

# Probability

Notes taken by Runqiu Ye

Lectures by Konstantin Tikhomirov

Carnegie Mellon University

Spring 2026

## Contents

<b>1</b>	<b>Measure theory review</b>	<b>3</b>
1.1	Measurable space and mapping . . . . .	3
1.2	Measure space . . . . .	4
1.3	$\pi$ - $\lambda$ theorem . . . . .	5
1.4	Extension theorems . . . . .	6
1.5	Lebesgue Integration . . . . .	7
1.6	Product measures and Fubini theorem . . . . .	9
<b>2</b>	<b>Probability theory basics</b>	<b>11</b>
2.1	Distributions and densities . . . . .	11
2.2	Independence . . . . .	13
2.3	Convolution . . . . .	15
2.4	Moments . . . . .	16
2.5	Convergence of random variable . . . . .	21
2.6	Law of rare events . . . . .	24
<b>3</b>	<b>Law of large numbers</b>	<b>29</b>

# 1 Measure theory review

## 1.1 Measurable space and mapping

**Definition** ( $\sigma$ -field). A collection of subsets  $\Sigma \subset 2^\Omega$  is a  $\sigma$ -field if

- $\emptyset \in \Sigma$ .
- If  $A \in \Sigma$ , then  $A^c \in \Sigma$ .
- If  $\{A_i\}_{i=1}^\infty \subset \Sigma$ , then  $\bigcup_{i=1}^\infty A_i \in \Sigma$ .

The pair  $(\Omega, \Sigma)$  is called a measurable space.

**Definition** (atom). Let  $\Sigma$  be a  $\sigma$ -field. Say  $A \in \Sigma$  is an atom if for all  $B \in \Sigma$  either  $A \subset B$  or  $A \cap B = \emptyset$ .

**Proposition.** For all  $\omega \in \Omega$ , there exists atom  $A \in \Sigma$  containing  $\omega$  if  $\Omega$  is finite or countable.

*Proof.* Define  $\tilde{A} = \bigcap \{B \in \Sigma : \omega \in B\}$ . We can check that  $\tilde{A} \in \Sigma$  and  $\tilde{A}$  is an atom containing  $\omega$ . □

**Corollary.** If  $\Omega$  is finite or countable, there exists a partition  $\Omega = \bigsqcup_i \Omega_i$ , where each  $\Omega_i$  is an atom of  $\Sigma$ . With this partition,  $\Sigma$  is just the power set with respect to  $\{\Omega_i\}_i$ .

**Definition.** If  $F \subset 2^\Omega$ , then the  $\sigma$ -field generated by  $F$  is the smallest  $\sigma$ -field containing all elements of  $F$ . Write this  $\sigma$ -field as  $\sigma(F)$ .

**Example.** Let  $\Omega = \{1, 2, 3, 4, 5\}$  and  $F = \{\{2, 3\}, \{3, 4\}\}$ . Construct  $\sigma$ -field  $\Sigma$  generated by  $F$ .  $\Sigma$  is all possible union of sets from the collection  $\{\{2\}, \{3\}, \{4\}, \{1, 5\}\}$ . △

**Definition** (measurable mapping). Given two measurable spaces  $(\Omega, \Sigma)$  and  $(\tilde{\Omega}, \tilde{\Sigma})$ . Then  $f : \Omega \rightarrow \tilde{\Omega}$  is measurable if  $f^{-1}(B) \in \Sigma$  for all  $B \in \tilde{\Sigma}$ .

**Definition** (Borel  $\sigma$ -field). Let  $(T, \tau)$  be a topological space. Then the Borel  $\sigma$ -field  $\mathcal{B}(T, \tau)$  is defined as the smallest  $\sigma$ -field containing all open sets.

**Definition** (product measurable space). Given two measurable spaces  $(\Omega, \Sigma)$  and  $(\tilde{\Omega}, \tilde{\Sigma})$ . We can define the product measurable space as follows: let the ground set be  $\Omega \times \tilde{\Omega}$ , and let  $\Sigma \otimes \tilde{\Sigma}$  be the smallest  $\sigma$ -field containing all rectangles  $B \times \tilde{B}$  where  $B \in \Sigma$  and  $\tilde{B} \in \tilde{\Sigma}$ .

More generally, let  $\Lambda$  be an index set and  $(\Omega_\lambda, \Sigma_\lambda)_{\lambda \in \Lambda}$ . Define the product  $\sigma$ -field  $\bigotimes_{\lambda \in \Lambda} \Sigma_\lambda$  be the smallest  $\sigma$ -field containing all elements in the form of  $\prod_{\lambda \in \Lambda} B_\lambda$  where  $B_\lambda \in \Sigma_\lambda$  and  $B_\lambda = \Omega_\lambda$  for all but countably many indices.

**Proposition.** Let  $(\Omega_i, \Sigma_i)_{i=1}^n$  be measurable spaces and  $(\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \Sigma_i)$  be the product space. Let  $(\Omega, \Sigma)$  be the domain and  $f = (f_1, \dots, f_n) : (\Omega, \Sigma) \rightarrow (\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \Sigma_i)$ . Suppose  $f$  is measurable, then every coordinate projection  $f_i : \Omega \rightarrow \Omega_i$  is measurable.

This is also true for arbitrary index set.

**Proposition.** If  $f : (\Omega, \Sigma) \rightarrow (\Omega_f, \Sigma_f)$  and  $g : (\Omega, \Sigma) \rightarrow (\Omega_g, \Sigma_g)$ , then the concatenation  $(f, g)$  is measurable w.r.t. the product space  $(\Omega_f \times \Omega_g, \Sigma_f \otimes \Sigma_g)$ .

*Proof.* Let  $A \times B$  be such that  $A \in \Sigma_f$  and  $B \in \Sigma_g$ . Then the preimage

$$(f, g)^{-1}(A \times B) = f^{-1}(A) \cap g^{-1}(B) \in \Sigma.$$

By definition, the product  $\sigma$ -field is generated by rectangles, so the proof is complete.  $\square$

## 1.2 Measure space

**Definition** (measure). Let  $(\Omega, \Sigma)$  be a measurable space. Then  $\mu : \Sigma \rightarrow [0, \infty]$  is a measure if

- $\mu(\emptyset) = 0$ .
- If  $A_i \in \Sigma$  is pairwise disjoint then  $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ .

**Proposition** (continuity of measure). If  $A_1 \subset A_2 \subset \dots$  is a nested sequence of elements of  $\Sigma$  and  $\mu$  be any measure on  $(\Omega, \Sigma)$ . Then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \rightarrow \infty} \mu(A_i).$$

If  $A_1 \supset A_2 \supset \dots$  is a nested sequence of elements of  $\Sigma$  and  $\mu(A_n) < \infty$  for some  $n$ . Then

$$\mu\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{i \rightarrow \infty} \mu(A_i).$$

**Definition.** Let  $(\Omega, \Sigma, \mu)$  be a measure space.

Say  $\mu$  is  $\sigma$ -finite if there is a representation  $\Omega = \bigcup_{i=1}^{\infty} \Omega_i$  where  $\Omega_i \in \Sigma$  and  $\mu(\Omega_i) < \infty$ .

Say  $\mu$  is a probability measure if  $\mu(\Omega) = 1$ .

**Definition** (completion of measure space). Let  $(\Omega, \Sigma, \mu)$  be a measure space. Let

$$\tilde{\Sigma} = \{A \cup B : A \in \Sigma, B \subset \Omega, \text{there exists } C \in \Sigma \text{ with } \mu(C) = 0 \text{ and } B \subset C\}.$$

We can check  $\tilde{\Sigma}$  is a  $\sigma$ -field. If  $\tilde{\mu}$  is a measure on  $(\Omega, \tilde{\Sigma})$  which agrees with  $\mu$  on  $\Sigma$ , then  $(\Omega, \tilde{\Sigma}, \tilde{\mu})$  is called a completion of  $(\Omega, \Sigma, \mu)$ .

### 1.3 $\pi$ - $\lambda$ theorem

**Definition** ( $\pi$ -system). Let  $\Omega$  be a set and  $\mathcal{P}$  be a collection of subsets of  $\Omega$ . Then  $\mathcal{P}$  is a  $\pi$ -system if it is closed with respect to taking finite intersections. That is,  $A, B \in \mathcal{P}$  implies  $A \cap B \in \mathcal{P}$ .

**Example.** On the real line  $\mathbb{R}$ , both  $\mathcal{P}_1 = \{(a, b) : a < b\}$  and  $\mathcal{P}_2 = \{(-\infty, a] : a \in \mathbb{R}\}$  are  $\pi$ -systems.

△

**Definition** ( $\lambda$ -system). Let  $\Omega$  be a set and  $\mathcal{L}$  be a collection of subsets of  $\Omega$ . Say  $\mathcal{L}$  is a  $\lambda$ -system if

- $\emptyset \in \mathcal{L}$ .
- $A \in \mathcal{L}$  implies  $A^c \in \mathcal{L}$ .
- for all countable collection of disjoint elements  $A_i \in \mathcal{L}$ , we have  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{L}$ .

For an alternative definition, say  $\mathcal{L}$  is a  $\lambda$ -system if

- $\Omega \in \mathcal{L}$ .
- If  $A, B \in \mathcal{L}$  and  $A \subset B$ , then  $B \setminus A \in \mathcal{L}$ .
- If  $A_n \in \mathcal{L}$  and  $A_n \uparrow A$ , then  $A \in \mathcal{L}$ .

**Theorem** ( $\pi$ - $\lambda$  theorem). Let  $\Omega$  be a set,  $\mathcal{P}$  be a  $\pi$ -system and  $\mathcal{L}$  be a  $\lambda$ -system. Also suppose  $\mathcal{P} \subset \mathcal{L}$ , then  $\sigma(\mathcal{P}) \subset \mathcal{L}$ .

*Proof.* Let  $\ell(\mathcal{P})$  be the smallest  $\lambda$ -system on  $\Omega$  containing  $\mathcal{P}$ . The goal is to show that  $\ell(\mathcal{P})$  is a  $\sigma$ -field. We need to show that if  $A_i \in \ell(\mathcal{P})$  for  $1 \leq i < \infty$ , then  $\bigcup_{i=1}^{\infty} A_i \in \ell(\mathcal{P})$ . Note that

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} \left( A_i \setminus \bigcup_{j=1}^{i-1} A_j \right),$$

so it suffices to show that  $A, B \in \ell(\mathcal{P})$  implies  $A \cap B \in \ell(\mathcal{P})$ .

For  $A \in \ell(\mathcal{P})$  we define

$$W_A = \{B \subset \Omega : A \cap B \in \ell(\mathcal{P})\}.$$

It can be directly verified that  $W_A$  is a  $\lambda$ -system.

Take  $A \in \mathcal{P}$ , then for any  $B \in \mathcal{P}$  we have  $A \cap B \in \mathcal{P} \subset \ell(\mathcal{P})$ . Hence,  $\mathcal{P} \subset W_A$  and thus  $\ell(\mathcal{P}) \subset W_A$  for all  $A \in \mathcal{P}$ , as  $\ell(\mathcal{P})$  is the smallest  $\lambda$ -system on  $\Omega$  containing  $\mathcal{P}$ . Now take  $A \in \ell(\mathcal{P})$ , we have  $A \in W_B$  for all  $B \in \mathcal{P}$ . It follows that  $A \cap B \in \ell(\mathcal{P})$  and thus  $B \in W_A$ . Hence similarly  $\ell(\mathcal{P}) \subset W_A$  for all  $A \in \ell(\mathcal{P})$ .

Now for any pair  $B, C \in \ell(\mathcal{P})$ , we have  $C \in W_B$  and thus  $B \cap C \in \ell(\mathcal{P})$ . This completes the proof. □

## 1.4 Extension theorems

**Definition** (semi-field). A collection of subsets  $S \subset 2^\Omega$  is a semi-field if

- $\emptyset \in S$  and  $\Omega \in S$ .
- $A, B \in S$  implies  $A \cap B \in S$ .
- If  $A \in S$ , then  $A^c$  is a finite disjoint union of sets in  $S$ .

**Theorem** (Caratheodory's extension theorem). Let  $S$  be a semi-field and let  $\mu$  be a non-negative function on  $S$  satisfying:

- $\mu(\emptyset) = 0$ .
- If  $A_1, \dots, A_n$  are disjoint and  $\bigcup_{i=1}^n A_i \in S$ , then  $\mu(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$ .
- If  $A_1, A_2, \dots$  are such that  $\bigcup_{i=1}^\infty A_i \in S$ , then  $\mu(\bigcup_{i=1}^\infty A_i) \leq \sum_{i=1}^\infty \mu(A_i)$ .

Then  $\mu$  admits a unique extension  $\bar{\mu}$  which is a measure on  $\bar{S}$ , the field (algebra) generated by  $S$ . Moreover, if  $\bar{\mu}$  is  $\sigma$ -finite then  $\bar{\mu}$  admits a unique extension  $\tilde{\mu}$  to  $\sigma(S)$ .

**Notation.** Let  $T$  be any set. Write

$$\mathbb{R}^T = \{(\omega_t)_{t \in T} : \omega_t \in \mathbb{R}\}.$$

Also write  $\mathcal{R}^T$  as the  $\sigma$ -field generated by rectangles of the form  $\prod_{t \in T} I_t$ , where for each  $t \in T$ ,  $I_t$  is either a semi-open interval of the form  $(a, b]$  with  $a < b$  or  $I_t = \mathbb{R}$ , and  $I_t = \mathbb{R}$  for all but finitely many  $t \in T$ .

**Theorem** (Kolmogorov's extension theorem). For each finite non-empty subset  $J \subset T$ , let  $\mu_J$  be a Borel probability measure in  $\mathbb{R}^J$ , and assume that the measures  $(\mu_J)_{J \subset T, |J| < \infty}$  are compatible, in the sense that whenever  $J_1 \subset J_2 \subset T$  with  $0 \leq |J_1| \leq |J_2| < \infty$ ,  $I_j \subset \mathbb{R}$  with  $j \in J_1$  are Borel subsets of  $\mathbb{R}$ , and

$$\tilde{I}_j = \begin{cases} I_j & (j \in J_1) \\ \mathbb{R} & (j \in J_2 \setminus J_1), \end{cases}$$

one has

$$\mu_{J_2} \left( \prod_{j \in J_2} \tilde{I}_j \right) = \mu_{J_1} \left( \prod_{j \in J_1} I_j \right).$$

Then there exists a unique probability measure  $\mu$  on  $(\mathbb{R}^T, \mathcal{R}^T)$  consistent with  $(\mu_J)_{J \subset T, |J| < \infty}$ . That is, one has

$$\mu \left( \prod_{t \in T} I_t \right) = \mu_J \left( \prod_{j \in J} I_j \right)$$

whenever  $J \subset T$  with  $|J| < \infty$  and  $I_t = \mathbb{R}$  for all  $t \notin J$ .

## 1.5 Lebesgue Integration

Here we provide a proof for dominated convergence theorem that uses the truncation technique, which will be a useful technique later in the course.

**Theorem** (dominated convergence theorem). Let  $\{f_n\}_{n=1}^\infty$  be a sequence of measurable functions on  $(\Omega, \Sigma, \mu)$  and  $g \geq 0$  be another measurable function. Suppose

1.  $\int g d\mu < \infty$ .
2.  $|f_n|(\omega) \leq g(\omega)$  for all  $\omega \in \Omega$  and  $n \geq 1$ .
3.  $f_n \rightarrow f$  pointwise.

Then

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int f d\mu.$$

*Proof.* **Claim 1.** If  $h$  is a function on  $(\Omega, \Sigma, \mu)$  with  $h \geq 0$  and  $\int h d\mu < \infty$ . Let  $\{A_n\}_{n=1}^\infty$  be any sequence of elements of  $\Sigma$  with  $\mu(A_n) \rightarrow 0$ . Then

$$\int_{A_n} h d\mu \rightarrow 0.$$

Proof of claim. WLOG assume  $\mu(A_n) \leq 2^{-n}$  for all  $n$ . Define  $h_n = h \mathbb{1}_{\bigcup_{i=n}^\infty A_i}$ . We then have

1. The sequence  $\{h_n\}_{n=1}^\infty$  is monotone.
2.  $h_n$  converges to 0 almost everywhere.

Monotone convergence theorem then implies  $\lim_{n \rightarrow \infty} \int h_n d\mu = 0$ . Meanwhile,

$$0 \leq \int_{A_n} h d\mu \leq \int h_n d\mu,$$

and the proof is complete.

**Claim 2.** Suppose  $h \geq 0$  and  $\int h d\mu < \infty$ . Let  $\{\varepsilon_n\}_{n=1}^\infty$  be a sequence of strictly positive numbers converging to zero. Define

$$B_n = \{\omega \in \Omega : h(\omega) \leq \varepsilon_n\} \in \Sigma.$$

Then

$$\int_{B_n} h d\mu \rightarrow 0.$$

Proof of this claim is left as an exercise.

Now we prove the theorem. Fix  $\varepsilon > 0$ . By the previous two claims, there exists  $M > 0$  and  $\delta > 0$  such that

$$\int_{\{g \geq M\}} g d\mu < \varepsilon, \quad \int_{\{g \leq \delta\}} g d\mu < \varepsilon.$$

Let  $U = \{\omega : \delta < g(\omega) < M\}$ . Since  $g$  is integrable,  $\mu(U) < \infty$ . For  $\omega \in U$ , let  $n_\varepsilon(\omega)$  be the smallest index such that  $n \geq n_\varepsilon(\omega)$  implies  $|f_n(\omega) - f(\omega)| \leq \varepsilon \mu(U)^{-1}$ . It follows that there exists  $N$  such that

$$\mu(\{\omega \in U : n_\varepsilon(\omega) > N\}) \leq \frac{\varepsilon}{M}.$$

Then, for  $n \geq N$ , we have

$$\left| \int_U (f_n - f) d\mu \right| \leq \int_{n_\varepsilon(\omega) \leq N} |f_n - f| d\mu + \int_{n_\varepsilon(\omega) > N} |f_n - f| d\mu \leq 3\varepsilon.$$

Now for  $n \geq N$ , we have

$$\left| \int (f_n - f) d\mu \right| \leq 3\varepsilon + \int_{U^c} |f - f_n| d\mu \leq 3\varepsilon + 2 \int_{U^c} g d\mu \leq 7\varepsilon.$$

□

**Theorem** (Markov-Chebyshev inequality). Suppose we have probability measure space  $(\Omega, \Sigma, \mathbb{P})$  and  $f \geq 0$ . Suppose also  $\int f d\mathbb{P} < \infty$ . Then

$$\mathbb{P}(\{\omega : f(\omega) > t\}) \leq \frac{1}{t} \int f d\mathbb{P}.$$

for all  $t > 0$ .

**Remark.** Let  $1 \leq p < \infty$ . Suppose  $f : (\Omega, \Sigma, \mathbb{P}) \rightarrow [0, \infty]$  and  $\int f^p d\mathbb{P} < \infty$ . Then

$$\mathbb{P}(\{\omega : f(\omega) > t\}) \leq \frac{1}{t^p} \int f^p d\mathbb{P}.$$

for all  $t > 0$ .



**Remark.** Suppose  $\int e^{\lambda f} d\mathbb{P} < \infty$  for all  $\lambda \in \mathbb{R}$  and  $f : (\Omega, \Sigma, \mathbb{P}) \rightarrow [0, \infty]$ . Then

$$\mathbb{P}(\{\omega : f(\omega) > t\}) \leq \frac{1}{e^{\lambda t}} \int e^{\lambda f} d\mathbb{P}$$

for all  $t > 0$  and  $\lambda > 0$ .

**Theorem** (Hölder inequality). Let  $p, q \in [1, \infty]$  and  $p^{-1} + q^{-1} = 1$ . Let  $(\Omega, \Sigma, \mu)$  be a probability space. For any measurable functions  $f, g$ , we have

$$\int |fg| d\mathbb{P} \leq \left( \int |f|^p d\mathbb{P} \right)^{1/p} \left( \int |g|^q d\mathbb{P} \right)^{1/q}.$$

**Theorem** (Jensen's inequality). Let  $(\Omega, \Sigma, \mathbb{P})$  be a probability space and  $f$  be integrable. Let  $\varphi : \overline{\mathbb{R}} \rightarrow \overline{\mathbb{R}}$  be convex and suppose  $\varphi(\infty) = \lim_{x \rightarrow \infty} \varphi(x)$  and  $\varphi(-\infty) = \lim_{x \rightarrow -\infty} \varphi(x)$ . Then

$$\varphi \left( \int f d\mathbb{P} \right) \leq \int \varphi(f) d\mathbb{P}.$$

## 1.6 Product measures and Fubini theorem

Let  $(\Omega_1, \Sigma_1, \mu_1), (\Omega_2, \Sigma_2, \mu_2)$  be  $\sigma$ -finite measure spaces. We already defined the product  $\Sigma_1 \otimes \Sigma_2$ . To define a product measure, we first consider the algebra of rectangles

$$S = \{A \in \Sigma_1 \otimes \Sigma_2 : A = A_1 \times A_2 \text{ for some } A_1 \in \Sigma_1, A_2 \in \Sigma_2\}.$$

Then we can define  $\mu = \mu_1 \times \mu_2$  on  $S$  by

$$\mu(A) = \mu_1(A_1)\mu_2(A_2)$$

for  $A = A_1 \times A_2$ . We can check that the definition is self-consistent. That is, if  $A = A_1 \times A_2$  is a countable union of disjoint rectangles  $\{A_1^{(j)} \times A_2^{(j)}\}_{j=1}^{\infty}$ , we have

$$\mu(A_1 \times A_2) = \sum_{j=1}^{\infty} \mu(A_1^{(j)} \times A_2^{(j)}).$$

This can be verified with monotone convergence theorem. Now  $\mu$  is a premeasure and can be uniquely extended to  $\Sigma_1 \otimes \Sigma_2$ .

**Theorem** (Fubini-Tonelli). Let  $(\Omega_1, \Sigma_1, \mu_1), (\Omega_2, \Sigma_2, \mu_2)$  be  $\sigma$ -finite measure spaces and let  $(\Omega, \Sigma, \mu)$  be the product space. Suppose  $f$  is measurable on the product space. Suppose either  $f$  is non-negative or  $\int_{\Omega} |f| d\mu < \infty$ . Then

$$- \quad y \mapsto f(x, y) \text{ is } \Sigma_2 \text{ measurable for all } x \in \Omega_1.$$

–  $x \mapsto \int_{\Omega_2} f(x, y) d\mu_2(y)$  is  $\Sigma_1$  measurable.

– We have

$$\int_{\Omega_1} \int_{\Omega_2} f(x, y) d\mu_2(y) d\mu_1(x) = \int_{\Omega} f(x, y) d\mu(x, y).$$

*Proof.* First suppose  $f = \mathbb{1}_A$  for  $A \in \Sigma$ . Also suppose  $\mu_1, \mu_2$  are finite. Define section

$$A_x = \{y \in \Omega_2 : (x, y) \in A\}.$$

The goal is to show that  $A_x \in \Sigma_2$  for all  $x \in \Omega_1$ . Define a family of sets

$$\mathcal{F}_x = \{B \in \Sigma : B_x \text{ is } \Sigma_2\text{-measurable}\}.$$

It can be verified that  $\mathcal{F}_x$  is a  $\sigma$ -field for all  $x \in \Omega_1$ . Also,  $\mathcal{F}_x$  contains all rectangles and thus  $\Sigma \subset \mathcal{F}_x$ . Hence, we have shown that  $y \mapsto \mathbb{1}_A(x, y) = \mathbb{1}_{A_x}(y)$  is measurable for all  $x \in \Omega_1$ .

Next we show  $x \mapsto \mu_2(A_x)$  is measurable and its integral over  $\Omega_1$  is equal to  $\mu(A)$ . Define

$$\mathcal{U} = \left\{ B \in \Sigma : x \mapsto \mu_2(B_x) \text{ is } \Sigma_1\text{-measurable and } \int_{\Omega_1} \mu_2(B_x) d\mu_1 = \mu(B) \right\}$$

It can be verified that  $\mathcal{U}$  is a  $\lambda$ -system. Note that  $\mathcal{U}$  also contains all rectangles in  $\Sigma$ . It follows that  $\mathcal{U} = \Sigma$  and the proof for indicator functions are complete.

Then use linearity to extend to simple functions, and use monotone convergence theorem to prove the statement for non-negative functions. For the case where  $f$  is integrable, consider the positive and negative part about  $f$  to complete the proof.

□

## 2 Probability theory basics

### 2.1 Distributions and densities

**Definition.** Let  $F : \mathbb{R} \rightarrow [0, 1]$ . Suppose  $F$  is

- right-continuous.
- non-decreasing.
- $\lim_{t \rightarrow -\infty} F(t) = 0$  and  $\lim_{t \rightarrow \infty} F(t) = 1$ .

Then  $F$  is a cumulative distribution function (CDF).

**Remark.** If we want to define CDF in  $\mathbb{R}^2$  then the axioms are

- right-continuous:  $F(\tilde{s}, \tilde{t}) \rightarrow F(s, t)$  as  $t \downarrow \tilde{t}$  and  $s \downarrow \tilde{s}$ .
- coordinate-wise non-decreasing.
- $\lim_{s, t \rightarrow \infty} F(s, t) = 1$ ,  $\lim_{s \rightarrow -\infty} F(s, t) = 0$  for any  $t$ , and  $\lim_{t \rightarrow -\infty} F(s, t) = 0$  for any  $s$ .
- For a rectangle with bottom left vertex  $(a_1, a_2)$  and top right vertex  $(b_1, b_2)$ ,

$$F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2) \geq 0.$$

Now we can connect the notion of CDF with randomness.

Suppose  $X$ -random real-valued variable on  $(\Omega, \Sigma, \mathbb{P})$  that is almost everywhere finite. Define

$$F_X(t) = \mathbb{P}(X(\omega) \leq t)$$

for  $-\infty < t < \infty$ . It can be verified that  $F_X$  is a CDF.

Conversely, for any CDF  $F$ , there exists a probability space  $(\Omega, \Sigma, \mathbb{P})$  and a real valued random variable on  $(\Omega, \Sigma, \mathbb{P})$  with CDF  $F$ .

**Definition.** If  $X$  is random variable on  $(\Omega, \Sigma, \mathbb{P})$  real valued and a.e. finite. Then we can define the induced Borel probability measure  $\mu_X$  on  $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$  by

$$\mu_X(B) := \mathbb{P}(X \in B)$$

for all  $B \in \mathcal{B}_{\mathbb{R}}$ .

Now suppose  $\mu$  is any Borel probability measure on  $\mathbb{R}$ . Consider probability space  $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mu)$  and formal identity mapping  $\text{id}$  on  $\mathbb{R}$ . Then  $\mu_{\text{id}} \equiv \mu$ .

**Theorem.** There is a one-to-one correspondence between the family of CDFs and the family of Borel probability measure on  $\mathbb{R}$ .

*Proof.* For any Borel probability measure  $\mu$ ,  $F_\mu(t) = \mu((-\infty, t])$  is a valid CDF.

Conversely, for any CDF  $F$ , there exists unique probability measure  $\mu_F$  on  $\mathbb{R}$  such that  $\mu_F((-\infty, t]) = F(t)$  for all  $-\infty < t < \infty$ . This is a corollary of Caratheorod extension theorem. For detailed proof see notes or textbook. □

**Remark.** Suppose  $X = (X_1, X_2)$  is a random vector in  $\mathbb{R}^2$ . We can define

$$F_X(s, t) = \mathbb{P}(X_1 \leq s, X_2 \leq t).$$

Corresponding results are also true.

**Definition** (Probability mass function). Let  $(\Omega, \Sigma, \mathbb{P})$  be a probability space and  $X : \Omega \rightarrow \mathbb{R}$  be random variable. Suppose there exists  $S \subset \mathbb{R}$  countable such that  $\mathbb{P}(X \in S) = 1$ . We can define the probability mass function (PMF)  $f_X$  via

$$f_X(t) = \mathbb{P}(X = t)$$

for  $t \in \mathbb{R}$ . Due to the restriction, this gives complete description of the distribution, and we can construct CDF  $F_X$  via

$$F_X(t) = \sum_{s \leq t} f_X(s).$$

This sum makes sense since the  $f_X(s) = 0$  for all but countably many  $s$ . Conversely, we can also reconstruct  $f_X$  from a CDF  $F_X$ .

**Definition** (Probability density function). Suppose  $F$  is a CDF which is absolutely continuous. That is, there exists Borel measurable non-negative function  $\rho$  on  $\mathbb{R}$  such that

$$F(t) = \int_{-\infty}^t \rho(s) ds$$

for all  $-\infty < t < \infty$ . This implies  $F$  is almost everywhere differentiable and the derivative is  $\rho$ . In this case, say  $\rho$  is the density function.

If random variable  $X$  is such that  $F_X$  is absolutely continuous, then the corresponding  $\rho_X$  is the probability density function for  $X$ .

**Remark.** Recall that a Borel  $\sigma$ -finite measure  $\mu$  on the real line is absolutely continuous w.r.t the Lebesgue measure  $m$  on  $\mathbb{R}$  if  $\mu(A) = 0$  whenever  $A \in \mathcal{B}_{\mathbb{R}}$  is Lebesgue null. In this case, Randon-Nikodym theorem implies existence of non-negative Borel measurable function  $f$  such that  $\mu(A) = \int_A f dm$ .

**Theorem.** Suppose  $X$  RV on  $(\Omega, \Sigma, \mathbb{P})$  is real-valued and a.e. finite. The following are equivalent:

1.  $F_X$  is absolutely continuous.
2.  $\mu_X$  is absolutely continuous w.r.t. Lebesgue measure.

Moreover,  $\rho_X$  is also the derivative of  $\mu_X$  w.r.t. Lebesgue measure. That is, for any  $A \in \mathcal{B}_{\mathbb{R}}$ ,

$$\mu_X(A) = \int_A \rho_X(t) dt.$$

## 2.2 Independence

**Definition.** Say two events  $A, B \in \Sigma$  are independent if  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ .

It is easy to verify that  $A, B$  are independent implies  $A^c, B$  are independent.

**Remark.** Suppose  $\mathbb{P}(B) > 0$ , then the conditional probability of  $A$  given  $B$  is defined as

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

Then, independence of  $A$  and  $B$  is equivalent to  $\mathbb{P}(A | B) = \mathbb{P}(A)$ .

**Definition.** Let  $A_1, \dots, A_n$  be events. Say they are mutually independent if for any  $\emptyset \neq I \subset [n]$ , we have

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

This is equivalent to saying that for every  $2 \leq i \leq n$ , the event  $A_i$  is independent from any event generated by  $A_1, \dots, A_{i-1}$ , or  $A_i$  is independent from  $\sigma(A_1, \dots, A_{i-1})$ .

**Remark.** The events  $A_1, \dots, A_n$  are called  $k$ -wise independent if any  $k$ -subset of the events are mutually independent. For  $k < n$ , this notion is strictly weaker than mutual independence of all  $n$  events. As an example, consider  $\mathbb{P}$  to be the uniform distribution on  $\{1, \dots, 4\}$ . Let  $A_1 = \{1, 2\}$ ,  $A_2 = \{1, 3\}$ , and  $A_3 = \{2, 3\}$ . Then they are pairwise independent but not mutually independent.

**Definition.** A collection of events  $\{A_\lambda\}_{\lambda \in \Lambda}$  on  $(\Omega, \Sigma, \mathbb{P})$  are mutually independent if any finite subset of events are mutually independent.

**Definition.** Let  $(\Omega, \Sigma, \mathbb{P})$  be a probability space. Two  $\sigma$ -subfields are independent if for any  $A \in \Sigma_1$  and  $B \in \Sigma_2$ ,  $A, B$  are independent.

**Definition.** Let  $(\Omega, \Sigma, \mathbb{P})$  be a probability space and  $X, Y$  be two real-valued random variables. Say  $X$  and  $Y$  are independent if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \mathbb{P}(Y \in B)$$

for any  $A, B \in \mathcal{B}_{\mathbb{R}}$ .

Equivalently, let  $\Sigma_X, \Sigma_Y$  be the  $\sigma$ -field generated by  $X$  and  $Y$ . Then independence of  $X$  and  $Y$  is equivalent to independence of  $\Sigma_X$  and  $\Sigma_Y$ .

Now we explore how this connect with product structure.

**Proposition.** Let  $(\Omega_1, \Sigma_1, \mathbb{P}_1)$  and  $(\Omega_2, \Sigma_2, \mathbb{P}_2)$  be two probability spaces and let  $(\Omega, \Sigma, \mathbb{P})$  be the product space. Let  $X$  and  $Y$  be two random variables on  $(\Omega, \Sigma, \mathbb{P})$ . Suppose there exists some measurable functions such that  $X(\omega_1, \omega_2) = g(\omega_1)$ , and  $Y(\omega_1, \omega_2) = h(\omega_2)$ . Then  $X$  and  $Y$  are independent.

*Proof.* Let  $A, B \in \mathcal{B}_{\mathbb{R}}$ . Then

$$\begin{aligned} \mathbb{P}(X \in A, Y \in B) &= \mathbb{P}((\omega_1, \omega_2) : \omega_1 \in g^{-1}(A), \omega_2 \in h^{-1}(B)) \\ &= \mathbb{P}(\{\omega_1 \in g^{-1}(A)\} \times \{\omega_2 \in h^{-1}(B)\}) \\ &= \mathbb{P}_1(\omega_1 \in g^{-1}(A)) \mathbb{P}_2(\omega_2 \in h^{-1}(B)). \end{aligned}$$

However,

$$\begin{aligned} \mathbb{P}_1(\omega_1 \in g^{-1}(A)) &= \mathbb{P}((\omega_1, \omega_2) : \omega_1 \in g^{-1}(A), \omega_2 \in \Omega_2) \\ &= \mathbb{P}(X \in A), \end{aligned}$$

and similarly for  $Y$ .

□

**Remark.** Let  $(\Omega, \Sigma, \mathbb{P})$  be a probability space and suppose  $X, Y$  be two random variables that are independent and a.e. finite. They then generate two Borel probability measure  $\mu_X$  and  $\mu_Y$  on  $\mathbb{R}$ . Define a product probability space as of  $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}, \mu_X \times \mu_Y)$ . Define  $\tilde{X}(x, y) = x$  and  $\tilde{Y}(x, y) = y$  as random variables on the product space. By definition,  $\tilde{X}$  is equidistributed with  $X$ . That is,  $\mu_{\tilde{X}} = \mu_X$  and  $F_{\tilde{X}} = F_X$ . Similarly  $\mu_{\tilde{Y}} = \mu_Y$ . Also,  $\tilde{X}, \tilde{Y}$  are independent. Now  $(X, Y)$  and  $(\tilde{X}, \tilde{Y})$  have the same distribution.

**Remark.** If  $X$  and  $Y$  are independent, then their joint distribution  $F_{(X,Y)}$  is uniquely determined by the individual distributions of  $F_X, F_Y$ . Indeed,

$$F_{(X,Y)}(s, t) = F_X(s)F_Y(t).$$

**Remark.** Let  $(\Omega, \Sigma, \mathbb{P})$  be a probability space and suppose  $X, Y$  be two random variables that are independent. Suppose they have densities  $\rho_X, \rho_Y$ , then the distribution density of vector  $(X, Y)$  is  $\rho_{(X,Y)}(s, t) = \rho_X(s)\rho_Y(t)$ .

**Remark.** If  $X$  and  $Y$  are independent random variable, and  $f, g$  are measurable functions. Then  $f(X)$  and  $g(Y)$  are independent as well.

**Remark.** Given probability space  $(\Omega, \Sigma, \mathbb{P})$  and random variable  $X$ . It may not exists another random variable  $Y$  that is independent from  $X$  on the same probability space. See the following example.

**Example.** As an example, consider  $([0, 1], \mathcal{B}_{[0,1]}, m)$  and  $X(\omega) = \omega$ , so  $X$  is uniform on  $[0, 1]$ . The goal is to construct variable  $Y$  such that  $Y \sim \text{Bernoulli}(\frac{1}{2})$ .

△

*Proof.* Let  $A \in \mathcal{B}_{[0,1]}$  and  $t \in [0, 1]$ . Define the density of set  $A$  at point  $t$  to be

$$\lim_{\varepsilon \rightarrow 0} \frac{m(A \cap [t - \varepsilon, t + \varepsilon])}{2\varepsilon}.$$

The Lebesgue density theorem says this is well defined and takes values in  $\{0, 1\}$  for  $m$ -a.e.  $t \in [0, 1]$ .

Suppose such  $Y$  exists, we can vies  $Y$  as an indicator of a set  $A \subset [0, 1]$  of probability  $\frac{1}{2}$ . That is,  $Y = \mathbb{1}_A$  and  $\mathbb{P}(A) = \frac{1}{2}$ . Choose any point  $t \in (0, 1)$  such that density of  $A$  at  $t$  is well-defined. WLOG assume the density is 1. Pick  $\varepsilon > 0$  such that  $m(A \cap [t - \varepsilon, t + \varepsilon]) \geq \frac{3}{2}\varepsilon$ . Now,

$$\begin{aligned} m(A \cap [t - \varepsilon, t + \varepsilon]) &= \mathbb{P}(Y = 1, X \in [t - \varepsilon, t + \varepsilon]) \\ &= \mathbb{P}(Y = 1) \mathbb{P}(X \in [t - \varepsilon, t + \varepsilon]) \\ &= \varepsilon, \end{aligned}$$

a contradiction.

Alternatively, we can derive a contradiction using  $m(A) = \mathbb{P}(Y = 1, X \in A)$ .

□

**Remark.** The goal is to have statements “independent” from the underlying probability space.

## 2.3 Convolution

**Definition** (convolution). Let  $\mu, \nu$  be Borel probability measures on  $\mathbb{R}$ . The convolution of  $\mu$  and  $\nu$  is a probability measure on  $\mathbb{R}$  such that

$$(\mu * \nu)(S) = \int_{\mathbb{R}^2} \mathbb{1}_S(x + y) d(\mu \times \nu)$$

for all  $S \in \mathcal{B}_{\mathbb{R}}$ .

**Remark.** Suppose both  $\mu$  and  $\nu$  have densities  $\rho_\mu$  and  $\rho_\nu$ . That is,  $\mu \ll m$  and  $\nu \ll m$ . It follows that

$$\begin{aligned} (\mu * \nu)(S) &= \int_{\mathbb{R}} \left( \int_{\mathbb{R}} \mathbb{1}_S(x+y) \rho_\mu(x) dx \right) \rho_\nu(y) dy \\ &= \int_{\mathbb{R}} \left( \int_S \rho_\mu(w-y) dw \right) \rho_\nu(y) dy \\ &= \int_S \left( \int_{\mathbb{R}} \rho_\mu(w-y) \rho_\nu(y) dy \right) dw \end{aligned}$$

Note that this implies  $\mu * \nu \ll m$  and

$$\rho_{\mu * \nu}(w) = \int_{\mathbb{R}} \rho_\mu(w-y) \rho_\nu(y) dy,$$

which is the convolution of function  $\rho_\mu$  and  $\rho_\nu$ .

**Remark.** If  $X$  and  $Y$  are two independent random variables and  $\mu_X$  and  $\mu_Y$  are the corresponding induced Borel measure  $\mathbb{R}$ , then  $\mu_{X+Y} = \mu_X * \mu_Y$ .

**Remark.** Suppose  $X$  and  $Y$  are independent and their CDF is  $F_X$  and  $F_Y$ , then

$$F_{X+Y}(t) = \int_{\mathbb{R}} F_X(t-w) d\mu_Y(w) = \int_{\mathbb{R}} F_X(t-w) dF_Y(w).$$

## 2.4 Moments

In this section we explore the computation and basic properties of moments.

**Definition** (moments). Let  $X$  be a random variable on  $(\Omega, \Sigma, \mathbb{P})$ . The  $p$ -th absolute moment of  $X$  is

$$\mathbb{E} |X|^p = \int_{\Omega} |X|^p d\mathbb{P}.$$

This is always well-defined but can be infinite.

If  $p$  is positive integer, the  $p$ -th moment of  $X$  is

$$\mathbb{E} X^p = \int_{\Omega} X^p d\mathbb{P}$$

whenever it is defined.

In particular,  $\mathbb{E} X$  is the mean or expectation, the variance is  $\text{var}(X) = \mathbb{E}(X - \mathbb{E} X)^2$  whenever the expectation is defined, and the standard deviation is  $\sqrt{\text{var}(X)}$ .

**Proposition.** Let  $X$  be random variable  $(\Omega, \Sigma, \mathbb{P})$  and  $0 < p < q < \infty$ . Then,

$$(\mathbb{E} |X|^p)^{1/p} \leq (\mathbb{E} |X|^q)^{1/q}.$$



*Proof.* Define  $Y = |X|^p$ , then we want to show that  $\mathbb{E}Y \leq (\mathbb{E}Y^{q/p})^{p/q}$ . Note that  $t \mapsto |t|^{q/p}$  is convex. Therefore, by Jensen's inequality,

$$(\mathbb{E}Y)^{q/p} \leq \mathbb{E}(Y^{q/p}).$$

□

Now we want to show that moments only depends on the distribution of the random variable, and it does not carry unnecessary information about the underlying probability space. To do this, we first show the following proposition.

**Proposition.** Let  $g$  be measurable and  $X$  a random variable. Then,

$$\mathbb{E}|g(X)| = \int_{\mathbb{R}} |g(t)| d\mu_X(t).$$

Moreover, if  $\mathbb{E}|g(X)| < \infty$ , then

$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(t) d\mu_X(t).$$

**Corollary.** If  $\mathbb{E}X$  is well-defined, then

$$\mathbb{E}X = \int_{\mathbb{R}} t d\mu_X(t).$$

Moreover, if  $X$  has distribution density  $\rho_X$ , then  $\mathbb{E}X = \int_{\mathbb{R}} t \rho_X(t) dt$ .

**Proposition.** If  $X \geq 0$ , then

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geq t) dt.$$

In particular, if  $X$  is non-negative and integer valued, then  $\mathbb{E}X = \sum_{i=1}^\infty \mathbb{P}(X \geq i)$ .

*Proof.* With Fubini-Tonelli, we have

$$\mathbb{E}X = \int_0^\infty s d\mu_X(s) = \int_0^\infty \int_0^s 1 dt d\mu_X(s) = \int_0^\infty \int_t^\infty 1 d\mu_X dt = \int_0^\infty \mathbb{P}(X \geq t) dt.$$

□

**Proposition.** Let  $X_1, \dots, X_n$  are random variables on  $(\Omega, \Sigma, \mathbb{P})$ . Suppose they are either all non-negative or all integrable. Suppose also that they are mutually independent. Then,

$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbb{E}X_i.$$

*Proof.* It suffices to show the statement for  $n = 2$ . Define independent variables  $\widetilde{X}_1$  and  $\widetilde{X}_2$  on the product space  $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}, \mu_{X_1} \times \mu_{X_2})$  by coordinate projection. We know  $\widetilde{X}_i$  is equidistributed with  $X_i$ , so

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[\widetilde{X}_1 \widetilde{X}_2] = \mathbb{E}[\widetilde{X}_1 \mathbb{E}\widetilde{X}_2] = \mathbb{E}[X_1 \mathbb{E}X_2],$$

where in the second equality we used Fubini-Tonelli. □

**Remark.** Recall that whenever  $\mathbb{E}X$  is finite,  $\text{var } X = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$ . If  $X_1, \dots, X_n$  are pairwise independent and have well-defined variances, then

$$\text{var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{var } X_i.$$

**Definition** (moment generating function). Let  $X$  be a random variable, and  $\lambda \in \mathbb{R}$ . Define the moment generating function of  $X$  via

$$M_X(\lambda) = \mathbb{E} \exp(\lambda X).$$

Suppose  $M_X(\lambda)$  is finite in some neighborhood of 0. Then,

$$M_X(\lambda) = \mathbb{E} \left[ \sum_{n=1}^{\infty} \frac{(\lambda x)^n}{n!} \right] = \sum_{n=1}^{\infty} \frac{\lambda^n \mathbb{E}X^n}{n!}.$$

It follows that  $M'_X(0) = \mathbb{E}X$ . Now to justify the exchange of integral and summation, note that

$$S_N = \sum_{i=1}^N \frac{(\lambda x)^n}{n!} \rightarrow e^{\lambda x},$$

and

$$|S_N| \leq \sum_{i=1}^N \frac{|\lambda x|^n}{n!} \leq e^{|\lambda x|}.$$

However,  $e^{|\lambda x|} \leq e^{\lambda x} + e^{-\lambda x}$ . The expectation of RHS is finite for small  $\lambda$  by assumption, so the claim follows from dominated convergence theorem.

**Proposition.** Let  $X$  be a non-negative random variable. Then TFAE:

1.  $M_X$  is finite in some neighborhood of 0.
2.  $\limsup_{p \rightarrow \infty} p^{-1}(\mathbb{E}X^p)^{1/p} < \infty$ .

*Proof.* (1)  $\implies$  (2). Suppose  $M_X(\varepsilon) = \mathbb{E} \left[ \sum_{n=1}^{\infty} \frac{\varepsilon^n X^n}{n!} \right] < \infty$ . This implies that  $\sup_{n \geq 1} \mathbb{E} \left[ \frac{\varepsilon^n X^n}{n!} \right] < \infty$ . Let  $1 \leq C < \infty$  be such that  $\mathbb{E} \left[ \frac{\varepsilon^n X^n}{n!} \right] < C$  for all  $n \geq 1$ . We then have

$$\left( \frac{\varepsilon}{n!} \right)^{1/n} (\mathbb{E} X^n)^{1/n} \leq C^{1/n}$$

It follows from Sterling's formula  $n! \sim (n/e)^n$  that

$$\frac{1}{n} (\mathbb{E} X^n)^{1/n} \leq C$$

for some other constant  $C$ .

(2)  $\implies$  (1). Suppose  $\limsup_{p \rightarrow \infty} p^{-1} (\mathbb{E} X^p)^{1/p} \leq C < \infty$ . It follows that for any  $n \geq 1$

$$\frac{1}{n!} (\mathbb{E} X^n) \leq C^n \frac{n^n}{n!} \leq C^n$$

for some other constant  $C$ . Take  $\varepsilon = \frac{1}{2C}$ , we then have

$$\mathbb{E} \left[ \frac{\varepsilon^n X^n}{n!} \right] \leq 2^{-n}.$$

Therefore,  $M_X(\varepsilon) = \mathbb{E} \left[ \sum_{n=1}^{\infty} \frac{\varepsilon^n X^n}{n!} \right] < \infty$ .

□

Now we present an example of moment method for approximating random variables.

**Example.** Let  $G \sim G(n, \frac{1}{2})$  be the Erdos-Renyi random graph. The goal is to estimate the number of triangles in  $G$ . Let  $N$  be the number of triangles in  $G$ . We have

$$N = \sum_{\substack{S \subset [n] \\ |S|=3}} b_S,$$

where  $b_S$  is the indicator that  $S$  is a triangle in  $G$ . It follows that  $\mathbb{E} N = \frac{1}{8} \binom{n}{3}$ . Also,

$$\mathbb{E} N^2 = \sum_{|S|=3} \sum_{|S'|=3} \mathbb{E} [b_S b_{S'}]$$

To compute this, we consider several cases.

1. If  $S \cap S' = \emptyset$ , then  $\mathbb{E} [b_S b_{S'}] = \frac{1}{64}$ .
2. If  $|S \cap S'| = 1$ , then  $\mathbb{E} [b_S b_{S'}] = \frac{1}{64}$ .
3. If  $|S \cap S'| = 3$ , then  $\mathbb{E} [b_S b_{S'}] = \frac{1}{8}$ .

4. If  $|S \cap S'| = 2$ , then  $\mathbb{E}[b_S b_{S'}] = \frac{1}{32}$ .

Hence,

$$\mathbb{E}N^2 = \frac{1}{64} \binom{n}{3}^2 \pm O(n^4),$$

where the  $O(n^4)$  term comes from the cases where  $|S \cap S'| = 2$  or  $3$ . Therefore,

$$\text{var } N = \mathbb{E}N^2 - (\mathbb{E}N)^2 = O(n^4).$$

Using Chebyshev's inequality, for each  $t > 0$  we have

$$\mathbb{P}(|N - \mathbb{E}N| \geq t \cdot \mathbb{E}N) \leq \frac{\text{var } N}{t^2 (\mathbb{E}N)^2} = t^{-2} O(n^{-2}) \rightarrow 0.$$

as  $n \rightarrow \infty$ .

△

We have the following proposition on sub-Gaussian decay of moments.

**Proposition.** Let  $X$  be a random variable, TFAE:

1.  $\sup_{p \geq 1} (\mathbb{E}|X|^p)^{1/p} p^{-1/2} < \infty$ .
2. there exists  $c > 0$  such that  $\mathbb{P}(|X| > t) \leq 2 \exp(-ct^2)$  for all  $t > 0$ .

If any of these statement is satisfied,  $X$  is called a *subgaussian variable*.

*Proof.* Exercise.

□

**Theorem** (Khintchine's inequality). There exists a universal constant  $C < \infty$  such that for any  $n \in \mathbb{N}$ ,  $a_1, \dots, a_n \in \mathbb{R}$ , and  $p \geq 1$ , we have

$$\left( \mathbb{E} \left| \sum_{i=1}^n a_i r_i \right|^p \right)^{1/p} \leq C \sqrt{p} \|a\|_2,$$

where  $r_1, \dots, r_n$  are i.i.d. Rademacher variables:  $\mathbb{P}(r_i = 1) = \mathbb{P}(r_i = -1) = \frac{1}{2}$ .

*Proof.* WLOG assume  $\|a\|_2 = 1$ . Let  $\lambda > 0$ . For any  $t > 0$ , we have

$$\mathbb{P} \left( \sum_{i=1}^n a_i r_i > t \right) = \mathbb{P} \left( \exp \left( \lambda \sum_{i=1}^n a_i r_i \right) > \exp(\lambda t) \right) \leq \frac{\mathbb{E} \exp \left( \lambda \sum_{i=1}^n a_i r_i \right)}{\exp(\lambda t)}$$

from Markov's inequality. By independence, we have

$$\mathbb{P}\left(\sum_{i=1}^n a_i r_i > t\right) \leq \frac{\prod_{i=1}^n \mathbb{E} \exp(\lambda a_i r_i)}{\exp(\lambda t)} = \frac{\prod_{i=1}^n \frac{1}{2} (\exp(\lambda a_i) + \exp(-\lambda a_i))}{\exp(\lambda t)}.$$

Now we use  $\exp(x) \leq x + \exp(x^2)$  to obtain

$$\mathbb{P}\left(\sum_{i=1}^n a_i r_i > t\right) \leq \frac{\prod_{i=1}^n \exp(\lambda^2 a_i^2)}{\exp(\lambda t)} = \exp\left(\lambda^2 \|a\|_2^2 - \lambda t\right).$$

This holds for any  $\lambda > 0$ . Substituting  $\lambda = \frac{1}{2} t \|a\|_2^{-2}$ , we have

$$\mathbb{P}\left(\sum_{i=1}^n a_i r_i > t\right) \leq \exp\left(-\frac{1}{4} \frac{t^2}{\|a\|_2^2}\right)$$

Similarly we can bound  $\mathbb{P}(\sum_{i=1}^n a_i r_i < -t)$  as the Rademacher variables are symmetric. Hence we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i r_i\right| > t\right) \leq 2 \exp\left(-\frac{1}{4} t^2\right).$$

The theorem follows from the previous proposition. □

**Theorem** (Anti-concentration inequalities). Let  $X$  be a non-negative random variable and  $1 \leq p < q < \infty$ . Suppose  $(\mathbb{E}X^p)^{1/p} \leq C (\mathbb{E}X^q)^{1/q}$  for some constant  $C$ . Then there exists  $\varepsilon > 0$  depending only on  $p, q, C$  such that

$$\mathbb{P}\left(X \geq \varepsilon (\mathbb{E}X^p)^{1/p}\right) \geq \varepsilon.$$

## 2.5 Convergence of random variable

**Definition.** Let  $\{X_n\}_{n=1}^\infty$  be a sequence of random variables and  $X$  is also a random variable. Say  $X_n$  converges to  $X$  *in distribution* or *weakly* if for any compactly supported continuous function  $f$ , we have

$$\int_{\mathbb{R}} f(t) d\mu_{X_n}(t) \rightarrow \int_{\mathbb{R}} f(t) d\mu_X(t).$$

Write  $X_n \Rightarrow X$  or  $X_n \xrightarrow{w} X$ .

**Proposition.**  $X_n \xrightarrow{w} X$  iff  $F_{X_n}(t) \rightarrow F_X(t)$  at every point  $t$  where  $F_X$  is continuous.

**Definition.** Let  $\{X_n\}_{n=1}^\infty$  be a sequence of random variables and  $X$  is also a random variable defined on the same probability space. Say  $X_n$  converges to  $X$  *in probability* if for any  $\varepsilon > 0$ ,

$$\mathbb{P}(|X_n - X| < \varepsilon) \rightarrow 1.$$

Write  $X_n \xrightarrow{p} X$ .

**Proposition.** If  $X_n \xrightarrow{p} X$ , then  $X_n \Rightarrow X$ .

*Proof.* Let  $t \in \mathbb{R}$  be such that  $F_X$  is continuous at point  $t$  and let  $\varepsilon > 0$ . Let  $\delta > 0$  be such that  $F_X(t + \delta) \leq F_X(t) + \varepsilon$  and  $F_X(t - \delta) \geq F_X(t) - \varepsilon$ . Since  $\mathbb{P}(|X - X_n| < \delta) \rightarrow 1$ , there exists  $n_0$  such that  $n \geq n_0$  implies

$$\mathbb{P}(|X_n - X| < \delta) \geq 1 - \varepsilon.$$

Take  $n \geq n_0$ . We have

$$\begin{aligned} F_{X_n}(t) &= \mathbb{P}(X_n \leq t) \\ &\leq \mathbb{P}(|X_n - X| > \delta) + \mathbb{P}(X \leq t + \delta) \\ &\leq \varepsilon + F_X(t) + \varepsilon \\ &\leq F_X(t) + 2\varepsilon. \end{aligned}$$

Similarly for the other side.

□

**Proposition.** If  $X_n \rightarrow X$  a.s., then  $X_n \xrightarrow{p} X$ .

*Proof.* Let  $\varepsilon > 0$ . For each  $\omega \in \Omega$ , define

$$n_\varepsilon(\omega) = \min \{n : |X_m(\omega) - X(\omega)| \leq \varepsilon \text{ for all } m \geq n\} < \infty.$$

By continuity of measure, for any  $\delta > 0$  there exists  $N$  such that  $\mathbb{P}(n_\varepsilon \leq N) \geq 1 - \delta$ . This implies that for all  $m \geq N$ , we have

$$\mathbb{P}(|X_m - X| \leq \varepsilon) \geq 1 - \delta.$$

Since  $\delta > 0$  is arbitrary, this completes the proof.

□

**Proposition.** Let  $c \in \mathbb{R}$  and suppose  $\{X_n\}_{n=1}^\infty$  are random variables on a common probability space  $(\Omega, \Sigma, \mathbb{P})$ . Then  $X_n \Rightarrow c$  iff  $X_n \xrightarrow{p} c$ .

*Proof.* Just need to check that  $X_n \Rightarrow c$  implies  $X_n \xrightarrow{p} c$ . We have

$$F_c(t) = \begin{cases} 1 & \text{if } t \geq c, \\ 0 & \text{otherwise.} \end{cases}$$

Take  $\varepsilon > 0$ , then  $c \pm \varepsilon$  are points of continuity. It follows that  $F_{X_n}(c + \varepsilon) \rightarrow 1$  and  $F_{X_n}(c - \varepsilon) \rightarrow 0$ . This implies that

$$\mathbb{P}(|X_n - c| > \varepsilon) \rightarrow 0.$$

□

Next we consider algebraic operations and convergence. It is clear that if  $X_n \rightarrow X$  and  $Y_n \rightarrow Y$  a.e., then  $X_n + Y_n \rightarrow X + Y$  a.e. Also, if  $X_n \xrightarrow{p} X$  and  $Y_n \xrightarrow{p} Y$ , then  $X_n + Y_n \xrightarrow{p} X + Y$ . But for convergence in distribution, the same statement need not be true without any extra assumptions. However, suppose  $X_n$  and  $Y_n$  are independent for each  $n$ ,  $X$  and  $Y$  are independent,  $X_n \Rightarrow X$ , and  $Y_n \Rightarrow Y$ . Then  $X_n + Y_n \Rightarrow X + Y$ .

**Example.** Suppose  $\{a_i\}_{i=1}^\infty$  are i.i.d. random variables and

$$\varepsilon_n \sum_{i=1}^n a_i \Rightarrow X$$

for some random variable  $X$  and sequence of positive numbers  $\{\varepsilon_n\}_{n=1}^\infty$ . Then we have

$$\varepsilon_{2n} \sum_{i=1}^{2n} a_i \Rightarrow X,$$

and

$$\frac{\varepsilon_{2n}}{\varepsilon_n} \left( \varepsilon_n \sum_{i=1}^n a_i \right) + \frac{\varepsilon_{2n}}{\varepsilon_n} \left( \varepsilon_n \sum_{i=n+1}^{2n} a_i \right) \Rightarrow X.$$

If we further suppose that there is  $\alpha > 0$  such that

$$\lim_{n \rightarrow \infty} \frac{\varepsilon_{2n}}{\varepsilon_n} = \alpha,$$

we would have  $\alpha X + \alpha \tilde{X} \sim X$ , where  $\tilde{X}$  is an independent copy of  $X$ . The only such distribution  $X$  with bounded second moment is the Gaussian distribution.

△

Recall that for independent random variables, we have a representation theorem that create new independent random variables on the product space. We also have a representation theorem for random variables converging in distribution. On the new probability space, they will converge a.e.

**Theorem** (Skorokhod's representation theorem). Suppose  $X_n \Rightarrow X$  on some probability space  $(\Omega, \Sigma, \mathbb{P})$ . Then there exists another probability space  $(\tilde{\Omega}, \tilde{\Sigma}, \tilde{\mathbb{P}})$  and random variables  $\tilde{X}_n, \tilde{X}$  on it such that the following holds:

- $\tilde{X}_n \sim X_n$  for all  $n$ .
- $\tilde{X} \sim X$ .
- $\tilde{X}_n \rightarrow \tilde{X}$  a.e.

*Proof.* Suppose for simplicity that  $F_{X_n}$  and  $F_X$  are all continuous and strictly increasing. With this, we can easily construct random variables with the same distribution. Let  $U \sim \text{unif}([0, 1])$ , then  $F_X^{-1}(U) \sim X$ . Indeed,

$$\mathbb{P}(F_X^{-1}(U) \leq t) = \mathbb{P}(F_X(F_X^{-1}(U)) \leq F_X(t)) = F_X(t).$$

Take  $([0, 1], \mathcal{B}_{[0,1]}, m)$ . Let  $\tilde{X}_n = F_{X_n}^{-1}(\omega)$  and  $\tilde{X} = F_X^{-1}(\omega)$ . Then  $\tilde{X}_n \sim X_n$  and  $\tilde{X} \sim X$ . Note that  $X_n \Rightarrow X$  implies  $F_{X_n} \rightarrow F_X$  pointwise. Checking  $\tilde{X}_n \rightarrow \tilde{X}$  a.e. is left as an exercise. □

## 2.6 Law of rare events

As an example for convergence of random variables, we have the following example presenting the law of rare events. First we need some definitions.

**Definition** (Poisson distribution). Say random variable  $X \sim \text{Poisson}(\lambda)$  has Poisson distribution with parameter  $\lambda$  for  $\lambda > 0$  if

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

Then,  $\mathbb{E}X = \lambda$ ,  $\mathbb{E}X^2 = \lambda^2 + \lambda$ , and  $\text{var } X = \lambda$ .

**Definition** (exponential distribution). Say random variable  $X \sim \text{Exp}(\lambda)$  has exponential distribution with parameter  $\lambda$  if  $X$  has density function

$$\rho_X(t) = \lambda \exp(-\lambda t) \quad \text{for } t \geq 0.$$

Then the corresponding CDF is

$$F_X(t) = \begin{cases} 1 - \exp(-\lambda t), & \text{if } t \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$



The exponential distribution is a memoryless distribution: suppose  $X \sim \text{Exp}(\lambda)$  and  $t > s > 0$ . Then,

$$\mathbb{P}(X \geq t \mid X \geq s) = \frac{\mathbb{P}(X \geq t)}{\mathbb{P}(X \geq s)} = \exp(-\lambda(t-s)) = \mathbb{P}(X \geq t-s).$$

**Example** (law of rare events). Let  $\lambda > 0$  be a fixed parameter. Let  $\{m_n\}_{n=1}^\infty$  be a non-decreasing sequence of integers such that

$$\lim_{n \rightarrow \infty} \frac{m_n}{n} = \lambda.$$

Create a triangular array as follows: for any  $n$ , let  $\{U_{n,j}\}_{j=1}^{m_n}$  be i.i.d. uniform random variables on  $[0, n]$ . Now let

$$X_n = |\{i \leq m_n : U_{n,i} \in [0, 1]\}|.$$

Then  $X_n \Rightarrow \text{Poisson}(\lambda)$ .

To prove this, let  $k \geq 0$ , we have

$$\begin{aligned} \mathbb{P}(X_n = k) &= \binom{m_n}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{m_n - k} \\ &= (1 + o(1)) \frac{m_n^k}{k!} n^{-k} \exp\left(-\frac{m_n}{n}\right) \\ &\rightarrow (1 + o(1)) \frac{\lambda^k e^{-\lambda}}{k!} \end{aligned}$$

as  $n \rightarrow \infty$ . Therefore,  $X_n \Rightarrow \text{Poisson}(\lambda)$  by the following proposition. △

**Proposition.** Let  $\{X_n\}_{n=1}^\infty$  and  $X$  be integer-valued random variables. Then  $X_n \Rightarrow X$  iff  $f_{X_n}(k) \rightarrow f_X(k)$  for any  $k \in \mathbb{Z}$ .

*Proof.* Verify directly using compactly supported continuous test functions. Since the test functions are compactly supported, only finite sums are involved. □

**Example.** Continuing the previous example, define

$$Y_n = \min_{1 \leq i \leq m_n} U_{n,i}.$$

Then  $Y_n \Rightarrow \text{Exp}(\lambda)$ .

Let  $t > 0$ , then

$$\begin{aligned}
 \mathbb{P}(Y_n \leq t) &= 1 - \mathbb{P}(Y_n > t) \\
 &= 1 - \mathbb{P}\left(\bigcap_{i=1}^{m_n} U_{n,i} > t\right) \\
 &= 1 - \mathbb{P}(U_{n,1} > t)^{m_n} \\
 &= 1 - \left(\frac{n-t}{n}\right)^{m_n} \\
 &\rightarrow 1 - \exp(-\lambda t)
 \end{aligned}$$

as  $n \rightarrow \infty$ .

△

**Example** (Poisson process). Fix integer parameter  $w > 0$ . For every large  $n$ , let

$$Z_n = (U_{n,1}^*, U_{n,2}^*, \dots, U_{n,w}^*)$$

be  $w$  smallest numbers from the set  $\{U_{n,i}\}_{i=1}^{m_n}$ , arranged in increasing order, then  $Z_n$  is a random vector in  $\mathbb{R}^w$ . We want to investigate the limiting distribution of  $\{Z_n\}_{n=1}^\infty$ , if any exists.

**Claim:** let  $\xi_1, \dots, \xi_w \sim \text{Exp}(\lambda)$  be i.i.d. Let the random vector

$$\xi = (\xi_1, \xi_1 + \xi_2, \dots, \xi_1 + \dots + \xi_w).$$

Then  $Z_n \Rightarrow \xi$ .

△

**Example** (Poisson process). Let  $\{X_n\}_{n=1}^\infty$  be a sequence of i.i.d. random variables that follows the exponential distribution  $\text{Exp}(\lambda)$ . Define Poisson process  $\{S_n\}_{n=0}^\infty$  via

$$S_n = \sum_{i=1}^n X_i.$$

This has two properties:

1. For each interval of length 1 on  $\mathbb{R}_+$ , the number of  $S_n$ 's within this interval follows the Poisson distribution with parameter  $\lambda$ .
2. For any finite collection of disjoint intervals, the variables corresponding to the number of  $S_n$ 's within the intervals are mutually independent. This is because of the memoryless property of the exponential distribution.

△

**Lemma.** If  $c_j \rightarrow 0$ ,  $a_j \rightarrow \infty$  and  $a_j c_j \rightarrow \lambda$ , then  $(1 + c_j)^{a_j} \rightarrow e^\lambda$ .

**Theorem** (Central limit theorem, basic form). Let  $\{b_n\}_{n=1}^\infty$  be i.i.d. random variables that are Bernoulli( $\frac{1}{2}$ ). Define  $S_n = \sum_{i=1}^n b_i$ . Then

$$\frac{S_n - \frac{n}{2}}{\sqrt{n/2}} \Rightarrow \mathcal{N}(0, 1).$$

*Proof.* Suppose  $n$  and  $k$  be some integer, then

$$\mathbb{P}(S_{2n} = n + k) = \frac{(2n)!}{(n+k)!(n-k)!} 2^{-2n}$$

Using Sterling's formula  $n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$ , we have

$$\frac{(2n)!}{(n+k)!(n-k)!} \sim \frac{(2n)^{2n}}{(n+k)^{n+k}(n-k)^{n-k}} \frac{\sqrt{2\pi n}}{\sqrt{2\pi(n+k)}\sqrt{2\pi(n-k)}}$$

Hence,

$$\mathbb{P}(S_n = n + k) \sim \left(1 + \frac{k}{n}\right)^{-n-k-\frac{1}{2}} \left(1 - \frac{k}{n}\right)^{-n+k-\frac{1}{2}} (\pi n)^{\frac{1}{2}}.$$

We also have

$$\left(1 + \frac{k}{n}\right)^{-n-k} \left(1 - \frac{k}{n}\right)^{-n+k} = \left(1 - \frac{k^2}{n^2}\right)^{-n} \left(1 + \frac{k}{n}\right)^{-k} \left(1 - \frac{k}{n}\right)^k.$$

Letting  $k = x\sqrt{n/2}$ , we know

$$\begin{aligned} \left(1 - \frac{k^2}{n^2}\right)^{-n} \left(1 + \frac{k}{n}\right)^{-k} \left(1 - \frac{k}{n}\right)^k &= \left(1 - \frac{x^2}{2n}\right)^{-n} \left(1 + \frac{x}{\sqrt{2n}}\right)^{-x\sqrt{n/2}} \left(1 - \frac{x}{\sqrt{2n}}\right)^{-x\sqrt{n/2}} \\ &\rightarrow \exp\left(\frac{x^2}{2}\right) \exp\left(-\frac{x^2}{2}\right) \exp\left(-\frac{x^2}{2}\right) \\ &= \exp\left(-\frac{x^2}{2}\right) \end{aligned}$$

by the lemma above.

Therefore, if  $2k/\sqrt{2n} \rightarrow x$ , then

$$\mathbb{P}(S_{2n} = n + k) \sim (\pi n)^{-1/2} e^{-x^2/2}.$$

Now,

$$\begin{aligned}\mathbb{P}\left(a \leq \frac{S_{2n} - n}{\sqrt{n/2}} \leq b\right) &= \mathbb{P}\left(n + a\sqrt{n/2} \leq S_{2n} \leq n + b\sqrt{n/2}\right) \\ &= \sum_{k \in [a\sqrt{n/2}, b\sqrt{n/2}] \cap \mathbb{Z}} \mathbb{P}(S_{2n} = n + k).\end{aligned}$$

Chaging variables  $m = k\sqrt{n/2}$ , we then have

$$\begin{aligned}\mathbb{P}\left(a \leq \frac{S_{2n} - n}{\sqrt{n/2}} \leq b\right) &= \sum_{m \in [a, b] \cap (2\mathbb{Z}/\sqrt{2n})} (2\pi)^{-1/2} e^{-x^2/2} (2/n)^{1/2} \\ &\approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx,\end{aligned}$$

using the definition of Riemann integral. The proof is now complete. □

**Definition** (Multivariate Gaussian). There are two equivalent definitions of multivariate Gaussian.

1.  $X \in \mathbb{R}^n$  has Gaussian distribution if there is non-random  $n$ -by- $n$  non-random matrix and  $\mu \in \mathbb{R}^n$  such that  $X \sim MG + \mu$ , where  $G$  is the standard Gaussian vector in  $\mathbb{R}^n$ . That is,  $G = (G_1, \dots, G_n)$  are random vector with i.i.d.  $\mathcal{N}(0, 1)$  components.
2.  $X \in \mathbb{R}^n$  has Gaussian distribution if for all non-random  $y \in \mathbb{R}^n$ ,  $\langle X, y \rangle$  is a scalar Gaussian random variables.

*Proof.* (1)  $\implies$  (2). Exercise. **\*\*\* TO-DO \*\*\***

(2)  $\implies$  (1). Later using characteristic functions. □

### 3 Law of large numbers

**Theorem** (Weak law of large numbers). Suppose  $\{X_n\}_{n=1}^\infty$  are mutually independent, identically distributed random variables. Suppose also  $\mathbb{E}|X_1| < \infty$ . Let  $S_n = \sum_{i=1}^n X_i$ , then  $S_n/n \xrightarrow{p} \mathbb{E}X_1$ .

*Proof.* We prove this using truncation method. Let  $\varepsilon > 0$  and choose truncation level  $M(\varepsilon) > 0$  satisfying the following:

$$\mathbb{E}|X_1 \mathbb{1}_{|X_1| > M}| < \varepsilon^2.$$

By dominated convergence theorem, there exists such an  $M$ . Note that this also implies

$$|\mathbb{E}X_1 - \mathbb{E}(X_1 \mathbb{1}_{|X_1| \leq M})| < \varepsilon.$$

□