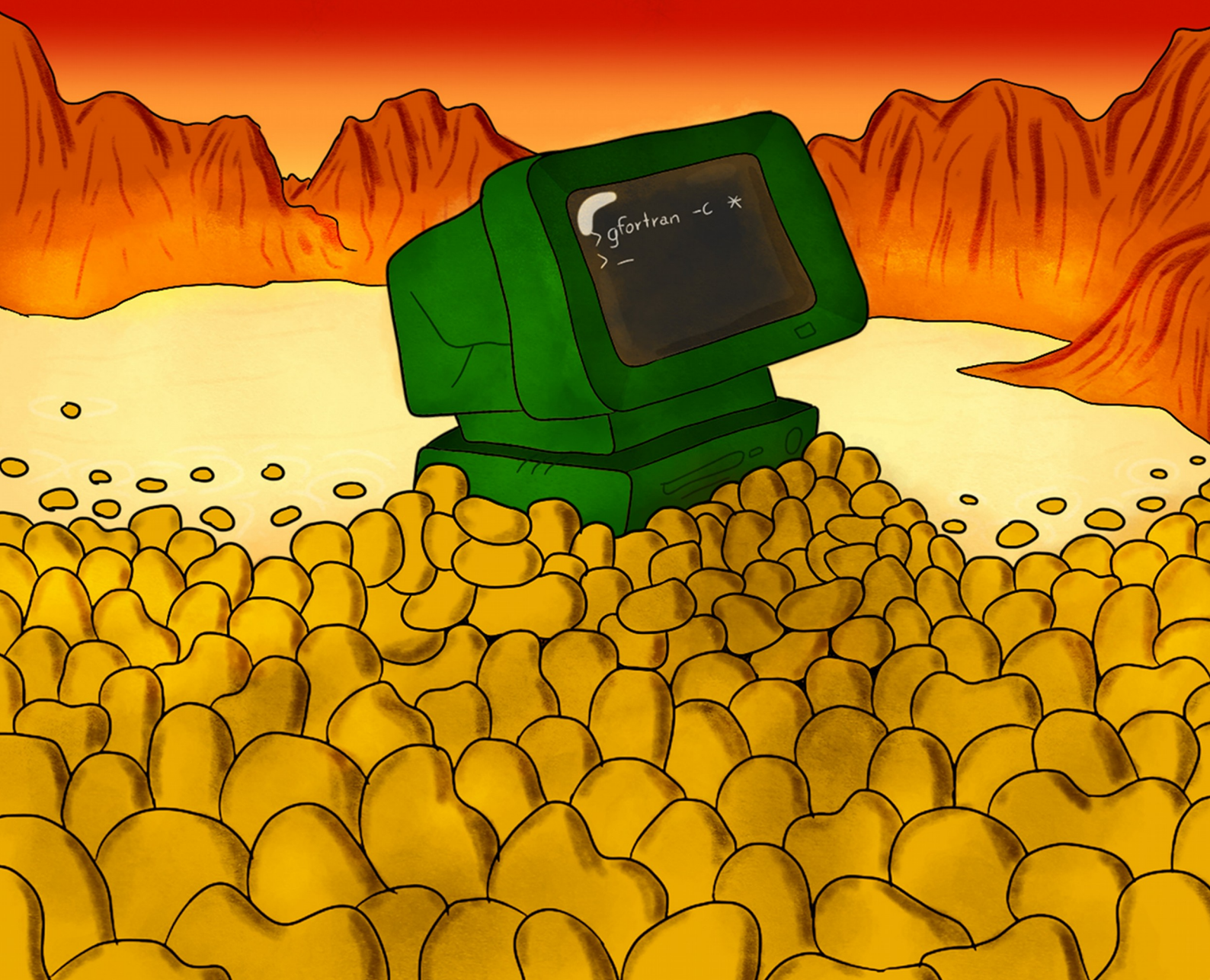


ВЫЧМ



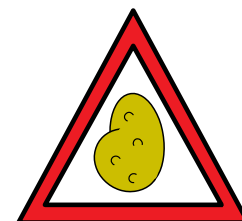
Конспект по методам вычислений

491 группа ММ СПбГУ (2020) и 491 группа ММ СПбГУ (2019)

Весенний семестр, 2020

Дисклеймер

Данный конспект является не более чем «тетрадкой с записями», которые я делал во время лекций. Плюс к этому, я в некотором смысле принял эстафету у прошлого круса и продолжил совершенствование их конспекта. Надеюсь, что читатель найдёт данные материалы хоть сколько-нибудь полезными. Источник мы собираемся открыть, дабы конспект можно было и дальше редачить и дополнять по мере необходимости. Вплоть до 20 билета я опирался на свой бумажный конспект, и изложил здесь ровно всё то, что нам успел дать преподаватель. После начался карантин, и поэтому билеты, начиная с 21, основаны на изложении прошлых конспектов (включая исправления преподавателя). Хочу выразить благодарность преподавателю за курс, составителям прошлого года конспекта за прекрасное разъяснение нужных аспектов функана и составителю ещё более старого и легендарного конспекта. Внимание! Данные материалы имеют ярко выраженный первокультурный оттенок изложения, однако в некоторых местах всё-таки возможно небольшое содержание картофана.



Осторожно, Картофан!

О нотации: Жирными символами (\mathbf{y}) обозначены вектора. Шапочка-циркумфлекс (\hat{A}) обозначает матрицы. Но в некоторых местах (где и так понятно) матрицы без шапочек. Начало доказательства обозначается как \square , а конец – как \blacksquare . Подглавы пронумерованы в соответствии со списком билетов.

Содержание

Дисклеймер	1
I. Краевые задачи для ОДУ (+ Жёсткие системы ДУ).	3
1. Краевая задача для ОДУ 2-го порядка: сведение к задаче Коши.	3
2. Метод дифференциальной прогонки для краевой задачи 2-го порядка.	4
3. Двухточечная краевая задача для системы уравнений 1-го порядка. Метод дифференциальной прогонки.	5
4. Ортогональная прогонка для систем уравнений 1-го порядка.	8
5. Разностный метод для краевой задачи 2-го порядка: составление разностных уравнений. . .	9
6. Метод разностной прогонки.	12
7. Лемма об оценке для системы разностных уравнений.	13
8. Теорема о сходимости разностного метода для обыкновенной краевой задачи.	15
9. Жесткие системы ОДУ. Простейшие методы. Понятие A-устойчивости.	16
10. Понятие L-устойчивости. Неявные методы Рунге-Кутты, общее понятие. Диагонально-неявные методы.	18
11. Асимптотический метод в задаче о быстрых колебаниях.	21
II. Задачи на собственные значения.	24
12. Вопрос об устойчивости собственных чисел и собственных векторов при возмущении матрицы. Отрицательный пример.	24
13. Теорема Бауэра-Файка о возмущении собственных чисел симметричной матрицы.	25
14. Устойчивость собственных векторов при возмущении матрицы.	26
15. Степенной метод для отыскания старшего собственного числа.	28
16. Обратный степенной метод.	30
17. Двумерные вращения, их виды.	31
18. Лемма о правиле знаков при исключении.	32
19. Метод Гивенса.	33

20. Метод Якоби.	33
21. Две леммы о факторизации матрицы.	35
22. Теорема о сходимости итерированных подпространств.	36
23. Треугольно-степенной метод. Сходимость.	39
24. Ортогонально-степенной метод.	39
25. LR-алгоритм. Практическая реализация.	40
26. QR-алгоритм. Практическая реализация.	41
III. Интегральные уравнения.	42
27. Интегральное уравнение 2-го рода. Метод замены ядра на вырожденное.	42
28. Метод квадратур для интегрального уравнения.	46
29. Вариационный принцип для ограниченного оператора. Метод Ритца для интегрального уравнения 2-го рода.	48
30. Интегральное уравнение 1-го рода. Понятие корректности. Некорректность уравнения 1-го рода.	50
31. Условная корректность по Тихонову. Метод квазирешений.	51
32. Метод регуляризации для уравнения 1-го рода. Сходимость.	52
IV. Вариационные методы.	54
33. Вариационный принцип для уравнения с неограниченным оператором.	54
34. Метод Ритца. Сходимость.	55
35. Метод Ритца для обыкновенной краевой задачи. Вид энергетического пространства. Естественные граничные условия.	55
36. ВРМ-1 для обыкновенной краевой задачи.	57
37. ВРМ-2 для обыкновенной краевой задачи.	58
38. Метод Ритца для эллиптического уравнения. Вид естественного граничного условия. Вид энергетического пространства.	59
V. Метод сеток для уравнений в частных производных.	62
39. Разностный метод для общего уравнения теплопроводности. Явная схема.	62
40. Неявная схема для уравнения теплопроводности.	63
41. Явная схема для простейшего уравнения теплопроводности. Решение разностных уравнений. Явление неустойчивости.	64
42. Общее определение устойчивости. Теорема об устойчивости и сходимости.	65
43. Разностные схемы для задач с начальными условиями. Дискретное преобразование Фурье.	66
44. Необходимое условие устойчивости по фон-Нейману.	71
45. Простейшие схемы для уравнения бегущей волны.	72
46. Схема Куранта-Рисса.	73
47. Явная схема для уравнения колебаний струны.	74
48. Явная и неявная схемы для двумерного уравнения теплопроводности.	75
49. Схема продольно-поперечной прогонки.	76
A. Введение в функциональный анализ.	78
A1. Пространства, отображения.	78
A2. Пара фактов о гильбертовых пространствах.	78
A3. Спектр оператора.	79
A4. Компактные операторы.	80
A5. Спектры компактных операторов.	80
A6. Альтернатива Фредгольма.	81

I. Краевые задачи для ОДУ (+ Жёсткие системы ДУ).

Ну... Давайте, начнём!
– А.Л.Г.

1. Краевая задача для ОДУ 2-го порядка: сведение к задаче Коши.

Рассмотрим простейшую задачу:

$$y'' + p(x)y' + q(x)y = f(x), \quad (1)$$

где $p(x), q(x), f(x) \in C(I \supset [a, b]);^1$ $y(x) \in C^2([a, b])$, для которой есть 3 различных варианта задания граничного условия:

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

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

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

Задача заключается в нахождении такого решения $y(x)$ дифференциального уравнения (1), чтобы выполнялось одно из условий (I, II, III).

Рассмотрим однородную задачу, то есть положим $f \equiv 0$ в (1) и $A = B = 0$ в (I), (II) или (III). Такая задача имеет тривиальное решение $y(x) \equiv 0$.

Теорема. Если однородная краевая задача имеет только одно решение (тривиальное), то любая соответствующая ей неоднородная краевая задача (с теми же $p(x)$ и $q(x)$) однозначно разрешима.

□ Пусть $y_0(x)$ – некоторое частное решение неоднородного уравнения (1), удовлетворяющее какому-нибудь из граничных условий (к примеру, (I)) а $y_1(x)$ и $y_2(x)$ – линейно-независимые решения однородного уравнения. Тогда общее решение неоднородного уравнения представимо в виде $y(x) = y_0(x) + C_1 y_1(x) + C_2 y_2(x)$ (вспоминаем 4-ую главу Басова). Где $C_1 y_1(x) + C_2 y_2(x)$ – общее решение однородной системы. Тривиальному решению соответствуют константы $C_1 = C_2 = 0$ и оно единственно для краевой задачи \Rightarrow Частное решение $y_0(x)$ тоже единственно для своей краевой задачи. Для задач вида (II) и (III) – Аналогично ▣.

Можно найти решение, задавая какие-нибудь начальные условия для y_0, y_1, y_2 и решая задачу Коши. Потом подставлять полученные решения в исходную систему. В этом заключается так называемый *метод смертельной пристрелки задачи Коши*

Пример 1. Пусть для y_0, y_1, y_2 заданы такие начальные условия:

- $y_0 : y_0(a) = 0, y'_0(a) = 0$; не забываем ($f(x) \not\equiv 0$);
- $y_1 : y_1(a) = 1, y'_1(a) = 0$; не забываем ($f(x) \equiv 0$);
- $y_2 : y_2(a) = 0, y'_2(a) = 1$; не забываем ($f(x) \equiv 0$).

Решения y_1 и y_2 независимы, потому что их определитель Вронского в точке a будет равен

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

значит он не ноль на всём отрезке $[a, b]$ (тоже по 4 главе Басова), а значит решения ЛНЗ.

Найдя все эти игреки, можно сконструировать $y(x) = y_0(x) + C_1 y_1(x) + C_2 y_2(x)$. Изменяя C_1 и C_2 , можем подобрать такие значения, что «попадём» в нужное нам правое граничное условие (потому это

¹Здесь I – это какой-нибудь интервал, содержащий в себе отрезок $[a, b]$. Просто формально функции p, q, f должны быть определёнными и непрерывными на связном открытом множестве (т.е. на области). Решение y , определённое на отрезке, продолжимо за его границы по лемме из первой главы Басова.

и называется методом пристрелки). Но этот пример не самый удачный. Тут надо варьировать аж две константы, а это неудобно.

Пример 2. На самом деле в y_2 мы вообще не нуждаемся.² Пусть задана краевая задача с условием (III). Для y_0 выберем начальные условия $y_0(a) = 0$, $y'_0(a) = A$. Для y_1 выберем $y_1(a) = 1$, $y'_1(a) = \alpha$. Получим решение $y(x) = y_0(x) + C_1 y_1(x)$, которое $\forall C_1$ удовлетворяет левому граничному условию. То есть мы нашли однопараметрическое частное семейство решений. Изменяя C_1 , можем «пристрелиться» и «попасть» в нужное правое граничное условие. Обозначим это семейство решений как (A):

$$y(x) = y_0(x) + C_1 y_1(x), \text{ где } y_0(a) = 0, y'_0(a) = A; y_1(a) = 1, y'_1(a) = \alpha. \quad (\text{A})$$

Пример 3. Теперь продемонстрируем, почему метод пристрелки не так уж хорош. Тем самым обоснуем необходимость метода из следующего билета. Пусть задана краевая задача

$$y'' - a^2 y = 0 \quad (a \neq 0), \quad y(0) = 1, \quad y(b) = 1.$$

Докажем, что $0 < y(x) < 1 \quad \forall x \in (0, b)$.

□ От обратного: Пусть $\exists c : \max y(x) = y(c) > 1 \Rightarrow y''(c) \leq 0 \Rightarrow$ уравнение не выполняется в точке c (получаем знак “ $<$ ” вместо равенства).

Аналогично, пусть $\exists c : \min y(x) = y(c) < 0 \Rightarrow y''(c) \geq 0 \Rightarrow$ уравнение не выполняется в точке c . Также y не может быть $\equiv 1$, потому что опять же не будет выполняться уравнение ■.

Зададим граничные условия для y_0 и y_1 :

$$y_0(0) = 1, y'_0(0) = 0; \Rightarrow y_0(x) = \operatorname{ch} ax = \frac{e^{ax} + e^{-ax}}{2}.$$

$$y_1(0) = 0, y'_1(0) = 1; \Rightarrow y_1(x) = \frac{e^{ax} - e^{-ax}}{2a} = \frac{1}{a} \operatorname{sh} ax.$$

$y(x) = y_0(x) + C_1 y_1(x)$, удовлетворим это правому граничному условию:

$$\operatorname{ch} ab + \frac{C_1}{a} \operatorname{sh} ab = 1 \Rightarrow C_1 = a \frac{1 - \operatorname{ch} ab}{\operatorname{sh} ab} \Rightarrow y(x) = \operatorname{ch} ax + \frac{1 - \operatorname{ch} ab}{\operatorname{sh} ab} \operatorname{sh} ax.$$

Если $x \approx b$, то $\operatorname{ch} ab = \frac{1}{2}(e^{ab} + e^{-ab}) \approx \frac{1}{2}e^{ab}$, и $\operatorname{sh} ab \approx \frac{1}{2}e^{ab}$. В итоге получаем, что решение – это разность двух больших чисел. Если $ab = 10$, то $e^{ab} \sim 10^4$; если $ab = 20$, то $e^{ab} \sim 10^8$, то есть теряем 8 знаков после запятой. Дело в том, что y_0 и y_1 оба растут. И постепенно они начинают «примерно терять линейную независимость».

2. Метод дифференциальной прогонки для краевой задачи 2-го порядка.

Рассматриваем задачу (1) с граничными условиями вида (III). И снова рассматриваем семейство решений (A). Теперь рассмотрим ЛНУ первого порядка, которое имеет вид

$$y'(x) = \alpha(x)y + \beta(x). \quad (2)$$

Хотим подобрать такие $\alpha(x)$ и $\beta(x)$,³ чтобы ДУ (2) описывало семейство (A).

$$y''(x) = \alpha'(x)y + \alpha(x)y' + \beta'(x) = \alpha(x)(\alpha(x)y + \beta(x)) + \alpha'(x)y + \beta'(x) = \alpha^2(x)y + \alpha(x)\beta(x) + \alpha'(x)y + \beta'(x).$$

Можем подставить полученное выражение для y'' , а также y' из (2) в (1), получим:

$$(\alpha^2(x) + \alpha'(x) + \alpha(x)\beta(x) + q(x))y + \beta'(x) + \alpha(x)\beta(x) + p(x)\beta(x) = f(x).$$

²Можем посчитать, что мы зафиксировали случайный C_2 и приплюсовали y_2 к y_0 . Поскольку y_2 – это решение однородного уравнения, то от этого y_0 как был, так и остался решением неоднородного, то есть ничего не поменялось.

³Не путать с α и β , определёнными в (III)! (Не люблю коллизии символов, но что поделать...)

В этом выражении полагаем, что то, что написано в скобках при y , равно нулю. А всё остальное, соответственно, равно f . Получаем ДУ для $\alpha(x)$ и $\beta(x)$:

$$\begin{aligned}\alpha'(x) + \alpha^2(x) + p(x)\alpha(x) + q(x) &= 0, \\ \beta'(x) + \alpha(x)\beta(x) + p(x)\beta(x) &= f(x).\end{aligned}\tag{3}$$

Если соотношения (3) выполняются, то любое решение ДУ (2) удовлетворяет (1). Нам также нужно, чтобы оно ещё и удовлетворяло граничным условиям.

Решим задачу Коши для уравнений (3) с начальными данными $\alpha(a) = \alpha$, $\beta(a) = A$. Тогда уравнение (2) в точке a примет вид

$$y'(a) = \alpha(a)y(a) + \beta(a) = \alpha y(a) + A,$$

то есть как раз получаем выполнение левого граничного условия. То, что мы проделали, называется *прямой прогонкой*.

Теперь мы знаем $\alpha(x)$ и $\beta(x)$. Обратим свой взор на правую границу:

$$\begin{aligned}y'(b) &= \alpha(b)y(b) + \beta(b), \\ y'(b) &= \beta y(b) + B.\end{aligned}$$

Из этой системы уравнений можем найти $y(b)$ и $y'(b)$.⁴ *Обратная прогонка* – решаем задачу Коши для уравнения (2) с начальными данными $y(b)$ (причём интегрируем в обратную сторону, $b \rightarrow a$). В результате находим $y(x)$, который является решением уравнения (1) и удовлетворяет граничным условиям (III). Таким образом, мы решили поставленную задачу. В этом заключается *метод дифференциальной прогонки*. Вообще, $\forall x$ и \forall правого граничного условия \exists условие задачи Коши в точке x , такое, что её решение удовлетворяет правому граничному условию. То есть существует соответствие между правым граничным условием и задачей Коши в какой-нибудь точке x .

Пример. Сейчас покажем, что метод дифференциальной прогонки не всегда хорош. Если $f \equiv 0$ и $A = 0$, то тогда $\beta \equiv 0$, и уравнение (2) превращается в

$$y'(x) = \alpha(x)y(x) \Rightarrow \alpha(x) = \frac{y'(x)}{y(x)}.$$

Если $\exists c : y(c) = 0$ (и при этом $y'(c) \neq 0$), то α в этой точке не существует. Таким образом, метод дифференциальной прогонки неприменим к однородной краевой задаче.

Также метод дифференциальной прогонки неприменим и к граничному условию (I). Для фиксированного x уравнение (2) задаёт прямую в фазовом пространстве.⁵ Первое граничное условие соответствует вертикальной прямой. Для неё мы не можем составить уравнение (2). В этом случае нужно производить замену переменных фазового пространства.⁶

3. Двухточечная краевая задача для системы уравнений 1-го порядка. Метод дифференциальной прогонки.

Любую систему дифференциальных уравнений можно свести к системе первого порядка:

$$\mathbf{y}' = \hat{A}(x)\mathbf{y} + \mathbf{f}(x),\tag{4}$$

⁴На самом деле не всегда: если $\alpha(b) = \beta$, то получаем условие $\beta(b) = B$. Если оно не выполнено, то решения, наверное, нет совсем. А если выполнено, то оно, скорее всего, может быть не единственно. Это моя личная догадка, а вообще стоит спросить у преподавателя.

⁵Фазовое пространство – пространство (топологическое) из пар (y, y') .

⁶Как я понял, либо создаём другой y , и из него получаем другой y' , либо вообще сводим уравнение 2 порядка к системе 1 порядка, но про это будет следующий билет. После этого в моей бумажной версии конспекта были какие-то фразы про обобщение на задачу 4 порядка. Не уверен, что это нужно.

где вектора имеют s компонент, а матрица \hat{A} – размера $s \times s$. Считаем, что $(\forall i, j \in \{1, \dots, s\}) f_i$ и $A_{ij} \in C(I \supset [a, b])$, а $y_i \in C^1([a, b])$ (с тем же самым замечанием про I , как в первом билете).

Сформулируем граничные условия:⁷

$$\hat{\alpha} \mathbf{y}(a) = \beta, \quad \hat{\gamma} \mathbf{y}(b) = \delta, \quad (5)$$

где $\hat{\alpha}$ – матрица размера $p \times s$, а $\hat{\gamma}$ – матрица размера $q \times s$, причём $q + p = s$; $\beta \in \mathbb{R}^p$, $\delta \in \mathbb{R}^q$. Общее решение системы (4) выглядит так:

$$\mathbf{y}(x) = \mathbf{y}^{(0)} + \sum_{i=1}^s C_i \mathbf{y}^{(i)}(x), \quad (6)$$

где $\{\mathbf{y}^{(i)}\}_{i=1}^s$ – это ФСР однородной системы, соответствующей (4), а $\mathbf{y}^{(0)}$ – это частное решение неоднородной системы (4) (всё как в главе 5 нашего любимого Басова).

Опишем нечто наподобие метода пристрелки, но уже для многомерного случая. Выберем начальные данные для $\mathbf{y}^{(0)}$ и $\mathbf{y}^{(i)}$. Выберем $\mathbf{y}^{(0)}(a)$ так, чтобы оно удовлетворяло соотношению $\hat{\alpha} \mathbf{y}^{(0)}(a) = \beta$. А $\mathbf{y}^{(i)}(a)$ выбираем так, чтобы $\hat{\alpha} \mathbf{y}^{(i)}(a) = 0$. Это недоопределённая однородная система. Надо полагать, что $\text{rang } \hat{\alpha} = p$. Всего у этой системы можно подобрать $s - p = q$ ЛНЗ решений, причём неважно, каких. Теперь интегрируем численно дифференциальные уравнения относительно $\mathbf{y}^{(0)}$ и $\mathbf{y}^{(i)}$

$$\mathbf{y}^{(0)'} = \hat{A} \mathbf{y}^{(0)} + \mathbf{f}, \quad \mathbf{y}^{(i)'} = A \mathbf{y}^{(i)}$$

с описанными ранее начальными данными. Можем составить комбинацию

$$\mathbf{y}(x) = \mathbf{y}^{(0)}(x) + \sum_{i=1}^q C_i \mathbf{y}^{(i)}(x), \quad (7)$$

она будет являться решением неоднородной системы (4) и $\forall \{C_i\}_{i=1}^q$ это решение будет удовлетворять левым граничным условиям (потому что мы так выбрали начальные данные). Посмотрим, что будет на правой границе:

$$\mathbf{y}(b) = \mathbf{y}^{(0)}(b) + \sum_{i=1}^q C_i \mathbf{y}^{(i)}(b),$$

подставляем это в правые граничные условия $\hat{\gamma} \mathbf{y}(b) = \delta$:

$$\hat{\gamma} \mathbf{y}^{(0)}(b) + \sum_{i=1}^q C_i \hat{\gamma} \mathbf{y}^{(i)}(b) = \delta.$$

Видимо, тут мы также должны, варьируя набор констант $\{C_i\}_{i=1}^q$, пристреливаться, чтобы удовлетворить правым граничным условиям.

У этого метода существует аналогичная вероятная беда с экспонентами и потерей точности: ищем решение однородной системы $\mathbf{y}^{(i)'} = \hat{A} \mathbf{y}^{(i)}$ в виде $\mathbf{y}^{(i)}(x) = \mathbf{c} e^{\lambda x} \Rightarrow \lambda \mathbf{c} e^{\lambda x} = \hat{A} \mathbf{c} e^{\lambda x} \Rightarrow \hat{A} \mathbf{c} = \lambda \mathbf{c}$, то есть λ – это собственное число, а \mathbf{c} – это собственный вектор матрицы \hat{A} . $|e^{\lambda x}| = e^{\text{Re } \lambda x}$. Если $\text{Re } \lambda$ сильно различаются для разных $\mathbf{y}^{(i)}$, то число обусловленности⁸ матрицы \hat{A} будет расти, вектора \mathbf{c} будут становиться «примерно линейно зависимыми», и мы получим потерю точности.

Теперь опишем метод дифференциальной прогонки. Вообще, ~~нережные~~ прогоночные соотношения можно рассматривать как перенос граничного условия из точки a в какую-нибудь точку $c \in (a, b)$. Рассмотрим подсемейство решений вида (7), удовлетворяющее левому граничному условию. Тогда в

⁷Граничные условия могут быть определены и в более общем виде, а это – линейные распадающиеся граничные условия.

⁸Обозначается как $\mu(\hat{A}) = \|\hat{A}\| \cdot \|\hat{A}^{-1}\|$, в частных случаях оно равно $|\frac{\lambda_{\max}}{\lambda_{\min}}|$ – число, характеризующее насколько точна система линейных уравнений с матрицей \hat{A} . Чем больше $\mu(\hat{A})$, тем менее точна система.

точке c оно также будет удовлетворять каким-то соотношениям. Их будет тоже p штук. Они будут неоднородными и линейными. То есть $\forall x \in (a, b) \exists \hat{\alpha}(x)$ и $\beta(x)$:

$$\hat{\alpha}(x)\mathbf{y}(x) = \beta(x). \quad (8)$$

Для любого семейства функций вида (7) (удовлетворяющих левым граничным условиям),

$$\begin{aligned} \hat{\alpha}'(x)\mathbf{y}(x) + \hat{\alpha}(x)\mathbf{y}'(x) = \beta'(x) &\Rightarrow \hat{\alpha}'(x)\mathbf{y}(x) + \hat{\alpha}(x)\hat{A}(x)\mathbf{y}(x) + \hat{\alpha}(x)\mathbf{f}(x) = \beta'(x) \Rightarrow \\ &\left(\hat{\alpha}'(x) + \hat{\alpha}(x)\hat{A}(x)\right)\mathbf{y} + \hat{\alpha}(x)\mathbf{f}(x) = \beta'(x). \end{aligned}$$

Как и в случае краевой задачи 2 порядка, полагаем

$$\hat{\alpha}' + \hat{\alpha}\hat{A} = \hat{0}, \quad \beta' = \hat{\alpha}\mathbf{f}. \quad (9)$$

Это система ОДУ 1 порядка. Можем расписать покомпонентно (картофаан):

$$\alpha'_{ik}(x) + \sum_{j=1}^s \alpha_{ij}(x)A_{jk}(x) = 0 \quad \forall i \in \{1, \dots, p\}.$$

Можем ввести $\alpha_i(x) = (\alpha_{i1}, \dots, \alpha_{is})(x)$ – (вектор-)строка $\hat{\alpha}$. Для него это же соотношение будет:

$$\alpha'_i(x) = -\alpha_i(x)\hat{A}(x),$$

можем транспонировать всё, и получим:

$$\alpha_i^{T'}(x) = -\hat{A}^T(x)\alpha_i^T(x),$$

иногда это называется методом сопряжённой системы. Для удовлетворения левым граничным условиям нужно потребовать $\hat{\alpha}(a) = \hat{\alpha}$ и $\beta(a) = \beta$. Это и есть прямая прогонка: мы задали начальные данные для задачи Коши системы (9). Причём здесь система всегда разрешима, поскольку линейна. «Это хорошо, но какой ценой?» Здесь точность также будет зависеть от числа обусловленности матрицы \hat{A} и, таким образом, эта система также будет подвержена потере точности. О том, как побороть эти потери – в следующем билете.

Из прямой прогонки находим $\hat{\alpha}(x)$ и $\beta(x)$, подставляем их значения в точке b в правое граничное условие:

$$\hat{\alpha}(b)\mathbf{y}(b) = \beta(b) \quad - p \text{ равенств},$$

$$\hat{\gamma}\mathbf{y}(b) = \delta \quad - q \text{ равенств}.$$

Всего имеем $q + p = s$ равенств. Из них находим $\mathbf{y}(b)$. И таким образом всё.⁹ На самом деле из соотношения (8) видно, что \mathbf{y} в каждой точке x можно найти как решение системы с матрицей $\hat{\alpha}$ и вектором β . То есть зная их, мы можем найти \mathbf{y} в любой точке x . Тогда вопрос: зачем мы вообще выводили последнюю систему, если найдя $\hat{\alpha}$ и β из прямой прогонки мы можем получить \mathbf{y} ? Ответ: потому что мы на самом деле не можем его так получить. Система (8) состоит из p соотношений, то есть она неполная. А вот последняя полученная нами система содержит s равенств и, таким образом, является полной относительно $\mathbf{y}(b)$. Я так понял, что для обратной прогонки мы должны интегрировать исходную систему (4) с начальными данными $\mathbf{y}(b)$. И на этом действительно будет всё.

⁹Заметьте, что мы при этом ничего не говорили про обратную прогонку. На самой лекции преподаватель именно на этом моменте сказал, что всё. Далее идут мои догадки.

4. Ортогональная прогонка для систем уравнений 1-го порядка.

Для начала быстренько вспомним, как производится ортогонализация. Пусть есть 2 вектора: \mathbf{x} и \mathbf{y} . Этот самый \mathbf{y} хотим разложить на две компоненты: одна коллинеарна \mathbf{x} , а другая – ортогональна. Хотим получить выражение для второй компоненты. Для этого мы должны вычесть из \mathbf{y} проекцию \mathbf{y} на \mathbf{x} . То есть получим

$$\mathbf{y} - \frac{(\mathbf{y}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})} \mathbf{x}, \quad \text{в матричном виде} - \quad \mathbf{y} - \mathbf{y} \mathbf{x}^T (\mathbf{x} \mathbf{x}^T)^{-1} \mathbf{x}.$$

Вид такого выражения нам потребуется дальше.¹⁰

Вместо дифференциального уравнения на $\hat{\alpha}$ из (9), которое мы ввели в прошлом билете, введём другое дифференциальное уравнение относительно $\hat{\alpha}$:

$$\hat{\alpha}' = -\hat{\alpha} \hat{A} + \hat{\alpha} \hat{A} \hat{\alpha}^T (\hat{\alpha} \hat{\alpha}^T)^{-1} \hat{\alpha}. \quad (10)$$

То, что мы видим в качестве второго слагаемого правой части этого выражения – проекция $\hat{\alpha} \hat{A}$ на $\hat{\alpha}$. Мы выразили $\hat{\alpha}'$ через матрицу, ортогональную $\hat{\alpha}$. Раз они ортогональны, то их скалярное произведение должны быть равно нулю. Проверим:

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

Утверждение. Если $\hat{\alpha}$ удовлетворяет уравнению (10), то $\hat{\alpha} \hat{\alpha}^T$ – постоянная матрица.

□ Рассмотрим производную «скалярного произведения» $\hat{\alpha}$ на себя:

$$(\hat{\alpha} \hat{\alpha}^T)' = \hat{\alpha}' \hat{\alpha}^T + \hat{\alpha} \hat{\alpha}^{T'} = \hat{\alpha}' \hat{\alpha}^T + (\hat{\alpha}' \hat{\alpha}^T)^T.$$

Выше мы уже показали, что $\hat{\alpha}' \hat{\alpha}^T = \hat{0}$. А значит и $(\hat{\alpha} \hat{\alpha}^T)' = \hat{0}$. Следовательно, $\hat{\alpha} \hat{\alpha}^T = \widehat{\text{const}}$ ▮.

Снова вводим прогоночные соотношения

$$\hat{\alpha}(x) \mathbf{y}(x) = \beta(x). \quad (11)$$

Можем продифференцировать их, чтобы найти дифференциальное уравнение для $\beta(x)$:

$$\hat{\alpha}(x) \mathbf{y}'(x) + \hat{\alpha}'(x) \mathbf{y}(x) = \beta'(x) \Rightarrow \hat{\alpha} \hat{A} \mathbf{y} + \hat{\alpha} \mathbf{f} - \hat{\alpha} \hat{A} \mathbf{y} + \hat{\alpha} \hat{A} \hat{\alpha}^T (\hat{\alpha} \hat{\alpha}^T)^{-1} \hat{\alpha} \mathbf{y} = \beta'(x).$$

Таким образом получаем

$$\beta'(x) = \hat{\alpha}(x) \mathbf{f}(x) + \hat{\alpha} \hat{A} \hat{\alpha}^T (\hat{\alpha} \hat{\alpha}^T)^{-1} \beta. \quad (12)$$

Из постоянства «скалярного квадрата» $\hat{\alpha}$ следует сохранение нормы $\hat{\alpha}$, а из него следует сохранение длин и углов. Получаем, что не происходит вырождения матрицы $\hat{\alpha}$. То есть у неё постоянно число обусловленности, и, таким образом, мы избавлены от тех проблем с потерями точности и ростом экспонент, которые были приведены в примерах к прошлым билетам.

Далее, можем применить процесс ортонормализации Грамма-Шмидта к строчкам $\hat{\alpha}$. Это нужно, потому что отыскание обратной матрицы – дорогая процедура. В результате получим $\hat{\alpha} \hat{\alpha}^T = E$. Тогда уравнения (10) и (12) превратятся в

$$\begin{aligned} \hat{\alpha}' &= -\hat{\alpha} \hat{A} + \hat{\alpha} \hat{A} \hat{\alpha}^T \hat{\alpha}, \\ \beta' &= \hat{\alpha} \mathbf{f} + \hat{\alpha} \hat{A} \hat{\alpha}^T \beta. \end{aligned} \quad (13)$$

¹⁰Здесь и далее под «скалярным произведением» матриц $\hat{\alpha}$ и $\hat{\beta}$ понимается $(\hat{\alpha}, \hat{\beta}) = \hat{\alpha} \hat{\beta}^T$. Забавно, что для такого «скалярного произведения» выполняется эрмитовость, если под сопряжением понимать транспонирование. Я так думаю, что преподаватель писал знак * вместо ^T именно потому, что он имел в виду Эрмитово сопряжение, как в квантах. Видимо, мы работаем только в \mathbb{R} , поэтому эрмитово сопряжение превратилось в обычное транспонирование.

Следует помнить, что раз мы ортонормировали $\hat{\alpha}$, то мы заменили переменные, поэтому \hat{A} и \mathbf{f} в (13) тоже подвергнутся соответствующим изменениям.

А теперь, чтобы решить систему, проводим привычные действия для метода дифференциальной прогонки: Решаем систему (13) с начальными условиями $\hat{\alpha}(a) = \hat{\alpha}$, $\beta(a) = \beta$; Искомое решение удовлетворяет прогоночным соотношениям (11) и правому граничному условию $\hat{\gamma}(b)\mathbf{y}(b) = \delta$. Вместе эти соотношения (в точке b) образуют систему, решая которую, мы можем найти $\mathbf{y}(b)$. И на этом (аналогично предыдущему билету) задача решится.

Также, можем ещё и решать ту же самую систему (13) с условиями задачи Коши для правого конца: $\hat{\gamma}(b) = \hat{\gamma}$, $\delta(b) = \delta$. Тогда получим два соотношения: $\hat{\alpha}(x)\mathbf{y}(x) = \beta(x)$ и $\hat{\gamma}(x)\mathbf{y}(x) = \delta(x)$. Можем решать и такую систему. Это называется встречной прогонкой.

5. Разностный метод для краевой задачи 2-го порядка: составление разностных уравнений.

Снова рассматриваем линейное дифференциальное уравнение 2-го порядка (1) с граничными условиями вида (I), (II), или (III). Опишем алгоритм решения этой краевой задачи разностным методом:

1. Выбираем сетку $\{x_k\}_{k=0}^n$. Простейший вид – равномерная сетка ($x_k = a + kh \ \forall k \in \{0, \dots, n\}$, где $h = \frac{b-a}{n}$ – её шаг).
2. Составляем разностные уравнения (Выражаем производные через формулы с конечными разностями и подставляем в уравнение (1) со значениями в узлах сетки).
3. Решение полученной системы разностных уравнений.

Формулы численного дифференцирования делятся на:

- Простейшие – используют минимальное число узлов ($n + 1$ для производной порядка n),
- Многоточечные.

Обозначение. $\mathcal{M}_m = \max |y^{(m)}(x)|$.

Приведём несколько формул численного дифференцирования и оценки для их остатков:

$$y'(x) = \frac{y(x+h) - y(x)}{h} + R, \quad |R| \leq \frac{\mathcal{M}_2}{2}h, \quad (\text{Д1})$$

$$y'(x) = \frac{y(x) - y(x-h)}{h} + R, \quad |R| \leq \frac{\mathcal{M}_2}{2}h, \quad (\text{Д2})$$

$$y'(x) = \frac{-y(x+2h) + 4y(x+h) - 3y(x)}{2h} + R, \quad |R| \leq \frac{\mathcal{M}_3}{3}h^2, \quad (\text{Д3})$$

$$y'(x) = \frac{3y(x) - 4y(x-h) + y(x-2h)}{2h} + R, \quad |R| \leq \frac{\mathcal{M}_3}{3}h^2, \quad (\text{Д4})$$

$$y'(x) = \frac{y(x+h) - y(x-h)}{2h} + R, \quad |R| \leq \frac{\mathcal{M}_3}{6}h^2, \quad (\text{Д5})$$

$$y''(x) = \frac{y(x+h) - 2y(x) + y(x-h)}{h^2} + R, \quad |R| \leq \frac{\mathcal{M}_4}{12}h^2. \quad (\text{Д6})$$

Формулы (Д1), (Д2) являются простейшими. По-моему, (Д6) – тоже.

Чуть поподробнее о том, каким образом эти формулы выводятся. На примере (Д5). Хотим представить y' в точке x как

$$y'(x) \approx A_1 y(x+h) + A_{-1} y(x-h).$$

При этом мы можем выразить $y(x+h)$ и $y(x-h)$ через разложение в ряд Тейлора:

$$y(x+h) = y + hy' + O(h^2), \quad y(x-h) = y - hy' + O(h^2),$$

тогда получаем

$$A_1 y(x+h) + A_{-1} y(x-h) = (A_1 + A_{-1})y + h(A_1 - A_{-1})y' + O(h^2).$$

Отсюда можем найти величины A_1 и A_{-1} :

$$A_1 + A_{-1} = 0, \quad h(A_1 - A_{-1}) = 1 \Rightarrow A_1 = \frac{1}{2h}, \quad A_{-1} = -\frac{1}{2h}.$$

Но если мы подставим эти величины в исходную формулу, то получим точность $O(h)$. Хотя из оценки остатка в (Д5) видно, что в данной формуле можно получить точность $O(h^2)$. В чём же дело?

На самом деле мы могли бы раскладывать $y(x+h)$ и $y(x-h)$ в ряд Тейлора до более высокого порядка

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

и тогда мы для A_1 и A_{-1} получили бы переопределённую систему:

$$A_1 + A_{-1} = 0, \quad h(A_1 - A_{-1}) = 1, \quad \frac{h^2}{2}(A_1 + A_{-1}) = 0,$$

которая, тем не менее, имеет единственное решение $A_1 = \frac{1}{2h}$, $A_{-1} = -\frac{1}{2h}$. Также мы могли бы в исходной формуле добавить член $A_0 y(x)$. И тогда в системе получилось бы 3 уравнения и 3 неизвестных. При этом A_0 получился бы равен нулю. В любом случае теперь получаем нужную точность $O(h^2)$.

А теперь об оценке остатка. Выразим его в форме Лагранжа¹¹:

$$\begin{aligned} y(x+h) &= y + hy' + \frac{h^2}{2}y'' + \frac{h^3}{6}y'''(\xi_1), \\ y(x-h) &= y - hy' + \frac{h^2}{2}y'' - \frac{h^3}{6}y'''(\xi_2), \\ \frac{y(x+h) - y(x-h)}{2h} &= y' + \frac{1}{2h}(y'''(\xi_1) + y'''(\xi_2))\frac{h^3}{6}, \\ \exists \xi : y'''(\xi) &= \frac{y'''(\xi_1) + y'''(\xi_2)}{2}. \end{aligned}$$

В конце концов, получаем:

$$y'(x) = \frac{y(x+h) - y(x-h)}{2h} - \frac{h^2}{6}y'''(\xi).$$

Очевидно, что $|y'''(\xi)| \leq \mathcal{M}_3$. Таким образом, мы вывели оценку остатка.

Теперь перейдём к следующему шагу алгоритма. Подставим формулы (Д6) и (Д5) в систему (1) и получим:

$$\frac{y(x_{k+1}) - 2y_k(x_k) + y(x_{k-1}))}{h^2} + R_k + p(x_k) \frac{y(x_{k+1}) - y(x_{k-1}))}{2h} + q(x_k)y(x_k) = f(x_k), \quad (15)$$

где $R_k = R_k'' + R_k'p(x_k)$ – это остаток от обеих формул численного дифференцирования. Он имеет порядок $O(h^2)$.

Обозначение. $f_k = f(x_k)$ – и так для каждой функции от x . Переписываем индекс элемента сетки подобным образом ради сокращения выкладок. С y_k ситуация немного другая: y_k приближенно равен $y(x_k)$, а не точно, потому что далее мы будем писать ту же формулу уже без остатка R . Собственно поиск $\{y_k\}_{k=0}^n$ – это и есть поиск (приближенного) численного решения, то есть та задача, которой мы занимаемся.

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

¹¹На этом моменте по-хорошему нужно требовать $y \in C^3([a, b])$.

Эти разностные уравнения определены $\forall k \in \{1, \dots, n-1\}$ а на краях они не работают. В случае граничных условий (I) мы знаем значения y_0 и y_n . А для (II) и (III) нужно составлять граничные разностные уравнения.

0) На ум приходит использовать формулы (Д1) и (Д2) для левой и правой границы соответственно (пример для (III) граничного условия):

$$\frac{y_1 - y_0}{h} = \alpha y_0 + A, \quad \frac{y_n - y_{n-1}}{h} = \beta y_n + B.$$

Но это плохая идея, потому что эти формулы имеют точность $O(h)$, то есть они испортят нам всё, ведь составленные нами соотношения для остальных точек сетки (Δ) имеют точность $O(h^2)$.

1) Используем формулы (Д3) и (Д4).

$$\frac{-y_2 + 4y_1 - 3y_0}{2h} = \alpha y_0 + A, \quad \frac{3y_n - 4y_{n-1} + y_{n-2}}{2h} = \beta y_n + B.$$

Они имеют порядок $O(h^2)$, но они испортят трёхдиагональность матрицы системы.

2) «Иногда, если нельзя, но очень хочется, то можно» (применить (Δ) к границам). Тут есть два подспособа:

2а) Вводим фиктивные узлы $y_{-1} \approx y(a-h)$ и $y_{n+1} \approx y(b+h)$ ¹² и применяем к ним (Д5):

$$\frac{y_1 - y_{-1}}{2h} = \alpha y_0 + A, \quad \frac{y_{n+1} - y_{n-1}}{2h} = \beta y_n + B.$$

2б) Снова вводим фиктивные узлы, но на этот раз сдвигаем сетку на $\frac{h}{2}$: $x_k = a - \frac{h}{2} + kh$, где шаг тот же ($h = \frac{b-a}{n}$), а $k \in \{0, \dots, n+1\}$. Разностные соотношения (Δ) будут верны $\forall k \in \{1, \dots, n\}$. Формулу (Д5) можно записать как

$$y'(a) \approx \frac{y(a + \frac{h}{2}) - y(a - \frac{h}{2})}{h} = \frac{y_1 - y_0}{h}.$$

Для этой сдвинутой сетки $y(a)$ будет находиться (по аргументу) между y_0 и y_1 . Применяем линейную интерполяцию: $y(a) = \frac{y_1 + y_0}{2}$. Аналогично с правой границей. Применяем всё это к граничным условиям:

$$\frac{y_1 - y_0}{h} = \alpha \frac{y_1 + y_0}{2} + A, \quad \frac{y_{n+1} - y_n}{h} = \beta \frac{y_{n+1} + y_n}{2} + B.$$

3) Используем ДУ для исключения главной части остатка R в формулах (Д1) и (Д2). Разложим $y(a+h)$ в ряд Тейлора:

$$y(a+h) = y(a) + hy'(a) + \frac{h^2}{2}y''(a) + O(h^3) \Rightarrow \frac{y(a+h) - y(a)}{h} = y'(a) + \frac{h}{2}y''(a) + O(h^2).$$

Из формулы (Д1) имеем:

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

При этом мы можем выразить $y''(a)$ из ДУ задачи (1):

$$y''(a) = -p(a)y'(a) - q(a)y(a) + f(a) = -p(a)\left(\frac{y(a+h) - y(a)}{h} + O(h)\right) - q(a)y(a) + f(a).$$

То есть

$$R = -\frac{h}{2}y''(a) + O(h^2) = \frac{h}{2}\left(-p(a)\frac{y(a+h) - y(a)}{h} - q(a)y(a) + f(a)\right) + O(h^2),$$

¹²Это имеет смысл, поскольку решение ДУ обычно продолжимо за отрезок (снова вспоминаем Басова).

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

Окончательно

$$y'(a) = \frac{y_1 - y_0}{h} \left(1 + \frac{h}{2}p_0\right) + \frac{h}{2}q_0y_0 - \frac{h}{2}f_0.$$

Для правой границы аналогичные действия дают формулу

$$y'(b) = \frac{y_n - y_{n-1}}{h} \left(1 - \frac{h}{2}p_n\right) - \frac{h}{2}q_ny_n + \frac{h}{2}f_n.$$

В эти формулы можем подставить $y'(a)$ и $y'(b)$ из граничных условий (II) и (III).

Таким образом, мы замкнули систему уравнений относительно $\{y_k\}$. Осталось её решить.

6. Метод разностной прогонки.

Запишем систему уравнений относительно $\{y_k\}$ в следующем виде:

$$\begin{cases} -b_0y_0 + c_0y_1 & = d_0, \\ a_ky_{k-1} - b_ky_k + c_ky_{k+1} & = d_k \quad \forall k \in \{1, \dots, n-1\}, \\ a_ny_{n-1} - b_ny_n & = d_n. \end{cases} \quad (16)$$

коэффициенты a_k, b_k, c_k, d_k для $k \in \{1, \dots, n-1\}$ можно найти из разностного соотношения (Δ):

$$a_k = 1 - \frac{h}{2}p_k, \quad b_k = 2 - h^2q_k, \quad c_k = 1 + \frac{h}{2}p_k, \quad d_k = h^2f_k. \quad (17)$$

А вот их граничные значения $b_0, c_0, d_0; a_n, b_n, d_n$ определяются исходя из того, какие разностные соотношения вы используете для границ (в прошлом билете расписали несколько способов).

В матричном виде система (16) выглядит так:

$$\begin{pmatrix} -b_0 & c_0 & & & \\ a_1 & -b_1 & c_1 & & \\ & a_2 & b_2 & c_2 & \\ & & \ddots & \ddots & \ddots \\ & & & a_n & -b_n \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{pmatrix} \quad (18)$$

В матрице системы заполнены только три диагонали, а всё остальное – нули. Такие матрицы называются *трёхдиагональными*. При программной реализации не нужно создавать матрицу целиком, потому что незачем выделять такое большое количество памяти, если большая часть матрицы заполнена нулями. Хранить нужно только 3 диагонали.

Такие СЛАУ можно решать методом разностной прогонки. Сначала идёт прямая прогонка. Поступаем по такой схеме:

$$\begin{pmatrix} * & * & & \\ * & * & * & \\ & * & * & * \\ & & * & * \end{pmatrix} \xrightarrow{\text{делим}} \begin{pmatrix} 1 & * & & \\ * & * & * & \\ & * & * & * \\ & & * & * \end{pmatrix} \xrightarrow{\text{вычитаем}} \begin{pmatrix} 1 & * & & \\ & * & * & * \\ & * & * & * \\ & & * & * \end{pmatrix} \xrightarrow{\text{делим}} \begin{pmatrix} 1 & * & & \\ & 1 & * & \\ & * & * & * \\ & & * & * \end{pmatrix} \rightarrow \dots \rightarrow \begin{pmatrix} 1 & * & & \\ & 1 & * & \\ & & 1 & * \\ & & & 1 \end{pmatrix}$$

Получаем двухдиагональную матрицу с единицами на главной диагонали. Элементы наддиагонали обозначаем как $\{\alpha_k\}_{k=0}^{n-1}$. Не забываем, что вектор-столбец $\{d_k\}$ из (18) тоже подвергается изменениям. Новый вектор невязок обозначим как $\{\beta_k\}$.

Система уравнений с этой матрицей будет состоять из таких соотношений:

$$y_k = \alpha_k y_{k+1} + \beta_k \quad \forall k \in \{0, \dots, n-1\}; \quad y_n = \beta_n. \quad (19)$$

Они похожи на прогоночные соотношения из метода дифференциальной прогонки (2). Выведем, как новые компоненты матрицы выражаются через старые.

$$y_k = \alpha_k y_{k+1} + \beta_k, \quad y_{k-1} = \alpha_{k-1} y_k + \beta_{k-1},$$

подставим y_{k-1} отсюда в среднее равенство из (16) и получим:

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

Таким образом, мы нашли выражение для α_k при $k \in \{1, \dots, n-1\}$ и β_k при $k \in \{1, \dots, n\}$:

$$\alpha_k = \frac{c_k}{b_k - a_k \alpha_{k-1}}, \quad \beta_k = \frac{a_k \beta_{k-1} - d_k}{b_k - a_k \alpha_{k-1}}. \quad (20)$$

А на краю (из первого уравнения системы (16)):

$$y_0 = \frac{c_0}{b_0} y_1 - \frac{d_0}{b_0} \Rightarrow \alpha_0 = \frac{c_0}{b_0}, \quad \beta_0 = -\frac{d_0}{b_0}.$$

Теперь мы знаем все α_k и β_k . На этом прямая прогонка завершена.

Обратная прогонка начинается с того, что $y_n = \beta_n$, а затем находим каждый предыдущий y_k по формуле (19). И на этом всё. Число операций – $8n$ (Хотя мне не совсем понятно, почему. По-моему $8n$ получается только для прямой прогонки. Если ещё добавить обратную, то получим $10n$. В любом случае $O(n)$, что хорошо).

Общее замечание к билету: Преподаватель полагал $a_0 = 0$ и $c_n = 0$, а также $\alpha_n = 0$. У меня получилось обойтись вообще без ввода этих величин, потому что я просто описывал алгоритм. Их можно ввести, если хочется выписывать все уравнения в более общем виде, и/или для упрощения кода.

7. Лемма об оценке для системы разностных уравнений.

Для того, чтобы метод разностной прогонки работал, естественно было бы потребовать, чтобы знаменатели в (20) не обращались в ноль. Полагаем $a_0 = 0$ и $c_n = 0$.

Утверждение. (Достаточное условие). Если

1. $a_k > 0, \quad c_i > 0 \quad \forall k \in \{1, \dots, n\} \quad \forall i \in \{0, \dots, n-1\}$,
2. Матрица системы (18) обладает диагональным преобладанием (т.е. $b_k \geq a_k + c_k \quad \forall k \in \{0, \dots, n\}$, причём хотя бы одно из этих неравенств – строгое),

Тогда метод прогонки осуществим.

□ Хотим доказать, что $0 < \alpha_k \leq 1$. Это, вкупе с условием диагонального преобладания, даст нам то, что в формулах (20) знаменатель точно не обратится в ноль. Докажем $0 < \alpha_k \leq 1$ по индукции:

База: $b_0 \geq c_0 > 0 \Rightarrow 0 < \alpha_0 = \frac{c_0}{b_0} \leq 1$.

Переход: считаем $\alpha_{k-1} \leq 1$, тогда $b_k - a_k \alpha_{k-1} \geq b_k - a_k \geq c_k > 0 \Rightarrow 0 < \alpha_k = \frac{c_k}{b_k - a_k \alpha_{k-1}} \leq 1$. Также, если $\alpha_{k-1} < 1$, то в этой цепочке самый первый знак ' \geq ' заменится на ' $>$ ' и мы получим $\alpha_k < 1$. То есть если неравенство стало строгим, то строгость сохраняется.

Вроде бы всё хорошо, но есть один нюанс. Это свойство ($0 < \alpha_k \leq 1$) мы доказали только для $k \in \{0, \dots, n-1\}$. А при $k = n$ оно ломается, потому что $c_n = 0$, а не больше нуля. С последним знаменателем мы ещё не разобрались. Вспомним условие, что $\exists k_0$: в нём $b_{k_0} > a_{k_0} + c_{k_0}$. Возможны два варианта:

1. $k_0 < n \Rightarrow \alpha_{n-1} < 1 \Rightarrow b_n - a_n \alpha_{n-1} > b_n - a_n \geq c_n = 0$;

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

В обоих случаях получили, что последний знаменатель тоже строго больше нуля \blacksquare .

Если вспомним формулы (17), то можем понять, какие ограничения на систему разностных уравнений накладывает это достаточное условие. Если $h < \frac{2}{\max |p_k|}$, то $a_k > 0$ и $c_k > 0$, как нужно. То есть мы должны выбирать соответствующий мелкий шаг. Также, нужно, чтобы $\boxed{q_k \leq 0}$, тогда получим $b_k \geq a_k + c_k$. Но формулы (17) говорят нам только о середине. Нужно ещё отдельно рассмотреть границы.

Если мы оценим производную на левой границе по простейшей схеме

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

то получим $b_0 = 1 + \alpha h$ и $c_0 = 1$. То есть нам нужно $\boxed{\alpha \geq 0}$. На правой границе, аналогично, $\boxed{\beta \leq 0}$. При этом одно из неравенств, обведённых прямоугольником, должно быть строгим.

Рассмотрим однородную задачу с однородным левым граничным условием ($A = 0$) и $\forall k < n d_k = 0 \Rightarrow \forall k < n \beta_k = 0$. Прогоночные соотношения примут вид

$$y_k = \alpha_k y_{k+1} \Rightarrow \alpha_k = \frac{y_k}{y_{k+1}}.$$

Если случится так, что $y_{k+1} \approx 0$, то получаем очень большой $|\alpha_k|$. Чем меньше сетка, тем больше вероятность попасть в такую окрестность нуля. Катастрофы в этом нет, просто следим за $\max |\alpha_k|$. Если он получается очень большой, сетку можно немного подвинуть.

Лемма. Пусть достаточное условие осуществимости прогонки выполнено в усиленном виде:

$$a_k > 0, c_k > 0, \exists \delta > 0 : b_k \geq a_k + c_k + \delta,$$

Тогда система (16) однозначно разрешима \forall правой части и справедлива оценка:

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

\square Пусть мы не знаем об однозначной разрешимости, но имеем некое решение $\{y_k\}$. Тогда (считаем $a_0 = 0$ и $c_n = 0$) для него выполнены соотношения (16):

$$a_k y_{k-1} - b_k y_k + c_k y_{k+1} = d_k \quad \forall k \in \{0, \dots, n\}.$$

Вводим $M = \max |y_k| = |y_{k_0}|$ (такое $k_0 \exists$, потому что $\{y_k\}$ конечно).

$$b_{k_0} y_{k_0} = a_{k_0} y_{k_0-1} + c_{k_0} y_{k_0+1} - d_{k_0},$$

берём от этого равенства модуль и получаем:

$$b_{k_0} M \leq a_{k_0} M + c_{k_0} M + |d_{k_0}| \Rightarrow$$

$$\delta M \leq (b_{k_0} - a_{k_0} - c_{k_0}) M \leq |d_{k_0}| \leq \max |d_k|.$$

То есть $M \leq \delta^{-1} \max |d_k|$. Если система однородная и с однородными граничными условиями (то есть все $d_k = 0$), то получаем, что все $y_k = 0$. То есть единственное решение однородной задачи – тривиальное. По **Теореме** из первого билета это означает, что неоднородная система однозначно разрешима \blacksquare .

8. Теорема о сходимости разностного метода для обыкновенной краевой задачи.

Теорема. Рассматриваем краевую задачу (1) с (III) граничным условием. Пусть $y(x) \in C^4([a, b])$ – единственное решение этой задачи. Пусть выполнены следующие условия на коэффициенты:

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

Тогда система однозначно разрешима и её точность на сетке можно оценить как:

$$\forall h \exists C = \text{const} : \forall k \in \{0, \dots, n\} \quad |y(x_k) - y_k| \leq Ch^2.$$

□ Вычтем из формулы (Δ) формулу (15), попутно введя обозначение $w_k = y_k - y(x_k)$. Получим

$$\frac{w_{k+1} - 2w_k + w_{k-1}}{h^2} + p_k \frac{w_{k+1} - w_{k-1}}{2h} + q_k w_k = R_k,$$

где $k \in \{1, \dots, n\}$.¹³ Первый пункт в условиях теоремы нужен, чтобы были определены числа M_4 и M_3 , и, как следствие, остатки R'_k и R''_k ¹⁴ из формул численного дифференцирования (Д5) и (Д6). Оценку остатка $R_k = R''_k + p_k R'_k$, как нетрудно убедиться из (Д5) и (Д6), можно написать в виде $|R_k| \leq L_k h^2$, где L_k – некоторая константа.

На границах воспользуемся формулами оценки производной с фиктивными узлами и сдвинутой на $h/2$ сеткой. Для ошибок получим формулы того же вида, но в правой части будут стоять R_0 и R_{n+1} , которые содержат ещё и интерполяционную ошибку.

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

Важно уточнить, что оценивание производной здесь производится не по простейшим формулам (Д1) и (Д2), как может показаться, а по формуле (Д5), в которой вместо h используется $h/2$. Таким образом остаток от этих формул численного дифференцирования здесь будет оцениваться как $\frac{M_3}{24} h^2$. Также остатки R_0 и R_{n+1} содержат ошибку интерполяции¹⁵, которая равна $M_2 h^2$. но, поскольку вместо h у нас в данной формуле $h/2$, то получим $\frac{M_2}{4} h^2$. В любом случае, опять получаем константы, умноженные на h^2 . То есть $|R_0| \leq L_0 h^2$ и $|R_{n+1}| \leq L_{n+1} h^2$.

Рассмотрим систему относительно $\{w_k\}$. Её коэффициенты (для $k \in \{1, \dots, n\}$) равны

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

Подгоним эту систему под условия Леммы:

$$b_k \geq a_k + c_k + \delta \Leftrightarrow q_k \leq -\delta.$$

В условии сказано, что $q_k \leq -q_0$. Если $\delta \leq q_0$, то получим и $q_k \leq -\delta$.

А на границах:

$$\begin{aligned} a_0 &= 0, & b_0 &= \frac{1}{h} + \frac{\alpha}{2}, & c_0 &= \frac{1}{h} - \frac{\alpha}{2}, & d_0 &= R_0, \\ a_{n+1} &= -\frac{1}{h} - \frac{\beta}{2}, & b_{n+1} &= \frac{\beta}{2} - \frac{1}{h}, & c_{n+1} &= 0, & d_{n+1} &= R_{n+1}. \end{aligned}$$

¹³ k меняется именно до n , а не до $n - 1$, потому что далее выяснится, что мы используем сетку, сдвинутую на полушаг.

¹⁴Эти штрихи – это не производные, а просто отличительные символы.

¹⁵Вывести её оценку можно тем же путём, как мы в 5 билете выводили оценку для (Д5). Только здесь нужно будет раскладывать $y(x + h)$ и $y(x - h)$ в формулу Тейлора с остатком в форме Лагранжа до второго порядка.

$$b_0 \geq a_0 + c_0 + \delta \Leftrightarrow \alpha \geq \delta,$$

$$b_{n+1} \geq a_{n+1} + c_{n+1} + \delta \Leftrightarrow -\beta \geq \delta.$$

Получили, что δ должно быть $\leq q_0$, α и $-\beta$ одновременно. Значит $\delta = \min\{q_0, \alpha, \beta\}$. Тогда условие Леммы выполняется.

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

То есть выбираем $C = \delta^{-1} \max(L_k)$ и на этом теорема доказана \blacksquare .

Далее шли какие-то непонятные слова про существование, оценку обратной матрицы и плохую обусловленность. Также там сказано, что компьютерная точность оценивается как $O(\varepsilon/h^2)$, в то время как точность метода – $O(h^2)$. Получается, что при очень маленьком шаге h ошибки компьютерной точности будут значительны. Обозначим ошибку как $C_1 \frac{\varepsilon}{h^2} + C_2 h^2$ и найдём минимум этой функции. Получим $h \sim \varepsilon^{1/4}$.

А ещё дальше говорили про разностный метод для задачи более высокого порядка. Я это не выписывал здесь, ибо не уверен, нужно ли. Возможно потом вынесу в приложения.

9. Жесткие системы ОДУ. Простейшие методы. Понятие А-устойчивости.

Сначала рассмотрим два примера, а определение приведём потом.

Пример 1. Рассмотрим дифференциальное уравнение

$$y' = Ay, \quad A < 0$$

с начальными данными $y(0) = 1$. Решением такой задачи Коши будет

$$y(x) = e^{Ax}.$$

Проинтегрируем это уравнение простейшим методом Эйлера (который явный):

$$y_{n+1} = y_n + hf_n, \quad f_n = f(x_n, y_n).$$

$$y_{n+1} = y_n + hAy_n = (1 + Ah)y_n \Rightarrow y_n = (1 + Ah)^n.$$

Если $nh \xrightarrow{n \rightarrow \infty} x_n$, то (Вспоминаем определение экспоненты через предел) $y_n \xrightarrow{n \rightarrow \infty} e^{Ax_n}$, то есть ряд сходится.

А теперь допустим, что $A \ll 0$. Пусть, к примеру, $A = -100, h = 0,1$. Тогда $(1 + Ah) = -9$, численное решение будет вести себя как $(-9)^n$. То есть это будут растущие осцилляции. Совсем не похоже на решение e^{-100x} , которое является просто убывающей экспонентой. Если же $h = 0,001$, то численное решение получится $(0,9)^n$.

Если A велико по модулю и отрицательно, то h должен быть малым, чтобы поведение численного решения было приличным. Должно быть $|A|h < 1$ (Даже \ll , чтобы ряд норм сходился).

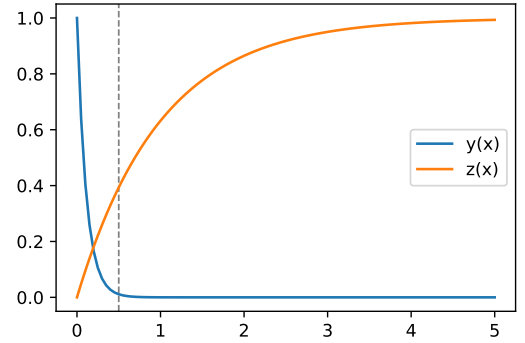
Пример 2. Теперь рассмотрим систему и начальные данные:

$$\begin{cases} y' = Ay, & y(0) = 1; \\ z' = -z + 1, & z(0) = 0. \end{cases}$$

Решение системы с этими начальными данными:

$$\begin{cases} y = e^{Ax}, \\ z = 1 - e^{-x}. \end{cases}$$

Взглянем на график этого решения (при $A = -9$) слева. Видно, что z меняется плавно, а y – резко. Причём чем больше A , тем более резким будет изменение y . Брать изначально большой шаг – неправильно, получим большие ошибки. При этом использовать везде маленький – накладно.



Определение. Жёсткая система ОДУ – такая система, в которой есть как быстрое убывание, так и плавное изменение, причём одновременно.¹⁶

Снова обратимся к уравнению $y' = Ay$ из примера 1, но на этот раз применим неявный метод Эйлера:

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}).$$

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

Если $A < 0$, то $0 < \frac{1}{1-hA} < 1$, то есть получаем сходимость. Этот метод как раз можно использовать на участке справа от пунктирной линии, когда y близок к нулю, и тогда не получим никаких осцилляций.

И вообще, приемлемые для жёстких систем методы следует искать среди неявных!

Рассматривая уравнение $y' = Ay$ и применяя к нему разные методы, будем получать формулу вида

$$y_{n+1} = R(Ah)y_n,$$

где $R(z)$ – некоторая функция комплексного аргумента.¹⁷ Для явного метода Эйлера $R(z) = 1 + z$. Для неявного – $R(z) = \frac{1}{1-z}$.

Определение. Метод численного решения называется A -устойчивым, если его $R(z)$ удовлетворяет условию $|R(z)| \leq 1$ при $\operatorname{Re}(z) \leq 0$ (по-пучковому: образ замыкания левой полуплоскости лежит внутри замкнутого единичного круга).

Теперь рассмотрим несколько методов интегрирования.¹⁸

1. Явный метод Эйлера – нет A -устойчивости

$$y_{n+1} = y_n + hf(x_n, y_n), \quad R(z) = 1 + z.$$

2. Неявный метод Эйлера – A -устойчив

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}), \quad R(z) = \frac{1}{1-z}.$$

3. Улучшенный метод Эйлера I – нет A -устойчивости

$$\begin{aligned} \tilde{y}_{n+1} &= y_n + hf(x_n, y_n), \\ y_{n+1} &= y_n + hf(x_{n+1}, \tilde{y}_{n+1}), \end{aligned} \quad R(z) = 1 + z + z^2.$$

Стоит отметить, что этот метод тоже явный.

¹⁶Это скорее размахивание руками. За строгим определением лучше обратиться к другим источникам.

¹⁷Рассматриваем комплексный аргумент, потому что в \mathbb{C} проще говорить о всяких сходимостях. А ещё иногда A может быть комплексным.

¹⁸Как я понял, функция $R(z)$ находится с предположением, что мы рассматриваем систему $y' = Ay$.

4. Улучшенный метод Эйлера II – не A-устойчивости

$$\begin{aligned} y_{n+\frac{1}{2}} &= y_n + \frac{h}{2}f(x_n, y_n), \\ y_{n+1} &= y_n + hf\left(x_n + \frac{h}{2}, y_{n+\frac{1}{2}}\right), \end{aligned} \quad R(z) = 1 + z + \frac{z^2}{2}.$$

И этот метод также является явным.

5. Метод средних прямоугольников I – A-устойчив

$$\begin{aligned} y_{n+\frac{1}{2}} &= y_n + \frac{h}{2}f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}\right), \\ y_{n+1} &= y_n + hf\left(x_n + \frac{h}{2}, y_{n+\frac{1}{2}}\right), \end{aligned} \quad R(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}.$$

6. Метод трапеций – A-устойчив

$$y_{n+1} = y_n + \frac{h}{2}(f(x_n, y_n) + f(x_{n+1}, y_{n+1})), \quad R(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}.$$

7. Метод средних прямоугольников II – A-устойчив

$$y_{n+1} = y_n + hf\left(x_n + \frac{h}{2}, \frac{y_n + y_{n+1}}{2}\right), \quad R(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}.$$

8. Весовая формула прямоугольников – A-устойчива при $\theta \geq \frac{1}{2}$

$$y_{n+1} = y_n + hf(x_n + \theta h, y_n + \theta(y_{n+1} - y_n)), \quad R(z) = \frac{1 + (1 - \theta)z}{1 - \theta z}.$$

9. Весовая формула трапеций¹⁹ – A-устойчива при $\theta \geq \frac{1}{2}$

$$y_{n+1} = y_n + h((1 - \theta)f(x_n, y_n) + \theta f(x_{n+1}, y_{n+1})), \quad R(z) = \frac{1 + (1 - \theta)z}{1 - \theta z}.$$

Методы 1, 2 – первого порядка. Методы 3-9 – второго. Порядок этот определяется так:

Пусть $y(x_n) = y_n$, тогда $y_{n+1} = y(x_{n+1}) + \mathcal{R}$, где \mathcal{R} – какой-то остаток (не путать с введённой ранее функцией $R(z)$, это совсем разные вещи). Тогда если \mathcal{R} допускает оценку $|\mathcal{R}| \leq kh^{p+1}$, то метод имеет порядок p .

10. Понятие L-устойчивости. Неявные методы Рунге-Кутты, общее понятие. Диагонально-неявные методы.

Среди методов более высокого порядка выделяют многоточечные методы (о них позднее) и методы Рунге-Кутты. Нам, в свою очередь, нужны неявные аналоги методов РК. Будем представлять методы Рунге-Кутты в виде таблицы. Например, классический метод РК ранга q выглядит так:

$$\begin{array}{c|cccc} \alpha_1 = 0 & \beta_{11} = 0 & & & \\ \alpha_2 & \beta_{21} & & & \\ \vdots & \vdots & \ddots & & \\ \alpha_q & \beta_{q1} & \cdots & \beta_{q,q-1} & \\ \hline & \gamma_1 & \cdots & \gamma_{q-1} & \gamma_q \end{array}$$

¹⁹В нашем случае преподаватель поменял θ и $1 - \theta$ местами. При этом соответственно поменяется и вид функции $R(z)$, а A-устойчивость возникнет при $\theta \leq \frac{1}{2}$.

На основании такой таблицы находятся коэффициенты k_i :

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}(x_n + \alpha_1 h, \mathbf{y}_n + h\beta_{11}\mathbf{k}_1) = \mathbf{f}(x_n, \mathbf{y}_n), \\ \mathbf{k}_2 &= \mathbf{f}(x_n + \alpha_2 h, \mathbf{y}_n + h\beta_{21}\mathbf{k}_1), \\ \mathbf{k}_i &= \mathbf{f}\left(x_n + \alpha_i h, \mathbf{y}_n + h \sum_{j=1}^{i-1} \beta_{ij}\mathbf{k}_j\right), \\ \mathbf{k}_q &= \mathbf{f}\left(x_n + \alpha_q h, \mathbf{y}_n + h \sum_{j=1}^{q-1} \beta_{qj}\mathbf{k}_j\right). \end{aligned}$$

а уже с их помощью находится \mathbf{y}_{n+1} :

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h(\gamma_1 \mathbf{k}_1 + \dots + \gamma_q \mathbf{k}_q).$$

Как видим, в явном методе РК матрица $\hat{\beta} = \{\beta_{ij}\}$ имеет нули на главной диагонали и над ней. Самый общий неявный метод полагает, что матрица может быть заполнена вся целиком:

$$\begin{array}{c|ccc} \alpha_1 & \beta_{11} & \dots & \beta_{1q} \\ \alpha_2 & \beta_{21} & \dots & \beta_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_q & \beta_{q1} & \dots & \beta_{qq} \\ \hline & \gamma_1 & \dots & \gamma_q \end{array} \quad \begin{aligned} \mathbf{k}_i &= \mathbf{f}\left(x_n + \alpha_i h, \mathbf{y}_n + h \sum_{j=1}^q \beta_{ij}\mathbf{k}_j\right), \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + h \sum_{i=1}^q \gamma_i \mathbf{k}_i. \end{aligned}$$

Если исходная система имела s уравнений, то эта система будет иметь sq уравнений.

Вообще, все эти коэффициенты $\{\alpha_i\}$, $\{\beta_{ij}\}$, $\{\gamma_i\}$ берутся не с потолка. Во-первых,

$$\sum_{j=1}^q \beta_{ij} = \alpha_i, \quad \sum_{i=1}^q \gamma_i = 1,$$

то есть сумма всех β_{ij} , которые лежат в одной (i -ой) строке, должна быть равна соответствующей α_i . А также сумма всех γ_i должна равняться единице.

А во-вторых, каждый метод РК основан на какой-нибудь квадратурной формуле.²⁰ Например, рассмотрим формулы

$$\int_0^1 f(x)dx \approx \frac{1}{4} \left(f(0) + 3f\left(\frac{2}{3}\right) \right); \quad \int_0^1 f(x)dx \approx \frac{1}{4} \left(3f\left(\frac{1}{3}\right) + f(1) \right).$$

Соответствующие им методы РК будут иметь коэффициенты:

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ 2 & \frac{1}{4} & \frac{5}{12} \\ 3 & \frac{1}{4} & \frac{3}{4} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array} \quad \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

²⁰Преподаватель нам просто объявил об этом. Я пытался понять, как именно это происходит, но так и не понял один момент: откуда берутся β_{ij} . Видимо, они имеют смысл весов и их можно брать произвольно, главное чтобы их сумма по j равнялась α_i .

11. Асимптотический метод в задаче о быстрых колебаниях.

Примечание: В этот билет я также включил многошаговые методы и понятие $A(a)$ -устойчивости.

Общий вид многошаговых методов такой:

$$\sum_{i=-p}^1 \alpha_i \mathbf{y}_{n+i} = h \sum_{j=-q}^1 \beta_j \mathbf{f}(x_{n+j}, \mathbf{y}_{n+j}).$$

Пример – известные нам методы Адамса. Все многошаговые методы не имеют A -устойчивости, за единственным исключением – методом дифференцирования назад второго порядка.

Обычно мы получали методы интегрирования путём аппроксимирования интеграла:

$$\mathbf{y}(x_{n+1}) = \mathbf{y}(x_n) + \int_{x_n}^{x_{n+1}} \mathbf{f}(x', \mathbf{y}(x')) dx'$$

(Именно этот интеграл мы заменяем квадратурной формулой при выводе методов Рунге-Кутты). Вместо этого мы можем аппроксимировать производную:

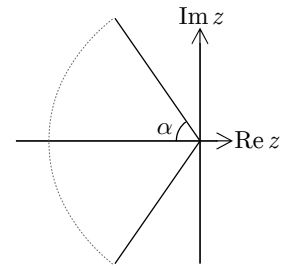
$$\frac{3\mathbf{y}_{n+1} - 4\mathbf{y}_n + \mathbf{y}_{n-1}}{2h} = \mathbf{f}(x_{n+1}, \mathbf{y}_{n+1}).$$

Мне хотелось бы проверить его A -устойчивость, но я не уверен, как определяется $R(z)$ для многошагового метода. Если как $y_{n+1} = R(Ah)y_n + \dots$ при условии рассмотрения уравнения $y' = Ay$, то тогда получаем $R(z) = \frac{4}{3-2z}$, и метод действительно A -устойчив.

Определение. Требование A -устойчивости в некотором смысле излишнее. Если $z = Ah$ – лежит близко к мнимой прямой, то $e^{Ah} \approx e^{i \operatorname{Im} Ah}$ – это решение не быстро растущее, а быстро колеблющееся. Поэтому мы хотим, чтобы z удовлетворяло:

$$|\arg z + \pi| \leq \alpha$$

(Это означает, что z лежит в секторе, отмеченном на рисунке справа). Тогда метод называется $A(\alpha)$ -устойчивым



Для колеблющихся систем нужно применять другие приёмы. Число точек интегрирования должно быть пропорционально частоте колебаний, то есть их должно быть много. Их лучше оценивать не в деталях, а в «общем виде». Для такого рода задач применяются *асимптотические методы*.

Определение. Рассмотрим ряд²¹ по малому параметру:

$$F(\alpha) \sim \sum_{n=0}^{\infty} a_n \alpha^n.$$

Этот ряд называется *асимптотическим*, если

$$\forall N \in \mathbb{N} \quad \exists C_N > 0 : \left| F(\alpha) - \sum_{n=0}^N a_n \alpha^n \right| \leq C_N \alpha^{N+1}.$$

Пример. Рассмотрим функцию $\sin \omega x$. Возмутим частоту: $\omega + \varepsilon$. Тогда

$$\underbrace{\sin(\omega x + \varepsilon x)}_{\text{Колеблется}} = \sin \omega x + \underbrace{\varepsilon x \cos \omega x}_{\text{Растёт}} + \dots$$

²¹В этой формуле α следует воспринимать не как константу, а как некую функцию от x . Вообще, определение асимптотического ряда в других источниках другое. Обычно раскладывают не по степеням α^n , а по семейству функций $\{\alpha_n\}$.

Последующие слагаемые будут ещё больше расти. Такое разложение неприемлемо.

Рассмотрим уравнение

$$y'' + \omega^2 y = \varepsilon f(y, y').$$

Невозмущённое уравнение (при $\varepsilon = 0$) описывается гармоническими колебаниями:

$$y(x) = a \cos \psi, \quad \frac{dy}{dx} = -a\omega \sin \psi, \quad \psi(x) = \omega x + \theta, \quad a, \theta = \text{const.}$$

Решение исходного уравнения ищем в виде

$$y(x) = a(x) \cos \psi(x), \quad \frac{dy}{dx} = -\omega a(x) \sin \psi(x), \quad \psi(x) = \omega x + \theta(x)$$

(Это мы просто так постулировали). Считаем, что $a(x)$ медленно меняется, а $\theta(x)$ просто мало.

Сейчас будет продемонстрирован асимптотический метод Крылова-Боголюбова. Продифференцируем y по x честно и получим:

$$\frac{dy}{dx} = \frac{da}{dx} \cos \psi - \omega a \sin \psi - \omega a \sin \psi \frac{d\theta}{dx}.$$

Вспомним, что мы ищем решение в такой форме, что dy/dx и $-\omega a \sin \psi$ просто сокращаются. Тогда

$$\frac{da}{dx} \cos \psi - \omega a \sin \psi \frac{d\theta}{dx} = 0, \quad (21)$$

$$\frac{d^2 y}{dx^2} = -\omega \frac{da}{dx} \sin \psi - \omega^2 a \cos \psi - \omega a \cos \psi \frac{d\theta}{dx} = -\omega^2 a \cos \psi + \varepsilon f(a \cos \psi, -a\omega \sin \psi).$$

В конце концов, получили

$$-\omega \frac{da}{dx} \sin \psi - \omega a \frac{d\theta}{dx} \cos \psi = \varepsilon f. \quad (22)$$

Из уравнений (21) и (22) можем получить систему:

$$\begin{cases} \frac{da}{dx} = -\frac{\varepsilon}{\omega} f(a \cos \psi, -a\omega \sin \psi) \sin \psi, \\ \frac{d\theta}{dx} = -\frac{\varepsilon}{\omega a} f(a \cos \psi, -a\omega \sin \psi) \cos \psi. \end{cases} \quad (A_{11})$$

Но это ещё не та система, которая нужна нам, потому что сюда входит ψ . Рассмотрим правые части как функции независимых переменных:

$$\begin{aligned} f(a \cos \chi, -a\omega \sin \chi) \sin \chi, \\ f(a \cos \chi, -a\omega \sin \chi) \cos \chi, \end{aligned}$$

разложим их в ряды Фурье:

$$\begin{aligned} f(a \cos \chi, -a\omega \sin \chi) \sin \chi &= \sum_{n=0}^{\infty} \alpha_n^{(1)}(a) \cos n\chi + \beta_n^{(1)}(a) \sin n\chi, \\ f(a \cos \chi, -a\omega \sin \chi) \cos \chi &= \sum_{n=0}^{\infty} \alpha_n^{(2)}(a) \cos n\chi + \beta_n^{(2)}(a) \sin n\chi. \end{aligned}$$

Нулевые члены разложения:

$$\begin{aligned} \alpha_0^{(1)}(a) &= \frac{1}{2\pi} \int_0^{2\pi} f(a \cos \chi, -a\omega \sin \chi) \sin \chi d\chi, \\ \alpha_0^{(2)}(a) &= \frac{1}{2\pi} \int_0^{2\pi} f(a \cos \chi, -a\omega \sin \chi) \cos \chi d\chi. \end{aligned}$$

Можем определить \bar{a} и $\bar{\theta}$, удовлетворяющие системе:

$$\begin{cases} \frac{d\bar{a}}{dx} = -\frac{\varepsilon}{2\pi\omega} \int_0^{2\pi} f(\bar{a}(\chi) \cos \chi, -\bar{a}(\chi)\omega \sin \chi) \sin \chi d\chi, \\ \frac{d\bar{\theta}}{dx} = -\frac{\varepsilon}{2\pi\omega\bar{a}(\chi)} \int_0^{2\pi} f(\bar{a}(\chi) \cos \chi, -\bar{a}(\chi)\omega \sin \chi) \cos \chi d\chi. \end{cases} \quad (\text{B}_{11})$$

Решения этой системы, \bar{a} и $\bar{\theta}$, можно подставить (зачем?) в систему (A₁₁). Но мы рассмотрели только нулевой член разложения. Далее, подставим решения системы (B₁₁) в ряды Фурье для системы (A₁₁):

$$g_1(x) \equiv \int \left(\sum_{n \neq 0} \alpha_n^{(1)}(\bar{a}) \cos n\psi + \beta_n^{(1)}(\bar{a}) \sin n\psi \right) dx,$$

где $\psi(x) = \omega x + \bar{\theta}(x)$. Возьмём интеграл грубо, игнорируя изменения $\bar{\theta}$ и \bar{a} :

$$g_1(x) \approx \frac{1}{\omega} \sum_{n \neq 0} \frac{1}{n} \alpha_n^{(1)}(\bar{a}) \sin n\psi - \frac{1}{n} \beta_n^{(1)}(\bar{a}) \cos n\psi.$$

Формула для $g_2(x)$ выглядит аналогично, только с двойками. В качестве решения принимаем

$$\begin{aligned} a(x) &= \bar{a}(x) - \frac{\varepsilon}{\omega} g_1(x), \\ \theta(x) &= \bar{\theta}(x) - \frac{\varepsilon}{a\omega} g_2(x). \end{aligned}$$

Эти приближения хороши в первом приближении по ε .

II. Задачи на собственные значения.

12. Вопрос об устойчивости собственных чисел и собственных векторов при возмущении матрицы. Отрицательный пример.

Сначала вспомним несколько фактов из Алгебры.

Определение. Собственные числа матрицы \hat{A} – это такие λ , для которых $\exists \mathbf{u} : \hat{A}\mathbf{u} = \lambda\mathbf{u}$. \mathbf{u} – собственный вектор матрицы \hat{A} , соответствующий собственному числу λ . Далее я не буду писать «шапочки» над матрицами и использовать жирное начертание для векторов.

Собственные числа можно искать как корни характеристического уравнения $p_s(\lambda) = \det(A - \lambda E) = 0$, где s – это порядок²² матрицы. $p_s(\lambda)$ – это характеристический полином. Он имеет корни $\lambda_1, \dots, \lambda_s$, каждому из которых соответствует свой собственный вектор. Если все λ_i разные, то все u_i тоже разные и линейно-независимые.

Определение. Алгебраическая кратность λ – кратность λ как корня $p_s(\lambda)$.

Геометрическая кратность λ – число линейно-независимых собственных векторов, ему соответствующих (Или размерность его собственного пространства, что одно и то же).

Всякую матрицу можно привести к Жордановой Нормальной Форме

$$C^{-1}AC = \Lambda = \begin{pmatrix} \boxed{J_{p_1}(\lambda_1)} & & \\ & \boxed{J_{p_2}(\lambda_2)} & \\ & & \ddots \\ & & & \boxed{J_{p_s}(\lambda_s)} \end{pmatrix}, \text{ где } J_p(\lambda) = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \lambda & 1 \\ & & & \lambda \end{pmatrix}_{p \times p}.$$

$J_p(\lambda)$ называется Жордановой клеткой. Может быть несколько Жордановых клеток с одинаковым λ .

Если матрица симметрична²³, то у неё вещественны все собственные числа. Также, её набор собственных векторов ортогонален. Эту матрицу можно привести к диагональному виду.

Есть теорема о том, что корни многочлена – непрерывные функции от его коэффициентов. При этом мы видим, что коэффициенты $p_s(\lambda)$ непрерывно зависят от элементов матрицы. Следовательно, *Собственные числа матрицы непрерывно зависят от её элементов.*

Пример. Рассмотрим матрицу

$$A = \begin{pmatrix} a & 1 & & \\ & a & \ddots & \\ & & \ddots & 1 \\ & & & a \end{pmatrix}; \det(A - \lambda E) = \begin{vmatrix} a - \lambda & 1 & & \\ & a - \lambda & \ddots & \\ & & \ddots & 1 \\ & & & a - \lambda \end{vmatrix} = (a - \lambda)^s, \lambda = a - \text{корень степени } s.$$

А теперь рассмотрим её возмущение вида

$$A + \varepsilon B = \begin{pmatrix} a & 1 & & \\ & a & \ddots & \\ & & \ddots & 1 \\ \varepsilon & & & a \end{pmatrix}; \det(A + \varepsilon B - \lambda E) = \begin{vmatrix} a - \lambda & 1 & & \\ & a - \lambda & \ddots & \\ & & \ddots & 1 \\ \varepsilon & & & a - \lambda \end{vmatrix} =$$

$$= (a - \lambda)(a - \lambda)^{s-1} \pm \varepsilon = 0 \Rightarrow \lambda = a \pm \varepsilon^{\frac{1}{s}},$$

То есть $\lambda_k = a + \omega_k \varepsilon^{\frac{1}{s}}$, где $|\omega_k| = 1$.

²²Имеется в виду, что матрица размера $s \times s$.

²³Тут можно было бы говорить и более общими словами, про самосопряжённые матрицы. Но раз мы в \mathbb{R} , то для нас и транспонирование – сопряжение. Такое замечание уже было в билете 4.

Пусть $\varepsilon = 10^{-16}$, тогда при $s = 16$ $\varepsilon^{\frac{1}{s}} = 10^{-1}$ – весьма грустный результат. При $s = 32$ получим $\varepsilon^{\frac{1}{s}} = \sqrt{0.1} \approx 0.3$. Малое возмущение матрицы породило большое возмущение собственного числа.

При этом у собственных векторов нет даже непрерывной зависимости от элементов матрицы (но про них ещё будет отдельный билет).

13. Теорема Бауэра-Файка о возмущении собственных чисел симметричной матрицы.

Определение. Хотелось бы напомнить, что такое норма матрицы²⁴:

$$\|A\| = \sup_{\mathbf{u} \neq \mathbf{0}} \left\{ \frac{\|A\mathbf{u}\|}{\|\mathbf{u}\|} \right\},$$

где $\|\mathbf{u}\|$ и $\|A\mathbf{u}\|$ – какая-то выбранная заранее норма в соответствующих векторных пространствах (которые являются областью определения и областью значений матрицы A как линейного отображения). В частности, если эти нормы являются евклидовыми, то соответствующую норму матрицы тоже называют евклидовой (она обозначается как $\|A\|_2$). Для комплексных матриц справедливо

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

Если же $A = A^* = A'$, то

$$\|A\|_2 = \max |A_{ij}|.$$

Для операторной нормы справедливо следующее свойство:

$$\|AB\| \leq \|A\| \|B\|.$$

Лемма. Матрица $E + T$ имеет обратную, если $\|T\| < 1$.²⁵ При этом существует такая оценка нормы:

$$\|(E + T)^{-1}\| \leq \frac{1}{1 - \|T\|}.$$

□ Пусть задан некий \mathbf{y} и мы решаем систему $(E + T)\mathbf{x} = \mathbf{y}$ методом простой итерации ($\mathbf{x}_{n+1} = f(\mathbf{x}_n)$). Для этого введём функцию $f(\mathbf{x}_n) = \mathbf{y} - T\mathbf{x}_n$. Покажем, что это сжимающее отображение.

$$\|f(\mathbf{x}) - f(\mathbf{x}')\| = \|T(\mathbf{x}' - \mathbf{x})\| \leq \|T\| \|\mathbf{x}' - \mathbf{x}\| < \|\mathbf{x}' - \mathbf{x}\|.$$

То есть мы показали, что метод сходится и решение существует $\forall \mathbf{y}$. Если мы вместо \mathbf{y} подставим строки единичной матрицы, то в качестве \mathbf{x} получим столбцы обратной матрицы. Таким образом, обратная матрица существует.

Чтобы оценить норму обратной матрицы, распишем следующую норму:

$$\mathbf{x} = \mathbf{y} - T\mathbf{x} \Rightarrow \|\mathbf{x}\| \leq \|\mathbf{y}\| + \|T\| \|\mathbf{x}\| \Rightarrow \frac{\|\mathbf{x}\|}{\|\mathbf{y}\|} \leq \frac{1}{1 - \|T\|}.$$

И это верно \forall пары \mathbf{x} и \mathbf{y} . Поскольку $\mathbf{x} = (E + T)^{-1}\mathbf{y}$, то $\max \frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}$ – это норма матрицы $(E + T)^{-1}$ по определению \blacksquare .

Теперь речь пойдёт о симметричных матрицах. Для них ситуация не такая грустная: возмущение собственного числа не превосходит евклидовой нормы матрицы возмущения. Хочу напомнить, что собственные числа симметричной матрицы вещественны, набор собственных векторов – ортогонален, и его можно даже сделать ортонормированным.

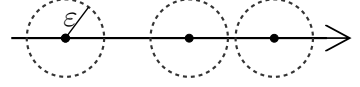
²⁴Иногда её ещё называют *операторной нормой*.

²⁵Эту Лемму преподаватель назвал теоремой Банаха и указал, что это учить не нужно.

Теорема. Пусть A – симметричная матрица, а B – возмущение, как в предыдущем билете. Пусть λ_{A+B} – собственное число матрицы $A + B$. Обозначим $\varepsilon = \|B\|_2$. Тогда

$$\forall \lambda_{A+B} (\in \mathbb{C}) \exists \lambda_A (\in \mathbb{R}) : |\lambda_A - \lambda_{A+B}| \leq \varepsilon.$$

□ На рисунке справа изображён спектр матрицы A (то есть её собственные числа), и его ε -окрестность (то есть объединение ε -окрестностей всех собственных чисел). Так вот всякое z вне этой окрестности не может быть собственным числом матрицы $A + B$. То есть можем переформулировать условие таким образом:



$$\forall z \in \mathbb{C} : \text{если } \forall \lambda_A \quad |z - \lambda_A| > \varepsilon \Rightarrow z \text{ не является собственным числом } A + B.$$

Будем доказывать именно это условие. Возьмём такой z . Нам нужно доказать, что $\det(A + B - zE) \neq 0$.

Матрица A симметрична, значит её можно привести к каноническому (диагональному) виду, то есть $\exists U$ – ортогональная:²⁶ $U'AU = \Lambda = \text{diag}\{\lambda_k\}$.

$$A + B - zE = U(U'AU + U'BU - zE)U' = U(\Lambda - zE + U'BU)U'.$$

Матрицы U, U' неособые. Если докажем неособость того, что в скобках, то докажем теорему.

$$\Lambda - zE + U'BU = (\Lambda - zE)(1 + (\Lambda - zE)^{-1}U'BU).$$

Матрица $(\Lambda - zE)$ неособая по условию (мы рассмотрели такой z , что $|z - \lambda_A| > \varepsilon$). Чтобы воспользоваться условием Леммы, нужно доказать, что $\|(\Lambda - zE)^{-1}U'BU\|_2 < 1$.

$$\|(\Lambda - zE)^{-1}\|_2 = \|\text{diag}\{(\lambda_k - z)^{-1}\}\|_2 = \max \frac{1}{|\lambda_k - z|} < \frac{1}{\varepsilon}.$$

Так как $U'U = E$, то все собственные числа $U'U$ – единицы. Значит $\|U\|_2 = \|U'\|_2 = 1$. Тогда

$$\|U'BU\|_2 \leq \|U\|_2 \|B\|_2 \|U'\|_2 = \|B\|_2 = \varepsilon.$$

В конце концов получим $\|(\Lambda - zE)^{-1}U'BU\|_2 < \frac{1}{\varepsilon}\varepsilon = 1$ ■.

Замечание. Если в условии теоремы положить, что A – не симметричная, но при этом диагонализированная матрица, то есть $\exists X : X^{-1}AX = \Lambda = \text{diag}\{\lambda_k\}$. Но здесь X уже не будут ортогональными, и тогда при оценке нормы $\|X^{-1}BX\|_2$ у нас не исчезнет множитель²⁷ $\|X^{-1}\|_2 \|X\|_2 = \mu(X)$. И в последних двух строчках доказательства теоремы мы получим

$$\|X^{-1}BX\|_2 \leq \|X^{-1}\|_2 \|B\|_2 \|X\|_2 = \mu(X)\varepsilon \equiv d.$$

В условии теоремы нужно поставить $|\lambda_A - \lambda_{A+B}| \leq d$ и тогда её доказательство также будет справедливым. Таким образом, мы получили, что для диагонализированных матриц собственные числа расположены в d -окрестности спектра. Но это не страшно, поскольку d линейно зависит от ε .

14. Устойчивость собственных векторов при возмущении матрицы.

Как мы уже говорили ранее, в общем случае у собственных векторов нет даже непрерывной зависимости от элементов матрицы. При этом разрывы возникают в окрестности кратных собственных чисел. Тогда было бы логично рассмотреть матрицы, у которых все собственные числа разные.

²⁶Ортогональная матрица – та, для которой $U'U = E \Leftrightarrow U' = U^{-1}$. Её столбцы образуют ортонормированный набор векторов.

²⁷Хочу на всякий случай напомнить, что эта величина называется числом обусловленности.

Утверждение. Пусть A – диагонализируемая вещественная матрица, у которой все собственные числа различны. Они могут быть комплексными, но тогда они идут парами со своими сопряжёнными. Матрица A' будет иметь все те же собственные числа (так как $\det(A - \lambda E) = \det(A' - \lambda E)$). Пусть \mathbf{u} – какой-то собственный вектор матрицы A : $A\mathbf{u} = \lambda\mathbf{u}$, а \mathbf{v} – собственный вектор матрицы A' : $A'\mathbf{v} = \mu\mathbf{v}$. Тогда если $\lambda \neq \bar{\mu}$, то $(\mathbf{u}, \mathbf{v}) = 0$.

□

$$\begin{aligned} (A\mathbf{u}, \mathbf{v}) &= (\lambda\mathbf{u}, \mathbf{v}) = \lambda(\mathbf{u}, \mathbf{v}) \\ (\mathbf{u}, A'\mathbf{v}) &= (\mathbf{u}, \mu\mathbf{v}) = \bar{\mu}(\mathbf{u}, \mathbf{v}) \Rightarrow (\lambda - \bar{\mu})(\mathbf{u}, \mathbf{v}) = 0 \Rightarrow (\mathbf{u}, \mathbf{v}) = 0 \end{aligned}$$

▣.

Можем выписать собственные числа матрицы A как $\lambda_1, \lambda_2, \dots, \lambda_s$, а у A' выпишем как $\bar{\lambda}_1, \bar{\lambda}_2, \dots, \bar{\lambda}_s$. Мы их выписали в таком порядке, чтобы было $\lambda_i = \bar{\mu}_i$. Тогда при $i \neq j$ получим $(\mathbf{u}_i, \mathbf{v}_j) = 0$. Это свойство называется *биортогональность*.

Пусть $\mathbf{x} \in \mathbb{R}^s$. Разложим его по базису собственных векторов A :

$$\mathbf{x} = \sum_{i=1}^s \gamma_i \mathbf{u}_i \Rightarrow (\mathbf{x}, \mathbf{v}_k) = \sum_{i=1}^s (\gamma_i \mathbf{u}_i, \mathbf{v}_k) = \gamma_k (\mathbf{u}_k, \mathbf{v}_k) \Rightarrow \gamma_i = \frac{(\mathbf{x}, \mathbf{v}_i)}{(\mathbf{u}_i, \mathbf{v}_i)}. \quad (23)$$

Рассмотрим возмущённую задачу:

$$(A + \Delta A)(\mathbf{u} + \Delta \mathbf{u}) = (\lambda + \Delta \lambda)(\mathbf{u} + \Delta \mathbf{u}).$$

Ограничимся линейными возмущениями (то есть отбросим члены второго порядка):²⁸

$$dA\mathbf{u}_i + A d\mathbf{u}_i = d\lambda_i \mathbf{u}_i + \lambda_i d\mathbf{u}_i. \quad (24)$$

Умножим это скалярно на \mathbf{v}_i :

$$\begin{aligned} (dA\mathbf{u}_i, \mathbf{v}_i) + \underbrace{(A d\mathbf{u}_i, \mathbf{v}_i)}_{\cong (d\mathbf{u}_i, A'\mathbf{v}_i)} &= d\lambda_i (\mathbf{u}_i, \mathbf{v}_i) + \lambda_i \underbrace{(d\mathbf{u}_i, \mathbf{v}_i)}_{\cong (\mathbf{u}_i, d\mathbf{v}_i)} \\ &\cong (d\mathbf{u}_i, A'\mathbf{v}_i) = \bar{\mu}_i (d\mathbf{u}_i, \mathbf{v}_i) \end{aligned}$$

То есть

$$d\lambda_i = \frac{(dA\mathbf{u}_i, \mathbf{v}_i)}{(\mathbf{u}_i, \mathbf{v}_i)} \Rightarrow |d\lambda_i| \leq \frac{\|dA\| \|\mathbf{u}_i\| \|\mathbf{v}_i\|}{|(\mathbf{u}_i, \mathbf{v}_i)|} = p_i \|dA\| \quad \left(p_i = \frac{\|\mathbf{u}_i\| \|\mathbf{v}_i\|}{|(\mathbf{u}_i, \mathbf{v}_i)|} \right)$$

Из неравенства Коши-Буняковского ($|(\mathbf{u}_i, \mathbf{v}_i)| \leq \|\mathbf{u}_i\| \|\mathbf{v}_i\|$) следует, что $p_i \geq 1$. Если матрица A симметрична, то $\mathbf{u}_i = \mathbf{v}_i \Rightarrow p_i = 1 \Rightarrow |d\lambda_i| \leq \|dA\|$.

Теперь умножим выражение (24) скалярно на \mathbf{v}_k :

$$\begin{aligned} (dA\mathbf{u}_i, \mathbf{v}_k) + \underbrace{(A d\mathbf{u}_i, \mathbf{v}_k)}_{\cong 0} &= d\lambda_i \underbrace{(\mathbf{u}_i, \mathbf{v}_k)}_{\cong 0} + \lambda_i (d\mathbf{u}_i, \mathbf{v}_k) \\ &\cong (d\mathbf{u}_i, A'\mathbf{v}_k) = \bar{\mu}_k (d\mathbf{u}_i, \mathbf{v}_k) \end{aligned}$$

То есть

$$(d\mathbf{u}_i, \mathbf{v}_k) = \frac{(dA\mathbf{u}_i, \mathbf{v}_k)}{\lambda_i - \lambda_k}.$$

Далее раскладываем $d\mathbf{u}_i$ по базису собственных векторов A и используем выражения для коэффициентов из (23):

$$d\mathbf{u}_i = \sum_{k=1}^s \gamma_k \mathbf{u}_k = \sum_{k=1}^s \frac{(d\mathbf{u}_i, \mathbf{v}_k)}{(\mathbf{u}_k, \mathbf{v}_k)} \mathbf{u}_k = \sum_{k=1}^s \frac{(dA\mathbf{u}_i, \mathbf{v}_k)}{(\lambda_i - \lambda_k)(\mathbf{u}_k, \mathbf{v}_k)} \mathbf{u}_k$$

²⁸Из-за того, что мы ограничились первым порядком, приращения можно обозначить как дифференциалы.

При $k = i$ получаем проблему – ноль в знаменателе. Решим её следующим образом: поскольку собственные вектора определяются с точностью до нормировки, выберем такое \mathbf{v}_i , что $\mathbf{v}_i \perp d\mathbf{u}_i$, тогда $\gamma_i = 0$ и слагаемое $k = i$ исчезает из суммы. Оценим норму:

$$\|d\mathbf{u}_i\| \leq \sum_{k \neq i} \frac{\|dA\| \|\mathbf{u}_i\| \|\mathbf{v}_k\|}{|\lambda_i - \lambda_k| |(\mathbf{u}_k, \mathbf{v}_k)|} \|\mathbf{u}_k\|,$$

Из этого получаем оценку относительной погрешности:

$$\frac{\|d\mathbf{u}_i\|}{\|\mathbf{u}_i\|} \leq \sum_{k \neq i} \frac{p_k}{|\lambda_i - \lambda_k|} \|dA\|$$

Из этой оценки видим, что когда два собственных числа находятся близко, то их собственные вектора будут плохо определены.

Далее речь пойдёт об алгоритмах отыскания собственных чисел.

15. Степенной метод для отыскания старшего собственного числа.

Note: Сначала я приведу рассуждения об алгоритмах, а также описание алгоритма Крылова. Не уверен что это нужно для ответа этого билета, но раз было в конспекте, я решил записать сюда. На экзамене этот момент лучше проскипать и сразу перейти к степенному методу.

Для собственных чисел не существует точных методов их нахождения. Их можно найти используя характеристический многочлен:

$$p_s(\lambda) = (-1)^s \lambda^s + p_1 \lambda^{s-1} + \dots + p_{s-1} \lambda + p_s.$$

Коэффициенты многочлена можно найти точно, но такие методы ушли в прошлое.

Примером такого метода является метод А. Н. Крылова. Из теоремы Кэли получаем равенство

$$p_s(A) = (-1)^s A^s + p_1 A^{s-1} + \dots + p_{s-1} A + p_s E = \mathbb{O},$$

которое является системой из s^2 уравнений с s неизвестными (p_1, \dots, p_s) . Она сильно переопределена, но имеет решения. Можем умножить её на произвольный \mathbf{x} :

$$(-1)^s A^s \mathbf{x} + p_1 A^{s-1} \mathbf{x} + \dots + p_{s-1} A \mathbf{x} + p_s \mathbf{x} = \mathbf{0}.$$

Тогда получается система из s уравнений с s неизвестными. Для её составления нужно $O(s^3)$ операций. Для решения – тоже $O(s^3)$.

Но этот метод сильно подвержен ошибкам округления. Как правило, система плохо обусловлена. По этим причинам этот метод вышел из употребления.

А теперь перейдём к итерационным методам, а именно к степенному. Снова рассмотрим векторы вида \mathbf{x} , $A\mathbf{x}$, $A(A\mathbf{x}) = A^2\mathbf{x}$, ..., $A^n\mathbf{x}$. Пусть $\{\mathbf{u}_i\}$ – полная система собственных векторов матрицы A , $A\mathbf{u}_i = \lambda_i \mathbf{u}_i$, пронумерованная в порядке убывания модулей соответствующих собственных чисел: $|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_s|$. Можем разложить вектор \mathbf{x} по базису собственных векторов:

$$\mathbf{x} = \sum_{i=1}^s \gamma_i \mathbf{u}_i \Rightarrow A\mathbf{x} = \sum_{i=1}^s \gamma_i \lambda_i \mathbf{u}_i, \quad A^n \mathbf{x} = \sum_{i=1}^s \gamma_i \lambda_i^n \mathbf{u}_i.$$

В этой сумме каждое слагаемое преобладает над следующим, потому что мы так расположили λ_i . Можем считать, что вся сумма $\approx \gamma_1 \lambda_1^n \mathbf{u}_1$, то есть первое слагаемое преобладает над всеми остальными, а $A^n \mathbf{x}$ – приближенный собственный вектор матрицы A . В этом заключается сама идея.

Распишем основной алгоритм степенного метода: дана матрица A и произвольный вектор \mathbf{x}_0 . Далее будем строить итерации следующего вида:

$$\mathbf{x}_n - \text{получен из прошлой итерации}; \quad \tilde{\mathbf{x}}_{n+1} = A\mathbf{x}_n, \quad \mathbf{x}_{n+1} = \frac{\tilde{\mathbf{x}}_{n+1}}{\{\tilde{\mathbf{x}}_{n+1}\}_1}.$$

Мы нормировали на первую компоненту, а не на норму вектора. Если попробуем построить итерации с нормированием на норму, то не всегда можем получить сходимость. Пример: пусть сначала первая компонента \mathbf{x}_0 равнялась -1 , а после умножения на A стала равна 1 . После следующего умножения получим снова -1 и так далее. То есть получили осцилляции. Вообще, сходимость степенного метода надо доказать.

Теорема. Пусть A – диагонализируемая матрица,²⁹ а её собственные числа подчинены условию

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_s|.$$

Пусть первая компонента первого собственного вектора ненулевая $u_{11} \neq 0$. Пусть \mathbf{x}_0 – произвольный вектор, для которого $\gamma_1 \neq 0$. Тогда степенной метод сходится, $\{\tilde{\mathbf{x}}_{n+1}\}_1 \xrightarrow{n \rightarrow \infty} \lambda_1$.

□

$$\mathbf{x}_n = \frac{A^n \mathbf{x}_0}{\{A^n \mathbf{x}_0\}_1} = \frac{\gamma_1 \lambda_1^n \mathbf{u}_1 + \gamma_2 \lambda_2^n \mathbf{u}_2 + \dots}{\gamma_1 \lambda_1^n u_{11} + \gamma_2 \lambda_2^n u_{21} + \dots} = \frac{\mathbf{u}_1 + \frac{\gamma_2}{\gamma_1} \left(\frac{\lambda_2}{\lambda_1}\right)^n \mathbf{u}_2 + \dots}{u_{11} + \frac{\gamma_2}{\gamma_1} \left(\frac{\lambda_2}{\lambda_1}\right)^n u_{21} + \dots} \xrightarrow{n \rightarrow \infty} \frac{\mathbf{u}_1}{u_{11}}.$$

То есть мы показали, что \mathbf{x}_n сходится к первому собственному вектору матрицы A , отнормированного на свою первую компоненту. А значит при умножении матрицы A на него мы первой компоненте получим λ_1 ■.

Метод будет сходиться со скоростью геометрической прогрессии $O(|\lambda_2/\lambda_1|^n)$.

Плюсы: простота. Минусы: работает только для старшего собственного числа; сходимость только в случае $|\lambda_1| > |\lambda_2|$; полагаемся на случай $\gamma_1 \neq 0$ и $u_{11} \neq 0$.

Что будет если нарушить условия сходимости?

- $\lambda_1 = \lambda_2$ и $|\lambda_2| > |\lambda_3|$. Тогда $A^n \mathbf{x} = \lambda_1^n (\gamma_1 \mathbf{u}_1 + \gamma_2 \mathbf{u}_2) + \gamma_3 \lambda_3^n \mathbf{u}_3 + \dots$, то есть по сути всё то же самое. Полученный вектор будет из линейной оболочки \mathbf{u}_1 и \mathbf{u}_2 . Чтобы получить линейно-независимый вектор, берём другой \mathbf{x}_0 .
- $\lambda_1 = -\lambda_2 = \lambda$ и $|\lambda_2| > |\lambda_3|$. Тогда $A^n \mathbf{x} = \lambda_1^n (\gamma_1 \mathbf{u}_1 + (-1)^n \gamma_2 \mathbf{u}_2) + \gamma_3 \lambda_3^n \mathbf{u}_3 + \dots$. То есть мы можем выделить две разные подпоследовательности:

$$\mathbf{x}_{2n} = \frac{A^{2n} \mathbf{x}_0}{\{A^{2n} \mathbf{x}_0\}_1} \xrightarrow{n \rightarrow \infty} \frac{\gamma_1 \mathbf{u}_1 + \gamma_2 \mathbf{u}_2}{\gamma_1 u_{11} + \gamma_2 u_{21}}; \quad \mathbf{x}_{2n+1} = \frac{A^{2n+1} \mathbf{x}_0}{\{A^{2n+1} \mathbf{x}_0\}_1} \xrightarrow{n \rightarrow \infty} \frac{\gamma_1 \mathbf{u}_1 - \gamma_2 \mathbf{u}_2}{\gamma_1 u_{11} - \gamma_2 u_{21}};$$

которые будут сходиться к разным собственным векторам, но при этом к одному и тому же квадрату собственного числа:

$$\tilde{\mathbf{x}}_{2(n+1)} = A^2 \mathbf{x}_{2n} \xrightarrow{n \rightarrow \infty} \lambda^2 \frac{\gamma_1 \mathbf{u}_1 + \gamma_2 \mathbf{u}_2}{\gamma_1 u_{11} + \gamma_2 u_{21}}; \quad \tilde{\mathbf{x}}_{2(n+1)+1} = A^2 \mathbf{x}_{2n+1} \xrightarrow{n \rightarrow \infty} \lambda^2 \frac{\gamma_1 \mathbf{u}_1 - \gamma_2 \mathbf{u}_2}{\gamma_1 u_{11} - \gamma_2 u_{21}}.$$

- Пусть $\lambda_1 = re^{i\theta} \Rightarrow \lambda_2 = re^{-i\theta}$. Так как матрица вещественная, то $\gamma_2 = \overline{\gamma_1}$ и $\mathbf{u}_2 = \overline{\mathbf{u}_1}$. Тогда получим, что никакой сходимости нет:

$$\mathbf{x}_n \xrightarrow{n \rightarrow \infty} \frac{\operatorname{Re}(\gamma_1 e^{i\theta n} \mathbf{u}_1)}{\operatorname{Re}(\gamma_1 e^{i\theta n} u_{11})}; \quad \{\tilde{\mathbf{x}}_{n+1}\}_1 \xrightarrow{n \rightarrow \infty} \frac{r \operatorname{Re}(\gamma_1 e^{i\theta(n+1)} \mathbf{u}_1)}{\operatorname{Re}(\gamma_1 e^{i\theta n} u_{11})}$$

²⁹Это требование \Leftrightarrow собственные вектора образуют базис в \mathbb{R}^s .

16. Обратный степенной метод.

Утверждение. Пусть λ_i – собственное число, а \mathbf{u}_i – соответствующий собственный вектор неособой матрицы A . Тогда λ_i^{-1} – собственное число матрицы A^{-1} с тем же собственным вектором.

□

$$A\mathbf{u}_i = \lambda_i\mathbf{u}_i \Rightarrow \lambda_i^{-1}\mathbf{u}_i = A^{-1}\mathbf{u}_i$$

▣.

Отсюда следует, что для наименьшего по модулю собственного числа верно:

$$\min |\lambda_A| = (\max |\lambda_{A^{-1}}|)^{-1}$$

Таким образом, если матрица A неособая и её собственные числа выстроены так, что $|\lambda_s| < |\lambda_{s-1}| \leq \dots \leq |\lambda_1|$, то мы можем применить обычный степенной метод к A^{-1} , и тогда получим λ_s^{-1} . В этом заключается *обратный степенной метод*.

На практике вместо нахождения обратной матрицы (весьма трудозатратная операция) и умножения лучше решать СЛАУ относительно $\tilde{\mathbf{x}}_{n+1}$ в итерациях:

$$\mathbf{x}_n - \text{получен из прошлой итерации}; \quad \tilde{\mathbf{x}}_{n+1} = A^{-1}\mathbf{x}_n \Leftrightarrow \boxed{A\tilde{\mathbf{x}}_{n+1} = \mathbf{x}_n}, \quad \mathbf{x}_{n+1} = \frac{\tilde{\mathbf{x}}_{n+1}}{\{\tilde{\mathbf{x}}_{n+1}\}_1}.$$

А что если мы применим обратный степенной метод не к матрице A , а к матрице $(A - tE)$, где t – заранее заданное число (сдвиг)? Тогда мы должны получить наименьшее собственное число $(A - tE)$: $(A - tE)\mathbf{u}_k = (\lambda_k - t)\mathbf{u}_k$. После возведения в минус первую степень оно окажется наибольшим:

$$\frac{1}{|\lambda_k - t|} > \frac{1}{|\lambda_i - t|} \quad \forall i \neq k.$$

Таким образом, мы можем найти собственное число матрицы A , ближайшее к t . В этом заключается *обратный степенной метод с постоянным сдвигом*. Скорость сходимости:

$$O\left(\frac{|\lambda_k - t|}{\min_{i \neq k} |\lambda_i - t|}\right),$$

то есть если мы в качестве начального t возьмём число, близкое к собственному числу, то сходимость будет быстрой. Этот метод хорош для уточнения собственных чисел.

Также мы можем менять сдвиг на каждом шаге. Обозначим $\{\tilde{\mathbf{x}}_{n+1}\}_1 = \mu_{n+1}$. Итерации будут такие:

$$\mathbf{x}_n, t_n - \text{получены из прошл. итерации}; \quad (A - t_n E)\tilde{\mathbf{x}}_{n+1} = \mathbf{x}_n, \quad \mathbf{x}_{n+1} = \mu_{n+1}^{-1}\tilde{\mathbf{x}}_{n+1}, \quad \boxed{t_{n+1} = t_n + \mu_{n+1}^{-1}}$$

Строго доказывать сходимость мы не будем. Но краткое пояснение дадим: поскольку $\tilde{\mathbf{x}}_{n+1}$ – приближение собственного вектора, то

$$(A - t_n E)\tilde{\mathbf{x}}_{n+1} \approx (\lambda_A - t_n)\tilde{\mathbf{x}}_{n+1},$$

$$\{\mathbf{x}_n\}_1 = 1 \Rightarrow \mu_{n+1} = \{\tilde{\mathbf{x}}_{n+1}\}_1 = (\lambda_A - t_n)^{-1} \Rightarrow \lambda_A = t_n + \mu_{n+1}^{-1}$$

В этом методе t_n сходится к собственному числу. Если t_0 и \mathbf{x}_0 близки к собственному числу и соответствующему собственному вектору, то сходимость будет быстрой. А иначе – может быть не быстрой. Если взять $t_0 > \|A\|$, то в результате найдём \mathbf{u}_1 – какой-то собственный вектор. Далее, берём $\mathbf{x}_0 \perp \mathbf{u}_1$ и снова решаем и получаем \mathbf{u}_2 . Потом выбираем \mathbf{x}_0 , ортогональный \mathbf{u}_1 и \mathbf{u}_2 одновременно, и так далее. Так можем найти все собственные числа, но сходимость будет медленной. Минус данного метода в том, что не существует оптимального способа выбора начального приближения t_0, \mathbf{x}_0 . Однако метод хорош для уточнения собственных чисел.

17. Двумерные вращения, их виды.

Для начала вспомним несколько фактов из линейной алгебры. Если матрица A симметрична, то существуют такие U – ортогональные,³⁰ что

$$A = U^{-1} \Lambda U = U' \Lambda U, \quad \Lambda = \text{diag} \{ \lambda_i \}.$$

Произведение ортогональных матриц ортогонально:

$$(UV)' = V'U' = V^{-1}U^{-1} = (UV)^{-1}.$$

А ещё для ортогональных матриц выполняется

$$(U\mathbf{x}, U\mathbf{y}) = (U'U\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y}),$$

то есть ортогональные матрицы сохраняют скалярное произведение \Rightarrow они сохраняют длины и углы. Сохранение длин и углов – это намёк о том, что такие ортогональные операторы имеют смысл вращений (или композиций вращений и отражений). Есть теорема которая подтверждает эти догадки. Её формулировка: для любой ортогональной матрицы U в \mathbb{R}^s существует базис, в котором эта матрица представлена в блочно-диагональном виде из блоков 3 типов: тождественного отображения, отражения и вращения, соответственно:

$$\boxed{1}, \quad \boxed{-1}, \quad \boxed{\begin{matrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{matrix}}.$$

Раз ортогональная матрица имеет смысл вращения, то имеет смысл искать её в виде композиции *простых (двумерных)* вращений:

$$V_{p,q}(\varphi) = \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ \dots & \dots & \dots & c & \dots & \dots & \dots & -s & \dots & \dots \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ \dots & \dots & \dots & s & \dots & \dots & \dots & c & \dots & \dots \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix}, \quad \text{где} \quad \begin{matrix} c = \cos \varphi \\ s = \sin \varphi \end{matrix}$$

Одно простое вращение симметричной матрицы A – это переход $A \mapsto C = V'AV$, где $V = V_{p,q}(\varphi)$. Видно, что матрица C симметрична. Разберём первый шаг: $C = V'B$, $B = AV$:

$$B^{(k)} = A^{(k)} \quad (\forall k \neq p, q); \quad B^{(p)} = cA^p + sA^q; \quad B^{(q)} = -sA^p + cA^q, \quad \text{или}$$

$$b_{ik} = a_{ik} \quad (\forall i \neq p, q); \quad b_{ip} = ca_{ip} + sa_{iq}; \quad b_{iq} = -sa_{ip} + ca_{iq}.$$

Рассмотрим элементы матрицы C :

$$c_{ik} = a_{ik} \quad (\forall i \neq p, q); \quad c_{pk} = ca_{pk} + sa_{qk}; \quad c_{qk} = -sa_{pk} + ca_{qk} \quad - \text{это всё } \forall k \neq p, q;$$

$$c_{pp} = a_{pp}c^2 + 2a_{pq}cs + a_{qq}s^2; \quad c_{pq} = (a_{qq} - a_{pp})cs + a_{pq}(c^2 - s^2); \quad c_{qq} = a_{qq}c^2 - 2a_{pq}cs + a_{pp}s^2.$$

Мы получили явные выражения для компонент C .

³⁰Если вы в \mathbb{C}^s , то «симметричная» \leftrightarrow «эрмитова» или «самосопряжённая», а «ортогональная» \leftrightarrow «унитарная».

С помощью вращений можно попытаться обнулить ~~срок президе~~ некоторые из внедиагональных элементов. Мы знаем два способа:

1. Вращение Гивенса – зануляем $c_{p-1,q}$:

$$c_{p-1,q} = -sa_{p,p-1} + ca_{q,p-1} = 0 \Rightarrow c = \cos \varphi = \frac{a_{p,p-1}}{\sqrt{a_{p,p-1}^2 + a_{1,p-q}^2}}, \quad s = \sin \varphi = \frac{a_{q,p-1}}{\sqrt{a_{p,p-1}^2 + a_{1,p-q}^2}}.$$

2. Вращение Якоби – зануляем c_{pq} :

$$c_{pq} = (a_{qq} - a_{pp})cs + a_{pq}(c^2 - s^2) = 0 \Rightarrow \frac{s}{c} = \operatorname{tg} \varphi = \frac{2a_{pq}}{a_{qq}^2 - a_{pp}^2}.$$

18. Лемма о правиле знаков при исключении.

Note: этого билета преподаватель нам уже не дал. Точнее дал, но очень не очень. Поэтому я компилировал его из прошлогоднего и прошлодесятилетнего конспектов.

Для начала стоит привести важные факты и определения из Алгебры:

Определение. Пусть $B(\mathbf{x}, \mathbf{y})$ – симметрическая билинейная функция. Тогда функция одного аргумента $B(\mathbf{x}, \mathbf{x})$ называется квадратичной формой. При этом

$$\forall B(\mathbf{x}, \mathbf{y}) \exists A - \text{симметричная: } (A\mathbf{x}, \mathbf{x}) = B(\mathbf{x}, \mathbf{x}),$$

эта A называется *матрицей квадратичной формы*. Для неё справедлив *закон инерции*: при приведении A к диагональному виду количество элементов одного знака не зависит от способа приведения.

Напомним, в чём заключается смысл метода [исключения] Гаусса. Мы решаем СЛАУ. В матрице системы сначала делим всю первую строку на a_{11} (а если он равен нулю, то можем поменять первую строку местами с какой-нибудь другой). После этого вычитаем первую строку из всех остальных строк, предварительно умножив на первый элемент соответствующей строки. Таким образом получим, что первый столбец превратился в $(1, 0, \dots, 0)$. Далее эта процедура повторяется для остальных строк и матрица превращается в треугольную, после чего нетрудно решить систему. Компоненты $a_{11}, a_{22}, \dots, a_{ss}$ называются *ведущими элементами*. Они могут не совпадать с диагональными элементами матрицы, ведь мы могли менять строки местами.

Лемма. Пусть A – симметричная матрица. Тогда число ведущих элементов одного знака совпадает с числом собственных чисел A того же знака.

□ Покажем, что метод исключения Гаусса аналогичен приведению матрицы к квадратичной форме

$$(A\mathbf{x}, \mathbf{x}) = \sum_{i=1}^s \sum_{j=1}^s a_{ij}x_i x_j = a_{11}x_1^2 + 2 \sum_{i=2}^s a_{1i}x_1 x_i + \sum_{i=2}^s \sum_{j=2}^s a_{ij}x_i x_j =$$

Приведём её к сумме квадратов стандартным способом (Лежандра).

$$\begin{aligned} &= a_{11}^{-1} \left(a_{11}^2 x_1^2 + 2a_{11}x_1 \sum_{i=2}^s a_{1i}x_i + \sum_{i=2}^s \sum_{j=2}^s a_{1i}a_{1j}x_i x_j \right) + \sum_{i=2}^s \sum_{j=2}^s \overbrace{\left(a_{ij} - \frac{a_{1i}a_{1j}}{a_{11}} \right)}^{\tilde{a}_{ij}} x_i x_j = \\ &= a_{11}^{-1} \underbrace{\left(\sum_{i=1}^s a_{1i}x_i \right)^2}_{\xi_1^2} + \sum_{i=2}^s \sum_{j=2}^s \tilde{a}_{ij}x_i x_j. \end{aligned}$$

Следующий шаг (аналогичный) применяется уже ко второй сумме. Таким образом форма приведётся к сумме квадратов:

$$(A\mathbf{x}, \mathbf{x}) = a_{11}^{-1}\xi_1^2 + \tilde{a}_{22}^{-1}\xi_2^2 + \tilde{\tilde{a}}_{33}^{-1}\xi_3^2 + \dots$$

То есть у матрицы системы на диагонали будут стоять обратные ведущие элементы. Взглянем на то, что мы делали при получении \tilde{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.

Приведём A к ЖНФ. Раз A симметрична, то J_A диагональна. На её диагонали стоят собственные числа. Сравниваем выражения для $(A\mathbf{x}, \mathbf{x})$ и $(J_A\mathbf{x}, \mathbf{x})$. Число положительных и отрицательных «квадратов» одинаково и зависит только от A (по закону инерции) \blacksquare .

19. Метод Гивенса.

Если мы проведём вращения Гивенса над симметричной матрицей A в следующем порядке:

$$\begin{aligned} (2, 3) &\rightarrow (2, 4) \rightarrow (2, 5) \rightarrow \dots \rightarrow (2, s) \rightarrow \\ &\rightarrow (3, 4) \rightarrow (3, 5) \rightarrow \dots \rightarrow (3, s) \rightarrow \\ &\rightarrow (4, 5) \rightarrow \dots \rightarrow (4, s) \rightarrow \\ &\rightarrow \dots \dots \dots \rightarrow \\ &\rightarrow (s-1, s), \end{aligned}$$

то матрица превратится в трёхдиагональную. Применим исключения Гаусса к матрице $(A - tE)$:

$$\begin{pmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & & \\ & b_2 & a_3 & b_3 & \\ & & \ddots & \ddots & \ddots \\ & & & b_{s-1} & a_s \end{pmatrix} \rightarrow \begin{pmatrix} 1 & \frac{b_1}{a_1} & & & \\ & a_2 - \frac{b_1^2}{a_1} & b_2 & & \\ & b_2 & a_3 & b_3 & \\ & & \ddots & \ddots & \ddots \\ & & & b_{s-1} & a_s \end{pmatrix}$$

Видим, что на первом шаге ведущий элемент $\alpha_1 = a_1$. А для следующего шага ведущий элемент изменился: $\alpha_2 = a_2 - \frac{b_1^2}{a_1}$. Для последующих будет верно $\alpha_{k+1} = a_{k+1} - \frac{b_k^2}{\alpha_k}$. Всего таких операций $2s$. В результате мы найдём количество положительных (и отрицательных) ведущих элементов. А из леммы из прошлого билета следует, что тогда найдём и количество положительных (и отрицательных) собственных чисел. Собственные числа матрицы $(A - tE)$ – это $\lambda_A - t$. Посчитав количество $\lambda_A - t > 0$, мы найдём количество $\lambda_A > t$. Можем найти собственные числа матрицы A методом бисекции (он же метод половинного деления): изменяя t , будем определять, при каких значениях меняется количество собственных чисел $\lambda_A > t$. Так как $|\lambda_A| \leq \|A\|$, то t разумно менять в пределах $[-\|A\|, \|A\|]$.

20. Метод Якоби.

Будем приближаться к диагональной матрице с помощью большого числа вращений Якоби. Введём следующие величины:

$$N^2(A) = \sum_{i,k=1}^s a_{ik}^2 = \|A\|_F^2; \quad d^2(A) = \sum_{i=1}^s a_{ii}^2; \quad t^2(A) = \sum_{i \neq k} a_{ik}^2 = N^2(A) - d^2(A).$$

Где $\|A\|_F$ – *Фробениусова норма*. Она не является операторной нормой. Одно из её свойств:

$$\|B\|_F^2 = \text{tr } BB',$$

так как у нас матрица A симметрична, то

$$N^2(A) = \|A\|_F^2 = \text{tr } A^2$$

Величина $t^2(A)$ есть сумма квадратов недиагональных элементов. Её можно трактовать как величину, которая измеряет «недиагональность».

Утверждение. Пусть A – симметричная матрица, а $V = V_{p,q}(\varphi)$ – вращение Якоби. Тогда после одного такого вращения ($C = V'AV$) сумма квадратов недиагональных элементов уменьшается следующим образом:

$$t^2(C) = t^2(A) - 2a_{pq}^2$$

□ Сначала покажем, что $N^2(C) = N^2(A)$. Выше мы уже показали, что это след квадрата матрицы A . А след – это второй коэффициент в характеристическом многочлене. Характеристический многочлен можно представить как $p_s(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)\dots(\lambda_s - \lambda)$. Поскольку подобные матрицы имеют одинаковые собственные числа, то у подобных матриц будут и одинаковые характеристические многочлены \Rightarrow следы у подобных матриц одинаковы. При этом $C^2 = V'AVV'AV = V'A^2V$, то есть C^2 и A^2 подобны. Значит у них одинаковый след $\Rightarrow N^2(C) = N^2(A)$.

$$\begin{aligned} t^2(C) - t^2(A) &= N^2(C) - d^2(C) - N^2(A) + d^2(A) = d^2(A) - d^2(C) = a_{pp}^2 + a_{qq}^2 - c_{pp}^2 - c_{qq}^2 = \\ &= (a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2) - 2a_{pq}^2 - c_{pp}^2 - c_{qq}^2. \end{aligned}$$

При этом из равенства $N^2(A) = N^2(C)$ следует

$$a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2 = c_{pp}^2 + c_{qq}^2 + 2c_{pq}^2.$$

Подставим это в выражение выше и, в конце концов, получим

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

и это выражение верно \forall вращений. А для вращения Якоби $c_{pq} = 0$ ▮.

Нужен какой-нибудь способ выбора (p, q) на данном шаге для приближения к диагональности.

Классический метод Якоби: выбираем (p, q) , соответствующие максимальному по модулю недиагональному элементу матрицы:

$$|a_{pq}| = \max_{i \neq j} |a_{ij}|.$$

В пределе (после бесконечного числа шагов) получим $t^2(C) \rightarrow 0$. Можем оценить скорость его сходимости:

$$\forall i, j \quad |a_{ij}| \leq |a_{pq}| \Rightarrow \frac{1}{2}t^2(A) = \sum_{i < j} a_{ij}^2 \leq \frac{s(s-1)}{2} a_{pq}^2 \Rightarrow a_{pq}^2 \geq \frac{1}{s(s-1)} t^2(A).$$

Таким образом,

$$t^2(C) \leq t^2(A) \underbrace{\left(1 - \frac{2}{s(s-1)}\right)}_{\theta \leq 1}.$$

То есть метод сходится со скоростью геометрической прогрессии со знаменателем θ (то есть $\sim O(s^2)$). Но это оценка грубая, на самом деле метод Якоби сходится быстрее. В примере классического метода Якоби поиск максимума – более ресурсозатратная операция ($\sim O(s^2)$), чем сам метод Якоби ($\sim O(s)$). Надо подумать, как можно это улучшить.

Циклический метод Якоби: проходим все наддиагональные элементы (их $s(s-1)/2$ штук). Но при этом мы будем проходить и очень малые элементы, что не имеет смысла.

Циклический метод Якоби с барьером: выбираем некий ε_1 . Обнуляем только те элементы, которые больше него. Если после цикла все элементы оказались меньше него, то выбираем $\varepsilon_2 < \varepsilon_1$, и повторяем процедуру, и так раз за разом.

Теперь разберёмся, как именно получить собственные числа и собственные векторы. Собственные числа будут лежать на диагонали получаемой матрицы. Корректность следует из теоремы Бауэра-Файка.

В качестве собственных векторов можно брать столбцы матрицы-произведения всех вращений $V_1 V_2 \dots V_k \equiv \mathcal{U}_k$. Так как $\Lambda = \mathcal{U}_k' A \mathcal{U}_k \Rightarrow \mathcal{U}_k \Lambda = A \mathcal{U}_k \Rightarrow \mathcal{U}_k \Lambda^{(i)} = A \mathcal{U}_k^{(i)} \Rightarrow \lambda_i \mathcal{U}_k^{(i)} = A \mathcal{U}_k^{(i)}$.

21. Две леммы о факторизации матрицы.

Note: На этом моменте начался карантин и занятий больше не было. Поэтому все последующие билеты, начиная с этого, написаны исключительно с опорой на конспекты прошлых лет, а также на материалы-исправления к прошлогоднему конспекту, присланные самим преподавателем.

Лемма (1). Пусть A – невырожденная матрица с ненулевыми диагональными минорами. Тогда

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

То есть её можно единственным образом разложить на произведение верхней треугольной и нижней треугольной матрицы с единицами в диагонали.

□ Эта теорема – матричная запись метода Гаусса. Матрицу R можем строить пошагово. Пусть мы проделали первый шаг метода Гаусса: поделили первый столбец на свой элемент, и вычли из каждого другого столбца первый, умноженный на первый элемент соответствующего столбца:

$$\tilde{l}_{ij} = a_{ij} - \frac{a_{1j}a_{i1}}{a_{11}}.$$

В результате получили матрицу, которую обозначим за \tilde{L} . Тогда можем записать

$$A = \tilde{L}\tilde{R} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ * & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & * \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1s} \\ 0 & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \dots & * \end{pmatrix}.$$


Вид матрицы \tilde{L} не вызывает вопросов: нули в первой строчке возникли из-за применения исключения Гаусса. А почему первый столбец и первая строчка матрицы \tilde{R} выглядят так? Если мы рассмотрим это равенство построчно, то получим, что первая строка матрицы \tilde{R} должна совпадать с первой строкой матрицы A . Далее, если мы рассмотрим равенство постолбцово, то можно понять, что первый столбец матрицы \tilde{R} должен состоять из нулей (кроме первого элемента).

Дальнейшее применение исключений Гаусса не затронет первую строчку и первый столбец матриц \tilde{L} и \tilde{R} . Далее можем аналогичным образом применить второй шаг для нижних правых блоков матрицы (без первых строки и столбца). Затем третий шаг для нижнего правого блока без первых двух строк и столбцов, и так далее. В конце концов получим

$$A = LR = \begin{pmatrix} 1 & & & \\ * & 1 & & \\ \vdots & \vdots & \ddots & \\ * & * & \dots & 1 \end{pmatrix} \begin{pmatrix} * & \dots & \dots & * \\ & * & \dots & * \\ & & \ddots & \vdots \\ & & & * \end{pmatrix}.$$

Можно расписать это равенство построчно:

$$\begin{aligned} A_1 &= R_1, \\ A_2 &= l_{21}R_1 + R_2 \Rightarrow R_2 = A_2 - l_{21}A_1, \\ &\dots \\ A_i &= l_{i1}R_1 + l_{i2}R_2 + \dots + l_{i,i-1}R_{i-1} + R_i, \end{aligned}$$

A_i известны как строки матрицы, а l_{ij} известны, так как L – результат применения метода Гаусса к A . Из этой цепочки равенств легко находятся строки матрицы последовательно R . Если A была трёхдиагональной, то L и R получатся двухдиагональными .

Лемма (2). Пусть A – неособая матрица, тогда

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

где R – верхняя треугольная матрица, а Q – ортогональная. Это разложение уже не единственно.

□ В доказательстве прошлой Леммы мы применяли исключения Гаусса к матрице A . А идея этого доказательства в том, что мы применяем ортогонализацию Грамма-Шмидта к строкам матрицы A . Будем расписывать равенство $A = QR$ по столбцово.

$$A^{(1)} = r_{11}Q^{(1)}.$$

Отсюда определяем r_{11} и $Q^{(1)}$.³¹ Если мы хотим ортонормированности Q , то должны потребовать $r_{11} = \pm \|A^{(1)}\|$. (Уже получаем два разных варианта). Следующая строка:

$$\begin{aligned} A^{(2)} &= r_{12}Q^{(1)} + r_{22}Q^{(2)} \Rightarrow \\ (A^{(2)}, Q^{(1)}) &= r_{12} \cdot 1 + r_{22} \cdot 0 = r_{12}, \end{aligned}$$

то есть r_{12} определено. Следовательно,

$$r_{22}Q^{(2)} = A^{(2)} - (A^{(2)}, Q^{(1)})Q^{(1)}.$$

Здесь как раз видна ортогонализация Грамма-Шмидта. Если хотим ортонормированности, то берём r_{22} , равный \pm норме правой части. Продолжаем этот процесс и дальше.

$$A^{(i)} = r_{1i}Q^{(1)} + r_{2i}Q^{(2)} + \dots + r_{ii}Q^{(i)}.$$

Скалярно умножаем это выражение на $Q^{(j)}$, где $j < i$:

$$(A^{(i)}, Q^{(j)}) = r_{ji} \quad \forall j < i.$$

Можем выразить $r_{ii}Q^{(i)}$:

$$r_{ii}Q^{(i)} = A^{(i)} - (A^{(i)}, Q^{(1)})Q^{(1)} - \dots - (A^{(i)}, Q^{(i-1)})Q^{(i-1)}.$$

Отсюда определяем r_{ii} и $Q^{(i)}$. Если хотим ортонормированности, то полагаем r_{ii} равным \pm норме правой части ■.

22. Теорема о сходимости итерированных подпространств.

Степенной метод обладал одним важным недостатком: он искал только одно собственное число (наибольшее) и только один собственный вектор. Можем обобщить его. Будем итерировать не один вектор, а целый базис $\{\mathbf{x}_i\}$. Рассмотрим итерационную формулу:

$$\mathbf{x} = \sum_{i=1}^s \gamma_i \mathbf{u}_i, \quad A^n \mathbf{x} = \sum_{i=1}^s \gamma_i \lambda_i^n \mathbf{u}_i$$

Пусть собственные числа выстроены в порядке убывания $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_s|$. Если при этом $|\lambda_1| > |\lambda_2|$, то первое слагаемое в сумме будет доминировать, и в результате мы получим \mathbf{x} , коллинеарный \mathbf{u}_1 . То есть получаемый собственный вектор будет лежать в одномерном пространстве $\langle \mathbf{u}_1 \rangle$.

Если же $|\lambda_1| = |\lambda_2| > |\lambda_3|$, то первые два слагаемых доминируют, и в результате \mathbf{x} будет лежать в линейной оболочке $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle$.

³¹Как я понял, r_{11} – произвольный множитель, а $Q^{(1)}$ – произвольный вектор, коллинеарный $A^{(1)}$.

Определение. Пусть $|\lambda_1| > |\lambda_2| > \dots > |\lambda_s| > 0 \Rightarrow A$ – обратима и её собственные вектора образуют базис. Тогда $U_k = \langle \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k \rangle$ – k -мерное старшее собственное подпространство. k -мерное приближение заключается в том, что при больших n итерации лежат в U_k .

Определение. Пусть $\mathbf{x}_1, \dots, \mathbf{x}_s$ – базис. Тогда $\{A^n \mathbf{x}_i\}_{i=1}^s$ – тоже образует базис. Он называется n -ый итерированный базис. Он будет весь целиком сходиться к \mathbf{u}_1 . А мы хотим получать в результате не один вектор, а базис.

Определение. $L_k^{(n)} = \langle A^n \mathbf{x}_1, \dots, A^n \mathbf{x}_k \rangle$ – k -мерное n -ное итерированное подпространство.

Определение. Говорят, что $P^{(n)} \rightarrow P$, если в $P^{(n)}$ существует базис, который сходится к базису P .

Теорема. Пусть A – диагонализуемая неособая матрица,

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_s| > 0.$$

Тогда $L_k^{(n)} \rightarrow U_k$.

□ Пусть $\{\mathbf{x}_i\}$ – исходный базис. Разложим каждый его элемент по $\{\mathbf{u}_j\}$ и взглянем на итерированный базис:

$$A^n \mathbf{x}_i = \sum_{l=1}^s c_{il} \lambda_l^n \mathbf{u}_l.$$

Хотим построить другой базис. Первый элемент трогать не будем:

$$\mathbf{y}_1^{(n)} = A^n \mathbf{x}_1 = \lambda_1^n c_{11} \mathbf{u}_1 + \lambda_2^n c_{12} \mathbf{u}_2 + \dots$$

Из второго элемента исключим первый член:

$$\mathbf{y}_2^{(n)} = A^n \mathbf{x}_2 - \gamma \mathbf{y}_1^{(n)} = \lambda_2^n c'_{22} \mathbf{u}_2 + \lambda_3^n c'_{23} \mathbf{u}_3 + \dots, \quad \gamma = c_{21}/c_{11}$$

Для третьего элемента исключаем первый и второй член:

$$\mathbf{y}_3^{(n)} = A^n \mathbf{x}_3 - \gamma' \mathbf{y}_1^{(n)} - \gamma'' \mathbf{y}_2^{(n)} = \lambda_3^n c''_{33} \mathbf{u}_3 + \dots$$

И так далее до $\mathbf{y}_k^{(n)}$. Можно заметить, что $\mathbf{y}_1^{(n)} \in L_1^{(n)}$, $\mathbf{y}_2^{(n)} \in L_2^{(n)}$, ..., $\mathbf{y}_k^{(n)} \in L_k^{(n)}$. Мы построили некий k -мерный базис. Но сходимости ещё нет. Рассмотрим такие векторы:

$$\lambda_1^{-n} \mathbf{y}_1^{(n)} = c_{11} \mathbf{u}_1 + \left(\frac{\lambda_2}{\lambda_1} \right)^n c_{12} \mathbf{u}_2 + \dots,$$

$$\lambda_2^{-n} \mathbf{y}_2^{(n)} = c'_{22} \mathbf{u}_2 + \left(\frac{\lambda_3}{\lambda_2} \right)^n c'_{23} \mathbf{u}_3 + \dots,$$

и так далее. Эти векторы сходятся:

$$\lambda_1^{-n} \mathbf{y}_1^{(n)} \rightarrow c_{11} \mathbf{u}_1, \quad \lambda_2^{-n} \mathbf{y}_2^{(n)} \rightarrow c'_{22} \mathbf{u}_2, \quad \dots$$

Здесь нужно, чтобы некоторые из коэффициентов не были равны нулю. Как я понимаю, логика такая: Если вдруг, например, c_{11} оказался равным нулю, то мы должны за \mathbf{x}_1 взять какой-нибудь другой из векторов \mathbf{x}_i . При этом первый коэффициент в разложениях \mathbf{x}_i по \mathbf{u}_i не может быть нулевым для всех подряд \mathbf{x}_i , потому что тогда $\{\mathbf{x}_i\}$ не будет базисом.

Таким образом, мы нашли нужный базис $L_k^{(n)}$, который сходится к базису U_k . Значит $L_k^{(n)} \rightarrow U_k$ ■.

Какова же скорость сходимости? У нас есть k разных скоростей $(\lambda_2/\lambda_1)^n$, $(\lambda_3/\lambda_2)^n$, Можем выбрать из них наименьшую. Но это не самая точная оценка. Поступим следующим образом: Снова будем

исключать некоторые члены, на этот раз из $\mathbf{y}_i^{(n)}$. Рассмотрим, для примера, подпространство с $k = 3$. 3-ий (k -ый) вектор не меняем:

$$\mathbf{z}_3^{(n)} = \mathbf{y}_3^{(n)} = \lambda_3^n c_{33}'' \mathbf{u}_3 + \lambda_4^n c_{34}'' \mathbf{u}_4 + \dots$$

Из второго вектора исключим член с λ_3 :

$$\mathbf{z}_2^{(n)} = \mathbf{y}_2^{(n)} - \rho \mathbf{y}_3^{(n)} = \lambda_2 c_{22}' \mathbf{u}_2 + \lambda_4^n \tilde{c}_{24}' \mathbf{u}_4 + \dots$$

А из первого исключим члены с λ_2 и λ_3 :

$$\mathbf{z}_1^{(n)} = \mathbf{y}_1^{(n)} - \sigma \mathbf{y}_2^{(n)} - \tau \mathbf{y}_3^{(n)} = \lambda_1^n c_{11} \mathbf{u}_1 + \lambda_4^n \tilde{c}_{14} \mathbf{u}_4 + \dots$$

Все эти векторы принадлежат $L_3^{(n)}$, что важно. Теперь поделим их на соответствующие λ_i^n :

$$\lambda_1^{-n} \mathbf{z}_1^{(n)} = c_{11} \mathbf{u}_1 + \left(\frac{\lambda_4}{\lambda_1} \right)^n \tilde{c}_{14} \mathbf{u}_4 + \dots$$

$$\lambda_2^{-n} \mathbf{z}_2^{(n)} = c_{22}' \mathbf{u}_2 + \left(\frac{\lambda_4}{\lambda_2} \right)^n \tilde{c}_{24}' \mathbf{u}_4 + \dots$$

$$\lambda_3^{-n} \mathbf{z}_3^{(n)} = c_{33}'' \mathbf{u}_3 + \left(\frac{\lambda_4}{\lambda_3} \right)^n \tilde{c}_{34}'' \mathbf{u}_4 + \dots$$

Итого имеем три скорости: $(\lambda_1/\lambda_4)^n$, $(\lambda_2/\lambda_4)^n$ и $(\lambda_3/\lambda_4)^n$. Из них наименьшей является $(\lambda_4/\lambda_3)^n$.

Абсолютно аналогичные действия можно провести и с системой с любым k . Таким образом, мы доказали, что $L_k^{(n)} \rightarrow U_k$ со скоростью $O(|\lambda_{k+1}/\lambda_k|)^n$.

Везде в наших рассуждениях мы использовали разложение по собственным векторам. Но на практике они нам неизвестны. Нужно построить какой-нибудь такой базис, чтобы его можно было вычислить и чтобы он сходиллся как нам нужно.

Определение. *Ступенчатый базис* –

$$\begin{aligned} \mathbf{c}_1 &= (1, *, *, \dots *), \\ \mathbf{c}_2 &= (0, 1, *, \dots *), \\ \mathbf{c}_i &= (0, \dots, 0, 1, *, \dots *), \\ &\quad \underbrace{\hspace{1.5cm}}_{i-1} \\ \mathbf{c}_s &= (0, \dots, 0, 1). \\ &\quad \underbrace{\hspace{1.5cm}}_{s-1} \end{aligned}$$

В $L_k^{(n)}$ мы можем получить ступенчатый базис $\{\mathbf{c}_i^{(n)}\}$ из базиса $\{\mathbf{y}_i^{(n)}\}$, если сначала исключим первые $i - 1$ компоненты из каждого вектора \mathbf{y}_i , а потом поделим его i -тую компоненту. Ступенчатый базис не вырождается.

Теорема. *Ступенчатый базис $L_k^{(n)}$ сходится к базису U_k , а каждый его отрезок $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_i\}$ сходится к базису U_i . При этом должны выполняться определённые условия невырожденности.*

□ Рассмотрим ступенчатый базис $\{\mathbf{c}_i^{(n)}\}_{i=1}^k$ подпространства $L_k^{(n)}$. Можем разложить каждый его вектор по базису $\{\tilde{\mathbf{z}}_i^{(n)}\} = \{\lambda_i^{-n} \mathbf{z}_i^{(n)}\}$ (который описан выше). Рассмотрим разложение:

$$\mathbf{c}_i^{(n)} = \gamma_{i1} \tilde{\mathbf{z}}_1^{(n)} + \dots + \gamma_{ik} \tilde{\mathbf{z}}_k^{(n)} = \gamma_{ii} \tilde{\mathbf{z}}_i^{(n)} + \dots + \gamma_{ik} \tilde{\mathbf{z}}_k^{(n)}.$$

Рассматривая равенство покомпонентно, можно определить γ_{ij} . Если честно расписывать, то в конце концов можно получить какое-то такое матричное равенство:

$$(\mathbf{c}_1, \dots, \mathbf{c}_k) = (\tilde{\mathbf{z}}_1^{(n)}, \dots, \tilde{\mathbf{z}}_k^{(n)}) \{\gamma_{ji}\}_{i,j=1}^k,$$

где первая матрица в качестве столбцов имеет вектора \mathbf{c}_i , вторая – вектора $\tilde{\mathbf{z}}_i^{(n)}$, а третья матрица состоит из компонент γ_{ij} . Первые две матрицы прямоугольные! Эту систему лучше рассматривать постолбцово, потому что так она превращается в k систем линейных алгебраических уравнений. Всего получаем ks равенств, в то время как неизвестных $\gamma_{ij} - s^2$ штук. То есть в каждой СЛАУ по s равенств, и при этом по k неизвестных. Может показаться, что каждая такая СЛАУ переопределена, но на самом деле это не так! Ведь матрица из векторов $\tilde{\mathbf{z}}_i$ имеет ранг k . Поскольку столбцовый и строчный ранг матрицы равны, то мы можем выбрать k линейно-независимых строк и составить СЛАУ с k уравнениями и k неизвестными. В этом и заключается то самое условие невырожденности.

Таким образом, мы показали, что можем найти γ_{ij} . Поскольку вектора $\tilde{\mathbf{z}}_i$ сходятся (со скоростью $O(|\lambda_{k+1}/\lambda_k|^n)$), то и матрица системы тоже сходится и с той же скоростью. Матрица, получаемая из предельных векторов $\tilde{\mathbf{z}}^{(n)}$, называется предельной матрицей. Эта матрица также невырожденная, поскольку её столбцы – базисные вектора. Значит и $\mathbf{c}_i^{(n)}$ тоже будут иметь предел. Скорость сходимости снова получится $O(|\lambda_{k+1}/\lambda_k|^n)$ ■.

23. Треугольно-степенной метод. Сходимость.

Как обычно, ищем собственные числа матрицы A . Пусть P_0 – произвольная невырожденная матрица. Строим итерации такого вида:

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

P_{n+1} – нижнетреугольная (с единицами на главной диагонали), а R_{n+1} – верхнетреугольная матрица.

Теорема. Пусть A – невырожденная диагонализируемая матрица

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_s| > 0$$

с ненулевыми диагональными минорами. Тогда треугольно-степенной метод сходится. То есть для P_n и $R_n \exists$ пределы P и R , при этом на диагонали R стоят собственные числа матрицы A .

□ Сначала докажем по индукции, что первые k столбцов P_{n+1} – это ступенчатый базис $L_k^{(n+1)}$.

База: матрица P_0 невырождена, значит её столбцы линейно-независимы и вообще образуют базис. Будем считать этот базис исходным. Тогда первые k столбцов – это базис $L_k^{(0)}$

Переход: Считаем, что первые k столбцов матрицы P_n – это базис $L_k^{(n)}$. После умножения каждого из этих k базисных векторов на матрицу A мы получим базис $L_k^{(n+1)}$. Далее производится LR-факторизация матрицы AP_n . По первой лемме о факторизации мы знаем, что это происходит единственным образом. При этом первые k столбцов матрицы P_{n+1} – это также базис $L_k^{(n+1)}$ (причём ступенчатый), поскольку эти вектора получаются путём исключений Гаусса, применённых к первым k столбцам матрицы AP_n .

Таким образом, мы доказали, что первые k столбцов матрицы P_{n+1} образуют ступенчатый базис $L_k^{(n+1)}$. А по теореме о сходимости ступенчатого базиса он сходится к базису U_k со скоростью $O(|\lambda_{k+1}/\lambda_k|^n)$. То есть мы доказали существование пределов $P = \lim_{n \rightarrow \infty} P_n$ и $R = \lim_{n \rightarrow \infty} R_n$.

Далее, мы знаем, что

$$R_n = P_n^{-1}AP_{n-1} \xrightarrow{n \rightarrow \infty} R = P^{-1}AP,$$

то есть мы получили, что матрицы R и A подобны. Значит у них одинаковые собственные числа. Так как матрица R треугольная, то она содержит свои собственные числа на диагонали ■.

24. Ортогонально-степенной метод.

И снова ищем собственные числа матрицы A . Пусть C_0 – произвольная невырожденная матрица. Итерации:

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

где C_{n+1} – ортогональная, а R_{n+1} – верхнетреугольная матрица. Для ортогонально-степенного метода нельзя ничего сказать про сходимость. Если вспомним вторую лемму о факторизации, то можем понять, что у нас существует неопределённость задания с точностью до знака. Но зато есть сходимость по форме.

Определение. Пусть матрица B квазитреугольная.³² Тогда говорят, что A_k *сходится по форме* к B , если все элементы A_k под квазидиагональю (той же формы) сходятся к 0. При этом нам не важно, к чему конкретно сходятся элементы на квазидиагонали и над ней.

Теорема. Пусть A – невырожденная диагонализируемая матрица

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_s| > 0$$

с ненулевыми диагональными минорами. Тогда $C'_n A C_n$ сходится по форме к верхнетреугольной матрице. При этом на диагонали этой матрицы расположены собственные числа.

□ Рассмотрим поддиагональные элементы $C'_n A C_n$:

$$\forall i > j \quad \{C'_n A C_n\}_{ij} = (A C_n^{(j)}, C_n^{(i)}).$$

Из соображений, аналогичных предыдущему билету, можно показать, что первые k столбцов матрицы C_n будут являться базисом (причём ортогональным) пространства $L_k^{(n)}$. Тогда столбцы C_n можно разложить по базису $\tilde{\mathbf{z}}_i^{(n)}$. В таком разложении будут присутствовать члены порядка $O(|\lambda_{k+1}/\lambda_k|^n)$, которые окажутся маленькими при больших n . Рассмотрим эти столбцы после умножения матриц $A C_n$: на этот раз получим (после деления на соответствующие λ_i) разложение по $\tilde{\mathbf{z}}_i^{(n+1)}$ с остаточными членами порядка $O(|\lambda_{k+1}/\lambda_k|^{n+1})$. То есть получаем, что при больших n пространства $L_k^{(n)}$ и $L_k^{(n+1)}$ близки. Из ортогональности столбцов C_n можем получить:

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

то есть при больших n вектора $A C_n^{(j)}$ и $C_n^{(i)}$ становятся примерно ортогональны ▣.

25. LR-алгоритм. Практическая реализация.

Снова рассматриваем A со стандартным условием $|\lambda_1| > |\lambda_2| > \dots > |\lambda_s| > 0$. На первом шаге производим LR-факторизацию матрицы A :

$$A = L_1 R_1,$$

а остальные итерационные шаги выглядят так:

$$R_n L_n = L_{n+1} R_{n+1},$$

где L_i – нижнетреугольная матрица с единицами на диагонали, а R_i – верхнетреугольная.

У этого алгоритма есть связь с треугольно-степенным: нужно подставить туда $P_0 = E$, $P_n = L_1 \dots L_n$. В результате получим:

$$\begin{aligned} A P_0 &= P_1 R_1 \Leftrightarrow A E = L_1 R_1, \\ A P_n &= P_{n+1} R_{n+1} \Leftrightarrow A L_1 \dots L_n = L_1 \dots L_{n+1} R_{n+1} \Leftrightarrow L_1^{-1} A L_1 \dots L_n = L_2 \dots L_{n+1} R_{n+1} \Leftrightarrow \\ &\Leftrightarrow \cancel{L_1} R_1 L_2 \dots L_n = \cancel{L_1} L_3 \dots L_{n+1} R_{n+1} \Leftrightarrow \cancel{L_2} R_2 L_3 \dots L_n = \cancel{L_2} L_4 \dots L_{n+1} R_{n+1} \Leftrightarrow \dots \Leftrightarrow \\ &\Leftrightarrow \cancel{L_n} R_n L_n = \cancel{L_n} L_{n+1} R_{n+1}. \end{aligned}$$

Поскольку мы свели метод к одному из предыдущих, то доказывать сходимость не нужно. Диагональные элементы R_n стремятся к собственным числам со скоростями, как и прежде, $O(|\lambda_{k+1}/\lambda_k|^n)$.

³² «Квази» \leftrightarrow «блочно».

Но на самом деле LR-факторизация происходит за время $O(s^3)$, что очень долго и невыгодно. Но если A – трёхдиагональная, то она факторизуется на двухдиагональные L и R , и этот процесс занимает время $O(s)$. Если матрица A симметричная, то её можно привести к трёхдиагональному виду с помощью метода Гивенса. На самом деле существуют методы, позволяющие приводить к трёхдиагональному виду и несимметричные матрицы.

Для ускорения сходимости можно применить сдвиг по Рэлею:

$$t_n = r_{ss}^{(n)}, \quad R_n L_n - t_n E = L_{n+1} R_{n+1}.$$

Поскольку последнее собственное число сходилось со скоростью $O(|\lambda_s/\lambda_{s-1}|^n)$, то теперь мы сильно уменьшили скорость за счёт сдвигов. В результате λ_s будет равняться сумме полученных сдвигов. А что с остальными собственными числами?

Так как $P_n = L_1 \dots L_n$, то $L_n = P_{n-1}^{-1} P_n \xrightarrow{n \rightarrow \infty} E$. Но если $r_{ss}^{(n)} \approx 0$, то вся последняя строка ≈ 0 . То есть остальные собственные числа находятся в подматрице без последней строки и столбца. Аналогичным образом можем применить сдвиг к предпоследнему собственному числу, и так далее.

26. QR-алгоритм. Практическая реализация.

Снова рассматриваем A со стандартным условием $|\lambda_1| > |\lambda_2| > \dots > |\lambda_s| > 0$. На первом шаге производим QR-факторизацию матрицы A :

$$A = Q_1 R_1,$$

а остальные итерационные шаги выглядят так:

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

где Q_i – ортогональная, а R_i – верхнетреугольная матрица.

С помощью замен $P_0 = E$, $P_n = Q_1 \dots Q_n$ этот алгоритм сводится к ортогонально-степенному, что можно показать выкладками, аналогичными таковым из прошлого билета.

QR-факторизация происходит за время $O(s^3)$, что, опять же, весьма долго. Во время QR-факторизации сохраняется трёхдиагональность A , но только если A симметрична. С помощью вращений Гивенса можем привести несимметричную матрицу к форме Хессенберга (похожа на нижнетреугольную, но с одной дополнительной диагональю над главной). Тогда число операций будет уже $O(s^2)$.

Для ускорения можем снова применить сдвиг

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

где в качестве t_n можем взять последний элемент $\{R_n Q_n\}_{ss}$ (сдвиг по Рэлею). Но эта тактика хороша только для вещественных матриц. А в противном случае лучше брать собственные числа подматрицы размера 2×2 в правом нижнем углу $\{R_n Q_n\}_{s-1, s-1}^{s, s}$. Берём в качестве сдвига эти собственные числа по очереди. При этом обходимся без комплексных вычислений. Это сдвиг по Уилкинсону.

III. Интегральные уравнения.

Note. Для понимания этого и последующих разделов сначала лучше ознакомиться с введением в функциональный анализ, приведённым в приложении. Благодарности прошлому курсу!

27. Интегральное уравнение 2-го рода. Метод замены ядра на вырожденное.

Определение. Интегральным уравнением Фредгольма II рода называется уравнение вида

$$\varphi(x) = f(x) + \mu \int_a^b K(x, t) \varphi(t) dt. \quad (25)$$

Непрерывная функция K – его ядро.

Через K также обозначим оператор

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

Он компактен как интегральный оператор. Уравнение теперь примет вид

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

Оператор $T = I - \mu K$ Фредгольмов по определению.

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

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

□ Распишем $(K\varphi, \psi)$ по определению, проделав замену $t \leftrightarrow x$ между выкладками:

$$(K\varphi, \psi) = \int_a^b \int_a^b K(x, t) \varphi(t) dt \psi(x) dx = \int_a^b \varphi(x) \int_a^b K(t, x) \psi(t) dt dx = (\varphi, K^* \psi).$$

Таким образом получили, что K^* имеет нужный вид \blacksquare .

У ядра меняются местами аргументы – точно так же, как транспонирование даёт матрицу сопряжённого оператора в вещественном конечномерном случае!

Определение. Оператор называется *симметричным*, если для него выполняется

$$(K\varphi, \psi) = (\varphi, K\psi).$$

Сформулируем альтернативу Фредгольма для такого интегрального уравнения при данном μ :

1. Либо однородное уравнение имеет только нулевое решение – и тогда неоднородное разрешимо при любой правой части;
2. Либо однородное уравнение имеет нетривиальное решение – и тогда неоднородное уравнение имеет решение лишь при условии, что его правая часть удовлетворяет линейным соотношениям. Такие μ называют характеристическими числами.

Пусть φ_0 – нетривиальное решение однородного уравнения. Тогда

$$(I - \mu K)\varphi_0 = 0 \Leftrightarrow \varphi_0 - \mu K\varphi_0 = 0 \Leftrightarrow K\varphi_0 = \mu^{-1}\varphi_0,$$

то есть μ^{-1} – собственное число оператора K , а φ_0 – его собственная функция. Собственные функции различных собственных чисел ортогональны. Собственные функции одного собственного числа можно ортогонализировать.

Итого можем получить ортонормальную последовательность. Если выполнено³³

$$\int_a^b K(x, t)g(t)dt = 0 \Leftrightarrow g(t) \equiv 0,$$

то эта последовательность называется *Фундаментальной системой*. Не стоит путать фундаментальную систему с базисом:

Определение. *Базис* в линейном нормированном пространстве – это такая бесконечная система функций, по которой любой элемент разлагается единственным образом в сходящийся ряд.

Фундаментальная система может не быть базисом, что можно продемонстрировать на следующем примере. Сумма целых неотрицательных степеней x^n на отрезке будет фундаментальной системой в пространстве $C([a, b])$, потому что любую функцию можно приблизить многочленом. Однако это не будет базисом, потому что не для всякой непрерывной функции существует разложение в ряд. Такая система будет являться базисом для пространства функций, аналитических в круге.

Можем представить произвольное решение однородного уравнения в виде линейной комбинации функций фундаментальной системы:

$$u(x) = \sum_{i=1}^{\infty} c_i \varphi_i(x).$$

Если $\varphi_i(x)$ ортонормированы, то $c_i = (u, \varphi_i)$. Тогда

$$Ku(x) = \sum_{i=1}^{\infty} c_i \mu_i^{-1} \varphi_i(x) = \int_a^b \left(\sum_{i=1}^{\infty} \mu_i^{-1} \varphi_i(x) \varphi_i(t) \right) u(t) dx dt,$$

то есть ядро имеет вид

$$K(x, t) = \sum_{i=1}^{\infty} \mu_i^{-1} \varphi_i(x) \varphi_i(t).$$

Если мы продифференцируем это разложение, то получим

$$\frac{\partial^{p+q} K}{\partial x^p \partial t^q} = \sum_{i=1}^{\infty} \mu_i^{-1} \varphi_i^{(p)}(x) \varphi_i^{(q)}(t).$$

Это разложение существует, если ряд сходится. Что может происходить далеко не всегда. Например, функция $\sin nx$ после дифференцирования только увеличивается. На самом деле если μ_i^{-1} – быстро убывающая последовательность, то она может скомпенсировать рост производных. Таким образом, *гладкость ядра связана с быстротой убывания последовательности собственных чисел*.

Теперь рассмотрим вырожденное ядро.

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

³³Вроде это свойство называется *полнотой* ядра.

где α_i и β_i можем НУО считать линейно-независимыми (иначе можем просто выразить одну функцию как линейную комбинацию других и избавиться от неё). Можем подставить это ядро в интегральное уравнение (25), получим

$$\varphi(x) = f(x) + \sum_{i=1}^n A_i \alpha_i(x), \quad \text{где } A_i = (\beta_i, \varphi) = \int_a^b \beta_i(t) \varphi(t) dt.$$

Подставим полученное выражение для φ в (25)

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

Чтобы записать это короче, введём обозначения

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

с ними 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 функции α_i линейно-независимы, коэффициенты при них справа и слева должны быть равны. Тогда равенство переписывается как

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

можем переписать его в векторном виде:

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

где A и f – векторы, β – матрица, а μ – число. Решение этой системы:

$$(I - \mu \beta) A = \mu f \Rightarrow A = \mu (I - \mu \beta)^{-1} f.$$

Мы хотим аппроксимировать произвольное ядро вырожденным. Это можно сделать тремя способами:

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

Заменяя ядро на вырожденное, мы надеемся, что решения изменятся не сильно. Это надо обосновать. Пусть есть уравнение

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

а приближающее его уравнение –

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

где K_n – вырожденное ядро. Нетрудно видеть, что

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

Оценим норму разности обратных операторов.

Определение. *Ограниченный оператор* – такой оператор, для которого $\exists C$:

$$\forall u \quad \|Au\| \leq C \|u\|.$$

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

$$(I - P)^{-1} = \sum_{i=0}^{\infty} P^i,$$

где сходимость – по операторной норме.

Отсюда следуют оценки

$$\begin{aligned} \|(I - P)^{-1}\| &= \left\| \sum_{i=0}^{\infty} P^i \right\| \leq \sum_{i=0}^{\infty} \|P\|^i = \frac{1}{1 - \|P\|}, \\ \|(I - P)^{-1} - I\| &= \left\| \sum_{i=1}^{\infty} P^i \right\| \leq \sum_{i=1}^{\infty} \|P\|^i = \frac{\|P\|}{1 - \|P\|}. \end{aligned}$$

Из этого утверждения следует следующее.

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

$$\begin{aligned} \|(P - H)^{-1}\| &\leq \frac{\|P^{-1}\|}{1 - \|H\|\|P^{-1}\|}, \\ \|(P - H)^{-1} - P^{-1}\| &\leq \frac{\|H\|\|P^{-1}\|^2}{1 - \|H\|\|P^{-1}\|}. \end{aligned}$$

А уже из этого следует

Утверждение. При достаточно больших n

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

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

Теперь рассмотрим задачу с симметричным ядром $K(x, t) = K(t, x)$. В ней существует ортонормированная фундаментальная система решений α_i однородного уравнения. Поэтому можем выписать

$$u = \sum_{i=1}^{\infty} (u, \alpha_i) \alpha_i, \quad Ku = \sum_{i=1}^{\infty} (u, \alpha_i) \lambda_i \alpha_i,$$

где λ_i – соответствующее собственное число. Расположим собственные числа в порядке убывания модуля и положим

$$K_n u = \sum_{i=1}^n (u, \alpha_i) \lambda_i \alpha_i.$$

Это интегральный оператор с вырожденным ядром

$$K(x, t) = \int_a^b \lambda_i \alpha_i(x) \alpha_i(t).$$

Разность выпишется как

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

Можем оценить её норму

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

При этом

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

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

При этом эта оценка достигается ровно когда u – собственный вектор числа λ_{n+1} . Поэтому

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

То есть мы получили, что чем быстрее убывают собственные числа, тем лучше оценка. Как мы показывали ранее, быстрое убывание собственных чисел связано с гладкостью. А ядра бывают гладкими не всегда.

Ядро можно сгладить, если подставить в уравнение (25)

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

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 лучше, но его надо считать.

28. 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:

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

то есть уравнение переписывается как

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

Пусть $\tilde{u}(x)$ – решение этого уравнения с отброшенным остатком, $u_k = u(x_k)$, $f_k = f(x_k)$ и $K_{ik} = K(x_i, x_k)$. Получаем СЛАУ

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

Её можем решать обычными методами. Зная u_k , можем оценить $u(x)$ в любой точке:

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

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

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

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

Чтобы оценить их, продифференцируем 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 (вспоминая утверждения предыдущего билета), как

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

где $\varkappa = (b - a) \max |K(s, t)|$ (оценка верна только при $\varkappa < 1$).

То есть мы нашли оценку для $\|u\|$, а значит можем найти оценку для $u^{(l)}$, и, в конце концов, $\frac{\partial^m}{\partial t^m}(K(x, t)u(t))$. Последнюю оценку обозначим как

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

Она зависит только от известных функций.

Теперь перейдём 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

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

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

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

отсюда видно, что $\|\tilde{K}\| \leq \varkappa$. Отсюда

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

Теперь оценим норму $\tilde{A}u - Au$:

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

В конечном итоге находим

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

29. Вариационный принцип для ограниченного оператора. Метод Ритца для интегрального уравнения 2-го рода.

В этом билете все действия происходят в гильбертовом пространстве, в вещественном государстве H . Оператор A всюду определён и ограничен. Основная идея заключается в том, чтобы свести решение уравнения

$$Au = f$$

к минимизации некоторого функционала.

Определение. Энергетическим функционалом для этого уравнения называется

$$\tilde{f}(u) = (Au, u) - 2(f, u).$$

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

Определение. Оператор A называют положительно определённым, если $(Au, u) \geq k^2(u, u)$.

Утверждение. Симметричный положительно определённый оператор A обратим, и обратный к нему оператор ограничен.

□ Положим в доказательстве $k^2 = 1$, ибо на обратимость это не влияет, можно просто разделить A на k^2 . Заметим, что ядро оператора тривиально (состоит только из нуля), поскольку

$$Au = 0 \Rightarrow (Au, u) = 0 \Rightarrow (u, u) = 0 \Rightarrow u \equiv 0.$$

При этом ядро есть ортогональное дополнение образа:

$$x \in \text{Im } A^\perp \Leftrightarrow \forall u \quad 0 = (x, Au) = (Ax, u) \Leftrightarrow Ax = 0.$$

Поэтому замыкание образа оператора A есть всё пространство H . То есть образ оператора всюду плотен в H . Докажем, что он на самом деле равен H . Для этого воспользуемся неравенством

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

Пусть $y \in H$. Поскольку образ A всюду плотен, то найдётся такая последовательность $\{x_n\}$, что $Ax_n \rightarrow y$. При этом

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

поэтому $\{x_n\}$ сходится в себе. Так как гильбертово пространство полно, то существует x , к которому эта последовательность сходится. Так как оператор непрерывен, то

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

то есть оператор A сюръективен, у него есть теоретико-множественный обратный. При этом

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

поэтому обратный оператор ограничен ▣.

Утверждение (Вариационный принцип для ограниченного оператора). Энергетический функционал достигает минимума в единственной точке, которая совпадает с решением уравнения $Au = f$, где A – симметричный и положительно-определённый.

□ Обозначим нужное нам решение за u^* . Существование и единственность решения следуют из обратимости оператора. Посчитаем значение функционала на векторе $u^* + h$:

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

Таким образом,

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

если $h \neq 0$, то неравенство будет строгим \Rightarrow единственность \blacksquare .

Метод Ритца заключается в следующем:

1. Выбираем в пространстве H ЛНЗ набор $\{\varphi_k\}$ – приближающие функции;
2. Рассматриваем конечномерное пространство H_n – линейную оболочку первых n элементов из $\{\varphi_k\}$;
3. Находим в нём точку минимума функционала \tilde{f} и считаем её приближением.

Точку минимума в H_n будем искать в виде

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

Утверждение. Координаты c_n точки минимума \tilde{f} в подпространстве H_n находятся из системы линейных уравнений

$$\sum_{k=1}^n (A\varphi_k, \varphi_i) c_k = (f, \varphi_i). \quad (26)$$

□ Если подставить

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

в выражение для энергетического функционала, то получится

$$\tilde{f}(u_n) = \sum_{k=1}^n \sum_{m=1}^n c_k c_m (A\varphi_k, \varphi_m) - 2 \sum_{m=1}^n c_m (f, \varphi_m).$$

Дифференцируя это выражение по c_i получим нужную СЛАУ \blacksquare .

Обозначим элемент матрицы системы как $a_{ij} = (A\varphi_i, \varphi_j)$. Она симметрична, так как

$$a_{ji} = (A\varphi_j, \varphi_i) = (\varphi_j, A\varphi_i) = (A\varphi_i, \varphi_j) = a_{ij}.$$

Более того, она положительно определена. Чтобы показать это, рассмотрим квадратичную форму

$$\begin{aligned} \sum_{i=1}^n \sum_{k=1}^n a_{ik} \xi_i \xi_k &= \sum_{i=1}^n \sum_{k=1}^n (A\varphi_i, \varphi_k) \xi_i \xi_k = \left(A \sum_{i=1}^n \xi_i \varphi_i, \sum_{k=1}^n \xi_k \varphi_k \right) = \\ &= (Av_n, v_n) \geq k^2 \|v_n\|^2 \geq 0, \end{aligned}$$

по пути мы ввели обозначение $\sum_{i=1}^n \xi_i \varphi_i \equiv v_n$. Знак « $=$ » появится только тогда, когда все $\xi_i = 0$.

Значит матрица положительно определена. А следовательно она обратима, а система с ней однозначно разрешима.

Поговорим о сходимости метода Ритца.

Утверждение. Если набор $\{\varphi_k\}$ является фундаментальной системой, то есть

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

то метод Ритца сходится, то есть $\|u_n - u^*\| \rightarrow 0$.

□ Поскольку оператор A симметричен и положительно определён, то билинейная форма $g(u, v) = (Au, v)$ является скалярным произведением. Будем обозначать норму, порождённую им, как $\|u\|_A = \sqrt{g(u, u)} = \sqrt{(Au, u)}$. Мы утверждаем, что u_n – элемент H_n , ближайший к u^* с точки зрения метрики, порождённой g . Докажем это. Для этого предположим, что

$$u_n = v_n + h,$$

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

$$\tilde{f}(u_n) = \tilde{f}(u^*) + (A(h - v_n^\perp), (h - v_n^\perp)) = \tilde{f}(u^*) + \|v_n^\perp\|_A^2 - \underbrace{2g(v_n^\perp, h)}_{=0} + \|h\|_A^2.$$

Видим, что это выражение минимально при $h = 0$ и $u_n = u^* - v_n^\perp$. Выберем теперь по ε такие N и $w \in H_N$, что $\|w - u^*\| < \varepsilon$. Тогда

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

Обоснуем переходы в этом неравенстве:

1. Потому что $g(x, x) \geq k^2(x, x)$;
2. Так как u_n – ближайший элемент к u^* ;
3. Потому что $(Ax, x) \leq \|Ax\| \|x\| \leq \|A\|(x, x)$.

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

Скорость сходимости метода Рунге не зависит от гладкости ядра, зато зависит от гладкости решения.

30. Интегральное уравнение 1-го рода. Понятие корректности. Некорректность уравнения 1-го рода.

Интегральное уравнение I рода имеет вид

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

Оно относится к классу так называемых некорректных задач. Адамар предложил называть *корректностью* не только однозначную разрешимость задачи, но и её устойчивость по отношению к возмущениям данных, то есть непрерывную зависимость её решения от данных. Примером *некорректной* задачи может послужить задача Коши для уравнения Лапласа.

Определение. будем рассматривать линейные уравнения вида

$$Au = f.$$

Задача решения такого уравнения называется *корректной*, если

1. Для любой правой части задача имеет единственное решение (то есть существует обратный оператор A^{-1} , он определён на области значений A).
2. Решение устойчиво по отношению к возмущениям правой части, то есть изменение решения сколь угодно мало при достаточно малом изменении правой части. По-сути это означает, что обратный оператор A^{-1} ограничен.

Некорректность же состоит в том, что обратный оператор не всюду существует и ограничен.

Покажем, что задача решения интегрального уравнения первого рода некорректна. Мы решаем уравнение вида $Ku = f$, где оператор K – компактный. У компактных операторов очень маленький образ, а следовательно маленькое возмущение правой части может привести к сильному изменению решения, если задача вообще будет разрешима.

Для примера рассмотрим оператор с полным симметричным ядром. Для него будет существовать фундаментальная система из собственных функций $\{\varphi_n\}$ с собственными числами μ_n^{-1} . Последовательность собственных чисел сходится к 0. А последовательность характеристических чисел μ_n стремится к бесконечности. Рассмотрим такие разложения по нашей фундаментальной системе:

$$u = \sum_{i=1}^{\infty} \alpha_i \varphi_i, \quad f = \sum_{i=1}^{\infty} c_i \varphi_i,$$

при этом

$$Ku = \sum_{i=1}^{\infty} \alpha_i \mu_i^{-1} \varphi_i = \sum_{i=1}^{\infty} c_i \varphi_i = f,$$

то есть $\alpha_i = c_i \mu_i$ и разложение решения по собственным функциям принимает вид

$$u = \sum_{i=1}^{\infty} c_i \mu_i \varphi_i,$$

где последовательность $\{\mu_i\}$ стремится к бесконечности, и поэтому этот ряд будет расходиться, если коэффициенты c_i убывают медленнее, чем μ_i растут. А даже если ряд $\sum |c_i|^2 |\mu_i|^2$ сходится, то всё равно мы не получим непрерывности. Рассмотрим, что получится при возмущении начальных данных

$$\tilde{f} = f + \delta f, \quad \tilde{u} = u + \delta u, \quad K\tilde{u} = \tilde{f},$$

возмущение решения примет вид

$$\delta u = \sum_{i=1}^{\infty} \delta c_i \mu_i \varphi_i.$$

Можем для примера выбрать $\delta f = \varepsilon \varphi_n$, где n таково, что $\mu_n^{-1} < \varepsilon \Rightarrow$ норма δu будет больше 1 для любого ε . То есть даже при стремлении δf к нулю мы получаем, что δu не стремится к нулю.

31. Условная корректность по Тихонову. Метод квазирешений.

Теорема (Об условной корректности). Пусть непрерывный оператор A таков, что $Au = 0$ только при $u = 0$ ³⁴. Тогда обратный оператор (он определён на области значений³⁵ A) будет непрерывен на образе каждого компакта.

□ Пусть $L \subset H$ – компакт, а $M = AL$ – его образ. Выберем в M сходящуюся последовательность $f_n \rightarrow f$. Тогда будут существовать такие u_n , что $Au_n = f_n$, причём, из инъективности оператора, соответствие $u_n \leftrightarrow f_n$ – взаимно-однозначное. Можем выбрать в последовательности $\{u_n\}$ сходящуюся подпоследовательность $u'_n \rightarrow u'$.

Поскольку оператор A непрерывен, то $Au'_n \rightarrow Au'$, но $Au'_n \rightarrow f$, то есть $u' = A^{-1}f$, на самом деле все частичные пределы равны. То есть u_n имеет предел $A^{-1}f$, что как раз говорит о том, что обратный оператор непрерывен ▣.

Определение. Это свойство (непрерывность обратного оператора на образе каждого компакта) называется *условной корректностью*.

³⁴Ядро оператора состоит только из нуля – то есть он инъективен. Или биективен на образ.

³⁵Мне кажется, что корректнее говорить на образе A .

Пусть нам поставлена задача $Au = f$. В реальности мы зачастую знаем правую часть лишь с некоторой точностью. То есть вместо f нам дано некоторое $f_\delta : \|f - f_\delta\| \leq \delta$. Но для некорректной задачи даже малое возмущение правой части может вывести уравнение из области значений оператора A . Поэтому такая задача, вообще говоря, неразрешима. Но мы можем решать немного другую задачу: искать приближенное решение, то есть чтобы невязка $\|Au - f_\delta\|$ была мала.

Определение. *Квазирешением* уравнения $Au = f_\delta$ на ограниченном множестве D называется такая функция u_δ , при которой достигается минимум

$$\min_{u \in D} \|Au - f_\delta\|,$$

обычно в качестве множества D берут замкнутый шар $D = \{u : \|u\| \leq R\}$.

Невязка должна минимизироваться при больших u . Поскольку мы ограничили область поиска шаром, то следует ожидать, чтобы минимум достигался на границе этого шара. Поиск минимума при условии $\|u\| = R$ – это задача об условном экстремуме по методу Лагранжа. Будем минимизировать функционал

$$\mathcal{F}_\alpha(u) = \alpha\|u\|^2 + \|Au - f_\delta\|^2,$$

где α – это множитель Лагранжа. Вычислим вариацию

$$\begin{aligned} \mathcal{F}_\alpha(u + v) &= \alpha(u + v, u + v) + (Au + Av - f_\delta, Au + Av - f_\delta) = \\ &= \alpha(u, u) + (Au - f_\delta, Au - f_\delta) + 2\alpha(u, v) + \alpha(v, v) + 2(Au - f_\delta, Av) + (Av, Av) = \\ &= \mathcal{F}_\alpha(u) + 2(A^*Au - A^*f, v) + (A^*Av, v) + 2\alpha(u, v) + \alpha(v, v) = \\ &= \mathcal{F}_\alpha(u) + 2(A^*(Au - f) + \alpha u, v) + o(\|v\|^2). \end{aligned}$$

В последнем равенстве второе слагаемое должно быть равно нулю. Из этого условия получим

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

Это называется регуляризованным уравнением. Его оператор симметричный и положительно-определённый:

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

и поэтому это уравнение разрешимо для любого α . Мы получили уравнение, похожее на исходное при малых α , но при этом относящееся ко второму роду. В этом заключается регуляризация. Конкретно такая регуляризация, которую мы произвели, называется *слабой*.

Можем ли мы рассчитывать, что при малых α решение регуляризованного уравнения будет похоже на решение исходного? На самом деле нет, при слабой регуляризации u_α могут не сходиться к u . Нужны более сильные требования (принадлежность u_α некоторому компакту), но об этом уже следующий билет.

32. Метод регуляризации для уравнения 1-го рода. Сходимость.

Идея регуляризации заключается в том, чтобы минимизировать функционал вида

$$\mathcal{F}_\alpha(u) = \alpha\Omega(u) + \|Au - f\|^2, \tag{27}$$

где $\Omega(u) \geq 0$, а $\{u : \Omega(u) < C\}$ – компактно. В таком случае *квазирешением* уже будет называться элемент, дающий минимум $\|Au - f\|^2$ при условии принадлежности u к компакту.

В методе квазирешений из предыдущего билета использовалась $\Omega(u) = \|u\|^2$. Для неё множества $\{u : \|u\|^2 < C\}$ – это открытые шары, они совсем не компактны.

Стандартный выбор – функционал

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

правда при этом мы ищем решения среди гладких функций.

Утверждение. Для такого функционала множества $\{u : \Omega(u) < C\} = D$ компактны.

□ Это верно по теореме Асколи-Арцела: множество непрерывных функций компактно, тогда и только тогда, когда оно равномерно ограничено³⁶ и равностепенно непрерывно.

Выберем ε . Рассмотрим

$$|u(t') - u(t'')| = \left| \int_{t'}^{t''} u'(t) dt \right| \leq \sqrt{\int_{t'}^{t''} u'^2(t) dt} \sqrt{\int_{t'}^{t''} 1^2 dt} \leq \sqrt{c} \sqrt{\delta} < \varepsilon,$$

это верно при условии $|x' - x''| < \delta$. Второе неравенство в цепочке – неравенство Коши-Буняковского. То есть берём $\delta < \varepsilon^2/C$, и тогда получим равностепенную непрерывность ▣.

Также подошли бы функционалы

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

Теорема (О сходимости регуляризованных решений). Пусть исходное уравнение имеет решение, оператор A обратим на своей области значений. Пусть снова у нас правая часть известна с какой-то точностью. То есть вместо f имеем $f_\delta : \|f - f_\delta\| < \delta$. Решаем приближённую задачу $Au = f_\delta$. Пусть u_α – элемент, на котором функционал (27) минимален. Если δ и α стремятся к нулю таким образом, что

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

то $u_\alpha \rightarrow u^*$ (точное решение).

□ Сначала докажем, что все u_α лежат в компактном множестве

$$\begin{aligned} \alpha\Omega(u_\alpha) &\leq \alpha\Omega(u_\alpha) + \|Au_\alpha - f_\delta\|^2 = \mathcal{F}_\alpha(u_\alpha) \leq \quad (\text{так как } u_\alpha \text{ — минимизирующий элемент}) \\ &\leq \mathcal{F}_\alpha(u^*) = \alpha\Omega(u^*) + \|Au^* - f_\delta\|^2 = \alpha\Omega(u^*) + \|f - f_\delta\|^2 \leq \alpha\Omega(u^*) + \delta^2 \Rightarrow \\ &\Rightarrow \Omega(u_\alpha) \leq \Omega(u^*) + \frac{\delta^2}{\alpha} = \Omega(u^*) + \gamma, \end{aligned}$$

то есть все $\Omega(u_\alpha)$ лежат в пределах $B_\gamma(\Omega(u^*))$. По условию Ω выбраны такие, что если $\Omega(u) \leq C$, то множество таких u компактно. Мы как раз получили такую ограничивающую константу $C = \Omega(u^*) + \gamma$.

Оценим невязку регуляризованного решения

$$\|Au_\alpha - f\| \leq \|Au_\alpha - f_\delta\| + \|f_\delta - f\| \leq \|Au_\alpha - f_\delta\| + \delta.$$

$$\begin{aligned} \|Au_\alpha - f_\delta\|^2 &\leq \alpha\Omega(u_\alpha) + \|Au_\alpha - f_\delta\|^2 = \mathcal{F}_\alpha(u_\alpha) \leq \mathcal{F}_\alpha(u^*) = \alpha\Omega(u^*) + \|Au^* - f_\delta\|^2 \leq \alpha\Omega(u^*) + \delta^2 \rightarrow 0 \Rightarrow \\ \|Au_\alpha - f\| &\leq \|Au_\alpha - f_\delta\| + \delta \rightarrow 0, \end{aligned}$$

то есть $Au_\alpha \rightarrow f$. Поскольку u_α лежат на компакте, то, по теореме об условной корректности, обратный оператор будет непрерывен на образе этого компакта. А значит $u_\alpha \rightarrow A^{-1}f$ ▣.

На практике коэффициент α обычно подбирают эмпирически: если он мал, то решение будет ближе к u^* , а если велик, то оно будет более гладким.

³⁶Тут очень странно, ибо равномерной ограниченности вообще нет. Можно выбрать функцию, удовлетворяющую нашему условию, и прибавить к ней константу.

IV. Вариационные методы.

33. Вариационный принцип для уравнения с неограниченным оператором.

Пусть A – симметричный и положительно определённый оператор, определённый на плотном линейном подпространстве $\mathcal{D}(A)$ гильбертова пространства H .

Напомним определения симметричности:

$$\forall u, v \in \mathcal{D}(A) \quad (Au, v) = (u, Av)$$

и положительной определённости

$$\forall u \in \mathcal{D}(A) \quad \exists k : (Au, u) \geq k^2(u, u).$$

Определение. Билинейная форма $(u, v)_A = (Au, v)$ называется *энергетическим скалярным произведением*, а норма $\|u\|_A = \sqrt{(u, u)_A}$, порождённая ею, называется *энергетической нормой*. Пополнение H_A множества $\mathcal{D}(A)$ называется *энергетическим пространством*.

Стоит немного поговорить о том, что такое пополнение. Полное пространство – это то, в котором каждая сходящая в себе последовательность сходится к элементу этого же пространства. *Кофинальными* последовательностями будем называть такие две сходящиеся в себе последовательности, что $\|u_n - v_n\| \rightarrow 0$. Кофинальность двух последовательностей – это отношение эквивалентности. В качестве элементов пополнения выступают классы эквивалентности сходящихся в себе последовательностей.

$$\|u\|^2 \leq k^{-2} \|u\|_A^2 \Rightarrow \|u\| \leq k^{-1} \|u\|_A,$$

поэтому сходящаяся в себе по энергетической норме последовательность будет сходиться в себе и по основной норме. А значит $\mathcal{D}(A) \subset H_A \subset H$.

Теорема (О вариационном принципе). *Рассмотрим энергетический функционал*

$$\mathcal{F}_A = (u, u)_A - 2(f, u).$$

1. Он достигает минимума в единственной точке u^* ;
2. Если $u^* \in \mathcal{D}(A)$, то $Au^* = f$;
3. Если $Au_0 = f$, то $u^* = u_0$.

(Экстремальный элемент существует всегда, но не всегда является решением). Решение из $\mathcal{D}(A)$ называется *классическим*.

□ Рассмотрим функционал $\Phi(u) = (f, u)$, действующий в пространстве H_A . Он ограничен, поскольку

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

По теореме Рисса он представим в виде скалярного произведения

$$\exists! u^* : (f, u) = (u^*, u)_A \Rightarrow$$

$$\Rightarrow \mathcal{F}_A(u) = (u, u)_A - 2(u^*, u)_A = (u - u^*, u - u^*)_A + 2(u^*, u)_A - (u^*, u^*)_A - 2(u^*, u)_A = \|u - u^*\|_A^2 - \|u^*\|_A^2.$$

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

Третий пункт очевиден, так как

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

То же самое равенство в обратную сторону даёт пункт 2 ▣.

34. Метод Ритца. Сходимость.

Метод Ритца для неограниченных операторов похож на обычный, но есть некоторые отличия. Набор $\{\varphi_k\}$ теперь мы будем выбирать в H_A . А ещё приближенное решение по Ритцу можно определить как $u_n^* \in H_n : (u_n^*, v)_A = (f, v) \forall v \in H_n$. Это то же самое, что искать минимум энергетического функционала. Также, система для коэффициентов c_k разложения u_n^* по $\{\varphi_k\}$ будет такой:

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

что слегка отличается от системы (26) для случая ограниченного оператора. Совпадение будет только если $\varphi_k \in \mathcal{D}(A)$.

Теорема. Если набор $\{\varphi_k\}$ является фундаментальной системой, то метод Ритца сходится по энергетической норме. То есть

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

□ Доказательство аналогично таковому для случая ограниченного оператора. Вначале также доказываем, что u_n^* – ближайший к u^* элемент H_n . А после можем по ε выбрать такое N и $w \in H_N$, что $\|w - u^*\|_A < \varepsilon$. Тогда при $n > N$ будет верно

$$\|u_n^* - u^*\|_A \leq \|w - u^*\|_A < \varepsilon,$$

что верно $\forall \varepsilon$ ■.

35. Метод Ритца для обыкновенной краевой задачи. Вид энергетического пространства. Естественные граничные условия.

Рассмотрим краевую задачу для уравнения

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

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. Классическое решение – лежит в $C^2([a, b])$, удовлетворяет уравнению в каждой точке.

То есть в качестве $\mathcal{D}(L)$ у нас выступает $C^2([a, b])$. Но на самом деле мы ищем решение в более узких пространствах. В случае граничного условия I типа нас интересует пространство

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

а в случае условия III типа – пространство

$$\mathcal{D}_{III} = \{y \in \mathcal{D}(L) : y'(a) = \alpha y(a); y'(b) = \beta y(b)\}.$$

Именно их мы будем пополнять, создавая соответствующее энергетическое пространство.

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

$$p(x) \geq 0, \quad q(x) \geq q_0 > 0,$$

то оператор получится симметричный и положительно определённый.

□ Энергетическое скалярное произведение будет иметь вид

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

Обозначим внеинтегральный член за $Q(y, z)$. Для условий I типа он равен нулю. Для условий III типа:

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

Видно, что эти выражения симметричны относительно y и z .

$$(Ly, y) = \int_a^b (py'^2 + qy^2) \, dx - \beta p(b)y(b)^2 + \alpha p(a)y(a)^2 \geq q_0 \int_a^b y^2 \, dx = q_0(y, y).$$

То есть оператор положительно определён \blacksquare .

Определение. Пусть u и f – суммируемые на $[a, b]$ функции. Тогда, если выполнено равенство

$$\int_a^b \varphi' u \, dx = \varphi u|_a^b - \int_a^b \varphi f \, dx \quad \forall \varphi \in C^1([a, b]),$$

то f называется *обобщённой производной* функции u .

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

Нас будет интересовать пространство $W_2^1([a, b])$. Оно гильбертово (в то время как произвольные пространства Соболева – банаховы). Скалярное произведение на нём таково:

$$(y, z)_{W_2^1} = \int_a^b (y'z' + yz) \, dx.$$

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

□ Пусть $P_m = \max p$, $Q_m = \max q$, $M = \max(P_m, Q_m)$. Тогда

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

Обратное утверждение очевидно для I типа граничных условий:

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

А для условия III типа нам понадобится лемма.

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

□ Оценка для 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 $(a + b)^2 \leq 2(a^2 + b^2)$, то

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

При этом интеграл

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

является L^2 -произведением y' и функции, тождественно равной единице. Поэтому можем записать

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

В итоге получаем, что

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

Проинтегрировав по ξ , получим

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

Лемма доказана \blacksquare .

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

Помимо прочего, выполняется *теорема вложения*: любая функция из W_2^1 непрерывна, причём отображение вложения $W_2^1 \rightarrow C([a, b])$ непрерывно. Это, вместе с доказанным утверждением, говорит о том, что $\mathcal{D}(L)$ гомеоморфно вкладывается в W_2^1 .

Утверждение. Энергетическое пространство H_L является подпространством W_2^1 .

\square Не очень важно, будем ли мы пополнять \mathcal{D}_I или \mathcal{D}_{III} : они оба лежат в $\mathcal{D}(L)$, которое вкладывается в W_2^1 гомеоморфно. Из гильбертовости W_2^1 следует, что пополнение $\mathcal{D}(L)$ по энергетической норме не выведет нас из W_2^1 \blacksquare .

Работает ли обратное включение $H_L \supset W_2^1$? Для \mathcal{D}_{III} работает, поскольку производная будет непрерывным оператором. Поэтому для III граничного условия верно, что $H_L = W_2^1$. А для I граничного условия это не так. Пополнение \mathcal{D}_I приведёт нас к пространству \dot{W}_2^1 функций из W_2^1 , удовлетворяющих граничному условию I типа. По этой причине условия I типа называют *главными*, а III типа – *естественными*.

36. ВРМ-1 для обыкновенной краевой задачи.

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

Пусть в сетке n элементов, $h = \frac{b-a}{n}$, $x_k = a + kh$; рассмотрим пространство, состоящее из сеточных функций $y_{(n)} = \{y_k\}_{k=0}^n$. Суть ВРМ-1 в том, чтобы заменить интегралы на суммы, производные – на разности, а функционал – на сеточный, то есть заданный на сеточных функциях, и потом минимизировать его. Наш функционал имеет вид

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

Сделаем численные замены:

$$\int_a^b p y'^2 dx \approx h \sum_{k=0}^{n-1} p \left(x_k + \frac{h}{2} \right) \left(\frac{y_{k+1} - y_k}{h} \right)^2;$$

$$\int (q y^2 - 2 f y) dx \approx h \sum_{k=0}^{n-1} (q_k y_k^2 - 2 f_k y_k),$$

где сумма со штрихом означает формулу трапеций (то есть крайние слагаемые домножены на $1/2$). Не представляет труда теперь выписать сеточный функционал:

$$\mathcal{F}(L) \approx h \sum_{k=0}^{n-1} p_{k+\frac{1}{2}} \left(\frac{y_{k+1} - y_k}{h} \right)^2 + h \sum_{k=0}^{n-1} (q_k y_k^2 - 2 f_k y_k) - \beta p_n y_n^2 + \alpha p_0 y_0^2.$$

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 разностного метода.

Для левого конца получится уравнение

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

Здесь стоило ожидать простейшее приближение $y'(a) = \alpha y(a)$, но мы получили неожиданное второе слагаемое. На самом деле оно компенсирует сдвиг:

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

37. ВРМ-2 для обыкновенной 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. Для восполнения 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 – это множество меры ноль, поэтому эта функция всё равно остаётся суммируемой, производная будет определена в смысле обобщённой производной наших восполненных сеточных функций. Поэтому они лежат в W_2^1 .

Можно ввести базисные функции – это восполнения сеточных функций, которые равны нулю всюду, кроме одной точки, а в ней равны единице, то есть

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

Получилось что-то очень похожее на метод Рунге, 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 пересекаются только с соседними носителями, поэтому матрица получится трёхдиагональной:

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

где

$$a_m = a_{m-1,m}; \quad b_m = a_{mm}.$$

Даже на гладких решениях мы не получим точности лучше, чем $O(h)$. Однако этот метод надёжнее для негладких решений, чем просто сеточный.

38. Метод Рунге для эллиптического уравнения. Вид естественного граничного условия. Вид энергетического пространства.

Рассмотрим уравнение

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

Коэффициенты должны удовлетворять нескольким условиям:

1. Все функции действуют в области $\Omega \subset \mathbb{R}^n$. Обычно решение считают дважды непрерывно дифференцируемым на Ω и непрерывным в $\bar{\Omega}$,³⁷ a_{ij} – один раз непрерывно дифференцируемы на $\bar{\Omega}$, а остальные коэффициенты просто непрерывны. Матрицу $\{a_{ik}\}$ можно считать симметричной, так как смешанные производные симметричны по аргументам.
2. Положительная определённость (эллиптичность):

$$\sum_{i=1}^n \sum_{k=1}^n a_{ik} \xi_i \xi_k \geq k^2 \sum_{i=1}^n \xi_i^2, \quad k^2 > 0.$$

Рассмотрим скалярное произведение:

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

вспомним, как выглядит интегрирование по частям:

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

³⁷Черта сверху – замыкание.

где $\cos(n, x_i)$ – это косинус угла между нормалью \mathbf{n} и ортом i -ой оси \mathbf{e}_i

$$\equiv \int_{\Omega} \left(\sum_{i=1}^n \sum_{k=1}^n a_{ik} \frac{\partial u}{\partial x_k} \frac{\partial v}{\partial x_i} + auv \right) dx + \int_{\partial\Omega} \sum_{i=1}^n \sum_{k=1}^n a_{ik} \frac{\partial u}{\partial x_k} \cos(n, x_i) v dS.$$

Введём обозначение

$$\frac{\partial u}{\partial n^*} \Big|_{\partial\Omega} = \sum_{i=1}^n \sum_{k=1}^n a_{ik} \frac{\partial u}{\partial x_k} \Big|_{\partial\Omega} \cos(n, x_i),$$

этот оператор называется *конормальной производной*. Если в нём положить $a_{ik} = \delta_{ik}$, то мы получим

$$\sum_{i=1}^n \sum_{k=1}^n \delta_{ik} \frac{\partial u}{\partial x_k} \Big|_{\partial\Omega} \cos(n, x_i) = \sum_{i=1}^n \frac{\partial u}{\partial x_i} \Big|_{\partial\Omega} \cos(u, x_i) = \frac{\partial u}{\partial n} \Big|_{\partial\Omega},$$

это уже является *нормальной производной*. Для оператора Лапласа конормальная и нормальная производные совпадают. В итоге можем выписать:

$$(Lu, v) = \int_{\Omega} \left(\sum_{i=1}^n \sum_{k=1}^n a_{ik} \frac{\partial u}{\partial x_k} \frac{\partial v}{\partial x_i} + auv \right) dx + \left(\frac{\partial u}{\partial n^*} \Big|_{\partial\Omega}, v \right).$$

Граничные условия бывают

I типа: $u|_{\partial\Omega} = 0$ – задача Дирихле;

II типа: $\frac{\partial u}{\partial n^*} \Big|_{\partial\Omega} = 0$;

III типа: $\frac{\partial u}{\partial n^*} \Big|_{\partial\Omega} = \sigma u|_{\partial\Omega}$.

Соответственно, энергетическое произведение будет иметь вид

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

второе слагаемое здесь обращается в ноль для граничных условий I и II типа.

Утверждение. 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 $a(x) \geq 0$, $\sigma(x) \geq \sigma_0 > 0$.

□

$$(Lu, u) = \int_{\Omega} \left(\sum_{i=1}^n \sum_{k=1}^n a_{ik} \frac{\partial u}{\partial x_k} \frac{\partial u}{\partial x_i} + au^2 \right) dx + \int_{\partial\Omega} \sigma u^2 dS.$$

Нам понадобится неравенство Фридрихса

$$\int_{\Omega} u^2 dx \leq c_1 \left(\int_{\Omega} \sum_{i=1}^n \left(\frac{\partial u}{\partial x_i} \right)^2 dx + \int_{\partial\Omega} u^2 dS \right).$$

1. Здесь работает совсем грубая оценка:

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

2. Ещё проще, как это ни странно.

$$\int_{\Omega} \left(\sum_{i=1}^n \sum_{k=1}^n a_{ik} \frac{\partial u}{\partial x_k} \frac{\partial u}{\partial x_i} + au^2 \right) dx \geq a_0 \int_{\Omega} u^2 dx.$$

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

$$\int_{\Omega} \left(\sum_{i=1}^n \sum_{k=1}^n a_{ik} \frac{\partial u}{\partial x_k} \frac{\partial u}{\partial x_i} + au^2 \right) dx + \int_{\partial\Omega} \sigma u^2 dS \geq k^2 \int_{\Omega} \sum_{i=1}^n \left(\frac{\partial u}{\partial x_i} \right)^2 dx + \sigma_0 \int_{\partial\Omega} u^2 dS \geq c_2 \int_{\Omega} u^2 dx,$$

где $c_2 = \frac{\min(k^2, \sigma_0)}{c_1}$ ■.

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

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.

V. Метод сеток для уравнений в частных производных.

39. Разностный метод для общего уравнения теплопроводности. Явная схема.

Определение. Общее уравнение теплопроводности выглядит 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 на отрезке $[a, b]$; временной отрезок будет $[0, T]$.

Определение. У уравнения теплопроводности бывает начальное условие:

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

а также 3 типа граничных условий:

I тип: $u(a, t) = \alpha_1(t), \quad u(b, t) = \alpha_2(t);$

II тип: $\frac{\partial u}{\partial x}(a, t) = \alpha_1(t), \quad \frac{\partial u}{\partial x}(b, t) = \alpha_2(t);$

III тип: $\frac{\partial u}{\partial x}(a, t) - \beta_1 u(a, t) = \alpha_1(t), \quad \frac{\partial u}{\partial x}(b, t) - \beta_2 u(b, t) = \alpha_2(t).$

Сетка характеризуется следующими величинами:

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

Обозначим $u_i^k = u(x_i, t_k)$ 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 производной по времени:

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

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

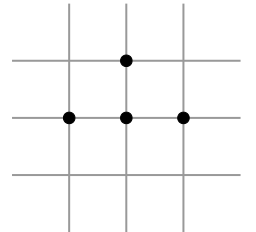
Для варианта (A) получается

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

Это простейшая явная схема. *Слой* называется совокупность точек (x_i, t_k) с одним t_k . Уравнение содержит только одно неизвестное с верхнего слоя.

В таком виде полученные уравнения можно писать для всех $i \in \{1, \dots, n-1\}$, $k \in \{0 \dots M-1\}$. Нужны дополнительные уравнения для границ.

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

$$\varphi(a) = \alpha_1(0); \quad \varphi(b) = \alpha_2(0).$$

Они необходимы для непрерывности решения.

2. Для типов II, III используются те же методы, как в обычных дифференциальных уравнениях. 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.

40. Неявная схема для уравнения теплопроводности.

Если мы для аппроксимации производной по времени выберем вариант (Б), то получим неявную схему. Сверху вниз посчитать не получится, потому что начальные условия заданы снизу. Формулы получатся такие:

$$A_i^k u_{i-1}^k - B_i^k u_i^k + C_i^k u_{i+1}^k = D_i^k,$$

где

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

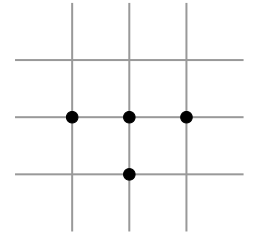
и $\sigma = \frac{\tau}{h^2}$. По сути, движение всё ещё послойное. Но нельзя просто так взять и посчитать значение на каждом слое, используя три значения с предыдущего слоя: наоборот, получается уравнение, которое связывает три значения с текущего слоя с одним уже известным. В итоге получается система с трёхдиагональной матрицей, которая замыкается добавлением граничных условий:

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

если они заданы по первому типу, в противном случае применяются стандартные аппроксимации. Система решается методом разностной прогонки (см. билет 6). Метод прогонки срабатывает, поскольку

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

и можно сослаться на Лемму из билета 7. Всего операций $O(n) \cdot M = O(n \cdot M)$. В обоих методах число операций одного порядка. При применении разностного метода решения могут расходиться. Причина тому – неустойчивость, связанная со свойствами системы уравнений.



41. Явная схема для простейшего уравнения теплопроводности. Решение разностных уравнений. Явление неустойчивости.

Рассмотрим уравнение

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

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

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

Для него разностные уравнения исключительно просты:

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

При этом

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

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

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

Подставим:

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

Несложными выкладками отсюда находится

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

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

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

– тоже будет решением. Оно удовлетворяет граничным условиям при целых m . В итоге получаем

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

Если мы рассмотрим $m \in \{1, \dots, n-1\}$, то получим набор из $n-1$ ЛНЗ решения. Из них можно собирать новые решения:

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

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

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

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

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

$$\cos \frac{(n-1)\pi}{n} = \cos \left(\pi - \frac{\pi}{n} \right) \approx -1 \Rightarrow \lambda(n-1) \approx -3.$$

При увеличении k решение

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

быстро меняет знак. Плохо! Реальное решение такой задачи – быстро убывающая колеблющаяся функция. Конечно, пространственный шаг взят большим: у начальных данных есть переменность на том же масштабе. Однако то, что при уменьшении шага по времени решение получает возможность становиться больше, уже вообще ни в какие ворота не лезет.

Чтобы побороть неустойчивость, можно наложить ограничение $|\lambda| \leq 1$:

$$1 + 2\sigma(\cos mh - 1) \geq -1 \Rightarrow 2\sigma(1 - \cos mh) \leq 2.$$

Чтобы это выполнялось при любых m , нужно, чтобы

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

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

$$\lambda = \frac{1}{1 + 2\sigma(1 - \cos mh)} \in [0, 1].$$

Таким образом, устойчивость всегда присутствует, можем выбирать любой шаг.

42. Общее определение устойчивости. Теорема об устойчивости и сходимости.

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

$$Au = f.$$

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

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

$$\begin{array}{ccc} U & \xrightarrow{A} & F \\ \varphi_h \downarrow & & \downarrow \psi_h \\ U_h & \xrightarrow{A_h} & F_h \end{array}$$

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

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

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

называется *естественной погрешностью* метода.

Далее, записывается и решается разностное уравнение

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

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

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

$$\|\varphi_h(u)\|_{U_h} \rightarrow \|u\|_U$$

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

Будем считать, что у нас нормы согласованы.

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

$$\|A_h \varphi_h(u) - \psi_h(Au)\| \xrightarrow{h \rightarrow 0} 0.$$

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

$$\|u_h - \varphi_h(u)\| \xrightarrow{h \rightarrow 0} 0.$$

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

Определение. Говорят, что сеточное приближение устойчиво, если

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

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

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

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

□ Запишем ошибку сеточного решения:

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

Заметим, что по свойству устойчивости

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

Оба слагаемых здесь стремятся к нулю по свойству аппроксимации ▣.

43. Разностные схемы для задач с начальными условиями. Дискретное преобразование Фурье.

Будем работать с многомерным уравнением

$$\frac{\partial \mathbf{u}}{\partial t} = \hat{L} \mathbf{u} + \mathbf{f},$$

где $\mathbf{x} = (x_1, \dots, x_s)$, $\mathbf{u} = \mathbf{u}(\mathbf{x}, t) = (u^1, \dots, u^p)$, а матричный дифференциальный оператор \hat{L} состоит из частных производных по x_i с постоянными коэффициентами. Считаем областью определения \mathbf{u} цилиндр $\Omega \times [0, T]$. Начальные условия: $\mathbf{u}(\mathbf{x}, 0) = \varphi(\mathbf{x})$.

Пусть $\Omega = [0, 2\pi]^s$ – куб, а решение периодически по каждой из переменных с периодом 2π . Решение на кубе допускает достаточно гладкое периодическое продолжение, если не периодична не только \mathbf{u} , но и её производные.

Сетку строим с одинаковым шагом по каждой из пространственных переменных

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

а по временной – шаг

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

Причём для пространственных переменных можно не рассматривать крайние правые значения, потому что они из-за периодичности совпадают с левыми. Поэтому для сеточной области можем ограничиться «кубом без правой и верхней границы»:

$$\Omega_h = \{(\mathbf{x}, t) : \mathbf{x} = (\alpha_1 h, \dots, \alpha_s h), t = t_k = k\tau\}, \text{ где } \alpha_j \in \{0, \dots, N-1\}.$$

$$\mathcal{D}_h = \{\mathbf{x} : \mathbf{x} = (\alpha_1 h, \dots, \alpha_s h) \text{ где } \alpha_j \in \{0, \dots, N-1\}.$$

Так как \hat{L} имеет постоянные коэффициенты, то и сеточное уравнение будет иметь постоянные коэффициенты (то есть не зависящие от точки). Отказываемся от рассмотрения граничных условий. Будем изучать условия аппроксимации самого оператора \hat{L} .

Пусть уравнения двухслойные: неизвестные будут с k -го и $k+1$ -го слоёв. Каждый слой в данном случае будет являться кубом с левой границей, но без правой. Всего точек в пространственной части куба N^s штук, в каждой точке записано $\dim \mathbf{u} = p$ уравнений и всего таких пар слоёв M штук. В итоге получаем MpN^s уравнений. А неизвестных $(M+1)pN^s$. Недостающие уравнения берём из начальных условий.

Обозначим $\mathbf{u}_h(k) = \{\mathbf{u}_h(\mathbf{x}, t_k) : \mathbf{x} \in \mathcal{D}_h\}$ – как я понял, собрали значения \mathbf{u}_h для всех точек \mathbf{x} с одного слоя в один длинный вектор. $\dim \mathbf{u}_h(k) = pN^s$. Пространство из таких векторов обозначим как V_h . Считаем, что на нём задана какая-нибудь норма, например l^∞ :

$$\|\mathbf{u}_h(k)\|_{V_h} = \max_{\substack{i \in \{1, \dots, p\} \\ \mathbf{x} \in \mathcal{D}_h}} (u_h^i(\mathbf{x}, t_k)).$$

А если рассмотрим целиком всё решение \mathbf{u}_h , то есть его значения для всех вообще точек сетки (\mathbf{x}, t_k) , то такие вектора будут сидеть в пространстве, которое мы обозначим как U_h . Норму на нём определяем через норму V_h :

$$\|\mathbf{u}_h\|_{U_h} = \max_{k \in \{0, \dots, M\}} \left(\|\mathbf{u}_h(k)\|_{V_h} \right)$$

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

$$\mathbf{u}_h(k+1) = \hat{R}_h \mathbf{u}_h(k) + \boldsymbol{\rho}_h(k),$$

где $\hat{R}_h : V_h \rightarrow V_h$ – большущая матрица с постоянными коэффициентами, называемая *оператором перехода для однородного уравнения*, а $\boldsymbol{\rho}_h(k)$ зависит от \mathbf{f} . При однородном уравнении $\mathbf{f} = 0$ также получим $\boldsymbol{\rho}_h = 0$, отсюда виден смысл названия матрицы \hat{R}_h .

Теорема (1). *Для устойчивости при $f = 0$ необходимо и достаточно, чтобы были ограничены \hat{R}_h^k : $\|\hat{R}_h^k\| \leq c_1$ (здесь k – это степень!) $\forall h$ и $\forall k : k\tau \leq T$.*

□ Достаточность: Из начальных условий $\mathbf{u}_h(0) = \varphi_h$. Тогда, поскольку уравнение однородное, на каждом шаге у нас получалось просто умножение на матрицу \hat{R}_h , что привело к

$$\mathbf{u}_h(k) = \hat{R}_h^k \varphi_h.$$

Тогда норма сеточного решения оценится через норму начальных условий

$$\|\mathbf{u}_h\|_{U_h} = \max_{k \in \{0, \dots, M\}} \|\mathbf{u}_h(k)\|_{V_h} \leq \|\hat{R}_h^{k_{\max}}\| \cdot \|\varphi_h\|_{V_h} \leq c_1 \|\varphi_h\|_{V_h}.$$

Необходимость: Пусть есть устойчивость, а ограниченности нет. Тогда возьмём такие h_l, k_l , что

$$k_l \tau \leq T, \text{ но } \|\hat{R}_{h_l}^{k_l}\| > l, \text{ тогда } \exists \varphi_l \in V_h : \|\varphi_l\|_{V_h} = 1$$

тогда если $\mathbf{u}_h(0) = \varphi_l$, то получим $\mathbf{u}_h(k) = \hat{R}_{h_l}^{k_l} \varphi_l$, норма оценится как³⁸

$$\|\mathbf{u}_h\|_{U_h} = \max_{k \in \{0, \dots, M\}} \|\mathbf{u}_h(k)\|_{V_h} \geq \|\mathbf{u}_h(k_l)\|_{V_h} = \|\hat{R}_{h_l}^{k_l}\| \cdot \|\mathbf{u}_h(0)\|_{V_h} > l,$$

и это всё верно $\forall l > 0$, то есть устойчивости нет. Получили противоречие, значит необходимость доказана ▣.

Теорема (2). Если $\mathbf{f} \neq 0$ и выполнены условия теоремы 1, то для устойчивости достаточно, чтобы выполнялось

$$\|\rho_h\|_{V_h} \leq c_2 \tau \|\mathbf{f}_h\|_{F_h}.$$

□ Покажем, что в неоднородном случае выполнено

$$\mathbf{u}_h(k) = \hat{R}_h^k \varphi_h + \sum_{j=1}^k \hat{R}_h^{k-j} \rho_h(j).$$

Индукционный переход:

$$\mathbf{u}_h(k+1) = \hat{R}_h \mathbf{u}_h(k) + \rho_h(k+1) = \hat{R}_h^{k+1} \varphi_h + \sum_{j=1}^k \hat{R}_h^{k-j+1} \rho_h(j) + \rho_h(k+1) = \hat{R}_h^{k+1} \varphi_h + \rho_h(k+1).$$

Оценим норму

$$\|\mathbf{u}_h(k)\|_{V_h} \leq c_1 \|\varphi_h\|_{V_h} + \sum_{j=1}^k c_1 \tau c_2 \|\mathbf{f}_h\|_{F_h} = c_1 \|\varphi_h\|_{V_h} + k \tau c_1 c_2 \|\mathbf{f}_h\|_{F_h} \leq c_1 \|\varphi_h\|_{V_h} + T c_1 c_2 \|\mathbf{f}_h\|_{F_h}.$$

раз ограничены $\|\mathbf{u}_h(k)\|_{V_h}$, то будут ограничены и $\|\mathbf{u}_h\|_{U_h}$ ▣.

Следствие. Если $\|\hat{R}_h\| \leq 1 + c_3 \tau$, то при $f = 0$ имеет место устойчивость. Действительно,

$$\|\hat{R}_h^k\| \leq \|\hat{R}_h\|^k \leq (1 + c_3 \tau)^k \leq e^{c_3 \tau k} \leq e^{c_3 T}.$$

Теперь перейдём к дискретному преобразованию Фурье. Пусть $s = 1$ и $p = 1$. Тогда пространство V_h будет N -мерным и комплексным, $V_h = \mathbb{C}^N$. Введём на нём скалярное произведение:

$$(v_h, w_h)_{V_h} = h \sum_{k=0}^{N-1} v_k \overline{w_k}, \quad \text{где } v_k = v(kh).$$

Рассмотрим набор функций $e_m(x) = e^{imx}$, где $x = nh$, $n \in \{0, \dots, N-1\}$. Таких функций всего N различных:

$$e_{m+N}(x) = e^{imx} \cdot e^{iNx} = e^{imx} \cdot e^{2\pi i} = e^{imx} = e_m(x).$$

³⁸ Второе равенство в следующей цепочке – именно равенство, а не « \leq », по определению операторной нормы.

Утверждение. Этот набор функций образует ортогональный базис в V_h , а $(e_m, e_m) = 2\pi$.

□ Посчитаем скалярное произведение

$$(e_m, e_l)_{V_h} = h \sum_{k=0}^{N-1} e_m(kh) \overline{e_l(kh)} = h \sum_{k=0}^{N-1} e^{i(m-l)kh} = h \sum_{k=0}^{N-1} z^k,$$

где $z = e^{i(m-l)h} = e^{i(m-l)\frac{2\pi}{N}}$. Получаем, что $z^N = 1$. А $z = 1$, если $m = l$ (или даже $m = \text{вычет } l \text{ по модулю } N$). В целом получаем

$$(e_m, e_l)_{V_h} = h \sum_{k=0}^{N-1} z^k = h \frac{z^N - 1}{z - 1} = 0 \text{ при } z \neq 1.$$

А если $z = 1$, то

$$(e_m, e_m)_{V_h} = h \sum_{k=0}^{N-1} 1 = Nh = 2\pi.$$

Этих функций N штук и они ортогональны \Rightarrow линейно-независимы, значит они образуют базис ▣.

Определение. Обратное дискретное преобразование Фурье:

$$v_h(x) = \{a_0, \dots, a_{N-1}\}(x) \mapsto \sum_{m=0}^{N-1} a_m e_m(x);$$

Прямое дискретное преобразование Фурье:

$$u_h(x) \mapsto \frac{1}{2\pi} \{(u_h, e_0), \dots, (u_h, e_{N-1})\};$$

видно, что эти операторы обратны друг другу.

Для дискретного преобразования Фурье выполняется *формула замкнутости*:

$$(v_h, w_h) = \left(\sum_{m=0}^{N-1} a_m e_m, \sum_{l=0}^{N-1} b_l e_l \right) = \sum_{m=0}^{N-1} \sum_{l=0}^{N-1} a_m \overline{b_l} (e_m, e_l) = 2\pi \sum_{m=0}^{N-1} a_m \overline{b_m}.$$

e^{imx} – собственная функция оператора сдвига

$$T_h u(x) = u(x + h)$$

с собственным числом e^{imh} . Это легко проверить прямой подстановкой:

$$T_h e^{imx} = e^{im(x+h)} = e^{imx} e^{imh}.$$

У нас периодические граничные условия, поэтому оператор сдвига может действовать, «переходя» через границу:

$$T_h \{u_0, u_1, \dots, u_{N-2}, u_{N-1}\} = \{u_1, u_2, \dots, u_{N-1}, u_0\}.$$

Можно представлять себе, что индекс i на самом деле меняется от $-\infty$ до ∞ , но $u_{i+N} = u_i$. Периодическую функцию можно восстановить, зная её значения внутри периода, вот и здесь так же.

В такой ситуации любой разумный разностный оператор можно собрать из операторов сдвига. Например, пусть

$$(Du)_i = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2},$$

его можно переписать в виде

$$Du = \frac{T_h u - 2u + T_{-h} u}{h^2}.$$

Общая формула, естественно, будет такая:

$$Lu = \sum_{\alpha \in A} c(\alpha) T_h^\alpha u,$$

где $\alpha \in A \subset \mathbb{Z}$ – показатель степени.

Теперь разберёмся с многомерным дискретным преобразованием Фурье. Здесь будет $\mathbf{x} = (x_1, \dots, x_s)$, на сетке: $\mathbf{x} = h\boldsymbol{\alpha}$, где $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_s) \in \{0, \dots, N-1\}^s = M_0$ – мультииндекс. $\mathbf{m} \in M_0$ – тоже мультииндекс.

$$e_{\mathbf{m}}(\mathbf{x}) = e^{i\mathbf{m} \cdot \mathbf{x}} = e^{im_1 x_1} e^{im_2 x_2} \dots e^{im_s x_s}.$$

Покажем, что $e_{\mathbf{m}}$ образуют ортогональный набор

$$\begin{aligned} (e_{\mathbf{m}}, e_{\mathbf{l}}) &= h^s \sum_{\mathbf{x} \in \mathcal{D}_h} e_{\mathbf{m}}(\mathbf{x}) \overline{e_{\mathbf{l}}(\mathbf{x})} = h^s \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} \dots \sum_{n_s=0}^{N-1} e^{im_1 n_1 h} e^{im_2 n_2 h} \dots e^{im_s n_s h} e^{-il_1 n_1 h} e^{-il_2 n_2 h} \dots e^{-il_s n_s h} = \\ &= h^s \sum_{n_1=0}^{N-1} \dots \sum_{n_s=0}^{N-1} e^{in_1(m_1-l_1)h} \dots e^{in_s(m_s-l_s)h} = h \sum_{n_1=0}^{N-1} e^{in_1(m_1-l_1)h} \dots h \sum_{n_s=0}^{N-1} e^{in_s(m_s-l_s)h}, \end{aligned}$$

то есть это всё сводится к произведению выражений, аналогичных таковым для одномерного дискретного преобразования Фурье. В итоге получаем, что эта сумма равна $(2\pi)^s$ в случае полного совпадения мультииндексов \mathbf{m} и \mathbf{l} , и нулю во всех остальных случаях.

Таким образом, мы показали, что эти функции ортогональны. Их всего N^s штук. Пространство V_h тоже размерности N^s , а значит они образуют базис, и по ним можно раскладывать другие функции

$$v_h(\mathbf{x}) = \sum_{\mathbf{m} \in M_0} a(\mathbf{m}) e_{\mathbf{m}}(\mathbf{x}).$$

Формула замкнутости:

$$(v_h(\mathbf{x}), w_h(\mathbf{x}))_{V_h} = (2\pi)^s \sum_{\mathbf{m} \in M_0} a(\mathbf{m}) \overline{b(\mathbf{m})}.$$

А если у нас ещё и $p > 1$, то про это вообще страшно думать. Но надо постараться. Пространство V_h по идее должно иметь размерность pN^s , и поэтому нам по идее нужно сконструировать из $e_{\mathbf{m}}$, которых было N^s штук, набор из pN^s векторов нехитрой схемой: расширением векторов и дополнением их нулями, смещая ненулевую часть в разные места. Тогда мультииндексы \mathbf{m} брались бы тоже из множества размерности pN^s , которое выглядело бы, наверное, примерно так: $M_0 = \bigsqcup_{j=1}^p \{0, \dots, N-1\}^s$.

Тогда предыдущие две формулы сохранили бы тот же самый вид, только буквы v_h и w_h стоило бы написать жирным начертанием. Также в этой формуле e стало бы жирным, а a осталось бы скалярной величиной. Однако, судя по конспектам, преподаватель поступил немного иначе. Разложение там выглядит так:

$$\mathbf{v}_h(\mathbf{x}) = \sum_{\mathbf{m} \in M_0} \mathbf{a}(\mathbf{m}) e_{\mathbf{m}}(\mathbf{x}),$$

отличие в том, что \mathbf{a} всё-таки стало вектором. Формула замкнутости принимает вид:

$$(\mathbf{v}_h(\mathbf{x}), \mathbf{w}_h(\mathbf{x}))_{V_h} = (2\pi)^s \sum_{\mathbf{m} \in M_0} (\mathbf{a}(\mathbf{m}) \overline{b(\mathbf{m})})_p.$$

Теперь применим метод Фурье к разностным уравнениям. Рассматриваем двухслойную схему с постоянными коэффициентами. Сеточное уравнение будет выглядеть так:

$$\sum_{\boldsymbol{\alpha} \in A_0} A(\boldsymbol{\alpha}) \mathbf{u}(\mathbf{x} + \boldsymbol{\alpha}h, k\tau) = \sum_{\boldsymbol{\beta} \in B_0} B(\boldsymbol{\beta}) \mathbf{u}(\mathbf{x} + \boldsymbol{\beta}h, (k+1)\tau).$$

Применим ДПФ. Пусть

$$\mathbf{u}(\mathbf{x}, k\tau) = \sum_{\mathbf{m} \in M_0} \mathbf{a}(\mathbf{m}, k) e^{i\mathbf{m} \cdot \mathbf{x}}.$$

Подставим это в суммы:

$$\sum_{\alpha \in A_0} e^{i\mathbf{m} \cdot \alpha h} \hat{A}(\alpha) \sum_{\mathbf{m} \in M_0} \mathbf{a}(\mathbf{m}, k) e^{i\mathbf{m} \cdot \mathbf{x}} = \sum_{\beta \in B_0} e^{i\mathbf{m} \cdot \beta h} \hat{B}(\beta) \sum_{\mathbf{m} \in M_0} \mathbf{a}(\mathbf{m}, k+1) e^{i\mathbf{m} \cdot \mathbf{x}}.$$

Слева и справа написаны два разложения по базису, коэффициенты в которых должны совпадать:

$$\mathbf{a}(\mathbf{m}, k) \sum_{\alpha \in A_0} e^{i\mathbf{m} \cdot \alpha h} \hat{A}(\alpha) = \mathbf{a}(\mathbf{m}, k+1) \sum_{\beta \in B_0} e^{i\mathbf{m} \cdot \beta h} \hat{B}(\beta).$$

В итоге получаем

$$\mathbf{a}(\mathbf{m}, k+1) = \hat{C}(\mathbf{m}) \mathbf{a}(\mathbf{m}, k), \quad \hat{C}(\mathbf{m}) = \left(\sum_{\beta \in B_0} e^{i\mathbf{m} \cdot \beta h} \hat{B}(\beta) \right)^{-1} \sum_{\alpha \in A_0} e^{i\mathbf{m} \cdot \alpha h} \hat{A}(\alpha).$$

$\hat{C}(\mathbf{m})$ – матрица перехода к следующему слою. Описывает переход в терминах коэффициентов Фурье.

44. Необходимое условие устойчивости по фон-Нейману.

Теорема (3). Для устойчивости при $f = 0$ необходимо и достаточно, чтобы

$$\|\hat{C}^k(\mathbf{m})\| \leq c_3 \quad \forall h, \tau, \mathbf{m} \quad (k\tau \leq T).$$

□ Достаточность:

$$\begin{aligned} \mathbf{a}(k, \mathbf{m}) &= \hat{C}^k \mathbf{a}(0, \mathbf{m}) = \hat{C}^k \mathbf{a}(\mathbf{m}), \\ \mathbf{u}_h(k) &= \sum_{\mathbf{m}} \mathbf{a}(k, \mathbf{m}) e^{i\mathbf{m} \cdot \mathbf{x}} = \sum_{\mathbf{m}} \hat{C}^k(\mathbf{m}) \mathbf{a}(\mathbf{m}) e^{i\mathbf{m} \cdot \mathbf{x}}. \end{aligned}$$

Применим формулу замкнутости

$$\begin{aligned} \|\mathbf{u}_h(k)\|^2 &= (2\pi)^s \sum_{\mathbf{m}} \|\hat{C}^k(\mathbf{m}) \mathbf{a}(\mathbf{m})\|^2 \leq \max_{\mathbf{m}} \|\hat{C}^k(\mathbf{m})\|^2 \cdot (2\pi)^s \cdot \sum_{\mathbf{m}} \|\mathbf{a}(\mathbf{m})\|^2 = \\ &= \max_{\mathbf{m}} \|\hat{C}^k(\mathbf{m})\|^2 \cdot \|\mathbf{u}_h(0)\|^2 \leq c_3^2 \cdot \|\mathbf{u}_h(0)\|^2 \Rightarrow \\ &\Rightarrow \|\mathbf{u}_h\|_{U_h} = \max_k \|\mathbf{u}_h(k)\|_{V_h} \leq c_3 \|\mathbf{u}_h(0)\|. \end{aligned}$$

Необходимость: Предположим, что устойчивость есть, но оценки для $\|\hat{C}^k(\mathbf{m})\|$ нет. Пусть $\mathbf{u}_h(0) = \mathbf{a} e^{i\mathbf{m}_0 \cdot \mathbf{x}}$, где \mathbf{a} – вектор с единичной нормой, а h_l, k_l и \mathbf{m}_l подобраны так, что $\|\hat{C}^{k_l}(\mathbf{m}_l)\| > l$. Тогда

$$\|\mathbf{u}_h\|_{U_h} = \max_k \|\mathbf{u}_h(k)\|_{V_h} \geq \|\mathbf{u}_h(k_l)\|_{V_h} = \|\hat{C}^{k_l}(\mathbf{m}_l)\| \cdot \|\mathbf{u}_h(0)\|_{V_h} > l$$

и это верно $\forall l$, значит никакой устойчивости нет, получили противоречие. Оно доказывает необходимость \blacksquare .

Теорема (4. Условие фон Неймана). Для устойчивости при $f = 0$ необходимо и достаточно, чтобы собственные числа матрицы перехода удовлетворяли условию

$$|\lambda_{\hat{C}}| \leq 1 + c_4 \tau \quad \forall h, \tau, \mathbf{m}.$$

□ Достаточность возникает только при $p = 1$. Тогда \hat{C} из матрицы превращается в число $c(\mathbf{m}) = \lambda_c$.

$$\|c^k\| = |\lambda_c|^k \leq (1 + c_4\tau)^k \leq e^{c_4\tau k} \leq e^{c_4T}.$$

Необходимость:

$$|\lambda_{\hat{C}^k(\mathbf{m})}| = |\lambda_{\hat{C}}|^k \leq c_3,$$

так как собственные числа не превосходят норму матрицы.

$$\ln |\lambda_{\hat{C}}| \leq \frac{c_3}{k} \quad \forall k \leq M \Rightarrow \ln |\lambda_{\hat{C}}| \leq \frac{c_3}{M} = c_3 \frac{\tau}{T}.$$

Пусть λ_{\max} – собственное число с максимальным модулем. Тогда³⁹

$$\ln |\lambda_{\hat{C}}| \geq \frac{\ln |\lambda_{\max}|}{|\lambda_{\max}| - 1} (|\lambda_{\hat{C}}| - 1) \Rightarrow |\lambda_{\hat{C}}| \leq 1 + \frac{\ln |\lambda_{\max}|}{|\lambda_{\max}| - 1} \ln |\lambda_{\hat{C}}| \leq 1 + \frac{\ln |\lambda_{\max}|}{|\lambda_{\max}| - 1} c_3 \frac{\tau}{T} = 1 + c_4\tau.$$

Теорема доказана ▣.

45. Простейшие схемы для уравнения бегущей волны.

Уравнение бегущей волны имеет вид

$$\frac{\partial u}{\partial t} = a \frac{\partial u}{\partial x},$$

любая функция вида

$$u(x, t) = f(x + at)$$

будет являться его решением. Пусть

$$u(x, 0) = \varphi(x) \Rightarrow \varphi(x + at) - \text{решение.}$$

Это бегущая со скоростью a волна. Считаем граничные условия периодическими. Запишем разностные уравнения:

$$\frac{u_n^{k+1} - u_n^k}{\tau} = a \frac{u_{n+1}^k - u_n^k}{h}.$$

Это обычная, явная схема. Чтобы исследовать устойчивость, найдём матрицу комплексный множитель перехода. Для этого подставим $u_n^k = e^{imx}$ и $u_n^{k+1} = c(m)e^{imx}$:

$$\frac{c(n)e^{imx} - e^{imx}}{\tau} = a \frac{e^{im(x+h)} - e^{imx}}{h}.$$

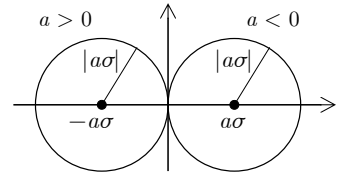
Отсюда легко найти

$$c(m) = 1 + a\sigma(e^{imh} - 1).$$

Нужно понять, когда $|c(m)| \leq 1$.

Утверждение. $|c| > 1$ при $a < 0$. Если $a > 0$ и $|a\sigma| \leq 1$, то $|c| \leq 1$.

□ e^{imh} пробегает единичную окружность. А $e^{imh} - 1$ – единичную окружность с центром в -1 . Умножение на $a\sigma$ либо просто растягивает (относительно нуля), либо растягивает и переворачивает. Когда $a < 0$, все точки полученной окружности, кроме 1, лежат вне единичного круга $e^{imh} - 1$, а это и есть неустойчивость. А если $a > 0$, то окружность будет лежать внутри единичного круга при $|a\sigma| \leq 1$. Прибавление единицы смещает всю картину вправо. Если полученная окружность лежит внутри единичного круга, то получаем $|c| \leq 1$ ▣.



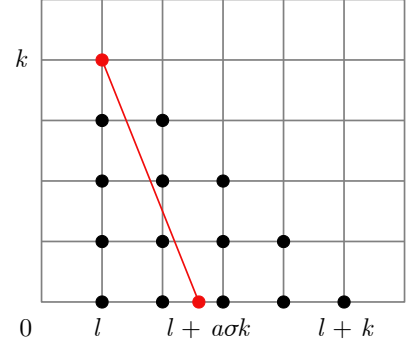
³⁹Это свойство легко можно увидеть, нарисовав графики $\ln(x + 1)$ и какой-нибудь прямой, проходящей через центр координат.

Заметим, что u постоянна на прямой $x + at = \text{const}$. Будем пока считать $a > 0$. Рассмотрим значение u_l^k . Оно должно определяться соответствующим значением на нулевом слое:

$$lh + ak\tau = sh \Rightarrow s = l + a\sigma k.$$

С другой стороны, в нашей схеме значение u_l^k определяется u_l^{k-1} и u_{l+1}^{k-1} . Если продолжить этот процесс до $k = 0$, увидим, что u_l^k зависит лишь от $u_l^0 \dots u_{l+k}^0$. Таким образом, чтобы вообще использовать нужное значение из начальных данных, надо

$$s \leq l + k \Rightarrow l + a\sigma k \leq l + k \Rightarrow a\sigma \leq 1.$$



Если $a < 0$, то мы вообще не будем использовать это значение, ибо красная прямая будет наклонена в другую сторону.

Для $a < 0$ работает схема

$$\frac{u_l^{k+1} - u_l^k}{\tau} = a \frac{u_l^k - u_{l-1}^k}{h}.$$

В ней информация распространяется в другую сторону, и оценки получаются те же с точностью до знака a .

46. Схема Куранта-Рисса.

Рассмотрим теперь систему

$$\frac{\partial \mathbf{u}}{\partial t} = \hat{A} \frac{\partial \mathbf{u}}{\partial x},$$

где \mathbf{u} – вектор из \mathbb{R}^p . Будем считать, что \hat{A} – симметричная матрица с постоянными коэффициентами (поэтому у неё все собственные числа вещественны). Собственные числа матрицы \hat{A} могут быть разных знаков, это приводит к появлению решений-волн, которые бегут в разные стороны. Это причина, по которой простейшие схемы будут неустойчивыми.

Рассмотрим две матрицы: \hat{A}_+ и \hat{A}_- . Первая из них имеет положительные собственные числа такие же, как у \hat{A} , а вместо отрицательных у неё нули. \hat{A}_- вместо отрицательных собственных чисел \hat{A} имеет их модули, а вместо положительных – нули. Если \mathbf{g} – собственный вектор \hat{A} , то $(\hat{A}_+ - \hat{A}_-)\mathbf{g} = \hat{A}\mathbf{g}$, то есть матрицы \hat{A} и $\hat{A}_+ - \hat{A}_-$ совпадают на собственных векторах \hat{A} . Поскольку собственные вектора образуют базис, то имеет место равенство $\hat{A} = \hat{A}_+ - \hat{A}_-$. Обе эти матрицы (\hat{A}_+ и \hat{A}_-) симметричные и неотрицательно-определённые. Схема будет устроена так:

$$\frac{\mathbf{u}_n^{k+1} - \mathbf{u}_n^k}{\tau} = \hat{A}_+ \frac{\mathbf{u}_{n+1}^k - \mathbf{u}_n^k}{h} - \hat{A}_- \frac{\mathbf{u}_n^k - \mathbf{u}_{n-1}^k}{h}.$$

Эти матрицы определены неоднозначно. Одно из представлений должно быть наиболее выгодным. Такая схема называется *схемой Куранта-Рисса*. Найдём матрицу перехода. Для этого вычислим её на

$$\mathbf{u}_n^k = \mathbf{f} e^{imx}, \text{ где } \mathbf{f} \text{ – произвольный постоянный вектор. } \mathbf{u}_n^{k+1} = \hat{C} \mathbf{f} e^{imx}.$$

Подставим это в схему и получим:

$$\frac{\hat{C} \mathbf{f} e^{imx} - \mathbf{f} e^{imx}}{\tau} = \hat{A}_+ \frac{\mathbf{f} e^{im(x+h)} - \mathbf{f} e^{imx}}{h} - \hat{A}_- \frac{\mathbf{f} e^{imx} - \mathbf{f} e^{im(x-h)}}{h}.$$

$$\hat{C} \mathbf{f} = \mathbf{f} + \sigma(e^{imh} - 1)\hat{A}_+ \mathbf{f} - \sigma(1 - e^{-imh})\hat{A}_- \mathbf{f}.$$

$$\hat{C} = \hat{E} + \sigma(e^{imh} - 1)\hat{A}_+ - \sigma(1 - e^{-imh})\hat{A}_-.$$

Так, слово за слово, мы нашли матрицу перехода. Теперь надо найти её собственные числа.

$$\mathbf{g} + \sigma(e^{imh} - 1)\hat{A}_+\mathbf{g} - \sigma(1 - e^{-imh})\hat{A}_-\mathbf{g} = \lambda_{\hat{C}}\mathbf{g}.$$

$$(e^{imh} - 1)\hat{A}_+\mathbf{g} + (e^{-imh} - 1)\hat{A}_-\mathbf{g} = \frac{\lambda_{\hat{C}} - 1}{\sigma}\mathbf{g}.$$

Слева стоит линейная комбинация матриц с известными собственными числами, причём там, где у первой ненулевое собственное число, у второй ноль, и наоборот. Поэтому получаем

$$\lambda_{\hat{C}} = 1 + \sigma\lambda_{\hat{A}_{\pm}}(e^{\pm imh} - 1).$$

Применим условие устойчивости из предыдущего билета. Вместо a у нас $\lambda_{\hat{A}_{\pm}}$ — всегда > 0 , а вместо m — $\pm m$. Знак m ни на что не повлияет. Получим $|\lambda_{\hat{C}}| \leq 1$ при $\sigma|\lambda_{\hat{A}}| \leq 1$ при всех $\lambda_{\hat{A}}$. В итоге, необходимое условие выглядит так:

$$\sigma \leq \frac{1}{\max|\lambda_{\hat{A}}|}.$$

В общем случае мы не доказывали достаточного условия, но здесь мы можем его доказать. Матрицы \hat{A}_+ и \hat{A}_- вещественны и симметричны,

$$\hat{A}_+\hat{A}_-\mathbf{g}_i = \lambda_{\hat{A}_+}\lambda_{\hat{A}_-}\mathbf{g}_i,$$

то есть произведения $\hat{A}_+\hat{A}_- = 0$ на всех собственных векторах матрицы \hat{A} . Матрица перехода \hat{C} является линейным многочленом от матриц \hat{A}_+ и \hat{A}_- , а значит она нормальна.

47. Явная схема для уравнения колебаний струны.

Рассмотрим одномерное уравнение колебаний

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}.$$

Вообще, надо было бы перейти к системе уравнений первого порядка, но мы будем искать множитель перехода λ , не переходя к ней. Схема получится такая:

$$\frac{u_l^{k+1} - 2u_l^k + u_l^{k-1}}{\tau^2} = \frac{u_{l+1}^k - 2u_l^k + u_{l-1}^k}{h^2}.$$

Положим

$$u_l^k = e^{imx}, \quad u_l^{k+1} = \lambda e^{imx}, \quad u_l^{k-1} = \lambda^{-1}e^{imx}.$$

После подстановки в схему, получим

$$(\lambda^{k+1} - 2\lambda^k + \lambda^{k-1})e^{imx} = \sigma^2\lambda^k(e^{im(x+h)} - 2e^{imx} + e^{im(x-h)}),$$

$$\lambda^2 - 2\lambda + 1 = \sigma^2\lambda(2\cos mh - 2),$$

$$\lambda^2 - 2 \underbrace{(1 + \sigma^2(\cos mh - 1))}_p \lambda + 1 = 0.$$

Решая, получим

$$\lambda = p \pm \sqrt{p^2 - 1}.$$

Неудивительно, что мы получили два значения. Это два собственных числа матрицы перехода \hat{C} , которая должна иметь размер 2×2 , поскольку система на самом деле второго порядка. Заметим, что по теореме Виета

$$\lambda_1\lambda_2 = 1,$$

то есть если $\lambda_1 < 1$, то $\lambda_2 > 1$. Нас устраивает ситуация, когда $|\lambda_1| = |\lambda_2| = 1$. Если корни вещественны, то это условие выполняется, когда $\lambda_1 = \lambda_2 = 1$ или $\lambda_1 = \lambda_2 = -1$. Для этого нужно $p = \pm 1$. Если корни комплексные, то они будут комплексно-сопряжёнными:

$$\lambda_{1,2} = p \pm i\sqrt{p^2 - 1}.$$

Модуль этого выражения равен 1 при $|p| < 1$. Вместе с вещественным случаем получаем общее условие

$$|p| \leq 1.$$

Из выражения для p видно, что оно выполнено при

$$\boxed{|\sigma| \leq 1}.$$

Получили необходимое условие устойчивости.

48. Явная и неявная схемы для двумерного уравнения теплопроводности.

Двумерное уравнение теплопроводности выглядит так:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Теперь нам понадобится два индекса внизу:

$$u_{np}^k \approx u(nh, ph, k\tau).$$

Рассмотрим явную схему

$$\frac{u_{np}^{k+1} - u_{np}^k}{\tau} = \frac{u_{n+1,p}^k - 2u_{np}^k + u_{n-1,p}^k}{h^2} + \frac{u_{n,p+1}^k - 2u_{np}^k + u_{n,p-1}^k}{h^2}.$$

Будем искать матрицу перехода. Для этого рассмотрим

$$u_{np}^k = e^{imx} e^{ily}, \quad x = nh, \quad y = ph.$$

После подстановки получим:

$$\frac{c-1}{\tau} e^{imx} e^{ily} = e^{ily} \frac{e^{im(x+h)} - 2e^{imx} + e^{im(x-h)} h^2}{h^2} + e^{imx} \frac{e^{il(y+h)} - 2e^{ily} + e^{il(y-h)} h^2}{h^2},$$

$$c(m, l) = 1 + 2\sigma(\cos mh - 1) + 2\sigma(\cos lh - 1), \quad \sigma = \frac{\tau}{h^2}.$$

Чтобы $|c(m, l)| \leq 1$ всегда, нужно

$$\boxed{\sigma \leq \frac{1}{4}}.$$

Получилось в два раза более жёсткое условие, чем для одномерного уравнения!

Теперь рассмотрим неявную схему. Для неё в уравнении для сетки поменяется левая часть:

$$\frac{u_{np}^k - u_{np}^{k-1}}{\tau} = \frac{u_{n+1,p}^k - 2u_{np}^k + u_{n-1,p}^k}{h^2} + \frac{u_{n,p+1}^k - 2u_{np}^k + u_{n,p-1}^k}{h^2}.$$

Аналогичными действиями находим

$$c(m, l) = \frac{1}{1 - 2\sigma(\cos mh - 1) - 2\sigma(\cos lh - 1)}.$$

Видно, что необходимое условие устойчивости выполнено всегда, как и в одномерном случае.

Теперь поговорим о том, как решать систему уравнений для неявной схемы. Если её переписать, выйдет

$$(1 + 4\sigma)v_{np} - \sigma v_{n+1,p} - \sigma v_{n,p+1} - \sigma v_{n-1,p} - \sigma v_{n,p-1} = \alpha_{np},$$

где

$$v_{np} = u_{np}^k, \quad \alpha_{np} = u_{np}^{k-1}.$$

Рассмотрим первую граничную задачу:

$$v_{00} = v_{0N} = v_{N0} = v_{NN} = 0.$$

Тогда n и p целесообразно считать изменяющимися в пределах от 1 до $N - 1$. Упорядочим элементы v_{ij} следующим образом:

$$v_{11}, v_{12}, \dots, v_{1,N-1}, v_{21}, \dots$$

Тогда матрица системы (для примера $N = 4$) будет выглядеть так:

$1 + 4\sigma$	$-\sigma$	0	$-\sigma$	0	0	0	0	0
$-\sigma$	$1 + 4\sigma$	$-\sigma$	0	$-\sigma$	0	0	0	0
0	$-\sigma$	$1 + 4\sigma$	0	0	$-\sigma$	0	0	0
$-\sigma$	0	0	$1 + 4\sigma$	$-\sigma$	0	$-\sigma$	0	0
0	$-\sigma$	0	$-\sigma$	$1 + 4\sigma$	$-\sigma$	0	$-\sigma$	0
0	0	$-\sigma$	0	$-\sigma$	$1 + 4\sigma$	0	0	$-\sigma$
0	0	0	$-\sigma$	0	0	$1 + 4\sigma$	$-\sigma$	0
0	0	0	0	$-\sigma$	0	$-\sigma$	$1 + 4\sigma$	$-\sigma$
0	0	0	0	0	$-\sigma$	0	$-\sigma$	$1 + 4\sigma$

В общем случае получается $(2N - 1)$ -диагональная матрица. Пусть

$$\mathbf{v}_n = (v_{n1}, \dots, v_{n,N-1}); \quad \boldsymbol{\alpha}_n = (\alpha_{n1}, \dots, \alpha_{n,N-1}).$$

Тогда можно записать систему, как

$$\hat{A}_n \mathbf{v}_{n-1} + \hat{B}_n \mathbf{v}_n + \hat{C}_n \mathbf{v}_{n+1} = \boldsymbol{\alpha}_n,$$

где \hat{A}_n , \hat{B}_n и \hat{C}_n – блоки из соответствующей строки. Дальше можно действовать так же, как обычной прогонкой. Это называется матричная прогонка.

К сожалению, обычная прогонка содержит умножения чисел, которые делаются за $O(1)$, и работает за $O(N)$, а матричная прогонка содержит умножения матриц, которые делаются за $O(N^3)$, и работает за $O(N^4)$. Долго! Поэтому нам такой неявный метод не подходит.

49. Схема продольно-поперечной прогонки.

Нужно сделать систему трёхдиагональной. Естественное желание – вынести часть переменных в правой части уравнения на соседний слой, чтобы сократить количество тех, что входит с текущего. Эта идея приводит к схеме

$$\frac{u_{np}^{k+1} - u_{np}^k}{\tau} = \frac{u_{n+1,p}^{k+1} - 2u_{np}^{k+1} + u_{n-1,p}^{k+1}}{h^2} + \frac{u_{n,p+1}^k - 2u_{np}^k + u_{n,p-1}^k}{h^2}.$$

Если переписать, получится

$$\sigma u_{n-1,p}^{k+1} - (1 + 2\sigma)u_{np}^{k+1} + \sigma u_{n+1,p}^{k+1} = \alpha_{np}.$$

Матрица трёхдиагональная, всё замечательно. Надо проверить на устойчивость. Теми же действиями, что в предыдущих билетах, получаем

$$c(l, m) = \frac{1 - 2\sigma(1 - \cos lh)}{1 + 2\sigma(1 - \cos mh)}.$$

Чтобы эта величина всегда была меньше или равна 1 по модулю, нужно

$$\boxed{\sigma \leq \frac{1}{2}}.$$

Только условная устойчивость! Можно рассмотреть аналогичную схему

$$\frac{u_{np}^{k+1} - u_{np}^k}{\tau} = \frac{u_{n+1,p}^k - 2u_{np}^k + u_{n-1,p}^k}{h^2} + \frac{u_{n,p+1}^{k+1} - 2u_{np}^{k+1} + u_{n,p-1}^{k+1}}{h^2}.$$

У неё свойства примерно такие же, только l и m меняются местами. Идея: чередовать схемы I и II:

$$2k \xrightarrow{I} 2k+1, \quad 2k+1 \xrightarrow{II} 2k+2.$$

Можем переобозначить номера слоёв. Пусть чётные слои имеют целые номера k , а нечётные – полуцелые $k + \frac{1}{2}$. Тогда переход $k \rightarrow k+1$ будет проходить сначала по I схеме, а потом по II:

$$\begin{aligned} \frac{u_{np}^{k+\frac{1}{2}} - u_{np}^k}{\tau/2} &= \frac{u_{n+1,p}^{k+\frac{1}{2}} - 2u_{np}^{k+\frac{1}{2}} + u_{n-1,p}^{k+\frac{1}{2}}}{h^2} + \frac{u_{n,p+1}^k - 2u_{np}^k + u_{n,p-1}^k}{h^2}, \\ \frac{u_{np}^{k+1} - u_{np}^{k+\frac{1}{2}}}{\tau/2} &= \frac{u_{n+1,p}^{k+\frac{1}{2}} - 2u_{np}^{k+\frac{1}{2}} + u_{n-1,p}^{k+\frac{1}{2}}}{h^2} + \frac{u_{n,p+1}^{k+\frac{1}{2}} - 2u_{np}^{k+\frac{1}{2}} + u_{n,p-1}^{k+\frac{1}{2}}}{h^2}. \end{aligned}$$

Коэффициенты перехода для каждой из схем:

$$c_I = \frac{1 - 2\sigma(1 - \cos lh)}{1 + 2\sigma(1 - \cos mh)}; \quad c_{II} = \frac{1 - 2\sigma(1 - \cos mh)}{1 + 2\sigma(1 - \cos lh)}$$

При последовательном применении схем I и II они должны перемножаться:

$$c = \frac{1 - 2\sigma(1 - \cos lh)}{1 + 2\sigma(1 - \cos mh)} \cdot \frac{1 - 2\sigma(1 - \cos mh)}{1 + 2\sigma(1 - \cos lh)}.$$

Тогда мы получаем, что $|c| \leq 1$ всегда. Наступает абсолютная устойчивость. Такую схему называют *схемой продольно-поперечной прогонки*.

К сожалению, если то же самое сделать для трёхмерного уравнения теплопроводности (а там будет три схемы и разбиение слоя на три подслоя), абсолютной устойчивости не выйдет. Это можно проверить прямым вычислением.

Вернёмся к двумерному уравнению. Уравнения для схем I и II в отдельности не аппроксимируют уравнение теплопроводности. А их совокупность – абсолютно устойчива и экономична. Существует так называемая *схема расщепления*, по которой продольно-поперечная прогонка распадается на уравнения по n или по p . Введём операторы:

$$L_1 u_{np} = \frac{u_{n+1,p} - 2u_{np} + u_{n-1,p}}{h^2}, \quad L_2 u_{np} = \frac{u_{n,p+1} - 2u_{np} + u_{n,p-1}}{h^2}.$$

$$\begin{aligned} u^{k+\frac{1}{2}} &= (I + \tau L_1) u^k, \quad u^{k+1} = (I + \tau L_2) u^{k+\frac{1}{2}} \Rightarrow \\ \Rightarrow u^{k+1} &= (I + \tau L_1)(I + \tau L_2) u^k = (I + \tau(L_1 + L_2) + \tau^2 L_2 L_1) u^k. \end{aligned}$$

$$\frac{\partial u}{\partial t} \leftarrow \frac{u^{k+1} - u^k}{\tau} = L_1 u^k + L_2 u^k + \tau L_2 L_1 u^k \rightarrow \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

$$L_2 L_1 u^k \rightarrow \frac{\partial^4 u}{\partial x^2 \partial y^2},$$

если четвёртая производная непрерывна (ввели дополнительное условие).

А. Введение в функциональный анализ.

А1. Пространства, отображения.

Бесконечномерные пространства во многом похожи на конечномерные, но есть и различия. Приведём наглядный пример:

Теорема. *В бесконечномерном пространстве с нормой единичный замкнутый шар не компактен.*⁴⁰

□ 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 ▮.

Определение. *Нормированное пространство* – векторное пространство с заданной на нём нормой.

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. *Для нормированных пространств непрерывность и ограниченность линейных отображений равносильны.*

В реальности многие (особенно определённые на всём пространстве) интересные отображения ограничены.

Определение. *Сепарабельное пространство* – топологическое пространство, для которого \exists всюду плотное счётное подмножество.

Определение. *Гильбертово пространство* – векторное пространство с введённым скалярным произведением, полное относительно метрики, порождённой скалярным произведением. Полнота по метрике означает, что любая фундаментальная последовательность сходится к элементу этого же пространства.

Для гильбертовых пространств сепарабельность \Leftrightarrow существованию в них счётного базиса. Будем считать все гильбертовы пространства сепарабельными.

А2. Пара фактов о гильбертовых пространствах.

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. *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. Это чем-то похоже на факт существования изоморфизма между обычным пространством и двойственным пространством в случае заданного скалярного произведения для конечномерных пространств.

⁴⁰Верно и обратное утверждение: если в нормированном пространстве единичный замкнутый шар компактен, то оно конечномерно.

Теорема (Рисса). 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.

А3. Спектр оператора.

Ещё одно различие, не столь наглядное, но очень важное, связано со спектром оператора.

Определение. 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. Говорят, что $\lambda \in \mathbb{C}$ – собственное число оператора A , если есть такой вектор $\mathbf{v} \in H$, что $A\mathbf{v} = \lambda\mathbf{v}$.

Собственные числа можно охарактеризовать в терминах оператора $A - \lambda$.

Утверждение. λ – собственное числа A тогда и только тогда, когда оператор $A - \lambda I$ неинъективен.

□ Пусть λ – собственное число A . Тогда $(A - \lambda I)\mathbf{v} = 0$. Но $(A - \lambda I)\mathbf{0}$ тоже равно нулю. Два разных вектора (\mathbf{v} и $\mathbf{0}$) имеют один образ, значит оператор неинъективен.

В обратную сторону: пусть $A - \lambda I$ неинъективен. Значит его ядро состоит из более чем одного элемента. Значит $\exists \mathbf{v} : (A - \lambda I)\mathbf{v} = 0 \Leftrightarrow A\mathbf{v} = \lambda\mathbf{v}$ ■.

Отсюда сразу следует такое утверждение:

Утверждение. Для конечномерных пространств спектр и множество собственных чисел – это одно и то же.

□ Как мы знаем,

$$\text{необратимость} \Leftrightarrow \text{несюръективность или неинъективность}.$$

Но в конечномерном случае

$$\text{несюръективность} \Rightarrow \text{неинъективность}.$$

Это связано с тем, что несюръективный оператор понижает размерность пространства, что вынуждает его склеивать вектора.

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

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

Что и требовалось доказать ■.

В бесконечномерном случае всё не так. Из необратимости неинъективность больше не следует, и у оператора появляются два разных способа быть необратимым:

1. Оператор склеивает векторы (он неинъективен).
2. Образ оператора меньше, чем всё пространство (он несюръективен).

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. 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 не лежат в $L^2([a, b])$ – свидетельство описанного феномена!

А4. Компактные операторы.

Обсудим один класс операторов, очень полезный на практике.

Определение. Пусть H – гильбертово пространство, B – единичный замкнутый шар в нём. Оператор $A : H \rightarrow H$ называют *компактным*, если замыкание множества $A(B)$ компактно.

Замечание. 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 конечномерности) свойств компактных операторов.

Утверждение. Если операторы A_n компактны и $\exists A : \|A_n - A\| \rightarrow 0$, то A компактен.

Следствие. Если операторы A_n конечного ранга (их образы конечномерны) и $\exists A : \|A_n - A\| \rightarrow 0$, то A компактен.

Главный пример компактного оператора – *интегральный оператор*:

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

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

где $K \in L^2([a, b]^2)$. Такой оператор называют *интегральным*, а функцию K – его *ядром*. В принципе, вместо L^2 можно жить в C – в пространстве непрерывных функций, но оно не гильбертово.

Утверждение. Интегральный оператор компактен.

□ *Note: это доказательство не строгое.* Разложим функцию K по базису (так можно, правда):

$$K(x, y) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} c_{nm} e_n(x) e_m(y).$$

Рассмотрим последовательность интегральных операторов A_N с ядрами

$$K_N(x, y) = \sum_{n=1}^N \sum_{m=1}^N c_{nm} e_n(x) e_m(y).$$

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 приближается операторами конечного ранга, а потому компактен ▣.

А5. Спектры компактных операторов.

Спектр компактного оператора обладает замечательным свойством:

Утверждение. Пусть A – компактный оператор. Тогда $\forall \delta > 0$ будет конечным множество таких собственных чисел A , что $|\lambda| \geq \delta$. Собственное пространство любого $\lambda \neq 0$ конечномерно.

Спектр произвольного самосопряжённого оператора, с другой стороны, обладает такими свойствами:

Утверждение. Собственные значения самосопряжённого оператора вещественны. Собственные векторы самосопряжённого оператора, соответствующие 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 (Гильберта-Шмидта). Пусть A – компактный и самосопряжённый оператор в гильбертовом пространстве H . \exists ортогональный базис $\{e_j\}$, состоящий из собственных векторов A .

А6. Альтернатива Фредгольма.

Определение. Оператор T на гильбертовом пространстве, равный $T = I - A$, где A компактен, называется *Фредгольмовым*.

Утверждение. Сопряжённый к компактному оператор компактен.

Теорема (Альтернатива Фредгольма).

1. Уравнение $T\varphi = f$ разрешимо $\Leftrightarrow f$ ортогонально любому решению $T^*\psi_0 = 0$.
2. Либо уравнение $T\varphi = f$ при любом f имеет только одно решение, либо уравнение $T\varphi_0 = 0$ имеет ненулевое решение.
3. Уравнения $T^*\psi_0 = 0$ и $T\varphi_0 = 0$ имеют одно и то же конечное число линейно-независимых решений.

Замечание. Поясним, как надо воспринимать эту теорему и почему она называется альтернативой. Представьте себе, что вы смотрите на уравнение $T\varphi = f$. Есть два варианта:

1. Уравнение $T\varphi_0 = 0$ не имеет ненулевых решений, и ваша задача однозначно разрешима. Прекрасно!
2. Оно их таки имеет, и всё не столь прекрасно.

Если вы попали во второй вариант, то 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.