МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
Белорусский национальный технический университет

Прикладная математика

Учебное пособие
для студентов
специальностей приборостроения

Учебное электронное издание

Минск БНТУ 2009

УДК 519.6(075.8)
ББК 22.193я7
П 75

А в т о р ы :
В.А.Нифагин, Н.Н.Роговцов,
В.М.Романчак, В.Н.Кушнир,
О.В.Титюра

Рецензенты:
И.Н. Мелешко, профессор, доктор физико-математических наук;
Н.В. Кулешов, зав. кафедрой “Лазерной техники и технологий”,
доктор физико-математических наук

В пособии излагаются основы классических численных методов, которые используются в математическом моделировании прикладных задач. Рассмотрены вопросы использования этих методов с применением пакета MathCad. Приводятся задачи для решения и контрольные вопросы проверки теоретических знаний.

Белорусский национальный технический университет
пр-т Независимости, 65, г. Минск, Республика Беларусь
Тел.(017)292-77-52 факс (017)292-91-37
E-mail: emd@bntu.by
http://www.bntu.by/ru/struktura/facult/psf/chairs/im/
Регистрационный № БНТУ/ПСФ85-5.2009

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Развитие математических методов и их использование при решении прикладных задач в значительной степени базируется на компьютерном моделировании. Для того чтобы современный инженер овладел принципами и методами компьютерного моделирования необходимо прежде всего, чтобы он освоил ряд качественных математических понятий и классические численные методы, которые и составляют содержание этого учебного пособия. В курсах математики и информатики на начальном этапе обучения эти разделы не занимают подобающего им места, а существующая литература в своем большинстве ориентирована на специалистов по вычислительной математике, а не на студентов инженерных специальностей. Настоящее пособие адресовано в первую очередь студентам и аспирантам высших технических учебных заведений. В учебных программах БНТУ эти вопросы излагаются в курсе прикладной математики и рассчитаны на студентов общетехнических специальностей, которые намерены применять персональные компьютеры (PC) для решения прикладных задач. В настоящее время при математическом моделировании широко применяются универсальные и специализированные пакеты программ, такие как: MathCad, MatLab, Maple. Поэтому знание теоретических аспектов применяемых в этих пакетах алгоритмов является обязательным для их осмысленного и эффективного использования.

В учебном пособии содержатся следующие разделы:

1. Элементарная теория погрешностей, особенности машинной арифметики.

2. Представление о корректной постановке вычислительной задачи, ее обусловленности и устойчивости вычислительных алгоритмов.

3. Основные методы интерполяции и аппроксимации функций.

4. Приближенные численные методы решения задач Коши для обыкновенных дифференциальных уравнений.

5. Аналитические и численные методы решения краевых задач для обыкновенных дифференциальных уравнений и уравнений в частных производных.

Изложение ведется достаточно строго, но без излишней формализации. Для удобства читателя, используемые дополнительные свойства и теоремы приведены со ссылками на соответствующие литературные источники, в некоторых приведены более подробные и углубленные сведения, относящиеся к описанным в пособии вопросам. Содержание подкреплено значительным количеством детально разобранных примеров, среди которых преобладают задачи качественного характера с их компьютерной реализацией. Контрольные вопросы в каждом разделе направлены на закрепление теоретического материала, а приведенные задачи и упражнения нацелены на активизацию самостоятельной работы студентов.

Пособие не претендует на полноту освещения всех используемых в прикладных задачах численных методов, но в нем приведены, с нашей точки зрения, ряд тем, лежащих в основе аппарата математического моделирования.

I. ПРОСТЕЙШИЕ ПОНЯТИЯ ТЕОРИИ ПОГРЕШНОСТЕЙ. КОРРЕКТНОСТЬ И ОБУСЛОВЛЕННОСТЬ ВЫЧИСЛИТЕЛЬНОЙ ЗАДАЧИ. ОБУСЛОВЛЕННОСТЬ ЗАДАЧИ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ (ОБУСЛОВЛЕННОСТЬ МАТРИЦЫ)

1.1. Истоки и общая классификация погрешностей

При решении разнообразных научно-технических проблем практически всегда приходится оперировать с величинами, конструкциями и моделями, которые не всегда полностью отражают свойства реальных явлений и процессов или которые задаются неточно (т.е. с некоторыми погрешностями). Источниками возникновения погрешностей решения той или иной задачи могут являться самые разнообразные причины. При решении данных задач с использованием компьютерной техники можно выделить следующие три основные причины:

a) математическое описание задачи (математическая модель) по сути является приближенным, т.е. в математической модели не учтены существенные факторы, присущие ее реальному прообразу, или исходные данные заданы неточно;

b) используемый для решения метод является приближенным (т.е. для получения точного решения математической задачи нужно использовать бесконечное или неприемлемо большое число операций, что реально осуществить невозможно);

c) при вводе исходных данных в компьютер, при осуществлении арифметических операций и при выводе полученных результатов по тому или иному алгоритму производятся округления, усечения чисел.

Погрешности, соответствующие пп. *a*), *b*), *c*), обычно называют неустранимой погрешностью, погрешностью метода и вычислительной погрешностью. Погрешность, которая отражена в п. а), состоит из суммы двух частей. Первая часть порождается неточностью задания исходных числовых

данных, фигурирующих в математической модели. Вторая часть отражает несоответствие самой математической модели реальному прототипу.

Определение 1.1. Пусть Φ — точное значение величины, подлежащей отысканию; $\widetilde{\Phi}$ — значение этой величины, соответствующее выбранной математической модели, $\widetilde{\Phi}_h$ — значение величины (она соответствует $\widetilde{\Phi}$), получаемое посредством численного метода в рамках отсутствия погрешностей, вызванных округлениями, усечениями; $\widetilde{\Phi}_h^*$ — приближенное значение величины $\widetilde{\Phi}$, полученные при реальных вычислениях. Тогда под неустранимой погрешностью, погрешностью метода и вычислительной погрешностью понимают соответственно величины $\varepsilon_1 = \widetilde{\Phi} - \Phi,\ \varepsilon_2 = \widetilde{\Phi}_h - \widetilde{\Phi},\ \varepsilon_3 = \widetilde{\Phi}_h^* - \widetilde{\Phi}_h$.

Определение 1.2. Под полной погрешностью понимают величину $\varepsilon_0 = \widetilde{\Phi}_h^* - \Phi$, которая равна разности между реально получаемым и точным значениями величины. Эта погрешность удовлетворяет равенству $\varepsilon_0 = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$.

Замечание 1.1. При решении конкретных задач зачастую под погрешностью того или иного типа понимают не разности $\varepsilon_0, \varepsilon_1, \varepsilon_2, \varepsilon_3$, а определенные меры близости между ними. В частности, при рассмотрении скалярной величины Φ используют соответствующие абсолютные погрешности, определяемые соотношениями $\left|\widetilde{\Phi}_h^* - \Phi\right|$, $\left|\widetilde{\Phi} - \Phi\right|$, $\left|\widetilde{\Phi}_h - \widetilde{\Phi}\right|$, $\left|\widetilde{\Phi}_h^* - \widetilde{\Phi}_h\right|$. При этом имеет место неравенство $\left|\widetilde{\Phi}_h^* - \Phi\right| \le \left|\widetilde{\Phi} - \Phi\right| + \left|\widetilde{\Phi}_h - \widetilde{\Phi}\right| + \left|\widetilde{\Phi}_h^* - \widetilde{\Phi}_h\right|$.

Следует отметить, что при рассмотрении векторных и иных математических объектов меры близости могут конструироваться и более сложным образом по сравнению с абсолютными погрешностями, указанными в замечании 1.1.

1.2. Представление чисел в компьютерах

Множество действительных чисел R является несчетным и поэтому не может быть точно и полностью представлено в памяти компьютеров. В настоящее

время наиболее широко используются различные позиционные системы счисления.

Определение 1.3. Пусть x — некоторое конечное вещественное число, а p и q — конечные натуральные числа, причем $p \neq 1$. Представлением числа x в p-ичной позиционной системе счисления (p — основание этой системы) называется представление этого числа в форме

$$x = \pm\, p^q \sum_{l=1}^{+\infty} \alpha_\ell p^{-\ell} = \left(\pm 0, \alpha_1 \alpha_2 ...\right) \times p^q , \qquad (1.1)$$

где для $\forall \ell \in N$ имеет место неравенства $0 \le \alpha_\ell \le p-1$ ($\times$ — знак умножения).

Замечание 1.2. Наиболее часто используются позиционные системы счисления с такими основаниями: $p = 2$ (двоичная система счисления), $p = 3$ (троичная), $p = 8$ (восьмеричная), $p = 10$ (десятичная), $p = 16$ (шеснадцатиричная).

Замечание 1.3. Любое конечное вещественное число x может быть представлено в виде (1.1), причем для любого конечного $p \in N \setminus \{1\}$.

Из (1.1) следует, что в общем случае вещественное число x требует для своего представления использования бесконечной последовательности символов $\{\alpha_1, \alpha_2, ...\}$, каждый из которых является натуральным числом, лежащим между 0 и $p-1$. В рамках теоретических рассуждений это допустимо. Но при проведении конкретных вычислений невозможно использовать бесконечный набор символов $\{\alpha_1, \alpha_2, ...\}$ и выполнить бесконечное число алгебраических операций с числами из R. Это приводит к необходимости использования конечного множества вещественных чисел. При этом использование конечного набора таких чисел порождает как погрешности в представлении чисел, так и накопление погрешностей в процессе вычислений.

Сейчас почти на всех компьютерах наиболее широко используется процедура, задающая числа с плавающей точкой.

Определение 1.4. Под множеством F чисел с плавающей точкой понимаются все вещественные числа, представимые в виде

$$x = \pm\left(\frac{\alpha_1}{p} + \frac{\alpha_2}{p^2} + \ldots + \frac{\alpha_r}{p^r}\right) \times p^q, \qquad (1.2)$$

где p — основание $\left(p \in N \setminus \{1\}\right)$, q — показатель, r — точность. При этом для $\forall \ell = \overline{1,r}$ $0 \le \alpha_\ell \le p-1$ и $q \in [L,U] \subset R$, причем L и U — целые конечные числа.

Определение 1.5. Если для $\forall x \in F$ справедливо неравенство $\alpha_1 \neq 0$, то плавающая система F называется нормализованной. Целое число q называется при этом показателем степени (показателем), а число $\frac{\alpha_1}{p} + \ldots + \frac{\alpha_r}{p^r}$ — мантиссой или дробной частью.

Замечание 1.4. В ряде случаев числа с плавающей точкой представляют в виде $x = \left(\pm\alpha_1.\alpha_2\alpha_3\ldots\right) \times p^{q-1}$, причем мантиссой называют число $\pm\alpha_1.\alpha_2\alpha_3\ldots$, а $(q-1)$ — показателем.

Замечание 1.5. Конечное множество F содержит ровно $2(p-1)p^{q-1}(U-L+1)+1$ чисел.

1.3. Элементы теории погрешностей

1.3.1. Пусть x — точное значение некоторой скалярной величины, а $x^\times$ — известное приближение к ней. Тогда под абсолютной погрешностью приближенного числа $x^\times$ понимают величину $\Delta\left(x^\times\right) = \left|x - x^\times\right|$. Любое конечное число $\overline{\Delta}\left(x^\times\right)$, удовлетворяющее неравенству $\Delta\left(x^\times\right) \le \overline{\Delta}\left(x^\times\right)$ называют предельной абсолютной погрешностью числа $x^\times$. В силу того, что $\Delta\left(x^\times\right)$ зависит от выбора системы единиц зачастую используют относительную погрешность числа $x^\times$,

которая определяется по формуле $\delta\left(x^{\times}\right)=\left(\left|x^{\times}-x\right|/\left|x\right|\right)$. Относительная погрешность $\delta\left(x^{\times}\right)$ не зависит от выбора системы единиц. Любое конечное число $\overline{\delta}\left(x^{\times}\right)$, удовлетворяющее неравенству $\delta\left(x^{\times}\right)\le\overline{\delta}\left(x^{\times}\right)$ называют предельной относительной погрешностью.

Под значащими цифрами числа будем понимать все цифры в его записи в десятичной системе счисления, начиная с первой ненулевой слева. Например, в числах $x^{\times}=0.00\underline{40201}$, $x^{\times}=0.0\underline{3800900}$ значащими цифрами являются подчеркнутые цифры. При этом количество значащих цифр соответственно равны 5 и 7.

Если абсолютная погрешность числа не превосходит половины единицы разряда выбранной значащей цифры, то эта цифра называется верной. Например, для точного числа 58.43 число 58.40 является приближением только с тремя верными значащими цифрами, т.к. $\left|58.40-58.43\right|=0.03>\frac{1}{2}0.01$. Следует отметить, что приводимые в математических таблицах численные данные таковы, что все помещенные в них значащие цифры являются верными.

Если $x^{\times}$ является приближением к числу x с абсолютной погрешностью $\Delta\left(x^{\times}\right)$, то часто используют такую запись $x=x^{\times}\pm\Delta\left(x^{\times}\right)$, причем числа $x^{\times}$ и $\Delta\left(x^{\times}\right)$ записывают с одинаковым числом знаков после точки (запятой). Например, запись $x=3.458\pm0.002=3.458\pm2\cdot10^{-3}$ означает, что верно двойное неравенство $3.458-0.002\le x\le3.458+0.002$.

Если известна относительная погрешность $\delta\left(x^{\times}\right)$ числа $x^{\times}$, то зачастую пишут $x=x^{\times}\left(1\pm\delta\left(x^{\times}\right)\right)$. Эта запись означает ,что верно такое двойное неравенство $x^{\times}-x^{\times}\delta\left(x^{\times}\right)\le x\le x^{\times}+x^{\times}\delta\left(x^{\times}\right)$.

При приведении реальных вычислений приходится прибегать к процедуре округления чисел.

Правило округления [1].

Для округления числа до n значащих цифр, отбрасывают все цифры, стоящие справа от n-й значащей цифры, или, если это необходимо для сохранения разрядов заменяют их нулями. Кроме этого осуществляются следующие действия:

a) если первая из отброшенных цифр меньше 5, то оставшиеся десятичные знаки не изменяются;

b) если первая из отброшенных цифр больше 5, то к последней оставшейся цифре прибавляется единица;

c) если первая из отброшенных цифр равна 5 и среди других отброшенных цифр есть ненулевые, то последняя оставшаяся цифра увеличивается на единицу;

d) если же первая из отброшенных цифр равна 5 и все другие отброшенные цифры являются нулями, то последняя оставшаяся цифра сохраняется неизменной, если она четная, и увеличивается на единицу, если она нечетная (правило четной цифры).

Замечание 1.6. При применении правила округления погрешность процедуры округления не превосходит половины единицы десятичного разряда, определяемого последней оставленной значащей цифры.

Количество верных значащих цифр зависит от относительной погрешности числа.

Теорема 1.1. [1,2] Если положительное приближенное число $x^{\times}$ имеет n верных десятичных знаков, то его относительная погрешность $\delta\left(x^{\times}\right)$ не превосходит 10^{1-n}.

1.3.2. Приведем теперь оценки погрешностей чисел, полученных посредством арифметических операций над двумя приближенными числами с заданными погрешностями. Эти оценки описываются следующими теоремами [1]:

Теорема 1.2. Абсолютная погрешность суммы (разности) приближенных чисел не превышает суммы абсолютных погрешностей этих чисел;

Теорема 1.3. Если слагаемые имеют один знак, то предельная относительная погрешность их суммы не превышает наибольшей из предельных относительных погрешностей слагаемых.

Теорема 1.4. Относительная погрешность произведения приближенных чисел, отличных от нуля, не превышает суммы относительных погрешностей этих чисел.

Теорема 1.5. Относительная погрешность частного не превышает суммы относительных погрешностей делимого и делителя.

Замечание 1.7. Если приближенные числа $x_1^{\times}$ и $x_2^{\times}$ имеют малые абсолютные погрешности и число $\left|x_1^{\times} - x_2^{\times}\right|$ достаточно мало, то относительная погрешность разности этих двух приближенных чисел может быть весьма большой, т.е. происходит потеря точности.

1.3.3. При решении разнообразных теоретических и прикладных задач зачастую приходится находить погрешности вычисления функций одной или нескольких переменных.

Пусть в некотором замкнутом параллелепипеде $D \subset R_n$ задана дифференцируемая функция $u = f(x_i, \ldots, x_n)$ и пусть $\Delta(x_i)$, где $i = \overline{1,n}$, абсолютные погрешности, с которыми заданы аргументы этой функции. Тогда по определению абсолютная погрешность функции будет равна

$$\Delta(u) = \left| f(x_1 + \Delta x_1, \ldots, x_n + \Delta x_n) - f(x_1, \ldots, x_n) \right|, \qquad (1.3)$$

где $\Delta x_1, ..., \Delta x_n$ — приращения аргументов $x_1, ..., x_n$.

Из формулы Тейлора для функции нескольких переменных следует, что имеет место равенство

$$f(x_1+\Delta x_1,\ldots,x_n+\Delta x_n)-f(x_1,\ldots,x_n)=\sum_{i=1}^{n}\left.\frac{\partial f(\tilde{x}_1,\ldots,\tilde{x}_n)}{\partial\tilde{x}_i}\right|_{\substack{\tilde{x}_1=x_1+\theta\Delta x_1\\ \vdots\\ \tilde{x}_n=x_n+\theta\Delta x_n}}\Delta x_i, \qquad (1.4)$$

где $\theta\in(0,1)$. Из (1.3) и (1.4) в свою очередь следует, что справедливо неравенство

$$\Delta(u)\le\sum_{i=1}^{n}\mathrm{M}_i|\Delta x_i|=\sum_{i=1}^{n}\mathrm{M}_i\Delta(x_i). \qquad (1.5)$$

Здесь $\mathrm{M}_i=\max\limits_{\tilde{\bar{x}}\in G_1}\left|\frac{\partial f(\tilde{x}_1,\ldots,\tilde{x}_n)}{\partial\tilde{x}_i}\right|$; $\Delta(x_i)=|\Delta x_i|$; $\tilde{\bar{x}}=(\tilde{x}_1,\ldots,\tilde{x}_n)$; G_1 — замкнутый параллелепипед, принадлежащий G и определяемый такой системой неравенств: $x_1\le\tilde{x}_1\le x_1+\Delta x_1,\ldots,x_n\le\tilde{x}_n\le x_n+\Delta x_n$. Если абсолютные погрешности $\Delta(x_i)=|\Delta x_i|$ аргументов функции u для $\forall i=\overline{1,n}$ достаточно малы (уровень малости значений $|\Delta x_i|$ определяется сутью рассматриваемой задачи), то величины M_i в (1.5) можно заменить соответственно на модули значений частных производных $\frac{\partial f(x_1,\ldots,x_n)}{\partial x_i}$.

Пример 1.1. Рассмотрим задачу об оценке абсолютной погрешности вычисления определителя $\det A^{\times}$, где $A^{\times}$ — матрица, приближенно задающая точную матрицу A (будем считать, что A и $A^{\times}$ имеет размерность $n\times n$). При решении следует рассматривать определитель $\det A^{\times}$ как функцию n^2 переменных.

Пусть $\Delta(a_{ij}^{\times})$ — абсолютные погрешности задания элементов матрицы $A^{\times}=(a_{ij}^{\times})$, а $A_{ij}^{\times}$ — алгебраические дополнения матрицы $A^{\times}$ $(i,j\in\{1,2,\ldots,n\})$. Если все $\Delta(a_{ij}^{\times})$ достаточно малы, то из неравенства (1.5) и формулы $\det A^{\times}=\sum\limits_{i=1}^{n}a_{ij}^{\times}A_{ij}^{\times}$, где $i=\overline{1,n}$, следует справедливость такой оценки для абсолютной погрешности вычисления определителя $\det A^{\times}$:

$$\Delta\left(\det A^{\times}\right)=\left|\det A-\det A^{\times}\right|\le\sum_{i=1}^{n}\sum_{j=1}^{n}\left|A_{ij}^{\times}\right|\Delta\left(a_{ij}^{\times}\right). \qquad (1.6)$$

Соответствующая программа вычисления правой части неравенства (1.6) приведена в примере решения задания 1.1. (см. ниже).

1.4. Понятие о корректности и обусловленности вычислительной задачи

1.4.1. При решении практически важных задач важную роль играет ряд общих понятий и оценок, характеризующих качество постановки самой вычислительной задачи и точность численных методов, применяемых для ее решения. В связи с этим ниже будут кратко описаны некоторые из них.

Одним из важнейших требований, предъявляемых к разнообразным прикладным задачам, является корректность постановки математической (вычислительной) задачи.

Обозначим через X множество входных данных задачи, соответствующих некоторой предметной области. Соответственно через x будем обозначать конкретный набор таких данных $(x\in X)$. Пусть y - решение поставленной задачи, которое соответствует набору x. Всю совокупность такого рода решений обозначим через Y.

Определение 1.6. Вычислительная задача называется **корректной**, если выполняются следующие требования:

1^0 решение $y(y\in Y)$ этой задачи существует для $\forall x\in X$;

2^0 это решение единственно;

3^0 решение **устойчиво** по отношению к малым возмущениям (отклонениям) входных данных.

Задача называется **некорректной**, если **не выполнено** хотя бы одно из перечисленных условий.

Замечание 1.8. Для каждой вычислительной задачи выбираются адекватные ей критерии отклонений входных данных и решений друг от друга.

1.4.2. Исключительно важной характеристикой вычислительной задачи является **обусловленность**, под которой понимают ее «чувствительность» к малым отклонениям (погрешностям) входных данных. Дадим не совсем строгое, но отвечающее сути вопроса определение обусловленности (конкретизация этого понятия будет произведена далее).

Определение 1.7. Вычислительная задача называется хорошо обусловленной, если «малым» погрешностям (отклонениям) входных данных соответствуют «малые» погрешности (отклонения) решения, и плохо обусловленной, если возможны «сильные» изменения решения при «малых» отклонениях входных данных.

Для количественной оценки обусловленности задачи в прикладной математике используют **число обусловленности**. Пусть $\Delta(x)$ и $\Delta(y)$ - абсолютные погрешности, соответствующие набору входных данных x и самому решению y рассматриваемой задачи (конкретный смысл, вкладываемый в символ Δ определяется сутью задачи и методом ее решения).

Определение 1.8. Под абсолютным числом обусловленности будем понимать коэффициент ν_Δ, связывающий между собой $\Delta(y)$ и $\Delta(x)$ с помощью соотношения $\Delta(y)=\nu_\Delta\Delta(x)$. Если $\nu_\Delta >>1$, то будем говорить, что исходная задача **плохо абсолютно обусловлена**.

Зачастую для приложений гораздо важнее дать связь между относительными погрешностями $\delta(x)$ č $\delta(y)$, соответствующих входным данным и решению задачи.

Определение 1.9. Относительным числом обусловленности ν_δ называют коэффициент пропорциональности, входящий в соотношение $\delta(y)=\nu_\delta\delta(x)$.

Замечание 1.9. При решении прикладных задач часто удается дать только верхние границы $\overline{\nu_\Delta}, \overline{\nu_\delta}$ чисел обусловленности, которые удовлетворяют неравенствам $\nu_\Delta \le \overline{\nu_\Delta}$, $\nu_\delta \le \overline{\nu_\delta}$.

1.5. Обусловленность задачи решения систем линейных алгебраических уравнений

1.5.1. Перейдем к конкретизации общих понятий, введенных выше, для случая систем n линейных алгебраических уравнений с n неизвестными. Запишем данную систему в виде

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2, \\ \quad ..... \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = b_n. \end{cases} \tag{1.7}$$

Введем такие обозначения:

$$A = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{bmatrix}, \ \bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix}, \ \bar{b} = \begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_n \end{pmatrix}. \tag{1.8}$$

Будем предполагать, что основная матрица A системы (1.7) является невырожденной $(\det A \neq 0)$. Далее под символом $\bar{x}^\times = \left(x_1^\times, x_2^\times, ..., x_n^\times\right)^T$ будем понимать приближенное решение системы (1.7) (T — знак транспонирования).

Определение 1.10. Вектор-столбец $\bar{e} = \bar{x} - \bar{x}^\times$ будем называть вектором-столбцом погрешности решения системы (1.7), а вектор $\bar{r} = \bar{b} - A\bar{x}^\times$ - вектором-столбцом невязки (или просто невязкой).

Замечание 1.10. Погрешность решения и невязка могут сильно отличаться друг от друга.

Для количественной оценки погрешностей решения и невязки полезно ввести более простые числовые характеристики. В качестве таких характеристик в математике используются различного рода нормы.

Определение 1.11. Пусть L_n - n-мерное линейное (векторное) пространство. Будем говорить, что в этом пространстве задана норма $\|\bar{x}\|$, если выполняются следующие условия:

1^0 $\forall \bar{x} \in L_n$ сопоставлено единственное число $\|\bar{x}\| \in R$;

2^0 для $\forall \bar{x} \in L_n$ $\|\bar{x}\| \geq 0$, причем $\|\bar{x}\| = 0$ тогда и только тогда, когда $\bar{x} = \bar{0}$;

3^0 для $\forall \bar{x} \in L_n$ и $\forall \alpha \in R$ $\|\alpha \bar{x}\| = |\alpha| \, \|\bar{x}\|$;

4^0 для $\forall \bar{x}, \bar{y} \in L_n$ выполняется неравенство треугольника $\|\bar{x} + \bar{y}\| \leq \|\bar{x}\| + \|\bar{y}\|$.

В математике пользуются разнообразными нормами. Одними из наиболее употребительных в вычислительной практике являются такие нормы:

$$\|\bar{x}\|_1 = \sum_{i=1}^{n} |x_i|, \tag{1.9}$$

$$\begin{aligned} \|\bar{x}\|_2 &= \sqrt{\sum_{i=1}^{n} |x_i|^2}, \\ \|\bar{x}\|_\infty &= \max_{1 \leq i \leq n} |x_i|. \end{aligned} \tag{1.10}$$

При этом имеют место неравенства $\|\bar{x}\|_\infty \leq \|\bar{x}\|_2 \leq \|\bar{x}\|_1 \leq n\|\bar{x}\|_\infty$. Норму $\|\bar{x}\|_2$ называют евклидовой нормой.

С помощью понятия нормы вектора можно уточнить определения абсолютной и относительной погрешностей вектора.

Определение 1.12. Абсолютной и относительной погрешностями приближенно заданного вектора $\bar{x}^\times$ будем называть соответственно такие выражения:

$$\Delta\left(\bar{x}^{\times}\right)=\left\|\bar{x}-\bar{x}^{\times}\right\|,\quad \delta\left(\bar{x}^{\times}\right)=\frac{\left\|\bar{x}-\bar{x}^{\times}\right\|}{\left\|\bar{x}\right\|},\tag{1.11}$$

где $\bar{x}$ — точно заданный вектор.

На основе понятия нормы векторов можно ввести понятие и нормы квадратной матрицы A.

Определение 1.13. Нормой матрицы A размерности $[n\times n]$, подчиненной норме вектора $\bar{x}\in L_n$ называется вещественное число $\|A\|=\sup\limits_{\bar{x}\neq\bar{0}}\dfrac{\|A\bar{x}\|}{\|\bar{x}\|}$, где символ $\sup(...)$ имеет смысл точной верхней грани [3] множества всех значений, принимаемых величиной $\left(\|A\bar{x}\|/\|\bar{x}\|\right)$, когда $\bar{x}\neq\bar{0}$.

Замечание 1.11. Норма матрицы обладает такими свойствами:

1^0 $\|A\|\geq 0$, $\|A\|=0$ тогда и только тогда, когда A – нулевая матрица;

2^0 для $\forall\alpha\in R$ и любой квадратной матрицы A выполняется равенство $\|\alpha A\|=|\alpha|\,\|A\|$;

3^0 для любых квадратных матриц A и B, имеющих одинаковую размерность $[n\times n]$, справедливо неравенство $\|A+B\|\leq\|A\|+\|B\|$;

4^0 для матриц, определенных в п. 3^0, верно неравенство $\|AB\|\leq\|A\|\cdot\|B\|$;

5^0 имеет место $\|A\bar{x}\|\leq\|A\|\cdot\|\bar{x}\|$.

Замечание 1.12. Нормам $\|\bar{x}\|_1$, $\|\bar{x}\|_2$, $\|\bar{x}\|_\infty$ подчинены нормы $\|A\|_1$, $\|A\|_2$, $\|A\|_\infty$, которые задаются следующими выражениями:

$$\|A\|_1=\max_{1\leq j\leq n}\sum_{i=1}^{n}\left|a_{ij}\right|,$$

$$\|A\|_2=\max_{1\leq j\leq n}\sqrt{\lambda_j\left(A^TA\right)},\tag{1.12}$$

$$\|A\|_{\infty} = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|,$$

где $\lambda_j(A^T A)$ - собственные числа матрицы $A^T A$ (T — знак транспонирования матрицы).

Поскольку собственные числа матрицы зачастую находить сложно, то иногда используют оценку $\|A\|_2 \le \|A\|_E = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}$, где $\|A\|_E$ - евклидова норма матрицы A.

1.5.2. Запишем систему (1.7) в матричной форме

$$A\bar{x} = \bar{b}. \qquad (1.13)$$

Если нам неизвестны точные значения элементов матрицы A и вектора-столбца $\bar{b}$, то вместо (1.13) нам придется решать систему

$$A^{\times}\bar{x}^{\times} = \bar{b}^{\times}, \qquad (1.14)$$

где $A^{\times}$ и $\bar{b}^{\times}$ - приближенные аналоги A и $\bar{b}$. Если $\det A^{\times} \neq 0$, то система (1.14) будет иметь единственное решение $\bar{x}^{\times}$, которое, однако, будет **приближенным решением** системы (1.13). Если $\det A = 0$, то решение системы (1.14) не будет иметь смысла. Это следует из того, что только по информации о приближенно заданном вектору-столбцу $\bar{b}^{\times}$ нельзя установить разрешимость системы (1.13).

Имеет место

Теорема 1.6. Верна такая оценка относительной погрешности приближенного решения $\bar{x}^{\times}$ системы (1.13):

$$\delta(\bar{x}^{\times}) \le \operatorname{cond}(A)\left(\delta(\bar{b}^{\times}) + \delta(A^{\times})\right), \qquad (1.15)$$

где $\delta(\bar{b}^{\times})=\dfrac{\|\bar{b}-\bar{b}^{\times}\|}{\|\bar{b}\|}$, $\delta(A^{\times})=\dfrac{\|A-A^{\times}\|}{\|A\|}$ – соответственно относительные погрешности вектора $\bar{b}^{\times}$ и матрицы $A^{\times}$, а $\mathrm{cond}(A)=\|A^{-1}\|\|A\|$ - стандартное число обусловленности.

Замечание 1.13. Стандартное число обусловленности $\mathrm{cond}(A)$ дает верхнюю оценку относительного числа обусловленности, введенного в опр.1.9.

Везде далее будем называть $\mathrm{cond}(A)$ числом обусловленности задачи решения системы линейных алгебраических уравнений или просто числом обусловленности матрицы A.

Из (1.15) видно, что для получения приближенного решения системы (1.13) с хорошей точностью при достаточно малых $\delta(\bar{b}^{\times})$ и $\delta(A^{\times})$ надо, чтобы число $\mathrm{cond}(A)$ было не слишком большим. Если $\mathrm{cond}(A)>>1$, то задача (1.13) будет плохо обусловленной и для ее решения потребуется задавать $\bar{b}$ и A с очень высокой точностью, что не всегда возможно сделать.

Корректное решение любой конкретной системы линейных алгебраических уравнений (1.14), в которой основная матрица $A^{\times}$ и вектор-столбец свободных членов $\bar{b}^{\times}$ заданы приближенно, невозможно произвести без указания отрезка, в который могут попасть значения определителя $\det A^{\times}$. Если данному отрезку принадлежит нуль, то система (1.14) не будет иметь никакого разумного смысла, о чем уже частично говорилось выше. В этом случае следует ужесточить ограничения на точность, с которой должны задаваться элементы матрицы $A^{\times}$. При этом эта точность должна быть такой, чтобы указанный выше отрезок не содержал нуль. Это условие является **необходимым** условием корректности системы (1.14).

1.6. Контрольные вопросы

1. Какая вычислительная задача называется корректно поставленной?

2. Как определяются нормы векторов и матриц?

3. По каким формулам вычисляются абсолютные и относительные погрешности приближенно заданных векторов и матриц?

4. Что такое число обусловленности и как оно вычисляется для систем линейных алгебраических уравнений (СЛАУ)?

5. С помощью какого неравенства оценивается относительная погрешность решения СЛАУ?

6. Каким образом формулируется необходимое условие разрешимости приближенно заданной СЛАУ, содержащего n уравнений и n неизвестных?

7. Посредством каких формул можно оценить погрешности, допускаемые при вычислении функций многих переменных?

1.7. Практические задания и пояснения к ним. Компьютерный практикум

1.7.1. При решении системы линейных алгебраических уравнений (1.14) надо сначала проверить выполнение необходимого условия ее разрешимости.

Задание 1.1. Пусть элементы основной квадратной матрицы $A^{\times} = \left(a_{ij}^{\times}\right)$ системы (1.14) размерности $m \times m$ заданы неточно. Предположим, что известна матрица $W = \left(w_{ij}\right)$ абсолютных ошибок, которые допускаются при отыскании элементов матрицы $A^{\times}$. В силу этого для $\forall i, j = \overline{1,m}$ имеют место неравенства $a_{ij}^{\times} - w_{ij} \le a_{ij} \le a_{ij}^{\times} + w_{ij}$, где $A = \left(a_{ij}\right)$ — точная матрица, которую приближенно задает матрица $A^{\times}$. Требуется дать с помощью средств пакета Mathcad и неравенства (1.5) оценку абсолютной погрешности вычисления определителя $\det A^{\times}$, т.е. оценить сверху величину $\Delta\left(\det A^{\times}\right) = \left|\det A^{\times} - \det A\right|$. Кроме этого

следует найти значение $\det A^{\times}$ и сделать вывод о выполнении или невыполнении необходимого условия корректности системы (1.14) (в качестве такого условия выступает выполнение неравенства $\det A \neq 0$).

Перечень вариантов к заданию 1.1.

Таблица 1.1

Вариант	Матрица A	Матрица W абсолютных ошибок
N	$\begin{bmatrix} 0.023406 & N.134051 & 0.001111 & -N.093427 \\ -N.122221 & 5.333354 & 1.112223 & 2.223956 \\ 0.003459 & 5.667891 & -2.555613 & 0.111112 \\ 0.NN0000 & -1.126781 & 2.267101 & 3.334512 \end{bmatrix}$	$\begin{bmatrix} 1.0\times10^{-8} & 1.0\times10^{-6} & 1.0\times10^{-8} & 1.0\times10^{-7} \\ 1.0\times10^{-7} & 1.0\times10^{-7} & 1.0\times10^{-6} & 1.0\times10^{-9} \\ 1.0\times10^{-6} & 1.0\times10^{-8} & 1.0\times10^{-7} & 1.0\times10^{-6} \\ 1.0\times10^{-8} & 1.0\times10^{-6} & 1.0\times10^{-6} & 1.0\times10^{-7} \end{bmatrix}$

Пример решения задания типа 1.1.

Решение

1°. Вводим элементы исходной матрицы B $(B = A^{\times})$, когда $N = 0$.

$$\text{dat} := \begin{pmatrix} 0.023406 & 0.134051 & 0.001111 & -0.093427 \\ -0.122221 & 5.333354 & 1.112223 & 2.223956 \\ 0.003459 & 5.667891 & -2.555613 & 0.111112 \\ 0.000000 & -1.126781 & 2.267101 & 3.334512 \end{pmatrix}$$

$$\text{B} := \text{dat}$$

2°. Вводим элементы матрицы ошибок ε $(\varepsilon = W)$

$$\text{dat1} := \begin{pmatrix} 1.0\cdot10^{-8} & 1.0\cdot10^{-6} & 1.0\cdot10^{-8} & 1.0\cdot10^{-7} \\ 1.0\cdot10^{-7} & 1.0\cdot10^{-7} & 1.0\cdot10^{-6} & 1.0\cdot10^{-9} \\ 1.0\cdot10^{-6} & 1.0\cdot10^{-8} & 1.0\cdot10^{-7} & 1.0\cdot10^{-6} \\ 1.0\cdot10^{-8} & 1.0\cdot10^{-6} & 1.0\cdot10^{-6} & 1.0\cdot10^{-7} \end{pmatrix}$$

$\varepsilon := dat1$

3°. Задаем размерность матриц B и ε

$m := 4$

4°. Пишем программу вычисления миноров элементов матрицы B и их абсолютных величин.

$$I(n,k) := \left| \begin{array}{l} \text{for } i \in 0..n-1 \\ \quad \text{for } j \in 0..n-1 \\ \quad\quad m_{i,j} \leftarrow 1 \text{ if } i = j \\ \quad\quad m_{i,j} \leftarrow 0 \text{ otherwise} \\ \quad m_{k,k} \leftarrow 0 \\ m \end{array} \right.$$

$$T(n,k,l) := \left| \begin{array}{l} \text{for } i \in 0..n-1 \\ \quad \text{for } j \in 0..n-1 \\ \quad\quad m_{i,l} \leftarrow 1 \text{ if } i = k \\ \quad\quad m_{i,j} \leftarrow 0 \text{ otherwise} \\ m \end{array} \right.$$

$$AA(m,s,f) := \left|\left| I(m,s) \cdot B \cdot I(m,f) + T(m,s,f) \right|\right|$$

Здесь внутренний символ $|\cdots|$ означает операцию вычисления определителей, а вторая пара символов $|\cdot\cdot|$ — операцию отыскания их абсолютных величин (не путать со знаком нормы $\|\cdots\|$).

5°.Оценка абсолютной ошибки $\Delta(\det B)$ вычисления определителя:

$$\sum_{f=0}^{m-1} \sum_{u=0}^{m-1} (AA(m,u,f)) \cdot \varepsilon_{u,f} = 4.609 \times 10^{-6}$$

6°. Вычисление определителя матрицы B:

$$|B| = -1.337$$

7°. Решение вопроса о выполнении или невыполнении необходимого условия корректности системы (1.14). Необходимым условием корректности системы (1.14) является выполнение отношения:

$$0 \notin (\det B - \Delta(\det B), \det B + \Delta(\det B)).$$

Это отношение выполняется так как

$$0 \notin (-1.337 - 4.609 \times 10^{-6}, -1.337 + 4.609 \times 10^{-6}.$$

1.7.2. Выполнение необходимого условия корректности системы (1.14) еще не позволяет судить об эффективности использования того или иного численного алгоритма решения этой системы. Очень важной характеристикой системы (1.14) является число обусловленности $cond(A)$ точной матрицы A, которую и приближает матрица $A^{\times}$. Без оценки этого числа затруднительно дать оценку относительной погрешности полученного с помощью какого-либо численного алгоритма решения $\overline{x}^{\times}$ системы (1.14).

Задание 1.2. С использованием встроенной подпрограммы *norm*1 пакета Mathcad и определения $\mathrm{cond}(A)$ дать оценки числа обусловленности основных матриц систем линейных алгебраических уравнений, все коэффициенты которой заданы с абсолютной погрешностью 10^{-7}. Выяснить с какой абсолютной погрешностью надо задавать элементы матрицы $A^{\times}$, чтобы можно было судить об ее однозначной разрешимости (т.е. следует проверить выполнение необходимого условия корректности системы (1.14) (см. пример решения задания 1.1). Провести исследование вопроса о достаточности задания всех числовых коэффициентов системы с абсолютной погрешностью 10^{-7} для того, чтобы относительная погрешность решения данной системы не превосходила 10^{-3}. Если абсолютная погрешность задания коэффициентов системы недостаточна, то надо указать какова она должна быть, чтобы удовлетворить данному условию (имеется ввиду относительная погрешность решения, не превосходящая 10^{-3}). Решить систему методом исключения неизвестных. Освоить схему решения этого задания, приведенную ниже.

Перечень вариантов к заданию 1.2.

Таблица 1.2

Вариант	Матрица $A^{\times}$		Вектор правой части $\bar{b}^{\times}$
1	100.000000	1.000000	-0.030456
	1.000000	0.010001	1.302905
2	100.000000	2.000000	8.134213
	0.500000	0.01001	-5.670312
3	10.000000	3.000000	0.100000
	0.(3)	0.110000	-3.000121
4	100.000000	4.000000	0.000000
	0.250000	0.020000	-4.219135
5	10.000000	4.000000	5.000000
	0.250000	0.110000	-7.000000
6	100.000000	2.000000	-2.013456
	0.500000	-0.010000	1.111111
7	1000.000000	4.000000	-0.345067
	0.250000	-0.003000	2.112891
8	-100.000000	5.000000	-3.000000
	0.200000	- 0.040000	2.134012
9	150.000000	6.000000	-4.121510
	0.1(6)	0.140000	8.100051
10	10.000000	2.000000	-1.000000
	0.500000	-0.300000	1.111111
11	-20.000000	10.000000	-5.555555
	0.100000	-0.050500	6.112233
12	30.000000	100.000000	-5.600346
	0.010000	-0.0(3)	1.222222
13	0.250000	1000.000000	-3.078245
	0.001000	3.996000	0.103456
14	50.000000	0.(3)	-1.000000
	3.000000	-0.100000	2.111111

Вариант	Матрица $A^{\times}$		Вектор правой части $\bar{b}^{\times}$
15	1000.000000	100.000000	0.000000
	0.010000	0.101000	-1.222333
16	250.000000	33.(3)	8.000000
	0.030000	-0.012000	-2.314560
17	200.000000	0.200000	10.000000
	5.000000	-0.002500	-3.621456
18	1000.000000	0.400000	0.001235
	2.500000	0.005000	1.(3)
19	-500.000000	1.000000	-2.(1)
	1.000000	0.004000	6.345121
20	-100.000000	2.000000	-1.341212
	0.500000	0.090000	2.161321
21	10000.000000	0.100000	0.000000
	10.000000	-0.000200	-3.(2)
22	-100.000000	100.000000	1.(1)
	0.010000	-0.090000	6.(6)
23	300.000000	1000.000000	1.345126
	0.001000	-0.00(3)	-2.(6)
24	1000.000000	10.000000	-3.121126
	0.100000	0.000900	0.000000
25	100.000000	1000.000000	-3.(4)
	0.001000	0.010100	8.(8)
26	10000.000000	10.000000	-1.123023
	0.100000	0.000105	3.567812
27	100.000000	100.000000	-2.123456
	0.010000	0.010010	0.234567
28	1000000.000000	10.000000	0.000001
	0.100000	0.000003	-2.480056
29	100.000000	10.000000	2.(1)
	0.100000	-0.010000	-3.(7)
30	10000.000000	100000.000000	1.234567
	0.000010	-0.000050	8.920134

Пример решения задания типа 1.2.

Пусть надо решить систему линейных алгебраических уравнений

$$\begin{cases} 1001x_1 + \quad 20x_2 = 1930, \\ 0.05x_1 + 0.0011x_2 = -0.0967 \end{cases} \tag{1.16}$$

с относительной погрешностью, не превышающей 10^{-2}, в предположении, что все числовые коэффициенты, входящие в нее, заданы с абсолютной погрешностью, не превосходящей 10^{-6}.

Решение.

Вычислим непосредственным образом сначала определитель основной матрицы $A^{\times}$ системы (1.16). Имеем $\det A^{\times} = 1001 \cdot 0.0011 - 1 = 0.1011 \neq 0$. Но элементы матрицы A отличаются от элементов приближенной матрицы $A^{\times}$. Поэтому $\det A$ приближенно задается $\det A^{\times}$ с некоторой погрешностью $\Delta_1 = \det A - \det A^{\times}$. Оценим ее. Пусть

$$a_{11} = a_{11}^{\times} + \Delta_{11},\ a_{12} = a_{12}^{\times} + \Delta_{12},\ a_{21} = a_{21}^{\times} + \Delta_{21},\ a_{22} = a_{22}^{\times} + \Delta_{22},$$

где $A = (a_{ij}),\ A^{\times} = (a_{ij}^{\times})\ (i, j = \overline{1,2})$, а погрешности Δ_{ij} не превосходят по абсолютной величине числа 10^{-6}. С учетом сказанного получим:

$$\Delta_1 = (a_{11}^{\times} + \Delta_{11})(a_{22}^{\times} + \Delta_{22}) - (a_{21}^{\times} + \Delta_{21})(a_{12}^{\times} + \Delta_{12}) - (a_{11}^{\times}a_{22}^{\times} - a_{21}^{\times}a_{12}^{\times}) = a_{11}^{\times}a_{22}^{\times} + a_{11}^{\times}\Delta_{22} +$$
$$+ a_{22}^{\times}\Delta_{11} + \Delta_{11}\Delta_{22} - a_{21}^{\times}a_{12}^{\times} - a_{21}^{\times}\Delta_{12} - a_{12}^{\times}\Delta_{21} - \Delta_{21}\Delta_{12} - a_{11}^{\times}a_{22}^{\times} + a_{21}^{\times}a_{12}^{\times} = a_{11}^{\times}\Delta_{22} +$$
$$+ a_{22}^{\times}\Delta_{11} + \Delta_{11}\Delta_{22} - a_{21}^{\times}\Delta_{12} - a_{12}^{\times}\Delta_{21} - \Delta_{21}\Delta_{12}.$$

Следовательно, с учетом (1.16) и оценок $|\Delta_{ij}| \le 10^{-6}$ для $\forall i, j = \overline{1,2}$ получим, что абсолютную погрешность вычисления $\det A$ с помощью $\det A^{\times}$ можно оценить таким образом:

$$|\Delta_1| = \left|\det A - \det A^{\times}\right| \le 1001 \cdot 10^{-6} + 0.0011 \cdot 10^{-6} + 10^{-12} + 0.05 \cdot 10^{-6} +$$
$$+ 20 \cdot 10^{-6} + 10^{-12} \approx 10^{-3}.$$

Поскольку $\det A^{\times} = 0.1011$, то $\det A = \det A^{\times} + \Delta_1 = 0.1011 + \Delta_1 \neq 0$, ибо $|\Delta_1|$ существенно меньше числа 0,1011. Итак, установлено, что точная система линейных алгебраических уравнений, соответствующая (1.16) является невырожденной. Следовательно, выполнено необходимое условие корректности системы (1.16), т.е. она имеет единственное решение.

Перейдем к оценке числа обусловленности $\mathrm{cond}(A) = \|A\|_1 \cdot \left\|A^{-1}\right\|_1$ точной матрицы. При этом для определенности выберем норму $\|...\|_1$. Для этого нам надо произвести оценки для $\|A_1\|$ è $\left\|A^{-1}\right\|_1$ на основе знания норм $\|A^{\times}\|_1$ č $\left\|\left(A^{\times}\right)^{-1}\right\|_1$. Найдем эти нормы, исходя из их определения (см. ***замечание 1.12***) и (1.16).

Имеем для $\|A\|_1$ оценки

$$\|A\|_1 = \max_{1\le j\le 2}\sum_{i=1}^{2}\left|a_{ij}\right| = \max_{1\le j\le 2}\sum_{i=1}^{2}\left|a_{ij}^{\times} + \Delta_{ij}\right| \le$$
$$\le \max\left\{1001 + |\Delta_{11}| + 0.05 + |\Delta_{21}|,\ 20 + |\Delta_{12}| + 0.0011 + |\Delta_{22}|\right\} \le \qquad (1.17)$$
$$\le \max\left\{1001 + 10^{-6} + 0.05 + 10^{-6},\ 20 + 10^{-6} + 0.0011 + 10^{-6}\right\} = 1001.050002.$$

Для $\|A\|_1$ справедлива и такая нижняя оценка

$$\|A\|_1 = \max_{1\le j\le 2}\sum_{i=1}^{2}\left|a_{ij}^{\times} + \Delta_{ij}\right| \ge$$
$$\ge \max\left\{1001 - |\Delta_{11}| + 0.05 - |\Delta_{21}|,\ 20 - |\Delta_{12}| + 0.0011 - |\Delta_{22}|\right\} \ge \qquad (1.18)$$
$$\ge \max\left\{1001 - 10^{-6} + 0.05 - 10^{-6},\ 20 - 10^{-6} + 0.0011 - 10^{-6}\right\} = 1001.049998.$$

Итак, норма $\|A\|_1$ удовлетворяет такому двойному неравенству:

$$1001.049998 \le \|A\|_1 \le 1001.050002. \qquad (1.19)$$

Для получения оценок нормы $\|A^{-1}\|_1$ надо записать обратную матрицу A^{-1} и выразить ее через элементы матрицы $A^{\times}$. Справедливы соотношения

$$\begin{aligned} A^{-1} &= \frac{1}{\det A}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} = \\ &= \frac{1}{\det A^{\times} + \Delta_1}\begin{pmatrix} a_{22}^{\times} + \Delta_{22} & -a_{12}^{\times} - \Delta_{12} \\ -a_{21}^{\times} - \Delta_{21} & a_{11}^{\times} + \Delta_{11} \end{pmatrix} = \begin{pmatrix} \tilde{a}_{11} & \tilde{a}_{12} \\ \tilde{a}_{21} & \tilde{a}_{22} \end{pmatrix}. \end{aligned} \qquad (1.20)$$

С учетом этого выражения запишем норму $\|A^{-1}\|_1$ в виде

$$\begin{aligned} \|A\|_1 &= \max_{1\le j\le 2}\sum_{i=1}^{2}|\tilde{a}_{ij}| = \max\{|\tilde{a}_{11}| + |\tilde{a}_{21}|,\ |\tilde{a}_{12}| + |\tilde{a}_{22}|\} = \\ &= \frac{1}{|\det A^{\times} + \Delta_1|}\max\{|a_{22}^{\times} + \Delta_{22}| + |a_{21}^{\times} + \Delta_{21}|,\ |a_{12}^{\times} + \Delta_{12}| + |a_{11}^{\times} + \Delta_{11}|\}. \end{aligned} \qquad (1.21)$$

Из (1.17) и (1.22) следует, что

$$\begin{aligned} \|A^{-1}\|_1 &= \frac{1}{|0.1011 + \Delta_1|}\max\{|0.0011 + \Delta_{22}| + |0.05 + \Delta_{21}|,\ \ |20 + \Delta_{12}| + |1001 + \Delta_{11}|\} = \\ &= \frac{|1001 + \Delta_{11}| + |20 + \Delta_{12}|}{|0.1011 + \Delta_1|}. \end{aligned} \qquad (1.22)$$

Так как $|\Delta_{ij}| \le 10^{-6}$ для $\forall i, j = \overline{1,2}$, а величина $|\Delta_1|$ приближенно равна 10^{-3}, то из (1.22) следует то, что норма $\|A^{-1}\|_1$ будет удовлетворять неравенствам

$$\frac{1020.999998}{0.1011 + 0.0011} \le \|A^{-1}\|_1 \le \frac{1021.000002}{0.1011 - 0.0011}. \qquad (1.23)$$

Из (1.19) и (1.23) и определения $\mathrm{cond}(A)$(см. *теорему 1.6*) получим, что

$$\mathrm{cond}(A) \approx 1001 \cdot 1021 \cdot 10 >> 1. \tag{1.24}$$

Итак, система (1.16) является **плохо обусловленной**.

Теперь оценим относительную погрешность $\delta(A^{\times})$, с которой задана матрица $A^{\times}$ системы (1.16). По определению $\delta(A^{\times}) = \left(\|A - A^{\times}\|_1 / \|A\|_1 \right)$. С учетом введенных выше обозначений получим, что

$$\delta(A^{\times}) = \frac{\|(\Delta_{ij})\|_1}{\|A\|_1}. \tag{1.25}$$

Несложно убедиться, что норма $\|(\Delta_{ij})\|_1$ матрицы $B = (\Delta_{ij})$ не превосходит 0.000002. С другой стороны норма $\|A\|_1$ согласно (1.19) приближенно равна 1001. Следовательно $\delta(A^{\times}) \approx 2 \cdot 10^{-6} \cdot 10^{-3} = 2 \cdot 10^{-9}$. Оценим еще относительную погрешность $\delta(\bar{b}^{\times}) = \left(\|\bar{b} - \bar{b}^{\times}\|_1 / \|\bar{b}\|_1 \right)$ вектора-столбца свободных членов в (1.16). С учетом определения (см. (1.9)) нормы $\|..\|_1$ вектора и (1.16) найдем, что

$$\delta(\bar{b}^{\times}) \le \frac{2 \cdot 10^{-6}}{\|\bar{b}\|_1}. \tag{1.26}$$

При получении (1.26) учтено, что все числа, входящие в (1.16), заданы с абсолютной погрешностью, не превышающей 10^{-6}. Нам осталось дать оценку для $\|\bar{b}\|_1$. Из (1.9) и (1.16) получим, что

$$\begin{aligned} \|\bar{b}\|_1 = \sum_{i=1}^{2} |b_i| = |b_1| + |b_2| \ge |b_1^{\times}| - 10^{-6} + |b_2^{\times}| - 10^{-6} = 1930 + \\ + 0.0967 + 2 \cdot 10^{-6} = 1930.096698. \end{aligned} \tag{1.27}$$

Из (1.26) и (1.27) следует, что

$$\delta\left(\bar{b}^{\times}\right) \le \frac{2 \cdot 10^{-6}}{1930.096698} \approx 10^{-9}. \qquad (1.28)$$

С помощью сделанных оценок и *теоремы 1.6* уже можно оценить относительную ошибку, которая будет допущена при решении системы (1.16). Из (1.15), (1.24), соотношения $\delta\left(A^{\times}\right) \approx 2 \cdot 10^{-9}$ и неравенства (1.28), получим такую оценку:

$$\delta\left(\bar{x}^{\times}\right) \le 1001 \cdot 1021 \cdot 10\left(10^{-9} + 2 \cdot 10^{-9}\right) = 0.03066063. \qquad (1.29)$$

Из (1.29) видно, что задание коэффициентов системы (1.16) с абсолютной погрешностью 10^{-6} недостаточно для получения ее решения с относительной точностью 10^{-2}. Наилучшая относительная погрешность решения системы (1.16) будет составлять примерно 3%, а не 1%. Поэтому коэффициенты системы (1.16) для получения относительной погрешности меньшей 1% надо задавать с абсолютной погрешностью порядка 10^{-7}. Действительно в данном случае относительные погрешности $\delta\left(A^{\times}\right)$ и $\delta\left(\bar{b}^{\times}\right)$ не превысят соответственно чисел $0{,}199791 \cdot 10^{-9}$ и $0{,}103622 \cdot 10^{-9}$. Из этих оценок и соотношений (1.15), (1.24) получим неравенство $\delta\left(\bar{x}^{\times}\right) \le 1001 \cdot 1021 \cdot 10\left(0{,}199791 + 0{,}103622\right) \cdot 10^{-9} \approx 3 \cdot 10^{-3}$, т.е. относительная погрешность решения системы (1.16) будет меньше, чем 1 %, что и доказывает верность высказанного выше утверждения.

Перейдем непосредственно к решению системы (1.16) методом исключения неизвестных в предположении, что все ее коэффициенты заданы с абсолютной погрешностью 10^{-6}. Умножим первое уравнение системы (1.16) на (-(0,05/1001)) и сложим полученное уравнение со вторым уравнением системы (1.16). В итоге получим соотношение

$$\left(-\frac{0.05\cdot 20}{1001}+0.0011\right)x_2=-\frac{0.05\cdot 1930}{1001}-0.0967\,. \qquad (1.30)$$

Из (1.30) следует, что $x_2^{\times}=-1911.916832$, где $x_2^{\times}$ — приближенное решение уравнения (1.30). В свою очередь из первого уравнения системы (1.16) найдем $x_1^{\times}=\left(\left(1930-20x_2^{\times}\right)/1001\right)=40.128208$. Данному решению соответствует вектор-столбец невязки $\bar{r}=(r_1,r_2)^T$ с компонентами

$$r_1=1930-1001\cdot x_1^{\times}-20x_2^{\times}=0.000026,$$
$$r_2=-0.0967-0.005\cdot x_1^{\times}-0.0011\cdot x_2^{\times}=-0.000002.$$

II. МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

2.1. Вводные замечания

Одной из основных задач линейной алгебры является численное решение системы линейных алгебраических уравнений (СЛАУ), выписанной ранее в развернутом виде (1.7) и в матричной форме (1.14). В разделе 1 были изложены теоретические понятия, необходимые для получения качественных и количественных оценок решений такого рода систем, когда ее основная матрица A и вектор-столбец свободных членов $\bar{b}$ заданы неточно. Ниже в сжатой форме будут описаны некоторые из наиболее известны и эффективных методов решения системы (1.14), а также будут приведены варианты заданий для студентов, которые должны будут выполнены ими на практических или лабораторных занятиях.

Основные методы численного решения СЛАУ можно разбить на прямые (точные) и итерационные. Прямыми методами называют методы, с помощью которых возможно за конечное число действий получить точное решение СЛАУ. Термин «точное решение» следует понимать как характеристику алгоритма получения решения, а не реального процесса вычислений. Это означает, что прямые методы дают точное решение, если все известные величины, фигурирующие в СЛАУ, заданы точно и все вычисления проводятся абсолютно точно. Под итерационными методами понимают методы, основанные на построении итерационной последовательности приближенных решений СЛАУ, которая сходится к точному решению. При этом за конечное число шагов (итераций) можно получить приближенное решение СЛАУ с любой наперед заданной точностью (при этом само число шагов зависит от этой точности).

Следует отметить, что выбор того или иного метода решения СЛАУ зависит от особенностей структуры матрицы A, порядка системы и характеристик компьютера (т.е. его архитектуры, операционной системы быстродействия,

памяти и программного обеспечения). Кроме того при этом выборе важную роль играют требования, предъявляемые к точности решения СЛАУ (1.7).

2.2. Прямые методы

2.2.1. К прямым методам относится метод Гаусса [1,5,6], специфическими вариантами которого являются методы прогонки (и матричной прогонки) [7]. К такого рода методам относятся также метод Холецкого (метод квадратных корней) [1,5] и методы вращений, отражений [5,8]. Методы Гаусса, вращений и отражений представляют собой методы исключения неизвестных. Важным свойством методов вращений и отражений является то, что они обладают гарантированной хорошей обусловленностью (см. раздел 1). Заметим еще, что метод Холецкого используется для решения СЛАУ (1.7), в которой матрица A является самосопряженной и положительно определенной, определения которых даны ниже.

Определение 2.1. Квадратная матрица $A=\left(a_{ij}\right)$, где $i,j=\overline{1,n}$, называется самосопряженной, если $A=A^*$, т.е. матрица A совпадает со своей сопряженной матрицей $A^*=\left(\widetilde{a_{ij}}\right)$ (для $\forall i,j=\overline{1,n}\ \ \tilde{a}_{ij}=\bar{a}_{ji}$, где черта сверху означает операцию комплексного сопряжения).

Определение 2.2. Квадратная матрица A называется положительно определенной, если для любого ненулевого вектор-столбца $\bar{x}=\left(x_1,x_2,...,x_n\right)^T$ выполнено неравенство $\left(A\bar{x},\bar{x}\right)>0$, где символ (…,…) имеет смысл скалярного произведения.

2.2.2. В современных программах, реализующих метод Гаусса без перестановок (т.е. без выполнения операций перестановок уравнений) на компьютерах, вычисления разбивают на два главных этапа. Первый, в основном состоит в построении LU -разложения основной матрицы A системы (1.13), т.е. по сути строится представление матрицы A в виде

$$A = LU\,, \qquad (2.1)$$

где L — нижняя треугольная, а U — верхняя треугольная матрицы (обе эти матрицы невырождены и имеют ту же размерность, что и матрица A). Для получения LU-разложения приходится затрачивать примерно $(2/3)n^3$ арифметических операций. Из (2.1) следует, что систему (1.13) можно переписать в виде

$$U\bar{x} = L^{-1}\bar{b} = \bar{c}\,. \qquad (2.2)$$

После отыскания матриц L, U, L^{-1} находят вектор-столбец $\bar{c}$. На втором этапе решают систему (2.2) с помощью обратных подстановок, т.е. из последнего n-го уравнения системы (2.2) находят $x_n = (c_n / u_{nn})$, где $\bar{c}^{\,T} = (c_1, c_2, ..., c_n)$, $U = (u_{ij})$, затем из $(n-1)$-го уравнения системы (2.2) находят x_{n-1} и т.д. Для получения решения $\bar{x}$ системы (1.13) на втором этапе нужно сделать около $2n^2$ арифметических операций. Следует отметить, что реализуемость LU-разложения основана на такой теореме [1].

Теорема 2.1. Если все главные миноры матрицы A отличны от нуля, то существуют единственные нижняя треугольная матрица L и верхняя треугольная матрица U такие, что $A = LU$.

2.2.3. Следует отметить, что кратко описанный выше вариант метода Гаусса без перестановок зачастую называют схемой единственного деления. Однако при проведении вычислений на основе этой схемы и наличия ошибок округления (усечения) имеет место процесс потери точности, порождаемый малыми значениями выбранных в этом варианте метода Гаусса ведущих (главных) элементов матрицы A (в качестве таковых выступают ее диагональные по предположению ненулевые элементы $a_{11}, a_{22}, ..., a_{nn}$). С целью противодействия процессу потери точности (накопления ошибок) была разработана более

эффективная и надежная модификация метода Гаусса, а именно метод Гаусса с выбором главных элементов по столбцу. В отличие от схемы единственного деления данная схема включает в себя одновременно процедуры перестановок уравнений системы и разложения типа LU-разложения, но не матрицы A, а матрицы $\widetilde{A}$, полученной из A в результате соответствующей перестановки ее строк. В остальном же этот вариант метода Гаусса аналогичен методу Гаусса без перестановок.

Кратко опишем основные идеи, лежащие в основе метода Гаусса с выбором главных элементов по столбцу. Перепишем систему (1.7) в виде

$$\begin{cases} a_{11}^{(0)}x_1 + a_{12}^{(0)}x_2 + a_{13}^{(0)}x_3 + ... + a_{1n}^{(0)}x_n = b_1^{(0)}, \\ a_{21}^{(0)}x_1 + a_{22}^{(0)}x_2 + a_{23}^{(0)}x_3 + ... + a_{2n}^{(0)}x_n = b_2^{(0)}, \\ \ldots\ldots\ldots\ldots \\ a_{n1}^{(0)}x_1 + a_{n2}^{(0)}x_2 + a_{n3}^{(0)}x_3 + ... + a_{nn}^{(0)}x_n = b_n^{(0)}, \end{cases} \qquad (2.3)$$

где для $\forall i, j = \overline{1,n}$ $a_{ij}^{(0)} = a_{ij}$ и $b_i^{(0)} = b_i$.

Теперь выберем из элементов $a_{11}^{(0)}, a_{21}^{(0)}, ..., a_{n1}^{(0)}$ первого столбца матрицы A наибольший по модулю элемент. Пусть это будет элемент $a_{i_1 1}^{(0)}$, где $i_1 \in \{1, 2, ..., n\}$. Данный элемент называют главным (ведущим) элементом первого столбца матрицы A (первого шага алгоритма). При этом в силу невырожденности матрицы A имеет место неравенство $a_{i_1 1} \neq 0$. Следующим действием является перестановка первого и i_1-го уравнений системы (2.3). В итоге система (2.3) преобразуется к виду $A_1 \bar{x} = \bar{b}_1$, где A_1 и $\bar{b}_1$ — матрица и вектор-столбец, полученные соответственно из A и $\bar{b}$ посредством перестановки в них первых и i_1-ой строк. Если последовательно исключить переменную x_1 из второго, третьего и т.д. уравнений данной системы по обычной схеме исключения (т.е. посредством

последовательного умножения первого уравнения этой системы на величины $\left(-\left(a_{i1}/a_{i_1 1}\right)\right)$, где $i \in \{1,2,...,i_1-1,i_1+1,...,n\}$, и последующего сложения полученных таким образом уравнений со вторым, третьим и т.д. уравнениями указанной системы). В итоге описанных выше действий первого шага алгоритма метода Гаусса с выбором главных элементов по столбцу система (2.3) приведется к такому виду:

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + ... + a_{1n}^{(1)}x_n = b_1^{(1)}, \\ \qquad a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + ... + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \ldots\ldots\ldots\ldots \\ \qquad a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + ... + a_{nn}^{(1)}x_n = b_n^{(1)}, \end{cases} \qquad (2.4)$$

где для $a_{11}^{(1)} = a_{i_1 1} \neq 0$, $a_{12}^{(1)} = a_{i_1 2}$, …, $a_{1n}^{(1)} = a_{i_1 n}$, $b_1^{(1)} = b_{i_1}^0$.

Совокупность всех уравнений системы (2.4), которая не содержит ее первого уравнений, сама является СЛАУ относительно неизвестных $x_1, x_2, ...x_n$. Данную СЛАУ можно преобразовать с помощью операций, использованных при приведении системы (2.3) к виду (2.4). Для этого выберем ведущий элемент второго шага алгоритма, под которым будем понимать максимальный по модулю элемент, принадлежащий первому столбцу (он получен из второго столбца матрицы A) матрицы

$$\begin{pmatrix} a_{22}^{(1)} & a_{23}^{(1)} ... a_{2n}^{(1)} \\ a_{32}^{(1)} & a_{33}^{(1)} ... a_{3n}^{(1)} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n2}^{(1)} & a_{n3}^{(1)} ... a_{nn}^{(1)} \end{pmatrix}.$$

Пусть этим элементом является элемент $a_{i_2 2}^{(1)}$, где $i_2 \in \{2,3,...,n\}$. Заметим, что в силу невырожденности исходной матрицы A имеет место неравенство $a_{i_2 2}^{(1)} \neq 0$. Теперь переставим второе уравнение системы (2.4) с i_2-ым уравнением этой же системы, а затем по аналогии с первым шагом алгоритма исключим переменную x_2 из полученной указанным выше образом СЛАУ порядка $(n-1)$. В итоге описанных преобразований с учетом первого уравнения в (2.4) приведем систему (2.4) к форме

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + ... + a_{1n}^{(1)}x_n = b_1^{(1)}, \\ \qquad a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + ... + a_{2n}^{(2)}x_n = b_2^{(2)}, \\ \qquad\qquad a_{33}^{(2)}x_3 + ... + a_{3n}^{(2)}x_n = b_3^{(2)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad a_{n3}^{(2)}x_3 + ... + a_{nn}^{(2)}x_n = b_n^{(2)}, \end{cases} \qquad (2.5)$$

причем имеют место следующие равенства: $a_{22}^{(2)} = a_{22}^{(1)}, ..., a_{2n}^{(2)} = a_{2n}^{(1)}, b_2^{(2)} = b_2^{(1)}$.

Повторяя аналогичным образом последующие шаги алгоритма приведем исходную систему (2.3) к следующему виду:

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + ... + a_{1n}^{(1)}x_n = b_1^{(1)}, \\ \qquad a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + ... + a_{2n}^{(2)}x_n = b_2^{(2)}, \\ \qquad\qquad a_{33}^{(2)}x_3 + ... + a_{3n}^{(2)}x_n = b_3^{(2)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad\qquad\qquad a_{nn}^{(n)}x_n = b_n^{(n)}. \end{cases} \qquad (2.6)$$

В системе (2.6) все ведущие элементы $a_{11}^{(1)}, a_{22}^{(2)}, ..., a_{nn}^{(n)}$ отличны от нуля. Приведение исходной системы (2.4) к виду (2.6) завершает прямой ход алгоритма Гаусса с выбором главных элементов по столбцу. Фактически целью прямого хода этого алгоритма было приведение исходной системы (2.3) к треугольному виду.

Вторым этапом данного алгоритма является решение треугольной системы (2.6). Из последнего уравнения (2.6) находится x_n, затем из предпоследнего уравнения находится x_{n-1} и т.д.

Следует подчеркнуть, что в отличие от формул Крамера, представляющих, в основном, теоретический интерес (при его прямом использовании необходимо совершить более $n!$ арифметических операций, что практически невозможно сделать при $n \gg 1$) описанный вариант метода Гаусса требует совершения при прямом ходе примерно $\left(n^2/3\right) + O\left(n^2\right)$ операций сложения, такое же количество операций умножения и $\left(n^2/2\right) + O(n)$ операций деления. Обратный же ход требует $\left(n^2/2\right) + \underline{\underline{O}}(n)$ операций сложения, столько же операций умножения и n операций деления. Сказанное выше указывает на практическую реализуемость метода Гаусса с выбором главных элементов по столбцу.

Отметим, что на практике широко используется и метод Гаусса с выбором главных элементов по всей матрице (схема полного выбора). В данной схеме допускается нарушение естественного порядка неизвестных, т.е. в этом варианте метода Гаусса используют еще операции перестановок этих неизвестных.

2.3. Метод прогонки

Данный метод является специфическим вариантом метода Гаусса, который был специально разработан для решения СЛАУ с трехдиагональной основной матрицей высокого порядка. Данную систему можно представить в виде

$$
\begin{cases}
b_1x_1 + c_1x_2 & = d_1, \\
a_2x_1 + b_2x_2 + c_2x_3 & = d_2, \\
\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots & \\
a_{n-1}x_{n-2} + b_{n-1}x_{n-1} + c_{n-1}x_n & = d_{n-1}, \\
a_nx_{n-1} + b_nx_n & = d_n.
\end{cases}
\qquad (2.7)
$$

Здесь $x_1, x_2, ..., x_n$ - искомые величины, а множества коэффициентов $\{b_1,...,b_n\}$, $\{c_1,...,c_{n-1}\}$, $\{a_2,...,a_n\}$, $\{d_1,...,d_n\}$ считаются заданными. Необходимость решения систем вида (2.7) возникает при рассмотрении различных задач математической физики (в частности такого рода системы приходится использовать при применении конечно-разностных методов решения краевых задач [6,7]). Следует отметить, что для получения решений многих важных прикладных проблем приходится с достаточной точностью решать системы (2.7), которые содержат сотни, тысячи, десятки тысяч и т.д. уравнений. Поэтому метод прогонки для такого рода проблем реализуем только с использованием современной компьютерной техники.

Опишем суть метода правой прогонки, который разбивается на два этапа. В этом методе сначала надо осуществить **прямой ход прогонки** («слева направо»), а затем произвести **обратный ход** («справа налево»).

Прямой ход. Из первого уравнения системы (2.7) выразим x_1 через x_2. Имеем

$$
x_1 = \frac{d_1 - c_1x_2}{b_1} = \frac{d_1}{b_1} - \frac{c_1}{b_1}x_2 = \alpha_1x_2 + \beta_1. \qquad (2.8)
$$

Подставим это выражение во второе уравнение системы (2.7). Получим

$$
a_2(\alpha_1x_2 + \beta_1) + b_2x_2 + c_2x_3 = d_2. \qquad (2.9)
$$

Из (2.9) находим

$$x_2 = -\frac{c_2}{b_2 + a_2\alpha_1}x_3 + \frac{d_2 - a_2\beta_1}{b_2 + a_2\alpha_1} = \alpha_2 x_3 + \beta_2. \qquad (2.10)$$

Данное соотношение для x_2 следует подставить в третье уравнение системы (2.7) и повторять эту процедуру и далее. На i-м шаге данного процесса $(1 < i < n)$ i-е уравнение системы приводится к тому же самому виду (см.(2.8) и (2.9)):

$$x_i = \alpha_i x_{i+1} + \beta_i, \qquad (2.11)$$

где

$$\alpha_i = -\frac{c_i}{b_i + a_i\alpha_{i-1}}, \quad \beta_i = \frac{d_i - a_i\beta_{i-1}}{b_i + a_i\alpha_{i-1}}. \qquad (2.12)$$

Осуществим теперь последний шаг прямого хода. Подставим $x_{n-1} = \alpha_{n-1}x_n + \beta_{n-1}$ (см. (2.11)) в последнее уравнение. Тогда получим

$$a_n\left(\alpha_{n-1}x_n + \beta_{n-1}\right) + b_n x_n = d_n. \qquad (2.13)$$

Из (2.13) найдем

$$x_n = \frac{d_n - a_n\beta_{n-1}}{b_n + a_n\alpha_{n-1}} = \beta_n. \qquad (2.14)$$

Итак, прямой ход прогонки фактически состоит в отыскании прогоночных коэффициентов, под которыми понимаются числа $\alpha_i, \gamma_i, \beta_i$. Эти коэффициенты вычисляются посредством таких рекуррентных формул:

$$\begin{cases} i=1: \ \alpha_1=-\dfrac{c_1}{\gamma_1}, \ \beta_1=\dfrac{d_1}{\gamma_1}, \ \gamma_1=b_1, \\ i=\overline{2,n-1}: \ \alpha_i=-\dfrac{c_i}{\gamma_i}, \ \beta_i=\dfrac{d_i-a_i\beta_{i-1}}{\gamma_i}, \ \gamma_i=b_i+a_i\alpha_{i-1}, \\ i=n: \ \beta_n=\dfrac{d_n-a_n\beta_{n-1}}{\gamma_n}, \ \gamma_n=b_n+a_n\alpha_{n-1}. \end{cases} \tag{2.15}$$

Обратный ход. После отыскания прогоночных коэффициентов можно найти значения неизвестных величин $x_1,...,x_n$ с помощью следующих формул:

$$\begin{aligned} &x_n=\beta_n, \\ &x_i=\alpha_i x_{i+1}+\beta_i \quad i=n-1,\ n-2,\ ...,\ 2,\ 1. \end{aligned} \tag{2.16}$$

Замечание 2.1. Для реализации вычислений по методу правой прогонки требуется осуществить приблизительно $8n$ арифметических операций. При этом для хранения матрицы коэффициентов системы требуется только $(3n-2)$ машинных слова.

При практическом использовании метода правой прогонки полезно принимать во внимание такое утверждение [7].

Теорема 2.2. Пусть коэффициенты системы (2.7) действительны и удовлетворяют таким условиям:

$|c_1|\geq 0, \ |a_n|\geq 0, \ |b_1|>0, \ |b_n|>0;$ для $\forall i=\overline{2,n-1} \ |a_i|>0, \ |c_i|>0;$

для $\forall i=\overline{2,n-1} \ |b_i|\geq|a_i|+|c_i|; \ |b_1|\geq|c_1|, \ |b_n|\geq|a_n|,$

причем хотя бы в одном из последних неравенств выполняется строгое неравенство. Тогда для алгоритма правой прогонки для $\forall i=\overline{1,n}$ имеют место неравенства $\gamma_i\neq 0, \ |\alpha_i|\leq 1$, гарантирующие корректность и устойчивость метода.

2.4. Метод простых итераций (метод Якоби)

Если система линейных алгебраических уравнений имеет высокий порядок $(n >> 1)$, а ее основная матрица не является трехдиагональной, то использование прямых методов решения такого рода систем не всегда оправдано или невозможно. В этом случае зачастую используют итерационные алгоритмы, которые позволяют, в принципе, получать решения с нужной точностью.

Ниже рассмотрен наиболее простой из методов такого типа, а именно метод Якоби. Пусть задана система линейных алгебраических уравнений $A\bar{x} = \bar{b}$, где A - невырожденная квадратная матрица, для которой $a_{ii} \neq 0$, когда $i = \overline{1,n}$. Такую систему можно преобразовать к виду

$$\bar{x} = B\bar{x} + \bar{c}, \tag{2.17}$$

где B - квадратная матрица той же размерности, что и A, а $\bar{c}$ - вектор-столбец. Распишем (2.17) в развернутом виде

$$\begin{cases} x_1 = \quad b_{12}x_2 + b_{13}x_3 + \ldots + b_{1n-1}x_{n-1} + b_{1n}x_n + c_1, \\ x_2 = b_{21}x_1 \quad + b_{23}x_3 + \ldots + b_{2n-1}x_{n-1} + b_{2n}x_n + c_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = b_{n1}x_1 + b_{n2}x_2 + b_{n3}x_3 + \ldots + b_{nn-1}x_{n-1} \quad + c_n. \end{cases} \tag{2.18}$$

На главной диагонали матрицы B стоят нули, а остальные элементы равны $b_{ij} = -\left(a_{ij}/a_{ii}\right) \quad \left(i, j = \overline{1,n},\ \text{í î}\ \ i \neq j\right)$. Соответственно компоненты вектора-столбца $\bar{c}$ находятся по формулам $c_i = \left(b_i/a_{ii}\right)$.

Итерационная схема метода Якоби строится на основе рекуррентного соотношения

$$\bar{x}^{(k+1)} = B\bar{x}^{(k)} + \bar{c} \quad \left(k = 0, 1, 2, \ldots\right), \tag{2.19}$$

причем $\overline{x}^{(k)}$ - k-е приближение к точному решению системы (2.17). При этом $\overline{x}^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(n)} \right)^T$ - некоторое нулевое приближение, выбор которого влияет на точность приближенного решения, полученного посредством конечного числа r итераций. При определенных ограничениях имеет место равенство $\lim\limits_{k \to +\infty} \overline{x}^{(k)} = \overline{x}$, где $\overline{x}$ - точное решение системы (2.18).

Справедлива

Теорема 2.3. [2,5] Пусть $\|B\| < 1$, где $\|B\|$ - некоторая норма матрицы B, подчиненной норме $\|\overline{x}\|$ вектор-столбца $\overline{x}$. Тогда существует единственное решение $\overline{x} = \overline{x}_{\text{òî÷í}}$ системы (2.17), причем при любом начальном приближении $\overline{x}^{(0)}$ метод Якоби сходится и справедлива оценка погрешности

$$\left\| \overline{x}^{(k)} - \overline{x}_{\text{òî÷í}} \right\| \le \|B\|^k \left\| \overline{x}^{(0)} - \overline{x}_{\text{òî÷í}} \right\|. \qquad (2.20)$$

Из (2.20) видно, что скорость сходимости и точность метода Якоби существенно зависят от того насколько $\|B\|$ меньше единицы, а также от величины нормы $\left\| \overline{x}^{(0)} - \overline{x}_{\text{òî÷í}} \right\|$, значение которой определяется выбором нулевого приближения $\overline{x}^{(0)}$. Оценка (2.20) является априорной.

При решении конкретных систем линейных алгебраических уравнений нам, вообще неизвестно точное ее решение. Поэтому приходится делать апостериорные оценки получаемых приближенных решений (т.е. оценки, отыскиваемые на основе уже проведенных расчетов).

Теорема 2.4. [2] Если $\|B\| < 1$, то справедлива оценка

$$\left\| \overline{x}^{(k)} - \overline{x}_{\text{òî÷í}} \right\| \le \frac{\|B\|}{1-\|B\|} \left\| \overline{x}^{(k)} - \overline{x}^{(k-1)} \right\| \qquad (2.21)$$

$k = 1,\ 2,\ 3, \ldots$

2.5. Контрольные вопросы

1. Каковы основные этапы решения системы линейных алгебраических уравнений с помощью метода правой прогонки?

2. Какие условия надо наложить на элементы основной матрицы системы линейных алгебраических уравнений, чтобы гарантировать корректность и устойчивость метода правой прогонки?

3. Какой смысл имеет вектор невязки?

4. При каких условиях применим методы простой итерации (метод Якоби) сходится?

5. Каким образом делается апостериорная оценка погрешности решения системы линейных алгебраических уравнений, получаемого с помощью метода Якоби?

2.6. Практические задания и пояснения к ним. Компьютерный практикум

Задание 2.1. С помощью метода правой прогонки непосредственным образом найти решение системы (2.7), когда $n = 3$, причем при расчетах использовать арифметику с шестью значащими цифрами (схема непосредственного решения такого рода систем проиллюстрирована в примере 2.1.). Применяя средства пакета MathCad и рекуррентные формулы (2.15) и (2.16) решить систему линейных алгебраических уравнений (2.7) с трехдиагональной основной матрицей, когда $n = 50$. Найти вектор-невязки (см. пример 2.2.).

Перечень вариантов к заданию 2.1.

Таблица 2.1

Вариант	Элементы матрицы A системы, стоящие на главной диагонали, поддиагонали и наддиагонали	Вектор правой части $\bar{b}$
N	$\forall i=\overline{1,n}\quad b_i=\dfrac{i+2+N}{i+1+N};$ $\forall i=\overline{2,n}\quad a_i=\dfrac{i}{2(i+1)N};$ $\forall i=\overline{1,n-1}\quad c_i=\dfrac{i}{3(i+1)N};$ $n=50.$	$\forall i=\overline{1,n}$ $d_i=(-1)^i\dfrac{i}{N}$

Пример 2.1. Методом правой прогонки решить систему

$$\begin{cases} x_1 - x_2 = 0.8, \\ x_1 + 4x_2 + x_3 = 3.6, \\ x_2 + 4x_3 + x_4 = -1.2, \\ x_3 + 4x_4 + x_5 = 3.6, \\ x_4 - x_5 = -0.8, \end{cases} \tag{2.22}$$

в которой все элементы основной матрицы заданы точно, а компоненты вектора-столбца свободных членов — с абсолютной погрешностью, не превышающей 10^{-6}.

После отыскания решения найти вектор-столбец невязки $\bar{r}=\bar{b}-\mathrm{A}\bar{x}^{\times}$ (см. опр. 1.10), где A — основная матрица системы (2.22), а $\bar{b}$ — вектор-столбец свободных членов этой системы.

Решение

Прямой ход. Имеем:

$$i=1,\ \gamma_1=b_1=1,\ \ \alpha_1=-\frac{c_1}{\gamma_1}=1,\ \ \beta_1=\frac{d_1}{\gamma_1}=0.8;$$

$$i=2,\ \gamma_2=b_2+a_2\alpha_1=4+1=5,\ \ \alpha_2=-\frac{c_2}{\gamma_2}=-\frac{1}{5}=-0.2;$$

$$\beta_2=\frac{d_2-a_2\beta_1}{\gamma_2}=\frac{3.6-0.8}{5}=0.56;$$

$$i=3,\ \gamma_3=b_3+a_3\alpha_2=4-0.2=3.8,\ \alpha_3=-\frac{c_3}{\gamma_3}=-\frac{1}{3.8}=-0.263158,$$

$$\beta_3=\frac{d_3-a_3\beta_2}{\gamma_3}=\frac{-1.2-0.56}{3.8}=-0.463158;$$

$$i=4,\ \gamma_4=b_4+a_4\alpha_3=4-0.263158=3.736842,\ \ \alpha_4=-\frac{c_4}{\gamma_4}=-0.267606,$$

$$\beta_4=\frac{d_4-a_4\beta_3}{\gamma_4}=\frac{3.6+0.463158}{3.736842}=1.087324;$$

$$i=5,\gamma_5=b_5+a_5\alpha_4=-1-0.267606=-1.267606,$$

$$\beta_5=\frac{d_5-a_5\beta_4}{\gamma_5}=\frac{-0.8-1.087324}{-1.267606}=1.488889.$$

Обратный ход.

$$x_5^{\times}=\beta_5=1.488889;$$

$$x_4^{\times}=\alpha_4x_5^{\times}+\beta_4=(-0.267606)\cdot(1.488889)+(1.087324)=0.688888;$$

$$x_3^{\times}=\alpha_3x_4^{\times}+\beta_3=(-0.263158)\cdot(0.688888)+(-0.463158)=-0.644444;$$

$$x_2^{\times}=\alpha_2x_3^{\times}+\beta_2=(-0.200000)\cdot(-0.644444)+0.560000=0.688889;$$

$$x_1^{\times}=\alpha_1x_2^{\times}+\beta_1=1.000000\cdot0.688889+0.800000=1.488889,$$

где $x_1^{\times}, x_2^{\times}, x_3^{\times}, x_4^{\times}, x_5^{\times}$ — приближенное решение системы (2.22).

Найдем вектор-столбец невязки (см. опр. 1.10), который для системы (2.22) имеет вид

$$\bar{r} = \begin{pmatrix} 0.8 - x_1^{\times} + x_2^{\times} \\ 3.6 - x_1^{\times} - 4x_2^{\times} - x_3^{\times} \\ -1.2 - x_2^{\times} - 4x_3^{\times} - x_4^{\times} \\ 3.6 - x_3^{\times} - 4x_4^{\times} - x_5^{\times} \\ -0.8 - x_4^{\times} + x_5^{\times} \end{pmatrix}.$$

Подставив вместо $x_1^{\times}, x_2^{\times}, x_3^{\times}, x_4^{\times}, x_5^{\times}$ найденные выше их численные значения, получим

$$\bar{r} = \begin{pmatrix} 0.000000 \\ -0.000001 \\ -0.000001 \\ 0.000003 \\ 0.000001 \end{pmatrix}.$$

Пример 2.2. Методом правой прогонки решить систему линейных алгебраических уравнений (2.7), в которой $n = 50$, а элементы основной матрицы и вектора-столбца приведены в перечне вариантов к заданию 2.1. При этом следует в этом перечне положить $N = 35$ и использовать средства пакета MathCad.

Решение

$n := 20 \qquad N := 35 \qquad i := 1..n$

$ORIGIN := 1$

$B_i := \dfrac{2 + i + N}{1 + i + N}$

$D_i := (-1)^i \dfrac{i}{N}$

$A_i := \dfrac{i}{2 \cdot (i + 1) \cdot N}$

$C_i := \dfrac{i}{3 \cdot (i + 1) \cdot N}$

$$
\begin{aligned}
&\text{progonka}(A, B, C, D) := \\
&\quad \gamma_1 \leftarrow B_1 \\
&\quad \beta_1 \leftarrow \frac{D_1}{\gamma_1} \\
&\quad \alpha_1 \leftarrow \frac{-C_1}{\gamma_1} \\
&\quad \text{for } i \in 2..n-1 \\
&\qquad \gamma_i \leftarrow B_i + A_i \cdot \alpha_{i-1} \\
&\qquad \alpha_i \leftarrow \frac{-C_i}{\gamma_i} \\
&\qquad \beta_i \leftarrow \frac{D_i - A_i \cdot \beta_{i-1}}{\gamma_i} \\
&\quad \gamma_n \leftarrow B_n + A_n \cdot \alpha_{n-1} \\
&\quad \beta_n \leftarrow \frac{D_n - A_n \cdot \beta_{n-1}}{\gamma_n} \\
&\quad x_{n,1} \leftarrow \beta_n \\
&\quad \text{for } i \in n-1, n-2..1 \\
&\qquad x_{i,1} \leftarrow \alpha_i \cdot x_{i+1,1} + \beta_i \\
&\quad \text{for } i \in n-1, n-2..2 \\
&\qquad x_{i,2} \leftarrow D_i - B_i \cdot x_{i,1} - C_i \cdot x_{i+1,1} - A_i \cdot x_{i-1,1} \\
&\quad x_{n,2} \leftarrow D_n - B_n \cdot x_{n,1} - A_n \cdot x_{n-1,1} \\
&\quad x_{1,2} \leftarrow D_1 - B_1 \cdot x_{1,1} - C_1 \cdot x_{2,1} \\
&\quad x
\end{aligned}
$$

$$x_{1,2} \leftarrow D_1 - B_1 \cdot x_{1,1} - C_1 \cdot x_{2,1}$$

$$x_{n,2} \leftarrow D_n - B_n \cdot x_{n,1} - A_n \cdot x_{n-1,1}$$

$$x_{i,2} \leftarrow D_i - B_i \cdot x_{i,1} - A_i \cdot x_{i-1,1} - C_i \cdot x_{i+1,1}$$

$$\text{progonka}(A, B, C, D)\ \text{float}, 10 \rightarrow \begin{pmatrix} -2.808134838 \cdot 10^{-2} & 0 \\ 5.646378554 \cdot 10^{-2} & 0 \\ -8.495173172 \cdot 10^{-2} & 0 \\ .1135016969 & 0 \\ -.1420975691 & 0 \\ .1707312488 & 0 \\ -.1993977635 & 0 \\ .2280936227 & 0 \\ -.2568161482 & 0 \\ .2855631618 & 0 \\ -.3143328224 & 0 \\ .3431235325 & 0 \\ -.3719338819 & 0 \\ .4007626091 & 0 \\ -.4296085762 & 0 \\ .4584707488 & 0 \\ -.4873481784 & 0 \\ .5162395861 & 0 \\ -.5450976725 & 0 \\ .5686896932 & 0 \end{pmatrix}$$

Задание 2.2. Методом простых итераций с точностью $\varepsilon = 10^{-7}$ решить систему линейных алгебраических уравнений, заданную в форме $A \cdot \overline{x} = \overline{b}$. Используя правую часть неравенства (2.21) найти абсолютную погрешность k-го приближения метода Якоби. При этом при вычислении нормы матрицы B применить встроенную подпрограмму *norm*1 и считать, что $k = 10$. (Схема решения этого задания проиллюстрирована в примере 2.2)

Перечень вариантов к заданию 2.2.

Таблица 2.2

Вариант	Матрица A				Вектор правой части $\overline{b}$
1	1.70	0.03	0.04	0.05	0.6810
	0.00	0.80	0.01	0.02	0.4803
	-0.03	-0.02	-0.10	0.00	-0.0802
	-0.05	-0.04	-0.03	-1.00	-1.0007
2	3.00	0.38	0.49	0.59	1.5136
	0.11	2.10	0.32	0.43	1.4782
	-0.05	0.05	1.20	0.26	1.0830
	-0.22	-0.11	-0.11	0.30	0.3280
3	0.77	0.04	-0.21	0.18	1.2400
	-0.45	1.23	-0.06	0.00	-0.8800
	-0.26	-0.34	1.11	0.00	-0.6200
	-0.05	0.26	-0.34	1.12	-1.1700
4	0.79	-0.12	0.34	0.16	-0.6400
	-0.34	1.08	-0.17	0.18	1.4200
	-0.16	-0.34	0.85	0.31	-0.4200
	-0.12	0.26	0.08	0.75	0.8300
5	0.99	-0.02	0.62	-0.08	-1.3000
	-0.03	0.72	-0.33	0.07	1.1000
	-0.09	-0.13	0.58	-0.28	-1.7000
	-0.19	0.23	-0.08	0.63	1.5000
6	3.68	0.16	0.18	0.22	1.1604
	0.12	3.59	0.18	0.21	1.2025
	0.11	0.14	3.50	0.21	1.2409
	0.11	0.14	0.17	3.11	1.2757

7	3.55	0.15	0.18	0.21	1.0834
	0.11	3.46	0.16	0.19	1.1239
	0.12	0.14	3.37	0.20	1.1607
	0.10	0.13	0.17	3.28	1.1938
8	2.38	0.10	0.12	0.14	5.0897
	0.08	2.29	0.11	0.14	5.3487
	0.07	0.09	2.20	0.15	5.5712
	0.06	0.08	0.11	1.10	5.7570
9	1.00	-0.17	0.33	-0.18	-1.2000
	0.00	0.82	-0.43	0.08	0.3300
	-0.22	-0.18	0.79	-0.07	0.4800
	-0.08	-0.07	-0.21	0.96	-1.2000
10	0.68	0.18	-0.02	-0.21	1.8300
	-0.16	0.88	0.14	-0.27	-0.6500
	-0.37	-0.27	1.02	0.24	2.2300
	-0.12	-0.21	0.18	0.75	-1.1300
11	0.58	0.32	-0.03	0.00	0.4400
	-0.11	1.26	0.36	0.00	1.4200
	-0.12	-0.08	1.14	0.24	-0.8300
	-0.15	0.35	0.18	1.00	-1.4200
12	0.82	0.34	0.12	-0.15	-1.3300
	-0.11	0.77	0.15	-0.32	0.8400
	-0.05	0.12	0.86	0.18	-1.1600
	-0.12	-0.08	-0.06	1.00	0.5700
13	0.87	-0.23	0.44	0.05	2.3000
	-0.24	1.00	0.31	-0.15	-0.1800
	-0.06	-0.15	1.00	0.23	1.4400
	-0.72	0.08	0.05	1.00	2.4200

14	0.85	-0.05	0.08	-0.14	-0.4800
	-0.32	1.13	0.12	-0.11	1.2400
	-0.17	-0.06	1.08	-0.12	1.1500
	-0.21	0.16	-0.36	1.00	-0.8800
15	0.97	0.05	-0.22	0.33	0.4300
	-0.22	0.45	0.08	-0.07	-1.8000
	-0.33	-0.13	1.08	0.05	-0.8000
	-0.08	-0.17	-0.29	0.67	1.7000
16	4.30	0.22	0.27	0.32	2.6632
	0.10	3.40	0.21	0.26	2.7779
	0.04	0.09	2.50	0.20	2.5330
	-0.03	0.03	0.08	1.60	1.9285
17	5.60	0.27	0.33	0.39	4.0316
	0.15	4.70	0.27	0.33	4.3135
	0.09	0.15	3.80	0.27	4.2353
	0.03	0.09	0.15	2.90	3.7969
18	6.90	0.32	0.39	0.46	5.6632
	0.19	6.00	0.33	0.41	6.1119
	0.13	0.21	5.10	0.35	6.2000
	0.08	0.15	0.22	4.20	5.9275
19	8.20	0.37	0.45	0.53	7.5591
	0.23	7.30	0.39	0.48	8.1741
	0.18	0.26	6.40	0.42	8.4281
	0.12	0.21	0.29	5.50	8.3210
20	9.50	0.42	0.51	0.60	9.7191
	0.28	8.60	0.46	0.55	10.5000
	0.22	0.32	7.70	0.50	10.9195
	0.17	0.26	0.35	6.80	10.9775

21	0.87	-0.22	0.33	-0.07	0.1100
	0.00	0.55	0.23	-0.07	-0.3300
	-0.11	0.00	1.08	-0.18	0.8500
	-0.08	-0.09	-0.33	0.79	-1.7000
22	0.68	0.16	0.08	-0.15	2.4200
	-0.16	1.23	-0.11	0.21	1.4300
	-0.05	0.08	1.00	-0.34	-0.1600
	-0.12	-0.14	0.18	0.94	1.6200
23	1.00	-0.08	0.23	-0.23	1.3400
	-0.16	1.23	-0.18	-0.16	-2.3300
	-0.15	-0.12	0.68	0.18	0.3400
	-0.25	-0.21	0.16	0.97	0.6300
24	10.80	0.05	0.06	0.07	12.1430
	0.03	9.90	0.05	0.06	13.0897
	0.04	0.04	9.00	0.08	13.6744
	0.02	0.03	0.04	8.10	13.8972
25	12.10	5.28	0.64	0.75	14.8310
	0.37	11.20	5.86	0.69	15.9430
	0.31	0.42	10.30	6.44	16.6926
	2.60	0.37	4.81	19.40	17.0800
26	13.40	5.81	0.70	0.82	17.7828
	0.41	12.50	6.50	0.77	19.0599
	0.36	0.48	11.60	7.18	19.9744
	0.31	0.43	0.54	10.70	20.5261
27	0.94	-0.18	-0.33	-0.16	2.4300
	-0.32	1.00	-0.23	0.05	-1.1200
	-0.16	0.08	1.00	0.12	0.4300
	-0.09	-0.22	0.13	1.00	0.8300

28	1.00	-0.34	-0.23	0.06	1.4200
	-0.11	1.23	0.18	-0.36	-0.6600
	-0.23	0.12	0.84	0.35	1.0800
	-0.12	-0.12	0.47	0.82	1.7200
29	0.68	0.23	-0.11	0.06	0.6700
	-0.18	0.88	0.33	0.00	-0.8800
	-0.12	-0.32	1.05	-0.07	0.1800
	-0.05	0.11	-0.09	1.12	1.4400
30	0.77	0.14	-0.06	0.12	1.2100
	-0.12	1.00	-0.32	0.18	-0.7200
	-0.08	0.12	0.77	-0.32	-0.5800
	-0.25	-0.22	-0.14	1.00	1.5600

Пример 2.2. Методом Якоби требуется с точностью $\varepsilon = 10^{-7}$ решить в пакете Mathcad следующую систему:

$$\begin{cases} 0.4000x_1 + 0.0003x_2 + 0.0008x_3 + 0.0014x_4 = 0.1220, \\ -0.0029x_1 - 0.5000x_2 - 0.0018x_3 - 0.0012x_4 = -0.2532, \\ -0.0055x_1 - 0.0050x_2 - 1.4000x_3 - 0.0039x_4 = -0.9876, \\ -0.0082x_1 - 0.0076x_2 - 0.0070x_3 - 2.3000x_4 = -2.0812. \end{cases} \quad (2.23)$$

Решение

$\text{ORIGIN} := 1 \quad \text{TOL} := 10^{-7}$

$$A := \begin{pmatrix} 0.4000 & 0.0003 & 0.0008 & 0.0014 \\ -0.0029 & -0.5000 & -0.0018 & -0.0012 \\ -0.0055 & -0.0050 & -1.4000 & -0.0039 \\ -0.0082 & -0.0076 & -0.0070 & -2.3000 \end{pmatrix} \quad b := \begin{pmatrix} 0.1220 \\ -0.2532 \\ -0.9876 \\ -2.0812 \end{pmatrix}$$

$$n := 4 \quad i := 1...n \quad c_i := \frac{b_i}{A_{i,i}} \quad c = \begin{pmatrix} 0.305 \\ 0.5064 \\ 0.7054286 \\ 0.9048696 \end{pmatrix}$$

$$i := 1...n \quad j := 1...n \quad B_{i,j} := \frac{-A_{i,j}}{A_{i,i}} \quad i := 1...n \quad B_{i,j} := 0$$

$$B = \begin{pmatrix} 0 & -7.5 \cdot 10^{-4} & -2 \cdot 10^{-3} & -3.5 \cdot 10^{-3} \\ -5.8 \cdot 10^{-3} & 0 & -3.6 \cdot 10^{-3} & -2.4 \cdot 10^{-3} \\ -3.928514 \cdot 10^{-3} & -3.5714286 \cdot 10^{-3} & 0 & -2.78571443 \cdot 10^{-3} \\ -3.5652174 \cdot 10^{-3} & -3.3043478 \cdot 10^{-3} & -3.0434783 \cdot 10^{-3} & 0 \end{pmatrix}$$

$$\text{norm}\,1(B) = 0.013$$

Встроенная подпрограмма $\text{norm}\,1$ позволяет вычислить нормы матриц A и B. Поскольку $\|B\|_1 = 0.013$, то согласно ***теореме 2.3*** метод простых итераций (метод Якоби) сходится при любом начальном приближении.

Предыдущие операторы программы приводят систему уравнений, которая задана в виде $A \cdot \overline{x} = \overline{b}$, к виду $\overline{x} = B \cdot \overline{x} + \overline{c}$. Процесс последовательных приближений метода Якоби записывается в векторно-матричной форме посредством одной строки программы:

$$xx^{\langle 1 \rangle} := c \quad k := 2...11 \quad xx^{\langle k \rangle} := c + B \cdot xx^{\langle k-1 \rangle}.$$

В качестве начального приближения взят вектор $\overline{c}$. Произведя расчет, получаем, что для достижения заданной точности 10^{-7}нужно произвести лишь три итерации. Отметим, что под $xx^{\langle k \rangle}$ понимается k-й столбец матрицы xx размерности (4×11), которая фиксирует (и сохраняет) все приближения к точному решению решаемой системы. Запишем часть элементов этой матрицы в виде

xx=

	1	2	3	4	5
1	0.305	0.3000423	0.3000754	0.300075	0.300075
2	0.5064	0.4999198	0.4999802	0.4999797	0.4999797
3	0.7054286	0.6999011	0.6999574	0.6999569	0.6999569
4	0.9048696	0.8999619	0.9000178	0.9000173	0.9000173

Ниже записаны две подпрограммы, реализующие приведение исходной системы $A \cdot \overline{x} = \overline{b}$ к виду $\overline{x} = B \cdot \overline{x} + \overline{c}$ и итерационные вычисления по методу Якоби.

$$\text{itera}(A,n) := \left| \begin{array}{l} \text{for } i \in 1...n \\ \quad \left| \begin{array}{l} \alpha \leftarrow \dfrac{1}{A_{i,i}} \\ \text{for } j \in 1...n \\ \quad \left| \begin{array}{l} B_{i,j} \leftarrow -A_{i,j} \cdot \alpha \\ B_{i,j} \leftarrow 0 \text{ if } i = j \end{array} \right. \end{array} \right. \\ B \end{array} \right.$$

$$\text{Jakobi}(A,b,\varepsilon) := \left|\begin{array}{l} n \leftarrow \text{rows}(A) \\ B \leftarrow \text{itera}(A,n) \\ \text{for } i \in 1..n \\ \quad \left|\begin{array}{l} c_i \leftarrow \dfrac{b_i}{A_{i,i}} \\ x_i \leftarrow 0 \\ x1_i \leftarrow c_i \end{array}\right. \\ \text{norm} \leftarrow \text{norm1}(B) \\ \text{return norm if norm} \geq 1 \\ \text{while norm} > \varepsilon \\ \quad \left|\begin{array}{l} \text{for } i \in 1..n \\ x_i \leftarrow c_i + \left[\displaystyle\sum_{j=1}^{n}\left(B_{i,j} \cdot x1_i\right)\right] \\ \text{norm} \leftarrow \sqrt{x1 \cdot x1} \\ x1 \leftarrow x - x1 \\ \text{norm} \leftarrow \dfrac{\sqrt{x1 \cdot x1}}{\text{norm}} \\ x1 \leftarrow x \end{array}\right. \\ x \end{array}\right.$$

$\varepsilon := 10^{-7}$ $z := Jakobi(A,b,\varepsilon)$

$$z = \begin{pmatrix} 0.3000750 \\ 0.4999797 \\ 0.6999569 \\ 0.9000173 \end{pmatrix}$$

III. АППРОКСИМАЦИЯ ФУНКЦИЙ

Изложим ряд общих понятий теории аппроксимации функций.

Определение 3.1. Аппроксимацией (приближением) функции называется замена заданной функции $f(x)$ некоторой функцией $\Phi(x)$ так, чтобы отклонение функции $\Phi(x)$ от $f(x)$ в заданной области (она может совпадать с областью определения функции $f(x)$ было в том или ином смысле наименьшим. Функция $\Phi(x)$ при этом называется а**ппроксимирующей**.

Типичными задачами аппроксимации функции являются задачи **интерполяции**, а так же **приближение функции по методу наименьших квадратов**.

3.1. Интерполяция

Постановка задачи интерполяции

Простейшая задача **интерполяции** заключается в следующем. На отрезке $[a, b]$ заданы $n + 1$ точек $x_i = x_0, x_1, \ldots, x_n$, которые называются **узлами интерполяции**, и значения некоторой функции $f(x)$ в этих точках

$$f(x_0) = y_0,\ f(x_1) = y_1, \ldots, f(x_n) = y_n. \qquad (3.1)$$

Определение 3.2. Функцию $\Phi(x)$ называется **интерполирующей,** если она принадлежит известному классу и принимает в узлах интерполяции те же значения, что и $f(x)$, т. е.

$$\Phi(x_0) = y_0, \Phi(x_1) = y_1, \ldots, \Phi(x_n) = y_n. \qquad (3.2)$$

Геометрически это означает, что нужно найти кривую $y = \Phi(x)$ некоторого определенного типа, проходящую через заданную систему точек

$M(x_i, y_i)$ ($i = 0, 1, \ldots, n$) (рис. 3.1.).

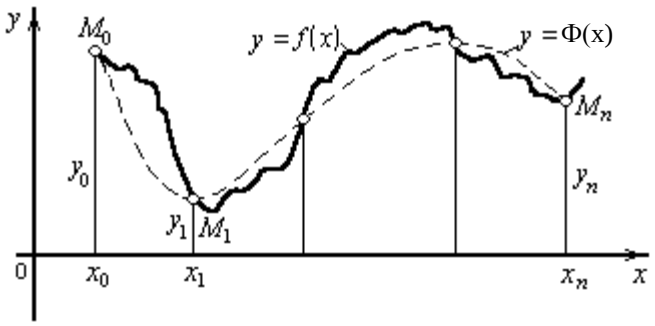


Рис. 3.1. Графическая иллюстрация к задаче интерполяции функции

Задача становится однозначной, если вместо произвольной функции Ф(x) искать полином $P_n(x)$ (**интерполяционный полином**) степени не выше n, удовлетворяющий условиям (3.2). Полученный полином $P_n(x)$ обычно используют для приближенного вычисления значений данной функции на замкнутом промежутке $x_0 \le x \le x_n$. Такая операция называется **интерполяцией** *функции*. Если полином используется для вычисления значений функции вне промежутка $x_0 \le x \le x_n$, то такая операция называется **экстраполяцией.**

Замечание 3.1. *Выбор того или иного вида интерполяционной схемы зависит от конкретной ситуации.*

1. Мы можем потребовать, чтобы интерполяционная кривая была гладкая на всем промежутке интерполяции. Тогда можно использовать полином Лагранжа [3,5] *соответствующей степени. Если число точек более четырех, то разумно использовать кубические сплайны* [5].

2. Если требование гладкости не является существенным, то вполне можно ограничиться классическими интерполяционными полиномами Лагранжа по двум (линейным), трем (квадратичным) или четырем (кубическим) ближайшим точкам.

3.1.1. Глобальная интерполяция полиномами Лагранжа

Интерполяционная формула Лагранжа

Пусть на отрезке [a, b] даны n + 1 различных значений аргумента: x_0, x_1, ..., x_n и известны для функции $y = f(x)$ соответствующие значения выражений

$$f(x_0) = y_0,\ f(x_1) = y_1,\ \ldots,\ f(x_n) = y_n\ . \qquad (3.3)$$

Требуется построить полином $L_n(x)$ степени не выше n, имеющий в заданных узлах $x_0, x_1, \ldots, x_n$ те же значения, что и функция $f(x)$, т. е. такой, что $L_n(x_i) = y_i\ (i = 0, 1, \ldots, n)$.

Интерполяционный полином в форме Лагранжа имеет вид [5]

$$L_n(x) = \sum_{i=0}^{n} y_i \frac{(x - x_0)(x - x_1)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)}. \qquad (3.4)$$

Пример 3.1. Положим $n = 1$. Ясно, что мы имеем в этом случае две точки и интерполяционная формула Лагранжа дает уравнение прямой, проходящей через две заданные точки.

$$L_1(x) = \frac{x - x_1}{x_0 - x_1} y_0 + \frac{x - x_0}{x_1 - x_0} y_1\ .$$

Пример 3.2. Пусть заданы значения функции

Таблица 3.1

X_i	0	0,5	1	3	7	12	15
Y_i	2,5	4,3	5,6	6,7	8,1	10,3.	11

Определить значение неизвестной функции $Y(x)$ в точке $x = 6{,}5$, используя кубический полином Лагранжа.

Для повышения точности интерполяции, выберем четыре точки расположенные рядом с точкой x=6,5. То

$$x_0 = 1,\quad x_1 = 3,\quad x_2 = 7,\quad x_3 = 12,$$
$$y_0 = 5,6,\quad y_1 = 6,7,\quad y_2 = 8,3,\quad y_3 = 10,3.$$

Интерполяционная формула Лагранжа имеет вид

$$L_3(x)=\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}\cdot y_0+\frac{(x-x_0)(x-x_2)(x-x_3)}{(x-x_0)(x_1-x_2)(x_1-x_3)}\cdot y_1+$$
$$+\frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}\cdot y_2+\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\cdot y_3.$$

После подстановки заданных значений в формулу Лагранжа получаем:

$$L_3(x)=\frac{(x-3)(x-7)(x-12)}{-132}\cdot 5,6+\frac{(x-1)(x-7)(x-12)}{72}\cdot 6,7+$$
$$+\frac{(x-1)(x-3)(x-12)}{-120}\cdot 8,1+\frac{(x-1)(x-3)(x-7)}{495}\cdot 10,3.$$

Определим значение $L_3(x)$ при $x = 6,5$:

$$L_3(6,5)=-\frac{3,5\cdot 0,5\cdot 5,5}{-132}\cdot 5,6+\frac{5,5\cdot 0,5\cdot 5,5}{72}\cdot 6,7+$$
$$+\frac{5,5\cdot 3,5\cdot 5,5}{120}\cdot 8,1-\frac{5,5\cdot 3,5\cdot 0,5}{495}\cdot 10,3=7,9.$$

3.1.2. Локальная интерполяция

Кусочно-полиномиальная интерполяция

В общем случае, когда число точек больше n+1, для вычисления значения функции можно использовать полином Лагранжа $P_n(x)$ степени n, выбирая для построения полинома $P_n(x)$ n+1 ближайших узлов (“бегущий полином”).

На рис. 3.4 приводятся графики кусочно-постоянной, кусочно-кубической интерполяции полиномами Лагранжа.

Кубическая сплайн-интерполяция

Кубическая сплайн-интерполяция позволяет провести кривую через набор точек таким образом, что первые и вторые производные кривой были непрерывны в каждой точке. Эта кривая образуется путем создания ряда кубических

полиномов, проходящих через две смежные точки. Кубические полиномы затем состыковываются друг с другом так, чтобы образовать одну гладкую кривую.

Определение 3.3. Для каждого отрезка $x_i \le x \le x_{i+1}$, функция сплайн-интерполяции S(x) имеет вид [9]

$$S(x) = \frac{1}{6h_i}\left[m_i \left(x_{i+1} - x \right)^3 + m_{i+1} \left(x - x_i \right)^3 \right] + \\ + \frac{1}{h_i}\left[\left(y_i - \frac{m_i h_i^2}{6} \right)\left(x_{i+1} - x \right) + \left(y_{i+1} - \frac{m_{i+1} h_i^2}{6} \right)\left(x - x_i \right) \right], \quad (3.5)$$

где $x_i \le x \le x_{i+1}$,

$h_i = x_{i+1} - x_i$,

$f_i(x) = y(x)$,

$m_i = f''''(x_i)$,

$y_i = F_i(t)$,

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

n - число узлов.

При известных x_i, y_i, m_i эта формула задает сплайн-аппроксимацию: кусочно-кубические полиномы, непрерывные вместе со своими производными первого и второго порядка. Если потребовать выполнение условия непрерывности вторых производных, то выражение (3.5) для кубических полиномов-сплайнов приведет к системе линейных уравнений, из которых находятся m_i - приближенные значения вторых производных в точке i:

$$h_i m_i + 2\left(h_i + h_{i+1}\right) m_{i+1} + h_i + m_{i+2} = 6\left(\frac{y_{i+2} - y_{i+1}}{h_{i+1}} - \frac{y_{i+1} - y_i}{h_i} \right). \quad (3.6)$$

MathCAD поставляется с тремя сплайн-функциями:

cspline(*X*,*Y*)

pspline(*X*,*Y*)

lspline(*X*,*Y*)

Они возвращают вектор коэффициентов вторых производных m_i**,** который мы будем называть *M*. Этот вектор *M* обычно используется в функции interp, описанной ниже. Значения вектора *X* должны быть расположены в порядке возрастания.

Эти три функции отличаются только граничными условиями:

- *функция* lspline *генерирует кривую сплайна, которая приближается к прямой линии в граничных точках;*
- *функция* pspline *генерирует кривую сплайна, которая приближается к параболе в граничных точках.*
- *функция* cspline *генерирует кривую сплайна, которая может быть кубическим полиномом в граничных точках.*

Функция interp (*X*,*Y*,*M*,*t*) возвращает интерполируемое значение *S*(*t*) соответствующее аргументу *t* по формуле **(**3.6**)**. Вектор *M* вычисляется на основе векторов данных *X* и *Y* одной из функций pspline, lspline или cspline.

Чтобы провести кубический сплайн через набор точек:

1. Создайте векторы X и Y, содержащие координаты x_i и y_i, $i = 1, 2,\ldots, n$, через которые нужно провести кубический сплайн. Элементы X должны быть расположены в порядке возрастания.

2. Вычислите вектор

M:=cspline(X,Y).

Вектор M содержит вторые производные интерполяционной кривой в рассматриваемых точках;

3.Чтобы найти интерполируемое значение в произвольной точке t, вычислите interp(X,Y,M,t), где X,Y,M- векторы, описанные ранее.

3.2. Метод наименьших квадратов

Аппроксимация данных с учетом их статистических параметров относится к задачам **регрессии**. Задачей регрессионного анализа является подбор математических формул, наилучшим образом описывающих экспериментальные данные. Нахождение неизвестных коэффициентов можно производить минимизируя сумму квадратов отклонений кривой регрессии от экспериментальных данных - метод наименьших квадратов.

3.2.1. Линейная регрессия в системе Mathcad

Линейная регрессия в системе Mathcad выполняется по векторам аргумента X и отсчетов Y функциями:

intercept(X,Y) – вычисляет параметр a, смещение линии регрессии по вертикали;

slope(X,Y) – вычисляет параметр b, угловой коэффициент линии регрессии.

Расположение отсчетов по аргументу X произвольное.

Функцией corr(X,Y) дополнительно можно вычислить коэффициент корреляции Пирсона. Чем он ближе к 1, тем точнее обрабатываемые данные соответствуют линейной зависимости.

Пример выполнения линейной регрессии приведен на рис. 3.2.

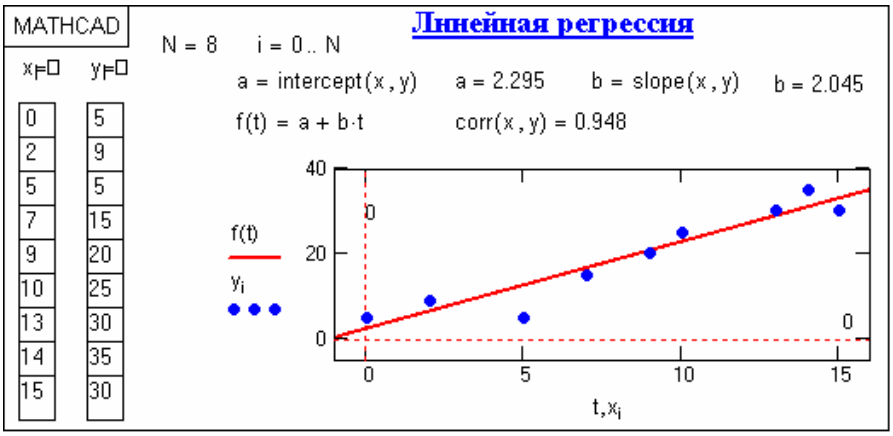


Рис. 3.2. Построение линейной регрессии с использованием пакета MathCad

3.2.2. Полиномиальная регрессия

Одномерная полиномиальная регрессия с произвольной степенью n полинома и с произвольными координатами отсчетов в Mathcad выполняется функциями:

regress(*X*,*Y*,*n*) – вычисляет вектор *S* для функции interp(…), в составе которого находятся коэффициенты k_i полинома *n*-й степени;

cinterp(*S*, *X*, *Y*, *x*) – возвращает значения функции аппроксимации по координатам x.

Функция interp(…) реализует вычисления по формуле:

$f(x) = k_0 + k_1 \cdot x^1 + k_2 \cdot x^2 + \ldots + k_n \cdot x^n \equiv \Sigma_i k_i \cdot x^i$.

Значения коэффициентов k_i могут быть извлечены из вектора *S* функцией submatrix (*S*, 3, length(*S*), 0, 0).

На рис. 3.3. приведен пример полиномиальной регрессии с использованием полиномов 2, 3 и 8-й степени. Степень полинома обычно устанавливают не более 4-6 порядка с последовательным повышением степени, контролируя среднеквадратическое отклонение функции аппроксимации от фактических данных. Нетрудно заметить, что по мере повышения степени полинома функция аппроксимации приближается к фактическим данным, а при степени полинома, равной количеству отсчетов данных минус 1, вообще превращается в функцию

интерполяции данных, что не соответствует задачам регрессии.

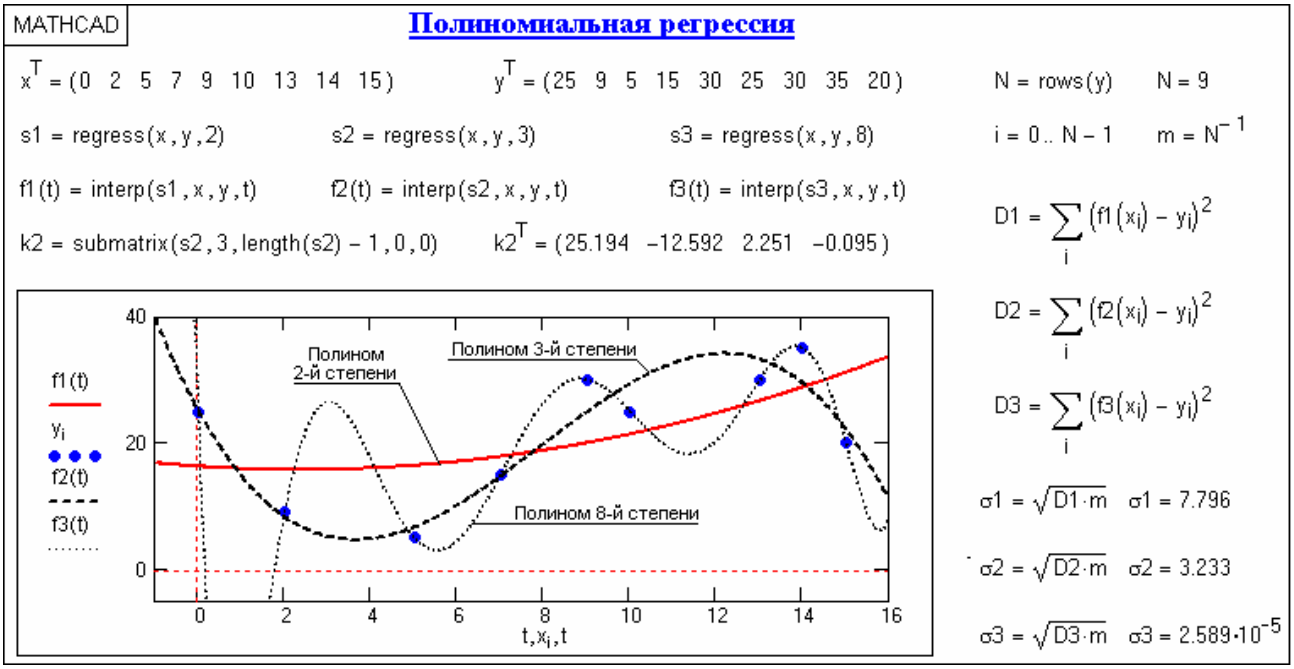
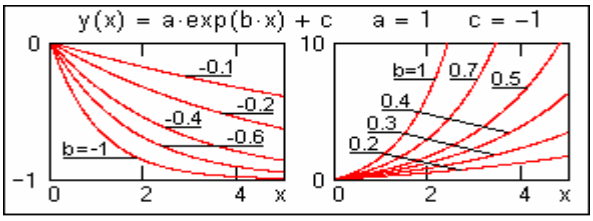


Рис. 3.3. Одномерная полиномиальная регрессия

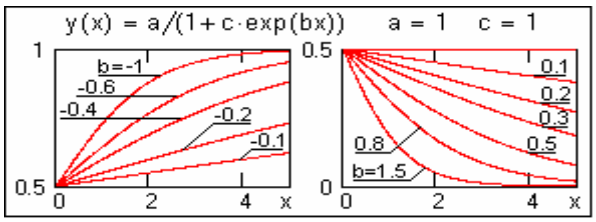
3.2.3. Типовые функции регрессии Mathcad

Для анализа эмпирических данных можно использовать некоторые простые типовые формулы.

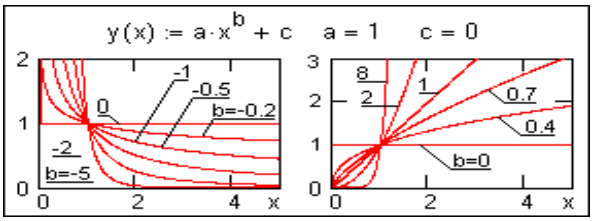
Для простых типовых формул аппроксимации предусмотрен ряд функций регрессии, в которых параметры функций подбираются программой Mathcad самостоятельно. К ним относятся следующие функции:



expfit(*X*,*Y*,*S*) – возвращает вектор, содержащий коэффициенты *a*, *b* и *c* экспоненциальной функции $y(x) = ae^{bx} + c$. В вектор *S* вводятся начальные значения коэффициентов *a*, *b* и *c* первого приближения. Для ориентировки по форме аппроксимирующих функций и задания соответствующих начальных значений коэффициентов на рисунках слева приводится вид функций при постоянных значениях коэффициентов *a* и *c*.

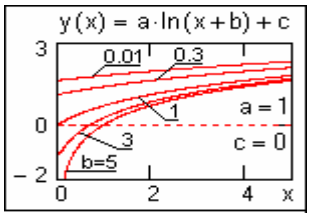


lgsfit(*X*,*Y*,*S*) – то же, для выражения $y(x) = a/(1 + ce^{bx})$.



pwrfit(*X*,*Y*,*S*) – то же, для выражения $y(x) = a \cdot x^b + c$.

sinfit(*X*,*Y*,*S*) – то же, для выражения $y(x) = a \cdot \sin(x+b) + c$. Подбирает коэффициенты для синусоидальной функции регрессии. Рисунок синусоиды общеизвестен.



logfit(*X*,*Y*) – то же, для выражения $y(x) = a \cdot ln(x+b) + c$. Задания начального приближения не требуется.

medfit(*X*,*Y*) – то же, для выражения $y(x) = a + b \cdot x$, т.е. для функции линейной регрессии. Задания начального приближения также не требуется. График – прямая линия.

На рис. 3.6 приведен пример реализации регрессии экспоненциального вида: $Y(t) = ae^{bt} + c$, а также вычисляются отклонения исходных значений функции от кривой регрессии в каждой точке и величина среднеквадратичного отклонения.

3.3. Контрольные вопросы

1) Что такое аппроксимация функций?
2) Какие виды аппроксимации вам известны
3) Для чего нужна интерполяция функций?
4) Что такое экстраполяция?
5) Охарактеризуйте виды интерполяции.
6) Сколько интерполяционных полиномов можно построить при заданном наборе узлов интерполяции?
7) Чем обуславливается выбор способов интерполяции?
8) Какие методы локальной интерполяции вам известны?
9) Какой из них наименее точный?
10) Какой метод локальной интерполяции проводится по трем точкам?
11) В чем преимущества сплайн-интерполяции по сравнению с интерполяционными полиномами?

12) Какая функция MathCAD реализует линейную интерполяцию?

13) Какие функции кубической сплайн-интерполяции вам известны, охарактеризуйте последовательность их использования?

14) Объясните разницу между глобальной и кусочно-полиномиальной интерполяцией на примере сплайн-функции.

15) Исправьте ошибки , допущенные программой Mathcad, на графиках функций на рисунках 3.5 (слева и справа).

16) В чем сущность метода наименьших квадратов?

17) Какие функции MathCAD реализует линейную аппроксимацию методом наименьших квадратов?

18) Какой диапазон изменения значений коэффициента корреляции?

19) Что такое эмпирическая формула и как ее подобрать?

20) Перечислите типовые функции регрессии.

3.4. Компьютерный практикум

$$x := \begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \\ 0.45 \\ 0.5 \end{pmatrix} \quad y := \begin{pmatrix} 0.7 \\ 0.0 \\ 0.7 \\ 0.0 \\ 0.7 \\ 0 \end{pmatrix}$$

Линейная интерполяция : $L(t) := linterp(x, y, t)$

Сплайн интерполяция: $M := cspline(x, y)$ $SP(t) := interp(M, x, y, t)$

Кубическая интерполяция : $N := length(x)$

$$s(j) := if(j > 3, j - 4, j) \qquad ni(t) := \sum_{i=2}^{N-3} if(t - x_i > 0, 1, 0)$$

$$I(t, i, j) := \frac{(t - x_{i+s(j+1)})(t - x_{i+s(j+2)}) \cdot (t - x_{i+s(j+3)})}{(x_{i+s(j)} - x_{i+s(j+1)})(x_{i+s(j)} - x_{i+s(j+2)}) \cdot (x_{i+s(j)} - x_{i+s(j+3)})}$$

$$P(t, i) := y_i \cdot I(t, i, 0) + y_{i+1} \cdot I(t, i, 1) + y_{i+2} \cdot I(t, i, 2) + y_{i+3} \cdot I(t, i, 3)$$

$$PL(t) := P(t, ni(t))$$

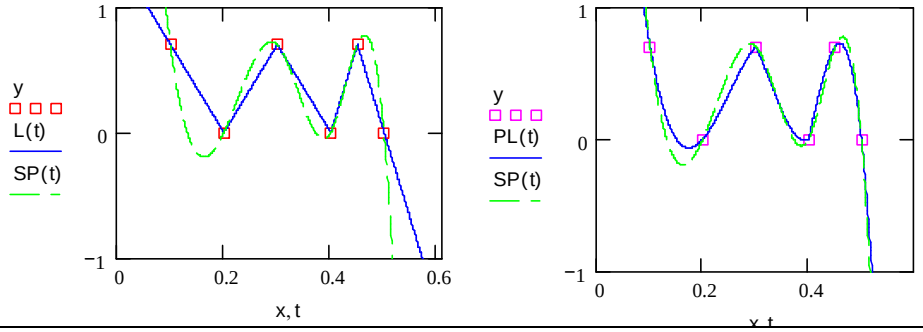


Рис. 3.4. Пример кусочно-линейной, кубической и сплайн-интерполяции

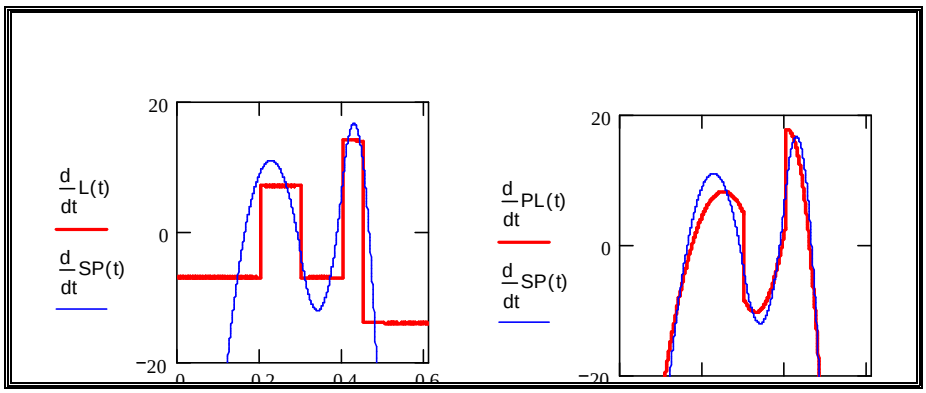


Рис. 3.5. Графики производных кусочно-линейной, кубической и сплайн-функции. Исходные функции взяты из примера на рис. 3.2.

Из рис. 3.3 видно, что только график производной сплайн-функции является гладким. На рис 3.3 слева неточность: график производной кусочно-постоянной функции D(L(t)), в точках, где производная не существует, программой Mathcad нарисован в виде вертикальных отрезков прямых.

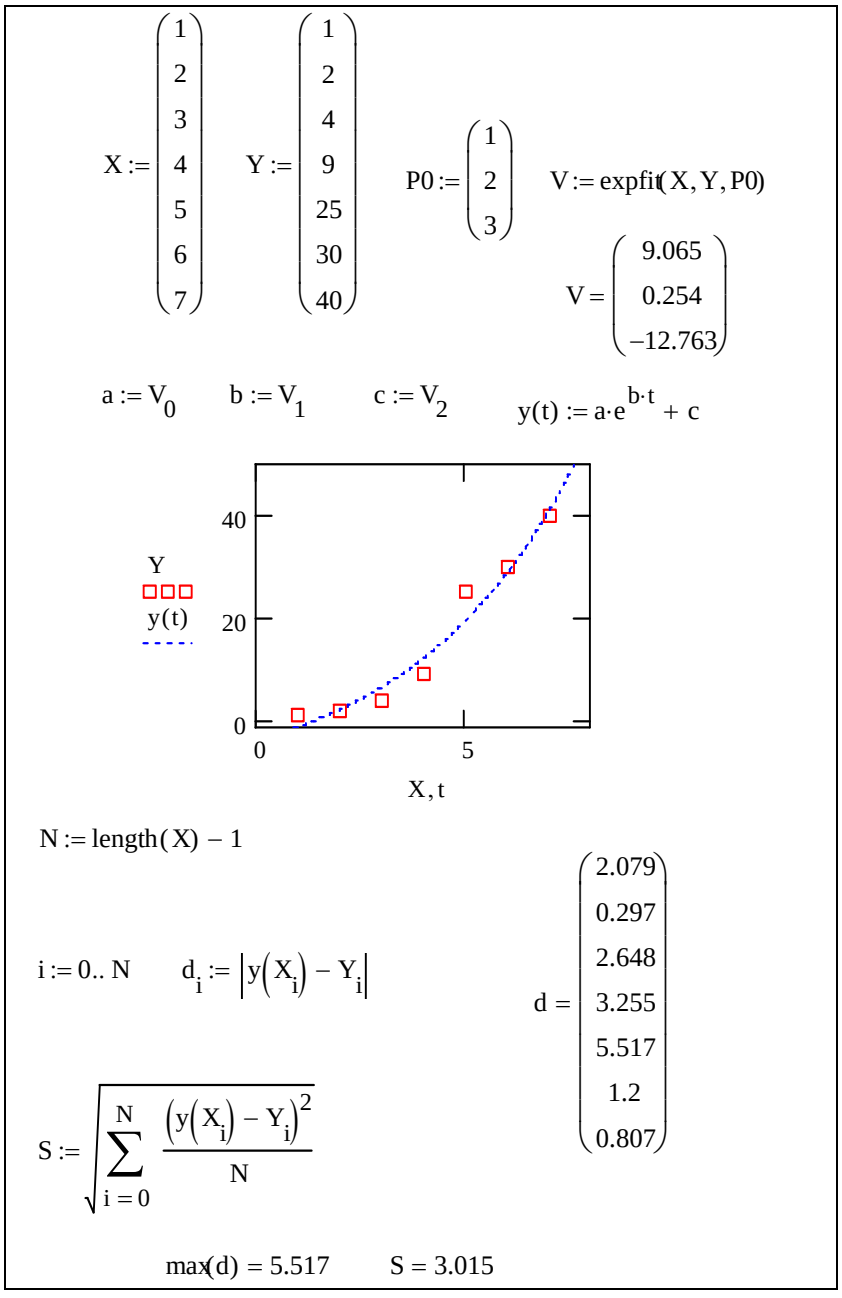


Рис. 3.6. Регрессия экспоненциального вида: $Y(t)=ae^{bt}+c$

3.5. Варианты заданий для самостоятельной работы

3.5.1. Задание по разделу интерполяция функции

1. Построить графики кусочно-постоянной, кубической и сплайн — интерполяции для таблично заданной функции. Данные взять из таблицы 3.1.

2. Построить графики производных кусочно-постоянной, кубической и сплайн-интерполяции для таблично заданной функции. Данные взять из таблицы 1.

3. Записать в общем виде формулу для вычисление значения X для $X_3<X<X_4$, в случае кусочно-постоянной, кусочно-кубической и сплайн – интерполяции.

4. Укажите на графике точки, в которых нарушается дифференцируемость функций.

Пример выполнения задания представлен на рисунках 3.2, 3.3.

3.5.2. Задание по разделу метод наименьших квадратов

1. Найти точечные оценки для параметров моделей

$$y = \beta_0 + \beta_1 x, \tag{3.7}$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{3.8}$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3, \tag{3.9}$$

$$y = \beta_0 + \beta_1 \exp(\text{alfa} \cdot x), \tag{3.10}$$

$$y = b \cdot \ln(x + a), \tag{3.11}$$

$$y = a x^{b \cdot x} + c, \tag{3.12}$$

$$y = \beta_0 + \beta_1 \sin(x). \tag{3.13}$$

2. Найти среднеквадратичное $\sqrt{\sum_i \frac{|Y(x_i) - Y_i|^2}{n}}$ отклонение и максимум

модуля разности $\max_i \left|Y(x_i) - Y_i\right|$ для каждой модели, построить графики. Выяснить — какая модель является наилучшей в смысле минимального среднеквадратичного отклонении, а какая модель является наилучшей в смысле минимального по модулю отклонения.

Данные взять из таблицы 3.1.

3. Записать в общем виде формулу для вычисление значения *y(x)* — для модели заданной выражениями 3.7—3.13.

4. Подготовьте ответы на вопросы:

1) В чем сущность метода наименьших квадратов?

2) Какие функции MathCAD реализует линейную аппроксимацию методом наименьших квадратов?

3) Какой диапазон изменения значений коэффициента корреляции?

4) Что такое эмпирическая формула и как ее подобрать?

5) Перечислите типовые функции регрессии.

5. Пример выполнения задания представлен на рисунках 3.4, 3.5, 3.6.

Перечень вариантов к заданиям 3.5.1. и 3.5.2.

Табличные значения функции (X_i,Y_i), i=0...N

Таблица 3.2

X	0.53	1.1	1.4	2.57	*CC*	2.8	3.7	4.5
Y	*BB*	0.33	1.0	1.7	0.0	3.4	4.1	*NN*

где *NN*- номер варианта;

CC – число, которое выбирается студентом самостоятельно;

BB – число, которое задается преподавателем.

Отчет должен содержать протокол выполнения задания и ответы на вопросы.

IV. МЕТОДЫ ЧИСЛЕННОГО РЕШЕНИЯ ЗАДАЧИ КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ (ОДУ) И СИСТЕМ ОДУ

4.1. Вводные замечания

В разделе «Дифференциальные уравнения» курса высшей математики были перечислены основные типы ОДУ и систем ОДУ, решаемых аналитически. В практических задачах такие уравнения – чаще исключение. В связи с этим были разработаны приближенные аналитические и полуаналитические методы, а также численные методы решения ОДУ [5]. Данный раздел можно рассматривать как краткое введение в численные методы решения задачи Коши для ОДУ и систем ОДУ.

Рассмотрим систему n дифференциальных уравнений 1-го порядка

$$\vec{Y}'(t) = \vec{F}(t;\vec{Y}(t)), \qquad (4.1)$$

заданную на отрезке $[t_0, T]$. В (4.1) $\vec{Y}(t) \equiv \left(y_1(t)\ \ y_2(t) \ldots \quad y_n(t)\right)^T$ — вектор-столбец неизвестных функций (напомним, что надстрочный символ T означает операцию транспонирования); правая часть в (4.1) $\vec{F}\left(t;\vec{Y}(t)\right) \equiv$

$\equiv \left(f_1\left(\vec{Y}(t)\right) \ldots f_n\left(\vec{Y}(t)\right)\right)^T = \left(f_1(t; y_1(t),\ldots, y_n(t)) \quad \ldots \quad f_n(t; y_1(t),\ldots, y_n(t))\right)^T$, где функции $f_1(\ldots),\ldots, f_n(\ldots)$ считаются заданными.

К частному случаю системы (4.1) легко сводится ОДУ n-го порядка относительно неизвестной функции $y(t)$:

$$y^{(n)}(t) = f(t;\ y'(t), y''(t),\ldots, y^{(n-2)}(t), y^{(n-1)}(t)). \qquad (4.2)$$

Это достигается введением неизвестных функций $y_1(t),\ ...,y_n(t)$ по правилу $y_1(t)\equiv y(t),\ y_2(t)\equiv y'(t),\ ...,\ y_n(t)\equiv y^{(n-1)}(t)$. Следовательно, можно, вообще говоря, не рассматривать отдельно уравнение (4.2).

Задав начальные условия (НУ)

$$\vec{Y}(t_0)=\vec{Y}_0, \tag{4.3}$$

где $\vec{Y}_0=(y_{10},...y_{n0})^T$ — известный вектор-столбец, получим для системы (4.1) задачу Коши (ЗК).

В дальнейшем будем полагать, что вектор-функция $\vec{F}\left(t,\vec{Y}(t)\right)$ удовлетворяет условиям классических *теорем существования и единственности* решения ЗК [10,11]. При этом будем принимать условия теоремы настолько жесткими, насколько это необходимо для существования и единственности решения, дифференцируемого желаемое количество раз на отрезке $[t_0,T]$.

4.2. Метод Эйлера, его сходимость и абсолютная погрешность

Суть численных методов решения задачи (4.1), (4.3) можно понять, рассмотрев самый простой из них – метод Эйлера. Он состоит в следующем:

1) Производная в (4.1) заменяется *разностным отношением*

$$\vec{Y}'(t)=\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left(\vec{Y}(t+\Delta t)-\vec{Y}(t)\right)\approx\frac{1}{\Delta t}\left(\vec{Y}(t+\Delta t)-\vec{Y}(t)\right). \tag{4.4}$$

Понятно, что приближение (4.4) оправдано для достаточно малых значений Δt.

2) Подставив (4.4) в (4.1), получим вместо системы дифференциальных уравнений систему уравнений в *конечных разностях*

$$\frac{1}{\Delta t}\left(\vec{Y}(t+\Delta t)-\vec{Y}(t)\right)\approx\vec{F}\left(t;\vec{Y}(t)\right). \tag{4.5}$$

Из (4.5) находим

$$\vec{Y}(t+\Delta t) \approx \vec{Y}(t) + \Delta t \cdot \vec{F}(t;\vec{Y}(t)). \qquad (4.6)$$

3) Из (4.6) следует каким образом строится простейший способ приближенного вычисления функции $\vec{Y}(t)$, а именно: задаем на отрезке $[t_0,T]$ последовательность точек $t_0,t_1,...,t_N$ $(t_N = T)$, и вычисляем в соответствии с (4.6) приближенные значения $\vec{Y}^{\times}(t_i)$ вектор-функции $\vec{Y}(t)$

$$\vec{Y}^{\times}(t_{i+1}) = \vec{Y}^{\times}(t_i) + \Delta t_i \cdot \vec{F}(t_i;\vec{Y}^{\times}(t_i)) \qquad (i = 0,1,..,N-1), \qquad (4.7)$$

где $\Delta t_i = t_{i+1} - t_i$.

Точки t_i называют узлами сетки разбиения отрезка $[t_0,T]$. *Шаг разбиения* Δt_i, вообще говоря, не обязан быть постоянным, однако здесь мы примем, что для $\forall i \in \{0,1,...,N-1\}$ $\Delta t_i \equiv h = (T - t_0)/N$.

Таким образом, формула (4.7) позволяет шаг за шагом построить решение системы (4.1) для дискретного набора точек отрезка $[t_0,T]$ исходя из начального условия (4.3). Разумно предположить, что построенное по правилу (4.7) приближенное решение должно в пределе $h \to 0$ (соответственно $N \to \infty$) сходиться к точному, коль скоро разностное отношение (4.4) в этом пределе дает производную $\vec{Y}'(t)$. Однако доказательство ***сходимости*** приближенного решения к точному решению оказывается не простым. Более того, исследование и доказательство сходимости можно считать одной из основных теоретических задач при построении численных методов. Данная задача будет рассмотрена здесь лишь на примере. Прежде, чем определить понятие сходимости, необходимо задать меру близости приближенного решения к точному решению. Естественный способ определить эту меру как максимальное по модулю отклонение

$\Delta y(t_i) \equiv y^{\times}{}_k(t_i) - y_k(t_i)$ (k = 1,2, …, n). Итак, примем за меру близости решений такое выражение:

$$\left\| \vec{Y}^{\times} - \vec{Y} \right\|_{(h)} \overset{def}{=} \max_{\substack{\{i \mid i=1,\ldots,N\} \\ \{k \mid k=1,\ldots,n\}}} \left| y^{\times}{}_k(t_i) - y_k(t_i) \right|, \qquad (4.8)$$

где $\vec{Y}^{\times} = \vec{Y}^{\times}(t)$ — вектор-функция приближенного решения ЗК (4.1), (4.3).

Иначе говоря, в качестве меры отклонения взята первая норма вектор-функции $\left(\vec{Y}^{\times}(t) - \vec{Y}(t)\right)$, заданной на дискретном множестве значений t. Кроме нормы (4.8) можно использовать и другие нормы (подробно см. гл.1).

Задав меру отклонения (4.8), легко определить понятие сходимости традиционным в анализе способом, а именно:

Определение 4.1. *Приближенное решение* $\vec{Y}^{\times}(t)$ назовем сходящимся к точному решению $\vec{Y}(t)$, если

$$\lim_{\substack{h \to 0 \\ (N \to \infty)}} \left\| \vec{Y}^{\times} - \vec{Y} \right\|_{(h)} = 0. \qquad (4.9)$$

Далее под абсолютной погрешностью вектор-функции $\vec{Y}^{\times}(t)$ приближенного решения задачи Коши будем понимать выражение (4.8).

Мера (4.8) есть и оценка *абсолютной* погрешности метода. Легко видеть, что погрешность метода Эйлера пропорциональна h. Действительно, в результате приближения (4.7) на первом шаге делается ошибка

$$\Delta\vec{Y}(t_1) = \| \vec{Y}^{\times}(t_0 + h) - \vec{Y}(t_0 + h) \| = \| \vec{Y}_0 + h \cdot \vec{Y}'(t_0) - \vec{Y}(t_0 + h) \| \approx \frac{h^2}{2} \| \vec{Y}''(t_0) \|.$$

(Обратим внимание, что в последнем соотношении используется норма $\left\|\vec{Y}^{\times}(t)-\vec{Y}(t)\right\| = \max\limits_{\{k|k=1,\ldots,n\}}\left|y_k^{\times}(t)-y_k(t)\right|$). При N-кратном применении формулы (4.7) по достижении узла t_N ошибка будет порядка $N\cdot h^2 = T\cdot h$. Строгая оценка погрешности метода приводит к тому же результату.

Теперь кратко обсудим вопросы, связанные со сходимостью и погрешностью метода Эйлера в рамках рассмотрения простой задачи Коши для ОДУ 1-го порядка.

Пример 4.1. Требуется решить численно на отрезке [0, T] ОДУ 1-го порядка

$$y'(t)=\alpha\, y(t) \tag{4.10}$$

при начальном условии

$$y(0)=y_0\,. \tag{4.11}$$

Дать оценки погрешности решения и доказать сходимость метода Эйлера.

Решение

Точное решение этой задачи Коши имеет вид

$$y(t)=y_0 e^{\alpha t}\,. \tag{4.12}$$

Для построения численного решения $y^{\times}(t)$ разбиваем область интегрирования [0,T] на N отрезков длиной $h=T/N$. Тогда узлы сетки находятся в точках $t_i = i\cdot h$ (i = 0,1,…,N). В соответствии с (4.7) для значений искомой функции в узлах получим рекуррентное соотношение

$$y^{\times}_{i+1} = y^{\times}{}_i + h\cdot\alpha\, y_i^{\times} \qquad (\text{i} = 0,1,\ldots,\text{N-1}), \tag{4.13}$$

где $y_i^{\times}\equiv y^{\times}(t_i)=y^{\times}(h\cdot i)$.

Рекуррентное соотношение (4.13) имеет следующее решение:

$$y_i^{\times} = (1 + h \cdot \alpha)^i \cdot y_0 . \qquad (4.14)$$

Покажем (не очень строго), что построенное приближенное решение в пределе $h \to 0$ сходится к точному решению. Действительно, для каждого значения $t > t_0$ найдется номер i такой, что $t_i \le t \le t_{i+1}$ или $t/h - 1 \le i \le t/h$. По фиксированному значению t и по найденному номеру i определим величину $\varepsilon_i(t,h)$ так, что $i(t,h) = i = t/h - \varepsilon_i(t,h)$ при этом $0 \le \varepsilon_i(t,h) \le 1$. Отсюда

$$\lim_{h\to 0} (1 + h \cdot \alpha)^i \, y_0 = y_0 \lim_{h\to 0} \left((1 + h \cdot \alpha)^{t/h} (1 + h \cdot \alpha)^{-\varepsilon_i(t,h)} \right) = y_0 e^{\alpha t},$$

т.е. приближенное решение стремится при $h \to 0$ к точному решению.

Для иллюстрации сходимости на рисунке 4.1 приведены графики численного решения задачи Коши (4.10), (4.11) с параметрами α=1, $y_0 = 1$ на отрезке [0,3] при значениях шага h = 0.15 и h = 0.0375.

Оценим теперь абсолютную $\Delta(y_i^{\times}) \equiv \left| y_i^{\times} - y(t_i) \right|$ и относительную $\delta(y_i^{\times}) = \left(\Delta(y_i^{\times}) \Big/ y(t_i) \right)$ погрешности решения (4.14) в узле t_i:

$$\begin{aligned} \Delta(y_i^{\times}) &= \left| (1 + h\alpha)^i - e^{\alpha i h} \right| \cdot |y_0| \approx \\ &\approx |y_0| \cdot e^{\alpha i h} \left| \exp\left(-\frac{i\alpha^2 h^2}{2} \right) - 1 \right| = |y_0| \cdot e^{\alpha t_i} \left(1 - \exp\left(-\frac{\alpha^2 t_i T}{2N} \right) \right), \end{aligned} \qquad (4.15)$$

$$\delta(y_i^{\times}) = 1 - \exp\left(-\frac{i\alpha^2 h^2}{2} \right) = 1 - \exp\left(-\frac{\alpha^2 t_i T}{2N} \right). \qquad (4.16)$$

При получении этих формул подразумевалось выполнение условия $h\alpha<<1$. Из формулы (4.16) видно, что для достижения приемлемой точности вычислений на всем отрезке $[0,T]$ необходимо потребовать выполнения условия

$$\delta_{\max} = \delta(y_N^{\times}) \approx \frac{\alpha^2 T^2}{2N} << 1 \qquad (4.17)$$

или, что то же

$$\delta_{\max} \approx \frac{\alpha^2 hT}{2} << 1. \qquad (4.17')$$

В этом случае выражения для погрешностей примут вид

$$\Delta(y_i^{\times}) \approx \frac{\alpha^2 h\, t_i}{2} |y_0| \exp(\alpha t_i), \qquad (4.18)$$

$$\delta\left(y_i^{\times}\right) \approx \frac{\alpha^2 h\, t_i}{2}. \qquad (4.19)$$

Таким образом, при выполнении условия (4.17) получаем, что погрешность вычислений пропорциональна шагу h разбиения (см. рис.4.2). Обратим внимание еще на два свойства данного решения. Во-первых, относительная погрешность приближенного (численного) решения растет квадратично относительно длины T отрезка интегрирования (при фиксированном N). Поэтому, для сохранения заданной точности, при увеличении T необходимо увеличивать количество узлов сетки разбиения по закону $N \cdot T^2$. Во-вторых, абсолютная погрешность указанного решения ведет себя по-разному в зависимости от знака параметра α. При $\alpha > 0$ $\Delta\left(y_i^{\times}\right)$ растет монотонно вместе с номером узла. При $\alpha < 0$ абсолютная

погрешность достигает максимума во внутренней точке отрезка $[0,T]$: $\max_i \Delta(y_i) \approx |\alpha| he \cdot |y_0|/2$ при $\left[|\alpha| \cdot t_i\right] = 1$ для $\alpha \cdot T > 1$ (см. рис. 4.3).

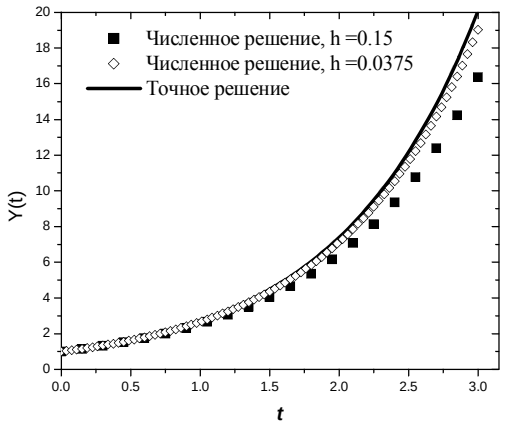


Рис. 4.1. Графики точного (сплошная кривая) и численных решений задачи Коши (4.10), (4.11) при h = 0.15 (точки) и h = 0.0375 (ромбы)

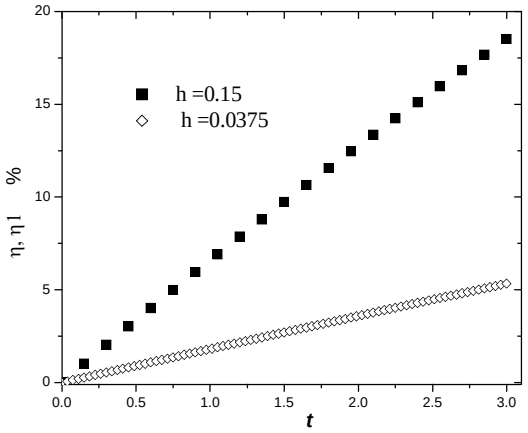


Рис. 4.2. Относительная погрешность (в %) численных решений, приведенных на рис. 4.1.

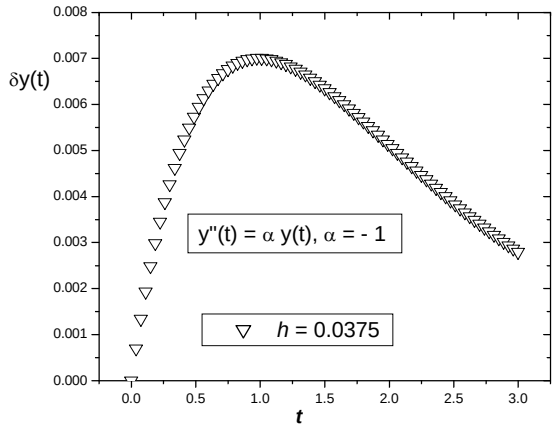


Рис. 4.3. Абсолютная погрешность численного решения ЗК (4.10), (4.11)

4.3. Метод Эйлера. Улучшение точности

Обобщение метода Эйлера связано с необходимостью повышения точности при заданном шаге сетки *h*.

Определение 4.2. Численный метод называют методом *k*-го порядка точности, если его погрешность пропорциональна h^k.

Метод Эйлера, следовательно, имеет *первый порядок точности*. Для уменьшения погрешности, очевидно, необходимо улучшить приближение (4.7). Этого можно достичь, например, с помощью использования разложения по формуле Тейлора в окрестности каждого из узлов t_i сетки. Например, для достижения второго порядка точности используем приближение

$$\vec{Y}^{\times}(t_{i+1}) = \vec{Y}^{\times}(t_i) + h\left(\vec{Y}^{\times}\right)'(t_i) + \frac{h^2}{2}\left(\vec{Y}^{\times}\right)''(t_i) = \vec{Y}_i^{\times} + h\vec{F}(t_i;\vec{Y}_i^{\times}) + \frac{h^2}{2}\left(\vec{Y}^{\times}\right)''_i. \quad (4.20)$$

Формула (4.20) позволяет вычислить $\vec{Y}^{\times}$ в каждой последующей точке по известным $\vec{Y}^{\times}$, $\left(\vec{Y}^{\times}\right)''$ в предыдущей точке. Следовательно, осталось определить $\left(\vec{Y}^{\times}(t)\right)''$. Естественный способ сделать это состоит в замене вторых производных

соответствующими разностными отношениями. Разностное отношение для второй производной (как и для высших производных) определяется по аналогии с (4.4). Следует при этом иметь в виду, что, поскольку знак Δt в правой части (4.4) произволен, разностное отношение определяется не единственным способом. Например, для второй производной можно использовать разностное отношение

$$\left(\vec{Y}^{\times}\right)''_{i} = \frac{1}{h^2}\left(\vec{Y}^{\times}_{i+1} - 2\vec{Y}^{\times}_{i} + \vec{Y}^{\times}_{i-1}\right). \qquad (4.21)$$

Тогда вместо (4.20) получим

$$\vec{Y}^{\times}_{i+1} = \vec{Y}^{\times}_{i-1} + 2\vec{F}(t_i; \vec{Y}^{\times}_{i}) \cdot h \qquad (i = 1,2,\ldots, N\text{-}1). \qquad (4.22)$$

Видно, что в рекуррентном соотношении (4.22) задействованы три соседних узла. Таким образом, желая повысить порядок точности метода, мы несколько усложняем процедуру вычислений. При построении по аналогии с (4.21), (4.22) численного метода *k-го порядка точности* получим рекуррентное соотношение, в котором величина $\vec{Y}^{\times}_{i+1}$ выражена через значения $\vec{Y}^{\times}_{i+1-j}$ $(j = 1,\ldots,m)$ в m предыдущих узлах. Численные схемы, приводящие к таким соотношениям, называются *явными* m-шаговыми. Численные схемы, которые вообще не приводят к рекуррентным соотношениям, называют *неявными*. Методы, в которых разностные отношения строятся на дополнительно введенных промежуточных узлах (между i-м и i+1-м), относятся к методам *Рунге – Кутта* [5].

Здесь мы рассмотрим более простой способ повышения точности численного решения ЗК. Он является одношаговым и основан на получении высших производных в ряде Тейлора повторным дифференцированием уравнения (4.1). Например, имея в виду (4.20), из (4.1) получим

$$\vec{Y}''(t) = \frac{\partial \vec{F}(t;\vec{Y}(t))}{\partial t} + \sum_{l=1}^{n} \frac{\partial \vec{F}(t;\vec{Y}(t))}{\partial y_l} \frac{dy_l(t)}{dt} = \\ = \frac{\partial \vec{F}(t;\vec{Y}(t))}{\partial t} + \sum_{l=1}^{n} \frac{\partial \vec{F}(t;\vec{Y}(t))}{\partial y_l} f_l(t;\vec{Y}(t)). \quad (4.23)$$

Продемонстрируем данный способ на примере ЗК (4.10), (4.11).

В этом случае

$$y_{i+1}^{\times} = y_i^{\times} + \alpha h y_i^{\times} + \frac{h^2}{2}\left(y_i^{\times}\right)''(t_i) = y_i^{\times} + \alpha h y_i^{\times} + \alpha^2 \frac{h^2}{2} y^{\times}{}_i. \quad (4.24)$$

Отсюда следует

$$y_i^{\times} = \left(1 + h \cdot \alpha + \frac{h^2}{2} \cdot \alpha^2\right)^i \cdot y_0. \quad (4.25)$$

Легко получить выражения для абсолютной и относительной погрешности в *i*-м узле:

$$\Delta(y_i^{\times}) \approx \frac{\alpha^3 h^2 t_i}{6} |y_0| \exp(\alpha t_i), \quad (4.26)$$

$$\delta(y_i^{\times}) \approx \frac{\alpha^3 h^2 t_i}{6}. \quad (4.27)$$

Из (4.26) и (4.27) видно, что действительно описанный метод и в самом деле имеет второй порядок точности. На рисунке 4.4. представлены графики приближенных решений первого порядка точности (черные квадраты), второго порядка точности (светлые квадраты), и точное решение ЗК (4.10), (4.11). И то, и другое приближенные решения получены на сетке с шагом $h = 0.15$ (количество

узлов $N = 20$). Рисунок 4.4. хорошо иллюстрирует эффективность численных методов высокого порядка точности.

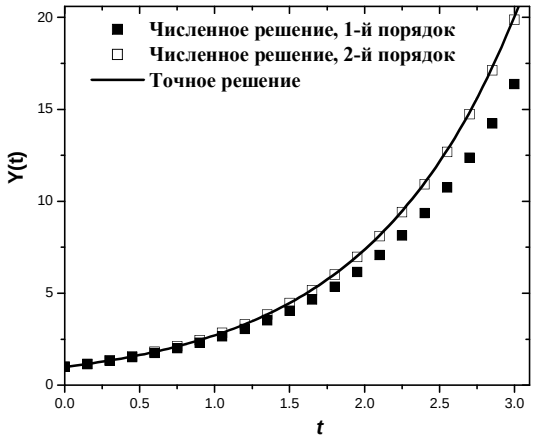


Рис.4.4. Численное сравнение решений задачи Коши

4.4. Контрольные вопросы

1. Для задачи Коши

$$\begin{aligned} &y' + 4y = \sin x, \\ &y(0) = 2, \end{aligned}$$

выписать формулы решения для явного и неявного метода Эйлера.

2. Для задачи Коши

$$\begin{cases} y' = y + 2z - 9x, \\ z' = 2y + z - 4e^x, \end{cases}$$
$$\begin{aligned} &y(0) = 1, \\ &z(0) = 2, \end{aligned}$$

выписать формулы решения для явного и неявного метода Эйлера.

3. Для задачи Коши

$$y' = 0,04y,$$
$$y(0)=1000,$$

найти точное решение и сравнить его с приближенными решениями, полученными явным методом Эйлера с шагами $h=1;0.5;0.25$ в точке $x=1$. Оценить погрешность.

4. Для задачи Коши

$$y' = -0,01y,$$
$$y(0)=100,$$

найти точное решение при $x=100$ и сравнить его с приближенными решениями, полученными явным методом Эйлера при $h=150.20$. Оценить погрешность.

5. Для задачи Коши

$$y' = 4-2x,$$
$$y(0)=2,$$

найти точное решение и сравнить его с приближенным, полученным явным методом Эйлера при $h=0.5;0.25$. Оценить погрешность.

6. Для задачи Коши

$$y' = -\frac{x}{y},$$
$$y(0)=20,$$

найти точное решение и приближенные решения модифицированным методом Эйлера на $[0,12]$ при $h=2$. Сравнить их. Оценить погрешность.

7. Для задачи Коши

$$y' = \frac{6y + 2x + 1}{x},$$
$$y(1) = -2,$$

найти приближенное решение модифицированное методом Эйлера на $[1,4]$ при шаге $h = 2{,}1$. Сравнить полученные решения.

4.5. Практические задания. Компьютерный практикум

4.5.1. Реализация метода Эйлера в MathCad

Рассмотрим самую простую реализацию численного решения задачи (4.10), (4.11) по методу Эйлера в MathCad.

1) Задаем правую часть уравнения (4.10), ее параметры (в данном случае параметр α), правую границу отрезка интегрирования T, начальное значение $y(0) = y_0$:

$\alpha := 1$

$f(x, y) := \alpha \cdot y$

$T := 3 \qquad y0 := 1$

2) Задаем количество узлов сетки разбиения N, вычисляем шаг разбиения h, определяем дискретную переменную (массив индексов) i:

$N := 20$

$h := \frac{T}{N}$

$h = 0.15$

$i := 0, 1 .. N - 1$

3) Реализуем алгоритм метода Эйлера:

$yeuler_0 := y0$

$yeuler_{i+1} := yeuler_i + h \cdot f(i \cdot h, yeuler_i)$

Здесь введена индексированная переменная yeuleri – приближенное решение задачи (4.10), (4.11) на сетке узлов.

4.5.2. Практические задания

Задание 4.1.

1) Получить точное решение ЗК для ОДУ, заданного на отрезке [0,*T*]. Варианты ОДУ, начальных условий и значения параметра *T* приведены в таблице 4.1.

2) Получить численное решение ЗК своего варианта методом Эйлера для нескольких значений шага. Построить графики точного и численного решений. Убедиться в сходимости метода Эйлера.

3) Определить абсолютную и относительную погрешности численного решения, построив соответствующие графики.

Задание 4.2.

1) Получить численное решение ЗК для ОДУ с помощью процедуры, встроенной в MathCad. Исследовать зависимость точности численного решения от величины шага сетки. Варианты заданий приведены в таблице 4.2.

2) Получить численное решение ЗК для ОДУ с помощью метода Эйлера и сравнить результаты с результатами п. 1.

3) Исследовать численные решения на сходимость и точность с помощью графиков.

Перечень вариантов к заданию 4.1.

Таблица 4.1

Вариант	ОДУ	*T*	Начальные условия
1	$y'' + 0.2\,y' + y = e^{-t}$	4π	$y(0)=1,\ y'(0)=0$
2	$y'' + 0.2\,y' + y = \cos(t)$	4π	$y(0)=1,\ y'(0)=0$
3	$y'' + 0.2\,y' + y = \sin(t)$	4π	$y(0)=1,\ y'(0)=0$
4	$y'' + 0.2\,y' + y = \cos(t)$	4π	$y(0)=0,\ y'(0)=1$

5	$y'' + 0.2\,y' + y = \sin(t)$	4π	$y(0)=0,\ y'(0)=1$
6	$y'' + 0.2\,y' + y = \cos(t)$	4π	$y(0)=1,\ y'(0)=0$
7	$y'' + 2\,y' + y = e^{-t}$	4π	$y(0)=0,\ y'(0)=1$
8	$y'' + 4\,y' + y = \cos(t)$	4π	$y(0)=1,\ y'(0)=0$
9	$y'' - 0.2\,y' + y = \sin(t)$	4π	$y(0)=0,\ y'(0)=1$
10	$y'' + y = \sin(t)$	4π	$y(0)=0,\ y'(0)=0$
11	$y'' + y = \cos(t)$	4π	$y(0)=0,\ y'(0)=0$
12	$y'' + y = \sin(t)$	4π	$y(0)=0,\ y'(0)=0$
13	$y'' + y = \cos(t)$	4π	$y(0)=0,\ y'(0)=0$
14	$y^{(4)} + y = 0$	4π	$y(0)=1,\ y'(0)=1/\sqrt{2}$, $y''(0)=0,\ y'''(0)=-1/\sqrt{2}$
15	$y^{(4)} + y = 0$	4π	$y(0)=1,\ y'(0)=-1/\sqrt{2}$, $y''(0)=0,\ y'''(0)=1/\sqrt{2}$

Перечень вариантов к заданию 4.2.

Таблица 4.2

Вариант	ОДУ	T	Начальные условия
1	$y'' + 0.2\,y' + y + y^3 = e^{-t}$	4π	$y(0)=1,\ y'(0)=0$
2	$y'' + 0.2\,y' + y - y^3 = \cos(t)$	4π	$y(0)=1,\ y'(0)=0$
3	$y'' + 0.2\,y' + y + y^3 = \sin(t)$	4π	$y(0)=1,\ y'(0)=0$
4	$y'' + 0.2\,(1+0.2\,y2)\,y' + y = \cos(t)$	4π	$y(0)=0,\ y'(0)=1$
5	$y'' + 0.2\,y' + \sin(y) = \sin(t)$	4π	$y(0)=0,\ y'(0)=1$
6	$y'' + 0.2\,y' + \sin(2t)\,y = \cos(t)$	4π	$y(0)=1,\ y'(0)=0$
7	$y'' + 2\,y' + \cos(2t)\,y = e^{-t}$	4π	$y(0)=0,\ y'(0)=1$
8	$y'' + 4\,y' + (1+0.3\cos(2t))\,y = \cos(t)$	4π	$y(0)=1,\ y'(0)=0$
9	$y'' - 0.2\,y\,y' + y = \sin(t)$	4π	$y(0)=0,\ y'(0)=1$
10	$y'' + \sin(y) = \sin(t)$	4π	$y(0)=0,\ y'(0)=0$
11	$y'' + \sin(y) = \cos(t)$	4π	$y(0)=0,\ y'(0)=0$
12	$y'' + \sin(y) = e^{-t}\sin(t)$	4π	$y(0)=0,\ y'(0)=0$
13	$y'' + \sin(y) = e^{-t}\cos(t)$	4π	$y(0)=0,\ y'(0)=0$
14	$y^{(4)} + 6\,y\,y' + y = 0$	4π	$y(0)=1,\ y'(0)=1/\sqrt{2}$, $y''(0)=0,\ y'''(0)=-1/\sqrt{2}$
15	$y^{(4)} + 6\,y\,y' + y = 0$	4π	$y(0)=1,\ y'(0)=-1/\sqrt{2}$, $y''(0)=0,\ y'''(0)=1/\sqrt{2}$

V. МЕТОД ФУРЬЕ ДЛЯ ЛИНЕЙНЫХ УРАВНЕНИЙ В ЧАСТНЫХ ПРОИЗВОДНЫХ ВТОРОГО ПОРЯДКА

5.1. Элементы общей теории уравнений в частных производных (УЧП)

Определение 5.1. Будем понимать под уравнением в частных производных второго порядка соотношение вида:

$$F\left(x_i, u(x_i), u_{x_i}, u_{x_i x_j}\right) = 0, \quad i, j = 1, 2. \tag{5.1}$$

Здесь u_{x_i}, $u_{x_i x_j}$, $u_{x_i^2}$ - обозначения частных производных первого и второго порядка функции $u(x_i)$ по соответствующим переменным $\frac{\partial u}{\partial x_i}, \frac{\partial^2 u}{\partial x_i \partial x_j}, \frac{\partial^2 u}{\partial x^2_i}$.

Определение 5.2. Классическим решением уравнения (5.1) назовем функцию $u(x_i)$ непрерывную и имеющую непрерывные частные производные, входящие в (5.1) и обращающую равенства (5.1) в тождество относительно x_i.

Определение 5.3. Если известная функция F линейна относительно искомой функции $u(x_i)$ и ее производных, то уравнение (5.1) называется линейным.

УЧП второго порядка с переменными коэффициентами

$$a(x_1, x_2)u_{x_1^2} + 2b(x_1, x_2)u_{x_1 x_2} + c(x_1, x_2)u_{x_2^2} = f\left(x_1, x_2, u, u_{x_1}, u_{x_2}\right), \tag{5.2}$$

невырожденным линейным преобразованием

$$\begin{cases} \xi_1 = \alpha_{11} x_1 + \alpha_{21} x_2; \\ \xi_2 = \alpha_{12} x_1 + \alpha_{22} x_2 \end{cases} \tag{5.3}$$

приводятся к так называемому каноническому виду, в соответствии со знаком выражения, называемого дискриминантом

$$D(x_1, x_2) = b^2(x_1, x_2) - a(x_1, x_2)c(x_1, x_2).$$

Возможны следующие случаи:

1) $D > 0$ $\qquad u_{\xi_1^2} - u_{\xi_2^2} = f(\xi_1, \xi_2, u, u_{\xi_1}, u_{\xi_2})$ (5.4)

(5.4) – уравнение гиперболического типа.

2) $D = 0$ $\qquad u_{\xi_2^2} = f(\xi_1, \xi_2, u, u_{\xi_1}, u_{\xi_2})$ (5.5)

(5.5) – уравнение параболического типа.

3) $D < 0$ $\qquad u_{\xi_1^2} + u_{\xi_2^2} = f(\xi_1, \xi_2, u_{\xi_1}, u_{\xi_2})$ (5.6)

(5.6) – уравнение эллиптического типа.

В качестве примеров УЧП уравнений математической физики соответствующих типов приведем:

1) гиперболический тип

$$u_{t^2} - a^2 u_{x^2} = f(x,t) \quad - \qquad (5.7)$$

уравнение малых поперечных колебаний математической струны или же малых продольных колебаний стержня [12]. В двумерном случае

$$u_{t^2} - a^2 \Delta u = f(x_1, x_2, t) \quad - \qquad (5.8)$$

уравнение малых поперечных колебаний мембраны [12];

2) параболический тип

$$u_t - a^2 \Delta_{x_i} u = f(x_i, t) \quad - \qquad (5.9)$$

уравнение теплопроводности или уравнение диффузии [12],

где $\Delta_{x_i} u = \sum_{i=1}^{n} u_{x_i^2}$ - n-мерный оператор Лапласа;

3) эллиптический тип

$$\Delta_{x_i} u = f(x_i) \quad - \tag{5.10}$$

уравнение Пуассона описывает установившиеся (стационарные) процессы колебаний или распространения тепла [12].

Еще один пример - уравнение Гельмгольца [12]

$$\Delta_{x_i} u + a^2 u = f(x_i).$$

Для корректной постановки задач о решении УЧП приведенных типов, рассмотрим два рода задач: задачи Коши и краевые задачи [2].

Формулировка задачи Коши: нужно найти решение уравнения (5.7), (5.8) для $t > 0, \quad x \in \Omega \subset R^n$ (здесь Ω - ограниченная область с гладкой границей S) при начальных условиях.

$$\begin{cases} u|_{t=0} = \varphi(x_i), \\ u_t|_{t=0} = \psi(x_i). \end{cases} \tag{5.11}$$

Формулировка краевой задачи: пусть уравнения (5.7), (5.8) рассматриваются в цилиндре $G = (0,T) \times \mathbf{\Omega}$ (знак $\times$ - обозначает декартово произведению множеств). Нужно найти функцию $u(x,t)$, удовлетворяющую уравнениям (5.7) или (5.8), начальным условиям (5.11) ,а также одной из следующих групп краевых условий:

$$u|_{x=0} = g_1(t), \quad u|_{x=l} = g_2(t), \tag{5.12}$$

$$u_x\big|_{x=0} = g_1(t), \quad u_x\big|_{x=l} = g_2(t), \tag{5.13}$$

$$\big(u_x + \alpha(t)u\big)\big|_{x=0} = g_1(t), \quad \big(u_x + \beta(t)u\big)\big|_{x=l} = g_2(t), \tag{5.14}$$

$$u\big|_{\Gamma} = g(t, x_i'), \tag{5.15}$$

$$u_\nu\big|_{\Gamma} = g(t, x_i'), \tag{5.16}$$

$$\big(u_\nu + \alpha(x_i', t)u\big)\big|_{\Gamma} = g(t, x_i'), \tag{5.17}$$

где $\Gamma = [0,T] \times S$ - боковая поверхность цилиндра G, $x' \in S$, ν - внешняя нормаль к Γ, u_ν - нормальная производная функции u.

Получим первую (5.7), (5.11), (5.12) или (5.8), (5.11), (5.15), вторую (5.7), (5.11), (5.13) или (5.8), (5.11), (5.16), и третью (5.7), (5.11), (5.14) или (5.8), (5.11), (5.17) - смешанную краевую задачу для УЧП.

Физически условия (5.11) соответствуют заданию начальных отклонений и начальных скоростей точек струны или мембраны. Краевые условия (5.12) означают, что концы струны движутся по заданному закону, (5.13) – по концам струны приложены определенные силы. (5.14) – на концах струны присутствуют и упругие силы.

Для уравнения (5.9) ставятся аналогичные задачи, только в начальных условиях (5.11) отсутствует производная по t. Для уравнения (5.10) задается только краевое условие одного из трех видов (5.15—5.17).

5.2. Метод Фурье для уравнения колебаний

Рассмотрим общую схему метода разделения переменных (Фурье) на примере краевой задачи для неоднородного уравнения колебаний [3].

Сформулируем краевую задачу для одномерного уравнения колебаний

$$u_{t^2} - a^2 u_{x^2}(x,t) = f(x,t), \tag{5.18}$$

$$u\big|_{t=0} = \varphi(x); \quad u_t\big|_{t=0} = \psi(x), \tag{5.19}$$

$$u\big|_{\Gamma} = 0. \tag{5.20}$$

Задача состоит в отыскании функции $u(x,t)$, удовлетворяющей уравнению (5.18) и условиям (5.19), (5.20) в области $G = (0,l)\times(0,T)$ - с границей $\Gamma = \{0,l\}\times[0,T]$.

Решим методом Фурье задачу (5.18) – (5.20). Будем искать нетривиальные решения в виде $u(x_i,t) = X(x_i)T(t)$, Подставив предполагаемую форму решения в уравнение (5.18) и разделив переменные, получим

$$\frac{T''(t)}{T(t)} = \frac{a^2 X''(x)}{X(x)} = -\lambda = \text{const}.$$

Поэтому функции $T(t)$ и $X(x)$ должны быть определены как решения дифференциальных уравнений

$$X''(x) + \lambda X(x) = 0; \tag{5.21}$$

$$T''(t) + a^2 T(t) = 0. \tag{5.22}$$

Граничные условия (5.20) дают:

$$u(0,t) = X(0)T(t) \equiv 0, \quad u(l,t) = X(l)T(t) \equiv 0.$$

Откуда заключаем, что функция $X(x)$ должна удовлетворять дополнительным условиям

$$X(0) = X(l) = 0. \tag{5.23}$$

Итак, мы приходим к задаче о собственных значениях дифференциального оператора, так называемой задаче Штурма-Лиувилля: найти те значения параметра λ, при которых существуют ненулевые решения граничной задачи (5.21), (5.23) [3].

Рассмотрим различные случаи, когда $\lambda < 0, \quad \lambda = 0$ или $\lambda > 0$.

1) При $\lambda < 0$ общее решение ОДУ (5.21)

$$X(x) = c_1 e^{\sqrt{-\lambda}x} + c_2 e^{-\sqrt{-\lambda}x}.$$

Граничные условия (5.23) дают

$$\begin{cases} c_1 + c_2 = 0; \\ c_1 e^{\alpha} + c_2 e^{-\alpha} = 0, \quad \text{где} \quad \alpha = l\sqrt{-\lambda}. \end{cases}$$

$\det A = e^{-\alpha} - e^{\alpha} \neq 0$, так как $\alpha > 0 \Rightarrow c_1 = c_2 = 0 \Rightarrow X(x) \equiv 0$, т.е. вспомогательная задача не имеет нетривиальных решений.

2) При $\lambda = 0$ общее решение ОДУ (5.21)

$$X(x) = c_1 x + c_2.$$

Граничные условия (5.23) дают

$$\begin{cases} c_2 = 0; \\ c_1 l = 0. \end{cases}$$

$\Rightarrow c_1 = c_2 = 0 \Rightarrow X(x) \equiv 0$.

3) При $\lambda > 0$ общее решение ОДУ (5.21) может быть записано в виде

$$X(x) = c_1 \cos\sqrt{\lambda}x + c_2 \sin\sqrt{\lambda}x.$$

Граничные условия (5.23) дают

$$\begin{cases} c_1 = 0; \\ c_1 \sin\sqrt{\lambda} l = 0. \end{cases}$$

Так как $X(x) \neq 0$, то $c_2 \neq 0$ и поэтому $\sin\sqrt{\lambda} l = 0$ или $\sqrt{\lambda} = \dfrac{n\pi}{l}$.

Таким образом, нетривиальные решения задачи (5.21), (5.23) возможны при собственных значениях

$$\lambda = \lambda_n = \left(\frac{n\pi}{l}\right)^2,$$

соответствующие собственные функции

$$X_n(x) = \sin\lambda_n x.$$

Эта задача имеет счетное множество собственных функций X_n и собственных значений λ_n, причем все $\lambda_n > 0$. Кроме того, все собственные функции $X_n, X_m\ (m \neq n)$ попарно ортогональны между собой и их можно нормировать, т.е. считать ортонормированными и систему $\{X_n\}$ — полной [15].

Примечание 1. Например, в качестве нормы ||.|| — может быть выбрана норма в классе функций, интегрируемых с квадратом, т.е.

$$\|X\| = \sqrt{\int_\Omega X^2(x_i)dx}.$$

Ортонормируемость системы функции $\{X_n\}_{n=1}^{\infty}$ в метрическом пространстве с выбранной нормой понимается как равенство нулю скалярного произведения $\int_\Omega X_k X_m dx = 0$, при всех $k \neq m$ и $\|X_n\| = 1$ для любого n.

Для отыскания функций $T_n(t)$ построим последовательность решений задач Коши

$$T_n''(t) + \lambda_n T_n(t) = f_n(x_i), \tag{5.24}$$

$$T_n(0) = \varphi_n; \quad T_n'(0) = \psi_n, \tag{5.25}$$

где

$$\varphi_n = \int_\Omega \varphi(x) X_n(x)\,dx, \quad \psi_n = \int_\Omega \psi(x) X_n(x)\,dx. \tag{5.26}$$

Тогда

$$T_n(t) = \varphi_n \cos\sqrt{\lambda_n}t + \frac{\psi_n}{\sqrt{\lambda_n}}\sin\sqrt{\lambda_n}t + \frac{1}{\sqrt{\lambda_n}}\int_0^\tau \sin\sqrt{\lambda_n}(t-\tau)\cdot f_n(\tau)d\tau. \tag{5.27}$$

Решение задачи (5.18) – (5.20) запишется в виде

$$u(x,t) = \sum_{n=1}^{\infty} T_n(t) X_n(x). \tag{5.27′}$$

Замечание 5.1.1. Для решения общей первой краевой задачи уравнения колебаний (5.7), (5.11), (5.12) ее можно привести к краевой задаче с однородными граничными условиями. Для этого построим функцию $v(x,t)$ для которой выполняются граничные условия (5.12). Например, можно взять функцию, линейную относительно переменной x:

$$v(x,t) = A(t)x + B(t).$$

Условия (5.12) дают

$$v(0,t) = g_1(t) = B(t),$$
$$v(l,t) = g_2(t) = A(t)l + B(t).$$

Следовательно

$$v(x,t) = g_1(t) + \frac{x}{l}(g_2(t) - g_1(t)).$$

Теперь введем новую неизвестную функцию $\omega(x,t)$, полагая, что

$$u(x,t) = v(x,t) + \omega(x,t).$$

Подставляя далее $u(x,t)$ в (5.7), (5.11), (5.12) получаем краевую задачу для определения $\omega(x,t)$:

$$v_{t^2} - a^2 v_{x^2} = F(x,t),\ 0 < x < l,\ t > 0;$$

$$v|_{t=0} = \Phi(x),\ v_t|_{t=0} = \Psi(x);$$

$$v|_{x=0} = 0,\ v|_{x=l} = 0,$$

которая аналогична задаче (5.18)—(5.20).

Примечание 2. Для обеспечения сходимости ряда решений (5.27′) и рядов, получаемых двукратным почленным дифференцированием этого ряда по t и x, установим ограничения на функции $\varphi(x)$ и $\psi(x)$.

Теорема 5.1. [1] Если $\varphi(x) \in C^{(2)}[0,l]$, $\psi(x) \in C^{(1)}[0,l]$, кроме того $\varphi(x)$ имеет третью, а $\psi(x)$ - вторую кусочно-непрерывную производную на $[0,l]$ и выполняются условия согласования $\varphi(0) = \varphi(l) = 0$, $\varphi''(0) = \varphi''(l) = 0$, $\psi(0) = \psi(l) = 0$. Тогда сумма ряда (5.27′) является классическим решением задачи (5.18) – (5.20).

Вообще говоря, условия теоремы 5.1 могут быть ослаблены. В этом случае ряд (5.27′) является так называемым обобщенным решением краевой задачи [4]. Тогда сходимость ряда решений и его производных следует понимать в смысле сходимости в среднем или слабой сходимости.

Решения задач о свободных колебаниях ограниченной струны или стержня с однородными граничными условиями второго или третьего рода могут быть построены в виде функциональных рядов аналогичной структуры, отличающихся лишь решениями соответствующих вспомогательных задач Штурма-Лиувилля.

5.3. Метод Фурье для уравнения теплопроводности

Рассмотрим общую схему метода разделения переменных (Фурье) на примере краевых задач для неоднородного уравнения теплопроводности [3].

Решение краевых задач для уравнения теплопроводности методом Фурье. Сформулируем задачу об отыскании нестационарного температурного поля $u(x,t)$ в тонком стержне длины l, имеющем в начальный момент времени температуру $\varphi(x)$, если на поверхностях $x=0$ и $x=l$ этого слоя происходит теплообмен с окружающей средой, имеющей нулевую температуру. Требуется найти решение линейного однородного параболического уравнения

$$u_t = a^2 u_{x^2},\ 0 < x < l,\ t > 0, \tag{5.28}$$

удовлетворяющее при $t=0$ начальному условию

$$u(x,0) = \varphi(x),\ 0 \le x \le l, \tag{5.29}$$

и однородными граничными условиями третьего рода

$$\begin{aligned} &\left(u_x + \alpha u\right)\big|_{x=0} = 0,\ t \ge 0; \\ &\left(u_x + \beta u\right)\big|_{x=l} = 0,\ t \ge 0. \end{aligned} \tag{5.30}$$

Следуя методу Фурье разделения переменных, нетривиальные решения уравнения (5.28), удовлетворяющие граничным условиям (5.30), будем искать в виде

$$u(x,t) = X(x)T(t) \ne 0\,. \tag{5.31}$$

Подставив предполагаемую форму решения (5.31) в уравнение (5.28) и разделив переменные, получим

$$\frac{1}{a^2}\frac{T'(t)}{T(t)}=\frac{X''(x)}{X(x)}=-\lambda=\text{const}.$$

Поэтому функции $T(t)$ и $X(x)$ должны быть определены как решения дифференциальных уравнений

$$T'(t)+\lambda a^2 T(t)=0; \tag{5.32}$$

$$X''(x)+\lambda X(x)=0. \tag{5.33}$$

Граничные условия (5.30) с учетом (5.31) дают условия для функции $X(x)$ в виде

$$\begin{aligned} &X'(0)+\alpha X(0)=0;\\ &X'(l)+\beta X(l)=0. \end{aligned} \tag{5.34}$$

Задача Штурма-Лиувилля (5.33), (5.34) имеет нетривиальные решения только при определенных, собственных значения $\lambda_n=\left(\frac{\mu_n}{l}\right)^2$, $n=1,2,\ldots$, которые можно выразить через неотрицательные корни μ_n трансцендентного уравнения

$$\operatorname{ctg}\mu=\frac{\mu^2-\alpha\beta l^2}{(-\beta l+\alpha l)\mu}, \tag{5.35}$$

а соответствующие им собственные функции $X_n(x)$ имеют вид

$$X_n(x)=\sin\left(\sqrt{\lambda_n}x+\theta_n\right),\ \theta_n=-\operatorname{arctg}\frac{\sqrt{\lambda_n}}{\alpha}.$$

Квадраты норм этих функций

$$\|X_n\|^2 = \frac{l}{2}\left\{1 + \frac{\left(-\mu_n^2 + \alpha\beta l^2\right)\left(-\beta l + \alpha l\right)}{\left(\mu_n^2 + \alpha^2 l^2\right)\left(\mu_n^2 + \beta^2 l^2\right)}\right\}.$$

При $\lambda = \lambda_n$ для выражения (5.32) запишем общее решение

$$T_n(t) = C_n e^{-\lambda_n a^2 t},\ C_n = \text{const}. \tag{5.36}$$

Подставив найденные функции $X_n(x)$ и $T_n(t)$ в выражение (5.31), получим частные решения уравнения (5.28), удовлетворяющие граничным условиям (5.30)

$$u_n(x,t) = T_n(t)X_n(x) = C_n e^{-\lambda_n a^2 t}\sin\left(\sqrt{\lambda_n}x + \theta_n\right).$$

Составим формально ряд, членами которого являются функции $u_n(x,t)$:

$$u(x,t) = \sum_{n=1}^{\infty} C_n e^{-\lambda_n a^2 t}\sin\left(\sqrt{\lambda_n}x + \theta_n\right). \tag{5.37}$$

Функция $u(x,t)$ удовлетворяет граничным условиям (5.30), так как этим условиям удовлетворяет каждый член ряда (5.37).

Определим коэффициенты C_n так, чтобы выполнялось начальное условие. Подставляя ряд (5.37) в (5.29), получаем

$$u(x,0) = \sum_{n=1}^{\infty} C_n \sin\left(\sqrt{\lambda_n}x + \theta_n\right) = \varphi(x). \tag{5.38}$$

Это соотношение представляет собой разложение функции $\varphi(x)$ в ряд Фурье по системе ортогональных на отрезке $0 \le x \le l$ собственных функций $X_n(x) = \sin\left(\sqrt{\lambda_n}x + \theta_n\right)$, $n = 1,2,\ldots$, а коэффициенты C_n являются коэффициентами Фурье и определяются по формуле

$$C_n = \frac{1}{\|X_n\|^2}\int_0^l \varphi(x)X_n(x)dx. \qquad (5.39)$$

Можно показать, что если функция $\varphi(x)$ кусочно-непрерывная на отрезке $[0,l]$, то ряд (5.37) с коэффициентами C_n, определяемыми по формуле (5.39), удовлетворяет уравнению (5.28) в области $0 < x < l, t > 0$, т.е. этот ряд сходится и его можно дифференцировать почленно дважды по x и один раз по t [15].

Рассмотрим задачу для неоднородного уравнения теплопроводности

$$u_t - a^2 u_{x^2} = f(x,t),\ 0 < x < l,\ t > 0, \qquad (5.40)$$

с начальным условием

$$u(x,0) = \varphi(x),\ 0 \le x \le l, \qquad (5.41)$$

и граничными условиями

$$\begin{aligned} &(u_x + \alpha u)\big|_{x=0} = 0; \\ &(u_x + \beta u)\big|_{x=l} = 0. \end{aligned} \qquad (5.42)$$

Решение этой задачи будем искать в виде ряда Фурье по системе собственных функций $X_n(x) = \sin\left(\sqrt{\lambda_n}x + \theta_n\right)$ задачи на собственные значения (5.33), (5.34), т.е. в форме разложения

$$u(x,t) = \sum_{n=1}^{\infty} T_n(t)X_n(x) = \sum_{n=1}^{\infty} T_n(t)\sin\left(\sqrt{\lambda_n}x + \theta_n\right). \qquad (5.43)$$

считая при этом t параметром.

Ряд (5.43) удовлетворяет граничным условиям (5.42). Поэтому функции $T_n(t)$ следует определить так, чтобы ряд (5.43) удовлетворял уравнению (5.40) и начальному условию (5.41).

Учитывая полноту системы собственных функций, представим функции $f(x,t)$ и $\varphi(x)$ в виде следующих рядов Фурье:

$$\begin{aligned} f(x,t) &= \sum_{n=1}^{\infty} f_n(t) X_n(x) = \sum_{n=1}^{\infty} f_n(t) \sin\left(\sqrt{\lambda_n}\, x + \theta_n\right), \\ \varphi(x) &= \sum_{n=1}^{\infty} \varphi_n X_n(x) = \sum_{n=1}^{\infty} \varphi_n \sin\left(\sqrt{\lambda_n}\, x + \theta_n\right), \end{aligned} \tag{5.44}$$

где $f_n(t)$ и φ_n — коэффициенты Фурье, определяемые по формулам

$$f_n(t) = \frac{1}{\|X_n\|^2} \int_0^l f(x,t) X_n(x)\, dx = \frac{1}{\|X_n\|^2} \int_0^l f(x,t) \sin\left(\sqrt{\lambda_n}\, x + \theta_n\right) dx; \tag{5.45}$$

$$\varphi_n = \frac{1}{\|X_n\|^2} \int_0^l \varphi(x) X_n(x)\, dx = \frac{1}{\|X_n\|^2} \int_0^l \varphi(x) \sin\left(\sqrt{\lambda_n}\, x + \theta_n\right) dx.$$

Подставляя предполагаемую форму решения (5.43) и разложение (5.44) для функции $f(x,t)$ в уравнение (5.40) и заменяя при этом $X_n''(x)$ на $-\lambda_n X_n(x)$, получаем

$$\sum_{n=1}^{\infty} \left[T_n'(t) + \lambda_n a^2 T_n(t) - f_n(t)\right] X_n(x) = 0.$$

Это соотношение, а значит, и уравнение (5.40) будут удовлетворены, если все коэффициенты разложения равны нулю, т.е.

$$T_n'(t) + \lambda_n a^2 T_n(t) = f_n(t). \tag{5.46}$$

Из начального условия (5.41) с учетом (5.43) и (5.44) находим

$$u(x,0)=\sum_{n=1}^{\infty}T_n(0)X_n(x)=\sum_{n=1}^{\infty}\varphi_n X_n(x),\ 0<x<l,$$

откуда

$$T_n(0)=\varphi_n. \tag{5.47}$$

Таким образом, для нахождения искомой функции $T_n(t)$ приходим к задаче Коши (5.46), (5.47) для обыкновенного линейного неоднородного дифференциального уравнения первого порядка. Решение этой задачи может быть найдено методом вариации постоянной. Оно имеет вид

$$T_n(t)=\varphi_n e^{-\lambda_n a^2 t}+\int_0^t f_n(\tau)\,e^{-\lambda_n a^2(t-\tau)}d\tau.$$

Подставляя функции $T_n(t),\ n=1,2,\ldots$, в разложение (5.43), находим решение исходной задачи (5.40)—(5.42) в следующей форме:

$$\begin{aligned}u(x,t)&=\sum_{n=1}^{\infty}\varphi_n e^{-\lambda_n a^2 t}X_n(x)+\sum_{n=1}^{\infty}\left[\int_0^t f_n(\tau)e^{-\lambda_n a^2(t-\tau)}d\tau\right]X_n(x)=\\&=\sum_{n=1}^{\infty}\varphi_n e^{-\lambda_n a^2 t}\sin\left(\sqrt{\lambda_n}x+\theta_n\right)+\sum_{n=1}^{\infty}\left[\int_0^t f_n(\tau)e^{-\lambda_n a^2(t-\tau)}d\tau\right]\sin\left(\sqrt{\lambda_n}x+\theta_n\right),\end{aligned} \tag{5.48}$$

где φ_n и $f_n(\tau)$ определены формулами (5.45).

Первое слагаемое в выражении (5.48) представляет собой решение краевой задачи для однородного уравнения (5.40), когда $f(x,t)=0$.

5.4. Контрольные вопросы

1. Записать задачу Коши для уравнения теплопроводности в области $G = \{(x,t) \mid x \in (0,l), t \geq 0\}$.

2. Записать задачу Коши для уравнения колебаний в области $G = \{(x,t) \mid x \in (0,l), t \geq 0\}$.

3. Записать первую краевую задачу для уравнения колебаний в области $G = \{(x,t) \mid x \in (0,l), t \geq 0\}$.

4. Записать смешанную краевую задачу для уравнения колебаний в области $G = \{(x,t) \mid x \in (0,l), t \geq 0\}$.

5. Записать вторую краевую задачу для уравнения теплопроводности в области $G = \{(x,t) \mid x \in (0,l), t \geq 0\}$.

6. Записать краевую задачу для уравнения Пуассона в области $G = \{(x,y) \mid x \in (a,b), y \in (c,d)\}$.

7. Привести уравнение к каноническому виду в каждой из областей, где его тип сохраняется:

$$u_{xx} + yu_{yy} + \frac{1}{2}u_y = 0.$$

8. Решить задачу Штурма-Лиувилля

$$X''(x) + \lambda X(x) = 0;$$
$$X'(0) + \alpha X(0) = a, X(l) = b.$$

9. Решить задачу Штурма-Лиувилля

$$X'(x) + \lambda X(x) = 0;$$
$$X(0) = a, X'(l) = b.$$

5.5. Компьютерный практикум

Пример 5.1.

Решим методом Фурье краевую задачу для неоднородного уравнения гиперболического типа

$$u_{t^2} - u_{x^2} = 2t, \quad 0 < x < 1, \quad t > 0, \tag{5.49}$$

$$u\big|_{t=0} = 0; \quad u_t\big|_{t=0} = x, \tag{5.50}$$

$$u\big|_{x=0} = 0; \quad u_x\big|_{x=1} = t. \tag{5.51}$$

Подберем вначале такую функцию w, чтобы она удовлетворяла граничным условиям (5.51) и однородному УЧП (5.49). Пусть, например, $w = xt$, тогда

$$w_{t^2} - w_{x^2} = 0,$$
$$w\big|_{t=0} = 0; \quad w_t\big|_{t=0} = x.$$

Тогда функция

$$v(x,t) = u(x,t) - xt \tag{5.52}$$

удовлетворяет уравнению

$$v_{t^2} - v_{x^2} = 2t, \tag{5.53}$$

однородным граничным условиям

$$v\big|_{x=0} = 0; \quad v_x\big|_{x=1} = 0 \tag{5.54}$$

и нулевым начальным условиям

$$v\big|_{t=0} = 0; \quad v_t\big|_{t=0} = 0. \tag{5.55}$$

Применяя общую схему метода Фурье для решения однородного уравнения $v_{t^2} - v_{x^2} = 0$ при условиях (5.54), (5.55) полагаем $v(x,t) = X(x)T(t)$. Приходим к задаче Штурма-Лиувилля

$$X''(x) + \lambda^2 X(x) = 0;$$
$$X(0) = 0;\ X'(1) = 0.$$

Решая ее, находим собственные значения $\lambda_n = \frac{\pi}{2} + \pi n, \quad n = 0, 1, 2, \ldots$ и соответствующие собственные функции

$$X_n(x) = \sin \lambda_n x. \tag{5.56}$$

Решение задачи (5.53) – (5.55) ищем в виде ряда

$$v(x,t) = \sum_{n=0}^{\infty} T_n(t) \sin \lambda_n x, \tag{5.57}$$

где
$$T_n(0) = 0, \quad T_n'(0) = 0. \tag{5.58}$$

Подставляя (5.57) в (5.53) имеем

$$\sum_{n=0}^{\infty} \left(T_n''(t) + \lambda_n^2 T_n(t)\right) \sin \lambda_n x = 2t. \tag{5.59}$$

Для нахождения функций $T_n(t)$ разложим функцию 5.1. в ряд Фурье по системе функций (5.56) на интервале (0, 1)

$$1 = \sum_{n=0}^{\infty} a_n \sin \lambda_n x. \tag{5.60}$$

Так как $\int_0^1 \sin^2 \lambda_n x dx = \frac{1}{2}$, то $a_n = 2\int_0^1 \sin \lambda_n x dx = \frac{2}{\lambda_n}$, и из (5.59) и (5.60) получаем

$$T_n''(t) + \lambda_n^2 T_n(t) = \frac{4t}{\lambda n}. \qquad (5.61)$$

Общее решение уравнения (5.61) будет $T_n(t) = \frac{4t}{\lambda_n^3} + A\sin \lambda_n t + B\cos \lambda_n t$. Используя условия (5.59), получим $B = 0;\ A = -\frac{4}{\lambda_n^4}$. Подставляя $T_n(t) = \frac{4t}{\lambda_n^3} - \frac{4}{\lambda_n^4}\sin \lambda_n t$ в (5.57) с учетом (5.52) находим решение исходной задачи (5.49)—(5.51):

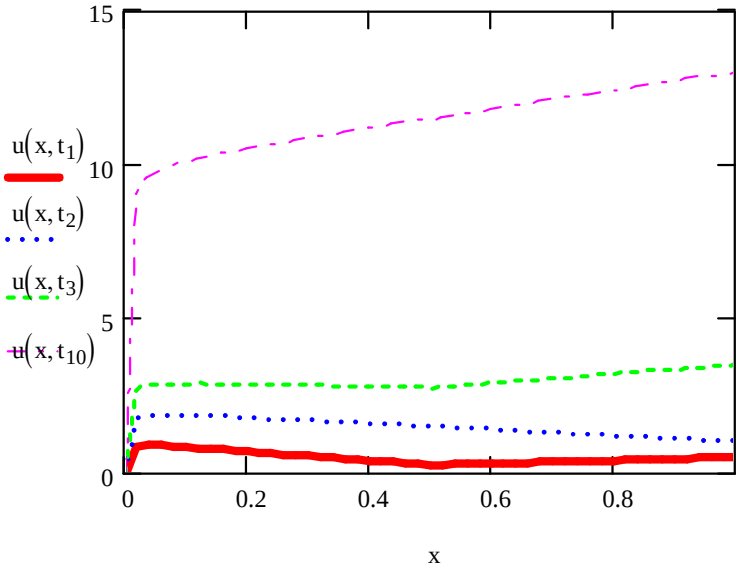
$$u(x,t) = xt + 4\sum_{n=0}^{\infty} \frac{1}{\lambda_n^4}(\lambda_n t - \sin \lambda_n t)\sin \lambda_n x, \text{ где } \lambda_n = \frac{\pi}{2} + n\pi.$$

Строим график решения при различных x и t в Mathcad.

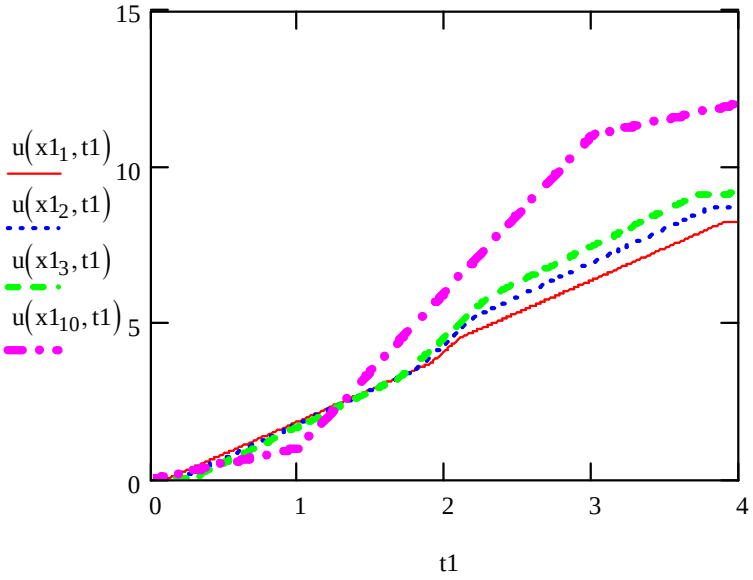
$m := 100$ $\quad n := 0..m$ $\quad lam_n := \frac{\pi}{2} + \pi \cdot n$

$$u(x,t) := x \cdot t + 4 \cdot \sum_{k=0}^{m} \left[\frac{1}{(lam_k)^2} \left(lam_k \cdot t - \sin\left(lam_k \cdot t\right) \right) \cdot \sin\left(lam_k \cdot x\right) \right]$$

$i := 0..10$ $\quad dt := 0.5$ $\quad t_i := i \cdot dt$ $\quad x := 0, 0.02..1$



$t1 := 0, 0.01 .. 4$ $\quad j := 0 .. 10$ $\quad dx := 0.1$ $\quad x1_j := j \cdot dx$



Пример 5.2.

Решим методом Фурье краевую задачу для неоднородного уравнения параболического типа.

$$u_t - u_{x^2} = t(x+1), \quad 0 < x < 1, t > 0, \tag{5.62}$$

$$u\big|_{t=0} = 0, \tag{5.63}$$

$$u_x\Big|_{x=0} = t^2; \quad u\Big|_{x=1} = t^2 . \tag{5.64}$$

Подберем сначала такую функцию w, чтобы она удовлетворяла граничным условиям (5.64) и начальному условию (5.63). Пусть, например, $w = xt^2$, тогда

$$w_t - w_{x^2} = 2xt,$$

$$w\big|_{t=0} = 0,$$

$$w_x\Big|_{x=0} = t^2; \quad w\Big|_{x=1} = t^2 .$$

Поэтому функция

$$v = u - xt^2 \tag{5.65}$$

удовлетворяет уравнению

$$v_t - v_{x^2} = (1-x)\cdot t \tag{5.66}$$

и условиям

$$v\big|_{t=0} = 0; \quad v_x\big|_{x=0} = 0; \quad v\big|_{x=1} = 0 . \tag{5.67}$$

Применяя метод Фурье для решения однородного уравнения $v_t - v_{x^2} = 0$ при условиях (5.67), полагаем $v(x,t) = X(x)T(t)$. Приходим к задаче Штурма-Лиувилля

$$X''(x) + \lambda^2 X(x) = 0;$$
$$X'(0) = 0;\ X(1) = 0,$$

собственными значениями которой являются $\lambda_n = \frac{\pi}{2} + n\pi, \quad n = 0, 1, 2, \ldots,$ а собственными функциями

$$X_n(x) = \cos \lambda_n x. \tag{5.68}$$

Решение задач (5.66), (5.67) ищем в виде

$$v(x,t) = \sum_{n=0}^{\infty} T_n(t) \cos \lambda_n x. \tag{5.69}$$

Подставляя (5.69) в (5.66), получаем

$$\sum_{n=0}^{\infty} \left(T_n'(t) + \lambda_n^2 T_n(t)\right) \cos \lambda_n x = (1 - x)t. \tag{5.70}$$

Разложим функцию $1 - x$ в ряд Фурье по системе функций (5.68) на интервале (0, 1):

$$1 - x = \sum_{n=0}^{\infty} a_n \cos \lambda_n x. \tag{5.71}$$

Так как $a_n = 2\int_0^1 (1 - x) \cos \lambda_n x dx = \frac{2}{\lambda_n^2}$, то из (5.70) и (5.71) находим

$$T_n'(t) + \lambda_n^2 T(t) = \frac{2t}{\lambda_n^2} \tag{5.72}$$

при условии

$$T_n(0) = 0. \tag{5.73}$$

Решением задач Коши (5.51), (5.52) является

$$T_n(t) = 2\lambda_n^{-6} \left(e^{-\lambda_n^2 t} + \lambda_n^2 t - 1\right). \tag{5.74}$$

Из (5.65), (5.69), (5.73) находим решение исходной задачи (5.62)—(5.64):

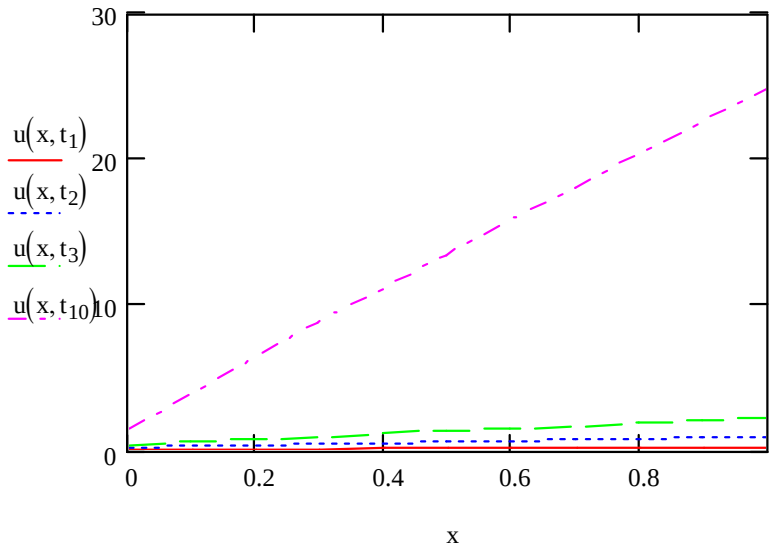
$$u(x,t)=xt^2+2\sum_{n=0}^{\infty}\lambda_n^{-6}\left(e^{-\lambda_n^2 t}+\lambda_n^2 t-1\right)\cos\lambda_n x\text{, где }\lambda_n=\frac{\pi}{2}+n\pi.$$

Строим график решения при различных t в Mathcad.

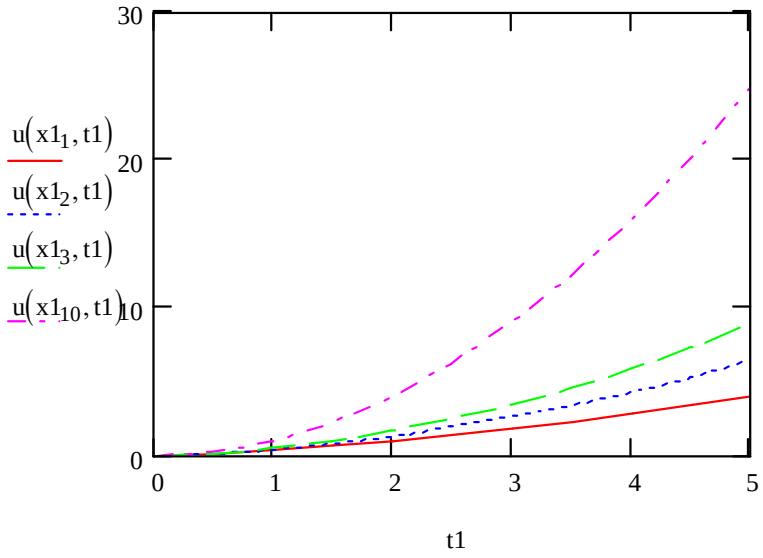
$$m := 100 \qquad n := 0..m \qquad lam_n := \frac{\pi}{2} + n\cdot\pi$$

$$u(x,t) := x\cdot t^2 + 2\left[\sum_{k=0}^{m}\left[\left(lam_k\right)^{-6}\cdot\left[\exp\left[-\left(lam_k\right)^2\cdot t\right]+\left(lam_k\right)^2\cdot t-1\right]\cdot\cos\left(lam_k\cdot x\right)\right]\right]$$

$$i := 0..10 \qquad dt := 0.5 \qquad t_i := i\cdot dt \qquad x := 0, 0.1..1$$



$$j := 0..10 \qquad dx := 0.1 \qquad x1_j := j\cdot dx \qquad t1 := 0, 0.5..5$$



5.6. Задания для самостоятельной работы

Решить методом Фурье краевую задачу для следующих УЧП. Построить график решения при различных t, а также при различных x в Mathcad.

Оценить погрешность по следующей формуле:

$$\Delta = \sqrt{\sum_{i=1}^{n} \left(S_n(t_i) - S_m(t_i)\right)^2} \ .$$

Перечень вариантов к заданию 5.1.

1.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty$$

$$u(x,0) = \sin(\pi x), \quad 0 \le x \le \infty$$

Полагая $x = \Delta x = 0{,}1$ найдите решение при $t_1 = 0{,}005,\ t_2 = 0{,}010,\ t_3 = 0{,}015$. Постройте график полученного решения на сетке при $x = 0; 01; 02 \ldots 0{,}9; 1$ при $t = 0{,}015$.

2.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u\ , \quad 0 < x < 1$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \quad 0 < t < \infty$$

$$u(x,0) = 1, \quad 0 \le x \le 1.$$

3.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \quad 0 < t < \infty$$

$$u(x,0) = 1, \quad 0 \le x \le 1.$$

4.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \sin(\pi x), \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \quad 0 < t < \infty$$

$$u(x,0) = \sin(2\pi x), \quad 0 \le x \le 1.$$

5.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1,$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty$$

$$u(x,0) = \sin(2\pi x) + \frac{1}{3}\sin(4\pi x) + \frac{1}{5}\sin(6\pi x), \quad 0 \le x \le 1.$$

6.

$$\frac{\partial u}{\partial t} = 5\frac{\partial^2 u}{\partial x^2} + \sin(8\pi x), \quad 0 < x < 1,$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty$$

$$u(x,0) = \cos(4\pi x), \quad 0 \le x \le 1$$

7.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - u + x, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty$$

$$u(x,0) = 0, \quad 0 \le x \le 1.$$

8.

$$\frac{\partial u}{\partial t} = 4\frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ \dfrac{\partial u}{\partial x}(1,t) = x^2, \end{cases} \qquad 0 < t < \infty$$

$u(x,0) = \sin(\pi x), \quad 0 \le x \le 1.$

9.

$$\frac{\partial^2 u}{\partial t^2} = 3\frac{\partial^2 u}{\partial x^2} + 28x, \quad 0 < x < 1$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \quad 0 < t < \infty$$

$$\begin{cases} u(x,0) = \sin(x) \\ \dfrac{\partial u(x,0)}{\partial t} = 0 \end{cases}, \quad 0 \le x \le 1$$

10.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x}, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \quad 0 < t < \infty$$

$u(x,0) = e^{x/2}, \quad 0 \le x \le 1.$

11.

$\Delta u = 0, \quad 0 < r < 2,$

$u(2, \theta) = \sin(\theta) \qquad 0 < \theta < 2\pi.$

12.

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = 0, \quad 0 < r < 1,$$

$$u(1,\theta) = 1 + \sin(\theta) + \frac{1}{2}\cos(\theta).$$

13.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = \cos(t), \end{cases} \qquad 0 < t < \infty$$

$$u(x,0) = 0, \quad 0 \le x \le \infty$$

14.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$u(0,t) = \sin(t), \qquad 0 < t < \infty$$

$$u(x,0) = 0, \quad 0 \le x \le \infty$$

15.

$$\frac{\partial^2 u}{\partial t^2} = 5\frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 8$$

$$\begin{cases} u(0,t) = 0, \\ u(8,t) = 0, \end{cases} \qquad 0 < t < \infty$$

$$\begin{cases} u(x,0) = \sin(\frac{\pi x}{8}) + \frac{1}{2}\sin(\frac{3\pi x}{8}) \\ \frac{\partial u(x,0)}{\partial t} = 0. \end{cases}, \quad 0 \le x \le 8$$

VI. ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ УРАВНЕНИЙ В ЧАСТНЫХ ПРОИЗВОДНЫХ НА КООРДИНАТНЫХ СЕТКАХ

6.1. Введение в разностные методы

При решении краевых задач, возникающих на практике аналитические методы решения трудно применимы. Одним из приближенных численных методов решения дифференциальных уравнений в частных производных является ***метод сеток***. Идея метода заключается в следующем. Для простоты, ограничимся случаем только функции двух переменных, и будем полагать, что решение уравнения ищется на квадратной области единичного размера (этого можно добиться, нормируя и обезразмеривая переменные задачи). Разобьем область сеткой. Шаг сетки по оси x и по оси y, вообще говоря, может быть разный. По определению частная производная равна

$$\frac{\partial u(x,y)}{\partial x} \equiv u_x(x_i, y_j) \equiv \lim_{\Delta x \to 0} \frac{u(x+\Delta x, y) - u(x)}{\Delta x} \approx \frac{u(x+\Delta x, y) - u(x)}{\Delta x}. \quad (6.1)$$

Если рассматривать функцию только в узлах сетки, то частную производную можно записать (аппроксимировать) в форме

$$u_x(x_i, y_j) \approx \frac{u_{i+1,j} - u_{i,j}}{h} \equiv \partial^+_{x^2} u_{i,j}\ , \quad (6.2)$$

где узел (i, j) соответствует точке (x_i, y_j), $x_i = h_x i$, $y_j = h_y j$.

Полученное выражение называется ***правой конечной разностью*** и имеет специальное обозначение. Название связано с тем, что для вычисления производной в точке используются значение функции в этой точке и точке, лежащей правее. Очевидно, что сходное выражение можно было бы получить, используя точку, лежащую слева.

$$u_x(x_i, y_j) \approx \frac{u_{i,j} - u_{i-1,j}}{h} \equiv \partial_x^- u_{i,j} \,. \tag{6.3}$$

Такое выражение называется ***левой конечной разностью***. Можно записать центральную конечную разность, найдя среднее арифметическое этих выражений. Введенные конечные разности называются разностными производными первого порядка.

Теперь получим выражения для вторых производных:

$$u_{xx}(x_i, y_j) \approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} \equiv \partial_{xx}^{+2} u_{i,j} \,. \tag{6.4}$$

В данном случае для нахождения производной мы использовали симметричные точки. Однако, очевидно, можно было бы использовать точки с несимметричным расположением.

6.2. Разностные уравнения, явная и неявная схемы

1. В качестве начального примера рассмотрим решение краевой задачи для волнового уравнения (уравнения гиперболического типа).

$$U_{t^2} = a^2 U_{x^2} + f(x,t), \tag{6.5}$$

$$U_{t=0} = \varphi(x) \qquad U_t\big|_{t=0} = \psi(x), \tag{6.6}$$

$$U\big|_{x=0} = \theta(t), \, U\big|_{x=l} = \vartheta(t) \,. \tag{6.7}$$

Уравнение колебаний математической струны с заданными начальными смещениями и скоростями всех точек и произвольными законами на концах.

Составим разностную схему, соответствующую задаче (6.5) — (6.7). Для

этого введем сеточную область. Выберем на временном интервале узловые точки

$$t_m = \tau m,\ m = 0,1,2 \ldots M\ , \tag{6.8}$$

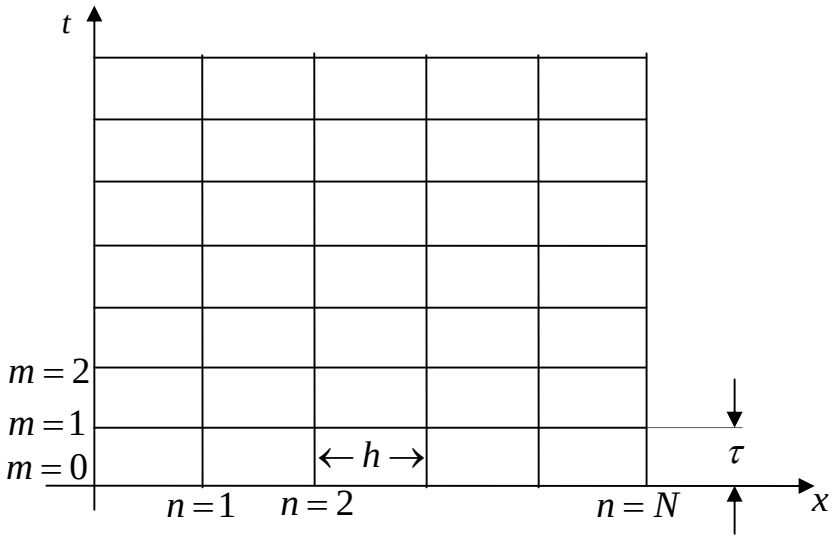


Рис. 6.1. Явная сеточная схема для уравнения колебаний

где $\tau = const$ - шаг по времени. Аналогично выберем координатные узлы $x_n = hn$, где h - шаг по длине, $n=0..N$. Тогда на плоскости $x0t$ получится прямоугольная сетка (рис. 6.1.). Прямую $t_m = const$ будем называть временным слоем m или просто m-слоем. Значения разностной функции в узлах сетки обозначим

$$u\left(x_m, t_m\right) = u_{m,n}, \tag{6.9}$$

т.е. первый нижний индекс относится к пространственной координате, а второй к временному слою. Аппроксимируя производные U_{xx} разностной производной $\partial_{xx}^{+2} u$, U_{tt} разностной производной $\partial_{tt}^{+2} u$, получим уравнение в конечных разностях

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + f_{i,j} = \frac{1}{a^2}\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\tau^2} \tag{6.10}$$

на следующей разностной явной схеме (см. рис. 6.2.)

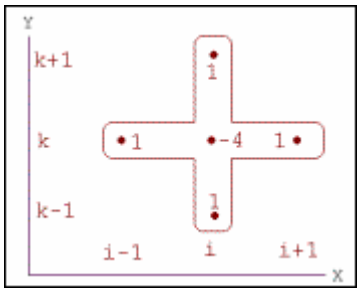


Рис.6.2. Элемент явной сеточной схемы

Полученное уравнение позволяет выразить значение функции u в момент времени $j+1$ через значения функции в предыдущие моменты времени.

$$u_{i,j+1} = a^2\left(\frac{\tau}{h}\right)^2\left(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\right) + 2u_{i,j} - u_{i,j-1} + \tau^2 f_{i,j}. \qquad (6.11)$$

Зададим начальные и граничные условия (6.6), (6.7) на сетке

$$u_{i,0} = \varphi_i, \qquad u_{i,1} - u_{i,0} = \tau\psi_i, \quad u_{0,j} = \theta_j, \quad u_{M,j} = \vartheta_j. \qquad (6.12)$$

Такая разностная схема называется ***явной***, так как искомая величина получается в явном виде. Последние уравнения представляют собой систему линейных алгебраических уравнений, решение которой может быть найдено методом простой итерации.

Примечание. Заметим, что о том насколько хорошо разностный оператор L_h аппроксимирует континуальный оператор L можно судить по разности $\left\|L_h U_h - r_h(LU)\right\|_{H_h}$, где L_h - разностный оператор, аппроксимирующий континуальный оператор L, задающий уравнение, $U(x)$ - точное решение, U_h - проекция решения на сеточное пространство. Говорят, что разностное уравнение аппроксимирует уравнение с порядком m или имеет место аппроксимация порядка m, если $\left\|L_h U_h - r_h(LU)\right\|_{H_h} = O\left(h^m\right)$. Таким образом, для того, чтобы имела место сходимость, т.е. решение разностного уравнения U_h стремилось бы к

точному решению U уравнения при стремлении $h \to 0$, необходимо, чтобы при $h \to 0$ стремилась к нулю разность $\delta u_h \equiv u_h - U_h$ где u_h является решением разносного уравнения соответствующего данному т.е. $\|\delta_h\|_{H_h} \underset{h\to 0}{\longrightarrow} 0$. Однако для сходимости (корректности) одной аппроксимации разностным уравнением оказывается недостаточно. Разностная схема является устойчивой, если малое изменение правой части влечет за собой малое изменение решения. Другими словами, разностная схема называется устойчивой, если $\|\delta f_h\|_{H_h} \to 0 \Rightarrow \|\delta u_h\| \to 0$. Условие же аппроксимации тогда можно переписать в виде $h \to 0 \Rightarrow \|\delta f_h\| \to 0$.

Итак, из аппроксимации и устойчивости следует сходимость. Это утверждение называется предельной теоремой Лакса. Условие устойчивости явных разностных схем (т. н. условно устойчивых) обеспечивается согласованием шагов сетки по осям, когда шаг по одной из осей выбирается в зависимости от величины шага по другой. В тоже время, условия согласования могут быть весьма обременительными из-за необходимости выбора слишком малого шага по одной из осей для получения устойчивого приближенного решения. Аппаратным методом решения этих проблем является применение других типов разностных схем (неявных), гарантирующих т. н. безусловную устойчивость. Для реализации неявной схемы часто используется специальный метод решения СЛАУ – метод прогонки. В последнее время актуализированными методами становятся универсальные многосеточные методы приближенного решения УЧП на структурированных сетках, в которых производится адаптация решаемых задач к численному методу, применяются несколько сеток на одном сеточном уровне для вычисления поправки, оригинальное построение грубых сеток, более эффективные типы дискретизаций исходной области.

Для устойчивости задачи (7.11 — 7.12) имеем условие согласования $\tau \le h / a$.

2. Еще один пример использования конечных разностей – уравнение теплопроводности (диффузии) (уравнение параболического типа).

На левом и правом торце стержня происходит тепловое взаимодействие с окружающей средой, и здесь необходимо задать граничные условия. Кроме того задана функция начальной температуры по всей длине стрежня.

Рассмотрим применение неявной схемы на примере уравнения теплопроводности вида

$$U_t = a^2 U_{x^2} + f(x,t); \qquad (6.13)$$

$$U|_{t=0} = \varphi(x_i); \tag{6.14}$$

$$\left(U_x + \alpha(t)U\right)\Big|_{x=0} = g_1(t), \left(U_x + \beta(t)U\right)\Big|_{x=l} = g_2(t). \tag{6.15}$$

Составим неявную разностную схему для этого уравнения

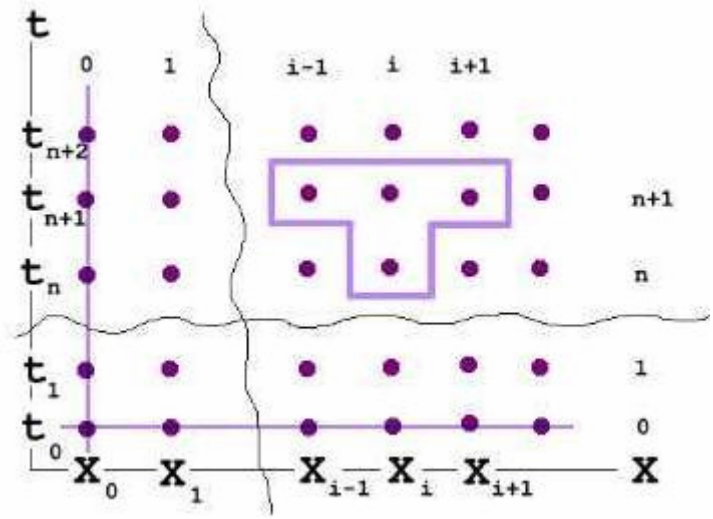


Рис.6.3. Неявная разностная схема для уравнения

Заменяем на сетке исходное дифференциальное уравнение его разностной аппроксимацией в соответствии с шаблоном (см. рис.6.3.).

$$\frac{u_{i,j+1} - u_{i,j}}{\tau} = \frac{a^2}{h^2}\left(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}\right). \tag{6.16}$$

Здесь первый индекс соответствует пространственной, а второй – временной координате. В отличие от явной схемы, для вычисления в правой части уравнения используются значения функции на том же самом временном шаге. Вводя обозначение $\mu = \frac{a^2\tau}{h^2}$, уравнение (6.16) можно переписать в виде

$$(1 + 2\mu)u_{i,j+1} - \mu\left(u_{i+1,j+1} + u_{i-1,j+1}\right) = u_{i,j}.$$

или в матричной форме

$$\begin{pmatrix} 1+2\mu & -\mu & & & \\ -\mu & 1+2\mu & -\mu & & \\ & \cdots & & & \\ & & 1+2\mu & -\mu & \\ & & -\mu & 1+2\mu & \end{pmatrix}\begin{pmatrix} u_{1,j+1} \\ u_{2,j+1} \\ \vdots \\ u_{n-1,j+1} \\ u_{n,j+1} \end{pmatrix}=\begin{pmatrix} u_{1,j}+\mu\alpha \\ u_{2,j} \\ \vdots \\ u_{n-1,j} \\ u_{n,j}+\mu\beta \end{pmatrix}.$$

Начальные и граничные условия (6.14), (6.15) на сетке имеют вид

$$u_{i,0}=\varphi_i \quad u_{0,j}=g1_j \quad u_{M,j}=g2_j .$$

Применять для таких разреженных матриц типовые алгоритмы (в частности, алгоритм Гаусса), например, встроенную функцию Lsolve Mathcad, расточительно. Для таких задач (а к ним приводит огромное количество неявных разностных схем для различных дифференциальных уравнений) применяется алгоритм *прогонки*.

3. Наконец, рассмотрим применение разностных методов к стационарным уравнениям эллиптического типа. Согласно идеям метода сеток уравнение Пуассона $U_{x^2}+U_{y^2}=F(x,y)$ может быть записано в разностной форме при помощи шаблона "крест" (рис.6.4.),

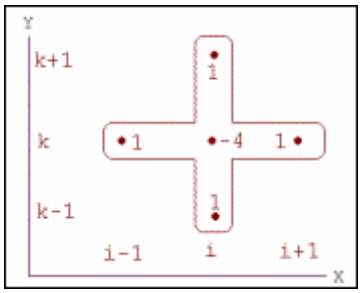


Рис. 6.4. Элемент неявной разностной схемы

которая после приведения подобных слагаемых в разностных уравнениях имеет вид

$$a_{i,j}u_{i+1,j}+b_{i,j}u_{i-1,j}+c_{i,j}u_{i,j+1}+d_{i,j}u_{i,j-1}+e_{i,j}u_{i,j}=f_{i,j}. \qquad (6.17)$$

Разностная схема для уравнения Пуассона уже не может быть решена явно.

Требуется решить систему линейных уравнений (6.17) т. н. методом релаксации. Для его запуска требуется задать начальное приближение к решению (нулевое приближение) и запустить итерационный процесс, уточняющий решение. Параметр численного алгоритма (число в пределах от 0 до 1) характеризует скорость сходимости итераций. Суть алгоритма релаксации сводится к тому, что в ходе итераций происходит проверка уравнений и соответствующая коррекция значений искомой функции в каждой точке. Если начальное приближение выбрано удачно, то можно надеяться, что алгоритм сойдется («срелаксирует») к правильному решению. Для решения стационарных уравнений Пуассона и Лапласа в Mathcad предназначена функция **relax*(a, b, c, d, e, f, u, rjac)***, реализующая метод релаксации. Поясним использование этой функции на простом примере. Фактически, эту функцию можно использовать для решения эллиптического уравнения общего вида

$$\Delta_{x_i} u = f(x_i) \tag{6.18}$$

с граничными условиями

$$\left(u_\nu + \alpha(x_i', t)u\right)\Big|_\Gamma = g(t, x_i'), \tag{6.19}$$

которое может быть сведено к уравнению в конечных разностях (6.17).

В частности, для уравнения Пуассона (6.18) коэффициенты $a_{i,j} = b_{i,j} = c_{i,j} = d_{i,j} = 1, e_{i,j} = -4$.

Метод релаксации заключается в следующем. Если нет источников (уравнение Лапласа), то значение функции в данном узле на текущем шаге $k+1$ (верхний индекс) определяется как среднее значение функции в ближайших узлах на предыдущем шаге *k* (шаблон разностной схемы имеет вид «крест»):

$$u_{i,j}^{k+1} = \frac{1}{4}\left(u_{i-1,j}^{k} + u_{i+1,j}^{k} + u_{i,j-1}^{k} + u_{i,j+1}^{k}\right). \tag{6.20}$$

При наличии источников разностная схема будет

$$u_{i,j}^{k+1} = \frac{1}{4}(u_{i-1,j}^{k} + u_{i+1,j}^{k} + u_{i,j-1}^{k} + u_{i,j+1}^{k}) - \frac{h^2}{4} f_{i,j}. \qquad (6.21)$$

Метод релаксации сходится достаточно медленно, так как фактически он использует разностную схему с максимально возможным для двумерного случая шагом $\tau = \frac{h^2}{4}$.

В методе релаксации необходимо задать начальное приближение, то есть значения функции во всех узлах области, а так же граничные условия.

6.3. Контрольные вопросы

1. Каким образом вычисляются значения приближенного решения уравнения колебаний на первом временном слое?

2. Запишите разностную схему для уравнения колебаний при замене первой производной в начальном условии разложением по формуле Тейлора по *t* с удержанием членов до 2-го порядка.

3. Используя Теорему Лакса, оцените количество временных слоев, возникающих в методе сеток для уравнения колебаний при h=0.01 и a=10^3, если необходимо найти решение в момент T=35с.

4. Какая разностная схема возникает (явная или неявная) для уравнения колебаний, если использовать следующие разностные аппроксимации, входящих в уравнение производных

$$U_{xx}(x_i, y_j) \approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}; \quad U_{tt}(x_i, y_j) \approx \frac{u_{i,j} - 2u_{i,j-1} + u_{i,j-2}}{\tau^2}$$

и уже применяемые начальные и граничные условия.

5. Запишите явную разностную схему для уравнения теплопроводности с уже применяемыми начальными и граничными условиями.

6. Постройте разностную схему для уравнения для уравнения Пуассона в прямоугольнике в случае задания нормальной производной на одной из граничных прямых.

7. Запишите разностную схему для уравнения Пуассона в круговой области, осуществив предварительно ее отображение на прямоугольник, переходя к полярным координатам. Граничные условия задаются на искомую функцию.

6.4. Компьютерный практикум

а) Уравнение колебаний

Указанные разностные схемы реализованы в специальном модуле пакета Mathcad для решения уравнений в частных производных PDE *Solver*.

Рассмотрим конкретный пример в Mathcad. Краевая задача для уравнения колебаний. Решение выполним, используя стандартные средства пакета. Для определенности выберем уже решенную ранее краевую задачу

$$u_{t^2} - u_{x^2} = 2t,\ 0 < x < 1,\ t > 0.$$

$$u\big|_{t=0} = 0;\ u_t\big|_{t=0} = x;$$

$$u\big|_{x=0} = 0;\ u_x\big|_{x=1} = t.$$

The Wave Equation

Given

$v_t(x,t) = a^2 w_{xx}(x,t)$ $\qquad$ $w_t(x,t) = v(x,t)$

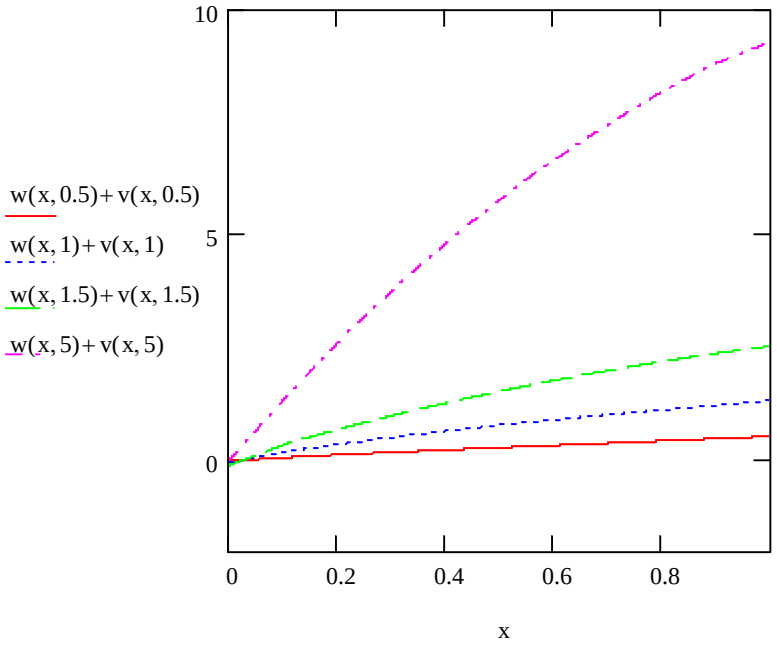
$w(x,0) = 0$ $\qquad$ $v(x,0) = 0$

$w(0,t) := \dfrac{-(t \cdot t \cdot t)}{3}$ $\qquad$ $w_x(L,t) := 0$

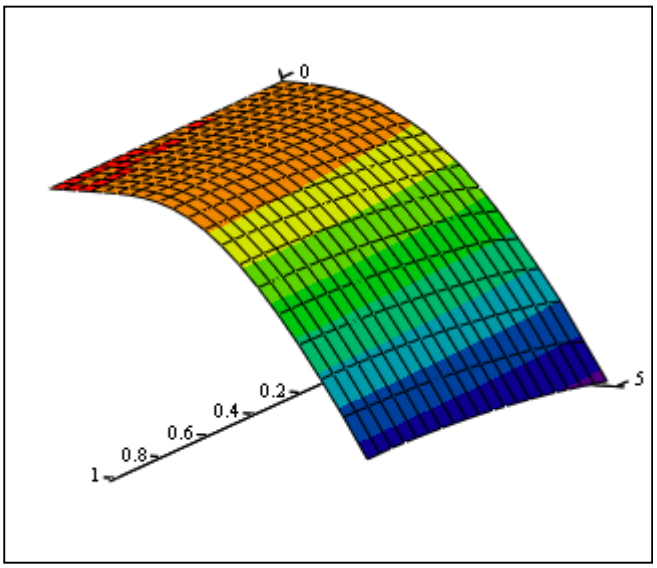
$a := 1 \quad L := 1 \quad T := 5$

$$\begin{pmatrix} w \\ v \end{pmatrix} := Pdesolve\left[\begin{pmatrix} w \\ v \end{pmatrix}, x, \begin{pmatrix} 0 \\ L \end{pmatrix}, t, \begin{pmatrix} 0 \\ T \end{pmatrix}\right]$$

$$v(x,t) := \frac{t \cdot t \cdot t}{3} + x \cdot t$$



$$M := CreateMesh(w, 0, L, 0, T)$$



M

b) Краевая задача для уравнения теплопроводности

Решение выполним, используя стандартные средства пакета. Для определенности выберем уже решенную ранее краевую задачу:

$$u_t - u_{x^2} = t\left(x+1\right),\ 0 < x < 1,\ t > 0.$$

$$u\big|_{t=0} = 0,$$

$$u_x\Big|_{x=0} = t^2;\ u\Big|_{x=1} = t^2.$$

The Heat Equation

$f(x,t) := t\cdot(x+1)$ $a := 1$ $T := 5$ $L := 1$

$spacepts := 10$ $timepts := 10$

Given

$u_t(x,t) = a^2 \cdot u_{xx}(x,t) + f(x,t)$

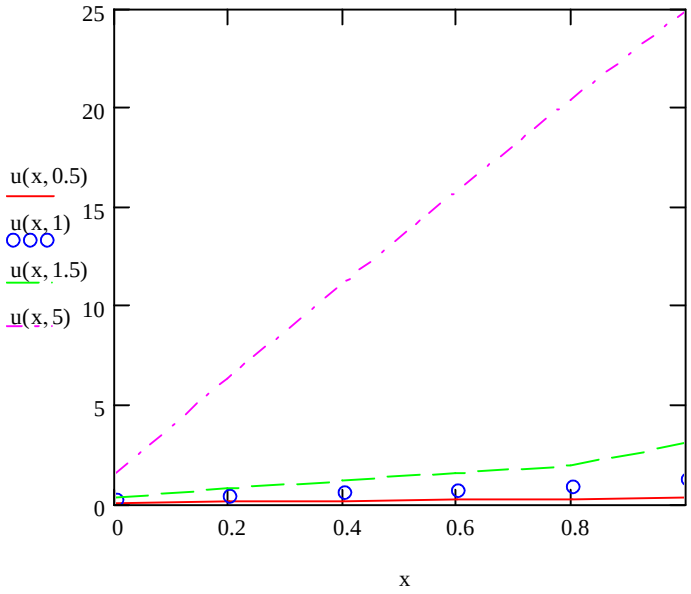
$u(x,0) = 0$

$u_x(0,t) = t\cdot t$

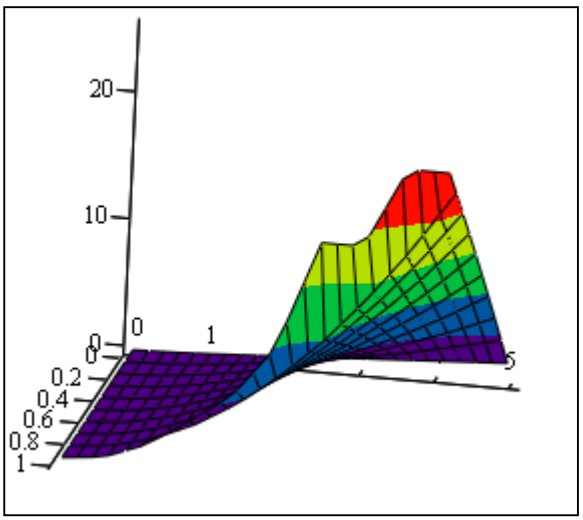
$u(L,t) = t\cdot t$

$$u := Pdesolve\left[u, x, \begin{pmatrix}0\\ L\end{pmatrix}, t, \begin{pmatrix}0\\ T\end{pmatrix}, spacepts, timepts\right]$$

$u(0.1,0.2) = 0.041$ $x := 0, 0.2..L$ $t := 0..T$



$$A1 := CreateMesh(u, 0, L, 0, T, 10 \cdot L, 5 \cdot T)$$



A1

c) Краевая задача для уравнения Пуассона

Рассмотрим конкретный пример в Mathcad. Решение выполним, используя стандартные средства пакета и представим его графически в виде поверхности и линий уровней.

Предварительно детализируем встроенную функцию пакета.

Функция ***relax*** реализует метод релаксации для приближения к решению.

Она возвращает квадратную матрицу, в которой:

1) расположение элемента в матрице соответствует его положению внутри квадратной области,

2) это значение приближает решение в этой точке.

Функция ***relax*** используется**,** если известны значения искомой функции ***u(x, y)*** на всех четырех сторонах квадратной области.

Ее аргументы:

a, b, c, d, e – квадратные матрицы одного и того же размера, содержащие коэффициенты дифференциального уравнения;

f – квадратная матрица, содержащая значения правой части уравнения в каждой точке внутри квадрата;

u – квадратная матрица, содержащая граничные значения функции на краях области, а также начальное приближение решения во внутренних точках области;

rjac – параметр, управляющий сходимостью процесса релаксации. Он может быть в диапазоне от 0 до 1, но оптимальное значение зависит от деталей задачи.

Решим следующую краевую задачу

$$u_{x^2} + u_{y^2} = 10(\delta(x - \frac{3}{16}, y - 0.25) + \delta(x - 5/16, y - 0.25)) \quad u\big|_{\Gamma} = 0, 0 < x < 1, y = 0$$

$$u\big|_{\Gamma} = -1,\ 0 < x < 1,\ y = 1\ u\big|_{\Gamma} = 1 - 2y,\ x = 0,\ 0 < y < 1\ u\big|_{\Gamma} = 1 - 2y,\ x = 1,\ 0 < y < 1$$

The Poisson's Equation

$n := 2^5$ $\quad i := 0..n$ $\quad j := 0..n$

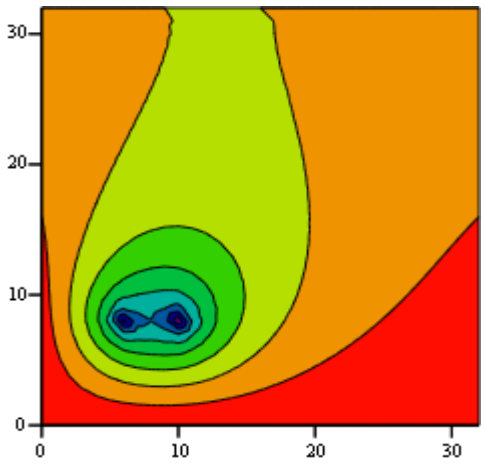
$M_{i,j} := 0$ $\quad M_{6,8} := 10$ $\quad M_{10,8} := 10$

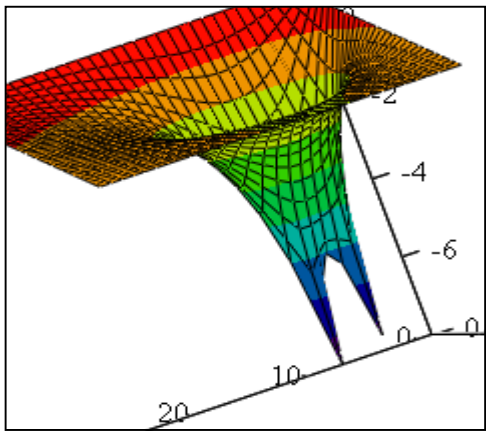
$a_{i,j} := 1$ $\quad b := a$ $\quad c := a$ $\quad d := a$ $\quad f := M$

$e := -4 \cdot a$

$u_{i,j} := 0$ $\quad u_{i,n} := -1$ $\quad u_{0,j} := 1 - 2 \cdot \frac{j}{n}$ $\quad u_{i,0} := 1$ $\quad u_{n,j} := 1 - 2 \cdot \frac{j}{n}$

$$Z := relax(a,b,c,d,e,f,u,0.95)$$





Z

Z Если граничные условия равны нулю на всех четырех сторонах квадрата, можно использовать функцию ***multigrid***.

$$Z := multigrid(M,3)$$

Алгоритм метода, реализуемый этой функцией является многосеточным.

6.5. Задания для самостоятельной работы

Решить методом сеток краевую задачу для следующих УЧП. Построить график решения при различных t, а также при различных x в Mathcad.

Оценить точность решений по следующим эмпирическим формулам оценки глобальной погрешности решения:

$\Delta = \max\limits_{0\le j\le N}\{\sqrt{\frac{1}{(M+1)}\sum\limits_{i=0}^{M}\left(U\left(x_i,t_j\right)-u_{i,j}\right)^2}\}$ - для гиперболических уравнений,

$\Delta = \max\limits_{\substack{0\le i\le M\\ 0\le j\le N}}\left|U(x_i,y_j)-u_{i,j}\right|$ - для параболических и эллиптических уравнений.

Перечень вариантов к заданию 6.1.

1.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty$$

$$u(x,0) = \sin(\pi x), \quad 0 \le x \le \infty .$$

Полагая $x = \Delta x = 0{,}1$ найдите решение при $t_1 = 0{,}005,\ t_2 = 0{,}010,\ t_3 = 0{,}015$. Постройте график полученного решения на сетке при $x = 0; 01; 02 \ldots 0{,}9; 1$ при $t = 0{,}015$.

2.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u \ , \quad 0 < x < 1,$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty ,$$

$$u(x,0) = 1, \quad 0 \le x \le 1 .$$

3.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \ , \quad 0 < x < 1, \quad 0 < t < \infty ,$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty ,$$

$$u(x,0) = 1, \quad 0 \le x \le 1 .$$

4.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \sin(\pi x), \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty,$$

$$u(x,0) = \sin(2\pi x), \quad 0 \le x \le 1.$$

5.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1,$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty,$$

$$u(x,0) = \sin(2\pi x) + \frac{1}{3}\sin(4\pi x) + \frac{1}{5}\sin(6\pi x), \quad 0 \le x \le 1.$$

6.

$$\frac{\partial u}{\partial t} = 5\frac{\partial^2 u}{\partial x^2} + \sin(8\pi x), \quad 0 < x < 1,$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty,$$

$$u(x,0) = \cos(4\pi x), \quad 0 \le x \le 1.$$

7.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - u + x, \quad 0 < x < 1, \quad 0 < t < \infty.$$

$$u(x,0) = 0, \quad 0 \le x \le 1.$$

8.

$$\frac{\partial u}{\partial t} = 4\frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < \infty.$$

$$\begin{cases} u(0,t) = 0, \\ \dfrac{\partial u}{\partial x}(1,t) = x^2, \end{cases} \qquad 0 < t < \infty.$$

$$u(x,0) = \sin(\pi x), \quad 0 \le x \le 1.$$

9.

$$\frac{\partial^2 u}{\partial t^2} = 3\frac{\partial^2 u}{\partial x^2} + 28x, \quad 0 < x < 1$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty$$

$$\begin{cases} u(x,0) = \sin(x) \\ \dfrac{\partial u(x,0)}{\partial t} = 0 \end{cases}, \quad 0 \le x \le 1$$

10.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x}, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = 0, \end{cases} \qquad 0 < t < \infty$$

$$u(x,0) = \mathrm{e}^{x/2}, \quad 0 \le x \le 1.$$

11.

$$\Delta u = 0, \quad 0 < r < 2,$$

$$u(2,\ \theta) = \sin(\ \theta) \qquad 0 < \theta < 2\pi.$$

12.

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = 0, \quad 0 < r < 1,$$

$$u(1,\theta) = 1 + \sin(\ \theta) + \frac{1}{2}\cos(\ \theta).$$

13.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$\begin{cases} u(0,t) = 0, \\ u(1,t) = \cos(\ t), \end{cases} \qquad 0 < t < \infty$$

$$u(x,0) = 0, \quad 0 \le x \le \infty$$

14.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < \infty$$

$$u(0,t) = \sin(t), \qquad 0 < t < \infty$$

$$u(x,0) = 0, \quad 0 \le x \le \infty$$

15.

$$\frac{\partial^2 u}{\partial t^2} = 5\frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 8$$

$$\begin{cases} u(0,t) = 0, \\ u(8,t) = 0, \end{cases} \quad 0 < t < \infty$$

$$\begin{cases} u(x,0) = \sin(\frac{\pi x}{8}) + \frac{1}{2}\sin(\frac{3\pi x}{8}) \\ \frac{\partial u(x,0)}{\partial t} = 0. \end{cases}, \quad 0 \le x \le 8$$

ЛИТЕРАТУРА

1. Демидович, Б.П. Основы вычислительной математики / Б.П. Демидович, И.А. Марон. – М.: Физматгиз, 1963. – 660 с.
2. Шапорев, С.Д. Методы вычислительной математики и их приложения / С.Д. Шапорев. – С.-Петербург: СМИО Пресс, 2003. – 230 с.
3. Ильин, В.А. Математический анализ / В.А. Ильин, В.А. Садовничий, Бл.Х. Сендов. – М.: Физматгиз, 1979. – 720 с.
4. Зверович, Э.И. Вещественный и комплексный анализ. Ч. 1. Введение в анализ и дифференциальное исчисление / Э.И. Зверович. – Минск: Вышэйшая школа, 2006. – 320 с.
5. Бахвалов, Н. Численные методы / Н. Бахвалов, Н. Жидков, Г. Кобельков. – М.: Бином. Лаборатория знаний, 2003. – 632 с.
6. Самарский, А.А. Введение в численные методы / А.А. Самарский. – М.: Лань, 2005. – 288 с.
7. Самарский, А.А. Методы решения сеточных уравнений / А.А. Самарский, Е.С. Николаев. – М.: Физматгиз, 1978. – 592 с.
8. Амосов, А.А. Вычислительные методы для инженеров. Учеб. пособие / А.А. Амосов, Ю.А. Дубинский, Н.В. Копченова. – М.: Высшая школа, 1994. – 544 с.
9. Алберг, Дж. Теория сплайнов и ее приложения / Дж. Алберг, Э. Нильсон, Дж. Уолш. – М.: Мир, 1972. – 316 с.
10. Петровский, И.Г. Лекции по теории обыкновенных дифференциальных уравнений. Т. 2 / И.Г. Петровский. – М.: Физматгиз, 1970. – 280 с.
11. Пискунов, Н.С. Дифференциальное и интегральное исчисления для втузов. Т. 2 / Н.С. Пискунов. – М.: Наука, 1985. – 560 с.
12. Кошляков, Н.С. Уравнения в частных производных математической физики. Учеб. пособие для мех.-мат. фак. ун-тов / Н.С. Кошляков. – М.: Высшая школа, 1970. – 712 с.
13. Михлин, С.Г. Линейные уравнения в частных производных. Учеб. пособие для вузов / С.Г. Михлин. – М.: Высшая школа, 1977. – 431 с.
14. Мартинсон, Л.К. Дифференциальные уравнения математической физики: Учеб. для вузов / Л.К. Мартинсон, Ю.И. Малов; под ред. В.С. Зарубина, А.П. Крищенко. – 2-е изд. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2002. – 368 с.
15. Владимиров, В.С. Уравнения математической физики / В.С. Владимиров. – 4-е изд. – М.: Наука. Главная редакция физико-математической литературы, 1981. – 512 с.
16. Самарский, А.А. Введение в теорию разностных схем / А.А. Самарский. – М.: Наука. Главная редакция физико-математической литературы, 1971. – 554 с.
17. Поттер Д. Вычислительные методы в физике / Д. Поттер. – М.: Высшая школа, 1999. – 394 с.
18. Wesseling, P. Introduction to Multigrid Methods / P. Wesseling. – Vesseling ft University of Technology, The Netherlands, 1991. – 275 p.