LECTURE 1

I. Orthogonal projection

I talked a bit about orthogonal projection last time and we saw that it was a useful tool for understanding the relationship between V and $V^{\perp}$. Now let's speak of it a little more cogently.

Our main goal today will be to understand orthogonal projection onto a line.

Draw two vectors $\vec{x}$ and $\vec{a}$. Let P be the matrix representing the transformation "orthogonal projection onto the line spanned by $\vec{a}$. Draw the picture. We can see that $P\vec{x}$ must be *some* multiple of $\vec{a}$, because it's on the line spanned by $\vec{a}$. But what multiple?

First example: $\vec{a} = \vec{e}_1$. We see that in this case,

$$P \begin{bmatrix} x_0 \\ \vdots \\ x_n \end{bmatrix} = x_0 \vec{e}_1.$$

Observe that we can also write

$$P\vec{x} = \vec{x} \cdot \vec{e}_1 = \vec{e}_1^T \vec{x}.$$

This may give us an idea for the general formula. So again. Suppose that

$$P\vec{x} = c\vec{a}.$$

Our goal is to compute c.

To find that formula, observe that $\vec{x} - P\vec{x}$ must be perpendicular to $\vec{a}$. In other words,

$$0 = \vec{a}^T(\vec{x} - P\vec{x}) = \vec{a}^T(\vec{x} - c\vec{a}) = \vec{a}^T\vec{x} - c\vec{a}^T\vec{a}$$

so

$$c = \frac{\vec{a}^T\vec{x}}{\vec{a}^T\vec{a}} = \frac{\vec{a}^T\vec{x}}{|\vec{a}|^2}. \tag{1}$$

Put a box around the above formula.

Observe that in case $\vec{a} = \vec{e}_1$, this indeed gives rise to the special case above.

Remark: If $\vec{x}$ is orthogonal to $\vec{a}$, we find that the projection of $\vec{x}$ onto the line spanned by $\vec{a}$ is 0. This agrees with our geometric intuition.

If there seems to be plenty of time, observe that we can also think of the angle θ between $\vec{x}$ and $\vec{a}$ and use trigonometry to compute c. We find

$$P\vec{x} = (\cos\theta|\vec{x}|/|\vec{a}|)\vec{a}$$

and if you compare this with (**??**), you find the usual formula for the cosine of the angle in terms of dot products (Strang, 3G).

Example: What is the projection of $\vec{x} = \begin{bmatrix} 6 \\ 5 \\ 4 \end{bmatrix}$ onto the line spanned by $\vec{a} = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}$?

Well, in this case, we get

$$P\vec{x} = \frac{\vec{a}^T\vec{x}}{\vec{a}^T\vec{a}}\vec{a} = \frac{12}{11}\vec{a} = \begin{bmatrix} 12/11 \\ -12/11 \\ 36/11 \end{bmatrix}.$$

Geometric remark: The projection $P\vec{x}$, in geometric terms, is the *closest point* to $\vec{x}$ on the line span$\vec{a}$.

II. Approximate solutions

This leads us to a basic statement of philosophy. Our goal up to this point has been: solve the Basic Problem $A\vec{x} = \vec{y}$.

> Sometimes there is no solution to $A\vec{x} = \vec{y}$. In that case, find $\vec{x}$ such that $A\vec{x}$ is *closest* to $\vec{y}$.

For example: suppose we were trying to solve

$$A\vec{x} = \begin{bmatrix} 6 \\ 5 \\ 4 \end{bmatrix} \tag{2}$$

where

$$A = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}.$$

Here, $\vec{x}$ would have to be a single number (or, if you like, a 1×1 matrix.) Now here the rank is just 1, much less than the number of rows, so there's no reason to expect a solution. Indeed, there isn't one. Another way to say this is that $\begin{bmatrix} 6 \\ 5 \\ 4 \end{bmatrix}$ is not in the column space of A.

But we might ask–what $\vec{x}$ is the best *approximate* solution to (**??**)? In other words, what is the *closest* point in the column space of A to $\begin{bmatrix} 6 \\ 5 \\ 4 \end{bmatrix}$? And this is the question we answered above.

Answer: $\vec{x} = 12/11$ is the best *approximate* solution to (**??**).

(Not that it's a very good one!)

III. Orthogonal projection as linear transformation.

Let me return to the fact that orthogonal projection is a linear transformation. That is, as we said above, there's a matrix P such that

$$P\vec{x} = \text{projection of } \vec{x} \text{ onto span}\vec{a} = \frac{\vec{a}^T\vec{x}}{\vec{a}^T\vec{a}}\vec{a}.$$

How can we find P? Well, the trick is to write the above equation in another way:

$$P\vec{x} = \vec{a}\frac{\vec{a}^T\vec{x}}{\vec{a}^T\vec{a}} = \frac{\vec{a}\vec{a}^T}{\vec{a}^T\vec{a}}\vec{x}.$$

Since this holds for *all* $\vec{x}$, we have

$$P = \frac{\vec{a}\vec{a}^T}{\vec{a}^T\vec{a}}$$

Isn't that a marvelous formula? Let's check it out: for $\vec{a} = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}$ as above, we get

$$\vec{a}\vec{a}^T = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} [1 - 13] = \begin{bmatrix} 1 & -1 & 3 \\ -1 & 1 & -3 \\ 3 & -3 & 9 \end{bmatrix}$$

while

$$\vec{a}^T\vec{a} = 11.$$

So

$$P = \begin{bmatrix} 1/11 & -1/11 & 3/11 \\ -1/11 & 1/11 & -3/11 \\ 3/11 & -3/11 & 9/11 \end{bmatrix}.$$

Contemplate: what is the rank of P? (We can see this either by doing Gaussian elimination or by thinking of what the column space must be.)

In fact, if there is time, discuss the rank and the four fundamental spaces attached to P.

LECTURE 2.

I. Least squares approximation and the normal equation

In general, how can we find an approximate solution to

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

Well. Let's try to do it.

Observe: If $\vec{y}$ is in the column space of A, we actually have an exact solution $\vec{x}$, so we are done.

If $\vec{y}$ is *not* in $\mathcal{R}(A)$, on the other hand, we have to find the closest solution; that is to say, the point of $\mathcal{R}(A)$ which is closest to $\vec{y}$.

MAIN QUESTION. How to find the point of $\mathcal{R}(A)$ closest to $\vec{y}$? Or: how to find $\vec{x}$ such that $|A\vec{x} - \vec{y}|$ is minimized?

Two ways to approach this question–the calculus way, and the geometric way. I took the geometric approach last time when we spoke about orthogonal projection onto a line; this time I will take the calculus approach. Strang (p.155) discusses the geometric approach to this problem in detail, if you would like to see it done that way.

The idea is as follows. Define a function

$$f(\vec{x}) = |A\vec{x} - \vec{y}|^2 = (A\vec{x} - \vec{y})^T (A\vec{x} - \vec{y}).$$

Remark: The fact that we are finding the *smallest* possible value for the *square* of the length of $A\vec{x} - \vec{y}$ is the reason for the term "least squares approximation."

Remark: f is *not* a linear transformation! We see here, not for the first time, that non-linear functions *do* play a role in making computations in linear algebra.

Anyway: our goal is to *minimize* f. We know from calculus that, if $\vec{x}$ is a local minimum for f, then the partial derivatives of f should all vanish at $\vec{x}$. We can work these partial derivatives out as follows. We have

$$\frac{\partial}{\partial x_i} f(\vec{x}) = \lim_{h \to 0} \frac{f(\vec{x} + h\vec{e}_i) - f(\vec{x})}{h}.$$

Let's compute this. Before we start, I want to remind you that

$$(AB)^T = B^T A^T$$

for any two matrices A, B. Also remember that

$$\vec{v}^T\vec{w} = \vec{v} \cdot \vec{w}$$

for any two vectors $\vec{v}, \vec{w}$.

Now

$$\frac{f(\vec{x} + h\vec{e_i})}{h} = (A(\vec{x} + h\vec{e_i}) - \vec{y})^T(A(\vec{x} + h\vec{e_i}) - \vec{y})$$

which we can expand as

$$(\vec{x}^T A^T + h\vec{e}_i^T A^T + \vec{y}^T)(A\vec{x} + hA\vec{e}_i - \vec{y}).$$

Collecting terms of h, we get

$$\begin{array}{cc} & \vec{x}^T A^T A\vec{x} - \vec{x}^T A^T \vec{y} - \vec{y}^T A^T \vec{x} + \vec{y}^T \vec{y} \\ + & h(-\vec{e}_i^T A^T \vec{y} - \vec{y}^T A^T \vec{e}_i + \vec{e}_i^T A^T A\vec{x} + \vec{x}^T A^T A\vec{e}_i) \\ + & h^2(\vec{e}_i^T A^T A\vec{e}_i^T) \end{array} .$$

The term which doesn't involve h is just $f(\vec{x})$. So can write

$$f(\vec{x} + h\vec{e_i}) = f(\vec{x}) + h(\text{big mess 1}) + h^2(\text{smaller mess 2}).$$

This makes sense: if $h = 0$, we should indeed get $f(\vec{x})$. Now we compute

$$\frac{\partial}{\partial x_i} f(\vec{x}) = \lim_{h \to 0} \frac{f(\vec{x} + h\vec{e_i}) - f(\vec{x})}{h} = \text{big mess 1}$$

or, to sum up,

$$\frac{\partial}{\partial x_i} f(\vec{x}) = -\vec{e}_i^T A^T \vec{y} - \vec{y}^T A\vec{e}_i + \vec{e}_i^T A^T A\vec{x} + \vec{x}^T A^T A\vec{e}_i.$$

Now if $\vec{x}$ is really a minimum for f, this mess must be 0 for all i. Well, let's analyze it further; we can write the mess as

$$-A^T\vec{y} \cdot \vec{e}_i - \vec{e}_i \cdot A^T\vec{y} + A^T A\vec{x} \cdot \vec{e}_i + \vec{e}_i \cdot A^T A\vec{x}$$

which, since dot product is commutative, amounts to

$$\frac{\partial}{\partial x_i} f(\vec{x}) = (A^T A\vec{x} - A^T\vec{y}) \cdot \vec{e}_i.$$

If this is 0 for all i, we are forced to conclude:

> **CONCLUSION:** If $\vec{x}$ is a minimum for $f(\vec{x}) = |A\vec{x} - \vec{y}|^2$, then $A^T A\vec{x} = A^T\vec{y}$.

The equation $A^T A\vec{x} = A^T\vec{y}$ is called the "normal equation." The minimum $\vec{x}$ is called the "least squares solution to $A\vec{x} = \vec{y}$.

Calculus remarks: True 203 warriors will complain that

- I haven't shown that $\vec{x}$ is a minimum for f; the fact that the partial derivatives vanish prove only that it is a *local* minimum.
- For that matter, how do I know it is a minimum and not a *maximum?* Geometric intuition should tell you that f does not have a maximum, or even a local maximum. To prove that $\vec{x}$ is a minimum and not a maximum, you probably know that we would have to compute second partial derivatives, and show that these partials were positive. We don't have the tools to attack this problem yet, but the central point is that $A^T A$ is (to use a word we'll define later) *positive semidefinite.*

Observation: If $\vec{x}$ is an honest solution to $A\vec{x} = \vec{y}$, then evidently it is a solution to the normal equation as well–just multiply both sides by A^T.

There's lots more to say about the theory behind and around the normal equation, but let's move first to a numerical example, which involves one of my favorite applications of linear algebra.

II. Linear regression

The population of the world is estimated to have been

- 980 million in 1800;
- 1.26 billion in 1850;
- 1.65 billion in 1900;
- 2.52 billion in 1950.

(Estimates from `http://www.census.gov`.)

We can write $P(t)$ for "population in year t." Imagine that it's 1950 and we want to project the world's population in the year 2000 from this data; i.e. we want to predict $P(2000)$.

A priori hypothesis: Population will grow exponentially. (This is reasonable, based on our model for how new people are made.)

In that case, we expect that $\log P(t)$ will grow linearly. That is, we make the

Prediction: There exist constants C and D such that

$$\log P(t) \sim C + Dt.$$

Problem: find C, D.

Another way of putting it: we are trying to solve the system of linear equations

$$\begin{aligned} C + 1800D &= log(P(1800)) = 20.70; \\ C + 1850D &= 20.95; \\ C + 1900D &= 21.22; \\ C + 1950D &= 21.65 \end{aligned}$$

or

$$\begin{bmatrix} 1 & 1800 \\ 1 & 1850 \\ 1 & 1900 \\ 1 & 1950 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 20.70 \\ 20.95 \\ 21.22 \\ 21.65 \end{bmatrix}$$

which we can write as $A \begin{bmatrix} C \\ D \end{bmatrix} = \vec{y}$.

Now we can check that this equation actually has no solution; which is to say that, unsurprisingly, the population of the world is not *exactly* described by a linear function. Still, we want to find the *closest* linear function, and this leads us to solve the normal equation

$$A^T A \begin{bmatrix} C \\ D \end{bmatrix} = A^T \vec{y}.$$

Can check that

$$A^T A = \begin{bmatrix} 4 & 7500 \\ 7500 & 14075000 \end{bmatrix}$$

and

$$A^T \vec{y} = \begin{bmatrix} 84.52 \\ 158553 \end{bmatrix}.$$

As it happens, one finds that $A^T A$ is *invertible.*

So the least squares solution is given by

$$\vec{x} = (A^T A)^{-1} A^T \vec{y} = \begin{bmatrix} 9.43 \\ .00624 \end{bmatrix}$$

which is to say that we estimate

$$\log P(t) \sim 9.43 + (.00624)t$$

or

$$P(t) \sim e^{9.43+(.00624)t}.$$

How close is this to the right answer? This estimate gives a population of 940 billion in 1800, 1.28 billion in 1850, 1.75 billion in 1900, and 2.4 billion in 1950. Not too bad! And according to our prediction, we would get

- $P(1960) = 2.55$ billion;
- $P(1980) = 2.89$ billion;
- $P(2000) = 3.27$ billion.

As it happens, these numbers are way off–world population growth exploded after 1950. But that's OK.

Lecture III I. Properties of $A^T A$

Note that if A is $m \times n$, then $A^T A$ is $n \times n$. We have seen above that it's quite useful in the normal equation for $A^T A$ to be invertible.

Fact: $A^T A$ and A have the same nullspace.

Proof: This theorem has an incredibly slick proof. Suppose $\vec{x} \in N(A^T A)$. Then $A^T A\vec{x} = \vec{0}$. So also

$$\vec{x}^T A^T A\vec{x} = 0.$$

But this can be rewritten as

$$(A\vec{x})^T A\vec{x} = |A\vec{x}|^2 = 0$$

and this means that $A\vec{x} = 0$. So $\vec{x} \in N(A)$.

On the other hand, if $A\vec{x} = 0$, then also $A^T A\vec{x} = 0$. So $\vec{x}$ is in $N(A)$ if and only if it is in $N(A^T A)$, which proves the Fact.

Theorem: Suppose $N(A) = 0$ (that is, the columns of A are linearly independent.) Then $A^T A$ is invertible.

Proof. In this case, $A^T A$ is a square matrix with zero nullspace. It follows from the rank theorem that the rank of $A^T A$ is n, so $A^T A$ is invertible.

To sum up: if $N(A) = 0$, we have that $A^T A$ is square, symmetric, and invertible. (You might have noticed the symmetry in the example we did last time...)

II. Orthogonal projection

What is the orthogonal projection of $\vec{y}$ to $\mathcal{R}(A)$? As we have discussed, the best solution to $A\vec{x}$ to $\vec{y}$ is given by

$$\vec{x} = (A^TA)^{-1}A^T\vec{y}.$$

So the point of $\mathcal{R}(A)$ approximating $\vec{y}$ is

$$A\vec{x} = A(A^TA)^{-1}A^T\vec{y}.$$

In other words, if P is the projection onto $\mathcal{R}(A)$, we have

$$P = A(A^TA)^{-1}A.$$

Draw a picture here for clarity, showing $\vec{y}$ and $A\vec{y}$.

Example: Suppose we want to understand the matrix corresponding to the linear transformation "projection onto the plane $x + 2y + 3z = 0$." Call this plane V. First of all, we need to find A such that $\mathcal{R}(A) = V$. **ASK:** How could I do that?

We can do it by finding some vectors that span V. We might as well take a basis

$$\begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix}$$

and then set

$$A = \begin{bmatrix} -2 & -3 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

One finds that

$$A^TA = \begin{bmatrix} 5 & 6 \\ 6 & 10 \end{bmatrix}$$

and

$$(A^TA)^{-1} = \begin{bmatrix} 5/7 & -3/7 \\ -3/7 & 5/14 \end{bmatrix}.$$

From there it is a simple matter to compute

$$P = A(A^TA)^{-1}A^T = \begin{bmatrix} 13/14 & -1/7 & -3/14 \\ -1/7 & 5/7 & -3/7 \\ -3/14 & -3/7 & 5/14 \end{bmatrix}.$$

For instance, if we take $\vec{v} = \begin{bmatrix} 5 \\ -2 \\ 0 \end{bmatrix}$, which is already pretty close to V, we expect that the projection $P\vec{v}$ will be pretty close to $\vec{v}$. Indeed,

$$P \begin{bmatrix} 5 \\ -2 \\ 0 \end{bmatrix} \sim \begin{bmatrix} 4.93 \\ -2.14 \\ -0.214 \end{bmatrix}.$$

III. Orthonormal bases

Definition. Let V be a subspace of $\mathbb{R}^n$. An *orthonormal basis* of V is a basis $\vec{v}_1, \ldots, \vec{v}_d$ such that

$$|\vec{v}_i| = 1$$

for all i, and

$$\vec{v}_i \cdot \vec{v}_j = 0$$

for all $i \neq j$.

Example: The standard basis $\vec{e}_1, \ldots, \vec{e}_n$ of $\mathbb{R}^n$.

Example: $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 3/5 \\ 4/5 \end{bmatrix}$ are an orthogonal basis for their span $W \subset \mathbb{R}^3$.

Such a basis is very useful, especially for orthogonal projection. For simplicity, let's see how it works when the subspace is 2-dimensional, like W.

We want to find $\vec{y}$ in W such that $\vec{x} - \vec{y}$ is perpendicular to the plane W, i.e. to every vector in W. It is enough for it to be perpendicular to $\vec{v}_1$ and $\vec{v}_2$. So write

$$\vec{y} = k_1\vec{v}_1 + k_2\vec{v}_2.$$

Now

$$\begin{aligned} (\vec{x} - \vec{y}) \cdot \vec{v}_1 &= 0 \\ (\vec{x} - k_1\vec{v}_1 - k_2\vec{v}_2) \cdot \vec{v}_1 &= 0 \\ (\vec{x} \cdot \vec{v}_1 - k_1\vec{v}_1 \cdot \vec{v}_1 - k_2\vec{v}_1 \cdot \vec{v}_2) &= 0 \\ \vec{x} \cdot \vec{v}_1 - k_1 &= 0. \end{aligned}$$

So $k_1 = \vec{x} \cdot \vec{v}_1$. Likewise, $k_2 = \vec{x} \cdot \vec{v}_2$.
CONCLUDE:

$$\vec{y} = (\vec{x} \cdot \vec{v}_1)\vec{v}_1 + (\vec{x} \cdot \vec{v}_2)\vec{v}_2.$$

THEOREM: Let W be a subspace of $\mathbb{R}^n$, and let $\vec{v}_1, \ldots, \vec{v}_d$ be an orthonormal basis for W. Then the projection of $\vec{x}$ onto W is given by

$$(\vec{x} \cdot \vec{v}_1)\vec{v}_1 + (\vec{x} \cdot \vec{v}_2)\vec{v}_2 + \ldots + \vec{x} \cdot \vec{v}_d)\vec{v}_d.$$

Now you might say–why did I need to do that? I mean, I could have just ground out the projection matrix above and used that to compute my orthogonal projection. Two reasons: first, this is easier to compute. Second, it is amenable to some very interesting generalizations.

If there's time, begin here to talk about the Fourier series.

IV. The Fourier series

This is the beginning of an incredibly wonderful subject called harmonic analysis. It is crucial in engineering–for instance, a proper understanding of Fourier analysis is what keeps resonant vibrations from building up in a skyscraper and ripping it apart (as once famously happened to a bridge in Washington State.)

Main idea: We have a function f on $[0, 2\pi]$, and we would like to approximate it as a linear combination of sine waves and cosine waves:

$$f(x) = a_1 \sin x + a_2 \sin 2x + a_3 \sin 3x + \ldots + b_0 + b_1 \cos x + \ldots$$

(Draw a picture of some random function f.) Amazing insight of Fourier–all functions can be expressed this way!

But let's walk before we run–suppose we try to express our function as a combination of a few cosine waves,

$$f(x) \sim b_0 + b_1 \cos x + b_2 \cos 2x.$$

Well. We could try to use the value of f at some points in order to estimate b_0, b_1, b_2. For instance, use

$$\begin{aligned}
b_0 + b_1 \cos \pi/2 + b_2 \cos \pi/2 &= f(\pi/2); \\
b_0 + b_1 \cos \pi + b_2 \cos \pi &= f(\pi); \\
b_0 + b_1 \cos 3\pi/2 + b_2 \cos 3\pi/2 &= f(3\pi/2); \\
b_0 + b_1 \cos 2\pi + b_2 \cos 2\pi &= f(2\pi).
\end{aligned}$$

And this is an equation, usually unsolvable, of the form

$$A \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \vec{y}$$

for A a 4×3 matrix. OK, so we could find a least squares solution for this. On the other hand, maybe we would get a better approximation if we used more points. We could use 8 points; that would make A an 8×3 matrix. Likewise, 10 points, 100 points... the problem is that the matrices will just get bigger and bigger and more and more unwieldy. And at each point we still may not have captured the "fine structure" of f, if f is very complicated.

One way to put it is this way. If we measure f at a sequence of points $P_1, \ldots, P_r$, then what is the quantity we are trying to minimize? Remember that the *error* in our estimate for f is the vector

$$\begin{bmatrix} f(P_1) - (b_0 + b_1 \cos P_1 + b_2 \cos 2P_1) \\ f(P_2) - (b_0 + b_1 \cos P_2 + b_2 \cos 2P_2) \\ \vdots \\ f(P_r) - (b_0 + b_1 \cos P_r + b_2 \cos 2P_r) \end{bmatrix}.$$

And the thing we're trying to minimize is the square of the length of this error, which is to say the sum of the squares of the coordinates, or

$$[f(P_1) - (b_0 + b_1 \cos P_1 + b_2 \cos 2P_1)]^2 + \ldots + [f(P_r) - (b_0 + b_1 \cos P_r + b_2 \cos 2P_r)]^2$$

And we wonder what happens as r gets larger and larger.

Great idea: That as we sum over more and more points, our sum becomes more and more like an integral.

In other words, we should try to minimize the integral

$$\int_0^{2\pi} [f(x) - (b_0 + b_1 \cos x + b_2 \cos x)]^2 dx.$$