ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ, 2025, том 61, № 2, с. 242–260

ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.622

К ОЦЕНКЕ ЛОКАЛЬНОЙ ПОГРЕШНОСТИ ЯВНОГО МЕТОДА ЭЙЛЕРА ЧИСЛЕННОГО РЕШЕНИЯ ЗАДАЧИ КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ, ПРЕОБРАЗОВАННЫХ К НАИЛУЧШЕМУ АРГУМЕНТУ

© 2025 г. Е. Б. Кузнецов[1], С. С. Леонов[2]

[1,2]Московский авиационный институт (национальный исследовательский университет)
[2]Российский университет дружбы народов имени Патриса Лумумбы, г. Москва

e-mail: [1]kuznetsov@mai.ru, [2]powerandglory@yandex.ru

Поступила в редакцию 26.06.2024 г., после доработки 18.10.2024 г.; принята к публикации 31.10.2024 г.

Рассмотрен вопрос численного решения задачи Коши для систем обыкновенных дифференциальных уравнений. Особое внимание уделено задачам, имеющим на интегральных кривых предельные особые точки. Известно, что традиционные явные методы решения задачи Коши малоэффективны для указанного класса задач. Неявные же методы многократно сложнее в использовании и не всегда приводят к результату желаемой точности. Поэтому совместно с традиционными методами численного интегрирования задачи Коши применяется метод продолжения решения по наилучшему аргументу (наилучшая параметризация, метод длины дуги), отсчитываемому по касательной вдоль интегральной кривой рассматриваемой задачи. В данной статье для преобразованных к наилучшему аргументу задач Коши приведены результаты исследования локальной погрешности численного решения, полученного явным методом Эйлера, выведена её оценка, с использованием которой найдена верхняя оценка локальной погрешности и доказано уменьшение локальной погрешности решения преобразованной задачи в окрестности предельных особых точек по сравнению с решением исходной задачи. Теоретические результаты согласуются с численным решением плохо обусловленной начальной задачи механики деформируемого твёрдого тела с одной предельной особой точкой.

Ключевые слова: явный метод Эйлера, локальная погрешность, задача Коши, система обыкновенных дифференциальных уравнений, наилучшая параметризация, предельная особая точка

DOI: 10.31857/S0374064125020093, EDN: HVZTKB

ВВЕДЕНИЕ

В настоящее время остаётся актуальной задача создания новых и совершенствования существующих методов численного решения задачи Коши (начальной задачи) для систем обыкновенных дифференциальных уравнений (ОДУ). Это связано с тем, что начальными задачами для таких систем описывается множество процессов практически во всех областях науки и техники, таких как биология, экономика, физика, химическая кинетика, прикладная механика, машиностроение, авиационно-космическая отрасль и др. Решать прикладные задачи аналитически удаётся только в исключительных случаях, поэтому, начиная с середины

XIX века, широкое развитие получили численные методы. Обзор всех значимых работ по численным методам решения задачи Коши для систем ОДУ является отдельной задачей, поскольку их количество на сегодняшний день исчисляется сотнями. По этой причине авторы хотели бы ограничиться лишь ссылками на известные монографии [1, 2], в которых исчерпывающе описаны все основные направления развития численных методов решения задачи Коши в XX веке. Также следует указать монографии Л.М. Скворцова [3] и Е.А. Новикова, Ю.В. Шорникова [4], в которых получили развитие современные направления численного решения жёстких начальных задач как для обыкновенных дифференциальных уравнений, так и для дифференциально-алгебраических уравнений.

При решении жёстких задач большинство явных схем дают только малую или умеренную точность. При расширении области устойчивости во многих случаях явная схема усложняется, сокращается её быстродействие. То же самое относится и к адаптивным явным методам, в которых возникают также трудности подбора структуры численной схемы. Тем не менее явные схемы предпочтительнее неявных, особенно для задач большой размерности. Одним из немногих подходов к решению жёстких и плохо обусловленных задач, сочетающих и быстродействие явных схем, и приемлемую точность, является метод продолжения решения по наилучшему аргументу [5, с. 50–104], называемый также методом наилучшей параметризации или методом длины дуги.

Подробный обзор литературы по методу продолжения решения по параметру и наилучшей параметризации, начиная с 30-х годов XX века, приведён в монографии Э.И. Григолюка и В.И. Шалашилина [6, с. 176–195]. Практика применения метода продолжения решения по параметру в численном анализе восходит к работам М. Лаэя и Д.Ф. Давиденко. Ими впервые была предложена идея замены параметра продолжения решения при численном решении систем алгебраических и трансцендентных уравнений, при этом у Д.Ф. Давиденко система нелинейных уравнений сводится к задаче Коши. В качестве параметров продолжения решения в указанных работах применялись как переменные рассматриваемой системы, так и параметры более общего вида, для которых выполняется условие сохранения ранга системы продолжения решения в окрестности предельных особых точек. Гипотеза о том, что наилучшим будет параметр продолжения решения, отсчитываемый по касательной к кривой множества решения рассматриваемой задачи, впервые сформулирована И.И. Воровичем и В.Ф. Зипаловой. Доказательство данной гипотезы было намечено в статье Э. Рикса, полное же доказательство дано лишь в работе В.И. Шалашилина и Е.Б. Кузнецова [7]. В монографии [5] и последующих работах Е.Б. Кузнецова и его учеников наилучшая параметризация получила дальнейшее развитие, найдя применение при решении начальных и краевых задач для систем уравнений смешанного типа, содержащих дифференциальные, функционально-дифференциальные, алгебраические и интегральные компоненты.

Отметим, что существенным для любого метода численного интегрирования задачи Коши является наличие возможности контроля устойчивости метода и погрешности вычислений. На многочисленных тестовых и прикладных плохо обусловленных начальных задачах показано, что метод продолжения решения по наилучшему аргументу имеет бóльшую устойчивость по сравнению с традиционными явными численными схемами и позволяет получать решение с меньшей погрешностью. В работах [5, с. 64–65] и [8] для скалярного ОДУ доказано, что в окрестности предельных особых точек метод продолжения решения по наилучшему аргументу даёт меньшую погрешность по сравнению с традиционными методами. В данной статье эти результаты будут обобщены на системы ОДУ, будет получена оценка локальной погрешности численного решения для начальных задач, преобразованных к наилучшему аргументу, и доказано уменьшение локальной погрешности решения преобразованной задачи в окрестности предельных особых точек.

Настоящая статья является значительно расширенным вариантом сообщения [9], которое содержит только формулировки основных результатов. Ниже доказываются сформулированные ранее утверждения, даётся верхняя оценка локальной погрешности явного метода Эйлера для преобразованной задачи Коши и приводится пример, демонстрирующий справедливость найденных теоретических оценок.

1. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим нормальную систему ОДУ вида

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}, t), \quad t \in [t_0, T], \tag{1}$$

с начальными условиями

$$\mathbf{y}(t_0) = \mathbf{y}_0, \tag{2}$$

где $\mathbf{y}(t) = (y_1(t), \ldots, y_n(t))^{\mathrm{T}}$ — вектор искомых функций, $\mathbf{f}(\mathbf{y}, t) = (f_1(\mathbf{y}, t), \ldots, f_n(\mathbf{y}, t))^{\mathrm{T}}$ — вектор-функция правой части, $\mathbf{y}_0 = (y_{10}, \ldots, y_{n0})^{\mathrm{T}}$ — вектор значений искомых функций в начальной точке t_0, T — правая граница отрезка изменения аргумента t. Роль аргумента t в прикладных задачах играет время. В дальнейшем будем полагать, что задача (1), (2) удовлетворяет условиям теоремы существования решения.

В общем случае из-за нелинейности правой части системы (1) задачу Коши (1), (2) решить аналитически или невозможно, или затруднительно. По этой причине основными инструментами исследования задачи Коши вида (1), (2) являются численные методы. Для нежёстких начальных задач существуют общие эффективные численные методы, позволяющие получить приближённое решение удовлетворительной точности. Но для жёстких, плохо обусловленных, осциллирующих задач, а также задач с контрастными структурами, общих численных методов, охватывающих широкий класс задач, нет. В монографии [5, с. 50–73] разработан метод, заключающийся в замене исходного аргумента задачи на наилучший аргумент λ, отсчитываемый вдоль интегральной кривой рассматриваемой задачи Коши. Этот метод, названный *наилучшей параметризацией*, показывает свою эффективность применительно к решению плохо обусловленных начальных задач, а также позволяет уменьшить показатель жёсткости для широкого класса жёстких начальных задач.

Для задачи Коши (1), (2) наилучший аргумент λ удовлетворяет следующему дифференциальному соотношению:

$$(d\lambda)^2 = d\mathbf{y}^{\mathrm{T}} d\mathbf{y} + (dt)^2, \tag{3}$$

где $d\mathbf{y} = (dy_1(t), \ldots, dy_n(t))^{\mathrm{T}}$ — дифференциал вектор-функции $\mathbf{y}(t)$.

Будем полагать, что вектор функция $\mathbf{y}$ и аргумент t зависят от наилучшего аргумента λ, определяемого в (3). Тогда задача (1), (2), преобразованная к аргументу λ, примет вид

$$\frac{d\mathbf{y}}{d\lambda} = \frac{1}{\sqrt{Q(\mathbf{y}, t)}}\mathbf{f}(\mathbf{y}, t), \quad \frac{dt}{d\lambda} = \frac{1}{\sqrt{Q(\mathbf{y}, t)}}, \quad \lambda \in [\lambda_0, \Lambda], \tag{4}$$

где λ_0 — начальное значение аргумента λ, Λ — правая граница отрезка изменения аргумента λ. Входящая в знаменатели правых частей уравнений системы (4) функция

$$Q(\mathbf{y}, t) = 1 + \mathbf{f}^{\mathrm{T}}\mathbf{f}. \tag{5}$$

Начальные условия (2) для системы (4) запишутся в форме

$$\mathbf{y}(\lambda_0) = \mathbf{y}_0, \quad t(\lambda_0) = t_0. \tag{6}$$

Замечание 1. Так как система (4) является автономной, т.е. аргумент λ не входит явно в правые части уравнений, то значение λ_0 можно выбирать произвольным образом (для удобства решения). В дальнейшем будем полагать, что $\lambda_0 = 0$.

По сравнению с исходной задачей Коши (1), (2) преобразованная задача (4), (6) имеет ряд преимуществ [5, с. 53–65], а именно:

1) задача Коши (4), (6) имеет наилучшую обусловленность;

2) правые части уравнений системы (4) по модулю не превосходят единицы, так как квадратичная (евклидова) норма правой части системы (4) равна единице;

3) показатель жёсткости системы (4) не больше, чем у системы (1).

Это означает, что для численного решения задачи (4), (6) можно использовать любые методы решения, в том числе и явные. Однако стоит учитывать, что применительно к задаче (4), (6) явные методы будут тем эффективнее, чем хуже обусловлена задача (1), (2) (т.е. имеет предельные особые точки, в которых правая часть системы (1) теряет смысл). Для жёстких задач наилучшая параметризация будет тем эффективнее, чем больше будет норма правой части системы (1).

Ранее в работах [5, с. 64–65; 8] для скалярной задачи Коши вида (1), (2) дана оценка локальной погрешности метода Эйлера для преобразованной к наилучшему аргументу задачи Коши вида (4), (6) в окрестности предельной особой точки. Сформулируем это утверждение.

Теорема 1. *В окрестности предельной особой точки локальные погрешности Δ_t и Δ_λ полученных явным методом Эйлера численных решений скалярных начальных задач вида* (1), (2) *и* (4), (6) *удовлетворяют неравенству*

$$\Delta_\lambda \leqslant \Delta_t, \tag{7}$$

т.е. локальная погрешность Δ_λ численного решения преобразованной начальной задачи не больше, чем локальная погрешность Δ_t численного решения исходной.

Вне окрестностей предельных особых точек неравенство (7) может нарушаться. Также данная оценка остаётся недоказанной для векторного случая. В рамках данной статьи для преобразованной задачи (4), (6) будет дана оценка локальной погрешности, которая обобщит ранее полученную (7) как на любую точку рассматриваемого интервала изменения аргумента, так и на векторный случай.

2. ОЦЕНКА ЛОКАЛЬНОЙ ПОГРЕШНОСТИ РЕШЕНИЯ, ПОЛУЧЕННОГО ЯВНЫМ МЕТОДОМ ЭЙЛЕРА

Сформулируем и докажем основное утверждение статьи.

Теорема 2. *Локальные погрешности Δ_t и Δ_λ полученных явным методом Эйлера численных решений задач* (1), (2) *и* (4), (6) *удовлетворяют неравенству*

$$\Delta_\lambda \leqslant \frac{1}{Q}\big(\|QE - \mathbf{f}\mathbf{f}^{\mathrm{T}}\|_2 + \|\mathbf{f}^{\mathrm{T}}\|_2\big)\Delta_t, \tag{8}$$

где E — единичная матрица порядка n и норма $\|\mathbf{a}\|_2 = \sqrt{a_1^2 + \ldots + a_n^2}$ для $\mathbf{a} = (a_1, \ldots, a_n)$.

Доказательство. Рассмотрим численное решение задач (1), (2) и (4), (6) явным методом Эйлера. Для задачи (1), (2) схема явного метода Эйлера запишется в виде системы рекуррентных соотношений [10, с. 363]

$$\mathbf{y}_{t,j+1} = \mathbf{y}_{t,j} + \tau \mathbf{f}(\mathbf{y}_{t,j}, t_j), \quad j = \overline{0, m-1}, \tag{9}$$

где $\mathbf{y}_{t,j}$ и $\mathbf{y}_{t,j+1}$ — приближённые решения явным методом Эйлера задачи (1), (2) в узловых точках t_j и $t_{j+1} = t_j + \tau$ соответственно, τ — шаг интегрирования по аргументу t, $t_m = T$, $\mathbf{y}_{t,0} = \mathbf{y}_0$.

По аналогии с [10, с. 365–366] получим выражение для локальной погрешности метода Эйлера. Рассмотрим поведение вектор-функции $\mathbf{y}(t)$ в правой полуокрестности точки t_j радиуса τ. Полагая, что функция $\mathbf{y}(t)$ достаточно гладкая, разложим её по формуле Тейлора в окрестности точки t_j до первой степени по τ с записью остаточного члена в форме Лагранжа:

$$\mathbf{y}(t_j+\tau)=\mathbf{y}(t_j)+\tau\frac{d\mathbf{y}}{dt}(t_j)+\frac{\tau^2}{2}\frac{d^2\mathbf{y}}{dt^2}(t_j+\theta_1\tau),\quad \theta_1\in(0,1). \tag{10}$$

Сопоставляя выражения (9) и (10) и учитывая, что $d\mathbf{y}/dt(t_j)=\mathbf{f}(\mathbf{y}(t_j),t_j)$, получаем абсолютную величину локальной погрешности решения задачи (1), (2) в точке t_{j+1} методом Эйлера в виде

$$\Delta_t=\|\mathbf{y}(t_j+\tau)-\mathbf{y}_{t,j+1}\|_2=\frac{\tau^2}{2}\left\|\frac{d^2\mathbf{y}}{dt^2}(t_j+\theta_1\tau)\right\|_2. \tag{11}$$

Для преобразованной задачи (4), (6) схема явного метода Эйлера примет вид

$$\mathbf{y}_{\lambda,k+1}=\mathbf{y}_{\lambda,k}+\frac{l}{\sqrt{Q_{\lambda,k}}}\mathbf{f}(\mathbf{y}_{\lambda,k},t_{\lambda,k}),\quad t_{\lambda,k+1}=t_{\lambda,k}+\frac{l}{\sqrt{Q_{\lambda,k}}},\quad k=\overline{0,p-1},$$

где $\{\mathbf{y}_{\lambda,k},t_{\lambda,k}\}$ и $\{\mathbf{y}_{\lambda,k+1},t_{\lambda,k+1}\}$ — приближённые решения явным методом Эйлера задачи (4), (6) в узловых точках λ_k и $\lambda_{k+1}=\lambda_k+l$ соответственно, l — шаг интегрирования по аргументу λ, $\lambda_p=\Lambda$, $Q_{\lambda,k}=Q(\mathbf{y}_{\lambda,k},t_{\lambda,k})$, $\mathbf{y}_{\lambda,0}=\mathbf{y}_0$, $t_{\lambda,0}=t_0$.

Используя для вектор-функции $\mathbf{y}(\lambda)$ и функции $t(\lambda)$ в правой полуокрестности точки λ_k радиуса l те же рассуждения, получаем оценку абсолютной величины локальной погрешности решения задачи (4), (6) в точке λ_{k+1} методом Эйлера в виде

$$\Delta_\lambda\leqslant\|\mathbf{y}(\lambda_k+l)-\mathbf{y}_{\lambda,k+1}\|_2+\|t(\lambda_k+l)-t_{\lambda,k+1}\|_2=$$
$$=\frac{l^2}{2}\left(\left\|\frac{d^2\mathbf{y}}{d\lambda^2}(\lambda_k+\theta_2 l)\right\|_2+\left\|\frac{d^2t}{d\lambda^2}(\lambda_k+\theta_2 l)\right\|_2\right),\quad \theta_2\in(0,1).$$

С учётом этих результатов найдём оценку локальной погрешности решения преобразованной задачи (4), (6), полученного явным методом Эйлера.

Вначале вычислим локальную погрешность для задачи (1), (2). Для этого найдём вторую производную по времени вектор-функции $\mathbf{y}$:

$$\frac{d^2\mathbf{y}}{dt^2}=\mathbf{f}_{\mathbf{y}}\frac{d\mathbf{y}}{dt}+\mathbf{f}_t, \tag{12}$$

где $\mathbf{f}_t=\partial\mathbf{f}/\partial t$ — частная производная вектор-функции $\mathbf{f}(\mathbf{y},t)$ по переменной t, а производная $\mathbf{f}_{\mathbf{y}}=\partial\mathbf{f}/\partial\mathbf{y}$ является матрицей Якоби вектор-функции $\mathbf{f}(\mathbf{y},t)$ вида

$$\mathbf{f}_{\mathbf{y}}=J_{\mathbf{f}}=\begin{pmatrix}\dfrac{\partial f_1}{\partial y_1} & \dfrac{\partial f_1}{\partial y_2} & \cdots & \dfrac{\partial f_1}{\partial y_n}\\ \dfrac{\partial f_2}{\partial y_1} & \dfrac{\partial f_2}{\partial y_2} & \cdots & \dfrac{\partial f_2}{\partial y_n}\\ \vdots & \vdots & \ddots & \vdots\\ \dfrac{\partial f_n}{\partial y_1} & \dfrac{\partial f_n}{\partial y_2} & \cdots & \dfrac{\partial f_n}{\partial y_n}\end{pmatrix}.$$

Таким образом, соотношение (12) с учётом равенства (1) будет иметь вид

$$\frac{d^2\mathbf{y}}{dt^2}=J_{\mathbf{f}}\mathbf{f}+\mathbf{f}_t, \tag{13}$$

а выражение для локальной погрешности (11) —

$$\Delta_t = \frac{\tau^2}{2}\|J_{\mathbf{f}}\mathbf{f} + \mathbf{f}_t\|_2\Big|_{t=t_j+\theta_1\tau}.$$

Для параметризованной задачи (4), (6) введём обозначения

$$\mathbf{F}(\mathbf{y},t) = G(\mathbf{y},t)\mathbf{f}(\mathbf{y},t), \quad G(\mathbf{y},t) = \frac{1}{\sqrt{Q(\mathbf{y},t)}}.$$

Тогда вторые производные вектор-функции $\mathbf{y}(\lambda)$ и функции $t(\lambda)$ по аргументу λ запишутся как

$$\frac{d^2\mathbf{y}}{d\lambda^2} = \mathbf{F}_{\mathbf{y}}\frac{d\mathbf{y}}{d\lambda} + \mathbf{F}_t\frac{dt}{d\lambda}, \quad \frac{d^2t}{d\lambda^2} = G_{\mathbf{y}}\frac{d\mathbf{y}}{d\lambda} + G_t\frac{dt}{d\lambda}. \tag{14}$$

Вычислим компоненты, входящие во вторые производные:

$$\mathbf{F}_{\mathbf{y}} = \frac{1}{\sqrt{Q}}\mathbf{f}_{\mathbf{y}} - \frac{1}{2\sqrt{Q^3}}\mathbf{f}\frac{\partial Q}{\partial \mathbf{y}}, \quad \mathbf{F}_t = \frac{1}{\sqrt{Q}}\mathbf{f}_t - \frac{1}{2\sqrt{Q^3}}\frac{\partial Q}{\partial t}\mathbf{f},$$
$$G_{\mathbf{y}} = -\frac{1}{2\sqrt{Q^3}}\frac{\partial Q}{\partial \mathbf{y}}, \quad G_t = -\frac{1}{2\sqrt{Q^3}}\frac{\partial Q}{\partial t}.$$

Используя функцию $Q(\mathbf{y},t)$ в виде (5), можно вычислить её первые производные:

$$\frac{\partial Q}{\partial \mathbf{y}} = 2\mathbf{f}^{\mathrm{T}}\mathbf{f}_{\mathbf{y}}, \quad \frac{\partial Q}{\partial t} = 2\mathbf{f}^{\mathrm{T}}\mathbf{f}_t.$$

Подставив все полученные соотношения в формулы (14), окончательно имеем

$$\frac{d^2\mathbf{y}}{d\lambda^2} = \frac{1}{Q^2}(QJ_{\mathbf{f}}\mathbf{f} - \mathbf{f}\mathbf{f}^{\mathrm{T}}J_{\mathbf{f}}\mathbf{f}) + \frac{1}{Q^2}(Q\mathbf{f}_t - \mathbf{f}\mathbf{f}^{\mathrm{T}}\mathbf{f}_t) = \frac{1}{Q^2}(QE - \mathbf{f}\mathbf{f}^{\mathrm{T}})(J_{\mathbf{f}}\mathbf{f} + \mathbf{f}_t),$$
$$\frac{d^2t}{d\lambda^2} = -\frac{1}{Q^2}(\mathbf{f}^{\mathrm{T}}J_{\mathbf{f}}\mathbf{f} + \mathbf{f}^{\mathrm{T}}\mathbf{f}_t) = -\frac{1}{Q^2}\mathbf{f}^{\mathrm{T}}(J_{\mathbf{f}}\mathbf{f} + \mathbf{f}_t).$$

Согласно равенству (13) вторые производные (14) примут вид

$$\frac{d^2\mathbf{y}}{d\lambda^2} = \frac{1}{Q^2}(QE - \mathbf{f}\mathbf{f}^{\mathrm{T}})\frac{d^2\mathbf{y}}{dt^2}(\lambda), \quad \frac{d^2t}{d\lambda^2} = -\frac{1}{Q^2}\mathbf{f}^{\mathrm{T}}\frac{d^2\mathbf{y}}{dt^2}(\lambda).$$

С учётом полученных результатов локальную погрешность решения параметризованной задачи в точке $\lambda = \lambda_{k+1}$ определим как

$$\Delta_\lambda \leqslant \frac{l^2}{2}\frac{1}{Q^2}\left(\left\|(QE - \mathbf{f}\mathbf{f}^{\mathrm{T}})\frac{d^2\mathbf{y}}{dt^2}(\lambda)\right\|_2 + \left\|\mathbf{f}^{\mathrm{T}}\frac{d^2\mathbf{y}}{dt^2}(\lambda)\right\|_2\right) =$$
$$= \frac{l^2}{2}\frac{1}{Q^2}\left(\|QE - \mathbf{f}\mathbf{f}^{\mathrm{T}}\|_2 + \|\mathbf{f}^{\mathrm{T}}\|_2\right)\left\|\frac{d^2\mathbf{y}}{dt^2}(\lambda)\right\|_2 = \left(\frac{l}{\tau}\right)^2\frac{1}{Q^2}\left(\|QE - \mathbf{f}\mathbf{f}^{\mathrm{T}}\|_2 + \|\mathbf{f}^{\mathrm{T}}\|_2\right)\frac{\tau^2}{2}\left\|\frac{d^2\mathbf{y}}{dt^2}(\lambda)\right\|_2. \tag{15}$$

Здесь все функции вычисляются при $\lambda = \lambda_k + \theta_2 l$.

Заметим, что поскольку правая часть второго уравнения системы (4) строго положительна, то существует взаимно-однозначное соответствие между значениями аргументов λ и t. Это означает, что существует пара чисел θ_1 и θ_2 таких, что $t(\lambda_k+\theta_2 l)=t_j+\theta_1\tau$, и можно записать равенство

$$\frac{\tau^2}{2}\left\|\frac{d^2\mathbf{y}}{dt^2}(\lambda_k+\theta_2 l)\right\|_2=\frac{\tau^2}{2}\left\|\frac{d^2\mathbf{y}}{dt^2}(t_j+\theta_1\tau)\right\|_2=\Delta_t$$

и получить оценку (15) в виде

$$\Delta_\lambda\leqslant\left(\frac{l}{\tau}\right)^2\frac{1}{Q^2}\big(\|QE-\mathbf{f}\mathbf{f}^{\mathrm{T}}\|_2+\|\mathbf{f}^{\mathrm{T}}\|_2\big)\Delta_t. \tag{16}$$

Остаётся получить выражение для отношения l/τ как значения в точке $\lambda_k+\theta_2 l$. Используем для этого явный метод Эйлера. При переходе от точки $\lambda=\lambda_k$ к точке $\lambda=\lambda_{k+1}=\lambda_k+l$ имеем

$$t(\lambda_k+l)=t(\lambda_k)+\frac{l}{\sqrt{Q_{\lambda,k}|_{\lambda=\lambda_k}}}.$$

Полагая, что $t(\lambda_k)=t_j$ и $t(\lambda_k+l)=t_j+\tau$, найдём соотношение

$$\frac{l}{\tau}=\sqrt{Q_{\lambda,k}|_{\lambda=\lambda_k}}. \tag{17}$$

При переходе от точки $\lambda=\lambda_k$ к точке $\lambda=\lambda_k+\theta_2 l$ получим

$$t(\lambda_k+\theta_2 l)=t_j+\frac{\theta_2 l}{\sqrt{Q_{\lambda,k}|_{\lambda=\lambda_k}}}.$$

Использование равенств (17) и $t(\lambda_k+\theta_2 l)=t_j+\theta_1\tau$ даёт

$$\frac{\theta_2 l}{\sqrt{Q_{\lambda,k}|_{\lambda=\lambda_k}}}=\theta_1\tau=\frac{\theta_1 l}{\sqrt{Q_{\lambda,k}|_{\lambda=\lambda_k}}},$$

т.е. в линейном приближении справедливо равенство

$$\theta_1=\theta_2. \tag{18}$$

При переходе от точки $\lambda=\lambda_k+\theta_2 l$ к точке $\lambda=\lambda_{k+1}=\lambda_k+l$ получим

$$t(\lambda_k+l)=t(\lambda_k+\theta_2 l+(1-\theta_2)l)=t(\lambda_k+\theta_2 l)+\frac{(1-\theta_2)l}{\sqrt{Q_{\lambda,k}|_{\lambda=\lambda_k+\theta_2 l}}}.$$

В результате найдём соотношение

$$\frac{l}{\tau}=\frac{(1-\theta_1)}{(1-\theta_2)}\sqrt{Q_{\lambda,k}|_{\lambda=\lambda_k+\theta_2 l}},$$

которое с учётом (18) запишем в виде

$$\frac{l}{\tau}=\sqrt{Q_{\lambda,k}|_{\lambda=\lambda_k+\theta_2 l}}. \tag{19}$$

Подставив равенство (19) в (16), получим оценку локальной погрешности решения преобразованной задачи (8). Теорема доказана.

Замечание 2. Нумерация точек, полученных при решении исходной и преобразованной задач, берётся различной (для исходной задачи — индекс j, для преобразованной — k). Это вызвано тем, что количество узловых точек для обеих задач может быть различным.

Замечание 3. Отметим, что при использовании оценки локальной погрешности (8) коэффициент

$$\frac{1}{Q}\left(\|QE-\mathbf{f}\mathbf{f}^{\mathrm{T}}\|_2+\|\mathbf{f}^{\mathrm{T}}\|_2\right)$$

вычисляется в точке $\lambda_k+\theta_2 l$ (или в точке $t_j+\theta_1\tau$). Поскольку значения параметров θ_1 и θ_2 неизвестны в процессе вычисления, то непосредственное использование оценки (8) невозможно. Для приближённого вычисления оценки (8) можно рассчитывать коэффициент не в точке $\lambda_k+\theta_2 l$ (или в точке $t_j+\theta_1\tau$), а в точке λ_k (или в точке t_j). Но более важной информацией из оценки (8) является не её значение в точке, а качественное поведение локальной погрешности на всём рассматриваемом интервале. Этот вопрос подробно рассмотрен в следующем пункте статьи.

3. СЛЕДСТВИЯ ТЕОРЕМЫ 2

В ряде случаев теорему 2 можно конкретизировать, используя оценки входящих в неравенство (8) норм или налагая на размерность задачи и норму правой части исходной системы уравнений дополнительные условия. Рассмотрим далее некоторые полученные следствия. При формулировке утверждений будем полагать, что все функции, входящие в оценку (8), вычисляются в некоторой окрестности точки $M(t_j,\mathbf{y}_j)$ интегральной кривой рассматриваемой задачи Коши.

3.1. ВЕРХНЯЯ ОЦЕНКА ЛОКАЛЬНОЙ ПОГРЕШНОСТИ ПАРАМЕТРИЗОВАННОЙ ЗАДАЧИ

Использовав свойства нормы векторов и матриц [11, с. 123–128], получим верхнюю оценку для локальной погрешности решения преобразованной задачи (4), (6) вида (8).

Следствие 1 (верхняя оценка локальной погрешности). *Локальные погрешности Δ_t и Δ_λ полученных явным методом Эйлера численных решений задач* (1), (2) *и* (4), (6), *соответственно, удовлетворяют верхней оценке*

$$\Delta_\lambda \leqslant R(\|\mathbf{f}\|_2)\Delta_t, \tag{20}$$

где функция $R(\|\mathbf{f}\|_2)$ имеет вид

$$R(\|\mathbf{f}\|_2)=1+\frac{\|\mathbf{f}\|_2^2+\|\mathbf{f}\|_2}{1+\|\mathbf{f}\|_2^2},$$

для неё справедливы оценка

$$1\leqslant R(\|\mathbf{f}\|_2)\leqslant\frac{1}{2}\,\frac{8+5\sqrt{2}}{2+\sqrt{2}}\approx 2{,}2071 \tag{21}$$

и две асимптотики:

$$\begin{aligned} R(\|\mathbf{f}\|_2)&=2+o(1)\quad \textit{при}\ \|\mathbf{f}\|_2\to+\infty,\\ R(\|\mathbf{f}\|_2)&=1+o(1)\quad \textit{при}\ \|\mathbf{f}\|_2\to 0. \end{aligned} \tag{22}$$

Доказательство. Учитывая, что

$$Q = 1 + \mathbf{f}^{\mathrm{T}}\mathbf{f} = 1 + \|\mathbf{f}\|_2^2,$$

оценим первую норму, входящую в (8), с помощью следующей цепочки неравенств:

$$\|QE - \mathbf{f}\mathbf{f}^{\mathrm{T}}\|_2 = \|E + \mathbf{f}^{\mathrm{T}}\mathbf{f}E - \mathbf{f}\mathbf{f}^{\mathrm{T}}\|_2 \leqslant \|E\|_2 + \|\mathbf{f}^{\mathrm{T}}\mathbf{f}\|_2 + \|\mathbf{f}\mathbf{f}^{\mathrm{T}}\|_2 \leqslant 1 + 2\|\mathbf{f}\|_2^2.$$

Здесь для единичной матрицы E используется норма, подчинённая квадратичной векторной норме $\|\|_2$, которая для произвольной квадратной матрицы A порядка m вычисляется по формуле [11, с. 126–128]

$$\|A\|_2 = \max_{1 \leqslant j \leqslant m} \sqrt{\lambda_j},$$

где $\lambda_1, \ \ldots, \ \lambda_m$ — собственные значения матрицы $A^{\mathrm{T}}A$.

Полученные результаты позволяют переписать локальную погрешность (8) в виде

$$\Delta_\lambda \leqslant \left(\frac{1 + 2\|\mathbf{f}\|_2^2 + \|\mathbf{f}\|_2}{1 + \|\mathbf{f}\|_2^2}\right)\Delta_t = \left(1 + \frac{\|\mathbf{f}\|_2^2 + \|\mathbf{f}\|_2}{1 + \|\mathbf{f}\|_2^2}\right)\Delta_t.$$

Заменим норму $\|\mathbf{f}\|_2$ на аргумент $x \in [0, +\infty)$ и рассмотрим функцию

$$R(x) = 1 + \frac{x + x^2}{1 + x^2}.$$

Максимум функции $R(x)$ достигается в точке $x^* = 1 + \sqrt{2} \approx 2{,}4142$ и равен

$$R^* = \frac{8 + 5\sqrt{2}}{4 + 2\sqrt{2}} \approx 2{,}2071.$$

Кроме того, для функции $R(x)$ справедливы две асимптотики:

$$R(x) = 2 + o(1) \quad \text{при } x \to +\infty \quad \text{и} \quad R(x) = 1 + o(1) \quad \text{при } x \to 0.$$

Таким образом, функция $R(x)$ монотонно возрастает от значения $R = 1$ при $x = 0$ до значения R^* при x^*, а затем монотонно стремится к значению 2.

Возвращаясь к исходной переменной, найдём, что оценка для функции $R(\|\mathbf{f}\|_2)$ примет вид (21), а асимптотики для неё — вид (22). Следствие доказано.

Замечание 4. Полученная верхняя оценка локальной погрешности (20) показывает, что погрешность решения преобразованной задачи (4), (6) не всегда меньше, чем погрешность решения исходной задачи (1), (2). При значениях нормы правой части задачи (1), (2) близкой к $1 + \sqrt{2}$ локальная погрешность решения задачи (4), (6) может быть значительно больше по сравнению с исходной задачей. Согласно верхней оценке при близких к нулю значениях нормы правой части задачи (1), (2) локальные погрешности решений обеих задач сопоставимы. При возрастании значения нормы правой части исходной задачи (1), (2) разница между локальными погрешностями обеих задач увеличивается до максимального значения R^* при $\|\mathbf{f}\|_2 = 1 + \sqrt{2} \approx 2{,}4142$, затем монотонно убывает до значения 2 при возрастании этой нормы вплоть до бесконечности. Как будет показано далее, верхняя оценка является

чрезмерно завышенной, в особенности в окрестности предельных особых точек. Однако она даёт понимание, насколько максимально могут отличаться локальные погрешности исходной и преобразованной задач.

3.2. Случай одного обыкновенного дифференциального уравнения

Оценка локальной погрешности для преобразованной задачи (4), (6), данная в следствии 1, является верхней. В ряде случаев её можно уточнить.

Рассмотрим задачу Коши для одного обыкновенного дифференциального уравнения первого порядка

$$\frac{dy}{dt} = f(y,t) \tag{23}$$

с начальным условием

$$y(t_0) = y_0. \tag{24}$$

Преобразованное к наилучшему аргументу λ, дифференциал которого удовлетворяет соотношению

$$(d\lambda)^2 = (dy)^2 + (dt)^2,$$

уравнение (23) запишем как систему

$$\frac{dy}{d\lambda} = \frac{f(y,t)}{\sqrt{1+f^2(y,t)}}, \quad \frac{dt}{d\lambda} = \frac{1}{\sqrt{1+f^2(y,t)}}, \tag{25}$$

а начальное условие (24) для неё примет вид

$$y(0) = y_0, \quad t(0) = t_0. \tag{26}$$

Для начальной задачи (25), (26) справедливо

Следствие 2. *Локальные погрешности Δ_t и Δ_λ полученных явным методом Эйлера численных решений задач* (23), (24) *и* (25), (26), *соответственно, удовлетворяют неравенству*

$$\Delta_\lambda \leqslant R(|f(y,t)|)\Delta_t, \tag{27}$$

где функция $R(|f(y,t)|)$ имеет вид

$$R(|f(y,t)|) = \frac{1+|f(y,t)|}{1+|f(y,t)|^2}$$

и для неё справедливы оценки

$$0 < R(|f(y,t)|) \leqslant 1 \quad \textit{при } |f(y,t)| \geqslant 1,$$

$$1 \leqslant R(|f(y,t)|) \leqslant \frac{1}{\sqrt{2}(2-\sqrt{2})} \approx 1{,}2071 \quad \textit{при } 0 \leqslant |f(y,t)| < 1$$

и асимптотики

$$R(|f(y,t)|) = o(1) \quad \textit{при } |f(y,t)| \to +\infty,$$

$$R(|f(y,t)|) = 1 + o(1) \quad \textit{при } |f(y,t)| \to 0.$$

Доказательство. Использовав неравенство (8), заменив в нём единичную матрицу на единицу, а норму правой части на её модуль, можем записать оценку локальной погрешности решения преобразованной задачи (25), (26) в виде (27).

Заменим абсолютное значение $|f(y,t)|$ на аргумент $x \in [0,+\infty)$ и рассмотрим функцию

$$R(x) = \frac{1+x}{1+x^2}, \quad x \in [0,+\infty).$$

На интервале $0 \leqslant x < 1$ функция $R(x)$ при $x^* = \sqrt{2}-1 \approx 0{,}4142$ достигает глобального максимума $R^* = 1/(\sqrt{2}(2-\sqrt{2})) \approx 1{,}2071$. Также для функции $R(x)$ справедлива асимптотика

$$R(x) = 1 + o(1) \quad \text{при } x \to 0.$$

При $x \geqslant 1$ справедливо неравенство

$$R(x) = \frac{1+x}{1+x^2} \leqslant \frac{1+x^2}{1+x^2} = 1.$$

Кроме того, справедлива асимптотика

$$R(x) = o(1) \quad \text{при } x \to +\infty.$$

Таким образом, функция $R(x)$ монотонно возрастает от значения $R=1$ в точке $x=0$ до максимального значения R^*, а затем монотонно убывает, стремясь к нулю при $x \to +\infty$.

Обратная замена x на $|f(y,t)|$ совместно с полученным неравенством (27) доказывают утверждение следствия.

Замечание 5. В п. 1 отмечалось, что применение наилучшей параметризации тем эффективнее, чем больше норма правой части исходной системы. В следствиях 1 и 2 это утверждение доказано. Если норма правой части исходной системы уравнений принимает близкие к единице значения, то применение наилучшей параметризации будет малоэффективным. Локальная погрешность численного решения преобразованной начальной задачи в этом случае будет выше, чем у решения исходной задачи.

Замечание 6. Результат следствия 2 аналогичен данной в работах [5, с. 64–65; 8] оценке локальной погрешности решения преобразованной задачи вида (25), (26), которая сформулирована в теореме 1. Оценка локальной погрешности следствия 2 обобщает на произвольные точки рассматриваемого интервала изменения аргумента результаты, полученные ранее для окрестности предельной особой точки. Далее даётся обобщение результатов теоремы 1 на случай систем ОДУ.

3.3. Системы обыкновенных дифференциальных уравнений с большими по норме правыми частями

Для практики одним из наиболее важных является класс задач, в которых норма правой части системы (1) принимает больши́е значения, т.е. $\|\mathbf{f}\|_2 \gg 1$. К подобным относятся задачи с пограничным слоем, контрастными структурами и предельными особыми точками. Однако при их численном решении могут возникать вычислительные сложности, связанные с плохой обусловленностью рассматриваемых задач, которые выражаются в необходимости значительного уменьшения шага интегрирования из-за возрастающей погрешности численного решения, полученного по явной схеме. Вследствие этого приходится либо переходить к использованию неявных схем, которые также имеют множество недостатков, либо применять наилучшую параметризацию.

Получим оценку локальной погрешности решения преобразованной задачи (4), (6) для явного метода Эйлера в окрестности предельной особой точки.

Предварительно сформулируем и докажем вспомогательное утверждение.

Лемма. *Для любого натурального значения n определитель n-го порядка*

$$\begin{vmatrix} a_1^2-a_0 & a_1a_2 & a_1a_3 & \dots & a_1a_n \\ a_2a_1 & a_2^2-a_0 & -a_2a_3 & \dots & a_2a_n \\ a_3a_1 & a_3a_2 & a_3^2-a_0 & \dots & a_3a_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_na_1 & a_na_2 & a_na_3 & \dots & a_n^2-a_0 \end{vmatrix} = (-1)^{n-1}a_0^{n-1}\left(\sum_{i=1}^{n} a_i^2 - a_0\right), \tag{28}$$

где $a_0,\ a_1,\ \dots,\ a_n$ — заданные вещественные числа.

Доказательство. Для доказательства данной леммы используем метод математической индукции. Сформируем базу индукции.

Для значения $n=1$ определитель вычисляется тривиально:

$$|a_1^2-a_0| = a_1^2-a_0,$$

для $n=2$

$$\begin{vmatrix} a_1^2-a_0 & a_1a_2 \\ a_2a_1 & a_2^2-a_0 \end{vmatrix} = -a_0(a_1^2+a_2^2-a_0),$$

для $n=3$, применив разложение по последней строке, получим

$$\begin{vmatrix} a_1^2-a_0 & a_1a_2 & a_1a_3 \\ a_2a_1 & a_2^2-a_0 & a_2a_3 \\ a_3a_1 & a_3a_2 & a_3^2-a_0 \end{vmatrix} = a_0^2(a_1^2+a_2^2+a_3^2-a_0).$$

Предположим, что для произвольного натурального n будет справедлива формула (28).

Проверим гипотезу, вычислив определитель $(n+1)$-го порядка, разложив его по последней строке:

$$D_{n+1} = \begin{vmatrix} a_1^2-a_0 & a_1a_2 & a_1a_3 & \dots & a_1a_{n+1} \\ a_2a_1 & a_2^2-a_0 & a_2a_3 & \dots & a_2a_{n+1} \\ a_3a_1 & a_3a_2 & a_3^2-a_0 & \dots & a_3a_{n+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n+1}a_1 & a_{n+1}a_2 & a_{n+1}a_3 & \dots & a_{n+1}^2-a_0 \end{vmatrix}_{n+1} =$$

$$= (-1)^{n-1}(a_{n+1}^2-a_0)a_0^{n-1}\left(\sum_{i=1}^{n} a_i^2 - a_0\right) +$$

$$+(-1)^{n+2}a_{n+1}^2a_1 \begin{vmatrix} a_1a_2 & a_1a_3 & \dots & a_1a_n & a_1 \\ a_2^2-a_0 & a_2a_3 & \dots & a_2a_n & a_2 \\ a_3a_2 & a_3^2-a_0 & \dots & a_3a_n & a_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_na_2 & a_na_3 & \dots & a_n^2-a_0 & a_n \end{vmatrix}_{n} + \dots$$

$$\dots + (-1)^{2n+1}a_{n+1}^2a_n \begin{vmatrix} a_1^2-a_0 & a_1a_2 & \dots & a_1a_{n-1} & a_1 \\ a_2a_1 & a_2^2-a_0 & \dots & a_2a_{n-1} & a_2 \\ a_3a_1 & a_3a_2 & \dots & a_3a_{n-1} & a_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_na_1 & a_na_2 & \dots & a_na_{n-1} & a_n \end{vmatrix}_{n}.$$

Остаётся лишь вычислить полученные n определителей порядка n. Так как все они вычисляются одинаково, продемонстрируем это на примере последнего определителя, обозначив его

$$D_n = a_1\begin{vmatrix} a_1 & a_1a_2 & \dots & a_1a_{n-1} & a_1 \\ a_2 & a_2^2-a_0 & \dots & a_2a_{n-1} & a_2 \\ a_3 & a_3a_2 & \dots & a_3a_{n-1} & a_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_n & a_na_2 & \dots & a_na_{n-1} & a_n \end{vmatrix}_n - a_0\begin{vmatrix} a_2^2-a_0 & a_2a_3 & \dots & a_2a_{n-1} & a_2 \\ a_3a_2 & a_3^2-a_0 & \dots & a_3a_{n-1} & a_3 \\ a_4a_2 & a_4a_3 & \dots & a_4a_{n-1} & a_4 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_na_2 & a_na_3 & \dots & a_na_{n-1} & a_n \end{vmatrix}_{n-1}.$$

Рассмотрим определитель в первом слагаемом. Его первый и последний столбцы совпадают, а значит, он равен нулю. Итоговый определитель по структуре аналогичен исходному, поэтому, повторив проделанные операции еще $n-3$ раз, получим

$$D_n = (-1)^{n-2}a_0^{n-2}\begin{vmatrix} -a_0 & a_{n-1} \\ 0 & a_n \end{vmatrix} = (-1)^{n-1}a_0^{n-1}a_n.$$

Таким же образом можно вычислить и оставшиеся определители, знаки которых будут чередоваться, на что указывает множитель $(-1)^{n-1}$. Таким образом, исходный определитель $(n+1)$-го порядка может быть записан в виде

$$D_{n+1} = (-1)^{n-1}(a_{n+1}^2-a_0)a_0^{n-1}\left(\sum_{i=1}^n a_i^2-a_0\right)+(-1)^n a_0^{n-1}a_{n+1}^2\sum_{i=1}^n a_i^2 =$$

$$= (-1)^n\left[-(a_{n+1}^2-a_0)a_0^{n-1}\left(\sum_{i=1}^n a_i^2-a_0\right)+a_0^{n-1}a_{n+1}^2\sum_{i=1}^n a_i^2\right] =$$

$$= (-1)^n\left[-a_{n+1}^2a_0^{n-1}\left(\sum_{i=1}^n a_i^2-a_0\right)+a_0^n\left(\sum_{i=1}^n a_i^2-a_0\right)+a_0^{n-1}a_{n+1}^2\sum_{i=1}^n a_i^2\right] =$$

$$= (-1)^n\left[a_{n+1}^2a_0^n+a_0^n\left(\sum_{i=1}^n a_i^2-a_0\right)\right] = (-1)^n a_0^n\left(\sum_{i=1}^{n+1} a_i^2-a_0\right).$$

Согласно методу математической индукции формула (28) справедлива для всех натуральных значений n. Лемма доказана.

Теперь докажем

Следствие 3. *В окрестности предельной особой точки локальная погрешность полученного явным методом Эйлера численного решения преобразованной задачи* (4), (6) *не превосходит локальную погрешность полученного явным методом Эйлера численного решения исходной задачи* (1), (2), *т.е.*

$$\Delta_\lambda \leqslant \Delta_t.$$

Доказательство. Рассмотрим окрестность предельной особой точки $M(\mathbf{y}_*(\lambda_*), t_*(\lambda_*))$. Поскольку в ней правая часть системы (1) теряет смысл (стремится к бесконечности), то становятся справедливыми следующие утверждения.

1. Не ограничивая общности доказательства, будем полагать, что существуют N $(N \leqslant n)$ компонент правой части системы (1), которые в окрестности предельной особой точки имеют наибольшую степень роста, т.е.

$$f_1(\mathbf{y}, t), \dots, f_N(\mathbf{y}, t) = O(\|\mathbf{f}\|_2).$$

Здесь и далее, без ограничения общности, будем полагать, что это первые N компонент правой части исходной системы. Все остальные компоненты правой части исходной системы будут бесконечно малыми по сравнению с ними:

$$f_{N+1}(\mathbf{y},t),\ldots,f_n(\mathbf{y},t)=o(\|\mathbf{f}\|_2).$$

2. В окрестности предельной особой точки локальная погрешность вычисления аргумента t преобразованной задачи (4), (6) стремится к нулю, т.е.

$$\frac{l^2}{2}\left\|\frac{d^2t}{d\lambda^2}(\lambda_k+\theta_2 l)\right\|_2=\frac{\|\mathbf{f}\|_2}{1+\|\mathbf{f}\|_2^2}\Delta_t\to 0.$$

Из второго утверждения следует, что вторым слагаемым в оценке локальной погрешности (8) преобразованной задачи можно пренебречь, записав её в виде

$$\Delta_\lambda\leqslant\frac{\|QE-\mathbf{f}\mathbf{f}^{\mathrm{T}}\|_2}{Q}\Delta_t=\left\|E-\frac{\mathbf{f}\mathbf{f}^{\mathrm{T}}}{Q}\right\|_2\Delta_t. \tag{29}$$

Получим оценку нормы $\|E-\mathbf{f}\mathbf{f}^{\mathrm{T}}/Q\|_2$, полагая, что функции наибольшей степени роста в окрестности предельной особой точки удовлетворяют предельным соотношениям

$$\lim_{(\mathbf{y},t)\to(\mathbf{y}_*,t_*)}\frac{f_j(\mathbf{y},t)}{f_1(\mathbf{y},t)}=C_j,\quad j=\overline{1,N},$$

где постоянные C_j принимают конечные ненулевые значения.

Используя всё сказанное выше, распишем норму

$$\left\|E-\frac{\mathbf{f}\mathbf{f}^{\mathrm{T}}}{Q}\right\|_2=\left\|\begin{array}{c|c}B_N & O_{N\times(n-N)}\\ \hline O_{(n-N)\times N} & E_{(n-N)}\end{array}\right\|_2, \tag{30}$$

где $E_{(n-N)}$ — единичная матрица $(n-N)$-го порядка, $O_{N\times(n-N)}$ и $O_{(n-N)\times N}$ — нулевые матрицы размерности $N\times(n-N)$ и $(n-N)\times N$ соответственно, матрица B_N порядка N имеет вид

$$B_N=\begin{pmatrix}\dfrac{\gamma-C_1^2}{\gamma} & -\dfrac{C_1C_2}{\gamma} & -\dfrac{C_1C_3}{\gamma} & \cdots & -\dfrac{C_1C_N}{\gamma}\\ -\dfrac{C_2C_1}{\gamma} & \dfrac{\gamma-C_2^2}{\gamma} & -\dfrac{C_2C_3}{\gamma} & \cdots & -\dfrac{C_2C_N}{\gamma}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ -\dfrac{C_NC_1}{\gamma} & -\dfrac{C_NC_2}{\gamma} & -\dfrac{C_NC_3}{\gamma} & \cdots & \dfrac{\gamma-C_N^2}{\gamma}\end{pmatrix}_N,$$

а параметр

$$\gamma=C_1^2+\ldots+C_N^2. \tag{31}$$

Поскольку матрица $E-\mathbf{f}\mathbf{f}^{\mathrm{T}}/Q$ является симметрической, то её норма будет равна максимальному по модулю собственному значению этой матрицы. Это непосредственно следует из свойства собственных значений: если матрица T имеет собственные значения $\mu_1,\ \ldots,\ \mu_n$, то $g(T)$ имеет собственные значения $g(\mu_1),\ \ldots,\ g(\mu_n)$, где $g(x)$ — произвольная целая рациональная функция от аргумента x [12, с. 19]. Таким образом, норма (30) равняется либо единице, либо максимальному по модулю собственному значению матрицы B_N, если $\mu_{\max}>1$.

Используя доказанную лемму, можно вычислить собственные значения матрицы B_N. Выпишем характеристическое уравнение

$$\begin{vmatrix} \frac{\gamma-C_1^2}{\gamma}-\mu & -\frac{C_1C_2}{\gamma} & -\frac{C_1C_3}{\gamma} & \dots & -\frac{C_1C_N}{\gamma} \\ -\frac{C_2C_1}{\gamma} & \frac{\gamma-C_2^2}{\gamma}-\mu & -\frac{C_2C_3}{\gamma} & \dots & -\frac{C_2C_N}{\gamma} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{C_NC_1}{\gamma} & -\frac{C_NC_2}{\gamma} & -\frac{C_NC_3}{\gamma} & \dots & \frac{\gamma-C_N^2}{\gamma}-\mu \end{vmatrix}_N = 0.$$

Преобразуем полученный определитель, вынеся из каждой его строки множитель $-1/\gamma$:

$$\begin{vmatrix} \frac{\gamma-C_1^2}{\gamma}-\mu & -\frac{C_1C_2}{\gamma} & -\frac{C_1C_3}{\gamma} & \dots & -\frac{C_1C_N}{\gamma} \\ -\frac{C_2C_1}{\gamma} & \frac{\gamma-C_2^2}{\gamma}-\mu & -\frac{C_2C_3}{\gamma} & \dots & -\frac{C_2C_N}{\gamma} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{C_NC_1}{\gamma} & -\frac{C_NC_2}{\gamma} & -\frac{C_NC_3}{\gamma} & \dots & \frac{\gamma-C_N^2}{\gamma}-\mu \end{vmatrix}_N =$$

$$= \frac{(-1)^N}{\gamma^N} \begin{vmatrix} C_1^2-\gamma(1-\mu) & C_1C_2 & C_1C_3 & \dots & C_1C_N \\ C_2C_1 & C_2^2-\gamma(1-\mu) & C_2C_3 & \dots & C_2C_N \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_NC_1 & C_NC_1 & C_NC_3 & \dots & C_N^2-\gamma(1-\mu) \end{vmatrix}_N .$$

Можно видеть, что он полностью удовлетворяет формуле (28) леммы при $n=N$ и $a_1=C_1, \dots, a_N=C_N$, $a_0=\gamma(1-\mu)$. С учётом леммы можно привести характеристическое уравнение к виду

$$-\frac{1}{\gamma}(1-\mu)^{N-1}[C_1^2+\ldots+C_N^2-\gamma(1-\mu)]=0.$$

Применив выражение (31) для γ, окончательно получим

$$-(1-\mu)^{N-1}\mu=0. \tag{32}$$

Корнями характеристического уравнения (32) будут

$$\mu_1=0, \quad \mu_2=\ldots=\mu_N=1,$$

а значит норма

$$\left\| E-\frac{\mathbf{f}\mathbf{f}^{\mathrm{T}}}{Q} \right\|_2 = 1. \tag{33}$$

Подстановка (33) в (29) доказывает утверждение следствия.

Замечание 7. Следствие 3 обобщает теорему 1 на случай систем ОДУ и имеет важное прикладное значение, поскольку доказывает эффективность метода продолжения решения по наилучшему аргументу для плохо обусловленных начальных задач для систем ОДУ с предельными особыми точками.

4. ЧИСЛЕННЫЙ ПРИМЕР

В качестве примера, иллюстрирующего полученные теоретические результаты, приведём задачу одноосного растяжения плоских образцов из титанового сплава ОТ-4 при постоянном напряжении и температуре 500 °C. Для её описания будем использовать определяющие уравнения энергетического варианта теории ползучести [13, с. 28–33]

$$\frac{d\varepsilon}{dt}=\frac{1}{\sigma_0}\frac{dA}{dt}=\frac{K\mathrm{e}^{\beta\sigma_0}}{\sigma_0(A_*-A)^m},\quad \frac{dA}{dt}=\frac{K\mathrm{e}^{\beta\sigma_0}}{(A_*-A)^m}. \tag{34}$$

Здесь ε — деформация ползучести; A и A_* — удельная работа рассеяния (конкретизация параметра повреждённости материала) и её предельное значение в момент разрушения; σ — действующее напряжение; t — время; K, β, m — материальные константы (характеристики ползучести).

В качестве начальных условий для системы (34) примем

$$\varepsilon(0)=0,\quad A(0)=0, \tag{35}$$

т.е. деформация ползучести и удельная работа рассеяния равны нулю.

В момент разрушения скорости деформации ползучести и удельной энергии рассеяния неограниченно возрастают (угол между касательными к кривым ползучести и осью абсцисс близок к прямому в окрестности момента разрушения). Это говорит о том, что момент разрушения является предельной особой точкой.

По результатам экспериментов для уравнений системы (34) были определены характеристики ползучести, которые для температуры $T=500\,^{\circ}\mathrm{C}$ имеют в системе СИ значения [13, с. 32]

$$A_*=88{,}2\ \text{МДж/м}^3,\quad m=3,\quad \beta=0{,}036\ \text{МПа}^{-1},\quad K=0{,}284\ \text{МПа}^4\cdot\text{с}^{-1}.$$

Для начальной задачи (34), (35), выберем наилучший аргумент, удовлетворяющий соотношению

$$(d\lambda)^2=(d\varepsilon)^2+(dA)^2+(dt)^2. \tag{36}$$

Тогда, используя процедуру λ-преобразования по аргументу (36), запишем параметризованную задачу в виде системы

$$\begin{cases}\dfrac{d\varepsilon}{d\lambda}=\dfrac{K\mathrm{e}^{\beta\sigma_0}}{\sqrt{\sigma_0^2(A_*-A)^{2m}+(1+\sigma_0^2)K^2\mathrm{e}^{2\beta\sigma_0}}},\\[2ex] \dfrac{dA}{d\lambda}=\dfrac{\sigma_0K\mathrm{e}^{\beta\sigma_0}}{\sqrt{\sigma_0^2(A_*-A)^{2m}+(1+\sigma_0^2)K^2\mathrm{e}^{2\beta\sigma_0}}},\\[2ex] \dfrac{dt}{d\lambda}=\dfrac{\sigma_0(A_*-A)^m}{\sqrt{\sigma_0^2(A_*-A)^{2m}+(1+\sigma_0^2)K^2\mathrm{e}^{2\beta\sigma_0}}}\end{cases} \tag{37}$$

с однородными начальными условиями

$$\varepsilon(0)=0,\quad A(0)=0,\quad t(0)=0. \tag{38}$$

Задача (37), (38) уже не имеет особенности в момент разрушения.

Рассматриваемая задача (34), (35) допускает аналитическое решение, полученное в работе [14]. Его наличие позволяет оценить погрешность найденных численных решений задачи (34), (35) и её преобразованного к наилучшему аргументу аналога (при использовании

явного метода Эйлера с постоянными шагами интегрирования $\tau = 10^{-1}; 10^{-4}$ — для исходной задачи и $l = \tau; 2\tau$ — для преобразованной). Анализ абсолютной погрешности показывает согласование с теоретическими результатами статьи. В частности, на участке интегральных кривых, предшествующем разрушению, погрешность решения преобразованной задачи не только уравнивается с погрешностью решения исходной задачи, но и становится меньше её. Снижение погрешности решения преобразованной задачи на последнем участке позволяет ближе подойти к моменту разрушения.

ЗАКЛЮЧЕНИЕ

В статье дана теоретическая оценка локальной погрешности полученного явным методом Эйлера численного решения задачи Коши для преобразованной к наилучшему аргументу λ системы обыкновенных дифференциальных уравнений вида (4), (6). Предложен общий вид оценки локальной погрешности (8) вычисленного в некоторой узловой точке явным методом Эйлера численного решения преобразованной задачи (4), (6). С использованием неравенства (8) получена верхняя оценка локальной погрешности (20) и доказано, что локальная погрешность решения задачи (4), (6) в окрестности предельной особой точки не превосходит локальную погрешность решения исходной задачи (1), (2). Представлены результаты, обобщающие ранее доказанные оценки локальной погрешности полученного явным методом Эйлера численного решения задачи Коши для одного обыкновенного дифференциального уравнения из работ [5, с. 64–65; 8].

Теоретические результаты проиллюстрированы на примере решения задачи одноосного растяжения плоских образцов из титанового сплава ОТ-4 при постоянном напряжении и постоянной температуре 500 °C в условиях ползучести. Анализ абсолютной погрешности полученных явным методом Эйлера с постоянным шагом численных решений исходной и преобразованной задач показал согласование с теоретическими результатами.

Полученная для явного метода Эйлера оценка локальной погрешности (8) может быть продолжена и на методы более высокого порядка точности, что планируется сделать в дальнейших исследованиях авторов.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы данной работы заявляют, что у них нет конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. Хайрер, Э. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи / Э. Хайрер, С. Нёрсетт, Г. Ваннер ; пер. с англ. И.А. Кульчицкой, С.С. Филиппова ; под ред. С.С. Филиппова. — М. : Мир, 1990. — 512 с.
2. Хайрер, Э. Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи / Э. Хайрер, Г. Ваннер ; пер. с англ. Е.Л. Старостина, И.А. Кульчицкой, А.В. Тыглияна, С.С. Филиппова ; под ред. С.С. Филиппова. — М. : Мир, 1999. — 685 с.
3. Скворцов, Л.М. Численное решение обыкновенных дифференциальных и дифференциально-алгебраических уравнений / Л.М. Скворцов. — М. : ДМК-Пресс, 2018. — 230 с.
4. Новиков, Е.А. Моделирование жестких гибридных систем / Е.А. Новиков, Ю.В. Шорников. — СПб. : Лань, 2019. — 420 с.
5. Шалашилин, В.И. Метод продолжения решения по параметру и наилучшая параметризация в прикладной математике и механике / В.И. Шалашилин, Е.Б. Кузнецов. — М. : Эдиториал УРСС, 1999. — 224 с.

6. Григолюк, Э.И. Метод продолжения решения по параметру в нелинейных задачах механики твёрдого деформируемого тела / Э.И. Григолюк, В.И. Шалашилин. — М. : Наука, 1988. — 232 с.
7. Кузнецов, Е.Б. Задача Коши как задача продолжения по наилучшему параметру / Е.Б. Кузнецов, В.И. Шалашилин // Дифференц. уравнения. — 1994. — Т. 30, № 6. — С. 964–971.
8. Некоторые количественные оценки эффективности преобразования задачи Коши для дифференциальных уравнений к наилучшему аргументу / А.Н. Данилин, Н.Н. Зуев, Е.Б. Кузнецов, В.И. Шалашилин // Журн. вычислит. математики и мат. физики. — 1999. — Т. 39, № 7. — С. 1134–1141.
9. Кузнецов, Е.Б. К оценке локальной погрешности численного решения параметризованной задачи Коши / Е.Б. Кузнецов, С.С. Леонов // Успехи мат. наук. — 2022. — Т. 77, № 3 (465). — С. 171–172.
10. Бахвалов, Н.С. Численные методы / Н.С. Бахвалов, Н.П. Жидков, Г.М. Кобельков. — М. : Лаборатория Базовых Знаний, 2002. — 632 с.
11. Амосов, А.А. Вычислительные методы для инженеров / А.А. Амосов, Ю.А. Дубинский, Н.В. Копченова. — М. : Высшая школа, 1994. — 544 с.
12. Курант, Р. Методы математической физики. Т. 1 / Р. Курант, Д. Гильберт. — М., Л. : Государственное технико-техническое издательство, 1933. — 525 с.
13. Соснин, О.В. Энергетический вариант теории ползучести / О.В. Соснин, Б.В. Горев, А.Ф. Никитенко. — Новосибирск : Институт гидродинамики СО АН СССР, 1986. — 96 с.
14. Описание процесса ползучести и разрушения современных конструкционных материалов с использованием кинетических уравнений в энергетической форме / Б.В. Горев, И.В. Любашевская, В.А. Панамарев, С.В. Иявойнен // Прикл. механика и техн. физика. — 2014. — Т. 55, № 6. — С. 132–144.

ON THE ESTIMATION OF THE EXPLICIT EULER METHOD LOCAL ERROR FOR THE NUMERICAL SOLUTION OF THE CAUCHY PROBLEM FOR ORDINARY DIFFERENTIAL EQUATIONS TRANSFORMED TO THE BEST ARGUMENT

© 2025 / E. B. Kuznetsov[1], S. S. Leonov[2]

[1,2]Moscow Aviation Institute (National Research University), Russia
[2]Peoples' Friendship University of Russia named after Patrice Lumumba, Moscow, Russia
e-mail: [1]kuznetsov@mai.ru, [2]powerandglory@yandex.ru

The paper considers the numerical solution of the Cauchy problem for systems of ordinary differential equations. Special attention is paid to problems with limiting singular points on integral curves. It is known that traditional explicit methods for solving the Cauchy problem are ineffective for this class of problems. Implicit methods are much more difficult to use and do not always lead to the result of the desired accuracy. Therefore, along with traditional methods of numerical integration of the Cauchy problem authors use the method of solution continuation with respect to the best argument (also known as the best parameterization and the arc length method). The best argument is calculated tangentially along the integral curve of the problem under consideration. For the Cauchy problems transformed to the best argument, the authors in this paper present the results of a study of the local error for the numerical solution obtained by the explicit Euler method. An estimate of the numerical solution local error of the numerical solution for the Cauchy problem transformed to the best argument is obtained for the explicit Euler method. Using it, an upper estimate of the local error was obtained and the effectiveness of using the best argument was proved. This is reflected in a decrease of the solution local error for the transformed problem in the neighborhood of the limiting singular points. The theoretical results are compatible with the numerical solution of the ill-conditioned initial value problem of deformable solid mechanics with one limiting singular point.

Keywords: explicit Euler method, local error, Cauchy problem, system of ordinary differential equations, best parameterization, limiting singular point

REFERENCES

1. Hairer, E., Nørsett, S.P., and Wanner, G., *Solving Ordinary Differential Equations. I: Nonstiff Problems*, Berlin: Springer-Verlag, 1987.
2. Hairer, E. and Wanner, G., *Solving Ordinary Differential Equations. II: Stiff and Differential-Algebraic Problems*, Berlin: Springer-Verlag, 1996.
3. Skvortsov, L.M., *Chislennoye resheniye obyknovennykh differentsial'nykh i differentsial'no-algebraicheskikh uravneniy* (Numerical Solution of Ordinary Differential and Differential-Algebraic Equations), Moscow: DMK-Press, 2018.
4. Novikov, E.A. and Shornikov, Yu.V., *Modelirovaniye zhestkikh gibridnykh sistem* (Modeling of Rigid Hybrid Systems), Saint Petersburg: Lan', 2019.
5. Shalashilin, V.I. and Kuznetsov, E.B., *Parametric Continuation and Optimal Parametrization in Applied Mathematics and Mechanics*, Dordrecht, Boston, London: Kluwer Academic Publishers, 2003.
6. Grigolyuk, E.I. and Shalashilin, V.I., *Problems of Nonlinear Deformation: The Continuation Method Applied to Nonlinear Problems in Solid Mechanics*, Dordrecht: Kluwer Academic Publishers, 1991.
7. Kuznetsov, E.B. and Shalashilin, V.I., The Cauchy problem as a problem of continuation with respect to the best parameter, *Differ. Equat.*, 1994, vol. 30, no. 6, pp. 893–898.
8. Danilin, A.N., Zuev, N.N., Kuznetsov, E.B., and Shalashilin, V.I., Some numerical efficiency estimates for the transformation of the Cauchy problem for differential equations to the best argument, *Comput. Math. Math. Phys.*, 1999, vol. 39, no. 7, pp. 1092–1099.
9. Kuznetsov, E.B. and Leonov S.S., On estimating the local error of a numerical solution of a parametrized Cauchy problem, *Russ. Math. Surv.*, 2022, vol. 77, no. 3, pp. 543–545.
10. Bakhvalov, N.S., Zhidkov, N.P., and Kobel'kov, G.M., *Chislennyye metody* (Numerical methods), Moscow: Laboratoriya Bazovykh Znaniy, 2002.
11. Amosov, A.A., Dubinskii, Yu.A., and Kopchenova, N.V., *Vychislitel'nyye metody dlya inzhenerov* (Computational Methods for Engineers), Moscow: Vysshaya shkola, 1994.
12. Courant, R. and Hilbert, D., *Methods of Mathematical Physics. Vol. 1*, New York, London, Sydney: Wiley & Sons, 1953.
13. Sosnin, O.V., Gorev, B.V., and Nikitenko, A.F., *Energeticheskiy variant teorii polzuchesti* (Energy Variant of the Creep Theory), Novosibirsk: Institut gidrodinamiki SO AN SSSR, 1986.
14. Gorev, B.V., Lyubashevskaya, I.V., Panamarev, V.A., and Iyavoynen, S.V., Description of creep and fracture of modern construction materials using kinetic equations in energy form, *J. Appl. Mech. Tech. Phys.*, 2014, vol. 55, no. 6, pp. 1020–1030.