Matrix Decompositions 3 of 3

Singular Value Decomposition

The singular value decomposition (SVD) of a matrix is a central matrix decomposition method in linear algebra.

- It has been referred to as the *fundamental theorem of linear algebra*
  - SVD can be applied to all matrices, even non square ones
  - The decomposition is also always possible.

Singular Value Decomposition

SVD Theorem

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a rectangular matrix of rank $r \in [0, \min(m, n)]$. The SVD of $\mathbf{A}$ is a decomposition of the form

$$\mathbf{A} = \underbrace{\mathbf{U}}_{Rotation} \underbrace{\mathbf{\Sigma}}_{Stretch} \underbrace{\mathbf{V^T}}_{Rotation} \tag{1}$$

- $\mathbf{U} \in \mathbb{R}^{m \times m}$ is an orthogonal matrix with column vectors $\mathbf{u}_i$, $i = 1, \ldots, m$, and
- $\mathbf{V} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix with column vectors $\mathbf{v}_j$, $j = 1, \ldots, n$.
- Lastly $\Sigma \in \mathbb{R}^{m \times n}$ matrix with $\Sigma_{ii} = \sigma_i \geq 0$ and $\Sigma_{ij} = 0$, $i \neq j$.

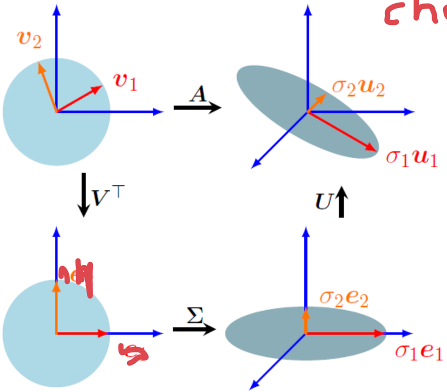
Singular Value Decomposition: Additional Terminology

The are a couple of important additional conventions and terminology with SVD. Specifically,

- The diagonal entries σ_i, $i = 1, \ldots, r$, of Σ are called the **singular values**.
- The vectors $\mathbf{u}_i$ are called the **left-singular vectors**.
- The vectors $\mathbf{v}_j$ are called the **right-singular vectors**.
- By convention, the singular values are ordered:
  $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$

Geometric Intuitions for the SVD

cf. Know our interp of change of basis



For a $\mathbf{A} \in \mathbb{R}^{3\times 2}$

Source: M.P. Deisenroth *et al*, Mathematics for Machine Learning (First Edition)

$R^2 \longrightarrow R^3$

Geometric Intuitions for the SVD

Ignore this

Focus on prev slide interp

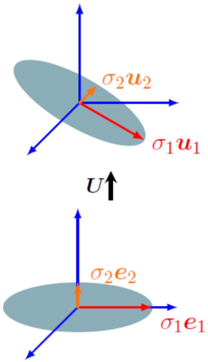
Assume we have transformation matrix of a linear mapping $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ with respect to the standard bases B and C of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Moreover, assume a second basis $\tilde{B}$ of $\mathbb{R}^n$ and $\tilde{C}$ of $\mathbb{R}^m$.

1. The matrix performs a basis change in the domain $\mathbb{R}^n$ from $\tilde{B}$ to standard basis B
   - $\mathbf{V}^T = \mathbf{V}^{-1}$ performs a basis change from B to $\tilde{B}$.
2. Having changed the coordinate system to $\tilde{B}$, $\boldsymbol{\Sigma}$ scales the new coordinates by the singular values σ_i (and adds or deletes dimensions)
   - $\boldsymbol{\Sigma}$ is the transformation matrix of Φ with respect to $\tilde{B}$ and $\tilde{C}$
   - If $m > n$ the scaling happens in a n-dimensional embedding within the m dimensional space. (In the example $m = 3$ and $n = 2$)
   - If $m < n$ the process is more akin to the scaling of a projection as we are mapping from a higher dimensional space into a lower one.

Geometric Intuitions for the SVD

Ignore

3. Lastly $\mathbf{U}$ performs a basis change in the codomain $\mathbb{R}^m$ from $\tilde{C}$ into the canonical basis of $\mathbb{R}^m$.



Source: M.P. Deisenroth *et al*, Mathematics for Machine Learning (First Edition)

Geometric Intuitions for the SVD: Example

$\mathbb{R}^2 \to \mathbb{R}^3$

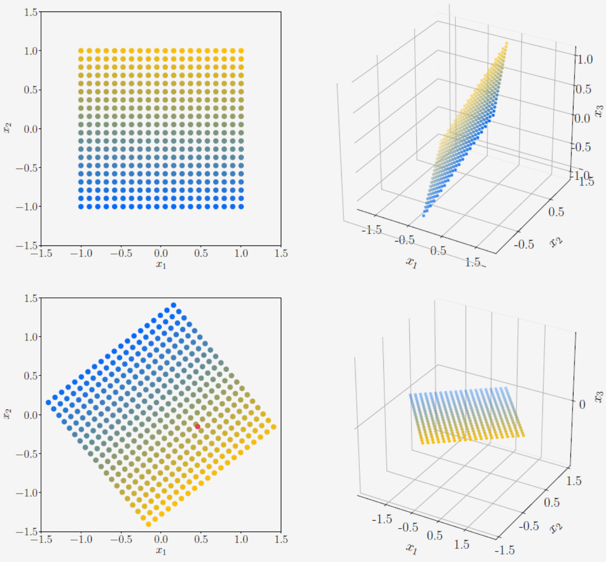
Consider

$$\mathbf{A} = \begin{bmatrix} 1 & -0.8 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \mathbf{U\Sigma V}^T \tag{2}$$

$$\begin{bmatrix} -0.79 & 0 & -0.62 \\ 0.38 & -0.78 & -0.49 \\ -0.48 & -0.62 & 0.62 \end{bmatrix} \begin{bmatrix} 1.62 & 0 \\ 0 & 1.0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.78 & 0.62 \\ -0.62 & -0.78 \end{bmatrix} \tag{3}$$

Consider a large number of vectors in the unit square centered around $\mathbf{0}$ we can easily visualize the transformation.

Geometric Intuitions for the SVD: Example



Source: M.P. Deisenroth *et al*, Mathematics for Machine Learning (First Edition)

SVD and Eigendecomposition equivalence for SPD matrices

There is a direct relationship between the SVD and the eigendecomposition.

- Let $\mathbf{S}$ be a symmetric, positive definite matrix then we have that

$$\mathbf{S} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1} \tag{4}$$
$$= \mathbf{P}\mathbf{D}\mathbf{P}^{T} \tag{5}$$

  where $\mathbf{P}$ is an orthogonal matrix and $\mathbf{D}$ is diagonal.
- If we then set

$$\mathbf{U} = \mathbf{P} = \mathbf{V}, \text{ and } \mathbf{D} = \mathbf{\Sigma} \tag{6}$$

  we see that the SVD of symmetric, positive definite matrices is their eigendecomposition.

Construction of the SVD

High level game plan for $\mathbf{A} \in \mathbb{R}^{m \times n}$:

- Find two sets of orthonormal bases $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ and $V = (\mathbf{v}_1, \ldots \mathbf{v}_n)$ of the codomain $\mathbb{R}^m$ and the domain $\mathbb{R}^n$, respectively.
  - From these ordered bases, we will construct the matrices $\mathbf{U}$ and $\mathbf{V}$.
- We are however looking for two specific orthonormal bases such that

$$\mathbf{AV} = \mathbf{U\Sigma} \tag{7}$$

  where $\mathbf{\Sigma}$ has only no-zero values for $\Sigma_{ii} = \sigma_i$ and that they decreases as i increases.

Construction of the SVD

We can solve for our $\mathbf{v}_i$s (right-singular vectors) by noting that $\mathbf{A}^T\mathbf{A}$ is symmetric, positive semi-definite and therefor diagonalizable. This means that

$$\mathbf{A}^T\mathbf{A} = \mathbf{P}\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}\mathbf{P}^T \tag{8}$$

where $\mathbf{P}$ is an orthogonal matrix, which is composed of the orthonormal eigenbasis. Where $\lambda_i \geq 0$ are the eigenvalues of $\mathbf{A}^T\mathbf{A}$

Construction of the SVD

Also observe, under the assuming that the SVD exists that,

$$\begin{aligned}
\mathbf{A}^T\mathbf{A} &= (\mathbf{U\Sigma V}^T)^T\mathbf{U\Sigma V}^T && (9)\\
&= \mathbf{V\Sigma}^T\mathbf{U}^T\mathbf{U\Sigma V}^T && (10)\\
&= \mathbf{V\Sigma}^T\mathbf{\Sigma V}^T && (11)\\
&= \mathbf{V}\begin{bmatrix}\sigma_1^2 & 0 & 0\\ \vdots & \ddots & \vdots\\ 0 & 0 & \sigma_n^2\end{bmatrix}\mathbf{V}^T && (12)
\end{aligned}$$

Now can see by equating equation (8) and equation (12) that

$$\mathbf{P}\begin{bmatrix}\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \lambda_n\end{bmatrix}\mathbf{P}^T = \mathbf{V}\begin{bmatrix}\sigma_1^2 & 0 & 0\\ \vdots & \ddots & \vdots\\ 0 & 0 & \sigma_n^2\end{bmatrix}\mathbf{V}^T \tag{13}$$

Construction of the SVD

It follows that

$$\mathbf{V}^T = \mathbf{P}^T \tag{14}$$

$$\sigma_i^2 = \lambda_i \tag{15}$$

- Therefore, the eigenvectors of $\mathbf{A}^T\mathbf{A}$ that compose $\mathbf{P}$ are the right-singular vectors $\mathbf{V}$ of $\mathbf{A}$
- The eigenvalues of $\mathbf{A}^T\mathbf{A}$ are the squared singular values of $\mathbf{\Sigma}$

Construction of the SVD

We can now follow a similar approach to the $\mathbf{u}_i$s (left-singular vectors). Namely

$$\mathbf{A}\mathbf{A}^T = \mathbf{Q} \begin{bmatrix} \alpha_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_m \end{bmatrix} \mathbf{Q} \tag{16}$$

where $\mathbf{Q}$ is an orthogonal matrix, which is composed of the orthonormal eigenbasis. Where $\alpha_i \geq 0$ are the eigenvalues of $\mathbf{A}\mathbf{A}^T$

Construction of the SVD

Also observe, under the assuming that the SVD exists that,

$$\begin{aligned}
\mathbf{A}\mathbf{A}^T &= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)^T && (17)\\
&= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T && (18)\\
&= \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\mathbf{U}^T && (19)\\
&= \mathbf{U}\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \sigma_m^2 \end{bmatrix}\mathbf{U}^T && (20)
\end{aligned}$$

Now can see by equating equation (16) and equation (17) that

$$\mathbf{Q}\begin{bmatrix} \alpha_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_m \end{bmatrix}\mathbf{Q}^T = \mathbf{U}\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \sigma_m^2 \end{bmatrix}\mathbf{U}^T \qquad (21)$$

Construction of the SVD

It follows that

$$\mathbf{U} = \mathbf{Q} \tag{22}$$

$$\sigma_i^2 = \alpha_i \tag{23}$$

- Therefore, the eigenvectors of $\mathbf{AA}^T$ that compose $\mathbf{Q}$ are the left-singular vectors $\mathbf{U}$ of $\mathbf{A}$
- The eigenvalues of $\mathbf{AA}^T$ are the squared singular values of $\mathbf{\Sigma}$

Construction of the SVD

Recall that a matrix $\mathbf{A}$ and its transpose $\mathbf{A}^T$ possess the same eigenvalues.

- This means that $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ have the same eigenvalues.
- This means that the nonzero entries of the $\mathbf{\Sigma}$ matrices in the SVD for both cases have to be the same. $(\lambda_i = \alpha_i = \sigma_i^2)$

Now in principle we already have our diagonalization.

- But it requires more calculation that we would like.
- It is possible to actually get $\mathbf{U}$ from $\mathbf{V}$ and $\mathbf{A}$.

Construction of the SVD (Full SVD since s is not square)

From $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ it follows that

$$\mathbf{A}\mathbf{V} = \mathbf{U}\mathbf{\Sigma} \tag{24}$$

this means that

$$\mathbf{A}\mathbf{v}_i = \mathbf{u}_i\sigma_i \tag{25}$$

so each $\mathbf{u}_i$ is

$$\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{A}\mathbf{v}_i \tag{26}$$

What is not clear is if these $\mathbf{u}_i$ will be orthogonal, but we can show this now

Construction of the SVD

(S1) Prove Orth

Note that $\mathbf{v}_i \perp \mathbf{v}_j$ $(i \neq j)$ still holds under the application of $\mathbf{A}$, namely

$$\mathbf{A}\mathbf{v}_i \perp \mathbf{A}\mathbf{v}_j \quad i \neq j \tag{27}$$

This can be shown by

$A^T A x = \lambda x$

$$\begin{aligned}
(\mathbf{A}\mathbf{v}_i)^T(\mathbf{A}\mathbf{v}_j) &= \mathbf{v}_i^T(\mathbf{A}^T\mathbf{A})\mathbf{v}_j && (28)\\
&= \mathbf{v}_i^T(\lambda_j \mathbf{v}_j) \text{ recall that } \mathbf{v}_j \text{ is an eigenvector of } \mathbf{A}^T\mathbf{A} && (29)\\
&= \lambda_j \mathbf{v}_i^T \mathbf{v}_j && (30)\\
&= 0 && (31)
\end{aligned}$$

This means that we can build a r dimensional orthogonal basis from $(\mathbf{A}\mathbf{v}_1, \ldots \mathbf{A}\mathbf{v}_r)$ where r is the rank of $\mathbf{A}$.

Construction of the SVD

52 Prove norm

We can replace

$$(\mathbf{Av}_1, \dots \mathbf{Av}_r) \text{ with } (\mathbf{u}_1, \dots \mathbf{u}_r) \tag{32}$$

to obtain an orthonormal basis, since

$$\|\mathbf{u}_i\| = \|\frac{1}{\sigma_i}\mathbf{Av}_i\| \tag{33}$$

$$= \frac{1}{|\sigma_i|}\|\mathbf{Av}_i\| \tag{34}$$

$$= \sqrt{(Av_i)^T(Av_i)}$$

$$= \sqrt{v_i^T A^T A v_i}$$

$$= \sqrt{v_i^T \lambda_i v_i}$$

$$= \frac{1}{|\sigma_i|}\sqrt{\lambda_i \mathbf{v}_i^T \mathbf{v}_1} \tag{35}$$

since $A^T A v_i = \lambda_i v_i$ by definition of Eig Decomp

$$= \frac{\sqrt{\lambda_i}}{|\sigma_i|}\|\mathbf{v}_i\| \tag{36}$$

$$= 1 \tag{37}$$

since $\|\mathbf{v}_i\|$ is unit length already and $\lambda_i = \sigma_i^2$

Construction of the SVD

We have a couple of cases to consider

- If $n < m$ (we have more rows than columns) then equation (25) only holds for $i \leq n$ and does not say anything about $\mathbf{u}_i$ for $i > n$
  - However, we know by construction that they are orthonormal.
- Conversely, for $m < n$ equation (25) only holds for $i \leq m$
  - For $i > m$, we have $\mathbf{Av}_i = \mathbf{0}$ and we still know that the $\mathbf{v}_i$ form an orthonormal set
    - These would correspond to the orthonormal basis of the kernel of $\mathbf{A}$.
    - $\{\mathbf{x} | \mathbf{Ax} = \mathbf{0}\}$

SVD Example

Work through Example 4.13 for a nice example.

- Try use $\mathbf{A}^T$ as an exercise (it will have more rows than columns)

Eigenvalue Decomposition vs. Singular Value Decomposition

It is well worth comparing and contrasting the eigendecomposition $\mathbf{A} = \mathbf{PDP}^{-1}$ and the SVD $\mathbf{A} = \mathbf{U\Sigma V}^T$

- The SVD always exists for any matrix $\mathbb{R}^{m \times n}$ where as eigendecomposition is only defined for square matrices $\mathbb{R}^{n \times n}$, and only if only exists if we can find a basis of eigenvectors of $\mathbb{R}^n$
- The vectors in the eigendecomposition matrix $\mathbf{P}$ are not necessarily orthogonal
  - So they are not necessarily a simple rotation and scaling.

  The vectors in the matrices $\mathbf{U}$ and $\mathbf{V}$ in the SVD are orthonormal, so they do represent rotations.

Eigenvalue Decomposition vs. Singular Value Decomposition

It is well worth comparing and contrasting the eigendecomposition $\mathbf{A} = \mathbf{PDP}^{-1}$ and the SVD $\mathbf{A} = \mathbf{U\Sigma V}^T$

- Both the eigendecomposition and the SVD are compositions of three linear mappings:
    1. Change of basis in the domain
    2. Independent scaling of each new basis vector and mapping from domain to codomain
    3. Change of basis in the codomain

  A key difference between the eigendecomposition and the SVD is that in the SVD, domain and codomain can be vector spaces of **different dimensions**.

Eigenvalue Decomposition vs. Singular Value Decomposition

It is well worth comparing and contrasting the eigendecomposition $\mathbf{A} = \mathbf{PDP}^{-1}$ and the SVD $\mathbf{A} = \mathbf{U\Sigma V}^T$

- In the SVD, the left- and right-singular vector matrices $\mathbf{U}$ and $\mathbf{V}$ are generally not inverses of each other
  - They perform basis changes in different vector spaces

  In the eigendecomposition, the basis change matrices $\mathbf{P}$ and $\mathbf{P}^{-1}$ are inverses of each other.

Eigenvalue Decomposition vs. Singular Value Decomposition

It is well worth comparing and contrasting the eigendecomposition $\mathbf{A} = \mathbf{PDP}^{-1}$ and the SVD $\mathbf{A} = \mathbf{U\Sigma V}^{T}$

- In the SVD, the entries in the diagonal matrix $\mathbf{\Sigma}$ are all real and nonnegative, which is not generally true for the diagonal matrix in the eigendecomposition.
- For symmetric matrices $A \in \mathbb{R}^{n \times n}$, the eigenvalue decomposition and the SVD are one and the same.

We will now investigate how the SVD allows us to represent a matrix $\mathbf{A}$ as a sum of simpler (low-rank) matrices $\mathbf{A}_i$, which lends itself to a matrix approximation scheme that is cheaper to compute than the full SVD.

- We can construct rank-1 matrices $\mathbf{A}_i \in \mathbb{R}^{m \times n}$ as

$$\mathbf{A}_i := \mathbf{u}_i \mathbf{v}_i^T \tag{38}$$

  where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith orthogonal column vectors of $\mathbf{U}$ and $\mathbf{V}$ respectively.

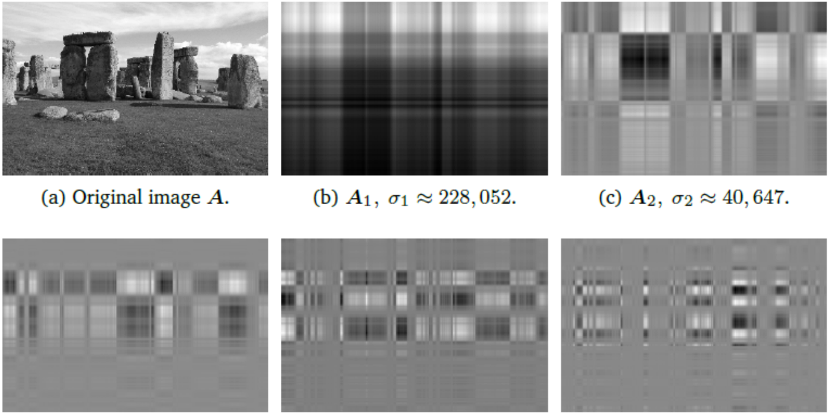
Matrix Approximation

A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ of rank r an be written as a sum of rank-1 matrices $\mathbf{A}_i$ so that

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \sum_{i=1}^{r} \sigma_i \mathbf{A}_i \tag{39}$$

where the outer-product matrices $\mathbf{A}_i$ are weighted by the ith singular value σ_i.

Matrix Approximation



(a) Original image $\boldsymbol{A}$. (b) $\boldsymbol{A}_1$, $\sigma_1 \approx 228,052$. (c) $\boldsymbol{A}_2$, $\sigma_2 \approx 40,647$.

(d) $\boldsymbol{A}_3$, $\sigma_3 \approx 26,125$. (e) $\boldsymbol{A}_4$, $\sigma_4 \approx 20,232$. (f) $\boldsymbol{A}_5$, $\sigma_5 \approx 15,436$.

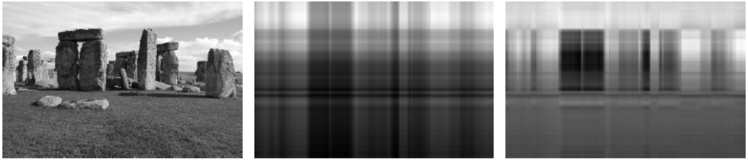
$$\mathbf{A} \in \mathbb{R}^{1432 \times 1910}$$

Source: M.P. Deisenroth *et al*, Mathematics for Machine Learning (First Edition)

Matrix Approximation

If we sum only up to $k < r$ we obtain a **rank-k approximation**

$$\hat{\mathbf{A}}(k) := \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \sum_{i=1}^{k} \sigma_i \mathbf{A}_i \tag{40}$$



(a) Original image $\boldsymbol{A}$. (b) Rank-1 approximation $\widehat{\boldsymbol{A}}(1)$. (c) Rank-2 approximation $\widehat{\boldsymbol{A}}(2)$.

(d) Rank-3 approximation $\widehat{\boldsymbol{A}}(3)$. (e) Rank-4 approximation $\widehat{\boldsymbol{A}}(4)$. (f) Rank-5 approximation $\widehat{\boldsymbol{A}}(5)$.

Source: M.P. Deisenroth *et al*, Mathematics for Machine Learning (First Edition)

Matrix Approximation

Can we quantify how close our approximation, $\hat{\mathbf{A}}(k)$, is from the original $\mathbf{A}$?

- All we need is a distance metric, in our case we will use a full matrix norm.
- There exists a couple of common matrix norms, for our current focus we will use the **spectral norm**

Spectral Norm of a Matrix

Spectral Norm of a Matrix

For $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, the spectral norm of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is defined as

$$\|\mathbf{A}\|_2 := \max_x \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2} \tag{41}$$

The spectral norm is the maximum 'scale', by which the matrix $\mathbf{A}$ can 'stretch' a vector.

Theorem 4.24

The spectral norm of $\mathbf{A}$ is its largest singular value σ_1.

Spectral Norm of a Matrix

Eckart-Young Theorem

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be of rank r and let $\mathbf{B} \in \mathbb{R}^{m \times n}$ be of rank k

- For all $k < r$ with $\hat{\mathbf{A}}(k) = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ it holds that

Optimal $\rightarrow$
$$\hat{\mathbf{A}}(k) = argmin_{rk(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_2 \tag{42}$$
$$\|\mathbf{A} - \hat{\mathbf{A}}(k)\|_2 = \sigma_{k+1} \tag{43}$$

The Eckart-Young theorem implies that we can use SVD to reduce a rank-r matrix $\mathbf{A}$ to a rank-k matrix $\hat{\mathbf{A}}$ in a principled, optimal (in the spectral norm sense) manner.

- We can interpret the approximation of $\mathbf{A}$ by a rank-k matrix as a form of lossy compression

Matrix Phylogeny

Source: M.P. Deisenroth *et al*, Mathematics for Machine Learning (First Edition)