Math 314/814 Section 5

Topics since the second exam

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 $w_2 = v_2 - \frac{< w_1, v_2 >}{< w_1, w_1 >} w_1$.
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 $w_3 = v_3 - \frac{< w_1, v_3 >}{< w_1, w_1 >} w_1 - \frac{< w_2, v_3 >}{< w_2, w_2 >} w_2$; 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, span$\{w_1, w_2, w_3\}$ =span$\{w_1, w_2, v_3\}$=span$\{v_1, v_2, v_3\}$).
Continuing this, we let $w_k = v_k - \frac{< w_1, v_k >}{< w_1, w_1 >} w_1 - \ldots - \frac{< w_{k-1}, v_k >}{< w_{k-1}, w_{k-1} >} w_{k-1}$
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 $W = \text{col}(A)$ for some matrix A (having columns a spanning set for W), and then $W^\perp = \text{null}(A^T)$, so $\dim(W)+\dim(W^\perp) = \text{rank}(A)+\text{nullity}(A^T) = \text{rowrank}(A)+\text{nullity}(A^T) = \text{rank}(A^T)+\text{nullity}(A^T)$ =(# 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 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 $\vec{w} \in \text{col}(A)$. If we use an orthonormal basis $\{\vec{w}_1, \dots, \vec{w}_k\}$ for 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 $\Sigma < \vec{w}_i, \vec{b} > \vec{w}_i = \text{proj}_{\text{col}(A)}(\vec{b})$, 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 col(A) in order to compute this; $\vec{w} = \text{proj}_{\text{col}(A)}(\vec{b})$ is the (unique) vector $\vec{w} \in \text{col}(A)$ such that $\vec{w} - \vec{b} \in (\text{col}(A))^\perp = \text{null}(A^T)$, 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 $A\vec{x} = \text{proj}_{\text{col}(A)}(\vec{b})$, 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 $\vec{x} = (A^TA)^{-1}(A^T\vec{b})$; $A\vec{x} = A(A^TA)^{-1}(A^T\vec{b})$.

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.

Symmetric matrices and Orthogonal Diagonalization.

A symmetric matrix $A$is one for which $A^T = A$. It is a fundamental fact (the "Spectral Theorem") that every symmetric matrix is diagonalizable, and in fact in a special way. This follows from the following facts:

Every eigenvalue of a symmetric matrix A is real. That is, the characteristic $\chi_A(\lambda)$ factors as a product of linear polynomials with coefficients in $\mathbb{R}$. This can be shown by supposing an eigenvalue λ_0 is complex, finding a complex eigenvector $\vec{v}$, and using $A^T = A$ to show that (where $\overline{a+bi} = a - bi$ means the complex conjugate)
$\lambda_0||\vec{v}|| = \lambda_0 < \vec{v}, \overline{\vec{v}} >=< \lambda_0\vec{v}, \overline{\vec{v}} >< A\vec{v}, \overline{\vec{v}} >= (A\vec{v})^T\overline{\vec{v}} = \vec{v}^T A^T \overline{\vec{v}}\vec{v}^T A\overline{\vec{v}} = \vec{v}^T\overline{A}\overline{\vec{v}} = \vec{v}^T\overline{A\vec{v}} = \vec{v}^T\overline{\lambda_0\vec{v}} = \vec{v}^T\overline{\lambda_0}\overline{\vec{v}} = \overline{\lambda_0}\vec{v}^T\overline{\vec{v}} = \overline{\lambda_0}||\vec{v}||$, so $\lambda_0 = \overline{\lambda_0}$, so λ_0 is real.

If $\vec{v}, \vec{w}$ are eigenectors for A with different eigenvalues $\lambda \neq \mu$, then $\vec{v} \perp \vec{w}$. This is because $\lambda < \vec{v}, \vec{w} >=< \lambda\vec{v}, \vec{w} >=< A\vec{v}, \vec{w} >= (A\vec{v})^T\vec{w} = \vec{v}^T A^T \vec{w} = \vec{v}^T A\vec{w} = \vec{v}^T(A\vec{w}) = \vec{v}^T(\mu\vec{w}) = \mu(\vec{v}^T\vec{w}) = \mu < \vec{v}, \vec{w} >$, so $(\lambda - \mu) < \vec{v}, \vec{w} >= 0$, so $< \vec{v}, \vec{w} >= 0$, since $\lambda - \mu \neq 0$.

If B is a symmmetric matrix and Q is an *orthogonal* matrix (so $Q^{-1} = Q^T$), then $A = Q^{-1}BQ$ is symmetric. This is because $A = Q^{-1}BQ = Q^TBQ = Q^TB^T(Q^T)^T = (Q^TBQ)^T = (Q^{-1}BQ)^T = A^T$. So <u>if</u> a matrix A has an orthonormal basis of eigenvectors, then $AQ = QD$ with D diagonal and Q orthogonal, so $A = (Q^{-1})^{-1}D(Q^{-1})$ is symmetric. But the <u>important</u> point is that the reverse is true!

To see this, we pick an eigenvalue (real) λ for A and an eigenvector with length 1, and extend the eigenvector to a orthonormal basis for $\mathbb{R}^n$ (by extending it to a basis and applying Gram-Schmidt). Assembling these into a matrix B, we then have that the first column of AB is *lambda* times the first column of B, so the first column of $C = B^{-1}AB$ is $\lambda\vec{e}_1$. But! C is also symmetric, so the first <u>row</u> of C is λ followed by 0s. This means that C looks like the beginnings of a diagonal matrix, the first row and column are right, and the rest is a (smaller!) diagonal matrix. The basic idea is to do the exact same thing to the smaller matrix; find an eigenvalue and eigenvector to extend to an orthonormal basis of $\mathbb{R}^{n-1}$ (which we think of as the last $n-1$ coordinates of $\mathbb{R}^n$) and "diagonalize" the next row. Essentially, we are diagonalizing the <u>next</u> row an column of C, by using the orthogonal matrix with columns e_1 and the $n-1$ vectors this step built. It's the end of the semester, so we will leave further details to your imagination (or reading), but applying this n times yields n orthogonal matrices $Q_1, \ldots, Q_n$ so that $Q_n^T \cdots Q_1^T AQ_1 \cdots Q_n = (Q_1 \cdots Q_n)^T A(Q_1 \cdots Q_n)$ is a diagonal matrix; but $Q_1 \cdots Q_n$ is a product of orthogonal matrices, so is orthogonal (each preserves length $||\cdot||$, so their product (i.e., composistion) does) establishng that A is *orthogonally diagonalizable*, as desired.