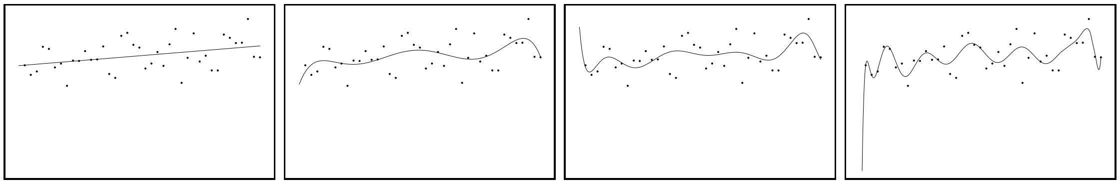
Least Squares 3/12/2001 Math 21b, O. Knill

Homework for Wednesday, March 14, 2001: Section 4.4, No 2,4,10,22,26,32

GOAL. We want to find solutions of linear systems $Ax = b$ in the case, when we have no consistent system, that is in the case, when we have no solution. The best possible solution will be called the **least square solution**. It is the orthogonal projection of b onto the image $\mathrm{im}(A)$ of A.

We will see in this lesson that some theory about the kernel and the image of linear transformations helps to understand the situation and gives us an explicit formula for the least square fit.

Why do we care about non-consistent systems? Often we have to solve linear systems of equations with more constraints then variables. An example is when we try to find the best polynomial which passes through a bunch of points. This is called **data fitting**. If we wanted to accommodate all data, the polynomial would become too large and wiggly. Taking a smaller degree polynomial will not only be more convenient but also give a better picture. In statistics for example one wants to fit data with lines (regression).



The above pictures show 30 data points which are fitted best with polynomials of degree 1, 6, 11 and 16. The first linear fit maybe tells most about the trend of the data.

THE ORTHOGONAL COMPLEMENT OF $\mathrm{im}(A)$. Because a vector is in the kernel of A^T if and only if it is orthogonal to the rows of A^T and so to the columns of A, the kernel of A^T is the orthogonal complement of $\mathrm{im}(A)$:

$$(\mathrm{im}(A))^{\perp} = \ker(A^T) \; .$$

EXAMPLES.

1) $A = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$. The kernel V of $A^T = \begin{bmatrix} a & b & c \end{bmatrix}$ consists of all vectors satisfying $ax+by+cz = 0$. V is a plane. The orthogonal complement is the image of A which is indeed spanned by the normal vector $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ to the plane.

2) $A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$. The image of A is spanned by $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ the kernel of A^T is spanned by $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

ORTHOGONAL PROJECTION. If b is a vector and V is a linear subspace, then $\mathrm{proj}_V(b)$ is the vector closest to b on V: given any other vector v on V, one can form the triangle $b, v, \mathrm{proj}_V(b)$ which has a right angle at $\mathrm{proj}_V(b)$ and invoke Pythagoras.

LEAST SQUARE SOLUTION. The **least square solution** of $Ax = b$ is the vector x^* such that Ax^* is closest to b from all other vectors Ax.

In other words, $x^* = \mathrm{proj}_V(b)$. Because $b - Ax^*$ is in $V^{\perp} = \mathrm{im}(A)^{\perp} = \ker(A^T)$, we have $A^T(b - Ax^*) = 0$. The last equation means that x^* is a solution of $A^T Ax = A^T b$, which is called the **normal equation** of $Ax = b$.
If the kernel of A is trivial, then $A^T A$ can be inverted and $x^* = (A^T A)^{-1} A^T b$ is the least square solution.

ORTHOGONAL PROJECTION If $v_1, \dots, v_n$ is a basis in V which is not orthonormal, then the orthogonal projection is $x \mapsto A(A^TA)^{-1}A^T(x)$, where $A = [w_1, \dots, w_n]$.
Proof. $x = (A^TA)^{-1}A^Tb$ is the least square solution of $Ax = b$. Therefore $Ax = A(A^TA)^{-1}A^Tb$ is the vector in $\mathrm{im}(A)$ closest to b.

Special case: If $w_1, \dots, w_n$ is a basis in V, we had seen earlier that AA^T with $A = [w_1, \dots, w_n]$ is the orthogonal projection onto V (this was just rewriting $Ax = (w_1 \cdot x)w_1 + \dots + (w_n \cdot x)w_n$ in matrix form.) This follows also from the above more general formula because $A^TA = 1$ in that case.

PROBLEM (see problem 6 in 4.4) Does the formula $\mathrm{im}(A) = \mathrm{im}(AA^T)$ hold? In the case, when the columns in A are orthonormal, then AA^T is the projection onto the image of A. The images are clearly the same. The image of AA^T is contained in the image of A because we can write $v = AA^Tx$ as $v = Ay$ with $y = A^Tx$. On the other hand, if v is in the image of A, then $v = Ax$. If $x = y + z$ where y in the kernel of A and z orthogonal to the kernel of A, then $Ax = Az$. Because z is orthogonal to the kernel of A, it is in the image of A^T. (You prove the formula

$$(\ker(A))^{\perp} = \mathrm{im}(A^T) \ .$$

in the homework.) Therefore, $z = A^Tu$ and $v = Az = AA^Tu$ is in the image of AA^T.

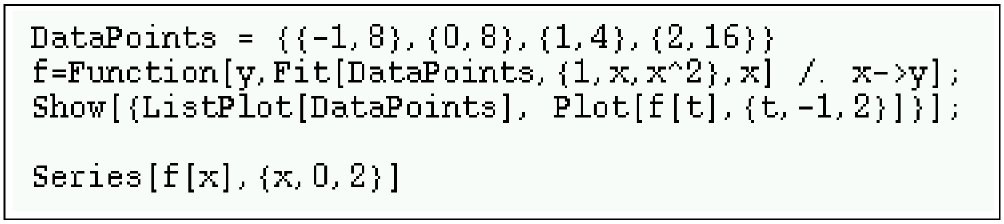
EXAMPLE Let $A = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$. The orthogonal projection onto $V = \mathrm{im}(A)$ is $b \mapsto A(A^TA)^{-1}A^Tb$. We have $A^TA = \begin{bmatrix} 5 & 0 \\ 2 & 1 \end{bmatrix}$ and $A(A^TA)^{-1}A^T = \begin{bmatrix} 1/5 & 2/5 & 0 \\ 2/5 & 4/5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.
For example, the projection of $b = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ is $x^* = \begin{bmatrix} 2/5 \\ 4/5 \\ 0 \end{bmatrix}$ and the distance to b is $1/\sqrt{5}$. The point x^* is the point on V which is closest to b.

Remember the formula for the distance of b to a plane V with normal vector n? It was $d = |n \cdot b|/||n||$. In our case, we can take $n = [-2, 1, 0]$ this formula gives the distance $1/\sqrt{5}$. Let's check: the distance of x^* and b is $||(2/5, -1/5, 0)|| = 1/\sqrt{5}$.

EXAMPLE. Let $A = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \end{bmatrix}$. Problem: find the matrix of the orthogonal projection onto the image of A. The image of A is a one-dimensional line spanned by the vector $v = (1, 2, 0, 1)$. We calculate $A^TA = 6$. Then

$$A(A^TA)^{-1}A^T = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 & 1 \end{bmatrix} /6 = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 2 & 4 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix} /6$$

DATA FIT. (see p. 226 in the book) Find a quadratic polynomial which best fits the data points. $(-1, 8), (0, 8), (1, 4), 2, 16)$. Mathematica has already built in the facility to fit such a thing:



```
DataPoints = {{-1,8},{0,8},{1,4},{2,16}}
f=Function[y,Fit[DataPoints,{1,x,x^2},x] /. x->y];
Show[{ListPlot[DataPoints], Plot[f[t],{t,-1,2}]}];

Series[f[x],{x,0,2}]
```

The series expansion of f showed that indeed, $f(t) = 5 - t + 3t^2$ is indeed best quadratic fit. Actually, Mathematica does the same to find the fit then what we do: **”Solving” an inconsistent system of linear equations as best as possible.**