ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 69 за 2013 г.

Белов А.А., Калиткин Н.Н.

Эволюционная факторизация и сверхбыстрый счет на установление

Рекомендуемая форма библиографической ссылки: Белов А.А., Калиткин Н.Н. Эволюционная факторизация и сверхбыстрый счет на установление // Препринты ИПМ им. М.В.Келдыша. 2013. № 69. 36 с. URL: http://library.keldysh.ru/preprint.asp?id=2013-69

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
ИМ. М.В. КЕЛДЫША

А. А. Белов, Н. Н. Калиткин

ЭВОЛЮЦИОННАЯ ФАКТОРИЗАЦИЯ И СВЕРХБЫСТРЫЙ СЧЕТ НА УСТАНОВЛЕНИЕ

Москва, 2013

УДК 519.6

А. А. Белов, Н. Н. Калиткин. Эволюционная факторизация и сверхбыстрый счет на установление. Препринт Института прикладной математики им. М. В. Келдыша РАН, Москва, 2013.

При разностном решении многомерных эллиптических уравнений возникают системы линейных алгебраических уравнений с сильно разреженными матрицами огромной размерности. Их решают итерационными методами, сходящимися довольно медленно. Для прямоугольных сеток при непостоянных коэффициентах и шагах сеток предложен гораздо более быстрый метод. В случае разностных схем для параболических уравнений построен экономичный метод, названный эволюционной факторизацией. Для эллиптических уравнений предлагается счет на установление по эволюционно факторизованным схемам. Это итерационный метод, имеющий логарифмическую скорость сходимости. Предложены набор шагов, практически оптимизирующий сходимость этого алгоритма, и процедура упорядочивания шагов, напоминающая метод Ричардсона. Она позволяет получить апостериорную асимптотически точную оценку погрешности итерационного процесса. Ранее подобные оценки для итерационных процессов были неизвестны.

Ключевые слова: эволюционная факторизация, логарифмический счет на установление.

A. A. Belov, N.N. Kalitkin. Evolutional factorization and superfast relaxation count.

In finite-difference solution of multi-dimensional elliptic equations the systems of linear algebraic equations with strongly rarefied matrices of enormous sizes appear. They are solved by iteratonal methods with slow convergence. For rectangular nets, variable coefficients and net steps much more fast method is proposed. In case of finite difference schemes for parabolic equations an efficient method, called evolutional factorisation, is built. For elliptic equations relaxation count for evolutionally factorized schemes is proposed. This iterational method has logarithmic convergence. A set of steps, that practically optimizes the method's convergence, and Richardson-like procedure of steps regulation are proposed. The procedure delivers an a posteriori asymptotically precise estimation for the iterational process error. Such estimations were not known before.

Keywords: evolutional factorization, logarithmic relaxation count.

Работа поддержана грантом РФФИ 11-01-00102.

Введение

Решение многомерных эллиптических уравнений занимает важное место в задачах математической физики. Наиболее распространены разностные методы решения [1]. При этом сеточные решения удовлетворяют системе линейных алгебраических уравнений с сильно разреженными матрицами огромной размерности. Решение таких систем является нетривиальной проблемой. Ей посвящена обширная литература [2].

Решение эллиптических уравнений разностными методами повышенной точности имеет два основных аспекта. 1) Чтобы решение разностной задачи стремилось к решению дифференциального эллиптического уравнения, нужно применять процедуры сгущения сеток по Ричардсону. Для получения хороших точностей нужно сгущать сетки многократно, что приводит к системам большого порядка. 2) На каждой сетке сеточное решение должно быть найдено с точностью, достаточной для применения метода Ричардсона. Поэтому от итерационных методов требуется очень высокая точность при умеренном числе итераций.

На произвольных сетках получаются линейные системы достаточно общего вида. Для них работоспособны только итерационные методы сопряженных направлений [3]. Однако такие методы сходятся сравнительно медленно: число итераций зависит от границ спектра по закону $S \approx 10\sqrt{\lambda_{\max}/\lambda_{\min}} \approx 10N$, где N – среднее число узлов по каждой координате. Для хорошей аппроксимации требуются большие $N \sim 300 - 1000$, что приводит к неприемлемо большим S. Поэтому нужно искать ограничения, при которых возможно построение более быстрых, но достаточно общих методов. Этому посвящена данная работа.

Ограничимся эллиптическими задачами без смешанных производных. В частности, это может быть уравнение теплопроводности или уравнение электростатики. В первом случае коэффициент уравнения k имеет смысл коэффициента теплопроводности, во втором – диэлектрической проницаемости. При этом коэффициенты уравнений будем считать переменными и гладкими, а сетки – прямоугольными и неравномерными. Такая постановка является достаточно содержательной. В этом случае консервативная разностная схема для трехмерного уравнения имеет следующий вид:

$$(\Lambda_x + \Lambda_y + \Lambda_z)\, u = -f. \tag{1}$$

Здесь трехточечный оператор

$$(\Lambda_x u)_n = \frac{2}{h_{x,n+1/2} + h_{x,n-1/2}} \left[\frac{k_{x,n+1/2}}{h_{x,n+1/2}} (u_{n+1} - u_n) - \frac{k_{x,n-1/2}}{h_{x,n-1/2}} (u_n - u_{n-1}) \right], \tag{2}$$

$$h_{x,n+1/2} = x_{n+1} - x_n;$$

в (2) оставлен только индекс по координате x. Выражения для Λ_y и Λ_z аналогичны.

Для системы (1), (2) наиболее быстро сходящимся процессом является счет на установление с логарифмическим набором шагов [4]. Параметрами итерационного процесса являются шаги счета на установление $\tau_s \in [\tau_{\min}, \tau_{\max}]$. Хорошие результаты дает сетка, равномерная по $\ln \tau$. При этом высокую точность обеспечивает число итераций $S \approx 10 \ln (\lambda_{\max}/\lambda_{\min}) \approx 20 \ln N$. Это число остается умеренным даже для $N \sim 1000$.

В данной работе сделано следующее. 1) Приведен экономичный алгоритм решения многомерных задач теплопроводности. Решение эллиптических задач рассматривается как счет на установление по этому алгоритму. 2) Найдены оптимальные значения $\tau_{\min}$ и $\tau_{\max}$ для двумерных и трехмерных задач. 3) Предложен набор шагов, уменьшающий число итераций в полтора раза по сравнению с логарифмически равномерным. 4) Улучшены имеющиеся теоретические оценки скорости сходимости логарифмического счета на установление. 5) Построена процедура упорядочивания шагов τ_s, напоминающая метод Ричардсона для разностных сеток. Это позволило получить апостериорные асимптотически точные оценки сходимости итерационного процесса. Ранее подобные оценки не были известны. Существовали только мажорантные оценки точности по невязке, которые могли отличаться от точных на несколько порядков. 6) Построены конструктивные оценки границ спектра $\lambda_{\min}$ и $\lambda_{\max}$. 7) Все методы проиллюстрированы на представительных примерах.

1 Методы решения эллиптических уравнений

Приведем некоторые известные прямые и итерационные методы решения разностных уравнений [2], [5], [6].

1.1 Быстрое преобразование Фурье

Этот метод является самым быстрым из известных прямых методов. Он применим к задаче Дирихле в прямоугольнике (прямоугольном параллелепипеде), решаемой на равномерных сетках при постоянных коэффициентах k_α. Метод экономичен в том случае, если число узлов N_α является произведением малых целых чисел. Особенно он эффективен при $N_\alpha = 2^{r_\alpha}$. Условия применимости этого метода к задачам математической физики являются слишком жесткими, но в задачах обработки изображений он применяется широко.

Суть метода в следующем. Пусть в одномерной задаче $k = const$, $h = const$. Разностная схема (1)-(2) с учетом граничных условий $u_0 = u_N = f_0$ содержит N неизвестных. Будем искать решение в виде суммы Фурье с таким же числом членов:

$$u_n = \sum_{q=0}^{N-1} a_q w^{nq}, \quad w = \exp(2\pi i/N). \tag{3}$$

Подставим ее в уравнение, домножим его на w^{-np} и просуммируем по n от 0 до $N-1$ с учетом ортогональности гармоник. Получим коэффициент Фурье решения a_p, выраженный через коэффициент Фурье правой части b_p:

$$a_p = \frac{b_p h^2}{4\sin^2(\pi p/N)}, \; b_p = \frac{1}{N}\sum_{n=0}^{N-1} f_n w^{-np}, \; 0 \leqslant p \leqslant N-1. \tag{4}$$

Обобщение на многомерный случай очевидно.

Нетрудно видеть, что в общем случае эти формулы неэкономичны: для нахождения коэффициентов a_p и b_p требуется $2N^2$ операций. Объем вычислений сокращается, если число интервалов сетки составное ($N = KL$). Если $N = L^r$, то $b(p)$ можно находить по рекуррентным формулам. Это заметно снижает трудоемкость алгоритма. Так, для $L = 2$ требуемое число операций составляет $4N\log_2(N)$.

Заметим также, что для быстрого преобразования Фурье, как и для других прямых методов, не возникает вопроса о точности сходимости итераций. Имеются только ошибки округления, которые невелики, поскольку в случае эллиптических уравнений матрица хорошо обусловлена.

1.2 Нечетно-четная редукция

Рассмотрим одномерную краевую задачу Дирихле

$$-Y_{n-1} + CY_n - Y_{n+1} = F_n, \quad 1 \leqslant n \leqslant N-1, \tag{5}$$

$$Y_0 = F_0, \quad Y_N = F_N. \tag{6}$$

Идея метода состоит в последовательном исключении из уравнений (5)-(6) неизвестных Y_n сначала с нечетными номерами n, затем из остальных уравнений с номерами n, равными произведению 2 на нечетное число, затем 4 на нечетное число и т.д. Каждый шаг процесса исключения уменьшает число неизвестных, и если $N = 2^r$, то в результате процесса исключения останется одно уравнение, из которого можно найти $Y_{N/2}$. Обратный ход метода заключается в последовательном нахождении неизвестных Y_n сначала с номерами

n, кратными $N/4$, затем $N/8$, $N/16$ и т.д. В случае большей размерности задачи решение строится аналогично.

Очевидно, этот метод есть модификация метода исключения Гаусса, в котором исключение неизвестных происходит в специальном порядке. Применительно к разностным эллиптическим задачам метод нечетно-четной редукции имеет такую же скорость сходимости и область применимости, как быстрое преобразование Фурье.

1.3 Чебышевский набор шагов

Счет на установление можно проводить по явной схеме. Эта схема не требует факторизации, поэтому ее можно применять к более широкому классу задач (уравнение со смешанными производными, криволинейная граница, непрямоугольные сетки и т.д.). Схема единообразно пишется при любом числе измерений. Каждая итерация содержит одно умножение матрицы на вектор. Поэтому каждая итерация нетрудоемка, а метод легко распараллеливается. Однако у явной схемы лишь условная сходимость с сильным ограничением на шаг τ: для простейшей явной схемы $\tau = O(h^2)$. Поэтому для расчета с постоянным шагом потребуется $S = O(N^2)$ шагов. Такая трудоемкость неприемлема.

Чебышевский набор шагов τ_s обеспечивает установление за $S = O(N)$ шагов: точность ε будет достигнута после

$$S = \frac{\ln(1/\varepsilon)}{2}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}. \tag{7}$$

итераций. Величины $1/\tau_s$ являются корнями многочлена Чебышева 1-го рода степени S, построенного на отрезке $[\lambda_{\min}, \lambda_{\max}]$. Для произвольной области получить оценки $\lambda_{\min}$ и $\lambda_{\max}$ крайне трудно. Кроме того, требуемое число шагов S весьма чувствительно к точности этих оценок. Заметим также, что S нужно задавать еще до начала расчета, что неудобно.

Отметим еще одну особенность чебышевского набора. Для части шагов не выполнено условие устойчивости, и поэтому ошибки нарастают. На остальных шагах это условие выполнено, и ошибки затухают. В конечном итоге затухание должно преобладать, но на промежуточных шагах ошибка может оказаться в принципе большой и даже выйти за пределы представимых чисел.

Для преодоления этой трудности нужно выполнять шаги в определенном порядке, например: τ_1, τ_S; τ_2, τ_{S-1}; τ_3, τ_{S-2} и т.д. В каждой паре один шаг дает нарастание ошибки, другой – соответствующее гашение. Существуют и другие варианты группировки. Это несколько усложняет алгоритм, но не

устраняет его ключевые недостатки. Поэтому на практике применяются более совершенные итерационные методы сопряженных направлений.

1.4 Метод сопряженных градиентов

Этот метод заключается в построении полного ортогонального базиса, минимизирующего некоторую квадратичную форму $\Phi(u)$ в пространстве $u \in R_M$. На практике M настолько велико, что вычисления не успевают дойти до исчерпывания. Но уже умеренное число итераций может обеспечить нужную точность ε. В качестве критерия остановки процесса выбирают малость невязки. Для хорошо обусловленных задач это дает разумные результаты.

Метод сопряженных градиентов применим к эрмитовым знакоопределенным матрицам, то есть охватывает значительно более широкий класс задач, чем преобразование Фурье или метод редукции. При этом он лишен недостатков явных схем с чебышевским набором шагов. В частности, каждый шаг метода устойчив, сам метод не требует задания границ спектра и числа итераций S. Кроме того, метод имеет простую одношаговую форму записи и легко распараллеливается.

Скорость сходимости этого метода составляет $S = O(N)$. Точнее, для достижения точности ε требуется

$$S \approx \frac{N}{\pi} \ln \frac{1}{\varepsilon} \tag{8}$$

итераций. Такая скорость считается лучшей для известных общих методов, но при больших $N > 1000$ алгоритм становится слишком трудоемким.

Контролировать сходимость этого метода можно косвенно по невязке. Но оценка погрешности по невязке имеет мажорантный характер, причем константа в этой мажорантной оценке неизвестна и велика. Поэтому аккуратно оценить погрешность практически невозможно.

Возможно также построение попеременно-треугольной схемы для треугольных операторов Самарского S_1 и S_2, $\Lambda = S_1 + S_2$. На прямоугольных сетках она легко факторизуется и обращается. Однако при этом возникает дополнительный член невязки $\tau^2/4\, S_1 S_2 \partial u/\partial t$. Даже в простейшей одномерной задаче это приводит к условной аппроксимации $O(\tau^2/h^2)$. Поэтому схема непригодна для расчета нестационарных задач.

При счете на установление для эллиптических задач таких проблем не возникает. Наилучшей стратегией здесь оказывается расчет с постоянным оптимальным шагом по времени. Однако этот процесс дает лишь такую же сходимость, как для метода сопряженных градиентов. При этом, в отличие от последнего, он требует знание границ спектра, что неудобно. Поэтому попеременно-треугольная схема представляется неперспективной.

Предлагаемый в данной работе счет логарифмический на установление по эволюционно факторизованной схеме обеспечивает сходимость принципиально нового класса $S \sim \ln \lambda_{\max}/\lambda_{\min}$ при достаточно содержательной постановке задачи, данной во введении. Кроме того, он позволяет построить апостериорную асимптотически точную оценку погрешности.

2 Эволюционная факторизация

Рассмотрим параболическое уравнение $u_t = Lu + f$ для пространственного оператора $L = \sum L_\alpha$, расщепляющегося на одномерные операторы. Для решения этой задачи можно написать схему "с полусуммой":

$$\frac{\hat{u} - u}{\tau} = \sum_\alpha \Lambda_\alpha \frac{\hat{u} + u}{2} + f, \tag{9}$$

где $\hat{u}$ есть решение на новом временном слое. Эта схема безусловно устойчива и имеет аппроксимацию $O\left(\tau^2 + \sum h_\alpha^2\right)$. Однако схема (9) приводит к решению системы с ленточной матрицей $\sim N^3$, у которой ширина ленты $\sim N$ в двумерном и $\sim N^2$ в трехмерном случаях, что чрезмерно трудоемко. Поэтому она неэкономична.

Двумерные задачи успешно решаются известной схемой переменных направлений Писмена-Рэкфорда. Однако эта схема не обобщается на трехмерный случай. Кроме того, для получения второго порядка точности нужна нетривиальная форма аппроксимации граничных условий на промежуточном шаге.

Для трехмерных задач созданы различные методы приближенной факторизации [7] и локально-одномерные методы [8]. В них также трудно получить второй порядок точности как из-за проблемы написания граничных условий на промежуточных шагах, так и из-за необходимости симметризовать порядок выполнения промежуточных шагов.

Эти недостатки удается преодолеть в методе эволюционной факторизации [4]. Заменим в уравнении (9) полусумму $0.5\left(\hat{u} + u\right) = u + 0.5\tau\left(\hat{u} - u\right)/\tau$ и перенесем слагаемые, содержащие производную по времени, в левую часть:

$$\left(E - \frac{\tau}{2}\sum_\alpha \Lambda_\alpha\right)\frac{\hat{u} - u}{\tau} = \sum_\alpha \Lambda_\alpha u + f. \tag{10}$$

Схема (10) эквивалентна схеме (9) и является неэкономичной.

Приближенно факторизуя оператор в левой части (10), получим новую схему, которую назовем ***эволюционно-факторизованной***:

$$\prod_\alpha \left(E - \frac{\tau}{2}\Lambda_\alpha\right)\frac{\hat{u} - u}{\tau} = \sum_\alpha \Lambda_\alpha u + f. \tag{11}$$

Исследуем свойства этой схемы.

2.1 Алгоритм

Рассмотрим наиболее сложный трехмерный случай. Введем вспомогательные сеточные функции v и w и перепишем (11) в виде трех уравнений:

$$(E - \tau\Lambda_x/2)\, w = (\Lambda_x + \Lambda_y + \Lambda_z)\, u + f, \tag{12}$$

$$(E - \tau\Lambda_y/2)\, v = w, \tag{13}$$

$$(E - \tau\Lambda_z/2)\, (\hat{u} - u)\, /\tau = v. \tag{14}$$

Правая часть (12) известна, поскольку вычисляется на предыдущем слое. Оператор в левой части этого уравнения является трехдиагональным по направлению x, а потому обращается одномерной прогонкой при каждых фиксированных сеточных y и z. Аналогично трехдиагональным в направлении y является оператор в левой части (13). Поэтому его можно обратить одномерной прогонкой при каждых фиксированных сеточных x и z и по вычисленному w найти v. Подставляя это значение v в правую часть (14), одномерной прогонкой по z при каждых фиксированных сеточных x и y можно найти $(\hat{u} - u)\, /\tau$. Таким образом, нахождение $\hat{u}$ сводится к последовательности трех одномерных прогонок. Это означает, что схема экономична.

Переход к двумерному случаю осуществляется вычеркиванием (14) и заменой v на $(\hat{u} - u)\, /\tau$ в (13). Заметим, что если в схеме Писмена-Рэкфорда исключить промежуточный слой, то она в точности совпадет с двумерной эволюционно-факторизованной схемой.

2.2 Аппроксимация

Вычтем схему (11) из схемы (10). Главный член разности есть $\tau^2/4 \sum \Lambda_\alpha \Lambda_\beta\, (\hat{u} - u)\, /\tau \approx \tau^2/4 \sum \Lambda_\alpha \Lambda_\beta u_t = O\left(\tau^2\right)$. Но схема (9) имеет аппроксимацию $O\left(\tau^2 + \sum h_\alpha^2\right)$. Следовательно, схема (11) также аппроксимирует дифференциальное уравнение с порядком $O\left(\tau^2 + \sum h_\alpha^2\right)$.

2.3 Граничные условия

Условия первого рода для прогонки по направлению z задаются тривиально, поскольку значения u на границе полагаются известными во все моменты времени. Для прогонки по направлению y граничные условия выразим из (14):

$$[v]_{\text{гран.}} = [(E - \tau\Lambda_z/2)\, (\hat{u} - u)\, /\tau]_{\text{гран.}} \approx$$
$$\approx [(E - \tau/2 \cdot \partial/\partial z\, (k_z \partial/\partial z))\, (\hat{u} - u)\, /\tau]_{\text{гран.}}\,. \tag{15}$$

В последнем переходе разностное дифференцирование заменено второй производной с точностью $O\left(h_z^2\right)$. Дифференцирование граничного условия нужно выполнять точно. Условие (15) вносит дополнительную погрешность $O\left(\tau h_z^2\right)$ третьего порядка малости, которой можно пренебречь.

Граничные условия для w получаются из (13) с учетом (14):

$$[w]_{\text{гран.}} = \left[\left(E - \tau\Lambda_y/2\right)\left(E - \tau\Lambda_z/2\right)\left(\hat{u} - u\right)/\tau\right]_{\text{гран.}} \approx$$
$$\approx \left[\left(E - \tau\Lambda_y/2 - \tau\Lambda_z/2\right)\left(\hat{u} - u\right)/\tau\right]_{\text{гран.}} \approx$$
$$\approx \left[\left(E - \tau/2 \cdot \partial/\partial y\left(k_y \partial/\partial y\right) - \tau/2 \cdot \partial/\partial z\left(k_z \partial/\partial z\right)\right)\left(\hat{u} - u\right)/\tau\right]_{\text{гран.}}. \qquad (16)$$

Здесь отброшен член $\tau^2\Lambda_y\Lambda_z/4$ порядка $O\left(\tau^2\right)$, разностное дифференцирование заменено точным. Очевидно, это не ухудшает порядок аппроксимации всей схемы. Таким образом, эволюционно-факторизованная схема с описанными граничными условиями обеспечивает аппроксимацию $O\left(\tau^2 + \sum h_\alpha^2\right)$.

Для двумерного случая граничные условия выписываются аналогично. Это показывает, что граничные условия для схемы Писмена-Рекфорда лучше писать без использования промежуточного слоя.

2.4 Устойчивость

Одномерные операторы $\Lambda_\alpha < 0$, но в качестве их собственных значений нам удобно выбрать положительные величины λ_α. Тогда множитель роста трехмерной гармоники в (11) имеет вид

$$\left(1 + \tau\lambda_x/2\right)\left(1 + \tau\lambda_y/2\right)\left(1 + \tau\lambda_z/2\right)\left(\rho - 1\right) = -\tau\left(\lambda_x + \lambda_y + \lambda_z\right). \qquad (17)$$

Из (17) получаем двумерный случай, полагая $\lambda_z = 0$. В одномерном случае надо также полагать $\lambda_y = 0$. Отсюда одномерные, двумерные и трехмерные множители роста имеют вид соответственно

$$\rho = \frac{1 - \tau\lambda_x/2}{1 + \tau\lambda_x/2}, \qquad (18)$$

$$\rho = \frac{1 - \tau\lambda_x/2}{1 + \tau\lambda_x/2}\frac{1 - \tau\lambda_y/2}{1 + \tau\lambda_y/2}, \qquad (19)$$

$$\rho = 1 - \frac{\tau\left(\lambda_x + \lambda_y + \lambda_z\right)}{\left(1 + \tau\lambda_x/2\right)\left(1 + \tau\lambda_y/2\right)\left(1 + \tau\lambda_z/2\right)}. \qquad (20)$$

Во всех случаях $|\rho| < 1$. Для (18) и (19) это очевидно, для (20) легко проверяется. Это обеспечивает безусловную устойчивость схемы (11) при любом числе измерений. Однако ее асимптотическая устойчивость является лишь условной. Это непосредственно видно для одномерного и двумерного случая по структуре множителей роста, которые таковы же, как для схемы "с полусуммой". Для трехмерного случая это нетрудно доказать.

3 Двойная факторизация

Отметим еще один способ факторизации, называемый ***двойной факторизацией*** [9]. Умножим обе части (9) на τ. Перенесем все слагаемые, содержащие $\hat{u}$, в левую часть, а все слагаемые, содержащие u, – в правую. Приближенно факторизуем операторы в обеих частях получившегося уравнения

$$\prod_{\alpha} (E - \tau\Lambda_\alpha/2)\,\hat{u} = \prod_{\alpha} (E + \tau\Lambda_\alpha/2)\,u + \tau f. \tag{21}$$

Доказательство экономичности, аппроксимации с порядком $O\left(\tau^2 + \sum h_\alpha^2\right)$ и построение граничных условий для этой схемы аналогичны схеме эволюционной факторизации.

Нетрудно убедиться, что для произвольного числа измерений множитель роста многомерной гармоники в точности равен произведению одномерных множителей

$$\rho = \prod_{\alpha} \frac{(E - \tau\lambda_\alpha/2)}{(E + \tau\lambda_\alpha/2)}, \tag{22}$$

что представляется заманчивым. В частности, из (22) следует безусловная устойчивость и условная асимптотическая устойчивость.

Нетрудно заметить, что в двумерном случае эволюционная и двойная факторизации точно совпадают. В трехмерном случае они различаются. В частности, схема двойной факторизации не подходит для счета на установление. В этом случае в пределе $\hat{u} = u$, и схема принимает вид

$$\left(\Lambda_x + \Lambda_y + \Lambda_z + \tau^2\Lambda_x\Lambda_y\Lambda_z/4\right) u + f = 0, \tag{23}$$

что означает лишь условную аппроксимацию исходного эллиптического уравнения. Перспективных приложений для этой схемы пока не найдено.

4 Счет на установление

4.1 Стационарное решение

Эллиптическое уравнение $Lu = -f$ можно рассматривать как стационарный предел при $t \to \infty$ для параболического уравнения $u_t = Lu + f$ со стационарными граничными условиями и правой частью. Для разностного уравнения (1) это означает решение эволюционной задачи по схеме (9) при выполнении условия асимптотической устойчивости. Стационарный предел решения задачи (9) есть решение системы (1). Решение эллиптического уравнения вычислением этого стационарного предела называют счетом на установление.

При счете на установление выбирают произвольные начальные данные. Для расчета берут какой-либо факторизованный вариант схемы (9) и считают до тех пор, пока левая часть факторизованной схемы не станет достаточно малой. Число необходимых для этого итераций существенно зависит от того, насколько удачно выбран набор шагов по времени. Поэтому актуальной является задача о построении набора, который обеспечивал бы наиболее быструю сходимость.

4.2 Оптимальный набор шагов

Для явных схем оптимальным является чебышевский набор шагов, но построенный не для τ, а для величины $1/\tau$. Эта схема лишь условно устойчива, поэтому не любые перестановки шагов допустимы. Для неявной продольно-поперечной схемы был найден постоянный оптимальный шаг $\tau_0 \approx \approx \sqrt{\tau_{\min}\tau_{\max}}$ [5]. Обобщение этой идеи дано в [4], где для набора шагов употребляется преобразование $\ln\tau$. Такой выбор обеспечивает логарифмическую скорость сходимости, которая, по-видимому, является наибольшей из достижимых скоростей.

Ранее логарифмической сходимостью обладали только узкоспециализированные методы вроде быстрого преобразования Фурье, но они применимы только для тепличных условий: постоянные k, h и специфические числа узлов 2^N. Область применимости логарифмического набора, описанная во введении, значительно шире.

Вопрос о построении оптимального логарифмического набора состоит из двух частей: 1) построение границ этого набора для разного числа измерений и 2) выбор порождающей функции.

5 Логарифмические наборы

5.1 Границы логарифмического набора

Оценим граничные шаги $\tau_{\min}$, $\tau_{\max}$ для разного числа измерений. Естественно выбрать их так, чтобы они максимально гасили множители роста, соответствующие границам спектрального интервала. Из формул множителей роста видно, что в одномерном случае

$$\tau_{\min} = 2/\lambda_{x,\max}, \quad \tau_{\max} = 2/\lambda_{x,\min}, \tag{24}$$

а в двумерном

$$\tau_{\min} = 2/\max\left\{\lambda_{x,\max}, \lambda_{y,\max}\right\}, \quad \tau_{\max} = 2/\min\left\{\lambda_{x,\min}, \lambda_{y,\min}\right\}. \tag{25}$$

При этом граничные множители роста точно равны нулю.

В трехмерном случае легко проверяется следующее. При $\tau > 0$ зависимость $\rho(\tau)$ имеет единственный минимум, удовлетворяющий приведенному кубическому уравнению

$$(2/\tau)^3 - b\,(2/\tau) - 2c = 0, \tag{26}$$

где $b = \lambda_x\lambda_y + \lambda_x\lambda_z + \lambda_y\lambda_z$, $c = \lambda_x\lambda_y\lambda_z$. Искомый (вещественный положительный) корень вычисляется явно по тригонометрическому варианту формулы Кардано:

$$\frac{2}{\tau} = -2\sqrt{\frac{b}{3}}\cos\left(\varphi + \frac{2\pi}{3}\right), \quad \varphi = \frac{1}{3}\arccos\left(-c\,(b/3)^{-3/2}\right). \tag{27}$$

Подставим в (26) $\lambda_{\alpha,\max}$. Если для полученного корня τ_* $\rho(\tau_*) \geqslant 0$, то положим $\tau_{\min} = \tau_*$ (см. рис. 1, а). Если $\rho(\tau_*) < 0$ (см. рис. 1, б), то множитель

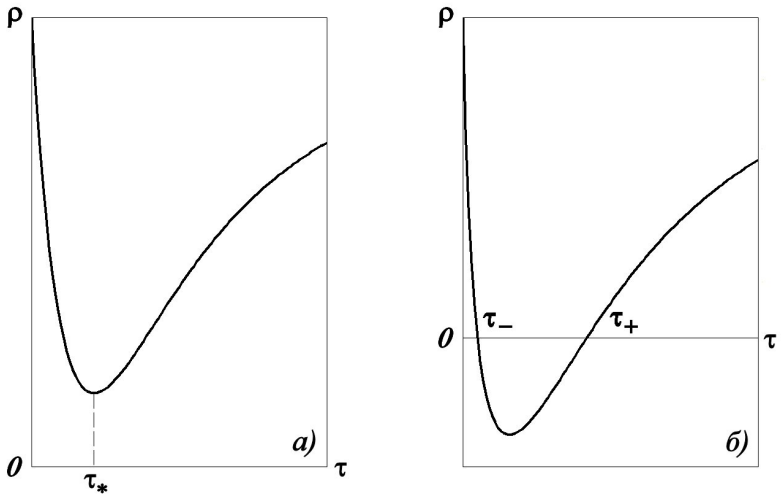

Рис. 1: Множитель роста трехмерной гармоники

роста имеет два вещественных положительных корня $0 < \tau_- < \tau_+$ (третий корень отрицателен). Эти корни удовлетворяют кубическому уравнению

$$(2/\tau)^3 - a\,(2/\tau)^2 + b\,(2/\tau) + c = 0, \tag{28}$$

где $a = \lambda_x + \lambda_y + \lambda_z$, и находятся чисто вещественными вычислениями по тригонометрическому варианту формулы Кардано:

$$\frac{2}{\tau_\pm} = \frac{a}{3} - 2\sqrt{\frac{f}{3}}\cos\left(\theta \mp \frac{2\pi}{3}\right), \quad \theta = \frac{1}{3}\arccos\left(g/2\,(f/3)^{-3/2}\right), \tag{29}$$

где

$$f = 1/2\,(\lambda_x - \lambda_y)^2 + 1/2\,(\lambda_x - \lambda_z)^2 + 1/2\,(\lambda_y - \lambda_z)^2, \tag{30}$$

$$g = \lambda_x^2(-2\lambda_x + 3\lambda_y + 3\lambda_z)/27 + \lambda_y^2(-2\lambda_y + 3\lambda_z + 3\lambda_x)/27 + \\ +\lambda_z^2(-2\lambda_z + 3\lambda_x + 3\lambda_y)/27 + 14/9\,\lambda_x\lambda_y\lambda_z. \qquad (31)$$

В этом случае положим $\tau_{\min} = \tau_-$. Аналогично выглядит процедура вычисления $\tau_{\max}$ по значениям $\lambda_{\alpha,\min}$. Если $\rho(\tau_*) \geqslant 0$, то аналогия полная. Если $\rho(\tau_*) < 0$, то следует положить $\tau_{\max} = \tau_+$.

Отметим частные случаи трехмерной задачи. Если $\lambda_x = \lambda_y = \lambda_z$, получим $\tau_{\min} = 1/\lambda_{\max}$, $\tau_{\max} = 1/\lambda_{\min}$. В этом случае для крайних гармоник $\rho = 1/9$, что обеспечивает хорошее убывание погрешности за один шаг. Если k_α/h_α^2 по разным направлениям многократно отличаются, то $\rho(\tau_*) < 0$ как для $\tau_{\min}$, так и для $\tau_{\max}$. Заметим также, что если $\rho(\tau_*) < 0$, то соответствующие гармоники гасятся лучше, чем при $\rho(\tau_*) > 0$, и следует ожидать более быстрой сходимости счета на установление.

5.2 Оценки границ спектра

Для построения границ логарифмического набора нужны оценки границ спектрального интервала. В одномерном случае имеет место

Лемма 1 *Для произвольных k, h*

$$\lambda_{\max} \leqslant \xi = 4 \max \frac{1}{h_{n+1/2} + h_{n-1/2}} \left(\frac{k_{n+1/2}}{h_{n+1/2}} + \frac{k_{n-1/2}}{h_{n-1/2}} \right). \qquad (32)$$

Доказательство. При $k, h = const$ эта оценка очевидна. В общем случае в силу неравенства треугольника из (2) следует

$$\|\lambda u\| = |\lambda| \cdot \|u\| \leqslant \\ \leqslant 2\|u\| \max \frac{2}{h_{n+1/2} + h_{n-1/2}} \left(\frac{k_{n+1/2}}{h_{n+1/2}} - \frac{k_{n-1/2}}{h_{n-1/2}} \right). \qquad (33)$$

После сокращения на $\|u\|$ получим утверждение леммы. ■

Оценка (32) без труда обобщается на многомерный случай, причем k_α может зависеть не только от x_α, но и от других переменных. Справедлива

Лемма 2 *Для произвольных k_α, h_α*

$$\lambda_{\max} \leqslant 4 \sum_\alpha \max_{x,y,z} \frac{2}{h_{\alpha,n+1/2} + h_{\alpha,n-1/2}} \left(\frac{k_{\alpha,n+1/2}}{h_{\alpha,n+1/2}} + \frac{k_{\alpha,n-1/2}}{h_{\alpha,n-1/2}} \right) \qquad (34)$$

Доказательство дословно повторяет доказательство леммы 1. ■

На квазиравномерных сетках $\xi/\lambda_{\max} \to 1$ при $N \to \infty$. Для иллюстрации возьмем пример с сильно пульсирующим коэффициентом теплопроводности и сильно неравномерной сеткой (см. рис. 2):

$$k(x) = 1 - 0.9 \sin^2 2\pi x, \quad x \in [0, 1]; \qquad (35)$$

$$h(x) = (25 + 20\cos 20x) \,/\, (N+1)\,\psi_1, \quad \psi_1 = 25 + \sin 20 \approx 25.31 \tag{36}$$

В этом случае поведение оценки (32) приведено на рис. 3. Видно, что наи-

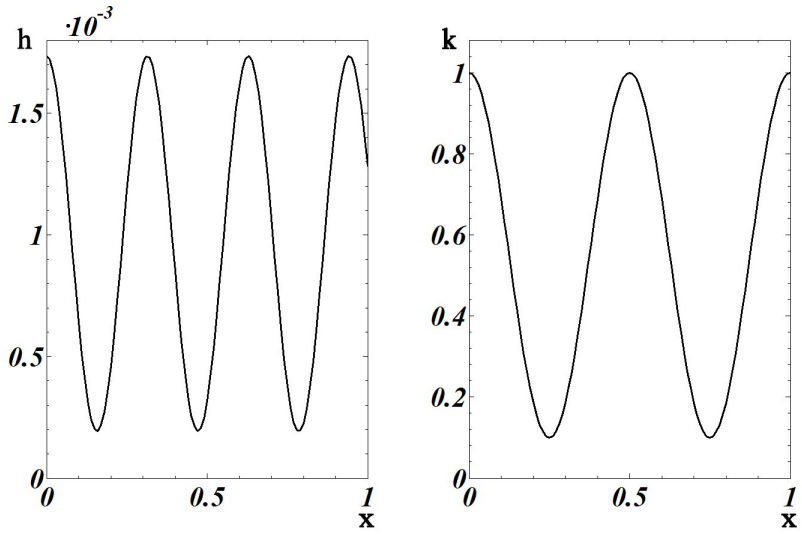

Рис. 2: Сильно неоднородная среда (35), (36)

большее отличие ξ от $\lambda_{\max}$ составляет 14%, что является приемлемым. Далее отношение $\xi/\lambda_{\max}$ быстро стремится к единице.

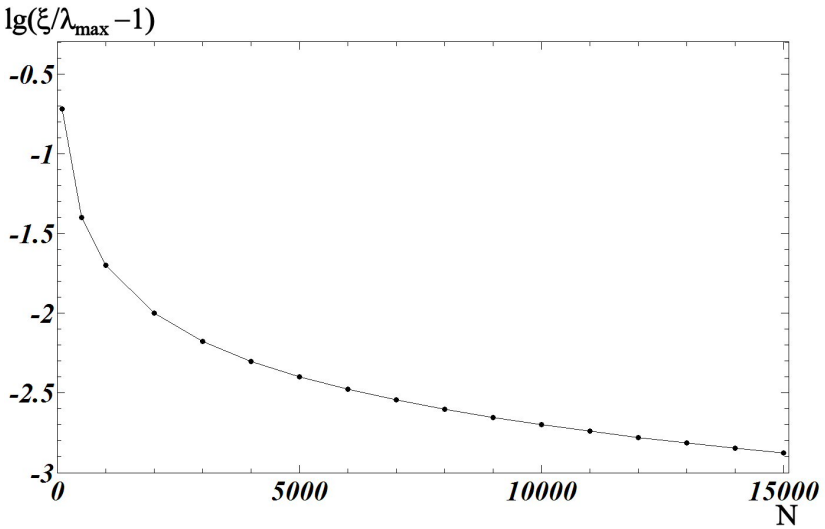

Рис. 3: Сильно неоднородная среда (35)-(36). Оценка (32) для $\lambda_{\max}$.

Для $\lambda_{\min}$ можно построить оценку снизу. Нетрудно доказывается

Лемма 3 *Для произвольного непрерывного k*

$$\lambda_{\min} \geqslant \frac{\pi^2}{l^2} \min k, \tag{37}$$

где l – длина отрезка.

Доказательство проведем, исходя из дифференциального представления:

$$\lambda u = -\frac{d}{dx}\left(k\frac{du}{dx}\right). \tag{38}$$

Домножим (38) скалярно на u и проинтегрируем по частям:

$$\lambda(u,u) = \left(k\frac{du}{dx}, \frac{du}{dx}\right). \tag{39}$$

Применяя теорему о среднем, вынесем $k(x^*)$, $x^* \in [0,l]$ за знак интеграла, разделим обе части (39) на (u,u) и перейдем к минимуму по u:

$$\lambda \geqslant k(x^*) \min_u \frac{\int (du/dx)^2\, dx}{\int u^2 dx} = k(x^*)\tilde{\lambda}, \tag{40}$$

где $\tilde{\lambda} = \pi^2/l^2$ – наименьшее собственное значение простейшей задачи. Поскольку k предполагается непрерывным, то $k(x^*) \geqslant \min k$, что завершает доказательство леммы. ■

На практике оценка (37) может оказаться не очень точной. Однако ее можно взять в качестве нулевого приближения в методе обратных итераций с переменным сдвигом. Тогда уже первая итерация дает хорошую точность (поскольку нижние собственные значения хорошо разнесены). Так, в примере (35)-(36) отличие первой итерации от $\lambda_{\min}$ не превышало 5%.

Сходимость метода обратных итераций с переменным сдвигом проиллюстрирована на примере простейшей задачи ($k = const$, $h = const$) при $N = 1000$. На графике (см. рис. 4) отложена величина $\lg(1 - \lambda_j/\lambda_{\min})$ в зависимости от номера итерации j. Видно, что уже третья итерация дает точность, сравнимую с фоном ошибок округления.

Лемма 3 допускает обобщение на многомерный случай. Именно, справедливо следующее утверждение:

Лемма 4 *Для произвольных непрерывных* k_α

$$\lambda_{\min} \geqslant \sum_\alpha \frac{\pi^2}{l_\alpha^2} \min_{x,y,z} k_\alpha, \tag{41}$$

где l_α *– длина отрезка по направлению* α.

Доказательство проводится по той же схеме, что и для леммы 3. ■

Для практического применения можно дать следующую рекомендацию. Нужно вычислить первые шаги метода обратных итераций с переменным сдвигом для оператора Λ_x при каждых фиксированных сеточных y, z. В качестве начального приближения выбрать $\lambda_x^{(0)}(y_n, z_m) = \pi^2/l_x^2 \min_x k_x(x, y_n, z_m)$.

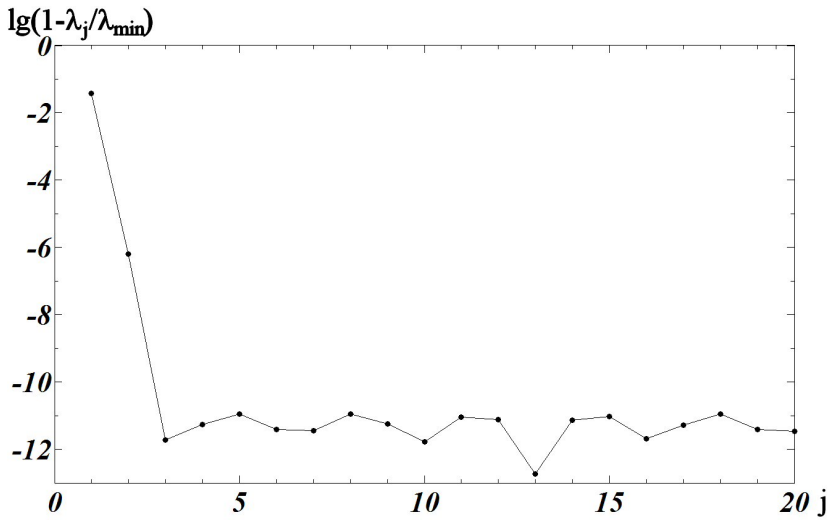

Рис. 4: Сходимость метода обратных итераций с переменным сдвигом.

После этого взять наименьший $\tilde{\xi}_x$ из полученных результатов. Провести аналогичную процедуру с операторами Λ_y при всех фиксированных сеточных x, z (наименьший результат $\tilde{\xi}_y$) и Λ_z при всех фиксированных сеточных x, y (наименьший результат $\tilde{\xi}_z$). Оценкой для наименьшего собственного значения будет величина $\tilde{\xi}_x + \tilde{\xi}_y + \tilde{\xi}_z$.

5.3 Порождающая функция

В работе [4] рассмотрено несколько вариантов логарифмической сетки. Общий вид такого набора можно записать как

$$\ln \tau_s = 1/2 \ln (\tau_{\max} \tau_{\min}) + 1/2 \ln (\tau_{\max}/\tau_{\min}) f(s), \quad 0 \leqslant s \leqslant S. \tag{42}$$

где порождающая функция $f(s) \in [-1, 1]$ является монотонной и нечетной. В [4] были рассмотрены следующие наборы: равномерный

$$f_{\text{р}}(s) = 2s/S - 1, \tag{43}$$

чебышевский

$$f_{\text{ч}}(s) = -\cos \pi s/S, \tag{44}$$

и интерполяционный

$$f_{\text{и}}(s) = \theta_s \left(1 + (1 - \theta_s)/2r\right)^r, \ \theta_s = 2s/S - 1, \ r = \left(1 + 1/8 \ln^2 (\lambda_{\max}/\lambda_{\min})\right)^{-1}. \tag{45}$$

Равномерный набор плохо подавлял граничные гармоники, а чебышевский – центральные. Интерполяционный примерно одинаково подавлял все гармоники, что обеспечивало лучшую точность. Однако он имеет громоздкий и непрозрачный вид.

Чтобы скомпенсировать указанные недостатки равномерного и чебышевского наборов, возьмем их линейную комбинацию с некоторым весом C:

$$f^{C}_{\text{ЛТ}}(s) = C f_{\text{р}}(s) + (1 - C) f_{\text{ч}}(s). \tag{46}$$

Выражение (46) представляет собой однопараметрическое семейство наборов, где параметр C определяет соотношение шагов в центре и на краях интервала. Для практического применения удобнее универсальный и простой набор, поэтому нужно выбрать некоторое фиксированное значение C.

Приведенные далее теоретические соображения позволяют рекомендовать значение

$$C = \pi / (\pi + 2). \tag{47}$$

Набор (46)-(47) будем называть ***линейно-тригонометрическим (ЛТ)***.

Было проведено большое количество численных расчетов для набора (46) с различными значениями константы C. На график выводились огибающие функции

$$R(\xi) = \lg \left| \prod_{s=1}^{S} \rho(\xi) \right|, \quad \xi = \frac{1}{2} \ln \lambda \tag{48}$$

для разных N и разных S. Примеры таких графиков для значения (47) приведены на рис. 5. Значения S подбирались так, чтобы во всех трех случаях $N = 100, 1000, 10000$ получать примерно одинаковые точности.

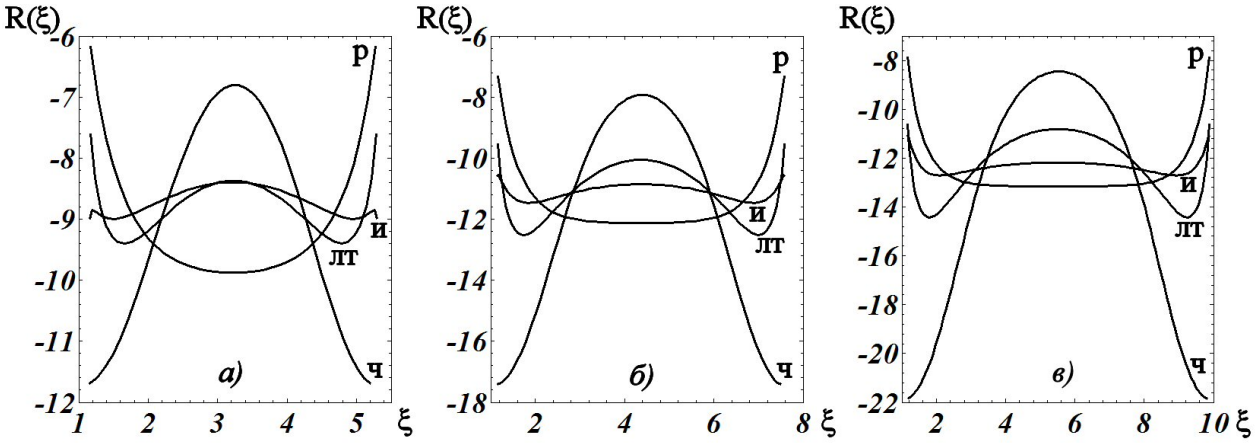

Рис. 5: Огибающие функции (48): а) $N = 100$, $S = 40$, б) $N = 1000$, $S = 75$, в) $N = 10000$, $S = 110$; **р** – равномерный набор, **ч** – чебышевский набор, **и** – интерполяционный набор, **лт** – линейно-тригонометрический набор.

Видно, что набор (46)-(47) обеспечивает примерно одинаковое подавление всех гармоник и дает лучшую сходимость. Соответствующие точности почти не уступают набору (45) и значительно превосходят (43) и (44) (см. также Табл. 1). При этом формулы являются простыми и прозрачными.

Заметим также, что качественное поведение набора (45) зависит от N: на рис. 5 *в)* края расположены выше, чем центр, а на рис. 5 *а)* – наоборот. Значит, этот набор может быть непредсказуемым, и его применение нежелательно. Качественное поведение набора (46)-(47) практически не меняется: края расположены примерно на том же уровне, что и центр.

N	удовлетворительная точность	хорошая точность	отличная точность
100	30	40	50
	-4.78	-6.16	-7.53
	-5.08	-6.81	-8.54
	-6.22	-8.40	-10.57
	-5.87	-7.60	-9.31
1000	55	75	95
	-5.54	-7.31	-9.05
	-5.77	-7.93	-10.10
	-7.88	-10.55	-13.19
	-7.20	-9.53	-11.84
10000	80	110	140
	-5.90	-7.84	-9.76
	-6.08	-8.45	-10.82
	-8.19	-11.06	-13.92
	-7.78	-10.59	-13.23

Таблица 1. Сходимость итераций. Числа в клетках: первое – S, остальные –логарифмы максимальной погрешности гармоник для наборов: равномерного (43), чебышевского (44), интерполяционного (45) и ЛТ (46)-(47).

6 Априорные оценки точности

Построим теоретические оценки сходимости, задавая набор $\{\tau_s\}$ в логарифмической шкале. Получим мажорантные оценки, которые являются почти строгими.

6.1 Средние гармоники

Пусть сетка равномерна по $\ln\tau$. Ее шаг $\delta = \ln\tau_s - \ln\tau_{s-1} = const$. Пусть число шагов достаточно велико. Пусть гармоника λ_k удалена от обоих краев спектрального интервала. Это означает, что с обеих сторон от нее лежит много шагов.

Если для одного шага случайно $\tau_s = 2/\lambda_k$, то $\rho = 0$ и гашение наилучшее. Гашение будет наихудшим, если величина $\ln(2/\lambda_k)$ равноудалена от двух соседних шагов $\ln \tau_s$. В этом случае

$$\ln \frac{\tau_s \lambda_k}{2} = \pm\frac{\delta}{2}, \pm\frac{3\delta}{2}, \pm\frac{5\delta}{2}. \tag{49}$$

Получим приближенную величину шага δ в этом наихудшем случае.

Разобьем набор $\{\tau_s\}$ на пары шагов τ_s, $\tau_{s'}$, расположенных левее и правее $2/\lambda_k$. Потребуем, чтобы гармоника λ_k после s-го шага гасилась в ε раз, т.е.

$$\prod_s |\rho(\tau_s \lambda_k)| = \varepsilon. \tag{50}$$

Множителей много, и они убывают достаточно быстро от центрального. Поэтому можно взять бесконечные пределы произведения. Тогда

$$\ln \varepsilon = \sum \ln |\rho(\tau\lambda)| = 2 \sum_{\tau\lambda/2>1} \ln |\rho(\tau\lambda)|. \tag{51}$$

Подставляя разложение

$$\begin{gathered} \ln|\rho| = \ln(1-\zeta) - \ln(1+\zeta) = \\ = -2\left(\zeta + 1/3\zeta^3 + 1/5\zeta^5 \ldots\right), \quad \zeta = 2/\tau_s\lambda_k < 1 \end{gathered} \tag{52}$$

в (51) и группируя члены, получим

$$\begin{gathered} \ln 1/\varepsilon = 4\left(e^{-\delta} + e^{-3\delta} + e^{-5\delta} + \ldots\right) + 4/3\left(e^{-3\delta} + e^{-9\delta} + e^{-15\delta} + \ldots\right) + \ldots = \\ = 2\left(\operatorname{sh}(\delta/2)\right)^{-1} + 2/3\left(\operatorname{sh}(3\delta/2)\right)^{-1} + \ldots \approx 4/\delta \sum_{n=0}^{\infty} (2n+1)^{-2} = \pi^2/2\delta. \end{gathered} \tag{53}$$

Отсюда нетрудно выразить δ через ε.

6.2 Крайние гармоники

Гармоники, расположенные по краям, гасятся шагами только с одной стороны, т.е. в 2 раза слабее, чем средние гармоники. Этим и объясняется указанный ранее недостаток равномерного набора. Значит, для получения гарантированной точности нужно подставлять в (53) не δ, а $\delta/2$. Поскольку для равномерной сетки $\delta = 1/S \ln \lambda_{\max}/\lambda_{\min}$, то

$$S = 4/\pi^2 \ln \lambda_{\max}/\lambda_{\min} \ln 1/\varepsilon \approx 0.4 \ln \lambda_{\max}/\lambda_{\min} \ln 1/\varepsilon. \tag{54}$$

Поскольку наши оценки были мажорантными, число итераций (54) заведомо достаточно для получения точности ε.

6.3 Неравномерная сетка

Пусть задан некоторый набор шагов, неравномерный в логарифмической сетке, и пусть число шагов S велико. Тогда каждый шаг эффективно гасит близкие гармоники и слабо – удаленные. Поэтому для фиксированной гармоники можно использовать приближение локальной равномерности набора шагов и оценку (54), где δ есть локальная величина шага. Для совокупности всех гармоник нужно подставить в (54) величину $\max \delta$.

Подчеркнем, это приближение хорошо работает лишь при больших S, т.е. для расчетов с высокой точностью. Для начальных S возникают нерегулярности.

6.4 Оценка для ЛТ-набора

Чтобы граничные гармоники гасились так же, как центральные, шаги на границах должны быть в 2 раза короче, чем шаги в центре. Применяя это условие к набору (46), нетрудно получить значение веса (47).

Из (53) нетрудно получить оценку гашения центральных гармоник ЛТ-набором, умножив это выражение на отношение шагов равномерного и ЛТ-наборов в центре интервала:

$$S = 4/\left(\pi^2 + 2\pi\right) \ln \lambda_{\max}/\lambda_{\min} \ln 1/\varepsilon \approx 0.25 \ln \lambda_{\max}/\lambda_{\min} \ln 1/\varepsilon. \qquad (55)$$

Для граничных гармоник в силу вдвое большей плотности шагов получается такая же оценка. Для промежутков между краями и центром доказать эту оценку не удается, но она хорошо оправдывается на практике.

6.5 Многомерный случай

Полученные результаты непосредственно переносятся на двумерный случай, поскольку двумерный множитель роста есть произведение одномерных. В трехмерном случае применимость этих результатов неочевидна, но проведенные вычисления показывают, что ЛТ-набор работает хорошо.

7 Расчеты

7.1 Графики точности

Было проведено сравнение разных наборов для одномерного случая. Выбиралась задача с известным точным сеточным решением (при $k = const$, $h = const$, $N = 1000$). Видно, что сетка весьма густая. Точное решение и

начальное приближение полагались гладкими: $u = x^2$, $u^{(0)} = 0$. Заметим, что выбор заведомо плохого нулевого приближения – набора псевдослучайных чисел – практически не ухудшает сходимости.

На рис. 6 представлены зависимости погрешности ε от S для различных наборов шагов. Это монотонно убывающие линии. При достаточно больших S

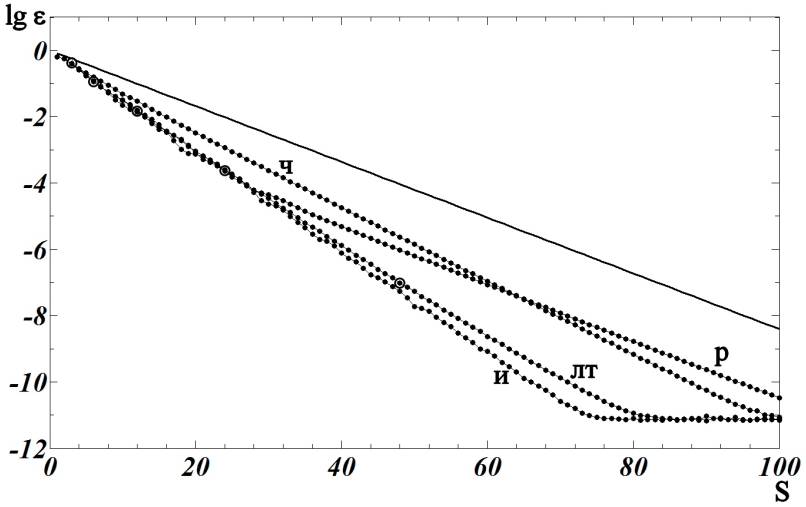

Рис. 6: Погрешность в одномерном случае; $N = 1000$, $k = 1$, $h = (N+1)^{-1}$; прямая – оценка (54); $\bullet$ – численные расчеты, обозначения наборов – см. рис. 5; $\circ$ – оценка (63)

они переходят в постоянный фон, обусловленный ошибками округления (расчеты производились с 64-разрядными числами). Видно, что чебышевский и равномерный наборы проигрывают по точности. Интерполяционный и ЛТ-наборы имеют почти одинаковую точность, но кривая ЛТ-набора практически прямолинейна. Далее мы увидим, что это позволяет получить апостериорные оценки точности. Поэтому все дальнейшие расчеты проводились с ЛТ-набором.

Участки с $S < 25$ следует назвать нерегулярными. На них еще не все кривые выходят на асимптотические режимы. При этом для ЛТ-набора наклон в регулярной области почти не отличается от приблизительного наклона в нерегулярной, что дополнительно свидетельствует о преимуществах этого набора.

На рис. 7 приведена зависимость погрешностей ε от S для ЛТ-набора при различных N. Видно, что линии очень быстро выходят на регулярный прямолинейный участок, который затем резко (почти изломом) переходит в горизонтальный фон. Этот фон тем выше, чем больше N, что объясняется ухудшением обусловленности линейной системы [10], [11]. Наклоны регулярных участков убывают с увеличением N примерно в соответствии с (55), хотя эта оценка является нестрогой.

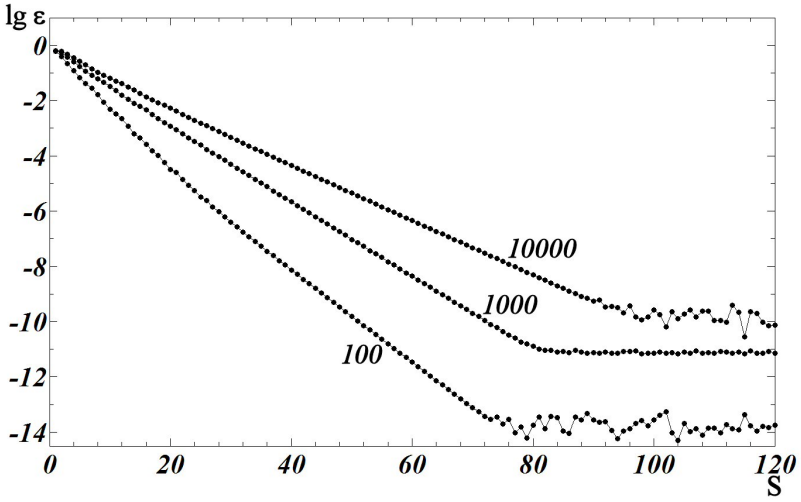

Рис. 7: Погрешность в одномерном случае; k, h – см. рис. 6; • – численные расчеты по ЛТ-набору; цифры около линий – значения N.

7.2 Теоретические оценки

Было проведено сравнение погрешностей равномерного и ЛТ-наборов с соответствующими теоретическими оценками. Оценки оказались хорошими. В приведенном примере коэффициенты наклона регулярных участков равномерного и ЛТ-наборов отличаются от оценок (54) и (55) не более, чем на 1%.

Отметим, что эти оценки определяют скорость сходимости в регулярной области и не учитывают нерегулярную область. Теоретическая оценка является прямой, исходящей из начала координат. Расчетная кривая может содержать нерегулярный участок. Для равномерного набора этот участок с крутым наклоном довольно велик. Поэтому для него оценка (54) лежит на два порядка выше регулярного участка. Для ЛТ-набора нерегулярный участок мал, и для него оценка (55) лежит выше регулярного участка примерно на половину порядка.

7.3 Трудные примеры

Приведем несколько представительных примеров: две задачи с сильно переменными средами и задачу в неограниченной области.

1) Рассмотрим задачу с сильно пульсирующим коэффициентом теплопроводности (35) и сильно неравномерной сеткой (36). Даже на достаточно подробной сетке ($N = 1000$) у кривых сходимостей равномерного и интерполяционного наборов вообще не наблюдается регулярных участков (см. рис. 8).

2) Рассмотрим следующую задачу с неоднородными $k(x)$ и $h(x)$:

$$k(x) = 0.1 + \pi/2 + \operatorname{arctg} 50(x - 1/2), \quad x \in [0, 1]; \tag{56}$$

$$h(x) = 3e^{3x}/\left(N+1\right)\psi_2, \quad \psi_2 = e^3 - 1 \approx 19.09. \tag{57}$$

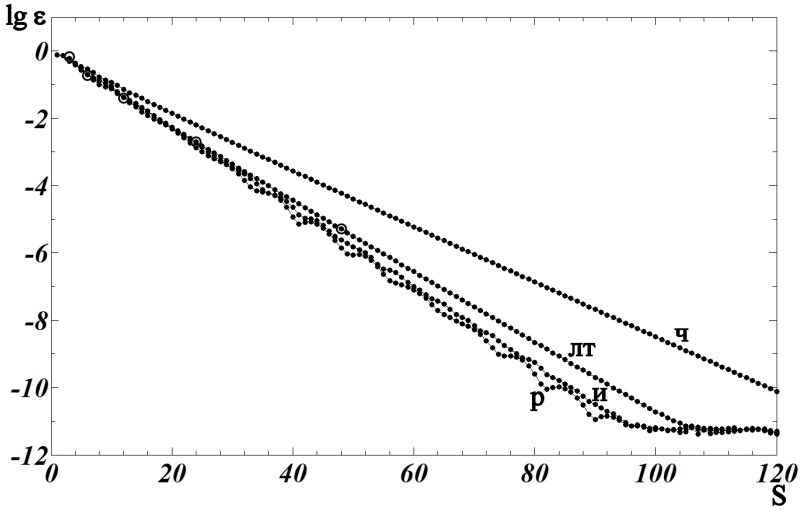

Рис. 8: Сильно неоднородная среда (35)-(36); $N = 1000$; обозначения соответствуют рис. 6.

Она сложна тем, что $k(x)$ меняется очень круто (практически скачком) (см. рис. 9), т.е. имитирует слоистую среду. Здесь равномерный, интерполяцион-

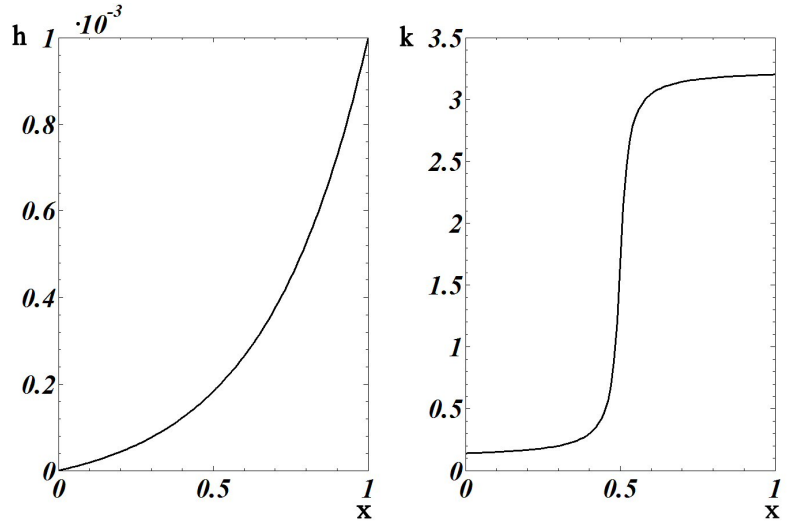

Рис. 9: Среда, изменяющаяся скачком (56)-(57).

ный и ЛТ-наборы обеспечивают примерно одинаковую скорость сходимости, но, также как в предыдущем примере, кривая равномерного набора не имеет регулярного участка (см. рис.10).

3) Рассмотрим задачу в однородной неограниченной области

$$k(x) = 1, \quad x \in [0,1]; \tag{58}$$

$$h(x) = (1-x^2)^{-1.5}. \tag{59}$$

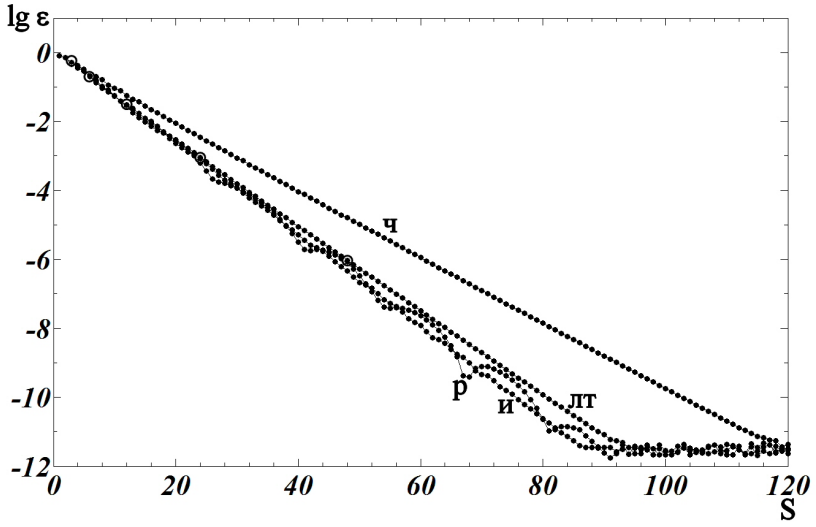

Рис. 10: Среда, изменяющаяся скачком (56)-(57); $N = 1000$; обозначения соответствуют рис. 6.

Счет на установление в неограниченных областях всегда считался очень трудным, поскольку при использовании квазиравномерных сеток $\lambda_{\max}/\lambda_{\min}$ хуже, чем $O(N^2)$. Очень часто это отношение составляет $O(N^4)$ [12]. В этом случае метод с постоянным оптимальным шагом или явным набором параметров давали бы $S = O(N^2)$, т.е. огромное число итераций. Однако для ЛТ-набора число итераций увеличивается всего в 1.5 раза по сравнению с простейшей задачей. Кривые сходимости разных наборов даны на рис. 11.

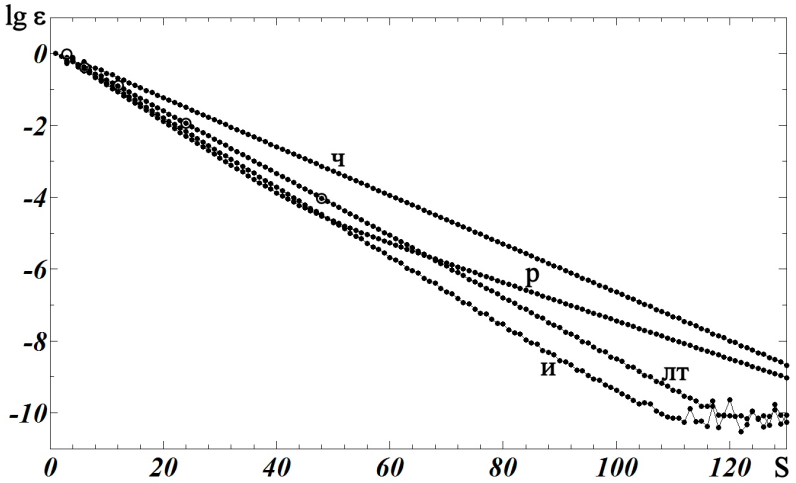

Рис. 11: Задача в неограниченной области (58)-(59); $N = 1000$; обозначения соответствуют рис. 6.

Отметим, что во всех этих примерах кривая ЛТ-набора имеет хороший регулярный участок. Это дополнительный довод в пользу ЛТ-набора.

Рассмотрим подробнее вопрос о сходимости различных методов в послед-

нем примере. Здесь использовались границы спектра, вычисленные методом обратных итераций с переменным сдвигом: $\lambda_{\min} = 3.2380 \cdot 10^{-3}$ и $\lambda_{\max} = 3.9976 \cdot 10^{6}$. Итерации обрывались при достижении относительной точности 10^{-6}, поэтому эти значения можно считать точными. Отношение границ спектра $\lambda_{\max}/\lambda_{\min} \approx 1.2 \cdot 10^{9}$, что означает плохую обусловленность матрицы линейной системы. Это отношение становится таким большим из-за того, что $\lambda_{\min}$ мало. Это можно понять по характеру точного решения: в бесконечной области оно раскладывается не в ряд, а в интеграл Фурье, причем значения λ начинаются от нуля.

Нетрудно убедиться, что при расчете по явной схеме с чебышевским набором параметров для для достижения точности $\varepsilon = 10^{-10}$ нужно сделать $S \approx 3.3 \cdot 10^{5}$ итераций. Такая же трудоемкость будет при расчете по неявной схеме с постоянным оптимальным шагом. Наконец, метод сопряженных градиентов даст такую точность за $S \approx 7.3 \cdot 10^{3}$ шагов. По рис. 11 видно, что при расчетах с логарифмическим набором и ЛТ-сеткой требуется значительно меньше – 115 итераций. Это говорит о больших преимуществах логарифмического набора при решении плохо обусловленных задач.

7.4 Двумерные расчеты

На рис. 12 представлены двумерные расчеты для разных N. По разным направлениям брались одни и те же сетки $N_x = N_y$, $h_x = h_y$, но полагалось $k_y = 10k_x$. Поэтому для гармоник с одинаковыми номерами $\lambda_y = 10\lambda_x$. Таким образом, границы спектров по обоим направлениям сильно отличались.

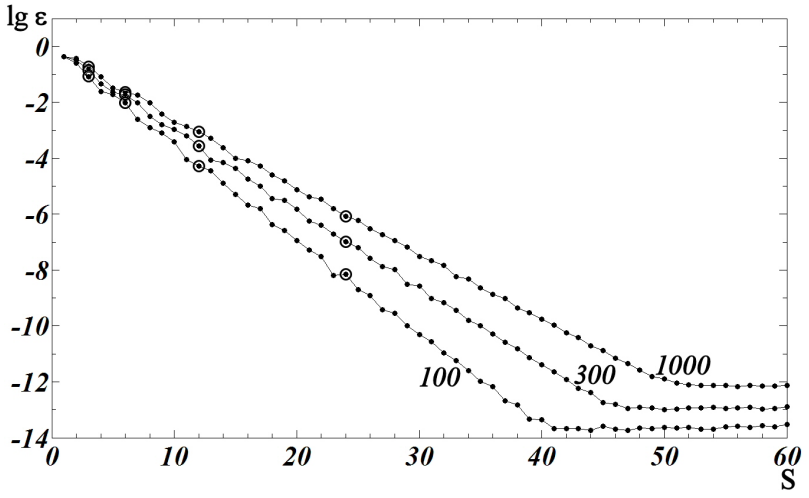

Рис. 12: Двумерный случай, сдвинутые спектры; $k_y = 10k_x$, $h_x = h_y = (N+1)^{-1}$; $\bullet$ – вычисления по ЛТ-набору; цифры около линий – значения N; $\circ$ – оценка (63).

Для двумерного случая справедливы те же выводы, что и для одномерного, но в двумерном случае скорость сходимости выше, чем в одномерном. Это

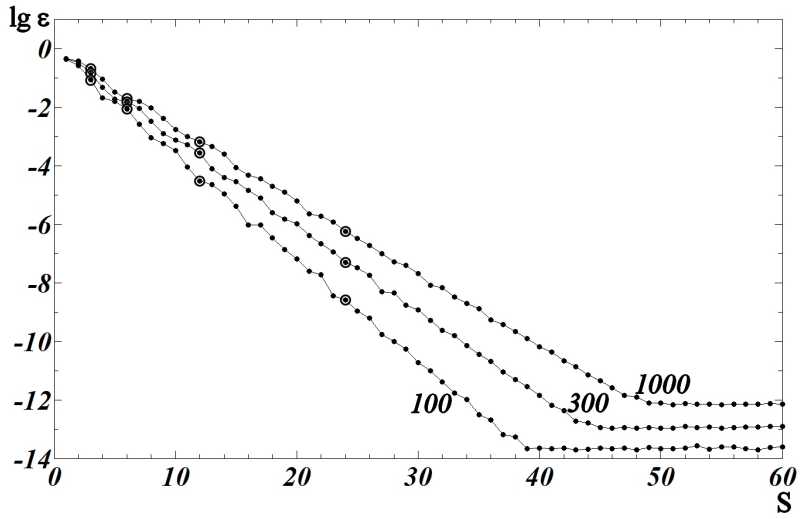

Рис. 13: Двумерный случай, совпадающие спектры; $k_x = k_y$, $h_x = h_y = (N+1)^{-1}$; обозначения соответствуют рис. 12.

объясняется тем, что каждую двумерную гармонику гасят сразу два множителя (19). Этот эффект особенно велик, если спектры по обоим направлениям одинаковы. Тогда сходимость ускоряется вдвое по сравнению с одномерным случаем (ср. рис. 7 и рис. 13).

7.5 Трехмерные расчеты

На рис. 14 показаны трехмерные расчеты. Сетки по разным направлениям

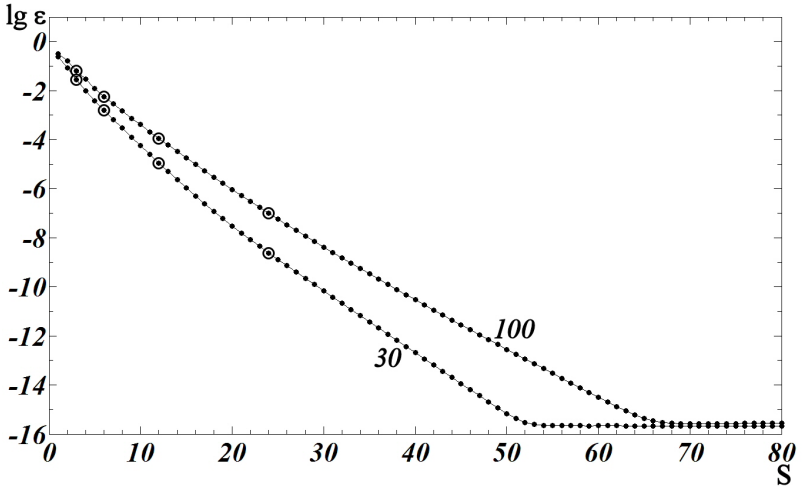

Рис. 14: Трехмерный случай, сдвинутые спектры; $k_y = 3k_x$, $k_z = 10k_x$, $h_x = h_y = h_z = (N+1)^{-1}$; Обозначения соответствуют рис. 12.

также одинаковы, но $k_y = 3k_x$, $k_z = 10k_x$. Поэтому спектры и их границы по разным направлениям сильно различаются. Результаты расчетов аналогичны двумерному случаю.

Проводились также вычисления для совпадающих спектров (при $k_x = k_y = k_z$). В этом случае сходимость оказывается медленнее (см. рис. 15).

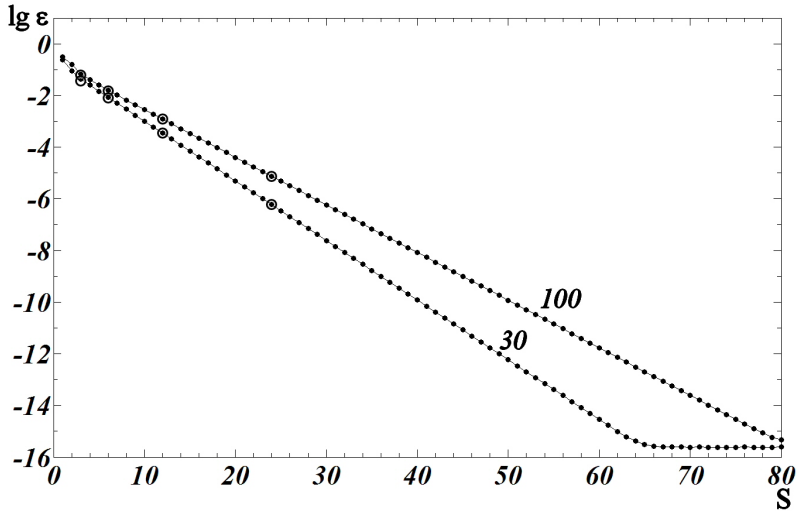

Рис. 15: Трехмерный случай, совпадающие спектры; $k_x = k_y = k_z$, $h_x = h_y = h_z = (N+1)^{-1}$; Обозначения соответствуют рис. 12.

Это можно объяснить тем, что для крайних гармоник множитель роста имеет положительный минимум $\rho = 1/9$, и эти гармоники гасятся не полностью. В примере со сдвинутыми спектрами минимум множителя роста оказывается отрицательным, и набор шагов строится по его нулям $\tau_\pm$. Это обеспечивает лучшее гашение.

7.6 Расширение границ спектра

Проводились вычисления с ЛТ-набором, построенным для спектра с измененными границами. Вместо точных границ брались $\lambda_{\min}/t$, $\lambda_{\max} t$. При $t < 1$ это эквивалентно сужению расчетных границ набора, при $t > 1$ – расширению.

Графики точности в зависимости от $\lg t$ при фиксированных N, S представлены на рис. 16-17. Видно, что сужение спектра дает сильное ухудшение точности. Это объясняется тем, что расчетные границы набора оказываются внутри истинных границ. Поэтому крайние гармоники гасятся плохо. Расширение спектра также ухудшает точность, но не столь заметно. Это происходит из-за увеличения длины каждого шага.

Отметим, что левая ($\lg t < 0$) и правая ($\lg t > 0$) ветви графика не образуют плавного перехода, а пересекаются под углом. Это означает, что ошибка в определении границ спектра приводит к заметному ухудшению точности. Поэтому оценки границ спектра должны быть насколько возможно аккуратными, причем соответствующими расширению.

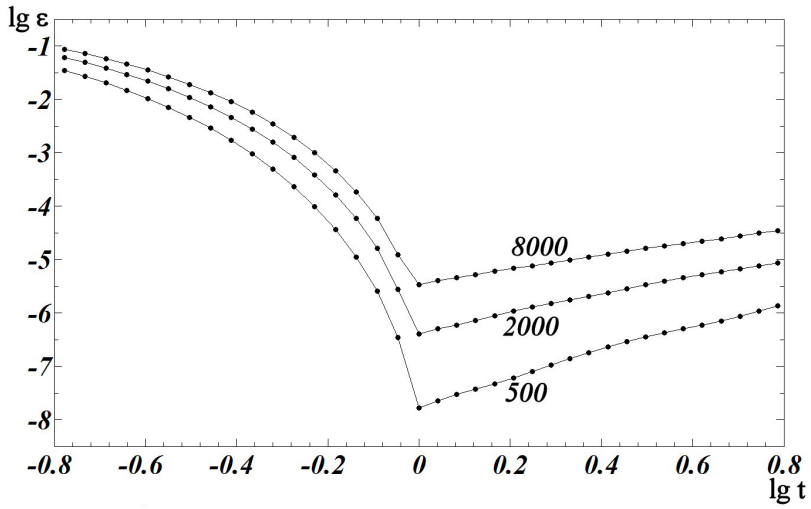

Рис. 16: Влияние границ расчетного спектра; $S = 50$, $k = 1$, $h = (N + 1)^{-1}$; цифры около линий – значения N.

Далее, видно, что правая ветвь графика хорошо аппроксимируется прямой, для которой можно вычислить коэффициент наклона.

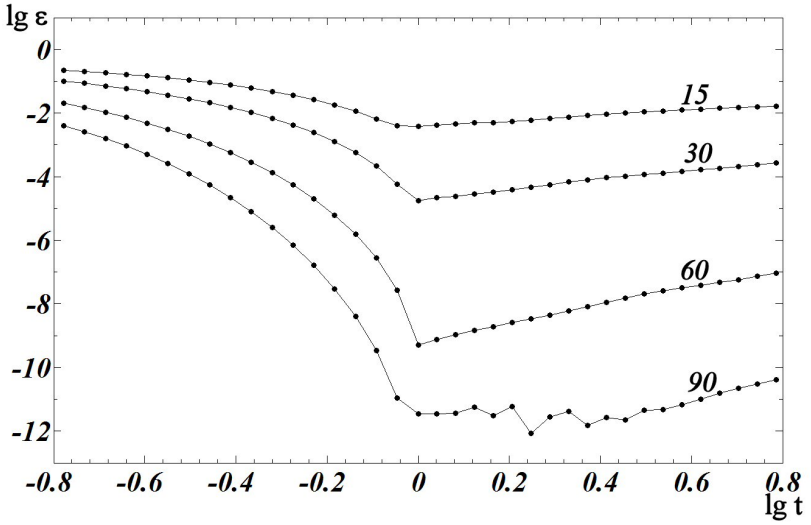

Рис. 17: Влияние границ расчетного спектра; $N = 500$, $k = 1$, $h = (N + 1)^{-1}$; цифры около линий – значения S.

Иными словами, точность при расширении можно представить в виде

$$\lg \varepsilon(t) = \lg \varepsilon_0 + \lg \varepsilon_1 \lg t. \tag{60}$$

Проведенные вычисления показывают, что $\lg \varepsilon_1$ зависит от N и от S, причем с неплохой точностью по N зависимость логарифмическая, а по S – линейная:

$$\lg \varepsilon_1 = (\eta_0 + \eta_1 \lg N)S, \tag{61}$$

где $\eta_0 \approx 0.12$, $\eta_1 \approx 0.02$. Эта эвристическая закономерность хорошо работает при умеренных расширениях (до $\lg t \approx 0.4$, т.е. $t \approx 2.5$).

По рис. 17 видно, что большим значениям S соответствует более резкое ухудшение точности при расширении. Так, для $S = 60$ излом более резкий, чем для $S = 30$, а для $S = 15$ график и вовсе имеет гладкий минимум. Значение $S = 90$ соответствует фону для задачи с истинным спектром. Однако, начиная с некоторого t, кривая точности возвращается на регулярный, дофоновый участок.

Расширение спектра может менять характер сходимости некоторых наборов (см. рис. 18). Так, чебышевский набор значительно теряет в скорости, а на кривой равномерного набора даже появляются участки немонотонности.

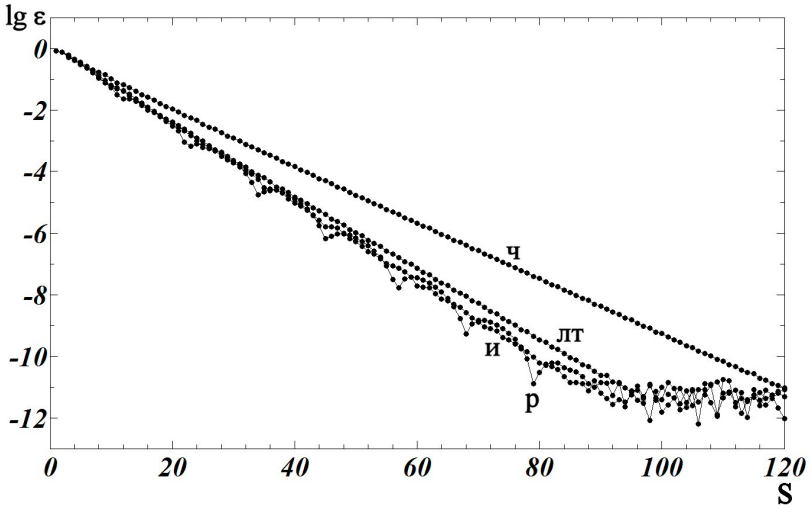

Рис. 18: Расчеты с расширенным спектром; $t = 4$; $S = 50$, $k = 1$, $h = (N+1)^{-1}$; обозначения наборов соответствуют рис. 5.

Это замечание особенно актуально, поскольку на практике границы спектра неизвестны, и их нужно оценивать с некоторым запасом. Заметим, что ЛТ-набор не имеет этих недостатков: для него кривая сходимости практически прямолинейна даже для заметного расширения ($t = 4$), а потеря скорости не так велика. Это дополнительный довод в пользу ЛТ-набора.

Сделанные выводы непосредственно переносятся на двумерный случай. В трехмерном случае вычисления также показывают нежелательность сужения или расширения границ спектра, поскольку это дает ухудшение точности (см. рис. 19).

В частных случаях (например, при совпадающих спектрах) расширение может давать преимущества (см. рис. 20). Но, во-первых, достигаемое таким образом улучшение точности не очень велико, и во-вторых, такие простые задачи не представляют практического интереса.

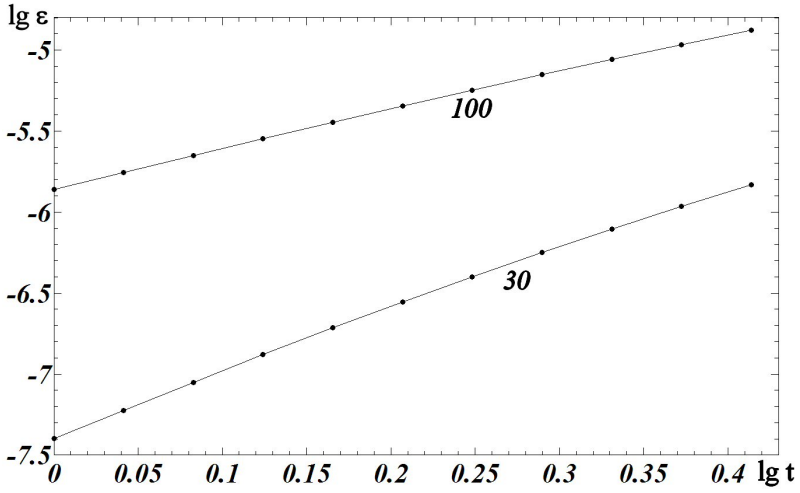

Рис. 19: Влияние границ расчетного спектра в трехмерном случае для несовпадающих спектров; $S = 20$; цифры около линий – значения N; k_α, h_α соответствуют рис. 14.

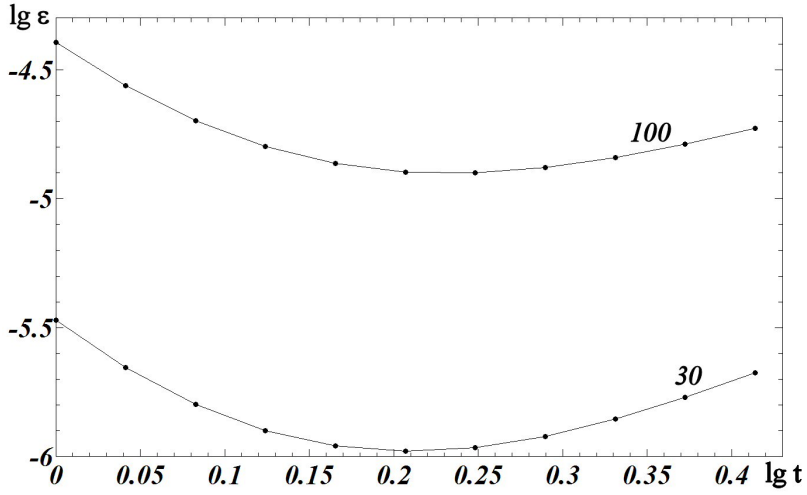

Рис. 20: Влияние границ расчетного спектра в трехмерном случае для совпадающих спектров; $S = 20$; цифры около линий – значения N; k_α, h_α соответствуют рис. 15.

8 Высокоточные расчеты

В литературе при иллюстрации сходимости итерационных методов обычно ограничиваются умеренной точностью $\varepsilon \sim 10^{-4}-10^{-6}$ и умеренными $N \sim 100$ (редко $N \sim 300$). Причина этого в медленной сходимости общеизвестных методов. Даже при таких скромных N и ε они требуют сотен итераций. Из рис. 6-14 видно, что эволюционно факторизованный счет на установление с ЛТ-набором обеспечивает такую точность за 15-35 итераций. Это на 2-3 порядка быстрее, чем для методов сопряженных направлений. Предложенный здесь метод позволяет достичь предельно возможной точности – фоновой $(10^{-11} - 10^{-13})$, причем за небольшое число итераций $(S \approx 40 - 80)$.

Решение сеточных уравнений с такой высокой точностью позволяет ре-

шать эллиптические уравнения на многократно сгущающихся сетках с применением уточнения по Ричардсону. Это дает возможность решать дифференциальные эллиптические уравнения с недостижимой ранее точностью.

9 Апостериорные оценки точности

9.1 Сгущение сеток

При итерационном решении сеточных уравнений непосредственно вычисляется невязка $R^{(S)}$. Погрешность решения в принципе можно оценить по невязке, поскольку

$$\left\| u^{(S)} - u \right\| \leqslant \lambda_{\max}/\lambda_{\min} \left\| R^{(S)} \right\|. \tag{62}$$

Оценка (62) мажорантная, а множитель $\lambda_{\max}/\lambda_{\min} = O(N^2) \gg 1$. Поэтому оценка (62) малопригодна.

Логарифмический счет на установление использует наборы с границами, определенными по границам спектра. Это порождает следующую неудобную особенность. Нужно заранее задать полное число итераций S и выполнять все шаги этого набора. Если выполнить только первую половину шагов, то будет эффективно погашена только половина гармоник, а остальные гармоники лишь слабо затухнут, и никакой сходимости не будет. Иными словами, нужно строить наборы шагов, плотно накрывающие весь спектральный интервал. Такие классы наборов будем называть допустимыми.

Кроме того, для алгоритмов описанного типа оценка S по ε (из (54) или (55)) является мажорантной и может давать большее значение S, чем реально нужно. При этом вопрос о фактической точности $u^{(S)}$ остается открытым. Воспользуемся двумя обстоятельствами. 1) Регулярные участки линий на рис. 6-7, 12, 14 практически прямолинейны. 2) Если набор шагов $\{\tau_s\}$ выбран, то сами шаги можно выполнять в произвольном порядке благодаря линейности процесса и устойчивости. Это позволяет строить последовательности двукратно сгущающихся сеток по τ, внешне напоминающие метод Ричардсона.

Для этого выберем некоторое небольшое значение S_0 ($1 \leqslant S_0 \leqslant 5$) и построим для него сетку $\left\{\tau_s^0\right\}$, $0 \leqslant s \leqslant S_0$ с помощью функции (46)-(47). Очевидно, эта сетка принадлежит классу допустимых сеток. Выполним на ней итерационный процесс и обозначим его результат через U_0.

Затем возьмем $S_1 = 2S_0$ и с помощью той же порождающей функции построим сетку $\left\{\tau_s^1\right\}$. Она также будет допустимой. Четные шаги $\left\{\tau_s^1\right\}$ совпадают с шагами сетки $\left\{\tau_s^0\right\}$, поэтому их можно не повторять. Достаточно

взять U_0 в качестве начального приближения и сделать счет на установление с нечетными шагами сетки $\left\{\tau_s^1\right\}$. Полученный результат обозначим U_1.

Потом возьмем сетку $\left\{\tau_s^2\right\}$ с $S_2 = 2S_1$ шагами, повторим описанную процедуру и так далее. В итоге получим последовательность решений U_q, $q = 1, 2, 3, ...$, соответствующих сгущающимся сеткам. При этом вычисления экономичны: суммарное число итераций во всех расчетах равно числу итераций последней сетки. Сама процедура сгущения эквивалентна перестановке шагов этой сетки.

Из экспоненциального характера сходимости логарифмического счета на установление следуют апостериорные асимптотически точные оценки норм погрешности:

$$\|U_q - u\| \approx \|U_{q+1} - U_q\|, \tag{63}$$

$$\|U_{q+1} - u\| \approx \|U_{q+1} - U_q\|^3 / \|U_q - U_{q-1}\|^2. \tag{64}$$

Оценка (64) означает экстраполяцию погрешности на U_{q+1}. Она учитывает, что регулярный участок кривой погрешности не проходит через начало координат. Этими эвристическими закономерностями можно пользоваться, пока расчеты не выйдут на ошибки округления. При этом (64) теряет применимость раньше, чем (63).

Формально для каждой сетки (кроме последней) можно пользоваться обеими оценками. Эти оценки для $S_0 = 3$ показаны на рис. 6, 8, 10-15 светлыми кружками. Видно, что они хорошо совпадают с действительными значениями погрешностей, вычисленными непосредственно по точному решению. Чем гуще сетка $\{\tau_s^q\}$, тем лучше совпадение, как и должно быть для асимптотически точной оценки.

Таким образом, предложенный метод дает апостериорную асимптотически точную оценку погрешности для счета на установление по эволюционно факторизованной схеме со сгущающейся ЛТ-сеткой по τ. Такой характер сходимости напоминает метод Ричардсона. Ранее оценок точности подобного типа не предлагалось.

Заметим, что аналогия с методом Ричардсона является неполной. Оценки (63)-(64) применимы лишь к нормам погрешности. Подобную поточечную оценку погрешности при сгущении сетки по τ построить невозможно.

9.2 Алгоритм расчета

В расчетах необходимо задавать S заранее. Поэтому вопрос о фактической оценке точности состоит из двух частей: 1) априорная оценка необходимого S, не влекущая за собой непроизводительных расчетов; 2) апостериорное подтверждение точности.

Для априорной оценки S необходимо задать требуемую точность ε. Здесь возможны две ситуации. 1) Точность $\varepsilon_{\text{п}}$, не превосходящая фоновую $\varepsilon_{\text{ф}}$, может быть задана пользователем. 2) Требуется рассчитать как можно точнее. В последнем случае необходимо оценить фон ошибок округления. Для этого существуют мажорантные оценки. Фон получается умножением ошибки единичного округления на число обусловленности матрицы. Минимально возможной оценкой является угловое число обусловленности [10], [11], но его трудно вычислить. В наших расчетах естественно вычисляется несколько завышенное спектральное число обусловленности $\varkappa_\lambda = \sum \lambda_{\alpha,\max} / \sum \lambda_{\alpha,\min}$. Тогда $\varepsilon_{\text{ф}} = 10^{-16.2}\varkappa_\lambda$ (для 64-разрядного программного обеспечения).

Положим $\varepsilon = \max\{\varepsilon_{\text{п}}, \varepsilon_{\text{ф}}\}$. Исходя из этой точности и оценок границ спектра, вычислим требуемое S. Будем делить это число рекуррентно на 2 до тех пор, пока не получится число, немного меньшее небольшого целого S_0 $(1 \ldots 5)$. Полученное S_0 следует взять в качестве начального и применять процедуру сгущения.

Процедура сгущения применяется до достижения требуемого S. При этом все погрешности, кроме последней, вычисляются по оценке (63). Эта оценка надежна, поскольку применяется вдали от фона. Для последнего вычисления применим экстраполяцию (64) и сравним полученную оценку $\varepsilon_{\text{э}}$ с фоновой $\varepsilon_{\text{ф}}$. В качестве погрешности последнего вычисления выберем величину $\max\{\varepsilon_{\text{э}}, \varepsilon_{\text{ф}}\}$.

Для визуального контроля удобно выводить оценки погрешностей на график. В реальных расчетах максимальное $S \leqslant 80$, поэтому при $S_0 > 5$ на этом графике будет слишком мало точек.

10 Актуальные задачи

Предложенный в данной статье метод не охватывает несколько типов задач, имеющих большое практическое значение. К ним относятся задачи для уравнения со смешанными производными и задачи в произвольной области с криволинейной границей. Основную трудность здесь представляет построение метода факторизации и конструктивных оценок границ спектра.

Вызывает трудности решение параболических и эллиптических уравнений в слоистой среде. Если границы раздела слоев параллельны границам области, то следует применять бикомпактные схемы. Если же граница раздела косая или имеет более сложную форму, то вопрос остается открытым.

Проблематично решение несамосопряженных задач с конвективным членом, поскольку не разработано корректного перехода от параболического уравнения к уравнению переноса в сингулярно возмущенной задаче.

Список литературы

1. *Самарский А. А., Андреев В. Б.* // Разностные методы для эллиптических уравнений. М: Наука, 1976.
2. *Самарский А. А., Николаев Е. С.* // Методы решения сеточных уравнений. М.: Наука, 1978.
3. *Фадеев Д. К., Фадеева В. Н.* // Вычислительные методы линейной алгебры. М.: Физматгиз, 1963.
4. *Болтнев А. А., Калиткин Н. Н., Качер О. А.* Логарифмически сходящийся счет на установление. // Доклады Академии Наук. 2005. Т. 404. N 2. С.177-180.
5. *Калиткин Н. Н.* // Численные методы. М.: Наука, 1978.
6. *Бахвалов Н. С., Жидков Н П., Кобельков Г. М.* // Численные методы. М.: БИНОМ, Лаборатория знаний, 2004.
7. *Яненко Н. Н.* // Метод дробных шагов решения многомерных задач математической физики. Новосибирск: Наука - Сибирское отделение, 1967.
8. *Самарский А. А.* // Теория разностных схем. М.: Наука, 1989.
9. *Калиткин Н. Н.* Улучшенная факторизация параболических схем. // Доклады Академии Наук. 2005. Т. 402. N 4. С. 467-471.
10. *Калиткин Н. Н., Юхно Л. Ф., Кузьмина Л. В.* Количественный критерий обусловленности систем линейных алгебраических уравнений. // Доклады Академии Наук. 2010. Т. 434. N 4. С. 464-467.
11. *Калиткин Н.Н., Юхно Л.Ф., Кузьмина Л.В.* Критерий обусловленности систем линейных алгебраических уравнений. // Математическое моделирование. 2011. Т. 23. N 2. С. 3-26.
12. *Калиткин Н. Н., Альшин А. Б., Альшина Е. А., Рогов Б. В.* // Вычисления на квазиравномерных сетках. М.: Физматлит, 2005.

Содержание