# Linear Algebra Review

Kaiyu Zheng

October 2017

Linear algebra is fundamental for many areas in computer science. This document aims at providing a reference (mostly for myself) when I need to remember some concepts or examples. Instead of a collection of facts as the Matrix Cookbook, this document is more gentle like a tutorial. Most of the content come from my notes while taking the undergraduate linear algebra course (Math 308) at the University of Washington. Contents on more advanced topics are collected from reading different sources on the Internet.

## Contents

## Notation

We denote vectors using bold lower case letters such as $\boldsymbol{x}$, matrices using bold upper case letters such as $\boldsymbol{X}$, and entries of matrices using normal upper case letters such as $X_{ij}$ or $X_{i,j}$ (The comma is used if the indices are expressed by equations).

# 1 Linear System of Equations

**Definition 1.1** (Row Echelon Form)**.** Each variable can be the leading variable for at most one equation.

For example,

$$\begin{aligned} x_1 + x_2 + x_3 - x_4 &= 0 \\ -x_2 + 7x_4 - x_5 &= -1 \\ x_4 + x_5 &= 2 \end{aligned} \tag{1}$$

**Definition 1.2.** Linear systems are *equivalent* if they are related by a sequence of elementary operations:

(1) Interchange position of rows

(2) Multiply an equal constant

(3) Add a multiple of one equation to another

**Definition 1.3** (Augmented Matrix)**.** The linear system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m &= b_1, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m &= b_n \end{aligned} \tag{2}$$

can be written as an *augmented matrix* as follows:

$$\begin{bmatrix} a_{11} & \dots & a_{1m} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \dots & a_{nm} & b_n \end{bmatrix} \tag{3}$$

**Definition 1.4** (Row Echelon Form)**.** A matrix is in *row echelon form* if

a) Every leading term is in a column to the left of the leading term of the row below it.

b) Any zero rows are at the bottom of the matrix

For example, the left matrix below is not an echelon form, because "0=7" has no leading variable. It is an *inconsistent* matrix. The right matrix is a echelon form.

$$\begin{bmatrix} 1 & 2 & 3 & 0 & 0 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 7 \end{bmatrix} \quad \begin{bmatrix} 1 & -2 & 5 & 2 & -1 \\ 0 & 3 & 4 & 5 & 6 \\ 0 & 0 & 22 & 14 & 4 \end{bmatrix}$$

The leading variable positions in the matrix are called *pivot positions*. A column in the matrix that contains a pivot position is a *pivot column*. The process of converting a linear system into echelon form is *Gaussian Elimination*.

**Definition 1.5** (Reduced Row Echelon Form)**.** A matrix is said to be in *reduced row echelon form* if:

a) all pivot positions have 1

b) the only nonzero term in each pivot column is the pivot

c) it is in row echelon form.

Try finding the reduced row echelon form of the following matrix:

$$\begin{bmatrix} 0 & 3 & 4 & 5 & 6 \\ 1 & -2 & 5 & 2 & -1 \\ 3 & 0 & 1 & 2 & 5 \end{bmatrix} \tag{4}$$

**Definition 1.6** (Homogeneity)**.** A *homogeneous linear equation* is

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0 \tag{5}$$

The equation is said to be in homogeneous form. A linear system where all equations are in homogeneous form is a *homogenous system*.

Every homogenous system is *consistent*, i.e. solvable.

## 2 Vectors

**Definition 2.1** (Norm)**.** The *norm*, or magnitude of a vector $\boldsymbol{a} \in \mathbb{R}^n$ is defined as the *L2-norm* of the vector.

$$|\boldsymbol{a}| = \sqrt{\sum_{i=1}^{n} a_i^2} \tag{6}$$

**Definition 2.2** (Dot Product)**.** *(Algebraic definition)* Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^n$. Then the dot product (or inner product) between $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as:

$$\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}^T \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i \tag{7}$$

*(Geometric definition)* The dot product of two Euclidean vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined by

$$\boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}||\boldsymbol{b}|cos(\theta_{\boldsymbol{a},\boldsymbol{b}}) \tag{8}$$

Also, The dot product $\boldsymbol{w} \cdot \boldsymbol{x} = b$ is a hyperplane, where $\boldsymbol{w}$ is normal to it.

**Definition 2.3** (Projection)**.** Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^n$. The projection of $\boldsymbol{b}$ onto $\boldsymbol{a}$ is defined

$$proj_{\boldsymbol{a}}\boldsymbol{b} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}|} \frac{\boldsymbol{a}}{|\boldsymbol{a}|} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}|^2} \boldsymbol{a} \tag{9}$$

**Definition 2.4** (Outer Product)**.** Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^n$. Then the outer product (or tensor product) between $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined such that $(\boldsymbol{a}\boldsymbol{b}^T)_{ij} = a_i b_j$:

$$\boldsymbol{a}\boldsymbol{b}^T = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_n \\ a_2 b_1 & a_2 b_2 & \cdots & a_2 b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n b_1 & a_n b_2 & \cdots & a_n b_n \end{bmatrix} \tag{10}$$

**Definition 2.5** (Linear Combination)**.** If $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots \boldsymbol{u}_m$ are vectors and $c_1, c_2 \cdots c_m$ are scalars, then $c_1\boldsymbol{u}_1 + c_2\boldsymbol{u}_2 + \cdots + c_m\boldsymbol{u}_m$ is a linear combination of the vectors.

**Definition 2.6** (Span)**.** Let $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ be a set of $m$ vectors in $\mathbb{R}^n$. The *span* of the set is the set of linear combinations of $\boldsymbol{u}_1 \cdots \boldsymbol{u}_m$.

For example, suppose $\boldsymbol{u}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ and $\boldsymbol{u}_2 = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$, what is the span of $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$? A vector $\boldsymbol{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \in \text{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ if and only if $\exists s, t \textbf{.} s\boldsymbol{u}_1 + t\boldsymbol{u}_2 = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$. $s, t$ exist if $\begin{bmatrix} 1 & 3 & a \\ 2 & 2 & b \\ 3 & 1 & c \end{bmatrix}$ has a solution. This matrix is reduced to $\begin{bmatrix} 1 & 3 & a \\ 0 & 4 & 2a - b \\ 0 & 0 & a - 2b + c \end{bmatrix}$, therefore it has a solution when $a - ab + c = 0$ holds. So the *span* of $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ is the plane $x - 2y + z = 0$.

**Definition 2.7** (Relation of Span and Augmented Matrix)**.** If a vector $\boldsymbol{v}$ is in the span of vectors $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ then the matrix $[\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_m \ \boldsymbol{v}]$ has at least 1 solution.

**Theorem 2.1** (Relation of Span and Linearly Independence)**.** *If* $\boldsymbol{u} \in \text{span}\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ *then* $\text{span}\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\} = \text{span}\{\boldsymbol{u}, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$

## 2.1 Linear independence

**Definition 2.8** (Linear Independence)**.** Let $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ be a set of vectors in $\mathbb{R}^n$. If the only solution to the equation $x_1\boldsymbol{u}_1 + \cdots + x_m\boldsymbol{u}_m = 0$ is the trivial solution (i.e. all zeros), then $\boldsymbol{u}_1 cdots \boldsymbol{u}_m$ are *linearly independent.*

Fact: If any set of vector contains $\mathbf{0}$, this set of vectors are not linearly independent.

**Definition 2.9** (Orthonormal Vectors)**.** Vectors in a set $\mathcal{U} = \{\boldsymbol{u}_1, \cdots \boldsymbol{u}_m\}$ are *orthonormal* if every vector in $\mathcal{U}$ is a unit vector and every pair $\boldsymbol{u_i}, \boldsymbol{u_j} \in \mathcal{U}$ of vectors are orthogonal, i.e. $\boldsymbol{u_i}^T\boldsymbol{u_j} = 0$.

**Theorem 2.2.** *Every set of orthonormal vectors is linearly independent (i.e. the vectors in the set are linearly independent).*

## 2.2 Linear dependence

**Theorem 2.3** (Linear Dependence)**.** *Let $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ be a set of vectors in $\mathbb{R}^n$. If $n < m$, the set is* linearly dependent*.*

**Corollary 2.3.1** (Relation of Span and Linearly Independence)**.** *If there is a set of $m$* linearly independent vectors *in $\mathbb{R}^n$ that spans all of $\mathbb{R}^n$, then $m = n$.*

**Theorem 2.4** (Relation of Linear Combination and Linearly Dependence)**.** *Let $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ be a set of vectors in $\mathbb{R}^n$. The vectors in this set are* linearly dependent *if one vector is a linear combination of others.*

## 2.3 Linear transformation

**Definition 2.10** (Linear Transformation)**.** Function $T : \mathbb{R}^m \to \mathbb{R}^n$ is a *linear transformation* if for all $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^m$ and for all $r \in \mathbb{R}$, $T(\boldsymbol{v} + \boldsymbol{u}) = T\boldsymbol{v} + T\boldsymbol{u}$ and $T(r\boldsymbol{v}) = rT(\boldsymbol{v})$. $\mathbb{R}^m$ is the *domain*, and $\mathbb{R}^n$ is the *co-domain*. For $\boldsymbol{u} \in \mathbb{R}^m$, $T(\boldsymbol{u})$ is the *image* of $\boldsymbol{u}$ under $T$.

**Definition 2.11** (Subspace)**.** A subset $S$ of $\mathbb{R}^n$ is a *subspace* if $S$ satisfies:

a) $S$ contains $\mathbf{0}$.

b) if $\boldsymbol{u}$ and $\boldsymbol{v}$ are in $S$ then $\boldsymbol{u} + \boldsymbol{v}$ is also in $S$. (*closure under addition*)

c) If $r$ is a real number, and $\boldsymbol{u} \in S$ then, $r\boldsymbol{u} \in S$. (*closure under multiplication*)

**Definition 2.12** (One-to-one and On-to)**.** Let $T : \mathbb{R}^m \to \mathbb{R}^n$, $T(\boldsymbol{v}) = \boldsymbol{A}\boldsymbol{v}$ thus $T$ is a linear transformation. $T$ is *one-to-one* (injective) if and only if $T(\boldsymbol{x}) = \mathbf{0}$ has only the trivial solution (i.e. $\boldsymbol{x} = \mathbf{0}$), or equivalently, $T(\boldsymbol{a}) = T(\boldsymbol{b})$ implies $\boldsymbol{a} = \boldsymbol{b}$. This means the columns of $\boldsymbol{A}$ are linearly independent. $T$ is *on-to* (surjective) if and only if columns of $\boldsymbol{A}$ span $\mathbb{R}^n$.

Note, $\boldsymbol{A}$ is a $n \times m$ matrix. If $m > n$, $T$ is *not* one-to-one. If $m < n$, $T$ is *not* on-to.

In more general terms, if a function is one-to-one (**injective**), every element of the co-domain is mapped to by *at most one* element of the domain. If a function is on-to (**surjective**) if every element of the co-domain is mapped to by at least one element of the domain. A function is *one-to-one and on-to* (**bijective**) if every element of the co-domain is mapped to by exactly one element of the domain.

# 3 Matrix Algebra

## 3.1 Addition

If $\boldsymbol{A}, \boldsymbol{B} \in M_{n\times m}(\mathbb{R})$ and $r \in \mathbb{R}$,

$$(\boldsymbol{A}+\boldsymbol{B})_{ij} = (\boldsymbol{A})_{ij} + (\boldsymbol{B})_{ij} \tag{11}$$

## 3.2 Scalar Multiplication

$$(r\boldsymbol{A})_{ij} = r(\boldsymbol{A})_{ij} \tag{12}$$

## 3.3 Matrix Multiplication

If $T : \mathbb{R}^m \to \mathbb{R}^n$ is represented by $\boldsymbol{A} \in M_{n\times m}(\mathbb{R})$ and $W : \mathbb{R}^n \to \mathbb{R}^l$ is represented by $\boldsymbol{B} \in M_{l\times m}(\mathbb{R})$, then $\boldsymbol{BA}$ should be represented as $W \circ T : \mathbb{R}^m \to \mathbb{R}^l$. So $\boldsymbol{BA} \in M_{l\times m}(\mathbb{R})$.

Matrix multiplication can be thought of as applying a series of linear transformation to vectors in an initial domain. For example, $\boldsymbol{BA}$ is illustrated as

$$\mathbb{R}^m \xrightarrow{T} \mathbb{R}^n \xrightarrow{W} \mathbb{R}^l$$

Notice that although the final transformation is $\mathbb{R}^m \to \mathbb{R}^l$ which reads "a transformation going from $\mathbb{R}^m$ (domain of $T$) to $\mathbb{R}^l$ (codomain of $W$)", the formal notation is "reversed", which is $W \circ T$.

<u>Alternative definition:</u> Let $\boldsymbol{A} \in M_{n\times p}(\mathbb{R})$ and $\boldsymbol{B} \in M_{p\times m}(\mathbb{R})$, then $\boldsymbol{AB} \in M_{n\times m}(\mathbb{R})$. We will look at several equivalent algebraic definitions of $\boldsymbol{AB}$ from different perspectives. But first of all, let us look at two interpretations of *matrix-vector multiplication* $\boldsymbol{Ax}$ where $\boldsymbol{x} \in \mathbb{R}^p$.

1) We consider $\boldsymbol{Ax}$ from the perspective of considering *row vectors* of $\boldsymbol{A}$, that is, we view $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{bmatrix} \underline{\boldsymbol{a}}_1 \\ \underline{\boldsymbol{a}}_2 \\ \vdots \\ \underline{\boldsymbol{a}}_n \end{bmatrix} \tag{13}$$

where each component $\underline{\boldsymbol{a}}_{\boldsymbol{i}}$ is a row vector. Then, $\boldsymbol{Ax}$ can be computed by performing dot product $\underline{\boldsymbol{a}}_{\boldsymbol{i}}^T \boldsymbol{x}$ for $i \in \{1, \cdots, n\}$, therefore $(\boldsymbol{Ax})_i = \underline{\boldsymbol{a}}_{\boldsymbol{i}}^T \boldsymbol{x}$. Specifically,

$$\boldsymbol{Ax} = \begin{bmatrix} \underline{\boldsymbol{a}}_1^T \boldsymbol{x} \\ \underline{\boldsymbol{a}}_2^T \boldsymbol{x} \\ \vdots \\ \underline{\boldsymbol{a}}_n^T \boldsymbol{x} \end{bmatrix} \tag{14}$$

2) We can also compute $\boldsymbol{Ax}$ by considering *column vectors* of $\boldsymbol{A}$, such that

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_{|1} & \boldsymbol{a}_{|2} & \cdots & \boldsymbol{a}_{|p} \end{bmatrix} \tag{15}$$

where each component $\boldsymbol{a}_{|i}$ is a column vector. Then, the matrix multiplication $\boldsymbol{Ax}$ can be viewed as a linear combination of columns of $\boldsymbol{A}$ with coefficients determined by entries $x_i$ for $i \in \{1, \cdots, k\}$.

$$\begin{aligned} \boldsymbol{Ax} &= x_1 \boldsymbol{a}_{|1} + x_2 \boldsymbol{a}_{|2} + \cdots + x_n \boldsymbol{a}_{|p} \\ &= \sum_{i=1}^{p} \boldsymbol{a}_{|i} x_i \end{aligned} \tag{16}$$

Now, let us look at *matrix-matrix multiplication* also from two perspectives.

1) When we consider row vectors of $\boldsymbol{A}$ and column vectors of $\boldsymbol{B}$, the multiplication $\boldsymbol{AB}$ can be viewed as

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{Ab}_{|1} & \boldsymbol{Ab}_{|2} & \cdots & \boldsymbol{Ab}_{|m} \end{bmatrix} \tag{17}$$

where $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_{|1} & \boldsymbol{b}_{|2} & \cdots & \boldsymbol{b}_{|m} \end{bmatrix}$. From Equation 14, we know $(\boldsymbol{Ab}_{|k})_i = \underline{\boldsymbol{a}}_i^T \boldsymbol{b}_{|k}$. Therefore, $(\boldsymbol{AB})_{ij} = \underline{\boldsymbol{a}}_i^T \boldsymbol{b}_{|j}$.

2) When we consider column vectors of $\boldsymbol{A}$ and row vectors of $\boldsymbol{B}$, the multiplication $\boldsymbol{AB}$ can be viewed as

$$\boldsymbol{AB} = \sum_{i=1}^{p} \boldsymbol{a}_{|i} \underline{\boldsymbol{b}}_i^T \tag{18}$$

where $\boldsymbol{a}_{|i} \underline{\boldsymbol{b}}_i^T$ is the *outer product* with output dimension of $n \times m$.

<u>*Properties of Matrix Multiplication*</u>:

1) $\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$

2) $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$

3) $(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$

4) $s\boldsymbol{AB} = \boldsymbol{A}s\boldsymbol{B}$

5) $\boldsymbol{IA} = \boldsymbol{AI} = \boldsymbol{A}$

*Caveats*:

1) $\boldsymbol{AB} \neq (\boldsymbol{BA})$ (usually)

2) $\boldsymbol{AC} = \boldsymbol{AB} \not\Rightarrow \boldsymbol{C} = \boldsymbol{B}$

## 3.4 Transpose

If $\boldsymbol{A} \in M_{n\times m}(\mathbb{R})$, then $\boldsymbol{A}^T \in M_{m\times n}(\mathbb{R})$.

*Properties of Transpose*:

1) $(\boldsymbol{A}+\boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$

2) $(s\boldsymbol{A})^T = s(\boldsymbol{A}^T)$

3) $(\boldsymbol{AC})^T = \boldsymbol{C}^T\boldsymbol{A}^T$

**Theorem 3.1.** *A matrix $\boldsymbol{A}$ has the property that for all $\boldsymbol{v}, \boldsymbol{W} \in \mathbb{R}^2$, $\boldsymbol{v}\cdot\boldsymbol{w} = \boldsymbol{Av}\cdot\boldsymbol{Aw}$ if and only if $\boldsymbol{A}$ is* orthogonal*, that is, $\boldsymbol{AA}^T = \boldsymbol{I}_n$, or equivalently, $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$.*

### 3.4.1 Conjugate Transpose

## 3.5 Inverse

**Definition 3.1** (Invertibility)**.** A linear map $T : \mathbb{R}^m \to \mathbb{R}^n$ is *invertible* if it is one-to-one and on-to. Two implications follows if $T : \mathbb{R}^m \to \mathbb{R}^n$ is invertible:

1) $m = n$ (required)

2) $T^{-1}$ is also linear.

**Theorem 3.2** (Invert of Matrix)**.** *An $n \times n$ matrix $\boldsymbol{A}$ is invertible if there exists a matrix $\boldsymbol{B}$ so that $\boldsymbol{BA} = \boldsymbol{I}_n$. If $\boldsymbol{A}$ is invertible, $\boldsymbol{B}$ is* unique *and define $\boldsymbol{A}^{-1} = \boldsymbol{B}$.*

$$\implies \boldsymbol{BA} = \boldsymbol{AB} = \boldsymbol{I}_n$$

To compute $\boldsymbol{A}^{-1}$, form an $n \times 2n$ matrix $\begin{bmatrix}\boldsymbol{A} & \boldsymbol{I}_n\end{bmatrix}$. Then convert it to reduced row echelon form, which results in $\begin{bmatrix}\boldsymbol{I}_n & \boldsymbol{A}^{-1}\end{bmatrix}$.

**Theorem 3.3** (Invertibility Implies Non-zero Determinant)**.** *An $n \times n$ matrix is invertible if and only if its* determinant *is not zero.*

**Theorem 3.4** (Invertibility and Positive-Definite)**.**

*Any* positive-definite *matrix is invertible.*

*Properties of Matrix Inverse*:
If $\boldsymbol{A}, \boldsymbol{B}$ are invertible $n \times n$ matrix, and $\boldsymbol{C}$, $\boldsymbol{D}$ are $n \times m$ matrix. Then:

a) $\boldsymbol{A}^{-1}$ is invertible. $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$

b) $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$

c) $\boldsymbol{AB}$ is invertible. $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$

d) If $\boldsymbol{AC} = \boldsymbol{AD}$, then $\boldsymbol{C} = \boldsymbol{D}$

e) If $\boldsymbol{AC} = \boldsymbol{0}$, then $\boldsymbol{C} = \boldsymbol{0}$

f) $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$

*Proof.* We will prove c) and d).

c) Show that $\boldsymbol{AB}(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}) = \boldsymbol{I}_n$:

$$\boldsymbol{AB}(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}) = \boldsymbol{I}_n = \boldsymbol{A}(\boldsymbol{BB}^{-1})\boldsymbol{A}^{-1} = \boldsymbol{AI}_n\boldsymbol{A}^{-1} = \boldsymbol{AA}^{-1} = \boldsymbol{I}_n \tag{19}$$

d) Show that $\boldsymbol{AC} = \boldsymbol{AD} \implies \boldsymbol{C} = \boldsymbol{D}$:

$$\boldsymbol{AC} = \boldsymbol{AD} \tag{20}$$
$$\implies \boldsymbol{A}^{-1}\boldsymbol{AC} = \boldsymbol{A}^{-1}\boldsymbol{AD} \tag{21}$$
$$\implies \boldsymbol{I_n C} = \boldsymbol{I_n D} \tag{22}$$
$$\implies \boldsymbol{C} = \boldsymbol{D} \tag{23}$$

From the above proof, we see that $\boldsymbol{A}$ being invertible is important, because otherwise $\boldsymbol{A}^{-1}$ does not exist.

□

### 3.6 Trace

**Definition 3.2** (Trace)**.** Let $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$. The trace of $\boldsymbol{A}$ is defined as the sum of entries along the main diagonal:

$$tr(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii} \tag{24}$$

*Properties of Trace*:

a) $tr(\boldsymbol{A}+\boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$

b) $tr(c\boldsymbol{A}) = c \cdot tr(\boldsymbol{A})$

c) $tr(\boldsymbol{A}\boldsymbol{B}) = tr(\boldsymbol{B}\boldsymbol{A})$

d) $tr(\boldsymbol{A}) = tr(\boldsymbol{A}^T)$

e) $tr(\boldsymbol{X}^T\boldsymbol{Y}) = tr(\boldsymbol{X}\boldsymbol{Y}^T) = tr(\boldsymbol{Y}^T\boldsymbol{X}) = \sum_{ij} X_{ij}Y_{ij}$

f) Similarity-invariant:

$$tr(\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}) = tr(\boldsymbol{P}^{-1}(\boldsymbol{A}\boldsymbol{P})) = tr((\boldsymbol{A}\boldsymbol{P})\boldsymbol{P}^{-1}) = tr(\boldsymbol{A}(\boldsymbol{P}\boldsymbol{P}^{-1})) = tr(\boldsymbol{A})$$

g) $d\ tr(\boldsymbol{X}) = tr(d\boldsymbol{X})$

*Trace and Eigenvalues*: In Section 5, we discuss eigenvectors and eigenvalues in more detail. For the sake of proximity, we describe the relation of trace and eigenvalues here.

**Theorem 3.5.** *If $\boldsymbol{A}$ is an $n \times n$ matrix with real or complex entries and if $\lambda_1, \cdots, \lambda_n$ are eigenvalues of $\boldsymbol{A}$, then*

$$tr(\boldsymbol{A}) = \sum_i \lambda_i \tag{25}$$

$$tr(\boldsymbol{A}^k) = \sum_i \lambda_i^k \tag{26}$$

### 3.7 Power

**Definition 3.3** (Integral Power). $\boldsymbol{A}^n$ is raising matrix $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$ to the power of $n$. It is defined as the multiplication of $n$ the same matrix $\boldsymbol{A}$:

$$\boldsymbol{A}^n = \boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A} \tag{27}$$

The matrix to the 0th power is defined to be the identity matrix, i.e. $\boldsymbol{A}^0 = \boldsymbol{I}$. The exponentiation of a non-square matrix is not well-defined; One reason is that the 0th power is undefined. Note that $\boldsymbol{A}^{-1} \neq 1/\boldsymbol{A}$, as it is the matrix inverse.

**Definition 3.4** (Square Root). Matrix $\boldsymbol{B} = \boldsymbol{A}^{1/2}$ if and only if $\boldsymbol{B}\boldsymbol{B} = \boldsymbol{A}$.

To compute the square root of an arbitrary square matrix, a method that involves Jordan Normal Form (Section **??**) can be used. We discuss the case when the matrix $\boldsymbol{A}$ is diagonalizable (Section 7.4), meaning there exist matrix $\boldsymbol{V}$ and diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1}$. The square root of $\boldsymbol{A}$ is $\boldsymbol{R}$ such that:

$$\boldsymbol{R} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^{-1} \tag{28}$$

where $\boldsymbol{S}$ is *any* square root of $\boldsymbol{D}$. To verify,

$$\boldsymbol{RR} = \boldsymbol{VS}(\boldsymbol{V}^{-1}\boldsymbol{V})\boldsymbol{SV}^{-1} = \boldsymbol{VSSV}^{-1} = \boldsymbol{VDV}^{-1} = \boldsymbol{A} \tag{29}$$

The square root of $\boldsymbol{D}$ is simply obtained by taking the square root of all entries along the diagonal. To raise a matrix $\boldsymbol{A}$ to an arbitrary real value $p$, we can follow

$$\boldsymbol{A}^p = \exp(p \ln(\boldsymbol{A})) \tag{30}$$

where $\ln(\boldsymbol{A})$ is defined in Section 3.8 below.

### 3.8 Exponential and Logarithm

**Definition 3.5** (Exponential of Matrix)**.** The exponential of matrix $\boldsymbol{A}$ is defined as

$$e^{\boldsymbol{A}} = \sum_{n=0}^{\infty} \frac{\boldsymbol{A}}{n!} \tag{31}$$

This is a generalization of ordinary exponential function $e^x$ which is

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \tag{32}$$

**Definition 3.6** (Logarithm of Matrix)**.** Matrix $\boldsymbol{B}$ is the logarithm of matrix $\boldsymbol{A}$ if

$$\ln(\boldsymbol{A}) = \boldsymbol{B} \tag{33}$$

which is equivalent as $e^{\boldsymbol{B}} = \boldsymbol{A}$.

The logarithm of $\boldsymbol{A}$ does not always exist; At least, $\boldsymbol{A}$ needs to be invertible, but this is not enough. For more, please refer to Wikipedia.

### 3.9 Conversion Between Matrix Notation and Summation

**Outer products** Suppose $\boldsymbol{x_i} \in \mathbb{R}^d$, and $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}]^T$. Then,

$$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T = \boldsymbol{X}^T \boldsymbol{X} \tag{34}$$

To understand this intuitively, note that the vertical vectors $\boldsymbol{x_i}$ are rows of $\boldsymbol{X}$. Then, recall from Equation 18, matrix multiplication $\boldsymbol{AB}$ can be viewed as the sum of outer products between column vectors of $\boldsymbol{A}$ and row vectors of $\boldsymbol{B}$. Therefore, we need to transpose $\boldsymbol{X}$ and multiply it by itself, yielding $\boldsymbol{X}^T\boldsymbol{X}$.

Similarly, if $\boldsymbol{y} \in \mathbb{R}^s$, and $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n]^T$, we have:

$$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{y}_i^T = \boldsymbol{X}^T \boldsymbol{Y} \tag{35}$$

*Examples*:

- Conversion from primal objective to dual objective for *kernel ridge regression.* In ridge regression, with $\boldsymbol{X} \in \mathbb{R}^{N \times d}$, $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{x}_i \in \mathbb{R}^d$ features each we can formulate the objective as:

$$\min_{\boldsymbol{w}} \frac{1}{N} \sum_{i=1}^{N} \left( y_i - \boldsymbol{w}^T \boldsymbol{x}_i \right)^2 + \lambda \boldsymbol{w}^T \boldsymbol{w} \tag{36}$$

According to the Representer Theorem, $\boldsymbol{w}^* = \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i$ is the optimal weights. Thus, with $\boldsymbol{\alpha} \in \mathbb{R}^N$, the above can be transformed into the following (kernel ridge regression objective), where $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$ is the kernel function:

$$\min_{\boldsymbol{\alpha}} \frac{1}{N} \sum_{i=1}^{N} \left( y_i - \sum_{j=1}^{N} \alpha_j \boldsymbol{x}_j^T \boldsymbol{x}_i \right)^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{37}$$

$$\Leftrightarrow \min_{\boldsymbol{\alpha}} \frac{1}{N} \sum_{i=1}^{N} \left( y_i - \sum_{j=1}^{N} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}_i) \right)^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{38}$$

To transform Equation 38 into matrix notation, first let $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ be the kernel matrix where $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Then, we have:

$$\Leftrightarrow \min_{\boldsymbol{\alpha}} \frac{1}{N} (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha})^T (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \tag{39}$$

$$\Leftrightarrow \min_{\boldsymbol{\alpha}} \frac{1}{N} (\boldsymbol{\alpha}^T \boldsymbol{K}^T \boldsymbol{K} \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \boldsymbol{K}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{K} \boldsymbol{\alpha} + \boldsymbol{y}^T \boldsymbol{y}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \tag{40}$$

Because $\boldsymbol{\alpha}^T \boldsymbol{K}^T \boldsymbol{y}$ and $\boldsymbol{y}^T \boldsymbol{K} \boldsymbol{\alpha}$ are just scalars, we can just write:

$$\Leftrightarrow \min_{\boldsymbol{\alpha}} \frac{1}{N} (\boldsymbol{\alpha}^T \boldsymbol{K}^T \boldsymbol{K} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T \boldsymbol{K}^T \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{y}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \tag{41}$$

The matrix notation conversion of the ridge regularization term is important.

**Singular value decomposition** Given matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$, how do you write its singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ using summation notation?

# 4 Vector Spaces

**Definition 4.1** (Vector Space)**.** A *vector space* $\mathcal{V}$ over a field, such as real numbers $\mathbb{R}$, is a set $\mathcal{V}$ with two functions:

$$\text{addition } + : V \times V \to V \qquad (\text{e.g. } \boldsymbol{v} + \boldsymbol{w}) \tag{42}$$

$$\text{scalar multiplication } \cdot : \mathbb{R} \times V \to V \qquad (\text{e.g. } a\boldsymbol{v},\ a \in \mathbb{R}) \tag{43}$$

and satisfy these properties (*axioms* for all $\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{u} \in \mathcal{V}$ and $s, t \in \mathbb{R}$:

1) $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$ (Associativity of addition)

2) $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$ (Commutativity of addition)

3) There exists an element $\mathbf{0} \in \mathcal{V}$, called the zero vector, such that $\boldsymbol{v} + \mathbf{0} = \boldsymbol{v} forall \boldsymbol{v} \in \mathcal{V}$. (Identity element of addition)

4) $\cdots$ For more, refer to the `Wikipedia's article on vector space`.

**Definition 4.2** (Subspace)**.** A *linear subspace* is a subset of $\mathbb{R}^n$ that is a vector space with the induced multiplication and addition from $\mathbb{R}^n$.

For example, $\mathcal{S} \in \mathbb{R}^n$ is a vector subspace if for all $\boldsymbol{v}, \boldsymbol{w} \in \mathcal{S}$, $\boldsymbol{v} + \boldsymbol{w} \in \mathcal{S}$, and for all $r \in \mathbb{R}$, $\boldsymbol{v} \in \mathcal{S}$, $r\boldsymbol{v} \in \mathcal{S}$. The latter implies $\mathbf{0} \in \mathcal{S}$.

$\left\{ \begin{bmatrix} a \\ b \\ 1 \end{bmatrix} \in \mathbb{R}^3, a, b \in \mathbb{R} \right\}$ is *not* a subspace.

**Definition 4.3** (Null space)**.** If $\boldsymbol{A}$ is an $n \times n$ matrix, the set of solutions to the system $\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$ is a subspace of $\mathbb{R}^n$, called the *null space* of $\boldsymbol{A}$ or *null($\boldsymbol{A}$)*.

*Proof.* Suppose $\boldsymbol{v}$, $\boldsymbol{w}$ are vectors in $\mathbb{R}^n$ that satisfy $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{A}\boldsymbol{w} = \mathbf{0}$. Then $\boldsymbol{A}(\boldsymbol{v} + \boldsymbol{w}) = \boldsymbol{A}\boldsymbol{v} + \boldsymbol{A}\boldsymbol{w} = \mathbf{0}$. And $\boldsymbol{A}(r\boldsymbol{v}) = r\boldsymbol{A}\boldsymbol{v} = \mathbf{0}$. Therefore, the set of solutions to $\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$ is closed both under addition and multiplication. □

## 4.1 Determinant

Before we formally define determinants, let us use $det(\boldsymbol{A})$ to refer to the determinant of matrix $\boldsymbol{A}$, which is a real value.

**Definition 4.4.** (Determinant and Minor) If $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, define $\boldsymbol{M}_{ij}$ as the $n-1 \times n-1$ matrix formed by deleting the $i$-th row and $j$-th column. $\boldsymbol{A}$. $det(\boldsymbol{M}_{ij})$ is called the *minor* of entry $a_{ij}$ in $\boldsymbol{A}$.

**Definition 4.5** (Cofactor)**.** If $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, the *cofactor* of $a_{ij}$, or $C_{ij} = (-1)^{i+j} det(\boldsymbol{M}_{ij})$.

**Definition 4.6** (Singularity)**.** A square matrix $\boldsymbol{A}$ that is invertible is called *nonsingular*. Otherwise, it is called *singular* or *degenerate*.

**Theorem 4.1** (Singularity and Determinant)**.** *A square matrix is* singular *if and only if its determinant is 0.*

Now, we formally introduce determinant of a matrix.

**Definition 4.7** (Determinant)**.** The determinant of $\boldsymbol{A}$ is an $n \times n$ matrix

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$$

The determinant of $\boldsymbol{A}$ is recursively defined as:

$$det(\boldsymbol{A}) = |\boldsymbol{A}| = a_{11}C_{11} + a_{12}C_{12} + \cdots + a_{1n}C_{1n} \tag{44}$$

And when $n = 1$, $det(a_{11}) = a_{11}$ (base case).

The above definition is recursive because the definition of cofactor contains determinant.

**Geometric Meaning of Determinants** First, we focus on 2D. Suppose

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \boldsymbol{x}_1 = \begin{bmatrix} a \\ c \end{bmatrix}, \boldsymbol{x}_2 = \begin{bmatrix} b \\ d \end{bmatrix}$$

We have $det(\boldsymbol{A}) = ad - bc$. This is the *signed* area of the parallelogram formed by vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. In the 3D case, the determinant represents the signed volume of the hexahedron formed by the three column vectors in the matrix.

**Theorem 4.2** (Invertibility and Determinant)**.** *For* $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, *it is invertible if and only if* $det(\boldsymbol{A} \neq 0$.

In other words, the determinant of an $n$ by $n$ matrix $\boldsymbol{A}$ is 0 if and only if the rows are linearly dependent (and not zero if and only if they are linearly independent).

*Properties of Determinants*:

a) The determinant equals to the product of eigenvalues $\lambda_i$:

$$det(\boldsymbol{A}) = \prod_i \lambda_i$$

b) $det(c\boldsymbol{A}) = c^n \cdot det(\boldsymbol{A})$

c) $det(\boldsymbol{AB}) = det(\boldsymbol{A})det(\boldsymbol{B})$

d) $det(\boldsymbol{A}^{-1}) = \dfrac{1}{det(\boldsymbol{A})}$

e) $det(\boldsymbol{A}^T) = det(\boldsymbol{A})$

f) $det(\boldsymbol{A}^n) = det(\boldsymbol{A})^n$

*Cool Facts about Determinants*[1]:

1) Interchanging any two rows of an $n$ by $n$ matrix $\boldsymbol{A}$ reverses the sign of its determinant.

2) If two rows of a matrix are equal, its determinant is 0. (Because $det(\boldsymbol{A}) = -det(\boldsymbol{A})$ implies $det(\boldsymbol{A}) = 0$.

3) If $\boldsymbol{A}$ is an $n$ by $n$ matrix, adding a multiple of one row to a different row does not affect its determinant.

4) An $n$ by $n$ matrix with a row of zeros has determinant zero.

## 4.2 Kernel

**Definition 4.8** (Kernel)**.** Suppose $T : \mathbb{R}^m \to \mathbb{R}^n$ is a linear transformation. The *kernel* of $T$ is the set of vectors $\boldsymbol{x}$ such that $T(\boldsymbol{x}) = \boldsymbol{0}$, denoted by $ker(T)$. In other words,

$$ker(T) = \{\boldsymbol{x} \in \mathbb{R}^m | T(\boldsymbol{x} = \boldsymbol{0})\} \tag{45}$$

**Theorem 4.3** (Kernel and Injectivity)**.** *Suppose $T : \mathbb{R}^m \to \mathbb{R}^n$ is a linear transformation. Then $T$ is one-to-one if and only if $ker(T) = \{\boldsymbol{0}\}$.*

This is rather intuitive. $T$ being one-to-one means $T(\boldsymbol{x}) = \boldsymbol{0}$ has only the trivial solution which is $\boldsymbol{x} = \boldsymbol{0}$. By definition of kernel, $ker(T) = \{\boldsymbol{0}\}$.

## 4.3 Basis

**Definition 4.9** (Basis)**.** A set $\mathcal{B} = \{\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}\}$ is a basis for a subspace $\mathcal{S}$ if

a) $\mathcal{B}$ spans $\mathcal{S}$.

b) $\mathcal{B}$ is linearly independent.

To find basis for $\mathcal{S} = span\{\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}\}$,

1. Use $\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}$ to form the rows of a matrix $\boldsymbol{A}$.
2. Transform $\boldsymbol{A}$ into row echelon form $\boldsymbol{B}$.
3. The nonzero rows give a basis for $\mathcal{S}$.

[1] Source: http://www.math.lsa.umich.edu/~hochster/419/det.html

### 4.4 Dimension, Row & Column Space, and Rank

**Definition 4.10** (Dimension)**.** Let $\mathcal{S}$ be a subspace of $\mathbb{R}^n$. Then the dimension of $\mathcal{S}$, denoted as $dim(\mathcal{S})$, is the number of vectors in any basis of $\mathcal{S}$.

**Definition 4.11** (Row Space, Column Space)**.** Suppose $\boldsymbol{A} \in M_{n\times m}(\mathbb{R})$. Then:

- $row(\boldsymbol{A})$ = span of rows of $\boldsymbol{A}$ (row space)
- $col(\boldsymbol{A})$ = span of columns of $\boldsymbol{A}$ (column space)

$row(\boldsymbol{A}) \subseteq \mathbb{R}^m$, $col(\boldsymbol{A}) \subseteq \mathbb{R}^n$.

**Theorem 4.4** (Basis for Row and Column Spaces)**.** *Let $\boldsymbol{A}$ be a matrix, and $\boldsymbol{B}$ be a row-echelon form of that matrix. Then*

*a) The nonzero rows of $\boldsymbol{B}$ form a basis for $row(\boldsymbol{A})$.*

*b) The columns of $\boldsymbol{A}$ corresponding to pivot columns of $\boldsymbol{B}$ form a basis for $col(\boldsymbol{A})$.*

**Theorem 4.5** (Dimension of Row and Column Spaces Are Equal)**.** *The following is always true for matrix $\boldsymbol{A}$:*

$$dim(col(\boldsymbol{A})) = dim(row(\boldsymbol{A})) \tag{46}$$

**Definition 4.12** (Rank)**.** The rank of a matrix $\boldsymbol{A}$ is defined by:

$$rank(\boldsymbol{A}) = dim(col(\boldsymbol{A})) = dim(row(\boldsymbol{A})) \tag{47}$$

**Definition 4.13** (Nullity)**.** The *nullity* of $\boldsymbol{A}$ is $dim(null(\boldsymbol{A}))$.

**Theorem 4.6** (Rank-Nullity Theorem)**.** *Let $\boldsymbol{A}$ be an $n \times m$ matrix. Then*

$$rank(\boldsymbol{A}) + nullity(\boldsymbol{A}) = m \tag{48}$$

## 5 Eigen

**Definition 5.1** (Eigenvector and Eigenvalue)**.** Let $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, then a nonzero vector $\boldsymbol{u}$ is an *eigenvector* of $\boldsymbol{A}$ if there exists a scalar $\lambda$ such that $\boldsymbol{Au} = \lambda\boldsymbol{u}$. The scalar $\lambda$ is called the *eigenvalue*

**0** is never an eigenvector.

**Theorem 5.1** (Scaled Eigenvectors)**.** *Suppose $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, and $\boldsymbol{u}$ is an eigenvector with eigenvalue $\lambda$. Then for any $r \neq 0$, $r \in \mathbb{R}$, $r\boldsymbol{u}$ is another eigenvector with eigenvalue $\lambda$.*

It is important to note that the theorem above does not imply that all eigenvectors with eigenvalue $\lambda$ should be related by the scalar $\lambda$. With this in mind, it is more intuitive to accept the following theorem.

**Theorem 5.2** (Eigenspace)**.** *If* $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$*, then the set of eigenvectors with eigenvalue* $\lambda$*, together with* $\mathbf{0}$ *is a subspace of* $\mathbb{R}^n$*, called the* eigenspace*.*

**Theorem 5.3** (Condition for an Eigenvalue)**.** *Let* $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$*. Then* $\lambda$ *is an eigenvalue of* $\boldsymbol{A}$ *if and only if*

$$det(\boldsymbol{A} - \lambda \boldsymbol{I_n}) = 0. \tag{49}$$

We refer to $det(\boldsymbol{A} - \lambda \boldsymbol{I_n}) = 0$ as the *characteristic polynomial.*

**Definition 5.2** (Characteristic Polynomial)**.** The *characteristic polynomial* of an $n \times n$ matrix $\boldsymbol{A}$, $char_{\boldsymbol{A}}(\lambda)$, is the degree $n$ polynomial $det(\boldsymbol{A} - \lambda \boldsymbol{I_n}) = 0$.

Caveat: Some linear maps do not have eigenvalues or eigenvectors, such as below:

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

The intuition of eigenvectors is to think of them as the axis of the corresponding linear transformation. The eigenvalue $\lambda$ helps to know if $\boldsymbol{x}$ is stretched or shrunk, when multiplied by a matrix $\boldsymbol{A}$ (i.e. $\boldsymbol{Ax}$).

## 5.1 Multiplicity of Eigenvalues

**Definition 5.3** (Algebraic Multiplicity)**.** The algebraic multiplicity of an eigenvalue $\alpha$ of $\boldsymbol{A}$ is found by $k$ in $char_{\boldsymbol{A}} = (\alpha - \lambda)^k Q(\lambda)$ where $Q(\lambda)$ is a polynomial with $Q(\lambda) \neq 0$.

For example, for $char_{\boldsymbol{A}} = -\lambda(\lambda-2)^2 = -(\lambda-0)(\lambda-2)^2$. Therefore, $\lambda = 0$ has algebraic multiplicity of 1, and $lambda = 2$ has algebraic multiplicity of 2.

**Definition 5.4** (Geometric Multiplicity)**.** The geometric multiplicity of an eigenvalue $\lambda$ is the dimension of the eigenspace associated with $\lambda$, i.e. number of linearly independent eigenvectors of that eigenvalue.

- 0 is eigenvalue if $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$ is *singular* (See definition 4.6).
- Geometric multiplicity $\leq$ algebraic multiplicity (of an eigenvalue).

## 5.2 Eigendecomposition

**Definition 5.5** (Eigendecomposition of a Matrix)**.** Let $\boldsymbol{A}$ be an $n \times n$ matrix, with $n$ linearly independent eigenvectors $\boldsymbol{u_i}$ for $i \in \{1, \cdots, n\}$. Then we can perform an eigendecomposition of $\boldsymbol{A}$ as follows

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1} \tag{50}$$

where $\boldsymbol{U}$ is an $n \times n$ matrix whose $i$th column is the eigenvector $\boldsymbol{u_i}$ of $\boldsymbol{A}$, and $\boldsymbol{\Lambda}$ is the diagonal matrix whose diagonal entries are the corresponding eigenvalues (i.e. $\Lambda_{ii} = \lambda_i$).

This definition implies that $\boldsymbol{A}$ must be *diagonalizable* (Section 7.4). It is usually convenient to have $\boldsymbol{U}$ be a orthonormal matrix.

# 6 The Big Theorem

**Theorem 6.1** (The Big Theorem)**.** *Let $\mathcal{A} = \{\boldsymbol{a_1}, \cdots, \boldsymbol{a_n}\}$ be a set of vectors in $\boldsymbol{R}^n$. Let $\boldsymbol{A} = [\boldsymbol{a_1} \quad \cdots \quad \boldsymbol{a_n}]$ be an $n \times n$ matrix, and let $T : \mathbb{R}^n \to \mathbb{R}^n$ be given by $T(\boldsymbol{X}) = \boldsymbol{Ax}$. Then the following statements are equivalent:*

*a)* $\mathcal{A}$ *spans* $\mathbb{R}^n$

*b)* $\mathcal{A}$ *is linearly independent (i.e.* $\boldsymbol{Ax} = \boldsymbol{0}$ *has only the trivial solution)*

*c)* $\mathcal{A}$ *is a basis for* $\mathbb{R}^n$

*d)* $\boldsymbol{Ax} = \boldsymbol{b}$ *has a unique solution for all* $\boldsymbol{b} \in \mathbb{R}^n$

*e)* $T$ *is onto (surjective)*

*f)* $T$ *is one-to-one (injective)*

*g)* $\boldsymbol{A}$ *is an invertible matrix*

*h)* $ker(T) = \{\boldsymbol{0}\}$

*i)* $col(\boldsymbol{A}) = \mathbb{R}^n$

*j)* $row(\boldsymbol{A}) = \mathbb{R}^n$

*k)* $rank(\boldsymbol{A}) = n$

*l)* $det(\boldsymbol{A}) \neq 0$

*m)* $\lambda = 0$ *is not an eigenvalue of* $\boldsymbol{A}$

# 7 Special Matrices

## 7.1 Block Matrix

**Definition 7.1** (Block Matrix)**.** A block matrix $\boldsymbol{M}$ is defined as

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$$

where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ are matrices (or block matrices) themselves.

Block matrices share many useful properties as normal matrices, by treating block entries as normal matrix entries. For example:

$$\boldsymbol{M}^2 = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \tag{51}$$

$$= \begin{bmatrix} \boldsymbol{A}^2 + \boldsymbol{BC} & \boldsymbol{AB} + \boldsymbol{BD} \\ \boldsymbol{CA} + \boldsymbol{DC} & \boldsymbol{CB} + \boldsymbol{D}^2 \end{bmatrix} \tag{52}$$

### 7.2 Orthogonal

**Definition 7.2** (Orthogonal Matrix)**.** An *orthogonal matrix* $\boldsymbol{Q}$ is a square matrix with real entries whose columns and rows are orthogonal unit vectors (i.e., *orthonormal* vectors), i.e.

$$\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I} \tag{53}$$

Therefore, we have $\boldsymbol{Q}^T = \boldsymbol{Q}^{-1}$. To fully understand why Equation 53 holds, we need to know that for two orthogonal vectors $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$, $\boldsymbol{u_1}^T\boldsymbol{u_2} = 0$. And $\boldsymbol{u_1}^T\boldsymbol{u_1} = |\boldsymbol{u_1}|^2 = 1$. Therefore, in the resulting matrix, all entries are 0 except for ones along the diagonal.

### 7.3 Diagonal

**Definition 7.3** (Diagonal Matrix)**.** A square matrix $\boldsymbol{D}$ is a diagonal matrix if all entries except for ones along the main diagonal are 0.

Simple fact: for two diagonal matrices $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$, their multiplication $\boldsymbol{D_1}\boldsymbol{D_2} = \boldsymbol{D_3}$ is also a diagonal matrix with each entry $\boldsymbol{D_3}[i]$ along[2] the main diagonal equals to $\boldsymbol{D_1}[i]\boldsymbol{D_2}[i]$.

Therefore, every diagonal matrix is invertible. The inverse $\boldsymbol{D}^{-1}$ of diagonal matrix $\boldsymbol{D}$ has entries $\boldsymbol{D}^{-1}[i] = 1/\boldsymbol{D}[i]$.

Another fact: The determinant of a diagonal matrix is the product of the diagonal entries.

Yet another fact: The column vectors of a diagonal matrix $\boldsymbol{D}$ are the eigenvectors of $\boldsymbol{D}$, and each diagonal entry is the eigenvalue for the eigenvector at the corresponding column, that is

$$\boldsymbol{D} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \tag{54}$$

This can be verified simply by solving the characteristic polynomial $det(\boldsymbol{D} - \lambda\boldsymbol{I}) = 0$.

### 7.4 Diagonalizable

**Definition 7.4** (Diagonalizable Matrix)**.** An $n \times n$ matrix $\boldsymbol{A}$ is *diagonalizable* if there exists an $n \times n$ matrix $\boldsymbol{P}$ such that

$$\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} \tag{55}$$

where $\boldsymbol{D}$ is a diagonal matrix.

Note that $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} \Longrightarrow \boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$

[2] The $[i]$ just means the $i$th entry along the main diagonal.

**Theorem 7.1** (The Diagonalization Theorem)**.**

a) *An* $n \times n$ *matrix* $\boldsymbol{A}$ *is diagonalizable if and only if* $\boldsymbol{A}$ *has* $n$ *linearly independent* eigenvectors*.*

b) $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$ *where* $\boldsymbol{D}$ *is a diagonal matrix* if and only if *all* $n$ *columns of* $\boldsymbol{P}$ *are linearly independent eigenvectors of* $\boldsymbol{A}$ and *the diagonal entries of* $\boldsymbol{D}$ *are their corresponding eigenvalues.*

If we can find $n$ linearly independent eigenvectors for an $n \times n$ matrix $\boldsymbol{A}$, then we know the matrix is diagonalizable. Furthermore, we can use those eigenvectors and their corresponding eigenvalues to find the invertible matrix $\boldsymbol{P}$ and diagonal matrix $\boldsymbol{D}$ necessary to show that $\boldsymbol{A}$ is diagonalizable.

**Theorem 7.2** (Power of Diagonalizable Matrix)**.** *If* $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$, *then* $\boldsymbol{A}^k = \boldsymbol{P}\boldsymbol{D}^k\boldsymbol{P}^{-1}$

## 7.5 Symmetric

**Definition 7.5** (Symmetric Matrix)**.** A square matrix $\boldsymbol{A}$ is symmetric if and only if

$$\boldsymbol{A} = \boldsymbol{A}^T \tag{56}$$

For any $n \times m$ matrix $\boldsymbol{B}$, the matrix $\boldsymbol{B}^T\boldsymbol{B} \in \mathbb{R}^{n \times n}$ is symmetric. Also, every square diagonal matrix is symmetric.

*Facts about symmetric matrix*

- Any symmtric matrix:
  - has only real eigenvalues;
  - is always *diagonalizable*;
  - has orthogonal eigenvectors;
- The symmetric matrix $\boldsymbol{A}$ is
  - positive definite if all its eigenvalues are positive.
  - positive semidefinite if all its eigenvalues are non negative..

## 7.6 Positive-Definite

We omit the discussion of complex matrices for now.

**Definition 7.6** (Positive-Definite)**.** A *symmetric* $n \times n$ real matrix $\boldsymbol{A}$ is *positive definite* if for all $\boldsymbol{x} \in \mathbb{R}^n$
$\{\boldsymbol{0}\}$,

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0 \tag{57}$$

The negative definite, positive semi-definite, and negative semi-definite matrices are defined analogously. For "$*$ semi-$*$", zero is allowed (e.g. $\boldsymbol{A}$ is positive semi-definite implies $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0$).

**Theorem 7.3.** *Covariance matrix is positive semi-definite.*

Given data $\boldsymbol{X} \in \ltimes \times$ , its covariance matrix $\boldsymbol{\Sigma}$ is computed by the following:

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T] \tag{58}$$

For nonzero $\boldsymbol{y} \in \mathbb{R}^d$

$$\begin{aligned}
\boldsymbol{y}^T \boldsymbol{\Sigma} \boldsymbol{y} &= \boldsymbol{y}^T \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T] \boldsymbol{y} && (59) \\
&= \mathbb{E}[\boldsymbol{y}^T (\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T \boldsymbol{y}] && (60) \\
&= \mathbb{E}[\boldsymbol{Q}^T \boldsymbol{Q}] && (61)
\end{aligned}$$

For $\boldsymbol{Q} = (\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T \boldsymbol{y}$. Therefore, $\boldsymbol{y}^T \boldsymbol{\Sigma} \boldsymbol{y} \geq 0$, which means $\boldsymbol{\Sigma}$ is *positive semi-definite*.

### 7.7 Singular Value Decomposition

**Theorem 7.4.** *For any given real matrix $A \in \mathbb{R}^{n \times m}$, there exists a unique set of matrices $U, S, V$ such that*

$$A = USV^T \tag{62}$$

*where $U \in \mathbb{R}^{n \times n}$ and $S \in \mathbb{R}^{n \times p}$ and $V \in \mathbb{R}^{p \times p}$ $U^T U = I$ and $V^T V = I$. This is called the* singular value decomposition *of $A$.*

$U$ and $V$ are orthonormal matrices. $S$ is a diagonal matrix[3]. The elements in $S$ are called *singular values* of $A$. The eigenvectors of $A^T A$ are columns of $V$, and the eigenvectors of $AA^T$ are columns of $U$. The entries in $S$ are positive, and sorted in decreasing order ($S_{11} \geq S_{22} \geq \cdots$).

### 7.8 Similar

### 7.9 Jordan Normal Form

### 7.10 Hermitian

## 8 Matrix Calculus

### 8.1 Differentiation

### 8.2 Hessian

[3]More precisely, it is a rectangular diagonal matrix because $n$ may not equal to $p$. Still, $S_{ij} = 0$ if $i \neq j$.