Numerical Calculus

Reading: Chapter 5 ("Integrals and Derivatives"); Chapter 4 ("Accuracy and Speed")

1 Review of derivatives and integrals

Let's start out with a reminder of what derivatives and integrals represent.

1.1 Derivatives

The derivative of function $f(x)$, written $\frac{df}{dx}$ or $f'(x)$, is the rate of change of f with respect to x (the "slope" of the function). For a line, this rate of change is constant (same for every value of x), so you can take $\Delta f/\Delta x$ between any two points on the line and get the derivative. For functions that aren't linear, the rate of change depends on x, so to get the rate of change at a particular point, we consider the change df for an "infinitesimal" change in variable x (dx):

$$\frac{df}{dx} = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \tag{1}$$

As an easy example, the derivative of a linear function $f(x) = mx + b$:

$$\begin{aligned}
\frac{df}{dx} &= \lim_{h\to 0} \frac{m(x+h) + b - (mx+b)}{h} \\
&= \lim_{h\to 0} \frac{mx + mh + b - mx - b}{h} \\
&= \lim_{h\to 0} \frac{mh}{h} \\
&= \lim_{h\to 0} m \\
&= m
\end{aligned}$$

All the standard derivative formulas (that we know by memory) come from evaluating the

limit above. For example, the derivative of a polynomial, $f(x) = x^n$:

$$\begin{aligned}
\frac{df}{dx} &= \lim_{h\to 0} \frac{(x+h)^n - x^n}{h} \\
&= \lim_{h\to 0} \frac{(x^n + nx^{n-1}h + \frac{n(n-1)}{2}x^{n-2}h^2 + ...nxh^{n-1} + h^n - x^n)}{h} \\
&= \lim_{h\to 0} \frac{(nx^{n-1}h + \frac{n(n-1)}{2}x^{n-2}h^2 + ... + nxh^{n-1} + h^n)}{h} \\
&= \lim_{h\to 0} nx^{n-1} + \frac{n(n-1)}{2}x^{n-2}h + ... + nxh^{n-2} + h^{n-1} \\
&= nx^{n-1}
\end{aligned} \tag{2}$$

where we used the binominal theorem to expand $(x+h)^n$:

$$(x+h)^n = \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} x^{n-k}h^k \tag{3}$$

To do a derivative numerically for a known function f, we calculate

$$\frac{df}{dx} \approx \frac{f(x+h) - f(x)}{h} \tag{4}$$

where h is very small. Exactly how small h should be depends on how precisely we want to know the derivative. More on this later.

1.2 Integrals

Integration is the inverse of the operation of differentiation (up to a constant). An indefinite integral, an integral with no bounds, refers to the function $F(x)$ defined as

$$F(x) = \int f(x)dx \tag{5}$$

$F(x)$ is also called the antiderivative of $f(x)$. ($f(x)$ is of course the derivative of $F(x)$.) Indefinite integrals are defined based on what we know about differentiation, i.e. the derivative of x^2 is $2x$ by the power rule we derived above, so if you ask me to find the integral of $2x$, I know the answer is x^2.

The definite integral of a function represents the area beneath the curve between the bounds of the integral. The integral of function $f(x)$ from $x = a$ to $x = b$ is the area of the region bounded by f, the x-axis and vertical lines at $x = a$ and $x = b$, as shown in Figure 1. Note that sign matters – the area above the x-axis is positive and area below the x-axis is negative.

The fundamental theorem of calculus defines a definite integral in terms of the antiderivative:

$$\int_a^b f(x)dx = F(b) - F(a) \tag{6}$$

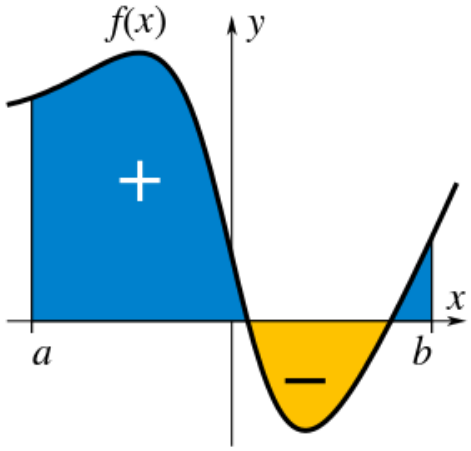


Figure 1: From: `https://en.wikipedia.org/wiki/Integral`

if $f(x)$ is continuous. This is usually referred to as performing an integral "analytically." However, not all integrals can be evaluated this way - you must know the antiderivative. (And I don't necessarily mean that you personally need to know the antiderivative. I mean that pretty much all continuous functions *have* an antiderivative, but it's not necessarily a function that can be expressed in terms of elementary functions, i.e. polynomials, trig functions, exponential functions, and so on. We'll see an example later.) This is when numerical integration is useful.

Definite integrals are commonly defined based on what's called the "Riemann sum". Basically, you divide up the region into intervals, approximate the area in that region with a rectangle, then add up the area of all the shapes to get the total. The more finely the region is divided, the closer the sum is to the true integral. The "Riemann integral" is therefore defined like

$$\int_a^b f(x)dx = \lim_{\Delta x \to 0} \sum_{i=1}^{n} f(x_i^*)\Delta x \tag{7}$$

where Δx is the size of each interval, n is the number of intevals between a and b, and x_i^* is some value of x in each interval so that $f(x_i^*)$ is the height of the rectangle and Δx is the width. You could make different choices for where x_i^* is located in the interval - left side? right side? center? - see Figure 2. Where n is small (and Δx is therefore large) the choice can make a big difference in the value of the sum, but as n gets large (and Δx is therefore small), sum approaches the true value of the integral, so the difference in the sum due to the different options becomes smaller.

Another way to do the Riemann sum is using the trapezoidal rule. In this case, the $f(x_i^*)$ in the equation above, which is the value of the function at some point in each interval, is replaced with the average of the function values at either end of the interval. Say the first

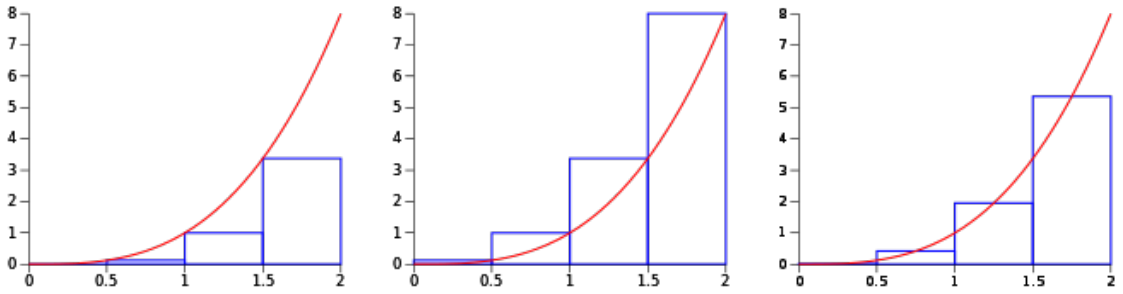


Figure 2: Using the left, right, and middle of the interval for the Riemann sum of x^3 from 0 to 2 using 4 subdivisions ($n = 4, \Delta x = 0.5$). From: `https://en.wikipedia.org/wiki/Riemann_sum`

interval extends from x_1 to x_2, where $x_2 - x_1 = \Delta x$. We take the average of the function values:

$$\begin{aligned} f_{avg} &= \frac{1}{2}\left(f(x_1) + f(x_2)\right) \\ &= \frac{1}{2}\left(f(x_1) + f(x_1 + \Delta x)\right) \end{aligned} \tag{8}$$

So the integral is given by

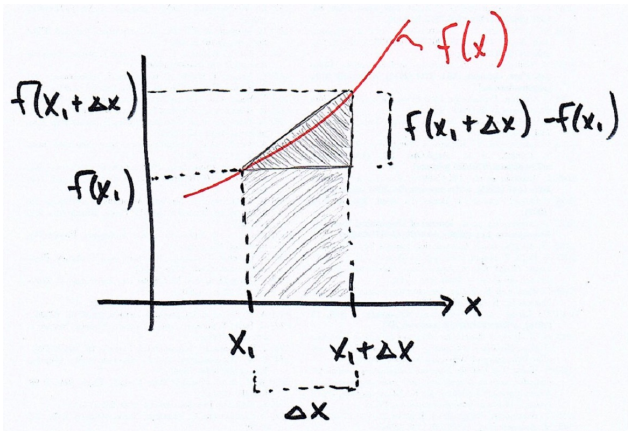
$$\int_a^b f(x)dx = \lim_{\Delta x \to 0} \sum_{i=1}^{n} \frac{1}{2}\left(f(x_1) + f(x_1 + \Delta x)\right)\Delta x \tag{9}$$

Let's look at each term of the sum and rearrange a bit:

$$\begin{aligned} \frac{1}{2}\left(f(x_1) + f(x_1 + \Delta x)\right)\Delta x &= \frac{1}{2}f(x_1)\Delta x + \frac{1}{2}f(x_1 + \Delta x)\Delta x \\ &= \left(f(x_1)\Delta x - \frac{1}{2}f(x_1)\Delta x\right) + \frac{1}{2}f(x_1 + \Delta x)\Delta x \\ &= f(x_1)\Delta x + \frac{1}{2}\left[f(x_1 + \Delta x) - f(x_1)\right]\Delta x \end{aligned} \tag{10}$$

The first term is the area of a rectangle with height $f(x_1)$ and width Δx. The second term is the area of a triangle with height $f(x_1 + \Delta x) - f(x_1)$ and base Δx. The sum of these two terms is the area of the trapezoid bounded by $x = x_1$, $x = x_1 + \Delta x$, the x-axis, and a straight line drawn between $f(x_1)$ and $f(x_1 + \Delta x)$, see the left side of Figure 3.The right figure in Figure 3 shows an illustration of the trapezoidal Riemann sum. This tends to be a better approximation to the integral than the rectangular sums shown in Figure 2. This is how we will do a basic numerical integral.

Section 2, Exercise 1



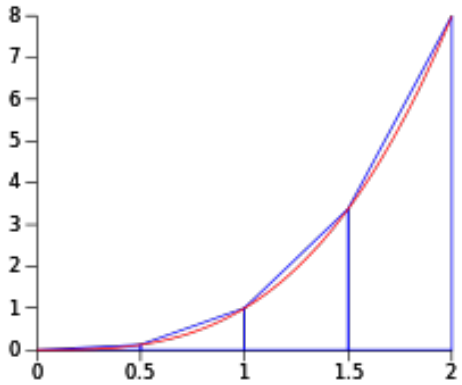


Figure 3: Left: Area of the trapezoid bounded by $x = x_1$, $x = x_1 + \Delta x$, the x-axis, and a straight line drawn between $f(x_1)$ and $f(x_1 + \Delta x)$. Right: The trapezoidal Riemann sum of x^3 from 0 to 2 using 4 subdivisions ($n = 4, \Delta x = 0.5$). From: `https://en.wikipedia.org/wiki/Riemann_sum`

2 Numerical Integrals

We're going to spend a good amount of time on numerical calculus, writing code to calculate numerical integrals using different methods. We do this not because I expect that you will use it much - there is a lot of existing software for this kind of thing. We do it because it's a good way to examine how algorithms work, how the errors can be estimated, how to write "searching" algorithms, "adaptive" algorithms, and other things.

2.1 Integrals using the Trapezoidal Rule

We saw above how to calculate the area of one interval. Now we'll sum up the areas for n intervals. Let's assume the intervals are evenly spaced between a and b, so that

$$\Delta x = (b - a)/n \tag{11}$$

and the left edge of i-th interval is defined by

$$x_i = a + (i - 1)\Delta x \tag{12}$$

so that

$$\begin{aligned} x_1 &= a \\ x_2 &= a + \Delta x \\ x_3 &= a + 2\Delta x \\ &\text{etc} \\ x_{n-1} &= a + (n-2)\Delta x \\ x_n &= a + (n-1)\Delta x \\ &= a + n\Delta x - \Delta x \\ &= a + (b-a) - \Delta x \\ &= b - \Delta x \end{aligned}$$

The approximation to the integral is therefore

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} \frac{1}{2}\left(f(x_i) + f(x_i + \Delta x)\right)\Delta x \tag{13}$$

Example: First Numerical Integral
Example: Integral With Negative Values

Writing out the sum and rearranging:

$$\begin{aligned} \int_a^b f(x)dx &\approx \sum_{i=1}^{n} \frac{1}{2}\left(f(x_i) + f(x_i + \Delta x)\right)\Delta x \\ &= \frac{1}{2}\Delta x\left[f(x_1) + f(x_1 + \Delta x) + f(x_2) + f(x_2 + \Delta x) + f(x_3) + f(x_3 + \Delta x)\right. \\ &\quad \left. + \ ... \ + f(x_{n-1}) + f(x_{n-1} + \Delta x) + f(x_n) + f(x_n + \Delta x)\right] \\ &\approx \frac{1}{2}\Delta x\left[f(a) + f(a + \Delta x) + f(a + \Delta x) + f(a + 2\Delta x) + f(a + 2\Delta x) + f(a + 3\Delta x)\right] \\ &\quad + \ ... \ + f(a + (n-2)\Delta x) + f(a + (n-1)\Delta x) + f(a + (n-1)\Delta x) + f(a + n\Delta x)\right] \\ &\approx \frac{1}{2}\Delta x\left[f(a) + 2f(a + \Delta x) + 2f(a + 2\Delta x) + \ ... \ 2f(a + (n-1)\Delta x) + f(a + n\Delta x)\right] \\ &\approx \frac{1}{2}\Delta x\left[f(a) + f(a + n\Delta x)\right] + \Delta x\left(\ f(a + \Delta x) + f(a + 2\Delta x) + \ ... \ + f(a + (n-1)\Delta x)\right) \end{aligned}$$

$$\int_a^b f(x)dx \approx \frac{1}{2}\Delta x\left[f(a) + f(b)\right] + \Delta x \sum_{k=1}^{n-1} f(a + k\Delta x) \tag{14}$$

The trapezoidal sum is typically written this way.

2.1.1 Integrating a Gaussian Function

Recall the Gaussian probability density function (PDF) that we discussed previously:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{15}$$

where μ is the mean and σ is the standard deviation. This represents the probability per unit length in the variable x, i.e.

$$P(a < x < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} dx \tag{16}$$

This is an example of an integral that cannot be done analytically. (The antiderivative of e^{-x^2} is not a known function.)

If we integrate this function over all x, the result should be 1 (the probability that x takes some value between $-\infty$ and ∞ is 1).

$$P(-\infty < x < \infty) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} dx = 1 \tag{17}$$

There is a "cute" way to prove this integral is 1. Let I be the integral of the function e^{-x^2} over all x:

$$\begin{aligned}
I &= \int_{-\infty}^{\infty} e^{-x^2} dx \\
I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right)^2 \\
&= \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right) \left(\int_{-\infty}^{\infty} e^{-y^2} dy \right) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2} e^{-y^2} dx dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy
\end{aligned}$$

Use cylindrical or polar coordinates, with radius s and polar angle ϕ. $s^2 = x^2 + y^2$, and the area element is $dxdy = sdsd\phi$. To integrate over all space, s goes from 0 to ∞ and ϕ goes from 0 to 2π:

$$\begin{aligned}
I^2 &= \int_0^{2\pi} \int_0^{\infty} e^{-s^2} s ds d\phi \\
&= \int_0^{2\pi} d\phi \int_0^{\infty} e^{-s^2} s ds \\
&= (2\pi) \int_0^{\infty} e^{-s^2} s ds
\end{aligned}$$

Let $u = s^2$, so $du = 2sds$. If $s = 0$ then $u = 0$, and if $s \to \infty$ then $u \to \infty$, so the limits

don't change.

$$\begin{aligned} I^2 &= (2\pi)\int_0^\infty e^{-u}\frac{du}{2} \\ &= (2\pi)\frac{1}{2}\left(-e^{-u}\right)\Big|_0^\infty \\ &= -(\pi)\left(e^{-\infty}-e^0\right) \\ &= -(\pi)(0-1) \\ I^2 &= \pi \\ I &= \sqrt{\pi} \end{aligned} \tag{18}$$

Therefore

$$\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi} \tag{19}$$

Given that, we can do the Gaussian integral with some simple substitutions:

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}dx \tag{20}$$

Let $y = (x-\mu)/(\sigma\sqrt{2})$, so that $dy = dx/(\sigma\sqrt{2})$ and $y^2 = (x-\mu)^2/(2\sigma^2)$. If $x \to \pm\infty$ then $y \to \pm\infty$, so the integration limits don't change.

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}dx &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-y^2}(dy\sigma\sqrt{2}) \\ &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-y^2}dy \\ &= \frac{1}{\sqrt{\pi}}(\sqrt{\pi}) \\ &= 1 \end{aligned} \tag{21}$$

That's nice to see, but it's not useful for integrating over a limited range in x. (The switch to cylindrical coordinates is only useful when you are integrating over all space. When trying to integrate over a limited range of x, you end up trying to describe a square region in cylindrical coordinates.)

Let's integrate a Gaussian PDF numerically.
Example: Integrating A Gaussian

This integral comes up so frequently that it's given a special name, the error function:

$$\mathrm{erf}(y) = \frac{2}{\sqrt{\pi}}\int_0^y e^{-x^2}dx \tag{22}$$

We'll revisit Gaussian functions several times this semester.

Section 2, Exercise 2

2.2 Simpson's Rule

Another way to think of the trapezoidal rule is this: it approximates the integral of a function between two points by finding the line between those two points and finding the area beneath the line. Consider the line connecting $f(x_1)$ and $f(x_1 + \Delta x)$ in Figure 3. The area beneath this line can be found by integrating the line from x_1 to $x_1 + \Delta x$:

$$\begin{aligned} I &= \int_{x_1}^{x_1+\Delta x} \ell(x)dx \\ &= \int_{x_1}^{x_1+\Delta x} (mx+b)dx \\ &= \left(\frac{1}{2}mx^2 + bx\right)\Big|_{x_1}^{x_1+\Delta x} \\ &= \frac{1}{2}m(x_1+\Delta x)^2 + b(x_1+\Delta x) - \left(\frac{1}{2}mx_1^2 + bx_1\right) \\ &= \frac{1}{2}m(x_1^2 + 2x_1\Delta x + \Delta x^2) + b(x_1+\Delta x) - \left(\frac{1}{2}mx_1^2 + bx_1\right) \\ I &= mx_1\Delta x + \frac{1}{2}m\Delta x^2 + b\Delta x \end{aligned} \tag{23}$$

where $\ell(x)$ is the equation of the line between two points. Two constants are need to specify the equation of the line (m and b). Once we know those, the above equation will give the area we want to know (the true area beneath the line, which we use as an approximation for the area underneath the function $f(x)$).

Let $f_1 = f(x_1)$ and $f_2 = f(x_1 + \Delta x)$. We have two points on the function with coordinates (x_1, f_1) and $(x_1 + \Delta x, f_2)$. We want to find the equation of the line $\ell(x)$ that pass through these two points. By plugging in these two points in $\ell(x)$, we can solve for m and b:

$$\begin{aligned} f(x_1) &= \ell(x_1) \\ f_1 &= mx_1 + b \end{aligned} \tag{24}$$

$$\begin{aligned} f(x_1+\Delta x) &= \ell(x_1+\Delta x) \\ f_2 &= m(x_1+\Delta x) + b \end{aligned} \tag{25}$$

Subtracting the two equations:

$$\begin{aligned} f_1 - f_2 &= mx_1 + b - (m(x_1+\Delta x) + b) \\ f_1 - f_2 &= -m\Delta x \\ m &= \frac{f_2 - f_1}{\Delta x} \end{aligned} \tag{26}$$

Plugging in m and solving for b:

$$\begin{aligned} f_1 &= \frac{(f_2 - f_1)}{\Delta x}x_1 + b \\ b &= f_1 - \frac{(f_2 - f_1)}{\Delta x}x_1 \end{aligned} \tag{27}$$

Now that we have m and b, we can plug back in to our formula for I:

$$\begin{aligned}
I &= mx_1\Delta x + \frac{1}{2}m\Delta x^2 + b\Delta x \\
&= \frac{(f_2 - f_1)}{\Delta x}x_1\Delta x + \frac{1}{2}\frac{(f_2 - f_1)}{\Delta x}\Delta x^2 + \left(f_1 - \frac{(f_2 - f_1)}{\Delta x}x_1\right)\Delta x \\
&= (f_2 - f_1)x_1 + \frac{1}{2}(f_2 - f_1)\Delta x + f_1\Delta x - (f_2 - f_1)x_1 \\
&= \frac{1}{2}(f_2 - f_1)\Delta x + f_1\Delta x \\
&= \frac{1}{2}(f_1 + f_2)\Delta x \\
&= \frac{1}{2}\left(f(x_1) + f(x_1 + \Delta x)\right)\Delta x
\end{aligned} \tag{28}$$

This is the same result we got previously, by approximating the area of a slice using the average function value between two points; we showed it is equal to the area underneath the line (the trapezoid) in Equation 10.

Simpson's rule does the same thing, but with a quadratic function instead of a linear one. Simpson's rule approximates the integral of some function $f(x)$ by finding the appropriate quadratic function and calculating the area beneath the quadratic function. A quadratic function has three unknown parameters ($q(x) = Ax^2 + Bx + C$, where A, B, and C are the constants to be determined), so we need three points of our function $f(x)$.

First of all, the area beneath a quadratic function $q(x)$ between $x = a - \Delta x$ and $x = a + \Delta x$ can be found by integrating $q(x)$ between the two points analytically:

$$\begin{aligned}
I &= \int_{a-\Delta x}^{a+\Delta x} \left(Ax^2 + Bx + C\right) \\
&= \left(\frac{1}{3}Ax^3 + \frac{1}{2}Bx^2 + Cx\right)\Big|_{a-\Delta x}^{a+\Delta x} \\
&= \frac{1}{3}A(a + \Delta x)^3 + \frac{1}{2}B(a + \Delta x)^2 + C(a + \Delta x) \\
&\quad - \left(\frac{1}{3}A(a - \Delta x)^3 + \frac{1}{2}B(a - \Delta x)^2 + C(a - \Delta x)\right) \\
&= \frac{1}{3}A(a^3 + 3a^2\Delta x + 3a\Delta x^2 + \Delta x^3) + \frac{1}{2}B(a^2 + 2a\Delta x + \Delta x^2) + Ca + C\Delta x \\
&\quad - \left(\frac{1}{3}A(a^3 - 3a^2\Delta x + 3a\Delta x^2 - \Delta x^3) + \frac{1}{2}B(a^2 - 2a\Delta x + \Delta x^2) + Ca - C\Delta x\right) \\
&= \frac{1}{3}Aa^3 + Aa^2\Delta x + Aa\Delta x^2 + \frac{1}{3}A\Delta x^3 + \frac{1}{2}Ba^2 + Ba\Delta x + \frac{1}{2}B\Delta x^2 + Ca + C\Delta x \\
&\quad - \frac{1}{3}Aa^3 + Aa^2\Delta x - Aa\Delta x^2 + \frac{1}{3}A\Delta x^3 - \frac{1}{2}Ba^2 + Ba\Delta x - \frac{1}{2}B\Delta x^2 - Ca + C\Delta x \\
&= 2Aa^2\Delta x + \frac{2}{3}A\Delta x^3 + 2Ba\Delta x + 2C\Delta x
\end{aligned} \tag{29}$$

In what follows, let $f_1 = f(a - \Delta x)$, $f_2 = f(a)$, and $f_3 = f(a + \Delta x)$ to make the algebra easier on the eyes.

Assume we know three points on the function $f(x)$ with coordinates $(a - \Delta x, f_1)$, (a, f_2) and $(a + \Delta x, f_3)$. We want to find the quadratic function $q(x)$ that passes through these three points. By plugging in these three points in $q(x) = Ax^2 + Bx + C$, we get three equations that we can use to solve for the three constants A, B, C. Once we know the constants A B, and C in terms of known quantities $a, \Delta x, f_1, f_2, f_3$, then Equation 29 above will give the area we want to know (the true area beneath $q(x)$, which we use as an approximation for the area underneath the function $f(x)$).

$$\begin{aligned} f(a - \Delta x) &= q(a - \Delta x) \\ f_1 &= A(a - \Delta x)^2 + B(a - \Delta x) + C \\ f_1 &= Aa^2 - 2Aa\Delta x + A\Delta x^2 + Ba - B\Delta x + C \end{aligned} \tag{30}$$

$$\begin{aligned} f(a) &= q(a) \\ f_2 &= Aa^2 + Ba + C \end{aligned} \tag{31}$$

$$\begin{aligned} f(a + \Delta x) &= q(a + \Delta x) \\ f_3 &= A(a + \Delta x)^2 + B(a + \Delta x) + C \\ f_3 &= Aa^2 + 2Aa\Delta x + A\Delta x^2 + Ba + B\Delta x + C \end{aligned} \tag{32}$$

Then

$$\begin{aligned} f_2 - f_1 &= Aa^2 + Ba + C - (Aa^2 - 2Aa\Delta x + A\Delta x^2 + Ba - B\Delta x + C) \\ &= 2Aa\Delta x - A\Delta x^2 + B\Delta x \end{aligned} \tag{33}$$

$$\begin{aligned} f_3 - f_2 &= Aa^2 + 2Aa\Delta x + A\Delta x^2 + Ba + B\Delta x + C - (Aa^2 + Ba + C) \\ &= 2Aa\Delta x + A\Delta x^2 + B\Delta x \end{aligned} \tag{34}$$

And

$$\begin{aligned} f_2 - f_1 - (f_3 - f_2) &= -2A\Delta x^2 \\ f_2 - f_1 - f_3 + f_2 &= -2A\Delta x^2 \\ 2f_2 - f_1 - f_3 &= -2A\Delta x^2 \\ A &= \frac{1}{2\Delta x^2}(f_1 + f_3 - 2f_2) \end{aligned} \tag{35}$$

Now to find B and C:

$$\begin{aligned} f_3 - f_1 &= Aa^2 + 2Aa\Delta x + A\Delta x^2 + Ba + B\Delta x + C \\ &\quad - (Aa^2 - 2Aa\Delta x + A\Delta x^2 + Ba - B\Delta x + C) \\ f_3 - f_1 &= 4Aa\Delta x + 2B\Delta x \\ \frac{1}{2\Delta x}(f_3 - f_1) &= 2Aa + B \\ B &= \frac{1}{2\Delta x}(f_3 - f_1) - 2Aa \\ B &= \frac{1}{2\Delta x}(f_3 - f_1) - \frac{2a}{2\Delta x^2}(f_1 + f_3 - 2f_2) \\ B &= \frac{1}{2\Delta x}(f_3 - f_1) - \frac{a}{\Delta x^2}(f_1 + f_3 - 2f_2) \end{aligned} \tag{36}$$

$$\begin{aligned} f_2 &= Aa^2 + Ba + C \\ C &= f_2 - Aa^2 - Ba \\ &= f_2 - \frac{a^2}{2\Delta x^2}(f_1 + f_3 - 2f_2) - \left(\frac{a}{2\Delta x}(f_3 - f_1) - \frac{a^2}{\Delta x^2}(f_1 + f_3 - 2f_2)\right) \\ &= f_2 - \frac{a^2}{2\Delta x^2}(f_1 + f_3 - 2f_2) - \frac{a}{2\Delta x}(f_3 - f_1) + \frac{a^2}{\Delta x^2}(f_1 + f_3 - 2f_2) \\ C &= f_2 + \frac{a^2}{2\Delta x^2}(f_1 + f_3 - 2f_2) - \frac{a}{2\Delta x}(f_3 - f_1) \end{aligned} \tag{37}$$

Now that we have the values for A, B, C, we can plug back in to our formula for the integral of a quadratic function:

$$\begin{aligned} I &= 2Aa^2\Delta x + \frac{2}{3}A\Delta x^3 + 2Ba\Delta x + 2C\Delta x \\ &= \frac{2a^2\Delta x}{2\Delta x^2}(f_1 + f_3 - 2f_2) + \frac{2}{3}\Delta x^3\frac{1}{2\Delta x^2}(f_1 + f_3 - 2f_2) + 2a\Delta x\left(\frac{1}{2\Delta x}(f_3 - f_1) - \frac{a}{\Delta x^2}(f_1 + f_3 - 2f_2)\right) \\ &\quad + 2\Delta x\left(f_2 + \frac{a^2}{2\Delta x^2}(f_1 + f_3 - 2f_2) - \frac{a}{2\Delta x}(f_3 - f_1)\right) \\ &= \frac{a^2}{\Delta x}(f_1 + f_3 - 2f_2) + \frac{\Delta x}{3}(f_1 + f_3 - 2f_2) + \frac{2a\Delta x}{2\Delta x}(f_3 - f_1) - (2a\Delta x)\frac{a}{\Delta x^2}(f_1 + f_3 - 2f_2) \\ &\quad + 2\Delta x f_2 + (2\Delta x)\frac{a^2}{2\Delta x^2}(f_1 + f_3 - 2f_2) - (2\Delta x)\frac{a}{2\Delta x}(f_3 - f_1) \\ &= \frac{a^2}{\Delta x}(f_1 + f_3 - 2f_2) + \frac{\Delta x}{3}(f_1 + f_3 - 2f_2) + a(f_3 - f_1) - \frac{2a^2}{\Delta x}(f_1 + f_3 - 2f_2) \\ &\quad + 2\Delta x f_2 + \frac{a^2}{\Delta x}(f_1 + f_3 - 2f_2) - a(f_3 - f_1) \\ &= \frac{\Delta x}{3}(f_1 + f_3 - 2f_2) + 2\Delta x f_2 \\ &= \frac{\Delta x}{3}(f_1 + f_3 - 2f_2 + 6f_2) \\ I &= \frac{\Delta x}{3}(f_1 + f_3 + 4f_2) \end{aligned} \tag{38}$$

Writing it in terms of the function values we have

$$I = \frac{\Delta x}{3}\left(f(a-\Delta x) + 4f(a) + f(a+\Delta x)\right) \tag{39}$$

Figure 4 shows the function $f(x) = \sin(x)$. In the shaded region, between $x = \pi/4$ and $x = \pi/2$, both a linear function (blue) and a quadratic function (red) are used to approximate the curve. Clearly the area of the shaded region calculated with the quadratic function will be a much better approximation than the area calculated with the blue line.

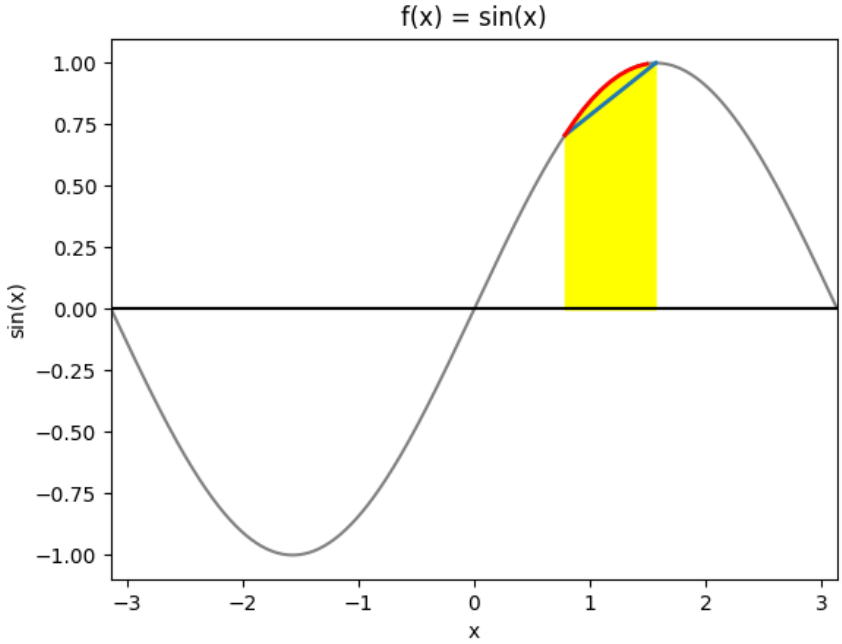


Figure 4: Approximating $f(x) = \sin(x)$ between $x = \pi/4$ and $x = \pi/2$ (the yellow region) using a linear function (blue) and a quadratic function (red).

Equation 39 gives us the area for two adjacent slices of a function. Suppose we want to integrate a function from $x = a$ to $x = b$ using Simpson's rule with two slices. The integral would be

$$\int_a^b f(x)dx \approx \frac{\Delta x}{3}\left(f(a) + 4f(a+\Delta x) + f(b)\right) \tag{40}$$

where $\Delta x = (b-a)/2$, i.e. $b = a + \Delta x$.

For four slices, the integral would be

$$\int_a^b f(x)dx \approx \frac{\Delta x}{3}\left(f(a) + 4f(a+\Delta x) + f(a+2\Delta x)\right) + \frac{\Delta x}{3}\left(f(a+2\Delta x) + 4f(a+3\Delta x) + f(b)\right) \tag{41}$$

where $\Delta x = (b-a)/4$, i.e. $b = a + 4\Delta x$.

The total number of slices must be even, since the calculation is done in pairs of (non-overlapping) slices. For n total slices (where n is even), the integral would be

$$\begin{aligned}\int_a^b f(x)dx &\approx \frac{\Delta x}{3}\left(f(a) + 4f(a+\Delta x) + f(a+2\Delta x)\right) + \frac{\Delta x}{3}\left(f(a+2\Delta x) + 4f(a+3\Delta x) + f(a+4\Delta x)\right) \\ &\quad + \ldots + \frac{\Delta x}{3}\left(f(a+(n-2)\Delta x) + 4f(a+(n-1)\Delta x) + f(b)\right) \\ &\approx \frac{\Delta x}{3}\left[f(a) + f(b) + \sum_{i=1,3,5,\ldots}^{n-1} 4f(a+i\Delta x) + \sum_{i=2,4,6,\ldots}^{n-2} 2f(a+i\Delta x)\right] \end{aligned} \tag{42}$$

Example: Simpson's Rule

2.3 Calculating an Electric Field by Integrating the Charge Distribution

Consider this problem, which should look familiar from Physics 2:

Find the electric field of a straight line segment of length L that carries a uniform line charge density (charge per unit length) λ. We'll assume the wire lies along the x axis, centered at the origin. We're going to restrict ourselves to 2D; we'll assume the wire lies in the $z = 0$ plane and only calculate the field in that plane. Figure 5 illustrates the problem.

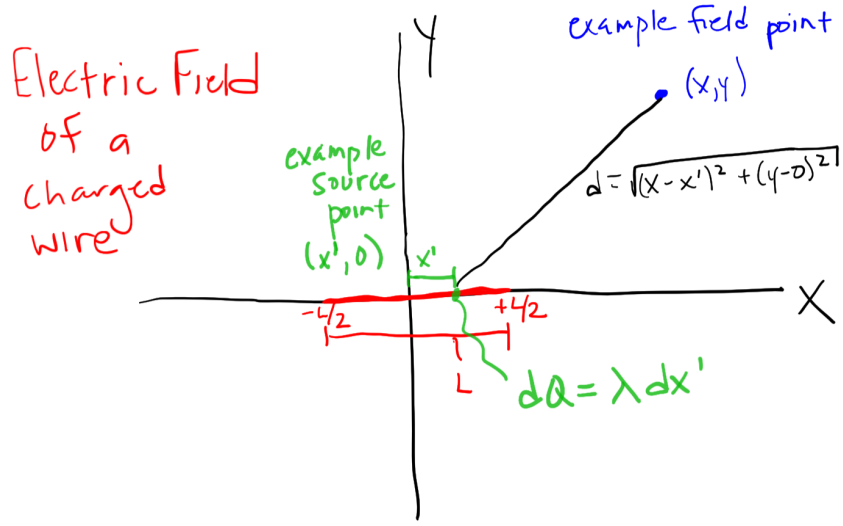


Figure 5:

Recall the electric field of a point charge Q located at the origin is given by

$$\vec{E} = \frac{1}{4\pi\epsilon_0}\frac{Q}{r^2}\,\hat{\mathrm{r}} \tag{43}$$

An infinitesimal bit of charge dQ along the wire will therefore create electric field

$$d\vec{E} = \frac{1}{4\pi\epsilon_0}\frac{dQ}{d^2}\ \hat{\mathrm{e}} \tag{44}$$

where d is the distance from dQ to the observation point and $\hat{\mathrm{e}}$ indicates the direction of the field. Given that the charge per unit length is constant along the wire and equal to λ, then a tiny bit of charge corresponds to a tiny length of the wire dx', i.e. $dQ = \lambda dx'$. The distance d is $d = \sqrt{(x-x')^2 + y^2}$, where (x, y) is the field point and x' is the location of the charge dQ. Therefore we have

$$d\vec{E} = \frac{1}{4\pi\epsilon_0}\frac{\lambda dx'}{(x-x')^2 + y^2}\ \hat{\mathrm{e}} \tag{45}$$

The direction of the electric field is in the direction connecting the tiny bit of charge and the field point. Note that $\hat{\mathrm{e}}$ is a unit vector.

$$\hat{\mathrm{e}} = \frac{(x-x')\ \hat{\mathrm{x}} + y\ \hat{\mathrm{y}}}{\sqrt{(x-x')^2 + y^2}} \tag{46}$$

so we have

$$\begin{aligned} d\vec{E} &= \frac{1}{4\pi\epsilon_0}\frac{\lambda dx'}{(x-x')^2 + y^2}\frac{(x-x')\ \hat{\mathrm{x}} + y\ \hat{\mathrm{y}}}{\sqrt{(x-x')^2 + y^2}} \\ &= \frac{\lambda}{4\pi\epsilon_0}\frac{(x-x')\ \hat{\mathrm{x}} + y\ \hat{\mathrm{y}}}{((x-x')^2 + y^2)^{3/2}}dx' \end{aligned} \tag{47}$$

Breaking this down into its x and y components:

$$dE_x = \frac{\lambda}{4\pi\epsilon_0}\frac{x-x'}{((x-x')^2 + y^2)^{3/2}}dx' \tag{48}$$

$$dE_y = \frac{\lambda}{4\pi\epsilon_0}\frac{y}{((x-x')^2 + y^2)^{3/2}}dx' \tag{49}$$

To find both the x and y components of the field $\vec{E}$, we integrate:

$$E_x(x,y) = \int dE_x = \int_{-L/2}^{L/2}\frac{\lambda}{4\pi\epsilon_0}\frac{x-x'}{((x-x')^2 + y^2)^{3/2}}dx' \tag{50}$$

$$E_y(x,y) = \int dE_y = \int_{-L/2}^{L/2}\frac{\lambda}{4\pi\epsilon_0}\frac{y}{((x-x')^2 + y^2)^{3/2}}dx' \tag{51}$$

where the integral is along the wire, represented by taking the source location x' from $-L/2$ to $L/2$. This gives us both vector components of the field at the point (x, y). To map out the entire field, we choose a grid of (x, y) points and perform this integral for every point.

For a real example, let $L = 10$ cm and $\lambda = 10^{-6}$ C/m. We'll use the trapezoidal rule for simplicity. We can calculate each component of the vector $\vec{E}$ at every point (x, y) by numerically integrating the equations above. Then we can visualize the vector field by drawing an arrow at each point representing the magnitude and direction of the field.

Example: Electric Field of a Wire

What happens when L is large compared to the distance between the wire and the point (x, y)? What happens when L is small compared to the distance between the wire and the point (x, y)?

Typically, when you do this problem in Physics 2 or in an upper-level E&M course, you calculate the field above the midpoint or the wire, or you assume the wire is infinitely long, to simplify the calculations. But it's a breeze to calculate the electric field numerically.

Section 2, Exercise 3

2.4 Errors on Numerical Integrals

Now we want to derive an expression for the error on our numerical integrals, i.e. how different is the area approximated with the trapezoidal rule or Simpson's rule from the true area under the curve?

Let's start with the trapezoidal rule:

$$\int_a^b f(x)dx \approx \frac{1}{2}\Delta x\left[f(a) + f(b)\right] + \Delta x \sum_{k=1}^{n-1} f(a + k\Delta x) \tag{52}$$

Let $x_k = a + k\Delta x$ so that x_{k-1} and x_k are the boundaries of a slice used in the trapezoidal rule. Let's perform a Taylor expansion of $f(x)$ about $x = x_{k-1}$:

$$\begin{aligned} f(x) &= \sum_{i=0}^{\infty} \frac{f^{(i)}(x_{k-1})}{i!}(x - x_{k-1})^i \\ &= f(x_{k-1}) + f'(x_{k-1})(x - x_{k-1}) + \frac{1}{2}f''(x_{k-1})(x - x_{k-1})^2 + ... \end{aligned} \tag{53}$$

where $f^{(i)}$ is the ith derivative of f.

We can use this series approximation to integrate $f(x)$ from x_{k-1} to x_k:

$$\begin{aligned} \int_{x_{k-1}}^{x_k} f(x)dx &= \int_{x_{k-1}}^{x_k} \left(f(x_{k-1}) + f'(x_{k-1})(x - x_{k-1}) + \frac{1}{2}f''(x_{k-1})(x - x_{k-1})^2 + ... \right) dx \\ &= f(x_{k-1}) \int_{x_{k-1}}^{x_k} dx + f'(x_{k-1}) \int_{x_{k-1}}^{x_k} (x - x_{k-1})dx + \frac{1}{2}f''(x_{k-1}) \int_{x_{k-1}}^{x_k} (x - x_{k-1})^2 dx + ... \end{aligned}$$

Note that the integration variable is x, while x_{k-1}, x_k, and the function and derivatives evaluated at $x = x_{k-1}$, $f(x_{k-1}), f'(x_{k-1})$ etc, are all constant under the integration. Let $u = x - x_{k-1}$, which means $du = dx$, $x = x_{k-1}$ corresponds to $u = 0$, and $x = x_k$ corresponds to $u = x_k - x_{k-1} = a + k\Delta x - (a + (k-1)\Delta x) = \Delta x$:

$$\begin{aligned}
\int_{x_{k-1}}^{x_k} f(x)dx &= f(x_{k-1})\int_0^{\Delta x} du + f'(x_{k-1})\int_0^{\Delta x} u du + \frac{1}{2}f''(x_{k-1})\int_0^{\Delta x} u^2 du + ... \\
&= f(x_{k-1})\Delta x + f'(x_{k-1})\frac{1}{2}\Delta x^2 + \frac{1}{2}f''(x_{k-1})\frac{1}{3}\Delta x^3 + ... \\
\int_{x_{k-1}}^{x_k} f(x)dx &= f(x_{k-1})\Delta x + \frac{1}{2}f'(x_{k-1})\Delta x^2 + \frac{1}{6}f''(x_{k-1})\Delta x^3 + ... \qquad (54)
\end{aligned}$$

Now we do a Taylor expansion of $f(x)$ about $x = x_k$, and use it to approximate the same integral. This time we let $u = x - x_k$, which means $du = dx$, $x = x_{k-1}$ corresponds to $u = x_{k-1} - x_k = a + (k-1)\Delta x - (a + k\Delta x) = -\Delta x$, and $x = x_k$ corresponds to $u = 0$:

$$\begin{aligned}
\int_{x_{k-1}}^{x_k} f(x)dx &= \int_{x_{k-1}}^{x_k} \left(f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2}f''(x_k)(x - x_k)^2 + ... \right) dx \\
&= f(x_k)\int_{x_{k-1}}^{x_k} dx + f'(x_k)\int_{x_{k-1}}^{x_k} (x - x_k)dx + \frac{1}{2}f''(x_k)\int_{x_{k-1}}^{x_k} (x - x_k)^2 dx + ... \\
&= f(x_k)\int_{-\Delta x}^{0} du + f'(x_k)\int_{-\Delta x}^{0} u du + \frac{1}{2}f''(x_k)\int_{-\Delta x}^{0} u^2 du + ... \\
&= f(x_k)(0 - (-\Delta x)) + f'(x_k)\left(0 - \frac{1}{2}(-\Delta x)^2\right) + \frac{1}{2}f''(x_k)\left(0 - \frac{1}{3}(-\Delta x)^3\right) + ... \\
\int_{x_{k-1}}^{x_k} f(x)dx &= f(x_k)\Delta x - \frac{1}{2}f'(x_k)\Delta x^2 + \frac{1}{6}f''(x_k)\frac{1}{3}\Delta x^3 + ... \qquad (55)
\end{aligned}$$

Now let's sum Equations 54 and 55:

$$\begin{aligned}
2\int_{x_{k-1}}^{x_k} f(x)dx &= \left(f(x_{k-1})\Delta x + \frac{1}{2}f'(x_{k-1})\Delta x^2 + \frac{1}{6}f''(x_{k-1})\Delta x^3 + ... \right) \\
&\quad + \left(f(x_k)\Delta x - \frac{1}{2}f'(x_k)\Delta x^2 + \frac{1}{6}f''(x_k)\frac{1}{3}\Delta x^3 + ... \right) \\
2\int_{x_{k-1}}^{x_k} f(x)dx &= \left(f(x_{k-1}) + f(x_k)\right)\Delta x + \frac{1}{2}\left(f'(x_{k-1}) - f'(x_k)\right)\Delta x^2 + \frac{1}{6}\left(f''(x_{k-1}) + f''(x_k)\right)\Delta x^3 + ... \\
\int_{x_{k-1}}^{x_k} f(x)dx &= \frac{1}{2}\left(f(x_{k-1}) + f(x_k)\right)\Delta x + \frac{1}{4}\left(f'(x_{k-1}) - f'(x_k)\right)\Delta x^2 + \frac{1}{12}\left(f''(x_{k-1}) + f''(x_k)\right)\Delta x^3 + ... \\
&\qquad (56)
\end{aligned}$$

This gives us an expression for the integral between $x = x_{k-1}$ and $x = x_k$. To find the integral between $x = a$ and $x = b$, we sum over all k (recalling that $x_k = a + k\Delta x$, so that $k = 1$ gives the interval from $x = x_{1-1} = a + (1-1)\Delta x = a$ to $x = x_1 = a + \Delta x$, and $k = n$

gives the interval from $x = x_{n-1} = a + (n-1)\Delta x = b - \Delta x$ to $x = x_n = a + n\Delta x = b$):

$$\begin{aligned}
\int_a^b f(x)dx &= \sum_{k=1}^{n} \left[\frac{1}{2} \left(f(x_{k-1}) + f(x_k) \right) \Delta x + \frac{1}{4} \left(f'(x_{k-1}) - f'(x_k) \right) \Delta x^2 \right. \\
&\quad \left. + \frac{1}{12} \left(f''(x_{k-1}) + f''(x_k) \right) \Delta x^3 + ... \right] \\
&= \frac{1}{2} \left(f(a) + f(a+\Delta x) \right) \Delta x + \frac{1}{4} \left(f'(a) - f'(a+\Delta x) \right) \Delta x^2 + \frac{1}{12} \left(f''(a) + f''(a+\Delta x) \right) \Delta x^3 \\
&\quad + \frac{1}{2} \left(f(a+\Delta x) + f(a+2\Delta x) \right) \Delta x + \frac{1}{4} \left(f'(a+\Delta x) - f'(a+2\Delta x) \right) \Delta x^2 \\
&\quad + \frac{1}{12} \left(f''(a+\Delta x) + f''(a+2\Delta x) \right) \Delta x^3 + ... \\
&\quad + \frac{1}{2} \left(f(a+(n-1)\Delta x) + f(a+n\Delta x) \right) \Delta x + \frac{1}{4} \left(f'(a+(n-1)\Delta x) - f'(a+n\Delta x) \right) \Delta x^2 \\
&\quad + \frac{1}{12} \left(f''(a+(n-1)\Delta x) + f''(a+n\Delta x) \right) \Delta x^3 + \mathcal{O}(\Delta x^4) \\
&= \frac{1}{2} \Delta x \left[f(a) + 2f(a+\Delta x) + 2f(a+2\Delta x) + ... + 2f(a+(n-1)\Delta x) + f(a+n\Delta x) \right] \\
&\quad + \frac{1}{4} \Delta x^2 \left[f'(a) - f'(a+\Delta x) + f'(a+\Delta x) - f'(a+2\Delta x) \right. \\
&\quad \left. + ... + f'(a+(n-1)\Delta x) - f'(a+n\Delta x) \right] \\
&\quad + \frac{1}{12} \Delta x^3 \left[f''(a) + 2f''(a+\Delta x) + 2f''(a+2\Delta x) + ... + 2f''(a+(n-1)\Delta x) \right. \\
&\quad \left. + f''(a+n\Delta x) \right] + \mathcal{O}(\Delta x^4) \\
&= \frac{1}{2} \Delta x \left[f(a) + f(b) \right] + \Delta x \sum_{k=1}^{n-1} f(a+k\Delta x) + \frac{1}{4} \Delta x^2 \left[f'(a) - f'(b) \right] \\
&\quad + \frac{1}{12} \Delta x^3 \left[f''(a) + f''(b) \right] + \frac{1}{6} \Delta x^3 \sum_{k=1}^{n-1} f''(a+k\Delta x) + \mathcal{O}(\Delta x^4)
\end{aligned} \tag{57}$$

The first two terms are the trapezoidal rule approximation to the integral. The rest of the terms, therefore, give the difference between the trapezoidal rule approximation and the true value of the integral. Notice the last two terms look very similar to the trapezoidal rule. In fact:

$$\begin{aligned}
\int_a^b f''(x)dx &= \frac{1}{2} \Delta x \left[f''(a) + f''(b) \right] + \Delta x \sum_{k=1}^{n-1} f''(a+k\Delta x) \\
f'(b) - f'(a) &= \frac{1}{2} \Delta x \left[f''(a) + f''(b) \right] + \Delta x \sum_{k=1}^{n-1} f''(a+k\Delta x) + \mathcal{O}(\Delta x^2)
\end{aligned} \tag{58}$$

Substituting this in the equation above

$$\begin{aligned}
\int_a^b f(x)dx &= \frac{1}{2}\Delta x\left[f(a)+f(b)\right] + \Delta x\sum_{k=1}^{n-1} f(a+k\Delta x) + \frac{1}{4}\Delta x^2\left[f'(a)-f'(b)\right] \\
&\quad + \frac{1}{6}\Delta x^2\left[\frac{1}{2}\Delta x\left[f''(a)+f''(b)\right] + \Delta x\sum_{k=1}^{n-1} f''(a+k\Delta x)\right] + \mathcal{O}(\Delta x^4) \\
&= \frac{1}{2}\Delta x\left[f(a)+f(b)\right] + \Delta x\sum_{k=1}^{n-1} f(a+k\Delta x) + \frac{1}{4}\Delta x^2\left[f'(a)-f'(b)\right] \\
&\quad + \frac{1}{6}\Delta x^2\left[f'(b)-f'(a)+\mathcal{O}(\Delta x^2)\right] + \mathcal{O}(\Delta x^4) \\
&= \frac{1}{2}\Delta x\left[f(a)+f(b)\right] + \Delta x\sum_{k=1}^{n-1} f(a+k\Delta x) + \Delta x^2\left[\frac{1}{4}f'(a)-\frac{1}{4}f'(b)+\frac{1}{6}f'(b)-\frac{1}{6}f'(a)\right] \\
&\quad + \mathcal{O}(\Delta x^4) \\
&= \frac{1}{2}\Delta x\left[f(a)+f(b)\right] + \Delta x\sum_{k=1}^{n-1} f(a+k\Delta x) + \frac{1}{12}\Delta x^2\left[f'(a)-f'(b)\right] + \mathcal{O}(\Delta x^4)
\end{aligned} \tag{59}$$

The first two terms are the trapezoidal rule approximation to the integral, and the rest of the terms give the difference between the trapezoidal rule approximation and the true value of the integral. Therefore, to leading order in Δx, the approximation error on the integral is given by

$$\epsilon = \frac{1}{12}\Delta x^2\left[f'(a)-f'(b)\right] \tag{60}$$

The trapezoidal rule is accurate up to and including terms proportional to Δx and the leading-order approximation error is of order Δx^2.

A similar treatment for Simpson's rule shows that the approximation error is given to leading order by

$$\epsilon = \frac{1}{90}\Delta x^4\left[f'''(a)-f'''(b)\right] \tag{61}$$

So Simpson's rule will produce a better approximation than the trapezoidal rule by two orders of magnitude, given the same Δx.

The error formulas above are only helpful if we know the function $f(x)$ and its derivatives. In some cases, we might want to integrate a function which is represented by a set of points (perhaps measurements of some quantity). So how would we calculate the error on our approximation of an integral if we don't know the functional form?

Assume we are evaluting an integral between $x = a$ and $x = b$ using the trapezoidal rule. First, we choose N_1 slices, so that $\Delta x_1 = (b-a)/N_1$. Let's call the resulting integral we get

I_1. As we saw above, the error on this approximation, or the difference between I_1 and the true value of the integral I, is proportional to Δx_1^2 at leading order:

$$\begin{aligned} \epsilon_1 &= \frac{1}{12}\left[f'(a) - f'(b)\right]\Delta x_1^2 \\ I - I_1 &= c\Delta x_1^2 \end{aligned} \tag{62}$$

where we've defined constant c as $(1/12)[f'(a) - f'(b)]$. (If we don't know the functional form of f or its derivatives, we won't know the exact value of c, but that's fine.)

Now we do the integral again with N_2 slices, where $N_2 = 2N_1$. The step size $\Delta x_2 = (b-a)/N_2 = (b-a)/(2N_1) = \Delta x_1/2$. Let's call the resulting integral we get in this case I_2. The error on this approximation, or the difference between I_2 and the true value of the integral I is

$$\begin{aligned} \epsilon_2 &= \frac{1}{12}\left[f'(a) - f'(b)\right]\Delta x_2^2 \\ I - I_2 &= c\Delta x_2^2 \end{aligned} \tag{63}$$

which depends on the same constant c as our first approximation.

Solving both equations for I (the true value) and setting equal, we get

$$\begin{aligned} I_1 + c\Delta x_1^2 &= I_2 + c\Delta x_2^2 \\ I_2 - I_1 &= c(\Delta x_1^2 + \Delta x_2^2) \\ I_2 - I_1 &= c(2\Delta x_2^2 + \Delta x_2^2) \\ I_2 - I_1 &= 3c\Delta x_2^2 \\ I_2 - I_1 &= 3\epsilon_2 \\ \epsilon_2 &= \frac{1}{3}(I_2 - I_1) \end{aligned} \tag{64}$$

In this case, you can find the error on your approximation by doing the approximation twice; once with N_1 slices and again with $2N_1$ slices. The difference between these two results determines the error in your second calculation.

A similar treatment for Simpson's rule yields

$$\epsilon_2 = \frac{1}{15}(I_2 - I_1) \tag{65}$$

Example: Approximation Error

2.5 Adaptive Integration

In the previous section, we found we could estimate the approximation error by calculating the integral twice, with a different number of slices. This sets us up to do an adaptive integration method, in which you repeatedly do the calculation, varying the parameters based

on the most recent result until you get a desired accuracy. Generally, if you are doing a numerical integral, you have some goal accuracy in mind, and the number of slices will be set based on that goal. We can formalize this procedure in such a way that saves computing time.

The generalization of Equation 64 for the trapezoidal rule is given by

$$\epsilon_i = \frac{1}{3}(I_i - I_{i-1}) \tag{66}$$

where ϵ_i is the approximation error for the trapezoidal rule with N_i slices and I_i is the integral approximation from the trapezoidal rule with N_i slices, where $N_i = 2N_{i-1}$ (so N_i is generally even) and $\Delta x_i = \frac{1}{2}\Delta x_{i-1}$.

$$\begin{aligned} I_i &= \frac{1}{2}\Delta x_i \left[f(a) + f(b)\right] + \Delta x_i \sum_{k=1}^{N_i-1} f(a + k\Delta x_i) \\ &= \Delta x_i \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{k=1}^{N_i-1} f(a + k\Delta x_i)\right) \\ &= \Delta x_i \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{k=2,4,6,\ldots}^{N_i-2} f(a + k\Delta x_i) + \sum_{k=1,3,5\ldots}^{N_i-1} f(a + k\Delta x_i)\right) \end{aligned} \tag{67}$$

The sum over even k:

$$\begin{aligned} \sum_{k=2,4,6,\ldots}^{N_i-2} f(a + k\Delta x_i) &= f(a + 2\Delta x_i) + f(a + 4\Delta x_i) + \ldots + f(a + (N_i - 2)\Delta x_i \\ &= f(a + 2(1)\Delta x_i) + f(a + 2(2)\Delta x_i) + \ldots + f(a + 2(\frac{N_i - 2}{2})\Delta x_i \\ &= f(a + 2(1)\Delta x_i) + f(a + 2(2)\Delta x_i) + \ldots + f(a + 2(\frac{N_i}{2} - 1)\Delta x_i \\ &= \sum_{k=1}^{\frac{N_i}{2}-1} f(a + 2k\Delta x_i) = \sum_{k=1}^{N_{i-1}-1} f(a + 2k\Delta x_i) \end{aligned} \tag{68}$$

Therefore, I_i can be written:

$$\begin{aligned} I_i &= \Delta x_i \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{k=2,4,6,\ldots}^{N_i-2} f(a + k\Delta x_i)\right) + \Delta x_i \sum_{k=1,3,5\ldots}^{N_i-1} f(a + k\Delta x_i) \\ &= \frac{1}{2}\left[\Delta x_{i-1}\left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{k=1}^{N_{i-1}-1} f(a + 2k\Delta x_i)\right)\right] + \Delta x_i \sum_{k=1,3,5\ldots}^{N_i-1} f(a + k\Delta x_i) \\ I_i &= \frac{1}{2}I_{i-1} + \Delta x_i \sum_{k=1,3,5\ldots}^{N_i-1} f(a + k\Delta x_i) \end{aligned} \tag{69}$$

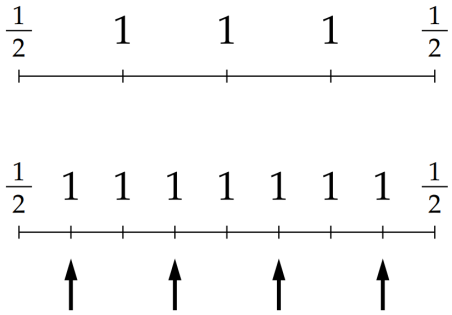


Figure 6:

The term in brackets in the middle line is just the trapezoidal rule approximation of the integral with N_{i-1} slices! What is the second term in I_i then? These are just the terms you add to the sum when you double the number of slices, see Figure 6.
This is handy - in calculating I_{i-1}, we have already performed half of the calculation for I_i. Therefore you add very little to the running time when you use this adaptive method, with the benefit of being able to easily choose the proper number of slices for the desired accuracy of the calculation.

The basic procedure:

1. Decide the target accuracy (typically to a certain number of decimal places, i.e. the error should be less than 0.001)
2. Choose and initial number of steps N_1 and calculate the first approximation to the integral (I_1) using the trapezoidal rule.
3. Double the number of steps and calculate the integral I_2 using Equation 69. Calculate the approximation error ϵ_2 using Equation 66.
4. If the error is less than the desired accuracy, stop. Otherwise, repeat starting from step 3 for I_3, ϵ_3, then I_4, ϵ_4, and so on until the error is less than the desired accuracy.

A similar procedure can be applied for Simpson's rule.

Example: Adaptive Integration

Section 2, Exercise 4

2.6 More Integration Methods

We're going to briefly touch on a few more integration methods. If you are going to be doing any heavy-duty numerical integration, you should learn more about these methods. But since our purpose for this class is only to introduce you to the concepts of numerical integration, we won't go into details.

2.6.1 Romberg Integration

Recall our expression for the true integral of a function $f(x)$ based on Taylor expansion:

$$\int_a^b f(x)dx = \frac{1}{2}\Delta x_i \left[f(a) + f(b)\right] + \Delta x_i \sum_{k=1}^{N_i-1} f(a + k\Delta x) + \frac{1}{12}\Delta x_i^2 \left[f'(a) - f'(b)\right] + \mathcal{O}(\Delta x_i^4)$$

$$I = I_i + \epsilon_i + \mathcal{O}(\Delta x_i^4) \tag{70}$$

where I is the true integral, I_i is the trapezoidal rule approximation with N_i steps, and ϵ_i is the leading-order approximation error, proportional to Δx_i^2. We found that we could write ϵ_i by doing the trapezoidal rule approximation twice, doubling the number of slices:

$$\epsilon_i = \frac{1}{3}(I_i - I_{i-1}) \tag{71}$$

Plugging that in, we have

$$I = I_i + \frac{1}{3}(I_i - I_{i-1}) + \mathcal{O}(\Delta x_i^4) \tag{72}$$

I_i has an approximation error proportional to Δx_i^2. But what if we use $I_i + (1/3)(I_i - I_{i-1})$ as the approximation of the integral? From the equation above, we can see that the approximation error is proportional to Δx_i^4. We have improved our error by two orders of magnitude (making it as accurate as Simpson's rule), but we only used the trapezoidal rule (to calculate I_i and I_{i-1})!

The process can be repeated. Let

$$R_{i,1} = I_i \tag{73}$$

$$R_{i,2} = I_i + \frac{1}{3}(I_i - I_{i-1}) = R_{i,1} + \frac{1}{3}(R_{i,1} - R_{i-1,1}) \tag{74}$$

$$\tag{75}$$

Using these definitions in the equation above, we have

$$I = R_{i,2} + \mathcal{O}(\Delta x_i^4)$$
$$I = R_{i,2} + c\Delta x_i^4 + \mathcal{O}(\Delta x_i^6) \tag{76}$$

where c is a constant that we have used to explicitly write the fourth-order terms.

We can write the same expression for $i-1$, and then make use of the fact that $\Delta x_i = \frac{1}{2}\Delta x_{i-1}$:

$$I = R_{i-1,2} + c\Delta x_{i-1}^4 + \mathcal{O}(\Delta x_{i-1}^6)$$
$$I = R_{i-1,2} + c(2\Delta x_i)^4 + \mathcal{O}((2\Delta x_i)^6)$$
$$I = R_{i-1,2} + 16c\Delta x_i^4 + \mathcal{O}(\Delta x_i^6) \tag{77}$$

(Since Δx_i and Δx_{i-1} are proportional to each other, terms of order Δx_{i-1}^6 are also of order Δx_i^6, i.e. we can ignore constant multiples in the "of order" function.)

Equating the two expressions for I and solving for the term proportional to Δx_i^4:

$$\begin{aligned} R_{i,2} + c\Delta x_i^4 + \mathcal{O}(\Delta x_i^6) &= R_{i-1,2} + 16c\Delta x_i^4 + \mathcal{O}(\Delta x_i^6) \\ 15c\Delta x_i^4 &= R_{i,2} - R_{i-1,2} + \mathcal{O}(\Delta x_i^6) \\ c\Delta x_i^4 &= \frac{1}{15}\left(R_{i,2} - R_{i-1,2}\right) + \mathcal{O}(\Delta x_i^6) \end{aligned} \tag{78}$$

Substituting this back into the expression for the true integral I:

$$\begin{aligned} I &= R_{i,2} + c\Delta x_i^4 + \mathcal{O}(\Delta x_i^6) \\ I &= R_{i,2} + \frac{1}{15}\left(R_{i,2} - R_{i-1,2}\right) + \mathcal{O}(\Delta x_i^6) \end{aligned} \tag{79}$$

Now we have an expression ($R_{i,2} + \frac{1}{15}\left(R_{i,2} - R_{i-1,2}\right)$) that approximates the integral with a sixth-order error, but again we have only used the trapezoidal rule!

This process can be repeated to achieve the desired accuracy. This is called Romberg integration.

2.6.2 Newton-Cotes

The trapezoidal rule works by approximating a function with a line between points. Simpson's rule works by approximating a function with a quadratic function. Both the trapezoidal rule (Equation 14) and Simpson's rule (Equation 42) have the same format:

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i), \tag{80}$$

a sum over the function values at equally spaced integrals with weighting factors. The weighting factors for the trapezoidal rule are $w_i = \frac{1}{2}$ for the first and last point and $w_i = 1$ for other points. For Simpson's rule, the weighting factors are $w_i = \frac{1}{3}$ for the first and last point, and $w_i = \frac{2}{3}, \frac{4}{3}$ for other points. We can extend the same logic to higher-orders, using a cubic ($\mathcal{O}(x^3)$) function, a quartic ($\mathcal{O}(x^4)$) function, and so on. Higher-order integration rules of this kind are called Newton-Cotes formulas. Just as the trapezoidal rule is exact for a linear function, and Simpson's rule is exact for a quadratic function, the kth Newton-Cotes rule is exact for a degree-k polynomial. With 2 equally-spaced sample points we can get an exact integral for a degree-1 polynomial (line); with 3 equally-spaced sample points we can get an exact integral for a degree-2 polynomial. Therefore, with N equally-spaced sample points, we could get an exact integral for an $(N-1)$-degree poynomial.

2.6.3 Gaussian quadrature

We can take this one step further if we allow the spacing between the intervals to vary, i.e. the position of the sample points in Equation 80 are not equally-spaced. Being able to vary the position of the N sample points gives us another N degrees of freedom, for a total of $2N$ (N from the weights and N from the sample point positions). Therefore it's

possible to create an integration rule that is exact for a $2N-1$ degree polynomial if all the degrees of freedom are chosen optimally. The method of Gaussian quadrature does exactly this.

The sum for Gaussian quadrature is just as easy to write (and easy to code) as that for the trapezoidal rule and Simpson's rule; the complexity comes in choosing the optimal weights and optimal position of the sample points. The optimal weights and sample points are not trivial to compute, and we're not going to go through the details here. Typically, the calculations are done for a standard range ($-1 < x < 1$), and then can easily be mapped to any other range. Lookup tables exist, and libraries have been written, so implementing this method is typically just as easy as implementing the trapezoidal rule or Simpson's rule.

2.7 Which method to use?

- Trapezoidal rule: Good choice for functions that aren't smooth, have singularities, are noisy, etc. Good if you are integrating measured data, where data points are equally spaced. Using adaptive integration can get you to a desired accuracy, though it may not be as quick as other methods.
- Simpson's rule: Similar benefits as trapezoidal rule, but more accurate with same number of points; doesn't work as well for non-smooth functions.
- Romberg integration: Gives the best accuracy for equally-spaced points, and allows you to set a desired accuracy. Doesn't work well for non-smooth functions.
- Gaussian quadrature: Highest accuracy, but you have to be able to use unequally spaced points. Also doesn't work well for non-smooth functions.

3 Rounding Error

We have discussed approximation error for numerical integrals, but there is another kind of error we have to contend with for all numerical calculations, related to the fact that computers cannot store numbers to an infinite number of decimal places.

The largest value for a floating-point number is around 10^{308} (the precise value is $2^{1024} = 1.79769 \times 10^{308}$). The corresponding limit for large negative numbers is around -10^{308}. If the value of a variable exceeds the limit, the variable would be printed as "inf" in Python, and any calculations with this variable will likely be incorrect. Similarly, there is a limit for the smallest value for a floating-point number, of around 10^{-308} (precisely, $2^{-1022} = 2.22507 \times 10^{-308}$). Any smaller than this, and the variable will just be set to zero, which could cause problems (if you are dividing by the variable for example).

For integers, Python can actually represent them to arbitrary precision; Python will store all digits of an integer. The problem here is that if you try to do a calculation with a very large integer, it can take a long time, so beware of using large integers.

Floating-point numbers are represented to 16 digits, which leads to rounding error. For example, the true value of π has an infinite number of digits, 3.14159265358979323846..., while the representation in Python is $\pi = 3.141592653589793$, leading to a round error of 0.00000000000000023846.... Any number whose true value has more than 16 digits will be rounded off.

In addition, the results of any arithmetic done with floating-point numbers is only guaranteed to 16 digits. So for example, if you add two numbers with 2 digits ($1.1 + 2.2 = 3.3$), the computer might give 3.299999999999999 instead of 3.3. This is why you should never test the equality of two floating-point numbers in a program, i.e. "if x==3.3". Because even if x is 3.3, it might be represented as 3.299999999999999, giving you an unexpected result from your if statement, a false when you expect a true.

Let's assume the rounding error is a uniformly distributed random number with standard deviation $\sigma = Cx$. In general if floating-point variable x is accurate to 16 digits, then the rounding error will have a typical size of $\sigma \sim x/10^{16}$, so that $C \sim 10^{-16}$. This seems small, but consider what happens when we perform mathematical with numbers that each have their own rounding error.

If we add or subtract two numbers x_1 and x_2 that have standard deviations σ_1 and σ_2, the variance of the sumσ^2 is equal to the sum of the individual variances:

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 \tag{81}$$

Using $\sigma_i = Cx_i$, we have:

$$\sigma = \sqrt{(Cx_1)^2 + (Cx_2)^2} = C\sqrt{x_1^2 + x_2^2} \tag{82}$$

If we are adding N numbers, then the variance on the final result is

$$\sigma^2 = \sum_{i=1}^{N} \sigma_i^2 = \sum_{i=1}^{N} (Cx_i^2) = C^2 \sum_{i=1}^{N} x_i^2 \tag{83}$$

If we calculate the mean squared value of x:

$$\overline{x^2} = \frac{1}{N} \sum_{i=1}^{N} x_i^2 \tag{84}$$

we can write the variance as

$$\sigma^2 = C^2 N \overline{x^2} \tag{85}$$

and the standard deviation as

$$\sigma = C\sqrt{N}\sqrt{\overline{x^2}} \tag{86}$$

The standard deviation in the sum increases as N increases; the more numbers we add together, the larger the error is on the result.

We can also think about this in terms of the fractional error, the total error divided by the sum:

$$\frac{\sigma}{\sum x_i} = \frac{C\sqrt{N}\sqrt{\overline{x^2}}}{N\overline{x}} = \frac{C}{\sqrt{N}}\frac{\sqrt{\overline{x^2}}}{\overline{x}} \tag{87}$$

where we wrote the mean value of x as $\overline{x} = (1/N)\sum x_i$. Therefore, the fractional error on the sum decreases as we add more terms.

One potential problem arises if the numbers you are adding vary a lot in magnitude, i.e. adding 0.001 to 1034382983342000.; the small numbers get lost in adding. A more substantial problem arises when you are subtracting very similar numbers. Take $x = 100000000000000$ and $y = 100000000000001.249476$. The second number gets rounded down to $y = 100000000000001.2$, so that the calculated difference is $y-x = 1.2$, rather than the true answer of $y-x = 1.249476$. The fractional error in this number is $(1.249476-1.2)/1.249476 = 0.049476/1.249476 \approx 4\%$, which is substantial.In calculations that involve the subtraction of nearly equal numbers, the rounding error can be significant.

Example: Subtraction

We went into some detail about the approximation error for numerical integration; how does the approximation error typically compare to the rounding error?

Consider the trapezoidal rule, which involves a sum over N terms. As shown above, the rounding error on the sum will be proportional to $C\sqrt{N}$, where $C \sim 10^{-16}$. However, the sum is multiplied by the factor Δx, which goes as $1/N$. (As N increases, Δx decreases.) Therefore, the cumulative error on the result of the calculation goes as $C/\sqrt{N}$. The error on the final operation in the calculation (adding the last term to the sum) is C times the final result. Since $C/\sqrt{N} < C$, it is safe to assume that the the final error is no greater than the error incurred by the final operation, i.e. we can neglect the contribution of the repeated additions and say that the rounding error is $\sim CI$, where I is the result for the integral.

The approximation error is proportional to Δx^2, and therefore decreases as we increase N. However, once the approximation error is as small as the rounding error, there's no point in

further increases in N. This happens when

$$\begin{aligned}
\frac{1}{12}\Delta x^2\left[f'(a)-f'(b)\right] &\approx CI \\
\Delta x &\approx \sqrt{\frac{12CI}{\left[f'(a)-f'(b)\right]}} \\
\frac{b-a}{N} &\approx \sqrt{\frac{12CI}{\left[f'(a)-f'(b)\right]}} \\
N &\approx (b-a)\sqrt{\frac{\left[f'(a)-f'(b)\right]}{12CI}}
\end{aligned} \tag{88}$$

To get an order of magnitude estimate, we assume every factor in this equation is ~ 1 except for C, then $N \sim C^{-1/2} = (10^{-16})^{-1/2} = 10^8$ is the number of slices at which the rounding error overtakes the approximation error as the main source of error. Since it would be fairly unusual to use 10^8 slices for the trapezoidal rule (because you could get a better approximation with fewer slices using another method), the rounding error can safely be neglected when using the trapezoidal rule.

What about Simpson's rule? Repeating the calculation,

$$\begin{aligned}
\frac{1}{90}\Delta x^4\left[f'''(a)-f'''(b)\right] &\approx CI \\
\Delta x &\approx \left(\frac{90CI}{\left[f'''(a)-f'''(b)\right]}\right)^{-1/4} \\
\frac{b-a}{N} &\approx \left(\frac{90CI}{\left[f'''(a)-f'''(b)\right]}\right)^{-1/4} \\
N &\approx (b-a)\left(\frac{\left[f'''(a)-f'''(b)\right]}{90CI}\right)^{-1/4}
\end{aligned} \tag{89}$$

So we get an order of magnitude estimate of $N \sim C^{-1/4} = (10^{-16})^{-1/4} = 10^4$ as the number of slices at which the rounding error overtakes the approximation error as the main source of error. Using 10,000 slices would not be so unusual, so this brings up an important point. When using Simpson's rule, there is no point in using more than a few thousand slices because the calculation will reach the limits of machine precision.

4 Derivatives

Derivatives can be defined in terms of the "forward difference":

$$\frac{df}{dx} = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} \tag{90}$$

or the "backward difference":

$$\frac{df}{dx} = \lim_{h\to 0}\frac{f(x)-f(x-h)}{h} \tag{91}$$

They typically give the same answer; the standard definition typically uses the forward difference. There are special cases where one is preferred over the other, especially when there is a discontinuity in the derivative of the the function at x or if you want the derivative of a bounded function on the boundary.

Unlike numerical integrals, the rounding error is more significant with numerical derivatives. The value of $f(x)$ can be calculated to an accuracy of $Cf(x)$, where $C = 10^{-16}$. The value of $f(x+h)$ can be calculated to an accuracy of $Cf(x+h)$. Therefore the maximum error on the difference $f(x+h) - f(x)$ is $\sim 2Cf(x)$ (note that $f(x+h) \approx f(x)$ since h is typically small). Therefore the rounding error on our calculation of $f'(x)$ using the forward difference rule is $\sim 2Cf(x)/h$.

To find the approximation error, let's consider the Tayor expansion of $f(x+h)$ about $x+h = x$:

$$\begin{aligned} f(x+h) &= \sum_{i=0}^{\infty} \frac{f^{(i)}(x)}{i!}(x+h-x)^i \\ f(x+h) &= \sum_{i=0}^{\infty} \frac{f^{(i)}(x)}{i!}h^i \\ f(x+h) &= f(x) + hf'(x) + \frac{1}{2}h^2 f''(x) + ... \end{aligned} \tag{92}$$

Rearranging this expression, we have

$$\begin{aligned} f'(x) &= \frac{1}{h}\left(f(x+h) - f(x) - \frac{1}{2}h^2 f''(x) - ... \right) \\ f'(x) &= \frac{f(x+h) - f(x)}{h} - \frac{1}{2}hf''(x) - ... \end{aligned} \tag{93}$$

The above expression gives us the true value of the derivative. When we use the forward difference formula to approximate the integral (Equation 90), we neglect the term proportion to the second derivative, plus all higher-order terms. Therefore, to leading order, the approximation error is given by $\sim \frac{1}{2}hf''(x)$.

The total error is therefore given by

$$\epsilon = \frac{2Cf(x)}{h} + \sim \frac{1}{2}hf''(x) \tag{94}$$

We want to find a value of h that minimizes the error, so we take the derivative of the

expression for ϵ with respect to h, set it to zero, and solve for h:

$$\begin{aligned}
\epsilon' &= 0 \\
-\frac{2Cf(x)}{h^2} + \frac{1}{2}f''(x) &= 0 \\
\frac{1}{2}f''(x) &= \frac{2Cf(x)}{h^2} \\
h^2 &= \frac{4Cf(x)}{f''(x)} \\
h &= \sqrt{\frac{4Cf(x)}{f''(x)}}
\end{aligned} \tag{95}$$

To get an order of magnitude estimate, we assume every factor in this equation is ~ 1 except for C, then $h \sim \sqrt{C} = \sqrt{10^{-16}} = 10^{-8}$. At this value of h, the error is $\epsilon \sim (10^{-16})/(10^{-8}) + 10^{-8} \sim 10^{-8}$. So with calculating derivative with the forward or backwared difference, we can't get estimates that are accurate to the machine precision. We can get accurate estimates to 8 digits.

Section 2, Exercise 5

Another way to calculate the derivative is to use the central difference:

$$\frac{df}{dx} \approx \frac{f(x+h/2) - f(x-h/2)}{h} \tag{96}$$

We can find the approximation using Taylor expansions as before:

$$\begin{aligned}
f(x+h/2) &= f(x) + \frac{h}{2}f'(x) + \frac{1}{2}\left(\frac{h}{2}\right)^2 f''(x) + \frac{1}{6}\left(\frac{h}{2}\right)^3 f'''(x) + ... \\
&= f(x) + \frac{1}{2}hf'(x) + \frac{1}{8}h^2 f''(x) + \frac{1}{48}h^3 f'''(x) + ...
\end{aligned} \tag{97}$$

$$\begin{aligned}
f(x-h/2) &= f(x) + \left(-\frac{h}{2}\right)f'(x) + \frac{1}{2}\left(-\frac{h}{2}\right)^2 f''(x) + \frac{1}{6}\left(-\frac{h}{2}\right)^3 f'''(x) + ... \\
&= f(x) - \frac{1}{2}hf'(x) + \frac{1}{8}h^2 f''(x) - \frac{1}{48}h^3 f'''(x) + ...
\end{aligned} \tag{98}$$

Subtracting the second equation from the first, we have

$$\begin{aligned}
f(x+h/2) - f(x-h/2) &= \left[f(x) + \frac{1}{2}hf'(x) + \frac{1}{8}h^2 f''(x) + \frac{1}{48}h^3 f'''(x) + ...\right] \\
&\quad - \left[f(x) - \frac{1}{2}hf'(x) + \frac{1}{8}h^2 f''(x) - \frac{1}{48}h^3 f'''(x) + ...\right] \\
f(x+h/2) - f(x-h/2) &= hf'(x) + \frac{1}{24}h^3 f'''(x) + ... \\
f'(x) &= \frac{f(x+h/2) - f(x-h/2)}{h} - \frac{1}{24}h^2 f'''(x) + ...
\end{aligned} \tag{99}$$

This indicates that the leading order approximation error for the central difference is $(1/24)h^2 f'''(x)$, which is an order higher in h than what we got with the forward or backward difference. The total error including rounding error is

$$\epsilon = \frac{2Cf(x)}{h} + \frac{1}{24}h^2 f'''(x) \tag{100}$$

Again, we take the derivative of ϵ, set it to zero, and solve for h to find the value of h that minimizes the error:

$$\begin{aligned} -\frac{2Cf(x)}{h^2} + \frac{1}{12}hf'''(x) &= 0 \\ \frac{1}{12}hf'''(x) &= \frac{2Cf(x)}{h^2} \\ h^3 &= \frac{24Cf(x)}{f'''(x)} \\ h &= \left(\frac{24Cf(x)}{f'''(x)}\right)^{1/3} \end{aligned} \tag{101}$$

An order of magnitude estimate gives us $h \sim C^{1/3} = (10^{-16})^{1/3} \sim 10^{-5}$. Plugging this value in to solve for ϵ, we get

$$\begin{aligned} \epsilon &= 2Cf(x)\left(\frac{f'''(x)}{24Cf(x)}\right)^{1/3} + \frac{1}{24}\left(\frac{24Cf(x)}{f'''(x)}\right)^{2/3} f'''(x) \\ &\sim C^{2/3} \sim 10^{-10} \end{aligned} \tag{102}$$

So the error using the central difference is about two orders of magnitude smaller than that we get from the forward/backward difference, but we achieve this by using a larger h (10^{-5} vs 10^{-8}).

Example: Derivatives

Essentially what we do when we use the forward or backward difference to approximate the derivative is to find the slope of the line between the two points $f(x)$ and $f(x+h)$. What if we used a quadratic approximation instead of a linear one? In this case you would need three points, $f(x-h/2), f(x), f(x+h/2)$ to find the quadratic function. Once you solve for the quadratic function that passes through the three points, you can approximate the derivative of the function f with the quadratic function derivative evaluated at x. Without going through the details here, it turns out this gives you the central difference approximation for the derivative. If necessary, you can make higher order approximations, using a cubic or quartic function for example, which will be more accurate.

In some cases, we might need a second or higher-order derivative, or a partial derivative, which can also be done numerically. We'll discuss these as we come across them in later problems.