МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ 2016 год, том 28, номер 11, стр. 97-112

ВЫБОР ШАГА ПО КРИВИЗНЕ ДЛЯ ЖЕСТКИХ ЗАДАЧ КОШИ

© *2016 г.* ***А.А. Белов[1,2], Н.Н. Калиткин[2]***

[1] Московский государственный университет им. М.В. Ломоносова, физический факультет, Москва
[2] Институт прикладной математики им. М.В. Келдыша РАН, Москва

belov_25.04.1991@mail.ru, kalitkin@imamod.ru

Работа поддержана грантами РФФИ 14-01-00161, 16-31-00062.

Предложен новый метод автоматического выбора шага для численного интегрирования задачи Коши для обыкновенных дифференциальных уравнений. Метод основан на использовании геометрических характеристик (кривизны и наклона) интегральной кривой.
Построены формулы кривизны интегральной кривой при различных выборах многомерного пространства. В двумерном случае они переходят в известные формулы, однако их общий многомерный вид нетривиален. Эти формулы имеют несложный вид, удобны для практического применения и представляют самостоятельный интерес для дифференциальной геометрии многомерных пространств.
Для построенных этим методом сеток разработан способ дробления шагов, позволяющий применить метод Ричардсона и находить апостериорную асимптотически точную оценку погрешности полученного решения (для традиционных алгоритмов автоматического выбора шага не найдено таких оценок). Поэтому предложенные методы существенно превосходят по надежности и достоверности результатов расчетов ранее известные алгоритмы. В существующих автоматах выбора шага наблюдаются резкие уменьшения величины шага на 2-4 порядка без видимых причин. Это ухудшает надежность алгоритмов. Объяснена причина этого явления.
Предлагаемые методы особенно эффективны на задачах высокой жесткости, что проиллюстрировано примерами расчетов.

Ключевые слова: жесткая задача Коши, автоматический выбор шага, кривизна в многомерном пространстве, оценки по методу Ричардсона.

MESH STEP SELECTION BASED ON CURVATURE FOR STIFF CAUCHY PROBLEMS

A.A. Belov[1,2], N.N. Kalitkin[2]

[1] Lomonosov Moscow State University, Faculty of Physics, Moscow
[2] Keldysh Institute of Applied Mathematics of RAS, Moscow

A new method of automatic step construction is proposed for numerical integration of Cauchy problem for ordinary differential equations. The method is based on using of geometrical properties (namely, curvature and slope) of the integration curve.
Formulae for curvature of the integration curve are constructed for different choices of the multi-

dimensional space. In two-dimensional case, they are equivalent to well-known formulae but their general multidimensional form is non-trivial.
For the meshes constructed by our method, a procedure of steps splitting is proposed that allows to apply Richardson method and to calculate a posteriori asymptotically precise error estimation for the obtained solutions. There are no such estimations for traditional automatic step selection algorithms. Consequently, the proposed methods sufficiently excel known before algorithms in reliability and trustworthiness. In existing automatic step algorithms, steps can be unexpectedly reduced by 2-4 orders of magnitude without observable reason. This reduces the algorithms' reliability. We have explained the cause of this phenomenon.
The methods proposed in this work are especially effective on highly stiff problems. This is illustrated by numerous calculations.

Key words: stiff Cauchy problem, automatic step selection, curvature in multidimensional space, Richardson method estimations.

1. Введение

1.1. Понятие жесткости. Жестким задачам посвящена обширная литература, наиболее полный обзор которой дан в [1]. Обычно рассматривают задачу Коши для системы обыкновенных дифференциальных уравнений порядка M

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u},t), \quad 0 \le t \le T; \quad \mathbf{u}(0) = \mathbf{u}^0; \qquad \mathbf{u}(t) = \{u_m(t), 1 \le m \le M\}, \quad \mathbf{f} = \{f_m, 1 \le m \le M\}. \tag{1}$$

Нередко задачу называют жесткой, если велико отношение границ спектра матрицы Якоби $\mathbf{f}_u$. Жесткость в чистом виде встречается редко. Обычно одновременно с жесткостью присутствует плохая обусловленность и, возможно, быстрые осцилляции решения. Однако мы не будем останавливаться на этих тонкостях.

Участки медленного изменения $\mathbf{u}(t)$ называют регулярными, а участки быстрого изменения – пограничными слоями. Пограничные слои, лежащие внутри промежутка интегрирования, называют внутренними пограничными слоями или контрастными структурами. Между каждым пограничным слоем и регулярным участком лежит зона большой кривизны κ интегральной кривой $\mathbf{u}(t)$. Назовем эту зону переходной.

1.2. Методы решения. Есть много работ по аналитическим методам решения жестких задач, собранных в обзоре [2]. Если в задаче (1) можно выделить только один характерный масштаб ширины пограничных слоев $\mu \ll T$, а порядок системы (1) невелик, то эти методы позволяют получить несложное аналитическое приближение к $\mathbf{u}(t)$ и численно его реализовать. При этом они одинаково хорошо передают как начальный, так и внутренние пограничные слои. Однако аналитические методы становятся слишком громоздкими и трудно реализуемыми в задачах с несколькими характерными масштабами, для систем большого порядка или при не слишком малых μ.

Различные численные методы, основанные на сеточном представлении решения, более просты и единообразны. Построены классы специальных схем, ориентированных на жесткие задачи [1]. В расчетах по этим схемам важную роль играет выбор сетки. Он имеет несколько аспектов.

• Расчеты на нелинейных тестах показали, что даже самые надежные схемы требуют, чтобы внутри пограничных слоев и переходных зон содержалось достаточно много шагов сетки. Поэтому расчеты на равномерных сетках возможны лишь при огромном числе узлов $S \gg T/\mu$, то есть неприемлемо трудоемки.

• Для выполнения расчета с гарантированной точностью существует строгая процедура вычислений на последовательности сгущающихся равномерных и квазиравномерных сеток с оценкой погрешности по Ричардсону [3]. Однако расчет на равномерных сетках неприемлемо трудоемок, а для априорного построения квазиравномерных сеток, адаптированных к пограничным слоям (особенно при наличии внутренних пограничных слоев), не найдено никаких алгоритмов.

• Существуют процедуры апостериорного построения сеток, адаптированных к решению. Они основаны на использовании вложенных разностных схем [1]. Однако известные алгоритмы в принципе не позволяют пользоваться методом Ричардсона, а выводимые в компьютерных программах оценки погрешности, которые называют *tolerance*, а не *accuracy* (!), могут быть на порядки больше или меньше фактической погрешности. Кроме того, эти алгоритмы не слишком надежны. Зачастую наблюдаются “срывы”: программа без видимых причин уменьшает шаг на 2-4 порядка, а затем шаг медленно увеличивается до прежнего значения. Такие срывы могут повторяться.

В данной работе предложен другой алгоритм апостериорного выбора шага. Он основан на геометрических характеристиках решения: наклоне и кривизне интегральной кривой. Этот алгоритм прост, надежен, экономичен и позволяет решать системы большого порядка и с различными характерными масштабами. Он хорошо рассчитывает начальный и внутренние пограничные слои в широком диапазоне ε. Одновременно с решением он находит апостериорную асимптотически точную оценку погрешности, чего не могут дать известные алгоритмы выбора шага.

Для реализации этого алгоритма были получены несложные формулы кривизны кривой в пространстве произвольной размерности. В двумерном случае они принимают известный из учебников вид, но их общий многомерный вид нетривиален. Эти формулы могут быть полезны в дифференциальной геометрии многомерных пространств.

2. Тестовая задача

В качестве демонстрационного примера удобно взять задачу с известным точным решением. Тогда можно вычислить фактическую погрешность на любой сетке и сравнить ее с оценкой погрешности по Ричардсону. Ограничимся здесь только одним тестом, который оказался достаточно сложным и содержательным, несмотря на его одномерность:

$$\frac{du}{dt} = -\frac{\lambda(t)(u^2 - a^2)^2}{(u^2 + a^2)}. \tag{2}$$

Точное решение имеет следующий вид:

$$u(t) = -\frac{2\Lambda(t)a^2}{1 + \sqrt{1 + 4a^2\Lambda(t)^2}}, \tag{3}$$

где

$$\Lambda(t) = \frac{u^0}{(u^0)^2 - a^2} + \int_0^t \lambda(\tau) d\tau. \qquad (4)$$

Задача (2) имеет 2 стационарных решения $u = \pm a$. Если $|\lambda(t)| \gg 1$, то задача относится к жестким.

Соответственным выбором $\lambda(t)$ можно получить решение с контрастными структурами. Например, при $\lambda(t) = \lambda_0 \cos t$ и $u^0 = 0$ получим $\Lambda(t) = \lambda_0 \sin t$. Тогда при $\lambda_0 \gg 1$ решение будет попеременно притягиваться к нижнему и верхнему стационарам. Пограничные слои шириной $\sim 1/\lambda_0$ будут располагаться у моментов $t_k = \pi k$. Каждый пограничный слой по существу состоит из двух частей: левая соответствует плохой обусловленности, а правая – жесткости. Положим $T = 6.5$, так что решение содержит 1 начальный и 2 внутренних пограничных слоя (см. рис.1).

В задачах с контрастными структурами решение может подходить очень близко к стационару, сливаясь с ним с точностью до ошибок округления. Чтобы тестирование выявило роль ошибок округления, рекомендуется использовать иррациональные значения a (мы брали $a = \pi$).

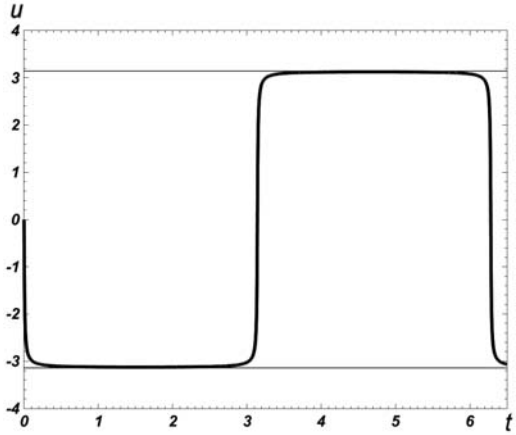

Рис.1. Тест (2) при $\lambda_0 = 30$. Кривая – точное решение, прямые – стационары.

3. Геометрические характеристики кривых

Эвристический алгоритм выбора шага мы строили из следующих соображений. Очевидно, шаг должен быть тем меньше, чем больше наклон интегральной кривой и чем больше ее кривизна. Нам не удалось найти в литературе конструктивных выражений для кривизны в многомерном пространстве. Поэтому мы построили такие выражения для трёх требующихся нам пространств. Приведем их.

3.1. Фазовое пространство. Это пространство состоит только из компонент вектора решения $\{\mathbf{u}(t)\}$ размерности M. В нем ненормированный вектор направления интегральной кривой есть

$$\mathbf{N} = \frac{d\mathbf{u}}{dt} = \mathbf{u}_t. \qquad (5)$$

Его модуль равен $|\mathbf{N}| = \sqrt{(\mathbf{u}_t, \mathbf{u}_t)}$. Поэтому единичный вектор направления имеет вид

$$\mathbf{n} = \frac{\mathbf{u}_t}{\sqrt{(\mathbf{u}_t, \mathbf{u}_t)}}. \tag{6}$$

Как известно, вектор кривизны κ выражается через производную единичного вектора направления $\mathbf{n}$, деленную на $|\mathbf{N}|$. Проводя дифференцирование, получим

$$\kappa = \frac{1}{|\mathbf{N}|}\frac{d\mathbf{n}}{dt} = \frac{\mathbf{u}_{tt}(\mathbf{u}_t, \mathbf{u}_t) - \mathbf{u}_t(\mathbf{u}_{tt}, \mathbf{u}_t)}{(\mathbf{u}_t, \mathbf{u}_t)^2}. \tag{7}$$

При этом нужно учитывать, что в скалярном произведении $(\mathbf{u}_t, \mathbf{u}_t)$ сомножители одинаковы, поэтому достаточно продифференцировать один из них и удвоить результат. Чтобы найти модуль вектора кривизны, скалярно умножим его самого на себя и извлечем квадратный корень

$$|\kappa| = \sqrt{(\kappa, \kappa)} = \sqrt{\frac{(\mathbf{u}_{tt}, \mathbf{u}_{tt})(\mathbf{u}_t, \mathbf{u}_t) - (\mathbf{u}_t, \mathbf{u}_{tt})^2}{(\mathbf{u}_t, \mathbf{u}_t)^3}}. \tag{8}$$

Непосредственным вычислением нетрудно убедиться, что векторы направления и кривизны ортогональны $(\mathbf{n}, \kappa) = 0$.

3.2. Пространство $\{t, \mathbf{u}(t)\}$**.** Это пространство содержит независимую переменную t в качестве первой компоненты. Оно чаще всего используется для представления интегральных кривых. Его размерность есть $M+1$. Ненормированный вектор направления интегральной кривой в нем равен

$$\mathbf{N} = \frac{d\{\mathrm{t}, \mathbf{u}\}}{dt} = \{1, \mathbf{u}_t\}. \tag{9}$$

Его модуль $|\mathbf{N}| = \sqrt{1 + (\mathbf{u}_t, \mathbf{u}_t)}$. Таким образом, единичный вектор направления имеет вид

$$\mathbf{n} = \frac{\{1, \mathbf{u}_t\}}{\sqrt{1 + (\mathbf{u}_t, \mathbf{u}_t)}}. \tag{10}$$

Аналогично предыдущему при помощи несложного дифференцирования можно получить вектор кривизны

$$\kappa = \frac{1}{|\mathbf{N}|}\frac{d\mathbf{n}}{dt} = \frac{\{-(\mathbf{u}_t, \mathbf{u}_{tt}), \mathbf{u}_{tt}[1 + (\mathbf{u}_t, \mathbf{u}_t)] - \mathbf{u}_t(\mathbf{u}_t, \mathbf{u}_{tt})\}}{[1 + (\mathbf{u}_t, \mathbf{u}_t)]^2}. \tag{11}$$

Наконец, модуль вектора кривизны равен

$$|\kappa| = \sqrt{\frac{(\mathbf{u}_{tt}, \mathbf{u}_{tt}) + (\mathbf{u}_{tt}, \mathbf{u}_{tt})(\mathbf{u}_t, \mathbf{u}_t) - (\mathbf{u}_t, \mathbf{u}_{tt})^2}{\left[1 + (\mathbf{u}_t, \mathbf{u}_t)\right]^3}}. \tag{12}$$

Снова перемножением векторов $\mathbf{n}$ и κ можно убедиться в их ортогональности $(\mathbf{n},\kappa)=0$. При $M=1$ решение u имеет только одну компоненту, и под скалярными произведениями понимаются произведения скалярных величин. Поэтому в числителе (12) второй и третий члены сокращаются, и формула принимает известный из учебников вид. Обратный переход от одномерного случая к многомерному нетривиален.

3.3. Переход к длине дуги. Элемент длины дуги равен $dl=\sqrt{1+(\mathbf{u}_t,\mathbf{u}_t)}dt$. Такую замену параметризации нередко применяют для интегрирования жестких систем. При этом возникает пространство $\{l,t,\mathbf{u}\}$ размерности $M+2$. Введем расширенные векторы $\mathbf{u}=\{u_0,u_1,\ldots,u_M;u_0\equiv t\}$, $\mathbf{f}=\{f_0,f_1,\ldots,f_M;f_0\equiv 1\}$. Тогда вместо (1) получим следующую систему размерности $M+1$:

$$\frac{d\mathbf{u}}{dl}=\mathbf{F}(\mathbf{u}),\quad \mathbf{F}=\frac{\mathbf{f}}{\sqrt{(\mathbf{f},\mathbf{f})}}. \tag{13}$$

Запись начальных условий очевидна. В системе (13) вектор правых частей единичный: $(\mathbf{F},\mathbf{F})=1$. Эта система автономна, так как $\mathbf{F}$ не зависит явно от l. Для интегральной кривой в этом $(M+2)$-мерном пространстве можно воспользоваться соотношениями (9)-(12), если сделать замену $\mathbf{u}_t\to\mathbf{U}_l=\mathbf{F}$, $\mathbf{u}_{tt}\to\mathbf{U}_{ll}=(\mathbf{F})_l=\mathbf{F}_\mathbf{U}\mathbf{U}_l=\mathbf{F}_\mathbf{U}\mathbf{F}$.

Тогда ненормированный вектор направления есть $\mathbf{N}=\{1,\mathbf{U}_l\}=\{1,\mathbf{F}\}$ и имеет длину $\sqrt{2}$. Единичный вектор направления равен

$$\mathbf{n}=\frac{1}{\sqrt{2}}\{1,\mathbf{F}\}, \tag{14}$$

Вектор кривизны и его модуль равны соответственно

$$\kappa=\frac{1}{4}\{-(\mathbf{F}_U\mathbf{F},\mathbf{F}),2\mathbf{F}_U\mathbf{F}-(\mathbf{F}_U\mathbf{F},\mathbf{F})\mathbf{F}\}, \tag{15}$$

$$|\kappa|=\frac{1}{2^{3/2}}\sqrt{2(\mathbf{F}_U\mathbf{F},\mathbf{F}_U\mathbf{F})-(\mathbf{F}_U\mathbf{F},\mathbf{F})}. \tag{16}$$

Все выражения (6)-(8), (10)-(12) и (14)-(16) применимы для любого числа измерений и легко реализуются в практических вычислениях. Наиболее трудоемким является вычисление матрицы Якоби $\mathbf{f}_u$ либо $\mathbf{F}_U$. Аналитическое вычисление матрицы Якоби не всегда бывает удобно. Можно заменить первые производные симметричными разностями. Шаг для численного дифференцирования нельзя брать слишком малым, так как начинают сказываться ошибки округления. При работе с 64-битовыми числами ошибка единичного округления составляет 10^{-16} от величины числа. При использовании симметричных разностей рекомендуется брать шаг порядка 10^{-5} от величины решения $\Delta u\sim 10^{-5}u$, что составляет треть от доступного числа разрядов. Этого вполне достаточно, поскольку кривизна используется лишь как вспомогательная величина для вычисления шага.

4. Выбор шага

Было опробовано много вариантов адаптивных сеток для тестовых задач, в которых $T \sim 1$, $\| u \| \sim 1$ (к таким величинам исходную задачу можно привести выбором характерных масштабов). Эти расчеты позволяют предложить следующие эвристические алгоритмы.

4.1. Аргумент l. При этом в интегральной кривой участки пограничного слоя и регулярного решения близки к прямолинейным. Кривизна значительна лишь в переходных зонах, где и нужно сгущать сетку. Обозначив шаг по l через h, выберем начальное значение шага h_*. Построим начальную сетку по закону

$$h_s = \frac{h_*}{1 + (\kappa, \kappa)_s^{1/4}}, \tag{17}$$

где кривизна берется из (15). Качественные соображения для выбора подобной зависимости были упомянуты выше, а количественное значение показателя степени подбиралось в расчетах различных задач.

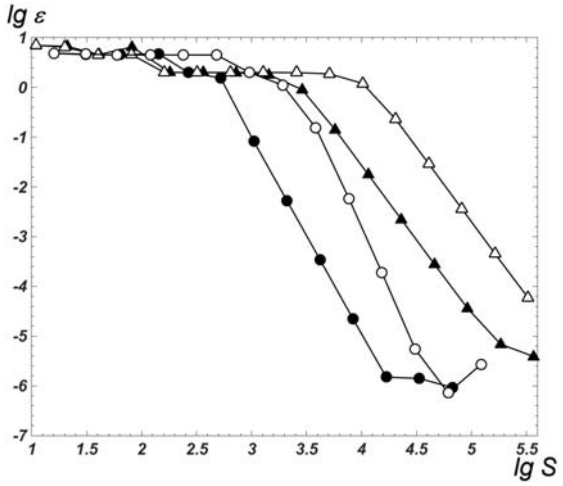

Рис.2. Погрешности на точном решении при $\lambda_0 = 10^5$, аргумент l; схемы: ○ – BORK4, Δ – CROS; темные маркеры – адаптивная сетка (17), светлые маркеры – равномерная сетка $h = \text{const}$.

Влияние знаменателя в (17), отвечающего за сгущение сетки в переходной зоне с большой кривизной, проиллюстрировано на рис.2. На нем показана зависимость погрешности на точном решении от фактического числа узлов S в двойном логарифмическом масштабе. В качестве тестовой выбрана задача (2) с большим $\lambda_0 = 10^5$. Расчеты велись по оптимальной обратной схеме Рунге-Кутты 4-го порядка BORK4 и комплексной схеме Розенброка CROS, свойства которых описаны ниже.

Видно, что на регулярном участке сходимости адаптивная сетка (17) дает на 2–2.5 порядка лучшую точность, чем равномерная сетка $h_s = \text{const}$. Это обусловлено тем, что участок нерегулярной сходимости оказывается более коротким, и раньше начинается быстрое убывание погрешности. На достаточно подробных сетках каждая схема выхо-

дит на фон ошибок округления, который при данном λ_0 составляет $\sim 10^{-6}$. Далее мы рассмотрим это явление подробнее.

Заметим также, что фактический порядок регулярной сходимости для схемы CROS превышает теоретический и равен 3. Схема BORK при использовании адаптивной сетки сходится с 4-м порядком в полном соответствии с теорией. Однако при использовании равномерной сетки эта схема дает ускоренную сходимость с 5-м порядком. Это связано с особенностями данного теста и объясняется тем, что в ошибке аппроксимации член $O(h^4)$ сокращается.

Начальное сгущение. На первом этапе будем уменьшать h_* вдвое и строить последовательность сеток. При этом узлы предыдущей (грубой) сетки довольно близки к четным узлам последующей (подробной) сетки, но точного совпадения нет. Это не позволяет применить метод Ричардсона и дать строгую апостериорную оценку погрешности. Поэтому такой расчет проводится до тех пор, пока соответствующие узлы соседних сеток не окажутся достаточно близкими.

На практике для оценки близости рекомендуется вычислять среднеквадратичное отличие указанных узлов двух соседних сеток:

$$\Delta = \sqrt{\frac{1}{S}\sum_{s=0}^{S}(l_s - \hat{l}_{2s})^2} < \delta. \tag{18}$$

Здесь l_s – узлы грубой сетки, соответствующей управляющему параметру h_*, $\hat{l}_{2s}$ – четные узлы подробной сетки, соответствующей управляющему параметру $h_*/2$. Первый этап сгущения проводится до тех пор, пока величина Δ не станет меньше некоторого δ, задаваемого пользователем. Рекомендуется выбирать $\delta = 10^{-1} \div 10^{-2}$. Влияние этого параметра на результаты расчетов будет продемонстрировано ниже.

Можно написать аналогичный критерий в норме C, но для него момент перехода оказывается примерно таким же, а норма L_2 дает большую стабильность.

Дробление шага. Второй этап начинается с последней сетки, полученной на первом этапе. Далее сгущение ведется по следующему правилу. Все узлы l_s последней сетки берутся в качестве четных узлов следующей сетки, а ее нечетные узлы вычисляются следующим образом. Если имеются 3 соседних шага h_{s-1}, h_s, h_{s+1}, то средний шаг h_s делится нечетным узлом новой сетки в отношении $\sqrt[4]{h_{s+1}/h_{s-1}}$. Тогда шаги подробной сетки

$$\hat{h}_{2s} = h_s \frac{\sqrt[4]{h_{s-1}}}{\sqrt[4]{h_{s+1}} + \sqrt[4]{h_{s-1}}}, \quad \hat{h}_{2s+1} = h_s \frac{\sqrt[4]{h_{s+1}}}{\sqrt[4]{h_{s+1}} + \sqrt[4]{h_{s-1}}}. \tag{19}$$

Если шаг h_{s-1} примыкает к левой границе, то он делится в отношении $\sqrt{h_s/h_{s-1}}$, то есть

$$\hat{h}_0 = h_0 \frac{\sqrt{h_0}}{\sqrt{h_0} + \sqrt{h_1}}, \quad \hat{h}_1 = h_0 \frac{\sqrt{h_1}}{\sqrt{h_0} + \sqrt{h_1}} \tag{20}$$

Процедура вблизи правой границы аналогична:

$$\hat{h}_{2S-1} = h_S \frac{\sqrt{h_{S-1}}}{\sqrt{h_{S-1}} + \sqrt{h_S}}, \quad \hat{h}_{2S} = h_S \frac{\sqrt{h_S}}{\sqrt{h_{S-1}} + \sqrt{h_S}} \tag{21}$$

Квазиравномерная сетка. Описанный алгоритм является простейшим способом построения квазиравномерной сетки. Напомним, что под квазиравномерной понимают сетку, полученную действием гладкой производящей функции на равномерную сетку. При этом разность двух соседних шагов есть величина порядка $O(h^2)$, а при сгущении все сетки строятся при помощи одной и той же производящей функции.

В нашем случае адаптивная сетка (17) строится по кривизне интегральной кривой. На достаточно подробных сетках бесконечно гладкая интегральная кривая передается хорошо, и при дальнейшем сгущении она (как и ее кривизна) меняется мало. Поэтому все последующие адаптивные сетки строятся по практически одной и той же производящей функции. Чем подробнее сетка, тем ближе фактическая производящая функция к предельной (точной) и тем меньше разница узлов текущей сетки и четных узлов следующей сетки. Таким образом, последняя сетка первого этапа практически удовлетворяет определению квазиравномерности.

На втором этапе соответственные узлы двух последовательных сеток точно совпадают, а разность двух соседних шагов

$$\hat{h}_{2s+1} - \hat{h}_{2s} = \hat{h}_{2s+1}\left(1 - \sqrt[4]{h_{s-1}/h_{s+1}}\right) \tag{22}$$

есть величина более высокого порядка малости, чем шаг $\hat{h}$. Поэтому определение квазиравномерности выполняется и здесь.

Метод Ричардсона. Поскольку построенные сетки являются квазиравномерными, то можно применить метод Ричардсона однократно к каждой паре сеток. Это позволяет проводить вычисления с автоматическим выбором шага и гарантированной оценкой погрешности, чего не было во всех ранее существовавших алгоритмах.

Однако пользоваться реккурентным вариантом метода Ричардсона, многократно повышающим порядок точности, в данном случае затруднительно. Это связано с тем, что нечетные узлы подробной сетки $\hat{l}$, построенные по правилам (19)–(21), не точно совпадают с нечетными узлами сетки, построенной по предельной производящей функции. Эта ошибка складывается из двух частей.

Во-первых, шаги h_s грубой сетки известны не точно, а строятся по приближенной производящей функции. Для того чтобы можно было применять рекуррентное уточнение по Ричардсону, погрешность производящей функции должна быть меньше, чем та погрешность, которую мы рассчитываем получить в результате уточнения. Это возможно только на чрезвычайно подробных сетках, поэтому на практике не реализуется.

Во-вторых, сами правила (19)–(21) вносят некоторую погрешность даже при применении к точной производящей функции. Эту часть погрешности можно оценить, пользуясь определением шага квазиравномерной сетки через производную x' производящей функции в полуцелом узле

$$h_s = \frac{1}{S} x'\left(\frac{s-0.5}{S}\right). \tag{23}$$

Запишем выражения для шагов h_{s-1} и h_{s+1}, аналогичные (23), подставим их в выражение для $\hat{h}_{2s}$ из (19) и разложим последнее в ряд по $1/S$ в окрестности точки $(s-0.5)/S$. Получим

$$\hat{h}_{2s} \approx \frac{x'}{2S}\left[1 + \frac{1}{4S}\frac{x''}{x'} + O\left(\frac{1}{S^3}\right)\right], \tag{24}$$

где все производные берутся в точке $(s-0.5)/S$. С другой стороны, точное выражение для этого шага имеет вид

$$\hat{h}_{2s} = \frac{1}{2S} x'\left(\frac{s-0.5}{S} + \frac{1}{4S}\right) \approx \frac{x'}{2S}\left[1 + \frac{1}{4S}\frac{x''}{x'} + \frac{1}{32S^2}\frac{x'''}{x'} + O\left(\frac{1}{S^3}\right)\right]. \tag{25}$$

Сравнивая (24) и (25), найдем ошибку "ручного" дробления шага

$$\Delta\hat{h}_{2s} \approx \frac{1}{64S^3} x''' + O\left(\frac{1}{S^4}\right). \tag{26}$$

О величине x''' можно получить представление, оценивая u'''. Для теста (2) это составляет несколько сотен λ_0. Для большого $\lambda_0 \sim 10^5$ ошибка $\Delta\hat{h}_{2s}$ будет небольшой уже на умеренных сетках с $S \sim 10^3$ при условии, что производящая функция сетки l близка к точной. Это означает, что правила дробления шага (19) – (21) выбраны удачно.

4.2. Аргумент t. В этом случае начальная сетка должна быть подробной не только в переходных зонах, но и в пограничных слоях. Хорошие результаты дает следующая начальная сетка:

$$\tau_s = \frac{\tau_*}{[1+(\mathbf{f},\mathbf{f})_s]^{5/8}[1+(\kappa,\kappa)_s^{1/4}]}. \tag{27}$$

Здесь кривизна берется согласно (11). Под $\mathbf{f}$ и κ понимаются M-мерные векторы. Дальнейшая процедура сгущения аналогична описанной выше.

Влияние различных множителей в (27) проиллюстрировано на рис.3 и 4 для схем BORK4 и CROS соответственно. Здесь по-прежнему $\lambda_0 = 10^5$. Видно, что по отдельности сгущение только в пограничном слое (множитель $[1+(\mathbf{f},\mathbf{f})_s]^{5/8}$) и только в переходной зоне (множитель $1+(\kappa,\kappa)_s^{1/4}$) не приводит к ощутимому повышению точности. Однако наличие обоих множителей улучшает точность на 4 порядка (!) для схемы BORK и на 1-2 порядка для схемы CROS. Заметим также, что для обеих схем фактический порядок регулярной сходимости в полтора раза превышает теоретический: он близок к 6 для BORK4 и с хорошей точностью равен 3 для CROS. Причина этого явления указана выше.

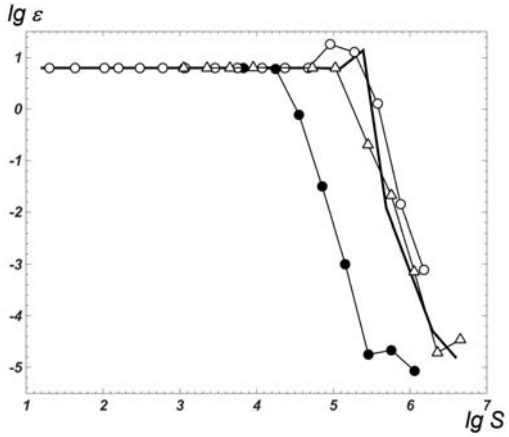

Рис.3. Погрешности на точном решении при $\lambda_0 = 10^5$, аргумент t, схема BORK4; ● – адаптивная сетка (27), Δ – сгущение только в пограничном слое, ○ – сгущение только в переходной зоне, жирная прямая – равномерная сетка $\tau = \text{const}$.

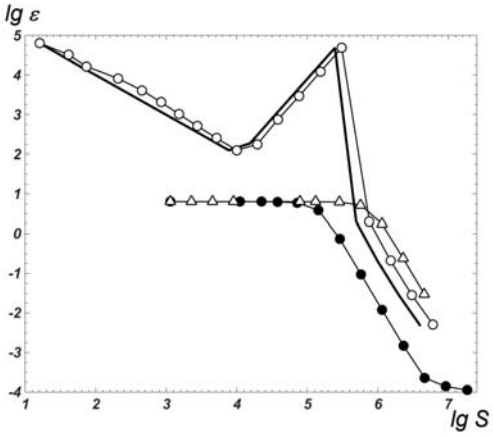

Рис.4. Погрешности на точном решении при $\lambda_0 = 10^5$, аргумент t, схема CROS. Обозначения соответствуют рис.3.

Сам выход на регулярную сходимость наступает лишь на достаточно подробных сетках: на схеме особенно высокой надежности BORK при $S \sim 10^4$, а для надежной (но не сверхнадежной) схемы CROS – даже при $S > 10^5$. Это показывает, что расчет с аргументом t существенно менее надежен, чем с аргументом l, и при его выполнении следует использовать только наиболее надежные схемы.

Поэтому для практических расчетов мы рекомендуем выбирать аргумент l, хотя сами формулы расчетов несколько усложняются.

Сопоставим формулы для шага в аргументах l и t. Множитель, зависящий от кривизны, входит в них одинаково. Глядя на формулу (27), можно подумать, что в (17) можно ввести множитель типа $[1+(\mathbf{f},\mathbf{f})_S]^{1/8}$. Однако на практике это не улучшало расчеты, а иногда даже ухудшало.

5. Сравниваемые методы

Основная масса расчетов велась по следующим схемам.

● Оптимальные обратные схемы Рунге-Кутты (BORK – Backward Optimal Runge-Kutta) были предложены в [4, 5]. Эти схемы привлекательны тем, что они обладают большой надежностью даже на задачах с очень высокой жесткостью. Они записываются через рекурсивные функции. Например, схема порядка $p = 1$ совпадает с обратной схемой Эйлера

$$\hat{\mathbf{u}} = \mathbf{u} + \tau \mathbf{f}(\hat{\mathbf{u}}). \qquad (28)$$

Здесь $\hat{\mathbf{u}}$ есть решение на новом шаге. Схема порядка $p = 2$ имеет вид

$$\hat{\mathbf{u}} = \mathbf{u} + \tau \mathbf{f}(\hat{\mathbf{u}} - \tau/2\ \mathbf{f}(\hat{\mathbf{u}})). \qquad (29)$$

Мы использовали схему порядка $p=4$, обладающую L_4-устойчивостью. Здесь она не приводится ввиду громоздкости. Для всех этих схем имеется пакет прикладных программ [6]. Мы вставляли в него предложенный выше алгоритм выбора шага.

- Схема DOPRI5 широко известна. Это явная схема типа Рунге-Кутты. Она имеет порядок точности $p=5$, и в ней присутствует традиционный алгоритм выбора шага, основанный на вложенных схемах. Хотя эта схема предназначена для нежестких задач, ее нередко применяют для задач умеренной жесткости.
- Программный пакет Гира (GEAR) предназначен для жестких задач. Он содержит набор схем с порядками точности с $p=1$ до $p=5$, которые имеют устойчивость $L_{1/p}$. Пакет снабжен некоторой автоматикой, выбирающей шаг и порядок точности. Он включен в библиотеку MatLab [7] и широко известен.
- В методических расчетах, приведенных на рис.2, 4 использовалась также схема CROS. Это одностадийная схема Розенброка с комплексным коэффициентом. Она имеет 2-й порядок точности и L_2-устойчивость и часто употребляется для жестких задач. Но поскольку ее порядок точности есть лишь $O(h^2)$, то сравнивать ее со схемами высокого порядка точности несправедливо.

Срывы автоматики. Традиционные автоматы выбора шага качественно работают следующим образом. При входе в пограничный слой они выбирают достаточно малые шаги $\tau \sim 1/\lambda$. По мере выхода из пограничного слоя они укрупняют шаг и регулярные участки решения считают со сравнительно крупным шагом. Однако именно на регулярных участках наблюдаются срывы шага, описанные в п.1.2: шаг без видимых причин внезапно уменьшается на 2-4 порядка. После этого автомат снова увеличивает шаг, но срыв шага может повториться, причем неонократно. Это явление описано [1], однако его причина не была объяснена.

Данное явление было исследовано А.А. Болтневым и О.А. Качер в 2005 году. Они обнаружили, что срывы связаны с процессом медленного увеличения шага на регулярных участках. Если вести расчеты регулярных участков равномерным шагом или увеличивать шаг в геометрической прогрессии со знаменателем, очень близким к 1 (~1.001), то срывов не происходит. Однако это не позволяет увеличить шаг до нужных пределов за разумное число шагов. Если же увеличивать шаг с разумным значением знаменателя ~1.1, то срывы возникают довольно часто. Поэтому ситуация напоминает известную поговорку “хвост вытащишь – голова увязнет”.

К сожалению, эта работа Болтнева и Качер не была опубликована.

Описанное явление напоминает потерю устойчивости схемами Гершфельдера-Кертисса (на которых построены программы Гира), если вести расчеты по ним не с постоянным шагом, а с увеличением шага в геометрической прогрессии. Потеря устойчивости происходит при тем меньшем знаменателе, чем выше порядок точности схемы. У схемы порядка точности $O(\tau)$ допустимый знаменатель превышает 2, а у схемы $O(\tau^5)$ он составляет ~1.04.

6. Результаты расчетов

Основная часть расчетов проводилась для аргумента l, так как для жестких задач этот аргумент обеспечивает существенно лучшую точность, чем t. При таком аргументе

решение теста (2) не выражается в элементарных функциях, поэтому погрешность вычислялась следующим образом. Для каждой сетки l_s вычисление по разностной схеме давало приближенные значения u_s и t_s. Решение u_s сравнивалось с точным решением (3), вычисленным в моменты t_s. Их разности давали значения погрешности в узлах, на график выводилась погрешность в норме C в зависимости от числа узлов S в двойном логарифмическом масштабе.

В схеме BORK процедура сгущения сетки автоматически производится нашей программой. В пакетах DOPRI5 и GEAR процедура сгущения сеток не предусмотрена и ввести ее в эти пакеты мы не сумели. Поэтому в них мы задавали последовательно уменьшающиеся значения точности пользователя tol, что приводило к генерации различных сеток с возрастающими S. Погрешность на таком решении вычислялась аналогично.

Кривые погрешности. Приведем типичные расчеты описанного выше теста. На рис.5 показаны погрешности по схеме BORK4 при разных λ_0. При $\lambda_0 = 10$ погрешность уже на грубых сетках убывает в соответствии с теоретическим порядком точности $p = 4$. При $\lambda_0 = 10^2$ кривая начинается с горизонтального участка. Это объясняется тем, что на грубых сетках сеточное решение хорошо описывает первый и второй пограничный слой, но вместо третьего пограничного слоя “срывается”, давая абсурдные результаты. При увеличении λ_0 начальный горизонтальный участок удлиняется. При $\lambda_0 = 10^3$ на первых трех сетках решение срывается на втором пограничном слое, на следующих двух сетках – лишь на третьем пограничном слое, а на дальнейших сетках правильно описывает все 3 слоя. При дальнейшем увеличении λ_0 нерегулярный участок захватывает все большее число сеток, но затем кривые выходят на участок теоретической сходимости.

На достаточно подробных сетках каждая линия выходит на горизонтальный фон ошибок округления. Этот фон весьма низок при $\lambda_0 = 10$, а с увеличением λ_0 довольно быстро возрастает. Фон связан с тем, что чем больше λ_0, тем ближе точное решение подходит к стационарам. Разность между точным решением и стационарным фактически является начальными данными для следующего пограничного слоя. Если в этой разности (с учетом конечной разрядности компьютера) осталось мало достоверных знаков, это ограничивает точность дальнейшего расчета.

Если фон ошибок округления оказался неприемлемо большим, это означает, что данным методом при данной разрядности чисел мы не можем решить задачу. В этом случае нужно либо попробовать более надежную разностную схему (напомним, что наиболее надежными из известных являются схемы BORK), либо провести вычисления с повышенной разрядностью чисел.

Сплошными линиями на рис.5 изображены погрешности, полученные путем прямого сравнения сеточного решения с точным. Наряду с этим на каждой паре соседних сеток (на втором этапе сгущения) проводилась оценка погрешности по Ричардсону, не использующая сравнение с точным решением. Эти оценки показаны кружками. Они отлично совпадают с непосредственным вычислением погрешности. Это показывает, что изложенная в п.4 процедура апостериорной оценки погрешности является правомерной,

то есть предложенный здесь метод одновременно дает и численное решение, и асимптотически точную оценку его погрешности.

Сравнение методов. На рис.6 приведено сравнение кривых погрешности схемы BORK4 с известными стандартными пакетами при достаточно большом $\lambda_0 = 10^5$. Видно, что пакет GEAR дает гораздо более высокий фон ошибок округления. В той области, где его сходимость соответствует теоретической, он близок к BORK4, но сама по себе эта область узка, так что пакет GEAR при такой высокой жесткости мало полезен, хотя он основан на L-устойчивых схемах.

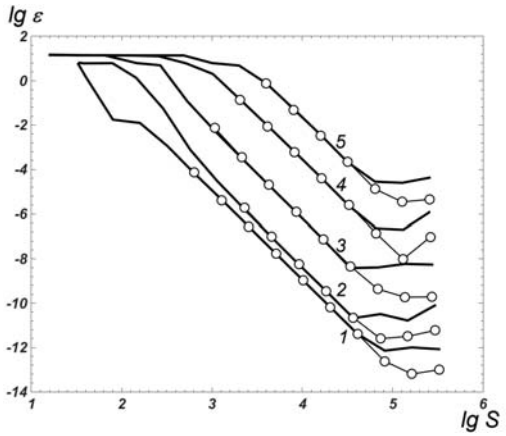

Рис.5. Погрешность BORK4; у линий указаны значения $\lg\lambda_0$. Жирные линии – оценки по точному решению, ○ – по методу Ричардсона.

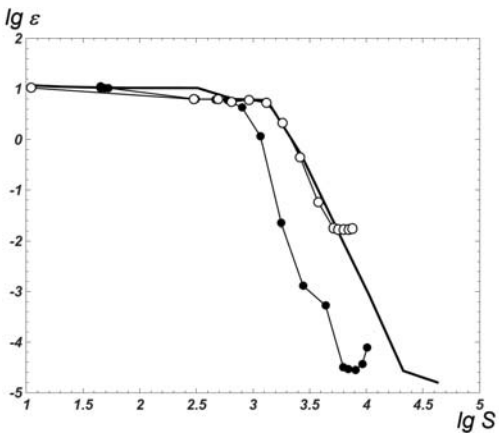

Рис.6. Сходимость разных схем при $\lambda_0 = 10^5$. Жирная линия – BORK4, ○ – GEAR, ● – DOPRI5.

Наоборот, явный метод DOPRI5 (не имеющий даже A-устойчивости) дает тот же фон, что и BORK4 и даже лучшую точность в области теоретической сходимости. Однако это мы смогли увидеть лишь благодаря точному решению. Что же произойдет, если надо считать задачу с неизвестным точным решением?

Представление об этом дает рис.7, на котором построена зависимость истинной погрешности ε от заданной пользователем точности tol при $\lambda_0 = 10^5$. Для нашего метода выбора шага и схемы BORK4 $\varepsilon = tol$ вплоть до выхода на ошибки округления (жирная линия). У пакетов DOPRI5 и GEAR величина ε превышает tol на 9-10 порядков в области теоретической сходимости! Чтобы получить разумную точность ε, приходится выбирать неправдоподобно малое tol (о чем пользователь обычно не подозревает). При этом мы не имеем никаких средств проверки фактической точности (что неоднократно отмечалось в литературе). Это убедительно показывает преимущества предложенного выше способа выбора шага по кривизне и модификации метода Ричардсона для оценки точности.

Критерий перехода. Исследовался вопрос, как влияет на точность выбор параметра δ в критерии перехода от первого этапа сгущения ко второму. На рис.8 показаны кривые погрешности схемы BORK4 при разных δ. Было выбрано $\lambda_0 = 10^3$.

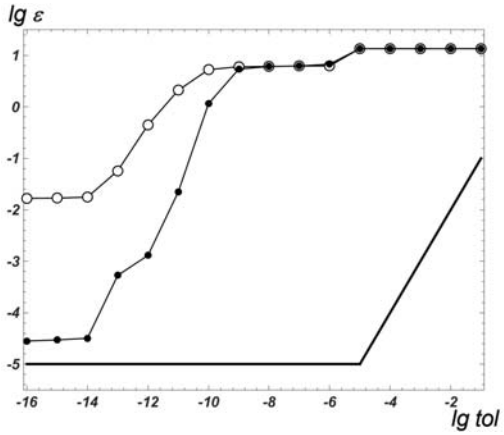

Рис.7. Соотношение *tol* и ε при $\lambda_0 = 10^5$. Жирная линия – BORK4, ○ – GEAR, ● – DOPRI5.

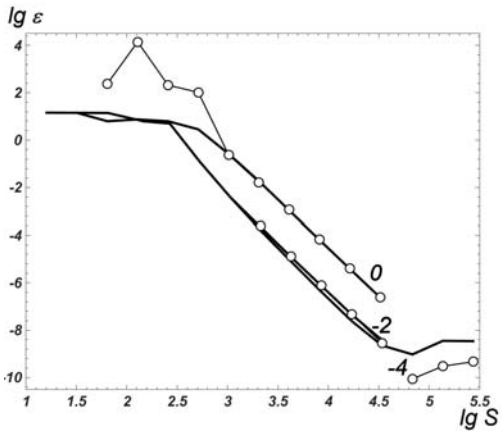

Рис.8. Погрешность BORK4 при разном выборе критерия перехода ко второму этапу сгущения; у линий указаны $\lg\delta$. Обозначения линий и маркеров соответствуют рис.5.

Если $\delta > 1$ велико, то первый этап сгущения прекращается слишком рано, и возможности адаптивной сетки используются не полностью. Это приводит к ощутимой потере точности (в данном случае до 2 порядков). Оценки погрешности по методу Ричардсона начинаются уже при небольших S и попадают на нерегулярный участок кривой сходимости. Однако там метод Ричардсона не является строгим, и оценки могут сильно отличаться от фактической погрешности.

При умеренных $\delta \sim 10^{-1} \div 10^{-2}$ переход от первого этапа сгущения ко второму уверенно попадает на регулярный участок, но не слишком далеко от его начала. Напомним, что на регулярном участке оценки по методу Ричардсона являются строго обоснованными и достоверными.

Наконец, при слишком малых $\delta < 10^{-4}$ оценки погрешности начинаются лишь на очень подробных сетках и приходятся на конец регулярного участка или вовсе попадают на фон ошибок округления. Это не позволяет аккуратно отследить порядок сходимости.

Таким образом, для практики оптимальны умеренные значения $\delta \sim 10^{-1} \div 10^{-2}$, как было указано выше. Завышение или занижение этого параметра нецелесообразно.

Такая рекомендация по выбору δ сохраняется по крайней мере вплоть до $\lambda_0 = 10^5$, что соответствует весьма значительной жесткости. При этом уверенно рассчитываются все 3 пограничных слоя. Разумеется, если потребуется существенно большая жесткость λ_0 или будет необходимо сосчитать большее число пограничных слоев, то в этот критерий придется вносить коррективы.

Авторы выражают благодарность А.Б. Альшину за ценные советы.

СПИСОК ЛИТЕРАТУРЫ

1. *Э. Хайрер, Г. Ваннер.* Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи. – М.: Мир, 1999, 685 с.;
E. Hairer, G. Wanner. Solving ordinary differential equations. Stiff and differential-algebraic prob-

lems / Springer-Verlag. – Berlin, Heidelberg, New York, London, Paris, Tokyo. 1999.

2. *А.Б. Васильева, В.Ф. Бутузов, Н.Н. Нефедов.* Контрастные структуры в сингулярно возмущенных задачах // Фундаментальная и прикладная математика, 1998, т.4, № 3, с.799-851;
   A.B. Vasileva, V.F. Butuzov, N.N. Nefedov. Kontrastnye struktury v singuliarno vozmushchennykh zadachakh // Fundamentalnaia i prikladnaia matematika, 1998, t.4, № 3, s.799-851.
3. *Н.Н. Калиткин, А.Б. Альшин, Е.А. Альшина, Б.В. Рогов.* Вычисления на квазиравномерных сетках. – М.: Физматлит, 2005, 224 с.;
   N.N. Kalitkin, A.B. Alshin, E.A. Alshina, B.V. Rogov. Vychisleniia na kvaziravnomernykh setkah. Fizmatlit, 2005, 224 s.
4. *Н.Н. Калиткин, И.П. Пошивайло.* Обратные Ls-устойчивые схемы Рунге-Кутты // Доклады академии наук, Информатика, 2012, т.442, 2, с.175-180;
   N.N. Kalitkin, I.P. Poshivaylo. Inverse Ls-stable Runge-Kutta schemes // Doklady Mathematics. 2012, v.85, №1, p.139-143.
5. *Н.Н. Калиткин, И.П. Пошивайло.* Вычисления с использованием обратных схем Рунге-Кутты // Математическое моделирование, 2013, т.25, №10, с.79-96;
   N.N. Kalitkin, I.P. Poshivaylo. Computations with inverse Runge–Kutta schemes // Mathematical Models and Computer Simulations, 2014, v.6, № 3, p.272-285.
6. *И.П. Пошивайло.* Жесткие и плохо обусловленные нелинейные модели и методы их расчета. Диссертация на соискание ученой степени кандидата физико-математических наук: 05.13.18. – М.: 2015, 89 с.;
   I.P. Poshivaylo. Zhestkie i plokho obuslovlennye nelineinye modeli i metody ikh rashcheta. Dissertaciia na soiskanie uchenoi stepeni kandidata fiziko-matematicheskikh nauk: 05.13.18. – M.: 2015, 89s.
7. *L.F. Shampine, M.W. Reichelt.* The Matlab ODE suite // SIAM Journal on Scientific Computing. 1997, v.18, № 1, p.1-22.

Поступила в редакцию 26.10.2015.