \dots & \dots & \dots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & \dots & 0 \end{pmatrix}.$$

Условие (4) означает, что $\|G\|_\infty \leq q$. Утверждение доказано.

Утверждение 3. *При выполнении условия (4) метод Гаусса – Зейделя сходится с произвольного начального приближения* $\mathbf{x}^0$ *с оценкой* $\|\mathbf{x}^k - \mathbf{x}\|_\infty \leq q^k \|\mathbf{x}^0 - \mathbf{x}\|_\infty$.

Доказательство. Обозначим вектор ошибки через $\mathbf{e}^k$. Для этого вектора имеет место соотношение (уравнение ошибки) $(D+L)\,\mathbf{e}^{k+1} + R\,\mathbf{e}^k = 0$. Пусть $\|\mathbf{e}^{k+1}\|_\infty = |e_m^{k+1}|$. Выпишем m-е уравнение

$$\sum_{j=1}^{m-1} a_{mj}\, e_j^{k+1} + a_{mm}\, e_m^{k+1} + \sum_{j=m+1}^{n} a_{mj}\, e_j^k = 0$$

и разрешим его относительно e_m^{k+1} :

$$e_m^{k+1} = -\sum_{j=1}^{m-1} \frac{a_{mj}}{a_{mm}}\, e_j^{k+1} - \sum_{j=m+1}^{n} \frac{a_{mj}}{a_{mm}}\, e_j^k\,.$$

Отсюда получим $\|\mathbf{e}^{k+1}\|_\infty = |e_m^{k+1}| \leq \alpha\, \|\mathbf{e}^{k+1}\|_\infty + \beta\, \|\mathbf{e}^k\|_\infty$, где

$$\alpha = \sum_{j=1}^{m-1} \left|\frac{a_{mj}}{a_{mm}}\right|, \quad \beta = \sum_{j=m+1}^{n} \left|\frac{a_{mj}}{a_{mm}}\right|.$$

Найденное соотношение можно переписать в виде

$$\|\mathbf{e}^{k+1}\|_\infty \leq \frac{\beta}{1-\alpha} \|\mathbf{e}^k\|_\infty.$$

Условие (4) означает, что $\alpha + \beta \leq q < 1$, следовательно,

$$\frac{\beta}{1-\alpha} \leq \frac{q-\alpha}{1-\alpha} \leq \frac{q-\alpha q}{1-\alpha} = q\,,$$

откуда и следует искомая оценка. Утверждение доказано.

Лекция 14

Метод верхней релаксации. — Метод наискорейшего градиентного спуска.

Метод верхней релаксации

Вернемся к случаю $A = A^T > 0$ и изучим сходимость метода верхней релаксации

$$(D + \omega L)\,\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\omega} + A\mathbf{x}^k = \mathbf{b}\,.$$

Утверждение 1. *Пусть $A = A^T > 0$, тогда для сходимости метода SOR с произвольного начального приближения необходимо и достаточно выполнение неравенства $0 < \omega < 2$.*

Доказательство. Достаточность. Пусть $0 < \omega < 2$. Напомним, что $A = L + D + R$, и в силу симметрии A имеем $L^T = R$. Кроме того, из положительной определенности A имеем $D > 0$. Действительно, $(A\mathbf{x}, \mathbf{x}) > 0 \;\; \forall \mathbf{x} \neq 0$. Возьмем в качестве $\mathbf{x}$ вектор $\mathbf{e}_i$, у которого все компоненты, кроме $i-$ой равны нулю, а $i-$ая — равна единице. Тогда $(A\mathbf{e}_i, \mathbf{e}_i) = a_{ii} = d_{ii} > 0$.

Вычислим величину

$$(A\mathbf{x}, \mathbf{x}) = ((L + D + R)\mathbf{x}, \mathbf{x}) = (D\mathbf{x}, \mathbf{x}) + 2(L\mathbf{x}, \mathbf{x}),$$

и воспользуемся достаточным условием сходимости обобщенного метода простой итерации $B - \dfrac{\omega}{2}A > 0$. Здесь $B = D + \omega L$, поэтому будем иметь

$$(B\mathbf{x}, \mathbf{x}) - \frac{\omega}{2}(A\mathbf{x}, \mathbf{x}) = (D\mathbf{x}, \mathbf{x}) + \omega(L\mathbf{x}, \mathbf{x}) -$$
$$-\frac{\omega}{2}[(D\mathbf{x}, \mathbf{x}) + 2(L\mathbf{x}, \mathbf{x})] = (D\mathbf{x}, \mathbf{x})\left(1 - \frac{\omega}{2}\right) > 0.$$

Достаточность доказана.

Необходимость. Пусть метод SOR сходится с произвольного начального приближения, тогда спектральный радиус оператора перехода $\rho(G)$, равный максимальному по модулю собственному значению, строго меньше единице.

Если формулу метода релаксации

$$(D + \omega L)\,\mathbf{x}^{k+1} + [\omega R + (\omega - 1)\,D]\,\mathbf{x}^k = \omega\mathbf{b}$$

умножить слева на матрицу D^{-1}, то оператор перехода G можно записать в следующем виде:

$$G = (I + \omega\, D^{-1}L)^{-1}\left[(1 - \omega)\, I - \omega\, D^{-1}R\right],$$

Здесь I — единичная, а $D^{-1}L$ и $D^{-1}R$ — строго нижняя и верхняя треугольные матрицы соответственно. Рассмотрим характеристический многочлен матрицы G : $d(\lambda) = \det(G - \lambda\, I)$. По теореме Виета имеет место равенство $(-1)^n\; d(0) = \prod\limits_{i=1}^{n} \lambda_i(G)$ (свободный член равен с точностью до знака произведению корней многочлена). Так как у треугольных матриц $D^{-1}L$ и $D^{-1}R$ на главной диагонали расположены нули, то $d(0) = \det(G) = (1-\omega)^n$. Отсюда для спектрального радиуса оператора перехода получим оценку

$$\rho(G) = \max_i |\lambda_i(G)| \geq \left|\prod_{i=1}^{n} \lambda_i(G)\right|^{1/n} = |\det(G)|^{1/n} = |1 - \omega|\,,$$

которая и приводит к искомому ответу. Утверждение доказано.

Отметим, что в качестве частного случая ($\omega = 1$) мы получили сходимость метода Гаусса – Зейделя для симметричных положительно определенных матриц.

Метод наискорейшего градиентного спуска

Обсудим вопрос о выборе итерационных параметров в случае, когда информация о границах собственных значений исходной матрицы отсутствует. Напомним сначала результаты, связанные с оптимальным одношаговым методом. Исходной является задача

$$A\mathbf{x} = \mathbf{b}, \quad A = A^T > 0\,.$$

Пусть известны минимальное и максимальное собственные значения:

$$\lambda_{\min}(A) = m, \quad \lambda_{\max}(A) = M.$$

Тогда если в методе

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}$$

положить $\tau = \dfrac{2}{m+M}$, то будет справедлива оценка погрешности

$$\|\mathbf{x}^k - \mathbf{x}\|_2 \le q^k \|\mathbf{x}^0 - \mathbf{x}\|_2, \quad q = \frac{M-m}{M+m}.$$

Недостатком метода является необходимость знания M и m. Наша цель — построение алгоритма (наискорейшего градиентного спуска), имеющего аналогичную скорость сходимости, но не использующего информацию о границах спектра. Для этого необходимо привлечь идеи вариационного исчисления.

Заменим исходную систему линейных алгебраических уравнений задачей отыскания минимума функционала

$$F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})\,.$$

Эти задачи равносильны, т.е. имеют одно и то же единственное решение $\mathbf{x}$. Действительно, рассмотрим функцию при фиксированном $\mathbf{x}$

$$(A(\mathbf{y}-\mathbf{x}), \mathbf{y}-\mathbf{x}) - (A\mathbf{x}, \mathbf{x}) = (A\mathbf{y}, \mathbf{y}) - 2(A\mathbf{y}, \mathbf{x}) \equiv F(\mathbf{y})\,.$$

Поскольку $A = A^T > 0$, то $(A(\mathbf{y}-\mathbf{x}), \mathbf{y}-\mathbf{x}) > 0$ при $\mathbf{y} \ne \mathbf{x}$, а величина $(A\mathbf{x}, \mathbf{x}) = (\mathbf{b}, \mathbf{x})$ не зависит от $\mathbf{y}$, то $F(\mathbf{y})$ имеет единственную точку минимума, совпадающую с $\mathbf{x}$.

Теперь для отыскания точки экстремума можно применить известные методы минимизации функционала. Простейшими из них являются методы градиентного спуска, в которых приближения определяются формулой

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \delta_k \mathrm{grad}(F(\mathbf{x}^k)) = \mathbf{x}^k - \Delta_k(A\mathbf{x}^k - \mathbf{b}),\ \Delta_k = 2\,\delta_k.$$

Здесь δ_k — параметр метода, и его выбор определяет конкретный алгоритм. Например, его можно определить из условия

$$\delta_k : F(\mathbf{x}^{k+1}) = F(\mathbf{x}^k - \delta_k \mathrm{grad}(F(\mathbf{x}^k)) \to \min\,.$$

В этом случае метод называется *методом наискорейшего градиентного спуска*. Когда функционал квадратичный, задача выбора параметра решается в явном виде. Определим функцию $\varphi(\Delta_k) \equiv F(\mathbf{x}^{k+1}) =$

$$= F(\mathbf{x}^k) - 2\Delta_k(A\mathbf{x}^k - \mathbf{b}, A\mathbf{x}^k - \mathbf{b}) + \Delta_k^2(A(A\mathbf{x}^k - \mathbf{b}), A\mathbf{x}^k - \mathbf{b})\,.$$

Введем обозначение невязки $\mathbf{r}^k = A\mathbf{x}^k - \mathbf{b}$ и найдем выражение для Δ_k из условия $\varphi'(\Delta_k) = 0$:

$$\Delta_k = \frac{(\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, \mathbf{r}^k)}\,.$$

Рассмотрим вопрос о скорости сходимости алгоритма. Введем обозначение

$$F_0(\mathbf{y}) = (A(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x}) = \|\mathbf{y} - \mathbf{x}\|_A^2,$$

где $\mathbf{x}$ — точное решение нашей задачи, и докажем

Утверждение 2. *Пусть* $A = A^T > 0$, $\lambda(A) \in [m, M]$, $m > 0$, *тогда приближения* $\mathbf{x}^k$ *метода наискорейшего градиентного спуска удовлетворяют оценке*

$$F_0(\mathbf{x}^k) \le \left(\frac{M-m}{M+m}\right)^{2k} F_0(\mathbf{x}^0)\,.$$

Заметим, что из этого неравенства следует

$$\|\mathbf{x}^k - \mathbf{x}\|_2 \le q^k\sqrt{\frac{M}{m}}\|\mathbf{x}^0 - \mathbf{x}\|_2\,, \quad q = \frac{M-m}{M+m}\,,$$

и асимптотическая скорость сходимости такая же, как у оптимального одношагового метода, но информация о величинах M и m здесь не требуется.

Доказательство. Зафиксируем точку $\mathbf{x}^k$. Из точки $\mathbf{y}^k \equiv \mathbf{x}^k$ сделаем один шаг оптимального одношагового метода:

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{2}{M+m}(A\mathbf{y}^k - \mathbf{b})\,.$$

Ошибка $\mathbf{e}^k = \mathbf{y}^k - \mathbf{x}$ удовлетворяет соотношению

$$\mathbf{e}^{k+1} = \left(I - \frac{2}{M+m}A\right)\mathbf{e}^k.$$

Пусть $\{\mathbf{z}_i\}_{i=1}^n$ — ортонормированная система собственных векторов матрицы A :

$$A\mathbf{z}_i = \lambda_i\mathbf{z}_i, \quad (\mathbf{z}_j, \mathbf{z}_i) = \delta_i^j, \quad \lambda_i \in [m, M],\ m > 0\,.$$

Легко проверить справедливость неравенства

$$\left|1 - \frac{2\lambda_i}{M+m}\right| \le q = \frac{M-m}{M+m} \quad \forall\lambda_i \in [m, M]\,.$$

Пусть $\mathbf{e}^k = \sum\limits_{i=1}^n c_i\mathbf{z}_i$. Тогда

$$(A\mathbf{e}^k, \mathbf{e}^k) = \left(\sum_{i=1}^n c_i\lambda_i\mathbf{z}_i, \sum_{i=1}^n c_i\mathbf{z}_i\right) = \sum_{i=1}^n c_i^2\lambda_i\,,$$

$$\mathbf{e}^{k+1} = \sum_{i=1}^n b_i\mathbf{z}_i\,, \quad b_i = \left(1 - \tfrac{2\lambda_i}{M+m}\right)c_i,$$

$$(A\mathbf{e}^{k+1}, \mathbf{e}^{k+1}) = \sum_{i=1}^n b_i^2\lambda_i = \sum_{i=1}^n c_i^2\lambda_i\left(1 - \tfrac{2\lambda_i}{M+m}\right)^2.$$

Отсюда имеем оценку

$$(A\mathbf{e}^{k+1}, \mathbf{e}^{k+1}) \le q^2 (A\mathbf{e}^k, \mathbf{e}^k)\,.$$

Поскольку $F_0(\mathbf{y}^k) = (A\mathbf{e}^k, \mathbf{e}^k)$, то

$$F_0(\mathbf{y}^{k+1}) \le q^2 F_0(\mathbf{y}^k) = q^2 F_0(\mathbf{x}^k)\,.$$

Отметим, что приближение $\mathbf{y}^{k+1}$ получено по формуле

$$\mathbf{y}^{k+1} = \mathbf{x}^k - \alpha\,\mathrm{grad}(F(\mathbf{x}^k))\,, \quad \alpha = \frac{1}{M+m}\,,$$

т.е. в результате шага градиентного спуска с неоптимальным значением параметра δ_k. Поэтому для приближения наискорейшего градиентного спуска $\mathbf{x}^{k+1}$ будет справедливо неравенство

$$F(\mathbf{x}^{k+1}) \le F(\mathbf{y}^{k+1})$$

в силу правильного выбора итерационного параметра. Напомним, что $F(\mathbf{y}) = F_0(\mathbf{y}) - (A\mathbf{x}, \mathbf{x})$, поэтому

$$F_0(\mathbf{x}^{k+1}) - (A\mathbf{x}, \mathbf{x}) \le F_0(\mathbf{y}^{k+1}) - (A\mathbf{x}, \mathbf{x})\,.$$

Отсюда следует

$$F_0(\mathbf{x}^{k+1}) \le F_0(\mathbf{y}^{k+1}) \le \left(\frac{M-m}{M+m}\right)^2 F_0(\mathbf{x}^k),$$

и соответственно справедливость искомой оценки. Утверждение доказано.

Расчетные формулы перехода $\mathbf{x}^k \to \mathbf{x}^{k+1}$ в методе наискорейшего градиентного спуска:

$$\mathbf{r}^k = A\,\mathbf{x}^k - \mathbf{b}, \quad \Delta_k = \frac{(\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, \mathbf{r}^k)}, \quad \mathbf{x}^{k+1} = \mathbf{x}^k - \Delta_k \mathbf{r}^k$$

имеют значимый в случае большой размерности задачи недостаток: два умножения матрицы на вектор. Поэтому на практике часто используют другие формулы перехода $\{\mathbf{x}^k, \mathbf{r}^k\} \to \{\mathbf{x}^{k+1}, \mathbf{r}^{k+1}\}$:

$$\Delta_k = \frac{(\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, \mathbf{r}^k)}, \quad \mathbf{r}^{k+1} = \mathbf{r}^k - \Delta_k A\,\mathbf{r}^k, \quad \mathbf{x}^{k+1} = \mathbf{x}^k - \Delta_k \mathbf{r}^k\,.$$

В этом случае требуется на каждой итерации вычислять только вектор $A\,\mathbf{r}^k$, но зато нужно постоянно хранить два рабочих вектора $\mathbf{x}^k$ и $\mathbf{r}^k$, т.е. чем-то обязательно приходится расплачиваться за незнание границ спектра (либо вычислительной работой, либо использованием дополнительной памяти).

Отметим, что методы градиентного спуска являются нелинейными, поскольку параметр δ_k выбирается как некоторая функция итерационного приближения $\mathbf{x}^k$. Это приводит к усложнению анализа сходимости, часто весьма существенному, особенно при учете ошибок округлений.

Лекция 15

Линейная задача наименьших квадратов. — Метод нормального уравнения. — Метод $QR-$разложения. — Метод сингулярного разложения.

Линейная задача наименьших квадратов

Пусть требуется решить систему линейных уравнений с прямоугольной матрицей A размерности $m \times n$:

$$A_{m\times n}\mathbf{x} = \mathbf{b}, \quad \mathbf{x} \in \mathbf{R}^n,\ \mathbf{b} \in \mathbf{R}^m\,.$$

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

В случае 1) задача имеет единственное решение $\mathbf{x} = A^{-1}\mathbf{b}$ и для вектора невязки $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ справедливо $\|\mathbf{r}\| = 0$.

В случае 2) задача недоопределена и исходная система имеет подпространство решений размерности $(n-m)$, причем 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 $\|\mathbf{b} - A\mathbf{x}\| = \|\mathbf{r}\| > 0$.

Наибольший интерес с точки зрения практики представляют задачи случая 3). Если не оговаривается иное, будем считать, что $m > n$ и $\mathrm{rank}(A) = n$. Для задач такого рода Гаусс предложил следующую постановку: решением системы $A\mathbf{x} = \mathbf{b}$ *в смысле наименьших квадратов* называется вектор $\mathbf{x}$, минимизирующий евклидову норму вектора невязки $\min\limits_{\mathbf{y}} \|\mathbf{b} - A\mathbf{y}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2$. Такая постановка называется *задачей наименьших квадратов* (ЗНК). Изучим основные подходы к ее решению.

Метод нормального уравнения

Рассмотрим следующую, называемую *нормальной*, систему уравнений $A^TA\mathbf{x} = A^T\mathbf{b}$ с квадратной матрицей A^TA размерности $n \times n$ и из нее найдем вектор $\mathbf{x}$.

Утверждение 1 (Теорема Гаусса). *Пусть $m \geq n$ и $\mathrm{rank}(A) = n$. Тогда нормальная система уравнений имеет единственное решение.*

Доказательство. Действительно, $A^TA = (A^TA)^T$ и $(A^TA\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) > 0$ при $A\mathbf{x} \neq 0$. Но $A\mathbf{x} \neq 0$ для всякого $\mathbf{x} \neq 0$, так как $\mathrm{rank}(A) = n$. Следовательно, матрица A^TA невырождена и нормальное уравнение имеет единственное решение. Утверждение доказано.

Утверждение 2. *Пусть $m \geq n$ и $\mathrm{rank}(A) = n$. 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 предыдущего утверждения следует существование и единственность такого вектора $\mathbf{x}$, что $A^T(\mathbf{b} - A\mathbf{x}) = 0$. Рассмотрим вектор $\mathbf{y} = \mathbf{x} + \mathbf{z}$, для него справедливо

$0 \leq \|\mathbf{b}-A\mathbf{y}\|_2^2 = (\mathbf{b}-A(\mathbf{x}+\mathbf{z}), \mathbf{b}-A(\mathbf{x}+\mathbf{z})) = (\mathbf{b}-A\mathbf{x}, \mathbf{b}-A\mathbf{x})+(A\mathbf{z}, A\mathbf{z})-2(A\mathbf{z}, \mathbf{b} - A\mathbf{x}) = \|\mathbf{b} - A\mathbf{x}\|_2^2 + (A^TA\mathbf{z}, \mathbf{z}) - 2(\mathbf{z}, A^T(\mathbf{b} - A\mathbf{x}))$.

Последнее слагаемое равно нулю. Остается сумма двух неотрицательных слагаемых. Эта сумма минимальна при $\mathbf{z} = 0$, так как первое слагаемое фиксировано. Поэтому $\mathbf{z} = 0$, $\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b}$ и $\mathbf{y} = \mathbf{x}$. Утверждение доказано.

Метод нормального уравнения прост в реализации, но чувствителен к ошибкам округления, так как, например, для невырожденных квадратных матриц справедливо $\mathrm{cond}_2(A^TA) = \mathrm{cond}_2^2(A)$. Таким образом, обусловленность исходной задачи возводится в квадрат, поэтому полученное приближенное решение может сильно отличаться от точного, если $\mathrm{cond}(A) \gg 1$.

Метод QR-разложения

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 ортогонализации Грама – Шмидта. Для полноты изложения приведем его.

Для $i = 1$ до n /вычислить i–е столбцы матриц Q и R/
  $q_i := a_i$
  Для $j = 1$ до $i - 1$ /вычесть из a_i компоненту в направлении q_j/
    $r_{ji} = q_j^Ta_i \equiv (q_j, a_i)$
    $q_i := q_i - r_{ji}q_j$
  Конец цикла j
  $r_{ii} = \|q_i\|_2$
  $q_i := q_i/r_{ii}$
Конец цикла i.

Следует отметить, что на практике предпочтение отдается модифицированному методу Грама – Шмидта, который отличается только формулой $r_{ji} = q_j^Tq_i$, однако является более устойчивым по отношению к ошибкам округлений. Рекомендуется в качестве самостоятельного упражнения убедиться, что при отсутствии ошибок округлений оба алгоритма эквивалентны.

Утверждение 3. *Пусть A — матрица размерности $m \times n$, причем $m \geq n$ и $\mathrm{rank}(A) = n$. Тогда существуют и единственны: $m \times n$ матрица Q с ортонормированными столбцами (т.е. $Q^TQ = I_n$) и верхнетреугольная $n \times n$ матрица R с положительными диагональными элементами (т.е. $r_{ii} > 0$) такие, что $A = QR$.*

Доказательство. Применим процесс ортогонализации Грама – Шмидта к столбцам a_i, $i = 1, 2, \ldots, n$ матрицы $A = [a_1, a_2, \ldots, a_n]$ в порядке возрастания их номеров. Так как столбцы линейно независимы ($\mathrm{rank}(A) = n$), то получим ортонормированные векторы $q_1, q_2, \ldots, q_n$. Они являются столбцами матрицы $Q_{m\times n}$ такой, что $Q^TQ = I_n$. В процессе Грама – Шмидта вычисляются также коэффициенты $r_{ji} = q_j^Ta_i$ в выражении каждого столбца a_i как линейной комбинации векторов $q_1, q_2, \ldots, q_i$, т.е. $a_i = \sum\limits_{j=1}^{i} r_{ji}q_j$. Эти числа r_{ji} — элементы матрицы $R_{n\times n}$. Утверждение доказано.

Утверждение 4. *Пусть $m \geq n$, $\mathrm{rank}(A) = n$ и известно представление $A = QR$. Тогда решением задачи наименьших квадратов является решение системы $R\mathbf{x} = Q^T\mathbf{b}$.*

Доказательство. Из метода нормального уравнения имеем $\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b}$. Поэтому $\mathbf{x} = (R^TQ^TQR)^{-1}R^TQ^T\mathbf{b} = (R^TR)^{-1}R^TQ^T\mathbf{b} = R^{-1}R^{-T}R^TQ^T\mathbf{b} = R^{-1}Q^T\mathbf{b}$, т.е. $R\mathbf{x} = Q^T\mathbf{b}$. Утверждение доказано.

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

Метод сингулярного (SVD) разложения

Метод применяют для решения гарантированно наилучшим образом плохо обусловленных и вырожденных задач.

Утверждение (без доказательства). *Пусть A — произвольная матрица размерности $m \times n$, причем $m \geq n$. Тогда справедливо сингулярное разложение $A = U\Sigma V^T$, где*
U имеет размерность $m \times n$ и удовлетворяет соотношению $U^TU = I_n$,
V имеет размерность $n \times n$ и удовлетворяет соотношению $V^TV = I_n$,
Σ — диагональная матрица размерности $n \times n$ с элементами $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$.

Столбцы $\mathbf{u}_1, \ldots, \mathbf{u}_n$ матрицы U называют *левыми сингулярными* векторами матрицы A, столбцы $\mathbf{v}_1, \ldots, \mathbf{v}_n$ матрицы V — *правыми сингулярными* векторами, величины σ_i — *сингулярными числами*.

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.

Утверждение 5. *Пусть $m \geq n$,* $\mathrm{rank}(A) = n$ *и известно представление $A = U\Sigma V^T$. Тогда решением задачи наименьших квадратов является вектор вида $\mathbf{x} = V\Sigma^{-1}U^T\mathbf{b}$.*

Доказательство. Так как $\mathrm{rank}(A) = n$, то матрица Σ невырождена. Из метода нормального уравнения следует, что решение $\mathbf{x}$ можно представить в виде

$\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b} = (V\Sigma U^TU\Sigma V^T)^{-1}V\Sigma U^T\mathbf{b} = V\Sigma^{-1}U^T\mathbf{b}$. Утверждение доказано.

Сформулируем правило решения задачи наименьших квадратов в приближенной арифметике. В реальных вычислениях все σ_i получатся (с учетом машинной точности) отличными от нуля, поэтому зафиксируем некоторое значение ε. Будем считать, что величины $\sigma_k < \varepsilon$ 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 16

Общая идея и примеры проекционных методов. — Пространства Крылова. Понятие о методе сопряженных градиентов.

Общая идея и примеры проекционных методов

Эффективными методами решения системы линейных алгебраических уравнений $A\mathbf{x} = \mathbf{b}$ большой размерности являются итерационные методы проекционного типа. Их общую идею можно сформулировать следующим образом:

в зависимости от текущего приближения $\mathbf{x}^l \in \mathbf{R}^n$ и номера итерации l выбирают два m-мерных ($m \leq n$) подпространства $\mathcal{K}$ и $\mathcal{L}$; следующее приближение $\mathbf{x}^{l+1}$ к точному решению $\mathbf{x}$ ищут в виде $\mathbf{x}^{l+1} = \mathbf{x}^l + \mathbf{k}$, $\mathbf{k} \in \mathcal{K}$ из условия $\mathbf{r}^{l+1} \perp \mathcal{L}$, где $\mathbf{r}^{l+1} = \mathbf{b} - A\mathbf{x}^{l+1}$.

Таким образом, следует построить вектор поправки $\mathbf{k}$ из подпространства $\mathcal{K}$, обеспечивающего ортогональность вектора невязки $\mathbf{r}^{l+1}$ подпространству $\mathcal{L}$. Различные правила выбора подпространств $\mathcal{K}$ и $\mathcal{L}$ приводят к различным расчетным формулам.

Пример 1 (метод наискорейшего градиентного спуска). В простейшем случае в качестве базовых пространств $\mathcal{K}$ и $\mathcal{L}$ выбирают одномерные подпространства. Определим $\mathcal{K} = \mathcal{L} = \{\mathbf{r}^l\}$, где $\mathbf{r}^l = \mathbf{b} - A\mathbf{x}^l$, и расссмотрим соответствующий проекционный метод. В данном случае имеем $\mathbf{x}^{l+1} = \mathbf{x}^l + \tau\mathbf{r}^l$, где итерационный параметр τ определяется из условия ортогональности $(\mathbf{b} - A(\mathbf{x}^l + \tau\mathbf{r}^l), \mathbf{r}^l) = 0$. Отсюда имеем $(\mathbf{r}^l - \tau A\mathbf{r}^l, \mathbf{r}^l) = 0$ и $\tau = (\mathbf{r}^l, \mathbf{r}^l)/(A\mathbf{r}^l, \mathbf{r}^l)$, т.е. значение итерационного параметра τ тождественно совпадает со значением Δ_l из метода наискорейшего градиентного спуска одновременно с формулой для $\mathbf{x}^{l+1}$.

Пример 2. 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 $\mathbf{R}^n$. Положим $\hat{\mathbf{x}} = \mathbf{x}^l$ и выполним n вспомогательных шагов. Последовательно для $i = 1, ..., n$ имеем: $\hat{\mathbf{x}} := \hat{\mathbf{x}} + c_i\mathbf{e}_i$, $(\mathbf{b} - A(\hat{\mathbf{x}} + c_i\mathbf{e}_i), \mathbf{e}_i) = 0$.

Это дает $c_i = \dfrac{1}{a_{ii}}\left(b_i - \sum\limits_{j=1}^{i-1} a_{ij}\hat{x}_j - \sum\limits_{j=i}^{n} a_{ij}x_j^l\right)$. Таким образом, за n вспомогательных шагов проекционного алгоритма получаем $\mathbf{x}^{l+1} = \mathbf{x}^l + \sum\limits_{i=1}^{n} c_i\mathbf{e}_i$, что соответствует одной итерации метода Гаусса–Зейделя: $(D + L)\mathbf{x}^{l+1} + R\mathbf{x}^l = \mathbf{b}$.

Пространства Крылова. Понятие о методе сопряженных градиентов

Пусть пространства $\mathcal{L}$ зависят от номера итерации и $\mathcal{L}^1 \subset \mathcal{L}^2 \subset \ldots \subset \mathcal{L}^l \subset \ldots \subset \mathcal{L}^n = \mathbf{R}^n$. Тогда точное решение системы будет получено не позже, чем за n шагов. Если же цепочка $\mathcal{L}^l$ задается некоторым оптимальным образом, то можно расчитывать, что требуемая точность $\|\mathbf{x} - \mathbf{x}^l\| \leq \varepsilon$, где $A\mathbf{x} = \mathbf{b}$, будет достигнута значительно раньше.

Эффективные алгоритмы удается построить, если в качестве пространства $\mathcal{K}^l$ выбрать пространство Крылова $\mathcal{K}^l = \mathrm{span}\{\mathbf{r}, A\mathbf{r}, ..., A^{l-1}\mathbf{r}\}$ размерности l для $\mathbf{r} = \mathbf{b} - A\mathbf{x}^0$. При этом пространство $\mathcal{L}^l$ определяется либо как $\mathcal{L}^l = \mathcal{K}^l$, либо $\mathcal{L}^l = A\mathcal{K}^l$. В **методе сопряженных градиентов** реализован первый подход в предположении $A = A^T > 0$. Опишем его структуру.

Построим проекционный метод для пары пространств

$$\mathcal{K}^l = \mathrm{span}\left\{\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{l-1}\mathbf{r}^0\right\}, \quad \mathcal{L}^l = \mathcal{K}^l, \quad \mathbf{r}^0 = \mathbf{b} - A\mathbf{x}^0,$$

очередное приближение найдем в виде $\mathbf{x}^l = \mathbf{x}^0 + \sum\limits_{i=1}^{l} c_i A^{i-1}\mathbf{r}^0$, а коэффициенты c_i определим из условия $\mathbf{r}^l = \left(\mathbf{b} - A\mathbf{x}^l\right) \perp \mathcal{L}^l$, т.е.

$$\left(\mathbf{r}^0 - \sum_{i=1}^{l} c_i A^i \mathbf{r}^0, A^{j-1}\mathbf{r}^0\right) = 0, \quad j = 1, 2, \ldots, l.$$

Такая форма алгоритма требует для нахождения $\{c_i\}$ решения системы линейных уравнений (при этом формально $\{c_i = c_i^{(l)}\}$, т.е. коэффициенты могут изменяться при увеличении l). Рассмотрим эквивалентную, но более удобную с практической точки зрения реализацию этого алгоритма.

Пусть в пространстве $\mathcal{K}^l = \mathrm{span}\left\{\mathbf{k}_1, \ldots, \mathbf{k}_l\right\}$ известен A-ортогональный базис, т.е. $(A\mathbf{k}_i, \mathbf{k}_j) = 0$ при $i \neq j$ и $\mathbf{k}_1 = \mathbf{r}^0$. Тогда $\mathbf{x}^l = \mathbf{x}^0 + \sum\limits_{i=1}^{l} \alpha_i \mathbf{k}_i$ и $\mathbf{r}^l = \mathbf{b} - A\mathbf{x}^l = \mathbf{b} - A\left(\mathbf{x}^0 + \sum\limits_{i=1}^{l} \alpha_i \mathbf{k}_i\right)$. В этом случае из условия $\mathbf{r}^l \perp \mathcal{L}^l$ имеем формулы для определения коэффициентов

$$(\mathbf{r}^l, \mathbf{k}_j) = (\mathbf{r}^0, \mathbf{k}_j) - \alpha_j (A\mathbf{k}_j, \mathbf{k}_j) = 0, \; \alpha_j = \frac{(\mathbf{r}^0, \mathbf{k}_j)}{(A\mathbf{k}_j, \mathbf{k}_j)}, \; j = 1, \ldots, l.$$

Таким образом, коэффициенты $\alpha_1, \alpha_2, \ldots$ не зависят от выбора l, поэтому могут быть найдены последовательно из условия $\mathbf{x}^l = \mathbf{x}^{l-1} + \alpha_l \mathbf{k}_l$. Отсюда получаем $\mathbf{r}^l = \mathbf{r}^{l-1} - \alpha_l A\mathbf{k}_l$ и $\alpha_l = (\mathbf{r}^{l-1}, \mathbf{k}_l)/(A\mathbf{k}_l, \mathbf{k}_l)$. 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 $\{\mathbf{k}_i\}_{i=1}^l$, в данном случае оказывается существенно менее эффективна. Будем строить базис последовательно. На шаге l имеем

$$\mathrm{span}\Big\{\left\{\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{l-1}\mathbf{r}^0\right\}, A^l\mathbf{r}^0\Big\} = \mathrm{span}\Big\{\left\{\mathbf{k}_1, \ldots, \mathbf{k}_l\right\}, \mathbf{k}_{l+1}\Big\}.$$

Отсюда следует, что $\mathbf{k}_{l+1} = A^l\mathbf{r}^0 + \sum\limits_{i=1}^{l} \tilde{\beta}_i \mathbf{k}_i$. Заметим, что из $\mathbf{x}^l = \mathbf{x}^0 + \sum\limits_{i=1}^{l} c_i A^{i-1}\mathbf{r}^0$ следует

$$\mathbf{r}^l = \mathbf{r}^0 - A\sum_{i=1}^{l} c_i A^{i-1}\mathbf{r}^0 = \mathbf{r}^0 - \sum_{i=1}^{l} c_i A^i \mathbf{r}^0,$$

следовательно, при $\mathbf{r}^l \neq 0$ и $c_l \neq 0$ (иначе метод сошелся!) вектор $\mathbf{k}_{l+1}$

можно искать в виде $\mathbf{k}_{l+1} = \mathbf{r}^l + \sum\limits_{i=1}^{l} \beta_i \mathbf{k}_i$. При указанных условиях вектор $A^l\mathbf{r}^0$ присутствует в конструкции вектора $\mathbf{r}^l$. В противном случае вектор $\mathbf{r}^l$

Так как $\mathbf{r}^l \perp \mathcal{L}^l$ (см. определение коэффициентов α_j выше) и векторы $A\mathbf{k}_1, A\mathbf{k}_2, \ldots, A\mathbf{k}_{l-1} \in \mathcal{L}^l$ (доказывается по индукции, начиная с $A\mathbf{k}_1 \in \mathcal{L}^2$), то $(\mathbf{r}^l, A\mathbf{k}_i) = 0$ при $i < l$. Отсюда и из A-ортогональности векторов $\mathbf{k}_i$ имеем $\beta_i = 0$ при $i < l$, т.е. $\mathbf{k}_{l+1} = \mathbf{r}^l + \beta_l \mathbf{k}_l$. Из данного представления определяем коэффициенты $\beta_l = -(\mathbf{r}^l, A\mathbf{k}_l)/(A\mathbf{k}_l, \mathbf{k}_l)$.

Окончательно набор расчетных формул имеет вид:

$$\mathbf{x}^l = \mathbf{x}^{l-1} + \alpha_l \mathbf{k}_l, \quad \alpha_l = \frac{(\mathbf{r}^{l-1}, \mathbf{k}_l)}{(A\mathbf{k}_l, \mathbf{k}_l)},$$

$$\mathbf{k}_{l+1} = \mathbf{r}^l + \beta_l \mathbf{k}_l, \quad \beta_l = -\frac{(\mathbf{r}^l, A\mathbf{k}_l)}{(A\mathbf{k}_l, \mathbf{k}_l)}, \ \mathbf{k}_1 = \mathbf{r}^0.$$

Сделаем несколько замечаний о свойствах рассматриваемого алгоритма.

Для метода сопряженных градиентов известно, что получаемые приближения удовлетворяют неравенству

$$F_0(\mathbf{x}^l) \le \frac{1}{T_l^2\left(-\dfrac{M+m}{M-m}\right)} F_0(\mathbf{x}^0), \quad F_0(\mathbf{x}^l) = \|\mathbf{x} - \mathbf{x}^l\|_A^2 \equiv (A(\mathbf{x} - \mathbf{x}^l), \mathbf{x} - \mathbf{x}^l),$$

где $T_l(x)$ — многочлен Чебышева l-й степени, M и m — соответственно максимальное и минимальное собственные значения матрицы A. При изучении оптимального циклического итерационного метода было получено представление

$$T_l\left(-\frac{M+m}{M-m}\right) = \frac{(-q)^l + (-q)^{-l}}{2},$$

где $q = \dfrac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}$, поэтому справедлива оценка

$$\|\mathbf{x} - \mathbf{x}^l\|_2 \le \sqrt{\frac{M}{m}}\, \frac{2q^l}{1+q^{2l}} \|\mathbf{x}^* - \mathbf{x}^0\|_2 .$$

Это означает, что метод сопряженных градиентов сходится к решению задачи с максимально возможной скоростью, определяемой нормой многочлена Чебышева.

Кроме того, для невязок $\mathbf{r}^l = \mathbf{b} - A\,\mathbf{x}^l$ справедливы соотношения ортогональности $(\mathbf{r}^l, \mathbf{r}^j) = 0$ для $j = 0, 1, \ldots, l-1$. Размерность же вектора решения $\mathbf{x}^*$ конечна и равна n, поэтому при отсутствии ошибок округлений не позже, чем за n итераций мы получим точное решение задачи. Необходимое количество итераций зависит от количества различных собственных векторов исходной матрицы A, присутствующих в разложении начальной ошибки. В качестве самостоятельного упражнения рекомендуется проанализировать поведение метода сопряженных градиентов в случае $\mathbf{x} - \mathbf{x}^0 = c_1\mathbf{e}_1 + \ldots + c_p\mathbf{e}_p$, $c_i \ne 0$, где $\mathbf{e}_i$ — ортонормированные собственные векторы матрицы A, соответствующие различным собственным значениям λ_i. Конечно, в реальных условиях округлений не избежать, и это влияет на устойчивость вычислений. Но, как правило, влияние вычислительной погрешности можно компенсировать дополнительными итерациями.

Таким образом, расмотренный алгоритм является одновременно и итерационным, и прямым, что делает его очень привлекательным для практических вычислений. В настоящее время метод сопряженных градиентов является основным алгоритмом решения систем уравнений с симметричными положительно определенными матрицами большой размерности.

Лекция 17

Частичная проблема собственных значений. — Полная проблема собственных значений. QR-алгоритм.

Частичная проблема собственных значений

Пусть A — заданная вещественная матрица размера $n \times n$, требуется определить ненулевой собственный вектор $\mathbf{x}$ и собственное значение λ, удовлетворяющие равенству $A\,\mathbf{x} = \lambda\,\mathbf{x}$.

Степенной метод вычисления максимального по модулю собственного значения матрицы A имеет вид:

$$\mathbf{x}^{k+1} = A\mathbf{x}^k, \;\; \lambda^{(k)} = \frac{(\mathbf{x}^{k+1}, \mathbf{x}^k)}{||\mathbf{x}^k||_2^2}, \;\; \mathbf{x}^k \neq 0; \;\; k = 0, 1, 2, \ldots .$$

При практической реализации разумно на каждом шаге нормировать текущий вектор: $\mathbf{x}^k := \mathbf{x}^k / \|\mathbf{x}^k\|_2$.

Утверждение 1. *Пусть A — матрица простой структуры (собственные векторы $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n$ матрицы образуют базис в $\mathbf{C}^n$). Пусть далее $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_n|$ и $L = \mathrm{span}\{\mathbf{e}_2, \mathbf{e}_3, ..., \mathbf{e}_n\}$. Тогда для степенного метода при условии $\mathbf{x}^0 \notin L$ справедлива оценка $\lambda^{(k)} = \lambda_1 + O(|\lambda_2/\lambda_1|^k)$.*

Доказательство. Разложим начальное приближение по собственным векторам

$$\mathbf{x}^0 = \sum_{i=1}^{n} c_i \mathbf{e}_i \,,$$

отсюда имеем

$$\mathbf{x}^k = \sum_{i=1}^{n} c_i \lambda_i^k \mathbf{e}_i \,, \quad \mathbf{x}^{k+1} = \sum_{i=1}^{n} c_i \lambda_i^{k+1} \mathbf{e}_i \,.$$

Оценим рост скалярных произведений:

$$\left(\mathbf{x}^k, \mathbf{x}^k\right) = \left(\sum_{i=1}^{n} c_i \lambda_i^k \mathbf{e}_i, \sum_{i=1}^{n} c_i \lambda_i^k \mathbf{e}_i \right) = c_1^2 \lambda_1^{2k} (\mathbf{e}_1, \mathbf{e}_1) + O(|\lambda_1^k \lambda_2^k|) \,,$$

$$\left(\mathbf{x}^{k+1}, \mathbf{x}^k\right) = c_1^2 \lambda_1^{2k+1} (\mathbf{e}_1, \mathbf{e}_1) + O(|\lambda_1^{k+1} \lambda_2^k|) \,.$$

Теперь рассмотрим величину

$$\lambda^{(k)} = \frac{(\mathbf{x}^{k+1}, \mathbf{x}^k)}{||\mathbf{x}^k||_2^2} = \frac{c_1^2 \lambda_1^{2k+1} (\mathbf{e}_1, \mathbf{e}_1) + O(|\lambda_1^{k+1} \lambda_2^k|)}{c_1^2 \lambda_1^{2k} (\mathbf{e}_1, \mathbf{e}_1) + O(|\lambda_1^k \lambda_2^k|)} \,.$$

После деления числителя и знаменателя на первое слагаемое в знаменателе получим

$$\lambda^{(k)} = \frac{\lambda_1 + O(|\lambda_2/\lambda_1|^k)}{1 + O(|\lambda_2/\lambda_1|^k)} = \lambda_1 + O(|\lambda_2/\lambda_1|^k) \,.$$

В числителе дроби множитель $|\lambda_1|$ был для удобства убран в константу асимптотики. Утверждение доказано.

Заметим, если матрица A является симметричной, то для степенного метода справедлива оценка $\lambda^{(k)} = \lambda_1 + O(|\lambda_2/\lambda_1|^{2k})$. Эта формула следует из приведенных выше выкладок с учетом ортогональности собственных векторов.

Для вычисления наименьшего по модулю собственного значения невырожденной матрицы можно применять **метод обратной итерации**, соответствующий степенному методу для матрицы A^{-1}:

$$\mathbf{x}^k := \mathbf{x}^k/\|\mathbf{x}^k\|_2,\ A\mathbf{x}^{k+1} = \mathbf{x}^k,\ \lambda^{(k)} = \frac{(\mathbf{x}^k, \mathbf{x}^{k+1})}{(\mathbf{x}^{k+1}, \mathbf{x}^{k+1})}.$$

Отметим, что при этом на каждом шаге алгоритма требуется решать систему $A\mathbf{x}^{k+1} = \mathbf{x}^k$.

Степенной метод и метод обратной итерации можно также применять к матрице $A - cI$, где I — единичная матрица, что позволяет влиять на сходимость с помощью сдвига c. Например, если исходная матрица вырождена или близка к вырожденной, то в качестве параметра сдвига можно использовать $c = 1/N$, где N — достаточно большое число. Если же с высокой точностью известно приближение $\tilde{\lambda}$ к некоторому собственному числу λ, тогда метод обратной итерации с параметром $c = \tilde{\lambda}$ обычно сходится за несколько итераций, т.е. очень быстро. Часто в качестве параметра сдвига применяют *отношение Рэлея* $R_A(\mathbf{x}) = (A\mathbf{x}, \mathbf{x})/(\mathbf{x}, \mathbf{x})$, что приводит к кубической сходимости, т.е.

$$\lim_{k\to\infty} \left| \frac{\varphi_{k+1}}{\varphi_k^3} \right| \le 1,$$

где φ_k — угол между $\mathbf{x}^k$ и $\mathbf{x}$ (приближением и точным собственным вектором). Отметим однако, что скорость сходимости всегда существенно замедляется при вычислении одного из группы близких собственных значений.

Полная проблема собственных значений. QR – алгоритм

Для вычисления **всех** собственных значений и векторов матрицы небольшого размера, как правило, применяется QR–**алгоритм**: пусть задана квадратная матрица A размера $n \times n$, положим $A_0 = A$ и вычислим $A_0 = Q_0 R_0$, где Q_0 — ортогональная (в комплексном случае — унитарная) матрица, R_0 — верхнетреугольная матрица; далее определим $A_1 = R_0 Q_0$, т.е. перемножим полученные в результате разложения матрицы в обратном порядке. Итак, на каждом шаге вычисляется QR–разложение матрицы $A_k = Q_k R_k$ и вычисляется $A_{k+1} = R_k Q_k$. Матрицы A_k и A_{k+1} ортогонально подобны, т.е. $A_{k+1} = Q_k^T A_k Q_k$, поэтому все их собственные значения совпадают с учетом кратности.

Целью алгоритма является получение предельной матрицы R_∞, с диагонали которой извлекается информация о модулях собственных значений исходной матрицы. Сами же значения определяются на основе структуры матрицы Q_∞. Отметим, что в случае невырожденной матрицы QR–разложение с положительными элементами r_{ii} треугольной матрицы R единственно (см. лекцию 15), поэтому будем считать в дальнейшем $r_{ii} \ge 0$ для произвольной матрицы.

Утверждение 2. *Если A — нормальная вещественная матрица ($A^T A = A\,A^T$), то последовательность треугольных матриц $\{R_k\}$ из QR–алгоритма сходится к диагональной матрице.*

Доказательство. Рассмотрим две соседние матрицы QR–алгоритма, обозначая их для простоты через A и B. Переход от A к B описывается

формулами:

$$A = Q\,R\,, \quad B = R\,Q\,. \tag{1}$$

Пусть $\mathbf{b}^i, \mathbf{a}^i, \mathbf{r}^i$ — столбцы, а $\mathbf{b}_i, \mathbf{a}_i, \mathbf{r}_i$ — строки соответственно матриц $B, A, R\,(i = 1, \ldots, n)$. Так как ортогональные преобразования не меняют евклидову длину вектора, то из (1) следует: $\|\mathbf{a}^i\|_2 = \|\mathbf{r}^i\|_2$, $\|\mathbf{r}_i\|_2 = \|\mathbf{b}_i\|_2\,(i = 1, \ldots, n)$. Из нормальности матриц A и B имеем: $\|\mathbf{a}^i\|_2 = \|\mathbf{a}_i\|_2$, $\|\mathbf{b}^i\|_2 = \|\mathbf{b}_i\|_2\,(i = 1, \ldots, n)$. Положим

$$\Delta_m = \sum_{i=1}^{m} (\|\mathbf{b}_i\|_2^2 - \|\mathbf{a}_i\|_2^2), \quad m = 1, \ldots, n-1\,.$$

Тогда

$$\begin{aligned} \Delta_1 &= \|\mathbf{b}_1\|_2^2 - \|\mathbf{a}_1\|_2^2 = \|\mathbf{r}_1\|_2^2 - \|\mathbf{r}^1\|_2^2 = |r_{12}|^2 + \cdots + |r_{1n}|^2\,, \\ \Delta_m &= \sum_{i=1}^{m} \|\mathbf{r}_i\|_2^2 - \sum_{i=1}^{m} \|\mathbf{r}^i\|_2^2 = \sum_{i=1}^{m} \sum_{j=m+1}^{n} |r_{ij}|^2\,, \end{aligned} \tag{2}$$

при $m = 2, \ldots, n-1$. Отметим, что все $\Delta_m > 0$, так как в противном случае можно уменьшить размерность и провести рассуждения для треугольных подматриц R меньшего размера. Если теперь составить для каждого k величины $\delta_m^{(k)} = \sum\limits_{i=1}^{m} \|\mathbf{a}_i^{(k)}\|_2^2$, где $\mathbf{a}_i^{(k)}$ — строки матрицы A_k и $m = 1, \ldots, n-1$, то получим $n-1$ последовательностей $\{\delta_m^{(k)}\}$. Из (2) следует, что каждая из этих последовательностей монотонно возрастает и каждая ограничена (например, общим значением квадрата евклидовой нормы матриц A_k, т.е. $\|A\|_E = \sum\limits_{i,j=1}^{n} |a_{ij}|^2$). Значит, последовательности $\{\delta_m^{(k)}\}$ сходятся. Соответствующие последовательности $\{\Delta_m^{(k)}\}$, где $\Delta_m^{(k)} = \delta_m^{(k+1)} - \delta_m^{(k)}$, сходятся к нулю, а вместе с тем сходятся к нулю все наддиагональные элементы матриц R_k.

Так как $\|\mathbf{b}_1\|_2^2 = \|\mathbf{r}_1^{(k)}\|_2^2 = \|\mathbf{a}_1^{(k+1)}\|_2^2$, то

$$|r_{11}^{(k)}|^2 = \delta_1^{(k+1)} - |r_{12}^{(k)}|^2 - \cdots - |r_{1n}^{(k)}|^2\,.$$

По соглашению, $r_{11}^{(k)} \geq 0$ для всех k, поэтому последовательность $\{r_{11}^{(k)}\}$ имеет предел. Точно так же из равенств

$$\begin{aligned} |r_{22}^{(k)}|^2 &= \|\mathbf{r}_2^{(k)}\|_2^2 - |r_{23}^{(k)}|^2 - \cdots - |r_{2n}^{(k)}|^2 = \\ &= \|\mathbf{a}_2^{(k+1)}\|_2^2 - |r_{23}^{(k)}|^2 - \cdots - |r_{2n}^{(k)}|^2 = \\ &= \delta_2^{(k+1)} - \delta_1^{(k+1)} - |r_{23}^{(k)}|^2 - \cdots - |r_{2n}^{(k)}|^2 \end{aligned}$$

выводим, что последовательность $\{r_{22}^{(k)}\}$ также сходится. Продолжая рассуждения аналогичным образом, установим существование предела матричной последовательности $\{R_k\}$. Утверждение доказано.

Покажем, что нормальная матрица A вида $A = Q\,D$, где Q — ортогональная матрица, а D — диагональная матрица с неотрицательными элементами, с точностью до симметричной перестановки строк и столбцов является блочно диагональной. При этом каждый диагональный блок лишь скалярным множителем отличается от ортогональной матрицы соответствующего порядка.

Напомним, что условие $A\,A^T = A^T A$ эквивалентно существованию собственного ортонормированного базиса у матрицы A, из него же имеем $D^2 = Q\,D^2 Q^T$, или $D^2 Q = Q\,D^2$. Отсюда получаем поэлементное равенство $q_{ij}(d_{ii}^2 - d_{jj}^2) = 0 \quad \forall i,j$ и $q_{ij} = 0$, если $d_{ii} \neq d_{jj}$. Матрица перестановок P, группирующая равные диагональные элементы матрицы D : $D \to \tilde{D} = P^T D P$, приводит Q к блочно диагональному виду: $Q \to \tilde{Q} = P^T Q P$. Но тогда и матрица $\tilde{A} = P^T A P = (P^T Q P)(P^T D P) = \tilde{Q}\tilde{D}$ — блочно диагональная, причем каждый диагональный блок есть произведение одноименного блока ортогональной матрицы Q на число d, отвечающее этому блоку.

При практическом применении метода сначала проводится масштабирование (уравновешивание) матрицы, сближающее ее норму со спектральным радиусом, и приведение к верхней форме Хессенберга H ($h_{ij} = 0$ при $i > j+1$), которая инвариантна относительно QR–итераций. Само же разложение используется *со сдвигами* c_k, т.е. применяется к матрицам вида $A_k = H_k - c_k\,I$.

В общем случае QR-алгоритм сходится *по форме* к блочно-треугольной матрице R, на диагонали которой находятся модули собственных значений исходной матрицы A.

Лекция 18

Метод простой итерации для нелинейных уравнений. — Метод Ньютона. — Методы установления для решения нелинейных уравнений.

Метод простой итерации для нелинейных уравнений

Пусть H — полное метрическое пространство с метрикой $\rho(x,y)$. Рассмотрим отображение $g: H \to H$. Методом простой итерации для решения уравнения $x = g(x)$ называется алгоритм вида $x^{n+1} = g(x^n)$ с некоторым заданным начальным приближением x^0.

Стандартная запись уравнения $f(x) = 0$ может быть приведена к виду $x = g(x)$ с помощью замены, например, $g(x) = x + \psi(x)f(x)$, где $\psi(x)$ — произвольная непрерывная знакопостоянная функция.

Напомним определение. Отображение $g(x)$ называется сжимающим, если для любых $x, y \in H$ справедливо неравенство $\rho(g(x), g(y)) \le q\rho(x,y)$ с постоянной $0 \le q < 1$.

Утверждение 1. *Если отображение $g(x)$ — сжимающее, то уравнение $x = g(x)$ имеет единственное решение z и справедливо неравенство*

$$\rho(z, x^n) \le \frac{q^n a}{1-q}, \qquad \text{где} \quad a = \rho(x^0, x^1)\,.$$

Доказательство. Имеем из сжимаемости отображения

$$\rho(x^{n+1}, x^n) = \rho(g(x^n), g(x^{n-1})) \le q\rho(x^n, x^{n-1})\,.$$

Продолжая цепочку неравенств, получим $\rho(x^{n+1}, x^n) \le q^n a$. При $l > n$ из неравенства треугольника ($\rho(x,t) \le \rho(x,y) + \rho(y,t)$) имеем цепочку неравенств

$$\rho(x^l, x^n) \le \rho(x^l, x^{l-1}) + \rho(x^{l-1}, x^{l-2}) + ... + \rho(x^{n+1}, x^n) \le$$
$$\le q^{l-1}a + q^{l-2}a + ... + q^n a \le q^n a(1 + q + ... + q^{l-1-n}) \le$$
$$\le q^n a \sum_{j=0}^{\infty} q^j = \frac{q^n a}{1-q}\,.$$

По критерию Коши ($\forall \varepsilon > 0\ \exists N = N(\varepsilon):\ \forall l \ge N, \forall n \ge N \quad \rho(x^l, x^n) \le \varepsilon$) в полном пространстве H последовательность $x^0, x^1, ..., x^l, ...$ сходится, т.е. $\lim\limits_{l\to\infty} x^l = z$, поэтому переходя к пределу по l получаем $\rho(z, x^n) \le q^n a/(1-q)$. В завершение доказательства надо показать, что z является решением уравнения $x = g(x)$, т.е. справедливо $\rho(z, g(z)) = 0$. Рассмотрим цепочку соотношений

$$\rho(z, g(z)) \le \rho(z, x^{n+1}) + \rho(x^{n+1}, g(z)) \le \rho(z, x^{n+1}) +$$
$$+\rho(g(x^n), g(z)) \le \rho(z, x^{n+1}) + q\rho(x^n, z) \le 2\frac{q^{n+1}a}{1-q}\,.$$

Так как n – произвольное, то $\rho(z, g(z)) = 0$, или $z = g(z)$.

Отметим единственность решения. Пусть $z = g(z),\ u = g(u),\ z \ne u$. Тогда $\rho(z,u) = \rho(g(z), g(u)) \le q\,\rho(z,u)$. Так как $0 \le q < 1$, из неравенства следует $\rho(z,u) = 0$. Утверждение доказано.

Различные методы решения нелинейных уравнений удобно сравнивать с помощью понятия *порядка метода*. Приведем его определение. *Наибольшее число k называется порядком метода, если существуют положительные конечные постоянные C_1 и C_2 такие, что справедливо неравенство*

$$\rho(x^{n+1}, z) \le C_2[\rho(x^n, z)]^k \quad \textit{при условии } \rho(x^n, z) \le C_1 \ \forall n \ge 0.$$

В соответствии с этим определением метод простой итерации имеет первый порядок с $C_2 = q < 1$. Чем выше порядок, тем быстрее сходимость к решению. Однако, методы высоких порядков часто требуют большого объема вычислительной работы. Поэтому на практике, как правило, применяют методы первого и второго порядков.

Метод Ньютона

Пусть $F(x)$ — оператор, отображающий линейное нормированное пространство H на линейное нормированное пространство Y (возможно их совпадение) — $F : H \to Y$. Линейный оператор $P : H \to Y$ назовем производной оператора $F(x)$ в точке x, если

$$\|F(x+\eta) - F(x) - P\,\eta\|_Y = o(\|\eta\|_H) \quad \text{при} \quad \|\eta\|_H \to 0\,.$$

Будем обозначать далее P через $F'(x)$. Пусть, например, $H = Y = R^m$, $x = (x_1, .., x_m)^T$, $F(x) = (f_1(x), \ldots, f_m(x))^T$ и функции f_i — непрерывно дифференцируемы в окрестности точки x. Тогда

$$f_i(x_1+\eta_1, x_2+\eta_2, \ldots, x_m+\eta_m) = f_i(x_1, .., x_m) + \sum_{j=1}^{m} \left.\frac{\partial f_i}{\partial x_j}\right|_{x=(x_1,..,x_m)^T} \eta_j + o(\|\eta\|),$$

т.е. оператор производной $F'(x)$является матрицей Якоби в точке x :

$$F'(x) = \left\{ \left[\frac{\partial f_i}{\partial x_j}\right]\Big|_{x=(x_1,..,x_m)^T} \right\}_{i,j=1}^{m}.$$

При $m = 1$ оператор $F'(x)$ превращается в обычную производную.

Пусть z — решение уравнения $F(x) = 0$, x^n — некоторое приближение к z. В предположении существования $F'(x)$ имеем

$$\|F(z) - F(x^n) - F'(x^n)(z - x^n)\|_Y = o(\|z - x^n\|_H)\,,$$

поэтому при малой правой части можно написать приближенное равенство

$$F(x^n) + F'(x^n)(z - x^n) \approx F(z) = 0\,.$$

Возьмем в качестве x^{n+1} решение линейного уравнения (при $m > 1$ — системы)

$$F(x^n) + F'(x^n)(x^{n+1} - x^n) = 0\,.$$

Это и есть метод Ньютона.

Обозначим через $\Omega_h = \{x : \|z - x\|_H < h\}$ открытую (!) h – окрестность решения, и пусть при некоторых $h, a_1, a_2 : 0 < h, 0 < a_1, a_2 < \infty$, выполнены условия:

$$1) \quad \|(F'(x))^{-1} y\|_H \le a_1 \|y\|_Y \quad \forall x \in \Omega_h,\ y \in Y;$$

$$2)\|F(u_1) - F(u_2) - F'(u_2)(u_1 - u_2)\|_Y \le a_2\|u_1 - u_2\|_H^2 \quad \forall u_1, u_2 \in \Omega_h\,.$$

Обозначим также $c = a_1 a_2$, $b = \min\left(h, c^{-1}\right)$.

Утверждение 2. *При условиях* 1), 2) *и* $x^0 \in \Omega_b$ *метод Ньютона сходится с оценкой погрешности*

$$\|x^n - z\|_H \le c^{-1}\left(c\|x^0 - z\|_H\right)^{2^n}\,,$$

т. е. имеет второй порядок сходимости.

Доказательство. Пусть $x^0 \in \Omega_b$. Индукцией по n покажем, что все $x^n \in \Omega_b$. Пусть это верно при некотором n. Так как $b \le h$, то $x^n \in \Omega_h$. Воспользуемся условием 2) при $u_1 = z$ и $u_2 = x^n$, получим

$$\|F(z) - F(x^n) - F'(x^n)(z - x^n)\|_Y \le a_2\|z - x^n\|_H^2\,.$$

Поскольку $F(x^n) = -F'(x^n)(x^{n+1} - x^n)$, а $F(z) = 0$, то из последней оценки получаем

$$\|F'(x^n)(z - x^{n+1})\|_Y \le a_2\|z - x^n\|_H^2.$$

Теперь воспользуемся условием 1) и получим:

$$\|z - x^{n+1}\|_H \le c\|z - x^n\|_H^2\,, \tag{1}$$

откуда следует

$$\|z - x^{n+1}\|_H < c\,b^2 = (cb)b \le b\,.$$

Таким образом, при $x^0 \in \Omega_b$ все $x^n \in \Omega_b$ и для них выполнено неравенство (1).

Обозначим $q_n = c\|z - x^n\|_H$. После умножения неравенства (1) на c имеем $q_{n+1} \le q_n^2 \le q_{n-1}^4 \le \ldots$. Индукцией по n покажем, что $q_n \le q_0^{2^n}$. При $n = 0$ оно очевидно. Предположив, что оно верно при $n = k$, получаем

$$q_{k+1} \le q_k^2 \le \left(q_0^{2^k}\right)^2 = q_0^{2^{k+1}}\,.$$

Таким образом, $q_n \le q_0^{2^n}$ при всех n. Это означает, что $c\|z - x^n\|_H \le \left(c\|z - x^0\|_H\right)^{2^n}$. По определению величин c и b имеем $c\|z - x^0\|_H < cb \le 1$, поэтому $x^n \to z$. Утверждение доказано.

Отметим, что в методе Ньютона требуется решать на каждой итерации новую систему линейных уравнений. Когда эта процедура трудоемка, возможно использование модифицированного метода

$$x^{n+1} - x^n = -\left[F'(x^0)\right]^{-1} F(x^n)\,.$$

Здесь требуется вычислить только один раз матрицу $[F'(x^0)]^{-1}$, но зато метод будет сходиться медленнее (его порядок равен $k = (\sqrt{5}+1)/2 \approx 1.62$).

Сделаем замечание о сходимости метода, если корень является кратным, например, $f(x) = x^2$. В этом случае сходимость замедляется, превращаясь из квадратичной в линейную.

Итерации x^n по методу Ньютона

$$x^{n+1} - x^n = -[F'(x^n)]^{-1}F(x^n) \equiv \Phi(x^n)$$

можно интерпретировать как значения некоторой функции $x(t)$ в моменты времени $t_n = n$. Это дает возможность записать алгоритм в виде

$$\frac{x^{n+1} - x^n}{\Delta t_n} \equiv \frac{x^{n+1} - x^n}{t_{n+1} - t_n} = \Phi(x^n)$$

и перейти к пределу при $\Delta t_n \to 0$. В предположении достаточной гладкости $x(t), \Phi(x)$ будем иметь

$$\frac{dx}{dt} = \Phi(x)\,.$$

Если z — решение уравнения $F(x) = 0$, то эта же точка является стационарной для полученной системы дифференциальных уравнений. Действительно,

$$0 = \frac{dz}{dt} = -[F'(z)]^{-1}F(z) = 0\,.$$

Проведем исследование ее устойчивости. Для этого положим $x(t) = z + \eta(t)$ и в окрестности стационарной точки применим формулу Тейлора. Пусть $g(x) = [F'(x)]^{-1}$ и $g(z + \eta) = g(z) + O(\|\eta\|)$, тогда

$$[F'(x)]^{-1} = [F'(z)]^{-1} + O(\|\eta\|)\,.$$

Аналогично для $g(x) = F(x)$ и $g(z+\eta) = g(z)+g'(z)\eta+O(\|\eta\|^2)$ при $F(z) = 0$ получаем

$$F(x) = F'(z)\eta + O(\|\eta\|^2)\,.$$

Отсюда следует, что

$$\frac{dx}{dt} = \frac{d\eta}{dt} = -\{[F'(z)]^{-1} + O(\|\eta\|)\} \cdot [F'(z)\eta + O(\|\eta\|^2)] = -\,\eta + O(\|\eta\|^2)\,.$$

Это означает, что в окрестности стационарной точки z возмущение $\eta(t)$ удовлетворяет дифференциальному уравнению

$$\frac{d\eta}{dt} = -\eta, \quad \text{т.е.} \quad \eta(t) = Ce^{-t},$$

следовательно точка $x = z$ является асимптотически устойчивой (все траектории будут стягиваться к z).

Этот факт позволяет обобщить метод Ньютона за счет введения переменного итерационного параметра Δ_n :

$$x^{n+1} - x^n = -\,\Delta_n[F'(x^n)]^{-1}F(x^n)\,.$$

Само значение Δ_n на каждом шаге можно выбирать из условия минимума функции $\Psi = \|\Phi(x)\|^2 \equiv \|[F'(x)]^{-1}F(x)\|^2$ (или $\Psi = \|F(x)\|^2$) в точке $x = x^{n+1}$. При этом нахождение искомого значение Δ_n эквивалентно поиску минимума функции одной переменной, так как в выражении для x^{n+1} неизвестно только значение скалярного параметра Δ_n. Конечно, следует

иметь в виду, что вычисление значений $\Psi(x)$ или их норм может существенно увеличить трудоемкость алгоритма. Поэтому использование такого подхода должно быть оправдано сложностью задачи.

Попробуем расширить применение обсуждаемой идеи, введя в рассмотрение уравнение второго порядка

$$\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} = -\left[F'(x)\right]^{-1}F(x).$$

Рассмотрим вопрос о выборе параметра γ при $[F'(x)]^{-1}F(x) \approx \eta$ (в окрестности стационарной точки z, т.е. при $x(t) = z + \eta(t)$). В этом случае будем иметь

$$\frac{d^2\eta}{dt^2} + \gamma\frac{d\eta}{dt} + \eta = 0\,.$$

Его характеристическое уравнение $\lambda^2 + \gamma\lambda + 1 = 0$ имеет корни $\lambda_{1,2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - 1}$. Отсюда, в частности, следует, что параметр γ должен быть положителен: в противном случае стационарная точка z не является асимптотически устойчивой.

Из приведенных соображений при $\gamma > 0$ можно рассмотреть трехслойный итерационный метод вида

$$\frac{x^{n+1} - 2x^n + x^{n-1}}{(\Delta_n)^2} + \gamma\frac{x^{n+1} - x^{n-1}}{2\Delta_n} = \Phi(x^n),$$

где параметр Δ_n выбирается из тех же соображений, что и выше (минимума $\|\Phi(x^{n+1})\|$ или $\|F(x^{n+1})\|$).

Методы установления, как правило, имеют более широкую область сходимости (относительно выбора начального приближения), но сходятся медленнее, чем метод Ньютона. Поэтому на практике сначала каким – либо методом интегрируют систему дифференциальных уравнений, а затем в окрестности корня используют метод Ньютона для увеличения скорости сходимости.

Лекция 19

Явный метод Эйлера для обыкновенных дифференциальных уравнений (ОДУ). Устойчивость. Локальная и глобальная ошибки. — Явные методы Рунге – Кутты.

Для изучения способов дискретизации по времени нестационарных задач наиболее удобной моделью является задача Коши для обыкновенного дифференциального уравнения (ОДУ):

$$y'(x) = f(x, y)\,, \quad y(x_0) = y_0\,. \tag{0}$$

Сразу сделаем замечание, что рассматриваемые подходы имеют одинаковую форму записи как для **одного** уравнения, так и для **системы**. Поэтому для упрощения изложения будем, где это возможно, рассматривать только одно уравнение.

Напомним свойство *асимптотической устойчивости* ОДУ. Для одного уравнения $y'(x) = f(x, y)$ рассмотрим два решения, порождаемые различными начальными данными. Значение $y_1(x_0) = \bar{y}_1$ определяет решение $y_1(x)$ и соответственно — $y_2(x_0) = \bar{y}_2$ приводит к $y_2(x)$. Если $\forall \bar{y}_1, \bar{y}_2$ выполнено $\lim\limits_{x\to\infty} |y_1(x) - y_2(x)| = 0$, то такое ОДУ называют асимптотически устойчивым. Простой пример: уравнение $y' + Ay = 0$, $A = \text{const}$ имеет общее решение $y(x) = C\,e^{-Ax}$, которое является асимптотически устойчивым при $A > 0$. Если $A < 0$, то решения $y_1(x)$ и $y_2(x)$ будут удаляться друг от друга для сколь угодно близких начальных значений $\bar{y}_1$ и $\bar{y}_2$. Уравнения с таким поведением решений называют неустойчивыми. Бывает ситуация еще хуже: задача $y' + y^2 = 0$, $y(0) = A$, $A = \text{const}$ имеет решение $y(x) = A/(1 + Ax)$. При $A \geq 0$ решение равномерно ограничено по x и асимптотически устойчиво; при $A < 0$ в момент $x_* = -1/A$ оно обращается в бесконечность. Про решение такого типа говорят, что оно имеет вертикальную асимптоту (носит blow-up (взрывной) характер). В дальнейшем для анализа различных методов решения задачи Коши для ОДУ мы будем использовать только гладкие решения асимптотически устойчивых уравнений.

Явный метод Эйлера для обыкновенных дифференциальных уравнений (ОДУ). Устойчивость. Локальная и глобальная ошибки

Познакомимся с простейшим методом решения ОДУ — **явным методом Эйлера**. Пусть $x_{n+1} = x_n + h\,, n \geq 0\,, h > 0$ — постоянный шаг интегрирования. Обозначим за y_n приближенное значение точного решения $y(x_n)$. Тогда метод Эйлера для задачи (0) можно записать в виде

$$y_{n+1} = y_n + hf(x_n, y_n), \quad y_0 \text{ задано}, \quad n \geq 0\,. \tag{1}$$

Проанализируем сначала устойчивость метода на модельном примере $f(x, y) = -Ay$, $y(0) = y_0$, $A > 0$:

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

Если $h > \dfrac{2}{A}$, то величина $1 - hA$ отрицательна и по модулю больше единицы. В этом случае приближенное решение y_n знакопеременно и возрастает по

абсолютной величине. Точное же решение знакопостоянно и стремится к нулю. Это означает неустойчивость метода. Если мы хотим иметь $|y_n| \to 0$, то должны выбирать h из условия $h < \dfrac{2}{A}$. Такую ситуацию называют **условной** устойчивостью метода. Отсюда следует важный вывод, что даже для *устойчивого* ОДУ метод может быть как *устойчивым*, так и *неустойчивым* в зависимости от величины шага интегрирования.

Рассмотрим теперь структуру погрешности метода Эйлера, считая h достаточно малым. Если точное решение $y(x)$ имеет вторую непрерывную производную, то справедливо

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(\bar{x}) = y(x_n) + hf(x_n, y(x_n)) + \frac{h^2}{2}y''(\bar{x})\,, \quad (2)$$

где $x_n \le \bar{x} \le x_{n+1}$. Сравнивая последнее выражение с формулой Эйлера, получаем, что, стартуя даже с **точного** значения $y(x_n) = y_n$, мы имеем ошибку в разности $y(x_{n+1}) - y_{n+1}$, пропорциональную величине h^2, т.е. $h^2y''(\bar{x})/2$. Эта часть погрешности метода называется **локальной ошибкой** (ошибкой на шаге). Отметим, что вместо термина "ошибка"часто применяют его синоним — "погрешность". Но значение y_n само получено по формуле Эйлера, т.е. содержит в себе уже n локальных ошибок, да и вычисления величин $f(x_n, y_n)$ все n раз могли быть проделаны неточно. Поэтому важно выражение для **глобальной** (полной) ошибки на **конечном** интервале интегрирования. К нему можно прийти, вычитая из (2) формулу (1):

$$y(x_{n+1}) - y_{n+1} = y(x_n) - y_n + h[f(x_n, y(x_n)) - f(x_n, y_n)] + \frac{h^2}{2}y''(\bar{x})\,.$$

По формуле конечных приращений Лагранжа можно записать

$$f(x_n, y(x_n)) - f(x_n, y_n) = (y(x_n) - y_n)f_y(x_n, \tilde{y}_n),$$

где $\tilde{y}_n$ находится между значениями $y(x_n)$ и y_n. Окончательно получим

$$y(x_{n+1}) - y_{n+1} = [1 + h\, f_y(x_n, \tilde{y}_n)](y(x_n) - y_n) + \frac{h^2}{2}y''(\bar{x})\,.$$

В этом выражении первое слагаемое в правой части называется **распространяемой** или переносимой ошибкой. В рассмотренном выше модельном примере с $f(x, y) = -Ay$ мы видели, что распространяемая ошибка может полностью исказить решение, если шаг h достаточно велик.

Таким образом, имеются два основных фактора, которые необходимо учитывать при построении методов решения задачи Коши для ОДУ: уменьшение локальной ошибки (как степени h) и невозрастание распространяемой ($|1 + h\, f_y(x, y)| \le 1$). Предположим, что нам удалось все это учесть, и мы хотим понять асимптотику ошибки по h на **конечном** интервале интегрирования $[x_0, X]$, содержащим N шагов длины h. Ответ несложен: если мы движемся с шагом h, то должны сделать $N = O(h^{-1})$ шагов; каждый шаг в методе Эйлера порождает ошибку порядка $O(h^2)$, значит суммарная ошибка оценивается произведением $N \cdot O(h^2) = O(h)$. Это означает, что для дюбого конечного интервала справедливо $y_N = y(X) + O(h)$. Поэтому метод Эйлера является методом **первого** порядка точности. Отметим, что для устойчивого метода порядок точности всегда есть величина на единицу меньше, чем порядок локальной ошибки.

Явные методы Рунге – Кутты

Пусть, как и ранее, в точке x считается известным значение $y(x)$ и требуется построить приближение к $y(x+h)$ в точке $x+h$, где h — шаг интегрирования. Такая постановка приводит к *одношаговым* методам. Типичными представителями этой группы методов являются **явные** методы Рунге-Кутты. Изложим идею их построения. Пусть для некоторого натурального фиксированного $q \geq 2$ заданы параметры:

$$\alpha_2, \alpha_3, \ldots, \alpha_q, \quad p_1, p_2, \ldots, p_q, \quad \beta_{i,j}, \quad 0 < j < i \leq q\,.$$

Будем последовательно вычислять

$$\begin{aligned}
&k_1 = h\,f(x,y)\,,\\
&k_2 = h\,f(x+\alpha_2 h, y+\beta_{2,1}k_1)\,,\\
&k_3 = h\,f(x+\alpha_3 h, y+\beta_{3,1}k_1+\beta_{3,2}k_2)\,,\\
&\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots\\
&k_q = h\,f(x+\alpha_q h, y+\beta_{q,1}k_1+\cdots+\beta_{q,q-1}k_{q-1})\,.
\end{aligned}$$

В качестве приближения к $y(x+h)$ возьмем величину $z(h)$:

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

Это общая схема вычислений. Откуда берутся значения параметров? Они определяются из соображений минимизации по порядку величины h локальной ошибки. Рассмотрим функцию

$$\varphi(h) = y(x+h) - z(h)\,,$$

т.е. локальную ошибку. Будем считать, что $f(x,y)$ настолько гладкая функция, что у $\varphi(h)$ существуют непрерывные производные до $(s+1)-$го порядка включительно, тогда искомые параметры можно выбрать из условий: $\varphi(0) = \varphi'(0) = \ldots = \varphi^{(s)}(0) = 0$. Так как справедлива формула Тейлора

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

то локальная ошибка по порядку равна $\varphi(h) = O(h^{s+1})$, и мы получили метод $s-$го порядка точности.

В качестве самостоятельного упражнения полезно разобрать, что явный метод Эйлера можно интерпретировать в качестве предшественника семейства Рунге – Кутты (его простейшего представителя с $q = 1$).

Рассмотрим несложный пример. Пусть $q = 2$, тогда получим

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

где обозначено: $\bar{x} = x + \alpha_2 h, \bar{y} = y(x) + \beta_{2,1}hf(x,y)$. Отметим, что $\varphi(0) = 0$, и выпишем соотношения для производных по параметру h:

$$\begin{aligned}
&f'_h(\bar{x},\bar{y}) = \alpha_2 f_x(\bar{x},\bar{y}) + \beta_{2,1}f_y(\bar{x},\bar{y})f(x,y),\\
&f''_h(\bar{x},\bar{y}) = \alpha_2^2 f_{xx}(\bar{x},\bar{y}) + 2\alpha_2\beta_{2,1}f_{xy}(\bar{x},\bar{y})f(x,y) + \beta_{2,1}^2 f_{yy}(\bar{x},\bar{y})(f(x,y))^2.
\end{aligned}$$

Из исходного уравнения имеем:

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

Теперь будем последовательно обнулять производные функции $\varphi(h)$ для $h = 0$ (при этом $\bar{x} = x$, $\bar{y} = y$):

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

В общем случае последнее слагаемое в выражении для $\varphi'''(0)$ в нуль не обращается, поэтому локальная ошибка будет $O(h^3)$, т.е. мы можем построить метод порядка не выше чем 2. Для определения самих параметров имеем систему трех уравнений с четырьмя неизвестными:

$$1 - p_1 - p_2 = 0\,, \quad 1 - 2\alpha_2 p_2 = 0\,, \quad 1 - 2p_2\beta_{2,1} = 0\,.$$

Выбирая какое–либо значение p_2, можно однозначно определить остальные. Наиболее употребительным здесь является выбор $p_2 = 1$ или $p_2 = 1/2$. Ситуация, приводящая сразу к группе методов при фиксированном q, является типичной для семейства Рунге–Кутты.

Рассмотренные методы имеют выраженные достоинства:
1. Собственно „одношаговость“, т.е. схожесть исходной постановки с дифференциальной задачей: по значению искомой функции в предыдущей точке находится ее значение в следующей.
2. Легко менять шаг интегрирования. Например, пусть локальная ошибка имеет вид $h^2 y''(\xi)/2$. Тогда если мы хотим добиться ее малости порядка ε, то достаточно выбрать шаг из условия $h \leq \sqrt{2\varepsilon/|y''|}$, где оценку величины $|y''|$ можно получить, дифференцируя исходное уравнение.
3. Вычисления ведутся по явным формулам, не требуя решения каких–либо вспомогательных задач.

К недостаткам методов Рунге–Кутты обычно относят трудоемкость (для метода $s-$го порядка требуется s вычислений правой части, что может быть обременительным при решении систем большого порядка) и ограничение на устойчивость. Для модельной задачи $y' + Ay = 0$, $A > 0$ оно, как правило, имеет вид $h \leq K/A$, где в лучшем случае для K справедлива оценка $2 \leq K \leq 3$.

Лекция 20

Неявные одношаговые методы решения ОДУ. — Многошаговые методы решения ОДУ.

Неявные одношаговые методы решения ОДУ

Рассмотрим другую идею построения одношаговых методов. Проинтегрируем уравнение $y' = f(x, y)$ на отрезке $[x_n, x_{n+1}]$, в результате получим

$$y(x_{n+1}) = y(x_n) + I, \quad \text{где} \quad I = \int\limits_{x_n}^{x_{n+1}} f(x, y(x))dx\,.$$

Известный нам **явный** метод Эйлера соответствует замене интеграла I в правой части на выражение $(x_{n+1}-x_n)f(x_n, y(x_n))$. Столь же равноправной является замена I на — $(x_{n+1} - x_n)f(x_{n+1}, y(x_{n+1}))$, приводящая к **неявному** методу Эйлера

$$y_{n+1} = y_n + h\, f(x_{n+1}, y_{n+1})\,.$$

Замена интеграла I на приближенное выражение по правилу трапеций приводит к методу Кранка–Николсон

$$y_{n+1} = y_n + \frac{h}{2}\big(f(x_n, y_n) + f(x_{n+1}, y_{n+1})\big)\,.$$

С помощью формулы Тейлора несложно установить порядки точности: $s = 1$ для неявного метода Эйлера и, соответственно, $s = 2$ для метода Кранка–Николсон.

Оба эти метода являются одношаговыми, но их реализация не так проста, так как неизвестное значение y_{n+1} входит в формулу неявно (через правую часть). Как правило, для решения таких уравнений используется метод Ньютона или метод *функциональной итерации*: если известно приближение $y_{n+1}^{(k)}$, то для неявного метода Эйлера следующее находится по формуле

$$y_{n+1}^{(k+1)} = y_n + h\, f(x_{n+1}, y_{n+1}^{(k)})\,,$$

Часто используют начальное приближение $y_{n+1}^{(0)} = y_n$.

Проанализируем устойчивость этих методов на модельной задаче $y' + Ay = 0$, $A > 0$. Для неявного метода Эйлера имеем соотношение $y_{n+1} = y_n - hAy_{n+1}$, это дает

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

Так как $|1+hA|^{-1} < 1$, то ошибки накопленные на предыдущем шаге будут не возрастать, а убывать, т.е. неявный метод Эйлера абсолютно устойчив. Метод Кранка – Николсон приводит к соотношению

$$y_{n+1} = \frac{1 - hA/2}{1 + hA/2}\, y_n\,.$$

Легко проверить, что при вещественных положительных A этот метод также абсолютно устойчив.

В завершение раздела сделаем вывод, что от ограничений, связанных с устойчивостью, можно избавиться за счет дополнительной работы — решения нелинейного уравнения (или в векторном случае — системы) для нахождения y_{n+1}.

Многошаговые методы решения ОДУ

Одним из недостатков одношаговых методов является то, что значения решения, вычисленные на предыдущих шагах $(y_{n-1}, y_{n-2}, \ldots, y_0)$, а также зависящие от них правые части $f(x, y)$, не используются в дальнейшем, даже если решение достаточно гладкое (т.е. меняется медленно). Идея использования нескольких предыдущих значений реализована в так называемых многошаговых методах. Пусть мы имеем дело с равноотстоящими узлами $x_n = x_0 + n\,h$, $n \geq 0$. Тогда общую формулу $k-$шагового метода интегрирования можно записать в виде

$$y_{n+1} = \Phi(f; x_{n+1}, x_n, \ldots, x_{n-k+1}, y_n, \ldots, y_{n-k+1})\,,$$

т.е. нужно знать k предыдущих значений $y_n, y_{n-1}, \ldots, y_{n-k+1}$, чтобы вычислить y_{n+1}. Рассмотрим в качестве примера общую формулу **двухшаговых** методов ($k = 2$)

$$y_{n+1} = \alpha_1 y_n + \alpha_2 y_{n-1} + h[\beta_0 f_{n+1} + \beta_1 f_n + \beta_2 f_{n-1}]\,,$$

где использовано обозначение $f_n = f(x_n, y_n)$, а постоянные коэффициенты выбираются из условий наиболее высокого порядка точности при сохранении устойчивости. При $\beta_0 = 0$ метод называется явным, иначе ($\beta_0 \neq 0$) — неявным.

Зафиксируем главное преимущество такого подхода. В многошаговом методе s-го порядка точности для нахождения значения y_{n+1} требуется только одно новое вычисление правой части $f(x, y)$ (в одношаговых методах требовалось s вычислений).

Рассмотрим построение алгоритмов интегрирования **методом неопределенных коэффициентов**. В качестве примера получим двухшаговый алгоритм Адамса ($\alpha_2 = 0$), содержащий четыре неизвестных величины $\alpha_1, \beta_0, \beta_1, \beta_2$, которые определим из условия точного интегрирования (так, что $y_n = y(x_n)$!) четырех уравнений

$$y' = l x^{l-1}: \quad y(x) = x^l, \quad l = 0, 1, 2, 3.$$

В этом случае величина шага не важна, положим для удобства $h = 1$, и при $n = 0$ получим $y(1) = \alpha_1 y(0) + \beta_0 f(1, y(1)) + \beta_1 f(0, y(0)) + \beta_2 f(-1, y(-1)))$. Подстановка различных l приводит к системе

$$\begin{array}{llll} l = 0, & f(x,y) = 0, & y = 1, & 1 = \alpha_1, \\ l = 1, & f(x,y) = 1, & y = x, & 1 = \beta_0 + \beta_1 + \beta_2, \\ l = 2, & f(x,y) = 2x, & y = x^2, & 1 = 2(\beta_0 - \beta_2), \\ l = 3, & f(x,y) = 3x^2, & y = x^3, & 1 = 3(\beta_0 + \beta_2)\,. \end{array}$$

Ее решение дает

$$\alpha_1 = 1,\ \beta_0 = \frac{5}{12},\ \beta_1 = \frac{8}{12},\ \beta_2 = -\frac{1}{12}\,.$$

Отсюда следует искомая формула

$$y_{n+1} = y_n + \frac{h}{12}[5f_{n+1} + 8f_n - f_{n-1}]\,.$$

Для оценки погрешности возьмем очередное значение l :

$$l = 4,\ \ y' = 4x^3,\ \ y = x^4,\ \ 1 \neq 4(\beta_0 - \beta_2)\,,$$

Отсюда следует $y(x_{n+1}) - y_{n+1} = Cy^{(4)}(\xi)h^4$, $C \neq 0$, т.е. порядок точности s построенного метода равен 3.

Исследование устойчивости многошаговых методов основано на изучении корней характеристических уравнений и весьма громоздко. Известен следующий важный результат (Дальквист, 1963): *никакой многошаговый метод не может быть абсолютно устойчивым (даже неявный!), если его порядок выше второго.*

Приведем иллюстрацию к этому утверждению на примере уравнения $y' + Ay = 0$, $A > 0$. Полученная ранее формула неявного двухшагового метода Адамса примет вид

$$y_{n+1} = y_n - \frac{hA}{12}[5y_{n+1} + 8y_n - y_{n-1}]\,.$$

Введем обозначение $p = hA/12 > 0$ и перепишем формулу Адамса в виде разностного уравнения:

$$(1 + 5p)y_{n+1} - (1 - 8p)y_n - py_{n-1} = 0\,.$$

Его характеристическое уравнение

$$(1 + 5p)\lambda^2 - (1 - 8p)\lambda - p = 0$$

имеет корни

$$\lambda_\pm = \frac{1 - 8p \pm \sqrt{D}}{2(1 + 5p)}\,, \quad D = 84p^2 - 12p + 1 > 0\,,$$

Для устойчивости метода корни характеристического уравнения должны удовлетворять неравенству $|\lambda_\pm| \leq 1$, т.е.

$$-1 \leq \lambda_- < \lambda_+ \leq 1\,.$$

Непосредственной проверкой можно убедиться, что $\lambda_+ \leq 1 \quad \forall p > 0$. Решение неравенства $\lambda_- \geq -1$ дает $-1/5 \leq p \leq 1/2$. Отсюда следует условие $0 < p \leq 1/2$, что означает условную устойчивость построенного метода, хотя он является неявным.

Отметим трудности использования многошаговых методов:
1) недостаточно значений для старта алгоритма (как правило, их получают на основе одношаговых методов);
2) смена шага интегрирования h приводит к изменению значений коэффициентов, т.е. самой формулы на нескольких (переходных) отрезках. После получения нужного количества стартовых значений с новым шагом происходит возврат к прежней формуле.

Лекция 21

Жесткие системы ОДУ. Пример. Определение. — Экспоненциальный метод решения жестких систем ОДУ.

Жесткие системы ОДУ. Пример. Определение

Прежде чем определять понятие жесткости, рассмотрим следующий пример

$$\begin{cases} u' - 998u - 1998v &= 0, \quad u(0) = 1, \\ v' + 999u + 1999v &= 0, \quad v(0) = 1, \end{cases}$$

или в матричной форме для вектора $y = (u, v)^T$

$$y' + Ay = 0, \quad y(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad A = \begin{pmatrix} -998 & -1998 \\ 999 & 1999 \end{pmatrix}.$$

Собственные значения матрицы A равны: $\lambda_1(A) = 1$, $\lambda_2(A) = 10^3$, поэтому решение можно записать в виде

$$u(x) = 4e^{-x} - 3e^{-10^3 x}, \quad v(x) = -2e^{-x} + 3e^{-10^3 x}.$$

Пусть требуется приближенно определить решение этой задачи на отрезке $[0, X]$ (например, с $X = 10$). Постараемся понять трудности интегрирования на примере явного метода Эйлера:

$$y_{n+1} = y_n - h\,Ay_n\,.$$

Вспомним условие его устойчивости $h \le 2/\max|\lambda_i(A)|$ и сразу же заметим главную неприятность ситуации. Компонента решения, пропорциональная $e^{-10^3 x}$, от которой зависит ограничение на величину шага интегрирования h, быстро исчезает, т.е. почти на всем отрезке, за исключением малого стартового этапа („пограничного слоя“), справедливы приближенные равенства:

$$u(x) \approx 4e^{-x}, \quad v(x) \approx -2e^{-x}.$$

Но при этом считать с более крупным шагом мы не имеем права, так как переносимая (вычислительная) погрешность сразу же исказит решение до неузнаваемости. Вышесказанное содержит главный смысл жесткости: величина шага интегрирования на **всем** отрезке зависит от того, „чего почти нет“, т.е. от той компоненты решения, норма которой на большей части отрезка интегрирования много меньше нормы всего решения. В общем-то, это не соответствует здравому смыслу. Поэтому для таких систем требуются более экономичные (эффективные) алгоритмы, чем имеющие ограничения на устойчивость вида $h \le \mathrm{const}/\max|\lambda_i(A)|$.

Линейная система вида $y' = Ay, y(x_0) = y_0$ *на отрезке* $[x_0, x_0 + X]$ *называется жесткой, если:*

$$1)\ X \max_i |\lambda_i(A)| \gg 1, \quad 2)\ X \max_i \mathrm{Re}[\lambda_i(A)] \sim 1, \quad 3)\ X \max_i |\mathrm{Im}[\lambda_i(A)]| \sim 1.$$

Устоявшегося определения жестких систем не существует, поэтому будем опираться приведенное как на одно из широко распространенных.

Ограничения, перечисленные в определении, соответственно означают, что в решении имеются быстро убывающие компоненты, но нет быстро растущих, а также, что решение быстро не колеблется. Для рассматриваемых нами асимптотически устойчивых дифференциальных уравнений удобно представлять, что длина отрезка интегрирования $X \sim 1$, а спектр матрицы A принадлежит прямоугольнику $\Lambda = \{-L \le \mathrm{Re}[\lambda] < 0, |\mathrm{Im}[\lambda]| \le l\}$ на комплексной плоскости, где $l \sim 1,\ L \gg 1$.

Про нелинейную систему общего вида $y' = f(x,y), y(x_0) = y_0$ говорят, что она жесткая, если для любого $x \in [x_0, x_0 + X]$ линейная система $y' = Ay, y(x_0) = y_0$ при $A = \dfrac{\partial f}{\partial y}(x, y(x))$ удовлетворяет определению жесткости.

Изначально на такие задачи обратили внимание в химической кинетике, но наиболее типичны они для дискретных аналогов по пространству нестационарных уравнений в частных производных.

Чтобы избавиться от ограничений на шаг h, в первую очередь используют **неявные** методы: Эйлера или Кранка – Николсон для уравнений в частных производных или Гира для небольших систем ОДУ. При этом на каждом шаге требуется решать, вообще говоря, нелинейную систему алгебраических уравнений. Далее мы познакомимся с одной из идей построения **явных** методов интегрирования, в которых привлекательны как простота, так и возможность их распараллеливания.

Экспоненциальный метод решения жестких систем ОДУ

Идею подхода предложил Ракитский Ю.В. Рассмотрим этот метод на примере системы $y' = f(y)$. Пусть имеется некоторое приближение $y_n \approx y(x_n)$. Используем ряд Тейлора

$$f(y) = f(y_n) + \left.\frac{\partial f}{\partial y}\right|_{y=y_n} (y - y_n) + O(\|y - y_n\|^2)\,,$$

и в качестве y_{n+1} примем значение $z(x_{n+1})$, определяемое уравнением

$$z' = f(y_n) + f_y(y_n)(z - y_n)\,, \quad z(x_n) = y_n\,.$$

Сделаем замену переменных:

$$z(x) - y_n = u(x)\,, \quad x - x_n = t\,,$$

и обозначим $b = f(y_n)$, $A = f_y(y_n)$. Таким образом, мы сформулировали модельную задачу, на основе решения которой будем строить решение нашей задачи. Требуется найти для $h = x_{n+1} - x_n$ значение $u(h)$ такое, что

$$u' = Au + b,\ \ u(0) = 0\,.$$

Имеется аналитическое решение этой задачи в виде

$$u(t) = \omega(t)b, \quad \text{где} \quad \omega(t) = \int\limits_0^t e^{A(t-\tau)} d\tau\,,$$

и можно попытаться вычислить ряд

$$\omega(h) = h \sum_{i=1}^{\infty} \frac{1}{i!} (Ah)^{i-1}\,.$$

Но для жестких систем справедливо $\|A\|\,h \gg 1$, поэтому
а) для получения хорошего приближения к $\omega(h)$ надо брать много слагаемых;
б) имеются большие по абсолютной величине члены ряда, и их вычислительные погрешности сильно искажают результат суммирования.

Другими словами, для корректного вычисления ряда требуется выполнение неравенства $\|A\|h < 1$, что, по сути, эквивалентно использованию для исходной задачи явного метода Эйлера ($h < 2/\|A\|$). Нашей же целью является избежать этого ограничения. Как это сделать? Используем замечательное свойство матрицы $\omega(t)$ при $\det A \neq 0$:

$$\omega(t) = \omega(t/2)[2I + A\omega(t/2)],$$

где I — единичная матрица. Действительно, справедливы соотношения:

$$\begin{aligned}\omega(t) &= -A^{-1}\int\limits_0^t de^{A(t-\tau)} = -A^{-1}e^{A(t-\tau)}\Big|_0^t = \\ &= -A^{-1}(I - e^{At}) = A^{-1}(e^{At} - I)\,.\end{aligned}$$

Поэтому

$$A^{-1}(e^{At/2} - I)[2I + e^{At/2} - I] = A^{-1}(e^{At} - I) = \omega(t)\,.$$

Главной идеей метода является использование рекуррентного соотношения: возьмем s такое, что справедливо $\|A\|h2^{-s} < 1$, и с любой заданной точностью вычислим быстросходящийся ряд

$$\omega\left(\frac{h}{2^s}\right) = \frac{h}{2^s}\sum_{i=1}^{\infty}\frac{1}{i!}\left(A\frac{h}{2^s}\right)^{i-1},$$

а затем последовательно определим

$$\omega\left(\frac{h}{2^{s-1}}\right), \omega\left(\frac{h}{2^{s-2}}\right), \ldots, \omega(h)\,.$$

При этом шаг h может быть достаточно большим, так как не связан с устойчивостью (накоплением вычислительной погрешности), а определяется только близостью применяемой линейной и исходной нелинейной задач.

Лекция 22

Основы метода конечных элементов: вариационная постановка, метод Ритца и базисные функции.

Основы метода конечных элементов: вариационная постановка, метод Ритца и кусочно полиномиальные базисные функции

Вариационная постановка задачи. Рассмотрим линейную краевую задачу для обыкновенного дифференциального уравнения 2-го порядка на отрезке $[0,1]$

$$L\,y \equiv -(k(x)y')' + p(x)y = f(x),$$
$$y(0) = 0,\ y'(1) = 0\,,$$

где $k(x), p(x), f(x)$ — достаточно гладкие функции, причем $0 < k_1 \le k(x) \le k_2, 0 \le p(x) \le p_2$. В данном случае под гладкостью функций понимается непрерывность $k'(x)$, $p(x)$, $f(x)$. Эти ограничения гарантируют существование, единственность и непрерывную зависимость решения $y(x)$ от входных данных задачи, причем решение является дважды непрерывно дифференцируемой функцией. Такое решение традиционно называют *классическим*.

Поставим в соответствие дифференциальной задаче вариационную. Сначала для этого возьмем функционал вида $J(v) = (Lv, v) - 2(f, v)$, или

$$J(v) = \int_0^1 \{k(x)[v'(x)]^2 + p(x)v^2(x) - 2f(x)v(x)\}dx\,.$$

Здесь было использовано обозначение $(f,g) = \int\limits_0^1 f(x)g(x)dx$. Отметим, что первое слагаемое под интегралом в $J(v)$ есть результат интегрирования по частям с учетом краевых условий, поэтому для функционала часто используют другое (более симметричное) обозначение $J(v) = a(v,v) - 2(f,v)$, где

$$a(u,v) = \int_0^1 [k(x)u'(x)v'(x) + p(x)u(x)v(x)]dx\,.$$

Затем определим пространство функций, в котором будем минимизировать функционал

$$H = \{v(x) :\ \int_0^1 [(v')^2 + v^2]dx < \infty,\ v(0) = 0\}\,.$$

Легко заметить, что из ограниченности коэффициентов исходного уравнения следует ограниченность сверху функционала $J(v)$ при фиксированных f и v.

Покажем, что исходная задача нахождения классического (т.е. дважды непрерывно дифференцируемого) решения дифференциального уравнения может быть заменена на задачу отыскания минимума функционала $J(v)$.

Для дальнейшего анализа нам потребуется неравенство типа Фридрихса

Утверждение 1. *Пусть $u \in H$, тогда справедливо*

$$\|u(x)\|^2_{L_2(0,1)} \le \|u'(x)\|^2_{L_2(0,1)}\,, \quad m.e. \quad \int\limits_0^1 u^2(x)dx \le \int\limits_0^1 (u'(x))^2 dx\,.$$

Доказательство. Из равенств

$$u(x_0) = \int\limits_0^{x_0} u'(x)dx, \quad u(0) = 0$$

на основании неравенства Коши – Буняковского следует, что

$$|u(x_0)|^2 \le \int\limits_0^{x_0} 1^2 dx \int\limits_0^{x_0} |u'(x)|^2 dx \le \int\limits_0^1 |u'(x)|^2 dx \quad (x_0 \in [0,1])\,.$$

Интегрируя по x_0 обе части неравенства, получаем

$$\int\limits_0^1 |u(x_0)|^2 dx_0 \le \int\limits_0^1 \int\limits_0^1 |u'(x)|^2 dx dx_0\,.$$

Отсюда следует искомое неравенство

$$\|u(x)\|^2_{L_2(0,1)} \le \|u'(x)\|^2_{L_2(0,1)}\,.$$

Доказательство закончено.

Из полученного неравенства следует важная оценка снизу

$$a(v,v) = \int\limits_0^1 [k(x)(v')^2 + p(x)v^2]dx \ge (k_{\min} + p_{\min}) \int\limits_0^1 v^2 dx > 0 \quad \forall 0 \ne v \in H,$$

где $k_{\min} \ge k_1$, $p_{\min} \ge 0$. Отсюда, в частности, имеем, что $a(v,v) = 0$ тогда и только тогда, когда $v = 0$.

Обратим внимание, что краевое условие второго рода не входит в определение H, но если $y(x)$ — точка минимума функционала $J(v)$, то соотношение $y'(1) = 0$ выполняется автоматически.

Утверждение 2. *Пусть достаточно гладкая функция y (например, $y \in C^{(2)}[0,1]$, что потребуется в доказательстве при интегрировании по частям) доставляет минимум функционалу $J(v)$ на пространстве H. Тогда справедливы равенства:*
1) $a(y,v) = (f,v) \quad \forall v \in H$;
2) $y'(1) = 0$.

Доказательство. Если y доставляет минимум функционалу $J(v)$ на пространстве H, то для произвольных величин — числа ε и элемента $v \in H$ имеем

$$J(y) \le J(y + \varepsilon v) = J(y) + 2\varepsilon[a(y,v) - (f,v)] + \varepsilon^2 a(v,v)\,.$$

Отсюда в силу произвольности знака ε следует

$$a(y,v)-(f,v)=0 \quad \forall v\in H\,.$$

Первое равенство получено. Из него, интегрируя по частям, имеем

$$0=\int\limits_0^1 (ky'v'+pyv-fv)dx=\int\limits_0^1 v[-(ky')'+py-f]dx+k(1)y'(1)v(1)\,.$$

Из этого равенства, в силу равенства интеграла нулю, произвольности $v(x)\in H$ и положительности $k(x)$, следует соотношение на правом конце отрезка, т.е. $y'(1)=0$. Например, для этого достаточно в качестве $v(x)$ использовать элементы последовательности x^n, учитывая, что в этом случае $v(1)=1$. Утверждение доказано.

Изучим вопрос о связи точки минимума функционала $J(v)$ на пространстве H и решения рассматриваемой краевой задачи.

Утверждение 3. *Пусть y — классическое решение краевой задачи, тогда оно доставляет единственный минимум функционалу $J(v)$ на пространстве H.*

Доказательство. Пусть $y(x)$ — классическое (т.е. дважды непрерывно дифференцируемое) решение дифференциального уравнения, тогда оно является элементом пространства H. Зафиксируем его и рассмотрим выражение $a(v-y,v-y)-a(y,y)$ как функционал от $v\in H$. Этот функционал имеет единственную точку минимума $v=y$, так как первое слагаемое неотрицательно и, в силу следствия к утверждению 1, обращается в нуль только тогда, когда аргумент равен нулю. При этом второе слагаемое от v не зависит. После раскрывания скобок получим

$$\begin{gathered}a(v-y,v-y)-a(y,y)=a(v,v)-2a(y,v)+a(y,y)-a(y,y)=\\=a(v,v)-2(f,v)\equiv J(v)\,.\end{gathered}$$

Равенство $a(y,v)=(f,v)$ является следствием дифференциальной постановки задачи. Утверждение доказано.

Из проведенных рассуждений следует вывод: если функция $y(x)$ удовлетворяет исходной дифференциальной задаче, то она доставляет единственный минимум нашему функционалу, так как пространство H шире, чем пространство дважды непрерывно дифференцируемых функций, другими словами, переходя к вариационной постановке, мы не теряем искомое решение дифференциальной задачи. Легко заметить, что обратное утверждение неверно: для этого достаточно рассмотреть случай разрывной функции $f(x)$.

Метод Ритца минимизации функционала $J(v)$. Следующим этапом в методе конечных элементов является поиск приближенного решения в виде

$$y^n(x)=\sum_{j=1}^n c_j\varphi_j(x)\in S^n\,,$$

где $\{\varphi_j(x)\}_{j=1}^n$ — фиксированный набор линейно независимых функций из H, c_j — коэффициенты, подлежащие определению. Часто конечномерное

подпространство S^n пространства H определяют как линейную оболочку набора функций $\{\varphi_j(x)\}_{j=1}^n$. Рассмотрим выражение

$$J(y^n) = a\left(\sum_{j=1}^n c_j\varphi_j, \sum_{j=1}^n c_j\varphi_j\right) - 2\left(f, \sum_{j=1}^n c_j\varphi_j\right) =$$

$$= \sum_{i,j=1}^n c_i c_j a(\varphi_i, \varphi_j) - 2\sum_{j=1}^n c_j(f, \varphi_j) \equiv (A\mathbf{c}, \mathbf{c}) - 2(\mathbf{b}, \mathbf{c})\,.$$

Здесь $A = A^T > 0$ — матрица Грама с элементами $a_{ij} = a(\varphi_i, \varphi_j)$, $\mathbf{b}$ — вектор проекций правой части f на базисную систему функций: $b_j = (f, \varphi_j)$, $(\mathbf{b}, \mathbf{c}) = \sum\limits_{j=1}^n b_j c_j$ — обычное скалярное произведение векторов. Выражение $J(y^n)$ является квадратичным относительно неизвестных коэффициентов c_j, поэтому нахождение y^n (минимизация $J(y^n)$ на подпространстве S^n) сводится к известной задаче — решению системы линейных уравнений

$$\frac{\partial J(y^n)}{\partial c_i} = 0, \;\; i = 1, \ldots, n \quad \text{или} \quad A\mathbf{c} = \mathbf{b}$$

с симметричной положительно определенной матрицей.

Кусочно полиномиальные базисные функции. Построение базисных функций состоит из двух этапов. Сначала на **одном** элементе строятся **функции формы** $\Phi_i(x), i = 1, 2, \ldots, k$, являющиеся локальными многочленами фиксированной степени $k-1$ и удовлетворяющие условиям $\Phi_i(x_j) = \delta_i^j$, где $x_j,\ j = 1, 2, \ldots, k$ — набор узлов элемента. Затем из функций форм "собираются"определенные на **всей** области **базисные (пробные) функции** $\varphi_j(x),\ j = 1, 2, \ldots, n$ формальным объединением всех функций форм, принимающих в данном узле значение 1.

Рассмотрим построение кусочно линейных базисных функций для нашего одномерного случая. Конечный элемент здесь — отрезок $[x_{j-1}, x_j]$ длины h, которому соответствует набор из двух узлов, являющихся концами отрезка. В локальных координатах, связанных с отрезком, эти узлы можно записать как $x_1 = 0$, $x_2 = h$. Тогда функции формы примут вид

$$\Phi_1(x) = \frac{h - x}{h}, \qquad \Phi_2(x) = \frac{x}{h},$$

а базисные функции для произвольного разбиения $0 = x_0 < x_1 < \cdots < x_n = 1$ отрезка $[0, 1]$ определим так:

$$\varphi_0(x) = \begin{cases} \dfrac{x_1 - x}{x_1 - x_0} & \text{при } x_0 \le x \le x_1\,, \\ 0 & \text{при } x_1 \le x \le x_n\,. \end{cases}$$

$$\varphi_n(x) = \begin{cases} 0 & \text{при } x_0 \le x \le x_{n-1}\,, \\ \dfrac{x - x_{n-1}}{x_n - x_{n-1}} & \text{при } x_{n-1} \le x \le x_n\,. \end{cases}$$

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}} & \text{при } x_{j-1} \le x \le x_j\,, \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j} & \text{при } x_j \le x \le x_{j+1}\,, \\ 0 & \text{при остальных } x\,. \end{cases}$$

для $j = 1, \ldots, n-1$. Отметим, что для нашей задачи $\varphi_0(x)$ не требуется, так как она не является функцией из H (в силу $\varphi_0(0) \neq 0$).

Построение набора кусочно квадратичных функций рекомендуется осуществить самостоятельно. Для определенности (и простоты!) следует считать, что кроме концов отрезка третьим узлом на каждом элементе $[x_{j-1}, x_j]$ является его середина.

Лекция 23

Оценка точности приближения кусочно линейными функциями. — Проекционная теорема в методе конечных элементов.

Оценка точности приближения кусочно линейными функциями

Пусть S^n — пространство непрерывных кусочно линейных на каждом элементе функций. Получим оценку точности для достаточно гладких функций из H при приближеннии элементами из S^n. Возьмем разбиение отрезка с постоянным шагом: $x_j = jh,\ 0 \le j \le n,\ h = 1/n$ и поставим в соответствие некоторой функции $y(x)$ ее линейный интерполянт, т.е. функцию

$$y_I(x) = \sum_{j=0}^{n} y(x_j)\varphi_j(x).$$

Здесь в качестве коэффициентов c_j взяты точные значения $y(x_j)$. Докажем

Утверждение. *Пусть*

$$\|y''\|^2 = \int_0^1 [y''(x)]^2 dx < \infty\,.$$

Тогда

$$\|y' - y_I'\| \le \frac{h}{\pi}\|y''\|\,, \quad \|y - y_I\| \le \left(\frac{h}{\pi}\right)^2 \|y''\|\,.$$

Доказательство. Рассмотрим какой–либо отрезок длины h, для простоты удобно взять — $[0, h]$. Построим на нем функцию

$$\Delta(x) = y(x) - y_I(x)\,.$$

По предположению о гладкости $y(x)$ функция $\Delta(x)$ обладает конечным интегралом

$$\int_0^h (\Delta'')^2 dx = \int_0^h (y'')^2 dx < \infty\,, \qquad \text{т.е.} \qquad \Delta''(x) \in L_2(0,h)\,.$$

Также выполнены равенства $\Delta(0) = \Delta(h) = 0$, поэтому справедливо формальное представление $\Delta(x)$ в виде ряда Фурье

$$\Delta(x) = \sum_{l=1}^{\infty} d_l \sin\frac{\pi l x}{h}\,.$$

Непосредственные вычисления дают

$$\int_0^h [\Delta'(x)]^2 dx = \frac{h}{2}\sum_{l=1}^{\infty}\left(\frac{\pi l}{h}\right)^2 d_l^2,\ \int_0^h [\Delta''(x)]^2 dx = \frac{h}{2}\sum_{l=1}^{\infty}\left(\frac{\pi l}{h}\right)^4 d_l^2 < \infty.$$

Заметим, что из сходимости ряда для второй производной следует как сходимость ряда для первой производной, так и для самой функции Δ.

Так как $l \geq 1$, то справедливо неравенство

$$\left(\frac{\pi l}{h}\right)^2 d_l^2 \leq \left(\frac{h}{\pi}\right)^2 \left(\frac{\pi l}{h}\right)^4 d_l^2 \,,$$

поэтому суммируя по l, получаем

$$\int_0^h [\Delta'(x)]^2 dx \leq \left(\frac{h}{\pi}\right)^2 \int_0^h [\Delta''(x)]^2 dx = \left(\frac{h}{\pi}\right)^2 \int_0^h [y'']^2 dx \,.$$

Полученное неравенство справедливо на **каждом** подотрезке, потому суммирование по всем j (т.е. подотрезкам длины h) дает

$$\|y' - y_I'\|^2 \leq \left(\frac{h}{\pi}\right)^2 \|y''\|^2 \,.$$

Аналогичным образом получим

$$\int_0^h [\Delta(x)]^2 dx = \frac{h}{2} \sum_{l=1}^{\infty} d_l^2 \leq \left(\frac{h}{\pi}\right)^4 \int_0^h [y'']^2 dx, \text{т.е. } \|y - y_I\| \leq \left(\frac{h}{\pi}\right)^2 \|y''\|.$$

Утверждение доказано.

Из доказательства легко понять, что при построении базиса S^n с помощью полиномов более высокой степени можно получить оценки близости более высокого порядка по h, однако это потребует соответствующей гладкости приближаемой функции.

Проекционная теорема МКЭ

Следующее утверждение имеет принципиально важное значение в теории метода конечных элементов. По сути, оно означает, что для сходимости приближенного решения к точному, достаточно "хороших"свойств только исходной дифференциальной задачи и правильно выбранного подпространства S^n.

Теорема. *Пусть y — точка минимума функционала $J(v) = a(v,v) - 2(f,v)$ на пространстве H, S^n — конечномерное подпространство H. Тогда y^n — точка минимума функционала $J(v)$ на S^n — существует, единственна и обладает следующими свойствами:*
1) функция y^n удовлетворяет условию

$$a(y^n, v^n) = (f, v^n) \quad \forall v^n \in S^n \,.$$

2) функция y^n есть проекция y на S^n по отношению к энергетическому скалярному произведению $a(u,v)$, или, другими словами, ошибка $y - y^n$ ортогональна S^n :

$$a(y - y^n, v^n) = 0 \quad \forall v^n \in S^n \,.$$

3) минимум $J(v^n)$ и минимум $a(y - v^n, y - v^n)$, где v^n пробегает подпространство S^n, достигаются на одной и той же функции y^n, так что

$$a(y - y^n, y - y^n) = \min_{v^n \in S^n} a(y - v^n, y - v^n) \,.$$

Доказательство.
1) Повторим уже известный нам прием. Если y^n минимизирует $J(v)$ на S^n, то для произвольных величин — числа ε и элемента $v^n \in S^n$ имеем

$$J(y^n) \le J(y^n + \varepsilon\, v^n) = J(y^n) + 2\varepsilon[a(y^n, v^n) - (f, v^n)] + \varepsilon^2 a(v^n, v^n)\,.$$

Отсюда получаем

$$0 \le 2\varepsilon[a(y^n, v^n) - (f, v^n)] + \varepsilon^2 a(v^n, v^n)\,.$$

Так как ε может иметь любой знак, а второе слагаемое строго положительно, то

$$a(y^n, v^n) = (f, v^n)\,.$$

Следует отметить, что при доказательстве этого свойства не использовалась единственность точки минимума функционала $J(v)$ на S^n.
2) Второе утверждение следует из первого. Напомним, что в предыдущей лекции (см. утверждение 2) было показано, что

$$a(y, v) = (f, v) \quad v \in H\,.$$

Вычитая первое из полученных равенств из второго, так как $v^n \in S^n \subset H$, получим

$$a(y - y^n, v^n) = 0 \quad \forall v^n \in S^n\,.$$

3) Рассмотрим следующее выражение для произвольного v^n

$$a(y - y^n - v^n, y - y^n - v^n) = a(y - y^n, y - y^n) - 2a(y - y^n, v^n) + a(v^n, v^n).$$

В силу предыдущего утверждения, второе слагаемое равно нулю, а третье неотрицательно, поэтому имеем

$$a(y - y^n, y - y^n) \le a(y - y^n - v^n, y - y^n - v^n) \quad \forall v^n \in S^n\,.$$

Это неравенство обращается в равенство только при $a(v^n, v^n) = 0$, т.е. при $v^n = 0$, поэтому

$$a(y - y^n, y - y^n) = \min_{v^n \in S^n} a(y - v^n, y - v^n)\,.$$

Существование и единственность $y^n \in S^n$ следует из конечномерности S^n, где

$$y^n(x) = \sum_{j=1}^{n} c_j \varphi_j(x) \in S^n.$$

Действительно, матрица Грама $A = A^T > 0$ в системе линейных уравнений $A\,\mathbf{c} = \mathbf{b}$ для определения коэффициентов c_j невырождена для произвольного конечного набора линейно независимых функций $\varphi_j(x)$, поэтому решение такой системы определено однозначно.

Теорема доказана.

Основной смысл теоремы заключается в том, что аппроксимация решения y элементами из S^n и устойчивость непрерывной задачи автоматически

обеспечивают сходимость в энергетической норме ($a(v,v) = \|v\|_*^2$). Это замечательное свойство позволяет переложить громоздкую техническую работу по вычислению коэффициентов схемы "на плечи"компьютера, т.е. порождает возможность высокой технологичности процесса построения схем МКЭ.

Применим выведенные выше оценки аппроксимации к третьему свойству из проекционной теоремы при $v^n = y_I \in S^n$, в результате получим сходимость в энергетической норме:

$$a(y - y^n, y - y^n) \le a(y - y_I, y - y_I) \le$$

$$\le k_{\max} \left(\frac{h}{\pi}\right)^2 \|y''\|^2 + p_{\max} \left(\frac{h}{\pi}\right)^4 \|y''\|^2 \le C\, h^2 \|y''\|^2 .$$

В последнем неравенстве постоянная C зависит от ограничений сверху на коэффициенты $k(x)$ и $p(x)$.

Напомним, что в приведенной оценке

$$y(x) = \arg\min_{v \in H} J(v)\,, \quad y^n(x) = \arg \min_{v^n \in S^n} J(v^n).$$

Заметим, что при использовании базисных функций $s-$го порядка величина в правой части будет меняться как $O(h^{2s})$, но от решения $y(x)$ потребуется большая гладкость.

Подведем итог. Для взятой дифференциальной задачи сформулировали вариационную постановку и минимизировали функционал в подпространстве кусочно полиномиальных функций. В результате получили приближение y^n к y, зависящее только от того, насколько хорошо достаточно гладкий произвольный элемент из H приближается линейной комбинацией функций из S^n. При этом для сходимости y^n к y оказалось достаточно только устойчивости исходной дифференциальной задачи.

Уточним про устойчивость дифференциальной задачи. В утверждении 2 предыдущей лекции было доказано, что $a(y,v) = (f,v) \quad \forall v \in H$. Полагая $v = y$ и используя оценку снизу

$$a(v,v) \ge (k_{\min} + p_{\min})\|v\|^2 \quad \forall 0 \ne v \in H,$$

приходим к искомой априорной оценке решения, означающей устойчивость,

$$\|y\| \le \frac{\|f\|}{k_{\min} + p_{\min}}\,.$$

Если дополнительно имеется оценка $\|y''\| \le C_1\|f\|$, то из проекционной теоремы следует неравенство

$$\|y - y^n\| \le \tilde{C} h^2 \|y''\|.$$

Лекция 24

Система уравнений в методе конечных элементов. — Решение краевой задачи методом Фурье (непрерывный случай).

Система уравнений в методе конечных элементов

Рассмотрим для простоты применявшееся выше разбиение на конечные элементы равной длины: $h = 1/n$, $x_j = jh$, $j = 0, 1, \ldots, n$ и выпишем явный вид матричных элементов в методе Ритца для случая кусочно – линейных базисных функций. Воспользуемся первым свойством из проекционной теоремы

$$a(y^n, v^n) = (f, v^n) \quad \forall v^n \in S^n \,.$$

Напомним, что

$$y^n(x) = \sum_{j=1}^{n} c_j \varphi_j(x) \,,$$

и будем перебирать в качестве v^n последовательно все базисные функции $\varphi_i(x)$, $i = 1, 2, \ldots, n$. В результате для матричных элементов системы $A\,\mathbf{c} = \mathbf{b}$ получим

$$a_{ij} = a(\varphi_i(x), \varphi_j(x)) = \int\limits_0^1 \left[k(x) \frac{d\varphi_i}{dx} \frac{d\varphi_j}{dx} + p(x)\, \varphi_i \varphi_j \right] dx \,.$$

Традиционно употребляется терминология, что матрица A состоит из двух частей: матрицы жесткости K и матрицы масс M так, что $A = K + M$. При этом матрицу K формирует первое слагаемое в правой части формулы, а матрицу M — второе.

Для постоянных коэффициентов

$$k(x) \equiv k > 0 \,, \quad p(x) \equiv p \geq 0$$

интегралы несложно вычислить в явном виде. Так как $\varphi_j \equiv 0$ вне отрезка $[x_{j-1}, x_{j+1}]$ и

$$\frac{d\varphi_j}{dx} = \begin{cases} \frac{1}{h} & \text{при } x_{j-1} < x < x_j \,, \\ -\frac{1}{h} & \text{при } x_j < x < x_{j+1} \,, \\ 0 & \text{при } x \notin [x_{j-1}, x_{j+1}] \,, \end{cases}$$

то интегрирование носит локальный характер. Для этого достаточно использовать выражения для функций форм в координатах, связанных с элементом $[x_{j-1}, x_j]$. Если x_j — внутренний узел ($j = 1, 2, \ldots, n-1$), то матричные элементы принимают вид

$$a_{ij} = \begin{cases} -\frac{k}{h} + \frac{ph}{6} & \text{при } i = j - 1 \,, \\ \frac{2k}{h} + \frac{2ph}{3} & \text{при } i = j \,, \\ -\frac{k}{h} + \frac{ph}{6} & \text{при } i = j + 1 \,, \\ 0 & \text{в остальных случаях} \quad . \end{cases}$$

С компонентами вектора правой части $\mathbf{b}$ дело обстоит несколько иначе. Вспомним формулу

$$b_j = \int\limits_0^1 f\, \varphi_j dx = \int\limits_{x_{j-1}}^{x_{j+1}} f\, \varphi_j dx \,.$$

Как правило, в явном виде получить значение интеграла затруднительно, поэтому на практике поступают так. Сначала заменяют функцию $f(x)$ ее кусочно – линейным интерполянтом

$$f_I(x) = \sum_{j=0}^{n} f(x_j)\varphi_j(x)\,,$$

а затем явно вычисляют скалярные произведения (φ_i, φ_j). В результате получается формула

$$\hat{b}_j = \int\limits_{x_{j-1}}^{x_{j+1}} f_I(x)\varphi_j dx = h\,\frac{f(x_{j+1}) + 4\,f(x_j) + f(x_{j-1})}{6}\,.$$

Несложно убедиться в справедливости оценки $|b_j - \hat{b}_j| = O(h^3)$. Поэтому замена точных значений b_j на приближенные $\hat{b}_j$ вполне допустима, в силу следствия проекционной теоремы о скорости сходимости $y^n(x)$ к $y(x)$.

Предлагается получить самостоятельно оценку погрешности для более простого приближения

$$b_j \approx f(x_j) \int\limits_{x_{j-1}}^{x_{j+1}} \varphi_j dx = h\,f_j,$$

а также сделать вывод о его применимости в расчетах.

Окончательный результат можно записать в виде $(0 < j < n)$:

$$-\,k\,\frac{c_{j+1} - 2\,c_j + c_{j-1}}{h} + p\,h\,\frac{c_{j+1} + 4\,c_j + c_{j-1}}{6} = h\,\frac{f(x_{j+1}) + 4\,f(x_j) + f(x_{j-1})}{6}\,.$$

Напомним, что $c_0 = 0$. Последнее уравнение будет несколько отличаться:

$$k\,\frac{c_n - c_{n-1}}{h} + \,p\,h\,\frac{2\,c_n + c_{n-1}}{6} = h\,\frac{2\,f(x_n) + f(x_{n-1})}{6}\,.$$

В результате система линейных уравнений, возникающая при решении методом конечных элементов рассматриваемой краевой задачи для дифференциального уравнения второго порядка, имеет трехдиагональную матрицу. Напомним, что ранее был рассмотрен метод прогонки для решения таких систем с затратами $O(n)$ арифметических операций. После проверки достаточных условий устойчивости и корректности метода прогонки можно сделать вывод о его применимости для решения системы $A\,\mathbf{c} = \hat{\mathbf{b}}$.

В случае систем линейных уравнений с трехдиагональными матрицами применяют не только метод прогонки. Например, при постоянных значениях элементов на каждой диагонали для матрицы системы можно решить полную задачу на собственные значения, а затем, используя собственные векторы и собственные значения, применить для решения системы $A\,\mathbf{y} = \mathbf{f}$ метод Фурье. Этот метод весьма удобен не только для нахождения решения, но и для изучения его свойств, например, устойчивости.

Решение краевой задачи методом Фурье (непрерывный случай)

Схему применения метода Фурье удобно вспомнить на примере краевой задачи для обыкновенного дифференциального уравнения второго порядка

$$-u'' = f(x)\,, \quad u(0) = u(t) = 0\,, \quad x \in [0, t].$$

Рассмотрим сначала вспомогательную дифференциальную задачу на собственные значения:

$$-v'' = \mu\, v, \quad v(0) = v(t) = 0\,,$$

Ее решение имеет вид

$$v^{(m)}(x) = \sqrt{\frac{2}{t}} \sin\Big(\frac{\pi m x}{t}\Big)\,, \quad \mu^{(m)} = \frac{\pi^2 m^2}{t^2}, \quad m = 1, 2, \ldots$$

Отметим, что собственные функции $v^{(m)}(x)$ образуют полную, ортогональную в метрике пространства $L_2[0, t]$ систему, т.е.

$$(v^{(m)}, v^{(k)}) \equiv \int\limits_0^t v^{(m)}(x) v^{(k)}\, dx = \delta_m^k\,.$$

Представим искомое решение $u(x)$ в виде ряда с неизвестными коэффициентами d_m:

$$u(x) = \sum_{m=1}^{\infty} d_m v^{(m)}(x)\,,$$

а заданную правую часть $f(x)$ в виде ряда с известными коэффициентами ψ_m:

$$f(x) = \sum_{m=1}^{\infty} \psi_m v^{(m)}(x)\,, \quad \psi_m = (f, v^{(m)})\,.$$

После подстановки рядов для решения и правой части в исходное уравнение, учитывая, что собственным функциям $v^{(m)}(x)$ соответствуют собственные значения $\mu^{(m)}$, получим

$$\sum_{m=1}^{\infty} d_m \mu^{(m)} v^{(m)}(x) = \sum_{m=1}^{\infty} \psi_m v^{(m)}(x)\,.$$

Так как собственные функции линейно независимы, то коэффициенты при каждой из них в левой и правой частях уравнения совпадают, отсюда получаем неизвестные величины $d_m = \psi_m / \mu^{(m)}$, $m = 1, 2, \ldots$. Подставляя их в формулу для решения, будем иметь

$$u(x) = \sum_{m=1}^{\infty} \frac{(v^{(m)}, f)}{\mu^{(m)}} v^{(m)}(x)\,.$$

Напомним, что часто вычисление коэффициентов разложения заданной функции $f(x)$ называют прямым преобразованием Фурье, а вычисление функции $u(x)$ по ее коэффициентам — обратным преобразованием Фурье.

Лекция 25

Решение модельной задачи методом Фурье. — Метод стрельбы для решения трехдиагональных систем.

Решение модельной задачи методом Фурье

Рассмотрим применение метода Фурье к модельной задаче с трехдиагональной матрицей, имеющей постоянные коэффициенты, $\Lambda y_h = \varphi_h$ при $h = t/N$, т.е.

$$-\frac{y_{i+1} - 2\,y_i + y_{i-1}}{h^2} = \varphi_i, \quad 1 \le i \le N-1, \quad y_0 = y_N = 0.$$

Как и в непрерывном случае, изучим сначала вспомогательную задачу на собственные значения:

$$-\frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} = \lambda v_i, \quad 1 \le i \le N-1,\ Nh = t,$$

$$v_0 = v_N = 0\,.$$

Ее можно записать в матричной форме относительно вектора $v_h = (v_1, v_2, \ldots, v_{N-1})^T$:

$$\Lambda v_h = \lambda v_h, \qquad \Lambda = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \ldots & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix}.$$

Найдем собственные векторы, с помощью проекции собственных функций дифференциальной задачи на сетку $x_i = ih, i = 1, 2, \ldots, N-1$, и приведем их в покомпонентном виде:

$$v_i^{(m)} = C_m \sin\left(\pi m \frac{x_i}{t}\right) = C_m \sin\left(\frac{\pi m i h}{t}\right) = C_m \sin\left(\frac{\pi m i}{N}\right).$$

Тогда соответствующие собственные значения вычисляются явно

$$\lambda^{(m)} = \frac{4}{h^2}\sin^2\frac{\pi m h}{2t} = \frac{4}{h^2}\sin^2\frac{\pi m}{2N}, \quad t = Nh\,.$$

Отметим, что здесь задача имеет конечную размерность $(N-1)$, поэтому достаточно взять $m = 1, \ldots, N-1$. А куда деваются остальные? Для m, кратных N, соответствующие собственные векторы являются тождественно нулевыми, а для остальных значений m с точностью до знака совпадают с уже имеющимися в силу периодичности.

Выберем для векторов норму, отличающуюся от обычной евклидовой, лишь скалярным множителем $\|v_h\|_2^2 = (v_h, v_h) \equiv h\sum\limits_{i=1}^{N-1} v_i^2$. Легко проверить, что все коэффициенты C_m, обеспечивающие ортонормированность собственных векторов $v^{(m)}$, равны $\sqrt{2/t}$, т.е. справедливо равенство $(v^{(m)}, v^{(k)}) = \delta_m^k$. Отметим, что в конечномерном случае проверять ортогональность собственных векторов необязательно, так как матрица системы симметрична.

Приступим непосредственно к решению системы $\Lambda y_h = \varphi_h$. Разложим искомое решение y_h и заданную правую часть φ_h по собственному ортонормированному базису матрицы Λ — векторам $v_i^{(m)} = \sqrt{2/t}\sin\dfrac{\pi m i}{N}$, $1 \le i \le N-1$:

$$y_i = \sum_{m=1}^{N-1} v_i^{(m)} d_m, \quad \varphi_i = \sum_{m=1}^{N-1} v_i^{(m)} \psi_m \,,$$

где известные коэффициенты ψ_m определяются как скалярные произведения

$$\psi_m = h \sum_{i=1}^{N-1} \varphi_i v_i^{(m)} = (\varphi_h, v_h^{(m)}) \,.$$

После подстановки разложений в левую и правую части уравнения $\Lambda y_h = \varphi_h$ получим

$$\sum_{m=1}^{N-1} \lambda^{(m)} v_h^{(m)} d_m = \sum_{m=1}^{N-1} \psi_m v_h^{(m)} \,,$$

где $\lambda^{(m)} = \dfrac{4}{h^2}\sin^2\dfrac{\pi m}{2N}$ — известные собственные значения, соответствующие векторам $v_h^{(m)}$. В силу ортонормированности векторов $v_h^{(m)}$, получаем уравнения для коэффициентов

$$\lambda^{(m)} d_m = \psi_m, \quad m = 1, 2, \ldots, N-1 \,.$$

Отсюда следует искомое выражение

$$y_i = \sum_{m=1}^{N-1} v_i^{(m)} \frac{\psi_m}{\lambda^{(m)}} \,.$$

Таким образом, вся процедура нахождения решения методом Фурье состоит из двух этапов: сначала вычисляются коэффициенты ψ_m (прямое преобразование Фурье), а затем находим решения, суммируя те же векторы $v_h^{(m)}$ с новыми коэффициентами $\dfrac{\psi_m}{\lambda^{(m)}}$ (обратное преобразование Фурье).

Оценим асимптотику объема вычислительной работы по сравнению с методом прогонки. Напомним, что там требовалось $O(N)$ арифметических операций. Здесь же вычисление **всех** коэффициентов ψ_m требует $O(N^2)$ операций, затем $O(N)$ делений для определения новых коэффициентов и еще $O(N^2)$ действий для осуществления обратного преобразования Фурье. Окончательно по порядку имеем — $O(N^2)$. Если для вычисления тригонометрических сумм использовать быстрое дискретное преобразование Фурье, то можно улучшить асимптотику метода до величины $O(N \log_2 N)$. Это все равно хуже, чем асимптотика метода прогонки, однако применение метода прогонки принципиально ограничено одномерными задачами. Кроме того, метод Фурье может быть полезен и для других целей, например, для исследования устойчивости разностной задачи.

Исследование устойчивости модельной задачи методом Фурье

Обсудим устойчивость задачи $\Lambda y_h = \varphi_h$. В силу невырожденности матрицы системы, имеем $y_h = \Lambda^{-1}\varphi_h$, откуда следует неравенство $\|y_h\| \le \|\Lambda^{-1}\|\|\varphi_h\|$. Следуя определению устойчивости для линейных задач, достаточно показать справедливость неравенства $\|\Lambda^{-1}\| \le C$ с постоянной C, не зависящей от h.

Для выбранной выше векторной нормы, отличающейся от евклидовой лишь постоянным множителем, определим подчиненную матричную норму по обычной формуле $\|A\| = \sup\limits_{x\neq 0} \frac{\|Ax\|}{\|x\|}$. Это дает $\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$, а в случае симметричных матриц имеем $\|A\|_2 = \max|\lambda(A)|$. У нас матрица Λ не только симметрична, но и положительно определена. Поэтому

$$\|\Lambda\|_2 = \lambda_{\max}(\Lambda), \quad \|\Lambda^{-1}\|_2 = \frac{1}{\lambda_{\min}(\Lambda)}.$$

Отсюда имеем, что неравенство $\|\Lambda^{-1}\|_2 \le C$ эквивалентно оценке снизу $\lambda_{\min}(\Lambda) \ge C^{-1}$ для минимального собственного значения матрицы Λ.

Проанализируем свойства собственных значений матрицы Λ. Сначала отметим их монотонность:

$$\frac{4}{h^2}\sin^2\frac{\pi}{2N} = \lambda^{(1)} \le \lambda^{(2)} \le \cdots \le \lambda^{(N-1)} = \frac{4}{h^2}\cos^2\frac{\pi}{2N}\,,$$

а затем получим оценку снизу для минимального из них. Напомним следствия выпуклости вниз (вогнутости) функции $\sin x$ на отрезке $[0, \pi/2]$:

$$\sin x \ge \frac{2}{\pi}x \quad \forall x \in [0,\pi/2] \quad \text{и} \quad \sin x \ge \frac{2\sqrt{2}}{\pi}x \quad \forall x \in [0,\pi/4]\,.$$

Из второго неравенства следует оценка снизу $\lambda^{(1)} \ge 8/t^2$, которая незначительно отличается от младшего собственного значения дифференциальной задачи $\mu^{(1)} = \pi^2/t^2$.

Из полученной оценки снизу $\lambda^{(1)} \ge 8/t^2 \equiv C^{-1}$, следует окончательное неравенство $\|\Lambda^{-1}\|_2 \le \dfrac{t^2}{8}$, т.е. наша одномерная разностная задача устойчива. При этом для решения справедлива априорная оценка $\|y_h\| \le \dfrac{t^2}{8}\|\varphi_h\|$.

Почему такой метод исследования устойчивости называют методом Фурье? Дело в том, что полученная априорная оценка непосредственно следует из формулы для решения, построенной методом Фурье. Действительно, рассмотрим выражение

$$y_h = \sum_{m=1}^{N-1} v_h^{(m)} \frac{\psi_m}{\lambda^{(m)}}\,, \quad \text{где} \quad \psi_m = (\varphi_h, v_h^{(m)}),$$

и вычислим квадрат нормы решения y_h. Получим

$$(y_h, y_h) = \sum_{m=1}^{N-1} \frac{\psi_m^2}{\left(\lambda^{(m)}\right)^2} \le \frac{1}{\left(\lambda^{(1)}\right)^2} \sum_{m=1}^{N-1} \psi_m^2 = \frac{1}{\left(\lambda^{(1)}\right)^2}(\varphi_h, \varphi_h)\,.$$

Отсюда следует

$$\|y_h\| \le \frac{1}{\lambda^{(1)}}\|\varphi_h\| \le \frac{t^2}{8}\|\varphi_h\|\,,$$

т.е. полученная ранее априорная оценка.

Метод стрельбы для решения трехдиагональных систем

Пусть требуется найти решение системы уравнений:

$$\begin{array}{cc} c_0 y_0 - b_0 y_1 = f_0\,, & i = 0\,, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i\,, & 1 \le i \le N-1\,, \\ -a_N y_{N-1} + c_N y_N = f_N\,, & i = N\,, \end{array} \tag{2}$$

или в векторном виде

$$A\, y_h = f_h\,,$$

где $y_h = (y_0, y_1, \ldots, y_N)$ — вектор неизвестных, $f_h = (f_0, f_1, \ldots, f_N)$ — заданный вектор правых частей, A — квадратная $(N+1) \times (N+1)$ матрица:

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

Без ограничения общности можно считать, что все коэффициенты a_i и b_i отличны от нуля, так как в противном случае система уравнений распадается на подсистемы меньшей размерности, матрицы которых также имеют трехдиагональный вид. Напомним, что ранее мы уже рассматривали один метод решения подобных систем — метод прогонки.

Идея другого алгоритма — метода стрельбы — наиболее просто излагается в терминах дифференциальных уравнений. Фактически этот метод решает краевую задачу на основе решений задач Коши, т.е. задач с начальными данными. Причем начальные данные подбирают так, чтобы удовлетворить обоим краевым условиям.

Пусть требуется решить краевую задачу:

$$\begin{gathered} u'' - p(x)\, u = f(x)\,, \quad 0 < x < 1\,, \\ u(0) = d\,, \quad u(1) = e\,. \end{gathered}$$

Построим для нее подходящие задачи Коши.

Для первой задачи Коши выберем решение $v(x)$, удовлетворяющее уравнению

$$v'' - p(x)\, v = f(x)\,,$$

и начальным условиям $v(0) = d$ и $v'(0) = \varphi$, где φ — некоторая постоянная.

Для второй задачи Коши — аналогично:

$$w'' - p(x)\, w = f(x), \quad w(0) = d, \quad w'(0) = \psi.$$

Постоянная ψ выбирается так, чтобы при $x = 1$ получаемые решения $v(x)$ и $w(x)$ были различными: $v(1) \ne w(1)$.

Далее решение исходной краевой задачи $u(x)$ будем искать в виде:

$$u(x) = \delta\, v(x) + (1 - \delta)\, w(x),$$

где постоянная δ определяется из условия:

$$u(1) = \delta\, v(1) + (1 - \delta)\, w(1) = e \quad \to \quad \delta = \frac{e - w(1)}{v(1) - w(1)}\,.$$

Подходящий выбор констант φ и ψ обеспечивает существование единственного значения постоянной δ, и, следовательно, — решения $u(x)$.

Применим этот подход к решению системы (2). Будем искать решение y_i в виде

$$y_i = \delta\, u_i + (1-\delta)\, v_i\,,$$

где δ — параметр, подлежащий определению, а сеточные функции u_i и v_i, удовлетворяют уравнениям:

$$\begin{gathered} c_0u_0 - b_0u_1 = f_0\,, \qquad c_0v_0 - b_0v_1 = f_0\,, \\ -a_iu_{i-1} + c_iu_i - b_iu_{i+1} = f_i\,, \quad -a_iv_{i-1} + c_iv_i - b_iv_{i+1} = f_i\,, \\ \text{при } 1 \le i \le N-1\,. \end{gathered}$$

К этим системам для однозначного определения u_i и v_i необходимо добавить при $b_0 \ne 0$ начальные условия u_0 и v_0 ($u_0 \ne v_0$). Если же $b_0 = 0$, то добавляются значения u_1 и v_1 ($u_1 \ne v_1$). Стартуя с них, теперь можно последовательно определить $u_2, u_3, \ldots, u_N$ и $v_2, v_3, \ldots, v_N$.

Для завершения алгоритма осталось определить δ из уравнения

$$-a_N(\delta\, u_{N-1} + (1-\delta)\, v_{N-1}) + c_N(\delta\, u_N + (1-\delta)\, v_N) = f_N\,,$$

или

$$\delta = \frac{f_N + a_Nv_{N-1} - c_Nv_N}{a_N(v_{N-1} - u_{N-1}) + c_N(u_N - v_N)}\,.$$

Метод стрельбы служит хорошим дополнением к методу прогонки: области их корректности и устойчивости практически не пересекаются.

Пример. Рассмотрим случай системы (2) с постоянными коэффициентами:

$$y_{i-1} - \frac{26}{5}y_i + y_{i+1} = 0, \quad y_0 = 3, \quad y_N = 4\,.$$

Требуется найти решение разностной краевой задачи методом стрельбы и проанализировать его устойчивость.

Рассмотрим вспомогательные функции u_i и v_i. Из исходной системы имеем $u_0 = 3$. Так как $b_0 = 0$, положим $u_1 = \varphi$, и далее будем вычислять

$$u_{i+1} = \frac{26}{5}\, u_i - u_{i-1}\,, \quad i = 1, 2, \ldots, N-1\,.$$

Это решение можно представить в виде:

$$u_i = \frac{5\,\varphi - 3}{24}\, 5^i + \frac{75 - 5\,\varphi}{24}\, 5^{-i}\,, \quad i = 0, 1, \ldots, N\,.$$

Аналогично, полагая $v_0 = 3$ и $v_1 = \psi \ne \varphi$, приходим к формуле:

$$v_i = \frac{5\,\psi - 3}{24}\, 5^i + \frac{75 - 5\,\psi}{24}\, 5^{-i}\,, \quad i = 0, 1, \ldots, N\,.$$

Теперь, используя вычисленные u_N и v_N, определяем δ из уравнения

$$\delta\, u_N + (1-\delta)\, v_N = 4\,.$$

Так как $\delta = (4 - v_N)/(u_N - v_N)$, следует отметить, что знаменатель гарантированно отличен от нуля при $\psi \neq \varphi$:

$$u_N - v_N = \frac{5}{24}(\varphi - \psi)5^N - \frac{5}{24}(\varphi - \psi)5^{-N}, \quad N > 1.$$

Подставляя значение δ в выражение

$$y_i = \delta\, u_i + (1 - \delta)\, v_i\,,$$

получаем ответ:

$$y_i = 3\,\frac{5^{N-i} - 5^{i-N}}{5^N - 5^{-N}} + 4\,\frac{5^i - 5^{-i}}{5^N - 5^{-N}}\,.$$

Отметим,что в данном случае алгоритм является вычислительно неустойчивым. Действительно, $\max\limits_i |u_i|$ и $\max\limits_i |v_i|$ растут, как 5^N. Поэтому малым возмущениям значений $u_1 = \varphi$ и $v_1 = \psi$ будут соответствовать большие возмущения в u_N и v_N, и, соответственно, — в δ. Для исходной системы выполнены достаточные условия корректности и устойчивости метода прогонки, который и является здесь более предпочтительным для нахождения y_i.

Для закрепления материала по методу стрельбы рекомендуется самостоятельно найти решение разностной краевой задачи

$$y_{i-1} - y_i + y_{i+1} = 0, \quad y_0 = 3, \quad y_N = 4\,.$$

и проанализировать его устойчивость.

Лекция 26

Пример аппроксимации уравнения и краевых условий. — Определения аппроксимации и устойчивости.

Для изучения метода конечных разностей (МКР) рассмотрим ту же самую задачу, которую использовали для изучения метода конечных элементов:

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

Напомним, что $k(x), p(x), f(x)$ — достаточно гладкие функции, причем $0 < k_1 \le k(x) \le k_2, 0 \le p(x) \le p_2$. Эти ограничения гарантируют существование, единственность и непрерывную зависимость классического решения от входных данных задачи.

В основе метода конечных разностей лежат три ключевых понятия: **аппроксимация, устойчивость, сходимость**. Но прежде чем давать формальные определения рассмотрим полезные примеры.

Пример аппроксимации уравнения и краевых условий

Познакомимся с понятием аппроксимации для частного случая нашей задачи при $k(x) \equiv 1$:

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

Введем на отрезке $[0,1]$ сетку $x_i = ih, \quad i = 0, \ldots, N, \quad Nh = 1$, т.е. заменим исходную непрерывную область на ее дискретный аналог, и попытаемся построить некоторое выражение (формулу), значение которого будет похоже на результат двухкратного дифференцирования достаточно гладкой функции $y(x)$ в некоторой фиксированной точке $x = x_i$.

Полезное соображение. Если некоторое приближение к $y'(x)$ можно получить, используя всего две соседние точки x и $x + h$

$$y'(x) \approx \frac{y(x+h) - y(x)}{h}\,,$$

то для приближения второй производной, скорее всего, потребуется больше точек. Возьмем три, которые расположены симметрично: x и $x \pm h$, и попробуем, подобрав коэффициенты a, b, c, получить приближение

$$y''(x) \approx ay(x-h) + by(x) + cy(x+h)\,.$$

Такой способ обычно называют **методом неопределенных коэффициентов**.

Выпишем разложения по формуле Тейлора для достаточно гладкой функции $y(x)$:

$$y(x \pm h) = y(x) \pm h\,y'(x) + \frac{h^2}{2!}y''(x) \pm \frac{h^3}{3!}y'''(x) + \frac{h^4}{4!}y^{(4)}(x_\pm).$$

Эти представления справедливы в предположении, что $y \in C^{(4)}([0,1])$.

Подставим полученные выражения в формулу с неопределенными пока коэффициентами a, b, c и сгруппируем множители при одинаковых производных функции y:

$$y(x)(a+b+c) + hy'(x)(a-c) + \frac{h^2}{2!}y''(x)(a+c) + \\ + \frac{h^3}{3!}y'''(x)(a-c) + \frac{h^4}{4!}[ay^{(4)}(x_-) + cy^{(4)}(x_+)]\,.$$

Так как линейная комбинация должна быть похожа на $y''(x)$, выпишем соответствующую систему уравнений

$$a+b+c=0, \quad a-c=0, \quad \frac{h^2}{2}(a+c)=1.$$

Ее решение дает

$$a=c=\frac{1}{h^2}, \quad b=-\frac{2}{h^2}\,.$$

Далее, используя непрерывность четвертой производной функции $y(x)$ для слагаемых в квадратных скобках, получим

$$\frac{y(x-h)-2y(x)+y(x+h)}{h^2} = y''(x) + \frac{h^2}{12}y^{(4)}(\xi), \quad \xi \in [x-h, x+h]\,.$$

Полученное выражение дает основание для произвольного $x = x_i$, где $0 < i < N$, записать

$$-\frac{y(x_i-h)-2y(x_i)+y(x_i+h)}{h^2} + \frac{h^2}{12}y^{(4)}(\xi_i) + p(x_i)y(x_i) = f(x_i)\,.$$

Эти соотношения — точные, но значения ξ_i неизвестны. Поэтому слагаемое порядка $O(h^2)$ отбрасывают, считая, что приближения y_i ($y_i \approx y(x_i)$) **точно** удовлетворяют **усеченным** уравнениям

$$-\frac{y_{i-1}-2y_i+y_{i+1}}{h^2} + p_i y_i = f_i\,, \quad i = 1, 2, \ldots, N-1.$$

Здесь введены обозначения для заданных функций, взятых в точке x_i : $p_i = p(x_i),\ f_i = f(x_i)$.

Построив систему линейных уравнений относительно y_i, мы заменили одно дифференциальное уравнение второго порядка на систему из $(N-1)$ линейных алгебраических уравнений. Эта система не замкнута, так как содержит $N+1$ неизвестных. Для ее замыкания следует построить разностные аналоги краевых условий.

Условие в точке $x = 0$ аппроксимируется точно: $y_0 = 0$, а для условия в точке $x = 1$, т.е. $y'(1) = 0$, применим следующий прием. Возьмем значение $y(x_{N-1})$ (напомним, что $Nh = 1$, т.е. $x_N = 1$) и запишем для него формулу Тейлора:

$$y(x_{N-1}) = y(x_N) - hy'(x_N) + \frac{h^2}{2}y''(x_N) - \frac{h^3}{3!}y'''(\xi_N)\,.$$

Здесь мы использовали более слабое предположение о гладкости: $y \in C^{(3)}([0,1])$.

Отсюда получаем

$$y'(x_N) = \frac{y(x_N) - y(x_{N-1})}{h} + \frac{h}{2} y''(x_N) + O(h^2)\,.$$

Теперь выразим $y''(x_N)$ из исходного уравнения, имеем $y''(1) = p(1)y(1) - f(1)$. Подстановка этого выражения в формулу для $y'(x_N)$ приводит к искомой аппроксимации краевого условия $y'(1) = 0$ со вторым порядком

$$\frac{y_N - y_{N-1}}{h} + \frac{h}{2}(p_N y_N - f_N) = 0\,.$$

Таким образом, аппроксимация уравнения и краевых условий приводит к замкнутой системе линейных алгебраических уравнений

$$\begin{gathered} -\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + p_i y_i = f_i\,, \quad 0 < i < N\,, \\ y_0 = 0, \quad \frac{y_N - y_{N-1}}{h} + \frac{h}{2}(p_N y_N - f_N) = 0\,. \end{gathered}$$

В эту систему входит параметр h, определяющий ее размерность (так как $Nh = 1$), поэтому фактически мы имеем дело не с одним уравнением, а — с семейством уравнений, зависящим от параметра.

Отметим еще раз, что величины y_i в общем случае не обязаны совпадать с точными значениями функции $y(x)$ в узлах сетки, т.е. с $y(x_i)$.

Перейдем к абстрактным формулировкам базовых определений метода конечных разностей.

Пусть в области D с границей Γ задана дифференциальная задача

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

с граничным условием

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

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

Первым шагом применения метода конечных разностей является построение *сетки* — конечного множество точек (узлов) $\overline{D}_h = D_h \bigcup \Gamma_h$, принадлежащее области $\overline{D} = D \bigcup \Gamma$. Как правило, $\Gamma_h \subset \Gamma$.

Сетка $\overline{D}_h$ характеризуется параметром h, в общем случае векторным, компоненты которого состоят из шагов сетки h_i по каждой переменной. Для изучения свойств сеточных задач вводится понятие *величины* шага сетки, в качестве которой принимается какая-либо норма вектора h, например, $||h||_\infty = \max\limits_{1 \le i \le n} h_i$, где n — число независимых переменных в дифференциальной задаче. В приводимых ниже оценках под h понимается величина шага сетки.

Если $X \subset Y$ и функция v определена на множестве Y, то ее *следом* на множестве X называется функция, определенная на X и совпадающая там с v. Если функция v определена на некотором множестве Y, содержащем Y_h, то ее след на Y_h будем обозначать $(v)_h$. Часто пространства F_h, Φ_h и U_h определяют как пространства следов функций из F, Φ и U (или следов

гладких функций из всюду плотных в этих пространствах подмножеств) на D_h, Γ_h и $\overline{D}_h$ соответственно. При этом задают *согласованные* нормы пространств. Так, норма пространства U_h согласована с нормой U, если для достаточно гладких функций $u \in U$ выполняется соотношение

$$\lim_{h \to 0} \| (u)_h \|_{U_h} = \| u \|_U .$$

Далее для построения разностной схемы все производные, входящие в уравнение и краевые условия, заменяют конечными разностями, или другими словами, *разностными аппроксимациями*. При записи этих аппроксимаций в некотором внутреннем узле сетки берут одно и то же количество соседних узлов, образующих строго определенную конфигурацию, называемую *шаблоном*. В результате дифференциальные операторы L и l заменяют разностными L_h и l_h .

Определим *разностную схему* как семейство сеточных 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 разностной схемы u_h, называемое *разностным*, принимается в качестве приближенного решения дифференциальной задачи (1),(2).

Определения аппроксимации и устойчивости

Аппроксимация. Оператор L_h из (3) *локально аппроксимирует* в точке x_i сетки D_h 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

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

Число p при этом называют *порядком аппроксимации*. Аналогично определяют порядок локальной аппроксимации для оператора l_h.

Говорят, что разностная схема (3), (4) *аппроксимирует* дифференциальную задачу (1), (2) с порядком аппроксимации $p = \min(p_1, p_2)$, если для любых достаточно гладких функций $u \in U$, $f \in F$ и $\varphi \in \Phi$ существуют такие положительные постоянные h_0, c_1, p_1, c_2 и p_2, не зависящие от h, что при всех $h \le h_0$ справедливы неравенства

$$\|L_h(u)_h - (Lu)_h\|_{F_h} + \|(f)_h - f_h\|_{F_h} \le c_1 h^{p_1},$$

$$\|l_h(u)_h - (lu)_h\|_{\Phi_h} + \|(\varphi)_h - \varphi_h\|_{\Phi_h} \le c_2 h^{p_2}.$$

Выражения, стоящие под знаками норм, называют *погрешностями аппроксимации*.

Также используется понятие аппроксимации на решении, позволяющее строить схемы более высокого порядка точности на фиксированном шаблоне. 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 положительные постоянные h_0, c_1, p_1, c_2 и p_2, не зависящие от h, что при всех $h \le h_0$ справедливы неравенства

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

Предполагается, что при этом выполнены условия нормировки сеточных функций f_h и φ_h: $\lim\limits_{h\to 0} \| f_h \|_{F_h} = \| f \|_F$, $\lim\limits_{h\to 0} \| \varphi_h \|_{\Phi_h} = \| \varphi \|_\Phi$.

Порядки аппроксимаций обычно оценивают с помощью разложения в ряды Тейлора. Порядок аппроксимации разностной схемы может быть разным по разным переменным. Если погрешность аппроксимации стремится к нулю при любом законе стремления шагов по различным переменным к нулю, то такую аппроксимацию называют *безусловной*. Если же погрешность аппроксимации стремится к нулю при одних законах убывания шагов и не стремится к нулю при других, то аппроксимацию называют *условной*.

Для того, чтобы усвоить разницу между определениями аппроксимации: „локальной“, „общей“ и „на решении“, рекомендуется самостоятельно проанализировать пример аппроксимации уравнения и краевых условий. Наибольший интерес представляет аппроксимация краевого условия при $x = 1$: в соответствии с общим определением аппроксимации порядок равен единице, а в соответствии с определением „на решении“ — двум.

Устойчивость. Разностную схему (3), (4) называют *устойчивой*, если ее решение существует, единственно и непрерывно зависит от входных данных f_h, φ_h, причем эта зависимость равномерна относительно величины шага сетки h. Уточним это определение, ограничившись для простоты только линейным случаем. Рассмотрим решения $u_h^{(i)}$ разностной схемы (3), (4) при соответствующих правых частях $f_h^{(i)}, \varphi_h^{(i)}$, $i = 1, 2$. Линейная разностная схема устойчива, если существуют такие положительные постоянные h_0, C_1 и C_2, не зависящие от h, что для всех $h \le h_0$ справедливо неравенство

$$\left\| u_h^{(1)} - u_h^{(2)} \right\|_{U_h} \le C_1 \left\| f_h^{(1)} - f_h^{(2)} \right\|_{F_h} + C_2 \left\| \varphi_h^{(1)} - \varphi_h^{(2)} \right\|_{\Phi_h}.$$

Это означает, что для решения (3), (4) в линейном случае имеет место *априорная оценка*

$$\|u_h\|_{U_h} \le C_1 \|f_h\|_{F_h} + C_2 \|\varphi_h\|_{\Phi_h},$$

из которой необходимо следует существование и единственность решения при любой правой части, т.е. векторах f_h и φ_h.

Устойчивость называют *безусловной*, если указанные неравенства выполняются при произвольном соотношении шагов по различным переменным. Если же для выполнения неравенств шаги должны удовлетворять дополнительным соотношениям, то устойчивость называют *условной*.

Лекция 27

Определение сходимости. Теорема А.Ф. Филиппова. — Интегро-интерполяционный метод.

Определение сходимости. Теорема А.Ф. Филиппова

Сходимость. Решение u_h разностной схемы (3), (4) *сходится* к решению u дифференциальной задачи (1), (2), если

$$\|(u)_h - u_h\|_{U_h} \to 0 \text{ при } h \to 0.$$

Если существуют такие положительные постоянные h_0, c и p, не зависящие от h, что для всех $h \le h_0$ справедливо неравенство

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

то разностная схема сходится *с порядком* p; говорят, также, что разностное решение u_h имеет порядок точности p.

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

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

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

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

Доказательство. Сфомулируем разностную задачу

$$L_h\hat{u}_h = \hat{f}_h, \quad l_h\hat{u}_h = \hat{\varphi}_h,$$

решением которой является погрешность в узлах сетки, т.е. $\hat{u}_h = u_h - (u)_h$.

Операторы L и L_h линейные, поэтому

$$\hat{f}_h = L_h(u_h - (u)_h) = L_hu_h - L_h(u)_h =$$

$$= f_h - L_h(u)_h \pm (Lu)_h = [(Lu)_h - L_h(u)_h] + [f_h - (f)_h]\,.$$

Аналогично для краевых условий находим

$$\hat{\varphi}_h = l_h(u_h - (u)_h) = [(lu)_h - l_h(u)_h] + [\varphi_h - (\varphi)_h]\,.$$

Решение разностной задачи устойчиво, поэтому, по определению, для линейных задач имеем

$$\|\hat{u}_h\|_{U_h} \le C_1 \left\|\hat{f}_h\right\|_{F_h} + C_2\,\|\hat{\varphi}_h\|_{\Phi_h}\,.$$

Подставляя в неравенство выражения для правых частей $\hat{f}_h$ и $\hat{\varphi}_h$ и используя определение аппроксимации, получаем

$$\|u_h - (u)_h\|_{U_h} \le C_1[\|(Lu)_h - L_h(u)_h\|_{F_h} + \|f_h - (f)_h\|_{F_h}] +$$

$$+C_2[\|(lu)_h - l_h(u)_h\|_{\Phi_h} + \|\varphi_h - (\varphi)_h\|_{\Phi_h}] \le C_1 c_1 h^{p_1} + C_2 c_2 h^{p_2} \equiv c\, h^p\,,$$

где $p = \min\{p_1, p_2\}$. Это неравенство означает сходимость с порядком не ниже p. Теорема доказана.

Для многомерных задач порядок аппроксимации по разным переменным может быть неодинаковым, поэтому порядки сходимости по разным переменным также могут быть различными. Если аппроксимация и (или) устойчивость разностной схемы условные, то сходимость имеет место только при тех соотношениях между шагами сетки по разным переменным, при которых выполнены условия аппроксимации и (или) устойчивости. В классе задач с решениями конечной гладкости требование устойчивости является необходимым условием сходимости.

Интегро – интерполяционный метод

Опишем применение этого метода к построению разностной схемы на равномерной сетке $\overline{D}_h = \{x_i = i\,h,\ i = 0, \ldots, N;\ \ N\,h = 1\}$ для рассматриваемой модельной задачи

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

По-прежнему будем считать, что $k(x), p(x), f(x)$ — достаточно гладкие функции, причем $0 < k_1 \le k(x) \le k_2, 0 \le p(x) \le p_2$.

Обозначим $x_{i\pm1/2} = x_i \pm h/2$, $\omega(x) = k(x)y'(x)$ и проинтегрируем исходное уравнение в пределах от $x_{i-1/2}$ до $x_{i+1/2}$. Будем иметь точное *уравнение баланса*

$$\omega(x_{i+1/2}) - \omega(x_{i-1/2}) = \int\limits_{x_{i-1/2}}^{x_{i+1/2}} p(x)\,y(x)\,dx - \int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx\,.$$

Для замены интегралов в уравнении баланса воспользуемся формулой прямоугольников с узлом в середине отрезка —

$$\int\limits_a^b \psi(x)\,dx = (b-a)\,\psi\left(\frac{a+b}{2}\right) + O((b-a)^3).$$

Получим

$$\int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx = h\,f(x_i) + O(h^3), \qquad \int\limits_{x_{i-1/2}}^{x_{i+1/2}} p(x)y(x)\,dx = h\,y(x_i)p(x_i) + O(h^3).$$

Теперь уравнение баланса можно записать в виде

$$-\frac{\omega(x_{i+1/2}) - \omega(x_{i-1/2})}{h} + p_i y(x_i) = f_i + O(h^2)\,,$$

где $p_i = p(x_i)$, $f_i = f(x_i)$.

Выразим далее величины $\omega(x_{i\pm1/2})$ через значения функции $y(x)$ в точках сетки. Для этого проинтегрируем соотношение $y'(x) = \omega(x)/k(x)$ сначала на отрезке $[x_{i-1}, x_i]$. Получим

$$y(x_i) - y(x_{i-1}) = \int\limits_{x_{i-1}}^{x_i} \frac{\omega(x)}{k(x)}\,dx = h\,\frac{\omega(x_{i-1/2})}{k(x_{i-1/2})} + O(h^3).$$

Отсюда следует

$$\omega(x_{i-1/2}) = k(x_{i-1/2})\frac{y(x_i) - y(x_{i-1})}{h} + O(h^2).$$

Аналогично получается выражение для $\omega(x_{i+1/2})$. Формально для этого достаточно заменить индекс i на $i+1$.

Подстановка найденных выражений в уравнение баланса дает

$$-\frac{1}{h}\left(k(x_{i+1/2})\frac{y(x_i+h) - y(x_i)}{h} - k(x_{i-1/2})\frac{y(x_i) - y(x_{i-1})}{h}\right) + p_i y(x_i) = f_i + O(h).$$

Теперь определяя приближенные значения $y_i \approx y(x_i)$ как точные решения уравнений

$$-\frac{1}{h}\left(k_{i+1/2}\frac{y_{i+1} - y_i}{h} - k_{i-1/2}\frac{y_i - y_{i-1}}{h}\right) + p_i y_i = f_i, \quad i = 1, 2, \ldots, N-1,$$

получаем семейство разностных уравнений, зависящих от параметра h. Формальный порядок аппроксимации равен единице, так как соответствует отбрасыванию слагаемого $O(h)$.

Отметим, что разностные уравнения по своему построению определены для значений индексов $i = 1, 2, \ldots, N-1$, но содержат $N+1$ неизвестное (дополнительно сюда входят y_0 и y_N). Поэтому его необходимо дополнить недостающими соотношениями, которые следуют из краевых условий. Первое получается просто — $y_0 = 0$, а второе может быть получено также интегро – интерполяционным методом. Для этого проинтегрируем исходное уравнение на отрезке $[1 - h/2, 1]$ (т.е. на отрезке $[x_{N-1/2}, x_N]$):

$$\omega(x_N) - \omega(x_{N-1/2}) = \int\limits_{x_{N-1/2}}^{x_N} p(x)\, y(x)\, dx - \int\limits_{x_{N-1/2}}^{x_N} f(x)\, dx\,.$$

Полагая, как и выше,

$$\omega(x_{N-1/2}) = k(x_{N-1/2})\frac{y(x_N) - y(x_{N-1})}{h} + O(h^2),$$

и учитывая, что $\omega(x_N) = 0$, а также заменяя интегралы по формуле прямоугольников с центральным узлом

$$\int\limits_{1-h/2}^{1} f(x)\, dx = \frac{h}{2}\, f\left(1 - \frac{h}{4}\right) + O(h^3), \quad \int\limits_{1-h/2}^{1} p(x) y(x)\, dx = \frac{h}{2}\, p\left(1 - \frac{h}{4}\right) y\left(1 - \frac{h}{4}\right) + O(h^3),$$

получим

$$k(x_{N-1/2})\frac{y(x_N) - y(x_{N-1})}{h} + \frac{h}{2}\, p\left(1 - \frac{h}{4}\right) y\left(1 - \frac{h}{4}\right) = \frac{h}{2}\, f\left(1 - \frac{h}{4}\right) + O(h^2).$$

Так как погрешность аппроксимации имеет второй порядок относительно h, то мы его не испортим сдвинув аргументы у функций y, p, f на величину

$h/4$ (если при этих функциях имеется множитель порядка h). Теперь, опять переходя к приближениям y_i, будем иметь

$$k_{N-1/2}\frac{y_N - y_{N-1}}{h} + \frac{h}{2}\,p_N y_N = \frac{h}{2}\,f_N.$$

В результате система алгебраических уравнений получилась замкнутой, т.е. содержащей $(N+1)$ уравнение относительно $(N+1)$ неизвестного.

Определим порядок локальной аппроксимации в построенной схеме. Пусть имеется результат дифференцирования $(k(x)y')'\big|_{x=x_i}$, взятый в точке $x = x_i$. Рассмотрим следующее выражение

$$\frac{1}{h}\left[k\left(x_{i+1/2}\right)\frac{y(x_{i+1}) - y(x_i)}{h} - k\left(x_{i-1/2}\right)\frac{y(x_i) - y(x_{i-1})}{h}\right],$$

и проанализируем с помощью формулы Тейлора его предел при $h \to 0$. В предположении достаточной гладкости функций $k(x)$ и $y(x)$ будем иметь:

$$k\left(x_{i\pm1/2}\right) = k(x_i) \pm \frac{h}{2}k'(x_i) + \frac{1}{2}\left(\frac{h}{2}\right)^2 k''(x_i) + O(h^3)\,,$$

$$y(x_i \pm h) = y(x_i) \pm hy'(x_i) + \frac{h^2}{2}y''(x_i) \pm \frac{h^3}{6}y^{(3)}(x_i) + O(h^4)\,.$$

Подставим разложения в первое слагаемое в квадратных скобках

$$k\left(x_{i+1/2}\right)\frac{y(x_i+h) - y(x_i)}{h} = \left[k(x_i) + \frac{h}{2}k'(x_i) + \frac{h^2}{8}k''(x_i) + \right.$$

$$\left. + O(h^3)\right]\left[y'(x_i) + \frac{h}{2}y''(x_i) + \frac{h^2}{6}y^{(3)}(x_i) + O(h^3)\right] =$$

$$= k(x_i)y'(x_i) + \frac{h}{2}[k(x_i)y''(x_i) + k'(x_i)y'(x_i)] +$$

$$+h^2\left[\frac{1}{6}k(x_i)y^{(3)}(x_i) + \frac{1}{4}k'(x_i)y''(x_i) + \frac{1}{8}k''(x_i)y'(x_i)\right] + O(h^3)\,.$$

Аналогичным образом преобразуем оставшееся выражение

$$k\left(x_{i-1/2}\right)\frac{y(x_i) - y(x_{i-1})}{h} =$$

$$= k(x_i)y'(x_i) - \frac{h}{2}[k(x_i)y''(x_i) + k'(x_i)y'(x_i)] +$$

$$+h^2\left[\frac{1}{6}k(x_i)y^{(3)}(x_i) + \frac{1}{4}k'(x_i)y''(x_i) + \frac{1}{8}k''(x_i)y'(x_i)\right] + O(h^3)\,.$$

Теперь после вычитания и деления на h имеем

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

$$= k(x_i)y''(x_i) + k'(x_i)y'(x_i) + O(h^2) = (k(x)y')'\big|_{x=x_i} + O(h^2)\,.$$

Из полученного выражения следует близость исходного результата дифференцирования и его приближения (аппроксимации) в точке $x = x_i$ при $h \to 0$. Более того, из введенного ранее определения следует, что эту близость можно охарактеризовать количественно — величиной порядка аппроксимации (в рассмотренном случае порядок равен двум).

Для анализа порядка аппроксимации краевого условия $y'(1) = 0$ следует привести разностное уравнение к виду

$$\frac{y_N - y_{N-1}}{h} = \frac{h}{2}\frac{1}{k_{N-1/2}}(f_N - p_N y_N)\,.$$

Рекомендуется в качестве самостоятельного упражнения проанализировать на этом примере разницу в определениях „аппроксимации“ и „аппроксимации на решении“.

Лекция 28

Исследование устойчивости методом априорных оценок. — Метод конечных разностей для уравнения Пуассона.

Исследование устойчивости методом априорных оценок

Рассмотрим этот метод сначала на примере дифференциальной задачи

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

$$0 < k_{\min} \le k(x) \le k_{\max}, \quad 0 \le p_{\min} \le p(x) \le p_{\max}\,.$$

Он состоит трех этапов.
1. Интегральное тождество. Возьмем интеграл по отрезку $[0,1]$ от обеих частей уравнения, предварительно умножив его на $y(x)$:

$$-\int\limits_0^1 (k(x)y')'ydx + \int\limits_0^1 p(x)y^2dx = \int\limits_0^1 fydx\,.$$

В результате интегрирования по частям первого слагаемого в левой части получим

$$-\int\limits_0^1 (k(x)y')'ydx = -k(x)y'y\Big|_0^1 + \int\limits_0^1 k(x)(y')^2dx\,.$$

В силу краевых условий, значения функции $k(x)y'(x)y(x)$ на границах отрезка равны нулю, отсюда имеем интегральное тождество

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

2. Неравенство для квадратов функции и ее производной. Ранее была показана справедливость неравенства для функций $u(x)$ таких, что $u(0) = 0$:

$$\int\limits_0^1 u^2(x)dx \le \int\limits_0^1 (u'(x))^2dx \quad \text{или} \quad \|u(x)\|^2_{L_2(0,1)} \le \|u'(x)\|^2_{L_2(0,1)}\,.$$

3. Априорная оценка. Для левой части интегрального тождества получим оценку снизу (напомним, что $p_{\min} = \min\limits_x p(x) \ge 0$, $k_{\min} = \min\limits_x k(x) > 0$)

$$\int\limits_0^1 k(x)(y'(x))^2dx + \int\limits_0^1 p(x)y^2(x)dx \ge (k_{\min}+p_{\min})\int\limits_0^1 y^2(x)dx = (k_{\min}+p_{\min})\|y\|^2_{L_2(0,1)}\,,$$

а для правой — оценку сверху, используя неравенство Коши – Буняковского,

$$\int\limits_0^1 f(x)y(x)dx \le \|f(x)\|_{L_2(0,1)}\|y(x)\|_{L_2(0,1)}\,.$$

Отсюда после деления обеих частей на $(k_{\min}+p_{\min})\|y(x)\|_{L_2}$ имеем априорную оценку

$$\|y(x)\|_{L_2(0,1)} \le \frac{1}{k_{\min}+p_{\min}}\|f(x)\|_{L_2(0,1)}\,,$$

из которой следует единственность решения и непрерывная зависимость от входных данных, т.е. устойчивость.

Перейдем к изучению метода априорных оценок в **разностном** случае. Введем обозначения для сеточных функций (векторов) вида $u_h = (u_0, u_1, \dots u_N)^T$:

$$(u_h, v_h) = h\left(\sum_{i=1}^{N-1} u_i v_i + \frac{u_0 v_0 + u_N v_N}{2}\right), \|u_h\|^2 = (u_h, u_h)\,.$$

Для получения интегрального тождества использовалась формула интегрирования по частям

$$\int_a^b v\,du = -\int_a^b u\,dv + uv\Big|_a^b.$$

Ее дискретным аналогом является формула Абеля

$$\sum_{i=1}^{N-1}(u_{i+1}-u_i)v_{i+1} = -\sum_{i=1}^{N-1}(v_{i+1}-v_i)u_i - u_1v_1 + u_Nv_N\,.$$

Проверим ее. Левая часть равенства имеет вид

$$(u_2v_2 - u_1v_2) + (u_3v_3 - u_2v_3) + \ldots + (u_Nv_N - u_{N-1}v_N)\,,$$

а первая сумма в правой части —

$$-[(v_2u_1 - v_1u_1) + (v_3u_2 - v_2u_2) + \ldots + (v_Nu_{N-1} - v_{N-1}u_{N-1})]\,.$$

Теперь сравнивая выписанные выражения с учетом слагаемых $-u_1v_1 + u_Nv_N$, убеждаемся в справедливости формулы.

В непрерывном случае применялось неравенство для функций $u(x)$ таких, что $u(0) = 0$:

$$\int_0^1 u^2(x)dx \le \int_0^1 (u'(x))^2dx.$$

Для получения дискретного аналога этого неравенства представим $k-$ую компоненту вектора u_h в виде суммы разностей предыдущих компонент:

$$u_k = \sum_{i=1}^{k}(u_i - u_{i-1}) = (u_1 - u_0) + (u_2 - u_1) + \ldots + (u_k - u_{k-1}).$$

Это равенство справедливо при условии $u_0 = 0$. Отсюда имеем неравенство для $1 \le k \le N$

$$u_k^2 \le \sum_{i=1}^{k} 1^2 \sum_{i=1}^{k}(u_i - u_{i-1})^2 \le N\sum_{i=1}^{N}(u_i - u_{i-1})^2.$$

Его суммирование по k дает

$$\sum_{k=1}^{N} u_k^2 \le N^2 \sum_{i=1}^{N} (u_i - u_{i-1})^2, \quad \text{откуда следует} \quad \frac{1}{h}\sum_{i=1}^{N} (u_i - u_{i-1})^2 \ge (u_h, u_h). \tag{1}$$

Теперь мы готовы к исследованию устойчивости схемы, построенной интегро – интерполяционным методом:

$$-\frac{1}{h}\left(k_{i+1/2}\frac{y_{i+1}-y_i}{h} - k_{i-1/2}\frac{y_i - y_{i-1}}{h}\right) + p_i y_i = f_i, \quad 1 \le i \le N-1\,, \quad Nh = 1,$$

$$y_0 = 0, \quad k_{N-1/2}\frac{y_N - y_{N-1}}{h} = \frac{h}{2}(f_N - p_N y_N)\,.$$

Для получения сумматорного (т.е. дискретного аналога интегрального) тождества умножим $i-$е уравнение на $h\,y_i$ для $1 \le i \le N-1$ и все полученные выражения сложим. В результате будем иметь

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

Для преобразования первого слагаемого в левой части применим формулу Абеля, в которой положим $u_i = k_{i-1/2}(y_i - y_{i-1}), v_i = y_{i-1}$. Это дает с учетом первого краевого условия $y_0 = 0$

$$\frac{1}{h}\sum_{i=1}^{N-1} k_{i-1/2}(y_i - y_{i-1})^2 - k_{N-1/2}\frac{y_N - y_{N-1}}{h} y_{N-1} + h\sum_{i=1}^{N-1} p_i y_i^2 = h\sum_{i=1}^{N-1} f_i y_i\,.$$

Умножим второе краевое условие на y_N и прибавим результат к полученному выше равенству. После группировки слагаемых будем иметь сумматорное (аналог интегрального) тождество

$$\frac{1}{h}\sum_{i=1}^{N} k_{i-1/2}(y_i - y_{i-1})^2 + h\left(\sum_{i=1}^{N-1} p_i y_i^2 + p_N\frac{y_N^2}{2}\right) = h\left(\sum_{i=1}^{N-1} f_i y_i + \frac{f_N y_N}{2}\right).$$

Вспоминая определение скалярного произведения, перепишем сумматорное тождество в удобной форме

$$\frac{1}{h}\sum_{i=1}^{N} k_{i-1/2}(y_i - y_{i-1})^2 + ((py)_h, y_h) = (f_h, y_h)\,.$$

Оценим снизу первое слагаемое в сумматорном тождестве, применяя неравенство (1):

$$\frac{1}{h}\sum_{i=1}^{N} k_{i-1/2}(y_i - y_{i-1})^2 \ge k_{\min}\frac{1}{h}\sum_{i=1}^{N}(y_i - y_{i-1})^2 \ge k_{\min}(y_h, y_h)\,.$$

Используем в сумматорном тождестве полученное неравенство и учтем оценку снизу $p(x) \ge p_{\min} \ge 0$, в результате получим $(k_{\min} + p_{\min})(y_h, y_h) \le (f_h, y_h)$. Отсюда, в силу неравенства Коши – Буняковского для векторов,

имеем $(k_{\min} + p_{\min})\|y_h\|^2 \le \|f_h\|\|y_h\|$, что приводит к искомой априорной оценке, означающей устойчивость,

$$\|y_h\| \le \frac{\|f_h\|}{k_{\min} + p_{\min}}\,.$$

Метод конечных разностей для уравнения Пуассона

Обобщением обыкновенного дифференциального уравнения 2-го порядка $-y''(x) = f(x)$ является **уравнение Пуассона**. В случае двух независимых переменных оно имеет вид :

$$-\Delta\, u \equiv -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f(x,y)\,.$$

Для простоты изучим случай, когда область решения D является прямоугольником $D = \{(x,y) : 0 < x < l_1,\, 0 < y < l_2\}$ и на ее границе Γ заданы однородные краевые условия Дирихле (1-го рода)

$$-\Delta\, u = f, \quad u|_\Gamma = 0\,.$$

Рассмотрим решение сформулированной задачи методом конечных разностей. Проанализируем сначала **аппроксимацию**. Введем в $\overline{D} = D \bigcup \Gamma$ равномерную сетку $\overline{D}_h = D_h \bigcup \Gamma_h = \{(x_i, y_j)\}$ с шагами h_1 и h_2 :

$$x_i = ih_1,\, h_1 N_1 = l_1,\, 0 \le i \le N_1, \quad y_j = jh_2,\, h_2 N_2 = l_2,\, 0 \le j \le N_2.$$

Напомним локальную аппроксимацию второй производной по x в точке x_i

$$y''(x_i) = \frac{y(x_i + h) - 2y(x_i) + y(x_i - h)}{h^2} + O(h^2)$$

и применим ее к нашему случаю. Вводя обозначение $u_{i,j} \approx u(x_i, y_j)$, получим разностный аналог оператора Δ —

$$\left(\Delta^h u_h\right)_{i,j} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h_1^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h_2^2}.$$

Полностью разностная задача (вместе с краевыми условиями) в покомпонентной форме примет вид

$$\begin{array}{ll} -\left(\Delta^h u_h\right)_{i,j} = f_{i,j}, & 1 \le i \le N_1 - 1,\ 1 \le j \le N_2 - 1, \\ u_{0,j} = u_{N_1,j} = 0, & 0 \le j \le N_2, \\ u_{i,0} = u_{i,N_2} = 0, & 0 \le i \le N_1. \end{array}$$

Эта схема имеет второй порядок **аппроксимации**, так как

$$\left(\Delta^h (u)_h\right)_{i,j} = \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)\Bigg|_{(x=x_i, y=y_j)} + O(h_1^2 + h_2^2)\,,$$

и носит название „крест“, отражающее взаимное расположение неизвестных значений функции. Отметим, что разностная задача может быть записана в виде системы уравнений $A_h u_h = f_h$ с симметричной положительно определенной матрицей A_h. Рекомендуется проделать это самостоятельно.

Для анализа **устойчивости** применим наши знания о методе Фурье к задаче

$$A_h u_h = f_h\,.$$

Воспользуемся свойствами алгебраических задач на собственные значения $\Lambda_k v_h = \lambda v_h,\ k = 1, 2:$

$$(\Lambda_k v_h)_i = -\frac{v_{i-1} - 2v_i + v_{i+1}}{h_k^2}, \quad 1 \le i \le N_k - 1, \quad v_0 = v_{N_k} = 0, \quad N_k h_k = l_k,$$

и представим решение задачи $A_h w_h = \mu\, w_h$ в виде

$$\mu^{(m,n)} = \lambda^{(m)}(\Lambda_1) + \lambda^{(n)}(\Lambda_2) = \frac{4}{h_1^2}\sin^2\frac{\pi m h_1}{2l_1} + \frac{4}{h_2^2}\sin^2\frac{\pi n h_2}{2l_2}\,,$$

$$w_{ij}^{(m,n)} = v_i^{(m)}(\Lambda_1) v_j^{(n)}(\Lambda_2) = 2\sin\frac{\pi i h_1 m}{l_1}\sin\frac{\pi j h_2 n}{l_2}\,,$$

где собственные векторы $w_h^{(m,n)}$ ортонормированы относительно скалярного произведения $(u_h, v_h) = h_1 h_2 \sum\limits_{i=1}^{N_1-1}\sum\limits_{j=1}^{N_2-1} u_{ij} v_{ij} \equiv (u_h, v_h)_2 h_1 h_2$, т.е.

$$\left(w_h^{(m,n)}, w_h^{(p,q)}\right) = \delta_m^p \delta_n^q$$

Теперь, как и в одномерном случае, исследование устойчивости сводится к проверке условия $\mu_{\min} \ge C^{-1}$:

$$\mu_{\min}(A_h) = \lambda_{\min}(\Lambda_1) + \lambda_{\min}(\Lambda_2) \ge \frac{8}{l_1^2} + \frac{8}{l_2^2} = C^{-1},$$

где постоянная C не зависит от h. Поэтому $\|A_h^{-1}\|_2 \le C$ и

$$\|u_h\|_2 \le C\|f_h\|_2, \quad \|v_h\|_2^2 = (v_h, v_h)/h_1 h_2,$$

т.е. линейная разностная схема $A_h u_h = f_h$ устойчива в смысле сформулированного ранее определения. Напомним, что из теоремы А.Ф.Филиппова (о связи аппроксимации, устойчивости и сходимости) следует, что решение дискретной задачи **сходится** к решению дифференциальной с порядком не ниже второго, в силу устойчивости и аппроксимации $O(h_1^2 + h_2^2)$.

Однако остался невыясненным вопрос о нахождении решения полученной системы линейных алгебраических уравнений. Здесь он сложнее, чем в одномерном случае, так как метод прогонки неприменим (матрица системы $A_h u_h = f_h$ является уже не просто трехдиагональной, а блочно-трехдиагональной, причем каждый блок имеет размерность одномерной задачи). Однако, как и в одномерном случае, для нахождения сеточного решения уравнения Пуассона имеются подходы, основанные на методе Фурье. Несложно привести эти формулы в виде двухкратных рядов (сумм). Это — полезное упражнение для усвоения метода Фурье и оценки вычислительных затрат.

Лекция 29

Спектральный признак устойчивости и примеры его применения для аппроксимаций гиперболических уравнений. — Принцип замороженных коэффициентов.

Спектральный признак устойчивости и примеры его применения для аппроксимаций гиперболических уравнений

Построение и исследование разностных схем для уравнений в частных производных гиперболического типа традиционно проводится в открытой полуплоскости

$$D = \{(x,t) \,:\, -\infty < x < \infty,\, 0 < t\}$$

на примере уравнения для оператора переноса

$$Lu \equiv \frac{\partial u}{\partial t} + a(x,t)\,\frac{\partial u}{\partial x} = f(x,t)$$

(часто используется сокращенная форма записи — $u_t + a\,u_x = f$), с начальным условием

$$lu \equiv u(x,0) = u_0(x) \quad \text{при} \quad t = 0\,.$$

Далее сетка выбирается равномерной по обеим переменным

$$x_m = m\,h, \quad m = 0, \pm 1, \,\ldots\,; \quad t^n = n\,\tau, \quad n = 0, 1, \,\ldots\,,$$

а для значения сеточной функции u в точке (x_m, t^n) используется обозначение u_m^n. Аппроксимация оператора L не имеет каких-либо примечательных особенностей по сравнению с рассмотренными ранее примерами использования метода конечных разностей, поэтому этот этап можно опустить без потери содержательности.

Важным инструментом исследования устойчивости разностных схем является **спектральный признак (СПУ)**. Познакомимся с ним поближе.

В его основе лежит явная форма решения однородного уравнения $u_t + a\,u_x = 0,\ a = \text{const}$ с достаточно гладким ограниченным начальным условием $u(x,0) = u_0(x)$. Это решение имеет вид — $u(x,t) = u_0(x - at)$. Поэтому представляется разумным, чтобы разностная схема также из ограниченных начальных условий порождала аналогичные решения, т.е. равномерно ограниченные по времени или хотя бы не сильно растущие. В связи с этим требуется легко проверяемый критерий качества разностных схем.

Разностные схемы для однородного уравнения переноса **с постоянным коэффициентом** a можно записать в виде

$$L_h\,u_m^n \equiv \sum_{k,l} b_{lk}\,u_{m+l}^{n+k} = 0\,.$$

Все частные решения схемы имеют вид

$$u_m^n = [\lambda(\varphi)]^n\,e^{\,\mathbf{i}\,m\,\varphi}\,, \quad \mathbf{i} = \sqrt{-1}.$$

Критерий качества разностной схемы, т.е. **спектральный признак устойчивости**, формулируется так: *если при заданных законах стремления* τ *и*

h к нулю существует постоянная $0 \le c < \infty$, такая, что для всех $0 \le \varphi \le 2\pi$ справедливо неравенство

$$|\lambda(\varphi)| \le e^{c\tau}\,,$$

то схема удовлетворяет СПУ и может быть применена для численного решения задачи Коши для уравнения $L\,u = f$.

Применим этот подход для исследования простейшей схемы

$$L_h\,u_m^n \equiv \frac{u_m^{n+1} - u_m^n}{\tau} + a\,\frac{u_m^n - u_{m-1}^n}{h} = 0, \quad a = \text{const}.$$

Данная схема имеет порядок аппроксимации $O(\tau, h)$. Подставим в нее общий вид частного решения $u_m^n = \lambda^n\,e^{\mathbf{i}\,m\,\varphi}$. В результате будем иметь

$$\frac{\lambda^{n+1}e^{\mathbf{i}\,m\,\varphi} - \lambda^n e^{\mathbf{i}\,m\,\varphi}}{\tau} + a\,\frac{\lambda^n e^{\mathbf{i}\,m\,\varphi} - \lambda^n e^{\mathbf{i}\,(m-1)\,\varphi}}{h} = 0\,.$$

После сокращения на $\lambda^n\,e^{\mathbf{i}\,m\,\varphi}$ получим

$$\frac{\lambda - 1}{\tau} + a\,\frac{1 - e^{-\mathbf{i}\,\varphi}}{h} = 0\,,$$

откуда следует

$$\lambda(\varphi) = 1 - \frac{a\tau}{h} + \frac{a\tau}{h}\,e^{-\mathbf{i}\,\varphi}\,.$$

Пусть $a \ge 0$. Тогда при $0 \le \dfrac{a\tau}{h} \le 1$ имеем $\bigl|\lambda(\varphi)\bigr| \le 1 - \dfrac{a\tau}{h} + \dfrac{a\tau}{h} = 1$, т.е. схема устойчива при выполнении указанных выше условий. Если $\dfrac{a\tau}{h} = 1 + \sigma > 1$ (при $\tau, h \to 0$ $\sigma = \text{const} > 0$) и $\varphi = \pi$ получим $\lambda(\pi) = -1 - 2\sigma < -1$, т.е. в этом случае схема неустойчива. Пусть теперь $a < 0$. Тогда при $\varphi = \pi$ имеем $\lambda(\pi) = 1 - 2\dfrac{a\tau}{h}$, т.е. в этом случае схема не удовлетворяет СПУ. Таким образом, рассмотренная разностная схема условно устойчива только при $a > 0$.

Отметим, что аналогичные рассуждения справедливы при $a < 0$ для схемы

$$L_h\,u_m^n \equiv \frac{u_m^{n+1} - u_m^n}{\tau} + a\,\frac{u_{m+1}^n - u_m^n}{h} = 0, \quad a = \text{const}.$$

Теперь рассмотрим более сложный случай: схему с центральной разностью по пространству

$$L_h\,u_m^n \equiv \frac{u_m^{n+1} - u_m^n}{\tau} + a\,\frac{u_{m+1}^n - u_{m-1}^n}{2h} = 0, \quad a = \text{const}.$$

Данная схема имеет порядок аппроксимации $O\bigl(\tau, h^2\bigr)$. Аналогично предыдущему случаю получим

$$\lambda(\varphi) = 1 - \frac{a\tau}{2h}\left(e^{\mathbf{i}\varphi} - e^{-\mathbf{i}\varphi}\right) = 1 - \mathbf{i}\frac{a\tau}{h}\sin\varphi\,,$$

откуда следует, что

$$\max\bigl|\lambda(\varphi)\bigr| = \left|\lambda\left(\frac{\pi}{2}\right)\right| = \sqrt{1 + \frac{a^2\tau^2}{h^2}}\,.$$

Пусть $\tau = Ah^2, A = \text{const}$. Тогда

$$\left|\lambda\left(\frac{\pi}{2}\right)\right| = \sqrt{1 + a^2 A\tau} = 1 + a^2 A\,\frac{\tau}{2} + O(\tau^2) \le e^{c\tau},$$

где $c = \dfrac{a^2 A}{2}$ — постоянная в правой части неравенства из спектрального признака устойчивости.

Обратим внимание, что исследование устойчивости с помощью спектрального признака фактически позволяет находить искомые законы стремления τ и h к нулю, хотя сам признак формально является только необходимым условием устойчивости.

Принцип замороженных коэффициентов

Пусть решается задача Коши для нелинейного уравнения

$$\frac{\partial u}{\partial t} + \frac{\partial f(x,t,u)}{\partial x} = \psi(x,t,u)\,, \quad u(x,0) = u_0(x)\,.$$

Рассмотрим для уравнения простейшую разностную схему

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{f(x_m, t^n, u_m^n) - f(x_{m-1}, t^n, u_{m-1}^n)}{h} = \psi(x_m, t^n, u_m^n). \qquad (1)$$

Принцип замороженных коэффициентов (ПЗК) состоит из четырех пунктов. Перечислим их последовательно.
1. Пишется уравнение в вариациях для разностной схемы, т.е. уравнение для разности двух бесконечно близких решений с отбрасыванием слагаемых квадратичного и более высоких порядков малости. В общем случае это приводит к линейному уравнению с *переменными* коэффициентами.
2. Собственно "замораживание коэффициентов": превращаем уравнение с переменными коэффициентами в уравнение с постоянными коэффициентами. Фиксируется точка P в области решения задачи, и **все** значения коэффициентов полагаются равными их значениям в этой точке.
3. Для полученного уравнения с постоянными коэффициентами используем спектральный признак устойчивости. Условие устойчивости будет зависеть от значения коэффициентов в точке $P:\ \gamma(\tau, h, P) \ge 0$.
4. За условие устойчивости схемы $\bar\gamma(\tau, h) \ge 0$ принимают некоторое условие, из которого следует выполнение всех ранее полученных частных: $\gamma(\tau, h, P) \ge 0\ \forall P$. Как правило, требуют выполнения неравенства $\bar\gamma(\tau, h) \ge 0$ с некоторым "запасом устойчивости".

Рассмотрим реализацию ПЗК для нашей схемы.
1. Обозначим одно из решений схемы (1) за u_m^n, а второе, которое может быть получено малым возмущением начального условия, представим в виде

$$v_m^n = u_m^n + \delta_m^n\,.$$

Вариациями (изменениями) решения называются значения функции δ_m^n; предполагается, что $|\delta_m^n| \ll 1$. Подставим v_m^n в (1):

$$\frac{v_m^{n+1} - v_m^n}{\tau} + \frac{f(x_m, t^n, v_m^n) - f(x_{m-1}, t^n, v_{m-1}^n)}{h} = \psi(x_m, t^n, v_m^n)\,. \qquad (2)$$

Вычитая почленно из (2) уравнение (1) и применяя формулу конечных приращений Лагранжа для достаточно гладкой функции $g(z)$

$$g(y+\delta) - g(y) = g_z(\bar{y})\delta = g_z(y)\delta + o(\delta),$$

получаем:

$$\frac{\delta_m^{n+1} - \delta_m^n}{\tau} + \frac{f_u(x_m, t^n, u_m^n)\delta_m^n - f_u(x_{m-1}, t^n, u_{m-1}^n)\delta_{m-1}^n}{h} = \\ = \psi_u(x_m, t^n, u_m^n)\delta_m^n\,.$$

2. Произведем "замораживание"коэффициентов, т.е. положим

$$a = f_u(x_m, t^n, u_m^n), \quad b = \psi_u(x_m, t^n, u_m^n).$$

Так как для гладких функций $f_u(x,t,u)$ и $u(x,t)$ справедливо

$$f_u(x_{m-1}, t^n, u_{m-1}^n) = f_u(x_m, t^n, u_m^n) + O(h),$$

то величина $|O(h)\delta_{m-1}^n/h| = |\text{const}\cdot\delta_{m-1}^n| \ll 1$. Это дает возможность заменить значение $f_u(x_{m-1}, t^n, u_{m-1}^n)$ также на a. В результате получим уравнение с постоянными коэффициентами:

$$\frac{\delta_m^{n+1} - \delta_m^n}{\tau} + a\frac{\delta_m^n - \delta_{m-1}^n}{h} = b\,\delta_m^n\,.$$

3. Мы уже имели дело с устойчивостью этой схемы при $b=0$ при знакомстве с СПУ, там было получено условие вида $0 \le \dfrac{a\tau}{h} \le 1$ при $a \ge 0$. В нашем случае в правую часть добавится слагаемое $|b|\tau$:

$$|\lambda(\varphi)| \le 1 + |b|\tau \le e^{c\tau},\ c = |b|\,.$$

4. Теперь нам нужно избавиться от точки $P = (x_m, t^n)$, для этого вернемся от a к исходным обозначениям:

$$0 \le \frac{\partial f(x,t,u)}{\partial u}\frac{\tau}{h} \le 1\,.$$

Это условие для всей области определения решения. Если оно выполнено, то для $\forall P$ справедливо $a \ge 0,\ 0 \le \dfrac{a\tau}{h} \le 1$.

Определим дополнительно коэффициент запаса æ :

$$æ \le \frac{\partial f(x,t,u)}{\partial u}\frac{\tau}{h} \le 1 - æ\,,$$

где $0 < æ < 1/2$. На практике величина æ подбирается путем численного эксперимента.

Лекция 30

Исследование устойчивости простейших схем для уравнения теплопроводности в равномерной метрике. — Исследование устойчивости схемы с весами для уравнения теплопроводности в интегральной метрике (в норме $L_{2,h}$).

Построение и исследование разностных схем для уравнений в частных производных параболического типа традиционно проводится в открытой полуполосе

$$D = \{(x,t)\ : 0 < x < l,\ 0 < t\}$$

на примере простейшего уравнения теплопроводности

$$Lu \equiv \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x,t)$$

с начальным

$$u(x,0) = u_0(x) \quad \text{при} \quad t = 0$$

и краевыми условиями

$$u(0,t) = u(l,t) = 0 \quad \text{при} \quad \forall t \geq 0\,.$$

Предполагается, что начальная функция $u_0(x)$ удовлетворяет краевым условиям.

Отметим, что в общем случае на любом из концов отрезка краевое условие может быть задано в виде линейной комбинации функции и производной, например, при $x = 0$:

$$a(t)\,u(0,t) + b(t)\,\frac{\partial u}{\partial x}(0,t) = c(t).$$

Тогда необходимо обратить внимание на способ его аппроксимации.

Характерная особенность параболической задачи — смешанный тип данных: краевые условия по x и начальные по t. Поэтому исследование аппроксимации такое же как и в гиперболических уравнениях, а исследование устойчивости — принципиально другое.

Пусть сетка является равномерной по обеим переменным

$$x_m = m\,h, \quad m = 0, 1, \ldots, M, \quad M\,h = l; \quad t^n = n\tau, \quad n = 0, 1, \ldots N,$$

а для сеточной функции u в точке (x_m, t^n) используется обозначение u_m^n.

Исследование устойчивости простейших схем для уравнения теплопроводности в равномерной метрике

Займемся анализом устойчивости схем в равномерной метрике. Определим норму сеточной функции u_m^n на n-ом временном слое следующим образом:

$$\|u^n\| = \max_{0 \leq m \leq M} |u_m^n|.$$

Будем называть схему *устойчивой в равномерной метрике* на отрезке $[0,T], T = N\tau$, если имеет место неравенство

$$\max_{0 \leq n \leq N} \|u^n\| \leq \|u^0\| + c \max_{0 \leq n \leq N} \|f^n\|\,,$$

где c не зависит от шагов сетки τ и h, но может зависеть от величины T.

Рассмотрим в качестве примера две схемы (явную и полностью неявную), имеющие одинаковый порядок аппроксимации $O\left(\tau, h^2\right)$.

Исследуем устойчивость явной схемы

$$\frac{u_m^{n+1}-u_m^n}{\tau}=\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2}+f_m^n\,,\quad 0<m<M,\ Mh=l,$$
$$u_m^0=u_0(x_m),\quad u_0^n=u_M^n=0\ \ \forall n\geq 0\,.$$

Утверждение 1. *Явная схема является устойчивой в равномерной метрике при выполении условия* $\tau/h^2\leq 1/2$.

Доказательство. Введем обозначение $\rho=\tau/h^2$ и перепишем схему в удобном для анализа виде

$$u_m^{n+1}=(1-2\rho)\,u_m^n+\rho\left(u_{m+1}^n+u_{m-1}^n\right)+\tau f_m^n\,.$$

Поскольку максимальные по модулю значения обеих частей равенства по m совпадают, то при $\rho\leq\dfrac{1}{2}$ имеем

$$\begin{gathered}\|u^{n+1}\|\leq(1-2\rho)\|u^n\|+2\rho\|u^n\|+\tau\|f^n\|=\|u^n\|+\tau\|f^n\|\leq\\ \leq\|u^{n-1}\|+\tau\big(\|f^n\|+\|f^{n-1}\|\big)\leq\ldots\leq\\ \leq\|u^0\|+\sum_{k=0}^{n}\tau\|f^k\|\leq\|u^0\|+(n+1)\,\tau\max_n\|f^n\|\,.\end{gathered}$$

Следовательно, схема удовлетворяет условию устойчивости с постоянной $c=n\,\tau=T$ при условии $\dfrac{\tau}{h^2}\leq\dfrac{1}{2}$. Утверждение доказано.

Рассмотрим теперь полностью неявную схему

$$\frac{u_m^{n+1}-u_m^n}{\tau}=\frac{u_{m-1}^{n+1}-2u_m^{n+1}+u_{m+1}^{n+1}}{h^2}+f_m^{n+1},\quad 0<m<M,\ Mh=l,$$
$$u_m^0=u_0(x_m),\quad u_0^n=u_M^n=0\,\forall n\geq 0.$$

Утверждение 2. *Полностью неявная схема является устойчивой в равномерной метрике при любых значениях* τ *и* h.

Доказательство. В данном случае удобная для анализа форма записи имеет вид

$$u_m^{n+1}+\rho\left(-u_{m-1}^{n+1}+2u_m^{n+1}-u_{m+1}^{n+1}\right)=u_m^n+\tau f_m^{n+1}\,.$$

Теперь из всех значений u_m^{n+1}, по модулю равных $\left\|u^{n+1}\right\|$, выберем такое, у которого индекс m принимает наименьшее значение. Заметим, что, в силу граничных условий, $1\leq m\leq M-1$. В этом случае имеем

$$\left|u_m^{n+1}\right|>\left|u_{m-1}^{n+1}\right|\qquad\text{и}\qquad\left|u_m^{n+1}\right|\geq\left|u_{m+1}^{n+1}\right|.$$

Отсюда $\left|2u_m^{n+1}\right|>\left(\left|u_{m-1}^{n+1}\right|+\left|u_{m+1}^{n+1}\right|\right)$, и знак выражения $2u_m^{n+1}-u_{m-1}^{n+1}-u_{m+1}^{n+1}$ совпадает со знаком u_m^{n+1}, т.е.

$$\left\|u^{n+1}\right\|=\left|u_m^{n+1}\right|<\left|u_m^{n+1}+\rho\big(2u_m^{n+1}-u_{m-1}^{n+1}-u_{m+1}^{n+1}\big)\right|,$$

а выражение справа равно $\left|u_m^n+\tau f_m^{n+1}\right|$. Поэтому при любых шагах сетки τ и h справедливо $\left\|u^{n+1}\right\|\leq\left\|u^n\right\|+\tau\left\|f^{n+1}\right\|$. Дальнейший вывод оценки

устойчивости выполняется по аналогии с предыдущим случаем. Утверждение доказано.

Отметим, что в неявной схеме шаги τ и h выбираются только из соображений аппроксимации, но за это приходится расплачиваться более сложным алгоритмом нахождения решения (требуется использование метода прогонки).

Исследование устойчивости схемы с весами для уравнения теплопроводности в интегральной метрике (в норме $L_{2,h}$)

Перейдем к иследованию устойчивости более сложных схем. Введем обозначение

$$\Lambda u_m = \frac{u_{m+1} - 2u_m + u_{m-1}}{h^2}$$

и рассмотрим схему *с весами* при $0 \le \delta \le 1$

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \delta \Lambda u_m^{n+1} + (1-\delta)\Lambda u_m^n\,, \quad 0 < m < M,\ Mh = l,$$
$$u_m^0 = u_0(x_m), \quad u_0^n = u_M^n = 0 \ \ \forall n \ge 0$$

для однородного уравнения $u_t = u_{xx}$. Отметим, что разложения в ряд Тейлора в точке $(x_m, t^{n+1/2})$ приводят к порядкам аппроксимации $O(\tau^2, h^2)$ при $\delta = \frac{1}{2}$ и $O(\tau, h^2)$ при остальных значениях δ (не путать с аппроксимацией на решении, где значение $\delta = 1/2 - h^2/(12\tau)$ приводит к погрешности $O(\tau^2, h^4)$). При этом для реализации схемы в случае $\delta \ne 0$ требуется метод прогонки, для которого выполнены достаточные условия устойчивости (как и для полностью неявной схемы).

Положим

$$\|u^n\|_{L_{2,h}} = \left(h \sum_{m=1}^{M-1} (u_m^n)^2 \right)^{1/2}$$

и назовем однородную разностную схему *устойчивой по начальным данным в интегральной метрике (в норме пространства $L_{2,h}$)* на отрезке $[0, T], T = N\tau$, если справедливо неравенство

$$\max_{0 \le n \le N} \|u^n\|_{L_{2,h}} \le \|u^0\|_{L_{2,h}}.$$

В общем случае в правой части в виде множителя может присутствовать постоянная c, не зависящая от шагов сетки τ и h.

Утверждение 3. *Схема с весами является устойчивой по начальным данным в интегральной метрике (в норме $L_{2,h}$):*

1) при любых значениях τ и h для $1/2 \le \delta \le 1$;

2) при выполнении условия $\tau/h^2 \le 1/(2-4\delta)$ для $0 \le \delta < 1/2$.

Доказательство. Во избежание избыточных обозначений будем понимать под Λ матрицу, которая ставит в соответствие вектору $u^n = (u_1^n, \ldots, u_{M-1}^n)^T$ вектор с отброшенными (удаленными) нулевыми граничными значениями $\Lambda u^n = (\Lambda u_1^n, \ldots, \Lambda u_{M-1}^n)^T$. Тогда рассматриваемую схему можно записать в виде

$$u^{n+1} = S\,u^n = S^{n+1} u^0\,, \qquad \text{где} \qquad S = (I - \tau\delta\Lambda)^{-1}(I + \tau(1-\delta)\Lambda)\,,$$

а I обозначает единичную матрицу.

Вспомним решение задачи на собственные значения

$$\Lambda v_m = -\nu v_m, \quad 1 \le m \le M-1,\, Mh = l, \quad v_0 = v_M = 0\,.$$

Его можно записать в форме

$$v_m^{(k)} = \sqrt{\frac{2}{l}} \sin\left(\frac{\pi m k}{M}\right), \quad \nu^{(k)} = \frac{4}{h^2}\sin^2\frac{\pi k h}{2l}, \quad 1 \le k \le M-1\,.$$

Воспользовавшись фактом, что матрица $(I+\beta\Lambda)^{\pm 1}$ имеет ту же систему собственных векторов, что и Λ, мы можем выразить собственные значения матрицы S через величины $\nu^{(k)}$. Действительно, пусть $S\,y = \lambda\,y$. Возьмем в качестве y вектор $v^{(k)}$, тогда для соответствующего $\lambda^{(k)}$ получим явное выражение

$$\lambda^{(k)}(S) = \frac{1-\tau(1-\delta)\nu^{(k)}}{1+\tau\delta\nu^{(k)}}\,.$$

Отсюда также следует, что матрица S является симметричной, так как имеет представление $S = QDQ^{-1}$, где столбцами ортогональной матрицы Q являются векторы $v^{(k)}$, а D — диагональная матрица, состоящая из соответствующих $\lambda^{(k)}$. Ранее мы уже встречали, что матричная норма, подчиненная векторной евклидовой норме, равна $\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$. Причем для симметричной матрицы выражение упрощается: $\|A\|_2 = \max|\lambda(A)|$. Метрика $L_{2,h}$ отличается от евклидовой только множителем h, поэтому справедливо

$$\|S\|_{L_{2,h}} = \max_k |\lambda^{(k)}(S)|\,.$$

Выясним теперь, когда $\|S\|_{L_{2,h}} \le 1$:

$$-1 \le \frac{1-\tau(1-\delta)\nu^{(k)}}{1+\tau\delta\nu^{(k)}} \le 1\,.$$

для $k = 1, 2, \ldots, M-1$. Так как $\tau, \nu^{(k)} > 0,\ \delta \ge 0$, то знаменатель всегда положителен, поэтому

$$-\left[1+\tau\delta\nu^{(k)}\right] \le 1-\tau(1-\delta)\nu^{(k)} \le 1+\tau\delta\nu^{(k)}\,.$$

Заметим, что правое неравенство выполнено всегда, значит, содержательным является левое. Перепишем его в виде

$$\tau(1-2\delta)\nu^{(k)} \le 2\,.$$

При $1/2 \le \delta \le 1$ это неравенство выполнено при любом τ, а при $0 \le \delta < 1/2$ возникает ограничение

$$\tau \le \frac{2}{(1-2\delta)\max\limits_k \nu^{(k)}} \le \frac{h^2}{2-4\delta} \quad \left(\text{так как } \max_k \nu^{(k)} \le \frac{4}{h^2}\right).$$

Из полученной выше формулы $u^n = S^n u^0$ следует

$$\|u^n\|_{L_{2,h}} \le \|S\|^n_{L_{2,h}} \|u^0\|_{L_{2,h}}\,.$$

Поэтому для всего диапазона $\delta \in [0,1]$ имеем устойчивость в метрике $L_{2,h}$ с постоянной $c = 1$. Утверждение доказано.

В завершение отметим, что такой метод исследования устойчивости неприменим, если бы в исходной задаче были переменные коэффициенты, более сложные граничные условия и т.п. В таких ситуациях используют наиболее общий метод — метод априорных оценок, с которым мы знакомились на примере более простой задачи для обыкновенного дифференциального уравнения второго порядка.