Week 5: Wednesday, Feb 23

Problem du jour

Recall that if $Q \in \mathbb{R}^{n \times n}$ is orthogonal (i.e. $Q^T Q = I$), then $\|Qx\|_2 = \|x\|_2$. Show that this implies

$$\|A\|_2 = \|AQ\|_2 = \|QA\|_2,$$
$$\|A\|_F = \|AQ\|_F = \|QA\|_F.$$

Answer: The first line follows from the definition of the operator norm:

$$\begin{aligned}
\|A\|_2 &= \max_{\|x\|_2=1} \|Ax\|_2 \\
&= \max_{\|x\|_2=1} \|QAx\|_2 = \|QA\|_2 \\
&= \max_{\|y\|_2=1} \|AQy\|_2 = \|AQ\|_2 \text{ where } y = Q^T x.
\end{aligned}$$

The second line follows because we can build the Frobenius norm up from vector norms:

$$\|QA\|_F^2 = \sum_{j=1}^{m} \|Qa(:,j)\|^2 = \sum_{j=1}^{n} \|a(:,j)\|^2 = \|A\|_F^2,$$
$$\|AQ\|_F^2 = \sum_{i=1}^{m} \|a(i,:)Q\|^2 = \sum_{i=1}^{m} \|a(:,i)\|^2 = \|A\|_F^2.$$

Logistics

1. Exam solutions will be posted this afternoon on CMS.
2. Comment on Google groups: the "edit my membership" task will let you turn off email (or get a digest). If you do this, you can always still see everything on the web.
3. Comment on office hours: I have official OH on WTh 11-12 or by appointment. In practice, I will often answer questions if you swing by and my door is open.
4. We will probably be skipping eigenvalue decompositions. This *does* make me sad — perhaps we will bring it in later.

From Monday's exciting episode...

1. We are given $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $m > n$. The columns of A might correspond to factors in a regression; the vector b is data we are trying to fit. The least squares problem is to minimize

   $$\|r\|^2 = \|b - Ax\|^2.$$

   We can characterize the problem geometrically as choosing Ax to be the orthogonal projection of b onto the span of A (so that the residual is orthogonal to the span of A). That is, r is normal to all the columns of A (or $A^T r = 0$).

2. If the columns of A are independent, we can characterize the least squares solution via the *normal equations*

   $$A^T A x = A^T b.$$

   Alternately, we can compute $A = QR = Q_1 R_1$; then

   $$R_1 x = Q_1^T b.$$

   The matrix $A^\dagger = (A^T A)^{-1} A = R_1^{-1} Q_1^T$ is called the *Moore-Penrose pseudoinverse*.

3. There are two ways that the solution to the least squares problem can be sensitive: either b might be nearly orthogonal to the span of A (in which case we have a poor fit) or the condition number $\kappa(A) = \|A\|\|A^\dagger\|$ may be large. If θ is the angle between b and the span of A, then the conditioning with respect to changes in b is given by

   $$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)}{\cos(\theta)} \frac{\|\Delta b\|}{\|b\|}.$$

   The conditioning with respect to changes in A is given by

   $$\frac{\|\Delta x\|}{\|x\|} \leq \left(\kappa(A)^2 \tan(\theta) + \kappa(A)\right) \frac{\|\Delta b\|}{\|b\|}.$$

4. Which method we choose depends on the conditioning in A and where we expect errors to come from (e.g. is A exact, or does it come from measured data?). In large problems, sparsity also plays a role.

Ill-conditioned problems

In regression problems, the columns of A correspond to explanatory factors. For example, we might try to use height, weight, and age to explain the probability of some disease. In this setting, ill-conditioning happens when the explanatory factors are correlated — for example, perhaps weight might be well predicted by height and age in our sample population. This happens reasonably often. When there is some correlation, we get moderate ill conditioning, and might want to use QR factorization. When there is a lot of correlation and the columns of A are truly linearly dependent (or close enough for numerical work), or when there A is contaminated by enough noise that a moderate correlation seems dangerous, then we may declare that we have a *rank-deficient* problem.

What should we do when the columns of A are close to linearly dependent (relative to the size of roundoff or of measurement noise)? The answer depends somewhat on our objective for the fit, and whether we care about x on its own merits (because the columns of A are meaningful) or we just about Ax:

1. We may want to balance the quality of the fit with the size of the solution or some similar penalty term that helps keep things unique. This is the *regularization* approach.

2. We may want to choose a strongly linearly dependent set of columns of A and leave the remaining columns out of our fitting. That is, we want to fit to a subset of the available factors. This can be done using the leading columns of a pivoted version of the QR factorization $AP = QR$. This is sometimes called *parameter subset selection*. MATLAB's backslash operator does this when A is numerically singular.

3. We may want to choose the "most important" directions in the span of A, and use them for our fitting. This is the idea behind *principal components analysis*.

We will focus on the "most important directions" version of this idea, since that will lead us into our next topic: the singular value decomposition. Still, it is important to realize that in some cases, it is more appropriate to add a regularization term or to reduce the number of fitting parameters.

Singular value decomposition

The *singular value decomposition* (SVD) is important for solving least squares problems and for a variety of other approximation tasks in linear algebra. For $A \in \mathbb{R}^{m \times n}$[1], we write

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

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal. The diagonal matrix Σ has non-negative diagonal entries

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0.$$

The σ_i are called the *singular values* of A. We sometimes also write

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T = U_1 \Sigma_1 V_1^T$$

where $U_1 \in \mathbb{R}^{m \times n}$, $\Sigma_1 \in \mathbb{R}^{n \times n}$, $V_1 \in \mathbb{R}^{n \times n}$. We call this the *economy* SVD.

We can interpret the SVD geometrically using the same picture we drew when talking about the operator two norm. The matrix A maps the unit ball to an ellipse. The axes of the ellipse are $\sigma_1 u_1$, $\sigma_2 u_2$, etc, where the σ_i give the lengths and the u_i give the directions (remember that the u_i are normalized). The columns of V are the vectors in the original space that map onto these axes; that is, $Av_i = \sigma_i u_i$.

We can use the geometry to define the SVD as follows. First, we look for the major axis of the ellipse formed by applying A to the unit ball:

$$\sigma_1^2 = \max_{\|v\|=1} \|Av\|^2 = \max_{\|v\|=1} v^T (A^T A) v.$$

Some of you may recognize this as an eigenvalue problem in disguise: σ_1^2 is the largest eigenvalue of $A^T A$, and v_1 is the corresponding eigenvector. We can then compute u_1 by the relation $\sigma_1 u_1 = Av_1$. To get σ_2, we restrict our attention to the spaces orthogonal to what we have already seen:

$$\sigma_2^2 = \max_{\|v\|=1, v \perp v_1, Av \perp u_1} \|Av\|^2.$$

We can keep going to get the other singular values and vectors.

[1] We will assume for the moment that $m \geq n$. Everything about the SVD still makes sense when $m < n$, though.

Norms, conditioning, and near singularity

Given an economy SVD $A = U\Sigma V^T$, we can give satisfyingly brief descriptions (in the two norm) of many of the concepts we've discussed so far in class. The two-norm of A is given by the largest singular value: $\|A\|_2 = \sigma_1$. The pseudoinverse of A, assuming A is full rank, is

$$(A^T A)^{-1} A = U\Sigma^{-1} V^T,$$

which means that $\|A^\dagger\|_2 = 1/\sigma_n$. The condition number for least squares (or for solving the linear system when $m = n$) is therefore

$$\kappa(A) = \sigma_1/\sigma_n.$$

Another useful fact about the SVD is that it gives us a precise characterization of what it means to be "almost" singular. Suppose $A = U\Sigma V^T$ and E is some perturbation. Using invariance of the matrix two norm under orthogonal transformations (the problem du jour), we have

$$\|A + E\|_2 = \|U(\Sigma + \tilde{E})V^T\| = \|\Sigma + \tilde{E}\|,$$

where $\|\tilde{E}\| = \|U^T E V\| = \|E\|$. For the diagonal case, we can actually characterize the smallest perturbation $\tilde{E}$ that makes $\Sigma + \tilde{E}$ singular. It turns out that this smallest perturbation is $\tilde{E} = -\sigma_n e_n e_n^T$ (i.e. something that zeros out the last singular value of Σ). Therefore, we can characterize the smallest singular value as the *distance to singularity*:

$$\sigma_n = \min\{\|E\|_2 : A + E \text{ is singular}\}.$$

The condition number therefore is a *relative distance to singularity*, which is why I keep saying ill-conditioned problems are "close to singular."

The SVD and rank-deficient least squares

If we substitute $A = U\Sigma V^T$ in the least squares residual norm formula, we can "factor out" U just as we pulled out the Q factor in QR decomposition:

$$\|Ax - b\| = \|U\Sigma V^T x - b\| = \|\Sigma\tilde{x} - \tilde{b}\|, \text{ where } \tilde{x} = V^T x \text{ and } \tilde{b} = U^T b.$$

Note that $\|\tilde{x}\| = \|x\|$ and $\|\tilde{b}\| = \|b\|$.

If A has rank r, then singular values $\sigma_{r+1}, \ldots, \sigma_n$ are all zero. In this case, there are many different solutions that minimize the residual — changing the values of $\tilde{x}_{r+1}$ through $\tilde{x}_n$ does not change the residual at all. One standard way to pick a unique solution is to choose the minimal norm solution to the problem, which corresponds to setting $\tilde{x}_{r+1} = \ldots = \tilde{x}_n = 0$. In this case, the Moore-Penrose pseudoinverse is defined as

$$A^\dagger = V_+ \Sigma_+^{-1} U_+^T$$

where $\Sigma_+ = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ and U_+ and V_+ consist of the first r left and right singular vectors.

If A has entries that are not zero but small, it often makes sense to use a *truncated SVD*. That is, instead of setting $\tilde{x}_i = 0$ just when $\sigma_i = 0$, we set $\tilde{x}_i = 0$ whenever σ is small enough. This corresponds, if you like, to perturbing A a little bit before solving in order to get an approximate least squares solution that does not have a terribly large norm.

Why, by the way, might we want to avoid large components? A few reasons come to mind. One issue is that we might be solving linear least squares problems as a step in the solution of some nonlinear problem, and a large solution corresponds to a large step — which means that the local, linear model might not be such a good idea. As another example, suppose we are looking at a model of patient reactions to three drugs, A and B. Drug A has a small effect and a horrible side effect. Drug B just cancels out the horrible side effect. Drug C has a more moderate effect on the problem of interest, and a different, small side effect. A poorly-considered regression might suggest that the best strategy would be to prescribe A and B together in giant doses, but common sense suggests that we should concentrate on drug C. Of course, neither of these examples requires that we use a truncated SVD — it might be fine to use another regularization strategy, or use subset selection.