Vector Spaces

January 31, 2025

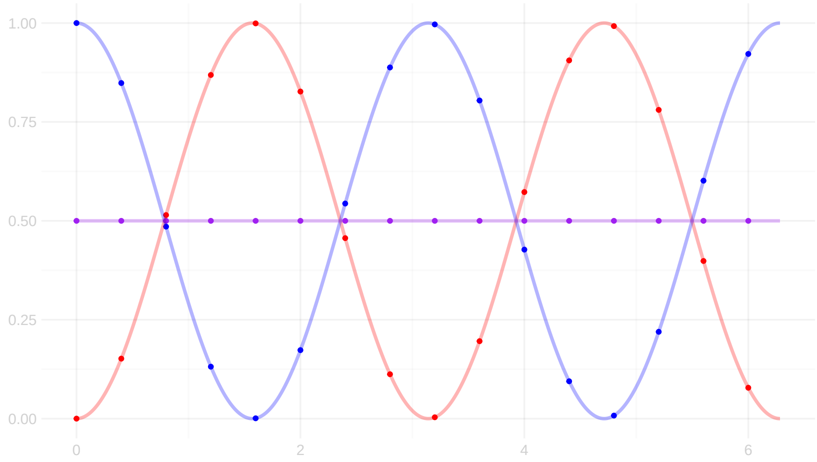
1 Introduction

Getting used to **thinking of functions as vectors**, conceptually and notationally, takes a bit of practice. But as we move forward, it'll be very useful. This won't be the most fun set of exercises we'll do, but it should pay off by making things smooth later on. And this is all standard notation, so it may be helpful in other contexts.

Throughout this problem set, we'll be thinking about a *vector space* $\mathcal{V}$. For our purposes, a vector space is a set of things that we can add together and multiply by scalars. Vectors spaces have a *zero element*. Here are the examples that'll be important for us.

- $\mathbb{R}$, the real numbers.
- $\mathbb{R}^n$, the n-dimensional vectors with real elements.
  - zero element: the zero vector $\vec{0} \in \mathbb{R}^n$
  - addition: for $x, y \in \mathbb{R}^n$, $x + y \in \mathbb{R}^n$
  - scalar multiplication: for $a \in \mathbb{R}, x \in \mathbb{R}^n$, $ax \in \mathbb{R}^n$
- Functions from some set to $\mathbb{R}$. We add and scale these *pointwise*
  - zero element: the function $f(x) = 0$ that's zero for all x.
  - addition: $f + g$ is a function with $(f + g)(x) = f(x) + g(x)$;
  - multiplication: for $\alpha \in \mathbb{R}$, αg is a function with $(\alpha f)(x) = \alpha f(x)$.

You may remember from high school the formula $\sin^2(x) + \cos^2(x) = 1$. You can think of this as a statement involving an addition of real numbers: for any x, $\sin(x)^2 + \cos(x)^2 = 1$. But you can also think of it as one involving an addition of functions, $\sin^2 + \cos^2 = 1$, which says that if you add the function $f(x) = \sin(x)^2$ and the function $g(x) = \cos(x)^2$, you get the constant function $h(x) = 1$. Below I've illustrated a version of this that's a bit easier to make sense of visually: the formula $(\sin^2 + \cos^2)/2 = 1/2$. Looking at a single dot illustrates addition of real numbers, looking at all the dots at once illustrates addition of vectors, and looking between the dots illustrates addition of functions.



Exercise 1 *Suppose we have a vector space $\mathcal{F}$ of differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}$. The set of* gradients *of these functions, which we might call $\nabla\mathcal{F}$, is $\{\nabla f \ : \ f \in \mathcal{F}\}$. Is this a vector space? If not, explain why. If so, explain how to add, subtract, and scale and describe the zero element.*

Solution 1 *It is a vector space. We use a vector version of the pointwise addition and scaling operations on functions. $(\nabla f + \nabla g)(x) = \nabla f(x) + \nabla g(x)$ and $\nabla(\alpha f)(x) = \alpha \nabla f(x)$. The zero, in this space, which we might write $0_{\nabla\mathcal{F}}$, is the function that maps x to the vector of zeros in $\mathbb{R}^n$. This is the gradient of the zero element $0_{\mathcal{F}}$ in our space of functions $\mathcal{F}$, which maps x to the scalar 0.*

2 Norms

A *seminorm* ρ on a vector space is a function that is *absolutely homogeneous* and satisfies a *triangle inequality*. That is, it's a function for which

$$\rho(\alpha v) = |\alpha|\rho(v) \quad \text{and} \quad \rho(u+v) \le \rho(u) + \rho(v).$$

Some seminorms are *norms*, which have the additional property that $\rho(v) = 0$ only if $v = 0$. We tend to write something like $\|v\|$ instead of $\rho(v)$ to indicate that we've got a norm and not a seminorm.[1] Here are some examples.

- On real numbers, i.e., vectors $v \in \mathbb{R}$, we have the magnitude $|v|$.
- On finite dimensional vectors $v \in \mathbb{R}^n$, there are a couple we use a lot.
  - $\|v\|_2 := \sqrt{\sum_{i=1}^n |v_i|^2}$, the two-norm.
  - $\|v\|_1 := \sum_{i=1}^n |v_i|$, the one-norm.
  - $\|v\|_\infty := \max_{i\in 1\ldots n} |v_i|$, the infinity norm.

We can, in fact, apply these to infinite sequences $v = v_1, v_2, v_3, \ldots$ as well. To do that, we just take $n = \infty$ above, i.e., we define $\|v\|_2 := \sqrt{\sum_{i=1}^{\infty} v_i^2}$, $\|v\|_1 := \sum_{i=1}^{\infty} |v_i|$, and $\|v\|_\infty := \max_{i \in 1 \ldots \infty} |v_i|$.

Exercise 2 *For the following vectors v, what are $\|v\|_1$, $\|v\|_2$, and $\|v\|_\infty$?*

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}.$$

Optional: what about for the infinite sequence $1/1, 1/2, 1/3, 1/4, 1/5, \ldots$?

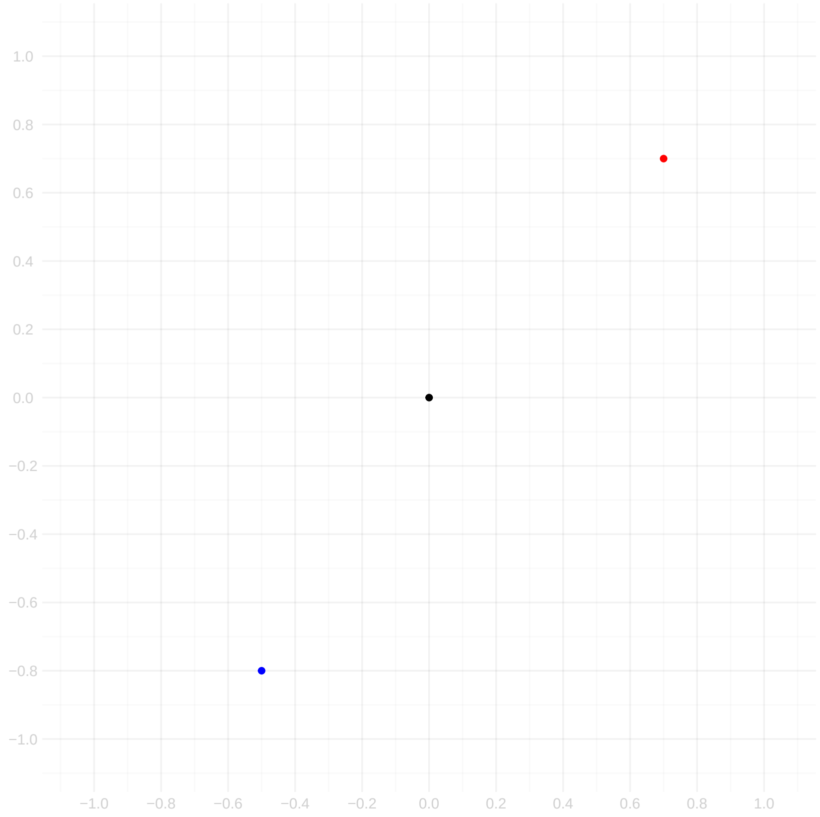
Solution 2 *Plug the vectors into the definitions. Here's what we get for the first vector.*

$$\|v\|_1 = |1|+|2|+|3| = 6,\ \|v\|_2 = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14},\ \|v\|_\infty = \max(|1|, |2|, |3|) = 3.$$

Here's what we get for the second.

$$\|v\|_1 = |2|+|2|+|2| = 6,\ \|v\|_2 = \sqrt{2^2 + 2^2 + 2^2} = \sqrt{12},\ \|v\|_\infty = \max(|2|, |2|, |2|) = 2.$$

The optional part involves some math trivia. The infinity norm of the infinite sequence $1/1, 1/2, \ldots$ is easy to work out: it's one, as the sequence is decreasing from one to zero. The two norm is $\sqrt{\pi^2/6}$: it's the square root of the sum $\sum_{i=1}^{\infty} 1/i^2$, which Euler calculated in the 1700s using the Taylor expansion of the sine. See wikipedia if you're interested. And the one norm is infinity: it's the sum of the harmonic series, $\sum_{i=1}^{\infty} 1/i$. You can read about this on wikipedia too.



Exercise 3 *It can be helpful to think about norms in physical terms. Imagine you're in a city laid out like a grid, like Manhattan. I've drawn a grid for you above to help you visualize what's going on. You're standing on one street corner* $a = (x_a, y_a)$ *and going to another corner* $b = (x_b, y_b)$. *Here, in some order, are interpretations of the one, two, and infinity norms* $\|a-b\|_1$, $\|a-b\|_2$, *and* $\|a-b\|_\infty$.

a. *The distance between a and b as the crow flies.*

b. *The distance you'd have to walk to get from a and b, i.e., distance when you can only move along the grid lines.*

c. *The longest distance you could walk along any one grid line (street) without going too far in some direction.*

Match each description to a norm.

Solution 3

a. *The distance between a and b as the crow flies is the two norm* $\|a-b\|_2$

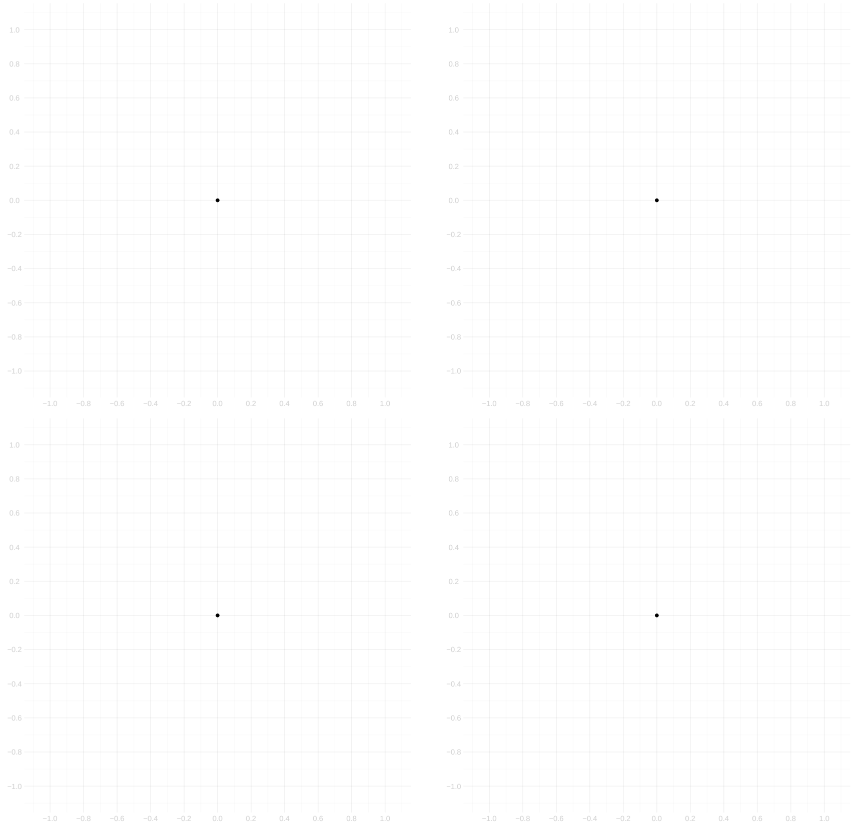
b. *The distance you'd have to walk to get from a and b, i.e., distance when you can only move along the grid lines, is the one norm* $\|a - b\|_1$

c. *The longest distance you could walk along any one grid line (street) without going too far in some direction is the infinity norm* $\|a - b\|_\infty$

Exercise 4 *If we have a norm* $\|\cdot\|$, *we say the set* $\mathcal{B} = \{x \ : \ \|x\| \leq 1\}$ *is its* unit ball. *For each of the norms* $\|\cdot\|_1$, $\|\cdot\|_2$, *and* $\|\cdot\|_\infty$ *on* $\mathbb{R}^2$, *draw its unit ball.*

Use the grids below. I've given you four grids, so you have one to spare. Use that one to draw all 3 on top of one another.

Tip. *To get started, work out where the boundary of the unit ball hits the x and y axes and then the diagonal lines* $y = \pm x$. *Then make up a few more lines and think about where they hit the ball's boundary, too.*

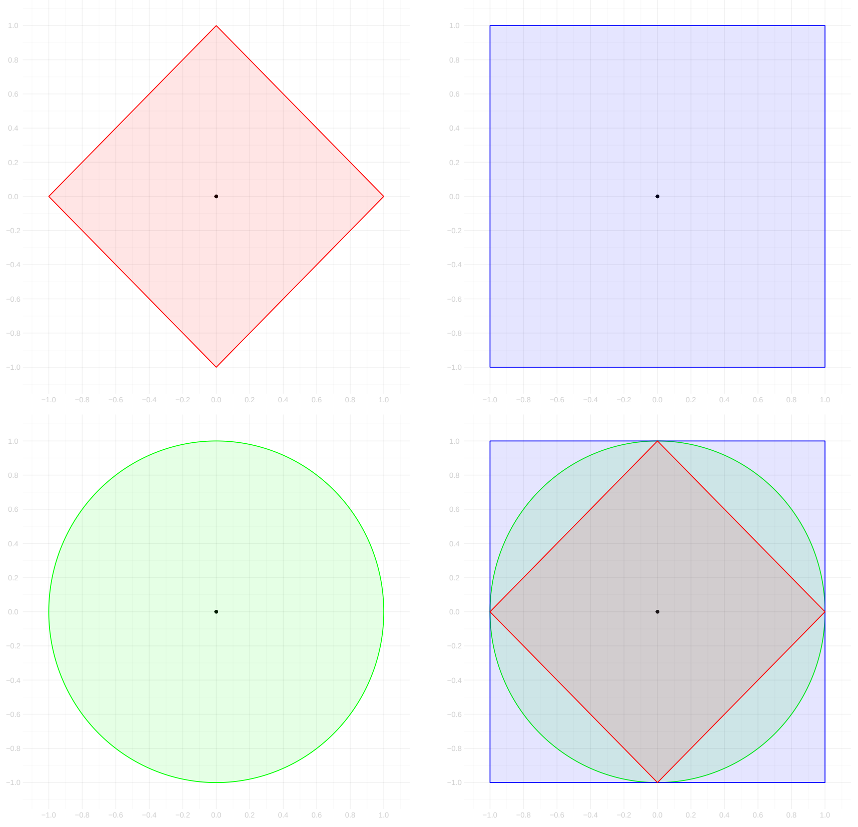


Solution 4

- *The ball for the one norm is a diamond with vertices at* $(1, 0)$, $(-1, 0)$, $(0, 1)$, *and* $(0, -1)$. *It's drawn in red.*

- *The ball for the two norm is a circle with radius one centered at the origin. It's drawn in green.*

- *The ball for the infinity norm is a square with vertices at* $(1, 1)$*,* $(-1, 1)$*,* $(-1, -1)$*, and* $(1, -1)$*. It's drawn in blue.*



2.1 Norms for Functions

For vector spaces of functions, v we tend to use analogous definitions, replacing sums with integrals. For example, for functions from the interval $[0, 1]$ to $\mathbb{R}$, we use these.

- $\|v\|_{L_2} := \sqrt{\int_0^1 |v(x)|^2}$, the two-norm.

- $\|v\|_{L_1} := \int_0^1 |v(x)|$, the one-norm.

More generally, for functions from any set $\mathcal{X}$ to $\mathbb{R}$, we tend to define these analogously using integrals over *probability distributions* on that set. That is, if P is the probability distribution of some random variable $X \in \mathcal{X}$, then

- $\|v\|_{L_2(\mathrm{P})} := \sqrt{\mathbb{E}\left[v(X)^2\right]}$, the population two-norm.
- $\|v\|_{L_1(\mathrm{P})} := \mathrm{E}\left[|v(X)|\right]$, the population one-norm.

Exercise 5 *Our definitions of $\|v\|_{L_2}$ and $\|v\|_{L_1}$ correspond to the case that* P *is the uniform distribution on* $[0,1]$. *Explain why.*

Solution 5 *Let's do the one-norm. For any continuous-valued random variable X, $\mathrm{E}[|v(X)|] = \int |v(x)| f(x) dx$ where $f(x)$ is the probability density of X. In the case of the uniform distribution on* $[0,1]$,

$$f(x) = \begin{cases} 1 & \textit{if } x \in [0,1] \\ 0 & \textit{otherwise} \end{cases},$$

so the latter integral is $\int_0^1 |v(x)| dx$. We can do the two-norm by replacing $|v(x)|$ with $v(x)^2$.

In the exercises below, we work with the following functions.

$$\begin{aligned} v_1(x) &= x^2 \\ v_2(x) &= \begin{cases} 1 & \text{if} \quad x = 0 \\ 0 & \text{otherwise} \end{cases} \\ v_3(x) &= e^x \end{aligned} \tag{1}$$

Exercise 6 *For each of these functions v, suppose they are defined from* $[0,1]$ *to* $\mathbb{R}$*, what are $\|v\|_{L_1}$ and $\|v\|_{L_2}$? And when P is the standard normal distribution, i.e. the distribution of random variable X that is normal with mean zero and variance one, what are $\|v\|_{L_1(P)}$ and $\|v\|_{L_2(P)}$? What if P is the distribution of a random variable X that is normal with mean one and variance one?*

If v_3 is giving you trouble, don't worry about it; just do v_1 and v_2.

Hint In your calculations, you'll need the values of the *moments* EX^k and the so-called *moment generating function* Ee^{tX} for normal random variables. For a normal random variable X with mean μ and variance σ^2,

$$\begin{aligned} E\left[e^{tX}\right] &= e^{\mu t + \sigma^2 t^2/2} \\ E\left[X\right] &= \mu \\ E\left[X^2\right] &= \mu^2 + \sigma^2 \\ E\left[X^3\right] &= \mu^3 + 3\mu\sigma^3 \\ E\left[X^4\right] &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4. \end{aligned}$$

I got this from the wikipedia article on the normal distribution, where you'll find a lot more information, none of which you'll need in this assignment. You might find it faster to calculate all the norms associated with one distribution and then move on to the second distribution.

Solution 6 *Let's start with v_1. For the calculations involving normally distributed random variables, we can use the table above, with $\mu = 0$ and $\sigma = 1$ in the first case and $\mu = 1$ and $\sigma = 1$ in the second.*

$$\begin{aligned}
\|v_1\|_{L_1} &= \int_0^1 |x^2| = \int_0^1 x^2 = x^3/3 \mid_0^1 = 1/3. \\
\|v_1\|_{L_2} &= \sqrt{\int_0^1 (x^2)^2} = \sqrt{x^5/5 \mid_0^1} = 1/\sqrt{5}. \\
\|v_1\|_{L_1(P)} &= \mathrm{E}[|X^2|] = \mathrm{E}[X^2] = 1 && \textit{for } X \sim N(0,1). \\
\|v_1\|_{L_2(P)} &= \sqrt{\mathrm{E}[X^4]} = \sqrt{3} && \textit{for } X \sim N(0,1). \\
\|v_1\|_{L_1(P)} &= \mathrm{E}[|X^2|] = \mathrm{E}[X^2] = 2 && \textit{for } X \sim N(1,1). \\
\|v_1\|_{L_2(P)} &= \sqrt{\mathrm{E}[X^4]} = \sqrt{1+6+3} = \sqrt{10} && \textit{for } X \sim N(1,1).
\end{aligned}$$

For v_2, all these norms are zero. That's because it's zero except at a single point and integrals aren't affected by the value of a function at a single point.

Now let's do v_3. For the calculations involving normally distributed random variables, we'll use its moment generating function from the table above, with $\mu = 0$ and $\sigma = 1$ in the first case and $\mu = 1$ and $\sigma = 1$ in the second.

$$\begin{aligned}
\|v_3\|_{L_1} &= \int_0^1 e^x = e^x \mid_0^1 = e - 1. \\
\|v_3\|_{L_2} &= \sqrt{\int_0^1 e^{2x}} = \sqrt{e^{2x}/2 \mid_0^1} = \sqrt{e^2/2 - 1/2} \\
\|v_3\|_{L_1(P)} &= \mathrm{E}[e^X] = e^{1/2} && \textit{for } X \sim N(0,1). \\
\|v_3\|_{L_2(P)} &= \sqrt{\mathrm{E}[e^{2X}]} = \sqrt{e^{2^2/2}} = e && \textit{for } X \sim N(0,1). \\
\|v_3\|_{L_1(P)} &= \mathrm{E}[e^X] = e^{1+1/2} = e^{3/2} && \textit{for } X \sim N(1,1). \\
\|v_3\|_{L_2(P)} &= \sqrt{\mathrm{E}[e^{2X}]} = \sqrt{e^{2+4/2}} = e^2 && \textit{for } X \sim N(1,1).
\end{aligned}$$

For good measure, here are two more seminorms we see a fair amount.

- $\mathrm{sd}_P(v) := \sqrt{\mathrm{E}[(v(X) - \mathrm{E}[v(X)])^2]}$, the population standard deviation.
- On differentiable functions $v(x)$ on $[0,1]$, the *total variation* $\rho_{TV}(v) = \int_0^1 |v'(x)| dx$.

Exercise 7 *For the functions v_1,v_2, and v_3 from Exercise 6, what is $\mathrm{sd}_P(v)$ when P is the standard normal distribution? What about when P is the distribution of a random variable X that is normal with mean one and variance one? And what is $\rho_{TV}(v)$?*

This last question doesn't totally make sense for v_2, as it isn't differentiable. But say what you think it should be anyway and briefly explain, in terms you'll be able to understand when you read it a few weeks from now, why you said what you said. We'll address this in our lecture on bounded variation regression, when I'll give a more general definition of the seminorm ρ_{TV} that will apply to v_2, and we'll talk about why that definition is what it is and how to think about it.

Solution 7 *It's convenient to work with another formula for the variance. Here it is.*

$$\begin{aligned}
\mathrm{E}[\{v(X) - \mathrm{E}[v(X)]\}^2] &= \mathrm{E}[v(X)^2] - \mathrm{E}[2v(X)\,\mathrm{E}[v(X)]] + \mathrm{E}[v(X)]^2 \\
&= \mathrm{E}[v(X)^2] - 2\,\mathrm{E}[v(X)]^2 + \mathrm{E}[v(X)]^2 \\
&= \mathrm{E}[v(X)^2] - \mathrm{E}[v(X)]^2.
\end{aligned}$$

That puts us in a good position to use our table of moments and the moment generating function.

Let's start with sd_P *where* P *is the distribution of a random variable* X *that's normal with mean zero and variance one.*

$$\begin{aligned}
&\mathrm{sd}_P(v_1) = \sqrt{\mathrm{E}[X^4] - \mathrm{E}[X^2]^2} = \sqrt{3-1} = \sqrt{2} \\
&\mathrm{sd}_P(v_2) = 0 \quad \textit{for the same reason as before} \\
&\mathrm{sd}_P(v_3) = \sqrt{\mathrm{E}[e^{2X}] - \mathrm{E}[e^X]^2} = \sqrt{e^2 - (e^{1/2})^2} = \sqrt{e^2 - e}.
\end{aligned}$$

Now let's do sd_P *where* P *is the distribution of a random variable* X *that's normal with mean one and variance one.*

$$\begin{aligned}
&\mathrm{sd}_P(v_1) = \sqrt{\mathrm{E}[X^4] - \mathrm{E}[X^2]^2} = \sqrt{10 - 2^2} = \sqrt{6} \\
&\mathrm{sd}_P(v_2) = 0 \quad \textit{for the same reason as before} \\
&\mathrm{sd}_P(v_3) = \sqrt{\mathrm{E}[e^{2X}] - \mathrm{E}[e^X]^2} = \sqrt{e^{2+4/2} - (e^{1+1/2})^2} = \sqrt{e^4 - e^3}.
\end{aligned}$$

Now let's do total variation.

$$\begin{aligned}
\rho_{TV}(v_1) &= \int_0^1 (x^2)' = \int_0^1 2x = x^2 \mid_0^1 = 1 \\
\rho_{TV}(v_2) &= 1 \\
\rho_{TV}(v_3) &= \int_0^1 (e^x)' = \int_0^1 e^x = e^x \mid_0^1 = e - 1.
\end{aligned}$$

I won't explain my answer for v_2 *here, but it should become clear in my lecture on total variation.*

Optional reading The corresponding generalization of the infinity norm is a little trickier and we won't really need it. In case you're interested, I put it in the Appendix A, along with a few related exercises. This may be a bit much if you haven't had real analysis, which is by no means required for you to understand what we're going to do in this class.

2.2 Norms associated with samples

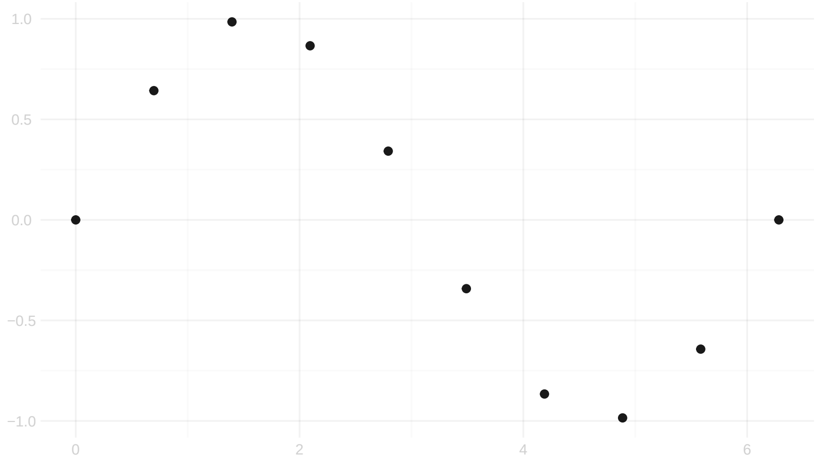
When we're working with a sample $X_1 \dots X_n$, sometimes we abuse notation by writing $\|v\|_2$ for a function v, meaning $\sqrt{\sum_{i=1}^n |v(X_i)|^2}$. When we do this, we're interpreting v as the vector $[v(X_1) \dots v(X_n)]$ of values it takes on the sample. We can do the same with the one and infinity norms. Up to a scale factor, we can also think of these as norms associated with the *empirical distribution* $\mathrm{P_n}$: the distribution of a random variable X that takes on each value $X_1 \dots X_n$ with probability $1/n$.

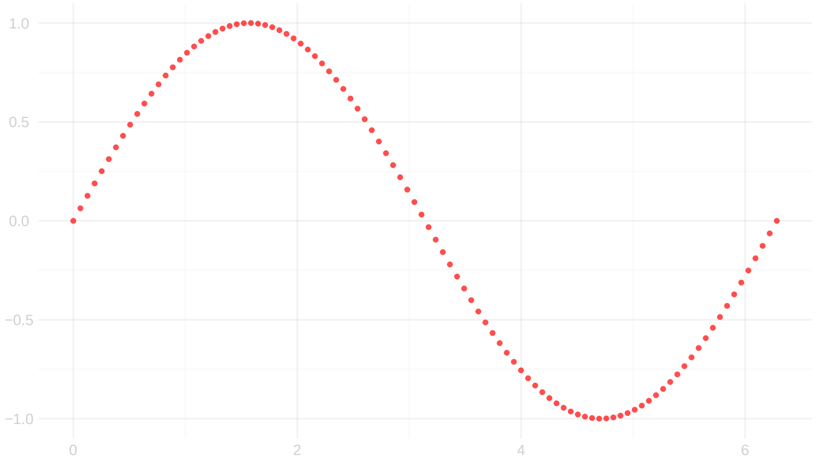
$$\begin{aligned}
\|v\|_{L_2(\mathrm{P_n})} &= \sqrt{\frac{1}{n}\sum_{i=1}^n |v(X_i)|^2} = \frac{\|v\|_2}{\sqrt{n}} && \text{the sample two-norm} \\
\|v\|_{L_1(\mathrm{P_n})} &= \frac{1}{n}\sum_{i=1}^n |v(X_i)| = \frac{\|v\|_1}{n}, && \text{the sample one-norm} \\
\|v\|_{L_\infty(\mathrm{P_n})} &= \max_{i \le n} |v(X_i)| = \|v\|_\infty && \text{the sample infinity norm.}
\end{aligned}$$

Similarly, the sample standard deviation is the population standard deviation associated with the empirical distribution.

The advantage of these norms based on the empirical distribution, relative to the analogous vector norm, is that they don't tend to vary much with sample size. For example, if we have a function $v(x)$ and a sample $X_1 \dots X_n$, then $\|v\|_1$ will be a sum $|v(X_1)| + \dots + |v(X_n)|$ of n values of $|v(x)|$ and therefore tends to be roughly proportional to n, whereas $\|v\|_{L_1(\mathrm{P_n})}$ is the average of these n values and therefore doesn't tend to grow with n. Same deal with $\|v\|_2^2$ and $\|v\|_{L_2(\mathrm{P_n})}^2$; the first is the sum of n values of $|v(x)|^2$ and the second is the average of them.

Below we see two sampled versions of the function $f(x) = \sin x$, one of size 10 and the other of size 100. The vector norm $\|v\|_1$ for is 5.671 for sample size $n = 10$ and 63.02 for sample size $n = 100$ On the other hand, the sample norm $\|v\|_{L_1(\mathrm{P_n})}$ is 0.567 for sample size $n = 10$ and 0.63 for sample size $n = 100$.





If we're thinking of $X_1 \dots X_n$ as a random sample, then these are *random norms*, and it makes sense to talk about the probability distribution of the norms $\|v\|_{L_2(\mathrm{P_n})}$, $\|v\|_{L_2(\mathrm{P_n})}$, and $\|v\|_{L_\infty(\mathrm{P_n})}$ for a function v. In particular, if each observation X_i is an independent draw from the distribution P, then we can relate them to the corresponding population norms. Let's do that.

Exercise 8 *Show the following.*

1. $\mathrm{E}\left[\|v\|_{L_1(\mathrm{P_n})}\right] = \|v\|_{L_1(\mathrm{P})}$.
2. $\mathrm{E}\left[\|v\|^2_{L_2(\mathrm{P_n})}\right] = \|v\|^2_{L_2(\mathrm{P})}$

Solution 8 *We just write things out using the definitions. Reordering our expectation and our sample average, we see that in both cases what we're doing is averaging a constant: the corresponding population norm or squared norm.*

$$\mathrm{E}\|v\|_{L_1(\mathrm{P_n})} = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}|v(X_i)|\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[|v(X_i)|] = \frac{1}{n}\sum_{i=1}^{n}\|v\|_{L_1(P)} = \|v\|_{L_1(P)}$$

$$\mathrm{E}\|v\|^2_{L_2(\mathrm{P_n})} = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}|v(X_i)|^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[|v(X_i)|^2] = \frac{1}{n}\sum_{i=1}^{n}\|v\|^2_{L_2(P)} = \|v\|^2_{L_2(P)}$$

Optional. There's an analogous exercise for the infinity norm in the appendix. If you want to explore infinity norms a bit more, it a shot. But don't feel uneasy skipping it. It's harder than the others and we won't need the result.

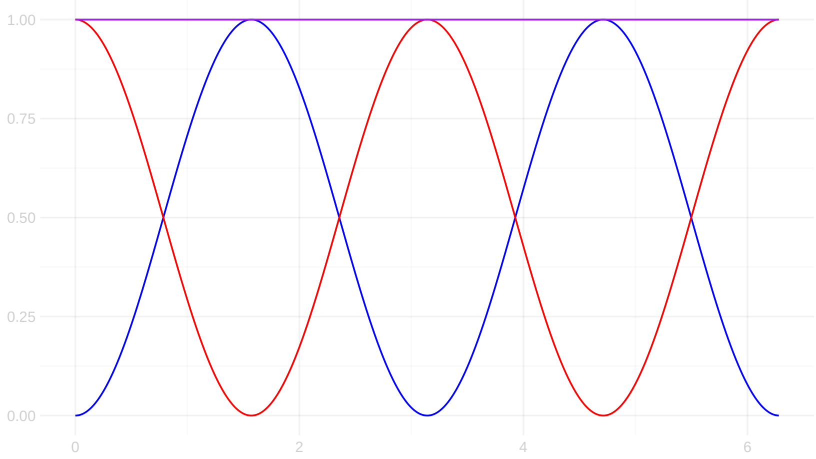
2.3 Checking that our examples are seminorms

Exercise 9 *Show that the population one-norm, the sample infinity-norm, and the total variation are seminorms. That is, show that they are absolutely homogeneous and satisfy a triangle inequality. Explain why this implies that the one-norm and infinity-norm on finite-dimensional vectors are also seminorms.*

You may assume that the magnitude (absolute value) is a seminorm on $\mathbb{R}$. Next week, we'll prove it.

It may be helpful to know that if the function u is always smaller than the function v, then the average of $u(X)$ will be smaller than $v(X)$ no matter what the distribution of X is, i.e., if $u(x) \leq v(x)$ for all x, then $\mathrm{E}[u(X)] \leq \mathrm{E}[v(X)]$ for all random variables X.

Hint You will need to think about the relationship between the maximum of two functions and their maximum of their sum. It may help to think about where that sum is maximized, i.e., the x at which $f(x) + g(x)$ is largest, and where the individual functions are maximized. Take a look at the graph below, in which the purple curve is the sum of the red and blue curves.



Solution 9 *Absolute homogeneity is essentially the same in each case. Because the absolute value is absolutely homogeneous by definition and expectations are linear in the sense that $\mathrm{E}\, aY = a\, \mathrm{E}\, Y$ and similarly for integration and differentiation,*

$$\|\alpha v\|_{L_1(\mathrm{P})} = \mathrm{E}[|\alpha v(X)|] = \mathrm{E}[|\alpha||v(X)|] = |\alpha|\, \mathrm{E}[|v(X)|] = |\alpha|\|v\|_{L_1(\mathrm{P})}.$$

$$\|\alpha v\|_{L_\infty(\mathrm{P_n})} = \max_{i \leq n} |\alpha v(X_i)| = \max_{i \leq n} |\alpha||v(X_i)|.$$

$$\rho_{TV}(\alpha v) = \int_0^1 |(\alpha v)'(x)| = \int_0^1 |\alpha v'(x)| = \int_0^1 |\alpha||v'(x)| = |\alpha|\rho_{TV}(v).$$

The triangle inequalities for the one norm and ρ_{TV} are essentially similar, relying on linearity and the triangle inequality for the magnitude.

$$\|u + v\|_{L_1(\mathrm{P})} = \mathrm{E}\left[|u(X) + v(X)|\right] \leq \mathrm{E}\left[|u(X)| + |v(X)|\right] = \|u\|_{L_1(\mathrm{P})} + \|v\|_{L_1(\mathrm{P})}.$$

$$\rho_{TV}(u + v) = \int_0^1 |(u + v)'(x)| = \int_0^1 |u'(x) + v'(x)| \leq \int_0^1 |u'(x)| + |v'(x)| = \rho_{TV}(u) + \rho_{TV}(v).$$

The triangle inequality for the infinity norm involves an extra step.

$$\begin{aligned}
\|u+v\|_{L_\infty(\mathrm{P_n})} &= \max_{i\le n} |u(X_i)+v(X_i)| \\
&\le \max_{i\le n} |u(X_i)| + |v(X_i)| \\
&\overset{\star}{\le} \max_{i,j:i\le n, j\le n} |u(X_i)| + |v(X_j)| \\
&= \|u\|_{L_\infty(\mathrm{P_n})} + \|v\|_{L_\infty(\mathrm{P_n})}.
\end{aligned}$$

That extra step, indicated with a star, is that in $\|u\|_{L_\infty(\mathrm{P_n})} + \|v\|_{L_\infty(\mathrm{P_n})}$ *we maximize the same quantity,* $|u(X_i)| + |v(X_j)|$*, that we were maximizing in the line above, but over a set of pairs of observations* $\{(i,j) : i \le n, j \le n\}$ *that includes the set* $\{(i,i) : i \le n\}$ *considered in the line above.*

To tackle the one-norm on finite-dimensional vectors, observe that the sample one norm is an instance of the sample population norm for $\mathrm{P} = \mathrm{P_n}$*, the empirical distribution. Thus, our work above implies the sample one-norm is a seminorm. And given any vector* $\vec{v}$*, we can find a function* $v(x)$ *for which* $\vec{v}_i = v(X_i)$*, and the one-norm of the vector will be* n *times the sample one-norm of any such function. Thus, the vector one-norm is homogenous and satisfies a triangle inequality if the sample one-norm does. The same goes for the vector and sample infinity norms, except there's no factor of* n*:* $\|\vec{v}\|_\infty = \|v\|_{L_2(\mathrm{P_n})}$*.*

2.4 Properties of Seminorms

2.4.1 Zero at Zero.

Seminorms are zero at zero, i.e., they satisfy $\rho(0) = 0$.

Exercise 10 *Prove it. If it takes more than one sentence, you're doing it wrong.*

Solution 10 *This follows from absolute homogeneity. For any* v *with* $\rho(v) < \infty$*,*

$$\rho(0) = \rho(0v) = |0|\rho(v) = 0.$$

2.4.2 Nonnegativity.

Seminorms are non-negative.

Exercise 11 *Prove it. This one shouldn't be much longer.*

Solution 11 *For any vector* v*, we can write* 0 *as* $v-v$*. Then, using the triangle inequality and absolute homogeneity, it follows that*

$$0 = \rho(0) = \rho(v-v) \le \rho(v) + \rho(-v) = |1|\rho(v) + |-1|\rho(v) = 2\rho(v).$$

2.4.3 Seminorms that aren't norms.

The population standard deviation and total variation are seminorms, but they are not norms.

Exercise 12 *Explain why.*

Hint *By definition, the norm of a function is zero if and only if the function is the zero element, which is the function that's always zero* $f(x) = 0$, $\forall x \in \mathbb{R}$. *Can you think of other functions whose population standard deviation or total variation is zero?*

Solution 12 *They measure variation: they're both zero for constant functions.*

A Generalization of the infinity norm

Here's your optional reading and exercises on the infinity norm. We could, of course, take $\|v\|_{L_\infty} = \max_{x \in [0,1]} |v(x)|$. But that would lead to a problem. Consider the function v_2 we've been working with: the discontinuous function v that's zero except at $x = 0$, where it's one. Using this definition, we get $\|v\|_{L_\infty} = 1$. But if we look at the L_1 and L_2 norms, or any other norms defined in terms of integrals, we'll get zero. Integrals are about the *area* under the curve, so they don't care what v looks like on a set with zero area, like a single point. What we want is something that's like the maximum, but doesn't care about that either. Here, in the general case, is what we wind up with. If P is the probability distribution of some random variable $\mathcal{X}$, then

$$\|v\|_{L_\infty(P)} := \inf \left\{ x \geq 0 : P(|v(X)| \leq x) = 1 \right\}.$$

And we define $\|v\|_{L_\infty}$ this way taking P to be the uniform distribution on $[0,1]$. Informally, this is the largest value of $|v(X)|$ that might actually occur when X is a random variable with distribution P. And it's smaller than the largest value outright, $\max_x |v(x)|$, so often even when being formal, you often won't need to think about the subleties. If you're not comfortable with what inf means, don't worry about this stuff, and either skip the following exercises or take a guess at them using the informal definition.

This one is a version of Exercise 6.

Exercise 13 ***(Optional).*** *For the functions v_1,v_2,and v_3 from Exercise 6, calculate $\|v\|_{L_\infty}$ and $\|v\|_{L_\infty(P)}$ where P is the standard normal distribution.*

Solution 13 *The reason we have this definition is so that functions that are zero except on a set of probability zero, like v_2, have an infinity norm of zero. That means $\|v_2\|_{L_\infty} = \|v_2\|_{L_\infty(P)} = 0$. For our other two functions, we can get away with thinking of our infinity norms as just ordinary maxima of $|v(x)|$, over $x \in [0,1]$ for $\|v\|_{L_\infty}$ and over the whole real line for $\|v\|_{L_\infty(P)}$. Assuming we can do this, we'd have $\|v_1\|_{L_\infty} = \max_{x \in [0,1]} |x^2| = 1$ and $\|v_3\|_{L_\infty} = \max_{x \in [0,1]} |e^x| = e$ and $\|v_1\|_{L_\infty(P)} = \max_{x \in \mathbb{R}} |x^2| = \infty$ and $\|v_3\|_{L_\infty(P)} = \max_{x \in \mathbb{R}} |e^2| = \infty$.*

Now let's talk about why we can do this. The maximum absolute values of the functions v_1 and v_3 occur at $x = 1$, and both functions get arbitrarily close to this value as we approach 1. This means that no matter how close we want to get to $|v(1)|$, there are sets of probability $\epsilon > 0$ under the uniform distribution, e.g. sets taking the form $[1 - \epsilon, 1]$, on which $|v(x)|$ is that close to it. Thus, the least probability-one upper bound on $|v(X)|$ will just be that maximum $|v(1)|$. The case of normally distributed X is analogous, except that $|v(x)|$ is getting arbitrarily large as x gets arbitrarily large.

This one is a version of Exercise 8

Exercise 14 ***(Optional).*** *Show that $\|v\|_{L_\infty(\mathrm{P_n})} \leq \|v\|_{L_\infty(P)}$ with probability one.*

Here's a hint. It's equivalent to say that the probability that $\|v\|_{L_\infty(\mathrm{P_n})} > \|v\|_{L_\infty(P)}$ is zero. And the probability that $|v(X_i)| > \|v\|_{L_\infty(P)}$ *for any* i in $1 \ldots n$ is no larger than the sum of the probabilities that $|v(X_i)| > \|v\|_{L_\infty(P)}$ for all i in $1 \ldots n$. That's a consequence of the Union Bound.

Solution 14 *Let's reframe this to focus on events of probability zero instead of probability one. It's either the case that $|v(X)| \le x$ or $|v(X)| > x$, so the probability that $|v(X)| \le x$ is one if and only if the probability that $|v(X)| > x$ is zero. Thus, we can equivalently say $\|v\|_{L_\infty(P)} = \inf\left\{x \ge 0 : P(|v(X)| > x) = 0\right\}$. Now let's take a look at the sample infinity norm.*

$\|v\|_{L_\infty(\mathrm{P_n})} = \max_{i \le n}|v(X_i)|$, *and the probability that this maximum exceeds any given bound x is the probability that $v(X_i) > x$ for any i, i.e., it's the probability of the* union of events $P(|v(X_1)| > x \quad or \quad |v(X_2)| > x \quad or \ \ldots)$. *The union bound tells us that this probability is less than the* sum *of the probabilities of the individual events, i.e., less than $\sum_{i=1}^n P(|v(X_i)| > x) = nP(|v(X_1)| > x)$. Thus, the probability that $\|v\|_{L_\infty(\mathrm{P_n})} > x$ is zero if the probability that $|v(X_1)| > x$ is zero, i.e., the probability that $\|v\|_{L_\infty(\mathrm{P_n})} > x$ is zero if $x > \|v\|_{L_\infty(P)}$.*

Notes

[1] Don't worry too much about whether a seminorm is a norm. In many vector spaces, there are many vectors that are almost zero, and some seminorms ρ that we tend to think of as norms aren't because $\rho(v) = 0$ for these almost-zero vectors. We still tend to write $\|v\|$ instead of $\rho(v)$ in this case.