Matrix Decompositons

Mathematics for Machine Learning

San Diego Machine Learning
Liam Barstad

Determinant

Measure of Volume

2D: $\det(A) = a_{11}a_{22} - a_{12}a_{21}$
3D: $\det(A) = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23}$
ND: $\det(A) = \Sigma^{n}_{k=1}(-1)^{k+j}a_{jk}\det(A_{k,j})$

- $\det(AB) = \det(A)\det(B)$
- $\det(A) = \det(A^T)$
- Similar matrices have the same determinant
- $\det(\lambda A) = \lambda^n\det(A)$
- Swapping 2 rows/cols changes the sign

Figure 4.2 The area of the parallelogram (shaded region) spanned by the vectors $\boldsymbol{b}$ and $\boldsymbol{g}$ is $|\det([\boldsymbol{b}, \boldsymbol{g}])|$.

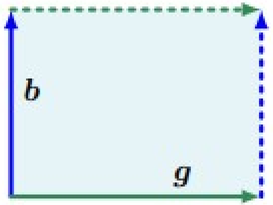
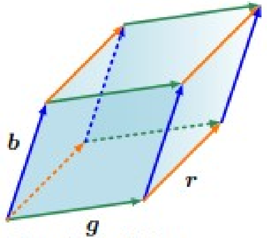


Figure 4.3 The volume of the parallelepiped (shaded volume) spanned by vectors $\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{g}$ is $|\det([\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{g}])|$.



The sign of the determinant indicates the orientation of the spanning vectors.

Trace

Sum of all diagonal elements of A

$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}$

- $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$
- $\mathrm{tr}(\lambda A) = \lambda \mathrm{tr}(A)$
- $\mathrm{tr}(I_n)$
- $\mathrm{tr}(AB) = \mathrm{tr}(BA)$

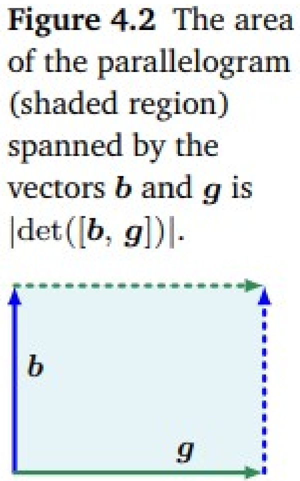


Figure 4.2 The area of the parallelogram (shaded region) spanned by the vectors $\boldsymbol{b}$ and $\boldsymbol{g}$ is $|\det([\boldsymbol{b}, \boldsymbol{g}])|$.

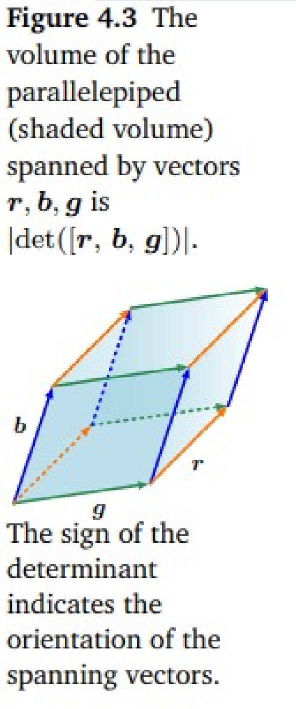
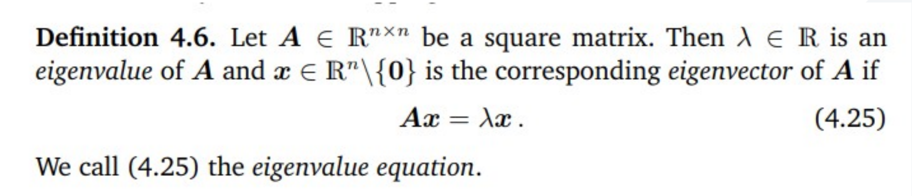


Figure 4.3 The volume of the parallelepiped (shaded volume) spanned by vectors $\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{g}$ is $|\det([\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{g}])|$.

The sign of the determinant indicates the orientation of the spanning vectors.

Eigenvectors + Eigenvalues



Definition 4.6. Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an *eigenvalue* of $\boldsymbol{A}$ and $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ is the corresponding *eigenvector* of $\boldsymbol{A}$ if

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x} \,. \tag{4.25}$$

We call (4.25) the *eigenvalue equation*.

Equivalent to $(A - \lambda I_n) = 0$

- $\mathrm{rk}(A - \lambda I_n) < n$ (**defective**)
- $\det(A - \lambda I_n) = 0$
- Product of eigenvalues = det(A)

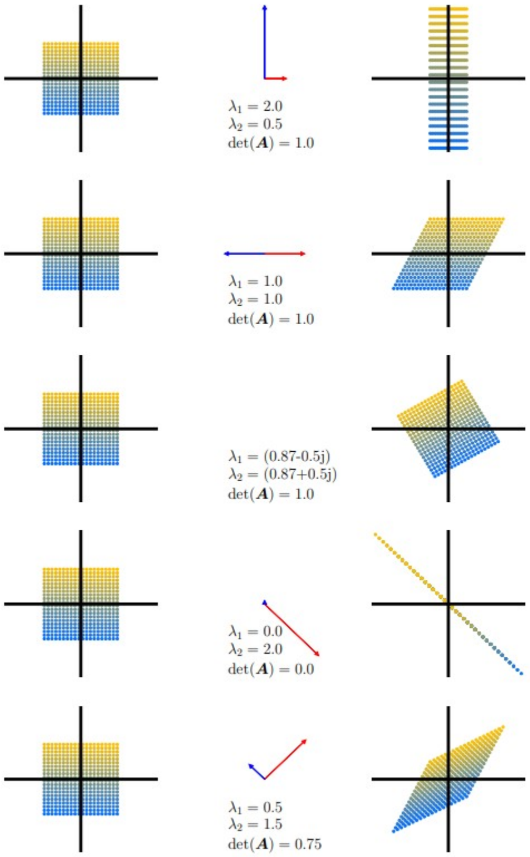
$A = P^{-1}DP$

Roots of characteristic polynomial

Set of all eigenvectors = **eigenspace**

Set of all eigenvalues = **eigenspectrum**

Geometric multiplicity = dims of eigenspectrum



$\lambda_1 = 2.0$
$\lambda_2 = 0.5$
$\det(\boldsymbol{A}) = 1.0$
$\lambda_1 = 1.0$
$\lambda_2 = 1.0$
$\det(\boldsymbol{A}) = 1.0$
$\lambda_1 = (0.87\text{-}0.5j)$
$\lambda_2 = (0.87{+}0.5j)$
$\det(\boldsymbol{A}) = 1.0$
$\lambda_1 = 0.0$
$\lambda_2 = 2.0$
$\det(\boldsymbol{A}) = 0.0$
$\lambda_1 = 0.5$
$\lambda_2 = 1.5$
$\det(\boldsymbol{A}) = 0.75$

Figure 4.4 Determinants and eigenspaces. Overview of five linear mappings and their associated transformation matrices $\boldsymbol{A}_i \in \mathbb{R}^{2 \times 2}$ projecting 400 color-coded points $\boldsymbol{x} \in \mathbb{R}^2$ (left column) onto target points $\boldsymbol{A}_i\boldsymbol{x}$ (right column). The central column depicts the first eigenvector, stretched by its associated eigenvalue λ_1, and the second eigenvector stretched by its eigenvalue λ_2. Each row depicts the effect of one of five transformation matrices $\boldsymbol{A}_i$ with respect to the standard basis .

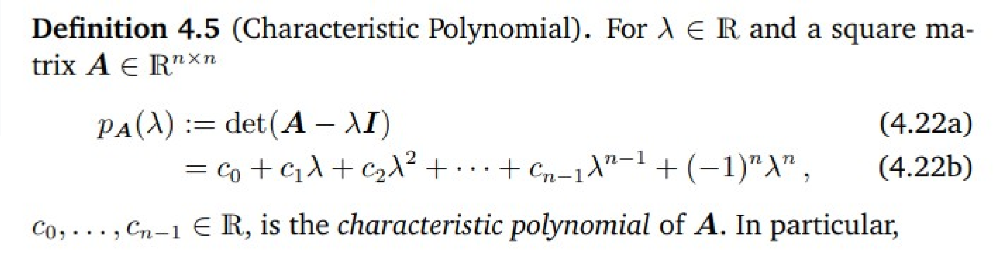
Characteristic Polynomial

Set $\det(A - \lambda I) = 0$ to get the solution space for $Ax = \lambda x$

- $c_0 = \det(A)$
- $c_{n-1} = (-1)^{n-1}\mathrm{tr}(A)$

Algebraic multiplicity – the # of times λ is the root of the characteristic polynomial, how strongly it influences the structure of A

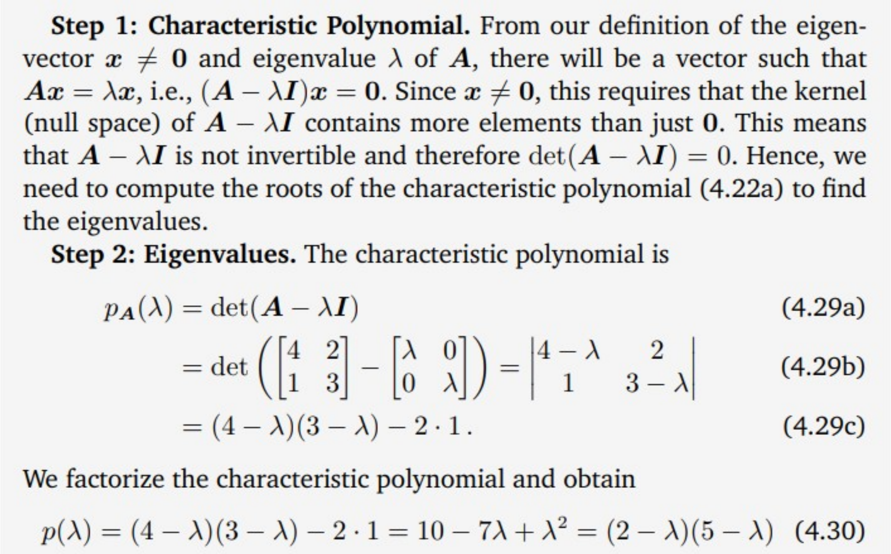
Defective matrix has at least one λ with algebraic multiplicity > 1



Definition 4.5 (Characteristic Polynomial)**.** For $\lambda \in \mathbb{R}$ and a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$

$$\begin{aligned} p_{\boldsymbol{A}}(\lambda) &:= \det(\boldsymbol{A} - \lambda \boldsymbol{I}) && (4.22a)\\ &= c_0 + c_1\lambda + c_2\lambda^2 + \cdots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n\,, && (4.22b) \end{aligned}$$

$c_0, \ldots, c_{n-1} \in \mathbb{R}$, is the *characteristic polynomial* of $\boldsymbol{A}$. In particular,



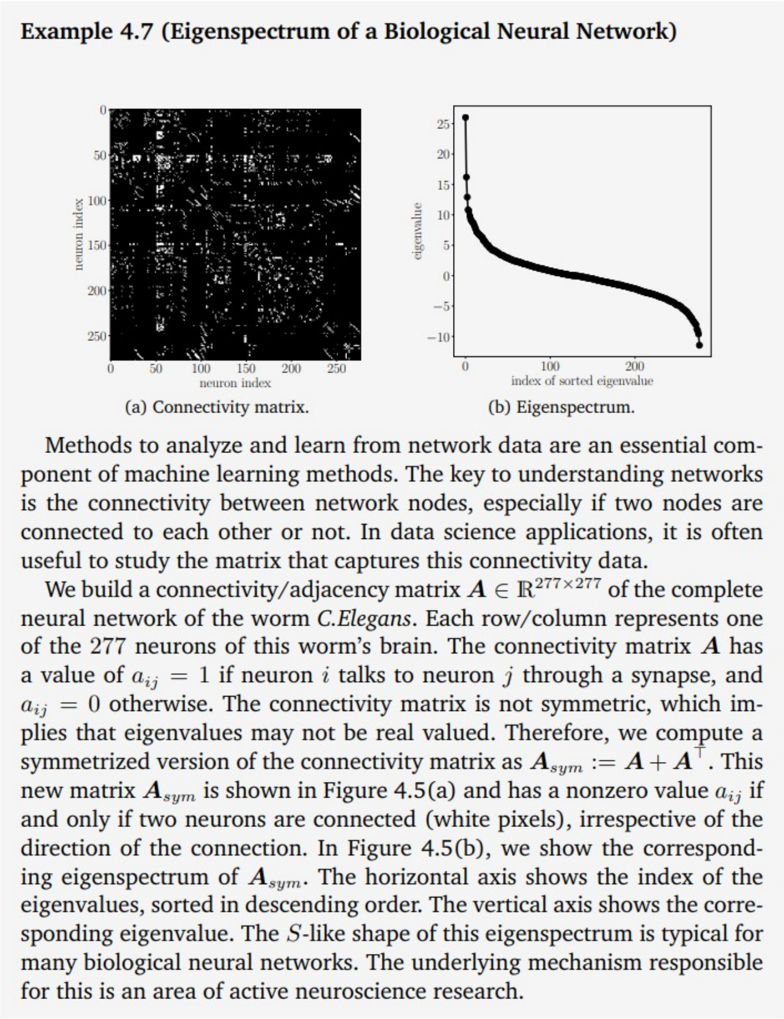
Step 1: Characteristic Polynomial. From our definition of the eigenvector $\boldsymbol{x} \neq \boldsymbol{0}$ and eigenvalue λ of $\boldsymbol{A}$, there will be a vector such that $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, i.e., $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$. Since $\boldsymbol{x} \neq \boldsymbol{0}$, this requires that the kernel (null space) of $\boldsymbol{A} - \lambda\boldsymbol{I}$ contains more elements than just $\boldsymbol{0}$. This means that $\boldsymbol{A} - \lambda\boldsymbol{I}$ is not invertible and therefore $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$. Hence, we need to compute the roots of the characteristic polynomial (4.22a) to find the eigenvalues.

Step 2: Eigenvalues. The characteristic polynomial is

$$\begin{aligned} p_{\boldsymbol{A}}(\lambda) &= \det(\boldsymbol{A} - \lambda\boldsymbol{I}) && (4.29a)\\ &= \det\left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = \begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} && (4.29b)\\ &= (4-\lambda)(3-\lambda) - 2\cdot 1\,. && (4.29c) \end{aligned}$$

We factorize the characteristic polynomial and obtain

$$p(\lambda) = (4-\lambda)(3-\lambda) - 2\cdot 1 = 10 - 7\lambda + \lambda^2 = (2-\lambda)(5-\lambda) \quad (4.30)$$



Example 4.7 (Eigenspectrum of a Biological Neural Network)

0
50
100
150
200
250
neuron index
0
50
100
150
200
250
neuron index
25
20
15
10
5
0
−5
−10
eigenvalue
0
100
200
index of sorted eigenvalue

(a) Connectivity matrix.

(b) Eigenspectrum.

Methods to analyze and learn from network data are an essential component of machine learning methods. The key to understanding networks is the connectivity between network nodes, especially if two nodes are connected to each other or not. In data science applications, it is often useful to study the matrix that captures this connectivity data.

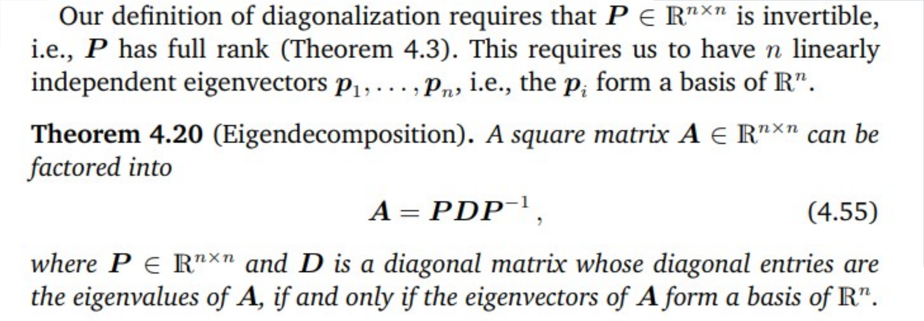
We build a connectivity/adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{277\times277}$ of the complete neural network of the worm *C.Elegans*. Each row/column represents one of the 277 neurons of this worm's brain. The connectivity matrix $\boldsymbol{A}$ has a value of $a_{ij} = 1$ if neuron i talks to neuron j through a synapse, and $a_{ij} = 0$ otherwise. The connectivity matrix is not symmetric, which implies that eigenvalues may not be real valued. Therefore, we compute a symmetrized version of the connectivity matrix as $\boldsymbol{A}_{sym} := \boldsymbol{A} + \boldsymbol{A}^\top$. This new matrix $\boldsymbol{A}_{sym}$ is shown in Figure 4.5(a) and has a nonzero value a_{ij} if and only if two neurons are connected (white pixels), irrespective of the direction of the connection. In Figure 4.5(b), we show the corresponding eigenspectrum of $\boldsymbol{A}_{sym}$. The horizontal axis shows the index of the eigenvalues, sorted in descending order. The vertical axis shows the corresponding eigenvalue. The S-like shape of this eigenspectrum is typical for many biological neural networks. The underlying mechanism responsible for this is an area of active neuroscience research.

Diagonalization

- Geometric multiplicity = algebraic multiplicity
- The eigenvectors form a basis in $\boldsymbol{R}^n$

If a matrix is diagonalizable, computations are much faster

Key concept in PCA

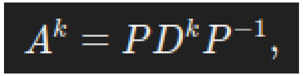


Our definition of diagonalization requires that $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ is invertible, i.e., $\boldsymbol{P}$ has full rank (Theorem 4.3). This requires us to have n linearly independent eigenvectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$, i.e., the $\boldsymbol{p}_i$ form a basis of $\mathbb{R}^n$.

Theorem 4.20 (Eigendecomposition)**.** *A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factored into*

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}\,, \tag{4.55}$$

where $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{D}$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\boldsymbol{A}$, if and only if the eigenvectors of $\boldsymbol{A}$ form a basis of $\mathbb{R}^n$.



$$A^k = PD^kP^{-1},$$

Singular Value Decomposition (SVD)

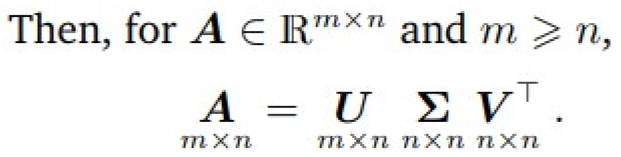
U = orthogonal basis of codomain
Σ = singular values, scales domain + codomain
V = orthogonal basis of domain

Fundamental theorem of linear algebra

For SPD matrices:

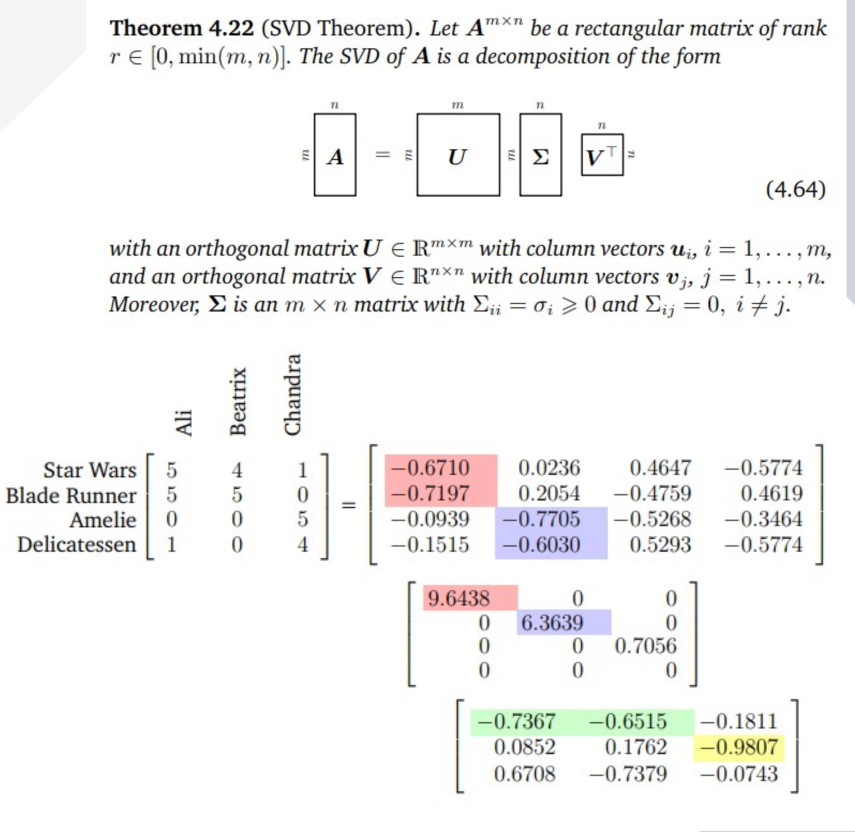
$PDP^{-1} = U \Sigma V$

Reduced SVD:



Then, for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $m \geqslant n$,

$$\underset{m\times n}{\boldsymbol{A}} = \underset{m\times n}{\boldsymbol{U}}\;\underset{n\times n}{\boldsymbol{\Sigma}}\;\underset{n\times n}{\boldsymbol{V}^\top}.$$



Theorem 4.22 (SVD Theorem). *Let $\boldsymbol{A}^{m\times n}$ be a rectangular matrix of rank $r \in [0, \min(m,n)]$. The SVD of $\boldsymbol{A}$ is a decomposition of the form*

$$\overset{n}{m\boxed{\boldsymbol{A}}} = \overset{m}{m\boxed{\boldsymbol{U}}}\;\overset{n}{m\boxed{\boldsymbol{\Sigma}}}\;\overset{n}{\boxed{\boldsymbol{V}^\top}n} \tag{4.64}$$

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

$$\begin{array}{r|ccc|} & \text{Ali} & \text{Beatrix} & \text{Chandra} \\ \text{Star Wars} & 5 & 4 & 1 \\ \text{Blade Runner} & 5 & 5 & 0 \\ \text{Amelie} & 0 & 0 & 5 \\ \text{Delicatessen} & 1 & 0 & 4 \end{array} = \begin{bmatrix} -0.6710 & 0.0236 & 0.4647 & -0.5774 \\ -0.7197 & 0.2054 & -0.4759 & 0.4619 \\ -0.0939 & -0.7705 & -0.5268 & -0.3464 \\ -0.1515 & -0.6030 & 0.5293 & -0.5774 \end{bmatrix}$$

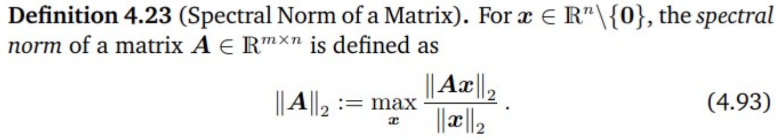
$$\begin{bmatrix} 9.6438 & 0 & 0 \\ 0 & 6.3639 & 0 \\ 0 & 0 & 0.7056 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -0.7367 & -0.6515 & -0.1811 \\ 0.0852 & 0.1762 & -0.9807 \\ 0.6708 & -0.7379 & -0.0743 \end{bmatrix}$$

Rank K Approximation

Takes outer product of each element of domain and codomain, multiplies by scaling factor, and sum together

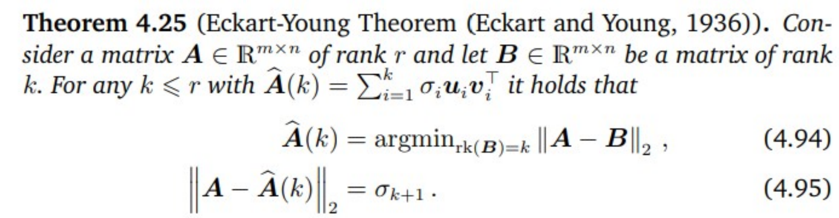
Spectral norm – maximum length any vector x can have when multiplied by A



Definition 4.23 (Spectral Norm of a Matrix). For $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$, the *spectral norm* of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is defined as

$$\|\boldsymbol{A}\|_2 := \max_{\boldsymbol{x}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} . \tag{4.93}$$

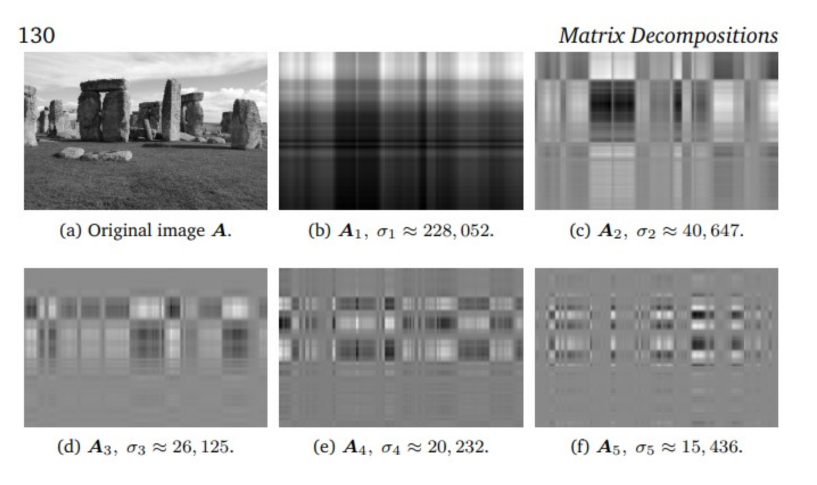
Eckart-Young Theorem – error of rank k approx.



Theorem 4.25 (Eckart-Young Theorem (Eckart and Young, 1936)). *Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ of rank r and let $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ be a matrix of rank k. For any $k \leqslant r$ with $\widehat{\boldsymbol{A}}(k) = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ it holds that*

$$\widehat{\boldsymbol{A}}(k) = \operatorname{argmin}_{\operatorname{rk}(\boldsymbol{B})=k} \|\boldsymbol{A} - \boldsymbol{B}\|_2 , \tag{4.94}$$

$$\left\|\boldsymbol{A} - \widehat{\boldsymbol{A}}(k)\right\|_2 = \sigma_{k+1} . \tag{4.95}$$



130 *Matrix Decompositions*

(a) Original image $\boldsymbol{A}$.

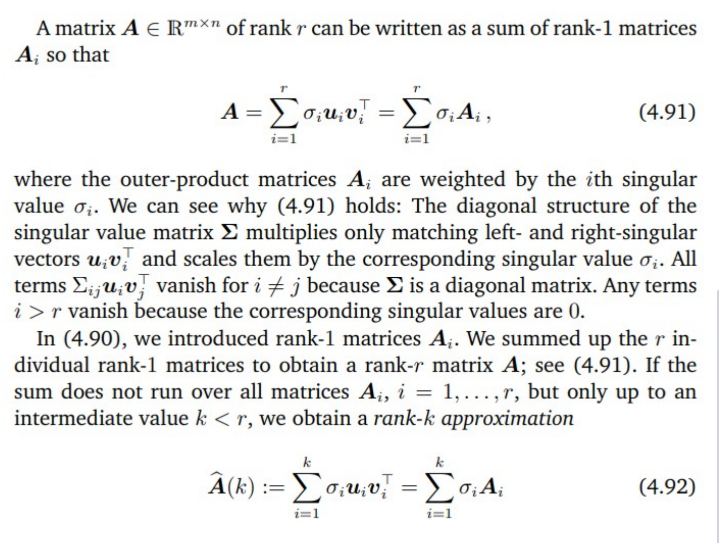
(b) $\boldsymbol{A}_1,\ \sigma_1 \approx 228,052$.

(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) $\boldsymbol{A}_5,\ \sigma_5 \approx 15,436$.



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

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top = \sum_{i=1}^{r} \sigma_i \boldsymbol{A}_i , \tag{4.91}$$

where the outer-product matrices $\boldsymbol{A}_i$ are weighted by the ith singular value σ_i. We can see why (4.91) holds: The diagonal structure of the singular value matrix $\boldsymbol{\Sigma}$ multiplies only matching left- and right-singular vectors $\boldsymbol{u}_i \boldsymbol{v}_i^\top$ and scales them by the corresponding singular value σ_i. All terms $\Sigma_{ij} \boldsymbol{u}_i \boldsymbol{v}_j^\top$ vanish for $i \neq j$ because $\boldsymbol{\Sigma}$ is a diagonal matrix. Any terms $i > r$ vanish because the corresponding singular values are 0.

In (4.90), we introduced rank-1 matrices $\boldsymbol{A}_i$. We summed up the r individual rank-1 matrices to obtain a rank-r matrix $\boldsymbol{A}$; see (4.91). If the sum does not run over all matrices $\boldsymbol{A}_i$, $i = 1, \ldots, r$, but only up to an intermediate value $k < r$, we obtain a *rank-k approximation*

$$\widehat{\boldsymbol{A}}(k) := \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top = \sum_{i=1}^{k} \sigma_i \boldsymbol{A}_i \tag{4.92}$$