Singular Value Decomposition (SVD)

Recommended reading:

- Lectures 4, 5 in [4]
- Sections I.8 and I.9 in [3]

Literature:

[1] R. Rannacher.
Numerik 0 - Einführung in die Numerische Mathematik.
Heidelberg University Publishing, 2017.

[2] G. Strang.
Introduction to Linear Algebra.
Wellesley-Cambridge Press, 2003.

[3] G. Strang.
Linear Algebra and Learning from Data.
Wellesley-Cambridge Press, 2019.

[4] L.N. Trefethen and D. Bau.
Numerical linear algebra.
SIAM, Soc. for Industrial and Applied Math., Philadelphia, 1997.

4 Singular Values and the Singular Value Decomposition (SVD)

We will extend the concept of eigenvalues and eigenvectors to general matrices $A \in \mathbb{R}^{m \times n}$.

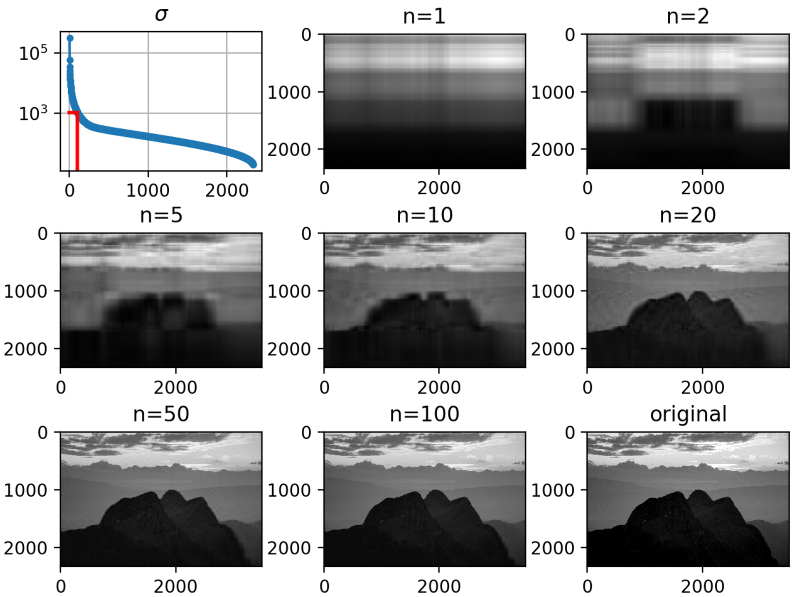
4.1 Motivation and Introduction

Gilbert Strang: *"The SVD $A = U\Sigma V^\top$ is the* ***most important*** *theorem in data science."*
([3] Linear Algebra and Learning from Data, p.31)

Importance and Applications:

- The SVD of a matrix reveals many properties about the matrix itself (representation of the image and kernel, rank, invertibility, condition,...)
- Low-Rank Approximation
  - Data compression (e.g., image data)
  - Principal Component Analysis
- Pseudoinverse (generalization of the inverse matrix) and relation to the minimum-norm least squares solution

Image and data compression:



3500×2333 greyscale image is interpreted as matrix

$$A \in [0,1]^{3500\times 2333}.$$

The singular values are shown in the figure with the title "σ".
The reconstructed image with the first 100 singular values only, i.e.,

$$A_{100} := U\mathrm{diag}(\sigma_1,\ldots,\sigma_{100},0,\ldots,0)V^\top$$

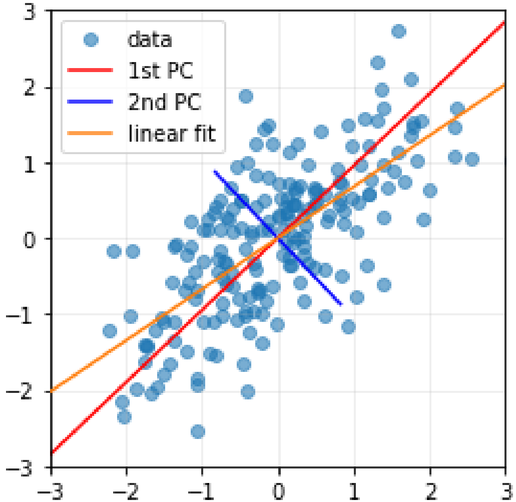
is quite close to the original image but takes only

$$\frac{3500\cdot 100 + 100 + 100\cdot 2333}{3500\cdot 2333} \approx 7\%$$

of the storage space.

Principal Component Analysis

Under the correct setup we have that the SVD equals the PCA, whose aim is dimension reduction:



The data represented by the blue dots can be fully explained by the red and blue line. However the red line might already capture a substantial part of the data's variance.

The Singular Value Decomposition (SVD)

For matrices $A \in \mathbb{R}^{m\times n}$ of general format, the equation $Av = \lambda v$ fails. Instead we define:

Definition 4.1 (*Singular Values and Vectors*) *Let $A \in \mathbb{R}^{m\times n}$ be a matrix. Then a positive number $\sigma > 0$ is called* ***singular value****, if there exist nonzero vectors $v \in \mathbb{R}^n \setminus \{0\}$ and $u \in \mathbb{R}^m \setminus \{0\}$, such that*

$$Av = \sigma u \quad \text{and} \quad A^\top u = \sigma v. \tag{4}$$

The vectors v and u are called right and left ***singular vectors of*** *A to the singular value σ.*

This will lead to the impactful theorem of the singular value decomposition:

Theorem 4.2 (*Singular value decomposition (SVD)*) *Let $A \in \mathbb{R}^{m\times n}$. Then there are orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ as well as a diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) \in \mathbb{R}^{m\times n}$, where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, $r \leq \min\{m, n\}$, are the sorted positive singular values, such that*

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

which is the so-called ***singular value decomposition of*** *A .*

4.2 Preparing Results

In order to understand and prove this central theorem we will put a few auxiliary results into position. The first one is about eigenvalues of symmetric and positive semi-definite matrices:

Lemma 4.3 (*Eigenvalues and Positivity*) *Let $B \in \mathbb{R}^{n \times n}$ be symmetric and positive definite (semi-definite), then $\lambda > 0$ (≥ 0) for all eigenvalues $\lambda \in \sigma(B)$.*

The next result is about the shared eigenvalues of product matrices:

Lemma 4.4 (*Shared Eigenvalues of Products*) *Let $A \in \mathbb{F}^{m \times n}$ and $B \in \mathbb{F}^{n \times m}$. Then the products $AB \in \mathbb{F}^{m \times m}$ and $BA \in \mathbb{F}^{n \times n}$ have the same nonzero eigenvalues.*

Remark:

- If $m \neq n$, then BA and AB have differently many eigenvalues. However the nonzero eigenvalues are the same. Thus both product matrices have at most $\ell := \min\{m, n\}$ nonzero eigenvalues!
- In the special case that $m = n$ and B invertible, we observe

$$B^{-1}(BA)B = (AB),$$

identifying the matrices AB and BA as being similar!

Now a special instance of the latter two results (choosing $B = A^\top$) leads us to the key lemma to prove the SVD Theorem 4.2:

Lemma 4.5 *Let $A \in \mathbb{R}^{m\times n}$, then the matrices $A^\top A$ and $AA^\top$ are symmetric, positive semi-definite and have the same positive eigenvalues.*

Remark:
Due to the symmetry of $A^\top A$ and $AA^\top$ we also know that we find orthonormal eigenvectors $v_1, \dots, v_n$ and $u_1, \dots, u_m$! The SVD will connect them!

4.3 From Reduced to Full SVD

Full, Reduced and Truncated SVD

The four fundamental subspaces revisited:

Summary and Remarks

$$A = \begin{pmatrix} | & & | & | & & | \\ u_1 & \cdots & u_r & u_{r+1} & \cdots & u_m \\ | & & | & | & & | \end{pmatrix} \left(\begin{array}{ccc|ccc} \sigma_1 & & & & \vdots & \\ & \ddots & & \cdots & 0 & \cdots \\ & & \sigma_r & & \vdots & \\ \hline & \vdots & & & \vdots & \\ \cdots & 0 & \cdots & \cdots & 0 & \cdots \\ & \vdots & & & \vdots & \end{array}\right) \begin{pmatrix} - & v_1 & - \\ & \vdots & \\ - & v_r & - \\ - & v_{r+1} & - \\ & \vdots & \\ - & v_n & - \end{pmatrix}$$

- we can show $\operatorname{Im}(A) = \operatorname{span}(u_1, \ldots, u_r)$ and $\ker(A) = \operatorname{span}(v_{r+1}, \ldots, v_n)$, in particular

$$\operatorname{rank}(A) = r$$

- columns of V are orthonormal eigenvectors of $A^\top A \in \mathbb{R}^{n \times n}$ and $A^\top A = V(\Sigma^\top \Sigma)V^\top$
- columns of U are orthonormal eigenvectors of $AA^\top \in \mathbb{R}^{m \times m}$ and $AA^\top = U(\Sigma\Sigma^\top)U^\top$
- σ_1^2 to σ_r^2 are the shared positive eigenvalues of both $A^\top A$ and $AA^\top$
- an SVD of the transpose $A^\top$ is easily found by

$$A^\top = (U\Sigma V^\top)^\top = V\Sigma^\top U^\top$$

- for square matrices singular values and eigenvalues are different in general, take for example $A = -I$
- however, for symmetric matrices $A = Q\Lambda Q^\top$, the singular values are the absolute values of the eigenvalues, i.e., $\sigma_i = \sqrt{\lambda_i^2}$ (see exercises)

Example 4.6 (*SVD by hand*)

Example: rank-1 pieces

Let $x \in \mathbb{R}^m \setminus \{0\}$ and $y \in \mathbb{R}^n \setminus \{0\}$, then

$$A := xy^\top = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} (y_1, \ldots, y_n) = \begin{pmatrix} | & & | \\ y_1 x & \cdots & y_n x \\ | & & | \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

What is the SVD of A?

4.4 The Geometry of the SVD

[Compare to the geometry of the eigendecomposition]

4.5 Matrix condition and rank

Situation:

Let $A = U\Sigma V^\top \in \mathbb{R}^{n\times n}$ be invertible (i.e., $\sigma_i \neq 0\ \forall i$) and assume we want to solve $Ax = b$. We also assume that the data is corrupted $\tilde{b} = b + \Delta b$ by some error Δb.

$\Rightarrow$ We obtain a perturbed solution $\tilde{x} = x + \Delta x$ with $\Delta x = A^{-1}\Delta b$.

Question:

How severe is the propagation of *data error* Δb to the resulting *solution error* Δx?
$\rightarrow$ Singular (eigen-) values give us this information!

Definition 4.7 (*Condition number*) *Let $A \in \mathbb{R}^{n \times n}$ be a matrix. Then we call*

$$\mathit{cond}_2(A) := \frac{\max\{\sigma_i\}}{\min\{\sigma_i\}}$$

*the **condition number** of the matrix A.*

Special Case: Symmetric Matrices (exercise)

Let $A \in \mathbb{R}^{n \times n}$ be a real symmetric matrix, then

$$\text{cond}_2(A) = \frac{\max\{|\lambda| \colon \lambda \in \sigma(A)\}}{\min\{|\lambda| \colon \lambda \in \sigma(A)\}}.$$

Remark:

If some of the singular values are actually zero or close to zero, the condition number is (almost) ∞. In this case, we cannot trust any numerical solver (for $Ax = b$) in finite precision, as errors in the data b (e.g., also due to rounding errors) may severely propagate to the computed solution x.

We also call such matrices *rank deficient*.

4.6 The Truncated SVD and its Best Approximation Property

Motivation:
Let the singular values be sorted $\sigma_1 \geq \ldots \geq \sigma_r > 0$, $r := \mathrm{rank}(A)$, then the reduced SVD reads as

$$A = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top + \cdots + \sigma_i u_i v_i^\top + \cdots + \sigma_{r-1} u_{r-1} v_{r-1}^\top + \sigma_r u_r v_r^\top$$

If a σ_i is small, then the matrix $u_i v_i^\top$ does not contribute much to A, and similarly for $\sigma_{i+1}, \ldots, \sigma_r$.

What about leaving them out?

This gives rise to the following definition:

Definition 4.8 (*Truncated SVD*) *Let $A = U\Sigma V^\top \in \mathbb{R}^{m\times n}$. For $k < r := \mathrm{rank}(A)$ define $\Sigma_k := \mathrm{diag}(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{k\times k}$, $U_k := [u_1, \ldots, u_k] \in \mathbb{R}^{m\times k}$ and $V_k := [v_1, \ldots, v_k] \in \mathbb{R}^{n\times k}$. Then*

$$A_k := U \mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0 \ldots, 0) V^\top = U_k \Sigma_k V_k^\top$$

*is called **truncated SVD of** A.*

We observe that

$$\mathrm{rank}(A_k) = k,$$

which is why A_k is also called *rank-k-approximation of* A.

Question: Leaving out some rank-1 summands, how much do we deviate from the original matrix?

With other words: In which sense does $A_k \in \mathbb{R}^{m\times n}$ *approximate* $A \in \mathbb{R}^{m\times n}$?

We first need to quantify the distance between matrices, i.e., we need a *norm* for matrices in $\mathbb{R}^{m\times n}$!

Here we consider the so–called Frobenius norm:
If we reshape a matrix $A \in \mathbb{R}^{m\times n}$ into a vector $v \in \mathbb{R}^{m\cdot n}$ (e.g., $v_{[(j-1)\cdot m+i]} := a_{ij}$), then we can use our norms for vectors, e.g.,

$$\|A\|_F := \|v\|_2.$$

This is precisely:
Definition 4.9 (*Frobenius norm*) *For any matrix $A \in \mathbb{R}^{m\times n}$, the* ***Frobenius norm*** *is defined as*

$$\|A\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}.$$

Exercise:

- One can show that
$$\|A\|_F^2 = \mathrm{tr}(A^\top A),$$
where tr:="trace" denotes the sum of the diagonal entries.

- Using this fact, for $A = U\Sigma V^\top$ with $r = \mathrm{rank}(A)$ we also find
$$\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i^2.$$

Finally, the truncated SVD satisfies a best approximation property:

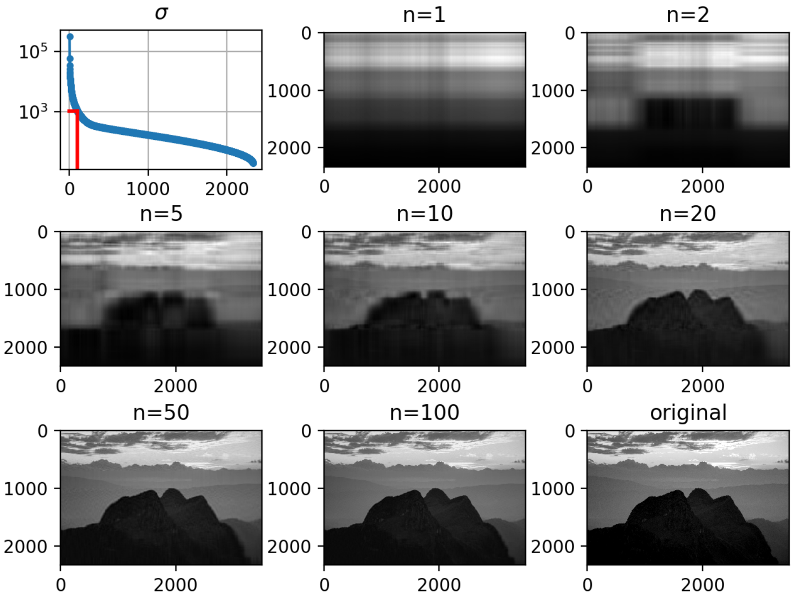
Theorem 4.10 (*Eckart-Young-Mirsky*) *Let $A \in \mathbb{R}^{m\times n}$ with SVD $A = U\Sigma V^\top$ and let $k \leq$ rank(A). Then, the truncated SVD A_k is the best approximation in the Frobenius norm among all matrices with rank k, i.e.*

$$\|A - A_k\|_F \leq \|A - B\|_F\,, \quad \forall B \in \mathbb{R}^{m\times n}, \text{rank}(B) = k.$$

In words:

Among all matrices with rank k, the truncated SVD is closest to A.

4.6.1 Image and Data Compression



3500×2333 greyscale image is interpreted as matrix

$$A \in [0,1]^{3500 \times 2333}.$$

The singular values are shown in the figure with the title "σ".
The reconstructed image with the first 100 singular values only, i.e.,

$$A_{100} := U\mathrm{diag}(\sigma_1, \ldots, \sigma_{100}, 0, \ldots, 0)V^\top$$

is quite close to the original image but takes only

$$\frac{3500 \cdot 100 + 100 + 100 \cdot 2333}{3500 \cdot 2333} \approx 7\%$$

of the storage space.

Note: The storage of A_k in general is $k \cdot (m + 1 + n)$.

Note: The same data compression can be performed with any matrix — and similarly with tensors.

4.6.2 Principal Component Analysis (PCA)

4.6.3 Pseudoinverses

With the help of the SVD one can define a generalized concept of an inverse matrix, called the *pseudoinverse*. This is closely related to the minimum-norm least-squares solution, so that we postpone a discussion to the section on least squares.

4.7 Numerical Computation of the SVD