Least Squares Again

This document will give another discussion of least squares. We are interested in attempting to solve the linear equation

$$Ax = b,$$

where A is an $m \times n$ matrix. There may not be a solution, so we try to find x that minimizes $\|Ax - b\|_2$. We have to prove that there is such an x and characterize it. Let $A = [a_{ij}]$ and

$$\begin{aligned} f(x) &= \sum_{i=1}^{m} \left[\sum_{l=1}^{n} a_{il} x_l - b_i \right]^2 \\ &= \|Ax - b\|_2^2 \\ &= x^T A^T A x - 2 x^T A^T b + \|b\|_2^2 . \end{aligned}$$

Let's compute.

$$f_{x_j} = 2 \sum_{i=1}^{m} \left[\sum_{l=1}^{n} a_{il} x_l - b_i \right] a_{ij}, \ j = 1 \ldots n.$$

Rewrite this as

$$Df = 2(A^T A x - A^T b).$$

A critical point x_0 (if there is a critical point) is a point x_0 such that

$$A^T A x_0 = A^T b.$$

Why is there such a point? It's the Fredholm alternative. Some people call it the *fundamental theorem of linear algebra.* This is one statement of it:

Theorem 1. *The linear equation $Lx = y$ is solvable exactly when $v^t A = 0$ implies $v^t y = 0$.*

Proof. In our case, if $v^T A^T A = 0$ then $v^T A^T A v = \left\|v^T A^T\right\|_2^2 = 0$. So $v^T A^T = 0$ and hence $v^T A^T b = 0$. This implies there will always be a critical point x_0. There may be infinitely many critical points. At each one of them f takes the same value, $f(x_0)$, and this value is the global minimum of f. It's instructive to prove by computing the Hessian that f assumes a relative minimum at each critical point. We will prove by a direct calculation that $f(x_0) \leq f(x)$ for all x. First we compute the Hessian.

$$\begin{aligned} (f_{x_j})_{x_k} &= 2 \frac{\partial}{\partial x_k} (\sum_i \sum_l a_{il} a_{ij} x_l) \\ &= 2 (\sum_i \sum_l a_{il} a_{ij} \delta_{lk}) \\ &= 2 \sum_i a_{ik} a_{ij} \\ &= 2 [A^T A]_{kj}, \end{aligned}$$

where $\delta_{kj} = 1$ if $k = j$ and 0 otherwise. Now $A^T A$ is positive semidefinite and doesn't depend on the critical point (it is constant), so each critical point is a local minimum (maybe not strict). Now we prove that the value at each critical point is the same, and is a global minimum.

$$\begin{aligned}
f(x) &= f(x_0 + x - x_0) \\
&= \|A(x - x_0) + Ax_0 - b\|^2 \\
&= \|Ax_0 - b\|^2 + 2\langle Ax_0 - b, A(x - x_0)\rangle + \|A(x - x_0)\|^2 \\
&= \|Ax_0 - b\|^2 + \|A(x - x_0)\|^2 + \langle A^T(Ax_0 - b), x - x_0\rangle \\
&= \|Ax_0 - b\|^2 + \|A(x - x_0)\|^2 \\
&\geq \|Ax_0 - b\|^2 = f(x_0).
\end{aligned}$$

This is true for any x, so if we take another critical point x_1 we find that $f(x_1) \geq f(x_0)$. The argument is symmetric, so all critical points have the same critical value. This argument also proves that all critical points differ by a vector x such that $Ax = 0$. If the kernel of A is nontrivial there is an entire affine subspace of critical points. □