Week 2 Homework: Vector Spaces

QTM 490R: Machine Learning Theory

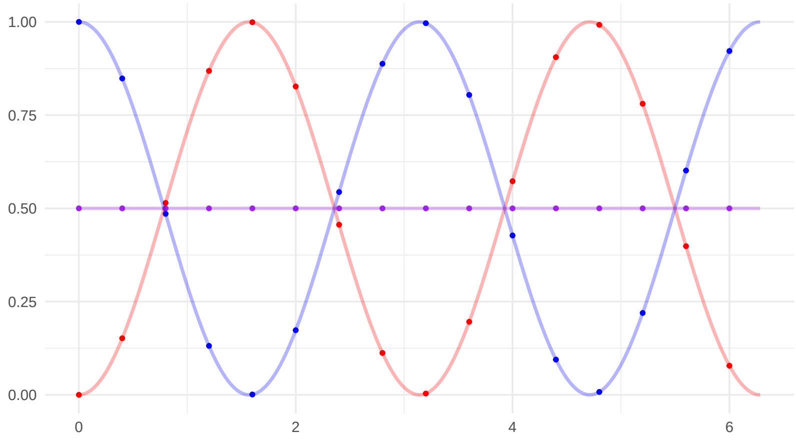
1 Introduction

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 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.*

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 $\|v\|_\infty := \max_{i \in 1 \ldots \infty} |v_i|$.

[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.

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 $1/1, 1/2, 1/3, 1/4, 1/5, \ldots$*?*

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 3 *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.*

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 4 *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.

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 5 *For the functions v_1,v_2, and v_3 from Exercise 4, 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.

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 $X_1 \ldots X_n$, 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 $[v(X_1) \ldots v(X_n)]$ of values it takes on the sample.

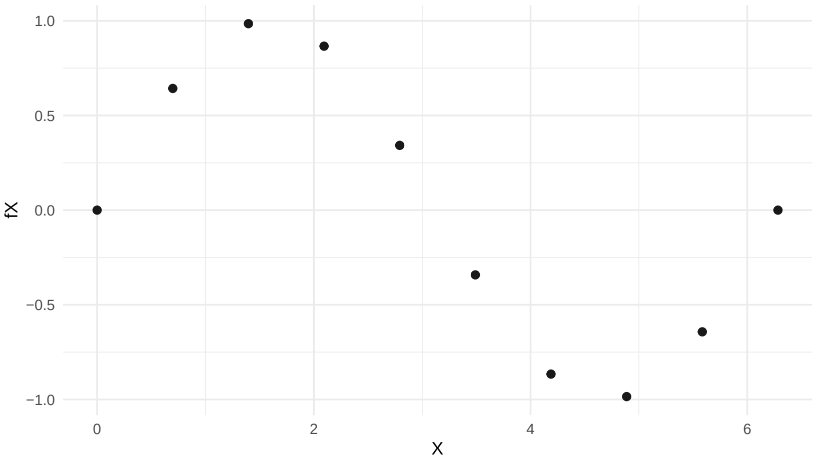
We can do the same with the one and infinity norms. Up to a scale factor, we can also think of these as norms associated with the *empirical distribution* $\mathrm{P_n}$: the distribution of a random variable X that takes on each value $X_1 \ldots X_n$ with probability $1/n$.

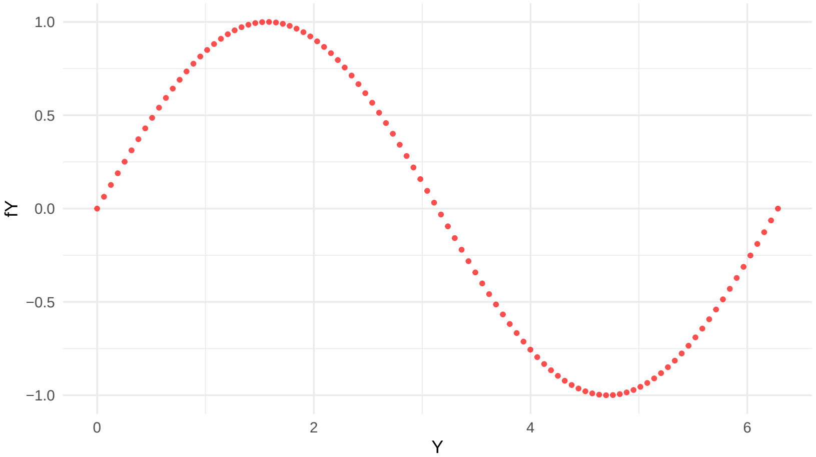
$$\begin{aligned}
\|v\|_{L_2(\mathrm{P_n})} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} v(X_i)^2} = \frac{\|v\|_2}{\sqrt{n}} && \text{the sample two-norm} \\
\|v\|_{L_1(\mathrm{P_n})} &= \frac{1}{n}\sum_{i=1}^{n} |v(X_i)| = \frac{\|v\|_1}{n}, && \text{the sample one-norm} \\
\|v\|_{L_\infty(\mathrm{P_n})} &=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

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 $X_1 \ldots X_n$, then $\|v\|_1$ will be a sum $|v(X_1)| + \ldots + |v(X_n)|$ 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 6 *Show the following.*

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

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

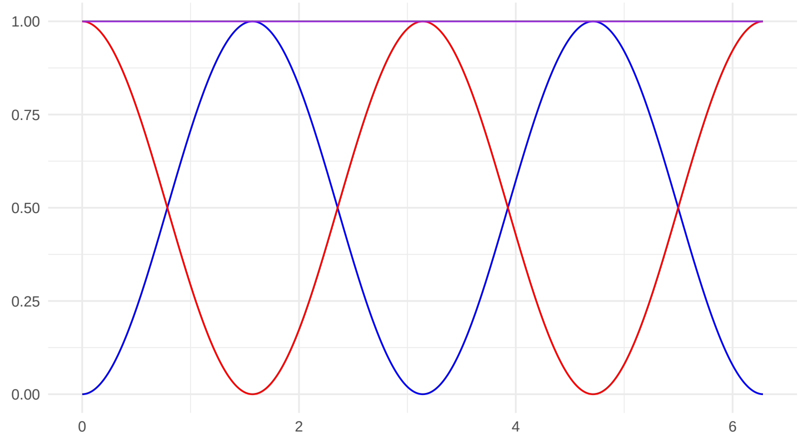
Optional. There's an analogous exercise for the infinity norm in the appendix. If you want to explore infinity norms a bit more, it a shot. But don't feel uneasy skipping it. It's harder than the others and we won't need the result.

2.3 Checking that our examples are seminorms

Exercise 7 *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 is a seminorm on complex numbers. In the next section, 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 $u(x) \le v(x)$ for all x, then $\mathrm{E}[u(X)] \le \mathrm{E}[v(X)]$ 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.



2.4 Properties of Seminorms

2.4.1 Zero at Zero.

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

Exercise 8 *Prove it. If it takes more than one sentence, you're doing it wrong.*

2.4.2 Nonnegativity.

Seminorms are non-negative.

Exercise 9 *Prove it. This one shouldn't be much longer.*

2.4.3 Seminorms that aren't norms.

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

Exercise 10 *Explain why.*

Hint *By definition, the norm of a funciton 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?*

3 Inner Products

A semi-inner-product $\langle u, v\rangle$ is a function of two vectors u, v that is *symmetric*, *linear* in its arguments, and *positive.* That is, for all vectors u, v, w and scalars α,

$$\langle u, v\rangle = \langle v, u\rangle, \quad \langle u + \alpha v, w\rangle = \langle u, w\rangle + \alpha\langle v, w\rangle, \quad \text{and} \quad \langle u, u\rangle \geq 0.$$

An inner product is a semi-inner-product that is *positive definite*, i.e., that satisfies $\langle u, u\rangle = 0$ if and only if $u = 0$. We tend to talk more about inner products than semi-inner products, but there are a few semi-inner products we use often that aren't positive-definite.

Here are some examples of semi-inner products.

- For real scalars, we have the product $\langle u, v\rangle = uv$.
- On finite dimensional vectors $v \in \mathbb{R}^n$, we have the dot product, $\langle u, v\rangle_2 := \sum_{i=1}^n u_i v_i = u^T v$.
- On functions $v(x)$, in terms of a random variable X with distribution P, we have the population inner product $\langle u, v\rangle_{L_2(P)} = \mathrm{E}[u(X)v(X)]$ and the covariance $\mathrm{Cov}_{\mathrm{P}}(u, v) = \mathrm{E}\left[\{u(X) - \mathrm{E}[u(X)]\}\{v(X) - \mathrm{E}[v(X)]\}\right]$.

Just like with seminorms, sometimes when working with a sample $X_1 \ldots X_n$, we thinking of functions as vectors: for functions u and v, $\langle u, v\rangle_2 = \sum_{i=1}^n u(X_i)v(X_i)$. And, as before, this is just a scaled version of the population inner product for the empirical distribution $\mathrm{P_n}$.

Exercise 11 *Prove that these examples are semi-inner-products.*

For each of these, there is an associated seminorm $\rho(v) = \sqrt{\langle v, v\rangle}$ included in our examples in Section 2. To work out which it is, you can write out $\langle v, v\rangle$ for the specific semi-inner-product you're thinking about, then compare to the example seminorms' definitions in Section 2.

Exercise 12 *For each of these examples of semi-inner-products, what is the corresponding seminorm?*

3.1 Cauchy-Schwarz Inequality

The Cauchy-Schwarz inequality is the first tool we reach for when bounding a semi-inner-product. For any semi-inner-product $\langle \cdot, \cdot\rangle$, $|\langle u, v\rangle| \leq \rho(u)\rho(v)$ where $\rho(v) = \sqrt{\langle v, v\rangle}$; furthermore, given any u, there is always a vector v of a given 'length' $\rho(v)$ for which this bound is attained.

Exercise 13 *Think about the Cauchy-Schwarz inequality in context of the inner product $\langle u, v\rangle = uv$ on scalars, the dot product $\langle u, v\rangle_2 = u^T v$, and the covariance inner product $\mathrm{Cov}_{\mathrm{P}}(u, v)$. In each context, what does it say? Be as context-specific as you can; repeating the definition three times is not an instructive exercise. A sentence or two will do for each.*

I will not ask you to prove the Cauchy-Schwarz inequality, but if you're interested, take a look at one of the proofs on Wikipedia.

3.2 Hölder's Inequality

To bound the dot product on vectors in $\mathbb{R}^n$, Hölder's inequality is the second tool we reach for. While this is a fairly general tool, we often use a simple special case that's easy to prove: the one for the dot product, $|\langle u, v\rangle_2| \le \|u\|_1 \|v\|_\infty$.[2]

Exercise 14 *Prove it! If it takes you more than one line, you're doing it wrong.*

There are also versions for some inner products on functions. We'll want one for sample inner products analogous to the one we have for the dot product on vectors above: $\langle u, v\rangle_{L_2(\mathrm{P_n})} \le \|u\|_{L_1(\mathrm{P_n})} \|v\|_{L_\infty(\mathrm{P_n})}$.

Exercise 15 *Prove it! If you want, you can write a new proof, but it may be more instructive to show that it's implied by the case for vectors in $\mathbb{R}^n$.*

3.3 Triangle Inequality

When we showed that a few of our examples of seminorms are in fact seminorms in Exercise 7, we didn't deal with any examples of seminorms associated with semi-inner-products. Let's do that now. Or the hard part, anyway.

Exercise 16 *Prove the triangle inequality for a seminorm $\rho(v) = \sqrt{\langle v, v\rangle}$ defined in terms of semi-inner-product.*

Hint. You want to show that $\rho(u+v)^2 \le (\rho(u) + \rho(v))^2$. You know that $\rho(u+v)^2 = \langle u+v, u+v\rangle$. Expand this as the sum of four terms using *linearity*, then see what you can work out using the Cauchy-Schwarz inequality.

Hint. If you are not entirely comfortable with notation $\langle u, v\rangle$ for inner products, use the more familiar notation $u^T v$.

[2] If you'd like to get a sense of Hölder's inequality in full generality, take a look at this wikipedia article.

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 4.

Exercise 17 *For the functions v_1,v_2,and v_3 from Exercise 4, calculate $\|v\|_{L_\infty}$ and $\|v\|_{L_\infty(P)}$ where P is the standard normal distribution.*

This one is a version of Exercise 6

Exercise 18 *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.