Lecure A3. Elementary Probability Theory

Lecturer: Jie Wang Date: Sept 22, 2022

Much of our life is based on the belief that the future is largely unpredictable. We often use the concept of **probability** to discuss an uncertain situation. In machine learning fields, probability theory is a powerful tool to describe the stochastic processes and formulate them into mathematical problems. In this lecture, we review some of the basics of elementary probability theory. The major references of this lecture are [1, 2]

1 Sample space and probability

We start with the probabilistic model in this section. A probabilistic model is a mathematical description of an uncertain situation, which helps us translate a practical problem into a mathematical problem to be solved.

Definition 1 (**Probabilistic models**)**.** A probabilistic model contains three parts $(\Omega, \mathcal{F}, \mathbb{P})$.

1. The sample space Ω is the set of all possible outcomes of an experiment.
2. The event field $\mathcal{F}$, a collection of some subsets of Ω, contains all the events that may occur.
3. A probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying
   (a) $\mathbb{P}(\varnothing) = 0$, $\mathbb{P}(\Omega) = 1$;
   (b) if $A_1, A_2, \ldots$ is a collection of disjoint members of $\mathcal{F}$, in that $A_i \cap A_j = \varnothing$ for all pairs i, j satisfying $i \neq j$, then

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

Example 1. Let $\Omega = \{a, b, c, d\}$ be a finite set, $\mathcal{F}$ be the collection containing all the subsets of Ω and $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ be a mapping from the subsets of Ω to real numbers. Suppose further that for any $A \in \mathcal{F}$, $\mathbb{P}(A) = |A|/4$, where $|A|$ is the cardinality of A. One can show that $(\Omega, \mathcal{F}, \mathbb{P})$ forms a probabilistic model.

Remark 1. Can we define a uniform probability measure on $\mathbb{N}$?

Many statements about chance take the form 'if an event B occurs, then the probability of another event A is p'. Now we describe it in mathematics.

Definition 2 (**Conditional probability**)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probabilistic model. $A, B \in \mathcal{F}$ are two events. If $\mathbb{P}(B) > 0$ then the conditional probability that A occurs given that B occurs is defined to be

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Divide-and-conquer is a vital thought in mathematics and machine learning areas. When the whole is difficult to deal with, we can think about breaking it down and solving it in parts. The total probability theorem reflects this idea in probability theory.

Theorem 1 (**Total probability theorem**). *Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space ($\Omega = \cup_{i=1}^{n} A_i$, each possible outcome is included in exactly one of the events $A_1, \ldots, A_n$) and assume that $\mathbb{P}(A_i) > 0$, for all i. Then, for any event B, we have*

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(A_1 \cap B) + \cdots + \mathbb{P}(A_n \cap B) \\ &= \mathbb{P}(A_1)\,\mathbb{P}(B \mid A_1) + \cdots + \mathbb{P}(A_n)\,\mathbb{P}(B \mid A_n).\end{aligned}$$

Combining the total probability theorem and conditional probability, we attain Bayes' rule and relates conditional probabilities of the form $\mathbb{P}(A \mid B)$ with conditional probabilities of the form $\mathbb{P}(B \mid A)$ as follows.

Theorem 2 (**Bayes' rule**). *Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $\mathbb{P}(A_i) > 0$, for all i. Then, for any event B such that $\mathbb{P}(B) > 0$, we have*

$$\begin{aligned}\mathbb{P}(A_i \mid B) &= \frac{\mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)}{\mathbb{P}(A_1)\,\mathbb{P}(B \mid A_1) + \cdots + \mathbb{P}(A_n)\,\mathbb{P}(B \mid A_n)}.\end{aligned}$$

Now we introduce the concept of independent events.

Definition 3 (**Independent events**). Events A and B are called independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

More generally, a family $\{A_i : i \in I\}$ is called independent if

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)$$

for all finite subsets J of I. Here I is an index set (probably infinite).

2 Random variables and their distributions

Given a sample space Ω and event field $\mathcal{F}$, a **random variable** associates a particular number with each element of sample space. This provides mathematical convenience in many situations.

Definition 4 (**Random variable**). Given a sample space Ω and event field $\mathcal{F}$, A random variable is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ for each subset $B \subset \mathbb{R}$ [1].

Example 2. Let $(\Omega, \mathcal{F})$ be defined the same as Example 1. Suppose $X : \Omega \to \mathbb{R}$ is a function such that $X(a) = X(b) = 1$, $X(c) = 2$ and $X(d) = 3$. Then X is a random variable.

Every random variable has a distribution function, which implies its value distribution information under a probability measure.

Definition 5 (**Distribution function**). Given a probabilistic model $(\Omega, \mathcal{F}, \mathbb{P})$, let X be a random variable defined on $(\Omega, \mathcal{F})$. The distribution function of X is the function $F : \mathbb{R} \to [0, 1]$ given by $F(x) = \mathbb{P}(X \leq x)$ for $x \in \mathbb{R}$. Here $\mathbb{P}(X \leq x)$ is an abbreviation of $\mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\})$.

[1] In fact, this definition is not rigorous in mathematics, in which the condition should be $\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ for every interval $B \subset \mathbb{R}$. But here we omit the difference between them.

Example 3. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be defined the same as Example 1 and X be defined in the same as Example 2. Then the distribution of X is

$$F_X(x) = \begin{cases} 0, & x < 1; \\ 1/2, & 1 \le x < 2; \\ 3/4, & 2 \le x < 3; \\ 1, & x \ge 3. \end{cases}$$

Remark 2. A distribution function F has the following properties:

1. $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$;
2. if $x < y$ then $F(x) \le F(y)$;
3. F is right-continuous, that is, $F(x+h) \to F(x)$ as $h \downarrow 0$.

We can also define the **joint distribution function** of two random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ by

$$F_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y), \quad x, y \in \mathbb{R}.$$

Remark 3. We should notice the following facts.

1. The definition of random variables does not include probability measure $\mathbb{P}$ in the probabilistic model, but the definition of distribution function does.
2. Let $f : \mathbb{R} \to \mathbb{R}$ and X is a random variable, then $f(X)$ is a random variable.
3. We say two random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if for any $x, y \in \mathbb{R}$, we have $\mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X \le x)\mathbb{P}(Y \le y)$. Notice that this is equivalent to

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \text{for all } x, y \in \mathbb{R}.$$

In the following context, we are going to study two typical kinds of random variables.

2.1 Discrete random variables

Discrete random variables are commonly seen in elementary probability. It is easy to understand and calculate.

Definition 6 (**Discrete random variables**)**.** The random variable X is called discrete if it takes values in some countable subset $\{x_1, x_2, \ldots\}$ of $\mathbb{R}$. We can define the (probability) mass function $p_X : \mathbb{R} \to [0, 1]$ of a discrete random variable X by $p_X(x) = \mathbb{P}(X = x)$. It is easy to see that $\sum_x p_X(x) = 1$.

Consider two discrete random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$. The probabilities of the values that X and Y can take are captured by the **joint probability mass function** of X and Y, denoted $p_{X,Y}$. In particular, if (x, y) is a pair of possible values of X and Y, the probability mass of (x, y) is the probability of the event $\{X = x, Y = y\}$:

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y).$$

We can calculate the probability mass functions of X and Y by using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x,y), \quad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

We sometimes refer to p_X and p_Y as the marginal mass functions, to distinguish them from the joint mass function.

With the probability mass functions, the independence of two random variables has a simpler description. Two discrete variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all x and y, i.e. $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$ or $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all $x, y \in \mathbb{R}$. Notice that if X and Y are independent and $g, h : \mathbb{R} \to \mathbb{R}$, then $g(X)$ and $h(Y)$ are independent also.

2.1.1 Expectation and variance

Definition 7 (**Expectation and variance**). 1. We define the expectation of a discrete random variable X, with probability mass function p_X, by

$$\mathbb{E}[X] = \sum_x x p_X(x)$$

whenever this sum is absolutely convergent.

2. We define the variance of a discrete random variable X, with probability mass function p_X, by

$$\operatorname{var}(X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2.$$

We have the following propositions of expectation.

Proposition 1. *Let X, Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Then we have the following statements.*

1. *If $a, b \in \mathbb{R}$ then $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.*
2. *If X and Y are independent then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

To calculate the expectation of functions of discrete random variables, the following theorem provides us with significant convenience.

Theorem 3. *Let X be a discrete random variable with mass function p_X and $g : \mathbb{R} \to \mathbb{R}$ is a function, then*

$$\mathbb{E}(g(X)) = \sum_x g(x) p_X(x)$$

whenever this sum is absolutely convergent.

We also have the following propositions of variance.

Proposition 2. *Let X, Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Then we have the following statements.*

1. *$\operatorname{var}(aX) = a^2 \operatorname{var}(X)$ for $a \in \mathbb{R}$.*
2. *If X and Y are uncorrelated, i.e., $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = 0$ or $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, then $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y)$.*

We have discussed the conditional probability $\mathbb{P}(B \mid A)$. This may be set in the more general context of the conditional distribution of one variable Y given the value of another variable X.

Definition 8 (Conditional distribution and conditional expectation)**.** Let X and Y be two discrete variables on $(\Omega, \mathcal{F}, \mathbb{P})$.

1. The conditional distribution function of Y given $X = x$, written $F_{Y|X}(\cdot \mid x)$, is defined by

$$F_{Y|X}(y \mid x) = \mathbb{P}(Y \leq y \mid X = x) = \sum_{v \leq y} \frac{p_{X,Y}(x, v)}{p_X(x)}$$

   for any x such that $\mathbb{P}(X = x) > 0$. The conditional (probability) mass function of Y given $X = x$, written $p_{Y|X}(\cdot \mid x)$, is defined by

$$p_{Y|X}(y \mid x) = \mathbb{P}(Y = y \mid X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

   for any x such that $\mathbb{P}(X = x) > 0$.

2. We define the conditional expectation of Y given X, written as $\mathbb{E}[Y \mid X]$, by

$$\mathbb{E}[Y \mid X] = \sum_y y p_{Y|X}(y \mid x).$$

2.1.2 Examples

We give two examples of discrete random variables.

Example 4 (The Poisson random variable)**.** A Poisson discrete random variable X has a probability mass function given by

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots,$$

where λ is a positive parameter.

The mean and the variance can be calculated to be

$$\mathbb{E}[X] = \lambda, \quad \text{var}(X) = \lambda.$$

Example 5 (The binomial random variable)**.** A coin is tossed n times. At each toss, the coin comes up with a head with probability p, and a tail with probability $1 - p$, independent of prior tosses. Let X be the number of heads in the n-toss sequence. We refer to X as a binomial random variable with parameters n and p. The probability mass function of X is

$$p_X(k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

The mean and the variance can be calculated to be

$$\mathbb{E}[X] = np, \quad \text{var}(X) = np(1-p).$$

2.2 Continuous random variables

Continuous random variables are another kind of commonly seen random variables.

Definition 9. The random variable X is called continuous if its distribution function can be expressed as

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du, \quad x \in \mathbb{R}$$

for some integrable function $f_X : \mathbb{R} \to [0, \infty)$ called the (probability) density function of X.

Remark 4. We say the random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ are (jointly) continuous with joint (probability) density function $f_{X,Y} : \mathbb{R}^2 \to [0, \infty)$ if

$$F_{X,Y}(x, y) = \int_{v=-\infty}^{y} \int_{u=-\infty}^{x} f_{X,Y}(u, v)dudv \quad \text{for each } x, y \in \mathbb{R}.$$

One can show that $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$. Furthermore, if F is sufficiently differentiable at the point (x, y), then we usually specify

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

For joint continuous random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$, X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

2.2.1 Expectation and variance

Definition 10 (Expectation and variance)**.**

1. The expectation of a continuous random variable X with density function f_X is given by

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

whenever this integral exists.

2. The variance of a continuous random variable X, with probability density function f_X, is defined by

$$\text{var}(X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2.$$

Remark 5. Similar to conditional distribution and conditional expectation of discrete random variables, we can define the conditional distribution and conditional expectation for continuous random variables X and Y on $\Omega, \mathcal{F}, \mathbb{P}$ with probability density functions as follows.

1. The conditional distribution function of Y given $X = x$ is the function $F_{Y|X}(\cdot \mid x)$ given by

$$F_{Y|X}(y \mid x) = \int_{-\infty}^{y} \frac{f_{X,Y}(x, v)}{f_X(x)} dv.$$

for any x such that $f_X(x) > 0$. It is sometimes denoted $\mathbb{P}(Y \leq y \mid X = x)$.

The conditional density function of $F_{Y|X}$, written $f_{Y|X}$, is given by

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

for any x such that $f_X(x) > 0$.

2. The conditional expectation of Y given X can be defined as in by

$$\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy.$$

The propositions of discrete random variables' expectation and variance we have studied can be applied to continuous random variables, where we should substitute the discrete summation weighted by mass functions with the integral weighted by the density functions.

Theorem 4. *If X is a continuous random variable and g is a continuous function, then*

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

2.2.2 Functions of continuous random variables

Let X be a continuous random variable with density function f_X, and let $g : \mathbb{R} \to \mathbb{R}$ be a function with differentiable inverse. Then $Y = g(X)$ is a random variable also. In order to calculate the distribution of Y, we calculate that

$$\begin{aligned}\mathbb{P}(Y \le y) &= \mathbb{P}(g(X) \le y) = \mathbb{P}(g(X) \in (-\infty, y]) \\ &= \mathbb{P}\left(X \in g^{-1}(-\infty, y]\right) = \int_{g^{-1}(-\infty,y]} f_X(x) dx.\end{aligned}$$

Differentiating the above equation with respect to y, we attain the expression of $f_Y(y)$. Moreover, for a strictly monotonic differentiable function g with continuously differentiable inverse, we can use the change of variable formula. Then, the probability density function of Y in the region where $f_Y(y) > 0$ is given by

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right|.$$

More generally, if X_1 and X_2 have joint density function f_{X_1,X_2}, and g, h are functions mapping $\mathbb{R}^2$ to $\mathbb{R}$, then what is the joint density function of the pair $Y_1 = g(X_1, X_2), Y_2 = h(X_1, X_2)$? Recall how to change variables within an integral. Let $y_1 = g(x_1, x_2), y_2 = h(x_1, x_2)$ be a one-one mapping $T : (x_1, x_2) \mapsto (y_1, y_2)$ taking some domain $D \subseteq \mathbb{R}^2$ onto some range $R \subseteq \mathbb{R}^2$. The transformation can be inverted as $x_1 = x_1(y_1, y_2), x_2 = x_2(y_1, y_2)$; the Jacobian of this inverse is defined to be the determinant

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \frac{\partial x_1}{\partial y_1}\frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2}\frac{\partial x_2}{\partial y_1}$$

which we express as a function $J = J(y_1, y_2)$. We assume that these partial derivatives are continuous.

Theorem 5 (Change of variable formula). *If X_1, X_2 have joint density function f_{X_1,X_2}, then the pair Y_1, Y_2 given by $(Y_1, Y_2) = T(X_1, X_2)$ has joint density function*

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} f_{X_1,X_2}(x_1(y_1, y_2), x_2(y_1, y_2)) \left|J(y_1, y_2)\right| & \text{if } (y_1, y_2) \text{ is in the range of } T, \\ 0 & \text{otherwise .} \end{cases}$$

Example 6. Let $Y = g(X) = X^2$, where X is a continuous uniform random variable on the interval $(0, 1]$. Within this interval, g is strictly monotonic, and its inverse is $h(y) = \sqrt{y}$. Thus, for any $y \in (0.1]$, we have

$$f_X(\sqrt{y}) = 1, \quad \left|\frac{dh}{dy}(y)\right| = \frac{1}{2\sqrt{y}}.$$

and

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}, & \text{if } y \in (0, 1] \\ 0, & \text{otherwise.} \end{cases}$$

What about X defined on the interval $[-1, 1]$? We leave it for an exercise.

2.2.3 Examples

Example 7 (The normal random variable). A continuous random variable X is said to be normal or Gaussian if it has a probability density function of the form

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2},$$

where μ and σ are two scalar parameters with σ assumed positive.

The mean and the variance can be calculated to be

$$\mathbb{E}[X] = \mu, \quad \text{var}(X) = \sigma^2.$$

Example 8 (The exponential random variable). An exponential random variable has a probability density function of the form

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where λ is a positive parameter.

The mean and the variance can be calculated to be

$$\mathbb{E}[X] = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2}.$$

3 Limit theorems

Machine learning lies in the fundamental of limit theorems. We train our model with a large amount of independent identically distributed data to attain accurate parameter estimations.

Theorem 6 (The weak law of large numbers). *Let $X_1, X_2, \ldots$ be independent identically distributed random variables $(\Omega, \mathcal{F}, \mathbb{P})$ with mean μ and $\mathbb{E}[|X_i|] < \infty$. For every $\epsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0, \qquad \text{as } n \to \infty.$$

The word 'weak' here means the convergence type of the average $(X_1 + \cdots + X_n)/n$ is not so strict as convergence point-wise or almost point-wise. We do have a strong version of the law of large numbers, in which the average converges to the expectation μ, except for a set of zero probability.

Theorem 7 (**The strong law of large numbers**). *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with mean μ and $\mathbb{E}[|X_i|] < \infty$. Then, we have*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

Remark 6. Let $Y_1, Y_2, \ldots$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say Y_n converges to a random variable Y in probability, if $\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| \geq \epsilon) = 0$ for every $\epsilon > 0$. We say Y_n converges to a random variable Y with probability 1, if $\mathbb{P}(\lim_{n\to\infty} Y_n = Y) = 1$. Convergence with probability 1 implies convergence in probability. But the inverse may not be true.

We need to distinguish these two kinds of convergence carefully. Roughly speaking, the set where Y_n and Y differ larger than any threshold ϵ becomes smaller across n does not mean that Y_n converges to Y almost everywhere.

Example 9. Let $\Omega = [0, 1)$, $\mathcal{F}$ be an event field containing all the intervals and any of their unions in $[0, 1)$, $\mathbb{P}$ be the uniform probability measure in Ω. Then $(\Omega, \mathcal{F}, \mathbb{P})$ forms a probabilistic model. For each $k \in \mathbb{N}_+$, we define k functions

$$f_i^{(k)}(x) = \begin{cases} 1, & x \in \left[\frac{i-1}{k}, \frac{i}{k}\right), \\ 0, & x \notin \left[\frac{i-1}{k}, \frac{i}{k}\right), \end{cases} \quad i = 1, 2, \cdots, k.$$

We rearrange these function as $\varphi_1(x) = f_1^{(1)}(x)$, $\varphi_2(x) = f_1^{(2)}(x)$, $\varphi_3(x) = f_2^{(2)}(x)$, $\varphi_4(x) = f_1^{(3)}(x), \cdots$. Then we can show that $\{\varphi_n(x)\}$ converges to 0 in probability but not with probability 1.

References

[1] D. Bertsekas and J. N. Tsitsiklis. *Introduction to probability*, volume 1. Athena Scientific, 2008.

[2] G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford university press, 2020.