Math 314/814

Topics since the second exam

Note: The final exam covers all three of our topics sheets for the semester.

Chapter 5: Orthogonality.

Length and inner product.

"Norm" means length! In $\mathbb{R}^n$ this is computed as $||x|| = ||(x_1, \ldots, x_n)|| = (x_1^2 + \cdots + x_n^2)^{1/2}$

Basic facts: $||x|| \geq 0$, and $||x|| = 0$ iff $x = \vec{0}$,
$||cu|| = |c| \cdot ||u||$, and $||u + v|| \leq ||u|| + ||v||$ (triangle inequality)

unit vector: the norm of $u/||u||$ is 1; $u/||u||$ is the *unit vector* in the direction of u.

Inner product:
idea: assign a number to a pair of vectors (think: angle between them?)
In $\mathbb{R}^n$, we use the *dot product*: $v = (v_1, \ldots, v_n)$, $w = (w_1, \ldots, w_n)$
$v \bullet w = \langle v, w \rangle = v_1w_1 + \cdots + v_nw_n = v^Tw$

Basic facts:
$\langle v, v \rangle = ||v||^2$ (so $\langle v, v \rangle \geq 0$, and equals 0 iff $v = \vec{0}$)
$\langle v, w \rangle = \langle w, v \rangle$; $\langle cv, w \rangle = \langle v, cw \rangle = c\langle v, w \rangle$ $\langle v_1 + v_2, w \rangle = \langle v_1, w \rangle + \langle v_2, w \rangle$

Orthogonality.

Two vectors are *orthogonal* if their angle is $\pi/2$, i.e., $\langle v, w \rangle$=0. Notation: $v \perp w$. (Also say they are *perpendicular*.)

A collection of vectors $\{v_1, \ldots, v_k\}$ is an *orthogonal set* if $v_i \perp v_j$ for every $i \neq j$. If all of the vectors in an orthogonal set are non-zero, then they are linearly independent.
An *orthogonal basis* for a subspace W is basis for W that is also an orthogonal set. If we have an orthogonal basis $v_1, \ldots, v_k$ for W, then determining the coordinates for a vector $w \in W$ is quick: $w = \Sigma a_iv_i$ for $a_i = \langle w, v_i \rangle / ||v_i||^2$.

A collection of vectors $\{v_1, \ldots, v_k\}$ is an *orthonormal set* (we write "o.n. set") if they are an orthogonal set and $||v_i|| = 1$ for every i. An *orthonormal basis* (o.n. basis) is a basis that is also an orthonormal set. For an o.n. basis for W, the coordinates of $w =\in W$ are even shorter: $w = \Sigma a_iv_i$ for $a_i = \langle w, v_i \rangle$.

Orthogonal Matrices.

We've seen that having a basis consisting of orthonormal vectors can simplify some of our previous calculations. Now we'll see where some of them come from.
An $n \times n$ matrix Q is called **orthogonal** if it's columns form an orthonormal basis for R^n. This means <(ith column of Q),(jth column of Q¿ $= 1$ if $i = j$, and is 0 otherwise . This in turn means that $Q^TQ = I$, which in turn means $Q^T = Q^{-1}$! So an orthogonal matrix is one whose inverse is equal to its own transpose.

Basic facts about an orthogonal matrix Q :
Q is orthogonal $\Leftrightarrow$ for every $v, w \in R^n$, $< Qv, Qw > = < v, w >$.
Q is orthogonal $\Leftrightarrow$ for every $v \in R^n$, $||Qv|| = ||v||$.
If Q is orthogonal, then Q^T is orthogonal. So the <u>rows</u> of Q form an o.n. basis!
If Q is orthogonal, then Q^{-1} is orthogonal.
If Q is orthogonal, then $\det(Q) = \pm 1$.

Orthogonal Complements.

This notion of orthogonal vectors can even be used to reinterpret some of our dearly-held results about systems of linear equations, where all of this stuff began.

Starting with $Ax = 0$, this can be interpreted as saying that $<$(every row of A),$x >$=0, i.e., x is orthogonal to every row of A. This in turn implies that x is orthogonal to every linear combination of rows of A, i.e., x is orthogonal to every vector in the row space of A.
This leads us to introduce a new concept: the **orthogonal complement** of a subspace W in a vector space V, denoted $W^\perp$, is the collection of vectors v with $v \perp w$ for **every** vector $w \in W$. It is not hard to see that these vectors form a subspace of V; the sum of two vectors orthogonal to w, for example, is orthogonal to w, so the sum of two vectors in $W^\perp$ is also in $W^\perp$. The same is true for scalar multiples.

Some basic facts:
For every subspace W, $W \cap W^\perp = \{0\}$ (since anything in both is orthogonal to *itself*, and only the 0-vector has that property).
Finally, $(W^\perp)^\perp = W$; this is because W is contained in $(W^\perp)^\perp$ (a vector in W is orthogonal to every vector that is orthogonal to things in W), and the dimensions of the two spaces are the same.

The importance that this has to systems of equations stems from the following facts:
$\text{null}(A) = \text{row}(A)^\perp$
$\text{row}(A) = \text{null}(A)^\perp$
$\text{col}(A) = \text{null}(A^T)^\perp$

This provides us with yet another (quicker?) way to decide if a system of equations $A\vec{x} = \vec{b}$ is consistent, or rather, for *which* $\vec{b}$ is it consistent; $\vec{b}$ must lie in col(A), i.e., in null$(A^T)^\perp$. So it must be $\perp$ to a basis for null(A^T). So we can compute a basis for null(A^T), $v_1 \dots , v_k$, and use this to check for consistency: $A\vec{x} = \vec{b}$ is consistent $\Leftrightarrow \langle\vec{b}, v_i\rangle = 0$ for every v_i.

And to compute a basis for $W^\perp$: start with a basis for W, writing them as the columns of a matrix A, so $W = \text{col}(A)$, then $W^\perp = \text{col}(A)^\perp = \text{row}(A^T)^\perp = \text{null}(A^T)$, which we know how to compute a basis for!

Or even better: start with a basis for W, writing them as the columns of a matrix A. The row reduce the superaugmented matrix $(A|I) \to (R|E)$ with R in REF or RREF. Then the transposes $V_1, \dots , v_k$ of the rows of E that stand opposite the all-0 rows of R form a basis for $\text{null}(A^T) = W^\perp$! This is because $R = EA$, so $R^T = A^T E^T$, which implies that $A^T v_i = \vec{0}$ for every i. But E is invertible, so E^T is invertible, so the v_i are linearly independent. <u>But</u>! k = thenumber of $\vec{0}$-rows of R = (# of rows of R) - (# of non-$\vec{0}$ rows of R = (# rows of A) - (row rank of A) =(# columns of A^T) - (rank of A^T) = nullity of A^T, so $v_1, \dots v_k$ are in null(A^T), are linearly independent, and there are as many of them as there is for a basis for null(A^T), so they form a basis for null(A^T) !

Orthogonal Projections.

Any vector $v \in V$ can be written, uniquely, as $v = w + w^\perp$, for $w \in W$ and $w^\perp \in W^\perp$; the uniqueness comes from the result above about intersections. That it can be written that way at all comes from orthogonal projections.
We've seen that if $w_1, \dots , w_k$ is an orthogonal basis for a subspace W of $\mathbb{R}^n$, and $w \in W$, then w $= \dfrac{< w_1, w >}{< w_1, w_1 >} w_1 + \dots + \dfrac{< w_k, w >}{< w_{k-1}, w_k >} w_k$
On the other hand, if $v \in V$, we can define the orthogonal projection

$$\text{proj}_W(v) = \frac{< w_1, v >}{< w_1, w_1 >} w_1 + \dots + \frac{< w_k, v >}{< w_k, w_k >} w_k$$

of v into W. This vector is in W, and we can show that $v - \text{proj}_W(v)$ is orthogonal to all of the w_i, so it is orthogonal to every linear combination, i.e., it is orthogonal to every vector in W. So $v - \text{proj}_W(v) = w' \in W^\perp$

In the case that the w_i are not just orthogonal but also *orthnormal*, we can simplify this somewhat: $\text{proj}_W(v) = < w_1, v > w_1 + \cdots + < w_n, v > w_n = (w_1 w_1^T + \cdots + w_n w_n^T)v = Pv$,
where $P = (w_1 w_1^T + \cdots + w_n w_n^T)$ is the **projection matrix** giving us orthogonal projection.

For any subspace $W \subseteq \mathbb{R}^n$, $\dim(W)+\dim(W^\perp) = n = \dim(\mathbb{R}^n)$. Even more, a basis for W and a basis for $W^\perp$ together form a basis for $\mathbb{R}^n$.

All that we need now is a <u>method</u> for building orthogonal bases for subspaces! (See also the formulation for the orthogonal projection which requires only a basis for W, later in these notes.)

Gram-Schmidt Orthogonalization.

Given a basis $v_1, \ldots, v_n$ for a subspace W, we can build an orthogonal basis for W by, essentially, repeatedly subtracting from w_i its orthogonal projection onto the span of the orthogonal vectors we have built up to that point.
To do so we repeatedly use the formula

$$(*)\ \text{proj}_{W_i}(v) = \frac{< w_1, v >}{< w_1, w_1 >} w_1 + \ldots + \frac{< w_i, v >}{< w_i, w_i >} w_i$$

for the projection of a vectors onto the span W_i of a collection of orthogonal vectors. Gram-Schmidt orthogonalization consists of repeatedly using this formula to replace a collection of vectors with ones that are orthogonal to one another, **without changing their span**. Starting with a collection $\{v_1, \ldots, v_n\}$ of vectors in V,

let $w_1 = v_1$, then let $w_2 = v_2 - \dfrac{< w_1, v_2 >}{< w_1, w_1 >} w_1$.
Then w_1 and w_2 are orthogonal, and since w_2 is a linear combination of $w_1 = v_1$ and v_2, while the above equation can also be rewritten to give v_2 as a linaear combination of w_1 and w_2, the span is unchanged. Continuing,
let $w_3 = v_3 - \dfrac{< w_1, v_3 >}{< w_1, w_1 >} w_1 - \dfrac{< w_2, v_3 >}{< w_2, w_2 >} w_2$; then since w_1 and w_2 are orthogonal, it is not hard to check that w_3 is orthogonal to **both** of them, and using the same argument, the span is unchanged (in this case, $\text{span}\{w_1, w_2, w_3\} = \text{span}\{w_1, w_2, v_3\} = \text{span}\{v_1, v_2, v_3\}$).
Continuing this, we let $w_k = v_k - \dfrac{< w_1, v_k >}{< w_1, w_1 >} w_1 - \ldots - \dfrac{< w_{k-1}, v_k >}{< w_{k-1}, w_{k-1} >} w_{k-1}$
Doing this all the way to n will replace $v_1, \ldots, v_n$ with orthogonal vectors $w_1, \ldots, w_n$, without changing the span.

One thing worth noting is that the if two vectors are orthogonal, then any scalar multiples of them are, too. This means that if the coordinates of one of our w_k are not to our satisfaction (having an ugly denomenator, perhaps), we can scale it to change the coordinates to something more pleasant. It is interesting to note that in so doing, the the later vectors w_k are unchanged, since our scalar, can be pulled out of both the top inner product and the bottom one in later calculations, and cancelled.

Once we know how to build an orthogonal basis for a subspace W, we know how to compute the orthogonal projection of a vector onto W; we use the formula (*) above. This in turn allows to compute the decomposition of any vector $\vec{v} \in \mathbb{R}^n$ as $\vec{v} = \vec{w} + \vec{w}'$ with $\vec{w} \in W$ and $\vec{w}' \in W^\perp$; $\vec{w} = \text{proj}_W(\vec{v})$ and $\vec{w}' = \vec{v} - \vec{w}$.

This in turn gives us the tools to establish some basic facts:

$(W^\perp)^\perp = W$, since if $\vec{w} \in W$, then $\vec{w} \perp \vec{v}$ for every $\vec{v} \in W^\perp$, so $w \in (W^\perp)^\perp$, while if $\vec{v} \in (W^\perp)^\perp$, then writing $\vec{v} = \vec{w} + \vec{w}'$ as above, we have $\vec{v} - \vec{w} \in (W^\perp)^\perp$, so $0 = <\vec{v} - \vec{w}, \vec{w}'> = <\vec{v} - \vec{w}, \vec{v} - \vec{w}> = ||\vec{v} - \vec{w}||^2$, so $\vec{v} - \vec{w} = \vec{0}$, so $\vec{v} = \vec{w} \in W$.

If $W \subseteq \mathbb{R}^n$ is a subspace, then $\dim(W) + \dim(W^\perp) = n$; this is because we can express $W = \mathrm{col}(A)$ for some matrix A (having columns a spanning set for W), and then $W^\perp = \mathrm{null}(A^T)$, so $\dim(W) + \dim(W^\perp) = \mathrm{rank}(A) + \mathrm{nullity}(A^T) = \mathrm{rowrank}(A) + \mathrm{nullity}(A^T) = \mathrm{rank}(A^T) + \mathrm{nullity}(A^T) = (\#$ of pivots for $A^T) + (\#$ free variables for $A^T) = \#$ of columns of $A^T = \#$ of rows of $A = n$.

Even more, any basis for W, together with a basis for $W^\perp$, forms a basis for $\mathbb{R}^n$. This is becuse such a collection of vectors will form n vectors in $\mathbb{R}^n$ and will be linearly independent. This is because if we express $\vec{0}$ as a linear combination of them, we can rearrange terms so that a linear combination of the W-vectors equals a linear combination of the $W^\perp$-vectors. This vector lies in $W \cap W^\perp = \{\vec{0}\}$, so since each basis is linearly independent, both sets of coefficients are 0.

Best approximations.

In the real world, the coefficients and target vector of a system of linear equations are only known up to some (measurement) error. But if the rank of the matrix is too small (e.g., we have more variables than equations), small changes in values can easily lead to an inconsistent system. In other words, our target, $\vec{b}$ might end up lying close to, but not in, the column space $\mathrm{col}(A)$, of our coefficient matrix. The appropriate solution, then, is to find the value of $A\vec{x}$, closest to $\vec{b}$, and treat $\vec{x}$ as our "solution" to the inconsisent system $A\vec{x} = \vec{b}$.

How? Minimize $||A\vec{x} - \vec{b}||^2$, i.e., minimize $\vec{w} - \vec{b}$ for $\vec{w} \in \mathrm{col}(A)$. If we use an orthonormal basis $\{\vec{w}_1, \dots, \vec{w}_k\}$ for $\mathrm{col}(A)$, then $< (\Sigma x_i \vec{w}_i) - \vec{b}, (\Sigma x_i \vec{w}_i) - \vec{b} > = < \vec{b}, \vec{b} > + \Sigma(x_i^2 - 2x_i < \vec{w}_i, \vec{b} >)$ is minimized when (the gradient of this function of the x_i is 0, i.e.) $x_i = < \vec{w}_i, \vec{b} >$ for each i, so the vector $\vec{w}$ closest to $\vec{b}$ is $\Sigma < \vec{w}_i, \vec{b} > \vec{w}_i = \mathrm{proj}_{\mathrm{col}(A)}(\vec{b})$, i.e., the orthogonal projection of $\vec{b}$ to the column space of A.

But! we don't need to build an orthogonal basis for $\mathrm{col}(A)$ in order to compute this; $\vec{w} = \mathrm{proj}_{\mathrm{col}(A)}(\vec{b})$ is the (unique) vector $\vec{w} \in \mathrm{col}(A)$ such that $\vec{w} - \vec{b} \in (\mathrm{col}(A))^\perp = \mathrm{null}(A^T)$, so we need to find $\vec{w} = A\vec{x}$ so that $A^T(A\vec{x} - \vec{b}) = \vec{0}$, i.e., $(A^TA)\vec{x} = A^T\vec{b}$.

This linear system is consistent (we know that the needed $A\vec{x}$ exists); solving the system for $\vec{x}$ gives us the vector $A\vec{x} = \mathrm{proj}_{\mathrm{col}(A)}(\vec{b})$, and so gives us a method for computing the orthogonal projection onto any subspace (that we have a spanning set for), without needing to compute an orthogonal basis for it first. $A\vec{x} = \vec{w}$ is also the closest vector to $\vec{b}$ for which $A\vec{x} = \vec{w}$ is consistent; $\vec{x}$ is called the *least squares solution* to the inconsistent system $A\vec{x} = \vec{b}$.

Note: if A^TA is invertible (need: $r(A)$=number of columnsof A), then we can write $\vec{x} = (A^TA)^{-1}(A^T\vec{b})$; $A\vec{x} = A(A^TA)^{-1}(A^T\vec{b})$. The gives us a general formula for the orthogonal projection onto a subspace W; $\mathrm{proj}_W(v) = A(A^TA)^{-1}(A^T\vec{v})$, where the columns of A form a basis for W.

Regression Lines.

We can apply this technology to produce a method for finding the "best fit" line to a collection of data. Suppose we have a collection $(x_1, y_1), \dots, (x_n, y_n)$ of data points, and we wish to find the line $L(x) = y = ax + b$ that best fits the data. Typically, this means that we want, on average, that the deviation between y_i and $L(x_i)$ to be as small as possible. In practice, what we minimize is the distance between the "value vector", $[y_1, \dots, y_n]^T$ and the "predicted vector" $[ax_1 + b, \dots, ax_n + b]^T$. Our unknowns are a, b, and our predicted vector can be expressed as a matrix product,

$$\begin{bmatrix} ax_1+b \\ \vdots \\ ax_n+b \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = A \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \vec{y}$$

So we want the vector $[a,b]^T$ so that $A[a,b]^T$ is closest to $\vec{y}$. But this is precisely the situation we just worked through; the slope (a) and intercept (b) of the best-fitting line are the solution to the system $A^TA\begin{bmatrix} a \\ b \end{bmatrix} = A^T\vec{y}$, which works out to $\begin{bmatrix} \Sigma x_i^2 & \Sigma x_i \\ \Sigma x_i & n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \Sigma x_i y_i \\ \Sigma y_i \end{bmatrix}$ (although we need not remember that....). The 2×2 matrix A^TA is invertible, unless all of the x_i are equal to one another.

We can do this sort of thing more generally, too. We can find the best-fitting quadratic $Q(x) = y = ax^2+bx+c$, or cubic $C(x) = y = ax^3+bx^2+cx+d$, or any polynomial, using the same basic approach. Let's illustrate this with a quadratic. As with linear regression, we wish to make the sum of the terms $[y_i - (ax_i^2+bx_i+c)]^2$ as small as possible, which means that we wish to make $A\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ as close to $\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \vec{y}$ as we can, where now

$A = \begin{bmatrix} x_1^2 & x_1 & 1 \\ \vdots & & \vdots \\ x_n^2 & x_n & 1 \end{bmatrix}$. This, again, has solution $\begin{bmatrix} a \\ b \\ c \end{bmatrix} = (A^TA)^{-1}A^T\vec{y}$. A^TA is invertible so long as at least three of the x_i are distinct. This last fact follows from a fact about the Vandermonde determinant, which is the determinant of the matrix

$$\begin{bmatrix} x_1^{n-1} & \dots & x_1 & 1 \\ \vdots & & \vdots & \\ x_n^{n-1} & \dots & x_n & 1 \end{bmatrix},$$

and which equals the product of all of the differences $x_i - x_j$ taken over pairs $i < j$. The determinant is therefore non-zero if all of the x_i are distinct, which means that the last $k+1$ columns (which are what we use for a degree k polynomial fitting) are linearly independent if at least $k+1$ of the x_i are distinct.