О.В. МАТЫСИК

В.В. МОРОЗОВ

В.Ф. САВЧУК

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

Электронный курс лекций для студентов физико-математического факультета

Учреждение образования
«Брестский государственный университет имени А.С. Пушкина»

Кафедра прикладной математики и информатики

О.В. МАТЫСИК, В.В. МОРОЗОВ, В.Ф. САВЧУК

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

Электронный курс лекций для студентов физико-математического факультета

Брест
БрГУ имени А.С. Пушкина
2017

Рекомендовано Научно-методическим советом
Учреждения образования
«Брестский государственный университет имени А.С. Пушкина»

Составители:

О.В. Матысик – доцент кафедры прикладной математики и информатики Учреждения образования «Брестский государственный университет имени А.С. Пушкина», кандидат физико-математических наук, доцент

В.Ф. Савчук – доцент кафедры прикладной математики и информатики Учреждения образования «Брестский государственный университет имени А.С. Пушкина», кандидат физико-математических наук, доцент

В.В. Морозов – старший преподаватель кафедры прикладной математики и информатики Учреждения образования «Брестский государственный университет имени А.С. Пушкина»

Рецензенты:

Д.В. Грицук – доцент кафедры алгебры, геометрии и математического моделирования Учреждения образования «Брестский государственный университет имени А.С. Пушкина», кандидат физико-математических наук, доцент

А.И. Серый – доцент кафедры общей и теоретической физики Учреждения образования «Брестский государственный университет имени А.С. Пушкина», кандидат физико-математических наук

Электронный курс лекций «Вычислительные методы и компьютерное моделирование» состоит из глав, в которых представлены такие математические дисциплины как математическая логика, числительные методы и теория алгоритмизации. В пособии изучаются численные методы решения линейных и нелинейных алгебраических уравнений, задач векторной алгебры, математического анализа и статистики. Сформулированы условия сходимости итерационных процессов к корню уравнения с доказательством их необходимости и/или достаточности.

Издание курса лекций инициировано, с одной стороны, большим количеством не всегда доступных студентам источников, с другой – разнообразием терминологии изложения теорий, цитируемых из смежных разделов математики. Адресуется студентам физико-математических специальностей, изучающим численные и функциональные методы решения операторных уравнений.

О Г Л А В Л Е Н И Е

Предисловие

Настоящий электронный курс лекций предназначен для студентов стационара физико- математического факультета специальностей «Математика и информатика» и «Физика и информатика». Он составлен в соответствии с действующей типовыми программами дисциплины «Вычислительные методы и компьютерное моделирование» по этим специальностям, утвержденным министром образования РБ.

В электронном курсе лекций излагается теория по темам дисциплины «Вычислительные методы и компьютерное моделирование»: теория погрешностей, методы решения нелинейных уравнений и систем уравнений, нахождение собственных векторов и собственных значений матрицы, приближение функций, численное дифференцирование, обработка данных эксперимента, приближённое вычисление определённых интегралов, аналитические и численные методы решения задачи Коши для обыкновенных дифференциальных уравнений, методы решения задач линейного программирования.

На задачах различного уровня сложности проиллюстрированы основные методы, изложенные в электронном издании. Предлагаются задания для самостоятельной работы. Имеется блок самоконтроля.

Данный электронный курс лекций ставит своей целью обучить будущих специалистов вычислительным методам обработки информации и облегчить самостоятельную работу студентов с теоретическим материалом при подготовке к лабораторным занятиям и экзамену.

Основными задачами пособия являются:

- изучение числительных методов решения задач векторной алгебры, математического анализа и дифференциальных уравнений;
- обеспечение формирования умений и навыков работы с компьютерной техникой с использованием современных информационных технологий.

В результате освоения теоретического материала, изложенного в электронном курсе лекций «Вычислительные методы и компьютерное моделирование», студенты должны знать:

- основные числительные методы решения задач векторной алгебры, математического анализа и дифференциальных уравнений;
- понятия и предпосылки разработки программного обеспечения для описания физических процессов с помощью компьютерных моделей.

Предлагаемые в пособии примеры исследования и практические задания должны повысить профессиональную компетенцию студентов в области овладения методологии численного эксперимента и привить требуемые образовательным стандартом высшего образования умения и навыки:

- при работе с компьютерной и оргтехникой на уровне опытного пользователя;
- обработки информации с использованием современных информационных технологий;
- применения прикладных программных и компьютерных средств в учебной и научно-исследовательской работе.

Издание электронного курса лекций инициировано, с одной стороны, большим количеством не всегда доступных студентам источников, с другой – разнообразием терминологии изложения теорий, цитируемых из смежных разделов математической логики, числительных методов и теории алгоритмизации.

Глава 1 Основы теории множеств, математической логики и технологий программирования

1.1 Операции над множествами. Линейное пространство. Отображение множеств в уравнении

Понятие ***множество*** определим как совокупность, собрание элементов, объединенных по какому-либо признаку. Такая замена первичного понятия синонимами не претендует на строгое определение, а лишь позволяет абстрактный объект наполнить реальным содержанием. Обозначать множества будем прописными буквами ***A, B, C, …***, а элементы множеств – малыми ***a, b, c, ….***

Множество, которое не содержит ни одного элемента, будем называть ***пустым множеством*** и обозначать символом $\varnothing$.

Утверждение «элемент ***a принадлежит*** множеству ***A***» можно записать с помощью символов $a \in A$. Запись $a \notin A$ означает, что «элемент ***a не принадлежит*** множеству ***A***».

Если все элементы множества ***A*** являются элементами множества ***B,*** то будем говорить, что «***A подмножество B***» и писать $A \subset B$. Любое множество содержит $\varnothing$ в качестве подмножества.

Определение 1.1. ***Объединением*** множеств ***A*** и ***B*** называется множество ***C***, состоящее из элементов, принадлежащих хотя бы одному из множеств ***A*** или ***B***. Это же можно записать с помощью символов

$$C = \{x : x \in A \text{ или } x \in B\} \equiv A \cup B.$$

Определение 1.2. ***Пересечением*** множеств ***A*** и ***B*** называется множество ***C***, состоящее из элементов, принадлежащих множествам ***A*** и ***B*** одновременно. Перепишем определение в символьном виде

$$C = \{x : x \in A \text{ и } x \in B\} \equiv A \cap B.$$

Операции объединения и пересечения множеств по своему определению коммутативны и ассоциативны

$$A \cup B = B \cup A,\ (A \cup B) \cup C = A \cup (B \cup C);$$

$$A \cap B = B \cap A,\ (A \cap B) \cap C = A \cap (B \cap C).$$

Кроме того, они взаимно дистрибутивны

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C);$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C).$$

Определение 1.3. ***Разностью*** множеств A и B назовем множество C, если оно состоит из тех элементов A, которые не принадлежат B, то есть

$$C = \{x : x \in A \text{ и } x \notin B\} \equiv A \setminus B.$$

В теории меры (раздел 3.2) будет использована симметрическая разность множеств A и B, которая обозначается $A \Delta B$.

Определение 1.4. Назовем ***симметрической разностью*** множеств A и B множество C, которое находится по формуле

$$C = (A \setminus B) \cup (B \setminus A) \equiv A \Delta B.$$

Непосредственно из определения следует, что

$$A \Delta B = (A \cup B) \setminus (B \cap A).$$

Определение 1.5. ***Дополнением*** множества $A \subset B$ до множества B называется разность $B \setminus A$, которая обозначается B_A.

В теории множеств и ее приложениях важную роль играет принцип двойственности, основанный на двух высказываниях:

- дополнение объединения множеств равно пересечению дополнений;
- дополнение пересечения множеств равно объединению дополнений.

Принцип двойственности состоит в том, что из любого равенства, относящегося к системе подмножеств множества X (см., например, раздел 3.2), может быть получено другое (двойственное) равенство путем замены всех рассматриваемых множеств их дополнениями, объединений – пересечениями, а пересечений – объединениями.

Одним из основных в математике является понятие линейного пространства. Термин «пространство» подчеркивает уникальность свойств данного множества. Наиболее значимые множества и пространства будем обозначать курсивом $A, B, \ldots$.

Определение 1.6. Множество L называется ***линейным*** (или ***векторным***) пространством над полем действительных чисел, если любые элементы этого множества удовлетворяют следующим двум постулатам.

I. Для всех элементов u и v из множества L однозначно определен элемент $s \in L$, который называется их ***суммой***. Операция сложения обозначается «+», а действие суммирования записывается как $u + v = s$, причем:

1) $u + v = v + u$ (коммутативность);

2) $u + (v + w) = (u + v) + w$, $\forall w \in L$ (ассоциативность);

3) в L существует такой элемент, обозначаемый 0 , что для любого $u \in L$ выполняется равенство $u + 0 = u$ (существование нуля);

4) для каждого $u \in L$ в L существует элемент, обозначаемый $-u$, такой, что $u + (-u) = 0$ (существование противоположного элемента).

II. Для любого действительного числа α и всех элементов $u \in L$ определен элемент $\alpha \cdot u \in L$ (***произведение*** элемента u на число α), причем:

1) $\alpha \cdot (\beta u) = (\alpha\beta)\, u, \ \forall \beta \in R$ (ассоциативность);

2) $1 \cdot u = u$ (существование единицы);

3) $(\alpha + \beta)\, u = \alpha u + \beta u$ (дистрибутивность чисел);

4) $\alpha \cdot (u + v) = \alpha u + \alpha v, \ \forall v \in L$ (дистрибутивность элементов).

Элементы L, удовлетворяющие аксиомам I и II, называются ***векторами***. Отсюда второе название-синоним этого пространства – *векторное*.

Приведем примеры линейных (векторных) пространств, которые будут рассматриваться в дальнейшем изложении:

- пространство R действительных чисел (множество точек на прямой) с обычными арифметическими операциями сложения и умножения;
- пространство последовательностей l_2, элементами которого являются бесконечные множества действительных чисел с суммируемым квадратом

$$u = (u_0, u_1, \ldots, u_n, \ldots)^T, \ \sum_{n=0}^{\infty} u_n^2 < \infty$$

и следующими операциями сложения элементов и умножения их на число

$$(u_0, \ldots, u_n, \ldots) + (v_0, \ldots, v_n, \ldots) = (u_0 + v_0, \ldots, u_n + v_n, \ldots),$$

$$\alpha \cdot (u_0, u_1, \ldots, u_n, \ldots) = (\alpha \cdot u_0, \alpha \cdot u_1, \ldots, \alpha \cdot u_n, \ldots)$$

(подразумеваемое транспонирование означает: если l_2 порождено числом x оси абсцисс OX координатной плоскости XOY, то все действительные значения u_n, $n = 0, 1, \ldots$ компонент $\forall u \in l_{2(x)}$ принадлежат оси ординат OY, а точки множества $\{(x, u_n), n = 0, 1, \ldots\}$ лежат на прямой $\{(x, y) : y \in R\}$);

- функциональное пространство $C_{[a, b]}$, состоящее из непрерывных на отрезке $[a, b]$ действительных функций, с обычными операциями сложения функций и умножения их на число

$$\overline{u + v}\,(x) = u(x) + v(x), \ \overline{k\,u}\,(x) = k \cdot u(x), \ x \in [a, b].$$

Определение 1.7. Множество, содержащее вместе с системой элементов $S = \{u, v, w, \ldots\}$ их произвольную линейную комбинацию с коэффициентами из R, назовем ***линейным многообразием*** системы S.

Пересечение всех линейных многообразий системы S называется ***оболочкой S***.

Определение 1.8. Элементы $u_1, u_2, \ldots, u_n$ линейного пространства L называются ***линейно зависимыми***, если существуют такие действительные числа $\alpha_1, \alpha_2, \ldots, \alpha_n$, не все равные 0, что

$$\alpha_1 \cdot u_1 + \alpha_2 \cdot u_2 + \ldots + \alpha_n \cdot u_n = 0.$$

Если же линейная комбинация равна 0 тогда и только тогда, когда $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$, то эти элементы называются ***линейно независимыми***.

Определение 1.9. Бесконечная система S (множество $\{u, v, \ldots, w, \ldots\}$) элементов линейного пространства L называется ***линейно независимой***, если любая ее конечная подсистема (подмножество) линейно независима.

Линейно независимая система Γ пространства L называется алгебраическим ***базисом*** (***Гамеля***) ***пространства L***, если ее линейная оболочка совпадает с L.

Определение 1.10. Если в линейном пространстве L найдется n линейно независимых элементов, а любые $(n + 1)$ элементов этого пространства линейно зависимы, то говорят, что пространство L имеет ***размерность n***, а n линейно независимых элементов образуют ***базис L***. Если же в L можно указать систему из произвольного конечного числа линейно независимых элементов, то говорят, что пространство L ***бесконечномерное***.

Из всего многообразия введения понятия функции (или отображения) на множествах заслуживает предпочтения следующая формулировка.

Определение 1.11. Пусть X и Y – два произвольных множества. Говорят, что на X определена ***функция*** F, принимающая значения из Y, если для любого x, принадлежащего X, поставлен в соответствие один и только один элемент y, принадлежащий Y.

Для множеств произвольной природы вместо термина «функция» часто пользуются термином «отображение», говоря об ***отображении*** одного множества в другое. В дальнейшем термин «функция» будет употребляться только как элемент функционального пространства.

Для обозначения отображения F из U в V используем запись

$$F : U \to V.$$

Определение 1.12. Если u – элемент из U, то соответствующий ему элемент $v = F(u)$ из V называется ***образом u*** (при отображении F).

Совокупность O всех элементов $u \in U$, образом которых является заданный элемент $v \in V$, называется множеством ***прообразов*** элемента v и обозначается $O \equiv \{F^{-1}(v)\}$ или без фигурных скобок $F^{-1}(v)$.

Закон, с помощью которого описывается отображение $\boldsymbol{F}$, должен формулироваться так, чтобы для $\forall \boldsymbol{u} \in \boldsymbol{U}$ было возможно определить (найти, вычислить и т.п.) единственный соответствующий ему элемент $\boldsymbol{v}$, принадлежащий $\boldsymbol{V}$. К сожалению, описать закон, по которому находится множество $\boldsymbol{O}$ прообразов данного элемента $\boldsymbol{v} \in \boldsymbol{V}$, удается не всегда.

Определение 1.13. Если элемент $\boldsymbol{v}$ принадлежит множеству $\boldsymbol{V}$, а совокупность элементов $\boldsymbol{u} \in \boldsymbol{O} \subset \boldsymbol{U}$ находится из условия

$$\boldsymbol{F(u)} = \boldsymbol{v}, \text{ где } \boldsymbol{F : U \to V}, \qquad (1.1)$$

то выражение (1.1) будем называть ***уравнением*** относительно искомой величины $\boldsymbol{u}$. Множество $\boldsymbol{O}$ прообразов элемента $\boldsymbol{v}$ назовем ***множеством решений (корней) уравнения*** (1.1).

Все дальнейшие исследования будут посвящены описанию и доказательству законов (методов), при помощи которых находится (вычисляется, определяется) хотя бы один элемент $\boldsymbol{u}$ из непустого множества $\boldsymbol{O}$.

В тех случаях, когда существование корня уравнения (1.1) доказано, но найти его в явном виде не представляется возможным, будем ограничиваться поиском приближенного решения $\tilde{\boldsymbol{u}}$, локализующего единственный корень $\boldsymbol{u^*}$ уравнения в некоторой окрестности $\boldsymbol{Q} \subset \boldsymbol{U}$ приближения $\tilde{\boldsymbol{u}}$.

Чаще всего это будут процессы поиска более точного приближенного решения, основанные на использовании предыдущих приближений и свойств отображения $\boldsymbol{F : U \to V}$. Генерируемые таким образом процессы решения уравнения (1.1) будем называть итерационными.

1.2 Элементы математической логики. Булево исчисление высказываний и предикатов

Теоретическое доказательство сходимости итерационного процесса к корню уравнения (1.1) является словесной или символьной *логической схемой* решения задачи. В математической логике все множество высказываний (первичное понятие предмета) делится на два непересекающихся класса. Один класс содержит лишь истинные высказывания (ИСТИНА), другой – только ложные (ЛОЖЬ). Переменную-высказывание, принимающую (в данный момент времени) одно из этих двух взаимодополняющих друг друга значений, назовем ***логической***. В зависимости от выбранной смысловой системы значения логических переменных могут быть следующими:

ДА ÷ НЕТ в системе речевого общения;

$1 \div 0$ в двоичной (числовой) СС;

T (TRUE) ÷ F (FALSE) в системе программирования и т.д.

Определение 1.14. ***Логическим умножением*** (***конъюнкцией***) переменных ***A*** и ***B*** называется операция (будем обозначать ее «***и***» или «$\wedge$»), удовлетворяющая следующим условиям:

A	***и***	***B***	=	***C***
1	$\wedge$	***1***	=	***1***
1	$\wedge$	***0***	=	***0***
0	$\wedge$	***1***	=	***0***
0	$\wedge$	***0***	=	***0***

Определение 1.15. ***Логическим сложением*** (***дизъюнкцией***) переменных ***A*** и ***B*** называется операция (будем обозначать ее «***или***» или «$\vee$»), удовлетворяющая следующим условиям:

A	***или***	***B***	=	***C***
1	$\vee$	***1***	=	***1***
1	$\vee$	***0***	=	***1***
0	$\vee$	***1***	=	***1***
0	$\vee$	***0***	=	***0***

Отметим, что введенные выше операции ассоциируются как с арифметическими действиями «×» и «+», так и с операциями над множествами «$\cap$» и «$\cup$».

Определение 1.16. Логическим ***следованием*** (***импликацией***) переменных ***A*** и ***B*** называется операция (будем обозначать ее «***влечет***» или «$\Rightarrow$»), удовлетворяющая следующим условиям:

A	***влечет***	***B***	=	***C***
1	$\Rightarrow$	***1***	=	***1***
1	$\Rightarrow$	***0***	=	***0***
0	$\Rightarrow$	***1***	=	***1***
0	$\Rightarrow$	***0***	=	***1***

Определение 1.17. ***Логической эквивалентностью*** (***равнозначностью***) переменных ***A*** и ***B*** называется операция (будем обозначать ее «***равносильно***» или «$\Leftrightarrow$»), удовлетворяющая следующим условиям:

A	***равносильно***	***B***	=	***C***
1	$\Leftrightarrow$	***1***	=	***1***
1	$\Leftrightarrow$	***0***	=	***0***
0	$\Leftrightarrow$	***1***	=	***0***
0	$\Leftrightarrow$	***0***	=	***1***

Определение 1.18. ***Логическим отрицанием*** (***дополнением***) переменной ***A*** называется операция (будем обозначать ее «***не***» или надчеркиванием «$\overline{\ldots}$»), удовлетворяющая следующим условиям:

не	***A***	**=**	***C***
не	***1***	**=**	***0***
не	***0***	**=**	***1***

Логические выражения (ЛВ), представляющие логические константы, переменные, обращение к логическим функциям, бинарные отношения сравнения и т.п., назовем *простыми*. С помощью логических операций и скобок формируются *сложные* ЛВ. Логические выражения, обособленные логическими скобками или являющиеся аргументами логических операций и функций, назовем *операндами* выражений, операций или функций.

Рассмотрим основные формулы, связывающие логические операции и выражения. В программировании нередко используется отрицание выражений. Найдем отрицание конъюнкции и дизъюнкции двух высказываний.

Теорема 1.1. Для логических выражений ***A*** и ***B*** справедлива формула

$$\overline{\boldsymbol{A}\ \textit{и}\ \boldsymbol{B}} = \overline{\boldsymbol{A}}\ \textit{или}\ \overline{\boldsymbol{B}}. \qquad (1.2)$$

Доказательство теоремы основывается на рассмотрении всех возможных вариантов значений логических переменных ***A*** и ***B***:

не	***(A***	***и***	***B)***	**=**	***не (A и B)***
не	***(1***	$\wedge$	***1)***	**=**	***0***
не	***(1***	$\wedge$	***0)***	**=**	***1***
не	***(0***	$\wedge$	***1)***	**=**	***1***
не	***(0***	$\wedge$	***0)***	**=**	***1***

(не	***A)***	***или***	***(не***	***B)***	**=**	***(не A) или (не B)***
(не	***1)***	$\vee$	***(не***	***1)***	**=**	***0***
(не	***1)***	$\vee$	***(не***	***0)***	**=**	***1***
(не	***0)***	$\vee$	***(не***	***1)***	**=**	***1***
(не	***0)***	$\vee$	***(не***	***0)***	**=**	***1***

Из равенства значений левой и правой частей формулы (1.2) при всех возможных значениях ***A*** и ***B*** следует истинность теоремы 1.1. 🗗

Теорема 1.2. Для логических выражений ***A*** и ***B*** справедлива формула

$$\overline{\boldsymbol{A}\ \textit{или}\ \boldsymbol{B}} = \overline{\boldsymbol{A}}\ \textit{и}\ \overline{\boldsymbol{B}}. \qquad (1.3)$$

Формулы (1.2) и (1.3) называются законами Моргана.

Упражнение. Применяя схему доказательства теоремы 1.1, докажите теорему 1.2 и истинность используемых в дальнейшем формул:

$$A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C) \quad (1.4)$$

$$A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C) \quad (1.5)$$

$$A \Leftrightarrow B = (A \Rightarrow B) \wedge (B \Rightarrow A) \quad (1.6)$$

Обозначим кванторы общности и существования символами $\forall$ и $\exists$ соответственно, а функциональные предикаты – как $F(u)$ и $F(u_1, \ldots, u_n)$. Тогда если $x \in \{a_1, \ldots, a_n\}$, то справедливы следующие равенства

$$\exists x\, F(x) = F(a_1) \vee \ldots \vee F(a_n) \quad (1.7)$$

$$\forall x\, F(x) = F(a_1) \wedge \ldots \wedge F(a_n) \quad (1.8)$$

$$\overline{\forall x\, F(x)} = \exists x\, \overline{F(x)} \quad (1.9)$$

$$\overline{\exists x\, F(x)} = \forall x\, \overline{F(x)} \quad (1.10)$$

1.3 Введение в программирование. Графические схемы логических конструкций алгоритма

Применяемый здесь как базовый алгоритмический язык программирования Pascal (Паскаль) был создан в конце $60^{-ых}$ годов XX века швейцарским математиком Н. Виртом. Язык предназначался для обучения студентов искусству программирования. Международный стандарт языка Pascal, который

- *строился на минимальном количестве базовых понятий,*
- *имел несложный синтаксис,*
- *осуществлял перевод программ в машинный код простым и быстрым компилятором*, был утвержден в 1982 году.

Сейчас в сфере среднего и высшего образования в качестве первого языка программирования используется ***Turbo Pascal***, а также визуальная версия этого языка фирмы Borland – ***Delphi***. Применяемый в Delphi язык *Object Pascal* сохранил основные черты Turbo Pascal, обогатившись новыми возможностями быстрой разработки приложений для *OS Windows*. Язык ***BPW*** будем использовать как для поиска корней математических уравнений, так и для графической их интерпретации.

Для решения сложных математических задач предназначены: система компьютерной алгебры и математического моделирования ***Maple*** и интегрированная среда программирования ***Fortran***, современные версии которых включают библиотеки из нескольких тысяч научных подпрограмм и совместимы со многими математическими пакетами.

Конечную последовательность команд пользователя, достаточную для конструктивного анализа математической модели и абстрактного решения соответствующего операторного уравнения, назовем *алгоритмом*.

Логически упорядоченную последовательность команд (предложений) языка программирования интегрированной среды, направленную на исполнение алгоритма, назовем *программой* (*кодом*).

В узком смысле язык программирования – это интегрированная среда (ИС), которая представляет собой комплексную программу, имеющую встроенный редактор текстов, подсистему работы с файлами, систему справочной информации (Help-систему), встроенный отладчик, подсистему управления компиляцией и редактированием связей и т.п.

Текстовый редактор осуществляет предпроцессорную подготовку исходного текста программы к компиляции. Затем отредактированный исходный модуль обрабатывается *компилятором*, который по мере возможности устраняет синтаксические и логические ошибки в тексте программы.

С помощью *отладчика* можно выполнить программу в пошаговом режиме, контролируя значения любых ее переменных. Задав в программе контрольные точки, можно, исходя из промежуточных значений той или иной переменной, вносить коррективы в процесс реализации алгоритма.

Компоновщик дополняет программу необходимыми встроенными подпрограммами из библиотеки объектных модулей и помещает ее в файл с расширением ***exe***, который в дальнейшем исполняется.

В широком смысле понятие «язык программирования» используется как ретранслятор алгоритма, разработанного пользователем, в текст программы, состоящий из слов и предложений, составленных с помощью символов и идентификаторов, операндов и операций языка программирования. Предложения программы могут быть законченными (в виде операторов) или незаконченными (в виде простых и сложных логических выражений).

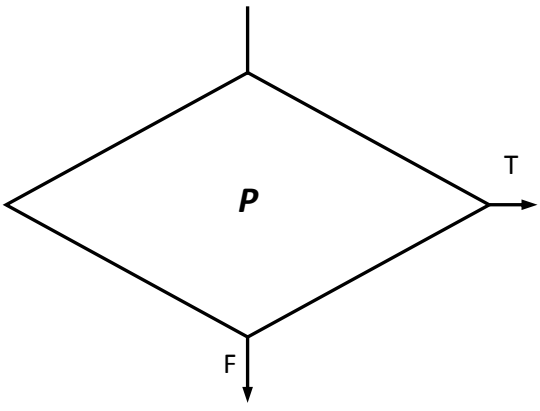


Рисунок 1

В программировании широко используется фундаментальный принцип управления сложными динамическими процессами «декомпозиция или структуризация алгоритма (библиотека модулей) – систематизация или иерархия классов (объектно-ориентированная среда)». Ознакомимся с методами детализации алгоритма и изучим возможности структурированного программирования.

Чтобы определить надлежащее направление движения алгоритма из множества возможных его продолжений, будем использовать графическое описание алгоритма. Элементами визуальных схем являются блоки, линии и стрелки. Основным блоком графических схем является Блок условного выбора ветви алгоритма (рисунок 1), предназначенный для вычисления значений алгебраических и/или логических выражений с последующим принятием решения в зависимости от истины (T) или лжи (F) предиката ***P***.

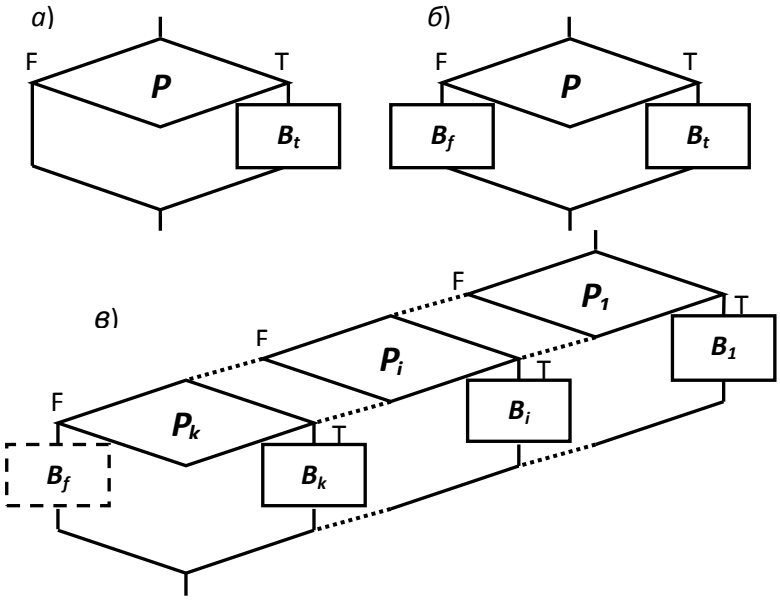


Рисунок 2

В Паскале применяются шесть логических конструкций (ЛК): три условные конструкции (УК) и три конструкции повторений (КП), инструкции (арифметико-логические блоки) которых, исполняемые при заданных условиях, заключают в операторные скобки «begin» и «end». В некоторых языках общее число ЛК сокращено до двух или трех, однако это либо усложняет разработку алгоритма программы, либо увеличивает время ее выполнения.

При описании алгоритмов все логические конструкции будем изображать в виде шестиугольников. Отметим, что в структурированном программировании эти конструкции всегда имеют одну точку входа и одну точку выхода. На рисунке 2 показаны условные конструкции «примыкание» (*а*) и «развилка» (*б*), осуществляющие простое ветвление алгоритма.

Опишем работу условной конструкции «выбор», представленной в виде графического объекта (*в*). Если условие $P_i \equiv$ «$\rho \in p_i$» выполнено при одном из $i = 1, \ldots, k$, то объект исполняет блок B_i. Иначе выполняется блок B_f, а при его отсутствии в объекте (по умолчанию) условная конструкция «выбор» пропускается.

Определим предикат P_i. Объединение всевозможных значений индикатора ρ обозначим P, а совокупность $p_i = \bigcup_s \rho_s$ (фаза слияния ветвей) попарно непересекающихся $p_i \cap p_j, i \neq j$ (фаза ветвления нелинейного алгоритма) подмножеств P обозначим S. Тогда для всех $\rho \in p_i$ осуществляется выбор одной корпоративной ветви алгоритма, соответствующей $P_i = True$.

Каждая условная конструкция имеет операторный аналог в Паскале:

Примыкание (*а*) – If **P** Then
begin
Bt
end;

Развилка (*б*) – If **P** Then
begin
Bt
end
Else
begin
Bf
end;

Выбор (*в*) – Case **ρ** Of
P_1: begin **B_1** end;
…
P_n: begin **B_n** end;
{Else begin **B_f** end;}
End;

При решении математических задач часто используются повторения некоторых фрагментов алгоритма. Причем КП (циклы) могут содержать не только УК, но и другие КП, что значительно усложняет разработку алгоритмов решения задач. Дополним множество графических объектов условных конструкций графическими объектами конструкций повторений.

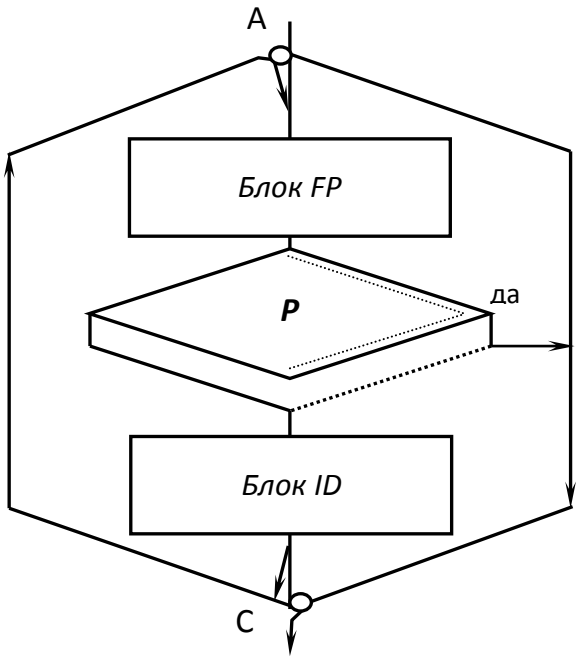


Рисунок 3

В качестве базовой используем циклическую конструкцию, изображенную на рисунке 3. Здесь A – точка входа в КП; блок ***FP*** формирует условие ***P*** выхода из цикла, то есть «если $P = True$, то выполняется оператор следующий за КП»; блок ***ID*** осуществляет изменение данных для блока ***FP*** очередного шага цикла; C – точка выхода из КП.

Основная цель циклического процесса – определение параметров, при которых предикат ***P*** принимает истинное значение. Поэтому во избежание зацикливания алгоритма некоторые (а может быть и все) переменные, входящие в логическое выражение ***P***, должны изменять свои значения в блоках ***FP*** и/или ***ID*** на каждом шаге повторений.

К сожалению, не всегда в языках программирования высокого уровня присутствует оптимальная базовая КП. Чаще ее представляют как цикл с предусловием (отсутствует блок ***FP***) или цикл с постусловием (отсутствует блок ***ID***). При черчении этих объектов в алгоритмах не будем применять стрелки, если вертикальная проекция движения алгоритма обращена вниз. В объектах повторений используем четвертую вершину блока принятия решения, служащую для возврата к точке входа в КП (без слияния с основным алгоритмом) или, наоборот, для выхода из КП (без слияния с блоком принятия решения). Если нет необходимости указывать направление движения алгоритма из блока слияния, то этот блок опускаем.

Из логических соображений следует, что блок ***FP*** в циклах с предусловием или постусловием должен быть сформирован до операторов ЦПред <условие> или ЦПост <условие>. Запишем данные циклы с помощью принятых сокращений (в скобках – изменения базового цикла):

```
НЦ
ЦПред <условие>
   НИн
   Блок ID
   (Блок FP)
   КИн
КЦ
```

```
НЦ
   НИн
   Блок FP
ЦОпт <условие>
   Блок ID
   КИн
КЦ
```

```
НЦ
   НИн
   Блок FP
   (Блок ID)
   КИн
ЦПост <условие>
КЦ
```

Если в *инструкции* логической конструкции только один оператор, то обособляющие *ее* операторные скобки можно опустить.

Определение 1.19. Если для формирования предиката ***P*** блока принятия решения конструкцию повторений (инструкцию тела цикла) требуется выполнить ***m*** раз, то такой циклический процесс назовем ***m****-ступенчатым*.

Изобразим изучаемые циклы в виде графических объект-схем. Так как Блоку ***P*** принятия решения предшествует блок ***FP***, то для применения цикла с предусловием (*а*) необходимо, в отличие от оптимального, предварительно выполнить Блок ***FP***. Существует также принципиальное различие в исполнении оптимальной конструкции и цикла с постусловием. Это связано с внедрением Блока ***ID*** (формирование предиката на основании $\boldsymbol{m-1}$ ступени повторений) из оптимального цикла в неоптимальный (*б*). Отметим, что до начала работы цикла с постусловием обязательно исполняется Блок ***ID***, иначе циклический процесс – бесступенчатый (обычный).

Для удобства использования циклов с предусловием и постусловием создадим единую логическую конструкцию объектов повторений. Во-первых, содержание Блока ***FP*** оформим в виде совокупности алгебраических и/или булевых функций так, чтобы их можно было напрямую применить в предикате ***P***. Блок принятия решения с таким ЛВ обозначим Блок ***Py*** (универсальный). Во-вторых, все объекты повторений изобразим в виде шестиугольников (рисунок 4), где левый контур объекта предназначен для возврата к началу цикла без слияния с основным алгоритмом.

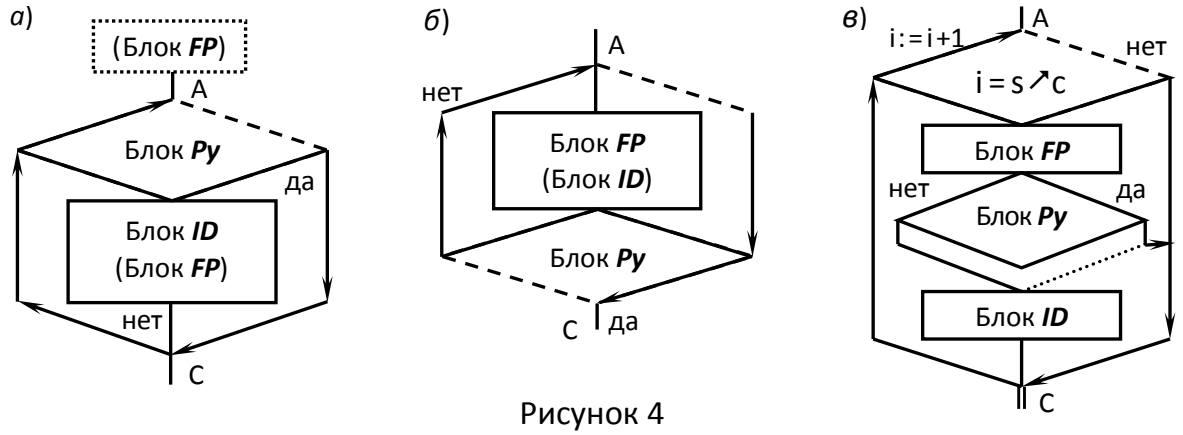


Рисунок 4

Штриховые звенья не являются ветвями алгоритма и предназначены для целостности восприятия объекта. Далее они будут изображены сплошными линиями. Правая часть объекта повторений завершает работу цикла по значению предиката ***P*** (*а*, *в*) или принудительно (*а*, *б*, *в*). Стрелка, направленная к левой границе объекта, означает переход на следующую итерацию (Continue), а стрелка, направленная к правой границе, – окончание цикла (Break). Выход из конструкции – точка *C*, где в случае (*а*) и (*в*) нет слияния ветвей, а в случае (*б*) левое нижнее звено формально.

В КП со счетчиком (рисунок 4, *в*) примем по умолчанию изменение значения счетчика с шагом ***1***. Этот цикл во многом идентичен оптимальному, так как в его инструкции находится блок принятия решений. Недостаток цикла-счетчика (если количество повторений заранее неизвестно) – два условия выхода, требующие дальнейшего распознавания. В языках низкого уровня данная проблема решается с помощью условных и безусловных переходов к меткам. В структурированных языках такие переходы осуществляют условные операторы ветвления и выбора.

Каждая КП с инструкцией ***B*** имеет операторный аналог в Паскале:

Предусловие (*а*) – While not(***P***) Do begin ***B*** end;

Постусловие (*б*) – Repeat ***B*** Until ***P***;

Счетчик (*в*) – For <счетчик> := ***s*** To ***c*** Do begin ***B*** end.

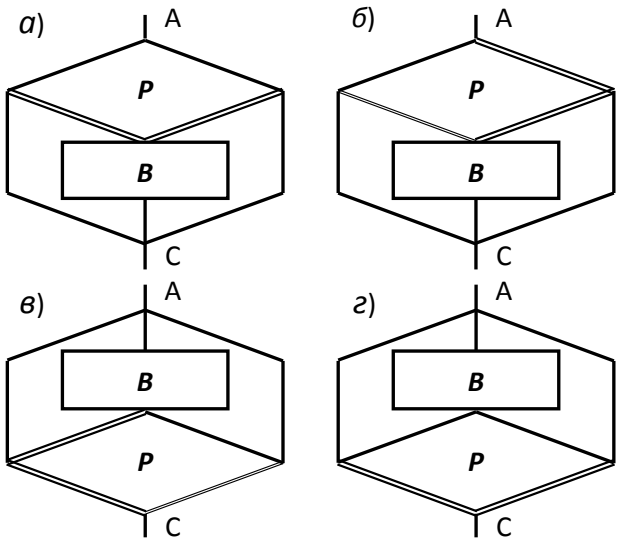


Рисунок 5

В языке VBA визуального объектно-ориентированного программирования (Visual Basic for Application) логические конструкции повторений «предусловие» и «постусловие» действуют с соглашениями While и Until. Это существенно расширяет возможности формирования условий выхода из КП в блоке принятия решений.

В частности, КП «счетчик» можно изобразить как цикл-предусловие (рисунок 5, *а*) с указателем «если ***P*** (истина), то выполнить ***B***». То есть инструкция этого цикла исполняется (соглашение While) при истинном значении логической переменной ***P*** = {***счетчик*** = ***s***↗ (или ↘) ***c***} в блоке принятия решений.

В конструкции повторений «предусловие» с соглашением Until (рисунок 5, *б*) указатель «≫» означает переход исполнения алгоритма в точку С, то есть окончание работы цикла. На рисунке 5 представлены также два варианта выполнения конструкции повторений «постусловие» с соглашениями While и Until, обозначенные буквами *в* и *г* соответственно. Для создания в языках программирования отсутствующих КП используем равносильности: While ***P*** ⇔ Until ***не*** (***P***) и Until ***P*** ⇔ While ***не*** (***P***).

Приведем основные логические конструкции, применяемые в VBA:

Примыкание: If ***P*** Then
Bt
End If

Развилка: If ***P*** Then
Bt
Else
Bf
End If

Счетчик: For <счетчик> = ***s*** To ***c***
B
Next <счетчик>

Цикл: (*с предусловием*)
Do <Until/While> ***P***
B
(*с постусловием*)
Loop <Until/While> ***P***

(в КП «цикл» используется одно из четырех предлагаемых соглашений).

1.4 Разработка, визуальное описание и анализ исполнения объектов алгоритма программы

Для упрощения записи алгоритма на языке программирования высокого уровня на стадии предкомпиляции будем применять словесные и/или визуальные формы его представления.

Задача 1. Построить блок-схему алгоритма решения уравнения

$$f(x) = 0, \text{ где } f \in C_{[-1;\,1]}, \qquad (1.11)$$

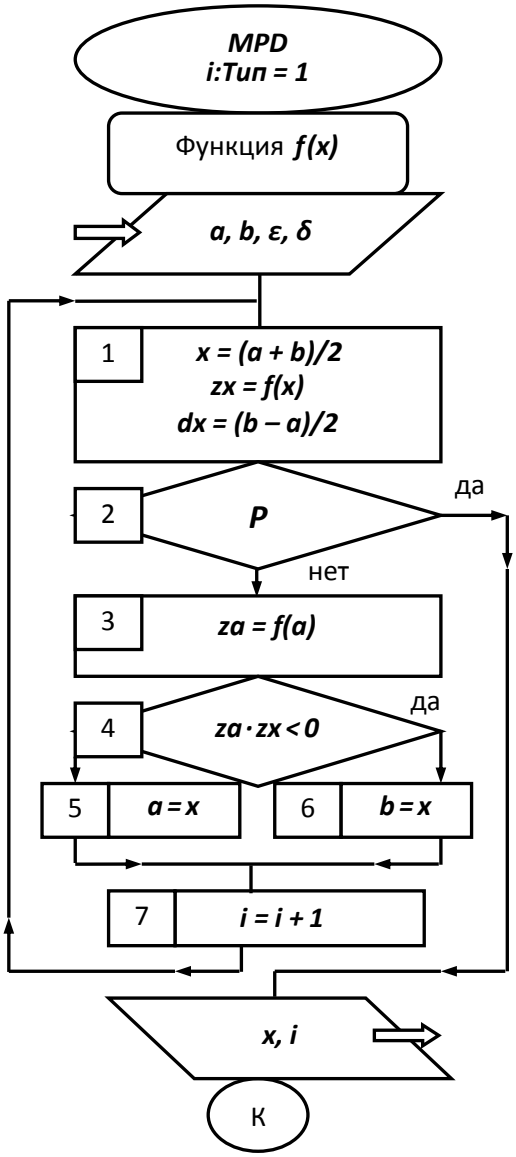


Рисунок 6

если на концах отрезка $[-1; 1]$ функция $y = f(x)$ принимает значения разных знаков.

Для вычисления корня x^* уравнения (1.11) разделим отрезок $[-1; 1]$ пополам и определим значение функции в полученной точке $x_i = 0$ $(i = 1)$. Если $P \equiv$ «$f(x_i) = 0$ или {$|f(x_i)|$ не больше $\varepsilon > 0$ и область существования корня достаточно мала $|x_i - x^*| \le \delta$}» – истина, то точку x_i будем считать приближенным решением (1.11) с заданной точностью «ε-δ» (рисунок 6). Иначе выберем из двух образовавшихся отрезков тот, на концах которого $f(x)$ принимает значения с разными знаками, и взятый отрезок вновь разделим пополам $(i = 2)$. Деление отрезков выполняем до тех пор, пока не станет истинным логическое выражение P в блоке принятия решения об окончании работы КП.

Отметим, что Останов итерационного процесса решения (1.11) по одному из условий ($|f(x_i)| \le \varepsilon$ – слабая сходимость к корню x^* или $|x_i - x^*| \le \delta$ – сильная сходимость к x^*) второго операнда P может привести к ошибке. В первом случае примером является функция $f(x) = \varepsilon\,(8\,x^3 - 1)$, принимающая в точке $x = 0$ значение $y = -\varepsilon$. Так как условие «Останова по норме невязки» выполнено, то 0 становится «приближенным» корнем (1.11), однако точным решением (1.11) с данной функцией $f(x)$ является $x^* = 1/2$.

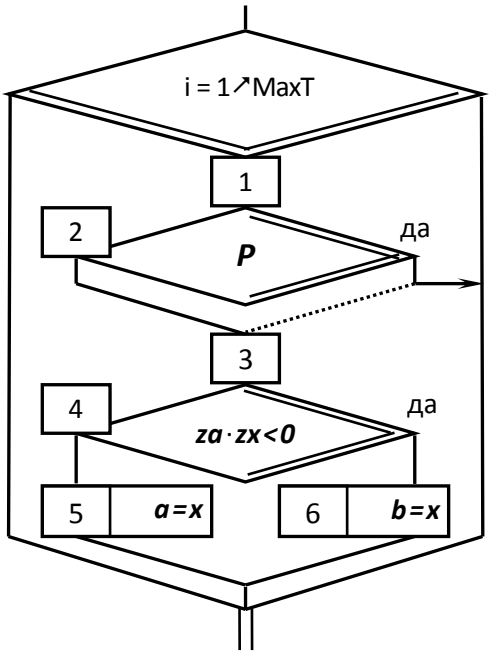


Рисунок 7

Некорректность второго способа Останова вытекает из решения уравнения (1.11) с $f(x) = \sqrt[3]{x-\delta}/\delta$, принимающей в точке $x = 0$ ($|x - x^*| = \delta$) значение $y = -\delta^{-2/3}$, существенно отличающееся от 0 при $\delta << 1$. Тем не менее, второй критерий более надежен, так как он точно определяет область существования корня. Поэтому, предусматривая все возможные контрпримеры при формировании условия выхода из КП, будем отдавать предпочтение второму критерию «Останова по локализации корня».

Описание алгоритма решения (1.11) методом половинного деления (МПД) основано на типичных постулатах: $a < b$; $f(a)\cdot f(b) < 0$, где $a = -1$, $b = 1$; перед проверкой условия завершения итерационного процесса каждый операнд в P должен быть сформирован и определен.

По логическому построению КП пронумерованные блоки делятся на три составляющих, которые совместно с операторами безусловного и условного переходов организуют оптимальные циклические процессы в языках низкого уровня, а также имеют свое специфическое назначение при организации циклов в современных языках высокого уровня.

До появления структурированных алгоритмических языков совокупность блоков 1–7 называли *телом цикла*. Чтобы не использовать (в явном виде) безусловные переходы в конструкции повторений, тело цикла поместили в цикл с целочисленным счетчиком, где блок 7 присутствует по умолчанию.

На рисунке 7 показана объект-схема *алгоритма* решения (1.11) при помощи цикла-счетчика. Принципиально *он* ничем не отличается от оптимального алгоритма, блок-схема которого изображена на рисунке 6. Но если исключить из последнего типизированную целочисленную переменную i, то ни один цикл-счетчик не сможет воспроизвести работу оптимальной логической конструкции, когда требуемое число повторений больше $2\,MaxT$ ($MaxT$ – предельное значение T-типа целого i).

Используя при решении (1.11) цикл-счетчик, в алгоритме появляются два варианта выхода из КП. Для их распознавания потребуется блок принятия решения, что приведет к созданию еще одного внешнего цикла. Значит, оптимальный цикл и цикл-счетчик с одним и тем же телом не эквивалентны.

При разработке КП в структурированных языках высокого уровня была предпринята попытка устранить этот недостаток цикла-счетчика, используя одно условие окончания работы цикла с неограниченным числом повторений. Для реализации этой идеи и «исключения» безусловного перехода стали разрабатывать циклы с постусловием и предусловием, смещая блок принятия решения в конец или начало цикла соответственно.

Телом цикла назвали множество операторов, расположенных между зарезервированными словами НЦ и КЦ, исключив из тела цикла блок принятия решения. Очевидно, этот блок и связанные с ним условные и безусловные переходы по-прежнему являются неотъемлемой частью тела цикла, повторяющейся на каждом шаге. Множество операторов, расположенных между указанными зарезервированными словами, составляют *инструкцию*.

При всех $\varepsilon > 0$ и $\delta > 0$ алгоритм решения (1.11) МПД на множестве R завершается выводом $x \approx x^*$. Говоря о «всех» значениях, будем подразумевать, что таковыми они являются лишь теоретически, а в ИС их представление зависит от объема памяти, отведенного для размещения этих данных.

Задача 2. Описать алгоритм процесса решения уравнения

$$f(x) = 0, \text{ где } f \in C^2_{[a,\,b]} \qquad (1.12)$$

итерационным методом Ньютона (методом касательных – рисунок 8).

Этот метод основан на принципе линейной аппроксимации образа функции $f(x_n + \Delta x_n)$ в окрестности $n^{\text{-го}}$ приближения и состоит в том, что, начиная с заданного нулевого приближения x_0, по рекуррентной формуле

$$x_{n+1} = x_n + \Delta x_n, \text{ где } \Delta x_n = -\frac{f(x_n)}{f'(x_n)},\ n = 0, 1, \ldots \qquad (1.13)$$

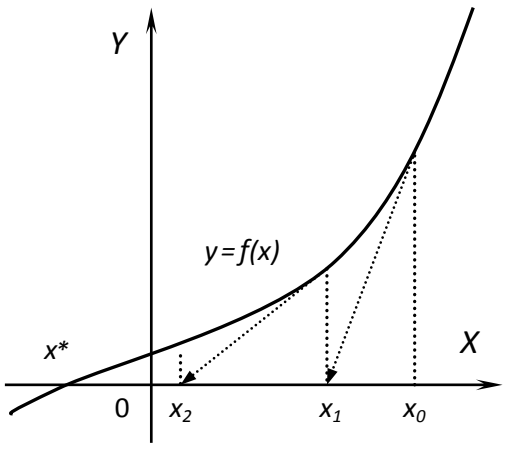


Рисунок 8

генерируется последовательность приближенных значений корня x^* уравнения (1.12). Изложим идею метода.

Из формулы Тейлора и теоремы о среднем следует, что для дважды непрерывно дифференцируемой в области $D_r = \{x : |x - x_0| \le r\}$ функции f и любого Δx ($|\Delta x| \le r$) справедливо

$$f(x_0 + \Delta x) = \frac{f''(\chi)}{2}\Delta^2 x + f'(x_0)\Delta x + f(x_0), \qquad (1.14)$$

где точка χ принадлежит отрезку $[x_0, x_0 + \Delta x]$.

Пусть x^* – единственный в области D_r корень уравнения (1.12) и $(f'(x_0))^2 \ge 2 f''(\chi) f(x_0)$. Тогда, зная χ, поправку $\Delta x \equiv x^* - x_0$ к приближению x_0 найдем из равенства (1.14). Однако в общем случае определить χ невозможно, поэтому для вычисления приближенного значения корня (1.12) используем линейную часть квадратичной аппроксимации $f(x_0 + \Delta x)$

$$f(x_0 + \Delta x) \approx f'(x_0)\Delta x + f(x_0). \qquad (1.15)$$

Оставив в силе предположение, что поправка Δx приведет приближение x_0 к x^*, найдем ее как Δx_0 с точностью $O(f^2(x_0))$ из уравнения

$$f(x_0) + f'(x_0)\Delta x_0 = 0, \qquad (1.16)$$

где $|\Delta x_0 - \Delta x| \le 2\left|\frac{f''(\chi)}{f'(x_0)}\right| |\Delta x_0|^2$.

Учитывая все допущенные упрощения при вычислении Δx_0, утверждение $x_0 + \Delta x_0 = x^*$ становится маловероятным. Тогда $x_0 + \Delta x_0 \equiv x_1$ примем за следующее приближение корня x^*, повторим описанную итерацию для x_1 и т.д. Можно сформулировать несколько теорем о сходимости метода Ньютона, что и будет сделано позже при решении нелинейных операторных уравнений. Здесь же рассмотрим простой случай, вытекающий из ограниченности второй производной $\max\limits_{x \in D_r} |f''(x)| \le K$.

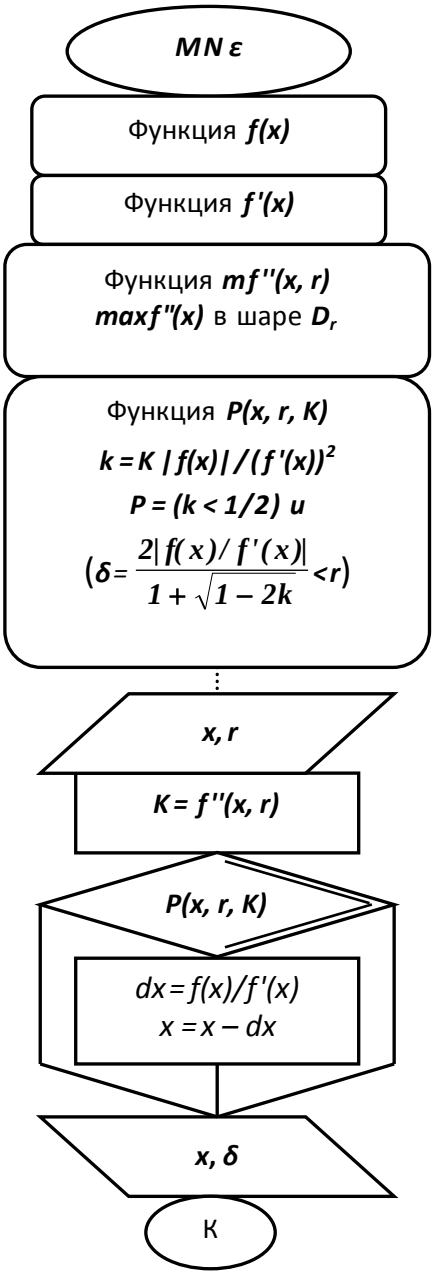


Рисунок 9

Теорема 1.3. Пусть

$$k = \frac{K|f(x_0)|}{(f'(x_0))^2} < \frac{1}{2}.$$

Тогда если $\delta = \frac{2|\Delta x_0|}{1+\sqrt{1-2k}} \le r$, то уравнение (1.12) имеет в шаре Q_δ единственный корень x^*, к которому сходится итерационный процесс Ньютона (1.13) с нуль-приближения x_0.

На рисунке 9 изображена объект-схема алгоритма метода Ньютона решения уравнения (1.12), по которому при выполнении условий теоремы 1.3 осуществляется локализация корня в δ-окрестности приближения. В отличие от алгоритма задачи 1, где в качестве КП был применен цикл-счетчик, здесь используется цикл-предусловие. При необходимости количество совершенных итераций будем подсчитывать с помощью вещественной переменной i.

Задача 3. Описать алгоритм решения уравнения

$$x = \varphi(x) \qquad (1.17)$$

методом последовательных приближений (МПП).

Теорема 1.4. Пусть функция $\varphi(x)$ определена на отрезке $Q_\delta = |x - x_0| \le \delta$ и удовлетворяет на нем условию

$$|\varphi(x_1) - \varphi(x_2)| \le q\,|x_1 - x_2| \ (0 \le q < 1).$$

Тогда если $|x_0 - \varphi(x_0)| \le m$ и $m \le (1-q)\,\delta$, то

1) уравнение (1.17) имеет корень в области Q_δ;

2) итерационная последовательность

$$x_n = \varphi(x_{n-1}) \qquad (1.18)$$

сходится в области Q_δ к пределу $x^* = \lim\limits_{n\to\infty} x_n$, который является корнем уравнения (1.17);

3) скорость сходимости x_n к x^* оценивается неравенством

$$|x_n - x^*| \le \frac{m q^n}{1-q},\ n = 1, 2, \ldots. \qquad (1.19)$$

С доказательством теоремы для операторных уравнений можно ознакомиться в монографии [15]. Что касается объект-схемы алгоритма, то из-за чрезвычайной простоты (иногда МПП называют методом простых итераций) она может быть представлена как в виде конструкции повторений с предусловием, так и в виде КП с постусловием. Отметим, что критерий завершения вычислительного процесса формируется из условия теоремы 1.4, а геометрический смысл пояснен на рисунке 10.

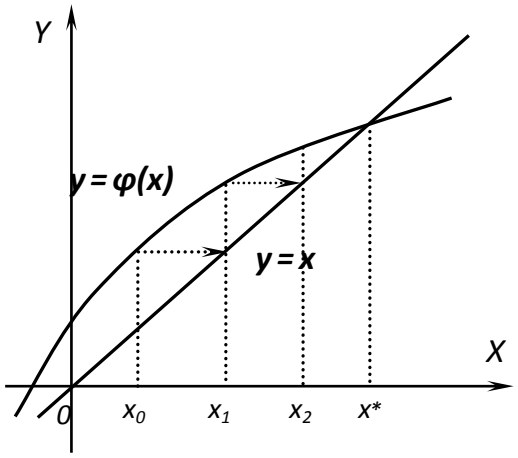


Рисунок 10

1.5 Понятие мощности множества. Пространство решений квадратных алгебраических уравнений

Определение 1.20. Пусть P – подмножество U. Совокупность всех элементов вида $F(u)$ при $u \in P$ называется ***образом*** P и обозначается

$$F(P) = \{F(u), u \in P\}.$$

Определим наиболее важные классы отображений $F : U \to V$:

- ***сюръекция***, если $F(U) = V$;
- ***инъекция***, если для любых u_1 и u_2, принадлежащих множеству U, из условия $u_1 \neq u_2$ следует, что $F(u_1) \neq F(u_2)$;
- ***биекция***, если F – сюръекция и инъекция одновременно (данное отображение называют также ***взаимно однозначным***).

Теорема 1.5. Образ объединения двух множеств равен объединению их образов

$$F(A \cup B) = F(A) \cup F(B).$$

Теорема 1.6. Прообраз объединения двух множеств равен объединению их прообразов

$$F^{-1}(A \cup B) = F^{-1}(A) \cup F^{-1}(B).$$

Теорема 1.7. Прообраз пересечения двух множеств равен пересечению их прообразов

$$F^{-1}(A \cap B) = F^{-1}(A) \cap F^{-1}(B).$$

Опишем методику, согласно которой осуществляется количественное сравнение множеств.

Пусть A и B – два конечных множества. В этом случае отношения «больше», «меньше» или «равно» для множеств можно определить, сравнивая количество элементов, из которых они состоят.

Существует и другой способ сравнения. Попытаемся установить взаимно однозначное соответствие F между элементами этих множеств. Это возможно лишь при равном числе элементов множеств A и B ($F(A) = B$). В противном случае возникает два варианта соответствий: множество A отображается в подмножество B или множество B в подмножество A, то есть

$$F(A) \subset B \text{ или } F^{-1}(B) \subset A. \qquad (1.20)$$

Это обстоятельство позволяет очевидным образом перенести понятие сравнения и на бесконечные множества, для которых вводится специальная шкала «мощностей». Среди всех бесконечных множеств наименьшим (самым слабым по мощности) является множество N ***натуральных чисел***, которое состоит из элементов счета ***1, 2, 3, ….***

Определение 1.21. Назовем ***счетным множеством*** всякое множество, элементы которого можно взаимно однозначно сопоставить со всеми натуральными числами множества N.

Иначе говоря, счетное множество – это такое множество, элементы которого можно пронумеровать, то есть представить в виде бесконечной последовательности $\{a_1, a_2, \ldots, a_n, \ldots\}$.

Утверждение 1.1. Теоремы 1.5, 1.6 и 1.7 остаются в силе для любого конечного или счетного числа множеств.

Примеры счетных множеств:

- множество $N_0 = \{N \cup 0\}$;
- множество Z ***целых*** чисел;
- множество Q ***рациональных*** чисел;
- множество P_M многочленов M переменных с коэффициентами из Q.

Для чисел множества N_0 разработана теория так называемой целочисленной арифметики, основанная на следующем утверждении.

Утверждение 1.2. Для любых натуральных чисел m и n существуют такие числа c и o из множества N_0, что справедливо равенство

$$m = c \cdot n + o, \text{ где } o < n. \qquad (1.21)$$

Если $o = 0$, то будем говорить, что число m *делится* на n *нацело*, и обозначать это как $m \vdots n$, в противном случае m *делится* на n *с остатком* o. Число c будем называть частным от целочисленного деления m на n.

В Паскале c находится как m **Div** n, а в VBA $c = Int(m/n)$ или $c = m \setminus n$. Остаток равен соответственно $o := m$ **Mod** n и $o = m - n*c$.

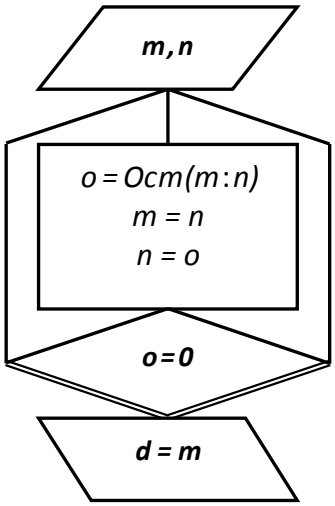


Рисунок 11

Опишем алгоритм нахождения наибольшего общего делителя (НОД) чисел ***m*** и ***n***. Для удобства вычисления значений ***c*** и ***o*** в VBA создадим функции ***Div(m, n)*** и ***Mod(m, n)***.

Из несложных умозаключений следует, что если НОД(***m***, ***n***) = ***d***, то при ***o*** ≡ ***Ocm(m : n)*** ≠ ***0*** вытекает равенство НОД(***n***, ***o***) = ***d***. В связи с этим критерием Останова циклического процесса при последовательном уменьшении остатка от деления чисел, для которых определяется НОД, является условие ***P*** = «***o*** = ***0***». Наибольшим общим делителем чисел ***m*** и ***n*** в этом случае будет последний, неравный нулю остаток (рисунок 11). Отметим, что

$$\text{НОД}(\boldsymbol{m}, \boldsymbol{n}) \cdot \text{НОК}(\boldsymbol{m}, \boldsymbol{n}) = \boldsymbol{m} \cdot \boldsymbol{n}. \qquad (1.22)$$

Важнейшими задачами целочисленной арифметики являются: установление количества делителей натурального числа (выясняющей также является ли данное число простым); перевод натуральных чисел из одной системы счисления (СС) в другие; представление любого натурального числа ***m*** > ***1*** в виде произведения степеней с натуральными показателями ***k₁***, ***…***, ***k_r***

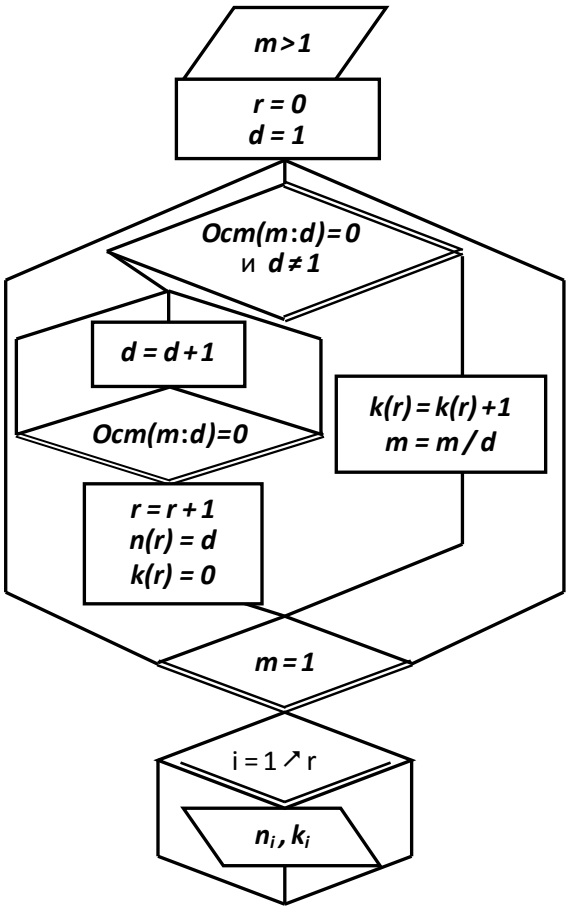


Рисунок 12

$$\boldsymbol{m} = n_1^{k_1} n_2^{k_2} \ldots n_r^{k_r}, \qquad (1.23)$$

где ***n₁***, ***n₂***, ***…***, ***n_r*** – все простые делители числа ***m*** и т.д.

Поиск простых делителей начнем с наименьшего ***d*** = ***2*** и при наличии у числа ***m*** какого-либо делителя установим его кратность переприсваиванием ***m*** := ***m/d***, затем в цикле найдем следующий делитель и т.д. На рисунке 12 изображена объект-схема алгоритма в виде конструкции повторений с постусловием, где критерием Останова является условие ***m*** = ***1*** (составьте еще три разные схемы алгоритмов).

Приведем некоторые общие свойства счетных множеств:

- каждое подмножество счетного множества конечное или счетное;
- объединение любого конечного или счетного множества счетных множеств снова счетное множество;
- всякое бесконечное множество содержит счетное подмножество.

Определение 1.22. Два множества A и B называются ***эквивалентными*** (обозначается $A \sim B$), если между их элементами можно установить взаимно однозначное соответствие.

Теорема 1.8. Множество $R_{(0;\,1)} = \{x : x \in R,\ 0 < x < 1\}$ действительных чисел, заключенных между нулем и единицей, несчетно.

Доказательство теоремы, основанное на диагональной процедуре Кантора, приведено в книге А.Н. Колмогорова и С.В. Фомина [10, с. 32].

Говорят, что множество $(0;1) \equiv R_{(0;1)}$ имеет мощность ***континуум*** (K). Примеры множеств, эквивалентных множеству точек интервала $(0;1)$:

- множество всех точек отрезка $[a, b]$ или интервала (a, b);
- множество всех точек на прямой (R);
- множество всех точек на плоскости (R^2) или в пространстве (R^3);
- множество всех функций M переменных непрерывных на множестве

$$X \subset R^M = \bigotimes_{j=1}^{M} R = R \otimes R \otimes \ldots \otimes R \quad (M \text{ сомножителей}).$$

Определение 1.23. Если множества A и B эквивалентны между собой, то говорят, что они имеют ***одинаковую мощность*** и пишут, $m(A) = m(B)$.

Отметим, что шкала мощностей бесконечных множеств не ограничена, то есть существуют множества с мощностью, превосходящей мощность континуума. Однако из всех бесконечных множеств мы будем рассматривать только счетные множества или множества, имеющие мощность K.

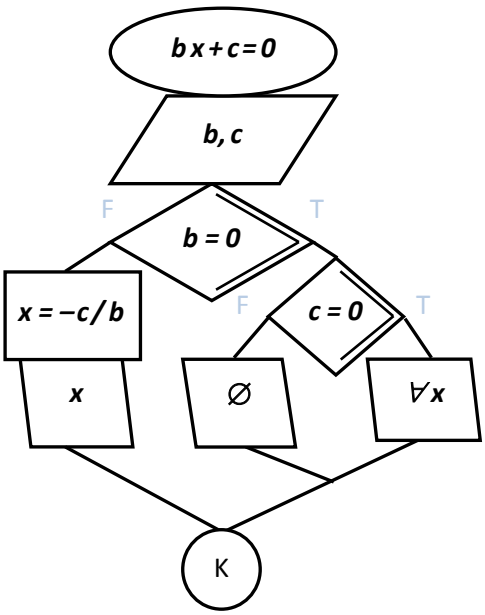


Рисунок 13

Разработаем алгоритм поиска корней уравнения вида

$$bx + c = 0,\ b \in R,\ c \in R. \qquad (1.24)$$

На рисунке 13 изображена объект-схема алгоритма решения уравнения (1.24) при различных значениях действительных переменных b и c. Существование единственного корня линейного уравнения следует из условия $b \neq 0$. В противном случае (1.24) не имеет корней (при $c \neq 0$) или его корнем (при $c = 0$) является любое вещественное число $x \in R$.

В процесс решения уравнения с действительными коэффициентами

$$a x^2 + b x + c = 0,\ \{a, b, c\} \subset R \qquad (1.25)$$

рассмотренный выше алгоритм решения (1.24) входит в качестве подпрограммы (при $a = 0$).

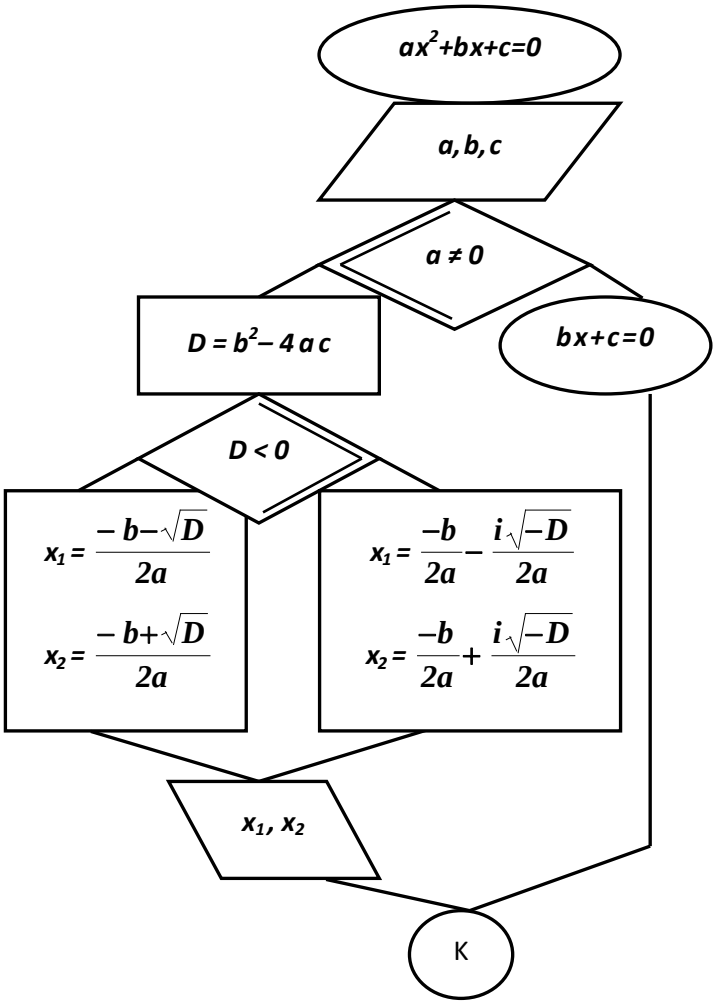


Рисунок 14

Представленную на рисунке 14 схему алгоритма решения уравнения (1.25) над полем C комплексных чисел можно разбить на четыре группы команд исполнителя. К первой группе отнесем команды постановки задачи и ввода заданных коэффициентов.

Вторую группу образуют команды линии указателя условной логической конструкции с блоком принятия решения $a \neq 0$, определяющие корни квадратного уравнения. Третья группа – это команда «решать уравнение $b x + c = 0$» с блоком команд предыдущего алгоритма. Четвертая группа команд организует вывод результатов и окончание работы алгоритма.

Большую роль при решении уравнений вида (1.1) играет выбор пространства решений U. Здесь нас будут интересовать ответы на два вопроса:

- возможно ли построение фундаментальной последовательности приближений корня в пространстве решений (и если это осуществимо, то);
- гарантирована ли сходимость этой последовательности в U.

Например, решение уравнения (1.25) в пространстве R действительных чисел возможно лишь при условии $D \geq 0$, а в пространстве Q еще и при других дополнительных условиях, налагаемых на коэффициенты a, b и c.

Глава 2 Отображение метрических пространств. Линейная система алгебраических уравнений

2.1 Норма и метрика. Скалярное произведение. Евклидово пространство $C_{2[a,b]}$ непрерывных функций

Важное место в анализе занимает понятие близости элементов пространства. Поэтому, кроме операций сложения элементов и умножения их на число, в линейном пространстве необходимо ввести понятия размера (норма) элемента и расстояния (метрика) между элементами.

Определение 2.1. Говорят, что в линейном пространстве $\boldsymbol{L}$ задана ***норма***, если каждому элементу $\boldsymbol{u} \in \boldsymbol{L}$ поставлено в соответствие действительное число $\|\boldsymbol{u}\|$, называемое *нормой* $\boldsymbol{u}$, так, что верны три аксиомы:

1) $\|\boldsymbol{u}\| \geq \boldsymbol{0}$ и $\|\boldsymbol{u}\| = \boldsymbol{0}$ тогда и только тогда, когда $\boldsymbol{u} = \boldsymbol{0}$;

2) $\|\boldsymbol{\alpha u}\| = |\boldsymbol{\alpha}| \cdot \|\boldsymbol{u}\|, \ \forall \boldsymbol{\alpha} \in \boldsymbol{R}$;

3) $\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|, \ \forall \boldsymbol{v} \in \boldsymbol{L}$ (неравенство треугольника).

Определение 2.2. Линейное пространство $\boldsymbol{L}$, в котором задана норма, назовем ***нормированным*** пространством. В случае линейного функционального пространства, в котором определена норма, говорят о ***функциональном нормированном*** пространстве, а норму элемента (функции) в таком пространстве называют ***нормой функции***.

Отметим, что для любых элементов $\boldsymbol{u}$ и $\boldsymbol{v}$ из нормированного пространства $\boldsymbol{L}$ справедливо следующее двойное неравенство

$$|\,\|\boldsymbol{u}\| - \|\boldsymbol{v}\|\,| \leq \|\boldsymbol{u} \pm \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|. \qquad (2.1)$$

Действительно, если в неравенстве треугольника (аксиома 3) сначала заменить элемент $\boldsymbol{u}$ на разность $\boldsymbol{u} - \boldsymbol{v}$, а затем заменить $\boldsymbol{v}$ на $\boldsymbol{v} - \boldsymbol{u}$, то получим *два неравенства*

$$\|\boldsymbol{u}\| - \|\boldsymbol{v}\| \leq \|\boldsymbol{u} - \boldsymbol{v}\| \text{ и } \|\boldsymbol{v}\| - \|\boldsymbol{u}\| \leq \|\boldsymbol{v} - \boldsymbol{u}\|.$$

Пусть $\|\boldsymbol{v}\| \leq \|\boldsymbol{u}\|$, тогда из *первого* неравенства вытекает, что

$$|\,\|\boldsymbol{u}\| - \|\boldsymbol{v}\|\,| \leq \|\boldsymbol{u} \pm \boldsymbol{v}\|. \qquad (2.2)$$

Если же $\|\boldsymbol{u}\| < \|\boldsymbol{v}\|$, то из *второго* неравенства следует

$$|\,\|\boldsymbol{v}\| - \|\boldsymbol{u}\|\,| < \|\boldsymbol{u} \pm \boldsymbol{v}\|. \qquad (2.3)$$

Из двух полученных неравенств (2.2) и (2.3) вытекает справедливость левой части двойного неравенства (2.1), правая часть которого непосредственно следует из определения нормы.

Понятие *расстояния* между элементами, присущее метрическим (не всегда линейным) пространствам, позволяет ввести одну из важнейших операций анализа – *предельный переход*. Раскроем суть этих и некоторых других понятий, рассматриваемых при изучении метрических пространств.

Определение 2.3. Множество Q назовем ***метрическим*** пространством, если в нем для всех u и v определена действительная неотрицательная функция $\rho(u, v)$, подчиненная следующим трем аксиомам *метрики*:

1) $\rho(u, v) = 0$ тогда и только тогда, когда $u = v$;

2) $\rho(u, v) = \rho(v, u)$ (аксиома симметрии);

3) $\rho(u, w) \le \rho(u, v) + \rho(v, w)$, $\forall w \in Q$ (аксиома треугольника).

Значение функции $\rho(u, v)$ будем называть ***расстоянием*** между двумя элементами u и v, принадлежащими метрическому пространству Q.

Определение 2.4. ***Отрезком*** $[u, w]$ назовем множество всех точек v, удовлетворяющих условию $\rho(u, v) + \rho(v, w) = \rho(u, w)$. Метрическое пространство Q, которое вместе с любыми двумя своими элементами содержит соединяющий их отрезок, называется ***выпуклым***.

Отметим, что любое выпуклое подмножество Q нормированного пространства L является метрическим пространством с метрикой индуцированной нормой, если для всех u и v из Q положить

$$\rho(u, v) = \|u - v\|. \qquad (2.4)$$

Окрестность точки, ***внутренняя***, ***изолированная***, ***граничная*** и ***предельная*** точки множества, ***фундаментальная*** последовательность, ***ограниченность***, ***плотность***, ***замкнутость*** и ***компактность*** множеств, а также многие другие понятия, определяемые для метрического пространства (см., например, [7] или [10]), применимы и для нормированного пространства.

Определение 2.5. ***Открытый шар*** $Q[u_0, r)$ с центром в точке u_0 и радиусом $r > 0$ в нормированном пространстве L – это множество элементов $u \in L$, удовлетворяющих условию $\|u - u_0\| < r$.

Замкнутый шар $Q[u_0, r]$ в нормированном пространстве L – это множество точек $u \in L$, удовлетворяющих условию $\|u - u_0\| \le r < \infty$.

В зависимости от контекста ***ε-окрестностью*** точки u_0 будем называть открытый или замкнутый шар с радиусом $r = \varepsilon$.

Определение 2.6. Точка $u \in L$ называется ***предельной*** точкой подмножества Q нормированного пространства L, если любая ее окрестность содержит бесконечно много точек из Q.

Определение 2.7. Совокупность всех предельных точек множества Q обозначим $[Q]$ и назовем ***замыканием*** Q.

Определение 2.8. Множество $P \subset L$ называется ***всюду плотным в L***, если его замыкание $[P]$ совпадает со всем множеством L.

Определение 2.9. Множество Q называется ***ограниченным в L***, если оно полностью содержится в замкнутом шаре множества L, то есть для всех u из Q существуют элемент $u_0 \in L$ и число $R \in R$ такие, что

$$\| u - u_0 \| \le R.$$

Определение 2.10. Множество $Q \subset L$ называется ***открытым*** в нормированном пространстве L, если выполнено одно из следующих условий:

- его дополнение $L_Q = L \backslash Q$ содержит все свои предельные точки, то есть замкнуто;
- для всякого элемента $u \in Q$ найдется открытый шар с центром в u, содержащийся в L.

Утверждение 2.1. Условия, приведенные в определении 2.10 открытого множества в нормированном пространстве, равносильны.

Совокупность открытых множеств $\{Q_k, k = 0, 1, 2, \ldots\}$ называется открытым *покрытием* метрического пространства Q, если

$$Q \subset \bigcup_k Q_k .$$

Определение 2.11. Метрическое пространство Q называется ***компактным***, если любое его открытое покрытие содержит конечное подпокрытие.

Очевидно, что любое замкнутое ограниченное подмножество конечномерного нормированного пространства компактно.

Под сходимостью последовательности $\{u_n, n = 0, 1, 2, \ldots\}$ по норме пространства L понимают сходимость этой последовательности в метрике ρ, индуцированной нормой $\| \ \|_L$. В связи с этим понятие *расстояния* будет всегда подразумеваться, когда речь идет о сходимости.

Определение 2.12. Элемент u нормированного пространства L называется ***пределом последовательности*** $\{u_n, n = 0, 1, 2, \ldots\} \subset Q \subset L$, сходящейся по норме L, если

$$\lim_{n\to\infty} \| u_n - u \| = 0 \text{ (то есть } \lim_{n\to\infty} \rho(u_n, u) = 0 \text{)}. \qquad (2.5)$$

Отметим, что предел u последовательности $\{u_n, n = 0, 1, \ldots\}$, являясь предельной точкой множества Q, может не принадлежать Q. Примером тому служат рациональные приближения иррациональных чисел из R .

Утверждение 2.2. Сходящаяся в нормированном пространстве L последовательность имеет единственный предел.

Приведем примеры нормированных (метрических) пространств:

- множество $\boldsymbol{R}$ действительных чисел станет нормированным (метрическим) пространством, если для любых $\boldsymbol{x}$ и $\boldsymbol{y}$ из $\boldsymbol{R}$ положить

$$\|x\| = |x| \text{ и } \rho(x, y) = |x - y|; \tag{2.6}$$

- множество $\boldsymbol{C_{[a, b]}}$ непрерывных действительных функций на отрезке $\boldsymbol{[a, b]}$ является нормированным пространством с *равномерной* нормой

$$\|u\| = \max_{a \le t \le b} |u(t)|; \tag{2.7}$$

- множество $\boldsymbol{C_{1[a, b]}}$ непрерывных действительных функций на отрезке $\boldsymbol{[a, b]}$ является нормированным пространством с *интегральной* нормой

$$\|u\| = \int_a^b |u(t)|\,dt; \tag{2.8}$$

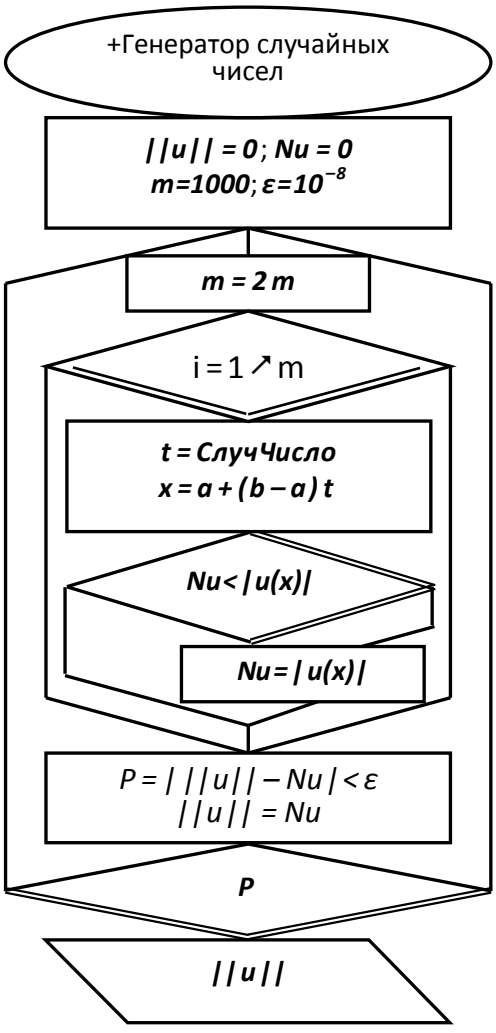


Рисунок 15

- множество $\boldsymbol{C_{2[a, b]}}$ непрерывных действительных функций на отрезке $\boldsymbol{[a, b]}$ является нормированным пространством с *квадратичной* нормой

$$\|u\| = \left(\int_a^b (u(t))^2 dt \right)^{1/2}. \tag{2.9}$$

Опишем алгоритм нахождения нормы функции $\boldsymbol{u(x)}$ из $\boldsymbol{C_{[a, b]}}$ статистическим методом Монте-Карло. Выбор точек множества $\boldsymbol{X = [a, b]}$ для вычисления значений функции $\boldsymbol{u(x)}$ осуществим генерацией последовательности случайных чисел отрезка $\boldsymbol{[0; 1]}$ (с периодом повторения более $\boldsymbol{10^9}$ и равномерным распределением по ста интервалам с погрешностью менее $\boldsymbol{5 \cdot 10^{-4}}$). Тогда точка $\boldsymbol{x = a + (b - a)\,t}$ с $\boldsymbol{t \in [0; 1]}$ принадлежит $\boldsymbol{X}$.

Критерием Останова конструкции повторений «постусловие» будем считать незначительное изменение $\boldsymbol{\|u(x)\|}$ при двукратном увеличении количества точек $\boldsymbol{m}$ подсчета $\boldsymbol{|u(x)|}$. На рисунке 15 приведена объект-схема алгоритма, с помощью которого вычисляется значение нормы $\boldsymbol{u(x)}$ в пространстве $\boldsymbol{C_{[a, b]}}$.

Определение 2.13. ***Скалярным произведением*** векторов в линейном пространстве L называется действительная функция $\langle\ ,\ \rangle$, определенная для всех пар элементов $\{u, v\} \subset L$ и удовлетворяющая постулатам:

1) $\langle u, v\rangle = \langle v, u\rangle$;
2) $\langle u + w, v\rangle = \langle u, v\rangle + \langle w, v\rangle,\ \forall w \in L$;
3) $\langle \alpha \circ u, v\rangle = \alpha \circ \langle u, v\rangle,\ \forall \alpha \in R$;
4) $\langle u, u\rangle \geq 0$, причем $\langle u, u\rangle = 0$ только при $u = 0$.

Один из способов введения нормы (назовем ее ***естественной***) в линейном пространстве − это задание в нем скалярного произведения.

Определение 2.14. Линейное пространство с естественной нормой

$$\| u \| = \langle u, u\rangle^{1/2} \tag{2.10}$$

называется ***евклидовым*** (или E-пространством).

В качестве примера E-пространства рассмотрим множество пар действительных чисел R^2. Скалярное произведение элементов ${}^2u = (u_1, u_2)$ и ${}^2v = (v_1, v_2)$ этого пространства определим по формуле

$$\langle {}^2u, {}^2v\rangle = u_1 v_1 + u_2 v_2. \tag{2.11}$$

Не будем использовать верхний левый индекс, указывающий на принадлежность элемента конечномерному пространству, если его значение очевидно из постановки задачи.

Докажем справедливость аксиом скалярного произведения в R^2.

1. $u_1 v_1 + u_2 v_2 = v_1 u_1 + v_2 u_2$.
2. $(u_1 + w_1)v_1 + (u_2 + w_2)v_2 = (u_1 v_1 + u_2 v_2) + (w_1 v_1 + w_2 v_2)$.
3. $(\alpha\, u_1)v_1 + (\alpha\, u_2)v_2 = \alpha\,(u_1 v_1 + u_2 v_2)$.
4. $\langle u, u\rangle = u_1 u_1 + u_2 u_2 = (u_1)^2 + (u_2)^2 \geq 0$, при этом $(u_1)^2 + (u_2)^2 = 0$ тогда и только тогда, когда $u_1 = u_2 = 0$.

Утверждение 2.3. Из свойств 1–4 скалярного произведения векторов и неравенства Коши-Буняковского

$$\langle u, u\rangle \langle v, v\rangle \geq \langle u, v\rangle^2. \tag{2.12}$$

вытекает справедливость всех аксиом нормы для *естественной нормы* E-пространства R^2.

После введения понятия «скалярное произведение элементов» E-пространство R^2 обретает свою геометрию. Сейчас в нем можно определить понятия *угла* между двумя векторами, *параллельности* и *перпендикулярности* ненулевых элементов.

Угол между ненулевыми векторами u и v найдем из соотношений

$$\langle u, v\rangle = \|u\|\,\|v\|\,\cos(u \wedge v) \text{ или } \langle u, v\rangle = \langle u, u\rangle^{1/2} \langle v, v\rangle^{1/2} \cos(u \wedge v). \tag{2.13}$$

Если быть последовательным, то изначально скалярное произведение векторов плоскости определялось в приложениях по формуле (2.13), затем с введением декартовой системы координат была получена формула (2.11).

Утверждение 2.4. Значение скалярного произведения двух векторов $\boldsymbol{u = (u_1, u_2)}$ и $\boldsymbol{v = (v_1, v_2)}$ выражается через их координаты по формуле:

$$\langle \boldsymbol{u, v} \rangle \equiv \|\boldsymbol{u}\| \, \|\boldsymbol{v}\| \cos(\boldsymbol{u \wedge v}) = \boldsymbol{u_1 \cdot v_1 + u_2 \cdot v_2}. \qquad (2.14)$$

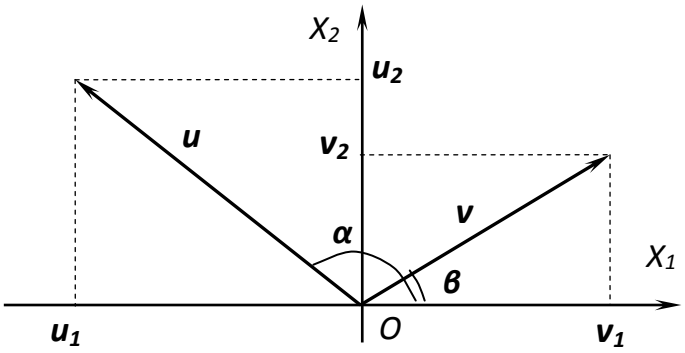


Рисунок 16

Так как угол между векторами $\boldsymbol{u}$ и $\boldsymbol{v}$ равен $\boldsymbol{\alpha - \beta}$ или $\boldsymbol{2\pi + \beta - \alpha}$, а $\boldsymbol{\cos(\beta - \alpha) = \cos(\alpha - \beta)}$, то

$$\|\boldsymbol{u}\| \, \|\boldsymbol{v}\| \cos(\boldsymbol{u \wedge v}) = \|\boldsymbol{u}\| \, \|\boldsymbol{v}\| \cos(\boldsymbol{\alpha - \beta}) =$$
$$= \|\boldsymbol{u}\| \, \|\boldsymbol{v}\| \cos\boldsymbol{\alpha} \cos\boldsymbol{\beta} + \|\boldsymbol{u}\| \, \|\boldsymbol{v}\| \sin\boldsymbol{\alpha} \sin\boldsymbol{\beta} =$$
$$= \|\boldsymbol{u}\| \cos\boldsymbol{\alpha} \, \|\boldsymbol{v}\| \cos\boldsymbol{\beta} + \|\boldsymbol{u}\| \sin\boldsymbol{\alpha} \, \|\boldsymbol{v}\| \sin\boldsymbol{\beta} =$$
$$= \boldsymbol{u_1 \, v_1 + u_2 \, v_2} \text{ (рисунок 16).}$$

Определение 2.15. Ненулевые векторы $\boldsymbol{u}$ и $\boldsymbol{v}$ из $\boldsymbol{R^2}$ называются *взаимно перпендикулярными* (обозначим $\boldsymbol{u \perp v}$), если $\langle \boldsymbol{u, v} \rangle = \boldsymbol{0}$.

Утверждение 2.5. Ненулевые векторы $\boldsymbol{v \in R^2}$ и $\boldsymbol{w \in R^2}$ параллельны между собой (обозначим $\boldsymbol{v \| w}$), то есть $\boldsymbol{w}$ и $\boldsymbol{v}$ сонаправлены ($\boldsymbol{v_1 w_2 = v_2 w_1}$) или противоположно направлены ($\boldsymbol{(-v_1) w_2 = (-v_2) w_1}$), если существует такой ненулевой вектор $\boldsymbol{u}$ из пространства $\boldsymbol{R^2}$, что

$$\langle \boldsymbol{u, v} \rangle = \boldsymbol{0} \text{ и } \langle \boldsymbol{u, w} \rangle = \boldsymbol{0}. \qquad (2.15)$$

Обобщив понятие скалярного произведения на пространство $\boldsymbol{R^n}$, найдем длины векторов $\boldsymbol{u}$ и $\boldsymbol{v}$, скалярное произведение и угол между векторами по заданным координатам. Чтобы иметь возможность работать с индексированными переменными (координатами вектора), необходимо описать векторы как массивы данных некоторой размерности (в нашем случае $\boldsymbol{n}$).

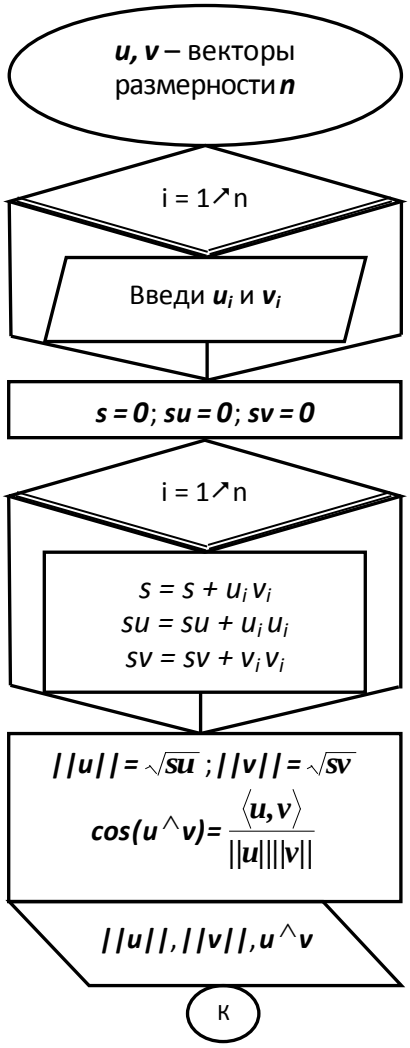


Рисунок 17

На рисунке 17 изображена объект-схема алгоритма вычисления длин векторов $\|\boldsymbol{u}\|$ и $\|\boldsymbol{v}\|$, а также скалярного произведения $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{s}$ и угла $\boldsymbol{u} \wedge \boldsymbol{v} = \boldsymbol{\arccos} \dfrac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|}$.

Введение нормы на множестве непрерывных функций по формуле (2.9) подводит нас к основному понятию геометрии бесконечномерных функциональных пространств – *скалярному произведению функций*.

Скалярное произведение непрерывных на отрезке $\boldsymbol{[a, b]}$ функций $\boldsymbol{u(x)}$ и $\boldsymbol{v(x)}$ в нормированных пространствах $\boldsymbol{C_{[a, b]}}$, $\boldsymbol{C_{1[a, b]}}$, $\boldsymbol{C_{2[a, b]}}$ (как, впрочем, и во многих других) можно определить по интегральной формуле

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \int_a^b \boldsymbol{u(t)\, v(t)\, dt}\,. \qquad (2.16)$$

Каждое из этих пространств содержит счетное всюду плотное по своей норме множество степенных и тригонометрических многочленов с рациональными коэффициентами. Отметим, что евклидовым пространством является лишь $\boldsymbol{C_{2[a, b]}}$, так как норма (2.9) есть естественная норма этого пространства.

2.2 Оператор. Операторное уравнение. Обратный оператор

В зависимости от постулирования пространств, связываемых отображением, существует разночтение при использовании термина «оператор». Из всего множества определений этого понятия будем придерживаться следующего.

Определение 2.16. Пусть в качестве $\boldsymbol{U}$ и $\boldsymbol{V}$ – произвольных множеств при отображении $\boldsymbol{F : U \to V}$ (определение 1.11) – используются *метрические* пространства, тогда такое отображение $\boldsymbol{F}$ называется ***оператором***, действующим из $\boldsymbol{U}$ в $\boldsymbol{V}$, а его результат находится по формуле

$$\boldsymbol{v = F(u)}, \qquad (2.17)$$

где $\boldsymbol{u \in U}$ – аргумент, $\boldsymbol{v \in V}$ – образ аргумента под действием $\boldsymbol{F}$.

Определение 2.17. Оператор $I: U \to U$, который переводит любой элемент $u \in U$ в себя, то есть $I(u) = u$, называется ***тождественным***.

Если оператор F связывает метрические пространства, то уравнение

$$F(u) = v, \text{ где } F: U \to V, \qquad (*)$$

назвается ***операторным*** уравнением с оператором или отображением F.

Определение 2.18. Оператор $F: U \to V$ будем называть ***взаимно однозначным***, если $F(U) = V$ и каждому элементу $v \in V$ отвечает единственный прообраз $u \in U$. Другими словами, для любого $v \in V$ существует, и притом единственный, корень уравнения (*).

Определение 2.19. Оператор, который каждому элементу $v \in V$ ставит в соответствие единственный элемент $u \in U$ такой, что выполняется равенство (*), называется ***обратным*** к оператору F.

Обратный оператор будем обозначать с помощью символов $[\ ]^{-1}$ или (если это не противоречит смыслу) без скобок. Из определения оператора, обратного к F, следуя этой символике, имеют место соотношения

$$F\,[F]^{-1} = F\,F^{-1} = I \text{ и } F^{-1}F = I.$$

Определение 2.20. Операторы

$$F_l^{-1}: V \to U \text{ и } F_r^{-1}: V \to U$$

называются соответственно ***левым*** и ***правым обратными*** операторами к оператору $F: U \to V$, если для всех $u \in U$ и $v \in V$ справедливо

$$F_l^{-1}(F(u)) = I_U u \text{ и } F(F_r^{-1}(v)) = I_V v,$$

где I_U и I_V – тождественные операторы в U и V соответственно.

Теорема 2.1. Если существуют левый и правый обратные операторы к оператору $F: U \to V$ и $F(U) = V$, то для $\forall v \in V$ операторное уравнение (*) имеет единственное решение и F имеет единственный обратный оператор F^{-1}, причем

$$F_l^{-1} = F_r^{-1} = F^{-1}.$$

Пусть существует оператор $F_r^{-1}: V \to U$ и $F(U) = V$, тогда для $\forall v$ из V найдется элемент $u = F_r^{-1}(v) \in U$, который обращает уравнение (*) в тождество, так как

$$F(F_r^{-1}(v)) = I_V v = v.$$

Из наличия $F_l^{-1}: V \to U$ следует, что этот корень единственный

$$F_l^{-1}(F(u)) = I_U u = u.$$

Упражнение. Проведите анализ взаимосвязи существования обратных операторов F_l^{-1} и F_r^{-1} с сюръективностью или инъективностью оператора F.

Сужение методами функционального анализа множества U существования корней уравнения (*) до области Q существования единственного корня операторного уравнения назовем локализацией корня.

Определение 2.21. Оператор F называется ***непрерывным*** в области $Q \subset U$, если он непрерывен в каждой точке $u \in Q$, то есть для любой ε-окрестности $Q[v, \varepsilon)$ точки $v = F(u)$ существует такая δ-окрестность $Q[u, \delta)$ точки u, что $F(w) \in Q[v, \varepsilon)$, как только

$$w \in \{Q[u, \delta) \cap Q\}.$$

2.3 Линейный оператор и линейный функционал. Норма ограниченного линейного оператора

Из множества отображений $F : U \to V$ метрических пространств выделим пространство линейных операторов $L(U, V)$. При помощи этих операторов и обратных к ним строится качественная и количественная теория решения как линейных, так и нелинейных операторных уравнений.

Определение 2.22. Оператор $A : U \to V$ называется ***линейным***, если для любых α и β из R , а также всех u и v из U выполняется условие

$$A(\alpha u + \beta v) = \alpha A(u) + \beta A(v), \qquad (2.18)$$

где $A(u)$ и $A(v)$ принадлежат пространству V.

Для линейных операторов принято (где это не противоречит смыслу) опускать скобки аргумента. Тогда условие из определения 2.22 примет вид

$$A(\alpha u + \beta v) = \alpha Au + \beta Av. \qquad (2.19)$$

Определение 2.23. Линейный оператор A нормированных пространств U и V называется ***ограниченным***, если он определен на всем пространстве U и каждое ограниченное множество переводит снова в ограниченное.

Между ограниченностью и непрерывностью линейного оператора $A : U \to V$, где U и V – нормированные пространства, существует тесная связь:

- все ограниченные линейные операторы $A : U \to V$ непрерывны в U;
- все непрерывные в пространстве U линейные операторы ограничены.

То есть для линейного оператора, отображающего нормированное пространство в нормированное, ограниченность равносильна непрерывности. Тогда понятию «ограниченность» можно дать эквивалентное 2.23

Определение 2.23'. Линейный оператор A называется *ограниченным*, если он переводит всякий шар в ограниченное множество, то есть существует такая константа $Ç \in R$, что для $\forall u \in U$ выполняется неравенство

$$\| Au \| \le Ç \| u \|. \qquad (2.20)$$

Определение 2.24. Пусть C – множество всех констант ограниченного оператора A, удовлетворяющих неравенству (2.20). Наименьшее из чисел множества C называется ***нормой*** (точнее *верхней гранью*, так как норма ассоциируется с понятием пространство) оператора A и обозначается

$$\| A \| = \inf C. \qquad (2.21)$$

Теорема 2.2. Линейный оператор A нормированных пространств U и V ограничен тогда и только тогда, когда он ограничен на единичном шаре с центром в точке $0 \in U$, то есть если существует такое число $C < \infty$, что

$$\| Au \| \le C \| u \| \text{ при } u \in Q[0; 1].$$

□ Пусть линейный оператор A является ограниченным, тогда норма $\| Au \|_V$ его действия на любой элемент u из единичного шара $Q[0; 1] \subset U$ ограничена по определению 2.23.

Если же для $\forall u$ в шаре $\| u \| \le 1$ выполнено условие $\| Au \| \le C \| u \|$, а D – произвольное ограниченное подмножество U ($D \subset Q[0, R]$, $1 < R < \infty$), то для всех элементов $u_D \in D$ следует, что

$$\frac{u_D}{R} \in Q[0; 1] \text{ и } \| Au_D \| = \| R\, A\left(\frac{u_D}{R}\right) \| = R \| A\left(\frac{u_D}{R}\right) \| \le RC \| \frac{u_D}{R} \|.$$

Следовательно, оператор A является ограниченным. ◻

Теорема 2.3. Для любого ограниченного линейного оператора A

$$\| A \| = \sup_{\| u \| \le 1} \| Au \| = \sup_{u \ne 0} \frac{\| Au \|}{\| u \|}. \qquad (2.22)$$

□ Пусть $a = \sup_{\| u \| \le 1} \| Au \|$, тогда в силу линейности A справедливо равенство $a = \sup_{\| u \| \le 1} \| Au \| = \sup_{u \ne 0} \frac{\| Au \|}{\| u \|}$. Поэтому для всех $u \in U$ следует, что $\| Au \| \le a \| u \|$ и по определению $\| A \| = \inf C \le a$.

С другой стороны, из непрерывности оператора A вытекает, что для $\forall \varepsilon > 0$ существует такой элемент $u_\varepsilon \ne 0$, для которого

$$a - \varepsilon \le \frac{\| Au_\varepsilon \|}{\| u_\varepsilon \|} \text{ или } (a - \varepsilon) \| u_\varepsilon \| \le \| Au_\varepsilon \| \le \inf C \, \| u_\varepsilon \|.$$

Так как ε произвольное положительное число, то $\| A \| \ge a$. ◻

Следствие 1. $\| A \| = \sup_{\| u \| = 1} \| Au \|$.

Следствие 2. Если единичный шар $\|u\| \le 1$ компактен (например, в конечномерном нормированном пространстве U), то

$$\|A\| = \max_{\|u\|=1} \|Au\|. \qquad (2.23)$$

Определение 2.25. Пусть A и B – линейные операторы, действующие из U в V. Назовем ***суммой*** A и B оператор S, ставящий любому u из U в соответствие элемент $(Au + Bu) \in V$, и введем обозначение $S = A + B$.

Теорема 2.4. Если линейные операторы A и B ограничены, то оператор $A + B$ также ограничен, причем

$$\|A + B\| \le \|A\| + \|B\|.$$

Доказательство теоремы 2.4 следует из теоремы 2.3.

Если произведение оператора A на число k определить как оператор, ставящий любому элементу $u \in U$ в соответствие элемент $k(Au) \in V$, то справедлива следующая

Теорема 2.5. Множество $L(U, V)$ всех ограниченных линейных операторов, действующих из U в V, образует нормированное пространство с описанной выше (определение 2.24) операторной нормой.

Из всех аксиом нормы требует доказательства лишь вторая:

$$\|kA\| = \sup_{u \ne 0} \frac{\|A(ku)\| \cdot \|ku\|}{\|u\| \cdot \|ku\|} = \|A\| \cdot |k|. \qquad (2.24)$$

Определение 2.26. Пусть A и B – линейные операторы, действующие из U в V и из V в W соответственно. ***Произведением*** операторов A и B называется оператор BA, который каждому элементу $u \in U$ ставит в соответствие элемент $B(A(u)) = w \in W$.

Из определения следует, что в общем случае $BA \ne AB$.

Композиция линейных операторов BA сохраняет свойство линейности, так как для всех элементов u, v из U и любых α, β из R справедливо

$$(BA)(\alpha u + \beta v) = B(\alpha Au + \beta Av) = \alpha (BA)u + \beta (BA)v.$$

Теорема 2.6. Если A и B – ограниченные операторы, то и оператор BA ограничен, причем для норм справедливо соотношение

$$\|BA\| \le \|B\| \cdot \|A\|.$$

Доказательство. По определению и теореме 2.2 для всех $u \in U$

$$\|(BA)u\| = \|B(Au)\| \le \|B\| \, \|A\| \, \|u\|.$$

Рассмотрим один из наиболее важных частных случаев линейного оператора – линейный функционал.

Определение 2.27. Линейный оператор $B : U \to R$, отображающий нормированное пространство U в пространство R действительных чисел, называется ***линейным функционалом***.

Теорема 2.7. Линейный функционал $B : U \to R$ непрерывен тогда и только тогда, когда его значения на единичной сфере $\|u\| = 1$ ограничены в совокупности.

Определение 2.28. Вещественное число

$$\| B \| = \sup_{\| u \| = 1} | Bu |, \qquad (2.25)$$

то есть точную верхнюю грань значений $|Bu|$ на единичной сфере пространства U, назовем ***нормой функционала*** B.

В силу общих свойств линейных операторов линейный функционал B непрерывен в области определения U, если он непрерывен в какой-либо одной точке (например, в точке $u = 0$).

2.4 Конечномерный оператор. Алгебраические операции над матрицами. Матричный анализ

Для изучения свойств бесконечномерных **M**-пространств потребуются компьютерные технологии, оперирующие базисными функциями всюду плотных в них множеств многочленов. Это позволит конструировать операторы, действующие на непрерывные элементы этих пространств, а преобразования, обусловленные не более чем *счетным* числом действий, могут рассматриваться как продолжение операций конечномерной алгебры.

Таким образом, матричный анализ будет полезен и в этом случае, не говоря о том, что в настоящем «конечномерном» изложении он просто необходим. Мы предпочитаем термин «матричный анализ» термину «линейная алгебра», поскольку он верно охватывает широту приложений и методологию области исследований, то есть дает возможность «соединить» алгебру, анализ и геометрию бесконечномерных M-пространств через заданное в них скалярное (предскалярное) произведение элементов.

Определение 2.29. *Квадратной матрицей* A порядка z будем называть таблицу (массив) размером $z \times z$, состоящую из множества элементов

$$^{z}A \equiv \{a_{ij},\ i, j = s, \ldots, c\},\ z = c - s + 1,$$

где i и j – номера строки и столбца, которым принадлежит элемент a_{ij}.

Изучение свойств матрицы ${}^{z}\boldsymbol{A}$ начнем с установления условий биективности линейного оператора $\boldsymbol{A : R^{z} \to R^{z}}$. Этот оператор, а также конечномерный оператор $\boldsymbol{A : {}^{n}U \to {}^{n}V}$, связывающий пространства $\boldsymbol{P^{n}}$ (степенных) и/или $\boldsymbol{T^{n}}$ (тригонометрических) многочленов, будем обозначать как и матрицу ${}^{z}\boldsymbol{A}$. Рассматриваемые полиномиальные пространства размерности $\boldsymbol{z = n + 1}$ ($\boldsymbol{c = n}$) равны с точностью до изоморфизма

$$\boldsymbol{{}^{n}U \cong R^{z}} \text{ и } \boldsymbol{{}^{n}V \cong R^{z}}.$$

Первым необходимым условием взаимно однозначного соответствия оператора ${}^{z}\boldsymbol{A}$ является его суръективность. Вторым необходимым (а в совокупности с первым и достаточным) условием биективности ${}^{z}\boldsymbol{A}$ есть его инъективность. То есть при $\boldsymbol{{}^{z}A({}^{n}U) = {}^{n}V}$ (что всегда предполагается), доказательство биективности оператора ${}^{z}\boldsymbol{A}$ будет вытекать из его инъективности.

Термин «матрица» будет использован и для приближения операторов бесконечномерных **M**-пространств: в разделе 3.3 при анализе линейной зависимости произвольной системы векторов; в разделе 9.6 при определении коэффициентов многочлена Фурье аппроксимации элементов этих пространств. В обоих случаях разговор пойдет о матрице Грама, которая наиболее отчетливо характеризует связь линейных операторов конечномерных и бесконечномерных пространств.

Определение 2.30. ***Ядром*** матрицы ${}^{z}\boldsymbol{A}$ называется множество

$$\boldsymbol{J({}^{z}A) = \{{}^{z}u \in R^{z} \cong {}^{n}U : {}^{z}A\ {}^{z}u = 0\}}. \qquad (2.26)$$

Совокупность значений $\boldsymbol{{}^{z}v = {}^{z}A\ {}^{z}u}$ является подмножеством $\boldsymbol{R^{z} \cong {}^{n}V}$, а совокупность элементов ядра $\boldsymbol{J({}^{z}A)}$ – подмножеством $\boldsymbol{R^{z} \cong {}^{n}U}$. Таким образом, размерность (***rank***) ядра матрицы ${}^{z}\boldsymbol{A}$ находится по формуле

$$\boldsymbol{rank(J({}^{z}A)) = z - rank({}^{z}A(R^{z}))}.$$

Значит, инъективность (невырожденность) ${}^{z}\boldsymbol{A}$ следует из условия

$$\boldsymbol{rank(J({}^{z}A)) = 0}.$$

Матричные операции называются так же, как и операторные. Однако определяются они (в отличие от операторных) независимо от элементов, на которые эти матрицы действуют.

Сумму матриц ${}^{z}\boldsymbol{A}$ и ${}^{z}\boldsymbol{B}$, определяемую как покомпонентное сложение соответствующих элементов массивов размерности $\boldsymbol{z \times z}$, обозначим ${}^{z}\boldsymbol{C}$

$$\boldsymbol{{}^{z}C = {}^{z}A + {}^{z}B}, \text{ если } \boldsymbol{c_{ij} = a_{ij} + b_{ij}, i, j = s, \ldots, n}. \qquad (2.27)$$

Операция сложения матриц отвечает операции сложения линейных операторов, заданных относительно одной и той же пары базисов, и является коммутативной и ассоциативной. Нейтральным элементом сложения матриц есть *нулевая* матрица, обозначаемая $^{z}\mathbf{0}$.

Умножение матрицы $^{z}\boldsymbol{A}$ ***на число*** $\boldsymbol{k}$ есть умножение всех элементов матрицы на это число. Результат умножения запишем

$$^{z}\boldsymbol{B} = \boldsymbol{k}\,{}^{z}\boldsymbol{A}, \text{ если } \boldsymbol{b_{ij} = k\,a_{ij},\ i, j = s, \ldots, n}. \qquad (2.28)$$

Отметим, что множество $^{z}\boldsymbol{M}$ квадратных матриц порядка $\boldsymbol{z}$ является линейным пространством с операциями (2.27)–(2.28).

Произведением матриц $^{z}\boldsymbol{A}$ и $^{z}\boldsymbol{B}$ называется массив $^{z}\boldsymbol{C}$, который состоит из элементов, определяемых по формуле

$$\boldsymbol{c_{ij} = \langle (a_{ik},\ k = s, \ldots, n),\ (b_{kj},\ k = s, \ldots, n) \rangle,\ i, j = s, \ldots, n}, \qquad (2.29)$$

то есть элемент $\boldsymbol{c_{ij}}$ матрицы $^{z}\boldsymbol{C}$ равен скалярному произведению (в $\boldsymbol{R}^{z}$) $\boldsymbol{i}^{\text{-ой}}$ строки матрицы $^{z}\boldsymbol{A}$ на $\boldsymbol{j}^{\text{-ый}}$ столбец матрицы $^{z}\boldsymbol{B}$.

Операцию умножения матриц $^{z}\boldsymbol{A}$ и $^{z}\boldsymbol{B}$, выполненную по правилу (2.29), будем записывать в виде

$$^{z}\boldsymbol{A}\cdot{}^{z}\boldsymbol{B} = {}^{z}\boldsymbol{C} \text{ (или без точки)}. \qquad (2.30)$$

Нейтральным элементом умножения является *единичная* матрица $^{z}\boldsymbol{I}$. Операция умножения матриц ассоциативна, но не коммутативна.

Алгоритм умножения матриц базируется на встроенных друг в друга трех счетчиках, внутренний из которых вычисляет скалярное произведение векторов по формуле (2.29).

Часто в математике объект, определяемый многими параметрами, в значительной степени можно охарактеризовать с помощью одной величины. Определитель матрицы – пример характеристики такого рода.

Определение 2.31. ***Определителем*** (или ***детерминантом***) матрицы $^{z}\boldsymbol{A}$ называется число $\boldsymbol{det(}^{z}\boldsymbol{A)}$, которое находится по формуле

$$\boldsymbol{det(}^{z}\boldsymbol{A)} = \sum_{k=1}^{z!} \boldsymbol{sign}\,\sigma_k \prod_{i=s}^{n} a_{i\,\sigma_k(i)}, \qquad (2.31)$$

где $\boldsymbol{\sigma_k} = \{\boldsymbol{\sigma_k(s), \sigma_k(s+1), \ldots, \sigma_k(n)}\}$ – одна из всех $\boldsymbol{z!}$ перестановок по $\boldsymbol{z}$ чисел и $\boldsymbol{sign}\,\boldsymbol{\sigma_k}$ – знак перестановки, то есть множитель $\boldsymbol{+1}$ или $\boldsymbol{-1}$ в зависимости от того, четное или нечетное количество транспозиций необходимо совершить для того, чтобы от $\boldsymbol{\sigma_k}$ перейти к перестановке $\{\boldsymbol{s, s+1, \ldots, n}\}$.

При помощи реорганизации строк (столбцов) любую матрицу можно привести к наглядной и однозначно определяемой канонической форме, удобной для изучения проблем матричного анализа.

$$\Pi_1 = \begin{pmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ & & & \vdots & & & \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & c & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ & & & \vdots & & & \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

(i-ый стл — над i-м столбцом; i-ая стр — справа от i-й строки)

Рисунок 18

Опишем, опираясь на строковые преобразования, алгоритмы решения задач линейной алгебры: вычисление ранга и определителя матрицы; обращение невырожденной матрицы и определение корней системы линейных алгебраических уравнений (СЛАУ). Затем перейдем к задачам метрического анализа: вычисление собственных чисел и векторов матрицы; нахождение s-грани (нормы) и i-грани матрицы.

Практика работы со строками унаследована из теории решения СЛАУ, где k-ый элемент вектор-столбца правой части (y_k) есть линейная комбинация k-ых элементов вектор-столбцов матрицы ${}^z A$ с искомыми коэффициентами $\{x_s, \dots, x_n\}$

$$x_s a_{ks} + \dots + x_n a_{kn} = y_k,\ k = s, \dots, n.$$

Преобразование 1. *Умножение i-ой строки матрицы на число c.*

Чтобы i-ую строку матрицы ${}^z A$ умножить на c, умножим ${}^z A$ слева на матрицу Π_1 (рисунок 18), где c находится в ячейке (i, i).

Преобразование 2. *Перестановка i-ой и j-ой строк матрицы.*

$$\Pi_2 = \begin{pmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ & & & \vdots & & & & & & & \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ & & & \vdots & & & & & & & \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ & & & \vdots & & & & & & & \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

(j-ый стл — над j-м столбцом; i-ый стл — над i-м столбцом; j-ая стр. — справа от j-й строки; i-ая стр. — справа от i-й строки)

Рисунок 19

Для того чтобы переставить местами i-ю и j-ую строки матрицы ${}^z A$, достаточно умножить эту матрицу слева на матрицу Π_2 (рисунок 19). Матрица Π_2 состоит из диагональных единиц, кроме двух нулей, расположенных в ячейках (i, i) и (j, j), и недиагональных нулей, за исключением двух единиц, занимающих ячейки (i, j) и (j, i).

Преобразование 3. *Сложение i-ой строки с j-ой строкой, умноженной на число c.*

Чтобы к i-ой строке матрицы ${}^z A$ прибавить j-ую строку, предварительно умноженную на число c, умножим ${}^z A$ слева на матрицу Π_3 (рисунок 20), в которой c занимает ячейку (i, j).

$$\Pi_3 = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \overset{j\text{-ый стл}}{\downarrow} \\ \begin{pmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ & & & \cdot & & \cdot & & \cdot & & & \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ & & & \cdot & & \cdot & & \cdot & & & \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & c & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ & & & \cdot & & \cdot & & \cdot & & & \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{pmatrix} \end{array} \begin{array}{l} \\ \\ \\ j\text{-ая стр.} \\ \\ \\ \\ \\ i\text{-ая стр.} \\ \\ \\ \end{array}$$

Рисунок 20

Отметим, что матрицы любого из трех элементарных преобразований получаются в результате применения соответствующего преобразования к ${}^{z}\boldsymbol{I}$.

Изложим метод вычисления определителя невырожденной матрицы ${}^{z}\boldsymbol{A}$, идея которого заключается в приведении ${}^{z}\boldsymbol{A}$ с помощью преобразований $\boldsymbol{\Pi_1}$, $\boldsymbol{\Pi_2}$ и $\boldsymbol{\Pi_3}$ к единичной матрице ${}^{z}\boldsymbol{I}$. Все преобразования элементов матрицы ${}^{z}\boldsymbol{A}$ производятся в области памяти, отведенной под массив размерности $\boldsymbol{z \times z}$.

Пусть дана таблица, состоящая из элементов матрицы ${}^{n}\boldsymbol{A}$ с $\boldsymbol{s = 1}$:

$$ {}^{n}\boldsymbol{A} \equiv \begin{pmatrix} a_{11} \dots a_{1j} \dots a_{1n} \\ \cdot \quad \cdot \quad \cdot \\ a_{i1} \dots a_{ij} \dots a_{in} \\ \cdot \quad \cdot \quad \cdot \\ a_{n1} \dots a_{nj} \dots a_{nn} \end{pmatrix}. \qquad (2.32) $$

Первый этап:

- начиная с *первого* элемента *первого* столбца матрицы ${}^{n}\boldsymbol{A}$, ищем *разрешающий* элемент $\boldsymbol{a_{i1}}$ (например, первый элемент, отличный от нуля). Обозначим $\boldsymbol{d_1 \equiv a_{i1}}$ и разделим $\boldsymbol{i}^{\text{-ую}}$ строку (2.32) на $\boldsymbol{d_1}$ (преобразование $\boldsymbol{\Pi_1}$);
- если $\boldsymbol{i \neq 1}$, то меняем местами *первую* и $\boldsymbol{i}^{\text{-ую}}$ строки таблицы (преобразование $\boldsymbol{\Pi_2}$, перемещающее *разрешающую* $\boldsymbol{1}$ на диагональ);
- обнуляем все отличные от нуля элементы *первого* столбца матрицы ${}^{n}\boldsymbol{A}$ (преобразование $\boldsymbol{\Pi_3}$), кроме диагонального элемента $\boldsymbol{a_{11} = 1}$.

Второй этап:

- то же для *второго* столбца, начиная со *второй* строки ($\boldsymbol{d_2 \equiv a_{i2}}$);
- те же *действия* со *второй* и $\boldsymbol{i}^{\text{-ой}}$ строками таблицы;
- те же *действия* с элементами *второго* столбца матрицы, кроме $\boldsymbol{a_{22}}$.

И так далее для всех столбцов матрицы ${}^{n}\boldsymbol{A}$, включая последний. Определитель невырожденной матрицы находится по формуле

$$ \boldsymbol{det({}^{n}A) = (-1)^{t}\, d_1\, d_2 \dots d_n}, \qquad (2.33) $$

где $\boldsymbol{t}$ – число выполненных преобразований $\boldsymbol{\Pi_2}$ ($\boldsymbol{t < n}$).

Отметим, что ранг невырожденной матрицы равен ее порядку.

2.5 Метод Крамера определения корней линейных систем алгебраических уравнений (ЛСАУ)

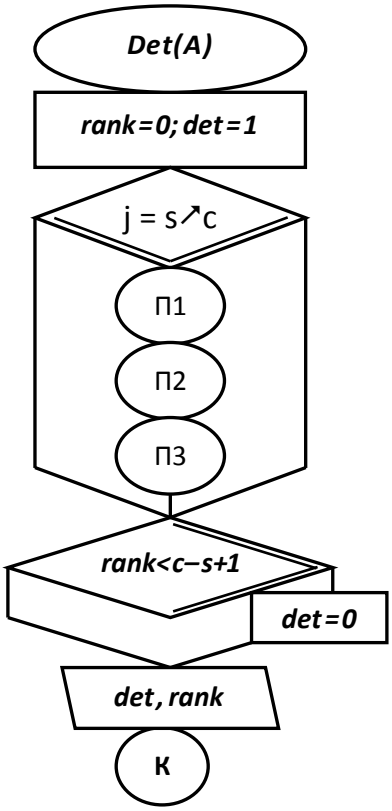


Рисунок 21

На рисунке 21 изображена объект-схема алгоритма вычисления ранга и определителя матрицы в виде функции ***Det(A)***. Этот алгоритм включает в себя три элементарных преобразования строк матрицы П1, П2 и П3 с изменением индексов от s до $n = c$ и $z = c - s + 1$, которые представлены на рисунке 22.

Теорема 2.8 . Определитель матрицы, равной произведению двух матриц ${}^{z}A$ и ${}^{z}B$, равен произведению определителей этих матриц

$$det({}^{z}A \cdot {}^{z}B) = det({}^{z}A)\ det({}^{z}B). \qquad (2.34)$$

Утверждение 2.6. В результате третьего преобразования определитель полученной матрицы не отличается от определителя исходной матрицы (эквивалентное преобразование); в результате второго преобразования значение определителя меняется на противоположное (***det := –det***); в результате первого преобразования значение определителя полученной матрицы изменяется в ***p1*** раз (***det := det·p1***).

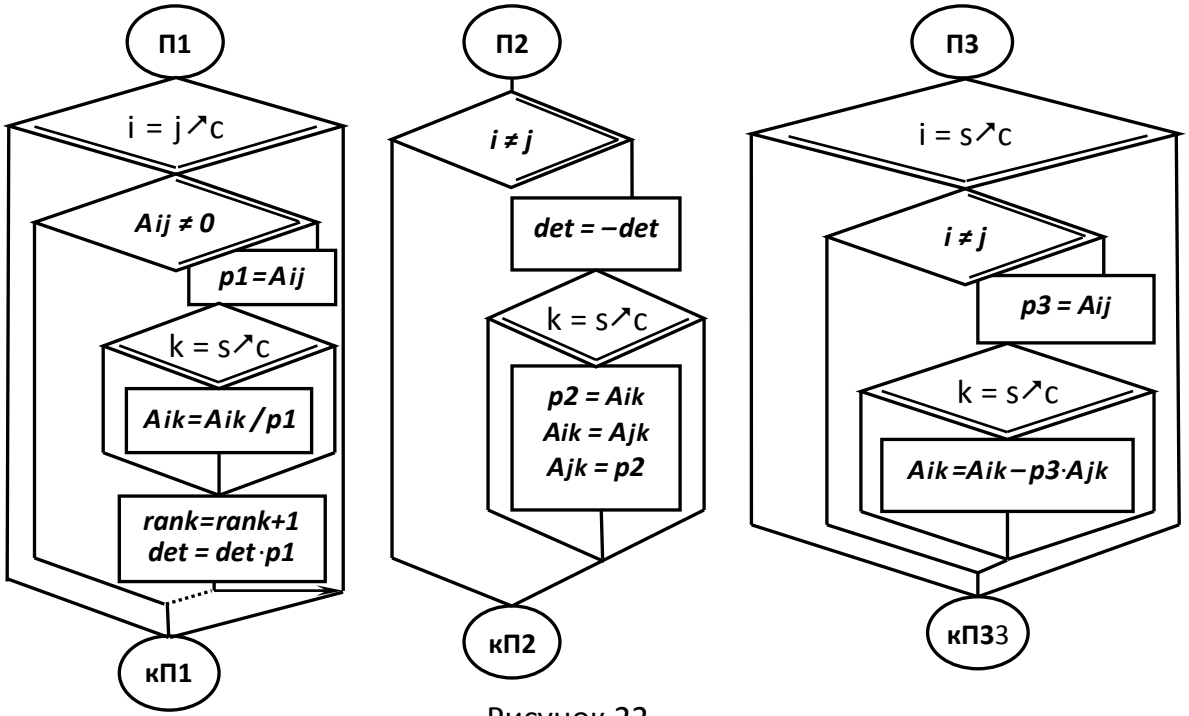


Рисунок 22

Если для преобразования матрицы использовать умножение ее на матрицы $\boldsymbol{П_1}$, $\boldsymbol{П_2}$ и $\boldsymbol{П_3}$ слева, то доказательство трех предложений утверждения следует из теоремы 2.8.

Решение линейных операторных уравнений часто сводится к вычислению корней систем линейных алгебраических уравнений (СЛАУ). В связи с тем, что нам предстоит решать и нелинейные алгебраические системы (где формулировка «система нелинейных алгебраических уравнений» не корректна), первые будем называть ЛСАУ, а вторые – НСАУ.

Пусть дана линейная система $\boldsymbol{n = c - s + 1}$ (введение параметров $\boldsymbol{c}$ и $\boldsymbol{s}$ связано с созданием программных модулей общего вида) алгебраических уравнений относительно $\boldsymbol{n}$ неизвестных $\{\boldsymbol{u_j, j = s, \ldots, c}\}$ с заданным в правой части вектором $\{\boldsymbol{v_i, i = s, \ldots, c}\}$ свободных членов:

$$\begin{cases} a_{ss}\,u_s + \ldots + a_{sj}\,u_j + \ldots + a_{sc}\,u_c = v_s; \\ \cdot \quad \cdot \quad \cdot \\ a_{is}\,u_s + \ldots + a_{ij}\,u_j + \ldots + a_{ic}\,u_c = v_i; \\ \cdot \quad \cdot \quad \cdot \\ a_{cs}\,u_s + \ldots + a_{cj}\,u_j + \ldots + a_{cc}\,u_c = v_c. \end{cases} \qquad (2.35)$$

Обозначим матрицу $\{a_{ij}\}_s^c$ коэффициентов при неизвестных как ${}^n\boldsymbol{A}$, вектор неизвестных $\{u_j\}_s^c \equiv {}^n\boldsymbol{u}$ и вектор свободных членов $\{v_i\}_s^c \equiv {}^n\boldsymbol{v}$. Здесь (и в некоторых других, обязательно оговариваемых случаях) индексы элементов матрицы $\boldsymbol{n}^{\text{-го}}$ порядка изменяются от $\boldsymbol{s = 1}$ до $\boldsymbol{c = n}$.

Если определитель матрицы ${}^n\boldsymbol{A}$ (обозначим $\boldsymbol{det({}^nA) \equiv D}$) отличен от $\boldsymbol{0}$, то линейное операторное уравнение ${}^n\boldsymbol{A}\,{}^n\boldsymbol{u} = {}^n\boldsymbol{v}$ можно решить методом Крамера. Найдем произведение матриц ${}^n\boldsymbol{A}$ и ${}^n\boldsymbol{U_j}$, где матрица ${}^n\boldsymbol{U_j}$ получена из единичной матрицы ${}^n\boldsymbol{I}$ заменой $\boldsymbol{j}^{\text{-го}}$ столбца вектор-столбцом ${}^n\boldsymbol{u}$.

Тогда произведение ${}^n\boldsymbol{A}\,{}^n\boldsymbol{U_j} = \begin{pmatrix} a_{ss} & \ldots & v_s & \ldots & a_{sc} \\ & \cdot & \cdot & \cdot & \\ a_{is} & \ldots & v_i & \ldots & a_{ic} \\ & \cdot & \cdot & \cdot & \\ a_{cs} & \ldots & v_c & \ldots & a_{cc} \end{pmatrix} \equiv {}^n\boldsymbol{A_j}$ и по теореме 2.8

$$\boldsymbol{det({}^nA)\,det({}^nU_j) = det({}^nA_j).} \qquad (2.36)$$

Так как $\boldsymbol{det({}^nU_j) = u_j}$ и $\boldsymbol{det({}^nA) = D}$, то справедлива формула

$$\boldsymbol{u_j = \frac{det({}^nA_j)}{D}, j = s, \ldots, c,} \qquad (2.37)$$

где матрица ${}^n\boldsymbol{A_j}$ получена путем замены $\boldsymbol{j}^{\text{-го}}$ столбца матрицы ${}^n\boldsymbol{A}$ вектор-столбцом ${}^n\boldsymbol{v}$. Такой способ вычисления координат вектора ${}^n\boldsymbol{u}$ матричного уравнения ${}^n\boldsymbol{A}\,{}^n\boldsymbol{u} = {}^n\boldsymbol{v}$ называется методом Крамера решения ЛСАУ (2.35).

Отметим, что требование $\det({}^n A) \neq 0$ является необходимым и достаточным условием существования единственного решения ЛСАУ.

При разработке алгоритма программы вычисления корней ЛСАУ (2.35) методом Крамера (с сохранением всех используемых двумерных массивов в одной области памяти ИС) для временного хранения изменяемых столбцов матриц ${}^n A_j$ будем применять массив-вектор ${}^n b$.

2.6 Обратная матрица. Матричный метод решения ЛСАУ. Метод Гаусса исключения неизвестных

Определение 2.32. Матрица ${}^n A$ называется ***невырожденной***, если этой матрицей в 0 переводится только 0, то есть ядро $J({}^n A) = \{0\}$. В противном случае матрица называется ***вырожденной***.

Определение 2.33. Матрица ${}^n A$ называется ***обратимой***, если существует такая матрица ${}^n B$, что ${}^n B\, {}^n A = {}^n I$. Матрица ${}^n B$ называется ***обратной*** к матрице ${}^n A$ и обозначается ${}^n A^{-1}$.

При разработке алгоритма и составлении программы вычисления обратной матрицы к невырожденной матрице ${}^n A$ будем использовать схему $({}^n A \mid {}^n I) \to ({}^n I \mid {}^n A^{-1})$, по которой единичная матрица ${}^n I$ переводится в обратную ${}^n A^{-1}$ с помощью дублирования элементарных преобразований ${}^n A \to {}^n I$ (см. объект-схему алгоритма вычисления определителя матрицы ${}^n A$ из раздела 2.5).

Введем еще одно преобразование матрицы ${}^n A$, которое осуществляет симметричное относительно главной диагонали отображение элементов.

Определение 2.34. Матрица ${}^n C = \{c_{ij},\ i, j = s, \ldots, n\}$ называется ***транспонированной к*** ${}^n A$, если $c_{ij} = a_{ji},\ i, j = s, \ldots, n$, и обозначается ${}^n A^T$.

Утверждение 2.7. Для операции транспонирования матриц справедливы следующие основные свойства:

1) $({}^n A + {}^n B)^T = {}^n A^T + {}^n B^T$;
2) $(a\ {}^n A)^T = a\ {}^n A^T$;
3) $({}^n A\ {}^n B)^T = {}^n B^T\ {}^n A^T$;
4) $({}^n A^T)^T = {}^n A$.

Доказательство формул утверждения 2.7 следует из определения 2.34.

Утверждение 2.8. Для невырожденной матрицы ${}^n A$ верно равенство

$$({}^n A^T)^{-1} = ({}^n A^{-1})^T.$$

Из соотношения ${}^n A^{-1} \cdot {}^n A = {}^n I$ вытекает, что $({}^n A^{-1} \cdot {}^n A)^T = {}^n I$. Тогда, следуя формуле 3 утверждения 2.7, $({}^n A^T) \cdot ({}^n A^{-1})^T = {}^n I$.

Приведем равносильные высказывания относительно обратимости ${}^n A$:

- матрица ${}^n A$ невырожденная;
- существует обратная к ${}^n A$ матрица ${}^n A^{-1}$;

- ранг и порядок матрицы ${}^{n}\boldsymbol{A}$ совпадают $\boldsymbol{rank({}^{n}A) = c - s + 1}$;
- ядро ${}^{n}\boldsymbol{A}$ состоит из одного элемента $\boldsymbol{ker({}^{n}A) = \{0\}}$;
- определитель ${}^{n}\boldsymbol{A}$ отличен от нуля $\boldsymbol{det({}^{n}A) \neq 0}$;
- собственные значения матрицы ${}^{n}\boldsymbol{A}$ отличны от нуля; (2.38)
- $\boldsymbol{i}$-грань матрицы ${}^{n}\boldsymbol{A}$ больше нуля $\boldsymbol{\lfloor {}^{n}A \rfloor > 0}$;
- размерность множества значений матрицы ${}^{n}\boldsymbol{A}$ равна $\boldsymbol{c - s + 1}$;
- вектор-столбцы матрицы ${}^{n}\boldsymbol{A}$ линейно независимы;
- вектор-строки матрицы ${}^{n}\boldsymbol{A}$ линейно независимы.

Рассматривая вектор-столбцы $\boldsymbol{{}^{n}u = (u_s, \ldots, u_c)^T}$ и $\boldsymbol{{}^{n}v = (v_s, \ldots, v_c)^T}$ как транспонированные соответствующие им вектор-строки, запишем соотношение (2.35) в матричной форме

$$\boldsymbol{{}^{n}A\ {}^{n}u = {}^{n}v}. \qquad (2.39)$$

В дальнейшем, где в этом не возникает необходимости, будем опускать символ $\boldsymbol{T}$, обозначающий транспонирование вектора. Домножим левую и правую части (2.39) слева на обратную матрицу ${}^{n}\boldsymbol{A^{-1}}$ к невырожденной матрице ${}^{n}\boldsymbol{A}$ и, используя определение обратной матрицы, получим

$$\boldsymbol{{}^{n}u = {}^{n}A^{-1}\ {}^{n}v}. \qquad (2.40)$$

Матричная форма (2.39) записи уравнения (2.35) – аналог операторной формы (6.24). Все выкладки из теории решения линейных операторных уравнений справедливы и для решения (2.35) с $\boldsymbol{A = {}^{n}A}$.

Рассмотрим наряду с системой (2.35) преобразование $\boldsymbol{{}^{n}\tilde{A}\,{}^{n}\tilde{u} = {}^{n}\tilde{v}}$, соответствующее тому же оператору $\boldsymbol{A}$, но в другом базисе $\boldsymbol{R^{n}}$. Пусть ${}^{n}\boldsymbol{T}$ – матрица перехода от одного базиса к другому, тогда

$$\boldsymbol{{}^{n}u = {}^{n}T\ {}^{n}\tilde{u}} \quad \text{и} \quad \boldsymbol{{}^{n}v = {}^{n}T\ {}^{n}\tilde{v}}$$

или

$$\boldsymbol{{}^{n}\tilde{v} = {}^{n}T^{-1} \cdot {}^{n}A \cdot {}^{n}T\ {}^{n}\tilde{u}}, \qquad (2.41)$$

что говорит о существовании целого класса матриц, представляющих данный оператор в различных базисах пространства $\boldsymbol{R^{n}}$.

Определение 2.35. Две матрицы ${}^{n}\boldsymbol{A}$ и ${}^{n}\boldsymbol{B}$, связанные соотношением

$$\boldsymbol{{}^{n}B = {}^{n}T^{-1} \cdot {}^{n}A \cdot {}^{n}T}, \qquad (2.42)$$

где ${}^{n}\boldsymbol{T}$ – некоторая невырожденная матрица, называются ***подобными***.

Непосредственно из определения 2.35 и теоремы 2.8 следует, что подобные матрицы имеют равные определители.

Часто при решении задач, в алгоритмах которых используются корни уравнения (2.35), нет необходимости применять обратную матрицу. В этом случае более удобен метод Гаусса (Гаусса-Жордана), который частично дублирует методику обращения матрицы, однако по количеству операций значительно быстрее матричного метода находит решение системы уравнений (2.35) при помощи реорганизации расширенной таблицы

$$({}^{n}A \mid {}^{n}v) \equiv \left(\begin{array}{ccccc|c} a_{ss} & \dots & a_{sj} & \dots & a_{sc} & v_s \\ & \cdot & \cdot & \cdot & & \dots \\ a_{is} & \dots & a_{ij} & \dots & a_{ic} & v_i \\ & \cdot & \cdot & \cdot & & \dots \\ a_{cs} & \dots & a_{cj} & \dots & a_{cc} & v_c \end{array}\right). \qquad (2.43)$$

Идея метода – приведение элементарными строковыми преобразованиями П1, П2 и П3 матрицы ${}^{n}A$ к треугольному виду с соответствующими преобразованиями вектор-столбца ${}^{n}v$.

В случае разрешимости системы (2.35) первая часть метода (***прямой ход***) заканчивается расширенной таблицей

$$\left(\begin{array}{ccccc|c} 1 & \dots & \tilde{a}_{sj} & \dots & \tilde{a}_{sc} & \tilde{v}_s \\ & \cdot & \cdot & \cdot & & \dots \\ 0 & \dots & 1 & \dots & \tilde{a}_{ic} & \tilde{v}_i \\ & \cdot & \cdot & \cdot & & \dots \\ 0 & \dots & 0 & \dots & 1 & \tilde{v}_c \end{array}\right). \qquad (2.44)$$

Вторая часть – процесс нахождения корня системы (2.35) по таблице (2.44) – называется ***обратным ходом*** метода Гаусса. Суть обратного хода метода Гаусса заключается в вычислении значений координат вектора ${}^{n}u = \{u_s, \dots, u_c\}$, начиная с последней координаты u_c.

Если в качестве разрешающего элемента (см. алгоритм вычисления определителя из раздела 2.4) использовать наибольший по модулю элемент столбца, то получим метод Гаусса решения ЛСАУ с выбором *главного элемента*. Объект-схема алгоритма поиска корней системы (2.35) методом Гаусса с выбором главного элемента приведена в приложении 1.

Исходя из различия идей решения ЛСАУ, метод Гаусса исключения неизвестных отличается от матричного метода (2.40) тем, что в нем:

1) нет необходимости приводить матрицу ${}^{n}A$ к единичной и поэтому преобразованием Π_3 достаточно *обнулить* только те *элементы* столбцов, которые расположены *ниже главной диагонали* матрицы;

2) преобразование строк матрицы левой части таблицы влечет аналогичные действия с элементами *одного* вектор-столбца правой части.

Глава 3 Полные нормированные (банаховы) пространства

3.1 Определение и примеры банаховых пространств. *B*-пространство $C_{[a, b]}$ непрерывных функций

Часто при описании закона, по которому в нормированном пространстве U находится точное решение операторного уравнения (*), используется итерационная последовательность

$$\{u_k, k = 0, 1, \ldots\} \tag{3.1}$$

приближенных значений корня уравнения.

Естественные требования, предъявляемые к этой последовательности приближений, следующие:

1) все члены последовательности (3.1) вместе с ее пределом

$$\lim_{k \to \infty} u_k = u^* \tag{3.2}$$

должны принадлежать множеству U;

2) она должна сходиться к корню уравнения (*), то есть

$$F(u^*) = v. \tag{3.3}$$

Добиться такого существенного продвижения в области анализа (условие 1) можно, лишь несколько сузив класс рассматриваемых нормированных пространств, налагая на их норму условие *полноты*. Предположение полноты вносит заметное упрощение в абстрактный анализ, и в то же время ему удовлетворяет широкий класс нормированных линейных пространств.

Определение 3.1. Последовательность $\{u_k, k = 0, 1, \ldots, K, \ldots\}$ называется ***фундаментальной*** (или последовательностью ***Коши***) в нормированном пространстве U, если для любого числа $\varepsilon > 0$ можно указать такое $K \in N$, что при всех l и m больше K будет выполняться неравенство

$$\| u_l - u_m \| < \varepsilon. \tag{3.4}$$

Определение 3.2. Метрическое пространство Q, в котором любая фундаментальная последовательность сходится, называется ***полным***. Полные метрические пространства будем обозначать символом M.

Утверждение 3.1. Последовательность (3.1) является фундаментальной в нормированном пространстве U, если

- она сходится в пространстве U;

- каждый элемент u_k, $k = 0, 1, \ldots$ последовательности принадлежит U и существует такое число $q < 1$, что при любом $k \in N$ верно неравенство

$$\|u_{k+1} - u_k\| \le q\,\|u_k - u_{k-1}\|. \qquad (3.5)$$

Теоретические исследования свойств полных нормированных пространств были систематизированы и получили дальнейшее развитие в трудах польского математика С. Банаха.

Определение 3.3. Полное нормированное пространство называется ***банаховым*** (или ***B***-пространством).

Примерами ***B***-пространств могут служить множества R^n и C^n, если норму в действительном и комплексном n-мерных E-пространствах определить по обобщенной формуле Пифагора

$$\|x\| = \sqrt{\sum_{i=1}^{n} |x_i|^2}\ . \qquad (3.6)$$

Теорема 3.1. Пространство l_2 бесконечных числовых последовательностей с суммируемым квадратом является полным по норме

$$\|u\| = \left(\sum_{n=0}^{\infty} |u_n|^2\right)^{1/2}. \qquad (3.7)$$

Корректность введения в пространстве l_2 естественной нормы по формуле (3.7) следует из двух неравенств [10, с. 60–64]:

1) $\sum_{n=0}^{\infty} |u_n v_n| \le \left(\sum_{n=0}^{\infty} u_n^2\right)^{1/2} \cdot \left(\sum_{n=0}^{\infty} v_n^2\right)^{1/2}$ (Гёльдера или Коши-Буняковского);

2) $\left(\sum_{n=0}^{\infty} (u_n + v_n)^2\right)^{1/2} \le \left(\sum_{n=0}^{\infty} u_n^2\right)^{1/2} + \left(\sum_{n=0}^{\infty} v_n^2\right)^{1/2}$ (Минковского).

Множество непрерывных функций на отрезке $[a, b]$, в зависимости от способа задания в нем нормы (раздел 2.1) может быть представлено как $C_{[a,b]}$, $C_{1[a,b]}$ или $C_{2[a,b]}$. Но только $C_{[a,b]}$ из них является ***B***-пространством по своей норме. Следовательно, для того чтобы любая фундаментальная последовательность функций из $C_{1[a,b]}$ или $C_{2[a,b]}$ сходилась в своем пространстве, необходимо каждое из них пополнить предельными элементами.

Очевидно, эти элементы не являются непрерывными функциями, поэтому норма в расширенных пространствах должна быть определена и для разрывных функций. Сделать это можно, применив в формуле задания нормы вместо римановского интеграла лебеговский. Пополненные таким образом пространства обозначаются $L^1_{[a, b]}$ и $L^2_{[a, b]}$ соответственно.

3.2 Мера Лебега. Суммируемость (интегрируемость) по Лебегу. *B*-пространство $L^1_{[a,\,b]}$ суммируемых функций

Один из важнейших классов полных нормированных пространств, используемых при решении функциональных уравнений с интегральным оператором, составляют пространства *суммируемых* (в литературе используется также термин «интегрируемых») по Лебегу функций. В первую очередь, это пространство $L^1_{[a,\,b]}$ всех суммируемых функций и пространство $L^2_{[a,\,b]}$ функций с суммируемым квадратом, являющихся расширением описанных ранее пространств $C_{1\,[a,\,b]}$ и $C_{2\,[a,\,b]}$.

Чтобы ввести понятие суммируемости функций из пространств L^1_X и L^2_X, определим термин «мера» на множестве X, где они заданы. Затем классифицируем элементы и рассмотрим свойства этих пространств.

Определение 3.4. Пусть $Ç$ – σ-алгебра множества X, то есть такая непустая система подмножеств X, что

1) для любых элементов A и B, принадлежащих $Ç$, справедливо

$$(A \,\Delta\, B) \in Ç \text{ и } (A \cap B) \in Ç;$$

2) для всех последовательностей $\{Ç_0, Ç_1, \ldots, Ç_n, \ldots\} \subset Ç$ следует

$$\bigcup_{n=0}^{\infty} Ç_n \in Ç.$$

Тогда ***мерой*** множества Y из σ-алгебры $Ç$ называется действительная неотрицательная функция $\mu(Y)$, для которой выполняются условия:

1) $\mu(Y) \geq 0$, причем $\mu(Y) = 0$, только если $Y = \varnothing$;

2) $\mu\left(\bigcup_{n=0}^{\infty} Ç_n\right) = \sum_{n=0}^{\infty} \mu(Ç_n)$ для всех попарно непересекающихся множеств $Ç_n$, $n = 0, 1, 2, \ldots$ (то есть $Ç_i \cap Ç_j = \varnothing$ при $i \neq j$).

Определение 3.5. Множество $A \subset X$ называется ***измеримым*** (в смысле Лебега), если для любого $\varepsilon > 0$ найдется такое множество $B \in Ç$, что

$$\mu(A \,\Delta\, B) < \varepsilon.$$

Функция μ, рассматриваемая на измеримых по Лебегу множествах, называется ***лебеговой мерой***. В дальнейшем изложении будут рассматриваться только измеримые мерой Лебега множества X.

Теорема 3.2. Объединение и пересечение двух измеримых множеств есть измеримые множества.

Доказательство теоремы 3.2 сразу следует из определения 3.5.

Определение 3.6. Мера μ называется ***мерой со счетным базисом***, если существует такая счетная система $Ç = \{Ç_n, n = 0, 1, 2, \ldots\}$ измеримых подмножеств X, что для всякого измеримого $A \subset X$ и любого $\varepsilon > 0$ найдется подмножество $Ç_n$, при котором станет справедливым неравенство

$$\mu(A \Delta Ç_n) < \varepsilon. \tag{3.8}$$

Определение 3.7. Пусть существует такая система попарно непересекающихся элементов $C = \{C_0, \ldots, C_n, \ldots\} \subset Ç$, что $\mu(X \setminus \bigcup_n C_n) = 0$ и $\mu(C_n) > 0$, $n = 0, 1, \ldots$. Тогда систему C назовем ***базисом*** меры μ на множестве X.

При решении уравнения (*) функциональными методами счетный базис меры μ будем исчерпывать конечной системой из $n+1$ элемента ${}^nC$, удовлетворяющей соотношениям определения 3.7 для $\forall n \in N$. Описанная *дискретизация* базиса меры позволит проектировать бесконечномерные функциональные пространства, заданные на множестве X, в конечномерные.

Определение 3.8. Пусть $Ç$ – некоторая σ-алгебра множества X. Действительная функция $u(x)$, $x \in X$ называется ***измеримой*** на множестве X, если для любого числа $\alpha \in R$ измеримо множество

$$X_\alpha = \{x : u(x) < \alpha\}.$$

Определение 3.9. Функция $u(x)$, определенная на множестве X, называется ***простой***, если она измерима и принимает не более чем счетное множество значений $\{u_n, n = 0, 1, 2, \ldots\}$.

Определение 3.10. Простая функция $u(x)$, определенная на множестве X с мерой μ, называется ***суммируемой***, если

$$\int_X |u(x)|\, d\mu = \sum_n |u_n|\, d\mu\{x : u(x) = u_n\} < \infty.$$

Функция $u(x)$ суммируема на X, если существует последовательность простых суммируемых функций, равномерно сходящаяся к $u(x)$.

Определение 3.11. Если $u(x)$ суммируема на X, то интегралом Лебега функции $u(x)$ по множеству X с мерой μ называется сумма ряда

$$\int_X u(x)\, d\mu = \sum_n u_n\, d\mu\{x : u(x) = u_n\}.$$

Определение 3.12. Две функции u и v, заданные на одном и том же измеримом множестве X, называются ***эквивалентными***, если

$$\mu\{x : u(x) \neq v(x)\} = 0.$$

Определение 3.13. ***Пространством*** L_X^1 называется нормированное пространство, элементами которого служат классы эквивалентных между собой суммируемых функций с обычными операциями сложения функций и умножения их на число. Норма элемента $u(x) \in L_X^1$ равна

$$\|u(x)\| = \int_X |u(x)|\, d\mu. \qquad (3.9)$$

Сходимость последовательности $\{u_n(x), n = 0, 1, 2, \ldots\}$ к функции $u(x)$ по норме L_X^1, определяемая условием

$$\lim_{n\to\infty} \int_X |u_n(x) - u(x)|\, d\mu = 0,$$

называется ***сходимостью в среднем***.

Теорема 3.3 [10, с. 431]. Пространство L_X^1 полное, то есть банахово.

Теорема 3.4 [10, с. 433]. Пространство C_{1X} непрерывных функций на множестве X с мерой $\mu(X) < \infty$ всюду плотно в L_X^1.

Теорема 3.5. Функция $u(x), x \in X$, принимающая не более чем счетное число различных значений $\{u_0, \ldots, u_n, \ldots\}$, измерима тогда и только тогда, когда измеримы все множества

$$u^{-1}(u_n) = \{x : u(x) = u_n\}.$$

Теорема 3.6 [10, с. 433]. Пространство l_X^1 простых суммируемых на множестве X функций всюду плотно в L_X^1.

Теорема 3.7 [10, с. 435]. Пусть μ – мера со счетным базисом множества X, тогда в пространстве L_X^1 существует счетное всюду плотное множество функций.

Если $X = [a, b]$, а μ – обычная мера Лебега на отрезке, то такими счетными всюду плотными множествами функций в пространстве $L_{[a, b]}^1$ являются множества степенных и тригонометрических многочленов с рациональными коэффициентами [10, с. 436].

3.3 Гильбертово пространство. *H*-пространство l_2 бесконечных числовых последовательностей

Перейдем к рассмотрению пространств, которые являются B и E-пространствами одновременно. Систематизированному изучению линейных бесконечномерных пространств, полных по естественной норме, посвящены труды немецкого ученого Д. Гильберта.

Определение 3.14. Бесконечномерное евклидово пространство, полное по естественной норме, называется ***гильбертовым*** и обозначается H.

Определение 3.15. Бесконечномерное банахово пространство с определенным в нем скалярным произведением будем называть ***предгильбертовым*** и обозначать символом Γ.

Гильбертово пространство – это классический пример абстрактного бесконечномерного пространства, имеющего практическое применение в прикладных дисциплинах. Наличие скалярного произведения значительно обогащает его геометрические свойства. Возможность введения в гильбертовом пространстве понятия перпендикулярности двух элементов позволяет трансформировать основные идеи евклидовой геометрии в современную теорию геометрии бесконечномерных пространств. Это обстоятельство имеет особое значение при изучении базисов гильбертовых пространств, так как в банаховых пространствах такой возможности нет.

Определение 3.16. Линейно независимую систему элементов

$$\boldsymbol{\Phi = \{\varphi_n,\ n = 0, 1, \ldots\}}$$

банахова пространства $\boldsymbol{B}$ будем называть ***полной***, если порожденное ею замкнутое векторное подпространство есть все $\boldsymbol{B}$.

Другими словами, система $\boldsymbol{\Phi}$ полна, если линейное многообразие над этой системой всюду плотно в $\boldsymbol{B}$, то есть каждый элемент $\boldsymbol{B}$-пространства с любой точностью можно приблизить конечной линейной комбинацией элементов из $\boldsymbol{\Phi}$. В связи с этим метрический анализ банаховых пространств осуществляется с помощью свойства сепарабельности.

Определение 3.17. $\boldsymbol{B}$-пространство называется ***сепарабельным***, если оно содержит счетное всюду плотное множество. То есть такое множество

$$\boldsymbol{\Psi = \{\psi_n,\ n = 0, 1, 2, \ldots\}}, \qquad (3.10)$$

что для каждого элемента $\boldsymbol{u \in B}$ и любого $\boldsymbol{\varepsilon > 0}$ найдется элемент $\boldsymbol{\psi_n \in \Psi}$, с которым будет справедливо неравенство $\boldsymbol{\| u - \psi_n \| < \varepsilon}$.

Понятие полной системы в бесконечномерных $\boldsymbol{B}$-пространствах не равносильно понятию базиса Шаудера, как метрической компоненты, присущей конечномерным $\boldsymbol{B}$-пространствам и бесконечномерным $\boldsymbol{H}$-пространствам. С различиями этих понятий можно ознакомиться в разделе 6.2.

Определение 3.18. Базис называется ***нормированным***, если норма каждого из его элементов равна одному. Базис называется ***ортогональным***, если скалярное произведение любых двух его элементов равно нулю.

В качестве примера $\boldsymbol{H}$-пространства рассмотрим множество $\boldsymbol{l_2}$ бесконечных числовых последовательностей $\boldsymbol{u = (u_0, u_1, \ldots, u_n, \ldots)}$ с суммируемым квадратом и обычными операциями сложения последовательностей и умножения их на число (см. раздел 1.1). Определим скалярное произведение двух векторов $\boldsymbol{u}$ и $\boldsymbol{v}$ этого пространства по формуле

$$\boldsymbol{\langle u, v \rangle = \sum_{n=0}^{\infty} u_n v_n}\,. \qquad (3.11)$$

Теорема 3.8. ***B***-пространство l_2 со скалярным произведением, заданным в виде (3.11), является гильбертовым.

Бесконечная линейно независимая система векторов

$$e_0 = (1, 0, 0, \ldots, 0, \ldots),$$
$$e_1 = (0, 1, 0, \ldots, 0, \ldots),$$
$$\ldots, \qquad (3.12)$$
$$e_n = (0, 0, 0, \ldots, 1, \ldots),$$
$$\ldots$$

образует простейший ортогональный (относительно скалярного произведения (3.11)) нормированный базис l_2. Система (3.12) полная, так как для $\forall u = (u_0, \ldots, u_n, u_{n+1}, \ldots) \in l_2$ найдется такой вектор ${}^n u = (u_0, \ldots, u_n, 0, \ldots)$ – линейная комбинация $\{e_0, \ldots, e_n, \ldots\}$, что по норме $\|u\| = \langle u, u\rangle^{1/2}$

$$\lim_{n\to\infty} \|{}^n u - u\| = 0.$$

Таким образом, l_2 – ***E***-пространство с базисом (3.12).

Подробное доказательство выполнения в l_2 всех аксиом гильбертова пространства, а также классификация произвольных систем векторов вида

$$\{\tilde{e}_0, \tilde{e}_1, \ldots, \tilde{e}_n, \ldots\} \qquad (3.13)$$

проводится с помощью матрицы Грама.

Определение 3.19. ***Матрицей Грама*** системы элементов (3.12) гильбертова пространства H называется множество (таблица) значений

$$G = \{a_{ij}, i, j = 0, 1, \ldots, n, \ldots\}, \text{ где } a_{ij} = \langle a_i, a_j\rangle.$$

В силу свойств скалярного произведения матрица Грама симметрична.

Определение 3.20. Ортогональный нормированный базис гильбертова пространства называется ***ортонормированным***.

Теорема 3.9. Для любого линейного ограниченного функционала B в гильбертовом пространстве H существует единственный элемент v с нормой $\|v\| = \|B\|$ такой, что

$$Bu = \langle u, v\rangle, \ \forall u \in H. \qquad (3.14)$$

Доказательство теоремы, основанно на применении теоремы о проекции.

С точки зрения приложений H-пространство уникально еще и тем, что в нем можно дать простое определение оператора A^*, сопряженного к линейному оператору A ($\langle Au, v\rangle = \langle u, A^*v\rangle$, $\{u, v\} \subset H$). На основе самосопряженности оператора ($A^* = A$) развита мощная теория решения функциональных уравнений. К сожалению, это свойство присуще узкому классу линейных операторов, поэтому мы не будем налагать условие самосопряженности на операторы линейных уравнений, к которым сводится решение нелинейных интегро-дифференциальных краевых задач общего вида.

3.4 H-пространство $L^2_{[a,b]}$ функций с суммируемым квадратом

Прежде чем доказывать, что пополнение множества $C_{2[a,b]}$ (раздел 2.1), именуемое пространством $L^2_{[a,b]}$ функций с суммируемым квадратом на отрезке $[a, b]$, является гильбертовым, перечислим аксиомы H-пространства.

1. H – бесконечномерное линейное пространство, то есть в нем для любого $n \in N$ можно найти систему из n линейно независимых элементов.

2. В пространстве H задано скалярное произведение $\langle , \rangle$, определяющее естественную норму произвольного элемента u этого пространства

$$\|u\| = \langle u, u\rangle^{1/2}.$$

3. Пространство H полно в смысле метрики

$$\rho(u, v) = \|u - v\|, \text{ где } u \text{ и } v \text{ – любые элементы из } H. \qquad (3.15)$$

4. H – сепарабельно, то есть содержит счетное всюду плотное множество.

Последняя аксиома для гильбертовых пространств не является обязательной, так как полнота H-пространства по его естественной норме часто достаточна для построения в H ортогонального базиса Шаудера.

Опишем классы эквивалентных функций на множестве X с $\mu(X) < \infty$, которые являются элементами пространства L^2_X, и введем понятие скалярного произведения элементов этого пространства.

Определение 3.21. Действительная функция $u(x)$, определенная на множестве X с мерой μ, называется функцией с суммируемым квадратом, если существует интеграл

$$\int_X u^2(x)\, d\mu < \infty. \qquad (3.16)$$

Определение 3.22. Пространством L_X^2 называется функциональное ***E***-пространство со скалярным произведением

$$\langle u, v\rangle = \int_X u(x)v(x)\,d\mu, \qquad (3.17)$$

элементами которого являются классы эквивалентных функций с суммируемым квадратом. Операции сложения элементов и умножения их на число в L_X^2 определяются как и для обычных функций (см. раздел 1.1).

Утверждение 3.2. Соотношение (3.17) удовлетворяет всем требованиям определения скалярного произведения элементов из раздела 2.1.

Задав в L_X^2 скалярное произведение, мы тем самым определили для функций этого пространства понятия нормы и расстояния, а для последовательностей функций – понятие сходимости.

Определение 3.23. Сходимость последовательности $\{u_n(x), n = 0, 1, \ldots\}$ к функции $u(x)$, определяемая в L_X^2 условием

$$\lim_{n\to\infty} \int_X (u_n(x) - u(x))^2\,d\mu = 0,$$

называется сходимостью ***в среднем квадратичном***.

Теорема 3.10. Пусть мера μ на множестве X имеет счетный базис, тогда в пространстве L_X^2 существует счетное всюду плотное множество.

Если $X = [a, b]$, то меру Лебега произвольного множества $A = \bigcup_{n=0}^{\infty} C_n$ для попарно непересекающихся $C_n \subset X$, $n = 0, 1, \ldots$ определим как

$$\mu(A) = \sum_{n=0}^{\infty} \mu(C_n),$$

где $\mu(C_n)$ – длина элементарных интервалов $C_n = \{a_n, b_n\}$, $a_{n+1} = b_n \in Q$ с рациональными окончаниями, на которые разбит отрезок $[a, b]$.

Из счетности множества Q следует, что мера μ на отрезке $[a, b]$ имеет счетный базис, а счетным всюду плотным множеством в $L_{[a, b]}^2$ является пространство $l_{[a, b]}^2$ простых функций с суммируемым квадратом.

Множества тригонометрических и степенных многочленов с рациональными коэффициентами также всюду плотны в $L_{[a, b]}^2$ ввиду того, что:

1) множество $C_{2[a, b]}$ является всюду плотным в $L_{[a, b]}^2$;

2) полнота систем тригонометрических и степенных полиномов в пространстве $C_{2[a, b]}$ вытекает из теорем Фейера [10, с. 476] и Вейерштрасса [10, с. 480] о *равномерной* (а значит, и среднеквадратичной) аппроксимации любой непрерывной на отрезке функции многочленами.

Теорема 3.11 [10, с. 439]. ***E***-пространство $L^2_{[a,\,b]}$ полно.

Следовательно, пространство $L^2_{[a,\,b]}$ является гильбертовым.

Отметим, что ***B***-пространство $L^1_{[a,\,b]}$ суммируемых функций таковым не является, так как его норму нельзя задать с помощью какого бы то ни было скалярного произведения.

3.5 Аппроксимация функций из $L^2_{[a,b]}$ тригонометрическими и степенными многочленами

Кроме линейной независимости, на полную систему элементов пред-гильбертовых пространств часто накладывают еще ряд условий. В частности, требуют ортогональности и нормированности ее элементов. Рассмотрим процесс ортогонализации полных систем функций, представляющих тригонометрические и степенные многочлены (или полиномы).

Определение 3.24. Функция вида

$$a_0 + \sum_{i=1}^{m}(a_i \cos i x + b_i \sin i x) \equiv {}^{2m}t(x) \in T^{2m}, \qquad (3.18)$$

где a_0, a_i и b_i ($i = 1, \ldots, m$) – произвольные действительные числа, называется ***тригонометрическим*** полиномом порядка дискретизации $z = 2m+1$ (по числу базисных функций) или многочленом гармонического порядка m.

Тригонометрический полином есть функция периода 2π, и при его изучении достаточно ограничиться рассмотрением изменения независимой переменной x на множествах $(0, 2\pi]$ или $(-\pi, \pi]$.

Теорема 3.12 . Тригонометрический полином порядка n ($n = 2m$), у которого не все коэффициенты a_i и b_i равны 0, имеет на множестве $(-\pi, \pi]$ не более $2m$ нулей с учетом их кратности.

Таким образом, если ${}^{n}t(x)$ равен нулю более чем в $2m$ точках, то на основании теоремы 3.12, все a_i и b_i в левой части (3.18) равны 0. Следовательно, в пространстве $C_{[-\pi,\,\pi]}$ система тригонометрических одночленов

$$1, \cos x, \sin x, \ldots, \cos mx, \sin mx, \ldots \qquad (3.19)$$

линейно независима для всех параметров дискретизации $z = 2m+1$, $m \in N$.

Линейная независимость системы (3.19) вытекает также из ортогональных свойств тригонометрических функций, если скалярное произведение в ***B***-пространстве $C_{[-\pi,\,\pi]}$ определить по формуле

$$\langle u, v\rangle = \int_{-\pi}^{\pi} u(x)v(x)dx.$$

Доказательство этого утверждения следует из соотношений:

$$2\cos ix\cos jx = (\cos(i-j)x + \cos(i+j)x),$$
$$2\sin ix\sin jx = (\cos(i-j)x - \cos(i+j)x), \qquad (3.20)$$
$$2\cos ix\sin jx = (\sin(i+j)x - \sin(i-j)x)$$

и свойств 2π-периодических функций $\sin x$ и $\cos x$.

Система элементов (3.19) является базисом в H-пространстве $L^2_{[-\pi,\pi]}$, а также в некоторых подмножествах B-пространства $C_{[-\pi,\pi]}$.

Теорема 3.13 [10, с. 474]. Если производная абсолютно непрерывной 2π-периодической функции $u(x)$ принадлежит $L^2_{[-\pi,\pi]}$, то ряд Фурье

$$\frac{a_0}{2} + \sum_{i=1}^{\infty}(a_i\cos ix + b_i\sin ix) \qquad (3.21)$$

функции $u(x)$ сходится к ней равномерно на отрезке $[-\pi, \pi]$.

Коэффициенты Фурье функции $u(x)$ определяются по формулам

$$a_0 = \frac{1}{\pi}\int_{[-\pi,\pi]} u(x)\,d\mu,\ a_i = \frac{1}{\pi}\int_{[-\pi,\pi]} u(x)\cos ix\,d\mu,$$
$$b_i = \frac{1}{\pi}\int_{[-\pi,\pi]} u(x)\sin ix\,d\mu,\ i = 1, 2, \ldots, m, \ldots. \qquad (3.22)$$

Однако во всем пространстве $C_{[-\pi,\pi]}$ по теореме Фейера система (3.19) лишь полна. Это означает, что линейное многообразие над ней хотя и всюду плотно в $C_{[-\pi,\pi]}$, но не содержит все предельные элементы из $C_{[-\pi,\pi]}$.

Теорема 3.14 [10, с. 477]. В Γ-пространстве $C_{[-\pi,\pi]}$ множество функций

$$\left\{\varphi_i(x) = \begin{bmatrix}\cos kx & \text{если } i - \text{четное};\\ \sin kx & \text{если } i - \text{нечетное},\end{bmatrix} \text{где } k = \left[\frac{i+1}{2}\right], i = 0, 1, \ldots\right\} \qquad (3.23)$$

образует полную ортогональную ***тригонометрическую*** систему.

Определим коэффициенты многочлена Фурье степени n наилучшего приближения функции $u(x) \in L^2_{[a,b]}$ методом наименьших квадратов (МНК). В качестве совокупности линейно независимых базисных функций конечномерного подмножества ${}^nU \subset L^2_{[a,b]}$ используем систему элементов

$$\{\varphi_i(x),\ i = 0, 1, \ldots, n\}. \qquad (3.24)$$

Тогда по условию многочлен ${}^nu(x)$ аппроксимации любой функции $u(x)$ из $L^2_{[a,b]}$ представляет собой линейную комбинацию

$$\phi(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x). \qquad (3.25)$$

Определение 3.25. Элемент $\boldsymbol{\phi(x)}$, полученный согласно (3.25), называется ***обобщенным*** многочленом порядка $\boldsymbol{n}$ по системе элементов (3.24).

Задача «о наилучшем приближении» состоит в том, чтобы для данной функции $\boldsymbol{u(x) \in U}$ среди всех линейных комбинаций вида (3.25) найти такой обобщенный многочлен $\boldsymbol{\phi(x)}$, для которого была бы минимальной норма

$$\|\boldsymbol{u(x) - (a_0 \varphi_0(x) + a_1 \varphi_1(x) + \ldots + a_n \varphi_n(x))}\|_U. \qquad (3.26)$$

Теорема 3.15 (о проекции). В замкнутом над базисом (3.24) подпространстве $\boldsymbol{^nU}$ $\boldsymbol{H}$-пространства $\boldsymbol{L^2_{[a,\,b]}}$ существует единственный элемент

$$\boldsymbol{\varphi(x) = \alpha_0 \varphi_0(x) + \alpha_1 \varphi_1(x) + \ldots + \alpha_n \varphi_n(x)} - \qquad (3.27)$$

многочлен ***наилучшего приближения*** (МНП), являющийся решением задачи «о наилучшем приближении» функции $\boldsymbol{u(x) \in U = L^2_{[a,\,b]}}$.

Используя определение нормы, представим (3.26) в виде

$$\boldsymbol{\|u(x) - \varphi(x)\| = \langle u(x) - \varphi(x),\, u(x) - \varphi(x)\rangle^{1/2}}$$

и возведем обе части этого равенства в квадрат

$$\boldsymbol{\|u(x) - \varphi(x)\|^2 = \langle A\alpha,\, \alpha\rangle - 2\langle v,\, \alpha\rangle + \|u(x)\|^2}, \qquad (3.28)$$

где $\boldsymbol{A = \{a_{kl} = \langle \varphi_k,\, \varphi_l\rangle,\, k,\, l = 0,\, 1,\, \ldots,\, n\}}$ – матрица Грама системы элементов (3.24), $\boldsymbol{\alpha = (\alpha_0,\, \ldots,\, \alpha_n)^T}$, $\boldsymbol{v = (\langle u,\, \varphi_0\rangle, \langle u,\, \varphi_1\rangle,\, \ldots,\, \langle u,\, \varphi_n\rangle)^T}$.

Задача нахождения МНП функции $\boldsymbol{u(x)}$ в $\boldsymbol{L^2_{[a,\,b]}}$ сводится к минимизации функционала $\boldsymbol{F(\alpha) = \langle A\alpha,\, \alpha\rangle - 2\langle v,\, \alpha\rangle}$, который по теореме [10, с. 580] и свойству $\boldsymbol{\langle A\alpha,\, \hbar\rangle = \langle \alpha,\, A^T \hbar\rangle}$ имеет единственную точку минимума

$$\boldsymbol{A\alpha = v}. \qquad (3.29)$$

Определение 3.26. Обобщенный многочлен $\boldsymbol{\varphi(x)}$ наилучшего приближения $\boldsymbol{u(x) \in L^2_{[a,\,b]}}$ в ортогональном базисе (3.24) назовем ***многочленом Фурье***, а его коэффициенты (элементы вектора $\boldsymbol{\alpha}$) – ***коэффициентами Фурье***.

Если $\boldsymbol{a = -\pi,\ b = \pi}$, а элементы системы (3.24) тригонометрические одночлены (3.19), то в силу тождеств (3.20) угол между любыми двумя векторами системы прямой. Таким образом, матрица Грама $\boldsymbol{A}$ по этой системе является диагональной, а все элементы диагонали (кроме $\boldsymbol{A_{00} = 2\pi}$) равны $\boldsymbol{\pi}$. Отсюда следует, что коэффициенты Фурье вычисляются по формулам (3.22).

Отметим, что среди всех тригонометрических полиномов $^{n}t(x) \in T^{n}$ с $n = 2m$ отрезок ряда Фурье гармонического порядка m дает наилучшую аппроксимацию функции $u(x)$ в метрике H-пространства $L^{2}_{[-\pi,\,\pi]}$ [10, с. 448].

Однако при аппроксимации непрерывных непериодических функций по норме $C_{[-\pi,\,\pi]}$ МНК всегда возникает неустранимая погрешность приближения, равная $|u(-\pi) - u(\pi)|$. В этом случае необходимо заменой $x := kx$ изменить период функций, входящих в систему (3.19). В тоже время любую 2π-периодическую функцию $u(x) \in C_{[-\pi,\,\pi]}$ можно представить в виде (3.21), но тогда коэффициенты a_i и b_i будут находиться не по формулам (3.22), а через средние арифметические сумм Фейера функции $u(x)$ [10, с. 477].

Следовательно, алгоритм генерации последовательности многочленов наилучшего приближения по *sup*-норме, равномерно аппроксимирующих порождающую функцию из $C_{[a,\,b]}$, отличается от алгоритмов интерполяционного процесса в $C^{k}_{[a,\,b]}$ и квадратичного проектирования в $L^{2}_{[a,\,b]}$ (реализуемого методом наименьших квадратов).

При приближении бесконечно дифференцируемых на отрезке функций процесс квадратичной аппроксимации и процесс интерполирования сходятся к порождающей функции (точнее, к ее многочлену Тейлора) равномерно. Например, аппроксимация МНК непрерывной непериодической функции $u(x) = x,\ x \in [0;\,1]$ тригонометрическим многочленом порядка $n = 16$ с точностью $\delta = 3{,}2\ 10^{-10}$ по норме $L^{2}_{[0;\,1]}$ и точностью $\delta = 2{,}3\ 10^{-9}$ по норме $C_{[0;\,1]}$ имеет вид:

$$^{16}u(x) = 0{,}3337724868\ldots + 0{,}7393725436\ldots \sin x -$$
$$- 0{,}1469448994\ldots \cos x + 0{,}3267341510\ldots \sin 2x - 0{,}5485640148\ldots \cos 2x -$$
$$- 0{,}1264346594\ldots \sin 3x + 0{,}4825248292\ldots \cos 3x + 0{,}0441978669\ldots \sin 4x -$$
$$- 0{,}0941897332\ldots \cos 4x - 0{,}1004920723\ldots \sin 5x - 0{,}0387255863\ldots \cos 5x +$$
$$+ 0{,}0733144751\ldots \sin 6x + 0{,}0073388391\ldots \cos 6x - 0{,}0203221236\ldots \sin 7x +$$
$$+ 0{,}0064654745\ldots \cos 7x + 0{,}0018125090\ldots \sin 8x - 0{,}0016773962\ldots \cos 8x.$$

Теорема 3.16 [10, с. 450]. Каждая из двух тригонометрических систем

$$1,\ \cos x,\ \ldots,\ \cos mx,\ \ldots; \tag{3.30}$$

$$\sin x,\ \ldots,\ \sin mx,\ \ldots \tag{3.31}$$

является полной и ортогональной в пространстве $L^{2}_{[0,\,\pi]}$.

Функцию $u(x) \in L^{2}_{[0,\,\pi]}$ доопределим на отрезке $[-\pi,\,0)$ по формуле $u(-x) = u(x)$ и разложим в ряд Фурье. Функция $u(x)$, определенная теперь на отрезке $[-\pi,\,\pi]$, четная, то есть все коэффициенты при синусах этого ряда равны 0.

Следовательно, функцию $u(x) \in L^2_{[0,\pi]}$ с любой точностью можно аппроксимировать линейной комбинацией элементов (3.30). Рассматривая систему (3.31), доопределим функцию $u(x)$ из $L^2_{[0,\pi]}$ на отрезке $[-\pi, 0)$ по формуле $u(-x) = -u(x)$ и разложим ее в ряд Фурье. Коэффициенты при косинусах этого ряда равны 0. Ортогональность данных систем очевидна.

Утверждение 3.3. В B-пространстве $C_{[0,\pi]}$ тригонометрическая система (3.30) является полной и ортогональной, а система (3.31) таковой не является. Чтобы система (3.31) стала полной, надо добавить к ней 1.

Рассмотрим линейную комбинацию степенных одночленов

$$1, x, x^2, \ldots, x^n, \ldots \tag{3.32}$$

с вещественными коэффициентами. Совокупность образованных таким образом функций назовем множеством P ***степенных*** многочленов.

В дальнейшем изложении используются только тригонометрические и степенные системы базисных функций, поэтому линейную комбинацию

$$a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \equiv {}^n p(x) \in P^n \tag{3.33}$$

будем иногда называть просто ***многочленом*** (или ***полиномом***) степени n.

Полнота системы (3.32) в пространстве $C_{[a,b]}$, а следовательно, и в $L^2_{[a,b]}$ вытекает из теоремы Вейерштрасса о равномерной аппроксимации любой непрерывной на отрезке $[a, b]$ функции многочленом [10, с. 480].

Пусть $a = 0$, $b = 1$, а элементами системы (3.24) являются степенные одночлены (3.32) с показателем степени не выше n, тогда угол между любыми двумя векторами φ_i и φ_j данной системы определяется по формуле

$$\cos(\varphi_i \wedge \varphi_j) = \frac{\sqrt{(2i+1)(2j+1)}}{i+j+1}, \; i, j = 0, 1, \ldots, n. \tag{3.34}$$

Косоугольный базис конечномерного пространства P^n не позволяет определить коэффициенты многочлена наилучшего приближения по норме $L^2_{[0;1]}$, так же как в случае с ортогональным базисом. Причем, чем больше показатель степени n, тем меньше параметр $£$ – минимальная величина из множества значений углов между элементами степенного базиса P^n.

Это обстоятельство ухудшает сходимость процесса квадратичной аппроксимации степенными многочленами множества P^n даже непрерывных функций из пространства $L^2_{[0;1]}$, так как определитель матрицы Грама (3.29) конечной системы (3.32) при $n \to \infty$ стремится к 0.

Для того чтобы получить аналогичные (3.22) формулы вычисления коэффициентов степенных многочленов, аппроксимирующих функции по норме $L^2_{[a,b]}$ в степенном базисе, требуется преобразовать систему (3.32) в ортогональный относительно скалярного произведения (3.17) базис.

Проводя процесс ортогонализации [1, с. 224] линейно независимой косоугольной системы (3.32) при $a = -1$ и $b = 1$ по отношению к скалярному произведению

$$\langle u, v\rangle = \int_{[-1;1]} u(x)v(x)d\mu, \tag{3.35}$$

получим полную ортогональную систему многочленов Лежандра

$${}^{i}p(x) = \frac{1}{i!\,2^i}\frac{d^i}{dx^i}(x^2-1)^i,\ i = 0, 1, \ldots, n, \ldots. \tag{3.36}$$

Приведем несколько первых членов этой системы

$$1,\ x,\ \frac{3}{2}x^2 - \frac{1}{2},\ \frac{5}{2}x^3 - \frac{3}{2}x,\ \frac{35}{8}x^4 - \frac{15}{4}x^2 + \frac{3}{8}, \ldots. \tag{3.37}$$

Разложение функции $u(x)$ по ортогональным (но не нормированным) многочленам Лежандра на отрезке $[-1; 1]$ представляет сходящийся ряд

$$u(x) = \sum_{i=0}^{\infty} c_i\,{}^{i}p(x), \text{ где } c_i = \frac{2i+1}{2}\int_{[-1;1]} u(x)\,{}^{i}p(x)d\mu. \tag{3.38}$$

Степенной ортонормированный базис Родриго множества P^n имеет вид

$${}^{i}s(x) = \frac{\sqrt{2i+1}}{i!\,2^i\sqrt{2}}\frac{d^i}{dx^i}(x^2-1)^i,\ i = 0, 1, \ldots, n.$$

Опишем алгоритм вычисления коэффициентов многочлена наилучшего приближения функции $u(x)$ в степенном базисе H-пространства $L^2_{[a,b]}$

$${}^{n}u(x) = \sum_{k=0}^{n} c_k x^k, \tag{3.39}$$

которое назовем *квадрополяционным* приближением функции $u(x) \in L^2_{[a,b]}$. Несмотря на то, что этот базис косоугольный, существование и единственность МНП функции $u(x)$ следует из существования и единственности соответствующего многочлена Лежандра в ортогональном базисе (3.36).

Сформулируем условие минимизации функционала

$$Z(c_0, \ldots, c_n) \equiv \int_{[a,b]} (u(x) - {}^{n}u(x))^2 d\mu \to \min. \tag{3.40}$$

Так как функция $Z(c_0, \ldots, c_n)$ достигает локальный минимум в точке, где частные производные по всем ее переменным обращаются в 0, то в условиях задачи выражение (3.40) можно заменить эквивалентным (3.29)

$$\int_{[a,b]} (u(x) - {}^{n}u(x))(-x^k)d\mu = 0,\ k = 0, \ldots, n. \tag{3.41}$$

Вынесем из-под знака интеграла Лебега постоянные множители, тогда (3.41) станет равносильно линейной системе алгебраических уравнений

$$\begin{cases} c_0 \int\limits_{[a,b]} d\mu + \ldots + c_n \int\limits_{[a,b]} x^n d\mu = \int\limits_{[a,b]} u(x) d\mu; \\ \cdot \quad \cdot \quad \cdot \\ c_0 \int\limits_{[a,b]} x^n d\mu + \ldots + c_n \int\limits_{[a,b]} x^{2n} d\mu = \int\limits_{[a,b]} x^n u(x) d\mu. \end{cases} \tag{3.42}$$

Определитель матрицы Грама (3.29), собранной из коэффициентов при искомых значениях $c_0, \ldots, c_n$, больше нуля. И хотя коэффициенты квадрополяционного многочлена сразу не определяются по формулам типа (3.22) и (3.38), их можно найти с помощью решения ЛСАУ (3.42).

Однако воспользоваться решением системы (3.42) при $n >> 100$ достаточно сложно, так как определитель матрицы Грама по системе одночленов $\{1, x, x^2, \ldots, x^n\}$ при $n \to \infty$ стремится к 0. Тем не менее, если МНП ${}^n u(x)$ функции $u(x) \in L^2_{[a,b]}$ найден в базисе (3.36), то через матрицу перехода (2.41) к базису (3.32) следует его представление в виде (3.39).

Утверждение 3.4. На множестве элементов вида (3.39) многочлен ${}^n u(x)$ минимизирует функционал $\|u(x) - {}^n u(x)\|^2$ по норме $L^2_{[a,b]}$ тогда и только тогда, когда ${}^n u(x)$ определен методом наименьших квадратов.

Для вычисления элементов матрицы Грама системы (3.32) в $L^2_{[0;\,1]}$ лебеговский интеграл заменим римановским. Тогда справедлива формула

$$G_{ij} = \frac{1}{1 + i + j}, \; i, j = 0, 1, \ldots, n. \tag{3.43}$$

Процесс квадрополяции непрерывных функций из $L^2_{[0;\,1]}$ неустойчив относительно их аппроксимации по норме $C_{[0;\,1]}$. Однако для функций из пространства $C^{\infty}_{[0;\,1]}$ он обладает равномерной сходимостью по *sup*-норме.

Например, степенной многочлен порядка $n = 12$, аппроксимирующий функцию $u(x) = \sin x$, $x \in [0; 1]$ из $C^{\infty}_{[0;\,1]}$ с точностью $\delta = 3{,}0 \cdot 10^{-10}$ по норме $L^2_{[0;\,1]}$ и точностью $\delta = 1{,}8 \cdot 10^{-9}$ по норме $C_{[0;\,1]}$, имеет вид, сходствующий с формулой Тейлора исследуемой функции: ${}^{12}u(x) = 0{,}0000000018\ldots +$

$$+ 0{,}9999997339\ldots x + 0{,}0000099857\ldots x^2 - 0{,}1668295165\ldots x^3 +$$
$$+ 0{,}0014385207\ldots x^4 + 0{,}0006338445\ldots x^5 + 0{,}0265668380\ldots x^6 -$$
$$- 0{,}0612322600\ldots x^7 + 0{,}0942991437\ldots x^8 - 0{,}0968074611\ldots x^9 +$$
$$+ 0{,}0633158266\ldots x^{10} - 0{,}0238726233\ldots x^{11} + 0{,}0039489522\ldots x^{12}.$$

Удобное для исследований представление решения (*) в виде (3.39) хотелось бы получить и в B-пространствах, однако в банаховом пространстве непрерывных на отрезке функций система (3.32) является лишь полной.

Тем не менее, аппроксимация функций ***B***-пространств, содержащих всюду плотное множество многочленов, МНП из $\boldsymbol{P^n}$ по норме ***B*** возможна.

Полная система пространства непрерывных функций ***M*** переменных

$$^{M}\boldsymbol{C_X} = \bigotimes_{j=1}^{M} \boldsymbol{C_{[a_j,b_j]}} \qquad (3.44)$$

определяется прямым произведением полных систем (3.32) по каждой переменной на отрезке $\boldsymbol{[a_j, b_j]}$. Таким образом, полная система $^{M}\boldsymbol{C_X}$ имеет вид

$$\boldsymbol{S} = \bigotimes_{j=1}^{M} \{\boldsymbol{1}, {}^{j}\boldsymbol{x}, {}^{j}\boldsymbol{x^2}, \ldots, {}^{j}\boldsymbol{x^n}, \ldots\}.$$

Определение 3.27. Если счетная линейно независимая система ***S*** элементов **B**-пространства ***U*** является полной в этом пространстве, то систему ***S*** назовем ***предбазисом U***.

Предбазис функционального ***B***-пространства $\boldsymbol{U_X}$ напрямую связан с понятием счетный базис меры множества ***X*** (раздел 3.2), на котором заданы функции пространства, а именно: любая дискретизация множества ***X*** влечет соответствующую дискретизацию предбазиса ***S***.

То есть каждую функцию ***u(x)*** из сепарабельного ***B***-пространства $\boldsymbol{U_X}$ можно с любой точностью $\boldsymbol{\delta}$ аппроксимировать линейной комбинацией элементов конечной подсистемы $^{n}\boldsymbol{S}$ предбазиса, выбрав специальным образом параметр $\boldsymbol{n = n(u, \delta)}$ и подсистему $^{n}\boldsymbol{C}$ сегментов множества ***X***.

ГЛАВА 4 ПРИБЛИЖЕННЫЕ ВЫЧИСЛЕНИЯ В ПРОСТРАНСТВЕ *R* ДЕЙСТВИТЕЛЬНЫХ ЧИСЕЛ

4.1 Учёт погрешностей вычислений

При решении математических задач погрешности появляются по нескольким различным причинам. Можно выделить следующие основные источники погрешностей.

1. При составлении математической модели какого-либо явления приходится принимать некоторые условия, упрощающие задачу. Таким образом, математическая формулировка задачи не точно отображает реальные явления, а лишь даёт в некоторой степени идеализированную картину. При этом возникает погрешность постановки задачи.

2. Бывает, что в данной постановке задачу решить трудно или вообще невозможно. Тогда применяют некоторый метод приближённого решения задачи, т.е. данную задачу заменяют задачей, решение которой в определённом смысле близко к решению данной задачи. Погрешность, возникающую при этом, называют погрешностью метода.

3. Погрешность могла быть вызвана тем, что при вычислениях приходится производить действия над приближёнными, а не точными значениями параметров, входящих в математическую формулу. Очевидно, что погрешность таких начальных данных (начальная погрешность) в некоторой степени переносится в результат. Такую погрешность называют погрешностью действий.

4. Погрешность возникает при округлении бесконечных и конечных десятичных чисел, имеющих большее количество значащих цифр или десятичных знаков, чем требуется при вычислениях. Это погрешность округления.

Таким образом, погрешность результата решения задачи складывается из погрешностей, возникающих по разным причинам.

Пусть x – некоторое число. *Приближённым значением (или приближением)* числа x называют некоторое число a, которое в определённом смысле мало отличается от x и заменяет его в вычислениях. Запись $x \approx a$ будет означать в дальнейшем, что число a есть приближённое значение числа x.

Определение 4.1. *Погрешностью* Δ_a *приближённого значения* a *числа* x *называют разность* $\Delta_a = x - a$.

Модуль погрешности (абсолютная погрешность) указывает, насколько отличается приближение от точного значения. По знаку погрешности можно определить, как взято приближение a с избытком или недостатком. Погрешность положительна, если приближение a взято с недостатком и отрицательна в противном случае.

Границей приближённого значения a числа *x* называют всякое неотрицательное число h_a, которое не меньше модуля погрешности $|\Delta_a| \le h_a$. Говорят, что число *a* является *приближённым значением числа x с точностью до* h_a, если выполнено неравенство $|x - a| \le h_a$. Отсюда следует, что число x заключено в границах

$$a - h_a \le x \le a + h_a.$$

Запись $x = a \pm h_a$ означает, что число *a* есть приближённое значение числа *x* с точностью до h_a.

Пример 4.1. Число $a = 0{,}273$ – приближённое значение числа *x* с точностью до 0,001.Указать границы, в которых заключено число *x*.

Решение

$$a - 0{,}001 \le x \le a + 0{,}001, \quad 0{,}272 \le x \le 0{,}274.$$

Определение 4.2. *Относительной погрешностью* ω_a *приближённого значения* $a\,(a \ne 0)$ *числа x называют отношение* $\omega_a = \dfrac{\Delta_a}{a}$.

При округлении чисел считают, что граница погрешности округления равна половине единицы округляемого разряда $h_{окр} = 0{,}5 \cdot 10^{\alpha}$, где α – порядок округляемого разряда.

Пример 4.2. Округлить до разряда единиц числа

а) 2,3; б) 1,998; в) 4,5.

Решение

Граница погрешности округления равна 0,5, $h_{окр} = 0{,}5$. После округления получим числа 2, 2, 5.

Пример 4.3. Округлить до десятых следующее число 27,52. Найти погрешность и относительную погрешность округления.

Решение

$$x = 27{,}52,\ a = 27{,}5; \quad \Delta_a = x - a = 0{,}02;$$

$$\omega_a = \frac{\Delta_a}{a} = \frac{0{,}02}{27{,}5} = \frac{1}{1375}.$$

Так же, как и погрешность, относительная погрешность не всегда может быть вычислена, поэтому приходится иногда лишь оценивать ее модуль. Модуль относительной погрешности часто выражают в процентах. Чем меньше модуль относительной погрешности, тем лучше приближённое значение характеризует точное значение числа или тем выше качество приближения.

Определение 4.3. *Границей относительной погрешности приближённого значения* a *числа* x *называют всякое неотрицательное число* ε_a, *которое не меньше модуля относительной погрешности:* $|\omega_a| \le \varepsilon_a$.

Связь между границами погрешности и границами относительной погрешности

Пусть известны a и его граница погрешности h_a, тогда

$$|\omega_a| = \left|\frac{\Delta_a}{a}\right| \le \frac{h_a}{|a|}.$$

Следовательно, за границу относительной погрешности ε_a можно принять отношение

$$\frac{h_a}{|a|}, \text{ т.е. } \varepsilon_a = \frac{h_a}{|a|}.$$

Пусть известны приближение a и граница его относительной погреш-ности ε_a. Тогда $|\Delta_a| \le |a|\varepsilon_a$, т.е. $h_a = |a|\varepsilon_a$ – граница погрешности.

Округление приближённых значений чисел

Пусть x – число; a – приближённое значение числа x.

При округлении приближенного значения a числа x получается новое приближённое значение a_1 того же числа x.

Вычислим погрешность этого нового приближения a_1:

$$|\Delta_{a_1}| = |x - a_1| = |x - a + a - a_1| = |(x - a) + (a - a_1)| = |\Delta_a + \Delta_{окр}| \le$$

$$\le |\Delta_a| + |\Delta_{окр}| \le h_a + |\Delta_{окр}|,$$

где $\Delta_{окр}$ – погрешность округления a до a_1;

h_a – граница погрешности приближения a.

Следовательно, за границу погрешности полученного приближения a_1 можно принять сумму границ погрешности округляемого приближённого значения и погрешности округления:

$$h_a + |\Delta_{окр}| = h_{a_1}.$$

Таким образом, доказана

Теорема 4.1. *При округлении приближённого значения числа получается новое приближённое значение, для которого границей погрешности является сумма границ погрешности округляемого приближения и погрешности округления.*

4.2 Оценка погрешностей результатов действий над приближёнными значениями чисел

Пусть

$$X = \sum_{i=1}^{n} x_i,$$

где x_i – некоторые числа любого знака, заданные приближениями a_i с точностью до h_{a_i}. Обозначим $A = \sum_{i=1}^{n} a_i$.

Тогда

$$X = A \pm h_A,$$

где h_A – граница погрешности суммы приближенных значений $a_{i.}$.

Теорема 4.2. Сумма границ погрешностей приближенных слагаемых является границей погрешности их алгебраической суммы:

$$\sum_{i=1}^{n} h_{a_i} = h_A.$$

Доказательство

$$|\Delta_A| = |X - A| = \left|\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} a_i\right| = \left|\sum_{i=1}^{n} (x_i - a_i)\right| \le \sum_{i=1}^{n} |x_i - a_i| \le \sum_{i=1}^{n} h_{a_i},$$

т.е. $\sum_{i=1}^{n} h_{a_i} = h_A$. Теорема доказана.

Теорема 4.3. Среди границ относительной погрешности суммы положительных приближённых слагаемых существует такая, которая не превышает наибольшей из границ относительных погрешностей слагаемых:

$$\varepsilon_A \le \max\left\{\varepsilon_{a_1}, \varepsilon_{a_2}, \ldots, \varepsilon_{a_n}\right\}.$$

Теорема 4.4. Сумма границ относительных погрешностей приближённых сомножителей является границей относительной погрешности их произведения:

$$\left|\omega_{ab}\right| \le \varepsilon_a + \varepsilon_b, \quad \left(x = a \pm h_a, \quad y = b \pm h_b\right).$$

Следствие 4.1. При умножении приближённого значения числа на точный множитель k граница относительной погрешности не изменяется, а граница погрешности увеличивается в $|к|$ раз.

Следствие 4.2. Произведение границы относительной погрешности приближённого значения a числа x на n является границей относительной погрешности результата возведения a в целую положительную степень n:

$$\left|\omega_{a^n}\right| \le n\,\varepsilon_a .$$

Следствие 4.3. Частное границы относительной погрешности приближённого значения a числа x и n является границей относительной погрешности корня $n-$ой степени из a:

$$\left|\omega_{\sqrt[n]{a}}\right| \le \frac{\varepsilon_a}{n}.$$

Теорема 4.5. Сумма границ относительных погрешностей приближённых значений делителя и делимого является границей относительной погрешности частного:

$$\left|\omega_{\frac{a}{b}}\right| \le \varepsilon_a + \varepsilon_b .$$

4.3 Приближённые вычисления без учёта погрешностей

Правило 4.1. Для того, чтобы вычислить алгебраическую сумму приближённых слагаемых, нужно:

1) *среди слагаемых выделить наименьшее точное,*
2) *все остальные слагаемые округлить, сохраняя один запасной разряд, следующий за последним разрядом в выделенном слагаемом,*
3) *сложить полученные после округления числа,*
4) *округлить полученный результат до предпоследнего разряда*.

Пример 4.4. Найти сумму приближённых слагаемых 2,737; 0,77974; 27,1; 0,293.

Решение

1) Наименее точным слагаемым является 27,1.
2) Округляем остальные слагаемые до сотых: 2,74; 0,78; 27,1; 0,29.
3) Сумма равна 39,91.
4) Округляем до десятых 30,9.

Определение 4.4. Значащими цифрами в десятичной записи числа называют все его цифры, кроме нулей, записанных слева от первой цифры, отличной от нуля.

Правило 4.2. Для того, чтобы произвести умножение (деление) приближённых значений чисел, следует:

1) *выделить сомножитель, имеющий наименьшее количество значащих цифр,*
2) *округлить остальные сомножители, сохраняя на одну запасную значащую цифру больше, чем в выделенном сомножителе,*
3) *произвести умножение (деление),*
4) *полученный результат округлить, сохраняя столько значащих цифр, сколько их в выделенном сомножителе.*

Пример 4.5. Вычислить $\rho = 4{,}748 \cdot 3{,}34 \cdot 0{,}7$.

Решение

1) Наименьшее число значащих цифр в числе 0,7 – только одна.
2) Остальные сомножители округляем, сохраняя по две значащие цифры 4,7 и 3,3.
3) $4{,}7 \cdot 3{,}3 \cdot 0{,}7 = 10{,}657$.
4) Округляем результат, сохраняя одну значащую цифру, как в выделенном сомножителе: $1 \cdot 10^1$. Следовательно, $\rho = 1 \cdot 10^1$.

Правило 4.3. При возведении приближённого значения в квадрат или куб, а также при извлечении корня квадратного или кубического из него в результате следует сохранить столько значащих цифр, сколько их имеет основание.

Правило 4.4. Если число является результатом промежуточных действий, то следует сохранить в нем на одну–две цифры больше, чем указано в правилах 4.1.–4.3.

Правило 4.5. При вычислениях с заданной точностью исходные данные следуют брать с таким количеством верных цифр, чтобы действуя по правилам 4.1.–4.4, *обеспечить количество верных цифр на одну больше, чем требуется. Иначе говоря, при небольшом количестве промежуточных действий количество значащих (десятичных) цифр в исходных данных следует брать на одну-две больше, чем требуется в результате.*

Правило 4.6. При вычислении по таблицам значения синуса и косинуса приближённого аргумента, выраженного в радианах, в результате следует сохранить столько десятичных знаков, сколько их имеет аргумент.

Правило 4.7. Если приближённое значение угла, выраженное в радианах, принадлежит промежутку $\left[-\pi/4, \pi/4\right]$, *то при вычислении значения тангенса этого угла в результате следует сохранить не больше десятичных знаков, чем их имеет приближённое значение угла.*

Правило 4.8. При вычислении мантиссы десятичного логарифма приближённого аргумента в результате следует сохранять на один десятичный знак меньше, чем количество значащих цифр аргумента. При этом в результате все цифры верные, если они верны в аргументе.

Правило 4.9. Если значение функции является результатом промежуточных действий, то следует сохранять в нем на одну цифру больше, чем рекомендуют правила 4.6–4.8.

4.4 Связь между количеством верных цифр и относительной погрешностью

Пусть $x \approx a$, $\Delta_a = x - a$.

Определение 4.5. Цифра приближённого значения, записанная в разряде α*, называется верной, если модуль его погрешности* Δ_a *не превосходит половины единицы этого разряда, т.е.* $|\Delta_a| \le 0{,}5 \cdot 10^{\alpha}$.

Очевидно, что слева от верной цифры находятся верные цифры.

Пример 4.6. x=27,421; a=27,384; Δ_a=0,037.

Решение

4 – неверная, так как Δ_a=0,037 $\overset{неверно}{\le}$ $0{,}5 \cdot 10^{-3}$=0,0005.

8 – неверная, так как Δ_a=0,037 $\overset{неверно}{\le}$ $0{,}5 \cdot 10^{-2}$=0,005.

3 – верная, так как Δ_a=0,037 $\le 0{,}5 \cdot 10^{-1}$=0,05.

Итак, 2, 7, 3 – верные цифры.

Пусть известно количество m верных значащих цифр приближенного значения числа, тогда в стандартной форме его можно записать так:

$$a = \pm\beta_0, \beta_{-1}\beta_{-2}..\beta_{-m+1} \cdot 10^n,$$

где $1 \le \beta_0 \le 9$, $0 \le \beta_{-i} \le 9$, $i = \overline{1, m-1}$, n – порядок числа а.

По предположению, m – ая цифра верна, поэтому $|\Delta_a| \le 0{,}5 \cdot 10^{-m+1} \cdot 10^n$.

Тогда $|\omega_a| \le \dfrac{ha}{|a|} = \dfrac{0{,}5 \cdot 10^{n-m+1}}{\beta_0, \beta_{-1}..\beta_{-m+1} \cdot 10^n} \le \dfrac{0{,}5}{\beta_0} 10^{-m+1} \le 0{,}5 \cdot (0{,}1)^{m-1}$.

Таким образом, доказана

Теорема 4.6. Если приближение имеет m *верных значащих цифр, то число* $0{,}5 \cdot (0{,}1)^{m-1}$ *является границей его относительной погрешности.*

Замечание 4.1. Пусть приближение имеет m верных значащих цифр. Если известна первая значащая цифра приближения, то за границу относительной погрешности можно принять число $\frac{1}{2}\beta_0\,(0{,}1)^{m-1}$, поскольку

$$\tfrac{1}{2}\beta_0\,(0{,}1)^{m-1} \geq 0{,}5\cdot(0{,}1)^{m-1}.$$

Граница относительной погрешности зависит от количества верных значащих цифр m, от величины первой из них β_0, но не зависит от порядка числа n.

Теорема 4.7. Если граница *относительной погрешности равна* $0{,}5\,(0{,}1)^m$, *то приближение a имеет не менее m верных значащих цифр.*

Доказательство

Пусть β_0 – первая значащая цифра приближённого значения a и n – порядок разряда этой цифры. Тогда

$$|\Delta_a| \leq \varepsilon_a \cdot |a| \leq 0{,}5\cdot(0{,}1)^m(\beta_0+1)\cdot 10^n \leq 0{,}5\cdot 10\cdot 10^{n-m} = 0{,}5\cdot 10^{-m+1}\cdot 10^n,$$

т.е. $|\Delta_a| \leq 0{,}5\cdot 10^{-m+1}\cdot 10^n$.

Значит, цифра, записанная в $-(m-1)$ разряде значащей части числа верная, следовательно, после запятой верных $m-1$ цифр и еще β_0 – верная, итого всех верных цифр не менее m. Теорема доказана.

Пример 4.7. Если известно, что относительная погрешность приближения по модулю не превосходит $0{,}03\%$, то согласно теореме это приближение имеет не менее трёх значащих верных цифр, так как

$$0{,}03\% = = 0{,}0003 < 0{,}5\cdot(0{,}1)^3.$$

4.5 Функция от приближённых значений аргументов

Пусть функция $y = f(x_1, x_2,\ldots,x_n)$ задана в области G и дифференцируема в ней по всем переменным x_i.

Пусть a_i – приближённое значение аргумента x_i, $(i=\overline{i,n})$, Δ_{a_i} – погрешность a_i, причем точка $A(a_1, a_2,\ldots,a_n)\in G$.

Требуется оценить погрешность приближённого значения функции $\tilde{y} = f(a_1, a_2,\ldots,a_n)$. Имеем

$$|\Delta_{\tilde{y}}| = |y-\tilde{y}| = |f(x_1, x_2,\ldots,x_n) - f(a_1, a_2,\ldots,a_n)| =$$

$$= |f(a_1+\Delta a_1,\ldots,a_n+\Delta a_n) - f(a_1, a_2,\ldots,a_n)|.$$

Если предположить, что Δa_i достаточно малые величины, то их произведениями, квадратами и высшими степенями можно пренебречь. Тогда получим

$$\left|\Delta_{\tilde{y}}\right| \approx \left|df(a_1, a_2, ..a_n)\right| = \left|\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)_A \cdot \Delta a_i\right| \le \sum_{i=1}^{n}\left|\left(\frac{\partial f}{\partial x_i}\right)_A\right| \cdot |\Delta a_i| \le \sum_{i=1}^{n}\left|\left(\frac{\partial f}{\partial x_i}\right)_A\right| h_{a_i}.$$

Поэтому за границу погрешности $\tilde{y}$ можно взять

$$h_{\tilde{y}} = \sum_{i=1}^{n}\left|\left(\frac{\partial f}{\partial x_i}\right)_A\right| h_{a_i}.$$

Если $f(a_1, a_2, ..., a_n) > 0,$ то, разделив последнее равенство на $f(a_1, a_2, ..., a_n),$ получим оценку модуля относительной погрешности $\tilde{y}$:

$$\frac{h_{\tilde{y}}}{\tilde{y}} = \sum_{i=1}^{n}\left|\frac{1}{f(a_1, a_2, ..., a_n)}\left(\frac{\partial f}{\partial x_i}\right)_A\right| h_{a_i} = \sum_{i=1}^{n}\left|\left(\frac{\partial \ln f}{\partial x_i}\right)_A\right| h_{a_i}.$$

Следовательно,

$$\varepsilon_{\tilde{y}} = \sum_{i=1}^{n}\left|\left(\frac{\partial \ln f(x_1, ..x_n)}{\partial x_i}\right)_A\right| h_{a_i}.$$

Пример 4.8. Вычислить величину погрешности приближённого значения большего корня уравнения $x^2 + \rho x + q = 0$ $(\rho \doteq 3{,}5,\ q \doteq -7{,}8)$, обусловленной погрешностями приближенных значений коэффициентов. (Точка над знаком «=» означает, что в записи числа участвуют только верные цифры).

Решение

$$x_1(\rho, q) = -\frac{\rho}{2} + \sqrt{\left(\frac{\rho}{2}\right)^2 - q}\ ;\ \rho \doteq 3{,}5;\ q \doteq -7{,}8;\ \tilde{\rho} = 3{,}5;\ \tilde{q} = -7{,}8;$$

$$x_1(\tilde{\rho}, \tilde{q}) = \tilde{x}_1 \text{ (обозначим)},\ h_{\tilde{\rho}} = 0{,}05,\ h_{\tilde{q}} = 0{,}05.$$

Тогда

$$\left|\Delta_{\tilde{x}_1}\right| \le \left|\left(\frac{\partial x_1(\rho, q)}{\partial \rho}\right)_{\tilde{\rho}, \tilde{q}}\right| \cdot h_{\tilde{\rho}} + \left|\left(\frac{\partial x_1(\rho, q)}{\partial q}\right)_{\tilde{\rho}, \tilde{q}}\right| \cdot h_{\tilde{q}} =$$

$$= \left|-\frac{1}{2} + \frac{\tilde{\rho}}{4 \cdot \sqrt{\left(\frac{\tilde{\rho}}{2}\right)^2 - \tilde{q}}}\right| h_{\tilde{\rho}} + \left|\frac{1}{2 \cdot \sqrt{\left(\frac{\tilde{\rho}}{2}\right)^2 - \tilde{q}}}\right| h_{\tilde{q}} =$$

$$= \left| -\frac{1}{2} + \frac{3{,}5}{4 \cdot \sqrt{\left(\frac{3{,}5}{2}\right)^2 + 7{,}8}} \right| \cdot 0{,}05 + \left| \frac{1}{2 \cdot \sqrt{\left(\frac{3{,}5}{2}\right)^2 + 7{,}8}} \right| \cdot 0{,}05 \le 0{,}02 .$$

Таким образом, $h_{\tilde{x}_1} = 0{,}02$.

4.6 Обратная задача теории погрешностей

Все задачи теории погрешностей можно разделить на два класса: прямые и обратные задачи.

Прямая задача. Определить погрешность данной функции от приближённых значений аргументов, заданных с известной относительной или абсолютной точностью (см. раздел 5).

Обратная задача. Какими должны быть абсолютные или относительные погрешности аргументов, чтобы модуль абсолютной или относительной погрешности значения функции от этих аргументов не превышал заданной величины.

Пусть $y = f(x_1, x_2, \ldots, x_n)$ – функция непрерывно-дифференцируемая в области G. Пусть точка $A(a_1, a_2, \ldots, a_n) \in G$ вместе с параллелепипедом

$$a_i - h_{a_i} \le x_i \le a_i + h_{a_i}, \quad (i = \overline{1,n}).$$

С какой точностью h_{a_i} следует взять приближенные значения a_i аргументов x_i, чтобы погрешность значения функции $\tilde{y} = f(a_1, a_2, \ldots, a_n)$ не превышала по модулю заданную величину $h_{\tilde{y}}$. Из условия задачи имеем

$$\left|\Delta_{\tilde{y}}\right| \le \sum_{i=1}^{n} \left| \left(\frac{\partial f}{\partial x_i}\right)_A \right| \cdot h_{a_i} \le h_{\tilde{y}} .$$

Существуют различные подходы к решению этой задачи. Рассмотрим принцип равных влияний, который заключается в дополнительном предположении, что погрешности всех аргументов вносят одинаковые доли в погрешность функции, т.е. все частные дифференциалы равны между собой по модулю.

Тогда

$$\left| \left(\frac{\partial f}{\partial x_i}\right)_A \right| \cdot h_{a_i} \le \frac{h_{\tilde{y}}}{n}, \quad i = \overline{1,n} .$$

Отсюда следует: для того, чтобы значение $\tilde{y}$ было вычислено с точностью до $h_{\tilde{y}}$, достаточно, чтобы погрешности аргументов не превосходили по модулю

$$h_{a_i} \le \frac{h_{\tilde{y}}}{n\left|\left(\frac{\partial f}{\partial x_i}\right)_A\right|}, \quad i=\overline{1,n}.$$

Иногда при решении обратной задачи предполагают, что погрешность всех аргументов одна и та же, т.е. $h_{a_1} = h_{a_2} = \ldots = h_{a_n} = h.$

Тогда

$$h_{a_i} = h \le \frac{h_{\tilde{y}}}{\sum\limits_{i=1}^{n}\left|\left(\frac{\partial f}{\partial x_i}\right)_A\right|}, \quad i=\overline{1,n}.$$

Пример 4.9. С каким числом десятичных знаков следует представить дроби, чтобы сумма

$$S = 1/23 - 1/28 + 1/3 - 1/7$$

могла быть получена с точно-стью до 0,001?

Решение

Обозначим

$$x_1 = 1/23,\ x_2 = 1/28,\ x_3 = 1/3,\ x_4 = 1/7.$$

Тогда $S = x_1 - x_2 + x_3 - x_4.$ Обозначим приближенное значение суммы, по-лучаемое как сумма приближенных значений $\tilde{x}_i$ аргументов x_i буквой $\tilde{S}$, $\left|\tilde{S} - S\right| \le h_{\tilde{S}}$. Тогда $h_{\tilde{S}}$ не должно превосходить $0{,}001$. Положим $h_{\tilde{S}} = 0{,}001.$

Требуется установить допустимые границы погрешностей $h_{\tilde{x}_i}$ $(i = 1,2,3,4)$. По принципу равных влияний

$$h_{\tilde{x}_i} \le \frac{h_{\tilde{S}}}{4} = 0{,}00025.$$

Таким образом, дроби следует представлять в десятичном виде так, чтобы модуль погрешности не превосходил 0,00025, т.е. с четырьмя десятичными знаками. При этом модуль погрешности не будет превосходить не только 0,00025 но и 0,00005, а следовательно, вычисление суммы с указанной точностью будет заведомо обеспечено.

4.7 Метод границ

Метод границ позволяет установить границы, в которых заключено значение, вычисляемое по формуле, если известны границы, в которых заключены значения параметров, содержащихся в этой формуле.

Рассмотрим сначала метод границ для четырех действий арифметики, а также действий возведения в целую положительную степень и извлечения корня.

Для этого нижнюю и верхнюю границы, в которых заключено значение некоторой переменной будем обозначать соответственно НГ и ВГ.

Например,

$НГ_x$, $ВГ_x$ – границы для x,

$НГ_y$, $ВГ_y$ – границы для y,

$НГ_{xy}$, $ВГ_{xy}$ – границы для произведения xy.

Теорема 4.8. Сумма верхних (нижних) границ слагаемых является верхней (нижней) границей суммы:

$$НГ_{x+y} = НГ_x + НГ_y ;$$

$$ВГ_{x+y} = ВГ_x + ВГ_y.$$

Доказательство

Действительно, если справедливы неравенства

$$НГ_x \le x \le ВГ_x \text{ и } НГ_y \le y \le ВГ_y,$$

то по теореме о сложении неравенств, будет справедливо неравенство

$$НГ_x + ВГ_x \le x + y \le ВГ_x + ВГ_y .$$

Теорема доказана.

Пример 4.10. Найти сумму $x+y$, если известны границы, в которых за-ключены x и y

$$5{,}7 \le x \le 8{,}4;\ 3{,}3 \le y \le 5{,}4.$$

Решение

$$НГ_x = 5{,}7;\quad ВГ_x = 8{,}4; НГ_y = 3{,}3;\ ВГ_y = 5{,}4.$$

Отсюда

$$НГ_{x+y} = НГ_x + НГ_y = 9.0,\quad ВГ_{x+y} = ВГ_x + ВГ_y = 13{,}8,$$

то есть $9{,}0 \le x + y \le 13{,}8$.

Теорема 4.9. Разность верхней (нижней) границы уменьшаемого и нижней (верхней) границы вычитаемого является верхней (нижней) границей разности:

$$НГ_{х\text{-}у} = НГ_х - ВГ_у;$$

$$ВГ_{х\text{-}у} = ВГ_х - НГ_у.$$

Доказательство

По теореме о вычитании неравенств противоположного смысла

$$НГ_х - ВГ_у \le x - y \le ВГ_х - НГ_у.$$

Теорема доказана.

Пример 4.11. Найти разность $x - y$, если известны границы, в которых заключены x и y: $5{,}2 \le x \le 8{,}8;\ 3{,}2 \le y \le 5{,}0.$

Решение

$$НГ_х = 5{,}2;\quad ВГ_х = 8{,}8;\ НГ_у = 3{,}2;\quad ВГ_у = 5{,}0.$$

Отсюда

$$НГ_{х\text{-}у} = НГ_х - ВГ_у = 0{,}2;\ ВГ_{х\text{-}у} = ВГ_х - НГ_у = 5{,}6$$

и $0{,}2 \le x - y \le 5{,}6$.

Теорема 4.10. Пусть нижние границы сомножителей неотрицательные. Тогда произведение верхних (нижних) границ сомножителей является верхней (нижней) границей произведения:

$$ВГ_{ху} = ВГ_х \cdot ВГ_у,$$

$$НГ_{ху} = НГ_х \cdot НГ_у.$$

Доказательство

Так как

$$НГ_х \le x \le ВГ_х \text{ и } НГ_у \le y \le ВГ_у,$$

то по теореме о произведении неравенств одного смысла имеем

$$НГ_х \cdot НГ_у \le xy \le ВГ_х \cdot ВГ_у.$$

Теорема доказана.

Пример 4.12. Вычислить произведение xy, если известно:

$$3{,}7 \le x \le 4{,}1;\ 1{,}1 \le y \le 1{,}4.$$

Решение

$$НГ_{ху} = НГ_х \cdot НГ_у = 4{,}07;\ ВГ_{ху} = ВГ_х \cdot ВГ_у = 5{,}74;$$

$$4{,}07 \le xy \le 5{,}74.$$

Теорема 4.11. Пусть нижняя граница числа x неотрицательна, а n *– целое положительное число. Тогда* $n^{-ая}$ *степень нижней (верхней) границы числа x является нижней (верхней) границей* $n^{-ой}$ *степени числа x*

$$НГ_x^n = (НГ_x)^n;\ ВГ_x^n = (ВГ_x)^n.$$

Пример 4.13. Найти границы x^3, если $3{,}7 \le x \le 3{,}8$.

Решение

$$(НГ_x)^3 = 50{,}653 > 50;$$

$$НГ_x^3 = 50;\ ВГ_x^3 = 55;$$

$$50 \le x^3 \le 55.$$

Теорема 4.12. Если нижняя граница числа x неотрицательна, то при извлечении корня $n^{-ой}$ *степени из него корень этой степени из верхней (нижней) границы является верхней (нижней) границей корня из числа x*

$$НГ_{\sqrt[n]{x}} = \sqrt[n]{НГ_x},\ ВГ_{\sqrt[n]{x}} = \sqrt[n]{ВГ_x}.$$

Пример 4.14. Вычислить $\sqrt{x}$, если $5{,}74 \le x \le 5{,}80$.

Решение

$$НГ_x = 5{,}74;\ ВГ_x = 5{,}80.$$

Отсюда $\sqrt{НГ_x} = 2{,}39\ldots > 2{,}39$ и $НГ_{\sqrt{x}} = 2{,}39$;

$$\sqrt{ВГ_x} = 2{,}40\ldots < 2{,}41;\quad ВГ_{\sqrt{x}} = 2{,}41;\quad 2{,}39 \le \sqrt{x} \le 2{,}41.$$

Теорема 4.13. Пусть нижняя граница делителя положительна. Тогда частное верхней (нижней) границы делимого и нижней (верхней) границы делителя является верхней (нижней) границей частного чисел:

$$ВГ_{\frac{x}{y}} = \frac{ВГ_x}{НГ_y},\ НГ_{\frac{x}{y}} = \frac{НГ_x}{ВГ_y}.$$

Доказательство

Так как

$$НГ_y > 0, \text{ то } \frac{1}{ВГ_y} \le \frac{1}{y} \le \frac{1}{НГ_y},$$

тогда

$$\frac{НГ_x}{ВГ_y} \le \frac{x}{y} \le \frac{ВГ_x}{НГ_y}.$$

Теорема доказана.

Пример 4.15. Вычислить $\frac{x}{y}$, если $5{,}7 \le x \le 8{,}7$ и $3{,}5 \le y \le 4{,}1$.

Решение

$$НГ_x = 5{,}7;\quad ВГ_x = 8{,}7;\quad НГ_y = 3{,}5;\quad ВГ_y = 4{,}1.$$

$$НГ_{\frac{x}{y}} = 1{,}39;\qquad ВГ_{\frac{x}{y}} = 2{,}49;\qquad 1{,}39 \le \frac{x}{y} \le 2{,}49.$$

Метод границ для функции

Теорема 4.14. Пусть функция $y = f(x_1, x_2, \ldots, x_n)$ *определена в области G, непрерывна и монотонна в ней по каждому аргументу* x_i, $i = \overline{1,n}$.

Пусть параллелепипед $G^*: НГ_{x_i} \le x_i \le ВГ_{x_i}$, $i = \overline{1,n}$, *целиком содержится в G.*

Тогда границы, в которых заключено значение функции $y = f(x_1, x_2, \ldots, x_n)$ *для точек, принадлежащих* G^*, *могут быть вычислены по формулам*:

$НГ_y = f(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$ и $ВГ_y = f(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n)$, *где*

$$\underline{x}_i = \begin{cases} НГ_{x_i}, & \text{если функция } f \text{ возрастающая по этому аргументу } x_i, \\ ВГ_{x_i}, & \text{если функция } f \text{ убывающая по этому аргументу } x_i; \end{cases}$$

$$\overline{x}_i = \begin{cases} ВГ_{x_i}, & \text{если функция } f \text{ возрастающая по этому аргументу } x_i, \\ НГ_{x_i}, & \text{если функция } f \text{ убывающая по этому аргументу } x_i. \end{cases}$$

При этом следует заметить, что значения $\underline{x}_i$ *(для всех* i*) и* $f(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$ *можно округлять лишь в сторону уменьшения нижней границы функции, а значения* $\overline{x}_i$ *и* $f(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n)$ *можно округлять лишь в сторону увеличения верхней границы функции.*

Доказательство

Действительно, если $f(x_1, x_2, \ldots, x_n)$ возрастающая по x_1, то для любой точки параллелепипеда G^* будет справедливо неравенство:

$$f(НГ_{x_1}, x_2, \ldots, x_n) \le f(x_1, x_2, \ldots, x_n) \le f(ВГ_{x_1}, x_2, \ldots, x_n).$$

Если $f(x_1, x_2, \ldots, x_n)$ убывающая по x_1, то

$$f(ВГ_{x_1}, x_2, \ldots, x_n) \le f(x_1, x_2, \ldots, x_n) \le f(НГ_{x_1}, x_2, \ldots, x_n).$$

Обозначим

$$\underline{x}_1 = \begin{cases} НГ_{x_1}, \text{если функция } f \text{ возрастающая по } x_1, \\ ВГ_{x_1}, \text{если функция } f \text{ убывающая по } x_1. \end{cases}$$

Тогда можно записать в общем виде

$$f\left(\underline{x_1}, x_2, \ldots, x_n\right) \le f\left(x_1, x_2, \ldots, x_n\right) \le f\left(\overline{x_1}, x_2, \ldots, x_n\right).$$

Так как функция монотонна по x_2, то в G^* справедливо:

$f\left(\underline{x_1}, \underline{x_2}, \ldots, x_n\right) \le f\left(\underline{x}_1, x_2, \ldots, x_n\right)$ и

$f\left(\overline{x_1}, x_2, x_3, \ldots, x_n\right) \le f\left(\overline{x_1}, \overline{x_2}, x_3, \ldots, x_n\right)$. Отсюда

$$f\left(\underline{x_1}, \underline{x_2}, x_3, \ldots, x_n\right) \le f\left(x_1, x_2, \ldots, x_n\right) \le f\left(\overline{x}_1, \overline{x_2}, x_3, \ldots, x_n\right).$$

Проводя аналогичные рассуждения для всех остальных аргументов, получим утверждение теоремы. Теорема доказана.

Пример 4.16. Вычислить границы, в которых заключены значения функции $A(x, y, z) = \dfrac{(x-y)z}{x+y}$, если $G^*: \begin{cases} 2{,}57 \le x \le 2{,}58 \\ 1{,}45 \le y \le 1{,}46 \\ 8{,}33 \le z \le 8{,}34. \end{cases}$

Решение

$$\frac{\partial A}{\partial z} = \frac{x-y}{x+y}; \quad \frac{\partial A}{\partial x} = \frac{2zy}{(x+y)^2}; \quad \frac{\partial A}{\partial y} = \frac{-2xz}{(x+y)^2}.$$

Так как для любой точки $(x, y, z) \in G^*$ имеют место неравенства $\dfrac{\partial A}{\partial z} > 0$; $\dfrac{\partial A}{\partial x} > 0$; $\dfrac{\partial A}{\partial y} < 0$, то функция $A(x, y, z)$ является в G^* возрастающей по x и z и убывающей по y.

Обозначим согласно теореме 4.14:

$$\underline{x} = 2{,}57; \ \overline{x} = 2{,}58;$$

$$\underline{y} = 1{,}46; \ \overline{y} = 1{,}45;$$

$$\underline{z} = 8{,}33; \ \overline{z} = 8{,}34.$$

Тогда $НГ_A = A(\underline{x}, \underline{y}, \underline{z}) = 2{,}29$; $ВГ_A = A(\overline{x}, \overline{y}, \overline{z}) = 2{,}34$.

Отсюда $A = 2{,}315 \pm 0{,}025$ или $A \doteq 2{,}3$, так как

$$0{,}025 + 0{,}015 = 0{,}04 \le 0{,}05.$$

ГЛАВА 5 РЕШЕНИЕ УРАВНЕНИЙ С ОДНИМ НЕИЗВЕСТНЫМ В ПРОСТРАНСТВЕ *R* ДЕЙСТВИТЕЛЬНЫХ ЧИСЕЛ

5.1 Понятие корректно и некорректно поставленных задач

При приближённом решении математических или прикладных задач весьма существенным является вопрос о том, корректна ли решаемая задача. Большинство некорректных задач может быть приведено к уравнению I рода, имеющему вид:

$$Ax = y,\ x \in X,\ y \in Y, \tag{5.1}$$

в котором по заданному, не обязательно линейному, оператору A, действующему из пространства X в пространство Y, и по заданному элементу $y \in Y$ требуется определить решение x в пространстве X.

Пространства X и Y будем считать метрическими, а в особо оговариваемых случаях, банаховыми или даже гильбертовыми.

Определение 5.1. Задача определения решения $x = R(y)$ *из пространства* X *по исходным данным* $y \in Y$ *называется устойчивой на пространствах* X *и* Y*, если для любого числа* $\varepsilon > 0$ *можно указать такое число* $\delta(\varepsilon) > 0$*, что из неравенства* $\rho_Y(y_1, y_2) \le \delta(\varepsilon)$ *следует, что*

$$\rho_X(x_1, x_2) \le \varepsilon,$$

где $x_1 = Ry_1, x_2 = Ry_2, x_1, x_2 \in X, y_1, y_2 \in Y.$

По-другому, если бесконечно малым вариациям правой части y соответствуют бесконечно малые вариации x. Помимо этого говорят вместо устойчивости x в пространстве X о непрерывной зависимости x от $y \in Y$.

Определение 5.2. Следуя Ж. Адамару, задачу отыскания $x \in X$ *из уравнения* (5.1) *называют корректной (корректно поставленной), если при любой фиксированной правой части* $y = y_0$ *из* Y *ее решение:*

а) существует в пространстве X *;*

б) единственно в X;

в) устойчиво в X.

Если хотя бы одно из условий не выполняется, то задачу называют некорректной (*некорректно поставленной*).

Таким образом, корректность задачи связана с наличием обратного оператора A^{-1}, определенного и непрерывного на всем пространстве Y.

Условия корректности задачи и особенно ее решения представляются настолько естественными с точки зрения приложений математики, что долгое время некорректные задачи считались не имеющими физического смысла и поэтому не изучались. Между тем практика и научные исследования стали одну за другой выдвигать некорректные задачи.

Так задача Коши для уравнения Лапласа оказалась важной для географических методов разведки полезных ископаемых, а задача Коши для эллиптических уравнений и систем – для сверхзвуковой аэродинамики. Уравнение Фредгольма I рода приходится решать в спектроскопии и в обратных задачах теории потенциала. Некорректна и задача теплопроводности с большинство так называемых обратных задач, в которых по результатам действий какого-либо физического поля или процесса определяются первоначальные характеристики самого этого поля или процесса.

Если не изменить постановку неустойчивых задач, то обычные методы, применяемые для решения корректных задач, оказываются, естественно, непригодными для решения некорректных задач, так как сколь бы малой ни была бы погрешность исходных данных нельзя быть уверенным в малости погрешности решения. Поэтому потребности практики в решении некорректных задач привели к необходимости пересмотреть классическое понятие корректности и выработать более широкий и приспособленный к реальным нуждам подход.

Приведем следующее определение.

Определение 5.3. Назовем задачу (5.1) *<u>корректной по Тихонову</u> на множестве* $M \subset X$, *а само множество* M – *её множеством корректности, если:*

а) точное решение задачи существует в классе M,

б) принадлежащее множеству M *решение задачи единственно для любой правой части* y *из множества* $N = AM \subset Y$,

в) принадлежащее множеству M *решение задачи устойчиво относительно любой правой части* y *из множества* N.

В случае нарушения любого из этих условий задачу (5.1) *называют <u>некорректной</u>.*

Начало изучению некорректных задач и методов их решения было положено в 1943 году А.Н. Тихоновым. После работ А.Н. Тихонова систематическое изучение некорректных задач началось в 60-х годах, но особый размах оно приняло в последние 50 лет. Основные результаты отражены в монографиях М.М. Лаврентьева, А.Н. Тихонова и В.Я. Тананы, О.А. Лисковца, В.Ф. Савчука, О.В. Матысика и других.

Для решения некорректных задач нашли большое применение метод квазирешений В.К. Иванова, метод регуляризации А.Н. Тихонова, метод невязки Д.Л. Филипса и В.К. Иванова. Особое место среди решения некорректных задач занимают итеративные методы. Их изучению посвящены работы М.М. Лаврентьева, В.Н. Страхова, Ю.Т. Антохина, Б.А. Андреева, О.А. Лисковца, Я.В. Константиновой и других. Во всех работах того времени число итераций выбирается априорно.

И.В. Емелин и М.А. Красносельский предложили для метода простых итераций выбирать число итераций апостериорно, используя правила останова по невязке и по соседним приближениям. Дальнейшее развитие идеи их работ получили в работах Г.М. Вайникко, предложившего правила останова, которые превращают метод итерации в регуляризующий алгоритм для задачи (8.1), не требуя при этом знания истокопредставимости точного решения, но в случае истокопредставимости обеспечивают оптимальную в классе скорость сходимости.

5.2 Метод дихотомии

Исследование уравнения.

Пусть задана непрерывная функция $f(x)$ и требуется найти все или некоторые корни уравнения

$$f(x)=0. \qquad (5.2)$$

Корнем уравнения (5.2) называется такое значение $x=x^*$, что

$$f(x^*)=0.$$

Если функция $f(x)$ непрерывна на $[a,b]$ и ее значения на концах отрезка разных знаков, $f(a)f(b)<0$, то на $[a,b]$ найдётся, по крайней мере, один корень уравнения.

Отделить корень уравнения – значит, найти такой интервал, внутри которого имеется корень данного уравнения и притом единственный на данном интервале. Для отделения корней уравнения (5.2) применяют следующий критерий: если на отрезке $[a,b]$ функция $f(x)$ непрерывна и монотонна, а ее значения на концах отрезка имеют разные знаки, то на рассматриваемом отрезке существует и только один корень данного уравнения. Достаточный признак монотонности функции на отрезке – сохранение знака производной функции.

Отделение корней уравнения (5.2) можно выполнить графически.

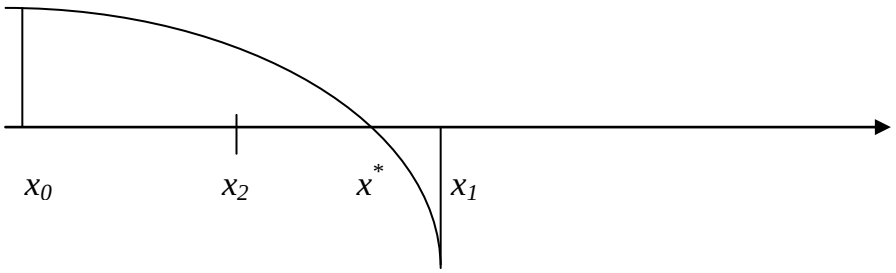


Рисунок 23

Для этого надо построить график функции уравнения $f(x)$, по которому можно судить, в каких интервалах находятся точки его пересечения с осью OX. Наиболее совершенным способом отделения корней является метод Штурма.

Дихотомия (метод деления отрезка пополам).

Пусть мы нашли такие точки x_0 и x_1, что $f(x_0)f(x_1)\le 0$, т.е. на отрезке $[x_0, x_1]$ лежит не менее одного корня уравнения. Найдем середину отрезка

$$x_2 = \frac{x_0 + x_1}{2}$$

и вычислим $f(x_2)$. Из двух половин отрезка выберем ту, для которой $f(x_2)\cdot f(x_{гран})\le 0$, ибо один из корней лежит на этой половине. Затем новый отрезок опять делим пополам и выберем ту половину, на концах которой функция имеет разные знаки и т.д. (рисунок 23).

Если требуется найти корень с точностью до ε, то продолжаем деление до тех пор, пока длина отрезка не станет меньше 2ε. Тогда середина последнего отрезка даёт значение корня с требуемой точностью.

Дихотомия проста и очень надёжна: к простому корню она сходится для любых непрерывных функций $f(x)$, в том числе недифференцируемых, при этом она устойчива к ошибкам округления. Скорость сходимости невелика: за одну итерацию точность увеличивается примерно вдвое.

Недостатки метода. Для начала расчетов надо найти отрезок, на котором функция меняет знак. Если в этом отрезке несколько корней, то заранее неизвестно, к какому из них сойдется процесс. Метод не применим к корням четной кратности. Для корней нечетной кратности он сходится, но менее точен и хуже устойчив к ошибкам округления возникающих при вычислениях $f(x)$. Наконец, на системы дихотомия не распространяется.

5.3 Метод простой итерации для алгебраических и трансцендентных уравнений

Применим принцип сжимающих отображений к исследованию сходимости итерационных методов решения скалярного уравнения

$$f(x)=0, \tag{5.3}$$

где $f(x)$ – вещественная функция вещественного аргумента.

Для решения алгебраических и трансцендентных уравнений вида (5.3) разработано много различных итерационных методов. Чаще всего по функции $f(x)$ строят функцию $\varphi(x)$ такую, что искомый корень $x=x^*$ уравнения (5.3) является и корнем уравнения

$$x=\varphi(x), \tag{5.4}$$

и затем строят последовательность $\{x_k\}$ с помощью соотношения

$$x_k=\varphi(x_{k-1}),\ k=1,2,\ldots, \tag{5.5}$$

начиная с некоторого приближения x_0. Сходимость последовательности обеспечивается соответствующим выбором функции φ и начального приближения x_0. Выбирая различными способами функцию φ (зависящую от $f(x)$), можно получить различные итерационные методы.

Теорема 5.1. Пусть уравнение (5.4) *имеет корень* $x=x^*$ *и в некоторой окрестности* R *этого корня* ($R=\{x:|x-x^*|\le r\}$) *функция* $\varphi(x)$ *удовлетворяет условию Липшица* $|\varphi(x)-\varphi(x')|\le q|x-x'|$, *где* $0<q<1$. *Тогда, каково бы ни было* $x_0\in R$, *последовательность* (5.5) *сходится к* x^*, *причем скорость сходимости характеризуется неравенством*:

$$|x-x^*|\le q^k|x_0-x^*|.$$

Доказательство

Совокупность точек отрезка R образует полное метрическое пространство, если расстояние между точками x и y определить соотношением $\rho(x,y)=|x-y|$. Если $x\in R$, то $y=\varphi(x)$ также принадлежит R, ибо

$$\rho(y,x^*)=|y-x^*|=|\varphi(x)-\varphi(x^*)|\le q|x-x^*|<r.$$

Таким образом, функция $y=\varphi(x)$ задаёт некоторое отображение отрезка R в себя.

Это отображение является сжимающим, поскольку для любых x, $x' \in R$,

$$\rho(\varphi(x), \varphi(x')) = |\varphi(x) - \varphi(x')| \le q|x - x'| = q\rho(x, x'),$$

где $0 < q < 1$. Тогда $x = x^*$, так как в силу принципа сжимающих отображений в R существует одна и только одна неподвижная точка отображения, определяемого функцией $\varphi(x)$. Эту точку можно получить как предел последовательности (5.5) при любом $x_0 \in R$.

Используя условие Липшица, имеем:

$$|x_k - x^*| = |\varphi(x_{k-1}) - \varphi(x^*)| \le q|x_{k-1} - x^*|, \text{ т.е.}$$

$$|x_k - x^*| \le q|x_{k-1} - x^*| \le q^2|x_{k-2} - x^*| \le \dots q^k|x_0 - x^*|.$$

Таким образом, $\{x_k\}$ сходится к x^* со скоростью геометрической прогрессии со знаменателем q. Скорость сходимости метода простой итерации линейная. Теорема доказана.

Замечание 5.1. Условие Липшица с константой $q < 1$ выполняется для функции $\varphi(x)$ на отрезке R, если эта функция имеет на R производную $\varphi'(x)$, удовлетворяющую неравенству:

$$|\varphi'(x)| \le q < 1. \qquad (5.6)$$

Действительно, в этом случае для любых x, y из R имеем:

$$|\varphi(x) - \varphi(y)| = |\varphi'(\xi)||x - y| \le q|x - y|, \ (\xi \in R).$$

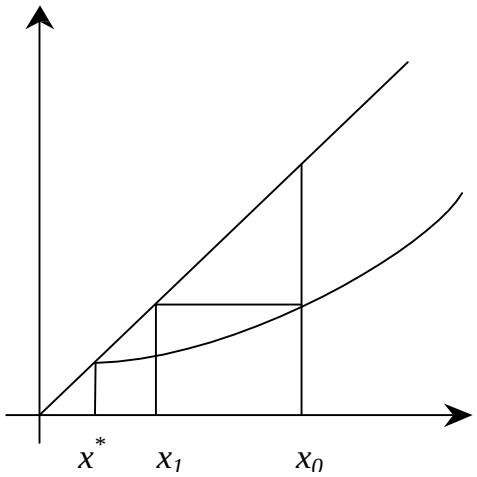


Рисунок 24

Формула (5.4) определяет численный метод решения уравнения $x = \varphi(x)$, который обычно называют *методом простой итерации.*

В нашем случае метод простой итерации имеет простую геометрическую интерпретацию. На рисунке 24 изображено поведение последовательных приближений для случая, когда в некоторой окрестности корня выполняется условие $0<\varphi'(x)\le q<1$. Из этой окрестности и выбирается начальное приближение x_0.

Отметим, что при выполнении условий доказанной теоремы 5.1 метод простой итерации сходится при любом выборе начального приближения x_0 из *R*. Благодаря этому он является *самоисправляющимся:* допущенная при вычислениях ошибка (не выводящая за пределы отрезка *R*) не влияет на конечный результат, поскольку ошибочное значение можно рассматривать как новое начальное приближение.

Свойство самоисправления делает метод простой итерации одним из надежнейших методов вычислений. Метод простой итерации очень удобен для реализации его на электронных вычислительных машинах.

В дальнейшем мы будем предполагать, что уравнение (5.3) имеет только *изолированные корни,* то есть для каждого корня уравнения $f(x)=0$ существует окрестность, не содержащая других корней этого уравнения.

Приближенное нахождение изолированных вещественных корней уравнения (5.3) обычно складывается из двух этапов:

1) *отделение корней,* то есть установление таких промежутков, в каждом из которых содержится один и только один корень уравнения (5.3);

2) *уточнение приближенных значений корней,* то есть вычисление каждого корня тем или иным численным методом с заданной точностью.

5.4 Метод хорд

Предположим теперь, что на некотором отрезке $[a,b]$ функция $f(x)$ непрерывна вместе с производными $f'(x)$ и $f''(x)$, причем $f(a)f(b)<0$, производные не меняют своего знака на интервале (a,b). Это означает, что на отрезке $[a,b]$ существует единственный корень $x=x^*$ уравнения (5.3).

Если функция $\psi(x)$ непрерывна в некоторой окрестности x^*, то уравнение (5.4), где

$$\varphi(x)=x-\psi(x)f(x), \qquad (5.7)$$

также имеет x^* своим корнем. Функцию $\psi(x)$ в (5.7) можно подобрать так, что итерационный процесс (5.5) для уравнения

$$x=\varphi(x)=x-\psi(x)f(x)$$

будет сходящимся.

Пусть x_0 – некоторая точка из отрезка $[a,b]$, выбранного, как указано выше, причем $f(x_0)f''(x_0)>0$. В (5.7) в качестве функции $\psi(x)$ возьмем функцию

$$\psi(x)=\frac{x-x_0}{f(x)-f(x_0)}.$$

Тогда

$$\varphi(x)=x-\frac{x-x_0}{f(x)-f(x_0)}f(x).$$

И уравнение

$$x=\frac{x_0 f(x)-xf(x_0)}{f(x)-f(x_0)}$$

также имеет корень x^*. Примем за начальное приближение любую, достаточно близкую к x^* точку x_1 из отрезка $[a,b]$, в которой $f(x_1)$ имеет знак, противоположный знаку $f(x_0)$, а последующие приближения будем строить обычным способом:

$$x_{k+1}=\frac{x_0 f(x_k)-x_k f(x_0)}{f(x_k)-f(x_0)},\quad k=1,2,\ldots. \qquad (5.8)$$

Для доказательства сходимости процесса оценим $\varphi'(x)$. Имеем

$$\varphi'(x^*)=\frac{f(x_0)+(x^*-x_0)f'(x^*)}{f(x_0)}.$$

По формуле Тейлора

$$f(x)=f(x^*)+(x-x^*)f'(x^*)+\frac{(x-x^*)^2}{2}f''(\xi),$$

где ξ заключено между x^* и x. Полагая $x=x_0$, получаем

$$f(x_0)+(x^*-x_0)f'(x^*)=\frac{(x_0-x^*)^2}{2}f''(\xi).$$

Таким образом,

$$\varphi'(x^*)=\frac{(x_0-x^*)^2}{2f(x_0)}f''(\xi).$$

Так как

$$\lim_{x_0\to x^*}\frac{(x_0-x^*)^2}{2f(x_0)}f''(\xi)=0,$$

то $\varphi'(x^*)$ – малое число, если x_0 достаточно близко к x^*. В силу непрерывности $\varphi'(x)$ на $[a,b]$ найдётся окрестность $\cup_{x^*}$ точки x^* такая, что для всех $x \in \cup_{x^*}$ будет иметь место неравенство:

$$|\varphi'(x)| \le q < 1;$$

и если x_1 взято из этой окрестности, то последовательность, построенная по формулам (11.2), будет сходиться к x^*.

Для оценки точности приближения пользоваться формулой:

$$|x_k - x^*| \le \frac{|f(x_k)|}{m_1}, \qquad (5.9)$$

где $m_1 = \min_{[a,b]} |f'(x)|$.

Действительно

$$f(x_k) = f(x_k) - f(x^*) = f'(\xi)(x_k - x^*),$$

где ξ заключено между x_k и x^*. Отсюда

$$|x_k - x^*| = \left|\frac{f(x_k)}{f'(\xi)}\right| \le \frac{|f(x_k)|}{m_1}.$$

Скорость сходимости метода хорд линейная.

На практике применяются следующие формулы. Если известно $(n-1)$-е приближение, то n - е можно вычислить по формуле

$$x_n = \frac{bf(x_{n-1}) - x_{n-1}f(b)}{f(x_{n-1}) - f(b)} (n = 1, 2, 3, \ldots) \qquad (5.10)$$

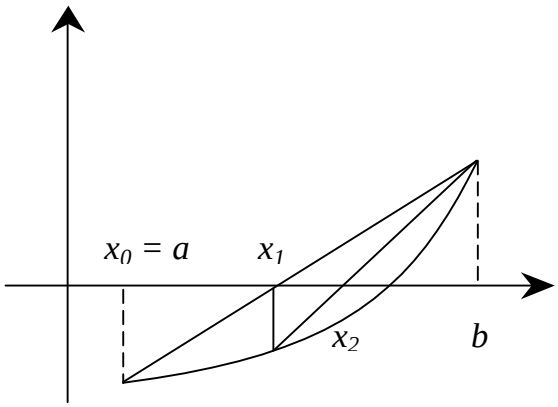


Рисунок 25

(рисунок 25), когда

$$f(b)f''(x) > 0. \qquad (5.11)$$

Или по формуле

$$x_n = \frac{af(x_{n-1}) - x_{n-1} f(a)}{f(x_{n-1}) - f(a)} \tag{5.12}$$

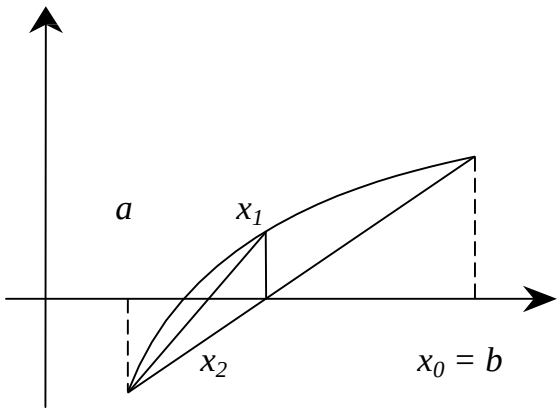


Рисунок 26

(рисунок 26), когда

$$f(a) f''(x) > 0. \tag{5.13}$$

В первом случае за начальное приближение принимается a, т.е. $x_0 = a$, во втором – b, т.е. $x_0 = b$

5.5 Метод касательных

Классическим методом решения скалярного уравнения $f(x) = 0$ является *метод Ньютона.*

Положим в (5.7) $\psi(x) = \frac{1}{f'(x)}$. Тогда

$$\varphi(x) = x - \frac{f(x)}{f'(x)}$$

и нужно найти корень x^* уравнения $x = x - \frac{f(x)}{f'(x)}$. По-прежнему будем предполагать, что на отрезке $[a,b]$ функция $f(x)$ имеет непрерывные, не обращающиеся в нуль производные $f'(x)$ и $f''(x)$ и $f(a) f(b) < 0$.

Имеем

$$\varphi'(x) = 1 - \frac{(f'(x))^2 - f(x) f''(x)}{(f'(x))^2},$$

так что $\varphi'(x^*) = 0$.

В силу непрерывности $\varphi'(x)$ найдется окрестность $\bigcup_{x^*}$ точки x^*, такая, что для всех $x \in \bigcup_{x^*}$ имеет место неравенство:

$$|\varphi'(x)| \le q < 1.$$

Если начальное приближение x_0 взято из этой окрестности, то последовательность $\{x_k\}$, построенная по формулам

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots,$$

будет сходиться к x^*. Начальное приближение x_0 целесообразно выбирать так, чтобы было выполнено условие:

$$f(x_0) f''(x_0) > 0.$$

Если за начальное приближение в методе Ньютона взять точку x_0, такую что $f(x_0) f''(x_0) < 0$, то мы можем не прийти к корню $x = x^*$, если только начальное приближение не очень хорошее.

Оценим скорость сходимости метода Ньютона. По формуле Тейлора:

$$f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2} f''(\xi)(x^* - x_k)^2$$

(ξ заключено между x^* и x_k). Так как $f(x^*) = 0$, то

$$\frac{f(x_k)}{f'(x_k)} = x_k - x^* - \frac{1}{2} \frac{f''(\xi)}{f'(x_k)} (x^* - x_k)^2.$$

Следовательно,

$$x_{k+1} - x^* = x_k - x^* - \frac{f(x_k)}{f'(x_k)} = \frac{1}{2} \frac{f''(\xi)}{f'(x_k)} (x^* - x_k)^2.$$

Если $m_1 = \min\limits_{[a,b]} |f'(x)|$, а $M_2 = \max\limits_{[a,b]} |f''(x)|$, то

$$|x_{k+1} - x^*| \le \frac{M_2}{2m_1} |x_k - x^*|^2. \qquad (5.14)$$

Неравенство (5.14) гарантирует быструю сходимость метода Ньютона, если начальное приближение x_0 таково, что

$$\frac{M_2}{2m_1} |x_0 - x^*| \le c < 1.$$

Действительно в этом случае из неравенства (5.14) следует

$$\left|x_k - x^*\right| \le \frac{2m_1}{M_2} c^{2k}.$$

Скорость сходимости метода Ньютона квадратичная.

Геометрический смысл метода касательных

Метод Ньютона, который часто называют также *методом касательных,* состоит в следующем. Пусть на отрезке $[a,b]$ находится единственный корень уравнения $f(x)=0$. Проведем касательную к кривой $y=f(x)$ в точке $A[a, f(a)]$ до пересечения с осью Ox (рисунок 27).

Уравнение касательной, проходящей через точку A, будет следующим

$$y = f(a) + f'(a)(x-a).$$

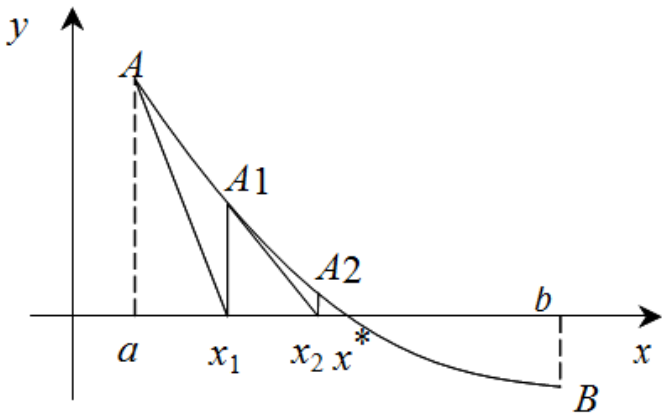


Рисунок 27

Если

$$f'(a) \neq 0,$$

то из этого уравнения (при $y=0$) находим абсциссу x_1 точки пересечения касательной с осью Ox:

$$x_1 = a - \frac{f(a)}{f'(a)}.$$

Абсциссу x_1 точки пересечения можно взять в качестве приближенного значения корня. Если проведем касательную через соответствующую точку $A_1[x_1, f(x_1)]$ и найдем точку пересечения с осью Ox, получим второе приближение корня x_2. Аналогично определяются последующие приближения.

Применяя метод касательных, n – e приближение находят по формуле

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}, \quad (n = 1,2,3,\ldots), \tag{5.15}$$

причем за нулевое приближение x_0 принимается такое значение из отрезка $[a,b]$, для которого выполняется условие

$$f(x_0) \cdot f''(x_0) > 0. \tag{5.16}$$

5.6 Метод секущих

К числу достаточно эффективных итерационных методов решения нелинейного уравнения $f(x) = 0$ можно отнести метод, описываемый формулой

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n),$$

который, в силу его геометрической интерпретации, носит название метода секущих. Скорость сходимости этого процесса сверхлинейная, точнее для $n-$го приближения x_n и x^* – решения уравнения $f(x) = 0$ имеет место соотношение

$$\left|x_n - x^*\right| \le \alpha \left|x_{n-1} - x^*\right|^{\alpha}, \quad \alpha = \frac{\sqrt{5}+1}{2}.$$

В связи с тем, что в методе секущих на каждом шаге процесса находится лишь одно значение функции, то при одинаковом объеме вычислительной работы метод эффективнее метода хорд и метода Ньютона.

Недостаток метода состоит в том, что при подходе к решению разность $f(x_n) - f(x_{n-1})$ может стать столь малой, что будет происходить потеря точности, что выразится в хаотическом блуждании элементов последовательности $\{x_n\}$, полученной по методу секущих вокруг корня. Этот феномен носит название “разболтки”.

Для предотвращения “разболтки” применяется принцип Гарвика. Суть его в следующем. Выбираем $0 \ne \varepsilon << 1$ и ведем счет, следя за тем, чтобы $\left|x_n - x_{n-1}\right|$ было меньше ε. Как только мы этого достигаем, делаем еще несколько шагов до момента, когда убывание $\left|x_к - x_{k-1}\right|$ прекратится и в качестве приближенного значения корня прини-маем x_{k_n}, предшествующее тому x_k, при котором началось возрастание $\left|x_к - x_{k-1}\right|$.

Метод секущих применяют в комплексе с методом Ньютона (или с методом хорд и Ньютона). Для методов итераций, Ньютона, хорд, секущих используются следующие правила останова.

Правило останова по соседним приближениям

Задается уровень останова $\varepsilon > 0$ и момент останова m итерационной процедуры определяется условием

$$|x_n - x_{n+1}| > \varepsilon, \quad (n < m), \qquad |x_m - x_{m+1}| \le \varepsilon.$$

Правило останова по невязке

Задается уровень останова $\varepsilon > 0$ и момент останова m итерационной процедуры определяется условием

$$|f(x_n)| > \varepsilon, \quad (n < m), \qquad |f(x_m)| \le \varepsilon.$$

Глава 6 Вычисление собственных векторов и собственных значений матрицы

6.1 Приведение симметрической матрицы к трёхдиагональной форме методом вращений

Некоторые определения и теоремы. Для того чтобы найти собственный вектор матрицы A (размерности $n \times n$), соответствующий собственному значению λ, нужно решить однородную систему

$$(A - \lambda E)x = 0.$$

Эта система имеет ненулевое решение тогда и только тогда, когда $|A - \lambda E| = 0$. Это алгебраическое уравнение $n^{-\text{ой}}$ степени относительно λ. Корни его и являются собственными значениями матрицы A.

Существует много приближенных методов для решения алгебраических уравнений: методы Ньютона, простых итераций, секущих, хорд и т.д.

При нахождении же собственных векторов (и собственных значений) сложности возникают при вычислении самих коэффициентов *характеристического многочлена* (если матрица A произвольная). Здесь мы разберем случай, когда $|A - \lambda E|$ матрица A симметричная.

Из курса линейной алгебры известно, что если $A = A^T$, то все ее собственные значения – действительные числа. Построение характеристического многочлена симметрической матрицы осуществляется весьма просто после предварительного ее преобразования; это преобразование подбирается так, чтобы полученная матрица имела те же собственные значения, что и исходная. Объясним, почему это возможно.

Определение 6.1. Матрица A называется подобной матрице B, если существует невырожденная матрица C, такая, что

$$B = C^{-1}AC.$$

Теорема 6.1. (о собственных значениях подобных матриц). *Собственные значения подобных матриц совпадают.*

Доказательство

Пусть матрицы A и B подобны, то есть существует матрица C, такая, что $B = C^{-1}AC$, и пусть λ_0 – собственное значение матрицы A. Докажем, что λ_0 – собственное значение матрицы B.

Действительно,

$$|B-\lambda_0 E| = |C^{-1}AC-\lambda_0 E| = |C^{-1}AC-\lambda_0 C^{-1}EC| =$$

$$= |C^{-1}||A-\lambda_0 E||C| = |C^{-1}||C||A-\lambda_0 E| = |A-\lambda_0 E| = 0,$$

Теорема доказана.

Собственные векторы подобных матриц A и B связаны соотношением: $x = Cy$, где $x -$ собственный вектор A, $y -$ собственный вектор B. Действительно, $By = \lambda y$; но $B = C^{-1}AC$, то есть $C^{-1}ACy = \lambda y$, или $ACy = \lambda Cy$, то есть $Cy = x -$ собственный вектор матрицы A.

Мы видим, что, зная собственные значения и собственные векторы матрицы, подобной исходной, можно легко найти собственные значения и собственные векторы исходной матрицы.

Ниже мы опишем, как симметричную матрицу A можно привести к подобной ей трёхдиагональной матрице. Для трёхдиагональной же матрицы выписать характеристический многочлен и вычислить собственные векторы уже легко.

Итак, займемся приведением симметрической матрицы к трёхдиагональной форме. Введем понятие *матрицы вращения.*

Определение 6.2. Матрица $C_{ij}\,(i<j)$ *вида*

$$\begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ . & . & . & . & . & . & . & . \\ 0 & 0 & \dots & c & \dots & -s & \dots & 0 \\ . & . & . & . & . & . & . & . \\ 0 & 0 & \dots & s & \dots & c & \dots & 0 \\ . & . & . & . & . & . & . & . \\ 0 & 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix}, \qquad (6.1)$$

в которой отличны от единицы i*-й и* j*-й столбцы и* i*-я и* j*-я строки и* $c^2 + s^2 = 1$, *называется матрицей вращения.*

Название связано с тем, что в R^2 матрица

$$\begin{pmatrix} c & -s \\ s & c \end{pmatrix}$$

является матрицей преобразования декартовых координат при повороте осей на угол $\varphi: \cos\varphi = c, \quad \sin\varphi = s$.

Обратная матрица к матрице вращений существует и совпадает с C_{ij}^T (это проверяется умножением C_{ij} на C_{ij}^T). Следовательно, матрица

$$C_{ij}^{-1} A C_{ij}$$

подобна матрице A.

Покажем, что любую симметричную матрицу последовательными умножениями на матрицы вращения можно привести к подобной ей трёхдиагональной.

Посмотрим вначале, что получится, если умножить симметричную матрицу A справа на C_{ij} и слева на C_{ij}^{-1}. Все столбцы матрицы A, кроме $i-$го и $j-$го, останутся без изменения. Обозначим $i-$ый и $j-$ый столбцы матрицы C_{ij} через B_i и B_j соответственно; тогда

$$B_i = cA_i + sA_j \ , \tag{6.2}$$

$$B_j = -sA_i + cA_j \, . \tag{6.3}$$

(A_i и $A_j - i$-ый и j-ый столбцы матрицы A). При умножении матрицы AC_{ij} на C_{ij}^{-1} (слева) будут изменяться только i-ая и j-ая строки, которые вычисляются по формулам аналогичным (6.2) и (6.3).

Но в силу симметричности матрицы A и структуры C_{ij} все элементы этих строк, кроме элементов, стоящих на пересечении с i-ым и j-ым столбцами, равны уже вычисленным элементам i-го и j-го столбцов.

Теперь подберем матрицы вращения так, чтобы при умножении A на эти матрицы получить трехдиагональную матрицу.

Вначале получим нули на месте элементов a_{13} и a_{31}. Для этого умножим матрицу A на матрицу C_{23} справа и C_{23}^{-1} слева (какую именно, мы скажем ниже):

$$AC_{23} = \begin{pmatrix} a_{11} & a_{12}c + a_{13}s & -a_{12}s + a_{13}c & a_{14} & \dots & a_{1n} \\ a_{21} & a_{22}c + a_{23}s & -a_{22}s + a_{23}c & a_{24} & \dots & a_{2n} \\ . & . & . & . & \dots & . \\ . & . & . & . & \dots & . \\ . & . & . & . & \dots & . \\ a_{n1} & a_{n2}c + a_{n3}s & -a_{n2}s + a_{n3}c & a_{n4} & \dots & a_{nn} \end{pmatrix},$$

$$C_{23}^{-1}AC_{23}=\begin{pmatrix} a_{11} & a_{12}c+a_{13}c & -a_{12}+a_{13}c & \dots & a_{1n} \\ a_{12}c+a_{13}s & a'_{22} & a'_{23} & \dots & a_{2n} \\ -a_{12}s+a_{13}c & a'_{32} & a'_{33} & \dots & a_{3n} \\ a_{41} & a'_{42} & a'_{43} & \dots & a_{4n} \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ a_{n1} & a'_{n2} & a'_{n3} & \dots & a_{nn} \end{pmatrix}$$

Мы хотим, чтобы $a'_{13}=0,$ поэтому элементы s и c матрицы C_{23} должны удовлетворять условиям:

$$\begin{cases} s^2+c^2=1, \\ -a_{12}s+a_{13}c=0. \end{cases}$$

Из этих условий $c=\dfrac{a_{12}}{\sqrt{a_{12}^2+a_{13}^2}}$, $s=\dfrac{a_{13}}{\sqrt{a_{12}^2+a_{13}^2}}$. Можно всегда вычислить арифметический корень $\sqrt{a_{12}^2+a_{13}^2}$.

Действительно, если $s=-\dfrac{a_{13}}{\sqrt{a_{12}^2+a_{13}^2}}$, то $c=-\dfrac{a_{12}}{\sqrt{a_{12}^2+a_{13}^2}}$, и оба слагаемые второго уравнения меняют знаки.

Обнулим теперь элементы $a_{i-1,\,j}\,(1<i<j)$. Для этого умножим A на матрицы C_{ij} справа и на $C_{ij}^{-1}=C_{ij}^{T}$ слева, такие, чтобы неизвестные элементы c и s матрицы C_{ij} удовлетворяли условиям:

$$\begin{cases} s^2+c^2=1, \\ -a_{i-1,\,i}s+a_{i-1,\,j}c=0, \end{cases}$$

то есть

$$c=\frac{a_{i-1,i}}{\sqrt{a_{i-1,i}^2+a_{i-1,\,j}^2}}, \tag{6.4}$$

$$s=\frac{a_{i-1,\,j}}{\sqrt{a_{i-1,i}^2+a_{i-1,\,j}^2}}. \tag{6.5}$$

Заметим, что описанные умножения (на матрицы вращения) нулевых элементов не меняют. Общее число преобразований для приведения симметричной матрицы к трёхдиагональной форме не превосходит числа

$$\frac{(n-1)(n-2)}{2}.$$

Контроль вычислений после каждого умножения производится путем вычислений *следа* матрицы, то есть суммы ее диагональных элементов. Выше мы доказали, что собственные значения подобных матриц совпадают; следовательно, должны быть равны и следы этих матриц.

6.2 Определение коэффициентов характеристического многочлена для трёхдиагональной матрицы

Коэффициенты характеристического многочлена для симметричной трёхдиагональной матрицы A находятся по рекуррентным формулам, связывающим характеристические многочлены $k-$ой и $k+1-$ой степеней главных миноров A_{kk} матрицы A.

Пусть $\varphi_k(\lambda)-$нормированный характеристический многочлен матрицы A (т.е. такой, что коэффициент при λ^k равен 1):

$$\varphi_k(\lambda)=(-1)^k|A-\lambda E|=$$

$$=(-1)^k\begin{vmatrix} a_{11}-\lambda & a_{12} & 0 & \dots & 0 & 0 \\ a_{12} & a_{22}-\lambda & a_{23} & \dots & 0 & 0 \\ 0 & a_{23} & a_{33}-\lambda & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ 0 & 0 & 0 & \dots & a_{k-1,k-1}-\lambda & a_{k-1,k} \\ 0 & 0 & 0 & \dots & a_{k-1,k} & a_{kk}-\lambda \end{vmatrix}.$$

Разложив определитель по элементам последнего столбца и отнормировав многочлены $\varphi_{k-1}(\lambda)$ и $\varphi_{k-2}(\lambda)$, получим:

$$\varphi_k(\lambda)=(-1)^k(a_{kk}-\lambda)(-1)^{k-1}\varphi_{k-1}(\lambda)-$$

$$-(-1)^k a_{k-1,k}\begin{vmatrix} a_{11}-\lambda & a_{12} & 0 & \dots & 0 & 0 \\ a_{12} & a_{22}-\lambda & a_{23} & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ 0 & 0 & 0 & \dots & a_{k-2,k-2}-\lambda & a_{k-2,k-1} \\ 0 & 0 & 0 & \dots & 0 & a_{k-1,k} \end{vmatrix} =$$

$$= -(a_{kk}-\lambda)\varphi_{k-1}(\lambda) - a_{k-1,k}^2\,\varphi_{k-2}(\lambda)\,. \qquad (6.6)$$

Формула (6.6) справедлива при $k \geq 3$. Если положить $\varphi_0 = 1$, то эта формула будет верна и для $k = 2$.

Действительно,

$$\varphi_1(\lambda) = \lambda - a_{11},\ \varphi_2(\lambda) = (-1)^2\begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{12} & a_{22}-\lambda \end{vmatrix} =$$

$$= (a_{11}-\lambda)(a_{22}-\lambda) - a_{12}^2 = (\lambda - a_{22})\varphi_1(\lambda) - a_{12}^2\varphi_0(\lambda).$$

Коэффициенты характеристического многочлена удобно вычислить по формуле (6.6), заполняя следующую таблицу:

Таблица 1 – Характеристический многочлен

Многочлен / Степень	φ_0	φ_1	φ_2	…	φ_{n-1}	φ_n
λ^n				…		1
λ^{n-1}				…	1	π_{n-1}^n
. .				…		
.			1	…	π_2^{n-1}	π_2^n
λ		1	π_1^2	…	π_1^{n-1}	π_1^n
λ_0	1	π_0^1	π_0^2	…	π_0^{n-1}	π_0^n

Здесь π_j^i – коэффициент при λ^j в многочлене φ_i, причем

$$\pi_0^0 = 1,\ \pi_j^i = -a_{i-1,i}^2 \pi_j^{i-2} - a_{ii}\pi_j^{i-1} + \pi_{j-1}^{i-1} (i \geq 2).$$

Вычисление коэффициентов характеристического многочлена для матрицы, приведённой к трёхдиагональной форме содержится в таблице 1.

6.3 Вычисление собственных векторов для симметричных матриц

Предположим, что симметричная матрица A приведена к подобной трёхдиагональной матрице B, пусть для B вычислены собственные значения λ_i. Тогда для вычисления соответствующих этим λ_i собственных векторов матрицы A нужно решить несколько систем однородных уравнений $(A - \lambda_i E)x = 0.$

Таблица 2 – Многочлен

Многочлен / Степень	φ_0	φ_1	φ_2	φ_3	φ_4
λ^4					1
λ^3				1	-42.3292
λ^2			1	-35.5918	434.1562
λ^1		1	-27.2941	195.3415	-1439.8524
λ^0	1	-3.31	50.8472	-150.5478	964.3942
$a_{i-1,i}^2$	1	28,5402	81,9840	0,9815	
a_{ii}	3.31	23.9841	8.2977	6.7374	

Поскольку при каждом $\lambda = \lambda_i$ каждая система имеет ненулевое решение, матрица $A - \lambda E$ $(\lambda = \lambda_i)$ вырождена. Поэтому вначале нужно определить, какое из уравнений системы является следствием остальных её уравнений. Используя же матрицы вращения C_{ij}, вычисленные при приведении A к трёхдиагональной форме, и зная собственные векторы матрицы B, можно получить собственные векторы матрицы A.

Действительно, если y собственный вектор B, где

$$B = \left(C_{n-1,n}^{-1} \cdot \ldots \cdot C_{23}^{-1}\right) A \left(C_{23} \cdot \ldots \cdot C_{n-1,n}\right), \tag{6.7}$$

то вектор

$$x = C_{23} C_{24} \cdot \ldots \cdot C_{n-1,n} y - \tag{6.8}$$

собственный вектор A.

После того, как вычислен вектор y, найти вектор x сравнительно легко, так как матрицы C_{ij} имеют очень простую структуру. При каждом умножении вектора y на матрицы C_{ij} будут меняться только две компоненты ранее полученного вектора: i-ая и j-ая (обозначим их через (y_i' и y_j'). Эти компоненты всякий раз определяются по формулам

$$\begin{aligned} y_i' &= c y_i - s y_j, \\ y_j' &= s y_i + c y_j \end{aligned} \tag{6.9}$$

(значения c и s соответствуют той матрице C_{ij}, на которую в данный момент мы умножаем).

Таким образом, для вычисления собственных векторов A нужно знать собственные векторы подобной ей трехдиагональной матрицы. Собственные векторы трехдиагональной матрицы B находятся в свою очередь из однородных систем с матрицей $B - \lambda_i E$ (λ_i – соответствующее искомому собственному вектору собственное значение):

$$\left(B - \lambda_i E\right) y = 0. \tag{6.10}$$

Положим $\lambda_i = \lambda$. Как и выше, матрица $B - \lambda E$ – вырожденная. При условии, что все $b_{i,i+1} \neq 0 \left(1 \le i \le n-1\right)$, последнее уравнение системы

$$\begin{aligned} &(b_{11} - \lambda) y_1 + b_{12} y_2 = 0, \\ &b_{21} y + (b_{22} - \lambda) y_2 + b_{23} y_3 = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &b_{n-1,n-2} y_{n-2} + (b_{n-1,n-1} - \lambda) y_{n-1} + b_{n-1,n} y_n = 0, \\ &b_{n,n-1} y_{n-1} + (b_{nn} - \lambda) y_n = 0 \end{aligned}$$

является следствием остальных уравнений. Действительно, при этом условии определитель треугольной матрицы, полученной из $B - \lambda E$ удалением 1-го столбца и n-ой строки, отличен от нуля, то есть строки 1-ая,…, $(n-1)$-ая линейно независимы.

Придавая компоненте y_1 вектора y произвольное, отличное от нуля значение, из образующейся системы с треугольной матрицей последовательно находятся компоненты $y_2, y_3, \ldots, y_n$. Последнее уравнение применяется для контроля вычислений.

Если некоторые $b_{i,i+1} = 0$, то матрицу системы (6.7) разбивают на блоки, после чего общее решение для каждого из блоков находят описанным выше способом.

6.4 Степенной метод вычисления наибольшего по модулю собственного значения матрицы

Рассмотрим метод решения частичных проблем собственных значений. Пусть известно, что у $A \in M_n(R)$ есть ровно n линейно независимых собственных векторов: $x_1 = \begin{bmatrix} x_{11} \\ x_{12} \\ \ldots \\ x_{1n} \end{bmatrix}, \ldots, x_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \ldots \\ x_{nn} \end{bmatrix}$. Пусть нумерация этих векторов отвечает упорядочению соответствующих им собственных значений по убыванию модулей (где первое из неравенств строгое):

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|. \qquad (6.11)$$

Ставим задачу приближённого вычисления наибольшего по модулю собственного числа λ_1 и соответствующего ему собственного вектора x_1 матрицы A.

Возьмем произвольный $y^{(0)} \neq \bar{0}$ и запишем его разложение по базису из собственных векторов $x_1, x_2, \ldots, x_n$:

$$y^{(0)} = c_1 x_1 + c_2 x_2 + \ldots + c_n x_n$$

и пусть $c_1 \neq 0$.

Далее

$$y^{(1)} = A y^{(0)} = c_1 A x_1 + c_2 A x_2 + \ldots + c_n A x_n.$$

Так как (λ_i, x_i) – собственные пары матрицы A, то имеем $y^{(1)} = c_1 \lambda_1 x_1 + c_2 \lambda_2 x_2 + \ldots + c_n \lambda_n x_n$, тогда

$$y^{(2)} = A y^{(1)} = A^2 y^{(0)} = c_1 \lambda_1^2 x_1 + c_2 \lambda_2^2 x_2 + \ldots + c_n \lambda_n^2 x_n.$$

Очевидно, k-ая итерация вектора $y^{(0)}$ с помощью матрицы A дает вектор

$$y^{(k)} = Ay^{(k-1)} = A^k y^{(0)} = c_1\lambda_1^k x_1 + c_2\lambda_2^k x_2 + \ldots + c_n\lambda_n^k x_n \Rightarrow$$

$$\Rightarrow y^{(k)} = A^k y^{(0)} = \begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \ldots \\ y_n^{(k)} \end{bmatrix} = c_1\lambda_1^k \begin{bmatrix} x_{11} \\ x_{12} \\ \ldots \\ x_{1n} \end{bmatrix} + \ldots + c_n\lambda_n^k \begin{bmatrix} x_{n1} \\ x_{n2} \\ \ldots \\ x_{nn} \end{bmatrix}. \qquad (6.12)$$

Рассмотрим отношение

$$\frac{y_i^{(k)}}{y_i^{(k-1)}} = \frac{c_1\lambda_1^k x_{1i} + c_2\lambda_2^k x_{2i} + \ldots + c_n\lambda_n^k x_{ni}}{c_1\lambda_1^{k-1} x_{1i} + c_2\lambda_2^{k-1} x_{2i} + \ldots + c_n\lambda_n^{k-1} x_{ni}} = \left\{ \begin{matrix} :c_1\lambda_1^{k-1}x_{1i} \\ :c_1\lambda_1^{k-1}x_{1i} \end{matrix} \right\} =$$

$$= \lambda_1 \frac{1 + \frac{c_2}{c_1}\cdot\frac{x_{2i}}{x_{1i}}\left(\frac{\lambda_2}{\lambda_1}\right)^k + \ldots + \frac{c_n}{c_1}\cdot\frac{x_{ni}}{x_{1i}}\left(\frac{\lambda_n}{\lambda_1}\right)^k}{1 + \frac{c_2}{c_1}\cdot\frac{x_{2i}}{x_{1i}}\left(\frac{\lambda_2}{\lambda_1}\right)^{k-1} + \ldots + \frac{c_n}{c_1}\cdot\frac{x_{ni}}{x_{1i}}\left(\frac{\lambda_n}{\lambda_1}\right)^{k-1}}.$$

Очевидно, что $\lim\limits_{k\to\infty} \frac{y_i^{(k)}}{y_i^{(k-1)}} = \lambda_1$ для $\forall i = \overline{1,n},$ при котором $x_{1i} \neq 0,$ в силу (6.1). Представляя вектор $y^{(k)}$ на основе (2) в виде

$$y^{(k)} = c_1\lambda_1^k \left[x_1 + \frac{c_2}{c_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k x_2 + \ldots + \frac{c_n}{c_1}\left(\frac{\lambda_n}{\lambda_1}\right)^k x_n \right],$$

можно сделать вывод, что в силу

$$\left|\frac{\lambda_i}{\lambda_1}\right|^k \underset{\substack{k\to\infty \\ (i\neq 1)}}{\to} 0$$

в последнем выражении для $y^{(k)}$ начнёт доминировать первое слагаемое.

Это означает, что вектор $y^{(k)}$ от итерации к итерации будет давать все более хорошие приближения к собственному вектору x_1 с точностью до скалярного множителя $c_1\lambda_1^k$.

Таким образом, приведенный метод нахождения «старшей» собственной пары матрицы называется *степенным методом*.

Как только установятся несколько первых цифр во всех этих отношениях (что выясняется проверкой выполнения приближённых равенств:

$$\frac{y_i^{(k)}}{y_i^{(k-1)}} \approx \frac{y_i^{(k-1)}}{y_i^{(k-2)}}),$$

так можно считать, что найдено наибольшее по модулю собственное число с точностью, определяемой последним установившимся в отношениях знаком после запятой, и соответствующий ему собственный вектор, за который принимается последний полученный вектор $y^{(k)}$.

Замечание 6.1. *Поскольку в процессе вычислений за счет множителя* λ_1^k *при* $k \to \infty$ *может произойти либо превышение допустимых для используемого ПЭВМ чисел, если* $|\lambda_1| > 1$, *либо пропадание значащих цифр полученных векторов* $y^{(k)}$, *если* $|\lambda_1| < 1$, *то целесообразно вводить в итерационный процесс нормирование найденных векторов* $y^{(k)}$.

Модификация: находим

$$A, A^2, ..., A^{2k}, ...; \quad y^{(2k)} = A^{2k} y^{(0)}; \quad \lambda_1 \approx \frac{trA^{2k+1}}{trA^{2k}}.$$

6.5 Метод обратных итераций

Известно, что если матрица A имеет собственные пары $(\lambda_1, x_1), ..., (\lambda_n, x_n)$, то собственными парами матрицы A^{-1} будут $\left(\frac{1}{\lambda_1}, x_1\right), ..., \left(\frac{1}{\lambda_n}, x_n\right)$. При этом упорядочиванию спектра матрицы A

$$|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_{n-1}| > |\lambda_n|$$

соответствует цепочка неравенств

$$\left|\frac{1}{\lambda_n}\right| > \left|\frac{1}{\lambda_{n-1}}\right| \geq ... \geq \left|\frac{1}{\lambda_2}\right| \geq \left|\frac{1}{\lambda_1}\right|$$

для собственных чисел: $\gamma_1 = \frac{1}{\lambda_n}, \gamma_2 = \frac{1}{\lambda_{n-1}}, ..., \gamma_n = \frac{1}{\lambda_1}$ матрицы A^{-1}.

Следовательно, наименьшим по модулю собственным числом матрицы A является величина, обратная наибольшему по модулю собственному числу матрицы A^{-1}. Последнее может быть получено прямыми итерациями степенного метода произвольного вектора $y^{(0)}$ посредством матрицы A^{-1} по формуле:

$$y^{(k)} = A^{-1}y^{(k-1)}, k = 1,2,.... \quad (6.13)$$

При $k \to \infty$ последовательность отношений $\frac{y^{(k)}}{y^{(k-1)}}$ должна давать приближенное значение $\frac{1}{\lambda_n}$, а вектор $y^{(k)}$ (желательно нормированный) можно принять за собственный вектор x_n.

Вместо прямых итераций (1), требующих предварительного обращения матрицы A, обычно строят ту же последовательность векторов $\{y^{(k)}\}$, решая при $k = 1,2,...$ линейные системы:

$$Ay^{(k)} = y^{(k-1)}. \quad (6.14)$$

Так как все эти системы имеют одну и туже матрицу коэффициентов, то самая трудоемкая часть метода Гаусса для их решения – LU-факторизация матрицы A – может быть выполнена один раз.

Построение последовательности векторов, приближающих собственный вектор x_n матрицы A, соответствующий наименьшему собственному числу λ_n называют *методом обратных итераций*.

6.6 Метод λ-разности

Этот метод позволяет находить собственное значение λ_2 после вычисления λ_1 при условии, что $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq ... \geq |\lambda_n|$.

Рассмотрим некоторый $y^{(0)}$ и предположим, что

$$y^{(0)} = c_1x_1 + c_2x_2 + ... + c_nx_n,$$

где $x_i (i = \overline{1,n})$ – собственные векторы матрицы A, и пусть $c_1 \neq 0, c_2 \neq 0$.

Далее вычислим последовательность:
$y^{(0)}, y^{(1)}, ..., y^{(k)}, ...$ по формуле степенного метода: $y^{(k)} = A^k y^{(0)}$, и найдем в итоге λ_1.

Введем обозначение

$$\Delta_\lambda y^{(k)} = y^{(k+1)} - \lambda y^{(k)} (k = 0,1,2,\ldots). \qquad (6.15)$$

Величину $\Delta_\lambda y^{(k)}$ будем называть λ-*разностью* от $y^{(k)}$. Тогда для некоторой компоненты *s* получим:

$$\Delta_{\lambda_1} y_s^{(k)} = y_s^{(k+1)} - \lambda_1 y_s^k = \left(c_1\lambda_1^{k+1}x_{1s} + c_2\lambda_2^{k+1}x_{2s} + \ldots + c_n\lambda_n^{k+1}x_{ns}\right) -$$

$$- \lambda_1\left(c_1\lambda_1^k x_{1s} + c_2\lambda_2^k x_{2s} + \ldots + c_n\lambda_n^k x_{ns}\right) =$$

$$= c_2(\lambda_2 - \lambda_1)\lambda_2^k x_{2s} + \ldots + c_n(\lambda_n - \lambda_1)\lambda_n^k x_{ns},$$

и, аналогично,

$$\Delta_{\lambda_1} y_s^{(k-1)} = y_s^{(k)} - \lambda_1 y_s^{(k-1)} = c_2(\lambda_2 - \lambda_1)\lambda_2^{k-1}x_{2s} + \ldots + c_n(\lambda_n - \lambda_1)\lambda_n^{k-1}x_{ns}.$$

Если $k \to \infty$, то в выражениях $\Delta_{\lambda_1} y_s^{(k)}$ и $\Delta_{\lambda_1} y_s^{(k-1)}$ преобладающими будут члены, содержащие λ_2^k. Значит,

$$\lambda_2 \approx \frac{\Delta_{\lambda_1} y_s^{(k)}}{\Delta_{\lambda_1} y_s^{(k-1)}}.$$

Заметим, что если

$$\lambda_1 \approx \frac{y_s^{(k+1)}}{y_s^{(k)}},$$

то λ_2 целесообразно искать:

$$\lambda_2 \approx \frac{\Delta_{\lambda_1} y_s^{(m)}}{\Delta_{\lambda_1} y_s^{(m-1)}}, \text{ где } m < k.$$

В качестве собственного вектора матрицы *A*, отвечающего λ_2, приближённо можно взять вектор $\Delta_{\lambda_1} y^{(k)}$.

Теоретически возможно метод λ-разности применять и к вычислению следующих собственных значений, однако результаты будут еще менее надёжными, чем в случае λ_2.

6.7 Ускорение сходимости степенного метода. δ^2-процесс Эйткена

При решении полной и частичной проблем собственных значений иногда возникает необходимость уточнить полученные результаты. Необходимость уточнения может быть обусловлена тем, что из-за неустойчивости применяемого метода к ошибкам округления полученный характеристический полином будет иметь неудовлетворительную точность, и, следовательно, нельзя будет хорошо вычислить и собственные значения матрицы.

Рассмотрим приём ускорения сходимости последовательностей, получающихся при использовании степенного метода – *δ^2-процесса Эйткена.*

Пусть дана последовательность чисел или функций $u_0, u_1, \ldots, u_n, \ldots$. Требуется преобразовать эту последовательность в новую последовательность $\{v_n\}$, которая сходилась бы к тому же самому пределу, что и $\{u_n\}$, но быстрее последней. Каждый член последовательности $\{v_n\}$ будем определять по *формуле Эйткена*:

$$v_n = \frac{u_{n+1} \cdot u_{n-1} - u_n^2}{u_{n+1} - 2u_n + u_{n-1}} \ (n = 1, 2, \ldots), \qquad (6.16)$$

где $u_{n+1} - 2u_n + u_{n-1} \neq 0$.

Если собственные значения матрицы A удовлетворяют условию: $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n|$, то преобразование Эйткена можно использовать для ускорения последовательности $\{\lambda_1^{(k)}\}$, возникающей в степенном методе:

$$\lambda_1^{(k)} = \frac{y_s^{(k+1)}}{y_s^{(k)}}.$$

Поскольку $\lambda_1^{(k)}$ изменяется по закону, близкому к геометрической прогрессии, то последовательность $\{v_k\}$, определяющаяся по формуле Эйткена (6.16)

$$v_k = \frac{\lambda_1^{(k+1)}\lambda_1^{(k-1)} - \left[\lambda_1^{(k)}\right]^2}{\lambda_1^{(k+1)} - 2\lambda_1^{(k)} + \lambda_1^{(k-1)}} \ (k = 1, 2, 3, \ldots)$$

и будет иметь более быструю сходимость к искомому пределу λ_1 (наибольшее собственное число матрицы A).

Будем считать, что λ_1 найдено достаточно точно. В этом случае δ^2-процесс Эйткена можно применить также к определению уточнённого собственного вектора, отвечающего λ_1. По формуле степенного метода: $y^{(k)} = A^k y^{(0)}$ и для s-ой компоненты $y^{(k)}$ справедливо записать

$$y_s^{(k)} = \beta_{s1}\lambda_1^k + \beta_{s2}\lambda_2^k + ... + \beta_{sn}\lambda_n^k, \quad \beta_{si} = c_i x_{is}, \quad \beta_{s1} \neq 1.$$

Составим величины

$$\lambda_1 y_s^{(k-1)}, 1 \cdot y_s^{(k)}, \frac{1}{\lambda_1} y_s^{(k+1)}$$

и применим к ним формулу Эйткена

$$v_s^{(k)} = \frac{\frac{1}{\lambda_1} y_s^{(k+1)} \lambda_1 y_s^{(k-1)} - \left[y_s^{(k)}\right]^2}{\frac{1}{\lambda_1} y_s^{(k+1)} - 2y_s^{(k)} + \lambda_1 y_s^{(k-1)}} = \beta_{s1}\lambda_1^k \left[1 + O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^k\right)\right].$$

Также можно записать

$$y_s^{(k)} = \beta_{s1}\lambda_1^k \left[1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right].$$

Поскольку $\left|\frac{\lambda_3}{\lambda_1}\right| < \left|\frac{\lambda_2}{\lambda_1}\right|$, то последовательность $\left\{v_s^{(k)}\right\}$ будет быстрее, чем последовательность $\left\{y_s^{(k)}\right\}$, сходится к искомому пределу – s-й компоненте собственного вектора, соответствующего λ_1. Причём эта сходимость будет тем быстрее, чем $|\lambda_3|$ меньше $|\lambda_2|$.

6.8 Метод Якоби, *LR* и *QR*-алгоритмы определения собственных векторов и собственных значений матрицы

В этом разделе мы вкратце опишем некоторые итерационные методы определения собственных пар матрицы A.

Как правило, итерационные методы позволяют определить собственные значения без предварительного вычисления коэффициентов характеристического многочлена исходной матрицы. В основе этих методов лежит построение последовательностей таких подобных матрице A матриц, для которых собственные значения определяются уже легко.

При этом собственные значения и собственные векторы исходной матрицы A являются пределами собственных значений и собственных векторов построенных последовательностей матриц.

1. *Метод вращений* (*метод Якоби*) предназначен для определения собственных значений симметричной матрицы. Известно, что всякая симметричная матрица A может быть записана в виде $A = TLT^*$, где T – унитарная матрица, L – диагональная с диагональными элементами $\lambda_1,...,\lambda_n$, равными собственным значениям матрицы A. Из представления $A = TLT^*$ получаем, что $AT = TL$ (напомним, что $T^* = T^{-1}$).

Если расписать это равенство по столбцам, то окажется что каждый i – ый столбец матрицы T является собственным вектором, соответствующим собственному значению λ_i. Преобразуя равенство $A = TLT^*$, получим $T^*AT = L$. В методе вращений матрица T строится как предел бесконечного произведения элементарных матриц вращения.

При этом получается последовательность матриц A^k:

$$A^1 = T_1^* A T_1,...,A^k = T_k^* A^{k-1} T_k,...,$$

где $T_1,...,T_k$ – элементарные матрицы вращения.

На k – ом шаге матрица вращения T_k строится в зависимости от матрицы A^{k-1} так, чтобы сумма квадратов внедиагональных элементов была как можно меньше.

Это возможно благодаря справедливости следующей теоремы:

***Теорема* 6.2.** *Всякую квадратную матрицу A можно привести к подобной ей почти диагональной матрице*

$$B = \begin{pmatrix} \lambda_1 & b_{12} & ... & b_{1n} \\ b_{21} & \lambda_2 & ... & b_{2n} \\ . & . & . & . \\ b_{n1} & b_{n2} & ... & \lambda_n \end{pmatrix},$$

где $B = C^{-1}AC, \lambda_1, \lambda_2, ..., \lambda_n$ – *собственные значения матрицы A, а* $|b_{ij}| < \varepsilon$ *(при* $i \neq j$*),* ε – *сколь угодно малое положительное число.*

Пусть S^k – матрица, получающаяся из A^k заменой всех ее внедиагональных элементов на нулевые. По построению матрицы A^k при больших k матрицы A^k и S^k близки (это можно доказать строго); поэтому близки и коэффициенты их характеристических многочленов.

В силу непрерывной зависимости нулей характеристического многочлена от его коэффициентов собственные значения матриц A^k и S^k также близки между собой.

Собственные же значения матрицы S^k – это диагональные элементы матрицы A^k, а собственные значения подобных матриц A^k и A совпадают. Таким образом, диагональные элементы матрицы A^k можно считать приближениями к собственным значениям матрицы A. Собственные же векторы матрицы A приближённо совпадают со столбцами $T^k = T_1 \cdot \ldots \cdot T_k$.

2. Метод Якоби укладывается в схему так называемых *степенных методов.* Другой тип схемы степенных методов имеет вид:

$$A = L_1 R_1,$$
$$R_1 L_1 = L_2 R_2,$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$R_{k-1} L_{k-1} = L_k R_k.$$

Известно несколько схем такого типа. Наиболее распространенные среди них – это записанный выше $LR-$ *алгоритм* и $QR-$ *алгоритм* (см. ниже). Несколько слов об этих двух алгоритмах (в предположении, что все главные миноры исходной матрицы A отличны от нуля и что все ее собственные значения различны):

а) В $LR-$ *алгоритме* матрицы L_k и R_k – треугольные: R_k – верхняя треугольная, L_k – нижняя треугольная с единичными элементами по диагонали.

Легко видеть, что $R_k L_k = (L_1 \cdot \ldots \cdot L_k)^{-1} A (L_1 \cdot \ldots \cdot L_k)$, так что матрицы $R_k L_k$ подобны матрице A.

Можно доказать, что при больших k матрица L_k близка к единичной матрице, а матрица R_k – к диагональной, у которой по диагонали стоят собственные значения $\lambda_1, \ldots, \lambda_n$ матрицы A; т. е., что при $k \to \infty$ последовательность матриц $\{R_k L_k\}$ сходится к верхней треугольной матрице, подобной A.

Так что в этом случае определение собственных значений матрицы A достаточно просто: они близки к диагональным элементам матрицы R_k при больших k. Собственные векторы матрицы A после определения собственных векторов матрицы $R_k L_k$ (k достаточно велико) определяются по формуле: $x = y \cdot L_1 \cdot \ldots \cdot L_k$ (здесь y – собственный вектор матрицы $R_k L_k$).

Учитывая же, что $L_k \to E$ при $k \to \infty$, получаем что собственные векторы исходной матрицы совпадают с собственными векторами матриц R_k при больших k (разумеется, мы всюду пользуемся тем фактом, что корни характеристического многочлена непрерывно зависят от его коэффициентов).

б) В $QR-$*алгоритме* разложение матрицы осуществляется по схеме:

$$A = Q_1 R_1,$$
$$R_1 Q_1 = Q_2 R_2.$$
$$\ldots\ldots\ldots\ldots$$
$$R_{k-1} Q_{k-1} = Q_k R_k,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$

Здесь Q_k – унитарные матрицы, R_k – верхние треугольные матрицы.

В этом алгоритме последовательность матриц $R_k Q_k$ при $k \to \infty$ сходится к верхней треугольной матрице, подобной A (мы предположили, что у A нет кратных собственных значений). После определения собственных векторов y предельной матрицы собственные векторы исходной находятся по формуле:

$$x = y \lim_{k \to \infty} Q_1 \cdot \ldots \cdot Q_k,$$

где $\lim_{k \to \infty} Q_1 \cdot \ldots \cdot Q_k$ – матрица, к которой сходится при $k \to \infty$ произведение унитарных матриц $Q_1 \cdot \ldots Q_k$.

ГЛАВА 7 ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЛИНЕЙНЫХ СИСТЕМ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

7.1 Метод исключения Гаусса. Схема с выбором главного элемента

Методы решения систем линейных алгебраических уравнений условно можно разделить на *точные* и *итерационные.*

Точные методы дают решение задачи за конечное число арифметических действий. Решение получается точным только в том случае, если исходные данные заданы точно и промежуточные вычисления выполняются без округлений.

Итерационные методы дают бесконечную последовательность приближенных решений, предел которой (если он существует) является решением системы.

В этой главе мы рассмотрим следующие точные методы решения линейных алгебраических систем: *метод исключения Гаусса, метод оптимального исключения, метод квадратного корня* (для систем с симметричной матрицей).

Выбор конкретного метода решения определяется спецификой рассматриваемой задачи и возможностями ЭВМ, находящейся в распоряжении вычислителя. Приведем некоторые суждения, влияющие на выбор метода.

1. Если матрица рассматриваемой системы алгебраических уравнений содержит большое число нулевых элементов, то необходимо выбрать такой метод решения, которых сохраняет при преобразованиях особенности структуры матрицы. Реализация многих методов существенно упрощается, если A, например, является симметричной или трехдиагональной.
2. На выбор метода существенно влияет требуемая точность решения системы. Иногда нужно определить немного верных цифр в результате (с малой точностью), но с минимальными затратами на сам счет. В других случаях, напротив, необходимо иметь достаточно точный результат, пусть и за счет увеличения времени счета.
3. При решении системы

$$Ax = b$$

необходимо знать, как влияет на результат изменение элементов матрицы A и вектора b. Слабая чувствительность к вариациям элементов матрицы и вектора правой части является важной характеристикой метода.

Идея метода исключения Гаусса заключается в том, чтобы исходную систему линейных уравнений

$$Ax = b \tag{7.1}$$

с произвольной матрицей A свести некоторыми *эквивалентными* преобразованиями (не меняющими множества решений системы (24.1)) к системе вида

$$\overline{A}x = \overline{b}\,, \tag{7.2}$$

где $\overline{A}$ – треугольная (например, верхняя треугольная) матрица. Ясно, что алгоритм решения систем вида (7.2) уже прост: из последнего уравнения легко находится x_n, затем из предпоследнего – x_{n-1} и т.д.

Рассмотрим одну из возможных реализаций метода Гаусса более подробно. Распишем систему (7.1) в виде:

$$\begin{aligned} a_{11}x_1 &+ a_{12}x_2 + \ldots + a_{1n}x_n = b_1,\\ a_{21}x_1 &+ a_{22}x_2 + \ldots + a_{2n}x_n = b_2,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ a_{n1}x_1 &+ a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{aligned} \tag{7.3}$$

(мы считаем, что матрица $A = (a_{ij})$ невырождена).

В предположении, что $a_{11} \neq 0$, делим первое уравнение системы (7.3) на коэффициент a_{11}, в результате чего получаем уравнение

$$x_1 + \frac{a_{12}}{a_{11}}x_2 + \ldots + \frac{a_{1n}}{a_{11}}x_n = \frac{b_1}{a_{11}}.$$

Затем из каждого из остальных уравнений вычитается первое, умноженное на соответствующий коэффициент a_{i1}. В результате первое неизвестное оказывается исключенным из всех уравнений, кроме первого, и мы приходим к системе, эквивалентной (7.3):

$$\begin{aligned} x_1 &+ a^1_{12}x_2 + \ldots + a^1_{1n}x_n = b^1_1,\\ &\quad a^1_{22}x_2 + \ldots + a^1_{2n}x_n = b^1_2,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &\quad a^1_{n2}x_2 + \ldots + a^1_{nn}x_n = b^1_n \end{aligned} \tag{7.4}$$

(здесь $a_{1j}^1 = \dfrac{a_{1j}}{a_{11}}$, $j = 2, \ldots, n$, $b_1^1 = \dfrac{b_1}{a_{11}}$, $a_{ij}^1 = a_{ij} - a_{i1}a_1^1$, $b_i^1 = b_i - b_1^1 a_{i1}$, $i = 2, 3, \ldots, n$, $j = 1, 2, \ldots, n$).

Далее предполагается, что $a_{22}^1 \neq 0$, делим второе уравнение системы (7.4) на коэффициент a_{22}^1 и исключаем неизвестное x_2 из всех уравнений, начиная с третьего. Получим систему, эквивалентную исходной

$$
\begin{array}{ccccccccccc}
x_1 & + & a_{12}^1 x_1 & + & a_{13}^1 x_3 & + & \ldots & + & a_{1n}^1 x_n & = & b_1^1, \\
 & & x_2 & + & a_{23}^2 x_3 & + & \ldots & + & a_{2n}^2 x_n & = & b_2^2, \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
 & & & & a_{n3}^2 x_3 & + & \ldots & + & a_{nn}^2 x_n & = & b_n^2,
\end{array}
$$

где $a_{2j}^2 = \dfrac{a_{2j}^1}{a_{22}^1}$, $a_{ij}^2 = a_{ij}^1 - a_{2j}^2 a_{i2}^1$, $b_2^2 = \dfrac{b_2^1}{a_{22}^1}$, $b_i^2 = b_i^1 - b_2^2 a_{i2}^1$, $i = 3, 4, \ldots, n$, $j = 2, 3, \ldots, n$.

Продолжая описанный процесс исключения неизвестных (т.е. исключая $x_3, x_4, \ldots, x_n$), мы вместо системы (7.3) получим эквивалентную систему:

$$
\begin{array}{ccccccccccc}
x_1 & + & a_{12}^1 x_2 & + & a_{13}^1 x_3 & + & \ldots & + & a_{1n}^1 x_n & = & b_1^1, \\
 & & x_2 & + & a_{23}^2 x_3 & + & \ldots & + & a_{2n}^2 x_n & = & b_2^2, \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
 & & & & & & & & x_n & = & b_n^n.
\end{array}
\tag{7.5}
$$

Выпишем формулы для коэффициентов преобразованной системы на k-ом шаге (когда x_k исключается из всех уравнений, начиная с ($k+1$)-го):

$$
a_{kj}^k = \frac{a_{kj}^{k-1}}{a_{kk}^{k-1}}, \quad a_{ij}^k = a_{ij}^{k-1} - a_{kj}^k a_{ik}^{k-1},
$$

$$
b_k^k = \frac{b_k^{k-1}}{a_{kk}^{k-1}}, \quad b_i^k = b_i^{k-1} - b_k^k a_{ik}^{k-1}, \tag{7.6}
$$

$i = k+1, \; k+2, \; \ldots, n$; $\quad j = k, \; k+1, \ldots, n$.

Процесс сведения системы (7.3) к системе с треугольной матрицей называется *прямым ходом метода Гаусса*. Выполнение всех указанных преобразований возможно, если получающиеся при расчетах коэффициенты a_{22}^{1}, …, a_{nn}^{n-1} отличны от нуля. В противном случае, например, если $a_{22}^{1} = 0$, нужно произвести перестановку уравнений и перенумерацию неизвестных: среди коэффициентов a_{32}^{1}, … , a_{n2}^{1} обязательно найдется хотя бы один, отличный от нуля, иначе матрица A была бы вырожденной.

Получив систему (7.5), последовательно находим ее решение:

$$\begin{aligned} x_n &= b_n^n, \\ x_{n-1} &= b_{n-1}^{n-1} - a_{n-1,n}^{n-1} x_n, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_k &= b_k^k - \sum_{i=k+1}^{n} a_{ki}^k x_i \quad . \end{aligned} \tag{7.7}$$

Нахождение решения системы (7.5) обычно называют *обратным ходом метода Гаусса.*

Контроль вычислений в методе Гаусса. При вычислениях, производимых по методу Гаусса, без помощи ЭВМ велика вероятность случайных ошибок. В таких случаях необходимо проводить текущий контроль правильности вычислений. Для проверки прямого хода метода Гаусса вводят *контрольный столбец системы*

$$K\sum = \begin{pmatrix} K\sum_1 \\ \vdots \\ K\sum_n \end{pmatrix},$$

состоящий из контрольных элементов уравнений системы:

$$K\sum_i = \sum_{j=1}^{n} a_{ij} + b_i \, .$$

При преобразовании уравнений системы над контрольными элементами производятся те же операции, что и над свободными членами уравнений. В результате этого контрольный элемент каждого нового уравнения должен равняться сумме коэффициентов этого уравнения; большое расхождение между ними указывает на ошибки в вычислениях или на неустойчивость алгоритма вычислений по отношению к вычислительной погрешности.

Однако совпадение этих сумм еще не дает гарантии правильности выполнения всех операций. Конечно, маловероятно, чтобы при «взаимном гашении» нескольких ошибок получался верный результат, но окончательный контроль все же необходим. Это можно сделать следующим образом. Рассмотрим вместе с системой $Ax = b$ систему $A\bar{x} = K\sum$. Легко заметить, что решения $x_1, \ldots, x_n$ и $\bar{x}_1, \ldots, \bar{x}_n$ этих двух систем связаны соотношением

$$\bar{x}_i = x_i + 1.$$

Поскольку со столбцом $K\sum$ свободных членов системы $A\bar{x} = K\sum$ мы производим (при контроле) все те же действия, что и со столбцом b свободных членов исходной системы (при приведении ее матрицы коэффициентов к треугольному виду), у нас в конце прямого хода метода Гаусса система $A\bar{x} = K\sum$ также имеет треугольный вид, и мы легко можем найти значения неизвестных $\bar{x}_1, \ldots, \bar{x}_n$. Если окажется, что $\bar{x}_i - x_i = 1$ для всех $i = 1, \ldots, n$, то решение $x_1, \ldots, x_n$ найдено правильно.

Задачи и примеры предыдущего параграфа решаются сравнительно просто. В случае, когда коэффициентами системы уравнений являются дробные числа или числа, достаточно большие по абсолютной величине, вычисления усложняются. Чтобы упростить их и иметь возможность контролировать производимые вычисления, применяют несколько видоизмененный метод Гаусса.

Пусть дана система линейных алгебраических уравнений

$$\left.\begin{array}{rcrcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & b_1; \\ a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & b_2; \\ \ldots\ldots & & \ldots\ldots & & \ldots\ldots & & \ldots\ldots & & \ldots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \ldots & + & a_{nn}x_n & = & b_n, \end{array}\right\} \qquad (7.8)$$

которую в матричном виде можно записать так:

$$Ax = b, \qquad (7.9)$$

где

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_n \end{bmatrix}. \qquad (7.10)$$

Метод Гаусса с выбором главного элемента состоит в следующем. В системе (7.1) выбирают сначала уравнение, в котором содержится наибольший по абсолютной величине коэффициент системы (главный элемент), и делят данное уравнение на этот коэффициент.

После этого так же, как и в простейшей схеме метода Гаусса, исключают из остальных уравнений то неизвестное, при котором был наибольший коэффициент в выбранном уравнении (для удобства главный элемент можно поместить в первую строку и первый столбец матрицы, над которой производятся соответствующие преобразования).

Далее, оставляя неизменным уравнение с главным элементом, ищут наибольший по абсолютной величине коэффициент в остальных уравнениях (новый главный элемент), делят на него уравнение, в котором он находится, и исключают из остальных уравнений соответствующее неизвестное и т.д., пока не останется одно уравнение с одним неизвестным, т.е. пока система (7.8) не будет приведена к диагональному виду.

Чтобы не сделать ошибок, применяют контрольные вычисления. Для этого поступают следующим образом. Делают замену

$$y = x + e, \tag{7.11}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad e = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \tag{7.12}$$

то есть

$$y_i = x_i + 1 \ (i = 1, 2, \ldots, n), \tag{7.13}$$

в результате чего приходят к новой системе

$$Ay = Ax + Ae = b + Ae = \sigma; \quad Ay = \sigma, \tag{7.14}$$

где

$$\sigma = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_n \end{bmatrix}, \ \sigma_i = b_i + \sum_{k=1}^{n} a_{ik} \ . \tag{7.15}$$

Одновременно решают обе системы (7.1), (7.7), приводят их к диагональному виду, все вычисления сводят в таблицу, контролируют их с помощью чисел контрольного столбца.

7.2 Метод Гаусса вычисления определителя матрицы и обратной матрицы

Идея способа Гаусса последовательного исключения неизвестных в системе уравнений может быть перенесена на задачу *вычисления определителей*, и здесь она переходит в способ последовательного понижения порядка n определителя. Рассмотрим схему единственного деления.

Пусть дан определитель

$$D = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & \dots & . \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

Выберем как-либо ведущий элемент первого шага преобразований. Он должен быть отличным от нуля; чтобы избежать сильного разброса в порядках чисел, за него принимают либо наибольший по модулю элемент D, либо наибольший элемент в избранной строке или избранном столбце. Выполняя, если нужно, перестановку строк и столбцов, можно считать, что за ведущий элемент принят a_{11}.

Вынося a_{11} из первой строки (первого столбца) за знак D, приведем определитель к виду

$$D = a_{11} \begin{vmatrix} 1 & b_{12} & \dots & b_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & \dots & . \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

Умножая первую строку последовательно на a_{21}, a_{31},…,a_{n1} и вычитая из второй, третьей и т.д. строк, получим

$$D = a_{11} \begin{vmatrix} 1 & b_{12} & \dots & b_{1n} \\ 0 & a_{22.1} & \dots & a_{2n.1} \\ . & . & \dots & . \\ 0 & a_{n2.1} & \dots & a_{nn.1} \end{vmatrix}. =$$

$$= a_{11}\begin{vmatrix} a_{22\cdot1} & a_{23\cdot1} \dots a_{2n\cdot1} \\ a_{32\cdot1} & a_{33\cdot1} \dots a_{3n\cdot1} \\ \cdot & \cdot \quad \dots \quad \cdot \\ a_{n2\cdot1} & a_{n3\cdot1} \dots a_{nn\cdot1} \end{vmatrix}. \qquad (7.16)$$

Этим мы понизим порядок определителя на единицу и можем перейти ко второму шагу преобразований, применяя к полученному порядку $n-1$ такие же преобразования. Выполняя все n шагов, найдем определитель D как произведение ведущих элементов:

$$D = a_{11} \cdot a_{22.1} \cdot a_{33.2} \dots a_{nn.n-1}. \qquad (7.17)$$

Можно было бы применить к вычислению определителя идеи метода оптимального исключения неизвестных, но, как легко видеть, в этом случае проделана излишняя, по сравнению с изложенным методом вычислительная работа. Это связано с тем, что в описанных выше преобразованиях таблица элементов, если не понижать порядка определителей, будет приведена к правой треугольной; определитель по этой причине будет вычислен просто, так как он равен произведению диагональных элементов.

При применении же метода оптимального исключения таблица будет приведена не к треугольной, а к диагональной, что при вычислении определителя является излишним упрощением.

Пусть дана невырожденная матрица

$$A = \lfloor a_{ij} \rfloor \ (i, j = 1, 2, \dots, n). \qquad (7.18)$$

Для нахождения ее обратной матрицы

$$A^{-1} = \lfloor x_{ij} \rfloor \qquad (7.19)$$

используем основное соотношение

$$AA^{-1} = E, \qquad (7.20)$$

где E – единичная матрица.

Перемножая матрицы A и A^{-1}, будем иметь n систем уравнений относительно n^2 неизвестных x_{ij}

$$\sum_{k=1}^{n} a_{ik} x_{kj} = \delta_{ij} \quad (i, j = 1, 2, \dots, n),$$

где

$$\delta_{ij} = \begin{cases} 1, \text{когда} & i = j, \\ 0, \text{когда} & i \neq j. \end{cases}$$

Полученные n систем линейных уравнений для $j = 1, 2, \ldots, n$, имеющих одну и ту же матрицу A и различные свободные члены, одновременно можно решить методом Гаусса.

Исправление элементов приближенной обратной матрицы. Пусть имеем неособенную матрицу A и требуется найти обратную матрицу A^{-1}. Положим, что мы получили приближенное значение обратной матрицы $D_0 \approx A^{-1}$. Тогда для улучшения точности можно воспользоваться методом последовательных приближений в специальной форме. В качестве предварительной меры погрешности используем разность

$$F_0 = E - AD_0.$$

Если $F_0 = 0$, то очевидно, что $D_0 = A^{-1}$, поэтому, если модули элементов матрицы F_0 малы, то матрицы A^{-1} и D_0 близки между собой. Будем строить последовательные приближения по формуле

$$D_k = D_{k-1} + D_{k-1}F_{k-1} \quad (k = 1, 2, 3, \ldots), \tag{7.21}$$

причем соответствующая погрешность есть $F_k = E - AD_k$.

Оценим скогость сходимости последовательных приближений. Имеем

$$F_1 = E - AD_1 = E - A(D_0 + D_0F_0) = E - AD_0(E + F_0) =$$

$$= E - (E - F_0)(E + F_0) = E - \left(E - F_0^2\right) = F_0^2.$$

Аналогично, $F_2 = F_1^2 = F_0^4$ и, вообще,

$$F_k = F_0^{2^k} \quad (k = 1, 2, 3, \ldots). \tag{7.22}$$

Докажем, что если

$$\|F_0\| \leq q < 1, \tag{7.23}$$

где $\|F_0\|$ – какая-нибудь каноническая норма матрицы F_0, то процесс итерации (7.4) сходится, т.е. $\lim\limits_{k \to \infty} D_k = A^{-1}$.

Действительно, из формулы (7.20) имеем

$$\|F_k\| \leq \|F_0\|^{2^k} \leq q^{2^k}.$$

Поэтому $\lim\limits_{k\to\infty}\|F_k\|=0$ и, следовательно,

$$\lim_{k\to\infty} F_k = \lim_{k\to\infty}\left(E - AD_k\right)=0$$

или $E - A\lim\limits_{k\to\infty} D_k = 0$, т.е. $\lim\limits_{k\to\infty} D_k = A^{-1}E = A^{-1}$.

Таким образом, утверждение доказано.

В частности, используя m-норму, получаем, что если элементы матрицы $F_0 = \lfloor f_{ij} \rfloor$ удовлетворяют неравенству

$$\left| f_{ij} \right| \le \frac{q}{n},$$

где n – порядок матрицы и $0 \le q < 1$, то процесс итерации (7.4) заведомо сходится.

Предполагая неравенство (7.6) выполненным, оценим погрешность

$$R_k = \left\| A^{-1} - D_k \right\| \le$$

$$= \left\| A^{-1} \right\| \| E - AD_k \| = \left\| A^{-1} \right\| \| F_k \| \le \left\| A^{-1} \right\| q^{2^k}.$$

Так как $AD_0 = E - F_0$, то

$$A^{-1} = D_0\left(E - F_0\right)^{-1} = D_0\left(E + F_0 + F_0^2 + \ldots\right).$$

Отсюда

$$\left\| A^{-1} \right\| \le \| D_0 \| \left\{ \| E \| + q + q^2 + \ldots \right\} = \| D_0 \| \left\{ \| E \| + \frac{q}{1-q} \right\}.$$

Для m –нормы или l-нормы имеем $\|E\| = 1$, и поэтому

$$\left\| A^{-1} \right\| < \frac{\| D_0 \|}{1-q}.$$

Таким образом,

$$\left\| A^{-1} - D_k \right\| \le \frac{\| D_0 \|}{1-q} \| F_k \| \qquad (7.24)$$

или

$$\left\| A^{-1} - D_k \right\| \le \frac{\| D_0 \|}{1-q} q^{2^k}, \qquad (7.25)$$

где норма понимается в смысле m – нормы или l–нормы. Из формулы (7.25) следует, что сходимость процесса (7.4) при $q << 1$ очень быстрая.

На практике процесс уточнения элементов обратной матрицы прекращают, когда обеспечено неравенство

$$\|D_k - D_{k-1}\| \le \varepsilon,$$

где ε – заданная точность.

7.3 Методы оптимального исключения, обратных итераций и квадратного корня

Метод оптимального исключения по существу является вариацией метода Гаусса. Идея этого метода состоит в том, что последовательным исключением неизвестных матрица системы приводится к диагональному виду. Возможность же таких эквивалентных преобразований следует из теоремы о приведении матрицы к диагональному виду, формулировку которой (без доказательства) мы сейчас приведем.

Теорема 7.1. Для любой квадратной вещественной матрицы A n-го порядка существуют такие квадратные вещественные матрицы U и V (n-го порядка), что матрица UAV – диагональная.

Опишем вкратце схему оптимального исключения. Пусть у нас есть система общего вида (24.1) с матрицей A. Допустим, что коэффициент $a_{11} \neq 0$. Разделив все коэффициенты первого уравнения на a_{11}, получим систему:

$$\begin{aligned}
x_1 &+ a_{12}^1 x_2 + \ldots + a_{1n}^1 x_n = b_1^1,\\
a_{21}x_1 &+ a_{22}x_2 + \ldots + a_{2n}x_n = b_2,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
a_{n1}x_1 &+ a_{n2}x_2 + \ldots + a_{nn}x_n = b_n.
\end{aligned}$$

Умножив первое уравнение на коэффициент $-a_{21}$, сложив его со вторым и поделив все коэффициенты второго уравнения на $a_{22} - a_{12}^1 a_{21}$ (в предположении, что $a_{22} - a_{12}^1 a_{21} \neq 0$), придем к системе:

$$\begin{aligned}
x_1 &+ a_{12}^1 x_2 + \ldots + a_{1n}^1 x_n = b_1^1,\\
&\quad x_2 + \ldots + a_{2n}^1 x_n = b_2^1,\\
a_{31}x_1 &+ a_{32}x_2 + \ldots + a_{3n}x_n = b_3,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
a_{n1}x_1 &+ a_{n2}x_2 + \ldots + a_{nn}x_n = b_n.
\end{aligned}$$

Умножив теперь второе уравнение на $-a_{12}^{1}$ и сложив его с первым, исключим неизвестное x_2 из первого уравнения.

Допустим, что таким образом мы преобразовали первые k уравнений $(k \geq 1)$, исключив из первого уравнения системы неизвестные $x_2, \cdots, x_k$, из второго неизвестные $x_1, x_3, \ldots, x_k$; наконец, из k-го уравнения – неизвестные $x_1, x_2, \ldots, x_{k-1}$. То есть допустим, что система $Ax = b$ приведена к эквивалентной системе вида:

$$
\begin{array}{l}
x_1 + a_{1,k+1}^{k} x_{k+1} + \ldots + a_{1n}^{k} x_n = b_1^{k}, \\
x_2 + a_{2,k+1}^{k} x_{k+1} + \ldots + a_{2n}^{k} x_n = b_2^{k}, \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_k + a_{k,k+1}^{k} x_{k+1} + \ldots + a_{kn}^{k} x_n = b_k^{k}, \\
a_{k+1,1} x_1 + a_{k+1,2} x_2 + \ldots + a_{k+1,k} x_k + a_{k+1,k+1} x_{k+1} + \ldots + a_{k+1,n} x_n = b_{k+1}, \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
a_{n1} x_1 + a_{n2} x_2 + \ldots + a_{nk} x_k + a_{n,k+1} x_{k+1} + \ldots + a_{nn} x_n = b_n.
\end{array}
$$

Исключим теперь неизвестные $x_1, x_2, \ldots, x_k$ из $k+1$-го уравнения. Для этого сложим его с первым, умноженным на $-a_{k+1,1}$; со вторым, умноженным на $-a_{k+1,2}$, и т.д. Разделим полученное в результате этих действий уравнение на коэффициент при x_{k+1}; $k+1$-ое уравнение примет вид:

$$x_{k+1} + a_{k+1,k+2}^{k+1} x_{k+2} + \ldots + a_{k+1,n}^{k+1} x_n = b_{k+1}^{k+1}.$$

Умножая это уравнение на

$$-a_{1,k+1}^{k}, \quad -a_{2,k+1}^{k}, \quad \ldots \quad , \quad -a_{k,k+1}^{k}$$

и складывая его соответственно с первым, вторым, … , k-ым уравнениями, получим систему, в первых k уравнениях которой коэффициенты при x_{k+1} равны нулю.

После аналогичных преобразований всех уравнений системы получим систему с диагональной матрицей:

$$x_i = b_i^{n} \quad (i = 1, \ldots, n).$$

Заметим, что количество арифметических операций для решения системы методом оптимального исключения примерно то же самое, что и в методе Гаусса. Однако этот метод требует несколько меньшего объема используемой памяти ЭВМ.

Теорема 7.2 (разложение Холецкого). *Если* $A = A^*$ *и главные миноры* A *отличны от нуля, то существует разложение*

$$A = C^* DC,$$

где C *– верхняя треугольная матрица,* D *– диагональная матрица*.

Этот метод применяется, если матрица A системы *симметрическая*, т. е. когда $a_{ij} = a_{ji},\ i, j = \overline{1,n}$

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

В нашем случае, если A симметрическая, то

$$A = C^T DC,$$

где

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ 0 & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & c_{nn} \end{bmatrix},\ c_{ii} > 0,\ i = \overline{1,n},$$

$$D = \begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn} \end{bmatrix},\ d_{ii} = \pm 1,\ i = \overline{1,n}.$$

Тогда вместо системы

$$Ax = b \qquad (7.26)$$

мы будем решать систему

$$C^T DCx = b. \tag{7.27}$$

Пусть $C^T D = T$, тогда $TCx = b$ и получим систему

$$\begin{cases} Cx = y, \\ Ty = b. \end{cases} \tag{7.28}$$

Решение системы (7.3) свелось к решению системы (7.26) с верхней треугольной матрицей C и нижней треугольной матрицей T (следовательно, это ускоряет процесс нахождения x методом Гаусса).

Итак, имеем

$$T = C^T D = \begin{bmatrix} c_{11} & 0 & \dots & 0 \\ c_{12} & c_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ c_{1n} & c_{2n} & \dots & c_{nn} \end{bmatrix} \begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn} \end{bmatrix} =$$

$$\begin{bmatrix} c_{11}d_{11} & 0 & \dots & 0 \\ c_{12}d_{11} & c_{22}d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ c_{1n}d_{11} & c_{2n}d_{22} & \dots & c_{nn}d_{nn} \end{bmatrix}.$$

Так как $A = TC$, то для того чтобы найти элемент a_{ij} надо i-ю строку матрицы T умножить на j-й столбец матрицы C:

$$a_{ij} = \begin{bmatrix} c_{1i}d_{11} & c_{2i}d_{22} & \dots & c_{ii}d_{ii} & 0 & \dots & 0 \end{bmatrix} \cdot \begin{bmatrix} c_{1j} \\ c_{2j} \\ \dots \\ c_{jj} \\ 0 \\ \dots \\ 0 \end{bmatrix}.$$

Таким образом, если $i \le j$, получим

$$a_{ij} = c_{1i}d_{11}c_{1j} + c_{2i}d_{22}c_{2j} + \dots + c_{ii}d_{ii}c_{ij} = \sum_{k=1}^{\min(i,j)} c_{ki}c_{kj}d_{kk}. \tag{7.29}$$

1) При $i=1, j=1$ из (7.27) имеем $a_{11}=c_{11}^2 d_{11},\ d_{11}=sign(a_{11}),$ $c_{11}=\sqrt{|a_{11}|}$.

2) При $i=1, j=\overline{2,n}$ из (28.4) имеем $a_{1j}=c_{11}c_{1j}d_{11}$ (c_{11}, d_{11} – найдны, a_{1j} – известны), поэтому

$$c_{1j}=\frac{a_{1j}}{c_{11}d_{11}},\ j=\overline{2,n}.$$

Итак, нашли первую строку матрицы C.

3) При $i=j$ из (28.4) $a_{ii}=c_{1i}^2 d_{11}+c_{2i}^2 d_{22}+\ldots+c_{ii}^2 d_{ii}$, отсюда

$$c_{ii}^2 d_{ii}=a_{ii}-\sum_{k=1}^{i-1} c_{ki}^2 d_{kk},$$

$$d_{ii}=sign\left(a_{ii}-\sum_{k}^{i-1} c_{ki}^2 d_{kk}\right),$$

$$c_{ii}=\sqrt{a_{ii}-\sum_{k=1}^{i-1} c_{ki}^2 d_{kk}}$$

(у нас по условию $c_{ii}>0$).

4) Из последнего слагаемого формулы (7.27):

$$c_{ij}=\frac{a_{ij}-\sum_{k=1}^{i-1} c_{ki}d_{kk}c_{kj}}{c_{ii}d_{ii}},\ i\neq j.$$

Таким образом, мы последовательно находим все строки матрицы C. Эти все полученные формулы для случая $i\le j$, а при $i>j$ $c_{ij}=0$.

Из формул видно, что матрица C получилась треугольного вида, а поскольку матрица $T=C^T D$, то матрица T тоже будет треугольного вида. Таким образом, решение системы вида $Ax=b$ мы свели к решению системы с треугольными матрицами

$$\begin{cases} Ty=b, \\ Cx=y. \end{cases}$$

Определение 7.1. Квадратная матрица называется положительно определённой $(A>0)$, *если* $\forall x\in R^n, x\neq 0$ $(Ax,x)>0$.

В случае реализации метода квадратного корня для матрицы $A>0$ имеем $d_{ii}=1$, т.е.

$$D=\begin{bmatrix}1 & 0 & 0 & \dots & 0\\ 0 & 1 & 0 & \dots & 0\\ \dots & \dots. & \dots & \dots & \dots\\ 0 & 0 & 0 & \dots & 1\end{bmatrix}.$$

Тогда наши формулы упростятся, так как будут отсутствовать d_{ii}, так как

$$A=C^TC.$$

Итак, вычисляется матрица C: $c_{11}=\sqrt{|a_{11}|}$, $c_{1j}=\dfrac{a_{1j}}{c_{11}}$,…, тогда

$$T=C^TD=\begin{bmatrix}c_{11} & 0 & \dots & 0\\ c_{12} & c_{22} & \dots & 0\\ \dots & \dots & \dots & \dots\\ c_{1n} & c_{2n} & \dots & c_{nn}\end{bmatrix}=C^T.$$

Сначала решается система $Ty=b$, т.е.

$$C^Ty=b:\ y_1=\frac{b_1}{c_{11}},$$

в общем, имеем равенство

$$y_i=\frac{b_i-\sum\limits_{k=1}^{i-1}c_{ki}y_k}{c_{ii}},\ i>1.$$

Затем решаем

$$Cx=y:\ x_n=\frac{y_n}{c_{nn}}$$

и получим её решения

$$x_i=\frac{y_i-\sum\limits_{k=i+1}^{n}c_{ik}x_k}{c_{ii}},\ i<n.$$

Контроль за ходом вычислений здесь такой же, как и в методе Гаусса (так как две системы решаются методом Гаусса).

7.4 Прогонка и метод ортогонализации

Наиболее важным частным случаем метода Гаусса является *метод прогонки,* применяемый к системам с трёхдиагональной матрицей (они часто встречаются при решениях краевых задач для дифференциальных уравнений второго порядка). Такие системы обычно записывают в каноническом виде

$$a_i x_{i-1} - b_i x_i + c_i x_{i+1} = d_i, \quad 1 \le i \le n, \quad a_1 = c_n = 0. \tag{7.30}$$

Формула (7.30) называется *разностным уравнением второго порядка, или трёхточечным уравнением.* В этом случае прямой ход (без выбора главного элемента) сводится к исключению элементов a_i. Получается треугольная система, содержащая в каждом уравнении только два неизвестных, x_i и x_{i+1}. Поэтому формулы обратного хода имеют следующий вид:

$$x_i = \xi_{i+1} x_{i+1} + \eta_{i+1}, \quad i = n, \quad n-1, \ldots, 1. \tag{7.31}$$

Уменьшим в формуле (7.31) индекс на единицу и подставим в уравнение (7.30):

$$a_i(\xi_i x_i + \eta_i) - b_i x_i + c_i x_{i+1} = d_i.$$

Выражая отсюда x_1 через x_{i+1}, получим

$$x_i = \frac{c_i}{b_i - a_i \xi_i} x_{i+1} + \frac{a_i \eta_i - d_i}{b_i - a_i \xi_i}.$$

Чтобы это выражение совпало с (7.31), надо, чтобы стоящие в его правой части дроби были равны соответственно ξ_{i+1} и η_{i+1}. Отсюда получим удобную запись формул прямого хода

$$\begin{aligned} &\xi_{i+1} = c_i / (b_i - a_i \xi_i), \\ &\eta_{i+1} = (a_i \eta_i - d_i)/(b_i - a_i \xi_i), \quad i = 1, 2, \ldots, n \end{aligned} \tag{7.32}$$

Попутно можно найти определитель трёхдиагональной матрицы

$$\det A = \prod_{i-1}^{n} (a_i \xi_i - b_i). \tag{7.33}$$

Вычисления по формулам прогонки (7.30)-(7.31) требуют всего $3n$ ячеек памяти и $9n$ арифметических действий, т.е. они гораздо экономнее общих формул метода исключения.

В формулах прямого и обратного хода начало счета «замаскировано»: для начала (*развязки)* расчета формально требуется задать величины ξ_1, η_1 и $x_{n+1,}$ которые неизвестны. Однако перед этими величинами в формулах стоят множители a_i или $\xi_{n+1} \sim c_n$, равные нулю. Это позволяет начать вычисления, полагая, например, $\xi_1 = \eta_1 = x_{n+1} = 0.$

Покажем, что если выполнено условие *преобладания диагональных элементов*

$$|b_i| \geq |a_i| + |c_i| \tag{7.34}$$

(причем хотя бы для одного i имеет место неравенство), то в формулах прямого хода (7.32) не возникает деления на нуль, и тем самым исходная система (7.3) имеет единственное решение. Для этого предположим, что $|\xi| < 1$ при некотором значении индекса. Тогда легко проверяется цепочка неравенств

$$|\xi_{i+1}| = |c_i| / |b_i - a_i\xi_i| \leq |c_i| / \left(|b_i| - |a_i||\xi_i|\right) \leq |c_i| / \left(|c_i| + |a_i| - |a_i| \times |\xi_i|\right) < 1.$$

Поскольку можно положить $\xi_1 = 0,$ отсюда по индукции следует $|\xi_i| < 1;$ значит, $|b_i - a_i\xi_i| > |c_i| \geq 0,$ что и требовалось доказать. При выполнении условия (7.34) формулы прогонки устойчивы относительно ошибок округления и позволяют успешно решать системы уравнений с несколькими сотнями неизвестных.

Условие (7.34) является достаточным, но не необходимым условием устойчивости прогонки. Конечно, можно построить примеры неустойчивости при несоблюдении этого условия. Но в практических расчетах для хорошо обусловленных систем типа (7.3) прогонка часто оказывается достаточно устойчивой даже при нарушении условия преобладания диагональных элементов.

Заметим, что к линейным системам с трёхдиагональной матрицей обычно приводят трёхточечные разностные схемы для дифференциальных уравнений второго порядка.

Будем считать, что определитель системы (7.3) отличен от нуля и, следовательно, система имеет единственное решение. Запишем её в виде

$$\begin{aligned} &a_{11}x_1 + \ldots + a_{1n}x_n + a_{1n+1} = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &a_{n1}x_1 + \ldots + a_{nn}x_n + a_{nn+1} = 0, \end{aligned} \tag{7.35}$$

где $a_{i\,n+1} = -b_i.$

Исключим из расстояния неинтересный случай, когда система (7.30) является однородной и все a_{in+1} равны нулю.

Введем $n+1-$мерные векторы $a^i=(a_{i1},\ldots,a_{in},a_{in+1})$ и $y=(x_1,\ldots,x_n,1)$. Это позволит придать уравнениям (7.30) форму условий ортогональности вектора y к векторам $a^i(i=1,\ldots,n)$:

$$(a^i,y)=0 \quad (i=1,\ldots,n), \tag{7.36}$$

что дает возможность применить к нахождению y хорошо известный процесс ортогонализации. Единственное ограничение, которому нужно заранее подчинить вектор y,– это равенство единице его последней составляющей.

Ортогональность y к a^i равносильна ортогональности y к линейному подпространству P_n, натянутому на векторы a^i, и, следовательно, к любому базису этого подпространства. Построим ортогональный базис при помощи известного *процесса ортогонализации* системы векторов:

$$u^1=a^1, v^k=\frac{1}{\|u^k\|}u^k,\ldots,u^k=a^k-\sum_{i=1}^{k-1}(a^k,v^i)v^i,\ldots$$

Векторы $v^i(i=1,2,\ldots,n)$ образуют в P_n ортонормированный базис.

Присоединим теперь к $v^i(i=1,\ldots,n)$ еще любой вектор a^{n+1}, от них линейно не зависящий. Таким будет, например, вектор $a^{n+1}=(0,\ldots,0,1)$. Выделим из него часть, ортогональную ко всем $v^i(1,\ldots,n)$, положив

$$u^{n+1}=a^{n+1}-\sum_{i=1}^{n}(a^{n+1},v^i)v^i.$$

Для u^{n+1} выполняются равенства $(u^{n+1},v^i)=0$ или равносильные им $(u^{n+1},a^i)=0$ $(i=1,\ldots,n)$.

Запишем их более подробно:

$$\begin{aligned}
&(u^{n+1},a^1)=a_{11}u_1{}^{n+1}+\ldots+a_{1n}u_n{}^{n+1}+u_{n+1}^{n+1}a_{1n+1}=0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(u^{n+1},a^n)=a_{n1}u_1{}^{n+1}+\ldots+a_{nn}u_n{}^{n+1}+u_{n+1}^{n+1}a_{nn+1}=0.
\end{aligned} \tag{7.37}$$

Последняя составляющая u_{n+1}^{n+1} вектора u^{n+1} отлична от нуля, ибо если она равна нулю, то составляющие $u_1^{n+1},\ldots,u_n^{n+1}$ были бы решением однородной системы с матрицей $A=\{a_{ij}\}$.

Но так как эта система может иметь только нулевое решение, то все составляющие вектора u^{n+1} были бы равны нулю и вектор a^{n+1} являлся бы линейной комбинацией векторов v^i, что противоречит выбору a^{n+1}.

Если уравнения системы (7.3) разделить на u_{n+1}^{n+1}, то будет видно, что вектор $y=(x_1,\ldots,x_n,1)$, для которого

$$x_i=\frac{u_i^{n+1}}{u_{n+1}^{n+1}}(i=1,\ldots,n),$$

будет решением системы (7.37), а вектор $x=(x_1,\ldots,x_n)$ – решением заданной системы (7.3). Метод ортогонализации требует выполнения большего числа умножений и делений, чем метод Гаусса, и в этом отношении уступает ему.

7.5 Плохо обусловленные системы

Учитывая распространённость СЛАУ (так как на определённом этапе к ним сводится процесс моделирования), попытаемся охарактеризовать степень неопределённости этих задач. Рассмотрим, как погрешность матрицы и столбца свободных членов (изменения элементов матрицы и столбца свободных членов) влияет на решение системы

$$Ax=b. \tag{7.38}$$

Пусть Δb_i – изменение i-ой компоненты столбца b, Δx_i – изменение x_i вектора x, а $\Delta A=[\Delta a_{ij}]$ – изменение матрицы A.

В качестве меры отклонения решения x от $x+\Delta x$ будем рассматривать отношения $\frac{\|\Delta x\|}{\|x\|}$ и $\frac{\|\Delta x\|}{\|x+\Delta x\|}$.

Если нам вместо точной правой части b известно приближение $b+\Delta b$, тогда вместо системы (7.38) мы будем решать

$$A(x+\Delta x)=b+\Delta b \tag{7.39}$$

и мерой отклонения Δx к x будет число. Убедимся в этом.

Подставим (7.37) в (7.38) и получим $A\Delta x = \Delta b$, следовательно

$$\Delta x = A^{-1}\Delta b. \qquad (7.40)$$

Нормируя (7.40) имеем $\|b\| \le \|A\|\|x\|$, $\|\Delta x\| \le \|A^{-1}\|\|\Delta b\|$, где матричная норма должна быть согласована с выбранной векторной нормой. Полученные неравенства перемножим

$$\|b\|\|\Delta x\| \le \|A\|\|x\|\|A^{-1}\|\|\Delta b\| \Rightarrow \frac{\|\Delta x\|}{\|x\|} \le \|A\|\|A^{-1}\|\frac{\|\Delta b\|}{\|b\|} = \frac{\mu_1}{\mu_2}\frac{\|\Delta b\|}{\|b\|},$$

то есть

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{\mu_1}{\mu_2}\cdot\frac{\|\Delta b\|}{\|b\|}$$

($\mu_1 = \|A\|$ – наибольшее, а $\mu_2 = \dfrac{1}{\|A^{-1}\|}$ – наименьшее сингулярное число матрицы A).

Примечание 7.1. *Сингулярные числа матрицы* A *удовлетворяют уравнению* $A^*Ax = \mu^2 x$.

Обозначим $cond(A) = \dfrac{\mu_1}{\mu_2}$ (от английского conditioned – «обусловленный»).

Определение 7.2. Число $cond(A)$ *называется числом (мерой) обусловленности для невырожденной матрицы* A.

Далее вместо (1) будет рассматриваться система $(A+\Delta A)(x+\Delta x) = b$. В этом случае мера отклонения

$$\frac{\|\Delta x\|}{\|x+\Delta x\|} \le cond(A)\frac{\|\Delta A\|}{\|A\|} = \frac{\mu_1}{\mu_2}\cdot\frac{\tilde{\mu}_1}{\mu_1},$$

где $\tilde{\mu}_1$ – наибольшее сингулярное число матрицы ΔA.

Если $\dfrac{\mu_1}{\mu_2}$ – относительно мало, то говорят, что матрица A *хорошо обусловлена по отношению к решению системы* (1), если же $\dfrac{\mu_1}{\mu_2}$ – относительно велико – *плохо обусловлена* (т.е. сильнее сказывается на решении СЛАУ ошибка в исходных данных).

Свойства меры обусловленности

1) $cond(A) = \|A\|\|A^{-1}\| \geq \|AA^{-1}\| = \|E\| = 1;$

2) $cond(AB) \leq cond(A)cond(B);$

3) Если $A = A^*$, то $cond(A) = \dfrac{|\lambda_{\max}|}{|\lambda_{\min}|};$

4) $cond(\alpha A) = \|\alpha A\|\|(\alpha A)^{-1}\| = |\alpha|\|\alpha^{-1}\|\|A\|\|A^{-1}\| = \|A\|\|A^{-1}\| = cond(A), \alpha \in P.$

Пример 7.1. Решить систему

$$\begin{cases} x_1 + 0{,}95x_2 = 1{,}95 \\ 0{,}95x_1 + 0{,}97x_2 = 1{,}92 \end{cases}$$

для симметрической матрицы A $(A = A^T)$. Определить меру обусловленности матрицы системы.

Решение.

Точное решение этой системы: $x_1 = 1, x_2 = 1$. Вместо b рассмотрим $b + \Delta b$, тогда получим систему:

$$\begin{cases} x_1 + 0{,}95x_2 = 2{,}031, \\ 0{,}95x_1 + 0{,}97x_2 = 1{,}8606, \end{cases} \text{ где } \begin{cases} \Delta b_1 = 2{,}031 - 1{,}95 = 0{,}081, \\ \Delta b_2 = 1{,}8606 - 1{,}92 = -0{,}0594. \end{cases}$$

Здесь решение: $x_1* = 3, x_2* = 2$, но $\Delta x_1 = x_1* - x_1 = 2$,

$$\Delta x_2 = x_2* - x_2 = -2{,}02.$$

Т.е. $\left.\begin{matrix} |\Delta x_1| > |x_1| \\ |\Delta x_2| > |x_2| \end{matrix}\right\} \Rightarrow$ малому изменению правой части соответствует большое изменение решения.

Посчитаем $\dfrac{\mu_1}{\mu_2} - ?$ Для этого найдём собственные значения для матрицы A (это μ_1 и μ_2, так как матрица A симметрическая).

Составим характеристический многочлен $P_A(x)$ и приравняем его к нулю: $\begin{vmatrix} 1-\lambda & 0{,}95 \\ 0{,}95 & 0{,}97-\lambda \end{vmatrix} = 0$, имеем $\lambda^2 - 1{,}97\lambda + 0{,}0675 = 0$. Отсюда $\lambda_1 \approx 1{,}955; \quad \lambda_2 \approx 0{,}015$, и, значит, $\mu_1 = 1{,}955; \quad \mu_2 = 0{,}015$. Тогда $\dfrac{\mu_1}{\mu_2} = 130{,}4 = cond(A)$, т.е. число обусловленности большое, в этом случае матрица системы плохо обусловлена по отношению к решению системы.

Пример 7.2. $H = \left(\frac{1}{i+j-1} \right)_{i,j=1}^{n}$ – матрица Гильберта. Данная матрица – плохо обусловлена. Так при $n = 8$ $cond(H_8) > 10^{10}$ (матрица встречается при нахождении коэффициентов многочленов наилучшего квадратического приближения методом наименьших квадратов).

7.6 Итерационные методы. Принцип сжимающих отображений в метрических пространствах

При использовании современной вычислительной техники очень удобны *итерационные методы.* Эти методы применяются для приближенного решения алгебраических и трансцендентных уравнений, систем уравнений и других задач вычислительной математики. Решение уравнения или системы уравнений при помощи итерационного метода получается как предел последовательности приближений, вычисляемых в ходе *процесса итераций.* Итерационные методы часто называют также *методами последовательных приближений.*

При решении задачи методом последовательных приближений с самого начала задают некоторые приближённые значения неизвестных. Из этих начальных приближений получают новые, «улучшенные» приближённые значения. Новые приближенные значения снова «улучшают» и т.д. При определенных условиях построенная таким образом последовательность приближений сходится к точному решению.

При применении того или иного итерационного метода прежде всего встает вопрос о том, сходится ли построенная данным методом последовательность приближений и является ли предел этой последовательности (если он существует) решением поставленной задачи. Множество начальных приближений, при которых последовательность приближений сходится к решению задачи, называют *областью сходимости* метода.

Кроме сходимости к решению, при применении итерационных методов существенной является и *скорость сходимости*. Для некоторых задач сходимость итерационного процесса может оказаться настолько медленной, что практически достигнуть удовлетворенной близости к решению невозможно. В связи с этим при применении итерационных методов важную роль играет *предварительная подготовка*, то есть сведение данной задачи к задаче, для которой бы выбранный итерационный процесс сходился по возможности быстро. Большое значение имеют также различные приемы ускорения сходимости.

В основе многих итерационных методов лежит так называемый *принцип сжимающих отображений.* Этот принцип широко применяется как для доказательства теорем существования и единственности решения уравнений и систем уравнений различных типов, так и при исследовании сходимости итерационных методов.

Прежде чем сформулировать и доказать принцип сжимающих отображений, напомним ряд понятий и фактов функционального анализа.

Определение 7.3. Множество R *элементов произвольной природы называется метрическим пространством, если каждой упорядоченной паре элементов* $x, y \in R$ *поставлено в соответствие неотрицательное число* $\rho(x, y)$*, называемое расстоянием между этими элементами или метрикой пространства* R *и удовлетворяющее следующим условиям (аксиомы метрики):*

1) $\rho(x, y) = 0$ *тогда и только тогда, когда* $x = y$;

2) $\rho(x, y) = \rho(y, x)$ *для любых* x *и* y *;*

3) $\rho(x, z) \le \rho(x, y) + \rho(y, z)$ *для любых* x, y, z *(неравенство треугольника).*

Элементы метрического пространства R называют также *точками* этого пространства.

Пример 7.3. Любое множество точек на вещественной прямой E' является метрическим пространством с расстоянием $\rho(x, y) = |x - y|$.

Пример 7.4. Любое множество точек $n-$мерного евклидова пространства E^n также является метрическим пространством. Расстояние между точками $x = (x_1, x_2, ..., x_n)$ и $y = (y_1, y_2, ..., y_n)$ этого пространства определяется формулой:

$$\rho(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Выполнение аксиом 1) и 2) здесь очевидно. Для проверки аксиом 3) (неравенства треугольника) воспользуемся *неравенством Коши:*

$$\left| \sum_{i=1}^{n} a_i b_i \right| \le \sqrt{\sum_{i=1}^{n} a_i^{\,2}} \sqrt{\sum_{i=1}^{n} b_i^{\,2}},$$

которое имеет место для любых чисел $a_1, a_2, ..., a_n, b_1, b_2, ..., b_n$.

В самом деле, если $z = (z_1, z_2, ..., z_n)$, то, полагая в этом неравенстве $a_i = x_i - y_i$, $b_i = y_i - z_i$, находим:

$$\rho^2(x,z)=\sum_{i=1}^{n}(x_i-z_i)^2=\sum_{i=1}^{n}((x_i-y_i)+(y_i-z_i))^2=$$

$$=\sum_{i=1}^{n}(x_i-y_i)^2+2\sum_{i=1}^{n}(x_i-y_i)(y_i-z_i)+\sum_{i=1}^{n}(y_i-z_i)^2\le$$

$$\le\sum_{i=1}^{n}(x_i-y_i)^2+2\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}\sqrt{\sum_{i=1}^{n}(y_i-z_i)^2}+$$

$$+\sum_{i=1}^{n}(y_i-z_i)^2=\left(\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}+\sqrt{\sum_{i=1}^{n}(y_i-z_i)^2}\right)^2=$$

$$=(\rho(x,y)+\rho(y,z))^2.$$

Пример 7.5. Любое множество $n-$мерных векторов становится метрическим пространством также в том случае, если расстояние между элементами этого множества определяется формулой:

$$\rho_1(y,z)=\max_{1\le i\le n}|x_i-y_i|.$$

Справедливость условий 1) – 3) определения 34.1 здесь очевидна.

Пример 7.6. Во множестве $n-$мерных векторов можно ввести расстояние и по такой формуле: $\rho_2(x,y)=\sum_{i=1}^{n}|x_i-y_i|$. Условия 1) – 3) определения 34.1, очевидно, выполняются, и, таким образом, любое векторное множество с метрикой ρ_2 также является метрическим пространством.

Последние три примера показывают, что одно и то же множество элементов может быть метризовано по-разному; при этом получаются различные метрические пространства.

Неравенство треугольника можно обобщить на случай любого числа элементов $x^{(1)}$, $x^{(2)},\ldots,x^{(m)}$:

$$\rho\left(x^{(1)},x^{(m)}\right)\le\rho\left(x^{(1)},x^{(2)}\right)+\rho\left(x^{(2)},x^{(3)}\right)+\ldots+\rho\left(x^{(m-1)},x^{(m)}\right).$$

Последнее неравенство получается путем последовательного применения неравенства треугольника.

Отметим еще одно простое свойство расстояний, которое можно назвать «неравенством четырехугольника»: для любых точек x,y,z,u метрического пространства

$$|\rho(x,z)-\rho(y,u)|\le\rho(x,y)+\rho(z,u).$$

Геометрически это означает, что разность длин двух сторон четырехугольника не превосходит суммы длин двух других его сторон. Доказательство вытекает из неравенств:

$$\rho(x,z)\le\rho(x,y)+\rho(y,u)+\rho(u,z) \text{ и } \rho(y,u)\le\rho(y,x)+\rho(x,z)+\rho(z,u).$$

Расстояние, определенное для элементов произвольной природы позволяет дать обобщение одного из важнейших понятий математического анализа – понятия предела.

Будем говорить, что *последовательность точек* $x^{(1)}, x^{(2)},\ldots,x^{(k)},\ldots$ *произвольного метрического пространства* R *сходится к точке* x *того же пространства* R, *если* $\lim\limits_{k\to\infty}\rho\left(x, x^{k}\right)=0.$

Точка x *в этом случае называется пределом последовательности* $\left\{x^{(k)}\right\}$ *и обозначается:* $\lim\limits_{k\to\infty}x^{(k)}=x.$

Следующее утверждение очевидно. Во-первых, никакая последовательность не может иметь двух различных пределов; во-вторых, если последовательность $\left\{x^{(k)}\right\}$ сходится в точке x, то и всякая её подпоследовательность сходится в этой точке.

В качестве следствия из неравенства четырехугольника можно получить доказательство непрерывности расстояния $\rho(x,y)$ (как функции от x и y) в том смысле, что если $x^{(k)}\to x$ и $y^{(k)}\to y$, то $\rho\left(x^{(k)},y^{(k),}\right)\to\rho(x,y)$ при $k\to\infty$. Действительно, с помощью неравенства четырехугольника имеем

$$\left|\rho\left(x^{(k)},y^{(k)}\right)-\rho(x,y)\right|\le\rho\left(x^{(k)},x\right)+\rho\left(y^{(k)},y\right)\to 0 \text{ при } k\to\infty.$$

Определение 7.4. Последовательность $\left\{x^{(k)}\right\}$ *точек метрического пространства* R *называется фундаментальной, если для любого положительного* ε *существует номер* N, *такой, что для любых* k *и* l, *больших* N, *выполняется неравенство:* $\rho\left(x^{(k)},\ x^{(l)}\right)<\varepsilon.$

Кратко в таком случае можно писать: $\lim\limits_{k,\,l\to\infty}\rho\left(x^{(k)},x^{(l)}\right)=0.$

Любая сходящаяся последовательность является фундаментальной. Действительно, пусть $\lim\limits_{k\to\infty}x^{(k)}=x.$ Тогда

$$\rho\left(x^{(k)},x^{(l)}\right)\le\rho\left(x^{(k)},x\right)+\rho\left(x,x^{(l)}\right)\to 0 \text{ при } k,l\to\infty.$$

Если $R-$ вещественная прямая с обычной метрикой, то понятие фундаментальной последовательности точек пространств R совпадает с классическим понятием фундаментальной числовой последовательности.

В теории числовых последовательностей имеется *критерий Коши*, в силу которого всякая фундаментальная числовая последовательность является сходящейся.

В общем метрическом пространстве критерий Коши уже несправедлив. Например, рассмотрим открытый интервал (0,1). Он представляет собой метрическое пространство с обычной метрикой числовой оси $\left(\rho(x,\, y)=|x-y|\right)$. Последовательность $\frac{1}{2},\frac{1}{3},...,\frac{1}{n},...,$ очевидно, является фундаментальной в этом метрическом пространстве, но она не является в нем сходящейся.

Определение 7.5. Метрическое пространство R называется полным, если в нем всякая фундаментальная последовательность является сходящейся.

Пример 7.7. Проверим что $n-$мерное евклидово пространство E^{n} с расстоянием

$$\rho(x,\, y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

является полным метрическим пространством. Пусть векторы $x^{(k)}=\left(x_1^{(k)},\ x_2^{(k)},...,x_n^{(k)}\right),\ \ k=1,\ 2,...,$ составляют фундаментальную последовательность. Поскольку

$$\left|x_i^{(k)}-x_i^{(l)}\right|\le\sqrt{\sum_{i=1}^{n}\left(x_i^{(k)}-x_i^{(l)}\right)^2}=\rho\left(x^{(k)},\, x^{(l)}\right),$$

числовая последовательность $\left\{x_i^{(k)}\right\}$ при каждом фиксированном $i=1,\ 2...,n$ является фундаментальной числовой последовательностью и имеет некоторый предел x_i. Числа $x_1,\, x_2,\ ...,x_n$ определяют вектор $x\in E^n$.

Поскольку при $k\to\infty,$

$$\rho\left(x,\, x^{(k)}\right)=\sqrt{\sum_{i=1}^{n}\left(x_i-x_i^{(k)}\right)^2}\to 0$$

вектор $x=\left(x_1,\, x_2,...,x_n\right)$ есть предел взятой фундаментальной последовательности. Итак, каждая фундаментальная последовательность $n-$мерного евклидова пространства E^{n} имеет предел, что и требовалось доказать.

Пример 7.8. Множество всех $n-$мерных векторов, в котором введена метрика

$$\rho_1(x, y)=\max_{1\le i\le n}|x_i - y_i|,$$

также является полным метрическим пространством. Доказательство аналогично приведенному для примера 7.7.

Пример 7.9**.** Совершенно аналогично примеру 7.8 доказывается, что метрическое пространство, состоящее из всех $n-$ мерных векторов с расстоянием

$$\rho_2(x, y)=\sum_{i=1}^{n}|x_i - y_i|,$$

полно.

Пусть X и $Y-$ два метрических пространства. Будем говорить, что оператор $y=\Phi x$ отображает X в Y, если каждому элементу $x\in X$ поставлен в соответствие элемент $y\in Y$.

Если $X = Y$, то этот оператор задаёт отображение Φ пространства X в себя.

Определение 7.6. Всякая точка x, которая переводится отображением Φ в себя (т.е. для которой $\Phi x = x$), называется неподвижной точкой *этого отображения.*

Определение 7.7. Пусть $R-$ метрическое пространство, а оператор $y=\Phi x$ отображает R в себя. Если при некотором α, $0<\alpha<1$, отображение $y=\Phi x$ удовлетворяет условию $\rho(\Phi x, \Phi x')\le\alpha\rho(x, x')$ для любых $x, x'\in R$, то такое отображение называют сжимающим*.*

Теорема 7.3 (*принцип сжимающих отображений). Если отображение $y=\Phi x$ сжимающее ($x\in R, \Phi: R\to R$, $R-$полное метрическое пространство), то существует одна и только одна неподвижная точка этого отображения, то есть уравнение $x=\Phi x$ имеет единственное решение. Решение этого уравнения может быть получено как предел последовательности*

$$x^{(k)}=\Phi x^{(k-1)},\quad k=1,2,\ldots,$$

где $x^{(0)}-$произвольный элемент из R.

Доказательство.

Пусть $x^{(0)}-$ произвольный элемент из R. Рассмотрим последовательность $x^{(k)}=\Phi x^{(k-1)}$, $k=1, 2,\ldots.$ Докажем, что она является фундаментальной.

Действительно, для любого k имеем:

$$\rho\left(x^{(k)}, x^{(k+1)}\right)=\rho\left(\Phi x^{(k-1)}, \quad \Phi x^{(k)}\right)\le \alpha\rho\left(x^{(k-1)}, x^{(k)}\right)\le$$

$$\le \alpha^2\rho\left(x^{(k-2)}, x^{(k-1)}\right)\le \ldots \le \alpha^k\rho\left(x^{(0)}, x^{(1)}\right).$$

Следовательно, при любых k и l, $k<l$,

$$\rho\left(x^{(k)}, x^{(l)}\right)\le \rho\left(x^{(k)}, x^{(k+1)}\right)+$$

$$+\rho\left(x^{(k+1)}, x^{(k+2)}\right)+\ldots+\rho\left(x^{(l-1)}, x^{(l)}\right)\le \left(\alpha^k+\alpha^{k+1}+\ldots+\alpha^{l-1}\right)\rho\left(x^{(0)}, x^{(1)}\right)=$$

$$=\frac{\alpha^k-\alpha^l}{1-\alpha}\rho\left(x^0, x^{(1)}\right)\le \frac{\alpha^k}{1-\alpha}\rho\left(x^{(0)}, x^{(1)}\right).$$

При достаточно большом k эта величина может быть сделана сколь угодно малой $(\alpha<1)$, что и доказывает фундаментальность выбранной последовательности $\left\{x^{(k)}\right\}$.

Так как R полно, то существует элемент $x^*=\lim\limits_{k\to\infty} x^k$. Для этого элемента x^* имеем:

$$\rho(\Phi x^*, x^*)\le \rho\left(\Phi x^*, x^{(k)}\right)+\rho\left(x^{(k)}, x^*\right)=\rho\left(\Phi x^*, \Phi x^{(k-1)}\right)+$$

$$+\rho\left(x^{(k)}, x^*\right)\le \alpha\rho\left(x^*, x^{(k-1)}\right)+\rho(x^{(k)}, x^{(*)})\to 0 \text{ при } k\to\infty.$$

Поскольку k произвольное, а $\rho(\Phi x^*, x^*)$ не зависит от k, получаем: $\rho(\Phi x^*, x^*)=0$, то есть $\Phi x^*=x^*$.

Остается показать, что x^*- единственная неподвижная точка отображения Φ. Допустим, что $z-$ вторая неподвижная точка, так что вместе с равенством $\Phi x^*=x^*$ имеет место и равенство $\Phi z=z$. Тогда

$$\rho\left(x^*, z\right)=\rho\left(\Phi x^*, \Phi z\right)\le \alpha\rho\left(x^*, z\right).$$

Если $\rho\left(x^*, z\right)>0$, то, сократив на $\rho\left(x^*, z\right)$, получим: $1\le\alpha-$ противоречие. Поэтому $\rho\left(x^*, z\right)=0$, $x^*=z$, то есть неподвижной точки, отличной от x^*, не существует. Теорема доказана.

Точки $x^{(k)}$, получаемые по формуле $x^{(k)}=\Phi x^{(k-1)}, k=1,2,\ldots$, есть *последовательные приближения решения* x^* *уравнения* $x=\Phi x$.

Из полученного выше неравенства

$$\rho\left(x^{(k)}, x^{(l)}\right) \le \frac{a^k}{1-\alpha}\rho\left(x^{(0)}, x^{(1)}\right), \text{ где } k < l,$$

переходом к пределу при $l \to \infty$ и фиксированном k, получаем (используя непрерывность расстояния):

$$\rho\left(x^{(k)}, x^{(*)}\right) \le \frac{a^k}{1-\alpha}\rho\left(x^{(0)}, x^{(1)}\right) \qquad (7.41)$$

Последнее неравенство дает оценку расстояния между точным решением x^* и его приближением $x^{(k)}$. Это расстояние убывает со скоростью геометрической прогрессии со знаменателем α. неравенство (7.41) позволяет также в конкретных задачах заранее оценить число шагов, необходимое для вычисления x^* с заданной точностью.

7.7 Метод простой итерации и метод Зейделя решения линейных систем алгебраических уравнений

Покажем, как применяется принцип сжимающих отображений к исследованию сходимости итерационных методов решения систем уравнений.

Пусть дана система уравнений специального вида

$$x = Bx + \tilde{b}, \qquad (7.42)$$

где

$$B = \begin{bmatrix} b_{11} & \dots & b_{1m} \\ \dots & \dots & \dots \\ b_{m1} & \dots & b_{mm} \end{bmatrix}, \; x = \begin{bmatrix} x_1 \\ \dots \\ x_m \end{bmatrix}, \; \tilde{b} = \begin{bmatrix} \tilde{b}_1 \\ \dots \\ \tilde{b}_m \end{bmatrix}.$$

Правую часть уравнения (7.42) обозначим через $\Phi(x)$, где $\Phi(x)$ можно рассматривать как отображение пространства R^m в R^m, где $\forall x = (x_1, x_2, \dots, x_m) \in R^m$ ставим в соответствии вектор $y = (y_1, y_2, \dots, y_m) \in R^m$, т.е. $y = \Phi(x)$.

Тогда координаты вектора y вычисляются по формуле:

$$y_i = \sum_{j=1}^{m} b_{ij} x_j + \tilde{b}_i, \; i = \overline{1,m}.$$

Решение уравнения (7.42) сведётся к отысканию неподвижной точки отображения Φ: $\Phi(x) = x$. Для того, чтобы отображение Φ имело бы одну неподвижную точку, нужно чтобы Φ было сжатием. Если Φ сжатие, то оно имеет в пространстве R^m единственную неподвижную точку $x*$ и к ней сходится итерационный процесс $x_{n+1} = \Phi(x_n)$, где $x_0 \in R^m$.

Имеем

$$x^{(n+1)} = Bx^{(n)} + \tilde{b}, \tag{7.43}$$

где (7.43) – *метод простой итерации решения СЛАУ* (7.1).

В координатной форме метод (7.43) запишется в виде

$$x_i^{(n+1)} = \sum_{j=1}^{m} b_{ij} x_j^{(n)} + \tilde{b}_i, \; i = \overline{1,m}. \tag{7.43`}$$

При каких условиях отображение Φ будет сжимающим. Ответ зависит не только от самого Φ, но и от выбора метрики в действительном пространстве R^m.

Рассмотрим первую метрику пространства R^m – кубическую.

$$\rho_1(x, x') = \max_{1 \le i \le m} |x_i - x_i'|,$$

где $x = (x_1, x_2, \ldots, x_m)$, $x' = (x_1', x_2', \ldots, x_m')$.

Рассмотрим расстояние между их образами $y = \Phi(x)$, $y_i = \sum_{j=1}^{m} b_{ij} x_j + \tilde{b}_i$, $y' = \Phi(x')$, $y_i' = \sum_{j=1}^{m} b_{ij} x_j' + \tilde{b}_i$.

$$\rho_1(\Phi(x), \Phi(x')) = \max_{1 \le i \le m} |y_i - y_i'| = \max_{1 \le i \le m} \left| \sum_{j=1}^{m} b_{ij} x_j + \tilde{b}_i - \sum_{j=1}^{m} b_{ij} x_j' - \tilde{b}_i \right| =$$

$$= \max_{1 \le i \le m} \left| \sum_{j=1}^{m} b_{ij} (x_j - x_j') \right| \le \max_{1 \le i \le m} \sum_{j=1}^{m} |b_{ij}| \max_{1 \le j \le m} |x_j - x_j'| = \max_{1 \le i \le m} \sum_{j=1}^{m} |b_{ij}| \; \rho_1(x, x').$$

Следовательно,

$$\|B\|_1 = \max_{1 \le i \le m} \sum_{j=1}^{m} |b_{ij}| \le q_1 < 1, \tag{7.44}$$

где (7.44) – *максимальная строчная матричная норма*.

Рассмотрим октаэдрическую метрику:

$$\rho_2(x,x') = \sum_{i=1}^{m} |x_i - x_i'|, \ \forall x, x' \in R^m.$$

Аналогично 1^0, получим, что Φ – сжатие, если

$$\|B\|_2 = \max_{1 \le j \le m} \sum_{i=1}^{m} |b_{ij}| \le q_2 < 1, \qquad (7.45)$$

где (7.45) – *максимальная столбцовая матричная норма*.

Рассмотрим сферическую метрику:

$$\rho_3(x,x') = \left\{ \sum_{i=1}^{m} (x_i - x_i')^2 \right\}^{1/2}, \ \forall x, x' \in R^m.$$

Аналогично 1^0, получим, что Φ – сжатие, если

$$\|B\|_3 = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} b_{ij}^{\ 2}} \le q_3 < 1, \qquad (7.46)$$

где (7.46) – *норма Фробениуса*.

Если выполнено одно из условий (7.44) – (7.46), то отображение Φ является сжатием и по принципу Банаха для него в пространстве R^m существует единственная неподвижная точка $x^* = \left(x_1^*, x_2^*, \ldots, x_m^*\right)$, к которой сходится итерационный процесс (7.4). В результате нами доказана

***Теорема* 7.4.** *Если для матрицы B системы* (36.1) *выполняется одно из условий:* $\|B\|_l \le q_l < 1, \ l = 1,2,3$, *то система линейных алгебраических уравнений вида* $x = Bx + \tilde{b}$ *имеет единственное решение* $x^* = \left(x_1^*, x_2^*, \ldots, x_m^*\right)$, *которое может быть получено как предел последовательности, вычисляемой по формуле* (36.2), *начиная с произвольного* $x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_m^{(0)}\right) \in R^m$, *причём скорость сходимости процесса* (36.2) *определяется соотношением*

$$\rho_l\left(x^{(n)}, x^*\right) \le \frac{q_l^n}{1 - q_l} \cdot \rho_l\left(x^{(1)}, x^{(0)}\right).$$

Если же мы имеем систему вида $Ax = b$, то преобразуем её: $Ax - b = 0$. Далее умножим на $-\tau$: $-\tau(Ax - b) = 0$ и прибавим к обеим частям уравнения по x: $-\tau(Ax - b) + x = x$.

Тогда итерационный процесс (7.43) запишется в виде:

$$x^{(n+1)} = x^{(n)} - \tau(Ax^{(n)} - b),\ x^{(n+1)} = (E - \tau A)x^{(n)} + \tau b. \qquad (7.43``)$$

Обозначим $B = E - \tau A \Rightarrow$ для сходимости итерационного процесса (7.43``) надо, чтобы $\|B\| = \|E - \tau A\| < 1.$

Замечание 7.1. *Если* $A > 0$ *(A – положительно определённая, т. е. все её собственные значения* $\lambda_i > 0$*), то* $\tau = 1/\|A\|$.

В процессе (7.43) до конца выполнения просчета $n+1$ шага должны сохраняться значения n-го шага. Этого недостатка избавлен *метод Зейделя*, который является модификацией метода простой итерации

$$x_1^{(n+1)} = \sum_{j=1}^{m} b_{1j} x_j^{(n)} + \tilde{b}_1,$$

$$x_2^{(n+1)} = b_{21} x_1^{(n+1)} + \sum_{j=2}^{m} b_{2j} x_j^{(n)} + \tilde{b}_2, \ldots, \qquad (7.44)$$

$$x_k^{(n+1)} = \sum_{j=1}^{k-1} b_{kj} x_j^{(n+1)} + \sum_{j=k}^{m} b_{kj} x_j^{(n)} + \tilde{b}_k.$$

Подробнее

$$x_k^{(n+1)} = b_{k1} x_1^{(n+1)} + b_{k2} x_2^{(n+1)} + \ldots + b_{kk-1} x_{k-1}^{(n+1)} + b_{kk} x_k^{(n)} + \ldots + b_{km} x_m^{(n)} + \tilde{b}_k.$$

Метод Зейделя (36.6) позволяет сразу же использовать при вычислении последних координат вектора $\bar{x}^{(n+1)}$ уже найденные его координаты. Условие сходимости для метода простой итерации и метода Зейделя одинаковы. Области сходимости для этих методов не совпадают, но если они пересекаются или совпадают, метод Зейделя сходится быстрее.

Если СЛАУ (36.1) решается с заданной точностью, то для определения момента останова в методах простой итерации и Зейделя целесообразно использовать правило останова по соседним приближениям (по поправкам).

Глава 8 Интерполирование функций. Численные методы интерполирования

8.1 Постановка задачи интерполирования функций

Пусть функция $f(x)$ задана таблицей значений для конечного множества x:

Таблица 3 – Задание функции

x	x_0	x_1	x_2	…	x_n
$f(x)$	$f(x_0)$	$f(x_1)$	$f(x_2)$	…	$f(x_n)$

Такая таблица может быть получена в результате наблюдений за ходом некоторого процесса. Если необходимо найти значение $f(x)$ для промежуточного значения аргумента, то строят функцию $\varphi(x)$, достаточно простую для вычислений, которая в заданных точках

$$x_0, x_1, \ldots, x_n$$

принимает значения

$$f(x_0),\ f(x_1), \ldots,\ f(x_n),$$

а в остальных точках отрезка $[a,b]$, принадлежащего области определения $f(x)$, приближенно представляет функцию $f(x)$ с той или иной степенью точности и при решении задачи вместо функции $f(x)$ используется $\varphi(x)$.

Задача построения функции $\varphi(x)$ называется *задачей интерполирования*. Чаще всего интерполяционную функцию представляют в виде алгебраического многочлена некоторой степени.

К интерполированию прибегают и в том случае, когда для функции $f(x)$ известна аналитическое выражение, с помощью которого можно вычислить её значение для любого значения x из отрезка, в котором она определена, но вычисление каждого значения сопряжено с большим объемом вычислений.

Если надо найти значение функции для большого количества значений аргумента, то прибегают к интерполированию, т.е. вычисляют несколько значений $f(x_i)$, $i = 0, 1, 2, \ldots, n$ и по ним строят простую интерполирующую функцию $\varphi(x)$, посредством которой и вычисляют приближенные значения $f(x)$ в других точках.

В качестве функции $\varphi(x)$ будем брать в дальнейшем алгебраический многочлен степени n, интерполяция в этом случае называется алгебраической. Алгебраическая интерполяция функции $y = f(x)$ на отрезке $[a,b]$ состоит в приближенной замене этой функции на данном отрезке многочленом $P_n(x)$ степени n, т.е.

$$f(x) \approx P_n(x),$$

причем $P_n(x)$ принимает в точках $x_0, x_1, ..., x_n$ те же значения, что и $f(x)$

$$f(x_i) = P_n(x_i),\ i = 0,\ 1,\ 2,\ ...,\ n.$$

Отметим, что для данной функции $f(x)$ не может существовать двух различных интерполяционных многочленов одной и той же степени n.

8.2 Интерполяционный многочлен Лагранжа

Поставим следующую задачу: построить многочлен $P_n(x)$ степени n, который в $n+1$ данных точках $x_0, x_1,...,x_n$ (узлах интерполирования) принимает значения $y_0, y_1,...,y_n$.

Для решения задачи определим так называемые фундаментальные многочлены $Q_n^k(x)$, т.е. многочлены n-ой степени относительно x, удовлетворяющие следующим условиям:

$$Q_n^k(x_i) = \begin{cases} 0 & при\ \ i \neq k, \\ 1 & при\ \ i = k. \end{cases} \qquad (8.1)$$

Искомый многочлен запишется в виде суммы при помощи фундаментальных многочленов

$$P_n(x) = y_0 Q_n^0(x) + y_1 Q_n^1(x) + ... + y_n Q_n^n(x). \qquad (8.2)$$

Поскольку $x_0,\ x_1,\ ...,\ x_{k-1},\ x_{k+1},\ ...,\ x_n$ – нули многочлена $Q_n^k(x)$, то

$$Q_n^k(x) = c(x - x_0)(x - x_1)...(x - x_{k-1})(x - x_{k+1})...(x - x_n).$$

Определим c из условия $Q_n^k(x_k) = 1$, получим:

$$Q_n^k(x) = \frac{(x - x_0)(x - x_1)...(x - x_{k-1})(x - x_{k+1})...(x - x_n)}{(x_k - x_0)(x_k - x_1)...(x_k - x_{k-1})(x_k - x_{k+1})...(x_k - x_n)}. \qquad (8.3)$$

Формула (8.2) с учетом (8.3) запишется:

$$P_n(x) = \sum_{k=0}^{n} y_k \frac{(x-x_0)(x-x_1)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_0)(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)}. \quad (8.4)$$

Многочлен, определенный формулой (40.4), называется интерполяционным *многочленом Лагранжа*, а фундаментальные многочлены (8.3) *коэффициентами Лагранжа*.

Формула

$$f(x) \approx P_n(x) = \sum_{k=0}^{n} y_k \frac{(x-x_0)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_0)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)} \quad (8.5)$$

называется интерполяционной формулой Лагранжа. Формулы (8.3–8.5) можно записать более компактно, если ввести следующее обозначение:

$$\omega(x) = (x-x_0)(x-x_1)\ldots(x-x_n). \quad (8.6)$$

Поскольку

$$\omega'(x_k) = (x_k-x_0)(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n),$$

то

$$Q_n^k(x) = \frac{\omega(x)}{(x-x_k)\omega'(x_k)}, \quad P_n(x) = \omega(x)\sum_{k=0}^{n}\frac{y_k}{(x-x_k)\omega'(x_k)},$$

$$f(x) \approx \omega(x)\sum_{k=0}^{n}\frac{y_k}{(x-x_k)\omega'(x_k)}.$$

Полином $P_n(x)$ совпадает с $f(x)$ в $n+1$ точке, а в остальных точках отрезка $[a,b]$ разность $R_n(x) = f(x) - P_n(x)$ называется остаточным членом интерполяции. Она отлична от нуля и представляет собой погрешность метода. Справедлива теорема для оценки остаточного члена.

***Теорема* 8.1.** *Если функция $f(x)$ в промежутке $[a,b]$ имеет непрерывные производные до $n+1$-го порядка, то остаточный член интерполяции $R_n(x)$ определяется формулой*

$$R_n(x) = f^{(n+1)}(\xi)\frac{\omega(x)}{(n+1)!},$$

где ξ – точка промежутка $[a,b]$, зависящая от x.

8.3 Конечные разности. Разделённые разности

Пусть

$$y_i = f(x_i) -$$

значения функции $y = f(x)$ в точках x_i, $i = 0,1,...,n$, тогда разности $y_1 - y_0$, $y_2 - y_1, ..., y_n - y_{n-1}$ называются конечными разностями 1-го порядка. Обозначим $\Delta y_i = y_{i+1} - y_i$. Разность второго порядка получается по формулам $\Delta^2 y_0 = \Delta y_1 - \Delta y_0$, $\Delta^2 y_1 = \Delta y_2 - \Delta y_1, ...$ Аналогично определяются последующие разности

$$\Delta^{n+1} y_0 = \Delta^n y_1 - \Delta^n y_0,\ \Delta^{n+1} y_1 = \Delta^n y_2 - \Delta^n y_1, ...$$

Для работы с конечными разностями удобно использовать таблицу 4.

Таблица 4 – Конечные разности

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
x_0	y_0				
		Δy_0			
x_1	y_1		$\Delta^2 y_0$		
		Δy_1		$\Delta^3 y_0$	
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$
		Δy_2		$\Delta^3 y_1$	
x_3	y_3		$\Delta^2 y_2$	·	·
		Δy_3			
x_4	y_4	·	·	·	·
…	…	…	…	…	…

Свойства конечных разностей

1. *Конечные разности суммы (разности) функций равны сумме (разности) конечных разностей этих функций.*

2. *При умножении функции на постоянный множитель конечные разности умножаются на этот множитель.*

3. *Конечные разности n-го порядка от многочленов степени n постоянны, а конечные разности n + 1 порядка равны нулю.*

Разделённые разности первого порядка определяются формулами:

$$f(x_1, x_0) = \frac{y_1 - y_0}{x_1 - x_0}, \quad f(x_2, x_1) = \frac{y_2 - y_1}{x_2 - x_1}, \quad \ldots, \quad f(x_n, x_{n-1}) = \frac{y_n - y_{n-1}}{x_n - x_{n-1}}.$$

Разделённые разности второго порядка получаются из разделённых разностей первого порядка по формулам:

$$f(x_2, x_1, x_0) = \frac{f(x_2, x_1) - f(x_1, x_0)}{x_2 - x_0}, \quad f(x_3, x_2, x_1) = \frac{f(x_3, x_2) - f(x_2, x_1)}{x_3 - x_1}.$$

Разделённые разности n-го порядка получаются из разделённых разностей $n-1$-го порядка по формулам:

$$f(x_n, x_{n-1}, \ldots, x_0) = \frac{f(x_n, x_{n-1}, \ldots, x_1) - f(x_{n-1}, x_{n-2}, \ldots, x_0)}{x_n - x_0}.$$

Разделённые разности в случае равностоящих узлов с шагом h, ($x_k = x_0 + kh, \; k = 0, \; 1, \; \ldots, \; n$) выражаются следующим образом:

$$f(x_1, x_0) = \frac{\Delta y_0}{h}, \quad f(x_2, x_1) = \frac{\Delta y_1}{h}, \ldots, \quad f(x_n, x_{n-1}) = \frac{\Delta y_{n-1}}{h}, \ldots$$

$$f(x_2, x_1, x_0) = \frac{f(x_2, x_1) - f(x_1, x_0)}{x_2 - x_0} = \frac{\frac{\Delta y_1}{h} - \frac{\Delta y_0}{h}}{2h} = \frac{\Delta y_1 - \Delta y_0}{2h^2} = \frac{\Delta^2 y_0}{2! h^2},$$

$$f(x_3, x_2, x_1) = \frac{\Delta^2 y_1}{2! h^2}, \ldots, \quad f(x_n, x_{n-1}, \ldots, x_0) = \frac{\Delta^n y_0}{n! h^n}.$$

Тем самым установлена связь между конечными и разделенными разностями.

8.4 Интерполяционный многочлен Ньютона

Интерполяционный многочлен Лагранжа, который можно построить при любом расположении узлов интерполяции, имеет один единственный недостаток. Если понадобится увеличить число знаков (следовательно, и степень многочлена) прибавлением нового узла, многочлен Лагранжа придется вычислять заново, так как каждый член зависит от узлов интерполирования. Указанным недостатком не обладает интерполяционный многочлен Ньютона.

Пусть дана функция $y = f(x)$, $y_k = f(x_k)$ – значения её в точках x_k, $k = 0, 1, \ldots, n$. Из первой разделенной разности $f(x, x_0) = \dfrac{f(x) - y_0}{x - x_0}$ получим

$$f(x) = y_0 + (x - x_0) f(x, x_0).$$

Поскольку

$$f(x, x_0, x_1) = \frac{f(x, x_0) - f(x_0, x_1)}{x - x_1},$$

то

$$f(x, x_0) = f(x_0, x_1) + (x - x_1) f(x, x_0, x_1),$$

следовательно,

$$f(x) = y_0 + (x - x_0) f(x, x_0) = y_0 + (x - x_0) f(x_0, x_1) + (x - x_0)(x - x_1) f(x, x_0, x_1).$$

Продолжая процесс, получим:

$$f(x) = y_0 + (x - x_0) f(x_0, x_1) + (x - x_0)(x - x_1) f(x_0, x_1, x_2) + \ldots + (x - x_0)(x - x_1) \times$$

$$\times \ldots (x - x_{n-1}) f(x_0, x_1, \ldots, x_n) + (x - x_0)(x - x_1) \ldots (x - x_n) f(x, x_0, x_1, \ldots, x_n) \quad (8.7)$$

или

$$f(x) = P_n(x) + (x - x_0)(x - x_1) \ldots (x - x_n) f(x, x_0, x_1, \ldots, x_n),$$

где

$$P_n(x) = y_0 + (x - x_0) f(x_0, x_1) + \ldots + (x - x_0)(x - x_1) \ldots (x - x_{n-1}) \times$$

$$\times f(x_0, x_1, \ldots, x_n). \quad (8.8)$$

Полагая в (8.7) $x = x_k$, получим $y_k = f(x_k) = P_n(x_k)$, $k = 0, 1, \ldots, n$, следовательно, многочлен (8.8) – интерполяционный многочлен для функции $y = f(x)$, построенный по $n + 1$ узлам $x_0, x_1, \ldots, x_n$.

Многочлен (8.8) называется *интерполяционным многочленом Ньютона*. Подставляя его в общую интерполяционную формулу, получим:

$$f(x) \approx y_0 + (x - x_0) f(x_0, x_1) + (x - x_0)(x - x_1) f(x_0, x_1, x_2) + \ldots$$

$$+ (x - x_0)(x - x_1) \ldots (x - x_{n-1}) f(x_0, x_1, \ldots, x_n). \quad (8.9)$$

В случае равностоящих узлов интерполяции $x_1 = x_0 + h$, $x_2 = x_0 + 2h,..,\ x_n = x_0 + nh$ из интерполяционной формулы Ньютона с учетом равенств $f(x_k, x_{k-1},...,x_0) = \dfrac{\Delta^k y_0}{h^k k!}$, $k = 0,\ 1,\ ...,\ n$, получается интерполяционная формула Ньютона «интерполирование вперед»

$$f(x) \approx y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)(x - x_1) + \frac{\Delta^3 y_0}{3!h^3}(x - x_0)(x - x_1)(x - x_2) +$$

$$+ ... + \frac{\Delta^n y_0}{n!h^n}(x - x_0)(x - x_1)...(x - x_{n-1}). \qquad (8.10)$$

«Интерполирование вперед» объясняется тем, что формула содержит заданные значения функции, соответствующие узлам интерполяции, находящимся только вправо от x_0. Формула (8.10) удобна при интерполировании функций для значений x, близких к наименьшему узлу x_0.

Положим $x = x_0 + ht$. Тогда

$$\frac{x - x_0}{h} = t,\ \frac{(x - x_0)(x - x_1)}{h^2} = t(t-1),$$

$$\frac{(x - x_0)(x - x_1)...(x - x_{n-1})}{h^n} = t(t-1)...(t - n + 1),$$

и формула (8.10) примет вид:

$$f(x) = f(x_0 + th) \approx y_0 + t\Delta y_0 + \frac{t(t-1)}{2!}\Delta^2 y_0 + \frac{t(t-1)(t-2)}{3!}\Delta^3 y_0 + ...$$

$$+ \frac{t(t-1)...(t-n+1)}{n!}\Delta^n y_0. \qquad (8.11)$$

Остаточный член для полинома (42.5) имеет вид:

$$R_n(x) = \frac{h^{n+1} f^{(n+1)}(\xi)}{(n+1)!} t(t-1)...(t-n), \text{ где } \xi \in [a,b].$$

Абсолютная погрешность метода по формуле Ньютона «интерполирование вперед» определяется неравенством:

$$|R_n(x)| \le \frac{M_{n+1} h^{n+1} |t(t-1)...(t-n)|}{(n+1)!}, \text{ где } M_{n+1} = \max_{a \le x \le b} |f^{(n+1)}(x)|.$$

Интерполяционную формулу Ньютона (8.10) можно записать так:

$$f(x) \approx y_n + (x - x_n) f(x_n, x_{n-1}) + (x - x_n)(x - x_{n-1}) f(x_n, x_{n-1}, x_{n-2}) + \ldots$$
$$+ (x - x_n)(x - x_{n-1}) \ldots (x - x_1) f(x_n, x_{n-1}, \ldots, x_0).$$

В случае равностоящих узлов из неё аналогично формуле Ньютона «интерполирование вперед» можно получить формулу Ньютона «интерполирование назад»

$$f(x) \approx y_n + t\Delta y_{n-1} + \frac{t(t+1)}{2!}\Delta^2 y_{n-2} + \frac{t(t+1)(t+2)}{3!}\Delta^3 y_{n-2} + \ldots + \frac{t(t+1)\ldots(t+n-1)}{n!} \times$$
$$\times \Delta^n y_0, t = \frac{x - x_n}{h}.$$

Формулу «интерполирование назад» используют при интерполировании функций в точках x, близких к наибольшему узлу x_n.

Абсолютная погрешность метода “интерполирование назад” определяется формулой:

$$|R_n(x)| \le \frac{M_{n+1} h^{n+1} |t(t+1)\ldots(t+n)|}{(n+1)!}, \text{ где } M_{n+1} = \max_{a \le x \le b} |f^{(n+1)}(x)|.$$

8.5 Интерполирование внутри таблицы. Интерполяционная формула Стирлинга

Предположим, что точка x лежит вблизи внутреннего узла x_k таблицы с любой стороны от него. Тогда табличные узлы можно привлекать для интерполирования в порядке удаленности от x_k, т.е. взять сначала узлы x_k и присоединить к нему пары узлов $(x_k + h, x_k - h)$, $(x_k + 2h, x_k - 2h)$, …, $(x_k + nh, x_k - nh)$. При таком порядке узлов интерполирования формула Ньютона будет иметь вид:

$$f(x) = y(x) = f(x_k) + (x - x_k) f(x_k, x_k + h) + (x - x_k)(x - x_k - h) f(x_k, x_k + h, x_k -$$
$$- h) + (x - x_k)(x - x_k - h)(x - x_k + h) f(x_k, x_k + h, x_k - h, x_k + 2h) + \ldots + (x - x_k)(x -$$
$$- x_k - h) \ldots (x - x_k - nh) f(x_k, x_k + h, x_k - h, \ldots, x_k - nh) + R_{2n}(x),$$

где

$$R_{2n}(x) = \frac{(x - x_k)(x - x_k - h) \ldots (x - x_k + nh)}{(2n+1)!} f^{(2n+1)}(\xi).$$

Здесь $\xi -$ есть точка отрезка, содержащего $x_k + nh, x_k - nh$ и x.

Заменим разностные отношения их выражениями через конечные разности

$$f(x_k) = y_k,\ f(x_k, x_k + h) = \frac{\Delta y_k}{1!h},$$

$$f(x_k, x_k + h, x_k - h) = f(x_k - h, x_k, x_k + h) = \frac{\Delta^2 y_{k-1}}{2!h^2}, \ldots$$

и введем переменную $t = \dfrac{x - x_k}{h}$, тогда получим:

$$y(x_k + th) = y_k + \frac{t}{1!}\Delta y_k + \frac{t(t-1)}{2!}\Delta^2 y_{k-1} + \frac{t(t-1)(t-2)}{3!}\Delta^3 y_{k-1} + \ldots$$
$$+ \frac{1}{(2n-1)!}(t-n+1)..t(t+1)\ldots(t+n-1)\Delta^{2n-1} y_{k-n+1} + \frac{1}{(2n)!}\times$$
$$\times (t+n-1)..t\ldots(t-n)\Delta^{2n} y_{k-n+1} + R_{2n}(x).$$

Для придания правой части симметричного вида перепишем равенство в форме

$$y(x_k + th) = y_k + t\left[\Delta y_k - \frac{1}{2}\Delta^2 y_{k-1}\right] + \frac{t^2}{2!}\Delta^2 y_{k-1} + \frac{t(t^2-1^2)}{3!}\left[\Delta^3 y_{k-1} - \frac{1}{2}\Delta^4 y_{k-2}\right] +$$
$$+ \ldots\frac{t(t^2-1^2)\ldots(t^2-(n-1)^2)}{(2n-1)!}\left[\Delta^{2n-1} y_{k-n+1} - \frac{1}{2}\Delta^{2n} y_{k-n}\right] +$$
$$+ \frac{t^2(t^2-1)\ldots(t^2-(n-1)^2)}{(2n)!}\Delta^{2n} y_{n-k} + R_{2n}(x).$$

Если из квадратных скобок исключить конечные разности четного порядка, пользуясь равенствами $\Delta^2 y_{k-1} = \Delta y_k - \Delta y_{k-1}$, $\Delta^4 y_{k-2} = \Delta^3 y_{k-1} - \Delta^3 y_{k-2}, \ldots$, то получим *интерполяционную формулу Ньютона-Стирлинга*:

$$y(x_k + th) = y_k + \frac{t}{1!}\frac{\Delta y_{k-1} + \Delta y_k}{2!} + \frac{t^2}{2!}\Delta^2 y_{k-1} + \frac{t(t^2-1^2)}{3!}\frac{\Delta^3 y_{k-2} + \Delta^3 y_{k-1}}{2} +$$
$$+ \frac{t^2(t^2-1^2)}{4!}\Delta^4 y_{k-2} + \ldots + \frac{t(t^2-1^2)\ldots(t^2-(n-1)^2)}{(2n-1)!}\frac{\Delta^{2n-1} y_{k-n} + \Delta^{2n-1} y_{k-n+1}}{2} +$$

$$+\frac{t^2(t^2-1^2)\ldots(t^2-(n-1)^2)}{(2n)!}\Delta^{2n}y_{k-n}+R_{2n}(x),$$

$$R_{2n}(x)=h^{2n+1}\frac{t^2(t^2-1^2)\ldots(t^2-n^2)}{(2n+1)!}y^{(2n+1)}(\xi).$$

Здесь $\xi-$ есть точка отрезка, содержащего x_n-kh+h, x_n+kh, x.

8.6 Численное дифференцирование (применение интерполирования к вычислению производных)

Численное дифференцирование применяется, если:

1) функция задана таблично,

2) функция задана неудобным для дифференцирования аналитическим выражением.

Задача численного дифференцирования некорректна, так как нарушается условие 3 корректности (решение непрерывно зависит от входных данных). При численном дифференцировании функцию $f(x)$ заменяют интерполяционным многочленом $P_n(x)$ и приближенно полагают

$$f'(x)=P'_n(x).$$

Близость значений функции $f(x)$ и полинома $P_n(x)$ не гарантирует близости их угловых коэффициентов φ_2 и φ_1 (см. рисунок 28).

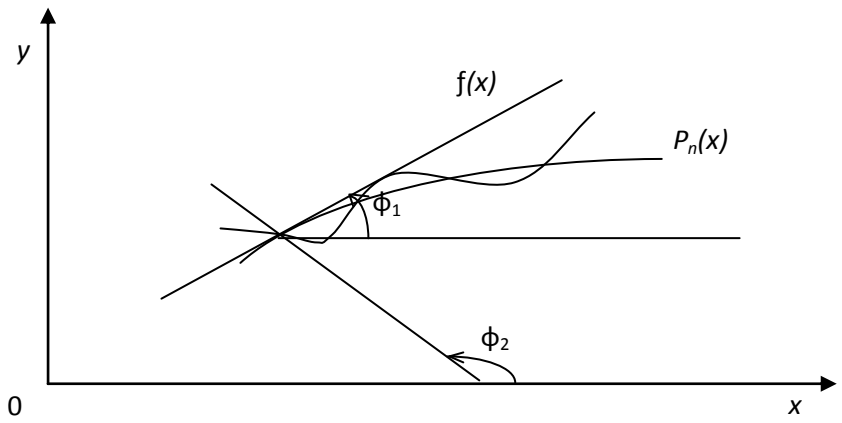


Рисунок 28

Пусть на отрезке $[a,b]$ рассматривается функция $f(x)$, имеющая непрерывную производную порядка $n+1$. Возьмем на $[a,b]$ $n+1$ различных узлов $x_0,x_1,\ldots,x_n$.

Для упрощения записи предположим, что они перенумерованы слева направо так, что $x_0 < x_1 < \ldots < x_n$. Интерполируем $f(x)$ по её значениям $f(x_0), f(x_1), \ldots, f(x_n)$ в узлах $x_0, x_1, \ldots, x_n$ посредством многочлена $P_n(x)$ степени n и обозначим $R_n(x)$ – погрешность интерполирования:

$$f(x) = P_n(x) + R_n(x), \text{ причём } P_n(x_i) = f(x_i), i = \overline{0,n}.$$

Вычислим производную от f порядка m:

$$f^{(m)}(x) = P_n^{(m)}(x) + R_n^{(m)}(x). \qquad (8.12)$$

Пренебрегая величиной $R_n^{(m)}(x)$, получим формулу для приближённого вычисления производной:

$$f^{(m)}(x) \approx P_n^{(m)}(x). \qquad (8.13)$$

Ее погрешность равна $R_n^{(m)}(x)$. Пользоваться ею целесообразно при небольших порядках m производной, во всяком случае, когда $m \le n$, так как все производные от $P_n(x)$ выше n тождественно равны 0.

Будем считать, что в точке x $\omega^{(m)}(x) \neq 0$ и на отрезке $[\alpha, \beta]$ производная

$$\varphi^{(m)}(t) = R_n^{(m)}(t) - \frac{K}{(n+1)!}\omega^{(m)}(t)$$

имеет $n + 2 - m$ нулей (отрезок $[\alpha, \beta]$ – наименьший отрезок, содержащий точки x_0, x_n, x). K выбирается из условия, чтобы точка x была нулем функции $\varphi^{(m)}(t)$ т.е. $\varphi^{(m)}(x) = 0$.

Тогда для погрешности $R_n^{(m)}(x)$ вычислительной формулы (8.13) верно представление:

$$R_n^{(m)}(x) = \frac{\omega^{(m)}(x)}{(n+1)!} f^{(n+1)}(\xi), \xi \in [\alpha, \beta],$$

где

$$\omega(x) = (x - x_0)(x - x_1)\ldots(x - x_n).$$

Глава 9 Суммарно-разностная аппроксимация операторов функциональных пространств

9.1 Интерполяционный многочлен, производная и интеграл сеточной функции

При численном решении уравнения (*) с дифференциальным оператором часто применяется разностный метод (см., например, [2] и [3]), или метод сеток. В этом разделе приводятся основные идеи разностного метода и излагается теория суммарного метода решения функциональных уравнений и краевых задач с интегральным оператором.

Определение 9.1. ***Сеткой*** на отрезке ***[a, b]*** называется любое конечное или счетное множество несовпадающих упорядоченных точек этого отрезка. Будем обозначать через $^{n}\Omega$ сетку, удовлетворяющую условиям

$$a = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = b. \qquad (9.1)$$

Определение 9.2. Точки $x_i \in {}^{n}\Omega$, $i = 0, \ldots, n$ назовем ***узлами*** сетки $^{n}\Omega$.

Определение 9.3. ***Равномерной сеткой*** на отрезке ***[a, b]*** будем называть множество равноотстоящих друг от друга точек

$$^{n}\Omega = \{x_i = a + i\,h,\ i = 0, 1, \ldots, n\}, \qquad (9.2)$$

где параметр h, называемый ***шагом*** сетки, находится из условия $x_n = b$:

$$h = \frac{b-a}{n},\ n \geq 2. \qquad (9.3)$$

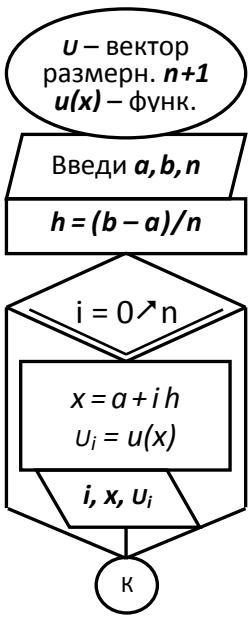


Рисунок 29

Определение 9.4. Множество всех узлов сетки x_k с индексами

$$0 \leq i - l \leq k \leq i + r \leq n,\ l + r \neq 0,$$

где l и r принадлежат N_0, называется ***шаблоном*** суммарно-разностной аппроксимации операторов проектирования, дифференцирования и интегрирования функции $u(x)$ в точке x_i или ее окрестности $\{x_{i-l}, x_{i+r}\}$.

Определение 9.5. Дискретная функция, определенная (заданная, вычисленная и т.п.) в точках сетки $^{n}\Omega$, называется ***сеточной функцией*** и обозначается

$$^{z}U \equiv \{u_i = u(x_i),\ i = 0, 1, \ldots, n\} \in R^{z},\ z = n+1. \qquad (9.4)$$

На рисунке 29 приведено графическое описание алгоритма табулирования функции $u(x)$ в точках сетки.

Здесь x – текущая точка сетки, которая играет роль индексированной переменной x_i из формулы (9.2). Идея использования параметра вместо массива связана с экономией памяти при больших значениях n и будет использоваться в случае несложного вычисления значения индексированной переменной.

Рассмотрим задачу интерполирования [12] бесконечно дифференцируемой функции $u(x) \in C^{\infty}_{[a,\,b]}$ на равномерной сетке.

Ее решением устанавливается взаимосвязь дискретного и полиномиального представления аналитических функций, то есть биективное отношение между значениями сеточной функции ${}^{z}U = \{u_i,\ i = 0,\ \ldots,\ n\}$ и коэффициентами интерполяционного многочлена ${}^{n}u(x) = c_0 + c_1 x + c_2 x^2 + \ldots + c_n x^n$. Основным условием задачи интерполяционного приближения является равенство значений функции $u(x)$ и многочлена ${}^{n}u(x)$ в узлах сетки ${}^{n}\Omega$

$$ {}^{n}u(x_i) \equiv c_0 + c_1 x_i + c_2 x_i^2 + \ldots + c_n x_i^n = u_i,\ i = 0,\ \ldots,\ n. \qquad (9.5) $$

Чтобы найти коэффициенты интерполяционного многочлена ${}^{n}u(x)$, запишем это условие в виде системы алгебраических уравнений:

$$ \begin{cases} c_0 + c_1 x_0 + c_2 x_0^2 + \ldots + c_n x_0^n = u(x_0) \\ \qquad\qquad \cdot\ \ \cdot\ \ \cdot \\ c_0 + c_1 x_i + c_2 x_i^2 + \ldots + c_n x_i^n = u(x_i) \\ \qquad\qquad \cdot\ \ \cdot\ \ \cdot \\ c_0 + c_1 x_n + c_2 x_n^2 + \ldots + c_n x_n^n = u(x_n) \end{cases}, \text{ где } x_i = a + i\,h,\ i = \overline{0,n},\ h = \frac{b-a}{n}. \qquad (9.6) $$

Полученная система уравнений линейна относительно координат искомого вектора ${}^{z}c = (c_0,\ \ldots,\ c_n)$, принадлежащего R^{z}. Решим систему (9.6) и найдем коэффициенты интерполяционного многочлена ${}^{n}u(x)$ заданной на $[a, b]$ сеточной функции ${}^{z}U$ ($s = 0,\ c = n$) с сеткой ${}^{n}\Omega$ (9.2) модифицированным методом Гаусса исключения неизвестных.

Перепишем систему (9.6) в матричном виде

$$ {}^{z}W\,{}^{z}c = {}^{z}U, \text{ где } {}^{z}W = \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ & & \cdot\ \cdot\ \cdot & & \\ 1 & x_i & x_i^2 & \ldots & x_i^n \\ & & \cdot\ \cdot\ \cdot & & \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix}. \qquad (9.7) $$

Матрица ${}^{z}W$, состоящая из коэффициентов при $\{c_j\}_0^n$, называется *матрицей Вандермонда*. Объект-схема алгоритма построения матрицы ${}^{z}W$ изображена на рисунке 30. Здесь n – число интервалов разбиения отрезка $[a, b]$, равное порядку многочлена ${}^{n}u(x)$. Отметим, что $\det({}^{z}W) \neq 0$.

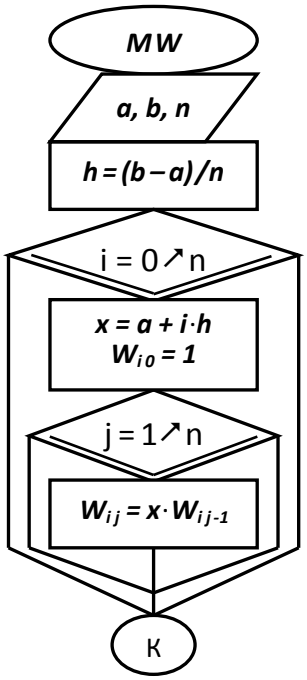


Рисунок 30

Процедура ***VKU(U, Ku)*** вычисления коэффициентов интерполяционного многочлена $^{n}u(x)$ заданной сеточной функции ^{z}U (со списком: U – вектор значений сеточной функции, ***Ku*** – вектор коэффициентов интерполяционного многочлена) включает построение матрицы Вандермонда $W = {}^{z}W$ и решение линейной системы с помощью процедуры ***RLS(0, n, W, U, Ku)***.

Пусть корень ***u*(x)*** уравнения (*) принадлежит функциональному нормированному пространству ***U*** с нормой $\| \|_U$, а $^{n}u(x)$ является элементом пробного пространства полиномиальных приближений $^{n}U \subset U$.

При увеличении ***n*** расстояние между этими функциями может стать меньше заданного параметра $\delta > 0$, которым разрешается пренебречь. Тогда решение задачи (*) сводится к определению ***N***, начиная с которого ^{N}U-пространство становится приближенным аналогом ***U***, а функция $^{N}u(x)$ – приближением (проекцией в ^{N}U) корня ***u*(x)*** с точностью $\acute{\varepsilon}(N) < \delta$, то есть

$$\acute{\varepsilon}(N) \equiv \rho(u^*(x), {}^{N}u(x)) < \delta. \qquad (9.8)$$

Для построения оператора проектирования непрерывных элементов функционального пространства ***U*** в множество ^{n}U будем использовать оператор вычисления значений функции в узлах сетки $P_y : U \to R^{n+1}$

$$^{z}U = P_y(u(x)) \equiv \{u_i = u(x)|_{x=x_i},\ i = 0, 1, \ldots, n\}. \qquad (9.9)$$

Рассмотрим задачу о приближенном вычислении правосторонней производной функции ***u(x)*** в точках сетки на отрезке ***[a, b]***. Здесь и далее будем считать, что функция ***u(x)*** обладает *необходимой* по ходу изложения *гладкостью*. Обозначим $\Delta x_i = x_{i+1} - x_i$ и $\Delta u_i = u_{i+1} - u_i$, $i = 0, 1, \ldots, n-1$.

Тогда искомая производная в точке x_i по определению равна

$$(u(x))'|_{x=x_i} = \lim_{\Delta x_i \to 0} \frac{\Delta u_i}{\Delta x_i}.$$

Отказавшись в формуле определения от предела и заменив Δx_i на ***h***, найдем приближенное значение $u'(x_i)$ с $1^{\text{-ым}}$ порядком точности по ***h***:

$$u'(x_i) \approx \frac{\Delta u_i}{h}. \qquad (9.10)$$

Определение 9.6. Правую часть (9.10) будем называть *правосторонней* ***разностной производной*** сеточной функции ${}^{z}U$ в точке x_i и обозначать

$$ {}^{+}u'_i = \frac{u_{i+1} - u_i}{h}. \qquad (9.11) $$

Аналогично определяется *левосторонняя* разностная производная ${}^{-}u'_i$ сеточной функции ${}^{z}U$ в узлах сетки x_i, $i = 1, 2, \ldots, n$. Увеличив количество точек шаблона аппроксимации до трех ($l = r = 1$), можно найти *центральную* производную ${}^{z}U$ в точке x_i, равную среднему значению ${}^{+}u'_i$ и ${}^{-}u'_i$.

Понятие «определенный интеграл (по Риману)» функции $u(x)$ на отрезке $[a, b]$ введем с помощью формулы предельного перехода

$$ \int_a^b u(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} u(x_{i-1} \le \xi_i \le x_i)(x_i - x_{i-1}) = \sum_{i=1}^{N} \overline{u}_i h_i, \qquad (9.12) $$

где $\overline{u}_i$ – среднее значение функции $u(x)$ на $i^{\text{-ом}}$ интервале.

Отказавшись в (9.12) от предела и заменив $u(\xi_i)$ на u_{i+1} или u_i, найдем приближенные значения интеграла с $1^{\text{-ым}}$ порядком точности по h:

$$ \int_a^b u(x)\,dx \approx h\sum_{i=0}^{n-1} u_i \quad \text{или} \quad \int_a^b u(x)\,dx \approx h\sum_{i=1}^{n} u_i. \qquad (9.13) $$

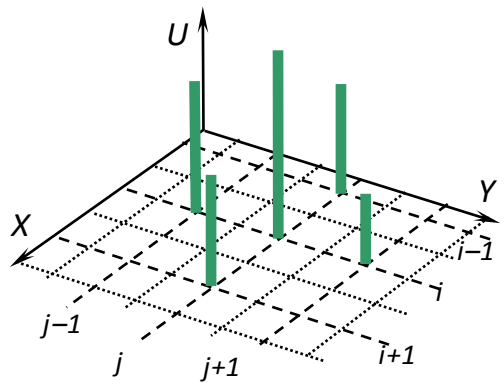


Рисунок 31

Определение 9.7. Левосторонним ***суммарным интегралом*** сеточной функции ${}^{z}U$ на отрезке $[a, b]$ назовем правую часть первого из соотношений (9.13) и обозначим

$$ \Sigma^{-}_{[a,b]} U = h\sum_{i=1}^{n-1} u_i + h\cdot u_0. \qquad (9.14) $$

Аналогично определим суммарный *правосторонний* интеграл $\Sigma^{+}_{[a,b]} U$.

По принципу одномерного пространства вводятся понятия «сетка», «узел», «шаг», «шаблон» и т.п. для ограниченной области M-мерного пространства

$$ X = \{x = ({}^{1}x, \ldots, {}^{M}x)\} \subset R^{M}. $$

Совокупность значений функции многих переменных, вычисленную в узлах M-мерной сетки на X, также назовем *сеточной функцией* ${}^{Z}U$. Частными производными по переменным ${}^{1}x, \ldots, {}^{M}x$ и интегралами этих функций являются суммарно-разностные соотношения вида (9.11) и (9.14).

Обычно при аппроксимации дифференциального оператора Лапласа ($M = 2,\ x = {}^1x,\ y = {}^2x$) разностным используется пятиточечный шаблон. На рисунке 31 он представлен в виде пяти *узлов шаблона* с координатами (x_m, y_k), где индексы переменной x

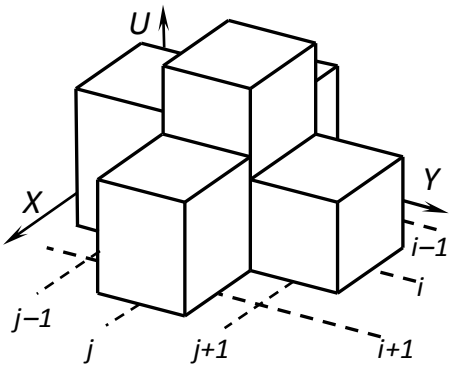


Рисунок 32

$$0 \le i - l \le m \le i + r \le n,\ l = r = 1 \text{ (если } k = j\text{)}$$

и переменной y

$$0 \le j - l \le k \le j + r \le n,\ l = r = 1 \text{ (если } m = i\text{)}.$$

Особенностью пятиточечного шаблона является его симметричность ($l = r$) относительно (x_i, y_j) и то, что хотя бы одна из координат любого его узла совпадает с соответствующей координатой (x_i или y_j) центральной точки шаблона. Значения сеточной функции двух переменных u_{ij} в узлах шаблона представлены в виде цилиндров высотой, равной этим значениям.

Как видно из рисунка 31, между штриховыми линиями сетки

$$x = x_i,\ i = 0, 1, \ldots, n \text{ и } y = y_j,\ j = 0, 1, \ldots, n$$

находятся воображаемые пунктирные линии, разбивающие все множество X на участки, где сеточная функция принимает постоянные значения (так называемые *ступеньки*, изображенные на рисунке 32).

Определение 9.8. Окрестность точки $x_{i_1 \ldots i_M} \in X$, равную основанию ступеньки, где сеточная функция ${}^zU = u_{i_1 \ldots i_M}$, назовем *сегментом* сетки ${}^n\Omega$ и обозначим его подчеркиванием $\underline{x}_{i_1 \ldots i_M}$. Совокупность сегментов сетки обозначим ${}^n\underline{\Omega}$.

9.2 Оператор дифференцирования функции и его дискретная аппроксимация

Каждый вид проецирования функций имеет свои преимущества и недостатки. Если для аппроксимации оператора дифференцирования функции использовать дискретное проектирование, то математическая модель исходной задачи будет выглядеть проще и поиск ее решения (даже при больших n) не вызовет больших затруднений. Однако соответствие полученных результатов корню операторного уравнения будет трудно оценить из-за того, что правомочность применения формул вида (9.11) достигается лишь при $h \to 0$. Тем не менее улучшить качество представления реального процесса с помощью математической модели возможно и в данном случае.

При определении левосторонней и правосторонней производных на равномерной сетке с шагом h мы указали, что, используя сеточную функцию и трехточечный шаблон, можно осуществить замену оператора дифференцирования центральной производной. Выясним, при каких условиях такая аппроксимация производной точнее, чем ее приближение с помощью односторонней производной достаточно гладкой функции.

Теорию аппроксимации дифференциальных операторов разностными D_h, заданными на сеточных функциях $\{u_{ij}\}$, рассмотрим для $M = 2$.

Определение 9.9. Оператор D дифференцирования функции $u(x, y)$

$$D(u) = \frac{\partial^n}{\partial x^k \partial y^l}(u(x, y)), \text{ где } k \geq 0, l \geq 0 \text{ и } k + l = n, \qquad (9.15)$$

назовем ***частной*** производной $u(x, y)$ порядка n. При k и l, неравных нулю, оператор D осуществляет ***смешанное*** дифференцирование функции $u(x, y)$.

Теорема 9.1. Оператор дифференцирования (9.15) линеен.

Доказательство соотношения

$$D(\alpha u + \beta v) = \alpha\, Du + \beta\, Dv \qquad (9.16)$$

следует из определения и свойств операции дифференцирования функций.

Теорема 9.2 [10, с. 564]. Если на множестве $X \subset R^2$ частные производные функции $u(x, y)$ непрерывны до n-го порядка включительно, то справедлива формула Тейлора в дифференциалах: $u(x + h, y + h) - u(x, y) =$

$$= h\left(\frac{\partial u(x,y)}{\partial x} + \frac{\partial u(x,y)}{\partial y}\right) + h^2\left(\frac{\partial^2 u(x,y)}{2!\,\partial x^2} + \frac{\partial^2 u(x,y)}{\partial x \partial y} + \frac{\partial^2 u(x,y)}{2!\,\partial y^2}\right) + \\ + \ldots + h^n \sum_{k=0}^{n} \frac{1}{k!\,(n-k)!}\,\frac{\partial^n u(x,y)}{\partial x^k \partial y^{n-k}} + o(h^n), \text{ где } (x, y) \in X. \qquad (9.17)$$

Определение 9.10. Разностный оператор D_h аппроксимирует в точке (x_i, y_j) дифференциальный оператор D с k-ым порядком точности по h, если

$$D_h\{u_{ij}\} - Du(x_i, y_j) = O(h^k). \qquad (9.18)$$

Из формулы Тейлора и теоремы о среднем следует, что

$$u(x_i + h, y_j) = u(x_i, y_j) + \frac{\partial u(x_i, y_j)}{\partial x} h + \frac{\partial^2 u(x_i, y_j)}{2!\,\partial x^2} h^2 + \\ + \frac{\partial^3 u(x_i, y_j)}{3!\,\partial x^3} h^3 + \frac{\partial^4 u(x_i + q_+ h, y_j)}{4!\,\partial x^4} h^4,\ 0 \leq q_+ \leq 1.$$

Аналогично доказывается соотношение

$$u(x_i - h, y_j) = u(x_i, y_j) - \frac{\partial u(x_i, y_j)}{\partial x} h + \frac{\partial^2 u(x_i, y_j)}{2!\,\partial x^2} h^2 - \frac{\partial^3 u(x_i, y_j)}{3!\,\partial x^3} h^3 + O(h^4).$$

В результате несложных вычислений получаем формулы

$$\frac{\partial u(x_i, y_j)}{\partial x} = \frac{u(x_i + h, y_j) - u(x_i - h, y_j)}{2h} + O(h^2) \qquad (9.19)$$

$$\frac{\partial^2 u(x_i, y_j)}{\partial x^2} = \frac{u(x_i + h, y_j) - 2u(x_i, y_j) + u(x_i - h, y_j)}{h^2} + O(h^2). \qquad (9.20)$$

Разностные операторы, стоящие в правых частях (9.19) и (9.20), аппроксимируют соответствующие дифференциальные операторы в точке (x_i, y_j) сетки со вторым порядком точности по h. Необходимым условием корректности этих формул является непрерывная дифференцируемость $u(x, y)$ по своим переменным до четвертого порядка включительно.

Теория аппроксимации операторов и разностные схемы решения задач с дифференциальным оператором изложены в книге И.С. Березина и Н.П. Жидкова [3].

Определение 9.11. Разностная схема (задача) $u_{ij} \approx u(x_i, y_j)$ аппроксимирует дифференциальную задачу с k-ым порядком точности по h, если

$$\max_{(x_i, y_j) \in X} |Du(x_i, y_j) - D_h\{u_{ij}\}| = O(h^k).$$

Определение 9.12. Разностная задача $u_{ij} \approx u(x_i, y_j)$ сходится к решению $u(x, y)$ дифференциальной задачи с k-ым порядком точности по h, если

$$\max_{(x_i, y_j) \in X} |u(x_i, y_j) - u_{ij}| = O(h^k).$$

9.3 Оператор интегрирования функции и его дискретная аппроксимация

Уравнение, которое содержит неизвестную функцию под знаком интеграла, назовем ***интегральным***. Таково, например, уравнение

$$u(x) - \lambda \int_a^b K(x,t)u(t)dt = v(x),\ x \in [a, b], \qquad (9.21)$$

где $u(x)$ – искомая, а $K(x, t)$ и $v(x)$ – заданные функции.

Уравнение (9.21) называется интегральным уравнением ***Фредгольма второго рода***.

Уравнение с переменным верхним пределом в интеграле

$$u(x) - \lambda \int_a^x K(x,t)u(t)dt = v(x),\ x \in [a, b] - \qquad (9.22)$$

интегральным уравнением ***Вольтерра второго рода***.

Определение 9.13. Оператор $\sum_{[a,x]} : U \to V$, определяемый формулой

$$\sum_{[a,x]} w(x) = \int_a^x w(t)dt, \qquad (9.23)$$

называется оператором ***интегрирования*** функции $w(x)$.

Теорема 9.3. Оператор интегрирования функций (9.23) линейный.

Из свойств римановского интеграла следует, что для любых чисел α, β из R и всех интегрируемых на отрезке $[a, b]$ функций u, v справедливо

$$\sum_{[a,x]} (\alpha u(x) + \beta v(x)) = \alpha \int_a^x u(t)dt + \beta \int_a^x v(t)dt.$$

Утверждение 9.1. Теорема 10.3 справедлива и для интеграла Лебега суммируемых функций из пространств L_X^1 и L_X^2.

Определение 9.14. Суммарный оператор Σ_h аппроксимирует на шаблоне $\{\underline{x}_{i-l}, \underline{x}_{i+r}\}$ интегральный оператор с $k^{\text{-ым}}$ порядком точности по h, если

$$\Sigma_h\{u_{i-l}, \ldots, u_{i+r}\} - \sum_{[x_{i-l}, x_{i+r}]} u(x) = O(h^k). \qquad (9.24)$$

При введении суммарного интеграла в разделе 9.1 использовалась аппроксимация определенного интеграла методами левых и правых прямоугольников. Рассмотрим понятие «центральный суммарный интеграл» и найдем оценку погрешности по $h = (b - a)/N$ аппроксимации

$$\int_a^b u(x)dx \approx \frac{h}{2}u_0 + h\sum_{i=1}^{N-1} u_i + \frac{h}{2}u_N. \qquad (9.25)$$

Суммарный интеграл, вычисленный по формуле (9.25), можно интерпретировать как интеграл Лебега простой функции $u_i = u(x_i)$, которая принимает постоянные значения в $\frac{h}{2}$-окрестности точек $\{x_i, i \in N_0\}$ при неограниченном увеличении параметра N.

Для интегрируемых по Риману функций справедливо

$$\int_a^b u(x)dx = \sum_{i=1}^N \int_{x_{i-1}}^{x_i} u(x)dx = h\sum_{i=1}^N \bar{u}_i, \qquad (9.26)$$

где $\bar{u}_i$ – средние значения $u(x)$ на $i^{\text{-ом}}$ интервале отрезка $[a, b]$.

Формула трапеций (9.25) – результат приближенного вычисления в (9.26) среднего значения $\bar{u}_i$ при помощи линейного интерполирования подынтегральной сеточной функции на шаблоне (9.24) с $l = 1$ и $r = 0$.

На интервале $\{x_{i-1}, x_i\}$, $i = 1, \ldots, N$ это приближение имеет вид

$$\int_{x_{i-1}}^{x_i} u(x)dx \approx \frac{h}{2}(u_{i-1} + u_i) \text{ с точностью } \Theta_i \le \frac{h^3}{12} \max_{x_{i-1} \le x \le x_i} |u''(x)|.$$

Доказательство формулы следует из соотношения

$$\bar{u}_i \approx \frac{1}{h}\int_{x_{i-1}}^{x_i} (\alpha x + \beta)dx = \left(\frac{\alpha x^2}{2h} + \frac{\beta x}{h}\right)\Bigg|_{x_{i-1}}^{x_i} = \frac{\alpha(x_{i-1} + x_i)}{2} + \beta = \frac{\alpha x_{i-1} + \beta + \alpha x_i + \beta}{2}.$$

Окончательная оценка погрешности формулы трапеций на всем отрезке интегрирования $[a, b]$ удовлетворяет неравенству [11, с. 144]

$$\Theta \le \frac{h^2(b-a)}{12} \max_{a \le x \le b} |u''(x)|. \qquad (9.27)$$

Теорема 9.4. Четвертый порядок точности аппроксимации по h оператора $\sum_{[a,b]}$ достигается квадратичным интерполированием подынтегральной сеточной функции на $n \in N$ интервалах $\{x_{i-1}, x_{i+1}\}$, $i = 1, \ldots, 2n-1$ с использованием их центральных точек $x_i = (x_{i+1} + x_{i-1})/2$, то есть при интегрировании парабол, проходящих через точки

$$\{(x_k, u(x_k)), k = i-1, i, i+1\}, i = 1, 3, \ldots, 2n-1.$$

Доказательство теоремы основано на разбиении отрезка $[a, b]$ на $N = 2n$ интервалов, где n – число шаблонов длины $2h$ с $l = r = 1$. Основная идея доказательства состоит в том, что

$$\bar{u}_i \approx \frac{1}{h}\int_{x_{i-1}}^{x_{i+1}} (\alpha x^2 + \beta x + \gamma)dx = \frac{u_{i+1} + 4u_i + u_{i-1}}{3}.$$

Окончательная формула данной аппроксимации – *формула Симпсона*

$$\int_a^b u(x)dx \approx \frac{h}{3}\left(u_0 + u_N + 2\sum_{k=2}^{2n-2} u_k + 4\sum_{k=1}^{2n-1} u_k\right), n > 1. \qquad (9.28)$$

Формула (9.28), погрешность которой оценивается как [11, с. 146]

$$\Theta \le \frac{h^4(b-a)}{180} \max_{a \le x \le b} |u^{IV}(x)| \text{ с } h = \frac{b-a}{N},$$

имеет место только для четных N.

Суммарный интеграл функции в этом случае есть лебеговский интеграл простой функции $\{{}^{N+1}u_k,\ k = 0, 1, \ldots, N\}$ (поточечно сходящейся к $u(x)$) с обычной мерой на отрезке $[a, b]$, причем

$$\mu\{x : u(x) = u_k\} = \begin{cases} h/3, k = 0, N; \\ 4h/3, k = 1,3, \ldots, N-1; \\ 2h/3, k = 2,4, \ldots, N-2. \end{cases}$$

Эта формула предназначена для вычисления приближенного значения интеграла сеточной функции ${}^{N+1}U$ с четным $N = 2n$ (9.28) и фиксированным шагом. Однако в полиномиальных методах подынтегральная функция почти всегда определена на всем отрезке $[a, b]$. Докажем теорему Симпсона применительно к вычислению интеграла достаточно гладкой функции $u(x)$ с заданной точностью и регулируемым шагом.

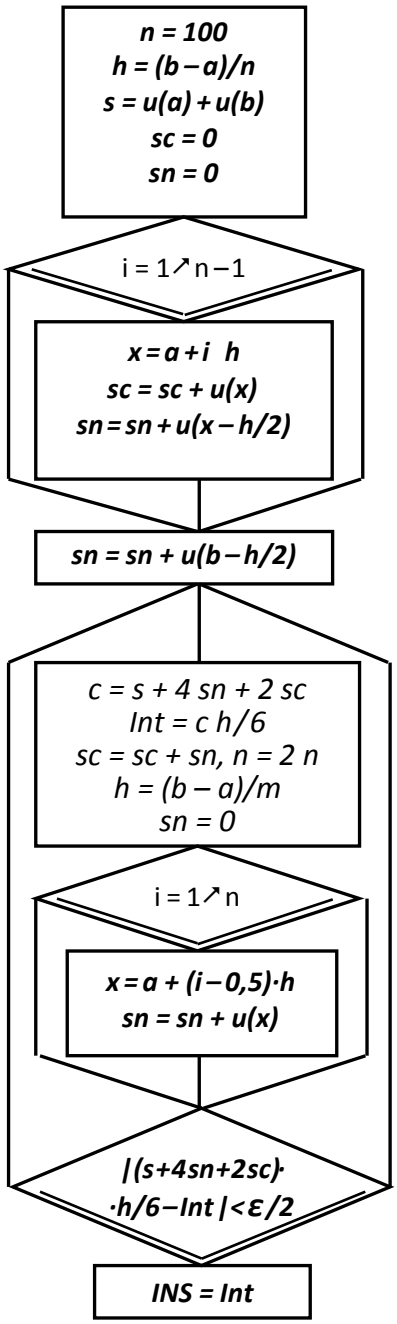


Рисунок 33

Пусть $\{x_i = a + ih,\ i = 0, \ldots, n\}$ – равномерная сетка на отрезке $[a, b]$ с шагом

$$h = (b – a)/n.$$

Проинтерполируем подынтегральную функцию $u(x)$ на элементарном интервале разбиения

$$[x_{i-1}, x_i],\ i = 1, \ldots, n$$

(с центральной точкой $x_{i-1/2}$) многочленом второй степени. Затем по теореме «о среднем» найдем среднее значение $\overline{u_i}$ функции $u(x) \in C^4_{[a, b]}$ при ее квадратичной интервальной аппроксимации:

$$\overline{u_i} = \frac{1}{h}\int_{x_{i-1}}^{x_i}(\alpha x^2 + \beta x + \gamma)dx = \left(\frac{\alpha x^3}{3h} + \frac{\beta x^2}{2h} + \frac{\gamma x}{h}\right)\Bigg|_{x_{i-1}}^{x_i} =$$

$$= \frac{\alpha}{3}(x_i^2 + x_i x_{i-1} + x_{i-1}^2) + \frac{\beta}{2}(x_i + x_{i-1}) + \gamma =$$

$$= \alpha x_{i-1/2}^2 + \beta x_{i-1/2} + \gamma + \frac{\alpha}{3}\left(\frac{h}{2}\right)^2 = u_{i-1/2} + \frac{u''_{i-1/2}}{6}\frac{h^2}{4} =$$

$$= u_{i-1/2} + \frac{u_i - 2u_{i-1/2} + u_{i-1}}{6(h/2)^2}\frac{h^2}{4} = \frac{u_i + 4u_{i-1/2} + u_{i-1}}{6}.$$

Вычислим сумму $I_s = \sum_{i=1}^{n} \bar{u}_i h$, $n \geq 2$, которая при квадратичной аппроксимации $u(x)$ на $[x_{i-1}, x_i]$ приближает значение интеграла Римана функции $u(x)$ с четвертым порядком точности по h

$$I_h = \frac{h}{6}\left(u_0 + u_n + 4 \sum_{i=1}^{n} u\left(x_i - \frac{h}{2}\right) + 2 \sum_{i=1}^{n-1} u(x_i) \right). \qquad (9.29)$$

В связи с тем, что алгоритм численного интегрирования функции по формуле (9.29) достаточно прост, разработаем алгоритм вычисления приближенного значения определенного интеграла функции $u(x) \in C^4_{[a,\,b]}$ методом Симпсона с заданной точностью $\varepsilon > \varepsilon_{ИС}$. Шаг интегрирования h, при котором сумма I_h будет отличаться от точного значения I_s определенного интеграла меньше чем на ε, находится из неравенства

$$\frac{h^4 (b-a)}{2880} \max_{a \leq x \leq b} |u^{IV}(x)| < \varepsilon. \qquad (9.30)$$

Чтобы найти h численно, будем в два раза увеличивать n (начиная с $n = 100$). Последовательность приближений I_s обозначим

$$I^k \equiv I_{h/2^k}, \text{ а } |I^{k+1} - I^k| \equiv \Delta^k,\ k = 0, 1, \ldots .$$

Если Δ^k с $k = i$ станет меньше $\varepsilon/2$ и при этом Δ^k, $k = i, i+1, \ldots$ мажорируется геометрической прогрессией с $q < 1/2$, то справедлива оценка погрешности $\|I_s - I^i\| < \varepsilon$.

На рисунке 33 показан фрагмент объект-схемы алгоритма, который сначала осуществляет подсчет составляющих I_h (s, sc и sn), а затем в цикле с постусловием по этим значениям определяет I_h и находит компоненты $I_{h/2}$ следующего приближения I_s. Отметим, что правомочность использования в цикле-постусловии алгоритма метода Симпсона при вычислении sc (сумма значений интегрируемой функции в «четных» узлах сетки) присваивания $sc := sc + sn$ вытекает из схемы, по которой внутренние точки сетки на очередном повторении становятся «четными».

9.4 Разностная и суммарная схемы поиска корней функциональных уравнений

Под ***разностной схемой*** понимают систему алгебраических уравнений, аппроксимирующих с помощью шаблона и сеточной функции M переменных

$$^{z}U = \{ u_{i_1 \ldots i_M},\ i_1, \ldots, i_M = 0, \ldots, n \}$$

дифференциальное уравнение и дополнительные условия краевой задачи.

Рассмотрим алгоритм построения разностной схемы решения дифференциальной краевой задачи при $M = 1$ для уравнения

$$u''(x) + u'(x) = 2 + 2x,\ x \in [0; 1] \tag{9.31}$$

с граничными условиями

$$u(0) = 0,\ u(1) = 1. \tag{9.32}$$

Пусть требуется найти значение функции $u(x)$ в точке $x = 1/3$.

Для решения задачи с $M = 1$ разобьем отрезок $[0; 1]$ на три равные части (то есть $n = 3$, $h = 1/3$) и обозначим значения функции $u(x)$ в точках сетки 0, $1/3$, $2/3$ и 1 соответственно u_0, u_1, u_2 и u_3.

Среднее значение левосторонней и правосторонней разностных производных (центральная производная) сеточной функции $\{u_i\}$

$$\frac{u_{i+1} - u_{i-1}}{2h} = \frac{(u_i - u_{i-1})/h + (u_{i+1} - u_i)/h}{2} \tag{9.33}$$

аппроксимирует дифференциальный оператор $u'(x)$ в точке x_i уравнения (9.31) со вторым порядком точности по h.

Утверждение 9.2. Из определения центральной производной следует, что приближения

$$u'(x_i + h/2) \approx (u_{i+1} - u_i)/h \ \text{ и } \ u''(x_i) \approx (u_{i-1} - 2u_i + u_{i+1})/h^2$$

аппроксимируют производные со вторым порядком точности по h.

Подставив в (9.31) приближенные значения $u''(x_i)$ и $u'(x_i)$ с i, равным 1 и 2, получим два уравнения разностной схемы:

$$\begin{cases} (-2u_1 + u_2)/h^2 + u_2/(2h) = 8/3; \\ (u_1 - 2u_2 + 1)/h^2 + (1 - u_1)/(2h) = 10/3. \end{cases} \tag{9.34}$$

Решение разностной задачи (системы уравнений 9.34) в точках $x_1 = 1/3$ и $x_2 = 2/3$ равно $u_1 = 1/9$ и $u_2 = 4/9$ соответственно. В данном случае оно совпадает с точным решением (9.31) $u^*(x) = x^2$ при $x = 1/3$ и $x = 2/3$.

Для визуального восприятия сеточной функции ${}^z U = \{u_{ij}\}$, соответствующей непрерывной по двум переменным функции $u(x, y)$, используется дискретное изображение (рисунок 31). Чтобы подчеркнуть непрерывность аппроксимируемой функции, точки ${}^z U$ обычно соединяют треугольниками.

При $M = 1$ сеточный аналог ${}^{z}U \in R^{z}$ функции $u(x) \in C_{[a,\,b]}$ часто представляют в виде ломаной линии (кусочно-линейная интерполяция), проходящей через точки $\{(x_i, u_i), i = 0, \ldots, n\}$ дискретной пары $({}^{z}X, u({}^{z}X))$ (метод Ритца с аппроксимирующими «функциями-колпаками»).

Кусочно-непрерывное изображение сеточной функции ${}^{z}U$ (см. рисунки 32 и 34) отвечает наглядному восприятию функции $u(x, y)$, *суммируемой* по Лебегу. Дискретная аппроксимация в R^{z^M}, $z = n + 1$ суммируемых функций реализуется с помощью осреднения заданной функции $u(x)$ на каждом сегменте множества ${}^{n}\underline{\Omega}$, то есть значение $u_{i_1 \ldots i_M}$ находится как частное значений лебеговского интеграла $u(x)$ на сегменте $\underline{x}_{i_1 \ldots i_M}$ и меры $\mu(\underline{x}_{i_1 \ldots i_M})$.

Теорема 9.5. Пусть X – конечный интервал в R и $u(x)$ – ограниченная функция на множестве X. Тогда если $u(x)$ интегрируема по Риману (почти всюду непрерывна), то она интегрируема и по Лебегу, причем значения обоих интегралов совпадают.

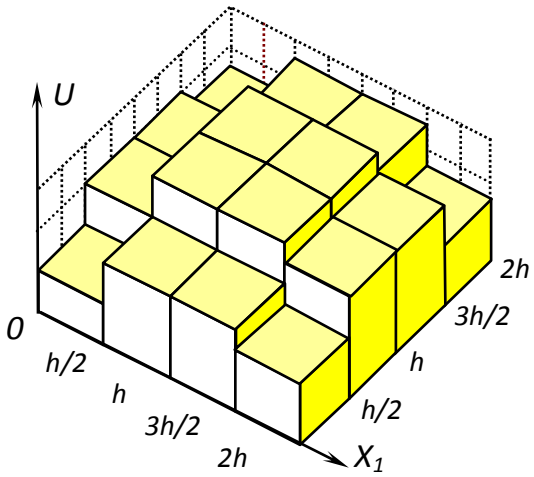


Рисунок 34

Утверждение теоремы справедливо для всех интегрируемых по Риману функций M переменных в ограниченной выпуклой области X пространства R^{M}.

Под ***суммарной схемой*** будем подразумевать совокупность алгебраических уравнений, аппроксимирующих при помощи шаблона и сеточной функции ${}^{z}U = \{u_{i_1 \ldots i_M}, i_1, i_2, \ldots, i_M = 0, 1, \ldots, n\}$ интегральное уравнение и дополнительные условия (если они заданы в интегральной краевой задаче).

Рассмотрим алгоритм решения интегрального уравнения с помощью суммарной схемы второго порядка точности по шагу сетки. Пусть функция $u(x)$ определена на отрезке $[0; 1]$ и является решением уравнения

$$u(x) = x \int_0^1 u(t)\,dt + \frac{x}{2}. \qquad (9.35)$$

Требуется найти значение функции $u(x)$ в точке $x = 1/2$.

Для решения данной задачи разобьем отрезок $[0; 1]$ на две равные части ($n = 2$, $h = 1/2$) и обозначим значения искомой функции $u(x)$ в точках сетки 0, $1/2$ и 1 соответственно u_0, u_1 и u_2. Будем аппроксимировать интегральный оператор в уравнении (9.35) средним значением лево и правостороннего суммарных интегралов (формула трапеций – 9.25).

Получим

$$\int_0^1 u(t)\,dt \approx h\left(\frac{u_0}{2} + u_1 + \frac{u_2}{2}\right).$$

Тогда суммарная схема примет вид системы трех уравнений

$$u_i = \frac{x_i}{2}\left(\frac{u_0}{2} + u_1 + \frac{u_2}{2}\right) + \frac{x_i}{2},\ i = 0, 1, 2. \qquad (9.36)$$

Решение $u_1 = 1/2$ системы (9.36), соответствующее $x_1 = 1/2$, совпадает со значением точного решения $u^*(x) = x$ задачи (9.35) при $x = 1/2$.

Здесь и далее для удобства изложения идеи метода в качестве примеров рассматриваются простейшие уравнения и задачи с очевидными решениями. Однако из этого ни в коем случае не следует, что данными методами нельзя решить более сложные задачи.

Дискретную аппроксимацию дифференциального и интегрального операторов будем использовать при решении краевых задач в предбазисе пространства $l_{[0;\,1]}$. Невысокая точность такой аппроксимации связана с тем, что значения сеточной функции ${}^{N+1}U$ лишь приближенно равны значениям функции $u(x)$ в точках сетки, а суммарно-разностные операторы имеют низкий порядок аппроксимации интегро-дифференциального оператора.

Все вышесказанное не позволяет надеяться на локализацию точного решения по сеточному приближению ${}^{N+1}U$. Однако на пути поиска приближенного значения корня уравнения (*) в полиномиальном виде этот шаг необходим, так как по данной сеточной функции можно построить многочлен ${}^{n}u(x),\ n \le N$ наилучшего приближения по норме пространства U, с которого начнут работать *функциональные* итерационные методы решения.

9.5 Таблица значений и многочлен наилучшего приближения сеточной функции

Как видно из приведенных примеров, при определенных условиях корень $u^*(x)$ операторного уравнения (*) можно получить в виде последовательности *точных* значений искомой функции в узлах сетки. Выясним, какими должны быть эти условия. Не нарушая общности рассуждений, будем считать, что все переменные $x^1, \ldots, x^M$ функции $u(x)$ по умолчанию изменяются в пределах от 0 до 1, а количество сегментов разбиения отрезка $[0; 1]$ по каждой координате равно $n + 1$ ($0 \le i_k \le n$, $k = 1, \ldots, M$).

Определение 9.15. ***Таблицей значений*** сеточной функции

$$^{Z}U = \{ u_{i_1 \dots i_M} \}$$

назовем M-мерную матрицу ^{Z}U размерности

$$\underbrace{(n+1)\times(n+1)\times\dots\times(n+1)}_{M \text{ сомножителей}},$$

элементами которой являются значения этой функции в соответствующих узлах сетки M-мерного базового множества X.

Для решения краевой задачи (*) матрицу преобразуем в вектор

$$^{Z}U = \{u(x_l),\ l = 0,\ \dots,\ Z-1\},\ Z = (n+1)^M,$$

в который входят все известные (граничные и внутренние) значения $u^*(x)$.

В одномерном случае таблица $^{z}U, z = n+1$ – это вектор пространства R^{z}, имеющий $n+1$ координату. Можно ли по таблице значений функции $u(x)$, принадлежащей $C_{[a,\,b]}$, $L^1_{[a,\,b]}$ или $L^2_{[a,\,b]}$, восстановить ее явный вид?

Ответ утвердительный лишь для многочленов степени не выше n и соответствующих им классов эквивалентных суммируемых функций.

Определение 9.16. Говорят, что некоторое соотношение выполнено ***почти всюду***, если оно справедливо на всем множестве X, кроме, быть может, точек, образующих множество меры нуль.

Таким образом, две суммируемые функции являются *эквивалентными* (определение 3.12), если они совпадают почти всюду на множестве X.

В связи с этим при решении задач, приводимых к виду (*), мы будем придерживаться следующей очередности изучения пространств на предмет принадлежности им корней операторного уравнения:

1. L-пространство $l_{[a,\,b]}$ простых ограниченных на отрезке $[a, b]$ функций, принимающих не более чем счетное число значений;
2. L-пространство $L^1_{[a,\,b]}$ всех суммируемых функций;
3. H-пространство $L^2_{[a,\,b]}$ функций с суммируемым квадратом;
4. L-пространство $C^n_{[a,\,b]}$ функций, имеющих непрерывные производные до $n^{\text{-го}}$ порядка включительно (кроме $C^0_{[a,\,b]} \equiv C_{[a,\,b]}$), с нормой

$$\| u \|_n = \max_{\substack{0 \le k \le n \\ a \le x \le b}} |u^{(k)}(x)|, \text{ где } n = 0,\ 1,\ \dots,\ N. \qquad (9.37)$$

В пространстве $l_{[a,\,b]}$ сегментарной сходимости есть возможность найти лишь нулевое приближение для функционального метода поиска корня (*) в виде многочлена.

Мера сегмента $\underline{x}_i$, $i = 0, 1, \ldots, n$ отрезка $[a, b]$ не равна 0, поэтому существенное различие значений *точного* и *сеточного* решений даже в одной точке сетки означает, что *они* не соответствуют друг другу.

Тем не менее, если корнем функционального линейного уравнения с постоянными коэффициентами является многочлен степени m, то после аппроксимации отображения F суммарно-разностным оператором (раздел 9.4) порядка $k \geq m$ значения сеточного ${}^{z}u$, $z > k$ и точного $u^*(x)$ решений могут совпасть в узлах x_i. Однако эта схема неприемлема при $m \to \infty$ (так как $z > m$).

Поиск корней (*) в бесконечномерных пространствах 2, 3, 4 с $m \leq \infty$ должен быть основан на корректной аппроксимации оператора (отвечающей точному определению F на элементе из U) и фундаментальной сходимости приближений

$$\{ {}^{n}u(x),\ n \in N \}$$

к функции $u^*(x)$. Оценка качества приближения δ в (9.8) зависит от n. Однако, увеличивая n, мы будем руководствоваться принципом решения (*) «с точностью до ***локализации*** корня», предусматривающим минимальный рост параметра дискретизации.

То есть итерационная последовательность $\{ {}^{n}u(x),\ n \in N \}$, состоящая при каждом n из наилучших приближений относительно минимума функционала $\|F({}^{n}u) - v\|$, генерируется до тех пор, пока не будет найдено ${}^{N}u(x)$, в δ-окрестности которого *существует изолированный корень* поставленной задачи (*).

Алгоритм восстановления полиномиального приближения ${}^{n}u(x) \in {}^{n}U$ корня (*) по таблице значений сеточной функции основан на определении нормы функции $u(x)$ в B-пространстве решений U. Естественно, проекцию элемента $u(x) \in U$ в конечномерное подпространство с системой линейно независимых базисных функций (3.24) можно найти, только располагая возможностью вычисления значений $u(x)$ в любой точке x базового множества X (см. раздел 3.5).

При решении функциональных уравнений количество узлов, используемых для аппроксимации корня, ограничено и важно выбрать последовательность точек сетки так, чтобы при увеличении параметра дискретизации соответствующий ей многочлен наилучшего приближения сеточной функции сходился к корню уравнения (*) по норме U.

В Γ-пространствах $C^{k}_{[a, b]}$, как впрочем, и в $L^{1}_{[a, b]}$ и $L^{2}_{[a, b]}$, к многочлену наилучшего приближения осредненной сеточной функции (раздел 9.4) с определенной точностью можно приблизиться методами интерполирования [12].

В любом случае коэффициенты степенного или тригонометрического многочлена наилучшего приближения (МНП) находятся из линейного операторного уравнения, в правую часть которого входят сеточные значения восстанавливаемой функции.

Замена входящих в уравнение (*) функций МНП при определенных условиях ε-аппроксимирует отображение $F : U \to V$ компактным оператором $F : U^p \to V^p$, где U^p и V^p – состоящие из элементов $C^{\infty}_{[a,\,b]}$ предкомпактные множества в U и V соответственно. В этом случае появляется возможность найти корень (*) в виде многочлена ${}^n u(x)$ с заданной точностью по норме каждого из изучаемых в пособии функциональных пространств.

Определение 9.17. ***Невязкой*** уравнения (*) на заданном приближении ${}^n u(x)$ называется элемент пространства V, вычисляемый по формуле

$$\Phi({}^n u(x)) = F({}^n u(x)) - v(x),\ x \in X. \qquad (9.38)$$

Расстояние между образами элементов ${}^n u(x)$ и $u^*(x)$ из U

$$\varepsilon(n) \equiv \rho(F({}^n u(x)),\ F(u^*(x))) = \rho(F({}^n u(x)),\ v(x)) - \qquad (9.39)$$

вторая оценка качества приближения ${}^n u(x)$. Вычисление невязки на элементе ${}^n u(x)$ с заданной точностью ε и δ-аппроксимация $\Phi(x) \in V$ многочлена наилучшего приближения по норме – необходимые условия генерации итерационной последовательности $\{{}^n u(x),\ n \in N\}$ приближенний корня (*), слабо сходящейся к $u^*(x)$. Это *стратегия* решения уравнения (*).

Несколько слов об общепринятой *тактике* поиска корней задачи (*). Пусть требуется найти функцию $u(x) \in U,\ x \in X$, удовлетворяющую уравнению задачи и некоторому дополнительному условию

$$G(u(x)) = 0, \text{ то есть } u(x) \in U_Д \subset U. \qquad (9.40)$$

Во-первых, каждый элемент последовательности приближенных решений уравнения (*) должен удовлетворять (9.40). Следовательно, в итерационном процессе могут использоваться только функции ${}^n u(x) \in U_Д$. *Во-вторых*, $F({}^n u(x))$ – образ приближения должен принадлежать V.

Проанализируем, как находится решение уравнения (9.31) с граничными условиями (9.32) из раздела 9.4. Здесь F – дифференциальный оператор, поэтому искомая функция $u(x)$ должна иметь первую и вторую производные по $x \in X = [0;\ 1]$, причем

$$\{u'(x),\ u''(x)\} \subset V,\ u(0) = 0 \text{ и } u(1) = 1.$$

Дискретное решение разностной задачи с $h = 1/3$ ($n = 3$) имеет вид ${}^4u = \{0; 1/9; 4/9; 1\}$. Дважды дифференцируемое решение задачи (9.31) определим с помощью интерполяционного многочлена Лагранжа ${}^3u(x) \in U_Д \subset C^2_{[0;\,1]}$, построенного по четырем значениям сеточного приближения ${}^4u$.

Коэффициенты полинома ${}^3u(x)$ в базисе (9.41) равны $\{0; 0; 1; 0\}$. Отсюда следует, что ${}^3u(x) = x^2$. Значит, сразу получено точное решение (9.31), которое можно найти и с $n = 2$. Однако описанный алгоритм неприемлем, если $u(x)$ не многочлен степени $m \le 3$.

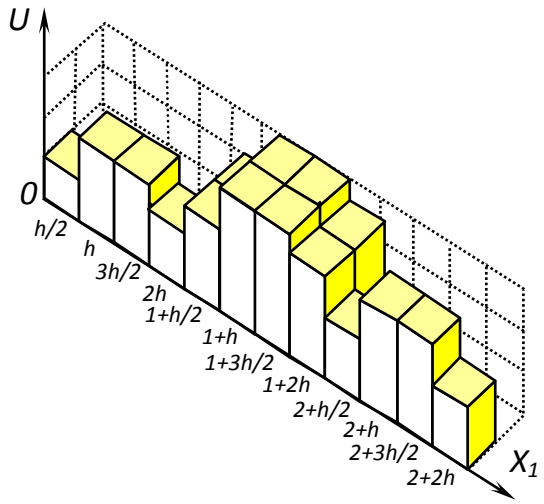


Рисунок 35

Отметим, что интерполяционный полином Лагранжа, найденный по $(n+1)^{\text{-му}}$ значению функции $u(x) \in C^n_{[a,\,b]}$ на сетке Чебышева [12] в базисе P^n

$$1, x, x^2, \ldots, x^n, \qquad (9.41)$$

аппроксимирует функцию $u(x)$ не хуже, чем многочлен Тейлора $n^{\text{-го}}$ порядка.

Условия тактики решения специально подобраны так, чтобы, с одной стороны, обе производные искомой функции принадлежали V, а с другой, – ограничения на область определения оператора F были минимальными. Этим самым создаются наилучшие шансы для существования сходящейся в $U_Д$ последовательности приближений корня уравнения (*). Другими словами, если пространство образов V есть $L^2_{[0;\,1]}$, то на отрезке $[0; 1]$ функция $u(x)$ должна быть дважды дифференцируемой, а при $V = C_{[0;\,1]}$ – дважды непрерывно дифференцируемой и т.д.

И в заключение о *практике* решения (*). Структура итерационного процесса не должна зависеть от M – размерности измеримого мерой μ множества X. Для этого расположим узлы (сегменты) сетки области $X \subset R^M$ и соответствующие им цилиндры (ступеньки) сеточной функции вдоль оси OX_1 в виде цепочки из Z звеньев (см. рисунки 33 и 35) с сохранением значений ${}^Z U$ в виде вектора с Z координатами.

В этом случае невязка $\Phi({}^n u(x))$ останется функцией, измеримой мерой μ на множестве сегментов X, и будет зависеть от одной переменной x, а таблица значений ${}^Z U(\Phi({}^n u))$ станет одноиндексной последовательностью из R^Z. Тогда линейные операторы, участвующие в функциональных преобразованиях, будут аппроксимированы квадратными матрицами (раздел 2.4).

В частности матрица Грама после принятых соглашений, связанных с реорганизацией M-мерной таблицы значений сеточной функции на единичном кубе $X \subset R^M$ в одномерный массив, станет квадратной порядка

$$Z = (n + 1)^M.$$

Также существенно, чтобы форма представления приближенного решения ${}^n u(x)$ уравнения (*) была удобна для качественной и количественной оценки приближения. Например, при аппроксимации действительных значений корней алгебраических уравнений используют рациональные числа из множества Q_n (n десятичных знаков), обеспечивающие погрешность округления $\varepsilon = 10^{-n}$.

Аналогично полиномы пространства P^n (T^n) могут являться приближенными с точностью $\acute{\varepsilon}(n)$ решениями задачи (*), если пространство U содержит всюду плотное множество P степенных (T тригонометрических) многочленов с рациональными коэффициентами.

Разумеется, не все задачи могут быть решены в предбазисах (3.19) или (3.32). В пространстве $L^2_{(-\infty,\infty)}$, к примеру, базисом является система функций

$$\{\varphi_i(x) = {}^i h(x) \cdot \exp(-x^2/2),\ i = 0, 1, \ldots, n, \ldots\}, \qquad (9.42)$$

где ${}^i h(x)$ – многочлены Эрмита, в $L^2_{(0,\infty)}$ – функции Лагерра и т.д.

9.6 Операторы проектирования функций пространства решений в множества R^z, P^n и T^n

Пусть $M = 1$ ($z = n + 1$), $u(x)$ – произвольный элемент H-пространства U и ${}^n U$ – некоторое $(n + 1)$-мерное подпространство U с базисом

$$\{\varphi_i(x),\ i = 0, 1, \ldots, n\}. \qquad (9.43)$$

Покажем, что элемент u можно (и притом единственным образом) представить в виде суммы

$$u(x) = {}^n u(x) + {}_n u(x), \qquad (9.44)$$

где ${}^n u \in {}^n U$ и ${}_n u \in {}_n U$ (${}^n U \cup {}_n U = U$, ${}^n U \cap {}_n U = \varnothing$).

Наличие элемента ${}^n u \in {}^n U$ гарантируется существованием многочлена наилучшего приближения. Докажем, что $(u - {}^n u) \in {}_n U$.

Рассмотрим случай, когда базис $\{\varphi_i(x), i \in N_0\}$ бесконечномерного гильбертова пространства U ортогональный. Тогда для каждого ${}^n v \in {}^n U$ и всех действительных $\alpha > 0$ линейная комбинация $({}^n u + \alpha\, {}^n v) \in {}^n U$ приближает u не лучше, чем многочлен наилучшего приближения ${}^n u$.

Таким образом, справедливо неравенство

$$\|u - {}^n u\|^2 \le \|u - {}^n u - \alpha\, {}^n v\|^2 = \|u - {}^n u\|^2 - 2\alpha \langle u - {}^n u, {}^n v\rangle + \alpha^2 \|{}^n v\|^2. \quad (9.45)$$

Отсюда следует, что

$$2\alpha \langle u - {}^n u, {}^n v\rangle \le \alpha^2 \|{}^n v\|^2. \quad (9.46)$$

Разделив обе части неравенства на α и устремив α к 0, получим

$$2 \langle u - {}^n u, {}^n v\rangle \le 0. \quad (9.47)$$

Повторение тех же рассуждений для $\alpha < 0$ показывает, что

$$2 \langle u - {}^n u, {}^n v\rangle \ge 0. \quad (9.48)$$

Значит, $\langle u - {}^n u, {}^n v\rangle = 0$ для всех ${}^n v \in {}^n U$, то есть $(u - {}^n u) \in {}_n U$.

Пусть сейчас система элементов $\{\varphi_i(x), i \in N_0\}$ просто линейно независима. В этом случае определитель Грама в уравнении (3.29) отличен от нуля. Следовательно, коэффициенты разложения ${}^n u(x)$ в базисе (9.43) определяются однозначно и любой элемент ${}^n v(x) \in {}^n U$ образует с (9.43) линейно зависимую систему, по которой определитель Грама равен нулю.

Теорема 10.6 . Если система функций (9.43) линейно независима, то для всех $0 < k \le n$ определитель Грама

$$\Gamma_k = \Gamma\{\varphi_i, i = 0, 1, \ldots, k\} > 0. \quad (9.49)$$

В противном случае найдется k, начиная с которого определитель Γ_k обращается в нуль.

Доказательство теоремы и много интересной информации об определителе Грама можно прочитать в монографии Ф.Р. Гантмахера [5].

Так как (9.43) образует с вектором ${}_n u(x) = u(x) - {}^n u(x)$ систему элементов Φ, по которой $\Gamma\{\Phi\} > 0$, то представление (9.44) единственно.

Операторы проектирования $P_T : U \to T^n$ в системе элементов (3.19), $P_C : U \to P^n$ (3.32) и $P_Л : U \to P^n$ (3.36) будем называть операторами ***непрерывного проектирования***, или P_H-операторами.

Если же проектирование осуществляется из пространства U в R^z, где $z = n + 1$, то соответствующий оператор $P_Д$ назовем оператором ***дискретного проектирования***. Таковым является «узловой оператор» $P_y : U \to R^z$

$${}^z U = P_y(u(x)) = \{u_i = u(x)|_{x=x_i}, i = 0, 1, \ldots, n\},$$

рассмотренный в разделе 9.1.

Оператором дискретного проектирования на отрезке ***[a, b]*** является также «оператор осреднения» $P_o : U \to R^z$

$$ {}^zU = P_o(u(x)) = \{u_i = \frac{1}{b_i - a_i} \int_{a_i}^{b_i} u(x)dx\}, \qquad (9.50) $$

где $a_i = max\{a, x_i - h/2\}$, $b_i = min\{b, x_i + h/2\}$, $i = 0, 1, \ldots, n$.

При дискретном проектировании функций из пространства $U = L^p_{[a, b]}$ в формуле (9.50) используется лебеговский интеграл с обычной мерой.

Дискретное решение сеточной задачи применяется для полиномиального

$$ P_П : R^z \to P^n \text{ (или } T^n) $$

и кусочно-непрерывного

$$ P_C : R^z \to {}^zl $$

представления дифференцируемых или интегрируемых на множестве X корней (*). Дискретному проектированию функций и разностным методам решения задач математической физики посвящены исследования И.С. Березина и Н.П. Жидкова. [3].

Без дискретного представления решения функционального уравнения (*) невозможно определить МНП корня по норме пространства U. Формула задания нормы в пространстве $U_{[a, b]}$ позволяет с помощью вычисления достаточного числа значений функции в точках отрезка ***[a, b]*** приближенно найти расстояние от функции до проективного многочлена ${}^nu(x) \in U$. Метод восстановления функции в виде МНП по ее сеточному приближению зависит от формулы задания нормы.

Найдем полиномиальное приближение функции $u(x)$, значения которой известны в $c+1$ точке сетки отрезка ***[a, b]***. Критерием близости выберем наименьшее отклонение значений функции и многочлена ${}^nu(x)$ в заданных точках ${}^c\Omega_{[a, b]}$. Если потребовать совпадение этих значений во всех узлах сетки, то получим метод интерполяционного приближения $u(x)$ с $n = c$ (раздел 9.1), алгоритм поиска коэффициентов МНП которого по логическому построению согласуется с описанием нормы в $C_{[a, b]}$.

Условие минимизации указанных отклонений во всех точках сетки по E-норме R^{c+1} можно записать в виде минимума функционала Z

$$ \min Z(c_0, c_1, \ldots, c_n) = \sum_{i=0}^{c} \left(\sum_{k=0}^{n} c_k x_i^k - u(x_i) \right)^2, \; x_i \in {}^c\Omega_{[a, b]}, \qquad (9.51) $$

где ${}^zc = (c_0, \ldots, c_n)$ – коэффициенты многочлена n-ой степени наилучшего приближения сеточной функции ${}^{c+1}U$. Данный метод поиска коэффициентов МНП по логической структуре согласуется с описанием нормы в пространстве $L^2_{[a, b]}$ функций с суммируемым квадратом.

Функция многих переменных достигает минимальное (максимальное) значение в точках экстремума, поэтому для вычисления коэффициентов ${}^{z}\boldsymbol{c}$ найдем частные производные по каждому $\boldsymbol{c_j, j = 0, \ldots, n}$ и составим систему

$$\frac{\partial Z(c_0, c_1, \ldots, c_n)}{\partial c_j} \equiv 2 \sum_{i=0}^{c} \left(\sum_{k=0}^{n} c_k \, x_i^k - u_i \right) x_i^j = 0, \; j = 0, \ldots, n,$$

которую после несложных преобразований можно привести к ЛСАУ относительно $(\boldsymbol{n+1})^{\text{-го}}$ неизвестного коэффициента:

$$\sum_{k=0}^{n} c_k \sum_{i=0}^{c} x_i^{k+j} = \sum_{i=0}^{c} x_i^j \, u_i, \; j = 0, \ldots, n. \qquad (9.52)$$

Или в матричном виде

$${}_{c}^{n}W^T \; {}_{c}^{n}W \; {}^{z}C = {}_{c}^{n}W^T \; {}^{c+1}U, \text{ где } {}_{c}^{n}W = \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ & . & . & . & \\ 1 & x_c & x_c^2 & \ldots & x_c^n \end{pmatrix}. \qquad (9.53)$$

Так как $\boldsymbol{d^2 Z(c_0, \ldots, c_n) > 0}$ при $\boldsymbol{dc_j > 0}$, то решение (9.53) аппроксимирует $\boldsymbol{u(x)}$ в точках сетки многочленом ${}^{n}\boldsymbol{u(x)}$ наилучшего приближения по норме $\boldsymbol{R}^{c+1}$ (см. пояснение к приложению 2). В математической статистике этот многочлен устанавливает регрессию зависимой и независимой случайных величин (СВ). Напомним, что в нашем изложении сеточная функция может определяться не только в точках сетки, но и на ее сегментах ${}^{c}\underline{\boldsymbol{\Omega}}_{\boldsymbol{[a, b]}}$.

Глава 10 Методология математического моделирования

10.1 Математическая модель исследуемого предмета. Постановка численного эксперимента

Модель – упрощенное представление явления, процесса или объекта действительности в виде схем и математических формул, описывающих реальные предметы, явления, процессы, изучаемое как их аналог.

Модели выполняют следующие функции:

- познания (дают возможность понять суть изучаемого процесса);
- прогнозирования (определяют перспективы развития процесса);
- принятия решений (позволяют планировать и управлять процессом).

Математической моделью назовем представление объекта действительности в виде совокупности и/или системы уравнений, формул и дополнительных условий, связывающих искомые переменные и функции.

Понятие *математическое моделирование* определим как адекватную замену исследуемого объекта (ИО) его математической моделью (ММ) с последующим изучением этой модели методами функционального и численного анализа в современной интегрированной среде программирования.

Физические явления часто описываются с помощью интегральных и дифференциальных уравнений или краевых задач с дополнительными условиями. Например, математическая модель электромагнитных процессов в сплошной среде представляет собой систему уравнений Максвелла

$$\begin{cases} \nabla E(t,x) + \dfrac{\partial B(t,x)}{\partial t} = 0,\ \nabla B(t,x) = 0; \\ \nabla H(t,x) - \dfrac{\partial D(t,x)}{\partial t} = j^{(\varepsilon)},\ \nabla D(t,x) = \rho_\varepsilon, \end{cases}$$

где t – время, $x \in R^3$, $E(t, x)$ и $H(t, x)$ – напряженности электрического и магнитного полей, $D(t, x) = \varepsilon \varepsilon_0 E$, $B(t, x) = 2\mu \mu_0 H$, $j^{(\varepsilon)}$ – вектор плотности электрического тока, ρ_ε – объемная плотность электрического заряда.

Математический аналог реального физического процесса может содержать не только уравнения с обыкновенными или частными производными искомых функций, но и интегральные и интегро-дифференциальные уравнения, в которые искомые функции входят под знаком интеграла. Характерным примером интегрального уравнения является формула Грина, содержащая объемный потенциал и потенциалы простого и двойного слоя.

Этому уравнению удовлетворяет решение краевой задачи для дифференциального уравнения Пуассона. Иногда математическая формулировка задачи в виде интегрального уравнения оказывается более простой и может быстрее привести к решению, чем соответствующая дифференциальная модель исходной задачи (см., например, [10, с. 527]).

Для *математической физики* характерно исследование физических объектов (процессов, явлений) в трехуровневых системах взаимодействия:

1) изучение свойств исследуемой величины в отдельно взятом элементарно малом объеме пространства;

2) изучение взаимодействия между элементарными объемами всей исследуемой системы;

3) изучение взаимосвязей рассматриваемой системы в целом с другими внутренними и внешними объектами.

Первый уровень соответствует установлению ***уравнений состояния*** среды в элементарном объеме области решения, второй – описанию взаимодействия элементарных объемов на основе ***законов сохранения*** физических субстанций и их переноса в рассматриваемое пространство, третий находит отражение в формулировке ***дополнительных условий***, налагаемых на исследуемый объект внутри и/или на границе области решения задачи.

Модели физических процессов, изучаемых в математической физике, в большинстве случаев предполагают непрерывную зависимость искомой величины u от аргумента $x \in X$. При данной гипотезе математическая модель, описывающая вынужденные гармонические колебания струны под действием внешней силы, имеет вид неоднородного интегрального уравнения Фредгольма второго рода (9.21) относительно непрерывной функции $u(x)$

$$u(x) = \eta(x)\int_a^b K(x,t)\,u(t)\,dt + \varphi(x),\ x \in X = [a, b], \qquad (10.1)$$

где ядро $K(x, t)$, функции $\eta(x)$ и $\varphi(x)$ непрерывны по своим аргументам.

Совокупность точек множества X, используемых для приближенного представления непрерывного распределения величины u из U, называют ***пространственной сеткой***, а определенные окрестности этих точек – ***сегментами*** сетки (аналогично математической дискретизации в разделе 9.1).

Сокращение расстояний между выбранными точками множества X должно приближать дискретное представление к непрерывному распределению искомой величины $u^*(x)$. То есть качество аппроксимации функций из пространства U при $M = 1$ зависит от малости h и расположения z узлов сетки ${}^n\Omega$ с таким же числом сегментов на отрезке $X = [a, b]$.

Счетная σ-алгебра сегментов определяет *меру* на множестве X и *метрику* в пространстве U. Замена счетного базиса меры (раздел 3.2) конечным позволит с точностью $\varepsilon(n)$ оценить поведение функции $u^*(x)$ в пространстве решений U по ее проекции в $(n+1)$-мерные подмножества степенных P^n $(n \in N)$ и тригонометрических T^n $(n = 2m,\ \ m \in N)$ многочленов.

Далее, опираясь на свойства полного метрического (чаще банахова) пространства U, описывается оператор F математической задачи так, чтобы для $\forall u \in U_Д \subset U$ существовал единственный образ $F(u) \in V$, где V – пространство, содержащее заданную функцию $v(x), x \in X$. Затем излагается идея метода решения и осуществляется поиск корней функционального уравнения $F(u) = v(x)$, то есть математическая модель *изучается* на предмет существования дискретного или непрерывного решения в области $Q \subset U_Д$.

Приближенное по норме $\|\ \ \|_U$ решение (*), полученное в виде функционального ряда в предбазисе (3.19) или (3.32) пространства U, будем называть *полиномиальным*. Вычисление полиномиального (всегда принадлежащего функциональному пространству решений) приближения связано с возможностью определения на нем нормы невязки заданного уравнения.

Предъявляемые к математическому моделированию требования создают реальные предпосылки для доказательства адекватности физической задачи и ее математической модели с помощью численного эксперимента (компьютерного аналога исследуемого процесса).

Численный эксперимент (ЧЭ) – это последовательность действий, направленная на создание числового аналога исследуемого объекта в виде математической модели и основанная на законах *математического* моделирования описания физических процессов. Следовательно, целью ЧЭ является доказательство того, что искомое решение математической задачи адекватно решению корректной исходной физической задачи.

Перечислим основные качественные характеристики ЧЭ.

I. ***Описание оператора*** $F: U \to V$ ***математической модели*** с классификацией функциональных метрических пространств U и V.

II. ***Вид представления*** и ***идея*** *численного* (*функционального*) ***метода*** решения операторного уравнения исходной задачи.

III. ***Вычислительный процесс*** (ВП) преобразования входных данных через метод решения в искомый конечный результат.

Таким образом, вычислительный процесс включает в себя:

1) ***итерационный метод***; 2) ***алгоритм***;

3) ***программу***; 4) ***счет***;

5) ***редактирование*** и ***анализ итогов***.

IV. ***Качественный и количественный анализ*** результатов численного эксперимента. Для доказательства адекватности (с заданной точностью ε) математической модели исследуемому объекту необходимо показать, что итоговый результат ${}^{N}u(x)$ ЧЭ удовлетворяет условиям реальной физической задачи с погрешностью приближения $\varepsilon_ф$. При этом операторное уравнение (*) требуется решить методом, гарантирующим существование корня u^* математической задачи в $\varepsilon_м$-окрестности приближения ${}^{N}u(x)$.

Пусть достигнута основная цель эксперимента, то есть доказано, что корни математической и физической задач совпадают с точностью ε

$$\|u_ф(x) - u^*(x)\| \le \|u_ф(x) - {}^{N}u(x)\| + \|{}^{N}u(x) - u^*(x)\| < \varepsilon_ф + \varepsilon_м = \varepsilon. \quad (10.2)$$

Тогда ММ будем считать *ε-адекватной* исследуемому объекту, а приближенное решение ${}^{N}u(x)$ – приближением, *моделирующим* физический процесс с точностью $\varepsilon_ф$. Если приближение ${}^{N}u(x)$ не является моделирующим, то для доказательства адекватности моделей следует внести изменения в одну из качественных характеристик и повторить ЧЭ.

На рисунке 36 изображена схема математического моделирования ИО (стрелками указаны повторы блоков схемы при соответствующем анализе).

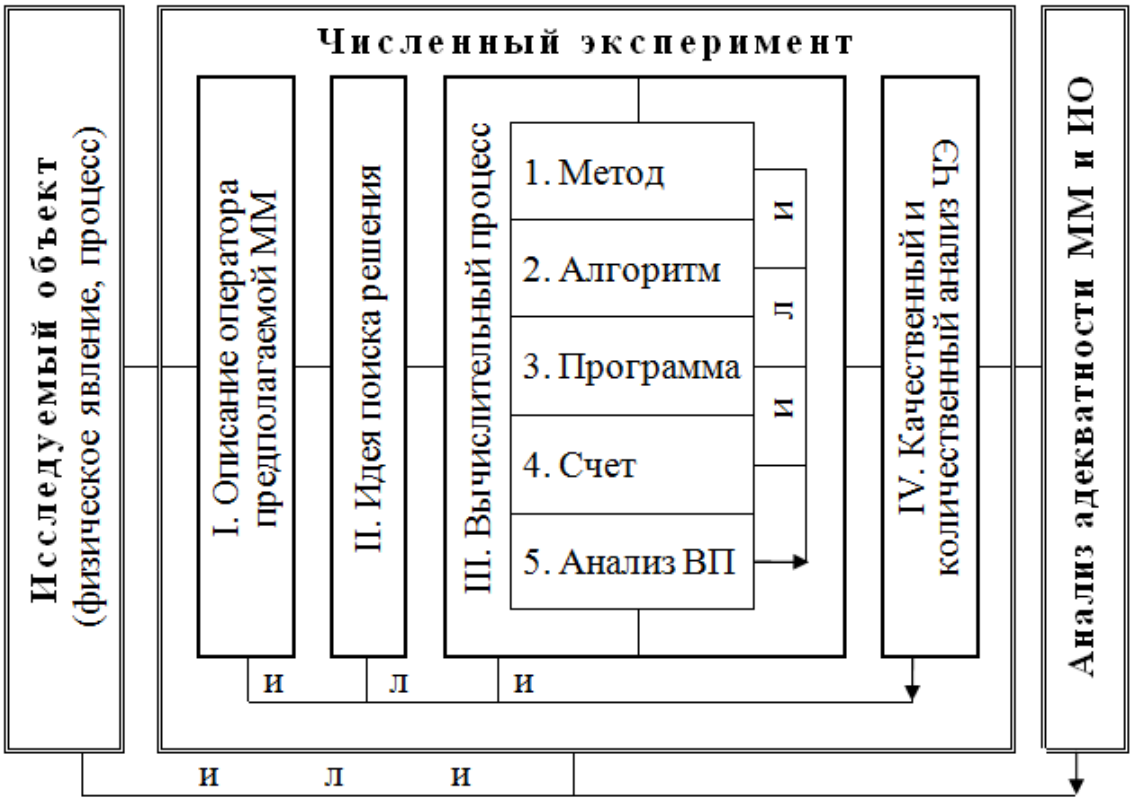


Рисунок 36

Приведем основные количественные характеристики ЧЭ.

1. ***Устойчивость*** вычислительного процесса по входным данным (начальным и дополнительным условиям).

2. ***Условия*** и ***скорость сходимости*** итерационного метода решения операторного уравнения математической задачи.

3. ***Оценка погрешности*** вычислительного процесса.

Количественные характеристики численного эксперимента зависят от качества каждого раздела ЧЭ и в первую очередь от точности ВП.

10.2 Вычислительный процесс. Функциональный метод решения операторного уравнения

Часть численного эксперимента, связывающая постановку задачи (в виде математической модели) и выбор идеи поиска корней (вид представления и метод решения) с получением итоговых численных результатов, назовем ***вычислительным процессом***. Перечислим основные этапы ВП.

1. ***Численный (полиномиальный) итерационный метод***, то есть теоретически обоснованный процесс обращения оператора математической задачи, основанный на результатах предыдущих приближений корня.

2. ***Алгоритм***, то есть совокупность действий, позволяющая приближенное решение операторного уравнения *суммарно-разностными* итерационными методами сколь угодно приблизить к корню математической задачи.

3. ***Программа***, то есть логически упорядоченная последовательность предложений языка (интегрированной среды) программирования в виде условий и команд, направленных на реализацию алгоритма.

4. ***Счет***, то есть программная обработка входных данных исходной задачи с выводом искомых количественных результатов.

5. ***Редактирование*** итогов и ***анализ*** работы всех этапов ВП.

Наиболее важным этапом ВП является выбор итерационного метода решения функционального уравнения. Причем теоретическое обоснование сходимости метода – только необходимое условие разработки алгоритма программы. Поскольку реально ПК может оперировать лишь с конечным числом узлов сетки, важно уметь правильно выбрать их количество.

С одной стороны, чем меньше узлов пространственной сетки, тем хуже дискретная модель описывает исходную задачу, с другой стороны, рост параметра дискретизации ведет к увеличению числа алгебраических уравнений в системе, нарушению устойчивости и ухудшению сходимости вычислительного процесса.

Рассмотрим нелинейное интегральное уравнение Фредгольма:

$$u^{1+\alpha}(x) = \eta(x) \int_{[a,b]} K(x,t)u(t)\,d\mu + \varphi(x), \qquad (10.3)$$

где $K(x, t)$ – непрерывная функция по t, а по x также, как $u(x)$, $\eta(x)$ и $\varphi(x)$, принадлежит $L^2_{[a, b]}$, μ – обычная мера Лебега на отрезке, $\alpha \neq 0$.

Корень $u^*(x)$ уравнения будем приближать элементами ${}^nU(x)$ ($n \in N$ определяет параметр дискретизации $z = n+1$) из всюду плотных в $L^2_{[a, b]}$:

1) множеств простых ограниченных функций $U \in l_{[a, b]}$, принимающих на отрезке $[a, b]$ не более чем счетное число значений;

2) множеств многочленов $U(x) \in P_{[a, b]}$, заданных на отрезке $[a, b]$.

Определим расстояние в H-пространстве $L^2_{[a, b]}$ по формуле:

$$\rho(u, v) = \|u - v\| = \left(\int_{[a,b]} (u(x) - v(x))^2 d\mu \right)^{1/2}. \qquad (10.4)$$

Тогда для вычисления невязки уравнения (10.3) на приближении ${}^nU(x)$

$$\Phi({}^nU(x)) = ({}^nU(x))^{1+\alpha} - \eta(x) \int_a^b K(x,t)\,{}^nU(t)\,dt - \varphi(x), \qquad (10.5)$$

в первом случае используем интерполяционный многочлен, построенный по $n+1$ значению сеточного решения ${}^zU = \{u_i, i = 0, \ldots, n\}$ при аппроксимации интегрального оператора суммарным (9.25), а во втором – прибли-жение ${}^nu(x)$ корня (10.3), сразу полученное в базисе P^n.

В зависимости от способа вычисления корня возникает две идеи его реализации. Первая – поиск *дискретного* решения в виде последовательности значений простой функции (здесь неизвестно $n+1$ значение сеточной функции ${}^zU$), вторая – поиск *непрерывного* решения в виде многочлена (здесь требуется найти $n+1$ коэффициент ${}^nu(x)$ интерполяционного многочлена Лагранжа).

Опишем алгоритм аппроксимации сеточной функции ${}^{m+1}U$ (простой ограниченной функции $u(x)$ из $l_{[a,b]}$) многочленом ${}^nu(x)$ наилучшего приближения по норме $L^2_{[a,b]}$ методом наименьших квадратов.

Коэффициенты МНП

$${}^nu(x) = \sum_{k=0}^{n} c_k x^k$$

вычислим из условия минимизации функционала Z

$$Z(c_0, c_1, \ldots, c_n) = \sum_{i=0}^{m} \left(\sum_{k=0}^{n} c_k x_i^k - u(x_i) \right)^2 \mu(\underline{x}_i),\ x_i \in {}^m\Omega_{[a, b]}. \qquad (10.6)$$

Здесь $\boldsymbol{Z}$ – лебеговский интеграл квадрата разности простой функции $\boldsymbol{u(x)}$ и ее МНП. Функция $\boldsymbol{Z(^{n+1}c)}$ достигает минимум в точке экстремума, для вычисления координат которой приравняем частные производные к $\boldsymbol{0}$

$$\frac{\partial Z(c_0, c_1, \ldots, c_n)}{\partial c_j} \equiv 2\sum_{i=0}^{m}\left(\sum_{k=0}^{n} c_k x_i^k - u_i\right) x_i^j \mu(\underline{x}_i) = 0,\ j = 0, \ldots, n. \quad (10.7)$$

Решение данной линейной системы позволяет сеточную функцию $\boldsymbol{^{m+1}U}$ аппроксимировать многочленом $\boldsymbol{^{n}u(x)}$ наилучшего приближения по норме $\boldsymbol{L^2_{[a,b]}}$. В основу алгоритма поиска коэффициентов МНП с помощью матрицы Грама положен метод наименьших квадратов, описанный в разделе 3.5, где интегральный оператор аппроксимируется суммарным.

Утверждение 5.1. МНП по норме $\boldsymbol{L^2_{[a,b]}}$ при аппроксимации интеграла (10.4) с точностью $\mu = \mu(\underline{x}_i) \equiv (b-a)/(n+1)$ методом прямоугольников и интерполяционный многочлен на равномерной сетке с $h = (b-a)/n$ порядка n, построенные для функции $u(x) \in C_{[a,\,b]}$, совпадают.

Независимо от вида представления решения, идея функционального метода подчинена основной цели ВП. Изучая возможности того или иного метода решения уравнения (*), необходимо указать *достаточные* условия сходимости итерационной последовательности приближений $\{^{n}\boldsymbol{u(x)},\ \boldsymbol{n \geq N}\}$ к точному решению $\boldsymbol{u^*(x)}$ в области локализации $\boldsymbol{Q[^{N}u,\ \acute{\varepsilon}] \subset U}$.

Результаты дискретного и непрерывного (полиномиального) способов решения функционального уравнения (10.3) при $\boldsymbol{\eta(x) \equiv 1}$, $K(x, t) \equiv 1$, $\varphi(x) = e^x - 2(\sqrt{e} - 1)$ и $\alpha = 1$ для различных значений параметра $\boldsymbol{n}$, связанного с дискретизацией множества $\boldsymbol{X = [0; 1]}$, сведем в таблицу.

Таблица 5 – Нормы погрешности приближений $\boldsymbol{\acute{\varepsilon}}$ и невязки $\boldsymbol{\varepsilon}$

$\boldsymbol{n}$	$\boldsymbol{\acute{\varepsilon}}_{\text{дискр}}$	$\boldsymbol{\acute{\varepsilon}}_{\text{дис-инт}}$	$\boldsymbol{\varepsilon}_{\text{дис-инт}}$	$\boldsymbol{\acute{\varepsilon}}_{\text{интерп}}$	$\boldsymbol{\varepsilon}_{\text{интерп}}$	$\boldsymbol{\acute{\varepsilon}}_{\text{дискр2}}$	$\boldsymbol{\acute{\varepsilon}}_{\text{дис-инт2}}$	$\boldsymbol{\varepsilon}_{\text{дис-инт2}}$
1	2	3	4	5	6	7	8	9
2	**0,25**	**0,12**	$\boldsymbol{2{,}0\cdot10^{-1}}$	$\boldsymbol{1{,}9\cdot10^{-3}}$	$\boldsymbol{2{,}9\cdot10^{-3}}$	**0,19**	$\boldsymbol{7{,}0\cdot10^{-3}}$	$\boldsymbol{9{,}0\cdot10^{-3}}$
4	**0,13**	$\boldsymbol{6{,}9\cdot10^{-2}}$	$\boldsymbol{8{,}6\cdot10^{-2}}$	$\boldsymbol{1{,}6\cdot10^{-6}}$	$\boldsymbol{2{,}0\cdot10^{-6}}$	**0,10**	$\boldsymbol{1{,}9\cdot10^{-3}}$	$\boldsymbol{2{,}1\cdot10^{-3}}$
8	**0,07**	$\boldsymbol{3{,}1\cdot10^{-2}}$	$\boldsymbol{4{,}4\cdot10^{-2}}$	$\boldsymbol{2{,}7\cdot10^{-13}}$	$\boldsymbol{3{,}1\cdot10^{-13}}$	**0,05**	$\boldsymbol{4{,}9\cdot10^{-4}}$	$\boldsymbol{4{,}5\cdot10^{-4}}$

Так как корнем тестового примера является бесконечно дифференцируемая функция $\boldsymbol{u^*(x) = \sqrt{e^x}}$, то теоретически оба интерполяционных процесса решения сходятся к ряду Тейлора функции $\boldsymbol{u^*(x)}$ на отрезке $\boldsymbol{[0; 1]}$.

Однако достижение точности $\boldsymbol{\acute{\varepsilon}}$ приближения интерполяционного метода при $\boldsymbol{n = 4}$ с помощью дискретного метода, то есть, используя аппроксимацию интегрального оператора суммарным $\boldsymbol{2^{\text{-го}}}$ порядка точности по $\boldsymbol{h}$ (9.25) и интерполирование сеточного решения (10.3) многочленом $\boldsymbol{^{n}u(x)}$ (см. графы 8, 9 таблицы 5), возможно лишь с параметром $\boldsymbol{n >> 16}$.

Что касается чисто сеточного решения ${}^{z}U = \{u_i, i = 0, \ldots, n\}$, то для достижения на этом приближении точности $\acute{\varepsilon} = 1{,}6 \cdot 10^{-6}$ потребуется взять число интервалов $n >> 10^3$ (графа 7). Таким образом, оценка $\|{}^{n}U(x) - u^*(x)\|$ зависит от вида представления решения при разных значениях параметра дискретизации $z = n + 1$.

Все этапы ВП имеют свою погрешность. Назовем их соответственно:

- ***погрешность метода*** (возникает при переходе от математической модели к численному или полиномиальному методу решения);
- ***погрешность дискретизации*** (является следствием использования конечного числа элементов базиса или предбазиса вместо бесконечного);
- ***погрешность интегрированной среды*** (зависит от качества размещения, хранения и использования информации в ИСП);
- ***погрешность вычислений*** (связана с техническими возможностями процессора и других арифметических устройств компьютера).
- ***погрешность редактирования результатов счета*** (зависит от носителя информации при передаче и представлении итогов).

При неудовлетворительных результатах анализа работы корректируется один или несколько этапов ВП.

10.3 Логический анализ итогов численного эксперимента. Статистические методы исследования

Результат численного эксперимента должен соответствовать исходной задаче с заданной допустимой погрешностью приближения (10.2). Поэтому для поиска корней краевых задач желательно использовать методы решения операторных уравнений (см. главу 8), которые не только находят достаточно хорошее приближение, но и гарантируют существование точного решения в определенной окрестности этого приближения.

Поэтому при локализации корня уравнения (*) мы будем ставить перед вычислительным процессом следующие цели:

- установить в пространстве U решений размеры области существования единственного корня (*), позволяющие сделать вывод об адекватности физической и математической моделей;
- найти в области локализации Q точного корня (*) его наилучшее приближение по норме пространства U решений в данной интегрированной среде программирования.

От правильного выбора вида математической модели и качества численного эксперимента зависит, как быстро будет найдено приближение ${}^{N}u(x)$ искомого результата $u^*(x)$ с заданной погрешностью ε.

Перечислим ключевые моменты постановки численного эксперимента:

1. Выбор абстрактного пространства ($L^1, L^2, C, C^1, \ldots$), содержащего решение реальной задачи, и конструирование математической модели.

2. Идея поиска корней операторного уравнения математической задачи (вид представления и метод решения).
3. Построение оптимального ВП решения задачи (итерационный метод, вычислительная техника, среда программирования).
4. Постановка и достижение потенциальных целей эксперимента:
   a) определить область существования решения;
   b) локализовать точное изолированное решение;
   c) найти наилучшее приближение корня в интегрированной среде;
   d) доказать, что в области Q_R корней уравнения не существует.

Наиболее удобное представление приближения корня (*) – многочлен из всюду плотного множества пространства решений U. Причем для удобства решения (*) функции, входящие в оператор F и правую часть уравнения, будем ε-аппроксимировать многочленами по соответствующим нормам. Тогда модифицированный оператор уравнения будет действовать из M-пространства $C^{\infty}_{[a,b]}$ в $C^{\infty}_{[a,b]}$, ограниченные подмножества которого предкомпактны в изучаемых B-пространствах. Это позволит решать уравнение (*) методами, схожими с методами математического анализа решения систем уравнений относительно искомых коэффициентов приближения корня.

Если достигнута поставленная цель эксперимента (то есть определен радиус r области существования точного решения, локализован корень u^* уравнения (*) в δ-окрестности приближения, минимизирован функционал $\|F(^n u) - v\|$ по заданному n или достигнута в области существования корня необходимая точность приближения по невязке), то полученное приближенное решение $^n u(x)$ уравнения (*) будем называть ***искомым приближением***. В противном случае (или если не достигнута какая-то другая цель эксперимента) необходимо изменить одну или несколько качественных характеристик и еще раз повторить численный эксперимент.

Во многих задачах при доказательстве адекватности математической модели некоторому естественному процессу используют статистические методы исследования. В первую очередь здесь определяют закон, по которому распределена изучаемая случайная величина (СВ), а затем на основании теории вероятности находят достоверность утверждаемых результатов численного эксперимента. Рассматриваемые в математической статистике гипотезы принимаются или отвергаются в зависимости от выбранной вероятности предположения их истинности.

Пусть СВ X распределена дискретно, то есть может принимать конечное число значений. Часто эти значения выбираются на основании оценочной шкалы, например, роста человека в сантиметрах. Покажем, как вычисляется вероятность того, что «случайно выбранный футболист из группы мальчиков 2000 года рождения окажется ростом не ниже 155 см».

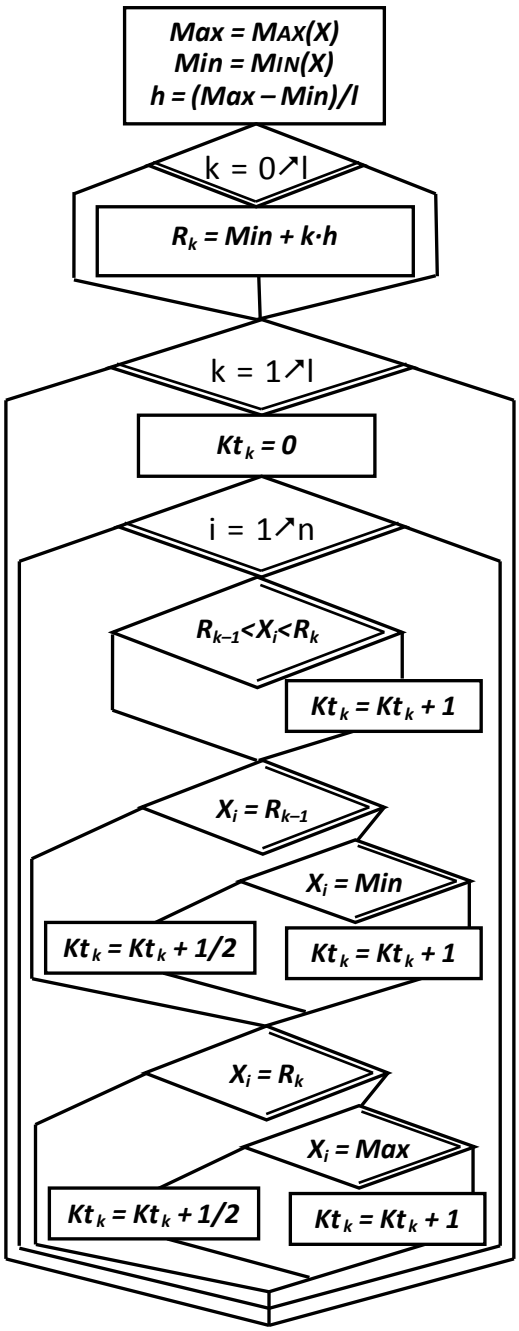


Рисунок 37

Результаты измерения роста учащихся следующие:
170 см – ***1***; 165 см – ***1***; 161 см – ***2***; 158 см – ***2***; 157 см – ***2***; 156 см – ***2***; 155 см – ***1***; 154 см – ***2***; 153 см – ***1***; 152 см – ***2***; 151 см – ***1***; 150 см – ***2***; 149 см – ***1***; 148 см – ***1***; 147 см – ***2***; 146 см – ***2***; 145 см – ***3***; 144 см – ***1***; 143 см – ***2***; 141 см – ***1***; 138 см – ***1***; 136 см – ***1***; 135 см – ***1***; 130 см – ***1***.

Так как данная шкала имеет достаточно мелкую градацию, осуществим переход от сантиметровой шкалы к интервальной длиной 5 см (при этом для чистоты эксперимента граничные значения будем поровну относить к обоим смежным интервалам). Объект-схема алгоритма сортировки приведена на рисунке 37.

Основными характеристиками СВ $\boldsymbol{X}$ являются среднее значение $\boldsymbol{Xs}$ и дисперсия $\boldsymbol{Dx}$ выборки. В дискретном случае среднее значение находится как

$$\bar{x} = \boldsymbol{Xs} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{i=1}^{n} \boldsymbol{X}_i ; \qquad (10.8)$$

а дисперсия вычисляется по формуле

$$\sigma_x^2 \equiv \boldsymbol{Dx} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{Xs})^2. \qquad (10.9)$$

Определение 5.1. Для случайной величины $\boldsymbol{X}$, заданной на $\boldsymbol{l}$ интервалах, среднее значение $\boldsymbol{Xs}$, дисперсия $\boldsymbol{Dx}$ и стандартное отклонение $\boldsymbol{Sx}$ выборки находятся по «интервальным» формулам

$$\boldsymbol{Xs} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{k=1}^{l} \boldsymbol{Kt}_k \boldsymbol{X}_k^s , \qquad (10.10)$$

$$\sigma_x^2 \equiv \boldsymbol{Dx} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{k=1}^{l} \boldsymbol{Kt}_k (\boldsymbol{X}_k^s - \boldsymbol{Xs})^2, \qquad (10.11)$$

где $\boldsymbol{X}_k^s$ – центр $\boldsymbol{k}^{\text{-го}}$ интервала, $\boldsymbol{Kt}_k$ – число точек в $\boldsymbol{k}^{\text{-ом}}$ интервале, $\boldsymbol{n} = \sum_{k=1}^{l} \boldsymbol{Kt}_k$.
Стандартное отклонение $\sigma_x \equiv \boldsymbol{Sx}$ найдем по формуле $\sigma_x = \sqrt{\boldsymbol{Dx}}$.

Минимальное значение выборки обозначим a, максимальное b.

Размах выборки находится как $R = b - a$, а длина интервала $h = R/l$. Тогда границы интервалов можно найти по формуле $R_k = a + k \cdot h,\ k = 0, \ldots, l$. Выражение $P = p(X \le R_k)$ означает, что вероятность случайного выбора спортсмена ростом не выше R_k равна P. Частота, частость распределения и другие результаты обработки данных СВ X при $l = 8$ сведены в таблицу 6.

Таблица 6 – Градация роста футболистов ЦОР 2000 года рождения

Рост (k) / ИП	130–135	135–140	140–145	145–150	150–155	155–160	160–165	165–170
Частота (Kt_k)	1,5	2,5	5,5	8	9	5,5	2,5	1,5
Частость (Kt_k/n)	0,042	0,069	0,153	0,236	0,208	0,181	0,069	0,042
Вероятность ($X \le R_k$)	0,042	0,111	0,264	0,500	0,708	0,889	0,958	1,000

Ответом на поставленный ранее вопрос является $p(X \ge 155) = 0{,}292$, который приближенно равен отношению благоприятных исходов выбора к их общему числу. Однако в математической статистике большое значение имеет обобщение результатов эксперимента на случай непрерывного (дискретно-непрерывного) изменения значений СВ X. Полученная таким образом модель некоторого процесса (объекта) в определенных условиях может быть применена для изучения аналогичного процесса (объекта) с произвольным числом элементов выборки и ее градации.

Решающим здесь является замена дискретной функции вероятности непрерывной функцией плотности распределения вероятности. Опишем алгоритм одного из вариантов адекватной (с точностью ε) замены. Примем значения частости повторений за образ достаточно гладкой функции в центрах интервалов разбиения выборки. Будем считать, что значения этой функции в точках, отстоящих от минимума и максимума выборки на расстоянии полшага в меньшую и большую стороны соответственно, равны нулю.

Если с ростом размерности выборки ее границы не меняются, то при увеличении числа точек сетки интерполяционный (аппроксимационный) процесс, соответствующий сеточной функции вероятности, будет сходиться к некоторой достаточно гладкой функции. Чтобы функция задавала плотность распределения вероятности на рассматриваемом отрезке требуется, чтобы она была неотрицательна и ее определенный интеграл от $-\infty$ до $+\infty$ был равен 1. Добиться этого можно разделив значения сеточной функции на значение определенного интеграла полученного многочлена.

Сейчас возможно более точно, чем в дискретном случае, установить: является ли данная СВ X нормально распределенной. Формула плотности распределения вероятности при нормальном распределении имеет вид

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} exp\left(- \frac{(x - Xs)^2}{2\sigma_x^2} \right). \qquad (10.12)$$

Сопоставляя по норме пространства $C_{1[a,b]}$ функцию (10.12) и многочлен плотности распределения, который в исследуемой задаче равен

$$^{9}f(x) = -928{,}6930582 + 22{,}59329853\,x - 0{,}204062818\,x^2 +$$
$$+ 0{,}001933675\,x^3 - 2{,}7878\cdot10^{-5}\,x^4 + 1{,}7828\cdot10^{-7}\,x^5 + 4{,}6939\cdot10^{-10}\,x^6 -$$
$$- 1{,}1162\cdot10^{-11}\,x^7 + 4{,}9744\cdot10^{-13}\,x^8 - 7{,}395\cdot10^{-16}\,x^9,$$

устанавливаем, что расстояние между ними не превосходит числа $\delta = 0{,}06$.

Интегральная формула вероятности события $X < x$ имеет вид

$$p(X < x) = \int_a^x f(t)dt.$$

Используя представление $f(x)$ в виде $^{9}f(x)$, получим $p(X \geq 155) = 0{,}289$.

Математическое ожидание СВ X (равное при нормальном распределении среднему значению Xs) находится по формуле

$$M[x] = \int_{-\infty}^{\infty} x\, f(x)dx.$$

Для двух случайных величин X и Y, кроме рассмотренных выше индивидуальных признаков выборки в статистике, изучаются парные закономерности свойств заданных СВ. Некоторые из них не предполагают установления функциональной зависимости одной случайной величины от другой. К таким характеристикам СВ X и СВ Y относятся коэффициенты ковариации Kxy и корреляции Rxy, определяемые по формулам

$$Kxy = \frac{1}{n} \sum_{i=1}^{n} (X_i - Xs)(Y_i - Ys) \qquad (10.13)$$

$$Rxy = \frac{Kxy}{\sigma_x\, \sigma_y}. \qquad (10.14)$$

Если одна из случайных величин (чаще это СВ Y) зависит от другой (СВ X), то математическая статистика изучает линейную или нелинейную регрессию соответствующих выборок. При установлении линейной зависимости СВ Y от СВ X находят коэффициенты α и β, входящие в формулу

$$y = \alpha x + \beta. \qquad (10.15)$$

Для решения этой задачи введем дополнительные параметры выборок

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \ – \text{среднее квадратов,} \tag{10.16}$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i \ – \text{среднее произведений,} \tag{10.17}$$

а также укажем их взаимосвязь с ранее определенными понятиями

$$Dx = \overline{x^2} - \left(\overline{x}\right)^2 \tag{10.18}$$

$$Kxy = \overline{xy} - \overline{x}\ \overline{y} \tag{10.19}$$

Теория вычисления коэффициентов (10.15) методом наименьших квадратов описана в разделе 3.5. Из-за важности данного частного случая рассмотрим его более подробно. Функционал минимизации задачи имеет вид

$$Z(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha x_i + \beta))^2 \to min. \tag{10.20}$$

Для определения точки экстремума $Z(\alpha, \beta)$ найдем частные производные функции двух переменных и приравняем их к нулю

$$\begin{cases} \dfrac{\partial Z}{\partial \alpha} \equiv \sum\limits_{i=1}^{n} 2(y_i - \alpha x_i - \beta)(-x_i) = 0 \\ \dfrac{\partial Z}{\partial \beta} \equiv \sum\limits_{i=1}^{n} 2(y_i - \alpha x_i - \beta)(-1) = 0 \end{cases}. \tag{10.21}$$

Полученная система уравнений – линейная относительно α и β:

$$\begin{cases} \alpha \sum\limits_{i=1}^{n} x_i^2 + \beta \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i\, y_i; \\ \alpha \sum\limits_{i=1}^{n} x_i + \beta \sum\limits_{i=1}^{n} 1 = \sum\limits_{i=1}^{n} y_i, \end{cases} \tag{10.22}$$

а ее основная матрица (как и любая другая матрица Грамма) – симметрична относительно главной диагонали.

Разделим обе части каждого из уравнений (10.22) на n. Тогда, учитывая равенства (10.18) и (10.19), корень данной системы вычисляется по формулам

$$\alpha = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2} = \frac{Kxy}{Dx} = \frac{\sigma_y}{\sigma_x} Rxy, \ \beta = \overline{y} - \alpha\, \overline{x}. \tag{10.23}$$

Дискретные и непрерывные нормально или близко к нормальному распределенные случайные величины (параметры, признаки) поддаются детальному изучению не только в задачах на доказательство существенного различия средних (критерий Стьюдента), в парном корреляционном анализе (коэффициент корреляции Пирсона), но и в теории распознавания образов (вычисление индивидуальной или совокупной информативности признаков с помощью модифицированной дивергенции Кульбака).

Задача диагностической классификации по ожидаемой надежности применительно к любым исследуемым объектам может быть сформулирована следующим образом. После изучения партии объектов на долговечность (надежность и т.п.) выявлено, что часть объектов (их число равно ***Ne***) в данной партии оказались удовлетворяющими предъявленным запросам (качественными), а другая часть (***Nk***) не соответствовала требуемым стандартам. Общее число объектов обозначим ***N = Ne + Nk.***

Рисунок 38

Будем называть совокупности ***Ne*** и ***Nk*** элементов соответственно экспериментальным (Э) и контрольным (К) классами (группами) партии из ***N*** объектов.

Диагностическая классификация совокупности объектов заключается в том, чтобы заблаговременно с достаточной степенью вероятности отнести каждый из изучаемых объектов к одному из двух классов (Э или К) по надежности с порогом разделения классов

$$Pr = \ln \frac{1 - P(Э)}{P(Э)}, \qquad (10.24)$$

где ***P(Э)*** – априорная вероятность попадания объекта выборки в экспериментальный класс Э, определяемая по формуле классической вероятности.

Диагностическая процедура распознавания образов состоит из следующих основных этапов:

1. Оценка индивидуальной информативности диагностических признаков (их количество равно ***Kp***);
2. Обучение ИС процессу распознавания принадлежности объектов классам по наиболее информативной совокупности диагностических признаков с минимизацией этих совокупностей;

3. Распознавание, то есть отнесение к классам Э или К отдельных экзаменуемых объектов, не входящих в обучающую выборку.

При необходимости можно выделить промежуточный класс объектов ***Np***, в этом случае задача распознавания распадается на несколько подзадач.

Рассмотрим основной алгоритм (рисунок 38) процедуры распознавания. В блоке 1 осуществляется ввод данных обучающей выборки (ОВ). Замеры экспериментального класса обозначим Re_{ik}, а контрольного – Rk_{ik}, где индекс i указывает порядковый номер объекта в классах Э или К, а индекс k – номер измеряемого параметра.

Кроме основных признаков, информативными могут оказаться производные от них признаки (блок 2). Например, при распознавании образов важную роль может играть отношение значений двух исходных признаков. Это объясняется тем, что обучающая выборка исследуется с помощью средних значений и коэффициентов ковариации признаков, которые не учитывают индивидуальные особенности объекта.

Информативность отдельных признаков (блок 3) вычисляется как

$$I_k = \frac{(Se_k - Sk_k)^2}{2\,K_{kk}}, \qquad (10.25)$$

где Se_k и Sk_k – средние $k^{\text{-го}}$ параметра в классах Э и К.

Диагональный элемент обобщенной (усредненной для двух классов) ковариационной матрицы K, используемый в (10.25), находится из формулы

$$K_{kj} = \frac{\sum_{i=1}^{Ne}(Re_{ik} - Se_k)(Re_{ij} - Se_j) + \sum_{i=1}^{Nk}(Rk_{ik} - Sk_k)(Rk_{ij} - Sk_j)}{Ne + Nk - 2}. \qquad (10.26)$$

Общую ковариационную матрицу можно построить по всей выборке.

В большинстве случаев самый информативный признак входит в набор параметров, имеющих наибольшую совокупную информативность. Однако при определенных условиях обучающая выборка лучше распознается при совмещении параметров, не имеющих максимальной информативности. Поэтому при компоновке совокупности, состоящей из k признаков (блок 4), в качестве первого будем поочередно использовать каждый параметр i, описанный в обучающей выборке (рисунок 39).

Вычисление совокупной информативности Si_{ik} набора $Mp_{ij}, j = 1, \ldots, k$ параметров осуществляется в два этапа. Сначала находим составляющую $C = (c_1, \ldots, c_k)$ – вектор размерности k с координатами

$$c_i = \sum_{j=1}^{k} \delta_{ij}\, u_j, \qquad (10.27)$$

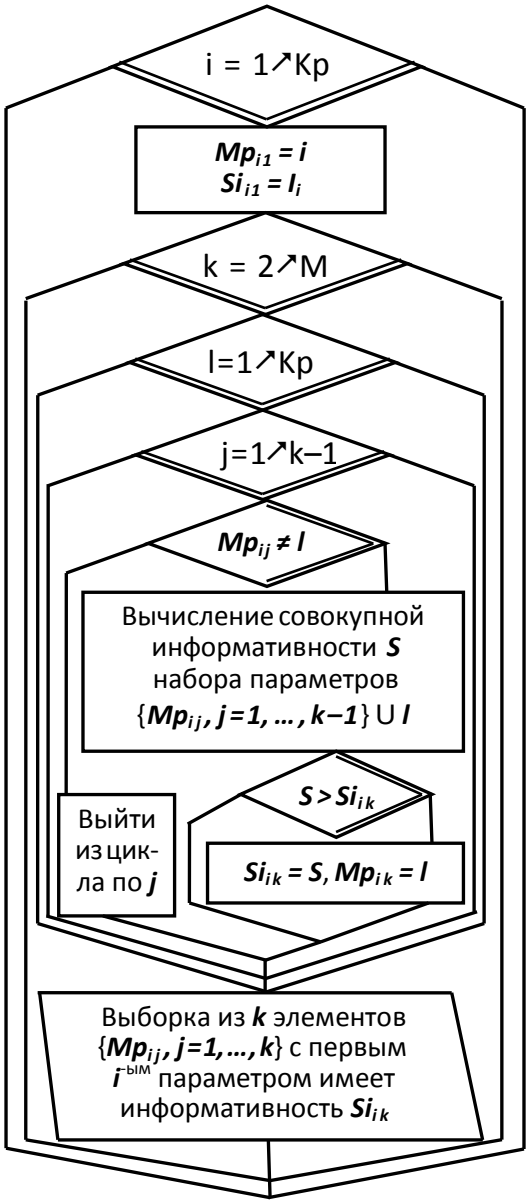


Рисунок 39

где $u_j = Se_j - Sk_j$ – разность средних значений $j^{\text{-го}}$ параметра в классах Э и К, а матрица $G = \{\delta_{ij}, i,j = 1, \ldots, k\}$ является обратной к матрице

$$S = \{s_{ij}, i,j = 1, \ldots, k\},$$

составленной из ковариационной матрицы K (10.26) путем отбора элементов, расположенных в строках и столбцах, соответствующих нумерации признаков в массиве совокупности $Mp_{ij}, j=1, \ldots, k$.

Следующим шагом вычисляется скалярное произведение векторов U и C

$$I^k = \sum_{i=1}^{k} u_i c_i, \qquad (10.28)$$

определяющее совокупную информативность указанного набора из k признаков.

Таким образом формируются Kp наборов признаков с увеличивающимися совокупной информативностью и надежностью распознавания, в которых первым располагается параметр, соответствующий номеру набора.

Наряду с информационной мерой Кульбака (направленное расхождение) будем использовать модифицированную дивергенцию Кульбака

$$div[p] = \sum_{i=1}^{N} (P(p_i \backslash Э) - P(p_i \backslash K)) \ln \frac{P(p_i \backslash Э)}{P(p_i \backslash K)}, \qquad (10.29)$$

где $P(p \backslash Класс)$ – вероятность диагноза $Класс$ при наличии признака p.

В работе приведена методика расчетов этих параметров и для модифицированной информационной меры (МИМ)

$$I_p(Э:K) = \int_{p_{\min}}^{p_{\max}} f_Э(p) \ln \frac{f_Э(p)}{f_K(p)} dp. \qquad (10.30)$$

Расчет вероятностей ошибок I и II рода:

$$P_I = P(Э)\ \Phi\left(\frac{PR - I(Э{:}К)}{\sqrt{2 \cdot I(Э{:}К)}}\right); \qquad (10.31)$$

$$P_{II} = (1 - P(Э))\ \Phi\left(-\frac{PR + I(Э{:}К)}{\sqrt{2 \cdot I(Э{:}К)}}\right), \qquad (10.32)$$

где $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx$ – функция Лапласа, а значение $I(Э{:}К)$ – количество информации о разделении классов – находится по всем Np признакам как

$$I(Э{:}К) = So\ K^{-1}\ U, \text{ где } So = \frac{1}{2}(Se + Sk). \qquad (10.33)$$

При равных ковариационных матрицах двух классов логарифм отношения правдоподобия для изучаемого $i^{\text{-го}}$ объекта вычисляется по формуле

$$\boldsymbol{Pr_i = \ln \eta_i = \sum_{k=1}^{Kp} (Rкласс_{ik} - So_k) c_k} \qquad (10.34)$$

Сравнивая $\boldsymbol{Pr_i}$ с порогом разделения $\boldsymbol{Pr}$ (10.24), устанавливаем: к какому из двух классов принадлежит данный объект.

Имея в массиве $\{(\boldsymbol{Mp_{ij}, j=1, \ldots, Kp}), \boldsymbol{i = 1, \ldots, Kp}\}$ все возможные наборы параметров с надежностью $\boldsymbol{P_н}$ и эффективностью $\boldsymbol{P_э}$ распознавания равной $\boldsymbol{100\%}$, можно в блоке 5 для любого значения этих параметров выбрать совокупность меньшего объема, чем $\boldsymbol{Kp}$. Надежность и эффективность распознавания в этом случае рассчитывается по формулам

$$\boldsymbol{P_н = (1 - (P_I + P_{II})) \cdot 100\%}; \qquad (10.35)$$

$$\boldsymbol{P_э = \left(1 - \frac{P_{II}}{1 - P(Ý)}\right) \cdot 100\%}; \qquad (10.36)$$

В блоке 6 раздельно вводятся данные экзаменуемой выборки, а в заключительном, седьмом, блоке осуществляется распознавание этой выборки с заданной надежностью и эффективностью.

Теория распознавания образов может использоваться не только для выявления наиболее информативного набора параметров, с помощью которого происходит разделение партии приборов на классы по качеству или долговечности эксплуатации. Одним из важнейших аспектов применения этой теории является селекционная работа по разведению районированных сортов семян растений, а также при отборе контингента индивидуумов, пригодных для определенной профессиональной деятельности.

10.4 Прикладная математическая программа. Объектно-ориентированное программирование

Будем считать, что ***математическая программа*** – это последовательность логически связанных предложений языка программирования, предназначенная для решения математической задачи по разработанному алгоритму, а также исполняемый файл, используемый для обработки данных и вывода результатов вычислений.

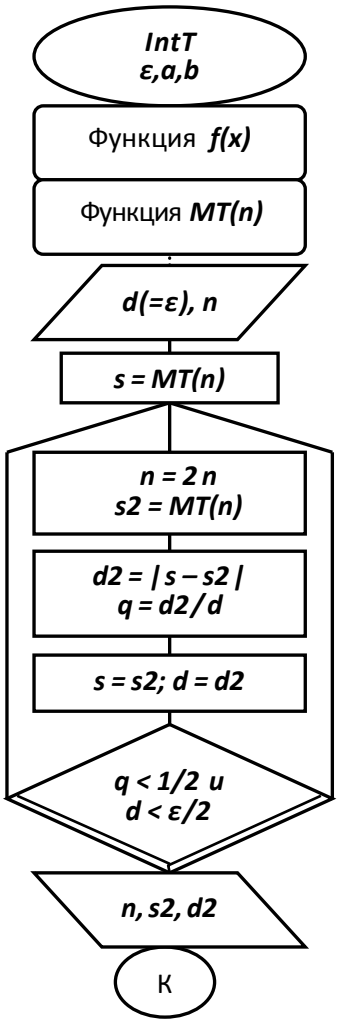


Рисунок 40

Разработаем алгоритм и составим программу вычисления определенного интеграла функции

$$y = f(x),\ x \in [a, b],\ f \in C_{[a, b]} \qquad (10.37)$$

методом трапеций (раздел 9.3) с точностью ε.

Критерием Останова вычислительного процесса (по заданной точности ε) в этом случае является близость значений $|I_h - I_{h/2}| < \varepsilon/2$ квадратурных сумм, полученных при интегрировании функции (10.37) с шагами h и $h/2$. Построив последовательность интегральных сумм $\{I^i \equiv I_{h/2}i,\ i = 0, 1, \ldots\}$, модуль разности соседних членов которой мажорируется бесконечно убывающей геометрической прогрессией, можно найти радиус области существования точного значения I_t вблизи приближения I^i.

Например, для всякой дважды непрерывно дифференцируемой функции существует i, начиная с которого знаменатель q геометрической прогрессии станет меньше $1/2$. Тогда при $\forall\varepsilon > \varepsilon_{ИС}$ точное значение I_t будет находиться в области $|I_t - I^i| < \varepsilon$.

Опишем алгоритм решения задачи с помощью одноступенчатого циклического процесса (рисунок 40). Приближенное значение $s = I_h$ интеграла вычисленного с шагом $h = (b - a)/n$, где n – количество интервалов разбиения $[a, b]$, используемых в методе трапеций, входит в предикат блока принятия решения.

Так как в рассматриваемом итерационном процессе определить значение логического выражения P по входным данным невозможно, то его формирование продолжается в теле цикла и для вычисления P требуется один раз выполнить функцию MT (с $n := 2n$) в конструкции повторений.

При проектировании сложной программной системы с помощью структурного программирования проводится алгоритмическая декомпозиция решаемой задачи (представление разрабатываемой системы в виде взаимодействующих подсистем, модулей или блоков).

Существование интеграла I_t функции $f(x)$ обеспечивается ее непрерывностью на отрезке интегрирования $[a, b]$. Метод трапеций сходится к I_t со вторым порядком точности по h для функций из $C^2_{[a, b]}$ (см. раздел 9.3), поэтому в данном случае погрешность формулы имеет первый порядок.

Чтобы составить программу по объект-схеме алгоритма решения задачи, надо предварительно разработать алгоритм функции **MT(n)** вычисления приближенного значения определенного интеграла функции $y = f(x)$ методом трапеций (9.25) по заданному числу n интервалов разбиения $[a, b]$. Эти подпрограммы можно отладить независимо друг от друга. СП ориентировано на составление программ с использованием процедур и функций.

Приведем текст Паскаль-программы решения задачи этого раздела.

```
PROGRAM IntMT;
USES WinCrt;
CONST e=1e-10; a=0; b=pi/2;
VAR n: LongInt;
Function f(x: Real): Real;
Begin
f:=Cos(x);
End;
Function MT(n: LongInt): Real;
Var i: LongInt; h,s: Real;
Begin
h:=(b-a)/n;
s:=(f(a)+f(b))/2;
For i:=1 To n-1 do
s:=s+f(a+i*h);
MT:=s*h;
End;
Function IntT(e: Real):Real;
Var n:LongInt;
    d,d2,s,s2,q:Real;
Begin
d:=e;
n:=10; s:=MT(n);
Repeat
n:=2*n; s2:=MT(n);
d2:=abs(s-s2);
q:=d2/d;
s:=s2; d:=d2;
Until (q < 0.5) and (d < e/2);
IntT:=s;
End;
BEGIN
Write('aSb f(x)dx= ',IntT(e):0:10);
END.
```

Многие важные аспекты языка программирования (машинное кодирование символов, числовых и логических значений) не описываются стандартом. Стандарт не определяет порядок создания программы для определенной среды выполнения.

Процесс создания программ, основанный на синтаксических и семантических особенностях интегрированной среды, включает четыре этапа:

1) предварительная запись и редактирование текста программы с последующим сохранением ее в виде исходного файла или модуля;

2) компиляция программы на установленном промежуточном языке и сохранение ее в виде объектного файла или модуля;

3) построение исполняемого файла или модуля путем объединения (компоновки) полученного объектного модуля программы с другими объектными модулями стандартных и специальных библиотек;

4) отладка программы посредством встроенного в интегрированную среду программирования отладчика, облегчающего обнаружение ошибок.

Рассмотрим основные этапы эволюции структурированного подхода в программировании, позволяющие выяснить взаимосвязь структурного (СП), модульного (МП) и объектно-ориентированного программирования (ООП).

Дальнейшее развитие структурного подхода привело к модульному программированию, предусматривающему декомпозицию прикладной задачи в виде иерархии взаимодействующих модулей и программ. Модуль, содержащий данные и процедуры их обработки, удобен для автономной разработки и отладки.

Специализация модулей по видам обработки, наличие в них данных определяемых типов – это свойства, отражающие генетическую связь модульного программирования и ООП. Результатом обобщения понятия «тип данных» являются классы объектов (C++) или объектные типы (Pascal), которые могут содержать в качестве элементов не только данные определенного типа, но и методы их обработки (функции и процедуры).

С позиции технологии программирования объект в ООП – это определенная программная структура, обладающая тремя важнейшими свойствами: инкапсуляции, наследования и полиморфизма.

Инкапсуляция (содержание) представляет собой объединение и локализацию в рамках объекта как единого целого данных и функций, обрабатывающих эти данные.

Наследование – способность объектов порождать своих потомков и наследовать свойства (элементы данных и методы изучения) своих родителей.

Полиморфизм – свойство объектов-родственников выполнять аналогичные действия (в зависимости от природы субъекта) по-разному.

Описание реального явления в форме взаимодействующих объектов естественнее, чем в форме иерархии подпрограмм, и поэтому облегчает программное моделирование процессов в изучаемой предметной области.

Процесс программирования в стиле ООП включает следующие этапы:

1. Определение основных понятий предметной области и соответствующих им объектов, имеющих определенные свойства (возможные состояния и действия);
2. Описание принципов взаимодействия изучаемых объектов в рамках программной системы;
3. Установление иерархии взаимосвязи свойств родственных объектов;
4. Реализация иерархии объектов с помощью механизмов инкапсуляции, наследования и полиморфизма;
5. Использование полного набора методов для управления свойствами изучаемых объектов.

Первые три этапа являются объектно-ориентированным анализом предметной области (создание библиотеки объектов). Четвертый и пятый этапы – моделирование, программирование и изучение реальных процессов.

Как интегрированная среда ООП обладает следующими достоинствами:

- возможность введения новых понятий на основе старых;
- отражение в библиотеке объектов общих свойств и отношений между объектами предметной области;
- естественность отображения пространства исходной задачи в пространство объектов программы;
- простота внесения изменений в объекты и программу;
- упрощение составления программ для родственных (аналогичных) предметных областей.

В качестве примера рассмотрим программу построения графика в VBA с возможностью динамического изменения параметра размерности массивов данных, категории графика, названия элементов и т.д. В описании объектов, соответствующих каждому состоянию или действию, определим их взаимодействие и иерархию в управлении свойствами изучаемых предметов.

```
n = Range("J3")
NamGraf = Range("C1")
Set Adx = Range(Cells(2, 2), Cells(2 + n, 2))
Set Ady = Range(Cells(2, 3), Cells(2 + n, 3))
   Charts.Add
   ActiveChart.ChartType = xlLineMarkers
   ActiveChart.SetSourceData Source:=Ady, PlotBy:=xlColumns
   ActiveChart.SeriesCollection(1).XValues = Adx
   With ActiveChart
      .HasTitle = True
      .ChartTitle.Characters.Text = NamGraf
      .Axes(xlCategory, xlPrimary).HasTitle = True
      .Axes(xlCategory, xlPrimary).AxisTitle.Characters.Text = "X"
      .Axes(xlValue, xlPrimary).HasTitle = True
      .Axes(xlValue, xlPrimary).AxisTitle.Characters.Text = "Y"
```

```
    End With
    With ActiveChart.Axes(xlCategory)
       .CrossesAt = 1
       .TickLabelSpacing = 2
       .TickMarkSpacing = 1
       .AxisBetweenCategories = False
       .ReversePlotOrder = False
    End With
```

В этой программе предполагается, что исходные данные размещены на рабочем листе MS Excel: в ячейке J3 записан параметр ***n***, задающий размерность (***n+1***) массива значений сеточной функции на отрезке ***[a, b]***; в ячейке C1 находится название графика; во втором и третьем столбцах расположены соответственно абсциссы и ординаты точек графика, количество которых зависит от ***n***. Результатом работы программы является построение графика заданного вида, подписи элементов и т.д.

Наряду с математическим и физическим моделированием (где модель и объект представляют реальные физические процессы) важную роль в изучении предметов и явлений играет *структурно-функциональное* моделирование, при котором аналоговыми моделями служат схемы, графики, таблицы, дополненные специальными правилами их изучения и преобразования.

Функциональная схема математического моделирования физического объекта, изображенная на рисунке 36, описывает структуру изучения адекватности физической и математической моделей. Свою структуру исследования имеют и процессы решения уравнений математической физики.

Например, если пространства решений задач с интегральным и/или дифференциальным оператором содержат всюду плотные множества (степенных или тригонометрических полиномов, функций Эрмита или Лагерра, многочленов Лежандра или Родриго и т.п.), то корни этих уравнений следует приближать элементами именно этих подмножеств.

Перечислим основные модули программы решения функциональных уравнений вида (*) полиномиальными методами:

VVOD: представление данных в предбазисе пространства ***U*** с учетом дополнительных (внутренних и/или граничных) условий;

FNV: описание функции ***Φ(U)*** и вычисление ее нормы ***||Φ||***;

VKU: представление ***U*** в виде вектора коэффициентов полинома;

FRECHEn: взятие первой (второй) производной оператора ***F***;

VVN: поиск вектора невязки ***VN*** в узлах итерационной сетки;

RLS: решение линейной системы алгебраических уравнений;

KRITERk: проверка выполнения условий критерия ***k***:

1 – определение радиуса ***r*** области существования решения;
2 – локализация решения;
3 – минимизация $\|\Phi(^{N}u)\|$ и т.д.

Программа, модульная структура которой обладает тремя основными свойствами ООП, состоит из четырех ключевых блоков (рисунок 41).

Опишем принципиальное назначение блоков алгоритма.

Блок 1 *«Ввод данных»* предназначен для выбора базиса всюду плотного множества пространства решений уравнения (*) и представления в этом базисе нуль-приближения u, удовлетворяющего дополнительным условиям задачи (*). Здесь же осуществляется аппроксимация оператора $F(u)$ и находится функция-невязка $\Phi(u)$.

Блок 2 *«Критерий Останова»* предназначен для выбора условия, по которому происходит завершение итерационного ВП решения математической задачи. Выход из вычислительного процесса может также произойти из-за превосходящего требуемую точность значения машинного ε (код *E*), недостатка памяти (код *П*) интегрированной среды программирования или по другим сообщаемым причинам.

Блок 3 *«Определение следующего приближения»*. В блоке находится очередное приближенное решение (*) функциональными методами, описанными в курсе лекций, с кодом ***n***: *1* – метод Ньютона и его модификации; *2* – метод второго порядка и его модификации; *3* – метод последовательных приближений.

Рисунок 41

Блок 4 *«Вывод результатов счета»*. В блоке определяется код ***k*** завершения программы и выводится приближенное решение уравнения (*), которое удовлетворяет критерию Останова или минимизирует невязку в условиях интегрированной среды. Модуль VYVOD организует сбор информации о степени адекватности математической модели исследуемому объекту при изучении этой модели методами функционального и численного анализа в интегрированной среде программирования.

10.5 Классификация семейств непрерывных на отрезке функций по гладкости элементов

При математическом моделировании физических процессов с помощью интегро-дифференциального оператора на искомые функции накладываются определенные условия гладкости. В зависимости от степени гладкости элементов пространства решений существуют различные способы их приближения многочленами. Аппроксимация суммируемых функций *многочленом наилучшего приближения* по норме $L^2_{[a,\,b]}$ (3.42) осуществляется матрицей Грама, а приближение бесконечно дифференцируемых функций по норме $C_{[a,\,b]}$ (9.6) *интерполяционным многочленом* – матрицей Вандермонда.

Проектирование элементов пространств $L^2_{[a,\,b]}$, $C^{\infty}_{[a,\,b]}$ и $C^k_{[a,\,b]}$ с $k \geq 1$ на множество степенных и тригонометрических многочленов достаточно хорошо изучено в функциональном анализе. Особый интерес с этой точки зрения представляет ***B***-пространство непрерывных функций $C_{[a,\,b]}$. Обнадеживает то, что каждую непрерывную на ***[a, b]*** функцию ***u(x)*** по теореме Вейерштрасса [10, с. 480] можно с любой точностью приблизить многочленом, причем последовательность $^n u(x)$ при $n \to \infty$ равномерно сходится к ***u(x)***.

По теореме Чебышева [12] среди многочленов степени не выше ***n*** существует единственный, являющийся МНП непрерывной функции ***u(x)*** по норме $C_{[a,\,b]}$. Усиление теоремы Вейерштрасса (теорема Фейера [10, с. 477]) позволяет сгенерировать последовательность тригонометрических многочленов $^n t(x)$, равномерно сходящуюся к $u(x) \in C_{[a,\,b]}$.

Так как множество $C_{[a,\,b]}$ всюду плотно в $L^p_{[a,\,b]}$, а множество $P_{[a,\,b]}$ всюду плотно в $C_{[a,\,b]}$, процесс аппроксимации непрерывных функций многочленами обретает особое значение. В учебнике [12] доказано, что качество интерполирования функций из $C_{[a,\,b]}$ и $C^1_{[a,\,b]}$ имеет существенное различие.

Проследим за свойствами элементов $C^{s}_{[a,\,b]}$ в диапазоне изменения гладкости от непрерывности функций до их дифференцируемости и изучим соответствующие этой классификации интерполяционные процессы.

Определение 10.2. Функция $u(x)$ называется ***непрерывной в точке*** $x_0 \in [a, b]$, если $\forall\varepsilon > 0 \;\; \exists\delta > 0$ такое, что справедлива импликация

$$\{|x - x_0| < \delta\} \Rightarrow \{|u(x) - u(x_0)| < \varepsilon\} \text{ при } x \in [a, b].$$

Определение 10.3. Функция $u(x)$ называется ***непрерывной на отрезке*** $[a, b]$, если она непрерывна в каждой точке этого отрезка.

Определение 10.4. Функция $u(x)$ называется *равномерно непрерывной на отрезке* $[a, b]$, если $\forall\varepsilon > 0 \;\; \exists\delta > 0$ такое, что справедлива импликация

$$\{|x_1 - x_2| < \delta\} \Rightarrow \{|u(x_1) - u(x_2)| < \varepsilon\} \text{ при } x_1 \text{ и } x_2, \text{ принадлежащих } [a, b].$$

Определение 10.5. Множество функций $S_{[a,\,b]}$ называется ***семейством равностепенно непрерывных функций***, если $\forall\varepsilon > 0 \;\; \exists\delta > 0$ такое, что из

$$\{|x_1 - x_2| < \delta\} \Rightarrow \{|u(x_1) - u(x_2)| < \varepsilon\} \text{ для } \{x_1, x_2\} \subset [a, b] \text{ и } \forall u(x) \in S.$$

Определение 10.6. Функция $u(x)$ называется *абсолютно непрерывной на отрезке* $[a, b]$, если $\forall\varepsilon > 0 \;\; \exists\delta > 0$ такое, что справедлива импликация

$$\{\sum_{i=1}^{n} (\beta_i - \alpha_i) < \delta\} \Rightarrow \{\sum_{i=1}^{n} |u(\beta_i) - u(\alpha_i)| < \varepsilon\},$$

где $S_n = \{\{a_i, b_i\}, i = 1, 2, \ldots, n\}$ – любая конечная система попарно непересекающихся интервалов отрезка $[a, b]$.

Определение 10.7. Функция $u(x)$ называется ***липшиц-непрерывной на отрезке*** $[a, b]$, если существует такое действительное число $L < \infty$, что для всех точек x_1 и x_2, принадлежащих $[a, b]$, выполняется неравенство

$$|u(x_1) - u(x_2)| < L|x_1 - x_2|.$$

Определение 10.8. Функция $u(x)$ называется непрерывно ***дифференцируемой в точке*** $x_0 \in [a, b]$, если $\forall\varepsilon > 0$ найдутся такие действительные числа $K < \infty$ и $\delta > 0$, что для указанных $x \in [a, b]$ будет справедлива импликация

$$\{|x - x_0| < \delta\} \Rightarrow \{|u(x) - u(x_0) - K(x - x_0)| \le \varepsilon|x - x_0|\}.$$

Описанные выше непрерывные функции и их совокупно ограниченные семейства классифицируем по гладкости элементов на отрезке $[a, b]$.

Таблица 7 – Классификация непрерывных функций и их семейств

Тип непрерывности функции $u(x)$	В точке $x_0 \in [a, b]$	На отрезке $[a, b]$	Класс ограниченных функций на $[a, b]$
Непрерывная	***Определение 5.2***	Если непрерывная во всех точках $[a, b]$	**M**-пространство, индуцированное нормой C. ε-аппроксимация $u(x)$ многочленом
Равномерно непрерывная	Непрерывная	***Определение 5.4***	
Равностепенно непрерывная	Непрерывная	Равномерно непрерывная	***Определение 5.5*** ($\exists$ конечная ε-сеть)
Абсолютно непрерывная	Почти всюду дифференцируемая	***Определение 5.6***	ε-аппроксимация $u(x)$ на сетке Чебышева
Липшиц-непрерывная	Почти всюду непрерывно дифференцируемая	***Определение 5.7***	Замыкание – множество дифференцируемых функций
Непрерывно дифференцируемая	***Определение 5.8***	Если непрерывно дифференцируемая во всех точках $[a, b]$	**M**-пространство, индуцированное нормой C^1

К сожалению, алгоритм вычисления коэффициентов многочлена наилучшего приближения непрерывных функций по норме $C_{[a, b]}$ индивидуален и зависит от особенностей каждой из них. Другими словами, для любого $n \in N$ в ε-окрестности $u(x)$ найдется функция $v(x) \in C_{[a, b]}$, многочлен ${}^m p(x)$ ε-аппроксимации которой будет иметь порядок $m >> n$. Значит, для аппроксимации функций из шара $\|w(x) - v(x)\| \le \varepsilon$ используется все множество P многочленов со счетной системой базисных функций (3.32).

Теорема 10.1 [12]. Какова бы ни была последовательность ${}^n\Omega$ сеток на отрезке $[a, b]$, существует функция $u(x) \in C_{[a, b]}$, для которой последовательность ${}^n p(x)$ соответствующих ${}^n\Omega$ интерполяционных многочленов не будет равномерно сходится к $u(x)$ при $n \to \infty$.

Тем не менее, если функция $u(x) \in C_{[a, b]}$, то для нее существует такая последовательность ${}^n\Omega$ сеток [12], что соответствующий ${}^n\Omega$ интерполяционный процесс для $u(x)$ будет сходиться к $u(x)$ равномерно на $[a, b]$. По теореме Арцела-Асколи для ограниченного множества равностепенно непрерывных функций выполняется важное для ε-аппроксимации элементов из $C_{[a, b]}$ свойство предкомпактности одного метрического пространства в другом.

Начиная с семейства равностепенно непрерывных функций, ограниченных на отрезке $[a, b]$ в совокупности, существует конечное множество P^n многочленов, ε-аппроксимирущих любую функцию этого семейства в конечном степенном базисе (9.41). Однако процесс аппроксимации равностепенно непрерывных функций достаточно сложный.

И только для абсолютно непрерывных функций разработан алгоритм их аппроксимации многочленами, аналогичный интерполяционному приближению k-дифференцируемых функций. Но для сходимости процесса интерполирования по норме B-пространства $C_{[a, b]}$ равномерную сетку требуется заменить чебышевской, а на лагранжевы коэффициенты наложить условие ограниченности частичных сумм [12].

ТЕСТ ДЛЯ САМОКОНТРОЛЯ ЗНАНИЙ

I. Пусть высказывание A ложно, а B истинно. Определить значение логических выражений

1) $\overline{A \text{ или } B}$;

2) $\overline{A} \text{ или } \overline{B}$;

3) $\overline{A \text{ и } B}$;

4) $\overline{A} \text{ и } \overline{B}$

и выбрать правильный ответ из предлагаемых:

а) первое и второе высказывания ложно;

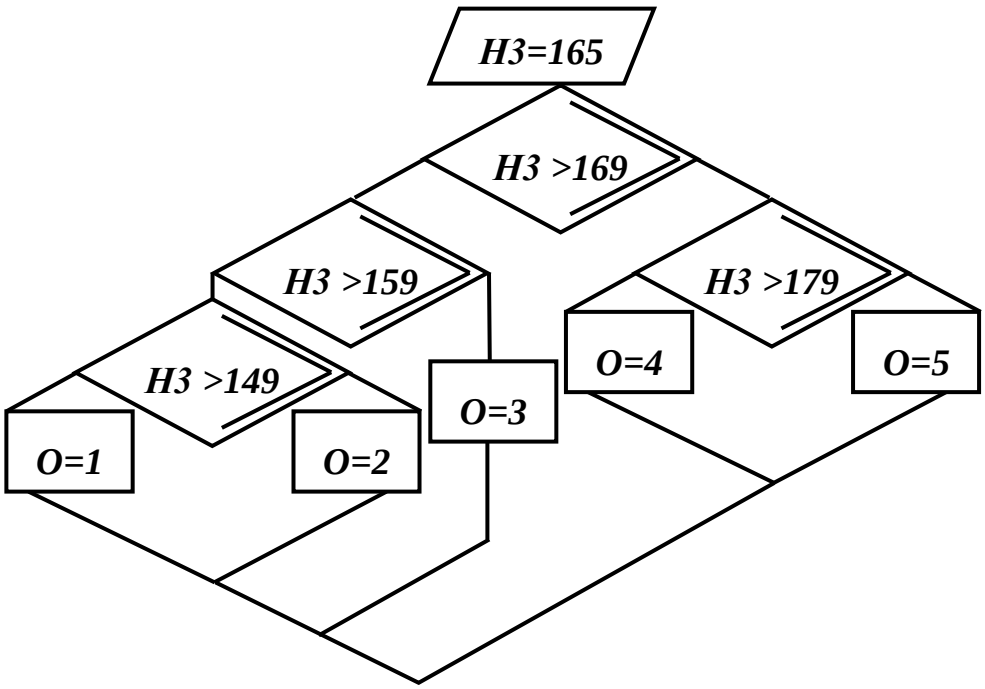
б) первое и третье высказывания ложно;

в) первое и четвертое высказывания ложно;

г) второе и третье высказывания ложно.

Ответ: в.

II. Какую оценку получил ученик по пятибалльной шкале, прыгнув в высоту на 165 см



а) 2;

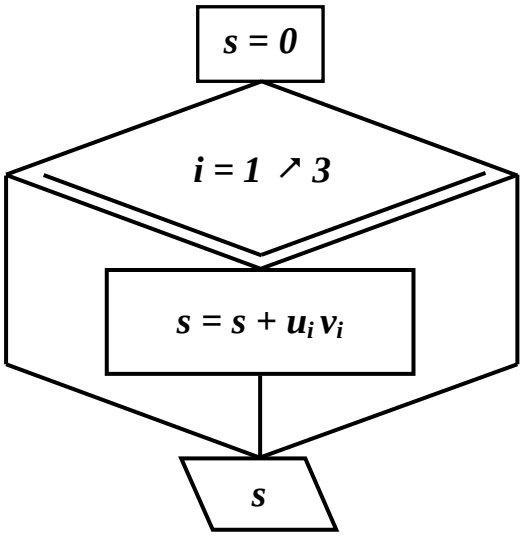
б) 3;

в) 4;

г) 5.

Ответ: б.

III. Определить значение ***s*** перед последним повторением с $i = 3$ при вычислении скалярного произведения векторов $U(1, 2, 3)$ и $V(2, 3, 4)$



а) 4;

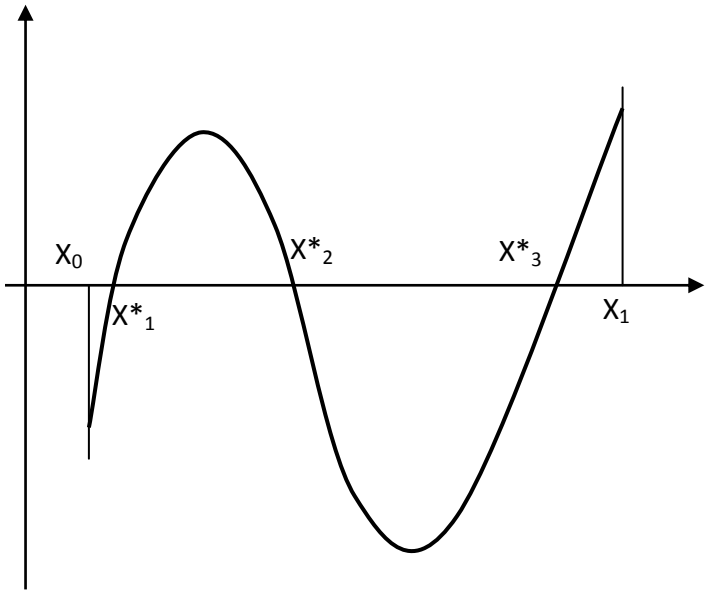
б) 6;

в) 8;

г) 10.

Ответ: в.

IV. Какой из трех корней функции определится методом дихотомии?



а) 1-ый;

б) 2-ой;

в) 3-ий;

Ответ: в.

V. Найти границы НГ и ВГ разности $x - y$, если известны границы, в которых заключены x и y:

$$4 \le x \le 5;\ 2 \le y \le 3.$$

а) 1 и 2;

б) 2 и 3;

в) 1 и 3;

г) 3 и 1.

Ответ: в.

VI. Найти собственные значения матрицы

$$\begin{pmatrix} 1 & 1 \\ 4 & 1 \end{pmatrix}$$

а) 3 и –1;

б) 1 и 3;

в) 1 и –3;

г) –3 и –1.

Ответ: а.

VII. Вычислить значение определителя матрицы методом Гаусса

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 0 \end{pmatrix}$$

а) –27;

б) 27;

в) 0.

Ответ: б.

VIII. Найти приближенное значение интеграла функции $y = x^2$ на отрезке [0; 1] с помощью метода трапеций

i	***0***	***1***	***2***	***3***	***4***	***5***
x_i	***0***	***0,2***	***0,4***	***0,6***	***0,8***	***1***
$f(x_i)$	***0***	***0,04***	***0,16***	***0,36***	***0,64***	***1***

а) 0,33;

б) 0,34;

в) 0,32.

Ответ: б.

IX. Дисперсия СВ $X = \{1, 2, 3\}$ равна

а) $\frac{\sqrt{2}}{2}$;

б) $\frac{2}{3}$;

в) $\frac{\sqrt{3}}{2}$;

г) 0.

Ответ: б.

X. Выборочная дисперсия СВ X, исходя из выборки $\{1, 2, 3\}$, равна

а) 1;

б) $\frac{2}{3}$;

в) $\frac{\sqrt{3}}{2}$;

г) 0.

Ответ: а.

Заключение

Методы решения функциональных уравнений, корнями которых являются элементы пространств $C^k_{[a,\,b]}$ или $\mathcal{L}^p_{[a,\,b]}$, по способу представления приближения корня делятся на две группы. Первую группу образуют дискретные формы записи приближения с последующей аппроксимацией сеточного решения кусочно-непрерывной функцией или полиномом. Большой вклад в популяризацию дискретных методов внесли ученые Н.П. Жидков и И.С. Березин [3], В.И. Крылов, В.В. Бобков и П.И. Монастырный [12,13].

В качестве координат искомого вектора в сеточных методах принимаются значения функции в точках сетки отрезка $[a, b]$. Операторы интегрирования и дифференцирования аппроксимируются суммарными или разностными формулами приближения. Суммарно-разностная аппроксимация, осуществляемая в основном на равномерной сетке, рассчитана на последующий поиск решений, представимых в виде сходящегося ряда Тейлора (к конечному отрезку которого можно приблизиться с помощью интерполяционного процесса).

Второй способ записи приближения корня функционального уравнения основан на представлении его в виде отрезка ряда Фурье. В монографии М.А. Красносельского, П.П. Забрейко и др. [11] изучаются методы Галеркина, где компонентами искомого вектора являются коэффициенты многочлена наилучшего приближения корня – проекции корня в конечномерное подмножество пространства решений со степенным или тригонометрическим базисами. Аппроксимация корня уравнения в этих процессах осуществляется методом наименьших квадратов, то есть по условию минимизации функционала (нормы невязки) в гильбертовом пространстве.

Ограничения применения методов Галеркина связаны с поиском решений в банаховых пространствах, где указанные полные системы функций в общем случае базисами не являются. Кроме того описанные дискретные и проекционные методы не являются итерационными относительно увеличения размерности искомого вектора, обеспечивающего сходимость последовательности приближений к корню

функционального уравнения по норме. Это связано с тем, что значения сеточной функции в узлах измененной сетки и коэффициенты многочлена наилучшего приближения не зависят от значений искомого вектора на предыдущей итерации [11].

Настоящий электронный курс лекций предназначен для студентов стационара физико- математического факультета специальностей «Математика и информатика» и «Физика и информатика». Он составлен в соответствии с действующей типовыми программами дисциплины «Вычислительные методы и компьютерное моделирование» по этим специальностям, утвержденным министром образования РБ.

В электронном курсе лекций излагается теория по темам дисциплины «Вычислительные методы и компьютерное моделирование»: теория погрешностей, методы решения нелинейных уравнений и систем уравнений, нахождение собственных векторов и собственных значений матрицы, приближение функций, численное дифференцирование, обработка данных эксперимента, приближённое вычисление определённых интегралов, аналитические и численные методы решения задачи Коши для обыкновенных дифференциальных уравнений, методы решения задач линейного программирования.

На задачах различного уровня сложности проиллюстрированы основные методы, изложенные в электронном издании. Предлагаются задания для самостоятельной работы. Имеется блок самоконтроля.

Данный электронный курс лекций дает возможность будущим специалистам ознакомиться и изучить численные методы решения задач векторной алгебры, математического анализа, дифференциальных уравнений и облегчить самостоятельную работу студентов с теоретическим материалом при подготовке к лабораторным занятиям и экзамену.

Издание электронного курса лекций инициировано, с одной стороны, большим количеством не всегда доступных студентам источников, с другой – разнообразием терминологии изложения теорий, цитируемых из смежных разделов математической логики, числительных методов и теории алгоритмизации.

Литература

1. Антоневич, А.Б. Функциональный анализ и интегральные уравнения / А.Б. Антоневич, Я.В. Радыно. – Минск : БГУ, 2006. – 430 с.
2. Бахвалов, Н. С. Численные методы : учеб. пособие : в 2 ч. / Н.С. Бахвалов. – М. : Наука, 1975. – Ч.1. – 632с.
3. Березин, И.С. Методы вычислений: в 2 т. / И.С. Березин, Н.П. Жидков. – М. : Наука, 1962. – Т. 2. – 620 с.
4. Богута, Л. И. Вычислительная математика и программирование : учеб. пособие : в 2 ч. / Л.И. Богута [и др]. – Л. : Изд-во ЛГПИ им. Герцена, 1973. – Ч.1. – 82 с.
5. Гантмахер, Ф.Р. Теория матриц / Ф.Р. Гантмахер. – М. : Наука, 1988. – 552 с.
6. Гусак, А. А. Элементы методов вычислений : учеб. пособие / А.А. Гусак. – Минск : Изд-во БГУ, 1982. – 166 с.
7. Иванова, Т. П. Программирование и вычислительная математика : уч. пособие / Т.П. Иванова, Г.В. Пухова. – М. : Просвещение, 1978. – 320 с.
8. Калиткин, Н. Н. Численные методы : учеб. пособие / Н.Н. Калиткин. – М. : Наука, 1978. – 512 с.
9. Канторович, Л.В. Функциональный анализ / Л.В. Канторович, Г.П. Акилов. – М. : Наука, 1977. – 742 с.
10. Колмогоров, А.Н. Элементы теории функций и функционального анализа / А.Н. Колмогоров, С.В. Фомин. – М. : Наука, 1989. – 624 с.
11. Красносельский, М.А. Приближенное решение операторных уравнений / М.А. Красносельский, Г.М Вайнико, П.П. Забрейко. – М. : Наука, 1969. – 456 с.
12. Крылов, В. И. Вычислительные методы : учеб. пособие : в 2 ч. / В.И. Крылов, В.В. Бобков, П.И. Монастырный. – М. : Наука, 1976. – Ч.1. – 304 с.
13. Крылов, В. И. Вычислительные методы : учеб. пособие : в 2 ч. / В. И. Крылов, В.В. Бобков, П.И. Монастырный. – М. : Наука, 1977. – Ч.2. – 400 с.
14. Марчук, Г. И. Методы вычислительной математики : учеб. пособие / Г.И. Марчук. – М. : Наука, 1980. – 534 с.
15. Морозов, В.В. Полиномиальные методы прикладного анализа : монография / В.В. Морозов. – Брест : БрГУ, 2011. – 200 с.
16. Фаддеев, Д. К. Вычислительные методы линейной алгебры : учеб. пособие/ Д.К. Фаддеев, В.Н. Фаддеева. – М. : Физматгиз, 1963. – 734 с.