

LECTURE 1

I. Fourier series

We ended last week talking about our desire to take a function f defined on $[0, 2\pi]$ and approximate it by a combination of sine waves:

$$f(x) \sim b_0 + a_1 \sin x + a_2 \sin 2x.$$

If I haven't said it already, talk about how a piece of meat in your ear (namely, the basilar membrane in the cochlea) carries out Fourier analysis so that you can hear the notes in a chord.

And how do we do this? We decided last time we wanted to minimize the integral

$$\int_0^{2\pi} [f(x) - (b_0 + a_1 \sin x + a_2 \sin 2x)]^2 dx.$$

Well. Let me observe that what we want to do is

“Orthogonally project f onto the space V spanned by $1, \sin x, \sin 2x$.”

Now what does this mean? The problem is that we do not have a good notion of what it means for two functions to be orthogonal. So we can't make any sense of the above statement. On the other hand, look up again at our least squares system from last time involving values at $0, \pi/2, \pi, 3\pi/2$. In that case, we had f represented by the vector

$$[f] = \begin{bmatrix} f(0) \\ f(\pi/2) \\ f(\pi) \\ f(3\pi/2) \end{bmatrix}$$

and if we had two functions f and g , we would have

$$[f] \cdot [g] = f(0)g(0) + f(\pi/2)g(\pi/2) + f(\pi)g(\pi) + f(3\pi/2)g(3\pi/2).$$

Now if we once again let the points get more and more finely distributed, we arrive at the following definition.

Definition. Let f and g be two functions defined on $[a, b]$. Then we define the “dot product” $f \cdot g$ to be

$$\int_a^b f(x)g(x)dx.$$

Remark that Strang calls this (f, g) (Strang, p.177) which avoids confusion with the dot product of vectors. But I will call it $f \cdot g$ for today to emphasize the analogy. It should rightly be called an *inner product* of the two functions.

Principle: This “dot product” behaves in every way like the dot product of vectors we are used to.

Example (Exercise) Prove $f \cdot f = 0$ if and only if $f = 0$. **Example** We say two functions f and g are *orthogonal* if $f \cdot g = 0$.

Fact: $1/\sqrt{2\pi}, \sin x/\sqrt{\pi}, \sin 2x/\sqrt{\pi}$ are an orthonormal basis for V , with respect to the above inner product.

Call these three functions f_1, f_2, f_3 . Then, for instance,

$$(\sqrt{2\pi})f_1 \cdot f_3 = (\sqrt{2\pi}) \int_0^{2\pi} f_1(x)f_3(x)dx = \int_0^{2\pi} \sin 2x dx = 0.$$

And

$$\begin{aligned} f_2 \cdot f_2 &= (1/\pi) \int_0^{2\pi} f_2^2(x)dx = (1/\pi) \int_0^{2\pi} \sin^2 x dx \\ &= (1/\pi) \int_0^{2\pi} (1 - \cos 2x)/2 dx \\ &= (1/\pi) \left[\int_0^{2\pi} 1/2 dx - \int_0^{2\pi} (\cos 2x)/2 dx \right] = (1/\pi)\pi = 1. \end{aligned}$$

So according to our formula from last time, we may state that the orthogonal projection of f onto V is just

$$(f \cdot f_1)f_1 + (f \cdot f_2)f_2 + (f \cdot f_3)f_3$$

For example, if we take $f(x) = x$, we get

$$(f \cdot f_1) = \int_0^{2\pi} x/\sqrt{2\pi} dx = (1/\sqrt{2\pi})(x^2/2)|_0^{2\pi}$$

which comes out to

$$(1/\sqrt{2\pi})(2\pi^2)$$

so $(f \cdot f_1)f_1 = \pi$. Similarly, we can compute

$$(f \cdot f_2)f_2 = -2 \sin x, (f \cdot f_3)f_3 = -\sin 2x.$$

So, to sum up, we estimate

$$x \sim \pi - 2 \sin x - 2 \sin 2x.$$

Draw a sketch of this to show it's pretty good. If we allowed more terms, it'd be even better.

Observe that there really doesn't seem to be a way to use our formula from last time, involving A , to produce this data; we would have to deal with matrices of infinite size. Yikes!

II. Orthogonal transformations and orthogonal matrices

We've now seen at least one vitally important example of the utility of having an orthonormal basis at hand. Now it's time to talk about how to produce such a basis for a given space—it seems like in the example above we got pretty lucky that the “obvious” basis was already orthonormal.

Definition: An *orthogonal matrix* is an $n \times n$ matrix whose columns are an orthonormal basis for \mathbb{R}^n .

Example: The identity matrix I_n . **Example:** The matrix

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

which, geometrically speaking, is “rotation by θ degrees counterclockwise.”

Example: Any permutation matrix is orthogonal.

Observe that an orthogonal matrix is automatically *invertible*, because its columns form a basis of \mathbb{R}^n . In fact, we can even say exactly what the inverse is:

Fact: If Q is an orthogonal matrix, then $Q^T = Q^{-1}$.

Proof: write out $Q^T Q$, and observe that the i, j entry is just $\vec{v}_i \cdot \vec{v}_j$, which is 1 if $i = j$ and 0 otherwise. In other words, $Q^T Q = I_n$, which is what we desire.

In fact, we have proved something more general (and also useful):

Fact: Suppose Q is an $m \times n$ matrix with orthonormal columns. Then $Q^T Q = I_n$.

In particular, if $\vec{y} \in \mathbb{R}^m$, then the projection of y onto $\mathcal{R}(Q)$ is given by

$$P = Q(Q^T Q)^{-1} Q^T = Q Q^T.$$

Orthogonal matrices have a very beautiful geometric meaning.

Fact: Orthogonal matrices preserve length. That is, if Q is an orthogonal matrix, then $|Q\vec{x}| = |\vec{x}|$.

Proof.

$$|Q\vec{x}|^2 = (Q\vec{x})^T Q\vec{x} = \vec{x}^T Q^T Q\vec{x} = \vec{x}^T \vec{x} = |\vec{x}|^2.$$

In fact, the converse is also true:

Fact: If a matrix preserves length, it is orthogonal.

But I won't prove that here. (Though it's not too hard.)

Fact: Orthogonal matrices preserve angles; that is, the angle between $Q\vec{x}$ and $Q\vec{y}$ is the same as the angle between \vec{x} and \vec{y} .

Because

$$Q\vec{x} \cdot Q\vec{y} = (Q\vec{x})^T(Q\vec{y}) = \vec{x}^T Q^T Q \vec{y} = \vec{x} \cdot \vec{y}.$$

And the angle is given by

$$\cos\theta = (\vec{x} \cdot \vec{y})/|\vec{x}||\vec{y}|.$$

Orthogonal transformations are often, for this reason, called "rigid motions." And, physically, lots of things behave this way. Length is, in many contexts (at least, non-relativistic contexts!) preserved by the physical transformations we want to study.

Ask: what are the rigid motions of the plane? We'll have a homework problem on this.

Now the idea is the following. Suppose we want to solve

$$A\vec{x} = \vec{y}$$

by least squares. Well, so we could find a basis for A and project \vec{y} onto $\mathcal{R}(A)$ via the formula $P = A(A^T A)^{-1} A^T$. But maybe it is better to find an *orthonormal* basis for $\mathcal{R}(A)$ instead of the one we started with.

Problem. Given a space V with a basis $\vec{v}_1, \dots, \vec{v}_d$. Find an orthonormal basis for V .

III. Gram-Schmidt process.

Idea: Suppose we start with a non-orthonormal basis. Like, say,

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

for a subspace $V \subset \mathbb{R}^3$. We desire to make these guys orthonormal.

Well. There's an algorithm.

Step 1. Make \vec{v}_1 orthonormal.

What does it mean to make *one vector* orthonormal? All it means, is, make its length 1. We can achieve this by multiplying by a scalar; in this case, $1/\sqrt{2}$.

That is:

Replace \vec{v}_1 by $\vec{v}_1/|\vec{v}_1|$.

So at this point we have

$$\vec{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}.$$

Step 2. Make \vec{v}_2 orthogonal to \vec{v}_1 . Here, draw a picture. Show how we have to subtract the “*vecv*₂-directional part” off of \vec{v}_2 to make the two orthogonal, and use this to motivate the step:

Replace \vec{v}_2 by $\vec{v}_2 - (\vec{v}_1^T \vec{v}_2) \vec{v}_1$.

In this case, we get

$$(\vec{v}_1^T \vec{v}_2) = 3/\sqrt{2}$$

so

$$\vec{v}_2 - (\vec{v}_1^T \vec{v}_2) \vec{v}_1 = \begin{bmatrix} 1 - 3/2 \\ 2 - 3/2 \\ 0 \end{bmatrix} = \begin{bmatrix} -1/2 \\ 1/2 \\ 0 \end{bmatrix}.$$

So we have

$$\vec{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} -1/2 \\ 1/2 \\ 0 \end{bmatrix}.$$

Step 3. Make \vec{v}_1, \vec{v}_2 orthonormal. We already have \vec{v}_1 length 1, and the two vectors are already orthogonal, so all that remains is to make \vec{v}_2 have length 1, which we can do by dividing by its length.

Replace Replace \vec{v}_2 by $\vec{v}_2/|\vec{v}_2|$.

And this gives

$$\vec{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \vec{v}_2 = \sqrt{2} \begin{bmatrix} -1/2 \\ 1/2 \\ 0 \end{bmatrix}.$$

Visibly these two vectors span the same space as do the original two—namely, the $x - y$ plane.

Gram-Schmidt process: Start with a basis $\vec{v}_1, \dots, \vec{v}_d$ for a space V . Then carry out the following series of steps, for each i from 1 to d .

- Replace \vec{v}_i by

$$\vec{v}_i - [(\vec{v}_1^T \vec{v})\vec{v}_1 + (\vec{v}_2^T \vec{v})\vec{v}_2 + \dots + (\vec{v}_{i-1}^T \vec{v})\vec{v}_{i-1}].$$

- Replace \vec{v}_i by $\vec{v}_2/|\vec{v}_2|$.

The result will be an orthogonal basis for V .

A good example is the space of polynomials—this is example 4 in Strang.

Something I *won't* talk about: the QR factorization. For this, I refer you to Strang. I will only say:

Fact: Let A be an $m \times n$ matrix with linearly independent columns. Then A can be written as a product QR , where Q is an $m \times n$ matrix with orthogonal columns, and R is an upper triangular $n \times n$ matrix with nonzero diagonal entries.

The construction of this factorization essentially *is* the Gram-Schmidt process. I refer you to p. 174-176 of Strang to learn more.

Lecture III.

I want to take our gaze away, for the moment, from questions of orthogonality, and revisit some concepts from the first part of the course with more sophisticated eyes.

I. Image and kernel.

This is a revisit of the idea of “nullspace” and “column space.”

Let $T : V \rightarrow W$ be a linear transformation. We say

- The *kernel* of T is the set of all \vec{v} such that $T(\vec{v}) = 0$. It is a subspace of V .
- The *image* of T is the set of all \vec{w} which are of the form $T(\vec{v})$ for some \vec{v} . It is a subspace of W .

If A is an $m \times n$ matrix, then we can think of A as a transformation from $\mathbb{R}^n \rightarrow \mathbb{R}^m$. Then the kernel is the nullspace and the image is the column space.

So it's natural to expect (but ask mid-sentence):

Thm: $\dim \ker T + \dim \text{ima} T = \dim V$.

And indeed this is true.

Example. Consider the map $T : P_3 \rightarrow \mathbb{R}^3$ given by

$$T(f) = \begin{bmatrix} f(0) \\ f(1) \\ f(2) \end{bmatrix}.$$

What is its kernel? What is its image? (This is essentially a question on the midterm.) Find that the kernel is 1-dimensional, and the image is 3-dimensional. The sum is 4, just as it should be. What about $T : P_3 \rightarrow \mathbb{R}^6$ given by

$$T(F) = \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \\ f(4) \\ f(5) \end{bmatrix} ?$$

What are the dimensions of the kernel and the image?

Remark: One way to do these problems is to find a matrix representing the linear transformation, and then compute rank. But often other methods are easier.

II. Intersection of vector spaces.

Now. Suppose somebody asks you: Do you think there is a nonzero cubic polynomial f such that $f(0) + f(1) + f(2) + f(3) + f(4) + f(5) = 0$?

That's not such an unreasonable kind of thing to ask. How would you think of it? Well, you might say something like the following. It's like saying: is there a vector

$$\begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \\ f(4) \\ f(5) \end{bmatrix} = T(f)$$

which lies in the space of vectors whose coordinates sum to 0?

Let V be $\text{ima}T$ and W be the space of vectors whose coordinates sum to 0—that is, those vectors \vec{v} such that $[111111]\vec{v} = 0$ —that is, the nullspace of

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} .$$

And the question is: is there some nonzero vector contained in both V and W ? That is, what is $V \cap W$?

Fact: $V \cap W$ is a subspace.

Idea: if the dimensions of the spaces V and W are pretty big, we expect to have some intersection. If they are pretty small, we expect maybe not.

Fact: Suppose V, W are subspaces of \mathbb{R}^n , and $\dim V + \dim W > n$. Then $V \cap W$ is larger than $\vec{0}$.

You might remember this as a homework problem, where $n = 6$ and V, W were 3-dimensional. Prove this at the board, if there is time.

Warning: It is *not* true that a basis for V and a basis for W must have a vector in common. For instance; let W be the $x - y$ plane and V be the space generated by and

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \vec{v}_2 = \text{vectwo121}.$$

Both are two-dimensional, so V and W must have a vector in common by the above Fact. But neither \vec{v}_1 nor \vec{v}_2 are in the $x - y$ plane. However, $\vec{v}_1 - \vec{v}_2$ is in both V and W .

Also by the above fact, there *is* a cubic with $f(0) + f(1) + f(2) + f(3) + f(4) + f(5) = 0$, because in that case V is 4-dimensional and W is 5-dimensional, and $4+5 > 6$. But it's a little tricky to *compute* the intersection of two vector spaces in general. Strang describes a good method on p.199.

ASK: Suppose I asked you to do this? Forget linear algebra for a moment—common-sensically, how might you come up with a cubic polynomial whose values at $0, \dots, 5$ summed to 0 as above? Do this in pairs.

III. Weighted least squares.

I want to return now to the idea of least squares. Recall (or maybe I'm saying this for the first time) that the average is the simplest example of a least-squares estimate. Namely, if we try to estimate

$$x = a_1, x = a_2, x = a_3, \dots, x = a_n$$

then we are trying to solve the equation

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} x = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

or

$$A\vec{x} = \vec{a}$$

which, as we now well know, has the least squares solution

$$\vec{x} = (A^T A)^{-1} A^T \vec{a}$$

which we can compute to be the average of a_1, \dots, a_n .

Now suppose we trust one of these measurements more than the others. For instance, suppose we trust the first one a lot more. How can we incorporate that into our system?

Idea:

Replace the first equation $x = a_1$ with $100x = 100a_1$. It seems exactly the same, right? But it's not. Because consider our measurement of error. Now any difference between x and a_1 will be highly magnified when we compute our overall error. So when we minimize the overall error, we must concentrate much more on minimizing the difference between x and a_1 . Which is as it should be!

So let's formalize this a little bit: we are replacing our original A with WA , where W is the diagonal matrix

$$W = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix}.$$

And we are trying to solve

$$WA\vec{x} = W\vec{a}.$$

So now the least squares solution is

$$\vec{x} = ([WA]^T WA)^{-1} (WA)^T W\vec{a} = (A^T W^T WA)^{-1} A^T W^T W\vec{a}.$$

And in fact this is in general how we compute the least squares solution when we want to place different amounts of trust in different measurements.