на БЭСМ-6 отлично от правильного, то относительная ошибка уменьшится, но ненамного, и будет примерно равна $3 \cdot 10^{-7}$.

Если суммирование выполнять группами по $n=1000$ чисел с последующим сложением n частичных сумм, то абсолютная ошибка оказывается равной примерно $8.62 \cdot 10^{-3}$, а относительная — примерно $8.7 \cdot 10^{-10}$, т. е. ошибка уменьшилась приблизительно в $n=10^3$ раз!

Если же применить алгоритм попарного суммирования, то результат получится с относительной точностью, приблизительно равной $6 \cdot 10^{-12}$. Эта точность превосходит теоретическую оценку уменьшения верхнего предела для ошибки, поскольку суммировались равные числа.

Другим примером накопления ошибок округления служит вычисление суммы

$$\sum_{i=1}^{N} (1/i)\, i.$$

При $N=10^6$ на ЭВМ БЭСМ-6 в режиме отбрасывания младших разрядов абсолютная ошибка суммы равна примерно 0.287, а относительная — примерно $2{,}9 \cdot 10^{-6}$. Если вычисления производить в режиме округления, то результат оказывается почти точным.

Еще раз подчеркнем здесь тот факт, что вычитание близких чисел может дать большую относительную ошибку, как это показано в примере из п. 9.1. Часто вычитания такого рода можно избежать преобразованием формул.

В заключение приведем список литературы, рекомендуемой к гл. 1: [1, 2, 7, 10, 22, 24, 27, 28, 33, 36, 37].

Глава 2

ОДНОШАГОВЫЕ МЕТОДЫ РЕШЕНИЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

В различных предметных областях часто встречаются в качестве математических моделей изучаемых процессов обыкновенные дифференциальные уравнения. Во многих случаях такие уравнения не интегрируются в явном виде. Поэтому необходимо использовать методы, позволяющие получать приближенное решение задачи. С примером таких методов знакомятся уже в общей теории обыкновенных дифференциальных уравнений, когда при доказательстве существования решения дифференциального уравнения

$$y' = f(x, y). \tag{1}$$

удовлетворяющего начальному условию

$$y(x_0) = y_0, \tag{2}$$

используется метод последовательных приближений Пикара.

Решение (1), (2) получается как предел последовательности

$$y_0(x), y_1(x), \ldots, y_n(x), \ldots,$$

где

$$y_n(x) = y_0 + \int_{x_0}^{x} f(\xi, y_{n-1}(\xi))\, d\xi, \quad y_0(x) \equiv y_0. \tag{3}$$

Функцию $y_n(x)$ с достаточно большим номером n можно принять за приближенное решение задачи (1), (2). Однако, если правая часть $f(x, y)$ уравнения (1) — сложная функция, то применение этого метода наталкивается на большие трудности, поскольку интеграл в (3) не берется в квадратурах и решение нельзя получить в аналитическом виде.

Мы будем применять *численные* методы, которые позволяют получать приближенное решение задачи в виде *таблицы* чисел.

1. Общее представление одношаговых методов. Пусть требуется найти решение задачи (1), (2) на отрезке $[x_0, x_0+X]$. Возьмем разбиение отрезка точками

$$x_0 < x_1 < x_2 < \ldots < x_N = x_0 + X.$$

Этот набор точек называется *сеткой*, а точки x_n — *узлами* сетки. Рассмотрим одношаговые численные методы, т. е. такие, которые последовательно дают приближения y_n к значениям точного решения $y(x_n)$ в каждом узле x_n сетки на основе известного приближения y_{n-1} к решению в предыдущем узле x_{n-1}. В общем виде их можно представить так:

$$y_{n+1}=F(f;\ x_{n+1},\ x_n;\ y_{n+1},\ y_n). \tag{4}$$

Мы займемся только явными одношаговыми методами, для которых функция F не зависит от y_{n+1}. Обозначая

$$h=x_{n+1}-x_n,$$

явные одношаговые методы будем записывать также в виде

$$y_{n+1}=F(f;\ h,\ x_n,\ y_n). \tag{5}$$

2. Метод рядов Тейлора. Предположим, что правая часть $f(x, y)$ дифференциального уравнения (1) имеет непрерывные частные производные до порядка s. Тогда искомое решение $y(x)$ имеет непрерывные производные до $(s+1)$-го порядка включительно. Точное значение решения в узле x_1 запишем по формуле Тейлора:

$$y(x_1)=y_0+hy_0'+\frac{h^2}{2}y_0''+\ldots+\frac{h^s}{s!}y_0^{(s)}+\frac{h^{s+1}}{(s+1)!}y^{(s+1)}(\xi), \tag{6}$$

где

$$y_0^{(k)}=y^{(k)}(x_0),\quad h=x_1-x_0,\quad x_0<\xi<x_1.$$

Может оказаться, что для получения решения с нужной точностью не требуется использовать все члены формулы (6). Производные, входящие в правую часть формулы (6), могут быть фактически найдены:

$$y_0'=f(x_0,\ y_0),$$
$$y_0''=\{f_x'+ff_y'\}_0,$$
$$y_0'''=\{f_{xx}''+2ff_{xy}''+f^2f_{yy}''+(f_x'+ff_y')f_y'\}_0.$$
$$\ldots$$

С увеличением порядка выражения для производных становятся все более громоздкими, что требует большого объема вычислений. Это является существенным недостатком данного метода.

3. Явные методы типа Рунге—Кутта. Рунге предложил следующую идею, основанную на вычислении приближенного решения y_1 в узле x_0+h в виде линейной комбинации с постоянными коэффициентами:

$$y_1=y_0+p_{q1}k_1(h)+p_{q2}k_2(h)+\ldots+p_{qq}k_q(h), \tag{7}$$

где

$$k_1(h)=hf(x_0,\ y_0),$$
$$k_2(h)=hf(x_0+\alpha_2h,\ y_0+\beta_{21}k_1(h)),$$
$$\ldots\quad\ldots\quad\ldots$$
$$k_q(h)=hf(x_0+\alpha_qh,\ y_0+\beta_{q1}k_1(h)+\ldots+\beta_{q,q-1}k_{q-1}(h)).$$

Числа α_i, β_{ij} и p_{qi} выбираются так, чтобы разложение выражения (7) по степеням h совпадало с разложением (6) до максимально возможной степени при произвольной правой части $f(x, y)$ и произвольном шаге h.

Это эквивалентно следующему. Если ввести вспомогательную функцию

$$\varphi_q(h)=y(x_0+h)-y_0-\sum_{i=1}^{q}p_{qi}k_i(h), \tag{8}$$

то ее разложение по степеням h должно начинаться с максимально возможной степени:

$$\varphi_q(h)=\frac{h^{s+1}}{(s+1)!}\varphi_q^{(s+1)}(0)+o(h^{s+1}). \tag{9}$$

Если можно определить эти постоянные так, чтобы разложение $\varphi_q(h)$ имело вид (9), то говорят, что формула (7) с выбранными коэффициентами имеет *порядок точности s*.

Величина

$$\rho_1=\varphi_q(h)=y(x_0+h)-y_1$$

называется *погрешностью метода на шаге* или *локальной погрешностью* метода, а первое слагаемое в (9)

$$\frac{h^{s+1}}{(s+1)!}\varphi_q^{(s+1)}(0) \tag{10}$$

называется *главным членом* локальной погрешности метода.

Доказано, что если $q=1, 2, 3, 4$, то всегда можно выбрать коэффициенты α_i, β_{ij}, p_{qi} так, чтобы получить метод типа Рунге—Кутта порядка точности q. При $q=5$ *невозможно* построить метод типа Рунге—Кутта (7) пятого порядка точности, необходимо брать в комбинации (7) более пяти членов.

3.1. Одночленная формула. Будем рассматривать такую формулу, когда приближенное решение в точке $x_1=x_0+h$ ищется в виде

$$y_1=y_0+p_{11}k_1(h),$$

где

$$k_1(h)=hf(x_0,\ y_0).$$

Единственный коэффициент p_{11} мы подбираем так, чтобы в разложении по степеням h функции (8), в данном случае равной

$$\varphi_1(h)=y(x_0+h)-y_0-p_{11}k_1(h),$$

максимальное количество членов обратилось в нуль. Этому требованию удовлетворяет единственное значение

$$p_{11}=1,$$

для которого разложение (9) имеет вид

$$\varphi_1(h)=\frac{h^2}{2}(f'_x+ff'_y)_0+o(h^2).$$

Таким образом, мы получили формулу первого порядка точности

$$y_1=y_0+hf(x_0, y_0).$$

Это известная формула Эйлера. Итак, метод типа Рунге—Кутта при $q=1$ есть метод Эйлера.

Геометрический смысл метода заключается в том, что интегральная кривая $y(x)$ приближается ломаной Эйлера (рис. 1).

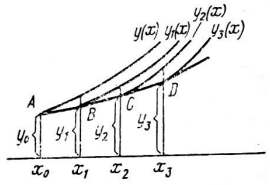

Рис. 1.
Геометрическая интерпретация метода Эйлера

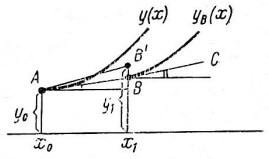

Рис. 2.
Геометрическая интерпретация метода Хойна

На $[x_0, x_1]$ интегральная кривая заменяется отрезком касательной AB, проведенной к интегральной кривой $y(x)$ в точке A. На $[x_1, x_2]$ интегральная кривая заменяется отрезком касательной BC, проведенной в точке B к интегральной кривой $y_1(x)$, проходящей через точку B. В качестве численного решения берутся ординаты вершин построенной ломаной $ABCD$. Таким образом, после выполнения каждого шага метода Эйлера приближенное решение переходит с одной интегральной кривой на другую.

3.2. Двухчленные формулы. Строим формулу следующего вида:

$$y_1=y_0+p_{21}k_1(h)+p_{22}k_2(h),$$

где

$$k_1(h)=hf(x_0, y_0),$$

$$k_2(h)=hf(x_0+\alpha_2 h, y_0+\beta_{21}k_1(h)).$$

В выписанной формуле четыре неизвестных параметра: α_2, β_{21}, p_{21}, p_{22}. Чтобы их найти, опять строим вспомогательную функцию (8), равную

$$\varphi_2(h)=y(x_0+h)-y_0-p_{21}k_1(h)-p_{22}k_2(h),$$

и подбираем коэффициенты так, чтобы в разложении этой функции по степеням h максимальное количество членов обратилось в нуль. Чтобы обратить в нуль первые две производные

$$\varphi'_2(0)=(1-p_{21}-p_{22})f(x_0, y_0),$$

$$\varphi''_2(0)=\{(1-2\alpha_2 p_{22})f'_x+(1-2\beta_{21}p_{22})ff'_y\}_{\substack{x=x_0\\y=y_0}},$$

необходимо, чтобы искомые коэффициенты удовлетворяли системе уравнений

$$\begin{aligned}1-p_{21}-p_{22}&=0,\\ 1-2\alpha_2 p_{22}&=0,\\ 1-2\beta_{21}p_{22}&=0.\end{aligned}\qquad(11)$$

Тогда разложение (9) имеет вид

$$\varphi_2(h)=\frac{h^3}{6}\varphi'''_2(0)+o(h^3).\qquad(12)$$

Как бы мы ни выбирали параметры, третью производную

$$\varphi'''_2(0)=\{(1-3\alpha_2^2p_{22})f''_{xx}+2(1-3\alpha_2\beta_{21}p_{22})ff''_{xy}+$$
$$+(1-3\beta_{21}^2p_{22})f^2f''_{yy}+(f'_x+ff'_y)f'_y\}_{\substack{x=x_0\\y=y_0}}\qquad(13)$$

обратить в нуль для произвольной функции $f(x, y)$ нельзя. Из последних двух уравнений в (11) следует, что $\alpha_2\neq 0$, $\beta_{21}\neq 0$, $\alpha_2=\beta_{21}$. Принимая α_2 за свободный параметр, имеем

$$\beta_{21}=\alpha_2,\ p_{22}=\frac{1}{2\alpha_2},\ p_{21}=1-\frac{1}{2\alpha_2}.$$

Таким образом, мы получили однопараметрическое семейство формул типа Рунге—Кутта второго порядка точности. Следует выбирать такие коэффициенты, которые дают удобные для вычисления формулы. Например, их можно выбирать так, чтобы главный член погрешности был простейшим или наименьшим. Рассмотрим несколько примеров.

1. Пусть $\alpha_2=1$. Тогда

$$\begin{aligned}y_1&=y_0+\frac{1}{2}(k_1+k_2),\\ k_1&=hf(x_0, y_0),\\ k_2&=hf(x_0+h, y_0+k_1).\end{aligned}\qquad(14)$$

Погрешность формулы (14), как следует из (12) и (13), имеет вид

$$\rho_1=\frac{h^3}{6}\left\{-\frac{1}{2}(f''_{xx}+2ff''_{xy}+f^2f''_{yy})+(f'_x+ff'_y)f'_y\right\}_{\substack{x=x_0\\y=y_0}}+o(h^3).\qquad(15)$$

Геометрический смысл формулы (14) понятен из рис. 2. Через точку A проводится не касательная к интегральной кривой $y(x)$, а прямая с угловым коэффициентом, равным среднему арифметическому угловых коэффициентов касательных AB и BC, которые строятся в методе Эйлера, и в качестве решения берется ордината точки B' пересечения этой прямой с прямой $x=x_1$. Данный метод называется методом Хойна.

Если правая часть уравнения (1) не зависит от y, то формула (14) переходит в квадратурную формулу трапеций.

2. Пусть $\alpha_2=1/2$. Тогда

$$y_1=y_0+k_2,$$
$$k_1=hf(x_0,\ y_0), \tag{16}$$
$$k_2=hf\left(x_0+\frac{1}{2}h,\ y_0+\frac{1}{2}k_1\right),$$

Погрешность формулы (16), как следует из (12) и (13), имеет вид

$$\rho_1=\frac{h^3}{6}\left\{\frac{1}{4}(f''_{xx}+2ff''_{xy}+f^2f''_{yy})+(f'_x+ff'_y)f'_y\right\}_{\substack{x=x_0\\y=y_0}}+o(h^3). \tag{17}$$

Геометрический смысл формулы (16) заключается в том, что через точку A (рис. 3) проводится прямая, угловой коэффициент которой равен угловому коэффициенту касательной BC к интегральной кривой $y_B(x)$, проходящей через промежуточную точку B, построенную по методу Эйлера с шагом $h/2$. В качестве ответа берется ордината точки D пересечения этой прямой с прямой $x=x_1$.

Если правая часть дифференциального уравнения (1) не зависит от y, то формула (16) переходит в квадратурную формулу средних прямоугольников.

3. Пусть $\alpha_2=2/3$. Тогда

$$y_1=y_0+\frac{1}{4}(k_1+3k_2),$$
$$k_1=hf(x_0,\ y_0), \tag{18}$$
$$k_2=hf\left(x_0+\frac{2}{3}h,\ y_0+\frac{2}{3}k_1\right).$$

Погрешность формулы (18), как следует из (12) и (13), имеет вид

$$\rho_1=\frac{h^3}{6}\{(f'_x+ff'_y)f'_y\}_{\substack{x=x_0\\y=y_0}}+o(h^3). \tag{19}$$

В данном случае прямая AD (рис. 4) имеет угловой коэффициент, равный не среднему арифметическому угловых коэффициентов касательных AB и BC, как в методе Хойна, а взвешенному среднему их значений. При этом касательная BC проведена к интегральной кривой $y_B(x)$, проходящей через промежуточную точку B, построенную по методу Эйлера с шагом $\frac{2}{3}h$.

Из трех формул (14), (16), (18) нельзя выбрать одну наилучшую (например, с точки зрения малости величины главного члена погрешности) для всех уравнений. Для одних уравнений предпочтительнее один метод, для других—другой.

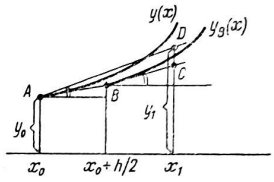

Рис. 3.
Геометрическая интерпретация метода второго порядка с $\alpha_2=1/2$

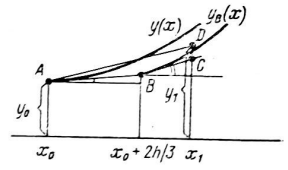

Рис. 4.
Геометрическая интерпретация метода второго порядка с $\alpha_2=2/3$

Например, в случае интегрирования уравнения

$$y'=y^\lambda,\quad y(0)>0,$$

при $\lambda>1$ главный член погрешности формулы (14) оказывается меньше главных членов погрешностей формул (16) и (18), а при $\lambda<1$ главный член погрешности формулы (16) меньше главных членов погрешностей формул (14) и (18) (в этом легко убедиться, сравнивая (15), (17) и (19) для $f(x,\ y)=y^\lambda$). В случае интегрирования уравнения

$$y'=f(x,\ y)=g(x) \tag{20}$$

первое слагаемое в локальной погрешности метода (18), как видно из (19), вообще равно нулю, поэтому метод (18) имеет для данного класса уравнений более высокий порядок точности по сравнению с методами (14) и (16).

3.3. Трехчленные формулы. Строим формулу следующего вида:

$$y_1=y_0+p_{31}k_1(h)+p_{32}k_2(h)+p_{33}k_3(h),$$

где

$$k_1(h)=hf(x_0,\ y_0),$$
$$k_2(h)=hf(x+\alpha_2h,\ y_0+\beta_{21}k_1(h)),$$
$$k_3(h)=hf(x_0+\alpha_3h,\ y_0+\beta_{31}k_1(h)+\beta_{32}k_2(h)).$$

Восемь неизвестных параметров выбираются так, чтобы разложение (9) функции (8) начиналось с максимально возможной степени.

Для отыскания этих восьми параметров получается система из шести уравнений

$$\alpha_2=\beta_{21},$$
$$\alpha_3=\beta_{31}+\beta_{32},$$
$$p_{31}+p_{32}+p_{33}=1,$$
$$2(\alpha_2 p_{32}+\alpha_3 p_{33})=1,$$
$$3(\alpha_2^2 p_{32}+\alpha_3^2 p_{33})=1,$$
$$6\alpha_2\beta_{32}p_{33}=1.$$

Она имеет два семейства решений: двухпараметрическое со свободными параметрами α_2 и α_3, причем $\alpha_2\neq\alpha_3$ и $\alpha_2\neq 2/3$, и однопараметрическое со свободным параметром β_{32} (при $\alpha_2=\alpha_3=2/3$). Для таким образом найденных параметров разложение (9) имеет вид

$$\varphi_3(h)=\frac{h^4}{24}\varphi_3^{(4)}(0)+o(h^4).$$

Пример. Пусть $\alpha_2=1/2$, $\alpha_3=1$. Тогда

$$\begin{aligned} y_1&=y_0+\frac{1}{6}(k_1+4k_2+k_3),\\ k_1&=hf(x_0,\ y_0),\\ k_2&=hf\left(x_0+\frac{1}{2}h,\ y_0+\frac{1}{2}k_1\right),\\ k_3&=hf(x_0+h,\ y_0-k_1+2k_2). \end{aligned} \qquad (21)$$

Для уравнения (20) формула (21) превращается в квадратурную формулу Симпсона, которая, как известно, имеет порядок точности $O(h^5)$. Следовательно, в этом частном случае порядок точности формулы (21) повышается.

3.4. Четырехчленные формулы. Существует семейство формул типа Рунге—Кутта (7) четвертого порядка точности, для которых $q=s=4$. В качестве примера приведем формулу классического метода Рунге—Кутта

$$\begin{aligned} y_1&=y_0+\frac{1}{6}(k_1+2k_2+2k_3+k_4),\\ k_1&=hf(x_0,\ y_0),\\ k_2&=hf\left(x_0+\frac{1}{2}h,\ y_0+\frac{1}{2}k_1\right),\\ k_3&=hf\left(x_0+\frac{1}{2}h,\ y_0+\frac{1}{2}k_2\right),\\ k_4&=hf(x_0+h,\ y_0+k_3). \end{aligned} \qquad (22)$$

Погрешность формулы (22), как и всех формул Рунге—Кутта четвертого порядка точности, представляется следующим образом:

$$\rho_1=\frac{h^5}{120}\varphi_4^{(5)}(0)+o(h^5),$$

где

$$\varphi_4(h)=y(x_0+h)-y_0-p_{41}k_1(h)-p_{42}k_2(h)-p_{43}k_3(h)-p_{44}k_4(h).$$

Геометрическая интерпретация классического метода Рунге—Кутта (22) дана на рис. 5. Прямая AG, проходящая через точку $(x_0,\ y_0)$, имеет угловой коэффициент, равный взвешенному среднему угловых коэффициентов касательных в точках A, B, D и F, проведенных к проходящим через эти точки интегральным кривым $y(x)$, $y_B(x)$, $y_D(x)$, $y_F(x)$.

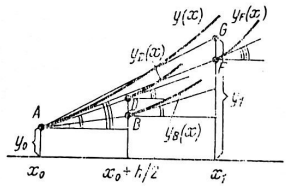

Рис. 5.
Геометрическая интерпретация классического метода Рунге—Кутта

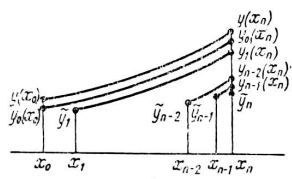

Рис. 6.
Полная погрешность приближенного решения $\tilde{y}_n$ в точке x_n

3.5. Формулы порядка выше четвертого. Чтобы построить метод типа Рунге—Кутта (7) порядка выше четвертого, для которого погрешность на шаге имеет порядок выше пятого для произвольного дифференциального уравнения (1) с достаточно гладкой правой частью $f(x,\ y)$, необходимо, чтобы в формуле (7) число слагаемых $k_i(h)$ было больше пяти: $q>5$. Примеры таких формул приводятся в п. 5.2.2.3.

Здесь существенно то, что значения правой части $k_i(h)$ дифференциального уравнения входят в формулу (7) в виде линейной комбинации. Может быть построен явный одношаговый метод пятого порядка точности $y_1=y_0+\Delta y$, использующий пять вычислений правой части, но не имеющий вида (7), так как значения $k_i(h)$ будут входить в Δy нелинейным образом.

Общим для всех рассмотренных методов численного интегрирования является то обстоятельство, что после выполнения очередного шага точка, изображающая полученное приближенное решение на плоскости $(x,\ y)$, переходит с одной интегральной кривой на другую.

Методы Рунге—Кутта без труда переносятся на системы обыкновенных дифференциальных уравнений

$$\bar{y}' = \bar{f}(x, \bar{y}). \tag{23}$$

где

$$\bar{y} = (y^1, y^2, \ldots, y^M)^T, \quad \bar{f} = (f^1(x, y^1, \ldots, y^M), \cdots, f^M(x, y^1, \ldots, y^M))^T,$$

Формулы Рунге—Кутта записываются в векторном виде

$$\bar{y}_1 = \bar{y}_0 + p_{q1}\bar{k}_1(h) + p_{q2}\bar{k}_2(h) + \ldots + p_{qq}\bar{k}_q(h), \tag{24}$$

где

$$\bar{k}_1(h) = h\bar{f}(x_0, \bar{y}_0),$$

$$\bar{k}_2(h) = h\bar{f}(x_0 + \alpha_2 h, \bar{y}_0 + \beta_{21}\bar{k}_1(h)),$$

. . . . . .

$$\bar{k}_q(h) = h\bar{f}(x_0 + \alpha_q h, \bar{y}_0 + \beta_{q1}\bar{k}_1(h) + \ldots + \beta_{q,q-1}\bar{k}_{q-1}(h)),$$

Всюду в дальнейшем черточка над обозначением векторов не ставится.

4. Сходимость явных одношаговых методов. *4.1. Классификация погрешностей.* Прежде чем перейти к классификации погрешностей, вернемся еще раз к обсуждению исходной задачи (1), (2). Предположим, что требуется найти функцию $y(x)$, которая является решением дифференциального уравнения (1)

$$y'(x) = f(x, y(x)), \quad x_0 \leqslant x \leqslant x_0 + X,$$

и принимает в точке x_0 некоторое определенное значение (2)

$$y(x_0) = y_0.$$

Может оказаться, что начальное условие $y(x_0)$ известно неточно, а определяется в результате эксперимента, например, с помощью измерений или в результате решения какой-либо другой задачи. В этом случае вместо точного начального условия $y(x_0)$ приходится использовать его приближенное значение $\tilde{y}_0$, а вместо задачи Коши (1), (2) решать задачу

$$y'_0(x) = f(x, y_0(x)), \tag{25}$$

$$y_0(x_0) = \tilde{y}_0 \tag{26}$$

с измененным начальным условием

$$y_0 - \tilde{y}_0 = R_0 \neq 0. \tag{27}$$

Решение задачи (25), (26) зависит от $\tilde{y}_0$ и не совпадает с искомым решением $y(x)$ задачи (1), (2). Разность

$$\zeta_n = y(x_n) - y_0(x_n) \tag{28}$$

называется *неустранимой погрешностью* решения $y_0(x)$.

Разность между значением решения $y_0(x_n)$ задачи (25), (26) и его приближенным значением y_n, полученным по формуле (5),

$$\varepsilon_n = y_0(x_n) - y_n \tag{29}$$

называется *погрешностью метода,* или *глобальной погрешностью* метода.

В действительности же вследствие ошибок округления и приближенного вычисления правой части $f(x, y)$ дифференциального уравнения вычисления значений y_{n+1} по формуле (5) выполняются, как правило, неточно. Фактически найденные значения $\tilde{y}_n$ удовлетворяют не соотношению (5), а условию

$$F(f;\ h,\ x_{n-1},\ \tilde{y}_{n-1}) - \tilde{y}_n = \delta_n. \tag{30}$$

Невязка δ_n называется *погрешностью округления на n-м шаге.*

Разность между точным решением $y(x_n)$ задачи (1), (2) и приближенным фактически найденным значением y_n

$$R_n = y(x_n) - \tilde{y}_n \tag{31}$$

называется *полной погрешностью* приближенного решения. Величина

$$\eta_n = y_n - \tilde{y}_n \tag{32}$$

называется *вычислительной погрешностью.*

Из соотношений (28), (29), (31) и (32) следует, что

$$R_n = \zeta_n + \varepsilon_n + \eta_n, \tag{33}$$

т. е. полная погрешность приближенного решения равна сумме неустранимой погрешности, погрешности метода и вычислительной погрешности.

Рассмотрим поведение полной погрешности. Представим R_n в следующем виде (рис. 6):

$$R_n = y(x_n) - \tilde{y}_n = (y(x_n) - y_{n-1}(x_n)) + (y_{n-1}(x_n) - \tilde{y}_n) =$$

$$= (y(x_n) - y_{n-2}(x_n)) + (y_{n-2}(x_n) - y_{n-1}(x_n)) + (y_{n-1}(x_n) - y_n(x_n)) =$$

$$= (y(x_n) - y_0(x_n)) + \sum_{j=1}^{n} (y_{j-1}(x_n) - y_j(x_n)). \tag{34}$$

Здесь $y_j(x)$ — интегральная кривая, проходящая через точку $(x_j, \tilde{y}_j)$. Разность

$$\omega_j(x) = y_{j-1}(x) - y_j(x)$$

двух решений уравнения (1) удовлетворяет линейному дифференциальному уравнению

$$(y_{j-1}(x) - y_j(x))' = f'_y(x, \hat{y}_j)(y_{j-1}(x) - y_j(x)),$$

где $\hat{y}_j(x)$ заключено между $y_{j-1}(x)$ и $y_j(x)$. Из этого уравнения находим

$$\omega_j(x)=y_{j-1}(x)-y_j(x)=\omega_j(x_j)\,e^{\int_{x_j}^{x}\frac{\partial f}{\partial y}(\xi,\widehat{y_j}(\xi))\,d\xi}. \qquad (35)$$

Точно так же с учетом (27)

$$y(x_n)-y_0(x_n)=R_0 e^{\int_{x_0}^{x_n}\frac{\partial f}{\partial y}(\xi,\widehat{y_0}(\xi))\,d\xi}. \qquad (36)$$

Значение

$$\omega_j(x_j)=y_{j-1}(x_j)-y_j(x_j)$$

представляет собой сумму локальной погрешности ρ_j метода и погрешности округления δ_j на шаге x_j-x_{j-1}:

$$\omega_j(x_j)=\rho_j+\delta_j. \qquad (37)$$

Подставляя представления (35) и (36) с учетом (37) в выражение (34) для R_n, получаем

$$R_n=\sum_{j=1}^{n}(\rho_j+\delta_j)\,e^{\int_{x_j}^{x_n}\frac{\partial f}{\partial y}(\xi,\widehat{y_j}(\xi))\,d\xi}+R_0 e^{\int_{x_0}^{x_n}\frac{\partial f}{\partial y}(\xi,\widehat{y_0}(\xi))\,d\xi}. \qquad (38)$$

Из (38) следует, что полная погрешность приближенного решения задачи (1), (2) в точке x_n равна сумме локальных погрешностей на каждом шаге, взятых с коэффициентами

$$e^{\int_{x_j}^{x_n}\frac{\partial f}{\partial y}(\xi,\widehat{y_j}(\xi))\,d\xi}.$$

Из соотношения (38) следует, что характер отклонения приближенного решения от точного, т. е. эволюция полной погрешности, зависит от поведения интегральных кривых уравнения (1). Если $f_y'=0$, а это бывает в том случае, когда правая часть уравнения не зависит от y (20), полная погрешность равна сумме локальных погрешностей:

$$R_n=R_0+\sum_{j=1}^{n}\omega_j(x_j)=R_0+\sum_{j=1}^{n}(\rho_j+\delta_j).$$

Если $f_y'>0$, т. е. интегральные кривые расходятся, влияние локальных погрешностей, полученных на предыдущих шагах, возрастает и полная погрешность больше суммы локальных погрешностей. Если $f_y'<0$, т. е. интегральные кривые сближаются, влияние локальных погрешностей ослабевает и глобальная ошибка как правило меньше суммы локальных ошибок.

Такое поведение характерно как для полной погрешности, так и для отдельных частей, из которых складывается полная погрешность: неустранимой погрешности, погрешности метода и вычислительной погрешности. Рис. 7 и 8 иллюстрируют влияние $\operatorname{sign}\left(\frac{\partial f}{\partial y}\right)$ на характер эволюции глобальной погрешности метода.

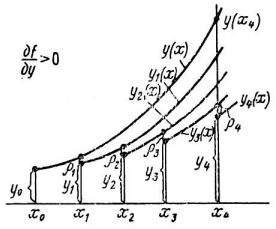

Рис. 7.
Интегральные кривые расходятся. Глобальная погрешность метода больше суммы локальных погрешностей:
$y(x_4)-y_4>\rho_1+\rho_2+\rho_3+\rho_4$

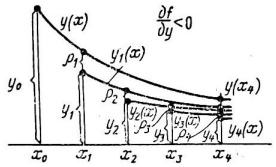

Рис. 8.
Интегральные кривые сближаются. Глобальная погрешность метода меньше суммы локальных погрешностей:
$y(x_4)-y_4<\rho_1+\rho_2+\rho_3+\rho_4$

4.2. Мажорантная оценка полной погрешности. Опираясь на (38), можно получить мажорантную оценку полной погрешности приближенного решения задачи Коши (1), (2):

$$|R_n|\leqslant e^{Lx}\Big(|R_0|+\sum_{j=1}^{n}(|\rho_j|+|\delta_j|)\Big), \qquad (39)$$

где

$$L=\sup\left|\frac{\partial f}{\partial y}\right|<\infty.$$

Учитывая оценку для локальной погрешности метода s-го порядка

$$|\rho_j|=O(h^{s+1})\leqslant C|x_j-x_{j-1}|^{s+1},$$

имеем

$$|R_n|\leqslant e^{Lx}(|R_0|+CXh^s+n\delta), \qquad (40)$$

где

$$\delta=\max_{1\leqslant j\leqslant n}|\delta_j|,\quad h=\max_{1\leqslant j\leqslant n}|x_j-x_{j-1}|.$$

В случае системы уравнений (23) для метода Рунге—Кутта (24) имеет место мажорантная оценка

$$\|R_n\| \leqslant e^{MLX}(\|R_0\| + CXh^s + n\delta), \tag{41}$$

где

$$\|R_n\| = \max_{1\leqslant i\leqslant M} |R_n^i|, \quad \delta = \max_{1\leqslant j\leqslant n} \|\delta_j\|.$$

Из соотношений (40), (41) следует, что приближенное решение задачи Коши, полученное по методу Рунге—Кутта порядка s, сводится к точному решению задачи при $h\to 0$, если

$$n\delta \to 0, \quad |R_0| \to 0 \quad (\|R_0\| \to 0). \tag{42}$$

Рассмотрим смысл каждого из трех слагаемых в (40), (41) и условия (42). Присутствие в (40), (41) слагаемого $|R_0|(\|R_0\|)$ означает, что погрешность от начального значения распространится на все узлы сетки. Часть полной погрешности приближенного решения, зависящая от ошибки в начальных условиях и называемая *неустранимой погрешностью*, не превосходит $|R_0|e^{LX}(\|R_0\| \times e^{MLX})$. Если начальные значения решения заданы точно, то этого члена нет.

Второй член получается за счет того, что мы находим не точное решение задачи, а приближение к нему по формуле Рунге—Кутта. Это погрешность метода, и она имеет порядок h^s.

Третий член получается за счет ошибок округления. Часть полной погрешности приближенного решения, источником которой являются ошибки округления, как уже указывалось, называется вычислительной погрешностью. Скорость возрастания вычислительной погрешности не превосходит

$$\frac{e^{LX_n}\delta}{nh} = \frac{e^{LX}\delta}{h} \quad \left(\text{или } \frac{e^{MLX}\delta}{h}\right).$$

Если величина δ ограничена снизу: $0<\delta_0\leqslant\delta$, а в практике вычислений на ЭВМ так обычно и бывает, и при этом длина шага h слишком мала, а значит, число шагов очень велико, то вычислительная погрешность может достигать больших значений.

В действительности же ошибки округления δ_j могут иметь различные знаки и частично компенсировать друг друга. Разные знаки могут иметь и погрешности метода ρ_j. Отдельные составляющие, входящие в полную погрешность R_n (38), могут давать отклонения от точного решения в разные стороны. Поэтому оценка (40) (или (41)) по сравнению с (38) является завышенной. Кроме того, ее применение затруднено еще и из-за сложности определения величины C, выражаемой через производные высокого порядка от правой части $f(x, y)$ уравнения (1). Поэтому данная оценка на практике не используется для определения точности окончательного результата.

Из приведенных оценок можно сделать следующий вывод, подтверждаемый практикой численного решения на ЭВМ дифференциальных уравнений. Если заранее обеспечена необходимая малость неустранимой погрешности, то в полной погрешности преобладает либо погрешность метода, либо вычислительная погрешность. Погрешность метода может быть сделана сколь угодно малой за счет уменьшения шага h. Вычислительная погрешность может быть снижена за счет увеличения числа используемых в промежуточных вычислениях *значащих цифр* (значащими цифрами называются все цифры в записи числа, начиная с первой ненулевой). Однако наши практические возможности в этом далеко не беспредельны из-за конечности разрядной сетки машины.

Если же погрешность (27) в начальных условиях велика, то неустранимая погрешность может также оказаться значительной и ее нельзя будет уменьшить никаким сокращением длины шага интегрирования, так как она не зависит от численного решения задачи. Неустранимую погрешность можно уменьшить только за счет более точного определения начальных условий. Поэтому остается только надеяться, что неустранимая погрешность будет незначительной по абсолютной величине по сравнению с другими видами погрешности.

Из вышесказанного следует, что вычислительный процесс должен быть организован таким образом, чтобы поддерживался баланс между всеми видами погрешности, составляющими полную погрешность приближенного решения. Этот баланс может быть достигнут, если надлежащим образом будут согласованы между собой требуемая точность решения задачи, точность задания начальных условий, порядок численного метода, величина шага интегрирования и используемая длина разрядной сетки ЭВМ.

4.3. Асимптотическая оценка погрешности метода. Пусть правая часть $f(x, y)$ уравнения (1) имеет непрерывные частные производные до порядка $s+2$. Предположим, что задача Коши (1), (2) решается с постоянным шагом

$$x_{i+1} - x_i = h_i = h$$

методом Рунге—Кутта порядка s (7), так что для локальной погрешности (8) метода справедлива оценка (9), которую запишем в следующем виде:

$$y(x_{n+1}) - y_{n+1} = \psi(x_n, y_n)h^{s+1} + O(h^{s+2}), \tag{43}$$

где

$$\psi(x_n, y_n) = \frac{1}{(s+1)!}\varphi_q^{(s+1)}(0)\Big|_{\substack{x=x_n\\ y=y_n}}.$$

Будем считать равными нулю погрешность начального условия и погрешности округлений: $R_0=0$ (27), $\delta_n=0$ (30). Тогда для рас-

бальной погрешности (29) метода имеет место асимптотическое представление

$$\varepsilon_n = z(x_n)h^s + O(h^{s+1}), \tag{44}$$

где

$$z(x_n) = \int_{x_0}^{x_n} \psi(\xi, y(\xi))\, e^{\int_{\xi}^{x_n} \frac{\partial f}{\partial y}(\tau, y(\tau))\,d\tau}\, d\xi. \tag{45}$$

Заметим, что $\psi(x, y) \not\equiv 0$ и $z(x) \not\equiv 0$. Более того, если $\psi(x, y)$ не меняет знака, то $z(x) \neq 0$ ни при каком x. Первое слагаемое в (44)

$$z(x_n)h^s, \tag{46}$$

которое указывает малую величину главного порядка, называется *главным членом погрешности* метода. При достаточно малых значениях h, таких, что в (44) можно пренебречь членом $O(h^{s+1})$, для погрешности метода справедлива формула

$$\varepsilon_n \cong z(x_n)h^s. \tag{47}$$

4.4. Реальная область асимптотики. Рассмотрим, когда главный член (46) погрешности метода хорошо отражает полную погрешность (33) приближенного решения. Для этого необходимо, чтобы члены порядка $O(h^{s+1})$, входящие в (44), наряду с вычислительной погрешностью η_n были малы по сравнению с главным членом. При этом неустранимая погрешность ζ_n также должна быть малой по отношению к нему. Условия, обеспечивающие близость главного члена погрешности метода к полной погрешности приближенного решения, математически можно записать в виде

$$R_0 = O(h^{s+1}), \quad \delta = O(h^{s+2}). \tag{48}$$

Если предположить, что условия (48) выполняются при $h \to 0$, то для полной погрешности приближенного решения имеет место асимптотическое разложение вида (44)

$$R_n = z(x_n)h^s + O(h^{s+1}). \tag{49}$$

При этом разложения (44) и (49) отличаются друг от друга только членами порядка $O(h^{s+1})$. При достаточно малых значениях h, таких, что в (49) можно пренебречь членом $O(h^{s+1})$, для полной погрешности справедлива формула

$$R_n \cong z(x_n)h^s. \tag{50}$$

Множество таких значений шага h называется *областью асимптотики.*

Однако в реальных условиях вычислительного процесса требования (48), как правило, не выполняются (по крайней мере, второе). Погрешность R_0 задания начальных условий вообще не зависит от длины шага интегрирования, а погрешности округления на шаге (30) при решении задачи на ЭВМ с фиксированной разрядной сеткой остаются ограниченными по абсолютной величине снизу

$$0 < \delta_0 \leqslant \delta.$$

Нарушение требований (48) при $h \to 0$ приводит к тому, что неустранимая погрешность будет ограничена снизу, а вычислительная погрешность может даже возрастать при $h \to 0$. В результате формулы (49) и (50) перестают быть справедливыми.

Поэтому в реальном процессе численного решения задачи Коши на ЭВМ множество значений шагов, для которых главный член (46) погрешности метода (29) хорошо представляет полную погрешность (33) приближенного решения (при достаточной малости неустранимой погрешности), т. е. для которых справедлива формула (50), ограничено не только сверху, но и снизу:

$$0 < \underline{h} < h < \bar{h}. \tag{51}$$

При переходе через верхнюю границу $\bar{h}$ в сторону увеличения значений шагов растет вклад членов в правой части формулы (49), содержащих высшие степени h, и формула (50) становится неверной. При переходе через нижнюю границу $\underline{h}$ в сторону уменьшения значений шагов увеличивается вычислительная погрешность η_n (32) и формула (50) снова перестает быть действительной. Таким образом, формула (50) для полной погрешности приближенного решения может оказаться несправедливой как при больших ($h > \bar{h}$), так и при очень малых ($h < \underline{h}$) значениях шага h.

Множество значений шагов (51) называется *реальной областью асимптотики.* Для фиксированной задачи (1), (2) и фиксированного метода решения (7) нижняя граница зависит от разрядной сетки вычислительной машины. Чем шире разрядная сетка, тем, вообще говоря, меньше нижняя граница реальной области асимптотики. Расширить разрядную сетку можно за счет выполнения вычислений с удвоенным числом значащих цифр.

5. Практические способы оценки погрешности приближенного решения. *5.1. Апостериорная оценка глобальной погрешности метода.* Предположим, что для погрешности приближенного решения справедливо асимптотическое разложение (49), (50). На этом разложении основывается важный для практики способ Рунге *апостериорной* оценки погрешности. Правило Рунге состоит в том, что решение задачи в некоторой точке x_n интервала интегрирования вычисляется дважды по одной и той же формуле (7) с разными малыми шагами и полученные значения решения используются для получения апостериорной оценки погрешности. Обычно в качестве шагов выбирают h и $h/2$.

Допустим, что в точке x_n по формуле Рунге—Кутта (7) с постоянным шагом h вычислено решение $\bar{y}_n$. На основании (50) погрешность этого решения приближенно равна

$$y(x_n) - \bar{y}_n \cong z(x_n)h^s.$$

Используя ту же формулу с шагом $h/2$, вычислим в точке x_n другое значение решения $\bar{y}_n$, для которого потребуется в два раза больше шагов и погрешность которого приближенно выражается равенством

$$y(x_n)-y_n \cong z(x_n)\left(\frac{h}{2}\right)^s.$$

Исключая из этих соотношений точное значение решения $y(x_n)$, имеем

$$\bar{\bar{y}}_n-\bar{y}_n \cong z(x_n)h^s\left(1-\frac{1}{2^s}\right).$$

Отсюда

$$z(x_n) \cong \frac{\bar{\bar{y}}_n-\bar{y}_n}{h^s\left(1-\frac{1}{2^s}\right)}. \qquad (52)$$

Окончательно получаем оценки погрешности для приближенных значений решения

$$\bar{R}_n=y(x_n)-\bar{y}_n \cong (\bar{\bar{y}}_n-\bar{y}_n)\Big/\left(1-\frac{1}{2^s}\right), \qquad (53)$$

$$\bar{\bar{R}}_n=y(x_n)-\bar{\bar{y}}_n \cong (\bar{\bar{y}}_n-y_n)/(2^s-1). \qquad (54)$$

Полученное приближенное значение можно уточнить, прибавив к нему величину главного члена погрешности, т. е. положив

$$y(x_n) \cong y_n=\bar{y}_n+\bar{R}_n$$

или

$$y(x_n) \cong y_n=\bar{\bar{y}}_n+\bar{\bar{R}}_n.$$

При этом

$$y(x_n)-y_n=O(h^{s+1}).$$

В случае системы уравнений (23) правило Рунге оценки погрешности (53), (54) может быть записано в координатной форме следующим образом:

$$\bar{R}_n^i=y^i(x_n)-\bar{y}_n^i \cong (\bar{\bar{y}}_n^i-\bar{y}_n^i)\Big/\left(1-\frac{1}{2^s}\right), \quad i=1, 2, \ldots, M, \qquad (55)$$

$$\bar{\bar{R}}_n^i=y^i(x_n)-\bar{\bar{y}}_n^i \cong (\bar{\bar{y}}_n^i-\bar{y}_n^i)/(2^s-1), \quad i=1, 2, \ldots, M. \qquad (56)$$

В качестве решения в точке x_n примем значение $\bar{\bar{y}}_n$ как более точное по сравнению с $\bar{y}_n$. Для него имеем оценку погрешности (54). Эта величина может быть как больше, так и меньше некоторого значения ε, являющегося наперед заданной *допустимой погрешностью*. Если $|\bar{\bar{R}}_n| \leqslant \varepsilon$, то заданная точность приближенного решения достигается, в противном случае — нет *. Если точность не достигается, то необходимо решить вопрос, какую нужно взять длину шага, чтобы все-таки достигнуть заданной точности.

Если же точность достигается, естественно поставить такой вопрос: можно ли увеличить длину шага интегрирования для того, чтобы уменьшить объем вычислительной работы и одновременно с этим сохранить заданную точность? В обоих случаях такую величину шага h_ε можно определить, если положить

$$|z(x_n)|h_\varepsilon^s=\varepsilon.$$

Отсюда находим

$$h_\varepsilon=\sqrt[s]{\varepsilon/|z(x_n)|}. \qquad (57)$$

Подставляя в (57) выражение (52) для $z(x_n)$, получаем

$$h_\varepsilon \cong \sqrt[s]{\frac{\varepsilon h^s(2^s-1)}{|\bar{\bar{y}}_n-y_n|2^s}}=\frac{h}{2}\sqrt[s]{\frac{(2^s-1)\varepsilon}{|\bar{\bar{y}}_n-\bar{y}_n|}}. \qquad (58)$$

Из (58) следует, что если $|\bar{\bar{R}}_n|>\varepsilon$, то новое значение шага уменьшается:

$$h_\varepsilon<h/2,$$

и если $|\bar{R}_n|<\varepsilon$, то новое значение шага увеличивается:

$$h_\varepsilon>h/2.$$

Таким образом, формула (58) дает более подходящее значение шага интегрирования.

5.2. Апостериорные оценки локальной погрешности метода. В предыдущем разделе мы познакомились с практическим способом оценки глобальной погрешности метода, когда решение задачи вычисляется в некоторой фиксированной точке интервала интегрирования. Теперь обсудим, как можно практически оценить локальную погрешность метода. Мы рассмотрим несколько способов, а начнем с известного нам правила Рунге.

5.2.1. Оценка погрешности по правилу Рунге.

Правило Рунге заключается в том, что по одной и той же формуле (7) вычисляются два приближения к решению в одной точке, но с разными малыми шагами, которые затем используются для получения апостериорной оценки погрешности.

Пусть в начальном узле x_0 известно решение y_0. Выполним из точки x_0 один шаг h по формуле Рунге—Кутта (7). Полученное в точке $x_1=x_0+h$ решение обозначим $\bar{y}_1$. Для локальной погрешности метода воспользуемся формулой (43). Если пренебречь членами порядка $O(h^{s+2})$, то для погрешности метода на данном шаге справедлива формула

$$y(x_0+h)-\bar{y}_1 \cong \psi(x_0, y_0)h^{s+1}. \qquad (59)$$

Вернемся в исходную точку x_0, и из точки x_0 сделаем подряд два шага, каждый величиной $h/2$. Сделав такой шаг первый раз,

* Контроль точности для системы уравнений (23) подробно излагается в п. 6.3.

получим приближение $\hat{y}$ к решению в точке $x_0+h/2$. Тогда погрешность метода на этом шаге согласно (43) равна

$$y(x_0+h/2)-\hat{y}\cong\psi(x_0,\ y_0)(h/2)^{s+1}. \tag{60}$$

Сделав шаг $h/2$ второй раз, но уже из точки $x_0+h/2$, получим приближение $\bar{y}_1$ к решению в точке $x_1=x_0+h$. На этом втором шаге погрешность метода равна

$$\hat{y}(x_0+h)-\bar{\bar{y}}_1\cong\psi(x_0+h/2,\ \hat{y})(h/2)^{s+1}, \tag{61}$$

где $\hat{y}(x)$ — точное решение уравнения (1), удовлетворяющее условию $\hat{y}(x_0+h/2)=\hat{y}$.

Так как мы используем малую длину шага h, то точка $(x_0+h/2,\ \hat{y})$ находится близко от точки $(x_0,\ y_0)$. Поэтому в силу ограниченности по предположению частных производных ψ_x' и ψ_y' главный член погрешности метода на втором шаге будет таким же, как и на первом:

$$\hat{y}(x_0+h)-\bar{\bar{y}}_1\cong\psi(x_0,\ y_0)(h/2)^{s+1}. \tag{61'}$$

Разность

$$y(x)-\hat{y}(x)$$

двух решений уравнения (1) удовлетворяет линейному дифференциальному уравнению

$$(y(x)-\hat{y}(x))'=f_y'(x,\ \check{y})(y(x)-\hat{y}(x)),$$

где $\check{y}(x)$ заключено между $y(x)$ и $\hat{y}(x)$, и может быть представлено в виде

$$y(x)-\hat{y}(x)=(y(x_0+h/2)-\hat{y}(x_0+h/2))\,e^{\int_{x_0+h/2}^{x}\frac{\partial f}{\partial y}(\xi,\check{y}(\xi))\,d\xi}.$$

Отсюда следует, что

$$y(x_0+h)-\hat{y}(x_0+h)=(y(x_0+h/2)-\hat{y}(x_0+h/2))\,e^{\int_{x_0+h/2}^{x_0+h}\frac{\partial f}{\partial y}(\xi,\check{y}(\xi))\,d\xi}=$$
$$=y(x_0+h/2)-\hat{y}(x_0+h/2)+O(h(y(x_0+h/2)-\hat{y}(x_0+h/2)))\cong$$
$$\cong y(x_0+h/2)-\hat{y}(x_0+h/2)=y(x_0+h/2)-\hat{y}.$$

Теперь может быть найдена погрешность метода на двух последовательных шагах $h/2$:

$$y(x_0+h)-\bar{\bar{y}}_1=(y(x_0+h)-\hat{y}(x_0+h))+(\hat{y}(x_0+h)-\bar{\bar{y}}_1)\cong$$
$$\cong(y(x_0+h/2)-\hat{y})+(\hat{y}(x_0+h)-\bar{\bar{y}}_1).$$

Подставляя в полученное соотношение правые части формул (60) и (61'), окончательно получаем

$$y(x_0+h)-\bar{\bar{y}}_1\cong 2\psi(x_0,\ y_0)(h/2)^{s+1}. \tag{62}$$

Из (59) и (62) вытекают представления главных членов погрешностей метода на шаге h и на двух последовательных шагах $h/2$:

$$\psi(x_0,\ y_0)h^{s+1}\cong(\bar{\bar{y}}_1-\bar{y}_1)/(1-1/2^s) \tag{63}$$

и

$$2\psi(x_0,\ y_0)(h/2)^{s+1}\cong(\bar{\bar{y}}_1-\bar{y}_1)/(2^s-1). \tag{64}$$

Если в качестве приближения к решению в точке x_1 принять $\bar{y}_1$, то локальная погрешность метода равна

$$y(x_0+h)-\bar{y}_1\cong(\bar{\bar{y}}_1-\bar{y}_1)/(1-1/2^s). \tag{65}$$

Если в качестве приближения к решению в точке x_1 принять $\bar{\bar{y}}_1$, то погрешность метода на двух последовательных шагах $h/2$ равна

$$y(x_0+h)-\bar{\bar{y}}_1\cong(\bar{\bar{y}}_1-\bar{y}_1)/(2^s-1). \tag{66}$$

Из (63), (64) видно, что выражения для погрешностей отличаются друг от друга только значениями знаменателей. При этом оценка (64) дает меньшее по абсолютной величине значение погрешности, чем оценка (63). Следовательно, значение $\bar{\bar{y}}_1$, полученное на двух последовательных шагах $h/2$, является более точным приближением к решению $y(x_1)$ по сравнению со значением $\bar{y}_1$, полученным за один шаг h. Конечно, сказанное справедливо только при достаточно малой величине шага интегрирования h.

В случае системы уравнений (23) правило Рунге оценки локальной погрешности (65), (66) может быть записано в координатной форме следующим образом:

$$y^i(x_0+h)-\bar{y}_1^i\cong(\bar{\bar{y}}_1^i-\bar{y}_1^i)/(1-1/2^s),\quad i=1,\ 2,\ \ldots,\ M, \tag{65'}$$

$$y^i(x_0+h)-\bar{\bar{y}}_1^i\cong(\bar{\bar{y}}_1^i-\bar{y}_1^i)/(2^s-1),\quad i=1,\ 2,\ \ldots,\ M. \tag{66'}$$

Полученное приближенное значение $\bar{y}_1$ или $\bar{\bar{y}}_1$ можно уточнить, прибавив к нему величину главного члена погрешности, т. е. положив

$$y(x_1)\cong y_1=\bar{y}_1+(\bar{\bar{y}}_1-\bar{y}_1)/(1-1/2^s) \tag{67}$$

или

$$y(x_1)\cong y_1=\bar{\bar{y}}_1+(\bar{\bar{y}}_1-\bar{y}_1)/(2^s-1). \tag{67'}$$

Тогда

$$y(x_1)-y_1=O(h^{s+2}).$$

В данном способе оценки погрешности формула Рунге—Кутта (7) применяется три раза и требует $3q-1$ вычислений правой части $f(x, y)$ дифференциального уравнения (1). Поэтому при сложных и трудоемких для вычисления правых частях этот способ влечет большие вычислительные затраты.

5.2.2. Оценка погрешности на основе комбинации формул разных порядков точности.

Рассмотрим второй способ оценки локальной погрешности метода. Этот способ также основан на использовании двух приближенных значений решения в одной точке. Однако эти приближения в отличие от правила Рунге вычисляются не по одной, а по двум формулам разных порядков точности p и s с одним и тем же шагом. Начнем обсуждение данного способа с общего случая, когда применяемые формулы Рунге—Кутта не связаны друг с другом, а потом перейдем к рассмотрению специально подобранных формул.

5.2.2.1. Комбинация независимых формул. Данный способ основан на комбинации двух формул вида (7) разных порядков точности p и s:

$$y_1^p=y_0+\sum_{i=1}^{r} p_i k_i, \tag{68}$$

где

$$k_1=hf(x_0, y_0),$$

$$k_i=hf\left(x_0+\alpha_i h,\ y_0+\sum_{j=1}^{i-1}\beta_{ij}k_j\right),$$

и

$$y_1^s=y_0+\sum_{i=1}^{\tilde r}\tilde p_i\tilde k_i, \tag{69}$$

где

$$\tilde k_1=hf(x_0, y_0),$$

$$\tilde k_i=hf\left(x_0+\tilde\alpha_i h,\ y_0+\sum_{j=1}^{i-1}\tilde\beta_{ij}\tilde k_j\right).$$

Пусть $p>s$, $r\geqslant\tilde r$. Локальные погрешности в этих формулах имеют вид

$$\rho^p=y(x_0+h)-y_1^p=O(h^{p+1})$$

и

$$\rho^s=y(x_0+h)-y_1^s=O(h^{s+1}).$$

Из последних равенств следует оценка локальной погрешности формулы (69)

$$\rho^s=y_1^p-y_1^s+O(h^{p+1}). \tag{70}$$

Оставляя в (70) только члены главного порядка, имеем

$$\rho^s\cong y_1^p-y_1^s. \tag{71}$$

Полученная оценка погрешности требует $r+\tilde r-1$ вычислений правой части уравнения (1).

5.2.2.2. Комбинация специально подобранных формул. Если коэффициенты в формулах (68) и (69) таковы, что

$$\alpha_i=\tilde\alpha_i,\quad \beta_{ij}=\tilde\beta_{ij},\quad i=1, 2, \ldots, \tilde r, \tag{72}$$

то

$$k_i=\tilde k_i,\quad i=1, 2, \ldots, \tilde r,$$

и для локальной погрешности (71) формулы (69) получается выражение следующего вида:

$$\rho^s\cong y_1^p-y_1^s=\sum_{i=1}^{r} q_i k_i, \tag{73}$$

где

$$\begin{aligned} q_i&=p_i-\tilde p_i,\quad i=1, 2, \ldots, \tilde r,\\ q_i&=p_i,\quad i=\tilde r+1, \ldots, r. \end{aligned} \tag{73'}$$

Оценка (73) помимо тех значений правой части, которые вычисляются на текущем шаге h и входят в формулу (69), включает дополнительные значения k_i, $i=\tilde r+1, \ldots, r$. Такой подход к оценке локальной погрешности позволяет уменьшить по сравнению с правилом Рунге (65) и оценкой (71) количество вычислений правой части уравнения (1).

Оценка (73), как и более общая оценка (71), является асимптотической, так как она учитывает только члены главного порядка, и справедлива при достаточно малых размерах шага интегрирования.

В практике вычислений в качестве приближенного значения решения принимается значение y_1^p как имеющее более высокий порядок точности.

Величина

$$E=\sum_{i=1}^{r} q_i k_i, \tag{74}$$

где коэффициенты q_i определяются с помощью (73'), называется *контрольным членом*. В случае системы уравнений (23) конт-

рольный член может быть записан в координатной форме следующим образом:

$$E^j=\sum_{i=1}^{r} q_i k_i^j, \quad j=1, 2, \ldots, M. \tag{74'}$$

5.2.2.3. Контрольные члены для методов Рунге—Кутта. Оценка погрешности в методах Мерсона, Ингленда, Фельберга. Рассмотрим несколько примеров.

1) Методы (21)

$$y_1=y_0+\frac{1}{6}(k_1+4k_2+k_3)$$

и (16)

$$y_1=y_0+k_2$$

удовлетворяют условию (72). Контрольный член (74) записывается в данном случае в виде

$$E=\frac{1}{6}(k_1-2k_2+k_3) \tag{75}$$

и имеет порядок $O(h^3)$. Здесь $p=3$, $s=2$, $r=3$, $\tilde{r}=2$. Если оценку локальной погрешности метода (16) вести по правилу Рунге (65), то для этого потребуется пять обращений к правой части вместо трех.

2) Классический метод Рунге—Кутта (22)

$$y_1=y_0+\frac{1}{6}(k_1+2k_2+2k_3+k_4)$$

и метод второго порядка

$$y_1=y_0+\frac{1}{2}(-k_1+2k_2+2k_3-k_4) \tag{76}$$

удовлетворяют условию (72). Контрольный член (74) записывается в данном случае в виде

$$E=\frac{2}{3}(k_1-k_2-k_3+k_4) \tag{77}$$

и имеет порядок $O(h^3)$. Здесь $p=4$, $s=2$, $r=4$, $\tilde{r}=4$. Он известен как контрольный член Егорова.

3) Классический метод Рунге—Кутта (22) и метод (16) также удовлетворяют условию (72). Контрольный член записывается в виде

$$E=\frac{1}{6}(k_1-4k_2+2k_3+k_4) \tag{78}$$

и имеет порядок $O(h^3)$. Здесь $p=4$, $s=2$, $r=4$, $\tilde{r}=2$.

4) Мерсон предложил следующую модификацию классического метода Рунге—Кутта:

$$\begin{aligned}
y_1&=y_0+\frac{1}{6}(k_1+4k_4+k_5),\\
k_1&=hf(x_0, y_0),\\
k_2&=hf\left(x_0+\frac{1}{3}h, y_0+\frac{1}{3}k_1\right),\\
k_3&=hf\left(x_0+\frac{1}{3}h, y_0+\frac{1}{6}k_1+\frac{1}{6}k_2\right),\\
k_4&=hf\left(x_0+\frac{1}{2}h, y_0+\frac{1}{8}k_1+\frac{3}{8}k_3\right),\\
k_5&=hf\left(x_0+h, y_0+\frac{1}{2}k_1-\frac{3}{2}k_3+2k_4\right).
\end{aligned} \tag{79}$$

Формула (79) и формула третьего порядка

$$y_1=y_0+\frac{1}{10}(k_1+3k_3+4k_4+2k_5). \tag{80}$$

удовлетворяют условию (72). Контрольный член (74) записывается в данном случае в виде

$$E=\frac{1}{30}(2k_1-9k_3+8k_4-k_5) \tag{81}$$

и имеет порядок $O(h^4)$. Здесь $p=4$, $s=3$, $r=5$, $\tilde{r}=5$. Если бы контроль точности для метода третьего порядка производился по правилу Рунге (65), то это потребовало бы восьми вычислений правой части вместо пяти в методе Мерсона.

На более узком, чем (1), классе линейных уравнений вида

$$y'=f(x, y)=ax+by+c \tag{82}$$

формула (80) имеет не третий, а пятый порядок точности. Порядок формулы (79) остается по-прежнему равным четырем. Поэтому величина

$$\rho=-E=-\frac{1}{30}(2k_1-9k_3+8k_4-k_5) \tag{83}$$

служит главным членом локальной погрешности формулы (79) и имеет порядок $O(h^5)$. Таким образом, метод Мерсона (79) обладает той особенностью, что на классе уравнений (82) главный член его локальной погрешности выражается с помощью линейной комбинации только тех значений правой части, которые непосредственно входят в формулу (79). Следовательно, для получения оценки погрешности отпадает необходимость в дополнительных вычислениях.

5) Фельберг разработал множество методов, удовлетворяющих условию (72). Приведем предложенные им формулы четвертого и пятого порядков:

$$y_1^{(5)}=y_0+\frac{16}{135}k_1+\frac{6656}{12\,825}k_3+\frac{28\,561}{56\,430}k_4-\frac{9}{50}k_5+\frac{2}{55}k_6, \quad (84)$$

$$y_1^{(4)}=y_0+\frac{25}{216}k_1+\frac{1408}{2565}k_3+\frac{2197}{4104}k_4-\frac{1}{5}k_5, \quad (85)$$

$$k_1=hf(x_0,\ y_0),$$

$$k_2=hf\left(x_0+\frac{1}{4}h,\ y_0+\frac{1}{4}k_1\right),$$

$$k_3=hf\left(x_0+\frac{3}{8}h,\ y_0+\frac{3}{32}k_1+\frac{9}{32}k_2\right),$$

$$k_4=hf\left(x_0+\frac{12}{13}h,\ y_0+\frac{1932}{2197}k_1-\frac{7200}{2197}k_2+\frac{7296}{2197}k_3\right),$$

$$k_5=hf\left(x_0+h,\ y_0+\frac{439}{216}k_1-8k_2+\frac{3680}{513}k_3-\frac{845}{4104}k_4\right),$$

$$k_6=hf\left(x_0+\frac{1}{2}h,\ y_0-\frac{8}{27}k_1+2k_2-\frac{3544}{2565}k_3+\frac{1859}{4104}k_4-\frac{11}{40}k_5\right).$$

Контрольный член (74) записывается в виде

$$E=\frac{1}{360}k_1-\frac{128}{4275}k_3+\frac{127}{6840}k_4+\frac{1}{50}k_5+\frac{2}{55}k_6 \quad (86)$$

и имеет порядок $O(h^5)$. Здесь $p=5$, $s=4$, $r=6$, $\tilde{r}=5$. Если оценку погрешности для метода четвертого порядка производить по правилу Рунге, то это потребует одиннадцати вычислений правой части вместо шести в методе Фельберга.

6) Инглендом построены следующие формулы четвертого и пятого порядков:

$$y_1^{(5)}=y_0+\frac{1}{336}(14k_1+35k_4+162k_5+125k_6), \quad (87)$$

$$y_1^{(4)}=y_0+\frac{1}{6}(k_1+4k_3+k_4), \quad (88)$$

$$k_1=hf(x_0,\ y_0),$$

$$k_2=hf\left(x_0+\frac{h}{2},\ y_0+\frac{1}{2}k_1\right),$$

$$k_3=hf\left(x_0+\frac{1}{2}h,\ y_0+\frac{1}{4}(k_1+k_2)\right),$$

$$k_4=hf(x_0+h,\ y_0-k_2+2k_3),$$

$$k_5=hf\left(x_0+\frac{2}{3}h,\ y_0+\frac{1}{27}(7k_1+10k_2+k_4)\right),$$

$$k_6=hf\left(x_0+\frac{1}{5}h,\ y_0+\frac{1}{625}(28k_1-125k_2+546k_3+54k_4-378k_5)\right).$$

Контрольный член

$$E=\frac{1}{336}(-42k_1-224k_3-21k_4+162k_5+125k_6) \quad (89)$$

имеет порядок $O(h^5)$. Здесь $p=5$, $s=4$, $r=6$, $\tilde{r}=4$. Выигрыш в количестве вычислений правой части по сравнению с правилом Рунге при оценке локальной погрешности метода такой же, как и в методе Фельберга.

5.2.3. Оценка погрешности с помощью нелинейного контрольного члена. Для приведенных в предыдущем разделе методов оценка локальной погрешности выражалась с помощью линейной комбинации (74) нескольких значений правой части дифференциального уравнения. Существуют ли другие оценки локальной погрешности, отличные от (74) и содержащие меньшее количество значений k_i правой части? Оказывается такие оценки существуют. В качестве примера можно привести формулу типа Рунге—Кутта четвертого порядка точности:

$$y_1=y_0+\frac{17}{162}k_1+\frac{81}{170}k_3+\frac{32}{135}k_4+\frac{250}{1377}k_5,$$

$$k_1=hf(x_0,\ y_0),$$

$$k_2=hf\left(x_0+\frac{2}{9}h,\ y_0+\frac{2}{9}k_1\right),$$

$$k_3=hf\left(x_0+\frac{1}{3}h,\ y_0+\frac{1}{12}k_1+\frac{1}{4}k_2\right),$$

$$k_4=hf\left(x_0+\frac{3}{4}h,\ y_0+\frac{3}{128}(23k_1-81k_2+90k_3)\right),$$

$$k_5=hf\left(x_0+\frac{9}{10}h,\ y_0+\frac{9}{10\,000}(-345k_1+2025k_2-1224k_3+544k_4)\right),$$

и выражение для погрешности формулы на шаге

$$\rho_1=y(x_0+h)-y_1=O(h^5)$$

в виде нелинейного контрольного члена Скрэтона:

$$\rho_1=\frac{uv}{w}+O(h^6),$$

где

$$u=-\frac{1}{18}k_1+\frac{27}{170}k_3-\frac{4}{15}k_4+\frac{25}{153}k_5,$$

$$v=\frac{19}{24}k_1-\frac{27}{8}k_2+\frac{57}{20}k_3-\frac{4}{15}k_4,$$

$$w=k_4-k_1.$$

Полученное значение решения y_1 можно уточнить, если прибавить к нему оценку локальной погрешности, т. е. в качестве приближенного решения взять сумму

$$y_1^{(5)}=y_1+\rho_1.$$

В результате получаем новую формулу вида (5) пятого порядка точности

$$y_1^{(5)}=y_0+\frac{17}{162}k_1+\frac{81}{170}k_3+\frac{32}{135}k_4+\frac{250}{1377}k_5+\frac{uv}{w},$$

использующую пять вычислений правой части на одном шаге. Но эта формула уже не является формулой типа Рунге—Кутта (7), поскольку величина

$$\Delta y_0=y_1^{(5)}-y_0$$

не выражается линейно через k_i.

6. Интегрирование с переменным шагом. Автоматический выбор шага интегрирования. До сих пор мы рассматривали такой процесс решения задачи (1), (2), когда формула Рунге—Кутта применялась с одной и той же величиной шага интегрирования во всей области вычисления решения вне всякой зависимости от характера поведения решения. Это были шаги h либо $h/2$ в случае повторного счета для оценки погрешности по правилу Рунге (53), либо h_ε (58). Какая бы из этих величин ни использовалась, она не изменялась от точки к точке.

В этом случае говорят, что решение задачи получено с *постоянным шагом интегрирования*. Применение *переменного шага интегрирования* позволяет учитывать характер поведения решения и уменьшить общее число шагов, сохранив при этом требуемую точность приближенного решения. Тем самым могут быть снижены объем работы и машинное время и замедлен рост вычислительной погрешности.

Имея в распоряжении способы (65), (66), (71), (73) оценки локальной погрешности метода, величину шага интегрирования можно выбирать автоматически в процессе счета. При этом можно исходить из того, чтобы на каждый шаг приходилась приблизительно одинаковая погрешность. Наиболее простой и распространенный алгоритм автоматического выбора шага является предметом нашего дальнейшего обсуждения.

6.1. Алгоритм выбора с помощью удвоения и деления шага пополам. Пусть ρ_{n+1} — оценка локальной погрешности метода на шаге h, допущенной при вычислении приближенного значения решения y_{n+1}^h в точке x_n+h. Если оценка превосходит некоторую наперед заданную границу ε:

$$|\rho_{n+1}|>\varepsilon, \tag{90}$$

то считается, что значение y_{n+1}^h решения не удовлетворяет предписанной точности и шаг h объявляется неприемлемым. Полученная точка x_n+h и значение y_{n+1}^h исключаются из рассмотрения. Выбирается новое значение шага

$$h^{(1)}=h/2,$$

и вновь по той же формуле Рунге—Кутта с шагом $h^{(1)}$ вычисляется новое значение решения $y_{n+1}^{h^{(1)}}$ в новой точке $x_n+h^{(1)}$.

Пусть $\rho_{n+1}^{(1)}$— оценка локальной погрешности метода на данном шаге $h^{(1)}$. Если оценка опять превосходит заданную границу ε:

$$|\rho_{n+1}^{(1)}|>\varepsilon,$$

то точка $x_n+h^{(1)}$ и значение $y_{n+1}^{h^{(1)}}$ опять исключаются из рассмотрения, шаг снова делится пополам:

$$h^{(2)}=h^{(1)}/2$$

и вычисления повторяются. Так происходит до тех пор, пока при какой-то величине шага (обозначим ее через h_n) оценка локальной погрешности не станет меньше ε:

$$|\rho_{n+1}|\leqslant\varepsilon. \tag{91}$$

После этого считается, что решение дифференциального уравнения продолжено до точки $x_{n+1}=x_n+h_n$. Дальнейшее интегрирование уравнения производится из точки x_{n+1} с шагом h_{n+1}, который выбирается описанным ниже способом.

Если оценка локальной погрешности на шаге $h_n=x_{n+1}-x_n$ удовлетворяет неравенству

$$|\rho_{n+1}|<\varepsilon/K, \tag{92}$$

где K — некоторая константа, то считается, что достигнута точность, превышающая заданную, и шаг интегрирования удваивается:

$$h_{n+1}=2h_n.$$

Если выполняется неравенство

$$\varepsilon/K\leqslant|\rho_{n+1}|\leqslant\varepsilon, \tag{93}$$

то считается, что полученное в точке x_{n+1} решение удовлетворяет заданной точности и шаг интегрирования остается без изменения

$$h_{n+1}=h_n.$$

Таким образом, на тех участках изменения независимой переменной, где достигается высокая точность приближенного решения, шаг интегрирования возрастает, а там, где точность не достигается, шаг интегрирования сокращается до необходимых для ее достижения значений. Тем самым обеспечивается выбор величины шага в зависимости от характера поведения решения дифференциального уравнения. Константа K обычно полагается равной 2^{ν}, где ν — порядок используемой оценки локальной погрешности метода. Константы, определяющие переход к удвоению шага для различных формул Рунге—Кутта, приведены в таблице.

Формула Рунге — Кутта и ее порядок s	Формула для уточнения решения и ее порядок p	Формула для оценки локальной погрешности метода и порядок оценки	Константа удвоения шага K	Способ оценки локальной погрешности метода
(16) $s=2$	(21) $p=3$	(75) 3	8	Контрольный член
(76) $s=2$	(22) $p=4$	(77) 3	8	Контрольный член Егорова
(16) $s=2$	(22) $p=4$	(78) 3	8	Контрольный член
(80) $s=3$	(79) $p=4$	(81) 4	16	Контрольный член в методе Мерсона
(79) $s=4$	(80) $p=5$	(83) 5	32	Контрольный член в методе Мерсона для $y'=ax+by+c$
(85) $s=4$	(84) $p=5$	(86) 5	32	Контрольный член в методе Фельберга
(88) $s=4$	(87) $p=5$	(89) 5	32	Контрольный член в методе Ингленда
(7) s	y s	(65), (66) $s+1$	2^{s+1}	Правило Рунге

Иногда для того чтобы сократить число неприемлемых шагов, в изложенный здесь алгоритм выбора шага вносится изменение, которое заключается в следующем. Если при продолжении решения из точки x_n в точку $x_{n+1}=x_n+h_n$ шаг интегрирования сокращался хотя бы один раз, то при выборе следующего значения шага h_{n+1} удвоения предыдущего шага h_n не происходит, даже если и выполняется соотношение (92).

6.2. Выбор максимальной для заданной точности длины шага. Рассмотрим еще один алгоритм выбора шага, применив ту же идею, которая была использована при выводе формулы (58). Так как оценка ρ_{n+1} локальной погрешности метода равна с точностью до членов более высокого порядка малости главному члену локальной погрешности метода, то в силу (43)

$$\rho_{n+1}\cong\psi(x_n, y_n)h^{s+1}. \tag{94}$$

Соотношение (94) справедливо для всех оценок, выведенных в п. 5.2 (см. также таблицу). Если оценка ρ_{n+1} погрешности превосходит заданную границу ε:

$$|\rho_{n+1}|>\varepsilon,$$

то считается, что на данном шаге h метод не достигает требуемой точности и вычисленное значение y_{n+1} решения вместе с точкой x_n+h исключается из рассмотрения. В этом случае выбирается новый размер шага, но не последовательным делением пополам, как в вышеописанном способе, а с помощью соотношения

$$h_{\varepsilon}=\alpha h, \tag{95}$$

где α находится из условия выполнения равенства

$$|\psi(x_n, y_n)h_{\varepsilon}^{s+1}|=\varepsilon. \tag{96}$$

Из (94), (96) получаем, что

$$\alpha^{s+1}=\varepsilon/|\rho_{n+1}|$$

и

$$\alpha=\sqrt[s+1]{\varepsilon/|\rho_{n+1}|}. \tag{97}$$

Здесь $\alpha<1$ и новое значение шага

$$h_{\varepsilon}=\sqrt[s+1]{\frac{\varepsilon}{|\rho_{n+1}|}\cdot h} \tag{98}$$

меньше предыдущего. Далее по формуле Рунге—Кутта из точки x_n выполняется один шаг h_{ε} и вычисляется приближение $y_{n+1}^{h_{\varepsilon}}$ к решению дифференциального уравнения в точке x_n+h_{ε}.

Если первоначальная оценка ρ_{n+1} локальной погрешности метода не превосходит заданную границу ε:

$$|\rho_{n+1}|\leqslant\varepsilon,$$

то считается, что полученное приближение y_{n+1} к решению удовлетворяет требуемой точности и значение x_n+h независимой переменной принимается в качестве следующего узла x_{n+1} интервала интегрирования. Дальнейшее интегрирование уравнения осуществляется из точки x_{n+1} с шагом h_{ε}, который определяется с помощью соотношений (95), (97). Теперь $\alpha\geqslant 1$ и $h_{\varepsilon}\geqslant h$.

Преимущество данного алгоритма выбора шага заключается в большей гибкости по сравнению с описанным в предыдущем разделе способом. Напомним, что в нем при достижении требуемой точности абсолютная величина шага интегрирования либо увеличивается в два раза, либо не изменяется в зависимости от того, выполняется или нет неравенство (92). Если это неравенство не выполняется из-за незначительного превышения оценки $|\rho_{n+1}|$ над величиной ε/K, то шаг интегрирования не увеличивается и остается прежним. В алгоритме, основанном на использовании формул (95), (97), имеется возможность увеличения шага в любое число α раз даже тогда, когда это число меньше двух. Это приводит к более сглаженному изменению шага интегрирования и, как следствие, к сокращению общего количества шагов и снижению вычислительных затрат.

В действительности берется несколько меньшее, чем определяемое с помощью (97), значение α, например

$$\alpha^* = 0.9\alpha = 0.9\sqrt[s+1]{\varepsilon/|\rho_{n+1}|}, \qquad (97')$$

и соответственно меньшее по сравнению с (95), (98) значение шага интегрирования

$$h_\varepsilon^* = \alpha^* h. \qquad (95')$$

Это делается для того, чтобы избежать тех шагов, для которых не достигается требуемая точность.

6.3. Использование различных характеристик точности. Автоматический выбор шага интегрирования имеет важный аспект, о котором надо помнить при практической реализации алгоритма на ЭВМ. Дело в том, что для достижения заданной точности может (в зависимости от алгоритма выбора шага) происходить большое число делений пополам шага интегрирования или величина α, вычисляемая по формуле (97), оказывается настолько малой, что вновь определяемая длина шага h не вызывает изменения независимой переменной, т. е. в условиях машинной арифметики выполняется равенство

$$x \oplus h = x, \qquad (99)$$

где $\oplus$ — машинная операция арифметического сложения, а точнее, сложение чисел с плавающей точкой, т. е. чисел вида

$$x = \pm \beta^p (a_1\beta^{-1} + a_2\beta^{-2} + \ldots + a_t\beta^{-t}), \qquad (100)$$

здесь β — основание системы счисления, p — порядок числа, удовлетворяющий неравенству

$$\underline{p} \leqslant p \leqslant \bar{p}.$$

Для ЭВМ БЭСМ-6 $\beta=2$, $\underline{p}=-64$, $\bar{p}=63$.

Равенство (99) выполняется тогда, когда текущее значение h станет по абсолютной величине меньше или минимального положительного числа, представимого на данной ЭВМ и равного $\sigma=\beta^{\underline{p}-1}$, или некоторого числа $r>0$, равного расстоянию от x до соседнего справа (при $h>0$) или слева (при $h<0$) вещественного числа, которое представимо на ЭВМ. Можно показать, что

$$r = r(x,\ t) = \beta^{1-t}\cdot\beta^{p-1} = \text{macheps}\cdot\beta^{p-1} \leqslant \text{macheps}\cdot|x|, \qquad (101)$$

где

$$\text{macheps} = \beta^{1-t},$$

называемое *машинным эпсилон*, равно расстоянию от 1 до соседнего справа вещественного числа, представимого на ЭВМ. Таким образом, если шаг интегрирования h удовлетворяет условию

$$|h| \geqslant \max\{\sigma,\ r(x,\ t)\}, \qquad (102)$$

то при этом h равенство (99) не выполняется. Без проверки условия (99) во время работы программы с автоматическим выбором шага может произойти зацикливание. Этого можно избежать, если вместо верхней границы локальной погрешности метода, которая может оказаться слишком малой в сравнении с порядком искомого решения, задавать число верных цифр в приближенном значении решения. Это позволяет более осторожно подходить к определению точности приближенного решения с учетом длины разрядной сетки машины.

Чтобы прояснить ситуацию, напомним определение *верных цифр* числа. Цифра a_k в приближенном числе

$$y^* = a_1\beta^{p-1} + a_2\beta^{p-2} + \ldots + a_m\beta^{p-m} + \ldots$$

считается *верной*, если *абсолютная погрешность* A_{y^*} числа y^* удовлетворяет неравенству

$$\rho = A_{y^*} \leqslant \omega\beta^{p-k}, \qquad (103)$$

где ω — некоторое число, удовлетворяющее условию $1/2 \leqslant \omega \leqslant 1$. Если $\omega=1$, то абсолютная погрешность числа y^* не превосходит единицы разряда, соответствующего цифре a_k, т. е. β^{p-k}.

Ясно, что, задавая ограничение сверху для ошибки в виде допустимой абсолютной погрешности, мы тем самым фиксируем разряд, соответствующий самой младшей верной цифре числа, а не число верных цифр в нем. В результате количество требуемых верных цифр в приближенном решении становится зависимым от порядка p искомого решения и может превзойти длину t разрядной сетки вычислительной машины и быть недостижимым на данной ЭВМ.

Поясним сказанное на примере. Предположим, что допустимая абсолютная погрешность метода составляет $1/2\cdot10^{-5}$, а точное решение задачи в некоторой точке интервала интегрирования принимает значение с десятичным порядком $p=5$, например 10000.33333... . Допустим, что вычисления ведутся на ЭВМ с

семью десятичными знаками. Заданная верхняя граница абсолютной погрешности соответствует требованию, чтобы приближенное значение решения имело по крайней мере пять верных знаков после запятой. Но так как решение имеет порядок $p=5$, то общее количество верных десятичных цифр должно быть не менее десяти, что превышает возможности условий проведения счета. Вообще, если допустимая абсолютная погрешность равна $\omega \cdot 10^{-k}$, порядок решения равен p, а количество используемых при вычислении десятичных знаков равно t, то, для того чтобы заданная точность была достижима, необходимо выполнение соотношения

$$t-p>k. \tag{104}$$

Как видно, для рассмотренного примера это соотношение не выполняется.

В таких случаях более целесообразно использовать в алгоритме выбора шага не абсолютную, а *относительную погрешность*

$$\rho/|y|, \tag{105}$$

так как, требуя, чтобы решение имело k верных цифр (десятичных), мы тем самым требуем, чтобы относительная погрешность этого решения не превосходила $\omega \cdot 10^{1-k}$. Однако здесь надо следить за тем, чтобы решение не обращалось в нуль.

Более гибким является использование меры погрешности. *Мерой погрешности* приближенного значения y называется величина μ, определяемая соотношением

$$\mu = \begin{cases} \rho/|y|, & |y|>P, \\ \rho, & |y|\leqslant P, \end{cases} \tag{106}$$

где P — некоторое фиксированное положительное число. Контроль точности по мере погрешности состоит в том, что на тех участках интервала интегрирования, где абсолютная величина решения не превосходит некоторого значения P, контроль точности ведется по абсолютной погрешности, а там, где абсолютная величина решения превосходит это значение, контроль точности ведется по относительной погрешности.

Другим эффективным средством может служить ограничение на число последовательных делений шага интегрирования, совершаемых в одной точке, или ограничение снизу на величину шага интегрирования. Например, ограничивая количество делений двадцатью, мы допускаем максимальное уменьшение шага в 10^6 раз. В большинстве случаев этого вполне достаточно, если, конечно, не нарушается требование (102), обеспечивающее невыполнение условия (99). Если и после двадцати делений точность по-прежнему не достигается, то, как правило, это свидетельствует о том, что явный метод типа Рунге—Кутта не подходит для решения данной задачи (конечно, если в программе нет ошибок).

Такая ситуация обычно встречается при интегрировании дифференциальных уравнений, у которых

$$\max \left| \frac{\partial f}{\partial y} \right| \gg 1.$$

К подобным уравнениям относятся уравнения и системы уравнений, правая часть которых удовлетворяет условию Липшица

$$|f^i(x, y^1, \ldots, y^M) - f^i(x, z^1, \ldots, z^M)| \leqslant L \cdot \sum_{i=1}^{M} |y^i - z^i| \tag{107}$$

с большой константой L:

$$L \gg 1, \tag{107'}$$

в частности *жесткие* системы (см. п. 7).

Указанный прием позволяет получить в процессе счета полезную информацию о характере решаемой задачи, обнаружить сильное измельчение шага интегрирования, избежать большого увеличения общего числа шагов и, как следствие, чрезмерного возрастания машинного времени.

Для системы уравнений (23) с проверкой на точность могут вычисляться либо все компоненты решения, либо некоторые из них, в частности одна компонента. При этом контроль точности может вестись покомпонентно или по норме. В первом случае каждая компонента проверяется на точность отдельно от остальных, причем для разных компонент могут использоваться как различные характеристики точности (абсолютная, относительная погрешность, мера погрешности), так и разные допустимые значения погрешности. В последнем случае ε — вектор $(\varepsilon^1, \varepsilon^2, \ldots, \ldots, \varepsilon^{M'})$, длина которого M' равна количеству проверяемых на точность компонент решения, а неравенства (90), (92), (93) будут своими для каждой компоненты решения:

$$|\rho_{n+1}^{i_j}| > \varepsilon^j, \tag{108}$$

$$|\rho_{n+1}^{i_j}| < \varepsilon^j/K, \tag{109}$$

$$\varepsilon^j/K \leqslant |\rho_{n+1}^{i_j}| \leqslant \varepsilon^j. \tag{110}$$

Здесь $j=1, \ldots, M'$, $M' \leqslant M$, i_j — номера проверяемых на точность компонент, $1 \leqslant i_j \leqslant M$. Решение о делении шага пополам принимается, если условие (108) выполняется хотя бы для одной компоненты, а решение об удвоении шага принимается тогда, когда условие (109) выполняется для всех этих компонент.

Те компоненты решения, которые проверяются на точность по мере погрешности (106), могут иметь каждая свое значение P для перехода от абсолютной погрешности к относительной и обратно. В этом случае P — вектор $(P^1, P^2, \ldots, P^{M''})$, длина которого

M'' равна количеству проверяемых на точность по мере погрешности компонент решения, а мера погрешности определяется отдельно для каждой компоненты:

$$\mu^{l_j}=\begin{cases}\rho^{l_j}/|y^{l_j}|, & |y^{l_j}|>P^j,\\ \rho^{l_j}, & |y^{l_j}|\leqslant P^j.\end{cases} \tag{111}$$

Здесь $j=1, \ldots, M''$, $M''\leqslant M'\leqslant M$, l_j — номера компонент, проверяемых на точность по мере погрешности, $i_1\leqslant l_j\leqslant i_{M'}$.

Контроль точности по норме означает, что контролируется некоторая норма оценки погрешности $\|\rho\|$. Часто используются нормы

$$\|\rho\|_\infty=\max_{1\leqslant i\leqslant M}|\rho^i|,$$

$$\|\rho\|_1=\sum_{i=1}^{M}|\rho^i|,$$

$$\|\rho\|_2=\sqrt{\sum_{i=1}^{M}|\rho^i|^2}.$$

В этом случае ε — скаляр, а неравенства (90), (92), (93) и соотношение (97′) записываются соответственно в виде

$$\|\rho_{n+1}\|>\varepsilon,$$

$$\|\rho_{n+1}\|<\varepsilon/K,$$

$$\varepsilon/K\leqslant\|\rho_{n+1}\|\leqslant\varepsilon,$$

$$\alpha^*=\underline{0.9}\alpha=\underline{0.9}\sqrt[s+1]{\varepsilon/\|\rho_{n+1}\|}.$$

Из вышеизложенного следует, что формулы Рунге—Кутта очень хорошо приспособлены для интегрирования с переменным шагом, так как они позволяют легко менять шаг интегрирования и при этом не требуют никаких дополнительных вычислений и преобразований. Этим они выгодно отличаются от других методов таких, как конечно-разностные методы (см. гл. 3), в которых при изменении шага интегрирования требуются дополнительные вычислительные затраты.

7. Чего не могут явные методы Рунге—Кутта? Рассмотрим случай линейных уравнений

$$y'=\lambda y, \tag{112}$$

$$y(0)=y_0 \tag{113}$$

с $\lambda<0$. Для метода Эйлера имеем

$$y_{n+1}=y_n+h\lambda y_n=(1+\lambda h)\,y_n. \tag{114}$$

Точное решение задачи (112), (113)

$$y(x)=e^{\lambda x}\cdot y_0$$

монотонно убывает, $y(x)\to 0$, сохраняя знак, когда x растет. Естественно требовать, чтобы решение разностной задачи (114), (113) обладало таким же свойством.

Очевидно, что условием такого поведения решения является выполнение неравенства

$$0<1+\lambda h<1,$$

или

$$|\lambda|h<1. \tag{115}$$

Для метода второго порядка точности (14) имеем

$$y_{n+1}=y_n+\frac{1}{2}(\lambda h y_n+\lambda h\,(y_n+\lambda h y_n))=(1+\lambda h+\lambda^2h^2/2)\,y_n. \tag{116}$$

Чтобы решение y_n задачи (116), (113) монотонно убывало, необходимо выполнение условия

$$0<1+\lambda h+\lambda^2h^2/2<1.$$

Отсюда следует неравенство, которому должна удовлетворять длина шага h:

$$-2<\lambda h<0,$$

или

$$|\lambda|h<2. \tag{117}$$

Такое же условие (117) на величину шага интегрирования получается и для метода (16). Ограничения, аналогичные (115) и (117), должны выполняться, хотя и с несколько большей константой, и для методов Рунге—Кутта более высокого порядка (7), которые, подобно (114) и (116), могут быть записаны в виде

$$y_{n+1}=F_q(\lambda h)\,y_n, \tag{118}$$

где $F_q(\lambda h)$ — многочлен степени q от λh.

Таким образом, применение методов типа Рунге—Кутта (7) для решения задачи (112), (113) возможно только при выполнении условия

$$|\lambda|h\leqslant\text{const}. \tag{119}$$

Рассмотрим систему линейных уравнений с постоянными коэффициентами

$$y'=Ay,\quad A=(a_{ij}),\quad i,\ j=1, \ldots, M, \tag{120}$$

$$y(0)=y_0,\quad y_0=(y_0^i),\quad i=1, \ldots, M. \tag{121}$$

Можно показать, что применение явных методов типа Рунге—Кутта (24) к (120), (121) приводит к соотношению

$$y_{n+1}=F_q(Ah)\,y_n, \tag{122}$$

где $F_q(Ah)$ — многочлен степени q от матрицы Ah.