УДК 518.5:519.95

ДИФФЕРЕНЦИРОВАНИЕ НА СЕТКЕ

И.А. Пахнутов

ФГБОУ ВПО «Калининградский государственный технический университет»,
Россия, 236022, г. Калининград, Советский проспект, 1
E-mail: IA-Pa2010@yandex.ru

Рассматривается методология оценок производных функций, заданных своими значениями на конечной сетке числовой прямой. Сравниваются результаты локальных оценок и оценок на всей сетке с вычислительной точки зрения. Обсуждается итерационный способ оценок одновременно всех производных. Приведены примеры, показывающее вычислительное преимущество рассматриваемого подхода перед локальными разностными формулами.

численное дифференцирование, аппроксимация, разностные формулы

К численным оценкам значений производных функции в заданной точке прибегают в том случае, когда функция слишком сложна и ее явное дифференцирование весьма трудоемко или невозможно, либо ее точные значения в принципе недоступны (см., например, [1] – [3]). Методологически здесь прослеживаются, в основном, две тенденции: выбор подходящего класса функций, которым достаточно хорошо представляются данные значения, и использование разностных формул. Класс подбираемых функций может быть достаточно широк (напр., [4], [5]) и определяется, в сущности, потребностями пользователя.

В том случае, когда информация о функции недостаточна, прибегают к решению соответствующих экстремальных задач [1]-[3], [6], которые часто сами порождают нужный класс функций. Так, в чертежной практике издавна использовалась гибкая металлическая линейка (spline) для проведения гладких кривых через заданные точки плоскости. Получается кривая y = f(x), для которой величина $\int_a^b (f''(x))^2 dx$ минимальна. Уравнение Эйлера $f^{(IV)}(x) = 0$ в этой экстремальной задаче имеет решение в классе дважды непрерывно дифференцируемых функций в виде кривых, гладко склеенных из кубических многочленов, которые по традиции называются сплайнами. Если значения $y_k = f(x_k)$, $k = 0, ..., n$, $a = x_0 < x_1 < ... < x_n = b$, известны лишь приближенно, то решается другая экстремальная задача (задача гладкой аппроксимации):

$$\int_a^b (f''(x))^2 dx + \lambda \sum_{k=0}^{n} (f(x_k) - y_k)^2 \rightarrow \min.$$

И в этом случае решение также можно построить в виде кубических сплайнов [7].

Во многих приложениях, особенно когда о функции информация недостаточна, сетка исходных для анализа данных $(x_k,\ y_k)_{k=0}^{n}$ выбирается равномерной, т. е. $x_k = x_0 + h \cdot k$, $k = 0,..., n$, для некоторого $h > 0$. Это оправдано [8] и это удобно с вычислительной точки зрения.

Во-первых, однотипные разностные формулы можно применять на любом участке сеточных данных (за исключением концов), причем порядок таких формул ограничивается лишь размером сетки и разумной целесообразностью. Некоторые популярные разностные формулы порядка k (точные для всех многочленов степени не выше k) приближенного вычисления первой и второй производных в узлах равномерной сетки с шагом h приведены в табл. 1, 2.

Таблица 1

Формулы для вычисления $y'_m h \cdot k!$ порядка k			
m	k = 2	k = 3	k = 4
0	$-3y_0 +4y_1 - y_2$	$-11y_0 +18y_1 -9y_2 +2y_3$	$2(-25y_0+48y_1-36y_2+16y_3-3y_4)$
1	$-y_0 +y_2$	$-2y_0 -3y_1 +6y_2 -y_3$	$2(-3y_0-10y_1+18y_2-6y_3+y_4)$
2	$y_0 - 4y_1 +3y_2$	$y_0 -6y_1 +3y_2 +2y_3$	$2(y_0-8y_1 +8y_3-y_4)$
3		$-2y_0 +9y_1 -18y_2 +11y_3$	$2(-y_0+6y_1-18y_2+10y_3+3y_4)$
4			$2(3y_0-16y_1+36y_2-48y_3+25y_4)$

Таблица 2

Формулы для вычисления $y''_m h^2$ порядка k			
m	k = 2	k = 3	k = 4
0	$y_0-2y_1+y_2$	$2y_0-5y_1+4y_2-y_3$	$(35y_0-104y_1+114y_2-56y_3+11y_4) / 12$
1	$y_0-2y_1+y_2$	$y_0-2y_1+y_2$	$(11y_0-20y_1+6y_2+4y_3-y_4) / 12$
2	$y_0-2y_1+y_2$	$y_1-2y_2+y_3$	$(- y_0+16y_1-30y_2+16y_3-y_4) / 12$
3		$-y_0+4y_1-5y_2+2y_3$	$(-y_0+4y_1+6y_2-20y_3+11y_4) / 12$
4			$(11y_0-56y_1+114y_2-104y_3+35y_4) / 12$

Во-вторых, с помощью распространенного в настоящее время аппарата приближения кубическими сплайнами оценки первой и второй производной сеточных значений функций можно получить из довольно простых линейных систем уравнений [3], [7]. Пусть, например, в точках $x_k = x_0 + kh$, $k = 0,..., n$, известны (с удовлетворительной погрешностью) значения y_k некоторой функции f. Пусть также m_k и M_k суть оценки первой и второй производной f в точке k, $0 \le k \le n$. Тогда с помощью кубических сплайнов эти оценки можно найти из систем уравнений:

$$\begin{cases} 4m_0 + 2m_1 = \dfrac{6}{h}(y_1 - y_0) - hA, \\ m_{k-1} + 4m_k + m_{k+1} = \dfrac{3}{h}(y_{k+1} - y_{k-1}),\ k = 1,..., n-1, \\ 2m_{n-1} + 4m_n = \dfrac{6}{h}(y_n - y_{n-1}) + hB, \end{cases} \tag{1}$$

и

$$\begin{cases} 4M_1 + M_2 = \dfrac{6}{h^2}(y_0 - 2y_1 + y_2) - A, \\ M_{k-1} + 4M_k + M_{k+1} = \dfrac{6}{h^2}(y_{k-1} - 2y_k + y_{k+1}),\ k = 2,..., n-2, \\ M_{n-2} + 4M_{n-1} = \dfrac{6}{h^2}(y_{n-2} - 2y_{n-1} + y_n) - B, \end{cases} \tag{2}$$

где A и B суть представления значений первых (соответственно вторых) производных f в точках x_0 и x_n (так называемые краевые условия), например, с помощью формул табл. 1, 2.

Если потребовать еще меньше, например, построить гладкий интерполирующий агрегат с возможно меньшими колебаниями, можно минимизировать суммарный "наклон", т.е. решать задачу $\int_a^b (f'(x))^2 dx \to \min$ при условиях $f(x_k) = y_k$, $\forall k$. Уравнение Эйлера $f''(x) = 0$ удовлетворяется кусочно-линейными функциями, но по информации (x_k, y_k, y'_k), k = 0,..., n, легко построить и более гладкое приближение [9]. В этом случае определяющая производные система уравнений ([9], (3) при $\alpha = 0.25$) аналогична предыдущим и не требует дополнительных граничных условий (заметим, тем не менее, что использование дополнительных условий на концах отрезка может существенно улучшить работу алгоритма):

$$\begin{cases} 5m_0 + m_1 = \dfrac{4}{h}(y_1 - y_0), \\ m_{k-1} + 6m_k + m_{k+1} = \dfrac{4}{h}(y_{k+1} - y_{k-1}),\ k = 1,..., n-1, \\ m_{k-1} + 5m_n = \dfrac{4}{h}(y_n - y_{n-1}). \end{cases} \tag{3}$$

Во всех рассмотренных случаях матрицы систем уравнений хорошо обусловлены, имеют доминирующую главную диагональ, системы легко решаются методом прогонки без накопления погрешности вычислений [2]. Объем вычислений при решении таких систем пропорционален количеству узлов сетки, так что вычисление производных на всей сетке не более трудоемко, чем при использовании соответствующих разностных формул.

Несколько выпадает из общего контекста метод [10], позволяющий на всей сетке оценить значения всех производных до заданного порядка (это бывает необходимо в задачах, использующих дифференцирование разных порядков, и в некоторых обратных задачах). Остановимся на нем подробнее.

Будем снова рассматривать равномерную сетку $x_k = x_0 + kh$, k = 0,..., n, и некоторый числовой ряд $y_0, y_1,..., y_n$, представляющий значения достаточно гладкой функции f на сетке. Фиксируем натуральное $m < n$ и обозначим $\sigma_{s,k} = f^{(s)}(x_k)\,\dfrac{h^s}{s!}$, k = 0,..., n, s = 1,..., m, $\sigma_{0,k} = f(x_k)$, $\forall k$. Найдем матрицу $B \in \mathbb{R}^{m\times m}$ и вектор $a \in \mathbb{R}^m$ такие, чтобы равенство

$$\sigma_{s,k+1} = \sum_{j=1}^{m} B_{s,j}\,\sigma_{j,k} + a_s(\sigma_{0,k+1} - \sigma_{0,k}),\ s = 1,..., m,\ k = 0,..., n \tag{4}$$

выполнялось для всех многочленов степени не выше m. Раскладывая обе части равенства (4) в ряд Тейлора по степеням h в окрестности точки x_k и сравнивая коэффициенты при одинаковых степенях, получим следующее представление матрицы B = B(a):

$$B_{s,j} = \binom{s}{j} - a_s,\ s, j = 1,\dots, m, \quad \text{где } \binom{s}{j} = \begin{cases} 0 \text{ при } s < j, \\ \dfrac{s!}{j!(s-j)!} \text{ при } s \ge j > 0, \end{cases}$$

- биномиальные коэффициенты (0! = 1 по определению). Можно распорядиться свободными параметрами (a_s) двояко: увеличить точность (4) либо повлиять нужным образом на свойства матрицы B(a). Введем обозначения $\sigma_k = (\sigma_{1,k}, \sigma_{2,k},\dots, \sigma_{m,k})^T$, $a = (a_1, a_2,\dots, a_m)^T$, $\beta_k = (y_{k+1} - y_k)\cdot a$. В этих обозначениях равенства (4) примут вид: $\sigma_{k+1} = B\sigma_k + \beta_k$. Отсюда следует, что погрешность θ_k, возникшая при вычислении в узле x_k, будет изменяться пропорционально степени матрицы B: $\theta_{k+p} = B^p\theta_k$, т.е. расти экспоненциально, если $\|B\| > 1$. Заметим ([11], с. 47), что если ρ – максимум модуля собственных значений матрицы B (ее спектральный радиус), то для любого $\varepsilon > 0$ найдется норма $\|\cdot\|_\varepsilon$ в $\mathbb{R}^m$ такая, что $\|B\|_\varepsilon < \rho + \varepsilon$. Покажем, что $\rho = \rho(a)$ можно сделать как угодно малым.

Пусть $\langle \cdot, \cdot \rangle$ - скалярное произведение в $\mathbb{R}^m$, $\pi_m(x)$ – множество всех многочленов степени m со старшим коэффициентом равным единице, E – единичная матрица, $P(\lambda, a) = \det(B(a) - \lambda E)$ – характеристический многочлен матрицы B = B(a).

ТЕОРЕМА 1. Отображение $P(\lambda, \cdot): \mathbb{R}^m \to \pi_m(\lambda)$ биективно.

ДОКАЗАТЕЛЬСТВО. Вычитая первый столбец определителя

$$P(\lambda, a) = \det\begin{pmatrix} 1 - a_1 - \lambda & \binom{2}{1} - a_1 & \dots & \binom{m}{1} - a_1 \\ -a_2 & 1 - a_2 - \lambda & \dots & \binom{m}{2} - a_2 \\ \vdots & \vdots & & \vdots \\ -a_m & -a_m & \dots & 1 - a_m - \lambda \end{pmatrix}$$

из остальных, получим

$$P(\lambda, a) = \det\begin{pmatrix} 1 - a_1 - \lambda & \binom{2}{1} - 1 + \lambda & \dots & \binom{m}{1} - 1 + \lambda \\ -a_2 & 1 - \lambda & \dots & \binom{m}{2} \\ \vdots & \vdots & & \vdots \\ -a_m & 0 & \dots & 1 - \lambda \end{pmatrix}, \tag{5}$$

откуда видно, что $P(\lambda, a) \in \pi_m(\lambda)$ и коэффициенты $P(\lambda, a)$ линейно зависят от a. С другой стороны, $P(\lambda, 0) = (1 - \lambda)^m$, так что найдутся $\alpha^i \in \mathbb{R}^m$, i = 1,..., m, для которых

$$P(\lambda, a) = (1-\lambda)^m + \sum_{k=1}^{m} \langle \alpha^k, a \rangle (1-\lambda)^{m-k}. \tag{6}$$

Докажем, что $\{\alpha^i\}_{i=1}^m$ – базис в $\mathbb{R}^m$. Пусть $\mathbb{R}^s$ – подпространство в $\mathbb{R}^m$, натянутое на

$\{\alpha^i\}_{i=1}^m$, $s < m$. Тогда найдется $a^* \in (\mathbb{R}^s)^\perp$, $a^* \neq 0$, такое, что $\langle \alpha^i, a^* \rangle = 0$, $\forall i$, $a_j^* = 0$, $j = s+1,\dots, m$. Из (6) следует, что $P(\lambda, a^*)$ имеет m-кратный корень $\lambda = 1$. Тогда из (5)

$$P(\lambda, a^*) = (1-\lambda)^m = \det\begin{pmatrix} 1-a_1^*-\lambda & \binom{2}{1}-1+\lambda & \dots & \binom{s}{1}-1+\lambda & \dots & \binom{m}{1}-1+\lambda \\ -a_2^* & 1-\lambda & \dots & \binom{s}{2} & \dots & \binom{m}{2} \\ \vdots & \vdots & & \vdots & & \vdots \\ -a_s^* & 0 & \dots & 1-\lambda & \dots & \binom{m}{s} \\ 0 & 0 & \dots & 0 & \dots & \binom{m}{s+1} \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1-\lambda \end{pmatrix}.$$

Не уменьшая общности, можно считать $a_s^* \neq 0$. Из предыдущего равенства имеем: $P(\lambda, a^*) = (1-\lambda)^{m-s}\, Q(\lambda)$, где

$$Q(\lambda) = \det\begin{pmatrix} 1-a_1^* & \binom{2}{1}-1+\lambda & \dots & \binom{s}{1}-1+\lambda \\ -a_2^* & 1-\lambda & \dots & \binom{s}{2} \\ \vdots & \vdots & & \vdots \\ -a_s^* & 0 & \dots & 1-\lambda \end{pmatrix}$$

и $Q(1) = -a_s^* s! \neq 0$. Противоречие показывает, что $s = m$, откуда следует утверждение теоремы.

ТЕОРЕМА 2. $\forall m \in \mathbb{N}$, $\varepsilon \geq 0$ $\exists$ $a \in \mathbb{R}^m$: $\rho(B(a)) = \varepsilon$ и разностное уравнение (4) имеет порядок $\geq m$.

ДОКАЗАТЕЛЬСТВО. Выберем произвольный многочлен $p(\lambda) \in \pi_m(\lambda)$, обладающий свойством: $\max\{|x|: p(x) = 0\} = \varepsilon$. Так как $P(\lambda, \cdot): \mathbb{R}^m \to \pi_m(\lambda)$ – изоморфизм (теор. 1), то уравнение $P(\lambda, a) = p(\lambda)$ однозначно разрешимо относительно a, что и требовалось.

Для практических целей удобно в теореме 2 брать $\varepsilon = 0$, т.е. решать уравнение $P(\lambda, a) = \lambda^m$. Так как в этом случае $B(a)^m = 0$ (матрица является нулем своего характеристического многочлена), то погрешности вычисления в конкретном узле сетки исчезают через m шагов. Таким образом, вычисление по (4) эквивалентно решению систем линейных уравнений с доминирующей диагональю.

Отметим несколько свойств решения уравнения $B(a)^m = 0$.

1) Нетрудно заметить из (5), что $P(\lambda, a) = \prod_{k=1}^m (1-a_k-\lambda) + p_{m-2}(\lambda)$, где $p_{m-2}(\lambda)$ – некоторый многочлен степени $\leq m-2$. Отсюда (раскрыв скобки) получим: $\sum_{k=1}^m a_k = m$.

2) Если в определителе $P(0, a)$ (см. (5)) сложить все строки, чередуя знаки, поместить результат в первую строку, то получим $P(0, a) = 1 - \sum_{k=1}^m (-1)^{k-1} a_k = 0$, откуда

$$\sum_{k=1}^{m}(-1)^{k-1}a_k = 1.$$

Непосредственные вычисления, кроме того, позволяют найти: $a_m = \frac{1}{m!}$, $a_{m-1} = = \frac{1}{m!}\binom{m+1}{2}$, $a_{m-2} = \frac{3m+2}{4m!}\binom{m+1}{3}$, $a_{m-3} = \frac{1}{m!}\binom{m+1}{2}\binom{m+1}{4}$, $a_{m-4} = \frac{15m^2(m+1)-2(5m+4)}{48m!}\binom{m+1}{5}$.

Трудно себе представить ситуацию, где требуется вычисление одновременно более чем трех-четырех производных в каждой точке. Тем не менее порядок точности формул (4) возрастает с m. С другой стороны, возрастают вычислительные погрешности на каждом шаге (оставаясь в силу свойств матрицы B в одних и тех же границах). Здесь ситуация напоминает использование разностных формул дифференцирования при неточных вычислениях – приходится выбирать компромисс между величиной шага сетки и порядком точности используемых формул.

Ниже приведена целочисленная матрица Z, с помощью которой для $m \le 10$ можно вычислить конкретные значения $a_k = \begin{cases} Z_{k,m} / Z_{m,k} \text{ при } k < m, \\ \frac{1}{m!} \text{ при } k = m. \end{cases}$

$$Z = \begin{pmatrix} 1 & 3 & 11 & 25 & 137 & 49 & 363 & 761 & 7129 & 7381 \\ 2 & 1 & 1 & 35 & 15 & 203 & 469 & 29531 & 6515 & 177133 \\ 6 & 1 & 1 & 5 & 17 & 49 & 967 & 267 & 4523 & 84095 \\ 12 & 24 & 12 & 1 & 1 & 35 & 7 & 1069 & 95 & 341693 \\ 60 & 8 & 24 & 8 & 1 & 7 & 23 & 9 & 3013 & 8591 \\ 20 & 90 & 48 & 144 & 240 & 1 & 1 & 13 & 5 & 7513 \\ 140 & 180 & 720 & 18 & 360 & 180 & 1 & 1 & 29 & 121 \\ 280 & 10080 & 160 & 1920 & 80 & 960 & 1120 & 1 & 1 & 11 \\ 2520 & 2016 & 2268 & 128 & 17280 & 192 & 12096 & 8064 & 1 & 11 \\ 2520 & 50400 & 36288 & 362880 & 34560 & 172800 & 24192 & 30240 & 725760 & 1 \end{pmatrix}$$

Например, для $m = 5$ имеем $a = \left(\frac{137}{60}, \frac{15}{8}, \frac{17}{24}, \frac{1}{8}, \frac{1}{120}\right)^T$. На начальном этапе полез-но выполнить не менее m шагов вперед с произвольными значениями производных в точке x_0, затем, вернувшись (обратным ходом) к началу, получить необходимые значения в исходной точке для запуска (4), после чего применить (4) ко всем сеточным значениям. Алгоритм прост, вычислительные затраты сравнимы с затратами метода прогонки для систем линейных уравнений соответствующего порядка.

Пример 1. Для аналитической функции sin(x), заданной своими значениями $y_i = \sin(x_i)$ на сетке $x_i = i \cdot h$, $i = 0, 1,...,20$, с шагом $h = 0.1$ получены приближенные значения u_i, v_i вторых производных сеточной функции с помощью кубических сплайнов (в соответствии с (2)) и итераций (4) при $m = 8$ соответственно. В евклидовой (сеточной) норме имеем $\| u + y \| = 9.3_{10}^{-3}$, $\| v + y \| = 1.3_{10}^{-7}$, что вполне согласуется с теорией. В тех же условиях вычислены приближенные значения четвертой производной p_i и q_i с помощью повторного дифференцирования (2) и итераций (4) соответственно, в результате $\|p - y\| > 3$, $\|q - y\| < 10^{-4}$ (плохой результат при использовании сплайнов связан с сильной зависимостью от выбора краевых условий, что приводит к необходимости решать задачу регуляризации; в последнем методе эта зависимость отсутствует).

Пример 2. Использование матрицы значений всех необходимых производных позволяет решать некоторые обратные задачи, например, по конструированию линейных дифференциальных операторов. В условиях предыдущего примера можно легко получить среднеквадратичные оценки представления $y'' = a + by + cy'$. В дан-ном случае $a = -5.9_{10}^{-6}$, $b = -1$, $c = -1.6_{10}^{-6}$, и если значениями a и c пренебречь в пределах допустимой погрешности, то $y'' + y = 0$, как и следовало ожидать. Можно попытаться и в более сложном случае попробовать сконструировать систему с заданным сигналом на выходе. Так, для функции $y(t) = \dfrac{te^{-t}}{1+t}$, $t \in [0, 2]$, аналогично предыдущему найдем оценки параметров линейного оператора вида $L(y, t) = (c_0 t + c_1)\, y'' + (c_2 t + c_3)\, y' + (c_4 t + 1)\, y$ так, чтобы $L(y, t) = 0$. В условиях примера 1 находим оценки $c = (0.141, 0.272, -0.14, 1.184, -0.321)^T$. Решение краевой задачи $L(w, t) = 0$, $w(0) = y(0)$, $w(2) = y(2)$, полученное в пакете MCad, отличается от y(t) не более чем на 4.1_{10}^{-4} (рисунок).

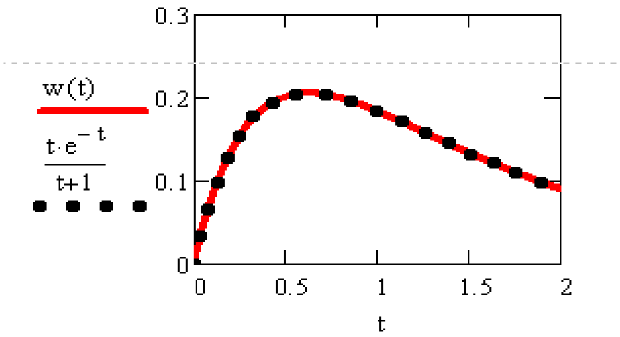

Рис.
Fig.

СПИСОК ИСПОЛЬЗОВАННЫХ ЛИТЕРАТУРНЫХ ИСТОЧНИКОВ

1. Арестов, В.В. Наилучшее приближение операторов дифференцирования. /В.В. Арестов // Математические заметки − 1967. − т.1. − №2. − С. 149-154.

2. Бахвалов, Н.С. Численные методы / Н.С. Бахвалов, Н.П. Жидков, Г.М. Кобельков. − М.: Наука, 1980. − 630 с.

3.Стечкин, С.Б. Сплайны в вычислительной математике / С.Б. Стечкин, Ю.Н. Субботин. − М.: Наука, 1976. − 248 с.

4. Kramer H.K. Decomposition of a function into a weighted sum of shifted replicas of another function. /H.K.Kramer, R.N. Lane // Journal of Mathematical Analysis and Applications. − 1974. − v.46. − N 3. − p. 395-608.

5. Rong Quing Jia. Linear independence of translates of a box spline. /Jia Quing Rong //Journal of Approximatin Theory − 1984 − v.40. − N 2. − p. 158-160.

6. Тихонов, А.Н. Методы решения некорректных задач / А.Н. Тихонов, В.Я. Арсенин. − М.: Наука, 1974. − 260 с.

7. Бор К.де. Практическое руководство по сплайнам / К. де Бор. – М.: Радио и связь, 1985.– 303 с.

8. Боянов, Б.Д. Наилучшие методы интерполирования для некоторых классов дифференцируемых функций / Б.Д. Боянов // Математические заметки. − 1975. − т.17. − №4. − С. 511-524.

9. Пахнутов, И.А. Аппроксимация пакетом ломаных / И.А. Пахнутов. //Известия КГТУ − 2007. − №12. − С. 190-199.

10. Nordsieck A. On numerical integration of ordinary differential equations / A. Nordsieck // Mathematics of Computations − 1962. − v. 16. − N.77. − p. 22-49.

11. Ортега, Дж. Итерационные методы решения нелинейных систем уравнений со многими неизвестными /Дж. Ортега, В. Рейнболдт. − М.: Мир. − 1975.

DIFFERENTIATION OF GRID FUNCTIONS

I.A. Pahnutov

Methodology of derivatives estimations is considered for functions given on a finite mesh of real line. Local and mesh estimates compared from computational point of view. Iterative method of simultaneous estimation of all necessary derivatives is discussed. Examples show some preferences of considered approaches compared to local difference formulas.

numerical differentiation, approximation, difference formulas