Contents

1 Linear Algebra Tools

This chapter introduces inner product to give geometric meaning to vectors and vector spaces, enabling calculations of length, distance, and angles.

Definition (Symmetric Positive Definitive Matrix). *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ that satisfies*

$$\text{for every nonzero vector } x : x^T A x > 0 \tag{1.1}$$

*is called **positive definite**. If only $\geq$ holds in 1.1, then A is called **positive semidefinite**.*

These properties helps in identifying positive definite matrices without having to check the definition explicitly:

1. The null space of A contains only the null vector;
2. The diagonal elements a_{ii} of A are positive;
3. The eigenvalues of A are real and positive.

1.1 Angles and Orthogonality

The angle ω between vectors x and y is computed as:

$$\cos\omega = \frac{\langle x, y\rangle}{\|x\|_2\|y\|_2}$$

Here, $\langle x, y\rangle$ denotes the inner product between x and y.
This angle indicated the vectors' similarity in orientation.

Definition (Orthogonal vectors). *Two vectors are orthogonal if $\langle x, y\rangle = 0$. If additionally $\|x\| = 1 = \|y\|$, then x and y are orthonormal.*

Definition (Orthogonal matrix). *A square matrix is an orthogonal matrix if and only if <u>its columns are orthonormal</u> so that*

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

which implies that

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

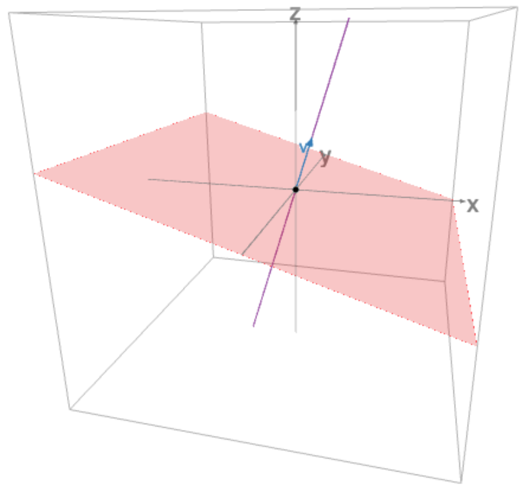
The length of a vector x is not changed when transforming it using an orthogonal matrix A.

$$\|Ax\|_2^2 = \|x\|_2^2$$

Moreover, the angle between any two vectors x, y is also unchanged when transforming both of them using an orthogonal matrix A.

Definition (Orthonormal Basis). *In an n-dimensional vector space V with a basis set $\{b_1, \ldots, b_n\}$, if all the basis vectors are orthogonal to each other, the basis is called as an **orthogonal basis**. Additionally, if the length of each basis vector is 1, the basis is referred to as an **orthonormal basis**.*

We can also have vector spaces that are orthogonal to each other. Given a vector space V of dimension D, let's consider a subspace U of dimension M such that $U \subseteq V$. Then its **orthogonal complement** $U^\perp$ is a $D - M$ dimensional subspace V and contains all vectors in V that are orthogonal to every vector in U.



1.1.1 Orthogonal Projections

Projections are key linear transformations in machine learning and are particularly useful for handling high-dimensional data. Often, only a few dimensions in such data are essential for capturing the most relevant information. By projecting the original high-dimensional data onto a lower dimensional feature space, we can work more efficiently to learn about the dataset and extract significant patterns.

Definition (Projection). *Let V be a vector space and $U \subseteq V$ a subspace of V. A linear mapping $\pi : V \to U$ is called **projection** if it satisfies $\pi^2 = \pi \circ \pi = \pi$.*

Given that linear mappings can be represented by transformation matrices, the above definition extends naturally to *projection matrices* P_π. These matrices exhibit the property that $P_\pi^2 = P_\pi$.
The projection $\pi_U(x)$ of a vector $x \in \mathbb{R}^n$ onto a subspace U is the closest point necessarily in U to x.

2 Matrix Decompositions

2.1 Eigenvalues and Eigenvectors

Eigenanalysis helps us understand linear transformations represented by a matrix A. Eigenvectors x are special vectors that only get scaled, not rotated, when multiplied by A. The scaling factor is the eigenvalue λ, which indicated how much x is stretched or shrunk. λ can also be zero.

Definition (Eigenvalue and Eigenvector)**.** *Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an **eigenvalue** of A and nonzero vector x is the corresponding **eigenvector** of A if*

$$Ax = \lambda x \tag{2.1}$$

*We call 2.1 the **eigenvalue equation**.*

The following statements are equivalent:

- λ is an eigenvalue of $A \in \mathbb{R}^{n \times n}$.
- A nonzero vector x exists such that $Ax = \lambda x$ or, equivalently, $(A - \lambda I_n)x = 0$ for $x \neq 0$.
- Then $A - \lambda I$ is a **singular matrix** and its determinant is **zero**.

Each eigenvector x has one unique eigenvalue λ, but each λ can have multiple eigenvectors.

Definition (Eigenspace and Eigenspectrum)**.** *For $A \in \mathbb{R}^{n \times n}$, the set of all eigenvectors of A associated with an eigenvalue λ spans a subspace of $\mathbb{R}^n$, which is called the **eigenspace** of A with respect to λ and is denoted by E_λ. The set of all eigenvalues of A is called the **eigenspectrum** of A.*

Definition. *Let λ_i be an eigenvalue of a square matrix A. Then the **geometric multiplicity** 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.*

Theorem. *The eigenvectors $x_1, \ldots, x_n$ of a matrix $A \in \mathbb{R}^{n \times n}$ with n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ are linearly independent.*

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

3 Vector calculus

Firstly, we'll explore partial derivatives and gradients, focusing on functions that take a vector as input and produce a single real number as output. These functions are formally represented as $f : \mathbb{R}^n \to \mathbb{R}$.
Subsequently, we will extend these ideas to functions that not only take a vector as input but also produce a vector as output. These functions can be written as $f : \mathbb{R}^n \to \mathbb{R}^m$.
When we deal with a function that depends on multiple variables, such as $f(x) = f(x_1, x_2)$, we use the **gradient** to represent its derivative. The gradient is a vector composed of **partial derivates** of the function. To compute each partial derivates, we differentiate the function with respect to one variable while keeping all other variables constant.

$$\nabla_x f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1\times n}$$

where n is the number of variables.

Basic Rules of Partial Differentiation

Product rule:

$$\frac{\partial}{\partial x}(f(x)g(x)) = \frac{\partial f}{\partial x}g(x) + f(x)\frac{\partial g}{\partial x}$$

Sum rule:

$$\frac{\partial}{\partial x}(f(x) + g(x)) = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x}$$

Chain rule:

$$\frac{\partial}{\partial x}(g \circ f)(x) = \frac{\partial}{\partial x}\left(g(f(x))\right) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$$

In the context of the chain rule, consider f as implicitly a composition $f \circ g$.
If a function $f(x_1, x_2)$ is a function of x_1 and x_2, where $x_1(t)$ and $x_2(t)$ are themselves functions of a single variable t, the chain rule yields the partial derivates

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial t}{\partial x_2}\frac{\partial x_2}{\partial t}$$

Example

Consider $f(x_1, x_2) = x_1^2 + 2x_2$, where $x_1 = \sin t$ and $x_2 = \cos t$, then

$$\text{with} \quad \frac{\partial f}{\partial x_1} = 2x_1, \quad \frac{\partial f}{\partial x_2} = 2$$

$$\begin{aligned} \frac{\mathrm{d}f}{\mathrm{d}t} &= 2\sin t\frac{\partial \sin t}{\partial t} + 2\frac{\partial \cos t}{\partial t} \\ &= 2\sin t \cos t - 2\sin t \end{aligned}$$

If a function $f(x_1, x_2)$ is a function of x_1 and x_2, where $x_1(s,t)$ and $x_2(s,t)$ are themselves functions of two variables s and t, the chain rule yields the partial derivates

$$\frac{\mathrm{d}f}{\mathrm{d}(s,t)} = \begin{bmatrix} \frac{\partial f}{\partial s} & \frac{\partial f}{\partial t} \end{bmatrix}$$

where

$$\begin{aligned} \frac{\partial f}{\partial s} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s} \\ \frac{\partial f}{\partial t} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \end{aligned}$$

Another way to obtain these two partial derivatives is to represent the previous formula as a row vector containing the partial derivatives of f with respect to x_1 and x_2. This row vector is then multiplied by a matrix composed of the partial derivatives of x_1 and x_2 with respect to s and t. When you perform this multiplication, you get the exact same result as above.

$$\begin{bmatrix} \frac{\partial f}{\partial s} & \frac{\partial f}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}$$

Example

Given the following functions:
$g : \mathbb{R}^2 \to \mathbb{R}^2 \quad g(s,t) = (\sin(t)s, \cos(s)t)$
$f : \mathbb{R}^2 \to \mathbb{R} \quad f(x_1,x_2) = x_1^2 + 2x_2$
$f \circ g : \mathbb{R}^2 \to \mathbb{R}$
Compute $\nabla_{(s,t)}(f \circ g)$ and evaluate $\nabla_{(s,t)}(f \circ g)(0,0)$.

$$= \begin{bmatrix} 2s\sin(t) & 2 \end{bmatrix} \begin{bmatrix} \sin(t) & s\cos(t) \\ -t\sin(s) & \cos(s) \end{bmatrix}$$

$$= \begin{bmatrix} 2s\sin^2(t) - 2t\sin(s) \\ 2s^2\sin(t)\cos(t) + 2\cos t \end{bmatrix} = (0,2)$$