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

Препринты ИПМ • Препринт № 138 за 2019 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Белов А.А., Вергазов А.С., Калиткин Н.Н.

Погрешность численного решения жестких задач Коши на геометрически-адаптивных сетках

Рекомендуемая форма библиографической ссылки: Белов А.А., Вергазов А.С., Калиткин Н.Н. Погрешность численного решения жестких задач Коши на геометрически-адаптивных сетках // Препринты ИПМ им. М.В.Келдыша. 2019. № 138. 23 с. http://doi.org/10.20948/prepr-2019-138
URL: http://library.keldysh.ru/preprint.asp?id=2019-138

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

А. А. Белов, А. С. Вергазов, Н. Н. Калиткин

ПОГРЕШНОСТЬ ЧИСЛЕННОГО РЕШЕНИЯ ЖЕСТКИХ ЗАДАЧ КОШИ НА ГЕОМЕТРИЧЕСКИ-АДАПТИВНЫХ СЕТКАХ

Москва, 2019

УДК 519.6

А. А. Белов, А. С. Вергазов, Н. Н. Калиткин. Погрешность численного решения жестких задач Коши на геометрически-адаптивных сетках

Уточнено понятие жесткости системы ОДУ. Указаны основные трудности, возникающие при решении соответствующей задачи Коши. Показаны преимущества перехода к новому аргументу – длине дуги интегральной кривой. Обсуждены разные критерии выбора шага и улучшена формула выбора шага по кривизне интегральной кривой. Изложена стратегия расчета, позволяющая а) строить последовательности сеток, асимптотически переходящих в квазиравномерные, б) одновременно с решением получать мажорантную оценку погрешности. Приведены иллюстративные примеры расчета.

Ключевые слова: жесткая задача Коши, автоматический выбор шага, оценки по методу Ричардсона.

A. A. Belov, A. S. Vergazov, N. N. Kalitkin. Numerical solution error of stiff Cauchy problems on geometrically adaptive meshes

The concept of stiffness of ODE system is refined. Major difficulties arising in solution of the corresponding Cauchy problems are pointed out. Advantages of the arc length arguments are shown. Different step selection criteria are discussed and step selection formula based on curvature of the integral curve is improved. A procedure is described which allows to a) construct mesh sequence tending to a quasi-uniform one, b) obtain majorant error estimate simultaneously with the solution. Illustrative calculation examples are given.

Keywords: stiff Cauchy problem, automatic step selection, Richardson method error estimates

Работа поддержана грантом РФФИ №18-01-00175

1 Проблема

Немного истории. До конца 1940-х годов задачи Коши для ОДУ успешно решались явными численными методами Адамса, Рунге-Кутты и др. Однако в конце 1940-х годов появился ряд новых прикладных задач (например, химическая кинетика горения ракетного топлива), на которых явные схемы требовали неприемлемо малого шага. Такие задачи получили название жестких, и для них стали разрабатываться новые численные методы. Этим методам посвящена обширная литература, обзор которой дан в классической монографии [1]. Однако, во-первых, в этой монографии не дано рекомендаций, какими методами лучше пользоваться. Во-вторых, есть немало задач, которые все методы, описанные в [1], не позволяют рассчитать, «срываясь» задолго до конечной точки. В-третьих, даже если расчет завершен, остается открытым вопрос о фактически достигнутой точности.

Понятие жесткости. Строгого математического определения жесткости систем ОДУ пока не дано. Предлагались формальные определения жесткости по спектру матрицы Якоби правых частей: если все спектральные числа отрицательны и среди них есть большие по модулю, то задачу относили к жестким. При этом неявно предполагалось, что для линейных задач все компоненты решения затухают в соответствии с величинами спектральных чисел. Однако это определение опровергается примером Винограда [2]. В нем для линейной неавтономной системы все собственные значения отрицательны и не зависят от аргумента t, а решение имеет экспоненциально нарастающую компоненту! Очевидно, для нелинейных задач строго определить жесткость еще труднее.

Разумнее определять жесткость как качественное понятие, описывающее структуру решения. Ракитский [3] определял жесткость как наличие у решения компонент, скорости изменения которых очень сильно отличаются (от очень быстрых до медленных). Это близко примыкает к теории пограничного слоя в работах Тихонова и его учеников (см., например, [4] – [6]).

В структуре решения жесткой задачи есть участки быстрого изменения, которые называются пограничными слоями, и участки плавного изменения, называемые регулярными. Это хорошо иллюстрируется на простейшем примере

$$\frac{du}{dt} = -\lambda u, \tag{1}$$

где $\lambda \gg 1$ (см. рис. 1). Мы предложили выделять еще один участок решения – переходную зону [7]: это участок перехода пограничного слоя в регулярное решение. Он характеризуется большой кривизной кривой $u(t)$.

Строго говоря, термин «пограничный слой» относится к быстрому изменению решения в начальный момент времени $t = 0$. Если быстрое изменение происходит в момент $t > 0$, то его называют внутренним пограничным слоем или контрастной структурой.

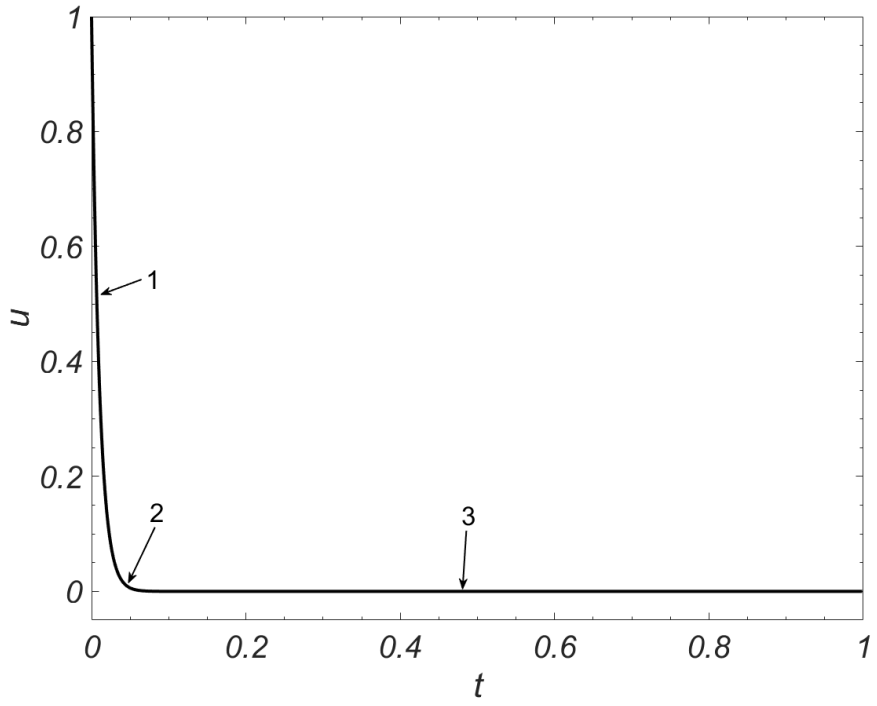

Рис. 1: Решение задачи (1) в аргументе t: 1 – пограничный слой, 2 – переходная зона, 3 – регулярный участок.

Очевидно, численный расчет жестких задач представляет значительные трудности. Во-первых, пограничный слой требует очень мелкого шага.

Во-вторых, не всякая схема даже при хорошем выборе шага позволяет считать без срывов (то есть позволяет выполнить расчет до конца, давая при этом хотя бы внешне правдоподобные результаты).

В-третьих, недостаточно только провести расчет до заданного момента. Нужно уметь надежно оценить достигнутую точность. Теоретические мажорантные оценки точности при этом неэффективны: они используют значения высоких производных решения, которые априори неизвестны и для жестких задач очень велики. Поэтому особое значение приобретают апостериорные вычисления погрешности, проводимые одновременно с нахождением самого решения.

Выбор шага. Заданной точности традиционно пытаются добиться с помощью некоторого алгоритма выбора шага. В настоящее время практически во всех пакетах программ используется один из двух методов выбора шага, описанных в [1], [8]. Чаще всего выбирают шаг с помощью вложенной схемы. Расчет проводят по схеме порядка точности p, из промежуточных

стадий которой можно составить схему порядка точности $p-1$. Эта схема называется вложенной. Вычисление каждого шага проводят по обеим схемам и сравнивают результаты. Разность этих результатов считают локальной погрешностью вложенной схемы. Если она примерно равна заданной пользователем погрешности tol (ее называют $tolerance$), то с этой же величиной h выполняется следующий шаг. В противном случае h увеличивают либо уменьшают по некоторому правилу. Именно так работает программа Дормана-Принса. Однако тестирование показывает, что на задачах высокой жесткости этот алгоритм нередко «срывается», даже отдаленно не обеспечивая требуемую пользователем точность.

Второй метод – локальное сгущение сетки. Используют только одну схему и проводят вычисления с шагом h и $h/2$. Локальную погрешность определяют по разности этих двух расчетов и дальше поступают аналогично предыдущему методу. Во-первых, поскольку этот метод требует вычисления двух дополнительных полушагов, его трудоемкость в среднем втрое больше предыдущего. Поэтому он менее употребителен. Во-вторых, этот метод, подобно предыдущему, также нередко срывается на задачах высокой жесткости.

В стандартных пакетах программ пользователь задает требуемую точность tol, и шаги сетки выбираются на основе вложенной схемы либо локального сгущения шага. На этой сетке проводится единственный расчет. Пользователю предлагается поверить, что погрешность этого расчета равна заданной tol. Это никак не доказывается.

Тестирование таких пакетов на задачах с известными точными решениями показывает следующее. Для мягких задач фактическая погрешность хотя и может превышать tol, но не в десятки раз. Напротив, на жестких задачах фактическая погрешность нередко бывает на 2-3 порядка больше, чем tol.

Оригинальный алгоритм выбора шага, напоминающий локальное сгущение, был предложен в [9]. Он основан на анализе сходимости метода Ньютона при решении системы нелинейных алгебраических уравнений относительно значения решения на новом шаге $\hat{\mathbf{u}}$. Если итерации сходятся медленно, то значение решения с предыдущего шага $\mathbf{u}$ (выбираемое в качестве начального приближения) далеко от $\hat{\mathbf{u}}$ и шаг следует уменьшить. Однако жесткие задачи могут быть линейными, а в них для нахождения $\hat{\mathbf{u}}$ не требуется никакого итерационного алгоритма.

В последние годы предложен принципиально другой метод, основанный на геометрических характеристиках интегральных кривых [10] – [12]. В нем адаптивная сетка строится по кривизне интегральной кривой. Этот метод представляется наиболее перспективным. Практическим аспектам применения этого метода посвящена данная работа.

2 Переход к длине дуги

Существует общий прием, позволяющий кардинально облегчить решение жестких задач. Он заключается в переходе к новому аргументу – длине дуги интегральной кривой. Рассмотрим исходную задачу Коши для системы ОДУ порядка M

$$\frac{du_m}{dt} = f_m(t, u_1, u_2, \ldots, u_M), \quad 1 \leqslant m \leqslant M, \quad 0 \leqslant t < T, \quad u_m(0) = u_m^0. \tag{2}$$

Для решения системы (2) используют численные методы некоторого порядка точности p. Их применение оправдано, если у решения существует $p+1$ непрерывная производная. Напомним, что для этого правые части должны иметь p-е непрерывные производные по всем аргументам. Будем полагать это условие выполненным, если не сказано обратное.

Жесткость означает, что есть быстро изменяющиеся компоненты и для них $|f_m| \gg 1$. Эта система в общем случае неавтономна. Формально мы можем считать аргумент t также некоторой функцией $u_0(t) \equiv t$. Тогда ей соответствует правая часть $f_0(t) \equiv 1$ и начальное условие $u_0^0 = 0$. Это простейшая автономизация системы (2).

Введем для автономизированной системы длину дуги интегральной кривой

$$dl = \left[\sum_{m=0}^{M} (du_m)^2\right]^{1/2} = \left[\sum_{m=0}^{M} f_m^2(u_0, u_1, \ldots, u_M)\right]^{1/2} dt. \tag{3}$$

Заменяя dt на dl в (2) с помощью формулы (3), получим

$$\frac{d\mathbf{u}}{dl} = \mathbf{F}(\mathbf{u}), \quad \mathbf{u} = \{u_0, \ldots u_M\}, \quad \mathbf{F} = \mathbf{f}/\rho, \quad \rho = (\mathbf{f}, \mathbf{f})^{1/2}. \tag{4}$$

Если правые части (2) имеют p-е непрерывные производные по всем аргументам, то то же справедливо и для системы (4).

Новая система (4) является автономной. Легко видеть, что $(\mathbf{F}, \mathbf{F}) = 1$. Тем самым, все компоненты правых частей невелики, так что пограничные слои системы (2) превращаются в регулярные участки системы (4). Это хорошо иллюстрируется рис. 2 для примера (1).

Таким образом, переход к длине дуги позволяет преодолеть трудности, связанные с численным расчетом пограничных слоев. Однако между участками бывших пограничных слоев и регулярными решениями по-прежнему остаются переходные зоны, в которых интегральная кривая для жестких задач имеет большую кривизну. Расчет переходной зоны по-прежнему представляет трудность и требует существенного измельчения шага.

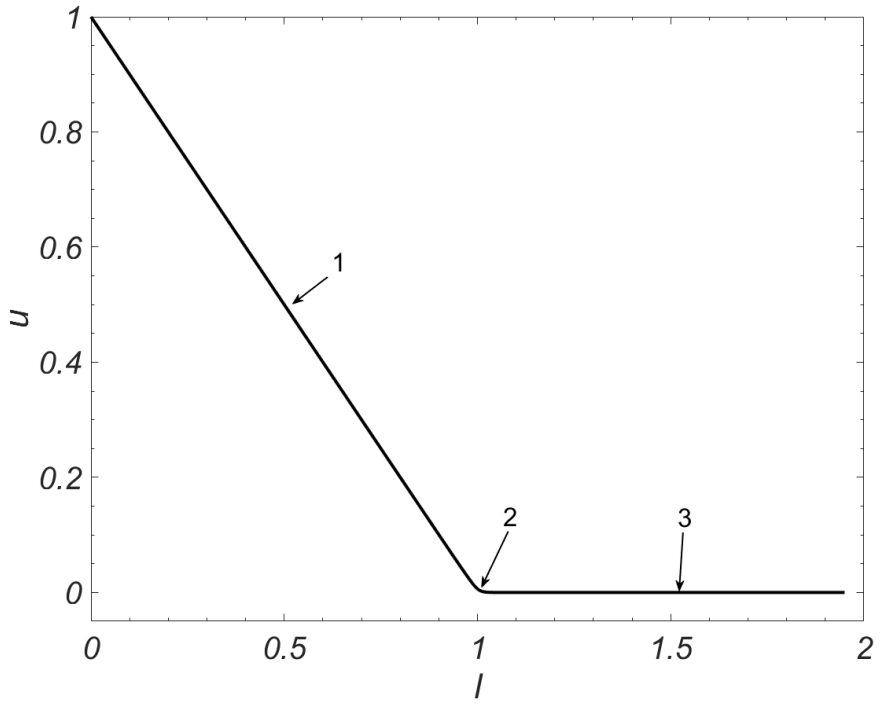

Рис. 2: Решение задачи (1) в аргументе l, обозначения соответствуют рис. 1.

По-видимому, переход к длине дуги был впервые предложен в [13]. Его изложение имеется в монографии [14]. Там же доказано, что такая параметризация является наилучшей с точки зрения обусловленности системы. Поэтому для жестких систем переход к длине дуги должен стать обязательной частью решения задачи.

Однако пока этот метод остается малоизвестеным широкому кругу вычислителей. В монографиях [1], [8] он отсутствует. По-видимому, в учебной литературе он имеется только в [15].

3 Построение геометрически-адаптивных сеток

Квазиравномерные сетки. Они были предложены Самарским в 1952 году, а впервые опубликованы Сидоровым в 1966 году [16]. Математическое определение таких сеток дано в [17]. Напомним его.

Семейство сеток $\omega_N = \{l_n,\ 0 \leqslant n \leqslant N\}$ на $[a, b]$ называется квазиравномерным, если существует такая строго возрастающая достаточно гладкая функция $\psi(\xi)$, $0 \leqslant \xi \leqslant 1$, $\psi(0) = a$, $\psi(1) = b$, что $l_n = \psi(\xi_n)$, $1 \leqslant n \leqslant N$ при любом N.

Отметим некоторые следствия. Шаг сетки $h_n = l_n - l_{n-1}$. Отношение соседних шагов $h_n/h_{n-1} \to 1$ при $N \to \infty$ (отсюда название квазиравномерные).

Середина интервала сетки определяется как $l_{n-1/2} = \psi((n - 1/2)/N)$.

Если из семейства квазиравномерных сеток выбрать семейство с удваивающимися значениями N, то узлы любой из этих сеток являются четными узлами следующей сетки с удвоенным числом шагов, а середины ее интервалов являются нечетными узлами следующей сетки.

Заметим, что квазиравномерные сетки можно строить в неограниченной области, что позволяет решать задачи в таких областях сравнительно простыми средствами с постановкой точных граничных условий непосредственно на бесконечности.

Близость кривых. Традиционно близость кривых $\mathbf{u}(l)$ и $\mathbf{v}(l)$ рассматривают через норму разности $\mathbf{u}-\mathbf{v}$. Такой подход неудобен для разрывных решений, потому что небольшое несовпадение местоположений разрывов двух кривых сильно меняет норму. Аналогичная ситуация имеет место для жестких задач, ибо их пограничные слои очень напоминают сильные разрывы. Поэтому необходимо иное определение близости кривых.

Рассмотрим интегральную кривую $\mathbf{u}(l)$ как множество точек в евклидовом $M+1$-мерном пространстве. Расстояние между двумя кривыми определим как расстояние между соответствующими множествами в метрике Хаусдорфа. Напомним это

Определение. *Пусть множества U, V состоят из точек $\mathbf{u}$, $\mathbf{v}$ соответственно. Тогда расстояние от U до V есть*

$$D(U,V) = \max\{\sup_{\mathbf{v}\in V}\inf_{\mathbf{u}\in U}|\mathbf{u}-\mathbf{v}|, \sup_{\mathbf{u}\in U}\inf_{\mathbf{v}\in V}|\mathbf{u}-\mathbf{v}|\}. \quad \bullet \tag{5}$$

Использование sup делает это определение аналогом нормы C. Если в этом определении заменить sup на интеграл по dl, взять $|\mathbf{u}-\mathbf{v}|$ в квадрате и извлечь из ответа квадратный корень, то получим аналог нормы L_2.

Оба определения достаточно естественно используются для решений с разрывами или пограничными слоями.

Адаптивная сетка. Пусть требуется решить задачу (4) на отрезке $0 \leqslant l \leqslant L$. Выбирая шаг, мы строим на этом отрезке сетку l_n, $0 = l_0 < < l_1 < \ldots < l_N = L$ из N интервалов. Обозначим через $\varkappa_n$ кривизну и через $R_n = 1/\varkappa_n$ – радиус кривизны интегральной кривой в узлах сетки.

Наша задача – построить сетку, обеспечивающую как можно более высокую точность. Такую сетку будем называть оптимальной. Интуитивно представляется, что шаги такой сетки должны сгущаться в областях большой кривизны, но соседние шаги таких сеток не должны сильно различаться. Поэтому естественно искать оптимум в классе квазиравномерных сеток.

Потребуем выполнения двух условий. Во-первых, кривизна выражается через вторые производные решения. Пусть правые части системы (4) имеют вторые непрерывные производные. Тогда кривизна $\varkappa(l)$ будет иметь

первую непрерывную производную. Во-вторых, ограничимся классом квазиравномерных сеток l_n с дважды непрерывно дифференцируемой производящей функцией.

Построим оптимальную сетку для схемы Эйлера. Шаг по этой схеме есть движение по касательной. Сравнивая расхождение кривой и касательной на шаге h_n, получаем величину локальной ошибки на одном шаге

$$\delta_n = \frac{h_n^2}{2R_n}. \tag{6}$$

Сама ошибка есть вектор, перпендикулярный кривой. Таким образом, эту ошибку нужно рассматривать в смысле метрики Хаусдорфа. Тогда аналог сеточной нормы L_2 погрешности Δ определяется выражением

$$\Delta^2 = \sum_{n=1}^{N} \delta_n^2 h_n = \frac{1}{4} \sum_{n=1}^{N} \frac{h_n^5}{R_n^2}. \tag{7}$$

Будем искать набор шагов h_n, минимизирующий Δ. При этом нужно учитывать, что значения R_n сами зависят от положения узлов l_n и, тем самым, от набора шагов. Удобнее приближенно перейти к непрерывному индексу n, тогда $h_n \approx dl/dn$ и $R_n = R(l)$. При сделанных выше предположениях о гладкости функций такой переход является асимптотически точным. При этом должно выполняться условие

$$\sum_{n=1}^{N} h_n \approx \int_0^N \frac{dl}{dn} dn = L. \tag{8}$$

Задача на условный экстремум $\Delta \to \min$ с ограничением (8) методом Лагранжа сводится к задаче на безусловный экстремум

$$\frac{1}{4} \int_0^N \frac{1}{R^2(l)} \left(\frac{dl}{dn} \right)^5 dn - \frac{\mu}{4} \left(\int_0^N \frac{dl}{dn} - L \right) \to \min, \tag{9}$$

где $\mu/4$ – множитель Лагранжа.

Вариационное уравнение Эйлера с краевыми условиями для задачи (9) принимает вид

$$\frac{d}{dn} \left[\frac{5}{R^2(l)} \left(\frac{dl}{dn} \right)^4 - \mu \right] + \frac{1}{2R^3(l)} \left(\frac{dl}{dn} \right)^5 \frac{dR}{dl} = 0, \quad l(0) = 0,\ l(N) = L. \tag{10}$$

Последнее выражение приводится к следующей форме:

$$\frac{d^2 l}{dn^2} - \frac{2}{5} \left(\frac{dl}{dn} \right)^2 \frac{d \ln R}{dl} = 0, \quad l(0) = 0,\ l(N) = L. \tag{11}$$

У этого уравнения нетрудно найти первый интеграл

$$h \equiv \frac{dl}{dn} = CR^{2/5}, \quad C = \text{const}. \tag{12}$$

Отсюда для положения узлов получаем

$$n(l) = C^{-1} \int_0^l R^{-2/5}(\tilde{l}) d\tilde{l}. \tag{13}$$

Константа определяется из условия $n(L) = N$. Находя эту константу, сформулируем полученный результат следующим образом:

Теорема. *При сделанных выше предположениях о гладкости оптимальная сетка для схемы Эйлера при $N \to \infty$ асимптотически удовлетворяет условию*

$$h_n = \frac{1}{N} \varkappa_n^{-2/5} \int_0^L \varkappa^{2/5}(l) dl, \quad 1 \leqslant n \leqslant N. \quad \bullet \tag{14}$$

Формула для шага. Построим расчетную формулу для шага. Обозначим через $N_{\max}$ число шагов, которое целесообразно взять на всей сетке с учетом переходных зон, а через $N_{\min}$ – число шагов, которое целесообразно взять на регулярных участках (без учета переходных зон). Очевидно, должно выполняться $N_{\min} \ll N_{\max}$. Тогда шаг ограничивается сверху выражением

$$h \leqslant L/N_{\min}. \tag{15}$$

В формуле (12) перейдем от радиуса кривизны к кривизне и подставим явное значение константы $\text{const} = 1/N_{\max}$. В качестве расчетной формулы выберем простую интерполяцию (12) и (15)

$$h = \left[\frac{N_{\min}}{L} + \frac{N_{\max} \varkappa^{2/5}}{\int_0^L \varkappa^{2/5}(l) dl} \right]^{-1}. \tag{16}$$

Способ вычисления интеграла в (16) будет описан далее.

Очевидно, сетка, построенная указанным образом, адаптирована к решению. Будем называть ее ***геометрически-адаптивной*** (GEAD mesh – Geometrically Adaptive mesh).

Замечание. Переход от аргумента t к аргументу l кажется усложнением задачи. Однако такой переход целесообразно делать даже для мягких задач. Поясним причину этого.

Во-первых, правые части формулы (4) очень просто выражаются через правые части формулы (2). Мягкие задачи решают явными схемами, в которых требуется вычислять только правые части (но не матрицу Якоби), поэтому для них никакого усложнения фактически не происходит.

Во-вторых, приведенный выше удачный выбор шага удалось построить только для аргумента l. Формулы выбора шага по аргументу t, используемые в методах вложенных схем или локального сгущения шага, далеко не столь надежны, особенно в случае жестких задач.

В-третьих, в аргументе l пограничные слои перестают быть трудными участками и легко рассчитываются крупными шагами. Измельчение шага требуется лишь в переходных зонах.

4 Вычисление кривизны

Для построения геометрически-адаптивной сетки нужно уметь вычислять кривизну многомерной кривой. Мы не встречали в литературе конструктивных формул для вычисления кривизны в многомерном пространстве. Однако есть общее

Определение. *Кривизна равна производной от единичного вектора направления касательной по длине дуги (то есть второй производной радиуса-вектора кривой по длине дуги).*

Реализации этого определения для явных и неявных схем различны. Опишем эти реализации.

Простейшее выражение. Вводя длину дуги в качестве аргумента, мы перешли от системы (2) к системе (4). В системе (4) правые части F_m суть компоненты вектора касательной к интегральной кривой. Напомним, что $\mathbf{F}$ есть вектор единичной длины. Таким образом, кривизна получается дифференцированием вектора $\mathbf{F}$ по скаляру l и является вектором

$$\varkappa = \frac{d\mathbf{F}}{dl}. \tag{17}$$

Правые части вычисляются на каждом шаге. Напишем простейшую разностную аппроксимацию

$$\hat{\varkappa} = \varkappa(l_n) = [\mathbf{F}(u_n) - \mathbf{F}(u_{n-1})]/h_n; \quad h_n = l_n - l_{n-1}. \tag{18}$$

Эта аппроксимация имеет первый порядок точности, что хорошо согласуется с точностью схемы Эйлера. Поэтому такая формула уже пригодна для построения геометрически-адаптивных сеток.

Кривизна (18) вычисляется после завершения текущего шага, поэтому она может использоваться для определения величины только следующего шага. На первом шаге значение кривизны неизвестно. Поэтому расчет первого шага нужно выполнить дважды: при первом вычислении найти величину шага по кривизне, взятой «с потолка», а при повторном вычислении скорректировать шаг по найденной кривизне.

Явные схемы Рунге-Кутты общеизвестны. Напомним их, чтобы ввести соответствующие обозначения. Схема с p стадиями имеет следующий вид:

$$\hat{\mathbf{u}} = \mathbf{u} + h\sum_{s=1}^{p} b_s \mathbf{w}_s, \quad \mathbf{w}_s = \mathbf{f}\left(\mathbf{u} + h\sum_{q=1}^{s-1} a_{sq}\mathbf{w}_q\right). \tag{19}$$

Порядок точности схемы не может превосходить числа стадий. Для $p \leqslant 4$ можно подобрать такие коэффициенты, что порядок точности равняется числу стадий. При $p \geqslant 4$ порядок точности «отстает» от числа стадий тем сильнее, чем больше p. Кроме того, при повышении p ухудшается надежность, поскольку растет порядок производных в остаточном члене схемы.

Выражение (18), по существу, получено для одностадийной схемы Рунге-Кутты. В многостадийных схемах можно использовать для построения кривизны величины $\mathbf{w}_s$ из промежуточных стадий (19), а также величину $\hat{\mathbf{w}} = \mathbf{F}(\hat{\mathbf{u}})$. Использование $\hat{\mathbf{w}}$ не увеличивает объем расчетов: выражение для кривизны относится к следующему шагу, на котором $\hat{\mathbf{w}}$ все равно нужно вычислять. Поэтому выражение кривизны ищем в следующем виде:

$$\hat{\varkappa} = h^{-1}\sum_{q=1}^{S+1} c_q \mathbf{w}_q, \quad \mathbf{w}_{S+1} = \mathbf{F}(\hat{\mathbf{u}}). \tag{20}$$

Коэффициенты c_q из (20) и b_q, a_{sq} из (19) нужно подбирать так, чтобы решение имело аппроксимацию $O(h^p)$, а аппроксимация кривизны (20) имела максимально возможный порядок точности. Этот анализ делается стандартным методом разложения схемы (19) и выражения для кривизны (20) по степеням h и сравнением с соответствующими разложениями для точного решения дифференциального уравнения (4).

Таблица 1: Коэффициенты схемы (19) и кривизны (20) для $p = 1$.

c_q	b_s	a_{sq}
-1	1	-
1	-	-

При этом оказывается, что для двухстадийной схемы возможно построить выражение кривизны лишь с первым порядком точности. Однако при этом остается один свободный параметр, выбором которого можно уменьшить коэффициент остаточного члена в кривизне. Для трех- и четырехстадийных схем возможно построить выражение кривизны лишь со вторым порядком точности. Такие ограничения напоминают известные пороги Бутчера для явных схем Рунге-Кутты. Рекомендуемые наборы коэффициентов приведены в табл. 1 – 4.

Таблица 2: Коэффициенты схемы (19) и кривизны (20) для $p = 2$.

c_q	b_s	a_{sq}	
0	0	0	0
-2	1	$1/2$	0
2	-	-	-

Таблица 3: Коэффициенты схемы (19) и кривизны (20) для $p = 3$.

c_q	b_s	a_{sq}		
$2/3$	$2/9$	0	0	0
-2	$1/3$	$1/2$	0	0
$-8/3$	$4/9$	0	$3/4$	-
4	-	-	-	-

Таблица 4: Коэффициенты схемы (19) и кривизны (20) для $p = 4$.

c_q	b_s	a_{sq}			
1	$1/6$	0	0	0	0
-2	$1/3$	$1/2$	0	0	0
-2	$1/3$	0	$1/2$	0	0
0	$1/6$	0	0	1	0
3	-	-	-	-	-

Отметим, что нахождение кривизны по приведенным выше формулам не увеличивает трудоемкости расчетов по схемам Рунге-Кутты.

Неявные схемы. Такие схемы делятся на два класса. Первый – истинно неявные схемы, например, неявные схемы Рунге-Кутты. Алгоритмически они сводятся к решению алгебраических систем нелинейных уравнений. Такие алгебраические системы решают итерационными методами. Простые итерации, как правило, не сходятся, и приходится применять метод Ньютона. Каждая итерация сводится к решению линейной системы того же порядка. При этом алгебраическая система линеаризуется, и на каждой итерации требуется вычислять матрицу Якоби системы и решать систему линейных уравнений. Это трудоемкая процедура.

Второй класс – явно-неявные схемы. Такие схемы предложил Розенброк. Он использовал линеаризацию неявных схем, но ограничился только одной итерацией. Это существенно упростило алгоритм, почти не ухуд-

шая надежности расчетов. Некоторые модификации схем Розенброка были предложены Ваннером.

В формулы как истинно неявных, так и явно-неявных схем входит матрица Якоби. Эту матрицу можно использовать для более точного расчета кривизны. Проведем преобразование формального определения кривизны

$$\varkappa = \frac{d\mathbf{F}}{dl} = \mathbf{F_u}\frac{d\mathbf{u}}{dl} = \mathbf{F_u}\mathbf{F}. \tag{21}$$

Таким образом, кривизна равна произведению матрицы Якоби от правых частей на вектор правых частей. Поскольку в явно-неявных и неявных схемах матрицу Якоби все равно приходится вычислять (это самая трудоемкая часть расчета), то попутное нахождение кривизны по формуле (21) не увеличивает общую трудоемкость вычислений. Поскольку матрица Якоби вычисляется ***до*** выполнения шага, то значение кривизны (21) можно использовать для определения величины текущего шага, а не следующего. В этом заключается качественное преимущество неявных схем перед явными.

5 Алгоритм

Расчет на единственной сетке в принципе не может дать гарантированную оценку погрешности. Единственный способ получения надежной оценки – это расчет на последовательности сгущающихся сеток и сравнение решений на этих сетках по методу Ричардсона. Этот способ дает асимптотически точное значение погрешности. Первоначально этот способ был предложен для равномерных сеток. Впоследствии было показано, что он применим на квазиравномерных сетках [15], [17], [18], а также на кусочно-равномерных и кусочно-квазиравномерных. Поэтому для гарантированной оценки погрешности нужно найти такую последовательность сеток, каждая из которых была бы геометрически-адаптивной, а вся последовательность была бы квазиравномерной.

Построение последовательности сеток. В формуле для шага (16) используются величины, которые до расчета неизвестны: полная длина дуги L и интеграл от кривизны. Построим алгоритм, позволяющий их найти.

Возьмем некоторые начальные не особенно большие значения $N_{\min}$ и $N_{\max}$ и проведем расчет по явной схеме Эйлера, используя для шага формулу (16). Перед началом расчета нам известно полное время T, но длина дуги L и значения интеграла (16) пока неизвестны. Поэтому зададим их «с потолка».

Расчет на первой сетке будем вести до тех пор, пока текущее расчетное время не станет больше либо равным T. В ходе этого расчета найдем полное L и вычислим значение интеграла от кривизны по любой квадратурной формуле. Затем удвоим $N_{\min}$ и $N_{\max}$, воспользуемся уже найденным значением L и интеграла и повторим расчет.

Полученная сетка не будет сгущением первой сетки, так как ее четные узлы не будут совпадать с узлами первой сетки. Далее снова удвоим $N_{\min}$ и $N_{\max}$ и повторим расчет. Такое удвоение будем повторять до тех пор, пока четные узлы новой сетки не окажутся достаточно близкими к узлам предыдущей сетки.

В тестовых расчетах было опробовано несколько критериев близости сеток. Наиболее удачным оказался критерий, в котором сравниваются отношения соответствующих шагов на двух соседних сетках

$$\sqrt{\zeta_n} - 1/\sqrt{\zeta_n}, \quad \zeta_n = \frac{\hat{h}_{2n} + \hat{h}_{2n+1}}{h_n}. \tag{22}$$

Сетки считаются близкими, если мала среднеквадратичная норма этой величины в расчете на один узел

$$D = \sum_{n=0}^{N-1} \left(\sqrt{\zeta_n} - 1/\sqrt{\zeta_n} \right). \tag{23}$$

Выбор схемы. В данной работе мы ограничимся явной схемой Эйлера по следующим причинам. Во-первых, она наименее трудоемка среди всех известных схем. Во-вторых, среди явных схем она наиболее надежна. Ее низкая точность не слишком существенна, поскольку результат расчета нужен только для построения геометрически-адаптивной сетки.

6 Апробация алгоритма

Тестовая задача. Для хорошей верификации данного метода тестовая задача должна обладать следующими свойствами. Во-первых, задача должна содержать параметр, позволяющий варьировать жесткость от малой до очень большой. Во-вторых, она должна иметь точное решение в элементарных функциях, причем как в аргументе «время» $u(t)$, так и в аргументе «длина дуги» $u(l)$, $t(l)$. Желательно также, чтобы и обратные зависимости $l(u)$ и $t(u)$ выражались через элементарные функции. Причина такого сурового требования заключается в том, что только для элементарных функций построены надежные стандартные программы, обеспечивающие вычисление с пренебрежимо малыми ошибками округления.

Тесты, в которых указанные зависимости выражаются через специальные функции, неконструктивны. В программах их расчета нередко значительную роль играют либо погрешности метода вычисления, либо ошибки компьютерного округления. В частности, многие специальные функции вычисляются как решение некоторых задач Коши для дифференциальных уравнений [19], а критику таких подходов мы дали во введении.

В литературе мы не нашли тестов, удовлетворяющих указанным требованиям. Нам удалось построить следующий тест.

$$\frac{du}{dt} = \operatorname{sh}(\lambda u) \tag{24}$$

$$\frac{du}{dl} = \operatorname{th}(\lambda u), \quad \frac{dt}{dl} = \frac{1}{\operatorname{ch}(\lambda u)}. \tag{25}$$

Здесь λ – управляющий параметр, который отвечает за жесткость задачи. Точное решение имеет следующий вид:

$$u(t) = \frac{1}{\lambda}\ln\frac{1+B}{1-B},\ B = e^{\lambda t}\operatorname{th}(\lambda u^0/2);\ t(u) = \frac{1}{\lambda}\ln\frac{\operatorname{th}(\lambda u/2)}{\operatorname{th}(\lambda u^0/2)}; \tag{26}$$

$$u(l) = \frac{1}{\lambda}\ln\left(A+\sqrt{A^2+1}\right),\ A = e^{\lambda t}\operatorname{sh}(\lambda u^0);\ l(u) = \frac{1}{\lambda}\ln\frac{\operatorname{sh}(\lambda u)}{\operatorname{sh}(\lambda u^0)}; \tag{27}$$

$$\begin{gathered} t(l) = \frac{1}{\lambda}\ln\frac{\operatorname{th}\left[(1/2)\ln\left(A+\sqrt{A^2+1}\right)\right]}{\operatorname{th}(\lambda u^0/2)}, \\ l(t) = \frac{1}{\lambda}\ln\frac{\operatorname{sh}\left[\ln(1+B)-\ln(1-B)\right]}{\operatorname{sh}(\lambda u^0)}. \end{gathered} \tag{28}$$

Длина дуги l определяется так, что она равна нулю в начальной точке интегрирования.

Качественный вид решения при $\lambda > 0$ в переменных $u(t)$ приведен на рис. 3. Оно знакопостоянно, стремится к нулю при $t \to -\infty$ и имеет сингулярность при $t = t_* = \lambda^{-1}\ln[\operatorname{cth}(\lambda u_0/2)]$. Кривизна решения определяется следующей формулой:

$$|\varkappa| = \frac{|\lambda \operatorname{sh}(\lambda u)|}{\operatorname{ch}^2(\lambda u)}. \tag{29}$$

Кривизна стремится к нулю как при $t \to -\infty$, так и при $t \to t_*$. В промежутке между ними кривизна достигает максимума

$$\max\varkappa = \frac{\lambda}{2}, \quad t_{extr} = \frac{1}{2\lambda}\ln\left[\frac{1}{3+2\sqrt{2}}\frac{\operatorname{ch}(\lambda u^0)+1}{\operatorname{ch}(\lambda u^0)-1}\right]. \tag{30}$$

Беря достаточно большое λ, можно сделать кривизну сколь угодно большой. Чтобы сделать задачу достаточно трудной, нужно задавать начальное условие заметно левее t_{extr} и заканчивать расчет, несколько не доходя до полюса.

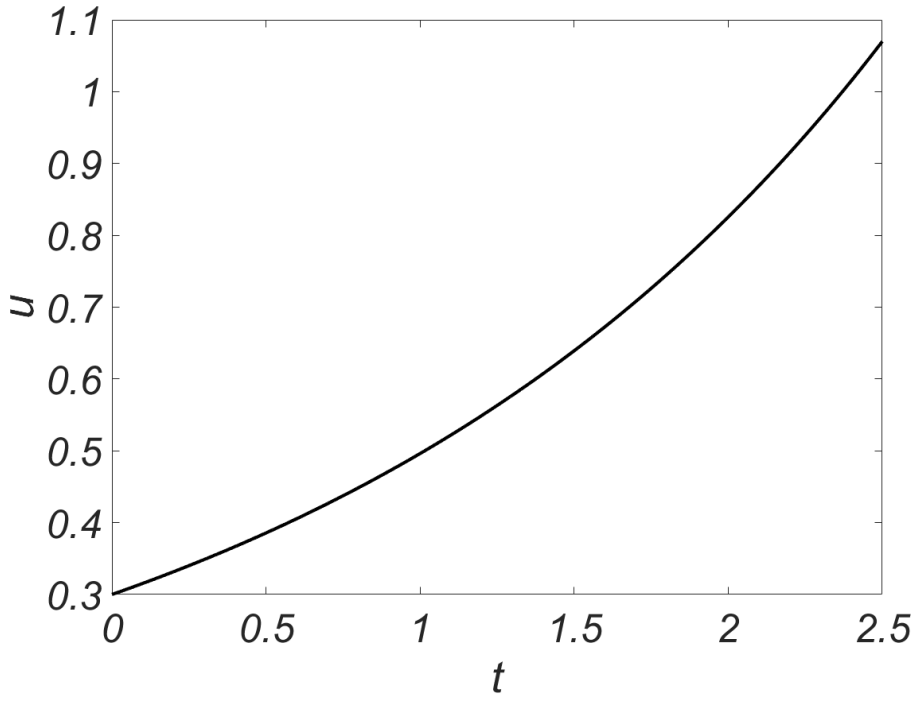

Рис. 3: Решение (26) теста (24), (25).

Был построен еще один тест с решением в элементарных функциях:

$$\frac{du}{dt} = \mathrm{tg}\,(\lambda u) \tag{31}$$

$$\frac{du}{dl} = \frac{\sin(\lambda u)}{\sqrt{\cos 2\lambda u}}, \quad \frac{dt}{dl} = \frac{\cos(\lambda u)}{\sqrt{\cos 2\lambda u}}. \tag{32}$$

$$u(t) = \frac{1}{\lambda}\mathrm{arcsin}\left[e^{\lambda t}\sin(\lambda u^0)\right], \quad t(u) = \frac{1}{\lambda}\ln\frac{\sin(\lambda u)}{\sin(\lambda u^0)}, \tag{33}$$

$$u(l) = \frac{1}{\lambda}2\,\mathrm{arctg}\left[e^{\lambda l}\,\mathrm{tg}\,(\lambda u^0/2)\right], \quad l(u) = \frac{1}{\lambda}\ln\left[\frac{\mathrm{tg}\,(\lambda u/2)}{\mathrm{tg}\,(\lambda u^0/2)}\right], \tag{34}$$

$$\begin{gathered} t(l) = \frac{1}{\lambda}\ln\frac{\sin\left[2\,\mathrm{arctg}\,(e^{\lambda l}\,\mathrm{tg}\,(\lambda u^0/2))\right]}{\sin(\lambda u^0)}, \\ l(t) = \frac{1}{\lambda}\ln\frac{\mathrm{tg}\,\left[(1/2)\,\mathrm{arcsin}\,(e^{\lambda t})\sin(\lambda u^0)\right]}{\mathrm{tg}\,(\lambda u^0/2)}. \end{gathered} \tag{35}$$

Однако он оказался менее представительным: в нем не удается сделать кривизну сколь угодно большой.

Результаты расчетов. Приведем пример расчета теста (25) с $\lambda = 0.5$ и начальными условиями $t^0 = 0$, $u^0 = 0.3$, $l^0 = 0$. Расчеты проводились с аргументом l до $l \approx 5$. Таким образом, начальная и конечная точки лежали по обе стороны от точки максимальной кривизны, причем достаточно далеко. Для первой сетки были выбраны $N_{\min} = 6$, $N_{\max} = 20$. Были проведены расчеты на 14 сетках с последовательным удвоением $N_{\min}$ и $N_{\max}$.

На каждой сетке определялись погрешности в каждом узле как разность численного и точного решений. Затем вычислялась истинная среднеквадратичная погрешность Δ. По парам соседних сеток рассчитывалось значение критерия качества сетки D.

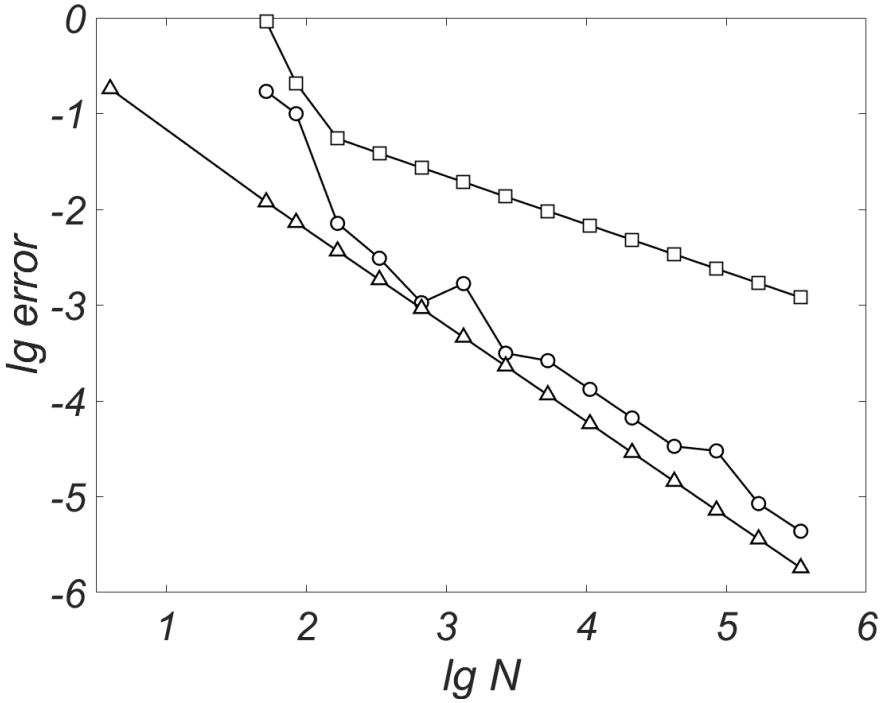

Рис. 4: Погрешности в тесте (24), (25): △ – истинная погрешность, ∘ – оценка (36), □ – критерий качества (23).

Кроме того, по парам сеток рассчитывались разности решений на этих сетках в соответствующих узлах: каждому узлу грубой сетки ставился в соответствие четный узел удвоенной сетки. Такое сопоставление принято делать при классической процедуре определения погрешности при расчете на квазиравномерных сетках, у которых соответствующие узлы имеют строго одинаковое значение координаты. Напомним, что согласно методу Ричардсона такие разности дают асимптотически точную оценку погрешности:

$$\delta_{Rich} = \frac{u - \hat{u}}{2^p - 1}. \tag{36}$$

Здесь u – решение на более грубой сетке, $\hat{u}$ – решение на более подробной сетке, p – порядок точности формулы интегрирования (для формулы Эйлера $p = 1$). В нашем случае соответствующие узлы имеют несколько различающиеся координаты. Тем не менее мы вычисляли погрешность (36) и находили для нее среднеквадратичное значение Δ_{Rich}.

Зависимость всех указанных величин от $N_{\max}$ показана на рис. 4 в двойном логарифмическом масштабе. Проанализируем полученные результаты.

График истинной погрешности на всех рассмотренных сетках образует прямую линию с наклоном -1. Это означает, что истинная погрешность убывает как $O(N^{-1})$ в соответствии с теоретическим порядком использованной схемы Эйлера. Это выполняется даже на очень грубой сетке с $N = 5$, что наглядно иллюстрирует достоинство формулы выбора шага по кривизне. В результате даже такая грубая схема, как схема Эйлера, позволяет получить неплохую точность ~ 0.01 при $N \sim 100$ шагов.

График критерия качества сетки стремится к прямой линии с наклоном -0.5 при возрастании N; такая скорость убывания соответствует закону $O(N^{-1/2})$. Выход на прямую наблюдается при $N \sim 200$. Плавный ход этого графика показывает, что критерий выбран разумно (прочие критерии давали неплавный вид кривой, поэтому они здесь не приводятся). Для хорошего качества сетки нужно, чтобы этот критерий был достаточно мал. При $N \sim 5 \cdot 10^3$ достигается значение критерия $D \sim 0.01$. Это значит, что среднеквадратичная несогласованность соответствующих шагов составляет 1%. Это можно считать удовлетворительным. С этого значения N дальнейшее сгущение сеток дает практически квазиравномерную последовательность.

Оценка погрешности по двум соседним сеткам Δ_{Rich} (36) ведет себя более сложно. Пока критерий $D > 0.1$, график погрешности ложится на прямую с наклоном -0.5. Это хуже порядка точности схемы, но совпадает со скоростью убывания критерия качества сетки. Такой эффект объясняется тем, что заметный вклад в оценку (36) в этом случае вносит несовпадение соответствующих узлов соседних сеток. При $N > 10^2$, когда качество сетки хорошее, наклон графика погрешности (36) выходит на -1. Это соответствует порядку точности схемы Эйлера. Однако на этом участке погрешность (36) превышает истинную погрешность решения. Таким образом, оценка по традиционной формуле Ричардсона является здесь не асимптотической, а мажорантной.

Истинная погрешность, оценка по методу Ричардсона и критерий качества сетки для теста (31), (32) приведены на рис. 5. Полученные результаты аналогичны таковым для теста (24), (25). Однако поскольку тест (31), (32) является более мягким, то кривые оказываются более плавными, чем на рис. 4.

Обсуждение. Любой численный метод должен в конечном итоге реализовываться в прикладные программы, ориентированные на практические применения. Эти программы должны обеспечивать заданную точность при возможно меньшей трудоемкости расчета. Требуемый уровень точности зависит от конкретной прикладной области. Можно условно выделить следующие характерные области.

1) Задачи химической кинетики, где не всегда известна даже система реакций, а скорости важнейших реакций определены с точностью не лучше

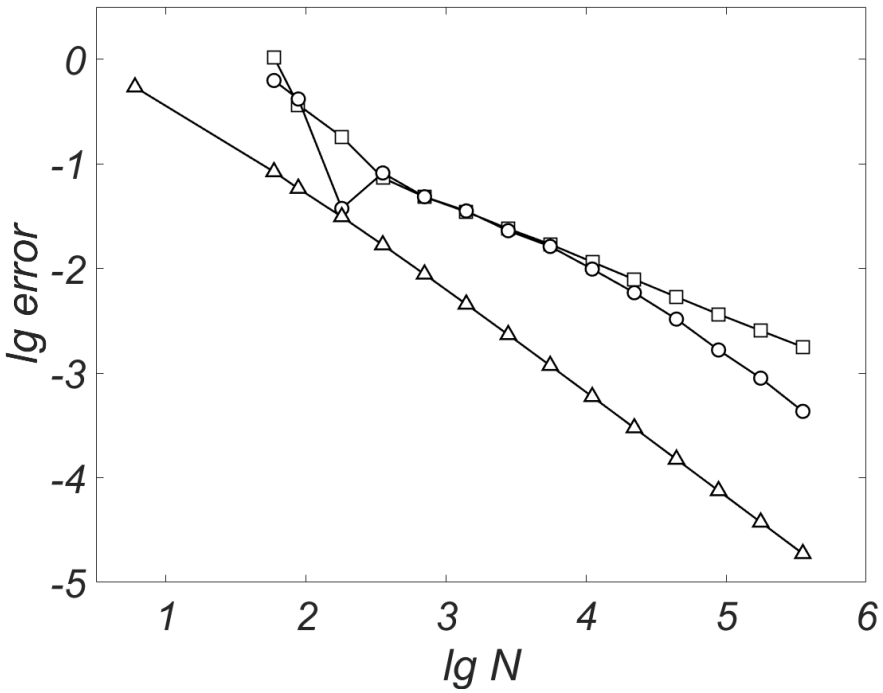

Рис. 5: Погрешности в тесте (31), (32), обозначения соответствуют рис. 4.

$\sim 10\%$ (обычно заметно хуже). В таких задачах достаточно относительной математической точности решения ~ 0.01, то есть на порядок точнее исходных данных.

2) Задачи кинематики механизмов и задачи баллистики, где относительная точность изготовления механизмов может достигать 10^{-4}. Здесь разумно требовать точности расчета $\sim 10^{-5}$.

3) Задачи космической баллистики и небесной механики, где относительные точности измерения орбит достигают $10^{-6} - 10^{-7}$. Соответственно, математическая точность расчета должна быть не хуже $10^{-7} - 10^{-8}$.

4) Эталонные математические расчеты, например, составление таблиц и стандартных программ для специальных функций. Точность расчета таких задач должна быть сопоставима с ошибками компьютерного округления и может достигать 10^{-16}, а в каких-то случаях и выше.

Видно, что использованные в данной работе методы первого порядка точности обеспечивают точность, соответствующую первой группе приложений. Изложенный алгоритм прост и пригоден для составления надежных и экономичных прикладных программ.

Можно использовать предложенный алгоритм совместно с явными схемами более высокого порядка точности, например, явными схемами Рунге-Кутты второго–четвертого порядков. Тогда при сохранении экономичности программ можно рассчитывать на обеспечение второго круга задач.

Чтобы обеспечить уровень точности для третьей и четвертой областей применения, может потребоваться переход к схемам еще более высокого

порядка точности (например, семистадийные явные схемы Рунге-Кутты шестого порядка точности), а также усложнение алгоритма. Описанный здесь алгоритм сгущения сеток следует использовать в качестве первой стадии. Когда будет достигнут критерий качества сетки $\sim 10^{-4}$, расчет первой стадии сгущения следует прекратить. Последняя полученная сетка берется за основу для дальнейшего квазиравномерного сгущения. Это видоизменение алгоритма описано в [10] – [12]. В этом случае формула Ричардсона (36) дает не мажоранту, а асимптотически точное значение погрешности, которое практически неотличимо от истинной погрешности. Разумеется, достижение высокой точности потребует многократного сгущения сеток, так что трудоемкость расчета существенно возрастет.

Работа поддержана грантом РФФИ №18-01-00175.

Список литературы

1. *Хайрер Э., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи. – М.: Мир, 1999.

2. *Виноград Р. Э.* Об одном критерии неустойчивости в смысле Ляпунова решений линейной системы обыкновенных дифференциальных уравнений // Доклады АН СССР – 1952 – Т. 84. – С. 201-204.

3. *Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г.* Численные методы решения жестких систем. М.: Наука, 1979.

4. *Тихонов А. Н.* О зависимости решений дифференциальных уравнений от малого параметра // Математический сборник – 1948. – Т. 22 (64), №2. – С. 193-204.

5. *Васильева А. Б., Бутузов В. Ф.* Асимптотические методы в теории сингулярных возмущений. – М.: Высшая школа, 1990.

6. *Нефедов Н. Н.* Метод дифференциальных неравенств для некоторых сингулярно возмущенных задач в частных производных // Дифференциальные уравнения – 1995 – Т. 31, №4. – С. 719-722.

7. *Белов А. А., Калиткин Н. Н.* Проблема нелинейности при численном решении сверхжестких задач Коши // Математическое моделирование – 2016 – Т. 28, №4 – С. 16-32.

8. *Хайрер Э., Нерсет С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. – М.: Мир, 1990.

9. *Галанин М.П., Конев С.А.* Об одном численном методе решения обыкновенных дифференциальных уравнений // Препринты ИПМ им. М.В. Келдыша – 2017 – №18, 28 с.
http://keldysh.ru/papers/2017/prep2017_18.pdf

10. *Белов А.А., Калиткин Н.Н., Пошивайло И.П.* Геометрически-адаптивные сетки для жестких задач Коши // Доклады Академии наук – 2016 – Т. 466, №3 – С. 276-281.

11. *Белов А.А., Калиткин Н.Н.* Выбор шага по кривизне для жестких задач Коши // Математическое моделирование – 2016 – Т. 28, №11 – С. 97-112.

12. *Белов А.А., Калиткин Н.Н.* Численные методы решения задач Коши с контрастными структурами // Моделирование и анализ информационных систем – 2016 – Т. 23, №5 – С. 528-537.

13. *Riks E.* The application of Newton's method to the problem of elastic stability // Journal of Applied Mechanics – 1972 – Т. 39, №4. – С. 1060-1065.

14. *Шалашилин В.И., Кузнецов Е.Б.* Метод продолжения решения по параметру и наилучшая параметризация. – М.: Эдиториал УРСС, 1999.

15. *Калиткин Н.Н., Корякин П.В.* Численные методы. В 2 кн. Кн. 2. Методы математической физики – М.: Академия, 2013.

16. *Сидоров А.Ф.* Об одном алгоритме расчета оптимальных разностных сеток // Труды МИАН СССР – 1966 – Т. 74 – С. 147-151.

17. *Калиткин Н.Н.* Численные методы. – М.: Наука, 1978.

18. *Калиткин Н.Н., Альшин А.Б., Альшина Е.А., Рогов Б.В.* Вычисления на квазиравномерных сетках. – М.: Физматлит, 2005.

19. NIST Digital Library of Mathematical Functions. https://dlmf.nist.gov

Содержание