Lecture 11: 3.3 Projections and Least-Square Problems. (The material below only took half a lecture so in the future I will do it with section 3.2.)

Recall that we last time derived the matrix for the orthogonal projection onto a subspace W. If W is spanned by a set of linearly independent vectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ we formed the matrix $A = [\,\mathbf{a}_1 \,\dots\, \mathbf{a}_k\,]$ with these vectors as columns. We showed that the projection is given by

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

In fact, if $\mathbf{b} \in C(A)$ then we can write $\mathbf{b} = A\mathbf{x}$ and hence

$$P\,\mathbf{b} = A(A^T A)^{-1} A^T A\mathbf{x} = A\mathbf{x} = \mathbf{b}$$

and if $\mathbf{b}$ is in the orthogonal complement $C(A)^{\perp} = N(A^T)$ then

$$P\mathbf{b} = A(A^T A)A^T\mathbf{b} = 0$$

In order for P to be defined we assume that $A^T A$ is invertible. This however follows from that the columns of A are linearly independent. In fact it suffices to prove that the nullspace of $A^T A$ is empty. Suppose that $A^T A\mathbf{x} = 0$ and take the dot product of this with $\mathbf{x}$ to get

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

so $A\mathbf{x} = 0$, which implies that $\mathbf{x} = 0$ since the columns of A are independent.

Next we want to show that the orthogonal projection $\mathbf{p} = P\mathbf{b}$ of $\mathbf{b}$ onto W is the point in W that is closest to $\mathbf{b}$. In fact if $\mathbf{v} \in W$ then we can write $\mathbf{b} - \mathbf{v} = \mathbf{b} - \mathbf{p} + \mathbf{p} - \mathbf{v}$. Since $\mathbf{b} - \mathbf{p}$ is orthogonal to any vector in W and since $\mathbf{p} - \mathbf{v}$ is in W it follows that

$$\|\mathbf{b} - \mathbf{v}\|^2 = \|\mathbf{b} - \mathbf{p}\|^2 + \|\mathbf{p} - \mathbf{v}\|^2 > \|\mathbf{b} - \mathbf{p}\|^2, \qquad \text{if} \quad \mathbf{v} \neq \mathbf{p}.$$

Least Square Problems.

A standard statistical technique called regression is to find a least square fit to data points in the by some simple curve e.g. a line. Since there might be errors in the measurements of the data we do not require the curve to pass through the points but instead be such that it is the optimal approximation to the data in the sense that the sum of squares of the error between the y values of the data points and the points on the curve should be minimized.

Suppose we have a set of data measurements (t_i, y_i), $i = 1, 2, 3$, $(1, 1)$, $(2, 2)$ and $(3, 2)$, of some quantities (t, y), that we want to approximate with some line $y = C + Dt$. The errors are defined by $e_i = C + Dt_i - y_i$, $i = 1, 2, 3$. We want to find the line that makes the total; error as small as possible. More specifically, we want to choose C and D so as to minimize the total error $E^2 = e_1^2 + e_2^2 + e_3^2$.

This can be formulated as a system:

$$A\mathbf{x} - \mathbf{b} = \mathbf{e}$$

where

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} C \\ D \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

A least square problem may be formulated as an over determined linear system. A system with more equations than unknowns usually is inconsistent. In general we can not solve a system $A\mathbf{x}{=}\mathbf{b}$, where A is an $m{\times}n$ matrix with $m{>}n$. Instead we want to find an $\mathbf{x}$ that makes $\|A\mathbf{x} - \mathbf{b}\|$ as small as possible. This is called the least square solution. By the previous section $\mathbf{x}$ is found by first orthogonally projecting $\mathbf{b}$ to the column space $\mathbf{p} = P\,\mathbf{b}$ and then solving the system $A\mathbf{x} = \mathbf{p}$.

Def The **least square solution** $\mathbf{x}$ of the system $A\mathbf{x} = \mathbf{b}$ is a vector such that

$$\|A\mathbf{x} - \mathbf{b}\| \le \|A\mathbf{y} - \mathbf{b}\|, \qquad \text{for all} \quad \mathbf{y} \in \mathbf{R}^n$$

Formulated differently, we want to find the vector $\mathbf{p} \in C(A)$ that is closest to $\mathbf{b}$. From the previous section we know that $\mathbf{p} = P\mathbf{b}$, is the orthogonal projection of $\mathbf{b}$ onto $C(A)$. Since $\mathbf{p} \in C(A)$, there is a vector $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{p}$.

We have now shown that there is a unique vector $\mathbf{p} \in C(A)$ that is closest to $\mathbf{b} \in \mathbf{R}^m$ and this is half the solution of the problem but it still remains to find $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{p}$. We also showed that $\mathbf{e} = \mathbf{b} - \mathbf{p}$ is orthogonal $C(A)$, i.e. in $C(A)^\perp$. But recall that $C(A)^\perp = N(A^T)$. Hence $\mathbf{e} = \mathbf{b} - A\mathbf{x} \in N(A^T)$, i.e.

$$A^T A\mathbf{x} = A^T\mathbf{b}$$

This so called **normal equation** represents an $n \times n$ system. We have:

Th If A is an $m \times n$ matrix of rank n then (5.3.1) has a unique solution

$$\mathbf{x} = (A^T A)^{-1} A^T \mathbf{b}$$

and $\mathbf{x}$ is the least square solution of the problem $A\mathbf{x} = \mathbf{b}$.

Note also that we can use the solution of the normal equation to construct the orthogonal projection onto the subspace spanned by the column vectors of A:

$$\mathbf{p} = A\mathbf{x} = A(A^T A)^{-1} A^T \mathbf{b}$$

Ex Find the least square solution to $A\mathbf{x}=\mathbf{b}$ where $A=\begin{bmatrix}1&1\\1&2\\1&3\end{bmatrix}$, $\mathbf{b}=\begin{bmatrix}1\\2\\2\end{bmatrix}$.

$$A^TA=\begin{bmatrix}1&1&1\\1&2&3\end{bmatrix}\begin{bmatrix}1&1\\1&2\\1&3\end{bmatrix}=\begin{bmatrix}3&6\\6&14\end{bmatrix}$$

and

$$(A^TA)^{-1}=\frac{1}{6}\begin{bmatrix}14&-6\\-6&3\end{bmatrix}$$

and

$$A^T\mathbf{b}=\begin{bmatrix}1&1&1\\1&2&3\end{bmatrix}\begin{bmatrix}1\\2\\2\end{bmatrix}=\begin{bmatrix}5\\11\end{bmatrix}$$

Hence

$$\mathbf{x}=(A^TA)^{-1}A^T\mathbf{b}=\frac{1}{6}\begin{bmatrix}14&-6\\-6&3\end{bmatrix}\begin{bmatrix}5\\11\end{bmatrix}=\frac{1}{6}\begin{bmatrix}4\\3\end{bmatrix}$$