

Lecture 1: Preliminaries I. Probability



School of Mathematical Sciences, Xiamen University

1. Discrete probability

- Ω : the sample space
- The *probability measure* or *probability function* $\mathbb{P}[\omega]$ maps the sample space Ω to the interval $[0, 1]$. This function has the so-called normalization property,

$$\sum_{\omega \in \Omega} \mathbb{P}[\omega] = 1.$$

- If $\mathcal{E} \subset \Omega$ is an event, then

$$\mathbb{P}[\mathcal{E}] = \sum_{\omega \in \mathcal{E}} \mathbb{P}[\omega],$$

namely the probability of an event is the sum of the probabilities of its elements.

- The union of two events:

$$\mathbb{P}[\mathcal{E}_1 \cup \mathcal{E}_2] = \mathbb{P}[\mathcal{E}_1] + \mathbb{P}[\mathcal{E}_2] - \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2].$$

- The union bound

$$\mathbb{P}\left[\bigcup_{i=1}^n \mathcal{E}_i\right] \leq \sum_{i=1}^n \mathbb{P}[\mathcal{E}_i].$$

- Disjoint or mutually exclusive events:

$$\mathcal{E}_1 \cap \mathcal{E}_2 = \emptyset.$$

This can be generalized to any number of events by necessitating that the events are all pairwise disjoint.

- Let $\overline{\mathcal{E}}$ denote the complement of the event \mathcal{E}

$$\mathbb{P}[\overline{\mathcal{E}}] = 1 - \mathbb{P}[\mathcal{E}].$$

1.1 Conditional probability

- For any two events \mathcal{E}_1 and \mathcal{E}_2 , the conditional probability $\mathbb{P}[\mathcal{E}_1|\mathcal{E}_2]$ is the probability that \mathcal{E}_1 occurs given that \mathcal{E}_2 occurs. Formally,

$$\mathbb{P}[\mathcal{E}_1|\mathcal{E}_2] = \frac{\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2]}{\mathbb{P}[\mathcal{E}_2]}.$$

- Bayes rules: for any two events \mathcal{E}_1 and \mathcal{E}_2 such that $\mathbb{P}[\mathcal{E}_1] > 0$ and $\mathbb{P}[\mathcal{E}_2] > 0$,

$$\mathbb{P}[\mathcal{E}_2|\mathcal{E}_1] = \frac{\mathbb{P}[\mathcal{E}_1|\mathcal{E}_2]\mathbb{P}[\mathcal{E}_2]}{\mathbb{P}[\mathcal{E}_1]}.$$

- By $\Omega = \mathcal{E}_2 \cup \overline{\mathcal{E}_2}$, we have

$$\begin{aligned}\mathbb{P}[\mathcal{E}_1] &= \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] + \mathbb{P}[\mathcal{E}_1 \cap \overline{\mathcal{E}_2}] \\ &= \mathbb{P}[\mathcal{E}_1|\mathcal{E}_2]\mathbb{P}[\mathcal{E}_2] + \mathbb{P}[\mathcal{E}_1|\overline{\mathcal{E}_2}]\mathbb{P}[\overline{\mathcal{E}_2}].\end{aligned}$$

1.2 Independent events

- Two events \mathcal{E}_1 and \mathcal{E}_2 are called independent if

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] = \mathbb{P}[\mathcal{E}_1] \cdot \mathbb{P}[\mathcal{E}_2].$$

This can be generalized to more than two events by necessitating that the events are all pairwise independent.

- For any two events \mathcal{E}_1 and \mathcal{E}_2 such that $\mathbb{P}[\mathcal{E}_1] > 0$ and $\mathbb{P}[\mathcal{E}_2] > 0$ the following three statements are equivalent:

$$\mathbb{P}[\mathcal{E}_1|\mathcal{E}_2] = \mathbb{P}[\mathcal{E}_1],$$

$$\mathbb{P}[\mathcal{E}_2|\mathcal{E}_1] = \mathbb{P}[\mathcal{E}_2],$$

and

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] = \mathbb{P}[\mathcal{E}_1] \cdot \mathbb{P}[\mathcal{E}_2].$$

2. Random variables

- Random variables are functions mapping the sample space Ω to the real numbers \mathbb{R} .

Note that even though they are called variables, in reality they are functions.

- Let Ω be the sample space of a random experiment. A formal definition for the random variable X would be as follows: let $\alpha \in \mathbb{R}$ be a real number (not necessarily positive) and note that the function

$$X^{-1}(\alpha) = \{\omega \in \Omega : X(\omega) = \alpha\}$$

returns a subset of Ω and thus is an event. Therefore, the function $X^{-1}(\alpha)$ has a probability.

- We will abuse notation and write $\mathbb{P}[X = \alpha]$ instead of the more proper notation $\mathbb{P}[X^{-1}(\alpha)]$, i.e.,

$$\begin{aligned}\mathbb{P}[X = \alpha] &= \mathbb{P}[X^{-1}(\alpha)] \\ &= \mathbb{P}[\omega \in \Omega : X(\omega) = \alpha].\end{aligned}$$

This function of α is of great interest and it is easy to generalize as follows:

$$\begin{aligned}\mathbb{P}[X \leq \alpha] &= \mathbb{P}[X^{-1}(\beta) : \beta \in (-\infty, \alpha)] \\ &= \mathbb{P}[\omega \in \Omega : X(\omega) \leq \alpha].\end{aligned}$$

- **Independent random variables:** Two random variables X and Y are independent if for all $a, b \in \mathbb{R}$,

$$\mathbb{P}[X = a \text{ and } Y = b] = \mathbb{P}[X = a] \cdot \mathbb{P}[Y = b].$$

2.1 PMF and CDF

- Probability mass function (PMF) measures the probability that a random variable X takes a particular value $\alpha \in \mathbb{R}$:

$$f(\alpha) = \mathbb{P}[X = \alpha].$$

- Cumulative distribution function (CDF) measures the probability that a random variable X takes any value below $\alpha \in \mathbb{R}$:

$$F(\alpha) = \mathbb{P}[X \leq \alpha].$$

It is obvious from the above definitions that

$$F(\alpha) = \sum_{x \leq \alpha} f(x).$$

2.2 Expectation (mean)

- Given a random variable X , its expectation $\mathbb{E}[X]$ is defined as

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x \cdot \mathbb{P}[X = x] = \sum_{\omega \in \Omega} X(\omega) \mathbb{P}[\omega],$$

where $X(\Omega)$ is the image of the random variable X over the sample space Ω . Note that $\mathbb{E}[f(X)] = \sum_{x \in X(\Omega)} f(x) \cdot \mathbb{P}[X = x]$.

- The most important property is linearity of expectation: for any random variables X and Y and real number λ ,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y], \quad \mathbb{E}[\lambda X] = \lambda \mathbb{E}[X].$$

- If two random variables X and Y are independent then we can manipulate the expectation of their product as follows:

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

2.3 Variance and covariance

- Given a random variable X , its variance $\mathbb{V}[X]$ is defined as

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Obviously, it holds $\mathbb{V}[X] \leq \mathbb{E}[X^2]$.

- The covariance of two random variables X and Y is defined as

$$\mathbb{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

X and Y are said to be uncorrelated if $\mathbb{Cov}(X, Y) = 0$.

- If the random variables X and Y are independent, then

$$\mathbb{Cov}(X, Y) = 0 \quad \text{and} \quad \mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y].$$

Also, for any real λ , it holds $\mathbb{V}[\lambda X] = \lambda^2 \mathbb{V}[X]$.

- The standard deviation is the square root of the variance and is often denoted by $\text{Std}(X) = \sqrt{\mathbb{V}\text{ar}(X)}$.

2.4 Markov's inequality

- Let X be a non-negative random variable. For any $\alpha > 0$,

$$\mathbb{P}[X \geq \alpha] \leq \frac{\mathbb{E}[X]}{\alpha}.$$

Proof. For any $\alpha > 0$, define the following function

$$f(X) = \begin{cases} 1, & \text{if } X \geq \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

Then $f(X) \leq X/\alpha$, which yields $\mathbb{E}[f(X)] \leq \mathbb{E}[X]/\alpha$. It follows from

$$\mathbb{E}[f(X)] = 1 \cdot \mathbb{P}[X \geq \alpha] + 0 \cdot \mathbb{P}[X < \alpha] = \mathbb{P}[X \geq \alpha]$$

that

$$\mathbb{P}[X \geq \alpha] \leq \frac{\mathbb{E}[X]}{\alpha}. \quad \square$$

3. Random vectors

- The expectation of an $n \times 1$ random vector, \mathbf{X} , is the vector of expectations of each entry (provided they exist):

$$\mathbb{E}\mathbf{X} = (\mathbb{E}X_1, \dots, \mathbb{E}X_n)^\top.$$

The expectation of a random matrix is also defined as the matrix consisting of the expectations of each entry.

- The variance of \mathbf{X} is defined as the $n \times n$ symmetric matrix:

$$\text{Var}(\mathbf{X}) = \mathbb{E} [(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^\top] = \mathbb{E}(\mathbf{X}\mathbf{X}^\top) - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{X}}^\top =: \boldsymbol{\Sigma},$$

with $\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}\mathbf{X}$. The covariance of two random variables X_i and X_j is the (i, j) entry of $\boldsymbol{\Sigma}$, i.e.,

$$\text{Cov}(X_i, X_j) = \Sigma_{ij}.$$

We also call $\boldsymbol{\Sigma}$ the covariance matrix of \mathbf{X} .

- The covariance (or cross-covariance) of \mathbf{X} with a second $m \times 1$ random vector, \mathbf{Y} , of mean $\boldsymbol{\mu}_Y$ is the $n \times m$ matrix,

$$\mathbb{C}\text{ov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E} [(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{Y} - \boldsymbol{\mu}_Y)^\top] = \mathbb{E}(\mathbf{X}\mathbf{Y}^\top) - \boldsymbol{\mu}_X\boldsymbol{\mu}_Y^\top,$$

and, as in the scalar case,

$$\mathbb{V}\text{ar}(\mathbf{X}) = \mathbb{C}\text{ov}(\mathbf{X}, \mathbf{X}).$$

Note that

$$\mathbb{C}\text{ov}(\mathbf{X}, \mathbf{Y}) = (\mathbb{C}\text{ov}(\mathbf{Y}, \mathbf{X}))^\top.$$

4. Properties of expectation, variance, and covariance

- \mathbb{E} is order preserving:

$$\mathbb{E}X \leq \mathbb{E}Y, \quad \text{if } X \leq Y.$$

- Cauchy-Schwarz inequality:

If X and Y have finite variances, then $|\mathbb{E}(XY)| < \infty$ and

$$|\mathbb{E}(XY)| \leq \mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

In particular,

$$|\text{Cov}(X, Y)| \leq \text{Std}(X)\text{Std}(Y).$$

More generally,

$$|\mathbb{E}(\mathbf{X}^\top \mathbf{Y})| \leq \mathbb{E}|\mathbf{X}^\top \mathbf{Y}| \leq \sqrt{\mathbb{E}(\|\mathbf{X}\|^2)\mathbb{E}(\|\mathbf{Y}\|^2)}.$$

- \mathbb{E} is linear: For any constants a and b ,

$$\mathbb{E}(a\mathbf{X} + b\mathbf{Y}) = a\mathbb{E}\mathbf{X} + b\mathbb{E}\mathbf{Y}.$$

- Cov is bilinear and shift invariant:

For any constants a and b and fixed vector \mathbf{c} ,

$$\text{Cov}(a\mathbf{X} + b\mathbf{Y} + \mathbf{c}, \mathbf{Z}) = a\text{Cov}(\mathbf{X}, \mathbf{Z}) + b\text{Cov}(\mathbf{Y}, \mathbf{Z}),$$

$$\text{Cov}(\mathbf{Z}, a\mathbf{X} + b\mathbf{Y} + \mathbf{c}) = a\text{Cov}(\mathbf{Z}, \mathbf{X}) + b\text{Cov}(\mathbf{Z}, \mathbf{Y}).$$

In particular,

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2\text{Cov}(X, Y),$$

and

$$\text{Var}(\mathbf{X} \pm \mathbf{Y}) = \text{Var}(\mathbf{X}) + \text{Var}(\mathbf{Y}) \pm (\text{Cov}(\mathbf{X}, \mathbf{Y}) + \text{Cov}(\mathbf{Y}, \mathbf{X})).$$

- Covariance transformation:

For any matrices \mathbf{A} and \mathbf{B} (of appropriate sizes),

$$\text{Cov}(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Y}) = \mathbf{A}\text{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^\top.$$

In particular,

$$\text{Var}(aX) = a^2\text{Var}(X), \quad \text{Var}(\mathbf{A}\mathbf{X}) = \mathbf{A}\text{Var}(\mathbf{X})\mathbf{A}^\top.$$

- Expectation of a quadratic form:

If $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$, then

$$\mathbb{E}(\mathbf{X}^\top \mathbf{A} \mathbf{X}) = \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu} + \text{tr}[\mathbf{A}\text{Var}(\mathbf{X})],$$

where tr denotes the trace of the matrix.

- Jensen's inequality:

If ψ is a convex function, then

$$\psi(\mathbb{E}X) \leq \mathbb{E}\psi(X).$$

In particular, $|\mathbb{E}X| \leq \mathbb{E}|X|$ and $\|\mathbb{E}\mathbf{X}\| \leq \mathbb{E}\|\mathbf{X}\|$.

- Markov's inequality:

If X is a random variable with $\mathbb{E}|X| < \infty$, then for any $t > 0$,

$$\mathbb{P}(|X| > t) \leq \mathbb{E}|X|/t.$$

- Association inequality:

If X is a random variable and f and g are nondecreasing functions, then

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