Optimization for Machine Learning

Ch1: Preliminaries

Dongyoung Lim
UNIST

AI51101, IE55101

Vector Spaces

- A **field** $\mathbb{F}$ is a set where basic arithmetic operations (addition, subtraction, multiplication, and division) are well defined.
- Here, "well-defined" means:
  - **Closure:** For all $a, b \in \mathbb{F}$, $a + b \in \mathbb{F}$ and $a \cdot b \in \mathbb{F}$.
  - **Identity:** There exist $0, 1 \in \mathbb{F}$ such that
  $$a + 0 = a, \quad a \cdot 1 = a.$$
  - **Inverses:** For each $a \in \mathbb{F}$, there exists $-a$ with $a + (-a) = 0$; for $a \neq 0$, there exists a^{-1} with $a \cdot a^{-1} = 1$.
  - **Subtraction and division:** defined using inverses,
  $$a - b := a + (-b), \quad a/b := a \cdot b^{-1},\ b \neq 0.$$
- Examples: $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Q}$.
- Counterexample: $\mathbb{Z}$ is not a field (e.g., 2 has no multiplicative inverse).
- In this course, we focus on $\mathbb{F} = \mathbb{R}$.

Vector Spaces

- **Exercise:** Is $\mathbb{R}^+$ a field?

Vector Spaces

Definition (Vector Space)

A vector space over a scalar field $\mathbb{F}$ is a set V such that:

(i) For $x, y \in V$, $x + y \in V$ (closed under addition),

(ii) For $x \in V$, $c \in \mathbb{F}$, $cx \in V$ (closed under scalar multiplication).

- Another name for a vector space is a *linear space*.
- Examples:
  - The set of all d-tuples of real numbers,

$$\mathbb{R}^d = \{(x_1, \ldots, x_d) : x_1, \ldots, x_d \in \mathbb{R}\}.$$

  - The set of all real-valued functions defined on $[0, 1]$,

$$\mathcal{F}[0, 1] = \{f : [0, 1] \to \mathbb{R}\}.$$

Vector Spaces

- **Exercise:** Let $C[0,1]$ be the set of all continuous real-valued functions defined on $[0,1]$. Is $C[0,1]$ a vector space?

Vector Spaces

Span

If A is a nonempty subset of a vector space V, then the **linear span** of A, denoted by $\operatorname{span}(A)$, is the set of all finite linear combinations of elements of A:

$$\operatorname{span}(A) := \left\{ \sum_{n=1}^{N} c_n x_n : N > 0,\ x_n \in A,\ c_n \in \mathbb{F} \right\}.$$

We say that A **spans** V if $\operatorname{span}(A) = V$.

Linear Independence

A nonempty subset A of a vector space V is **linearly independent** if, for any finite choice of distinct vectors $x_1, \ldots, x_N \in A$,

$$\sum_{n=1}^{N} c_n x_n = 0 \quad \Leftrightarrow \quad c_1 = \cdots = c_N = 0.$$

- **Exercise 1:** In $\mathbb{R}^3$, determine whether the following pairs of vectors are linearly independent.
  - $x_1 = (1, 2, 3),\ x_2 = (2, 4, 6)$

    Answer: They are linearly dependent, because

    $$2x_1 - x_2 = 0,$$

    so there exists a nontrivial linear combination ($c_1 = 2$, $c_2 = -1$) that gives the zero vector.
  - $x_1 = (1, 0, 0),\ x_2 = (0, 1, 0)$

    Answer: They are linearly independent, because if

    $$c_1x_1 + c_2x_2 = 0,$$

    then $c_1 = c_2 = 0$ is the only solution.
- **Exercise 2:** Let $\mathcal{M} = \{x^k\}_{k=0}^{\infty}$. Show that $\mathcal{M}$ is a linearly independent subset.
- **Exercise 3:** Let $\mathcal{P}$ be the set of all polynomials. Show that $\operatorname{span}(\mathcal{M}) = \mathcal{P}$.

Vector Spaces

- A **basis** $\mathcal{B}$ for a vector space V is a set of vectors that is both linearly independent and spans V:
  (i) $\mathcal{B}$ is linearly independent,
  (ii) $\operatorname{span}(\mathcal{B}) = V$.
- **Example:** the set of monomials $\{1, x, x^2, \dots\}$ is a basis for the space of polynomials $\mathcal{P}$.
- If $\mathcal{B}$ has finitely many elements, say $\mathcal{B} = \{x_1, \dots, x_d\}$, we call this number d the **dimension** of V, written $\dim(V) = d$.
- If V has a basis consisting of infinitely many elements, we say that V is **infinite-dimensional**, i.e., $\dim(V) = \infty$.

Vector Spaces

- **Exercise:** Give an example of an infinite-dimensional vector space.
- **Answer:** The set $C[0,1]$ of all continuous real-valued functions on $[0,1]$ is an infinite-dimensional vector space. Why? For any $n \geq 0$, the monomials $1, x, x^2, \ldots, x^n$ are linearly independent in $C[0,1]$. However, we can always take one more function, namely x^{n+1}, which is not contained in the span of $\{1, x, \ldots, x^n\}$. This means that no finite set of vectors can span $C[0,1]$. Therefore $C[0,1]$ cannot have a finite basis, and so its dimension must be infinite.

Metric Spaces

- A **metric** on a set V is a function that assigns a distance to each pair of elements of V.

Definition (Metric Space)

Let V be a nonempty set. A metric on V is a function $d : V \times V \to \mathbb{R}$ such that for all $x, y, z \in V$:

(a) $d(x,y) \geq 0$ (non-negativity),

(b) $d(x,y) = 0 \iff x = y$,

(c) $d(x,y) = d(y,x)$ (symmetry),

(d) $d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality).

A **metric space** is a set endowed with a metric d.

- For $\mathbb{R}^d$, common metrics include:

$$d_1(x,y) = \sum_{i=1}^{d} |x_i - y_i| \quad (L^1 \text{ metric}),$$

$$d_2(x,y) = \sqrt{\sum_{i=1}^{d} |x_i - y_i|^2} \quad (\text{Euclidean } / \ L^2 \text{ metric}).$$

Metric Spaces

- **Exercise:** Show that d_1 and d_2 are valid metrics on $\mathbb{R}^d$.

Normed Vector Spaces

- A **norm** assigns to each vector $x \in V$ a length $\|x\|$.

Definition (Normed Vector Space)

Let V be a vector space. A norm on V is a function $\|\cdot\| : V \to \mathbb{R}$ such that for all $c \in \mathbb{R}$ and $x, y \in V$:

(a) $\|x\| \geq 0$ and $\|x\| = 0 \iff x = 0$,

(b) $\|cx\| = |c| \, \|x\|$,

(c) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality).

A vector space V together with a norm $\|\cdot\|$ is called a **normed vector space**.

- Examples on $\mathbb{R}^d$:
  (a) ℓ^1-norm: $\|x\|_1 = \sum_{i=1}^d |x_i|$,
  (b) ℓ^2-norm: $\|x\|_2 = \sqrt{\sum_{i=1}^d |x_i|^2}$ (Euclidean norm),
  (c) ℓ^∞-norm: $\|x\|_\infty = \max\{|x_1|, \ldots, |x_d|\}$.

Normed Vector Spaces

- Every normed linear space is automatically a metric space, with the metric defined by

$$d(u, v) = \|u - v\|.$$

  This is called the **natural metric** induced by the norm.

Normed Vector Spaces

- Conversely, is every metric space a normed space? If we define

$$\|u\| = d(u, 0),$$

does this always give a norm?

- **No!** Not every metric can be induced by a norm. For example, the *discrete metric*

$$d(x, y) = \begin{cases} 0, & x = y, \\ 1, & x \neq y \end{cases}$$

is a metric, but the induced function $\|x\| = d(x, 0)$ fails homogeneity since for $x \neq 0$, $\|2x\| = \|x\| \neq 2\|x\|$ in general.

Normed Vector Spaces

Definition (ℓ^p space)

For $1 \le p < \infty$, the ℓ^p space is defined as

$$\ell^p = \left\{ x = (x_1, x_2, \dots) : \sum_{i=1}^{\infty} |x_i|^p < \infty \right\},$$

with norm

$$\|x\|_p = \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{1/p}.$$

For $p = \infty$,

$$\ell^\infty = \{x = (x_1, x_2, \dots) : \sup_i |x_i| < \infty\}, \quad \|x\|_\infty = \sup_i |x_i|.$$

- ℓ^2 is the space of square-summable sequences, with norm

$$\|x\|_2 = \left(\sum_{i=1}^{\infty} |x_i|^2 \right)^{1/2}.$$

Normed Vector Spaces

Definition (L^p space)

For $1 \leq p < \infty$, the L^p space is defined as

$$L^p(X) = \left\{ f : X \to \mathbb{R} \ : \ \int_X |f(x)|^p \, d\mu(x) < \infty \right\},$$

with norm

$$\|f\|_p = \left(\int_X |f(x)|^p \, d\mu(x) \right)^{1/p}.$$

where μ is a measure (e.g., Lebesgue measure, probability measure). For $p = \infty$,

$$L^\infty(X) = \{ f : X \to \mathbb{R} \ : \ \sup_{x \in X} |f(x)| < \infty \}, \quad \|f\|_\infty = \sup_{x \in X} |f(x)|.$$

- $L^2(X)$ is the space of square-integrable functions, with norm

$$\|f\|_2 = \left(\int_X |f(x)|^2 \, d\mu(x) \right)^{1/2}.$$

Equivalent Norms

- Suppose that $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on a vector space V. We say that these norms are **equivalent** if there exist constants $C_1, C_2 > 0$ such that for all $x \in V$,

$$C_1\|x\|_a \;\le\; \|x\|_b \;\le\; C_2\|x\|_a.$$

- **Exercise:** In $\mathbb{R}^d$, any two ℓ^p norms with $1 \le p < \infty$ are equivalent.

Theorem

If V is a finite-dimensional vector space, then any two norms on V are equivalent.

- Recall: The dimension of a vector space is the number of vectors in a basis, i.e., the minimum number of linearly independent vectors needed to span the space. For example, $\{e_1, e_2, \ldots, e_n\}$ with e_i the i-th standard unit vector is a basis of $\mathbb{R}^n$.

Equivalent Norms

- In $L^p[0,1]$ spaces, the norms
$$\|f\|_1 = \int_0^1 |f(x)|\,dx, \quad \|f\|_2 = \left(\int_0^1 |f(x)|^2\,dx\right)^{1/2}$$
are not equivalent.
- Example: Let $f_n(x) = \mathbf{1}_{[0,1/n]}(x)$, the indicator of the interval $[0,1/n]$.
$$\|f_n\|_1 = \int_0^{1/n} 1\,dx = \tfrac{1}{n}, \quad \|f_n\|_2 = \left(\int_0^{1/n} 1^2\,dx\right)^{1/2} = \tfrac{1}{\sqrt{n}}.$$
The ratio
$$\frac{\|f_n\|_2}{\|f_n\|_1} = \sqrt{n} \;\to\; \infty \quad (n \to \infty).$$
Hence, there do not exist constants $C_1, C_2 > 0$ such that
$$C_1\|f\|_1 \le \|f\|_2 \le C_2\|f\|_1$$
for all $f \in L^1[0,1] \cap L^2[0,1]$.
Therefore, $L^1[0,1]$ and $L^2[0,1]$ norms are not equivalent.

Convergence

Definition (Convergence of a sequence)

A sequence $(x_n)_{n\in\mathbb{N}}$ **converges** to x if for all $\varepsilon > 0$ there exists an integer $N > 0$ such that for all $n \geq N$,

$$d(x_n, x) < \varepsilon.$$

In this case, we write $x_n \to x$ or $\lim_{n\to\infty} x_n = x$.

Definition (Cauchy sequence)

A sequence $(u_n)_{n\in\mathbb{N}}$ is a **Cauchy sequence** if for all $\varepsilon > 0$ there exists an integer $N > 0$ such that for all $m, k > N$,

$$d(u_m, u_k) < \varepsilon.$$

- Every convergent sequence is a Cauchy sequence (in any metric space).
- Not every Cauchy sequence converges: counterexample in $\mathbb{Q}$ (approximations of $\sqrt{2}$).

Banach Spaces

- A metric space X is **complete** if every Cauchy sequence in X converges to a point in X.
  - Example: $\mathbb{R}$ is complete, but $\mathbb{Q}$ is not. e.g., decimal approximations of $\sqrt{2}$.
- A **Banach space** is a normed linear space that is complete with respect to its natural metric.
  - Example: $C[a, b]$, the space of continuous functions on $[a, b]$, is a Banach space with the sup norm

$$\|f\|_\infty = \sup_{x \in [a,b]} |f(x)|.$$

Mode of Convergence: Pointwise Convergence

- Consider a sequence of functions $f_n : X \to \mathbb{R}$ $(n = 1, 2, 3, \ldots)$ and a function $f : X \to \mathbb{R}$.

Definition (Pointwise Convergence)

We say that f_n converges **pointwise** to f on X if

$$\forall x \in X, \quad \lim_{n\to\infty} f_n(x) = f(x).$$

In this case, we write $f_n \to f$ pointwise.

Mode of Convergence: Uniform Convergence

- Consider a sequence of functions $f_n : X \to \mathbb{R}$ $(n = 1, 2, 3, \ldots)$ and a function $f : X \to \mathbb{R}$.
- The **sup norm** (also called the uniform norm) is defined by

$$\|f\|_\infty = \sup_{x \in X} |f(x)|.$$

Definition (Uniform Convergence)

We say that f_n converges **uniformly** to f on X if

$$\lim_{n \to \infty} \|f_n - f\|_\infty = 0,$$

equivalently,

$$\sup_{x \in X} |f_n(x) - f(x)| \;\to\; 0 \quad \text{as } n \to \infty.$$

In this case, we write $f_n \to f$ uniformly.

Mode of Convergence: L^p Convergence

- Consider a sequence of measurable functions $f_n : X \to \mathbb{R}$ ($n = 1, 2, 3, \ldots$) and a function $f : X \to \mathbb{R}$.
- For $1 \leq p < \infty$, the L^p **norm** is defined by

$$\|f\|_p = \left(\int_X |f(x)|^p \, d\mu(x) \right)^{1/p},$$

where μ is a measure on X (e.g., Lebesgue measure, probability measure).

Definition (L^p Convergence)

We say that f_n converges to f in L^p if

$$\lim_{n\to\infty} \|f_n - f\|_p = 0.$$

In this case, we write $f_n \to f$ in L^p.

- Intuition: L^p convergence measures how small the *average p-th power error* is.

Mode of Convergence: Exercise

- Define the sequence of functions $f_n : [0,1] \to \mathbb{R}$ by

$$f_n(x) = \begin{cases} 0, & x = 0, \\ 2nx, & 0 < x \leq \frac{1}{2n}, \\ 2 - 2nx, & \frac{1}{2n} < x \leq \frac{1}{n}, \\ 0, & \frac{1}{n} < x \leq 1, \end{cases}$$

  and $f(x) = 0$.
- **Exercise 1:** Show that $f_n \to f$ pointwise.
- **Exercise 2:** Show that $f_n \not\to f$ uniformly.
- **Exercise 3:** Show that $f_n \to f$ in L^p with $1 \leq p < \infty$.

Modes of Convergence

- Uniform convergence always implies pointwise convergence. However, the converse does not hold in general.
- For a function $f : X \to \mathbb{R}$, recall
$$\|f\|_\infty = \sup_{x \in X} |f(x)|.$$
This value can be infinite; in fact, $\|f\|_\infty < \infty$ if and only if f is bounded on X.
- If $K \subset \mathbb{R}^d$ is closed and bounded (compact), define
$$C(K) := \{f : K \to \mathbb{R} \ : \ f \text{ is continuous on } K\}.$$
Then $\| \cdot \|_\infty$ is a well-defined norm on $C(K)$, since continuous functions on compact sets are bounded.

Modes of Convergence

- **Uniform limit theorem:** If (f_n) converges uniformly to f with each $f_n \in C(K)$, then f is also continuous.
- **Consequence:** Every Cauchy sequence in $(C(K), \|\cdot\|_\infty)$ converges to a continuous function in $C(K)$. Hence, $(C(K), \|\cdot\|_\infty)$ is a **Banach space**.

Inner Product Space

- In a normed vector space, each vector has a length. An inner product refines this structure by also defining the angle between vectors.
- In particular, the inner product allows us to determine whether two vectors are perpendicular (orthogonal).

Definition (Inner Product)

An inner product $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ on a vector space V is a function satisfying, for all $u, v, w \in V$ and $\alpha \in \mathbb{R}$:

- Symmetry: $\langle u, v \rangle = \langle v, u \rangle$,
- Linearity in the first argument: $\langle \alpha u + v, w \rangle = \alpha \langle u, w \rangle + \langle v, w \rangle$,
- Positive-definiteness: $\langle u, u \rangle \geq 0$ with equality iff $u = 0$.

A vector space with an inner product is called an **inner product space**.

Inner Product Space

Theorem (Cauchy–Schwarz Inequality)

For any two vectors u, v in an inner product space V, we have

$$|\langle u, v\rangle| \leq \|u\| \, \|v\|,$$

with equality if and only if u and v are linearly dependent.

- We can always define
$$\|x\| = \langle x, x\rangle^{1/2}, \quad x \in V.$$
Thus, an inner product space is automatically a normed linear space with metric
$$d(u, v) = \|u - v\| = \langle u - v, u - v\rangle^{1/2}.$$
The norm $\| \cdot \|$ is called the **norm induced by the inner product**.

Inner Product Space

- **Exercise 1:** Using the definition of the inner product, compute

$$\langle u, \alpha v \rangle, \quad \langle u, v + w \rangle, \quad \langle u, 0 \rangle.$$

- **Exercise 2:** Show that

$$\|x + y\|^2 = \|x\|^2 + 2\langle x, y \rangle + \|y\|^2.$$

Bounded Linear Functionals and Dual Space

- A **linear functional** on a normed vector space $(X, \|\cdot\|)$ is a map $f : X \to \mathbb{R}$ such that

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y), \quad \forall x, y \in X,\ \alpha, \beta \in \mathbb{R}.$$

- The functional f is **bounded** (or continuous) if there exists $C > 0$ such that

$$|f(x)| \leq C\|x\| \quad \forall x \in X.$$

- The set of all bounded linear functionals on X is called the **dual space** of X, denoted X^*.

Bounded Linear Functionals and Dual Space

- Let $(X, \|\cdot\|)$ be a normed vector space. The **dual norm** of a vector $x \in X$ is defined by

$$\|x\|_* = \sup\{\, |f(x)| : f \in X^*,\ \|f\| \le 1\},$$

  where X^* is the dual space (the set of all bounded linear functionals on X). Thus X^* itself is a normed space with this norm.
- **Intuition:** Dual norm is the size of a vector $x \in X$ measured from the perspective of the dual space.
- If X is an inner product space, then by the Riesz representation theorem every $f \in X^*$ can be written as

$$f(x) = \langle x, y\rangle \quad \text{for some } y \in X.$$

  In this case,

$$\|x\|_* = \sup\{\, |\langle x, y\rangle| : \|y\| \le 1\}.$$

- **Fact**: In $\mathbb{R}^p$ with the $\|\cdot\|_p$ norm, the dual norm is the $\|\cdot\|_q$ norm, where

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Bounded Linear Functionals and Dual Space

- Consider the optimization problem

$$\max_{\|\delta\|_p \le \epsilon} \langle g, \delta \rangle,$$

  where $g \in \mathbb{R}^d$ is fixed.
- Recall the definition of the dual norm:

$$\|g\|_q = \sup_{\|x\|_p \le 1} \langle g, x \rangle, \qquad \frac{1}{p} + \frac{1}{q} = 1.$$

- For any δ with $\|\delta\|_p \le \epsilon$, write $\delta = \epsilon x$ with $\|x\|_p \le 1$. Then

$$\max_{\|\delta\|_p \le \epsilon} \langle g, \delta \rangle = \epsilon \sup_{\|x\|_p \le 1} \langle g, x \rangle = \epsilon \|g\|_q.$$

- **Interpretation:** The dual norm $\|g\|_q$ quantifies the maximum alignment of g with any vector inside the p-norm unit ball.

Hilbert Spaces

Definition (Hilbert Space)

An inner product space that is complete in its induced norm is called a **Hilbert space**.

- Example 1 (finite-dimensional vector space): $\mathbb{R}^d$ with the dot product

$$\langle x, y \rangle = \sum_{i=1}^{d} x_i y_i$$

  is a Hilbert space.

- Example 2 (function space): $L^2[a, b]$ with inner product

$$\langle f, g \rangle = \int_a^b f(t) g(t)\, dt$$

  is a Hilbert space with norm $\|f\|_2 = \langle f, f \rangle^{1/2}$.

- Example 3 (probability space): Let $(\Omega, \mathcal{F}, P)$ be a probability space. Define

$$L^2(\Omega) = \{X : \Omega \to \mathbb{R} \;\; : \;\; \mathbb{E}[X^2] < \infty\}.$$

  With inner product $\langle X, Y \rangle = \mathbb{E}[XY]$, $L^2(\Omega)$ is a Hilbert space of square-integrable random variables.

Orthogonality

Definition (Orthogonality)

Two vectors u, v are **orthogonal** (perpendicular) if $\langle u, v \rangle = 0$.

- An **orthogonal set**: every pair of distinct vectors is orthogonal.
- An **orthonormal set**: orthogonal and each vector has unit norm.

- By the Cauchy–Schwarz inequality, we can define the angle θ_{uv} between u and v:

$$\cos \theta_{uv} = \frac{\langle u, v \rangle}{\|u\| \|v\|}.$$

Orthogonal Projection

Definition (Orthogonal Projection)

Let M be a closed subspace of a Hilbert space H. For $x \in H$, the unique vector $p \in M$ closest to x is called the orthogonal projection of x onto M:

$$p = \arg\min_{z \in M} \|x - z\|.$$

- Properties:
  - **Characterization:** $p \in M$ is the projection of x iff $x - p \perp M$, that is $\langle x - p, z \rangle = 0$ for all $z \in M$.
  - **Pythagorean identity:** $\|x\|^2 = \|p\|^2 + \|x - p\|^2$.

Orthogonal Projection

- **Exercise:** Let $a, b \in \mathbb{R}^n$. Compute the orthogonal projection of a onto the subspace $\operatorname{span}\{b\}$.

Basic Calculus: Gradient and Hessian

Definition (Gradient)

For $f : \mathbb{R}^n \to \mathbb{R}$, the gradient at x is

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}.$$

Definition (Hessian)

For $f : \mathbb{R}^n \to \mathbb{R}$, the Hessian at x is

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

Basic Calculus: Gradient and Hessian

- **Example:** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be

$$f(x, y) = x^2 y + y^3.$$

- Gradient:

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} 2xy \\ x^2 + 3y^2 \end{bmatrix}.$$

- Hessian:

$$\nabla^2 f(x, y) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2y & 2x \\ 2x & 6y \end{bmatrix}.$$

Basic Calculus: Jacobian

Definition (Jacobian)

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ with $f = (f_1, \ldots, f_m)$. The **Jacobian** of f at x is the $m \times n$ matrix

$$J_f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

Chain Rule

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^k$. For the composition $h = g \circ f : \mathbb{R}^n \to \mathbb{R}^k$, we have

$$J_h(x) = J_g(f(x))\, J_f(x).$$

Basic Calculus: Jacobian

- Let
$$f : \mathbb{R}^2 \to \mathbb{R}^2, \quad f(x,y) = \begin{bmatrix} x^2 + y \\ e^x + y \end{bmatrix},$$
and
$$g : \mathbb{R}^2 \to \mathbb{R}, \quad g(u,v) = u + v^2.$$
Define $h = g \circ f : \mathbb{R}^2 \to \mathbb{R}$.
- **Jacobian:**
$$J_f(x,y) = \begin{bmatrix} 2x & 1 \\ e^x & 1 \end{bmatrix}, \quad J_g(u,v) = \begin{bmatrix} 1 & 2v \end{bmatrix}.$$
- **Chain Rule:** Substitute $f(x,y) = (x^2 + y, e^x + y)$:
$$J_g(f(x,y)) = \begin{bmatrix} 1 & 2(e^x + y) \end{bmatrix}.$$
Thus,
$$J_h(x,y) = \begin{bmatrix} 1 & 2(e^x + y) \end{bmatrix} \begin{bmatrix} 2x & 1 \\ e^x & 1 \end{bmatrix} = \begin{bmatrix} 2x + 2e^x(e^x + y) & 1 + 2(e^x + y) \end{bmatrix}.$$

Definition (Directional Derivative)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at x, and let u be a unit vector. The directional derivative of f at x in the direction u is

$$\nabla_u f(x) = \lim_{h \to 0} \frac{f(x + hu) - f(x)}{h}.$$

Proposition

The directional derivative can be computed as

$$\nabla_u f(x) = \langle \nabla f(x), u \rangle.$$

Basic Calculus: Directional Derivative

- **Example:** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be

$$f(x, y) = x^2 + xy + y^2.$$

Compute the directional derivative of f at $(1, 2)$ in the direction

$$u = \tfrac{1}{\sqrt{2}}(1, 1).$$

- Gradient:

$$\nabla f(x, y) = \begin{bmatrix} 2x + y \\ x + 2y \end{bmatrix}, \quad \nabla f(1, 2) = \begin{bmatrix} 4 \\ 5 \end{bmatrix}.$$

- Directional derivative:

$$\nabla_u f(1, 2) = \langle \nabla f(1, 2), u \rangle = \left\langle \begin{bmatrix} 4 \\ 5 \end{bmatrix}, \tfrac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\rangle = \tfrac{9}{\sqrt{2}}.$$

Basic Calculus: Positive Semi-Definite Matrices

- A symmetric matrix $M \in \mathbb{R}^{d \times d}$ is **positive semi-definite (PSD)** if

$$x^\top M x \geq 0 \quad \text{for all } x \in \mathbb{R}^d.$$

Notation: $M \succeq 0$.

- Caution: A PSD matrix does *not* mean all entries of M are non-negative.

Theorem (Characterization of PSD matrices)

A symmetric matrix $M \in \mathbb{R}^{d \times d}$ is PSD if and only if all its eigenvalues λ_i are non-negative.

Basic Calculus: Positive Semi-Definite Matrices

- **Exercise:** Find a matrix with all positive entries that is not positive semi-definite.

Basic Calculus: Taylor Expansion

Theorem (One-dimensional Taylor Expansion)

For a smooth function $f : \mathbb{R} \to \mathbb{R}$, the Taylor expansion of f at $a \in \mathbb{R}$ is

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \mathcal{O}(|x-a|^3).$$

Theorem (Multidimensional Taylor Expansion)

For a smooth function $f : \mathbb{R}^n \to \mathbb{R}$, the Taylor expansion of f at $\mathbf{a} \in \mathbb{R}^n$ is

$$f(\mathbf{x}) = f(\mathbf{a}) + \nabla f(\mathbf{a})^\top(\mathbf{x}-\mathbf{a}) + \tfrac{1}{2}(\mathbf{x}-\mathbf{a})^\top \nabla^2 f(\mathbf{a})(\mathbf{x}-\mathbf{a}) + \mathcal{O}(\|\mathbf{x}-\mathbf{a}\|^3).$$

Basic Probability

- A probability space is $(\Omega, \mathcal{F}, P)$.
- A d-dimensional random variable is a measurable function

$$X : \Omega \to \mathbb{R}^d.$$

- For $S \subset \mathbb{R}^d$,

$$P(X \in S) = P(\{\omega \in \Omega : X(\omega) \in S\}).$$

- If X has pdf f, then

$$P(X \in S) = \int_S f(x)\, dx.$$

- Expectation:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

- Moment generating function (mgf):

$$M_X(t) = \mathbb{E}[e^{tX}].$$

Basic Probability

Union bound (Boole's inequality)

For events $E_1, E_2, \ldots$,

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i).$$

Markov's inequality

For nonnegative X and $a > 0$,

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Chebyshev's inequality

If $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$, then for $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Basic Probability

Chernoff bound

For random variable X with mgf $M_X(t)$,

$$P(X \geq a) \leq \inf_{t>0} e^{-ta} M_X(t).$$

Hoeffding's inequality

If $X_1, \ldots, X_n$ are independent with $X_i \in [a, b]$, then for $\epsilon > 0$,

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \geq \epsilon\right) \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

Jensen's inequality

For convex g,

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]).$$

Tower property (Law of total expectation)

For random variables X, Y,

$$\mathbb{E}[X] = \mathbb{E}\big[\,\mathbb{E}[X \mid Y]\,\big].$$

Basic Probability

- In probability and information theory, we often need to measure how *different* two probability distributions P and Q are.
- A common choice is the **Kullback–Leibler (KL) divergence**:

$$\mathrm{KL}(Q\|P) = \int \log\left(\frac{dQ}{dP}\right) dQ,$$

  where Q is absolutely continuous w.r.t. P. (For $A \in \mathcal{F}$, $P(A) = 0$ implies $Q(A) = 0$.)
- Properties:
  - $\mathrm{KL}(Q\|P) \geq 0$ (non-negativity).
  - $\mathrm{KL}(Q\|P) = 0$ if and only if $Q = P$ (a.e.).
  - Not symmetric: in general $\mathrm{KL}(Q\|P) \neq \mathrm{KL}(P\|Q)$.
- KL divergence is not a true metric, but it plays the role of a "distance" between probability measures in many learning bounds.
- Other metrics for probability distributions: Jensen-Shannon divergence, Total variation distance, Wasserstein distance.