Week 11: Wednesday, Apr 11

Truncation versus rounding

Last week, we discussed two different ways to derive the centered difference approximation to the first derivative

\[ f'(x) \approx f[x + h, x - h] = \frac{f(x + h) - f(x - h)}{2h}. \]

Using Taylor series, we were also able to write down an estimate of the truncation error:

\[ f[x + h, x - h] - f'(x) = \frac{h^2}{6} f'''(x) + O(h^4). \]

As \( h \) grows smaller and smaller, \( f[x + h, x - h] \) becomes a better and better approximation to \( f'(x) \) — at least, it does in exact arithmetic. If we plot the truncation error \( |h^2/6 f'''(x)| \) against \( h \) on a log-log scale, we expect to see a nice straight line with slope 2. But Figure 1 shows that something rather different happens in floating point. Try it for yourself!

The problem, of course, is cancellation. As \( h \) goes to zero, \( f(x + h) \) and \( f(x - h) \) get close together; and for \( h \) small enough, the computed value of \( f(x + h) - f(x - h) \) starts to be dominated by rounding error. If the values of \( f(x + h) \) and \( f(x - h) \) are computed in floating point as \( f(x + h)(1 + \delta_1) \) and \( f(x - h)(1 + \delta_2) \), then the computed finite difference is approximately

\[ \hat{f}[h, -h] = f[h, -h] + \frac{\delta_1 f(x + h) - \delta_2 f(x - h)}{2h}, \]

and if we manage to get the values of \( f(x+h) \) and \( f(x-h) \) correctly rounded, we have

\[ \left| \frac{\delta_1 f(x + h) - \delta_2 f(x - h)}{2h} \right| \leq \frac{\epsilon_{\text{mach}}}{h} \left( \max_{x - h \leq \xi \leq x + h} |f(\xi)| \right) \approx \frac{\epsilon_{\text{mach}}}{h} f(x). \]

The total error in approximating \( f'(x) \) by \( f[x + h, x - h] \) in floating point therefore consists of two pieces: truncation error proportional to \( h^2 \), and rounding error proportional to \( \epsilon_{\text{mach}}/h \). The total error is minimized when these two effects are approximately equal, at

\[ h \approx \left( \frac{6f(x)}{f'''(x)\epsilon_{\text{mach}}} \right)^{1/3}, \]
i.e. when $h$ is close to $\epsilon_{\text{mach}}^{1/3}$. From the plot in Figure 1, we can see that this is right — the minimum observed error occurs for $h$ pretty close to $\epsilon_{\text{mach}}^{1/3}$ (around $10^{-5}$).

Of course, the analysis in the previous paragraph assumed the happy circumstance that we could get our hands on the correctly rounded values of $f(x + h)$ and $f(x - h)$. In general, we might have a little more error inherited from the evaluation of $f$ itself, which would just make the optimal $h$ (and the corresponding optimal accuracy) that much larger.

**Richardson extrapolation**

Let’s put aside our concerns about rounding error for a moment, and just look at the truncation error in the centered difference approximation of $f'(x)$. We have an estimate of the form

$$f[x + h, x - h] - f'(x) = \frac{h^2}{6} f'''(x) + O(h^4).$$

Usually we don’t get to write down such a sharp estimate for the error. There is a good reason for this: if we have a very sharp error estimate, we can use the estimate to reduce the error! The general trick is this: if we have $g_h(x) \approx g(x)$ with an error expansion of the form

$$g_h(x) = g(x) + Ch^p + O(h^{p+1}),$$

then we can write

$$a g_h(x) + b g_{2h}(x) = (a + b) g(x) + C(a + 2^p b) h^p + O(h^{p+1}).$$

Now find coefficients $a$ and $b$ so that

$$a + b = 1$$

$$a + 2^p b = 0;$$

the solution to this system is

$$a = \frac{2^p}{2^p - 1}, \quad b = -\frac{1}{2^p - 1}.$$

Therefore, we have

$$\frac{2^p g_h(x) - g_{2h}(x)}{2^p - 1} = g(x) + O(h^{p+1});$$
Figure 1: Actual error and estimated truncation error for a centered difference approximation to $\frac{d}{dx} \sin(x)$ at $x = 1$. For small $h$ values, the error is dominated by roundoff rather than by truncation error.

```matlab
% Compute actual error and estimated truncation error
% for a centered difference approximation to sin'(x)
% at x = 1.
%
h = 2.^-(1:50);
f = (sin(1+h)-sin(1-h))/h/2;
err = fd-cos(1);
errest = -h.^2/6 * cos(1);
%
% Plot the actual error and estimated truncation error
% versus h on a log–log scale.
%
loglog(h, abs(err), h, abs(errest));
legend('Error', 'Error estimate');
xlabel('h');
```

Figure 2: Code to produce Figure 1.
that is, we have cancelled off the leading term in the error.

In the case of the centered difference formula, only even powers of $h$ appear in the series expansion of the error; so we actually have that

$$\frac{4f[x+h,x-h] - f[x+2h,x-2h]}{3} = f'(0) + O(h^4).$$

An advantage of the higher order of accuracy is that we can get very small truncation errors even when $h$ is not very small, and so we tend to be able to reach a better optimal error before cancellation effects start to dominate; see Figure 3.
Problems to ponder

1. Suppose that $f(x)$ is smooth and has a single local maximum between $[h,-h]$, and let $p_h(x)$ denote the quadratic interpolant through $0$, $h$, and $-h$. Argue that if the second derivative of $f$ is bounded away from zero near $0$, then the actual maximizing point $x_*$ for $f$ satisfies

$$x_* = -\frac{p'_h(0)}{p''_h(0)} + O(h^2).$$

2. Suppose we know $f(x)$, $f(x+h)$, and $f(x+2h)$. Both by interpolation and by manipulation of Taylor series, find a formula to estimate $f'(x)$ of the form $c_0 f(x) + c_1 f(x+h) + c_2 f(x+2h)$. Using Taylor expansions about $x$, also estimate the truncation error.

3. Consider the one-sided finite difference approximation

$$f'(x) \approx f[x + h, x] = \frac{f(x+h) - f(x)}{h}.$$  

(a) Show using Taylor series that

$$f[0,h] - f'(0) = \frac{1}{2} f''(0)h + O(h^2).$$

(b) Apply Richardson extrapolation to this approximation.

4. Verify that the extrapolated centered difference approximation to $f'(x)$ is the same as the approximation derived by differentiating the quartic that passes through $f$ at $\{x - 2h, x - h, x, x + h, x + 2h\}$.

5. Richardson extrapolation is just one example of an acceleration technique that can turn a slowly-convergent sequence of estimates into something that converges more quickly. We can use the same idea in other cases. For example, suppose we believe a one-dimensional iteration $x_{k+1} = g(x_k)$ converges linearly to a fixed point $x_*$. Then

(a) Suppose the rate constant $C = g'(x_*)$ is known. Using

$$e_{k+1} = Ce_k + O(e_k^2),$$

show that

$$\frac{x_{k+1} - Cx_k}{1 - C} = x_* + O(e_k^2)$$
(b) Show that the rate constant $g'(x_*)$ can be estimated by

$$C_k \equiv \frac{x_{k+2} - x_{k+1}}{x_{k+1} - x_k} \to g'(x_*)$$

(c) If you are bored and feel like doing algebra, show that

$$y_k \equiv \frac{x_{k+1} - C_k x_k}{1 - C_k} = \frac{x_k x_{k+2} - x_{k+1}^2}{x_{k+2} - 2x_{k+1} + x_k},$$

and using the techniques developed in the first two parts, that

$$y_k - x_* = O((x_k - x_*)^2).$$

The transformation from the sequence $x_k$ into the (more rapidly convergent) sequence $y_k$ is sometimes known as Aitken’s delta-squared process. The process can sometimes be applied repeatedly. You may find it entertaining to try running this transformation repeatedly on the partial sums of the alternating harmonic series

$$S_n = \sum_{j=1}^{n} \frac{(-1)^{j+1}}{j},$$

which converges very slowly to $\ln(2)$. Without any transformation, $S_{20}$ has an error of greater than $10^{-2}$; one step of transformation reduces that to nearly $10^{-5}$; and with three steps, one is below $10^{-7}$. 