

УДК 519.622

МАКСИМАЛЬНЫЙ ПОРЯДОК ТОЧНОСТИ $(m, 1)$ -МЕТОДОВ РЕШЕНИЯ ЖЁСТКИХ ЗАДАЧ

Е. А. Новиков^{1,2}

¹ Институт вычислительного моделирования СО РАН, 660036, Красноярск, Академгородок.

² Сибирский федеральный университет, 660041, Красноярск, пр. Свободный, 79.

E-mail: novikov@icm.krasn.ru

Исследованы $(m, 1)$ -методы решения жёстких задач, в которых на каждом шаге один раз вычисляется правая часть системы дифференциальных уравнений. Показано, что максимальный порядок точности L -устойчивого $(m, 1)$ -метода равен двум, и построен метод максимального порядка.

Ключевые слова: жёсткие задачи, схемы Розенброка, (m, k) -методы, A -устойчивость, L -устойчивость.

Введение. При решении задачи Коши для жёстких систем обыкновенных дифференциальных уравнений широкое распространение получили методы типа Розенброка [1] благодаря простоте реализации и достаточно хорошим свойствам точности и устойчивости. Данные численные схемы получены из полуживных методов типа Рунге—Кутта, в которых для решения нелинейной системы алгебраических уравнений, возникающей при вычислении каждой стадии, используется одна итерация метода Ньютона. Все остальные проблемы решаются выбором величины шага интегрирования.

Наибольшее распространение получили методы типа Розенброка, в которых при вычислении каждой стадии применяется одна и та же матрица Якоби. Известно [2], что в этом случае для m -стадийного метода Розенброка максимальный порядок точности равен $m + 1$, причем схема максимального порядка может быть только A -устойчивой. Если отказаться от максимального порядка, то можно построить L -устойчивую численную формулу m -того порядка точности. В практических расчётах, как правило, отказываются от максимального порядка в пользу L -устойчивости. Заметим, что на основе методов типа Розенброка нельзя построить схему с замораживанием матрицы Якоби выше второго порядка точности [3], что ограничивает применение данных методов расчётами с небольшой точностью или задачами небольшой размерности.

В [4] предложен класс (m, k) -методов, в которых нахождение стадий не связывается с обязательным вычислением правой части системы дифференциальных уравнений. Числа m и k означают соответственно число стадий и количество вычислений правой части системы дифференциальных уравнений на шаг интегрирования. Реализация (m, k) -методов так же проста, как и методов Розенброка, однако (m, k) -схемы имеют лучшие свойства точности и устойчивости. В рамках (m, k) -методов значительно проще решаются

Евгений Александрович Новиков (д.ф.-м.н., профессор), главный научный сотрудник, отд. вычислительной математики¹; зав. кафедрой, каф. математического обеспечения дискретных устройств и систем².

проблемы замораживания матрицы Якоби и ее численной аппроксимации.

Здесь исследуются $(m, 1)$ -методы решения жёстких задач, в которых на каждом шаге один раз вычисляется правая часть системы дифференциальных уравнений. Показано, что максимальный порядок точности L -устойчивого $(m, 1)$ -метода равен двум, и построен метод максимального порядка.

1. Схемы типа Розенброка. Далее будет рассматриваться задача Коши для автономной системы обыкновенных дифференциальных уравнений

$$y' = f(y), \quad y(t_0) = y_0, \quad t_0 \leq t \leq t_k, \quad (1)$$

где y и f — вещественные N -мерные вектор-функции, t — независимая переменная. Рассмотрение автономной задачи не снижает общности, потому что введением дополнительной переменной $y'_{N+1} = 1$, $y_{N+1}(t_0) = t_0$ неавтономную задачу можно привести к автономному виду. Для решения задачи (1) будем применять методы типа Розенброка вида

$$y_{n+1} = y_n + \sum_{i=1}^m p_i k_i, \quad D_n k_i = hf \left(y_n + \sum_{j=1}^{i-1} \beta_{ij} k_j \right), \quad (2)$$

где $D_n = E - ahf'_n$; E — единичная матрица; $f'_n = \partial f(y_n)/\partial y$ — матрица Якоби системы (1); k_i — стадии метода; a , p_i , β_{ij} — коэффициенты, определяющие свойства точности и устойчивости (2); $1 \leq i \leq m$, $1 \leq j \leq i - 1$. В настоящее время методы типа Розенброка трактуются более широко. Под ними понимаются все численные схемы, в которых матрица Якоби или ее аппроксимация вводятся непосредственно в формулу интегрирования.

Рассмотрим в качестве примера одностадийный метод типа Розенброка

$$y_{n+1} = y_n + p_1 k_1, \quad D_n k_1 = hf(y_n), \quad (3)$$

где матрица D_n определена в (2). Разлагая приближенное решение y_{n+1} в ряд Тейлора по степеням h до членов с h^2 включительно, получим

$$y_{n+1} = y_n + p_1 h f_n + ap_1 h^2 f'_n f_n + O(h^3),$$

где $f_n = f(y_n)$. Представление точного решения $y(t_{n+1})$ в виде ряда Тейлора в окрестности точки t_n имеет вид

$$y(t_{n+1}) = y(t_n) + hf + \frac{1}{2} h^2 f' f + O(h^3),$$

где элементарные дифференциалы f и $f'f$ вычислены на точном решении $y(t_n)$. Сравнивая ряды для точного и приближенного решений при условии $y_n = y(t_n)$, видим, что схема (3) будет иметь второй порядок точности, если $p_1 = 1$ и $ap_1 = 0,5$, то есть $p_1 = 1$ и $a = 0,5$. Теперь исследуем устойчивость метода (3). Для этого применим его для решения скалярного тестового уравнения $y' = \lambda y$, где λ есть произвольное комплексное число, $\text{Re}(\lambda) < 0$. Смысл λ — некоторое собственное число матрицы Якоби задачи (1). Обозначая $x = h\lambda$, получим $y_{n+1} = Q(x)y_n$, где функция устойчивости $Q(x)$ имеет следующий вид:

$$Q(x) = \frac{1 + (p_1 - a)x}{1 - ax}.$$

Подставляя сюда значения коэффициентов $p_1 = 1$ и $a = 0,5$, имеем $Q(x) = (1 + 0,5x)/(1 - 0,5x)$, то есть схема (3) второго порядка является A -устойчивой. Из вида функции $Q(x)$ следует, что схема (3) будет L -устойчивой, если $p_1 = a = 1$, что противоречит второму порядку точности. Обычно отказываются от второго порядка в пользу L -устойчивости, что приводит к более эффективному методу, хотя и первого порядка.

В случае большой размерности задачи (1) основные вычислительные затраты связаны с обращением матрицы D_n . Обычно вместо обращения решается линейная система алгебраических уравнений $D_n k_1 = hf(y_n)$ с применением LU -разложение матрицы D_n с выбором главного элемента по строке или столбцу, а иногда и по всей матрице, то есть при вычислении приближенного решения по формуле (3) осуществляется декомпозиция матрицы D_n (порядка N^3 арифметических операций). Обратный ход метода Гаусса стоит порядка N^2 операций. Таким образом, при большой размерности исходной задачи общие вычислительные затраты определяются временем декомпозиции матрицы D_n . Возникает естественный вопрос: нельзя ли без значительного увеличения вычислительных затрат исправить схему (3) таким образом, чтобы она была L -устойчивой и имела второй порядок точности? Эта проблема решается в рамках (m, k) -методов.

2. Класс (m, k) -методов. Пусть заданы $m, k \in \mathbb{Z}$, $k \leq m$. Обозначим через M_m множество чисел $i \in \mathbb{Z}$ таких, что $1 \leq i \leq m$, а через M_k и J_i подмножества из M_m вида

$$M_k = \{m_i \in M_m \mid 1 = m_1 < m_2 < \dots < m_k \leq m\},$$

$$J_i = \{m_{j-1} \in M_m \mid j > 1, m_j \in M_k, m_j \leq i\}, \quad 1 < i \leq m.$$

Рассмотрим следующие численные схемы:

$$y_{n+1} = y_n + \sum_{i=1}^m p_i k_i,$$

$$D_n k_i = hf \left(y_n + \sum_{j=1}^{i-1} \beta_{ij} k_j \right) + \sum_{j \in J_i} \alpha_{ij} k_j + hf'_n \sum_{j=1}^{i-1} c_{ij} k_j, \quad i \in M_k, \quad (4)$$

$$D_n k_i = k_{i-1} + \sum_{j \in J_i} \alpha_{ij} k_j + hf'_n \sum_{j=1}^{i-1} c_{ij} k_j, \quad i \in M_m \setminus M_k,$$

где $D_n = E - ahf'_n$; $a, p_i, \beta_{ij}, \alpha_{ij}, c_{ij}$ — постоянные коэффициенты; h — шаг интегрирования; k и m — соответственно количество вычислений функции f и число обратных ходов в методе Гаусса (число стадий). На каждом шаге интегрирования осуществляются одно вычисление матрицы Якоби и одна декомпозиция матрицы D_n . Допускается аппроксимация матрицы Якоби f'_n матрицей A_n , удовлетворяющей условию

$$A_n = f'_n + hB_n + O(h^2),$$

где матрица B_n не зависит от шага интегрирования. Данное условие позволяет применять методы (4) с замораживанием как численной, так и аналитической матрицы Якоби. Так как k и m полностью определяют затраты на

шаг, а набор чисел m_1, m_2, \dots, m_k из множества M_k только распределяет их внутри шага, то методы типа (4) называются (m, k) -методами.

Отметим, что при $k = m$ и $\alpha_{ij} = c_{ij} = 0$ методы (4) совпадают со схемами типа Розенброка (2), а при $k = m$ и $\alpha_{ij} = 0$ — с ROW-методами [2]. Заметим также, что при рассмотрении методов такого типа все авторы изучали случай $k = m$, то есть когда число стадий и количество вычислений правой части системы дифференциальных уравнений (1) совпадают. В этом случае k -стадийную схему (4) можно поставить в соответствие k -стадийной полуявной формуле типа Рунге—Кутты, при реализации которой на каждом шаге используется одна матрица размерности N . Относительно таких численных формул известно, что нельзя построить k -стадийную схему выше $(k + 1)$ -го порядка точности, причем схема максимального порядка является A -устойчивой. Очевидно, этот факт распространяется и на (m, k) -методы при $m = k$.

3. Численные схемы с одним вычислением правой части. Выберем $M_m = \{1, 2, \dots, m\}$ и $M_k = \{1\}$ при $k = 1$, тогда J_i есть пустое множество. Рассмотрим семейство методов следующего вида:

$$y_{n+1} = y_n + \sum_{i=1}^m p_i k_i, \quad D_n k_1 = hf(y_n), \quad D_n k_i = k_{i-1}, \quad 2 \leq i \leq m, \quad (5)$$

где матрица D_n определена в (4). Для изучения (5) введём в рассмотрение матрицу B с элементами b_{ij} :

$$b_{1i} = b_{i1} = 1, \quad i \geq 1, \quad b_{ij} = b_{i-1,j} + b_{i,j-1}, \quad i, j \geq 2. \quad (6)$$

ЛЕММА 1. *Элементы матрицы B представимы в виде*

$$b_{ij} = \sum_{k=1}^j b_{i-1,k}, \quad b_{ij} = \sum_{k=1}^i b_{k,j-1}, \quad i, j \geq 2. \quad (7)$$

Доказательство. Для доказательства достаточно расписать второе рекуррентное соотношение (6), используя первое условие. \square

Ниже через $B_{s,k}$ будем обозначать матрицу с элементами (6), составленную из первых s строк и k столбцов матрицы B .

ЛЕММА 2. *Матрица $B_{m,m}$ невырожденная.*

Доказательство. Для доказательства введем в рассмотрение матрицы R_k , $2 \leq k \leq m$, с элементами r_k^{ij} , у которых все элементы равны нулю, за исключением следующих:

$$r_k^{ii} = 1, \quad 1 \leq i \leq m, \quad r_k^{i,i-1} = -1, \quad k \leq i \leq m, \quad (8)$$

и матрицу R вида $R = R_m R_{m-1} \dots R_2$. Очевидно, матрица R невырожденная. Покажем, что $RB_{m,m}$ есть верхняя треугольная матрица с единицами на главной диагонали, что доказывает лемму 2.

Сначала умножим R_2 на $B_{m,m}$. Учитывая (8), для этого нужно из второй строки матрицы $B_{m,m}$ вычесть первую, из третьей вторую и т. д., а результат

следует поместить на место соответственно второй, третьей и т. д. строк. Тогда с использованием первого соотношения (6) получим, что в первом столбце матрицы $R_2 B_{m,m}$ на первом месте стоит единица, а на остальных нули. Из второго равенства (6) имеем

$$b_{i,j-1} = b_{ij} - b_{i-1,j}, \quad 2 \leq i \leq m, \quad 3 \leq j \leq m,$$

то есть, если в матрице $R_2 B_{m,m}$ вычеркнуть первую строку и столбец, то она совпадает с $B_{m,m}$, у которой вычеркнуты первая строка и последний столбец. Умножая последовательно матрицу $R_2 B_{m,m}$ на R_3, \dots, R_m и проводя аналогичные рассуждения, получим доказательство леммы. \square

Заметим, что так как матрицы $B_{m,m}$ и R_k , $2 \leq k \leq m$, целочисленные, то R и $R B_{m,m}$ также целочисленные.

ЛЕММА 3. Стадии метода (5) представимы в виде

$$k_j = \sum_{i=1}^{\infty} a^{i-1} b_{ij} h^i f_n^{i(i-1)} f_n. \quad (9)$$

Доказательство. Доказательство проведем с использованием метода математической индукции. Для этого запишем D_n^{-1} в виде ряда Тейлора

$$D_n^{-1} = E + ahf'_n + a^2 h^2 f_n'^2 + a^3 h^3 f_n'^3 + \dots \quad (10)$$

Представление (9) при $j = 1$ очевидно. Пусть (9) выполняется при некотором j . Для определения k_{j+1} умножим (10) на (9) и соберем слагаемые при одинаковых степенях h . В результате имеем

$$k_{j+1} = \sum_{i=1}^{\infty} a^{i-1} \left(\sum_{k=1}^i b_{kj} \right) h^i f_n^{i(i-1)} f_n.$$

Воспользовавшись вторым равенством (7), завершим доказательство. \square

С применением (9) приближенное решение по схеме (5) представимо в виде

$$y_{n+1} = y_n + \sum_{i=1}^{\infty} a^{i-1} \left(\sum_{j=1}^m b_{ij} p_j \right) h^i f_n^{i(i-1)} f_n. \quad (11)$$

ТЕОРЕМА. При любом значении m нельзя построить m -стадийную схему (5) выше второго порядка точности.

Доказательство. Запишем ряд Тейлора для точного решения $y(t_{n+1})$ задачи (1) в окрестности точки t_n по степеням h до членов с h^3 включительно, то есть

$$y(t_{n+1}) = y(t_n) + hf + \frac{1}{2} h^2 f' f + \frac{1}{6} h^3 (f'^2 f + f'' f^2) + O(h^4), \quad (12)$$

где элементарные дифференциалы вычислены на точном решении $y(t_n)$. Положим $y_n = y(t_n)$. Тогда из сравнения (11) и (12) следует, что в разложении

точного решения в ряд Тейлора есть слагаемое вида $f''f^2$, а в представлении (11) оно отсутствует. \square

Отметим, что в случае линейной задачи (1) вида

$$y' = Ay + b, \quad y(t_0) = y_0,$$

где A и b — соответственно матрица и вектор с постоянными коэффициентами, имеет место соотношение

$$B_{m,m}P_m = V_m(a)$$

с невырожденной матрицей $B_{m,m}$, которое можно применять для построения методов заданного порядка точности. Здесь m — число стадий в методе (5),

$$P_m = (p_1, \dots, p_m)^\top, \quad V_m(a) = \left(1, \frac{a^{-1}}{2!}, \dots, \frac{a^{1-m}}{m!}\right)^\top.$$

4. Метод второго порядка точности. Пусть $m = 2$, то есть рассмотрим схему вида

$$y_{n+1} = y_n + p_1k_1 + p_2k_2, \quad D_nk_1 = hf(y_n), \quad D_nk_2 = k_1. \quad (13)$$

Подставляя ряды Тейлора для k_1 и k_2 в первую формулу (13), получим

$$y_{n+1} = y_n + (p_1 + p_2)hf_n + a(p_1 + 2p_2)h^2f'_nf_n + a^2(p_1 + 3p_2)h^3f''_nf_n + O(h^4).$$

Полагая $y_n = y(t_n)$ и сравнивая полученное соотношение с (12) до членов с h^2 включительно, получим условия второго порядка точности (13), то есть

$$p_1 + p_2 = 1, \quad a(p_1 + 2p_2) = 0,5. \quad (14)$$

Исследуем устойчивость (13). Для этого применим его для решения скалярного тестового уравнения $y' = \lambda y$. Учитывая условия порядка, получим $y_{n+1} = Q(x)y_n$, где функция устойчивости $Q(x)$ имеет следующий вид:

$$Q(x) = \frac{1 + (1 - 2a)x + a(a - p_1)x^2}{(1 - ax)^2}.$$

Тогда схема (13) будет L -устойчивой, если $p_1 = a$. Решая систему $p_1 = a$ и (14), имеем набор коэффициентов $p_1 = a$ и $p_2 = 1 - a$, где a удовлетворяет условию L -устойчивости:

$$a^2 - 2a + 0,5 = 0.$$

Данное уравнение имеет два корня: $a_1 = 1 - \sqrt{2}/2$ и $a_2 = 1 + \sqrt{2}/2$. Обычно в расчётах применяется корень $a = a_1$, потому что в этом случае меньше коэффициент в локальной ошибке.

Контроль точности вычислений численной схемы (13) построим по аналогии [5]. Для этого введём обозначение

$$\varepsilon(j_n) = D_n^{1-j_n}(k_2 - k_1), \quad (15)$$

где k_1 и k_2 вычисляются по формулам (13). Тогда согласно [5] для контроля точности вычислений на каждом шаге нужно проверять неравенство

$$\|\varepsilon(j_n)\| \leq \varepsilon, \quad 1 \leq j_n \leq 2, \quad (16)$$

где ε — требуемая точность расчётов, $\|\cdot\|$ — некоторая норма в \mathbb{R}^N , а целочисленная переменная j_n выбирается наименьшей, при которой выполняется неравенство (16).

Отметим одну важную особенность построенной оценки ошибки (15). Схема (13) L -устойчивая, то есть для ее функции устойчивости $Q(x)$ имеет место соотношение $Q(x) \rightarrow 0$ при $x \rightarrow -\infty$. Так как для точного решения $y(t_{n+1}) = \exp(x)y(t_n)$ задачи $y' = \lambda y$, $y(t_0) = y_0$ выполняется аналогичное свойство, то естественным будет требование стремления к нулю оценки ошибки при $x \rightarrow -\infty$. Однако для $\varepsilon(1)$ это свойство не выполняется — данная оценка ведет себя A -устойчивым образом. С целью исправления асимптотического поведения ошибки вместо $\varepsilon(1)$ введена оценка $\varepsilon(j_n)$, $1 \leq j_n \leq 2$. В этом случае поведение оценки ошибки при $j_n = 2$ будет согласовано с поведением точного решения тестовой задачи $y' = \lambda y$, $y(t_0) = y_0$ при $x \rightarrow -\infty$.

Подчеркнем, что в смысле главного члена оценки $\varepsilon(1)$ и $\varepsilon(2)$ совпадают. Использование $\varepsilon(j_n)$ фактически не приводит к увеличению вычислительных затрат. Это связано с тем, что $\varepsilon(j_n)$ при $j_n = 2$ проверяется только в том случае, если оно нарушено при $j_n = 1$. Такая ситуация встречается достаточно редко, в основном при быстром росте величины шага интегрирования. Однако это позволяет выбирать шаг более правильно и тем самым уменьшается количество неоправданных повторных вычислений решения (возвратов).

В расчётах норма $\|\xi\|$ в левой части неравенства (16) вычисляется по формуле

$$\|\xi\| = \max_{1 \leq i \leq N} \frac{|\xi_i|}{|y_n^i| + \mu}.$$

В случае выполнения неравенства $|y_n^i| < \mu$ по i -той компоненте решения контролируется абсолютная ошибка $\mu\varepsilon$, в противном случае контролируется относительная ошибка ε . Иногда с целью повышения надежности расчётов задают набор параметров μ_i , $1 \leq i \leq N$, что позволяет более квалифицированно контролировать точность расчётов.

5. Заключение. Из сравнения численной формулы типа Розенброка (3) и (2,1)-метода (13) следует, что по вычислительным затратам схема (13) отличается от (3) на один дополнительный обратный ход метода Гаусса. Однако (14) имеет второй порядок точности и, как показывают расчёты, примерно в три раза эффективнее (3) по времени счета. Кроме того, для (14) построено неравенство для контроля точности вычислений, что позволяет проводить расчёты с переменным шагом интегрирования. Это приводит к дополнительному повышению эффективности.

Численную схему (13) можно рассматривать как способ реализации неявного одностадийного метода типа Рунге—Кутта, причем в (13) нет необходимости применять итерационный процесс Ньютона, что позволяет оценить вычислительные затраты на шаг интегрирования до начала расчётов и значительно упрощает реализацию алгоритма интегрирования.

Работа выполнена при поддержке РФФИ (проект № 11-01-00106-а).

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Rosenbrock H. H.* Some general implicit processes for the numerical solution of differential equations // *Computer*, 1963. Vol. 5, no. 4. Pp. 329–330.
2. *Hairer E., Wanner G.* Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems / Springer Series in Computational Mathematics. Vol. 14. Berlin: Springer-Verlag, 1996. 614 pp.; русск. пер.: *Хайрер Э., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Жёсткие и дифференциально-алгебраические задачи. М.: Мир, 1999. 685 с.
3. *Новиков Е. А., Двинский А. Л.* Замораживание матрицы Якоби в $(3, 2)$ -методе решения жёстких систем / В сб.: *Совместный выпуск журналов «Вычислительные технологии» и «Региональный вестник Востока»*: Труды международной конференции «Вычислительные и информационные технологии в науке, технике и образовании». Часть II. Новосибирск, Алматы, Усть-Каменогорск, 2003. С. 272–278. [*Novikov E. A., Dvinskiy A. L.* Freezing of the Jacobi matrix in $(3, 2)$ -method of solving stiff systems / In: *Joint issue of “Computational Technologies” and “Regional Bulletin of the East”*: Proceedings of International Conference “Computational and Informational Technologies for Science, Engineering and Education”. Part II. Novosibirsk, Almaty, Ust’-Kamenogorsk, 2003. Pp. 272–278].
4. *Новиков Е. А., Шитов Ю. А., Шокин Ю. И.* Одношаговые безытерационные методы решения жёстких систем // *ДАН СССР*, 1988. Т. 301, № 6. С. 1310–1314; англ. пер.: *Novikov E. A., Shitov Yu. A., Shokin Yu. I.* One-step noniterative methods for solving stiff systems // *Soviet Math. Dokl.*, 1989. Vol. 38, no. 1. Pp. 212–216.
5. *Новиков Е. А.* Явные методы для жёстких систем. Новосибирск: Наука, 1997. 197 с. [*Novikov E. A.* Explicit methods for stiff systems. Novosibirsk: Nauka, 1997. 195 pp.]

Поступила в редакцию 28/I/2011;
в окончательном варианте — 17/VIII/2011.

MSC: 65L20; 65L05, 34A34

MAXIMAL ORDER OF ACCURACY OF $(m, 1)$ -METHODS FOR SOLVING STIFF PROBLEMS

E. A. Novikov^{1,2}

¹ Institute of Computational Modelling, Siberian Branch of the Russian Academy of Sciences, Akademgorodok, Krasnoyarsk, 660036.

² Siberian Federal University, 79, Svobodniy, Krasnoyarsk, 660041.

E-mail: novikov@icm.krasn.ru

We investigate $(m, 1)$ -methods for solving stiff problems in which the right part of system of the differential equations is calculated one times on each step. It is shown that the maximal order of accuracy of the L -stability $(m, 1)$ -method is equal to two, and the method of the maximal order is constructed.

Key words: *stiff problems, Rosenbrock schemes, (m, k) -methods, A -stability, L -stability.*

Original article submitted 28/I/2011;
revision submitted 17/VIII/2011.

Evgeniy A. Novikov (Dr. Sci. (Phys. & Math.)), Chief Research Scientist, Dept. of Computational Mathematics¹; Head of Dept., Dept. of Mathematical Software for Systems and Discrete Devices².