Lecture 2. Random Variables

Tiejun Li$^{1,2}$

$^1$School of Mathematical Sciences (SMS),
&
$^2$Center for Machine Learning Research (CMLR),
Peking University,
Beijing 100871,
P.R. China
tieli@pku.edu.cn

Office: No. 1 Science Building, Room 1376E
Table of Contents

Elementary Random Variables

Axiomatic Probability Theory Setup

Conditional Expectation

Characteristic and Generating Functions

Borel-Cantelli Lemma
Discrete Examples: Bernoulli distribution $\mathcal{B}er(p)$

We will first consider the elementary and intuitive aspects of probability here. In the discrete case, the function $\mathbb{P}(X)$ is called the probability mass function (pmf).

Bernoulli distribution $\mathcal{B}er(p)$.

- Bernoulli distribution:

  $\mathbb{P}(X) = \begin{cases} p, & X = 1, \\ q, & X = 0. \end{cases}$

  where $p > 0, q > 0, p + q = 1$. 

  If $p = q = \frac{1}{2}$, it is the well-known fair-coin tossing game.

  The mean and variance are $E(X) = p$, $\text{Var}(X) = pq$. 
Discrete Examples: Bernoulli distribution $\mathcal{B}er(p)$

We will first consider the elementary and intuitive aspects of probability here. In the discrete case, the function $\mathbb{P}(X)$ is called the probability mass function (pmf).

Bernoulli distribution $\mathcal{B}er(p)$.

- Bernoulli distribution:

  \[ \mathbb{P}(X) = \begin{cases} p, & X = 1, \\ q, & X = 0. \end{cases} \]

  where $p > 0$, $q > 0$, $p + q = 1$.

- If $p = q = \frac{1}{2}$, it is the well-known fair-coin tossing game.
Discrete Examples: Bernoulli distribution $Ber(p)$

We will first consider the elementary and intuitive aspects of probability here. In the discrete case, the function $P(X)$ is called the probability mass function (pmf).

Bernoulli distribution $Ber(p)$.

- Bernoulli distribution:

$$P(X) = \begin{cases} 
p, & X = 1, 
q, & X = 0. 
\end{cases}$$

where $p > 0, q > 0, p + q = 1$.

- If $p = q = \frac{1}{2}$, it is the well-known fair-coin tossing game.

- The mean and variance are

$$\mathbb{E}X = p, \text{Var}(X) = pq.$$
Discrete Examples: Categorical distribution $Cat(p)$

Categorical distribution $Cat(p)$.

- A generalization of Bernoulli distribution, in which each trial results in exactly one of some fixed number $r$ possible outcomes with probability $p_1, p_2, \ldots, p_r$, where

$$\sum_{i=1}^{r} p_i = 1, \quad 0 \leq p_i \leq 1, \quad i = 1, \ldots, r,$$
Discrete Examples: Categorical distribution $Cat(p)$

Categorical distribution $Cat(p)$.

- A generalization of Bernoulli distribution, in which each trial results in exactly one of some fixed number $r$ possible outcomes with probability $p_1, p_2, \ldots, p_r$, where

$$\sum_{i=1}^{r} p_i = 1, \quad 0 \leq p_i \leq 1, \quad i = 1, \ldots, r,$$

- Denote $X = e_k = (\delta_{kj})_{j=1:r}$ for $k = 1 : r$ instead of $X \in \{1, 2, \ldots, r\}$ if the outcome is $k$. And denote

$$X = (X_1, \ldots, X_r).$$
Discrete Examples: Categorical distribution $Cat(p)$

Categorical distribution $Cat(p)$.

- A generalization of Bernoulli distribution, in which each trial results in exactly one of some fixed number $r$ possible outcomes with probability $p_1, p_2, \ldots, p_r$, where

  \[ \sum_{i=1}^{r} p_i = 1, \quad 0 \leq p_i \leq 1, \quad i = 1, \ldots, r, \]

- Denote $X = e_k = (\delta_{kj})_{j=1:r}$ for $k = 1 : r$ instead of $X \in \{1, 2, \ldots, r\}$ if the outcome is $k$. And denote

  \[ X = (X_1, \ldots, X_r). \]

- The pmf is:

  \[ \mathbb{P}(X = e_k) = p_k, \quad k \in \{1, 2, \ldots, r\} \]
Discrete Examples: Categorical distribution \( \text{Cat}(p) \)

Categorical distribution \( \text{Cat}(p) \).

- A generalization of Bernoulli distribution, in which each trial results in exactly one of some fixed number \( r \) possible outcomes with probability \( p_1, p_2, \ldots, p_r \), where

\[
\sum_{i=1}^{r} p_i = 1, \quad 0 \leq p_i \leq 1, \quad i = 1, \ldots, r,
\]

- Denote \( X = e_k = (\delta_{kj})_{j=1:r} \) for \( k = 1 : r \) instead of \( X \in \{1, 2, \ldots, r\} \) if the outcome is \( k \). And denote

\[
X = (X_1, \ldots, X_r).
\]

- The pmf is:

\[
P(X = e_k) = p_k, \quad k \in \{1, 2, \ldots, r\}
\]

- The mean and variance are

\[
\mathbb{E}(X_i) = p_i, \quad \text{Var}(X_i) = p_i(1 - p_i).
\]
Discrete Examples: Binomial distribution $B(n, p)$

Binomial distribution $B(n, p)$:

- Consider $n$ independent experiments of Bernoulli distribution $X_k$
Discrete Examples: Binomial distribution $B(n, p)$

Binomial distribution $B(n, p)$:

- Consider $n$ independent experiments of Bernoulli distribution $X_k$
- A binomially distributed random variable $X$ can be viewed as the sum of $n$ independent Bernoulli trials $X_k$. Define

$$X := X_1 + \ldots + X_n$$
Discrete Examples: Binomial distribution $B(n, p)$

Binomial distribution $B(n, p)$:

- Consider $n$ independent experiments of Bernoulli distribution $X_k$
- A binomially distributed random variable $X$ can be viewed as the sum of $n$ independent Bernoulli trials $X_k$. Define

$$X := X_1 + \ldots + X_n$$

- Then

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}.$$
Discrete Examples: Binomial distribution $B(n, p)$

Binomial distribution $B(n, p)$:

- Consider $n$ independent experiments of Bernoulli distribution $X_k$
- A binomially distributed random variable $X$ can be viewed as the sum of $n$ independent Bernoulli trials $X_k$. Define

\[ X := X_1 + \ldots + X_n \]

- Then

\[ \mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}. \]

- The mean and variance are

\[ \mathbb{E}X = np, \ Var(X) = npq. \]
Discrete Examples: Multinomial distribution $M(n, p)$

Multinomial distribution $M(n, p)$.

- A generalization of binomial distribution, in which each trial is a categorically distributed RV with parameter $p$. 

The pmf of the multinomial distribution is:

$$P(X_1 = x_1, \ldots, X_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r},$$

where $n = x_1 + \cdots + x_r$.

The mean, variance and covariance are

$$E(X_i) = np_i,$$

$$\text{Var}(X_i) = np_i(1 - p_i),$$

$$\text{Cov}(X_i, X_j) = -np_i p_j (i \neq j).$$
Discrete Examples: Multinomial distribution $M(n, p)$

Multinomial distribution $M(n, p)$.

- A generalization of binomial distribution, in which each trial is a categorically distributed RV with parameter $p$.

- Let $X_i$ indicate the number of times the $i$-th outcome was observed over the $n$ trials. Then

$$X = (X_1, \ldots, X_r).$$
Discrete Examples: Multinomial distribution $M(n, \mathbf{p})$

Multinomial distribution $M(n, \mathbf{p})$.

- A generalization of binomial distribution, in which each trial is a categorically distributed RV with parameter $\mathbf{p}$.
- Let $X_i$ indicate the number of times the $i$-th outcome was observed over the $n$ trials. Then

  $$X = (X_1, \ldots, X_r).$$

- The pmf of the multinomial distribution is:

  $$\mathbb{P}(X_1 = x_1, \ldots, X_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r},$$

  where $n = x_1 + \cdots + x_r$. 
Discrete Examples: Multinomial distribution $M(n, p)$

Multinomial distribution $M(n, p)$.

- A generalization of binomial distribution, in which each trial is a categorically distributed RV with parameter $p$.
- Let $X_i$ indicate the number of times the $i$-th outcome was observed over the $n$ trials. Then

$$X = (X_1, \ldots, X_r).$$

- The pmf of the multinomial distribution is:

$$\mathbb{P}(X_1 = x_1, \ldots, X_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r},$$

where $n = x_1 + \cdots + x_r$.

- The mean, variance and covariance are $\mathbb{E}(X_i) = np_i$,

$$\text{Var}(X_i) = np_i(1 - p_i), \quad \text{Cov}(X_i, X_j) = -np_ip_j \ (i \neq j).$$
Discrete Examples: Poisson distribution $\mathcal{P}(\lambda)$

Poisson distribution $\mathcal{P}(\lambda)$.

- The number $X$ of radiated particles in a fixed time $\tau$ obeys

$$
\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},
$$

where $\lambda$ is the average number of radiated particles each time.
Discrete Examples: Poisson distribution $\mathcal{P}(\lambda)$

Poisson distribution $\mathcal{P}(\lambda)$.

- The number $X$ of radiated particles in a fixed time $\tau$ obeys

  $$
  \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},
  $$

  where $\lambda$ is the average number of radiated particles each time.

- The mean and variance are

  $$
  \mathbb{E}X = \lambda, \text{Var}(X) = \lambda.
  $$
Discrete Examples: Poisson distribution $\mathcal{P}(\lambda)$

Poisson distribution $\mathcal{P}(\lambda)$.

- The number $X$ of radiated particles in a fixed time $\tau$ obeys

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

where $\lambda$ is the average number of radiated particles each time.

- The mean and variance are

$$\mathbb{E}X = \lambda, \ Var(X) = \lambda.$$

- Poisson distribution may be viewed as the limit of binomial distribution (the law of rare events)

$$\binom{n}{k} p^k q^{n-k} \rightarrow \frac{\lambda^k}{k!} e^{-\lambda} \quad (n \rightarrow \infty, np = \lambda).$$
Discrete Examples: Poisson distribution $\mathcal{P}(\lambda)$

- Poisson distribution can also describe the spatial distribution of randomly scattered points.

\[ \mathbb{P}(X_A = n) = \frac{(\lambda \cdot \text{meas}(A))^n}{n!} e^{-\lambda \cdot \text{meas}(A)}. \]

$A$: a set in $R^2$,  
$X_A(\omega)$: number of points in $A$.  
$\lambda$: scattering density.
Continuous Examples: Uniform distribution $\mathcal{U}[0, 1]$

In continuous case, the function $p(x)$ is called the **probability density function** (pdf).

Uniform distribution $\mathcal{U}[0, 1]$:  

- The pdf  

\[ p(x) = \begin{cases} 
1 & \text{if } x \in [0, 1] \\
0 & \text{otherwise} 
\end{cases} \]
Continuous Examples: Uniform distribution $\mathcal{U}[0, 1]$

In continuous case, the function $p(x)$ is called the probability density function (pdf).

Uniform distribution $\mathcal{U}[0, 1]$:

- The pdf

\[
p(x) = \begin{cases} 
1 & \text{if } x \in [0, 1] \\
0 & \text{otherwise}
\end{cases}
\]

- The mean and variance are

\[
\mathbb{E}X = \frac{1}{2}, \text{Var}(X) = \frac{1}{12}.
\]
Continuous Examples: Exponential distribution: $\mathcal{E}xp(\lambda)$

Exponential distribution: $\mathcal{E}xp(\lambda)$

- The pdf with $(\lambda > 0)$

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

- The mean and variance are $E_X = \frac{1}{\lambda}$, $\text{Var}(X) = \frac{1}{\lambda^2}$.

- Waiting time for continuous time Markov process also has exponential distribution, where $\lambda$ is the rate of the process.
Continuous Examples: Exponential distribution: $\mathcal{E}xp(\lambda)$

Exponential distribution: $\mathcal{E}xp(\lambda)$
- The pdf with $(\lambda > 0)$

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

- The mean and variance are

$$\mathbb{E}X = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$
Continuous Examples: Exponential distribution: $\mathcal{E}xp(\lambda)$

Exponential distribution: $\mathcal{E}xp(\lambda)$

- The pdf with ($\lambda > 0$)

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

- The mean and variance are

$$\mathbb{E}X = \frac{1}{\lambda}, \text{Var}(X) = \frac{1}{\lambda^2}.$$ 

- Waiting time for continuous time Markov process also has exponential distribution, where $\lambda$ is the rate of the process.
Continuous Examples: Gaussian distribution $N(\mu, \Sigma)$

- Normal distribution (Gaussian distribution) ($N(0, 1)$):

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

or more generally $N(\mu, \sigma)$

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

where $\mu$ is the mean (expectation), $\sigma^2$ is the variance.
Continuous Examples: Gaussian distribution $N(\mu, \Sigma)$

- Normal distribution (Gaussian distribution) ($N(0, 1)$):

\[
p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}
\]

or more generally $N(\mu, \sigma)$

\[
p(x) = \frac{1}{\sqrt{2\pi} \sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

where $\mu$ is the mean (expectation), $\sigma^2$ is the variance.

- High dimensional case ($N(\mu, \Sigma^2)$)

\[
p(x) = \frac{1}{(2\pi)^{n/2} (\det \Sigma)^{1/2}} e^{-(x-\mu)^T \Sigma^{-1} (x-\mu)}
\]

where $\mu$ is the mean, $\Sigma$ is the covariance matrix of $X$. 
Continuous Examples: Gaussian distribution \( N(\mu, \Sigma) \)

- Normal distribution (Gaussian distribution) \( (N(0, 1)) \):
  \[
p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}
\]

or more generally \( N(\mu, \sigma) \)

\[
p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

where \( \mu \) is the mean (expectation), \( \sigma^2 \) is the variance.

- High dimensional case \( (N(\mu, \Sigma^2)) \)
  \[
p(x) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} e^{-(X-\mu)^T\Sigma^{-1}(X-\mu)}
\]

where \( \mu \) is the mean, \( \Sigma \) is the covariance matrix of \( X \).

- More general Gaussian distribution with \( \det \Sigma = 0 \)?
Remarks on Gaussian distribution

► In 1D case, the normal distribution $N(np, npq)$ may be viewed as the limit of the Binomial distribution $B(n, p)$ when $n$ is large. This is the famous De Moivre-Laplace limit theorem. It is a special case of the central limit theorem (CLT). Notice that

$$B(n, p) - np \over \sqrt{npq} \to N(0, 1) \text{ as } n \to \infty.$$
Remarks on Gaussian distribution

In 1D case, the normal distribution $N(np, npq)$ may be viewed as the limit of the Binomial distribution $B(n, p)$ when $n$ is large. This is the famous De Moivre-Laplace limit theorem. It is a special case of the central limit theorem (CLT). Notice that

$$\frac{B(n, p) - np}{\sqrt{npq}} \to N(0, 1) \text{ as } n \to \infty.$$  

In 1D case, the normal distribution $N(\lambda, \lambda)$ may be viewed as the limit of the Poisson distribution $P(\lambda)$ when $\lambda$ is large. Notice the simple fact that the sum of two independent $P(\lambda)$ and $P(\mu)$ is $P(\lambda + \mu)$ (why?), we can decompose $P(\lambda)$ into the sum of $n$ i.i.d. $P(\lambda/n)$, we have

$$\frac{P(\lambda) - \lambda}{\sqrt{\lambda}} \to N(0, 1) \text{ when } \lambda \text{ is large.}$$

Question: What if $n \to \infty$?
Table of Contents

Elementary Random Variables

Axiomatic Probability Theory Setup

Conditional Expectation

Characteristic and Generating Functions

Borel-Cantelli Lemma
Axiomatic Setup: Probability Space

- Sample space $\Omega$: the set of all outcomes $\omega$. 

- Event space: $\sigma$-algebra $\mathcal{F}$ is a collection of subsets of $\Omega$:
  1. $\Omega \in \mathcal{F}$;
  2. If $A \in \mathcal{F}$, then $\overline{A} = \Omega \setminus A \in \mathcal{F}$;
  3. If $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

$(\Omega, \mathcal{F})$ is called a measurable space.

- Probability measure $P$:
  1. (Positive) For all $A \in \mathcal{F}$, $P(A) \geq 0$;
  2. (Countably additive) If $A_1, A_2, \ldots \in \mathcal{F}$ and they are disjoint, then $P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$;
  3. (Normalization) $P(\Omega) = 1$.

- Probability space — Triplet $(\Omega, \mathcal{F}, P)$
Axiomatic Setup: Probability Space

- Sample space $\Omega$: the set of all outcomes $\omega$.
- Event space: $\sigma$-algebra $\mathcal{F}$
  $\mathcal{F}$ is a collection of subsets of $\Omega$: 
  
- Probability space — Triplet $(\Omega, \mathcal{F}, P)$
Axiomatic Setup: Probability Space

- Sample space $\Omega$: the set of all outcomes $\omega$.
- Event space: $\sigma$-algebra $\mathcal{F}$
  $\mathcal{F}$ is a collection of subsets of $\Omega$:
  1. $\Omega \in \mathcal{F}$;
Axiomatic Setup: Probability Space

- Sample space $\Omega$: the set of all outcomes $\omega$.
- Event space: $\sigma$-algebra $\mathcal{F}$
  $\mathcal{F}$ is a collection of subsets of $\Omega$:
  1. $\Omega \in \mathcal{F}$;
  2. If $A \in \mathcal{F}$, then $\bar{A} = \Omega \setminus A \in \mathcal{F}$;

$(\Omega, \mathcal{F}, P)$ is called a measurable space.
Axiomatic Setup: Probability Space

- Sample space $\Omega$: the set of all outcomes $\omega$.
- Event space: $\sigma$-algebra $\mathcal{F}$
  $\mathcal{F}$ is a collection of subsets of $\Omega$:
  1. $\Omega \in \mathcal{F}$;
  2. If $A \in \mathcal{F}$, then $\bar{A} = \Omega \setminus A \in \mathcal{F}$;
  3. If $A_1, A_2, \cdots, A_n, \cdots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

$(\Omega, \mathcal{F})$ is called a measurable space.

- Probability measure $P$
  1. (Positive) $\forall A \in \mathcal{F}$, $P(A) \geq 0$;
  2. (Countably additive) If $A_1, A_2, \cdots \in \mathcal{F}$, and they are disjoint, then $P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$;
  3. (Normalization) $P(\Omega) = 1$.

- Probability space — Triplet $(\Omega, \mathcal{F}, P)$
Axiomatic Setup: Probability Space

- Sample space $\Omega$: the set of all outcomes $\omega$.
- Event space: $\sigma$-algebra $F$
  $F$ is a collection of subsets of $\Omega$:
  1. $\Omega \in F$;
  2. If $A \in F$, then $\bar{A} = \Omega \setminus A \in F$;
  3. If $A_1, A_2, \cdots, A_n, \cdots \in F$, then $\bigcup_{j=1}^{\infty} A_j \in F$.

$(\Omega, F)$ is called a measurable space.
Axiomatic Setup: Probability Space

- Sample space $\Omega$: the set of all outcomes $\omega$.

- Event space: $\sigma$-algebra $\mathcal{F}$
  
  $\mathcal{F}$ is a collection of subsets of $\Omega$:
  
  1. $\Omega \in \mathcal{F}$;
  2. If $A \in \mathcal{F}$, then $\bar{A} = \Omega \setminus A \in \mathcal{F}$;
  3. If $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

  $(\Omega, \mathcal{F})$ is called a measurable space.

- Probability measure $P$
Axiomatic Setup: Probability Space

- Sample space $\Omega$: the set of all outcomes $\omega$.
- Event space: $\sigma$-algebra $\mathcal{F}$
  $\mathcal{F}$ is a collection of subsets of $\Omega$:
  1. $\Omega \in \mathcal{F}$;
  2. If $A \in \mathcal{F}$, then $\bar{A} = \Omega \setminus A \in \mathcal{F}$;
  3. If $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

$(\Omega, \mathcal{F})$ is called a measurable space.

- Probability measure $P$
  1. (Positive) $\forall A \in \mathcal{F}, P(A) \geq 0$;
Axiomatic Setup: Probability Space

- **Sample space** $\Omega$: the set of all outcomes $\omega$.

- **Event space**: $\sigma$-algebra $\mathcal{F}$
  $\mathcal{F}$ is a collection of subsets of $\Omega$:
  1. $\Omega \in \mathcal{F}$;
  2. If $A \in \mathcal{F}$, then $\bar{A} = \Omega \setminus A \in \mathcal{F}$;
  3. If $A_1, A_2, \cdots, A_n, \cdots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

($\Omega, \mathcal{F}$) is called a measurable space.

- **Probability measure** $P$
  1. (Positive) $\forall A \in \mathcal{F}, P(A) \geq 0$;
  2. (Countably additive) If $A_1, A_2, \cdots \in \mathcal{F}$, and they are disjoint, then $P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$;
Axiomatic Setup: Probability Space

▶ Sample space $\Omega$: the set of all outcomes $\omega$.

▶ Event space: $\sigma$-algebra $\mathcal{F}$

$\mathcal{F}$ is a collection of subsets of $\Omega$:

1. $\Omega \in \mathcal{F}$;
2. If $A \in \mathcal{F}$, then $\bar{A} = \Omega \setminus A \in \mathcal{F}$;
3. If $A_1, A_2, \cdots, A_n, \cdots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

$(\Omega, \mathcal{F})$ is called a measurable space.

▶ Probability measure $P$

1. (Positive) $\forall A \in \mathcal{F}, \, P(A) \geq 0$;
2. (Countably additive) If $A_1, A_2, \cdots \in \mathcal{F}$, and they are disjoint, then $P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$;
3. (Normalization) $\mathbb{P}(\Omega) = 1$. 

Probability space — Triplet $(\Omega, \mathcal{F}, P)$
Axiomatic Setup: Probability Space

▶ Sample space $\Omega$: the set of all outcomes $\omega$.

▶ Event space: $\sigma$-algebra $\mathcal{F}$
$\mathcal{F}$ is a collection of subsets of $\Omega$:
1. $\Omega \in \mathcal{F}$;
2. If $A \in \mathcal{F}$, then $\bar{A} = \Omega \setminus A \in \mathcal{F}$;
3. If $A_1, A_2, \cdots, A_n, \cdots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.
$(\Omega, \mathcal{F})$ is called a measurable space.

▶ Probability measure $P$
1. (Positive) $\forall A \in \mathcal{F}$, $P(A) \geq 0$;
2. (Countably additive) If $A_1, A_2, \cdots \in \mathcal{F}$, and they are disjoint, then $P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$;
3. (Normalization) $\mathbb{P}(\Omega) = 1$.

▶ Probability space — Triplet $(\Omega, \mathcal{F}, \mathbb{P})$
Radon-Nikodym Theorem

Theorem
Suppose $\mu$ is a $\sigma$-finite measure, $\nu$ is a signed measure on measurable space $(\Omega, \mathcal{F})$. If $\nu$ is absolutely continuous w.r.t. $\mu$, then there exists a measurable function $f$, such that for any $A \in \mathcal{F}$

$$\nu(A) = \int_A f(\omega) \mu(d\omega),$$

and $f$ is unique in the $\mu$-a.e. sense.

$f$ is defined as the Radon-Nikodym derivative $d\nu/d\mu = f$.

---

For any $A \in \mathcal{F}$, if $\mu(A) = 0$, then $\nu(A) = 0$. It is usually denoted as $\nu \ll \mu$. 
Random Variables

- Random variable: a measurable function $X : \Omega \to \mathbb{R}$.
Random Variables

- Random variable: a measurable function $X : \Omega \to \mathbb{R}$.
- Distribution (or law): a probability measure $\mu$ on $\mathbb{R}$ defined for any set $B \subset \mathbb{R}$ by

$$\mu(B) = \text{Prob}(X \in B) = \mathbb{P}\{\omega \in \Omega : X(\omega) \in B\}.$$
Random Variables

- Random variable: a measurable function \( X : \Omega \to \mathbb{R} \).
- Distribution (or law): a probability measure \( \mu \) on \( \mathbb{R} \) defined for any set \( B \subset \mathbb{R} \) by

  \[
  \mu(B) = \text{Prob}(X \in B) = \mathbb{P}\{\omega \in \Omega : X(\omega) \in B\}.
  \]

- Probability density function (pdf): an integrable function \( p(x) \) on \( \mathbb{R} \) such that for any set \( B \subset \mathbb{R} \),

  \[
  \mu(B) = \int_B p(x) \, dx.
  \]
Random Variables

- Random variable: a measurable function $X : \Omega \rightarrow \mathbb{R}$.

- Distribution (or law): a probability measure $\mu$ on $\mathbb{R}$ defined for any set $B \subset \mathbb{R}$ by

  $$\mu(B) = \text{Prob}(X \in B) = \mathbb{P}\{\omega \in \Omega : X(\omega) \in B\}.$$ 

- Probability density function (pdf): an integrable function $p(x)$ on $\mathbb{R}$ such that for any set $B \subset \mathbb{R}$,

  $$\mu(B) = \int_B p(x)dx.$$ 

- Mean (expectation):

  $$\mathbb{E}f(X) = \int_{\Omega} f(X(\omega))P(d\omega) = \int_{\mathbb{R}} f(x)d\mu(x) = \int_{\mathbb{R}} f(x)p(x)dx.$$
Random Variables

- **Random variable**: a measurable function $X : \Omega \rightarrow \mathbb{R}$.
- **Distribution (or law)**: a probability measure $\mu$ on $\mathbb{R}$ defined for any set $B \subset \mathbb{R}$ by
  $$\mu(B) = \text{Prob}(X \in B) = \mathbb{P}\{\omega \in \Omega : X(\omega) \in B\}.$$ 
- **Probability density function (pdf)**: an integrable function $p(x)$ on $\mathbb{R}$ such that for any set $B \subset \mathbb{R}$,
  $$\mu(B) = \int_B p(x)dx.$$
- **Mean (expectation)**:
  $$\mathbb{E}f(X) = \int_{\Omega} f(X(\omega))P(d\omega) = \int_{\mathbb{R}} f(x)d\mu(x) = \int_{\mathbb{R}} f(x)p(x)dx.$$ 
- **Variance**:
  $$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$
Moments, Covariance, etc.

- $p$-th moment: $\mathbb{E}|X|^p$. 
Moments, Covariance, etc.

- \( p \)-th moment: \( \mathbb{E}|X|^p \).
- Covariance:

\[
\text{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).
\]
Moments, Covariance, etc.

- $p$-th moment: $\mathbb{E}|X|^p$.
- Covariance:
  \[
  \text{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).
  \]
- Independence:
  \[
  \mathbb{E}f(X)g(Y) = \mathbb{E}f(X)\mathbb{E}g(Y).
  \]
  for all continuous functions $f$ and $g$. 

Notions of Convergence

Probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\{X_n\}$ — a sequence of random variables, $\mu_n$ — the distribution of $X_n$. $X$ — another random variable with distribution $\mu$.

Definition (Almost sure convergence)

$X_n$ converges to $X$ almost surely as $n \to \infty$, $(X_n \to X, \text{a.s.})$ if

$$\mathbb{P}\{\omega \in \Omega, \ X_n(\omega) \to X(\omega)\} = 1$$
Notions of Convergence

Probability space \((\Omega, \mathcal{F}, \mathbb{P})\), \(\{X_n\}\) — a sequence of random variables, \(\mu_n\) — the distribution of \(X_n\). \(X\) — another random variable with distribution \(\mu\).

Definition (Almost sure convergence)
\(X_n\) converges to \(X\) almost surely as \(n \to \infty\), \((X_n \to X, \text{ a.s.})\) if

\[
\mathbb{P}\{\omega \in \Omega, \ X_n(\omega) \to X(\omega)\} = 1
\]

Definition (Convergence in probability)
\(X_n\) converges to \(X\) in probability if for any \(\epsilon > 0\),

\[
\mathbb{P}\{\omega |X_n(\omega) - X(\omega)| > \epsilon\} \to 0
\]

as \(n \to +\infty\).
Notions of Convergence

Definition (Convergence in distribution)

$X_n$ converges to $X$ in distribution ($X_n \xrightarrow{d} X$) (i.e. $\mu_n \rightarrow \mu$ or $\mu_n \xrightarrow{d} \mu$, weak convergence), if for any bounded continuous function $f$

$$\mathbb{E}f(X_n) \rightarrow \mathbb{E}f(X).$$
Notions of Convergence

Definition (Convergence in distribution)

$X_n$ converges to $X$ in distribution $(X_n \xrightarrow{d} X)$ (i.e. $\mu_n \rightarrow \mu$ or $\mu_n \xrightarrow{d} \mu$, weak convergence), if for any bounded continuous function $f$

$$\mathbb{E}f(X_n) \rightarrow \mathbb{E}f(X).$$

Definition (Convergence in $L^p$)

If $X_n, X \in L^p$, and

$$\mathbb{E}|X_n - X|^p \rightarrow 0.$$

If $p = 1$, that is convergence in mean; if $p = 2$, that is convergence in mean square.
Relation between different convergence concepts

**Relation:**

Almost sure convergence \(\iff\) Converge in probability \(\rightarrow\) Converge in distribution

\[\uparrow\]

\(L^p\) convergence
Conditional Expectation: Naive definition

Let $X$ and $Y$ be two discrete random variables with joint probability

$$p(i, j) = \mathbb{P}(X = i, Y = j).$$
Conditional Expectation: Naive definition

Let $X$ and $Y$ be two discrete random variables with joint probability

$$p(i, j) = \mathbb{P}(X = i, Y = j).$$

The *conditional probability* that $X = i$ given that $Y = j$ is given by

$$p(i|j) = \frac{p(i, j)}{\sum_i p(i, j)} = \frac{p(i, j)}{\mathbb{P}(Y = j)}$$

if $\sum_i p(i, j) > 0$ and conventionally taken to be zero if $\sum_i p(i, j) = 0$. 
Let $X$ and $Y$ be two discrete random variables with joint probability
\[ p(i, j) = \mathbb{P}(X = i, Y = j). \]

The conditional probability that $X = i$ given that $Y = j$ is given by
\[ p(i|j) = \frac{p(i, j)}{\sum_i p(i, j)} = \frac{p(i, j)}{\mathbb{P}(Y = j)} \]
if $\sum_i p(i, j) > 0$ and conventionally taken to be zero if $\sum_i p(i, j) = 0$.

The natural definition of the conditional expectation of $f(X)$ given that $Y = j$ is
\[ \mathbb{E}(f(X)|Y = j) = \sum_i f(i)p(i|j). \]
Conditional Expectation: Abstract definition

- The axiomatic definition of the conditional expectation $Z = E(X|G)$ is defined with respect to a sub-$\sigma$-algebra $G \subset \mathcal{F}$ as follows.
The axiomatic definition of the conditional expectation $Z = E(X|\mathcal{G})$ is defined with respect to a sub-$\sigma$-algebra $\mathcal{G} \subset \mathcal{F}$ as follows.

**Definition (Conditional expectation)**
For any random variable $X$ with $\mathbb{E}|X| < \infty$, the conditional expectation $Z$ of $X$ given $\mathcal{G}$ is defined as
The axiomatic definition of the conditional expectation \( Z = E(X|\mathcal{G}) \) is defined with respect to a sub-\( \sigma \)-algebra \( \mathcal{G} \subset \mathcal{F} \) as follows.

**Definition (Conditional expectation)**

For any random variable \( X \) with \( \mathbb{E}|X| < \infty \), the conditional expectation \( Z \) of \( X \) given \( \mathcal{G} \) is defined as

(i) \( Z \) is a random variable which is measurable with respect to \( \mathcal{G} \);
The axiomatic definition of the conditional expectation $Z = E(X|G)$ is defined with respect to a sub-$\sigma$-algebra $G \subset \mathcal{F}$ as follows.

**Definition (Conditional expectation)**

For any random variable $X$ with $\mathbb{E}|X| < \infty$, the conditional expectation $Z$ of $X$ given $G$ is defined as

(i) $Z$ is a random variable which is measurable with respect to $G$;

(ii) for any set $A \in G$,

$$\int_A Z(\omega) \mathbb{P}(d\omega) = \int_A X(\omega) \mathbb{P}(d\omega).$$
The existence of $Z = E(X|\mathcal{G})$ comes from the Radon-Nikodym theorem by considering the measure $\mu$ on $\mathcal{G}$ defined by $\mu(A) = \int_A X(\omega)P(d\omega)$ (see Billingsley: Probability and measure).
Conditional Expectation: Existence

- The existence of $Z = E(X|\mathcal{G})$ comes from the Radon-Nikodym theorem by considering the measure $\mu$ on $\mathcal{G}$ defined by $\mu(A) = \int_A X(\omega)P(d\omega)$ (see Billingsley: Probability and measure).

- One can easily find that $\mu$ is absolutely continuous with respect to the measure $P|_{\mathcal{G}}$, the probability measure confined in $\mathcal{G}$. Thus $Z$ exists and is unique up to the almost sure equivalence in $P|_{\mathcal{G}}$. 
The existence of $Z = E(X|\mathcal{G})$ comes from the Radon-Nikodym theorem by considering the measure $\mu$ on $\mathcal{G}$ defined by $\mu(A) = \int_A X(\omega) P(d\omega)$ (see Billingsley: Probability and measure).

One can easily find that $\mu$ is absolutely continuous with respect to the measure $P|_\mathcal{G}$, the probability measure confined in $\mathcal{G}$. Thus $Z$ exists and is unique up to the almost sure equivalence in $P|_\mathcal{G}$.

For the conditional expectation of a random variable $X$ with respect to another random variable $Y$, it is natural to define it as

$$\mathbb{E}(X|Y) := \mathbb{E}(X|\mathcal{G})$$

where $\mathcal{G}$ is the $\sigma$-algebra $Y^{-1}(\mathcal{B})$ generated by $Y$. 


Theorem (Properties of conditional expectation)

Suppose $X, Y$ are random variables with $\mathbb{E}|X|, \mathbb{E}|Y| < \infty$, $a, b \in \mathbb{R}$. Then
Theorem (Properties of conditional expectation)

Suppose $X, Y$ are random variables with $\mathbb{E}|X|, \mathbb{E}|Y| < \infty$, $a, b \in \mathbb{R}$. Then

(i) $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$
Theorem (Properties of conditional expectation)

Suppose $X, Y$ are random variables with $E|X|, E|Y| < \infty$, $a, b \in \mathbb{R}$. Then

(i) $E(aX + bY|\mathcal{G}) = aE(X|\mathcal{G}) + bE(Y|\mathcal{G})$

(ii) $E(E(X|\mathcal{G})) = E(X)$
Conditional Expectation: Properties

Theorem (Properties of conditional expectation)

Suppose \( X, Y \) are random variables with \( \mathbb{E}|X|, \mathbb{E}|Y| < \infty \), \( a, b \in \mathbb{R} \). Then

(i) \( \mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G}) \)

(ii) \( \mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X) \)

(iii) \( \mathbb{E}(X|\mathcal{G}) = X \), if \( X \) is \( \mathcal{G} \)-measurable
Conditional Expectation: Properties

Theorem (Properties of conditional expectation)
Suppose \( X, Y \) are random variables with \( \mathbb{E}|X|, \mathbb{E}|Y| < \infty \), \( a, b \in \mathbb{R} \). Then

(i) \( \mathbb{E}(aX + bY|G) = a\mathbb{E}(X|G) + b\mathbb{E}(Y|G) \)

(ii) \( \mathbb{E}(\mathbb{E}(X|G)) = \mathbb{E}(X) \)

(iii) \( \mathbb{E}(X|G) = X \), if \( X \) is \( G \)-measurable

(iv) \( \mathbb{E}(X|G) = \mathbb{E}X \), if \( X \) is independent of \( G \)
Conditional Expectation: Properties

Theorem (Properties of conditional expectation)

Suppose \( X, Y \) are random variables with \( \mathbb{E}|X|, \mathbb{E}|Y| < \infty \), \( a, b \in \mathbb{R} \). Then

(i) \( \mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G}) \)

(ii) \( \mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X) \)

(iii) \( \mathbb{E}(X|\mathcal{G}) = X \), if \( X \) is \( \mathcal{G} \)-measurable

(iv) \( \mathbb{E}(X|\mathcal{G}) = \mathbb{E}X \), if \( X \) is independent of \( \mathcal{G} \)

(v) \( \mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G}) \), if \( Y \) is \( \mathcal{G} \)-measurable
Theorem (Properties of conditional expectation)

Suppose $X$, $Y$ are random variables with $\mathbb{E}|X|, \mathbb{E}|Y| < \infty$, $a, b \in \mathbb{R}$. Then

(i) $\mathbb{E}(aX + bY | G) = a\mathbb{E}(X | G) + b\mathbb{E}(Y | G)$

(ii) $\mathbb{E}(\mathbb{E}(X | G)) = \mathbb{E}(X)$

(iii) $\mathbb{E}(X | G) = X$, if $X$ is $G$-measurable

(iv) $\mathbb{E}(X | G) = \mathbb{E}X$, if $X$ is independent of $G$

(v) $\mathbb{E}(XY | G) = Y\mathbb{E}(X | G)$, if $Y$ is $G$-measurable

(vi) $\mathbb{E}(X | G) = \mathbb{E}(\mathbb{E}(X | H) | G)$ for the sub-$\sigma$-algebras $G \subset H$. 

Lemma (Conditional Jensen’s inequality)

Let $X$ be a random variable such that $\mathbb{E}|X| < \infty$ and $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function such that $\mathbb{E} |\phi(X)| < \infty$. Then

$$\mathbb{E}(\phi(X) | \mathcal{G}) \geq \phi(\mathbb{E}(X | \mathcal{G})).$$

The readers may be referred to (K.L. Chung: A course in probability theory) for the details of the proof.
To realize the equivalence between the abstract definition
\[ \mathbb{E}(X|Y) := \mathbb{E}(X|\mathcal{G}) \]
and
\[ \mathbb{E}(f(X)|Y = j) = \sum_i f(i)p(i|j) \]
when \( Y \) only takes finitely discrete values, we suppose the following decomposition

\[ \Omega = \bigcup_{j=1}^{n} \Omega_j \]

and \( \Omega_j = \{ \omega : Y(\omega) = j \} \). Then the \( \sigma \)-algebra \( \mathcal{G} \) is simply the sets of all possible unions of \( \Omega_j \).
Conditional Expectation: Abstract vs Naive definition

To realize the equivalence between the abstract definition
\[ E(X|Y) := E(X|G) \] and \[ E(f(X)|Y = j) = \sum_i f(i)p(i|j) \]
when \( Y \) only takes finitely discrete values, we suppose the following decomposition

\[ \Omega = \bigcup_{j=1}^{n} \Omega_j \]

and \( \Omega_j = \{\omega : Y(\omega) = j\} \). Then the \( \sigma \)-algebra \( G \) is simply the sets of all possible unions of \( \Omega_j \).

The measurability of conditional expectation \( E(X|Y) \) with respect to \( G \) means \( E(X|Y) \) takes constant on each \( \Omega_j \), which exactly corresponds to \( E(X|Y = j) \) as we will see.
Conditional Expectation: Abstract vs Naive definition

By definition, we have

\[ \int_{\Omega_j} \mathbb{E}(X|Y) \mathbb{P}(d\omega) = \int_{\Omega_j} X(\omega) \mathbb{P}(d\omega) \]

which implies

\[ \mathbb{E}(X|Y) = \frac{1}{\mathbb{P}(\Omega_j)} \int_{\Omega_j} X(\omega) \mathbb{P}(d\omega). \]

This is exactly \( \mathbb{E}(X|Y = j) \) when \( f(X) = X \) and \( X \) also takes discrete values.
Conditional Expectation: Optimal Approximation

The conditional expectation has the following important property as the optimal approximation in $L^2$ norm among all of the $Y$-measurable functions.

**Proposition**

Let $g(Y)$ be any measurable function of $Y$, then

$$\mathbb{E}(X - \mathbb{E}(X|Y))^2 \leq \mathbb{E}(X - g(Y))^2.$$
Conditional Expectation: Optimal Approximation

Proof.
We have

$$\mathbb{E}(X - g(Y))^2 = \mathbb{E}(X - E(X|Y))^2 + \mathbb{E}(E(X|Y) - g(Y))^2$$
$$+ 2\mathbb{E}\left[(X - E(X|Y))(E(X|Y) - g(Y))\right].$$

and

$$\mathbb{E}\left[(X - \mathbb{E}(X|Y))\mathbb{E}(X|Y) - g(Y))\right]$$
$$= \mathbb{E}\left[\mathbb{E}\left[(X - \mathbb{E}(X|Y))(E(X|Y) - g(Y))|Y\right]\right]$$
$$= \mathbb{E}\left[\mathbb{E}(X|Y) - \mathbb{E}(X|Y))(E(X|Y) - g(Y))\right] = 0$$

by properties (ii),(iii) and (v) in properties of conditional expectation. The proof is done.
Table of Contents

Elementary Random Variables

Axiomatic Probability Theory Setup

Conditional Expectation

Characteristic and Generating Functions

Borel-Cantelli Lemma
The *characteristic function* of a random variable $X$ or its distribution $\mu$ is defined as

$$f(\xi) = \mathbb{E}e^{i\xi X} = \int_{\mathbb{R}} e^{i\xi x} \mu(dx).$$
The characteristic function of a random variable $X$ or its distribution $\mu$ is defined as

$$f(\xi) = \mathbb{E}e^{i\xi X} = \int_{\mathbb{R}} e^{i\xi x} \mu(dx).$$

Obviously, when $X$, $Y$ are independent and has characteristic functions $f(\xi), g(\xi)$, then we have the characteristic function for $Z = X + Y$

$$h(\xi) = \mathbb{E}e^{i\xi Z} = \mathbb{E}e^{i\xi(X+Y)} = f(\xi)g(\xi).$$
Characteristic Function: Examples

The characteristic functions of some typical distributions are as below.

- Bernoulli distribution: \( f(\xi) = q + pe^{i\xi} \).

- Binomial distribution \( B(n, p) \): \( f(\xi) = (q + pe^{i\xi})^n \).

- Poisson distribution \( P(\lambda) \): \( f(\xi) = e^{\lambda(e^{i\xi} - 1)} \).

- Exponential distribution \( \text{Exp}(\lambda) \): \( f(\xi) = (1 - \lambda - i\xi)^{-1} \).

- Normal distribution \( N(\mu, \sigma^2) \): \( f(\xi) = \exp(i\mu\xi - \frac{\sigma^2}{2}\xi^2) \).
The characteristic functions of some typical distributions are as below.

- Bernoulli distribution: \( f(\xi) = q + pe^{i\xi} \).
- Binomial distribution \( B(n, p) \): \( f(\xi) = (q + pe^{i\xi})^n \).
The characteristic functions of some typical distributions are as below.

- **Bernoulli distribution**: \( f(\xi) = q + pe^{i\xi} \).
- **Binomial distribution** \( B(n, p) \): \( f(\xi) = (q + pe^{i\xi})^n \).
- **Poisson distribution** \( \mathcal{P}(\lambda) \): \( f(\xi) = e^{\lambda(e^{i\xi} - 1)} \).
The characteristic functions of some typical distributions are as below.

- **Bernoulli distribution**: \( f(\xi) = q + pe^{i\xi} \).
- **Binomial distribution** \( B(n,p) \): \( f(\xi) = (q + pe^{i\xi})^n \).
- **Poisson distribution** \( \mathcal{P}(\lambda) \): \( f(\xi) = e^{\lambda(e^{i\xi} - 1)} \).
- **Exponential distribution** \( \mathcal{E}xp(\lambda) \): \( f(\xi) = (1 - \lambda^{-1}i\xi)^{-1} \).
The characteristic functions of some typical distributions are as below.

- Bernoulli distribution: $f(\xi) = q + pe^{i\xi}$.
- Binomial distribution $B(n, p)$: $f(\xi) = (q + pe^{i\xi})^n$.
- Poisson distribution $\mathcal{P}(\lambda)$: $f(\xi) = e^{\lambda(e^{i\xi}-1)}$.
- Exponential distribution $\mathcal{E}xp(\lambda)$: $f(\xi) = (1 - \lambda^{-1}i\xi)^{-1}$.
- Normal distribution $N(\mu, \sigma^2)$: $f(\xi) = \exp\left(i\mu\xi - \frac{\sigma^2\xi^2}{2}\right)$. 
Proposition

The characteristic function has the following properties:

1. \( \forall \xi \in \mathbb{R}, |f(\xi)| \leq 1 \), \( f(\xi) = f(-\xi) \), \( f(0) = 1 \);
2. \( f \) is uniformly continuous on \( \mathbb{R} \);
3. \( f(n) (0) = i^n E_X^n \) provided \( E|X|^n < \infty \).

Proof. The proof of statements 1 and 3 are straightforward. The second statement is valid by
\[
|f(\xi_1) - f(\xi_2)| = |E(e^{i\xi_1 X} - e^{i\xi_2 X})| = |E(e^{i\xi_1 X}(1 - e^{i(\xi_2 - \xi_1)X})| \\
\leq E|1 - e^{i(\xi_2 - \xi_1)X}|
\]
Dominated convergence theorem concludes the proof.
Proposition

The characteristic function has the following properties:

1. $\forall \xi \in \mathbb{R}, \ |f(\xi)| \leq 1$, $f(\xi) = f(-\xi)$, $f(0) = 1$;
Characteristic Function: Property

Proposition

The characteristic function has the following properties:

1. $\forall \xi \in \mathbb{R}, |f(\xi)| \leq 1, f(\xi) = f(-\xi), f(0) = 1$;
2. $f$ is uniformly continuous on $\mathbb{R}$.
Proposition

The characteristic function has the following properties:

1. \( \forall \xi \in \mathbb{R}, \left| f(\xi) \right| \leq 1, f(\xi) = f(-\xi), f(0) = 1; \)
2. \( f \) is uniformly continuous on \( \mathbb{R}; \)
3. \( f^{(n)}(0) = i^n \mathbb{E}X^n \) provided \( \mathbb{E}|X|^n < \infty. \)
Proposition

The characteristic function has the following properties:

1. \( \forall \xi \in \mathbb{R}, \ |f(\xi)| \leq 1, \ f(\xi) = f(-\xi), \ f(0) = 1; \)
2. \( f \) is uniformly continuous on \( \mathbb{R}; \)
3. \( f^{(n)}(0) = i^n \mathbb{E}X^n \) provided \( \mathbb{E}|X|^n < \infty. \)

Proof.

The proof of statements 1 and 3 are straightforward. The second statement is valid by

\[
|f(\xi_1) - f(\xi_2)| = |\mathbb{E}(e^{i\xi_1 X} - e^{i\xi_2 X})| = |\mathbb{E}(e^{i\xi_1 X} (1 - e^{i(\xi_2-\xi_1)X}))|
\leq \mathbb{E}|1 - e^{i(\xi_2-\xi_1)X}|.
\]

Dominated convergence theorem concludes the proof. \( \square \)
Lévy’s continuity theorem

Theorem (Lévy’s continuity theorem)

Let \( \{\mu_n\}_{n \in \mathbb{N}} \) be a sequence of probability measures, and \( \{f_n\}_{n \in \mathbb{N}} \) be their corresponding characteristic functions.

Assume that

1. \( f_n \) converges everywhere on \( \mathbb{R} \) to a limiting function \( f \).
2. \( f \) is continuous at \( \xi = 0 \).

Then there exists a probability distribution \( \mu \) such that \( \mu_n \rightharpoonup \mu \).

Moreover, \( f \) is the characteristic function of \( \mu \).

Conversely, if \( \mu_n \rightharpoonup \mu \), where \( \mu \) is some probability distribution, then \( f_n \) converges to \( f \) uniformly in every finite interval, where \( f \) is the characteristic function of \( \mu \).

For a proof, see K.L. Chung: A course in probability theory.
Lévy’s continuity theorem

Theorem (Lévy’s continuity theorem)
Let \( \{\mu_n\}_{n \in \mathbb{N}} \) be a sequence of probability measures, and \( \{f_n\}_{n \in \mathbb{N}} \) be their corresponding characteristic functions. Assume that

1. \( f_n \) converges everywhere on \( \mathbb{R} \) to a limiting function \( f \).

Moreover, \( f \) is the characteristic function of \( \mu \). Conversely, if \( \mu_n \rightharpoonup \mu \) where \( \mu \) is some probability distribution, then \( f_n \) converges to \( f \) uniformly in every finite interval, where \( f \) is the characteristic function of \( \mu \).
Lévy’s continuity theorem

Theorem (Lévy’s continuity theorem)

Let \( \{ \mu_n \}_{n \in \mathbb{N}} \) be a sequence of probability measures, and \( \{ f_n \}_{n \in \mathbb{N}} \) be their corresponding characteristic functions. Assume that

1. \( f_n \) converges everywhere on \( \mathbb{R} \) to a limiting function \( f \).
2. \( f \) is continuous at \( \xi = 0 \).

Then there exists a probability distribution \( \mu \) such that \( \mu_n \xrightarrow{d} \mu \). Moreover, \( f \) is the characteristic function of \( \mu \).

Conversely, if \( \mu_n \xrightarrow{d} \mu \), where \( \mu \) is some probability distribution then \( f_n \) converges to \( f \) uniformly in every finite interval, where \( f \) is the characteristic function of \( \mu \).

For a proof, see K.L. Chung: A course in probability theory.
Lévy’s continuity theorem

Theorem (Lévy’s continuity theorem)

Let \( \{ \mu_n \}_{n \in \mathbb{N}} \) be a sequence of probability measures, and \( \{ f_n \}_{n \in \mathbb{N}} \) be their corresponding characteristic functions. Assume that

1. \( f_n \) converges everywhere on \( \mathbb{R} \) to a limiting function \( f \).
2. \( f \) is continuous at \( \xi = 0 \).

Then there exists a probability distribution \( \mu \) such that \( \mu_u \xrightarrow{d} \mu \). Moreover \( f \) is the characteristic function of \( \mu \).
Lévy’s continuity theorem

Theorem (Lévy’s continuity theorem)

Let \( \{\mu_n\}_{n \in \mathbb{N}} \) be a sequence of probability measures, and \( \{f_n\}_{n \in \mathbb{N}} \) be their corresponding characteristic functions. Assume that

1. \( f_n \) converges everywhere on \( \mathbb{R} \) to a limiting function \( f \).
2. \( f \) is continuous at \( \xi = 0 \).

Then there exists a probability distribution \( \mu \) such that \( \mu_n \xrightarrow{d} \mu \). Moreover \( f \) is the characteristic function of \( \mu \).

Conversely, if \( \mu_n \xrightarrow{d} \mu \), where \( \mu \) is some probability distribution then \( f_n \) converges to \( f \) uniformly in every finite interval, where \( f \) is the characteristic function of \( \mu \).

For a proof, see K.L. Chung: A course in probability theory.
Characteristic Function: Positive Semi-definite Function

As in Fourier transforms, one can also define the inverse transform

\[ \rho(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\xi x} f(\xi) d\xi. \]

An interesting question arises as to when this gives the density of a probability measure. To answer this we define

Definition
A function \( f \) is called positive semi-definite if for any finite set of values \( \{\xi_1, \ldots, \xi_n\} \), \( n \in \mathbb{N} \), the matrix \( (f(\xi_i - \xi_j))_{i,j=1}^n \) is positive semi-definite, i.e.

\[ \sum_{i,j} f(\xi_i - \xi_j)v_i \bar{v}_j \geq 0, \]

for any \( v_1, \ldots, v_n \in \mathbb{C} \).
Bochner’s Theorem

Theorem (Bochner’s Theorem)

A function $f$ is the characteristic function of a probability measure if and only if it is a positive semi-definite and continuous at 0 with $f(0) = 1$. 

Proof. We only give the necessity part. Suppose $f$ is a characteristic function, then

$$\sum_{i,j=1}^{n} f(\xi_i - \xi_j) v_i \overline{v_j} = \int \left| \sum_{i=1}^{n} v_i e^{i \xi_i x} \right|^2 \mu(dx) \geq 0.$$ 

The sufficiency part is difficult and the readers may be referred to (K.L. Chung: A course in probability theory).
Bochner’s Theorem

Theorem (Bochner’s Theorem)

A function $f$ is the characteristic function of a probability measure if and only if it is a positive semi-definite and continuous at 0 with $f(0) = 1$.

Proof.
We only gives the necessity part. Suppose $f$ is a characteristic function, then

$$
\sum_{i,j=1}^{n} f(\xi_i - \xi_j) v_i \bar{v}_j = \int_{\mathbb{R}} \left| \sum_{i=1}^{n} v_i e^{i\xi_i x} \right|^2 \mu(dx) \geq 0.
$$

The sufficiency part is difficult and the readers may be referred to (K.L. Chung: A course in probability theory).
Generating function

For discrete R.V. taking integer values, the generating function has the central importance

\[ G(x) = \sum_{k=0}^{\infty} P(k) x^k. \]

One immediately has the formula:

\[ P(k) = \frac{1}{k!} G^{(k)}(0). \]

Some generating functions:

- Bernoulli distribution: \( G(x) = q + px. \)
Generating function

For discrete R.V. taking integer values, the generating function has the central importance

\[ G(x) = \sum_{k=0}^{\infty} P(k) x^k. \]

One immediately has the formula:

\[ P(k) = \frac{1}{k!} G^{(k)}(x) \bigg|_{x=0}. \]

Some generating functions:

- Bernoulli distribution: \( G(x) = q + px. \)
- Binomial distribution: \( G(x) = (q + px)^n. \)
Generating function

For discrete R.V. taking integer values, the generating function has the central importance

$$G(x) = \sum_{k=0}^{\infty} P(k) x^k.$$ 

One immediately has the formula:

$$P(k) = \frac{1}{k!} G^{(k)}(x) \bigg|_{x=0}. $$

Some generating functions:

- Bernoulli distribution: $G(x) = q + px$.
- Binomial distribution: $G(x) = (q + px)^n$.
- Poisson distribution: $G(x) = e^{-\lambda} + \lambda x$. 
Generating function

Definition
Define the convolution of two sequences \( \{a_k\}, \{b_k\} \) as \( \{c_k\} = \{a_k\} * \{b_k\} \), the components are defined as

\[
c_k = \sum_{j=0}^{k} a_j b_{k-j}.
\]
Generating function

**Definition**

Define the convolution of two sequences \( \{ a_k \}, \{ b_k \} \) as \( \{ c_k \} = \{ a_k \} \ast \{ b_k \}, \) the components are defined as

\[
c_k = \sum_{j=0}^{k} a_j b_{k-j}.
\]

**Theorem**

*Consider two independent R.V. \( X \) and \( Y \) with PMF

\[
P(X = j) = a_j, \quad P(Y = k) = b_k
\]

and \( \{ c_k \} = \{ a_k \} \ast \{ b_k \}. \) Suppose the generating functions are \( A(x), B(x) \) and \( C(x), \) respectively, then the generating function of \( X + Y \) is \( C(x). \)
Moment Generating Function

▶ The moment generating function of a random variable \( X \) is defined for all values of \( t \) by

\[
M(t) = \mathbb{E}e^{tX} = \begin{cases} 
\sum_{x} p(x)e^{tx}, & X \text{ is discrete-valued} \\
\int_{\mathbb{R}} p(x)e^{tx} \, dx, & X \text{ is continuous}
\end{cases}
\]

provided that \( e^{tx} \) is integrable. It is obvious \( M(0) = 1 \).
Moment Generating Function

The moment generating function of a random variable $X$ is defined for all values of $t$ by

$$M(t) = \mathbb{E}e^{tx} = \begin{cases} \sum_x p(x)e^{tx}, & X \text{ is discrete-valued} \\ \int_{\mathbb{R}} p(x)e^{tx} dx, & X \text{ is continuous} \end{cases}$$

provided that $e^{tx}$ is integrable. It is obvious $M(0) = 1$.

Once $M(t)$ can be defined, one can show $M(t) \in C^\infty$ in its domain and its relation to the $n$th moments

$$M^{(n)}(t) = \mathbb{E}(X^n e^{tx})$$

and $\mu_n := \mathbb{E}X^n = M^{(n)}(0), \ n \in \mathbb{N}.$

This gives

$$M(t) = \sum_{n=0}^{\infty} \mu_n \frac{t^n}{n!},$$

which tells why $M(t)$ is called the moment generating function.
Moment Generating Function: Property

Theorem
Denote $M_X(t)$, $M_Y(t)$ and $M_{X+Y}(t)$ the moment generating functions of random variables $X, Y$ and $X + Y$, respectively. If $X, Y$ are independent, we have

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

The proof is straightforward.
The following moment generating functions of typical random variables can be obtained by direct calculations.

(a) Binomial distribution: \( M(t) = (pe^t + 1 - p)^n \).
The following moment generating functions of typical random variables can be obtained by direct calculations.

(a) Binomial distribution: \( M(t) = (pe^t + 1 - p)^n \).

(b) Poisson distribution: \( M(t) = \exp[\lambda(e^t - 1)] \).
The following moment generating functions of typical random variables can be obtained by direct calculations.

(a) Binomial distribution: \( M(t) = (pe^t + 1 - p)^n \).

(b) Poisson distribution: \( M(t) = \exp[\lambda(e^t - 1)] \).

(c) Exponential distribution: \( M(t) = \lambda/(\lambda - t) \) for \( t < \lambda \).
The following moment generating functions of typical random variables can be obtained by direct calculations.

(a) Binomial distribution: \( M(t) = (pe^t + 1 - p)^n \).

(b) Poisson distribution: \( M(t) = \exp[\lambda(e^t - 1)] \).

(c) Exponential distribution: \( M(t) = \lambda/\left(\lambda - t\right) \) for \( t < \lambda \).

(d) Normal distribution \( N(\mu, \sigma^2) \): \( M(t) = \exp \left(\mu t + \frac{\sigma^2 t^2}{2}\right) \).
The cumulant generating function $K(t)$ is defined based on $M(t)$ by

$$K(t) = \ln M(t) = \ln \mathbb{E} e^{tX} = \sum_{n=1}^{\infty} \frac{\kappa_n t^n}{n!}.$$ 

With such definition, we have the cumulants $\kappa_0 = 0$ and

$$\kappa_n = K^{(n)}(0), \quad n \in \mathbb{N}.$$
Cumulants Generating Function

- The cumulant generating function $K(t)$ is defined based on $M(t)$ by

$$K(t) = \ln M(t) = \ln \mathbb{E}e^{tX} = \sum_{n=1}^{\infty} \frac{\kappa_n t^n}{n!}.$$  

With such definition, we have the cumulants $\kappa_0 = 0$ and $\kappa_n = K^{(n)}(0), \quad n \in \mathbb{N}$.

- The moment and cumulant generating functions have explicit meaning in statistical physics, in which

$$Z(\beta) = \mathbb{E}e^{-\beta E}, \quad F(\beta) = -\beta^{-1} \ln Z(\beta)$$

are called partition function and Helmholtz free energy, respectively. They can be connected to $M$ and $K$ by

$$Z(\beta) = M_X(-\beta), \quad F(\beta) = -\beta^{-1}K_X(-\beta)$$

if $X$ is taken as $E$, the energy of the system.
Table of Contents

Elementary Random Variables

Axiomatic Probability Theory Setup

Conditional Expectation

Characteristic and Generating Functions

Borel-Cantelli Lemma
Let \( \{ A_n \} \) be a sequence of events, \( A_n \in \mathcal{F} \). Define

\[
\limsup_{n \to \infty} (A_n) = \{ \omega \in \Omega, \ \omega \in A_n \ \text{infinitely often (i.o.)} \}
\]

\[
= \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k
\]

Question: What is the set

\[
\liminf_{n \to \infty} (A_n) := \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k
\]
Lemma (First Borel-Cantelli Lemma)

If \( \sum_{n=1}^{\infty} \mathbb{P}(A_n) < +\infty \), then

\[
\mathbb{P}(\limsup_{n \to \infty} A_n) = \mathbb{P}\{\omega : \omega \in A_n, \text{i.o.}\} = 0.
\]
First Borel-Cantelli Lemma

Lemma (First Borel-Cantelli Lemma)

If \( \sum_{n=1}^{\infty} P(A_n) < +\infty \), then

\[
P(\limsup_{n \to \infty} A_n) = P\{\omega : \omega \in A_n, \text{i.o.}\} = 0.
\]

Proof.
We have

\[
P\left\{ \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \right\} \leq P\left\{ \bigcup_{k=n}^{\infty} A_k \right\} \leq \sum_{k=n}^{\infty} P(A_k)
\]

for any \( n \), but the last term goes to 0, as \( n \to \infty \). \( \square \)
As an example of the application of this result, we prove

**Proposition (BCL-Application)**

Let \( \{X_n\} \) be a sequence of identically distributed (not necessarily independent) random variables, such that

\[
\mathbb{E}|X_n| < +\infty.
\]

Then

\[
\lim_{n \to \infty} \frac{X_n}{n} = 0 \quad \text{a.s.}
\]
Lemma (Chebyshev Inequality)

Let $X$ be a random variable such that $\mathbb{E}|X|^k < +\infty$, for some integer $k$. Then

$$P\{|X| > \lambda\} \leq \frac{1}{\lambda^k} \mathbb{E}|X|^k$$

for any positive constant $\lambda$.

Proof. For any $\lambda > 0$,

$$\mathbb{E}|X|^k = \int_{-\infty}^{\infty} |x|^k d\mu \geq \int_{|X| \geq \lambda} |X|^k d\mu$$

$$\geq \lambda^k \int_{|X| \geq \lambda} d\mu = \lambda^k P\{|X| \geq \lambda\}.$$
Proof. For any $\epsilon > 0$, define

$$A_n = \{ \omega \in \Omega : \left| \frac{X_n(\omega)}{n} \right| > \epsilon \}$$

$$\sum_n P(A_n) = \sum_n P\{|X_n| > n\epsilon\}$$

$$= \sum_n \sum_{k=n} P\{k\epsilon < |X_n| < (k + 1)\epsilon\}$$

$$= \sum_k kP\{k\epsilon < |X_n| < (k + 1)\epsilon\}$$

$$\leq \frac{1}{\epsilon} \mathbb{E}|X| < +\infty$$
Therefore if we define

\[ B_\epsilon = \{ \omega \in \Omega, \quad \omega \in A_n \text{ i.o.} \} \]

then \( P(B_\epsilon) = 0 \). Let \( B = \bigcup_{n=1}^{\infty} B_{\frac{1}{n}} \). Then \( P(B) = 0 \), and

\[ \lim_{n \to \infty} \frac{X_n(\omega)}{n} = 0, \quad \text{if } \omega \notin B. \]

The proof is done.
Convergence in Probability implies A.S. Convergence in subsequence: Proof

Here we give the proof by 1st BCL lemma. Without loss of generality (W.L.G.), we assume \( X = 0 \).
Convergence in Probability implies A.S. Convergence in subsequence: Proof

Here we give the proof by 1st BCL lemma. Without loss of generality (W.L.G.), we assume $X = 0$.

Convergence in probability implies that for any $k$, we can choose subsequence $X_{n_k}$ ($n_k$ is increasing in $k$) such that

$$
\mathbb{P}(X_{n_k} \geq 1/k) \leq 1/2^k
$$
Convergence in Probability implies A.S. Convergence in subsequence: Proof

Here we give the proof by 1st BCL lemma. Without loss of generality (W.L.G.), we assume $X = 0$.

- Convergence in probability implies that for any $k$, we can choose subsequence $X_{n_k}$ ($n_k$ is increasing in $k$) such that

$$\mathbb{P}(X_{n_k} \geq 1/k) \leq 1/2^k$$

- For any $\epsilon > 0$, we have

$$\sum_{k=1}^{\infty} \mathbb{P}(|X_{n_k}| \geq \epsilon) = \sum_{k=1}^{k_{\epsilon}} + \sum_{k=k_{\epsilon}}^{\infty} \mathbb{P}(|X_{n_k}| \geq \epsilon) < \infty, \quad 1/k_{\epsilon} \leq \epsilon$$
Convergence in Probability implies A.S. Convergence in subsequence: Proof

Here we give the proof by 1st BCL lemma. Without loss of generality (W.L.G.), we assume $X = 0$.

- Convergence in probability implies that for any $k$, we can choose subsequence $X_{n_k}$ ($n_k$ is increasing in $k$) such that

$$
P(X_{n_k} \geq 1/k) \leq 1/2^k$$

- For any $\epsilon > 0$, we have

$$
\sum_{k=1}^{\infty} P(|X_{n_k}| \geq \epsilon) = \sum_{k=1}^{k_\epsilon} + \sum_{k=k_\epsilon}^{\infty} P(|X_{n_k}| \geq \epsilon) < \infty, \quad 1/k_\epsilon \leq \epsilon
$$

- From the 1st BCL lemma, we have

$$
P(|X_{n_k}| \geq \epsilon, i.o.) = 0 \quad \text{for any } \epsilon > 0$$
Convergence in Probability implies A.S. Convergence in subsequence: Proof

Here we give the proof by 1st BCL lemma. Without loss of generality (W.L.G.), we assume $X = 0$.

- Convergence in probability implies that for any $k$, we can choose subsequence $X_{n_k}$ ($n_k$ is increasing in $k$) such that

  $$\mathbb{P}(X_{n_k} \geq 1/k) \leq 1/2^k$$

- For any $\epsilon > 0$, we have

  $$\sum_{k=1}^{\infty} \mathbb{P}(|X_{n_k}| \geq \epsilon) = \sum_{k=1}^{k_{\epsilon}} + \sum_{k=k_{\epsilon}}^{\infty} \mathbb{P}(|X_{n_k}| \geq \epsilon) < \infty, \quad 1/k_{\epsilon} \leq \epsilon$$

- From the 1st BCL lemma, we have

  $$\mathbb{P}(|X_{n_k}| \geq \epsilon, i.o.) = 0 \quad \text{for any } \epsilon > 0$$

- With similar argument as before, we have the almost sure convergence of $\{X_{n_k}\}$ to 0.
Second Borel-Cantelli Lemma

Lemma (Second Borel-Cantelli Lemma)

If $\sum_{n=1}^{\infty} P(A_n) = +\infty$, and $A_n$ are mutually independent, then

$$P\{\omega \in \Omega, \omega \in A_n \text{ i.o.}\} = 1.$$