Chapter 2: Finite-Difference Approximations to Derivatives

Copyright 2015, David A. Randall

2.1 Finite-difference quotients

Consider the derivative \( \frac{df}{dx} \), where \( f = f(x) \), and \( x \) is the independent variable (which could be either space or time). Finite-difference methods represent the continuous function \( f(x) \) by a set of values defined at a finite number of discrete points in a specified (spatial or temporal) region. Thus, we usually introduce a “grid” with discrete points where the variable \( f \) is defined,

\[
\begin{align*}
\Delta x &= j + \frac{1}{2} \\
j - 1 & \quad j \quad j + 1
\end{align*}
\]

Figure 2.1: An example of a grid, with uniform grid spacing \( \Delta x \). The grid points are denoted by the integer index \( j \). Half-integer grid points can also be defined, as shown by the x at \( j + \frac{1}{2} \).

as shown in Figure 2.1. Sometimes the words “mesh” or “lattice” are used in place of “grid”. The interval \( \Delta x \) is called the grid spacing, grid size, mesh size, etc. We assume that the grid spacing is constant for the time being; then \( x_j = j \Delta x \), where \( j \) is the “index” used to identify the grid points. Note that \( x \) is defined only at the grid points denoted by the integers \( j, j+1 \), etc.

Using the notation \( f_j = f(x_j) = f(j \Delta x) \), we can define

the forward difference at the point \( j \) by \( f_{j+1} - f_j \),

\[
(1)
\]
the *backward difference* at the point \( j \) by \( f_j - f_{j-1} \), and

\[
\text{backward difference at point } j = f_j - f_{j-1}.
\]

(2)

the *centered difference* quotient at the point \( j + \frac{1}{2} \) by \( f_{j+1} - f_j \).

\[
\text{centered difference quotient at point } j + \frac{1}{2} = \frac{f_{j+1} - f_j}{\Delta x}.
\]

(3)

Note that \( f \) itself is not defined at the point \( j + \frac{1}{2} \). From (1) - (3) we can define the following “finite-difference quotients:” the forward-difference quotient at the point \( j \):

\[
\left( \frac{df}{dx} \right)_{j, \text{approx}} = \frac{f_{j+1} - f_j}{\Delta x},
\]

(4)

the backward-difference quotient at the point \( j \):

\[
\left( \frac{df}{dx} \right)_{j, \text{approx}} = \frac{f_j - f_{j-1}}{\Delta x};
\]

(5)

and the centered-difference quotient at the point \( j + \frac{1}{2} \):

\[
\left( \frac{df}{dx} \right)_{j+\frac{1}{2}, \text{approx}} = \frac{f_{j+1} - f_j}{\Delta x}.
\]

(6)

In addition, the centered-difference quotient at the point \( j \) can be defined by

\[
\left( \frac{df}{dx} \right)_{j, \text{approx}} = \frac{f_{j+1} - f_{j-1}}{2\Delta x}.
\]

(7)

Since (4) and (5) employ the values of \( f \) at two points, and give an approximation to \( \frac{df}{dx} \) at one of the same points, they are sometimes called *two-point approximations*. On the other hand, (6) and (7) are *three-point approximations*, because the approximation to \( \frac{df}{dx} \) is defined at a location different from the locations of the two values of \( f \) on the right-hand side of the equals sign.
When $x$ is time, the time point is frequently referred to as a “level.” In that case, (4) and (5) can be referred to as two-level approximations and (6) and (7) as three-level approximations.

How accurate are these finite-difference approximations? “Accuracy” can be measured in a variety of ways, as we shall see. One measure of accuracy is truncation error. The term refers to the truncation of an infinite series expansion. As an example, consider the forward difference quotient

$$
\frac{df}{dx}_{j, \text{approx}} = \frac{f_{j+1} - f_j}{\Delta x} = f\left( (j+1)\Delta x \right) - f(j\Delta x) \equiv \frac{f\left( (j+1)\Delta x \right) - f(j\Delta x)}{\Delta x}.
$$

(8)

Expand $f$ in a Taylor series about the point $x_j$, as follows:

$$
f_{j+1} = f_j + \Delta x \left( \frac{df}{dx} \right)_j + \frac{\Delta x^2}{2!} \left( \frac{d^2 f}{dx^2} \right)_j + \frac{\Delta x^3}{3!} \left( \frac{d^3 f}{dx^3} \right)_j + \cdots + \frac{\Delta x^{n-1}}{(n-1)!} \left( \frac{d^{n-1} f}{dx^{n-1}} \right)_j + \cdots.
$$

(9)

The expansion (9) can be derived without any assumptions or approximations except that the indicated derivatives exist (Arfken, 1985; for a very brief review see the QuickStudy on Taylor's Series). Eq. (9) can be rearranged to

$$
\frac{f_{j+1} - f_j}{\Delta x} = \left( \frac{df}{dx} \right)_j + \epsilon,
$$

(10)

where

$$
\epsilon \equiv \Delta x \left( \frac{d^2 f}{dx^2} \right)_j + \frac{\Delta x^2}{2!} \left( \frac{d^3 f}{dx^3} \right)_j + \cdots + \frac{\Delta x^{n-2}}{(n-1)!} \left( \frac{d^{n-1} f}{dx^{n-1}} \right)_j + \cdots.
$$

(11)

is called the “truncation error.” If $\Delta x$ is small enough, the leading term on the right-hand side of (11) will be the largest part of the error. The lowest power of $\Delta x$ that appears in the truncation error is called the “order of accuracy” of the corresponding difference quotient. For example, the leading term of (11) is of order $\Delta x$, abbreviated as $O(\Delta x)$, and so we say that (10) is a first-order approximation or an approximation of first-order accuracy. Obviously (5) is also first-order accurate. Just to be as clear as possible, a first-order scheme for the first derivative has the form

$$
\left( \frac{df}{dx} \right)_{j, \text{approx}} = \left( \frac{df}{dx} \right)_j + O[\Delta x], \text{ where } \left( \frac{df}{dx} \right)_{j, \text{approx}} \text{ is an approximation to the first derivative, and}
$$

An Introduction to Numerical Modeling of the Atmosphere
\[
\left( \frac{df}{dx} \right)_j \text{ is the true first derivative. Similarly, a second-order accurate scheme for the first}
\]
\[
derivative has the form \( \left( \frac{df}{dx} \right)_{j, \text{approx}} = \left( \frac{df}{dx} \right)_j + O[(\Delta x)^2] \), and so on for higher orders of accuracy.
\]

Similar analyses of (6) and (7) show that they are of second-order accuracy. For example, we can write
\[
f_{j+1} - f_j = \left( \frac{df}{dx} \right)_j (\Delta x) + \left( \frac{d^2 f}{dx^2} \right)_j \left( \frac{-\Delta x}{2!} \right) + \left( \frac{d^3 f}{dx^3} \right)_j \left[ \frac{-\Delta x}{3!} \right] + \cdots.
\]

Subtracting (12) from (9) gives
\[
f_{j+1} - f_{j-1} = 2\left( \frac{df}{dx} \right)_j (\Delta x) + \frac{2}{3!} \left( \frac{d^3 f}{dx^3} \right)_j \left[ (\Delta x)^3 \right] + \cdots \text{ odd powers only,}
\]

which can be rearranged to
\[
\left( \frac{df}{dx} \right)_j = \frac{f_{j+1} - f_{j-1}}{2\Delta x} - \left( \frac{d^3 f}{dx^3} \right)_j \frac{\Delta x^2}{3!} + O[(\Delta x)^4].
\]

Similarly,
\[
\left( \frac{df}{dx} \right)_{j+\frac{1}{2}} = \frac{f_{j+1} - f_j}{\Delta x} - \left( \frac{d^3 f}{dx^3} \right)_{j+\frac{1}{2}} \frac{(\Delta x / 2)^2}{3!} + O[(\Delta x)^4].
\]

From (14) and (15), we see that
\[
\frac{\text{Error of (14)}}{\text{Error of (15)}} \equiv \frac{\left( \frac{d^3 f}{dx^3} \right)_j \frac{\Delta x^2}{3!}}{\left( \frac{d^3 f}{dx^3} \right)_{j+\frac{1}{2}} \frac{(\Delta x / 2)^2}{3!}} = \frac{4 \left( \frac{d^3 f}{dx^3} \right)_j}{\left( \frac{d^3 f}{dx^3} \right)_{j+\frac{1}{2}}} \equiv 4.
\]

This shows that the error of (14) is about four times as large as the error of (15), even though both finite-difference quotients have second-order accuracy. The point is that the "order of
accuracy” tells how rapidly the error changes as the grid is refined, but it does not tell how large the error is for a given grid size. It is possible for a scheme of low-order accuracy to give a more accurate result than a scheme of higher-order accuracy, if a finer grid spacing is used with the low-order scheme.

Suppose that the leading (and dominant) term of the error has the form

$$\varepsilon \equiv C(\Delta x)^p.$$  

(17)

From (17) we see that

$$\ln(\varepsilon) \equiv p \ln(\Delta x) + \ln(C),$$

and so

$$\frac{d[\ln(\varepsilon)]}{d[\ln(\Delta x)]} \equiv p.$$  

(18)

The implication is that if we plot $\ln(\varepsilon)$ as a function of $\ln(\Delta x)$ (i.e., plot the error as a function of the grid spacing on “log-log” paper), we will get (approximately) a straight line whose slope is $p$. This is a simple way to determine empirically the order of accuracy of a finite-difference quotient. Of course, in order to carry this out in practice it is necessary to compute the error of the finite-difference approximation, and that can only be done when the exact derivative is known. For that reason, this empirical approach is usually implemented by using an analytical “test function.” Can you think of a way to do it approximately even when the exact solution is not known?

### 2.2 Difference quotients of higher accuracy

Suppose that we write

$$\frac{f_{j+2} - f_{j-2}}{4\Delta x} = \left(\frac{df}{dx}\right)_j + \frac{1}{3!}\left(\frac{d^3 f}{dx^3}\right)_j (2\Delta x)^2 + \cdots \text{ even powers only.}$$

(19)

Here we have written a centered difference using the points $j + 2$ and $j - 2$ instead of $j + 1$ and $j - 1$, respectively. It should be clear that (19) is second-order accurate, although for any given value of $\Delta x$ the error of (19) is expected to be larger than the error of (14). We can combine (14) and (19) with a weight, $w$, so as to obtain a “hybrid” approximation to $\left(\frac{df}{dx}\right)_j$.
\[
\left( \frac{df}{dx} \right)_j = w \left( \frac{f_{j+1}-f_{j-1}}{2\Delta x} \right) + (1-w) \left( \frac{f_{j+2}-f_{j-2}}{4\Delta x} \right) \\
- \frac{w}{3!} \left( \frac{d^3 f}{dx^3} \right)_j (\Delta x)^2 - \frac{(1-w)}{3!} \left( \frac{d^3 f}{dx^3} \right)_j (2\Delta x)^2 + O\left[ (\Delta x)^4 \right].
\]

(20)

Inspection of (20) shows that we can force the coefficient of \((\Delta x)^2\) to vanish by choosing

\[ w + (1-w)4 = 0, \text{ or } w = 4/3. \]

(21)

With this choice, (20) reduces to

\[
\left( \frac{df}{dx} \right)_j = \frac{4}{3} \left( \frac{f_{j+1}-f_{j-1}}{2\Delta x} \right) - \frac{1}{3} \left( \frac{f_{j+2}-f_{j-2}}{4\Delta x} \right) + O\left[ (\Delta x)^4 \right].
\]

(22)

This is a fourth-order accurate scheme.

The derivation given above, in terms of a weighted combination of two second-order schemes, can be interpreted as a linear extrapolation of the value of the finite-difference
expression to a smaller grid size, as illustrated in Fig. 2.2. Both extrapolation and interpolation use weights that sum to one. In the case of interpolation, both weights lie between zero and one, while in the case of extrapolation one of the weights is larger than one and the other weight is negative. The concepts of extrapolation and interpolation will be discussed in more detail, later in this chapter.

Is there a more systematic way to construct schemes of any desired order of accuracy? The answer is “Yes,” and one such approach is as follows. Suppose that we write a finite-difference approximation to \( \frac{df}{dx} \) in the following somewhat generalized form:

\[
\frac{df}{dx} \approx \frac{1}{\Delta x} \sum_{k=-\infty}^{\infty} a_k f(x_j + k\Delta x).
\]

(23)

Here the \( a_k \) are coefficients or “weights,” which are undetermined at this point. We can design a scheme by choosing suitable expressions for the \( a_k \). In most schemes, all but a few of the \( a_k \) will
be zero, so that the sum (23) in will actually involve only a few (non-zero) terms. In writing (23), we have assumed for simplicity that $\Delta x$ is a constant; this assumption will be relaxed soon. The index $k$ in (23) is a counter that is zero at our “home base,” grid point $j$. For $k < 0$ we count to the left, and for $k > 0$ we count to the right. According to (23), our finite-difference approximation to $\left( \frac{df}{dx} \right)_j$ has the form of a weighted sum of values of $f(x)$ at various grid points in the vicinity of point $j$. Every finite-difference approximation that we have considered so far does indeed have this form, but you should be aware that there are (infinitely many!) schemes that do not have this form; a few of them will be discussed later.

Introducing a Taylor series expansion, we can write

$$f(x_j + k\Delta x) = f_j^0 + f_j^1 (k\Delta x) + f_j^2 \frac{(k\Delta x)^2}{2!} + f_j^3 \frac{(k\Delta x)^3}{3!} + \cdots$$

(24)

Here we introduce a new short-hand notation: $f_j^n$ is the $n$th derivative of $f$, evaluated at the point $j$. Using (24), we can rewrite (23) as

$$\left( \frac{df}{dx} \right)_{j, \text{approx}} \equiv \frac{1}{\Delta x} \sum_{k=-\infty}^{\infty} a_k \left[ f_j^0 + f_j^1 (k\Delta x) + f_j^2 \frac{(k\Delta x)^2}{2!} + f_j^3 \frac{(k\Delta x)^3}{3!} + \cdots \right].$$

(25)

By inspection of (25), we see that in order to have at least first-order accuracy, we need

$$\sum_{k=-\infty}^{\infty} a_k = 0 \quad \text{and} \quad \sum_{k=-\infty}^{\infty} ka_k = 1.$$  

(26)

To have at least second-order accuracy, we must impose an additional requirement:

$$\sum_{k=-\infty}^{\infty} k^2 a_k = 0.$$  

(27)

In general, to have at least $n$th-order accuracy, we must require that

$$\sum_{k=-\infty}^{\infty} k^m a_k = \delta_{m,1} \quad \text{for} \quad 0 \leq m \leq n.$$  

(28)
Here $\delta_{m,1}$ is the Kronecker delta. In order to satisfy (28), we must solve a system of $n+1$ linear equations for the $n+1$ unknown coefficients $a_k$.

According to (28), a scheme of $n$th-order accuracy can be constructed by satisfying $n+1$ equations. In particular, because (26) involves two equations, a first-order scheme has to involve at least two grid points, i.e., there must be at least two non-zero values of $a_k$. Pretty obvious, right? Note that we could make a first-order scheme that used fifty grid points if we wanted to -- but then, why would we want to? A second-order scheme must involve at least three grid points. A scheme that is parsimonious in its use of points is called “compact.”

Consider a simple example. Still assuming a uniform grid, a first order scheme for $f_j^1$ can be constructed using the points $j$ and $j+1$ as follows. From (26), we get $a_0 + a_1 = 0$ and $a_1 = 1$. Obviously we must choose $a_0 = -1$. Substituting into (23), we find that the scheme is given by $f_j^1 \equiv \frac{f(x_{j+1}) - f(x_j)}{\Delta x}$, i.e., the one-sided forward difference discussed earlier. In a similar way, we can also construct a first-order scheme using the points $j$ and $j-1$, with another one-sided result.

If we choose the points $j+1$ and $j-1$, imposing the requirements for first-order accuracy, i.e., (27), will actually give us the centered second-order scheme, i.e., $f_j^1 \equiv \frac{f(x_{j+1}) - f(x_{j-1})}{2\Delta x}$ because (27) is satisfied “accidentally” or “automatically” -- by luck, we manage to satisfy three equations using only two unknowns. If we choose the three points $j-1$, $j$, and $j+1$, and require second-order accuracy, we get exactly the same centered scheme, because $a_0$ turns out to be zero.

It should now be apparent that (28) can be used to construct schemes of arbitrarily high accuracy, simply by allocating enough grid points and solving the resulting system of linear equations for the $a_k$. Schemes that use a lot of grid points involve lots of arithmetic, so there is a “law of diminishing returns” at work. As a rough rule of thumb, it is not useful to go beyond about 5th order accuracy.

Next, we work out a generalization of the family of schemes given above, for the case of (possibly) non-uniform grid spacing. Eq. (23) is replaced by

$$\left( \frac{df}{dx} \right)_{j, \text{approx}} \equiv \sum_{k=\infty}^{\infty} b_{j,k} f(x_{j+k}).$$  

(29)
Note that, since $\Delta x$ is no longer a constant, the factor of $\frac{1}{\Delta x}$ that appears in (23) has been omitted in (29), and in view of this, in order to avoid notational confusion, we have replaced the symbol $a_k$ by $b_{j,k}$. Naturally, it is going to turn out that $b_k \sim \frac{1}{\Delta x}$. The subscript $j$ is included on $b_{j,k}$ because the grid can be non-uniform; as will be demonstrated below, for a given value of $k$ the coefficients are different for different values of $j$, i.e., at different places on the non-uniform grid. Similarly, (24) is replaced by

$$f(x_{j+k}) = f_j^0 + f_j^1(\delta x)_{j,k} + f_j^2(\delta x)_{j,k}^2 + f_j^3(\delta x)_{j,k}^3 + \ldots$$

(30)

Here $(\delta x)_{j,k} \equiv x_{j+k} - x_j$ takes the place of $k\Delta x$ in (24). Note that $(\delta x)_{j,0} = 0$, and $(\delta x)_{j,k} < 0$ for $k < 0$. To achieve second-order accuracy with (31), we must require, in addition to (32), that

$$\sum_{k=-\infty}^{\infty} b_{j,k} (\delta x)_{j,k}^2 = 0 \text{ for all } j .$$

(33)
Eq. (33) is the requirement that the first-order part of the error vanishes, so that we have at least second-order accuracy. In general, to have at least $n$th-order accuracy, we must require that

$$
\sum_{k=\infty}^{m} b_k (\delta x)_j^m = \delta_{m,1} \text{ for all } j, \text{ and for } 0 \leq m \leq n.
$$

(34)

As an example, consider the first-order accurate scheme using the points $j$ and $j + 1$. Since we are using only those two points, the only non-zero coefficients are $b_{j,0}$ and $b_{j,1}$, and they must satisfy the two equations corresponding to (32), i.e., $b_{j,0} + b_{j,1} = 0$ and $b_{j,1} = \frac{1}{(\delta x)_{j,1}}$. We see immediately that $b_{j,0} = -\frac{1}{(\delta x)_{j,1}}$. Referring back to (29), we see that the scheme is

$$
f_j^1 \equiv \frac{f(x_{j+1}) - f(x_j)}{(\delta x)_{j,1}},
$$

which, not unexpectedly, has the same form as the result that we obtained for the case of the uniform grid. Naturally, the non-uniformity of the grid is irrelevant when we consider only two points.

To obtain a second-order accurate approximation to $f_j^1$ on an arbitrary grid, using the three points, $j - 1$, $j$, and $j + 1$, we must require, from (32) that

$$
b_{j-1} + b_{j,0} + b_{j,1} = 0 \text{ and } b_{j-1}(\delta x)_{j-1} + b_{j,1}(\delta x)_{j,1} = 1,
$$

(35)

which suffice for first-order accuracy, and additionally from (34) that

$$
b_{j-1}(\delta x)_{j-1}^2 + b_{j,1}(\delta x)_{j,1}^2 = 0.
$$

(36)

Note that $(\delta x)_{j-1} = x_{j-1} - x_j = -(\delta x)_{j-1,1}$. The solution of this system of three linear equations can be written as

$$
b_{j-1} = \frac{(\delta x)_{j,1}}{(\delta x)_{j-1}[(\delta x)_{j,1} - (\delta x)_{j-1,1}]],
$$

(37)
\[ b_{j,0} = -\left( \frac{\delta x}{\delta x_j} \right) \left. \frac{\delta x}{\delta x_{j-1}} \right|_{j,1} \left( \frac{\delta x}{\delta x_j} \right) \left. \frac{\delta x}{\delta x_{j-1}} \right|_{j,-1} \], \quad (38) \\
\[ b_{j,1} = \frac{-\left( \frac{\delta x}{\delta x_j} \right)_{j,-1}}{\left( \frac{\delta x}{\delta x_j} \right)_{j,1} - \left( \frac{\delta x}{\delta x_j} \right)_{j,-1}} \]. \quad (39) 

You should confirm that for the case of uniform grid-spacing this reduces to the familiar centered second-order scheme.

Here is a simple and very practical question: Suppose that we use a scheme that has second-order accuracy on a uniform grid, but we go ahead and apply it on a non-uniform grid. What happens? As a concrete example, we use the scheme

\[ \left( \frac{df}{dx} \right)_{\text{approx}} \equiv \frac{f(x_{j+1}) - f(x_{j-1})}{x_{j+1} - x_{j-1}}. \quad (40) \]

By inspection, we have

\[ b_{j,-1} = -\frac{1}{x_{j+1} - x_{j-1}}, \quad (41) \]
\[ b_{j,0} = 0, \quad (42) \]
\[ b_{j,1} = \frac{1}{x_{j+1} - x_{j-1}}. \quad (43) \]

Eqs. (41) - (43) do satisfy both of the conditions in (35), so that the scheme does have first-order accuracy, even on the non-uniform grid. Eq. (37) - (39) are not satisfied, however, so it appears that second-order accuracy is lost.

This argument is a bit too hasty, however. Intuition suggests that, if the grid-spacing varies slowly enough, the scheme given by (40) should be nearly as accurate as if the grid-spacing were strictly constant. Intuition can never prove anything, but it can suggest ideas. Let’s pursue this idea to see if it has merit. Define \((\Delta x)_{j+1/2} \equiv x_{j+1} - x_j\) for all \(j\), and let \(\delta\) be the grid spacing at some reference grid point. We write the centered second-order scheme appropriate to a uniform grid, but apply it on a non-uniform grid:
Obviously, it is going to turn out that (possibly) non-uniform grid, we write provided that second-order. For example, if \( i.e., it is a \text{“} \text{difference of a difference}. \) If the grid spacing varies slowly enough, this term will be

There is indeed a \text{“}first-order term\text{”} in the error, as expected, but notice that it is proportional to

Here \((\Delta x)_{j+1/2} \equiv x_{j+1} - x_j\) and \((\Delta x)_{j-1/2} \equiv x_j - x_{j-1}\). Eq. (44) can be simplified to

There is indeed a \text{“}first-order term\text{”} in the error, as expected, but notice that it is proportional to

\[
\left[ (\Delta x)_{j+1/2} - (\Delta x)_{j-1/2} \right] = \delta(1 + \alpha x_{j+1/2}) - \delta(1 + \alpha x_{j-1/2}) = \delta \alpha (x_{j+1/2} - x_{j-1/2}) \sim O(\delta^2),
\]

provided that \( \alpha \sim O(1) \) or smaller.

Next, we observe that (29) can be generalized to derive approximations to higher-order derivatives of \( f \). For example, to derive approximations to the second derivative, \( f_j^2 \), on a (possibly) non-uniform grid, we write

Obviously, it is going to turn out that \( c_{j,k} \sim \frac{1}{(\Delta x)^2} \). Substitution of (30) into (47) gives

\[
\left( \frac{d^2 f}{dx^2} \right)_{j, \text{approx}} \equiv \sum_{k=-\infty}^{\infty} c_{j,k} f_{j+k}(x_{j+k}).
\]

\[
\left( \frac{d^2 f}{dx^2} \right)_{j, \text{approx}} = \sum_{k=-\infty}^{\infty} c_{j,k} \left[ f_j^0 + f_j^1(\delta x)_{j,k} + f_j^2(\delta x)_{j,k}^2 \frac{1}{2!} + f_j^3(\delta x)_{j,k}^3 \frac{1}{3!} + \ldots \right],
\]
A first-order accurate approximation to the second derivative is ensured if we enforce the three conditions
\[
\sum_{k=-\infty}^{\infty} c_{j,k} = 0, \quad \sum_{k=-\infty}^{\infty} c_{j,k} (\delta x)_{j,k} = 0, \quad \text{and} \quad \sum_{k=-\infty}^{\infty} c_{j,k} (\delta x)^2_{j,k} = 2!, \text{ for all } j.
\]

(49)

To achieve a second-order accurate approximation to the second derivative, we must also require that
\[
\sum_{k=-\infty}^{\infty} c_{j,k} (\delta x)^3_{j,k} = 0, \text{ for all } j.
\]

(50)

In general, to have an \(n\)th-order accurate approximation to the second derivative, we must require that
\[
\sum_{k=-\infty}^{\infty} c_{j,k} (\delta x)^m_{j,k} = (2!)^{\delta m,j} \text{ for all } j, \text{ and for } 0 \leq m \leq n + 1.
\]

(51)

We thus have to satisfy \(n + 2\) equations.

Earlier we showed that, in general, a second-order approximation to the first derivative must involve a minimum of three grid points, because three conditions must be satisfied [i.e., (35) and (36)]. Now we see that, in general, a second-order approximation to the second derivative must involve four grid points, because four conditions must be satisfied, i.e., (46) and (47). Five points may be preferable to four, from the point of view of symmetry. In the special case of a uniform grid, three points suffice.

At this point, you should be able to see (“by induction”) that on a (possibly) non-uniform grid, an \(n\)th-order accurate approximation to the \(l\)th derivative of \(f\) takes the form
\[
\left( \frac{d^lf}{dx^l} \right)_{j,\text{approx}} \equiv \sum_{k=-\infty}^{\infty} d_k f(x_{j+k}),
\]

(52)

where
\[
\sum_{k=-\infty}^{\infty} (\delta x)_{j,k}^m d_k = (l!)^{\delta m,j} \text{ for } 0 \leq m \leq n + l - 1.
\]

(53)
This is a total of \( n+l \) requirements, so in general a minimum of \( n+l \) points will be needed. It is straightforward to write a computer program that will automatically generate the coefficients for a compact \( n \)th-order-accurate approximation to the \( l \)th derivative of \( f \).

What happens for \( l = 0 \) ?

### 2.3 Extension to two dimensions

The approach presented above can be generalized to multi-dimensional problems. We will illustrate this using the two-dimensional Laplacian operator. The Laplacian appears, for example, in the diffusion equation with a constant diffusion coefficient, which is

\[
\frac{\partial f}{\partial t} = K \nabla^2 f,
\]

where \( t \) is time and \( K \) is a constant positive diffusion coefficient. A later chapter is devoted entirely to the diffusion equation.

Earlier in this chapter, we discussed one-dimensional differences that could represent either space or time differences, but our discussion of the Laplacian is unambiguously about space differencing.

Consider a fairly general finite-difference approximation to the Laplacian, of the form

\[
(\nabla^2 f)_{j,\text{approx}} \equiv \sum_{k=-\infty}^{\infty} e_{j,k} f(x_{j+k},y_{j+k}).
\]

Here we use one-dimensional indices even though we are on a two-dimensional grid. The grid is not necessarily rectangular, and can be non-uniform. The subscript \( j \) denotes a particular grid point ("home base" for this calculation), whose coordinates are \((x_j, y_j)\). Similarly, the subscript \( j+k \) denotes a grid point in the neighborhood of point \( j \), whose coordinates are \((x_{j+k}, y_{j+k})\). In practice, a method is needed to compute (and perhaps tabulate for later use) the appropriate values of \( k \) for the grid points in the neighborhood of each \( j \); for purposes of the present discussion this is an irrelevant detail.

The two-dimensional Taylor series is
$$f(x_{jk}, y_{jk}) = f(x_j, y_j) + \left[ (\delta x)_{jk} \frac{\partial}{\partial x} + (\delta y)_{jk} \frac{\partial}{\partial y} \right] f + \frac{1}{2!} \left[ (\delta x)_{jk} \frac{\partial}{\partial x} + (\delta y)_{jk} \frac{\partial}{\partial y} \right]^2 f + \frac{1}{3!} \left[ (\delta x)_{jk} \frac{\partial}{\partial x} + (\delta y)_{jk} \frac{\partial}{\partial y} \right]^3 f + \frac{1}{4!} \left[ (\delta x)_{jk} \frac{\partial}{\partial x} + (\delta y)_{jk} \frac{\partial}{\partial y} \right]^4 f + \cdots$$

which can be written out in gruesome detail as

$$f(x_{jk}, y_{jk}) = f(x_j, y_j) + \left[ (\delta x)_{jk} f_x + (\delta y)_{jk} f_y \right] + \frac{1}{2!} \left[ (\delta x)_{jk}^2 f_{xx} + 2(\delta x)_{jk} (\delta y)_{jk} f_{xy} + (\delta y)_{jk}^2 f_{yy} \right] + \frac{1}{3!} \left[ (\delta x)_{jk}^3 f_{xxx} + 3(\delta x)_{jk}^2 (\delta y)_{jk} f_{xxy} + 3(\delta x)_{jk} (\delta y)_{jk}^2 f_{xyy} + (\delta y)_{jk}^3 f_{yyy} \right] + \frac{1}{4!} \left[ (\delta x)_{jk}^4 f_{xxxx} + 4(\delta x)_{jk}^3 (\delta y)_{jk} f_{xxyy} + 6(\delta x)_{jk}^2 (\delta y)_{jk}^2 f_{xxyy} + 4(\delta x)_{jk} (\delta y)_{jk}^3 f_{xxyy} + (\delta y)_{jk}^4 f_{yyyy} \right] + \cdots$$

(56)

Here we use the notation

$$(\delta x)_{jk} \equiv x_{jk} - x_j \quad \text{and} \quad (\delta y)_{jk} \equiv y_{jk} - y_j ,$$

(57)

and it is understood that all of the derivatives are evaluated at the point $(x_j, y_j)$. Note the “cross terms,” that involve products of $(\delta x)_k$ and $(\delta y)_k$, and the corresponding cross-derivatives. A more general form of (56) is

$$f(r + a) = f(r) + \sum_{n=1}^{\infty} \frac{1}{n!} (a \cdot \nabla)^n f(r) ,$$

(58)

where $r$ is a position vector, and $a$ is a displacement vector (Arfken, 1985, p. 309). In (58), the operator $(a \cdot \nabla)^n$ acts on the function $f(r)$. You should confirm for yourself that the general form (58) is consistent with (56). Eq. (58) has the advantage that it does not make use of any particular coordinate system, but it can be used to work out the series expansion using any coordinate system, e.g., spherical coordinates or polar coordinates, by using the appropriate form of \( \nabla \).

Substituting from (56) into (55), we find that
\[(f_x + f_y)_{j,\text{approx}} \equiv \sum_{k=-\infty}^{\infty} e_{j,k} \left\{ f(x_j, y_j) + \left[ (\delta x)_{j,k} f_x + (\delta y)_{j,k} f_y \right] + \frac{1}{2!} \left[ (\delta x)_{j,k}^2 f_{xx} + 2(\delta x)_{j,k}(\delta y)_{j,k} f_{xy} + (\delta y)_{j,k}^2 f_{yy} \right] + \frac{1}{3!} \left[ (\delta x)_{j,k}^3 f_{xxx} + 3(\delta x)_{j,k}^2(\delta y)_{j,k} f_{xxy} + 3(\delta x)_{j,k}(\delta y)_{j,k}^2 f_{xyy} + (\delta y)_{j,k}^3 f_{yyy} \right] + \frac{1}{4!} \left[ (\delta x)_{j,k}^4 f_{xxxx} + 4(\delta x)_{j,k}^3(\delta y)_{j,k} f_{xxyy} + 6(\delta x)_{j,k}^2(\delta y)_{j,k}^2 f_{xxyy} + 4(\delta x)_{j,k}(\delta y)_{j,k}^3 f_{xyyy} + \cdots \right] \right\} \]

Notice that we have expressed the Laplacian on the left-hand side of (59) in terms of Cartesian coordinates. The motivation is that we are going to use the special case of Cartesian coordinates \((x, y)\) as an example, and in the process we are going to “match up terms” on the left and right sides of (59). Note, however, that the use of Cartesian coordinates in (59) does not limit its applicability to Cartesian grids. Eq. (59) can be used to analyze the truncation errors of a finite-difference Laplacian on any planar grid, regardless of how the grid points are distributed.

To have first-order accuracy, we need

\[\sum_{k=-\infty}^{\infty} e_{j,k} = 0, \text{ for all } j,\]

\[\sum_{k=-\infty}^{\infty} e_{j,k} (\delta x)_k = 0, \text{ for all } j,\]

\[\sum_{k=-\infty}^{\infty} e_{j,k} (\delta y)_k = 0, \text{ for all } j, \text{ and}\]

\[\sum_{k=-\infty}^{\infty} e_{j,k} (\delta x)_k^2 = 2!, \text{ for all } j,\]

\[\sum_{k=-\infty}^{\infty} e_{j,k} (\delta x)_k (\delta y)_k = 0, \text{ for all } j,\]
\[
\sum_{k=\infty}^{\infty} e_{j,k} (\delta y)_{k}^2 = 2!, \text{ for all } j .
\]

(65)

From (55) and (59), it is clear that \( e_{j,k} \) is of order \( \delta^{-2} \), where \( \delta \) denotes \( \delta x \) or \( \delta y \). Therefore, the quantities inside the sums in (61-62) are of order \( \delta^{-1} \), and those inside the sums in (63-65) are of order one. This is why (63-65) are required, in addition to (60-62), to obtain first-order accuracy.

Eq. (60) implies [with (55)] that a constant field has a Laplacian of zero, as it should. That’s nice.

So far, (60-65) involve six equations. This means that to ensure first-order accuracy for a two-dimensional grid, six grid points are needed for the general case. There are exceptions to this. If we are fortunate enough to be working on a highly symmetrical grid, it is possible that the conditions for second-order accuracy can be satisfied with a smaller number of points. For example, if we satisfy (60-65) on a square grid, we will get second-order accuracy “for free,” and, as you will show when you do the homework, it can be done with only five points. More generally, with a non-uniform grid, we need the following four additional conditions to achieve second-order accuracy:

\[
\sum_{k=\infty}^{\infty} e_{j,k} (\delta x)_{k}^3 = 0 , \text{ for all } j ,
\]

(66)

\[
\sum_{k=\infty}^{\infty} e_{j,k} (\delta x)_{k}^2 (\delta y)_{k} = 0 , \text{ for all } j ,
\]

(67)

\[
\sum_{k=\infty}^{\infty} e_{j,k} (\delta x)_{k} (\delta y)_{k}^2 = 0 , \text{ for all } j ,
\]

(68)

\[
\sum_{k=\infty}^{\infty} e_{j,k} (\delta y)_{k}^3 = 0 , \text{ for all } j .
\]

(69)

In general, a total of ten conditions, i.e., Eqs. (60-69), must be satisfied to ensure second-order accuracy on a non-uniform two-dimensional grid.

If we were working in more than two dimensions, we would simply replace (56) by the appropriate multi-dimensional Taylor series expansion. The rest of the argument would be
parallel to that given above, although of course the requirements for second-order accuracy would be more numerous.

### 2.4 Integral properties of the Laplacian

Here comes a digression, which deals with something other than the order of accuracy of the scheme.

For the continuous Laplacian on a closed or periodic domain, we can prove the following two important properties:

\[
\int_A (\nabla^2 f) dA = 0, \tag{70}
\]

\[
\int_A f (\nabla^2 f) dA \leq 0. \tag{71}
\]

Here the integrals are with respect to area, over the entire domain. These results hold for any sufficiently differentiable function \( f \). For the diffusion equation, (54), Eq. (70) implies that diffusion does not change the area-averaged value of \( f \), and the inequality (71) implies that diffusion reduces the area average of the square of \( f \).

The finite-difference requirements corresponding to (70) and (71) are

\[
\sum_{all \, j} (\nabla^2 f)_{j, \text{approx}} A_j \equiv \sum_{all \, j} \left[ \sum_{k=-\infty}^{\infty} e_{j,k} f(x_{j+k}, y_{j+k}) \right] A_j = 0, \tag{72}
\]

\[
\sum_{all \, j} f_j (\nabla^2 f)_{j, \text{approx}} A_j \equiv \sum_{all \, j} f_j \left[ \sum_{k=-\infty}^{\infty} e_{j,k} f(x_{j+k}, y_{j+k}) \right] A_j \leq 0, \tag{73}
\]

where \( A_j \) is the area of grid-cell \( j \). Suppose that we want to satisfy (72-73) regardless of the numerical values assigned to \( f(x_j, y_j) \). This may sound impossible, but it actually can be done by suitable choice of the \( e_{j,k} \). As an example, consider what is needed to ensure that (72) will be satisfied for an arbitrary distribution of \( f(x_j, y_j) \). Each value of \( f(x_j, y_j) \) will appear more than once in the sum on the right-hand side of (72). We can “collect the coefficients” of each value of \( f(x_j, y_j) \), and require that the sum of the coefficients is zero. To see how this works, define \( j' \equiv j + k \), and write
\[
\sum_{j, k} \sum_{i, j, k = -\infty}^{\infty} e_{j, k} f(x_{j+k}, y_{j+k}) A_j = \sum_{j, j'} \sum_{j, j'} e_{j, j'} f(x_{j'}, y_{j'}) A_j \\
= \sum_{j'} \sum_{j} e_{j, j'} f(x_{j'}, y_{j'}) A_j \\
= \sum_{j'} \left( f(x_{j'}, y_{j'}) \left( \sum_{j} e_{j, j'} A_j \right) \right) \\
= 0 .
\]

(74)

In the second line above, we simply change the order of summation. The last equality above is a re-statement of the requirement (72). The only way to satisfy it for an arbitrary distribution of \( f(x_j, y_j) \) is to write

\[
\sum_{j} e_{j, j'} A_j = 0 \text{ for each } j',
\]

(75)

which is equivalent to

\[
\sum_{j} e_{j, k} A_j = 0 \text{ for each } k .
\]

(76)

If the \( e_{j, k} \) satisfy (76), then (72) will also be satisfied.

Similar (but more complicated) ideas were used by Arakawa (1966), in the context of energy and enstrophy conservation with a finite-difference vorticity equation. This will be discussed in Chapter 11.

2.5 Laplacians on rectangular grids

Consider a nine-points on a rectangular grid, as shown in Fig. 2.3. Because this grid has a high degree of symmetry, is possible to obtain second-order accuracy with just five points, and in fact this can be done in two different ways, corresponding to the two five-point stencils shown
by the grey boxes in the figure. Based on their shapes, one of the stencils can be called “+”, and the other one “x”. We assume a grid spacing $d$ in both the $x$ and $y$ directions, and use a two-dimensional indexing system, with counters $i$ and $j$ in the $x$ and $y$ directions, respectively. Using the methods explained above, it can be shown that the second-order finite-difference Laplacians are given by

$$
(\nabla^2 f)_{i,j,\text{approx}} \equiv \frac{f_{i,j+1} + f_{i,j-1} + f_{i+1,j} + f_{i-1,j} - 4f_{i,j}}{d^2} \quad \text{with the + stencil},
$$

(77)

and

$$
(\nabla^2 f)_{i,j,\text{approx}} \equiv \frac{f_{i+1,j+1} + f_{i-1,j+1} + f_{i-1,j-1} + f_{i+1,j-1} - 4f_{i,j}}{(\sqrt{2}d)^2} \quad \text{with the x stencil}.
$$

(78)

The Laplacian based on the x stencil cannot “see” a checkerboard pattern in the function represented on the grid, as shown by the plus and minus symbols in the top two panels. What I mean by this is that the scheme returns a zero for the Laplacian even when a checkerboard is present. In contrast, the + stencil feels the checkerboard. The checkerboard solution can be called a “computational mode” because it is a solution of a finite-difference approximation to $\nabla^2 f = 0$. 

Figure 2.3: The grey shading shows two five-point stencils that can be used to create second-order Laplacians on a rectangular grid. In the upper two panels, the plus and minus symbols represent an input function that has the form of a checkerboard. In the lower two panels the plus and minus symbols represent an input function that has the form of a “rotated checkerboard,” which can also be viewed as a set of horizontal stripes. (The checkerboard in the top two panels can be viewed as “rotated stripes.”)
but it does not correspond to any true solution of the differential equation. Computational modes are bad. This is our first encounter with them, but they will be discussed in much more detail later.

The + stencil is not necessarily the clear winner here. It under-estimates the strength of the “rotated checkerboard” (which can also be called “diagonal stripes”) shown in the bottom two panels of Fig. 2.3, while the x stencil feels it more strongly. The best second-order Laplacian uses all nine points, and can be obtained by writing a weighted sum of the two Laplacians given by (77) and 78). What principle would you suggest for choosing the values of the weights? The nine-point stencil will involve more arithmetic than either of the two five-point stencils, but the benefit may justify the cost in this case.

2.6 Why be square?

As is well known, only three regular polygons tile the plane: equilateral triangles, squares, and hexagons. Fig. 2.4 shows planar grids made from each of these three possible polygonal elements.

On the triangular grid and the square grid, some of the neighbors of a given cell lie directly across cell walls, while others lie across cell vertices. As a result, finite-difference operators constructed on these grids tend to use “wall neighbors” and “vertex neighbors” in different ways. For example, the simplest second-order finite-difference approximation to the gradient, on a square grid, uses only “wall neighbors;” vertex neighbors are ignored. Although it is certainly possible to construct finite-difference operators on square grids (and triangular grids) in which information from all neighboring cells is used [e.g., the Arakawa Jacobian, as discussed by Arakawa (1966) and in Chapter 11], the essential anisotropies of these grids remain, and are unavoidably manifested in the forms of the finite-difference operators.

In contrast, hexagonal grids have the property that all neighbors of a given cell lie across cell walls; there are no “vertex neighbors.” In this sense, hexagonal grids are quasi-isotropic. As a result, the most natural finite-difference Laplacians on hexagonal grids treat all neighboring cells in the same way; they are as symmetrical and isotropic as possible.

Further discussion of non-quadrilateral grids will be given later in the course.
Summary

It is straightforward to design finite-difference schemes to approximate a derivative of any order, with any desired order of accuracy, on irregular grids of any shape, and in multiple dimensions. The schemes can also be designed to satisfy rules based on properties of the differential operators that they approximate; for example, we can force the area-average of a two-dimensional finite-difference Laplacian to vanish on a periodic domain. Some finite-difference schemes suffer from an inability to recognize small-scale, “noisy” computational modes on the grid, such as checkerboard patterns.

Figure 2.4: The upper row of figures shows small sections of grids made up of equilateral triangles (left), squares (center), and hexagons (right). These are the only regular polygons that tile the plane. The hexagonal grid has the highest symmetry. For example, all neighboring cells of a given hexagonal cell are located across cell walls. In contrast, with either triangles or squares some neighbors are across walls, while others are across corners. The lower row of figures shows the “checkerboards” associated with each grid. The triangular and quadrilateral checkerboards have two colors, while the hexagonal checkerboard has three colors.

2.7 Summary

It is straightforward to design finite-difference schemes to approximate a derivative of any order, with any desired order of accuracy, on irregular grids of any shape, and in multiple dimensions. The schemes can also be designed to satisfy rules based on properties of the differential operators that they approximate; for example, we can force the area-average of a two-dimensional finite-difference Laplacian to vanish on a periodic domain. Some finite-difference schemes suffer from an inability to recognize small-scale, “noisy” computational modes on the grid, such as checkerboard patterns.
Problems

1. Prove that a finite-difference scheme with errors of order $n$ gives exact derivatives for polynomial functions of degree $n$ or less. For example, a first-order scheme gives exact derivatives for linear functions.

2. Choose a simple differentiable function $f(x)$ that is not a polynomial. Find the exact numerical value of $\frac{df}{dx}$ at a particular value of $x$, say $x_i$. Then choose

   a) a first-order scheme, and

   b) a second-order scheme to approximate $\left(\frac{df}{dx}\right)_{x=x_i}$.

   For each case, plot the log of the absolute value of the total error of these approximations as a function of the log of $\Delta x$. You can find the total error by subtracting the approximate derivative from the exact derivative. By inspection of the plot, verify that the errors of the schemes decrease, with the expected slopes (which you should estimate from the plots), as $\Delta x$ decreases.

3. Prove that the only regular convex polygons that tile the plane are the equilateral triangle, the square, and the equilateral hexagon.

4. For this problem, use equations (60)-(69), which state the requirements for second-order accuracy of the Laplacian.

   To standardize the notation, in all cases let $d$ be the distance between grid-cell centers, as measured across cell walls. Write each of your solutions in terms of $d$, as was done in (76) and (77).

   a) Consider a plane tiled with perfectly hexagonal cells. The dependent variable is defined at the center of each hexagon. Find a second-order accurate scheme for the Laplacian that uses just the central cell and its six closest neighbors, i.e., just seven cells in total. Make a sketch to show explain your notation for distances in the $x$ and $y$ directions. *Hint:* You can drastically simplify the problem by taking advantage of the high degree of symmetry.

   b) Repeat for a plane tiled with equilateral triangles. You will have to figure out which and how many cells to use, in addition to the central cell. Use the centroids of the triangles to assign their positions. Make a sketch to explain your notation for distances in the $x$ and $y$ directions. *Hint:* You can drastically simplify the problem by taking advantage of the high degree of symmetry.
c) Repeat for a plane tiled with squares, using the central cell and the neighboring cells across the cell walls. Make a sketch to show explain your notation for distances in the \( x \) and \( y \) directions. \textit{Hint:} You can drastically simplify the problem by taking advantage of the high degree of symmetry.

d) Repeat for a plane tiled with squares, using the central cell and the neighboring cells that lie diagonally across cell corners. Make a sketch to show explain your notation for distances in the \( x \) and \( y \) directions. \textit{Hint:} You can drastically simplify the problem by taking advantage of the high degree of symmetry.

e) For each of the four cases discussed above, and \textit{assuming a doubly periodic domain on a plane}, discuss whether it is possible to have \( \text{Laplacian} = 0 \) throughout the entire domain even when the input field is not constant. Fig. 2.4 may help you to figure out the answer to this question.

5. a) Repeat the analysis of Section 2.3, up to Eq. (69), using cylindrical (polar) coordinates.

b) Use the results of part a to find a second-order accurate scheme for the Laplacian on a hexagonal grid. Compare with the scheme that you found when you did part a) of problem 4 above. Are the two schemes mathematically equivalent?

6. a) Invent a way to “index” the points on a hexagonal grid with periodic boundary conditions. As a starting point, I suggest that you make an integer array like this: \texttt{NEIGHBORS(I,J)}. The first subscript would designate which point on the grid is “home base,” i.e., it would be a one-dimensional counter that covers the entire grid. The second subscript would range from 0 to 6 (or, alternatively, from 1 to 7). The smallest value (0 or 1) would designate the central point, and the remaining 6 would designate its six neighbors. The indexing problem then reduces to generating the array \texttt{NEIGHBORS(I,J)}, which need only be done once for a given grid.

b) Set up a hexagonal grid to represent a square domain with periodic boundary conditions. It is not possible to create an \textit{exactly} square domain using a hexagonal grid. Therefore you have to set up a domain that is \textit{approximately} square, with 100 points in one direction and the appropriate number of points (you get to figure it out) in the other direction. The total number of points in the approximately square domain will be very roughly 8000. This (approximately) square domain has periodic boundary conditions, so it actually represents one “patch” of an infinite domain. The period in the \( x \)-direction cannot be exactly the same as the period in the \( y \)-direction, but it can be pretty close. Make sure that the boundary conditions are truly periodic on your grid, so that no discontinuities occur. Fig. 2.5 can help you to understand how to define the computational domain and how to implement the periodic boundary conditions.
c) Using the tools created under parts a) and b) above, write a program to compute the Laplacian of a given doubly periodic function.

d) Empirically demonstrate the order of accuracy of your program using an analytical test function, for which you can compute the Laplacian exactly. To do this, you have to define a useful measure of the overall error; justify your definition.

7. Consider a scheme for the Laplacian on a square grid that is given by a weighted sum of (77) and (78). Propose a rationale for choosing the weight. Explain the motivation for your proposal. There can be more than one good answer to this question.

8. The Laplacian can be written using Cartesian coordinates $(x,y)$ as $\nabla^2 f = f_{xx} + f_{yy}$. Consider a second Cartesian coordinate system, $(x',y')$, that is rotated, with respect to the
first, by an angle $\theta$. Prove by direct calculation that $f_{xx} + f_{yy} = f_{xx'} + f_{yy'}$. This demonstrates that the Laplacian is invariant with respect to rotations of a Cartesian coordinate system. In fact, it can be demonstrated that the Laplacian takes the same numerical value no matter what coordinate system is used.