Numerical Analysis

You could say that some of the equations that you encounter in describing physical systems can't be solved in terms of familiar functions and that they require numerical calculations to solve. It would be misleading to say this however, because the reality is quite the opposite. *Most* of the equations that describe the real world are sufficiently complex that your only hope of solving them is to use numerical methods. The simple equations that you find in introductory texts are there because they *can* be solved in terms of elementary calculations. When you start to add reality, you quickly reach a point at which no amount of clever analytical ability will get you a solution. That becomes the subject of this chapter. In all of the examples that I present I'm just showing you a taste of the subject, but I hope that you will see the essential ideas of how to extend the concepts.

11.1 Interpolation

Given equally spaced tabulated data, the problem is to find a value between the tabulated points, and to estimate the error in doing so. As a first example, to find a value midway between given points use a linear interpolation:

$$f(x_0+h/2)\approx\frac{1}{2}\big[f(x_0)+f(x_0+h)\big].$$

This gives no hint of the error. To compute an error estimate, it is convenient to transform the variables so that this equation reads

$$f(0)\approx\frac{1}{2}\big[f(k)+f(-k)\big],$$

where the interval between data points is now $2k$. Use a power series expansion of f to find the error.

$$\begin{aligned}f(k)&=f(0)+kf'(0)+\frac{1}{2}k^2f''(0)+\cdots\\ f(-k)&=f(0)-kf'(0)+\frac{1}{2}k^2f''(0)+\cdots\end{aligned}$$

Then

$$\frac{1}{2}\big[f(k)+f(-k)\big]\approx f(0)+\big[\frac{1}{2}k^2f''(0)\big],\tag{1}$$

where the last term gives an estimate of the error:

$$\text{error}=\text{estimate}-\text{exact}=+k^2f''(0)/2=+h^2f''(0)/8$$

As an example, interpolate the function $f(x)=2^x$ between 0 and 1. Here $h=1$.

$$2^{1/2}\approx\frac{1}{2}\big[2^0+2^1\big]=1.5$$

The error term is

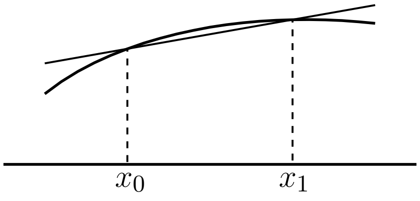
$$\begin{aligned}\text{error}&\approx(\ln 2)^2 2^x/8\qquad\text{for}\quad x=.5\\ &=(.693)^2(1.5)/8=.090,\end{aligned}$$

and of course the true error is $1.5 - 1.414 = .086$

You can write a more general interpolation method for an arbitrary point between x_0 and $x_0 + h$. The solution is a simple extension of the above result.

The line passing through the two points of the graph is

$$y - f_0 = (x - x_0)(f_1 - f_0)/h,$$



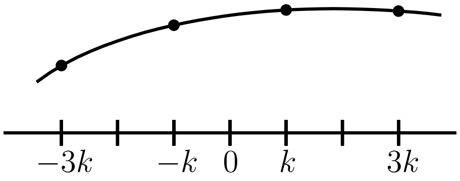
$$f_0 = f(x_0), \qquad f_1 = f(x_0 + h).$$

At the point $x = x_0 + ph$ you have

$$y = f_0 + (ph)(f_1 - f_0)/h = f_0(1 - p) + f_1 p.$$

As before, this approach doesn't suggest the error, but again, the Taylor series allows you to work it out to be $\big[h^2 p(1-p) f''(x_0 + ph)/2\big]$.

The use of only two points to do an interpolation ignores the data available in the rest of the table. By using more points, you can greatly improve the accuracy. The simplest example of this method is the 4-point interpolation to find the function halfway between the data points. Again, the independent variable has an increment $h = 2k$, so the problem can be stated as one of finding the value of $f(0)$ given $f(\pm k)$ and $f(\pm 3k)$.



$$f(k) = f(0) + kf'(0) + \frac{1}{2}k^2 f''(0) + \frac{1}{6}k^3 f'''(0) + \cdots. \tag{2}$$

I want to isolate $f(0)$ from this, so take

$$\begin{aligned} f(k) + f(-k) &= 2f(0) + k^2 f''(0) + \frac{1}{12}k^4 f''''(0) + \cdots \\ f(3k) + f(-3k) &= 2f(0) + 9k^2 f''(0) + \frac{81}{12}k^4 f''''(0) + \cdots. \end{aligned}$$

The biggest term after the $f(0)$ is in $k^2 f''(0)$, so I'll eliminate this.

$$\big[f(3k) + f(-3k)\big] - 9\big[f(k) - f(-k)\big] \approx -16f(0) + \left[\frac{81}{12} - \frac{9}{12}\right] k^4 f''''(0)$$

$$f(0) \approx \frac{1}{16}\big[-f(-3k) + 9f(-k) + 9f(k) - f(3k)\big] - \big[-\frac{3}{8}k^4 f''''(0)\big]. \tag{3}$$

The error estimate is then $-3h^4 f''''(0)/128$.

To apply this, take the same example as before, $f(x) = 2^x$ at $x = .5$

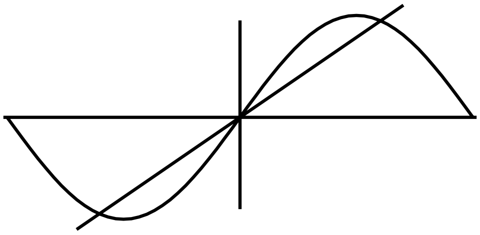
$$2^{1/2} \approx \frac{1}{16}\left[-2^{-1} + 9 \cdot 2^0 + 9 \cdot 2^1 - 2^2\right] = \frac{45}{32} = 1.40625,$$

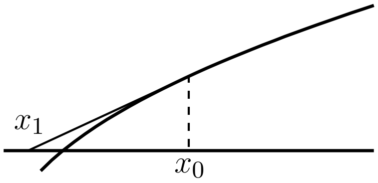
and the error is $1.40625 - 1.41421 = -.008$, a tenfold improvement over the previous interpolation despite the fact that the function changes markedly in this interval and you shouldn't expect interpolation to work very well here.

11.2 Solving equations

Example: $\sin x - x/2 = 0$

From the first graph, the equation clearly has three real solutions, but finding them is the problem. The first method for solving $f(x) = 0$ is* Newton's method. The basic idea is that over a small enough region, *everything* is more or less linear. This isn't true of course, so don't be surprised that this method doesn't always work.





A general picture of a function with a root is the second graph. In that case, observe that if x_0 is a first approximation to the root of f, the straight line tangent to the curve can be used to calculate an improved approximation. The equation of this line is

$$y - f(x_0) = f'(x_0)(x - x_0).$$

The root of this line is $y = 0$, with solution

$$x = x_0 - f(x_0)/f'(x_0).$$

Call this solution x_1. You can use this in an iterative procedure to find

$$x_2 = x_1 - f(x_1)/f'(x_1), \tag{4}$$

and in turn x_3 is defined in terms of x_2 etc.

Example: Solve $\sin x - x/2 = 0$. From the graph, a plausible guess for a root is $x_0 = 2$.

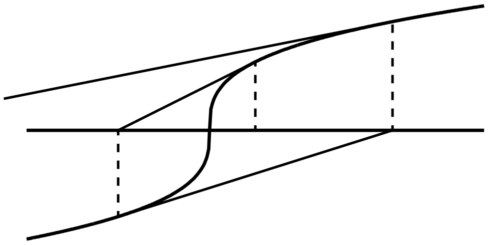
$$\begin{aligned}
x_1 &= x_0 - (\sin x_0 - x_0/2)/(\cos x_0 - 1/2) \\
&= 1.900995594 & f(x_1) &= .00452 \\
x_2 &= x_1 - (\sin x_1 - x_1/2)/(\cos x_1 - 1/2) \\
&= 1.895511645 & f(x_2) &= -.000014 \\
x_3 &= x_2 - (\sin x_2 - x_2/2)/(\cos x_2 - 1/2) \\
&= 1.895494267 & f(x_3) &= 2 \times 10^{-10}
\end{aligned}$$

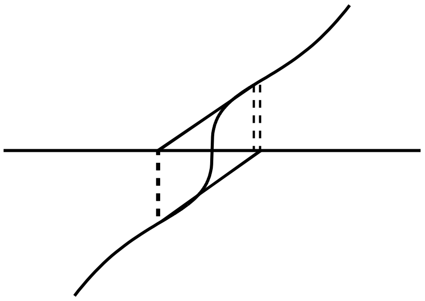
* Though attributed to Newton and Raphson, it was probably discovered by Simpson.

Such iterative procedures are ideal for use on a computer, but use them with caution, as a simple example shows:

$$f(x) = x^{1/3}.$$

Instead of the root $x = 0$, the iterations in this first graph carry the supposed solution infinitely far away. This happens here because the higher derivatives neglected in the straight line approximation are large near the root.





A milder form of non-convergence can occur if at the root the curvature changes sign and is large, as in the second graph. This can lead to a limit cycle where the iteration simply oscillates from one side of the root to the other without going anywhere. As I said earlier this doesn't *always* work.

A non-graphical derivation of this method starts from a Taylor series: If z_0 is an approximate root and $z_0 + \epsilon$ is a presumed exact root, then

$$f(z_0 + \epsilon) = 0 = f(z_0) + \epsilon f'(z_0) + \cdots.$$

Neglecting higher terms then,

$$\epsilon = -f(z_0)/f'(z_0), \qquad \text{and} \qquad z_1 = z_0 + \epsilon = z_0 - f(z_0)/f'(z_0), \tag{5}$$

as before. I use z instead of x this time to remind you that this method is just as valid for complex functions as for real ones (and has as many pitfalls).

There is a simple variation on this method that can be used to speed convergence where it is poor or to bring about convergence where the technique would otherwise break down.

$$x_1 = x_0 - wf(x_0)/f'(x_0). \tag{6}$$

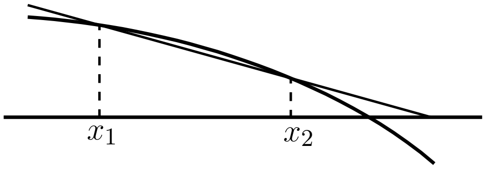
W is a factor that can be chosen greater than one to increase the correction or less than one to decrease it. Which one to do is more an art than a science (1.5 and 0.5 are common choices). You can easily verify that any choice of w between 0 and 2/3 will cause convergence for the solution of $x^{1/3} = 0$. You can also try this method on the solution of $f(x) = x^2 = 0$. A straight-forward iteration will certainly converge, but with painful slowness. The choice of $w > 1$ improves this considerably.

When Newton's method works well, it will typically double the number of significant figures at each iteration.

A drawback to this method is that it requires knowledge of $f'(x)$, and that may not be simple. An alternate approach that avoids this starts from the picture in which a secant through the curve is used in place of a tangent at a point.

Given $f(x_1)$ and $f(x_2)$, construct a straight line

$$y - f(x_2) = \left[\frac{f(x_2) - f(x_1)}{x_2 - x_1}\right](x - x_2).$$



This has its root at $y = 0$, or

$$x = x_2 - f(x_2)\frac{x_2 - x_1}{f(x_2) - f(x_1)}. \tag{7}$$

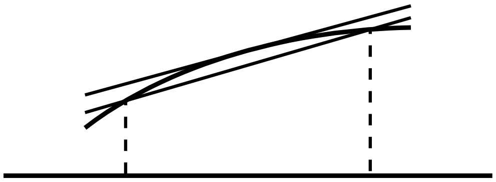
This root is taken as x_3 and the method is iterated, substituting x_2 and x_3 for x_1 and x_2. As with Newton's method, when it works, it works very well, but you must look out for the same type of non-convergence problems. This is called the secant method.

11.3 Differentiation

Given tabular or experimental data, how can you compute its derivative?

Approximating the tangent by a secant, a good estimate for the derivative of f at the midpoint of the (x_1, x_2) interval is

$$\big[f(x_2) - f(x_1)\big]/(x_2 - x_1)$$



As usual, the geometric approach doesn't indicate the size of the error, so it's back to Taylor's series.

Given data at points $x = 0$, $\pm h$, $\pm 2h$, $\ldots$. I want the derivative $f'(0)$.

$$f(h) = f(0) + hf'(0) + \frac{1}{2}h^2 f''(0) + \frac{1}{6}h^3 f'''(0) + \cdots$$
$$f(-h) = f(0) - hf'(0) + \frac{1}{2}h^2 f''(0) - \frac{1}{6}h^3 f'''(0) + \cdots$$

In order to isolate the term in $f'(0)$, it's necessary to eliminate the larger term $f(0)$, so subtract:

$$f(h) - f(-h) = 2hf'(0) + \frac{1}{3}h^3 f'''(0) + \cdots,$$
$$\text{giving} \qquad \frac{1}{2h}\big[f(h) - f(-h)\big] \approx f'(0) + \Big[\frac{1}{6}h^2 f'''(0)\Big] \tag{8}$$

and the last term, in brackets, estimates the error in the straight line approximation.

The most obvious point about this error term is that it varies as h^2, and so indicates by how much the error should decrease as you decrease the size of the interval. (How to estimate the factor $f'''(0)$ I'll come to presently.) This method evaluates the derivative at one of the data points; you can make it more accurate if you evaluate it between the points, so that the distance from where the derivative is being taken to where the data is available is smaller. As before, let $h = 2k$, then

$$\frac{1}{2k}\big[f(k) - f(-k)\big] = f'(0) + \frac{1}{6}k^2 f'''(0) + \cdots,$$

or, in terms of h with a shifted origin,

$$\frac{1}{h}\big[f(h) - f(0)\big] \approx f'(h/2) + \frac{1}{24}h^2 f'''\big(\frac{h}{2}\big), \tag{9}$$

and the error is only 1/4 as big.

As with interpolation methods, you can gain accuracy by going to higher order in the Taylor series,

$$\begin{aligned} f(h) - f(-h) &= 2hf'(0) + \frac{1}{3}h^3 f'''(0) + \frac{1}{60}h^5 f^v(0) + \cdots \\ f(2h) - f(-2h) &= 4hf'(0) + \frac{8}{3}h^3 f'''(0) + \frac{32}{60}h^5 f^v(0) + \cdots. \end{aligned}$$

To eliminate the largest source of error, the h^3 term, multiply the first equation by 8 and subtract the second.

$$8\big[f(h) - f(-h)\big] - \big[f(2h) - f(-2h)\big] = 12hf'(0) - \frac{24}{60}h^5 f^v(0) + \cdots,$$

or

$$f'(0) \approx \frac{1}{12h}\big[f(-2h) - 8f(-h) + 8f(h) - f(2h)\big] - \Big[-\frac{1}{30}h^4 f^v(0)\Big]. \tag{10}$$

with an error term of order h^4.

As an example of this method, let $f(x) = \sin x$ and evaluate the derivative at $x = 0.2$ by the 2-point formula and the 4-point formula with h=0.1:

2-point: $\frac{1}{0.2}[0.2955202 - 0.0998334] = 0.9784340$

4-point: $\frac{1}{1.2}[0.0 - 8 \times 0.0998334 + 8 \times 0.2955202 - 0.3894183]$

$$= 0.9800633$$

$$\cos 0.2 = 0.9800666$$

Again, you have a more accurate formula by evaluating the derivative between the data points: $h = 2k$

$$\begin{aligned} f(k) - f(-k) &= 2kf'(0) + \frac{1}{3}k^3 f'''(0) + \frac{1}{60}k^5 f^v(0) \\ f(3k) - f(-3k) &= 6kf'(0) + \frac{27}{3}k^3 f'''(0) + \frac{243}{60}k^5 f^v(0) \\ 27\big[f(k) - f(-k)\big] - \big[f(3k) - f(-3k)\big] &= 48kf'(0) - \frac{216}{60}k^5 f^v(0). \end{aligned}$$

Changing k to $h/2$ and translating the origin gives

$$\frac{1}{24h}\big[f(-h)-27f(0)+27f(h)-f(2h)\big]=f'(h/2)-\frac{3}{640}h^4f^v(h/2), \qquad (11)$$

and the coefficient of the error term is much smaller.

The previous example of the derivative of $\sin x$ at $x=0.2$ with $h=0.1$ gives, using this formula:

$$\frac{1}{2.4}[0.0499792-27\times0.1494381+27\times0.2474040-0.3428978]=0.9800661,$$

and the error is less by a factor of about 7.

You can find higher derivatives the same way.

$$\begin{aligned} f(h)&=f(0)+hf'(0)+\frac{1}{2}h^2f''(0)+\frac{1}{6}h^3f'''(0)+\frac{1}{24}h^4f''''(0)\\ f(h)+f(-h)&=2f(0)+h^2f''(0)+\frac{1}{12}h^4f''''(0)+\cdots\\ f''(0)&=\frac{f(-h)-2f(0)+f(h)}{h^2}-\frac{1}{12}h^2f''''(0)+\cdots \qquad (12)\end{aligned}$$

Notice that the numerical approximation for $f''(0)$ is even in h because the second derivative is unchanged if x is changed to $-x$.

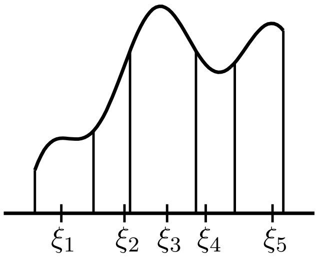
You can get any of these expressions for higher derivatives recursively, though finding the error estimates requires the series method. The above expression for $f''(0)$ can be viewed as a combination of first derivative formulas:

$$\begin{aligned} f''(0)&\approx\big[f'(h/2)-f'(-h/2)\big]/h\\ &\approx\frac{1}{h}\left[\frac{f(h)-f(0)}{h}-\frac{f(0)-f(-h)}{h}\right]\\ &=\big[f(h)-2f(0)+f(-h)\big]/h^2. \qquad (13)\end{aligned}$$

Similarly, the third and higher derivatives can be computed. The numbers that appear in these numerical derivatives are simply the binomial coefficients, Eq. (2.17).

11.4 Integration

The basic definition of an integral is a limit of the sum,



$$\sum f(\xi_i)(x_{i+1}-x_i) \qquad (x_i\le\xi_i\le x_{i+1}), \qquad (14)$$

and this is the basis for the numerical evaluation of any integral, as in section 1.6.

The simplest choices to evaluate the integral of $f(x)$ over the domain x_0 to x_0+h would be to take the position of ξ at one of the endpoints or maybe in the middle (here I assume h is small).

$$\begin{aligned}
\int_{x_0}^{x_0+h} f(x)\,dx \approx{}& f(x_0)h && && \text{(a)} \\
\text{or}\qquad & f(x_0+h)h && && \text{(b)} \qquad (15) \\
\text{or}\qquad & f(x_0+h/2)h && \text{(midpoint rule)} && \text{(c)} \\
\text{or}\qquad & \big[f(x_0)+f(x_0+h)\big]h/2 && \text{(trapezoidal rule)} && \text{(d)}
\end{aligned}$$

The last expression is the average of the first two.

I can now compare the errors in all of these approximations. Set $x_0=0$.

$$\begin{aligned}
\int_0^h dx\, f(x) &= \int_0^h dx\big[f(0)+xf'(0)+\frac{1}{2}x^2f''(0)+\frac{1}{6}x^3f'''(0)+\cdots\big] \\
&= hf(0)+\frac{1}{2}h^2f'(0)+\frac{1}{6}h^3f''(0)+\frac{1}{24}h^4f'''(0)+\cdots.
\end{aligned}$$

This immediately gives the error in formula (a):

$$\text{error (a)} = hf(0)-\int_0^h dx\, f(x) \approx -\frac{1}{2}h^2f'(0). \tag{16}$$

The error for expression (b) requires another expansion,

$$\begin{aligned}
\text{error (b)} &= hf(h)-\int_0^h dx\, f(x) \\
&= h\big[f(0)+hf'(0)+\cdots\big]-\big[hf(0)+\frac{1}{2}h^2f'(0)+\cdots\big] \\
&\approx \frac{1}{2}h^2f'(0).
\end{aligned} \tag{17}$$

Since this is the opposite sign from the previous error, it is immediately clear that the error in (d) will be less, because (d) is the average of (a) and (b).

$$\begin{aligned}
\text{error (d)} = \big[f(0)+f(0)+hf'(0)+\frac{1}{2}h^2f''(0)+\cdots\big]\frac{h}{2}& \\
-\big[hf(0)+\frac{1}{2}h^2f'(0)+\frac{1}{6}h^3f''(0)+\cdots\big]& \\
\approx \left(\frac{1}{4}-\frac{1}{6}\right)h^3f''(0) = \frac{1}{12}h^3f''(0).&
\end{aligned} \tag{18}$$

Similarly, the error in (c) is

$$\begin{aligned}
\text{error (c)} = h\big[f(0)+\frac{1}{2}hf'(0)+\frac{1}{8}h^2f''(0)+\cdots\big]& \\
-\big[hf(0)+\frac{1}{2}h^2f'(0)+\frac{1}{6}h^2f''(0)+\cdots\big]& \\
\approx -\frac{1}{24}h^3f''(0).&
\end{aligned} \tag{19}$$

The errors in the (c) and (d) formulas are both therefore the order of h^3.

Notice that just as the errors in formulas (a) and (b) canceled to highest order when you averaged them, the same thing can happen between formulas (c) and (d). Here however you need a weighted average, with twice as much of (c) as of (d). $[1/12 - 2/24 = 0]$

$$\frac{1}{3}(d) + \frac{2}{3}(c) = \big[f(x_0) + f(x_0 + h)\big]\frac{h}{6} + f\big(x_0 + h/2\big)\frac{4}{6}h. \tag{20}$$

This is known as Simpson's rule.

Simpson's Rule

Before applying this last result, I'll go back and derive it in a more systematic way, putting it into the form you'll see most often.

Integrate Taylor's expansion over a symmetric domain to simplify the algebra:

$$\int_{-h}^{h} dx\, f(x) = 2hf(0) + \frac{2}{6}h^3 f''(0) + \frac{2}{120}h^5 f''''(0) + \cdots. \tag{21}$$

I'll try to approximate this by a three point formula $\alpha f(-h) + \beta f(0) + \gamma f(h)$ where α, β, and γ, are unknown. Because of the symmetry of the problem, you can anticipate that $\alpha = \gamma$, but let that go for now and it will come out of the algebra.

$$\begin{aligned}
&\alpha f(-h) + \beta f(0) + \gamma f(h) = \\
&\qquad \alpha\big[f(0) - hf'(0) + \frac{1}{2}h^2 f''(0) - \frac{1}{6}h^3 f'''(0) + \frac{1}{24}h^4 f''''(0) + \cdots\big] \\
&\qquad +\beta f(0) \\
&\qquad +\gamma\big[f(0) + hf'(0) + \frac{1}{2}h^2 f''(0) + \frac{1}{6}h^3 f'''(0) + \frac{1}{24}h^4 f''''(0) + \cdots\big]
\end{aligned}$$

You now determine the three constants by requiring that the two series for the same integral agree to as high an order as is possible for any f.

$$\begin{aligned}
2h &= \alpha + \beta + \gamma \\
0 &= -\alpha h + \gamma h \qquad \Longrightarrow \qquad \alpha = \gamma = h/3, \qquad \beta = 4h/3 \\
\frac{1}{3}h^3 &= \frac{1}{2}(\alpha + \gamma)h^2
\end{aligned}$$

$$\text{and so,} \qquad \int_{-h}^{h} dx\, f(x) \approx \frac{h}{3}\big[f(-h) + 4f(0) + f(h)\big]. \tag{22}$$

The error term (the "truncation error") is

$$\frac{h}{3}\big[f(-h)+4f(0)+f(-h)\big] - \int_{-h}^{h} dx\, f(x) \approx \frac{1}{12}\cdot\frac{1}{3}h^5 f''''(0) - \frac{1}{60}h^5 f''''(0) = \frac{1}{90}h^5 f''''(0). \tag{23}$$

Simpson's rule is exact up through cubics, because the fourth and higher derivatives vanishes in that case. It's worth noting that there is also an elementary derivation of Simpson's rule: Given three points, there is a unique quadratic in x that passes through all of them. Take the three points to be $\big(-h, f(-h)\big)$, $\big(0, f(0)\big)$, and $\big(h, f(h)\big)$, then integrate the resulting polynomial. Express your answer for the integral in terms of the values of f at the three points, and you get the above Simpson's rule. This has the drawback that it gives no estimate of the error.

To apply Simpson's rule, it's necessary to divide the region of integration into an even number of pieces and apply the above formula to each pair.

$$\begin{aligned}\int_a^b dx\, f(x) \approx \frac{h}{3}\big[f(x_0)+4f(x_1)+f(x_2)\big] + \frac{h}{3}\big[f(x_2)+4f(x_3)+f(x_4)\big] + \cdots \\ + \frac{h}{3}\big[f(x_{N-2})+4f(x_{N-1})+f(x_N)\big] \\ = \frac{h}{3}\big[f(x_0)+4f(x_1)+2f(x_2)+4f(x_3)+\cdots+4f(x_{N-1})+f(x_N)\big] \qquad (24)\end{aligned}$$

Example:

$$\int_0^1 \frac{4}{1+x^2}dx = 4\tan^{-1} x\bigg|_0^1 = \pi$$

Divide the interval 0 to 1 into four pieces, then

$$\int_0^1 \frac{4}{1+x^2}dx \approx \frac{4}{12}\left[1 + 4\frac{1}{1+(1/4)^2} + 2\frac{1}{1+(1/2)^2} + 4\frac{1}{1+(3/4)^2} + \frac{1}{1+1}\right] = 3.1415686$$

as compared to $\pi = 3.1415927\ldots$.

When the function to be integrated is smooth, this gives very accurate results.

Gaussian Integration

If the integrand is known at all points of the interval and not just at discrete locations as for tabulated or experimental data, there is more freedom that you can use to gain higher accuracy even is you use only a two point formula:

$$\int_{-h}^{h} f(x)\,dx \approx \alpha\big[f(\beta)+f(-\beta)\big].$$

I could try picking two arbitrary points, not symmetrically placed, in the interval, but the previous experience with Simpson's rule indicates that the result will come out as indicated. (Though it's easy to check what happens if you pick two general points in the interval.)

$$2hf(0) + \frac{1}{3}h^3 f''(0) + \frac{1}{60}h^5 f''''(0) + \cdots = \alpha\big[2f(0) + \beta^2 f''(0) + \frac{1}{12}\beta^4 f''''(0) + \cdots\big]$$

To make this an equality through the low orders implies

$$\text{or} \qquad \begin{aligned} 2h &= 2\alpha \\ \alpha &= h \end{aligned} \qquad \begin{aligned} \frac{1}{3}h^3 &= \alpha\beta^2 \\ \beta &= h/\sqrt{3}. \end{aligned} \qquad (25)$$

with an error term

$$\frac{1}{12}\cdot\frac{1}{9}h^5 f''''(0) - \frac{1}{60}h^5 f''''(0) = -\frac{1}{135}h^5 f''''(0),$$

and

$$\int_{-h}^{h} f(x)\,dx \approx h\left[f\left(h/\sqrt{3}\right)+f\left(-h/\sqrt{3}\right)\right]-\left[-\frac{1}{135}h^5 f''''(0)\right] \tag{26}$$

With only two points, this expression yields an accuracy equal to the three point Simpson formula.

Notice that the two points found in this way are roots of a certain quadratic

$$\left(x-\frac{1}{\sqrt{3}}\right)\left(x+\frac{1}{\sqrt{3}}\right)=x^2-1/3,$$

which is proportional to

$$\frac{3}{2}x^2-\frac{1}{2}=P_2(x), \tag{27}$$

the Legendre polynomial of second order.

This approach to integration, known as Gaussian integration, or Gaussian quadrature, can be extended to more points, as for example

$$\int_{-h}^{h} f(x)\,dx \approx \alpha f(-\beta)+\gamma f(0)+\alpha f(\beta).$$

The same expansion procedure leads to the result

$$\frac{h}{9}\left[5f\left(-h\sqrt{\frac{3}{5}}\right)+8f(0)+f\left(h\sqrt{\frac{3}{5}}\right)\right], \tag{28}$$

with an error proportional to $h^7 f^{(6)}(0)$. The polynomial with roots $0, \pm\sqrt{3/5}$ is

$$\frac{5}{2}x^3-\frac{3}{2}x=P_3(x), \tag{29}$$

the third order Legendre polynomial.

For an integral $\int_a^b f(x)\,dx$, let $x=[(a+b)/2]+z[(b-a)/2]$. Then for the domain $-1<z<1$, x covers the whole integration interval.

$$\int_a^b f(x)\,dx=\frac{b-a}{2}\int_{-1}^{1} dz\,f(x)$$

When you use an integration scheme such as Gauss's, it is in the form of a weighted sum over points. The weights and the points are defined by equations such as (26) or (28).

$$\int_{-1}^{1} dz\,f(z)\ \to\ \sum_k w_k f(z_k) \tag{30}$$

$$\text{or}\qquad \int_a^b f(x)\,dx=\frac{b-a}{2}\sum_k w_k f(x_k),\quad x_k=[(a+b)/2]+z_k[(b-a)/2]$$

Many other properties of Gaussian integration are discussed in the two books by C. Lanczos, “Linear Differential Operators,” “Applied Analysis,” both available in Dover reprints. The general expressions for the integration points as roots of Legendre polynomials and expressions for the coefficients are there. The important technical distinction he points out between the Gaussian method and generalizations of Simpson's rule involving more points is in the divergences for large numbers of points. Gauss's method does not suffer from this defect. In practice, there is rarely any problem with using the ordinary Simpson rule as indicated above, though it will require more points than the more elegant Gauss's method. When problems do arise with either of these methods, they often occur because the function is ill-behaved, and the high derivatives are very large. In this case it can be more accurate to use a method with a lower order derivative for the truncation error.

11.5 Differential Equations

To solve the first order differential equation

$$y' = f(x,y) \qquad y(x_0) = y_0, \tag{31}$$

the simplest algorithm is Euler's method. The initial conditions are $y(x_0) = y_0$, and $y'(x_0) = f(x_0, y_0)$, and a straight line extrapolation is

$$y(x_0 + h) = y_0 + hf(x_0, y_0). \tag{32}$$

You can now iterate this procedure using this newly found value of y as a new starting condition to go from $x_0 + h$ to $x_0 + 2h$.

Runge-Kutta

Euler's method is not very accurate. For an improvement, change from a straight line extrapolation to a parabolic one. Take $x_0 = 0$ to keep the algebra down, and try a solution near 0 in the form $y(x) = \alpha + \beta x + \gamma x^2$; evaluate α, β, and γ so that the differential equation is satisfied near $x = 0$,

$$y' = \beta + 2\gamma x = f(x, \alpha + \beta x + \gamma x^2).$$

Recall the Taylor series expansion for a function of two variables, section 2.5:

$$\begin{aligned} f(x,y) = f(x_0,y_0) + (x-x_0)D_1 f(x_0,y_0) &+ (y-y_0)D_2 f(x_0,y_0) + \frac{1}{2}(x-x_0)^2 D_1 D_1 f(x_0,y_0) \\ &+ \frac{1}{2}(y-y_0)^2 D_2 D_2 f(x_0,y_0) + (x-x_0)(y-y_0) D_1 D_2 f(x_0,y_0) + \cdots \end{aligned} \tag{33}$$

$$\Longrightarrow \beta + 2\gamma x = f(0,\alpha) + x D_1 f(0,\alpha) + (\beta x + \gamma x^2) D_2 f(0,\alpha) + \cdots. \tag{34}$$

The initial condition is at $x = 0, y = y_0$, so $\alpha = y_0$. Equate coefficients of powers of x as high as is possible (here through x^1).

$$\beta = f(0,\alpha) \qquad 2\gamma = D_1 f(0,\alpha) + \beta D_2 f(0,\alpha).$$

(If you set $\gamma = 0$, this is Euler's method.)

$$y(h) = y_0 + hf(0,y_0) + \frac{h^2}{2}\big[D_1 f(0,y_0) + f(0,y_0) D_2 f(0,y_0)\big]. \tag{35}$$

The next problem is to evaluate these derivatives. If you can easily do them analytically, you can choose to do that. Otherwise, since they appear in a term that is multiplied by h^2, it is enough to use the simplest approximation for the numerical derivative,

$$D_1 f(0,y_0) = \bigl[f(h,y_0) - f(0,y_0)\bigr]/h. \tag{36}$$

You cannot expect to use the same interval, h, for the y variable — it might not even have the same dimensions,

$$D_2 f(0,y_0) = \bigl[f(j,y_0+k) - f(j,y_0)\bigr]/k. \tag{37}$$

where j and k are the order of h. Note that because this term appears in an expression multiplied by h^2, it doesn't matter what j is. You can choose it for convenience. Possible values for these are

$$\begin{array}{llll} (1)\ j=0 & k=hf(0,y_0) & (3)\ j=h & k=hf(0,y_0) \\ (2)\ j=0 & k=hf(h,y_0) & (4)\ j=h & k=hf(h,y_0). \end{array}$$

The third of these choices for example gives

$$\begin{aligned} y &= y_0 + hf(0,y_0) + \frac{h^2}{2}\left[\frac{1}{h}\bigl[f(h,y_0) - f(0,y_0)\bigr] + f(0,y_0)\frac{f(h,y_0+k) - f(h,y_0)}{hf(0,y_0)}\right] \\ &= y_0 + \frac{h}{2}f(0,y_0) + \frac{h}{2}f\bigl(h, y_0 + hf(0,y_0)\bigr). \end{aligned} \tag{38}$$

This procedure, a second order Runge-Kutta method, is a moderately accurate method for advancing from one point to the next in the solution of a differential equation. It requires evaluating the function twice for each step of the iteration.
Example: $y' = 1 + y^2$ $y(0) = 0.$ Let h=0.1

x	y(Euler)	y(RK2)	tan x
0.	0.	0.	0.
0.1	0.10	0.10050	0.10053
0.2	0.20100	0.20304	0.20271
0.3	0.30504	0.30981	0.30934
0.4	0.41435	0.42341	0.42279
0.5	0.53151	0.54702	0.54630

(39)

The error at $x = 0.5$ with RK2 is 0.13% and with Euler it is 2.7%. A commonly used version of this is the fourth order Runge-Kutta method:

$$y = y_0 + \frac{1}{6}\bigl[k_1 + 2k_2 + 2k_3 + k_4\bigr] \tag{40}$$

$$\begin{array}{ll} k_1 = hf\bigl(0,y_0\bigr) & k_2 = hf\bigl(h/2, y_0 + k_1/2\bigr) \\ k_3 = hf\bigl(h/2, y_0 + k_2/2\bigr) & k_4 = hf\bigl(h, y_0 + k_3\bigr). \end{array}$$

You can look up a fancier version of this called the Runge-Kutta-Fehlberg method. It's one of the better techniques around.

Higher Order Equations

How can you use either the Euler or the Runge-Kutta method to solve a second order differential equation? Answer: Turn it into a pair of first order equations.

$$y'' = f(x, y, y') \longrightarrow y' = v, \qquad \text{and} \qquad v' = f(x, y, v) \tag{41}$$

The Euler method, Eq. (32) becomes

$$y(x_0 + h) = y(x_0) + hv(x_0), \qquad \text{and} \qquad v(x_0 + h) = v(x_0) + hf\big(x_0, y(x_0), v(x_0)\big)$$

The construction for Runge-Kutta is essentially the same.

Adams Methods

The Runge-Kutta algorithm has the advantage that it is self-starting; it requires only the initial condition to go on to the next step. It has the disadvantage that it is inefficient. In going from one step to the next, it ignores all the information available from any previous steps. The opposite approach leads to the Adams methods, though these are not as commonly used any more. I'm going to develop a little of the subject mostly to show that the methods that I've used so far can lead to disaster if you're not careful.

Shift the origin to be the point at which you want the new value of y. Assume that you already know y at $-h$, $-2h$, $\ldots$, $-Nh$. Because of the differential equation $y' = f(x, y)$, you also know y' at these points.

Assume

$$y(0) = \sum_1^N \alpha_k y(-kh) + \sum_1^N \beta_k y'(-kh). \tag{42}$$

With $2N$ parameters, you can get this accurate to order h^{2N-1},

$$y(-kh) = \sum_0^\infty (-kh)^n \frac{y^{(n)}(0)}{n!}.$$

Substitute this into the equation for $y(0)$:

$$y(0) = \sum_{k=1}^N \alpha_k \sum_{n=0}^\infty (-kh)^n \frac{y^{(n)}(0)}{n!} + h \sum_{k=1}^N \beta_k \sum_{n=0}^\infty (-kh)^n \frac{y^{(n+1)}(0)}{n!}.$$

This should be an identity to as high an order as possible. The coefficient of h^0 gives

$$1 = \sum_{k=1}^N \alpha_k. \tag{43}$$

The next orders are

$$\begin{aligned} 0 &= \sum_k \alpha_k(-kh) + h\sum_k \beta_k \\ 0 &= \sum_k \frac{1}{2}\alpha_k(-kh)^2 + h\sum_k \beta_k(-kh) \\ &\vdots \end{aligned} \tag{44}$$

$N = 1$ is Euler's method again.

$N = 2$ gives

$$\alpha_1 + \alpha_2 = 1 \qquad\qquad \alpha_1 + 2\alpha_2 = \beta_1 + \beta_2$$
$$\alpha_1 + 4\alpha_2 = 2(\beta_1 + 2\beta_2) \qquad\qquad \alpha_1 + 8\alpha_2 = 3(\beta_1 + 4\beta_2).$$

The solution of these equations is

$$\alpha_1 = -4 \qquad \alpha_2 = +5 \qquad \beta_1 = +4 \qquad \beta_2 = +2$$
$$y(0) = -4y(-h) + 5y(-2h) + h\big[4y'(-h) + 2y'(-2h)\big]. \tag{45}$$

To start this algorithm off, you need two pieces of information: the values of y at $-h$ and at $-2h$. This is in contrast to Runge-Kutta, which needs only one point.

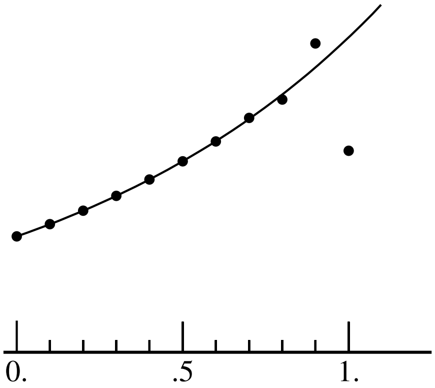
Example: Solve $y' = y \quad y(0) = 1 \qquad (h = 0.1)$

I could use Runge-Kutta to start and then switch to Adams as soon as possible. For the purpose of this example, I'll just take the exact value of y at $x = 0.1$.

$$\begin{aligned} e^{.1} &= 1.105170918 \\ y(.2) &= -4y(.1) + 5y(0) + .1\big[4f\big(.1, y(.1)\big) + 2f\big(0, y(0)\big)\big] \\ &= -4y(.1) + 5y(0) + .4y(.1) + .2y(0) \\ &= -3.6y(.1) + 5.2y(0) \\ &= 1.22138\underline{8}4695 \end{aligned}$$

The exact value is $e^{.2} = 1.221402758$; the first error is in the underlined term. Continuing the calculation to higher values of x,

x	y
.3	$1.3499\underline{0}38$
.4	$1.491\underline{5}47$
.5	$1.648\underline{9}31$
.6	$1.81\underline{9}88$
.7	$2.0\underline{2}28$
.8	$2.1\underline{8}12$
.9	$2.\underline{6}66$
1.0	$\underline{1}.74$
1.1	7.59
1.2	−18.26
1.3	105.22



Everything is going very smoothly for a while, though the error is creeping up. At around $x = 1$, the numerical solution goes into wild oscillation and is completely unstable. The reason for this is in the coefficients -4 and $+5$ of $y(-h)$ and $y(-2h)$. Small errors are magnified by these large factors. (The coefficients of y' are not any trouble because of the factor h in front.)

Instability

You can compute the growth of this error explicitly in this simple example. The equation (45) together with $y' = y$ is

$$y(0) = -3.6y(-h) + 5.2y(-2h),$$

or in terms of an index notation

$$y_n = -3.6y_{n-1} + 5.2y_{n-2}.$$

This is a linear, constant coefficient, difference equation, and the method for solving it is essentially the same as for a linear differential equation — assume an exponential form $y_n = k^n$.

$$\begin{gathered} k^n = -3.6k^{n-1} + 5.2k^{n-2} \\ k^2 + 3.6k - 5.2 = 0 \\ k = 1.11 \qquad \text{and} \qquad -4.71 \end{gathered}$$

Just as with the differential equation, the general solution is a linear combination of these two functions of n:

$$y_n = A(1.11)^n + B(-4.71)^n,$$

where A and B are determined by two conditions, typically specifying y_1 and y_2. If $B = 0$, then y_n is proportional to 1.11^n and it is the well behaved exponential solution that you expect. If, however, there is even a little bit of B present (perhaps because of roundoff errors), that term will eventually dominate and cause the large oscillations. If B is as small as 10^{-6}, then when $n = 9$ the unwanted term is greater than 1.

When I worked out the coefficients in Eq. (45) the manipulations didn't look all that different from those leading to numerical derivatives or integrals, but the result was useless. This is a warning. You're in treacherous territory here; tread cautiously.

Are Adams-type methods useless? No, but you have to modify the development in order to get a stable algorithm. The difficulty in assuming the form

$$y(0) = \sum_1^N \alpha_k y(-kh) + \sum_1^N \beta_k y'(-kh)$$

is that the coefficients α_k are too large. To cure this, you can give up some of the $2N$ degrees of freedom that the method started with, and pick the α_k *a priori* to avoid instability. There are two common ways to do this, consistent with the constraint that must be kept on the α's,

$$\sum_{k=1}^N \alpha_k = 1.$$

One way is to pick all the α_k to equal $1/N$. Another way is to pick $\alpha_1 = 1$ and all the others $= 0$, and both of these methods are numerically stable. The book by Lanczos in the bibliography goes into these techniques, and there are tabulations of these and other methods in Abramowitz and Stegun.

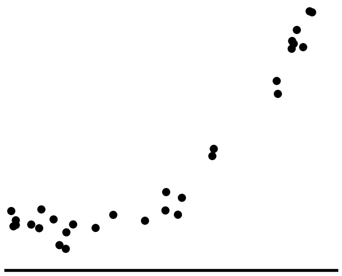
Backwards Iteration

Before leaving the subject, there is one more kind of instability that you can encounter. If you try to solve $y'' = +y$ with $y(0) = 1$ and $y'(0) = -1$, the solution is e^{-x}. If you use any stable numerical algorithm to solve this problem, it will soon deviate arbitrarily far from the desired one. The reason is that the general solution of this equation is $y = Ae^x + Be^{-x}$. Any numerical method will, through rounding errors, generate a little bit of the undesired solution, e^{+x}. Eventually, this must overwhelm the correct solution. No algorithm, no matter how stable, can get around this.

There is a clever trick that sometimes works in cases like this: backwards iteration. Instead of going from zero up, start at some large value of x and iterate downward. In this direction it is the desired solution, e^{-x}, that is unstable, and the e^{+x} is damped out. Pick an arbitrary value, say $x = 10$, and assign an arbitrary value to $y(10)$, say 0. Next, pick an arbitrary value for $y'(10)$, say 1. Use these as initial conditions (terminal conditions?) and solve the differential equation moving left; necessarily the dominant term will be the unstable one, e^{-x}, and independent of the choice of initial conditions, it will be *the* solution. At the end it is only necessary to multiply all the terms by a scale factor to reduce the value at $x = 0$ to the desired one; automatically, the value of $y'(0)$ will be correct. What you are really doing by this method is to replace the initial value problem by a two point boundary value problem. You require that the function approach zero for large x.

11.6 Fitting of Data

If you have a set of data in the form of independent and dependent variables $\{x_i, y_i\}$ $(i = 1, \ldots, N)$, and you have proposed a model that this data is to be represented by a linear combination of some set of functions, $f_\mu(x)$



$$y = \sum_{\mu=1}^{M} \alpha_\mu f_\mu(x), \tag{46}$$

what values of α_μ will represent the observations in the "best" way? There are several answers to this question depending on the meaning of the word "best." The most commonly used one, largely because of its simplicity, is Gauss's method of least squares.

Here there are N data and there are M functions that I will use to fit the data. You have to pick the functions for yourself. You can choose them because they are the implications of a theoretical calculation; you can choose them because they are simple; you can choose them because your daily horoscope suggested them. The sum of functions, $\sum \alpha_\mu f_\mu$, now depends only on the M parameters α_μ. The fs are fixed. The difference between this sum and the data points y_i is what you want to be as small as possible. You can't use the differences themselves because they will as likely be negative as positive. The least squares method uses the sum of the squares of the differences between your sum of functions and the data. This criterion for best fit is that the sum

$$\sum_{i=1}^{N} \left[y_i - \sum_{\mu=1}^{M} \alpha_\mu f_\mu(x_i) \right]^2 = N\sigma^2 \tag{47}$$

be a minimum. The mean square deviation of the theory from the experiment is to be least. This quantity σ^2 is called the variance.

Some observations to make here: $N \geq M$, for otherwise there are more free parameters than data to fit them, and almost any theory with enough parameters can be forced to fit any data. Also, the functions f_μ must be linearly independent; if not, then you can throw away some and not alter the result — the solution is not unique. A further point: there is no requirement that all of the x_i are different; you may have repeated the measurements at some points.

Minimizing this is now a problem in ordinary calculus with M variables.

$$\frac{\partial}{\partial\alpha_\nu}\sum_{i=1}^{N}\left[y_i - \sum_{\mu=1}^{M}\alpha_\mu f_\mu(x_i)\right]^2 = -2\sum_i\left[y_i - \sum_\mu \alpha_\mu f_\mu(x_i)\right] f_\nu(x_i) = 0$$

$$\text{rearrange:}\qquad \sum_\mu\left[\sum_i f_\nu(x_i)f_\mu(x_i)\right]\alpha_\mu = \sum_i y_i f_\nu(x_i). \tag{48}$$

These linear equations are easily expressed in terms of matrices.

$$Ca = b,$$

where

$$C_{\nu\mu} = \sum_{i=1}^{N} f_\nu(x_i)f_\mu(x_i). \tag{49}$$

a is the column matrix with components α_μ and b has components $\sum_i y_i f_\nu(x_i)$.

The solution for a is

$$a = C^{-1}b. \tag{50}$$

If C turned out singular, so this inversion is impossible, the functions f_μ were not independent.

Example: Fit to a straight line

$$f_1(x) = 1 \qquad\qquad f_2(x) = x.$$

Then $Ca = b$ is

$$\begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} = \begin{pmatrix}\sum y_i \\ \sum y_i x_i.\end{pmatrix}$$

The inverse is

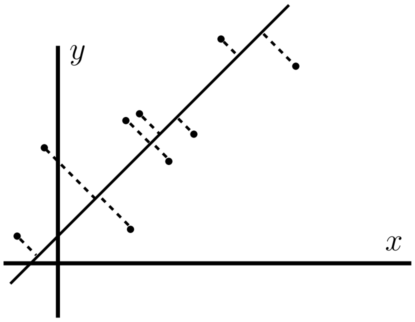
$$\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} = \frac{1}{\left[N\sum x_i^2 - \left(\sum x_i\right)^2\right]}\begin{pmatrix}\sum x_i^2 & -\sum x_i \\ -\sum x_i & N\end{pmatrix}\begin{pmatrix}\sum y_i \\ \sum x_i y_i.\end{pmatrix} \tag{51}$$

and the best fit line is

$$y = \alpha_1 + \alpha_2 x$$

11.7 Euclidean Fit

In fitting data to a combination of functions, the least squares method used Eq. (47) as a measure of how far the proposed function is from the data. If you're fitting data to a straight line (or plane if you have more variables) there's another way to picture the distance. Instead of measuring the distance from a point to the curve *vertically* using only y, measure it as the *perpendicular* distance to the line. Why should this be any better? It's not, but it does have different uses, and a primary one is data compression.



Do this in two dimensions, fitting the given data to a straight line, and to describe the line I'll use vector notation, where the line is $\vec u + \alpha\vec v$ and the parameter α varies over the reals. First I need to answer the simple question: what is the distance from a point to a line? The perpendicular distance from $\vec w$ to this line requires that

$$d^2 = \big(\vec w - \vec u - \alpha\vec v\big)^2$$

be a minimum. Differentiate this with respect to α and you have

$$\big(\vec w - \vec u - \alpha\vec v\big)\cdot\big(-\vec v\big) = 0 \qquad \text{implying} \qquad \alpha v^2 = \big(\vec w - \vec u\big)\cdot\vec v$$

For this value of α what is d^2?

$$\begin{aligned} d^2 &= \big(\vec w - \vec u\big)^2 + \alpha^2 v^2 - 2\alpha\vec v\cdot\big(\vec w - \vec u\big) \\ &= \big(\vec w - \vec u\big)^2 - \frac{1}{v^2}\big[(\vec w - \vec u)\cdot\vec v\big]^2 \end{aligned} \tag{52}$$

Is this plausible? (1) It's independent of the size of $\vec v$, depending on its direction only. (2) It depends on only the *difference* vector between $\vec w$ and $\vec u$, not on any other aspect of the vectors. (3) If I add any multiple of $\vec v$ to $\vec u$, the result is unchanged. See problem 37. Also, *can you find an easier way to get the result?* Perhaps one that simply requires some geometric insight?

The data that I'm trying to fit will be described by a set of vectors $\vec w_i$, and the sum of the distances squared to the line is

$$D^2 = \sum_1^N \big(\vec w_i - \vec u\big)^2 - \sum_1^N \frac{1}{v^2}\big[(\vec w_i - \vec u)\cdot\vec v\big]^2$$

Now to minimize this among all $\vec u$ and $\vec v$ I'll first take advantage of some of the observations that I made in the preceding paragraph. Because the magnitude of $\vec v$ does not matter, I'll make it a unit vector.

$$D^2 = \sum\big(\vec w_i - \vec u\big)^2 - \sum\big[(\vec w_i - \vec u)\cdot\hat v\big]^2 \tag{53}$$

Now to figure out $\vec{u}$, I note that I expect the best fit line to go somewhere through the middle of the set of data points, so move the origin to the "center of mass" of the points.

$$\vec{w}_{\text{mean}} = \sum \vec{w}_i/N \qquad \text{and let} \qquad \vec{w}_i^{\,\prime} = \vec{w}_i - \vec{w}_{\text{mean}} \qquad \text{and} \qquad \vec{u}^{\,\prime} = \vec{u} - \vec{w}_{\text{mean}}$$

then the sum $\sum \vec{w}_i^{\,\prime} = 0$ and

$$D^2 = \sum w_i'^2 + Nu'^2 - \sum (\vec{w}_i^{\,\prime} \cdot \hat{v})^2 - N(\vec{u}^{\,\prime} \cdot \hat{v})^2 \tag{54}$$

This depends on four variables, u'_x, u'_y, v_x and v_y. If I have to do derivatives with respect to all of them, so be it, but maybe I can use some geometric insight to simplify the calculation. I can still add any multiple of $\hat{v}$ to $\vec{u}$ without changing this expression. That means that for a given $\vec{v}$ the derivative of D^2 as I change $\vec{u}^{\,\prime}$ in *that* particular direction is zero. It's only as I change $\vec{u}^{\,\prime}$ perpendicular to the direction of $\vec{v}$ that D^2 changes. The second and fourth term involve $u'^2 - (\vec{u}^{\,\prime} \cdot \hat{v}\,)^2 = u'^2(1 - \cos^2\theta) = u'^2 \sin^2\theta$, where this angle θ is the angle between $\vec{u}^{\,\prime}$ and $\vec{v}$. This *is* the perpendicular distance to the line (squared). Call it $u'_\perp = u' \sin\theta$.

$$D^2 = \sum w_i'^2 - \sum (\vec{w}_i^{\,\prime} \cdot \hat{v})^2 + Nu'^2 - N(\vec{u}^{\,\prime} \cdot \hat{v})^2 = \sum w_i'^2 - \sum (\vec{w}_i^{\,\prime} \cdot \hat{v})^2 + Nu_\perp'^2$$

The minimum of this obviously occurs for $\vec{u}'_\perp = 0$. Also, because the component of $\vec{u}'$ along the direction of $\vec{v}$ is arbitrary, I may as well take it to be zero. That makes $\vec{u}' = 0$. Remember now that this is for the shifted $\vec{w}^{\,\prime}$ data. For the original $\vec{w}_i$ data, $\vec{u}$ is shifted to $\vec{u} = \vec{w}_{\text{mean}}$.

$$D^2 = \sum w_i'^2 - \sum (\vec{w}_i^{\,\prime} \cdot \hat{v})^2 \tag{55}$$

I'm not done. I still have to find the direction of $\hat{v}$. That is, I have to find the minimum of D^2 subject to the constraint that $|\hat{v}| = 1$. Use Lagrange multipliers (section 8.12).

$$\text{Minimize} \quad D^2 = \sum w_i'^2 - \sum (\vec{w}_i^{\,\prime} \cdot \vec{v}\,)^2 \qquad \text{subject to} \qquad \phi = v_x^2 + v_y^2 - 1 = 0$$

The independent variables are v_x and v_y, and the problem becomes

$$\nabla\big(D^2 + \lambda\phi\big) = 0, \qquad \text{with} \qquad \phi = 0$$

Differentiate with respect to the independent variables and you have linear equations for v_x and v_y,

$$-\frac{\partial}{\partial v_x} \sum \left(w'_{xi} v_x + w'_{yi} v_y\right)^2 + \lambda 2 v_x = 0 \qquad \text{or} \qquad \begin{aligned} -\sum 2\big(w'_{xi} v_x + w'_{yi} v_y\big) w_{xi} + \lambda 2 v_x = 0 \\ -\sum 2\big(w'_{xi} v_x + w'_{yi} v_y\big) w_{yi} + \lambda 2 v_y = 0 \end{aligned} \tag{56}$$

Correlation, Principal Components

The correlation matrix of this data is

$$(C) = \frac{1}{N} \begin{pmatrix} \sum w_{xi}'^2 & \sum w_{xi} w'_{yi} \\ \sum w'_{yi} w'_{xi} & \sum w_{yi}'^2 \end{pmatrix}$$

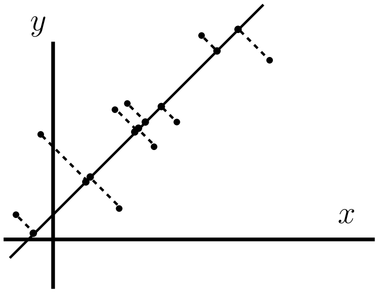
The equations (56) are

$$\begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} = \lambda' \begin{pmatrix} v_x \\ v_y \end{pmatrix} \tag{57}$$

where $\lambda' = \lambda/N$. This is a traditional eigenvector equation, and there is a non-zero solution only if the determinant of the coefficients equals zero. Which eigenvalue do I pick? There are two of them, and one will give the best fit while the other gives the *worst* fit. Just because the first derivative is zero doesn't mean you have a minimum of D^2; it could be a maximum or a saddle. Here the answer is that you pick the largest eigenvalue. You can see why this is plausible by looking at the special case for which all the data lie along the x-axis, then $C_{xx} > 0$ and all the other components of the matrix $= 0$. The eigenvalues are C_{xx} and zero, and the corresponding eigenvectors are $\hat x$ and $\hat y$ respectively. Clearly the best fit corresponds to the former, and the best fit line is the x-axis. The general form of the best fit line is (now using the original coordinate system for the data)

$$\alpha\hat v + \frac{1}{N}\sum \vec w_i = \alpha\hat v + \vec w_{\text{mean}}$$

and this $\hat v$ is the eigenvector having the largest eigenvalue. More generally, look at Eq. (55) and you see that that lone negative term is biggest if the $\vec w$s are in the same direction (or opposite) as $\hat v$.



This establishes the best fit to the line in the Euclidean sense. What good is it? It leads into the subject of Principal Component Analysis and of Data Reduction. The basic idea of this scheme is that if this fit is a good one, and the original points lie fairly close to the line that I've found, I can replace the original data with the points on this line. The nine points in this figure require $9 \times 2 = 18$ coordinates to describe their positions. The nine points that approximate the data, but that lie on the line and are closest to the original points require $9 \times 1 = 9$ coordinates along this line. Of course you have some overhead in the data storage because you need to know the line. That takes three more data ($\vec u$ and the angle of $\hat v$), so the total data storage is 12 numbers. See problem 38

This doesn't look like much of a saving, but if you have 10^6 points you go from $2\,000\,000$ numbers to $1\,000\,003$ numbers, and that starts to be significant. Remember too that this is only a two dimensional problem, with only two numbers for each point. With more coordinates you will sometimes achieve far greater savings. You can easily establish the equation to solve for the values of α for each point, problem 38. The result is

$$\alpha_i = \big(\vec w_i - \vec u\big) \cdot \hat v$$

11.8 Differentiating noisy data

Differentiation involves dividing a small number by another small number. Any errors in the numerator will be magnified by this process. If you have to differentiate experimental data this will always happen. If it is data from the output of a Monte Carlo calculation the same problem will arise.

Here is a method for differentiation that minimizes the sensitivity of the result to the errors in the input. Assume equally spaced data where each value of the dependent variable $f(x)$ is a random variable with mean $\big\langle f(x)\big\rangle$ and variance σ^2. Follow the procedure for differentiating smooth data and expand in a power series. Let $h = 2k$ and obtain the derivative between data points.

$$\begin{aligned} f(k) &= f(0) + kf'(0) + \frac{1}{2}k^2 f''(0) + \frac{1}{6}k^3 f'''(0) + \cdots \\ f(k) - f(-k) &= 2kf'(0) + \frac{1}{3}k^3 f'''(0) + \cdots \\ f(3k) - f(-3k) &= 6kf'(0) + \frac{27}{3}k^3 f'''(0) + \cdots \end{aligned}$$

I'll seek a formula of the form

$$f'(0) = \alpha\big[f(k) - f(-k)\big] + \beta\big[f(3k) - f(-3k)\big]. \tag{58}$$

I am assuming that the variance of f at each point is the same, σ^2, and that the fluctuations in f at different points are uncorrelated. The last statement is, for random variables f_1 and f_2,

$$\big\langle\big(f_1 - \langle f_1\rangle\big)\big(f_2 - \langle f_2\rangle\big)\big\rangle = 0 \qquad \text{which expands to} \qquad \big\langle f_1 f_2\big\rangle = \big\langle f_1\big\rangle\big\langle f_2\big\rangle. \tag{59}$$

Insert the preceding series expansions into Eq. (58) and match the coefficients of $f'(0)$. This gives an equation for α and β:

$$2k\alpha + 6k\beta = 1. \tag{60}$$

One way to obtain another equation for α and β is to require that the $k^3 f'''(0)$ term vanish; this leads back to the old formulas for differentiation, Eq. (11). Instead, require that the variance of $f'(0)$ be a minimum.

$$\begin{aligned} \big\langle\big(f'(0) - \langle f'(0)\rangle\big)^2\big\rangle &= \big\langle\big[\alpha\big(f(k) - \langle f(k)\rangle\big) + \alpha\big(f(-k) - \langle f(-k)\rangle\big) + \cdots\big]^2\big\rangle \\ &= 2\sigma^2\alpha^2 + 2\sigma^2\beta^2 \end{aligned} \tag{61}$$

This comes from the fact that the correlation between say $f(k)$ and $f(-3k)$ vanishes, and that all the individual variances are σ^2. That is,

$$\Big\langle\big(f(k) - \langle f(k)\rangle\big)\big(f(-k) - \langle f(-k)\rangle\big)\Big\rangle = 0$$

along with all the other cross terms. Problem: minimize $2\sigma^2(\alpha^2 + \beta^2)$ subject to the constraint $2k\alpha + 6k\beta = 1$. It's hardly necessary to resort to Lagrange multipliers for this problem.

Eliminate α:

$$\frac{d}{d\beta}\left[\left(\frac{1}{2k}-3\beta\right)^2+\beta^2\right]=0 \quad\Longrightarrow\quad -6\left(\frac{1}{2k}-3\beta\right)+2\beta=0$$
$$\Longrightarrow\quad \beta=3/20k,\quad \alpha=1/20k$$

$$f'(.5h)\approx\frac{-3f(-h)-f(0)+f(h)+3f(2h)}{10h}, \tag{62}$$

and the variance is $2\sigma^2(\alpha^2+\beta^2)=\sigma^2/5h^2$. In contrast, the formula for the variance in the standard four point differentiation formula Eq. (10), where the truncation error is least, is $65\sigma^2/72h^2$, which is 4.5 times larger.

When the data is noisy, and most data is, this expression will give much better results for this derivative. Can you do even better? Of course. You can for example go to higher order and both decrease the truncation error and minimize the statistical error.

11.9 Partial Differential Equations

I'll illustrate the ideas involved here and the difficulties that occur in only the simplest example of a PDE, a first order constant coefficient equation in one space dimension

$$\partial u/\partial t+c\,\partial u/\partial x=u_t+cu_x=0, \tag{63}$$

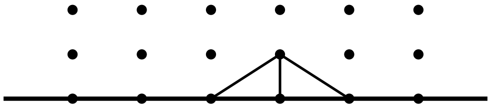
where the subscript denotes differentiation with respect to the respective variables. This is a very simple sort of wave equation. Given the initial condition that at $t=0$, $u(0,x)=f(x)$, you can easily check that the solution is

$$u(t,x)=f(x-ct). \tag{64}$$

The simplest scheme to carry data forward in time from the initial values is a generalization of Euler's method for ordinary differential equations

$$\begin{aligned}u(t+\Delta t,x)&=u(t,x)+u_t(t,x)\Delta t\\&=u(t,x)-u_x(t,x)c\Delta t\\&=u(t,x)-\frac{c\Delta t}{2\Delta x}\big[u(t,x+\Delta x)-u(t,x-\Delta x)\big],\end{aligned} \tag{65}$$

where to evaluate the derivative, I've used the two point differentiation formula.



In this equation, the value of u at point $(\Delta t,4\Delta x)$ depends on the values at $(0,3\Delta x)$, $(0,4\Delta x)$, and $(0,5\Delta x)$. This diagram shows the scheme as a picture, with the horizontal axis being x and the vertical axis t. You march the values of u at the grid points forward in time (or backward) by a set of simple equations.

The difficulties in this method are the usual errors, and more importantly, the instabilities that can occur. The errors due to the approximations involved can be classified in this case by how they manifest themselves on wavelike solutions. They can lead to dispersion or dissipation.

I'll analyze the dispersion first. Take as initial data $u(t,x) = A\cos kx$ (or if you prefer, e^{ikx}). The exact solution will be $A\cos(kx-\omega t)$ where $\omega = ck$. Now analyze the effect of the numerical scheme. If Δx is very small, using the discrete values of Δx in the iteration give an approximate equation

$$u_t = -\frac{c}{2\Delta x}\big[u(t,x+\Delta x) - u(t,x-\Delta x)\big].$$

A power series expansion in Δx gives, for the first two non-vanishing terms

$$u_t = -c\left[u_x + \frac{1}{6}(\Delta x)^2 u_{xxx}\right]. \tag{66}$$

So, though I started off solving one equation, the numerical method more nearly represents quite a different equation. Try a solution of the form $A\cos(kx-\omega t)$ in this equation and you get

$$\omega = c\left[k - \frac{1}{6}(\Delta x)^2 k^3\right], \tag{67}$$

and you have dispersion of the wave. The velocity of the wave, ω/k, depends on k and so it depends on its wavelength or frequency.

The problem of instabilities is more conveniently analyzed by the use of an initial condition $u(0,x) = e^{ikx}$, then Eq. (65) is

$$\begin{aligned} u(\Delta t,x) &= e^{ikx} - \frac{c\Delta t}{2\Delta x}\left[e^{ik(x+\Delta x)} - e^{ik(x-\Delta x)}\right] \\ &= e^{ikx}\left[1 - \frac{ic\Delta t}{\Delta x}\sin k\Delta x\right]. \end{aligned} \tag{68}$$

The n-fold iteration of this, therefore involves only the n^{th} power of the bracketed expression; that's why the exponential form is easier to use in this case. If $k\Delta x$ is small, the first term in the expansion of the sine says that this is approximately

$$e^{ikx}\big[1 - ikc\Delta t\big]^n,$$

and with small Δt and $n = t/\Delta t$ a large number, this is

$$e^{ikx}\left[1 - \frac{ikct}{n}\right]^n \approx e^{ik(x-ct)}.$$

Looking more closely though, the object in brackets in Eq. (68) has magnitude

$$r = \left[1 + \frac{c^2(\Delta t)^2}{(\Delta x)^2}\sin^2 k\Delta x\right]^{1/2} > 1. \tag{69}$$

so the magnitude of the solution grows exponentially. This instability can be pictured as a kind of negative dissipation. This growth is reduced by requiring $kc\Delta t \ll 1$.

Given a finite fixed time interval, is it possible to get there with arbitrary accuracy by making Δt small enough? With n steps $= t/\Delta t$, r^n is

$$\begin{aligned} r &= \left[1 + \frac{c^2(\Delta t)^2}{(\Delta x)^2}\sin^2 k\Delta x\right]^{t/2\Delta t} \qquad = [1+\alpha]^\beta \\ &= \left[[1+\alpha]^{1/\alpha}\right]^{\alpha\beta} \approx e^{\alpha\beta} \\ &= \exp\left[\frac{c^2 t\Delta t}{2(\Delta x)^2}\sin^2 k\Delta x\right], \end{aligned}$$

so by shrinking Δt sufficiently, this is arbitrarily close to one.

There are several methods to avoid some of these difficulties. One is the Lax-Friedrichs method:

$$u(t+\Delta t, x) = \frac{1}{2}\big[u(t,x+\Delta x) + u(t,x-\Delta x)\big] - \frac{c\Delta t}{2\Delta x}\big[u(t,x+\Delta x) - u(t,x-\Delta x)\big]. \tag{70}$$

By appropriate choice of Δt and Δx, this will have $r \leq 1$, causing a dissipation of the wave. Another scheme is the Lax-Wendroff method.

$$\begin{aligned} u(t+\Delta t, x) = u(t,x) &- \frac{c\Delta t}{2\Delta x}\big[u(t,x+\Delta x) - u(t,x-\Delta x)\big] \\ &+ \frac{c^2(\Delta t)^2}{2(\Delta x)^2}\big[u(t,x+\Delta x) - 2u(t,x) + u(t,x-\Delta x)\big]. \end{aligned} \tag{71}$$

This keeps one more term in the power series expansion.

Problems

11.1 Show that a two point extrapolation formula is

$$f(0) \approx 2f(-h) - f(-2h) + h^2 f''(0).$$

11.2 Show that a three point extrapolation formula is

$$f(0) \approx 3f(-h) - 3f(-2h) + f(-3h) + h^3 f'''(0).$$

11.3 Solve $x^2 - a = 0$ by Newton's method, showing graphically that in this case, no matter what the initial guess is (positive or negative), the sequence will always converge. Draw graphs. Find $\sqrt{2}$. (This is the basis for the library square root algorithm on some computers.)

11.4 Find all real roots of $e^{-x} = \sin x$ to $\pm 10^{-4}$.

11.5 The first root r_1 of $e^{-ax} = \sin x$ is a function of the variable $a > 0$. Find dr_1/da at $a = 1$ by two means. First find r_1 for some values of a near 1 and use a four-point differentiation formula. Second, use analytical techniques on the equation to solve for dr_1/da and evaluate the derivative in terms of the known value of the root from the previous problem.

11.6 Evaluate $\operatorname{erf}(1) = \frac{2}{\sqrt{\pi}} \int_0^1 dt\, e^{-t^2}$

11.7 The principal value of an integral is $(a < x_0 < b)$

$$\mathsf{P}\int_a^b \frac{f(x)}{x - x_0}\,dx = \lim_{\epsilon\to 0}\left[\int_a^{x_0-\epsilon} \frac{f(x)}{x - x_0}\,dx + \int_{x_0+\epsilon}^b \frac{f(x)}{x - x_0}\,dx\right].$$

(a) Show that an equal spaced integration scheme to evaluate such an integral is (using points $0, \pm h$)

$$\mathsf{P}\int_{-h}^{+h} \frac{f(x)}{x}\,dx = f(h) - f(-h) - \frac{2}{9}h^3 f'''(0).$$

(b) Also, an integration scheme of the Gaussian type is

$$\sqrt{3}\big[f(h/\sqrt{3}) - f(-h/\sqrt{3})\big] + \frac{h^5}{675} f^{v}(0).$$

11.8 Devise a two point Gaussian integration with errors for the class of integrals

$$\int_{-\infty}^{+\infty} dx\, e^{-x^2} f(x).$$

Find what standard polynomial has roots at the points where f is to be evaluated.
Ans: $\frac{1}{2}\sqrt{\pi}\big[f(-h/\sqrt{2}) + f(h/\sqrt{2})\big]$

11.9 Same as the previous problem, but make it a three point method.

11.10 Find two and three point Gauss methods for

$$\int_0^\infty dx\, e^{-x} f(x).$$

What polynomials are involved here? Look up Laguerre.

11.11 In numerical differentiation it is possible to choose the interval *too* small. Every computation is done to a finite precision. (a) Do the simplest numerical differentiation of some specific function and take smaller and smaller intervals. What happens when the interval gets very small? (b) To analyze the reason for this behavior, assume that every number in the two point differentiation formula is kept to a fixed number of significant figures (perhaps 7 or 8). How does the error vary with the interval? What interval gives the most accurate answer? Compare this theoretical answer with the experimental value found in the first part of the problem.

11.12 The same phenomenon caused by roundoff errors occurs in integration. For any of the integration schemes discussed here, analyze the dependence on the number of significant figures kept and determine the most accurate interval. (Surprise?)

11.13 Compute the solution of $y' = 1 + y^2$ and check the numbers in the table where that example was given, (39).

11.14 If in the least square fit to a linear combination of functions, the result is constrained to pass through one point, so that $\sum \alpha_\mu f_\mu(x_0) = K$ is a requirement on the α's, show that the result becomes

$$a = C^{-1}\big[b + \lambda f_0\big],$$

where f_0 is the vector $f_\mu(x_0)$ and λ satisfies

$$\lambda\big\langle f_0, C^{-1} f_0\big\rangle = K - \big\langle f_0, C^{-1} b\big\rangle.$$

11.15 Find the variances in the formulas (8) and (10) for f', assuming noisy data. Ans: $\sigma^2/2h^2$, $65\sigma^2/72h^2$

11.16 Derive Eqs. (59), (60), and (61).

11.17 The Van der Pol equation arises in (among other places) nonlinear circuits and leads to self-exciting oscillations as in multi-vibrators

$$\frac{d^2x}{dt^2} - \epsilon(1 - x^2)\frac{dx}{dt} + x = 0.$$

Take $\epsilon = .3$ and solve subject to any non-zero initial conditions. Solve over many periods to demonstrate the development of the oscillations.

11.18 Find a formula for the numerical third derivative. Cf. (2.17)

11.19 The equation resulting from the secant method, Eq. (7), can be simplified by placing everything over a common denominator, $\big(f(x_2)-f(x_1)\big)$. Explain why this is a bad thing to do, how it can lead to inaccuracies.

11.20 Rederive the first Gauss integration formula Eq. (25) without assuming the symmetry of the result

$$\int_{-h}^{+h} f(x)\,dx \approx \alpha f(\beta)+\gamma f(\delta).$$

11.21 Derive the coefficients for the stable two-point Adams method.

11.22 By putting in one more parameter in the differentiation algorithm for noisy data, it is possible both to minimize the variance in f' *and* to eliminate the error terms in $h^2 f'''$. Find such a 6-point formula for the derivatives halfway between data points OR one for the derivatives at the data points (with errors and variance).

11.23 In the same spirit as the method for differentiating noisy data, how do you *interpolate* noisy data? That is, use some extra points to stabilize the interpolation against random variations in the data. To be specific, do a midpoint interpolation for equally spaced points. Compare the variance here to that in Eq. (3). Ans: $f(0)\approx[f(-3k)+f(-k)+f(k)+f(3k)]/4$, σ^2 is 4.8 times smaller

11.24 Find the dispersion resulting from the use of a four point formula for u_x in the numerical solution of the PDE $u_t+cu_x=0$.

11.25 Find the exact dispersion resulting from the equation

$$u_t=-c\big[u(t,x+\Delta x)-u(t,x-\Delta x)\big]/2\Delta x.$$

That is, don't do the series expansion on Δx.

11.26 Compute the dispersion and the dissipation in the Lax-Friedrichs and in the Lax-Wendroff methods.

11.27 In the simple iteration method of Eq. (69), if the grid points are denoted $x=m\Delta x$, $t=n\Delta t$, where n and m are integers ($-\infty<n,m<+\infty$), the result is a linear, constant-coefficient, partial difference equation. Solve subject to the initial condition

$$u(0,m)=e^{ikm\Delta x}.$$

11.28 Lobatto integration is like Gaussian integration, except that you require the end-points of the interval to be included in the sum. The interior points are left free. Three point Lobatto is the same as Simpson; find the four point Lobatto formula. The points found are roots of P'_{n-1}.

11.29 From the equation $y' = f(x,y)$, one derives $y'' = f_x + f f_y$. Derive a two point Adams type formula using the first and second derivatives, with error of order h^5 as for the standard four-point expression. This is useful when the analytic derivatives are easy. The form is

$$y(0) = y(-h) + \beta_1 y'(-h) + \beta_2 y'(-2h) + \gamma_1 y''(-h) + \gamma_2 y''(-2h)$$

Ans: $\beta_1 = -h/2$, $\beta_2 = 3h/2$, $\gamma_1 = 17h^2/12$, $\gamma_2 = 7h^2/12$

11.30 Using the same idea as in the previous problem, find a differential equation solver in the spirit of the original Euler method, (32), but doing a parabolic extrapolation instead of a linear one. That is, start from (x_0, y_0) and fit the initial data to $y = \alpha + \beta(x - x_0) + \gamma(x - x_0)^2$ in order to take a step. Ans: $y(h) = y_0 + hf(0, y_0) + \big(h^2/2\big)\big[f_x(0, y_0) + f_y(0, y_0)f(0, y_0)\big]$

11.31 Show that the root finding algorithm of Eq. (7) is valid for analytic functions of a complex variable with complex roots.

11.32 In the Runge-Kutta method, pick one of the other choices for the value of $D_2 f(0, y_0)$ in Eq. (37). How many function evaluations will it require at each step?

11.33 Sometimes you want an integral where the data is known outside the domain of integration. Find an integration scheme for $\int_0^h f(x)\,dx$ in terms of $f(h)$, $f(0)$, and $f(-h)$. Ans: $[-f(-h) + 8f(0) + 5f(h)]h/12$, error $\propto h^4$

11.34 When you must subtract two quantities that are almost the same size, you can find yourself trying to carry ridiculously many significant figures in intermediate steps. If a and b are very close and you want to evaluate $\sqrt{a} - \sqrt{b}$, devise an algorithm that does not necessitate carrying square roots out to many more places than you want in the final answer. Write $a = b + \epsilon$.
Ans: $\epsilon/2\sqrt{b}$, error: $\epsilon^2/8b^{3/2}$

11.35 Repeat the previous problem but in a more symmetric fashion. Write $a = x + \epsilon$ and $b = x - \epsilon$. Compare the sizes of the truncation errors. Ans: $\epsilon/\sqrt{x}$, $\;-\epsilon^3/8x^{5/2}$

11.36 The value of π was found in the notes by integrating $4/(1 + x^2)$ from zero to one using Simpson's rule and five points. Do the same calculation using Gaussian integration and two points. Ans: 3.14754

11.37 Derive Eq. (52).
(b) Explain why the plausibility arguments that follow it actually say something.

11.38 After you've done the Euclidean fit of data to a straight line and you want to do the data reduction described after Eq. (57), you have to find the coordinate along the line of the best fit to each point. This is essentially the problem: Given the line ($\vec u$ and $\hat v$) and a point ($\vec w$), the new reduced coordinate is the α in $\vec u + \alpha \hat v$ so that this point is closest to $\vec w$. What is it? You can do this the hard way, with a lot of calculus and algebra, or you can draw a picture and write the answer down.

11.39 Data is given as $(x_i, y_i) = \{(1,1),\,(2,2),\,(3,2)\}$. Compute the Euclidean best fit line. Also find the coordinates, α_i, along this line and representing the reduced data set.
Ans: $\vec{u} = (2, 5/3)$ $\hat{v} = (0.88167, 0.47186)$ $\alpha_1 = -1.1962$ $\alpha_2 = 0.1573$ $\alpha_3 = 1.0390$
The approximate points are $(0.945, 1.102)$, $(2.139, 1.741)$, $(2.916, 2.157)$
[It may not warrant this many significant figures, but it should make it easier to check your work.]

11.40 In the paragraph immediately following Eq. (23) there's mention of an alternate way to derive Simpson's rule. Carry in out, though you already know the answer.