Week 9: Monday, Oct 19

Rank-deficient problems

At the end of last time, we got as far as describing the minimum l^2 norm solution to the rank-deficient least squares problem. If A has rank r, we write

$$A = U_1 \Sigma_1 V_1^T$$

where $U_1 \in \mathbb{R}^{m \times r}$, $\Sigma_1 \in \mathbb{R}^{r \times r}$, and $V_1 \in \mathbb{R}^{n \times r}$. Then

$$x_{LS} = V_1 \Sigma_1^{-1} U_1^T b$$

is the minimum l^2 norm minimizer of $\|Ax - b\|$. To see why, note that the projection of b onto the range space of A is

$$z = U_1 U_1^T b$$

and the set of x values that map to z satisfy

$$U_1 \Sigma_1 V_1^T x = U_1 U_1^T b$$

or

$$V_1^T x = \Sigma_1^{-1} U_1^T b.$$

Note that any x can be written as $x = V_1 y_1 + V_2 y_2$, where V_2 is the remainder of the V matrix in the full SVD. Because $V_1^T V_2 = 0$, we have $(V_1 y_1)^T (V_2 y_2) = 0$, and thus

$$\|x\|_2^2 = \|V_1 y_1\|_2^2 + \|V_2 y_2\|_2^2 = \|y_1\|^2 + \|y_2\|_2^2.$$

Also, note that

$$V_1^T (V_1 y_1 + V_2 y_2) = y_1,$$

so $y_1 = \Sigma_1^{-1} U_1^T b$ is completely constrained and y_2 is completely unconstrained. Therefore, the solution x with minimal l^2 norm corresponds to $y_1 = \Sigma_1^{-1} U_1^T b$ and $y_2 = 0$.

What if $A = U \Sigma V^T$ is close to rank r? That is, what if $\sigma_{r+1}, \ldots, \sigma_n$ are nonzero, but are much smaller than σ_1? In this case, the closest rank r matrix is

$$\hat{A} = U \hat{\Sigma} V^T$$

where $\hat{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) \in \mathbb{R}^{n \times n}$ is a diagonal matrix. Note that $\|\hat{A} - A\|_2 = \|\hat{\Sigma} - \Sigma\|_2 = \sigma_{r+1}$. If σ_1/σ_r is not too large, then solving the rank-deficient least squares problem with the truncated $\hat{A}$ is likely to be "well-behaved" in the sense that entries of the minimal norm solution x will not be too huge, and small changes to the matrix and the right hand side will not cause huge changes to the matrix. The same cannot necessarily be said if we solve the least squares problem with the original matrix A, which is very ill conditioned. Solving using the truncated SVD is one way of *regularizating* the problem — that is, trading off fidelity to the original equation against stability under small perturbations. There are many other methods of regularization as well.

Cheerful facts of calculus

Differentiation and matrix functions

First-order perturbation theory is mostly an exercise in differential calculus. I've described the process in terms of throwing away higher-order terms, but I could equally well describe things in terms of relating directional derivatives. For example, consider the problem of finding $Y = X^{-1}$. If $X = X(t)$ and $dX/dt = \delta_X$, and similarly with Y, then differentiating the relation $XY = I$ gives

$$\delta X \, Y + X \, \delta Y = 0,$$

so that

$$\delta Y = -X^{-1}(\delta X)X^{-1}.$$

Similarly, suppose I want to differentiate $y = x^T A x$ where $x = x(t)$ and the symmetric matrix A is considered fixed. Then

$$\delta y = \delta x^T \, Ax + x^T A \delta x = 2\delta x^T Ax,$$

where the second step involves the fact that the transpose of a scalar is the same as the original value, so that $\delta x^T Ax = x^T A \, \delta x$.

Gradients vs derivatives

Suppose $F : \mathbb{R}^n \to \mathbb{R}$ is differentiable. The derivative $F'(x)$ is a linear map from $\mathbb{R}^n$ into $\mathbb{R}$, conventionally represented as a row vector, so that we can

write the directional derivatives as linear functions of the direction:

$$\frac{d}{dt}F(x + t\,\delta x) = F'(x)\delta x.$$

In contrast, the gradient is a vector which is dual to the derivative. In Euclidean space, this just means

$$\nabla F(x) = F'(x)^T,$$

so that the directional derivatives can be expressed as dot products between $\nabla F(x)$ and a direction vector.

Please keep this distinction in mind in what follows.

Constrained optimization and Lagrange multipliers

Most people remember how to find the extrema of a differentiable multivariate function $F(x)$. The condition is that all directional derivatives are zero:

$$F'(x)\delta x = 0 \text{ for all } \delta x$$

This is usually rephrased as the statement that $\nabla F(x) = 0$ at the extrema.

To find extrema subject to the constraint, we need that the directional derivatives are zero *in every direction consistent with the constraint*:

$$F'(x)\delta x = 0 \text{ whenever } G'(x)\delta x = 0.$$

Put differently, $\nabla F(x)$ is orthogonal to the constraint surface $G(x) = 0$ at x. The directions orthogonal to this surface are precisely the directions in which $G(x)$ varies at first order — which is everything in the range space of $G'(x)^T$. Therefore, the condition for a constrained critical point is

$$\nabla F(x) \in \mathcal{R}(G'(x)^T)$$

or, equivalently,

$$\nabla F(x) = -G'(x)^T\mu$$

for some vector μ. Therefore, we have a constrained critical point when

$$\begin{aligned} \nabla F(x) + G'(x)^T\mu &= 0 \\ G(x) &= 0. \end{aligned}$$

With the additional variables μ — the Lagrange multipliers — this system now has as many unknowns as it has equations.

Now, suppose we construct the *augmented* objective function

$$\hat{L}(x, \mu) = F(x) + \mu^T G(x).$$

Then

$$\begin{aligned} \nabla_x \hat{L}(x, \mu) &= \nabla F(x) + G'(x)^T \mu, \\ \nabla_\mu \hat{L}(x, \mu) &= G(x), \end{aligned}$$

and so setting the gradient of the augmented function to zero is equivalent to finding a constrained critical point for the original function F. Setting the gradient of $\hat{L}$ to zero, in turn, is equivalent to saying that all the directional derivatives of $\hat{L}$ are zero — as we did in our discussion of linearly constrained least squares at the start of Lecture 21 on Friday. In that case, we used $F(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $G(x) = Bx - d$.