EECS 127
Convex Optimization
Notes

Instructor: Laurent El Ghaoui

Kelvin Lee

UC Berkeley

Contents

1 Introduction **3**
- 1 Standard Form of Optimization . . . 3
- 2 Convex Problems . . . 8
- 3 Non-convex Problems . . . 8

2 Vectors and Functions **9**
- 4 Basics . . . 9
- 5 Vector Spaces . . . 9
- 6 Inner Product . . . 11
- 7 Orthogonality and Orthonormality . . . 13
- 8 Functions and Maps . . . 17
- 9 Hyperplanes and Halfspaces . . . 18
- 10 Gradients . . . 19

3 Matrices and Linear Maps **22**
- 11 Matrix Basics . . . 22
- 12 Matrices as linear maps . . . 24
- 13 Determinants . . . 25

4 Convex Sets **28**
- 14 Affine Sets . . . 28
- 15 Convex Sets . . . 28
- 16 Cones . . . 29
- 17 Hyperplanes and Halfspaces . . . 30
- 18 Operations preserving convexity . . . 31

1

Introduction

1 Standard Form of Optimization

$$
\begin{array}{ll}
p^* = \min_{\boldsymbol{x}} & f_0(\boldsymbol{x}) \\
\text{subject to:} & f_i(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, m
\end{array}
$$

where

- vector $\boldsymbol{x} \in \mathbb{R}^n$ is the **decision variable**;
- $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the **objective function**, or **cost**;
- $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, m,$ represent the **constraints**;
- p^* is the **optimal value**.

1.1 Least-squares Regression

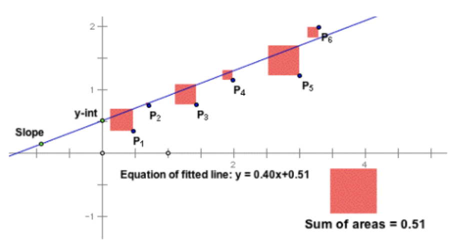

Figure 1.1: Least-squares regression.

$$
\min_{\boldsymbol{x}} \sum_{i=1}^{m} \left(\boldsymbol{y}_i - \boldsymbol{x}^\top \boldsymbol{z}^{(i)} \right)^2
$$

where

- $\boldsymbol{z}^{(i)} \in \mathbb{R}^n, i = 1, \ldots, n$ are **data points**;
- $\boldsymbol{y} \in \mathbb{R}^m$ is a **response vector**;

- $\boldsymbol{x}^\top \boldsymbol{z}$ is the **scalar product** $z_1x_1 + \ldots + z_nx_n$ between the two vectors $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$.
- Many variants exist.
- Once $\boldsymbol{x}$ is found, allows to predict the output $\hat{\boldsymbol{y}}$ corresponding to a new data point $z : \hat{\boldsymbol{y}} = \boldsymbol{x}^\top \boldsymbol{z}$.

1.2 Linear Classification

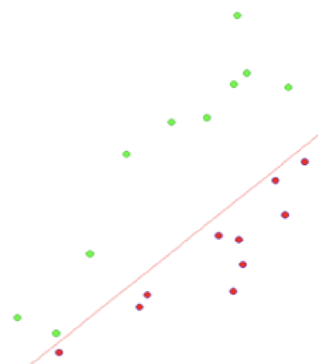

Figure 1.2: Linear classification.

Support Vector Machine (SVM):

$$\min_{\boldsymbol{x},b} \sum_{i=1}^{m} \max\left(0, 1 - y_i\left(\boldsymbol{x}^\top \boldsymbol{z}^{(i)} + b\right)\right)$$

where

- $\mathbf{0}z^{(i)} \in \mathbb{R}^n, i = 1, \ldots, n$ are data points;
- $\boldsymbol{y} \in \{-1, 1\}^m$ is a **binary response vector**;
- $x^\top z + b = 0$ defines a **separating hyperplane** in data space.
- Once $\boldsymbol{x}, n$ are found, we can predict the binary output $\hat{y}$ corresponding to a new data point $\boldsymbol{z}$: $\hat{y} = \textbf{sign}\left(\boldsymbol{x}^T\boldsymbol{z} + b\right)$.
- Very useful for classifying data.

1.3 Nomenclature

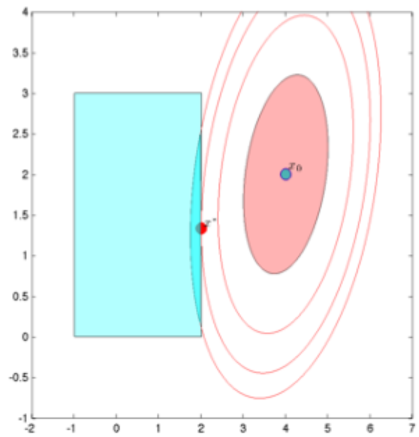

Figure 1.3: A toy optimization problem.

$$\begin{array}{ll} \min_{\boldsymbol{x}} & 0.9x_1^2 - 0.4x_1x_2 - 0.6x_2^2 - 6.4x_1 - 0.8x_2 \\ \text{s.t.} & -1 \le x_1 \le 2, 0 \le x_2 \le 3. \end{array}$$

- **Feasible set**: a set of possible values that satisfy the constraints. (light blue region)
- **Unconstrained minimizer**: x_0.
- **Optimal Point**: $\boldsymbol{x}^*$.
- **Level sets** of objective functions: $\{\boldsymbol{x} \mid g(\boldsymbol{x}) = c\}$ for some c. (red lines)
- **Sub-level sets**: $\{\boldsymbol{x} \mid g(\boldsymbol{x}) \le c\}$ for some c. (red region)

1.4 Problems with equality constraints

Sometimes the problem may have equality constraints, along with inequality ones:

$$\begin{array}{rll} p^* = \min_{\boldsymbol{x}} & f_0(\boldsymbol{x}) & \\ \text{s.t.} & f_i(\boldsymbol{x}) \le 0, & i = 1, \dots, m \\ & h_i(\boldsymbol{x}) = 0, & i = 1, \dots, p \end{array}$$

where h_i's are given functions.
However, we can always reduce it to a **standard form** with inequality constraints only, using the following method:

$$h_i(\boldsymbol{x}) = 0 \quad \Longrightarrow \quad h_i(\boldsymbol{x}) \le 0, \quad h_i(\boldsymbol{x}) \ge 0.$$

1.5 Problems with set constraints

Sometimes, the constraints of the problem are described abstractly via a set-membership condition of the form $\boldsymbol{x} \in \mathcal{X}$, for some subset $\mathcal{X}$ of $\mathbb{R}^n$.

The corresponding notation is

$$p^* = \min_{\boldsymbol{x} \in \mathcal{X}} f_0(\boldsymbol{x}),$$

or, equivalently,

$$\begin{aligned} p^* = \min_{\boldsymbol{x}} \quad & f_0(x) \\ \text{s.t.} \quad & x \in \mathcal{X}. \end{aligned}$$

1.6 Problems in maximization form

Some optimization problems come in the form of maximization (instead of minimization) of an objective function, i.e.,

$$p^* = \max_{\boldsymbol{x} \in \mathcal{X}} g_0(\boldsymbol{x}).$$

We can recast it as a **standard minimization form** using the following fact:

$$\max_{\boldsymbol{x} \in \mathcal{X}} g_0(\boldsymbol{x}) = -\min_{\boldsymbol{x} \in \mathcal{X}} -g_0(\boldsymbol{x}).$$

Thus we can reformulate the problem as the following:

$$-p^* = \min_{\boldsymbol{x} \in \mathcal{X}} f_0(\boldsymbol{x}),$$

where $f_0 = -g_0$.

1.7 Feasible Set

The **feasible set** of a problem is defined as

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n \mid f_i(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, m\}.$$

Definition 1 (Infeasible)**.** A problem is **infeasible** if the feasible set is empty, i.e., the constraints cannot be satisfied simultaneously.

Remark. We take the convention that the optimal value is $p^* = +\infty$ for infeasible minimization problems, while $p^* = -\infty$ for infeasible maximization problems.

1.8 Feasibility Problems

Sometimes an objective function is not provided. This means that we are just interested in finding a feasible point, or determine that the problem is infeasible.

In this case, we set f_0 to be a constant to reflect the fact that we are indifferent to the choice of a point $\boldsymbol{x}$, as long as it is feasible.

1.9 Solution to an optimization problem

The optimal value p^* is *attained* if there exists a feasible $\boldsymbol{x}^*$ such that

$$f_0(\boldsymbol{x}^*) = p^*.$$

1.9.1 Optimal Set

Definition 2 (Optimal Set)**.** The **optimal set** is defined as

$$\mathcal{X}_{\mathsf{OPT}} = \{\boldsymbol{x} \in \mathbb{R}^n \mid f_0(\boldsymbol{x}) = p^*, f(\boldsymbol{x}) \leq 0, i = 1, \ldots, m\},$$

or equivalently,

$$\mathcal{X}_{\mathsf{OPT}} = \arg\min_{\boldsymbol{x} \in \mathcal{X}} f_0(\boldsymbol{x}).$$

A point $\boldsymbol{x}$ is **optimal** if $\boldsymbol{x} \in \mathcal{X}_{\mathsf{OPT}}$.

1.9.2 Empty Optimal Set

The optimal set can be empty for two reasons:

1. The problem is infeasible.
2. The optimal value is only reached in the limit.
   - For example, the problem
     $$p^* = \min_x e^{-x}$$
     has no optimal points because $p^* = 0$ is only reached in the limit for $x \to +\infty$.
   - Another example is when constraints include strict inequalities:
     $$p^* = \min_x x \quad \text{s.t } 0 < x \leq 1.$$
     In this case, $p^* = 0$ but cannot be attained by any $\boldsymbol{x}$ that satisfies the constraints.

1.9.3 Sub-optimality

Definition 3 (Suboptimal)**.** We say that a point $\boldsymbol{x}$ is **ϵ-suboptimal** for a problem if it is feasible, and satisfies

$$p^* \leq f_0(\boldsymbol{x}) \leq p^* + \epsilon.$$

In other words, $\boldsymbol{x}$ is ϵ-close to p^*.

1.9.4 Local vs. global optimal points

Definition 4 (Locally optimal)**.** A point $\boldsymbol{z}$ is **locally optimal** if there exist a value $R > 0$ such that $\boldsymbol{z}$ is optimal for problem

$$\min_{\boldsymbol{x}} f_0(\boldsymbol{x}) \text{ s.t. } f_i(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, m, \quad |x_i - z_i| \leq R, \quad i = 1, \ldots, n.$$

1.10 Problem Transformations

Sometimes we can cast a problem in a tractable formulation. For example, consider

$$\max_{\boldsymbol{x}} x_1 x_2^3 x_3 \quad \text{s.t.} \quad x_i \geq 0, \quad i = 1, 2, 3, \quad x_1 x_2 \leq 2, \quad x_2^2 x_3 \leq 1$$

can be transformed into the following by taking the log, in terms of $z_i = \log x_i$:

$$\max_{\boldsymbol{z}} z_1 + 3z_2 + z_3 \quad \text{s.t.} \quad z_1 + z_2 \leq \log 2, \quad 2z_2 + z_3 \leq 0.$$

Now the objective function and the constraints are all linear.

2 Convex Problems

Convex optimization problems are problems where the objective and constraint functions have the special property of **convexity**.

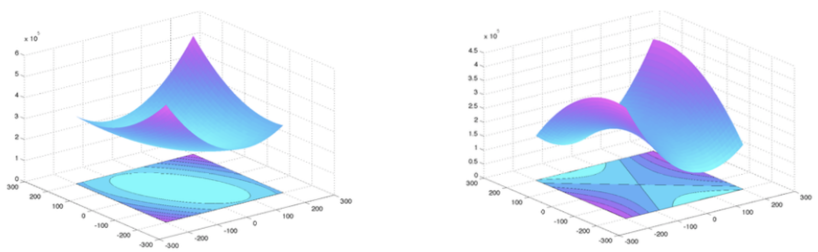

Figure 1.4: *Left.* Convex function. *Right.* Non-convex function.

For a convex function, any local minimum is global.

2.1 Special convex models

Convex optimization problems with special structure:

- Least-Squares (LS)
- Linear Programs (LP)
- Convex Quadratic Programs (QP)
- Geometric Programs (GP)
- Second-order Cone Programs (SOCP)
- Semi-definite Programs (SDP).

3 Non-convex Problems

- **Boolean/integer optimization**: some variables are constrained to be Boolean or integers. Convex optimization can be used for getting good approximations.
- **Cardinality-constrained problems**: we seek to bound the number of non-zero elements in a vector variable. Convex optimization can be used for getting good approximations.
- **Non-linear programming**: usually non-convex problems with differentiable objective and functions. Algorithms provide only local minima.

Remark. Most non-convex problems are hard.

2

Vectors and Functions

4 Basics

Definition 5 (Vector)**.** A **vector** is a collection of numbers, arranged in a column or a row, representing the **coordinates** of a point in n-dimensional space. We write vectors in column format:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

where each element x_i is the i-**th component** of vector $\boldsymbol{x}$ and n is the **dimension** of $\boldsymbol{x}$. If $\boldsymbol{x}$ is a *real* vector, then we write $\boldsymbol{x} \in \mathbb{R}^n$. If $\boldsymbol{x}$ is a *complex* vector, then we write $\boldsymbol{x} \in \mathbb{C}^n$.

Definition 6 (Transpose)**.** The **transpose** of a vector $\boldsymbol{x}$ is defined as

$$\boldsymbol{x}^\top = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array}\right]$$

and the transpose of the transpose of $\boldsymbol{x}$ is itself, i.e., $\boldsymbol{x}^{\top\top} = \boldsymbol{x}$.

5 Vector Spaces

Definition 7 (Vector Space)**.** A **vector space** $\mathcal{V}$ is a set of vectors on which two operations: **vector addition** and **scalar multiplication**, are defined.

5.1 Subspaces and Span

Definition 8 (Subspace)**.** A nonempty subset $\mathcal{S}$ of a vector space $\mathcal{V}$ is a **subspace** of $\mathcal{V}$ if, for $\boldsymbol{x}, y \in \mathcal{S}$ and any scalars $\alpha, \beta \in \mathbb{R}$,

$$\alpha\boldsymbol{x} + \beta y \in \mathcal{S}.$$

In other words, $\mathcal{S}$ is *closed* under addition and scalar multiplication.

Definition 9 (Linear Combination)**.** A **linear combination** of a set of vectors $S = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ in a vector space $\mathcal{X}$ is a vector

$$\boldsymbol{x} = \sum_{i=1}^{m} \alpha\boldsymbol{x}^{(i)},$$

where each α_i is a given scalar.

Definition 10 (Span)**.** The **span** of a set of vectors $S = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ in a vector space $\mathcal{X}$ is the set of all vectors that is a linear combination of that set of vectors

$$\text{span}(S) = \left\{ \boldsymbol{x} \mid \exists \alpha_1, \ldots, \alpha_m \text{ s.t. } \boldsymbol{x} = \sum_{i=1}^{m} \alpha_i \boldsymbol{x}^{(i)} \right\}.$$

Definition 11 (Direct Sum)**.** Given two subspaces $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$, the **direct sum** of $\mathcal{X}, \mathcal{Y}$, denoted by $\mathcal{X} \oplus \mathcal{Y}$, is the set of vectors of the form $\boldsymbol{x} + y$, where $\boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}$. The direct sum is itself a subspace.

5.2 Bases and Dimensions

Definition 12 (Linearly Independent)**.** A set of vectors $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ in a vector space $\mathcal{X}$ is **linearly independent** if

$$\sum_{i=1}^{m} \alpha_i \boldsymbol{x}^{(i)} = 0 \implies \alpha_1 = \ldots = \alpha_n = 0.$$

Definition 13 (Basis)**.** Given a subspace of $\mathcal{S}$ of a vector space $\mathcal{X}$, a **basis** of $\mathcal{S}$ is a set $\mathcal{B}$ of vectors of minimal cardinality, such that $\text{span}(\mathcal{B}) = \mathcal{S}$.

Definition 14 (Dimension)**.** The **dimension** of a subspace is the cardinality of a basis of that subspace. If we have a basis $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)}\}$ for a subspace $\mathcal{S}$, then any element in the subspace can be expressed as a linear combination of the elements in the basis. That is, any $\boldsymbol{x} \in \mathcal{S}$ can be written as

$$\boldsymbol{x} = \sum_{i=1}^{d} \alpha_i \boldsymbol{x}^{(i)},$$

for some scalars α_i.

5.3 Affine Sets

Definition 15 (Affine Set)**.** An **affine set** is a set of the form

$$\mathcal{A} = \left\{ \boldsymbol{x} \in \mathcal{X} \mid \boldsymbol{x} = v + x^{(0)}, v \in \mathcal{V} \right\}$$

where $\boldsymbol{x}^{(0)}$ is a given point and $\mathcal{V}$ is a given subspace of $\mathcal{X}$. Subspaces are just affine spaces containing the origin.

Geometric interpretation: An affine set is a flat plane passing through $\boldsymbol{x}^{(0)}$.

The dimension of an affine set $\mathcal{A}$ is defined as the dimension of its generating subspace $\mathcal{V}$.

5.4 Euclidean Length

Definition 16 (Euclidean Length)**.** The **Euclidean length** of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\|\boldsymbol{x}\|_2 \doteq \sqrt{\sum_{i=1}^{n} x_i^2}.$$

5.5 Norms

Definition 17 (Norm)**.** A **norm** on a vector space $\mathcal{X}$ is a real-valued function with special properties that maps any element $\boldsymbol{x} \in \mathcal{X}$ into a real number $\|\boldsymbol{x}\|$.

Definition 18. A function from $\mathcal{X}$ to $\mathbb{R}$ is a **norm**, if

- $\forall \boldsymbol{x} \in \mathcal{X}, \|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = 0$;
- $\forall \boldsymbol{x}, y \in \mathcal{X}, \|\boldsymbol{x} + y\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ (**triangle inequality**);
- $\forall \boldsymbol{x} \in \mathcal{X}, \|\alpha \boldsymbol{x}\| = |\alpha| \|\boldsymbol{x}\|$ for any scalar α.

Definition 19 (ℓ_p norms)**.** ℓ_p **norms** are defined as

$$\|\boldsymbol{x}\|_p \doteq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \qquad 1 \leq p < \infty.$$

For $p = 2$, we have the **Euclidean length**

$$\|\boldsymbol{x}\|_2 \doteq \sqrt{\sum_{i=1}^{n} x_i^2},$$

or $p = 1$ we get the **sum-of-absolute-values length**

$$\|\boldsymbol{x}\|_1 \doteq \sum_{i=1}^{n} |x_i|.$$

The limit case $p = \infty$ defines the ℓ_∞ **norm** (**max absolute value norm**, or **Chebyshev norm**)

$$\|\boldsymbol{x}\|_\infty \doteq \max_{i=1,\ldots,n} |x_i|.$$

The cardinality of a vector $\boldsymbol{x}$ is called the ℓ_0 (**pseudo**) **norm** and denoted by $\|\boldsymbol{x}\|_0$.

6 Inner Product

Definition 20 (Inner Product)**.** An **inner product** on a real vector space $\mathcal{X}$ is a real-valued function which maps any pair of elements $\boldsymbol{x}, y \in \mathcal{X}$ into a scalar denoted as $\langle \boldsymbol{x}, y \rangle$. It satisfies the following axioms: for any $\boldsymbol{x}, y, z \in \mathcal{X}$ and scalar α

(i) $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$;

(ii) $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ if and only if $\boldsymbol{x} = 0$;

(iii) $\langle \boldsymbol{x} + y, z \rangle = \langle \boldsymbol{x}, z \rangle + \langle y, z \rangle$;

(iv) $\langle \alpha \boldsymbol{x}, y \rangle = \alpha \langle \boldsymbol{x}, y \rangle$;

(v) $\langle \boldsymbol{x}, y \rangle = \langle y, z \rangle$.

Definition 21 (Standard Inner Product)**.** The **standard inner product**, also called the **dot product** is defined as

$$\langle \boldsymbol{x}, y \rangle = \boldsymbol{x}^\top y = \sum_{i=1}^{n} x_i y_i.$$

An inner product naturally induces an associated **norm**: $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$.

6.1 Angle between vectors

The angle between $\boldsymbol{x}$ and y is defined via the relation

$$\cos = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2}.$$

There is a right angle between $\boldsymbol{x}$ and $\boldsymbol{y}$ when $\boldsymbol{x}^\top \boldsymbol{y} = 0$, i.e., $\boldsymbol{x}$ and $\boldsymbol{y}$ are **orthogonal**.
When $= 0°$, or $\pm 180°$, then $\boldsymbol{y} = \alpha \boldsymbol{x}$ for some scalar α, i.e. $\boldsymbol{x}$ and $\boldsymbol{y}$ are **parallel**. Then $|\boldsymbol{x}^\top \boldsymbol{y}|$ achieves its **maximum value** $|\alpha| \|\boldsymbol{x}\|_2^2$.

6.2 Cauchy-Schwartz and Hölder Inequality

Theorem 22 (Cauchy-Schwartz's Inequality)**.** For any vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, it holds that

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| = |\boldsymbol{x}^\top \boldsymbol{y}| \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2,$$

Proof. Note that $|\cos| \leq 1$, then using the angle equation, we have

$$|\cos| = \frac{|\boldsymbol{x}^\top \boldsymbol{y}|}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2} \leq 1 \implies |\boldsymbol{x}^\top \boldsymbol{y}| \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2.$$

□

Theorem 23 (Hölder's Inequality)**.** For any vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and for any $p, q \geq 1$ such that $1/p + 1/q = 1$, it holds that

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| = |\boldsymbol{x}^\top \boldsymbol{y}| \leq \sum_{i=1}^n |x_i y_i| \leq \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q.$$

6.3 Maximization of inner product over norm balls

Given a nonzero vector $\boldsymbol{y} \in \mathbb{R}^n$, we want to find some vector $\boldsymbol{x} \in \mathcal{B}_p$ (the unit ball in ℓ_p norm) that maximizes the inner product $\boldsymbol{x}^\top \boldsymbol{y}$, i.e., we want to solve the following:

$$\max_{\|\boldsymbol{x}\|_p \leq 1} \boldsymbol{x}^\top \boldsymbol{y}.$$

If the level set $\alpha = 0$, then we are solving for

$$\boldsymbol{x}^\top \boldsymbol{y} = 0,$$

which are the set of vectors that are on a line that is orthogonal to $\boldsymbol{y}$ and passes through the origin. However, if we have $\neq 0$, then we have

$$\boldsymbol{x}^\top \boldsymbol{y} = \alpha \implies \boldsymbol{x}_0 = \alpha \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_p}.$$

Note that $\boldsymbol{x}_0$ is parallel to $\boldsymbol{y}$. Then we can rewrite the equation as

$$\boldsymbol{y}^\top (\boldsymbol{x} - \boldsymbol{x}_0) = 0.$$

Geometrically, $\boldsymbol{x} - \boldsymbol{x}_0$ represents the vectors on the that are shifted $away(towards \boldsymbol{y} if 0 and away from \boldsymbol{y} otherwise) anda$

Question. What is the distance (margin) between the two separating hyperplanes $\boldsymbol{w}^\top \boldsymbol{x} + b = 1$ and $\boldsymbol{w}^\top \boldsymbol{x} + b = -1$?

Answer. $\frac{2}{\|\boldsymbol{w}\|_2}$. (why?)

7 Orthogonality and Orthonormality

7.1 Orthogonal Vectors

Definition 24 (Orthogonal)**.** Two vectors $\boldsymbol{x}, \boldsymbol{y}$ in an inner product space $\mathcal{X}$ are **orthogonal** if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$, i.e., $\boldsymbol{x} \perp \boldsymbol{y}$.

Definition 25 (Mutually Orthogonal)**.** Nonzero vectors $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)}$ are said to be **mutually orthogonal** if $\left\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \right\rangle = 0$ whenever $i \neq j$. In other words, each vector is orthogonal to all other vectors in the collection.

Proposition 26. Mutually orthogonal vectors are linearly independent.

Proof. Suppose for the sake of contradiction that $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)}$ are orthogonal but linearly dependent vectors. Then this implies that there exist scalars $\alpha_1 \ldots, \alpha_d$ that are not all identically zero, such that

$$\sum_{i=1}^{d} \alpha_i \boldsymbol{x}^{(i)} = 0.$$

□

Taking the linear product of both sides of this equation with $\boldsymbol{x}^{(j)}$ for $j = 1, \ldots, d$, we have

$$\left\langle \sum_{i=1}^{d} \alpha_i \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \right\rangle = 0.$$

Since

$$\left\langle \sum_{i=1}^{d} \alpha_i \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \right\rangle = 0,$$

this means that $\alpha_i = 0$ for all $i = 1, \ldots, d$, hence a contradiction.

Definition 27 (Orthonormal)**.** A collection of vectors $S = \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)}\right\}$ is **orthonormal** if, for $i, j = 1, \ldots, d$

$$\left\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \right\rangle = \begin{cases} 0 & \text{if } i \neq j; \\ 1 & \text{if } i = j, \end{cases}$$

i.e., S is orthonormal if every element has **unit norm**, and all elements are **orthogonal** to each other. A collection of orthonormal vectors S forms an **orthonormal basis** for the span of S.

7.2 Orthogoanl Complement

Definition 28 (Orthogonal Complement)**.** The set of vectors in $\mathcal{X}$ that are orthogonal to $\mathcal{S}$ is called the **orthogonal complement** of $\mathcal{S}$, denoted by $S^\perp$.

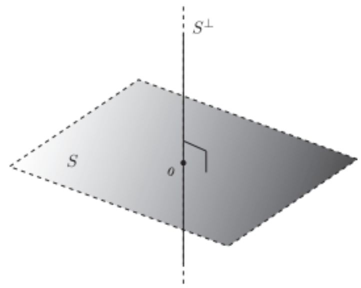

Figure 2.1: Orthogonal complement of $\mathcal{S}$.

Theorem 29 (Orthogonal Decomposition)**.** If $\mathcal{S}$ is a subspace of an inner product space $\mathcal{X}$, then any vector $\boldsymbol{x} \in \mathcal{X}$ can be written in an **unique** way as the sum of an element in $\mathcal{S}$ and one in the orthogonal complement $\mathcal{S}^{\perp}$:

$$\mathcal{X} = \mathcal{S} \oplus \mathcal{S}^{\perp}$$

for any subspace $\mathcal{S} \subseteq \mathcal{X}$.

Proof. □

7.3 Projections

Definition 30 (Projection)**.** Given a vector $\boldsymbol{x}$ in an inner product space $\mathcal{X}$ and a closed set $\mathcal{S} \subseteq \mathcal{X}$, the projection of $\boldsymbol{x}$ onto $\mathcal{S}$, denoted as $\Pi_{\mathcal{S}}(\boldsymbol{x})$, is defined as the point in $\mathcal{S}$ at minimal distance from $\boldsymbol{x}$:

$$\Pi_{\mathcal{S}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{y} \in \mathcal{S}} \|\boldsymbol{y} - \boldsymbol{x}\|,$$

called **Euclidean projection**.

Theorem 31 (Projection Theorem)**.** Let $\mathcal{X}$ be an inner product space, let $\boldsymbol{x}$ be a given element in $\mathcal{X}$, and let $\mathcal{S}$ be a subspace of $\mathcal{X}$. Then, there exists a unique vector $\boldsymbol{x}^* \in \mathcal{S}$ which is solution to the problem

$$\min_{\boldsymbol{y} \in \mathcal{S}} \|\boldsymbol{y} - \boldsymbol{x}\|$$

Moreover, a necessary and sufficient condition for $\boldsymbol{x}^*$ being the optimal solution for this problem is that

$$\boldsymbol{x}^* \in \mathcal{S}, \quad (\boldsymbol{x} - \boldsymbol{x}^*) \perp \mathcal{S}.$$

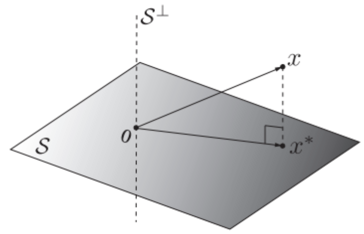

Figure 2.2: Projection onto a subspace.

Proof. □

Theorem 32 (Projection on affine set)**.** Let $\mathcal{X}$ be an inner product space, let x be a given element in $\mathcal{X}$, and let $\mathcal{A} = x^{(0)} + \mathcal{S}$ be the affine set obtained by translating a given subspace $\mathcal{S}$ by a given vector $x^{(0)}$. Then, there exists a unique vector $x^* \in \mathcal{A}$ which is solution to the problem

$$\min_{y \in \mathcal{A}} \|y - x\|$$

Moreover, a necessary and sufficient condition for x^* to be the optimal solution for this problem is that

$$x^* \in \mathcal{A}, \quad (x - x^*) \perp \mathcal{S}.$$

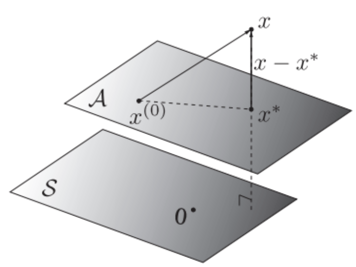

Figure 2.3: Projection on affine set.

Proof. □

7.3.1 Euclidean projection of a point onto a line

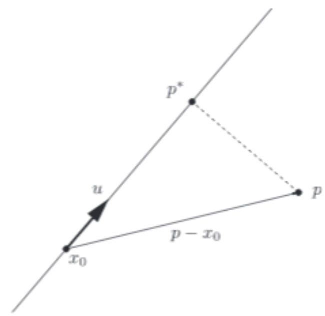

Figure 2.4: Euclidean projection of a point onto a line.

Let $\boldsymbol{p} \in \mathbb{R}^n$ be a given point. We want to compute the Euclidean projection $\boldsymbol{p}^*$ of $\boldsymbol{p}$ onto a line $L = \{\boldsymbol{x}_0 + \text{span}(\boldsymbol{u})\}$, where $\|\boldsymbol{u}\|_2 = 1$:

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{x} \in L} \|\boldsymbol{x} - \boldsymbol{p}\|_2.$$

Since any point $\boldsymbol{x} \in L$ can be written as $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{v}$, for some $\boldsymbol{v} \in \text{span}(\boldsymbol{u})$, the problem is equivalent to finding a value $\boldsymbol{v}^*$ such that

$$\boldsymbol{v}^* = \arg \min_{\boldsymbol{v} \in \text{span}(\boldsymbol{u})} \|\boldsymbol{v} - (\boldsymbol{p} - \boldsymbol{x}_0)\|_2.$$

7.3.2 Euclidean projection of a point onto an hyperplane

A **hyperplane** is an affine set defined as

$$H = \left\{\boldsymbol{z} \in \mathbb{R}^n \mid \boldsymbol{a}^\top \boldsymbol{z} = b\right\}$$

where $\boldsymbol{a} \neq 0$ is called a **normal direction** of the hyperplane, since for any two vectors $\boldsymbol{z}_1, \boldsymbol{z}_2 \in H$ it holds that $(\boldsymbol{z}_1 - \boldsymbol{z}_2) \perp \boldsymbol{a}$.

Our goal is that given $\boldsymbol{p} \in \mathbb{R}^n$ we want to determine the Euclidean projection $\boldsymbol{p}^*$ of $\boldsymbol{p}$ onto H.

The projection theorem requires $\boldsymbol{p} - \boldsymbol{p}^*$ to be orthogonal to H. Since $\boldsymbol{a}$ is a direction orthogonal to H, the condition $(\boldsymbol{p} - \boldsymbol{p}^*) \perp H$ is equivalent to saying that $\boldsymbol{p} - \boldsymbol{p}^* = \alpha\boldsymbol{a}$, for some $\alpha \in \mathbb{R}$.

To find α, consider that $\boldsymbol{p}^* \in H$, thus $\boldsymbol{a}^\top \boldsymbol{p}^* = b$, then consider the optimality condition

$$\boldsymbol{p} - \boldsymbol{p}^* = \alpha a$$

and multiply it on the left by $\boldsymbol{a}^\top$, obtaining

$$\boldsymbol{a}^\top \boldsymbol{p} - b = \alpha\|\boldsymbol{a}\|^2$$

whereby

$$\alpha = \frac{\boldsymbol{a}^\top \boldsymbol{p} - b}{\|\boldsymbol{a}\|_2^2}$$

and

$$\boldsymbol{p}^* = \boldsymbol{p} - \frac{\boldsymbol{a}^\top \boldsymbol{p} - b}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a}.$$

The distance from $\boldsymbol{p}$ to H is

$$\|\boldsymbol{p} - \boldsymbol{p}^*\|_2 = |\alpha| \cdot \|\boldsymbol{a}\|_2 = \frac{\left|\boldsymbol{a}^\top \boldsymbol{p} - b\right|}{\|\boldsymbol{a}\|_2}.$$

7.3.3 Projection on a vector span

Suppose we have a basis for a subspace $\mathcal{S} \subseteq \mathcal{X}$, that is

$$\mathcal{S} = \operatorname{span}\left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)}\right).$$

Given $\boldsymbol{x} \in \mathcal{X}$, the Projection Theorem states that the unique projection $\boldsymbol{x}^*$ of $\boldsymbol{x}$ onto $\mathcal{S}$ is characterized by $(\boldsymbol{x} - \boldsymbol{x}^*) \perp \mathcal{S}$.

Since $\boldsymbol{x}^* \in \mathcal{S}$, we can write $\boldsymbol{x}^*$ as some (unknown) linear combination of the elements in the basis of $\mathcal{S}$, that is

$$\boldsymbol{x}^* = \sum_{i=1}^{d} \alpha_i \boldsymbol{x}^{(i)}$$

Then $(\boldsymbol{x} - \boldsymbol{x}^*) \perp \mathcal{S} \Leftrightarrow \left\langle \boldsymbol{x} - \boldsymbol{x}^*, \boldsymbol{x}^{(k)} \right\rangle = 0, k = 1, \ldots, d:$

$$\sum_{i=1}^{d} \alpha_i \left\langle \boldsymbol{x}^{(k)}, \boldsymbol{x}^{(i)} \right\rangle = \left\langle \boldsymbol{x}^{(k)}, \boldsymbol{x} \right\rangle, \quad k = 1, \ldots, d$$

Solving this system of linear equations (Gram equations) provides the coefficients α, and hence the desired $\boldsymbol{x}^*$.

8 Functions and Maps

Definition 33 (Function)**.** A **function** takes a vector argument in $\mathbb{R}^n$, and returns a unique value in $\mathbb{R}$. We write

$$f : \mathbb{R}^n \to \mathbb{R}.$$

Definition 34 (Domain)**.** The **domain** of a function f, denoted f, is defined as the set of points where the function is finite.

Definition 35 (Map)**.** **Maps** are functions that return a vector of values. We write

$$f : \mathbb{R}^n \to \mathbb{R}^m.$$

8.1 Sets related to functions

Definition 36 (Graph)**.** The **graph** of f is the set of input-output pairs that f can attain, that is:

$$f = \left\{(\boldsymbol{x}, f(\boldsymbol{x})) \in \mathbb{R}^{n+1} \mid \boldsymbol{x} \in \mathbb{R}^n\right\}$$

Definition 37 (Epigraph)**.** The **epigraph**, denoted f, describes the set of input-output pairs that f can achieve, as well as *anything above*:

$$f = \left\{(\boldsymbol{x}, t) \in \mathbb{R}^{n+1} \mid \boldsymbol{x} \in \mathbb{R}^n, t \geq f(\boldsymbol{x})\right\}.$$

Definition 38 (Level Set)**.** A **level set** (or **contour line**) is the set of points that achieve exactly some value for the function f. For $t \in \mathbb{R}$, the t-level set of the function f is defined as

$$C_f(t) = \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = t\}.$$

Definition 39 (t-sublevel set)**.** The ***t*-sublevel set** of f is the set of points that achieve at most a certain value for f :

$$L_f(t) = \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leq t\}.$$

8.2 Linear and Affine Functions

Definition 40 (Linear)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is **linear** if and only if

- $\forall \boldsymbol{x} \in \mathbb{R}^n, \alpha \in \mathbb{R}$, $f(\alpha \boldsymbol{x}) = \alpha f(\boldsymbol{x})$;
- $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$, $f(\boldsymbol{x}_1 + \boldsymbol{x}_2) = f(\boldsymbol{x}_1) + f(\boldsymbol{x}_2)$.

Definition 41 (Affine)**.** A function f is **affine** if and only if the function $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{0})$ is linear (affine = linear + constant). In addition, f is affine if and only if it can be expressed as

$$f(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} + b,$$

for some unique pair $(\boldsymbol{a}, b)$ where $\boldsymbol{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

For any affine function f, we can obtain $\boldsymbol{a}$ and b as follows:

$$b = f(\boldsymbol{0}),$$

$$a_i = f(\boldsymbol{e}_i) - b, \quad \text{for } i = 1, \ldots, n.$$

9 Hyperplanes and Halfspaces

Definition 42 (Hyperplane)**.** A **hyperplane** in $\mathbb{R}^n$ is a set of the form

$$\mathcal{H} = \left\{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{a}^\top \boldsymbol{x} = b\right\},$$

where $\boldsymbol{a} \in \mathbb{R}^n, \boldsymbol{a} \neq \boldsymbol{0}$, and $b \in \mathbb{R}$ are given.

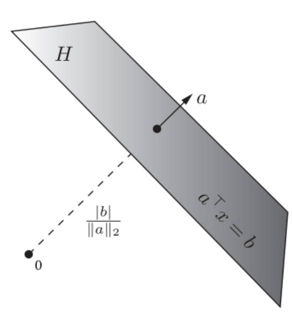

Figure 2.5: Hyperplane.

Definition 43 (Halfspace)**.** A hyperplane $\mathcal{H}$ separates the whole space in two regions called **halfspaces** ($\mathcal{H}_{-}$ is a **closed halfspace**, $\mathcal{H}$ is an **open halfspace**).

$$\mathcal{H}_{-}=\left\{\boldsymbol{x} \mid \boldsymbol{a}^{\top} \boldsymbol{x} \leq b\right\}, \quad \mathcal{H}_{++}=\left\{\boldsymbol{x} \mid \boldsymbol{a}^{\top} \boldsymbol{x}>b\right\}.$$

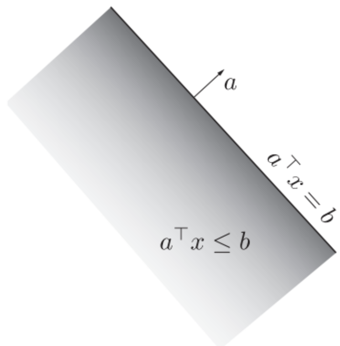

Figure 2.6: Halfspace.

10 Gradients

Definition 44 (Gradient)**.** The **gradient** of a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ at a point $\boldsymbol{x}$ where f is differentiable, denoted with $\nabla f(\boldsymbol{x})$, is a column vector of first derivatives of f with respect to $x_1, \ldots, x_n$

$$\nabla f(\boldsymbol{x})=\begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix}^{\top}$$

An affine function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, represented as $f(\boldsymbol{x})=\boldsymbol{a}^{\top} \boldsymbol{x}+b$, has a very simple gradient: $\nabla f(\boldsymbol{x})=\boldsymbol{a}$.

Example 45. The distance function $\rho(x)=\|\boldsymbol{x}-\boldsymbol{p}\|_2=\sqrt{\sum_{i=1}^n\left(x_i-p_i\right)^2}$ has gradient

$$\nabla \rho(\boldsymbol{x})=\frac{1}{\|\boldsymbol{x}-\boldsymbol{p}\|_2}(\boldsymbol{x}-\boldsymbol{p}).$$

10.1 Affine approximation of non-linear functions

A non-linear function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ can be approximated locally via an affine function, using a **first-order Taylor series expansion**:

Theorem 46 (First-order Taylor Series Expansion)**.** If f is differentiable at point $\boldsymbol{x}_0$, then for all points $\boldsymbol{x}$ in a neighborhood of $\boldsymbol{x}_0$, we have that

$$f(\boldsymbol{x})=f\left(\boldsymbol{x}_0\right)+\nabla f\left(\boldsymbol{x}_0\right)^{\top}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)+\epsilon(\boldsymbol{x})$$

where the error term $\epsilon(\boldsymbol{x})$ goes to zero faster than first order, as $\boldsymbol{x} \rightarrow \boldsymbol{x}_0$, that is

$$\lim_{\boldsymbol{x} \rightarrow \boldsymbol{x}_0} \frac{\epsilon(\boldsymbol{x})}{\left\|\boldsymbol{x}-\boldsymbol{x}_0\right\|_2}=0$$

In practice, this means that for $\boldsymbol{x}$ sufficiently close to $\boldsymbol{x}_0$, we can write the approximation

$$\boxed{f(\boldsymbol{x}) \simeq f\left(\boldsymbol{x}_0\right)+\nabla f\left(\boldsymbol{x}_0\right)^{\top}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)}.$$

10.2 Geometric interpretation of the gradient

Geometrically, the gradient of f at a point $\boldsymbol{x}_0$ is a vector $\nabla f(\boldsymbol{x}_0)$ perpendicular to the contour line of f at level $\alpha = f(\boldsymbol{x}_0)$, pointing from $\boldsymbol{x}_0$ outwards the α-sublevel set (i.e., it points towards higher values of the function).

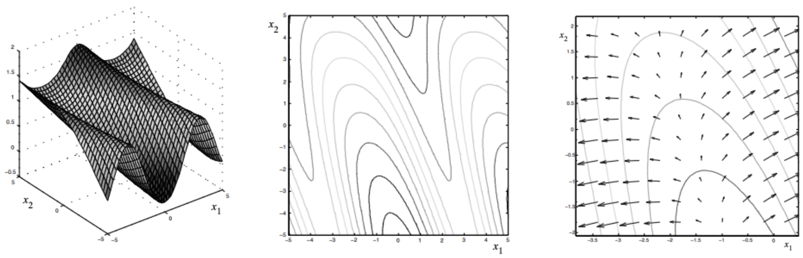

Figure 2.7: *Left.* Graph of a function. *Center.* Its contour lines. *Right.* Gradient vectors (arrows) at some grid points.

The gradient $\nabla f\left(\boldsymbol{x}_0\right)$ also represents the direction along which the function has the **maximum rate of increase** (**steepest ascent direction**).

Let $\boldsymbol{v}$ be a unit direction vector (i.e., $\|\boldsymbol{v}\|_2 = 1$), let $\epsilon \geq 0$, and consider moving away at distance ϵ from $\boldsymbol{x}_0$ along direction $\boldsymbol{v}$, that is, consider a point $\boldsymbol{x} = \boldsymbol{x}_0 + \epsilon\boldsymbol{v}$. We have that

$$f\left(\boldsymbol{x}_0+\epsilon \boldsymbol{v}\right) \simeq f\left(\boldsymbol{x}_0\right)+\epsilon \nabla f\left(\boldsymbol{x}_0\right)^{\top} \boldsymbol{v}, \text{ for } \epsilon \rightarrow 0,$$

equivalently,

$$\lim_{\epsilon \rightarrow 0} \frac{f\left(\boldsymbol{x}_0+\epsilon \boldsymbol{v}\right)-f\left(\boldsymbol{x}_0\right)}{\epsilon}=\nabla f\left(\boldsymbol{x}_0\right)^{\top} \boldsymbol{v}.$$

Whenever $\epsilon > 0$ and v is such that $\nabla f\left(\boldsymbol{x}_0\right)^{\top} \boldsymbol{v} > 0$, then f is increasing along the direction $\boldsymbol{v}$, for small ϵ.

Remark. The inner product $\nabla f\left(\boldsymbol{x}_0\right)^{\top} \boldsymbol{v}$ measures the rate of variation of f at $\boldsymbol{x}_0$, along direction $\boldsymbol{v}$, and it is called the **directional derivative** of f along $\boldsymbol{v}$.

If $\boldsymbol{v}$ is orthogonal to $\nabla f(\boldsymbol{x}_0)$, the rate of variation is zero: along such a direction the function value remains constant. Contrary, the rate of variation is maximal when $\boldsymbol{v}$ is parallel to $\nabla f\left(\boldsymbol{x}_0\right)$, hence along the normal direction to the contour line at $\boldsymbol{x}_0$.

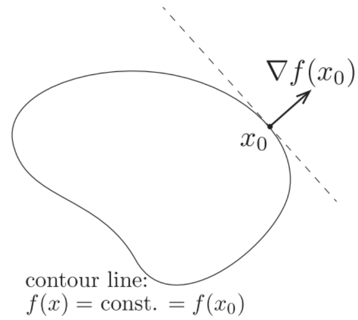

Figure 2.8: The gradient $\nabla f(\boldsymbol{x}_0)$ is normal to the contour line of f at $\boldsymbol{x}_0$, and defines the direction of maximum increase rate.

3

Matrices and Linear Maps

11 Matrix Basics

Definition 47 (Matrix)**.** A **matrix** is a collection of numbers, arranged in columns and rows in a tabular format:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

where m is the number of rows and n is the number of columns. If A contains only real elements, we write $A \in \mathbb{R}^{m,n}$ and $A \in \mathbb{C}^{m,n}$ if A contains complex elements.

Definition 48 (Transpose)**.** The **transposition** operation is defined as

$$A^\top_{ij} = A_{ji},$$

where A_{ij} is the element of A positioned in row i and column j.

11.1 Matrix Products

Definition 49 (Matrix Multiplication)**.** Two matrices can be multiplied if conformably sized, i.e., if $A \in \mathbb{R}^{m,n}$ and $B \in \mathbb{R}^{n,p}$, then the matrix product $AB \in \mathbb{R}^{m,p}$ is defined as a matrix whose (i, j)-th entry is

$$AB_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

Remark. The matrix product is **non-commutative**, i.e., $AB \neq BA$.

Definition 50 (Identity Matrix)**.** The $n \times n$ **identity matrix** (denoted I_n, or I), is a matrix with all zero elements, except for the elements on the diagonal, which are equal to one. This matrix satisfies $AI_n = A$ for every matrix A with n columns, and $I_nB = B$ for every matrix B with n rows.

11.2 Matrix-vector Product

Definition 51 (Matrix-vector Product)**.** Let $A \in \mathbb{R}^{m,n}$ be a matrix with columns $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^m$ and $\boldsymbol{b} \in \mathbb{R}^n$ a vector. The **matrix-vector product** is defined as

$$A\boldsymbol{b} = \sum_{k=1}^{n} \boldsymbol{a}_k b_k, \quad A \in \mathbb{R}^{m,n}, \boldsymbol{b} \in \mathbb{R}^n$$

which is a linear combination of the columns of A, using the elements in $\boldsymbol{b}$ as coefficients.
Similarly, we can multiply matrix $A \in \mathbb{R}^{m,n}$ on the left by (the transpose of) vector $\boldsymbol{c} \in \mathbb{R}^m$ as follows:

$$\boldsymbol{c}^\top A = \sum_{k=1}^{m} c_k \alpha_k^\top, \quad A \in \mathbb{R}^{m,n}, \boldsymbol{c} \in \mathbb{R}^m$$

forming a linear combination of the rows α_k of A, using the elements in $\boldsymbol{c}$ as coefficients.

11.3 Matrix Representations

A matrix $A \in \mathbb{R}^{m,n}$ can be expressed in the following two forms:

$$A = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_n \end{bmatrix}, \text{ or } A = \begin{bmatrix} \boldsymbol{\alpha_1}^\top \\ \boldsymbol{\alpha_2}^\top \\ \vdots \\ \boldsymbol{\alpha_m}^\top \end{bmatrix},$$

where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^m$ denote the columns of A, and $\boldsymbol{\alpha_1}^\top, \ldots, \boldsymbol{\alpha_m}^\top \in \mathbb{R}^n$ denote the rows of A.

AB can be written as

$$AB = A\begin{bmatrix} \boldsymbol{b}_1 & \ldots & \boldsymbol{b}_p \end{bmatrix} = \begin{bmatrix} A\boldsymbol{b}_1 & \ldots & A\boldsymbol{b}_p \end{bmatrix}.$$

In other words, AB results from transforming each column $\boldsymbol{b}_i$ of B into $A\boldsymbol{b}_i$.
Similarly, we can also write

$$AB = \begin{bmatrix} \boldsymbol{\alpha_1}^\top \\ \vdots \\ \boldsymbol{\alpha_m}^\top \end{bmatrix} B = \begin{bmatrix} \boldsymbol{\alpha_1}^\top B \\ \vdots \\ \boldsymbol{\alpha_m}^\top B \end{bmatrix}.$$

Finally, the product AB can be given the interpretation as the sum of so-called **dyadic** matrices (matrices of rank one, of the form $\boldsymbol{a}_i \boldsymbol{\beta_i}^\top$, where $\boldsymbol{\beta_i}^\top$ denote the rows of B :

$$AB = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{\beta_i}^\top, \quad A \in \mathbb{R}^{m,n}, B \in \mathbb{R}^{n,p}.$$

For any two conformably sized matrices A, B, it holds that

$$(AB)^\top = B^\top A^\top.$$

Then for a generic chain of n products, we have

$$(A_1 A_2 \cdots A_p)^\top = A_p^\top \cdots A_2^\top A_1^\top.$$

12 Matrices as linear maps

We can interpret matrices as linear maps (vector-valued functions), or **operators**, acting from an **input** space to an **output** space.

Recall that a map $f : \mathcal{X} \to \mathcal{Y}$ is **linear** if any points $\boldsymbol{x}$ and $\boldsymbol{z}$ in $\mathcal{X}$ and any scalars λ, μ satisfy $f(\lambda \boldsymbol{x} + \mu \boldsymbol{z}) = \lambda f(x) + \mu f(\boldsymbol{z})$.
Any linear map $f : \mathbb{R}^n \to \mathbb{R}^m$ can be represented by a matrix $A \in \mathbb{R}^{m,n}$, mapping input vectors $\boldsymbol{x} \in \mathbb{R}^n$ to output vectors $\boldsymbol{y} \in \mathbb{R}^m$:

$$y = Ax.$$

Figure 3.1: Linear map defined by a matrix A.

Affine maps are simply linear functions plus a constant term, thus any affine map $f : \mathbb{R}^n \to \mathbb{R}^m$ can be represented as

$$f(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b},$$

for some $A \in \mathbb{R}^{m,n}, \boldsymbol{b} \in \mathbb{R}^m$.

12.1 Range, rank, and nullspace

Definition 52 (Range)**.** The **range** of a matrix A is defined as

$$\mathcal{R}(A) = \{A\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^n\},$$

which is a subspace.

Definition 53 (Rank)**.** The **rank** of $\mathcal{R}(A)$, denoted by rank(A), is the **dimension** of A, which is the number of linearly independent columns of A.

Remark. The rank is also equal to the number of linearly independent rows of A; that is,

$$\text{rank}(A) = \text{rank}(A^\top).$$

Thus,

$$1 \leq \text{rank}(A) \leq \min(m, n).$$

Definition 54 (Nullspace)**.** The **nullspace** of a matrix A, denoted $\mathcal{N}(A)$ is defined as:

$$\mathcal{N}(A) = \{\boldsymbol{x} \in \mathbb{R}^n \mid A\boldsymbol{x} = \boldsymbol{0}\},$$

which is also a subspace.

Corollary 55. $\mathcal{R}(A^\top)$ and $\mathcal{N}(A)$ are mutually orthogonal subspaces, i.e., $\mathcal{N}(A) \perp \mathcal{R}(A^\top)$.

Corollary 56.

$$\mathbb{R}^n = \mathcal{N}(A) \oplus \mathcal{N}(A)^\perp = \mathcal{N}(A) \oplus \mathcal{R}(A^\top).$$

Theorem 57 (Fundamental Theorem of Linear Algebra)**.** For any given matrix $A \in \mathbb{R}^{m,n}$, it holds that $\mathcal{N}(A) \perp \mathcal{R}\left(A^{\top}\right)$ and $\mathcal{R}(A) \perp \mathcal{N}\left(A^{\top}\right)$, hence

$$\mathcal{N}(A) \oplus \mathcal{R}\left(A^{\top}\right) = \mathbb{R}^n$$
$$\mathcal{R}(A) \oplus \mathcal{N}\left(A^{\top}\right) = \mathbb{R}^m.$$

Consequently, we can decompose any vector $\boldsymbol{x} \in \mathbb{R}^n$ as the sum of two vectors orthogonal to each other, one in the range of $A^{\top}$, and the other in the nullspace of A :

$$\boldsymbol{x} = A^{\top}\xi + \boldsymbol{z}, \quad \boldsymbol{z} \in \mathcal{N}(A)$$

Similarly, we can decompose any vector $\boldsymbol{w} \in \mathbb{R}^m$ as the sum of two vectors orthogonal to each other, one in the range of A, and the other in the nullspace of $A^{\top}$:

$$\boldsymbol{w} = A\boldsymbol{\varphi} + \boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \in \mathcal{N}\left(A^{\top}\right).$$

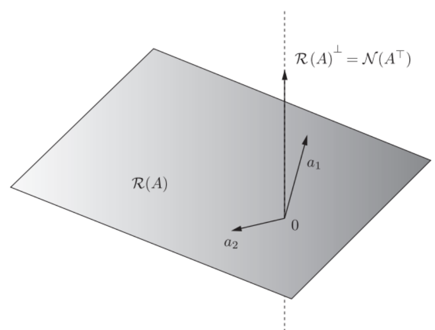

Figure 3.2: Illustration of the fundamental theorem of linear algebra in $\mathbb{R}^3$. Here, $A = [\boldsymbol{a}_1 \boldsymbol{a}_2]$. Any vector in $\mathbb{R}^3$ can be written as the sum of two orthogonal vectors, one in the range of A, the other in the nullspace of $A^{\top}$.

13 Determinants

Definition 58 (Determinants)**.** The **determinant** of a generic (square) matrix $A \in \mathbb{R}^{n,n}$ can be computed by defining $\det a = a$ for a scalar a, and then applying the following inductive formula (**Laplace's determinant expansion**):

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{(i,j)},$$

where i is any row, chosen at will, and $A_{(i,j)}$ denotes a $(n-1) \times (n-1)$ submatrix of A obtained by eliminating row i and column j from A.

$$A \in \mathbb{R}^{n,n} \text{ is singular } \iff \det A = 0 \iff \mathcal{N}(A) \text{ is not equal to } \{0\}.$$

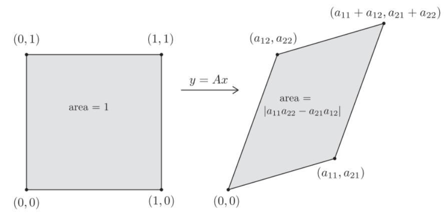

Figure 3.3: Linear mapping of the unit square. The absolute value of the determinant equals the area of the transformed unit square.

For any square matrices $A, B \in \mathbb{R}^{n,n}$ and scalar α:

$$\begin{aligned} \det A &= \det A^\top \\ \det AB &= \det BA = \det A \det B \\ \det \alpha A &= \alpha^n \det A. \end{aligned}$$

13.1 Matrix Inverses

If $A \in \mathbb{R}^{n,n}$ is **nonsingular** (i.e., $\det A \neq 0$), then the inverse matrix A^{-1} is defined as the unique $n \times n$ matrix such that

$$AA^{-1} = A^{-1}A = I_n.$$

If A, B are **square** and **nonsingular**, then

$$(AB)^{-1} = B^{-1}A^{-1}.$$

If A is **square** and **nonsingular**, then

$$\begin{aligned} (A^\top)^{-1} &= (A^{-1})^\top \\ \det A = \det A^\top &= \frac{1}{\det A^{-1}}. \end{aligned}$$

For a generic matrix $A \in \mathbb{R}^{m,n}$, a **generalized inverse** (**pseudoinverse**) can be defined:

13.2 Similar Matrices

Definition 59 (Similar)**.** Two matrices $A, B \in \mathbb{R}^{n,n}$ are **similar** if there exist a nonsingular matrix $P \in \mathbb{R}^{n,n}$ such that

$$B = P^{-1}AP.$$

13.3 Eigenvalues and Eigenvectors

Definition 60 (Eigenvalue/Eigenvector)**.** $\lambda \in \mathbb{C}$ is an **eigenvalue** of matrix $A \in \mathbb{R}^{n,n}$, and $\boldsymbol{u} \in \mathbb{C}^n$ is a corresponding **eigenvector**, if it holds that

$$A\boldsymbol{u} = \lambda \boldsymbol{u}, \quad \boldsymbol{u} \neq 0,$$

or equivalently, $(\lambda I_n - A)\boldsymbol{u} = 0, \boldsymbol{u} \neq 0.$

Definition 61 (Characteristic Polynomial)**.** Eigenvalues can be characterized as those real or complex numbers that satisfy the equation

$$p(\lambda) \doteq \det\left(\lambda I_n - A\right) = 0$$

where $p(\lambda)$ is a polynomial of degree n in λ, known as the **characteristic polynomial** of A

Any matrix $A \in \mathbb{R}^{n,n}$ has n eigenvalues $\lambda_i, i = 1, \ldots, n$, counting multiplicities. To each distinct eigenvalue $\lambda_i, i = 1, \ldots, k$, there corresponds a whole subspace $\phi_i \doteq \mathcal{N}\left(\lambda_i I_n - A\right)$ of eigenvectors associated to this eigenvalue, called the eigenspace.

4

Convex Sets

14 Affine Sets

Definition 62 (Affine Set)**.** A set $C \subseteq \mathbb{R}^n$ is **affine** if the line through any two distinct points in C lies in C, i.e. if for any $x_1, x_2 \in C$ and $\theta \in \mathbb{R}$, we have $\theta x_1 + (1-\theta)x_2 \in C$.

Definition 63 (Affine Combination)**.** A point of the form

$$\theta_1 x_1 + \cdots + \theta_k x_k$$

where $\sum_{i=1}^k \theta_i = 1$ is an **affine combination** of the points $x_1, \ldots, x_k$.

Definition 64 (Affine Hull)**.** The set of all affine combinations of points in some set $C \subseteq \mathbb{R}^n$ is called the **affine hull** of C, denoted **aff** C:

$$\textbf{aff } C = \left\{ \sum_{i=1}^k \theta_i x_i \mid x_1, \ldots, x_k \in C, \sum_{i=1}^k \theta_i = 1 \right\}.$$

15 Convex Sets

Definition 65 (Convex Set)**.** A set C is **convex** if for any $x_1, x_2 \in C$ and any $\theta \in [0,1]$:

$$\theta x_1 + (1-\theta)x_2 \in C,$$

i.e., the **line segment** between any two points in C lies in C.

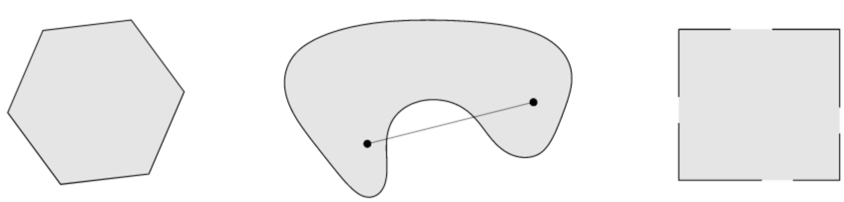

Figure 4.1: Convex and nonconvex sets. *Left.* Convex. *Middle.* Not convex as the line segment between the two points in the set is not contained. *Right.* Not convex as it contains some boundary points but not other.

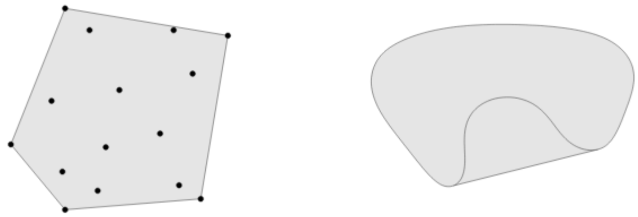

Figure 4.2: *Left.* The convex hull of a set of 15 points is the pentagon. *Right.* The convex hull of the kidney shped set is the shaded set.

Remark. Subspaces, halfspaces, and affine sets, such as lines and hyperplanes are convex, as they contain the entire line between any two distinct points in it, and thus also the line segment in $\mathbb{R}^2$. Hence, **every affine set is convex. However, not every convex set is affine.**

Definition 66 (Convex Combination)**.** A point of the form

$$\theta_1 x_1 + \cdots + \theta_k x_k$$

where $\sum_{i=1}^k \theta_i = 1$ and $\theta_i \geq 0$ is a **convex combination** of the points $x_1, \ldots, x_k$.

Definition 67 (Convex Hull)**.** The **convex hull** of a set C, denoted **conv** C, is the set of all convex combinations of points in C:

$$\textbf{conv } C = \left\{ \sum_{i=1}^k \theta_i x_i \mid x_i \in C, \theta_i \geq 0, i = 1, \ldots, k, \sum_{i=1}^k \theta_i = 1 \right\}.$$

Remark. conv C is always convex and it is the smallest convex set that contains C, i.e., if B is any convex set that contains C, then **conv** $C \subseteq B$.

16 Cones

Definition 68 (Cone)**.** A set C is a **cone** if $\theta x \in C$ for every $x \in C$ and $\theta \geq 0$.

Definition 69 (Convex Cone)**.** A set C is a **convex cone** if it is convex and a cone. Mathematically, it means that for any $x_1, x_2 \in C$ and $\theta_1, \theta_2 \geq 0$, we have

$$\theta_1 x_1 + \theta_2 x_2 \in C.$$

Definition 70 (Conic Combination)**.** A point of the form

$$\theta_1 x_2 + +\theta_k x_k$$

with $\theta_1, \ldots, \theta_k \geq 0$ is a **conic combination** of $x_1 \ldots, x_k$.

Definition 71 (Conic Hull)**.** The **conic hull** of a set C is the set of all conic combinations of points in C, i.e.,

$$\left\{ \sum_{i=1}^k \theta_i x_i \mid x_i \in C, \theta_i \geq 0, i = 1, \ldots, k \right\},$$

which is also the smallest convex cone that contains C.

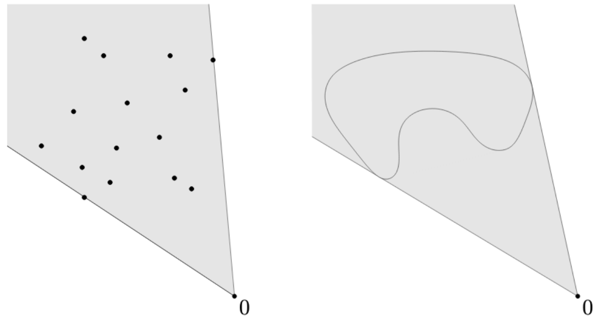

Figure 4.3: The conic hulls of the two sets of figure 1.2.

17 Hyperplanes and Halfspaces

Definition 72 (Hyperplane)**.** A **hyperplane** is a set of the form

$$\{\boldsymbol{x} \mid a^\top \boldsymbol{x} = b\},$$

where $a \in \mathbb{R}^n, a \neq 0$, and $b \in \mathbb{R}$, i.e., the solutions set of a nontrivial linear equation among the components of $\boldsymbol{x}$ (and hence an affine set).

Geometric interpretation: The hyperplane is a set of points with a constant inner product to a given vector a, which can also be viewed as a **normal vector**; the constant b determines the offset from the origin.

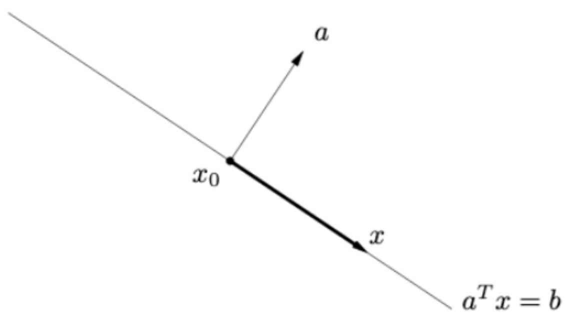

Figure 4.4: Hyperplane in $\mathbb{R}^2$, with normal vector a. $x - x_0$ (arrow) is orthogonal to a for any x in the hyperplane.

A hyperplane divides $\mathbb{R}^n$ into two **halfspaces**, defined as follows:

Definition 73 (Halfspace)**.** A **halfspace** is a set of the form

$$\{\boldsymbol{x} \mid a^\top \boldsymbol{x} \leq b\},$$

where $a \neq 0$, i.e., the solution set of a nontrivial linear inequality.

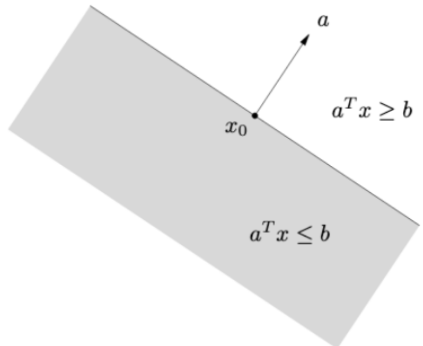

Figure 4.5: The halfspace determined by $a^\top \boldsymbol{x} \leq b$ (shaded) extends in the direction $-a$.

Remark. Halfspaces are **convex**, but not affine.

Proof. Let $\boldsymbol{x}_1, \boldsymbol{x}_2$ be two points in a halfspace. Then for any $\theta \in [0, 1]$, we have

$$\begin{aligned} a^\top(\theta x_1 + (1-\theta)x_2) &= \theta a^\top x_1 + (1-\theta)a^\top x_2 \\ &\leq \theta b + (1-\theta)b \\ &= b. \end{aligned}$$

Thus, halfspaces are convex. □

18 Operations preserving convexity

18.1 Intersection

Theorem 74. If $C_1, \ldots, C_m$ are convex sets, then their intersection

$$C = \bigcap_{i}^{m} C_i$$

is also a convex set.

Proof. Let $\{C_i\}_{i=1}^m$ be convex sets. For any $x_1, x_2 \in \bigcap_{i=1}^m C_i, \theta \in [0, 1]$, $x_1 \in C_i$ and $x_2 \in C_i$ implies

$$\theta x_1 + (1-\theta)x_2 \in C_i$$

for $i = 1, 2, \ldots, m$ by convexity of C_i. Hence,

$$\theta x_1 + (1-\theta)x_2 \in \bigcap_{i=1}^{m} C_i.$$

Thus, $\bigcap_{i=1}^m C_i$ is convex. □

Remark. This also holds for possibly infinite families of convex sets.