УДК 519.63

Л. Е. Довгилович, И. Л. Софронов

Московский физико-технический институт (государственный университет)
Московский исследовательский центр Шлюмберже

Анализ явных и неявных центрально-разностных операторов для вычисления второй производной на равномерных сетках

В работе рассмотрена задача вычисления второй производной от достаточно гладкой функции с помощью конечных разностей высокого порядка на равномерной сетке. Формулируются критерии сравнения различных алгоритмов вычисления по точности, памяти и количеству операций. Проанализированы два семейства конечно-разностных операторов с порядками аппроксимации от 4-го до 20-го: центрально-разностные и неявные центрально-разностные. Выявлено, что неявные операторы высокого порядка аппроксимации с трехдиагональной матрицей в левой части обладают преимуществом по всем критериям. Также показано, что такие операторы можно эффективно применять для вычислений на высокопроизводительных системах с распределенной памятью с помощью разбиения на подобласти.

Ключевые слова: вторая производная, аппроксимация, центрально-разностные операторы, неявные центрально-разностные операторы, компактные схемы, волновое уравнение.

Введение

Оператор вычисления второй производной является одним из наиболее используемых в конечно-разностных методах. Например, при решении волнового уравнения. Чтобы конкретизировать предмет нашего исследования, мы рассмотрим *одномерный случай* и *равномерные сетки*. Обобщение на многомерность, т.е. на оператор Лапласа, можно сделать путем последовательного применения одномерных производных по разным направлениям.

Нетривиальность предпринятого анализа заключается в исследовании разностных операторов *высокого порядка аппроксимации*. Занимая нишу между классическим трехточечным оператором второго порядка и спектральным оператором [1], они позволяют получать алгоритмы вычисления второй производной, оптимальные с точки зрения вычислительных затрат, памяти и времени вычислений. В многоядерных процессорах, в том числе в GPU («графических картах»), число арифметических операций, которое может выполнить процессор за время чтения/записи одного числа из основной памяти, исчисляется десятками и сотнями. Иными словами, выгоднее становятся те алгоритмы, которые за счет б*о*льшего числа операций на точку сетки позволяют использовать меньшие объемы сеток. Поэтому разностные схемы высокого порядка стали уже привычными и вытесняют традиционные трехточечные схемы.

Данная работа посвящена подробному теоретическому сравнению двух подходов вычисления второй производной функции на отрезке: центрально-разностными операторами (CDO) и неявными центрально-разностными операторами ($ICDO$), часто называемыми «компактными конечно-разностными операторами». Мы задались целью показать основные *возможные* преимущества $ICDO$ над CDO. Большинство приводимых результатов было получено при разработке алгоритмов высокого порядка точности для решения волнового уравнения. Тем не менее эти результаты обладают более широким спектром применений. Мы ограничились рассмотрением $ICDO$ с *трехдиагональными* операторами в левой части. Другие случаи $ICDO$ можно исследовать аналогично. Но предварительный анализ показывает, что они не так привлекательны.

Неявные центрально-разностные операторы не новы, и им посвящено много исследований, см., например, [2–7]. Однако, не все возникшие у нас вопросы освещены в указанных публикациях. Первые результаты нашего собственного анализа были представлены в [8]. Наш собственный опыт использования $ICDO$ стимулировал дальнейший анализ, который и приводится в данной работе.

Мы рассматриваем разностные операторы и проводим их сравнение, исходя из трех основных критериев:

1) Точность решения на заданной сетке.

2) Размер сетки, обеспечивающей заданную точность.

3) Количество операций с плавающей точкой, затраченных на решение задачи с заданной точностью.

Для каждого из приведенных критериев предлагается инструментарий, позволяющий помочь с оптимальным выбором типа оператора.

1. Постановка задачи

Вначале обговорим обозначения. На протяжении всей работы мы используем только одну непрерывную функцию $u(x)$ и ее производные. Эта функция определена на отрезке $[x_{\min}; x_{\max}]$, на котором задана равномерная сетка:

$$x_i = x_{\min} + ih, \quad h = \frac{x_{\max} - x_{\min}}{N}, \quad i = 0, ..., N. \tag{1}$$

Остальные функции в работе — дискретные со значениями, приписанными к точкам x_i, т.е.

$$\bar{g} = \{g_i\}, \quad i = 0, ..., N.$$

Для функции $u(x)$ вводим обозначения

$$\bar{u} = \{u_i\}, \quad u_i = u(x_i),$$

$$\bar{u}'' = \{u_i''\}, \quad u_i'' = u''(x_i),$$

в соответствии с этими обозначениями всегда будет понятно, о какой функции — дискретной или непрерывной — идет речь.

Задача

Пусть функция $u(x)$ имеет достаточное количество производных на отрезке $[x_{\min}; x_{\max}]$. По значениям функции $\bar{u}$ на равномерной сетке (1) мы хотим вычислить вектор $\bar{g}$ аппроксимации второй производной $\bar{u}''$, такой, что ошибка

$$\bar{\varepsilon} = \bar{g} - \bar{u''}$$

оценивается как

$$\max_i(|\varepsilon_i|) \leqslant O(h^P),$$

где четное целое P обозначает порядок аппроксимации.

Неявные центрально-разностные операторы

Предположим вначале, что $u(x)$ — периодическая функция с периодом $x_{\max} - x_{\min}$. В данной работе мы рассмотрим два вида операторов аппроксимации второй производной. Первый вид — это центрально-разностные операторы (CDO_P), которые можно представить как

$$g_i = CDO_P(u_i) = \frac{\beta_0}{h^2}u_i + \sum_{j=1}^{P/2}\frac{\beta_j}{h^2}(u_{i+j} + u_{i-j}). \tag{2}$$

Здесь и далее для периодического случая считаем, что все индексы берутся по модулю N, то есть $i \pm j \to i \pm j \mp N$, если $i \pm j \notin [0, N-1]$.

Второй вид — это неявные центрально-разностные операторы ($ICDO_P$), часто называемые компактными операторами, которые можно получить из уравнения

$$g_i + \alpha(g_{i-1} + g_{i+1}) = \frac{\beta_0}{h^2}u_i + \sum_{j=1}^{P/2-1}\frac{\beta_j}{h^2}(u_{i+j} + u_{i-j}), \tag{3}$$

или

$$A_P\bar{g} = B_P\bar{u}, \tag{4}$$

где A_P и B_P — сеточные операторы с коэффициентами из (3); индекс P в коэффициентах α и β_j мы опускаем.

Согласно (3) мы исследуем случай трехдиагональных $ICDO_P$, хотя можно вводить два и больше дополнительных коэффициентов в операторе левой части (4) и рассматривать пятидиагональные $ICDO_P$ и т.д.

Обозначим через $\bar{\delta}$ вектор погрешности аппроксимации:

$$\bar{\delta} = A_P\bar{u''} - B_P\bar{u}. \tag{5}$$

Тогда ошибку $\bar{\varepsilon}$ вычисления второй производной можно выразить следующим образом:

$$\bar{\varepsilon} = A_P^{-1}\bar{\delta}. \tag{6}$$

Для единообразия полагаем, что (5), (6) применяются и для CDO_P (2) с $A_P = 1$.

2. Анализ на основе рядов Тейлора

Данный вид анализа позволяет нам сравнить поведение операторов в асимптотическом приближении $h \to 0$.

Во многих статьях, в частности в [3, 4], явные и неявные операторы сравниваются по погрешности аппроксимации $\bar{\delta}$. Но этого, очевидно, недостаточно: сравнение необходимо проводить по $\bar{\varepsilon}$, см. (6), так как именно вектор $\bar{\varepsilon}$ характеризует точность приближения второй производной. Для CDO_P эти векторы совпадают и находятся разложением в ряд Тейлора выражения $CDO_P(u_i) - u_i''$, см. (2). Например, для центрально-разностной схемы шестого порядка, как известно, $\max|\delta_i| = 72R_6$, где введено обозначение

$$R_P = \max|u^{(P+2)}|\frac{h^P}{(P+2)!}. \tag{7}$$

Для $ICDO_P$ оценку ε_i надо проводить исходя их формулы (6). Проще всего это сделать на классе периодических функций, чтобы исключить краевые эффекты. В этом случае матрица A_P является циклической трехдиагональной:

$$A_P = \begin{pmatrix} 1 & \alpha & & & \alpha \\ \alpha & 1 & \alpha & & \\ & & \ldots & & \\ & & \alpha & 1 & \alpha \\ \alpha & & & \alpha & 1 \end{pmatrix} \tag{8}$$

со значением α, зависящим от P и лежащим в интервале $0 < \alpha < 0.5$. Приведем здесь без доказательства следующую теорему.

Теорема. Для трехдиагонального оператора $ICDO_P$ ошибка $\bar{\varepsilon}$ в (6) оценивается при $h \to 0$ посредством

$$\max_i |\varepsilon_i| \leqslant \left(\frac{1}{1+2\alpha} + O(h) \right) \max_i |\delta_i|,$$

где погрешность аппроксимации $\bar{\delta}$ определяется из (7).

На рис. 1 мы видим теоретически полученное предельное отношение ошибок CDO_P к ошибке $ICDO_P$, посчитанное из сравнения величин $\bar{\delta}$ в тейлоровских разложениях и последующим применением результатов теоремы для $ICDO_P$.

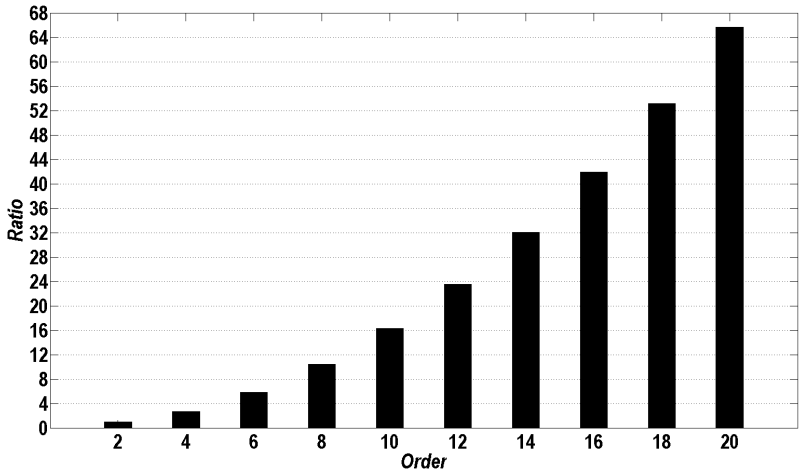

Рис. 1. Отношение ошибки CDO_P к ошибке $ICDO_P$ в зависимости от порядка P

Из данного анализа мы можем сравнить, во сколько раз решение, полученное с помощью $ICDO_P$, будет точнее решения, полученного с помощью CDO_P, при использовании одинаковой сетки. Это помогает выбрать оптимальный тип оператора для заданной сетки по первому критерию.

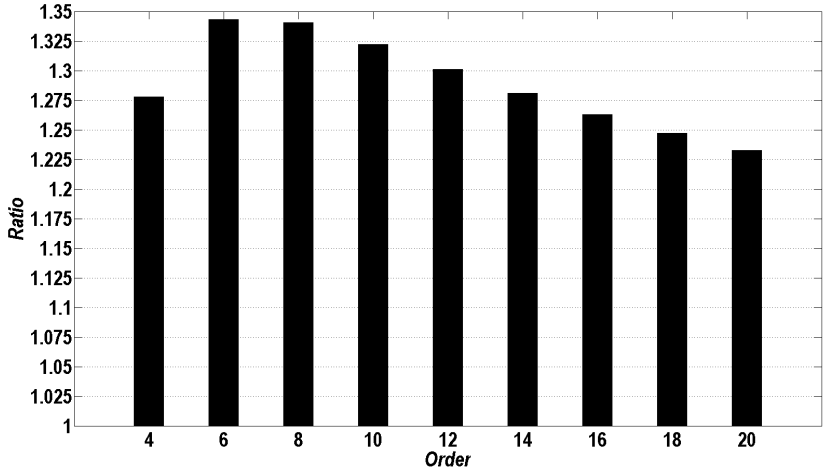

Рис. 2. Отношение шагов сетки, необходимых для достижения заданной точности при использовании $ICDO_P$ и CDO_P, в зависимости от порядка P

Теперь обратимся ко второму критерию. Для этого нам нужно взять корень степени P из отношения ошибок, показанных на рис. 1. На рис. 2 мы можем видеть результаты

данных оценок, из которых видно, что использование $ICDO_P$ позволяет увеличить шаг сетки на 23–34% при сохранении точности в зависимости от порядка.

3. Вычисления в приграничных точках

Здесь мы рассмотрим вопрос о построении операторов $ICDO_P$ в непериодическом случае. Чтобы определиться с записью операторов вблизи границ отрезка, будем считать, что для CDO_P этот вопрос решен (например, с использованием односторонних разностей или граничных условий исходной дифференциальной задачи). Таким образом, предположим, что для правой части (4) мы можем вычислить вторые производные вплоть до граничных точек $i = 0$, $i = N$ с порядком аппроксимации в граничных точках не ниже, чем у $ICDO_P$. Тогда эти «явные» значения будем использовать в качестве граничных условий для вычисления вторых производных внутри отрезка, а сам оператор A_P модифицируем на решение получаемой задачи Дирихле:

$$A'_P = \begin{pmatrix} 1 & & & & \\ \alpha & 1 & \alpha & & \\ & .. & .. & .. & \\ & & \alpha & 1 & \alpha \\ & & & & 1 \end{pmatrix}. \tag{9}$$

Проанализируем на периодической задаче, как данный прием повлияет на величины ошибок $\bar{\delta}'$ и $\bar{\varepsilon}'$ для $ICDO_P$. То есть мы хотим сравнить решение двух задач: первой

$$A_P \bar{g} = \bar{f},$$

где A_P определяется формулой (8), и второй

$$A'_P \bar{g}' = \bar{f}',$$

где

$$f'_i = \begin{cases} f_i, & i = 1..N-1, \\ g_i + \eta_i, & i = 0, N, \end{cases}$$

а η_i — есть ошибка, с которой мы вычислили граничные значения исходной (первой) задачи. Тогда ошибка, вносимая такой модификацией, может быть вычислена как

$$\bar{\varepsilon} = \bar{g} - \bar{g}' = {A'}_P^{-1}({A'}_P A_P^{-1} \bar{f} - \bar{f}'). \tag{10}$$

Матрица A'_P может быть представленная как $A'_P = A_P - M_P$, где

$$M_P = \begin{pmatrix} 0 & \alpha & 0 & & \alpha \\ 0 & 0 & 0 & & \\ & .. & .. & .. & \\ & & 0 & 0 & 0 \\ \alpha & & 0 & \alpha & 0 \end{pmatrix}.$$

Учтем также, что

$$M_P A_P^{-1} \bar{f} = \begin{vmatrix} \alpha(g_1 + g_N) \\ 0 \\ 0 \\ \alpha(g_0 + g_{N-1}) \end{vmatrix} = \begin{vmatrix} g_0 + \alpha(g_1 + g_N) - g_0 \\ 0 \\ 0 \\ g_N + \alpha(g_0 + g_{N-1}) - g_N \end{vmatrix} = \begin{vmatrix} f_0 - g_0 \\ 0 \\ 0 \\ f_N - g_N \end{vmatrix}.$$

Тогда (10) можно записать как

$$\bar{\varepsilon} = \bar{g} - \bar{g}' = {A'}_P^{-1}(\bar{f} - M_P A_P^{-1} \bar{f} - \bar{f}') = {A'}_P^{-1} \bar{\delta}, \tag{11}$$

где

$$\bar{\delta} = -\begin{vmatrix} \eta_0 \\ 0 \\ 0 \\ \eta_N \end{vmatrix}.$$

Из линейности (11) следует, что влияние ошибки на границе можно оценить с помощью решения следующей системы уравнений:

$$A'_P \bar{v} = \bar{f} = \begin{vmatrix} 0 \\ \dots \\ 0 \\ 1 \end{vmatrix}. \tag{12}$$

Решение данной системы прогонкой записывается следующим образом:

$$a_{i+1} = \tfrac{-\alpha}{1+\alpha a_i},\ \ i = 1..N-1,\ a_1 = 0;$$
$$v_N = 1,\ v_i = a_{i+1} v_{i+1} = \textstyle\prod_{k=i+1}^{N} a_k.$$

Обратимся теперь к последовательности a_i. При условии $\frac{1}{2} > \alpha > 0$ можно показать, что последовательность a_i убывает и ограничена снизу, следовательно, она сходится. Ее предел, вычисляемый из уравнения $a = -\alpha/(1+\alpha a)$, равен

$$a = \frac{-1+\sqrt{1-4\alpha^2}}{2\alpha}\ . \tag{13}$$

Также можно оценить скорость сходимости a_i:

$$|a_{i+1} - a_i| = |a_{i+1}| \left(\prod_{k=3}^{i} a_k^2 \right) \leqslant |a|^{2(i-3)+1}\ . \tag{14}$$

Теперь вернемся к решению задачи (12):

$$v_N = 1,$$
$$|v_i| = \textstyle\prod_{k=i+1}^{N} |a_k| < \prod_{k=i+1}^{N} |a| = |a|^{N-i-1}\ \cdot$$

То есть $|v_i|$ убывает быстрее, чем $|a|^{N-i-1}$. Значения предела a_i при $i \to \infty$ для рассматриваемых операторов принадлежат промежутку от $a = -0.445$ для 20-го порядка до $a = -0.101$ для 4-го.

Таким образом, погрешность, генерируемая неточностью вычисления второй производной функции в граничной точке, убывает со скоростью геометрической прогрессии с достаточно малым знаменателем по мере удаления от границы внутрь области.

Этот результат важен также для организации параллельных вычислений, т.е. исходный отрезок можно разбивать на отрезки с достаточно небольшим перекрытием и вычислять на них $ICDO_P$ независимо с помощью (9).

4. Оценки количества операций с плавающей точкой для получения заданной точности

Здесь мы проведем анализ третьего критерия. Для начала нам необходимо оценить количество операций, затрачиваемых на вычисления с помощью рассматриваемых операторов. Для CDO_P эта оценка делается довольно просто; с учетом симметрии коэффициентов получаем, что с точностью до некоторой константы количество арифметических операций составляет

$$N_{op}^{CDO_P} = \left(1 + 3\frac{P}{2}\right) N, \tag{15}$$

где $N \gg P$ — число точек сетки на отрезке. Расчет с помощью $ICDO_P$ происходит в два этапа, первый этап — это вычисление правой части в (3) и совпадает по сложности с CDO_{P-2}, второй — это решение трехдиагональной системы линейных уравнений методом прогонки:

$$A'_P \bar{g} = \bar{f}. \tag{16}$$

Формулы для прогонки записываются следующим образом:

$$g_i = a_{i+1} g_{i+1} + b_{i+1},$$

$$a_{i+1} = \frac{-\alpha}{(1 + \alpha a_i)}, \quad i = 2...N-1,$$

$$b_{i+1} = (b_i - \tilde{f}_i) a_{i+1}, \quad i = 2...N-1,$$

$$a_2 = 0; \quad b_2 = \alpha \tilde{f}_1; \quad g_N = \alpha \tilde{f}_N; \quad \tilde{f}_i = \frac{f_i}{\alpha}.$$

Заметим, что a_i не зависят от правой части и, как было показано ранее (см. (13) и (14)), быстро сходится к своему пределу. Полагая, что система (16) будет решаться многократно с различными правыми частями, мы не учитываем операции, затраченные на вычисление коэффициентов a_i. Также мы видим, что в формулах фигурирует только $\tilde{f}_i = \frac{f_i}{\alpha}$. Поэтому при вычислении правой части по формуле

$$f_i = \frac{\beta_0}{h^2} u_i + \sum_{j=1}^{P/2-1} \frac{\beta_j}{h^2} (u_{i+j} + u_{i-j}),$$

см. (3), нам достаточно заменить β_j на $\tilde{\beta}_j = \frac{\beta_j}{\alpha}$ для вычисления $\tilde{f}_i$. Таким образом, мы сэкономим еще одну операцию из расчета на точку. После всех подсчетов мы получаем, что количество операций, требуемых для вычисления с помощью $ICDO_P$, оценивается формулой

$$N_{op}^{ICDO_P} = \left(2 + 3\frac{P}{2}\right) N. \tag{17}$$

Оценим величину отношения количества операций, требуемых для достижения заданной точности при использовании CDO_P и $ICDO_P$ в зависимости от порядка P. Для этого отмасштабируем диаграммы на рис. 2, используя (15) и (17). Результат представлен на рис. 3. Как и в случае первых двух критериев, мы получаем, что использование $ICDO_P$ предпочтительнее, нежели чем CDO_P.

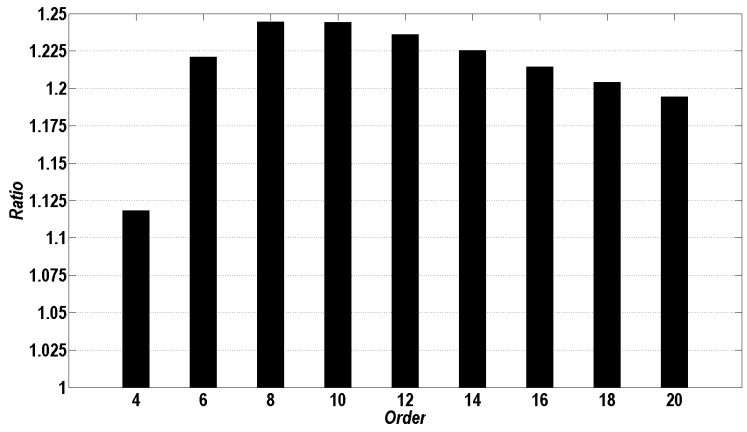

Рис. 3. Отношение количества операций, необходимого для достижения заданной точности при использовании CDO_P и $ICDO_P$ в зависимости от порядка P

Заключение

В данной работе мы провели анализ точности и вычислительной сложности центрально-разностных и неявных центрально-разностных операторов высокого порядка аппроксимации для вычисления второй производной на равномерной сетке. Использовались оценки, полученные на основе рядов Тейлора.

Во введении сформулированы три критерия, по которым можно выбирать конкретный оператор для решения прикладных задач. Они отражают проблемы точности, памяти и времени вычислений. На практике приходится находить компромисс для одновременного выполнения этих критериев, а также для учета эффективности программной реализации операторов, архитектуры вычислительной системы и т.п.

На основе предложенного анализа показано, что использование $ICDO$ с *трехдиагональными* операторами в левой части предпочтительнее с точки зрения всех трех критериев в асимптотическом приближении $h \to 0$, см. рис. 1, рис. 2, рис. 3;

Исследован вопрос эффективного вычисления $ICDO$ с *трехдиагональными* операторами в левой части, см. (3). Благодаря сильному диагональному преобладанию обращаемого оператора оказалось возможным применять те же принципы разбиения вычислительной области при реализации параллельных вычислений на системах с распределенной памятью, что и при использовании центрально-разностных операторов. В случае $ICDO$ область перекрытия соседних сеток оценивается длиной шаблона оператора правой части; на практике эта область ненамного больше, чем для CDO того же порядка.

Работа выполнена при поддержке РФФИ, грант 10-01-00567.

Литература

1. *Самарский А.А., Гулин А.В.* Численные методы. — М.: Наука, 1989.

2. *Толстых А.И.* Компактные разностные схемы и их применение в задачах аэрогидродинамики. — М.: Наука, 1990.

3. *Chu P.C., Fan C.* A three-point combined compact difference scheme // J. Comp. Phys. — 1998. — V. 140. — P. 370—399.

4. *Chu P.C., Fan C.* A three-point sixth-order no uniform combined compact difference scheme // J. Comp. Phys. — 1999. — V. 148. — P. 663–674.

5. *Zhang J., Zhao J.J.* Truncation error and oscillation property of the combined compact difference scheme // Applied Mathematics and Computation. — 2005. — V. 161, N. 1. — P. 241–251.

6. *Lele S.K.* Compact finite difference schemes with spectral-like resolution //J. Comp. Phys. — 1992. — V. 103. — P. 16—42.

7. *Liu Y., Sen M.K.* A practical implicit finite-difference method: examples from seismic modeling // Journal of Geophysics and Engineering. — 2009. — V. 6. — P. 231.

8. Довгилович Л.Е., Софронов И.Л. О применении компактных схем для решения волнового уравнения: препринт / ИПМ им. Келдыша РАН. — М., 2008. — № 84.

Поступила в редакцию 23.04.2012.