# Probability

> The note of HKUST course MATH2431 Honors Probability in Spring 2023, by Zory ZHANG. In case of any broken math renderring in Github preview, please open this markdown file using your own markdown editor. PDF version (may not up-to-date) can be found here.

# Reminder

- Expectation exists: absolutely convergence.

# Ch0 Notations

$A \setminus B$: the set difference/minus. $A \triangle B$: the symmetric difference, $(A \setminus B) \cup (B \setminus A)$.

# Ch1 Events

## 1.1 Space

> Def. (**Sample space** $\Omega$) The set of all possible outcomes of an experiment(elementary events).

> Def. (**Event**) A subset of the sample space.

> Def. (**Field** $\mathcal{F}$) A collection of subsets of $\Omega$ which satisfies
>
> 1. $\emptyset \in \mathcal{F}$
> 2. If $A \in \mathcal{F}$, then $\overline{A} \in \mathcal{F}$
> 3. If $A, B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$

Corollary. If $A, B \in \mathcal{F}$, then $A \cap B \in \mathcal{F}$; $\Omega \in \mathcal{F}$; i.e. **Field is closed under finite unions & intersections.**

Rmk. Field is a collection of events that are of interest.

> Def. ($\sigma$-**Field** $\mathcal{F}$) A Field satisfies that (closed under countable unions & intersections)
>
> $$A_1, A_2, \cdots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$$

E.g. The smallest $\sigma$-Field $= \{\emptyset, \Omega\}$, largest= $2^{\Omega}$ .

Rmk. With any experiment, we may associate a pair $(\Omega, \mathcal{F})$.

Def. (**Measurable space**) A pair $(\Omega, \mathcal{F})$.

| Typical notation | Set jargon | Probability jargon |
|---|---|---|
| $\Omega$ | Collection of objects | Sample space |
| $\omega$ | Member of $\Omega$ | Elementary event, outcome |
| $A$ | Subset of $\Omega$ | Event that some outcome in $A$ occurs |
| $A^c$ | Complement of $A$ | Event that no outcome in $A$ occurs |
| $A \cap B$ | Intersection | Both $A$ and $B$ |
| $A \cup B$ | Union | Either $A$ or $B$ or both |
| $A \setminus B$ | Difference | $A$, but not $B$ |
| $A \triangle B$ | Symmetric difference | Either $A$ or $B$, but not both |
| $A \subseteq B$ | Inclusion | If $A$, then $B$ |
| $\varnothing$ | Empty set | Impossible event |
| $\Omega$ | Whole space | Certain event |

Table 1.1. The jargon of set theory and probability theory.

## 1.2 Assignment

Def. (**General measure**) Given a measurable space $(\Omega, \mathcal{F})$, a measure $\mu$ is a set function $\mu : \mathcal{F} \to [0, \infty]$, s.t.

1. $\mu(\emptyset) = 0$
2. (**countable additivity**) If $A_1, A_2, \ldots$ is a collection of **disjoint** members of $\mathcal{F}$, i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$, then $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

Def. (**Measure space**) A triple $(\Omega, \mathcal{F}, \mu)$.

Def. (**Probability measure**) A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying

1. $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1$.
2. (**countable additivity**) If $A_1, A_2, \ldots$ is a collection of **disjoint** members of $\mathcal{F}$, i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Ext.

1. A measure $\mu$ is probability measure iff $\mu(\Omega) = 1$.
2. Probability measure is a finite measure.
3. Lebesgue measure is a $\sigma$-finite measure.

> Def. (**Probability space**) A triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Rmk.

1. $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$
2. If $B \supset A$, then $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$, notice that $B \setminus A \in \mathcal{F}$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
4. Inclusion-exclusion principle
5. $\mathbb{P}$ is a continuous set function

# 1.3 Continuity

Recall:

1. $f$ is continuous at x if $\forall \{x_n\}, x_n \to x, n \to \infty \Longrightarrow \lim_{n\to\infty} f(x_n) = f(\lim_{n\to\infty} x_n) = f(x)$.
2. $\limsup_n x_n = \lim_{m\to\infty} \sup_{n\geq m} x_n = \lim_{m\to\infty} c_m$, where $c_m$ is monotonic, so that it must converge if we include $\pm\infty$.

> Def. (**Set limit**) Given $A_1, A_2, \cdots \in \mathcal{F}$,
>
> $$\limsup_{n\to\infty} A_n := \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n = \bigcap_{m=1}^{\infty} B_m = \{\omega \in \Omega : \omega \in A_n \text{ for infinitely many n}\}$$

Proof. Consider $\omega \in RHS$ or not. If yes, $\omega \in B_m, \forall m$; if not, disappear eventually.

$$\liminf_{n\to\infty} A_n := \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n = \bigcup_{m=1}^{\infty} C_m = \{\omega \in \Omega : \omega \in A_n \text{ for all but finitely many n}\}$$

Proof. Consider $\omega \in RHS$ or not. If yes, appear eventually; otherwise fail.

Rmk. $\liminf A_n \subset \limsup A_n$; if equal, we say $A_n$ converges.

E.g. Monotonic set sequence converges (if including $\infty$).

> Def. (**Continuity of general measure**) $\mu$ is continuous if
> $\forall \{A_n\}, A_n \to A, n \to \infty \longrightarrow \lim_{n\to\infty} \mu(A_n) = \mu(\lim_{n\to\infty} A_n) = \mu(A)$.

Rmk. Notice the closeness under union&intersection gives that $A := \limsup_n A_n \in \mathcal{F}$.

Thm. (**Countable additivity implies continuity**)

Proof. For all convergent sequence $\{A_n\}$, which means

1. Case1: monotonic increasing An ($A_{n-1} \subset A_n$) Recall countable additivity (*), construct $D_n = A_n \setminus A_{n-1}$, then

$$\begin{aligned}\mu(A) = \mu(\lim_{n\to\infty} A_n) &:= \mu(\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} A_m)\\ &= \mu(\bigcup_{n=1}^{\infty} A_n) = \mu(\bigcup_{n=1}^{\infty} D_n) \overset{(*)}{=} \sum_{n=1}^{\infty}\mu(D_n)\\ &= \lim_{n\to\infty}\sum_{i=1}^{n}\mu(D_i) = \lim_{n\to\infty}\mu(\bigcup_{i=1}^{n} D_i) = \lim_{n\to\infty}\mu(A_n)\end{aligned}$$

2. Case2: monotonic decreasing An ($A_{n-1} \supset A_n$) Construct $E_n = A_n \setminus A_{n+1}$, then

$$\begin{aligned}\mu(A) = \mu(\lim_{n\to\infty} A_n) &:= \mu(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m)\\ &= \mu(\bigcup_{n=1}^{\infty} A_n) = \mu(\bigcup_{n=1}^{\infty} E_n) \overset{(*)}{=} \sum_{n=1}^{\infty}\mu(E_n)\\ &= \lim_{n\to\infty}\sum_{i=1}^{n}\mu(E_i) = \lim_{n\to\infty}\mu(\bigcup_{i=1}^{n} E_i) = \lim_{n\to\infty}\mu(A_n)\end{aligned}$$

3. Case3: general An Recall $B_n = \bigcup_{m=n}^{\infty} A_m, C_n = \bigcap_{m=n}^{\infty} A_m$. Clearly $C_n \subset A_n \subset B_n$, and that $B_n$ is monotonic decreasing, $C_n$ is monotonic increasing. From case1, we know that

$$\begin{aligned}\limsup_{n\to\infty}\mu(A_n) \le \lim_{n\to\infty}\mu(B_n) =& \mu(\lim_{n\to\infty} B_n)\\ =& \mu(B) = \mu(A) = \mu(C)\\ =& \mu(\lim_{n\to\infty} C_n) = \lim_{n\to\infty}\mu(C_n) \le \liminf_{n\to\infty}\mu(A_n)\end{aligned}$$

However, $\limsup_{n\to\infty} A_n \ge \liminf_{n\to\infty} A_n$, therefore
$\lim_{n\to\infty}\mu(A_n) = \limsup_{n\to\infty}\mu(A_n) = \liminf_{n\to\infty}\mu(A_n) = \mu(A)$.

Conclusion: $\mu$ is a continuous set function.

Prop. (**Finite additivity + continuity <=> countable additivity**) Proof. (=>) Recall continuity: $\forall\{A_n\}, A_n \to A, n \to \infty \longrightarrow \lim_{n\to\infty}\mu(A_n) = \mu(\lim_{n\to\infty} A_n) = \mu(A)$ and (countable additivity) If $A_1, A_2, \ldots$ is a collection of disjoint members of $\mathcal{F}$, then

$$\mu(\bigcup_{i=1}^{\infty} A_i) = \mu(\lim_{n\to\infty}\bigcup_{i=1}^{n} A_i) = \lim_{n\to\infty}\mu(\bigcup_{i=1}^{n} A_i) = \lim_{n\to\infty}\sum_{i=1}^{n}\mu(A_i) = \sum_{i=1}^{\infty}\mu(A_i)$$

E.g. $\Omega = \mathbb{R}$, $\mathcal{F}$ is some $\sigma$-field including all intervals. If we know $\mathbb{P}([a,b])$ for all $[a,b] \subset \mathbb{R}$, then

$$\lim_{n\to\infty}\mathbb{P}([a+\frac{1}{n}, b-\frac{1}{n}]) = \mathbb{P}(\lim_{n\to\infty}[a+\frac{1}{n}, b-\frac{1}{n}]) = \mathbb{P}((a,b))$$

, which uniquely extends all open intervals.

# 1.4 Conditional probability, independence

Rmk. "B has occurred" means $\omega \in B$. We may consider B as another adjusted sample space.

> Def. (**Conditional probability**) If $\mathbb{P}(B) > 0$, then $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ (pronounced as probability of A given B).

Lemma. (**Law of total probability**) Let $B_1, B_2, \ldots, B_n$ be a partition of $\Omega$(disjoint $B_i$, $\sqcup B_i = \Omega$) and $\mathbb{P}(B_i) > 0$, then $\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$.

Rmk. $\mathbb{P}(\cdot|B) = \mathbb{Q}(\cdot)$ is also a probability measure, therefore B can be viewed as another adjusted sample space.

Cor. (**Bayes' Theorm**) Skipped.

> Def. (**Independence of events**) Let $A, B \in \mathcal{F}$, that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, denoted as $A \perp B$.

> Def. (**Mutually independence**) Let $\{A_i\}, A_i \in \mathcal{F}$, that $\mathbb{P}(\bigcap_{i \in J} A_i) = \prod_{i \in J} \mathbb{P}(A_i), \forall J \subset I$.

> Def. (**Pairwise independence**) Let $\{A_i\}, A_i \in \mathcal{F}$, that $\mathbb{P}(\bigcap_{i \in J} A_i) = \prod_{i \in J} \mathbb{P}(A_i), \forall J \subset I, |J| = 2$.

Cor. If $A \perp B$, then $A \perp \overline{B}, \overline{A} \perp B, \overline{A} \perp \overline{B}$.

Cor. If A, B, and C are mutually independent, then A is independent of whatever events formed from B and C.

> Def. (**Generated $\sigma$-field**) $\sigma(A)$ is the smallest $\sigma$-field generated by A, a collection of subset of $\Omega$, by keeping taking countable union, intersection, or complement.

> Def. (**Product space**) Given $(\Omega_1, \mathcal{F}_1, \mathbb{P}1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}2)$, then their product space is $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = \Omega_1 \times \Omega_2, \mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$, $\mathbb{P}$ is defined on $\Omega$ by $\mathbb{P}(A_1 \times A_2) = \mathbb{P}1(A_1)\mathbb{P}2(A_2)$ and then continuity.

> Def. (**Independent trials**) If a combined experiment is modeled by the product space, the two sub-experiments are independent.

# Ch2 Random variable

## 2.1 "Preimage" viewpoint

Rmk. To answer the question that $\mathbb{P}(X \in B) = \mathbb{P}(\{\omega : X(\omega) \in B\})$, random variable acts as a "translation", where $\{\omega : X(\omega) \in B\}$ is the preimage $X^{-1}(B) \in \mathcal{F}$.

> Def. (**Random variable**) Given $(\Omega, \mathcal{F}, \mathbb{P})$, consider a function $X : \Omega \to \mathbb{R}$, s.t. $X^{-1}(B) \in \mathcal{F}, \forall$ intervals $B \subset \mathbb{R}$, called "X is $\mathcal{F}$-measurable"(*).

Observ.

1. Preimage can be interchanged with any set operation, e.g. $X^{-1}(B_1 \cap B_2) = X^{-1}(B_1) \cap X^{-1}(B_2)$.
2. $\mathcal{F}$ is a $\sigma$-field: closed under taking countable intersections...

Rmk. Equivalently, holding on the following B will be sufficient instead of the condition (*):

1. All open intervals
2. All closed intervals
3. All (a,b] intervals
4. All $(-\infty, x]$ intervals (simplest)

Prop. From Prop8.2 in Frederick Fong's book, if $\mu$ is $\mathcal{F}$-measurable, and $\varphi$ is continuous, then $\varphi(\mu)$ is also $\mathcal{F}$-measurable.

# 2.2 Borel Measure space

Rmk. Eventually, by taking countable set operations, we get a **Borel set**. The collection of them are a $\sigma$-field generated by all open intervals, called **Borel $\sigma$-field**, e.g. $(a, b) \in \mathcal{B}(\mathbb{R}), \mathbb{Q} \in \mathcal{B}(\mathbb{R}), \mathbb{R} \setminus \mathbb{Q} \in \mathcal{B}(\mathbb{R})$.

Rmk. Prefer to define by "open set" instead of "open interval", because the latter only works in $\mathbb{R}$.

Rmk. (level 2 understanding) $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a measurable map, s.t. $X^{-1}(B) \in \mathcal{F}, \forall B \in \mathcal{B}(\mathbb{R})$.

# 2.3 Borel Probability space

Motiv. (level 3 understanding) Is there a probability measure Q, such that Q(B) can answer the question of $\mathbb{P}(X \in B)$? That is consider $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), Q)$.

Rmk. Define $Q(B) := \mathbb{P}(X^{-1}(B))$, or $Q = \mathbb{P} \circ X^{-1}$, and $Q$ is called the **push forward measure, induced measure**, or **(cumulative) distribution function(c.d.f.)** of X.

Proof. (countable additivity) Notice that $X^{-1}(B_n) = \{\omega : X(\omega) \in B_n\}$, and must be disjoint if $B_n$ disjoint, therefore $Q(\bigcup_{i=1}^{\infty} B_i) = \bigcup_{i=1}^{\infty} \mathbb{P}(X^{-1}(B_i)) = \sum_{i=1}^{\infty} Q(B_i)$.

# 2.4 Distribution function

Rmk. Because of the continuity, it's enough to know the **(cumulative) distribution function** $F_X(x) := \mathbb{P} \circ X^{-1}((-\infty, x]), \forall x$, given Caratheodory's extension theorem, to understand the prabability distribution of a random variable.

Prop. c.d.f is **right continuous**.

Proof. Given the monotonicity of $F_X$,

$$\lim_{h\to 0^+} F_X(x+h) = \lim_{h\to 0^+} \mathbb{P}\circ X^{-1}((-\infty, x+h]) = \mathbb{P}\circ X^{-1}(\lim_{h\to 0^+}(-\infty, x+h])$$

$$= \mathbb{P}\circ X^{-1}(\lim_{n\to\infty}(-\infty, x+\frac{1}{n}])$$

$$= \mathbb{P}\circ X^{-1}(\cap_{n=1}^{\infty}(-\infty, x+\frac{1}{n}])$$

$$= \mathbb{P}\circ X^{-1}((-\infty, x]) = F_X(x)$$

In contrast, for $\lim_{h\to 0^-} F_X(x+h) = \mathbb{P}\circ X^{-1}((-\infty, x)) = F_X(x-)$.

> Def. (**Constant r.v.**) for some constant c, $\mathbb{P}(X=c) = \mathbb{P}(\omega\in\Omega : X(\omega)=c) = 1$.

> Def. (**Bernoulli/indicator r.v.**) $A\in\mathcal{F}$, r.v. $\mathbb{1}_{\mathbb{A}}$ that $\mathbb{1}_{\mathbb{A}}(\omega) = [\omega\in A]$.

> Def. (**Discrete r.v.**) X takes values in some countable subset of $\mathbb{R}$, as a property of distribution func.
>
> Def. (**Probability mass func p.m.f.**)
> $f(x) = \mathbb{P}\circ X^{-1}(\{x\}) = \mathbb{P}(X=x) = F_X(x) - F_X(x-)$.

> Def. (**(Absolutely) continuous r.v.**) c.d.f. $F_X(x)$ can be written as an integral of some Lebesgue integrable function $f : \mathbb{R}\to[0,\infty)$. $f$ is called **probability density function**(p.d.f.).

# 2.5 Random vector

> Def.1 (**Random vector**) Given $(\Omega, \mathcal{F}, \mathbb{P})$, consider a function $\vec{X} : \Omega\to\mathbb{R}^n$, s.t. $X^{-1}(D)\in\mathcal{F}, \forall D\in B(\mathbb{R}^n)$.

Rmk. We can replace all Borel sets by all "rectangles".

> Def.2 (**Random vector**) Given $(\Omega, \mathcal{F}, \mathbb{P})$, consider n random variables $X_i$, which means $X_i^{-1}(D_i)\in\mathcal{F}, \forall D_i\in B(\mathbb{R})$, to constitute $\vec{X} = [X_1, X_2, \dots, X_n]$.

> Def. (**Random function**) By new metric space -> open set -> Borel $\sigma$-field.

> Def. (**Joint distribution func**) $F_{X_1,X_2}(x_1,x_2) = \mathbb{P}\circ\vec{X}^{-1}((-\infty,x_1]\times(-\infty,x_2])$.

> Def. (**Marginal distribution func**) $F_X(x) = \lim_{y\uparrow\infty} F_{X,Y}(x,y)$.

Rmk. The marginal distribution of a subset of a collection of random variables is the probability distribution of the variables contained in the subset. It gives the probabilities of various values of the variables in the subset without reference to the values of the other variables.

> Def. (**Jointly (absolutely) continuous r.v.s.**) distribution func $F_{X,Y}(x,y)$ can be written as an Lebesgue double integral of some functions $f:\mathbb{R}^2\to[0,\infty)$. $f$ is called **joint probability density function**.

Rmk. (important) If $X=Y$, they're not jointly continuous, because $\iint f(x)\mathbb{1}(x=y)\,\mathrm{d}x\,\mathrm{d}y=\int f(x)\left[\int \mathbb{1}(x=y)\,\mathrm{d}y\right]\mathrm{d}x=0$.

## 2.6 Independence of r.v.s.

> Def. (**Independence of r.v.s.**) $X,Y:\Omega\to\mathbb{R}$ are independent($X\perp Y$) if $\mathbb{P}(X\in E,Y\in F)=\mathbb{P}(X\in E)\mathbb{P}(Y\in F),\forall E,F\in B(\mathbb{R})$.

Rmk. The following two are also equivalent to the definition:

1. Holding on all half spaces is enough: $F_{X,Y}(x,y)=F_X(x)F_Y(y)$.
2. (important) Holding on all single points is enough: p.m.f if discrete; otherwise p.d.f, i.e., $f_{X,Y}(x,y)=f_X(x)f_Y(y)$.

> Def. (**Mutually independence of r.v.s.**) $F_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n)=F_{X_1}(x_1)F_{X_2}(x_2)\dots F_{X_n}(x_n)$. By choosing $x_i=\infty$, we have the freedom to restrict on all subsets of r.v.s, not only all of them.

Rmk. Independence of events is a special case of indepence of r.v., given that $A\perp B\iff I_A\perp I_B$.

Thm. If $X\perp Y$, and two Borel measurable functions $g,h:\mathbb{R}\to\mathbb{R}$, then $g(X),h(Y)$ are still r.v.s., and that $g(X)\perp h(Y)$. Prove by measure theory.

E.g. $\omega=(\omega_1,\omega_2),X_1(\omega)\equiv\tilde{X}_1(\omega_1),X_2(\omega)\equiv\tilde{X}_2(\omega_2)$, where $X_1,X_2$ are defined on the product space, then $X_1\perp X_2$.

# Ch3 Discrete world

## 3.1 Discrete distributions

> Def. (**Bernoulli distribution**, $X\sim Be(p)$ ) Bernoulli trial is that $A\in\mathcal{F}$, and we call the trial a success if $A$ occurs. Bernoulli distribution is based on single Bernoulli trial.

> Def. (**Binomial distribution**, $Y\sim Bin(n,p)$ ) Perform n independent Bernoulli trials with $p=\mathbb{P}(A)$, and let Bernoulli r.v.s. $X_1,X_2\dots X_n$ be the indicator funciton of success of the experiments. Let $Y:=\sum X_i$, then the p.m.f. $f_Y(k)=\binom{n}{k}p^k(1-p)^{n-k},k=1,2,\dots,n$. **Binomial process** is that $Y_n$ to be the number of successes in the first n $Be(p)$ trials.

> Def. (**Geometric distribution**, $W \sim Geom(p)$ ) Keep performing Bernoulli trials until the first success, and let $W :=$ the waiting time, then $f_W(k) = (1-p)^{k-1}p, k = 1, 2, \dots$

> Def. (**Negative Binomial distribution**, $W_r \sim NB(r, p)$ ) Let $W_r :=$ the waiting time for r successes, then $f_{W_r}(k) = \binom{k-1}{k-r}p^{r-1}(1-p)^{k-r}p, k = 1, 2, \dots$

> Def. (**Poisson distribution**, $X \sim poisson(\lambda)$ ) $f_X(k) := \frac{\lambda^k}{k!}e^{-\lambda}, k = 1, 2, \dots$ The Poisson distribution is an approximation of a binomial distribution of a rare event in the case that **n is large, p is small**, and $\lambda = np$ is moderate.

# 3.2 Expectation of discrete r.v.

> Def. (**Expectation**) A discrete r.v. X taking values from $x_1, x_2, \dots$ with p.m.f. $f_X(x)$. The expectation $\mathbb{E}X := \sum_i x_i f_X(x_i) = \sum_{x:f_X(x)>0} x f_X(x)$ exists if the sum is absolutely convergent. The summation range is indicating the countability.

Rmk. Our best guess of X to minimize $(X - \mathbb{E}X)^2$.

Rmk. Absolutely convergence ensures the reorderability.

Lemma. (**Change of variable for Lebesgue integral**) $\mathbb{E}g(X) = \sum_{x:f_X(x)>0} g(x)f_X(x)$.

Lemma. (3.6.6) Let $X, Y$ be two r.v.s. jointly discrete, and $g(X, Y)$ is still a r.v, then $\mathbb{E}g(X, Y) = \sum_{x,y:f_{X,Y}(x,y)>0} g(x, y)f_{X,Y}(x, y)$.

Rmk. Expectation has finite linearity.

> Def. (**Uncorrelated**) $\mathbb{E}(XY) = \mathbb{E}(X)E(Y)$. Independence implies uncorrelation.

# 3.3 Variance of discrete r.v.

> Def. (**k-th moments**) $m_k := \mathbb{E}X^k$.

> Def. (**k-th central moments**) $\sigma_k := \mathbb{E}(X - \mathbb{E}X)^k$.

> Def. (**Variance**) 2-nd central moments $\sigma_2 = m_2 - m_1^2$, which describes the randomness of a r.v.

> Def. (**Covariance**) $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}XY - \mathbb{E}X \cdot \mathbb{E}Y$.

Rmk. Properties of variance.

1. $Var(aX + b) = a^2 VarX$
2. $VarX = Cov(X, X)$
3. $Var(X + Y) = VarX + VarY + 2Cov(X, Y)$

## 3.4 Conditional distribution & expectation

> Joint p.d.f is a symmetric viewpoint, here we try an asymmetric viewpoint.

> Def. (**Conditional distribution** of Y given event "X=x") To answer $\mathbb{P}(Y \in \cdot | X = x) : B(\mathbb{R}) \to [0,1]$, define the c.m.f. $F_{Y|X}(\cdot|x) := \mathbb{P}(Y \leq \cdot | X = x)$. The p.m.f $f_{Y|X}(\cdot|x) = \frac{f_{X,Y}(x,\cdot)}{f_X(x)}$, if $f_X(x) > 0$.

> Def. (**Conditional expectation** of Y given event " $X = x := \cdot$ ") $\psi_{Y|X} : \mathbb{R} \to \mathbb{R}$, $\psi_{Y|X}(\cdot) := \mathbb{E}(Y|X = \cdot) := \sum_y y f_{Y|X}(y|\cdot)$.

Rmk. Our best guess of Y to minimize $(Y - \mathbb{E}Y)^2$, after know that $X = x$.

> Def. (**Conditional expectation** of Y given a r.v. X) $\psi_Y : r.v. \to r.v.$, $\psi_Y(X) := \mathbb{E}(Y|X) = \sum_x \mathbb{1}(X = x) \cdot \mathbb{E}(Y|X = x) = \sum_x \mathbb{1}(X = x)\psi_{Y|X}(x)$.

Rmk. $\psi_Y(X)$ is from the same probability space as $X$, and it's our best guess of the "strategy" of Y to minimize $(Y - \mathbb{E}Y)^2$, after knowing the distribution of $X$.

> Def. ($\sigma$-**field induced by a r.v. X**) $\sigma(X) := \{X^{-1}(B) : \forall B \in \mathcal{B}(\mathbb{R})\}$, which characterizes all information that can be obtained by observing the value of X.

E.g. For dice rolling, $X(\omega) := \mathbb{1}(\omega \text{ is odd})$, then $\sigma(X) := \{\emptyset, \Omega, \{1,3,5\}, \{2,4,6\}\}$.

Rmk. $\sigma(Y) \subset \sigma(\sigma(Y), \sigma(Z))$, here we use the above defined concept of generated $\sigma$-field. The meaning is that the information from both Y and Z is more than the information from Y.

> Def. (**Conditional expectation** of Y given a $\sigma$-field) skipped.

Rmk. (**Tower property**) Consider two $\sigma$-field $H_1 \subset H_2 \subset \mathcal{F}$, then $\mathbb{E}(\mathbb{E}(X|H_1)|H_2) = \mathbb{E}(\mathbb{E}(X|H_2)|H_1) = \mathbb{E}(X|H_1)$. It's reasonable if we consider $\mathbb{E}(Y|X) := \psi_Y(X)$ as a function that maps a r.v. X to another r.v. in the same probability space. Then $\sigma(\psi_Y(X)) \subset \sigma(X)$ because there will at most no information lost during mapping.

> Thm. (**Law of total expectation**) $\mathbb{E}\mathbb{E}(Y|X) = \mathbb{E}\psi_Y(X) = \sum_x \mathbb{P}(X = x) \cdot \mathbb{E}(Y|X = x) = \mathbb{E}Y$.

Rmk. $\mathbb{E}(\mathbb{E}(Y|X)g(X)) = \mathbb{E}(Yg(X))$.

## 3.6 Sum of discrete r.v.s.

- (Convolution) If $X \perp Y$, $f_{X+Y}(z) = \sum_x f_X(x) f_Y(z - x)$.

# Ch4 Continuous world

$\mathbb{P}(u \le x \le u + \mathrm{d}u) = f_X(u)\,\mathrm{d}u$. More generally, borel set $B \in \mathbb{R}^n, \mathbb{P}(\vec{X} \in B) = \mathbb{P} \circ \vec{X}^{-1}(B) = \int_B f_X(u)\,\mathrm{d}u$.

# 4.1 Expectation of continuous r.v.

> Def. (**Expectation of continuous r.v.** X) $\mathbb{E}X := \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x$. It exists if absolutely convergent.

Lemma. (**Tail probability**) $\mathbb{E}X = \int_{-\infty}^{\infty}(1 - F_X(x))\,\mathrm{d}x = \int_{-\infty}^{\infty}\mathbb{P}(X \ge x)\,\mathrm{d}x$.

Thm. (**Change of variable for Lebesgue integral**) Suppose X and g(X) are both continuous r.v., $g(X) \ge 0, \mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x$.

Proof.

$$\begin{aligned}\mathbb{E}g(X) &= \int_0^{\infty}\mathbb{P}(g(X) > x)\,\mathrm{d}x = \int_0^{\infty}\left[\int_{B:=\{y:g(y)>x\}} f_X(y)\,\mathrm{d}y\right]\mathrm{d}x, \\ &= \int_{-\infty}^{\infty}\left[\int_0^{g(y)} f_X(y)\,\mathrm{d}x\right]\mathrm{d}y = \int_{-\infty}^{\infty} g(y) f_X(y)\,\mathrm{d}y\end{aligned}$$

# 4.2 Variance of continuous r.v.

> Def. (**k-th moments**) $m_k := \mathbb{E}X^k = \int x^k f_X(x)\,\mathrm{d}x$.

> Def. (**k-th central moments**) $\sigma_k := \mathbb{E}(X - \mathbb{E}X)^k = \int (x - \mathbb{E}X)^k f_X(x)\,\mathrm{d}x$.

E.g. for normal distribution, $\sigma_i = \begin{cases} 0, \text{ if i odd,} \\ \sigma^i(i-1)!!, \text{ o.w.} \end{cases}$

# 4.3 Continuous distributions

> Def. (**Uniform distribution** $X \sim \mathcal{U}[0,1]$ ) $f_X(x) = 1, F_X(x) = x$ over $[0,1]$, otherwise nature.

Thm. (**Inverse transform sampling**) Given a distribution function G, we can generate a r.v. Y with $F_Y = G$ by only generating a r.v. $U \sim \mathcal{U}[0,1]$.

Proof. We know G, as a distribution function, is non-decreasing and right continuous, with $G(x) \in [0,1]$.

1. Case 1: G is strictly increasing, and therefore invertible. Claim that $Y := G^{-1}(U)$, then

$$F_Y(x) = \mathbb{P}(Y \le x) = \mathbb{P}(G^{-1}(U) \le x) = \mathbb{P}(U \le G(x)) = G(x)$$

2. Case 2: otherwise, what we really want of $G^{-1}$ is the property that $G^{-1}(z) \le x \iff z \le G(x)$. Consider to construct a generalized inverse function to satisfy this property. Claim that $Y := invG(U)$, where

$invG(z) := \inf_{t\in\mathbb{R}}\{G(t) \geq z\} = \inf H_z$. Due to right continuity and monotone, $invG(z) \in H_z$. To preserve the $F_Y = G$, we only need to verify the desired property.

$$invG(z) \leq x \iff \inf H_z \leq x \iff x \in H_z \iff G(x) \geq z$$

Rmk. It can help generate psuedo-random numbers following specific distribution given a uniformly distributed psuedo-random numbers generator.

> Def. (**Exponential distribution**, $X \sim Exp(\lambda)$ )
>
> $$F_X(x) = \begin{cases} 0, x < 0 \\ \\ 1 - e^{-\lambda x}, x \geq 0 \end{cases}, f_X(x) = \begin{cases} 0, x < 0 \\ \\ \lambda e^{-\lambda x}, x \geq 0 \end{cases}$$

Rmk.

1. X is the waiting time for the first success in a poisson process.
2. $\mathbb{E}X = \frac{1}{\lambda}$.
3. (Memoryless) $\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s)$.
4. (Failture/hazard rate for positive r.v.) $r(t) = \lim_{h\downarrow 0} \frac{1}{h}\mathbb{P}(X \leq t + h | X > t) = \frac{f(t)}{1-F(t)} = \lambda$.

> Def. (**Normal/Gaussian distribution**, $X \sim \mathcal{N}(\mu, \sigma^2)$ ) $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$.
> Standard normal distribution $Y \sim \mathcal{N}(0, 1)$, denote the p.d.f. as $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(\frac{-x^2}{2})$.

Rmk. To compute $I = \int_{-\infty}^{\infty} \phi(x)\,\mathrm{d}x$, consider

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \phi(x)\,\mathrm{d}x \int_{-\infty}^{\infty} \phi(y)\,\mathrm{d}y \\ \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-\frac{x^2+y^2}{2})\,\mathrm{d}x\,\mathrm{d}y \\ \\ &= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} \exp(-\frac{r^2}{2})r\,\mathrm{d}r\,\mathrm{d}\theta = 1 \end{aligned}$$

> Def. (**(standard) Cauchy distribution**, $X \sim \mathcal{N}(\mu, \sigma^2)$ ) p.d.f.
> $f_X(x) = \frac{1}{\pi(1+x^2)}, F_X(x) = \arctan(x) + \frac{\pi}{2}$.

Rmk. (Heavy tail distribution without moment) Notice the $\mathbb{E}X = \int_{-\infty}^{\infty} |x| f_X(x)\,\mathrm{d}x = \infty$ does not exist.

# 4.4 Dependence & joint distribution(symmetric)

Thm. (**Change of variable for Lebesgue integral**) $g : \mathbb{R}^2 \to \mathbb{R}$ is Lebesgue integrable, then $\mathbb{E}g(X, Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y$.

Rmk. By marginal distribution func, linearity of expectation can be proved.

> Def. (**Correlation coefficient**) $\frac{Cov(X,Y)}{\sqrt{VarX \cdot VarY}}$, which is invariant under scaling.

> Def. (**Bivariate normal**) $f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2))$.

Property. $X, Y$ are bivariate normal and uncorrelated($\rho = 0$), then they're independent.
Examples: Wiki

# 4.5 Conditional distribution(asymmetric)

> Def. (**Regular conditional dist**) If $f_X(x) > 0$,
> $F_{Y|X}(y|x) := \mathbb{P}(Y \le y | X = x) := \int_{-\infty}^{y} \frac{f_{X,Y}(x,u)}{f_X(x)} \, \mathrm{d}u$; the cond. p.d.f $f_{Y|X}(\cdot|x) = \frac{f_{X,Y}(x,\cdot)}{f_X(x)}$.

Rmk. To compute $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = C_y g_y(x)$, we can view it as a p.d.f. of x given fixed y, because we know that $C_y = \frac{1}{\int g_y(x)}$, so we may ignore $f_Y(y)$, which is included in $C_y$.

E.g. For bivariate normal, $f_{X|Y}(x|y)$, given Y=y, $X \sim \mathcal{N}(\rho y, 1-\rho^2)$. When $\rho = 0$, they're independent. In contrast, when $\rho \to 1$, $X \to Y$. In general, $X = \rho Y + \sqrt{1-\rho^2} Z$, where $Z \sim \mathcal{N}(0,1), Y \perp Z$. Furthermore,

$$\begin{pmatrix} X \\ \\ Y \end{pmatrix} = \begin{bmatrix} \rho & \sqrt{1-\rho^2} \\ 1 & 0 \end{bmatrix} \begin{pmatrix} Y \\ \\ Z \end{pmatrix}$$

> Def. (**Conditional expectation** of Y given event " $X = x := \cdot$ ") $\psi_{Y|X} : \mathbb{R} \to \mathbb{R}$,
> $\psi_{Y|X}(\cdot) := \mathbb{E}(Y|X = \cdot) := \int y f_{Y|X}(y|\cdot) \, \mathrm{d}y$.
>
> Def. (**Conditional expectation** of Y given a r.v. X) $\psi_Y : r.v. \to r.v.$,
> $\psi_Y(X) := \mathbb{E}(Y|X) = \int \mathbb{1}(X = x) \cdot \mathbb{E}(Y|X = x) \, \mathrm{d}x = \int \mathbb{1}(X = x) \psi_{Y|X}(x) \, \mathrm{d}x$.
>
> Def. (**Conditional expectation** of Y given a $\sigma$-field) skipped.
>
> Thm. (**Law of total expectation**)
> $\mathbb{E}\mathbb{E}(Y|X) = \mathbb{E}\psi_Y(X) = \int \mathbb{P}(X = x) \cdot \mathbb{E}(Y|X = x) \, \mathrm{d}x = \mathbb{E}Y$.
>
> Rmk. $\mathbb{E}(\mathbb{E}(Y|X)g(X)) = \mathbb{E}(Yg(X))$.

# 4.6 Change of variable

$(U,V) \xrightarrow{g} (X,Y)$, where $g^{-1}$ exists. Furthermore, the Jacobian
$J = \det \frac{\partial(x,y)}{\partial(u,v)} = \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix} \neq 0$. Then $f_{X,Y}(x,y) = f_{U,V}(g^{-1}(x,y)) \frac{1}{\left|\det \frac{\partial(x,y)}{\partial(u,v)}\right|}$.

# 4.7 Sum of continuous r.v.s.

$$F_{X+Y}(z) = \mathbb{P}(X+Y \le z) = \iint_{x+y\le z} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

$$\overset{t=x+y}{=} \iint_{t\le z} f_{X,Y}(x,t-x)\left|\det\frac{\partial(x,t)}{\partial(x,y)}\right|\mathrm{d}x\,\mathrm{d}t$$

$$= \int_{-\infty}^{z}\left(\int f_{X,Y}(x,t-x)\,\mathrm{d}x\right)\mathrm{d}t$$

$$f_{X+Y}(z) = \int f_{X,Y}(x,z-x)\,\mathrm{d}x$$

Eventually, convolution if independent.

Rmk. Sum of independent normal is still normal.

# Ch5 Generating function

Motivation. $G_X(s) = \mathbb{E}s^X, M_X(t) = \mathbb{E}e^{tX}, \phi_X(t) = \mathbb{E}e^{itX}$. Then $X \perp Y$, we have $G_{X+Y} = G_X G_Y$. The converse may not be true.

> Def. (**Generating function**) $G_a(s) = \sum_{i=0}^{\infty} a_i s^i$, then $a_i = \frac{G_a^{(i)}(0)}{i!}$ if within the radius of convergence.

Rmk. (**Radius of convergence** R) Recall various convergence tests.

1. When $s \in (-R, R)$, $G_a(s)$ is absolutely convergent and can be differentiated or integrated term by term.
2. If $\exists R' \in (0, R], \forall s \in [-R', R'], G_a(s) = G_b(s)$, then $a_i = b_i, \forall i$.
3. If $R > 0$ for $G_a(s)$, then $\{a_n\}$ is **uniquely determined** by taking derivative.

Thm. (**Abel's theorem**) If $a_i > 0$, $G_a(s)$ has $R = 1$, and $G_a(1)$ exists (diverging to positive infinity is included), then $G_a(s)$ is left continuous at $s = 1$, i.e. $G_a(1) := \sum_{i=0}^{\infty} a_i = \lim_{s\uparrow 1} G_a(s)$.

## 5.1 Probability generating function

> Def. (**Probability generating function**) For nonnegative interger-valued discrete r.v. X, the p.g.f. $G_X(s) = \mathbb{E}s^X = \sum_{i=0}^{\infty} f_X(i)s^i$ is the g.f. of $a_i = f_X(i)$.

Rmk.

1. $\sum_{i=0}^{\infty} f_X(i)s^i \le \sum_{i=0}^{\infty} s^i$, therefore $R \ge 1$.
2. $G_X(1) = \sum_i \mathbb{P}(X = i) = 1 - \mathbb{P}(X = \infty)$. If not 1, we say that X is **defective** with defective distribution func, and all moments equal $+\infty$.
3. $\mathbb{E}X = \lim_{s\uparrow 1} G'_X(s) := G'_X(1)$.

4. $\mathbb{E}X^{\underline{k}} := \mathbb{E}[X(X-1)\dots(X-k+1)] = \lim_{s\uparrow 1} G_X^{(k)}(s) := G_X^{(k)}(1)$.
5. Limitation: hard to directly generate moment. E.g. $VarX = G''(1) + G'(1) - G'(1)^2$.

E.g.

1. $X \sim Be(p), G_X(s) = 1 - p + ps$.
2. $Y \sim Bin(n,p), G_Y(s) = G_X^n(s)$.
3. $W \sim Geom(p), G_W(s) = \frac{ps}{1-s(1-p)}$.
4. $X \sim poisson(\lambda), G_X(s) = e^{\lambda(s-1)}$.

Thm 5.1.1. (**Sum of a random number of i.i.d. r.v.s.**) N i.i.d. r.v. $X_i$ are independent of N, $T = \sum_{i=1}^N X_i$, then $G_T(s) = G_N(G_X(s))$. Prove by the law of total expectation.

E.g. The sum of a poisson number of i.i.d. Bernoulli r.v.s. is still poisson.

> Def. (**Joint probability generating function**) For nonnegative interger-valued discrete r.v.s. $X_1, X_2$, the j.p.g.f. $G_{X,Y}(s,t) = \mathbb{E}s^X t^Y = \sum_{i=0}^\infty \sum_{j=0}^\infty f_{X,Y}(i,j)s^i t^j$.

Thm 5.1.2. $X \perp Y \iff G_{X,Y}(s,t) = G_X(s)G_Y(t)$.

## 5.1.1 Application 1: Recurrence of random walk

Problem.

- $S_0 = 0, S_n = \sum_{i=1}^n X_i$, where $X_i$ are i.i.d. $Be(p)$ r.v.s. within $\{-1, 1\}$.
- Define $T_0 := \min\{i \geq 1 : S_i = 0\}$. Notice that $T_0$ technically can be $\infty$, therefore it's a defective r.v. if $\mathbb{P}(T_0 = \infty) > 0$;

Qes. The transience $\mathbb{P}(T_0 = \infty)$; if none, the $\mathbb{E}T_0$.

Sol.

1. $f_{T_0}(n) := \mathbb{P}(S_1 \dots S_{n-1} \neq 0; S_n = 0)$, then $\mathbb{P}(T_0 = \infty) = 1 - G_{T_0}(1)$.
2. Consider $p_0(n) = \mathbb{P}(S_n = 0) = \mathbb{1}(\text{n is even})\binom{n}{n/2}(pq)^{n/2}$, then $G_{p_0}(s) = (1 - 4pqs^2)^{-1/2}$.
3. By law of total probability, that first return happens at the k-th step, then

$$\begin{aligned} p_0(n) &= \sum_{k=1}^n p_0(n-k) f_{T_0}(k) \\ \implies G_{p_0}(s) &= 1 + G_{p_0}(s)G_{T_0}(s) \end{aligned}$$

4. $G_{T_0}(s) = 1 - (1 - 4pqs^2)^{1/2}$, following (2) and (3).
5. $P(T_0 = \infty) = |p - q|$. If $p = q = \frac{1}{2}$, then transience is none. In such case, $G_{T_0}(s) = 1 - (1 - s^2)^{1/2}$, and $\mathbb{E}T_0 = G'_{T_0}(1) = +\infty$.

## 5.1.2 Application 2: Branching process (Galton-Watson tree)

Problem.

- Each member of the n-th generation gives birth to a family of members of the (n+1)-th generation.
- $Z_0 = 1, Z_{n+1} = X_1^{(n)} + X_2^{(n)} + \cdots + X_{Z_n}^{(n)}$, where the branching r.v. X are i.i.d.
- Once $Z_n = 0$, then $Z_{n+1} = 0$.
- $X_i^{(m)}$ is identical to $Z_1 := Z$.

Qes. What's the expectation and variance of $Z_n$?

Sol.

1. Denote $\mu := \mathbb{E}Z = \mu, \sigma^2 = VarZ$.
2. p.g.f. $G_n = G_1(G_{n-1}) = G(G_{n-1})$ by Thm 5.1.1.
3. $\mathbb{E}Z_n = \frac{\mathrm{d}}{\mathrm{d}s} G(G_{n-1}(s))|_{s=1} = \mu \mathbb{E}Z_{n-1}$, i.e. $\mathbb{E}Z_n = \mu^n$.
4. $VarZ_n = \frac{\mathrm{d}^2}{\mathrm{d}s^2} G(G_{n-1}(s))|_{s=1} + \frac{\mathrm{d}}{\mathrm{d}s} G_n(s)|_{s=1} - (\frac{\mathrm{d}}{\mathrm{d}s} G_n(s)|_{s=1})^2$.
5. If $\mu = 1, VarZ_n = n\sigma^2$.
6. If $\mu \neq 1, VarZ_n = \frac{\sigma^2(\mu^n - 1)\mu^{n-1}}{\mu - 1}$.

Qes. Does the process eventually extinct?

Sol.

1. $\{\text{ultimate extinction}\} = \bigcup_n \{Z_n = 0\} = \lim_{n\to\infty} \{Z_n = 0\}$, as an increasing sequence. Therefore $\mathbb{P}(\text{extinction}) = \lim_{n\to\infty} \mathbb{P}(Z_n = 0) = \lim_{n\to\infty} G_n(0)$. Let $\eta_n := G_n(0) \to \eta := \mathbb{P}(\text{extinction})$, then $\mu_n = G(\mu_{n-1})$.
2. Take limit, then $\eta = G(\eta)$ by continuity. Claim that $\eta$ is the smallest non-negative number that makes $s = G(s)$. Proof can be done by induction.
3. $G(s)$ is convex on $[0, 1]$ because $G''(s) := \mathbb{E}(Z(Z-1)s^{Z-2}) > 0$ if $s \geq 0$.
4. When $\mu = 1$, then $\eta = 1$, i.e. must extinct given $\sigma > 0$.
5. When $\mu < 1$, then $\eta = 1$, i.e. must extinct.
6. When $\mu > 1$, then $\eta < 1$.

# 5.2 Moment generating function

Limitation of p.g.f is that it only applys to nonnegative interger-valued discrete r.v.. We ask for a unified framework.

> Notation. (**Expectation for general r.v.**) $\mathbb{E}g(X) = \int g(x)\, \mathrm{d}F_X(x) = \int g(x) \mathbb{P} \circ X^{-1}(\mathrm{d}x)$.

E.g. $g(x)$ is the Dirichlet function.

Motivation. (**Lebesgue integral**) We ask for intervals such that function value in this interval is quite close. But we don't have to decompose the domain into intervals according

to the natural ordering. Instead, we can decompose according to the function value using measure theory.

> Def. (**Moment generating function**) $M_X(t) = \mathbb{E}e^{tX}$.

Rmk. $\frac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(t)|_{t=0} = \int x^k \, \mathrm{d}F_X(x)$. When k is even, we can directly claim existence.

# 5.3 Characteristic function

Motivation. We need to keep x on the exponential. But convergence of m.g.f. easily requires X to decay fast. The only left choice is to make the exponential complex. By Euler's formula, $e^{iy} = \cos y + i \sin y \implies \mathbb{E}e^{itx} = \mathbb{E}(\cos tX) + \mathbb{E}(\sin tX)$, therefore bounded. Essentially it becomes Fourier transform.

> Def. (**Characteristic function**) $\phi_X(t) = \mathbb{E}e^{itX} = \int e^{itx} \, \mathrm{d}F = \int (\cos tx + i \sin tx) \, \mathrm{d}F$.
> (**Cumulants generating function**) $\log \phi_X(t) = \sum_j \frac{i^j c_j t^j}{j!}$.

Thm. (Bochner's thm) A function is a characteristic func of some r.v. iff (1), (2), (3) hold.

1. $\phi(0) = 1, |\phi(t)| = |\int e^{itx} \, \mathrm{d}F| \le \int |e^{itx}| \, \mathrm{d}F \le \int \, \mathrm{d}F = 1$.
2. $\phi(t)$ is uniformly continuous. $|\phi(t+h) - \phi(t)| \le \mathbb{E}|e^{itX}(e^{ihX} - 1)| \le \mathbb{E}|e^{ihX} - 1|$.
   Proved by dominated convergence thm.
3. $\phi$ is non-negative definite.

Thm. (c.f. can generate moments) $\phi^{(k)}(t) = \int (ix)^k e^{itx} \, \mathrm{d}F$.

1. If $\phi^{(k)}(0)$ exists, which is $i^k \mathbb{E}X^k$, then $\mathbb{E}X^{\lfloor \frac{k}{2} \rfloor * 2}$ exists.
2. If $\mathbb{E}X^k$ exists, then $\phi^{(k)}(0)$ exists. Then by Taylor expansion,
   $\phi(t) = \sum_{j=0}^k \phi^{(j)}(0) \frac{t^j}{j!} + o(t^k) = \sum_{j=0}^k \mathbb{E}X^j \frac{(it)^j}{j!} + o(t^k)$.

Thm. $\forall s, t, \phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t) \iff X \perp Y$.

E.g.

- $X \sim Be(p), \phi(t) = q + pe^{it}$.
- $X \sim Exp(\lambda), \phi(t) = \frac{\lambda}{\lambda - it}$.
- $X \sim Cauchy, \phi(t) = e^{-|t|}$.
- $X \sim N(\mu, \sigma^2), \phi(t) = \exp(i\mu t - \frac{1}{2}\sigma^2 t^2)$.

Prop. Normal distribution is the only one whose cumulants expansion has finitely many non-zero terms: $\log \phi_X(t) = i\mu t - \frac{1}{2}\sigma^2 t^2$.

## 5.3.1 Inversion thm

Thm. (**Fourier inversion thm**) At all points where f is differentiable, $f_X(x) = \frac{1}{2\pi}\int e^{-itx}\phi_X(t)\,\mathrm{d}t$. If the integral fails to converge absolutely, we interpret it as its principal value.

Thm. (**Inversion thm**) $\bar{F}(x) := \frac{1}{2}(F(x) + F(x-))$. Notice that there's a one-to-one correspondence between $F, \bar{F}$. Then,

$$\bar{F}(b) - \bar{F}(a) := \lim_{N\to\infty}\int_{-N}^{N}\frac{e^{-ita} - e^{-itb}}{2\pi it}\phi_X(t)\,\mathrm{d}t$$

## 5.3.2 Lévy's continuity theorem

Motiv. In certain sense, the function that maps probability measure on $\mathbb{R}$ to its c.f. (Fourier transform) is continuous and has a continuous inverse. Hence the convergence of probability measures can be checked with the aid of the c.f.

> Def. (**Convergence of distribution function sequence**) $F_{X_n} \to F_X$ if $F_{X_n}(x) \to F_X(x)$ at every point x where $F_X$ is continuous (*weak convergence*). We say $X_n$ converges to $X$ in distribution/weakly/in law (denoted as $X_n \xrightarrow{D} X$ or $X_n \Rightarrow X$).

Motiv. We expect that $X_n(\omega) = \frac{1}{n}$ and $Y_n(\omega) = \frac{-1}{n}$ have the same limit. Therefore we drop the requirement at discontinuous points.

> Def. (**Vague convergence**) Given a set of d.f. $F_n$, if $F_n(x) \to G(x)$ at all continuous point of G, then $F_n$ converges to $G$ vaguely (because $G$ may be not a d.f.).

E.g. $X_n(\omega) \in \{\frac{1}{n}, n\}$, G is not a d.f.

Rmk. Convergence in distribution is equivalent to convergence with respect to the Lévy metric.

Thm. (**Lévy's continuity theorem**)

1. If $F_n \to F$ vaguely and $F$ is a d.f. with c.f. $\phi$ , then $\phi_n \to \phi$ pointwise.
2. If $\phi_n \to \phi$ pointwise and the c.f. $\phi$ is continuous at 0 with d.f. $F$, then $F_n \to F$.

Rmk. the statement "$\phi$ is continuous at 0" can be replaced by

1. $\phi$ is continuous (as a pointwise limit of c.f., 0 is the only point that can become discontinuous).
2. $\phi$ is a c.f. with d.f. $F$.
3. $\{F_n\}_{n=0}^{\infty}$ is tight, i.e. the tail probability is uniformly bounded, $\forall\epsilon > 0, \exists M_\epsilon > 0, s.t. \sup_n[F_n(-M_\epsilon) + (1 - F_n(M_\epsilon))] \le \epsilon$.

E.g. $X_n \sim N(0, n^2)$, $\phi_n(t) = \exp(-\frac{1}{2}n^2t^2)$.

## 5.3.3 Two limiting thm

### Thm.

n i.i.d. r.v. $X_n$ with $\mu := \mathbb{E}X < \infty$: (Abuse of notation here)

1. (**Weak law of large number**, WLLN) $\frac{1}{n}\sum_{i=1}^n X_i \overset{D}{\to} \mu$.
2. (**Central limit thm**, CLT) Suppose $VarX_1 = \sigma$, then $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right) \overset{D}{\to} N(0, \sigma^2)$.
3. In other word, $\frac{1}{n}S_n := \frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n})$.

### Proof: characteristics function

$$\phi_{X_1}(t) := \mathbb{E}e^{itX_1} = [k \leq 1] \sum_{j=0}^k \mathbb{E}X_1^j \frac{(it)^j}{j!} + o(t^k) = 1 + i\mu t + o(t)$$

$$\begin{aligned}\phi_n(t) := \mathbb{E}e^{it\frac{1}{n}S_1} = (\mathbb{E}e^{it\frac{1}{n}X_1})^n = \left(\phi_{X_1}\left(\frac{t}{n}\right)\right)^n \\ = \left(1 + i\mu\frac{t}{n} + o(\frac{t}{n})\right)^n \to e^{it\mu} = \phi_\mu(t), \forall t\end{aligned}$$

$\phi_\mu$ is continuous at 0, by Lévy's continuity theorem, $\frac{1}{n}\sum_{i=1}^n X_i \overset{D}{\to} \mu$. The same trick for Central limit thm.

### Alternative proof of CLT: moment method

Say $\mu = 0, \sigma = 1, \mathbb{E}|X_i|^k < \infty$ for simplicity, then it becomes showing
$Y_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \overset{D}{\to} N(0,1) \sim Y$.

Thm. (**Portmanteau thm**) Let $X_1, X_2, \ldots$ be r.v.s. with d.f. $F_1, F_2, \ldots$, then $X_n \overset{D}{\to} X$ iff $\forall$bounded continuous func $g : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}g(X_n) := \int g(x)\,\mathrm{d}F_n(x) \to \int g(x)\,\mathrm{d}F_X(x) := \mathbb{E}g(X)$ as $n \to \infty$.

Rmk. It's sufficient to only check for all Lipschitz continuious functions.

Thm. (**Carleman's thm**) It's sufficient to only check for $\forall k \in \mathbb{N}, g_k(x) := x^k$ if given X satisfies $\sum_{k=1}^\infty (\int g_{2k}(x)\,\mathrm{d}F_X(x))^{-\frac{1}{2k}} = \infty$.

Alternative proof for CLT: It's suffices to show $\forall k \in \mathbb{N}, \mathbb{E}Y_n^{2k} \to (2k-1)!!, \mathbb{E}Y_n^{2k+1} \to 0$.

$$\mathbb{E}Y_n^p = n^{-\frac{p}{2}} \sum_{S \in 2^{[n]}} \mathbb{E}\left(\prod_{j=1}^p X_{S_j}\right)$$

Due to independence and $\mathbb{E}X_j = 0$, this become a combinatorics problem. Some terms are negligible. Details are skipped.

### Alternative proof of CLT: comparison method(Lindeberg swapping)

Say $\mu = 0, \sigma = 1, \mathbb{E}|X_i|^3 \le C_3 < \infty$ for simplicity, then it becomes showing
$Y_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{D} N(0,1)$.

Construct $g_i \sim N(0,1)$ as i.i.d., and let $Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i \sim N(0,1)$. It suffices to show
$\forall t, F_{Y_n}(t) - F_{Z_n}(t) \to 0$ as $n \to \infty$.

$$
\begin{aligned}
&\forall t, F_{Y_n}(t) - F_{Z_n}(t) \to 0 \text{ as } n \to \infty \\
\Leftarrow &\forall t, \mathbb{E}\mathbb{1}(Y_n \le t) - \mathbb{E}\mathbb{1}(Z_n \le t) \to 0 \text{ as } n \to \infty \\
\Leftarrow &\forall t, \phi_t(\cdot) := \mathbb{1}(\cdot \le t), \mathbb{E}\phi_t(Y_n) - \phi_t(Z_n) \to 0 \text{ as } n \to \infty \\
\Leftarrow &\forall \text{bounded smooth func with bounded derivative } \phi, \\
&\mathbb{E}\phi(Y_n) - \phi(Z_n) \to 0 \text{ as } n \to \infty
\end{aligned}
$$

Construct $H_{n,i} := n^{-\frac{1}{2}}(g_1 + g_2 \dots g_i + X_{i+1} + \dots + X_n)$, then $Y_n = H_{n,0}, Z_n = H_{n,n}$. Through this construction, we can bounded the difference between the two summations term by term.

$$
\begin{aligned}
&\mathbb{E}\phi(Y_n) - \phi(Z_n) \\
&= \mathbb{E}\phi(H_{n,0}) - \mathbb{E}\phi(H_{n,n}) \\
&= \sum_{i=0}^{n-1} \mathbb{E}\phi(H_{n,i}) - \mathbb{E}\phi(H_{n,i+1}) \in o(1) \\
&\Leftarrow \forall n, i, \mathbb{E}\phi(H_{n,i}) - \mathbb{E}\phi(H_{n,i+1}) \in o(\frac{1}{n})
\end{aligned}
$$

Notice that the consecutive two terms $\phi(H_{n,i}), \phi(H_{n,i+1})$ only differ in the $i+1$-th entry. We can estimate by Tarlor expansion (with Lagrange remainder) w.r.t. this entry.

$$
\begin{aligned}
&H^\circ_{n,i} := n^{-\frac{1}{2}}(g_1 + g_2 \dots g_i + 0 + X_{i+2} + \dots + X_n) \\
&\phi(H_{n,i}) = \phi(H^\circ_{n,i}) + \phi'(H^\circ_{n,i})(\frac{X_{i+1}}{\sqrt{n}}) + \frac{1}{2}\phi''(H^\circ_{n,i})(\frac{X_{i+1}}{\sqrt{n}})^2 + \frac{1}{3!}\phi'''(t_1)(\frac{X_{i+1}}{\sqrt{n}})^3 \\
&\phi(H_{n,i+1}) = \phi(H^\circ_{n,i}) + \phi'(H^\circ_{n,i})(\frac{g_{i+1}}{\sqrt{n}}) + \frac{1}{2}\phi''(H^\circ_{n,i})(\frac{g_{i+1}}{\sqrt{n}})^2 + \frac{1}{3!}\phi'''(t_2)(\frac{g_{i+1}}{\sqrt{n}})^3 \\
&\mathbb{E}\phi(H_{n,i}) - \mathbb{E}\phi(H_{n,i+1}) \le C_3 \sup_{x \in \mathbb{R}} |\phi'''(x)| n^{-\frac{3}{2}} \in O\left(n^{-\frac{3}{2}}\right) \in o(\frac{1}{n})
\end{aligned}
$$

## Alternative proof of CLT: Stein's method

A powerful method to show weak convergence to certain dist, especially Gaussian.

Prop. $X \sim N(0,1)$ iff $\forall f : \mathbb{R} \to \mathbb{R}$ continuously differentiable with bounded $f, f'$,
$\mathbb{E}[f'(X) - Xf(X)] = 0$.

Thm. (**Stein continuity thm**) $X_n$ is a sequence of real. r.v.s. with $\mathbb{E}|X_i|^2 \le C_2 < \infty$, then
$X_n \xrightarrow{D} N(0,1)$ iff $\forall f : \mathbb{R} \to \mathbb{R}$ continuously differentiable with bounded $f, f'$,
$\mathbb{E}[f'(X_n) - X_n f(X_n)] \to 0$ as $n \to \infty$.

Again, define $Y_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$. We assume $\mathbb{E}|Y_i|^2 \le C_2 < \infty$ required by the thm. Then, to show the convergence between $f$ and $f'$, we estimate by Tarlor expansion (with Lagrange

remainder) again:

$$
\begin{aligned}
Y_n^{(j)} :=& \frac{1}{\sqrt{n}} \sum_{j \neq i} X_i \\
& \mathbb{E} Y_n f(Y_n) \\
=& \mathbb{E} \frac{1}{\sqrt{n}} \sum_j X_j f(Y_n) \\
=& \frac{1}{\sqrt{n}} \sum_j \mathbb{E} X_j (\sum_{k=0}^{\infty} f^{(k)}(Y_n^{(j)}) (\frac{X_j}{\sqrt{n}})^k) \\
\sim& \sum_j \sum_{k=0}^{\infty} n^{-\frac{k+1}{2}} \mathbb{E} f^{(k)}(Y_n^{(j)}) X_j^{k+1} \\
=& \sum_j \sum_{k=0}^{\infty} n^{-\frac{k+1}{2}} \mathbb{E} f^{(k)}(Y_n^{(j)}) \mathbb{E} X_j^{k+1}, \text{due the independence} \\
=& \sum_j \sum_{\text{odd } k} n^{-\frac{k+1}{2}} \mathbb{E} f^{(k)}(Y_n^{(j)}) \mathbb{E} X_j^{k+1}, \text{ because odd order moment are } 0 \\
=& \sum_j (n^{-1} \mathbb{E} f'(Y_n^{(j)}) + O(n^{-3/2})) \\
=& O(n^{-1/2}) + n^{-1} (\sum_j \mathbb{E} f'(Y_n^{(j)}) - \mathbb{E} f'(Y_n) + \sum_j \mathbb{E} f'(Y_n)) \\
=& O(n^{-1/2}) + \mathbb{E} f'(Y_n), \text{due to bounded } f'
\end{aligned}
$$

### Thm: A general CLT

Motiv. When every error is so minor that doesn't dominate the total error.

Thm. Let $X_1, X_2 \dots$ be independent; define $\sigma_n^2 = Var(S_n) = \sum_{j=1}^n \sigma_j^2$ with $\mathbb{E} X_j = 0, Var X_j = \sigma_j^2, \mathbb{E}|X_j^3| < \infty$, and such that

$$
\frac{1}{\sigma_n^3} \sum_{j=1}^n \mathbb{E}|X_j^3| \to 0 \text{ as } n \to \infty
$$

Then $\frac{1}{\sigma_n} S_n \xrightarrow{D} N(0,1)$.

# Ch6 Markov

Skipped.

# Ch7 Convergence of r.v.

Convergence in distribution only reflects one aspect of r.v. It's reasonable to consider the nature of r.v.: a function from $[0,1]$ to $\mathbb{R}$.

Recall. Convergence of real functions $f_n, f : [0,1] \to \mathbb{R}$,

1. (**Pointwise convergence**) $\forall x, f_n(x) \to f(x)$ as $n \to \infty$.

2. (**Convergence in norm**) $||f_n - f|| \to 0$ as $n \to \infty$. (bad point is not too bad in norm)
3. (**Convergence in measure** $\mu$) Bad point is not too much.
    - For $\epsilon > 0, E_\epsilon = \{u \in [0,1] : |g(u) - h(u)| > \epsilon\}, d_\epsilon(g,h) = \mu(E_\epsilon) := \int_{E_\epsilon} \mathrm{d}\mu$.
    - $d_\epsilon(f_n, f) \to 0$ as $n \to \infty$ for all $\epsilon > 0$.
    - Notice that proving on $\exists \delta > 0, \forall \epsilon \in (0, \delta)$ will be sufficient.

Rmk.

1. Convergence in $L_p$ norm is stronger than convergence in measure.
2. Convergence pointwise is stronger than convergence in measure.
3. The pointwise convergence and convergence in $L_1$ norm are not comparable.

---

# 7.1 Convergence mode of r.v.s.

We have introduced convergence in distribution, which is the only one that allows r.v.s. living in different probability spaces. Now let's go into the convergence of r.v.s.

> Def. (**Almost sure convergence** $X_n \overset{a.s.}{\to} X$, $X_n \overset{a.e.}{\to} X$, or $X_n \to X$ w.p. 1) $X_n$ converges to $X$ almost surely, or almost everywhere, or with probability 1, when $\mathbb{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}) = 1$.

> Def. (**Convergence in r-th mean** $X_n \overset{r}{\to} X$) $\mathbb{E}|X_n - X|^r \to 0$ as $n \to \infty$. Note that $\mathbb{E}|Y|^r = \int_\Omega |y|^r \, \mathrm{d}F_Y$, and $(\mathbb{E}|Y|^r)^{\frac{1}{r}}$ is $L_r$ norm. (Care about the degree of badness of bad events)

> Def. (**Convergence in probability (measure)** $X_n \overset{\mathbb{P}}{\to} X$) $\exists \delta > 0, \forall \epsilon \in (0, \delta)$, $\mathbb{P}(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \epsilon\}) \to 0$ as $n \to \infty$. (Care about the size of bad events)

Rmk. Almost sure convergence is an adoption of pointwise convergence, where the requirement of convergence over zero-probability events is drop.

Rmk. Convergence in probability is essentially using $d_\epsilon(g,h) = \int_{\mathbb{E}} \mathrm{d}\mathbb{P}(x)$ in the convergence in measure.

### Thm. 7.1.1

Let $A_n(\epsilon) := \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}, B_m(\epsilon) := \bigcup_{n \geq m} A_n(\epsilon)$. Here $\{B_m\}$ is decreasing, and therefore we denote by $A(\epsilon) := \lim_{m \to \infty} B_m(\epsilon) = \bigcap_{m=0}^{\infty} \bigcup_{n \geq m} A_n(\epsilon) = \limsup_n A_n(\epsilon)$, and denoted by $A(\epsilon) := \{A_n(\epsilon) \text{ i.o.}\}$, standing for infinitely often. Then the following hold:

1. $X_n \overset{a.s.}{\to} X \iff \forall \epsilon > 0, \mathbb{P}(A(\epsilon)) = 0$.
2. $X_n \overset{a.s.}{\to} X \Leftarrow \forall \epsilon > 0, \sum_{n=1}^{\infty} \mathbb{P}(A_n(\epsilon)) < \infty$.

## Proof.

Proof of 1. Notice that $A(\epsilon)$ is decreasing as $\epsilon$ getting larger.

$$
\begin{aligned}
& \mathbb{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}) = 1 \\
\iff & \mathbb{P}(\{\omega \in \Omega : \forall \epsilon > 0, \exists n_0, \forall n > n_{0,} |X_n(\omega) - X(\omega)| \leq \epsilon\}) = 1 \\
\iff & \mathbb{P}(\{\omega \in \Omega : \exists \epsilon_0 > 0, \forall n_0, \exists n > n_{0,} |X_n(\omega) - X(\omega)| > \epsilon_0\}) = 0 \\
\iff & \mathbb{P}(\cup_{\epsilon > 0}\{\omega \in \Omega : \forall n_0, \exists n > n_{0,} |X_n(\omega) - X(\omega)| > \epsilon\}) = 0 \\
\iff & \mathbb{P}(\cup_{\epsilon > 0}\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon \text{ for infinitely many n}\}) = 0 \\
\iff & \mathbb{P}(\cup_{\epsilon > 0} \limsup_n A_n(\epsilon)) = 0
\end{aligned}
$$

From left to right: notice that $\forall \epsilon > 0, \mathbb{P}(\limsup_n A_n(\epsilon)) \leq \mathbb{P}(\cup_{\epsilon > 0} \limsup_n A_n(\epsilon)) = 0$. From right to left is obvious too.

Proof of 2. $\mathbb{P}(B_m(\epsilon)) \leq \sum_{n=m}^{\infty} \mathbb{P}(A_n(\epsilon))$, and the latter goes to 0 as m go to infinity.

## Thm. 7.1.2

1. $X_n \overset{a.s.}{\to} X$ implies $X_n \overset{\mathbb{P}}{\to} X$.
2. $X_n \overset{r}{\to} X$ implies $X_n \overset{\mathbb{P}}{\to} X$.
3. $X_n \overset{\mathbb{P}}{\to} X$ implies $X_n \overset{D}{\to} X$.
4. $X_n \overset{r}{\to} X$ implies $X_n \overset{s}{\to} X$, if $r \geq s \geq 1$.
5. No other implication hold in general.

## Proof of 5 (No other implication hold in general)

($X_n \overset{D}{\to} X$ doesn't imply $X_n \overset{\mathbb{P}}{\to} X$) Trivial.

($X_n \overset{\mathbb{P}}{\to} X$ doesn't imply $X_n \overset{r}{\to} X$) Tiny bad point with very bad value. E.g. pick $X_n := n^3$ with proba $n^{-2}$, otherwise 0.

($X_n \overset{\text{s}}{\to} X$ doesn't imply $X_n \overset{r}{\to} X$, if $r > s \geq 1$) We want small s-th moment with large r-th moment. E.g. pick $X_n := n$ with proba $n^{-\frac{1}{2}(r+s)}$, otherwise 0.

($X_n \overset{\mathbb{P}}{\to} X$ doesn't imply $X_n \overset{a.s.}{\to} X$):

- $I_{i,j} := [\frac{j}{i}, \frac{j+1}{i}], I_n := ordering_n(I_{i,j}) = (I_{1,0}, I_{2,0}, I_{2,1}, I_{3,0}, \dots)_n$.
- $X_n := \mathbb{1}_{I_n}, X := 0$.
- Take $\delta = 1, \mathbb{P}(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \epsilon\}) = \mathbb{P}(I_n) \to 0$.
- For no $\omega$, $X_n(\omega) \to X(\omega)$ as $n \to \infty$, i.e. $\mathbb{P} = 0$ instead.
- Intuition: a.s. convergence want to make sure many $\omega$ that can run away from bad events in a limiting sence.

($X_n \overset{r}{\to} X$ doesn't imply $X_n \overset{a.s.}{\to} X$) $X_n = \begin{cases} 1, w.p. \, n^{-1} \\ 0, w.p. \, 1 - n^{-1} \end{cases}$

($X_n \xrightarrow{a.s.} X$ doesn't imply $X_n \xrightarrow{r} X$) $X_n = \begin{cases} n^3, w.p. \, n^{-2} \\ 0, w.p. \, 1 - n^{-2} \end{cases}$

### Proof of 3

($X_n \xrightarrow{\mathbb{P}} X$ implies $X_n \xrightarrow{D} X$)

$$\forall \epsilon > 0, \mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X \leq x + \epsilon) + \mathbb{P}(X_n \leq x, X > x + \epsilon)$$

$$\leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon)$$

$$\mathbb{P}(X \leq y) = \mathbb{P}(X \leq y, X_n \leq y + \epsilon) + \mathbb{P}(X \leq y, X_n > y + \epsilon)$$

$$\leq \mathbb{P}(X_n \leq y + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon)$$

$$\begin{gathered} \text{Let } y = x - \epsilon, \\ \mathbb{P}(X \leq x - \epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \leq \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \\ F_X(x - \epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \leq F_n(x) \leq F_X(x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \\ n \to \infty, F_X(x - \epsilon) \leq \liminf_n F_n(x) \leq \limsup_n F_n(x) \leq F_X(x + \epsilon) \end{gathered}$$

When $F_x$ is continuous at x, let $\epsilon \to 0$, we get $F_n(x) \to F_X(x)$ and concludes the proof.

### Proof of 2

($X_n \xrightarrow{1} X$ implies $X_n \xrightarrow{\mathbb{P}} X$) Directly result from Markov's inequality.

$$\begin{gathered} \text{Markov's inequality } X \geq 0, \forall a > 0, P(X \geq a) \leq E(X)/a \\ \text{Chebyshev's inequality } \forall k > 0, P(|X - \mu| \geq k\sigma) \leq 1/k^2 \end{gathered}$$

### Proof of 4

($X_n \xrightarrow{r} X$ implies $X_n \xrightarrow{s} X$)

$$\begin{gathered} \text{Lyapunov's inequality } (\mathbb{E}|Z|^r)^{\frac{1}{r}} \geq (\mathbb{E}|Z|^s)^{\frac{1}{s}}, r \geq s > 0 \\ \text{Hölder's inequality } ||fg||_1 \leq ||f||_p ||g||_q, \frac{1}{p} + \frac{1}{q} = 1 \\ \text{Jensen's inequality } \phi(\mathbb{E}X) \leq \mathbb{E}(\phi(X)), \phi \text{ is convex} \end{gathered}$$

In Hölder, let $X := |Z|^s, Y = 1, p = \frac{r}{s}, q = \frac{r}{r-s}$. In Jensen, let $X := |Z|^s, \phi(x) = x^{\frac{r}{s}}$.

### Proof of 1

($X_n \xrightarrow{a.s.} X$ implies $X_n \xrightarrow{\mathbb{P}} X$) $X_n \xrightarrow{\mathbb{P}} X \iff \exists \delta > 0, \forall 0 < \epsilon < \delta, \mathbb{P}(A_n(\epsilon)) \to 0$ is true whenever $\mathbb{P}(B_n(\epsilon)) \to 0$ is true, which holds iff $X_n \xrightarrow{a.s.} X$ by Thm 7.1.1(1).

# 7.2 Partial converse statements

### Thm.

1. If $X_n \xrightarrow{D} c$, where c is a constant, then $X_n \xrightarrow{\mathbb{P}} c$.
2. If $X_n \xrightarrow{\mathbb{P}} X$, $\mathbb{P}(|X_n| \leq k) = 1$ for all n and some constant k, then $X_n \xrightarrow{r} X, r \geq 1$.
3. If $X_n \xrightarrow{\mathbb{P}} X$, there's a non-random subsequence such that $X_{n_i} \xrightarrow{a.s.} X$ as $i \to \infty$.
4. If $\forall \epsilon > 0, \sum_n \mathbb{P}(|X_n - X| > \epsilon) < \infty$, then $X_n \xrightarrow{a.s.} X$.

### Proof of 1

(If $X_n \xrightarrow{D} c$, where c is a constant, then $X_n \xrightarrow{\mathbb{P}} c$) Intuition from counter example: because no way to switch the $\omega$ if only one possible choice, i.e. $c$.

Application. to prove the law of large number, no need to prove the convergence in probability anymore.

### Proof of 2

(If $X_n \xrightarrow{\mathbb{P}} X$, $\mathbb{P}(|X_n| \leq k) = 1$ for all n and some constant k, then $X_n \xrightarrow{r} X, r \geq 1$)

### Proof of 3

#TODO

### Proof of 4

#TODO

## 7.3 More applications of WLLN

### Thm. (Continuous mapping theorm)

Suppose function $g$ has the discontinuity points $D_g$ such that $\mathbb{P}(X \in D_g) = 0$, then:

$$
\begin{aligned}
X_n \xrightarrow{D} X &\implies g(X_n) \xrightarrow{D} g(X) \\
X_n \xrightarrow{\mathbb{P}} X &\implies g(X_n) \xrightarrow{\mathbb{P}} g(X) \\
X_n \xrightarrow{a.s.} X &\implies g(X_n) \xrightarrow{a.s.} g(X)
\end{aligned}
$$

### Application. 1 (Bernstein approximation)

Let $f$ be continuous on $[0, 1]$ (i.e. uniform continuous and bounded). Define the Bernstein polynomial of degree n:

$$
f_n(x) = \sum_{m=0}^{n} \binom{n}{m} x^m (1-x)^{n-m} f(\frac{m}{n})
$$

Then $\sup_{x \in [0,1]} |f_n(x) - f(x)| \to 0$ as $n \to \infty$ (uniform convergence).

Proof.

1. For $x \in [0,1]$, define $X_{x,i} \sim Be(x), S_{x,n} := \sum_{i=1}^n X_{x,i} \sim Bin(n,x)$, then $f_n(x) = \mathbb{E} f(\frac{S_{x,n}}{n})$. Now we want to show $\mathbb{E} f\left(\frac{S_{x,n}}{n}\right) \rightrightarrows f(x)$.
2. By WLLN and partial converse statement 1, $\frac{S_{x,n}}{n} = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\mathbb{P}} x$. By continuous mapping thm, we can get $f\left(\frac{S_{x,n}}{n}\right) \xrightarrow{\mathbb{P}} f(x)$. But this's not used because such convergence depends on $x$.
3. We look back at the uniform continuity.
   $\forall \epsilon > 0, \exists \delta := \delta(\epsilon) > 0, \forall x, n, |\frac{S_{x,n}}{n} - x| < \delta \implies |f(\frac{S_{x,n}}{n}) - f(x)| < \epsilon$.
4. Note that $\mathbb{E}\frac{S_{x,n}}{n} = x, Var(\frac{S_{x,n}}{n}) = x(1-x)/n$, then:

$$
\begin{aligned}
&\forall \epsilon > 0, \exists \delta := \delta(\epsilon) > 0, \forall x, n, \\
&\left|\mathbb{E}\left[f\left(\frac{S_{x,n}}{n}\right)\right] - f(x)\right| \\
\le &\mathbb{E}\left|f\left(\frac{S_{x,n}}{n}\right) - f(x)\right| \\
= &\mathbb{E}\left|f\left(\frac{S_{x,n}}{n}\right) - f(x)\right| \mathbb{1}(|\frac{S_{x,n}}{n} - x| < \delta) \\
&+ \mathbb{E}\left|f\left(\frac{S_{x,n}}{n}\right) - f(x)\right| \mathbb{1}(|\frac{S_{x,n}}{n} - x| \ge \delta) \\
\le &\epsilon \mathbb{E}\mathbb{1}(|\frac{S_{x,n}}{n} - x| < \delta) + 2\sup_x |f| \cdot \mathbb{E}\mathbb{1}(|\frac{S_{x,n}}{n} - x| \ge \delta) \\
= &\epsilon \mathbb{P}(|\frac{S_{x,n}}{n} - x| < \delta) + 2\sup_x |f| \cdot \mathbb{P}(|\frac{S_{x,n}}{n} - x| \ge \delta) \\
\le &\epsilon + 2\sup_x |f| \cdot \frac{Var(\frac{S_{x,n}}{n})}{\delta^2}, \text{by Chebyshev ineq} \\
= &\epsilon + 2\sup_x |f| \cdot \frac{x(1-x)}{n\delta^2} \le \epsilon + \frac{\sup_x |f|}{2n\delta^2}
\end{aligned}
$$

5. $\forall \epsilon > 0, \exists \delta := \delta(\epsilon) > 0, \forall n, \sup_x |f_n(x) - f(x)| \le \epsilon + \frac{\sup_x |f|}{2n\delta^2}$.
6. $\forall \epsilon > 0, \exists \delta := \delta(\epsilon) > 0, \limsup_n \sup_x |f_n(x) - f(x)| \le \epsilon$.
7. $\lim_n \sup_x |f_n(x) - f(x)| \epsilon = 0$.

## Application 2. (Borel's geometric concentration)

Let $\mu_n$ be the uniform probability measure on the n-dimentional cube $[-1,1]^n$. Let $H$ be a hyperplane that's orthogonal to a principal diagonal of $[-1,1]^n$, i.e. $H = (1,1,\dots,1)^\perp$. Let $H_r = \{x \in [-1,1]^n : dist(x,H) \le r\}$. Then $\forall \epsilon > 0, \lim_{n\to\infty} \mu_n(H_{\epsilon\sqrt{n}}) = 1$.

Proof.

$$
\begin{aligned}
&\mu_n(H_{\epsilon\sqrt{n}}) = \mathbb{P}(dist(\vec{x}, H) \le \epsilon\sqrt{n}) \\
= &\mathbb{P}\left(\frac{|<\vec{x}, (1,1,\dots,1)>|}{\sqrt{n}} \le \epsilon\sqrt{n}\right) = \mathbb{P}(\frac{|\sum x_i|}{n} \le \epsilon)
\end{aligned}
$$

#TODO

# 7.4 Other versions of WLLN

### Thm. $L^2 - WLLN$

Let $X_1, X_2, \dots$ be **uncorrelated** r.v.s. with $\forall i, \mathbb{E}X_i = \mu, VarX_i \le C < \infty$ Then $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{2} \mu$.

Proof. By def, $\mathbb{E}(\frac{S_n}{n} - \mu)^2 = Var(\frac{S_n}{n}) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) \le \frac{C}{n} \to 0$ as $n \to \infty$.

### Thm. WLLN for triangular array

Thm. Let $\{X_{n,i}\}$ be a triangular array, where in each level, $X_{n,i}$ are i.i.d. with $VarX_i = \sigma^2$. Let $S_n := \sum_{i=1}^n X_{n,i}, \mu_n := \mathbb{E}S_n = n\mathbb{E}X_{n,1}$. Suppose $\frac{Var(S_n)}{b_n^2} \to 0$ for some real sequence chosen by ourselves (non-random) $\{b_n\}$, then $\frac{S_n - \mu_n}{b_n} \xrightarrow{2} 0$.

E.g. $X_{n,i} := \frac{Y_i}{n}$, then $S_n = \sum_{i=1}^n \frac{Y_i}{n}$.

Proof. $\mathbb{E}(\frac{S_n - \mu_n}{b_n} - 0)^2 = b_n^{-2} Var(S_n) \to 0$ as $n \to \infty$.

Motiv. We use $b_n = n$ in the L2-WLLN, and eventually $\frac{C}{n} \to 0$ very fast. But we only need an $o(1)$ for it.

Rmk. We can try an extreme case inspired from CLT, where $b_n = n^{1/2+\epsilon}$. Then $\frac{S_n - n\mathbb{E}X_{n,1}}{n^{1/2+\epsilon}} \xrightarrow{\mathbb{P}} 0 \Rightarrow S_n = n\mathbb{E}X_{n,1} + o_{\mathbb{P}}(n^{1/2+\epsilon})$. The principle is that if possible, we shall not choose $b_n >> \mu_n = n\mathbb{E}X_{n,1}$, otherwise meaningless.

### E.g.1 Coupon collection problem

$Y_{n,i}$: i.i.d. uniform on $\{1, 2, \dots, n\}$. Let $\psi_k^n := \inf_m\{|\{Y_{n,1}, Y_{n,2} \dots Y_{n,m}\}| = k\}$, $X_{n,k} := \psi_k^n - \psi_{k-1}^n \sim Geom(1 - \frac{k-1}{n})$ and $X_{n,k}$ are independent. #TODO By WLLN for triangular array, #TODO

### E.g.2 Occupacy problem

#TODO

# 7.5 Borel-Cantelli lemma

### Thm.

For a sequence of events $A_n$ from the same probability space, $A := \bigcap_{m=0}^{\infty}\bigcup_{n\ge m} A_n = \{A_n \text{ i.o.}\}$.

1. $\mathbb{P}(A) = 0$ if $\sum_n \mathbb{P}(A_n) < \infty$.
2. $\mathbb{P}(A) = 1$ if $\sum_n \mathbb{P}(A_n) = \infty$ and $A_i$ are mutually independent events.

Rmk. We can state the lemma as an example of "zero-one law": when $A_i$ are mutually independent events, $\mathbb{P}(A)$ equals either 0 or 1.

## Proof.

Proof of 1. $\forall m, A \subset \cup_{n=m}^{\infty} A_n$, therefore $\mathbb{P}(A) \leq \sum_{n=m}^{\infty} \mathbb{P}(A_n)$, and the latter goes to 0 as $n \to \infty$ (Cauchy criterion).

Proof of 2. To prove $\mathbb{P}(A) = 1$, we typically prove $\mathbb{P}(A^C) = 0$.

$$\begin{aligned}
\mathbb{P}(\bigcap_{n=m}^{\infty} A_n^C) &= \prod_{n=m}^{\infty} (1 - \mathbb{P}(A_n)), \text{by independence} \\
&\leq \prod_{n=m}^{\infty} \exp(-\mathbb{P}(A_n)), \text{by } 1 - x \leq e^{-x} \\
&= \exp(-\prod_{n=m}^{\infty} \mathbb{P}(A_n)) \to 0 \text{ as } m \to \infty
\end{aligned}$$

## Application. Infinite Monkey

Consider an infinite-length string produced from a finite alphabet by picking each letter independently at random, uniformly from the alphabet (say the alphabet has n letters). Fix a string S of length m from the same alphabet. Let $E_k$ be the event "the m-substring starting at position k is the string S". Then, infinitely many of the $E_k$ occur with probability 1.

Proof. Note that the events $\{E_{m \cdot j}\}_{j=0}^{\infty}$, a subsequence of $\{E_k\}$, are independent because disjoint, where $\sum_{j=1}^{\infty} \mathbb{P}(E_{mj}) = \sum_{j=1}^{\infty} (\frac{1}{n})^m = \infty$ and $\mathbb{P}(E_{mj} \text{ i.o.}) = 1$. Therefore $\mathbb{P}(E_k \text{ i.o.}) = 1$.

## Another zero-one law

Thm. For a sequence of events $A_n$ from the same probability space, let $\mathcal{A} := \sigma(A_1, A_2, \ldots)$. If $A \in \mathcal{A}$ and for any n, A is independent of the finite collection $A_1, A_2, \ldots, A_n$, then $P(A)$ is either 0 or 1.

Proof. $A \in \mathcal{A}$ means A is definable in terms of $A_1, A_2, \ldots$ From measure theory, that means $\exists \{C_n\}$ s.t. $C_n \in \mathcal{A}_n$ and $\mathbb{P}(A \triangle C_n) \to 0$ as $n \to \infty$, i.e. $\mathbb{P}(A \cap C_n) \to \mathbb{P}(A)$. On the other hand, A being independent of events in $\mathcal{A}_n$, $\mathbb{P}(A \cap C_n) = \mathbb{P}(A)\mathbb{P}(C_n) \to P(A)^2$, therefore $\mathbb{P}(A) = \mathbb{P}(A)^2$.

## Kolmogorov's zero-one law

Def. (**Tail $\sigma$-field**): Consider a sequence of r.v.s. $X_n$ in the same probability space. We define $\mathcal{H}_n := \sigma(X_n, X_{n+1}, \ldots)$ be the $\sigma$-field induced by r.v.s. defined above in Part 3.4. Then

$\mathcal{H}_n$ is decreasing ($\mathcal{H}_n \supset \mathcal{H}_{n+1} \supset \mathcal{H}_{n+2} \ldots$), as the number of r.v.s. is decreasing. $\mathcal{H}_\infty := \cap_n \mathcal{H}_n$ , called the tail $\sigma$-field.

Def. (**Tail events**): Elements in $\mathcal{H}_\infty$, which must be events that do not refer to any finite subcollections $\{X_1, X_2, \ldots, X_n\}$. For examples,
$\{X_n > 0, \text{ i.o.}\}, \{\limsup_n X_n = \infty\}, \{\sum_n X_n \text{ converges}\}$.

Thm. (**Kolmogorov's zero-one law**) If $\{X_n\}$ is a independent r.v. sequence, then $\forall H \in \mathcal{H}_\infty$, $\mathbb{P}(H)$ is either 0 or 1.

Def. (**Tail function**) A r.v. Y defined based on independent r.v. sequence $\{X_n\}$ that is $\mathcal{H}_\infty$-measurable.

E.g. $Y := \limsup_n X_n$ is a tail func, while $Y' := X_1 + X_2$ is not. By Kolmogorov's zero-one law, when $\{\omega : Y < y\} \in \mathcal{H}_\infty$, $\mathbb{P}(Y \leq y)$ is either 0 or 1, i.e., $Y$ is a constant r.v.

Appli. Let Y be a tail func, then $\exists k, -\infty \leq k \leq +\infty$, s.t., $\mathbb{P}(Y = k) = 1$.

# 7.6 Strong LLN

Motiv. In WLLN, the condition is i.i.d. sequence $\{X_n\}$ that $\mathbb{E}X_i = \mu, |\mu| < \infty$, and the result is $\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$. But what's the necessary condition?

Thm. For i.i.d. sequence $\{X_n\}$ and some constant $\mu$, $\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$ iff (1) or (2).

1. $\mathbb{P}(|X_1| > n) \in o(\frac{1}{n})$ and $\int_{[-n,n]} x \, \mathrm{d}F \to \mu$ as $n \to \infty$. The latter can also be written as $\mathbb{E}(X_1 \mathbb{1}(|X_1| \leq n))$.
2. The c.f. of $X_j$ is differentiable at $t = 0$, and $\phi'(0) = i\mu$.

Thm. (**Strong law of large number**) For i.i.d. sequence $\{X_n\}$,

1. $\frac{S_n}{n} \xrightarrow{a.s.} \mu$ iff $|\mathbb{E}X_i| < \infty$.
2. $\frac{S_n}{n} \xrightarrow{2} \mu$ iff $|\mathbb{E}X_i^2| < \infty$.

# 7.7 The law of the iterated logarithm

Skipped.