методы отчислений

Лектор: Б. А. Самокиш

10.01.2019
версия от 09.01.2019 05:39:35, до экзамена 29 часов

Оглавление

1 Однородные дифференциальные уравнения

§ 1. Краевая задача для ОДУ 2 порядка и сведение к задаче Коши

Определение 1. Рассмотрим ОДУ 2 порядка

$$y'' + p(x)y' + q(x)y = f(x) \qquad y \in C^2([a;b])$$

и 3 варианта условий на y

I $y(a) = A, \quad y(b) = B$

II $y'(a) = A, \quad y'(b) = B$

III $y'(a) = \alpha y(a) + A, \quad y'(b) = \beta y(b) + B$

Если y — решение для которого выполнено какое-то из условий выше, то y — решение граничной задачи.

Определение 2 (Однородная краевая задача)**.** Положим $f \equiv 0$ в 1.1.1.

Определение 3 (Однородные граничные условия)**.** Положим $A = B = 0$ в граничных условиях в 1.1.1

Теорема 1 (об альтернативе)**.** Рассмотрим однородную граничную задачу с однородными граничными условиями. Пусть y_H — решение однородной задачи.

Тогда

1. $y_0 \equiv 0$ — единственное решение однородной задачи $\Rightarrow$ неоднородная краевая задача имеет единственное решение
2. $y_0 \equiv 0$ — неединственное решение однородной задачи $\Rightarrow$ неоднородная краевая задача имеет бесконечно много или не имеет решений вовсе

□ Рассмотреть решение неоднородной краевой в виде $y(x) = y_0(x) + c_1y_1(x) + c_2y_2(x)$ и подставить граничные условия, а дальше все следует из линейной алгебры. ■

Разберёмся как численно найти y_0, y_1, y_2, потребовавшиеся в предыдущем доказательстве. Будем считать что p, q, f определены на $I \ni [a;b]$, так что y можно продолжить на $(a-\varepsilon; b+\varepsilon)$.

1. $y(a) = 0, y'(a) = 0$. Поскольку 0 явно решение однородной задачи, то что мы найдем будет как раз частным решением неоднородной задачи (Коши!).
2. $y_H(a) = 1, y_H'(a) = 0$ и решаем мы тут однородную задачу (Коши!). Будем считать то что нашлось y_1
3. $y_H(a) = 0, y_H'(a) = 1$. Скажем что это y_2. Здесь важно заметить про линейную независимость y_1 и y_2. Найдем определитель Вронского в точке a

$$W = \begin{vmatrix} y_1(a) & y_2(a) \\ y_1'(a) & y_2'(a) \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1 \neq 0$$

   А тогда он нигде не ноль. А значит y_1 и y_2 линейно независимы.

Всё это называется *методом начальных данных* для решения краевой задачи.

В рассуждении выше можно было бы взять другие начальные данные дабы упростить себе жизнь. Ведь никто не запрещает запихать, например, кусок y_2 в y_0 (если мы уже знаем правильное c_2). Нам просто были нужны какие-то линейно независимые решения однородной задачи.

Рассмотрим граничную задачу в форме III

1. $y(a) = 0, y'(a) = A$, нашли y_0.
2. $y_H(a) = 1, y'_H(a) = \alpha$, нашли y_1.
3. $y_H(a) = 0, y'_H(a) = 0$. Мы просто решили что $y_2 \equiv 0$. Эту ЗК мы даже не решаем, а сразу знаем ответ.

При таком раскладе $y(x) = y_0(x) + c_1 y_1(x)$. Проверим левое граничное условие

$$\begin{aligned} y(a) &= y_0(a) + c_1\, y_1(a) = 0 + c_1\, 1 = c_1 \\ y'(a) &= y'_0(a) + c_1\, y'_1(a) = A + c_1\, \alpha = A + \alpha\, y(a) \end{aligned}$$

Как видно, всё получилось.

В случае I можно сделать так:

1. $y(a) = A, y'(a) = 0$, нашли y_0.
2. $y_H(a) = 0, y'_H(a) = 1$, нашли y_1.

Как видно, свободы в выборе c_1 хватает чтобы разобраться с правой границей.

Пример 1. $y'' - q^2 y = 0, y(0) = 1, y(b) = 1$

⟨⚒⟩, а он важный вообще-то, из него необходимость метода прогонки следует.

§ 2. Метод дифференциальной прогонки

Здесь будем решать краевую задачу с граничными условиями в форме III.

Рассмотрим $\alpha(x), \beta(x)$: (прогоночные коэффициенты)

$$y'(x) = \alpha(x)\, y(x) + \beta(x) \tag{1.1}$$

Такая форма напрашивается при вспоминании трюка, который мы делали в прошлом параграфе. Там как раз $y'(a) = y_0(a) + c_1\, y_1(a)$, а $c_1 = y(a)$. Здесь мы пока вводим прогоночные коэффициенты формально, а существование покажем конструктивно.

Найдем уравнения на α, β

$$\begin{aligned} y' &= \alpha y + \beta \\ y'' &= \alpha y' + \alpha' y + \beta' \end{aligned} \quad \Rightarrow \quad \begin{aligned} & y'' + py' + qy = f \\ \Leftrightarrow\ & \alpha y' + \alpha' y + \beta' + p(\alpha y + \beta) + qy = f \\ \Leftrightarrow\ & y\,\underbrace{(\alpha^2 + \alpha' + p\alpha + q)}_{0} + \underbrace{\alpha\beta + \beta' + p\beta}_{f} = f \end{aligned}$$

В итоге получаем систему ОДУ первого порядка

$$\begin{aligned} \alpha' &= -\alpha^2 - p\alpha - q \\ \beta' &= f - p\beta - \alpha\beta \end{aligned} \tag{1.2}$$

Посмотрим что происходит на правом конце[1]

$$\begin{aligned} y'(b) &= \alpha(b) y(b) + \beta(b) \\ y'(b) &= \beta y(b) + B \end{aligned} \quad \Rightarrow \quad y(b) = \frac{B - \beta(b)}{\alpha(b) - \beta}$$

Сам метод выглядит так:

прямая прогонка: решаем систему (1.2) с начальными данными $\alpha(a) = \alpha, \beta(a) = A$.

обратная прогонка: уже зная $\alpha(x), \beta(x)$ решаем (1.1) с начальными данными $y(b) = \frac{B - \beta(b)}{\alpha(b) - \beta}$.

Замечание 1. Рассмотрим однородную задачу с однородными граничными условиями. Тогда (1.1) переходит в $y'(x) = \alpha(x)\, y(x)$. Если при этом $\exists\, c \in (a; b)$: $y(c) = 0 \wedge y'(c) \neq 0$, то $\alpha(c)$ не существует. Так что, как видно, не всякое решение краевой задачи можно найти методом прогонки.

[1] а что делать если $\alpha(b) = \beta$ неясно

§ 3. Метод прогонки для систем ОДУ

Определение 1. Рассмотрим ОДУ

$$\boldsymbol{y}' = \hat{A}(x)\,\boldsymbol{y} + \boldsymbol{f}(x) \qquad \boldsymbol{y} \in C^2([a;b]), \quad y\colon [a;b] \to \mathbb{R}^s \tag{1.1}$$

и условия вида

$$\begin{aligned} x &= a \quad \hat{\alpha}\boldsymbol{y}(a) = \boldsymbol{\beta} \qquad && \hat{\alpha}\colon \mathbb{R}^s \to \mathbb{R}^p \\ x &= b \quad \hat{\gamma}\boldsymbol{y}(b) = \boldsymbol{\delta} \qquad && \hat{\gamma}\colon \mathbb{R}^s \to \mathbb{R}^q \\ & && s = p + q \end{aligned}$$

Если y — решение для которого выполнено условие выше, то y — решение граничной задачи.

Замечание 1. Вообще, граничные условия бывают куда более общего вида, но мы их не рассматриваем. То, что у нас — это линейные распадающиеся граничные условия.

А вот так выглядят нераспадающиеся:

$$\hat{\alpha}\boldsymbol{y}(a) + \hat{\gamma}\boldsymbol{y}(b) = \boldsymbol{\beta}, \qquad \hat{\alpha}, \hat{\gamma}\colon \mathbb{R}^s \to \mathbb{R}^s$$

Общее решение задачи Коши 1.1 имеет вид

$$\boldsymbol{y}(x) = \boldsymbol{y}_0(x) + \sum_{j=1}^{s} c_j \boldsymbol{y}_j(x)$$

где как обычно y_0 — решение неоднородной задачи Коши, а $\{y_j\}$ — фундаментальная система решений однородной.

1. Метод начальных данных такой же в §1 — находим из граничных условий $\{c_j\}$ в общем решении.

Чтобы добыть решения задач Коши можно взять $\boldsymbol{y}_j(a) = \boldsymbol{e}_j$ (это единичный вектор с 1 на jом месте), $\boldsymbol{y}_0(a) = 0$

Можно снова уменьшить количество работы

1. в качестве начальных данных для y_0 — какое-нибудь решение системы $\hat{\alpha}y = \beta$
2. в качестве начальных данных для y_j, $j \in p+1 \mathinner{..} s$ — q линейно независимых решений $\hat{\alpha}y = 0$
3. $y_j \equiv 0$, $j \in 1 \mathinner{..} p$

В итоге решение примет вид

$$\boldsymbol{y}(x) = \boldsymbol{y}_0(x) + \sum_{j=p+1}^{s} c_j \boldsymbol{y}_j(x)$$

`<+здесь снова этот понятный кусок про экспоненты и беды вычислений+>`⟨⚒⟩

2. Метод прогонки, в котором снова зададим $\hat{\alpha}, \boldsymbol{\beta}$

$$\hat{\alpha}(x)\,\boldsymbol{y}(x) = \boldsymbol{\beta}(x), \qquad \hat{\alpha}\colon [a;b] \times \mathbb{R}^s \to \mathbb{R}^p \tag{1.2}$$

При таком условии $\forall\, x \;::\; y(x) \in M \subset \mathbb{R}^s$, $\dim M = s - p$ (предполагая что $\operatorname{rk}\hat{\alpha}(x) = p$)

Найдем уравнения на $\hat{\alpha}(x), \boldsymbol{\beta}(x)$

$$\begin{aligned} \hat{\alpha}\boldsymbol{y} &= \boldsymbol{\beta} \\ \hat{\alpha}\boldsymbol{y}' + \hat{\alpha}'\boldsymbol{y} &= \boldsymbol{\beta}' \end{aligned} \quad \Rightarrow \quad \begin{aligned} & \boldsymbol{y}' = \hat{A}\boldsymbol{y} + \boldsymbol{f} \\ & \hat{\alpha}\hat{A}\boldsymbol{y} + \hat{\alpha}\boldsymbol{f} + \hat{\alpha}'\boldsymbol{y} = \boldsymbol{\beta}' \\ \Leftrightarrow\ & \underbrace{(\hat{\alpha}\hat{A} + \hat{\alpha}')}\,\boldsymbol{y} - \underbrace{\boldsymbol{\beta}'} = -\hat{\alpha}\boldsymbol{f} \end{aligned}$$

Пусть α_j — строка $\hat{\alpha}$. Тогда мы получаем систему ОДУ первого порядка

$$\begin{aligned} \alpha_j' &= -\hat{A}^T \alpha_j \\ \beta_j' &= (\alpha_j, \boldsymbol{f}) \end{aligned} \tag{1.3}$$

Посмотрим что происходит на правом конце

$$\begin{cases} \hat{\alpha}(b)\,\boldsymbol{y}(b) = \boldsymbol{\beta}(b) \\ \hat{\gamma}\,\boldsymbol{y}(b) = \boldsymbol{\delta} \end{cases} \tag{1.4}$$

это просто линейная система порядка s на $\boldsymbol{y}(b)$, решаем и находим.

Сам метод выглядит так:

прямая прогонка: решаем прогоночные уравнения(1.3) с начальными данными $\hat{\alpha}(a) = \hat{\alpha}$, $\boldsymbol{\beta}(a) = \boldsymbol{\beta}$.

обратная прогонка: уже зная $\alpha(x)$, $\beta(x)$ решаем (1.2) с начальными данными $y(b)$, найденным из системы (1.4).

Замечание 2. $\langle ? \rangle$ Метод с заменой $\hat{A}$ на сопряженную в прогоночных уравнениях уже пафосно называется методом *сопряжённых систем*, но ничем кроме названия по сути не отличается.

Замечание 3. Вообще, этот метод накладывает слишком жёсткие условия на $\hat{\alpha}$, β. Например краевая задача из §1 им не решается. Проблема возникает в том месте, где из $(\hat{\alpha}\hat{A} + \hat{\alpha}')\,\boldsymbol{y} = 0$ выводится $\hat{\alpha}\hat{A} + \hat{\alpha}' = 0$. Произвольностью $\boldsymbol{y}$ мы вообще-то пользоваться не можем, так как на него есть условие $\hat{\alpha}(x)\,\boldsymbol{y}(x) = \boldsymbol{\beta}(x)$.

Замечание 4. У вышеописанного метода есть ещё пара недостатков:

1. $\alpha_j' = -\hat{A}^T \alpha_j$ отличается от исходной системы только отсутствием неоднородности, так от проблем связанных с потерей точности из-за собственных чисел разного знака в решениях задач Коши мы убежать не смогли.

2. $\hat{\alpha}\hat{\alpha}^T$ может быть плохо обусловленной и ища $\boldsymbol{y}(b)$ мы потеряем точность. [1]

Собственно, для того чтобы обойти эти проблемы и нужен §4.

§4. Ортогональная прогонка

«будем решать немного другую задачу»

Заменим уравенение для $\hat{\alpha}$ в методе выше.

$$\hat{\alpha}' = -\hat{\alpha}\hat{A} \longrightarrow \hat{\alpha}' = -\hat{\alpha}\hat{A} + \hat{\alpha}\hat{A}\hat{\alpha}^T \left(\hat{\alpha}\hat{\alpha}^T\right)^{-1} \hat{\alpha}$$

Крокодил в формуле сверху — ортогонанальная проекция $\hat{\alpha}\hat{A}$ на $\hat{\alpha}$, а скалярное произведение имеет вид $(\hat{\alpha}, \hat{\beta}) = \hat{\alpha}\hat{\beta}^T$[2]. Так что по идее, раз мы проекцию на $\hat{\alpha}$ вычли, $(\hat{\alpha}', \hat{\alpha}) = 0$. Проверим:

$$\hat{\alpha}'\hat{\alpha}^T = -\hat{\alpha}\hat{A}\hat{\alpha}^T + \hat{\alpha}\hat{A}\hat{\alpha}^T \left(\hat{\alpha}\hat{\alpha}^T\right)^{-1} \hat{\alpha}\hat{\alpha}^T = -\hat{\alpha}\hat{A}\hat{\alpha}^T + \hat{\alpha}\hat{A}\hat{\alpha}^T = 0$$

Отсюда следует, что $\frac{\mathrm{d}}{\mathrm{d}x}\left(\hat{\alpha}\hat{\alpha}^T\right) = 0$, так что матрица $\hat{\alpha}\hat{\alpha}^T$ постоянна на всём $[a; b]$

Получим уравнения на прогоночные коэффициенты

$$\begin{aligned} &\hat{\alpha}\hat{A}\boldsymbol{y} + \hat{\alpha}\boldsymbol{f} + \hat{\alpha}'\boldsymbol{y} = \boldsymbol{\beta}' \\ \Leftrightarrow\ &\hat{\alpha}\hat{A}\boldsymbol{y} + \hat{\alpha}\boldsymbol{f} - \hat{\alpha}\hat{A}\boldsymbol{y} + \hat{\alpha}\hat{A}\hat{\alpha}^T \left(\hat{\alpha}\hat{\alpha}^T\right)^{-1} \underbrace{\hat{\alpha}\boldsymbol{y}}_{\boldsymbol{\beta}} = \boldsymbol{\beta}' \end{aligned}$$

В итоге

$$\begin{aligned} \hat{\alpha}' &= -\hat{\alpha}\hat{A} + \hat{\alpha}\hat{A}\hat{\alpha}^T \left(\hat{\alpha}\hat{\alpha}^T\right)^{-1} \hat{\alpha} \\ \boldsymbol{\beta}' &= \hat{\alpha}\boldsymbol{f} + \hat{\alpha}\hat{A}\hat{\alpha}^T \left(\hat{\alpha}\hat{\alpha}^T\right)^{-1} \boldsymbol{\beta} \end{aligned} \tag{1.1}$$

Разберёмся что делать с $\left(\alpha\alpha^T\right)^{-1}$. Не очень приятно каждый раз искать обратную матрицу.

На левой границе $\hat{\alpha}(a)\boldsymbol{y}(a) = \boldsymbol{\beta}(a)$. Проведём процесс Грамма-Шмидта и ортогонализуем строчки $\hat{\alpha}(a)$. При этом заменили переменную в исходном уравнении, соответственно поменялись $\hat{A} \to \hat{B}$, $\boldsymbol{f} \to \boldsymbol{g}$. Зато $\hat{\alpha}(a)\,\hat{\alpha}(a)^T = I$. Так что прогоночные уравнения принимают вид

$$\begin{aligned} \hat{\alpha}' &= -\hat{\alpha}\hat{B} + \hat{\alpha}\hat{B}\hat{\alpha}^T\hat{\alpha} \\ \boldsymbol{\beta}' &= \hat{\alpha}\boldsymbol{g} + \hat{\alpha}\hat{B}\hat{\alpha}^T\boldsymbol{\beta} \end{aligned} \tag{1.2}$$

[1] $\operatorname{rk}\hat{\alpha}\hat{\alpha}^T \geqslant 2\operatorname{rk}\hat{\alpha} - s$ из теоремы Сильвестра о ранге, так что так в одну сторону вроде можно

[2] на самом деле оно несимметрично. Нужно здесь понимать матрицу как набор векторов-строк. Тогда какой-то смысл есть.

Поскольку $\hat{\alpha}\hat{\alpha}^T$ постоянна на $[a; b]$ (всюду I), то проблем с её плохой обусловленностью в $x = b$ нет. Правое граничное условие решится.

Судя по всему, это же условие исключает быстрый рост компонент $\hat{\alpha}$. Так что обе проблемы из замечания в конце предыдущего параграфа снимаются. $\langle ? \rangle$[1]

Вышеописанный метод ещё называется методом Абрамова.

§ 5. Разностный метод для краевой задачи 2 порядка

Предупреждение: в силу повышенной техничности этого параграфа он написан в соответствующем стиле. Что поделать. Приятного прочтения.

Решать краевую задачу для дифференциального уравнения второго порядка

$$y'' + p(x)y' + q(x)y = f(x) \qquad y \in C^2([a; b]) \tag{1.1}$$

1. Алгоритм

Определение 1 (Метод разностной прогонки). Пусть задано дифференциальное уравнение с граничными условиями. Методом разностной прогонки называет следующий алгоритм:

1. Выбор сетки: узлы, шаг (если она равномерная)
2. Построение сеточных уравнений
   (a) Диффур в узлах сетки
   (b) Все производные через конечные разности
3. Решение получившейся линейной системы

Будем дальше всюду считать, что решение задано на $[a; b]$

$$n \text{ узлов} \qquad h = \frac{b-a}{n} \qquad x_k = a + kh$$

2. Формулы численного дифференцирования

Здесь $M_n = \max\left|y^{(m)}(x)\right|$

$$y'(x) = \frac{y(x+h) - y(x)}{h} + R \qquad |R| \leqslant \tfrac{hM_2}{2} \tag{1.2}$$

$$y'(x) = \frac{y(x) - y(x-h)}{h} + R \qquad |R| \leqslant \tfrac{hM_2}{2} \tag{1.3}$$

$$y'(x) = \frac{-y(x+2h) + 4y(x+h) - 3y(x)}{2h} + R \qquad |R| \leqslant \tfrac{h^2M_3}{3} \tag{1.4}$$

$$y'(x) = \frac{3y(x) - 4y(x-h) + y(x-2h)}{2h} + R \qquad |R| \leqslant \tfrac{h^2M_3}{3} \tag{1.5}$$

$$y'(x) = \frac{y(x+h) - y(x-h)}{2h} + R \qquad |R| \leqslant \tfrac{h^2M_3}{6} \tag{1.6}$$

$$y''(x) = \frac{y(x+h) - 2y(x) + y(x-h)}{h^2} + R \qquad |R| \leqslant \tfrac{h^2M_4}{12} \tag{1.7}$$

Схемы 1.2 и 1.3 называются простейшими.

3. Разностное уравнение

$$\frac{y(x_{k+1}) - 2y(x_k) + y(x_{k-1})}{h^2} + R_1 + p(x_k)\frac{y(x_{k+1}) - y(x_{k-1})}{2h} + p(x_k)R_2 + q(x_k)\,y(x_k) = f(x_k)$$

Можно заметить, что $R_1 + p(x_k)R_2 = O(h^2)$. Так что можно вместо $y(x_k)$ получить приближённое решение y_k (по сути, решение уже совсем другой задачи). Попутно, обозначим

$$p(x_k) = p_k, \qquad q(x_k) = q_k, \qquad f(x_k) = f_k.$$

Получится

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + p_k\frac{y_{k+1} - y_{k-1}}{2h} + q_k y_k = f_k \tag{1.8}$$

[1] про это два слова в Крылове написано и больше нигде нет.

4. Граничные условия будут рассматриваться III типа, но вообще это неважно. Всё равно раскрывать не будем.

1. *Трёхточечная односторонняя аппроксимация*

$$y'(x) = \frac{-y(x+2h) + 4y(x+h) - 3y(x)}{2h} + O(h^2)$$

Запишем это выражение для границ:

$$y'(a) = \frac{-y(a+2h) + 4y(a+h) - 3y(a)}{2h} + O(h^2) \quad \to \quad \alpha y_0 + A = \frac{-y_2 + 4y_1 - 3y_0}{2h}$$
$$y'(b) = \frac{3y(b) - 4y(b-h) + y(b-2h)}{2h} + O(h^2) \quad \to \quad \beta y_n + B = \frac{3y_n - 4y_{n-1} + y_{n-2}}{2h}$$

2. *Метод фиктивных узлов*

(a) Введём $y_{-1} = y(a-h)$, $y_{n+1} = y(b+h)$

$$y'(x) = \frac{y(x+h) - y(x-h)}{2h} + O(h^2)$$

Запишем это выражение для границ:

$$y'(a) = \frac{y(a+h) - y(a-h)}{2h} + O(h^2) \quad \to \quad \alpha y_0 + A = \frac{y_1 - y_{-1}}{2h}$$
$$y'(b) = \frac{y(b+h) - y(b-h)}{2h} + O(h^2) \quad \to \quad \beta y_n + B = \frac{y_{n+1} - y_{n-1}}{2h}$$

Как правило, решение можно продолжить с отрезка на интервал побольше, подберём $h : \ y(a-h) \in I \wedge y(b+h) \in I$. Так что такой метод имеет смысл.

(b) Сдвинем сетку на $h/2$, $x_0 = a - h/2$, $x_{n+1} = b + h/2$

$$x_k = a - h/2 + kh \qquad k = 0, 1 \,..\, n+1$$

Значения в узлах сетки при этом придётся вводить с помощью интерполяции

$$y(a) = \frac{y(a - h/2) + y(a + h/2)}{2}$$

Сами выражения для производной имеют вид

$$y'(x) = \frac{y(x + h/2) - y(x - h/2)}{h} + O(h^2)$$

Запишем это выражение для границ:

$$y'(a) = \frac{y(a + h/2) - y(a - h/2)}{h} + O(h^2) \quad \to \quad \alpha\frac{y_0 + y_1}{2} + A = \frac{y_1 - y_0}{h}$$
$$y'(b) = \frac{y(b + h/2) - y(b - h/2)}{h} + O(h^2) \quad \to \quad \beta\frac{y_{n+1} + y_n}{2} + B = \frac{y_{n+1} - y_n}{h}$$

Такой подход не очень удобен если нужны значения в узлах. Придётся уменьшать шаг в 2 раза.

3. *Использование ДУ для исключения главного члена простейшей формулы*

$$y'(x) = \frac{y(x+h) - y(x)}{h} + R, \quad R = O(h)$$

Теперь запишем разложение в ряд Тейлора:

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2}y''(x) + O(h^3) \ \Leftrightarrow \ y'(x) = \frac{y(x+h) - y(x)}{h} - \frac{h}{2}y''(x) + O(h^2)$$

Из исходного уравнения (1.1) подстановкой простейшей формулы получаем

$$y''(x) = -\big(p(x)y'(x) + q(x)y(x)\big) + f(x) = -\Big(p(x)\,\tfrac{y(x+h)-y(x)}{h} + p(x)O(h) + q(x)y(x)\Big) + f(x)$$

$$\Rightarrow\ -\tfrac{h}{2}y''(x) = \tfrac{p(x)}{2}\Big(y(x+h) - y(x)\Big) + \tfrac{h}{2}\Big(q(x)y(x) - f(x)\Big) + O(h^2)$$

Оценка производной на краю

$$y'(a) = \frac{y(a+h)-y(a)}{h} + \frac{p(a)}{2}\Big(y(a+h) - y(a)\Big) + \frac{h}{2}\Big(q(a)y(a) - f(a)\Big) + O(h^2)$$
$$y'(b) = \frac{y(b)-y(b-h)}{h} - \frac{p(b)}{2}\Big(y(b) - y(b-h)\Big) - \frac{h}{2}\Big(q(b)y(b) - f(b)\Big) + O(h^2)$$

На сетке оно имеет вид

$$\alpha y_0 + A = \frac{y_1 - y_0}{h} + \frac{p_0}{2}\Big(y_1 - y_0\Big) + \frac{h}{2}\Big(q_0 y_0 - f_0\Big)$$
$$\beta y_n + B = \frac{y_n - y_{n-1}}{h} - \frac{p_n}{2}\Big(y_n - y_{n-1}\Big) - \frac{h}{2}\Big(q_n y_n - f_n\Big)$$

5. Составление системы линейных уравнений

$$\begin{aligned} c_0 y_1 - b_0 y_0 &= d_0 \\ c_k y_{k+1} - b_k y_k + y_{k-1} a_k &= d_k, \quad k = 1\,..\,n-1 \\ -\, b_n y_n + a_n y_{n-1} &= d_n \end{aligned} \tag{1.9}$$

Из разностного уравнения (1.8) можно найти a_k, c_k, b_k, d_k

$$a_k = 1 - \tfrac{h}{2}\,p_k \quad b_k = 2 - h^2\,q_k \quad c_k = 1 + \tfrac{h}{2}\,p_k \quad d_k = h^2\,f_k$$

Явные выражения для a_0, b_0, c_0, d_0 и a_n, b_n, c_n, d_n зависят от способа учета граничных условий. Разве что $a_0 = 0 \wedge c_0 = 0$. Оставим остальные читателям из Беларуси в качестве упражнения.

§ 6. Метод разностной прогонки

Вспомним систему линейную систему 1.9. Её матрица, как видно чуть ниже, трёхдиагональная.

$$\begin{pmatrix} -b_0 & c_0 & & & 0 \\ a_1 & -b_1 & c_1 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & -b_{n-1} & c_{n-1} \\ 0 & & & a_n & -b_n \end{pmatrix}$$

Такие СЛУ можно решать не за $O(n^3)$, а за $O(n)$. Опишем, как именно.

Преобразум систему: уберём поддиагональ, на самой диагонали оставим всюду 1. Тогда можно написать прогоночное соотношение, очень похожее на такое же для дифференциальной прогонки (в § 2).

$$y_k = \alpha_k y_{k+1} + \beta_k \tag{1.1}$$

Свяжем α, β с a, b, c, d.

$$\begin{cases} y_{k-1} = \alpha_{k-1} y_k + \beta_{k-1} \\ a_k y_{k-1} - b_k y_k + c_k y_{k+1} = d_k \end{cases} \Rightarrow \quad \begin{aligned} & a_k\alpha_{k-1} y_k + a_k\beta_{k-1} - b_k y_k + c_k y_{k+1} = d_k \\ \Leftrightarrow\ & y_k = \frac{c_k}{b_k - a_k\alpha_{k-1}} y_{k+1} + \frac{a_k\beta_{k-1} - d_k}{b_k - a_k\alpha_{k-1}} \end{aligned}$$

Отсюда получаются удобные рекурсивные соотношения для α_k, β_k

$$\alpha_k = \frac{c_k}{b_k - a_k\alpha_{k-1}}; \qquad \beta_k = \frac{a_k\beta_{k-1} - d_k}{b_k - a_k\alpha_{k-1}} \tag{1.2}$$

Начальные значения α_k, β_k легко определяются из линейной системы на a_k, b_k, c_k, d_k (1.9).

Итак, алгоритм, следующий

прямая прогонка: решаем прогоночные уравнения (1.2) с начальными данными $\alpha_0 = \frac{c_0}{b_0}$, $\beta_0 = -\frac{d_0}{b_0}$.

обратная прогонка: уже зная α_k, β_k решаем прогоночное соотношение (1.1) с начальными данными $y_n = \beta_n$ ($c_n = 0 \Rightarrow \alpha_n = 0$).

Как видно, что прямая, что обратная прогонка имеют асимптотику $O(n)$, что не может не радовать. Теперь подумаем над корректностью метода.

§ 7. Лемма об оценке для системы разностных уравнений

Утверждение 1 (Достаточное условие). Пусть $\forall k :: a_k, c_k > 0$ ($c_n, a_0 = 0$), и СЛУ имеет диагональное преобладание: $b_k \geqslant a_k + c_k$.

Тогда если $\exists\, k : b_k > a_k + c_k$ прогоночные уравнения 1.2 разрешимы

▶ Проблемы у нас возникнут только если знаменатели обратятся в 0. Учитывая диагональное преобладание, это эквивалентно $\alpha_{k-1} \geqslant 1$.

(!) $0 < \alpha_{k-1} \leqslant 1$ (по индукции)

база: $\alpha_0 = \frac{c_0}{b_0}$, $b_0 \geqslant c_0 > 0$, $c_0 \leqslant b_0$. Кажется, все верно.[1]

переход: Знаем что $\alpha_{k-1} \leqslant 1$. Так что из условий теоремы

$$b_k - \alpha_{k-1} a_k \geqslant b_k - a_k \geqslant c_k > 0 \;\Rightarrow\; 0 < \alpha_k = \frac{c_k}{b_k - a_k \alpha_{k-1}} \leqslant 1$$

Если $\alpha_{k-1} < 1$, то $\alpha_k < 1$, поскольку $b_k - \alpha_{k-1} a_k > b_k - a_k \geqslant c_k$. Это означает, что если уж неравенство стало строгим, оно таким и останется.

Таким образом, все знаменатели кроме последнего > 0. В нём $c_n = 0 \not> 0$.

Разберёмся, что с ним делать. Поскольку мы предположили что диагональное преобладание хоть где-то строгое (в k_0), возможны варианты:

1. $k_0 < n \Rightarrow \alpha_{n-1} < 1$. Тогда $b_n - a_n \alpha_{n-1} > b_n - a_n \geqslant 0$
2. $k_0 = n \Rightarrow \alpha_{n-1} \leqslant 1$. Тогда $b_n - a_n \alpha_{n-1} \geqslant b_n - a_n > 0$

Как видно, даже последний знаменатель $\neq 0$ ◀

Посмотрим, какие условия утверждение выше накладывает на уравнение. Вспомним выражения для $a_k, \ldots$

$$a_k = 1 - \frac{h}{2} p_k \quad b_k = 2 - h^2 q_k \quad c_k = 1 + \frac{h}{2} p_k \quad d_k = h^2 f_k$$

Отсюда

$$\begin{aligned} a_k > 0 &\quad\Leftarrow\quad h < \frac{2}{\max|p_k|} \\ c_k > 0 &\quad\Leftarrow\quad h < \frac{2}{\max|p_k|} \\ b_k \geqslant c_k + a_k &\quad\Leftrightarrow\quad \boxed{q_k \leqslant 0} \end{aligned}$$

Можно ещё подумать про граничные условия. Тут всё зависит от способа вычисления производной на границе.

Пример 1. Оценим производную по простейшей схеме

$$\frac{y_1 - y_0}{h} = \alpha y_0 + A$$

Тогда

$$y_1 + y_0(-1 - \alpha h) = Ah \;\Rightarrow\; \begin{aligned} b_0 &= 1 + \alpha h \\ c_0 &= 1 \end{aligned}$$

И по сути нам нужно $\boxed{\alpha \geqslant 0}$. Аналогично с правой границей, там $\boxed{\beta \leqslant 0}$.

[1] здесь бы и хотелось сразу $\alpha_0 < 1$, да $a_0 = 0$, так что неравенство $c_0 \leqslant b_0$ нестрогое

Если хотя бы одно из трёх условий в рамке строгое, прямая прогонка сработает.

Замечание 1. Рассмотрим однородную задачу с однородным левым граничным условияем ($A = 0$). Тогда $d_k = 0 \Rightarrow \beta_k = 0$ при $k < n$. Прогоночное соотношение примет вид $y_k = \alpha_k y_{k+1}$. Если один из узлов находится вблизи $y(x) = 0$, α_k окажется большим, что не очень хорошо с вычислительной точки зрения. Но, вообще, вероятность этого низкая, и можно просто узлы сдвинуть если что-то сломается.

Теперь видимо то, что в названии

Лемма 1. Пусть выполено условие достаточности прогонки 1.7.1 в усиленном виде:

$$b_k \geqslant a_k + c_k + \delta, \quad \delta > 0.$$

Тогда

$$\max_k |y_k| \leqslant \delta^{-1} \max_k |d_k|$$

▼

Пусть $M = \max |y_k| = y_{k_0}$ (их же конечное число). Тогда из разностной СЛУ (1.9)

$$\begin{aligned} b_{k_0} y_{k_0} &= a_{k_0} y_{k_0-1} + c_{k_0} y_{k_0+1} - d_{k_0} \\ \Rightarrow\ b_{k_0} M &\leqslant a_{k_0} M + c_{k_0} M - d_{k_0} \\ &\Rightarrow\ \delta M \leqslant (b_{k_0} - a_{k_0} - c_{k_0}) M \leqslant \left|d_{k_0}\right| \leqslant \max |d_k| \end{aligned}$$

◀

§ 8. Теорема о сходимости разностного метода

Теорема 1. Рассмотрим краевую задачу III типа для уравнения второго порядка. Пусть

1. $p, q, f \in C^2([a; b])$
2. $q(x) \leqslant -q_0, q_0 > 0$
3. $\alpha < 0, \beta > 0$

Тогда разностный метод сходится:

$$\forall\, k\ ::\ |y(x_k) - y_k| < Ch^2, \quad C = \text{const}$$

☐ Запишем уравнение на сетке:

$$\frac{y(x_{k+1}) - 2y(x_k) + y(x_{k-1})}{h^2} + R_1 + p(x_k)\frac{y(x_{k+1}) - y(x_{k-1})}{2h} + p(x_k) R_2 + q(x_k)\, y(x_k) = f(x_k)$$

Вычтем теперь из него разностное (которое (1.8)), вводя $w_k = y(x_k) - y_k$

$$\frac{w_{k+1} - 2w_k + w_{k-1}}{h^2} + p_k \frac{w_{k+1} - w_{k-1}}{2h} + q_k\, w_k = R, \quad R = -R_1 - p_k R_2$$

Если вспомнить как выглядят оценки R_1 и R_2 то становится понятно зачем нужен первый пункт в условиях теоремы. Но мы вспоминать точный вид не будем, а просто запишем $R = Lh^2$

Чтобы оценить произодную на границе, можно воспользоваться формулой с фиктивными узлами и сдвинутой сеткой (на $h/2$), тогда:

$$\frac{w_1 - w_0}{h} = \alpha \frac{w_0 + w_1}{2} + R_0 \qquad \frac{w_{n+1} - w_n}{h} = \beta \frac{w_{n+1} + w_n}{2} + R_{n+1}$$

Выражения для R снова квадратичны по h, запишем их так: $R_0 = L_0 h^2$, $R_{n+1} = L_{n+1} h^2$.

Запишем выражения для коэффициентов линейной системы с w_k:

$$a_k = \tfrac{1}{h^2} - \tfrac{p_k}{2h} \quad b_k = \tfrac{2}{h^2} - q_k \quad c_k = \tfrac{1}{h^2} + \tfrac{p_k}{2h} \quad d_k = h^2 L$$

(они такие же как и раньше, только правая часть в уравнении поменялась)

Подгоним под условия леммы 1.7.1

$$b_k \geqslant a_k + c_k + \delta \;\Leftrightarrow\; -q_k \geqslant \delta \;\Leftarrow\; \delta = q_0$$

На границах

$$\begin{array}{llll} a_0 = 0 & b_0 = \frac{1}{h} + \frac{\alpha}{2} & c_0 = \frac{1}{h} - \frac{\alpha}{2} & d_0 = h^2 L_0 \\ a_{n+1} = \frac{1}{h} + \frac{\beta}{2} & b_{n+1} = \frac{1}{h} - \frac{\beta}{2} & c_{n+1} = 0 & d_{n+1} = -h^2 L_{n+1} \end{array}$$

Снова найдём δ

$$\begin{array}{lccc} b_0 \geqslant a_0 + c_0 + \delta & \Leftrightarrow \quad \alpha \geqslant \delta & \Leftarrow & \delta = \alpha \\ b_{n+1} \geqslant a_{n+1} + c_{n+1} + \delta & \Leftrightarrow \quad -\beta \geqslant \delta & \Leftarrow & \delta = -\beta \end{array}$$

Выберем:

$$\delta = \min\left\{q_0, \alpha, -\beta\right\} \quad C = \max\left\{|L|, |L_0|, |L_1|\right\}$$

Тогда по лемме

$$\max|w_k| \leqslant \delta^{-1} \max|d_k| \;\Rightarrow\; \max|y(x_k) - y_k| \leqslant \delta^{-1} C h^2$$

■

<+дальше идет каша из всяких обобщений, не буду пока их писать+>

§ 9. Жёсткие системы ОДУ

Будем рассматривать задачу Коши для системы обыкновенных дифференциальных уравнений.

$$\boldsymbol{y}'(x) = f(x, \boldsymbol{y}(x)), \qquad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

Пусть x_n — узлы равномерной сетки с шагом h.

1. Методы численного интегрирования ОДУ

1. Метод Эйлера

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h f(x_n, \boldsymbol{y}_n)$$

2. Неявный метод Эйлера

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h f(x_{n+1}, \boldsymbol{y}_{n+1})$$

оба метода выше имеют ошибку $\sim O(h)$. Поэтому они и называются простейшими.

Рассмотрим более точные методы

3. Улучшенный метод Эйлера

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{h}{2}\Big(f(x_n, \boldsymbol{y}_n) + f\big(x_{n+1}, \boldsymbol{y}_n + h f(x_n, \boldsymbol{y}_n)\big)\Big)$$

4. Метод трапеций

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{h}{2}\Big(f(x_n, \boldsymbol{y}_n) + f(x_{n+1}, \boldsymbol{y}_{n+1})\Big)$$

5. Метод средних прямоугольников

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h f\left(x_n + \tfrac{h}{2}, \tfrac{\boldsymbol{y}_n + \boldsymbol{y}_{n+1}}{2}\right)$$

6. Весовая формула трапеций

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\Big((1-\theta) f(x_n, \boldsymbol{y}_n) + \theta f(x_{n+1}, \boldsymbol{y}_{n+1})\Big), \quad 0 \leqslant \theta \leqslant 1$$

7. Весовая формула прямоугольников

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h f\left(x_n + \theta h, (1-\theta)\,\boldsymbol{y}_n + \theta \boldsymbol{y}_{n+1}\right), \quad 0 \leqslant \theta \leqslant 1$$

Как видно, только первый метод явный, остальные неявные в той или иной степени.

<+можно их точность оценить+>

2. Жёсткие системы

Определение 1. Формального полного определения жёсткости нет.

Замечание 1. Обычно под жёсткими системами понимают следующую ситуацию: пусть в решении системы есть две области

1. «переходный слой» где решение быстро изменяется, как правило небольшой
2. область плавного изменения решения

Проблема исключительно вычислительная — хочется интегрировать переходный слой малым шагом, а область плавного изменения большим. А она может быть довольно большой. И времени у нас не вечность. И вот в процессе перехода от малого шага к большому и возникают некоторые трудности. Для систем эти области ещё могут перекрываться.

Если такие подобные трудности возникают, то система — жёсткая.

Пример 1. Рассмотрим $f(x, y) = Ay$, $y(0) = 1$ (всё одномерное). Его решение — $y = e^{Ax}$.

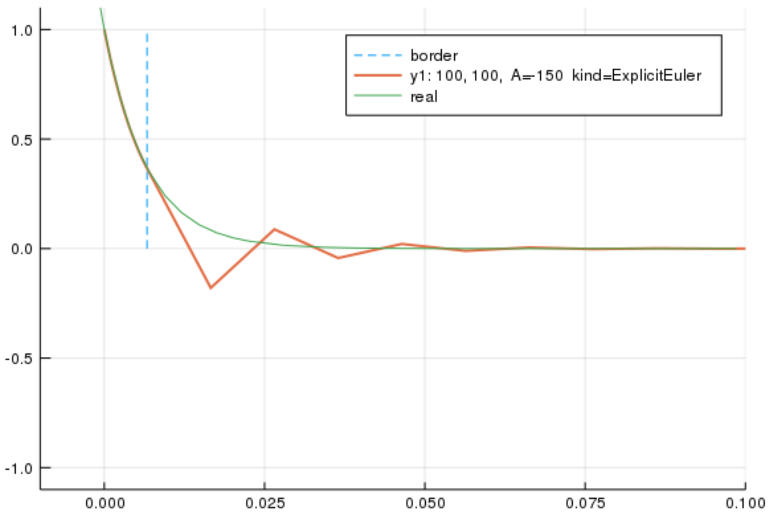
Попробуем решить методом Эйлера

$$y_{n+1} = y_n + hAy_n = y_n(1 + hA) = y_0(1 + hA)^{n+1}$$

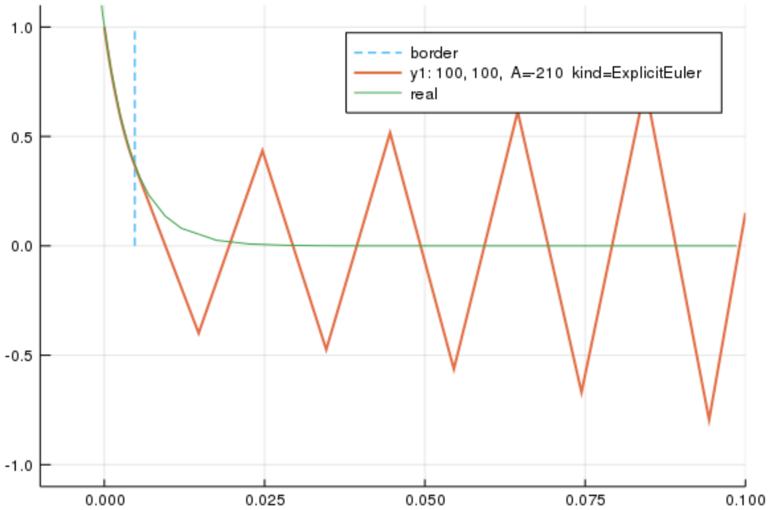
Из определения e при $h \to 0$ получается истинное решение. Однако, рассмотрим что происходит при $A \ll 0$.

1. $A < 0, 1 < |Ah| < 2$. Тогда $-1 < 1 + hA < 0$, $(1 + hA)^n$ меняет знак. При этом численное решение осциллирует, но осцилляции затухают. Истинное решение, как мы помним, убывающая экспонента. С ней такого точно не бывает.

   Выглядит это примерно вот так:



2. $A < 0, |Ah| \geqslant 2$. Тогда $1 + hA \leqslant -1$. Здесь колебания даже не затухают, а вообще растут. Что-то никак не связанное с экспонентой.

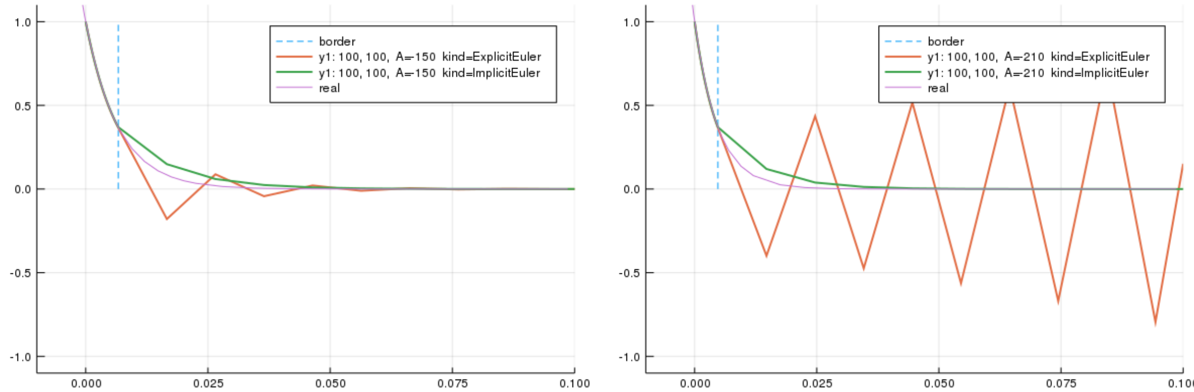


Пример 2. Снова рассмотрим $f(x, y) = Ay$, $y(0) = 1$ (всё одномерное). Его решение — $y = e^{Ax}$. Такую штуку называют пробным уравнением

Попробуем решить неявным методом Эйлера

$$y_{n+1} = y_n + hAy_{n+1} \Rightarrow y_{n+1} = \frac{1}{1 - Ah} y_n = \left(\frac{1}{1 - Ah}\right)^{n+1} y_0$$

И вот здесь никаких проблем с $A < 0$ нету, какое бы оно большое не было. Сравним его с явным методом Эйлера



Пример 3.

$$f(x, \mathbf{y}) = \begin{pmatrix} A & 0 \\ 0 & a \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad -A \gg 1, \ a \sim 1$$

вот такая штука точно жёсткая: на одном и том же отрезке один кусок решения резко изменяется, а другой ведёт себя весьма плавно.

В качестве некоторой попытки формализации рассуждений иногда вводят такое определение:

Определение 2. Пусть $f(x, \mathbf{y}) = \hat{A}\mathbf{y}$, λ_i — собственные числа $\hat{A}$. Тогда если

$$\frac{\max_i \{|\operatorname{Re} \lambda_i|\}}{\min_i \{|\operatorname{Re} \lambda_i|\}} \gg 1$$

систему называют жёсткой.

Определение 3. Рассмотрим одношаговый метод для $y' = Ay$, $R(z)\ :\ \ y_{n+1} = R(Ah)y_n$ — функция устойчивости ⟨?⟩переходный множитель ⟨?⟩ ⟨⚒⟩

Определение 4. Одношаговый метод называется A-устойчивым, если для него $|R(z)| \leqslant 1$ в левой полуплоскости.

Определение 5. $\{z \mid |R(z) \leqslant 1| \leqslant 1\}$ называется областью A-устойчивости метода.

Пример 4. Явный метод Эйлера устойчив в круге $|z + 1| < 1$: для него $R(z) = 1 + z$.

§ 10. Неявные методы Рунге-Кутты

Определение 1. Одношаговый метод называется L-устойчивым, если он A-устойчив и $\lim\limits_{z\to\infty} R(z) = 0$. например, тот же неявный метод Эйлера.

Прежде чем вводить методы Рунге-Кутты, разберёмся с устойчивостью оставшихся методов 3–7

- Улучшенный метод Эйлера
$$R(z) = 1 + z + \frac{z^2}{2}$$
совсем неустойчив
- Метод трапеций/метод средних прямоугольников
$$R(z) = \frac{1 + z/2}{1 - z/2}$$
- Весовая формула
$$R(z) = \frac{1 + (1-\theta)\,z}{1 - \theta\,z}$$
A-устойчива при $\theta \geqslant \frac{1}{2}$. Просто при таком преобразовании прообразом единичного круга будет круг/внешность круга с центром в $\theta - \frac{1}{2}$ и радиусом $\left|\theta - \frac{1}{2}\right|$.

Всё, можно бросаться сеять паслёновые определения

Определение 2 (q-этапный метод РК).
$$k_i = f\left(x_n + \alpha_i h, y_n + h\sum_{j=1}^{i-1}\beta_{ij}k_j\right)$$
$$y_{n+1} = y_n + h\sum_{i=1}^{q}\gamma_i k_i$$
а α_i, β_{ij}, γ_i уже зависят от метода.

Определение 3. *Неявным методом называется* такой вариант метода РК, где у матрицы β_{ij} ненулевые диагональные и/или наддиагональные члены.

Определение 4. *Диагональным неявным методом* называется такой вариант метода РК, где у матрицы β_{ij} ненулевые диагональные члены, а наддиагональные все нулевые.

Закопаемся в варианты реализации этих методов. Все примеры будут иметь такой вид: $\begin{array}{c|c} \alpha & \beta \\ \hline & \gamma \end{array}$
Сначала стоит заметить чем хороши неявные методы РК.

Утверждение 1. Существует реализация неявного метода с $p = 2q$, где p — порядок точности

Собственно, просто берём узлы и коэффициенты гауссовой квадратурной формулы.
Ещё стоит отметить что неявные методы A-устойчивы, а вот диагональные как повезёт.

Пример 1 (обычный `rk4`).
$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & \frac{1}{2} & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Пример 2. (основанная на методе Гаусса)

$$
\begin{array}{c|cc}
\frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\
\frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}-\frac{\sqrt{3}}{6} & \frac{1}{4} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
$$

Пример 3. (дигонального неявного РК)

$$
\begin{array}{c|cc}
\gamma & \gamma & 0 \\
1-\gamma & 1-2\gamma & \gamma \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
\qquad \gamma = \frac{1}{2}+\frac{\sqrt{6}}{3}
$$

<+ещё примеры?+>

2 Методы линейной алгебры

§ 1. Устойчивость собственных чисел при возмущении матрицы

Пусть A — линейный оператор $\mathbb{R}^s \to \mathbb{R}^s$, x, b — векторы-столбцы в $\mathbb{R}^s$. Здесь будет столько матриц и векторов, что рисовать шляпы не будем, и так понятно кто есть кто.

Какие задачи вообще можно здесь решать

1. Решение линейной системы $Ax = b$
2. Поиск собственных чисел $Ax = \lambda x$

Какие при этом могут возникнуть ошибки

1. Ошибки округления (алгоритма)
2. Ошибки начальных данных (неустранимые)

Посмотрим, как оценить ошибки вычисления. Пусть $\circ$ — какая-то операция, а $\circledcirc$ — её машинное представление. Существуют два подхода

1. *Прямой анализ ошибок*

   Просто учитываем погрешность $a \circ b$ как ошибку округления. Часто делают так (ε_M — «машинный эпсилон»):
   $$a \circledcirc b = a \circ b(1+\varepsilon), \quad \varepsilon \leqslant \varepsilon_M$$
2. *Обратный анализ ошибок*(метод эквивалентных возмущений)

   Сводим все ошибки к возмущениям начальных данных:
   $$a \circledcirc b = \tilde{a} \circ \tilde{b}, \quad \tilde{a} = a + \Delta a, \ \tilde{b} = b + \Delta b.$$
   (a) оцениваем эквивалентные возмущения

   (b) оцениваем влияние возмущений

   получается, что мы все ошибки записали в неустранимые погрешности начальных данных

Первый метод частно выдает неправданно большие оценки погрешности, так что займёмся в основном вторым.

Разберёмся с корректностью задач.

1. Решение ЛСУ

Определение 1 (мера обусловленности). $\mu = \|A\| \, \|A^{-1}\|$

Почему она так выглядит? Посмотрим какие вообще есть способы оценки вырожденности A

1. $\det A$. Почти не бывает равным 0. К тому же, перемешивает большие и маленькие собственные числа.
2. $\dfrac{\|Ax\|}{\|x\|}$. Здесь мы пытаемся смотреть на ЛЗ строчек матрицы. Но не очень понятно с чем сравнить, чтобы понять близость к ЛЗ. Может быть компоненты матрицы маленькие.

3. $\dfrac{\max \frac{\|Ax\|}{\|x\|}}{\min \frac{\|Ax\|}{\|x\|}}$ уже выглядит разумно. Преобразуем, используя определение нормы (конечномерного) оператора

$$\max \frac{\|Ax\|}{\|x\|} = \|A\|$$
$$\min \frac{\|Ax\|}{\|x\|} = \min \frac{\|y\|}{\|A^{-1}y\|} = \frac{1}{\frac{\|A^{-1}y\|}{\|y\|}} = \left\|A^{-1}\right\|^{-1}.$$

А это очень похоже на определение выше.

Лемма 1. $\|B\| < 1 \;\Rightarrow\; \exists\,(I-B)^{-1} \wedge \left\|(I-B)^{-1}\right\| \leqslant \dfrac{1}{1-\|B\|}$

▼

Рассмотрим систему $x - Bx = y$. Будем искать решение методом простой итерации: $x_{n+1} = f(x_n) = Bx_n + y$. Покажем, что он сходится. Для этого нужно убедиться что f — сжимающее отображение.

$$\|f(x) - f(x')\| = \|B(x-x')\| \leqslant \|B\|\,\|x-x'\| < \|x-x'\|$$

Решение нашлось $\forall\, y \;\Rightarrow\; \exists\,(I-B)^{-1}$. Теперь получим оценку нормы

$$\forall\, x \;::\; x = Bx + y \;\Rightarrow\; \|x\| \leqslant \|B\|\,\|x\| + \|y\| \;\Rightarrow\; \|x\| \leqslant \frac{1}{1-\|B\|}\,\|y\|$$

Тогда это верно и для $\max \|x\| / \|y\| = \left\|(I-B)^{-1}\right\|$ ◀

Теперь, оценим, наконец, погрешность решения СЛУ.

Теорема 1. Рассмотрим возмущенную задачу: $\widetilde{A}x = \tilde{b}$. Введём относительную и абсолютную погрешность A, x, b:

$$\Delta A = \widetilde{A} - A, \quad \Delta x = x - x^*, \qquad \Delta b = \tilde{b} - b$$
$$\delta_A = \frac{\|\Delta A\|}{\|A\|}, \quad \delta_x = \frac{\|\Delta x\|}{\|x*\|}, \qquad \delta_b = \frac{\|\Delta b\|}{\|b\|}$$

x^* — невозмущенное решение

Тогда

$$\delta_x \leqslant \frac{\mu(A)}{1-\mu(A)\delta_A}\,(\delta_A + \delta_b)$$

□ Раз x^* — решение $Ax^* = b$, провернём пару скучных манёвров

$$A'x = b' \;\Leftrightarrow\; (A+\Delta A)(x^* + \Delta x) = b + \Delta b \;\Leftrightarrow\; (A+\Delta A)\Delta x = -\Delta A x^* + \Delta b$$
$$\Leftrightarrow\; \left(I - (-A^{-1}\Delta A)\right)\frac{\Delta x}{x^*} = -A^{-1}\Delta A + A^{-1}\frac{\Delta b}{x^*}$$

Из леммы выше

$$\left\|\frac{\Delta x}{x^*}\right\| \leqslant \frac{1}{1-\|A^{-1}\|\,\|\Delta A\|}\left(\|A^{-1}\|\,\|\Delta A\| + \|A^{-1}\|\left\|\frac{\Delta b}{x^*}\right\|\right)$$

Из невозмущённой системы $\|x^*\| \geqslant \|A\|^{-1}\,\|b\|$, вспомнив определение числа обусловленности осознаем $\|A^{-1}\|\,\|\Delta A\| = \mu(A)\,\delta_A$. Осталось переписать остальное через δ и получить утверждение теоремы. ■

Замечание 1. Из этой теоремы можно прикинуть ошибку решения ЛСУ. Будем, как и обещали, использовать обратный анализ ошибок. Из-за неточного представления в памяти $\delta_A, \delta_b \sim \varepsilon_M$ (ну никак не меньше), так что $\delta_x \sim C(s)\,\mu(A)\,\varepsilon_M$, $C(s)$ — функция параметров задачи.

Замечание 2. На оценку погрешности ещё влияют индивидуальные особенности методов. Например, в методе исключения Гаусса часто накапливается ошибка из-за деления на маленькие ведущие элементы.

2. Поиск собственных чисел

Некий полезный набор фактов из линейной алгебры, который совсем не стоит забывать

1. $Au - \lambda u$ — уравнение на собственные числа и собственные вектора.
2. $p_A(t) = \det(A - tI)$ — характеристический многочлен.
3. матрицы можно приводить к ЖНФ
4. ЖНФ — диагональ из жордановых клеток:
$$J_p(a) = \begin{pmatrix} a & 1 & \\ & \ddots & 1 \\ & & a \end{pmatrix} : \ p \times p, \qquad p_{J_p(a)}(t) = (a-t)^p$$
5. алгебраическая кратность собственного числа — кратность его как корня характеристического многочлена. Совпадает с размерностью корневого подпространства ($V(\lambda)$).
6. геометрическая кратность — размерность собственного подпространства (V_λ).
7. геометрическая кратность $\leqslant$ алгебраической, ибо $\dim V_\lambda \leqslant V(\lambda)$.
8. собственные числа самосопряженных операторов вещественные.
9. из собственных векторов самосопряжённого оператора можно собрать ортогональный базис.

Утверждение 1. Для самосопряжённого положительно определённого оператора
$$\max \lambda_A = \max \frac{(Au, u)}{(u, u)}, \quad \min \lambda_A = \min \frac{(Au, u)}{(u, u)}$$

▶ Например, через теорему об условном экстремуме ◀

Утверждение 2. $\|A\|_2 = \sqrt{\max \lambda_{A^*A}}$

▶ Эвклидова норма согласована со скалярным произведением, так что
$$\max \frac{\|Ax\|_2}{\|x\|_2} = \max \frac{(Ax, Ax)}{(x, x)} = \max \frac{(A^*Ax, x)}{(x, x)}.$$
А дальше можно глянуть утверждение выше. ◀

В принципе это сработает для любой нормы, согласованной со скалярным произведением. А если это не так, то мы уже явно переусложнили себе жизнб для линейной алгебры.

Теперь наконец обсудим устойчивость

Пример 1. Пусть
$$A = J_p(a), \quad \varepsilon B : \ (\varepsilon B)_{ij} = \delta_{ip}\delta_{j1}, \quad \widetilde{A} = A + \varepsilon B$$
Оценим ошибку собственного числа. Попреобразовываем систему..
$$Ax + \varepsilon Bx = \lambda x \ \Leftrightarrow \ \begin{cases} ax_k + x_{k+1} = \lambda x_k, & k \in 1 \mathinner{..} p-1 \\ \varepsilon x_1 + ax_p = \lambda x_p \end{cases} \ \Rightarrow \ \varepsilon x_1 = (\lambda - a)^p x_1$$
В итоге получается, что $\lambda = a + \varepsilon^{1/s}$[1]

Пусть $\varepsilon = 10^{-16}$ (удвоенная точность). Тогда уже на матрицах порядка 15 ошибка ~ 0.1. Грустная оценка получилась.

[1] можно конечно корень из 1 в $\mathbb{C}$ посчитать, но идея не изменится

§ 2. Теорема Бауэра-Файка

ситуация немного лучше, когда матрицы симметричные. Можно придумать не такие грустные оценки, как в примере в предыдущем параграфе.

Теорема 1. Пусть A — диагонализуемая матрица, $D^{-1}AD = \Lambda$. Тогда

$$\lambda_{A+B} - \text{с.ч. } A + B \;\Rightarrow\; \exists\, \lambda_A \;:\; |\lambda_{A+B} - \lambda_A| \leqslant \mu(D)\,\|B\|$$

☐ Построим отрицание

$$\forall\, \lambda_A \;::\; |z - \lambda_A| > \mu(D)\,\|B\| \;\Rightarrow\; z - \text{не с.ч. } A + B$$

и будем его доказывать. Пусть $|z - \lambda_A| > \|B\|$. (!) $A + B - zI$ — неособая

$$A - zI + B = D^{-1}\left(\Lambda - zI + DBD^{-1}\right)D = D^{-1}\left(\Lambda - zI\right)\big(I + \underbrace{(\Lambda - zI)^{-1}DBD^{-1}}_{C}\big)\,D$$

D, $(\Lambda - zI)$ неособые по условию. Воспользуемся леммой об обратимости (2.1.1). Для этого нужно $\|C\| < 1$:

$$\|C\| \leqslant \|(\Lambda - zI)^{-1}\|\;\underbrace{\|D^{-1}\|\,\|B\|\,\|D\|}_{\mu(D)\|B\|}$$

Из утверждения в предыдущем параграфе, и отрицания к предположению теоремы

$$\|(\Lambda - zI)^{-1}\| = \sqrt{\max_k |\lambda_k - z|^{-2}} = \sqrt{\frac{1}{\min_k |\lambda_k - z|^2}} < \frac{1}{\mu(D)\,\|B\|}$$

Как видно, у нас как раз получилось что $\|C\| < 1$. А тогда и матрица выше обратима. ■

Следствие 1. Для самосопряженных матриц

$$\lambda_{A+B} - \text{с.ч. } A + B \;\Rightarrow\; \exists\, \lambda_A \;:\; |\lambda_{A+B} - \lambda_A| \leqslant \|B\|$$

Для них просто D — унитарная, $\|D\| = \|D^{-1}\| = 1$.

Минутка троллинга: что будет, если сразу выбрать базис, где A диагональная. Куда число обусловленности подевалось?

По сути мы сейчас доказали что собственные числа устойчивы к возмущениям матрицы. А вот что там с собственными векторами?

§ 3. Устойчивость собственных векторов при возмущении матрицы

Сразу поясним, какие вообще возникнут проблемы

Пример 1. Пусть A_1, A_2 имеют разные с.ч. и с.в, а

$$C = \begin{cases} I + \varepsilon A_1, & \varepsilon \geqslant 0 \\ I + \varepsilon A_2, & \varepsilon < 0 \end{cases}.$$

Тогда, как видно

$$\lambda_C = \begin{cases} 1 + \varepsilon\lambda_1, & \varepsilon \geqslant 0 \\ 1 + \varepsilon\lambda_2, & \varepsilon < 0 \end{cases}, \quad u_C = \begin{cases} u_1, & \varepsilon \geqslant 0 \\ u_2, & \varepsilon < 0 \end{cases}$$

и в u никакого ε нету. Направление у них изменяется скачком при проходе через 0. А вот с λ всё хорошо.

Проблемы, как видно, возникают в окрестности кратных собственных чисел, снятие вырождения радикально меняет собственные подпространства. Давайте не делать кратных собственных чисел. Может быть так всё будет хорошо?

Утверждение 1. Пусть (λ_i, u_i) — собственные числа и векторы A, (μ_i, v_i) — A^*, все λ_i разные. Короче говоря, A диагонализуема, но может быть не самосопряжённой. Рассмотрим возмущенную задачу на собственные числа и векторы: $(A + \Delta A)\, x = (\lambda + \Delta\lambda)\, x$.

Тогда в линейном приближении[1]

- $p_i = \dfrac{\|u_i\|\,\|v_i\|}{(u_i, v_i)}$

1. $\|\Delta\lambda_i\| \leqslant p_i\,\|\Delta A\|$
2. $\delta u_i \leqslant \displaystyle\sum_{k=1}^{s} \frac{p_k}{|\lambda_i - \lambda_k|}\,\|\Delta A\|$

▶ Пойдём по порядку.

1. $\lambda_i = \overline{\mu_i}$
2. $(u_i, v_k) = 0$ при $i \neq k$

$$\begin{matrix} (Au_i, v_k) = \lambda_i(u_i, v_k) \\ (u_i, A^* v_k) = \overline{\mu_k}(u_i, v_k) \end{matrix} \;\Rightarrow\; (\lambda_i - \overline{\mu_k})\,(u_i, v_k) = 0 \;\Rightarrow\; (\lambda_i - \lambda_k)\,(u_i, v_k) = 0$$

3. $(Ax, v_i) = \overline{\mu_i}(x, v_i) = \lambda_i(x, v_i)$
4. $\Delta\lambda_i(u_i, v_i) = (\Delta A u_i, v_i)$

$$(\widetilde{A}\tilde{u}_i, v_i) = (\tilde{\lambda}_i \tilde{u}_i, v_i) \;\Rightarrow\; (\Delta A u_i, v_i) + (A\,\overbrace{\Delta u_i, v_i) = \Delta\lambda_i(u_i, v_i) + \lambda_i(\Delta u_i}^{=}, v_i)$$

отсюда уже легко вывести первый пункт.

5. $(\Delta u_i, v_k) = (1 - \delta_{ik})\,\dfrac{(\Delta A u_i, v_k)}{(\lambda_i, \lambda_k)}$, аналогично предыдущему пункту. Здесь выбрали $(\Delta u_i, v_i) = 0$ пожертвовав нормированностью v_i. Всё равно одна лишняя степень свободы была.
6. $\Delta u_i = \displaystyle\sum_{k=1}^{s} \gamma_k u_k$ [2], $\gamma_k = \dfrac{(\Delta u_i, v_k)}{(u_k, v_k)}$
7. $\Delta u_i = \displaystyle\sum_{k=1}^{s} \frac{(\Delta A u_i, v_k)}{(\lambda_i - \lambda_k)}\, \frac{1}{(u_i, v_k)}$, отсюда очевиден второй

◀

§ 4. Степенной метод

Определение 1 (Степенной метод)**.** Пусть A — диагонализуемая матрица порядка s. Построим итерации такого сорта, x_0 выбирается случайно.

$$\tilde{x}_{n+1} = Ax_n, \quad x_{n+1} = \frac{\tilde{x}_{n+1}}{\tilde{x}^1_{n+1}} \quad \text{(делим на первую компоненту)}$$

Будем брать $\tilde{x}^1_{n+1}$ как оценку наибольшего собственного числа A

Мотивировка у такого определения понятная — если x_n разложить по собственным векторам A, то через много шагов наибольшее собственное число забьет все остальные. Однако, нужно аккуратно сформулировать условия сходимости.

Утверждение 1. Пусть для собственных чисел A выполнено условие:

$$|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_s|$$

Тогда степенной метод сходится к λ_1

[1] а если нет, то надо думать

[2] u_i образуют базис, раз матрица диагонализуема; корневые подпространства совпадают с собственными

▶ Из определения степенного метода

$$x_n = \frac{\tilde{x}_n}{\tilde{x}_n^1} = \frac{Ax_{n-1}}{\left\{Ax_{n-1}\right\}^1} = \frac{\left(\tilde{x}_{n-1}^1\right)^{-1} A\tilde{x}_{n-1}}{\left(\tilde{x}_{n-1}^1\right)^{-1} \left\{A\tilde{x}_{n-1}\right\}^1} = \dots = \frac{A^n x_0}{\left\{A^n x_0\right\}^1}$$

Разложим x_0 по собственным векторам A, тогда

$$x_0 = \sum_{k=1}^{s} c_k u_k \;\Rightarrow\; A^n x_0 = c_1 \lambda_1^n u_1 + \sum_{k=2}^{s} c_k \lambda_k^n u_k$$

Отсюда переходим к пределу

$$x_n = \frac{A^n x_0}{\left\{A^n x_0\right\}^1} = \frac{c_1\lambda_1^n u_1 + \sum_{k=2}^{s} c_k \lambda_k^n u_k}{c_1\lambda_1^n u_1^1 + \sum_{k=2}^{s} c_k \lambda_k^n u_k^1} = \frac{\frac{u_1}{u_1^1} + \sum_{k=2}^{s} c_k' \left(\frac{\lambda_k}{\lambda_1}\right)^n u_k'}{1 + \sum_{k=2}^{s} c_k' \left(\frac{\lambda_k}{\lambda_1}\right)^n {u_k'}^1} \xrightarrow[n\to\infty]{} \frac{u_1}{u_1^1}$$

$$\Rightarrow\; \tilde{x}_{n+1}^1 = \left\{Ax_n\right\}^1 \xrightarrow[n\to\infty]{} \frac{\left\{\lambda_1 u_1\right\}^1}{u_1^1} = \lambda_1$$

Так сработает для комплексных λ, поскольку

$$\lim_{x\to\infty} \left|\left(\frac{\lambda_k}{\lambda_1}\right)^n - 0\right| = \lim_{x\to\infty} \left(\frac{|\lambda_k|}{|\lambda_1|}\right)^n = 0$$

◀

Посмотрим, что будет если нарушить условия утверждения выше

Пример 1. $\lambda_1 = \lambda_2 = \lambda$: ну это одно и тоже число, так неинтересно

Пример 2. $\lambda_1 = -\lambda_2 = \lambda$

$$x_{2n} \to \tfrac{c_1u_1 + c_2u_2}{c_1u_1^1 + c_2u_2^1}, \quad \tilde{x}_{2n+1}^1 \to \lambda \tfrac{c_1u_1^1 - c_2u_2^1}{c_1u_1^1 + c_2u_2^1}$$
$$x_{2n+1} \to \tfrac{c_1u_1 - c_2u_2}{c_1u_1^1 - c_2u_2^1}, \quad \tilde{x}_{2n+2}^1 \to \lambda \tfrac{c_1u_1^1 + c_2u_2^1}{c_1u_1^1 + c_2u_2^1}$$

подпоследовательности сходятся к разным числам

Пример 3. $\lambda_1 = Re^{i\theta}$, $\lambda_2 = Re^{-i\theta}$, раз матрица вещественная $c_2 = \overline{c_1}$, $u_2 = \overline{u_1}$

$$x_n \to \tfrac{2\,\mathrm{Re}\left(c_1u_1e^{in\theta}\right)}{2\,\mathrm{Re}\left(c_1u_1^1e^{in\theta}\right)}, \quad \tilde{x}_{n+1}^1 \to R\tfrac{\mathrm{Re}\left(c_1u_1^1e^{i(n+1)\theta}\right)}{\mathrm{Re}\left(c_1u_1^1e^{in\theta}\right)}$$

кажется это вообще никуда не сходится.

Для недиагонализуемых может сходиться, но медленно. `<+пример с матрицей-производной+>`

Определение 2 (Степенной метод со сдвигом). Рассмотрим в степенном методе матрицу $A - tI$ вместо A. При этом ищется наиболее удалённое по модулю от t собственное число.

какой-то не очень полезный метод

§ 5. Обратный степенной метод

Определение 1 (Обратный степенной метод). Пусть матрица A — неособая. Будем применять степенной метод для A^{-1}. Решим, что то, что нашлось — наименьшее по модулю собственное число.

Утверждение 1. Пусть для собственных чисел A выполнено условие:

$$|\lambda_1| < |\lambda_2| \leqslant \dots \leqslant |\lambda_s|$$

Тогда обратный степенной метод сходится к λ_1^{-1}

▶

$$A^{-1}x = \lambda_* x \;\Leftrightarrow\; \lambda x = \lambda_*^{-1} x = Ax$$
$$|\lambda_1| < |\lambda_2| \leqslant \dots \leqslant |\lambda_s| \;\Leftrightarrow\; |\lambda_1^{-1}| > |\lambda_2^{-1}| \geqslant \dots \geqslant |\lambda_s^{-1}|$$

◀

Определение 2 (Обратный степенной метод со сдвигом). Рассмотрим в обратном степенном методе матрицу $A - tI$ вместо A. Метод при этом будет искать ближайшее к t собственное число

вот это уже полезно. Можно взять грубую оценку с.ч. и уточнить её таким методом.

Определение 3 (Обратный степенной метод с переменным сдвигом). Возьмём обратный степенной метод и слегка изменим шаг итерации. Помимо махинаций с $\tilde{x}_{n+1}$,

$$t_{n+1} = t_n + \mu_n^{-1}, \qquad \mu_n = \tilde{x}_{n+1}^1$$

Метод при этом будет искать ближайшее к t_0 собственное число

Доказывать что такой алгоритм сходится мы не будем. Зато можно понять почему он так устроен. Поскольку $\tilde{x}_{n+1}$ — текущее приближение собственного вектора

$$\tilde{x}_{n+1} = (A - tI)^{-1} x_n \;\Leftrightarrow\; (A - tI)\tilde{x}_{n+1} = x_n \;\Leftrightarrow\; \tilde{x}_{n+1} = (\lambda_A - t)^{-1} x_n$$

На каждом шаге$x_n^1 = 1$, так что

$$\mu = \tilde{x}_{n+1}^1 = (\lambda_A - t)^{-1} \;\Leftrightarrow\; \lambda_A = t + \mu^{-1}$$

§ 6. Двумерные вращения

Будем рассматривать матрицы самосопряженных операторов. На всякий случай, снова приведём набор полезных фактов из линейной алгебры.

1. Унитарные операторы — такие, что сохраняют скалярное произведение:

$$U \;:\; (Ux, Uy) = (x, y)$$

2. $U^* U = 1 \;\Rightarrow\; U^{-1} = U^*$
3. Если A — самосопряженный, $\exists\, U \;:\; A = U^{-1} \Lambda U$, Λ тут диагональная.
4. Произведение унитарных операторов — унитарный оператор

Вернёмся на скучную вещественную прямую. Матрицы операторов сменили названия

унитарные	→	ортогональные
эрмитовы	→	симметричные

Все утверждения выше спокойно сохранились. Запишем ещё пару специфических для $\mathbb{R}^n$ фактов

5. Существует базис, в котором матрица ортогонального оператора — диагональ из блоков такого сорта:

   тождество: $\boxed{1}$

   отражение: $\boxed{-1}$

   вращение: $\boxed{\begin{matrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{matrix}}$

6. Если вся ортогональная матрица единичная кроме одного одинокого грустного блока, то она простое отражение/вращение.

Простые вращения ещё называются двумерными вращениями. Все потому, что блоки такого сорта соответствуют двумерным инвариантным подпространстрам.

Будем потихоньку приводить матрицу A к (по возможности) диагональному виду. Посмотрим, как выглядит один шаг такого приведения

$$A \to C = O^T A O, \quad O \text{ — ортогональная}$$

Если продраться через паслёновые заросли (для краткости переобозначив cos, sin), получится явное выражение для компонент C

$$C: \quad \begin{aligned} c_{ij} &= a_{ij}, & i,j \neq p,q \\ c_{pj} = c_{jp} &= c\,a_{pj} + s\,a_{qj}, & j \neq p,q \\ c_{qj} = c_{jq} &= -s\,a_{pj} + c\,a_{qj}, & j \neq p,q \\ c_{pp} &= a_{pp}\,c^2 + 2\,a_{pq}\,cs + a_{qq}s^2 & \\ c_{pq} = c_{qp} &= (a_{qq} - a_{pp})\,cs + a_{pq}\,(c^2 - s^2) & \\ c_{qq} &= a_{qq}\,c^2 - 2\,a_{pq}\,cs + a_{pp}s^2 & \end{aligned} \tag{2.1}$$

Как можно избавиться от внедиагональных членов:

1. $c_{p-1,q} = 0$: вращение Гивенса

$$\begin{matrix} c_{p-1,q} = -s\,a_{p,p-1} + c\,a_{q,p-1}, \\ c = \cos\varphi, s = \sin\varphi \end{matrix} \Rightarrow \cos\varphi = \frac{a_{p-1,p}}{\sqrt{a^2_{p-1,p} + a^2_{p-1,q}}}; \quad \sin\varphi = \frac{a_{p-1,q}}{\sqrt{a^2_{p-1,p} + a^2_{p-1,q}}}$$

2. $c_{p,q} = 0$: вращение Якоби

$$\begin{matrix} c_{p,q} = (a_{qq} - a_{pp})\,cs + a_{pq}\,(c^2 - s^2) \\ c = \cos\varphi, s = \sin\varphi \end{matrix} \Rightarrow \tan\varphi = \frac{2a_{pq}}{a^2_{qq} - a^2_{pp}}$$

§ 7. Лемма о правиле знаков при исключении

Вспомним пару фактов из линейной алгебры:

Определение 1. Пусть $B(x,y)$ — симметрическая билинейная функция. Тогда функция одного аргумента $B(x,x)$ называется квадратичной формой.

1. $\forall B(x,x)\ \exists A\ :\ (Ax,x) = B(x,x)$, A — самосопряженный. Это означает, что можно записать матрицу квадратичной формы и она симметрична.
2. *Закон инерции*: если привести матрицу квадратичной формы к диагональному виду, то количество элементов одного знака не зависит от способа приведения.

Теперь можно и лемму сформулировать.

Лемма 1. Пусть A — симметричная матрица. Тогда число ведущих элементов одного в методе исключения Гаусса для такой матрицы совпадает с числом собственных чисел того же знака.

честно говоря ушло немало времени чтобы понять что формулировка именно такая.

▼
Рассотрим квадратичную форму (Ax,x). Напишем её в координатах для экономии чернил:

$$(Ax,x) = \sum_{ij} a_{ij}x_i x_j = a_{11}x_1^2 + 2\sum_i a_{1j}x_1x_j + \sum_{i,j\geqslant 2} a_{ij}x_i.x_j$$

Будем приводить её к сумме квадратов стандартным способом (Лежандра).

$$(Ax,x) = a_{11}^{-1}\underbrace{\left(\sum_j a_{1j}x_j\right)^2}_{\xi_1^2} + \sum_{i,j\geqslant 2}\underbrace{\left(a_{ij} - \tfrac{a_{1i}a_{1j}}{a_{11}}\right)}_{a'_{ij}} x_i x_j$$

Внимательно присмотримся к a'_{ij}. Мы вычитаем из элемента строки такой же элемент первой строки, поделённый на первый элемент первой строки, умноженный на первый элемент данной строки. А это как раз шаг метода исключения Гаусса.

Теперь приведём A к ЖНФ. Поскольку она симметричная, J_A диагональная. На диагонали стоят собственные числа. Сравнивая их с a_{jj} и припоминая закон инерции, приходим к утверждению леммы. ◀

§ 8. Метод Гивенса

Вспомним, как выглядело вращение Гивенса

$$\begin{array}{c} c_{p-1,q} = -s\,a_{p,p-1} + c\,a_{q,p-1} = 0, \\ c = \cos\varphi, s = \sin\varphi \end{array} \Rightarrow \cos\varphi = \frac{a_{p-1,p}}{\sqrt{a_{p-1,p}^2 + a_{p-1,q}^2}}; \quad \sin\varphi = \frac{a_{p-1,q}}{\sqrt{a_{p-1,p}^2 + a_{p-1,q}^2}}$$

Будем строить повороты с (p,q) в таком порядке, как на картинке ниже

$$\begin{tikzcd} (2,3) \arrow[r] & (2,4) \arrow[r] & (2,5) \arrow[r] & \cdots \arrow[r] & (2,s) \arrow[dllll, rounded corners] \\ & (3,4) \arrow[r] & (3,5) \arrow[r] & \cdots \arrow[r] & (3,s) \arrow[dlll, rounded corners] \\ & & (4,5) \arrow[r] & \cdots & \end{tikzcd}$$

На каждом шаге метода столбцы/строки с индексами p и q заменяются их линейными комбинациями. При этом явно зануляется $(p-1,q)$ элемент. А после предыдущих шагов $\forall j > 1$ $(p-j,q)$, $(p-j,p)$ уже нули. Либо $p = 2$ и выше просто ничего нет. Так что и линейные комбинации «верхушек» столбцов будут нулями, и ничего испортиться не сможет. Про область нулей под диагональю можно особо не думать, она получится автоматически, так как матрица симметричная на каждом шаге.

После вращений Гивенса матрица стала трёхдиагональной.

$$(A - tI) = \begin{pmatrix} a_1 - t & b_1 & & \\ b_1 & a_2 - t & \ddots & \\ & & \ddots & b_{s-1} \\ & & b_{s-1} & a_s - t \end{pmatrix}$$

Введём p_k — угловые миноры. порядка k Из формулы разложения определителя по строке получаются рекуррентные формулы для p

$$\begin{aligned} &p_1(t) = a_1 - t \\ &p_2(t) = (a_2 - t)p_1(t) - b_1^2 \\ &p_k(t) = (a_k - t)p_{k-1}(t) - b_{k-1}^2 p_{k-2}(t) \end{aligned}$$

Как нетрудно заметить, последовательность $p_k(t)$ для фиксированного t — это тоже самое что и a_{kk} в лемме в предыдущем параграфе (2.7.1). Ну ведь правда, после k шагов метода Гаусса на диагонали вплоть до k строки стоят 1. Так что угловой минор просто равен a_{kk}.

Разберёмся теперь как искать собственные числа.

1. Как корни характеристического многочлена, $\chi(t) = p_s(t)$

2. Методом бисекции

   Про этот пункт придётся написать чуть подробнее. Выберем какие-то 2 начальных приближения λ, чтобы искать его между ними. Будем считать число перемен знака в последовательности a_{kk}. Нам нужно добиться чтобы перемена знака была всегда одна. Обычным методом половинного деления как раз можно к этому прийти.

§ 9. Метод Якоби

Вспомним, как выглядело вращение Якоби

$$\begin{matrix} c_{p,q} = (a_{qq} - a_{pp})\,cs + a_{pq}\,(c^2 - s^2) \\ c = \cos\varphi, s = \sin\varphi \end{matrix} \Rightarrow \tan\varphi = \frac{2a_{pq}}{a_{qq}^2 - a_{pp}^2}$$

Будем пытаться прийти к почти диагональной матрицей. Для этого надо как-то измерять «недиагональность». Введём набор величин

- $N^2(A) = \sum_{i,k} a_{ik}^2 = \mathrm{Tr}(A^2)$
- $d^2(A) = \sum_{i,k} a_{ii}^2$
- $t^2(A) = \sum_{i\neq k} a_{ik}^2 = N^2(A) - d^2(A)$

Утверждение 1. После одного двумерного вращения ($C = O_{pq}^T A O_{pq}$)

$$t^2(C) = t^2(A) - 2a_{p,q} + 2c_{p,q}^2$$

▶ Пойдем по порядку

1. $N^2(C) = N^2(A)$, поскольку $C^2 = O^T AO\,O^T AO = O^T AO$, а след подобных матриц совпадает.
2. $t^2(C) = N^2(C) - d^2(C) = t^2(A) + d^2(A) - d^2(C)$
3. $d^2(A) - d^2(C) = a_{p,p}^2 + a_{q,q}^2 - c_{p,p}^2 - c_{q,q}^2$, просто все остальные элементы на диагонали не поменялись.
4. $a_{p,p}^2 + a_{q,q}^2 + 2a_{p,q}^2 = c_{p,p}^2 + c_{q,q}^2 + 2c_{p,q}^2$.

   Это можно либо явно проверить из формулы (2.1), либо вспомнить что вращения квадраты норм матриц не изменяют, а эти 4 элемента преобразуются независимо от других. Разве что мы норму оператора не так определяли.

◀

Метод Якоби как раз зануляет c_{pq}^2 на шаге, оптимально уменьшая таким образом t^2. Разберёмся как выбирать здесь p и q.

1. Классический метод Якоби: $p, q\ :\ |a_{p,q}| = \max\limits_{i\neq k} |a_{i,k}|$.

   Оценим, как быстро он сходится

   $$a_{p,q}^2\,\tfrac{s(s-1)}{2} \geqslant \sum_{i,k} a_{i,k}^2 \;\Rightarrow\; t^2(C) \leqslant \left(1 - \tfrac{2}{s(s-1)}\right) t^2(A)$$

   Сходится, конечно, неплохо, но поиск максимума $\sim O(s^2)$, а сам метод Якоби $\sim O(s)$. Неловко выходит.

2. Циклический метод Якоби: просто проходим по всем наддиагональным элементам много раз. Занулять почти нули приятно, но немного бесполезно…
3. Циклический метод Якоби с барьером: выбираем $\varepsilon_i > 0$ и зануляем всё что больше него. Потом выбираем $\varepsilon_{i+1} < \varepsilon_i$ и повторяем.

Разберёмся, как искать собственные числа и собственные векторы.

1. С λ всё просто — они на диагонали матрицы. Корректность следует из теоремы Бауэра-Файка (2.2.1), просто вычтем внедиагональные члены
2. в качестве собственных векторов можно просто взять строки матрицы произведения всех двумерных вращений.

   Это сработает, поскольку собственные векторы диагональной формы — e_k,

   $$\lambda_k e_k = \Lambda e_k = OAO^T e_k \;\Leftrightarrow\; A\,O^T e_k = \lambda_k O^T e_k,$$

   а $O^T e_k$ как раз k-ая строчка O.[1]

[1] то ли мне спать пора, то ли правда везде порядок матриц другой, и у них там столбцы, а не строки нужно брать…

§ 10. Две леммы о факторизации матрицы

Тут я придумал что-то по мотивам конспекта и Голуба. Кажется оно работает.

Лемма 1. Пусть A — неособая матрица с ненулевыми диагональными минорами Тогда

$$\exists!\, L, R :\ A = LR, \quad L = \begin{pmatrix} 1 & & \\ \bullet & \ddots & \\ \bullet & \bullet & 1 \end{pmatrix}, \qquad R = \begin{pmatrix} \bullet & \bullet & \bullet \\ & \ddots & \bullet \\ & & \bullet \end{pmatrix}.$$

То есть раскладывается на произведение верхней/нижней треугольной.

▼
Эта теорема — матричная запись метода Гаусса. Запишем явное выражение для первого шага

$$a_{ij} = a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}}, \quad i = 2, \dots, s$$

Посмотрим на эту формулу как на преобразование jго столбца. Тогда матрица такого преобразования имеет вид

$$\begin{pmatrix} 1 & & & \\ \bullet & 1 & & \\ \vdots & 0 & 1 & \\ \bullet & 0 & 0 & 1 \end{pmatrix}$$

понятно, что в случае k-го шага будет просто k-й столбик. Если мы будем перемножать такие столбики, они просто будут пристраваться рядом. Ну в самом деле, умножим нижнетреугольную матрицу на такой столбик. $c_{ij} = \sum_p a_{ip}b_{pj}$, а при всех $j \neq k$ вместо b_{pj} просто такой же член от единичной матрицы. А сохранение треугольности следует из треугольности a_{ip}.

Чтобы убедиться в единственности, можно рассмотреть матричное равенство построчно. А у последовательного набора этих равенств получается всего одно решение. ◀

она везде называется LU факторизация и непонятно в чем разница.

Лемма 2. Пусть A — неособая матрица. Тогда

$$\exists\, Q, R :\ A = QR$$

Здесь R как в лемме выше, а Q — ортогональная.

▼
Прогоним процесс ортогонализации Грамма-Шмидта для строчек A, строчки Q это полученный ортогональный базис. При этом $Q = LA$, только на диагонали не обязательно 1. А L^{-1} уже будет верхнетреугольной. ◀

§ 11. Теорема о сходимости итерированных подпространств

Вспомним про степенной метод из § 4. У него была беда, он не умел искать больше одного собственного числа. Но мы и итерировали всего один вектор. Давайте обобщим.

Определение 1. Пусть A — диагонализуемая неособая матрица, $\{x_j\}$ — базис в $\mathbb{R}^s$,

1. $\{x_j^{(n)}\} = \{A^n u_j\}$ — n-ный итерированный базис
2. $L_s^{(n)} = \langle A^n x_1, \dots, A^n x_s \rangle$ — n-ное итерированое подпространство.

Замечание 1. Можно итерировать не весь базис, а, например, только k векторов из s. Помимо добавления эпитетов, характеризующих размерность, вводят $U_k = \langle u_1, \dots, u_k \rangle$ — k-мерное старшее собственное подпространство, а $\{u_j\}$ — базис из собственных векторов A.

Определение 2. Говорят, что $P^{(n)} \to P$, если в $P^{(n)}$ существует базис, сходящийся к базису P.

Теорема 1. Пусть A — диагонализуемая неособая матрица,

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_s| > 0 \qquad \text{(все разные)}$$

Тогда $L_k^{(n)} \to U_k$.

☐ Соорудим базис в $L_k^{(n)}$ который будем сходится к базису U_k. Пусть ${x_j}^k$ — исходный базис,[1]

$$\langle x_1, \dots, x_k \rangle \supset \langle u_1, \dots, u_k \rangle.$$

Разложим его по u_j и посмотрим на итерированный

$$x_i^{(n)} = \sum_{\ell=1}^{k} c_{i\ell}\,\lambda_\ell^n u_\ell + \sum_{\ell=k+1}^{s} c_{i\ell}\,\lambda_\ell^n u_\ell$$

Домножим обе части на $D = \left(\widetilde{C}\right)^{-1}$, $\widetilde{C}$ — квадратный кусок C размерами $k \times k$ из перых коэффициентов.

Рассмотрим $\widetilde{Z}_m^{(n)} = \sum_{i=1}^{k} x_i^{(n)}$, они явно базис $L_k^{(n)}$ в силу невырожденности D.

$$\widetilde{Z}_m^{(n)} = \sum_{\ell,i=1}^{k} \underbrace{d_{m,i}\,c_{i,\ell}}_{\delta_{m\ell}}\,\lambda_\ell^n u_\ell + \sum_{\ell=k+1,i=1}^{s,k} d_{m,i}\,c_{i,\ell}\,\lambda_\ell^n u_\ell$$

Теперь поделим: $Z_m^{(n)} = \widetilde{Z}_m^n\,\lambda_m^{-n}$, они всё ещё базис $L_k^{(n)}$.

$$Z_m^{(n)} = u_m + \sum_{\ell \geqslant k+1, i} d_{m,i}\,c_{i,\ell}\left(\frac{\lambda_\ell}{\lambda_m}\right)^n u_\ell \to u_m$$

■

Поплодим ещё сущностей

Определение 3 (Ступенчатый базис).

$$e_1 = (1, x_{12}, \dots) \tag{2.1}$$
$$e_k = (0, \dots, 0, 1, x_{k,k+1}, \dots) \tag{2.2}$$
$$\tag{2.3}$$

Утверждение 1. Обычно базис пространства приводится к ступенчатому.

▶ метод Гаусса ◀

можно ещё рассматривать, например, $e_1, \dots, e_k$ как базис L_k, нам ведь неважно что там в следущих компонентах происходит.

Теорема 2. Ступенчатый базис $L_k^{(n)}$ сходится к базису U_k при грамотно заданных условиях невырожденности.

☐ Сузим на подпространство и разложим по $Z_m^{(n)}$. Невероятно увлекательно ■

§ 12. Треугольно-степенной метод и его сходимость

Определение 1. Рассмотрим A со стандартными условиями на собственные вектора, произвольную невырожденную P_0. Шаг итерации выглядит так:

$$AP_n = P_{n+1}R_{n+1},$$

где P_{n+1} нижнетреугольная, а R_{n+1} верхнетреугольная.

[1] этого условия у нас не было, но без него не доказать невырожденность $\widetilde{C}$.

Теорема 1. При стандартных предположениях и неравенстве нулю диагональных миноров A на λ_A, P_k, R_k сходятся к P, R. При этом на диагонали R оказываются собственные числа.

□ 1. Первые k столбцов P_n образуют ступенчатый базис $L_k^{(n)}$

   (a) AP_n переводит его снова в базис, так как его можно через $Z_m^{(n)}$ выразить.

   (b) LR-факторизация выражает строчку L через предыдущие.

2. Ступенчатый базис сходится к базису U_k

3. $R_n = P_n^{-1}AP_n \xrightarrow[n\to\infty]{} P^TAP$, а у подобных матриц собственные числа совпадают.

■

скорость сходимости здесь степенная, что видно из теоремы в § 11.

§ 13. Ортогонально-степенной метод

Определение 1. Рассмотрим A со стандартными условиями на собственные вектора, произвольную невырожденную $_0$. Шаг итерации выглядит так:

$$AC_k = C_{n+1}R_{n+1},$$

где C_{n+1} ортогональная, а R верхнетреугольная.

Определение 2 (Сходимость по форме). Пусть B — блочная треугольная матрица. Тогда говорят, что A_k сходится по форме к B, если все элементы ниже квазидиагонали сходятся к 0. А что на диагонали и выше нас не интересует, главное чтобы хоть куда-то сходилось.

Теорема 1. При стандартных предположениях на λ_A и неравенстве нулю диагональных миноров A, $C_n^*AC_n$ по форме сходится к $\hat{A}$, которая верхнетреугольная. При этом на диагонали $\hat{A}$ оказываются собственные числа.

□ Для сходимости по форме нужно просто чтобы вся поддиагональ сходилась к 0. Т.е для $j > k$

$$\{C_n^*AC_n\}_{j,k} \to 0 \ \Leftrightarrow\ \left(AC_n^{(k)}, C_n^{(j)}\right) \to 0$$

Первые k строк C образуют ортогональный базис $L_k^{(n)}$.

Вспомним доказательство теоремы про итерируемые подпространства и скажем что x_i оттуда это $C^{(i)}$ сейчас. Тогда, $C^{(i)}$ раскладываются по Z^m + ещё какие-то члены порядка $O\left(\left|\frac{\lambda_{k+1}}{\lambda_k}\right|^n\right)$.

При умножении на A разложение по Z_m не испортилось, так что и $AC_n^{(i)}$ примерно в $L_k^{(n+1)}$. Тогда, из ортогональности всех столбцов C_n

$$\left(AC_n^{(k)}, C_n^{(j)}\right) = O\left(\left|\frac{\lambda_{k+1}}{\lambda_k}\right|^n\right) \xrightarrow[\to\infty]{} 0$$

■

§ 14. LR-алгоритм. Практическая реализация

Определение 1. Рассмотрим A со стандартными условиями на собственные вектора, произвольную невырожденную $_0$. Шаг итерации выглядит так:

$$A = L_1R_1,$$
$$R_nL_n = L_{n+1}R_{n+1},$$

где L_{n+1} нижнетреугольная, а R верхнетреугольная.

Нетрудно заметить, что есть связь этого алгоритма со треугольным степенным

$$P_0 = I$$
$$P_n = L_1 \cdots L_n$$

Подставим

$$\begin{aligned} AP_0 &= P_1 R_1 &\to&\quad AI = L_1 R_1 \\ AP_n &= P_{n+1} R_{n+1} &\to&\quad A\,L_1 \cdots L_n = L_1 \cdots L_{n+1} R_{n+1} \end{aligned}$$

Разберёмся с $A\,L_1 \cdots L_n$

$$A = L_1 R_1 = L_1 L_2\, R_2\, L_1^{-1} = \cdots = L_1 \cdots L_n\, R_n\, L_1^{-1} \cdots L_{n-1}$$

А подставив такого крокодила как раз получаем $R_n L_n = L_{n+1} R_{n+1}$ Поскольку мы свели метод к предущему, доказывать сходимость уже не нужно.

Приведём пару фактов, нужных для расчетов

1. LR факторизация занимает $O(s^3)$ времени. Это очень больно. Даже классический метод Якоби на шаге делает $O(s^2)$ работы.
2. Трехдиагональную матрицу можно факторизовать (прогонкой:) за $O(s)$ на двудиагональные.
3. Если A не трехдиагональная, но симметричная, вращения Гивенса нам помогут.
4. Можно привести A к трёхдиагональному виду даже если она несимметричная.
5. Можно ускорить сходимость взяв сдвиг по Реллею

$$R_n L_n - t_n I = L_{n+1} R_{n+1}, \qquad t_n = r_{ss}$$

   (последний элемент).

§ 15. QR-алгоритм. Практическая реализация

Определение 1. Рассмотрим A со стандартными условиями на собственные вектора, произвольную невырожденную $_0$. Шаг итерации выглядит так:

$$A = Q_1 R_1,$$
$$R_n Q_n = Q_{n+1} R_{n+1},$$

где C_{n+1} ортогональная, а R верхнетреугольная.

сходимость доказывается аналогично LR.

Приведём пару фактов, нужных для расчетов

1. QR-факторизация занимает $O(s^3)$ времени. И это всё больно.
2. QR-факторизация сохраняет трёхдиагональность A, но только для симметричных
3. Чтобы ускорить жизнь до $O(s^2)$ привести Q к форме Хессенберга вращениями Гивенса. В такой форме просто есть ещё одна диагональ по сравнению с верхнетреугольной.
4. Можно ускорить сходимость, соорудив сдвиг.

$$R_n Q_n - t_n I = Q_{n+1} R_{n+1}$$

   (a) По Реллею: $\{R_n Q_n\}_{s,s}$
   (b) По Уилкинсону: собтвенные числа матрицы 2×2 $\{R_n Q_n\}_{s-1,s-1}^{s,s}$

Господи, какая же это гадость. Сходите лучше в Голуба и почитайте это там. Всё равно Самокиш подробно только про их связь с предыдущими методами рассказал. Если он имеет в виду алгоритм написать, то это очевидно, а если все штуки доказать, то можно сразу накрываться простынёй и ползти на пересдачу.

3 Интегральные уравнения

§ 1. Интегральное уравнение II рода, метод замены ядра на вырожденное

Определение 1. *Интегральным уравнением Фредгольма II рода* называется уравнение вида

$$\varphi(x) = f(x) + \mu \int\limits_a^b K(x,\, t)\varphi(t)\,\mathrm{d}t. \tag{3.1}$$

Функция K — его *ядро*, а μ — *характеристическое число*.[2]

Обозначим через K (хм, да, вольность) оператор

$$\varphi(t) \mapsto \int\limits_a^b K(x,\, t)\varphi(t)\,\mathrm{d}t.$$

Ясно, что он компактен. Уравнение теперь примет вид

$$(I - \mu K)\varphi = f.$$

Оператор $T = I - \mu K$, конечно, фредгольмов.

Утверждение 1. Сопряжённый в $L^2([a,\, b])$ оператор к K выражается следующим образом:

$$K^*\varphi(x) = \int\limits_a^b \overline{K(t,\, x)}\varphi(t)\,\mathrm{d}t.$$

Доказательство. Прямым вычислением (ну, там внутри ещё теорема Фубини) проверяется, что

$$\langle K\varphi,\, \psi\rangle = \langle \varphi,\, K^*\psi\rangle.$$

□

Замечание 1. У ядра меняются местами аргументы и оно сопрягается — точно так же, как транспонирование вместе с комплексным сопряжением дают матрицу сопряжённого оператора в конечномерном случае!

Сформулируем альтернативу Фредгольма 1.6.1 для такого уравнения:

Утверждение 2.

1. Уравнение $T\varphi = f$ разрешимо однозначно тогда и только тогда, когда μ^{-1} — не собственное число оператора K.
2. В противном случае уравнение $T\varphi = f$ разрешимо тогда и только тогда, когда функция f ортогональна всем собственным векторам оператора K^*, соответствующим числу $\overline{\mu}^{-1}$.
3. μ^{-1} и $\overline{\mu}^{-1}$ — собственные числа операторов K и K^* соответственно одинаковой конечной кратности.

[2] Кажется, иногда в определении полагают $\mu = 1$, но всегда ведь можно внести его в ядро. Мы иногда тоже будем на него забивать.

Замечание 2. Для симметричного ядра (т.е. когда $K = K^*$) то же самое несложно доказать, используя разложение по собственному базису оператора K (которое есть по теореме Гильберта-Шмидта 1.5.1). Так можно быстро понять, что если μ^{-1} — собственное число K, то решений либо нет, либо их бесконечно много.

Рассмотрим уравнение 3.1 с вырожденным ядром

$$K(x, t) = \sum_{i=1}^{n} \alpha_i(x)\beta_i(t).$$

Функции α_i и β_i можно считать ЛНЗ: если это не так, нетрудно выразить одну из них через другие и избавиться от неё. Подставляя ядро в уравнение 3.1, получим

$$\varphi(x) = f(x) + \sum_{j=1}^{n} A_j\alpha_i(x), \text{ где } A_j = \mu \int_a^b \beta_j(t)\varphi(t)\,\mathrm{d}t. \tag{3.2}$$

Это представление для функции φ теперь подставим в исходное уравнение:

$$f(x) + \sum_{i=1}^{n} A_i\alpha_i(x) = f(x) + \mu \int_a^b \sum_{i=1}^{n} \alpha_i(x)\beta_i(t)\left(f(t) + \sum_{j=1}^{n} A_j\alpha_j(t)\right)\mathrm{d}t$$

Чтобы переписать это покороче, введём обозначения

$$\beta_{ij} = \int_a^b \beta_i(t)\alpha_j(t)\,\mathrm{d}t, \quad f_i = \int_a^b f(t)\beta_i(t)\,\mathrm{d}t.$$

и получим

$$\sum_{i=1}^{n} A_i\alpha_i(x) = \mu\sum_{i=1}^{n}\left(f_i + \sum_{j=1}^{n}\beta_{ij}A_j\right)\alpha_i(x).$$

Поскольку α_i линейно независимы, коэффициенты при них слева и справа должны быть равны. Записав эти равенства, мы приходим к системе линейных уравнений

$$A_i = \mu f_i + \mu\sum_{j=1}^{n}\beta_{ij}A_j.$$

В векторном виде она будет выглядеть так:

$$A = \mu(\beta A + f),$$

где A и f — векторы, β — матрица, а μ всё ещё число.

Эта система решается так:

$$(I - \mu\beta)A = \mu f \;\Rightarrow\; \boxed{A = \mu(I - \mu\beta)^{-1}f}, \text{ если } \det(I - \mu\beta) \neq 0.$$

Пусть $\Delta = \det(I - \mu\beta)$ и Δ_{ij} — алгебраическое дополнение элемента $\delta_{ij} - \mu\beta_{ij}$. Тогда можно записать явную формулу для A[1]:

$$A_i = \frac{\mu}{\Delta}\sum_{j=1}^{n}\Delta_{ji}f_j$$

Подставляя теперь найденные A_i в 3.2, найдём, что

$$\varphi(x) = f(x) + \lambda\int_a^b \Gamma(x, t)f(t)\,\mathrm{d}t,$$

где *резольвента* Γ имеет вид

$$\Gamma(x, t) = \frac{1}{\Delta}\sum_{i,j=1}^{n}\Delta_{ji}\alpha_i(x)\beta_j(s).$$

Трудная задача — приблизить произвольное ядро вырожденным. Есть несколько способов:

[1]Это просто формула для обратной матрицы через алгебраические дополнения.

1. Разложить ядро в ряд Тейлора.
2. Интерполировать ядро.
3. Разложить ядро по ортогональной системе функций.

Подробнее про них можно прочитать в книге [?].

Заменяя ядро на вырожденное, мы надеемся, что и решения тоже изменятся не сильно. Надо бы это обосновать (хотя бы как-то). Пусть есть уравнение

$$Au = f, \quad A = I - K$$

и приближающее его уравнение

$$A_n u_n = f, \quad A_n = I - K_n.$$

Нетрудно видеть, что

$$u - u_n = (A^{-1} - A_n^{-1})f \;\Rightarrow\; \|u - u_n\| \leqslant \left\|A^{-1} - A_n^{-1}\right\| \cdot \|f\|.$$

Поэтому интересно оценить норму разности обратных операторов. Займёмся этим.

Утверждение 3. Пусть P — ограниченный оператор, $\|P\| < 1$. Тогда оператор $I - P$ обратим, причём

$$(I - P)^{-1} = \sum_{i=1}^{\infty} P^n,$$

где сходимость — по операторной норме.

Утверждение 4. Пусть P и H — ограниченные операторы, P обратим, а $\|H\| < \|P^{-1}\|^{-1}$. Тогда элемент $P - H$ обратим, причём

$$\left\|(P - H)^{-1}\right\| \leqslant \frac{\|P^{-1}\|}{1 - \|H\|\,\|P^{-1}\|}.$$

и

$$\left\|(P - H)^{-1} - P^{-1}\right\| \leqslant \frac{\|H\|\,\|P^{-1}\|^2}{1 - \|H\|\,\|P^{-1}\|}.$$

Доказательство. Позволим себе иногда использовать дроби и 1 вместо I, как если бы операторы были числами. Не составит труда переписать всё через обратные!

Заметим, что первое из двух утверждений теоремы для $P = I$ следует из 3.1.3:

$$\left\|(I - H)^{-1}\right\| = \left\|\sum_{i=1}^{\infty} H^n\right\| \leqslant \sum_{i=1}^{\infty} \|H\|^n = \frac{1}{1 - \|H\|}. \tag{3.3}$$

Далее,

$$\left\|\frac{1}{P - H}\right\| = \left\|P^{-1}\frac{1}{1 - P^{-1}H}\right\| \leqslant \|P^{-1}\| \cdot \left\|\frac{1}{1 - P^{-1}H}\right\| \leqslant \frac{\|P^{-1}\|}{1 - \|P^{-1}H\|} \leqslant \frac{\|P^{-1}\|}{1 - \|H\|\,\|P^{-1}\|}.$$

В предпоследнем переходе используется соотношение 3.3, где $H \to P^{-1}H$.

Наконец,

$$\left\|\frac{1}{P - H} - \frac{1}{P}\right\| = \left\|\frac{1}{P}\left(\frac{1}{1 - P^{-1}H} - 1\right)\right\| = \left\|\frac{1}{P}\,\frac{P^{-1}H}{1 - P^{-1}H}\right\| \leqslant \frac{\|H\|\,\|P^{-1}\|^2}{1 - \|H\|\,\|P^{-1}\|}.$$

□

Отсюда сразу же следует утверждение

Утверждение 5. При достаточно больших n

$$\left\|A^{-1} - A_n^{-1}\right\| \leqslant \frac{\rho\,\|A^{-1}\|^2}{1 - \rho\,\|A^{-1}\|} \text{ и } \left\|A^{-1} - A_n^{-1}\right\| \leqslant \frac{\rho\,\|A_n^{-1}\|^2}{1 - \rho\,\|A_n^{-1}\|},$$

где $\rho = \|A - A_n\| = \|K - K_n\|$.

Замечание 3. Рассмотрим теперь задачу с симметричным ядром (т.е. с самосопряжённым K). В ней есть ортонормированный собственный базис α_i, поэтому

$$u = \sum_{i=1}^{\infty} \langle u, \alpha_i \rangle \alpha_i \Rightarrow Ku = \sum_{i=1}^{\infty} \langle u, \alpha_i \rangle \lambda_i \alpha_i,$$

где λ_i — соответствующее собственное число. Расположим λ_i в порядке убывания модуля и положим

$$K_n u = \sum_{i=1}^{n} \langle u, \alpha_i \rangle \lambda_i \alpha_i$$

Это интегральный оператор с вырожденным ядром

$$K_n(x, t) = \sum_{i=1}^{n} \lambda_i \alpha_i(x) \overline{\alpha_i(t)}. \tag{3.4}$$

Можно доказать, что он является лучшей аппроксимацией ранга n для оператора K по операторной L^2-норме.

Посмотрим на разность:

$$(K - K_n)u = \sum_{i=n+1}^{\infty} \langle u, \alpha_i \rangle \lambda_i \alpha_i.$$

Найдём её норму:

$$\left\| (K - K_n)u \right\|^2 = \sum_{i=n+1}^{\infty} |u_i|^2 \, |\lambda_i|^2, \quad u_i = \langle u, \alpha_i \rangle.$$

При этом

$$\|K - K_n\| = \sup \frac{\left\| (K - K_n)u \right\|}{\|u\|},$$

и

$$\frac{\left\| (K - K_n)u \right\|^2}{\|u\|^2} = \frac{\sum\limits_{i=n+1}^{\infty} |u_i|^2 \, |\lambda_i|^2}{\sum\limits_{i=n+1}^{\infty} |u_i|^2} \leqslant \frac{\sum\limits_{i=n+1}^{\infty} |u_i|^2 \, |\lambda_{n+1}|^2}{\sum\limits_{i=n+1}^{\infty} |u_i|^2} = |\lambda_{n+1}|^2.$$

С другой стороны, эта оценка достигается, когда u — собственный вектор числа λ_{n+1}. Поэтому

$$\boxed{\|K - K_n\| = |\lambda_{n+1}|}.$$

Отсюда и из утверждения 3.1.5 ясно: чем быстрее убывают собственные числа, тем лучше наша оценка! Из уравнения 3.4 видно, что собственные числа — что-то вроде коэффициентов в ряде Фурье по собственным функциям для ядра. Видимо, поэтому скорость их убывания возрастает, если ядро становится более гладким... А ядра гладкие не всегда.

Замечание 4. Есть способ сгладить ядро. Надо в уравнение 3.1 подставить

$$\varphi(t) = f(t) + \mu \int_a^b K(t, \xi) \varphi(\xi) \, \mathrm{d}\xi.$$

Получится уравнение

$$\varphi(x) = f_2(x) + \mu \int_a^b K_2(x, \xi) \varphi(\xi) \, \delta\xi,$$

где

$$f_2(x) = f(x) + \mu \int_a^b K(x, t) f(t) \, \mathrm{d}t, \quad K_2(x, \xi) = \mu \int_a^b K(x, t) K(t, \xi) \, \mathrm{d}t.$$

У K_2 с гладкостью получше, но его надо считать.

§ 2. Метод квадратур для интегрального уравнения

Идея заключается в том, чтобы в уравнении

$$u(x) = f(x) + \int_a^b K(x,\, t)u(t)\,\mathrm{d}t$$

заменить интегрирование на вычисление по какой-нибудь квадратурной формуле:

$$\int_a^b u(x)\,\mathrm{d}x = \sum_{k=1}^{n} A_k u(x_k) + R.$$

Получится

$$u(x) = f(x) + \sum_{k=1}^{n} A_k K(x,\, x_k)u(x_k) + R.$$

Пусть $\tilde{u}$ — решение этого уравнения с отброшенным R, $u_k = \tilde{u}(x_k)$, $f_k = f(x_k)$ и $K_{ik} = K(x_i,\, x_k)$. Получаем систему линейных уравнений

$$u_i = f_i + \sum_{k=1}^{n} A_k K_{ik} u_k.$$

Её можно решить обычными методами; зная u_k, можно оценить $u(x)$ в любой точке:

$$u(x) = f(x) + \sum_{k=1}^{n} A_k K(x,\, x_k)u_k.$$

Попробуем оценить погрешность результата. Для многих стандартных квадратурных методов верна формула

$$R[\theta] = \delta(n)\max|\theta^{(m)}(x)|.$$

Нас интересует $R\big[K(x,\, t)u(t)\big]$ при фиксированном x. m-е производные функции $K(x,\, t)u(t)$ выражаются через производные известной $K(x,\, t)$ и через производные $u(t)$ порядка не более m.

Чтобы оценить их, продифференцируем наше интегральное уравнение:

$$u^{(l)}(x) = f^{(l)}(x) + \int_a^b K_x^{(l)}(x,\, t)u(t)\,\mathrm{d}t.$$

Отсюда можно найти оценку для $u^{(l)}$ через известные f и K и максимум модуля решения. Решение же можно записать, как

$$u = (I-K)^{-1}f \;\Rightarrow\; \|u\| \leqslant \left\|(I-K)^{-1}\right\| \cdot \|f\| \leqslant \frac{\|f\|}{1-\|K\|} \leqslant \frac{\|f\|}{1-\varkappa},$$

где

$$\varkappa = (b-a)\max\big|K(s,\, t)\big|.$$

Предпоследний переход обусловлен утверждением 3.1.4.

Замечание 1. Во-первых, сейчас у нас все нормы — L^1, от этого ничего не портится. Во-вторых, мы только что неявно предположили, что $|\varkappa| < 1$.

Получив оценку для модуля решения, мы можем найти оценку

$$\left|\frac{\partial^m}{\partial t^m}\big(K(x,\, t)u(t)\big)\right| \leqslant M,$$

зависящую только от известных функций.

Перейдём теперь непосредственно к оценке ошибки. У нас есть два уравнения

$$Au = f, \quad A = I - K;$$
$$\tilde{A}\tilde{u} = f, \quad \tilde{A} = I - \tilde{K},$$

где
$$\tilde{K}\varphi(x) = \sum_{i=1}^{n} A_i K(x,\, x_i)\varphi(x_i).$$
Заметим, что
$$\tilde{A}(u-\tilde{u}) = \tilde{A}u - Au \;\Rightarrow\; \|u-\tilde{u}\| \leqslant \boxed{\|\tilde{A}^{-1}\|\,\|\tilde{A}u - Au\|}\,.$$
Оценим норму $\tilde{A}^{-1}$. Для этого сначала оценим норму $\tilde{K}$:
$$\left|\sum_{i=1}^{n} A_i K(x,\, x_i)\varphi(x_i)\right| \leqslant \max|K|\cdot\|\varphi\|\cdot\sum_{i=1}^{n} A_i = (b-a)\max|K|\cdot\|\varphi\|,$$
поэтому $\|\tilde{K}\| \leqslant \varkappa$.

Отсюда
$$\|\tilde{A}^{-1}\| = \left\|(I-\tilde{K})^{-1}\right\| \leqslant \frac{1}{1-\varkappa}.$$
Теперь оценим $\|\tilde{A}u - Au\|$:
$$\|\tilde{A}u - Au\| = \max\left|R\big[K(x,\, t)u(t)\big]\right| \leqslant M\delta(n).$$
В конечном итоге находим
$$\boxed{\|u-\tilde{u}\| \leqslant \frac{M\delta(n)}{1-\varkappa}}\,.$$

Подробнее про этот метод можно прочитать в книгах [?] и [?].

§ 3. Вариационный принцип для ограниченного оператора; метод Ритца для интегрального уравнения II рода

Замечание 1. В этом параграфе все гильбертовы пространства вещественны.

Основная идея заключается в том, чтобы свести решение уравнения
$$Au = f$$
к минимизации некоторого функционала.

Определение 1. *Энергетическим* функционалом для такого уравнения называется
$$\tilde{f}(u) = (Au,\, u) - 2(f,\, u).$$

Чтобы работать с энергетическим функционалом, нужны дополнительные ограничения на оператор A.

Определение 2. Оператор A называют *положительно определённым*, если $(Au,\, u) \geqslant k^2(u,\, u)$[1].

Утверждение 1. Самосопряжённый положительно определённый оператор A обратим.

Доказательство. Положим в доказательстве $k^2 = 1$, ибо на обратимость это не влияет, можно просто разделить A на k^2. Заметим, что $\ker A = \{0\}$, поскольку
$$Au = 0 \;\Rightarrow\; (Au,\, u) = 0 \;\Rightarrow\; (u,\, u) = 0 \;\Rightarrow\; u = 0.$$
При этом ортогональное дополнение образа A — его ядро:
$$x \in \operatorname{Im} A^{\perp} \Leftrightarrow \forall u \quad 0 = (x,\, Au) = (Ax,\, u) \Leftrightarrow Ax = 0.$$
Поэтому
$$\overline{\operatorname{Im} A} = \ker A^{\perp} = H,$$
и образ оператора A плотен в H.

[1]Это необычное название, кажется. Их называют ещё *полуограниченными снизу*.

Докажем, что он на самом деле равен H. Для этого нам пригодится неравенство

$$\|u\|^2 \leqslant (Au,\, u) \leqslant \|Au\|\,\|u\| \;\Rightarrow\; \boxed{\|u\| \leqslant \|Au\|}.$$

Пусть $y \in H$. Поскольку образ плотен, найдётся последовательность $\{x_n\}$ такая, что $Ax_n \to y$. Однако

$$\|x_n - x_m\| \leqslant \|Ax_n - Ax_m\|,$$

поэтому $\{x_n\}$ сходится в себе; гильбертово пространство полно, поэтому $x_n \to x$. Но оператор A непрерывен, и

$$x_n \to x \;\Rightarrow\; Ax_n \to Ax \;\Rightarrow\; Ax = y.$$

Таким образом, A сюръективен, и у него есть теоретико-множественный обратный.

При этом

$$\|A^{-1}y\| \leqslant \|y\|,$$

поэтому обратный оператор ограничен.

□

Утверждение 2. Если A — самосопряжённый и положительно определённый, то существует единственное решение u^* уравнения $Au = f$, которое совпадает с единственным минимумом энергетического функционала.

Доказательство. Существование и единственность решения следуют из обратимости оператора. Посчитаем значение функционала на векторе $u^* + h$:

$$\tilde{f}(u^* + h) = \big(A(u^* + h),\, u^* + h\big) - 2(f,\, u^* + h) = \tilde{f}(u^*) + (Au^*,\, h) + (Ah,\, u^*) + (Ah,\, h) - 2(f,\, h) =$$
$$= \tilde{f}(u^*) + (h,\, f) - (f,\, h) + (Ah,\, h).$$

Мы считаем всё вещественным, поэтому $(h,\, f) = (f,\, h)$ и

$$\tilde{f}(u^* + h) = \tilde{f}(u^*) + (Ah,\, h) \geqslant \tilde{f}(u^*).$$

□

Метод Ритца устроен примерно так:

1. Выбрать в пространстве H линейно независимый набор $\{\varphi_k\}$.
2. Рассмотреть конечномерное подпространство H_n, натянутое на первые n векторов базиса.
3. Найти в нём минимум функционала $\tilde{f}$ и считать его приближением.

Минимум в H_n будем искать в виде

$$u_n = \sum_{k=1}^{n} c_k \varphi_k.$$

Утверждение 3. Координаты c_n минимума $\tilde{f}$ в подпространстве H_n находится из системы линейных уравнений

$$\sum_{k=1}^{n} (A\varphi_k,\, \varphi_i)\, c_n = (f,\, \varphi_i)$$

Доказательство. Если подставить

$$u_n = \sum_{k=1}^{n} c_k \varphi_k$$

в формулу для функционала

$$\tilde{f}(u_n) = (Au_n,\, u_n) - 2(f,\, u_n),$$

получится

$$\tilde{f}(u_n) = \sum_{k,\,m} c_k c_m (A\varphi_k,\, \varphi_m) - 2\sum_{m} c_m (f,\, \varphi_m).$$

Дифференцируя это выражение по c_i и приравнивая к нулю, получим нужную СЛУ. □

Замечание 2. Симметричная матрица $a_{ij} = (A\varphi_i, \varphi_j)$ — матрица Грама положительно определённой симметрической билинейной формы $g(u, v) = (Au, v)$. Известно, что определитель матрицы Грама равен квадрату объёма параллелипипеда, натянутого на базисные вектора, в соответствующей метрике. Он, конечно, ненулевой, а потому система линейных уравнений разрешима однозначно.

Поговорим о сходимости метода Ритца.

Утверждение 4. Если набор $\{\varphi_k\}$ таков (это по сути означает, что он является базисом), что

$$\forall v \in H \quad \forall \varepsilon > 0 \quad \exists\, n, \alpha_i : \left\| v - \sum_{i=1}^{n} \alpha_i \varphi_i \right\| < \varepsilon,$$

то метод Ритца сходится, т.е. $\|u_n - u^*\| \to 0$.

Доказательство. Поскольку оператор A положительно определён, форма $g(u, v) = (Au, v)$ является настоящим скалярным произведением. Мы утверждаем, что u_n — элемент из H_n, ближайший к u^* с точки зрения метрики g. Докажем это. Для этого предположим, что

$$u_n = v_n + h,$$

где $v_n = u^* - v_n^\perp$ — ближайший к u^* элемент из H_n, а $v_n^\perp \perp H_n$. Тогда

$$\begin{aligned} \tilde{f}(u_n) &= \tilde{f}(u^*) + \left(A(h - v_n^\perp), h - v_n^\perp\right) = \\ &= \tilde{f}(u^*) + g(v_n^\perp, v_n^\perp) + g(h, h). \end{aligned}$$

Видно, что это выражение минимально, когда $h = 0$ и $u_n = u^* - v_n^\perp$.

Найдём теперь по ε такое N и $w \in H_N$, что $\|w - u^*\| < \varepsilon$. Тогда

$$\|u_n - u^*\| \leqslant \frac{1}{k}\, \|u_n - u^*\|_A \leqslant \frac{1}{k} \|w - u^*\|_A \leqslant \frac{\sqrt{\|A\|}}{k} \|w - u^*\| < \frac{\sqrt{\|A\|}}{k} \varepsilon,$$

где $\|x\|_A = \sqrt{g(x, x)}$. Объясним переходы по пунктам:

1. Потому что $g(x, x) \geqslant k^2 (x, x)$.
2. Потому что u_n — самый близкий элемент к u^*.
3. Потому что $(Ax, x) \leqslant \|Ax\|\, \|x\| \leqslant \|A\|\, (x, x)$.
4. Прост)00

Эпсилон домножился на константу, но это не страшно: стремление к нулю всё равно есть. □

Замечание 3. Видно, что скорость сходимости метода от гладкости ядра не зависит (только от его нормы). По сути она определяется тем, насколько быстро убывают коэффициенты разложения u^* по базису φ_k, что связано с гладкостью решения. Зато ограничения на оператор сильные.

§ 4. Интегральное уравнение I рода и его некорректность

Определение 1. *Интегральным уравнением I рода* называют уравнение вида

$$\int_a^b K(x, t) u(t)\, \mathrm{d}t = f(x).$$

Определение 2. Говорят, что задача *корректна*, если при малых изменениях исходных данных решение меняется слабо.

Определение 3. Задачу вида $Au = f$ называют *корректной*, если у оператора A есть ограниченный обратный.

Кажется, эти два определения почти одинаковые. :)

Утверждение 1. Задача о решении уравнения Фредгольма I рода некорректна.

Доказательство. Интуитивно это понятно: мы решаем уравнение вида $Ku = f$, где K — компактный оператор. Его образ маленький, и логично, что слегка изменив f мы можем получить задачу с совсем другим решением или, скорее, вовсе неразрешимую.

Покажем это для случая симметричного ядра. Симметричность позволит нам выбрать собственный базис $\{\varphi_n\}$ с собственными числами λ_n. Пусть

$$u = \sum u_i \varphi_i; \quad u = \sum f_i \varphi_i.$$

Тогда уравнение перепишется, как

$$\sum u_i \lambda_i \varphi_i = \sum f_i \varphi_i \Leftrightarrow \boxed{u_i \lambda_i = f_i}.$$

Формально решение имеет вид

$$u = \sum_{i=1}^{\infty} \frac{f_i}{\lambda_i} \varphi_i.$$

Если $\lambda_i = 0$, а $f_i \neq 0$, то задача наверняка не имеет решения. Всё плохо, даже если это не так: известно, что собственные числа компактного оператора стремятся к нулю, поэтому ряд для u будет сходиться только если f_i убывают ещё быстрее.

Посмотрим, что будет при небольшом изменении начальных данных; пусть

$$Ku = f; \quad k\tilde{u} = \tilde{f}; \quad \tilde{f} = f + \delta f; \quad \tilde{u} = u + \delta u,$$

причём $\|\delta f\| < \varepsilon$. Функция δu удовлетворяет уравнению $K\delta u = \delta f$. Решение должно выглядеть как

$$\delta u = \sum_{i=1}^{\infty} \frac{\delta f_i}{\lambda_i} \varphi_i.$$

Даже если этот ряд сходится, нельзя гарантировать, что при $\varepsilon \to 0$ δu тоже будет стремиться к нулю.

Действительно, всегда можно выбрать $\delta f = \varepsilon \varphi_n$, где n таково, что $\lambda_n < \varepsilon$. Тогда $\|\delta u\|$ будет больше 1. □

Подробнее про это можно прочитать в книге [?].

§ 5. Условная корректность по Тихонову, метод квазирешений

Теорема 1. (об условной корректности) Пусть оператор A на гильбертовом пространстве H некорректен, но инъективен (устанавливает взаимно однозначное отображение на образ). Рассмотрим компакт $L \in H$, пусть M — его образ. Отображение A^{-1} непрерывно на M[1].

Доказательство. Возьмём какую-нибудь последовательность $f_n \to f$ в M. Оператор A инъективен, поэтому элементы u_n такие, что $Au_n = f_n$ определены однозначно. Выберем в $\{u_n\}$ какую-нибудь сходящуюся подпоследовательность $u'_n \to u'$.

Поскольку оператор A непрерывен, $Au'_n \to Au'$; но $Au'_n \to f$, поэтому и $Au' = f \;\Rightarrow\; u' = A^{-1}f$. Но тогда выходит, что пределы всех сходящихся подпоследовательностей в $\{u_n\}$ одинаковы! Поэтому все частичные пределы совпадают, и u_n имеет предел, который равен $A^{-1}f$, что и даёт нам непрерывность. □

Определение 1. Это свойство — иметь непрерывный обратный на образах компактов — и называется *условной корректностью*.

Пусть мы решаем задачу $Au = f$, причём правая часть известна с погрешностью:

$$\|f - f_\delta\| \leqslant \delta,$$

но уравнение $Au = f_\delta$ не всегда имеет решение даже когда f_δ из этого шара. Приходим к определению *квазирешения*:

[1]На самом деле это просто стандартная теорема про то, что непрерывное отображение из компактного пространства в хаусдорфово является гомеоморфизмом на образ.

Определение 2. Зафиксируем конкретную f_δ. Тогда *квазирешением* уравнения $Au = f_\delta$ называется вектор u_δ, при котором достигается

$$\min_{u \in D} \|Au - f_\delta\|, \quad D = \{u \mid \|u\| \leqslant R\}.$$

Если искать не при $\|u\| \leqslant R$, а при $\|u\| = R$, получится задача на условный экстремум. Используя метод множителей Лагранжа, будем минимизировать функционал

$$F(u) = \alpha\|u\|^2 + \|Au - f_\delta\|^2.$$

Утверждение 1. Минимум этого функционала удовлетворяет уравнению

$$(\alpha I + A^*A)u = A^*f.$$

Доказательство. Обычный поиск вариации, нужно расписать $F(u + th)$ через скалярные произведения, продифференцировать по t, а после положить t равным нулю. □

Мы получили уравнение, похожее на исходное, но оно уже второго типа, а при малых α похоже на исходное. Произошла *регуляризация*! Более того, оператор $\alpha I + A^*A$ самосопряжён, и

$$\big((\alpha I + A^*A)u,\, u\big) = \alpha(u,\, u) + (Au,\, Au) \geqslant \alpha(u,\, u),$$

поэтому применим вариационный принцип.

Замечание 1. Насколько я понимаю, метод квазирешений нам рассказан в основном для того, чтобы прийти к регуляризации. Из квазирешений следует, что при некотором α минимум функционала должен хорошо приближать решение — мотивация! Плюс демонстрация того, что не совсем очевидно, что альфу можно просто к нулю стремить.

§ 6. Метод регуляризации для уравнения I рода, сходимость

Замечание 1. Кажется, этого билета почти нет у Оли, поэтому я опускаю доказательства. Это хорошо написано у Ангелины.

Идея регуляризации заключается в том, чтобы минимизировать функционал вида

$$F(u) = \alpha\Omega(u) + \|Au - f\|^2,$$

где $\Omega(u) \geqslant 0$ и множества $\Omega(u) < C$ компактны.

Замечание 2. В методе квазирешений у нас получился $\Omega(u) = \|u\|^2$, для него эти множества — открытые шары, они совсем не компактны.

Стандартный выбор — функционал

$$\Omega(u) = \int\limits_a^b u'^2 \,\mathrm{d}t.$$

Правда, при этом мы начинаем искать решение среди гладких функций.

Утверждение 1. Для такого функционала Ω множества $\Omega(u) < C$ компактны.

Доказательство. Стандартное рассуждение, использующее теорему Арцела-Асколи: подмножество в пространстве непрерывных функций на отрезке компактно тогда и только тогда, когда оно равномерно ограничено и равностепенно непрерывно. Из компактности в смысле топологии пространства непрерывных функций следует компактность в смысле L^2-нормы. □

Годятся и функционалы

$$\Omega(u) = \int\limits_a^b {u^{(p)}}^2 \,\mathrm{d}t, \quad \Omega(u) = \int\limits_a^b u'^2 - u^2 \,\mathrm{d}t.$$

Теорема 1. Пусть $\|f - f_\delta\| < \delta$, и мы решаем приближённую задачу $A\tilde{u} = f_\delta$ вместо точной. Если δ и α стремятся к нулю так, что

$$\frac{\delta^2}{\alpha} \leqslant \gamma < \infty,$$

то $\tilde{u} \to u$.

Доказательство. См. конспект Ангелины. $\square$

Коэффициент α обычно подбирают эмпирически: если он мал, то решение будет ближе к $\tilde{u}$, если велик, оно будет глаже... Стандартный функционал приводит к вариационной задаче

$$K^*Ku - \alpha u'' = K^*f.$$

4 Вариационные методы

§ 1. Вариационный принцип для уравнения с неограниченным оператором

Определение 1. *Неограниченным* называется оператор A на гильбертовом пространстве, определённый на его всюду плотном линейном подпространстве $\mathcal{D}(A)$.

Мы будем требовать от A также симметричности (самосопряжённости) и положительной определённости.

Определение 2. Билинейная форма $(u, v)_A = (Au, v)$ называется *энергетическим скалярным произведением*, норма $\|u\|_A = (u, u)_A$ — *энергетической нормой*. Пополнение H_A пространства $\mathcal{D}(A)$ по энергетической норме называется *энергетическим пространством*.

Замечание 1. В доказательстве 3.3.4 мы видели, что

$$\|u\| \leqslant k^{-1}\|u\|_A.$$

Поэтому если последовательность сходится в себе по энергетической норме, то она сходится и по обычной; кофинальные последовательности тоже одинаковые и там, и там. Поэтому пополнение по энергетической норме можно рассматривать, как подмножество H.

Теорема 1. (О вариационном принципе) Рассмотрим энергетический функционал $F(u) = (u, u)_A - 2(f, u)$.

1. $F(u)$ имеет единственный минимум u^*.
2. Если $u^* \in \mathcal{D}(\mathcal{A})$, то $Au^* = f$.
3. Если $Au_0 = f$, то $u^* = u_0$.

Доказательство. Рассмотрим функционал $\Phi(u) = (f, u)$. Он ограничен на H_A, поскольку

$$\left|(f, u)\right| \leqslant \|f\|\,\|u\| \leqslant k^{-1}\|f\|\,\|y\|_A.$$

По теореме Рисса он представим в виде $\Phi(u) = (u^*, u)_A$. Тогда

$$F(u) = (u, u)_A - 2(u^*, u)_A = \|u - u^*\|^2 - \|u^*\|^2.$$

Ясно, что минимум достигается при $u = u^*$.

Третий пункт очевиден, ибо

$$(f, u) = (Au_0, u) = (u_0, u)_A \;\Rightarrow\; u_0 = u^*$$

То же самое в обратную сторону даёт пункт 2. □

§ 2. Метод Ритца, сходимость

Метод Ритца и в энергетической Африке метод Ритца. Ну да, φ_n теперь лежат в H_A, а в остальном — никакой разницы.

Теорема 1. Если набор $\{\varphi_k\}$ таков (это по сути означает, что он является базисом), что

$$\forall v \in H \quad \forall \varepsilon > 0 \quad \exists\, n,\, \alpha_i : \left\| v - \sum_{i=1}^{n} \alpha_i \varphi_i \right\|_A < \varepsilon,$$

то метод Ритца сходится, т.е. $\|u_n - u^*\|_A \to 0$.

Доказательство. Доказательство теоремы аналогично доказательству 3.3.4. Первый абзац такой же, по сути, а дальше надо оставить только оценки, содержащие энергетическую норму (я случайно ей воспользовался, когда ещё не знал, что это, извините. то доказательство было непонятно, а то, что я придумал — скорее отсюда). □

§ 3. Метод Ритца для обычной краевой задачи, вид энергетического пространства, естественные граничные условия

Рассмотрим краевую задачу для уравнения

$$L(y)(x) = -(p(x)y')' + q(x)y = f(x)$$

на отрезке $[a,\, b]$ с граничными условиями

- I типа: $y(a) = 0$, $y(b) = 0$.
- III типа: $y'(a) = \alpha y(a)$, $y'(b) = \beta y(b)$.

Определение 1. *Классическое решение* — лежит в $C^2\big([a,\, b]\big)$, ~~никого не трогает~~ удовлетворяет уравнению в каждой точке.

Ну и $\mathcal{D}(L) = C^2\big([a,\, b]\big)$.

Замечание 1. На самом деле, мы ищем решения не в $\mathcal{D}(L)$, а в более узких пространствах. В случае условия I типа нас интересует простанство

$$D_I = \big\{ y \in \mathcal{D}(L) \,|\, y(a) = y(b) = 0 \big\},$$

а в случае условия III типа

$$D_{III} = \big\{ y \in \mathcal{D}(L) \,|\, y'(a)\alpha y(a),\ y'(b) = \alpha y(b) \big\}.$$

Именно их мы будем пополнять, создавая соответствующее энергетическое пространство.

Утверждение 1. Если $\alpha \geqslant 0$ и $\beta \leqslant 0$ в добавок к условиям

$$p(x) \geqslant p_0 > 0, \quad q(x) \geqslant 0,$$

то оператор получится симметричный и положительно определённый.

Доказательство. Посмотрим, как будет выглядеть энергетическое скалярное произведение:

$$(Ly,\, z) = \int\limits_a^b \big(-(py')' + qy \big) z \,\mathrm{d}x = -py'z\big|_a^b + \int\limits_a^b (py'z' + qyz) \,\mathrm{d}z.$$

Интеграл обозначим через $I(X)$, а внеинтегральный член — $Q(x)$. Если условия первого типа, то $Q = 0$, а если третьего, то

$$Q(y,\, z) = -\beta p(b)y(b)z(b) + \alpha p(a)y(a)z(a).$$

Симметричность уже видна, и

$$(Ly,\, y) = \int\limits_a^b (py'^2 + qy^2) \,\mathrm{d}x - \beta p(b)y(b)^2 + \alpha p(a)y(a)^2.$$

Если, $\alpha \geqslant 0$ и $\beta \leqslant 0$, то и положительная определённость будет. □

Определение 2. *Пространством Соболева* $W_p^k(Q) \subset L^p(Q)$ называют пространство функций, обобщённые производные которых вплоть до k-й лежат в $L_p(Q)$.

Замечание 2. На пространствах Соболева есть норма. Нас будет интересовать пространство $W_2^1\big([a,\, b]\big)$; на нём эта норма имеет вид

$$\|f\|_{W_2^1}^2 = \int_a^b (f^2 + f'^2)\,\mathrm{d}x.$$

Можно доказать, что с такой нормой является гильбертовым (а произвольные пространства Соболева — банаховы).

Утверждение 2. Энергетическая норма для оператора L эквивалентна норме в W_2^1.

Доказательство. Пусть $P_m = \max p$, $Q_m = \max q$, $M = \max(P_m,\, Q_m)$. Нетрудно доказывается, что $\|y\|_{W_2^1} \leqslant C\|y\|_L$:

$$\int_a^b (f^2 + f'^2)\,\mathrm{d}x \leqslant \frac{1}{M}\int_a^b (pf^2 + qf'^2)\,\mathrm{d}x \leqslant \frac{1}{M}\|f\|_L.$$

Обратное утверждение очевидно для I типа граничных условий:

$$\int_a^b (pf^2 + qf'^2)\,\mathrm{d}x \leqslant M\|f\|_{W_2^1}.$$

Чтобы разобраться с граничными условиями III типа, нам понадобится лемма:

Лемма 1. Для любой точки x значение $y(x)^2$ не превосходит константы, умноженной на $\|y\|_{W_2^1}^2$.

Доказательство. Ограничение соболевской нормы даёт ограничение на интеграл от квадрата функции + не позволяет ей расти слишком быстро, поэтому есть надежда, что значения и правда будут ограничены нормой. Займёмся оценкой. Очевидно, что

$$y(x) = y(\xi) + \int_\xi^x y(t)\,\mathrm{d}t.$$

Поскольку $(a+b)^2 \leqslant 2(a^2+b^2)$,

$$y(x)^2 \leqslant 2y(\xi)^2 + 2\left(\int_\xi^x y'(t)\,\mathrm{d}t\right)^2.$$

При этом интеграл

$$\int_\xi^x y'(t)\,\mathrm{d}t$$

является L^2-произведением (в отрезке от ξ до x) $\big(y'(t),\, 1\big)_{L^2}$, и

$$\big(y'(t),\, 1\big)_{L^2}^2 \leqslant \|1\|_{L^2([x,\,\xi])}^2\, \|y'\|_{L^2([x,\,\xi])}^2 = (x-\xi)\int_\xi^x y'(t)^2\,\mathrm{d}t \leqslant (b-a)\|y'\|_{L^2}^2$$

В итоге получаем, что

$$y(x)^2 \leqslant 2y(\xi)^2 + 2(b-a)\|y'\|_{L^2}^2$$

Навесив слева и справа интегралы по ξ, получим, что

$$y(x)^2 \leqslant \frac{2}{b-a}\|y\|_{L^2}^2 + 2(b-a)\|y'\|_{L^2}^2 \leqslant C\|y\|_{W_2^1}^2.$$

□

Используя полученную оценку, нетрудно оценить отвечающий за граничные условия член $Q(x)$ через соболевскую норму. $\square$

Замечание 3. Ещё выполняется *теорема вложения*: все функции из W_2^1 непрерывны, при этом отображение вложения $W_2^1 \to C([a,\, b])$ непрерывно.

Утверждение 3. Энергетическое пространство H_L является подпространством в W_2^1.

Доказательство. Не очень важно, D_I или D_{III} придётся пополнять: они оба лежат в $\mathcal{D}(L)$, про которое мы доказали, что с энергетической нормой оно гомеоморфно вкладывается в W_2^1. Поскольку W_2^1 гильбертово, пополнение нас из него не выведет. $\square$

Замечание 4. Пополнение пространства D_I приведёт нас к пространству $\overset{\circ}{W}{}_1^2$ элементов W_1^2, удовлетворяющих граничному условию I типа. С условием III так не получится, поскольку производная — не непрерывный функционал, и мы придём ко всему W_1^2. По этой причине условия I типа называют *главными*, а III типа — *естественными*.

О пространствах Соболева в контексте вычислительных методов можно прочитать в книге [?], и ещё подробнее в книге [?].

§ 4. ВРМ-1 для обычной краевой задачи

Идея *вариационно-разностных методов* заключается в том, чтобы использовать сетку и минимизацию функционала одновременно.

Пусть в сетке n элементов, $h = \dfrac{b-a}{n}$, $x_k = a + kh$; рассмотрим пространство, состоящее из сеточных функций $y_{(n)} = \{y_k\}_0^n$. Суть ВРМ-1 в том, чтобы заменить интегралы на суммы, а производные — на разности, и минимизировать функционал на сеточных функциях.

Наш функционал имеет вид

$$F(y) = (y,\, y)_L - 2(f,\, y) = \int_a^b (py'^2 + qy^2 - 2fy)\,\mathrm{d}x - \beta p(b)y^2(b) + \alpha p(a)y^2(a).$$

Сделаем численные замены:

$$\int_a^b py'\,\mathrm{d}x \approx h\sum_{k=0}^{n-1} p\left(x_k + \frac{h}{2}\right)\cdot\left(\frac{y_{k+1}-y_k}{h}\right)^2;$$
$$\int_a^b (qy^2 - 2fy)\,\mathrm{d}x \approx k\sum_{k=0}^{n-1}{}'(q_k y_k^2 - 2f_k y_k),$$

где сумма со штрихом означает, что это формула трапеций (т.е. крайние слагаемые домножены на 1/2).

Не представляет труда теперь выписать сеточный функционал. Далее минимум ищется дифференцированием по y_k и приравниванием всех производных к нулю. В итоге для внутренних точек получаются уравнения

$$-\frac{1}{h}\left(p_{i+\frac{1}{2}}\frac{y_{i+1}-y_i}{h} - p_{i-\frac{1}{2}}\frac{y_i - y_{i-1}}{h}\right) + q_i y_i = f_i.$$

Они напоминают уравнения разностной прогонки.

Для левого конца получится уравнение

$$-p_{\frac{1}{2}}\frac{y_1-y_0}{2} + \frac{h}{2}(q_0y_0 - f_0) + \alpha p_0 y_0 = 0$$

Второе слагаемое неожиданное! Ведь здесь стоило ожидать простейшее приближение $y'(a) = \alpha y(a)$. Оказывается, что оно компенсирует сдвиг:

$$\begin{aligned} p_{\frac{1}{2}}\frac{y_1 - y_0}{2} &= [py']\left(a + \frac{h}{2}\right) + O(h^2) = \\ &= p(a)y'(a) + \frac{h}{2}(py')'|_a + O(h^2) = \\ &= p(a)y'(a) + \frac{h}{2}\big(q(a)y(a) - f(a)\big) + O(h^2). \end{aligned}$$

§ 5. ВРМ-2 для обычной краевой задачи

Идея ВРМ-2 заключается в том, чтобы «поднять» сеточные функции до каких-нибудь функций из W_2^1 (с помощью некоторого сорта интерполяции), а потом минимизировать функционал на получившемся пространстве.

Будем работать с граничными условиями I типа.

Используем кусочно-линейную интерполяцию:

$$\tilde{y}_{(n)}(x) = \frac{x_{k+1} - x}{h}y_k + \frac{x - x_k}{h}y_{k+1}.$$

Производная определена всюду, кроме узлов:

$$\tilde{y}'_{(n)}(x) = \frac{y_{k+1} - y_k}{h}.$$

Однако узлы — множество меры ноль, поэтому производная всё равно определена, как обобщённая функция. Можно доказать, что это и будет производная в смысле обобщённых функций от восполненной сеточной функции. Поэтому наши восполненные функции находятся в W_2^1.

Можно ввести базисные функции — это восполнения сеточных функций, которые равны нулю всюду, кроме одной точки, а в ней равны единица, т.е.

$$\psi_k(x) = \begin{cases} \dfrac{x_{k+1} - x}{h}, & [x_k,\, x_{k+1}]; \\ \dfrac{x - x_{k-1}}{h}, & [x_{k-1},\, x_k]; \\ 0. \end{cases}$$

Получилось что-то очень похожее на метод Ритца, но только теперь у нас не фиксированный бесконечный набор $\{\varphi_k\}$, а для каждого n есть набор $\{\psi_k\}$ с понятным геометрическим смыслом.

Уравнение для минимизации получится такое же:

$$\sum_{k=1}^{n-1}(\psi_k,\, \psi_m)_A y_k = (f,\, \psi_m),$$

матрица системы — $\{a_{km}\} = (\psi_k,\, \psi_m)_A$.

Носители базисных функций почти не пересекаются, поэтому

$$|k - m| > 1 \ \Rightarrow\ a_{km} = 0.$$

Поэтому система уравнений снова выходит трёхдиагональной:

$$a_m y_{m-1} + b_m y_m + a_{m+1} y_{m+1} = (f,\, \psi_m),$$

где

$$a_m = a_{m-1,\,m} \text{ и } b_m = a_{mm}.$$

На негладких решениях мы не получим точности лучше, чем $O(h)$. Однако этот метод для них надёжнее, чем просто сеточный.

§ 6. Метод Ритца для эллиптического уравнения, энергетическое пространство и естественные условия

Рассмотрим уравнение

$$Lu = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ik}\frac{\partial u}{\partial x_k}\right) + au = f.$$

Коэффициенты должны удовлетворять нескольким условиям:

1. Все функции действуют в области $\Omega \subset \mathbb{R}^n$. В классическом сеттинге решение считается дважды непрерывно дифференцируемыми в Ω и непрерывным на $\overline{\Omega}$; a_{ij} — один раз непрерывно дифференцируемы на $\overline{\Omega}$, а остальные коэффициенты просто непрерывны.
2. Симметричность (эллиптичность): $a_{ik}(x) = a_{ki}(x)$.
3. Положительная определённость:

$$\sum_i \sum_k a_{ik}\xi_i\xi_k \geqslant k^2 \sum_i \xi_i^2.$$

Граничные условия бывают

1. $u|_{\partial\Omega} = 0$ — I типа (задача Дирихле).
2. Пусть

$$\left.\frac{\partial u}{\partial \nu}\right|_{\partial\Omega} = \sum_{i=1}^{n} a_{ij}\cos(n,\, x_i)\left.\frac{\partial u}{\partial x_i}\right|_{\partial\Omega}.$$

   Этот оператор называется *конормальной производной*. Тогда граничное условие выглядит, как

$$\frac{\partial u}{\partial \nu} = \sigma u|_{\partial\Omega}.$$

   Это — задача третьего рода (задача Фон-Неймана).
3. Задача второго рода — когда $\sigma = 0$.

Найдём вид энергетического произведения.

Утверждение 1.

$$(Lu,\, v) = \int\limits_{\Omega}\left(\sum_{i,\,j=1}^{n} a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + auv\right)\partial x + \int\limits_{\partial\Omega}\sigma uv\,\mathrm{d}S.$$

Доказательство.

$$(Lu,\, v) = \int\limits_{\Omega}\left(-\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ik}\frac{\partial u}{\partial x_k}\right) + au\right)v\,\mathrm{d}x$$

Используя формулу интегрирования по частям

$$\int\limits_{\Omega}\frac{\partial u}{\partial x_i}\varphi\,\mathrm{d}x = -\int\limits_{\Omega} u\frac{\partial \varphi}{\partial x_i}\,\mathrm{d}x + \int\limits_{\partial\Omega} u\varphi\cos(n,\, x_i)\,\mathrm{d}S,$$

найдём искомый результат. □

Утверждение 2. Оператор положительно определён, если

1. Задача первого типа: $a(x) \geqslant 0$;
2. Задача второго типа: $a(x) \geqslant a_0 > 0$;
3. Задача третьего типа: $a(x) \geqslant a_0 > 0$, $\sigma(x) \geqslant 0$ или $a(x) \geqslant 0$, $\sigma(x) \geqslant \sigma_0 > 0$.

Доказательство.

$$(Lu,\, u) = \int\limits_{\Omega} \left(\sum_{i,\, j=1}^{n} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} + a u^2 \right) \partial x + \underbrace{\int\limits_{\partial\Omega} \sigma u^2 \, \mathrm{d}S}_{\mathrm{III}}.$$

Нам понадобится неравенство Фридрихса

$$\int\limits_{\Omega} u^2 \, \mathrm{d}x \leqslant c_1 \left(\int\limits_{\Omega} \sum \left(\frac{\partial u}{\partial x_i} \right)^2 \mathrm{d}x + \int\limits_{\partial\Omega} u^2 \, \mathrm{d}S \right).$$

1. Здесь работает совсем грубая оценка:

$$\int\limits_{\Omega} \left(\sum_{i,\, j=1}^{n} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} + a u^2 \right) \partial x \geqslant \int\limits_{\Omega} \sum_{i,\, j=1}^{n} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} \partial x \geqslant k^2 \int\limits_{\Omega} \sum_{i=1}^{n} \left(\frac{\partial u}{\partial x_j} \right)^2 \partial x \geqslant \frac{k^2}{c_1} \int\limits_{\Omega} u^2 \, \mathrm{d}x.$$

2. Ещё проще, как это ни странно.

$$\int\limits_{\Omega} \left(\sum_{i,\, j=1}^{n} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} + a u^2 \right) \partial x \geqslant a_0 \int\limits_{\Omega} u^2 \, \mathrm{d}x$$

3. Первый вариант доказывается точно так же, как для II типа, а второй:

$$\int\limits_{\Omega} \left(\sum_{i,\, j=1}^{n} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} + a u^2 \right) \partial x + \int\limits_{\partial\Omega} \sigma u^2 \, \mathrm{d}S \geqslant k^2 \int\limits_{\Omega} \sum_{i=1}^{n} \left(\frac{\partial u}{\partial x_j} \right)^2 \partial x + \sigma_0 \int\limits_{\partial\Omega} u^2 \, \mathrm{d}S \geqslant \frac{\min(k^2,\, \sigma_0)}{c_1} \int\limits_{\Omega} u^2 \, \mathrm{d}x.$$

□

Замечание 1. Можно доказать, что эта энергетическая норма эквивалентна норме в $W_2^1(\Omega)$:

$$\|v\|^2_{W_2^1} = \int\limits_{\Omega} \left(\sum \left(\frac{\partial u}{\partial x_i} \right)^2 + u^2 \right) \mathrm{d}x.$$

Энергетическое пространство для II и III типов совпадёт с W_2^1, а для типа I унаследует граничное условие и будет состоять из элементов W_2^1, обращающихся в ноль на границе.

Замечание 2. Вообще всё это очень похоже на обычную краевую задачу, только многомерную. При подборе базиса $\{\varphi_k\}$ для метода Ритца в задаче I типа нужно как-то заставить φ_k обращаться в ноль на границе области Ω, которая может быть некрасивой. Чтобы это сделать, можно найти функцию $\omega(x,\, y)$ — это на плоскости — которая положительна в Ω и равна нулю на границе. Читатель сможет придумать такие функции для квадрата/круга/сектора круга, но вообще это, видимо, искусство.

По поводу этого билета стоит заглянуть в книгу [?].

5 Уравения в частных производных

§ 1. Разностный метод для общего уравнения теплопроводности, явная схема

Определение 1. Общее *уравнение теплопроводности* выглядит вот так:

$$\frac{\partial u}{\partial t} = a_0 \frac{\partial^2 u}{\partial x^2} + a_1 \frac{\partial y}{\partial x} + a_2 u + f.$$

Функции a_i и f зависят от x и t.

Работать будем, как всегда, на отрезке $[a,\, b]$; временной отрезок будет $[0,\, T]$.

Определение 2. У уравнения теплопроводности бывает *начальное условие*:

$$u(x,\, 0) = \varphi(x),$$

а также три типа *граничных условий*

1. $u(a,\, t) = \psi_0(t)$, $u(b,\, t) = \psi_1(t)$.
2. $\dfrac{\partial u}{\partial x}(a,\, t) = \psi_0(t)$, $\dfrac{\partial u}{\partial x}(b,\, t) = \psi_1(t)$.
3. $\dfrac{\partial u}{\partial x} - \alpha u\Big|_{x=a} = \psi_0(t)$, $\dfrac{\partial u}{\partial x} - \beta u\Big|_{x=b} = \psi_1(t)$.

Сетка характеризуется такими же, как обычно, величинами:

$$\begin{aligned} x_i &= a + ih, & h &= \frac{b-a}{h}, & i &\in 0 \dots n; \\ t_k &= k\tau, & \tau &= \frac{T}{M}, & k &\in 0 \dots M. \end{aligned}$$

Положим $u_i^k = u(x_i,\, t_k)$ и

$$Lu = a_0 \frac{\partial^2 u}{\partial x^2} + a_1 \frac{\partial y}{\partial x} + a_2 u.$$

Тогда

$$(\tilde{L}u)_i^k = a_0 \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{h^2} + a_1 \frac{u_{i+1}^k - u_{i-1}^k}{2h} + a_2 u_i^k.$$

Есть два варианта для производной по времени:

$$\text{A:} \quad \frac{\partial u}{\partial t}(x_i,\, t_k) \approx \frac{u_i^{k+1} - u_i^k}{\tau}, \tag{5.1}$$

$$\text{B:} \quad \frac{\partial u}{\partial t}(x_i,\, t_k) \approx \frac{u_i^k - u_i^{k-1}}{\tau}. \tag{5.2}$$

Для варианта A получается

$$\boxed{\frac{u_i^{k+1} - u_i^k}{\tau} = \tilde{L}u_i^k + f(x_i,\, t_k)}.$$

Это простейшая явная схема.

В таком виде уравнения можно писать для $i \in 1 \dots n-1$, $k \in 0 \dots M-1$; нужны дополнительные с граничными условиями.

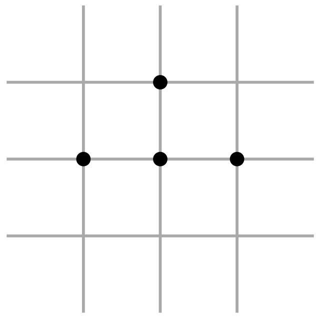


Рис. 5.1: Простейшая явная схема для уравнения теплопроводности.

- Начальные условия: $u_i^0 = \varphi(x_i)$.
- Граничные условия:
  1. $u_0^k = \alpha_1(t_k)$, $u_n^k = \alpha_2(t_k)$; при этом выполняются условия согласования *нулевого порядка*
  $$\varphi(a) = \alpha_1(0), \quad \varphi(b) = \alpha_2(0).$$
  2. Для типов II, III используются такие же трюки, как в обычных диффурах. Надо аппроксимировать производные. Можно применять метод фиктивных точек или метод исключения главного члена погрешности.
  В угловых точках снова возникнет два разных условия:
  $$u_0^0 = \varphi(a) \text{ и } \frac{\partial u}{\partial x}(a,\, 0) = \beta_1(0)u_0^0 + \alpha_1(0).$$
  Будет ли выполняться равенство
  $$\varphi'(a) = \beta_1(0)u_0^0 + \alpha_1(0)?$$
  Оно называется *условием согласования I порядка*. Без него уравнения не станут формально противоречивы.

Если разрешить уравнения относительно u_i^{k+1}, получится
$$u_i^{k+1} = A_i^k u_{i-1}^k + B_i^k u_i^k + C_i^k u_{i+1}^k + D_i^k.$$
Коэффициенты выражаются по формулам
$$\begin{aligned} A_i^k &= \sigma a_0 - \sigma a_1 \frac{h}{2}, & C_i^k &= \sigma a_0 + \sigma \frac{h}{2} a_1, \\ B_i^k &= 1 - 2\sigma a_0 + \tau a_2, & D_i^k &= \tau f(x_i,\, t_k), \end{aligned}$$
где $\sigma = \dfrac{\tau}{h^2}$.

Можно просто двигаться вперёд по *слоям* — множествам точек с постоянным временем; значения находятся последовательно.

§ 2. Неявная схема для уравнения теплопроводности

Неявная схема получается, если в 5.1 выбрать вариант B. Сверху вниз (т.е. назад по времени) просчитать не получится, поскольку начальные данные даются в начале, а не в конце.

Формулы получатся такие:
$$A_i^k u_{i-1}^k - B_i^k u_i^k + C_i^k u_{i+1}^k = D_i^k,$$
где
$$\begin{aligned} A_i^k &= \sigma a_0 - \sigma a_1 \frac{h}{2}, & C_i^k &= \sigma a_0 + \sigma \frac{h}{2} a_1, \\ B_i^k &= 1 + 2\sigma a_0 - \tau a_2, & D_i^k &= -u_i^{k-1} - \tau f(x_i,\, t_k), \end{aligned}$$

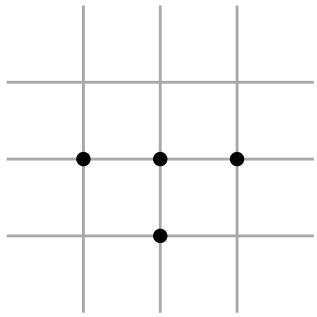


Рис. 5.2: Неявная схема для уравнения теплопроводности.

и $\sigma = \dfrac{\tau}{h^2}$.

По сути, движение всё ещё послойное. Но на каждом слое я не могу просто посчитать значение, используя три значения с предыдущего слоя: наоборот, получается уравнение, которое связывает три значения с текущего слоя с одним уже известным. В итоге получается система с трёхдиагональной матрицей, которая замыкается добавлением граничных условий:

$$u_0^k = \alpha_1(t_k),\ u_n^k = \alpha_2(t_k),$$

если они заданы по первому типу, в противном случае применяются стандартные аппроксимации.

Система решается методом разностной прогонки § 6.

Кажется, тут утверждается, что метод прогонки срабатывает, поскольку

$$A_i^k + C_i^k = 2\sigma a_0 = B_i^k + \tau a_2 - 1,$$

и можно сослаться на 1.7.1. Наверное, на практике это правда так, потому что $\tau a_2 \ll 1$, но выглядит сомнительно, я чего-то не понял.

§ 3. Явная схема для простейшего уравнения теплопроводности, решение разностных уравнений, неустойчивость

Рассмотрим уравнение

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial t^2}, \quad t \in [0, \pi].$$

с начальным условием $u(x, 0) = \varphi(x)$ и граничными условиями

$$u(0, t) = u(\pi, t) = 0$$

Для него разностные уравнения исключительно просты:

$$\frac{u_l^{k+1} - u_l^k}{\tau} = \frac{u_{l+1}^k - 2u_l^k + u_{l-1}^k}{h^2}, \quad u_0^k = u_n^k = 0.$$

При этом

$$h = \frac{\pi}{n}, \quad x_l = lh, \quad u_l^0 = \varphi(x_l).$$

Решим наше разностное уравнение методом разделения переменных, будем искать решение в виде

$$u_l^k = \lambda^k e^{imx}, \quad x = x_l = lh.$$

Подставим:

$$\frac{\lambda^{k+1} e^{imx} - \lambda^k e^{imx}}{\tau} = \frac{\lambda^k e^{im(x+h)} - 2\lambda^k e^{imx} + \lambda^k e^{im(x-h)}}{h^2}.$$

Несложными выкладками отсюда находится

$$\lambda = 1 + 2\sigma\big(\cos mh - 1\big), \quad \sigma = \frac{\tau}{h^2}.$$

Но такое решение не удовлетворяет граничным условиям; можно рассмотреть какую-нибудь комбинацию решений! Заметим, что $\lambda(m)$ — чётная функция, поэтому

$$\lambda^k(m)\big(e^{imx} - e^{-imx}\big) = 2i\lambda^k(m)\sin(mx)$$

тоже решение. Оно удовлетворяет граничным условиям при целых m; в итоге получаем

$$\boxed{u_l^k = \lambda^k(m)\cdot\sin(mx), \quad m \in \mathbb{Z}}.$$

У нас теперь есть $n-1$ ЛНЗ решение, из которых можно собирать новые:

$$\varphi(x) = \sum_{m=1}^{n-1} C_m\lambda^k(m)\cdot\sin(mx).$$

Замечание 1. Остальные значения m нам не интересны, поскольку у нас набралась $n-1$ базисная функция: действительно, изначально наши разностные уравнения решались однозначно, а сейчас мы их решили, учитывая граничные условия, но отпустив начальные. А их как раз $n-1$ — от u_1^0 до u_{n-1}^0, они и создают все степени свободы.

Рассмотрим $\tau = h^2 \Rightarrow \sigma = 1$:

$$\lambda(m) = -1 + 2\cos mh.$$

При $m = n-1$ и густой сетке (большом n)

$$\cos\frac{(n-1)\pi}{n} = \cos\Big(\pi - \frac{\pi}{n}\Big) \approx -1 \Rightarrow \lambda(n-1) \approx -3!$$

При увеличении k решение

$$(-3)^k\sin(n-1)x$$

очень быстро растёт по модулю и всё время меняет знак. Кажется, что-то пошло не так!

Замечание 2. Реальное решение такой задачи — быстро убывающая колебашка. Конечно, пространственный шаг взят большим: у начальных данных есть переменность на том же масштабе. Однако то, что при уменьшении шага по времени k получает возможность становиться больше, уже вообще ни в какие ворота не лезет.

Чтобы решения не вымирали подобным образом, можно наложить ограничение $|\lambda| \leqslant 1$:

$$1 + 2\sigma\big(\cos mh - 1\big) \leqslant 1 \Rightarrow 2\sigma\cos mh - 1 \leqslant 0 \Rightarrow 2\sigma\cos mh \leqslant 1.$$

Чтобы это выполнялось при любых m, нужно, чтобы

$$\boxed{\sigma \leqslant \frac{1}{2} \Leftrightarrow \tau \leqslant \frac{h^2}{2}}.$$

Если точно так же решить разделением переменных систему уравнений для простейшей неявной схемы, получим

$$\lambda = \frac{1}{1 + 2\sigma(1 - \cos mh)} \leqslant 1,$$

и устойчивость всегда присутствует.

§ 4. Общее определение устойчивости, теорема об устойчивости и сходимости

В начале книги [?] есть общие рассуждения про вычислительные методы и всякие пространства. В книге [?] есть про устойчивость, аппроксимацию, сходимость, их связь между собой, и про схемы для уравнения теплопроводности.

С какой ситуацией мы сталкиваемся, занимаясь сеточными методами? У нас есть оператор $A\colon U \to F$, и мы решаем уравнение вида

$$Au = f.$$

Выбирая сетку с шагом h на отрезке, мы вместо функций на отрезке начинаем рассматривать функции на самой сетке они образуют другое, гораздо более маленькое пространство U_h. При этом по любому элементу U можно легко найти элемент U_h, просто вычислив его значения на сетке. Аналогично строится пространство F_h[1].

Наконец, есть оператор A_h: $U_h \to F_h$ — приближение A, которое получается при переходе к конечным разностям. Для иллюстрации полезна диаграмма

$$\begin{tikzcd} U \arrow[r, "A"] \arrow[d, "\varphi_h"'] & F \arrow[d, "\psi_h"] \\ U_h \arrow[r, "A_h"] & F_h \end{tikzcd}$$

Определение 1. Операторы $\varphi_h(u)(x_l) = u(x_l)$ и такой же ψ_h называются *операторами (простого) сноса*.

Замечание 1. Понятно, что диаграмма должна быть почти коммутативна, но не совсем: если мы сначала продифференцируем функцию, а потом возьмём результат на сетке, и если мы сначала возьмём её на сетке, а потом посчитаем разностный аналог производной, получатся близкие, но разные вещи. Разность

$$A_h\varphi_h(u) - \psi_h(Au)$$

называется *естественной погрешностью метода*.

Далее, записывается разностное уравнение

$$A_h\tilde{u} = \psi_h(f)$$

и решается.

Во всех четырёх пространствах надо ввести нормы. В пространствах функциональной природы U, F они уже и так есть, вероятно.

Определение 2. Говорят, что норма на U_h *согласована* с нормой на U, если верно, что

$$\|\varphi_h u\|_{U_h} \to \|u\|_U,$$

когда $h \to 0$ хотя бы для $u \in K \subset U$, где K плотно в U.

Будем считать, что у нас нормы согласованы.

Определение 3. Говорят, что A_h *аппроксимирует* A на $u \in U$, если

$$\left\|A_h\varphi_h(u) - \psi_h(Au)\right\| \to 0 \text{ при } h \to 0.$$

Определение 4. Говорят, что сеточные функции u_h сходятся к функции $u \in U$, если

$$\left\|u_h - \varphi_h(u)\right\| \to 0 \text{ при } h \to 0.$$

Определение 5. Говорят, что *сеточное приближение* обладает *свойством аппроксимации*, если A_h аппроксимирует A, и сеточные функции f_h сходятся к f.

Определение 6. Говорят, что сеточное приближение *устойчиво*, если

1. Уравнение $A_h u_h = f_h$ однозначно разрешимо для всех $f_h \in F_h$;
2. Для этого решения $\|u_h\| \leqslant k\|f_h\|$, где k не зависит от h.

Теорема 1 (Основная теорема теории разностных методов). Пусть дана некоторая краевая задача, и сеточная аппроксимация удовлетворяет следующим свойствам:

1. u^* — единственное решение уравнения $Lu^* = f$.
2. Сеточное приближение обладает свойством аппроксимации.
3. Сеточная задача устойчива.

[1]Зачастую $U_h = F_h$ и даже $U = F$, но может быть и не так, в принципе — вдруг, например, оператор действует в пространство функций на другом отрезке, или просто там другие ограничения на гладкость/непрерывность.

Тогда есть сходимость сеточных решений: $u_h^* \to u$.

Доказательство. Запишем ошибку сеточного решения:

$$w_h = u_h^* - \varphi_h u^*.$$

Заметим, что по свойству устойчивости

$$\begin{aligned}\|w_h\| \leqslant k\|L_h w_h\| = k\|L_h u_h^* - L_h\varphi_h u^*\| = \\ = k\|f_h - \psi_h f + \psi_h f - L_h\varphi_h u^*\| \leqslant k\|f_h - \psi_h f\| + k\|\psi_h L u^* - L_h\varphi_h u^*\|.\end{aligned}$$

Оба слагаемых в правой части стремятся к нулю по свойству аппроксимации. □

§ 5. Разностные схемы для задач с начальными условиями, дискретное преобразование Фурье

Замечание 1. ПРОИЗОШЛО СЖАТИЕ С ПОТЕРЯМИ ШАКАЛОМЕТР ЗАШКАЛИВАЕТ БИП

В этом параграфе в целом посмотрим на уравнение

$$\frac{\partial u}{\partial t} = Lu + f,$$

где всё многомерное (т.е. u — вектор, а L — «матрица» из частных производных), и L — линейный дифференциальный оператор с постоянными коэффициентами. Область определения u — цилиндр $D \times [0, T] \subset \mathbb{R}^{p+1}$. Начальные условия — $u(x, 0) = \varphi(x)$.

Ограничимся теперь ситуацией, когда D — куб $[0, 2\pi]^p$, а граничные условия периодические по каждой из переменных (т.е. $u(x_1, \dots, 0, \dots, x_p; t) = u(x_1, \dots, 2\pi, \dots, x_p; t)$).

По каждой из пространственных переменных выберем одинаковые шаги

$$h = \frac{2\pi}{N},$$

а по временной — шаг

$$\tau = \frac{T}{M}.$$

По пространственной причём рассматриваем только от 0 до $M-1$, потому что справа снова будет то же граничное значение. Уравнения будут двухслойными, с k-го и $k+1$-го слоя.

Даже в неявном случае с помощью разностной прогонки можно выразить все следующие слои через предыдущие и получить уравнения

$$u_h(k+1) = R_h u_h(k) + \rho_h(k),$$

где R_h — *оператор перехода в однородном случае*. R_h — просто матрица с постоянными коэффициентами, а $\rho_h(k)$ зависит от f.

Замечание 2. Всё-таки скажу про эти обозначения пространств... V_h — пространство сеточных функций *на фиксированном слое* (т.е. оно N-мерное), а F_h — видимо, аналогичное пространство, которое мы отличаем только по формальным причинам, в котором лежат f_h — сеточные версии f. Нормы в обоих пространствах — просто l^∞, т.е.

$$\left\|\{u_i\}\right\| = \max|u_i|.$$

Теорема 1. Для устойчивости при $f = 0$ необходимо и достаточно, чтобы были ограничены $\|R_h^k\|$ (здесь k — степень!) при $k\tau \leqslant T$.

Доказательство. Оно в целом понятно: когда нет f-ок, нет и ρ-шек, а без них переход на следующий слой — тупо умножение на матрицу R_h. Ясно, что если нормы этих матриц в совокупности ограничены, то и

$$\|u_h(k+1)\| \leqslant C\|\varphi\|,$$

где φ задаёт начальные условия.

Обратно тоже понятно: если ограниченности норм матриц нет, можно просто пойти от противного и сконструировать мерзкую последовательность. □

Теорема 2. Если $f \neq 0$ и $\|R_h^k\|$ ограничены, то для устойчивости достаточно, чтобы

$$\|\rho_h\|_{V_h} \leqslant c_2 \tau \|f_h\|_{F_h}$$

Доказательство. Обычная оценка, см. конспект Ангелины, страница 75. □

Следствие 1. Если $\|R_h\| \leqslant 1 + c_3\tau$, то есть устойчивость (при $f = 0$).

Доказательство.

$$\|R_h^k\| \leqslant \|R_h\|^k \leqslant (1 + c_3\tau)^k \leqslant e^{c_3\tau k} \leqslant e^{c_3 T}.$$

□

По поводу этих теорем можно ещё заглянуть в следующий параграф § 6, там доказаны очень похожие вещи.

Перейдём теперь к дискретному преобразованию Фурье. Пусть размерность p пока равна 1. Введём на пространстве V_h функций на фиксированном слое скалярное произведение:

$$(u_h, v_h) = h \sum_{i=0}^{N-1} u_h \overline{v_h}$$

Утверждение 1. Набор функций $e_m(x) = e^{imx}$, где $m \in 0 \dots N-1$, образует ортогональный базис в V_h, причём

$$(e_m, e_m) = 2\pi.$$

Доказательство. Чтобы увидеть, что они ортогональны, достаточно посчитать скалярное произведение. Отсюда следует, в принципе, что они ЛНЗ. Ну а дальше — их N, пространство N-мерное, потому и базис. □

Определение 1. *Обратное дискретное преобразование Фурье* —

$$\{a_1, \dots, a_N\} \mapsto \sum_{i=1}^{N-1} a_i e_i(x).$$

Прямое ДПФ —

$$u_h \mapsto \frac{1}{2\pi} \big\{(u_h, e_1), \dots, (u_h, e_n)\big\}.$$

Ясно, что это взаимно обратные операторы.

Утверждение 2 (Формула замкнутости). Дискретное преобразование Фурье — почти унитарный оператор, т.е.

$$\left(\sum_{i=1}^{N-1} a_i e_i, \sum_{i=1}^{N-1} b_i e_i\right) = 2\pi \sum_{i=0}^{N-1} a_i \overline{b_i}.$$

Доказательство. Проверяется прямым вычислением. □

Утверждение 3. e^{imx} — собственная функция оператора сдвига

$$T_h u(x) = u(x + h)$$

с собственным числом e^{imh}.

Доказательство. Действительно,

$$T_h e^{imx} = e^{im(x+h)} = e^{imh} e^{imx}.$$

□

Замечание 3. У нас периодические граничные условия, поэтому оператор сдвига может действовать, «переходя» через границу:

$$T_h\{u_0, \dots, u_{N-1}\} = \{u_1, u_2, \dots, u_{N-1}, u_0\}.$$

Можно представлять себе, что индекс i на самом деле меняется от $-\infty$ до ∞, но $u_{i+N} = u_i$. Периодическую функцию можно восстановить, зная её значения внутри периода, вот и здесь так же.

Замечание 4. В такой ситуации любой разумный разностный оператор можно собрать из операторов сдвига. Например, пусть

$$(Du)_i = \frac{u_{i+1} - 2u_i + u_{i-1}}{h}.$$

Это можно переписать просто как

$$Du = \frac{T_h u - 2u + T_{-h} u}{h}.$$

Общая формула, естественно, будет такая:

$$Lu = \sum_{\alpha} c(\alpha) T_h^{\alpha}(u),$$

где $\alpha \in \mathbb{Z}$ — показатель степени.

Тем удобнее будет применять эти операторы к экспонентам — от них ведь сдвиг считать легко.

Это всё можно написать и в многомерии, при $p > 1$, но, кажется, У Оли этого нет. См. конспект Ангелины, билет и так длинный.

Применим теперь построенную теорию к сеткам. Сеточное уравнение будет выглядеть примерно так:

$$\sum_{\alpha \in A_0} A(\alpha) u(x + \alpha h,\, k\tau) = \sum_{\beta \in B_0} B(\beta) u\big(x + \beta h,\, (k+1)\tau\big).$$

Ну, это просто два слоя, k-й и $k+1$-й. Теперь сделаем ДПФ, пусть

$$u(x,\, k\tau) = \sum a^k(m) e^{imx}.$$

Подставим это в суммы:

$$\sum_{\alpha \in A_0} e^{im\alpha h} A(\alpha) \sum_m a^k(m) e^{imx} = \sum_{\beta \in B_0} e^{im\beta h} B(\beta) \sum_m a^{k+1}(m) e^{imx}.$$

Слева и справа написаны два разложения по базису, коэффициенты в которых должны совпадать:

$$a^k(m) \sum_{\alpha \in A_0} e^{im\alpha h} A(\alpha) = a^{k+1}(m) \sum_{\beta \in B_0} e^{im\beta h} B(\beta)$$

В итоге получаем

$$a^{k+1}(m) = c(m) a^k(m), \quad c(m) = \frac{\sum\limits_{\alpha \in A_0} e^{im\alpha h} A(\alpha)}{\sum\limits_{\beta \in B_0} e^{im\beta h} B(\beta)}.$$

§ 6. Необходимое условие устойчивости по фон Нейману

Замечание 1. Коэффициент c — по сути, видимо, диагональная матрица. Это матрица перехода между слоями в терминах коэффициентов Фурье.

Я не совсем понял, где в конспекте проходит грань между матрицей и числом (всё усложняется тем, что в высших размерностях m — мультииндекс, а u^k и c, видимо, «тензоры»). Поэтому я буду исходить из того, что $c(m)$ — просто число, а c — набор этих чисел, причём

$$\|c\| = \|c\|_{l^\infty} = \max_m c(m).$$

Ну и вообще, пусть все доказательства будут одномерными.

Теорема 1. Для устойчивости при $f = 0$ необходимо и достаточно, чтобы

$$\|c^k\| \leqslant c_3, \quad k\tau \leqslant T.$$

Доказательство. Интересно, видимо, доказывать достаточность. Попробуем просто найти оценку на норму $u_h(k)$.

$$a(k,\, m) = c^k(m)a(0,\, m),$$

поэтому

$$u_h(k) = \sum_m a(k,\, m)e^{imx} = \sum_m a(0,\, m)c^k(m)e^{imx}.$$

Далее K — произвольная неотрицательная константа, в которую можно вносить другие. По формуле замкнутости

$$\Big\|u_h(k)\Big\|^2_{l^2} = 2\pi \sum_m |a(0,\, m)c^k(m)|^2 \leqslant$$
$$\leqslant K \max_m |a(0,\, m)|^2 |c^k(m)|^2 \leqslant K \max_m |a(0,\, m)|^2.$$

Последний переход возможен, поскольку $\|c^k\| \leqslant c_3$. При этом

$$\max_m |a(0,\, m)|^2 = \Big(\max_m |a(0,\, m)|\Big)^2 = \|a^0\|^2_{l^\infty} \leqslant$$
$$\leqslant K\|a^0\|^2_{l^2} = K\Big\|u_h(0)\Big\|^2_{l^2} \leqslant K\Big\|u_h(0)\Big\|^2_{l^\infty}.$$

В итоге получаем

$$\Big\|u_h(k)\Big\|^2_{l^\infty} \leqslant K\Big\|u_h(k)\Big\|^2_{l^2} \leqslant K\Big\|u_h(0)\Big\|^2_{l^\infty}.$$

Это и есть устойчвость, по сути. Чтобы доказать необходимость, предположим, что нет такой оценки $\|c^k\| \leqslant c_3$, не зависящей от h и τ. Рассмотрим $u_h(0) = e^{imx}$. Тогда

$$a(0,\, l) = \delta_{ml} \;\Rightarrow\; u_h(k) = c^k(m)e^{imx}.$$

По предположению мы можем так подобрать $h,\, \tau,\, m,\, k$, что $|c^k(m)|$ станет сколь угодно большим; но тогда это произойдёт и с $\|u_h(k)\|$! Понятно, что никакой устойчивости нет и в помине. □

Теорема 2 (условие фон Неймана). Для устойчивости при $f = 0$ необходимо и достаточно, чтобы собственные числа c удовлетворяли условию

$$|\lambda| \leqslant 1 + c_4\tau.$$

Доказательство. Ну, достаточность не слишком сложна:

$$\|c^k\| = \max_m |c(m)|^k \leqslant |1 + c_4\tau|^k \leq e^{kc_4\tau} \leqslant e^{c_4 T}.$$

Необходимость, впрочем, тоже. Пусть этого условия нет; тогда для любого c_4 можно подобрать такие h, τ и m, что $c(m) > 1 + c_4\tau$. Но тогда

$$\|c^k\| \geqslant |c(m)^k| \leqslant (1 + c_4\tau)^k.$$

Ясно, что увеличивая c_4, можно неограниченно увеличивать $\|c^k\|$. □

Замечание 2. Кажется, в многомерии это условие только необходимое, но я не очень понимаю, почему.

§ 7. Простейшие схемы для уравнения бегущей волны

Определение 1. *Уравнение бегущей волны* имеет вид

$$\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x}.$$

Замечание 1. Любая функция вида

$$u(x,\, t) = f(x + at)$$

является решением.

Замечание 2. Видимо, мы будем работать с периодическим граничными условиями I типа, как и с простейшим уравнением теплопроводности.

Запишем разностные уравнения:

$$\frac{u_l^{k+1} - u_l^k}{\tau} = a\frac{u_{l+1}^k - u_l^k}{h}.$$

Это обычная, явная схема. Чтобы исследовать устойчивость, найдём матрицу перехода. Для этого подставим $u_l^k = e^{imx}$ и $u_l^{k+1} = c(m)e^{imx}$:

$$\frac{c(m)e^{imx} - e^{imx}}{\tau} = a\frac{e^{imx}e^{imh} - e^{imx}}{h}.$$

Отсюда быстро находим

$$\boxed{c(m) = 1 + a\sigma(e^{imh} - 1), \quad \sigma = \frac{\tau}{h}}. \tag{5.1}$$

Нужно понять, когда $|c(m)| \leqslant 1$.

Замечание 3. Этого будет достаточно, но мне не очень понятно, почему не проверяется более точное условие с $1 + c\tau$... Впрочем, в данном конкретном случае это, видимо, то же самое.

Утверждение 1.

1. Если $a < 0$, то $|c| > 1$.

2. Если $a > 0$ и $|a\sigma| \leqslant 1$, то $|c| < 1$.

Доказательство. Вменяемое доказательство требует рисовать много картинок, поэтому обратитесь к конспекту Ангелины. Если кратко, то

1. e^{imh} пробегает единичную окружность.

2. $e^{imh} - 1$ пробегает единичную окружность с центром в -1 (правой стороной она касается нуля).

3. Умножение на $a\sigma$ либо просто растягивает (относительно нуля), либо растягивает и переворачивает.

4. Когда $a < 0$, переворачивает, и получается окружность справа от нуля. После прибавления 1 — окружность справа от единицы, фэйл.

5. Если $a > 0$, получается окружность слева от единицы, которая через неё проходит, радиуса $a\sigma$. Логично, что этот радиус можно увеличить до 1 — тогда центр будет в нуле. А дальше нельзя.

□

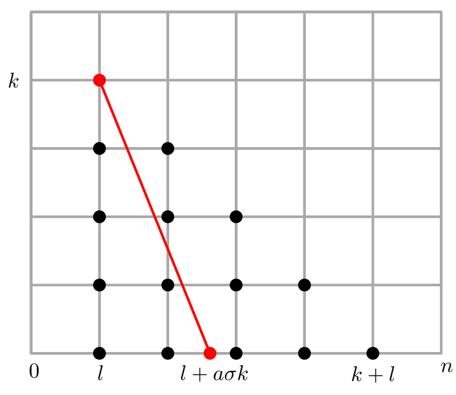


Рис. 5.3: К замечанию 5.7.4.

Замечание 4. Заметим, что u постоянна на прямой $x + at = const$. Будем пока считать $a > 0$. Рассмотрим значение u_l^k. Оно должно определятся соответствущим значением на прямой $t = 0$:

$$lh + ak\tau = sh \;\Rightarrow\; s = l + a\sigma k.$$

С другой стороны, в нашей схеме значение u_l^k определяется u_l^{k-1} и u_{l+1}^{k-1}. Если продолжить этот процесс до $k = 0$, увидим, что u_l^k зависит лишь от $u_l^0 \dots u_{l+k}^0$. Таким образом, чтобы вообще использовать нужное значение из начальных данных, надо

$$s \leqslant l + k \;\Rightarrow\; l + a\sigma k \leqslant l + k \;\Rightarrow\; a\sigma \leqslant 1.$$

Если $a < 0$, то мы вообще не будем использовать это значение, ибо красная прямая будет наклонена в другую сторону.

Именно с этим связано то, что для $a < 0$ срабатывает схема

$$\frac{u_l^{k+1} - u_l^k}{\tau} = a\frac{u^k - u_{l-1}^k}{h}.$$

В ней информация распространяется в другую сторону, и оценки получаются те же с точностью до знака a.

§ 8. Схема Куранта-Рисса

Рассмотрим теперь систему

$$\frac{\partial u}{\partial t} = A\frac{\partial u}{\partial x},$$

где u — вектор из $\mathbb{R}^p$. Будем считать, что A — симметричная матрица с постоянными коэффициентами (поэтому у неё все собственные числа вещественны).

Собственные числа матрицы A могут быть разных знаков, это приводит к появлению решений-волн, которые бегут в разные стороны. Это причина, по которой ни одна из простейших схем, вероятно, не будет работать.

Рассмотрим две матрицы: A_+ и A_-. Первая из них имеет положительные собственные числа такие же, как у A, а вместо отрицательных у неё нули. A_- вместо отрицательных собственных чисел A имеет их модули, а вместо отрицательных — нули.

Замечание 1. Можно эти две матрицы построить в собственном базисе A, а потом вернуть их оттуда назад.

В итоге $A = A_+ - A_-$. Схема будет устроена так:

$$\frac{u_l^{k+1} - u_l^k}{\tau} = A_+\frac{u_{l+1}^k - u_l^k}{h} - A_-\frac{u_l^k - u_{l-1}^k}{h}.$$

Естественно ожидать от неё хорошего поведения. Такая схема называется *схемой Куранта-Рисса*. Найдём матрицу перехода; для этого вычислим её на

$$u_l^k = fe^{imx}, \quad f \text{ — произвольный постоянный вектор.}$$

Такой же подстановкой, как в прошлом пункте, находим

$$c(m) = I + \sigma\Big((e^{imh} - 1)A_+ - (1 - e^{-imh})A_-\Big), \quad \sigma = \frac{\tau}{h}.$$

Теперь нам надо искать собственные числа «матрицы» C (мы тут всё-таки вляпались в многомерие, но не сильно). Запишем условие на собственные числа:

$$g + \sigma\Big((e^{imh} - 1)A_+ - (1 - e^{-imh})A_-\Big)g = \lambda g.$$

Отсюда

$$\Big((e^{imh} - 1)A_+ - (1 - e^{-imh})A_-\Big)g = \frac{\lambda - 1}{\sigma}g$$

Слева стоит линейная комбинация матриц с известными собственными числами, причём там, где у первой ненулевое собственное число, у второй ноль, и наоборот. Поэтому получаем

$$\lambda = 1 + \sigma\lambda_{A_\pm}(e^{\pm imh} - 1).$$

Очень похоже на 5.1, но только теперь $a = \lambda_{A_\pm}$ и $m = \pm m$.

Ясно, что знак m ни на что не повлияет, и теперь $a > 0$. Второе условие получится аналогичным.

Замечание 2. Мы не доказывали достаточность условия для многомерности, но в этом конкретном случае её можно доказать.

§ 9. Явная схема для уравнения колебаний струны

Рассмотрим уравнение колебаний

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}.$$

Вообще, надо было бы перейти к системе уравнений первого порядка, но мы попробуем найти $c(m)$ не переходя — вдруг получится!..

Схема

$$\frac{u_l^{k+1} - 2u_l^k + u_l^{k-1}}{\tau^2} = \frac{u_{l+1}^k - 2u_l^k + u_{l-1}^k}{h^2}.$$

Положим

$$u_l^k = e^{imx}, \quad u_l^{k+1} = ce^{imx}, \quad u_l^{k-1} = c^{-1}e^{imx}.$$

Выйдет уравнение на c:

$$\frac{c + c^{-1}}{2} = 1 + \sigma^2\big(\cos(mh) - 1\big) = p.$$

Решая, получим

$$c = p \pm \sqrt{p^2 - 1}.$$

Неудивительно, что нашлись два значения для каждого m: на самом деле это собственные числа матрицы перехода, λ_1 и λ_2, потому что система второго порядка!

Заметим, что по теореме Виета (ну или руками)

$$\lambda_1\lambda_2 = 1,$$

поэтому, чтобы была устойчивость, нам нужно, чтобы $|\lambda_1| = |\lambda_2| = 1$. Это условие выполняется, когда корни комплексны:

$$\lambda_{1,2} = p \pm i\sqrt{1 - p^2} \;\Rightarrow\; |\lambda_{1,2}| = 1.$$

Для этого необходимо, чтобы $|p| < 1$. Ещё оно выполняется, когда корни равны. Чтобы они были равны, нужно $p = \pm 1$, поэтому общее условие:

$$|p| \leqslant 1.$$

Посмотрев на формулу для p, поймём, что это гарантированно происходит при

$$\boxed{|\sigma| \leqslant 1}.$$

Замечание 1. У Ангелины есть ещё много про достаточное условие, у нас этого, к счастью, не было. Можно ещё сказать, что при $\sigma = 1$ на самом деле проявляется слабая неустойчивость, но она не влияет на сходимость.

§ 10. Явная и неявная схемы для двумерного уравнения теплопроводности

Определение 1. *Двумерное уравнение теплопроводности*

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Теперь нам понадобится два индекса внизу:

$$u_{np}^k \approx u(nh,\ ph,\ k\tau).$$

Рассмотрим явную схему

$$\frac{u_{np}^{k+1} - u_{np}^k}{\tau} = \frac{u_{n+1\ p}^k - 2u_{np}^k + u_{n-1\ p}^k}{h^2} + \frac{u_{n\ p+1}^k - 2u_{np}^k + u_{n\ p-1}^k}{h^2}. \tag{5.1}$$

Попробуем найти матрицу перехода. Для этого рассмотрим

$$u_{np}^k = e^{imx}e^{ily}, \quad x = nh, \quad y = ph.$$

Делая стандартную подстановку и преобразования, получаем

$$c(m,\ l) = 1 + 2\sigma\big(cos(mh) - 1\big) + 2\sigma\big(\cos(lh) - 1\big), \quad \sigma = \frac{\tau}{h^2}.$$

Чтобы $\big|c(m,\ l)\big| \leqslant 1$ всегда, нужно

$$\boxed{\sigma \leqslant \frac{1}{4}}.$$

Получилось в два раза жёсткое условие, чем для одномерного уравнения!

Посмотрим теперь на простейшую неявную схему, в ней левая часть уравнения 5.1 просто заменится на

$$\frac{u_{np}^k - u_{np}^{k-1}}{\tau}.$$

Тем же путём находим

$$c(m,\ l) = \frac{1}{1 - 2\sigma\big(cos(mh) - 1\big) - 2\sigma\big(\cos(lh) - 1\big)}.$$

Видно, что необходимое условие устойчивости выполнено всегда, как и в одномерном случае.

Поговорим о том, как решать систему уравнений для неявной схемы, которая целиком выглядит вот так:

$$\frac{u_{np}^k - u_{np}^{k-1}}{\tau} = \frac{u_{n+1\ p}^k - 2u_{np}^k + u_{n-1\ p}^k}{h^2} + \frac{u_{n\ p+1}^k - 2u_{np}^k + u_{n\ p-1}^k}{h^2}.$$

Если её переписать, выйдет

$$(1 + 4\sigma)v_{np} - \sigma v_{n+1\ p} - \sigma v_{n\ p+1} - \sigma v_{n-1\ p} - \sigma v_{n\ p-1} = \alpha_{np},$$

где

$$v_{np} = u_{np}^k \text{ и } \alpha_{np} = u_{np}^{k-1}.$$

Граничные условия, как обычно в последнее время, I типа:

$$v_{00} = v_{0N} = v_{N0} = v_{NN} = 0.$$

Поэтому будем рассматривать n и p от 1 до $N - 1$.

Упорядочим элементы v следующим образом:

$$v_{11},\ v_{12},\ \dots,\ v_{1\,N-1},\ v_{21},\ \dots$$

Тогда матрица системы (для примера $N = 4$) будет выглядеть так:

$$\begin{array}{ccc|ccc|ccc}
1+4\sigma & -\sigma & 0 & -\sigma & 0 & 0 & 0 & 0 & 0 \\
-\sigma & 1+4\sigma & -\sigma & 0 & -\sigma & 0 & 0 & 0 & 0 \\
0 & -\sigma & 1+4\sigma & 0 & 0 & -\sigma & 0 & 0 & 0 \\
\hline
-\sigma & 0 & 0 & 1+4\sigma & -\sigma & 0 & -\sigma & 0 & 0 \\
0 & -\sigma & 0 & -\sigma & 1+4\sigma & -\sigma & 0 & -\sigma & 0 \\
0 & 0 & -\sigma & 0 & -\sigma & 1+4\sigma & 0 & 0 & -\sigma \\
\hline
0 & 0 & 0 & -\sigma & 0 & 0 & 1+4\sigma & -\sigma & 0 \\
0 & 0 & 0 & 0 & -\sigma & 0 & -\sigma & 1+4\sigma & -\sigma \\
0 & 0 & 0 & 0 & 0 & -\sigma & 0 & -\sigma & 1+4\sigma
\end{array}$$

В общем случае получается $(2N-1)$-диагональная матрица.

Пусть

$$v_n = (v_{n1}, \ldots, v_{n\,N-1}), \quad \alpha_n = (\alpha_{n1}, \ldots, \alpha_{n\,N-1}).$$

Тогда можно записать систему, как

$$A_n v_{n-1} + B_n v_n + C_n v_{n+1} = \alpha_n,$$

где A_n, B_n и C_n — блоки из соответствующей строки. Дальше можно действовать так же, как обычной прогонкой. Это называется *матричная прогонка*.

К сожалению, обычная прогонка содержит умножения чисел, которые делаются за $O(1)$, и работает $O(N)$, а матричная прогонка содержит умножения матриц, которые делаются за $O(N^3)$, и работает за $O(N^4)$. Долго! Поэтому нам такой неявный метод не подходит.

§ 11. Схема продольно-поперечной прогонки

Нужно сделать систему трёхдиагональной. Естественное желание — вынести часть переменных в правой части уравнения на соседний слой, чтобы сократить количество тех, что входит с текущего. Эта идея приводит к схеме

$$\frac{u_{np}^{k+1} - u_{np}^{k}}{\tau} = \frac{u_{n+1\,p}^{k+1} - 2u_{np}^{k+1} + u_{n-1\,p}^{k+1}}{h^2} + \frac{u_{n\,p+1}^{k} - 2u_{np}^{k} + u_{n\,p-1}^{k}}{h^2}.$$

Если переписать, получится

$$\sigma u_{n-1\,p}^{k+1} - (1+2\sigma)u_{np}^{k+1} + \sigma u_{n+1\,p}^{k+1} = \alpha_{np}.$$

Матрица трёхдиагональная, всё замечательно.

Надо проверить на устойчивость. Обычной техникой получаем

$$c(l,\, m) = \frac{1 - 2\sigma(1 - \cos lh)}{1 + 2\sigma(1 - \cos mh)}.$$

Чтобы эта штука всегда была меньше 1 по модулю, нужно

$$\boxed{\sigma \leqslant \frac{1}{2}}.$$

Только условная устойчивость!

Можно рассмотреть аналогичную схему

$$\frac{u_{np}^{k+1} - u_{np}^{k}}{\tau} = \frac{u_{n+1\,p}^{k} - 2u_{np}^{k} + u_{n-1\,p}^{k}}{h^2} + \frac{u_{n\,p+1}^{k+1} - 2u_{np}^{k+1} + u_{n\,p-1}^{k+1}}{h^2}.$$

У неё свойства примерно такие же, только l и m меняются местами.

Идея: чередовать схемы I и II:

$$2k \xrightarrow{\text{I}} 2k+1 \quad 2k+1 \xrightarrow{\text{II}} 2k+2.$$

Пусть у чётных слоёв будет номер k, у нечётных — $k + 1/2$. Ну и просто пишем последовательно формулы для I сначала, потом для II, и получаем переход

$$k \to k + \frac{1}{2} \to k + 1.$$

Ясно, что при этом коэффициенты перехода перемножатся:

$$C = C_{\text{I}} C_{\text{II}} = \frac{1 - 2\sigma(1 - \cos lh)}{1 + 2\sigma(1 - \cos mh)} \cdot \frac{1 - 2\sigma(1 - \cos mh)}{1 + 2\sigma(1 - \cos lh)}.$$

У этой штуки получается $|c| \leqslant 1$ всегда, наступает абсолютная устойчивость.

Такую схему называют *схемой продольно-поперечной прогонки.*

Замечание 1. К сожалению, если то же самое сделать для трёхмерного уравнения теплопроводности (а там будет три схемы и разбиение слоя на три подслоя), абсолютной устойчивости не выйдет. Это можно проверить прямым вычислением.

Замечание 2. Более подробно все выкладки можно прочитать у Ангелины, там всё понятно. Ещё там рассказывается про *схему расщепления*, которая всегда работает. Потом есть разговоры про то, как избавиться от наших общих ограничений вроде периодичности граничных условий, кубической формы области и постоянных коэффициентов.

§ 12. Задача Дирихле для двумерного эллиптического уравнения, составление разностных уравнений

Определение 1. Рассмотрим уравнение в частных производных 2 порядка

$$Lu = f \, : \, \sum_{i,j} a_{ij} \frac{\partial u}{\partial x_i \partial x_j} + \sum_i a_j \frac{\partial u}{\partial x_i} + au = f$$

Пусть все функции заданы на области $\Omega \subset \mathbb{R}^n$.

Если квадратичная форма, соответствующая L знакоопределена, то

- L называют эллиптическим оператором.
- $Lu = f$ называют эллиптическим уравнением.

Пример 1. Уравнение Пуассона: $L = \Delta$

$$\Delta u = f \, : \, \sum_i \frac{\partial^2 u}{\partial {x_i}^2} = f \qquad a_{ii} = 1 > 0$$

Задачу Коши для такого уравнения не поставить — нету выделенной переменной. Так что будем решать граничные задачи. Пусть $\Gamma = \partial\Omega$

I $u|_{\Gamma} = \varphi$ — задача Дирихле

II $\frac{\partial u}{\partial n}\big|_{\Gamma} = \psi$ — задача Неймана

Утверждение 1. Для эллиптических уравнений работает принцип максимума:

$$\max_{\Omega \cup \partial\Omega} u = \max_{\partial\Omega} u$$

Замечание 1. Для разностных уравнений он работает лишь если нету перекрёстных членов. Надеюсь, дефицита крахмала мы не испытываем и сможем обойтись без доказательства этого славного факта.

Рассмотрим пока $n = 2$. Будем решать эти задачи методом сеток

$$\begin{aligned}
\Omega_h &= \left\{ (ih, kh) \in \overline{\Omega} \right\} \\
\Omega_h^0 &= \left\{ (ih, kh) \wedge ((i \pm 1)\, h, kh) \wedge (ih, (k \pm 1)\, h) \in \overline{\Omega} \right\} \\
\Gamma_h &= \Omega_h \setminus \Omega_h^0
\end{aligned}$$

Как вычислять производные в Ω_h^0:

$$\frac{\partial^2 u}{\partial x^2} \longrightarrow \quad \begin{array}{cccc} & 0 & 0 & 0 \\ k & 1/h^2 & -2/h^2 & 1/h^2 \\ & 0 & 0 & 0 \\ & & i & \end{array} \qquad \frac{\partial^2 u}{\partial y^2} \longrightarrow \quad \begin{array}{cccc} & 0 & 1/h^2 & 0 \\ k & 0 & -2/h^2 & 0 \\ & 0 & 1/h^2 & 0 \\ & & i & \end{array}$$

$$\frac{\partial^2 u}{\partial x\,\partial y} \longrightarrow \quad \begin{array}{cccc} & -h^2/4 & 0 & h^2/4 \\ k & 0 & 0 & 0 \\ & h^2/4 & 0 & -h^2/4 \\ & & i & \end{array}$$

Таким образом, видно что шаблон схемы состоит из 9 точек, а если перекрёстных членов нету, то из пяти. Будем считать что их всё-таки нету. Можно же привести к сумме квадратов.

Запишем разностное уравнение как в первой главе

$$A_{ik}u_{i+1,k} + B_{ik}u_{i-1,k} + C_{ik}u_{i,k} + D_{ik}u_{i,k+1} + E_{ik}u_{i,k-1} = f_{ik} \tag{5.1}$$

Нетрудно выразить и его коэффициенты

$$A_{ik} = \frac{a_{11}}{h^2} + \frac{a_1}{2h} \qquad B_{ik} = \frac{a_{11}}{h^2} - \frac{a_1}{2h}$$
$$C_{ik} = -\frac{2a_{11}}{h^2} - \frac{2a_{22}}{h^2} + a$$
$$D_{ik} = \frac{a_{22}}{h^2} + \frac{a_2}{2h} \qquad E_{ik} = \frac{a_{22}}{h^2} - \frac{a_2}{2h}$$

Поскольку $a_{11}, a_{22} > 0$ (положительно определённый) C_{ik} скорее всего < 0 Для сходимости разностных схем мы обычно требовали диагонального преобладания, а тут это выливается в условие на a

$$|C_{ik}| > |A_{ik} + B_{ik} + D_{ik} + E_{ik}| \Leftarrow a < 0$$

Осталось понять что делать с Γ_h.

1. Простой снос на границу: выбираем $M\ :\ u_{pq} = \varphi(M)$. Точность тут $O(h)$. А нам бы лучше извернуться и сделать $O(h^2)$.

2. Снос с интерполяцией: найдем поточнее точку пересечения с границей.

   Например, между $p-1$ и p

   $$p_1(x) = \frac{x - h(p-1)}{h}\,u_{p,q} + \frac{hp - x}{h}\,u_{p-1,q}$$

   Пусть граница пересекается в $p+d$, тогда у нас получается уравнение на d

   $$\left(1 + \frac{d}{h}\right) u_{p,q} + \frac{d}{h}\,u_{p-1,q} = \varphi(M), \quad M = h(p+d), hq$$

   Вот тут уже $O(h^2)$, первый порядок мы убираем выбором d.

   Можно похожим на снос с интерполяцией метод посчитать вторую производную на граю. По сути меняем шаг в сетке. Но мы этого делать не будем. Надеюсь.

   А вот как решать задачу Неймана мы не разбирали. Вопрос называется задача Дирихле, так пусть кроме неё ничего и не будет

§ 13. Итерационный метод решения сеточной системы

Посмотрим на систему (5.1). Здесь ровно та же проблема, что и в неявной схеме в уравнении теплопроводности: прогоночные коэффициенты стали матрицами. По слоям решать не выйдет: граница какой угодно формы.[1] А делать $O(N^4)$ как-то не хочется.

[1] на задачу коши наверное можно посмотреть как на граничное условие на дне цилиндра.

Говоря об этом уравнении как о линейной системе, мы имеем в виду что собрали одномерный вектор из u_{ik} просто расположив строки друг за другом. Как двумерные массивы в фортране.

Будем использовать итеративные методы

$$Au = f \;\Leftrightarrow\; u = Bu + g \qquad u_{n+1} = Bu_n + g$$

Утверждение 1 (Теорема о сжимающем отображении)**.** Если $||B|| < 1$ итерации выше сходятся.

Замечание 1. Это достаточное условие. Необходимым и достаточным будет $\max|\lambda_B| < 1$

Утверждение 2. $||B||_\infty = \max_i \sum_j |b_{ij}|$

Преобразуем нашу систему к пригодному для итераций виду

$$u_{ik} = -\frac{1}{C_{ik}}\left(A_{ik}u_{i+1,k} + B_{ik}u_{i-1,k} + D_{ik}u_{i,k+1} + E_{ik}u_{i,k-1}\right) - \frac{f_{ij}}{C_{ij}}$$

Граничное условие тоже можно итеративно решать[1]

$$u_{p,q} = -\frac{d}{h+d}\,u_{p-1,q} + \frac{h}{h+d}\,\varphi(M)$$

Докажем, что метод сходится.

□ С граничным условием всё понятно: $\frac{d}{h+d} < \frac{1}{2}$ а вот с серединкой чуть хитрее

1. $a < 0$. Из диагонального преобладания сходу следует что $||B||_\infty < 1$.
2. $a = 1$. Тут уже нужно возиться с более точным критерием.

   Пойдём от противного: пусть $\exists\,\lambda_B\;:\;|\lambda_B| = 1$. Пусть $M = \max|v_{ik}|$, $Bv = \lambda_B v$. Обозначим его $|v_{i_0,k_0}|$ Тогда

   $$M = |\lambda_B|\,|v_{i_0,k_0}| \leqslant \frac{|A_{ik}|}{|C_{ik}|}|v_{i_0+1,k_0}| + \frac{|B_{ik}|}{|C_{ik}|}|v_{i_0-1,k_0}| + \frac{|D_{ik}|}{|C_{ik}|}|v_{i_0,k_0+1}| + \frac{|E_{ik}|}{|C_{ik}|}|v_{i_0,k_0-1}| \leqslant M$$

   Тогда со всей неизбежностью мы получаем что и весь шаблон (крестик) равен M по модулю. Таким способом мы неизбежно дойдём до границы

   $$\frac{d}{d+h}M = M, \qquad \frac{d}{d+h} < \frac{1}{2}$$

   Что-то тут не сходится...

■

§ 14. Анализ сходимости простейшего итерационного метода для модельной задачи.

а тут у меня сел ноут а до зарядки далеко

§ 15. Метод оптимальной верхней релаксации, описание

ы

[1] тут кто-то перепутал знак

А Введение в функциональный анализ

§ 1. Пространства, отображения

Бесконечномерные пространства во многом похожи на конечномерные, но есть и различия. Приведём наглядный пример:

Теорема 1. (Рисса) В бесконечномерном пространстве с нормой единичный замкнутый шар не компактен. [2]

Доказательство. Чтобы доказать, что что-то не компактно, нужно найти там последовательность, у которой нет сходящейся подпоследовательности. Здесь это нетрудно: подойдёт любой счётный ортнормированный набор векторов!

Представьте себе: у вас есть n единичных ортогональных друг другу векторов. Вы можете добавить ещё один, и ещё, и ещё... Конечно, в такой последовательности не выбрать сходящейся. □

В том, что касается линейных отображений, тоже есть тонкости. Мы знаем, что любое линейное отображение конечномерных пространств непрерывно и *ограничено* (т.е. образ единичного замкнутого шара при нём ограничен). В бесконечномерном случае это не так! Однако выполняется такое утверждение:

Утверждение 1. Для нормированных пространств непрерывность и ограниченность линейных отображений равносильны.

В реальности почти все интересные отображения ограничены. Да и у неограниченных слишком плохие свойства, поэтому в большинстве теорем ограниченность предполагается.

Замечание 1. Будем все гильбертовы пространства считать *сепарабельными*. Это по сути равносильно тому, что в них есть счётный базис.

§ 2. Пара фактов про гильбертовы пространства

Замечание 1. В бесконечномерных пространствах не все подпространства замкнуты; в частности, там бывают всюду плотные подпространства (как, например, многочлены в пространстве непрерывных функций). Об этом не стоит забывать.

Оказывается, в гильбертовых пространствах ортогональные дополнения устроены почти так же, как и в конечномерной ситуации.

Утверждение 1. Ортогональное дополнение любого множества является замкнутым линейным подпространством. Если $A \subset H$ — замкнутое линейное подпространство, то $H = A \oplus A^{\perp}$.

Этот факт используется для того, чтобы доказать теорему Рисса: линейные функционалы в гильбертовом пространстве — просто скалярные умножения на какие-то вектора.

Теорема 1 (Рисс). Пусть H — гильбертово пространство. Тогда каждый вектор e задаёт ограниченный функционал $f_e\colon H \to \mathbb{C}$ по правилу $x \mapsto (x, e)$, и каждый ограниченный функционал на H есть f_e для некоторого однозначно определённого вектора $e \in H$. Определённая этим биекция $H \to H^*$ есть сопряжённо-линейный изометрический изоморфизм нормированных пространств.

[2] Верно и обратное утверждение: если в нормированном пространстве единичный замкнутый шар компактен, то оно конечномерно.

§ 3. Спектр оператора

Ещё одно различие, не столь наглядное, но очень важное, связано со *спектром* оператора.

Определение 1. Пусть H — гильбертово пространство, $A\colon H \to H$ — ограниченный оператор. *Спектром* A называют множество таких $\lambda \in \mathbb{C}$, что оператор $A - \lambda I$ необратим.

Понятие спектра тесно связано с собственными числами:

Определение 2. Говорят, что $\lambda \in \mathbb{C}$ — *собственное число* оператора A, если есть такой вектор $v \in H$, что $Av = \lambda v$.

Собственные числа можно охарактеризовать в терминах оператора $A - \lambda I$:

Утверждение 1. λ — собственное число A тогда и только тогда, когда оператор $A - \lambda I$ не инъективен (то есть склеивает какие-то векторы в один).

Доказательство. Пусть λ — собственное число, v — собственный вектор. Тогда $(A - \lambda I)v = 0 = A0$, поэтому оператор не инъективен.

Докажем в обратную сторону. Пусть оператор $A - \lambda I$ не инъективен. Тогда есть вектор из ядра — такой, что $(A - \lambda I)v = 0$, т.е. $Av = \lambda v$. ☐

Отсюда сразу следует утверждение:

Утверждение 2. Для конечномерных пространств спектр и множество собственных чисел — одно и то же.

Доказательство. Как мы знаем,

$$\text{необратимость} \Leftrightarrow \text{неинъективность или несюръективность.}$$

Но в конечномерном случае

$$\text{несюръективность} \;\Rightarrow\; \text{неинъективность.}$$

Это связано с тем, что несюръективный оператор понижает размерность пространства, что вынуждает его склеивать вектора.

Поэтому необратимость либо сразу влечёт неинъективность, либо сначала влечёт несюръективность, а потом уже неинъективность. Отсюда

$$\text{необратимость} \Leftrightarrow \text{неинъективность,}$$

что и требовалось доказать. ☐

В бесконечномерном случае всё не так. Из необратимости неинъективность больше не следует, и у оператора появляются два разных способа быть необратимым:

1. Оператор склеивает векторы.
2. Образ оператора меньше, чем всё пространство.

Поэтому спектр оператора A в бесконечномерном пространстве разбивается на собственные числа и те точки, в которых $A - \lambda I$ не является сюръективным (хоть и векторы не склеивает).

Замечание 1. Это не мифическая ситуация: обычный оператор умножения на координату (т.е. $Af(x) = xf(x)$) в $L^2([a, b])$ не имеет собственных чисел, но его спектр равен всему отрезку!

Когда мы занимались квантовой механикой, мы находили «собственные вектора» — дельта-функции. То, что они на самом деле не функции и в L^2 не лежат — свидетельство описанного феномена!

§ 4. Компактные операторы

Обсудим один класс операторов, очень полезный на практике.

Определение 1. Пусть H — гильбертово пространство, B — единичный замкнутый шар в нём. Оператор $A\colon H \to H$ называют *компактным*, если замыкание множества $A(B)$ компактно.

Замечание 1. На самом деле, компактный оператор переводит любое ограниченное множество в множество с компактным замыканием.

Мы знаем, что даже единичный шар в H не компактен. Это значит, что A — оператор с очень маленьким образом, он сжимает всё пространство во что-то крохотное! Это объясняет простоту (и близость к конечномерию) свойств компактных операторов.

Утверждение 1. Если операторы A_n компактны и $\|A_n - A\| \to 0$, то оператор A компактен.

Следствие 1. Если операторы A_n конечного ранга (т.е. их образы конечномерны), и $\|A - A_n\| \to 0$, то оператор A компактен.

Главный пример компактного оператора — *интегральный оператор*.

Пример 1. Пусть $\square = [a,\, b]\times[a,\, b]$. Рассмотрим оператор A на $L^2\big([a,\, b]\big)$, действующий по правилу

$$Af(x) = \int\limits_a^b K(x, y) f(y)\,\mathrm{d}y,$$

где $K \in L^2(\square)$. Такой оператор называют *интегральным*, а функцию K называют его *ядром*. В принципе, вместо L^2 можно жить в C — пространстве непрерывных функций, но оно не гильбертово.

Утверждение 2. Интегральный оператор компактен.

Почти доказательство. Разложим функцию K по базису (так можно, правда):

$$K(x,\, y) = \sum_{n,\, m=0}^{\infty} c_{nm} e_n(x) e_m(y).$$

Рассмотрим последовательность интегральных операторов A_N с ядрами

$$K_N(x,\, y) = \sum_{n,\, m=0}^{N} c_{nm} e_n(x) e_m(y).$$

Простым преобразованием находим, что

$$A_N f(x) = \sum_{n=1}^{N} \left(\sum_{m=1}^{N} c_{nm} \int\limits_a^b e_m(y) f(y)\,\mathrm{d}y \right) e_n(x).$$

Образ оператора A_N находится внутри линейной оболочки векторов $e_1, \ldots, e_N$! Это значит, что наш оператор A приближается операторами конечного ранга, а потому компактен. □

§ 5. Спектры компактных операторов

Спектр компактного оператора обладает замечательным свойством:

Утверждение 1. Пусть A — компактный оператор. Для любого $\delta > 0$ множество собственных чисел A таких, что $|\lambda| \geqslant \delta$ конечно. Собственное пространство любого $\lambda \neq 0$ конечномерно.

Спектр произвольного самосопряжённого оператора, с другой стороны, обладает такими свойствами:

Утверждение 2.

1. Собственные значения самосопряжённого оператора вещественны.

2. Собственные векторы самосопряжённого оператора, отвечающие разным собственным значениям, ортогональны.

Для операторов, одновременно компактных и самосопряжённых, удаётся доказать вариант *спектральной теоремы* — бесконечномерного аналога утверждения о том, что симметричную матрицу можно привести к диагональному виду:

Теорема 1 (Гильберта-Шмидта). Пусть A — компактный и самосопряжённый оператор в гильбертовом пространстве H. Существует ортогональный базис $\{e_i\}$, состоящий из собственных векторов A.

§ 6. Альтернатива Фредгольма

Определение 1. *Фредгольмовым* называют такой оператор T на гильбертовом пространстве, что $T = I - A$, где A компактен.

Утверждение 1. Сопряжённый к компактному оператор компактен.

Теорема 1 (Альтернатива Фредгольма).

1. Уравнение $T\varphi = f$ разрешимо тогда и только тогда, когда f ортогонально любому решению уравнения $T^*\psi_0 = 0$.

2. Либо уравнение $T\varphi = f$ имеет при любом f ровно одно решение, либо уравнение $T\varphi_0 = 0$ имеет ненулевое решение.

3. Уравнения $T^*\psi_0 = 0$ и $T\varphi_0 = 0$ имеют одно и то же конечное число линейно независимых решений.

Замечание 1. Эту теорему называют *альтернативой*, потому что трудно вынести безальтернативность приближения сдачи вычей. Представьте себе, что вы смотрите на уравнение $T\varphi = f$. Есть два варианта:

1. Уравнение $T\varphi_0$ не имеет ненулевых решений, и ваша задача разрешима единственным способом. Всё прекрасно!

2. Оно их таки имеет, и всё не столь прекрасно.

Пусть вы попали во второй вариант. Снова выбор:

1. f ортогонально всем решениям уравнения $T^*\psi_0 = 0$ (которые теперь уже точно есть по третьему пункту). Тогда ваша задача разрешима, но не одним способом (видимо, их будет бесконечно много).

2. f не такое. Тогда ваша задача неразрешима.

B Обозначения