Numerical Analysis Hilary Term 2025
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

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

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

$$\mathbb{R}^{n\times n} \ni \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \quad (= \mathrm{diag}(\sigma_1, \dots, \sigma_n))$$

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 view the SVD decomposition as a generalization of the diagonalisation of symmetric matrices: $A = Q^\top DQ$, where Q is orthogonal and D is diagonal whose entries are the eigenvalues of A. In fact, if A is symmetric positive (semi)definite, then the $Q^\top DQ$ is an SVD of A since the entries of D are nonnegative.

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

The diagonal entries $\{\sigma_i\}_{i=1}^n$ 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 and 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 $A^\top A \in \mathbb{R}^{n\times n}$ is symmetric. This is straightforward to verify, either by direct calculations or from the general identity $(XY)^\top = Y^\top X^\top$.

2. The eigenvalues of $A^\top A$ are all real and nonnegative (such matrices are called symmetric positive semidefinite). To see this, suppose $A^\top Ax = \lambda x$, $x \neq 0$. Then,

$$x^\top A^\top Ax = \lambda x^\top x \implies \lambda = \frac{x^\top A^\top Ax}{x^\top x} = \frac{\|Ax\|_2^2}{\|x\|_2^2} \geq 0,$$

where we recall that $\|\cdot\|_2$ denotes the standard Euclidean norm (length): $\|z\|_2^2 := z_1^2 + z_2^2 + \cdots + z_n^2$ for $z \in \mathbb{R}^n$.

3. Let $A^\top A = VD^2V^\top$ be the symmetric eigenvalue decomposition, with $V \in \mathbb{R}^{n\times n}$ orthogonal and $D \in \mathbb{R}^{n\times n}$ diagonal. Taking $B := AV \in \mathbb{R}^{m\times n}$, we have

$$B^\top B = V^\top A^\top AV = (V^\top V)D^2(V^\top V) = D^2,$$

and so the columns of B are pairwise orthogonal.

4. Let's write $B^\top B = D^2 = \mathrm{diag}(\lambda_1, \ldots, \lambda_\ell, 0, \ldots, 0)$, where $\lambda_\ell > 0$. Note that $\mathrm{rank}(A) = \ell$. There are two cases:

   (a) $\ell = n$: D is invertible and $D^{-1} = \mathrm{diag}(1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_n})$. Defining $U \in \mathbb{R}^{m\times n}$ by $U := BD^{-1} = AVD^{-1}$, we have

$$U^\top U = D^{-1}B^\top BD^{-1} = D^{-1}D^2D^{-1} = I_n,$$

   and so the columns of U are orthonormal. Setting $\Sigma := D$ completes the decomposition $A = U\Sigma V^\top$ since

$$U\Sigma V^\top = BD^{-1}DV^\top = BV^\top = AVV^\top = A.$$

   (b) $\ell < n$ (the rank-deficient case): The last $n - \ell$ columns of B are all 0. Let $D_\ell := \mathrm{diag}(\lambda_1, \ldots, \lambda_\ell)$. Then, we can still proceed in as (a) in the sense that

$$B \begin{bmatrix} D_\ell^{-1} & \\ & I_{n-\ell} \end{bmatrix} = \begin{bmatrix} U_1 & 0 \end{bmatrix},$$

   where $U_1 \in \mathbb{R}^{m\times l}$ denotes the first l columns of the above product. U_1 has orthonormal columns since

$$\begin{aligned}
\begin{bmatrix} U_1^\top U_1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} U_1^\top \\ 0 \end{bmatrix} \begin{bmatrix} U_1 & 0 \end{bmatrix} &= \begin{bmatrix} D_\ell^{-1} & \\ & I_{n-\ell} \end{bmatrix} B^\top B \begin{bmatrix} D_\ell^{-1} & \\ & I_{n-\ell} \end{bmatrix} \\
&= \begin{bmatrix} D_\ell^{-1} & \\ & I_{n-\ell} \end{bmatrix} D^2 \begin{bmatrix} D_\ell^{-1} & \\ & I_{n-\ell} \end{bmatrix} \\
&= \begin{bmatrix} I_\ell & 0 \\ 0 & 0 \end{bmatrix}
\end{aligned}$$

   and we have

$$\begin{bmatrix} U_1 & 0 \end{bmatrix} \begin{bmatrix} D_\ell & \\ & I_{n-\ell} \end{bmatrix} V^\top = B \begin{bmatrix} D_\ell^{-1} & \\ & I_{n-\ell} \end{bmatrix} \begin{bmatrix} D_\ell & \\ & I_{n-\ell} \end{bmatrix} V^\top = BV^\top = AVV^\top = A.$$

   Let $U_2 \in \mathbb{R}^{m\times(n-l)}$ be a matrix with orthonormal columns $U_2^\top U_2 = I_{n-l}$ satisfying $U_2^\top U_1 = 0$ (the columns of U_2 are orthogonal to the columns of U_1). Such a U_2 exists by e.g. Gram-Schmidt.
   Let $U := \begin{bmatrix} U_1 & U_2 \end{bmatrix}$ and $\Sigma := D$. Then,

$$U\Sigma V^\top = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} D_\ell & \\ & 0 \end{bmatrix} V^\top = \begin{bmatrix} U_1 & 0 \end{bmatrix} \begin{bmatrix} D_\ell & \\ & I_{n-l} \end{bmatrix} V^\top = A,$$

   which completes the proof. □

Some comments:

1. Analogous to the full QR factorisation, there is a 'full SVD' $A = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$, 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.
2. If A has rank ℓ and $A = \tilde{U}\tilde{\Sigma}V^\top$ is the full SVD of A, then
   - the first ℓ columns of $\tilde{U}$ form a basis for the column space of A;
   - the last $m-\ell$ columns of $\tilde{U}$ form a basis for the null space of $A^\top$;
   - the first ℓ columns of V form a basis for the columns space of $A^\top$;
   - the last $n-\ell$ columns of V form a basis for the null space of A.
3. Fat matrices: the assumption $m \geq n$ is just for convenience; if $m < n$, one still has $A = U\Sigma V^\top$ 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 $m \geq n$.
4. The SVD extends directly to matrices with complex entries: $A = U\Sigma V^*$, where U, V are unitary matrices and * denotes the conjugate transpose. Comment 2 above still holds in this case if we replace $A^\top$ by A^*.

Matrix spectral norm

Let us briefly introduce the *spectral norm*[1] for matrices $A \in \mathbb{R}^{m\times n}$:

$$\|A\|_2 := \max_{0\neq x\in\mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2}.$$

This is the "standard" way to extend norms on vectors to norms on matrices (more generally, norms on vector spaces to norms on linear operators). It is a nonnegative scalar that measures 'how large' the matrix is. Note that

$$\|Ax\|_2 = \frac{\|Ax\|_2}{\|x\|_2} \cdot \|x\|_2 \leq \|A\|_2\|x\|_2.$$

Lemma. For all $A \in \mathbb{R}^{m\times n}$, there holds $\|A\|_2 = \sigma_1(A)$, the largest singular value of A.

Proof. Let $x \in \mathbb{R}^n$ and let $A = U\Sigma V^\top$ be the SVD factorisation of A. Then,

$$\|Ax\|_2^2 = \|U\Sigma V^\top x\|_2^2 = \|\Sigma \underbrace{V^\top x}_{=:y}\|_2^2 = \|\Sigma y\|_2^2 = \sum_{i=1}^n (\sigma_i y_i)^2 \leq \sigma_i^2\|y\|_2^2,$$

where used that U has orthonormal columns. Since V is orthogonal, $\|y\|_2 = \|V^\top x\|_2 = \|x\|_2$, there holds $\|Ax\|_2 \leq \sigma_1\|x\|$, and so $\|A\|_2 \leq \sigma_1(A)$. Note that we obtain equality for $x = v_1$, the first column of V, and so $\|A\|_2 = \sigma_1(A)$. □

Lemma. For all $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times k}$, there holds $\|AB\|_2 \leq \|A\|_2\|B\|_2$.

Proof. Note that

$$\|AB\|_2 = \max_{0\neq x\in\mathbb{R}^k} \frac{\|ABx\|_2}{\|x\|_2} \leq \|A\|_2 \max_{0\neq x\in\mathbb{R}^k} \frac{\|Bx\|_2}{\|x\|_2} = \|A\|_2\|B\|_2,$$

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

which completes the proof. $\square$

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 $A = U\Sigma V^\top$ be the SVD and write $U = [u_1, \dots, u_n], V = [v_1, \dots, v_n]$. Let r be any integer $r \le n$, and define the "tall-skinny matrices" $U_r = [u_1, \dots, u_r]$, $V_r = [v_1, \dots, v_r]$, and $\Sigma_r = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$. Then set

$$A_r = U_r \Sigma_r V_r^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top.$$

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

We are now ready to state the result.

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

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

In other words, the truncated SVD 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 $A - A_r = \sum_{i=r+1}^{n} \sigma_i u_i v_i^\top$ with singular values $\sigma_{r+1}, \dots, \sigma_n$, along with r 0's.

Now let $B \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(B) \le r$.

1. Let $B = \hat{U}\hat{\Sigma}\hat{V}^\top$ denote the SVD decomposition of B. By comments 2 and 4 above, the columns of the matrix $W = \begin{bmatrix} \hat{v}_{r+1} & \hat{v}_{r+2} & \cdots & \hat{v}_n \end{bmatrix} \in \mathbb{R}^{n\times(n-r)}$, where $\{\hat{v}_i\}$ denote the columns of $\hat{V}$, span the nullspace of B. Hence, $BW = 0$.

2. Note that $\|W\|_2 = 1$ since the columns of W are orthonormal and $\|Wx\|_2 = \|x\|_2$ for all $x \in \mathbb{R}^{n-r}$. Consequently, there holds $\|(A-B)W\|_2 \le \|A-B\|_2\|W\|_2 = \|A-B\|_2$.

3. $\|A - B\|_2 \ge \|(A-B)W\|_2 = \|AW\|_2 = \|U\Sigma(V^\top W)\|_2$. Let $\{v_1, v_2, \dots, v_{r+1}\}$ denote the leading $r+1$ right singular vectors (the first $r+1$ columns of V). Then, the set $\{w_1, w_2, \dots, w_{n-r}, v_1, \dots, v_{r+1}\} \subset \mathbb{R}^n$ consists of $n+1$ vectors and hence is linearly dependent. That is, there exists $x \in \mathbb{R}^{n-r}$ and $y \in \mathbb{R}^{r+1}$ such that

$$0 = x_1 w_1 + \cdots x_{n-r} w_{n-r} + y_1 v_1 + \cdots y_{r+1} v_{r+1} = Wx + V_{r+1} y.$$

   By rescaling x and y, we may assume that $\|x\|_2 = 1$. Since W and V have orthonormal columns,

$$\|y\|_2 = \|V_{r+1} y\|_2 = \|Wx\|_2 = \|x\|_2 = 1.$$

4. Since U has orthonormal columns, $\|U\Sigma V^\top Wx\|_2 = \|\Sigma V^\top Wx\|_2$. Note that
$$Wx = -V_{r+1}y \implies V^\top Wx = -V^\top V_{r+1}y = -\begin{bmatrix} I_{r+1} \\ 0 \end{bmatrix} y$$
and so $\Sigma V^\top Wx = -\Sigma_{r+1}y$, where $\Sigma_{r+1} := \mathrm{diag}(\sigma_1, \ldots, \sigma_{r+1})$. Consequently,
$$\|\Sigma V^\top Wx\|_2^2 = \sigma_1^2 y_1^2 + \cdots + \sigma_{r+1} y_{r+1}^2 \geq \sigma_{r+1}^2 \|y\|_2^2 = \sigma_{r+1}^2.$$

5. Combining everything together, we have $\|A - B\|_2 \geq \|AW\|_2 = \|\Sigma V^\top W\|_2 = \|\Sigma_{r+1}y\|_2 \geq \sigma_{r+1}$, which completes the proof. □

In fact, more generally it is known that
$$\|A - A_r\| \leq \|A - B\| \tag{3}$$
for any so-called unitarily invariant norm $\|\cdot\|$: $\|A\| = \|Q_1 A Q_2\|$ for all orthogonal $Q_1 \in \mathbb{R}^{m\times m}$ and $Q_2 \in \mathbb{R}^{n\times n}$ (non-examinable).

In many applications $\sigma_{r+1} \ll \sigma_1$ for some $r \ll n$, in which case $A \approx U_r \Sigma_r V_r^\top$. 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!

Application of low-rank approximation

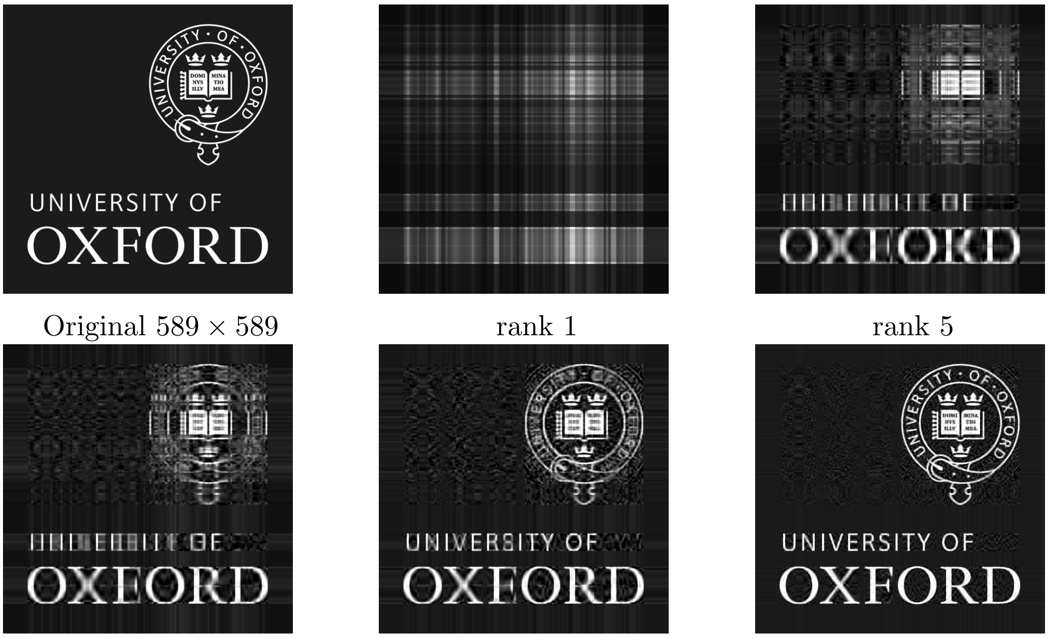
A traditional example to illustrate low-rank approximation via the truncated SVD is image compression. A grayscale image can be represented by a matrix A, with each entry representing the intensity of a pixel. One can then approximate A by a truncated SVD, and use that to get a compressed image that 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 significant. (This is however not how images are usually compressed in practice; e.g. the algorithm behind the jpg format is completely different).

Concluding remarks on factorisations

The SVD $A = U\Sigma V^\top$ and symmetric eigenvalue decomposition $A = V\Lambda V^\top$ 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 $A^\top A$ (this is not exactly how the SVD is computed in practice, but that is outside the scope).

Thus far, we have discussed three different matrix factorisations: LU, QR, and SVD. A summary of some of their properties is below:



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.

	LU	QR	SVD
Square A	Y	Y	Y
Rectangular A	N (Y)	Y	Y
Solve $Ax = b$	Y	Y	Y
Solve $\min_x \|Ax - b\|_2$	N	Y	Y

For the remainder of the linear algebra section of the course, we will turn to eigenvalue problems $Ax = \lambda x$ and describe an algorithm for solving them.