ГЛАВА III

ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

В главе III рассмотрено численное дифференцирование функции, заданной на некоторой сетке. Введены квазиравномерные сетки, полезные во многих приложениях. Обсуждена некорректность задачи дифференцирования, проявляющаяся при сильном уменьшении шага, и изложены некоторые способы регуляции. Показано, как можно повышать точность и оценивать погрешность при сгущении сетки.

1. Полиномиальные формулы. Численное дифференцирование применяется, если функцию $y(x)$ трудно или невозможно продифференцировать аналитически — например, если она задана таблицей. Оно нужно также при решении дифференциальных уравнений при помощи разностных методов.

При численном дифференцировании функцию $y(x)$ аппроксимируют легко вычисляемой функцией $\varphi(x; \boldsymbol{a})$ и приближенно полагают $y'(x) = \varphi'(x; \boldsymbol{a})$. При этом можно использовать различные способы аппроксимации, изложенные в главе II. Сейчас мы рассмотрим простейший случай — аппроксимацию интерполяционным многочленом Ньютона (2.8). Вводя обозначение $\xi_i = x - x_i$, запишем этот многочлен и продифференцируем его почленно:

$$\varphi(x) = y(x_0) + \xi_0 y(x_0, x_1) + \xi_0\xi_1 y(x_0, x_1, x_2) + \\ + \xi_0\xi_1\xi_2 y(x_0, x_1, x_2, x_3) + \dots$$

$$\varphi'(x) = y(x_0, x_1) + (\xi_0 + \xi_1)\, y(x_0, x_1, x_2) + \\ + (\xi_0\xi_1 + \xi_0\xi_2 + \xi_1\xi_2)\, y(x_0, x_1, x_2, x_3) + \dots,$$

$$\varphi''(x) = 2y(x_0, x_1, x_2) + 2(\xi_0 + \xi_1 + \xi_2)\, y(x_0, x_1, x_2, x_3) + \dots$$

Общая формула имеет следующий вид:

$$\varphi^{(k)}(x) = k!\left[y(x_0, x_1, \dots, x_k) + \left(\sum_{i=0}^{k} \xi_i\right) y(x_0, x_1, \dots, x_{k+1}) + \right. \\ + \left(\sum_{i>j\geqslant 0}^{i=k+1} \xi_i\xi_j\right) y(x_0, x_1, \dots, x_{k+2}) + \\ \left. + \left(\sum_{i>j>l\geqslant 0}^{i=k+2} \xi_i\xi_j\xi_l\right) y(x_0, x_1, \dots, x_{k+3}) + \dots\right]. \quad (1)$$

Обрывая ряд на некотором числе членов, получим приближенное выражение для соответствующей производной. Наиболее простые выражения получим, оставляя в формуле (1) только первый член:

$$y'(x) \approx y(x_0, x_1) = [y(x_0) - y(x_1)] / (x_0 - x_1),$$
$$\frac{1}{2} y''(x) \approx y(x_0, x_1, x_2) = \frac{1}{x_0 - x_2} \left(\frac{y_0 - y_1}{x_0 - x_1} - \frac{y_1 - y_2}{x_1 - x_2} \right), \qquad (2)$$
$$\frac{1}{k!} y^{(k)}(x) \approx y(x_0, x_1, \ldots, x_k) = \sum_{p=0}^{k} y_p \prod_{\substack{i=0 \\ i \neq p}}^{k} (x_p - x_i)^{-1}.$$

При написании последней формулы использованы результаты задачи 1 к главе II. Все формулы (1) — (2) рассчитаны на произвольную неравномерную сетку.

Исследование точности полученных выражений при численных расчетах удобно делать при помощи апостериорной оценки, по скорости убывания членов ряда (1). Если шаг сетки достаточно мал, то погрешность близка к первому отброшенному члену. Пусть мы используем узлы x_i, $0 \leqslant i \leqslant n$. Тогда первый отброшенный член содержит разделенную разность $y(x_0, x_1, \ldots, x_{n+1})$, которая согласно (2) примерно равна $y^{(n+1)}(x) / (n+1)!$. Перед ней стоит сумма произведений различных множителей ξ_i; каждое произведение содержит $n+1-k$ множителей, а вся сумма состоит из C_{n+1}^k слагаемых. Отсюда следует оценка погрешности формулы (1) с $n+1$ узлами:

$$R_n^{(k)} \lesssim \frac{M_{n+1}}{(n+1-k)!} \max_i |\xi_i|^{n+1-k}, \qquad M_{n+1} = \max |y^{(n+1)}|. \qquad (3)$$

В частности, если сетка равномерная, то $\max |\xi_i| < nh$, откуда

$$R_n^{(k)} < M_{n+1} \left(\frac{en}{n+1-k} h \right)^{n+1-k} = O(h^{n+1-k}). \qquad (4)$$

Эти оценки можно несколько улучшить за счет более детального рассмотрения множителей ξ_i. Заметим, что строгое априорное исследование погрешности формулы (1), аналогичное выводу остаточного члена многочлена Ньютона в форме Коши (2.10), для произвольного расположения узлов приводит к той же оценке (3).

Таким образом, *порядок точности формулы* (1) *по отношению к шагу сетки равен числу оставленных в ней членов*, или, что то же самое, он равен числу узлов интерполяции минус порядок производной. Поэтому минимальное число узлов, необходимое для вычисления k-й производной, равно $k+1$; оно приводит к формулам (2) и обеспечивает первый порядок точности. Эти выводы соответствуют общему принципу: при почленном дифференцировании ряда скорость его сходимости уменьшается.

В главе II рекомендовалось использовать в формулах интерполяции не более 4—6 узлов. Если еще учесть ухудшение сходимости ряда при дифференцировании, то можно сделать вывод: даже если функция задана хорошо составленной таблицей на довольно подробной сетке, то практически численным дифференцированием можно хорошо определить первую и вторую производные, а третью и четвертую — лишь удовлетворительно. Более высокие производные редко удается вычислить с приемлемой точностью.

Замечание 1. Кубическая сплайновая интерполяция (2.20) обладает тем свойством, что первая и вторая производные интерполяционного многочлена всюду непрерывны. Обычно дифференцирование кубического сплайна позволяет определить эти производные с хорошей точностью. Если надо вычислить более высокие производные, то целесообразно строить сплайны высоких порядков. Из-за большой трудоемкости этот способ редко используется; теоретически он мало исследован.

Замечание 2. Если табулирована не только функция, но и ее производные, то следует составлять и дифференцировать интерполяционный многочлен Эрмита. Производные при этом вычисляются намного точнее, чем при дифференцировании интерполяционного многочлена Ньютона с тем же числом свободных параметров по формулам (1).

2. Простейшие формулы. Чаще всего используются равномерные сетки, на которых вид формул (1) заметно упрощается, а точность нередко повышается.

Рассмотрим сначала причину повышения точности. Остаточный член общей формулы (1) есть многочлен $\sum\prod(x-x_i)$ степени $n+1-k$ относительно x. Если x равен корню этого многочлена, то главный остаточный член обращается в нуль, т. е. в этой точке формула имеет порядок точности на единицу больше, чем согласно оценке (4). Эти *точки повышенной точности* будем обозначать $x_k^{(p)}$, где k — порядок производной, а $p=n+1-k$ — число оставленных в формуле (1) членов. Очевидно, p-членная формула имеет p точек повышенной точности.

У одночленной формулы (2) для k-й производной точка повышенной точности на произвольной сетке определяется условием $\sum\xi_i=\sum(x-x_i)=0$, что дает

$$x_k^{(1)}=(x_0+x_1+\ldots+x_k)/(k+1); \tag{5}$$

в этой точке одночленная формула имеет погрешность $O(h^2)$ вместо обычной $O(h)$. Для двучленной формулы задача нахождения точек повышенной точности приводит к квадратному уравнению, корни которого действительны, но формула для их нахождения громоздка (см. задачу 2). Если $p>2$, то найти точки

повышенной точности очень сложно, за исключением одного частного случая, который мы сейчас рассмотрим.

Пусть p нечетно, а узлы в формуле (1) *выбраны так, что они расположены симметрично относительно точки x; тогда x является одной из точек повышенной точности $x_k^{(p)}$.*

Доказательство. В самом деле, при этом величины $\xi_i = x - x_i$ имеют попарно равные абсолютные величины, но противоположные знаки. В остаточном члене множитель $w = \sum \prod \xi_i$ имеет нечетную степень, и при одновременном изменении знаков всех ξ_i он должен изменить знак. Но поскольку одновременное изменение знаков ξ_i сводится при таком расположении узлов лишь к перемене их нумерации, то величина w должна сохраниться, что возможно только при $w = 0$. Утверждение доказано.

Замечание 1. Доказательство справедливо для неравномерной сетки.

Замечание 2. Число узлов предполагалось произвольным; очевидно, симметричное расположение узлов относительно точки $x_k^{(p)}$ означает, что при нечетном числе узлов точка $x_k^{(p)}$ совпадает с центральным узлом, а при четном — лежит между средними узлами.

Замечание 3. Повышение точности достигается не только в самих точках повышенной точности, но и в достаточно малой их окрестности, где изменение производной не превышает погрешности формулы; для точки $x_k^{(1)}$ это окрестность размером $O(h^2)$, для $x_k^{(2)} - O(h^3)$ и т. д.

На произвольной сетке условие симметрии реализуется только в исключительных случаях. Но если сетка равномерна, то каждый ее узел симметрично окружен соседними узлами. Это позволяет составить несложные формулы хорошей точности для вычисления производных в узлах сетки.

Например, возьмем три соседних узла x_0, x_1, x_2 и вычислим первую и вторую производные в среднем узле. Выражая в одночленных формулах (2) разделенные разности через узловые значения функции, легко получим

$$y'(x_1) = (y_2 - y_0)/2h + O(h^2), \qquad h = x_{i+1} - x_i = \text{const}, \tag{6}$$

$$y''(x_1) = (y_2 - 2y_1 + y_0)/h^2 + O(h^2). \tag{7}$$

Формулу (6) часто записывают в несколько ином виде, удобном для определения производной в средней точке интервала сетки:

$$\begin{aligned} y'_{i+1/2} &\equiv y'(x_{i+1/2}) = (y_{i+1} - y_i)/h + O(h^2), \\ x_{i+1/2} &= x_i + {}^1/_2 h. \end{aligned} \tag{8}$$

Аналогично можно вывести формулы более высокого порядка точности или для более высоких производных. Например, трех-

членная формула (1) для первой производной в середине интервала по четырем соседним узлам дает

$$y'_{5/2} = (-y_3 + 27y_2 - 27y_1 + y_0)/(24h) + O(h^4), \qquad (9)$$

а для второй производной в центральном узле по пяти узлам

$$y''_2 = (-y_4 + 16y_3 - 30y_2 + 16y_1 - y_0)/(12h^2) + O(h^4). \qquad (10)$$

Все формулы (6) — (10) имеют четный порядок точности. Заметим, что все эти формулы написаны для случая равномерной сетки; применение их на произвольной неравномерной сетке для первой производной приводит к низкой точности $O(h)$, а для второй производной — к грубой ошибке.

На равномерной сетке для априорной оценки точности формул часто применяют способ разложения по формуле Тейлора — Маклорена. Предположим, например, что функция $y(x)$ имеет непрерывную четвертую производную, и выразим значения функции в узлах $x_{i\pm1}$ через значения функции и ее производных в центре симметрии узлов (в данном случае этим центром является узел x_i):

$$y(x_{i\pm1}) = y(x_i \pm h) = y_i \pm hy'_i + \frac{1}{2}h^2y''_i \pm \frac{1}{6}h^3y'''_i + \\ + \frac{1}{24}h^4y^{IV}(\eta_\pm), \qquad y(x_i) = y_i, \quad (11)$$

где η_+ есть некоторая точка интервала (x_i, x_{i+1}), а η_- есть некоторая точка интервала (x_{i-1}, x_i). Подставляя эти разложения во вторую разность, стоящую в правой части формулы (7) для второй производной, получим

$$\frac{1}{h^2}(y_{i+1} - 2y_i + y_{i-1}) = y''_i + \frac{h^2}{24}[y^{IV}(\eta_+) + y^{IV}(\eta_-)] = y''_i + O(h^2). \quad (12)$$

Это подтверждает ранее сделанную оценку и уточняет величину остаточного члена, который оказался равным $h^2y^{IV}(\eta)/12$. Такой способ получения остаточного члена проще, чем непосредственное вычисление по формуле (1). Особенно часто он применяется при исследовании аппроксимации разностных схем (см. главу IX).

3. Метод Рунге — Ромберга. При вычислении одной и той же величины формулы с бо́льшим числом узлов дают более высокий порядок точности, но они более громоздки. Для оценки их точности надо привлекать дополнительный узел, что требует еще более сложных вычислений. Рассмотрим более простой способ получения высокого порядка точности.

Из формулы (12) видно, что погрешность простейшей формулы (7) для четырежды дифференцируемой функции имеет вид $R = h^2\psi(\eta)$, где η — некоторая точка вблизи узла x_i. Если $y^{IV}(x)$ липшиц-непрерывна, то оценку нетрудно уточнить: $R = h^2\psi(x_i) +$

$+O(h^3)$. Пусть в общем случае имеется некоторая приближенная формула $\zeta(x, h)$ для вычисления величины $z(x)$ по значениям на равномерной сетке с шагом h, а остаточный член этой формулы имеет следующую структуру:

$$z(x) - \zeta(x, h) = \psi(x) h^p + O(h^{p+1}). \quad (13)$$

Произведем теперь расчет по той же приближенной формуле для той же точки x, но используя равномерную сетку с другим шагом rh. Тогда получим значение $\zeta(x, rh)$, связанное с точным значением соотношением

$$z(x) - \zeta(x, rh) = \psi(x)(rh)^p + O((rh)^{p+1}). \quad (14)$$

Заметим, что $O((rh)^{p+1}) \approx O(h^{p+1})$. Имея два расчета на разных сетках, нетрудно оценить величину погрешности. Для этого вычтем (13) из (14) и получим *первую формулу Рунге*:

$$R \approx \psi(x) h^p = \frac{\zeta(x, h) - \zeta(x, rh)}{r^p - 1} + O(h^{p+1}). \quad (15)$$

Первое слагаемое справа есть главный член погрешности. Таким образом, расчет по второй сетке позволяет оценить погрешность расчета на первой сетке (с точностью до членов более высокого порядка).

Можно исключить найденную погрешность (15) из формулы (13) и получить результат с более высокой точностью по *второй формуле Рунге*:

$$z(x) = \zeta(x, h) + \frac{\zeta(x, h) - \zeta(x, rh)}{r^p - 1} + O(h^{p+1}). \quad (16)$$

Этот метод оценки погрешности и повышения точности результата очень прост, применим в большом числе случаев и исключительно эффективен. Рассмотрим два примера его применения к численному дифференцированию.

Таблица 7

x	$y = \lg x$
1	0,000
2	0,301
3	0,478
4	0,602
5	0,699

Пример 1. Пусть функция $y(x) = \lg x$ задана таблицей 7 и требуется вычислить $y'(3)$. Выберем для вычислений простейшую формулу (6). Полагая $h = 1$, т. е. производя вычисления по точкам $x = 2$ и $x = 4$, получим $y'(3) \approx 0{,}151$. Увеличивая шаг вдвое ($r = 2$), т. е. вычисляя производную по точкам $x = 1$ и $x = 5$, получим $y'(3) \approx 0{,}175$. Проводя вычисления по формуле Рунге (16), где согласно оценке (6) берется $p = 2$, получим уточненное значение $y'(3) \approx 0{,}143$; это всего 2% отличается от искомого значения $y'(3) \approx 0{,}145$.

Пример 2. Выведем формулу высокой точности из формулы низкой точности. Возьмем простейшую формулу для вычисления

первой производной в середине интервала (8) и запишем ее, выбирая сначала соседние узлы, а затем более удаленные:

$$y'_{3/2}(h) \approx (y_2 - y_1)/h, \quad y'_{3/2}(3h) \approx (y_3 - y_0)/3h.$$

Порядок точности формулы $p=2$, а коэффициент увеличения шага $r=3$, поэтому уточнение методом Рунге дает формулу (9):

$$y'_{3/2} \approx y'_{3/2}(h) + \frac{1}{8}[y'_{3/2}(h) - y'_{3/2}(3h)] = \frac{1}{24h}(y_0 - 27y_1 + 27y_2 - y_3).$$

Отсюда видно, что для получения высокого порядка точности не обязательно производить вычисления непосредственно по формулам высокого порядка точности; можно произвести вычисления по простым формулам низкой точности на разных сетках и затем уточнить результат методом Рунге. Последний способ предпочтительней еще потому, что величина поправки (15) дает апостериорную оценку точности.

Метод Рунге обобщается на случай произвольного числа сеток. Пусть функция $y(x)$ имеет достаточно высокие непрерывные производные. Тогда в разложениях Тейлора типа (11) можно удерживать большое число членов и подстановка их в формулы типа (6)—(10) приводит к представлению остаточного члена в виде ряда

$$z(x) - \zeta(x, h) = \sum_{m \geqslant p} \psi_m(x) h^m. \tag{17}$$

Пусть расчет проведен на q различных сетках с шагами h_j, $1 \leqslant j \leqslant q$. Тогда из остаточного члена можно исключить первые $q-1$ слагаемых. Для этого перепишем соотношение (17), оставляя первые $q-1$ члены погрешности:

$$z(x) - \sum_{m=p}^{p+q-2} \psi_m(x) h_j^m = \zeta(x, h_j) + O(h^{p+q-1}), \quad 1 \leqslant j \leqslant q.$$

Это система линейных уравнений относительно величин $z(x)$, $\psi_m(x)$. Решая ее по правилу Крамера, получим уточненное значение по *формуле Ромберга*

$$z(x) = \begin{vmatrix} \zeta(x, h_1) & h_1^p & h_1^{p+1} & \dots & h_1^{p+q-2} \\ \zeta(x, h_2) & h_2^p & h_2^{p+1} & \dots & h_2^{p+q-2} \\ \dots & \dots & \dots & \dots & \dots \\ \zeta(x, h_q) & h_q^p & h_q^{p+1} & \dots & h_q^{p+q-2} \end{vmatrix} \times \begin{vmatrix} 1 & h_1^p & h_1^{p+1} & \dots & h_1^{p+q-2} \\ 1 & h_2^p & h_2^{p+1} & \dots & h_2^{p+q-2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & h_q^p & h_q^{p+1} & \dots & h_q^{p+q-2} \end{vmatrix}^{-1} + O(h^{p+q-1}). \tag{18}$$

Эта формула приводит к повышению порядка точности результата

на $q-1$ по сравнению с исходной формулой $\zeta(x, h)$, т. е. каждая лишняя сетка позволяет повысить порядок точности на единицу.

Формула Ромберга удобна тем, что ее можно применять при любом числе равномерных сеток и любом соотношении их шагов. Ее недостатками являются сравнительная громоздкость и отсутствие в промежуточных выкладках апостериорных оценок точности. Если сетки выбраны так, что сгущение сеток происходит всегда в одно и то же число раз (т. е. $h_j = rh_{j-1} = \ldots = r^{j-1}h_1$), то вместо формулы Ромберга удобнее рекуррентно применять метод Рунге.

Для этого берут последовательные пары сеток (h_1, h_2), (h_2, h_3), (h_3, h_4) и т. д. По каждой паре производят уточнение методом Рунге, исключая тем самым главный член погрешности $\psi_p(x)\,h^p$. Поэтому в уточненных величинах главный член погрешности будет иметь вид $\tilde{\psi}_{p+1}(x)\,h^{p+1}$, где шаг можно условно принять для первой пары сеток за h_1, для второй — за h_2 и т. д. (это верно, только если h_j/h_{j-1} одинаково для всех пар сеток). Уточненные значения таким же образом группируют в пары и исключают ошибку следующего порядка $O(h^{p+1})$. Всего можно произвести $q-1$ уточнение, на единицу меньше числа сеток. При каждом уточнении вычисляется погрешность (15), дающая апостериорную оценку точности на данном этапе вычислений. Пример такого вычисления будет дан в главе IV.

Замечание 1. Если исходная формула для вычисления $\zeta(x, h)$ имеет симметричный вид, то на равномерной сетке обычно все нечетные члены ряда (17) обращаются в нуль. При этом пользоваться общей формулой (18) можно, но невыгодно, ибо она не учитывает дополнительной информации о нулевых коэффициентах. Следует оставить в сумме (17) только степени h^p, h^{p+2}, h^{p+4}, ... и соответственно изменить формулу Ромберга. Аналогично изменяется рекуррентная процедура Рунге: при очередном исключении ошибки порядок точности повышается не на 1, а на 2. Примером может служить данный выше вывод формулы (9) из формулы (8), когда после первого уточнения погрешность уменьшилась с $O(h^2)$ сразу до $O(h^4)$.

Замечание 2. Допустимое число членов суммы (17) связано с количеством существующих у функции непрерывных производных. Поэтому для недостаточно гладких функций бессмысленно брать большое число сеток. Практически даже для «хороших» функций используют не более 3 — 5 сеток; обычно отношение r их шагов стараются выбрать равным 2.

Замечание 3. Метод Рунге — Ромберга можно применять только в том случае, если ошибка представима в виде (17), где коэффициенты $\psi_m(x)$ одинаковы для *всех* сеток. Строго говоря, при численном дифференцировании эти коэффициенты зависят от положения узлов сетки. Но если выбранные конфигурации узлов

на всех сетках подобны относительно точки x (рис. 14, *а*), то зависимость от узлов одинакова для всех сеток и сводится к величине шага. Тогда метод Рунге — Ромберга применим. Если же правило подобия нарушено (рис. 14, *б*, *в*), то метод применять нельзя.

Поэтому при численном дифференцировании метод Рунге—Ромберга удается применять только для нахождения производных в узлах или серединах интервалов равномерных (или описанных далее квазиравномерных) сеток. Но эти случаи являются достаточно важными в практических приложениях. Особенно широко применяется описанный метод при численном интегрировании и разностных методах решения задач для дифференциальных и интегральных уравнений.

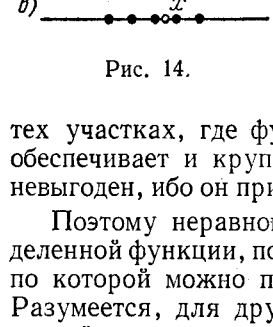

Рис. 14.

4. Квазиравномерные сетки. При тех значениях аргумента, где функция резко меняется, шаг таблицы должен быть малым, иначе точность вычисления по этой таблице будет плохой. А на тех участках, где функция меняется медленно, хорошую точность обеспечивает и крупный шаг таблицы; мелкий шаг при этом даже невыгоден, ибо он приводит к сильному увеличению объема таблицы.

Поэтому неравномерная сетка, удачно подобранная для определенной функции, позволяет построить таблицу небольшого объема, по которой можно производить вычисления с хорошей точностью. Разумеется, для других функций эта сетка может быть малопригодной.

Каждая конкретная сетка либо равномерна (т. е. ее шаг $h_i = x_{i+1} - x_i$ постоянен), либо неравномерна. Но нам нередко приходится сгущать сетку, т. е. рассматривать на $[a, b]$ последовательность сеток $x_i^{(N)}$, $0 \leqslant i \leqslant N$, с возрастающим числом интервалов N. Разумеется, если таблица уже задана, то сетку сгущать невозможно, но можно ее разреживать, выбрасывая из таблицы половину, две трети и т. д. точек. Это также является некоторым способом построения последовательности сеток. Сгущение сеток широко применяется при численном решении дифференциальных и интегральных уравнений.

Среди последовательностей сеток важное место занимают *квазиравномерные сетки*. Будем называть сетки квазиравномерными, если существует дважды непрерывно дифференцируемая функция $x = \xi(t)$, преобразующая отрезок $0 \leqslant t \leqslant 1$ в отрезок $a \leqslant x \leqslant b$ так, что каждой сетке $x_i^{(N)}$ соответствует равномерная сетка $t_i^{(N)} = i/N$, причем на этом отрезке $\xi'(t) \geqslant \varepsilon > 0$, а $\xi''(t)$ ограничена.

Если эти условия выполнены, то шаг сетки $h_i \approx \xi'(t_i)/N$, а разность двух соседних шагов есть $h_i - h_{i-1} \approx \xi''(t_i)/N^2$. Значит,

при большом числе узлов разность соседних шагов $\Delta h \sim h^2$, т. е. много меньше длины шага, и соседние интервалы почти равны. Поэтому такие сетки и называют квазиравномерными или почти равномерными. Однако отношение длин далеких друг от друга интервалов $h_i/h_j \approx \xi'(t_i)/\xi'(t_j)$ может быть большим.

Для того чтобы сгустить квазиравномерную сетку x_i, надо сгустить равномерную сетку t_i (увеличить N) и по ней вычислить новую сетку. Середину интервала $x_{i+1/2}$ квазиравномерной сетки надо вычислять при помощи того же преобразования, полагая

$$x_{i+1/2} = \xi\left(\frac{i+1/2}{N}\right);$$

брать $x_{i+1/2}$ равной полусумме соседних узлов x_i, x_{i+1} нельзя.

Рассмотрим некоторые примеры.

а) Если надо детально передать поведение функции вблизи одного из концов отрезка $[a, b]$, то удобно преобразование

$$x = a + (b-a)(e^{\alpha t} - 1)/(e^{\alpha} - 1). \tag{19}$$

Значение $\alpha > 0$ соответствует малому шагу сетки у левого конца отрезка, значение $\alpha < 0$ — у правого. Шаги этой сетки составляют геометрическую прогрессию со знаменателем $q = h_{i+1}/h_i = e^{\alpha/N}$. Отношение первого и последнего шага сетки примерно равно e^{α}; при большом α оно может быть очень большим. Такая сетка полезна, например, в задачах атомной физики, где волновые функции наиболее быстро меняются вблизи ядра.

б) На полупрямой $[a, \infty)$ тоже можно построить квазиравномерную сетку; например, таким преобразованием:

$$x = a + \alpha \operatorname{tg}(\pi t/2), \quad 0 \leqslant t < 1. \tag{20}$$

Параметр α управляет сеткой; чем он меньше, тем гуще узлы сетки при $x \to a$ и реже при $x \to \infty$. Последний интервал этой сетки (x_{N-1}, x_N) бесконечно велик, ибо точка x_N — бесконечно удаленная (отсюда ясно, что середину интервала квазиравномерной сетки надо находить при помощи основного преобразования!). Эта сетка полезна при вычислении интегралов с бесконечным верхним пределом.

в) Преобразование $x = \alpha \operatorname{tg}(\pi t/2)$ при $-1 < t < 1$ позволяет построить квазиравномерную сетку на бесконечной прямой. Первый и последний интервалы этой сетки бесконечны.

г) Преобразование $x = t^2$, $0 \leqslant t \leqslant 1$, определяет не квазиравномерную сетку. Здесь не выполнено условие строгой положительности $\xi'(t)$. По этому преобразованию строится такая сетка:

$$x_0 = 0, \quad x_1 = 1/N^2, \quad x_2 = 4/N^2, \ldots; \quad h_0 = 1/N^2, \quad h_1 = 3/N^2, \ldots$$

В результате разность двух соседних шагов — первого и второго — вдвое больше одного из них при любом N. Значит, вблизи точки $x = 0$ сетка не стремится к равномерной при $N \to \infty$.

Если сетка квазиравномерна, то производные на ней вычисляются либо проще, либо точнее, чем на произвольной неравномерной сетке. Например, если на такой сетке взять подряд три узла x_0, x_1, x_2, то $x_1 = (x_0 + x_2)/2 + (h_0 - h_1)/2 = (x_0 + x_2)/2 + O(h^2)$

и аналогично $x_1 = (x_0 + x_1 + x_2)/3 + O(h^2)$. Это означает, что узел x_1 расположен вблизи точки повышенной точности для этих узлов, в окрестности размером $O(h^2)$. Из сделанного в п. 2 замечания следует, что одночленные формулы (2), рассчитанные на произвольную сетку:

$$y_1' \approx y(x_0, x_2), \quad y_1'' \approx 2y(x_0, x_1, x_2),$$

в узлах квазиравномерной сетки обеспечивают точность $O(h^2)$. Пользоваться в этом случае формулами типа (7), рассчитанными на равномерную сетку, не следует — на квазиравномерной сетке их точность будет хуже.

На квазиравномерных сетках справедливо разложение остаточного члена в ряд (17), если порождающее сетки преобразование $x = \xi(t)$ достаточное число раз непрерывно дифференцируемо. В этом случае для повышения точности расчетов можно употреблять метод Рунге — Ромберга, подставляя в формулы (16) — (18) вместо h величину $1/N$. Для квазиравномерных сеток этот метод особенно выгоден, ибо для них прямые формулы высокого порядка точности очень громоздки.

Только в одном пункте квазиравномерные сетки уступают равномерным. На них ряд для остаточного члена (17) даже в случае симметричной формулы содержит обычно все степени $1/N$, поэтому каждая лишняя сетка позволяет повысить порядок точности лишь на единицу, а не на двойку.

Квазиравномерные сетки часто используют при решении сложных задач математической физики, когда необходимо при малом числе узлов детально передать особенности решения.

5. Быстропеременные функции. Если функция (точнее, ее разделенные разности) значительно меняется на протяжении нескольких интервалов сетки, то интерполяция обобщенным многочленом обычно недостаточно точна для дифференцирования этой функции. Для таких функций особенно полезна квазилинейная интерполяция, производимая при помощи выравнивающих переменных.

Если $\xi(x)$, $\eta(y)$ — выравнивающие переменные, то для искомой производной справедливо соотношение

$$y_x' = \xi_x' \eta_\xi' / \eta_y'. \tag{21}$$

Выравнивающие преобразования подбирают несложными, чтобы их производные ξ_x', η_y' находились точно. Остается только численно найти η_ξ' способами, изложенными в предыдущих пунктах.

Например, пусть имеются таблицы энергии $E(T, \rho)$ многократно ионизованной плазмы тяжелых атомов. Рассмотрим нахождение теплоемкости $c_v = (\partial E/\partial T)_v$; она отличается от теплоемкости идеального газа, поскольку в нее входит энергия, идущая на отрыв от ионного остова новых электронов при повышении температуры. Ранее упоминалось, что зависимость $E(T)$ напоминает степенную со слабо переменным показателем и выравнивающим является пре-

образование $\eta = \ln E$, $\xi = \ln T$. Легко видеть, что формула (21) принимает вид

$$c_v = (E/T)\,\eta'_\xi = (E/T)\,[\partial(\ln E)/\partial(\ln T)];$$

последнюю производную находят численным дифференцированием (см. задачу 7).

Если в исходных переменных сетка была равномерной или квазиравномерной, то обычно она квазиравномерна и в выравнивающих переменных, ибо выравнивающее преобразование на ограниченном отрезке почти всегда обладает требуемыми свойствами производных. В этом случае результат можно уточнять методом Рунге — Ромберга.

Формула двукратного дифференцирования при помощи выравнивающих переменных достаточно сложна:

$$y''_{xx} = [\xi''_{xx}\eta'_x + (\xi'_x)^2\,\eta''_{\xi\xi} - \eta''_{yy}\,(\xi'_x\eta'_\xi/\eta'_y)^2]/\eta'_y, \qquad (22)$$

и ее применение не всегда обеспечивает хорошую точность. Но для быстропеременной функции двукратное дифференцирование интерполяционного многочлена Лагранжа еще более ненадежно. Поэтому вторую и более высокие производные быстропеременной функции трудно найти численно.

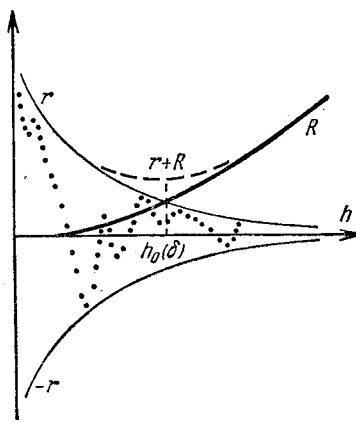

Рис. 15.

6. Регуляризация дифференцирования. При численном дифференцировании приходится вычитать друг из друга близкие значения функции. Это приводит к уничтожению первых значащих цифр, т. е. к потере части достоверных знаков числа. Если значения функции известны с малой точностью, то встает естественный вопрос — останется ли в ответе хоть один достоверный знак?

Для ответа на этот вопрос исследуем ошибки при численном дифференцировании. При интерполировании обобщенным многочленом производная k-го порядка определяется согласно (2)—(3) формулой типа

$$y^{(k)}(x) = h^{-k}\sum_q C_q(x)\,y(x_q) + R_k(x). \qquad C_q(x) = O(1). \qquad (23)$$

Если формула имеет порядок точности p, то, значит, ее остаточный член равен $R_k(x) \approx C(x)\,h^p$. Этот остаточный член определяет погрешность метода, и он неограниченно убывает при $h \to 0$. Его зависимость от шага изображена на рис. 15 жирной линией.

Но есть еще неустранимая погрешность, связанная с погрешностью функции $\delta y(x)$. Поскольку точный вид этой погрешности неизвестен, можно оценить только мажоранту неустранимой погрешности $r_k = \delta \cdot h^{-k} \sum_q |C_q|$; она неограниченно возрастает при $h \to 0$ (тонкая линия на рис. 15). Фактически же неустранимая погрешность будет нерегулярно зависеть от величины шага, беспорядочно осциллируя в границах, определяемых мажорантой (точки на рис. 15).

Пока шаг достаточно велик, при его убывании неустранимая погрешность мала по сравнению с погрешностью метода; поэтому полная погрешность убывает. При дальнейшем уменьшении шага неустранимая погрешность становится заметной, что проявляется в не вполне регулярной зависимости результатов вычислений от величины шага. Наконец, при достаточно малом шаге неустранимая погрешность становится преобладающей, и при дальнейшем уменьшении шага результат вычислений становится все менее достоверным.

Полная погрешность мажорируется суммой $R_k + r_k$ (штриховая кривая на рисунке). Оптимальным будет шаг, соответствующий минимуму этой кривой. Нетрудно подсчитать, что

$$h_0(\delta) = \left(k\delta \sum_q |C_q|/pC\right)^{1/(p+k)} = O\left(\delta^{1/(p+k)}\right),$$
$$\min(R_k + r_k) \doteq C h_0^p \left(1 + \frac{p}{k}\right) = O\left(\delta^{p/(p+k)}\right). \tag{24}$$

Меньший шаг невыгоден, а меньшая погрешность, вообще говоря, недостижима (хотя отдельные вычисления случайно окажутся более точными, но мы этого не сможем узнать). Эта минимальная ошибка тем меньше, чем меньше погрешность входных данных и порядок вычисляемой производной и чем выше порядок точности формулы.

Очевидно, *при* $\delta y(x) \to 0$ *можно получить сколь угодно высокую точность результата, если шаг стремится к нулю, будучи всегда не менее* $h_0(\delta)$. Но если допустить $h < h_0(\delta)$, то результат предельного перехода может быть неправильным.

Эта тонкость связана с некорректностью задачи дифференцирования. Рассмотрим погрешность входных данных вида $\delta y(x) = m^{-1} \sin m^2 x$. Она приводит к погрешности первой производной $\delta y'(x) = m \cos m^2 x$. При $m \to \infty$ погрешность функции в $\|\cdot\|_C$ неограниченно убывает, а погрешность производной в той же норме неограниченно растет. Значит, нет непрерывной зависимости производной от функции, т. е. *дифференцирование некорректно.* Особенно сильно это сказывается при нахождении производных высокого порядка.

Изложенный выше способ определения оптимального шага и запрещение вести расчет шагом меньше оптимального есть некоторый способ регуляризации дифференцирования, так называемая *регуляризация по шагу*. Этот способ в простейшей форме давно применялся физиками, которые при однократном численном дифференцировании всегда выбирали такой шаг, чтобы $|y(x+h) - y(x)| \geqslant \delta$.

К этой задаче применим и метод регуляризации А. Н. Тихонова; он будет изложен в главе XIV, § 2.

Физики издавна употребляют (без строгого обоснования) еще один способ регуляризации — дифференцирование предварительно сглаженной кривой, причем сглаживание обычно выполняют методом наименьших квадратов. Роль параметра регуляризации здесь играет отношение числа свободных параметров n аппроксимирующей кривой к числу узлов сетки N; для хорошего сглаживания должно выполняться условие $n \ll N$.

Рассмотрим, как это делается в простейшем случае. Выберем около искомой точки не очень большой интервал изменения аргумента, чтобы двучленная аппроксимация $y(x) \approx a + bx$ обеспечивала удовлетворительную точность. Но этот интервал должен содержать довольно много узлов сетки, т. е. быть не слишком малым.

Система уравнений (2.43) для определения коэффициентов среднеквадратичной аппроксимации принимает следующий вид:

$$\begin{aligned} a\sum \rho_i + b\sum \rho_i x_i &= \sum \rho_i y_i, \\ a\sum \rho_i x_i + b\sum \rho_i x_i^2 &= \sum \rho_i x_i y_i, \end{aligned} \tag{25}$$

где сумма берется по узлам сетки x_i, лежащим в этом интервале. Введем на этом интервале средние значения

$$\bar{x} = \Big(\sum \rho_i x_i\Big)\Big/\Big(\sum \rho_i\Big), \quad \bar{y} = \Big(\sum \rho_i y_i\Big)\Big/\Big(\sum \rho_i\Big). \tag{26}$$

Тогда первое уравнение (25) можно записать в виде $a + b\bar{x} = \bar{y}$ (см. задачу 8 к главе II). Умножая его на $\bar{x}\sum \rho_i$ и вычитая из второго уравнения (25), получим

$$y'(x) \approx b = \Big[\sum \rho_i (x_i y_i - \bar{x}\bar{y})\Big]\Big/\Big[\sum \rho_i (x_i^2 - \bar{x}^2)\Big]. \tag{27}$$

Пользуясь определением средних (26), произведем несложное преобразование знаменателя в (27):

$$\begin{aligned} \sum \rho_i (x_i^2 - \bar{x}^2) &= \sum \rho_i (x_i^2 + \bar{x}^2) - 2\bar{x}^2 \sum \rho_i = \\ &= \sum \rho_i (x_i^2 + \bar{x}^2) - 2\bar{x} \sum \rho_i x_i = \sum \rho_i (x_i - \bar{x})^2 \end{aligned}$$

и аналогично преобразуем числитель. Тогда выражение (27)

приводится к виду, напоминающему коэффициент парной корреляции величин x и y:

$$y'(x) \approx b = \left[\sum \rho_i (x_i - \bar{x})(y_i - \bar{y})\right] / \left[\sum \rho_i (x_i - \bar{x})^2\right]. \qquad (28)$$

Последняя формула несколько выгодней для численных расчетов, чем предыдущая, ибо ошибки округления в ней меньше.

Двучленная среднеквадратичная аппроксимация дает удовлетворительные результаты, только если $\delta/\sqrt{N} \ll \Delta$, где δ — погрешности отдельных значений функции, N — число точек в выбранном участке, а Δ — нелинейная часть приращения функции на данном участке. Если это соотношение нарушено, то надо строить сглаживающие аппроксимации с 3—4 членами и дифференцировать их.

В заключение отметим, что выравнивающие переменные позволяют вести расчет крупным шагом или с малым числом свободных параметров. Поэтому предварительное приведение к выравнивающим переменным существенно ослабляет влияние погрешности начальных данных и позволяет теми же способами регуляризации добиться большей точности.

ЗАДАЧИ

1. Составить формулу вычисления $y'(x)$ на основании интерполяционного многочлена Эрмита (2.18) и сравнить ее с простейшей формулой $y'(x) \approx y'(x_0) + (x - x_0)\, y'(x_0, x_1)$. Найти погрешности обеих формул. Какая из формул точнее и почему?

2. Показать, что у двучленной формулы (1) есть две точки повышенной точности, определяемые соотношением

$$x_k^{(2)} = \left[\sqrt{k+1} \sum_{i=0}^{k+1} x_i \pm \sqrt{\sum_{i>j\geqslant 0}^{i=k+1} (x_i - x_j)^2}\right] \Big/ \left[(k+2)\sqrt{k+1}\right],$$

в которых достигается третий порядок точности.

3. Аналогично (6)—(10) получить формулы для вычисления y^{III} и y^{IV} в среднем узле по пяти узлам равномерной сетки.

Таблица 8

T, эв	E, $\frac{\text{кдж}}{\text{г}}$
2,04	2250
1,15	720
0,646	303
0,363	176
0,204	64,8
0,115	24,8

4. Способом разложения по формуле Тейлора найти остаточные члены формул (8)—(10).

5. Получить для второй производной формулу высокой точности (10) из простейшей формулы (7) методом Рунге.

6. Строго обосновать рекуррентное применение метода Рунге.

7. В таблице 8 приведены данные по энергии плазмы алюминия при плотности 10^{19} атом/см^3. Составить таблицу теплоемкости c_v и оценить ее точность, полагая, что: а) значения энергии вычислены точно, б) значения энергии имеют погрешность $\pm 10\%$.

8. Используя среднеквадратичную аппроксимацию функции параболой $y(x) \approx a + bx + cx^2$, найти выражение для ее первой и второй производной через значения функции в узлах.