Calculus for Machine Learning

Daniel O'Connor

Contents

Preface

Part 1

Calculus

CHAPTER 1

Rate of change

What one fool can do, so can another.

An "ancient Simian proverb" appearing in *Calculus Made Easy* and sometimes repeated by Richard Feynman

The name "calculus" doesn't tell you what the subject is about, so here it is: the main idea of calculus is *instantaneous rate of change*. Everyone who has read a speedometer understands this concept intuitively. If you are driving a car, you might see the speedometer steadily increase from 0 meters per second to 30 meters per second. Along the way, there was a single instant in time at which your speed was 20 meters per second.

Before writing down a precise definition of "instantaneous rate of change", let's first review the concept of *average* rate of change. Let $f : \mathbb{R} \to \mathbb{R}$ be a function. For example, f could be the function that takes as input the number t of seconds that have passed since you began driving, and returns as output the number $f(t)$ of meters that the car has traveled during this time. (This function f is like an *odometer* — it tells you the total distance that you have traveled so far.) The average velocity of the car during the time interval from a time t_0 to a later time t is

$$\text{average velocity} = \frac{\text{distance traveled}}{\text{time elapsed}} = \frac{f(t) - f(t_0)}{t - t_0}.$$

For example, suppose that at time $t_0 = 30$ the odometer reports that the car has traveled a total distance of 120 meters, so

that $f(30) = 120$, and suppose also that two seconds later the odometer reports a total distance traveled of 160 meters, so that $f(32) = 160$. Then the car has traveled $f(32) - f(30) = 160 - 120 = 40$ meters during the two second time interval from time $t_0 = 30$ to time $t = 32$. The average velocity of the car during this time interval is $(160 - 120)/(32 - 30) = 40/2 = 20$ meters per second.

Now imagine that the time interval from t_0 to t is very short, so that time t is only a split second later than time t_0. Then the average velocity during the time interval from t_0 to t is a good approximation to the instantaneous velocity of the car at time t_0. Even better approximations can be obtained by taking t to be closer and closer to t_0. In fact, we can approximate the instantaneous velocity *as closely as we like* by taking t to be sufficiently close to t_0. We express this fact concisely by saying that the instantaneous velocity at time t_0 is equal to the *limit* as t approaches t_0 of the average velocity $(f(t) - f(t_0))/(t - t_0)$. To save writing, the instantaneous velocity at time t_0 is denoted $f'(t_0)$. In summary:

$$f'(t_0) = \lim_{t \to t_0} \frac{f(t) - f(t_0)}{t - t_0}. \tag{1.1}$$

This idea of taking a *limit* is illustrated with a numerical example in table 1.

The number $f'(t_0)$ has been given an undescriptive and unnecessarily intimidating name: it is called the **derivative** of the function f at t_0. The function f', which takes a number t_0 as input and returns the number $f'(t_0)$ as output, is called the derivative of f. Another common notation for f' is $\frac{df}{dt}$. The notation $\frac{df}{dt}$ arguably has more mnemonic power, as it reminds us that if t is close to t_0 then

$$f'(t_0) \approx \frac{\Delta f}{\Delta t},$$

t	$f(t)$	$\frac{f(t)-f(t_0)}{t-t_0}$
2.1	4.41	4.1
2.01	4.0401	4.01
2.001	4.004001	4.001

Table 1. Here we illustrate the idea of taking a *limit* in equation (1.1). As t approaches t_0, the average rate of change approaches the instantaneous rate of change. In this example, $f(t) = t^2$ and $t_0 = 2$. As we try out values of t that get closer and closer to t_0, we observe that the average rate of change (shown in the rightmost column) appears to be getting closer and closer to 4. This suggests that $f'(2) = 4$.

where $\Delta t = t - t_0$ is a small change in value of the input to f and $\Delta f = f(t) - f(t_0)$ is the corresponding change in the value of the output.

In the above discussion, we used a moving car as an example, but f can be any function that takes a real number as input and returns a real number as output. A car's velocity is only one example of the concept of instantaneous rate of change of a quantity. And while we used the letter t as a name for the function input, we may of course use any letter we want, such as x. We would then write equation (1.1) as

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}. \tag{1.2}$$

Example 1.1. To make this discussion more concrete, let's compute the derivative of the function $f(x) = x^2$. We must evaluate the limit on the right in equation (1.2). Notice that if

$x \neq x_0$ then

$$\begin{aligned}\frac{f(x)-f(x_0)}{x-x_0} &= \frac{x^2-x_0^2}{x-x_0}\\ &= \frac{(x-x_0)(x+x_0)}{x-x_0}\\ &= x+x_0.\end{aligned}$$

Clearly, $x+x_0$ approaches $x_0+x_0=2x_0$ as x approaches x_0. This shows that

$$f'(x_0) = 2x_0. \tag{1.3}$$

For example, $f'(2) = 2\cdot 2 = 4$. This confirms the result that we observed in table 1.

One of the main goals of a calculus course is to derive a large number of rules like this for computing the derivatives of specific functions. We will derive more such rules in section 2.

EXERCISE 1.2. Derive formulas for the derivatives of the functions x, x^3, and $1/x$.

1.1. Geometric interpretation of the derivative

There is a nice geometric interpretation of the derivative that is illustrated in figure 1. If x is close to x_0, then $(x_0, f(x_0))$ and $(x, f(x))$ are nearby points on the graph of f. The line connecting these two points, shown in figure 1a, is called a "secant line". The slope of this line is given by the formula $\text{slope} = \frac{\text{rise}}{\text{run}}$, and as can be seen in figure 1a, the run is $x-x_0$ and the rise is $f(x)-f(x_0)$. Thus:

$$\text{slope of secant line} = \frac{\text{rise}}{\text{run}} = \frac{f(x)-f(x_0)}{x-x_0}.$$

As x approaches x_0, the point $(x, f(x))$ approaches the point $(x_0, f(x_0))$, and the slope of the secant line approaches the slope of the "tangent line" that is shown in figure 1b. So, the slope of the tangent line in figure 1b is

$$\text{slope of tangent line} = \lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}.$$

The quantity on the right is none other than the derivative $f'(x_0)$. We have found our geometric interpretation of the derivative:

The derivative is the slope of the tangent line.

1.2. The fundamental strategy of calculus

The definition of the derivative (equation (1.1)) tells us that the approximation

$$f'(x_0) \approx \frac{f(x) - f(x_0)}{x - x_0} \tag{1.4}$$

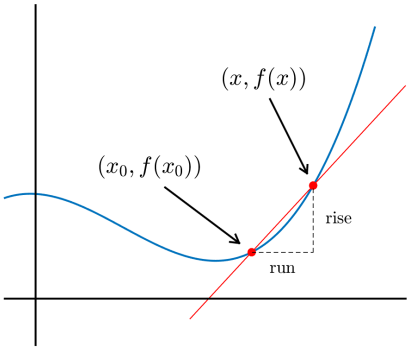
becomes more and more accurate as we select values of x that are closer and closer to x_0, and that any desired level of accuracy can be obtained by restricting x to be sufficiently close to x_0. Visually, we are approximating the slope of the tangent line by computing the slope of a secant line. Multiplying both sides of (1.4) by $x - x_0$, we obtain

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0). \tag{1.5}$$

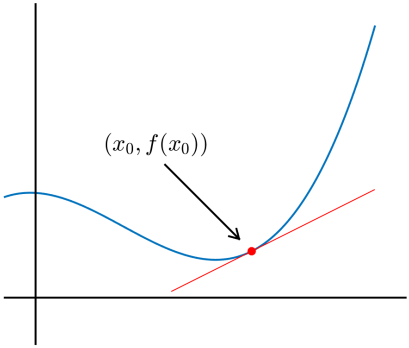
The approximation is good when x is close to x_0.

Equation (1.5) is called "Newton's approximation", and it is extremely useful for the following reason: Although f itself might be a complicated nonlinear function, f can be approximated accurately by the very simple linear function $L(x) = f(x_0) + f'(x_0)(x - x_0)$ which appears on the right in equation (1.5). The graph of L is a *straight line* that passes through the point $(x_0, f(x_0))$ and has slope $f'(x_0)$. In other words, the graph of L is the tangent line shown in figure 1b. The function L is called the linear approximation to f near x_0.

The *fundamental strategy of calculus* is to replace f (which is difficult to work with) with a linear approximation to f (which is easy to work with). When we do this, whatever calculations we want to perform are greatly simplified, and often the approximation is accurate enough that the result of the calculation is useful. This strategy is used again and again throughout calculus,



(a) The line connecting the points $(x_0, f(x_0))$ and $(x, f(x))$. is called a "secant line". We use the formula slope $= \frac{\text{rise}}{\text{run}}$ to compute the slope of the secant line. The run is $x - x_0$ and the rise is $f(x) - f(x_0)$, so the slope is $\frac{f(x)-f(x_0)}{x-x_0}$.



(b) As x approaches x_0, the point $(x, f(x))$ approaches the point $(x_0, f(x_0))$, and the slope of the secant line approaches the slope of the tangent line. Thus, the slope of the tangent line is $\lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}$.

Figure 1. The derivative is the slope of the tangent line.

and it makes calculus easy. It is the key to calculus. Newton's approximation (1.5) should be internalized so that you can use it effortlessly and reflexively.

Suppose that you are driving and at a particular moment the speedometer reads 20 meters per second, and you are asked to estimate how far the car travels during the next two seconds. Even if you don't know calculus, you will estimate $20 \times 2 = 40$ meters. You already use the approximation (1.5) even if you do not realize it. It is so intuitive that a child would give the same answer.

1.3. Warning: Sometimes the derivative does not exist

The geometric interpretation of the derivative helps us to understand visually something that can go wrong when computing $f'(x_0)$. Figure 2a shows the graph of the ramp function f defined by

$$f(x) = \begin{cases} x & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

This ramp function is also called the "ReLU" function (yet another undescriptive name), and it plays an important role in neural networks. For this example, let $x_0 = 0$ (and note that $f(x_0) = 0$). As shown in figure 2a, if $x > x_0$, then the slope of the secant line connecting $(x_0, f(x_0))$ and $(x, f(x))$ is $\frac{f(x)-f(x_0)}{x-x_0} = \frac{x-0}{x-0} = 1$. Thus, as x approaches x_0 *from the right*, the slope of the secant line approaches 1:

$$\lim_{x \searrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = 1.$$

The symbol $\searrow$ indicates that x approaches x_0 from the right (in other words, x *decreases* towards x_0).

However, we get a different result if x approaches x_0 *from the left*. As shown in figure 2b, if $x < x_0$, then the slope of the secant line connecting $(x_0, f(x_0))$ and $(x, f(x))$ is $\frac{f(x)-f(x_0)}{x-x_0} = \frac{0-0}{x-0} = 0$. So, as x approaches x_0 from the left, the slope of the secant line

approaches 0:

$$\lim_{x \nearrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = 0.$$

The fact that the slope of the secant line approaches different values depending on whether x approaches x_0 from the right or from the left means that in this example $\frac{f(x)-f(x_0)}{x-x_0}$ simply does not have a unique limit as x approaches x_0:

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \text{ does not exist.}$$

The function f *does not have* a derivative at 0.

Imagine that a car is driving along at 10 meters per second, and then at time $t = 30$ the car's velocity magically jumps to 20 meters per second (perhaps due to a glitch in the matrix). What is the car's instantaneous velocity at time $t = 30$? There is no correct answer. Both values 10 meters per second and 20 meters per second would be equally valid. The car simply does not have a velocity at that instant in time.

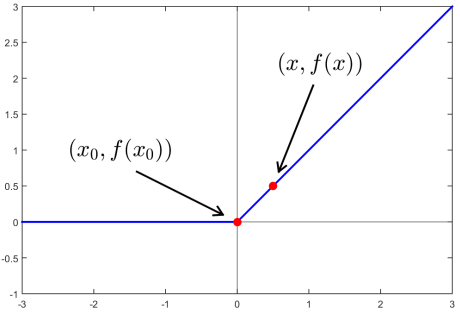
When we make the statement

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = L,$$

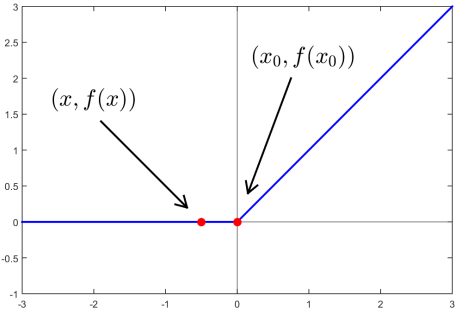
we insist that $\frac{f(x)-f(x_0)}{x-x_0}$ approaches the same limiting value L *no matter what path x follows* as x approaches x_0. If that is not the case, then the statement is not true, and $f'(x_0)$ simply does not exist.

Before we compute the derivative of a function, we should be careful to first check that the derivative exists. But don't worry, most functions we encounter have perfectly smooth graphs, with no sharp corners where the tangent line is not well defined.

Here is a bit more terminology. The process of computing the derivative of a function is called "differentiation". If f has a derivative at x_0, then f is said to be "differentiable" at x_0. We have seen in this section that the ramp function is not differentiable at 0. However, the ramp function is differentiable everywhere else.



(a) If $x > x_0$, then the slope of the secant line connecting $(x_0, f(x_0))$ and $(x, f(x))$ is 1. Thus, $\lim_{x \searrow x_0} \frac{f(x)-f(x_0)}{x-x_0} = 1$.



(b) However, if $x < x_0$, then the slope of the secant line is 0. Thus, $\lim_{x \nearrow x_0} \frac{f(x)-f(x_0)}{x-x_0} = 0 \neq 1$.

Figure 2. The ramp function does not have a derivative at $x_0 = 0$. The derivative is supposed to be the slope of the tangent line, but there is not a unique tangent line at this point.

CHAPTER 2

Formulas for the derivative

In this chapter we will discover some useful rules for computing derivatives. In the process we will see our first examples of the fundamental strategy of calculus in action.

2.1. Derivative of a constant function

Suppose that f is a constant function. In other words, there is some number c such that $f(x) = c$ for all possible values of x. Then, if $x \neq x_0$, we have

$$\frac{f(x) - f(x_0)}{x - x_0} = \frac{c - c}{x - x_0} = 0.$$

It follows that

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \to x_0} 0 = 0.$$

In words:

The derivative of a constant function is 0.

If a car's position is not changing, then its velocity is 0.

2.2. Derivative of a sum

Suppose that

$$f(x) = g(x) + h(x),$$

and both f and g are differentiable at x_0. If $x \neq x_0$ then

$$\begin{aligned}\frac{f(x)-f(x_0)}{x-x_0} &= \frac{g(x)+h(x)-(g(x_0)+h(x_0))}{x-x_0} \\ &= \underbrace{\frac{g(x)-g(x_0)}{x-x_0}}_{\text{approaches } g'(x_0)} + \underbrace{\frac{h(x)-h(x_0)}{x-x_0}}_{\text{approaches } h'(x_0)}.\end{aligned}$$

As x approaches x_0, the first term on the right approaches $g'(x_0)$ and the second term on the right approaches $h'(x_0)$. Thus,

$$f'(x_0) = g'(x_0) + h'(x_0).$$

In words:

The derivative of a sum is the sum of the derivatives.

EXAMPLE 2.1. If

$$f(x) = \underset{\uparrow\, g(x)}{1} + \underset{\uparrow\, h(x)}{x^2}$$

then

$$f'(x_0) = \underset{\uparrow\, g'(x_0)}{0} + \underset{\uparrow\, h'(x_0)}{2x_0} = 2x_0.$$

2.3. Derivative of $f(x) = cg(x)$

Suppose that there is a number c such that

$$f(x) = cg(x)$$

for all real numbers x, and that g is differentiable at x_0. If $x \neq x_0$ then

$$\begin{aligned}\frac{f(x)-f(x_0)}{x-x_0} &= \frac{cg(x)-cg(x_0)}{x-x_0} \\ &= c\underbrace{\left(\frac{g(x)-g(x_0)}{x-x_0}\right)}_{\text{approaches } g'(x_0)}.\end{aligned}$$

It follows that

$$f'(x_0) = cg'(x_0).$$

EXAMPLE 2.2. If

$$f(x) = \overset{c}{\underset{\downarrow}{5}}\, \underset{\uparrow}{x^2}\,, \qquad g(x)$$

then

$$f'(x_0) = \overset{c}{\underset{\downarrow}{5}} \cdot \underset{\uparrow}{(2x_0)} = 10x_0. \qquad g'(x_0)$$

2.4. The product rule

Suppose that

$$f(x) = g(x)h(x),$$

and that g and h are differentiable at x_0. If $x \neq x_0$ then

$$\frac{f(x) - f(x_0)}{x - x_0} = \frac{g(x)h(x) - g(x_0)h(x_0)}{x - x_0}. \tag{2.1}$$

We invoke the fundamental strategy of calculus to simplify the expression on the right. Using the approximations

$$\begin{aligned} g(x) &\approx g(x_0) + g'(x_0)(x - x_0) \\ \text{and} \quad h(x) &\approx h(x_0) + h'(x_0)(x - x_0), \end{aligned}$$

we obtain

$$\begin{aligned} g(x)h(x) &\approx \Big(g(x_0) + g'(x_0)(x - x_0)\Big)\Big(h(x_0) + h'(x_0)(x - x_0)\Big) \\ &= g(x_0)h(x_0) \\ &\quad + g'(x_0)h(x_0)(x - x_0) + g(x_0)h'(x_0)(x - x_0) \\ &\quad + g'(x_0)h'(x_0)(x - x_0)^2. \end{aligned}$$

It follows that

$$\frac{g(x)h(x) - g(x_0)h(x_0)}{x - x_0} \approx g'(x_0)h(x_0) + g(x_0)h'(x_0) + \underbrace{g'(x_0)h'(x_0)(x - x_0)}_{\text{approaches } 0}.$$

As x approaches x_0, the final term on the right approaches 0. We discover that

$$f'(x_0) = g'(x_0)h(x_0) + g(x_0)h'(x_0).$$

This rule is known as the "product rule".

EXERCISE 2.3. Use the product rule to compute the derivative of the function $f(x) = x^3$.

Solution: Notice that $f(x) = g(x)h(x)$, where $g(x) = x$ and $h(x) = x^2$. We saw earlier (in equation (1.3)) that $h'(x_0) = 2x_0$, and from Exercise (1.2) we know that $g'(x_0) = 1$. The product rule tells us that

$$f'(x_0) = \underset{\uparrow\, g'(x_0)}{1} \cdot \underset{\uparrow\, h(x_0)}{x_0^2} + \underset{\uparrow\, g(x_0)}{x_0} \cdot \underset{\uparrow\, h'(x_0)}{2x_0} = 3x_0^2.$$

EXERCISE 2.4. Use the product rule and the result of the previous exercise to compute the derivative of the function $f(x) = x^4$. Conjecture a formula for the derivative of $f(x) = x^n$, where n is a nonnegative integer.

Solution: We can use the same approach again, writing $f(x) = g(x)h(x)$ where $g(x) = x$ and $h(x) = x^3$. The product rule tells us that

$$f'(x_0) = 1 \cdot x_0^3 + x_0 \cdot (3x_0^2) = 4x_0^3.$$

The pattern is now clear. The derivative of $f(x) = x^n$, where n is any nonnegative integer, is

$$f'(x_0) = nx_0^{n-1}. \tag{2.2}$$

EXERCISE 2.5. Assume that the functions g and h are differentiable at x_0, and furthermore that $h(x_0) \neq 0$. Let f be the function defined by

$$f(x) = \frac{g(x)}{h(x)}$$

for all numbers x such that $h(x) \neq 0$. Use the product rule to compute the derivative $f'(x_0)$.

Solution: Notice that

$$f(x)h(x) = g(x). \tag{2.3}$$

Differentiating both sides of (2.3), and using the product rule to compute the derivative of the left-hand side, we obtain

$$\begin{aligned}
& f'(x_0)h(x_0) + f(x_0)h'(x_0) = g'(x_0) \\
\implies \quad & f'(x_0)h(x_0) + \frac{g(x_0)}{h(x_0)}h'(x_0) = g'(x_0) \\
\implies \quad & f'(x_0)h(x_0)^2 + g(x_0)h'(x_0) = h(x_0)g'(x_0) \\
\implies \quad & f'(x_0) = \frac{h(x_0)g'(x_0) - g(x_0)h'(x_0)}{h(x_0)^2}. \qquad (2.4)
\end{aligned}$$

This formula is known as the quotient rule.

EXERCISE 2.6. Use the quotient rule (2.4) and formula (2.2) to compute the derivative of the function $f(x) = 1/x^m$, where m is a positive integer.

Solution: Assume that $x_0 \neq 0$ (otherwise f is not defined at x_0). The quotient rule with $g(x) = 1$ and $h(x) = x^m$ tells us that

$$f'(x_0) = \frac{\overset{h(x_0)}{\downarrow}\ x_0^m \cdot \overset{g'(x_0)}{\downarrow}\ 0 - \overset{g(x_0)}{\downarrow}\ 1 \cdot \overset{h'(x_0)}{\downarrow}\ (mx_0^{m-1})}{\underset{h(x_0)^2}{\uparrow}\ (x_0^m)^2} = -mx_0^{-m-1}.$$

This shows that the formula (2.2) holds also when n is negative.

2.5. The chain rule

One important way to build a new function out of simpler functions is to take the output of one function and plug it in as input to another function. Suppose that

$$f(x) = g(h(x))$$

and that h is differentiable at x_0 and g is differentiable at $h(x_0)$. The function f is said to be the "composition" of g and h. For example, if $h(x) = 1 + x^2$ and $g(y) = 1/y$, then $f(x) = g(h(x)) = 1/(1 + x^2)$.

If $x \neq x_0$ then

$$\frac{f(x) - f(x_0)}{x - x_0} = \frac{g(h(x)) - g(h(x_0))}{x - x_0}.$$

We will use the fundamental strategy to simplify the expression on the right. If x is very close to x_0, then the point $y = h(x)$ is close to the point $y_0 = h(x_0)$. We plug these values of y and y_0 into Newton's approximation

$$g(y) \approx g(y_0) + g'(y_0)(y - y_0)$$

to obtain

$$g(h(x)) \approx g(h(x_0)) + g'(h(x_0))(h(x) - h(x_0)),$$

which implies that

$$\frac{g(h(x)) - g(h(x_0))}{x - x_0} \approx g'(h(x_0)) \underbrace{\left(\frac{h(x) - h(x_0)}{x - x_0} \right)}_{\text{approaches } h'(x_0)}.$$

As x approaches x_0, the quotient on the right approaches $h'(x_0)$. We discover that

$$f'(x_0) = g'(h(x_0))h'(x_0).$$

This fact is known as the chain rule.

2.6. Derivative of b^x

Suppose that $f(x) = b^x$, where $b > 1$ is a number. If $x \neq x_0$, then

$$\begin{aligned}\frac{f(x) - f(x_0)}{x - x_0} &= \frac{b^x - b^{x_0}}{x - x_0} \\ &= b^{x_0}\left(\frac{b^{x-x_0} - 1}{x - x_0}\right).\end{aligned}$$

The term b^{x_0} does not depend on x. Thus,

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = b^{x_0} \underbrace{\left(\lim_{x \to x_0} \frac{b^{x-x_0} - 1}{x - x_0}\right)}_{\text{annoying number}}.$$

The "annoying number" on the right needs a name, so for the moment let's call it c. The expression for c might look unpleasant, but keep in mind that c is just some number, and we can approximate c by plugging in a particular value of x that is very close to x_0.

It may seem that we are stuck with this unpleasant number, but there is a ray of hope: the value of c depends on the value of b. This raises a question: *Is it possible to find a special value of b such that $c = 1$?* It would be great if this were the case, because then c would vanish from sight and we would have the following simple and neat result:

$$f(x_0) = b^{x_0}, \qquad f'(x_0) = b^{x_0}$$

for every real number x_0. In other words, the function f would be equal to its own derivative.

In fact, the answer to the question is *yes*: there is a special value of b for which the number c turns out to be equal to 1. Let's find this special value of b. Assume that b is chosen so that $c = 1$. Select a value of x that is very close to x_0, and to save

Δx	$(1+\Delta x)^{1/\Delta x}$
.1	2.5937424
.01	2.7048138
.001	2.7169239
.0001	2.7181459

Table 1. As Δx approaches 0, the quantity $(1+\Delta x)^{1/\Delta x}$ approaches $e \approx 2.718$.

writing let $\Delta x = x - x_0$. Then

$$\begin{aligned} & \frac{b^{\Delta x}-1}{\Delta x} \approx 1 \\ \implies \quad & b^{\Delta x} \approx 1+\Delta x \\ \implies \quad & b \approx (1+\Delta x)^{1/\Delta x}. \end{aligned}$$

The approximation improves as $\Delta x = x - x_0$ approaches 0. So we have found that

$$b = \lim_{\Delta x \to 0} (1+\Delta x)^{1/\Delta x}.$$

This special value of b is universally known as "e", or Euler's number. The number e is a fundamental mathematical constant, like π. We estimate the value of e in table 1, where we see that $e \approx 2.718$. In conclusion: if $f(x) = e^x$, then

$$f'(x_0) = e^{x_0}.$$

The function e^x is called the "exponential function", and thanks to this special property it is arguably the most important function in math:

The exponential function is its own derivative.

EXERCISE 2.7. Compute the derivative of the function $f(x) = e^{-x}$.

Solution: Note that $f(x) = g(h(x))$ where $h(x) = -x$ and $g(y) = e^y$. The chain rule tells us that

$$f'(x_0) = \underset{\uparrow\, g'(h(x_0))}{e^{-x_0}} \cdot \underset{\uparrow\, h'(x_0)}{(-1)} = -e^{-x_0}.$$

EXERCISE 2.8. The *sigmoid* function S is defined by

$$S(x) = \frac{e^x}{1 + e^x}.$$

Its graph is shown in figure 3. The output of the sigmoid function

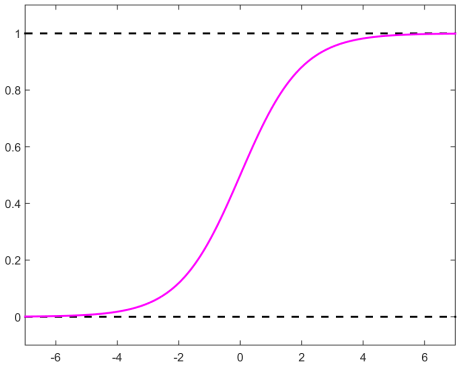


Figure 3. The sigmoid function $S(x) = e^x/(1 + e^x)$.

is always between 0 and 1, which makes the sigmoid function useful for estimating probabilities (which are required to be between 0 and 1). For this reason, the sigmoid function plays an important role in machine learning, for example in logistic regression and neural networks. (A typical application might be computing the probability that an email is spam.)

Use the quotient rule to compute the derivative $S'(x_0)$.

Solution: The quotient rule tells us that

$$\begin{aligned} S'(x_0) &= \frac{(1+e^{x_0})e^{x_0} - e^{x_0}\cdot e^{x_0}}{(1+e^{x_0})^2} \\ &= \frac{e^{x_0}}{(1+e^{x_0})^2}. \end{aligned}$$

We are done, but if we notice that

$$1 - S(x) = \frac{1+e^x}{1+e^x} - \frac{e^x}{1+e^x} = \frac{1}{1+e^x},$$

then our formula for $S'(x_0)$ can be written as

$$S'(x_0) = \frac{e^{x_0}}{1+e^{x_0}} \cdot \frac{1}{1+e^{x_0}} = S(x_0)(1 - S(x_0)).$$

This formula is convenient because it shows that if we have already evaluated $S(x_0)$, then hardly any additional arithmetic operations are required to evaluate $S'(x_0)$. We can compute $S'(x_0)$ very efficiently.

2.7. Derivative of $\log(x)$

The defining property of the natural logarithm function $f(x) = \log(x)$ is

$$e^{f(x)} = x \tag{2.5}$$

for every positive number x. Differentiating both sides of (2.5), and using the chain rule to compute the derivative of the left-hand side, we find that

$$\begin{aligned} & e^{f(x)} f'(x) = 1 \\ \implies \quad & x f'(x) = 1 \\ \implies \quad & \boxed{f'(x) = \tfrac{1}{x}.} \end{aligned}$$

2.8. The power rule

Let r be any real number and let f be the function defined by $f(x) = x^r$. (Here x can be any positive real number.) To derive a formula for the derivative $f'(x_0)$, we first observe that

$$f(x) = (e^{\log(x)})^r = e^{r\log(x)} = e^{h(x)},$$

where $h(x) = r\log(x)$. The chain rule tells us that

$$\begin{aligned} f'(x) &= e^{h(x)}h'(x) \\ &= x^r\left(\frac{r}{x}\right) \\ &= rx^{r-1}. \end{aligned}$$

This formula is known as the power rule. As we have seen previously (in exercises 2.4 and 2.6), if r is an integer then there is no need to restrict x to be positive. (If r is not an integer, then x^r might not even be defined when x is negative. For example, $(-1)^{1/2} = \sqrt{-1}$ is not a real number.)

Exercise 2.9. Use Newton's approximation to estimate $\sqrt{50}$.

Solution: Let $f(x) = \sqrt{x} = x^{1/2}$. From the power rule, $f'(x_0) = (1/2)x_0^{-1/2}$. Newton's approximation (1.5) with $x = 50$ and $x_0 = 49$ yields

$$\underset{\uparrow f(x)}{\sqrt{50}} \approx \underset{\uparrow f(x_0)}{\sqrt{49}} + \underset{\uparrow f'(x_0)}{\frac{1}{2\sqrt{49}}} \cdot \underbrace{(50-49)}_{\uparrow (x-x_0)}.$$

Simplifying, we obtain

$$\sqrt{50} \approx 7 + \frac{1}{14} \approx 7.07142.$$

The true value of $\sqrt{50}$ is $7.07106\ldots$, so the approximation is not bad.

Part 2

Vector calculus and linear algebra

CHAPTER 3

Points and vectors

> One of the miseries of life is that everybody names things a little bit wrong, and so it makes everything a little harder to understand.
>
> Richard Feynman

The word "vector" can have different meanings in different mathematical contexts, which can be confusing. The same goes for the word "point". But it's a shame for such simple concepts to be a source of confusion. Here we will explain the meanings of the words "point" and "vector" and pin down the definitions that will be used throughout this book.

A classic example of a data science problem is predicting the value of a house based on information such as the house's square footage, the age of the house, the distance of the house from downtown, the number of restaurants within walking distance, etc. When collecting data about a house, inevitably we write down a list of numbers like this. An *n-tuple* is simply a list of n numbers. For example, here is a 4-tuple: $(850, 10, 15.7, 4)$. The order in which the numbers are written down matters, so that a rearrangement such as $(15.7, 10, 4, 850)$ is considered to be a different n-tuple. Two n-tuples $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are equal if and only if $x_1 = y_1, x_2 = y_2, \ldots,$ and $x_n = y_n$.

The set of all n-tuples of real numbers is denoted $\mathbb{R}^n$. The notation $x \in \mathbb{R}^n$ means that x is an element of the set $\mathbb{R}^n$, so

$x = (x_1, \ldots, x_n)$ for some real numbers $x_1, \ldots, x_n$. The numbers $x_1, \ldots, x_n$ are called the "components" of x.

If we have carefully collected a large amount of data about a house, we can easily find ourselves working with n-tuples where n is some large number such as 50. Much larger values of n are typical in other applications. For example, if instead of predicting housing prices we are working on a computer vision problem, such as developing an algorithm to recognize people in images, we might describe an image by listing the RGB values for each pixel in the image. This gives us an n-tuple where n is perhaps one million.

When $n = 2$ or $n = 3$, it is possible to visualize an n-tuple. In fact, there are two different methods to visualize an n-tuple: the *point picture* and the *vector picture*, which we explain below.

3.1. Method 1: The point picture

The first method to visualize a 2-tuple such as $(3, 2)$ is simply to draw a coordinate system and then draw the point whose coordinates are $(3, 2)$, as shown in figure 4. In this viewpoint, the ordered pair $(3, 2)$ specifies a *location*. If we draw a third axis coming out of the page, then we are able to visualize 3-tuples.

When we use Method 1 to visualize an n-tuple, we often refer to the n-tuple as a "point". So in our terminology, a "point" really is nothing more than an n-tuple, but using the term "point" provides a hint that it will be helpful to visualize the n-tuple as a location in space.

3.2. Method 2: The vector picture

In the second method, we visualize a 2-tuple such as $(3, 2)$ by drawing a coordinate system and then drawing an arrow that connects a starting point (selected arbitrarily) to an ending point which is 3 units to the right and 2 units up from the starting point. In this viewpoint, which is illustrated in figure 5, the ordered pair $(3, 2)$ tells us the *displacement* from a starting point

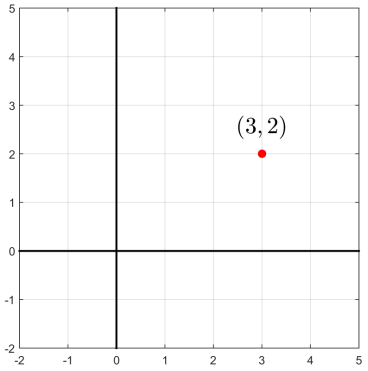


Figure 4. The point picture: We visualize the point in space whose coordinates are $(3, 2)$. Starting at the origin, you move 3 units to the right and 2 units upwards to arrive at the location shown in red.

to an ending point. If we draw a third axis coming out of the page, then we can use the same method to visualize 3-tuples.

When we use Method 2 to visualize an n-tuple, we often refer to the n-tuple as a vector. In our terminology, a "vector" is nothing more than an n-tuple, but using the term "vector" suggests that we should visualize the n-tuple as the displacement (drawn as an arrow) from one point to another.

3.3. Vector operations

The vector picture suggests doing certain things with vectors that we would never think of doing with points. A shift in viewpoint leads to new ideas.

3.3.1. Adding vectors. It does not seem to make sense visually to add together two points. However, there is a perfectly logical way to add two displacements: if $x = (x_1, x_2)$ is the displacement from location A to location B, and $y = (y_1, y_2)$ is

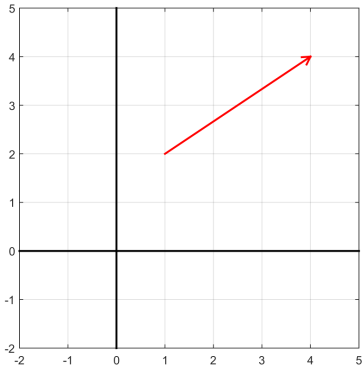


Figure 5. The vector picture: We visualize the *displacement* from a starting point (selected arbitrarily) to an ending point. In this example, the arbitrarily selected starting point has coordinates $(1, 2)$. You move 3 units to the right and 2 units upwards to arrive at the ending point, which has coordinates $(4, 4)$.

the displacement from location B to location C, then $x + y$ is the total displacement from location A to location C. If an object is located at point A and experiences a displacement of x, followed by a displacement of y, then the object's total displacement is $x + y$. This is illustrated in figure 6. Hopefully figure 6 makes it clear that

$$(x_1, x_2) + (y_1, y_2) = (x_1 + x_2, y_1 + y_2).$$

A similar formula (and a similar picture) holds for vectors in $\mathbb{R}^3$. Although we cannot visualize n-tuples when n is greater than 3, we still define the sum of vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ as follows:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n).$$

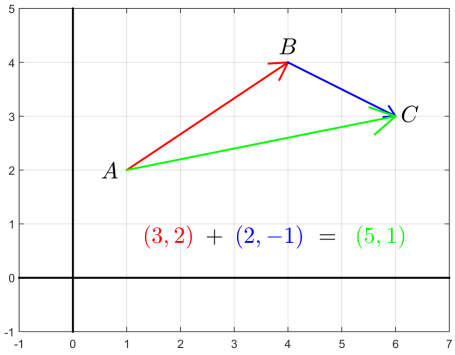


Figure 6. Adding two vectors: The displacement from A to B plus the displacement from B to C is equal to the displacement from A to C.

3.3.2. Multiplying a vector by a number. Visually, it would not make sense to multiply a point by a number. But it is natural to multiply a vector x by a number c. Picturing x as an arrow (representing a displacement), you just scale the length of the arrow by c without changing the arrow's direction. If x is visualized as a displacement, then $2x$ is twice the displacement, in the same direction. This is illustrated in figure 7. Hopefully figure 7 makes it clear that if $x = (x_1, x_2)$ and c is a number, then

$$cx = (cx_1, cx_2).$$

Although we can't visualize vectors in $\mathbb{R}^n$ when $n > 3$, we still define

$$c(x_1, \ldots, x_n) = (cx_1, \ldots, cx_n).$$

If c is negative, then cx points in the direction opposite to that of x. When a number c is multiplied by a vector x, often c is called a "scalar" because the length of x gets scaled by the factor c.

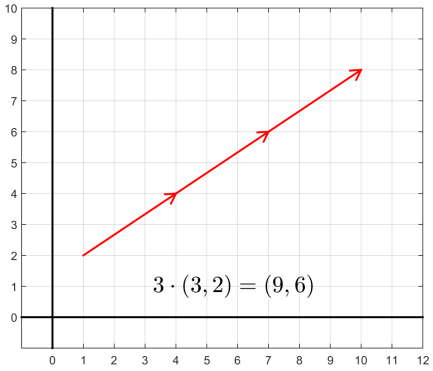


Figure 7. Multiplying a vector by a scalar: The magnitude (length) of the displacement is multiplied by c, while the direction of the displacement is unchanged.

3.3.3. The length or "norm" of a vector. A point has no "size", and it would not make sense to describe one point as being somehow larger or smaller than another point. But some *displacements* are larger than other displacements. If a vector $x = (x_1, x_2)$ is visualized as an arrow (representing a displacement from one point to another), then the length of the arrow tells us the size of the displacement. The length of a vector x is also called the "norm" of x, and is denoted as $\|x\|$. Using the Pythagorean theorem, we can see that the norm of the vector $x = (x_1, x_2)$ is

$$\|x\| = \sqrt{x_1^2 + x_2^2},$$

as illustrated in figure 8.

Figure 9 shows how we can use the Pythagorean theorem twice to compute the length of a vector $x = (x_1, x_2, x_3)$, obtaining the formula

$$\|x\| = \sqrt{x_1^2 + x_2^2 + x_3^2}.$$

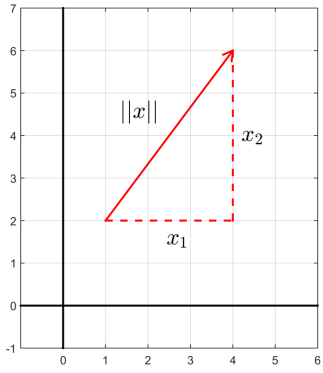


Figure 8. The length or "norm" of a vector $x = (x_1, x_2)$: The Pythagorean theorem tells us that the length of the arrow is $\|x\| = \sqrt{x_1^2 + x_2^2}$. In this example, $x = (3, 4)$ and $\|x\| = \sqrt{3^2 + 4^2} = 5$.

Although we can't visualize vectors in $\mathbb{R}^n$ when $n > 3$, we define the norm of a vector $x = (x_1, \ldots, x_n)$ by

$$\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}.$$

Let's check that if we multiply a vector $x = (x_1, \ldots, x_n)$ by a scalar c, the norm of the resulting vector is $|c|\|x\|$. By definition, $cx = (cx_1, \ldots, cx_n)$. It follows that

$$\begin{aligned} \|cx\| &= \sqrt{(cx_1)^2 + \cdots + (cx_n)^2} \\ &= |c|\sqrt{x_1^2 + \cdots + x_n^2} \\ &= |c|\|x\|. \end{aligned}$$

A "unit vector" is a vector u whose norm is 1. Unit vectors are convenient for specifying a *direction* in space. If u is a unit

vector, and t is a scalar, then the norm of tu is equal to $|t|$:

$$\|tu\| = |t|\|u\| = |t| \cdot 1 = |t|.$$

So tu is a vector that points in the direction u and has length t. The vector tu represents a displacement of length t in the direction u.

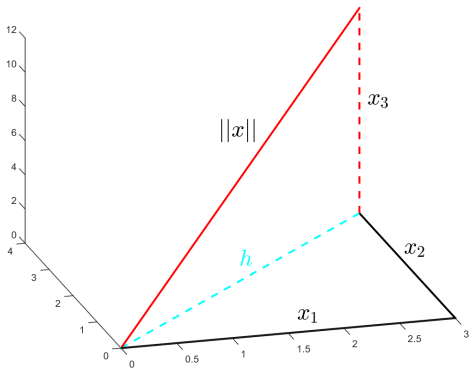


Figure 9. The length of a vector $x = (x_1, x_2, x_3)$: We use the Pythagorean theorem to see that the length of the blue line is $h = \sqrt{x_1^2 + x_2^2}$. We then use the Pythagorean theorem again to see that $\|x\| = \sqrt{h^2 + x_3^2} = \sqrt{x_1^2 + x_2^2 + x_3^2}$. In this particular example, $h = \sqrt{3^2 + 4^2} = 5$, and $\|x\| = \sqrt{5^2 + 12^2} = 13$.

There are other popular ways to measure the "size" of a vector $x = (x_1, \ldots, x_n)$. For example, the quantity $|x_1| + \cdots + |x_n|$ is called the "taxicab norm" of x. If x is the displacement from point A to point B, then the taxicab norm tells us the distance we must travel in order to get from A to B if we are only allowed to move in directions that are parallel to one of the coordinate axes. (Imagine driving in a city where all roads go either East-West or

North-South.) The taxicab norm is more commonly called the ℓ_1-norm (a less descriptive name), and is denoted $\|x\|_1$. Another way to measure the size of a vector x is using the "worst-case norm", which is equal to the largest (in absolute value) component of x and is denoted $\|x\|_\infty$. For example, if $x = (-3, 5, -9)$, then $\|x\|_\infty = 9$. The worst-case norm is usually called the ℓ_∞-norm. The length of the vector x, which we have been simply calling the "norm" of x, is more precisely called the ℓ_2-norm of x, in order to distinguish it from these other norms that we have now discussed. The ℓ_2-norm of x is commonly denoted $\|x\|_2$:

$$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} = \text{length of } x,$$

but we will continue to use the notation $\|x\|$ for the length of x.

3.3.4. Adding a vector to a point. It also makes sense visually to add a vector to a point, to obtain a new *point*. If x is a point and y is a vector, then $x + y$ is the point you would arrive at by starting at the location x and moving through a displacement of y. This is illustrated in figure 10. Although the geometric interpretation is different, the addition *formula* remains the same:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n).$$

In calculus, we often want to compare the value of a function f at a point x with the value of f at some nearby point that is very close to x. Let u be a unit vector (so that $\|u\| = 1$) and let t be a tiny number. Notice that $\|tu\| = |t|$, so that tu is a tiny vector that represents a tiny displacement in the direction u. Then $x + tu$ is a point nearby x. Specifically, $x + tu$ is the point you would arrive at by starting at the location x and moving a distance t in the direction u. If that is not perfectly clear, take a moment to let it sink in, as this picture will be crucial in multivariable calculus.

3.3.5. Subtracting a point from a point. Although it does not make sense visually to add a point to a point, it does

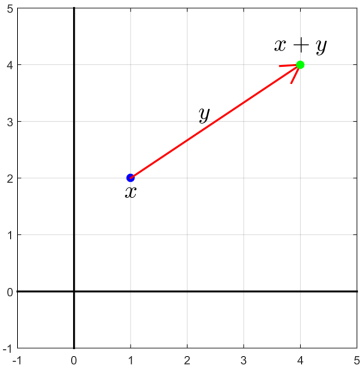


Figure 10. Adding a point x and a vector y: $x+y$ is the *point* whose displacement from x is y. In this example, $x = (1,2), y = (3,2)$, and $x + y = (4,4)$.

make sense to *subtract* a point from a point to obtain a vector. The previous section could be summarized as

$$\text{point} + \text{vector} = \text{new point}.$$

We can rewrite this equation as

$$\text{new point} - \text{point} = \text{vector}.$$

If x and y are points, then $y - x$ is the displacement vector from x to y. This is illustrated in figure 11. Notice that

$$\underbrace{x}_{\text{point}} + \underbrace{(y-x)}_{\text{vector}} = \underbrace{y}_{\text{point}}.$$

3.3.6. The dot product of two vectors. There is something called the "dot product" of vectors x and y that turns out to be very useful. If $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ then the dot product of x and y is denoted $\langle x, y \rangle$ and is defined as

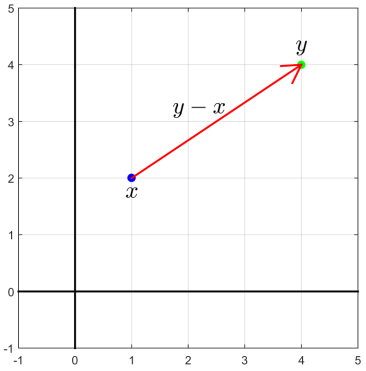


Figure 11. Subtracting a point x from a point y: $y - x$ is the displacement *vector* from x to y. In this example, $x = (1, 2), y = (4, 4)$, and $y - x = (3, 2)$.

follows:

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n. \tag{3.1}$$

Notice that the dot product of x and y is a number, not a vector. And why should we care about this number? The main reason is that this number has a nice geometric interpretation: If θ is the angle between the vectors x and y, then

$$\langle x, y \rangle = \|x\| \|y\| \cos(\theta).$$

This is illustrated in figure 12. The geometric interpretation of the dot product is not obvious, but we'll show where it comes from at the end of this section.

If x and y are perpendicular, then $\theta = \pi/2$, and so $\cos(\theta) = 0$. Using the geometric interpretation of the dot product, we see that $\langle x, y \rangle = 0$. This gives us a convenient way to check whether or not two vectors are perpendicular — we simply check if their dot product is equal to 0. Usually the word "orthogonal" is used

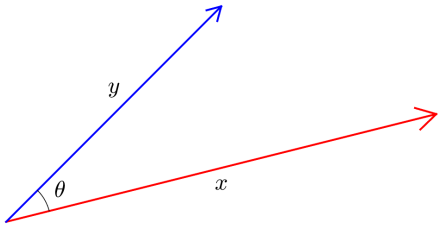


Figure 12. The geometric interpretation of the dot product.

instead of "perpendicular", perhaps because it sounds fancier, but both words mean the same thing:

$$x \text{ is orthogonal to } y \iff \langle x, y\rangle = 0.$$

Here is a question that will be important for vector calculus: which way should y be pointing in order for $\langle x, y\rangle$ to be as large as possible? Choosing y to be orthogonal to x would be a bad choice, because then $\langle x, y\rangle = 0$. Recalling that $\langle x, y\rangle = \|x\|\|y\|\cos(\theta)$, and noting that $-1 \leq \cos(\theta) \leq 1$, we see that $\langle x, y\rangle$ is as large as possible when $\cos(\theta) = 1$, or equivalently when θ is 0. This means that y should point in the *same direction* as x. In this case, $\langle x, y\rangle = \|x\|\|y\|$. On the other hand, if we want $\langle x, y\rangle$ to be as negative as possible, then we should pick y to point in the opposite direction as x, so that $\cos(\theta) = -1$ and $\langle x, y\rangle = -\|x\|\|y\|$.

We now mention a few rules that tend to be useful when working with vectors. Using equation (3.1), we see immediately that

$$\langle x, x\rangle = \|x\|^2$$

for all vectors $x \in \mathbb{R}^n$. We can also use equation (3.1) to easily check that

- $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$ for all $x, y, z \in \mathbb{R}^n$.
- $\langle cx, y\rangle = c\langle x, y\rangle$ for all $x, y \in \mathbb{R}^n, c \in \mathbb{R}$.

Because $\langle x, y\rangle = \langle y, x\rangle$ for all $x, y \in \mathbb{R}^n$, we have the following counterparts for the previous two rules:

- $\langle x, y + z\rangle = \langle x, y\rangle + \langle x, z\rangle$ for all $x, y, z \in \mathbb{R}^n$.
- $\langle x, cy\rangle = c\langle x, y\rangle$ for all $x, y \in \mathbb{R}^n, c \in \mathbb{R}$.

The fact that the dot product formula is simple and has these nice properties is one of the reasons that the dot product is so ubiquitous. It's important to internalize these simple rules so that vector arithmetic becomes effortless.

Finally let's justify the geometric interpretation of the dot product. In figure 13, it would be incorrect to invoke the

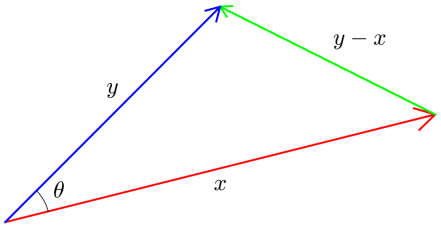


Figure 13. The law of cosines is a generalization of the Pythagorean theorem that holds for non-right triangles. The geometric interpretation of the dot product follows from the law of cosines.

Pythagorean theorem to conclude that $\|y - x\|^2 = \|x\|^2 + \|y\|^2$,

because the triangle in figure 13 is not a right triangle. However, there is a generalization of the Pythagorean theorem known as the law of cosines which tells us that

$$\|y - x\|^2 = \|x\|^2 + \|y\|^2 - 2\|x\|\|y\| \cos(\theta). \tag{3.2}$$

Notice that when $\theta = \pi/2$, the law of cosines reduces to the Pythagorean theorem. We can expand the left-hand side of (3.2) as follows:

$$\begin{aligned} \|y - x\|^2 &= \langle y - x, y - x \rangle \\ &= \langle y - x, y \rangle - \langle y - x, x \rangle \\ &= \langle y, y \rangle - \langle x, y \rangle - \langle y, x \rangle + \langle x, x \rangle \\ &= \|x\|^2 + \|y\|^2 - 2\langle x, y \rangle. \end{aligned}$$

Now comparing the left-hand side of (3.2) (in expanded form) with the right-hand side, and canceling like terms from both sides, we discover that $\langle x, y \rangle = \|x\|\|y\| \cos(\theta)$.

CHAPTER 4

The gradient vector

So far we have worked with functions that take a single number as input (for example, the number t of seconds for which a car has been driving) and return a single number as output (for example, the car's position in meters at time t). Our next step is to consider functions that take a *list* of numbers as input and return a single number as output. For example, we could have a function f that takes the coordinates $x = (x_1, x_2, x_3)$ of a point in space as input and returns the temperature at the point x as output. If you are a mosquito who likes cool weather, you might want to find a point x for which the temperature $f(x)$ is as small as possible, and fly to that location.

EXAMPLE 4.1. Here is an example of how functions that take a list of numbers as input appear in prediction problems such as predicting the price of a house. Suppose that we have collected the following data about a large number N of recently sold houses:

- S_i is the number of square feet of the ith house.
- T_i is the age of the ith house.
- R_i is the distance of the ith house from downtown.
- y_i is the selling price of the ith house.

The numbers S_i, T_i, and R_i describe the ith house, and our goal is to use these numbers somehow to predict the selling price y_i of the ith house.

A simple and popular approach, called **linear regression**, is to try to find numbers w_0, w_1, w_2, and w_3 such that $w_0 + S_i w_1 +$

$T_i w_2 + R_i w_3$ is a good approximation to y_i, the price of the ith house. We want:

$$y_i \approx w_0 + S_i w_1 + T_i w_2 + R_i w_3 \tag{4.1}$$

for $i = 1, \ldots, N$. The challenge is to select the numbers w_0, w_1, w_2, w_3 so that the approximation (4.1) is accurate for *all* of the houses for which we have collected data. It is not good enough for the prediction error $w_0 + S_i w_1 + T_i w_2 + R_i w_3 - y_i$ to be small for just some houses but not others. We want this prediction error to be small for *all* houses. In other words, we want the *total* prediction error

$$f(w_0, w_1, w_2, w_3) = \sum_{i=1}^{N} (w_0 + S_i w_1 + T_i w_2 + R_i w_3 - y_i)^2$$

to be as small as possible.

This example shows how a goal of making accurate predictions naturally leads us to the goal of minimizing an error function such as f that takes a list of numbers as input. Many of the most popular machine learning algorithms, such as neural networks, are variations of this simple, classic idea.

4.1. The directional derivative

It will help the mosquito find a cool spot if he can feel how rapidly the temperature is changing in various directions. Suppose that we are currently located at a point

$$x = (x_1, \ldots, x_n),$$

and the temperature at this point is $f(x)$. (This function f is like a thermometer.) We are wondering how rapidly the temperature will increase or decrease if we move away from x in the direction u. (Here u is a unit vector.) If we move a short distance of t meters in the direction u, then our new location is $x + tu$, and the temperature at our new location is $f(x + tu)$. The change in temperature is $f(x + tu) - f(x)$. The average rate of change in

the temperature is

$$\frac{\text{change in temperature}}{\text{distance traveled}} = \frac{f(x+tu)-f(x)}{t} \quad \frac{\text{degrees}}{\text{meter}}.$$

If we select values of t that are shorter and shorter, and approaching 0, this average rate of change approaches the *instantaneous* rate of change of f in the direction u. This instantaneous rate of change is denoted $D_u f(x)$:

$$D_u f(x) = \lim_{t\to 0} \frac{f(x+tu)-f(x)}{t}. \tag{4.2}$$

The number $D_u f(x)$ is called the "directional derivative" of f at x in the direction u.

EXERCISE 4.2. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x_1, x_2) = x_1 x_2.$$

Compute $D_u f(3,2)$, where u is the unit vector $(1/\sqrt{2}, 1/\sqrt{2})$.

Solution:

$$\begin{aligned} D_u f(3,2) &= \lim_{t\to 0} \frac{(3+t/\sqrt{2})(2+t/\sqrt{2})-6}{t} \\ &= \lim_{t\to 0} \frac{6+5t/\sqrt{2}+t^2/2-6}{t} \\ &= \lim_{t\to 0} 5/\sqrt{2}+t/2 \\ &= 5/\sqrt{2}. \end{aligned}$$

4.2. Partial derivatives

In the special case that $u = (1, 0, \ldots, 0)$, computing the directional derivative $D_u f(x)$ is particularly easy. With this choice of u, we have

$$\begin{aligned} x + tu &= (x_1, x_2, \ldots, x_n) + t(1, 0, \ldots, 0) \\ &= (x_1 + t, x_2, \ldots, x_n) \end{aligned}$$

and

$$D_u f(x) = \lim_{t \to 0} \frac{f(x_1 + t, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{t}. \quad (4.3)$$

The quantity on the right is the derivative of the function $g : \mathbb{R} \to \mathbb{R}$ defined by

$$g(x_1) = f(x_1, x_2, \ldots, x_n). \quad (4.4)$$

(When defining g, we are thinking of the numbers $x_2, \ldots, x_n$ as being held fixed, whereas x_1 can be any real number.) In other words,

$$D_u f(x) = g'(x_1).$$

This is nice because g is a function of a *single variable*, which means that $g'(x_1)$ can be computed using the arsenal of formulas for derivatives that we derived in chapter 2.

When $u = (1, 0, \ldots, 0)$, an alternative notation for the directional derivative $D_u f(x)$ is $D_1 f(x)$. Likewise, in the special case where the vector u has a 1 in the ith position and zeros elsewhere, an alternative notation for the directional derivative $D_u f(x)$ is $D_i f(x)$. These numbers $D_i f(x)$ (for $i = 1, \ldots, n$) are called the **partial derivatives** of f at x. Computing $D_i f(x)$ is easy for the same reason that computing $D_1 f(x)$ is easy:

> To compute $D_i f(x)$, think of f as a function of x_i alone (with the other components of x held fixed to constant values), and then take the derivative using single-variable calculus techniques from chapter 2.

Note that another very common notation for $D_i f(x)$ that you will see in other books is $\frac{\partial f(x)}{\partial x_i}$.

EXERCISE 4.3. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined by

$$f(x_1, x_2, x_3) = x_1 e^{x_2 x_3}.$$

Compute the partial derivatives of f.

Solution: Thinking of $x_1 e^{x_2 x_3}$ as a function of x_1 alone, with x_2 and x_3 held fixed, we see that

$$D_1 f(x) = e^{x_2 x_3}.$$

On the other hand, thinking of $x_1 e^{x_2 x_3}$ as a function of x_2 alone, we see that

$$D_2 f(x) = x_1 e^{x_2 x_3} x_3.$$

Finally, by thinking of $x_1 e^{x_2 x_3}$ as a function of x_3 alone, we see that

$$D_3 f(x) = x_1 e^{x_2 x_3} x_2.$$

4.3. Newton's approximation for partial derivatives

Newton's approximation for the function g in equation (4.4) tells us that $g(x_1 + \Delta x_1) \approx g(x_1) + g'(x_1)\Delta x_1$, or equivalently

$$f(x_1 + \Delta x_1, x_2, \ldots, x_n) \approx f(x) + D_1 f(x)\Delta x_1.$$

Similarly, if we increase x_i by a small amount Δx_i and leave the other components of x unchanged, then we have the following version of Newton's approximation for partial derivatives:

$$f(x_1, \ldots, x_i + \Delta x_i, \ldots, x_n) \approx f(x) + D_i f(x)\Delta x_i. \tag{4.5}$$

In words, if you start at a point x and move a distance Δx_i in the direction of the ith axis, then the change in the value of f is approximately $D_i f(x)\Delta x_i$.

4.4. Newton's approximation when $f : \mathbb{R}^n \to \mathbb{R}$

Newton's approximation answers the fundamental question of calculus: how much does the value of f change when its input changes by a small amount Δx? We now address this question in the case where $f : \mathbb{R}^n \to \mathbb{R}$ and $\Delta x \in \mathbb{R}^n$. To keep notation simple we look at the case where $n = 2$.

Let Δf be the amount that f changes when its input changes from point $A = (x_1, x_2)$ to point $C = (x_1 + \Delta x_1, x_2 + \Delta x_2)$. Notice that

$$\Delta f = \Delta f_1 + \Delta f_2,$$

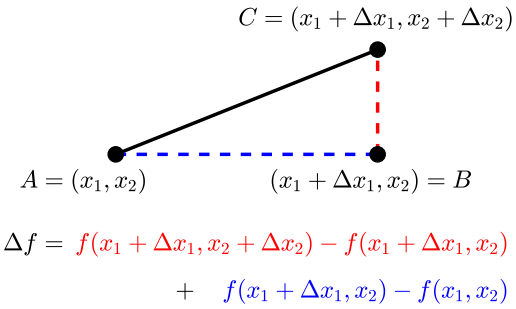


Figure 14. The change in temperature when moving from point A to point C is equal to the change in temperature from point A to point B plus the change in temperature from point B to point C. Algebraically, the second term in red cancels with the first term in blue.

where Δf_1 is the amount that f changes when its input changes from point A to point $B = (x_1 + \Delta x_1, x_2)$, and Δf_2 is the amount that f changes when its input changes from point B to point C. This is illustrated in figure 14.

According to equation (4.5), if Δx_1 and Δx_2 are small numbers then

$$\Delta f_1 \approx D_1 f(x_1, x_2) \Delta x_1$$

and

$$\begin{aligned} \Delta f_2 &\approx D_2 f(x_1 + \Delta x_1, x_2) \Delta x_2 \\ &\approx D_2 f(x_1, x_2) \Delta x_2. \end{aligned}$$

Putting these pieces together, we find that

$$\Delta f \approx D_1 f(x_1, x_2) \Delta x_1 + D_2 f(x_1, x_2) \Delta x_2. \tag{4.6}$$

In words, as you move from A to B to C, the value of f changes first by $D_1 f(x) \Delta x_1$ and then by $D_2 f(x) \Delta x_2$.

The expression on the right in (4.6) looks like the dot product of two vectors. This suggests that equation (4.6) can be written

more concisely if we introduce the vector $(D_1 f(x), \ldots, D_n f(x))$, which we shall call the **gradient** of f at x and which is denoted by $\nabla f(x)$. The gradient vector $\nabla f(x)$ is just a list of all the partial derivatives of f at x. With this notation, equation (4.6) becomes

$$\Delta f \approx \langle \nabla f(x), \Delta x \rangle. \tag{4.7}$$

Equivalently:

$$f(x + \Delta x) \approx f(x) + \langle \nabla f(x), \Delta x \rangle. \tag{4.8}$$

This is Newton's approximation in the case that $f : \mathbb{R}^n \to \mathbb{R}$. (Although we took $n = 2$ for simplicity, a similar derivation works for any value of n.)

4.5. A formula for directional derivatives

We can easily discover a formula for directional derivatives by using Newton's approximation with $\Delta x = tu$:

$$\begin{aligned} D_u f(x) &= \lim_{t \to 0} \frac{f(x + tu) - f(x)}{t} \\ &= \lim_{t \to 0} \frac{f(x) + \langle \nabla f(x), tu \rangle - f(x)}{t} \\ &= \lim_{t \to 0} \frac{t \langle \nabla f(x), u \rangle}{t} \\ &= \langle \nabla f(x), u \rangle. \end{aligned}$$

According to this formula, to compute the directional derivative $D_u f(x)$ we can just take the dot product of $\nabla f(x)$ with u:

$$D_u f(x) = \langle \nabla f(x), u \rangle.$$

4.6. The direction of steepest ascent

For which direction u is the directional derivative $D_u f(x)$ as large as possible? Notice that if u is a unit vector then

$$D_u f(x) = \langle \nabla f(x), u \rangle = \|\nabla f(x)\| \|u\| \cos(\theta) = \|\nabla f(x)\| \cos(\theta),$$

where θ is the angle between $\nabla f(x)$ and u. The term $\cos(\theta)$ is always between -1 and 1, so the largest possible value that $D_u f(x)$ can have is $\|\nabla f(x)\| \cdot 1$. This occurs when $\theta = 0$, which means that u points in the same direction as $\nabla f(x)$. Thus:

> The gradient vector points in the direction of steepest ascent.

Moreover, the magnitude of the gradient has a meaning also: if u is a unit vector that points in this direction of steepest ascent, then $D_u f(x) = \|\nabla f(x)\|$.

So the gradient vector tells you two things: the direction of steepest ascent, and the rate of change of f in that direction. With this visual interpretation, the gradient vector has sprung to life as a geometric object.

Similarly, $-\nabla f(x)$ points in the direction of steepest *descent*. If $f(x)$ is the temperature at the point x, and you are a mosquito who likes cool temperatures, you will want to fly in the direction $-\nabla f(x)$.

CHAPTER 5

The Jacobian matrix

Often we encounter functions that take a list of numbers as input and return a *list* of numbers as output. For example, the input could be an image of a handwritten digit (stored as a list of pixel intensity values), and the output could be a list of ten probabilities: the probability that the digit is a 0, the probability that it is a 1, etc. Our goal now is to discover a version of Newton's approximation for such a function $f : \mathbb{R}^n \to \mathbb{R}^m$. As always, the purpose of Newton's approximation is to estimate the value of $f(x + \Delta x)$.

To be concrete, suppose that $f : \mathbb{R}^3 \to \mathbb{R}^3$. If x is a point in $\mathbb{R}^3$, then $f(x)$ is a list of three numbers, which could be called $f_1(x), f_2(x)$, and $f_3(x)$. The functions $f_i : \mathbb{R}^3 \to \mathbb{R}$ defined in this way are called the **component functions** of f:

$$f(x) = (f_1(x), f_2(x), f_3(x)) \quad \text{for all points } x \text{ in } \mathbb{R}^3.$$

In other words, $f_i(x)$ is by definition the ith component of the point $f(x)$.

If $\Delta x = (\Delta x_1, \Delta x_2, \Delta x_3)$ is a small vector in $\mathbb{R}^3$, then Newton's approximation (4.8) for the function f_1 tells us that

$$f_1(x+\Delta x) \approx f_1(x) + D_1 f_1(x)\Delta x_1 + D_2 f_1(x)\Delta x_2 + D_3 f_1(x)\Delta x_3.$$

We have similar approximations for $f_2(x + \Delta x)$ and $f_3(x + \Delta x)$. Using these approximations for the component functions of f allows us to approximate $f(x+\Delta x)$. At this point, our equations will start to look nicer if we sometimes write vectors vertically

instead of horizontally:

$$f(x+\Delta x) = \begin{bmatrix} f_1(x+\Delta x) \\ f_2(x+\Delta x) \\ f_3(x+\Delta x) \end{bmatrix}$$

$$\approx \begin{bmatrix} f_1(x) + D_1 f_1(x)\Delta x_1 + D_2 f_1(x)\Delta x_2 + D_3 f_1(x)\Delta x_3 \\ f_2(x) + D_1 f_2(x)\Delta x_1 + D_2 f_2(x)\Delta x_2 + D_3 f_2(x)\Delta x_3 \\ f_3(x) + D_1 f_3(x)\Delta x_1 + D_2 f_3(x)\Delta x_2 + D_3 f_3(x)\Delta x_3 \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} f_1(x) \\ f_2(x) \\ f_3(x) \end{bmatrix}}_{f(x)} + \underbrace{\begin{bmatrix} D_1 f_1(x)\Delta x_1 + D_2 f_1(x)\Delta x_2 + D_3 f_1(x)\Delta x_3 \\ D_1 f_2(x)\Delta x_1 + D_2 f_2(x)\Delta x_2 + D_3 f_2(x)\Delta x_3 \\ D_1 f_3(x)\Delta x_1 + D_2 f_3(x)\Delta x_2 + D_3 f_3(x)\Delta x_3 \end{bmatrix}}_{\text{Need concise notation for this.}}.$$

The messy-looking expression on the right can be written more concisely if we introduce some new notation. There is something wasteful about the expression on the right, because the symbols $\Delta x_1, \Delta x_2$, and Δx_3 are written repeatedly, once on each row. **We have wasted ink.** The same information can be conveyed with less writing if we only write down the arrays of numbers

$$\begin{bmatrix} D_1 f_1(x) & D_2 f_1(x) & D_3 f_1(x) \\ D_1 f_2(x) & D_2 f_2(x) & D_3 f_2(x) \\ D_1 f_3(x) & D_2 f_3(x) & D_3 f_3(x) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \end{bmatrix}$$

and place them side by side. In other words, we will now declare that the expression

$$\begin{bmatrix} D_1 f_1(x) & D_2 f_1(x) & D_3 f_1(x) \\ D_1 f_2(x) & D_2 f_2(x) & D_3 f_2(x) \\ D_1 f_3(x) & D_2 f_3(x) & D_3 f_3(x) \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \end{bmatrix} \tag{5.1}$$

means the same thing as

$$\begin{bmatrix} D_1 f_1(x)\Delta x_1 + D_2 f_1(x)\Delta x_2 + D_3 f_1(x)\Delta x_3 \\ D_1 f_2(x)\Delta x_1 + D_2 f_2(x)\Delta x_2 + D_3 f_2(x)\Delta x_3 \\ D_1 f_3(x)\Delta x_1 + D_2 f_3(x)\Delta x_2 + D_3 f_3(x)\Delta x_3 \end{bmatrix}. \tag{5.2}$$

A 3×3 array of numbers such as the one on the left in (5.1) is called a **matrix**. With this new "matrix notation", Newton's

approximation for our function $f : \mathbb{R}^3 \to \mathbb{R}^3$ can now be written as

$$\underbrace{\begin{bmatrix} f_1(x+\Delta x) \\ f_2(x+\Delta x) \\ f_3(x+\Delta x) \end{bmatrix}}_{f(x+\Delta x)} \approx \underbrace{\begin{bmatrix} f_1(x) \\ f_2(x) \\ f_3(x) \end{bmatrix}}_{f(x)} + \underbrace{\begin{bmatrix} D_1 f_1(x) & D_2 f_1(x) & D_3 f_1(x) \\ D_1 f_2(x) & D_2 f_2(x) & D_3 f_2(x) \\ D_1 f_3(x) & D_2 f_3(x) & D_3 f_3(x) \end{bmatrix}}_{\text{matrix}} \underbrace{\begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \end{bmatrix}}_{\Delta x}$$

Comparing this with our familiar way of writing Newton's approximation,

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x,$$

suggests that the matrix above should in fact be denoted as $f'(x)$ and should be called the *derivative* of f at x. In summary, when $f : \mathbb{R}^3 \to \mathbb{R}^3$, we define $f'(x)$ to be the 3×3 matrix:

$$f'(x) = \begin{bmatrix} D_1 f_1(x) & D_2 f_1(x) & D_3 f_1(x) \\ D_1 f_2(x) & D_2 f_2(x) & D_3 f_2(x) \\ D_1 f_3(x) & D_2 f_3(x) & D_3 f_3(x) \end{bmatrix}.$$

Similarly, when $f : \mathbb{R}^n \to \mathbb{R}^m$, we define $f'(x)$ to be the following rectangular array of numbers:

$$f'(x) = \begin{bmatrix} D_1 f_1(x) & \cdots & D_n f_1(x) \\ \vdots & \ddots & \vdots \\ D_1 f_m(x) & \cdots & D_n f_m(x) \end{bmatrix}.$$

A rectangular array of numbers with m rows and n columns is called an $m \times n$ **matrix**. The immediate triumph of our new matrix notation is that Newton's approximation for a function $f : \mathbb{R}^n \to \mathbb{R}^m$ looks identical to Newton's approximation for a function $f : \mathbb{R} \to \mathbb{R}$:

$$\underbrace{f(\overset{\text{vector in } \mathbb{R}^n}{\downarrow} x + \overset{\text{vector in } \mathbb{R}^n}{\downarrow} \Delta x)}_{\uparrow \text{ vector in } \mathbb{R}^m} \approx \underbrace{f(x)}_{\uparrow \text{ vector in } \mathbb{R}^m} + \underbrace{f'(x)}_{\uparrow\, m \times n \text{ matrix}} \overset{\text{vector in } \mathbb{R}^n}{\downarrow} \Delta x \,. \tag{5.3}$$

We emphasize the following fact:

If $f : \mathbb{R}^n \to \mathbb{R}^m$, then $f'(x)$ is an $m \times n$ matrix.

This matrix $f'(x)$ is often called the "Jacobian" or the "Jacobian matrix" of f at x. However, I think a much better name for $f'(x)$ is simply "the derivative of f at x".

EXAMPLE 5.1. If $f : \mathbb{R}^n \to \mathbb{R}$, then $f'(x)$ is a $1 \times n$ matrix. In detail,

$$f'(x) = \begin{bmatrix} D_1 f(x) & D_2 f(x) & \cdots & D_n f(x) \end{bmatrix}.$$

A matrix with just one row is also called a "row vector".

EXAMPLE 5.2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be the function defined by

$$f(x) = \frac{1}{2}\|x\|^2 = \frac{1}{2}(x_1^2 + x_2^2 + \cdots + x_n^2).$$

(The numbers $x_1, \ldots, x_n$ are the components of the vector x.) Thinking of f as a function of x_1 alone, with $x_2, \ldots, x_n$ held fixed, we see that

$$D_1 f(x) = x_1.$$

Likewise, the ith partial derivative of f is $D_i f(x) = x_i$ (for $i = 1, \ldots, n$). Thus,

$$f'(x) = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}.$$

CHAPTER 6

Matrix multiplication

6.1. A matrix wants to operate on a vector

Our "matrix notation" has already saved writing, but matrices come to life if we take the viewpoint that the matrix in (5.1) is *performing an operation* on the vector Δx, resulting in the new vector (5.2). In this viewpoint, a matrix is not an inert object. It has a mission in life: to operate on a vector. We shall call this operation "multiplication" (reusing an old word), and we shall say that in the expression (5.1) we are "multiplying" the matrix on the left by the vector Δx.

An $m \times n$ matrix is, by definition, a rectangular array of numbers with m rows and n columns. The set of all possible $m \times n$ matrices is denoted $\mathbb{R}^{m \times n}$. If a matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

is "multiplied" by a vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$$

then the result is the vector $Ax \in \mathbb{R}^m$ defined by

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix}}_{Ax}.$$

This is the definition of multiplying a matrix by a vector, an operation that we invented when we declared that expression (5.1) means the same thing as expression (5.2). Notice the following pattern:

> The ith entry of the vector Ax is the dot product of the ith row of A with the vector x.

Throughout math, calculations which can be expressed as multiplying a matrix by a vector are ubiquitous. This is true in part because Newton's approximation is so ubiquitious.

6.2. Some useful rules of arithmetic

It's not hard to check that the following basic arithmetic rules are satisfied.

6.2.1. Distributive rule. If x and y are vectors in $\mathbb{R}^n$ then

$$A(x + y) = Ax + Ay. \tag{6.1}$$

This rule is called the "distributive" property of matrix-vector multiplication.

We have a similar distributive rule for multiplying A by a sum of any finite number of vectors. For example, if $x, y, z \in \mathbb{R}^n$ then

$$A(x + y + z) = Ax + Ay + Az.$$

(Proof: Just use rule (6.1) twice, to obtain $A(x + y + z) = A(x + y) + Az = Ax + Ay + Az$.)

6.2.2. Multiplying by a scalar. If we define the product of a scalar c with a matrix A in the obvious way, so that cA is the matrix obtained by multiplying each entry of A by c, then we have

$$A(cx) = c(Ax) = (cA)x.$$

It follows from this rule that the expression cAx is unambiguous, because either interpretation $(cA)x$ or $c(Ax)$ yields the same result.

EXERCISE 6.1. Suppose that x_1, x_2, x_3 are vectors in $\mathbb{R}^n$ and c_1, c_2, and c_3 are scalars (that is, real numbers). Explain why

$$A(c_1x_1 + c_2x_2 + c_3x_3) = c_1Ax_1 + c_2Ax_2 + c_3Ax_3.$$

6.3. Another perspective on matrix-vector multiplication

There is a different, more visual way to think about multiplying a matrix by a vector that turns out to be very useful. Notice that

$$\begin{aligned} Ax &= \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix} \\ &= x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix}. \end{aligned} \tag{6.2}$$

Thus,

Ax is a linear combination of the columns of A.

This is, in fact, my *favorite* way to think about multiplying a matrix by a vector. I find that it is usually the most illuminating viewpoint. I will call this the "visual interpretation" of matrix-vector multiplication.

We can state equation (6.2) more concisely if we let a_j denote the jth column of A, i.e.,

$$a_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix} \qquad \text{for } j = 1, \ldots, n.$$

Then we can write A using "block notation" as

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}$$

and equation (6.2) becomes

$$Ax = x_1 a_1 + x_2 a_2 + \cdots + x_n a_n. \tag{6.3}$$

EXERCISE 6.2. Suppose that M_1 and M_2 are $m \times n$ matrices and $M_1 x = M_2 x$ for all vectors $x \in \mathbb{R}^n$. Show that $M_1 = M_2$.

Solution: Take $x = (1, 0, \ldots, 0)$. Then from equation (6.3) we see that $M_1 x$ is equal to the first column of M_1 and $M_2 x$ is the first column of M_2. So the first column of M_1 is equal to the first column of M_2. A similar argument shows that the second column of M_1 is equal to the second column of M_2, and so on. Thus, $M_1 = M_2$.

6.4. Multiplying a matrix by a matrix

Now suppose that $B \in \mathbb{R}^{k \times m}$. If $x \in \mathbb{R}^n$, then

$$\begin{aligned} B(Ax) &= B(x_1 a_1 + x_2 a_2 + \cdots + x_n a_n) \\ &= x_1 Ba_1 + x_2 Ba_2 + \cdots + x_n Ba_n \qquad (6.4) \\ &= Mx \qquad (6.5) \end{aligned}$$

where M is the matrix whose jth column is Ba_j. (In going from (6.4) to (6.5), we have used the "visual interpretation" of matrix-vector multiplication.)

In other words, multiplying first by A and then by B yields the same result as simply multiplying by M:

$$B(Ax) = Mx$$

for all $x \in \mathbb{R}^n$. This matrix M is called the "product" of B and A and is denoted as BA. We again reuse the word "multiply", and say that M is the result of "multiplying" B by A. Using block notation, the definition of BA is

$$BA = \begin{bmatrix} Ba_1 & Ba_2 & \cdots & Ba_n \end{bmatrix}.$$

We can only multiply a matrix B by a matrix A if their shapes are compatible, meaning that the number of columns of B is equal to the number of rows of A. Otherwise, BA is not defined.

6.5. When multiplying matrices, order matters

Warning: The order in which we multiply matrices matters. It is usually not true that $BA = AB$. In fact, the product AB might not even be defined (even if BA is defined).

However, it is always true that

$$C(BA) = (CB)A,$$

provided that A, B, and C are matrices with compatible shapes. To see this, just check that multiplying a vector x by the matrix on the left always yields the same result as multiplying x by the matrix on the right. The result of multiplying x by the matrix $C(BA)$ is

$$\begin{aligned} (C(BA))\,x &= C\,((BA)x) \\ &= C\,(B(Ax))\,. \end{aligned}$$

Meanwhile, the result of multiplying x by the matrix $(CB)A$ is

$$\begin{aligned} ((CB)A)\,x &= (CB)(Ax) \\ &= C(B(Ax)). \end{aligned}$$

So we get the same result either way.

6.6. Conventions about column vectors and row vectors

A "column vector" is a matrix with just one column. From now on, we shall declare that the elements of $\mathbb{R}^n$ are column vectors. With this convention, matrix-vector multiplication is a special case of matrix-matrix multiplication. If $x \in \mathbb{R}^n$, then x is an $n \times 1$ matrix (that is, column vector), and when we compute Ax we are multiplying an $m \times n$ matrix A by an $n \times 1$ matrix x.

A "row vector" is a matrix with just one row. Notice that if u and v are vectors in $\mathbb{R}^n$ (so u and v are column vectors), we can compute their dot product by first flipping u sideways, turning it into a row vector, and then multiplying the resulting row vector by v:

$$\begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = u_1v_1 + u_2v_2 + \cdots + u_nv_n = \langle u, v \rangle.$$

On the left we are multiplying a $1 \times n$ matrix by an $n \times 1$ matrix.

The row vector obtained by flipping u sideways is called the "transpose" of u and is denoted u^T. With this notation, the above equation tells us that

$$\langle u, v \rangle = u^T v. \tag{6.6}$$

Notice that $u^T v = v^T u$, because $\langle u, v \rangle = \langle v, u \rangle$.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathbb{R}^n$. Because we have declared that the elements of $\mathbb{R}^n$ are column vectors, the row vector $f'(x) = \begin{bmatrix} D_1 f(x) & \cdots & D_n f(x) \end{bmatrix}$ is not an element of $\mathbb{R}^n$. We shall adopt the convention the gradient of f at x is a column vector:

$$\nabla f(x) = \begin{bmatrix} D_1 f(x) \\ D_2 f(x) \\ \vdots \\ D_n f(x) \end{bmatrix} = f'(x)^T.$$

With this convention, Newton's approximation $f(x + \Delta x) \approx f(x) + f'(x)\Delta x$ can be written equivalently as

$$f(x + \Delta x) \approx f(x) + \nabla f(x)^T \Delta x.$$

Defining the gradient to be a column vector is convenient because it will allow us to make statements such as the following: if we are located at x and we move a short distance in the direction of steepest descent for f, then our new location is $x - t\nabla f(x)$ (for some scalar t). Repeatedly moving in the direction of steepest descent is a good strategy for finding a point x^* at which f has a minimum value. This strategy is called "gradient descent" and is fundamental in machine learning.

EXERCISE 6.3. Suppose that $u, v \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Explain why

$$(u + v)^T = u^T + v^T \quad \text{and} \quad (cu)^T = cu^T.$$

Solution: What difference does it make if we flip then add or add then flip? Likewise, what difference does it make if we flip then scale by c or if we scale by c and then flip? We get the same result either way.

EXERCISE 6.4. Suppose that $u_1, u_2, u_3 \in \mathbb{R}^n$ and $c_1, c_2, c_3 \in \mathbb{R}$. Explain why

$$(c_1u_1 + c_2u_2 + c_3u_3)^T = c_1u_1^T + c_2u_2^T + c_3u_3^T.$$

Solution: As with the previous exercise, this fact might seem obvious, because what difference does it make if our column vectors are tipped sideways before or after we combine them? Another way to think about it is to use the results of the previous exercise repeatedly:

$$\begin{aligned}(c_1u_1 + c_2u_2 + c_3u_3)^T &= (c_1u_1 + c_2u_2)^T + (c_3u_3)^T \\ &= (c_1u_1)^T + (c_2u_2)^T + (c_3u_3)^T \\ &= c_1u_1^T + c_2u_2^T + c_3u_3^T.\end{aligned}$$

6.7. Transposing matrices

In the previous section we transposed a column vector to obtain a row vector. Now we will ask what is the transpose of Ax, where A is an $m \times n$ matrix A and $x \in \mathbb{R}^n$. This question leads us to discover the "transpose" of a matrix A.

6.7.1. Visual interpretation of $z^T A$. The "visual interpretation" of matrix-vector multiplication tells us that Ax is a linear combination of the columns of A. (See section 6.3.) If $z \in \mathbb{R}^m$, there is a similar "visual interpretation" of the product $z^T A$:

$$z^T A \text{ is a linear combination of the rows of } A. \tag{6.7}$$

To see this concretely, let's look at the special case where $m = 3$ and $n = 2$. Then z and A can be written in detail as

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_2 \end{bmatrix}, \quad A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$$

and

$$\begin{aligned} z^T A &= \begin{bmatrix} z_1 & z_2 & z_3 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \\ &= \begin{bmatrix} z_1 a_{11} + z_2 a_{21} + z_3 a_{31}, & z_1 a_{12} + z_2 a_{22} + z_3 a_{32} \end{bmatrix} \\ &= z_1 \underbrace{\begin{bmatrix} a_{11} & a_{12} \end{bmatrix}}_{\substack{\text{first row} \\ \text{of } A}} + z_2 \underbrace{\begin{bmatrix} a_{21} & a_{22} \end{bmatrix}}_{\substack{\text{second row} \\ \text{of } A}} + z_3 \underbrace{\begin{bmatrix} a_{31} & a_{32} \end{bmatrix}}_{\substack{\text{third row} \\ \text{of } A}}. \end{aligned}$$

6.7.2. The transpose of a matrix. Sometimes we might need to compute the transpose of the column vector Ax. Let a_j be the jth column of an $m \times n$ matrix A. Then

$$\begin{aligned} (Ax)^T &= (x_1 a_1 + x_2 a_2 + \cdots + x_n a_n)^T \\ &= x_1 a_1^T + x_2 a_2^T + \cdots + x_n a_n^T. \end{aligned}$$

But here we have a linear combination of row vectors. According to the "visual interpretation" above, this linear combination of

row vectors is equal to $x^T M$, where M is the matrix whose rows are $a_1^T, a_2^T, \ldots, a_n^T$. This matrix M is called the "transpose" of A, and is denoted A^T. With this notation, we have

$$(Ax)^T = x^T A^T.$$

Notice that the first column of A is the first row of A^T, the second column of A is the second row of A^T, and so on. For example,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}.$$

EXERCISE 6.5. Suppose that A is an $m \times n$ matrix, and that $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Show that

$$\langle Ax, y \rangle = \langle x, A^T y \rangle.$$

(As you go deeper in math, this turns out to be in some sense the most essential fact about the transpose of a matrix. When I think about A^T, I usually think about this equation.)

Solution: Using equation (6.6), we have

$$\begin{aligned} \langle Ax, y \rangle &= (Ax)^T y \\ &= (x^T A^T) y \\ &= x^T (A^T y) \\ &= \langle x, A^T y \rangle. \end{aligned}$$

EXERCISE 6.6. Suppose that A is an $m \times n$ matrix and B is a $k \times m$ matrix. Show that

$$(BA)^T = A^T B^T.$$

Solution: If $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^k$, then

$$\begin{aligned}\langle (BA)x, y\rangle &= \langle B(Ax), y\rangle \\ &= \langle Ax, B^T y\rangle \\ &= \langle x, A^T B^T y\rangle.\end{aligned}$$

On the other hand,

$$\langle (BA)x, y\rangle = \langle x, (BA)^T y\rangle.$$

So we have discovered that

$$\langle x, (BA)^T y\rangle = \langle x, (A^T B^T)y\rangle \tag{6.8}$$

for all $x \in \mathbb{R}^n, y \in \mathbb{R}^k$. This suggests that $(BA)^T = A^T B^T$.

To finish off the argument, take

$$y = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^k \quad \text{and} \quad x = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^n.$$

Then $(BA)^T y$ is the first column of $(BA)^T$, and $\langle x, (BA)^T y\rangle$ is the first entry of the first column of $(BA)^T$. Likewise, $\langle x, (A^T B^T)y\rangle$ is the upper left entry of $A^T B^T$. Thus, $(BA)^T$ and $A^T B^T$ have the same upper left entry. A similar argument shows that all of the corresponding entries of $(BA)^T$ and $A^T B^T$ are equal.

6.8. Matrix addition

Having discussed matrix multiplication, we should also mention the simpler operation of matrix addition. Suppose A and B are $m \times n$ matrices. We define $A + B$ to be the $m \times n$ matrix obtained by adding together the corresponding entries of A and B. For example, if $m = n = 2$ then

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}}_{A} + \underbrace{\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}}_{B} = \underbrace{\begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{bmatrix}}_{A+B}.$$

Note that we can only add together matrices which have the same shape. We could not, for example, add a row vector to a column vector.

EXERCISE 6.7. Suppose that $x \in \mathbb{R}^n$. Explain why

$$(A + B)x = Ax + Bx.$$

Solution: It's probably more clear to check this for yourself than to read my explanation. Nevertheless, let a_i^T and b_i^T be the ith rows of A and B, respectively. Then the ith row of $A + B$ is $a_i^T + b_i^T$, and the ith entry of $(A+B)x$ is $(a_i^T + b_i^T)x = a_i^T x + b_i^T x$. But this is the sum of the ith entries of Ax and Bx.

EXERCISE 6.8. Suppose C is an $n \times k$ matrix. Explain why $(A + B)C = AB + AC$.

Solution: Let c_i be the ith column of C. The ith column of $(A + B)C$ is $(A + B)c_i = Ac_i + Bc_i$. But this is the same as the ith column of $AC + BC$.

EXERCISE 6.9. Suppose C is a $k \times m$ matrix. Explain why $C(A + B) = CA + CB$.

Solution: Let a_j and b_j be the jth columns of A and B, respectively. The jth column of $C(A + B)$ is $C(a_j + b_j) = Ca_j + Cb_j$. But this is the same as the jth column of $CA + CB$.

6.9. Additional exercises

EXERCISE 6.10. The arithmetic rules given in section 6.2 can be summarized as stating that if $A \in \mathbb{R}^{n\times m}$, then the function $L : \mathbb{R}^n \to \mathbb{R}^m$ defined by $L(x) = Ax$ is a "linear transformation", which means that

(1) $L(x + y) = L(x) + L(y)$ for all vectors $x, y \in \mathbb{R}^n$.
(2) $L(cx) = cL(x)$ for all scalars $c \in \mathbb{R}$ and vectors $x \in \mathbb{R}^n$.

Suppose that $L : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation. Show that there exists a matrix A such that $L(x) = Ax$ for all $x \in \mathbb{R}^n$.

Solution: Let e_j be the jth standard basis vector for $\mathbb{R}^n$, so that e_j has a 1 in the jth position and zeros elsewhere. If $x \in \mathbb{R}^n$, then

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{can be written as} \quad x = x_1e_1 + x_2e_2 + \cdots + x_ne_n.$$

It follows that

$$\begin{aligned} L(x) &= L(x_1e_1 + x_2e_2 + \cdots + x_ne_n) \\ &= x_1L(e_1) + x_2L(e_2) + \cdots + x_nL(e_n) \\ &= Ax, \end{aligned}$$

where A is the matrix whose jth column is $L(e_j)$.

CHAPTER 7

The chain rule

We will now discover one of the most useful rules for computing derivatives in multivariable calculus. Suppose that $h : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^k$, and suppose that

$$f(x) = g(h(x))$$

for all $x \in \mathbb{R}^n$. Notice that f takes as input a point in $\mathbb{R}^n$ and returns as output a point in $\mathbb{R}^k$. Let Δx be a small vector in $\mathbb{R}^n$. We will approximate $f(x+\Delta x)$ by using Newton's approximation twice, first with h and then with g:

$$\begin{aligned} f(x + \Delta x) &= g(h(x + \Delta x)) \\ &\approx g(h(x) + h'(x)\Delta x) \\ &\approx g(h(x)) + g'(h(x))h'(x)\Delta x. \end{aligned}$$

In the final step, we used the approximation $g(z + \Delta z) \approx g(z) + g'(z)\Delta z$ with $z = h(x)$ and $\Delta z = h'(x)\Delta x$. Comparing the approximation

$$f(x + \Delta x) \approx \underbrace{g(h(x))}_{f(x)} + g'(h(x))h'(x)\Delta x$$

with Newton's approximation

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x$$

reveals (or at least suggests) that

$$f'(x) = g'(h(x))h'(x). \tag{7.1}$$

This fact is known as the chain rule.

It is a triumph of matrix notation that the multivariable chain rule can be discovered so easily and that it looks identical to the chain rule in single-variable calculus. Notice that in equation (7.1) $f'(x)$ is a matrix, and on the right we are multiplying two matrices:

$$\underbrace{f'(x)}_{\substack{k\times n\\ \text{matrix}}} = \underbrace{g'(h(x))}_{\substack{k\times m\\ \text{matrix}}}\underbrace{h'(x)}_{\substack{m\times n\\ \text{matrix}}}.$$

EXERCISE 7.1. Suppose that a function $f : \mathbb{R}^n \to \mathbb{R}$ tells us the temperature at each point in $\mathbb{R}^n$. A mosquito is located at a point x at time $t = 0$ and is moving with constant velocity vector $v \in \mathbb{R}^n$. So the mosquito's position at time t is $x + tv$, and the mosquito's temperature at time t is

$$F(t) = f(x + tv).$$

a) How rapidly is the mosquito's temperature changing at time $t = 0$?

b) Suppose the temperature is as cold as possible at the point x (so x is a minimizer for the function f). What can we conclude about the value of $\nabla f(x)$?

Solution: Notice that $F(t) = f(h(t))$ where $h : \mathbb{R} \to \mathbb{R}^n$ is the function defined by

$$h(t) = x + tv.$$

It is straightforward to check that $h'(t) = v$. By the chain rule,

$$F'(t) = f'(x + tv)h'(t) = \underbrace{f'(x + tv)}_{1\times n}\overset{\overset{n\times 1}{\downarrow}}{v}$$

so the rate of change of the mosquito's temperature at time $t = 0$ is

$$F'(0) = \underbrace{f'(x)}_{1\times n}\overset{\overset{n\times 1}{\downarrow}}{v} = \langle \nabla f(x), v\rangle.$$

This formula $\langle \nabla f(x), v \rangle$ is the same directional derivative formula that we derived in section 4.5.

If x is a minimizer for f, then the function F has a minimum value at $t = 0$. So, from single-variable calculus, we know that $F'(0) = \langle \nabla f(x), v \rangle = 0$. But notice that v could be any vector in $\mathbb{R}^n$. The only way that $\langle \nabla f(x), v \rangle$ can be 0 for every vector $v \in \mathbb{R}^n$ is if $\nabla f(x) = 0$. So we can conclude that $\nabla f(x) = 0$.

CHAPTER 8

Minimizing a function

Suppose again that a function $f : \mathbb{R}^n \to \mathbb{R}$ tells us the temperature at each point $x \in \mathbb{R}^n$. (I am visualizing the case where $n = 3$.) Imagine that a mosquito who likes cool weather has found a point x^* in the shade where the temperature is as low as possible. This point x^* is a minimizer for the function f. If the mosquito were to move a slight bit in the direction of any given unit vector u, the value of f could not decrease. Thus, $D_u f(x^*) \geq 0$ for every unit vector $u \in \mathbb{R}^n$. Moreover, if $u \in \mathbb{R}^n$, it is impossible that $D_u f(x^*) > 0$, because then the directional derivative of f in the opposite direction $-u$ would be negative. Therefore,

$$D_u f(x^*) = \langle \nabla f(x^*), u \rangle = 0 \quad \text{for all } u \in \mathbb{R}^n.$$

This is only possible if $\nabla f(x^*) = 0$. So, in conclusion:

> If $x^* \in \mathbb{R}^n$ is a minimizer for a function $f : \mathbb{R}^n \to \mathbb{R}$ then $\nabla f(x^*) = 0$.

See exercise (7.1) for a slightly different derivation of this fact. This conclusion holds regardless of whether x^* is a "global" minimizer for f (which means that $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$) or just a "local" minimizer for f (which means that $f(x) \geq f(x^*)$ for all x in the near vicinity of x^*).

This suggests the following strategy for finding a minimizer of f: we compute the gradient of f, then set the gradient of f equal to 0 and solve the resulting system of equations for x. However, we must be careful, because not every point x which

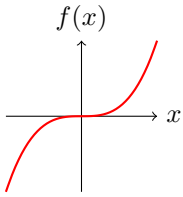


Figure 15. If $f(x) = x^3$, then $f'(0) = 0$, but 0 is neither a minimizer nor a maximizer for f.

satisfies $\nabla f(x) = 0$ is a minimizer for f. Indeed, a similar argument to the one given above shows that if x is a maximizer for f then $\nabla f(x) = 0$. It is also possible for a point x which satisfies $\nabla f(x) = 0$ to be neither a maximizer nor a minimizer for f. For example, if $n = 1$ and $f(x) = x^3$, then $f'(0) = 0$ but 0 is neither a minimizer nor a maximizer for f. This is illustrated in figure 15.

Another example is provided by the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x_1, x_2) = x_1 x_2$. If x_1 and x_2 are both positive, then $f(x_1, x_2) > 0$. On the other hand, if x_1 is positive and x_2 is negative, then $f(x_1, x_2) < 0$. Thus, any ball centered at the origin contains points where f is positive and also points where f is negative. So the origin is neither a minimizer nor a maximizer for f.

Nevertheless, i

APPENDIX A

Algebra review

Calculus is much easier to learn if computations using high school algebra are effortless. (If not, that's ok, calculus provides a good chance to practice algebra and you can always backtrack to fill in any gaps in knowledge.) We review a few formulas from algebra below.

A.1. FOIL

The FOIL ("first-outer-inner-last") formula states that if x, y, z and w are numbers then

$$(x+y)(z+w) = xz + xw + yz + yw.$$

This rule can be derived by repeated use of the distributive rule

$$a(b+c) = ab + ac.$$

In detail, we derive FOIL as follows:

$$\begin{aligned}(x+y)(z+w) &= (x+y)z + (x+y)w \\ &= xz + yz + xw + yw.\end{aligned}$$

We can use FOIL to expand expressions such as $(x+y)^2$. If x and y are numbers, then

$$\begin{aligned}(x+y)^2 &= (x+y)(x+y) \\ &= x^2 + xy + yx + y^2 \\ &= x^2 + 2xy + y^2.\end{aligned}$$

We can also expand $(x-y)^2$:

$$\begin{aligned}(x-y)^2 &= (x-y)(x-y)\\ &= x^2 - xy - yx + y^2\\ &= x^2 - 2xy + y^2.\end{aligned}$$

Let's expand $(x+y)^3$:

$$\begin{aligned}(x+y)^3 &= (x+y)(x+y)^2\\ &= (x+y)(x^2+2xy+y^2)\\ &= x(x^2+2xy+y^2) + y(x^2+2xy+y^2)\\ &= x^3 + 2x^2y + xy^2 + yx^2 + 2xy^2 + y^3\\ &= x^3 + 3x^2y + 3xy^2 + y^3.\end{aligned}$$

A.2. Difference of squares

The "difference of squares" formula states that if x and y are numbers then

$$x^2 - y^2 = (x-y)(x+y).$$

This rule can be proved by applying FOIL to simply the expression on the right:

$$\begin{aligned}(x-y)(x+y) &= x^2 + xy - xy - y^2\\ &= x^2 - y^2.\end{aligned}$$

APPENDIX B

The equation of a line

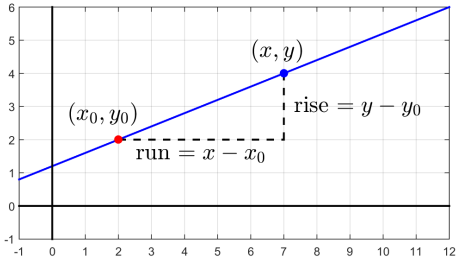


Figure 16. Here is a line with slope m that passes through the point (x_0, y_0). If (x, y) is a point on this line, then $m = \frac{\text{rise}}{\text{run}} = \frac{y-y_0}{x-x_0}$. In this example, $(x_0, y_0) = (2, 2)$ and $m = 2/5$.