1 The least squares problem

Given an $m \times n$ matrix A, we want to solve

$$x_* = arg\min_x ||Ax - b||_2^2.$$

Here, $arg\min_x F(x)$ is "the vector x minimizing F", to be distinguished from $\min_x F(x)$, which is the *objective value* at the minimum, that is, $\min_x F(x) = F(x_*)$, where $x_* = arg\min_x F(x)$. So the problem $arg\min_x ||Ax - b||_2^2$ is to find a vector x as close as possible to a solution to the equations in the least squares sense.

2 The normal equations

The normal equations for the least squares problem are given by

$$x_* = arg\min_x ||Ax - b||_2^2 \iff A^T A x_* = A^T b;$$

the fastest route to these equations is via taking the gradient. Since we do not assume knowledge of calculus here, we will take a different approach.

3 A solution

For now, assume A has n independent columns. We will use the fact that we can write

$$A = QR,$$

where Q is an $m \times n$ matrix with orthogonal rows, and R is an $n \times n$ upper triangular invertible matrix (Q and R are obtained by Gram-Schmidt applied to A).

First note that we can write

$$b = b_1 + b_2,$$

with $b_1 = QQ^T b$, and $b_2 = b - QQ^T b$. Since $Q^T Q = I_n$,

$$Q^T b_1 = Q^T Q Q^T b = Q^T b;$$

but on the other hand,

$$Q^T b_2 = Q^T(b - QQ^T b) = Q^T b - Q^T Q Q^T b = Q^T b - Q^T b = 0.$$

Thus for any $x \in \mathbb{R}^n$,

$$\langle Ax, b_2 \rangle = x^T A^T b_2 = x^T R^T Q^T b_2 = 0,$$

and

$$\langle b_1, b_2 \rangle = b^T Q Q^T b_2 = 0.$$

So

$$\begin{aligned} arg\min_x ||Ax - b||_2^2 &= arg\min_x ||(Ax - b_1) - b_2||_2^2 \\ &= arg\min_x ||Ax - b_1||_2^2 + ||b_2||_2^2 - 2\langle Ax - b_1, b_2 \rangle \\ &= arg\min_x ||Ax - b_1||_2^2 + ||b_2||_2^2 \\ &= arg\min_x ||Ax - b_1||_2^2 \end{aligned}$$

because $||b_2||_2^2$ is independent of x; that is, adding $||b_2||_2^2$ changes the objective value at the minimum, but does not change the $arg\min$. Now make the substitution $a = Rx$:

$$arg\min_x ||Ax - b||_2^2 = arg\min_a ||Qa - b_1||_2^2.$$

Since $b_1 = Q(Q^T b)$, $arg\min_a ||Qa - b_1||_2^2$ has solution given by $a = Q^T b$ (with objective value $||Qa - b_1||_2^2 = 0$), and so $arg\min_x ||Ax - b_1||_2^2 = R^{-1}Q^T b_1$. Recall $Q^T b_1 = Q^T b$, so

$$arg\min_x ||Ax - b||_2^2 = R^{-1}Q^T b.$$

We have solved the problem with neither calculus nor the normal equations. However, we can recover the normal equations: Since R is invertible, R^T is invertible, and since $A^T Ax = R^T Q^T QRx = R^T Rx$,

$$A^T Ax = A^T b \iff Rx = Q^T b.$$

In fact, the right hand side version of the normal equations is "better", because inverting R is computationally more stable than inverting $A^T A$.

4 Some geometry

The seemingly sneaky factorization of b into $(b - QQ^T b) + (QQ^T b)$ is extremely natural from a geometric standpoint: Q is an orthnormal basis for the column space col(A), $b_1 = QQ^T b$ is the (orthogonal) projection of b into col(A), and so b_2 is the component of b orthogonal to col(A). Thus b_2 is the component of b unreachable by A, and so we can ignore it when looking for x. We will more carefully define these terms (orthogonal projection) in the next few lectures.

5 What if A is not full rank?

If the columns of A are not linearly independent, then Q will be $m \times r$, with $r < n$, and there is a submatrix R_J of R corresponding to the the columns of A that had nonzero residual during the Gram-Schmidt process; denote the indices of these columns by J. In this case, a solution to the problem is given by x, where $x_J = R_J^{-1}Q^T b$, and setting $x_i = 0$ for $i \in \{1, ..., n\} \setminus J$. The general solution is given by $x + N(A)$, where $N(A)$ is the null space of A.

6 Some examples:

Suppose we have scalars $x_1, ...x_m$ and $y_1, ..., y_m$, and we suspect that y can be (approximately) obtained from x via the formula $y = ax + b$ for some a and b. We might then wish to find

$$arg\min_{a,b} \sum_{i=1}^{m} |ax_i + b - y_i|^2.$$

Set

$$M = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{pmatrix}, \; y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \text{ and } u = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Then

$$arg\min_{a,b} \sum_{i=1}^{m} |ax_i + b - y_i|^2 = arg\min_u ||Mu - y||_2^2,$$

which we now know how to solve: write $M = QR$, and set $u = R^{-1}Q^T y$.

If we suspected that y_i was a degree n polynomial of x_i, we could do the same thing:

$$arg\min_a \sum_{i=1}^{m} \left|(a_1 + a_1 x_i + a_2 x_i^2 + ... + a_n x_i^{n-1}) - y_i\right|^2 = arg\min_a ||Ma - y||_2^2,$$

this time with

$$M = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^{n-1} \end{pmatrix}, \text{ and } a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}.$$

Again, the solution is to write $M = QR$, and set $a = R^{-1}Q^T y$.

Now suppose $x_1, ..., x_m$ are d-vectors (but y_i are still scalars). The analagous version of the affine mapping from the first example is to find $b_1, ..., b_d$ and $b_{d+1} = c$ such that $b^T x_i + c \sim y_i$. That is, we want to find

$$arg \min_b \sum_{i=1}^{m} |(b^T x_i + c) - y_i|^2.$$

If we place each x_i in a row of the $m \times (d+1)$ matrix X, and set the last column of X to be the vector of ones, we get the problem $arg \min_b ||Xb - y||^2$. In your homework, you may think of having 64×32 regression problems, each of which has $d = 64 \times 32$ and $m = 12000$.