Notes for Section 4.1

In Section 3.7, we learned how the antiderivative of the velocity can be used to find the position of a moving object. I suggest you start reading Section 4.1 with the Distance Problem on p. 206. Example 4 considers the challenge of finding the displacement, that is the distance covered between some point in time a and some other point in time b, of a moving object, in this case, a car, based on knowing its velocity at certain times. Suppose the velocity function $v(t)$ is not known, but we know the velocity at certain points of time

$$a = t_0 < t_1 < t_2 < \cdots < t_n = b.$$

Note that the counting starts at 0, which is so that we have n time intervals. Also note that the book uses $f(t)$ for the velocity, not $v(t)$. I will stick with $v(t)$. The example walks you through how you can estimate the distance traveled between t_i and t_{i+1} by multiplying the velocity at t_i by the amount of time that elapses, namely $t_{i+1} - t_i$. The idea is that we assume that the velocity does not change much over a short enough time interval and use $v(t_i)$ as an estimate for the velocity over the entire time interval $[t_i, t_{i+1}]$. This results in $v(t_i)(t_{i+1} - t_i)$ as our estimate for the distance covered during the time interval $[t_i, t_{i+1}]$. When we add these estimates, we get an estimate for the distance covered between $t_0 = a$ and $t_n = b$. As the example points out, we could just as well use $v(t_{i+1})$, the velocity at the end of the time interval $[t_i, t_{i+1}]$, to estimate how quickly the car was moving between t_i and t_{i+1}. That would give us $v(t_{i+1})(t_{i+1} - t_i)$ as our estimate for the distance covered during the time interval $[t_i, t_{i+1}]$. Adding these over all of the time intervals also gives us an estimate for the distance covered between t_0 and t_n. The two estimates are usually different. Which one is better depends on how the car actually moves.

In this example, the times at which v is measured are evenly spaced, that is $t_{i+1} - t_i$ is always the same for each i. But this is not necessary. The same ideas work even if the time intervals have different widths. It should make intuitive sense that using shorter time intervals gives a more accurate estimate because the velocity of the car is likely to vary less over a short time interval than over a longer time interval. This is similar to the observation we made about measuring speed (or velocity) in the fall: the shorter the time interval of the measurement, the better the estimate. Ultimately, if we want to make the estimate perfect, that is no longer an estimate but an accurate calculation of the distance covered, we need to let the widths of the time intervals approach 0, while we look at what number the estimated distances are getting closer and closer to. That is the distance covered is the limit of the sum

$$v(t_0)(t_1 - t_0) + v(t_1)(t_2 - t_1) + \cdots + v(t_{n-1})(t_n - t_{n-1}) = \sum_{i=0}^{n-1} v(t_i)(t_{i+1} - t_i)$$

or of the sum

$$v(t_1)(t_1 - t_0) + v(t_2)(t_2 - t_1) + \cdots + v(t_n)(t_n - t_{n-1}) = \sum_{i=1}^{n} v(t_i)(t_i - t_{i-1})$$

as the widths of the time intervals all approach 0. A good way to say that the widths of the time intervals all approach 0 is to say $\max_{0 \le i \le n-1}(t_{i+1} - t_i) \to 0$, that is even the longest one among them should be getting closer and closer to 0. So the distance covered is

$$d = \lim_{\max(t_{i+1} - t_i) \to 0} \sum_{i=0}^{n-1} v(t_i)(t_{i+1} - t_i)$$

or

$$d = \lim_{\max(t_{i+1} - t_i) \to 0} \sum_{i=1}^{n} v(t_i)(t_i - t_{i-1}).$$

That these two limits have the same value is certainly not obvious at this point, but it is not hard to show for a function f that is sufficiently well-behaved. Right now, let's not get bogged down on that detail. Let's just accept that it is so, and work on developing our conceptual understanding of what such a limit of a sum actually means.

These formulas become simpler if we assume that the time points are evenly spaced. Then

$$t_1 - t_0 = t_2 - t_1 = \cdots = t_n - t_{n-1} = \frac{b-a}{n}$$

Then we can write

$$d = \lim_{\Delta t \to 0} \sum_{i=0}^{n-1} v(t_i)\Delta t$$

or

$$d = \lim_{\Delta t \to 0} \sum_{i=1}^{n} v(t_i)\Delta t.$$

where $\Delta t = \frac{b-a}{n}$. Since Δt gets small when n gets large, another way to write this is

$$d = \lim_{n \to \infty} \sum_{i=0}^{n-1} v(t_i)\Delta t$$

or

$$d = \lim_{n \to \infty} \sum_{i=1}^{n} v(t_i)\Delta t.$$

Once you have digested these ideas for estimating displacement from velocity, go back to the beginning of the section and read about how similar ideas can be used to estimate the area under the graph of a function f. Actually, they are not just similar ideas. They are the exact same ideas. Estimating the displacement of a moving object between two points a and b in time based on the velocity $v(t)$ is the same problem as finding the area under the graph of $v(t)$ for $t \in [a, b]$. The book addresses the general problem of finding the area under the graph of a function $f(x)$ for $x \in [a, b]$. Example 1 shows you how you can estimate the area under the graph of $f(x) = x^2$ between $x = 0$ and $x = 1$ by dividing $[0, 1]$ into a few subintervals and using rectangles over these subintervals. It does this two ways: once evaluating f at the right endpoints of the subintervals, and once using the left endpoints of the subintervals. This is exactly how it was done in Example 4 too. Let's give these two approaches names, so we can refer to them more easily. The sum

$$L_n = f(x_0)(x_1 - x_0) + f(x_1)(x_2 - x_1) + \cdots + f(x_{n-1})(x_n - x_{n-1}) = \sum_{i=0}^{n-1} f(x_i)(x_{i+1} - x_i)$$

uses the values of f at the left endpoints of the subintervals and is called the *left-hand sum*, while the sum

$$R_n = f(x_1)(x_1 - x_0) + f(x_2)(x_2 - x_1) + \cdots + f(x_n)(x_n - x_{n-1}) = \sum_{i=1}^{n} f(x_i)(x_i - x_{i-1})$$

uses the values of f at the right endpoints of the subintervals and is called the *right-hand sum*. It is clear from the graphs that in the case of $f(x) = x^2$ between $x = 0$ and $x = 1$, the left-hand sum always underestimates the area, while the right-hand sum always overestimates it. This is because f is an increasing function on the interval $[0, 1]$.

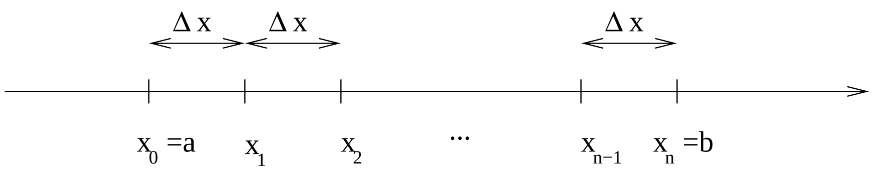
Of course, we could have used the midpoints of the subintervals as well. In that case, the sum would be

$$\begin{aligned} M_n &= f\left(\frac{x_0+x_1}{2}\right)(x_1-x_0) + f\left(\frac{x_1+x_2}{2}\right)(x_2-x_1) + \cdots + f\left(\frac{x_{n-1}+x_n}{2}\right)(x_n-x_{n-1}) \\ &= \sum_{i=1}^{n} f\left(\frac{x_{i-1}+x_i}{2}\right)(x_i-x_{i-1}) \end{aligned}$$

and would be called the *midpoint sum* or *midpoint rule*. This will appear in Example 3 and again in Section 4.2.

Example 2 illustrates how you can calculate the right-hand sum R_n for the area under the same function $f(x) = x^2$ using n rectangles over n evenly spaced subintervals in general. In this simple case, it is actually possible to turn the sum into a closed form, using a well-known formula for the sum of the first n perfect squares. The example then shows you how you can calculate the limit of R_n as $n \to \infty$. This will give you the exact area under the graph of $f(x)$ between $x = 0$ and $x = 1$. If you feel up to the challenge, you could now try to do the analogous calculations with the left-hand sum L_n. Do you get the same result?

In fact, we can do the same to find the area under the graph of a general function $f(x)$ over the interval $[a, b]$. If we divide this interval into n evenly spaced subintervals



then each of these will have width $\Delta x = \frac{b-a}{n}$. We know $x_0 = a$ and $x_n = b$. It is easy enough to find x_i in general:

$$x_i = a + i\Delta x = a + i\,\frac{b-a}{n}.$$

So we can write the left-hand sum as

$$L_n = f(a)\Delta x + f(a+\Delta x)\Delta x + \cdots + f(a+(n-1)\Delta x)\Delta x = \sum_{i=0}^{n-1} f(a+i\Delta x)\Delta x,$$

and the right-hand sum as

$$R_n = f(a+\Delta x)\Delta x + f(a+2\Delta x)\Delta x + \cdots + f(\underbrace{a+n\Delta x}_{b})\Delta x = \sum_{i=1}^{n} f(a+i\Delta x)\Delta x.$$

For the midpoint sum, you would use

$$\begin{aligned} M_n &= f\left(a+\frac{1}{2}\,\Delta x\right)\Delta x + f\left(a+\frac{3}{2}\,\Delta x\right)\Delta x + \cdots + f\left(a+\frac{2n-1}{2}\,\Delta x\right)\Delta x \\ &= \sum_{i=0}^{n-1} f\left(a+\left(i+\frac{1}{2}\right)\Delta x\right)\Delta x. \end{aligned}$$

In fact, there is nothing magical about the left endpoint or the right endpoint or the midpoint of a subinterval. We could use any point x_i^* in $[x_{i-1}, x_i]$ for estimating the height of the function f over the subinterval. The $x_1^*, \ldots, x_n^*$ are called *sample points*. This would give us the following sum using n evenly spaced subintervals of width $\Delta x = \frac{b-a}{n}$:

$$f(x_1^*)\Delta x + f(x_1^*)\Delta x + \cdots + f(x_{n-1}^*)\Delta x = \sum_{i=1}^{n} f(x_i^*)\Delta x.$$

Or if the subintervals are of different widths:

$$f(x_1^*)(x_1 - x_0) + f(x_1^*)(x_2 - x_1) + \cdots + f(x_{n-1}^*)(x_n - x_{n-1}) = \sum_{i=1}^{n} f(x_i^*)(x_i - x_{i-1})$$

If your head is spinning with all these indices and stars, staring Figure 13 in the book may help you sort out what is what.

While we all have an intuitive understanding of what the area under the graph of $f(x)$ between $x = a$ and $x = b$ means, if we want to have a precise definition of what it means to measure that area, we need exactly the ideas we have talked about. In fact, the limit of one of these sums as the widths of the subintervals approach 0 is exactly the definition of the area under the graph of f. Of course, the limit may not exist if f is a sufficiently unruly function. Another problem we could run into is that the limits may be different for different choices of sample points. Then which of those limits should we accept as the area under f? This is why Definition 2 defines the area under the graph of f only for a continuous function f. It can be shown that a continuous function is sufficiently well-behaved that the limits of these sums do exist and they are the same for any choice of sample points. Of the possible sums, Definition 2 chooses the right-hand sum. In fact, we could use any sum if f is continuous. There are also plenty of discontinuous functions that are sufficiently well-behaved that the limits of these sums do exist and they are the same for any choice of sample points. But if f is continuous, then we know for sure that the limits of the sum exist and are all equal, whereas if f is not continuous, we just cannot be sure.

We have talked about choosing the sample points $x_1^*, \ldots, x_n^*$ so that they are at the left endpoints or right endpoints or at the midpoints of the subintervals. Another notable choice would be to choose them so that each $f(x_i^*)$ is the smallest value of f in the subinterval $[x_{i-1}, x_i]$. This would give us what is called the *lower sum*. One advantage of working with the lower sum is that we can be certain that it underestimates the area under f. Similarly, choosing the sample points so that each $f(x_i^*)$ is the largest value of f in the subinterval $[x_{i-1}, x_i]$ results in an *upper sum*, which always overestimates the area under f. In Example 1, the left-hand sum was also the lower sum and the right-hand sum was also the upper sum because the function f was increasing. In general, these are four different things.

Example 3 further illustrates how these sums can be used to estimate the area under the graph of a function.