Contents

1 **Probability** **2**
1.1 Probability Spaces 2
1.2 Markov Chains 6
1.3 Probability Bounds 9
1.4 Convexity 11

1 Probability

1.1 Probability Spaces

In an elementary level, we have been viewing probability as the quotient between the number of desired outcomes and the number of all possible outcomes. This definition, though intuitive, is not very solid when it comes to an infinite sample space. In this introductory chapter, we would establish the theories of probability using a more modern and rigorous structure.

Definition 1.1.1 ▸ Set Algebra

Let X be a set. A **set algebra** over X is a family $\mathcal{F} \subseteq \mathcal{P}(X)$ such that

- $X \backslash F \in \mathcal{F}$ for all $F \in \mathcal{F}$ (closed under complementation);
- $X \in \mathcal{F}$;
- $X_1 \cup X_2 \in \mathcal{F}$ for any $X_1, X_2 \in \mathcal{F}$ (closed under binary union).

There are several immediate implications from the above definition.

First, by closure under complementation, we know that an algebra over any set X must contain the empty set.

Second, by De Morgan's Law, one can easily check that if the first 2 axioms hold, the closure under binary union is equivalent to

- $X_1 \cap X_2 \in \mathcal{F}$ for any $X_1, X_2 \in \mathcal{F}$;
- $\bigcup_{i=1}^{n} X_i \in \mathcal{F}$ for any $X_1, X_2, \cdots, X_n \in \mathcal{F}$ for all $n \in \mathbb{N}$;
- $\bigcap_{i=1}^{n} X_i \in \mathcal{F}$ for any $X_1, X_2, \cdots, X_n \in \mathcal{F}$ for all $n \in \mathbb{N}$.

$(X, \mathcal{F})$ is known as a *field of sets*, where the elements of X are called *points* and those of $\mathcal{F}$, *complexes* or *admissible sets* of X.

In probability theory, what we are interested in is a special type of set algebras known as *σ-algebras*.

Definition 1.1.2 ▸ σ-Algebra

A **σ-Algebra** over a set A is a non-empty set algebra over A that is closed under countable union.

Of course, by the same argument as above, we known that any σ-algebra is closed under countable intersection as well.

Now, as we all know, we can take some set Ω as a *sample space* and denote an *event* by some subset of Ω. Roughly speaking, we could now define the probability of an event $E \subseteq \Omega$ as the ratio between the sets' volumes. The remaining question now is: how do we define the volume of a set properly?

Definition 1.1.3 ▸ Measure

Let X be a set and Σ be a σ-algebra over X. A **measure** over Σ is a function

$$\mu : \Sigma \to \mathbb{R} \cup \{-\infty, +\infty\}$$

such that

- $\mu(E) \geq 0$ for all $E \in \Sigma$ (non-negativity);
- $\mu(\varnothing) = 0$;
- $\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$ for any countable collection of pairwise disjoint elements of Σ (countable additivity or σ-additivity).

The triple (X, Σ, μ) is known as a **measure space** and the pair (X, Σ), a **measurable space**.

One thing to note here is that if at least one $E \in \Sigma$ has a finite measure, then $\mu(\varnothing) = 0$ is automatically guaranteed for obvious reasons.

Definition 1.1.4 ▸ Probability Space

Let Ω be a sample space and $\mathcal{F}$ be a σ-algebra over Ω. A **probability space** is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathbb{P} : \mathcal{F} \to [0, 1]$, known as a **probability measure**, is such that $\mathbb{P}(\Omega) = 1$.

Obviously, the above definition immediately guarantees that

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;
2. $\mathbb{P}(A) \leq \mathbb{P}(B)$ if $\mathbb{P}(A) \subseteq \mathbb{P}(A)$;
3. $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

The third result follows from a direct application of the principle of inclusion and exclusion.

By induction, one can easily check that

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

for any finitely many events. The following proposition extends this result to countable collections of events:

Proposition 1.1.5 ▸ Union Bound of Countable Collections of Events

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $E_1, E_2, \cdots, E_n, \cdots \in \mathcal{F}$ is any countable sequence of events, then

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

Proof. Define $F_1 \coloneqq E_1$ and $F_k \coloneqq E_k \setminus \bigcup_{i=1}^{k-1} E_i$ for $k \geq 2$. Clearly, the F_i's are pairwise disjoint. By Definition 1.1.2, the F_i's are elements of $\mathcal{F}$. Note that $\mathbb{P}(F_i) \leq \mathbb{E}_{\mathbb{i}}$ for all $i \in \mathbb{N}^+$, so

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) &= \mathbb{P}\left(\bigcup_{i=1}^{\infty} F_i\right) \\ &= \sum_{i=1}^{\infty} \mathbb{P}(F_i) \\ &\leq \sum_{i=1}^{\infty} \mathbb{P}(E_i).\end{aligned}$$

□

Next, we will introduce the notion of *random variables* formally. For this purpose, we first establish the notion of a *Borel algebra*.

Definition 1.1.6 ▸ Borel Algebra

Let X be a topological space. A **Borel set** on X is a set which can be formed via countable union, countable intersection and relative complementation of open sets in X. The smallest σ-algebra over X containing all Borel sets on X is known as the **Borel algebra** over X.

Clearly, the Borel algebra over X contains all open sets in X according to the above axioms from Definition 1.1.2. This helps us define the following:

Definition 1.1.7 ▸ Random Variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{X}, \mathcal{B})$ be a measurable space where $\mathcal{B}$ is the Borel algebra over $\mathcal{X}$. A **random variable** is a function $X : \Omega \to \mathcal{X}$ such that

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

for all $B \in \mathcal{B}$.

Remark. Rigorously, such a random variable X is a *measurable function* or *measurable mapping* from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{B})$.

The probability measure $\mathbb{P}$ thus induces a probability measure P_X over $(\mathcal{X}, \mathcal{B})$.

Definition 1.1.8 ▸ Distribution

Let $X : \Omega \to \mathcal{X}$ be a random variable over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{B}$ be the Borel algebra over $\mathcal{X}$, the **distribution** of X is the probability measure P_X on $(\mathcal{X}, \mathcal{B})$ given by

$$P_X(B) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in B\}\right).$$

Remark. Often times, we write $\Pr(X \in B) = P_X(B)$.

In the context of information theory, we mostly are concerned with real-valued random variables only.

Definition 1.1.9 ▸ Real-Valued Random Variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, a **real-valued random variable** over the space is a mapping $X : \Omega \to \mathbb{R}$ such that

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$.

Note that the Borel set over $\mathbb{R}$ is just the family of all open intervals.

Clearly, if X is a real-valued random variable, we have $\{\omega \in \Omega : X(\omega) > x\} \in \mathcal{F}$. Moreover, we claim that

$$\{\omega \in \Omega : X(\omega) < x\} = \bigcup_{y<x} \{\omega \in \Omega : X(\omega) \leq y\}.$$

The proof is quite straightforward and is left to the reader as an exercise. By Definition

1.1.2, this means that

$$\{\omega \in \Omega : X(\omega) < x\} \cup \{\omega \in \Omega : X(\omega) > x\} \in \mathcal{F}.$$

Therefore, $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$. This argument justifies the probabilities $\Pr(X < x)$ and $\Pr(X = x)$. We give a special name to the range of a random variable in computer science.

Definition 1.1.10 ▸ Alphabet

Let X be a random variable, the range of X is called an **alphabet**, denoted as $\mathcal{X}$.

Recall that we have defined expectations for discrete and continuous random variables in elementary probability theory. In terms of measure theory, the two formulae can be unified as the Lebesgue integral

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega).$$

Note that $\mathbb{E}[X]$ is a real number while $\mathbb{E}[X \mid Y]$ is a **random variable** formed as a function of Y. In a way, Y partitions the sample space into regions where $\mathbb{E}[X \mid Y = y_i]$ gives the expectation of X in the region induced by $Y = y_i$ for each $y_i \in \mathcal{Y}$. In general, the following result holds:

Theorem 1.1.11 ▸ Law of Iterated Expectations

Let X and Y be random variables, then $\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \mathbb{E}[X]$.

The above formula can be interpreted as the fact that $\mathbb{E}[X \mid Y]$ is a best estimator for X.

1.2 Markov Chains

Recall that 2 random variables X and Z are *independent* if and only if $P_{X,Z}(x,z) = P_X(x)P_Z(z)$ or $P_{X|Z}(x \mid z) = P_X(x)$ for all $(x,z) \in \mathcal{X} \times \mathcal{Z}$. We will extend this definition with a third random variable.

Definition 1.2.1 ▸ Markov Chain

Let X, Y, Z be random variables. If

$$P_{X,Y,Z}(x,y,z) = P_X(x)P_{Y|X}(y \mid x)P_{Z|Y}(z \mid y)$$

for all $(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, then we say that X, Y, Z forms a **Markov chain** in this order, or that X and Z are conditionally independent on Y.

Recall also that the *Bayes's Rule* states the following:

Theorem 1.2.2 ▶ Bayes's Rule

For any random variables X and Y,

$$P_{X|Y}(x \mid y) = \frac{P_{Y|X}(y \mid x)P_X(x)}{\sum_{x' \in \mathcal{X}} P_{Y|X}(y \mid x')P_X(x')}.$$

Based on Theorem 1.2.2, we have

$$P_{X,Y}(x,y) = P_{X|Y}(x \mid y)P_Y(y) = P_X(x)P_{Y|X}(y \mid x).$$

By applying the formula repeatedly, we have

$$\begin{aligned} P_{X,Y,Z}(x,y,z) &= P_{X,Y}(x,y)P_{Z|X,Y}(z \mid x,y) \\ &= P_X(x)P_{Y|X}(y \mid x)P_{Z|X,Y}(z \mid x,y). \end{aligned}$$

Therefore, a Markov chain simply states that the distribution of Z is no longer dependent on X, but depends on Y solely. Therefore, this allows us to remove one condition when applying Theorem 1.2.2. Thus, it actually suffices to prove $P_{Z|X,Y} = P_{Z|Y}$ when proving that X-Y-Z forms a Markov chain.

We can denote a Markov chain by X-Y-Z. Intuitively, such a relationship should be symmetric.

Proposition 1.2.3 ▶ Symmetricity of Markov Chains

If X-Y-Z is a Markov chain, then Z-Y-X is also a Markov chain.

Proof. By Definition 1.2.1,

$$P_{X,Y,Z}(x,y,z) = P_X(x)P_{Y|X}(y \mid x)P_{Z|Y}(z \mid y).$$

By Theorem 1.2.2, we have

$$\begin{aligned} P_{X|Y}(x \mid y) &= \frac{P_X(x)P_{Y|X}(y \mid x)}{P_Y(y)} \\ &= \frac{P_{X,Y,Z}(x,y,z)}{P_Y(y)P_{Z|Y}(z \mid y)} \\ &= \frac{P_{X,Y,Z}(x,y,z)}{P_{Z,Y}(z,y)} \\ &= P_{X|Z,Y}(x \mid z,y). \end{aligned}$$

Therefore, Z-Y-X is a Markov chain. □

One obvious case where dependence exists between the random variables in a Markov chain is that one of the random variables is a function of another one.

Proposition 1.2.4 ▶ Markov Chain Involving Functions of a Random Variable

Let X and Y be any random variables and $Z \coloneqq f(Y)$ for some function f, then X-Y-Z is a Markov chain.

Proof. Notice that

$$P_{Z|X,Y}(z \mid x, y) = P_{f(Y)|X,Y}(z \mid x, y) = \begin{cases} 1 & \text{if } z = f(y) \\ 0 & \text{otherwise} \end{cases},$$

$$P_{Z|Y}(z \mid y) = P_{f(Y)|Y}(z \mid y) = \begin{cases} 1 & \text{if } z = f(y) \\ 0 & \text{otherwise} \end{cases}$$

for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Therefore, $P_{Z|X,Y} = P_{Z|Y}$ and so X-Y-Z forms a Markov chain. □

Note that if X and Z are independent, they are naturally conditionally independent given any Y. However, the inverse may not be true.

Proposition 1.2.5 ▶ Conditional Independence Does Not Imply Independence

There exists random variables X, Y, Z such that X and Z are dependent but conditionally independent given Y.

Proof. Let N_1, N_2, N_3 be pairwise independent random variables such that

$$\mathcal{N}_1 = \mathcal{N}_2 = \mathcal{N}_3 = \{0, 1\}.$$

Take $X = N_1 + N_2$, $Y = N_2$ and $Z = N_2 + N_3$. Clearly, X and Z are dependent, but

$$\begin{aligned} P_{Z|X}(z \mid x) &= P_{N_2+N_3|N_1+N_2}(z \mid x) \\ &= P_{N_3|N_1,N_2}(z - y \mid x - y, y) \\ &= P_{N_2+N_3|N_1+N_2,N_2}(z \mid x, y) \\ &= P_{Z|X,Y}(z \mid x, y), \end{aligned}$$

which implies that X and Z are conditionally independent given Y. □

1.3 Probability Bounds

We use various bounds to make estimates and approximations for probability distributions. The first commonly used bound is *Markov's Inequality*.

Theorem 1.3.1 ▶ Markov's Inequality

If X is a non-negative random variable, then $\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$ *for all* $a > 0$.

Proof. It suffices to prove for the continuous case. Notice that

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty x f_X(x)\,\mathrm{d}x \\ &\geq \int_a^\infty x f_X(x)\,\mathrm{d}x \\ &\geq a \int_0^\infty f_X(x)\,\mathrm{d}x \\ &= \Pr(X \geq a). \end{aligned}$$

Therefore, $\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$. □

Note that the bound given by Markov's inequality is a rather loose bound. The following inequality proposes a better bound:

Theorem 1.3.2 ▶ Chebyshev's Inequality

For any real-valued random variable X with finite variance,

$$\Pr\left(|X - \mathbb{E}[X]| > a\sqrt{\mathrm{Var}(X)}\right) \leq \frac{1}{a^2}$$

for all $a > 0$.

Proof. Define $g(X) : (X - \mathbb{E}[X])^2$, which is clearly non-negative. By Theorem 1.3.1, we have

$$\Pr\bigl(g(X) > a^2\mathrm{Var}(X)\bigr) \leq \frac{\mathbb{E}[g(X)]}{a^2\mathrm{Var}(X)}.$$

Note that $\mathbb{E}[g(X)] = \mathrm{Var}(X)$, so

$$\Pr\left(|X - \mathbb{E}[X]| > a\sqrt{\mathrm{Var}(X)}\right) = \Pr\bigl(g(X) > a^2\mathrm{Var}(X)\bigr) \leq \frac{1}{a^2}.$$

□

Finally, we state the following law of large numbers:

Theorem 1.3.3 ▶ Weak Law of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be pairwise independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\operatorname{Var}(X_i) = \sigma^2 \in \mathbb{R}$ for every $i \in \mathbb{N}^+$. For every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0.$$

Proof. Note that $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$ and that

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sum_{i=1}^{n}\operatorname{Var}(X_i)}{n^2} = \frac{\sigma^2}{n}.$$

By Theorem 1.3.2, we have

$$0 \leq \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

By Squeeze Theorem, this clearly implies that

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0.$$

□

Alternatively, we may phrase Theorem 1.3.3 as "$\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ in probability". When a sequence $\{S_n\}_{n=1}^{\infty}$ converges to b in probability, we write $S_n \xrightarrow{\text{p}} b$.

Remark. Essentially, what Theorem 1.3.3 says is that when n is large, the sample mean from n measurements of the same data converges to the expectation of the distribution.

Under some mild conditions, this convergence occurs exponentially fast, i.e., the probability $\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right)$ decreases at least as fast as $\exp(-ng(\epsilon))$ for some real-valued function $g\colon \mathbb{R}^+ \to \mathbb{R}^+$. In terms of asymptotic analysis, we write this as

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \leq \exp(-ng(\epsilon) + o(n)).$$

Equivalently, this means that there exists a function $g\colon \mathbb{R} \to \mathbb{R}$ with $g(\epsilon) > 0$ for every

$\epsilon > 0$ such that

$$\liminf_{n\to\infty} -\frac{1}{n} \log \Pr\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \epsilon \right) \geq g(\epsilon) + o(1).$$

There is a strong version for the law, which shall be stated without proof:

Theorem 1.3.4 ▸ Strong Law of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be pairwise independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 \in \mathbb{R}$ for every $i \in \mathbb{N}^+$, then

$$\Pr\left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu \right) = 1.$$

1.4 Convexity

Recall the following definition:

Definition 1.4.1 ▸ Convex Function

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for any $\lambda \in [0, 1]$ and any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$f\big(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\big) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$

From a graphical perspective, a convex function is an overestimate of all linear functions whose values are bounded above by it. The following proposition set this result in a rigorous context:

Proposition 1.4.2 ▸ Convex Functions as Overestimates for Linear Functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and define

$$\mathcal{L} \coloneqq \left\{ \ell \in \mathrm{Maps}\left(\mathbb{R}^n, \mathbb{R}\right) : \ell(\boldsymbol{u}) = \boldsymbol{a}^{\mathrm{T}} \cdot \boldsymbol{u} + b \leq f(\boldsymbol{u}) \text{ for all } \boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{a} \in \mathbb{R}^n, b \in \mathbb{R} \right\}$$

to be the set of all linear functions bounded above by f, then for each $\boldsymbol{x} \in \mathbb{R}^n$,

$$f(\boldsymbol{x}) = \sup_{\ell \in \mathcal{L}} \ell(\boldsymbol{x}).$$

Proof. It suffices to prove that for all $\boldsymbol{x} \in \mathbb{R}^n$, there exists some linear function $\ell \in \mathcal{L}$

such that $\ell(\boldsymbol{x}) = f(\boldsymbol{x})$. Take any $\boldsymbol{h} \in \mathbb{R}^n$. Since f is convex, we have

$$\begin{aligned} 2f(\boldsymbol{x}) &= 2f\left(\frac{1}{2}(\boldsymbol{x}+\boldsymbol{h}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{h})\right) \\ &\leq f(\boldsymbol{x}+\boldsymbol{h}) + f(\boldsymbol{x}-\boldsymbol{h}). \end{aligned}$$

Therefore, we have

$$L_1 = \lim_{\|\boldsymbol{h}\|\to 0} \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}-\boldsymbol{h})}{\|\boldsymbol{h}\|} \leq \lim_{\|\boldsymbol{h}\|\to 0} \frac{f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x})}{\|\boldsymbol{h}\|} = L_2.$$

Take some $a \in [L_1, L_2]$ and let $\ell(\boldsymbol{y}) = a\|\boldsymbol{y}-\boldsymbol{x}\| + f(\boldsymbol{x})$. Observe that $\ell(\boldsymbol{x}) = f(\boldsymbol{x})$. Take $\boldsymbol{h} = \boldsymbol{y} - \boldsymbol{x}$, then

$$\begin{aligned} \ell(\boldsymbol{y}) &= a\|\boldsymbol{y}-\boldsymbol{x}\| + f(\boldsymbol{x}) \\ &\leq \frac{f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x})}{\|\boldsymbol{h}\|}\|\boldsymbol{y}-\boldsymbol{x}\| + f(\boldsymbol{x}) \\ &= f(\boldsymbol{x}+\boldsymbol{h}) \\ &= f(\boldsymbol{y}). \end{aligned}$$

Therefore, $\ell \in \mathcal{L}$ as desired. □

The following proposition gives a simple test for convexity in one-dimensional case, which is a special case of the Hessian matrix test:

Proposition 1.4.3 ▸ Second Derivative Test for Convexity

If a real-valued function f is twice-differentiable on $[a, b]$, then it is convex if and only if $f''(x) \geq 0$ for all $x \in (a, b)$.

Convex functions produce the following interesting result regarding expectation:

Theorem 1.4.4 ▸ Jensen's Inequality

Let f be a convex function and X be a random variable, then $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.

Proof. Let $\mathcal{L}$ be the set of all linear functions bounded above by f. By Proposition

1.4.2, we have

$$\begin{aligned}\mathbb{E}[f(X)] &= \mathbb{E}\left[\sup_{\ell\in\mathcal{L}} \ell(X)\right] \\ &\geq \sup_{\ell\in\mathcal{L}} \mathbb{E}[\ell(X)] \\ &= \sup_{\ell\in\mathcal{L}} \ell(\mathbb{E}[X]) \\ &= f(\mathbb{E}[X]).\end{aligned}$$

□

Remark. If f is strictly convex, the equality holds if and only if X is constant.