LINEAR ALGEBRA

- When formalizing intuitive concepts, a common approach is to construct a set of objects (symbols) and a set of rules to manipulate these objects. This is know as an algebra. Linear algebra is the study of vectors and certain rules to manipulate vectors. The vectors we know from school are called "geometric vectors", which are usually denoted by a small arrow above the letter, e.g., $\vec{x}$ and $\vec{y}$. In this lectures, we discuss more general concepts of vectors and use a bold letter (or a underline) to represent them, e.g., $\underline{x}$ and $\underline{y}$.

1. MATRICES

↳ Matrices play a central role in linear algebra. They can be used to compactly represent systems of linear equations, but they also represent linear functions (linear mappings).

▸ DEFINITION[1] (Matrix): With $m, n \in \mathbb{N}$ a real-valued (m,n) matrix $\underline{A}$ is an $m \cdot n$-tuple of elements a_{ij}, $i = 1, \dots, m$, $j = 1, \dots, n$, which is general ordered according to a rectangular scheme consisting of m rows and n columns:

$$\underline{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}_{m \times n}, \quad a_{ij} \in \mathbb{R}. \tag{1.1}$$

↳ By convention $(1,n)$-matrices are called rows and $(m,1)$-matrices are called columns. These special matrices are called row/column vectors. Here, $\mathbb{R}^{m \times n}$ is the set of all real-valued (m,n)-matrices. Also, $\underline{A} \in \mathbb{R}^{m \times n}$ can be equivalently represented as $\underline{a} \in \mathbb{R}^{mn}$ by stacking all n columns of the matrix into a long vector.

1.1) MATRIX ADDITION AND MULTIPLICATION

↳ The sum of two matrices $\underline{A} \in \mathbb{R}^{m \times n}$, $\underline{B} \in \mathbb{R}^{m \times n}$ is defined as the element-wise sum, i.e.,

$$\underline{A} + \underline{B} := \begin{pmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n}. \tag{1.2}$$

For matrices $\underline{A} \in \mathbb{R}^{m \times n}$, $\underline{B} \in \mathbb{R}^{n \times k}$, the elements c_{ij} of the product $\underline{C} = \underline{A}\,\underline{B} \in \mathbb{R}^{m \times k}$ are computed as:

$$c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj} \,, \quad \begin{cases} i = 1, \dots, m \\ j = 1, \dots, k \end{cases} \quad . \tag{1.3}$$

↳ This means, to compute element c_{ij} we multiply the elements of the i^{th} row of $\underline{A}$ with the j^{th} column of $\underline{B}$ and sum them up. Later, we will call this the dot product of the corresponding row and column. In some cases, where we need to be explicit that we are performing multiplication, we use the notation $\underline{A} \cdot \underline{B}$ to denote multiplication (explicitly showing the symbol "$\cdot$").

*) REMARK[1]: Matrices can only be multiplied if their "neighboring" dimensions match. For instance, an $n \times k$-matrix $\underline{A}$ can be multiplied with a $k \times m$-matrix $\underline{B}$, but only from the left side: $\underline{A}_{n \times k} \underline{B}_{k \times m} = \underline{C}_{n \times m}$. The product $\underline{B}\,\underline{A}$ is not defined if $m \neq n$ since the neighboring dimensions do not match. ◇

*) REMARK[2]: Matrix multiplication is NOT defined as an element-wise operation on matrix elements, i.e., $c_{ij} \neq a_{ij} b_{ij}$ (even if the size of $\underline{A}$, $\underline{B}$ was chosen appropriately). This kind of element-wise multiplication often appears in programing languages when we multiply (multi-dimensional) arrays with each other, and is called a Hadamard product. ◇

▶ DEFINITION[2] (Identity Matrix) In $\mathbb{R}^{n \times n}$, we define the identity matrix as:

$$\mathbb{1}_n \equiv \underline{I}_n := \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}_{n \times n} \equiv \begin{pmatrix} 1 & & & \mathbb{0} \\ & 1 & & \\ & & \ddots \; 1 & \\ \mathbb{0} & & & \ddots \; 1 \end{pmatrix} \in \mathbb{R}^{n \times n} \,, \tag{1.4}$$

as the $n \times n$-matrix containing 1 on the diagonal and 0 everywhere else. Note that $\underline{I}_m \neq \underline{I}_n$ for $m \neq n$.

↳ Now that we defined matrix multiplication, matrix addition and the identity matrix, let us have a look at some properties of matrices:

- Associativity: $\forall\, \underline{A} \in \mathbb{R}^{m \times n}, \underline{B} \in \mathbb{R}^{n \times p}, \underline{C} \in \mathbb{R}^{p \times q}: (\underline{A}\,\underline{B})\underline{C} = \underline{A}(\underline{B}\,\underline{C})$;
- Distributivity: $\forall\, \underline{A}, \underline{B} \in \mathbb{R}^{m \times n}, \underline{C}, \underline{D} \in \mathbb{R}^{n \times p}: \begin{cases} (\underline{A}+\underline{B})\underline{C} = \underline{A}\,\underline{C} + \underline{B}\,\underline{C} \\ \underline{A}(\underline{C}+\underline{D}) = \underline{A}\,\underline{C} + \underline{A}\,\underline{D} \end{cases}$;
- Multiplication with the identity matrix: $\forall\, \underline{A} \in \mathbb{R}^{m \times n}: \underline{I}_m \underline{A} = \underline{A}\,\underline{I}_n = \underline{A}$;

1.2) INVERSE AND TRANSPOSE

▸ DEFINITION[3] (Inverse): Consider a square matrix $\underline{A} \in \mathbb{R}^{n\times n}$. Let a matrix $\underline{B} \in \mathbb{R}^{n\times n}$ have the property that $\underline{A}\underline{B} = \underline{I}_n = \underline{B}\underline{A}$. The matrix $\underline{B}$ is called the inverse of $\underline{A}$ and denoted by $\underline{A}^{-1}$.

↳ Unfornately, not every matrix $\underline{A}$ possesses an inverse $\underline{A}^{-1}$. If this inverse does exist, $\underline{A}$ is called regular, invertible, or non-singular, otherwise is singular or non-invertible. When the matrix inverse exists, it is unique.

*) REMARK[3] (Existence of the Inverse of a 2x2 matrix): Consider the matrices

$$\underline{A} := \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad , \quad \underline{B} := \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} \quad , \text{ where : } \underline{A}, \underline{B} \in \mathbb{R}^{2\times 2} \quad , \quad \text{so :}$$

$$\Rightarrow \underline{A}\underline{B} = \begin{pmatrix} a_{11}a_{22} - a_{12}a_{21} & 0 \\ 0 & a_{11}a_{22} - a_{12}a_{21} \end{pmatrix} = (a_{11}a_{22} - a_{12}a_{21})\, \underline{I}_2 \text{ , therefore :}$$

$$\therefore \quad \underline{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} = (a_{11}a_{22} - a_{12}a_{21})^{-1}\, \underline{B} \quad , \qquad \underline{(1.5)}$$

if and only if $a_{11}a_{22} - a_{12}a_{21} \neq 0$. In fact, the term $(a_{11}a_{22} - a_{12}a_{21})$ is the determinant of a 2x2 matrix. Furthermore, we can generally use the determinant to check whether a matrix is invertible.

◇

▸ DEFINITION[4] (Transpose): For a matrix $\underline{A} \in \mathbb{R}^{m\times n}$ the matrix $\underline{B} \in \mathbb{R}^{n\times m}$ with $b_{ij} = a_{ji}$ is called the transpose of $\underline{A}$. We write $\underline{B} = \underline{A}^T$.

↳ In general, $\underline{A}^T$ can be obtained by writing the columns of $\underline{A}$ as the rows of $\underline{A}^T$. The following are important properties of inverses and transposes:

$$\left\{ \begin{array}{l} 1)\ \underline{A}\underline{A}^{-1} = \underline{A}^{-1}\underline{A} = \underline{I} \ , \\ 2)\ (\underline{A}\underline{B})^{-1} = \underline{B}^{-1}\underline{A}^{-1} \ , \\ 3)\ (\underline{A}+\underline{B})^{-1} \neq \underline{A}^{-1} + \underline{B}^{-1} \ , \end{array} \right. \qquad \left\{ \begin{array}{l} 4)\ (\underline{A}^T)^T = \underline{A} \ , \\ 5)\ (\underline{A}+\underline{B})^T = \underline{A}^T + \underline{B}^T \ ; \\ 6)\ (\underline{A}\underline{B})^T = \underline{B}^T\underline{A}^T \ , \end{array} \right.$$

▸ **DEFINITION[5]** (Symmetric Matrix): A matrix $\underline{A} \in \mathbb{R}^{n \times n}$ is symmetric if $\underline{A} = \underline{A}^T$.

↳ Note that only (n,n)-matrices can be symmetric. Generally, we call (n,n)-matrices also square matrices because they possess the same number of rows and columns. Moreover, if $\underline{A}$ is invertible, then so is $\underline{A}^T$, and $(\underline{A}^{-1})^T = (\underline{A}^T)^{-1} =: \underline{A}^{-T}$.

*) **REMARK[4]** (Sum and Product of Symmetric Matrices): The sum of symmetric matrices $\underline{A}, \underline{B} \in \mathbb{R}^{n \times n}$ is always symmetric. However, although their product is always defined, it is generally not symmetric. For example:

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

◇

1.3) MULTIPLICATION BY A SCALAR

↳ Let us look at what happens to matrices when they are multiplied by a scalar $\lambda \in \mathbb{R}$. Let $\underline{A} \in \mathbb{R}^{m \times n}$ and $\lambda \in \mathbb{R}$. Then, $\lambda \underline{A} = \underline{K}$, with $K_{ij} = \lambda a_{ij}$. Pratically, λ scales each element of $\underline{A}$. For $\lambda, \psi \in \mathbb{R}$, the following holds:

- 1) Associativity: $(\lambda\psi)\underline{C} = \lambda(\psi\underline{C})$, $\underline{C} \in \mathbb{R}^{m \times n}$;
- 2) $\lambda(\underline{B}\underline{C}) = (\lambda\underline{B})\underline{C} = \underline{B}(\lambda\underline{C}) = (\underline{B}\underline{C})\lambda$, $\underline{B} \in \mathbb{R}^{m \times n}$, $\underline{C} \in \mathbb{R}^{n \times k}$;
- 3) $(\lambda\underline{C})^T = \underline{C}^T\lambda^T = \underline{C}^T\lambda = \lambda\underline{C}^T$, since $\lambda = \lambda^T$ for all $\lambda \in \mathbb{R}$;
- 4) Distributivity: $\begin{cases} (\lambda+\psi)\underline{C} = \lambda\underline{C} + \psi\underline{C}, & \underline{C} \in \mathbb{R}^{m \times n} \\ \lambda(\underline{A}+\underline{B}) = \lambda\underline{A} + \lambda\underline{B}, & \underline{B}, \underline{C} \in \mathbb{R}^{m \times n} \end{cases}$;

*) **REMARK[5]**: Generally, a system of linear equations can be compactly represented in their matrix form as $\underline{A}\underline{x} = \underline{b}$, where the product $\underline{A}\underline{x}$ is a (linear) combination of the columns of $\underline{A}$. In symbols:

$$\begin{cases} a_{11}x_1 + \dots + a_{1n}x_n = b_1 \\ \vdots \qquad\qquad \vdots \qquad \vdots \\ a_{m1}x_1 + \dots + a_{mn}x_n = b_m \end{cases}, \tag{1.6}$$

where $\underline{A} \in \mathbb{R}^{m \times n}$ with $a_{ij} \in \mathbb{R}$; $\underline{x} \in \mathbb{R}^{n \times 1}$, and $\underline{b} \in \mathbb{R}^{m \times 1}$ with $b_i \in \mathbb{R}$.

◇

2. LINEAR ALGEBRA

↳ In a vector space V, we are particularly interested in sets of vectors $\mathcal{A}$ that possess the property that any vector $\underline{v} \in V$ can be obtained by a linear combination of vectors in $\mathcal{A}$. These vectors are special vectors, and in the following, we will caracterize them.

2.1) GENERATING SETS AND BASIS

↳ Generating sets are sets of vectors that span vector (sub)spaces, i.e., every vector can be represented as a linear combination of the vectors in the generating set. Now, we will be more specific and characterize the smallest generating set that spans a vector (sub) space.

▶ DEFINITION[6] (Generating Set and Span): Consider a vector space $V=(\mathcal{V},+,\cdot)$ and a set of vectors $\mathcal{A}=\{\underline{x}_1,\dots,\underline{x}_k\} \subseteq \mathcal{V}$. If every vector $\underline{v} \in \mathcal{V}$ can be expressed as a linear combination of $\underline{x}_1,\dots,\underline{x}_k$, $\mathcal{A}$ is called a generating set of V. The set of all linear combinations of vectors in $\mathcal{A}$ is called span of $\mathcal{A}$. If $\mathcal{A}$ spans the vector space V, we write $V = \text{span}[\mathcal{A}] = \text{span}\{\mathcal{A}\}$ or $V = \text{span}[\underline{x}_1,\dots,\underline{x}_k] = \{\underline{x}_1,\dots,\underline{x}_k\}$.

▶ DEFINITION[7] (Basis): Consider a vector space $V=(\mathcal{V},+,\cdot)$ and $\mathcal{A} \subseteq \mathcal{V}$. A generating set $\mathcal{A}$ of V is called minimal if there exists no smaller set $\tilde{\mathcal{A}} \subseteq \mathcal{A} \subseteq \mathcal{V}$ that spans V. Every linearly independent generating set of V is minimal and is called a basis of V.

↳ Let $V=(\mathcal{V},+,\cdot)$ be a vector space and $\mathcal{B} \subseteq \mathcal{V}$, $\mathcal{B} \neq \emptyset$. Then, the following statements are equivalent:

- $\mathcal{B}$ is a basis of V.
- $\mathcal{B}$ is a minimal generating set.
- $\mathcal{B}$ is a maximal linearly independent set of vectors in V, i.e., adding any other vector to this set will make it linearly dependent.
- Every vector $\underline{x} \in V$ is a linear combination of vectors from $\mathcal{B}$, and every linear combination is unique, i.e., with:

$$\underline{x} = \sum_{i=1}^{k} \lambda_i \underline{b}_i = \sum_{i=1}^{k} \psi_i \underline{b}_i \,, \qquad (1.7)$$

and $\lambda_i, \psi_i \in \mathbb{R}$, $\underline{b}_i \in \mathcal{B}$ it follows that $\lambda_i = \psi_i$, with $i=1,\dots,k$.

*) REMARK[6]: Every vector space V possesses a basis $\mathcal{B}$. The canonical/standard basis show that there can be many bases of a vector space V, i.e., there is no unique basis. However, all bases possess the same number of elements, the basis vectors.

◊

↳ We only consider finite-dimensional vector spaces V. In this case, the <u>dimension</u> of V is the number of basis vectors of V, and we write $\dim(V)$. If $U \subseteq V$ is a subspace of V, then $\dim(U) \leq \dim(V)$ and $\dim(U) = \dim(V)$ if and only if $U = V$. Intuitively, the dimension of a vector space can be thought of as the number of independent directions in this vector space.

*) <u>REMARK[7]</u>: The dimension of a vector space is not necessarily the number of elements in a vector. For instance, the vector space $V = \mathrm{span}\left\{\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right\}$ is one-dimensional, although the basis vector possesses two elements.

◇

*) <u>REMARK[8]</u>: A basis of a subspace $U = \mathrm{span}\{\underline{x}_1, \dots, \underline{x}_m\} \subseteq \mathbb{R}^n$ can be found by executing the following steps:

- 1) Write the spanning vectors as columns of a matrix $\underline{A}$;
  2) Determine the row-echelon form of $\underline{A}$;
  3) The spanning vectors associated with the pivot columns are a basis of U.

◇

2.2) <u>RANK</u>

↳ The number of linearly independent columns of a matrix $\underline{A} \in \mathbb{R}^{m \times n}$ equals the number of linearly independent rows and is called the <u>rank</u> of $\underline{A}$ and is denoted by $\mathrm{rank}(\underline{A}) \equiv \mathrm{rk}(\underline{A})$. The rank of a matrix has some important properties:

1) $\mathrm{rk}(\underline{A}) = \mathrm{rk}(\underline{A}^T)$, i.e., the column rank equals the row rank ;

2) The columns of $\underline{A} \in \mathbb{R}^{m \times n}$ span a subspace $U \subseteq \mathbb{R}^m$ with $\dim(U) = \mathrm{rk}(\underline{A})$. Later, we will call this subspace the <u>image</u> or <u>range</u>. A basis of U can be found by applying Gaussian elimination to $\underline{A}$ to identify the pivot columns ;

3) The rows of $\underline{A} \in \mathbb{R}^{m \times n}$ span a subspace $W \subseteq \mathbb{R}^n$ with $\dim(W) = \mathrm{rk}(\underline{A})$. A basis of W can be found by applying Gaussian elimination to $\underline{A}^T$;

4) For all $\underline{A} \in \mathbb{R}^{n \times n}$ it holds that $\underline{A}$ is regular (invertible) if and only if $\mathrm{rk}(\underline{A}) = n$,

5) For all $\underline{A} \in \mathbb{R}^{m \times n}$ and all $\underline{b} \in \mathbb{R}^m$ it holds that the linear equation $\underline{A}\underline{x} = \underline{b}$ can be solved if and only if $\mathrm{rk}(\underline{A}) = \mathrm{rk}(\underline{A}|\underline{b})$, where $\underline{A}|\underline{b}$ denotes the augmented system ;

6) For $\underline{A} \in \mathbb{R}^{m \times n}$ the subspace of solutions for $\underline{A}\underline{x} = \underline{0}$ possesses dimension $n - \mathrm{rk}(\underline{A})$. Later, we will call this subspace the <u>kernel</u> or the <u>null space</u>.

7) A matrix $\underline{A} \in \mathbb{R}^{m \times n}$ has <u>full rank</u> if it rank equals the largest possible rank for a matrix of the same dimensions. This means that the rank of a full-rank matrix is the lesser of the number of rows and columns, i.e., $\mathrm{rk}(\underline{A}) = \min(m, n)$. A matrix is said to be <u>rank deficient</u> if it does not have a full rank.

2.3) Linear Mappings

↳ In the following, we will study mappings on vector spaces that preserve their structure, which will allow us to define the concept of a coordinate. In the beginning of this notes, we said that vectors are objects that can be added together and multiplied by a scalar, and the resulting object is still a vector. We wish to preserve this property when applying the mapping: Consider two real vector spaces V, W. A mapping $\Phi: V \longrightarrow W$ preserves the structure of the vector space if:

$$\bullet \begin{cases} \Phi(\underline{x}+\underline{y}) = \Phi(\underline{x}) + \Phi(\underline{y}) \\ \Phi(\lambda \underline{x}) = \lambda \Phi(\underline{x}) \end{cases} \qquad (1.8)$$

for all $\underline{x}, \underline{y} \in V$ and $\lambda \in \mathbb{R}$. We can summarize this in the following definition:

▸ **Definition[8]** (Linear Mapping): For vector spaces V, W, a mapping $\Phi: V \to W$ is called a <u>linear mapping</u> (or <u>vector space homomorphism</u> / <u>linear transformation</u>) if:

$$\forall\, \underline{x}, \underline{y} \in V,\ \forall \lambda, \psi \in \mathbb{R} : \Phi(\lambda \underline{x} + \psi \underline{y}) = \lambda \Phi(\underline{x}) + \psi \Phi(\underline{y}) \ . \qquad (1.9)$$

↳ It turns out that we can represent linear mappings as matrices. Recall that we can also collect a set of vectors as columns of a matrix. When working with matrices, we have to keep in mind what the matrix represents: a linear mapping or a collection of vectors. Before we continue, we will briefly introduce special mappings.

▸ **Definition[9]** (Injective, Surjective, Bijective): Consider a mapping $\Phi: \mathcal{V} \to \mathcal{W}$, where $\mathcal{V}, \mathcal{W}$ can be arbitrary sets. Then Φ is called:

- <u>Injective</u> if $\forall \underline{x}, \underline{y} \in \mathcal{V}: \Phi(\underline{x}) = \Phi(\underline{y}) \Rightarrow \underline{x} = \underline{y}$;
- <u>Surjective</u> if $\Phi(\mathcal{V}) = \mathcal{W}$;
- <u>Bijective</u> if it is injective and surjective ;

↳ If Φ is surjective, then every element in $\mathcal{W}$ can be "reached" from $\mathcal{V}$ using Φ. A bijective Φ can be "undone", i.e., there exists a mapping $\Psi: \mathcal{W} \to \mathcal{V}$ so that $\Psi \circ \Phi(\underline{x}) = \underline{x}$. This mapping Ψ is then called the inverse of Φ and normally denoted by Φ^{-1}. With these definitions, we introduce the following special cases of linear mappings between vector spaces V and W:

- <u>Isomorphism</u>: $\Phi: V \to W$ linear and bijective ,
- <u>Endomorphism</u>: $\Phi: V \to V$ linear ;
- <u>Automorphism</u>: $\Phi: V \to V$ linear and bijective ;
- We define $\mathrm{id}_V: V \to V$, $\underline{x} \mapsto \underline{x}$ as the <u>identity mapping</u> or <u>identity automorphism</u> in V.

THEOREM[1] (Theorem 3.59 in Axler, 2015*): Finite-dimensional vector spaces V and W are isomorphic if and only if $\dim(V) = \dim(W)$.

*Axler, Sheldon. 2015. *Linear Algebra Done Right*. Springer.

↪ The THEOREM[1] states that there exists a linear, bijective mapping between two vector spaces of the same dimension. Intuitively, this means that vector spaces of the same dimension are kind of the same thing, as they can be transformed into each other without incurring any loss.

The THEOREM[1] also gives us the justification to treat $\mathbb{R}^{m\times n}$ (the vector space of $m\times n$-matrices) and $\mathbb{R}^{mn}$ (the vector space of vectors of length mn) the same, as their dimensions are mn, and there exists a linear, bijective mapping that transforms one into the other.

*) REMARK[9]: Consider vector spaces V, W, X. Then:

- For linear mappings $\Phi: V \to W$ and $\Psi: W \to X$, the mapping $\Psi \circ \Phi: V \to X$ is also linear;
- If $\Phi: V \to W$ is an isomorphism, then $\Phi^{-1}: W \to V$ is an isomorphism, too.
- If $\Phi: V \to W$, $\Psi: V \to W$ are linear, then $\Phi + \Psi$ and $\lambda\Phi$, $\lambda \in \mathbb{R}$, are linear, too.

◇

2.4) MATRIX REPRESENTATION OF LINEAR MAPPINGS

↪ Any n-dimensional vector space is isomorphic to $\mathbb{R}^n$ (THEOREM[1]). We consider a basis $\{\underline{b}_1, \ldots, \underline{b}_n\}$ of an n-dimensional vector space V. In the following, the order of the basis vectors will be important. Therefore, we write:

$$B = (\underline{b}_1, \ldots, \underline{b}_n), \tag{1.10}$$

and call this n-tuple and ordered basis of V.

*) REMARK[10] (Notation): We are at the point where notation gets a bit tricky. Therefore, we summarize some parts here. $B = (\underline{b}_1, \ldots, \underline{b}_n)$ is an ordered basis, $\mathcal{B} = \{\underline{b}_1, \ldots, \underline{b}_n\}$ is an (unordered) basis, and $\underline{B} = [\underline{b}_1, \ldots, \underline{b}_n]$ is a matrix whose columns are the vectors $\underline{b}_1, \ldots, \underline{b}_n$.

◇

▸ DEFINITION[10] (Coordinates): Consider a vector space V and an ordered basis $B = (\underline{b}_1, \ldots, \underline{b}_n)$ of V. For any $\underline{x} \in V$ we obtain a unique representation (linear combination):

$$\underline{x} = \sum_{i=1}^{n} \alpha_i \underline{b}_i = \alpha_1 \underline{b}_1 + \ldots + \alpha_n \underline{b}_n, \tag{1.11}$$

of $\underline{x}$ with respect to B. Then $\alpha_1, \ldots, \alpha_n$ are the coordinates of $\underline{x}$ with respect to B, and the vector: [...]

[...]

$$\underline{\alpha} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \in \mathbb{R}^n \quad , \tag{1.12}$$

is the coordinate vector / coordinate representation of $\underline{x}$ with respect to the ordered basis B.

↳ A basis effectively defines a coordinate system. We are familiar with the Cartesian coordinate system in two dimensions, which is spanned by the canonical basis vector $\underline{e}_1, \underline{e}_2$. In this coordinate system, a vector $\underline{x} \in \mathbb{R}^2$ has a representation that tells us how to linearly combine $\underline{e}_1$ and $\underline{e}_2$ to obtain $\underline{x}$. However, any basis of $\mathbb{R}^2$ defines a valid coordinate system, and the same vector $\underline{x}$ from before may have a different coordinate representation in the $(\underline{b}_1, \underline{b}_2)$ basis.

*) REMARK[11]: For an n-dimensional vector space V and an ordered basis B of V, the mapping $\Phi : \mathbb{R}^n \to V$, and $\Phi(\underline{e}_i) = \underline{b}_i$, $i = 1, \dots, n$ is linear (and because of THEOREM', p. 8 an isomorphism), where $(\underline{e}_1, \dots, \underline{e}_n)$ is the standard basis of $\mathbb{R}^n$.

◇

▸ DEFINITION[11] (Transformation Matrix): Consider vector spaces V, W with corresponding (ordered) basis given by $B = (\underline{b}_1, \dots, \underline{b}_n)$ and $C = (\underline{c}_1, \dots, \underline{c}_m)$. Moreover, we consider a linear mapping $\Phi : V \to W$. For $j \in \{1, \dots, n\}$,

$$\Phi(\underline{b}_j) = \alpha_{1j}\underline{c}_1 + \dots + \alpha_{mj}\underline{c}_m = \sum_{i=1}^{m} \alpha_{ij}\underline{c}_i \quad , \tag{1.13}$$

is the unique representation of $\Phi(\underline{b}_j)$ with respect to C. Then, we call the $m \times n$-matrix $\underline{A}_\Phi$, whose elements are given by:

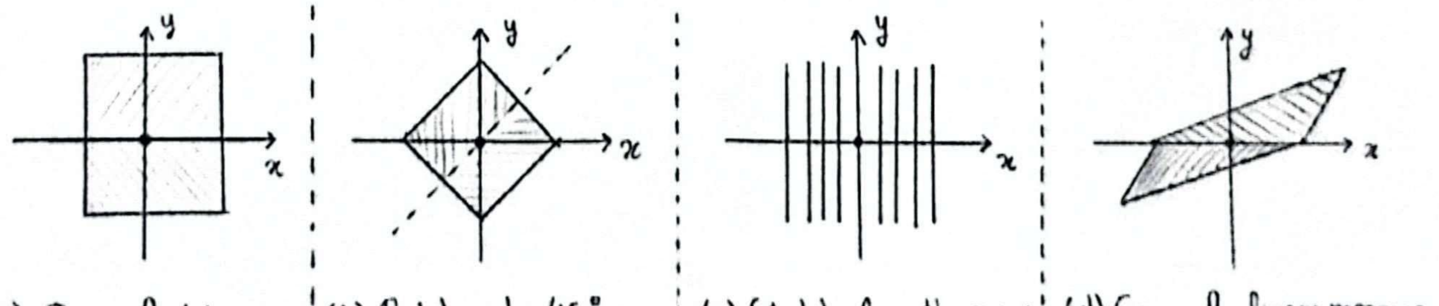
$$A_\Phi(i,j) = \alpha_{ij} \quad , \tag{1.14}$$

the transformation matrix of Φ (with respect to the ordered basis B of V and C of W).

↳ The coordinates of $\Phi(\underline{b}_j)$ with respect to the ordered basis C of W are the j^{th} column of $\underline{A}_\Phi$. Consider (finite-dimensional) vector spaces V, W with ordered basis B, C and a linear mapping $\Phi : V \to W$ with transformation matrix $\underline{A}_\Phi$. If $\hat{\underline{x}}$ is the coordinate vector of $\underline{x} \in V$ with respect to B and $\hat{\underline{y}}$ the coordinate vector of $\underline{y} = \Phi(\underline{x}) \in W$ with respect to C, then:

$$\hat{\underline{y}} = \underline{A}_\Phi \hat{\underline{x}} \quad . \tag{1.15}$$

This means that the transformation matrix can be used to map coordinates with respect to an ordered basis in V to coordinates with respect to an ordered basis in W.



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

2.5) EIGENVALUES AND EIGENVECTORS

↳ We will now get to know a new way to characterize a matrix and its associated linear mapping. We known that every linear mapping has a unique transformation matrix given an ordered basis. We can interpret linear mappings and their associated transformation matrices by performing an "eigen" analysis.

▸ DEFINITION[12] (Eigenvalue Equation): Let $\underline{A} \in \mathbb{R}^{n\times n}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an eigenvalue of $\underline{A}$ and $\underline{x} \in \mathbb{R}^n \setminus \{\underline{0}\}$ is the corresponding eigenvector of $\underline{A}$ if:

$$\underline{A}\,\underline{x} = \lambda \underline{x} \;. \tag{1.16}$$

We call Eq. (1.16) the eigenvalue equation.

*) REMARK[11]: In the linear algebra literature and software, it is often a convention that eigenvalues are sorted in descending order, so that the largest eigenvalue and associated eigenvector are called the first eigenvalue and its associated eigenvector, and the second largest called the second eigenvalue and its associated eigenvector, and so on. However, textbooks and publications may have different or no notion of orderings.

◇

The following statements are equivalent:

- λ is an eigenvalue of $\underline{A} \in \mathbb{R}^{n\times n}$;
- There exists an $\underline{x} \in \mathbb{R}^n \setminus \{\underline{0}\}$ with $\underline{A}\underline{x} = \lambda \underline{x}$, or equivalently, $(\underline{A} - \lambda \underline{I}_n)\underline{x} = \underline{0}$ can be solved non-trivially, i.e., with $\underline{x} \neq \underline{0}$;
- $\mathrm{rk}(\underline{A} - \lambda \underline{I}_n) < n$;
- $\det(\underline{A} - \lambda \underline{I}_n) = 0$;

▸ DEFINITION[13] (Collinearity and Codirection): Two vectors that point in the same direction are called codirected. Two vectors are collinear if they point in the same or the opposite direction.

*) REMARK[12] (Non-uniqueness of eigenvectors): If $\underline{x}$ is an eigenvector of $\underline{A}$ associated with eigenvalue λ, then for any $c \in \mathbb{R}\setminus\{0\}$ it holds that $c\underline{x}$ is an eigenvector of $\underline{A}$ with the same eigenvalue since:

$$\underline{A}(c\underline{x}) = c\underline{A}\,\underline{x} = c\lambda\underline{x} = \lambda(c\underline{x}) \;. \tag{1.17}$$

Thus, all vectors that are collinear to $\underline{x}$ are also eigenvectors of $\underline{A}$.

◇

■ THEOREM[2]: $\lambda \in \mathbb{R}$ is an eigenvalue of $\underline{A} \in \mathbb{R}^{n\times n}$ if and only if λ is a root of the characteristic polynomial $p_A(\lambda)$ of $\underline{A}$.

▸ <u>DEFINITION</u>[14] : Let a square matrix $\underline{A}$ have an eigenvalue λ_i. The <u>algebraic multiplicity</u> of λ_i is the number of times the root appears in the characteristic polynomial.

▸ <u>DEFINITION</u>[15] (Eigenspace and Eigenspectrum) : For $\underline{A} \in \mathbb{R}^{n\times n}$, the set of all eigenvectors of $\underline{A}$ associated with an eigenvalue λ spans a subspace of $\mathbb{R}^n$, which is called the <u>eigenspace</u> of $\underline{A}$ with respect to λ and is denoted by E_λ. The set of all eigenvalues of $\underline{A}$ is called the <u>eigenspectrum</u>, or just <u>spectrum</u>, of $\underline{A}$.

↳ If λ is an eigenvalue of $\underline{A} \in \mathbb{R}^{n\times n}$, then the corresponding eigenspace E_λ is the solution space of the homogeneous system of linear equations $(\underline{A}-\lambda\underline{I})\underline{x} = \underline{0}$. Geometrically, the eigenvector corresponding to a nonzero eigenvalue points in a direction that is stretched by the linear mapping. The eigenvalue is the factor by which it is stretched. If the eigenvalue is negative, the direction of the stretching is flipped.

Useful properties regarding eigenvalues and eigenvectors include the following:

- A matrix $\underline{A}$ and its transpose $\underline{A}^T$ possess the same eigenvalues, but not necessarily the same eigenvectors ;
- The eigenspace E_λ is the null space of $(\underline{A}-\lambda\underline{I})$ since :

$$\underline{A}\underline{x} = \lambda\underline{x} \iff \underline{A}\underline{x} - \lambda\underline{x} = \underline{0} \iff (\underline{A}-\lambda\underline{I})\underline{x} = \underline{0} \iff \underline{x} \in \ker(\underline{A}-\lambda\underline{I}) \quad . \tag{1.18}$$

- Similar matrices possess the same eigenvalues. Therefore, a linear mapping Φ has eigenvalues that are independent of the choice of basis of its transformation matrix. This makes eigenvalues, together with the determinant and the trace, key characteristic parameters of a linear mapping as they are all invariant under basis change.
- Symmetric, positive definite matrices always have positive, real eigenvalues.

▸ <u>DEFINITION</u>[16] : Let λ_i be an eigenvalue of a square matrix $\underline{A}$. Then the <u>geometric multiplicity</u> of λ_i is the number of linearly independent eigenvectors associated with λ_i. In other words, it is the dimensionality of the eigenspace spanned by the eigenvectors associated with λ_i

*) <u>REMARK</u>[13] : A specific eigenvalue's geometric multiplicity must be at least one because every eigenvalue has at least one associated eigenvector. An eigenvalue's geometric multiplicity cannot exceed its algebraic multiplicity, but it may be lower.

◇

■ <u>THEOREM</u>[3] : The eigenvectors $\underline{x}_1, \dots, \underline{x}_n$ of a matrix $\underline{A} \in \mathbb{R}^{n\times n}$ with n distinct eigenvalues $\lambda_1, \dots, \lambda_n$ are linearly independent.

→ This theorem states that eigenvectors of a matrix with n distinct eigenvalues form a basis of $\mathbb{R}^n$.

▶ DEFINITION[17] : A square matrix $\underline{A} \in \mathbb{R}^{n \times n}$ is defective if it possesses fewer than n linearly independent eigenvectors.

↪ A non-defective matrix $\underline{A} \in \mathbb{R}^{n \times n}$ does not necessarily require n distinct eigenvalues, but it does require that the eigenvectors form a basis of $\mathbb{R}^n$. Looking at the eigenspaces of a defective matrix, it follows that the sum of the dimensions of the eigenspaces is less than n. Specifically, a defective matrix has at least one eigenvalue λ_i with an algebraic multiplicity $m > 1$ and a geometric multiplicity of less than m.

*) REMARK[14] : A defective matrix cannot have n distinct eigenvalues, as distinct eigenvalues have linearly independent eigenvectors (THEOREM[3]). ◇

■ THEOREM[4] : Given an matrix $\underline{A} \in \mathbb{R}^{m \times n}$, we can always obtain a symmetric, positive and semi-definite matrix $\underline{S} \in \mathbb{R}^{n \times n}$ by defining:

$$\underline{S} := \underline{A}^T \underline{A} \quad . \tag{1.19}$$

*) REMARK[15] : If $\mathrm{rk}(\underline{A}) = n$, then $\underline{S} := \underline{A}^T \underline{A}$ is symmetric, positive and definite.

↪ Understanding why THEOREM[4] holds is insightful for how we can uses symmetrized matrices: Symmetry requires that $\underline{S} = \underline{S}^T$, and by inserting Eq. (1.19) we obtain $\underline{S} = \underline{A}^T\underline{A} = \underline{A}^T(\underline{A}^T)^T = (\underline{A}^T\underline{A})^T = \underline{S}^T$. Moreover, positive semi-definiteness requires that $\underline{x}^T \underline{S}\, \underline{x} \geq 0$ and inserting Eq. (1.19) we obtain $\underline{x}^T \underline{S}\, \underline{x} = \underline{x}^T \underline{A}^T \underline{A}\, \underline{x} = (\underline{x}^T \underline{A}^T)(\underline{A}\underline{x}) = (\underline{A}\underline{x})^T(\underline{A}\underline{x}) \geq 0$, because the dot product computes a sum of squares (which are themselves non-negative).

■ THEOREM[5] (Spectral Theorem) : If $\underline{A} \in \mathbb{R}^{n \times n}$ is symmetric, there exists an othernormal basis of the corresponding vector space V consisting of eigenvectors of $\underline{A}$, and each eigenvalue is real.

↪ A direct implication of the spectral theorem is that the eigendecomposition of a symmetric matrix $\underline{A}$ exists (with real eigenvalues), and that we can find an ONB of eigenvectors so that $\underline{A} = \underline{P}\,\underline{D}\,\underline{P}^T$, where $\underline{D}$ is diagonal and the columns of $\underline{P}$ contain the eigenvectors.

■ THEOREM[6] : The determinant of a matrix $\underline{A} \in \mathbb{R}^{n \times n}$ is the product of its eigenvalues, i.e.,

$$\det(\underline{A}) = \prod_{i=1}^{n} \lambda_i \quad , \tag{1.20}$$

where λ_i are (possibly repeated) eigenvalues of $\underline{A}$.

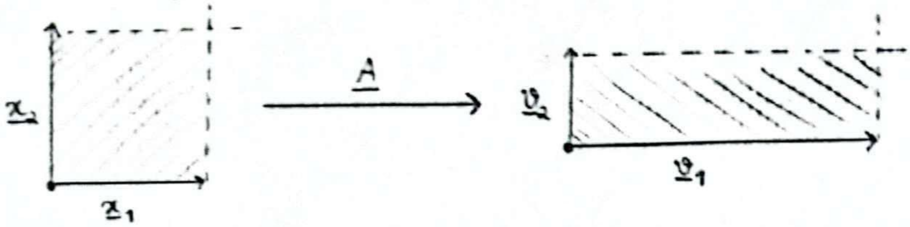
[...]

■ <u>THEOREM[7]</u>: The trace of a matrix $\underline{A} \in \mathbb{R}^{n \times n}$ is the sum of its eigenvalues, i.e.,

$$tr(\underline{A}) = \sum_{i=1}^{n} \lambda_i \,, \tag{1.21}$$

where λ_i are (possibly repeated) eigenvalues of $\underline{A}$.

↳ Let us provide a geometric intuition of these last two theorems. Consider a matrix $\underline{A} \in \mathbb{R}^{2 \times 2}$ that possesses two linearly independent eigenvectors $\underline{x}_1, \underline{x}_2$. For this example, we assume $(\underline{x}_1, \underline{x}_2)$ are an ONB of $\mathbb{R}^2$ so that they are orthogonal and the area of the square they span is 1. We know that the determinant computes the change of area of unit square under the transformation $\underline{A}$. In this example, we can compute the change of area explicitly: Mapping the eigenvectors using $\underline{A}$ gives us vectors $\underline{v}_1 = \underline{A}\underline{x}_1 = \lambda_1 \underline{x}_1$ and $\underline{v}_2 = \underline{A}\underline{x}_2 = \lambda_2 \underline{x}_2$, i.e., the new vectors $\underline{v}_i$ are scaled versions of the eigenvectors $\underline{x}_i$, and the scaling factors are the corresponding eigenvalues λ_i. Also, $\underline{v}_1, \underline{v}_2$ are still orthogonal, and the area of the rectangle they span is $|\lambda_1 \lambda_2|$.



↳ Given that $\underline{x}_1, \underline{x}_2$ (in this example) are orthonormal, we can directly compute the circumference of the unit square as $2(1+1)$. Mapping the eigenvectors using $\underline{A}$ creates a rectangle whose circumference is $2(|\lambda_1| + |\lambda_2|)$. Therefore, the sum of the absolute values of the eigenvalues tells us how the circumference of the unit square changes under the transformation matrix $\underline{A}$.

2.6) <u>EIGENDECOMPOSITION AND DIAGONALIZATION</u>

↳ A <u>diagonal matrix</u> is a matrix that has value zero on all off-diagonal elements, i.e., they are of the form:

$$\underline{D} = \begin{pmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_n \end{pmatrix} = \begin{pmatrix} c_1 & & \mathbb{O} \\ & \ddots & \\ \mathbb{O} & & c_n \end{pmatrix} = diag(c_i) \,. \tag{1.22}$$

They allow fast computation of determinants, powers, and inverses. The determinant is the product of its diagonal entries, a matrix $\underline{D}^k$ is given by each diagonal element raised to the power of k, and the inverse $\underline{D}^{-1}$ is the reciprocal of its diagonal elements if all of them are nonzero.

Recall that two matrices $\underline{A}, \underline{D}$ are similar if there exists an invertible matrix $\underline{P}$, such that $\underline{D} = \underline{P}^{-1}\underline{A}\underline{P}$. More specifically, we will look at matrices $\underline{A}$ that are similar to diagonal matrices $\underline{D}$ that contain the eigenvalues of $\underline{A}$ on the diagonal.

▸ DEFINITION[18] (Diagonalizable): A matrix $\underline{A} \in \mathbb{R}^{n \times n}$ is <u>diagonalizable</u> if it is similar to a diagonal matrix, i.e., if there exists an invertible matrix $\underline{P} \in \mathbb{R}^{n \times n}$ such that $\underline{D} = \underline{P}^{-1}\underline{A}\,\underline{P}$.

↳ Let $\underline{A} \in \mathbb{R}^{n \times n}$, let $\lambda_1, \dots, \lambda_n$ be a set of scalars, and let $\underline{p}_1, \dots \underline{p}_n$ be a set of vectors in $\mathbb{R}^n$. We can define $\underline{P} := (\underline{p}_1, \dots, \underline{p}_n)$ and let $\underline{D} \in \mathbb{R}^{n \times n}$ be a diagonal matrix with diagonal entries $\lambda_1, \dots, \lambda_n$. Then we can show that:

$$\underline{A}\,\underline{P} = \underline{P}\,\underline{D} \,, \tag{1.23}$$

if and only if $\lambda_1, \dots, \lambda_n$ are the eigenvalues of $\underline{A}$ and $\underline{p}_1, \dots, \underline{p}_n$ are corresponding eigenvectors of $\underline{A}^T$. We can see that this statement holds because:

•) $$\underline{A}(\underline{p}_1, \dots, \underline{p}_n) = (\underline{p}_1, \dots, \underline{p}_n) \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \Rightarrow \underline{A}(\underline{p}_1, \dots, \underline{p}_n) = (\lambda_1 \underline{p}_1, \dots, \lambda_n \underline{p}_n) \,. \tag{1.24}$$

Thus, Eq. (1.23) implies that:

•) $$\begin{cases} \underline{A}\,\underline{p}_1 = \lambda_1 \underline{p}_1 \,, \\ \ \vdots \quad \vdots \quad \vdots \\ \underline{A}\,\underline{p}_n = \lambda_n \underline{p}_n \,. \end{cases} \tag{1.25}$$

Therefore, the columns of $\underline{P}$ must be eigenvectors of $\underline{A}$.

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

■ THEOREM[8] (Eigendecomposition): A square matrix $\underline{A} \in \mathbb{R}^{n \times n}$ can be factored into:

$$\underline{A} = \underline{P}\,\underline{D}\,\underline{P}^{-1} \,, \tag{1.26}$$

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

↳ The THEOREM[8] implies that only non-defective matrices can be diagonalized and that the columns of $\underline{P}$ are the n eigenvectors of $\underline{A}$. For symmetric matrices we can obtain even stronger outcomes for the eigenvalue decomposition.

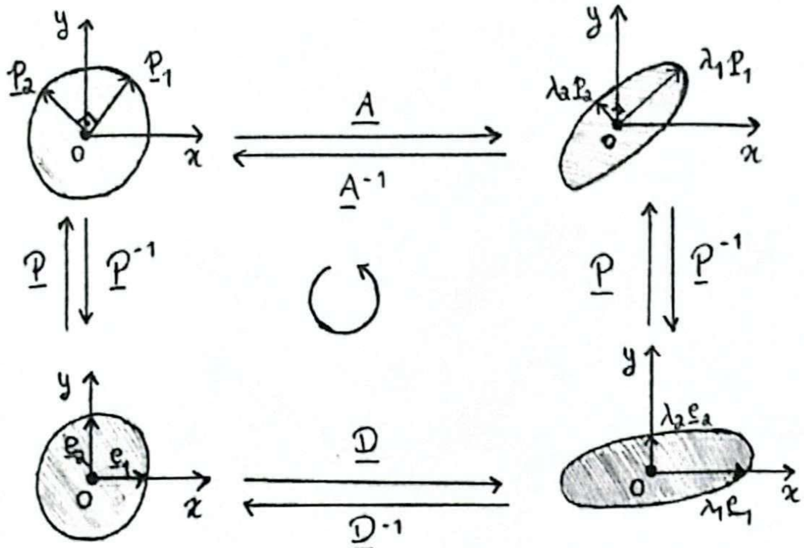
■ THEOREM[9]: A symmetric matrix $\underline{S} \in \mathbb{R}^{n \times n}$ can always be diagonalized.

↳ THEOREM[9] follows directly from the spectral theorem (THEOREM[5], p. 12). Moreover, the spectral theorem states that we can find an ONB of eigenvectors of $\mathbb{R}^n$. This makes $\underline{P}$ an orthogonal matrix so that $\underline{D} = \underline{P}^T \underline{A}\,\underline{P}$.

*) REMARK[16]: The Jordan normal form of a matrix offers a decomposition that works for defective matrices (Long, 1987) but is beyond the scope of this minicourse. ◇

2.7) Geometric Intuition for the Eigendecomposition

↳ We can interpret the eigendecomposition of a matrix as follows: Let $\underline{A}$ be the transformation matrix of a linear mapping with respect to the standard basis. The matrix $\underline{P}^{-1}$ performs a basis change from the standard basis into the eigenbasis. This identifies the eigenvectors $\underline{p}_i$ onto the standard basis vectors $\underline{e}_i$. Then, the diagonal $\underline{D}$ scales the vectors along these axes by the eigenvalues λ_i. Finally, $\underline{P}$ transforms these scaled vectors back into the standard/canonical coordinates yielding $\lambda_i \underline{p}_i$.



↳ Diagonal matrices $\underline{D}$ can efficiently be raised to a power. Therefore, we can find a matrix power for a matrix $\underline{A} \in \mathbb{R}^{n\times n}$ via the eigenvalue decomposition (if it exists) so that:

$$\underline{A}^k = (\underline{P}\,\underline{D}\,\underline{P}^{-1})^k = \underline{P}\,\underline{D}^k\underline{P}^{-1}. \tag{1.27}$$

Computing $\underline{D}^k$ is efficient because we apply this operation individually to any diagonal element. Assume that the eigendecomposition $\underline{A} = \underline{P}\,\underline{D}\,\underline{P}^{-1}$ exists. Then:

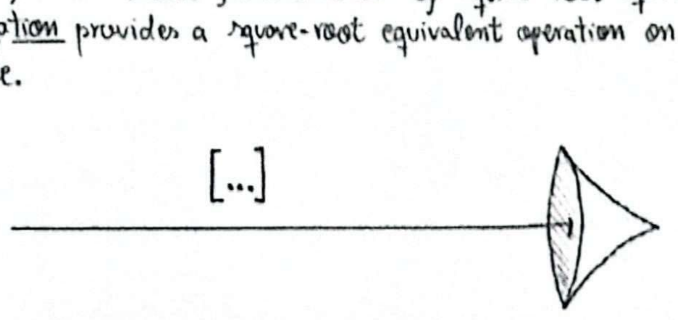
•) $$\det(\underline{A}) = \det(\underline{P}\,\underline{D}\,\underline{P}^{-1}) = \det(\underline{P})\det(\underline{D})\det(\underline{P}^{-1}) = \det(\underline{D}) = \prod_{i=1}^{n} d_{ii}, \tag{1.28}$$

allows for an efficient computation of the determinant of $\underline{A}$.

2.8) Cholesky Decomposition

↳ There are many ways to factorize special types of matrices that we encounter often in machine learning. In linear algebra, we need to be careful that we compute a square-root equivalent operation on positive quantities. For symmetric, positive and definite matrices, we can choose from a number of square-root equivalent operations. The Cholesky decomposition / Cholesky factorization provides a square-root equivalent operation on symmetric, positive and definite matrices that is useful in practice.

[...]



■ <u>THEOREM</u>[10] (Cholesky Decomposition): A symmetric, positive and definite matrix $\underline{A}$ can be factorized into a product $\underline{A} = \underline{L}\,\underline{L}^T$, where $\underline{L}$ is a lower-triangular matrix with positive diagonal elements:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & & \mathbb{O} \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} l_{11} & \cdots & l_{n1} \\ & \ddots & \vdots \\ \mathbb{O} & & l_{nn} \end{pmatrix} . \tag{1.29}$$

Here, $\underline{L}$ is called the Cholesky factor of $\underline{A}$, and $\underline{L}$ is unique.

↳ Consider a symmetric, positive and definite matrix $\underline{A} \in \mathbb{R}^{3\times3}$. We are interested in finding its Cholesky factorization given by $\underline{A} = \underline{L}\,\underline{L}^T$, i.e.:

$$\underline{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \underline{L}\,\underline{L}^T = \begin{pmatrix} l_{11} & & \mathbb{O} \\ l_{21} & l_{22} & \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ & l_{22} & l_{32} \\ \mathbb{O} & & l_{33} \end{pmatrix} . \tag{1.30}$$

Multiplying out the right-hand side yields:

$$\underline{A} = \begin{pmatrix} l_{11}^2 & l_{21}l_{11} & l_{31}l_{11} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & l_{31}l_{21} + l_{32}l_{22} \\ l_{31}l_{11} & l_{31}l_{21} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{pmatrix} . \tag{1.31}$$

Comparing the left-hand side of Eq. (1.30) and the right-hand of Eq. (1.31) shows that there is a simple pattern in the diagonal elements l_{ii}:

$$l_{11} = \sqrt{a_{11}} \quad , \quad l_{22} = \sqrt{a_{22} - l_{21}^2} \quad , \quad l_{33} = \sqrt{a_{33} - (l_{31}^2 + l_{32}^2)} . \tag{1.32}$$

Similarly for the elements below the diagonal (l_{ij}, where $i > j$), there is also a repeating pattern:

$$l_{21} = \frac{1}{l_{11}} a_{21} \quad , \quad l_{31} = \frac{1}{l_{11}} a_{31} \quad , \quad l_{32} = \frac{1}{l_{22}} (a_{32} - l_{31} l_{21}) . \tag{1.33}$$

Thus, we constructed the Cholesky decomposition for any symmetric, positive and definite 3×3 matrix. The key realization is that we can backward calculate what the components l_{ij} for the $\underline{L}$ should be, given the values a_{ij} for $\underline{A}$ and previously computed values of l_{ij}.

↳ The Cholesky decomposition is an important tool for the numerical computations underlying machine learning. Here, symmetric positive and definite matrices require frequent manipulation, e.g., the covariance matrix of a multivariate Gaussian variable is symmetric, positive and definite.

The Cholesky decomposition also allows us to compute determinants very efficiently. Given the Cholesky decomposition $\underline{A} = \underline{L}\,\underline{L}^T$, we know that $\det(\underline{A}) = \det(\underline{L})\det(\underline{L}^T) = [\det(\underline{L})]^2$. Since $\underline{L}$ is a triangular matrix, the determinant is simply the product of its diagonal entries so that $\det(\underline{A}) = \prod_{i=1}^{n} l_{ii}^2$. Thus, many numerical software packages use the Cholesky decomposition to make computations more efficient.

2.9) Singular Value Decomposition

▪ THEOREM[11] (SVD Theorem): Let $\underline{A}^{m\times n}$ be a rectangular matrix of rank $r \in [0, \min(m,n)]$. The SVD of $\underline{A}$ is a decomposition of the form:

$$\underline{A}^{m\times n} = \underline{U}^{m\times m}\,\underline{\Sigma}^{m\times n}\,(\underline{V}^{T})^{n\times n} \quad , \tag{1.34}$$

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

↪ The diagonal entries σ_i, $i=1,\dots,r$, of $\underline{\Sigma}$ are called the <u>singular values</u>, $\underline{u}_i$ are called the <u>left-singular vectors</u>, and $\underline{v}_j$ are called the <u>right-singular vectors</u>. By convention, the singular values are ordered, i.e., $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$. The <u>singular value matrix</u> $\underline{\Sigma}$ is unique, but it require some attention. Observe that the $\underline{\Sigma} \in \mathbb{R}^{m\times n}$ is rectangular. In particular, $\underline{\Sigma}$ is of the same size as $\underline{A}$. This means that $\underline{\Sigma}$ has a diagonal submatrix that contains the singular values and need additional zero padding. Specifically, if $m > n$, then the matrix $\underline{\Sigma}$ has diagonal structure up to row n and then consists of $\underline{0}^T$ row vectors from $n+1$ to m below so that:

$$\underline{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} \sigma_1 & & \mathbb{O} \\ & \ddots & \sigma_n \\ \mathbb{O} & & \\ & \mathbb{O} & \end{pmatrix} . \tag{1.35}$$

If $m < n$, the matrix $\underline{\Sigma}$ has a diagonal structure up to column m and columns that consist of $\underline{0}$ from $m+1$ to n:

$$\underline{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \sigma_m & 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} \sigma_1 & & \mathbb{O} & \\ & \ddots & & \mathbb{O} \\ \mathbb{O} & & \sigma_m & \end{pmatrix} . \tag{1.36}$$

*) REMARK[17]: The SVD exists for any matrix $\underline{A} \in \mathbb{R}^{m\times n}$.

○

2.10) Geometric Intuitions for the SVD

↪ The SVD offers geometric intuitions to describe a transformation matrix $\underline{A}$. In the following, we will discuss the SVD as sequential linear transformations performed on the bases. The SVD of a matrix can be interpreted as a decomposition of a corresponding linear mapping $\Phi: \mathbb{R}^n \to \mathbb{R}^m$ into three operations.

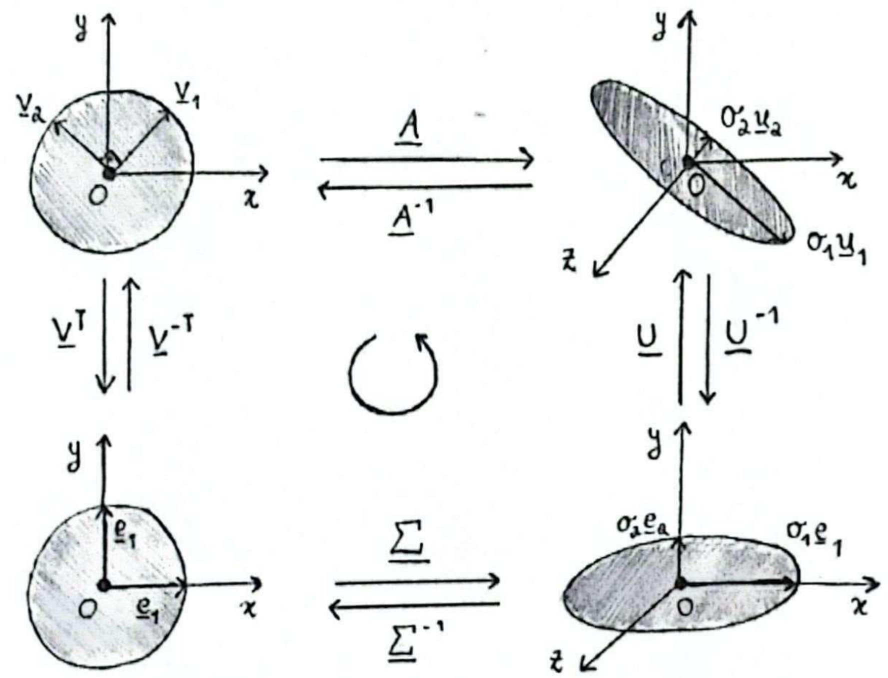
The SVD of a matrix can be interpreted as a decomposition of a corresponding linear mapping $\Phi: \mathbb{R}^n \to \mathbb{R}^n$ into three operations (see the next figure). The SVD intuition follows superficially a similar structure to our eigendecomposition intuition. Broadly speaking, the SVD performs a basis change via $\underline{V}^T$ followed by a scaling and augmentation (or reduction) in dimensionality via the singular value matrix $\underline{\Sigma}$. Finally, it performs a second basis change via $\underline{U}$.

↳ Assume we are given a 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 $\mathbb{R}^m$. Then:

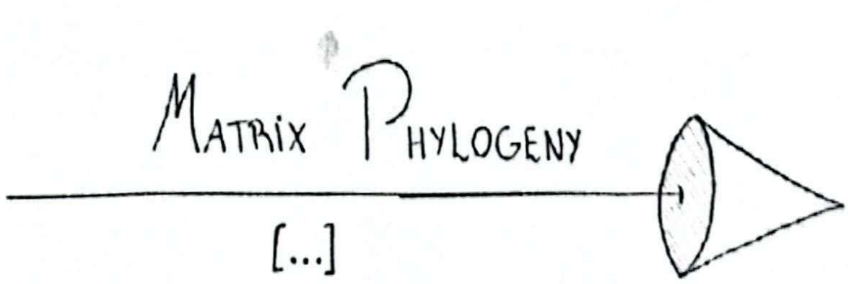
•) 1.) The matrix $\underline{V}$ performs a basis change in the domain $\mathbb{R}^n$ from $\tilde{B}$ to the standard basis B. Also, $\underline{V}^T = \underline{V}^{-1}$ performs a basis change from B to $\tilde{B}$. The vectors $\underline{v}_1$ and $\underline{v}_2$ are now aligned with the canonical basis.

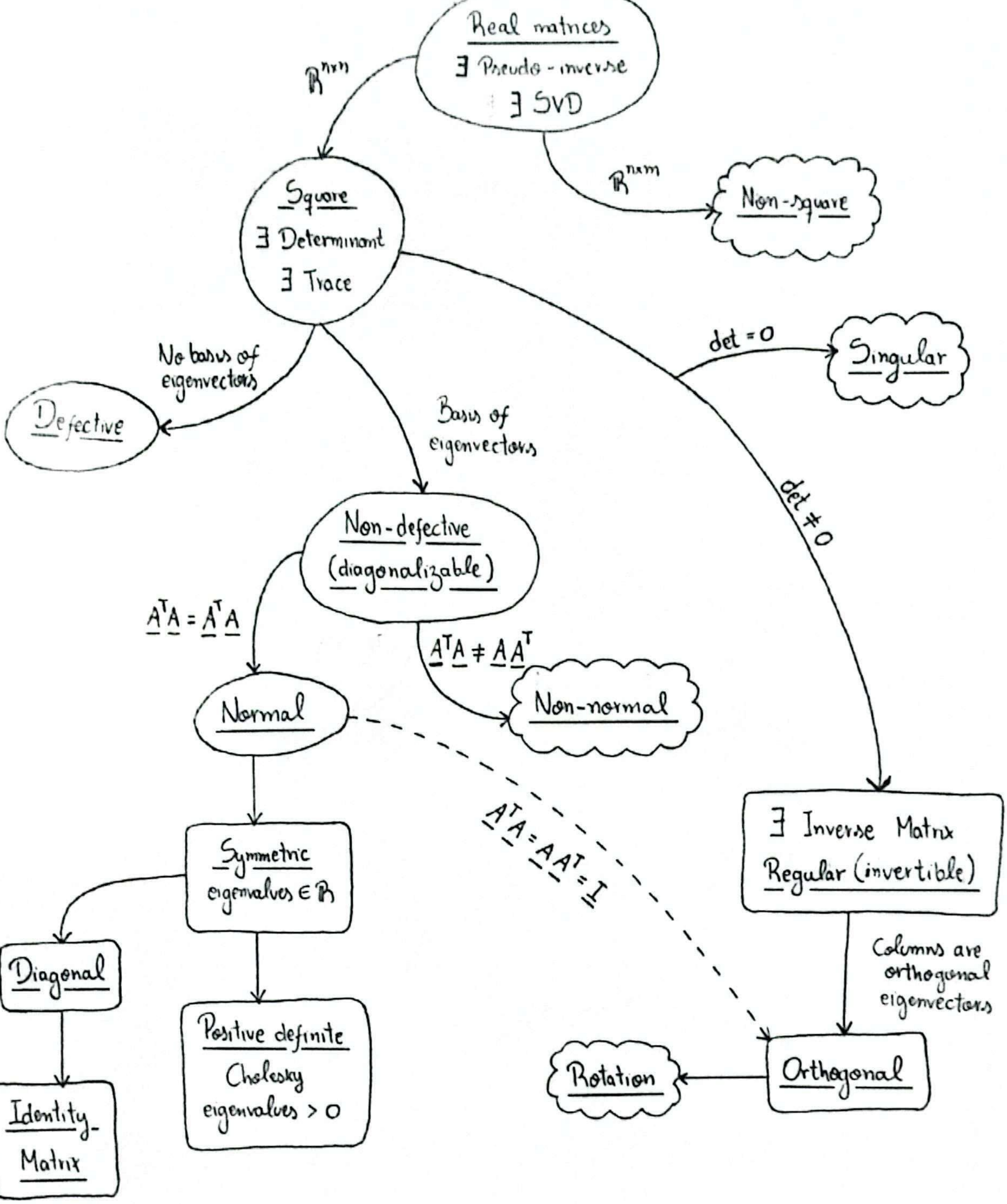
2.) Having changed the coordinate system to $\tilde{B}$, and $\underline{\Sigma}$ scales the new coordinates by the singular values σ_i (and adds or delete dimensions), i.e., $\underline{\Sigma}$ is the transformation matrix of Φ with respect to $\tilde{B}$ and $\tilde{C}$, represented by the vectors $\underline{v}_1$ and $\underline{v}_2$ being stretched and lying in the $\underline{e}_1$-$\underline{e}_2$ plane, which is now embedded in a third dimension.

3.) The matrix $\underline{U}$ performs a basis change in the codomain $\mathbb{R}^m$ of $\tilde{C}$ into the canonical basis of $\mathbb{R}^m$, represented by a rotation of the vectors $\underline{v}_1$ and $\underline{v}_2$ out of the $\underline{e}_1$-$\underline{e}_2$ plane.



↳ The SVD expresses a change of basis in both the domain and codomain. This is in contrast with eigendecomposition that operates within the same vector space, where the same basis change is applied and then undone. What makes the SVD special is that these two different bases are simultaneously linked by the singular value matrix $\underline{\Sigma}$.




Real matrices
∃ Pseudo-inverse
∃ SVD
$\mathbb{R}^{n \times n}$
$\mathbb{R}^{n \times m}$
Non-square
Square
∃ Determinant
∃ Trace
No basis of eigenvectors
Defective
Basis of eigenvectors
det = 0
Singular
det ≠ 0
Non-defective (diagonalizable)
$A^T A = A^T A$
$A^T A \neq A A^T$
Normal
Non-normal
$A^T A = A A^T = I$
∃ Inverse Matrix
Regular (invertible)
Symmetric
eigenvalues ∈ $\mathbb{R}$
Diagonal
Identity Matrix
Positive definite
Cholesky
eigenvalues > 0
Columns are orthogonal eigenvectors
Orthogonal
Rotation