

методы отчислений

Лектор: Б. А. Самокиш

10.01.2019

версия от 08.01.2019 20:16:44, до экзамена 38 часов

Оглавление

1	Однородные дифференциальные уравнения	3
§ 1	Краевая задача для ОДУ 2 порядка и сведение к задаче Коши	3
§ 2	Метод дифференциальной прогонки	4
§ 3	Метод прогонки для систем ОДУ	5
1.	Метод начальных данных	5
2.	Метод прогонки,	5
§ 4	Ортогональная прогонка	6
§ 5	Разностный метод для краевой задачи 2 порядка	7
1.	Алгоритм	7
2.	Формулы численного дифференцирования	7
3.	Разностное уравнение	7
4.	Граничные условия	8
5.	Составление системы линейных уравнений	9
§ 6	Метод разностной прогонки	9
§ 7	Лемма об оценке для системы разностных уравнений	10
§ 8	Теорема о сходимости разностного метода	11
§ 9	Жёсткие системы ОДУ	12
1.	Методы численного интегрирования ОДУ	12
2.	Жёсткие системы	13
§ 10	Неявные методы Рунге-Кутты	15
2	Методы линейной алгебры	17
§ 1	Устойчивость собственных чисел при возмущении матрицы	17
1.	Решение ЛСУ	17
2.	Поиск собственных чисел	19
§ 2	Теорема Бауэра-Файка	20
§ 3	Устойчивость собственных векторов при возмущении матрицы	20
§ 4	Степенной метод	21
§ 5	Обратный степенной метод	22
§ 6	Двумерные вращения	23
§ 7	Лемма о правиле знаков при исключении	24
§ 8	Метод Гивенса	25
§ 9	Метод Якоби	26
§ 10	Две леммы о факторизации матрицы	27
§ 11	Теорема о сходимости итерированных подпространств	27
§ 12	Треугольно-степенной метод и его сходимость	28
§ 13	Ортогонально-степенной метод	29
§ 14	LR-алгоритм. Практическая реализация	29
§ 15	QR-алгоритм. Практическая реализация	30
3	Интегральные уравнения	31
§ 1	Интегральное уравнение II рода, метод замены ядра на вырожденное	31
§ 2	Метод квадратур для интегрального уравнения	35
§ 3	Вариационный принцип для ограниченного оператора; метод Ритца для интегрального уравнения II рода	36
§ 4	Интегральное уравнение I рода и его некорректность	38

§ 5	Условная корректность по Тихонову, метод квазирешений	39
§ 6	Метод регуляризации для уравнения I рода, сходимость	40
4	Вариационные методы	42
§ 1	Вариационный принцип для уравнения с неограниченным оператором	42
§ 2	Метод Ритца, сходимость	42
§ 3	Метод Ритца для обычной краевой задачи, вид энергетического пространства, естественные граничные условия	43
§ 4	ВРМ-1 для обычной краевой задачи	45
§ 5	ВРМ-2 для обычной краевой задачи	46
§ 6	Метод Ритца для эллиптического уравнения, энергетическое пространство и естественные условия	47
5	Уравнения в частных производных	49
§ 1	Разностный метод для общего уравнения теплопроводности, явная схема	49
A	Введение в функциональный анализ	51
§ 1	Пространства, отображения	51
§ 2	Пара фактов про гильбертовы пространства	51
§ 3	Спектр оператора	52
§ 4	Компактные операторы	53
§ 5	Спектры компактных операторов	53
§ 6	Альтернатива Фредгольма	54
B	Обозначения	55

1 Однородные дифференциальные уравнения

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

Определение 1. Рассмотрим ОДУ 2 порядка

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

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

□ Рассмотрим решение неоднородной краевой в виде $y(x) = y_0(x) + c_1 y_1(x) + c_2 y_2(x)$ и подставим 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 записать, 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

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

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)

$$y'(x) = \alpha(x) y(x) + \beta(x) \quad (1.1)$$

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

$$\begin{aligned} y' &= \alpha y + \beta \\ y'' &= \alpha y' + \alpha' y + \beta' \Rightarrow y'' + p y' + q y = f \\ &\Leftrightarrow \alpha y' + \alpha' y + \beta' + p(\alpha y + \beta) + q y = f \\ &\Leftrightarrow y \underbrace{(\alpha^2 + \alpha' + p\alpha + q)}_0 + \underbrace{(\alpha\beta + \beta' + p\beta)}_f = f \end{aligned}$$

В итоге получаем систему ОДУ первого порядка

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

Посмотрим что происходит на правом конце¹

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

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.

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

§ 3. Метод прогонки для систем ОДУ

Определение 1. Рассмотрим ОДУ

$$\mathbf{y}' = \hat{A}(x)\mathbf{y} + \mathbf{f}(x) \quad \mathbf{y} \in C^2([a; b]), \quad \mathbf{y}: [a; b] \rightarrow \mathbb{R}^s \quad (1.1)$$

и условия вида

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

Если \mathbf{y} — решение для которого выполнено условие выше, то \mathbf{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:

$$\hat{\alpha}\mathbf{y}(a) + \hat{\gamma}\mathbf{y}(b) = \boldsymbol{\beta}, \quad \hat{\alpha}, \hat{\gamma}: \mathbb{R}^s \rightarrow \mathbb{R}^s$$

Общее решение задачи Коши 1.1 имеет вид

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

где как обычно \mathbf{y}_0 — решение неоднородной задачи Коши, а $\{\mathbf{y}_j\}$ — 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 $\mathbf{y}_j(a) = \mathbf{e}_j$ (это единичный вектор с 1 на j ом месте), $\mathbf{y}_0(a) = 0$

Можно снова уменьшить количество работы

1. в качестве начальных данных для \mathbf{y}_0 — какое-нибудь решение системы $\hat{\alpha}\mathbf{y} = \boldsymbol{\beta}$
2. в качестве начальных данных для \mathbf{y}_j , $j \in p + 1 \dots s - q$ линейно независимых решений $\hat{\alpha}\mathbf{y} = 0$
3. $\mathbf{y}_j \equiv 0$, $j \in 1 \dots p$

В итоге решение примет вид

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

<здесь снова этот понятный кусок про экспоненты и беды вычислений> (✖)

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

$$\hat{\alpha}(x)\mathbf{y}(x) = \boldsymbol{\beta}(x), \quad \hat{\alpha}: [a; b] \times \mathbb{R}^s \rightarrow \mathbb{R}^p \quad (1.2)$$

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

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

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

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

$$\begin{aligned} \boldsymbol{\alpha}_j' &= -\hat{A}^T \boldsymbol{\alpha}_j \\ \boldsymbol{\beta}_j' &= (\boldsymbol{\alpha}_j, \mathbf{f}) \end{aligned} \quad (1.3)$$

Посмотрим что происходит на правом конце

$$\begin{cases} \hat{\alpha}(b) \mathbf{y}(b) = \beta(b) \\ \hat{\gamma} \mathbf{y}(b) = \delta \end{cases} \quad (1.4)$$

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

Сам метод выглядит так:

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

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

Замечание 2. (?) 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 крайняя задача из §1 им не решается. Проблема возникает в том месте, где из $(\hat{\alpha}\hat{A} + \hat{\alpha}') \mathbf{y} = 0$ выводится $\hat{\alpha}\hat{A} + \hat{\alpha}' = 0$. Произвольностью \mathbf{y} мы вообще-то пользоваться не можем, так как на него есть условие $\hat{\alpha}(x) \mathbf{y}(x) = \beta(x)$.

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 $\mathbf{y}(b)$ мы потеряем точность.¹
- Собственно, для того чтобы обойти эти проблемы и нужен §4.

§4. Ортогональная прогонка

«будем решать немного другую задачу»

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

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

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

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

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

Получим уравнения на прогоночные коэффициенты

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

В итоге

$$\begin{aligned} \hat{\alpha}' &= -\hat{\alpha}\hat{A} + \hat{\alpha}\hat{A}\hat{\alpha}^T (\hat{\alpha}\hat{\alpha}^T)^{-1} \hat{\alpha} \\ \beta' &= \hat{\alpha}\mathbf{f} + \hat{\alpha}\hat{A}\hat{\alpha}^T (\hat{\alpha}\hat{\alpha}^T)^{-1} \beta \end{aligned} \quad (1.1)$$

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

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

$$\begin{aligned} \hat{\alpha}' &= -\hat{\alpha}\hat{B} + \hat{\alpha}\hat{B}\hat{\alpha}^T \hat{\alpha} \\ \beta' &= \hat{\alpha}\mathbf{g} + \hat{\alpha}\hat{B}\hat{\alpha}^T \beta \end{aligned} \quad (1.2)$$

¹ $\text{rk } \hat{\alpha}\hat{\alpha}^T \geq 2 \text{ rk } \hat{\alpha} - s$ 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 $\hat{\alpha}^T$ 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. 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

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

1. Алгоритм

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

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

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

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

2. Формулы численного дифференцирования

Здесь $M_n = \max |y^{(n)}(x)|$

$$y'(x) = \frac{y(x+h) - y(x)}{h} + R \quad |R| \leq \frac{hM_2}{2} \quad (1.2)$$

$$y'(x) = \frac{y(x) - y(x-h)}{h} + R \quad |R| \leq \frac{hM_2}{2} \quad (1.3)$$

$$y'(x) = \frac{-y(x+2h) + 4y(x+h) - 3y(x)}{2h} + R \quad |R| \leq \frac{h^2M_3}{3} \quad (1.4)$$

$$y'(x) = \frac{3y(x) - 4y(x-h) + y(x-2h)}{2h} + R \quad |R| \leq \frac{h^2M_3}{3} \quad (1.5)$$

$$y'(x) = \frac{y(x+h) - y(x-h)}{2h} + R \quad |R| \leq \frac{h^2M_3}{6} \quad (1.6)$$

$$y''(x) = \frac{y(x+h) - 2y(x) + y(x-h)}{h^2} + R \quad |R| \leq \frac{h^2M_4}{12} \quad (1.7)$$

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

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

Получится

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

¹про 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:

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

2. Метод фиктивных узлов

(а) 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:

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

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.

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

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

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:

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

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

$$\begin{aligned} y''(x) &= -(p(x)y'(x) + q(x)y(x)) + f(x) = -\left(p(x)\frac{y(x+h)-y(x)}{h} + p(x)O(h) + q(x)y(x)\right) + f(x) \\ \Rightarrow -\frac{h}{2}y''(x) &= \frac{p(x)}{2}(y(x+h) - y(x)) + \frac{h}{2}(q(x)y(x) - f(x)) + O(h^2) \end{aligned}$$

Оценка производной на краю

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

На сетке оно имеет вид

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

5. Составление системы линейных уравнений

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

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

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

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.

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

$$y_k = \alpha_k y_{k+1} + \beta_k \quad (1.1)$$

Свяжем α, β с a, b, c, d .

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

Отсюда получаются удобные рекурсивные соотношения для α_k, β_k

$$\alpha_k = \frac{c_k}{b_k - a_k \alpha_{k-1}}; \quad \beta_k = \frac{a_k \beta_{k-1} - d_k}{b_k - a_k \alpha_{k-1}} \quad (1.2)$$

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: $b_k \geq a_k + c_k$.

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 $\alpha_{k-1} \geq 1$.

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

база: $\alpha_0 = \frac{c_0}{b_0}, b_0 \geq c_0 > 0, c_0 \leq b_0$. Кажется, все верно.¹

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

$$b_k - \alpha_{k-1}a_k \geq b_k - a_k \geq c_k > 0 \Rightarrow 0 < \alpha_k = \frac{c_k}{b_k - a_k\alpha_{k-1}} \leq 1$$

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

Таким образом, все знаменатели кроме последнего > 0 . В нём $c_n = 0 \neq 0$.

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 $b_n - a_n\alpha_{n-1} > b_n - a_n \geq 0$

2. $k_0 = n \Rightarrow \alpha_{n-1} \leq 1$. Тогда $b_n - a_n\alpha_{n-1} \geq b_n - a_n > 0$

Как видно, даже последний знаменатель $\neq 0$

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

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

Отсюда

$$a_k > 0 \Leftrightarrow h < \frac{2}{\max |p_k|}$$

$$c_k > 0 \Leftrightarrow h < \frac{2}{\max |p_k|}$$

$$b_k \geq c_k + a_k \Leftrightarrow \boxed{q_k \leq 0}$$

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

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

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

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

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

Замечание 1. Рассмотрим однородную задачу с однородным левым граничным условием ($A = 0$). 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 выполнено условие достаточности прогонки 1.7.1 в усиленном виде:

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

Тогда

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

▼

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

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

◀

§ 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. $q(x) \leq -q_0, q_0 > 0$
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 производную на границе, можно воспользоваться формулой с фиктивными узлами и сдвинутой сеткой (на $h/2$), тогда:

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

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 :

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

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

Подгоним под условия леммы 1.7.1

$$b_k \geq a_k + c_k + \delta \Leftrightarrow -q_k \geq \delta \Leftrightarrow \delta = q_0$$

На границах

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

Снова найдём δ

$$\begin{aligned} b_0 \geq a_0 + c_0 + \delta &\Leftrightarrow \alpha \geq \delta \Leftrightarrow \delta = \alpha \\ b_{n+1} \geq a_{n+1} + c_{n+1} + \delta &\Leftrightarrow -\beta \geq \delta \Leftrightarrow \delta = -\beta \end{aligned}$$

Выберем:

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

Тогда по лемме

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



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

§ 9. Жёсткие системы ОДУ

Будем рассматривать задачу Коши для системы обыкновенных дифференциальных уравнений.

$$\mathbf{y}'(x) = f(x, \mathbf{y}(x)), \quad \mathbf{y}(x_0) = \mathbf{y}_0$$

Пусть x_n — узлы равномерной сетки с шагом h .

1. Методы численного интегрирования ОДУ

1. Метод Эйлера

$$\mathbf{y}_{n+1} = \mathbf{y}_n + hf(x_n, \mathbf{y}_n)$$

2. Неявный метод Эйлера

$$\mathbf{y}_{n+1} = \mathbf{y}_n + hf(x_{n+1}, \mathbf{y}_{n+1})$$

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

Рассмотрим более точные методы

3. Улучшенный метод Эйлера

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2} \left(f(x_n, \mathbf{y}_n) + f(x_{n+1}, \mathbf{y}_n + hf(x_n, \mathbf{y}_n)) \right)$$

4. Метод трапеций

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

5. Метод средних прямоугольников

$$\mathbf{y}_{n+1} = \mathbf{y}_n + hf \left(x_n + \frac{h}{2}, \frac{\mathbf{y}_n + \mathbf{y}_{n+1}}{2} \right)$$

6. Весовая формула трапеций

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

7. Весовая формула прямоугольников

$$\mathbf{y}_{n+1} = \mathbf{y}_n + hf \left(x_n + \theta h, (1 - \theta) \mathbf{y}_n + \theta \mathbf{y}_{n+1} \right), \quad 0 \leq \theta \leq 1$$

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.

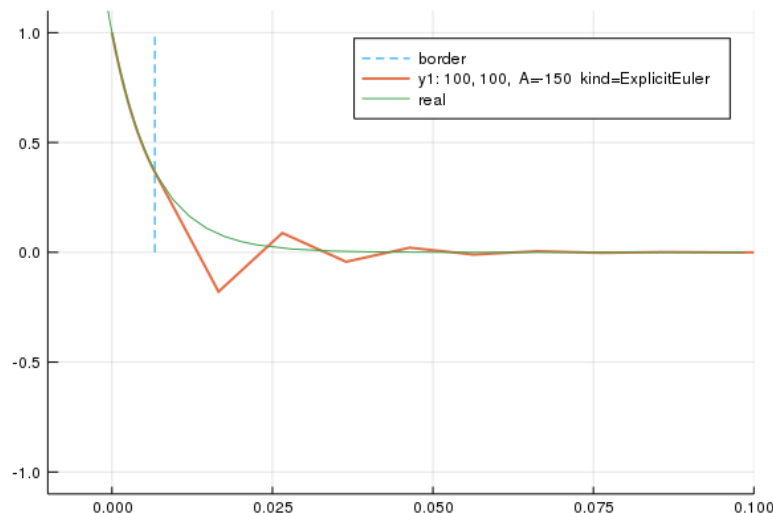
Пример 1. Рассмотрим $f(x, y) = Ay$, $y(0) = 1$ (всё одномерное). Его решение — $y = e^{Ax}$. Попробуем решить методом Эйлера

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

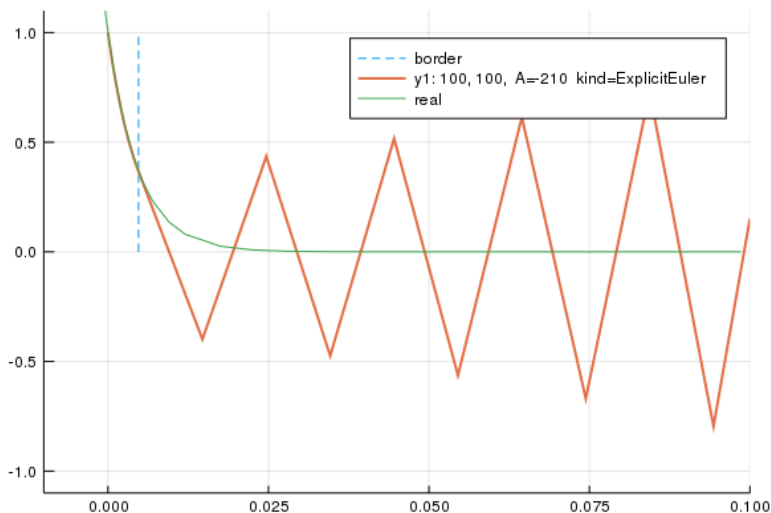
Из определения ϵ при $h \rightarrow 0$ получается истинное решение. Однако, рассмотрим что происходит при $A \ll 0$.

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

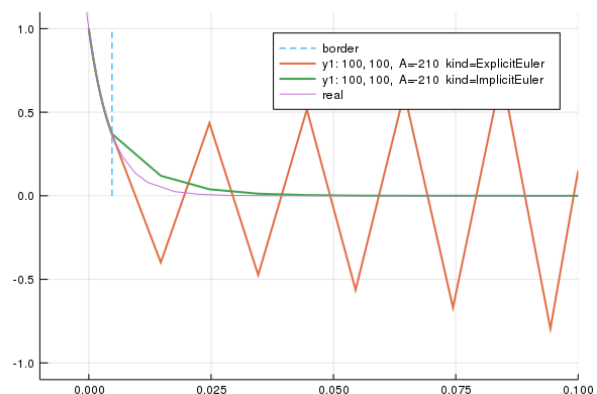
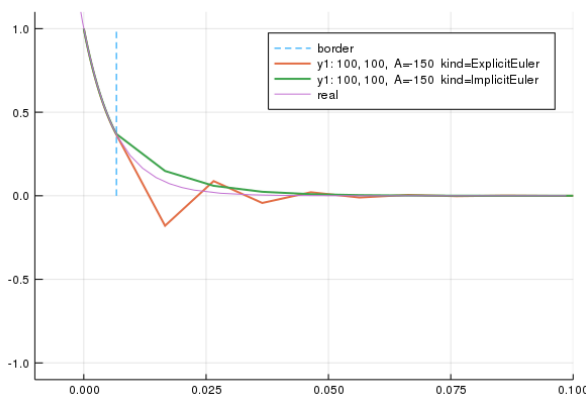


Пример 2. Снова рассмотрим $f(x, y) = Ay$, $y(0) = 1$ (всё одномерное). Его решение — $y = e^{Ax}$. Такую штуку называют пробным уравнением

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.

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

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 $f(x, y) = \hat{A}y$, 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 $|R(z)| \leq 1$ в левой полуплоскости.

Определение 5. $\{z \mid |R(z)| \leq 1\}$ 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 $\lim_{z \rightarrow \infty} R(z) = 0$. 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 $\theta \geq \frac{1}{2}$. Просто при таком преобразовании прообразом единичного круга будет круг/внешность круга с центром в $\theta - \frac{1}{2}$ и радиусом $|\theta - \frac{1}{2}|$.

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 $\alpha_i, \beta_{ij}, \gamma_i$ 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: $\alpha \frac{\beta}{\gamma}$
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)

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

Пример 3. (диагонального неявного РК)

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

<+ещё примеры?+>

2 Методы линейной алгебры

§ 1. Устойчивость собственных чисел при возмущении матрицы

Пусть A — линейный оператор $\mathbb{R}^s \rightarrow \mathbb{R}^s$, 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 \odot — 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»):

$$a \odot b = a \circ b (1 + \varepsilon), \quad \varepsilon \leq \varepsilon_M$$

2. Обратный анализ ошибок (метод эквивалентных возмущений)

Сводим все ошибки к возмущениям начальных данных:

$$a \odot b = \tilde{a} \circ \tilde{b}, \quad \tilde{a} = a + \Delta a, \quad \tilde{b} = b + \Delta b.$$

- (a) оцениваем эквивалентные возмущения
- (b) оцениваем влияние возмущений

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

Первый метод часто выдает неправомерно 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. $\frac{\|Ax\|}{\|x\|}$. Здесь мы пытаемся смотреть на ЛЗ строчек матрицы. Но не очень понятно с чем сравнивать, чтобы понять близость к ЛЗ. Может быть компоненты матрицы маленькие.

3. $\frac{\max_{\|x\|} \frac{\|Ax\|}{\|x\|}}{\min_{\|x\|} \frac{\|Ax\|}{\|x\|}}$ уже выглядит разумно. Преобразуем, используя определение нормы (конечно-мерного) оператора

$$\begin{aligned}\max \frac{\|Ax\|}{\|x\|} &= \|A\| \\ \min \frac{\|Ax\|}{\|x\|} &= \min \frac{\|y\|}{\|A^{-1}y\|} = \frac{1}{\frac{\|A^{-1}y\|}{\|y\|}} = \|A^{-1}\|^{-1}.\end{aligned}$$

А это очень похоже на определение выше.

Лемма 1. $\|B\| < 1 \Rightarrow \exists (I - B)^{-1} \wedge \|(I - B)^{-1}\| \leq \frac{1}{1 - \|B\|}$

▼

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.

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

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

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

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

◀

Теперь, оценим, наконец, погрешность решения СЛУ.

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

$$\begin{aligned}\Delta A &= \tilde{A} - A, \quad \Delta x = x - x^*, \quad \Delta b = \tilde{b} - b \\ \delta_A &= \frac{\|\Delta A\|}{\|A\|}, \quad \delta_x = \frac{\|\Delta x\|}{\|x^*\|}, \quad \delta_b = \frac{\|\Delta b\|}{\|b\|} \\ x^* &\text{ — невозмущенное решение}\end{aligned}$$

Тогда

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

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

$$\begin{aligned}A'x &= b' \Leftrightarrow (A + \Delta A)(x^* + \Delta x) = b + \Delta b \Leftrightarrow (A + \Delta A)\Delta x = -\Delta Ax^* + \Delta b \\ &\Leftrightarrow \left(I - (-A^{-1}\Delta A)\right) \frac{\Delta x}{x^*} = -A^{-1}\Delta A + A^{-1} \frac{\Delta b}{x^*}\end{aligned}$$

Из леммы выше

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

Из невозмущённой системы $\|x^*\| \geq \|A\|^{-1} \|b\|$, 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:

$$J_p(a) = \begin{pmatrix} a & 1 & & \\ & \ddots & \ddots & \\ & & a & 1 \\ & & & a \end{pmatrix} : p \times p, \quad p_{J_p(a)}(t) = (a - t)^p$$

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 \leq алгебраической, ибо $\dim V_\lambda \leq V(\lambda)$.
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 жизнь для линейной алгебры.

Теперь наконец обсудим устойчивость

Пример 1. Пусть

$$A = J_p(a), \quad \varepsilon B : (\varepsilon B)_{ij} = \delta_{ip}\delta_{j1}, \quad \tilde{A} = A + \varepsilon B$$

Оценим ошибку собственного числа. Попробуем преобразовать систему..

$$Ax + \varepsilon Bx = \lambda x \Leftrightarrow \begin{cases} ax_k + x_{k+1} = \lambda x_k, & k \in 1 \dots p-1 \\ \varepsilon x_1 + ax_p = \lambda x_p \end{cases} \Rightarrow \varepsilon x_1 = (\lambda - a)^p x_1$$

В итоге получается, что $\lambda = a + \varepsilon^{1/p}$

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

¹можно 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

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

□ 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

$$A - zI + B = D^{-1}(\Lambda - zI + DBD^{-1})D = D^{-1}(\Lambda - zI) \underbrace{(I + (\Lambda - zI)^{-1}DBD^{-1})}_C D$$

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

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

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

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

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

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

Тогда, как видно

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

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 , $(\mu_i, v_i) — A^*$, 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 приближении¹

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

$$1. \|\Delta\lambda_i\| \leq p_i \|\Delta A\|$$

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

► Пойдём по порядку.

$$1. \lambda_i = \overline{\mu_i}$$

$$2. (u_i, v_k) = 0 \text{ при } i \neq k$$

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

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

$$4. \Delta\lambda_i (u_i, v_i) = (\Delta A u_i, v_i)$$

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

отсюда уже легко вывести первый пункт.

$$5. (\Delta u_i, v_k) = (1 - \delta_{ik}) \frac{(\Delta A u_i, v_k)}{(\lambda_i, \lambda_k)}, \text{ аналогично предыдущему пункту. Здесь выбрали } (\Delta u_i, v_i) = 0 \text{ пожертвовав нормированностью } v_i. \text{ Всё равно одна лишняя степень свободы была.}$$

$$6. \Delta u_i = \sum_{k=1}^s \gamma_k u_k, \gamma_k = \frac{(\Delta u_i, v_k)}{(u_k, v_k)}$$

$$7. \Delta u_i = \sum_{k=1}^s \frac{(\Delta A u_i, v_k)}{(\lambda_i - \lambda_k)} \frac{1}{(u_i, v_k)}, \text{ отсюда очевиден второй}$$

◀

§ 4. Степенной метод

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

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

Будем брать \tilde{x}_{n+1}^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 заберёт все остальные. Однако, нужно аккуратно сформулировать условия сходимости.

Утверждение 1. Пусть для собственных чисел A выполнено условие:

$$|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_s|$$

Тогда степенной метод сходится к λ_1

¹а если нет, то надо думать

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

► Из определения степенного метода

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

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

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

$$\Rightarrow \tilde{x}_{n+1}^1 = \{Ax_n\}^1 \xrightarrow{n \rightarrow \infty} \frac{\{\lambda_1 u_1\}^1}{u_1^1} = \lambda_1$$

Так работает для комплексных λ , поскольку

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

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

$$x_{2n} \rightarrow \frac{c_1 u_1 + c_2 u_2}{c_1 u_1^1 + c_2 u_2^1}, \quad \tilde{x}_{2n+1}^1 \rightarrow \lambda \frac{c_1 u_1^1 - c_2 u_2^1}{c_1 u_1^1 + c_2 u_2^1}$$

$$x_{2n+1} \rightarrow \frac{c_1 u_1 - c_2 u_2}{c_1 u_1^1 - c_2 u_2^1}, \quad \tilde{x}_{2n+2}^1 \rightarrow \lambda \frac{c_1 u_1^1 + c_2 u_2^1}{c_1 u_1^1 + c_2 u_2^1}$$

подпоследовательности сходятся к разным числам

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

$$x_n \rightarrow \frac{2 \operatorname{Re}(c_1 u_1 e^{in\theta})}{2 \operatorname{Re}(c_1 u_1^1 e^{in\theta})}, \quad \tilde{x}_{n+1}^1 \rightarrow R \frac{\operatorname{Re}(c_1 u_1^1 e^{i(n+1)\theta})}{\operatorname{Re}(c_1 u_1^1 e^{in\theta})}$$

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:

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

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



$$A^{-1}x = \lambda_* x \Leftrightarrow \lambda x = \lambda_*^{-1} x = Ax$$

$$|\lambda_1| < |\lambda_2| \leq \dots \leq |\lambda_s| \Leftrightarrow |\lambda_1^{-1}| > |\lambda_2^{-1}| \geq \dots \geq |\lambda_s^{-1}|$$



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} ,

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

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

$$\begin{array}{ll} \text{унитарные} & \rightarrow \text{ортогональные} \\ \text{эрмитовы} & \rightarrow \text{симметричные} \end{array}$$

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: $\begin{bmatrix} 1 \end{bmatrix}$

отражение: $\begin{bmatrix} -1 \end{bmatrix}$

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

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 подпространствам.

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

$$A \rightarrow C = O^T A O, \quad O - \text{ортогональная}$$

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

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

Как можно избавиться от внедиагональных членов:

1. $c_{p-1,q} = 0$: вращение Гивенса

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

2. $c_{p,q} = 0$: вращение Якоби

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

§ 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:

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

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

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

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.

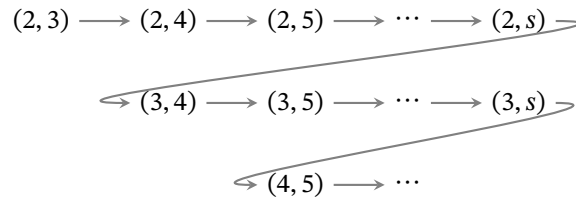
Теперь приведём A к ЖНФ. Поскольку она симметричная, J_A диагональная. На диагонали стоят собственные числа. Сравнивая их с a_{jj} и припоминая закон инерции, приходим к утверждению леммы. ◀

§ 8. Метод Гивенса

Вспомним, как выглядело вращение Гивенса

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

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



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.

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

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 переменна знака была всегда одна. Обычным методом половинного деления как раз можно к этому прийти.

§ 9. Метод Якоби

Вспомним, как выглядело вращение Якоби

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

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

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

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 $C^2 = O^T A O O^T A O = O^T A O$, 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: $p, q : |a_{p,q}| = \max_{i \neq k} |a_{i,k}|$.

Оценим, как быстро он сходится

$$a_{p,q}^2 \frac{s(s-1)}{2} \geq \sum_{i,k} a_{i,k}^2 \Rightarrow t^2(C) \leq \left(1 - \frac{2}{s(s-1)}\right) t^2(A)$$

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 ,

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

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

¹то 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

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

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 j -го столбца. Тогда матрица такого преобразования имеет вид

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

понятно, что в случае k -го шага будет просто k -й столбик. Если мы будем перемножать такие столбики, они просто будут приставаться 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 итерированное 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 $P^{(n)} \rightarrow P$, если в $P^{(n)}$ существует базис, сходящийся к базису P .

Теорема 1. Пусть A — диагонализуемая неособая матрица,

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

Тогда $L_k^{(n)} \rightarrow U_k$.

□ Соорудим базис в $L_k^{(n)}$ который будем сходиться 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 $D = (\tilde{C})^{-1}$, \tilde{C} — квадратный кусок C размерами $k \times k$ из первых коэффициентов.

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

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

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

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

■

Поплодим ещё сущностей

Определение 3 (Ступенчатый базис).

$$e_1 = (1, x_{12}, \dots) \quad (2.1)$$

$$e_k = (0, \dots, 0, 1, x_{k,k+1}, \dots) \quad (2.2)$$

$$(2.3)$$

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 следующих 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.

¹этого условия у нас не было, но без него не доказать невырожденность \tilde{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. $R_n = P_n^{-1}AP_n \xrightarrow{n \rightarrow \infty} P^TAP$, 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 Q . 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 $j > k$

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

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

Вспомним доказательство теоремы про итерируемые подпространства и скажем что x_i отсюда это $C^{(i)}$ сейчас. Тогда, $C^{(i)}$ раскладывают 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

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

§ 14. LR-алгоритм. Практическая реализация

Определение 1. Рассмотрим A со стандартными условиями на собственные вектора, произвольную невырожденную Q . 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

$$AP_0 = P_1 R_1 \quad \rightarrow \quad AI = L_1 R_1$$

$$AP_n = P_{n+1} R_{n+1} \quad \rightarrow \quad AL_1 \cdots L_n = L_1 \cdots L_{n+1} R_{n+1}$$

Разберёмся с $AL_1 \cdots L_n$

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

А подставив такого крокодила как раз получаем $R_n L_n = L_{n+1} R_{n+1}$ Поскольку мы свели метод к предыдущему, 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. Трёхдиагональную матрицу можно факторизовать (прогонкой:) за $O(s)$ на двудиагональные.
3. Если A не трёхдиагональная, 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

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

(последний элемент).

§ 15. QR-алгоритм. Практическая реализация

Определение 1. Рассмотрим A со стандартными условиями на собственные вектора, произвольную невырожденную Q_0 . Шаг итерации выглядит так:

$$A = Q_1 R_1,$$

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

где Q_{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.

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

(а) По Реллею: $\{R_n Q_n\}_{s,s}$

(б) По Уилкинсону: собственные числа матрицы $2 \times 2 \{R_n Q_n\}_{s-1,s-1}^{s,s}$

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 простыней 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

$$\varphi(x) = f(x) + \mu \int_a^b K(x, t) \varphi(t) dt. \quad (3.1)$$

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

Обозначим через K (хм, да, вольность) оператор

$$\varphi(t) \mapsto \int_a^b K(x, t) \varphi(t) dt.$$

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:

$$K^* \varphi(x) = \int_a^b \overline{K(t, x)} \varphi(t) dt.$$

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 $\bar{\mu}^{-1}$.
3. μ^{-1} и $\bar{\mu}^{-1}$ — собственные числа операторов K и K^* соответственно одинаковой конечной кратности.

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

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

$$\varphi(x) = f(x) + \sum_{j=1}^n A_j \alpha_j(x), \text{ где } A_j = \mu \int_a^b \beta_j(t) \varphi(t) dt. \quad (3.2)$$

Это представление для функции φ теперь подставим в исходное уравнение:

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

Чтобы переписать это покороче, введём обозначения

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

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:

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

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

$$\varphi(x) = f(x) + \lambda \int_a^b \Gamma(x, t) f(t) dt,$$

где резольвента Γ имеет вид

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

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

¹Это 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

$$u - u_n = (A^{-1} - A_n^{-1})f \Rightarrow \|u - u_n\| \leq \|A^{-1} - A_n^{-1}\| \cdot \|f\|.$$

Поэтому интересно оценить норму разности обратных операторов. Займёмся этим.

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

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

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

$$\|(P - H)^{-1}\| \leq \frac{\|P\|^{-1}}{1 - \|H\| \|P^{-1}\|}.$$

и

$$\|(P - H)^{-1} - P^{-1}\| \leq \frac{\|H\| \|P^{-1}\|^2}{1 - \|H\| \|P^{-1}\|}.$$

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:

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

Далее,

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

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

Наконец,

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

□

Отсюда сразу же следует утверждение

Утверждение 5. При достаточно больших n

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

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

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

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:

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

При этом

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

и

$$\frac{\|(K - K_n)u\|^2}{\|u\|^2} = \frac{\sum_{i=n+1}^{\infty} |u_i|^2 |\lambda_i|^2}{\sum_{i=n+1}^{\infty} |u_i|^2} \leq \frac{\sum_{i=n+1}^{\infty} |u_i|^2 |\lambda_{n+1}|^2}{\sum_{i=n+1}^{\infty} |u_i|^2} = |\lambda_{n+1}|^2.$$

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

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

Получится уравнение

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

где

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

У K_2 с гладкостью получше, но его надо считать.

§ 2. Метод квадратур для интегрального уравнения

Идея заключается в том, чтобы в уравнении

$$u(x) = f(x) + \int_a^b K(x, t)u(t) dt$$

заменить интегрирование на вычисление по какой-нибудь квадратурной формуле:

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

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 $R[K(x, t)u(t)]$ 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:

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

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

$$u = (I - K)^{-1} f \Rightarrow \|u\| \leq \|(I - K)^{-1}\| \cdot \|f\| \leq \frac{\|f\|}{1 - \|K\|} \leq \frac{\|f\|}{1 - \chi},$$

где

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

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 $|\chi| < 1$.

Получив оценку для модуля решения, мы можем найти оценку

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

зависящую только от известных функций.

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

$$\begin{aligned} Au &= f, & A &= I - K; \\ \tilde{A}\tilde{u} &= f, & \tilde{A} &= I - \tilde{K}, \end{aligned}$$

где

$$\tilde{K}\varphi(x) = \sum_{i=1}^n A_i K(x, x_i) \varphi(x_i).$$

Заметим, что

$$\tilde{A}(u - \tilde{u}) = \tilde{A}u - Au \Rightarrow \|u - \tilde{u}\| \leq \|\tilde{A}^{-1}\| \|\tilde{A}u - Au\|.$$

Оценим норму \tilde{A}^{-1} . Для этого сначала оценим норму \tilde{K} :

$$\left| \sum_{i=1}^n A_i K(x, x_i) \varphi(x_i) \right| \leq \max |K| \cdot \|\varphi\| \cdot \sum_{i=1}^n A_i = (b-a) \max |K| \cdot \|\varphi\|,$$

поэтому $\|\tilde{K}\| \leq \kappa$.

Отсюда

$$\|\tilde{A}^{-1}\| = \|(I - \tilde{K})^{-1}\| \leq \frac{1}{1 - \kappa}.$$

Теперь оценим $\|\tilde{A}u - Au\|$:

$$\|\tilde{A}u - Au\| = \max |R[K(x, t)u(t)]| \leq M\delta(n).$$

В конечном итоге находим

$$\|u - \tilde{u}\| \leq \frac{M\delta(n)}{1 - \kappa}.$$

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 $(Au, u) \geq k^2(u, u)$ 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 .

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

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

$$\|u\|^2 \leq (Au, u) \leq \|Au\| \|u\| \Rightarrow \|u\| \leq \|Au\|.$$

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

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

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

$$x_n \rightarrow x \Rightarrow Ax_n \rightarrow Ax \Rightarrow Ax = y.$$

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

При этом

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

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 $u^* + h$:

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

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

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

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 находятся из системы линейных уравнений

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

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 приравнявая 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 параллелепипеда, 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. $\|u_n - u^*\| \rightarrow 0$.

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

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

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

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

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

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

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

$$\int_a^b K(x, t)u(t) dt = f(x).$$

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 переписывается, 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 $\varepsilon \rightarrow 0$ 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. \square

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 $f_n \rightarrow f$ 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 $u'_n \rightarrow u'$.

Поскольку оператор A непрерывен, $Au'_n \rightarrow Au'$; но $Au'_n \rightarrow f$, 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. \square

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:

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

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

¹На 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

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

Если искать не при $\|u\| \leq R$, 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. \square

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

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

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 $\Omega(u) \geq 0$ 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

$$\Omega(u) = \int_a^b u'^2 dt.$$

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. \square

Годятся и функционалы

$$\Omega(u) = \int_a^b u^{(p)^2} dt, \quad \Omega(u) = \int_a^b u'^2 - u^2 dt.$$

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

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

то $\tilde{y} \rightarrow u$.

Доказательство. См. конспект Ангелины. □

Коэффициент α обычно подбирают эмпирически: если он мал, то решение будет ближе к \tilde{y} , 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

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

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 $u^* \in \mathcal{D}(A)$, 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

$$|(f, u)| \leq \|f\| \|u\| \leq k^{-1} \|f\| \|u\|_A.$$

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. $\|u_n - u^*\|_A \rightarrow 0$.

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). \square

§ 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 $C^2([a, b])$, нигде не трогает удовлетворяет уравнению в каждой точке.

Ну и $\mathcal{D}(L) = C^2([a, b])$.

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 пространство

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

а в случае условия III типа

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

Именно их мы будем пополнять, создавая соответствующее энергетическое пространство.

Утверждение 1. Если $\alpha \geq 0$ и $\beta \leq 0$ в добавок к условиям

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

то оператор получится симметричный и положительно определённый.

Доказательство. Посмотрим, как будет выглядеть энергетическое скалярное произведение:

$$(Ly, z) = \int_a^b (-(py')' + qy)z \, dx = -py'z|_a^b + \int_a^b (py'z' + qyz) \, dz.$$

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

$$(Ly, y) = \int_a^b (py'^2 + qy^2) \, dx - \beta p(b)y(b)^2 + \alpha p(a)y(a)^2.$$

Если, $\alpha \geq 0$ и $\beta \leq 0$, то и положительная определённость будет. \square

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 $W_2^1([a, b])$; на нём эта норма имеет вид

$$\|f\|_{W_2^1}^2 = \int_a^b (f^2 + f'^2) dx.$$

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 $\|y\|_{W_2^1} \leq C\|y\|_L$:

$$\int_a^b (f^2 + f'^2) dx \leq \frac{1}{M} \int_a^b (pf^2 + qf'^2) dx \leq \frac{1}{M} \|f\|_L^2.$$

Обратное утверждение очевидно для I типа граничных условий:

$$\int_a^b (pf^2 + qf'^2) dx \leq M \|f\|_{W_2^1}^2.$$

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

$$y(x) = y(\xi) + \int_{\xi}^x y'(t) dt.$$

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

$$y(x)^2 \leq 2y(\xi)^2 + 2 \left(\int_{\xi}^x y'(t) dt \right)^2.$$

При этом интеграл

$$\int_{\xi}^x y'(t) dt$$

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

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

В итоге получаем, что

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

Навесив слева и справа интегралы по ξ , получим, что

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

□

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 $W_2^1 \rightarrow C([a, b])$ 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, $h = \frac{b-a}{n}$, $x_k = a + kh$; 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

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

Сделаем численные замены:

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

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

$$-p_{\frac{1}{2}} \frac{y_1 - y_0}{2} + \frac{h}{2} (q_0 y_0 - f_0) + \alpha p_0 y_0 = 0$$

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

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

§ 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 выполненной сеточной функции. Поэтому наши выполненные функции находятся в W_2^1 .

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

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

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 $\bar{\Omega}$; a_{ij} — один раз непрерывно дифференцируемы на $\bar{\Omega}$, а остальные коэффициенты просто непрерывны.
2. Симметричность (эллиптичность): $a_{ik}(x) = a_{ki}(x)$.
3. Положительная определённость:

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

Граничные условия бывают

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

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

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.

$$(Lu, v) = \int_{\Omega} \left(\sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + auv \right) dx + \int_{\partial\Omega} \sigma uv dS.$$

Доказательство.

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

Используя формулу интегрирования по частям

$$\int_{\Omega} \frac{\partial u}{\partial x_i} \varphi dx = - \int_{\Omega} u \frac{\partial \varphi}{\partial x_i} dx + \int_{\partial\Omega} u \varphi \cos(n, x_i) dS,$$

найдем искомый результат. □

Утверждение 2. Оператор положительно определён, если

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

Доказательство.

$$(Lu, u) = \int_{\Omega} \left(\sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} + au^2 \right) dx + \underbrace{\int_{\partial\Omega} \sigma u^2 dS}_{III}.$$

Нам понадобится неравенство Фридрихса

$$\int_{\Omega} u^2 dx \leq c_1 \left(\int_{\Omega} \sum \left(\frac{\partial u}{\partial x_i} \right)^2 dx + \int_{\partial\Omega} u^2 dS \right).$$

1. Здесь работает совсем грубая оценка:

$$\int_{\Omega} \left(\sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} + au^2 \right) dx \geq \int_{\Omega} \sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} dx \geq k^2 \int_{\Omega} \sum_{i=1}^n \left(\frac{\partial u}{\partial x_j} \right)^2 dx \geq \frac{k^2}{c_1} \int_{\Omega} u^2 dx.$$

2. Ещё проще, как это ни странно.

$$\int_{\Omega} \left(\sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} + au^2 \right) dx \geq a_0 \int_{\Omega} u^2 dx$$

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

$$\int_{\Omega} \left(\sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} + au^2 \right) dx + \int_{\partial\Omega} \sigma u^2 dS \geq k^2 \int_{\Omega} \sum_{i=1}^n \left(\frac{\partial u}{\partial x_j} \right)^2 dx + \sigma_0 \int_{\partial\Omega} u^2 dS \geq \frac{\min(k^2, \sigma_0)}{c_1} \int_{\Omega} u^2 dx.$$

□

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

$$\|v\|_{W_2^1}^2 = \int_{\Omega} \left(\sum \left(\frac{\partial u}{\partial x_i} \right)^2 + u^2 \right) dx.$$

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 Уравнения в частных производных

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

Определение 1. Общее уравнение теплопроводности выглядит вот так:

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

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

Сетка характеризуется такими же, как обычно, величинами:

$$x_i = a + ih, \quad h = \frac{b-a}{n}, \quad i \in 0 \dots n;$$

$$t_k = k\tau, \quad \tau = \frac{T}{M}, \quad k \in 0 \dots M.$$

Положим $u_i^k = u(x_i, t_k)$ и

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

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:

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

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

Для варианта А 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.

забыли

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 $\sigma = \frac{\tau}{h^2}$.

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

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 ортонормированный 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. \square

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 $f_e: H \rightarrow \mathbb{C}$ 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 $H \rightarrow H^*$ есть сопряжённо-линейный изометрический изоморфизм нормированных пространств.

²Верно 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, $A: H \rightarrow H$ — 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 $(A - \lambda I)v = 0 = Av$, 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. $Av = \lambda v$. \square

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 векторы.

Поэтому необратимость либо сразу влечёт неинъективность, либо сначала влечёт несюръективность, а потом уже неинъективность. Отсюда

$$\text{необратимость} \Leftrightarrow \text{неинъективность,}$$

что и требовалось доказать. \square

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 $A: H \rightarrow H$ 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 $\|A_n - A\| \rightarrow 0$, то оператор A компактен.

Следствие 1. Если операторы A_n конечного ранга (т.е. их образы конечномерны), и $\|A - A_n\| \rightarrow 0$, то оператор A компактен.

Главный пример компактного оператора — *интегральный оператор*.

Пример 1. Пусть $\square = [a, b] \times [a, b]$. Рассмотрим оператор A на $L^2([a, b])$, действующий по правилу

$$Af(x) = \int_a^b K(x, y)f(y) dy,$$

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

$$A_N f(x) = \sum_{n=1}^N \left(\sum_{m=1}^N c_{nm} \int_a^b e_m(y) f(y) dy \right) e_n(x).$$

Образ оператора A_N находится внутри линейной оболочки векторов e_1, \dots, e_N ! Это значит, что наш оператор A приближается операторами конечного ранга, а потому компактен. \square

§ 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 $|\lambda| \geq \delta$ 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 оператору 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 $T\varphi_0$ 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.

В Обозначения