Week 8: Friday, Oct 17

Logistics

1. HW 3 errata: in Problem 1, I meant to say $p_i < i$, not that p_i is strictly ascending — my apologies. You would want $p_i > i$ if you were simply forming the matrices and handing them off to MATLAB's sparse solvers. I also mixed up the definition of A slightly. An updated version of the homework has been posted.

2. HW 3 is due on Wednesday, Oct 21. The midterm will be handed out on the same day.

Constrained least squares

We've worked for a while on the problem of minimizing $\|Ax - b\|_2$. What about doing the same minimization in the presence of constraints? For equality constraints, we can use the method of Lagrange multipliers. For example, to minimize $\frac{1}{2}\|Ax - b\|^2$ subject to $Bx = d$, we would look for stationary points of the Lagrangian

$$L(x, \lambda) = \frac{1}{2}\langle Ax - b, Ax - b\rangle + \lambda^T (Bx - d).$$

If we take derivatives in the direction δx and $\delta\lambda$, we have

$$\begin{aligned}
\delta L &= \langle A\delta x, Ax - b\rangle + \lambda^T B\delta x + \delta\lambda^T (Bx - d) \\
&= \begin{bmatrix} \delta x \\ \delta\lambda \end{bmatrix}^T \left(\begin{bmatrix} A^T A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} - \begin{bmatrix} A^T b \\ d \end{bmatrix} \right) = \begin{bmatrix} \delta x \\ \delta\lambda \end{bmatrix}^T \begin{bmatrix} \nabla_x L \\ \nabla_\lambda L \end{bmatrix}.
\end{aligned}$$

Setting the gradient equal to zero (or equivalently setting the directional derivative to zero in all directions) gives us the constrained normal equations

$$\begin{bmatrix} A^T A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} A^T b \\ d \end{bmatrix}.$$

Of course, in the case of linear equality constraints, we could also simply find a basis in which the some components of x are completely constrained and the remaining components of x are completely free (this is done in the book).

Orthogonal Procrustes

Not all least squares problems involve individual vectors. We can have more exotic problems in which the unknown is a linear transformation, too. For example, consider the following alignment problem: suppose we are given two sets of coordinates for m points in n-dimensional space, arranged into rows of $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{m \times n}$. Suppose the two matrices are (approximately) related by a rigid motion that leaves the origin fixed; how can we recover that transformation? That is, we want to find an orthogonal $W \in O(n)$ that minimizes $\|AW - B\|_F$.

This is a constrained least squares problem in disguise. Define the Frobenius inner product

$$\langle X, Y \rangle_F = \operatorname{tr}(X^T Y) = \sum_{i,j} x_{ij} y_{ij},$$

and note that $\|X\|_F^2 = \langle X, X \rangle_F$. Then to minimize $\frac{1}{2}\|AW - B\|_F^2$ subject to $W^T W = I$, we look for points where

$$L(W) = \frac{1}{2} \langle AW - B, AW - B \rangle_F$$

has zero derivative in the directions tangent to the constraint $W^T W = I$. We could do this with Lagrange multipliers, but in this case it is simpler to work with the constraints directly. When W is orthogonal, we have

$$\begin{aligned} L(W) &= \frac{1}{2} \left(\|AW\|_F^2 + \|B\|_F^2 \right) - \langle AW, B \rangle_F \\ &= \frac{1}{2} \left(\|A\|_F^2 + \|B\|_F^2 \right) - \langle AW, B \rangle_F, \end{aligned}$$

so minimizing L is equivalent to maximizing $\langle AW, B \rangle_F$. Now, note that if we differentiate $W^T W = I$, we find

$$\delta W^T W + W^T \delta W = 0;$$

that is, $\delta W = WS$ where S is a *skew symmetric* matrix ($S = -S^T$). So, the constrained stationary point should satisfy

$$\delta L = -\langle AWS, B \rangle_F = 0$$

for all skew symmetric S.

Now, note that

$$\langle AWS, B\rangle_F = \operatorname{tr}(S^T W^T A^T B) = \langle S, W^T A^T B\rangle_F.$$

This says that $W^T A^T B$ should be in the orthogonal subspace to the space of skew-symmetric matrices — which is the same as saying $W^T A^T B$ should be symmetric. Now write a *polar decomposition* for $A^T B$, $A^T B = QH$ where Q is orthogonal and H is symmetric and positive definite (assuming $A^T B$ has full rank). Then we need an orthogonal matrix W such that $W^T QH$ is symmetric. There are two natural choices, $W = \pm Q$. The choice $W = Q$ maximizes $\langle AW, B\rangle_F = \operatorname{tr}(W^T A^T B)$. Note that the polar decomposition of $A^T B$ can be computed through the singular value decomposition:

$$A^T B = U\Sigma V^T = (UV^T)(V\Sigma V^T) = QH.$$

There is another argument (used in the book) that the polar factor Q is the right matrix to maximize $\langle AW, B\rangle_F = \operatorname{tr}(W^T A^T B)$. Recall that for any matrices X and Y such that XY is square, $\operatorname{tr}(XY) = \operatorname{tr}(YX)$. Therefore if we write $A^T B = U\Sigma V^T$, we have

$$\operatorname{tr}(W^T A^T B) = \operatorname{tr}(W^T U\Sigma V^T) = \operatorname{tr}(VWU^T\Sigma) = \operatorname{tr}(Z\Sigma),$$

where $Z = VWU^T$ is again orthogonal. Note that

$$\operatorname{tr}(\Sigma Z) = \sum_i \sigma_i z_{ii}$$

is greatest when $z_{ii} = 1$ for each i (since the entries of an orthogonal matrix are strictly bounded by one). Therefore, the trace is maximized when $Z = I$, which again corresponds to $W = UV^T = Q$.

While the SVD argument is slick, I actually like the argument in terms of constrained stationary points better. It brings in the relationship between orthogonal matrices and skew-symmetric matrices, and it invokes again the idea of decomposing a vector space into orthogonal spaces (in this case the symmetric and the skew-symmetric matrices). It also gives me the excuse to introduce the polar factorization of a matrix.

This least squares problem of finding a best-fitting orthogonal matrix is sometimes called the *orthogonal Procrustes problem*. It is named in honor of the Greek legend of Procrustes, who had a bed on which he would either stretch guests or amputate them in order to make them fit perfectly.

Rank-deficient problems

Suppose A is not full rank. In this case, we have

$$A = U\Sigma V^T$$

where

$$\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r, 0, \ldots, 0).$$

There are now multiple minimizers to $\|Ax-b\|_2$, and we need some additional condition to make a unique choice. A standard choice is the x with smallest norm, which satisfies

$$x = A^\dagger b = V\Sigma^\dagger U^T b$$

where

$$\Sigma^\dagger = \operatorname{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0).$$

The matrix $A^\dagger$ is the *Moore-Penrose pseudoinverse.*

The Moore-Penrose pseudoinverse is discontinuous precisely at the rank-deficient matrices. This is not good news for computation. On the other hand if we have a *nearly* rank-deficient problem in which $\sigma_{r+1}, \ldots, \sigma_n \ll \sigma_r$, then we might want to perturb to the rank-deficient case and apply the Moore-Penrose pseudoinverse. That is, we use the *truncated* SVD of A to approximately solve the minimization problem.

One common source of rank-deficient or nearly rank-deficient problems is fitting problems in which some of the explanatory factors (represented by columns of A) are highly correlated with other factors. In such cases, we might prefer an (approximate) minimization of $\|Ax - b\|_2$ that does not use redundant factors — that is, we might want x to have only r nonzeros, where r is the effective rank. A conventional choice of the r columns is via QR with column pivoting: that is, we write

$$A\Pi = QR,$$

where the column permutation Π is chosen so that the diagonal elements of R are in order of descending magnitude. If there is a sharp drop from σ_r to σ_{r+1}, then the $n - r$ by $n - r$ trailing submatrix of R is likely to be small (on the order of σ_{r+1} in size), and discarding it corresponds to a small permutation of A. Pivoted QR is *not* foolproof; there are nearly singular matrices for which the diagonal elements of R never get too small. But pivoted QR does a good job at revealing the rank, and at constructing sparse approximate minimizers, for many practical problems.

Underdetermined problems

So far we have talked about *overdetermined* problems in which A has more rows than columns ($m > n$). Sometimes there is also a call to solve *underdetermined* problems ($m < n$). Assuming that A has full row rank, an underdetermined problem will have an $(n-m)$-dimensional space of solutions. We can use QR or SVD decompositions to find the minimum l^2 norm solution to an underdetermined problem; if we write "economy" decompositions $A^T = QR = U\Sigma V^T$, then

$$x_* = QR^{-T}b = U\Sigma^{-1}V^Tb$$

is the minimal l^2 norm solution to $Ax = b$. In many circumstances, though, we may be interested in some other solution. For example, we may want a sparse solution with as few nonzeros as possible; this is often the case for applications in *compressive sensing*. It turns out that the sparsest solution often minimizes the l^1 norm. Needless to say, the matrix decompositions we have discussed do not provide such a simple approach to finding the minimal l^1-norm solution to an underdetermined system.