Hal Daumé III *28 August 2009*

Math for Machine Learning

The goal of this document is to provide a "refresher" on continuous mathematics for computer science students. It is by no means a rigorous course on these topics. The presentation, motivation, etc., are all from a machine learning perspective. The hope, however, is that it's useful in other contexts. The two major topics covered are linear algebra and calculus (probability is currently left off)).

1 Calculus

Calculus is classically the study of the relationship between variables and their rates of change. However, this is *not* what we use calculus for. We use *differential calculus* as a method for finding extrema of functions; we use *integral calculus* as a method for probabilistic modeling.

1.1 Differential Calculus

linear regression.

Example 1. *To be more concrete, a classical statistics problem is* **linear regression.** *Suppose that I have a bunch of points* $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$*, and I want to fit a line of the form* $y = mx + b$*. If I have a lot of points, it's pretty unlikely that there is going to be a line that actually passes exactly through all of them. So we can ask instead for a line* $y = mx + b$ *that lies as close to the points as possible. See Figure* **??**.

One easy option is to use squared error *as a measure of closeness. For a point* (x_n, y_n) *and a line defined by* m *and* b*, we can measure the squared error as* $[(mx_n + b) - y_n]^2$*. That is: our predicted value minus the true value, all squared.*[1] *We can easily* sum *all of the point-wise errors to get a total error (which, for some strange reason, we'll call "J") of:*

$$J(m, b) = \sum_{n=1}^{N} \left[(mx_n + b) - y_n\right]^2 \tag{1}$$

Note that we have written the error J *as a function of* m *and* b*, since, for any setting of* m *and* b*, we will get a different error.*

Now, our goal is to find values of m *and* b *that* minimize *the error. How can we do this? Differential calculus tells us that the minimum of the* J *function can be computed by finding the zeros of its derivatives. (Assuming it is* convex*: see Section 1.3.)*

derivative
velocity

The **derivative** of a function at a point is the *slope* of the function at that point (a derivative is like a **velocity**). See Figure **??**. To be precise, suppose we have a function f that maps real numbers to real numbers. (That is: $f : \mathbb{R} \to \mathbb{R}$; see Section **??**). For instance, $f(x) = 3x^2 - e^x$. The derivative of f with respect to x, denoted $\partial f / \partial x$, is[2]:

$$\frac{\partial f}{\partial x}(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{2}$$

This essentially says that the derivative of f with respect to x, evaluated at a point x_0, is the rate of change of f at x_0. It is fairly common to see $\partial f / \partial x$ denoted by f'. The disadvantage to this notation is that when f is a function of multiple variables (such as J in linear regression; see Example 1), then f' is ambiguous as

[1] As discussed in the notation (see Section **??**), we count data points by $1 \ldots N$ and index them by n. In general, things that range from $1 \ldots X$ will be indexed by x.

[2] Note that there are other definitions that catch interesting corner cases.

to which variable the derivative is being taken with respect to. Nevertheless, when clear from context, we will also use f'.

Also regarding notation, if we want to talk about the derivative of a function without naming the function, we will write something like:

$$\frac{\partial}{\partial x}\left[3x^2 - e^x\right] \tag{3}$$

Or, if we're really trying to save space, will write ∂_x for the derivative with respect to x, yielding: $\partial_x[3x^2 - e^x]$.

In case you are a bit rusty taking derivatives by hand, the important rules are given below:

- Scalar multiplication: $\partial_x[af(x)] = a[\partial_x f(x)]$
- Polynomials: $\partial_x[x^k] = kx^{k-1}$
- Function addition: $\partial_x[f(x) + g(x)] = [\partial_x f(x)] + [\partial_x g(x)]$
- Function multiplication: $\partial_x[f(x)g(x)] = f(x)[\partial_x g(x)] + [\partial_x f(x)]g(x)$
- Function division: $\partial_x\left[\frac{f(x)}{g(x)}\right] = \frac{[\partial_x f(x)]g(x) - f(x)[\partial_x g(x)]}{[g(x)]^2}$
- Function composition: $\partial_x[f(g(x))] = [\partial_x g(x)][\partial_x f](g(x))$
- Exponentiation: $\partial_x[e^x] = e^x \quad$ and $\quad \partial_x[a^x] = \log(a)e^x$
- Logarithms: $\partial_x[\log x] = \frac{1}{x}$

Note that throughout this document, log means **natural log** – that is, logarithm base e. You may have seen this previously as ln, but we do not use this notation. If we intend a log base *other than* e, we will write, eg., $\log_{10} x$, which can be converted into natural log as $\log x / \log 10$.

natural log

Exercise 1. *Compute derivatives of the following functions:*

1. $f(x) = e^{x+1}$
2. $f(x) = e^{-\frac{1}{2}x^2}$
3. $f(x) = x^a x^{1-a}$
4. $f(x) = (e^x + x^2 + 1/x)^3$
5. $f(x) = \log(x^2 + x - 1)$
6. $f(x) = \frac{e^x+1}{e^{-x}}$

Example 2. *Returning to Example 1, we have a function $J(m, b)$ and we want to compute its derivative with respect to m and its derivative with respect to b. Working through the case for m, we have:*

$$\partial_m J(m,b) = \partial_m \left(\sum_{n=1}^{N} \left[(mx_n + b) - y_n\right]^2 \right) \tag{4}$$

$$= \sum_{n=1}^{N} \partial_m \left[(mx_n + b) - y_n\right]^2 \tag{5}$$

$$= \sum_{n=1}^{N} \left[2\left[(mx_n + b) - y_n\right]\right] \partial_m \left[(mx_n + b) - y_n\right] \tag{6}$$

$$= \sum_{n=1}^{N} \left[2\left[(mx_n + b) - y_n\right]\right] x_n \tag{7}$$

In the first step, we apply the function addition rule; in the second step, we apply the composition rule; in the third step, we apply the polynomial rule.

Exercise 2. *Compute $\partial_b J(m, b)$.*

One nice thing about derivatives is that they allow us to find **extreme points** of functions in a straightforward way. (Usually you can think of an exteme point as a **maximum** or **minimum** of a function.) Consider again Figure **??**; here, we can easily see that the point at which the function is minimized has a derivative (slope) of zero. Thus, we we can find zeros of the derivative of a function, we can also find minima (or maxima) of that function.

extreme points
maximum
minimum

Example 3. *The example plotted in Figure* **??** *is of the function $f(x) = 2x^2 - 3x + 1$. We can compute the derivative of this function as $\partial_x f(x) = 4x - 3$. We equate this to zero ($4x - 3 = 0$) and apply algebra to solve for x, yielding $x = 3/4$. As we can see from the plot, this is indeed a minimum of this function.*

Exercise 3. *Using $\partial_m J$ and $\partial_b J$ from previous examples and exercises, compute the values of m and b that minimize the function J, thus* solving *the linear regression problem!*

1.2 Integral Calculus

An integral is the "opposite" of a derivative. Its most common use, at least by us, is in computing areas under a curve. We will never actually have to compute integrals by hand, though you should be familiar with their properties.

The "area computing" integral typically has two bounds, a (the lower bound) and b (the upper bound). We will write them as $\int_a^b \mathrm{d}x f(x)$ to mean the area under the curve given by the function f between a and b.[3] You should think of these integrals as being the *continuous analogues* of simple sums. That is, you can "kind of" read such an integral as $\sum_{x=a}^{b} f(x)$.

The interpretation of an integral as a sum comes from the following thought experiment. Suppose we were to discretize the range $[a, b]$ into R many units of width $(a - b)/R$. Then, we could *approximate* the area under the curve by a sum over these units, evaluating $f(x)$ at each position (to get the height of a rectangle there) and multiplying by $(a - b)/R$, which is the width. Summing these rectangles (see Figure **??**) will approximate the area. As we let $R \to \infty$, we'll get a better and better approximation. However, as $R \to \infty$, the width of each rectangle will approach 0. We name this width "$\mathrm{d}x$," and thus the integral notation mimics almost exactly the "rectangular sum" notation (we have width of $\mathrm{d}x$ times height of $f(x)$, summed over the range).

An common integral is that over an unbounded range, for instance $\int_{-\infty}^{\infty} \mathrm{d}x f(x)$. While it may seem crazy to try to sum up things over an infinite range, there are actually many functions f for which the result of this integration is *finite*. For instance, a half-bounded integral of $1/x^2$ *is* finite:

$$\int_1^\infty \mathrm{d}x \frac{1}{x^2} = \lim_{b\to\infty} \int_1^b \mathrm{d}x \frac{1}{x^2} = \lim_{b\to\infty} \left[-\frac{1}{b} - (-\frac{1}{1}) \right] = 0 + 1 = 1 \tag{8}$$

A similar calculation can show the following (called Gauss' integral):

$$\int_{-\infty}^{\infty} \mathrm{d}x e^{-x^2} = \sqrt{\pi} \tag{9}$$

1.3 Convexity

The notion of a **convex function** and a **convex set** will turn out to be *incredibly* important in our studies.

convex function
convex set

[3] You may be more used to the notation $\int_a^b f(x)\mathrm{d}x$ – the reason for putting the d on the left is so that your brain doesn't have to "find it" when $f(x)$ is some long expression.

A convex function is, in many ways, "well behaved." Although not a precise definition, you can think of a convex function as one that has a single point at which the derivative goes to zero, and this point is a minimum. For instance, the function $f(x) = 2x^2 - 3x + 1$ from Figure **??** is convex. One usually thinks of context functions as functions that "hold water" – i.e., if you were to pour water into them, it wouldn't spill out.

The opposite of a convex function is a **concave function**. A function f is concave if the function $-f$ is convex. So convex functions look like valleys, concave functions like hills.

concave function

The reason we care about convexity is because it means that finding minima is *easy.* For instance, the fact that $f(x) = 2x^2 - 3x + 1$ is convex means that once we've found a point that has a zero derivative, we have found the *unique, global minimum.* For instance, consider the function $f(x) = x^4 + x^3 - 4x^2$, which is plotted in Figure **??**. This function is *non-convex.* It has *three* points at which the derivative goes to zero. The left-most corresponds to a *global minimum*, the middle to a *local maximum* and the right-most to a *local minimum.* What this means is that even if we are able to find a point x for which $\partial_x f(x) = 0$, it is *not* necessarily true that x is a minimum (or maximum) of f.

More formally, a function f is **convex** on the range $[a, b]$ if its *second derivative* is positive everywhere in that range. The second derivative is simply the derivative of the derivative (and is physically associated with **acceleration**). The second derivative of f with respect to x is typically denoted by one of the following:

convex

acceleration

$$\frac{\partial^2 f}{\partial x \partial x} \quad = \quad \frac{\partial^2 f}{\partial x^2} \quad = \quad \frac{\partial}{\partial x}\left[\frac{\partial f}{\partial x}\right] \quad = \quad \partial_x \partial_x f \tag{10}$$

A function f is **convex everywhere** if f is convex on the range $(-\infty, \infty)$.

convex everywhere

Example 4. *Consider the function $f(x) = 2x^2 - 3x + 1$. We've already computed the first derivative of this function: $\partial_x f(x) = 4x - 3$. To compute the second derivative of f, we just* re-differentiate *the derivative, yielding $\partial_x \partial_x f(x) = 4$. Clearly, the function that maps everything to 4 is positive everywhere, so we know that f is convex.*

Example 5. *Now, consider the non-convex function $f(x) = x^4 + x^3 - 4x^2$. The first derivative is $\partial_x f(x) = 4x^3 + 3x^2 - 4x$ and the second derivative is $12x^2 + 6x - 4$. It's fairly easy to find a value of x for which the second derivative is negative: 0 is such an example. It is moderately interesting to note that while this f is not convex everywhere, it* is *convex in certain ranges, for instance the open intervals $(-\infty, -1)$ and $(0.5, \infty)$ are ranges over which f* is *convex.*

Exercise 4. *Verify whether the functions from Exercise 1 are convex, concave or neither.*

An analogous notion to a convex function is a **convex set**. Consider some subset A of the real line. We'll denote the real line by $\mathbb{R}$, so we have $A \subset \mathbb{R}$. We say that A is **convex** whenever the following holds: for all $x, y \in A$ and $\lambda \in [0, 1]$, the point $\lambda x + (1 - \lambda)y$ is also in A. In more mathy terms, A is convex if it is *closed* under *convex combination.*

convex set

convex

The way to think of this is as follows. Given two points x and y on the plane, the function $f(\lambda) = \lambda x + (1-\lambda)y$ on the range $\lambda \in [0, 1]$ denotes the **line segment** that joins x and y. A set A is convex if all points on all such line segments are also contained in A.[4]

line segment

Example 6. *For example, the closed interval $[1, 3]$ is convex. To show this, let $x, y \in [1, 3]$ be given, and let $\lambda \in [0, 1]$ be given. Let $z = \lambda x + (1 - \lambda)y$. First, we show that $z \geq 1$. Without loss of generality, $x \leq y$, so $z = \lambda x + (1 - \lambda)y \geq \lambda x + (1 - \lambda)x = x \geq 1$. Next, we show that $z \leq 3$. Similarly, $z = \lambda x + (1 - \lambda)y \leq \lambda y + (1 - \lambda)y = y \leq 3$.*

In general, all open and closed intervals of the real line are convex.

[4] Except under strange conditions, it is sufficient to check that for all $x, y \in A$, the point $(x + y)/2$ is also in A. This is equivalent to just checking the case for $\lambda = 0.5$.

Exercise 5. *Show that* $[-3,-1] \cup [1,3]$ *(the union of the closed interval* $[-3,-1]$ *and the closed interval* $[1,3]$*) is* not *convex.*

Why do we care about convex sets? A lot of times we're going to be trying to minimize some function $f(x)$, but under a constraint that x lies in some set A. If A is convex, the life is much easier. This is because it means that if we have two solutions, both in A, we can try to find a better solution between them, and this is guaranteed to still be in A (by convexity). (We'll come back to this later in Section 3.1.

An immediate question is: convex sets and convex functions share the word "convex." This implies that they have something in common. They do, but we'll need to get to multidimensional analogues before we can see this (see Section 3.1.

1.4 Wrap-up

The important concepts from this section are:

- Differentiation as a tool to finding maxima/minima of a function.
- Integration as a tool for computing area under a function.
- Convex functions hold water.

If you feel comfortable with these issues and can solve most of the exercises, you're in good shape!

2 Linear Algebra

A large part of statistics and machine learning has to do with modeling *data.* Although not always the case, for many problems, it is useful to think of data points as being points in some high dimensional space. For instance, we might characterize a car by it's length, width, height and maximum velocity. A given car can then be realized by a point in 4-dimensional space, where the value in each dimension corresponds to one of the properties we are measuring. Linear algebra gives us a set of tools for describing and manipulating such objects.

2.1 Vector Spaces

Our presentation here is going to be focused on on particular *type* of vector space, namely D-dimensional **Euclidean space**; that is, the space $\mathbb{R}^D$. It's very important to realize, however, that the ideas in linear algebra are much more general. All of our examples will be from $D \in \{2,3\}$, since these are the only ones for which we have a chance of drawing examples.

Euclidean space

We begin by defining a **vector**. This is simply an element $\boldsymbol{x}$ that lives in our vector space. In two dimensions, vectors are simply points on the plane. The vector $\boldsymbol{x} \in \mathbb{R}^D$ has D-many components, denoted by $\langle x_1, x_2, \ldots, x_D \rangle$. We'll typically use small d to denote one particular dimension; thus, x_d is the (real value!) corresponding to the dth dimension of $\boldsymbol{x}$.

vector

There are several things we might want to do with vectors: add them, and multiply them by a scalar. These operations are defined by the following rules:

- For $\boldsymbol{x} \in \mathbb{R}^D$ and $a \in \mathbb{R}$, the **scalar product** of $\boldsymbol{x}$ and a, denoted $a\boldsymbol{x}$ given by the vector in $\mathbb{R}^D$ defined component-wise by $\langle ax_1, ax_2, \ldots, ax_D \rangle$.

scalar product

- For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$, the **vector sum** (or simply the sum) of $\boldsymbol{x}$ and $\boldsymbol{y}$, denoted $\boldsymbol{x} + \boldsymbol{y}$, is again a vector in $\mathbb{R}^D$ defined component-wise by $\langle x_1 + y_1, x_2 + y_2, \ldots, x_D + y_D \rangle$. vector sum

Figure **??** shows a simple example of vector addition and scalar product.

As you can deduce from the above, there is a vector, called the **zero vector**, often denoted $\mathbf{0}$, defined as $\langle 0, 0, \ldots, 0 \rangle \in \mathbb{R}^D$ that is the **additive identity**: that is, $\boldsymbol{x} + \mathbf{0} = \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^D$. zero vector additive identity

Figure **??** shows an example of a scalar product (left) of the vector $\langle 1, 2 \rangle$ with two values of a: 0.5 and 2. Note that multiplying by 0.5 brings the point "closer" to the origin, while multiplying by 2 pushes it out further. On the right, we see an example of vector addition between the vector $\langle 1, 2 \rangle$ and the vector $\langle 2, 1 \rangle$ to yield the vector $\langle 3, 3 \rangle$. As can be seen from this figure, it is often helpful to think of vectors as "rays" that point from the origin to the value $\boldsymbol{x}$. This makes visualizing vector addition more straightforward.

Given this construction, one can also define, for instance, subtraction. $\boldsymbol{x} - \boldsymbol{y}$ is just $\boldsymbol{x} + (-1)\boldsymbol{y}$, where (-1) is the real value of negative one.

The rules that govern scalar produces and vector sums agree with what one might imagine:

- $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$
- $a\mathbf{0} = \mathbf{0}$
- $0\boldsymbol{x} = \mathbf{0}$
- $a(\boldsymbol{x} + \boldsymbol{y}) = a\boldsymbol{x} + a\boldsymbol{y}$

TODO... Should we generalize here and talk about vector spaces in general, like ℓ_2 and Ł$_2$ and ℓ_1, etc.?

2.2 Vector Norms

In many cases, we care about measuring the **length** of a vector, or the **distance** between two vectors. The notion of a **vector norm** allows us to do this. length distance vector norm

For instance, in the (uninteresting) vector space $\mathbb{R}$, the standard norm would be **absolute value**. Using just the notion of absolute value, we can define the size of a number x as $|x|$ and the distance between two numbers x and y as $|x - y|$. absolute value

We now need to generalize this notion to arbitrary vector spaces, such as $\mathbb{R}^D$. A **norm** is any function g that maps vectors to real numbers that satisfies the following conditions: norm

- **Non-negativity**: for all $\boldsymbol{x} \in \mathbb{R}^D$, $g(\boldsymbol{x}) \geq 0$ Non-negativity
- **Strictly positive**: for all $\boldsymbol{x}$, $g(\boldsymbol{x}) = 0$ implies that $\boldsymbol{x} = \mathbf{0}$ Strictly positive
- **Homogeneity**: for all $\boldsymbol{x}$ and a, $g(a\boldsymbol{x}) = |a|\, g(\boldsymbol{x})$, where $|a|$ is the absolute value. Homogeneity
- **Triangle inequality**: for all $\boldsymbol{x}, \boldsymbol{y}$, $g(\boldsymbol{x} + \boldsymbol{y}) \leq g(\boldsymbol{x}) + g(\boldsymbol{y})$ Triangle inequality

These conditions state, in turn, the following. First, lengths are always positive. Second, a length of zero implies that you are zero. Third, scalar multiplication extends lengths in a predictable way. Fourth, distances add "reasonably."

Exercise 6. *Verify that for the vector space $\mathbb{R}$, the absolute value norm $g(x) = |x|$ satisfies all four conditions.*

An immediate question is: are norms *unique* (i.e., for every space, is there a *single, unique norm* for that space)? The answer is a resounding **no**!. For $\mathbb{R}^D$, there are *lots* of functions g that satisfy the above conditions. Here are some examples:

1. Euclidean norm: $g(\boldsymbol{x}) = \sqrt{\sum_{d=1}^{D} x_d^2}$
2. Manhattan norm: $g(\boldsymbol{x}) = \sum_{d=1}^{D} |x_d|$
3. Maximum norm: $g(\boldsymbol{x}) = \max_d |x_d|$
4. Zero norm: $g(\boldsymbol{x}) = \sum_{d=1}^{D} \mathbf{1}[x_d \neq 0]$

(Here, to define the zero norm, we have used an **indicator function**, denoted by "$\mathbf{1}[\bullet]$". The value of $\mathbf{1}[\bullet]$ is *one* whenever "$\bullet$" is "true" and *zero* otherwise.)

indicator function

These three norms behave quite differently. Euclidean norm is probably the one you're most familiar with. It corresponds in two dimensions to the Pythagorean theorem. Essentially, it measures length by walking in a straight line from the original to the point $\boldsymbol{x}$.

Manhattan norm (named because of the grid system for laying out streets in Manhattan—not unlike Salt Lake City!) measure length by walking along each dimension separately. You are not allowed to "cut across" diagonally.

Maximum norm measures the size of a vector as just the size of the maximum element in that vector.

Zero norm simply counts the number of non-zero elements in the vector.

Exercise 7. *Compute each of the four norms on the following vectors:*

1. $\langle 1, 2, 3 \rangle$
2. $\langle 1, -1, 0 \rangle$
3. $\langle 0, 0, 0 \rangle$
4. $\langle 1, 5, -6 \rangle$

Example 7. *Let's verify that the Euclidean norm is actually a norm (i.e., it satisfies the four conditions).*

- *Non-negativity: Let $\boldsymbol{x}$ be some vector; then look at $\sqrt{\sum_{d=1}^{D} x_d^2}$. We know this value will be non-negative so long as the sum is non-negative. But the sum is the sum of a bunch of values* squared*, so each of them has to be positive. Thus, Euclidean norm is non-negative.*
- *Strictly positive: suppose $\boldsymbol{x}$ is such that $g(\boldsymbol{x}) = 0$. For contradiction, suppose that $\boldsymbol{x} \neq \mathbf{0}$, which means there is some dimension d for which $x_d \neq 0$. But now it cannot be the case that $g(\boldsymbol{x}) = 0$ because $x_d^2 > 0$. This is a contradiction, so Euclidean norm is strictly positive.*
- *Homogeneity: let $\boldsymbol{x}$ and a be given. Then compute: $g(a\boldsymbol{x}) = \sqrt{\sum_{d=1}^{D} (ax_d)^2} = \sqrt{\sum_{d=1}^{D} a^2 x_d^2} = \sqrt{a^2 \sum_{d=1}^{D} x_d^2} = |a| \sqrt{\sum_{d=1}^{D} x_d^2} = |a|\, g(\boldsymbol{x})$.*
- *Triangle inequality: this follows directly from the Pythagorean theorem.*

Exercise 8. *Verify that the Manhattan norm, the Maximum norm and the Zero norm are actually norms. (These are actually easier than the Euclidean case.)*

One nice thing about these four norms is that they're actually specific cases of a *family* of norms called the **ell-p** norms, sometimes denoted by ℓ_p. (Note that it's a script ℓ, not a roman l.) This is either pronounced "ell p" or "little ell p", depending on the context. (There is another set of norms called the L_p norms, which

ell-p

are often called the "big ell p" norms.) These are defined as follows. Let p be in the range $[0, \infty]$; then the ℓ_p norm of $\boldsymbol{x}$, denoted by $||\boldsymbol{x}||_p$, is defined by:

$$||\boldsymbol{x}||_p = \left(\sum_{d=1}^{D} |x_d|^p \right)^{1/p} \tag{11}$$

Given this definition, it is easy to see that Euclidean norm is the ℓ_2 norm and Manhattan norm is the ℓ_1 norm.

Maximum norm is a bit harder to see: it is actually the ℓ_∞ norm. The way to think about this is as follows. Take a vector of (positive) numbers and raise them all to some gigantic (almost infinite; say 1000000) power. Even if the vector is $\langle 2, 2, 2, 2.001, 2, 2 \rangle$, after raising it to a gigantic power, the fourth element, $2.001^{1000000}$ is going to totally dominate: it will be almost-infinitely bigger than the others. So then when we add them up, we'll end up with a sum that is not really any different from $2.001^{1000000}$. So then when we raise the sum to $1/p$, we'll just get 2.001 back. (Yes, this is hand-wavy. But you can prove it formally by taking a limit as p tends toward infinity.)

Zero norm is also a bit tricky. Again, to prove it formally you have to argue in terms of limits as p approaches zero (from above). The intuition, however, is that as p goes to zero, any element x_d that is non-zero will map to $x_d^0 = 1$. On the other hand, any element x_d that *is* zero, will map just to zero. So the non-zero elements map to one and the zero elements map to zero and we sum the resulting vector. This gives us precisely the Zero norm.

Which norm should you use? It depends on your application. Definitely the most common are Euclidean (ℓ_2) and Manhattan (ℓ_1).

Once we've chosen a norm, we immediately get a method for computing distances. We define the **distance** between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ as $||\boldsymbol{x} - \boldsymbol{y}||$, where $||\bullet||$ denotes the norm of our choosing. Thus, you can think of the length of the vector $||\boldsymbol{x}||$ as the distance of that point to the origin. distance

We will often make use of **unit vectors**. These are vectors $\boldsymbol{x}$ that have unit norm: $||\boldsymbol{x}|| = 1$ (for whatever particular norm we are using). So long as $\boldsymbol{x}$ is not the zero vector, we can **normalize** $\boldsymbol{x}$ by multiplying it by $1/||\boldsymbol{x}||$. This yields a unit vector $\boldsymbol{x}/||\boldsymbol{x}||$ in the "same direction" (see Section 2.3) as $\boldsymbol{x}$, but with unit norm. unit vectors normalize

Exercise 9. *Verify that for non-zero $\boldsymbol{x}$, we have that the norm of $(1/||\boldsymbol{x}||)\boldsymbol{x} = 1$. For simplicity, first show that this is true for Euclidean norm. Next, show that it is true for* any *norm that satisfies the required properties.*

2.3 Dot Products

One thing that has been noticible absent from our discussion thus far is any notion of *multiplying* two vectors together. A standard *variety* of multiplication of vectors is the **dot product**. Before we define it, however, let's motivate it a bit. dot product

Let's say I hand you two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ and I want to know if they are **perpendicular** to each other or not. In two dimensions, they are perpendicular if the angle between them is 90 degrees. That doesn't answer our question, though, because we don't know how to generalize the notion of angle! Moreover, we might want to know if $\boldsymbol{x}$ and $\boldsymbol{y}$ are roughly in the same direction. The dot product allows us to answer these questions. perpendicular

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be vectors in $\mathbb{R}^D$. We define the **dot product** between $\boldsymbol{x}$ and $\boldsymbol{y}$, denoted $\boldsymbol{x} \cdot \boldsymbol{y}$ as: dot product

$$\boldsymbol{x} \cdot \boldsymbol{y} = \sum_{d=1}^{D} x_d y_d \tag{12}$$

Note that the dot product returns a *real value*, not another vector. (Sometimes you will see different notation for the dot product. The two other standard notations are $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ or $\boldsymbol{x}^\top \boldsymbol{y}$. We will actually use both of these in the future, for different purposes, so be fore-warned! The latter will make more sense when we talk about matrices—see Section 2.4).

Example 8. *We can compute the dot product between* $\langle 5, 3, -1 \rangle$ *and* $\langle 2, 0, 1 \rangle$ *as:*

$$(5 \times 2) + (3 \times 0) + (-1 \times 1) = 10 + 0 - 1 = 9 \tag{13}$$

Exercise 10. *Compute the following dot products:*

1. $\langle 1, 2, 3 \rangle \cdot \langle 4, 5, 6 \rangle$
2. $\langle 4, -1, 2 \rangle \cdot \langle 1, 1, 1 \rangle$
3. $\langle 0, 0, 1 \rangle \cdot \langle 1, -1, 1 \rangle$

Note that the dot product has a nice relationship to the Euclidean norm:

$$||\boldsymbol{x}||_2^2 = \boldsymbol{x} \cdot \boldsymbol{x} \tag{14}$$

The dot product (at least in Euclidean space) has a nice relationship to the **angle** between two vectors. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two vectors. Then it is easy to show that: angle

$$\boldsymbol{x} \cdot \boldsymbol{y} = ||\boldsymbol{x}||_2 \, ||\boldsymbol{y}||_2 \cos\theta \tag{15}$$

Where θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. Figure **??** has an example.

Thus, now that we have the notion of a dot product in hand, we can answer a whole host of questions. We say that two (non-zero) vectors are perpendicular if $\boldsymbol{x} \cdot \boldsymbol{y} = 0$ (this means that $\cos\theta = 90$). And, we can measure the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$ by their angle:

$$\theta = \arccos\left(\frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \, ||\boldsymbol{y}||}\right) \tag{16}$$

The dot product satisfies the following useful properties:

- **Commutativity**: $\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{y} \cdot \boldsymbol{x}$ Commutativity
- **Distributivity**: $\boldsymbol{x} \cdot (\boldsymbol{y} + \boldsymbol{z}) = \boldsymbol{x} \cdot \boldsymbol{y} + \boldsymbol{x} \cdot \boldsymbol{z}$ Distributivity

The more mathy term for "perpendicular" is **orthogonal**. We will stick with this word from now on. orthogonal

Exercise 11. *We claimed earlier that you can normalize a (non-zero) vector* $\boldsymbol{x}$ *by multiplying it by the scalar value* $(1/||\boldsymbol{x}||)$*, and that these two vectors are in the "same direction." Show that this is true by demonstrating that the angle between* $\boldsymbol{x}$ *and* $(1/||\boldsymbol{x}||)\boldsymbol{x}$ *is zero.*

One significant use of dot products is to evaluate **projections**. That is, $\boldsymbol{u} \cdot \boldsymbol{v}$ can be interpreted as a projection of the vector $\boldsymbol{u}$ onto the vector $\boldsymbol{v}$. It gives a scalar value that represents the distance that $\boldsymbol{u}$ goes in the direction of $\boldsymbol{v}$. See Figure **??** for a geometric intuition. This makes the most sense when $\boldsymbol{v}$ is a unit vector. See Section 2.5 for more details. projections

2.4 Matrices

A real-valued **matrix** is a rectangular collection of real values. For instance, we might define the following matrix: matrix

$$\mathbf{A} = \begin{bmatrix} 5 & 10 \\ -2 & 0 \\ 1 & -1 \end{bmatrix} \tag{17}$$

An index into a matrix is in row-by-column notation. This example matrix $\mathbf{A}$ has 3 rows and 2 columns. The value in the 2nd row and 1st column, denoted $A_{2,1}$ is -2. (This is a bit confusing at first for those used to thinking of x-by-y notation: it seems backwards.)

A matrix with N rows and M columns is usually called an $N \times M$ ("N by M") matrix, and we write $\mathbf{A} \in \mathbb{R}^{N \times M}$.

Note that we can think of *vectors* as matrices where one of the dimensions is 1. The standard is to say that a vector of length D is a matrix with D rows and 1 column. Thus, vectors are "tall skinny" matrices that live in $\mathbb{R}^{D \times 1}$.

Matrices can be manipulated in very similar ways to vectors, in terms of sums and products:

- For $\mathbf{A} \in \mathbb{R}^{N \times M}$ and $a \in R$, the matrix $a\mathbf{A}$ is also in $\mathbb{R}^{N \times M}$ with values given by $(a\mathbf{A})_{n,m} = aA_{n,m}$. (Here, we denote the n, mth element of $a\mathbf{A}$ as $(a\mathbf{A})_{n,m}$.)
- For $\mathbf{A}, \mathbf{B} \in R^{N \times M}$, the sum $\mathbf{A} + \mathbf{B}$ is of the same size with values given by $(\mathbf{A} + \mathbf{B})_{n,m} = A_{n,m} + B_{n,m}$

One convenient piece of notation is the **matrix cut**. Let $\mathbf{A}$ be $N \times M$. Then, we write $\boldsymbol{A}_{n,\bullet}$ to denote the *row vector* obtained by taking the nth row of $\mathbf{A}$. Similarly, $\boldsymbol{A}_{\bullet,m}$ is the *column vector* obtained by taking the mth column of $\mathbf{A}$. matrix cut

Just as we defined many different norms over vectors, we can also define different **matrix norms**. However, since we won't make use of these properties in any great depth, we'll just mention the most common. This is analogous to the Euclidean (ℓ_2) norm on vectors and is called the **Frobenius norm**: matrix norms Frobenius norm

$$||\mathbf{A}||_{\text{Fro}} = \sqrt{\sum_{n=1}^{N} \sum_{m=1}^{M} A_{n,m}^2} \tag{18}$$

There are *two* important notions of multiplication of matrices, each of which has different properties. The easiest to understand is the **Hadamard product**, also called the **element-wise product**, typically denoted $\odot$: Hadamard product element-wise product

$$(\mathbf{A} \odot \mathbf{B})_{n,m} = A_{n,m} B_{n,m} \tag{19}$$

The Hadamard product is only defined over matrices of equal size (say, $N \times M$) and returns a matrix of that size (again, $N \times M$). The value in each cell of the returned matrix is just the product of the values in the corresponding cells of the original two matrices.

Example 9. *An easy example of Hadamard product:*

$$\begin{bmatrix} 5 & 10 \\ -2 & 0 \\ 1 & -1 \end{bmatrix} \odot \begin{bmatrix} 2 & -1 \\ -1 & 5 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 \times 2 & 10 \times -1 \\ -2 \times -1 & 0 \times 5 \\ 1 \times 0 & -1 \times 1 \end{bmatrix} = \begin{bmatrix} 10 & -10 \\ 2 & 0 \\ 0 & -1 \end{bmatrix} \tag{20}$$

Exercise 12. *The two two matrices from the previous example and compute the Hadamard product between each and itself. How does this relate to Frobenius norm?*

Hadamard multiplication is commutative: $\mathbf{A} \odot \mathbf{B} = \mathbf{B} \odot \mathbf{A}$; associative: $\mathbf{A} \odot (\mathbf{B} \odot \mathbf{C}) = (\mathbf{A} \odot \mathbf{B}) \odot \mathbf{C}$); and distributive: $\mathbf{A} \odot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \odot \mathbf{B} + \mathbf{A} \odot \mathbf{C}$.

The second, perhaps more common, type of **matrix multiplication** is the **inner product**. Let $\mathbf{A}$ be $N \times K$ and let $\mathbf{B}$ be $K \times M$. Then the inner product of $\mathbf{A}$ and $\mathbf{B}$, written $\mathbf{AB}$ is a matrix *of size* $N \times M$. It is called an inner product because the inner dimensions of the multiplication must "match up." The product is defined as:

matrix multiplication

inner product

$$(\mathbf{AB})_{n,m} = \sum_{k=1}^{K} A_{n,k} B_{k,m} \tag{21}$$

This can be though of in terms of matrix cuts. The n, mth cell in the resulting matrix is given by the (vector) dot product $\langle \boldsymbol{A}_{n,\bullet}, \boldsymbol{B}_{\bullet,m} \rangle$. Note that since the internal dimension size K must match up, this vector dot product is well defined.

Example 10. *As an example of matrix multiplication, we have:*

$$\underbrace{\begin{bmatrix} 5 & -2 & 1 \\ 10 & 0 & -1 \end{bmatrix}}_{dims\ 2\times3} \underbrace{\begin{bmatrix} 2 & -1 \\ -1 & 5 \\ 0 & 1 \end{bmatrix}}_{dims\ 3\times2} = \overbrace{\begin{bmatrix} 5\times2 + (-2)\times(-1) + 1\times0 & 5\times(-1) + (-2)\times5 + 1\times1 \\ 10\times2 + 0\times(-1) + -1\times0 & 10\times(-1) + 0\times5 + -1\times1 \end{bmatrix}}^{dims\ 2\times2}$$

$$= \begin{bmatrix} 12 & -14 \\ 20 & -11 \end{bmatrix}$$

Exercise 13. *Compute the following matrix products:*

1. $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} -1 \\ -2 \\ -3 \end{bmatrix}$
2. $\begin{bmatrix} -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ -2 & 2 \\ -3 & 3 \\ -4 & 4 \end{bmatrix}$

Matrix multiplication is associative: $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$); and distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$, but is **not commutative**! In general, it is *not true* that $\mathbf{AB} = \mathbf{BA}$ (even if $\mathbf{A}$ and $\mathbf{B}$ are square).

A useful matrix operation is the **transpose** operator, denoted "$^\top$". The transpose of a matrix $\mathbf{A}$ is simply the matrix you get by "rotating" $\mathbf{A}$. That is, if we start out with $\mathbf{A} \in \mathbb{R}^{N \times M}$, then the transpose, denoted $\mathbf{A}^\top$, is now in $\mathbb{R}^{M \times N}$. The values are defined by:

transpose

$$(\mathbf{A}^\top)_{m,n} = A_{n,m} \tag{22}$$

Example 11. *For instance, we can compute the following transpose:*

$$\begin{bmatrix} 5 & 10 \\ -2 & 0 \\ 1 & -1 \end{bmatrix}^\top = \begin{bmatrix} 5 & -2 & 1 \\ 10 & 0 & -1 \end{bmatrix} \tag{23}$$

The notion of transposition is where we get the previously mention notation for vector **dot products**. The vector $\langle 1, 2, 3 \rangle$ corresponds to the 3×1 matrix $[1 \quad 2 \quad 3]^\top$. Similarly, the vector $\langle 4, 5, 6 \rangle$ corresponds to the

dot products

3×1 matrix $[4 \;\; 5 \;\; 6]^\top$. Taking the *transpose* of the first, we get a 3×1 matrix and a 1×3 matrix. These can be multiplied (in the matrix multiplication sense) to yield a scalar value:

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}^\top \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = [1 \;\; 2 \;\; 3] \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = (1 \times 4) + (2 \times 5) + (3 \times 6) = 5 + 10 + 18 = 33 \tag{24}$$

Which is precisely the dot product between these two vectors. This is where the notation $\boldsymbol{x}^\top \boldsymbol{y}$ for the dot product between two vectors comes from.

Exercise 14. *Let* $\mathbf{A}$ *be the left-most matrix from Eq* (20)*. What is the value* $\mathbf{A}^\top \mathbf{A}$*? What is the value* $\mathbf{A}\mathbf{A}^\top$*? How does these relate to the Frobenius norm?*

We say that a matrix is **square** if it has the same number of rows as columns. A square matrix $\mathbf{A}$ is **symmetric** if $\mathbf{A} = \mathbf{A}^\top$.

square

symmetric

Exercise 15. *Which of the following matrices are symmetric:*

1. $\begin{bmatrix} 1 & 4 \\ 4 & 1 \end{bmatrix}$
2. $\begin{bmatrix} 1 & 4 & 5 \\ 4 & 2 & 7 \\ 5 & 2 & 6 \end{bmatrix}$
3. $\begin{bmatrix} 1 & 4 & 5 \\ 4 & 2 & 7 \\ 5 & 7 & 6 \end{bmatrix}$

2.5 Projections

A very important geometric concept is the notion of an **orthogonal projection**. We've already (almost) seen an example of this for vectors. In Figure **??**, we have two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. One question we might want to ask is: how far does $\boldsymbol{x}$ reach "in the direction of" $\boldsymbol{y}$? The more mathy way to state this is: what is the length of the **projection** of $\boldsymbol{x}$ onto $\boldsymbol{y}$?

orthogonal projection

projection

One way of thinking about this is the case when $\boldsymbol{y}$ is simply one of the axes. I.e., $\boldsymbol{y} = \langle 1, 0 \rangle$ (the "x-axis") or $\boldsymbol{y} = \langle 0, 1 \rangle$ (the "y-axis"). Then, the question we're asking is: how far does $\boldsymbol{x}$ point in the x direction or the y direction? In the axis-aligned case, this is easy! It points x_1 units along the first direction and x_2 units along the second direction! See Figure **??**.

Things become only slightly more complicated when $\boldsymbol{y}$ is an arbitrary vector. The notion of vector projection is shown pictorially in Figure **??**. The projection of $\boldsymbol{x}$ onto $\boldsymbol{y}$ is the component of $\boldsymbol{x}$ that points in the direction of $\boldsymbol{y}$. It turns out that the *length* of this projection (sometimes called the **scalar component**) is precisely given by the dot project of the two vectors: $\boldsymbol{x}^\top \boldsymbol{y}$!

scalar component

That is, the dot product $\boldsymbol{x}^\top \boldsymbol{y}$ gives us the length of the component of $\boldsymbol{x}$ along $\boldsymbol{y}$. It *doesn't* give us the actual vector. However, this is also easy to come by. We just need a vector that points in the same direction as $\boldsymbol{y}$ but has length $\boldsymbol{x}^\top \boldsymbol{y}$. As we've seen previously, we can obtain this by just multiplying $\boldsymbol{y}$ by the scalar value $(\boldsymbol{x}^\top \boldsymbol{y}) / ||\boldsymbol{y}||$.

Example 12. *Take the simple case where* $\boldsymbol{y} = \langle 1, 0 \rangle$*. Let's say* $\boldsymbol{x} = \langle 0.5, 6 \rangle$*. It's fairly intuitive in this case that* $\boldsymbol{x}$ *points* 0.5 *units in the direction of* $\boldsymbol{y}$*. We can verify this by computing* $\boldsymbol{x}^\top \boldsymbol{y} = 0.5 \times 1 + 6 \times 0 = 0.5$*, precisely as we wanted. Moreover, we can find the actual vector corresponding to the projection of* $\boldsymbol{x}$ *onto* $\boldsymbol{y}$ *as* $(0.5)\boldsymbol{y} = \langle 0.5, 0 \rangle$*, again, as expected.*

Exercise 16. *Take $\boldsymbol{x}$ from the previous example and project it onto the* other *axis. What is the length of this projection? What is the vector corresponding to the projection?*

Exercise 17. *Compute the projection of $\boldsymbol{x}$ onto $\boldsymbol{y}$ (as a vector) for each of the following pairs:*

1. $\boldsymbol{x} = \langle 6, 0.5\rangle$ *and* $\boldsymbol{y} = \langle 1, 2\rangle$
2. $\boldsymbol{x} = \langle 1, 2\rangle$ *and* $\boldsymbol{y} = \langle 1, 2\rangle$
3. $\boldsymbol{x} = \langle 2, 3\rangle$ *and* $\boldsymbol{y} = \langle 1, 0\rangle$
4. $\boldsymbol{x} = \langle -2, 3\rangle$ *and* $\boldsymbol{y} = \langle 1, 0\rangle$

One way to think about projections of $\boldsymbol{x}$ onto $\boldsymbol{y}$ is in terms of projections of vectors onto the one-dimensional subspace defined by $\{\boldsymbol{y}\}$. If we think of the $\boldsymbol{y}$ a one dimensional subspace, then the projection of $\boldsymbol{x}$ onto $\boldsymbol{y}$ is precisely the projection of $\boldsymbol{x}$ onto the **subspace** defined by $\boldsymbol{y}$. subspace

Now, let's consider some special cases. Let $\boldsymbol{u} = \langle 2, 0.5$ and let $\boldsymbol{v} = \langle 1, 0\rangle$ and $\boldsymbol{w} = \langle 0, 1\rangle$. Note that both $\boldsymbol{v}$ and $\boldsymbol{w}$ are unit vectors. What is the projection of $\boldsymbol{u}$ onto $\boldsymbol{v}$? 2. And onto $\boldsymbol{w}$? 0.5. We note an interesting fact: we can write $\boldsymbol{u}$ as $2\boldsymbol{v} + 0.5\boldsymbol{w}$. In fact, this is a general fact, so long as $\boldsymbol{v}$ and $\boldsymbol{w}$ are unit and orthogonal (which you can easily verify). See Figure **??**.

Let's take this a step further. Let $\boldsymbol{v} = \langle \sqrt{2}/2, \sqrt{2}/2\rangle$ and $\boldsymbol{w} = \langle -\sqrt{2}/2, \sqrt{2}/2\rangle$. Again, you can verify that these are unit vectors and that they are orthogonal. Now, let's compute:

$$\boldsymbol{u} \cdot \boldsymbol{v} = \sqrt{2} + \sqrt{2}/4 = \tfrac{5}{4}\sqrt{2} \tag{25}$$

$$\boldsymbol{u} \cdot \boldsymbol{w} = \sqrt{2} - \sqrt{2}/4 = \tfrac{3}{4}\sqrt{2} \tag{26}$$

This implies that we can write $\boldsymbol{u} = \frac{5}{4}\sqrt{2}\boldsymbol{u} + \frac{3}{4}\sqrt{2}\boldsymbol{w}$, which we can verify geometrically in Figure **??**.

This way of thinking is useful to generalize the notion of projecting a vector onto a vector to that of projecting a vector onto a matrix. Let $\boldsymbol{x}$ be a vector in $\mathbb{R}^D$ and let $\mathbf{A}$ be an $N \times D$ matrix. As before, we think of the span of $\mathbf{A}$ as the span of the N-many D-dimensional vectors obtained through cuts of $\mathbf{A}$. The projection of $\boldsymbol{x}$ onto $\mathbf{A}$ is just the projection of $\boldsymbol{x}$ onto the subspace spanned by $\mathbf{A}$: see Figure **??**.

Computing the projection of $\boldsymbol{x} \in \mathbb{R}^D$ onto $\mathbf{A} \in \mathbb{R}^{N \times D}$ for arbitrary $\mathbf{A}$ is actually difficult. However, if we can assume that $\mathbf{A}$ is **orthonormal**, then it is easy. For $\mathbf{A}$ to be orthonormal means that two things hold: orthonormal

1. Normalization constraint: for all n, $||\boldsymbol{A}_{n,\bullet}|| = 1$. That is, each D-vector in $\mathbf{A}$ is normalized.
2. Orthogonality constraint: for all $1 \leq n \neq n' \leq N$, $\boldsymbol{A}_{n,\bullet}$ and $\boldsymbol{A}_{n',\bullet}$ are orthogonal: their dot product is zero.

The nice thing about orthonormal sets is that they make various computations very easy. If A is orthonormal, then we can project $\boldsymbol{x}$ independently onto each cut $\boldsymbol{A}_{n,\bullet}$ of A. This is just a vector projection, so has length $\boldsymbol{A}_{n,\bullet}\boldsymbol{x}$. Then, for each component n in A, we have the projection of $\boldsymbol{x}$ along that vector as $(\boldsymbol{A}_{n,\bullet}\boldsymbol{x})\boldsymbol{A}_{n,\bullet}$ (just as before, in the vector projection case).

Putting this all together, the length of the projection of $\boldsymbol{x}$ onto $\mathbf{A}$ is $||\mathbf{A}\boldsymbol{x}||$ (this is just the length of each of the individual vectors) and the projection itself is: $\sum_n (\boldsymbol{A}_{n,\bullet}\boldsymbol{x})\boldsymbol{A}_{n,\bullet}$. This projection indeed lies in the span of $\mathbf{A}$, since it is clearly a linear combination of the rows of $\mathbf{A}$.

If $\mathbf{A}$ is *not* orthonormal, then we must first orthonormal-ize it and then do this projection. We will talk about how to do this in Section 2.7.

2.6 Important Matrices

There are a few matrices that will come up over an over again. The first is the **identity matrix**. The identity matrix is always **square** (that is, it has dimension $D \times D$). The D-dimensional identity matrix is the matrix that has zeros in every cell *except* the diagonal. It is denoted $\mathbf{I}_D$, or, when the dimensionality is clear from context, just $\mathbf{I}$. A few examples:

identity matrix
square

$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad ; \quad \mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad ; \quad \mathbf{I}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{27}$$

The identity matrix has several nice properties. Assuming dimensions match up, we have:

- $\mathbf{I}^\top = \mathbf{I}$
- $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$

Exercise 18. *Prove the second claim.*

Another important matrix is the **zero matrix**, denoted $\mathbf{0}$ (not to be confused with the zero vector, $\mathbf{0}$). This is the (not necessarily square) matrix of all zeros. Again, assuming dimensions match up, the zero matrix has the nice properties that:

zero matrix

- $\mathbf{0}^\top = \mathbf{0}$
- $\mathbf{0A} = \mathbf{A0} = \mathbf{0}$

Finally, there's the **ones matrix**, denoted $\mathbf{1}$ (not to be confused with the indicator function $\mathbf{1}[\bullet]$). This is the (not necessarily square) matrix of all ones. It only obeys the symmetry property (when it is square).

ones matrix

2.7 Matrix Properties: Trace, Determinant and Rank

At this point, you're going to have to suspend disbelief and just acknowledge that what we're about to talk about is at all important. As we progress, it will become obvious that these things are *very* important. But for now, just have faith.

Before progressing, we need to define the **diagonal** of a matrix $\mathbf{A}$. Let $\mathbf{A}$ be an arbitrary square $D \times D$ matrix. The diagonal of $\mathbf{A}$, denoted $\text{diag}(\mathbf{A})$ is the D-dimensional vector with compondent d equal to $A_{d,d}$.

diagonal

Example 13. *We can compute:*

$$\text{diag}\left(\begin{bmatrix} 1 & 4 & 5 \\ 4 & 2 & 7 \\ 5 & 2 & 6 \end{bmatrix}\right) = \langle 1, 2, 6 \rangle \tag{28}$$

We say that a matrix $\mathbf{A}$ is **diagonal** if all of its off-diagonal elements are zero. For instance, the identity matrix and the zeros matrix are both diagonal, but the ones matrix is not. We say that a matrix is **upper-triangular** if all of the elements below the diagonal are zero. Similarly, it is **lower-triangular** if all of the elements above the diagonal are zero. It is **triangular** if either condition holds.

diagonal
upper-triangular
lower-triangular
triangular

The **trace** of a matrix $\mathbf{A}$, denoted $\text{tr}\,\mathbf{A}$, is the sum of the elements along the diagonal of $\mathbf{A}$. That is: $\text{tr}\,\mathbf{A} = \sum_{d=1}^{D} A_{d,d}$.

trace

Exercise 19. *Compute the trace of the matrix from the previous example.*

The trace of a matrix, although simplistic, is often used as a measure of the "size" of a matrix. Note that it is not a norm, in that it doesn't satisfy the necessary requirements, but it is sort of reasonable.

Exercise 20. *What is the value of* $\mathrm{tr}[\mathbf{A}\mathbf{A}^\top]$*? What about* $\mathrm{tr}[\mathbf{A}^\top\mathbf{A}]$*? What do these remind you of?*

While the trace is fairly easy to comprehend, the **determinant** is quite a bit more complicated. We'll start with a special case of 2×2 matrices. The determinant of a matrix $\mathbf{A}$, denoted $\det \mathbf{A}$, is defined as follows for 2×2 matrices:

determinant

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc \tag{29}$$

Note that if $\mathbf{A}$ is diagonal, the the determinant is the *product* of the elements along its diagonal. However, if it's not diagonal, the determinant is quite a bit different.

Exercise 21. *Compute the determinant of the following matrices:*

1. $\mathbf{I}_2$
2. $\begin{bmatrix} 5 & 1 \\ 2 & 4 \end{bmatrix}$
3. $\begin{bmatrix} 2 & 10 \\ 3 & 6 \end{bmatrix}$

Extending the notion of a determinant to larger matrices is a bit involved. We'll give a recursive definition and then give an example.

Let $\mathbf{A}$ be a $D \times D$ matrix. Denote by $\mathbf{A}_{-i,-j}$ the matrix obtained by *removing* the ith row and jth column from $\mathbf{A}$. (Thus, $\mathbf{A}_{-i,-j}$ is a $(D-1) \times (D-1)$ matrix.) For arbitrary $\mathbf{A}$, the determinant of $\mathbf{A}$ is:

$$\det \mathbf{A} = \sum_{d=1}^{D} (-1)^{1+d} A_{1,d} \det \mathbf{A}_{-1,-d} \tag{30}$$

We can end the recursion once we've gotten down to a 2×2 matrix, for which we know how to perform the computation.

What this computation is doing is the following. We loop (d) over every column of $\mathbf{A}$. On column d, we remove the first row and dth column from $\mathbf{A}$ and recursively compute the determinant on that smaller matrix. We multiply this value by $A_{1,d}$. Then, if d is odd, we *add* the resulting value to the running sum; if d id even, we *subtract* the resulting value. (This last part is encoded in the $(-1)^{1+d}$ term.)

It is not terribly important for our purposes that you know how to compute a determinant (there are much more efficient methods than actually evaluating the recursion).

Here are some important observations about determinants:

- If $\mathbf{A}$ is triangular, then the determinant is the product of the diagonal of $\mathbf{A}$.
- If we take some row $\boldsymbol{A}_{d,\bullet}$ of $\mathbf{A}$ and *add* it to another row (perhaps scaled in some way), then the determinant is *unchanged*!
- If two rows of $\mathbf{A}$ are swapped to produce $\mathbf{B}$, then $\det \mathbf{A} = -\det \mathbf{B}$.
- If one row of $\mathbf{A}$ is multiplied by a scalar a to produce $\mathbf{B}$, then $\det \mathbf{B} = k \det \mathbf{A}$.

This last property tells us that $\det(a\mathbf{A}) = a^D\mathbf{A}$, since the standard scalar product over matrices multiples *every* row by the scalar a, and there are D-many rows.

These properties may seem mysterious at first (the certainly did to me!), but their importance will become clear later on when we start using determinants to do computation.

The difficulty arises with $\mathbf{A}$ is *not* orthonormal. (Actually, the normalization aspect is irrelevant. We only are worried with $\mathbf{A}$ is not orthogonal.) In this case, we have to ask ourselves: what is the size of the smallest matrix $\mathbf{B}$ that has the same span as $\mathbf{A}$. Such a matrix $\mathbf{B}$ will not be unique, but its dimensionality will be.

Example 14. *Consider the matrix defined by the set of vectors in R^3: $\{\langle 1,0,1\rangle, \langle 0,1,0\rangle, \langle 1,2,1\}$. This does* not *define an orthogonal set. In particular, the third element can be obtained as a linear combination of the first two, and is therefore not orthogonal. In this case, we can see that if we simply* drop *the third element, we are left with a set of* 2 *points that* are *orthogonal and* do *span the same space as the original set. This means the the* rank *of the original matrix was actually* 2*.*

The important property of the rank of a matrix is that it tells us the true size of the subspace spanned by this matrix. We say that a matrix $\mathbf{A} \in \mathbb{R}^{N\times D}$ (with $N \leq D$) is **full rank** if the rank of $\mathbf{A}$ is equal to N. In general, we like to work with full rank matrices: a matrix that is not full rank is somehow wasteful.

full rank

2.8 Matrix Inversion

We've seen previously (Section 2.6) that the identity matrix $\mathbf{I}$ behaves in a reasonable way: $\mathbf{IA} = \mathbf{A}$ for all $\mathbf{A}$. This is just like how, in real numbers, $1x = x$ for all x. Another property that 1 (the real number) has is that for any $x \neq 0$, there exists a y such that $xy = 1$. We call y the "inverse" of x and write it either as $1/x$ or as x^{-1}.

The question that arises is: given a matrix $\mathbf{A}$, does there exist a matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{I}$? If $\mathbf{A}$ is $D \times D$, it turns out that $\mathbf{A}$ is **invertible** if and only if $\mathbf{A}$ is **full rank** (see Section 2.7)!

invertible
full rank

So, provided that $\mathbf{A}$ is full rank and square, we denote by $\mathbf{A}^{-1}$ the **inverse** of $\mathbf{A}$. The inverse satisfies the following properties:

inverse

1. Commutativity: $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$
2. Scalar multiplication: $(a\mathbf{A})^{-1} = a^{-1}\mathbf{A}^{-1}$, provided $a \neq 0$
3. Transposition: $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$
4. Product: $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

Since matrix multiplication is not commutative, it is important to be careful when trying to use matrix inverses to solve equations.

Example 15. *Given the equation $\mathbf{X} = \mathbf{AB}$ for given matrices $\mathbf{A}$ and $\mathbf{X}$, we wish to solve for $\mathbf{B}$. (Suppose everything is square and full rank.) In order to do this, we* multiply *both sides of the equation by $\mathbf{A}^{-1}$* on the left*. This yields: $\mathbf{A}^{-1}\mathbf{X} = \mathbf{A}^{-1}\mathbf{AB} = \mathbf{B}$, which gives us a solution.*

Exercise 22. *Given $\mathbf{X} = \mathbf{A}(\mathbf{B} + \mathbf{C})$, where everything except $\mathbf{B}$ is known, solve for $\mathbf{B}$. (Assume everything is square and full rank.)*

One interesting fact is that the determinant of a square matrix $\mathbf{A}$ is *non-zero* if and only if it has full rank. This means that an alternative way of checking to see if a matrix $\mathbf{A}$ is invertible is to check that $\det \mathbf{A} \neq 0$.

Sometimes it is necessary to have an inverse-like quantity for non-square matrices. Suppose that $\mathbf{A}$ is $M \times N$. We say that $\mathbf{A}^\dagger$ is a **pseudo-inverse** of $\mathbf{A}$ if the following four criteria hold:

pseudo-inverse

1. $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A}$
2. $\mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}$
3. $(\mathbf{A}\mathbf{A}^\dagger)^\top = \mathbf{A}\mathbf{A}^\dagger$
4. $(\mathbf{A}^\dagger\mathbf{A})^\top = \mathbf{A}^\dagger\mathbf{A}$

It turns out that the pseudo-inverse exists and is unique for *any* matrix $\mathbf{A}$. In the case that $\mathbf{A}$ is invertible, $\mathbf{A}^{-1} = \mathbf{A}^\dagger$, so the pseudo-inverse extends the notion of inverse.

Sometimes the pseudo-inverse is called the **Moore-Penrose inverse**. Moore-Penrose inverse

2.9 Eigenvectors and Eigenvalues

TODO... write this!

2.10 Wrap-up

The most important concepts from this section are:

- Definition of basic operations on vectors and matrices
- Interpreting dot products as projections
- Vector and matrix norms
- The properties of the matrix inverse (and that you can't always invert a matrix!)

The matrix cookbook (`http://matrixcookbook`) provides lots more detail on all of these things and more!

3 Multidimensional Calculus

We've already hinted several times that differentiation in high dimensional spaces is going to be important.

For example, suppose we have vectors $\boldsymbol{w}$ and $\boldsymbol{x}$ and a function $f(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$. We might want to differentiate f with respect to $\boldsymbol{w}$ so that we can (for instance) maximize it. This is where differentiation with respect to vectors becomes important.

We'll start with a definition and then just convince you that differentiating in high dimensions is exactly the same as differentiating in one dimension.

Suppose $f(\boldsymbol{w})$ is some function of a D-dimensional vector $\boldsymbol{w} = \langle w_1, w_2, \ldots, w_D \rangle$. We can compute *partial derivatives* of f with respect to each component of $\boldsymbol{w}$:

$$\frac{\partial f}{\partial w_1} \quad , \quad \frac{\partial f}{\partial w_2} \quad , \quad \frac{\partial f}{\partial w_3} \quad , \quad \cdots \quad , \quad \frac{\partial f}{\partial w_D} \tag{31}$$

Each of these are just univariate derivatives. If we package them together into a vector, we get the **gradient** of f with respect to the *vector* $\boldsymbol{w}$. Gradient is just the fancy term for multidimensional derivative. It is denoted by $\nabla_{\boldsymbol{w}} f$, and is formally defined as: gradient

$$\nabla_{\boldsymbol{w}} f = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \frac{\partial f}{\partial w_3} \\ \vdots \\ \frac{\partial f}{\partial w_D} \end{bmatrix} \tag{32}$$

Example 16. *Let's take a simple example. Suppose* $f(\boldsymbol{w}) = \boldsymbol{w}^\top \langle 1, 2, 3, 0, -1 \rangle = 1w_1 + 2w_2 + 3w_3 + 0w_4 - 1w_5$. *We can compute:*

$$\frac{\partial f}{\partial w_1} = 1 \tag{33}$$
$$\frac{\partial f}{\partial w_2} = 2 \tag{34}$$
$$\frac{\partial f}{\partial w_3} = 3 \tag{35}$$
$$\frac{\partial f}{\partial w_4} = 0 \tag{36}$$
$$\frac{\partial f}{\partial w_5} = -1 \tag{37}$$
$$\tag{38}$$

So the gradient is simply $\nabla_{\boldsymbol{w}} f = \langle 1, 2, 3, 0, -1 \rangle$. *Note that this is* complete analogous *to a standard derivative.* f *is just the "product" of* $\boldsymbol{w}$ *and* $\langle 1, 2, 3, 0, -1$ *and it's "derivative" is just the vector* $\langle 1, 2, 3, 0, -1 \rangle$.

Example 17. *Let's take another example that's slightly more complicated:* $f(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{w}$ *(note that this is just squared Euclidean norm). Rewriting in non-vector notation, we have* $f(\boldsymbol{w}) = \sum_i w_i^2$. *Let's compute partial derivatives:*

$$\frac{\partial f}{\partial w_1} = 2w_1 \tag{39}$$
$$\frac{\partial f}{\partial w_2} = 2w_2 \tag{40}$$
$$\vdots \tag{41}$$
$$\frac{\partial f}{\partial w_D} = 2w_D \tag{42}$$
$$\tag{43}$$

Note, that, in general, $\partial f / \partial w_i = 2w_i$, *which means that* $\nabla_{\boldsymbol{w}} f = 2\boldsymbol{w}$*! Again,* $\boldsymbol{w}^\top \boldsymbol{w}$ *is basically like* w^2 *and the derivative is just like* $2w$*!*

Example 18. *Let's do one more really simple example. Note that we can express "the sum of the elements of* $\boldsymbol{w}$*" as* $f(\boldsymbol{w}) = \mathbf{1}^\top \boldsymbol{w}$, *where* $\mathbf{1}$ *is a vector of ones. Can you guess what* $\nabla_{\boldsymbol{w}} f$ *is? Let's work it out.* $f(\boldsymbol{w}) = \sum_i 1w_i$, *so* $\partial f / \partial w_j = 1$ *for all* j. *Thus,* $\nabla_{\boldsymbol{w}} f = \mathbf{1}$. **Note** *that this is the* ones vector, *and is* not *the constant one.*

As you might expect, function composition works in exactly the same way as in univariate calculus.

Exercise 23. *Suppose $f(\boldsymbol{w}) = (\boldsymbol{w}^\top \boldsymbol{x})^2$ for some constant vector $\boldsymbol{x}$. Compute $\nabla_{\boldsymbol{w}} f$.* Hint: *Begin by writing f in non-vector form. Then take the derivative with respect to some component w_j. Package these up into a vector to get the gradient.*

Exercise 24. *Similar to the previous exercise, verify that when $f(\boldsymbol{w}) = \exp[\boldsymbol{w}^\top \boldsymbol{x}]$, we get $\nabla_{\boldsymbol{w}} f = \exp[\boldsymbol{w}^\top \boldsymbol{x}]\boldsymbol{x}$.*

We can generalize this as follows. Suppose $f(\boldsymbol{w}) = g(h(\boldsymbol{w}))$ for some arbitrary functions g and h. We wish to compute $\nabla_{\boldsymbol{w}} f$. First, we compute the gradient of the "inside" and let $\boldsymbol{u} = \nabla_{\boldsymbol{w}} h$. Now, we take the derivative of g – which we'll call g' here. The gradient of f is then just $\boldsymbol{u} g'(h(\boldsymbol{w}))$, as in univarite calculus.

As you can see, most of these conform to our intuitions based on univariate calculus. There are lots of identities listed in the matrix cookbook (`http://matrixcookbook`):

Just as gradients are the multidimensional equivalents of derivatives, the **Hessian matrix** is the multidimensional equivalents of second derivatives. Let $f(\boldsymbol{w})$ be a D-dimensional function. The Hessian of f is the *matrix* of second derivatives: Hessian matrix

$$\mathbf{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_D} \\ \frac{\partial^2 f}{\partial w_1 \partial w_2} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 \partial w_D} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial w_1 \partial w_D} & \frac{\partial^2 f}{\partial w_2 \partial w_D} & \cdots & \frac{\partial^2 f}{\partial w_D^2} \end{bmatrix} \tag{44}$$

That is, it is a matrix where $\Big(\mathbf{H}(f)\Big)_{i,j} = \frac{\partial^2 f}{\partial w_i \partial w_j}$. By physical analogy, the derivative of a function is it's rate of change ("velocity"). The gradient of a function is its velocity in each available dimension. The second derivative of a function is the rate of change of it's velocity, aka it's acceleration. The *Hessian* of a function is the rate at which different dimension accelerate together. In other words, if $H_{2,2}$ is high, then it means that there is a high (positive) rate of change in the second dimension. If $H_{1,2}$ is high, then it means that as we accelerate in the first dimension, we simultaneously accelerate in the second dimension. If $H_{1,2}$ is negative, it means that as we accelerate in the first dimension, we *decelerate* in the second dimension.

3.1 Multidimensional Convexity

We discussed the idea of convex subsets of $\mathbb{R}$ in Section 1.3. Namely, a set $A \subseteq \mathbb{R}$ is convex if and only if $\forall x, y \in A, \forall \lambda \in [0,1]$ it holds that $\lambda x + (1-\lambda) y \in A$. The definition for high dimensional spaces is identical. Let $\mathbb{R}^D$ be D-dimensional Euclidean space, then $A \subset \mathbb{R}^D$ is convex if and only if for all vectors $\boldsymbol{x}, \boldsymbol{y} \in A$ and all $\lambda \in [0,1]$, we have that $\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in A$. Figure **??** shows an example of a convex set and a non-convex set. Convex sets look like balls, non-convex sets look like potatoes (or worse!).

Example 19. *Let A_p be a D dimensional ℓ_p ball centered at the origin. Namely A_p is the set of all points with ℓ_p norm less than or equal to one; that is: $A_p = \{\boldsymbol{x} : ||\boldsymbol{x}||_p \leq 1\}$. Figure **??** shows examples of such balls for different p. As we can see from these figures, the ℓ_p balls for $p < 1$* are non-convex *but the ℓ_p balls for $p \geq 1$* are convex!.

Exercise 25. *Prove the above claim for $p = 0.5$, $p = 1$ and $p = 2$.*

Exercise 26. *Prove the above claim for all p. (This is kind of hard.)*

Convex functions in higher dimensional space are actually harder to define that convex sets.

Now, recall from one dimension that f is convex if $f'' \geq 0$ (it's strictly convex if $f'' > 0$). This was easy, because f'' is just a scalar value. How we have a matrix and we need some equivalent statement. The

equivalent statement is that the matrix $\mathbf{H}$ is **positive semi-definite**, written $\mathbf{H} \succeq 0$ (in Latex, this is `succeq`). (Similarly, it is strictly convex if $\mathbf{H} \succ 0$, or $\mathbf{H}$ is **positive definite**.)

positive semi-definite
positive definite

3.1.1 Positive Semi-Definite-ness

The whole notion of being positive semi-definite is something that is hard to grasp, but it comes up all the time. The reason it comes up all the time is because begin positive semi-definite is the analogue for matrices of just being a non-negative real number: something that comes up all the time! It's terribly annoying to write "positive semi-definite" all the time, so we will just write **psd** in the future.

psd

There are actually lots of ways to define what it means to be psd. We'll do it by making an analogy to just regular positive numbers. Let a be a real value. We'll define a to be "fluffy" if the following holds: for all real values x, we have that $ax^2 \geq 0$. This is a slightly odd definition, but the important thing to realize is that a is fluffy *if and only if* a is non-negative! That is, we've essentially redefined non-negativeness in a very strange way. Note that we can equivalently write the fluffy requirement as being: for all x, $xax \geq 0$.

Now, let's move to matrices. Define a matrix $\mathbf{A}$ to be "multi-dimensional fluffy" if for all vectors $\boldsymbol{x}$, we have that $\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} \geq 0$. If we carry out the computation, this is just saying that for all $\boldsymbol{x}$, we have that $\sum_{i,j} A_{i,j} x_i x_j \geq 0$. Hopefully it's clear that being multi-dimensional fluffy is a strict generalization of being fluffy.

Well, and here's the magic: multi-dimensional fluffy is just the definition of being positive semi-definite! That is, a psd matrix is just the multi-dimensional analogue of a non-negative real value!

3.1.2 Convex Sets and Convex Functions

So what do convex functions have to do with convex sets? The first thing they have in common is that they're easy to deal with. However, the real reason is that you can equivalently define convex functions *in terms of* convex sets. To see this, let f be some function. Define the set $A = \{(x, y) \in \mathbb{R}^2 : f(x) \leq y\}$. In effect, this set A is the set of all points that *lie above* the plot of the function f. It can be shown that f is a convex function if and only if A is a convex set.

Exercise 27. *Prove the above claim. (This is kind of hard!)*

3.2 Wrap-up

The things you should know are:

- Gradients are multidimensional derivatives
- How to compute gradients
- Hessians are multidimensional second derivatives
- Convex sets are not potatoes
- Positive definite matrices are just like positive numbers
- Convex functions are those whose Hessians are psd

4 Probability & Statistics

PRML actually has a fairly good discussion of basic probability and statistics. See Section 1.2 (though but not including the part on "Bayesian probabilities") and Section **TODO...**.

The important things that you should really know are:

- The difference between computing the probability of some event and sampling from the associated distribution
- Joint and conditional distributions
- Marginal distributions
- Chain rule and Bayes' rule
- Standard distributions: Bernoulli, Binomial, Multinomial, Uniform and Gaussian
- Expectation, variance and standard deviation
- Covariance

One thing that always seems to trip students up who aren't used to these this is that it is *okay* for a continuous distribution to have density greater than one at some point.

Example 20. *Let p be a Gaussian distribution with zero mean and variance* 0.1*. Let's compute it's density at* 0*:*

$$\begin{aligned} p(0) &= \mathcal{N}or(0 \mid 0, 0.1) && (45)\\ &= \tfrac{1}{\sqrt{2\pi 0.1}} \exp\left[-\tfrac{1}{2(0.1)} 0^2\right] && (46)\\ &= \sqrt{\tfrac{1}{2\pi 0.1}} && (47)\\ &= \sqrt{\tfrac{1}{0.6283}} && (48)\\ &= \sqrt{1.5915} && (49)\\ &= 1.2615 && (50) \end{aligned}$$

This is very definitely above one! In fact, if we replace 0.1 *with* 0.01*, we get a value of almost* 4*!*

The thing to remember is that continuous probability densities are *not* like discrete densities. That is, for a continuous density, $p(0)$ is *not* the probability of drawing a zero from this distribution. If you think of a Gaussian, the probability of drawing any value at all is always zero! This is because you're drawing real numbers, and you're never going to draw *exactly* some given value (it will always differ in the 100th decimal place or something). The right way to interpret it is that the probability of drawing some value in the *range* $[a, b]$ is $\int_a^b \mathrm{d}x p(x)$. It is *this* value that should always be less than (or equal to) one.[5]

[5] Keep in mind that it's easy to have a function that takes values greater than one, but still integrates to one. Simply recall the example from Section 1.2 of $f(x) = 1/x^2$, which takes value 4 at $x = 0.5$, but still integrates to 1.