УЧРЕЖДЕНИЕ
РОССИЙСКОЙ АКАДЕМИИ НАУК
ИНСТИТУТ ЯДЕРНОЙ ФИЗИКИ
им. Г.И. Будкера СО РАН
СИБИРСКОГО ОТДЕЛЕНИЯ РАН
(ИЯФ СО РАН)

Э.А. Бибердорф, Н.И. Попова

КОНТРОЛЬ ТОЧНОСТИ РЕШЕНИЯ КРАЕВОЙ ЗАДАЧИ МЕТОДОМ ОРТОГОНАЛЬНОЙ ПРОГОНКИ

ИЯФ 2009-1

НОВОСИБИРСК
2009

Контроль точности решения краевой задачи методом ортогональной прогонки

Э.А. Бибердорф

Институт математики им. С.Л. СОболева
630090, Новосибирск, Россия

Н.И. Попова

Институт ядерной физики им. Г.И.Будкера
630090, Новосибирск, Россия

Аннотация

Цель нашей работы заключается в развитии и популяризации современных алгоритмов нового поколения, позволяющих проводить вычисления с гарантированной оценкой точности. В данной работе мы впервые переходим от задач линейной алгебры к гораздо более сложной проблеме – оценке точности выполнения алгоритма решения задачи математической физики. В качестве объекта исследования рассмотрен алгоритм ортогональной прогонки решения краевой задачи для системы обыкновенных дифференциальных уравнений, который включает в себя такие элементы, как дискретизация, интерполяция, численное интегрирование и решение серий задач линейной алгебры. Здесь мы используем полученные нами ранее гарантированные оценки точности ортогонализации и решения линейных алгебраических систем [3]. В результате впервые практически получена гарантированная оценка погрешности (как алгоритмической, так и погрешности машинных округлений) решения краевой задачи. Создана новая версия **GALA-2.1** пакета программ **GALA-2.0** (Guaranteed Accuracy in Linear Algebra), описанного в монографии [3]. В новую версию входят программы реализующие такие вычислительные методы, как "ортогональная прогонка", "регуляризация", QR-разложение матриц с контролем оценки точности результата. Приведены примеры.

©*Институт ядерной физики им. Г.И.Буджера СО РАН*

Accuracy control of solving the boundary-value problem by the orthogonal sweep method

E.A. Biberdorf

Sobolev Institute of Mathematics
630090, Novosibirsk, Russia

N.I. Popova

Budker Institute of Nuclear Physics
630090, Novosibirsk, Russia

Abstract

The aim of our paper is to develop and popularize the modern algorithms permitting one to carry out computations with the guaranteed accuracy estimation. In this paper we first pass from the linear algebra problem to the more complicated problem – the accuracy estimation of fulfilling the algorithm of solving the mathematical physics problem. The subject of our study is the orthogonal sweep algorithm for solving the boundary-value problem of the system of ordinary differential equations. The algorithm contains such elements as discretization, interpolation, numerical integration and solution of series of linear algebra problems. Here we used the guaranteed accuracy estimations of the orthogonalization and the solution of linear algebra systems obtained in [3]. As a result we first practically obtained the guaranteed estimation of error (both the algorithmic and round-off errors) for solving the boundary-value problem. The new version **GALA-2.1** of the program package **GALA-2.0** (Guaranteed Accuracy in Linear Algebra) described in the monograph [3] is created. The programs realizing such numerical methods as the orthogonal sweep method, regularization, QR-decomposition of matrixes with the guaranteed estimation of the result accuracy enter into new version. The numerical examples are given.

1 Введение

Данная работа продолжает начатую в 1999 году (см. [1, 2, 3]) деятельность по программной реализации алгоритмов, позволяющих проводить вычисления с гарантированной оценкой точности. Цель нашей работы заключается в развитии и популяризации этих современных алгоритмов.

Все вычисления на компьютере проводятся приближенно из-за округлений при выполнении элементарных арифметических операций (погрешности появляются уже при вводе данных в память компьютера). Поэтому при реализации на компьютере какого-либо численного алгоритма возникают вычислительные погрешности. В результате накопления машинных погрешностей вычисленное решение может сильно отличаться от истинного (см. вычислительные парадоксы в [8]) и важно уметь оценивать границы интервала, в котором находится точное решение.

В начале 50-х годов XX века Гивенс ввел метод, называемый методом обратного анализа погрешностей [14], который позволяет контролировать накопление ошибок округлений в процессе выполнения вычислительного алгоритма на компьютере. Суть метода заключается в том, что арифметические погрешности, возникающие в процессе вычислений трактуются как возмущения исходных данных задачи, что позволяет оценить погрешность решения, используя соответствующую теорию возмущений. Оказалось, однако, что для подавляющего большинства алгоритмов метод обратного анализа практически не примерим из-за его громоздкости.

В работах [8, 11, 9, 13] опубликованы результаты исследования ряда алгоритмов линейной алгебры с точки зрения возможности применения метода обратного анализа для вычисления погрешности. Продолжая эту линию, мы получили формулы гарантированных оценок погрешностей выполнения алгоритмов решения основных задач линейной алгебры [1, 2, 3]. Эти алгоритмы были реализованы в виде программного пакета **GALA-2.0** (Guaranteed Accuracy in Linear Algebra), (Свидетельство об официальной регистрации программы для ЭВМ № 2007614804 Федеральной службы по интеллектуальной собственности, патентам и товар-

ным знакам). Основная особенность процедур этого пакета в том, что вычисленный результат сопровождается гарантированной оценкой его точности. Таким образом, результат приобретает характер не "гипотезы" (предположительного значения), а "теоремы" (приводятся границы интервала, в котором лежит точное значение решения).

В данной работе мы впервые переходим от задач линейной алгебры к гораздо более сложной проблеме – оценке точности выполнения алгоритма решения задачи математической физики. Эти алгоритмы имеют комплексный характер и включают в себя такие элементы, как дискретизация, интерполяция, численное интегрирование и решение серий задач линейной алгебры. Следовательно, при получении гарантированной оценки решения необходимо учитывать погрешности всех этих операций.

Так как это **первая** работа данного направления, то мы в качестве объекта исследования выбрали алгоритм ортогональной прогонки решения краевой задачи для системы обыкновенных дифференциальных уравнений. Такой выбор обусловлен следующими причинами. Первая, это – простота алгоритма. Весь алгоритм состоит из цепочки однотипных шагов (ортогонализация + интегрирование задачи Коши) и решения цепочки линейных алгебраических систем уравнений. Такая структура облегчает нашу задачу, тем более, что мы можем использовать полученные нами ранее гарантированные оценки точности ортогонализации и решения линейных алгебраических систем [3]. Вторая, это – свобода в выборе алгоритмов для промежуточных шагов. Как уже говорилось, для решения алгебраических задач мы выбрали алгоритмы, с которым мы работали ранее, а для решения задач Коши на данном начальном этапе был выбран простейший алгоритм Эйлера. Интерполяцию начальных данных мы тоже выбираем самую простую – линейную. Третья причина, это – надежность алгоритма. За основу нами была взята работа [9]. В ней доказано, что результат ортогональной прогонки устойчив к ошибкам округления. Фактически, в этой работе показано, что к ортогональной прогонке применим метод обратного анализа погрешностей и итоговая погрешность решения оценивается пропорционально вносимой погрешности. Порядок же коэффициента пропорциональности не оценивается. Таким образом, наша задача свелась к тому, чтобы для конкретных алгоритмов внутренних шагов оценить вносимую погрешность и вычислить коэффициент пропорциональности. Результат этой деятельности мы представляем ниже.

Во втором разделе дана постановка краевой задачи и краткое описание метода ортогональной прогонки. В третьем разделе приведены факты из теории возмущений, на которых базируются итоговые оценки. В

четвертом разделе мы последовательно разбираем шаги ортогональной прогонки на предмет накопления арифметической погрешности. И наконец, пятый раздел посвящен численным примерам.

Очевидно, что поскольку мы вычисляем **гарантированные оценки** погрешности, то они должны были получиться завышенными. Забегая вперед, отметим, что действительность превзошла наши ожидания. Полученные оценки оказались достаточно грубыми. Основную причину такого положения мы видим в несовершенстве подхода, предложенного в [9], который мы взяли за основу. Дальнейшими шагами по преодолению сложившейся ситуации должны стать как модификация способа применения метода обратного анализа к ортогональной прогонке, так и совершенствование оценок погрешностей промежуточных шагов.

Тем не менее еще раз подчеркнем, что в данной работе впервые получены гарантированные оценки погрешности решения задачи математичекой физики. Это говорит о том, что несмотря на сложность соответствующих алгоритмов, решение задач математической физики **может** сопровождаться гарантированной оценкой.

Данная работа связана с участием авторов в Проекте № 46 (Конкурса междисциплинарных, интеграционных проектов фундаментальных исследований СО РАН) "Исследование и моделирование физиологических, молекулярногинетических и биофизических механизмов формирования артериальной гипертонии с целью создания оптимальных программ ранней диагностики, прогнозирования осложнений и их профилактики". Впервые предпринята попытка компьютерного моделирования сердечно-сосудистой системы человека как численного решения системы гемодинамики методом прогонки. Работа завершена изданием монографии [4] совместно с биологами, медиками, физиками и математиками.

2 Метод ортогональной прогонки

2.1 Предварительные замечания

Метод ортогональной прогонки предназначен для численного решения краевой задачи (1). Он зарекомендовал себя на практике как эффективное и надежное средство численного решения краевой задачи. Метод обладает повышенной устойчивостью к погрешностям округлений при реализации метода на компьютере в предположении, что решаемая краевая задача имеет матрицы Грина $G(x,s), G_L(x), G_R(x)$ ограниченные не слишком большой постоянной K (11).

Опишем коротко метод численного решения краевой задачи

$$\begin{aligned} &\frac{du}{dx} = A(x)u(x) + f(x), \\ &Lu(x_0) = \varphi, \qquad Ru(x_m) = \psi, \\ &x_0 \le x \le x_m, \qquad d = x_m - x_0, \end{aligned} \tag{1}$$

где A – матрица размера $n \times n$, f – вектор-функция размера n, L, R – прямоугольные матрицы размера $k \times n$ и $p \times n$ соответственно:

$$L = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1n} \\ l_{21} & & & \\ \vdots & & & \\ l_{k1} & l_{k2} & \dots & l_{kn} \end{pmatrix}, \quad R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & & & \\ \vdots & & & \\ r_{p1} & r_{p2} & \dots & r_{pn} \end{pmatrix},$$

причем $k + p = n$ и ранги матриц L и R соответственно равны k и p. Элементы матрицы $A(x)$ и вектор-функции $f(x) = (f_1(x), \dots, f_n(x))^T$ предполагаются непрерывными на отрезке $[x_0, x_m]$. Метод основан на представлении решения задачи (1) через решения серий задач Коши.

Пусть $z_1, z_2, \dots, z_k$ - полная ортонормированная система линейно независимых векторов удовлетворяющих краевому условию $Lz(x_0) = 0$. Пусть вектор z_f удовлетворяет левому условию $Lz_f(x_0) = \varphi$ и ортогонален всем z_j $(j = 1, 2, \dots, k)$: $(z_f, z_j) = 0$. Очевидно, что если решение задачи (1) существует, то оно представимо в виде:

$$u(x) = y_f(x) + \sum_{j=1}^{p} \alpha_j y_j(x),$$

где вектор-функции $y_f(x), y_1(x), \dots, y_p(x)$ – решения следующих задач Коши:

$$\begin{aligned} &\frac{dy_f(x)}{dx} = A(x)y_f(x) + f(x), \quad y_f(x_0) = z_f(x_0), \\ &\frac{dy_j(x)}{dx} = A(x)y_j(x), \quad y_j(x_0) = z_j(x_0), \quad j = 1, \dots, p, \end{aligned} \tag{2}$$

а коэффициенты $\alpha_j \quad j = 1, \dots, p$ определяются из правого граничного условия $Ru(x_m) = \psi$ как решения систем линейных уравнений

$$\sum_{j=1}^{p} \alpha_j R y_j(x_m) = \psi - R y_f(x_m).$$

Но во многих случаях описанная процедура практически невыполнима из-за ряда причин (например, "сплющивания" системы векторов $y_f(x), y_1(x), \ldots, y_p(x)$). Для преодоления явления "сплющивания" базиса, можно применить несколько раз процедуру его ортогонализации. Что и было сделано С. К. Годуновым в [5] (ортогонализация Грама-Шмита), затем С. В. Кузнецовым (ортогональные отражения Хаусхолдера) [9]. При этом весь отрезок $[x_0, x_m]$ разбивается на m интервалов. На каждом интервале решается серия задач Коши (4) и проводиться ортогонализация базисов векторов. Такой метод решения краевой задачи называется методом ортогональной прогонки (левое граничное условие как бы "перегоняется" с одного интервала на следующий за ним справа).

Здесь в методе ортогональной прогонки мы используем преобразования отражения Хаусхолдера, что дает следующие преимущества:

1) такой способ ортогонализации характеризуется повышенной устойчивостью к ошибкам округления по сравнению с методом Грама-Шмидта;

2) точность получения ортогональной системы векторов при использовании преобразований отражения не зависит от числа обусловленности прямоугольной матрицы, составленной из векторов, подлежащих ортогонализации;

3) позволяет осуществлять более крупные шаги ортогонализации;

4) использование метода отражений для ортогонализации систем векторов позволяет реально провести оценку точности полученного численного решения, то есть оценить близость численного решения к истинному.

Ниже дан алгоритм численного решения краевой задачи (1) методом ортогональной прогонки.

2.2 Алгоритм

Дано: краевая задача (1).

Разбивка отрезка $[x_0, x_m]$**.**

Разобьем отрезок $[x_0, x_m]$ на m участков точками

$$x_0 < x_1 < x_2 < \cdots < x_{s-1} < x_s < \cdots < x_m$$

так, что

$$|x_s - x_{s-1}| \le \frac{C}{\max\limits_{x\in[x_0,x_m]} \|A(x)\|}, \qquad C \approx 1 \div 3. \tag{3}$$

Определение базиса решений однородного левого граничного условия.

Определяем $z_j(x_0),\ j = 1, \ldots, p$ – полную ортонормированную систему векторов

$$(z_j, z_i) = \begin{cases} 1, & \text{если} \quad j = i, \\ 0, & \text{если} \quad j \neq i, \end{cases}$$

удовлетворяющих условию

$$Lz_j(x_0) = 0.$$

Вычисление решения неоднородного левого граничного условия.

Находим вектор $z_f(x_0)$, ортогональный ко всем $z_j(x_0)$ (т.е. $(z_f, z_j) = 0$), удовлетворяющий уравнению

$$Lz_f(x_0) = \varphi.$$

Прямая прогонка.

Цикл по $s = 1, \ldots, m$, отрезок $[x_{s-1}, x_s]$.

Решение задач Коши.

Путем численного интегрирования следующей серии задач Коши находим векторы $y_f(x_s)$, $y_j(x_s)$:

$$\begin{aligned} \frac{dy_f(x)}{dx} &= A(x)y_f(x) + f(x), \quad y_f(x_{s-1}) = z_f(x_{s-1}), \\ \frac{dy_j(x)}{dx} &= A(x)y_j(x), \quad y_j(x_{s-1}) = z_j(x_{s-1}), \quad j = 1, \ldots, p. \end{aligned} \tag{4}$$

Ортогонализация.

В результате ортогонализации набора векторов $y_1(x_s), \ldots, y_p(x_s), y_f(x_s)$, используя, например, QR-разложение, получаем векторы $z_1(x_s), \ldots, z_p(x_s), z_f(x_s)$ и матрицу Ω_s размера $(p+1) \times (p+1)$ такую, что

$$Y_s = Z_s\Omega_s, \tag{5}$$

где

$$Y_s = [y_1(x_s), \ldots, y_p(x_s), y_f(x_s)], \quad Z_s = [z_1(x_s), \ldots, z_p(x_s), z_f(x_s)],$$

причем $Z_s^* Z_s = I_n$, а Ω_s – верхнетреугольная матрица вида

$$\Omega_s = \left(\begin{array}{cccc|c} & & & & \theta_1^{(s)} \\ & & \Theta_s & & \vdots \\ & & & & \theta_p^{(s)} \\ \hline 0 & 0 & \dots & 0 & 1 \end{array} \right).$$

Конец цикла.

Обратная прогонка.

Правое граничное условие. Вектор-коэффициент α однозначно определяется из граничного условия

$$Ru(x_m) = RZ_m\alpha = \psi,$$

т. е. из решения системы линейных уравнений

$$\sum_{j=1}^{p} \alpha_j R z_j(x_m) = \psi - R z_f(x_m). \tag{6}$$

Замечание. Если решение системы единственно, то можно утверждать, что решение задачи (1) единственно и наоборот.

Вычисление решения $u(x)$.

Находим значения искомой вектор-функции $u(x)$ в точках

$$x_m, x_{m-1}, \dots, x_s, \dots, x_1, x_0$$

следующим образом. Пусть

$$\beta_s = \begin{pmatrix} \beta_s^{(1)} \\ \beta_s^{(2)} \\ \vdots \\ \beta_s^{(p)} \\ 1 \end{pmatrix}, \qquad s = m, m-1, \dots, 2, 1.$$

Положим $\beta_m^{(j)} = \alpha_j,\ j = 1, \dots, p$. Тогда, учитывая, что на каждом шаге s при прямой прогонке уже вычислены Z_s и Ω_s (5), находим

$$u(x_s) = Z_s \beta_s, \tag{7}$$

где β_s определяется через β_{s+1} путем решения системы уравнений с хорошо обусловленной верхнетреугольной матрицей Ω_{s+1}:

$$\Omega_{s+1}\beta_s = \beta_{s+1}. \tag{8}$$

Оценка вычислительных погрешностей.

Определить погрешность результата по формуле (75).

Результат выполнения алгоритма – вычисленное решение краевой задачи (1) с оценкой точности результата.

Схематично этот алгоритм представлен на рис. 1.

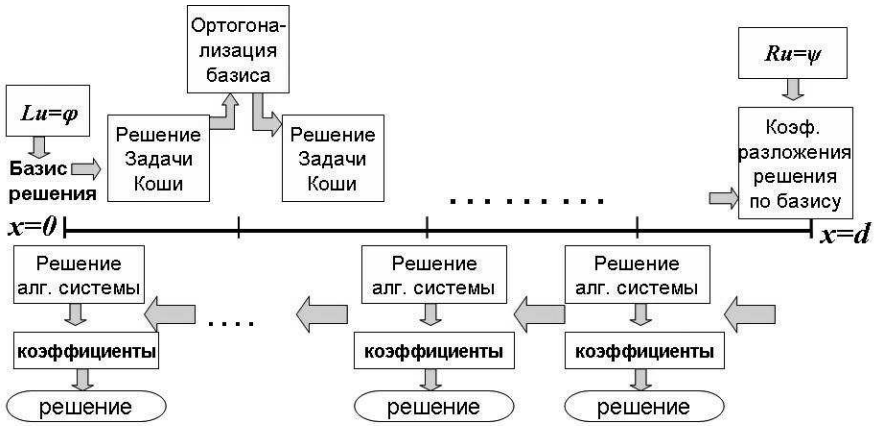

Рис. 1. Схема метода ортогональной прогонки.

3 Теория возмущений

Пусть краевая задача

$$\begin{aligned} &\frac{du(x)}{dx} = A(x)u(x) + f(x), \quad 0 \le x \le d \\ &Lu(0) = \varphi, \qquad Ru(d) = \psi, \end{aligned} \tag{9}$$

правильно поставлена. (Здесь $A(x)$ – матрица размера $n \times n$, L, R – прямоугольные матрицы размера $k \times n$ и $p \times n$ соответственно, причем $k + p = n$ (ранги матриц L, R равны k и p), вектор-функция $f(x)$ – размера n).

Это означает, что краевая задача (9) имеет единственное решение для любых φ, ψ и любой непрерывной на отрезке $[0, d]$ вектор-функции $f(x)$. Тогда решение задачи представляется с помощью матриц Грина $G(x, s), G_L(x), G_R(x)$ (см. [6])

$$u(x) = G_L(x)\varphi + \int_0^d G(x, s)f(s)ds + G_R(x)\psi, \tag{10}$$

где $G_L(x)$ – матрица-функция размера $n \times k$, решение задачи

$$\begin{cases} \dfrac{dG_L(x)}{dx} = A(x)G_L(x), \\ \\ LG_L(0) = I_k, \qquad RG_L(d) = O_{p\times k}, \end{cases}$$

матрица-функция $G_R(x)$ размера $n \times p$, решение задачи

$$\begin{cases} \dfrac{dG_R(x)}{dx} = A(x)G_R(x), \\ \\ LG_R(0) = O_{k\times p}, \qquad RG_R(d) = I_p, \end{cases}$$

и матрица-функция $G(x, s)$ размера $n \times n$, решение задачи

$$\begin{cases} \dfrac{dG(x,s)}{dx} = A(x)G(x, s), \\ \\ LG(0, s) = O_{k\times n}, \qquad RG(d, s) = O_{p\times n}, \\ \\ G(s + 0, s) - G(s - 0, s) = I_n, \end{cases}$$

где $O_{l\times m}$ – нулевая матрица размера $l \times m$.

Будем говорить, что задача (9) хорошо обусловленна, если для любых x и s из интервала $[0, d]$ верны оценки

$$\|G_L(x)\| \le K, \quad \|G(x, s)\| \le K, \quad \|G_R(x)\| \le K. \tag{11}$$

Из (10) следует, что для решения хорошо обусловленной задачи (9) справедлива оценка

$$\max_{x\in[0,d]} \|u(x)\| \le K(\|\varphi\| + \|\psi\| + d \max_{x\in[0,d]} \|f(x)\|). \tag{12}$$

Это позволяет утверждать, что решение задачи (9) устойчиво по отношению к возмущениям.

Предположим, что мы одновременно рассматриваем две "близкие" краевые задачи на отрезке $0 \le x \le d$:

$$\begin{cases} \dfrac{du(x)}{dx} = A(x)u(x) + f(x), \\ \\ Lu(0) = \varphi, \quad Ru(d) = \psi; \end{cases} \qquad \begin{cases} \dfrac{d\widetilde{u}(x)}{dx} = \widetilde{A}(x)\widetilde{u}(x) + \widetilde{f}(x), \\ \\ \widetilde{L}\widetilde{u}(0) = \widetilde{\varphi}, \quad \widetilde{R}\widetilde{u}(d) = \widetilde{\psi}. \end{cases} \tag{13}$$

Матрицы $\widetilde{L}, \widetilde{R}$, входящие в граничные условия, предполагаются имеющими то же число строк и столбцов, что и L, R соответственно. Утверждение, что первая и вторая задачи "близкие", можно понимать как утверждение, что имеют место неравенства

$$\begin{array}{ll} \|\widetilde{A} - A\| \le \epsilon_A, & \|\widetilde{f} - f\| \le \epsilon_f, \\ \|\widetilde{\varphi} - \varphi\| \le \epsilon_\varphi, & \|\widetilde{\psi} - \psi\| \le \epsilon_\psi, \\ \|\widetilde{L} - L\| \le \epsilon_L, & \|\widetilde{R} - R\| \le \epsilon_R, \end{array} \tag{14}$$

где

$\max\{\epsilon_A, \epsilon_f, \epsilon_\varphi, \epsilon_\psi, \epsilon_L, \epsilon_R\} = \varepsilon$ – достаточно маленькое число.

Всюду $f(x), \widetilde{f}(x)$ – непрерывные вектор-функции. Оказывается, что из разрешимости первой задачи при не слишком большом ε вытекает разрешимость второй.

Теорема 1 *Пусть $u(x)$ – решение правильно поставленной задачи (9), для функций Грина, которой справедливы оценки (11). Пусть наряду с краевой задачей (9) имеется близкая к ней краевая задача (13), причем для элементов, определяющих эту краевую задачу, справедливы оценки (14). Тогда при*

$$\varepsilon < \min\{1/[2K(2+d)], 1/(2+d)\} \tag{15}$$

существует и единственно $\widetilde{u}(x)$ – решение задачи (13) и для него верна оценка

$$\max_{x\in[0,d]} \|\widetilde{u}(x) - u(x)\| \le \varepsilon\mu, \tag{16}$$

причем μ – число обусловленности задачи равно

$$\mu = K(2+d)(1 + \max_{x\in[0,d]} \|\widetilde{u}(x)\|) \le K(2+d)(1+2KF), \tag{17}$$

где

$$F = 1 + \|\varphi\| + \|\psi\| + d \max_{x\in[0,d]} \|f(x)\|.$$

(Доказательство – в [9].)

Оценка означает, что решение $u(x)$ краевой задачи непрерывно зависит от коэффициентов A, L, R и правых частей $\varphi, \psi, f(x)$. В оценку непрерывности входит длина d отрезка, на котором ищется решение, и оценка K норм матриц Грина. Видно, что первое выражение для числа обусловленности μ в (17) удобно использовать, когда уже есть вычисленное решение $\widetilde{u}(x)$, второе – при прогнозировании погрешности, когда численное решение еще не известно.

Вспомогательные леммы. Нам понадобятся следующие леммы.

Лемма 1 *Пусть Z и Δ –матрицы размера $n \times p$ $(n > p)$, обладающие свойствами*

$$Z^*Z = I_p, \qquad \|\Delta\| \le \varepsilon,$$

где ε – достаточно мало. Тогда для любой квадратной $(n \times n)$-матрицы M существует квадратная $(n \times n)$-матрица $\widetilde{M}$ такая, что

$$MZ + \Delta = \widetilde{M}Z$$

и верна оценка

$$\|\widetilde{M} - M\| \le p\varepsilon. \tag{18}$$

Лемма 2 *Пусть даны матрицы Z и Δ размера $n \times p$ $(n > p)$, такие, что*

$$Z^*Z = I_p, \qquad \|\Delta\| \le \varepsilon,$$

где ε – достаточно мало. Матрица $U(x_s)$ размера $n \times p$ – решение задачи Коши:

$$\frac{dU(x)}{dx} = A(x)U(x), \qquad U(x_{s-1}) = Z,$$

причем

$$|x_s - x_{s-1}| \le \frac{C}{\max\limits_{x \in [x_{s-1}, x_s]} \|A(x)\|}.$$

Тогда матрица $\widetilde{U}(x_s) = U(x_s) + \Delta$ может быть получена, как решение задачи Коши:

$$\frac{d\widetilde{U}(x)}{dx} = \widetilde{A}(x)\widetilde{U}(x), \qquad \widetilde{U}(x_{s-1}) = Z,$$

причем, при $\varepsilon \le \dfrac{1}{2pe^C}$ справедлива оценка

$$\max_{x \in [x_{s-1}, x_s]} \|\widetilde{A}(x) - A(x)\| \le 2e^C(C+1)p\varepsilon/(x_s - x_{s-1}) \le \frac{1}{2SK}, \tag{19}$$

где $S = \max\{2, d\}$. (Доказательство – в [9].)

4 Детали алгоритма и его погрешности

Численное решение задачи (1), полученное по Алгоритму 2.2, обозначим $u^{[c]}(x)$ ("c"от английского "computation"). Вычисленное решение $u^{[c]}(x)$ можно рассматривать как решение краевой задачи

$$\begin{aligned} &\frac{du^{[c]}(x)}{dx} = \widetilde{A}(x)u^{[c]}(x) + \widetilde{f}(x), \\ &Lu^{[c]}(x_0) = \widetilde{\varphi}, \qquad Ru^{[c]}(x_m) = \widetilde{\psi}, \end{aligned} \tag{20}$$

где возмущения

$$\|\widetilde{\varphi} - \varphi\| = \|\triangle\varphi\| \le \epsilon_\varphi, \qquad \|\widetilde{\psi} - \psi\| = \|\triangle\psi\| \le \epsilon_\psi,$$

$$\max_{x\in[x_0,x_m]} \|A - \widetilde{A}\| = \|\triangle A\| \le \epsilon_A, \max_{x\in[x_0,x_m]} \|f(x) - \widetilde{f}(x)\| = \|\triangle f_1 + \triangle f_2\| \le \epsilon_f,$$

невелики. Следовательно, краевая задача (20) может считаться достаточно близкой к исходной (1) и применение Теоремы 1 дает оценку решения

$$\max_{x\in[x_0,x_m]} \|u^{[c]}(x) - u(x)\| \le \varepsilon\mu, \tag{21}$$

причем число обусловленности μ равно

$$\mu = K(2+d)(1 + \max_{x\in[x_0,x_m]} \|\widetilde{u}(x)\|).$$

Видно, что для вычисления погрешности решения (21) необходимо знать ε и K. Для оценки K (11) существуют алгоритмы вычисления матриц Грина методом ортогональной встречной прогонки (см. [6]). Чтобы оценить погрешности в неравенствах (14) и выбрать $\varepsilon = \max\{\epsilon_\varphi, \epsilon_\psi, \epsilon_L, \epsilon_R, \epsilon_A, \epsilon_f\}$ нужно повторить процесс решения краевой задачи по Алгоритму 2.2, учитывая погрешности, которые возникают в промежуточных вычислениях.

Забегая вперед скажем, что методика описанная в [9] позволяет рассматривать вычислительные погрешности, возникающие при прямой прогонке, как возмущения матрицы $\triangle A$ и правой части $\triangle f_1$, погрешности, возникающие при обратной прогонке – как возмущение правой части $\triangle f_2$, в граничных же точках их можно представить как возмущения $\triangle\varphi$, $\triangle\psi$ и, применяя метод обратного анализа погрешностей, оценить. Мы используем эту методику только при прямой прогонке. Погрешности обратной прогонки оцениваются напрямую.

4.1 Ввод данных

В зависимости от исследуемой задачи природа погрешности входных данных для алгоритма прогонки может быть самой разной. Если входные данные вычисляются на основе точных данных, то погрешность этих вычислений порядка ε_1 (ε_1 – параметр разрядной сетки компьютера и при вычислениях с двойной точностью $\varepsilon_1 \sim 10^{-16}$). А так как эта погрешность вносится в решение один раз, то не нужно следить за ее накоплением, т. е. мы пренебрегаем погрешностями $\|\widetilde{L} - L\| \le \epsilon_L$ и $\|\widetilde{R} - R\| \le \epsilon_R$.

4.2 Левое граничное условие

Сначала коротко изложим способ определения z_j и z_f из левого граничного условия, затем оценим погрешности.

Допустим, необходимо решить систему

$$LZ = 0, \tag{22}$$

где матрица L размера $k \times n$, $(k + p = n)$, вектор Z размера n.

С помощью ортогональных преобразований отражения получаем ортогональную матрицу Q такую, что матрица L^* представима в виде произведения двух матриц – верхнетреугольной Ω размера $n \times k$ и квадратной матрицы Q размера $n \times n$. Для построения отражений воспользуемся рекомендациями из Главы 3 монографии [3]. В результате получим верхнетреугольную матрицу Ω_0

$$\Omega_0 = Q_0 L^* \qquad \text{или} \qquad L^* = Q_0^* \Omega_0.$$

При этом система (22) запишется в виде

$$\Omega_0^* Q_0 Z_0 = 0 \quad \text{или} \quad \Omega_0^* Y_0 = 0, \quad \text{где} \quad Y_0 = Q_0 Z_0. \tag{23}$$

Для системы (23) полный ортогональный базис подпространства векторов, удовлетворяющих $\Omega_0^* Y_0 = 0$, составляют векторы $e_{k+1}, \ldots, e_n$, где e_i – вектор, у которого все компоненты, кроме i-ой, равны 0, а i-я компонента равна 1.

Отсюда получаем решение в (22) $Z_0 = Q_0^* Y_0$. Столбцы $z_j(x_0)$, $j = 1, ..., p$, матрицы Z_0 являются искомым начальным базисом.

Вектор $z_f(x_0)$ находится из решения недоопределенной системы линейных уравнений

$$Lz_f = \varphi. \tag{24}$$

В результате вычисления на компьютере вместо векторов $z_j(x_0)$ и $z_f(x_0)$ будут найдены близкие к ним $z_j^{[c]}(x_0) = z_j(x_0) + \zeta_j^{(0)}$ и $z_f^{[c]}(x_0) = z_f(x_0) + \zeta_f^{(0)}$.

Так как векторы z_j получаются из столбцов матрицы Q_0, то, очевидно, погрешности $\zeta_j^{(0)}$ связаны с погрешностью вычисления матрицы Q_0:

$$\|\zeta_j^{(0)}\| \le \|Q_0^{[c]} - Q_0\| \le \epsilon_z, \tag{25}$$

где ϵ_z вычисляется по следующим формулам (см.[3], стр. 66,96):

$$\begin{aligned} \epsilon_z &= (p+2)\sqrt{n}\,\widetilde{\epsilon}_P(n), \\ \widetilde{\epsilon}_P(n) &= \varepsilon_1(5n + 2\gamma + 10). \end{aligned} \tag{26}$$

Здесь параметры компьютера: γ – основание машинной арифметики, ε_1 – относительная погрешность единицы. Обычно для персональных компьютеров $\gamma = 2$ и $\varepsilon_1 \approx 10^{-16}$.

Для оценки точности решения недоопределенной системы (24) воспользуемся упрощеным вариантом оценки (8.71) из [3]:

$$\|z_f - z_f^{[c]}\| \le [\widetilde{\epsilon}_P(n)(3(k+1)\sqrt{6k}+1)\mu(L) + \widetilde{\epsilon}_P(n)(n+1)]\|z_f^{[c]}\| \le$$

$$\le [\widetilde{\epsilon}_P(n)(3(k+1)\sqrt{6k}+1)\mu(L) + \widetilde{\epsilon}_P(n)(n+1)]\|u^{[c]}\| = \epsilon_{z_f}^{(0)}.$$

Здесь $\mu(L)$ – число обусловленности матрицы L.

В итоге погрешность вычисления начального базиса оценивается величиной

$$\max\{\epsilon_{z_f}^{(0)}, \epsilon_z\} = \epsilon_0, \tag{27}$$

т. е. вычислительный процесс при прямой прогонки (интегрирование системы (4)) начинается с неточных векторов $z_j^{[c]}(x_0)$ и $z_f^{[c]}(x_0)$, определенных с погрешностью (27).

4.3 Оценка погрешности ортогонализации

В результате интегрирования системы (4) на отрезке $[x_{s-1}, x_s]$ мы получим приближенные значения $y_1^{[c]}(x_s), \ldots, y_p^{[c]}(x_s), y_f^{[c]}(x_s)$ в точке x_s, которые рассматриваем как точные решения системы (62) и следующим шагом в Алгоритме 2.2 является их ортогонализация.

Для ортогонализации базисных векторов мы использовали метод ортогональных отражений, подробно описанный в Главах 3, 4 монографии [3]. Согласно этому методу, если

$$Y_s^{[c]} = [y_1^{[c]}(x_s), \ldots, y_p^{[c]}(x_s), y_f^{[c]}(x_s)],$$

то точное QR-разложение

$$\bar{R}_s = \bar{Q}_s^* Y_s^{[c]} \quad \text{или} \quad Y_s^{[c]} = \bar{Q}_s \bar{R}_s,$$

где

$$\bar{Q}_s = P_{p+1} P_p \ldots P_1,$$

P_j - преобразования отражения, аннулирующие элементы j-го столбца матрицы $P_{j-1} \ldots P_1 \bar{Y}_s$. Отсюда, если $\bar{q}_1, \bar{q}_2, \ldots, \bar{q}_{p+1}$ – первые $p+1$ столбцов матрицы $\bar{Q}_s$, то стартовые значения для интегрирования на следующем шаге берутся в виде

$$\bar{z}_i(x_s) = q_i, \qquad \bar{z}_f(x_s) = \bar{r}_s q_{p+1},$$

а матрица $\bar{\Omega}_s$, использующаяся при обратной прогонке, совпадает с верхней квадратной частью матрицы $\bar{R}_s$ за исключением последнего элемента на диагонали: $\bar{\omega}_{p+1,p+1}^{(s)} = 1$. В действительности при реализации этого подхода на компьютере возникают арифметические погрешности. В результате вместо векторов $\bar{z}_i$, $\bar{z}_f$ и матрицы $\bar{\Omega}_s$ будут вычислены $z_i^{[c]}(x_s)$, $z_f^{[c]}(x_s)$ и $\Omega_s^{[c]}$.

Для оценок погрешностей вычисления $\bar{Q}_s$ и $\bar{\Omega}_s$ мы как и при выводе оценки (25) воспользуемся неравенствами (3.18) и (4.18) из монографии [3]:

$$\|\bar{z}_i(x_s) - z_i^{[c]}(x_s)\| \le \|Q_s^{[c]} - \bar{Q}_s\| \le \epsilon_z, \tag{28}$$

где ϵ_z определяется по формуле (26).

Чтобы подготовить оценку точности для $z_f^{[c]}(x_s)$ воспользуемся неравенством треугольника

$$\|\bar{z}_f - z_f^{[c]}\| = \|\bar{q}_{p+1} \cdot \bar{r}_s \pm \bar{q}_{p+1} \cdot r_s^{[c]} - q_{p+1}^{[c]} \cdot r_s^{[c]}\| \le$$

$$\le \|\bar{q}_{p+1}\| \, |\bar{r}_s - r_s^{[c]}| + \|\bar{q}_{p+1} - q_{p+1}^{[c]}\| \, |r_s^{[c]}| \tag{29}$$

Далее оцениваем отдельно каждое слагаемое.

Напомним, что точность QR-разложения оценивается по формуле (4.18) из [3]:

$$\|R_s^{[c]} - \bar{R}_s\| \le (p+2)\sqrt{p+1}\,\widetilde{\epsilon}_P(n)\|\bar{R}_s\| \le \epsilon_z\|\bar{R}_s\|. \tag{30}$$

Также воспользуемся оценками

$$\|z_f(x)\| \le \|u(x)\|, \qquad \|z_f^{[c]}(x)\| \le \|u^{[c]}(x)\|.$$

По формуле (2.35) из [3] (оценка точности умножения вектора на скаляр)

$$\|z_f^{[c]}\| \le (1+\alpha)\|q_{p+1}^{[c]} \cdot r^{[c]}\|,$$

где $|\alpha| \le \varepsilon_1$

$$(1+\alpha)(1-\epsilon_z) \le \|z_f^{[c]} \cdot \frac{1}{r_s^{[c]}}\| = (1+\alpha)\|q_{p+1}^{[c]}\| \le (1+\alpha)(1+\epsilon_z).$$

Следовательно, для всех точек x_s верна оценка

$$|r_s^{[c]}| \le \frac{\|z_f^{[c]}\|}{(1-\varepsilon_1)(1-\epsilon_z)} \le \frac{\|u^{[c]}\|}{(1-\varepsilon_1)(1-\epsilon_z)}.$$

Для верхнетреугольной матрицы верна цепочка неравенств

$$\|\bar{R}_s\| = \|Y_s^{[c]}\| \le \|\bar{Y}_s\| + \epsilon_I \le e^C\|Z_{s-1}^{[c]}\| + \epsilon_I \le e^C(1+\epsilon_z + |r_{s-1}^{[c]}|) + \epsilon_I,$$

где ϵ_I – погрешность интегрирования задачи Коши (см. формулу (44)).

Подставим сюда оценку (30):

$$|r_s^{[c]} - \bar{r}_s| \le \|R_s^{[c]} - \bar{R}_s\| \le \epsilon_z\|\bar{R}_s\| \le \epsilon_z\left[e^C\left(1+\epsilon_z + \frac{\|u^{[c]}(x_{s-1})\|}{(1-\varepsilon_1)(1-\epsilon_z)}\right) + \epsilon_I\right]$$

Для получения окончательного результата используем также факт

$$\|\bar{q}_{p+1} - q_{p+1}^{[c]}\| \le \epsilon_z$$

и подставляем найденные оценки в (29):

$$\|z_f^{[c]}(x_s) - \bar{z}_f(x_s)\| \le \|\bar{q}_{p+1}\|\,|\bar{r} - r^{[c]}| + \|\bar{q}_{p+1} - q_{p+1}^{[c]}\|\,|r^{[c]}| \le$$

$$\le \epsilon_z\left[e^C\left(1+\epsilon_z + \frac{\|u^{[c]}(x_{s-1})\|}{(1-\varepsilon_1)(1-\epsilon_z)}\right) + \epsilon_I\right] + \epsilon_z\frac{\|u^{[c]}(x_s)\|}{(1-\varepsilon_1)(1-\epsilon_z)} =$$

$$= \epsilon_z \left[e^C(1+\epsilon_z) + \epsilon_I + (1+e^C)\frac{K_{sol}}{(1-\varepsilon_1)(1-\epsilon_z)} \right] = \epsilon_{z_f}^{(s)} \qquad (31)$$

Для обратной прогонки понадобиться также оценка точности матрицы Θ_s – подматрицы матрицы Ω_s (см. Алгоритм 2.2). Вновь используя формулу (4.18) из [3], получаем неравенство:

$$\|\bar{\Theta}_s - \Theta_s^{[c]}\| \le \epsilon_z \|\bar{\Theta}_s\|. \qquad (32)$$

4.4 Кусочно-линейное приближение матрицы $A(x)$ и правой части $f(x)$

Часто алгоритм ортогональной прогонки применяется к задачам, для которых матричная и векторная функции $A(x)$ и $f(x)$ заданы только в отдельных точках, т. е. матрицу и правую часть мы сможем вычислить только в точках x_s: $x_s - x_{s-1} = d/m$, $0 = x_0 < x_1 < \cdots < x_s < \cdots < x_m = d$. Для того, чтобы приблизить $A(x)$ и $f(x)$ непрерывными функциями, применим простейший метод интерполяции при помощи кусочно-линейных функций. Таким образом, задача (1) заменяется на следующую

$$\begin{aligned} &\frac{du_{lin}(x)}{dx} = A_{lin}(x)u_{lin} + f_{lin}(x), \\ &Lu_{lin}(x_0) = \varphi, \qquad Ru_{lin}(x_N) = \psi, \end{aligned} \qquad (33)$$

где на отрезке $[x_{s-1}, x_s]$ функции $A_{lin}(x)$ и $f_{lin}(x)$ представляются в виде

$$A_{lin}(x) = A_1^{(s)}x + A_0^{(s)}, \qquad f_{lin}(x) = f_1^{(s)}x + f_0^{(s)},$$

так что

$$\begin{aligned} A_{lin}(x_{s-1}) &= A(x_{s-1}), \qquad A_{lin}(x_s) = A(x_s) \\ f_{lin}(x_{s-1}) &= f(x_{s-1}), \qquad f_{lin}(x_s) = f(x_s) \end{aligned}$$

Для того, чтобы оценить погрешности $\|f(x) - f_{lin}(x)\|$ и $\|A(x) - A_{lin}(x)\|$, используем теорему о полиномиальной интерполяции.

Теорема 2 *Пусть $f(x)$ имеет $n+1$ непрерывную производную, тогда для фиксированных узлов $x_0 < x_1 < \cdots < x_n$ существует единственный полином $p(x)$ степени не больше n такой, что $p(x_j) = f(x_j)$. Причем для любой точки x, $x_0 < x < x_n$ имеется представление*

$$f(x) - p(x) = \frac{(x-x_0)(x-x_1)\dots(x-x_n)}{(n+1)!} f^{(n+1)}(z), \qquad \text{где } z \in [x_0, x_n].$$

В нашем случае интерполяция производится полиномами первого порядка. Поэтому введем величину K_{dif}, которая оценивает вторые производные элементов матричной и векторной функций $A(x)$ и $f(x)$ так, что

$$\|A''(x)\| \le K_{dif} \quad \text{и} \quad \|f''(x)\| \le K_{dif} \tag{34}$$

при $x \in [x_0, x_m]$. Тогда по теореме о полиномиальной интерполяции получаем требуемые оценки

$$\|f(x) - f_{lin}(x)\| \le K_{dif}\frac{h^2}{2} = \epsilon_{lin}, \qquad \|A(x) - A_{lin}(x\| \le K_{dif}\frac{h^2}{2} = \epsilon_{lin}. \tag{35}$$

На основании этих неравенств по Теореме 1 получаем оценку для разности решений исходной задачи (1) и задачи с кусочно-линейными коэффициентами

$$\max_{x\in[x_0,x_m]} \|u - u_{lin}\| \le \frac{h^2}{2} K_{dif}\mu, \tag{36}$$

где

$$\mu = K_{lin}(2+d)(1 + \max_{x\in[x_0,x_m]} \|u_{lin}\|).$$

Здесь K_{lin} означает оценку матричной функции Грина задачи (33).

Далее в тех случаях, когда кусочная линейность функций несущественна, мы будем опускать обозначение "lin".

4.5 Погрешности интегрирования

Метод интегрирования

Для интегрирования мы используем самый простой метод – метод Эйлера. У этого метода есть недостаток – уменьшение его погрешности (т. е. соответствующее уменьшение шага интегрирования) требует больших вычислительных затрат. Несмотря на это мы используем этот метод по следующим соображениям. Во-первых, во многих задачах интегрируются не сами функции, а их приближения. В данном случае используется также наиболее простое кусочно-линейное приближение (см. Раздел 4.4). Следовательно, уже на этапе интерполяции коэффициентов задачи вносится определенная погрешность. И при выборе метода интегрирования достаточно следить за тем, чтобы погрешность интегрирования не превышала погрешности интерполяции. В предлагаемом варианте это условие будет соблюдено.

Замечание. Из этих рассуждений также следует, что для того, чтобы повысить общую точность решения задачи, нужно параллельно повышать порядок интерполяции и точность интегрирования.

Вторая причина использования именно метода Эйлера – это его простота, которая позволяет относительно легко проследить за накоплением и распространением погрешностей в процессе интегрирования.

При этом метод Эйлера можно все же несколько видоизменить,если учесть специальный кусочно-линейный вид матрицы коэффициентов и правой части с тем, чтобы минимизировать погрешности интегрирования.

Погрешность метода интегрирования на одном шаге

Рассмотрим отрезок $[z_0, z_1]$, где $z_1 - z_0 = \Delta$. Пусть $v(z)$ есть решение задачи Коши

$$\frac{dv}{dz} = A_{lin}(z)v + f_{lin}(z), \qquad v(z_0) = v_0.$$

тогда

$$v(z_1) = v_0 + \int_{z_0}^{z_1} (A_{lin}(z)v(z) + f_{lin}(z))dz$$

Сравним это выражение со следующим:

$$w(z_1) = v_0 + (A_{mid}v_0 + f_{mid})\Delta.$$

Здесь введено обозначение

$$A_{mid} = \frac{A_{lin}(z_1) - A_{lin}(z_0)}{2}, \qquad f_{mid} = \frac{f_{lin}(z_1) - f_{lin}(z_0)}{2}.$$

Так как

$$A_{mid} = (A_0 + A_1 z_1)(z_1 - z_0) - A_1 \frac{(z_1 - z_0)^2}{2}$$

и

$$\int_{z_0}^{z_1} f_{lin}(z)dz = f_{mid},$$

то

$$v(z_1) - w_{z_1} = \int_{z_0}^{z_1} A_{lin}(z)v(z)dz - A_{lin}(z_1)v_0\Delta + A_1 v_0 \frac{\Delta^2}{2}.$$

Перегруппируем слагаемые и введем для них обозначения

$$v(z_1) - w(z_1) = A_0 \left(\int_{z_0}^{z_1} v(z)dz - v_0\Delta \right) +$$

$$+A_1\left(\int_{z_0}^{z_1} v(z)zdz - z_1v_0\Delta + v_0\frac{\Delta}{2}\right) = A_0S_0 + A_1S_1.$$

Оценка первого слагаемого дает следующий результат

$$\|S_0\| = \left\|\int_{z_0}^{z_1} (v(z) - v_0)\, dz\right\| \le \max|v'(z)|\frac{\Delta^2}{2}.$$

А так как

$$z_1 = \frac{z_1 - z_0}{2} + \frac{z_1 + z_0}{2} \quad \text{и} \quad z_1\Delta - \Delta/2 = \frac{z_1^2 - z_0^2}{2},$$

то

$$\|S_1\| = \left\|\int_{z_0}^{z_1} (v(z) - v_0)\, zdz\right\| =$$

$$= \left\|\int_{z_0}^{z_1} v'(\xi(z))(z - z_0)zdz\right\| \le \max|v'(z)|\left(z_0\frac{\Delta^2}{2} + \frac{\Delta^3}{3}\right).$$

Учтем также, что для средней точки z_{mid} отрезка $[z_0, z_1]$ имеют место выражения:

$$z_{mid} = \frac{\Delta}{2} + z_0, \qquad z_0\frac{\Delta^2}{2} + \frac{\Delta^3}{3} = \frac{2}{3}\Delta^2\left(\frac{3}{4}z_0 + \frac{\Delta}{2}\right) \le \frac{2\Delta^2}{3}z_{mid}.$$

Тогда общая погрешность одного шага представленной модификации метода Эйлера оценивается по формуле

$$\|v(z_1) - w(z_1)\| \le \frac{2}{3}\Delta^2 K_{sol}\|A_{mid}\|, \tag{37}$$

где

$$\|A_{mid}\| = \|A_0\| + |z_{mid}|\|A_1\|,$$

$$\max|v'(z)| \le \|v\|_{C^1} = \max|v(z)| + \max|v'(z)| = K_{sol}, \tag{38}$$

где $z_{mid} = (z_1 - z_0)/2$.

Арифметическая погрешность на одном шаге интегрирования

Сравним результат точного выражения

$$w(z_1) = v_0 + (A_{mid}v_0 + f_{mid})\Delta$$

с результатом, полученным при вычислении на компьютере

$$w^{[c]}(z_1) = v_0 \oplus (A_{mid}^{[c]} \otimes v_0 \oplus f_{mid}^{[c]}) \otimes \Delta,$$

где

$$A_{mid}^{[c]} = (A_{lin}(z_1) \ominus A_{lin}(z_0)) \oslash 2, \qquad f_{mid}^{[c]} = (f_{lin}(z_1) \ominus f_{lin}(z_0)) \oslash 2.$$

Здесь и далее в кружочках обозначены машинные (компьютерные) бинарные арифметические операции над машинными числами. Так как моделирование погрешностей машинного вычитания происходит по правилу

$$a \ominus b = (1+\alpha)(a-b) + \beta, \qquad |\alpha| \le \varepsilon_1, \quad |\beta| \le \varepsilon_0,$$

(см. [3]), а деление на 2 происходит с абсолютной погрешностью не более ε_0, то

$$\|\Omega_A\| = \|A_{mid}^{[c]} - A_{mid}\| \le \varepsilon_0(1+n/2) + \varepsilon_1\sqrt{n}\|A_{mid}\|,$$
$$\|\omega_f\| = \|f_{mid}^{[c]} - f_{mid}\| \le \varepsilon_0(1+\sqrt{n}/2) + \varepsilon_1\|f_{mid}\|,$$

где $n \times n$ – размер матрицы A. Следовательно,

$$w^{[c]}(z_1) = v_0 \oplus [(A_{mid} + \Omega_A) \otimes v_0 \oplus (f_{mid} + \omega_f)] \otimes \Delta.$$

При дополнительных условиях

$$\|A_{mid}\| \ge \frac{\varepsilon_0(1+n/2)}{\varepsilon_1}$$

и

$$\|f_{mid}\| \ge \frac{\varepsilon_0(1+\sqrt{n}/2)}{\varepsilon_1}$$

оценки имеют относительный характер

$$\|\Omega_A\| \le \varepsilon_1(\sqrt{n}+1)\|A_{mid}\|, \qquad \|\omega_f\| \le 2\varepsilon_1\|f_{mid}\|.$$

Именно их и имеет смысл использовать, так как дополнительные условия нарушаются только в исключительных случаях для практически нулевой матрицы A и правой части f.

Оценка точности машинного умножения матрицы на вектор имеет вид

$$\|A \otimes v - Av\| \le (n+1)\sqrt{n}\varepsilon_1\|A\|\|v\|$$

при условии

$$\sqrt{2n\varepsilon_0/\varepsilon_1} \le \|v\| \le \sqrt{\varepsilon_\infty/2},$$
$$\sqrt{2n\varepsilon_0/\varepsilon_1} \le \|a_i\| \le \sqrt{\varepsilon_\infty/2},$$

где a_i – i-я строка матрицы A. Эти условия не являются слишком обременительными и, как правило, выполняются. Следовательно,

$$w^{[c]}(z_1) = v_0 \oplus [((A_{mid} + \Omega_A)v_0 + \nu) \oplus (f_{mid} + \omega_f)] \otimes \Delta,$$

где

$$\|\nu\| \le (n+1)\sqrt{n}\varepsilon_1(1 + \varepsilon_1(\sqrt{n}+1))\|A_{mid}\|\|v_0\|.$$

Для моделирования погрешностей машинного сложения и умножения на скаляр будем использовать формулы (см. Главу 2 в [3]):

$$a \oplus b = (1+\alpha)(a+b), \qquad |\alpha| \le \varepsilon_1,$$

$$a \otimes v = (1+\alpha)(av), \qquad |\alpha| \le \varepsilon_1.$$

Замечание. Мы не учитываем абсолютные погрешности порядка ε_0 не только потому, что они малы по сравнению с другими погрешностями, но и потому, что они возникают только в единичных случаях, когда результат операции близок к нулю, и, следовательно, не накапливаются.

В результате применения этих формул получим

$$w^{[c]}(z_1) = v_0 + [(A_{mid} + \Omega_A)v_0 + \nu + f_{mid} + \omega_f + \xi]\,\Delta + \zeta + \eta,$$

где

$$\|\xi\| \le \varepsilon_1\|(A_{mid} + \Omega_A)v_0 + \nu + f_{mid} + \omega_f\|,$$
$$\|\zeta\| \le \varepsilon_1\Delta\|(A_{mid} + \Omega_A)v_0 + \nu + f_{mid} + \omega_f + \xi\|,$$
$$\|\eta\| \le \varepsilon_1\|v_0 + [(A_{mid} + \Omega_A)v_0 + \nu + f_{mid} + \omega_f + \xi]\,\Delta + \zeta\|.$$

Группируя все оценки погрешностей, получим

$$w^{[c]}(z_1) = w(z_1) + \zeta + \eta + [\Omega_A v_0 + \nu + \omega_f + \xi]\,\Delta = w(z_1) + \chi,$$

а для величины χ, которая оценивает арифметические погрешности на одном шаге интегрирования, нетрудно получить неравенство.

$$\|w(z_1) - w^{[c]}(z_1)\| = \|\chi\| \le \varepsilon_1\Big[\big(1 + \Delta[4 + c_1]\|A_{mid}\|\big)\|v_0\| + 8\|f_{mid}\|\Big], \quad (39)$$

где $c_1 = (1 + \varepsilon_1\sqrt{n})(n+1)(\sqrt{n}+1) \approx n^{3/2}$, n - размер вектора.

Накопление погрешностей интегрирования

Рассмотрим отрезок $[z_0, z_1]$, где $z_1 - z_0 = \Delta$. Предположим теперь, что значение функции $v(z)$ в начальной точке z_0 задано с погрешностью: $\widetilde{v}_0 = v(z_0) + \vartheta_0$. Вычислим, как это скажется на погрешности в точке z_1. Для этого будем сравнивать величины

$$v(z_1) = v_0 + \int_{z_0}^{z_1} (A_{lin}(z)v(z) + f_{lin}(z))dz \tag{40}$$

и

$$\widetilde{w}^{[c]}(z_1) = \widetilde{v}_0 \oplus (A^{[c]}_{mid} \otimes \widetilde{v}_0 \oplus f^{[c]}_{mid}) \otimes \Delta.$$

Применим очевидное неравенство

$$\|v(z_1) - \widetilde{w}^{[c]}(z_1)\| \le \|v(z_1) - w(z_1)\| + \|w(z_1) - \widetilde{w}(z_1)\| + \|\widetilde{w}(z_1) - \widetilde{w}^{[c]}(z_1)\|. \tag{41}$$

Первое слагаемое в правой части этого неравенства было оценено ранее (37). Для того, чтобы оценить последнее слагаемое, модифицируем (39):

$$\|\widetilde{w}(z_1) - \widetilde{w}^{[c]}(z_1)\| \le \varepsilon_1 \Big[\big(1 + \Delta[4 + c_1]\|A_{mid}\|\big)\|\widetilde{v}_0\| + 8\|f_{mid}\| \Big] \le$$
$$\le \varepsilon_1 \Big[\big(1 + \Delta[4 + c_1]\|A_{mid}\|\big)\|v_0\| + 8\|f_{mid}\| \Big] + \varepsilon_1 \Big[\big(1 + \Delta[4 + c_1]\|A_{mid}\|\big)\|\vartheta_0\|.$$

Легко также вывести оценку для среднего слагаемого в (41):

$$\|w(z_1) - \widetilde{w}(z_1)\| = \|(I + \Delta A_{mid})\vartheta_0\| \le (1 + \Delta\|A_{mid}\|)\|\vartheta_0\|.$$

Следовательно, оценка имеет вид

$$\|v(z_1) - \widetilde{w}^{[c]}(z_1)\| \le (1 + (\varepsilon_1 + \Delta(1 + \varepsilon_1(4 + c_1))\|A_{mid}\|)\|\vartheta_0\| +$$
$$+ (\frac{2}{3}\Delta^2\|A_{mid}\| + \varepsilon_1 + \varepsilon_1\Delta(4 + c_1)\|A_{mid}\|)K_{sol} + \varepsilon_1 8\|f_{mid}\|.$$

Усиливая неравенство, получим

$$\|v(z_1) - \widetilde{w}^{[c]}(z_1)\| = \|\vartheta_1\| \le (1 + r)\|\vartheta_0\| + q,$$

где

$$r = \varepsilon_1 + \Delta(1 + \varepsilon_1(4 + c_1))K_A\varepsilon_1, \qquad K_A = \max\|A(z)\|,$$
$$q = (\varepsilon_1 + (\frac{2}{3}\Delta^2 + \varepsilon_1\Delta(4 + c_1))K_A)K_{sol} + \varepsilon_1 8K_f, \qquad K_f = \max\|f(z)\|,$$

а K_{sol} определяется в (38).

Таким образом, мы получили погрешность $\|\vartheta_1\|$ вычисления значения функции $v(z)$ по формуле (40) с учетом погрешности метода Эйлера, машинных погрешностей и начальной погрешности $\|\vartheta_0\|$.

Перейдем теперь к исследованию вопроса о накоплении алгоритмической и арифметической погрешностей в процессе интегрирования по отрезку. Итак, отрезок $[x_0, x_m]$ разбит на m равных частей $[x_{s-1}, x_s]$, $s = 1, 2, \ldots, m+1$, длина которых составляет $x_s - x_{s-1} = (x_m - x_0)/m = h$. Рассмотрим отдельно некоторый отрезок $[x_{s-1}, x_s]$, по которому мы будем интегрировать векторную функцию $y(x)$, являющуюся решением задачи Коши

$$\frac{dy}{dx} = A_{lin}(x)y + f_{lin}(x), \qquad y(x_{s-1}) = y_{s-1}.$$

С этой целью мы произведем дополнительное разбиение отрезка $[x_{s-1}, x_s]$, поделив его на N равных частей с шагом интегрирования $\Delta = h/N = x^{(k)} - x^{(k-1)}$:

$$x_{s-1} = x^{(0)} < x^{(1)} < x^{(2)} < \cdots < x^{(N)} = x_s.$$

Из предыдущих рассуждений получаем

$$\|\vartheta_N\| \le (1+r)\|\vartheta_{N-1}\| + q \le (1+r)^2\|\vartheta_{N-2}\| + (1+r)q + q \le \ldots$$

$$\le (1+r)^N\|\vartheta_0\| + q\sum_{i=0}^{N-1}(1+r)^i,$$

где

$$\vartheta_i = y(x^{(i)}) - \widetilde{y}^{[c]}(x^{(i)}).$$

Так как

$$(1+r)^N \le \frac{1}{1-Nr} \qquad \text{и} \qquad \sum_{i=0}^{N-1}(1+r)^i = \frac{(1+r)^N - 1}{r} \le \frac{N}{1-(N-1)r/2},$$

то в итоге получаем

$$\|\vartheta_N\| \le \frac{1}{1-Nr}\|\vartheta_0\| + \frac{N}{1-(N-1)r/2}q.$$

Заметим, что $rN \approx \Delta\varepsilon_1 K_A N = \varepsilon_1 h K_A$, $q \approx 2/3\Delta^2 K_A K_{sol}$, тогда можно прояснить порядок накопленной погрешности

$$\|\vartheta_N\| \lesssim \frac{\|\vartheta_0\|}{1-\varepsilon_1 h K_A} + \frac{2}{3}\frac{\Delta h}{1-\varepsilon_1 h K_A/2}K_A K_{sol} = k_b\|\vartheta_0\| + \epsilon_I = \epsilon_C. \quad (42)$$

Таким образом, мы получили погрешность вычисления векторной функции $y(x)$ в точке x_s с учетом накопления всех погрешностей, возникающих при интегрировании на отрезке $[x_{s-1}, x_s]$: начальной в точке x_{s-1}, машинной и алгоритмической. В оценке (42) мы используем следующие обозначения, которые понадобятся в дальнейшем:

$$k_b = 1/(1 - \varepsilon_1 h K_A) \tag{43}$$

– коэффициент влияния погрешности начальных данных на результат интегрирования задачи Коши ("b"от слова "begin"),

$$\epsilon_I = \frac{2}{3}\frac{\Delta h}{1 - \varepsilon_1 h K_A/2} K_A \tag{44}$$

– алгоритмическая и арифметическая погрешность интегрирования задачи Коши модифицированным метом Эйлера (индекс "I"означает "Integral"), ϵ_C – общая погрешность интегрирования задачи Коши ("C"по имени "Cauchy").

4.6 Обратный анализ погрешностей прямой прогонки

Рассмотрим погрешности, возникающие при прямой прогонке на шаге s (Алгоритм 2.2). Если бы вычисления выполнялись точно, то формулы s шага прямой прогонки выглядели бы следующим образом

$$y_f(x_s) = X(x_s, x_{s-1}) z_f(x_{s-1}) + X(x_s, x_{s-1}) \int\limits_{x_{s-1}}^{x_s} [X(\xi, x_{s-1})]^{-1} f(\xi) d\xi, \tag{45}$$

и

$$y_j(x_s) = X(x_s, x_{s-1}) z_j(x_{s-1}), \tag{46}$$

где $X(x, x_s)$ – матрицант, т. е. матрица-функция размера $n \times n$, решение задачи Коши

$$\frac{dX(x, x_s)}{dx} = A(x) X(x, x_s), \quad X(x_s, x_s) = I_n, \tag{47}$$

но из левого граничного условия начальный базис определяется с погрешностью (27). Таким образом, дальнейший вычислительный процесс начинается с этих неточных векторов. Далее при прямой прогонке путем численного интегрирования на отрезке $[x_{s-1}, x_s]$ серии задач Коши (4)

с учетом погрешностей ортогонализации и интегрирования в точке x_s будут вычислены значения векторов $y_f^{[c]}(x_s)$, $y_j^{[c]}(x_s)$:

$$y_f^{[c]}(x_s) = X(x_s, x_{s-1})\bar{z}_f(x_{s-1}) + X(x_s, x_{s-1}) \int\limits_{x_{s-1}}^{x_s} [X(\xi, x_{s-1})]^{-1} f(\xi) d\xi + \eta_f^{(s)}, \tag{48}$$

и

$$y_j^{[c]}(x_s) = X(x_s, x_{s-1})\bar{z}_j(x_{s-1}) + \eta_j^{(s)}, \tag{49}$$

где

$$\|\eta_f^{(s)}\| \le k_b \epsilon_{z_f}^{(s)} + \epsilon_I, \qquad \|\eta_j^{(s)}\| \le k_b \epsilon_z + \epsilon_I,$$

$\epsilon_{z_f}^{(s)}$, ϵ_z, k_b, ϵ_I определяются по формулам (31), (26), (43), (44). Обозначим

$$\max\{\|\eta_f^{(s)}\|, \|\eta_j^{(s)}\|\} = \delta. \tag{50}$$

Рассмотрим уравнение (49). Из утверждения Леммы 1 следует, что существует матрица $\widetilde{X}(x_s, x_{s-1})$ для которой верно

$$X(x_s, x_{s-1})\bar{z}_j(x_{s-1}) + \eta_j^{(s)} = \widetilde{X}(x_s, x_{s-1})\bar{z}_j(x_{s-1}) \tag{51}$$

и

$$\|\widetilde{X}(x_s, x_{s-1}) - X(x_s, x_{s-1})\| \le p\delta.$$

Это означает, что существует такая система дифференциальных уравнений

$$\frac{du(x)}{dx} = \widetilde{A}(x)u(x), \tag{52}$$

что векторы $y_1^{[c]}(x_s), y_2^{[c]}(x_s), \ldots, y_p^{[c]}(x_s)$ есть решения следующих задач Коши

$$\frac{dy_j^{[c]}(x)}{dx} = \widetilde{A}(x)y_j^{[c]}(x), \quad y_j^{[c]}(x_{s-1}) = \bar{z}_j(x_{s-1})$$

и

$$y_j^{[c]}(x_s) = \widetilde{X}(x_s, x_{s-1})\bar{z}_j(x_{s-1}),$$

где $\widetilde{X}(x_s, x_{s-1})$ – матрицант системы (52).

При этом погрешности вычислений интерпретируются как малые возмущения кусочно-линейного приближения матрицы A и при $\delta \le 1/2pe^C$ для любого интервала $[x_{s-1}, x_s]$ верна оценка (19) из Леммы 2:

$$\max_{x \in [x_{s-1}, x_s]} \|\widetilde{A}(x) - A_{lin}(x)\| \le \delta p 2e^C (C+1)/h = \epsilon_A. \tag{53}$$

Теперь рассмотрим уравнение (48) и получим оценку возмущения правой части f. Положим для остальных $x \in [x_{s-1}, x_s]$

$$\widetilde{X}(x, x_{s-1}) = X(x, x_{s-1}) + \frac{x - x_{s-1}}{x_s - x_{s-1}} \left(\widetilde{X}(x_s, x_{s-1}) - X(x_s, x_{s-1}) \right). \quad (54)$$

Перепишем выражение (48) следующим образом

$$y_f^{[c]}(x_s) = \widetilde{X}(x_s, x_{s-1}) \bar{z}_f(x_{s-1}) + \eta_f^{(s)} + \left(X(x_s, x_{s-1}) - \widetilde{X}(x_s, x_{s-1}) \right) \bar{z}_f(x_{s-1})$$

$$+ \widetilde{X}(x_s, x_{s-1}) \int_{x_{s-1}}^{x_s} [\widetilde{X}(\xi, x_{s-1})]^{-1} f_1(\xi) d\xi, \quad (55)$$

где

$$f_1(x) = \widetilde{X}(x, x_{s-1}) [\widetilde{X}(x_s, x_{s-1})]^{-1} X(x_s, x_{s-1}) [X(x, x_{s-1})]^{-1} f(x).$$

С целью оценить разность $\|f(x) - f_1(x)\|$ выполним следующие преобразования:

$$\begin{aligned}
f(x) - f_1(x) = & \left(I - \widetilde{X}(x, x_{s-1}) [\widetilde{X}(x_s, x_{s-1})]^{-1} X(x_s, x_{s-1}) [X(x, x_{s-1})]^{-1} \right) f(x) \\
\vDash & \left(X(x_{s-1}, x) [X(x_{s-1}, x_s)]^{-1} - \widetilde{X}(x, x_{s-1}) [\widetilde{X}(x_s, x_{s-1})]^{-1} \right) \\
& \times X(x_s, x_{s-1}) [X(x, x_{s-1})]^{-1} f(x) \\
= & \left(X(x, x_{s-1}) [X(x_s, x_{s-1})]^{-1} - X(x, x_{s-1}) [\widetilde{X}(x_s, x_{s-1})]^{-1} \right. \\
& \left. - \frac{x - x_{s-1}}{x_s - x_{s-1}} \left(\widetilde{X}(x_s, x_{s-1}) - X(x_s, x_{s-1}) \right) [\widetilde{X}(x_s, x_{s-1})]^{-1} \right) \\
& \times X(x_s, x_{s-1}) [X(x, x_{s-1})]^{-1} f(x) \\
= & \left(X(x, x_{s-1}) \left([X(x_s, x_{s-1})]^{-1} - [\widetilde{X}(x_s, x_{s-1})]^{-1} \right) \right. \\
& \left. - \frac{x - x_{s-1}}{x_s - x_{s-1}} \left(\widetilde{X}(x_s, x_{s-1}) - X(x_s, x_{s-1}) \right) [\widetilde{X}(x_s, x_{s-1})]^{-1} \right) \\
& \times X(x_s, x_{s-1}) [X(x, x_{s-1})]^{-1} f(x).
\end{aligned}$$

Обозначим для краткости $X = X(x_s, x_{s-1})$, $\widetilde{X} = \widetilde{X}(x_s, x_{s-1})$, $\mathcal{X} = X - \widetilde{X}$, тогда нетрудно получить следующие оценки

$$\|\widetilde{X}^{-1}\| = \|(X - \mathcal{X})^{-1}\| = \|X^{-1}(I - \mathcal{X} X^{-1})\|.$$

Если выполнено условие $\|\mathcal{X}\| \cdot \|X^{-1}\| < 1$, а в нашем случае это означает $p\delta e^C < 1$, то можно применить ряд Неймана и получить оценку

$$\|\widetilde{X}^{-1}\| = \|X^{-1}\sum_{j=0}^{\infty}(\mathcal{X}X^{-1})^j\| \le \|X^{-1}\|\sum_{j=0}^{\infty}(\|\mathcal{X}\| \cdot \|X^{-1}\|)^j =$$

$$= \frac{\|[X(x_s, x_{s-1})]^{-1}\|}{1 - \|\mathcal{X}\|\|[X(x_s, x_{s-1})]^{-1}\|}.$$

Аналогичным образом получаем

$$\|\widetilde{X}^{-1} - X^{-1}\| = \|(X - \mathcal{X})^{-1} - X^{-1}\| = \|X^{-1}(\sum_{j=0}^{\infty}(\mathcal{X}X^{-1})^j - I)\| \le$$

$$\le \frac{\|[X(x_s, x_{s-1})]^{-1}\|^2\|\mathcal{X}\|}{1 - \|\mathcal{X}\|\|[X(x_s, x_{s-1})]^{-1}\|}.$$

Поскольку в нашем случае

$$\|\widetilde{X}(x_s, x_{s-1}) - X(x_s, x_{s-1})\| = \|\mathcal{X}\| \le p\delta,$$

$$\|X(x_s, x_{s-1})\| \le e^C, \;\; \|[X(x_s, x_{s-1})]^{-1}\| \le e^C,$$

то

$$\|[\widetilde{X}(x_s, x_{s-1})]^{-1}\| \le \frac{e^C}{1 - p\delta e^C},$$

$$\|[\widetilde{X}(x_s, x_{s-1})]^{-1} - [X(x_s, x_{s-1})]^{-1}\| \le p\delta\frac{e^{2C}}{1 - p\delta e^C}.$$

Объединяя полученные оценки, выводим итоговое неравенство

$$\max_{x\in[x_{s-1},x_s]}\|f(x) - f_1(x)\| \le$$

$$\le \Big(p\delta\frac{e^{3C}}{1 - p\delta e^C} + p\delta\frac{e^C}{1 - p\delta e^C}\Big)e^{2C}\max_{x\in[x_{s-1},x_s]}\|f(x)\| =$$

$$= p\delta\frac{e^{3C}}{1 - p\delta e^C}(e^{2C} + 1)\max_{x\in[x_{s-1},x_s]}\|f(x)\|. \qquad (56)$$

Введем обозначение

$$\chi_s = \eta_f^{(s)} + \Big(X(x_s, x_{s-1}) - \widetilde{X}(x_s, x_{s-1})\Big)\,\bar{z}_f(x_{s-1}).$$

Очевидно, что

$$\|\chi_s\| \le \delta + p\delta\|\bar{z}_f(x_{s-1})\| \le \delta(1 + pK_{sol}^{[c]}),$$

так как

$$\|\bar{z}_f(x_{s-1})\| \le \max_{x\in[x_0,x_m]} \|u^{[c]}(x)\| \le K_{sol}^{[c]}. \tag{57}$$

Рассмотрим на интервале $[x_{s-1}, x_s]$ функцию

$$f_2(x) = \widetilde{X}(x, x_{s-1})[\widetilde{X}(x_s, x_{s-1})]^{-1}\chi_s/(x_s - x_{s-1}).$$

На основе проведенных выше оценок, получаем неравенство

$$\|f_2(x)\| \le (e^C + p\delta)\frac{e^C}{1 - p\delta e^C}\frac{\delta(1 + pK_{sol}^{[c]})}{h}. \tag{58}$$

Тогда выражение (55) можно переписать в виде

$$y_f^{[c]}(x_s) = \widetilde{X}(x_s, x_{s-1})\bar{z}_f(x_{s-1}) + \widetilde{X}(x_s, x_{s-1})\int\limits_{x_{s-1}}^{x_s} [\widetilde{X}(\xi, x_{s-1})]^{-1}\widetilde{f}(\xi)d\xi,$$

где

$$\widetilde{f}(\xi) = f_1(\xi) + f_2(\xi).$$

Это означает, что вектор $y_f^{[c]}(x_s)$ является значением решения задачи Коши:

$$\frac{dy_f^{[c]}(x)}{dx} = \widetilde{A}(x)\bar{y}_f(x) + \widetilde{f}(x), \quad \bar{y}_f(x_{s-1}) = \bar{z}_f(x_{s-1})$$

в точке x_s. Отсюда, исходя из (56), (58), (35) имеем оценку

$$\max_{x\in[x_{s-1},x_s]} \|f_{lin}(x) - \widetilde{f}(x)\| \le \max_{x\in[x_{s-1},x_s]} \|f_{lin}(x) - f_1(x)\| + \max_{x\in[x_{s-1},x_s]} \|f_2(x)\| \le$$

$$\le \frac{\delta}{1 - p\delta e^C}\left(pe^{3C}(e^{2C} + 1)\max_{x\in[x_{s-1},x_s]} \|f(x)\| + \frac{1}{h}e^C(e^C + p\delta)(1 + pK_{sol}^{[c]})\right) = \epsilon_F, \tag{59}$$

где δ определяется по формуле (50). Очевидно, последняя оценка может быть сделана только при условии

$$p\delta e^C < 1. \tag{60}$$

Таким образом, обратный анализ позволяет нам рассматривать вычисленные векторы как точные решения серии задач Коши со слегка возмущенными матрицей $\widetilde{A}$ (53) и правой частью $\widetilde{f}$ (59).

Резюме. В результате прямой прогонки в каждой точке x_s будут вычислены матрицы $Z_s^{[c]}$ и $\Omega_s^{[c]}$. Согласно Алгоритму 2.2 решение задачи в каждой точке x_s представимо в виде линейной комбинации столбцов матрицы $Z_s^{[c]} = [\hat{Z}_s, z_f(x_s)]$, где $\hat{Z}_s = [z_1(x_s), \ldots, z_p(x_s)]$.

Вычисленные верхне-треугольные матрицы $\Omega_s^{[c]}$ имеют специальный вид

$$\Omega_s^{[c]} = \begin{bmatrix} \Theta_s^{[c]} & \theta_s^{[c]} \\ 0 & 1 \end{bmatrix}. \tag{61}$$

Если из правого условия и при обратной прогонке будут найдены векторы $\bar{\alpha}_s$ (черта сверху означает, что величины не являются точными) такие, что

$$[R\bar{\hat{Z}}_m]\alpha_m = \psi - R\bar{z}_f(x_s), \qquad \bar{\Theta}_s\bar{\alpha}_{s-1} = \alpha_s - \bar{\theta}_s,$$

то функция $\widetilde{u}(x_s) = \bar{\hat{Z}}_s\bar{\alpha}_s + \bar{z}_f(x_s)$ в точках x_s совпадает с решением системы

$$\begin{cases} \dfrac{d\widetilde{u}}{dx} = \widetilde{A}(x)\widetilde{u}(x) + \widetilde{f}(x), \\ \\ L\widetilde{u}(x_0) = \varphi, \qquad R\widetilde{u}(x_m) = \psi, \end{cases} \tag{62}$$

которое согласно Теореме 1 отличается от u_{lin} на величину

$$\|u_{lin}(x_s) - \widetilde{u}(x_s)\| \le \max\{\epsilon_A, \epsilon_F\}K_{lin}(2+d)(1+\max\|\widetilde{u}(x_s)\|). \tag{63}$$

4.7 Правое граничное условие

Мы уже оценили влияние вычислительных и алгоритмических погрешностей прямой прогонки на решение, используя метод обратного анализа погрешностей. Влияние погрешностей обработки правого граничного условия и обратной прогонки мы в отличие от [9] оценим напрямую.

Обратная прогонка начинается с того, что $u(x_m)$ представляется в виде

$$u(x_m) = \bar{z}_f(x_m) + \bar{\hat{Z}}(x_m)\bar{\alpha}_m,$$

а вектор-коэффициент $\bar{\alpha}_m$ находятся из правого граничного условия (6):

$$[R\bar{\hat{Z}}(x_m)]\bar{\alpha}_m = \psi - R\bar{z}_f(x_m).$$

В действительности же, с учетом погрешностей, решение ищется в виде

$$u^{[c]}(x_m) = z_f^{[c]}(x_m) + \hat{Z}^{[c]}(x_m)\alpha_m^{[c]} \tag{64}$$

так, что

$$[R\hat{Z}^{[c]}(x_m)]\alpha_m^{[c]} = \psi - Rz_f^{[c]}(x_m). \tag{65}$$

Очевидно, что погрешность решения в точке x_m есть

$$u(x_m) - u^{[c]}(x_m) = \bar{z}_f(x_m) + \bar{\hat{Z}}(x_m)\bar{\alpha}_m - z_f^{[c]}(x_m) - \hat{Z}^{[c]}(x_m)\alpha_m^{[c]}.$$

Преобразуем скобки в последнем выражении к удобному виду

$$u(x_m) - u^{[c]}(x_m) = \left(\bar{z}_f(x_m) - z_f^{[c]}(x_m)\right) + \bar{\hat{Z}}(x_m)(\bar{\alpha}_m - \alpha_m^{[c]}) +$$

$$+ \left(\bar{\hat{Z}}(x_m) - \hat{Z}^{[c]}(x_m)\right)\alpha_m^{[c]}. \tag{66}$$

Следовательно

$$\|u(x_m) - u^{[c]}(x_m)\| \leq$$

$$\leq \|\bar{z}_f(x_m) - z_f^{[c]}(x_m)\| + \|\bar{\hat{Z}}(x_m)\|\,\|\bar{\alpha}_m - \alpha_m^{[c]}\| + \|\bar{\hat{Z}}(x_m) - \hat{Z}^{[c]}(x_m)\|\,\|\alpha_m^{[c]}\|. \tag{67}$$

Оценим каждое слагаемое в отдельности.

Погрешности$\|\bar{z}_f(x_m) - z_f^{[c]}(x_m)\|$ и $\|\bar{\hat{Z}}(x_m) - \hat{Z}^{[c]}(x_m)\|$ возникают в процессе ортогонализации и оцениваются по формулам (28) и (31).

Далее, из представления (64)следует, что

$$\|\alpha_m^{[c]}\| \leq \|u^{[c]}(x_m)\|.$$

Осталось оценить погрешность вычисления $\alpha^{[c]}$. Так как вектор $\alpha_m^{[c]}$ является решением системы алгебраических уравнений (65), то можно применить следующую оценку. Если есть две системы линейных алгебраических уравнений: исходная $Ax = f$ и возмущенная $(A + \Delta A)(x + \Delta x) = f + \Delta f$, то возмущение решения оценивается по формуле

$$\|\Delta x\| \leq (\|\Delta f\| + \|\Delta A\|\,\|x\|)\,\|(A + \Delta A)^{-1}\| \leq$$

$$\leq (\|\Delta f\| + \|\Delta A\|\,\|x\|)\frac{\|A^{-1}\|}{1 - \|\Delta A\|\|A^{-1}\|}. \tag{68}$$

Рассмотрим вспомогательную краевую задачу

$$\begin{cases} \dfrac{du^R}{dx} = A(x)u^R, \\ \\ Lu^R(x_0) = 0, \quad Ru^R(x_m) = \psi \end{cases}.$$

Структура решения этой задачи $u^R(x_m) = \hat{Z}^R(x_m)\alpha_m^R$, так как компонента $z_f^R = 0$ ввиду отсутствия правых частей в уравнении и в левом краевом условии. При этом имеет место оценка $\|u^R(x)\| \le K\|\psi\|$, где K как и ранее означает оценку нормы матрицы Грина. Так как вследствие ортогональности матрицы $\hat{Z}^R(x_m)$ имеет место неравенство

$$\|u^R(x_m)\| = \|\hat{Z}^R(x_m)\alpha_m^R\| = \|\alpha_m^R\| \le K\|\psi\|,$$

выполненное при любом векторе ψ, то $\| \left[R\hat{Z}^R(x_m)\right]^{-1} \| \le K$. А по построению решение задачи с произвольной правой частью

$$\begin{cases} \dfrac{du}{dx} = A(x)u + f(x), \\ \\ Lu^R(x_0) = \varphi, \quad Ru^R(x_m) = \psi \end{cases}$$

в точке x_m имеет структуру $u(x_m) = \hat{Z}(x_m)\alpha_m + z_f(x_m)$, причем $\hat{Z}(x_m) = \hat{Z}^R(x_m)$, то $\| \left[R\hat{Z}(x_m)\right]^{-1} \| \le K$.

Рассмотрим теперь возмущенную задачу

$$\begin{cases} \dfrac{d\widetilde{u}^R}{dx} = \widetilde{A}(x)\widetilde{u}^R, \\ \\ L\widetilde{u}^R(x_0) = 0, \quad R\widetilde{u}^R(x_m) = \psi. \end{cases}$$

Воспользуемся фактом, что если $\mathcal{L}$ – некий обратимый линейный оператор, то

$$\|(\mathcal{L} + \Delta\mathcal{L})^{-1}\| \le \frac{\|\mathcal{L}^{-1}\|}{1 - \|\mathcal{L}^{-1}\|\|\Delta\mathcal{L}\|}$$

при достаточно малом возмущении $\Delta\mathcal{L}$: $\|\mathcal{L}^{-1}\|\|\Delta\mathcal{L}\| < 1$. В нашей ситуации $\mathcal{L}$ это дифференциальный оператор задачи и норма обратного оператора оценивается через норму функции Грина $\|\mathcal{L}^{-1}\| \le K(2+d)$. Для возмущения оператора используем оценку $\|\Delta\mathcal{L}\| = \max\|A(x) - \widetilde{A}(x)\| \le \epsilon_A$

и после подстановки получаем

$$\|(\mathcal{L}+\Delta\mathcal{L})^{-1}\| \le \frac{K(2+d)}{1-\epsilon_A K(2+d)}.$$

Возвращаясь к решению возмущенной задачи и используя полученный результат, приходим к оценке

$$\|\widetilde{u}^R\| \le \frac{K(2+d)}{1-\epsilon_A K(2+d)}\|\psi\|,$$

а следовательно,

$$\|(R\bar{\hat{Z}}(x_m))^{-1}\| \le \frac{K(2+d)}{1-\epsilon_A K(2+d)}, \qquad \|\bar{\alpha}_m\| \le \frac{K(2+d)}{1-\epsilon_A K(2+d)}\|\psi\|.$$

Так как имеет место неравенство

$$\|R\hat{Z}^{[c]}(x_m) - R\bar{\hat{Z}}(x_m)\| = \|R(\hat{Z}^{[c]}(x_m) - \bar{\hat{Z}}(x_m))\| \le p\epsilon_z\|R\| = p\epsilon_z\|R\bar{\hat{Z}}(x_m)\|,$$

то на основании оценки (4.7) получаем в точке x_m

$$\|\bar{\alpha}_m - \alpha_m^{(c)}\| \le (\epsilon_F + p\epsilon_z\|\bar{\alpha}_m\|)\frac{K(2+d)\|R\|\|\psi\|}{1-K(2+d)(\epsilon_A+p\epsilon_z\|R\|\|\psi\|)} = \epsilon_\alpha \quad (69)$$

при условии

$$K(2+d)(\epsilon_A + p\epsilon_z\|R\|\|\psi\|) < 1.$$

При выводе оценки (69) мы не учли погрешность с которой $\alpha_m^{[c]}$ определяется при численном решении системы (65) на компьютере, а также погрешность перемножения матриц R и $\hat{Z}^{[c]}(x_m)$ размера $p\times n$, $n\times p$ и погрешность вычисления правой части. Чтобы сделать это нужно сравнить решения следующих систем

$$[R\otimes\hat{Z}^{[c]}(x_m)]\otimes\alpha_m^{[c]} = \psi\ominus R\otimes z_f^{[c]}(x_m),$$

$$[R\otimes\hat{Z}^{[c]}(x_m)]\cdot\hat{\alpha}_m = \psi\ominus R\otimes z_f^{[c]}(x_m),$$

$$[R\hat{Z}^{[c]}(x_m)]\cdot\widetilde{\alpha}_m = \psi - Rz_f^{[c]}(x_m),$$

$$[R\bar{\hat{Z}}(x_m)]\cdot\bar{\alpha}_m = \psi - R\bar{z}_f(x_m)$$

(кружочками обозначены машинные операции, см. [3]) и оценить выражение

$$\|\alpha_m^{[c]} - \bar{\alpha}_m\| \le \|\alpha_m^{[c]} - \hat{\alpha}_m\| + \|\hat{\alpha}_m - \widetilde{\alpha}_m\| + \|\widetilde{\alpha}_m - \bar{\alpha}_m\|. \quad (70)$$

Сначала рассмотрим второе слагаемое. Для простоты введем обозначения

$$W = R\hat{Z}^{[c]}(x_m), \quad x = \widetilde{\alpha}_m, \quad w = \psi - Rz_f^{[c]}(x_m)$$

и запишем систему (65) в виде

$$Wx = w. \tag{71}$$

Погрешность машинного (компьютерного) вычисления правой части равна (см. формулы (2.37), (2.40) в [3])

$$\|\triangle w\| = \|[\psi \ominus R \otimes z_f^{[c]}(x_m)] - [\psi - Rz_f^{[c]}(x_m)]\|$$

$$\leq 2\varepsilon_1\|\psi\| + [2\varepsilon_1(1+\delta_0) + \delta_0]\|R\|K_{sol}^{[c]},$$

где $\delta_0 = (n+1)\sqrt{p}\varepsilon_1$ и $\|\triangle w\|/\|w\| \leq \delta_1$. Погрешность машинного перемножения двух матриц A, B размера $p \times n$, $n \times p$ определяем используя формулу (2.40) из [3]:

$$\|(A \otimes B) - AB\| \leq (n+1)p\varepsilon_1\|A\|\|B\|.$$

Таким образом, матрица W вычислена с погрешностью

$$\frac{\|W_{\text{маш}} - W\|}{\|W\|} \leq (n+1)p\varepsilon_1 = \delta_2$$

и в действительности решается система

$$[W + \triangle W](x + \triangle x) = w + \triangle w,$$

из которой x определяется с погрешностью (см. формулу (8.33) в [3]):

$$\frac{\|\triangle x\|}{\|x\|} \leq (\delta_2 + \delta_1)\frac{\mu(W)}{1 - \delta_2\mu(W)} = \delta_3,$$

где $\mu(W)$ – число обусловленности матрицы W. Отсюда для второго слагаемого в (70) имеем

$$\frac{\|\hat{\alpha}_m - \widetilde{\alpha}_m\|}{\|\widetilde{\alpha}_m\|} \leq \delta_3 \quad \text{или} \quad \|\hat{\alpha}_m - \widetilde{\alpha}_m\| \leq \frac{\delta_3}{1+\delta_3}\|\hat{\alpha}_m\|.$$

Первое слагаемое в (70) оценивает погрешность решения на компьютере системы с квадратной матрицей W размера n (см. формулу (8.7) и таблицу 8.1 в [3]), которая при упрощении равна

$$\frac{\|\alpha_m^{[c]} - \hat{\alpha}_m\|}{\|\hat{\alpha}_m\|} \leq \widetilde{\epsilon}_P(p)\mu(W)[3(p+1)\sqrt{p}+1] + \widetilde{\epsilon}_P(p)(p+1) = \delta_4,$$

где $\widetilde{\epsilon}_P(p)$ определяется по формуле (26). Следовательно,

$$\|\alpha_m^{[c]} - \hat{\alpha}_m\| \le \frac{\delta_4}{1+\delta_4}\|\alpha_m^{[c]}\|.$$

Третье слагаемое в (70) есть ϵ_α (69). Складывая все оценки, окончательно получаем

$$\|\bar{\alpha}_m - \alpha_m^{[c]}\| \le \frac{(1+\delta_3)\delta_4 + \delta_3}{(1+\delta_3)(1+\delta_4)}\|\alpha_m^{[c]}\| + \epsilon_\alpha = \epsilon_\alpha(x_m).$$

Отсюда, используя (67), имеем оценку погрешности решения в точке x_m:

$$\|u(x_m) - u^{[c]}(x_m)\| \le \epsilon_{zf}^{(m)} + \epsilon_\alpha(x_m) + \epsilon_z(1+\epsilon_\alpha(x_m)) = \epsilon_{U_m}. \qquad (72)$$

4.8 Обратный анализ погрешностей обратной прогонки

При обратной прогонке особое влияние оказывают погрешности вычисления матриц Ω_s (см. (32)). На интервале $[x_{s-1}, x_s]$ вместо решения систем (8)

$$\bar{\Omega}_s\bar{\beta}_{s-1} = \bar{\beta}_s$$

будет получено решение системы

$$\Omega_s^{[c]}\beta_{s-1}^{[c]} = \beta_s^{[c]}.$$

Оценку погрешностей удобнее проводить в терминах подматриц (см. (61))

$$\Theta_s^{[c]}\alpha_{s-1}^{[c]} = \alpha_s^{[c]} - \theta_s^{[c]}.$$

Тогда

$$\|\alpha_{s-1}^{[c]} - \bar{\alpha}_{s-1}\| \le$$

$$\le \Big(\|\alpha_{s-1}^{[c]} - \bar{\alpha}_{s-1}\| + \|\theta_s^{[c]} - \bar{\theta}_s\| + \|\Theta_s^{[c]} - \bar{\Theta}_s\|\|\bar{\alpha}_{s-1}\|\Big)\frac{\|\bar{\Theta}_s^{-1}\|}{1 - \|\Theta_s^{[c]} - \bar{\Theta}_s\|\|\bar{\Theta}_s^{-1}\|}.$$

Так как

$$\|\Omega_s^{[c]} - \bar{\Omega}_s\| \le \epsilon_z\|\bar{\Omega}_s\| \le \epsilon_z\|\widetilde{X}(x_{s-1}, x_s)\| \le \epsilon_z(e^C + p\delta),$$

то

$$\|\Theta_s^{[c]} - \bar{\Theta}_s\| \le \epsilon_z(e^C + p\delta),$$
$$\|\theta_s^{[c]} - \bar{\theta}_s\| \le \epsilon_z(e^C + p\delta).$$

С учетом неравенств

$$\|\bar{\Theta}_s^{-1}\| \le \frac{e^C}{1 - p\delta e^C} \qquad \text{и} \qquad \|\bar{\alpha}_{s-1}\| \le \|\widetilde{u}(x_{s-1})\|$$

имеем

$$\|\alpha_{s-1}^{[c]} - \bar{\alpha}_{s-1}\| \le$$
$$\le \Big(\|\alpha_s^{[c]} - \bar{\alpha}_s\| + \epsilon_z(e^C + p\delta(1 + K_{sol})\Big)\frac{e^C}{1 - (1+\epsilon_z)p\delta - \epsilon_z e^C}.$$

Отсюда по индукции получаем оценку

$$\|\alpha_{s-1}^{[c]} - \bar{\alpha}_{s-1}\| \le \|\alpha_m^{[c]} - \bar{\alpha}_m\| C_\alpha^{m-s} + p\delta(1+K_{sol})\frac{C_\alpha^{m-s} - 1}{C_\alpha - 1} = \epsilon_\alpha(x_{s-1}), \quad (73)$$

где

$$C_\alpha = \frac{e^C}{1 - (1+\epsilon_z)p\delta - \epsilon_z e^C}$$

при условии

$$(1+\epsilon_z)p\delta - \epsilon_z e^C < 1.$$

Решение в точке x_s определяется по формуле

$$u^{[c]}(x_s) = z_f^{[c]}(x_s) + Z_s^{[c]}\alpha_s^{[c]}.$$

Используя имеющиеся оценки (28), (73), (31), нетрудно вывести неравенство

$$\|u^{[c]}(x_s) - \widetilde{u}(x_s)\| \le \|Z_s^{[c]} - \bar{Z}_s\|\|\bar{\alpha}_s\| + \|\bar{Z}_s\|\|\alpha_s^{[c]} - \bar{\alpha}_s\| + \|\bar{z}_f(x_s) - z_f^{[c]}(x_s)\| \le$$
$$\le \epsilon_z K_{sol} + \epsilon_\alpha(x_s) + \epsilon_{z_f}^{(s)} \quad (74)$$

– погрешность вносимая в решение задачи при обратной прогонки.

4.9 Оценка погрешности решения

Таким образом, в результате вычислений по Алгоритму 2.2 в каждой точке x_s будет получено решение $u^{[c]}(x_s)$ краевой задачи близкой к исходной (1). Для определения погрешности решения необходимо собрать все полученные оценки (74),(63), (36). При этом нужно, чтобы были выполнены условия (например, типа (60)) при которых они выводились.

Окончательная оценка погрешности решения в точке x_s имеет вид

$$\|u^{[c]}(x_s)-u(x_s)\| \le \|u^{[c]}(x_s)-\widetilde{u}(x_s)\|+\|\widetilde{u}(x_s)-u_{lin}(x_s)\|+\|u_{lin}(x_s)-u(x_s)\| \le$$

$$\le \epsilon_z K^{[c]}_{sol} + \epsilon_\alpha(x_s) + \epsilon^{(s)}_{z_f} + \varepsilon\mu, \tag{75}$$

где

$$\varepsilon = \max\{\epsilon_A, \epsilon_F\} + \frac{h^2}{2}K_{dif} \qquad \mu = K_{lin}(2+d)(1+K^{[c]}_{sol}).$$

Здесь K_{lin} – оценка норм матриц Грина задачи (33), ϵ_A находится из (53), ϵ_f – из (59), ϵ_z – из (28), $\epsilon^{(s)}_{z_f}$ – из (31), $\epsilon_\alpha(x_s)$ – из (73), K_{dif} – из (34), $h = (x_0 - x_m)/m = x_s - x_{s-1}$ – длина интервалов на которые разбит отрезок $[x_0, x_m]$, а $K^{[c]}_{sol}$ оценивается из (57) как

$$\max_{x\in[x_0,x_m]} \|u^{[c]}(x)\| \le K^{[c]}_{sol}.$$

5 Примеры

5.1 Пример 1

Приведем простой модельный пример использования описанного выше метода численного решения краевой задачи. Рассмотрим следующую краеву задачу

$$\begin{aligned} &\frac{du}{dx} = A(x)u(x) + f(x), \\ &Lu(x_0) = \varphi, \qquad Ru(x_m) = \psi, \\ &x_0 \le x \le x_m, \quad x_0 = 0.0, \quad x_m = 1.0, \quad d = x_m - x_0 = 1.0, \end{aligned} \tag{76}$$

где

$$A = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}, \qquad f(x) = \begin{pmatrix} 0 \\ -2x \end{pmatrix},$$

$$L = (0\ 1), \qquad R = (1\ 0), \qquad \varphi = 1, \qquad \psi = 1.$$

Сравним точное решение

$$u = \begin{pmatrix} x \\ 1 \end{pmatrix}$$

с численным решением задачи методом ортогональной прогонки, полученным с помощью программы Sweep из пакета **GALA-2.1**.

Результаты вычислений. В результате вычислений мы получили, что реальная погрешность вектора решения равна

$$\max_{x\in[x_0,x_m]} \|u^{[c]}(x) - u(x)\| \le 3.28\,10^{-12}. \tag{77}$$

Оценка погрешности решения $u^{[c]}(x_s)$ в точке x_s определяется по формуле (75), где число обусловленности $\mu = K(2+d)(1+K_{sol}^{[c]}) = 23.18$ для данной задачи величина постоянная (оценка K для матриц Грина вычислена программой EstimGreenMat из пакета **GALA-2.1**), а остальные величины зависят от шага $h = d/m$ разбивки отрезка $[x_0, x_m]$ и от шага интегрирования $\Delta = h/N$ на каждом интервале $[x_s, x_{s+1}]$. Видно, что для получения приемлемой оценки необходимо, чтобы выполнялось и условие (15) Теоремы 1. Для этого пришлось значительно уменьшать Δ.

Например, оценка погрешности вычисленного решения равна

$$\max_{x\in[x_0,x_m]} \|u^{[c]}(x) - u(x)\| \le 7.5\,10^{-2} \tag{78}$$

при $m = 8$ и $N = 500$.

Видно, что даже в таком простом примере гарантированная погрешность (78) решения задачи оказалась значительно завышенной по сравнению с реальной (77). Это говорит о том, что оценка погрешности решения краевой задачи, основанная на Теореме 1 и на методике, описанной в работе [9], не является удовлетворительной с практической точки зрения. Видимо необходимо разрабатывать другой способ оценки погрешности решения краевой задачи (1), что мы и предполагаем сделать в дальнейшем.

5.2 Пример 2

Следующий пример – это компьютерное моделирование сердечно-сосудистой системы человека. Описанный алгоритм решения краевой задачи использовался для решения системы гемодинамики (одномерной

модели кровотока в сосуде) [4]. Мы рассматривали упрощенную гидродинамическую модель течения крови в сосудах. Ее математические свойства (гиперболичность) позволяют формулировать для нее корректные краевые задачи [4]. Одномерная модель выводится с помощью интегрирования уравнений Навье-Стокса по произвольному осевому сечению S (усреднением этих уравнений по поперечному направлению) и представляет собой систему из двух дифференциальных уравнений в частных производных (см. [4, 16, 15]):

$$\begin{aligned} \frac{\partial A}{\partial t} + \frac{\partial Q}{\partial z} &= 0, \\ \frac{\partial Q}{\partial t} + \alpha \frac{\partial}{\partial z}\left(\frac{Q^2}{A}\right) + \frac{A}{\rho}\frac{\partial p}{\partial p} + K_r\left(\frac{Q}{A}\right) &= 0, \end{aligned} \tag{79}$$

где A, Q – неизвестные. Здесь $A = A(t, z)$ – площадь поперечного сечения сосуда, $Q = Q(t, z)$ – поток массы, $p = p(t, z)$ – давление, $\alpha = \text{const}$ – коэффициент Кориолиса, $\rho = \text{const}$ – плотность, $K_r = \text{const}$ – коэффициент трения, t – время, z – осевая координата.

Методом прямых задача сводится к решению системы обыкновенных дифференциальных уравнений на каждом временном слое, которая затем решается описанным методом ортогональной прогонки. Эта математическая модель *благодаря* своей простоте позволяет моделировать кровоток во всем артериальном сосудистом русле (глобальное моделирование), т. е. для всех 55 артерий сосудистого дерева человека. При этом, *несмотря* на простоту модели она позволила провести моделирование основных параметров кровотока как в норме, так и в ряде патологий. **Результаты вычислений** подробно изложены в Главе 4 монографии [4] и дипломной работе [10].

Как уже отмечалось выше, метод ортогональной прогонки, допускает *контроль за накоплением вычислительной погрешности*. В своей работе мы делаем на этом особый акцент. Связано это с тем, что увеличение объемов вычислений при современном уровне математического моделирования приводит к усилению роли вычислительной погрешности. Но многие алгоритмы не позволяют оценить ее вклад в решение, вследствие чего результат математического моделирования принимает характер *гипотезы*, которая в ряде случаев может достаточно сильно отличаться от точного решения исходной задачи.

В данном случае такой неопределенный характер численного решения неприемлем по двум основным причинам. Во-первых, пользователями данного алгоритма являются биологи, физиологи и другие специалисты,

далекие от вычислительной математики. В силу своей специальности они не знают и не обязаны знать особенности внутреннего устройства вычислительных программ. Но при этом пользователи должны иметь возможность делать обоснованные выводы из результатов математического моделирования. Во-вторых, в данном случае моделируются процессы, связанные с жизнью и здоровьем человека. Следовательно, используемые алгоритмы должны вести себя предсказуемым и контролируемым образом.

Список литературы

[1] *Бибердорф Э.А., Попова Н.И.* Решение линейных систем с гарантированной оценкой точности результатов (часть первая). Новосибирск, 1999 (Препринт ИЯФ СО РАН; 99-49).

[2] *Бибердорф Э.А., Попова Н.И.* Вычисления с гарантированной оценкой точности (часть вторая). Решение спектральных задач. Новосибирск, 2001 (Препринт ИЯФ СО РАН; 2001-21).

[3] *Бибердорф Э.А., Попова Н.И.* Гарантированная точность современных алгоритмов линейной алгебры. Новосибирск: Издательство СО РАН, 2006.

[4] *Бибердорф Э.А., Блохин А.М., Попова Н.И., Трахинин Ю.Л.* Система кровообращения и артериальная гипертония: биофизические и генетико-физиологические механизмы, математическое и компьютерное моделирование. Глава 4: Глобальное моделирование артериальной системы человека. Новосибирск: Издательство СО РАН, 2008 (в печати).

[5] *Годунов С.К.* О численном решении краевых задач для систем линейных обыкновенных дифференциальных уравнений // Успехи мат. наук, 1961, т.16, N3, с.171-175.

[6] *Годунов С.К.* Обыкновенные дифференциальные уравнения с постоянными коэффициентами. Новосибирск: Издательство НУ, 1994.

[7] *Годунов С.К., Антонов А.Г., Кирилюк О.П., Костин В.И.* Гарантированная точность решения систем линейных уравнений в евклидовых пространствах. Новосибирск: Наука, 1992.

[8] *Годунов С.К.* Лекции по современным аспектам линейной алгебры. Новосибирск: Научная книга, 2002.

[9] *Кузнецов С.В.* Развитие метода ортогональной прогонки. Диссертация. Институт математики СО РАН, Новосибирск: 1988.

[10] *Леонова Т.И.* Построение, численное моделирование и анализ одномерной модели гемодинамики. Дипломная работа, механико-математический факультет НГУ, Новосибирск: 2008.

[11] *Уилкинсон Дж.Х.* Алгебраическая проблема собственных значений/ Пер. с англ. М.: Наука, 1970.

[12] *Форсайт Дж., Мальколъм М., Молер К.* Машинные методы математических вычислений/ Пер. с англ. М.: Мир, 1980.

[13] *Хаусхолдер А.С.* Основы численного анализа. М.: Изд-во иностр. лит., 1956.

[14] *Givens W.* Numerical Computation of the Characteristic Values of a Real Symmetric Matrix. Report ORNL 1574, Oak Ridge, Tenn., Oak Ridge National Laboratory, 1954.

[15] *Lamponi D.N.* One dimensional and multiscale models for blood flow circulation. Thesis, Lausanne, 2004.

[16] *Quarteroni A., Formaggia L.* Mathematical modelling and numerical simulation of the cardiovascular system. In: *Ciarlet P.G. (ed.)*. Handbook of numerical analysis, v.12, special volume: *Ayache N. (ed.)*. Computational Models for the Human Body. Amsterdam, Elsevier Science & Technology, 2003, p.3-129.

Э.А. Бибердорф, Н.И. Попова

Контроль точности решения краевой задачи методом ортогональной прогонки

E.A. Biberdorf, N.I. Popova

Accuracy control of solving the boundary-value problem by the orthogonal sweep method

ИЯФ 2009-1

Ответственный за выпуск А.В. Васильев
Работа поступила 30.12.2008 г.

Сдано в набор 11.01.2009 г.
Подписано в печать 12.01.2009 г.
Формат бумаги 60×90 1/16 Объем 3.0 печ.л., 2.4 уч.-изд.л.
Тираж 90 экз. Бесплатно. Заказ № 1

Обработано на IBM PC и отпечатано на
ротапринте ИЯФ им. Г.И. Будкера СО РАН
Новосибирск, 630090, пр. академика Лаврентьева, 11.