Numerical Analysis Hilary Term 2021
Lecture 5: Singular Value Decomposition

We now introduce the Singular Value Decomposition (SVD), an extremely important matrix decomposition applicable to any matrix, including nonsymmetric and rectangular ones.

Theorem. (SVD) Every matrix $A \in \mathbb{R}^{m\times n}$ with $m \geq n$ can be written as

$$A = U\Sigma V^T, \tag{1}$$

where $U \in \mathbb{R}^{m\times n}$ and $V \in \mathbb{R}^{n\times n}$ are matrices with orthonormal columns, i.e., $U^TU = I_n$ and $V^TV = I_n = VV^T$ (V is square orthogonal; note that $UU^T \neq I_m$), and

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \ (= \mathrm{diag}(\sigma_1, \ldots, \sigma_n))$$

is a diagonal matrix with nonnegative diagonal entries. In short, the SVD is a decomposition of A into a product of 'orthonormal-diagonal-orthogonal' matrices; when A is square $m = n$, 'orthogonal-diagonal-orthogonal'.

One can think of orthogonal matrices as a length-preserving rotation, so the SVD indicates that applying a matrix performs a rotation, followed by shrinkage or amplification of the elements, followed by another (different) rotation.

σ_i are called the *singular values* and usually arranged in decreasing order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. The columns of U, V are called the (left and right) *singular vectors* of A. The *rank* of a matrix A is the number of its positive singular values (this is equivalent e.g. to the number of linearly independent columns or rows).

Proof. Let's prove the existence of the SVD (1) by the following steps.

1. The matrix $A^TA \in \mathbb{R}^{n\times n}$ is symmetric. This is straightforward to verify, either by direct calculations or from the general identity $(XY)^T = Y^TX^T$.

2. The eigenvalues of A^TA are all real and nonnegative (such matrices are called symmetric positive definite). To see this, suppose $A^TAx = \lambda x$, $x \neq 0$. Then $x^TA^TAx = \lambda x^Tx$, so $\lambda = \frac{x^TA^TAx}{x^Tx} = \frac{y^Ty}{x^Tx} \geq 0$, where $y = Ax$.

3. Let $A^TA = VD^2V^T$ be the symmetric eigenvalue decomposition, with $V \in \mathbb{R}^{n\times n}$ orthogonal and D diagonal. Then let $B = AV$. Now $B^TB = D^2$ is a diagonal matrix, implying that the columns of B are pairwise orthogonal.

4. Let's write $B^TB = D^2 = \mathrm{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$, where $\lambda_r > 0$.

   (a) It is possible that $r = n$, and this is an important case (happens iff $\mathrm{rank}(A) = n$) where there is no 0 diagonal entry in D^2. We then have $D^{-1} = \mathrm{diag}(1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_r})$. Take $U := BD^{-1} = AVD^{-1}$, which has orthonormal columns $U^TU = I_n$. We are then done, as taking $\Sigma = D$, $A = U\Sigma V^T$.

(b) When $r < n$ (the rank-deficient case), B has columns that are 0. Let $D_r = \text{diag}(\lambda_1, \ldots, \lambda_r)$. We still have $B \begin{bmatrix} D_r^{-1} & \\ & I_{n-r} \end{bmatrix} = [U_1,\ 0]$, and so

$$A = [U_1,\ 0] \begin{bmatrix} D_r & \\ & I_{n-r} \end{bmatrix} V^T = [U_1 U_2] \begin{bmatrix} D_r & \\ & 0 \end{bmatrix} V^T$$

for any U_2; we take it to be orthonormal $U_2^T U_2 = I_{n-r}$ and $U_2^T U_1 = 0$ (U_2 is any orthonormal matrix in the orthogonal complement of U_1; its existence can be verified e.g. using Householder reflectors). Taking $U = [U_1, U_2]$ completes the proof, again with $\Sigma = D$.

□

Some comments:

- Analogous to the full QR factorisation, there is a 'full SVD' $A = \tilde{U}\tilde{\Sigma}\tilde{V}^T$, where $\tilde{U} = [U\ U_\perp] \in \mathbb{R}^{m \times m}$ is orthogonal and $\tilde{\Sigma} \in \mathbb{R}^{m \times n} = \begin{bmatrix} \Sigma \\ 0_{(m-n)\times n} \end{bmatrix}$ and $\tilde{V} = V$. This can be obtained by starting from (1) and finding an orthogonal complement $U_\perp$ of U.
- Fat matrices: the assumption $m \geq n$ is just for convenience; if $m < n$, one still has $A = U\Sigma V^T$ where $\Sigma \in \mathbb{R}^{m \times m}$ is diagonal, $U \in \mathbb{R}^{m \times m}$ is orthogonal, and $V \in \mathbb{R}^{n \times m}$ has orthonormal columns. Below we continue with the assumption $m \geq n$.
- The SVD extends directly to matrices with nonreal entries: $A = U\Sigma V^*$, where U, V are unitary matrices and * denotes the conjugate transpose.

Matrix spectral norm Let us briefly introduce the *spectral norm*[1] for matrices $A \in \mathbb{R}^{m \times n}$: $\|A\|_2 = \sigma_1(A)$, i.e., the largest singular value. It is a nonnegative scalar that measures 'how large' the matrix is. It has the equivalent characterisation $\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$, where the norms in the right-hand side are the standard Euclidean norm (length) for vectors $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$.

Low-rank approximation The SVD is useful for theoretical purposes, as it identifies e.g. the range (column space), null space, rank, and many more. In applications, the primary reason SVD is so important is that it gives the optimal low-rank approximation.

Let $A = U\Sigma V^T$ be the SVD and write $U = [u_1, \ldots, u_n], V = [v_1, \ldots, v_n]$, and define the "tall-skinny matrices" $U_k = [u_1, \ldots, u_k]$, $V_k = [v_1, \ldots, v_k]$, and $\Sigma_k = \text{diag}(\sigma_1, \ldots, \sigma_k)$. Let k be any integer $k \leq n$. Then set

$$A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

Note that $\text{rank}(A_r) = r$. Also note that $A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$, which is another way of expressing the SVD. A_r is called the *truncated SVD* of A, as A_r is obtained by truncating the trailing components of the SVD of A.

[1] Also known as the 2-norm or the operator norm. We return to the topic of norms later in the course.

We are now ready to state the result.

Theorem. Let $r \leq n$ be an integer. For any $B \in \mathbb{C}^{m\times n}$ with $\text{rank}(B) \leq r$,

$$\|A - A_r\|_2 = \sigma_{r+1} \leq \|A - B\|_2. \tag{2}$$

In other words, A_r *is the best rank-r approximant to* A in the spectral norm.

Proof. The first equality $\|A - A_r\|_2 = \sigma_{r+1}$ can be seen by noting that $A - A_r = \sum_{i=r+1}^{n} \sigma_i u_i v_i^T$ with singular values $\sigma_{r+1}, \ldots, \sigma_n$, along with r 0's. For the inequality:

1. Since $\text{rank}(B) \leq r$, we can write $B = B_1 B_2^T$ where B_1, B_2 have r columns. Therefore, there exists an orthonormal null space $W \in \mathbb{C}^{n\times(n-r)}$ s.t. $BW = 0$.

2. Then $\|A - B\|_2 \geq \|(A - B)W\|_2 = \|AW\|_2 = \|U\Sigma(V^T W)\|_2$. Now since W is $(n - r)$-dimensional, there is an interesection between W and $[v_1, \ldots, v_{r+1}]$, the $(r + 1)$-dimensional subspace spanned by the leading $r + 1$ left singular vectors ($[W, v_1, \ldots, v_{r+1}][x_1, x_2]^T = 0$ has a solution; then Wx_1 is such a vector).

3. Scale x_1 to have unit norm, and by orthogonal invariance $\|U\Sigma V^T W x_1\|_2 = \|\Sigma V^T W x_1\|_2 = \|\Sigma_{r+1} y_1\|_2$, where $\|y_1\|_2 = 1$ (b.c. Wx_1 lies in $\text{span}[v_1, \ldots, v_{r+1}]$) and Σ_{r+1} is the leading $r + 1$ part of Σ.

4. Then $\|U\Sigma_{r+1} y_1\|_2 \geq \sigma_{r+1}$ can be verified by direct calculations.

□

In fact, more generally it is known that

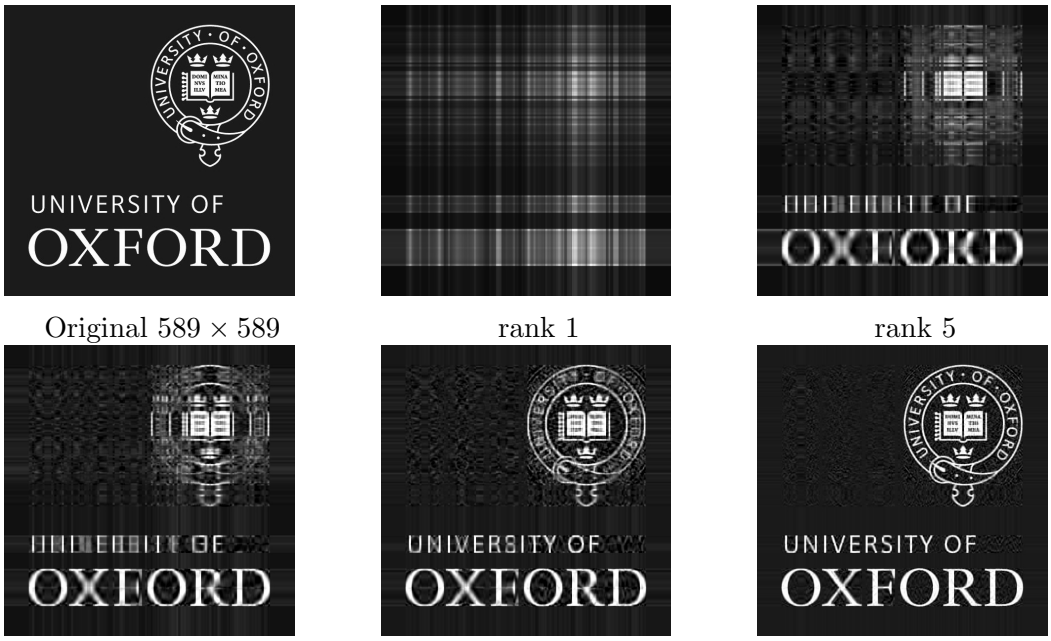
$$\|A - A_r\| \leq \|A - B\| \tag{3}$$

for any so-called unitarily invariant norm $\|\cdot\|$ (non-examinable).

In many applications $\sigma_{r+1} \ll \sigma_1$ for some $r \ll n$, in which case $A \approx U_r \Sigma_r V_r^T$. Now, storing U_r, Σ_r, V_r requires $\approx (m + n + 1)r$ memory, as opposed to mn for the full A, so when $r \ll \min(m, n)$, this can be used for data compression; this fact is used everywhere e.g. in data science!

Illustration of low-rank approximation: A traditional example to illustrate low-rank approximation via the truncated SVD is image compression. A grayscale image can be represented by a matrix A, with each entry representing the intensity of a pixel. One can then approximate A by a truncated SVD, and use that to get a compressed image that hopefully looks similar to the original image to human eyes. Images tend to have structure that lends A to be approximately low-rank.

Below we take an image of the Oxford logo, represent it as a matrix $A \in \mathbb{R}^{589\times 589}$ and compute its SVD (just `[U,S,V] = svd(A)` in MATLAB). Using the truncated SVD we then compute a rank-r approximation for different values of r. With a rank-1 matrix the rows (and columns) are all parallel so the image is uninformative; but as r increases the image becomes clear, and with rank 50 the image is almost indistinguishable from the original, while still giving some data compression. For larger images, such savings can be



Original 589×589 rank 1 rank 5

rank 10 rank 20 rank 50

Figure 1: The Oxford logo and its low-rank approximations via the truncated SVD.

significant. (This is however not how images are usually compressed in practice; e.g. the algorithm behind the jpg format is completely different).

The SVD $A = U\Sigma V^T$ and symmetric eigenvalue decomposition $A = V\Lambda V^T$ have many properties and results in common (e.g. Courant-Fisher min-max theorem; nonexaminable), stemming from the fact that they are both decompositions of the form "orthogonal-diagonal-orthogonal". In fact the SVD proof given above suggests an algorithm for computing the SVD via a symmetric eigenvalue decomposition of $A^T A$ (this is not exactly how the SVD is compute in practice, but this is outside the scope); we now turn to eigenvalue problems $Ax = \lambda x$ and describe an algorithm for solving them.