МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ 2002г., том 14, номер 2, стр. 3-17

ДИАГОНАЛЬНО НЕЯВНЫЕ FSAL-МЕТОДЫ РУНГЕ-КУТТЫ ДЛЯ ЖЕСТКИХ И ДИФФЕРЕНЦИАЛЬНО-АЛГЕБРАИЧЕСКИХ СИСТЕМ

© *Л.М. Скворцов*

Московский государственный технический университет им. Н.Э. Баумана

Рассматриваются неявные методы Рунге-Кутты, первая стадия которых совпадает с последней стадией предыдущего шага. Предложены методы порядка 3, 4, 5. Показано преимущество этих методов по сравнению с однократно диагонально неявными методами Рунге-Кутты.

DIAGONALLY IMPLICIT RUNGE-KUTTA FSAL METHODS FOR STIFF AND DIFFERENTIAL-ALGEBRAIC SYSTEMS

L.M. Skvortsov

Moscow State Technical University

Implicit Runge-Kutta methods are considered which first stage coincides with the last stage of previous step. The methods of order 3, 4, 5 are proposed. Advantage of these methods in comparison with singly diagonally implicit Runge-Kutta methods is shown.

1. Введение

Будем рассматривать задачу Коши для системы обыкновенных дифференциальных уравнений

$$y' = f(t,y), \quad y(t_0) = y_0. \tag{1}$$

Один шаг численного решения задачи (1) s-стадийным методом Рунге-Кутты в общем случае задается расчетными формулами

$$y_1 = y_0 + h\sum_{i=1}^{s} b_i k_i; \quad k_i = f(t_0 + c_i h, g_i), \quad g_i = y_0 + h\sum_{j=1}^{s} a_{ij} k_j, i = \overline{1,s}. \tag{2}$$

Часто приводят еще одну формулу

$$\hat{y}_1 = y_0 + h\sum_{i=1}^{s} \hat{b}_i k_i ,$$

которая используется для получения оценки погрешности численного решения $e = y_1 - \hat{y}_1$. Конкретный метод Рунге-Кутты определяется набором коэффициентов a_{ij}, b_i, c_i и может быть представлен в виде таблицы

$$\begin{array}{c|c} c & A \\ \hline & b^{\mathrm{T}} \end{array} = \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \cdots & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}$$

(в случае метода с оцениванием погрешности добавляется строка коэффициентов $\hat{b}_i$).

Для решения жестких и дифференциально-алгебраических систем обычно используют неявные методы. Среди неявных методов Рунге-Кутты наиболее просто реализуются диагонально неявные (DIRK – Diagonally Implicit Runge-Kutta) [1-3], у которых матрица A имеет нижнюю треугольную форму. Обычно также требуют, чтобы все диагональные элементы матрицы A были равны между собой, что позволяет выполнять только одно LU-разложение на шаге интегрирования. Такие методы называют однократно диагонально неявными (SDIRK – Singly DIRK). Таблица коэффициентов метода SDIRK имеет вид

$$\begin{array}{c|cccc} c_1 & \gamma & & & \\ c_2 & a_{21} & \gamma & & \\ \vdots & \vdots & \vdots & \ddots & \\ c_s & a_{s1} & a_{s2} & \cdots & \gamma \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \qquad (3)$$

Будем рассматривать методы DIRK, таблица коэффициентов которых имеет вид

$$\begin{array}{c|cccc} 0 & 0 & & & \\ c_2 & a_{21} & \gamma & & \\ \vdots & \vdots & \vdots & \ddots & \\ c_s & a_{s1} & a_{s2} & \cdots & \gamma \\ 1 & b_1 & b_2 & \cdots & b_s & \gamma \\ \hline & b_1 & b_2 & \cdots & b_s & \gamma \end{array} \qquad (4)$$

Формально метод (4) является (s+1)-стадийным, но по вычислительным затратам он равноценен s-стадийному методу SDIRK (3), поскольку первая стадия является явной и не требует никаких дополнительных вычислений. Действительно, на первом шаге $k_1 = f(t_0, y_0)$, а на каждом последующем можно принять

$$k_1 = k_{s+1\,\mathrm{old}}. \qquad (5)$$

т.е. первая стадия совпадает с последней стадией предыдущего шага. Благодаря этому такие методы получили название FSAL (First Same As Last) [4]. В наших обозначениях s - число неявных стадий, тогда при заданном s методы SDIRK (3) и FSAL-DIRK (4) равноценны по вычислительным затратам и могут иметь одинаковую функцию устойчивости. Реализуются методы FSAL-DIRK так же просто, как и методы SDIRK.

Неявный метод Рунге-Кутты называется жестко точным, если последняя строка матрицы A совпадает с b^{T}. Это свойство имеет важное значение при решении жестких и дифференциально-алгебраических систем [1, 3]. Методы FSAL являются жестко точными по своему построению.

При $s = 1$ существуют единственные методы SDIRK и FSAL-DIRK второго порядка (средней точки и трапеций, соответственно):

$$\begin{array}{c|cc} 1/2 & 1/2 & \\ \hline & 1 & \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

При $s=2$ существует единственный A-устойчивый метод SDIRK третьего порядка [1]

$$\begin{array}{c|cc} \gamma & \gamma & \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline & 1/2 & 1/2 \end{array} \qquad \gamma = \frac{3+\sqrt{3}}{6}.$$

Аналогичный метод FSAL-DIRK (также единственный) имеет вид

$$\begin{array}{c|ccc} 0 & 0 & & \\ 2\gamma & \gamma & \gamma & \\ 1 & 1-b_2-\gamma & b_2 & \gamma \\ \hline & 1-b_2-\gamma & b_2 & \gamma \end{array} \qquad \gamma = \frac{3+\sqrt{3}}{6}, \quad b_2 = \frac{1-2\gamma}{4\gamma}.$$

Основное преимущество приведенных методов FSAL-DIRK состоит в том, что они являются жестко точными, в то время как аналогичные методы SDIRK не являются таковыми.

При решении жестких задач весьма желательно, чтобы метод был жестко точным и L-устойчивым. При $s=2$ существует два таких метода SDIRK второго порядка [1]:

$$\begin{array}{c|cc} \gamma & \gamma & \\ 1 & 1-\gamma & \gamma \\ \hline & 1-\gamma & \gamma \end{array} \qquad \gamma = 1 \pm \sqrt{2}/2.$$

Из них метод с меньшим значением γ более точен. Аналогичный метод FSAL-DIRK имеет вид

$$\begin{array}{c|ccc} 0 & 0 & & \\ 2\gamma & \gamma & \gamma & \\ 1 & (1-\gamma)/2 & (1-\gamma)/2 & \gamma \\ \hline y_1 & (1-\gamma)/2 & (1-\gamma)/2 & \gamma \\ \hat{y}_1 & (1+\gamma)/6 & (5-3\gamma)/6 & \gamma/3 \end{array} \qquad \gamma = 1-\sqrt{2}/2$$

(формула оценивания погрешности имеет третий порядок). Этот метод можно интерпретировать как последовательное применение правила трапеций и формулы дифференцирования назад второго порядка, поэтому он получил название TR-BDF2. Метод TR-BDF2 реализован в системе математических вычислений MATLAB и подробно рассмотрен в [4]. В настоящей статье рассматриваются методы FSAL-DIRK более высокого порядка.

2. Функция устойчивости

Простейший тест, на котором исследуется устойчивость и точность методов решения задачи Коши – скалярное уравнение

$$y' = \lambda y, \quad y(t_0) = y_0. \tag{6}$$

Применяя к этому уравнению один шаг метода Рунге-Кутты, получим

$$y_1 = R(h\lambda)y_0,$$

где $R(z)$ – функция устойчивости. Первая задача, которая возникает при построении неявного метода Рунге-Кутты – выбор подходящей функции устойчивости.

Для методов SDIRK (3) и FSAL-DIRK (4) функция устойчивости имеет вид [3]

$$R(z) = P(z)/(1-\gamma z)^s,$$

где

$$P(z) = \det\left(I - zA + z\mathbf{1}b^{\mathrm{T}}\right), \quad \mathbf{1} = [1,\ldots,1]^{\mathrm{T}}, \quad I = \mathrm{diag}(\mathbf{1}).$$

Функция устойчивости аппроксимирует показательную функцию с порядком p, если при $z \to 0$

$$\exp(z) - R(z) = Cz^{p+1} + O\left(z^{p+2}\right), \quad C \neq 0.$$

Порядок аппроксимации p и константа погрешности C функции устойчивости определяют точность метода при решении уравнения (6).

Степень полинома $P(z)$ не превышает s, и при $p \geq s$ коэффициенты этого полинома однозначно определяются по заданному значению γ [3]. Из необходимого условия L(α)-устойчивости $R(\infty) = 0$ следует, что для L(α)-устойчивых методов степень полинома $P(z)$ должна быть не более $s-1$. В этом случае коэффициенты полинома $P(z)$ однозначно определяются по заданному значению γ при $p \geq s-1$. Диапазоны значений γ, обеспечивающих A-устойчивость при $p \geq s$ и L-устойчивость при $p \geq s$-1, приведены в [3].

В табл.1 приведены некоторые подходящие аппроксимации, которые можно использовать при построении методов SDIRK и FSAL-DIRK. В качестве показателя точности предложена константа глобальной погрешности, определяемая через константу погрешности (локальной) по формуле

$$C_g = |C|s^p.$$

Смысл этого показателя заключается в том, что при решении уравнения (6) на интервале $[0,T]$ с малым постоянным шагом относительная ошибка в конце интервала будет приближенно равна $|T\lambda| \cdot C_g \cdot |\tau\lambda|^p$, где $\tau = h/s$ – средняя длина неявной стадии интегрирования (эта оценка получена в результате предельного перехода при $h \to 0$). Константа глобальной погрешности позволяет проводить корректное сравнение функций устойчивости для методов с различным числом стадий.

Таблица 1

№	s	p	γ	Устойчивость	C_g
1	2	3	0.788675	A(90°)	0.718
2	3	3	0.435866	L(90°)	0.699
3	3	3	0.158984	L(75.6°)	0.106
4	3	4	1.06858	A(90°)	13.3
5	4	4	0.220428	L(89.55°)	0.288
6	5	4	0.25	L(90°)	0.512
7	5	4	0.174484	L2(89.97°)	0.397
8	5	5	0.141127	L(72.3°)	0.182
9	6	5	0.2	L(90°)	1.04
10	6	5	0.119061	L2(82.5°)	0.281

Из табл.1 видно, что лучшая точность обеспечивается при малых γ, однако в этом случае уменьшается величина сектора L(α)-устойчивости. Две из приведенных аппроксимаций (позиции 7 и 10) имеют второй порядок L-затухания [5].

3. Точность

Построение неявного метода Рунге-Кутты сводится к выбору подходящей функции устойчивости и последующему определению коэффициентов метода, обеспечивающих заданный порядок аппроксимации. Условия порядка удобно формировать, используя помеченные деревья [6]. Тогда для метода порядка p должны выполняться равенства

$$e(t_{ij}) = 0, \quad t_{ij} \in LT_i, \quad i = \overline{1,p},$$

где $e(t_{ij})$ – коэффициенты погрешности; LT_i – упорядоченное множество всех помеченных деревьев порядка i. При выполнении условия $c = A\mathbf{1}$ выражения для коэффициентов погрешности до 4-го порядка включительно имеют вид

$$\begin{aligned} &e(t_{11}) = 1 - b^{\mathrm{T}}\mathbf{1}, \quad e(t_{21}) = 1 - 2b^{\mathrm{T}}c, \\ &e(t_{31}) = 1 - 3b^{\mathrm{T}}c^2, \quad e(t_{32}) = 1 - 6b^{\mathrm{T}}Ac, \\ &e(t_{41}) = 1 - 4b^{\mathrm{T}}c^3, \quad e(t_{42}) = 1 - 8(bc)^{\mathrm{T}}Ac, \\ &e(t_{43}) = 1 - 12b^{\mathrm{T}}Ac^2, \quad e(t_{44}) = 1 - 24b^{\mathrm{T}}A^2c. \end{aligned} \tag{7}$$

В этих формулах предполагается покомпонентное выполнение операций умножения векторов и возведения вектора в степень.

Согласно классическим представлениям, точность метода Рунге-Кутты определяется его порядком p и значениями коэффициентов погрешности $e(t_{p+1,j})$. Свободные параметры обычно выбирают из условия минимизации этих коэффициентов. Такой подход вполне оправдан при построении методов решения нежестких задач, однако для жестких задач он не всегда оправдан. Действительно, классическое определение порядка аппроксимации основано на асимптотическом поведении решения при $h \to 0$, но тогда все задачи становятся нежесткими. Для жестких задач реальный порядок метода может быть меньше классического [2, 3]. В этом случае важное значение имеет стадийный порядок, т.е. наибольшее целое число q, для которого выполняются равенства

$$c^i = iAc^{i-1}, \quad 1 - ib^{\mathrm{T}}c^{i-1} = 0, \quad i = \overline{1,q}. \tag{8}$$

Стадийный порядок метода SDIRK ограничен порядком первой стадии (т.е. порядком метода Эйлера) и не может быть более единицы. Стадийный порядок метода FSAL-DIRK ограничен порядком второй стадии (т.е. порядком метода трапеций) и может быть равен двум. Более высокий стадийный порядок – основное преимущество методов FSAL-DIRK по сравнению с методами SDIRK, позволяющее обеспечить более высокую точность при решении жестких задач. Второй стадийный порядок позволяет также упростить построение методов, поскольку в этом случае некоторые коэффициенты погрешности равны между собой. Например, для (7) имеем

$$e(t_{31}) = e(t_{32}), \quad e(t_{41}) = e(t_{42}), \quad e(t_{43}) = e(t_{44}),$$

вследствие чего соответственно уменьшается число условий порядка 3 и 4. Далее, вместо девяти дополнительных условий порядка 5 можно использовать только четыре и т.д.

Простейшая модель, на которой исследовался феномен снижения порядка – уравнение Протеро-Робинсона [1-3]

$$y' = \lambda\big(y - \varphi(t)\big) + \varphi'(t), \quad y(t_0) = \varphi_0 = \varphi(t_0), \quad \mathrm{Re}\lambda \le 0, \tag{9}$$

решение которого $y(t)=\varphi(t)$. Применяя один шаг метода Рунге-Кутты (2) к уравнению (9), получим

$$y_1 = \varphi_0 + b^{\mathrm{T}}(I - zA)^{-1}\left(z(\mathbf{1}\varphi_0 - \Phi) + h\Phi'\right), \tag{10}$$

где

$$z = h\lambda, \quad \Phi = \begin{bmatrix} \varphi(t_0 + c_1 h) \\ \cdots \\ \varphi(t_0 + c_s h) \end{bmatrix}, \quad \Phi' = \begin{bmatrix} \varphi'(t_0 + c_1 h) \\ \cdots \\ \varphi'(t_0 + c_s h) \end{bmatrix}.$$

Рассмотрим поведение численного решения (10) при $\lambda \to \infty$. В случае неявного метода с обратимой матрицей A получим

$$y_1^* = \lim_{z\to\infty} y_1(z) = (1 - b^{\mathrm{T}}A^{-1}\mathbf{1})\varphi_0 + b^T A^{-1}\Phi.$$

Для жестко точного метода справедливо $b^{\mathrm{T}}A^{-1} = [0,\ldots,0,1]$, тогда $y_1^* = \varphi_1 = \varphi(t_0 + h)$. Для методов FSAL матрица A вырождена. В этом случае следует использовать представление

$$A = \begin{bmatrix} 0 & 0\ldots 0 \\ \widetilde{a} & \widetilde{A} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ \widetilde{b} \end{bmatrix}, \quad c = \begin{bmatrix} 0 \\ \widetilde{c} \end{bmatrix}, \quad \Phi = \begin{bmatrix} \varphi_0 \\ \widetilde{\Phi} \end{bmatrix}, \tag{11}$$

тогда, учитывая что $\widetilde{b}^{\mathrm{T}}\widetilde{A}^{-1} = [0,\ldots,0,1]$, $\widetilde{b}^{\mathrm{T}}\widetilde{A}^{-1}\widetilde{a} = b_1$, получим

$$y_1^* = (1 - \widetilde{b}^{\mathrm{T}}\widetilde{A}^{-1}\mathbf{1})\varphi_0 + (b_1 - \widetilde{b}^{\mathrm{T}}\widetilde{A}^{-1}\widetilde{a})\varphi_0' + \widetilde{b}^{\mathrm{T}}\widetilde{A}^{-1}\widetilde{\Phi} = \varphi_1.$$

Таким образом, жестко точные методы обеспечивают асимптотически точное решение уравнения (9) при $\lambda \to \infty$. Более детальный анализ показывает, что ошибка численного решения в этом случае пропорциональна z^{-1}.

Исследуем теперь поведение локальной погрешности при $h \to 0$. Используя разложение $\varphi(t)$ в ряд Тейлора, получим

$$\varphi_1 - y_1 = \sum_{i=1}^{\infty} e_i(z)\frac{d^i\varphi(t_0)\,h^i}{dt^i \quad i!}, \tag{12}$$

$$e_i(z) = zb^{\mathrm{T}}(I - zA)^{-1}\left(c^i - iAc^{i-1}\right) + \left(1 - ib^{\mathrm{T}}c^{i-1}\right).$$

Для метода стадийного порядка q выполняются равенства (8), поэтому $e_i(z) \equiv 0$ при $i \le q$. Это означает, что главный член погрешности в разложении (12) пропорционален $e_{q+1}(z)$ и соответствующему члену разложения $\varphi(t)$ в ряд Тейлора. Глобальная погрешность при $|R(z)| < 1$ асимптотически ведет себя так же, как и локальная [3], причем при $|R(z)| << 1$ она практически равна локальной. Таким образом, глобальная погрешность может вести себя как $O(h^{q+1})$, а не как $O(h^p)$, что объясняет феномен снижения порядка. Для жестко точного L(α)-устойчивого метода глобальная погрешность при $z \to \infty, h \to 0$ пропорциональна $O(z^{-1}h^{q+1})$ и асимптотически равна локальной. Аналогичные оценки справедливы также и для сингулярно возмущенных задач [3].

Функцию $e_{q+1}(z)$, определяющую главный член локальной погрешности при решении уравнения (9), назовем функцией погрешности. Функция погрешности играет такую же роль по отношению к стадийному порядку q, что и коэффициенты погрешности по отношению к

классическому порядку p. Минимизация этой функции, как и минимизация коэффициентов погрешности, позволяет повысить точность метода, имеющего заданный порядок (стадийный или классический).

Для методов SDIRK функция погрешности

$$e_2(z) = zb^{\mathrm{T}}(I - zA)^{-1}\left(c^2 - 2Ac\right) + \left(1 - 2b^{\mathrm{T}}c\right). \tag{13}$$

Для методов FSAL-DIRK второго стадийного порядка (далее рассматриваются только такие методы), используя (11), можно записать

$$e_3(z) = z\tilde{b}^{\mathrm{T}}(I - z\tilde{A})^{-1}\left(\tilde{c}^3 - 3\tilde{A}\tilde{c}^2\right) + \left(1 - 3\tilde{b}^{\mathrm{T}}\tilde{c}^2\right). \tag{14}$$

Асимптотическое поведение функции погрешности при $z \to 0$ в общем случае соответствует $O(z^{p-q})$. При $z \to \infty$ для жестко точных (в том числе и FSAL-DIRK) методов функция погрешности ведет себя как $O(z^{-1})$, причем в некоторых случаях эту оценку можно улучшить. Например, если потребовать для метода FSAL-DIRK выполнения условия

$$\tilde{b}^T \tilde{A}^{-2}\tilde{c}^3 = 3, \tag{15}$$

то $e_3(z)$ будет вести себя как $O(z^{-2})$. Важность условия (15) также и в том, что оно совпадает с одним из условий порядка для дифференциально-алгебраических уравнений индекса 2 [3].

Помимо хорошего асимптотического поведения функции $e_{q+1}(z)$ желательно обеспечить малую величину ее модуля при изменении z в широких пределах. Поэтому в качестве показателей точности будем использовать нормы

$$\begin{aligned} \left\|e_{q+1}(z)\right\|_{\mathrm{R}} &= \max_{\mathrm{Re}\, z \le 0,\, \mathrm{Im}\, z = 0} \left|e_{q+1}(z)\right|, \\ \left\|e_{q+1}(z)\right\|_{\mathrm{C}} &= \max_{\mathrm{Re}\, z \le 0} \left|e_{q+1}(z)\right| = \max_{\mathrm{Re}\, z = 0} \left|e_{q+1}(z)\right| \end{aligned} \tag{16}$$

(при $\gamma > 0$ функция $e_{q+1}(z)$ не имеет полюсов в левой полуплоскости, поэтому максимум ее модуля достигается на мнимой оси).

4. Методы

Начнем с методов третьего порядка при $s = 3$. L(α)-устойчивые жестко точные методы SDIRK задаются в виде [1]

$$\begin{array}{c|ccc} \gamma & \gamma & & \\ c_2 & c_2 - \gamma & \gamma & \\ 1 & b_1 & b_2 & \gamma \\ \hline & b_1 & b_2 & \gamma \end{array} \qquad \begin{aligned} c_2 &= (1+\gamma)/2, \\ b_2 &= (6\gamma^2 - 20\gamma + 5)/4, \\ b_1 &= 1 - b_2 - \gamma, \end{aligned} \tag{17}$$

где γ – один из корней уравнения

$$x^3 - 3x^2 + \frac{3}{2}x - \frac{1}{6} = 0. \tag{18}$$

Погрешность будем оценивать с помощью формулы

$$\hat{y}_1 = y_0 + h\left(\hat{b}_0 f(t_0, y_0) + \hat{b}_1 k_1 + \hat{b}_2 k_2\right),$$

$$\hat{b}_1 = \frac{3\gamma - 1}{6\gamma(1-\gamma)}, \hat{b}_2 = \frac{2(2-3\gamma)}{3\gamma(1-\gamma^2)}, \hat{b}_0 = 1 - \hat{b}_1 - \hat{b}_2$$

Аналогичные методы FSAL-DIRK имеют вид

$$\begin{array}{c|cccc} 0 & 0 & & & \\ 2\gamma & \gamma & \gamma & & \\ c_3 & a_{31} & a_{32} & \gamma & \\ 1 & b_1 & b_2 & b_3 & \gamma \\ \hline y_1 & b_1 & b_2 & b_3 & \gamma \\ \hat{y}_1 & \hat{b}_1 & \hat{b}_2 & \hat{b}_3 & 0 \end{array} \qquad (19)$$

где γ – один из корней уравнения (18), c_3 – свободный параметр. Из условия второго стадийного порядка имеем $c_2 = 2\gamma$, а коэффициенты третьей стадии равны

$$a_{32} = \frac{c_3(c_3 - 2\gamma)}{4\gamma}, a_{31} = c_3 - a_{32} - \gamma. \qquad (20)$$

При нахождении коэффициентов заключительной стадии дополнительно используется условие $e(t_{31}) = 0$, обеспечивающее третий порядок метода. В результате получим

$$b_2 = \frac{6\gamma(1-c_3) + 3c_3 - 2}{12\gamma(c_3 - 2\gamma)}, \quad b_3 = \frac{6\gamma^2 - 6\gamma + 1}{3c_3(c_3 - 2\gamma)}, \quad b_1 = 1 - b_2 - b_3.$$

Коэффициенты формулы вычисления $\hat{y}_1$ также определяются из условия третьего порядка:

$$\hat{b}_2 = \frac{3c_3 - 2}{12\gamma(c_3 - 2\gamma)}, \quad \hat{b}_3 = \frac{2 - 6\gamma}{6c_3(c_3 - 2\gamma)}, \quad \hat{b}_1 = 1 - \hat{b}_2 - \hat{b}_3.$$

Зададим $\gamma = 0.158984$ (табл.1, позиция 3) и выберем c_3. Из условия минимизации коэффициентов погрешности $e(t_{41}) = e(t_{42}) = 0$ получим

$$c_3 = \frac{24\gamma^2 - 20\gamma + 3}{24\gamma^2 - 24\gamma + 4} = 0.539746. \qquad (21)$$

Близкое к этому значение дает условие L-устойчивости третьей стадии:

$$c_3 = \left(2 + \sqrt{2}\right)\gamma = 0.542805. \qquad (22)$$

Можно также задать свободный параметр из условия минимизации показателей (16) (тогда $c_3 = 0.818$), но для метода третьего порядка это не дает ощутимого преимущества. Из двух значений (21), (22) выберем второе, которое дает наиболее удобные для реализации коэффициенты. Их значения:

$$\gamma = 0.158983899988677,$$

$$c_3 = \left(2 + \sqrt{2}\right)\gamma, \quad a_{31} = a_{32} = (c_3 - \gamma)/2,$$

$$b_3 = \left(\sqrt{2} - 1\right)\frac{6\gamma^2 - 6\gamma + 1}{6\gamma^2}, \quad b_1 = b_2 = (1 - b_3 - \gamma)/2, \qquad (23)$$

$$\hat{b}_2 = \left(\sqrt{2}+1\right)\frac{\sqrt{2}-2+3\gamma}{12\gamma^2},\quad \hat{b}_3 = \left(\sqrt{2}-1\right)\frac{1-3\gamma}{6\gamma^2},\quad \hat{b}_1 = 1-\hat{b}_2-\hat{b}_3-\gamma.$$

Первые три стадии метода (19), (23) выполняются по формулам TR-BDF2, а последняя – по формуле дифференцирования назад третьего порядка, поэтому построенный метод можно назвать TR-BDF2-BDF3.

L(α)-устойчивые методы порядка 4 при $s=4$ можно получить, если задать γ равным одному из корней уравнения

$$x^4 - 4x^3 + 3x^2 - \frac{2}{3}x + \frac{1}{24} = 0$$

(за исключением $\gamma = 0.106$). Жестко точных методов SDIRK такого типа не существует [1]. При заданном γ методы FSAL-DIRK второго стадийного порядка образуют три семейства:

1) c_3, c_4 – свободные параметры, $c_2 = 2\gamma$, $c_3 \neq c_2$, $c_4 \neq c_2$, $c_3 \neq c_4$;

2) b_4 – свободный параметр, $b_4 \neq 0$, $c_4 = c_2 = 2\gamma$, $c_3 = \dfrac{24\gamma^2 - 20\gamma + 3}{4(6\gamma^2 - 6\gamma + 1)}$;

3) b_4 – свободный параметр, $b_4 \neq 0$, $c_2 = 2\gamma$, $c_3 = c_4 = \dfrac{24\gamma^2 - 20\gamma + 3}{4(6\gamma^2 - 6\gamma + 1)}$.

Рассмотрим первый случай. Весовые коэффициенты находим из уравнений

$$b_1 + b_2 + b_3 + b_4 = 1 - \gamma,$$

$$b_2 c_2 + b_3 c_3 + b_4 c_4 = 1/2 - \gamma,$$

$$b_2 c_2^2 + b_3 c_3^2 + b_4 c_4^2 = 1/3 - \gamma,$$

$$b_2 c_2^3 + b_3 c_3^3 + b_4 c_4^3 = 1/4 - \gamma.$$

Коэффициенты третьей стадии определяем по формулам (20), а четвертой – решая уравнения

$$a_{41} + a_{42} + a_{43} = c_4,$$

$$a_{42} c_2 + a_{43} c_3 + \gamma c_4 = c_4^2/2,$$

$$b_2 \gamma c_2^2 + b_3 (a_{32} c_2^2 + \gamma c_3^2) + b_4 (a_{42} c_2^2 + a_{43} c_3^2 + \gamma c_4^2) + \gamma/3 = 1/12.$$

Мы выбрали $\gamma = 0.2204$ (табл.1, позиция 5) а параметры c_3, c_4 задали из условия L-устойчивости третьей и четвертой стадий. В результате получили следующие значения коэффициентов:

$$\gamma = a_{21} = 0.220428410259212,$$

$$c_3 = 0.752589667839344, \qquad a_{31} = a_{32} = 0.266080628790066,$$

$$c_4 = 0.610097451414243, \qquad a_{41} = a_{42} = 0.227031047465079,$$

$$a_{43} = -0.064393053775127, \quad b_1 = b_2 = 0.175575441883476, \qquad (24)$$

$b_3 = -0.415534431720558, \quad b_4 = 0.843955137694394,$

$\hat{b}_1 = \hat{b}_2 = 0.217113586697490, \quad \hat{b}_3 = 0.414811674412460,$

$b_4 = 0.150961152192560, \quad \hat{b}_5 = 0.$

Исследовалось также двупараметрическое семейство L(α)-устойчивых методов FSAL-DIRK порядка 5 при $s = 5$ и $\gamma = 0.141$ (табл.1, позиция 8). Один из подходящих наборов коэффициентов получен при $c_4 = 0.8$, $c_5 = 1$:

$\gamma = a_{21} = 0.141127125787053,$

$c_3 = 0.732905744297517, \quad c_4 = 0.8, \quad c_5 = 1,$

$a_{31} = 0.006694309148835, \quad a_{32} = 0.585084309361629,$

$a_{41} = 0.168415634641113, \quad a_{42} = 0.338089701918851,$

$a_{43} = 0.152367537652983, \quad a_{51} = -0.258119533121494,$

$a_{52} = 1.069536753666977, \quad a_{53} = -0.283586950067325,$ (25)

$a_{54} = 0.331042603734788, \quad b_1 = 0.085667539849126,$

$b_2 = 0.422665716195131, \quad b_3 = 0.431493500913056,$

$b_4 = -0.021417480601987, \quad b_5 = -0.059536402142379,$

$\hat{b}_1 = 0.080558017906371, \quad \hat{b}_2 = 0.440554894684905,$

$\hat{b}_3 = 0.288630408509544, \quad \hat{b}_4 = 0.130720276756800,$

$\hat{b}_5 = 0.059536402142380, \quad \hat{b}_6 = 0.$

Рассмотрим теперь методы, параметры которых выбирались из условия (15) и условия минимизации показателей (16). Для увеличения числа свободных параметров мы принимали $s = 5, p = 4$ либо $s = 6, p = 5$.

При построении метода порядка 4 использовалось значение $\gamma = 1/4$ (табл.1, позиция 6), выбранное в [3] для метода SDIRK. Полученная схема:

0	0					
1/2	1/4	1/4				
1/4	1/16	−1/16	1/4			
3/4	1/16	−1/16	1/2	1/4		
1	−9/62	−77/124	143/124	45/124	1/4	
1	7/90	2/15	16/45	16/45	−31/180	1/4
y_1	7/90	2/15	16/45	16/45	−31/180	1/4
$\hat{y}_1$	0	−1/3	2/3	2/3	0	0

(26)

Аналогичный метод SDIRK, параметры которого получены из условия минимизации коэффициентов погрешности, приведен в [3, с.118].

Для метода порядка 5 мы выбрали $\gamma = 1/5$ (табл.1, позиция 9), в результате получили

$$
\begin{array}{c|cccccc}
0 & 0 & & & & & \\
\frac{2}{5} & \frac{1}{5} & \frac{1}{5} & & & & \\
\frac{3}{5} & \frac{1}{4} & \frac{3}{20} & \frac{1}{5} & & & \\
1 & \frac{87}{140} & -\frac{27}{28} & \frac{8}{7} & \frac{1}{5} & & \\
\frac{4}{5} & \frac{156973}{590625} & \frac{58202}{118125} & -\frac{35852}{118125} & \frac{12236}{84375} & \frac{1}{5} & \\
\frac{1}{5} & \frac{2436319}{54337500} & -\frac{2573719}{10867500} & \frac{690136}{2716875} & -\frac{272248}{1940625} & \frac{9}{115} & \frac{1}{5} \\
1 & \frac{19}{288} & \frac{25}{144} & \frac{25}{144} & -\frac{193}{1440} & \frac{25}{96} & \frac{25}{96} & \frac{1}{5} \\
\hline
y_1 & \frac{19}{288} & \frac{25}{144} & \frac{25}{144} & -\frac{193}{1440} & \frac{25}{96} & \frac{25}{96} & \frac{1}{5} \\
\hat{y}_1 & -\frac{26107}{103776} & -\frac{28625}{51888} & -\frac{16925}{51888} & -\frac{23767}{103776} & \frac{116525}{103776} & \frac{5575}{4512} & 0
\end{array}
\qquad (27)
$$

Расчет коэффициентов методов выполнялся с использованием системы MathCAD 6.0.

5. Реализация

Эффективность решения задачи Коши зависит не только от выбранного метода, но и от его программной реализации. Особенно это касается неявных методов, реализация которых не определяется однозначно их коэффициентами. К числу вопросов, которые необходимо решать при реализации неявных методов, относятся: выбор начального приближения для итераций метода Ньютона; критерии окончания итераций и обновления матрицы Якоби; оценка погрешности и управление величиной шага.

Систему алгебраических уравнений, возникающую при реализации неявного метода (2), целесообразно решать относительно приращений. Применительно к методу FSAL-DIRK (4) обозначим

$$z_1 = h\gamma k_1, \quad z_i = g_i - y_0, \quad i = \overline{2,s+1}.$$

Тогда

$$\begin{bmatrix} z_1 \\ \vdots \\ z_{s+1} \end{bmatrix} = \left(A_0 + \gamma I\right) \begin{bmatrix} hk_1 \\ \vdots \\ hk_{s+1} \end{bmatrix}, \qquad (28)$$

где A_0 – матрица, полученная путем замены диагональных элементов матрицы A нулями. Используя (28), получим неявные стадии в виде

$$z_i = \sum_{j=1}^{i-1} \gamma_{ij} z_j + h\gamma f\left(t_0 + c_i h, y_0 + z_i\right), \qquad i = \overline{2,s+1}, \qquad (29)$$

где коэффициенты матрицы $\left[\gamma_{ij}\right] = A_0 (A_0 + \gamma I)^{-1}$ можно найти по рекуррентным формулам

$$\gamma_{ij} = \frac{1}{\gamma}\left(a_{ij} - \sum_{k=j+1}^{i-1} a_{ik}\gamma_{kj}\right), \quad i = \overline{2,s+1}, \quad j = \overline{1,i-1}.$$

На первом шаге интегрирования принимаем $z_1 = h\gamma f(t_0, y_0)$. На каждом последующем шаге, в соответствии с (5), принимаем

$$z_1 = \frac{h_{new}}{h_{old}}\left(z_{s+1\,\mathrm{old}} - \sum_{j=1}^{s}\gamma_{s+1,j} z_{j\,\mathrm{old}}\right).$$

Такая "сглаженная первая стадия" [4] повышает эффективность решения жестких задач.

Итерации Ньютона для решения на каждой неявной стадии системы алгебраических уравнений (29) запишутся в виде

$$\left(I - h\gamma J\right)\left(z_i^{k+1} - z_i^k\right) = \sum_{j=1}^{i-1}\gamma_{ij} z_j + h\gamma f\left(t_0 + c_i h, y_0 + z_i\right) - z_i^k,$$

где J - матрица Якоби, k - индекс итерации. Начальные приближения для итераций примем в виде

$$z_i^0 = \sum_{j=1}^{i-1}\beta_{ij} z_j, \quad i = \overline{2, s+1}. \tag{30}$$

Для определения коэффициентов β_{ij} будем рассматривать на каждой неявной стадии вложенную в нее явную стадию

$$\hat{g}_i = y_0 + h\sum_{j=1}^{i-1}\hat{a}_{ij} k_j, \quad i = \overline{2, s+1}. \tag{31}$$

Коэффициенты $\hat{a}_{ij}$ выбираются такими, чтобы минимизировать погрешность $g_i - \hat{g}_i$. На последней стадии в качестве начального приближения имеет смысл использовать $\hat{y}_1$. Матрица коэффициентов формулы (30) определяется из (28), (31) и равна $\left[\beta_{ij}\right] = \left[\hat{a}_{ij}\right]\left(A_0 + \gamma I\right)^{-1}$. Для расчета этих коэффициентов можно использовать рекуррентные формулы

$$\beta_{ij} = \frac{1}{\gamma}\left(\hat{a}_{ij} - \sum_{k=j+1}^{i-1}\hat{a}_{ik}\gamma_{kj}\right), \quad i = \overline{2, s+1}, \quad j = \overline{1, i-1}.$$

Как и другие жестко точные методы, методы FSAL-DIRK хорошо приспособлены для решения дифференциально-алгебраических систем. В этом случае на каждой неявной стадии решается система уравнений относительно дифференциальных и алгебраических переменных, а результат выполнения последней стадии принимается как численное решение на текущем шаге. Для уменьшения числа итераций важно использовать хорошее начальное приближение для всех переменных. Используя метод ε-вложения [3], можно показать, что формулы (30) подходят также и для алгебраических переменных.

В нашей программе в качестве критерия останова итераций используется оценка погрешности, вычисляемая в соответствии с [3]. Эксперименты показали, что выгоднее делать немного итераций, поэтому допускается не более 2...4 (в зависимости от метода) итераций. Пересчет матрицы Якоби перед началом очередного шага производится только в том случае, когда на одной из стадий предыдущего шага выполняется хотя бы одно из двух условий:

a) скорость сходимости итераций ниже допустимой; b) выполнено максимальное число итераций, но оценка погрешности итераций больше требуемого значения.

Для управления величиной шага используется стандартная процедура [3]. При этом, в соответствии с [3, 4], оценка погрешности вычисляется по формуле

$$e = (I - h\gamma J)^{-1}(y_1 - \hat{y}_1),$$

что ограничивает рост жестких компонент оценки.

6. Эксперименты

Для проведения экспериментов с предложенными методами была написана программа, позволяющая реализовать произвольный метод FSAL-DIRK, достаточно только задать коэффициенты γ_{ij}, β_{ij}, а также некоторые настраиваемые параметры. Поскольку жестко точный метод SDIRK можно представить как метод FSAL-DIRK с нулевым первым столбцом матрицы A, эта же программа использовалась для реализации методов SDIRK, с которыми производилось сравнение.

Основные характеристики используемых методов приведены в табл.2. После названия метода в скобках указываются число неявных стадий и порядок. Во второй графе таблицы даны номера формул, задающих коэффициенты метода. Далее приводятся:
$\hat{p}$ - порядок формулы вычисления $\hat{y}$;

$\|e(t_{p+1})\| = \left(\sum_i e(t_{p+1,i})^2\right)^{1/2}$ - норма вектора коэффициентов погрешности;

$\|e_{q+1}(z)\|_R$, $\|e_{q+1}(z)\|_C$ - нормы (16) функции погрешности (13) или (14).

Методы FSAL(3,3) и FSAL(4,4) реализованы также в программном комплексе "МВТУ" (Моделирование в технических устройствах).

Таблица 2

Метод	Формула	γ	$\hat{p}$	$\|e(t_{p+1})\|$	$\|e_{q+1}(z)\|_R$	$\|e_{q+1}(z)\|_C$
SDIRK(3,3)	(17)	0.159	2	0.103	3.07E-2	7.98E-2
SDIRK(5,4)	[3, с.118]	0.25	3	0.134	2.04E-2	6.34E-2
FSAL(3,3)	(23)	0.159	3	0.133	2.91E-2	6.12E-2
FSAL(4,4)	(24)	0.2204	3	0.250	2.04E-3	1.16E-2
FSAL(5,4)	(26)	0.25	4	0.233	5.14E-4	4.20E-3
FSAL(5,5)	(25)	0.141	4	0.107	7.55E-3	2.43E-2
FSAL(6,5)	(27)	0.2	4	0.342	2.27E-4	4.65E-3

Со всеми этими методами были проведены эксперименты по определению реального порядка аппроксимации в зависимости от жесткости задачи. В качестве тестовой выбрана задача Капса [2]

$$y_1' = -(E+2)y_1 + Ey_2^2, \quad y_1(0) = 1,$$
$$y_2' = y_1 - y_2 - y_2^2, \quad y_2(0) = 1, \qquad (32)$$

которая имеет гладкое решение $y_1(t) = \exp(-2t)$, $y_2(t) = \exp(-t)$, не зависящее от параметра жесткости E. Реальный порядок определялся по формуле [2]

$$p_r = \log_2(\|e_N\|/\|e_{2N}\|),$$

где $\|e_N\|, \|e_{2N}\|$ – евклидова норма абсолютной погрешности в конце интервала [0, 1] при интегрировании с шагом $h = 1/N$ и $h = 1/(2N)$. Чтобы поставить все методы в одинаковые условия, мы задавали $N = 60/s$, в этом случае число неявных стадий, выполняемых на заданном интервале, одинаково для всех методов. Результаты приведены в табл. 3.

Таблица 3

	SDIRK(3,3)		FSAL(3,3)		SDIRK(5,4)		FSAL(5,4)		FSAL(4,4)		FSAL(5,5)		FSAL(6,5)	
E	$\|e_{40}\|$	p_r	$\|e_{40}\|$	p_r	$\|e_{24}\|$	p_r	$\|e_{24}\|$	p_r	$\|e_{30}\|$	p_r	$\|e_{24}\|$	p_r	$\|e_{20}\|$	p_r
10^1	2.1E-7	3.0	8.8E-8	3.0	1.4E-7	3.8	2.9E-9	4.5	7.7E-9	3.9	9.7E-11	5.2	1.9E-9	4.6
10^2	1.3E-6	3.0	8.6E-8	3.2	3.1E-6	2.9	2.0E-9	5.1	1.4E-8	3.0	1.7E-8	4.3	5.3E-9	2.6
10^3	4.1E-6	1.5	7.8E-8	2.4	4.0E-6	1.3	2.4E-9	3.4	2.7E-9	2.6	4.9E-8	2.6	2.5E-11	8.9
10^4	4.9E-7	1.2	3.1E-8	3.0	4.9E-7	1.1	1.2E-9	4.0	6.8E-10	3.7	1.2E-8	2.2	2.2E-10	3.4
10^5	6.7E-8	2.1	2.9E-8	3.0	5.1E-8	1.2	1.2E-9	4.0	5.8E-10	4.1	1.4E-9	2.1	3.9E-11	4.2
10^6	3.1E-8	2.9	2.8E-8	3.0	5.8E-9	2.2	1.2E-9	4.0	6.3E-10	4.0	1.4E-10	2.0	2.0E-11	4.9
10^7	2.9E-8	3.0	2.8E-8	3.0	1.5E-9	3.7	1.2E-9	4.0	6.3E-10	4.0	1.3E-11	2.8	1.9E-11	5.0
10^8	2.8E-8	3.0	2.8E-8	3.0	1.2E-9	4.0	1.2E-9	4.0	6.3E-10	4.0	2.8E-12	5.2	1.9E-11	5.0

При анализе этих результатов следует учитывать, что погрешность численного решения складывается из двух составляющих: жесткой и мягкой. Жесткая составляющая невелика при малых и больших значениях E, и достигает максимального значения при умеренной жесткости, что согласуется с поведением функции погрешности. Сравнение результатов применения методов SDIRK и FSAL убедительно показывает преимущество более высокого стадийного порядка. Эффект снижения порядка наиболее заметно проявился у методов SDIRK(5,4) и FSAL(5,5), поскольку они имеют более высокий порядок, а также потому, что показатели (16) этих методов не были минимизированы.

При больших E наибольшее по модулю собственное значение мало изменяется на интервале интегрирования и примерно равно $-E$. Это позволяет сравнить ошибку численного решения, как функцию от $z=-hE$, с функцией погрешности. Такое сравнение показало хорошее совпадение (с точностью до постоянного множителя) этих функций, что вполне объясняет полученные результаты. Снижение порядка практически до первого у методов SDIRK и до второго у метода FSAL(5,5) объясняется тем, что при $E \to \infty$ жесткая составляющая погрешности ведет себя как $O(E^{-1}h^q)$.

Задача (32) решалась также с автоматическим выбором шага. Результаты при $E = 10^4$ и трех значениях задаваемой относительной погрешности *Tol* приведены в табл. 4. Здесь N_f - число вычислений правой части; $\|e(1)\|$ – евклидова норма ошибки в конце интервала.

Таблица 4

	Tol=10^{-3}		*Tol*=10^{-5}		*Tol*=10^{-7}	
Метод	N_f	$\|e(1)\|$	N_f	$\|e(1)\|$	N_f	$\|e(1)\|$
SDIRK(3,3)	42	1.0E-5	380	5.0E-7	9579	5.1E-9
SDIRK(5,4)	53	2.7E-5	243	2.6E-6	5786	4.1E-8
FSAL(3,3)	27	1.5E-4	105	2.8E-6	412	6.3E-9
FSAL(4,4)	32	1.2E-4	69	1.9E-7	267	2.6E-10
FSAL(5,4)	50	4.9E-6	61	8.4E-7	231	5.7E-10
FSAL(5,5)	46	3.1E-8	145	4.0E-9	766	5.0E-10
FSAL(6,5)	42	9.9E-6	59	2.8E-7	426	4.4E-10

Эксперименты с другими жесткими задачами полностью подтверждают качественную картину, полученную при решении задачи Капса. Приведем результаты решения уравнения Ван-дер-Поля

$$y_1' = y_2, \quad y_1(0) = 2,$$
$$y_2' = 10^6\left(\left(1 - y_1^2\right)y_2 - y_1\right), \quad y_2(0) = -0.66$$

на интервале [0, 2] с начальным шагом $h_0 = 10^{-6}$. Эта задача выбрана не только потому, что жесткое уравнение Ван-дер-Поля признано одной из самых трудных тестовых задач, но также и потому, что результат решения этой задачи программой RADAU5 приведен в приложении книги [3]. Это позволяет сравнить нашу программу с одной из самых эффективных программ решения жестких систем. Для расчета ошибки мы табулировали решение с шагом $\Delta t = 0.2$ (как и в [3]) и вычисляли среднеквадратичное значение относительной ошибки из 20 полученных значений (в качестве точного принималось решение, полученное при $Tol = 10^{-14}$). Результаты приведены в табл.5, где N_J – число вычислений якобиана; N_{step} – число выполненных шагов.

Таблица 5

Tol	Метод	Ошибка	N_f	N_J	N_{step}
10^{-4}	SDIRK(3,3)	2.4E-4	3789	102	610
	SDIRK(5,4)	9.2E-4	3482	236	371
	FSAL(3,3)	5.4E-4	2197	81	361
	FSAL(4,4)	1.2E-4	2834	57	318
	FSAL(5,4)	4.7E-4	2438	197	260
	FSAL(6,5)	4.9E-5	2917	175	263
	RADAU5	4.1E-5	2263	182	276
10^{-7}	SDIRK(5,4)	7.0E-6	34426	992	4397
	FSAL(5,4)	4.9E-7	9969	677	1123
	FSAL(6,5)	6.1E-8	11291	411	1127

Эксперименты показали, что методы SDIRK показывают удовлетворительные результаты только при низкой точности, а при повышении точности их эффективность резко снижается, причем особенно неэффективны эти методы для задач весьма умеренной жесткости. Поэтому неудивительно, что и в экспериментах, проведенных с другими задачами, метод SDIRK(5,4) "дает весьма разочаровывающие результаты" [3, с.183].

СПИСОК ЛИТЕРАТУРЫ

1. *Alexander R.* Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s // SIAM J. Numer. Anal., 1977, v.14, no.6, p.1006-1021.
2. *Деккер К., Вервер Я.* Устойчивость методов Рунге-Кутты для жестких нелинейных дифференциальных уравнений. – М.: Мир, 1988, 334 с.
3. *Хайрер Э., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи. – М.: Мир, 1999, 685 с.
4. *Hosea M.E., Shampine L.F.* Analysis and implementation of TR-BDF2 // Applied Numerical Mathematics, 1996, v.20, №1-3, p.21-37.
5. *Кочетков К.А., Ширков П.Д.* L-затухающие ROW-методы третьего порядка точности // ЖВМ и МФ, 1997, т.37, № 6, с.699-710.
6. *Хайрер Э., Нёрсетт С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. – М.: Мир, 1990, 512 с.

Поступила в редакцию 29.11.2000