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 $A \in \mathbb{R}^{m \times n}$.

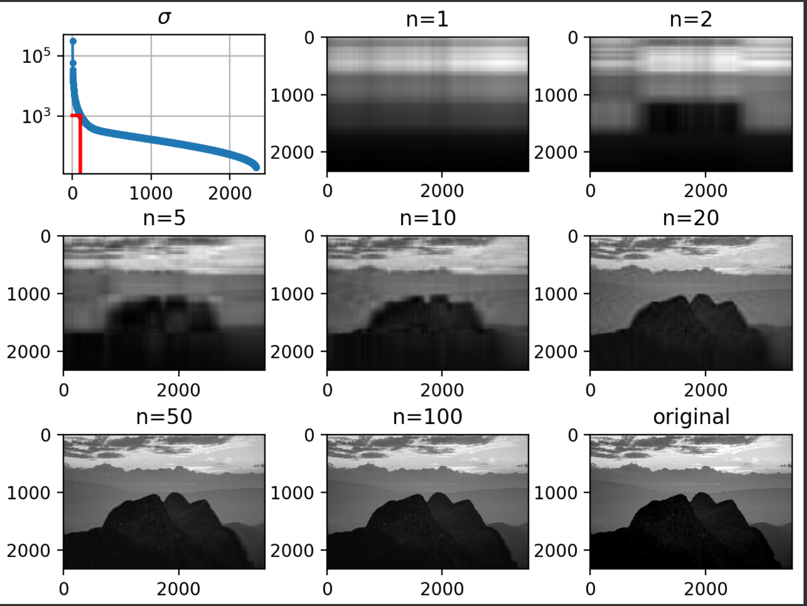
4.1 Motivation and Introduction

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.,

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

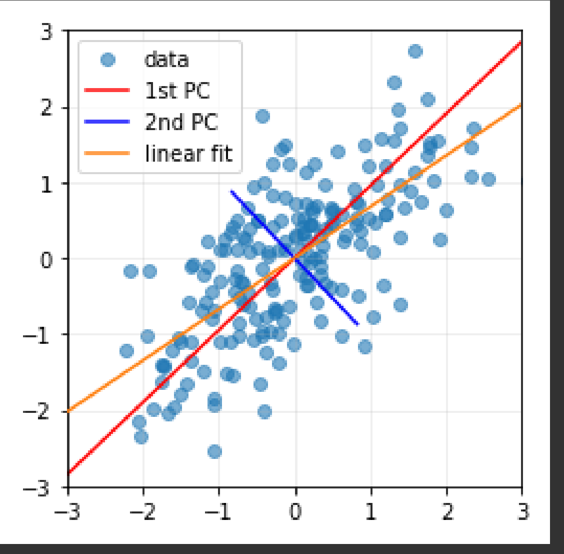
is quite close to the original image but takes only

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

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.*

Assume we had singular vectors v_i, u_i and values σ_i and put them into matrices V, U, Σ (as we did for the eigendecomposition). Then we find

$$AV = U\Sigma$$

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 $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0) \in \mathbb{R}^{m\times n}$, where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$, $r \leq \min\{m, n\}$, 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 $\lambda \in \sigma(B)$.*

Proof. First of all we note that due to symmetry $\sigma(B) \subset \mathbb{R}$ and we can choose eigenvectors with real coefficients. We now perform a proof by contradiction:
Let B be $\underbrace{\text{positive definite}}_{:\Leftrightarrow x^\top Bx>0\ \forall x\neq 0}$ and assume $\lambda \leq 0$ for some $\underbrace{\lambda \in \sigma(B)}_{:\Leftrightarrow \exists v\neq 0: Bv=\lambda v}$ with eigenvector $v \in \mathbb{R}^n, v \neq 0$.
Then we find

$$v^\top \underbrace{Bv}_{=\lambda v} = \lambda v^\top v = \underbrace{\lambda}_{\leq 0} \underbrace{\|v\|_2^2}_{>0} \leq 0 \quad \text{[contradiction to the positivity of A].}$$

(Analogous proof for B positive semi-definite.)
(Alternative proof via Rayleigh quotient.)

□

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.*

Proof. We prove this by mutual subset relation:
First let $\lambda \in \sigma(AB), \lambda \neq 0$ be a nonzero eigenvalue of AB with eigenvector $v \in \mathbb{F}^n, v \neq 0$, i.e.,

$$ABv = \lambda v.$$

Now multiply both sides by B to obtain

$$BA(Bv) = \lambda Bv,$$

which implies that Bv is an eigenvector of BA with the *same* eigenvalue λ. To see this, note that $\lambda \neq 0$ implies that $ABv = \lambda v \neq 0$ and thus $Bv \neq 0$.

Similarly, let now $\lambda \in \sigma(BA), \lambda \neq 0$ be a nonzero eigenvalue of BA with eigenvector $v \in \mathbb{F}^n, v \neq 0$, i.e., $BAv = \lambda v$. Then we multiply both sides by A to proceed along the same lines. □

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.*

Proof. We find:

1) Symmetry: $(A^\top A)^\top = A^\top (A^\top)^\top = A^\top A$ and similarly $(AA^\top)^\top = AA^\top$
2) p(s)d: $x^\top A^\top A x = \|Ax\|_2^2 \geq 0, \quad x^\top AA^\top x = \|A^\top x\|_2^2 \geq 0$
3) The same positive eigenvalues:
   - By Lemma 4.3 we know that the matrices only have nonnegative eigenvalues
   - By lemma 4.4 we know that the nonzero, i.e., positive, eigenvalues are the same

□

Remark:

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

4.3 From Reduced to Full SVD

Recall:

- $\operatorname{Im}(A) \perp \ker(A^\top)$ and $\operatorname{Im}(A^\top) \perp \ker(A)$
- $A^\top A$, $AA^\top$ are
  - symmetric $\Rightarrow$ real eigenvalues and we find orthonormal basis of eigenvectors
  - positive semi-definite $\Rightarrow$ their eigenvalues are nonnegative, i.e., $\lambda \geq 0$
  - they have the *same* positive eigenvalues λ_i for $1 \leq i \leq r \leq \min(m, n)$
  - $\ker(A) = \ker(A^\top A)$ and $\ker(A^\top) = \ker(AA^\top)$

Proof of SVD: We are looking for nonzero vectors $u \in \mathbb{R}^m, v \in \mathbb{R}^n$ and positive numbers $\sigma > 0$, such that

$$Av = \sigma u \quad \Longleftrightarrow \quad u = \frac{1}{\sigma} Av \in \operatorname{Im}(A), \tag{5}$$

$$A^\top u = \sigma v \quad \Longleftrightarrow \quad v = \frac{1}{\sigma} A^\top u \in \operatorname{Im}(A^\top). \tag{6}$$

1) So we have two equations for two unknown vectors. By inserting one into the other we obtain two equivalent formulations (this is *elimination*). Here, we insert (5) into (6) which gives

$$A^\top A v = \sigma^2 v \quad \Longleftrightarrow \quad (\sigma^2, v) \text{ eigenpair of } A^\top A. \tag{7}$$

(Note: Inserting (6) into (5) would give (σ^2, u) eigenpair of $AA^\top$)
2) Let $\lambda_1, \ldots, \lambda_r > 0$ $(r \leq \min(m, n))$ be the positive eigenvalues of $A^\top A$ with orthonormal eigenvectors $v_1, \ldots, v_r$ $(\in \operatorname{Im}(A^\top))$. Then according to (5) and (7) we set

$$\sigma_i := \sqrt{\lambda_i}, \qquad u_i := \frac{1}{\sigma_i} A v_i \; (\in \operatorname{Im}(A)).$$

We then find:

- By construction v_i, u_i are singular vectors to the singular value σ_i, i.e., we have

$$Av_i = \sigma_i u_i$$

  and indeed

$$A^\top u_i = \frac{1}{\sigma_i} \underbrace{A^\top A v_i}_{=\lambda_i v_i} = \frac{\lambda_i}{\sigma_i} v_i = \sigma_i v_i.$$

- For the SVD we want the u_i to be orthonormal. Let us check this:

$$u_i^\top u_j = \frac{1}{\sigma_i} \frac{1}{\sigma_j} (Av_i)^\top A v_j = \frac{1}{\sigma_i} \frac{1}{\sigma_j} v_i^\top \underbrace{A^\top A v_j}_{=\lambda_j v_j} = \underbrace{\frac{\sigma_j}{\sigma_i}}_{} \underbrace{v_i^\top v_j}_{=\delta_{ij}} = \delta_{ij}.$$

3) Finally, choose orthonormal bases

$$v_{r+1}, \dots, v_n \in \ker(A) \quad (\perp \operatorname{Im}(A^\top)),$$
$$u_{r+1}, \dots, u_m \in \ker(A^\top) \quad (\perp \operatorname{Im}(A)).$$

We note that these are eigenvectors of $A^\top A$ and $AA^\top$, respectively, to the eigenvalue 0. Then let us collect everything:

$$V := \begin{pmatrix} \underbrace{\begin{matrix} | & & | \\ v_1 & \cdots & v_r \\ | & & | \end{matrix}}_{=V_r,\ \in \mathrm{Im}(A^\top)} & \underbrace{\begin{matrix} | & & | \\ v_{r+1} & \cdots & v_n \\ | & & | \end{matrix}}_{\in \ker(A)} \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad U := \begin{pmatrix} \underbrace{\begin{matrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{matrix}}_{=U_r,\ \in \mathrm{Im}(A)} & \underbrace{\begin{matrix} | & & | \\ u_{r+1} & \cdots & u_m \\ | & & | \end{matrix}}_{\in \ker(A^\top)} \end{pmatrix} \in \mathbb{R}^{m \times m}$$

$$\Sigma := \left(\begin{array}{ccc|ccc} \sigma_1 & & & & \vdots & \\ & \ddots & & \cdots & 0 & \cdots \\ & & \sigma_r & & \vdots & \\ \hline & \vdots & & & \vdots & \\ \cdots & 0 & \cdots & \cdots & 0 & \cdots \\ & \vdots & & & \vdots & \end{array}\right) \in \mathbb{R}^{m \times n}.$$

With $\Sigma_r := \operatorname{diag}(\sigma_1, \dots, \sigma_r)$ we can write

$$AV = (AV_r | 0) = (U_r \Sigma_r | 0) = U\Sigma.$$

Now, since $V \in \mathbb{R}^{n \times n}$ is orthogonal (i.e., $V^{-1} = V^\top$), we can multiply with $V^\top$ from the right and finally obtain the desired SVD

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

Remark: The zeros in Σ may justify to also allow for zero singular values $\sigma_{r+1} = \ldots = \sigma_\ell = 0$ with $\ell = \min(m, n)$ in Definition 4.1. However, we require singular values to be positive here. At this point the literature is not uniform.

Full, Reduced and Truncated SVD

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

$$= (U_r\Sigma_r \mid 0) \begin{pmatrix} V_r^\top \\ - \\ * \end{pmatrix}$$

$$= U_r\Sigma_r V_r^\top \qquad \text{(reduced SVD)}$$

$$= \begin{pmatrix} | & & | \\ \sigma_1 u_1 & \cdots & \sigma_r u_r \\ | & & | \end{pmatrix} \begin{pmatrix} - & v_1 & - \\ & \vdots & \\ - & v_r & - \end{pmatrix}$$

$$= \sum_{j=1}^{r} \sigma_j u_j v_j^\top = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top + \cdots + \sigma_r u_r v_r^\top \qquad \text{(sum of rank-1 matrices)}$$

$$\approx \sigma_1 u_1 v_1^\top + \cdots + \sigma_k u_k v_k^\top \qquad \text{(truncated SVD } (k < r))$$

The four fundamental subspaces revisited:

By Lemma ?? (note: $U_r\Sigma_r$ is injective and $\Sigma_r V_r^\top$ is surjective) we find

$$\operatorname{Im}(A) = \operatorname{Im}(U_r\Sigma_r V_r^\top) = \operatorname{Im}(U_r) = \operatorname{span}(u_1,\dots,u_r),$$
$$\ker(A) = \ker(U_r\Sigma_r V_r^\top) = \ker(V_r^\top) = \operatorname{Im}(V_r)^\perp = \operatorname{span}(v_{r+1},\dots,v_n)$$

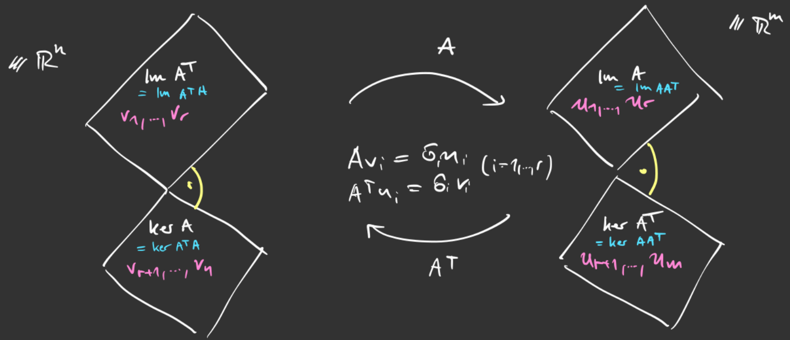
and by considering $A^\top = V\Sigma^\top U^\top$we find

$$\operatorname{Im}(A^\top) = \operatorname{span}(v_1,\dots,v_r),$$
$$\ker(A^\top) = \operatorname{span}(u_{r+1},\dots,u_m).$$

With other words:

The SVD contains orthonormal bases for all four fundamental subspaces.
And even more than that, they are connected via

$$Av = \sigma u, \quad A^\top u = \sigma v.$$



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 $\mathrm{Im}(A) = \mathrm{span}(u_1, \ldots, u_r)$ and $\ker(A) = \mathrm{span}(v_{r+1}, \ldots, v_n)$, in particular

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

- 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*)

$A = \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix}, A^\top = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$

$$A^\top A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix} = \begin{bmatrix} 17 & 8 \\ 8 & 17 \end{bmatrix}$$

- Compute eigenvalues of $A^\top A$:

$$0 \overset{!}{=} \det(A^\top A - \lambda I) = \det \begin{pmatrix} 17-\lambda & 8 \\ 8 & 17-\lambda \end{pmatrix} = (17-\lambda)^2 - 64$$

$$\begin{aligned} &\Leftrightarrow \quad 17-\lambda = \pm 8 \\ &\Leftrightarrow \quad \lambda = 17 \pm 8 \\ &\Leftrightarrow \quad \lambda_1 = 25, \lambda_2 = 9 \end{aligned}$$

- Compute corresponding normalized eigenvectors:
  a) $(A^\top A - \lambda_1 I)v_1 = \begin{pmatrix} -8 & 8 \\ 8 & -8 \end{pmatrix} v_1 \overset{!}{=} 0 \quad \Rightarrow \quad v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
  b) $(A^\top A - \lambda_2 I)v_2 = \begin{pmatrix} 8 & 8 \\ 8 & 8 \end{pmatrix} v_2 \overset{!}{=} 0 \quad \Rightarrow \quad v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

- Compute left singular vectors:

$$\sigma_1 := \sqrt{\lambda_1} = 5, \qquad \sigma_2 := \sqrt{\lambda_2} = 3,$$

$$\begin{aligned} u_1 &:= \frac{1}{\sigma_1} A v_1 \\ &= \frac{1}{5}\frac{1}{\sqrt{2}} \begin{pmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \frac{1}{5\sqrt{2}} \begin{pmatrix} 5 \\ 5 \\ 0 \end{pmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \end{aligned} \qquad \begin{aligned} u_2 &:= \frac{1}{\sigma_2} A v_2 \\ &= \frac{1}{3}\frac{1}{\sqrt{2}} \begin{pmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &= \frac{1}{3\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 4 \end{pmatrix} \end{aligned}$$

Find $u_3 \in \ker(A^\top)$:

$$A^\top u_3 = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix} \begin{pmatrix} u_3^1 \\ u_3^2 \\ u_3^3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$u_3 = \frac{1}{3} \begin{pmatrix} 2 \\ -2 \\ -1 \end{pmatrix}$$

All in all:

$$V = \begin{pmatrix} | & | \\ v_1 & v_2 \\ | & | \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \in \mathbb{R}^{n \times n} = \mathbb{R}^{2 \times 2}$$

$$U = \begin{pmatrix} | & | & | \\ u_1 & u_2 & u_3 \\ | & | & | \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} & \frac{2}{3} \\ \frac{1}{\sqrt{2}} & -\frac{1}{3\sqrt{2}} & -\frac{2}{3} \\ 0 & \frac{4}{3\sqrt{2}} & -\frac{1}{3} \end{pmatrix} \in \mathbb{R}^{m \times m} = \mathbb{R}^{3 \times 3}$$

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 5 & 0 \\ 0 & 3 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m \times n} = \mathbb{R}^{3 \times 2}$$

$$\Rightarrow \quad A = U \Sigma V^\top$$

Example: rank-1 pieces

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

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

What is the SVD of A?

$$A^\top A = (xy^\top)^\top xy^\top = y \underbrace{x^\top x}_{=\|x\|^2} y^\top = \|x\|^2 yy^\top$$

Compute eigenpairs: We find $A^\top Ay = \|x\|^2 y \underbrace{y^\top y}_{=\|y\|^2} = \|x\|^2\|y\|^2 y$

$v_1 := \frac{y}{\|y\|}$ is eigenvector to the eigenvalue $\lambda_1 := \|x\|^2\|y\|^2$

Set

$$\sigma_1 := \sqrt{\lambda_1} \overset{(\neq 0, \text{since} x \neq 0 \neq y)}{=} \|x\|\,\|y\|$$

and

$$u_1 := \frac{1}{\sigma_1} A v_1 = \frac{1}{\|x\|\|y\|} xy^\top \frac{y}{\|y\|} = \frac{x}{\|x\|}$$

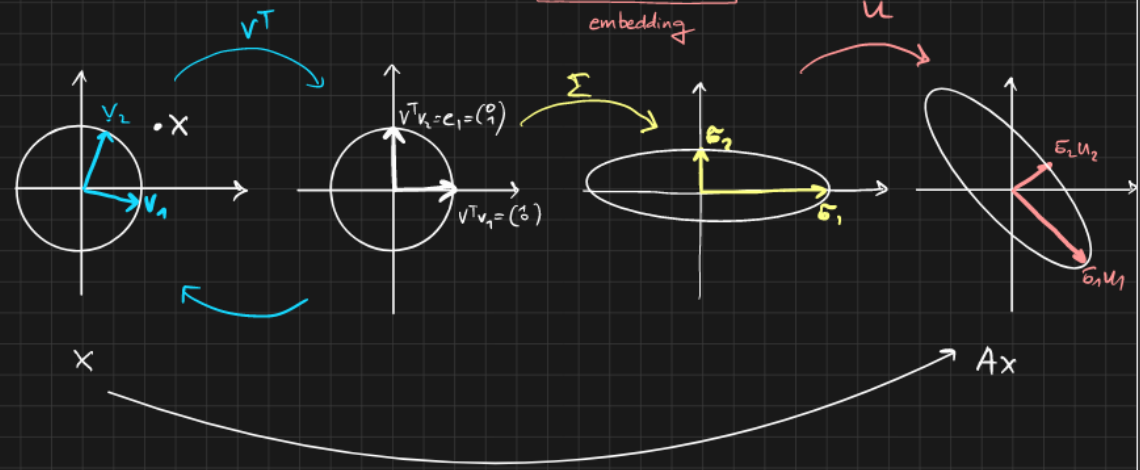
then

$$A = U\Sigma V^\top = \frac{x}{\|x\|}(\|x\|\,\|y\|)\frac{y^\top}{\|y\|} = xy^\top \ \checkmark \quad (\to r = 1, \text{ thus } \operatorname{rank}(A) = 1)$$

4.4 The Geometry of the SVD

[Compare to the geometry of the eigendecomposition]

∘ GEOMETRY: $$Ax = \underbrace{U \cdot \underbrace{\Sigma \cdot \underbrace{V^T \cdot x}_{\text{projection}}}_{\text{scaling}}}_{\text{embedding}}$$



$$x = \mu_1 v_1 + \mu_2 v_2$$

$$\Rightarrow \quad Ax = \mu_1 \sigma_1 u_1 + \mu_2 \sigma_2 u_2$$

- The orthonormal bases V and U are connected via $Av_j = \sigma_j u_j$.
- Using these orthonormal bases, one can regard *any* matrix as a diagonal matrix.

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!

$$b = Ax \Rightarrow \|b\|_2 = \|Ax\|_2 = \|U\Sigma V^\top x\|_2 = \|\Sigma V^\top x\|_2 = \|\Sigma_{j=1}^{r}\sigma_j v_j^\top x\|_2 \leq \sigma_1 \|V^\top x\|_2 = \sigma_1\|x\|_2$$

$$\Delta x = A^{-1}\Delta b \Rightarrow \|\Delta x\|_2 = \|A^{-1}\Delta b\|_2 = \|V\Sigma^{-1}U^\top \Delta b\|_2 = \|\Sigma^{-1}U^\top\Delta b\|_2 \leq \frac{1}{\sigma_n}\|\Delta b\|_2$$

$$\Rightarrow \frac{\|\Delta x\|_2}{\|x\|_2} \leq \frac{1}{\sigma_n}\frac{\|\Delta b\|_2}{\|x\|_2} \leq \frac{\sigma_1}{\sigma_n}\frac{\|\Delta b\|_2}{\|b\|_2}$$

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

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

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 $\sigma_1 \geq \ldots \geq \sigma_r > 0$, $r := \operatorname{rank}(A)$, 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

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

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 $k \leq \text{rank}(A)$. 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.

Proof. We use the so-called Weyl inequality (see (8) below): For matrices $C, D \in \mathbb{R}^{m \times n}$ with decreasingly ordered singular values, we denote by $\sigma_i(C), \sigma_i(C), \sigma_i(C + D)$ the i-th singular value of the respective matrix. Then Weyl's inequality gives us the relation

$$\sigma_{i+\ell-1}(C + D) \leq \sigma_i(C) + \sigma_\ell(D), \text{ with } i, \ell, i + \ell - 1 \in \{1, \dots, p\}\,,\ p := \min\{m, n\}. \tag{8}$$

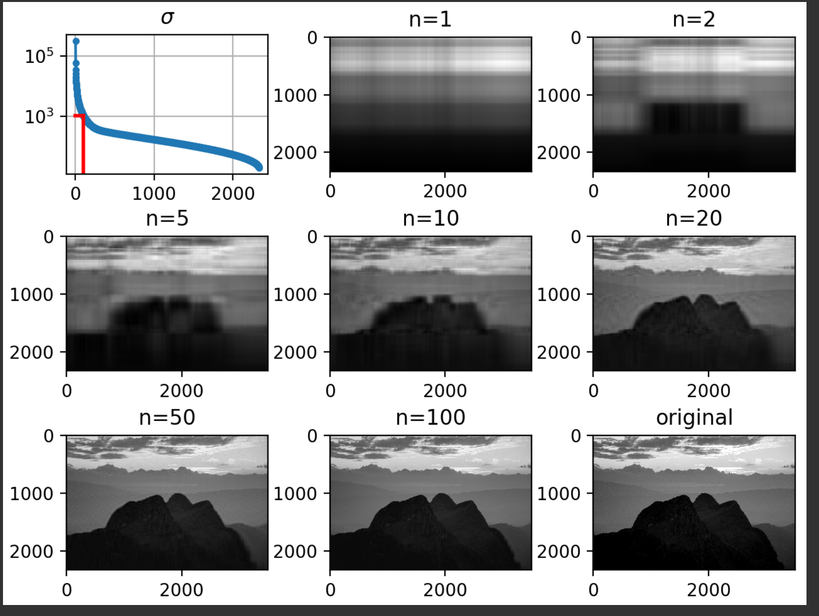
We assume $\text{rank}(B) = k$, which results in $\sigma_l(B) = 0$ for $l > k$ and thus we conclude from Weyl's inequality (8) for $C := A - B, D := B, \ell := k + 1$ that

$$\sigma_{i+k}(A) \leq \sigma_i(A - B) + \sigma_{k+1}(B) = \sigma_i(A - B) \text{ for } i = 1, \dots, p - k$$

$$\Rightarrow \|A - B\|_F^2 = \sum_{i=1}^{p} \sigma_i(A - B)^2 \geq \sum_{i=1}^{p-k} \sigma_i(A - B)^2 \geq \sum_{i=k+1}^{p} \sigma_i(A)^2 = \|A - A_k\|_F^2$$

for all B with $\text{rank}(B) = k$. $\square$

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.,

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

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 $k\cdot(m+1+n)$.

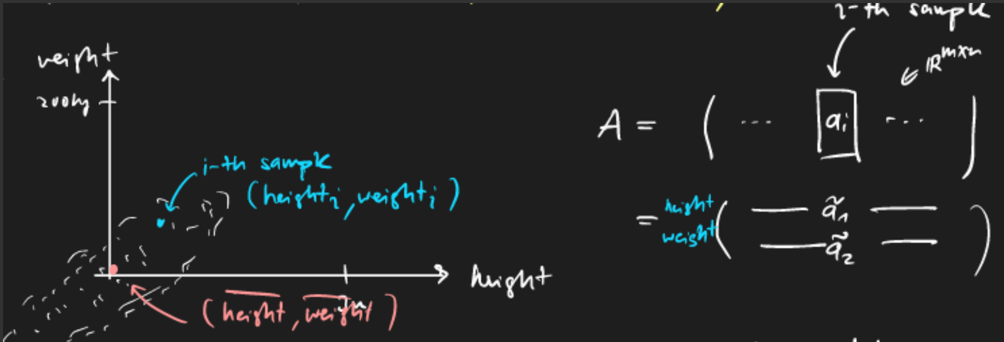
Note: The same data compression can be performed with any matrix — and similarly with tensors.

4.6.2 Principal Component Analysis (PCA)

Situation:

n measurements / samples (e.g., questioning n persons)
m features / variables (e.g., height and weight)

Example:



Without loss of generality we can center the data by substracting the mean from each sample

Observation:

Height and weight are proportional in some sense (i.e., they correlate), however there is some spread/variance.

Aim:

Can we explain "most" of the variance with a lower dimensional subspace?
(In the example above, e.g., a line may capture most of the variance)

More on the statistics: ($\mathrm{Var}(X) = E(X - E(X))^2$)

statistical variance = "normalized" sum of squared distances from the mean

$$\text{statistical variance in height} = \frac{1}{n-1}\sum_{i=1}^{n}(\text{height}_i - \underbrace{\overline{\text{height}}}_{\text{w.l.o.g.}=0})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\widetilde{\text{height}_i}^2 = \frac{1}{n-1}\tilde{a}_1^T\tilde{a}_1$$

$$A = \begin{array}{c} m \text{ feats} \\ \downarrow \end{array} \quad \underbrace{\begin{pmatrix} -\tilde{a}_1- \\ -\tilde{a}_2- \end{pmatrix}}_{\xrightarrow{} \; n \text{ people}} \leftarrow \text{centered} \quad \begin{array}{l} \text{height measurements} \\ \text{weight measurements} \end{array}$$

Then:

$$\frac{1}{n-1}AA^T = \frac{1}{n-1}\begin{pmatrix} -\tilde{a}_1- \\ -\tilde{a}_2- \end{pmatrix}\begin{pmatrix} | & | \\ \tilde{a}_1 & \tilde{a}_2 \\ | & | \end{pmatrix} = \frac{1}{n-1}\begin{pmatrix} \tilde{a}_1^T\tilde{a}_1 & \tilde{a}_1^T\tilde{a}_2 \\ \tilde{a}_2^T\tilde{a}_1 & \tilde{a}_2^T\tilde{a}_2 \end{pmatrix}$$

(diagonals: variances, off-diagonals: co-variance)

Using SVD: $A = U\Sigma V^T$

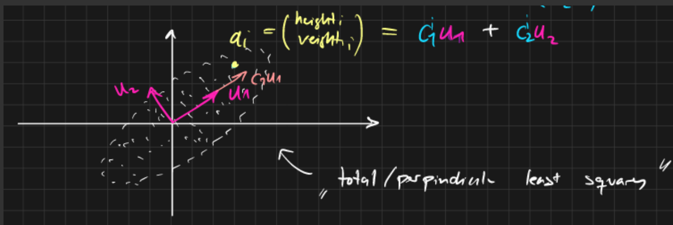
$$\frac{1}{n-1}AA^T = \frac{1}{n-1}U\begin{pmatrix}\sigma_1^2 & & 0\\ & \ddots & \\ 0 & & \sigma_r^2\end{pmatrix}U^T = \frac{1}{n-1}\sum_{i=1}^{r}\sigma_i^2 u_i u_i^T$$

Thus, the first few summands explain most of AA^T, i.e., the variance
The singular vectors $u_1, \dots, u_r$ are called principal components in this setting.
(Remark: $\|A\|_F = \mathrm{tr}(AA^T) = \sum_{i=1}^{m} \tilde{a}_i^T \tilde{a}_i$ = sum of variances)

Now to the geometry of the SVD:

$$A = \begin{array}{c} m \text{ feats} \\ \downarrow \end{array} \begin{array}{c} n \text{ samples} \rightarrow \\ \begin{pmatrix} | & & | & & | \\ a_1 & \cdots & a_i & \cdots & a_n \\ | & & | & & | \end{pmatrix}\end{array} = U\Sigma V^T = \underbrace{\begin{pmatrix} | & & | \\ u_1 & \cdots & u_m \\ | & & | \end{pmatrix}}_{\text{orthonormal basis}} \underbrace{(\Sigma V^T)}_{\text{coordinates of } a_i \text{ in terms of this basis}}$$

Thus, each sample $a_i \in \mathbb{R}^m$ is a linear combination of $u_1, \dots, u_m$ with coefficients $(\Sigma V^T)_i = c_i = \begin{pmatrix} c_1^i \\ c_2^i \end{pmatrix}$



The speciality about the particular orthonormal system $u_1, \ldots, u_m$ $(m = 2)$ is this:

If we only take the first $u_1, \ldots, u_k$ $(k = 1)$ then among all orthonormal systems which are composed of k vectors, these give the best approximation to A (= the measurements) in the $\| \cdot \|_F$-sense.

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

Let us write equation (4) in matrix form:

$$\begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} Av \\ A^\top u \end{pmatrix} = \begin{pmatrix} \sigma u \\ \sigma v \end{pmatrix} = \sigma \begin{pmatrix} u \\ v \end{pmatrix}.$$

Then this reads as an eigenvalue problem for the symmetric matrix $S := \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}$.

Thus we already identify r eigenpairs for S, namely,

$$(\sigma_1, \begin{pmatrix} u_1 \\ v_1 \end{pmatrix}), \ldots, (\sigma_r, \begin{pmatrix} u_r \\ v_r \end{pmatrix}),$$

where $(\sigma_i, \begin{pmatrix} u_i \\ v_i \end{pmatrix})$ are the r singular values and vectors of A, respectively.

Also we easily find that

$$(-\sigma_1, \begin{pmatrix} -u_1 \\ v_2 \end{pmatrix}), \ldots, (-\sigma_r, \begin{pmatrix} -u_r \\ v_r \end{pmatrix})$$

are eigenpairs of S.
For the remaining $(m-r)+(n-r)$ eigenpairs take orthonomal bases $u_{r+1}, \ldots, u_m \in \ker A^\top$ and $v_{r+1}, \ldots, v_n \in \ker A$, then the $(0, \begin{pmatrix} u_i \\ 0 \end{pmatrix})$ and $(0, \begin{pmatrix} 0 \\ v_i \end{pmatrix})$ give the remaining eigenpairs (with eigenvalue 0).

Implications:
→ We can compute the SVD without computing $A^\top A$ or $AA^\top$.
→ Goes back to Gene Golub in the 1960s (→ see his license plate)

Final Remark:

The SVD is a powerful tool and being able to compute it efficiently further facilitates, among others, the following:

- standard method for computing matrix norms $\|A\|_F$ (or $\|A\|_2 := \sigma_1$)
- the best method for determining the rank of a matrix is to count the number of singular values greater than a judiciously chosen tolerance (note: the fundamental problem is distinguishing a small float which is prone to rounding errors from an actual zero!)
- most accurate method for finding an orthonormal basis of a range or a nullspace
- standards for computing low-rank approximations w.r.t to $\|\cdot\|_F$
- ingredient in robust algorithms for least squares fitting via pseudoinverse