Homework 0
Machine Learning and Nonparametric Estimation

David A. Hirshberg

August 23, 2022

As we move forward, there are some concepts and related notation that it'll help to feel comfortable with. 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.

This is meant to review and generalize things you've probably seen before. That itself can be hard work, but it'll be harder if what it's meant to be reviewing is new to you. If this is more work than you want, let me know. This isn't what most of the class will look like, so I hope it won't dissuade you from taking it.

1 Complex Exponentials and Trigonometric Functions

Hopefully you know the ingredients for this one. You'll need to know the exponential function e^x, the trigonometric functions $\sin(x)$ and $\cos(x)$, and the complex numbers $z = x + iy$. If you don't, wikipedia is good for stuff like that.

While the exponential function and trigonometric functions seem pretty different when you're thinking about real numbers, the complex numbers help to connect them. Here's something you can take as a definition.

$$e^{x+iy} = e^x\{\cos(y) + i\sin(y)\} \quad \text{for} \quad x, y \in \mathbb{R}. \tag{1}$$

We tend to think of a complex number $z = x + iy$ as vectors in the plane, with magnitude $|x + iy| = \sqrt{x^2 + y^2}$ and angle $\tan^{-1}(y/x)$. In the case of the exponential, mathematical and everyday language intersect nicely. Exponential growth in an imaginary direction has no impact on magnitude; it's just spinning around the circle. We call $\bar{z} = x - iy$ the complex conjugate of $z = x + iy$, which allows us a convenient expression for the magnitude of z: $|z| = \sqrt{z\bar{z}}$.[1]

As is often the case, we're not interested in working with complex numbers for their own sake, but we do it because it's much more convenient to work

[1] We will also refer to the complex conjugate of a vector or a function, which is interpreted *elementwise*. For a vector $v \in \mathbb{C}^n$ with elements v_i, $\bar{v}$ is the vector with elements $\bar{v}_i$; for a complex-valued function v, $\bar{v}$ is the function with $\bar{v}(x) = \overline{v(x)}$ for all x.

with complex exponentials than with trigonometric functions directly. You can use all the arithmetic you're used to using for powers, so you don't need to use complicated formulas.

$$\begin{aligned}\cos(a+b) + i\sin(a+b) = e^{i(a+b)} &= e^{ia}e^{ib} \\ &= \{\cos(a) + i\sin(a)\}\{\cos(b) + i\sin(b)\} \\ &= \cos(a)\cos(b) + i^2\sin(a)\sin(b) + i\{\cos(a)\sin(b) + \sin(a)\cos(b)\}\end{aligned} \tag{2}$$

so, substituting $i^2 = -1$ and matching up real and imaginary parts,

$$\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b) \quad \text{and} \quad \sin(a+b) = \cos(a)\sin(b) + \sin(a)\cos(b).$$

1.1 Properties of Pure-Imaginary Exponentials

1. Find formulas for $\cos(x)$ and $\sin(x)$ in terms of e^{ix} and e^{-ix}. Here's a hint. $\cos(x) = \cos(-x)$, but $\sin(-x) = -\sin(x)$.
2. Calculate $|e^{ix}|$. Here's a hint. $\overline{e^{ix}} = e^{-ix}$. Why is this true?
3. Show that, for integers k and j, $\int_{-1}^{1} e^{i\pi kx}\overline{e^{i\pi jx}}dx = 2$ if $k = j$ and 0 otherwise.
4. Show that $\overline{wz} = \bar{w}\bar{z}$ for any complex numbers w, z and use this to show that, for integers k and j, $\int_{-1}^{1} ce^{i\pi kx}\overline{de^{i\pi jx}} = 2c\bar{d}$ if $k = j$ and 0 otherwise.
5. Show that $\int_{-1}^{1} \cos(\pi kx)\sin(\pi kx) = 0$ for integers k.
6. Show that $\int_{-1}^{1} \cos(\pi kx)^2 = \int_{-1}^{1} \sin(\pi kx)^2 = 1$ for integers k.

2 Simple Differential Equations

Complex exponentials are great for differential equations. It's still the case that $\frac{d}{dx}e^{ax} = ae^{ax}$ when a is a complex number and $x \in \mathbb{R}$, so in particular we can see that $\phi(x) = e^{ix}$ solves the differential equation $\frac{d^2}{dx^2}\phi = -\phi$: $\frac{d^2}{dx^2}e^{ix} = i^2e^{ix} = -e^{ix}$. More generally,

$$\phi(x) = e^{ikx} \quad \text{solves} \quad \frac{d^2}{dx^2}\phi = -k^2\phi. \tag{3}$$

That isn't the only solution. For example, $\phi(x) = e^{-ix}$ also solves $\frac{d^2}{dx^2}\phi = -\phi$. And given any set of solutions, we can find more by multiplying by them constants and adding them together.

1. If ϕ is a solution, so is $c\phi$ for any complex constant $c \in \mathbb{C}$.
2. If $\phi_1 \ldots \phi_k$ are solutions, so is $\phi_1 + \ldots + \phi_k$.

We can combine our solutions like this to find ones that satisfy constraints.

2.1 Warm Up Exercise: First Order Differential Equations

Find a solution to the differential equation $\frac{d}{dx}\phi = -\phi$ satisfying $\int_0^\infty \phi(x)dx = 1$. Then find one satisfying $\int_0^1 \phi(x)dx = 1$ and one satisfying $\int_0^1 \phi(x)^2 dx = 1$.

2.2 Second Order Differential Equations

Find a solution to the differential equation $\frac{d^2}{dx^2}\phi = -\pi^2\phi$ that is real-valued ($\phi(x) \in \mathbb{R}$) and satisfies $\int_{-1}^1 \phi(x)^2 dx = 1$ and $\phi(0) = 0$. Find another with $\phi(0) = 1$.

Now find a general formula for real-valued solutions to this differential equation. Express this as a linear combination $\phi_{a,b} = a\phi_0 + b\phi_1$ of your previous solutions ϕ_0 and ϕ_1 that satisfied $\phi_0(0) = 0$ and $\phi_1(0) = 1$. And find the specific combination satisfying $\phi_{a,b}(0) = c_0$ of minimal 'length' $\sqrt{(1/2)\int_{-1}^1 \phi_{a,b}(x)^2}$.

3 Seminorms

In this problem and the ones remaining, 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, subtract, and multiply by scalars. Here are the important examples.

- Complex numbers $\mathbb{C}$.
- Finite-dimensional vectors $v \in \mathbb{C}^n$. You know this stuff.
- Functions from some set to $\mathbb{C}$. We add, subtract, and scale these *pointwise*
  - $f + g$ is a function with $(f + g)(x) = f(x) + g(x)$;
  - $f - g$ is a function with $(f - g)(x) = f(x) - g(x)$;
  - αg for $\alpha \in \mathbb{C}$ is a function with $(\alpha f)(x) = \alpha f(x)$.

Vectors spaces have a *zero element*, which we will write 0, with the property that $f + 0 = f - 0 = f$. For finite dimensional vectors, it's the vector of all zeros; for functions, it's the function $f(x) = 0$ that's zero for all x.

During the semester, we'll mostly be working with real numbers/vectors/functions. But just like in high school algebra, it's occasionally useful to work with complex ones. To get in some practice, we'll do everything in the complex case in this homework. Later on, unless we're explicitly working with something complex, we'll not bother writing in all complex conjugates that appear here.

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

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

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.

Here are some examples.

- On complex numbers, i.e., vectors $v = a + ib$ in the space $\mathbb{C}$, we have the magnitude $|v| = \sqrt{v\bar{v}}$. This reduces to the absolute value for real numbers.
- On finite dimensional vectors $v \in \mathbb{C}^n$, we have
  - $\|v\|_2 := \sqrt{\sum_{i=1}^n |v_i|^2}$, 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.
- On functions $v(x)$, given a random variable X with probability distribution P, we have
  - $\|v\|_{L_2(\mathrm{P})} := \sqrt{\mathrm{E}|v(X)|^2}$, the population two-norm.
  - $\|v\|_{L_1(\mathrm{P})} := \mathrm{E}|v(X)|$, the population one-norm.
  - $\|v\|_{L_\infty(\mathrm{P})} := \inf\{x : P(|v(X)| > x) = 0\}$, the population infinity norm. Informally, this is the largest value of $|v(X)|$ that might actually occur. And it's smaller than the largest value outright, $\max_x |v(x)|$, so often even when being formal, you need to think about the subleties.
  - $\mathrm{sd}_\mathrm{P}(v) := \sqrt{\mathrm{E}|v(X) - \mathrm{E}\, v(X)|^2}$, the population standard deviation.
- On differentiable functions $v(x)$ on $[0, 1]$, the *total variation* $\rho_{TV}(v) = \int_0^1 |v'(x)| dx$.

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 $1/n$.

$$\|v\|_{L_2(\mathrm{P_n})} = \sqrt{\frac{1}{n}\sum_{i=1}^n |v(X_i)|^2} = \|v\|_2/\sqrt{n} \qquad \text{the sample two-norm}$$

$$\|v\|_{L_1(\mathrm{P_n})} = \frac{1}{n}\sum_{i=1}^n |v(X_i)| = \|v\|_1/n, \qquad \text{the sample one-norm}$$

$$\|v\|_{L_\infty(\mathrm{P_n})} = \max_{i \le n} |v(X_i)| = \|v\|_\infty \qquad \text{the sample infinity norm.}$$

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

3.1 Checking Properties

Show that the population one-norm, the population 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 seminorms as well.

You may assume that the magnitude is a seminorm on complex numbers. In the next problem, we will prove it.

The population infinity norm is a little tricky. Try the infinity norm on finite dimensional vectors first. If you're having trouble adapting your argument to deal with the population infinity norm, or just not used to thinking about infima, feel free to skip it. We can get by the more informal understanding mentioned above.

3.2 Zero at zero

Seminorms are zero at zero, i.e., they satisfy $\rho(0) = 0$. Prove it. If your proof is more than one sentence, you're doing it wrong.

3.3 Non-Negativity

Seminorms are non-negative. Prove it. This one shouldn't be much longer.

3.4 Positivity

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

3.5 Balls, Spheres, and Convexity

We call the set $\mathcal{B}_{r,\rho} := \{v : \rho(v) \leq r\}$ the seminorm's *ball* of radius r and the set $\mathcal{B}^{\circ}_{r,\rho} := \{v : \rho(v) = r\}$ the seminorm's *sphere* of radius r. We will often use balls like this in the definition of our regression models. Prove that for any seminorm ρ and any radius r, the ball $\mathcal{B}^{\rho}_{r}$ is a convex set—a set with the property that, if it contains points u and v, it also contains every point on the line segment $\{\lambda u + (1-\lambda)v : \lambda \in [0,1]\}$ between them. And explain why the sphere is not convex.

4 Inner Products

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

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

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 generally, but there are some semi-inner-products we use a fair amount.

Here are some examples.

- On complex scalars, we have the product $\langle u, v \rangle = u\bar{v}$. Since a real number's complex conjugate is itself, it reduces to uv on the reals.
- On finite dimensional vectors $v \in \mathbb{C}^n$, we have the dot product, $\langle u, v \rangle_2 := \sum_{i=1}^n u_i \bar{v}_i = u^T \bar{v}$.
- On functions $v(x)$, in terms of a random variable X with distribution P, we have the population inner product $\langle u, v \rangle_{L_2(\mathrm{P})} := \mathrm{E}\, u(X)\bar{v}(X)$ and covariance $\mathrm{Cov}_{\mathrm{P}}(u, v) = \mathrm{E}\, u(X)\bar{v}(X) - \mathrm{E}\, u(X)\, \mathrm{E}\, \bar{v}(X)$.

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, $\langle u, v \rangle_2 = \sum_{i=1}^n u(X_i)\bar{v}(X_i)$. And, as before, this is just a scaled version of the population inner product for the empirical distribution $\mathrm{P_n}$.

4.1 Checking Properties

Prove that these are semi-inner-products.

4.2 Associated Seminorms

For each of these, there is an associated seminorm $\rho(v) = \sqrt{\langle v, v \rangle}$ included in our examples. Which is it?

4.3 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. In the context of each of the examples above, interpret that statement. To keep it simple, think of the vectors and functions as real-valued.

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.

4.4 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: $|\langle u, v \rangle_2| \leq \|u\|_1 \|v\|_\infty$. 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. If you'd like to get a sense of Hölder's inequality in full generality, take a look at wikipedia. For our purposes, we'll want this one simple case.

4.5 Triangle Inequality

Prove the triangle inequality for a seminorm $\rho(v) = \sqrt{\langle v, v\rangle}$ defined in terms of semi-inner-product. Hint: $\rho(u+v)^2 = \langle u+v, u+v\rangle$. Expand this as the sum of four terms using *linearity*, then see what you can work out using the Cauchy-Schwarz inequality.

5 Norms of Symmetric Matrices

Every real symmetric $n \times n$ matrix A has a diagonalization,

$$A = \sum_{k=1}^{n} \lambda_k u_k u_k^T$$

where the eigenvalues λ_k are real and the eigenvectors $u_k \in \mathbb{R}^n$ are orthonormal, i.e., $u_k^T u_k = 1$ and $u_j^T u_k = 0$ for $j \neq k$.[2] What this means is that if you multiply a vector in the direction of u_k by A, it comes out λ_k-times longer pointing in the same direction — or the opposite direction is λ_k is negative. This helps us think about what the matrix-vector product Av looks like. First, we decompose the vector v as a linear combination of the vectors u_k, i.e., we write $v = \sum_k \alpha_k u_k$ then; think about how each term comes out as a scaled version of u_k; and finally, add those together.

5.1 Operator Norm

We call the largest absolute value $\max_k |\lambda_k|$ of an eigenvalue of A the *operator norm* because it is the largest scale factor we can get when 'operating on' a vector of $v \in \mathbb{R}^n$ by multiplication. That is, $\|A\|_{op} := \max_{v\in\mathbb{R}^n} \|Av\|_2/\|v\|_2$ is equal to $\max_k |\lambda_k|$. Prove it. And prove that it's a seminorm, i.e., that it satisfies $\|A+B\|_{op} \leq \|A\|_{op} + \|B\|_{op}$ for matrices A, B and $\|\alpha A\|_{op} = |\alpha|\|A\|$ for $\alpha \in \mathbb{C}$.

5.2 Frobenius Norm

When we multiply a random vector by a symmetric matrix, it gets scaled by a random combination of its eigenvalues. For a gaussian vector $g \in \mathbb{R}^n$ with mean zero and identity covariance, i.e. with $g_i \overset{iid}{\sim} N(0,1)$, it's fairly easy to characterize a sort of typical scaling $\sqrt{\mathrm{E}\|Ag\|_2^2 / \mathrm{E}\|g\|_2^2}$ in terms of the eigenvalues of A. What is it? Prove that it's a seminorm. We call the numerator $\mathrm{E}\|Ag\|_2^2$ the *Frobenius norm* of the matrix A, and write it $\|A\|_F$.

[2] It's conventional to order the terms in this sum so the eigenvalues λ_k are decreasing.

5.3 Eigenvectors of Self-Adjoint Operators

We can think of a symmetric matrix as a *self-adjoint linear operator* on the space of vectors $v \in \mathbb{R}^n$ with the usual inner product $\langle \cdot, \cdot \rangle_2$. That is, it satisfies $\langle Au, v\rangle_2 = \langle u, Av\rangle_2$. The first is $(Au)^T v = u^T A^T v$ and the second is $u^T Av$, and they are equal because $A^T = A$.

This idea applies more generally. For example, we can talk about self-adjoint linear operators on spaces of functions. A classic example is the differential operator $-\frac{d^2}{dx^2}$ on the space of real-valued twice-differentiable functions on $[-1, 1]$ that are periodic in the sense that $v(-1) = v(1)$ with inner product $\langle u, v\rangle = (1/2)\int_{-1}^{1} u(x)v(x)dx$.

An operator has eigenvalues and eigenvectors, just like a matrix.[3] In our example, they are defined by the differential equation $-\frac{d^2}{dx^2}v = \lambda v$. And like a symmetric matrix, a self-adjoint linear operator's eigenvalues are real and the eigenvectors corresponding to distinct eigenvalues are orthogonal. Prove it. And having done this, explain why this implies that, for integers j and k with $j \neq k$,

$$\int_{-1}^{1} \sin(\pi kx)\sin(\pi jx) = \int_{-1}^{1} \cos(\pi kx)\cos(\pi jx) = \int_{-1}^{1} \cos(\pi kx)\sin(\pi jx)dx = 0.$$

[3]The eigenvectors of operators on vector spaces of functions are sometimes called eigenfunctions.