Week 3: Wednesday, Sep 5

Cauchy-Schwarz: a quick reminder

For any inner product,

$$0 \leq \|su + v\|^2 = \langle su + v, su + v \rangle = s^2\|u\|^2 + 2s\langle u, v\rangle + \|v\|^2$$

So we have a quadratic in s with at most one real root. Therefore, the discriminant must be nonpositive, i.e.

$$4\langle u, v\rangle^2 - 4\|u\|^2\|v\|^2 \leq 0.$$

With a little algebra, we have the *Cauchy-Schwarz inequality*,

$$|\langle u, v\rangle| < \|u\|\|v\|.$$

Furthermore, $|\langle u, v\rangle| = \|u\|\|v\|$ iff $\|su - v\|^2 = 0$ for some s, in which case u and v are parallel.

I hope you will have seen the Cauchy-Schwarz inequality before, but I remind you of it because I will want to use it repeatedly. In particular, I want to use it right now to prove that $\|A\|_2 = \|A^*\|_2$. By definition,

$$\|A\|_2 = \max_{\|v\|=1} \|Av\|_2.$$

Let v_1 be a unit vector such that $\|Av_1\|$ is maximal and define $u_1 = Av_1/\|Av_1\|_2$. Then by Cauchy-Schwarz, together with the definition of the 2-norm, we have

$$\|A\|_2 = \langle Av_1, u_1\rangle = \langle v_1, A^*u_1\rangle \leq \|v_1\|\|A^*u_1\|_2 = \|A^*u_1\|_2 \leq \|A^*\|_2.$$

Now define $w_1 = A^*u_1/\|A^*u_1\|_2$, and use the same argument to get that

$$\|A^*\|_2 = \langle A^*u_1, w_1\rangle = \langle u_1, Aw_1\rangle \leq \|u_1\|\|Aw_1\|_2 = \|Aw_1\|_2 \leq \|A\|_2.$$

Therefore, $\|A\|_2 = \|A_*\|_2$, and all the inequalities in the previous two linear are actually equalities. Note that this also means that both v_1 and w_1 are parallel to A^*u_1, and hence to each other. In fact, both v_1 and w_1 are vectors that form a zero angle with A^*u_1 — which means that $v_1 = w_1$.

Orthogonal matrices

To develop fast, stable methods for matrix computation, it will be crucial that we understand different types of structures that matrices can have. This includes both "basis-free" properties, such as orthogonality, singularity, or self-adjointness; and properties such as the location of zero elements that are really associated with a *matrix* rather than with a linear transform.

Orthogonal matrices will be important throughout our work. The usual definition says that square matrix Q is orthogonal if $Q^*Q = I$, but there are other ways to characterize orthogonality as well. For example, a real square matrix Q is orthogonal iff $\|Qv\|_2 = \|v\|_2$ for all v. Why? Recall that for a real vector space,

$$\langle u+v, u+v \rangle = \langle u, u \rangle + \langle v, v \rangle + 2\langle u, v \rangle.$$

With a little algebra, we have

$$\langle u, v \rangle = \frac{1}{2}\left(\|u+v\|_2^2 - \|u\|_2^2 - \|v\|_2^2 \right).$$

Therefore, if $\|Qv\|_2 = \|v\|_2$ for every v, we have

$$\begin{aligned}
\langle Qu, Qv \rangle &= \frac{1}{2}\left(\|Q(u+v)\|_2^2 - \|Qu\|_2^2 - \|Qv\|_2^2 \right) \\
&= \frac{1}{2}\left(\|u+v\|_2^2 - \|u\|_2^2 - \|v\|_2^2 \right) = \langle u, v \rangle.
\end{aligned}$$

In particular, that means that if e_i denotes the ith column of the identity, then $\langle Qe_i, Qe_j \rangle = \langle e_i, e_j \rangle = \delta_{ij}$, or $Q^*Q = I$.

Because a matrix is orthogonal iff it preserves lengths in the two-norm, we have that

$$\begin{aligned}
\|QA\|_2 &= \|A\|_2, & \|AQ\|_2 &= \|A\|_2, \\
\|QA\|_F &= \|A\|_F, & \|AQ\|_F &= \|A\|_F.
\end{aligned}$$

There are other important cases of things that remain invariant under orthogonal transformation, too. For example, suppose $Z_1, \ldots, Z_n$ are independent standard normal random variables; then their joint probability density is

$$f(z_1, z_2, \ldots, z_n) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \right) = \frac{e^{-\|z\|_2^2/2}}{(2\pi)^{n/2}}.$$

Because the density depends only on the length of the vector z, we find that $Y = QZ$ has the same density for any orthogonal matrix Q.

Scalar multiples of orthogonal matrices are also the only perfectly conditioned matrices. That is, if $\kappa_2(A) = 1$, then $A = \alpha Q$, where Q is some orthogonal matrix. To see this, recall that

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\max_{\|v\|_2=1} \|Av\|_2}{\min_{\|u\|_2=1} \|Au\|_2},$$

so if $\kappa_2(A) = 1$, the images of all unit vectors under A have the same length – which means that the lengths of all vectors are scaled by the same amount by the action of A. Define $\alpha = \|Av\|/\|v\|$ to be the scaling factor; then $Q = \alpha^{-1}A$ scales the length of every vector by one, which means that Q is orthogonal.

The singular value decomposition

The fact that orthogonal transforms leave so many metric properties of matrices unchanged suggests the following: find orthogonal transformations that, when applied to a matrix A, result in a matrix that is as structurally simple as possible. The result of this is the singular value decomposition (SVD), which is discussed in 2.5.3–2.5.5 in the third edition of Golub and Van Loan. That is, we can write

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

where U and V are unitary matrices and Σ is a diagonal matrix with nonnegative diagonal entries that — according to convention — appear in descending order. If A is rectangular, we will sometimes distinguish the "full SVD" (in which Σ is a rectangular matrix with the same dimensions as A) from the "economy SVD" (in which one of U or V is a rectangular matrix with orthonormal columns).

There are a few ways to derive the SVD. The most fundamental approach is via a sequence of optimization problems. Recall that the 2-norm of A is defined via

$$\sigma_1 \equiv \|A\|_2 = \max_{\|v\|_2=1} \|Av\|_2.$$

Let v_1 be a vector at which $\|Av\|_2$ is maximal, and let $u_1 = Av/\|Av\|_2$. Let $[v_1, V_2]$ and $[u, U_2]$ be orthonormal bases for the row and column spaces,

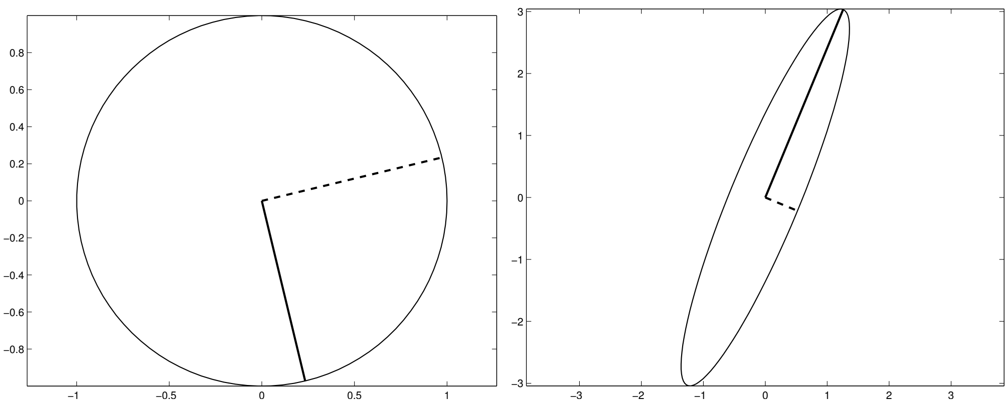

Figure 1: Graphical depiction of an SVD of $A \in \mathbb{R}^{2\times 2}$. The matrix A maps the unit circle (left) to an oval (right); the vectors v_1 (solid, left) and v_2 (dashed, left) are mapped to the major axis $\sigma_1 u_1$ (solid, right) and the minor axis $\sigma_2 u_2$ (dashed, right) for the oval.

respectively, and write

$$\tilde{A} \equiv \begin{bmatrix} \sigma_1 & f^* \\ g & A_{22} \end{bmatrix} = \begin{bmatrix} u_1 & U_2 \end{bmatrix}^* A \begin{bmatrix} v_1 & V_2 \end{bmatrix}$$

Because we only used length-preserving orthogonal operations, we must have $\|\tilde{A}\|_2 = \|A\| = \sigma_1$. But $\|\tilde{A}e_1\|^2 = \sigma_1^2 + \|g\|^2$, where e_1 is the first column of the identity (and is therefore a unit length vector). So $g = 0$. Applying a similar argument to $\tilde{A}^*$ (recall that $\|\tilde{A}^*\| = \sigma_1$, too), we also have $f = 0$. Applying the same process recursively to A_{22}, we can get the rest of the SVD.

Geometry of the SVD

How should we understand the singular value decomposition? We've already described the basic algebraic picture:

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

where U and V are orthonormal matrices and Σ is diagonal. But what does this say about the geometry of A? It says that v_1 is the vector that is stretched the most by multiplication by A, and σ_1 is the amount of stretching. More generally, we can *completely* characterize A by an orthonormal basis of right singular vectors that are each transformed in the same special way: they get scaled, then rotated or reflected in a way that preserves lengths. Viewed differently, the matrix A maps vectors on the unit sphere into an ovoid shape, and the singular values are the lengths of the axes. In Figure 1, we show this for a particular example, the matrix

$$A = \begin{bmatrix} 0.8 & -1.1 \\ 0.5 & -3.0 \end{bmatrix}.$$

Conditioning and the distance to singularity

We have already seen that the condition number for matrix multiplication is

$$\kappa(A) = \|A\| \|A^{-1}\|$$

When the norm in question is the operator two norm, we have that $\|A\| = \sigma_1$ and $\|A^{-1}\| = \sigma_n^{-1}$, so

$$\kappa(A) = \frac{\sigma_1}{\sigma_n}$$

That is, $\kappa(A)$ is the ratio between the largest and the smallest amounts by which a vector can be stretched through multiplication by A.

There is another way to interpret this, too. If $A = U\Sigma V^T$ is a square matrix, then the smallest E (in the two-norm) such that $A - E$ is *exactly* singular is $A - \sigma_n u_n v_n^T$. Thus,

$$\kappa(A)^{-1} = \frac{\|E\|}{\|A\|}$$

is the *relative distance to singularity* for the matrix A. So a matrix is ill-conditioned exactly when a relatively small perturbation would make it exactly singular.

Numerical low rank

The rank of a matrix A is given by the number of nonzero singular values. In computational practice, we would say that a matrix has *numerical* rank k if exactly k singular values are "sufficiently" greater than zero. If $k < \min(m, n)$, then we say that the matrix is numerically singular.

The rank of a matrix is theoretically interesting and useful, but it is also computationally useful to realize when a matrix is low rank, because that low rank structure can be used for fast multiplication. Suppose the SVD for $A \in \mathbb{R}^{n \times n}$ is

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

where $U_1, V_1 \in \mathbb{R}^{n \times k}$ and $\Sigma_1 \in \mathbb{R}^{k \times k}$. If we don't use anything about the structure of A, then we take $O(n^2)$ time to compute $y = Ax$. If we write $y = U_1(\Sigma_1(V_1^* x))$, then it takes $O(nk)$ time to compute y. If $k \ll n$, this may be a substantial savings! So there is a potential efficiency win in recognizing when a matrix has low rank, particularly when the matrix can be written as an outer product from the outset so that we don't have to compute an SVD or similar factorization.