Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Thang Vu

Math for Machine Learning

Resources

This lecture builds on content from:

- Mathematics for Machine Learning [1]

Additional free resources if you don't understand a specific topic, or want to learn more:

- Linear Algebra MIT course on Youtube
  `https://www.youtube.com/playlist?list=PL49CF3715CB9EF31D`

Outline

1

Contents

Recap

2

Math Notation

- Natural numbers (positive whole numbers): $\mathbb{N} = \{1, 2, 3, \dots\}$
- Real numbers: $\mathbb{R} = \{-1.3, 0.0, 0.2, 1, \pi, \dots\}$
- Sum from 1 to N: $\sum_{i=1}^{N} i = 1 + 2 + \cdots + N$
- Product from 1 to N: $\prod_{i=1}^{N} i = 1 \cdot 2 \cdot \cdots \cdot N$
- Absolute value (turns negative values into positive values): $|-10| = 10$
- Set inclusion (what set does an element (not) belong to): $5.5 \notin \mathbb{N}$, but $\in \mathbb{R}$
- Cartesian product (all possible ordered combinations - tuples - of the elements of two sets):
  $A = \{1, 2\}, B = \{a, b, c\} : A \times B = \{(1, a), (2, a), (1, b), (2, b), (1, c), (2, c)\}$
  Because of the order: $(1, a) \in A \times B$, but $(a, 1) \notin A \times B$

Important Objects

- A scalar is a single number: $\lambda \in \mathbb{R}$, e.g. $\lambda = 3.14$
- A vector is a a tuple of numbers: $\mathbf{v} \in \mathbb{R}^n$, e.g. $n = 3, \mathbf{v} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$

  We can refer to individual components of a vector by indices, e.g. $v_1 = 0$
- A matrix is a "table" of numbers with m rows and n columns: $\mathbf{A} \in \mathbb{R}^{m \times n}$, e.g. $m = 2, n = 3, A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$

  We can refer to individual components of a matrix using two indices (row and column), e.g. $A_{2,3} = 6$

Important Vector Operations

- Scalar multiplication: $3 \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \cdot 1 \\ 3 \cdot 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$
- Addition: $\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1+3 \\ 2+4 \end{bmatrix} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$
- Element-wise product (Hadamard product): $\begin{bmatrix} 1 \\ 2 \end{bmatrix} \odot \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \cdot 3 \\ 2 \cdot 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 8 \end{bmatrix}$
- Scalar product: $\langle \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \rangle = \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 1 \cdot 0 + 2 \cdot 1 = 2$

Important Matrix Operations

- Scalar multiplication: $3 \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 3 \cdot 1 & 3 \cdot 2 \\ 3 \cdot 3 & 3 \cdot 4 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 9 & 12 \end{bmatrix}$
- Addition: $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}$
- Element-wise product (Hadamard product):
  $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \odot \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1 \cdot 5 & 2 \cdot 6 \\ 3 \cdot 7 & 4 \cdot 8 \end{bmatrix} = \begin{bmatrix} 5 & 12 \\ 21 & 32 \end{bmatrix}$
- Matrix product: $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1 \cdot 5 + 2 \cdot 7 & 1 \cdot 6 + 2 \cdot 8 \\ 3 \cdot 5 + 4 \cdot 7 & 3 \cdot 6 + 4 \cdot 8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$

Recap

Matrices continued

3

Identity Matrix

Definition (Identity Matrix)

In $\mathbb{R}^{n\times n}$ it is defined as:

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 \end{bmatrix}$$

A $n \times n$ matrix that contains ones on the diagonal and zero everywhere else.

Important property: $A \cdot I_n = A = I_m \cdot A, A \in \mathbb{R}^{m,n}$

Inverse Matrix

Definition (Inverse Matrix)

Matrix $\mathbf{A}$ is called invertible if there exists a matrix $\mathbf{A}^{-1}$, such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n = \mathbf{A}^{-1}\mathbf{A}$

Example

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 4 & 4 & 5 \\ 6 & 7 & 7 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -7 & -7 & 6 \\ 2 & 1 & -1 \\ 4 & 5 & -4 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 1 \\ 4 & 4 & 5 \\ 6 & 7 & 7 \end{bmatrix} \begin{bmatrix} -7 & -7 & 6 \\ 2 & 1 & -1 \\ 4 & 5 & -4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -7 & -7 & 6 \\ 2 & 1 & -1 \\ 4 & 5 & -4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 4 & 4 & 5 \\ 6 & 7 & 7 \end{bmatrix}$$

$$\mathbf{AB} = \mathbf{I} = \mathbf{BA}$$

Transpose

Definition (Transpose)

For $\mathbf{A} \in \mathbb{R}^{m \times n}$ the matrix $\mathbf{B} \in \mathbb{R}^{n \times m}$ with $b_{ij} = a_{ji}$ is called the transpose of $\mathbf{A}$. We write $\mathbf{B} = \mathbf{A}^T$.

$\mathbf{A}^T$ can be obtained by switching rows with columns of $\mathbf{A}$.

Example

Let $\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 4 & 4 & 5 \\ 6 & 7 & 7 \end{bmatrix}$, then $\mathbf{A}^T = \begin{bmatrix} 1 & 4 & 6 \\ 2 & 4 & 7 \\ 1 & 5 & 7 \end{bmatrix}$

Matrices in NLP

Examples

- Trainable parameters of neural networks are matrices
- Word embeddings, where each word is represented as a vector, the whole sentence as a matrix in sentence classification tasks
- Adjacency matrix, which contains the edge information of a graph, e.g. dependency graphs in NLP

Live Voting

Linear Independence

4

Linear Combination

Definition (Linear Combination)

Consider a vector space $\mathcal{V}$ and a finite number of vectors $x_1, x_2, \ldots, x_k \in \mathcal{V}$. Then every $v \in \mathcal{V}$ of the form

$$v = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_k x_k = \sum_{i=1}^{k} \lambda_i x_i \in \mathcal{V}$$

with $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$is a linear combination of the vectors $x_1, \ldots, x_k$.

Linear Independence

The zero vector can also be written as a linear combination, because $0 = \sum_{i=1}^{k} 0 \cdot x_i$ is always true.

Definition (Linear Dependence)

If there exists a non-trivial solution to generate $0 = \sum_{i=1}^{k} \lambda_i x_i$ with at least one $\lambda_i \neq 0$ then the vectors $x_1, \ldots, x_k$ are linearly dependent.

Definition (Linear Independence)

If only the trivial solution exists, i.e. $\lambda_1 = \ldots = \lambda_i = 0$, then the vectors $x_1, \ldots, x_k$ are linearly independent.

Intuition of Linear Independence

Vectors being linearly independent means that no vector in the set can be written as a linear combination of the others.

Intuitively, a set of linearly independent vectors consists of vectors that have no redundancy, i.e. if we remove any of those vectors from the set, we will loose something.

In $\mathbb{R}^2$:

- Two vectors are linearly dependent if they lie on the same line or are parallel to each other.
- Two vectors are linearly independent if they are not parallel and do not lie on the same line.

Rank of a Matrix

Definition (Rank)

The number of linearly independent columns of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ equals the number of linearly independent rows and is called the rank of $\mathbf{A}$ and is denoted by rk($\mathbf{A}$).

Example

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$\mathbf{A}$ has a rk($\mathbf{A}$)$= 2$ since it has two linearly independent rows/columns.

Basis

Definition (Basis)

The basis vectors of a vector space $\mathcal{V}$ is the minimal set of of vectors $\mathcal{A}$ that can generate any vector in $\mathcal{V}$ as a linear combination.

- Canonical/standard base in $\mathbb{R}^3$

$$\mathcal{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Different bases in $\mathbb{R}^3$

$$\mathcal{B}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \mathcal{B}_2 = \begin{bmatrix} 0.5 & 1.8 & -2.2 \\ 0.8 & 0.3 & -1.3 \\ 0.4 & 0.3 & 3.5 \end{bmatrix}$$

Live Voting

Figure: Link to Ilias live voting: `https://ilias3.uni-stuttgart.de/vote/PPIA`

Linear Mappings

5

Linear Mappings (Functions)

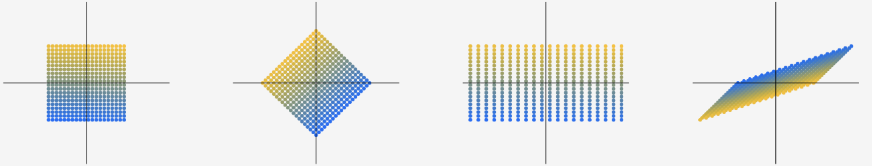
Definition (Linear Mappings)

Given two vector spaces $\mathbf{V}, \mathbf{W}$, a mapping $\Phi : \mathbf{V} \rightarrow \mathbf{W}$ is a linear mapping if, and only if:
$\forall x, y \in \mathbf{V}, \forall \lambda, \psi \in \mathbb{R} : \Phi(\lambda x + \psi y) = \lambda \Phi(x) + \psi \Phi(y)$

Examples

- $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$ is a linear regression model that maps an input signal with n dimensions to an output scalar
- A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ applied to $\mathbf{x} \in \mathbb{R}^n$: $\mathbf{A}\mathbf{x} = \mathbf{y} \in \mathbb{R}^m$ is a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$

Examples



(a) Original data. (b) Rotation by 45°. (c) Stretch along the horizontal axis. (d) General linear mapping.

We consider three linear transformations of a set of vectors in $\mathbb{R}^2$ with the transformation matrices

$$\boldsymbol{A}_1 = \begin{bmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{bmatrix}, \ \boldsymbol{A}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \ \boldsymbol{A}_3 = \frac{1}{2}\begin{bmatrix} 3 & -1 \\ 1 & -1 \end{bmatrix}. \quad (2.97)$$

Figure: Examples of linear transformations as matrices. Figure taken from [1].

Live Voting

Figure: Link to Ilias live voting: `https://ilias3.uni-stuttgart.de/vote/PPIA`

Matrix Decomposition

6

Determinants of Matrices

Definition (Determinant of a Square Matrix)

The determinant of a square matrix $A \in \mathbb{R}^{n \times n}$ is a function that maps A onto a real number, denoted as $\det(A)$, that represents how A scales areas / volumes.

Determinants of Small Matrices:

- For $n = 1$:

$$\det(A) = a_{11}$$

- For $n = 2$:

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}$$

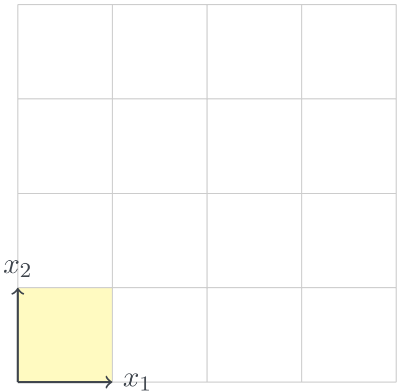
- For $n = 3$ (Sarrus' rule):

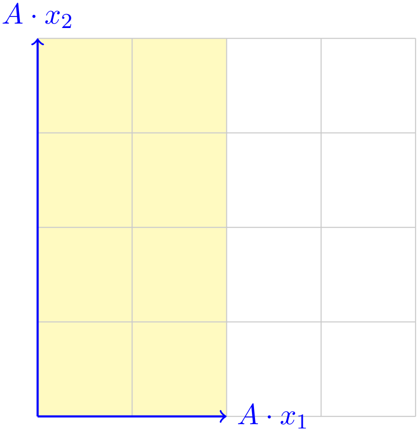
$$\det(A) = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33}$$

- For $n > 3$: Laplace Expansion

Determinant Examples

$$A = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}, det(A) = 4 \cdot 2 - 0 \cdot 0 = 8$$





Live Voting

Figure: Link to Ilias live voting: `https://ilias3.uni-stuttgart.de/vote/PPIA`

Diagonal Matrices

Definition (Diagonal Matrix)

A diagonal matrix is a square matrix that has zeros on all off-diagonal elements. It is of the form:

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}$$

where d_i are the diagonal elements.

Properties of Diagonal Matrices:

- **Determinant:** The determinant of D is the product of its diagonal entries:

$$\det(D) = \prod_{i=1}^{n} d_i$$

Properties of Diagonal Matrices:

- **Matrix Powers:** The k-th power of D is obtained by raising each diagonal element to the k-th power:

$$D^k = \begin{pmatrix} d_1^k & 0 & \cdots & 0 \\ 0 & d_2^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n^k \end{pmatrix}$$

- **Inverse:** If all $d_i \neq 0$, the inverse of D is given by the reciprocals of its diagonal elements:

$$D^{-1} = \begin{pmatrix} d_1^{-1} & 0 & \cdots & 0 \\ 0 & d_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n^{-1} \end{pmatrix}$$

Live Voting

Figure: Link to Ilias live voting: `https://ilias3.uni-stuttgart.de/vote/PPIA`

Eigenvalues and Eigenvectors

Definition (Eigenvalues and Eigenvectors)

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

$$Ax = \lambda x$$

Key Points:

- x is a non-zero vector ($x \neq 0$).
- The equation $Ax = \lambda x$ means that applying A to x scales x by λ.
- Eigenvalues can be found by solving the characteristic equation:

$$\det(A - \lambda I) = 0$$

Eigendecomposition

Definition (Eigendecomposition)

A square matrix $A \in \mathbb{R}^{n \times n}$ can be factored into

$$A = PDP^{-1}$$

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

Key Points:

- The columns of P are the eigenvectors of A.
- D is a diagonal matrix with the eigenvalues of A on the diagonal.
- Eigendecomposition is possible when A is diagonalizable.

Eigendecomposition: Intuition

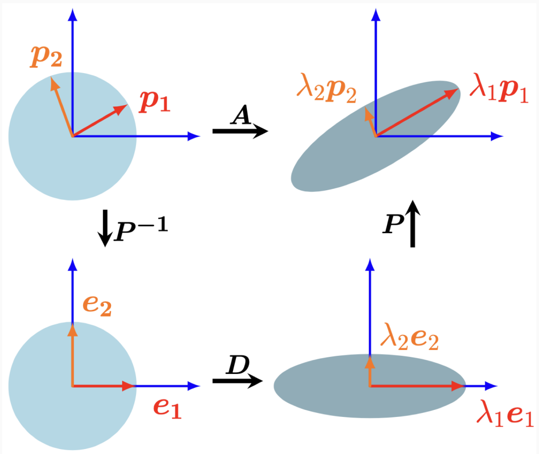


Figure: Intuition behind the eigendecomposition sequential transformations, taken from Princeton COS 302 (2020)
https://www.cs.princeton.edu/courses/archive/spring20/cos302/files/COS_302_Precept_4.pdf

Eigendecomposition: An Example

Given $A = \begin{bmatrix} \frac{5}{2} & -1 \\ -1 & \frac{5}{2} \end{bmatrix}$

Step 1: Compute eigenvalues and eigenvectors

$$\begin{aligned} \det(A - \lambda I) &= \det \begin{bmatrix} \frac{5}{2} - \lambda & -1 \\ -1 & \frac{5}{2} - \lambda \end{bmatrix} \\ &= (\frac{5}{2} - \lambda)^2 - 1 \\ &= \lambda^2 - 5\lambda + \frac{21}{4} \\ &= (\lambda - \frac{7}{2})(\lambda - \frac{3}{2}). \end{aligned}$$

Thus, the eigenvalues of A are $\lambda_1 = \frac{7}{2}$ and $\lambda_2 = \frac{3}{2}$.

Eigendecomposition: An Example

Given $A = \begin{bmatrix} \frac{5}{2} & -1 \\ -1 & \frac{5}{2} \end{bmatrix}$

Step 1: Compute eigenvalues and eigenvectors

$$Ap_1 = \frac{7}{2}p_1$$

$$Ap_2 = \frac{3}{2}p_2$$

This yields

$$p_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$p_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Eigendecomposition: An Example

Given $A = \begin{bmatrix} \frac{5}{2} & -1 \\ -1 & \frac{5}{2} \end{bmatrix}$

Step 2: Check whether p_1 and p_2 form a basis of R^2.

Answer: Yes!

Eigendecomposition: An Example

Given $A = \begin{bmatrix} \frac{5}{2} & -1 \\ -1 & \frac{5}{2} \end{bmatrix}$

Step 3: Construct P and D

$$P = [p_1, p_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$D = \begin{bmatrix} \frac{7}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} \frac{5}{2} & -1 \\ -1 & \frac{5}{2} \end{bmatrix}}_{A} = \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} \frac{7}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}}_{D} \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}}_{P^{-1}}$$

Live Voting

Figure: Link to Ilias live voting: `https://ilias3.uni-stuttgart.de/vote/PPIA`

Orthogonal Matrix

Definition (Orthogonal Matrix)

A square matrix $A \in \mathbb{R}^{n \times n}$ is an orthogonal matrix if and only if its columns are orthonormal so that

$$AA^T = I = A^T A,$$

which implies that

$$A^{-1} = A^T,$$

i.e., the inverse is obtained by simply transposing the matrix.

Transformation with orthogonal matrix is special because it preserves the length and the angle between vectors

Singular Value Decomposition (SVD)

Theorem (SVD Theorem)

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

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

where:

- $U \in \mathbb{R}^{m \times m}$ is an orthogonal matrix.
- $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with non-negative real numbers on the diagonal.
- $V \in \mathbb{R}^{n \times n}$ is an orthogonal matrix.

Structure of the Singular Value Matrix Σ – I

Singular Value Matrix Σ:

- Σ is an $m \times n$ diagonal matrix.
- The diagonal entries $\Sigma_{ii} = \sigma_i$, for $i = 1, \ldots, r$.
- All off-diagonal entries are zero: $\Sigma_{ij} = 0$, for $i \neq j$.

Case 1: If $m > n$, then Σ has the form:

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n \\ \hline & \mathbf{0}_{(m-n)\times n} & & \end{pmatrix}$$

Structure of Σ – II

Case 2: If $m < n$, then Σ has the form:

$$\Sigma = \left(\begin{array}{cccc|c} \sigma_1 & 0 & \dots & 0 & \mathbf{0} \\ 0 & \sigma_2 & \dots & 0 & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma_m & \mathbf{0} \end{array}\right)$$

where the zeros after the vertical line extend from column $m+1$ to n.

Observation:

- Σ is of the same size as A ($m \times n$).
- The singular value matrix Σ is unique for a given A.

Existence of the SVD

Remark

The Singular Value Decomposition exists for any matrix $A \in \mathbb{R}^{m \times n}$, regardless of its rank.

Key Points:

- The SVD provides a way to factor any matrix into simpler, interpretable components.
- It is widely used in applications like signal processing, statistics, and machine learning.

How to construct SVD

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

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

Because $U^\top U = I$

$$A^\top A = V\Sigma^\top \Sigma V^\top = V \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n^2 \end{bmatrix} V^\top \implies \text{Eigendecomposition!}$$

Thus, the eigenvectors of $A^\top A$ will form V and the eigenvalues of $A^\top A$ are the squared singular values of Σ.
$Av_i = \sigma_i u_i \implies u_i = \frac{1}{\sigma_i} A v_i$

Live Voting

Figure: Link to Ilias live voting: `https://ilias3.uni-stuttgart.de/vote/PPIA`

References

[1] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, Mathematics for Machine Learning. Cambridge University Press, 2020.