Probability Notes 2024

Ruan Yuanlong

BUAA, Beijing, China
Email address: `ruanyl@buaa.edu.cn`

Contents

1. 单调类定理

Review:

- $\mathscr{A}$ is a field, $\mathscr{M}$ is a monotone class. Then

$$\mathscr{A} \subset \mathscr{M} \Longrightarrow \sigma(\mathscr{A}) \subset \mathscr{M}.$$

- $\mathscr{P}$ is a π-system, $\mathscr{L}$ is a λ-system. Then

$$\mathscr{P} \subset \mathscr{L} \Longrightarrow \sigma(\mathscr{P}) \subset \mathscr{L}.$$

- measurable spaces $(E, \mathscr{F}_E), (F, \mathscr{F}_F)$, $f : (E, \mathscr{F}_E) \mapsto (F, \mathscr{F}_F)$. f is $\mathscr{F}_E/\mathscr{F}_F$-measurable if

$$\sigma(f) \triangleq f^{-1}(\mathscr{F}_F) \subset \mathscr{F}_E.$$

  Call it $\mathscr{F}_E$-measurable if

$$(F, \mathscr{F}_F) = (\mathbb{R}, \mathscr{B}(\mathbb{R})).$$

- $f : (E, \mathscr{F}_E) \mapsto (F, \sigma(\mathscr{E}))$, f is $\mathscr{F}_E/\sigma(\mathscr{E})$-measurable if

$$f^{-1}(\mathscr{E}) \subset \mathscr{F}_E.$$

Thm 1 (π-λ theorem). *$\mathscr{P}$ is a π-system, $\mathscr{L}$ is a λ-system. If $\mathscr{P} \subset \mathscr{L}$, then $\sigma(\mathscr{P}) \subset \mathscr{L}$.*

Def 1 (**Simple function**). *$i = 1, ..., n$, $A_i \in \mathscr{F}$ (pairwise) disjoint, $c_i \in \mathbb{R}$. f is (measurable) simple if $f = \sum_{i=1}^{n} c_i 1_{A_i}$.*

Alt. $i = 1, ..., n$, $A_i \in \mathscr{F}$, $c_i \in \mathbb{R}$ non-zero distinct, f is simple if $f = \sum_{i=1}^{n} c_i 1_{A_i}$.

$\triangleright$ 1. *$a, b \in \mathbb{R}$, g simple, then $af + bg$ simple*

Thm 2 (**Simple approximation**). (1) *$f \geqslant 0$ measurable. There exist simple $\{f_n\}$, $0 \leqslant f_n \uparrow f$, uniform if f is bounded.*

(2) *f measurable. There exist simple $\{f_n\}$, $f_n \to f$, uniform if f is bounded.*

PROOF. **1**. Let

$$f_n = \frac{[2^n f]}{2^n} \wedge n = \sum_{i=0}^{n2^n-1} \frac{i}{2^n} 1_{\{i/2^n \leqslant f < (i+1)/2^n\}} + n 1_{\{f \geqslant n\}}.$$

Then

$$0 \leqslant f - f_n \leqslant \frac{1}{2^n} \text{ if } f < n;\ f_n = n \leqslant f \text{ otherwise.}$$

2. $f = f^+ - f^-$. $\square$

Thm 3 (Doob). $f : (E, \mathscr{F}_E) \mapsto (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, *$g$ measurable* $(E, \mathscr{F}_E) \mapsto (F, \mathscr{F}_F)$. *If f is $\sigma(g)$-measurable, then $f = h \circ g$ for some measurable h.*

PROOF. **1**. $f = 1_A$, $A = g^{-1}(B) \in \sigma(g)$, $B \in \mathscr{F}_F$. Then $x \in A$ if and only if $g(x) \in B$, i.e.,

$$f = 1_A = 1_B \circ g.$$

2. f simple, $f = \sum_{i=1}^{n} c_i 1_{A_i}$, $c_i \in \mathbb{R}$, $A_i \in \sigma(g)$ disjoint. Let $A_i = g^{-1}(B_i)$, $B_i \in \mathscr{F}_F$, then

$$C_i = B_i \Big\backslash \Big(\bigcup\nolimits_{j<i} B_j \Big) \in \mathscr{F}_F \text{ disjoint}$$

and

$$f^{-1}(C_i) = A_i \Big\backslash \Big(\bigcup\nolimits_{j<i} A_j \Big) = A_i.$$

By step **1**,

$$f = \sum_{i=1}^{n} c_i 1_{A_i} = \sum_{i=1}^{n} c_i 1_{C_i} \circ g = \left(\sum_{i=1}^{n} c_i 1_{C_i} \right) \circ g \triangleq h \circ g.$$

3. $f \geqslant 0$ is $\sigma(g)$-measurable, there exist $\sigma(g)$-measurable simple f_n with $0 \leqslant f_n \uparrow f$. It follows $f_n = h_n \circ g$ for some h_n,

$$h \triangleq \sup_n h_n$$

is $\sigma(g)$-measurable,

$$f = \lim_n f_n = \sup_n (h_n \circ g) = \left(\sup_n h_n\right) \circ g = h \circ g.$$

4. f is $\sigma(g)$-measurable. f^+, f^- are $\sigma(g)$-measurable. Use **3**. $\square$

Thm 4. *$\mathscr{A}$ is a π-system, $\Omega \in \mathscr{A}$, $\mathcal{H}$ is a collection of real-valued functions. Suppose*

(1) *If $A \in \mathscr{A}$, then $1_A \in \mathcal{H}$*
(2) *If $f, g \in \mathcal{H}$, $c \in \mathbb{R}$, then $f + g$, $cg \in \mathcal{H}$*
(3) *If $f_n \in \mathcal{H}$, $0 \leqslant f_n \uparrow f$ with f bounded, then $f \in \mathcal{H}$*

Then

$$\{f : f \text{ bounded } \sigma(\mathscr{A})\text{-measurable}\} \subset \mathcal{H}$$

Proof. The system of sets

$$\mathscr{G} = \{A : 1_A \in \mathcal{H}\}$$

is a λ-system and $\mathscr{A} \subset \mathscr{G}$. Hence

$$\sigma(\mathscr{A}) \subset \mathscr{G}.$$

(2) implies that $\mathcal{H}$ contains all $\sigma(\mathscr{A})$-measurable simple functions, (3) implies that $\mathcal{H}$ contains all bounded $\sigma(\mathscr{A})$-measurable functions. $\square$

Thm 5. *$\mathscr{A}$ is a π-system, $\Omega \in \mathscr{A}$, $\mathcal{H}$ is a collection of real-valued functions. Suppose*

(1) *If $A \in \mathscr{A}$, then $1_A \in \mathcal{H}$*
(2) *If $f, g \in \mathcal{H}$, $a, b \geqslant 0$, then $af + bg \in \mathcal{H}$*
(3) *If $f, g \in \mathcal{H}$ are bounded, $f \geqslant g$, then $f - g \in \mathcal{H}$*
(4) *If $f_n \in \mathcal{H}$, $0 \leqslant f_n \uparrow f$, then $f \in \mathcal{H}$*

Then

$$\{f : f \text{ nonnegative } \sigma(\mathscr{A})\text{-measurable}\} \subset \mathcal{H}$$

2. 集函数与测度

2.1. 集函数. $\mathscr{E}$ is a collection of subsets of E.

Def 2. *Set function,* $\mu : \mathscr{E} \mapsto \mathbb{R} \cup \{\pm\infty\}$.

Def 3. *Nonnegative set function,* $\mu : \mathscr{E} \mapsto \mathbb{R} \cup \{\infty\}$.

Def 4. μ *is finite if,* $\forall A \in \mathscr{E}$, $|\mu(A)| < \infty$.

Def 5. μ *is* σ*-finite on* $\mathscr{E}$ *if,* $\forall A \in \mathscr{E}$, *there exist* $\{A_n\} \subset \mathscr{E}$, $A = \bigcup_n A_n$ *with* $|\mu(A_n)| < \infty$.

Def 6. μ *is additive if,* $\forall A, B \in \mathscr{E}$, $AB = \varnothing$,

$$\mu(A + B) = \mu(A) + \mu(B).$$

Def 7. μ *is countably additive if,* $\forall A_i \in \mathscr{E}$, $i = 1, 2, \ldots$, *disjoint,*

$$\mu\left(\sum_i A_i\right) = \sum_i \mu(A_i).$$

Def 8. *$\varnothing \in \mathscr{E}$. μ is a measure on $\mathscr{E}$ if it is nonnegative, countably additive, $\mu(\varnothing) = 0$.*

Example 1. *$(X, \mathscr{F})$ measurable space, $x \in X$,*

$$\delta_x(A) = 1_A(x), \ \forall A \in \mathscr{F}.$$

$x_1, ..., x_n \in X$,

$$\mu(A) = \sum_i \delta_{x_i}(A), \ \forall A \in \mathscr{F}.$$

Example 2. *F real-valued nonnegative, non-decreasing, right continuous. Semi-ring on $\mathbb{R}$,*

$$\mathscr{A} = \{(a, b] : a, b, \in \mathbb{R}\}.$$

Then

$$\mu((a, b]) = F(b) - F(a)$$

defines a measure $\mathscr{A}$. It is unique on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

PROOF. **1**. Additivity. $(a_i, b_i]$, $i = 1, ..., n$, disjoint, $(a, b] = \bigcup_i^n (a_i, b_i]$, then

$$\mu((a, b]) = \sum_{i=1}^n \mu((a_i, b_i]).$$

2. $(a_i, b_i]$, $i = 1, ...$, disjoint, $\bigcup_i (a_i, b_i] \subset (a, b]$, then

$$\sum_{i=1}^\infty \mu((a_i, b_i]) \leqslant \mu((a, b]).$$

3. $(a_i, b_i]$, $i = 1, ..., n$, $(a, b] \subset \bigcup_i^n (a_i, b_i]$, then

$$\mu((a, b]) \leqslant \sum_{i=1}^n \mu((a_i, b_i]).$$

4. $(a_i, b_i]$, $i = 1, ...,$ disjoint, $\bigcup_i (a_i, b_i] = (a, b]$, then

$$\mu((a,b]) = \sum_{i=1}^{\infty} \mu((a_i, b_i]).$$

$\forall \varepsilon > 0$, there is $\delta_i > 0$,

$$F(b_i + \delta_i) - F(b_i) < \frac{\varepsilon}{2^i}.$$

$\forall \theta > 0$, $\{(a_i, b_i + \delta_i) : i\}$ is an open cover of $[a + \theta, b]$, there exists n_0

$$(a + \theta, b] \subset \bigcup_i^{n_0} (a_i, b_i + \delta_i].$$

By **3**.,

$$
\begin{aligned}
\mu((a+\theta,b]) &\leqslant \sum_{i=1}^{n_0}\mu((a_i,b_i+\delta_i]) \\
&= \sum_{i=1}^{n_0}(F(b_i+\delta_i)-F(b_i)) \\
&\leqslant \sum_{i=1}^{n_0}(F(b_i)-F(b_i)) + \sum_{i=1}^{n_0}\frac{\varepsilon}{2^i} \\
&\leqslant \sum_{i=1}^{\infty}(F(b_i)-F(b_i)) + \varepsilon.
\end{aligned}
$$

$\square$

2.2. 半环上非负集函数. $\mathscr{E}$ is a collection of subsets of E, μ is a nonnegative set function on $\mathscr{E}$.

Def 9. *Monotonicity:* $\forall A \subset B \in \mathscr{E}$,

$$\mu(A) \leqslant \mu(B).$$

Def 10. *Countably subadditive:* $\forall A_i \in \mathscr{E}$, $i = 1, 2, ...$, $\bigcup_{i=1}^{\infty} A_i \in \mathscr{E}$,

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

Def 11. *Continuity from below:* $A_i \in \mathscr{E}$, $A_i \uparrow A \in \mathscr{E}$,

$$\lim_n \mu(A_i) = \mu(A).$$

Def 12. *Continuity from above:* $A_i \in \mathscr{E}$, $A_i \downarrow A \in \mathscr{E}$, $\mu(A_1) < \infty$,

$$\lim_n \mu(A_i) = \mu(A).$$

REMARK 1. ***Note*** *finiteness is part of the defintion of continuity from above.*

$\mathscr{S}$ is a semi-ring on E, μ is a nonnegative set function on $\mathscr{S}$.

Suppose μ is **additive**.

1. $\mu(\varnothing) = 0, +\infty$.

PROOF. $\varnothing \in \mathscr{S}$. By additivity

$$\mu(\varnothing) = \sum_{i=1}^{n} \mu(\varnothing).$$

$\mu(\varnothing)$ equals 0, or ∞. □

2. Monotonicity.

PROOF. $A, B \in \mathscr{S}$, $A \subset B$. There exist disjoint $C_1, ..., C_k \in \mathscr{S}$,

$$B \backslash A = \bigcup_{i=1}^{k} C_i.$$

$$B = A \cup (B \backslash A) = A \cup \left(\bigcup_{i=1}^{k} C_i \right).$$

By additivity

$$\mu(B) = \mu(A) + \sum_{i=1}^{k} \mu(C_i) \geqslant \mu(A).$$

□

Suppose μ is **countably additive**.

3. Continuity from below.

Proof. $A_i \in \mathscr{S}$, $A_i \uparrow A \in \mathscr{S}$. There exist disjoint $C_{n,1}, ..., C_{n,k_n} \in \mathscr{S}$,

$$B_n \triangleq A_n \backslash A_{n-1} = \bigcup_{i=1}^{k_n} C_{n,i}.$$

$(A_0 = \varnothing)$

$$\begin{aligned}
\mu(A) &= \mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \mu\left(\bigcup_{n=1}^{\infty}\bigcup_{i=1}^{k_n} C_{n,i}\right) \\
&= \sum_{n=1}^{\infty}\sum_{i=1}^{k_n} \mu(C_{n,i}) = \lim_N \sum_{n=1}^{N}\sum_{i=1}^{k_n} \mu(C_{n,i}) \\
&= \lim_N \mu\left(\bigcup_{n=1}^{N}\bigcup_{i=1}^{k_n} C_{n,i}\right) = \lim_n \mu(A_n).
\end{aligned}$$

$\square$

4. Continuity from above.

PROOF. (**WRONG PROOF**)$A_i \in \mathscr{S}$, $A_i \downarrow A \in \mathscr{S}$, $\mu(A_1) < \infty$. Clearly

$$\mu\left(\bigcap_{n=1}^{\infty} A_n\right) \leqslant \mu(A_i) \leqslant \mu(A_1) < \infty.$$

$$\lim_n \mu(A_n) = \mu\left(\bigcap_{n=1}^{\infty} A_n\right)$$

$$\Longleftrightarrow$$

$$\mu(A_1) - \lim_n \mu(A_n) = \mu(A_1) - \mu\left(\bigcap_{n=1}^{\infty} A_n\right)$$

$$\Longleftrightarrow$$

$$\lim_n \mu(A_1 \backslash A_n) = \mu\left(A_1 \Bigg\backslash \bigcap_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} (A_1 \backslash A_n)\right).$$

$\square$

5. Subadditivity.

PROOF. Analogous to continuity from below. $\square$

2.3. 环上非负集函数.

Thm 6. *$\mathscr{R}$ is a ring. μ is nonnegative additive.*
(1) μ *countably additive*

$$\Longleftrightarrow$$

(2) μ *countably subadditive*

$$\Longleftrightarrow$$

(3) μ *continuity from below*

$$\Longrightarrow$$

(4) μ *continuity from above*

$$\Longrightarrow$$

(5) μ *continuity from above at $\varnothing$.*

If μ is finite, (5) implies (1).

PROOF. **1**. Already have: (1) $\Longrightarrow$ (2), (1) $\Longrightarrow$ (3), (1) $\Longrightarrow$ (4), (4) $\Longrightarrow$ (5).

2. (2) $\Longrightarrow$ (1). Suppose $A_i \in \mathscr{R}$, $i = 1, 2, ...$, disjoint, $\bigcup\limits_{i=1}^{\infty} A_i \in \mathscr{R}$. By countable subadditivity,

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

By monotonicity and additivity,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \geqslant \mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i),\ \forall n.$$

Sending $n \to \infty$,

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

3. (3) $\Longrightarrow$ (1). Suppose $A_i \in \mathscr{R}$, $i = 1, 2, \ldots$, disjoint, $\bigcup_{i=1}^{\infty} A_i \in \mathscr{R}$. Since

$$\bigcup_{i=1}^{n} A_i \uparrow \bigcup_{i=1}^{\infty} A_i,$$

by continuity from below,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_n \mu\left(\bigcup_{i=1}^{n} A_i\right) = \lim_n \sum_{i=1}^{n} \mu(A_i) = \sum_{i=1}^{\infty} \mu(A_i).$$

4. (5) $\Longrightarrow$ (1). Suppose $A_i \in \mathscr{R}$, $i = 1, 2, \ldots$, disjoint, $\bigcup_{i=1}^{\infty} A_i \in \mathscr{R}$. Then, $\forall n$,

$$\bigcup_{i=1}^{n} A_i \in \mathscr{R} \text{ and } \bigcup_{i=n+1}^{\infty} A_i = \bigcup_{i=1}^{\infty} A_i \Big\backslash \bigcup_{i=1}^{n} A_i \in \mathscr{R}.$$

By additivity

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

Since μ is finite

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

The continuity from above at $\varnothing$ yields,

$$\lim_n \mu\left(\bigcup_{i=n+1}^{\infty} A_i\right) = 0.$$

Hence

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_n \mu\left(\bigcup_{i=1}^{n} A_i\right) + \lim_n \mu\left(\bigcup_{i=n+1}^{\infty} A_i\right)$$

$$= \lim_n \sum_{i=1}^{n} \mu(A_i) = \sum_{i=1}^{\infty} \mu(A_i).$$

□

3. Carathéodory's 延拓

3.1. 外测度.

Def 13. *μ^* is an outer measure on E if*
(1) $\mu^*(\varnothing) = 0$
(2) $\forall A, B \in 2^E$, *if* $A \subset B$, *then*

$$\mu^*(A) \leqslant \mu^*(B)$$

(3) *If* $A_i \in 2^E$,$i = 1, 2, ...,$

$$\mu^*\left(\bigcup_{i=1}^{\infty} A\right) \leqslant \sum_{i=1}^{\infty} \mu^*(A_i)$$

Thm 7. *Let $\mathscr{E}$ be a collection of sets on E, $\varnothing \in \mathscr{E}$. μ is a nonnegative set function on $\mathscr{E}$ with $\mu(\varnothing) = 0$. Define, $\forall A \in 2^E$,*

$$\mu^*(A) = \inf\left\{\sum_{i=1}^{\infty} \mu(A_i) : A_i \in \mathscr{E},\ A \subset \bigcup_{i=1}^{\infty} A_i\right\}.$$

Then $\mu^*(A)$ is an outer measure.

Proof. **1**. $\mu^*(\varnothing) = 0$ since $\varnothing \in \mathscr{E}$, $\varnothing \subset \bigcup\limits_{i=1}^{\infty} \varnothing$.

2. If $A \subset B$, $B \subset \bigcup\limits_{i=1}^{\infty} B_i$, then $A \subset \bigcup\limits_{i=1}^{\infty} B_i$, from the definition $\mu^*(A) \leqslant \mu^*(B)$.

3. Let $A_i \in 2^E, i = 1, 2, ..., \varepsilon > 0$. There are $A_{i,k} \in \mathscr{E}$, $A_i \subset \bigcup\limits_{k=1}^{\infty} A_{i,k}$,

$$\sum_{k=1}^{\infty} \mu(A_{i,k}) \leqslant \mu^*(A_i) + \frac{\varepsilon}{2^i},\ \forall i.$$

Since

$$\bigcup_{i=1}^{\infty} A_i \subset \bigcup_{i=1}^{\infty} \bigcup_{k=1}^{\infty} A_{i,k},$$

$$
\begin{aligned}
\mu^*\left(\bigcup_{i=1}^{\infty} A_i\right) &\leqslant \sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\mu(A_{i,k}) \\
&\leqslant \sum_{i=1}^{\infty}\left[\mu^*(A_i)+\frac{\varepsilon}{2^i}\right] \leqslant \sum_{i=1}^{\infty}\mu^*(A_i)+\varepsilon.
\end{aligned}
$$

$\square$

Def 14. *μ^* is an outer measure on E. $A \in 2^E$ is μ^*-measurable if*

$$\mu^*(D) = \mu^*(D \cap A) + \mu^*(D \cap A^c),\ \forall D \in 2^E.$$

The class of μ^*-measurable sets is denoted by $\mathscr{F}_\mu^*$.

Def 15. *Let μ be a measure on a σ-field $\mathscr{F}$ of E, the measure space $(E, \mathscr{F}, \mu)$ is complete if*

$$A \in \mathscr{F},\ \mu(A) = 0 \Longrightarrow B \in \mathscr{F},\ \forall B \subset A.$$

Thm 8 (Carathéodory). *Let $\mathscr{E}$ be a collection of sets on E, $\varnothing \in \mathscr{E}$. μ is a nonnegative set function on $\mathscr{E}$ with $\mu(\varnothing) = 0$.*

(1) $\mathscr{F}_\mu^*$ is a σ-field.

(2) $\left(E, \mathscr{F}_\mu^*, \mu^*\right)$ is a complete measure space.

Proof. **1**. Obviously, $E \in \mathscr{F}_\mu^*$ and $A^c \in \mathscr{F}_\mu^*$ if $A \in \mathscr{F}_\mu^*$.

2. If $A_1, A_2 \in \mathscr{F}_\mu^*$, then $A_1 \cup A_2$, $A_1 \cap A_2 \in \mathscr{F}_\mu^*$.

$\forall D \in 2^E$, we note

$$D \cap (A_1 \cup A_2) = (D \cap A_1) \cup (D \cap A_1^c \cap A_2).$$

Then

$$\begin{aligned}
&\mu^*(D \cap (A_1 \cup A_2)) + \mu^*(D \cap (A_1 \cup A_2)^c) \\
&\leqslant \mu^*(D \cap A_1) + \mu^*(D \cap A_1^c \cap A_2) + \mu^*(D \cap A_1^c \cap A_2^c) \text{ (subadditivity)} \\
&\leqslant \mu^*(D \cap A_1) + \mu^*(D \cap A_1^c) \ (A_2 \in \mathscr{F}_\mu^*) \\
&= \mu^*(D) \ (A_1 \in \mathscr{F}_\mu^*).
\end{aligned}$$

Hence

$$A_1 \cup A_2 \in \mathscr{F}_\mu^*.$$

It follows that

$$(A_1 \cap A_2)^c = A_1^c \cup A_2^c \in \mathscr{F}_\mu^*.$$

3. Finite additivity. If $A_1, ..., A_n \in \mathscr{F}_\mu^*$ disjoint, then $\forall D \in 2^E$,

$$\mu^* \left(D \cap \left(\bigcup_{i=1}^{n} A_i \right) \right) = \sum_{i=1}^{n} \mu^*(D \cap A_i).$$

Indeed, since $A_1 \in \mathscr{F}_\mu^*$,

$$\begin{aligned}
&\mu^*\left(D \cap \left(\bigcup_{i=1}^n A_i\right)\right)\\
&= \mu^*\left(D \cap \left(\bigcup_{i=1}^n A_i\right) \cap A_1\right) + \mu^*\left(D \cap \left(\bigcup_{i=1}^n A_i\right) \cap A_1^c\right)\\
&= \mu^*(D \cap A_1) + \mu^*\left(D \cap \left(\bigcup_{i=2}^n A_i\right)\right) = \cdots = \sum_{i=1}^n \mu^*(D \cap A_i)
\end{aligned}$$

4. If $A_1, A_2, ..., \in \mathscr{F}_\mu^*$, then $A \triangleq \bigcup_{i=1}^{\infty} A_i \in \mathscr{F}_\mu^*$.

We can assume that $A_1, A_2, ..., \in \mathscr{F}_\mu^*$ are disjoint. Indeed, by **1** and **2**, $B_i = A_i \Big\backslash \left(\bigcup_{j<i} A_j\right) \in \mathscr{F}_\mu^*$, are disjoint and $\bigcup_{i=1}^n B_i = \bigcup_{i=1}^n A_i$,

$\forall n$. Let

$$C_n = \bigcup_{i=1}^{n} A_i \in \mathscr{F}_\mu^*, \ \forall n.$$

Since $A_1, A_2, ...$ are disjoint, we can use **3** (the finite additivity). $\forall D \in 2^E$,

$$\begin{aligned}\mu^*(D) &= \mu^*(D \cap C_n) + \mu^*(D \cap C_n^c) \\ &= \sum_{i=1}^{n} \mu^*(D \cap C_i) + \mu^*(D \cap C_n^c) \\ &\geqslant \sum_{i=1}^{n} \mu^*(D \cap C_i) + \mu^*(D \cap A^c), \ \forall n.\end{aligned}$$

Let $n \to \infty$, note $A \subset \bigcup_{i=1}^{\infty} C_i$ and use subadditivity of outer measure

$$\mu^*(D) \geqslant \sum_{i=1}^{\infty} \mu^*(D \cap C_i) + \mu^*(D \cap A^c) \geqslant \mu^*(D \cap A) + \mu^*(D \cap A^c).$$

5. Countable additivity.

If $A_1, A_2, ..., \in \mathscr{F}_\mu^*$ are disjoint, use **3** and send $n \to \infty$,

$$\mu^*\left(\bigcup_{i=1}^{\infty} A_i\right) \geqslant \mu^*\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu^*(A_i),\ \forall n.$$

The opposite inequality is subadditivity of outer measure.

6. Completeness. If $A \in \mathscr{F}_\mu^*$, $\mu^*(A) = 0$ and $B \subset A$, then $\mu^*(B) = 0$. $\forall D \in 2^E$,

$$\mu^*(D) \geqslant \mu^*(D \cap B^c) = \mu^*(D \cap B) + \mu^*(D \cap B^c).$$

So $B \in \mathscr{F}_\mu^*$. □

3.2. 域上测度的延拓.

Thm 9. *If μ is a measure on a field $\mathscr{A}$ with the generated outer measure μ^*. Then*

(1) $\mathscr{A} \subset \mathscr{F}_\mu^*$ *thus* $\sigma(\mathscr{A}) \subset \mathscr{F}_\mu^*$.
(2) μ^* *is an extension of* μ *to* $\sigma(\mathscr{A})$ *in the sense that*

$$\mu(A) = \mu^*(A), \ \forall A \in \mathscr{A}.$$

Proof. **1**. Let $A \subset \mathscr{A}$. If $A_i \in \mathscr{A}$, $A \subset \bigcup_{i=1}^{\infty} A_i$, then

$$\mu(A) \leqslant \sum_{i=1}^{\infty} \mu(A_i). \tag{3.1}$$

Indeed,

$$\mu\left(A \cap \bigcup_{i=1}^{n} A_i\right) \leqslant \mu\left(\bigcup_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} \mu(A_i).$$

Let $n \to \infty$ and use that μ is a measure to get (3.1). So

$$\mu(A) \leqslant \mu^*(A).$$

Since $A \subset \mathscr{A}$, $A_1 = A$, $A_2 = A_3 ... = \varnothing$ form a countable cover of A, so

$$\mu^*(A) \leqslant \mu(A).$$

2. Fix $A \subset \mathscr{A}$, will prove $A \in \mathscr{F}_\mu^*$. $\forall D \in 2^E$, it is enough to show that

$$\mu^*(D) \geqslant \mu^*(D \cap A) + \mu^*(D \cap A^c).$$

There is nothing to prove if $\mu^*(D) = \infty$, so we assume that $\mu^*(D) < \infty$. Then, $\forall \varepsilon > 0$, there exist $A_i \in \mathscr{A}$, $D \subset \bigcup\limits_{i=1}^{\infty} A_i$ so that

$$\sum_{i=1}^{\infty} \mu(A_i) \leqslant \mu^*(D) + \varepsilon.$$

Since $\mathscr{A}$ is a field,

$$A_i \cap A, A_i \cap A^c \in \mathscr{A}.$$

By **1** and the additivity of μ,

$$\begin{aligned}\mu(A_i) &= \mu(A_i \cap A) + \mu(A_i \cap A^c) \\ &= \mu^*(A_i \cap A) + \mu^*(A_i \cap A^c).\end{aligned}$$

Summing over i gives

$$\begin{aligned}\sum_{i=1}^{\infty} \mu(A_i) &= \sum_{i=1}^{\infty} \mu^*(A_i \cap A) + \sum_{i=1}^{\infty} \mu^*(A_i \cap A^c) \\ &\geqslant \mu^*(D \cap A) + \mu^*(D \cap A^c).\end{aligned}$$

So

$$\mu^*(D) + \varepsilon \geqslant \sum_{i=1}^{\infty} \mu(A_i) \geqslant \mu^*(D \cap A) + \mu^*(D \cap A^c).$$

$\square$

Thm 10 (Uniqueness)**.** *Let $\mathscr{P}$ be a π-system on E, μ and ν measures on $\sigma(\mathscr{P})$. Assume that*

(1) *μ and ν agree on $\mathscr{P}$.*

(2) *There are* $B_i \in \mathscr{P}$, $i = 1, 2, ...,$ *disjoint so that* $\bigcup_{i=1}^{\infty} B_i = E$ *and* $\mu(B_i) < \infty$.

Then μ *and* ν *are equal on* $\sigma(\mathscr{P})$.

PROOF. **1**. Let $B \in \mathscr{P}$ have $\mu(B) < \infty$. Define

$$\mathscr{L} = \{A \in \sigma(\mathscr{P}) : \mu(A \cap B) = \nu(A \cap B)\}.$$

$\mathscr{L}$ is a λ-system (finiteness is needed to justify sets subtraction!), $\mathscr{P} \subset \mathscr{L}$. So

$$\sigma(\mathscr{P}) \subset \mathscr{L},$$

i.e.

$$\mu(A \cap B) = \nu(A \cap B), \ \forall A \in \sigma(\mathscr{P}).$$

2. $\forall A \in \sigma(\mathscr{P})$, use (2) to write it as disjoint union,

$$A = \bigcup_{i=1}^{\infty} (A \cap B_i), \ \mu(A \cap B_i) \leqslant \mu(B_i) < \infty.$$

Then, by **1**,

$$\mu(A) = \mu\left(\bigcup_{i=1}^{\infty}(A \cap B_i)\right) = \sum_{i=1}^{\infty}\mu(A \cap B_i)$$
$$= \sum_{i=1}^{\infty}\nu(A \cap B_i) = \nu\left(\bigcup_{i=1}^{\infty}(A \cap B_i)\right) = \nu(A).$$

□

▷ 2. *The condition Therem 10 (2) can be replaced with either one of the following:*
(2′) $\mathscr{P}$ *is a semi-ring,* $E \in \mathscr{P}$ *and* μ *is* σ*-finite on* $\mathscr{P}$*.*
(2″) *there are* $B_1, B_2, \ldots \in \mathscr{P}$*, so that* $B_i \uparrow E$ *and* $\mu(B_i) < \infty$*.*

3.3. 半环上测度的延拓.

Thm 11. *Let* μ *be a measure on the semi-ring* $\mathscr{S}$ *with the generated outer measure* μ^**. Then*
(1) $\mathscr{S} \subset \mathscr{F}_\mu^*$ *thus* $\sigma(\mathscr{S}) \subset \mathscr{F}_\mu^*$*.*

(2) *μ^* is an extension of μ to $\sigma(\mathscr{S})$ in the sense that*

$$\mu(A) = \mu^*(A), \ \forall A \in \mathscr{S}. \tag{3.2}$$

(3) *Assume that there are $B_i \in \mathscr{S}$, $i = 1, 2, ...,$ disjoint so that $\bigcup_{i=1}^{n} B_i = E$ and $\mu(B_i) < \infty$, then the extension of μ to $\sigma(\mathscr{S})$ is unique.*

PROOF. Let $\bar{\mu}$ be the outer measure generated by μ.

1. $\bar{\mu}$ agrees with μ on $\mathscr{S}$.

The proof is identical to Theorem 9 (1).

2. Fix $A \subset \mathscr{S}$, will prove $A \in \mathscr{F}^*_\mu$.

The proof is identical to Theorem 9 (2). The difference is $A_i \cap A^c$ is replaced with disjoint union of sets in $\mathscr{S}$.

3. Uniqueness. Apply Theorem 10 to conclude. $\square$

3.4. Approximating $\mu^*|_{\mathscr{F}_\mu^*}$ by $\mu^*|_{\sigma(\mathscr{S})}$.

Thm 12. *Let μ be a measure on the semi-ring $\mathscr{S}$ with the generated outer measure μ^*. Suppose $E \in \mathscr{S}$.*

(1) $\forall A \in \mathscr{F}_\mu^*$, *there is $B \in \sigma(\mathscr{S})$ such that $A \subset B$ and*

$$\mu^*(A) = \mu^*(B).$$

(2) *If μ is σ-finite on $\mathscr{S}$, then $\forall A \in \mathscr{F}_\mu^*$, there is $B \in \sigma(\mathscr{S})$ such that $A \subset B$ and*

$$\mu^*(B \backslash A) = 0.$$

Proof.

1. There is nothing to prove if $\mu^*(A) = \infty$, we assume that $\mu^*(A) < \infty$. There are $B_{n,i} \in \mathscr{S}$, $A \subset \bigcup_{i=1}^{\infty} B_{n,i}$,

$$\sum_{i=1}^{\infty} \mu(B_{n,i}) < \mu^*(A) + \frac{1}{n}.$$

Set

$$B = \bigcap_{n=1}^{\infty} \bigcup_{i=1}^{\infty} B_{n,i}.$$

Then $A \subset B \in \sigma(\mathscr{S})$,

$$\mu^*(A) \leqslant \mu^*(B).$$

Moreover

$$\mu^*(B) \leqslant \mu^*\left(\bigcup_{i=1}^{\infty} B_{n,i}\right) \leqslant \sum_{i=1}^{\infty} \mu(B_{n,i}) \leqslant \mu^*(A) + \frac{1}{n}.$$

It follows that

$$\mu^*(B) \leqslant \mu^*(A).$$

2. If μ is *finite* on $\mathscr{S}$, then by **1**, $\forall A \in \mathscr{F}_\mu^*$, there is $B \in \sigma(\mathscr{S})$ such that $A \subset B$ and

$$\mu^*(A) = \mu^*(B).$$

Since μ^* is a measure on $\mathscr{F}_\mu^*$, this gives

$$\mu^*(B \backslash A) = 0.$$

The σ-finite case follows from similar argument as in step **3** of Theorem 11. $\square$

3.5. Approximating $\mu|_{\sigma(\mathscr{A})}$ by $\mu|_{\mathscr{A}}$.

Thm 13. *Let μ be a measure on the field $\mathscr{A}$ with the generated outer measure μ^*. For any $A \in \sigma(\mathscr{A})$ with $\mu^*(A) < \infty$, $\forall \varepsilon > 0$, there is $B \in \mathscr{A}$ such that $\mu^*(A \Delta B) < \varepsilon$.*

If, in the last Theorem, the measure μ is defined on $\sigma(\mathscr{A})$ and σ-finite on $\mathscr{A}$, then μ must equal μ^* on $\sigma(\mathscr{A})$ by uniqueness, we can use μ in place of μ^* in the conclusion.

Thm 14. *Let $\mathscr{A}$ be a field, μ a measure on $\sigma(\mathscr{A})$ and σ-finite on $\mathscr{A}$. For any $A \in \sigma(\mathscr{A})$ with $\mu(A) < \infty$, $\forall \varepsilon > 0$, there is $B \in \mathscr{A}$ such that $\mu(A \Delta B) < \varepsilon$.*

3.6. Completion of a measure space.

Thm 15. *Let $(X, \mathscr{F}, \mu)$ be a measure space,*

$$\bar{\mathscr{F}} \triangleq \{A \cup N : A \in \mathscr{F}, N \subset B \text{ for some } B \in \mathscr{F} \text{ with } \mu(B) = 0\}.$$

Define

$$\bar{\mu}(A \cup N) = \mu(A),\ \forall A \in \bar{\mathscr{F}}.$$

Then $\left(X, \bar{\mathscr{F}}, \bar{\mu}\right)$ *is a complete measure space.*

Clearly the Theorem says

$$\bar{\mu}(A) = \mu(A),\ \forall A \in \mathscr{F}.$$

PROOF. **1**. $\bar{\mathscr{F}}$ is a σ-field.

Suppose $A \cup N \in \bar{\mathscr{F}}$ where $A \in \mathscr{F}$, $N \subset B$, $B \in \mathscr{F}$ with $\mu(B) = 0$. Then

$$(A \cup N)^c = (A^c \cap B^c) \cup (B \cap A^c \cap N^c) \in \bar{\mathscr{F}}.$$

Suppose $A_i \cup N_i \in \bar{\mathscr{F}}$ where $A_i \in \mathscr{F}$, $N_i \subset B_i$, $B_i \in \mathscr{F}$ with $\mu(B_i) = 0$. Then

$$\bigcup_{i=1}^{\infty}(A_i \cup N_i) = \left(\bigcup_{i=1}^{\infty} A_i\right) \cup \left(\bigcup_{i=1}^{\infty} N_i\right) \in \bar{\mathscr{F}},$$

since

$$\bigcup_{i=1}^{\infty} N_i \subset \bigcup_{i=1}^{\infty} B_i \in \mathscr{F}$$

and

$$\mu\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mu(B_i) = 0.$$

2. The definition of $\bar{\mu}$ nonambiguous, i.e.

$$A_1 \cup N_1 = A_2 \cup N_2 \in \bar{\mathscr{F}} \Longrightarrow \bar{\mu}(A_1 \cup N_1) = \bar{\mu}(A_2 \cup N_2).$$

Here $N_i \subset B_i$ for some $B_i \in \mathscr{F}$ with $\mu(B_i) = 0$, $i = 1, 2$.

$$\bar{\mu}(A_1 \cup N_1) = \mu(A_1) = \mu(A_1 \cup B_1 \cup B_2) \geqslant \mu(A_2) = \bar{\mu}(A_2 \cup N_2).$$

By symmetry,

$$\bar{\mu}(A_1 \cup N_1) \leqslant \bar{\mu}(A_2 \cup N_2).$$

(In fact

$$A_1 \cup B_1 \cup B_2 = A_1 \cup N_1 \cup B_1 \cup B_2 = A_2 \cup N_2 \cup B_1 \cup B_2 = A_2 \cup B_1 \cup B_2$$

so

$$\mu(A_1 \cup B_1 \cup B_2) = \mu(A_2).$$

)

3. Countable additivity. Suppose $A_i \cup N_i \in \bar{\mathscr{F}}$ disjoint, where $A_i \in \mathscr{F}$, $N_i \subset B_i$, $B_i \in \mathscr{F}$ with $\mu(B_i) = 0$. Then

$$\bar{\mu}\left(\bigcup_{i=1}^{\infty}(A_i \cup N_i)\right) = \mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) = \sum_{i=1}^{\infty} \bar{\mu}(A_i \cup N_i).$$

4. Completeness. Let $A \cup N \in \bar{\mathscr{F}}$, $N \subset B$, $B \in \mathscr{F}$ with $\mu(B) = 0$ and $\bar{\mu}(A \cup N)$, then

$$\mu(A \cup B) = \mu(A) = \bar{\mu}(A \cup N) = 0.$$

So for any $C \subset A \cup N$, $C \subset A \cup B$,

$$C = \varnothing \cup C \in \bar{\mathscr{F}}.$$

$\square$

Thm 16. *Suppose that μ is σ-finite on the semi-ring $\mathscr{S}$ with the generated outer measure μ^*. Then $\left(X, \mathscr{F}_\mu^*, \mu^*\right)$ is the completion of $(X, \sigma(\mathscr{S}), \mu^*)$.*

Proof. Let

$\bar{\mathscr{F}} \triangleq \{A \cup N : A \in \sigma(\mathscr{S}), N \subset B$ for some $B \in \sigma(\mathscr{S})$ with $\mu(B) = 0\}$.

It is enough to show that

$$\mathscr{F}_\mu^* = \bar{\mathscr{F}}.$$

Since $\left(X, \mathscr{F}_\mu^*, \mu^*\right)$ is a complete measure space,

$$\bar{\mathscr{F}} \subset \mathscr{F}_\mu^*.$$

Let $A \in \mathscr{F}_\mu^*$, by Theorem 12 there exist $B, C \in \sigma(\mathscr{S})$ so that

$$A \subset B,\ \mu^*(B \backslash A) = 0;\ B \backslash A \subset C,\ \mu^*(C) = \mu^*(B \backslash A) = 0.$$

Writing

$$A = (B \cap C^c) \cup (A \cap C),$$

we get that $B \cap C^c \in \sigma(\mathscr{S})$, $(A \cap C) \subset C$, $\mu^*(C) = 0$, so $A \in \bar{\mathscr{F}}$. $\square$

4. 收敛

4.1. 可测函数的收敛. $(E, \mathscr{F}, \mu)$ a measure space, $f_n \in \mathscr{F}$, $i = 1, 2, ...,$ $f \in \mathscr{F}$

Def 16. *Almost everywhere convergence,* $f_n \xrightarrow{a.e.} f$*:*

$$\mu\Big(\lim_n f_n \neq f\Big) = 0.$$

Def 17. *Convergence in measure,* $f_n \xrightarrow{\mu} f$*:* $\forall \varepsilon > 0$,

$$\lim_n \mu(|f_n - f| > \varepsilon) = 0.$$

Evidently

$$\begin{aligned} f_n \xrightarrow{a.e.} f &\iff \forall \varepsilon > 0,\ \mu\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{|f_m - f| > \varepsilon\}\right) = 0 \\ &\iff \forall \varepsilon > 0,\ \mu(\{|f_n - f| > \varepsilon\} \text{ i.o.}) = 0. \end{aligned}$$

Recall

$$x \in \limsup A_n \iff x \in A_n \text{ i.o.}$$

Thm 17. *If μ is finite, then*

$$f_n \xrightarrow{a.e.} f \implies f_n \xrightarrow{\mu} f.$$

PROOF. Indeed,

$$\mu(|f_n - f| > \varepsilon) \leqslant \mu\left(\bigcup_{m=n}^{\infty} \{|f_m - f| > \varepsilon\}\right), \ \forall n.$$

Let $n \to \infty$ and use continuity from above (requires finiteness of μ)

$$\begin{aligned} \limsup_n \mu(|f_n - f| > \varepsilon) &\leqslant \lim_n \mu\left(\bigcup_{m=n}^{\infty} \{|f_m - f| > \varepsilon\}\right) \\ &= \mu\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{|f_m - f| > \varepsilon\}\right) = 0. \end{aligned}$$

(or use

$$\limsup_n \mu(A_n) \leqslant \mu\left(\limsup_n A_n\right).$$

) $\square$

Def 18. *Almost uniform convergence,* $f_n \xrightarrow{a.u.} f$*:* $\forall \varepsilon > 0$*, there is* $A_\varepsilon \in \mathscr{F}$ *so that* $\mu(A_\varepsilon) < \varepsilon$*,*

$$\lim_n \sup_{x \notin A_\varepsilon} |f_n - f| = 0.$$

Compare with Egoroff's Theorem on *finite* measure!

Thm 18. $f_n \xrightarrow{a.u.} f$ *if and only if* $\forall \varepsilon > 0$,

$$\lim_n \mu\left(\bigcup_{m=n}^{\infty} \{|f_m - f| > \varepsilon\}\right) = 0.$$

Proof. **1**. "$\Longrightarrow$". $\forall \varepsilon > 0$, there is A_ε so that $\mu(A_\varepsilon) < \varepsilon$ and

$$\lim_m \sup_{x \notin A_\varepsilon} |f_m - f| = 0.$$

So, $\forall \varepsilon' > 0$, there is $n_0 \in \mathbb{N}$ such that

$$\sup_{x \notin A_\varepsilon} |f_m - f| \leqslant \varepsilon', \ \forall m \geqslant n_0.$$

This translates to

$$\bigcup_{m=n_0}^{\infty} \{|f_m - f| > \varepsilon'\} \subset A_\varepsilon.$$

Therefore

$$\mu\left(\bigcup_{m=n_0}^{\infty} \{|f_m - f| > \varepsilon'\}\right) \leqslant \mu(A_\varepsilon) < \varepsilon.$$

2. " $\Longleftarrow$ ". $\forall \varepsilon > 0$ and $k \in \mathbb{N}$, there is $n_{\varepsilon,k} \in \mathbb{N}$ such that

$$\mu\left(\bigcup_{m=n_{\varepsilon,k}}^{\infty} \left\{|f_m - f| > \frac{1}{k}\right\}\right) < \frac{\varepsilon}{2^k},\ \forall m \geqslant n_{\varepsilon,k}.$$

Denote (the set of all possible divergence points! measurable!)

$$A_\varepsilon = \bigcup_{k=1}^{\infty} \bigcup_{m=n_{\varepsilon,k}}^{\infty} \left\{|f_m - f| > \frac{1}{k}\right\}.$$

Then $\mu(A_\varepsilon) < \varepsilon$ and for any $x \notin A_\varepsilon$, we have $\forall k$,

$$|f_m - f| \leqslant \frac{1}{k},\ \forall m > n_{\varepsilon,k}.$$

$\square$

We have proved:

Thm 19. (1)

$$f_n \xrightarrow{a.u.} f \implies f_n \xrightarrow{a.e.} f \text{ and } f_n \xrightarrow{\mu} f$$

(2) *If μ is finite, then*

$$f_n \xrightarrow{a.u.} f \iff f_n \xrightarrow{a.e.} f \implies f_n \xrightarrow{\mu} f$$

Example 3.

$$f_n(x) = \begin{cases} 1, & x \in (0, 1/n), \\ 0, & x \in [1/n, 1]. \end{cases}$$

Example 4.

$$f_n(x) = x^n, x \in [0, 1]$$

$\triangleright$ 3. *Let* $f = 0$ *and* $f_n = 1_{A_n}$. *Then* $f_n \xrightarrow{\mu} f$ *is equivalent to* $\mu(A_n) \to 0$ *and* $\left(\lim_n f_n \neq f\right) = (A_n \ i.o.)$.

Any sequence $\{A_n\}$ so that $\mu(A_n) \to 0$ but $\mu(A_n \text{ i.o.}) > 0$ gives an exmple that $f_n \xrightarrow{\mu} f \nRightarrow f_n \xrightarrow{a.e.} f$. It is enough to have $\mu(A_n) \to 0$ and

$$\sum_{i=1}^{\infty} 1_{A_n}(x) = \infty,\ \sum_{i=1}^{\infty} 1_{A_n^c}(x) = \infty.$$

Example 5. *For each* $n = 1, 2, ...$ *there is a unique decomposition* $n = k(k-1)/2 + i$ *with* $k = 1, 2, ...,\ i = 1, 2, ..., k.$

$$f_n(x) = \begin{cases} 1, & x \in (((i-1)/k, i/k]), \\ 0, & \textit{otherwise.} \end{cases}$$

Example 6. *Consider*

$$A_k^i = \left[\frac{i-1}{k}, \frac{i}{k}\right],\ h_k^i(x) = 1_{A_k^i}(x),\ i = 1, ..., k.$$

Let f_n be the sequence

$$\{h_1^1; h_2^1, h_2^2; h_3^1, h_3^2; h_3^3; ...\}$$

Thm 20. $f_n \xrightarrow{\mu} f \iff$ *for any subsequence there is a further subsequence* $f_{n_k} \xrightarrow{a.u.} f$.

PROOF. " $\Longrightarrow$ ". Since any subsequence of f_n converges in measure to f, it is enough to show there is a subsequence $f_{n_k} \xrightarrow{a.u.} f$. To see this, for any $k > 0$, by definition of convergence in measure, we can choose $n_k > n_{k-1}$ so that

$$\mu\left(|f_{n_k} - f| > \frac{1}{k}\right) \leqslant \frac{1}{2^k}.$$

Then

$$\mu\left(\bigcup_{k=m}^{\infty} |f_{n_k} - f| > \frac{1}{k}\right) \leqslant \sum_{k=m}^{\infty} \frac{1}{2^k} = \frac{1}{2^{m-1}}.$$

$\forall \varepsilon > 0$, for large m,

$$\bigcup_{k=m}^{\infty} \{|f_{n_k} - f| > \varepsilon\} \subset \bigcup_{k=m}^{\infty} \left\{|f_{n_k} - f| > \frac{1}{k}\right\}.$$

So

$$\lim_m \mu\left(\bigcup_{k=m}^{\infty} |f_{n_k} - f| > \varepsilon\right) \leqslant \lim_m \mu\left(\bigcup_{k=m}^{\infty} |f_{n_k} - f| > \frac{1}{k}\right) = 0.$$

" $\Longleftarrow$ " Suppose $f_n \xrightarrow{\mu} f$ does not hold, i.e. there are $n_k \to \infty$, $\varepsilon_0 > 0$, $\delta_0 > 0$ so that

$$\mu(|f_{n_k} - f| > \varepsilon_0) > \delta_0.$$

Then

$$\liminf_m \mu\left(\bigcup_{k=m}^{\infty} |f_{n_k} - f| > \varepsilon_0\right) \geqslant \delta_0,$$

Contradicting Theorem 18. □

Theorem 19 and Theorem 20 indicate that if $f_n \xrightarrow{\mu} f$, then there is a subsequence $f_{n_k} \xrightarrow{a.e.} f$.

4.2. 随机变量的分布函数.

Def 19. $(\Omega, \mathscr{F}, P)$ *is a probability space if* P *is a nonnegative measure on the* σ*-field* $\mathscr{F}$ *with* $P(\Omega) = 1$.

Def 20. *A random variable (r.v.)* X *on* $(\Omega, \mathscr{F}, P)$ *is a real-valued mapping,* $X : \omega \in \Omega \mapsto X(\omega) \in \mathbb{R}$.

Def 21. *The distribution function of a r.v.* X *is*

$$F(x) = P(X \leqslant x).$$

Denoted by $X \sim F$.

Thm 21. *Any distribution function* F *has the following properties.*
(1) *non-decreasing,* $F(-\infty) = 0$ *and* $F(\infty) = 1$
(2) *right continuity:* $\lim\limits_{y \downarrow x} F(y) = F(x)$.
(3) *left limit exists:* $F(x-) = \lim\limits_{y \uparrow x} F(y) = P(X < x)$.

(4) $P(X = x) = F(x) - F(x-)$.

The **inverse of the distrbution function** F is defined as below. $\forall z \in (0,1)$,

$$F^{-1}(z) = \inf\{x \in \mathbb{R} : F(x) \geqslant z\}. \tag{4.1}$$

▷ 4. *Also equivalently defined as,*

$$F^{-1}(z) = \sup\{x \in \mathbb{R} : F(x) < z\}. \tag{4.2}$$

LEMMA 22. *F^{-1} has the properties,*
(1) *F^{-1} is real-valued non-decreasing.*
(2) *F^{-1} is left-continuous and has right limit.*
(3) *$F^{-1}(F(x)) \leqslant x$, $F\big(F^{-1}(z)\big) \geqslant z$.*
(4) *$F^{-1}(z) \leqslant x$ iff $F(x) \geqslant z$.*

PROOF. Exercise. □

Thm 23. *If F satisfies (1)(2)(3) of Theorem 21, there is a r.v. X with distribution F.*

PROOF. Let $\Omega = (0,1)$, $\mathscr{F} = \mathscr{B}_{(0,1)}$ (i.e. $(0,1) \cap \mathscr{B}_{\mathbb{R}}$), $P =$ Lebesgue measure. Define

$$X(\omega) = F^{-1}(\omega).$$

Then X is $\mathscr{F}$-measurable (check this!) and

$$\begin{aligned} P(\omega : X(\omega) \leqslant x) &= P(\omega : F(x) \geqslant \omega) \\ &= \text{Lebesgue measure of } (0, F(x)) = F(x). \end{aligned}$$

So X is a r.v. with distribution function F. □

▷ 5. *Another construction of a r.v. X with distribution F is to take $(\Omega, \mathscr{F}) = (\mathbb{R}, \mathscr{B})$, $P =$ the Lebesgue measure induced by F and consider the coordinate map $X(\omega) = \omega$.*

4.3. 随机变量的收敛. Probability space $(\Omega, \mathscr{F}, P)$, r.v. X_n, X,

$$X_n \xrightarrow{a.s.} X \iff P(X_n = X) = 1.$$

$$X_n \xrightarrow{P} X \iff \forall \varepsilon > 0,\ \lim_n P(|X_n - X| > \varepsilon) = 0.$$

Def 22. $X_n \sim F_n$, $X \sim F$. *Convergence in distribution (weak convergence): $F_n(x) \to F(x)$ for all x where F is continuous, written $X_n \xrightarrow{d} X$.*

Thm 24. $X_n \sim F_n$, $X \sim F$.

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X.$$

PROOF. **1**. The first implication is a special case of Theorem 17.

2. Check the second implication. $\forall \varepsilon, x \in \mathbb{R}, n \in \mathbb{N}$,

$$\begin{aligned} & P(X \leqslant x - \varepsilon) - P(|X_n - X| > \varepsilon) \\ & \leqslant P(X_n \leqslant x) \\ & \leqslant P(X_n \leqslant x, |X_n - X| \leqslant \varepsilon) + P(X_n \leqslant x, |X_n - X| > \varepsilon) \\ & \leqslant P(X \leqslant x + \varepsilon) + P(|X_n - X| > \varepsilon). \end{aligned}$$

So $n \to \infty$, $\varepsilon \to 0$ yield

$$F(x-) \leqslant \liminf_n P(X_n \leqslant x) \leqslant \limsup_n P(X_n \leqslant x) \leqslant F(x).$$

$\square$

LEMMA 25. $F_n \xrightarrow{w} F \iff F_n^{-1} \xrightarrow{w} F^{-1}$.

PROOF OF " $\Longrightarrow$ ". Construct r.v.s' $X_n \sim F_n$, $X \sim F$ as Theorem 23. Fix any ω.

1. Choose any $\varepsilon > 0$ so that F is continuous at $X(\omega) - \varepsilon$ (the discontinuities of F are at most countable, ε can be arbitrarily small). By the definition (the infimum!) of $X(\omega)$,

$$F(X(\omega) - \varepsilon) < \omega.$$

Then, for large n,

$$F_n(X(\omega) - \varepsilon) < \omega.$$

so (note the above inequality is strict)

$$X(\omega) - \varepsilon \leqslant X_n(\omega).$$

Hence

$$X(\omega) \leqslant \liminf_n X_n(\omega).$$

2. To see the opposite. Choose any $\varepsilon, \delta > 0$ so that X is continuous at ω and F is continuous at $X(\omega) + \varepsilon$, then by Lemma 22

$$F(X(\omega + \delta) + \varepsilon) \geqslant F(X(\omega + \delta)) \geqslant \omega + \delta > \omega.$$

For large n ($\delta > 0$),

$$F_n(X(\omega + \delta) + \varepsilon) \geqslant \omega.$$

By Lemma 22 again,

$$X(\omega + \delta) + \varepsilon \geqslant X_n(F_n(X(\omega + \delta) + \varepsilon)) \geqslant X_n(\omega).$$

Let $n \to \infty$, $\varepsilon \to 0$, $\delta \to 0$ (continuity at ω),

$$X(\omega) \geqslant \limsup_n X_n(\omega).$$

□

Thm 26 (Skorohod)**.** $X_n \sim F_n$, $X \sim F$. *Suppose* $X_n \xrightarrow{d} X$. *There exist r.v.* $\bar{X}_n$, $\bar{X}$ *on a common probability space so that* $\bar{X}_n \stackrel{d}{=} X_n$, $\bar{X} \stackrel{d}{=} X$, $\bar{X}_n \xrightarrow{a.s.} \bar{X}$.

Proof. Let $\Omega = (0,1)$, $\mathscr{F} = \mathscr{B}_{(0,1)}$, $P =$ Lebesgue measure. By Theorem 23 there exist r.v. on $(\Omega, \mathscr{F}, P)$ so that $\bar{X}_n \sim F_n$, $\bar{X} \sim F$. Lemma 25 then says $F_n^{-1} \xrightarrow{w} F^{-1}$. Since the discontinuity set of F^{-1} is countable, $F_n^{-1}(\omega) \to F^{-1}(\omega)$ for almost all $\omega \in \Omega$, i.e. $\bar{X}_n(\omega) \xrightarrow{a.s.} \bar{X}(\omega)$. $\square$

5. 积分

5.1. 非负可测函数积分. $(E, \mathscr{F}, \mu)$ a measure space, $f \in \mathscr{F}$ with values in $[0, \infty]$,. A *finite (measurable) partition* of E is a finite collection of $\mathscr{F}$-measurable sets $\{A_i : i = 1, ..., m\}$ with $\displaystyle\bigcup_{i=1}^{m} A_i = E$.

$$\int f d\mu \triangleq \sup_{\text{finite partitions}} \sum_i \left[\inf_{x \in A_i} f(x)\right] \mu(A_i). \tag{5.1}$$

Convention: $0 \cdot \infty = 0$.

▷ 6. *Consider*

$$\int f d\mu \triangleq \inf_{\textit{finite partitions}} \sum_i \left[\sup_{x \in A_i} f(x)\right] \mu(A_i). \tag{5.2}$$

Is (5.2) a good definition of integration?

Properties: f, $g \in \mathscr{F}$ *nonnegative.*

(1) If $f = 0$, μ-*a.e.*, then $\displaystyle\int f d\mu = 0$.

(2) If $\mu(f > 0) > 0$, then $\displaystyle\int f d\mu > 0$.

(3) If $\displaystyle\int f d\mu < \infty$, then $f < \infty$, μ-*a.e.*

(4) If $f \leqslant g$, μ-*a.e.*, then $\displaystyle\int f d\mu \leqslant \int g d\mu$.

(5) If $f = g$, μ-*a.e.*, then $\displaystyle\int f d\mu = \int g d\mu$.

Thm 27 (Monotone convergence Theorem)**.** *If* $0 \leqslant f_n \uparrow f$, μ-*a.e.*, *then* $0 \leqslant \displaystyle\int f_n d\mu \uparrow \int f d\mu$.

Proof. **1**. First prove it under the assumption that

$$0 \leqslant f_n(x) \uparrow f(x),\ \forall x.$$

Integration is monotonic, so $\displaystyle\int f_n d\mu \leqslant \int f d\mu$. It remains to show

$$\lim_n \int f_n d\mu \geqslant \int f d\mu \tag{5.3}$$

or

$$\lim_n \int f_n d\mu \geqslant S = \sum_{i=1}^{m} c_i \mu(A_i)$$

for any finite measurable partition $\{A_i : i = 1, ..., m\}$ and $c_i = \inf_{A_i} f$.

For such a partition, assume that the sum S, c_i and $\mu(A_i)$ are all finite. Fix $\alpha < 1$, define

$$A_{i,n} = \{x \in A_i : f_n(x) > \alpha c_i\}.$$

Since $f_n \uparrow f$, $A_{i,n} \uparrow A_i$. Consider the *measurable* partition

$$\{A_{i,n} : i = 1, ..., m\} \cup \left\{ \left(\bigcup_{i=1}^{m} A_{i,n} \right)^c \right\}.$$

Then

$$\int f_n d\mu \geqslant \sum_{i=1}^{m} \alpha c_i \mu(A_{i,n}).$$

Let $n \to \infty$ and use continuity from below,

$$\lim_n \int f_n d\mu \geqslant \sum_{i=1}^{m} \alpha c_i \mu(A_i).$$

Finally let $\alpha \to 1$, (5.3) is proved.

Now suppose S is finite but not all of c_i, $\mu(A_i)$. Then $c_i\mu(A_i)$, $i = 1, ..., m$ are finite. c_i or $\mu(A_i)$ may be infinity, but then $c_i\mu(A_i)$ must be zero. Use the adjusted parition $\{A_i : c_i\mu(A_i) > 0\} \cup \{\text{complement}\}$.

Lastly suppose S is infinite. Then there is some i_0, $c_{i_0}\mu(A_{i_0}) = \infty$, i.e., $c_{i_0} > 0$, $\mu(A_{i_0}) > 0$ and at least one of them is ∞. In this case

$$\int f d\mu = \infty.$$

To prove (5.3), let a, b satisfy

$$0 < a < c_{i_0} \leqslant \infty,\ 0 < b < \mu(A_{i_0}) \leqslant \infty.$$

Define

$$A_{i_0,n} = \{x \in A_{i_0} : f_n(x) > a\}.$$

Since $f_n \uparrow f$, $A_{i_0,n} \uparrow A_{i_0}$ and $\mu(A_{i_0,n}) > b$ for n larger than some $n_{a,b}$. For the partition $\{A_{i_0,n}, A_{i_0,n}^c\}$, we have

$$\int f_n d\mu \geqslant a\mu(A_{i_0,n}) > ab, \forall n > n_{a,b}.$$

Let $a \to \infty$ if $c_{i_0} = \infty$, $b \to \infty$ if $\mu(A_{i_0,n}) = \infty$, we get

$$\lim_n \int f_n d\mu = \infty.$$

2. If $0 \leqslant f_n \uparrow f$ on A with $\mu(A^c) = 0$, then $0 \leqslant f_n 1_A \uparrow f 1_A$ holds everywhere. Then apply step **1**. □

5.2. 可测函数积分. $f \in \mathscr{F}$ with values in $[-\infty, \infty]$,

$$\int f d\mu \triangleq \int f^+ d\mu - \int f^- d\mu.$$

f is said to be integrable if $\displaystyle\int f^+ d\mu, \int f^- d\mu$ are finite. So f integrable iff $|f|$ integrable.

Properties: f, $g \in \mathscr{F}$ integrable.

(1) If $f \leqslant g$, μ-*a.e.*, then $\int f d\mu \leqslant \int g d\mu$.

(2) If $\alpha, \beta \in \mathbb{R}$, then $\alpha f + \beta g$ is integrable,

$$\int (\alpha f + \beta g) d\mu = \alpha \int f d\mu + \beta \int g d\mu.$$

Example 7. *Let* $E = \{1, 2, 3, ...\}$, $\mathscr{F} = \{$*all subsets of* $E\}$, $\mu =$ *counting measure. A function on* E *is a sequence* $x_1, x_2, ...$ *Any function is* $\mathscr{F}$*-measurable.* $\{x_k : k = 1, 2, ...\}$ *is* μ*-integrable if and only if* $\sum_{k=1}^{\infty} |x_k|$ *converges. When* μ*-integrable,*

$$\sum_{k=1}^{\infty} |x_k| = \sum_{k=1}^{\infty} x_k^+ - \sum_{k=1}^{\infty} x_k^-.$$

The function $x_k = (-1)^{k+1}/k$, $k = 1, 2, ...$ *is not* μ*-integrable, although*

$$\lim_m \sum_{k=1}^{m} (-1)^{k+1} \frac{1}{k} = \ln 2.$$

Thm 28 (Fatou's lemma)**.** *Given f_n measurable.*
(1) *If g integrable, $f_n \geqslant g$, μ-a.e, then $\liminf_n f_n$ is integrable and*

$$\int \liminf_n f_n d\mu \leqslant \liminf_n \int f_n d\mu.$$

(1) *If g integrable, $f_n \leqslant g$, μ-a.e, then $\limsup_n f_n$ is integrable and*

$$\limsup_n \int f_n d\mu \leqslant \int \limsup_n f_n d\mu.$$

Thm 29 (Lebesgue's dominated convergence theorem)**.** *Given g nonnegative integrable, $|f_n| \leqslant g$, μ-a.e.. If $f_n \xrightarrow{a.e.} f$ or $f_n \xrightarrow{\mu} f$, then*

$$\int f_n d\mu \longrightarrow \int f d\mu.$$

The following is a generalized dominated convergence theorem.

Thm 30. *Given g_n nonnegative integrable, $|f_n| \leqslant g_n$, μ-a.e. with $g_n \xrightarrow{a.e.} g$ and $\int g_n d\mu \longrightarrow \int g d\mu$. If $f_n \xrightarrow{a.e.} f$ or $f_n \xrightarrow{\mu} f$, then*

$$\int f_n d\mu \longrightarrow \int f d\mu.$$

Example 8 (Weierstrass M-test)**.** *If $|x_{n,m}| \leqslant M_m$, $\sum_{m=1}^{\infty} M_m < \infty$, $\lim_n x_{n,m} = x_m$ for each m. Then*

$$\lim_n \sum_{m=1}^{\infty} x_{n,m} = \sum_{m=1}^{\infty} x_m.$$

Example 9 (Bounded convergence theorem)**.** *Suppose μ is finite, $M > 0$. $|f_n| \leqslant M$, μ-a.e.. If $f_n \xrightarrow{a.e.} f$ or $f_n \xrightarrow{\mu} f$, then*

$$\int f_n d\mu \longrightarrow \int f d\mu.$$

Example 10. *If* $f_n \geqslant 0$ *or* $\sum_{n=1}^{\infty} \int |f_n| d\mu < \infty$, *then*

$$\int \sum_{n=1}^{\infty} f_n d\mu = \sum_{n=1}^{\infty} \int f_n d\mu.$$

From this we get

Example 11. *If* $x_{n,m} \geqslant 0$ *or* $\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} |x_{n,m}| < \infty$, *then*

$$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} x_{n,m} = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} x_{n,m}.$$

Example 12 (Abel's theorem)**.** *Suppose that the series* $\sum_{k=1}^{\infty} |c_k|$ *is convergent. Then*

$$\lim_{x \to 1-} \sum_{k=1}^{\infty} c_k x^k = \sum_{k=1}^{\infty} c_k.$$

5.3. Change of variables. $(E_1, \mathscr{F}_1)$, $(E_2, \mathscr{F}_2)$ are measurable spaces, μ is a measure on $\mathscr{F}_1$. T is measurable mapping from $(E_1, \mathscr{F}_1)$ to $(E_2, \mathscr{F}_2)$. Define

$$\nu(B) = \mu\big(T^{-1}(B)\big), \ \forall B \in \mathscr{F}_2. \tag{5.4}$$

Then $\nu(B)$ is a measure on $\mathscr{F}_2$ and for any $f \in \mathscr{F}_2$,

$$\int_{E_2} f d\nu = \int_{E_1} f \circ T d\mu. \tag{5.5}$$

Note if $f = 1_B$, then $f \circ T(x) = 1_B(T(x)) = 1_{T^{-1}(B)}(x)$, since $T(x) \in B$ iff $x \in T^{-1}(B)$. So in this case (5.5) reduces to (5.4).

6. L_p 空间

6.1. Inequlities.

Lemma 31 (Jensen's inequality). *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space with $\mu(\Omega) = 1$, X a μ-integrable function on Ω, φ convex on $\mathbb{R}$. Then*

$$\varphi\Big(\int_\Omega X d\mu\Big) \leqslant \int_\Omega \varphi(X) d\mu. \tag{6.1}$$

Equality holds iff φ is linear on some convex set $A \subset \mathbb{R}$ with $\mu\big(X^{-1}A\big) = 1$.

Proof. Denote by μ_X the induced measure of X on $\mathbb{R}$ (ref section 5.3), then (6.1) is equivalent to

$$\varphi\left(\int_{\mathbb{R}} x d\mu_X\right) \leqslant \int_{\mathbb{R}} \varphi(x) d\mu_X \tag{6.2}$$

(Apply (5.5) with $f(x) = x$, $T = X$). It is enough to prove (6.2).

1. Denote $\bar{x} = \displaystyle\int_{\mathbb{R}} x d\mu_X$. Since φ is convex, there is a supporting line $L(x) = ax + b$ through $\bar{x}$, i.e. $L(\bar{x}) = \varphi(\bar{x})$ and

$$L(x) \leqslant \varphi(x), \ \forall x.$$

Then

$$\int_{\mathbb{R}} L(x) d\mu_X \leqslant \int_{\mathbb{R}} \varphi(x) d\mu_X. \tag{6.3}$$

The LHS equals $\varphi\left(\displaystyle\int_{\mathbb{R}} x d\mu_X\right)$, hence (6.2) follows.

2. Suppose the equality in (6.2) holds, then by the above computation

$$\int_{\mathbb{R}} [\varphi(x) - L(x)] d\mu_X = 0.$$

The integrand is nonnegative, so the measurable set

$$A = \{x \in \mathbb{R} : \varphi(x) - L(x) = 0\}$$

has full measure, i.e. $\mu_X(A) = 1$. Moreover the set A is convex (verify directly!). On the other hand, if φ is linear on some convex $A \subset \mathbb{R}$ with $\mu\big(X^{-1}A\big) = 1$, then $\mu_X(A) = 1$,

$$\int_{\mathbb{R}} L(x) d\mu_X = \int_A L(x) d\mu_X, \quad \int_{\mathbb{R}} \varphi(x) d\mu_X = \int_A \varphi(x) d\mu_X.$$

Hence by (6.3),

$$\int_A [\varphi(X) - L(X)] d\mu \geqslant 0.$$

But the integrand $\varphi - L$ is nonnegative and linear on A. Since $A \subset \mathbb{R}$ is convex, it must be an interval. So the above integral is zero, hence the equality of (6.2) holds. $\square$

Notice that Lemma 31 does not require $\varphi(X)$ to be μ-integrable. From (6.3) it is clear that either $\int_\Omega \varphi(X)d\mu$ exists or equals infinity, in the latter case (6.1) trivially holds.

LEMMA 32. $a, b \in \mathbb{R}$, $1 \leqslant p < \infty$,

$$|a+b|^p \leqslant 2^{p-1}(|a|^p + |b|^p).$$

PROOF. Apply Jensen's inequality with $\varphi(x) = |x|^p$,

$$\left|\frac{a+b}{2}\right|^p \leqslant \frac{|a|^p + |b|^p}{2}.$$

$\square$

LEMMA 33 (Young's inequality). $a, b \geqslant 0$, $1 < p, q < \infty$, $\frac{1}{p} + \frac{1}{q} = 1$,

$$a^{1/p} b^{1/q} \leqslant \frac{a}{p} + \frac{b}{q}.$$

Equal iff $a = b$.

Proof. The inequality holds if $ab = 0$. In this case equality holds iff $a = b = 0$. Now suppose $ab > 0$. Apply Jensen's inequality with $\varphi(x) = -\ln x$,

$$-\ln\left(\frac{a}{p} + \frac{b}{q}\right) \leqslant -\frac{1}{p}\ln a - \frac{1}{q}\ln b.$$

Since φ is strictly convex (can touch a linear function at exactly one point), equality holds iff $a = b$. $\square$

$(E, \mathscr{F}, \mu)$ is a measure space in the following definitions.

Def 23. $p = 1$, *let*

$$L_1 \triangleq \{f \in \mathscr{F} : |f| \text{ is } \mu\text{-integrable}\}$$

and

$$\|f\|_1 = \|f\|_{L_1} = \int |f| d\mu.$$

Def 24. $1 < p < \infty$, *let*

$$L_p \triangleq \{f \in \mathscr{F} : |f|^p \in L_1\}$$

and

$$\|f\|_p = \|f\|_{L_p} = \left(\int |f|^p d\mu\right)^{1/p}.$$

Def 25. $p = \infty$, *let*

$$L_\infty \triangleq \{f \in \mathscr{F} : \textit{there is } C > 0 \textit{ such that } |f| \leqslant C,\ \textit{a.e.}\}$$

and

$$\|f\|_\infty = \|f\|_{L_\infty} = \inf\{C : |f| \leqslant C,\ \textit{a.e.}\}.$$

We could have written $L_p(\mu)$ to emphasize the dependence of the spaces L_p on the measure μ. But, when no ambiguity arises from the contexts, we will simply drop μ from the notation.

Thm 34 (Hölder inequality)**.** $1 \leqslant p,\ q \leqslant \infty$, $\dfrac{1}{p} + \dfrac{1}{q} = 1$, $f \in L_p$, $g \in L_q$, *then* $fg \in L_1$ *and*

$$\|fg\|_1 \leqslant \|f\|_p \|g\|_q. \tag{6.4}$$

If $p = 1$, *equality iff* $|g| = \|g\|_\infty$, *a.e. on the set where* $f \neq 0$.

If $p=\infty$, equality iff $|f| = \|f\|_\infty$, a.e. on the set where $g \neq 0$.

If $1<p<\infty$, equality iff there are nonnegative constants α, β such that $(\alpha, \beta) \neq (0,0)$, $\alpha|f|^p = \beta|g|^q$, a.e.

PROOF. **1**. The inequality easily follows if $p=1$ or $p=\infty$. To see the equaliy, suppose $p=1$, then $q=\infty$. (6.4) is equivalent to

$$\int |f|(\|g\|_\infty - |g|) \geqslant 0.$$

It is equality iff $|g| = \|g\|_\infty$, *a.e.* on the set where $f \neq 0$.

2. Suppose $1<p$, $q<\infty$. The conclusion is obvious if $\|f\|_p = 0$ or $\|g\|_q = 0$. Hence we assume that $0 < \|f\|_p$, $\|g\|_q < \infty$. Using Young's inequality with

$$a = \left(\frac{|f|}{\|f\|_p}\right)^p, \ b = \left(\frac{|g|}{\|g\|_q}\right)^q,$$

we have

$$\frac{|fg|}{\|f\|_p\|g\|_q} \leqslant \frac{1}{p}\left(\frac{|f|}{\|f\|_p}\right)^p + \frac{1}{q}\left(\frac{|g|}{\|g\|_q}\right)^q, \ a.e.$$

Integrating on both sides gives

$$\int \frac{|fg|}{\|f\|_p\|g\|_q} d\mu \leqslant \frac{1}{p} + \frac{1}{q} = 1,$$

which is the desired inequality. The equality holds iff $a = b$, $a.e.$ i.e.,

$$\|g\|_q^q |f|^p = \|f\|_p^p |g|^q, \ a.e.$$

$\square$

A familiar case of Hölder inequality is the following.

Thm 35 (Cauchy–Schwarz inequality)**.** *f, $g \in L_2$, then $fg \in L_1$ and*

$$\|fg\|_1 \leqslant \|f\|_2\|g\|_2.$$

Thm 36 (Minkowski inequality)**.** *$1 \leqslant p \leqslant \infty$, f, $g \in L_p$, then $f+g \in L_p$ and*

$$\|f+g\|_p \leqslant \|f\|_p + \|g\|_p. \tag{6.5}$$

If $p=1$ or $p=\infty$, equality iff $fg \geqslant 0$, a.e..

If $1<p<\infty$, equality iff there are nonnegative constants α, β such that $(\alpha,\beta) \neq (0,0)$, $\alpha f = \beta g$, a.e.

Proof. **1**. The case $p=1$ or $p=\infty$ is immediate.

2. Suppose $1<p<\infty$. Let $q>1$, $\dfrac{1}{p}+\dfrac{1}{q}=1$. By Hölder inequality

$$\begin{aligned}
\|f+g\|_p^p = \int |f+g||f+g|^{p-1} &\leqslant_{(e1)} \int |f||f+g|^{p-1} + \int |g||f+g|^{p-1} \\
&\leqslant_{(e2)} \|f\|_p \big\||f+g|^{p-1}\big\|_q + \|g\|_p \big\||f+g|^{p-1}\big\|_q \\
&= \|f\|_p \|f+g\|_p^{p-1} + \|g\|_p \|f+g\|_p^{p-1}
\end{aligned}$$

Here

$$\left\||f+g|^{p-1}\right\|_q=\left(\int\left(|f+g|^{p-1}\right)^q\right)^{1/q}=\left(\int|f+g|^p\right)^{1/q}$$
$$=\|f+g\|_p^{p/q}=\|f+g\|_p^{p-1}.$$

$(e1)$ is equality iff $fg\geqslant 0$, *a.e.*, $(e2)$ is equality iff there are nonegative constants a, b, c, d such that $(a,b)\neq(0,0)$, $(c,d)\neq(0,0)$,

$$a|f|^p=b\left(|f+g|^{p-1}\right)^q,\ c|g|^p=d\left(|f+g|^{p-1}\right)^q,\ a.e.$$

Hence

$$a|f|=b|f+g|,\ c|g|=d|f+g|,\ a.e.$$

The conclusion follows by combining the equality conditions of $(e1)(e2)$.

$\square$

Def 26. $0<p<1$, *let*

$$L_p\triangleq\left\{f\in\mathscr{F}:\int|f|^p d\mu<\infty\right\}$$

and

$$\|f\|_p = \int |f|^p d\mu.$$

LEMMA 37. *Let* a, $b \in \mathbb{R}$, $0 < p < 1$. $|a+b|^p \leqslant |a|^p + |b|^p$.

PROOF. Since $||a| + |b||^p \leqslant |a|^p + |b|^p$ implies the desired inequality, we assume w.l.g. that a, b are of the same sign. Suppose $a \neq 0$, otherwise there is nothing to prove. Finally it suffices to show that

$$(1+s)^p \leqslant 1 + s^p, \;\; s \geqslant 0,$$

which is verified by elementary calculus. ☐

Lemma 32 and Lemma 37 can be merged into the compact form,

$$|a+b|^p \leqslant C_p(|a|^p + |b|^p), \;\; 0 < p < \infty, \tag{6.6}$$

where $C_p = 2^{p-1} \vee 1$.

Thm 38. $0 < p < 1$, $\|f+g\|_p \leqslant \|f\|_p + \|g\|_p$.

6.2. Completeness.

Thm 39. *Let* $0 < p \leqslant \infty$, L_p *is complete.*

Proof for $p = \infty$. Let $f_n \in L_\infty$. Suppose that f_n is Cauchy. Given $k \geqslant 1$, there is n_k such that

$$\|f_m - f_n\|_\infty \leqslant \frac{1}{k}, \ \forall m, n > n_k.$$

Hence there is a null set[1] A_k such that

$$|f_m - f_n| \leqslant \frac{1}{k}, \ \forall x \in A_k^c, \ m, n > n_k.$$

Then $A = \bigcup_{k=1}^{\infty} A_k$ is a null set and $f_n(x)$ is Cauchy for each $x \in A^c$. Hence there exist f, $f_n \to f$ for $x \in A^c$. Let $m \to \infty$ in the above inequality we get

$$|f_n - f| \leqslant \frac{1}{k}, \ \forall x \in A^c, \ n > n_k.$$

[1] A null set is a measurable set with measure zero.

So $f \in L_\infty$ and

$$\|f_n - f\|_\infty \leqslant \frac{1}{k}, \ \forall n > n_k.$$

Therefore f_n converges to f in L_∞. $\square$

Proof for $0 < p < \infty$. Let $f_n \in L_p$. Suppose that f_n is Cauchy in L_p,

$$\lim_{m,n\to\infty} \|f_m - f_n\|_p = 0. \tag{6.7}$$

We intend to show that $\lim_{n\to\infty} \|f_n - f\|_p = 0$ for some $f \in L_p$. Owing to (6.7), we have a subsequence $n_k \to \infty$ so that

$$\left\|f_{n_{k+1}} - f_{n_k}\right\|_p < \frac{1}{2^k}. \tag{6.8}$$

We claim that

(a) there is $h \in L_p$ such that $|f_{n_k}| \leqslant h$, $a.e.$
(b) $\lim_k f_{n_k} \to f$, $a.e.$ for some $f \in L_p$.
(c) $\lim_k \|f_{n_k} - f\|_p = 0$.

The conclusion of the Theorem clearly follows once (c) is proved, since a Cauchy sequence converges iff it has a convergent subsequence. Let

$$g_k = \sum_{i=1}^{k} \left| f_{n_{i+1}} - f_{n_i} \right|, \; g = \sum_{i=1}^{\infty} \left| f_{n_{i+1}} - f_{n_i} \right|.$$

Then $0 \leqslant g_k \uparrow g$ and $\|g_k\|_p \leqslant 1$ by (6.8) (Theorem 36 or Theorem 38). Using monontone convergence theorem,

$$\int g^p d\mu = \lim_k \int (g_k)^p d\mu \leqslant 1.$$

This shows $g \in L_p$ and that $g < \infty$, $a.e.$ Therefore

$$f_{n_k} = f_{n_1} + \sum_{i=1}^{k} \left(f_{n_{i+1}} - f_{n_i} \right)$$

converges almost everywhere to some measurable function f and

$$|f_{n_k}| \leqslant |f_{n_1}| + g.$$

Let $k \to \infty$, we have

$$|f| \leqslant |f_{n_1}| + g, \ a.e.$$

hence $f \in L_p$. $(a)(b)$ follows with $h = |f_{n_1}| + g$. By inequality (6.6),

$$\begin{aligned} |f_{n_k} - f|^p &\leqslant C_p(|f_{n_k}|^p + |f|^p) \leqslant C_p(||f_{n_1}| + g|^p + |f|^p) \\ &\leqslant C_p(C_p(|f_{n_1}|^p + g^p) + |f|^p). \end{aligned}$$

Therefore (c) is a result of the dominated convergence theorem. $\square$

COROLLARY 1. (1) $0 < p < 1$, L_p *is a complete metric space.*
(2) $1 \leqslant p \leqslant \infty$, L_p *is a Banach space.*

6.3. L_p and weak convergence.

Thm 40. *Let* $0 < p < \infty$, $f_n \in L_p$, $f \in L_p$.
(1) $f_n \xrightarrow{L_p} f \implies f_n \xrightarrow{\mu} f$ *and* $\|f_n\|_p \to \|f\|_p$.
(2) $f_n \xrightarrow{a.e.} f$ *or* $f_n \xrightarrow{\mu} f$, *then*

$$\|f_n\|_p \to \|f\|_p \iff f_n \xrightarrow{L_p} f.$$

Proof. **1**. To prove (1), use Markov inequality

$$\mu(|f_n - f| > \varepsilon) \leqslant \frac{1}{\varepsilon^p}\|f_n - f\|_p^p.$$

and the triangle inequality

$$\left|\|f_n\|_p - \|f\|_p\right| \leqslant \|f_n - f\|_p.$$

2. " $\Longleftarrow$ " of (2) is included in step **1**.

3. " $\Longrightarrow$ " of (2). In view of Theorem 20, it is enough to prove the case where $f_n \xrightarrow{a.e.} f$. Define

$$g_n = C_p(|f_n|^p + |f|^p) - |f_n - f|^p,$$

where $C_p = 2^{p-1} \vee 1$. Then $g_n \geqslant 0$ by inequality (6.6) and $\lim_n g_n = 2C_p|f|^p$, *a.e.* Using Fatou's lemma

$$\int 2C_p|f|^p d\mu = \int \lim_n g_n d\mu \leqslant \liminf_n \int g_n d\mu$$
$$= \int 2C_p|f|^p d\mu - \limsup_n \int |f_n - f|^p.$$

Canceling $\int 2C_p|f|^p d\mu$ from both side gives

$$\lim_n \int |f_n - f|^p = 0.$$

$\square$

Def 27. *$(E, \mathscr{F}, \mu)$ is a measure space. $1 \leqslant p < \infty$, $\frac{1}{p} + \frac{1}{q} = 1$, f_n converges weakly to f in L_p, denoted by $f_n \xrightarrow{w\text{-}L_p} f$, if*

$$\lim_n \int f_n g d\mu = \int f g d\mu, \ \forall g \in L_q.$$

μ is additionally assumed to be σ-finite if $p = 1$.

Thm 41. $1 \leqslant p < \infty$. *$f_n \xrightarrow{L_p} f$ implies $f_n \xrightarrow{w\text{-}L_p} f$.*

PROOF. By Hölder inequality (Theorem 34), $\forall g \in L_q$, q conjugate to p,

$$\int |f_n - f||g| d\mu \leqslant \|f_n - f\|_p \|g\|_q.$$

$\square$

Thm 42. *$(E, \mathscr{F}, \mu)$ is a measure space. Let $1 < p < \infty$, $\{f_n\}$ bounded in L_p. If $f_n \xrightarrow{a.e.} f$ or $f_n \xrightarrow{\mu} f$ for some measurable f, then $f \in L_p$ and $f_n \xrightarrow{w\text{-}L_p} f$.*

PROOF. Let $g \in L_q$, q conjugate to p. As before, it is enough to prove it for $f_n \xrightarrow{a.e.} f$.

1. $f \in L_p$ is a consequence of Fatou's lemma,

$$\int |f|^p d\mu = \int \lim_n |f_n|^p d\mu \leqslant \liminf_n \int |f_n|^p d\mu \leqslant \sup_n \|f_n\|_{L_p} < \infty.$$

It follows that $\{f_n - f\}$ is bounded in L_p.

2. Fix $\varepsilon > 0$, let $\delta > 0$, define $A_\delta = \{x \in E : \delta \leqslant |g|^q \leqslant 1/\delta\}$ and write

$$\int |f_n - f||g| d\mu = \int_{A_\delta \cap B} + \int_{A_\delta \cap B^c} + \int_{A_\delta^c} .$$

Choose δ small so that

$$\int_{A_\delta^c} \leqslant \|f_n - f\|_p \big\|g 1_{A_\delta^c}\big\|_q < \frac{\varepsilon}{3}.$$

With δ fixed, we have

$$\int_{A_\delta \cap B^c} \leqslant \|f_n - f\|_p \|g 1_{A_\delta \cap B^c}\|_q < \frac{\varepsilon}{3},$$

as soon as $B \subset A_\delta$ is such that $\mu(A_\delta \cap B^c)$ is smaller than some ε'.

Note $|g| \leqslant 1/\delta^{1/q}$ on A_δ. Since $\mu(A_\delta)$ is finite by Markov inequality, so a subset $B \subset A_\delta$ can be chosen so that $\mu(A_\delta \cap B^c) < \varepsilon'$ and $|f_n - f|$

converges uniformly to 0 on $A_\delta \cap B$ (Theorem 19). Hence for large n,

$$\int_{A_\delta \cap B} \leqslant \frac{1}{\delta^{1/q}} \int_{A_\delta \cap B} |f_n - f| d\mu < \frac{\varepsilon}{3}.$$

$\square$

Note the above proof does not get through if $p = 1$ (so that $q = \infty$). The example below demonstates, in general, Theorem 42 does not for $p = 1$.

Example 13. *$E = (0, 1)$ with the usual Lebesgue measure, $f_n = n1_{(0,1/n)}$. Clearly $\|f_n\|_1 = 1$, $f_n \xrightarrow{\mu} f = 0$. But with $g = 1 \in L_\infty$, $\lim_n \int f_n g d\mu = 1 \neq 0 = \int f g d\mu$, hence $f_n \xrightarrow{w\text{-}L_1} f$ does not hold.*

However we have

Thm 43. *$(E, \mathscr{F}, \mu)$ is a measure space. Let $\{f_n\} \in L_1$. Suppose $f_n \xrightarrow{a.e.} f$ or $f_n \xrightarrow{\mu} f$. Then*

$$f \in L_1,\ \|f_n\|_1 \to \|f\|_1 \iff f_n \xrightarrow{L_1} f.$$

Either of them gives $\int_A f_n d\mu \to \int_A f d\mu,\ \forall A \in \mathscr{F}$.

PROOF. The first conclusion is contained in Theorem 40. So $f_n \xrightarrow{w\text{-}L_2} f$ by Theorem 41. To complete the proof, take $1_A \in L_\infty$ as test function. $\square$

6.4. Uniform integrability. Let $(E, \mathscr{F}, \mu)$ be a measure space.

Def 28. $\mathcal{H} = \{f_t : t \in T\}$ *is uniformly integrable if*

$$\lim_{a\to\infty} \sup_{f\in\mathcal{H}} \int_{\{|f|\geqslant a\}} |f| d\mu = 0. \tag{6.9}$$

Def 29. $\mathcal{H} = \{f_t : t \in T\}$ *is absolutely continuous if,* $\forall \varepsilon > 0$, *there is* $\delta > 0$ *so that*

$$\sup_{f\in\mathcal{H}} \int 1_A |f| d\mu < \varepsilon \textit{ for any } A \textit{ with } \mu(A) < \delta.$$

Thm 44. *Suppose $(E, \mathscr{F}, \mu)$ is a measure space with μ finite. $\mathcal{H} = \{f_t : t \in T\}$ is uniformly integrable if and only if $\mathcal{H}$ is absolutely continuous and bounded in L_1.*

PROOF. **1**. If $\mathcal{H}$ is uniformly integrable, $\forall \varepsilon > 0$, there is $a_0 > 0$ so that

$$\sup_{f\in\mathcal{H}} \int_{\{|f|\geqslant a\}} |f| d\mu \leqslant \frac{\varepsilon}{2},\ \forall a \geqslant a_0.$$

For any measurable A, $a \geqslant a_0$,

$$\begin{aligned}\sup_{f\in\mathcal{H}} \int 1_A|f|d\mu &\leqslant \sup_{f\in\mathcal{H}} \int_{\{|f|<a\}} 1_A|f|d\mu + \sup_{f\in\mathcal{H}} \int_{\{|f|\geqslant a\}} 1_A|f|d\mu \\ &\leqslant a\mu(A) + \sup_{f\in\mathcal{H}} \int_{\{|f|\geqslant a\}} |f|d\mu \leqslant a\mu(A) + \frac{\varepsilon}{2}.\end{aligned}$$

That $\mathcal{H}$ is bounded in L_1 follows by setting $A = E$ and using the fact that μ is finite. Fix $a \geqslant a_0$. For any A with $\mu(A) \leqslant \varepsilon/(2a)$, we get that $\sup_{f\in\mathcal{H}} \int 1_A|f|d\mu$ is bounded from above by ε, hence the absolute continuity.

2. Suppose that $\mathcal{H}$ is absolutely continuous and bounded in L_1. Denote the uniform L_1 bound of $\mathcal{H}$ by M. By Markov inequality, $\forall a > 0$,

$$\mu(|f| > a) \leqslant \frac{1}{a}\int |f| d\mu \leqslant \frac{1}{a}M, \ \forall f \in \mathcal{H}.$$

$\forall \varepsilon > 0$, by absolute continuity, $\sup_{f\in\mathcal{H}} \int 1_A |f| d\mu < \varepsilon$ as soon as $\mu(A)$ is less than some $\delta > 0$. Fix a with $M/a < \delta$. Then setting $A = \mu(|f| > a)$ gives the uniform integrability. $\square$

Thm 45 (Vitali convergence theorem)**.** *Suppose that μ is finite, $f_n \xrightarrow{a.e.} f$ or $f_n \xrightarrow{\mu} f$.*

(1) *If $\{f_n\}$ is uniformly integrable, then $f \in L_1$ and*

$$\int f_n d\mu \to \int f d\mu. \tag{6.10}$$

(2) *f_n, f are nonnegative integrable, then* (6.10) *implies that $\{f_n\}$ is uniformly integrable.*

PROOF. The proof is given for $f_n \xrightarrow{a.e.} f$.

1. If f_n is uniformly integrable, then f is integrable by Theorem 44 and Fatou's lemma. Define

$$f_{n,a} = 1_{\{|f_n|<a\}} f_n, \;\; f_a = 1_{\{|f|<a\}} f.$$

It follows that $f_{n,a} \to f_a$, $a.e.$ provided $\mu(|f| = a) = 0$. By bounded dominated convergence,

$$\int f_{n,a} d\mu \to \int f_a d\mu.$$

Writing

$$\int_{\{|f_n| \geqslant a\}} f_n d\mu = \int f_n d\mu - \int f_{n,a} d\mu \tag{6.11}$$

and

$$\int_{\{|f| \geqslant a\}} f d\mu = \int f d\mu - \int f_a d\mu, \tag{6.12}$$

we see that

$$\begin{aligned}
&\limsup_n \left| \int f_n d\mu - \int f d\mu \right| \\
&\leqslant \limsup_n \left| \int f_{n,a} d\mu - \int f_a d\mu \right| + \sup_n \int_{\{|f_n| \geqslant a\}} |f_n| d\mu + \int_{\{|f| \geqslant a\}} |f| d\mu \\
&= \sup_n \int_{\{|f_n| \geqslant a\}} |f_n| d\mu + \int_{\{|f| \geqslant a\}} |f| d\mu.
\end{aligned}$$

Note $\mu(|f| = a) = 0$ for all but countably many a. Sending $a \to \infty$ proves (6.10).

2. Suppose f_n, f are nonnegative integrable and (6.10) holds. Write

$$\int_{\{|f_n| \geqslant a\}} f_n d\mu = \int_{\{|f| \geqslant a\}} f d\mu + \left(\int_{\{|f_n| \geqslant a\}} f_n d\mu - \int_{\{|f| \geqslant a\}} f d\mu \right).$$

Since f is integrable, the first term is less than $\varepsilon/2$ when a is larger than some a_0. If $\mu(|f| = a) = 0$, (6.11) and (6.12) indicate the term in the bracket is also less than $\varepsilon/2$ when n is larger than some n_0.

Therefore,

$$\sup_{n>n_0} \int_{\{|f_n| \geqslant a\}} f_n d\mu \leqslant \varepsilon, \ \forall a > a_0 \text{ with } \mu(|f| = a) = 0.$$

Since the finite family $\{f_1, ..., f_{n_0}\}$ is uniformly integrable, the uniform integrability of $\{f_n, n \geqslant 1\}$ follows. $\square$

Additional details on the proof of Theorem 45. Suppose $|f_n(x)| \to |f(x)| < a$. Then for large n, $|f_n(x)| < a$. So $1_{\{|f_n|<a\}}$ and $1_{\{|f|<a\}}$ are both equal to 1, it follows $f_{n,a} \to f_a$ at x. The same is true for x with $|f(x)| > a$. If $|f(x)| = a \neq 0$, then $f_{n,a}(x) \to f_a(x)$ may not happen, since in this case $f_a(x) = 0$ while there could be a subsequence n_k with $f_{n_k}(x) < a$ so that

$$f_{n_k,a}(x) = f_{n_k}(x) \to f(x) \neq 0.$$

But if the set $\{x : |f(x)| = a\}$ has zero μ-measure, then $f_{n,a} \to f_a$, $a.e.$ Fortunately the set of a for which $\mu(|f| = a)$ is not zero is at most countable. Indeed, let

$$F(x) = \mu(|f| \leqslant x).$$

Then $F(x)$ is non-decreasing, hence has at most countably many discontinuities. F is (right-continuous and thus) discontinuous at $x = a$ if and only if

$$\mu(|f| = a) = F(a) - F(a-) \neq 0.$$

This verifies that $\mu(|f| = a) = 0$ for all but countably many a.

COROLLARY 2. *Suppose that μ is finite, f_n, f are integrable. If $f_n \xrightarrow{a.e.} f$ or $f_n \xrightarrow{\mu} f$, then these are equivalent:*
(1) *$\{f_n\}$ is uniformly integrable;*
(2) $\displaystyle\int |f_n - f| d\mu \to 0$;
(3) $\displaystyle\int |f_n| d\mu \to \int |f| d\mu$.

6.5. Summary of various convergences.

$$
\begin{array}{ccccccc}
f_n \xrightarrow{\mu} f & \underset{\text{Thm } 20}{\overset{\text{sub subseq}}{\Longleftrightarrow}} & f_n \xrightarrow{a.u} f & & & & \\
\text{Markov} \Uparrow & & & & & & \\
f_n \xrightarrow{L_p} f & \begin{array}{c} \overset{\text{has a subseq}}{\Longrightarrow} \\ \underset{\|f_n\|_p \to \|f\|_p}{\Longleftarrow} \\ \text{Thm } 40 \end{array} & f_n \xrightarrow{a.e.} f & \begin{array}{c} \overset{\mu \text{ finite}}{\Longleftrightarrow} \\ \Longleftarrow \\ \text{Thm } 19 \end{array} & f_n \xrightarrow{a.u} f & & \\
& & & & \Downarrow \text{Thm } 18 & & \\
& & f_n \xrightarrow{d} f & \underset{\text{Thm } 24}{\overset{\text{in a prob space}}{\Longleftarrow}} & f_n \xrightarrow{\mu} f & &
\end{array}
$$

7. 概率空间的积分

7.1. Expected value. $(\Omega, \mathscr{F}, P)$ is a probability space, X a r.v.

Def 30. *Expectation, written* EX,

$$EX = \int X dP.$$

Suppose X is discrete, i.e., X takes values in a finite or infinitely countable *distinct* sequence $\{x_1, x_2, ...\}$. Then its expectation ($\int X dP$ computed according to (5.1)) equals

$$EX = \sum_i x_i P(X = x_i).$$

The mapping $i \mapsto P(X = x_i)$ is called the probability mass function of X. If $Y = g(X)$ for some measurable function g, then Y is discrete with values in, say, $\{y_1, y_2, ...\}$. The expectation of Y, computed in the same way as EX, is

$$EY = \sum_i y_i P(Y = y_i).$$

To calculate EY, we first need to find its probability mass function $i \mapsto P(Y = y_i)$. This can be complicated, and it is avoided by using the "*law of the unconscious statistician*",

$$EY = \sum_i g(x_i) P(X = x_i).$$

This turns out to be a change of variables formula (see also Theorem 48).

Thm 46 (Change of variables formula)**.** *Let $(\Omega, \mathscr{F}, P)$ be a probability space, X a r.v, and $g \in \mathscr{B}_{\mathbb{R}}$. If $g \geqslant 0$ or $\int_\Omega |g(X)| dP < \infty$, then*

$$Eg(X) = \int_\Omega g(X) dP = \int_{\mathbb{R}} g(x) d\mu_X. \tag{7.1}$$

*Here $\mu_X(A) = PX^{-1}(A) = P(X \in A)$, $\forall A \in \mathscr{B}_{\mathbb{R}}$ is the probability induced by X (Section 5.3), which will be called the **distribution** of X.*

Proof. **1**. The nonnegative case $g \geqslant 0$. If $g = 1_A$, then $g(X(\omega)) = 1_A(X(\omega)) = 1_{X^{-1}(A)}(\omega)$, so (7.1) reduces to the definition of μ_X. By linearity, (7.1) holds for simple functions. If g_n are simple functions such that $0 \leqslant g_n(x) \uparrow g(x)$, then $0 \leqslant g_n(X(\omega)) \uparrow g(X(\omega))$, then (7.1) follows by monontone convergence theorem.

2. The case $\displaystyle\int_\Omega |g(X)|dP < \infty$. Applying step **1** to $|g(X)|$ shows that g is integrable with respect to μ_X, hence the integrability of g^+, g^-, and (7.1) follows from subtracting $Eg^-(X) = \displaystyle\int_{\mathbb{R}} g^-(x)d\mu_X$ from $Eg^+(X) = \displaystyle\int_{\mathbb{R}} g^+(x)d\mu_X$. □

The probability μ_X equals (as a result of the uniqueness Theorem 10) the measure μ constructed from the distribution function F

of X : $\mu((a,b]) = F(b) - F(a)$, $\forall a, b$. The measure μ is called a Lebesgue-Stieltjes measure and its integral is the Lebesgue-Stieltjes integral (section 7.3). The above formula thus relates integral on a probability space to Lebesgue-Stieltjes integral over $\mathbb{R}$. The rightmost term of (7.1) is also written as $\displaystyle\int g dF$, i.e.

$$Eg(X) = \int_{\mathbb{R}} g(x) dF.$$

REMARK 2. *An implication of Theorem 46 is that the integration (e.g. the expection and variance) of a random variable is a distributional property, i.e., it depends on the random variable only through its distribution. This lays the basis for applying probability theory tools such as Skorohod Theorem (Theorem 26).*

Def 31. *Variance, written* $Var(X)$,

$$Var(X) = \int (X - EX)^2 dP = E(X - EX)^2.$$

It is easy to see that

$$\mathrm{Var}(X) = EX^2 - (EX)^2.$$

Def 32. *k-th moment, $k = 1, 2, ...,$*

$$E\big(X^k\big) = \int X^k dP.$$

Example 14 (Bernoulli distribution)**.** *Let $0 < p < 1$. $X \sim$ Bernoulli(p) if $P(X = 1) = p$, $P(X = 0) = 1 - p$. Then*

$$EX = 1 \cdot p + 0 \cdot (1 - p) = p.$$

$$Var(X) = EX^2 - (EX)^2 = p - p^2 = p(1 - p).$$

Example 15 (Poisson distribution)**.** *Let $\lambda > 0$. $X \sim Poisson(\lambda)$ if*

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, 1, 2, ...$$

Then

$$EX = \sum_{k=0}^{\infty} k \cdot e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} = \lambda.$$

$$E(X(X-1)) = \sum_{k=0}^{\infty} k(k-1) \cdot e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} = \lambda^2.$$

Hence $EX^2 = \lambda^2 + \lambda$, *and*

$$Var(X) = \left(\lambda^2 + \lambda\right) - \lambda^2 = \lambda.$$

Example 16 (Geometric distribution). *Repeatedly flip a coin with head probability* p *and stop only when the head appears. The number of tosses* X *has the distribution*

$$P(X = k) = (1-p)^{k-1} p, \ \ k = 1, 2, ..$$

The distribution of X *is called geometric, denoted by* $X \sim Geom(p)$,

$$EX = \frac{1}{p}, \ \ Var(X) = \frac{1}{p^2}.$$

7.2. Properties of expectation. X, Y are random variables. The following are immediate from section 6.1.

Jensen inequality: if X integrable, φ convex, then

$$\varphi(EX) \leqslant E\varphi(X).$$

Hölder inequality: if $p, q \geqslant 1, 1/p + 1/q = 1$, then

$$E|XY| \leqslant \|X\|_p \|Y\|_q.$$

Minkowski inequality: if $p \geqslant 1$, then

$$\|X + Y\|_p \leqslant \|X\|_p + \|Y\|_p.$$

Thm 47. $0 < s < t < \infty$, X *is a r.v. Then* $\|X\|_s \leqslant \|X\|_t$.

PROOF. By Hölder inequality with $p = \dfrac{t}{s}$, $q = \dfrac{t}{t-s}$,

$$\|X\|_s^s = E|X|^s \leqslant (E|X|^{sp})^{1/p} (E1^q)^{1/q} = \left(E|X|^t\right)^{s/t} = \|X\|_t^s.$$

$\square$

Example 17. *If* X *has* $EX^2 < \infty$, *then its expectation and variance exist, since* $E|X| \leqslant \|X\|_2 < \infty$, *and*

$$0 \leqslant Var(X) \leqslant EX^2.$$

7.3. Lebesgue-Stieltjes and Riemann-Stieltjes integrals. Let G be a **generalized distribution function**, i.e., nondecreasing, right-continuous on $\mathbb{R}$. There is a unique measure μ such that

$$\mu((a,b]) = G(b) - G(a), \ \forall a, b. \tag{7.2}$$

The measure μ constructed this way is called a **Lebesgue-Stieltjes measure**. Integration with respect to Lebesgue-Stieltjes measure is called **Lebesgue-Stieltjes integral**, denoted by $\displaystyle\int f d\mu$ or $\displaystyle\int f dG$.

REMARK 3. *Under suitable conditions (see below),* $\displaystyle\int f dG$ *may be interpreted as Riemann-Stieltjes integral. Since this does not provide anything new in the context of general measure theory,* $\displaystyle\int f dG$ *is best understood as a notional variant of* $\displaystyle\int f d\mu$*, and hence by convention (see (7.2))* $\displaystyle\int_a^b f dG$ *means* $\displaystyle\int_{(a,b]} f dG$ *or* $\displaystyle\int_{(a,b]} f d\mu$.

Here we recall a few facts about Riemann-Stieltjes integration. Let G be the function as in (7.2), f a bounded function on $[a, b]$. Corresponding to each partition $\mathcal{P} : a = x_0 < x_1 < \cdots < x_n = b$, we consider

$$L(\mathcal{P}, f) = \sum_{i=1}^{n} \inf_{x \in [x_{i-1}, x_i]} f(x) \Delta G_i, \ U(\mathcal{P}, f) = \sum_{i=1}^{n} \sup_{x \in [x_{i-1}, x_i]} f(x) \Delta G_i.$$

Here $\Delta G_i = G(x_i) - G(x_{i-1})$. Define

$$R_* f = \sup_{\mathcal{P}} L(\mathcal{P}, f), \ R^* f = \inf_{\mathcal{P}} U(\mathcal{P}, f).$$

If $R_* f = R^* f$, then f is Riemann-Stieltjes integrable with respect to G, the common value, written $(R\text{-}S) \displaystyle\int f$, is called the Riemann-Stieltjes integral. For simplicity we have omitted the dependence of the integral on G in the notations.

A sufficient condition for Riemann-Stieltjes integrability is this: Suppose f is bounded on $[a, b]$, has at most finitely many discontinuities, G is continuous at every point where f is discontinuous. Then f is Riemann-Stieltjes integrable with respect to G.

Example 18. *If $a < s < b$, f is bounded on $[a, b]$, continuous at s and $G(x) = 1_{[s,\infty)}(x)$. Then*

$$(R\text{-}S) \int_a^b f dG = f(s).$$

Indeed, consider paritions $\mathcal{P} = \{x_0, x_1, x_2, x_3\}$, $a = x_0$ and $x_1 < x_2 = s < x_3 = b$. Then $\Delta G_2 = 1$, $\Delta G_i = 0$ if $i \neq 2$,

$$L(\mathcal{P}, f) = \inf_{x \in [x_1, x_2]} f(x),\ U(\mathcal{P}, f) = \sup_{x \in [x_1, x_2]} f(x).$$

Since f is continuous at s, we see that $L(\mathcal{P}, f)$ and $U(\mathcal{P}, f)$ converge to $f(s)$ as $x_1 \to s$.

Thm 48. *Suppose* $c_n \geqslant 0$, $\sum c_n < \infty$, $\{s_n\}$ *is a sequence of distinct points in* (a, b), *and*

$$G(x) = \sum_{n=1}^{\infty} c_n 1_{[s_n,\infty)}(x).$$

If f *is continuous on* $[a, b]$, *then*

$$(R\text{-}S) \int_a^b f dG = \sum_{n=1}^{\infty} c_n f(s_n).$$

Proof. Exercise. □

If we denote by $L_* f$ the integral in (5.1) with the G-induced Lebesgue-Stieltjes measure in the role of μ, and by $L^* f$ the integral in (5.2). Then

$$R_* f \leqslant L_* f \leqslant L^* f \leqslant R^* f.$$

Therefore if, for instance, f is continuous on $[a, b]$, then it is Riemann-Stieltjes integrable, hence Lebesgue-Stieltjes integrable.

7.4. L_p convergence and uniform integrability.

Thm 49. $(\Omega, \mathscr{F}, P)$ *is a probability space,* $0 < p < \infty$, $X_n \in L_p$, $X \in \mathscr{F}$. *If* $X_n \xrightarrow{P} X$, *then these are equivalent:*

(1) $\{|X_n|^p\}$ *is uniformly integrable;*
(2) $X \in L_p$, $E(|X_n - X|^p) \to 0$;
(3) $X \in L_p$, $E(|X_n|^p) \to E(|X|^p)$.

Proof. **1**. Suppose that $\{|X_n|^p\}$ is uniformly integrable. Observe that $X \in L_p$ by Theorem 45, hence $\{|X_n - X|^p\}$ is uniformly integrable since $|X_n - X|^p \leqslant C_p(|X_n|^p + |X|^p)$ where $C_p = 2^{p-1} \vee 1$. Note also that $|X_n - X|^p \xrightarrow{P} 0$. Therefore (1) implies (2) is a consequence of Theorem 45 with $f_n = |X_n - X|^p$.

2. (2) implies (3) because $\left| \|X_n\|_p - \|X\|_p \right| \leqslant \|X_n - X\|_p$, $0 < p < \infty$ (Theorem 36, Theorem 38).

3. (3) implies (2) follows from an application of Theorem 45 with $f_n = |X_n|^p$. $\square$

We notice another criterion for uniform integrability, in addition to Theorem 44.

Lemma 50. *Let* $(\Omega, \mathscr{F}, P)$ *be a probability space,*

$$\mathcal{H} = \{X_t : t \in T,\ E|X_t| < \infty\}.$$

Suppose that $g \geqslant 0$ *is an increasing function on* $[0, \infty)$ *such that*

$$\lim_{s \to \infty} \frac{g(s)}{s} = \infty$$

and

$$\sup_{X \in \mathcal{H}} \int g(|X|) dP < \infty.$$

Then $\mathcal{H}$ *is uniformly integrable.*

Proof. $\forall \varepsilon > 0$. Fix $a > 0$ so that

$$\frac{1}{a} \sup_{X \in \mathcal{H}} \int g(|X|) dP < \varepsilon.$$

There is $s_0 > 0$ such that $g(s) \geqslant as$ for all $s \geqslant s_0$. Hence, $\forall X \in \mathcal{H}$, $s \geqslant s_0$,

$$\int_{\{|X| \geqslant s\}} |X| dP \leqslant \frac{1}{a} \int_{\{|X| \geqslant s\}} g(|X|) dP \leqslant \frac{1}{a} \sup_{X \in \mathcal{H}} \int g(|X|) dP < \varepsilon.$$

$\square$

8. 乘积测度空间

Let $(X, \mathscr{X}, \mu)$, $(Y, \mathscr{Y}, \nu)$ be measure spaces. The problem is to construct a **product measure** π on $X \times Y$ such that

$$\pi(A \times B) = \mu(A)\nu(B) \text{ for } A \in \mathscr{X},\ B \in \mathscr{Y}. \tag{8.1}$$

8.1. Product σ-field.

Def 33. *In the product space $X \times Y$, a **measurable rectangle** is a product of the form*

$$A \times B,\ A \in \mathscr{X},\ B \in \mathscr{Y}.$$

Let

$$\mathscr{S} = \{A \times B : A \in \mathscr{X},\ B \in \mathscr{Y}\} \tag{8.2}$$

*be the class of measurable rectangles on $X \times Y$. The **product σ-field** on $X \times Y$ is then defined as*

$$\mathscr{X} \times \mathscr{Y} = \sigma(\mathscr{S}).$$

The space $X \times Y$ equipped with this product σ-field is called ***product measurable space****.*

As the example below shows, the product σ-field is generally larger than the class of measurable rectangles.

Example 19. $\mathscr{B}(\mathbb{R}^2) = \mathscr{B}(\mathbb{R}) \times \mathscr{B}(\mathbb{R})$. *Here $\mathscr{B}(\mathbb{R}^2)$ is the usual σ-field on $\mathbb{R}^2$ generated by the class of products of one-dimensional intervals.*

Example 20. *If $A \times B \in \mathscr{S}$, then*

$$(A \times B)^c = A^c \times Y + A \times B^c \in \mathscr{S}.$$

From this it is easy to check that $\mathscr{S}$ is a semi-ring and $X \times Y \in \mathscr{S}$.

Def 34. *The* ***section of a set*** *$E \in X \times Y$ at $x \in X$ is*

$$E_x = \{y : (x, y) \in E\}.$$

Similarly $E_y = \{x : (x, y) \in E\}$ is the section at $y \in Y$. The ***section of a function*** *$f(x, y)$ at $x \in X$ is the mapping*

$$y \mapsto f(x, y).$$

The section at $y \in Y$ is $x \mapsto f(x, y)$.

Example 21. *If E, $E_k \in X \times Y$, $x \in X$, then*

$$(E^c)_x = (E_x)^c, \ \left(\bigcup_k E_k\right)_x = \bigcup_k (E_k)_x, \ \left(\bigcap_k E_k\right)_x = \bigcap_k (E_k)_x$$

Thm 51. (1) *Sections of $\mathscr{X} \times \mathscr{Y}$-measurable set are measurable.* (2) *Sections of $\mathscr{X} \times \mathscr{Y}$-measurable function are measurable.*

Proof. **1**. Fix $x \in X$. Consider the mapping $R_x : Y \mapsto X \times Y$ defined by $R_x(y) = (x, y)$. We intend to prove that R_x is $\mathscr{Y}$-measurable so that the conclusion follows immediately:

$$E_x = R_x^{-1} E \in \mathscr{Y}, \ \forall E \in \mathscr{X} \times \mathscr{Y}.$$

If $E = A \times B$ is a measurable rectangle, then $R_x^{-1} E = B \in \mathscr{Y}$. This shows that R_x is $\mathscr{Y}$-measurable, since $\mathscr{X} \times \mathscr{Y}$ is generated by measurable rectangles. So the first part is proved.

2. If $f \in \mathscr{X} \times \mathscr{Y}$, then $f(x, \cdot) = f \circ R_x(\cdot)$ is $\mathscr{Y}$-measurable by measurable composition.

3. The conclusion for fixed $y \in Y$ is proved similarly. □

8.2. Product measure space. Let $(X, \mathscr{X}, \mu)$, $(Y, \mathscr{Y}, \nu)$ be measure spaces.

LEMMA 52. *Suppose that μ and ν are finite. If $E \in \mathscr{X} \times \mathscr{Y}$, then the mapping $x \mapsto \nu(E_x)$ is $\mathscr{X}$-measurable, $y \mapsto \mu(E_y)$ is $\mathscr{Y}$-measurable.*

PROOF. Let $\mathscr{L}$ be the class of $E \in \mathscr{X} \times \mathscr{Y}$ that has the stated property. Then $\mathscr{L}$ is a λ-system. Indeed, it is easy to see that $X \times Y \in \mathscr{L}$. If E, $F \in \mathscr{L}$ with $E \subset F$, then

$$x \mapsto \nu((F \backslash E)_x) = \nu(F_x \backslash E_x) = \nu(F_x) - \nu(E_x)$$

is $\mathscr{X}$-measurable (the finiteness of ν is used to justify subtraction). If $E_k \in \mathscr{L}$, $E_k \subset E_{k+1}$, then

$$x \mapsto \nu\Big(\Big(\bigcup_k E_k\Big)_x\Big) = \nu\Big(\bigcup_k (E_k)_x\Big) = \lim_k \nu((E_k)_x)$$

is $\mathscr{X}$-measurable. This shows that $\mathscr{L}$ is a λ-system. For any measurable rectangle $E = A \times B$, the function

$$x \mapsto \nu(E_x) = 1_A(x)\nu(B)$$

is $\mathscr{X}$-measurable. So $\mathscr{L}$ contains the π-system of measurable rectangles, thus coincides with $\mathscr{X} \times \mathscr{Y}$ by π-λ Theorem. $\square$

LEMMA 53. *Let $(X, \mathscr{X}, \mu)$, $(Y, \mathscr{Y}, \nu)$ be σ-finite measure spaces. Define*

$$\pi_{21}(E) = \int_X \nu(E_x) d\mu, \ E \in \mathscr{X} \times \mathscr{Y}$$

and

$$\pi_{12}(E) = \int_Y \mu(E_y) d\nu, \ E \in \mathscr{X} \times \mathscr{Y}.$$

Then $\pi_{21}(E) = \pi_{12}(E)$ for $E \in \mathscr{X} \times \mathscr{Y}$. Moreover, π_{21}, π_{12} satisfy (8.1).

PROOF. **1**. First suppose μ, ν are finite. Let $\mathscr{L}$ be the class of $E \in \mathscr{X} \times \mathscr{Y}$ such that $\pi_{21}(E) = \pi_{12}(E)$. $\mathscr{L}$ contains all measurable rectangle $A \times B$, since

$$\pi_{21}(A \times B) = \int_X 1_A(x)\nu(B)d\mu = \mu(A)\nu(B)$$
$$= \int_Y \mu(A)1_B(y)d\nu = \pi_{12}(A \times B).$$

It is easy to check that $\mathscr{L}$ is a λ-system, hence equals $\mathscr{X} \times \mathscr{Y}$ by π-λ Theorem.

2. Now suppose μ, ν are σ-finite, then there are $\{A_m\}$, $\{B_n\}$ that partition X and Y into disjoint sets of finite measure. Define

$$\mu_m(E) = \mu(E \cap A_m),\ \nu_n(E) = \nu(E \cap B_n),\ E \in \mathscr{X} \times \mathscr{Y}.$$

Step **1** is valid for these finite measures,

$$\pi_{21}^{(mn)}(E) = \pi_{12}^{(mn)}(E),\ E \in \mathscr{X} \times \mathscr{Y}, \tag{8.3}$$

where

$$\pi_{21}^{(mn)}(E) = \int_X \nu_n(E_x)d\mu_m,$$

and

$$\pi_{12}^{(mn)}(E) = \int_Y \mu_m(E_y)d\nu_n.$$

In addition, for measurable rectangle $A \times B$,

$$\pi_{21}^{(mn)}(A \times B) = \mu_m(A)\nu_n(B) = \pi_{12}^{(mn)}(A \times B). \tag{8.4}$$

From Lemma 52, $x \mapsto \nu_n(E_x)$ is measurable. Since $\nu = \sum_n \nu_n$, $x \mapsto \nu(E_x)$ is measurable. The same can be said for $y \mapsto \mu(E_y)$. Therefore π_{21}, π_{12} are well-defined for the σ-finite case. By Example 10, 11 and (8.3), for $E \in \mathscr{X} \times \mathscr{Y}$,

$$\begin{aligned}
\pi_{21}(E) &= \int_X \nu(E_x)d\mu = \sum_m \int_X \nu(E_x)d\mu_m = \sum_m \int_X \sum_n \nu_n(E_x)d\mu_m \\
&= \sum_{m,n} \pi_{21}^{(mn)}(E) = \sum_{m,n} \pi_{12}^{(mn)}(E) = \sum_n \int_Y \sum_m \mu_m(E_y)d\nu_n \\
&= \pi_{12}(E).
\end{aligned}$$

Particularly, this together with (8.4) yields that, for measurable rectangle $A \times B$,

$$\pi_{21}(A \times B) = \sum_{m,n} \mu_m(A)\nu_n(B) = \pi_{12}(A \times B).$$

$\square$

Thm 54. *Let $(X, \mathscr{X}, \mu)$, $(Y, \mathscr{Y}, \nu)$ be σ-finite measure spaces. Then*

$$\pi(E) \triangleq \pi_{21}(E) = \pi_{12}(E)$$

defines the unique σ-finite measure on $X \times Y$ that satisfies (8.1).

PROOF. Decompose X and Y into disjoint sets $\{A_m\}$, $\{B_n\}$ of finite measure, and define μ_m, ν_n as before. Then $X \times Y$ is the disjoint union of $\{A_m \times B_n\}$ and each $A_m \times B_n$ has finite π-measure: $\pi(A_m \times B_n) = \mu_m(A_m)\nu_n(B_m)$. It follows that π is σ-finite. The uniqueness is a consequence of Theorem 10, since measurable rectangles form a π-system (Example 20). $\square$

In the future, the product measure π of μ and ν will be written as $\mu \times \nu$.

8.3. Fubini's Theorem.

Thm 55. *Let $(X, \mathscr{X}, \mu)$, $(Y, \mathscr{Y}, \nu)$ be σ-finite measure spaces, π the product measure constructed in Theorem 54, f a $\mathscr{X} \times \mathscr{Y}$-measurable function. If f is nonnegative or $\int_{X\times Y} |f| d\pi < \infty$. Then*

$$\int_X \left[\int_Y f(x,y) d\nu \right] d\mu = \int_Y \left[\int_X f(x,y) d\mu \right] d\nu = \int_{X\times Y} f d\pi.$$

It is implicit in the statement that all integrands are integrable. In the nonnegative case, if one of the above integrals is infinite, so it is with the other two.

Proof. The conclusion holds for measurable indicator function by Theorem 54, and hence simple function by linearity of integration.

Then the monotone convergence theorem gives the conclusion for nonnegative measurable function. If $\displaystyle\int_{X\times Y}|f|d\pi<\infty$, then applying the nonnegative case to $|f|$,

$$\int_X\left[\int_Y|f|d\nu\right]d\mu=\int_{X\times Y}|f|d\pi<\infty.$$

It follows that

$$\int_Y|f|d\nu<\infty,\ \ a.e.\ x.$$

Hence it makes sense (outside a set of zero μ-measure) to write

$$\int_Y f d\nu=\int_Y f^+d\nu-\int_Y f^-d\nu.$$

Now the desired property follows by integrating over X and using the result for nonnegative integrand. The same reasoning applies to $\displaystyle\int_Y\left[\int_X|f|d\mu\right]d\nu$. $\square$

8.4. Applications.

Example 22 (Euler-Poisson integral). *Let* $I = \int_{\mathbb{R}} e^{-x^2} dx$. *By Fubini's theorem*

$$I^2 = \iint_{\mathbb{R}^2} e^{-(x+y)^2} dxdy = \iint_{\substack{r \geqslant 0, \\ 0 \leqslant \theta < 2\pi}} e^{-r^2} r dr d\theta.$$

Again by Fubini's theorem, the double integral on the RHS is written as an iterated integral and evaluated to give $I^2 = \pi$, *so*

$$I = \int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}.$$

Thm 56. *Let* $(X, \mathscr{X}, \mu)$ *be a* σ*-finite measure space,* $f \geqslant 0$ *measurable. Then*

$$\int f d\mu = \int_0^\infty \mu(f \geqslant t) dt = \int_0^\infty \mu(f > t) dt$$

Proof. Since f is nonnegative, we may write (recall our convention Remark 3), $\forall x$,

$$f(x) = \int_0^{f(x)} dt = \int_0^\infty 1_{(0,f(x)]}(t)dt.$$

Notice that

$$1_{(0,f(x)]}(t) = 1_{\{x:f(x)\geqslant t\}}(x).$$

Then using Fubini theorem

$$\begin{aligned}\int f d\mu &= \int \int_0^\infty 1_{(0,f(x)]}(t)dtd\mu = \int_0^\infty \int 1_{(0,f(x)]}(t)d\mu dt \\ &= \int_0^\infty \int 1_{\{x:f(x)\geqslant t\}}(x)d\mu dt = \int_0^\infty \mu(\{x : f(x) \geqslant t\})dt.\end{aligned}$$

Since the set of t such that $\mu(\{x : f(x) = t\})$ is non-zero is at most countable, hence has zero Lebesgue measure. Thus the two integrals are equal,

$$\int_0^\infty \mu(\{x : f(x) \geqslant t\})dt = \int_0^\infty \mu(\{x : f(x) > t\})dt.$$

$\square$

If f takes values in $\{y_1, y_2, ...\}$ and $0 \leqslant y_1 < y_2 < \cdots$. Then $t \mapsto \mu(f \geqslant t)$ is a step function

$$\mu(f \geqslant t) = \begin{cases} \mu(f \geqslant y_1), & 0 \leqslant t \leqslant y_1; \\ \mu(f \geqslant y_n), & y_{n-1} < t \leqslant y_n. \end{cases}$$

Hence Theorem 56 reduces to

$$\int f d\mu = y_1 \mu(f \geqslant y_1) + \sum_{n=2}^{\infty} (y_n - y_{n-1}) \mu(f \geqslant y_n).$$

A particular case of this is f taking values in nonnegative integers.

Corollary 3. *Let $(X, \mathscr{X}, \mu)$ be a σ-finite measure space, f measurable with values in $\{0, 1, 2, ...\}$. Then*

$$\int f d\mu = \sum_{n=1}^{\infty} \mu(f \geqslant n) = \sum_{n=0}^{\infty} \mu(f > n).$$

Example 23. *Example 11 has validated interchanging the order of summation as an application of dominated convergence theorem. The same can also be proved directly using Fubini theorem.*

Thm 57 (**Integration by parts**). *Let F, G be two nondecreasing, right-continuous functions on $\mathbb{R}$, then, for $a < b$,*

$$F(b)G(b) - F(a)G(a) = \int_{(a,b]} G(x)dF(x) + \int_{(a,b]} F(x-)dG(x),$$

or equivalently

$$F(b)G(b) - F(a)G(a) = \int_{(a,b]} G(-x)dF(x) + \int_{(a,b]} F(x-)dG(x) + \sum_{a<x\leqslant b} \Delta F(x)\Delta G(x),$$

where $\Delta F(x) = F(x) - F(x-)$.

Proof. Denote respectively by μ, ν the Lebesgue-Stieltjes measure induced by F, G. Let $\pi = \mu \times \nu$ be the product measure of μ, ν.

Using Fubini Theorem we have

$$
\begin{aligned}
\pi((x,y) : a < x < y \leqslant b) &= \int_{(a,b]} \int_{(a,y)} d\mu(x) d\nu(y) \\
&= \int_{(a,b]} [F(y-) - F(a)] d\nu(y) \\
&= \int_{(a,b]} F(y-) d\nu(y) - F(a)[G(b) - G(a)]
\end{aligned}
$$

and similarly

$$
\begin{aligned}
\pi((x,y) : a < y \leqslant x \leqslant b) &= \int_{(a,b]} \int_{(a,x]} d\nu(y) d\mu(x) \\
&= \int_{(a,b]} [G(x) - G(a)] d\mu(x) \\
&= \int_{(a,b]} G(x) d\mu(x) - G(a)[F(b) - F(a)]
\end{aligned}
$$

By the construction of π,

$$(F(b) - F(a))(G(b) - G(a)) = \pi((x,y) \in (a,b] \times (a,b]).$$

The first conclusion follows by putting together these equations. To complete the proof, it suffices to note that

$$\begin{aligned}
\pi((x,y) : a < y = x \leqslant b) &= \int_{(a,b]} \nu(\{x\}) d\mu(x) \\
&= \int_{(a,b]} [G(x) - G(x-)] d\mu(x) \\
&= \sum_{a<x\leqslant b} \Delta F(x) \Delta G(x).
\end{aligned}$$

□

8.5. Finite-dimensional product space. The discussion for two-dimensional product space obviously extends to finite dimensions. Denote by $\mathscr{B}(\mathbb{R}^n)$ the σ-field on $\mathbb{R}^n$ generated by open sets. The product

σ-field $\prod_{i=1}^{n} \mathscr{B}(\mathbb{R})$ is defined as being generated by measurable rectangles with sides in $\mathscr{B}(\mathbb{R})$. Similar to Example 19, we have on $\mathbb{R}^n$,

$$\mathscr{B}(\mathbb{R}^n) = \prod_{i=1}^{n} \mathscr{B}(\mathbb{R}). \tag{8.5}$$

9. 独立性 Independence

Let $(\Omega, \mathscr{F}, P)$ be a probability space. A subset in $\mathscr{F}$ is called an **event**, and an element of Ω is called a **sample**.

9.1. Independence of events and random variables.

Def 35. *The events A and B are independent if*

$$P(A \cap B) = P(A)P(B).$$

Thm 58. *A, B are independent if and only if one of the three pairs are independent: (i) A^c, B; (ii) A, B^c; (iii) A^c, B^c.*

Example 24. *An event A is independent of any event if and only if $P(A) = 0$ or 1. In particular, Ω and $\varnothing$ are independent of any event.*

Example 25 (Gambler's fallacy). *In some situations, an individual erroneously think that certain event is more or less likely to happen in the future based on the outcome of the past events. This incorrect belief may lead a gambler in a coin flipping game to believe that after 100 successive heads, the next toss would be more likely to* *come up tail. The fallacy roots from the ignorance of the independence between tosses.*

Def 36. *σ-fields $\mathscr{F}, \mathscr{G}$ are independent if*

$$P(A \cap B) = P(A)P(B),\ \forall A \in \mathscr{F},\ B \in \mathscr{G}.$$

Def 37. *Random variables X, Y are independent if $\sigma(X)$ and $\sigma(Y)$ are independent, where $\sigma(X) = X^{-1}(\mathscr{B}(\mathbb{R}))$.*

Thm 59. *A, B are independent if and only if 1_A and 1_B are independent.*

Proof. Note that $\sigma(1_A) = \{A, A^c, \varnothing, \Omega\}$ and the same for $\sigma(1_B)$. $\square$

Def 38. *A family of events $\{A_1, ..., A_n\}$ is independent if for any $I \subset \{1, ..., n\}$,*

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

Pairwise independence is weaker than independence.

Example 26. *Flip a fair coin twice and consider the events,*

$$A_1 = \{ \text{ head-head, head-tail } \},$$
$$A_2 = \{ \text{ head-head, tail-head } \},$$
$$A_3 = \{ \text{ head-head, tail-tail } \}.$$

Then A_1, A_2, A_3 are pairwise independent but not independent, since

$$P(A_i \cap A_j) = P(\text{ head-head }) = \frac{1}{4} = P(A_i)P(A_j),\ i \neq j.$$

$$P(A_1 \cap A_2 \cap A_3) = P(\text{ head-head }) = \frac{1}{4} \neq \frac{1}{8} = P(A_1)P(A_2)P(A_3).$$

Def 39. *$\mathscr{A}_1, ..., \mathscr{A}_n$ are classes of sets. They are independent if for any $I \subset \{1, ..., n\}$,*

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i),\ \forall A_i \in \mathscr{A}_i.$$

If we denote by $\mathscr{A}_i'$ the class formed by augmenting $\mathscr{A}_i$ with Ω. Then it is easy to see that $\mathscr{A}_1, ..., \mathscr{A}_n$ are independent if and only if

$\mathscr{A}_1', ..., \mathscr{A}_n'$ are independent. Definition 39 is thus equivalent to the *full-product form*,

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i),\ \forall A_i \in \mathscr{A}_i'. \tag{9.1}$$

This form may bring added convenience when independence is to be verifed. Since Ω is contained in σ-field, the independence of random variables can be defined in this full-product form.

Def 40. *$X_1, ..., X_n$ are independent if $\sigma(X_1), ..., \sigma(X_n)$ are independent, i.e.,*

$$P\left(\bigcap_{i=1}^n \{X_i \in A_i\}\right) = \prod_{i=1}^n P(X_i \in A_i),\ \forall A_i \in \mathscr{B}(\mathbb{R}),\ i = 1, ..., n.$$

Thm 60. *Suppose that $\mathscr{A}_1, \mathscr{A}_2, ..., \mathscr{A}_n$ are independent π-systems. Then $\sigma(\mathscr{A}_1), \sigma(\mathscr{A}_2), ..., \sigma(\mathscr{A}_n)$ are independent.*

PROOF. **1**. Clearly it suffices to show that $\sigma(\mathscr{A}_1), \mathscr{A}_2, ..., \mathscr{A}_n$ are independent, since the conclusion applies to itself and would yield that $\sigma(\mathscr{A}_1)$, $\sigma(\mathscr{A}_2), ..., \mathscr{A}_n$ are independent, and so on.

2. Now we show that $\sigma(\mathscr{A}_1), \mathscr{A}_2, ..., \mathscr{A}_n$ are independent. Fix $A_i \in \mathscr{A}_i$, $i = 2, ..., n$. Let $E = \bigcap_{i=2}^{n} A_i$ and

$$\mathscr{L}_E = \{A \in \sigma(\mathscr{A}_1) : P(A \cap E) = P(A)P(E)\}.$$

Then $\mathscr{A}_1 \subset \mathscr{L}_E$. In view of Example 24, $\Omega \in \mathscr{L}_E$. If B_1, $B_2 \in \mathscr{L}_E$ and $B_1 \subset B_2$, then

$$\begin{aligned} P((B_2 - B_1) \cap E) &= P(B_2 \cap E) - P(B_1 \cap E) \\ &= P(B_2)P(E) - P(B_1)P(E) \\ &= P(B_2 - B_1)P(E). \end{aligned}$$

Hence $B_2 - B_1 \in \mathscr{L}_E$. Finally let $B_k \in \mathscr{L}_E$, $B_k \subset B_{k+1}$, then

$$\begin{aligned}
P\left(\left(\bigcup_{k=1}^{\infty} B_k\right) \cap E\right) &= \lim_k P(B_k \cap E) \\
&= \lim_k P(B_k)P(E) = P\left(\bigcup_{k=1}^{\infty} B_k\right) P(E).
\end{aligned}$$

Thus $\bigcup_{k=1}^{\infty} B_k \in \mathscr{L}_E$. Therefore $\mathscr{L}_E$ is a λ-system and $\sigma(\mathscr{A}_1) \subset \mathscr{L}_E$. The desired conclusion follows. $\square$

Since $\mathscr{B}(\mathbb{R})$ is generated by the class $\mathscr{S} = \{(-\infty, a] : a \in \mathbb{R}\}$, the σ-field $\sigma(X)$ generated by X equals $\sigma(\{X \leqslant a : a \in \mathbb{R}\})$, which together with Theorem 60 gives the following criterion for independence in terms of distribution functions.

Thm 61. *$X_1, ..., X_n$ are independent if and only if $\forall x_1, ..., x_n \in \mathbb{R}$,*

$$P(X_1 \leqslant x_1, ..., X_n \leqslant x_n) = \prod_{i=1}^{n} P(X_i \leqslant x_i).$$

Note the class $\{X_i \leqslant x_i : x_i \in \mathbb{R}\}$ may not contain but can approximate Ω, so it is still legal to use the full-product form (9.1).

Recall from Theorem 46 that each random variable X induces a probability μ on $\mathbb{R}$, which is called the distribution of X. When random vector $(X_1, ..., X_n)$ is involved, the same can be said. From Section 5.3, we see that the random vector induces a probability on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, where $\mathscr{B}(\mathbb{R}^n)$ equals the σ-field generated by measurable rectangles with sides in $\mathscr{B}(\mathbb{R})$ (see (8.5)). The induced probability on $\mathscr{B}(\mathbb{R}^n)$, denoted by $P_{X_1,...,X_n}$, satisfies, $\forall B_1, ..., B_n \in \mathscr{B}(\mathbb{R})$,

$$P_{X_1,...,X_n}(B_1 \times \cdots \times B_n) = P(X_1 \in B_1, ..., X_n \in B_n).$$

$P_{X_1,...,X_n}$ is called the **joint distribution** of $(X_1, ..., X_n)$. If B_i takes the form $(-\infty, x_i]$, then we get an associated mapping

$$F_{X_1,...,X_n} : (x_1, ..., x_n) \mapsto P(X_1 \leqslant x_1, ..., X_n \leqslant x_n)$$

which is called the **joint distribution function**[2] of the random vectors $(X_1, ..., X_n)$. Moreover, the general change of variables formula from Section 5.3 tells us that, for measurable $g : \mathbb{R}^n \mapsto \mathbb{R}$,

$$(9.2) \qquad Eg(X_1, ..., X_n) = \int_{\mathbb{R}^n} g(x_1, ..., x_n) dP_{X_1,...,X_n}.$$

whenever one of the integrals exists.

Example 27. *The discrete random variables $X \in \{x_1, x_2, ...\}$ and $Y \in \{y_1, y_2, ...\}$ are independent if and only if*

$$(9.3) \qquad P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j), \ \forall i, \ j.$$

[2] In fact, by uniqueness $P_{X_1,...,X_n}$ equals the Lebesgue-Stieltjes measure determined by the joint distribution function on $\mathscr{B}(\mathbb{R}^n)$.

PROOF. **1**. Suppose that (9.3) holds. $\forall x,\ y \in \mathbb{R}$,

$$P(X \leqslant x, Y \leqslant y) = \sum_{\substack{i: x_i \leqslant x \\ j: y_j \leqslant y}} P(X = x_i, Y = y_j)$$

$$= \sum_{\substack{i: x_i \leqslant x \\ j: y_j \leqslant y}} P(X = x_i) P(Y = y_j).$$

The double summation may contain infinite number of terms, but we can invoke Fubini theorem to write it as iterated summation (see also Example 23)

$$\sum_{\substack{i: x_i \leqslant x \\ j: y_j \leqslant y}} P(X = x_i) P(Y = y_j) = \sum_{i: x_i \leqslant x} P(X = x_i) \sum_{j: y_j \leqslant y} P(Y = y_j)$$

$$= P(X \leqslant x) P(Y \leqslant y).$$

Hence

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

By Theorem 61 X and Y are independent.

2. Conversely suppose X and Y are independent. Then $\forall i$, j, the events $\{X = x_i\}$, $\{Y = y_j\}$ are independent, so (9.3) holds. $\square$

In general, the probability of the event $\{X \in (a, b], Y \in (c, d]\}$ can be expressed in terms of joint distribution function F,

$$P(X \in (a, b], Y \in (c, d]) = F(b, d) - F(b, c) - F(a, d) + F(a, c).$$

Thm 62. *Suppose that the collection of events*

$$A_{IJ} = \{A_{ij} : i \in I,\ j \in J\}$$

are independent. Here I, J are finite or infinite index sets. Let

$$\mathscr{F}_i = \sigma(A_{ij} : j \in J),\ \forall i \in I.$$

Then $\mathscr{F}_1, \mathscr{F}_2, \ldots$ are independent.

Remark 4. *Independence of infinite number of events is defined as any finite subcollection being independent.*

PROOF. $\forall i \in I$, denote

$$\mathscr{A}_i = \{\text{all finite intersections of } A_{i1}, A_{i2}, ...\}.$$

Then $\mathscr{A}_1$, $\mathscr{A}_2$,... are π-systems and $\mathscr{F}_i = \sigma(\mathscr{A}_i)$. Hence the conclusion follows from Theorem 60. $\square$

By inspecting the proof, we see that the above theorem extends to the case where the index set J varies with $i \in I$. As an application, we have the following useful result which states that functions of disjoint subgroups of independent random variables are independent.

For ease of writing, we introduce the notation of indexing by set, for example,

$$\text{if } I = \{i_1, ..., i_l\}, \text{ then } X_I \text{ means } (X_{i_1}, ..., X_{i_l}).$$

Thm 63. *Divide the independent random variables $X_1, ..., X_n$ into disjoint subgroups $X_{I_1}, ..., X_{I_k}$, where $I_1, I_2, ..., I_k \subset \{1, 2, ..., n\}$ are disjoint, $\bigcup_{i=1}^{k} I_i = \{1, 2, ..., n\}$. If $g_1(x_{I_1}), ..., g_k(x_{I_k})$ are measurable functions, then $g_1(X_{I_1}), ..., g_k(X_{I_k})$ are independent.*

PROOF. Note $\sigma(g_s(X_{I_s})) \subset \sigma(X_{I_s})$, $s \in \{1, ..., k\}$, hence it is enough to show that $\sigma(X_{I_1}), ..., \sigma(X_{I_k})$ are independent. Each $\sigma(X_{I_i})$ can be generated by $\{\sigma(X_j) : i \in I_i\}$. Now the proof is completed by applying Theorem 62 to

$$A_{IJ} = \{\sigma(X_j) : i = 1, ..., k,\ j \in I_i\}.$$

$\square$

9.2. Independence and expectation.

Thm 64. *Suppose that $X_1, ..., X_n$ are independent with respective distribution μ_i. Then $(X_1, ..., X_n)$ has the joint distribution $\mu_1 \times \cdots \times \mu_n$.*

PROOF. $\forall B_1, ..., B_n \in \mathscr{B}(\mathbb{R})$, using independence and the definition of product measure

$$\begin{aligned} P(X_1 \in B_1, ..., X_n \in B_n) &= \prod_{i=1}^{n} P(X_i \in B_i) = \prod_{i=1}^{n} \mu_i(B_i) \\ &= \mu_1 \times \cdots \times \mu_n(B_1 \times \cdots \times B_n). \end{aligned}$$

The class of measurable rectangles

$$\{B_1 \times \cdots \times B_n : B_1, ..., B_n \in \mathscr{B}(\mathbb{R})\}$$

is a π-system. Therefore by uniqueness (Theorem 10), $\mu_1 \times \cdots \times \mu_n$ agrees with the joint distribution of $(X_1, ..., X_n)$. $\square$

Thm 65. *Suppose that X, Y are independent with respective distribution μ and ν, $h : \mathbb{R}^2 \mapsto \mathbb{R}$ is measurable. If $h \geqslant 0$ or $E|h(X,Y)| < \infty$, then*

$$Eh(X,Y) = \int_{\mathbb{R}^2} h(x,y) d\mu(x) d\nu(y).$$

In particular if $h(x,y) = f(x)g(y)$, then

$$Ef(X)g(Y) = Ef(X) \cdot Eg(Y).$$

PROOF. By Theorem 64, the induced probability of (X,Y) is given by the product of μ and ν. Then using the general change of variables formula (Section 5.3, Formula 9.2), we can write

$$Eh(X,Y) = \int_{\mathbb{R}^2} h(x,y) d\mu(x) d\nu(y)$$

The remaining conclusion follows easily. □

Inductively using the above theorem we get the following expectation formula for independent random variables.

Thm 66. *Suppose that* $X_1, ..., X_n$ *are independent and either* (a) $X_i \geqslant 0$, $\forall i$ *or* (b) $E|X_i| < \infty$, $\forall i$. *Then*

$$EX_1 \cdots X_n = EX_1 \cdots EX_n.$$

Proof. The nonnegative case is immediate from Theorem 65. To prove case (b), applying the nonnegative case to $|X_1|$ and $|X_2|$, we have

$$E|X_1 X_2| = E|X_1| \cdot E|X_2|.$$

Since the RHS is finite, so $E|X_1 X_2| < \infty$ and by Theorem 65,

$$EX_1 X_2 = EX_1 \cdot EX_2.$$

Now the nonnegative case is again applicable to $|X_1 X_2|$ and $|X_3|$,

$$E|X_1 X_2 X_3| = E|X_1 X_2| \cdot E|X_3|.$$

So $E|X_1X_2X_3| < \infty$ and Theorem 65 can be invoked to get

$$EX_1X_2X_3 = EX_1 \cdot EX_2 \cdot EX_3.$$

The procedure continues until X_n is processed, then the proof is completed. $\square$

Def 41. *The covariance of X and Y is defined as*

$$Cov(X, Y) = E(X - EX)(Y - EY) = EXY - EX \cdot EY.$$

Def 42. *X and Y are uncorrelated if $Cov(X, Y) = 0$, i.e.,*

$$EXY = EX \cdot EY.$$

That X and Y are independent implies that they are uncorrelated, but not vice versa.

Example 28. *Suppose that X, Y are jointly distributed as below*

	$Y = -1$	$Y = 0$	$Y = 1$
$X = -1$	0	1/4	0
$X = 0$	1/4	0	1/4
$X = 1$	0	1/4	0

Then $EX = EY = EXY = 0$, *but* X, Y *are not independent by Example 27, since*

$$P(X=0, Y=0) = 0 \neq \frac{1}{4} = P(X=0)P(Y=0).$$

Thm 67. *Suppose that* $X_1, ..., X_n$ *are pairwise uncorrelated and* $EX_i^2 < \infty$. *Then*

$$Var(X_1 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n).$$

Proof. Denote $X = X_1 + \cdots + X_n$. We have

$$\begin{aligned}
\mathrm{Var}(X) &= E(X - EX)^2 = E\left(\sum_{i=1}^{n}(X_i - EX_i)\right)^2 \\
&= E\left(\sum_{i=1}^{n}(X_i - EX_i)^2 + \sum_{i \neq j}(X_i - EX_i)(X_j - EX_j)\right) \\
&= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n).
\end{aligned}$$

We have used that $X_1, ..., X_n$ are pairwise uncorrelated, hence $\forall i \neq j$,

$$E((X_i - EX_i)(X_j - EX_j)) = EX_iX_j - EX_iEX_j = 0.$$

$\square$

9.3. Sum of independent random variables. "Independent and identically distributed" is abbreviated as **i.i.d**.

Example 29. ***(Binomial distribution)*** *Let $p \in (0, 1)$, $X_1, ..., X_n$ i.i.d. $\sim$ Bernoulli(p). Define $S_n = \sum_{i=1}^{n} X_i$. Then*

$$P(S_n = k) = C_n^k p^k (1-p)^{n-k}, \ k = 0, ..., n.$$

The distribution of S_n is the binomial distribution, written $S_n \sim Bin(n, p)$. By employing linearity of integration, Example 14 and Theorem 67,

$$ES_n = np, \ Var(S_n) = np(1-p).$$

Example 30 (The problem of points). *A coin with head probability p is flipped repeatedly. Gambler A wins one point if head appears* *on a toss, otherwise gambler B wins one point. Whoever reaches first the finishing line wins the game. Suppose that gambler A and B are m and n points away from the finishing line. We intend to find the probability $W(m,n)$ that gambler A wins the game. Imagine tossing the coins $m+n-1$ times, then gambler A wins the game if and only if heads show up at least m times, the probability is*

$$W(m,n) = \sum_{k=m}^{m+n-1} C_{m+n-1}^{k} p^k (1-p)^{m+n-1-k}.$$

The probability of A winning the game can be categorized based on the outcome of the first toss. If the first toss is a head, the probability of A winning the game afterwards would be $W(m-1,n)$, otherwise $W(m,n-1)$. Therefore the recursion holds

$$W(m,n) = p \cdot W(m-1,n) + (1-p) \cdot W(m,n-1).$$

The equation may be solved by observing the boundary conditions,

$$W(0,j) = 1 \text{ for } j = n, n-1, ..., 1$$

and

$$W(i,0) = 0 \textit{ for } i = m, m-1, ..., 1.$$

Example 31 (**Sum of Binomials**). *Suppose that* $X \sim Bin(m,p)$ *and* $Y \sim Bin(n,p)$ *are independent, then* $X+Y \sim Bin(m+n,p)$.

PROOF. For any $k \in \{0,1,...,m+n\}$,

$$\begin{aligned}
P(X+Y=k) &= \sum_{i=0}^{k} P(X=i, Y=k-i) = \sum_{i=0}^{k} P(X=i)P(Y=k-i) \\
&= \sum_{i=0}^{k} C_m^i p^i (1-p)^{m-i} \cdot C_n^{k-i} p^{k-i} (1-p)^{n-k+i} \\
&= p^k (1-p)^{m+n-k} \sum_{i=0}^{k} C_m^i C_n^{k-i} = C_{m+n}^k p^k (1-p)^{m+n-k}.
\end{aligned}$$

The last equality is due to Vandermonde identity. $\square$

Thm 68 (**Convolution**). *Suppose that X, Y are independent with distribution functions F and G. Then*

$$P(X+Y \leqslant z) = \int_{\mathbb{R}} G(z-x)dF(x) = \int_{\mathbb{R}} F(z-y)dG(y)$$

Recall that dF, dG are notational variants for the corresponding Lebesgue-Stieltjes measure (Remark 3).

PROOF. Denote by μ, ν the distribution of X, Y. By Fubini theorem

$$\begin{aligned} P(X+Y \leqslant z) &= \int_{\mathbb{R}} \left(\int_{(-\infty, z-x]} d\nu(y) \right) d\mu(x) \\ &= \int_{\mathbb{R}} \left(\int_{(-\infty, z-y]} d\mu(x) \right) d\nu(y). \end{aligned}$$

The integrands respectively equal $G(z-x)$ and $F(z-y)$. $\square$

Def 43. *The random vector* $(X_1, ..., X_n)$ *has continuous distribution if there exists a function* $p \geqslant 0$ *such that*

$$P((X_1, ..., X_n) \in A) = \int_A p(x_1, ..., x_n) dx_1 \cdots dx_n, \ \forall A \in \mathscr{B}(\mathbb{R}^n).$$

The function p *is called the (joint) density of* $(X_1, ..., X_n)$.

The definition can be equivalently[3] stated as: The random vector $(X_1, ..., X_n)$ has continuous distribution if there exists a function $p \geqslant 0$ such that the joint distribution function F has

$$F(x_1, ..., x_n) = \int_{(-\infty, x]} p(s_1, ..., s_n) ds_1 \cdots ds_n, \ \forall x = (x_1, ..., x_n) \in \mathbb{R}^n,$$

where $(-\infty, x] = (-\infty, x_1] \times \cdots \times (-\infty, x_n]$.

In view of the definition, if X has density p, then its distribution function F has

$$F(x) = \int_{-\infty}^{x} p(s) ds, \ \forall x.$$

[3] See previous section.

Example 32 (Exponential distribution). *X has exponential distribution with parameter $\lambda > 0$, written $X \sim Exp(\lambda)$, if X has density*

$$p(x) = \begin{cases} \lambda e^{\lambda x}, & x \geqslant 0, \\ 0, & x < 0. \end{cases}$$

Then $EX = \lambda^{-1}$, $Var(X) = \lambda^{-2}$. In probability and statistics, the exponential distribution models the distribution of the waiting time before an event occurs.

Example 33 (Normal distribution). *X has normal distribution with parameter $\mu \in \mathbb{R}$, $\sigma > 0$, written $X \sim \mathcal{N}\big(\mu, \sigma^2\big)$, if X has density*

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

Then $EX = \mu$, $Var(X) = \sigma^2$.

Thm 69. *Suppose that X, Y are independent with distribution functions F and G. If X has density p_X, then $Z = X + Y$ has density*

$$h(z) = \int p_X(z-y)dG(y).$$

If also Y has density p_Y, then

$$h(z) = \int p_X(z-y)p_Y(y)dy.$$

PROOF. The convolution Theorem 68 now becomes

$$P(X+Y \leqslant z) = \int_{\mathbb{R}} \left(\int_{-\infty}^{z-y} p_X(x)dx \right) dG(y).$$

Combining a change of variable $u = x + y$ with Fubini theorem, we obtain

$$\begin{aligned} P(X+Y \leqslant z) &= \int_{\mathbb{R}} \left(\int_{-\infty}^{z} p_X(u-y)du \right) dG(y) \\ &= \int_{-\infty}^{z} \left(\int_{\mathbb{R}} p_X(u-y)dG(y) \right) du \end{aligned}$$

$\square$

Example 34 (**Bivariate normal distribution**). *X, Y are jointly normal, denoted by $(X, Y) \sim \mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, if the joint density $p(x, y)$ is given by*

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

Find the density of X, Y and $Z = X + Y$.

Proof. Tediuous calculations are omitted. The answers are

$$X \sim \mathcal{N}(\mu_1, \sigma_1^2), \ Y \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

and

$$Z \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2).$$

$\square$

Def 44. *Gamma function is defined for $\alpha > 0$, $\beta > 0$,*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx = \beta^\alpha \int_0^\infty x^{\alpha-1}e^{-\beta x}dx.$$

Note the first integral does not depend on β.

The following properties of Gamma functions are easy to verify that $\Gamma(1) = 1$ and

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \ \forall\alpha > 0.$$

If n is a positive integer, then

$$\Gamma(n + 1) = n!.$$

Def 45. *Beta function is defined for $\alpha > 0$, $\beta > 0$,*

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1 - x)^{\beta-1}dx.$$

LEMMA 70. *Beta function is related to Gamma function by the equation*

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

PROOF. We start with from the definition of Gamma function,

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty u^{\alpha-1}e^{-u}du \int_0^\infty v^{\beta-1}e^{-v}dv \\ &= \int_0^\infty \int_0^\infty u^{\alpha-1}v^{\beta-1}e^{-u-v}dudv.\end{aligned}$$

Now perform a change of variables,

$$u = st,\ v = s(1-t),\ \text{for } s > 0,\ 0 < t < 1.$$

The Jacobian determinant

$$\frac{\partial(u,v)}{\partial(s,t)} = \det\begin{pmatrix} t & s \\ 1-t & -s \end{pmatrix} = -s.$$

Hence by Fubini theorem

$$\begin{aligned}
\Gamma(\alpha)\Gamma(\beta) &= \int_0^1 \int_0^\infty s^{\alpha-1} t^{\alpha-1} s^{\beta-1} (1-t)^{\beta-1} e^{-s} s ds dt \\
&= \int_0^\infty s^{\alpha+\beta-1} e^{-s} ds \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} dt \\
&= \Gamma(\alpha+\beta) B(\alpha,\beta).
\end{aligned}$$

$\square$

Example 35 (Gamma distribution). *A random variable follows a Gamma distribution with parameter $\alpha > 0$, $\beta > 0$, if it has the density*

$$p(x) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & \text{for } x > 0, \\ 0, & \text{for } x \leqslant 0. \end{cases}.$$

Written Gamma(α, β).

Note Gamma$(1, \beta)$ is exponential distribution with parameter β, Exp(β).

Thm 71 (Sum of Gamma). *Suppose that $X_i \sim Gamma(\alpha_i, \beta)$, $i = 1, ..., n$ are independent. Then*

$$Y = X_1 + \cdots + X_n \sim Gamma(\alpha_1 + \cdots + \alpha_n, \beta).$$

If $\alpha_i = 1$, we can imagine n customers in a queue, each must wait time X_i for service once reaching the head of the queue. The average service rate is β. Then Y is the total waiting time of all n customers.

PROOF. It suffices to prove for $i = 2$. Write p_{X_1}, p_{X_2} for the densities of X_1, X_2. Then the density of $Y = X_1 + X_2$ is given by Theorem 69,

$$p_Y(y) = \int_0^y p_{X_1}(y - x) p_{X_2}(x) dx, \;\; y > 0.$$

The integration is from 0 to y, since the densities $p_{X_1}(x_1)$ and $p_{X_2}(x_2)$ are non-zero only if $x_1 > 0$, $x_2 > 0$. Hence

$$\begin{aligned}
p_Y(y) &= \int_0^y \frac{\beta^{\alpha_1}}{\Gamma(\alpha_1)}(y-x)^{\alpha_1-1}e^{-\beta(y-x)}\frac{\beta^{\alpha_2}}{\Gamma(\alpha_2)}x^{\alpha_2-1}e^{-\beta x}dx \\
&= \frac{\beta^{\alpha_1+\alpha_2}e^{-\beta y}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^y (y-x)^{\alpha_1-1}x^{\alpha_2-1}dx \\
&=_{(x=yt)} \frac{\beta^{\alpha_1+\alpha_2}e^{-\beta y}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^y (y-yt)^{\alpha_1-1}(yt)^{\alpha_2-1}ydt \\
&= \frac{\beta^{\alpha_1+\alpha_2}y^{\alpha_1+\alpha_2-1}e^{-\beta y}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1 (1-t)^{\alpha_1-1}t^{\alpha_2-1}dt.
\end{aligned}$$

Using Lemma 70, we get

$$p_Y(y) = \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)}y^{\alpha_1+\alpha_2-1}e^{-\beta y} \sim \mathrm{Gamma}(\alpha_1+\alpha_2, \beta).$$

$\square$

Def 46. *A random variable X has Beta distribution with parameter $\alpha > 0$, $\beta > 0$, written $X \sim Beta(\alpha, \beta)$, if it has the density*

$$p(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1},\ 0 < x < 1,$$

where $B(\alpha, \beta)$ is the Beta function. Lemma 70 confirms that the integration of $p(x)$ over $(0, 1)$ equals one.

Example 36. *If $X \sim Beta(\alpha, \beta)$ with $\alpha > 0$, $\beta > 0$, then using the properties of Gamma functions and Lemma 70, it is easy to check that*

$$EX = \frac{\alpha}{\alpha + \beta},\ \ Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

9.4. Construction of independent sequence. Given a probability measure μ, we intend to construct a sequence of independent random variables so that each has distribution μ. Consider the probability space

$$(\Omega, \mathscr{F}, P) = ((0, 1], \mathscr{B}((0, 1]), \text{Lebesgue measure}).$$

Let $X \sim U((0,1])$ and take its **dyadic expansion**

$$X(\omega) = \sum_{k=1}^{\infty} \frac{X_k(\omega)}{2^k}. \tag{9.4}$$

The sequence X_k is determined by the algorithm: $X_1 = 0$ if $X \in (0, 1/2]$ (we can omit the left boundary point 0 since it carries zero Lebesgue measure), and 1 if $X \in (1/2, 1]$. Since X is distributed uniformly, it has equal probability of landing in $(0, 1/2]$ or $(1/2, 1]$, hence the random variable X_1 has Bernoulli distribution with parameter $1/2$,

$$P(X_1 = 0) = P(X \in (0, 1/2]) = 1/2,$$

and

$$P(X_1 = 1) = P(X \in (1/2, 1]) = 1/2.$$

If $X_1, ..., X_{k-1}$ are already determined, then split into two halves the interval where X locates and define $X_k = 0$ if X is on the left half, and 1 on the right half. As early,

$$X_k \sim \text{Bernoulli}(1/2).$$

Then we see by induction that for all $\omega \in \Omega$, $X(\omega)$ is bracketed by a sequence of intervals: $X(\omega) \in D_n$ for all $n \geqslant 1$ where D_n is the **dyadic interval of rank** n,

$$D_n = \left(\sum_{k=1}^{n} \frac{X_k(\omega)}{2^k}, \sum_{k=1}^{n} \frac{X_k(\omega)}{2^k} + \frac{1}{2^n} \right]. \tag{9.5}$$

This implies that every dyadic expansion defined this way is **non-terminating**, otherwise if there is n_0 such that $X_n = 0$ for $n > n_0$, then

$$X(\omega) = \sum_{n=1}^{n_0} \frac{X_n(\omega)}{2^n}.$$

But this contradicts that $X(\omega) \in D_{n_0}$. A consequence of the property is the (natural) uniqueness of dyadic expansion. As an example, between the two mathematically equivalent expressions of $1/2$,

$$\frac{1}{2} + \frac{0}{2} + \frac{0}{2} + \cdots \text{ and } \frac{0}{2} + \frac{1}{2} + \frac{1}{2} + \cdots,$$

the algorithm always chooses the one with infinitely many 1s, i.e., the former is *not* a dyadic expansion by the algorithm.

In numerics, the dyadic expansion of $X \in (0,1]$ is nothing but the bindary representation

$$X = 0.X_1X_2X_3\cdots$$

with $X_1, X_2, \cdots$ interpreted as binary digits.

LEMMA 72. *Every $X \sim U((0,1])$ has a unique (non-terminating) dyadic expansion, i.e., two expansions of X generated by the algorithm must necessarily have equal coefficients X_ks.*

PROOF. The uniqueness follows from the algorithm itself. Another way to prove it is to compare the coefficients of the expansion. Observe that if X has the dyadic expansion (9.4), then the coefficients $X_1, X_2, \cdots$ are uniquely determined as functions of X. For this, we define an operator similar to the floor function. To be consistent with our algorithm, the operator should map $x \in (n, n+1]$ to n for any

$n \in \{0, 1, ...\}$. Thus the function is

$$[[x]] = \text{ceil}(x) - 1 \text{ for } x > 0,$$

where $\text{ceil}(x) = \inf\{n \in \mathbb{N} : n \geqslant x\}$ is the smallest integer no less than x. Now we can succinctly write X_1 as a function of X. Multiply (9.4) by 2,

$$2X = 2\left(\frac{X_1}{2}\right) + 2\left(\frac{X_2}{2^2} + \frac{X_3}{2^3} + \cdots\right).$$

Since $X_k \in \{0, 1\}$ and the expansion generated by the algorithm is non-terminating, so the second term cannot be zero,

$$2\left(\frac{X_2}{2^2} + \frac{X_3}{2^3} + \cdots\right) \in (0, 1].$$

Therefore

$$[[2X]] = 2\left(\frac{X_1}{2}\right) = X_1.$$

Similarly we have

$$\big[\big[2^2 X\big]\big] = 2^2\left(\frac{X_1}{2} + \frac{X_2}{2^2}\right).$$

Hence

$$X_2 = 2^2\left(\frac{X_1}{2} + \frac{X_2}{2^2}\right) - 2 \cdot 2\left(\frac{X_1}{2}\right) = \big[\big[2^2 X\big]\big] - 2[[2X]].$$

By induction that, for $k \geqslant 1$,

$$X_k = \big[\big[2^k X\big]\big] - 2\big[\big[2^{k-1} X\big]\big].$$

Therefore $X_1, X_2, \cdots$ are uniquely determined by X. $\square$

REMARK 5. *If the base space is* $\Omega = [0, 1)$ *and* $X_1 = 0$ *if* $X \in [0, 1/2)$*, and* 1 *if* $X \in [1/2, 1)$*, and so on, then the generated dyadic expansion of* $1/2$ *would be*

$$\frac{1}{2} = \frac{1}{2} + \frac{0}{2} + \frac{0}{2} + \cdots.$$

In particualy the expansion coefficients $X_1, X_2, \cdots$ are uniquely determined as

$$X_k = \left[2^k X\right] - 2\left[2^{k-1} X\right] \text{ for } k \geqslant 1.$$

Recall $[\cdot]$ is the floor function that takes out the integer part.

LEMMA 73. *If $X \sim U((0,1])$ has the dyadic expansion (9.4), then $X_1, X_2, \cdots$ i.i.d $\sim$ Bernoulli$(1/2)$.*

PROOF. That each X_k has Bernoulli distribution with parameter $1/2$ is clearly from the algorthm. To show that $X_1, X_2, \cdots$ are independent, it is enough to verify that for any $n \geqslant 1$, $i_1, i_2, ..., i_n \in \{0,1\}$,

$$P(X_1 = i_1, ..., X_n = i_n) = P(X_1 = i_1) \cdots P(X_n = i_n).$$

But this is immediate once we observe that the RHS equals $1/2^n$ by construction and

$$\{X_1 = i_1, ..., X_n = i_n\} = \left\{X \in \left(\sum_{k=1}^{n} \frac{i_k}{2^k}, \sum_{k=1}^{n} \frac{i_k}{2^k} + \frac{1}{2^n}\right]\right\}.$$

Since X is uniform, the event on the RHS has probability $1/2^n$. Hence the desired independence follows. ☐

Thm 74. *If $X \sim U((0,1])$, then there are i.i.d random variables $X_1, X_2, \ldots \sim$ Bernoulli$(1/2)$ so that*

$$X = \sum_{k=1}^{\infty} \frac{X_k}{2^k}. \tag{9.6}$$

Conversely, if (9.6) holds for independent $X_1, X_2, \ldots$ with distribution Bernoulli$(1/2)$, then $X \sim U((0,1])$.

Proof. The first conclusion is contained in Lemma 73. Suppose that (9.6) holds for a sequence of independent Bernoulli random variables $X_1, X_2, \ldots$, with parameter $1/2$. Define $Y_n = \sum_{k=1}^{n} X_k/2^k$. Clearly $Y_n \to X$, *a.s.*, hence converges in distribution. Thus it suffices to show that the pointwise limit of the distribution functions of Y_n is indeed the distribution function of $U((0,1])$, so that $X \sim U((0,1])$. For each

$n \geqslant 1$, Y_n takes 2^n distinct values,

$$Y_n \in \left\{ \frac{0}{2^n}, \frac{1}{2^n}, ..., \frac{2^n - 1}{2^n} \right\},$$

so the distribution function of Y_n is a step function with jumps at $i/2^n$, $i = 0, 1, ..., 2^n - 1$ and

$$P\left(Y_n = \frac{i}{2^n}\right) = \frac{1}{2^n}. \tag{9.7}$$

For any $y \in (0, 1]$, the dyadic expansion generation algorithm tells us that for all $r \geqslant 1$, y is contained in a dyadic interval D_r of rank r (Imagine that y plays the role of ω in (9.5)). The boundary points of D_r can be explicitly written down in terms of y. Each dyadic interval of rank r has length $1/2^r$, the number of these intervals that come before y is $[y/(1/2^r)]$, so

$$D_r = \left(\frac{1}{2^r}[2^r y], \frac{1}{2^r}[2^r y] + \frac{1}{2^r} \right].$$

It follows for all $r \geqslant 1$,

$$P\left(Y_n \leqslant \frac{1}{2^r}[2^r y]\right) \leqslant P(Y_n \leqslant y) \leqslant P\left(Y_n \leqslant \frac{1}{2^r}[2^r y] + \frac{1}{2^r}\right).$$

Now we take $r = n$ and show that both of the extreme terms converge to y as soon as $n \to 0$, it would follow that $P(Y_n \leqslant y)$ converges to the identity function $y \mapsto y$ on $(0, 1]$, the distribution function of $U((0, 1])$. We only compute the LHS, the RHS is handled similarly. By (9.7),

$$p_n(y) \triangleq P\left(Y_n \leqslant \frac{1}{2^n}[2^n y]\right) = \sum_{i=0}^{[2^n y]} P\left(Y_n = \frac{i}{2^n}\right) = \frac{[2^n y] \wedge (2^n - 1)}{2^n},$$

Note if $y = 1$, $[2^n y] = 2^n$, then the effective upper bound of the summation is $2^n - 1$, since the maximal value of Y_n is $(2^n - 1)/2^n$. Since

$$\frac{2^n y - 1}{2^n} < \frac{[2^n y]}{2^n} \leqslant \frac{2^n y}{2^n},$$

so $p_n(y) \to y$. The proof is completed. $\square$

Thm 75. *Given a probability measure μ, there exist a sequence of i.i.d random variables with μ being the common distribution.*

Proof. Let $X(\omega) = \omega$, then $X \sim U((0,1])$. By Theorem 74, there are independent $X_1, X_2, \ldots \sim Bernoulli(1/2)$ so that (9.6) holds. Let α be the one-to-one mapping from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$ and define

$$Y_{ij} = X_{\alpha(i,j)} \text{ and } U_i = \sum_{j=1}^{\infty} \frac{Y_{ij}}{2^j}, \ \ i = 1, 2, \ldots$$

Then $U_1, U_2, \ldots$ are independent by Theorem 62. By Theorem 74 again, we see that $U_1, U_2, \ldots$ have uniform distribution $U((0,1])$. Let F be the distribution function associated with μ and F^{-1} the inverse distribution function, then Theorem 63 and the proof of Theorem 23 tell us that $F^{-1}(U_1)$, $F^{-1}(U_2)$, ... are i.i.d with common distribution F. $\square$

10. 大数律 Law of large numbers

10.1. L_2 weak law.

Thm 76 (**L_2 weak law**). *Suppose that $X_1, ..., X_n$ are pairwise uncorrelated with $EX_i = \mu$ and $Var(X_i) \leqslant C < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to \mu \text{ in } L_2 \text{ and probability.}$$

PROOF. We only show L_2 convergence, which will give convergence of probability via Markov inequality. Using the variance of sum formula (Theorem 67), we have

$$E\left|\frac{S_n}{n} - \mu\right|^2 = \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\mathrm{Var}(S_n)}{n^2} = \frac{\sum_i \mathrm{Var}(X_i)}{n^2} \leqslant \frac{C}{n} \to 0.$$

□

An important special case of the L_2 weak law is the following.

Thm 77. *Suppose that $X_1, ..., X_n$ are i.i.d with $EX_i = \mu$ and $Var(X_i) = \sigma^2$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to \mu \text{ in } L_2 \text{ and probability.}$$

Below is a probabilistic proof of Weierstrass approximation theorem.

Example 37 (Bernstein polynomial). *f is continuous on $[0,1]$. Define*

$$f_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) C_n^k x^k (1-x)^{n-k}, \ \forall x \in [0,1].$$

Then

$$\sup_{x\in[0,1]} |f_n(x) - f(x)| \to 0.$$

PROOF. Observe that, if we let $S_n \sim Bin(n,x)$, then

$$f_n(x) = Ef(S_n/n).$$

Hence

$$\begin{aligned} |f_n(x) - f(x)| &= |Ef(S_n/n) - f(x)| = |E[f(S_n/n) - f(x)]| \\ &\leqslant E|f(S_n/n) - f(x)|. \end{aligned} \tag{10.1}$$

$\forall \varepsilon > 0$, we want to bound the rightmost expectation in terms of ε. Since f is continuous on $[0,1]$ and hence uniformly coninuous, we can fix δ small so that

$$|f(s) - f(t)| < \varepsilon \text{ if } |s - t| < \delta.$$

Let $M = \sup_{x\in[0,1]} f(x)$ and $A_n = \{\omega : |S_n(\omega)/n - x| < \delta\}$. Then (10.1) continues

$$\begin{aligned}
|f_n(x) - f(x)| &\leqslant E(|f(S_n/n) - f(x)|1_{A_n}) + E\big(|f(S_n/n) - f(x)|1_{A_n^c}\big) \\
&\leqslant \varepsilon + 2MP(|S_n/n - x| \geqslant \delta) \\
&\leqslant_{(e_1)} \varepsilon + 2M\frac{\mathrm{Var}(S_n/n)}{\delta^2} =_{(e_2)} \varepsilon + 2M\frac{x(1-x)}{n\delta^2} \\
&\leqslant \varepsilon + \frac{M}{2n\delta^2} \leqslant 2\varepsilon,
\end{aligned}$$

as soon as n is large so that $M/\big(2n\delta^2\big) \leqslant \varepsilon$, where (e_1) uses Markov inequality, and (e_2) Example 29. $\square$

LEMMA 78. *If b_n satisfies $Var(S_n)/b_n^2 \to 0$, then*

$$\frac{S_n - ES_n}{b_n} \to 0 \text{ in } L_2 \text{ and probability.}$$

PROOF. We have

$$\mathrm{Var}\left(\frac{S_n - ES_n}{b_n}\right) = \frac{\mathrm{Var}(S_n)}{b_n^2} \to 0.$$

□

Example 38 (Coupon collector's problem). *Suppose there are n types of coupons. You get one coupon each time you open a box of candy, and the coupon is equally likely to be any of the n types. We are interested in the time T_n to collect a complete set of coupons. Let $\tau_0^n = 0$ and*

$$\tau_k^n = \text{the first time we have } k \text{ different coupons}, \ k = 1, ..., n.$$

Then

$$T_n = \tau_n^n = \sum_{k=1}^{n} \left(\tau_k^n - \tau_{k-1}^n\right).$$

It is readily seen that the waiting times $\left\{\tau_k^n - \tau_{k-1}^n\right\}_{k=1}^n$ *between two types of coupons are independent and each has geometric distribution,*

$$\tau_k^n - \tau_{k-1}^n \sim Geom\left(1 - \frac{k-1}{n}\right).$$

Example 16 tells us that

$$ET_n = \sum_{k=1}^n \left(1 - \frac{k-1}{n}\right)^{-1} = n \sum_{m=1}^n m^{-1} \approx n \log n.$$

and

$$Var(T_n) = \sum_{k=1}^n \left(1 - \frac{k-1}{n}\right)^{-2} = n^2 \sum_{m=1}^n m^{-2} \leqslant n^2 \sum_{m=1}^\infty m^{-2}.$$

Since $\displaystyle\sum_{m=1}^\infty m^{-2}$ *is convergent, if we take* $b_n = n \log n$*, then*

$$\frac{Var(T_n)}{b_n^2} \leqslant \frac{n^2 \sum_{m=1}^\infty m^{-2}}{(n \log n)^2} \to 0.$$

So Lemma 78 gives

$$\frac{T_n - n\sum_{m=1}^{n} m^{-1}}{n\log n} \to 0 \textit{ in probability.}$$

It follows that

$$\frac{T_n}{n\log n} \to 1 \textit{ in probability.}$$

This tells us that T_n is roughly $n\log n$.

Example 39 (Random permutation). *A permutation of $\{1, ..., n\}$ is a one-to-one mapping from $\{1, ..., n\}$ to itself. There are $n!$ permutations in total. We are interested in the expected number of cycles in a randomly chosen permutation. As an example, we look at the permutation $i \mapsto \pi(i)$,*

$$\begin{array}{r cccccc} i: & 1 & 2 & 3 & 4 & 5 & 6 \\ \pi(i): & 2 & 5 & 6 & 4 & 1 & 3 \end{array}$$

Starting with 1, we follow the route of mapping

$$1 \to \pi(1) \to \pi^2(1) \to \pi^3(1) \to \cdots$$

Since $\pi^3(1) = 1$, *we get a cycle*

$$1 \to 2 \to 5 \to 1.$$

We use brackets to indicate cycles, so we have the first cycle (125), *and the remaining cycles are* (36), (4). *The original permutation can now be simply written as the decomposition*

$$(125)(36)(4).$$

The representations of a permutation as a mapping and block decomposition are equivalent . This is indeed how random permutation generation algorithm works: decompose $1, ..., n$ *into disjoint blocks, at the* k*-th position of the decomposition, the algorithm has choices with equal probability among the* $n - k$ *numbers that have not been seen so far (if the block is to grow) plus the first number of the current block the algorithm is in (if the block is to close so that a cycle is formed).*

Therefore, if we define the indicator random variables

$$X_{n,k} = \begin{cases} 1, & \text{a closing bracket occurs after the } k\text{-th position in the decomposition,} \\ 0, & \text{otherwise.} \end{cases}$$

then

$$S_n \triangleq X_{n,1} + \cdots + X_{n,n}$$

gives the total number of cycles in the permutation and

$$P(X_{n,k} = 1) = \frac{1}{n-k+1}.$$

It can also be verified that for $k < l$,

$$P(X_{n,k} = 1, X_{n,l} = 1) = \frac{1}{n-k+1} \cdot \frac{1}{n-l+1},$$

which implies that $X_{n,k}$, $X_{n,l}$ are independent (by Example 27 and Theorem 58). The same direct computation can prove the independence *of* $\{X_{n,1}, ..., X_{n,n}\}$. *Then*

$$ES_n = \sum_{k=1}^{n} EX_{n,k} = \sum_{k=1}^{n} k^{-1}$$

and noting $X_{n,k}^2 = X_{n,k}$ *we have*

$$Var(S_n) \leqslant \sum_{k=1}^{n} EX_{n,k}^2 = \sum_{k=1}^{n} EX_{n,k} = \sum_{k=1}^{n} k^{-1}.$$

Now applying Lemma 78 with $b_n = (\log n)^{0.5+\varepsilon}, \varepsilon > 0,$

$$\frac{S_n - \sum_{k=1}^{n} k^{-1}}{b_n} \to 0 \text{ in probability.}$$

It follows that, if $\varepsilon = 0.5,$

$$\frac{S_n}{(\log n)^{0.5+\varepsilon}} \to 1 \text{ in probability.}$$

The arbitrariness of ε *indicates that* $(\log n)^{0.5}$ *is a threshold for the convergence.*

Random permutation is commonly used in applications from coding to games, one example is the **100 prisoners riddle**. 100 prisoners, who are numbered from 1 to 100, are offered a last chance to be pardoned. At a room, there is a cupboard with 100 drawers. 100 numbers from 1 to 100 are randomly put into these drawers. The prisoners enter the room one by one. Each prisoner can open up to 50 drawers. No communications are allowed. If every prisoner finds their numbers, all prisoners are set free, otherwise all will be sentenced. If every prisoner randomly opens 50 drawers, the survival probability would be $(1/2)^{100}$. The prisoners need to figure out the best strategy to follow.

The numbers in the drawers form a permutation π of $\{1, ..., 100\}$, the drawer labelled with i contains the number $\pi(i)$. The permutation is decomposed as collections of cycles. The strategy is thus to enter the correct cycle containing the wanted number. For the prisoner with number i_0, the first drawer to open is the one labelled with i_0, subsequently with label $\pi(i_0)$, $\pi^2(i_0)$, ... Since every number is in some cycle , there is k, $1 \leqslant k \leqslant n$, so that $\pi^k(i_0) = i_0$, i.e., the wanted number i_0 would be found after opening the drawer labelled with $\pi^{k-1}(i_0)$. The prisoners survive the test if the random permutation in the drawer contains no cycle of length strictly greater than 50 (there is at most one in every permutation). The probability of a random permutation containing a cycle of length k is

$$\frac{C_{100}^k \cdot (k-1)! \cdot (100-k)!}{100!}.$$

Therefore the survival probability of all prisoners is then equal to

$$1 - \sum_{k=51}^{100} \frac{C_{100}^k \cdot (k-1)! \cdot (100-k)!}{100!} = 1 - \sum_{k=51}^{100} \frac{1}{k}.$$

10.2. Weak law of large numbers.

Thm 79 (Weak law for triangular arrays)**.** *Consider the triangular array of random variables* $X_{n,k}$, $k = 1, ..., n$,

$$\begin{array}{lllll} X_{1,1} & & & & \\ X_{2,1} & X_{2,2} & & & \\ \cdots & & & & \\ X_{n,1} & \cdots & X_{n,k} & \cdots & X_{n,n} \end{array}$$

Random variables in each row are pairwise independent. Let $b_n > 0$ *satisfies*
(10.2)

$$(i)\ \sum_{k=1}^{n} P(|X_{n,k}| > b_n) \to 0;\ (ii)\ \frac{\sum_{k=1}^{n} Var\Big(X_{n,k} 1_{\left|X_{n,k}\right| \leqslant b_n}\Big)}{b_n^2} \to 0.$$

If we set $S_n = \sum_{k=1}^{n} X_{n,k}$, $a_n = \sum_{k=1}^{n} E\Big(X_{n,k} 1_{\left|X_{n,k}\right| \leqslant b_n}\Big)$, *then*

$$\frac{S_n - a_n}{b_n} \to 0 \text{ in probability.}$$

Proof. Let

$$\bar{S}_n = \sum_{k=1}^{n} X_{n,k} 1_{\left|X_{n,k}\right| \leqslant b_n} \text{ and } Z_n = \frac{S_n - a_n}{b_n}.$$

We have $\forall \varepsilon > 0$,

$$\begin{aligned} P(|Z_n| > \varepsilon) &= P\big(|Z_n| > \varepsilon,\ S_n \neq \bar{S}_n\big) + P\big(|Z_n| > \varepsilon,\ S_n = \bar{S}_n\big) \\ &\leqslant P\big(S_n \neq \bar{S}_n\big) + P\big(|Z_n| > \varepsilon,\ S_n = \bar{S}_n\big). \end{aligned}$$

By assumption (i),

$$P\big(S_n \neq \bar{S}_n\big) \leqslant \sum_{k=1}^{n} P(|X_{n,k}| > b_n) \to 0.$$

Now using assumption (ii) and that $X_{n,i}, X_{n,j}, i \neq j$ are independent, we have

$$P\big(|Z_n| > \varepsilon,\ S_n = \bar{S}_n\big) \leqslant P\left(\left|\frac{\bar{S}_n - a_n}{b_n}\right| > \varepsilon\right) \leqslant \frac{\operatorname{Var}\big(\bar{S}_n\big)}{\varepsilon^2 b_n^2}$$

$$= \frac{\sum_{k=1}^{n} \operatorname{Var}\Big(X_{n,k} 1_{|X_{n,k}| \leqslant b_n}\Big)}{\varepsilon^2 b_n^2} \to 0.$$

The proof is complete. $\square$

LEMMA 80. *If* $X \geqslant 0$, φ *differentiable with* $\varphi' > 0$ *and* $\varphi(0) = 0$, *then*

$$\int \varphi(X) dP = \int_0^\infty \varphi'(t) P(X > t) dt.$$

Proof. An application of Theorem 56 with $Y = \varphi(X)$ gives,

$$\int Y dP = \int_0^\infty P(Y > s) ds$$
$$=_{(e_1)} \int_0^\infty \varphi'(t) P(Y > \varphi(t)) dt = \int_0^\infty \varphi'(t) P(X > t) dt.$$

We have performed in (e_1) a change of variable $s = s(t) = \int_0^t \varphi'$. $\square$

Thm 81 (Weak law of large numbers). *Let $X_1, ..., X_n$ be i.i.d with*

$$xP(|X_1| > x) \to 0 \text{ as } x \to 0. \tag{10.3}$$

If we set $S_n = X_1 + \cdots + X_n$, $\mu_n = E\big(X_1 1_{|X_1| \leqslant n}\big)$, then

$$\frac{S_n}{n} - \mu_n \to 0 \text{ in probability.}$$

Proof. We want to apply Theorem 79 with $X_{n,k} = X_k$ and $b_n = n$. To do this, we need to verify condition (10.2). First note that

$$\sum_{k=1}^{n} P(|X_{n,k}| > b_n) = nP(|X_1| > n) \to 0$$

and

$$\frac{\sum_{k=1}^{n} \operatorname{Var}\Big(X_{n,k} 1_{|X_{n,k}| \leqslant b_n}\Big)}{b_n^2} = \frac{\operatorname{Var}(X_1 1_{|X_1| \leqslant n})}{n} \leqslant \frac{E\Big(\big(X_1 1_{|X_1| \leqslant n}\big)^2\Big)}{n},$$

recalling Example 17. Thus the proof would be completed if we show that

$$\frac{E\Big(\big(X_1 1_{|X_1| \leqslant n}\big)^2\Big)}{n} \to 0.$$

By Lemma 80,

$$E\Big(\big(X_1 1_{|X_1| \leqslant n}\big)^2\Big) = \int_0^\infty 2tP\big(|X_1| 1_{|X_1| \leqslant n} > t\big)dt. \tag{10.4}$$

Note the integrand has

$$P\big(|X_1|1_{|X_1|\leqslant n} > t\big) = P(|X_1| > t, |X_1| \leqslant n),$$

which gives the expression,

$$P\big(|X_1|1_{|X_1|\leqslant n} > t\big) = \begin{cases} P(|X_1| > t) - P(|X_1| > n), & t < n \\ 0, & t \geqslant n \end{cases}$$

Hence upon substituting the above in (10.4) we obtain

$$\frac{E\Big(\big(X_1 1_{|X_1|\leqslant n}\big)^2\Big)}{n} \leqslant \frac{1}{n}\int_0^n 2tP(|X_1| > t)dt.$$

Using the assumption that $tP(|X_1| > t) \to 0$ as $t \to \infty$, we see that the RHS converges to zero, which completes the proof. $\square$

A sufficient condition for (10.3) is $E|X_1| < \infty$. Indeed, by dominated convergence theorem,

$$xP(|X_1| > x) \leqslant E\big(|X_1|1_{|X_1|>x}\big) \to 0 \text{ as } x \to \infty.$$

So $E|X_1| < \infty$ implies (10.3), and is thus a stronger condition, but the latter is not much weaker since by Lemma 80, for $0 < \varepsilon < 1$,

$$
\begin{aligned}
E|X|^{1-\varepsilon} &= \int_0^\infty (1-\varepsilon)t^{-\varepsilon}P(X > t)dt \\
&= \int_0^1 (1-\varepsilon)t^{-\varepsilon}P(X > t)dt + \int_1^\infty (1-\varepsilon)t^{-\varepsilon}P(X > t)dt \\
&\leqslant \int_0^1 t^{-\varepsilon}dt + \int_1^\infty t^{-(1+\varepsilon)}tP(X > t)dt < \infty.
\end{aligned}
$$

Thm 82. *Let $X_1, ..., X_n$ be i.i.d with $E|X_1| < \infty$. If we set $S_n = X_1 + \cdots + X_n$, then*

$$\frac{S_n}{n} \to EX_1 \textit{ in probability.}$$

Proof. Let $\mu = EX_1$, $\mu_n = E\big(X_1 1_{|X_1| \leqslant n}\big)$. As we have already seen that $E|X_1| < \infty$ implies (10.3), so we can employ Theorem 81 to conclude that, $\forall \varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu_n\right| > \varepsilon\right) \to 0.$$

Since $\mu_n \to \mu$ by dominated convergence theorem, we have $|\mu_n - \mu| < \varepsilon$ for large n, therefore

$$P\left(\left|\frac{S_n}{n} - \mu\right| > 2\varepsilon\right) \leqslant P\left(\left|\frac{S_n}{n} - \mu_n\right| > \varepsilon\right).$$

It follows that $S_n/n - \mu \to 0$ in probability. $\square$

In the example below, we will see that weak law can exist even if the condition of Theorem 82 fails: $E|X_1| = \infty$.

Example 40 (St. Petersburg paradox). *A single-player game begins with an initial wager of 2 dollars and a fair coin. The coin is tossed repeatedly. Each time a tail comes up, the wager is doubled. The game ends if head appears. So if the first toss is head, the game ends and the player receives 2 dollars. The expected amount the player* *would receive is*

$$2 \cdot \frac{1}{2} + 2^2 \cdot \frac{1}{2^2} + 2^3 \cdot \frac{1}{2^3} + \cdots = \infty.$$

Paradoxically, no one would pay an infinite amount to play a game. We want to use the weak law Theorem 79 to find the right value of the game. The idea is to see where the average game value goes after playing several rounds of the game. Let X_1, X_2, ... be independent with values in $\{2^m : m = 1, 2, ...\}$ and satisfy

$$P(X_k = 2^m) = 2^{-m}.$$

To apply Theorem 79, we need to find $b_n > 0$ so that (10.2) is satisfied with $X_{n,k} = X_k$. Since

$$nP(X_1 > b_n) = n \sum_{m:2^m > b_n} 2^{-m}$$

and

$$nb_n^{-2}\, Var(X_1 1_{X_1 \leqslant b_n}) \leqslant nb_n^{-2} E(X_1 1_{X_1 \leqslant b_n})^2 \leqslant nb_n^{-2} \sum_{m:2^m \leqslant b_n} 2^{2m} \cdot 2^{-m}$$

So (10.2)(*i*) *and* (*ii*) *translate as requiring*

$$n \sum_{m:2^m > b_n} 2^{-m} \to 0 \text{ and } nb_n^{-2} \sum_{m:2^m \leqslant b_n} 2^{2m} \cdot 2^{-m} \to 0. \tag{10.5}$$

We assume that

$$m(n) \triangleq \log_2 b_n = \log_2 n + K(n),$$

where $K(n)$ *is chosen so that* $m(n)$ *is an integer. Then*

$$n \sum_{m:2^m > b_n} 2^{-m} \leqslant n2^{-m(n)} = 2^{-K(n)}$$

and

$$\begin{aligned} nb_n^{-2} \sum_{m:2^m \leqslant b_n} 2^{2m} \cdot 2^{-m} &= nb_n^{-2} \frac{2\left(2^{m(n)+1} - 1\right)}{2-1} \\ &\leqslant 4nb_n^{-2} 2^{m(n)} \leqslant 4nb_n^{-1} = 4 \cdot 2^{-K(n)}. \end{aligned}$$

Hence all it takes for (10.5) to hold is $K(n) \to \infty$ while keeping $m(n)$ an integer. Thus Theorem 79 tells us that, with $S_n = \sum_{k=1}^{n} X_k$,

$$\frac{S_n - a_n}{n2^{K(n)}} \to 0 \text{ in probability.} \tag{10.6}$$

where

$$a_n = nE\big(X_1 1_{|X_1| \leqslant b_n}\big) = n \sum_{m:2^m \leqslant b_n} 2^m \cdot 2^{-m} = nm(n) = n(\log_2 n + K(n)).$$

To draw a meaningful conclusion, we choose $K(n)$ so that

$$\frac{a_n}{n2^{K(n)}} = \frac{\log_2 n + K(n)}{2^{K(n)}} \to 1.$$

In particular, if $K(n) \approx \log_2 \log_2 n$ for large n, then the above is satisfied and (10.6) gives

$$\frac{S_n}{n \log_2 n} \to 1 \text{ in probability.}$$

This says that the average S_n/n of n rounds of the game is close to $\log_2 n$, which should therefore be a reasonable price for the game.

10.3. Borel-Cantelli lemma and applications.

LEMMA 83 (**The first Borel-Cantelli lemma**). *We have*

$$\sum_{k=1}^{\infty} P(A_n) < \infty \text{ implies } P(A_n \text{ i.o.}) = 0.$$

PROOF. We have

$$P\left(\limsup_n A_n\right) = \lim_n P\left(\bigcup_{k=n}^{\infty} A_k\right) \leqslant \lim_n \sum_{k=n}^{\infty} P(A_k) = 0.$$

□

The next is a typical application of Borel-Cantelli lemma, the application to the strong law of large numbers is postponed to Theorem 88.

Lemma 84. *Suppose that $\varepsilon_n \geqslant 0$ satisfies $\sum_n \varepsilon_n < \infty$ and the random variables X_n have*

$$\sum_{n=1}^{\infty} P(|X_{n+1} - X_n| > \varepsilon_n) < \infty.$$

Then there exits a finite random variable X so that $X_n \to X$, a.s.

Proof. Let $A_n = \{|X_{n+1} - X_n| > \varepsilon_n\}$ and $A^* = \limsup_n A_n$. Then for $\omega \in (A^*)^c = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_k^c$ if and only if there is $m(\omega)$ satisfying

$$|X_{n+1}(\omega) - X_n(\omega)| \leqslant \varepsilon_n \text{ for } n \geqslant m(\omega),$$

hence $\{X_n(\omega)\}$ is Cauchy and converges to some finite limit, say $X^*(\omega)$. Define $X = 0$ for $\omega \in A^*$ and $X = X^*$ otherwise. Then X is a random variable since A^* is measurable. Using the first Borel-Cantelli lemma, we have $P(A^*) = 0$ which shows that $X_n \to X$, *a.s.* □

Lemma 85 (**The second Borel-Cantelli lemma**). *If $A_1, ..., A_n$ are independent, then*

$$\sum_{n=1}^{\infty} P(A_n) = \infty \textit{ implies } P(A_n \textit{ i.o.}) = 1.$$

Proof. It suffices to show $P\left(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k^c\right) = 0$, which clearly follows if we can show that $P\left(\bigcap_{k=n}^{\infty} A_k^c\right) = 0$ for all n. By independence and $1 - x \leqslant e^{-x}$,

$$P\left(\bigcap_{k=n}^{N} A_k^c\right) = \prod_{k=n}^{N}(1 - P(A_k)) \leqslant \prod_{k=n}^{N} e^{-P(A_k)} = \exp\left\{-\sum_{k=n}^{N} P(A_k)\right\}.$$

The latter converges to zero as $N \to \infty$. Hence

$$P\left(\bigcap_{k=n}^{\infty} A_k^c\right) = \lim_N P\left(\bigcap_{k=n}^{N} A_k^c\right) = 0.$$

$\square$

COROLLARY 4 (**Zero-One law**). *If $A_1, ..., A_n$ are independent, then $P(A_n \text{ i.o.}) = 0$ or 1 according as $\sum_{n=1}^{\infty} P(A_n)$ converges or diverges.*

Borel-Cantelli lemmas are easier to understand when translated into the language of random variables. Let

$$S_n = \sum_{k=1}^{n} 1_{A_k} \text{ and } S = \sum_{k=1}^{\infty} 1_{A_k}, \tag{10.7}$$

then we have the translation of Borel-Cantelli lemmas in the language of random variables,

$$\text{Lemma 83}: \quad ES < \infty \text{ implies } S < \infty \ a.s.$$

$$\text{Lemma 85}: \quad \begin{array}{l} \text{If } A_1, A_2, ... \text{ are independent, then} \\ ES = \infty \text{ implies } S = \infty \ a.s. \end{array}$$

Note by monotone convergence theorem $\lim_n ES_n = ES$.

With the random variable translation, we can easily show that the second Borel-Cantelli lemma (Lemma 85) continues to hold if independence is replaced with pairwise independence.

LEMMA 86 (**The second Borel-Cantelli lemma**). *If $A_1, ..., A_n$ are pairwise independent, then*

$$\sum_{n=1}^{\infty} P(A_n) = \infty \text{ implies } P(A_n \ i.o.) = 1.$$

Proof. Let $S_n = \sum_{k=1}^{n} 1_{A_k}$ and $S = \sum_{k=1}^{\infty} 1_{A_k}$. We see that the desired conclusion is equivalent to $P(S < \infty) = 0$. By pairwise independence,

$$\mathrm{Var}(S_n) = \sum_{k=1}^{n} \mathrm{Var}(1_{A_k}) \leqslant \sum_{k=1}^{n} P(A_k) = ES_n.$$

Since $S_n \leqslant S$, we have

$$\begin{aligned} P(S < ES_n/2) &\leqslant P(S_n < ES_n/2) \\ &\leqslant P(|S_n - ES_n| > ES_n/2) \leqslant \frac{4\mathrm{Var}(S_n)}{(ES_n)^2} \leqslant \frac{4}{ES_n}. \end{aligned}$$

Noting that $ES_n \to \infty$ by assumption, we proceed to write

$$P(S < \infty) = \lim_n P(S < ES_n/2) \leqslant \lim_n \frac{4}{ES_n} = 0.$$

$\square$

REMARK 6. *Assume the same conditions as Lemma 86. A slight modification of the proof of Lemma 86 yields that,* $\forall \delta > 0$,

$$P(|S_n - ES_n| > \delta ES_n) \leqslant \frac{Var(S_n)}{\delta^2 (ES_n)^2} \leqslant \frac{1}{\delta^2 ES_n} \tag{10.8}$$

So we have

$$ES_n \to \infty \text{ implies } \frac{S_n}{ES_n} \to 1 \text{ in probability.}$$

This can also be derived directly from Lemma 78.

Through a useful technique which we call the **method of subsequence**, we show that the above convergence can be strengthened and it is indeed almost sure. We will again see the use of the method of subsequence in the proof of the strong law (Theorem 89).

Thm 87. *If* $A_1, ..., A_n$ *are pairwise independent, then*

$$\sum_{n=1}^{\infty} P(A_n) = \infty \text{ implies } \frac{\sum_{k=1}^{n} 1_{A_k}}{\sum_{k=1}^{n} P(A_k)} \to 1, \ a.s.$$

PROOF. Let $N_n = \sum_{k=1}^{n} 1_{A_k}$, we want to prove that $N_n/EN_n \to 1$, $a.s.$ To do this, we proceed in two steps.

1. First we show that the conclusion is true for a subsequence. Let

$$\tau_k = \inf\{n : EN_n \geqslant k\} \text{ and } S_k = N_{\tau_k}.$$

By the definition $EN_{\tau_{k^2}-1} < k^2$ (otherwise τ_{k^2} would be no greater than $\tau_{k^2} - 1$, a contradiction). So

$$k^2 \leqslant ES_{k^2} = EN_{\tau_{k^2}-1} + E1_{A_{\tau_{k^2}}} < k^2 + 1. \tag{10.9}$$

Notice that the assumption $EN_n \to \infty$ ensures that $\tau_k < \infty$ and $\tau_k \to \infty$ so that (10.9) holds for all $k \geqslant 1$, otherwise if $\sup_n EN_n < \infty$, then for large k, $\tau_k = \infty$ ($\inf \varnothing = \infty$) and $k^2 \leqslant ES_{k^2}$ would not hold. Now applying (10.8) to the subsequence $k \mapsto S_{k^2}$ gives, $\forall \delta > 0$,

$$P(|S_{k^2} - ES_{k^2}| > \delta ES_{k^2}) \leqslant \frac{1}{\delta^2 ES_{k^2}} \leqslant \frac{1}{\delta^2 k^2}.$$

So the first Borel-Cantelli Lemma (Lemma 83) shows that

$$\frac{S_{k^2}}{ES_{k^2}} \to 1, \; a.s.$$

2. Next we extend the conclusion from S_{k^2} to the whole sequence. Note that for any n with $\tau_{k^2} \leqslant n \leqslant \tau_{(k+1)^2}$,

$$\frac{S_{k^2}}{ES_{(k+1)^2}} \leqslant \frac{N_n}{EN_n} \leqslant \frac{S_{(k+1)^2}}{ES_{k^2}},$$

which can be rewritten as

$$\frac{S_{k^2}}{ES_{k^2}} \cdot \frac{ES_{k^2}}{ES_{(k+1)^2}} \leqslant \frac{N_n}{EN_n} \leqslant \frac{S_{(k+1)^2}}{ES_{(k+1)^2}} \cdot \frac{ES_{(k+1)^2}}{ES_{k^2}}.$$

Thus the desired conclusion follows if $ES_{(k+1)^2}/ES_{k^2} \to 1$, but this is immediate since by the definition of S_{k^2}, $S_{(k+1)^2}$ and (10.9),

$$1 \leqslant \frac{ES_{(k+1)^2}}{ES_{k^2}} < \frac{(k+1)^2 + 1}{k^2} \to 1.$$

Therefore the proof is completed. $\square$

Example 41 (Record value). *Suppose that i.i.d random variables $X_1, X_2, ...$ from a continuous distribution function F are observed sequentially. Denote by $A_k = \{X_k > X_i \text{ for } i = 1, .., k-1\}$ the event that a record occurs at the k-th random variable. We want to determine the asymptotics of the count*

$$R_n = \sum_{k=1}^{n} 1_{A_k}$$

of record events in the first n random variables. Since the distribution function is continuous, the values of $X_1, X_2, ..., X_n$ are almost surely distinct [4]. *By rearranging $X_1, X_2, ..., X_n$ in decreasing order, we obtain a permutation π_n over $1, ..., n$, where all $n!$ permutations are equally likely. The event A_k occurs if and only if the k-th position is the greatest among the first k, this is, in the language of permutation, $\pi_k(k) = 1$. Note that the permutation after the k-th position does not affect that of the first k. There are only one way to put the greatest (of*

[4] Durrett 5th Exercise 2.1.5

the first k) at the k-th position and the remaining can be permuated in any of $(k-1)!$ ways. Hence

$$P(A_k) = P(\pi_k(k) = 1) = \frac{1 \cdot (k-1)!}{k!} = \frac{1}{k}.$$

The same idea generalizes to multiple record events, for example, for $k < l$,

$$P(A_k A_l) = \frac{1 \cdot (l-1)!}{l!} \cdot \frac{1 \cdot (k-1)!}{k!} = \frac{1}{k} \cdot \frac{1}{l} = P(A_k)P(A_l).$$

With these it can be verified that $A_1, A_2, ..., A_n$ are independent. Now we can employ Theorem 87 to conclude that

$$\frac{R_n}{\log n} \to 1, \ a.s.$$

Note the conclusion is independent of F as long as it is continuous.

10.4. Strong law of large numbers. Our first version of strong law of large numbers is a typical application of the first Borel-Cantelli lemme (Lemma 83)

Thm 88. *Let $X_1, ..., X_n$ be i.i.d with $EX_1^4 < \infty$. If we set $S_n = X_1 + \cdots + X_n$, then*

$$\frac{S_n}{n} \to EX_1 \ a.s.$$

Proof. Assuming without loss of generality that $EX_1 = 0$, we observe that the desired conclusion amounts to, $\forall \varepsilon > 0$,

$$P(|S_n| > n\varepsilon \text{ i.o.}) = 0. \tag{10.10}$$

We have by Markov inequality that

$$P(|S_n| > n\varepsilon) \leqslant \frac{ES_n^4}{(n\varepsilon)^4}. \tag{10.11}$$

Now

$$ES_n^4 = E\left(\sum_{1\leqslant i,j,k,l\leqslant n} X_i X_j X_k X_l\right) = \sum_{1\leqslant i,j,k,l\leqslant n} E(X_i X_j X_k X_l).$$

By the i.i.d assumption and that $EX_1 = 0$, we see from Theorem 66 that $E(X_i X_j X_k X_l)$ is zero unless it is of either one of the form EX_i^4,

$EX_i^2X_j^2$ with $i \neq j$. There are respectively n and $C_4^2 \cdot C_n^2 = 3n(n-1)$ of these terms (for the latter, pick two indices out of i, j, k, l and then two distinct random variables out of $X_1, ..., X_n$). Hence

$$ES_n^4 = nEX_i^4 + 3n(n-1)EX_i^2X_j^2 = nEX_1^4 + 3n(n-1)\big(EX_1^2\big)^2 \leqslant Cn^2,$$

where C is a constant independent of n. Plugging this into (10.11), we obtain

$$P(|S_n| > n\varepsilon) \leqslant \frac{C}{n^2\varepsilon^4}.$$

Hence $\displaystyle\sum_{n=1}^{\infty} P(|S_n| > n\varepsilon) < \infty$, so (10.10) follows from the first Borel-Cantelli lemma (Lemma 83). $\square$

The i.i.d and fourth order moment assumption of Theorem 88 can be weakened. Next we give Etemadi's proof of **Kolmogorov's strong law of large numbers** under pairwise independence and finite first order moment.

Thm 89 (Strong law of large numbers). *Suppose that $X_1, ..., X_n$ are pairwise independent identically distributed with $E|X_1| < \infty$. If we set $S_n = X_1 + \cdots + X_n$, then*

$$\frac{S_n}{n} \to EX_1 \ a.s.$$

PROOF. We start by observing that if the theorem holds for non-negative random variable, then

$$\frac{S_n}{n} = \frac{1}{n}\left(\sum_{k=1}^{n} X_k^+ - \sum_{k=1}^{n} X_k^-\right) \to EX_1^+ - EX_1^- = EX_1 \ a.s.$$

So we can assume from now on that $X_k \geqslant 0$, $k \geqslant 1$. As in Theorem 79, we define the truncated partial sum

$$\bar{S}_n = \sum_{k=1}^{n} X_k 1_{X_k \leqslant k}.$$

Let $\alpha > 1$ and $\tau_n = [\alpha^n]$.

1. We first show that

$$\sum_{n=1}^{\infty} P\left(\left|\bar{S}_{\tau_n} - E\bar{S}_{\tau_n}\right| > \varepsilon\tau_n\right) < \infty.$$

As usual

$$\operatorname{Var}\left(\bar{S}_{\tau_n}\right) = \sum_{k=1}^{\tau_n} \operatorname{Var}(X_k 1_{X_k \leqslant k}) \leqslant \sum_{k=1}^{\tau_n} E\left(X_k^2 1_{X_k \leqslant k}\right) \leqslant \tau_n E\left(X_1^2 1_{X_1 \leqslant \tau_n}\right).$$

Hence

$$\sum_{n=1}^{\infty} P\left(\left|\bar{S}_{\tau_n} - E\bar{S}_{\tau_n}\right| > \varepsilon\tau_n\right) \leqslant \sum_{n=1}^{\infty} \frac{\operatorname{Var}\left(\bar{S}_{\tau_n}\right)}{\varepsilon^2\tau_n^2} \leqslant \frac{1}{\varepsilon^2} E\left[X_1^2 \sum_{n=1}^{\infty} \frac{1_{X_1 \leqslant \tau_n}}{\tau_n}\right].$$

For $x = X_1(\omega) > 0$, let $n_x = \min\{n \in \mathbb{N} : \tau_n \geqslant x\}$. By the definition we have $\tau_n \geqslant \alpha^n/2$ [5] and $\alpha^{n_x} \geqslant \tau_{n_x} \geqslant x$, it follows that

$$\sum_{n=1}^{\infty} \frac{1_{X_1 \leqslant \tau_n}}{\tau_n} = \sum_{n \geqslant n_x} \frac{1}{\tau_n} \leqslant 2 \sum_{n \geqslant n_x} \alpha^{-n} = \frac{2\alpha^{-n_x}}{1-\alpha^{-1}} \leqslant \frac{2x^{-1}}{1-\alpha^{-1}} = \frac{2X_1^{-1}}{1-\alpha^{-1}}.$$

[5] For $z \geqslant 1$, $z/2 < [z]$: if $z \in [1,2)$, $z/2 < 1 = [z]$; if $z \geqslant 2$, $z - [z] < z/2$.

Therefore

$$\sum_{n=1}^{\infty} P\big(\big|\bar{S}_{\tau_n} - E\bar{S}_{\tau_n}\big| > \varepsilon\tau_n\big) \leqslant \frac{2}{\varepsilon^2(1-\alpha^{-1})} EX_1 < \infty.$$

2. Next we claim that

$$\frac{S_{\tau_n}}{\tau_n} \to EX_1,\ a.s.$$

With what we already have from step **1**, we can invoke the first Borel-Cantelli lemma to obtain that

$$\frac{\bar{S}_{\tau_n} - E\bar{S}_{\tau_n}}{\tau_n} \to 0,\ a.s.$$

But $EX_k 1_{X_k \leqslant k} \to EX_1$ by dominated convergence theorem, it follows that $E\bar{S}_{\tau_n}/\tau_n \to EX_1$. Hence $\bar{S}_{\tau_n}/\tau_n \to EX_1$, $a.s.$ Since

$$\begin{aligned}\sum_{k=1}^{\infty} P(X_k 1_{X_k \leqslant k} \neq X_k) \leqslant \sum_{k=1}^{\infty} P(X_k > k) &\leqslant \int_0^{\infty} P(X_1 > t)dt \\ &= EX_1 < \infty,\end{aligned}$$

invoking again the first Borel-Cantelli lemma we get that

$$P(X_k 1_{X_k \leqslant k} \neq X_k \text{ i.o.}) = 0,$$

hence $\big(S_{\tau_n} - \bar{S}_{\tau_n}\big)/\tau_n \to 0$, $a.s.$ It follows that

$$\frac{S_{\tau_n}}{\tau_n} = \frac{S_{\tau_n} - \bar{S}_{\tau_n}}{\tau_n} + \frac{\bar{S}_{\tau_n}}{\tau_n} \to EX_1, \; a.s.$$

3. Finally we conclude via the use of subsequence method. For any k satisfying $\tau_n \leqslant k \leqslant \tau_{n+1}$, since $X_k \geqslant 0$, we have

$$\frac{S_{\tau_n}}{\tau_{n+1}} \leqslant \frac{S_k}{k} \leqslant \frac{S_{\tau_{n+1}}}{\tau_n},$$

which we rewrite as

$$\frac{S_{\tau_n}}{\tau_n} \cdot \frac{\tau_n}{\tau_{n+1}} \leqslant \frac{S_k}{k} \leqslant \frac{S_{\tau_{n+1}}}{\tau_{n+1}} \cdot \frac{\tau_{n+1}}{\tau_n}.$$

But by the definition $\tau_{n+1}/\tau_n \to \alpha$, so it follows from step **2** that

$$\frac{1}{\alpha} EX_1 \leqslant \liminf \frac{S_k}{k} \leqslant \limsup \frac{S_k}{k} \leqslant \alpha EX_1, \; a.s.$$

The proof is completed by sending $\alpha \to 1$. $\square$

The next theorem shows that for i.i.d sequence, finite first moment $E|X_1| < \infty$ in Theorem 89 is not only sufficient but also necessary for the strong law to hold.

Thm 90. *Let $X_1, ..., X_n$ be i.i.d with $E|X_1| = \infty$, then*

$$P(|X_n| > n \ i.o.) = 1$$

and setting $S_n = X_1 + \cdots + X_n$,

$$P\left(\lim_n \frac{S_n}{n} \text{ exists and is finite}\right) = 0.$$

PROOF. We have by Lemma 80

$$E|X_1| = \int_0^\infty P(|X_1| > t)dt \leqslant \sum_{n=0}^{\infty} P(|X_1| > n)$$

Since $E|X_1| = \infty$, we infer from the second Borel-Cantelli lemma that $P(|X_n| > n \text{ i.o.}) = 1$. To prove the remaining conclusion, let

$$C = \left\{\omega : \lim_n \frac{S_n(\omega)}{n} \text{ exists and is finite}\right\}.$$

We claim that C does not intersect $\{|X_n| > n \text{ i.o.}\}$, thus has $P(C) = 0$. If $\omega \in C \cap \{|X_n| > n \text{ i.o.}\}$, then

$$\begin{aligned}\left|\frac{S_{n+1}(\omega)}{n+1} - \frac{S_n(\omega)}{n}\right| &= \left|\frac{S_n(\omega)}{n+1} - \frac{S_n(\omega)}{n} + \frac{X_{n+1}(\omega)}{n+1}\right| \\ &\geqslant \left|\frac{X_n(\omega)}{n+1}\right| - \left|\frac{S_n(\omega)}{n+1} - \frac{S_n(\omega)}{n}\right|.\end{aligned}$$

Whence

$$\limsup_n \left|\frac{S_{n+1}(\omega)}{n+1} - \frac{S_n(\omega)}{n}\right| \geqslant 1.$$

But this contradicts $\omega \in C$, therefore $C \cap \{|X_n| > n \text{ i.o.}\}$ must be empty. $\square$

The strong law of large numbers holds whenever EX_1 exists in the extended sense, i.e., at least one of EX_1^+, EX_1^- is finite.

Thm 91. *Let $X_1, ..., X_n$ be i.i.d with $EX_1^+ = \infty$, $EX_1^- < \infty$. Set $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to \infty \ a.s.$$

PROOF. Define $X_k^M = X_k \wedge M$, $S_n^M = X_1^M + \cdots + X_n^M$, then by the assumption, $X_1^M, ..., X_n^M$ are i.i.d with finite expectation. Hence Theorem 89 tells us that for $M > 0$,

$$\liminf_n \frac{S_n}{n} \geqslant \lim_n \frac{S_n^M}{n} \to EX_1^M, \ a.s.$$

Now let $M \to \infty$. □

The following is an application of the strong law to statistics, particularly we will prove the Glivenko-Cantelli theorem which is usually referred to as the fundamental theorem of statistics.

Example 42 (Empirical distribution function). *Let $X_1, ..., X_n$ be i.i.d with distribution function F. Fix $x \in \mathbb{R}$. We can estimate the value $F(x)$ as below. Define*

$$F_n(x, \omega) = \frac{\sum_{k=1}^n 1_{\{X_k \leqslant x\}}(\omega)}{n}.$$

The mapping $x \mapsto F_n(x, \omega)$ is the so called empirical distribution function. By the strong law (Theorem 89), for each x, there is an exception set A_x of zero probability,

$$\lim_n F_n(x, \omega) = E1_{\{X_1 \leqslant x\}} = F(x), \ \omega \in A_x^c.$$

But the theorem below says more: the exception set A_x can be independent of x.

In the following, we will drop ω and simply write F_n for empirical distribution function. But keep in mind F_n is a random function.

Before we state the main theorem, it is worthwhile recalling the exercise [6]: if G_n, G are nondecreasing functions and G is bounded and

[6] Pku Textbook Chapter 2 Exercise 32

continuous, then G_n converges to G uniformly. But if G has discontinuities, then it is not immediately obvious that the convergence is uniform.

Thm 92 (The Glivenko-Cantelli theorem). *Let $X_1, ..., X_n$ be i.i.d with distribution function F. Then the associated empirical distribution function F_n has*

$$\sup_{x \in \mathbb{R}} |F_n(x, \omega) - F(x)|, \ \omega\text{-}a.s.$$

Proof. **1**. As with Example 42, an application of the strong law shows that, for each x, there is an exception set B_x of zero probability so that

$$F_n(x-, \omega) \triangleq \frac{\sum_{k=1}^{n} 1_{\{X_k < x\}}(\omega)}{n} \to F(x-), \ \omega \in B_x^c.$$

Let $m \geqslant 1$ and $F^{-1}(z)$, $z \in (0,1)$, be the inverse distribution function (see (4.1)). Define

$$x_{m,k} = \begin{cases} F^{-1}(k/m), & 1 \leqslant k \leqslant m-1, \\ -\infty, & k = 0, \\ +\infty, & k = m. \end{cases}$$

Since

$$F\big(F^{-1}(z)-\big) \leqslant z \leqslant F\big(F^{-1}(z)\big), \ \forall z \in (0,1),$$

we have for $2 \leqslant k \leqslant m-1$,

$$F(x_{m,k}-) - F(x_{m,k-1}) \leqslant k/m - (k-1)m = 1/m.$$

The inequality remains true for $k = 1$ or m with the understanding that $F(x_{m,0}) = 0$ and $F(x_{m,m}-) = 1$. For $1 \leqslant k \leqslant m$ and $x_{m,k-1} \leqslant x < x_{m,k}$, we get by monotonicity,

$$\begin{aligned} F_n(x) - F(x) &\leqslant F_n(x_{m,k}-) - F(x_{m,k-1}) \\ &\leqslant F(x_{m,k}-) - F(x_{m,k-1}) + |F_n(x_{m,k}-) - F(x_{m,k}-)| \\ &\leqslant 1/m + |F_n(x_{m,k}-) - F(x_{m,k}-)|. \end{aligned}$$

Similarly,

$$\begin{aligned} F(x) - F_n(x) &\leqslant F(x_{m,k}-) - F_n(x_{m,k-1}) \\ &\leqslant F(x_{m,k}-) - F(x_{m,k-1}) + |F(x_{m,k-1}) - F_n(x_{m,k-1})| \\ &\leqslant 1/m + |F(x_{m,k-1}) - F_n(x_{m,k-1})| \end{aligned}$$

Note $|F(x_{m,0}) - F_n(x_{m,0})| = 0$, $|F_n(x_{m,m}-) - F(x_{m,m}-)| = 0$.

2. Let $\varepsilon > 0$. For $1 \leqslant k \leqslant m$, denote by $A_{m,k}$ the exception set for the convergence of $F_n(x_{m,k-1})$ ($A_{m,0} \triangleq \varnothing$), and $B_{m,k}$ for the convergence of $F_n(x_{m,k}-)$ ($B_{m,m} \triangleq \varnothing$). Also let $E = \bigcup_{m=1}^{\infty} \bigcup_{k=1}^{m} (A_{m,k} \cup B_{m,k})$, then $P(E) = 0$. If

$$D_n(m) \triangleq \frac{1}{m} + \max_{1 \leqslant k \leqslant m} \{|F_n(x_{m,k}-) - F(x_{m,k}-)|, |F(x_{m,k-1}) - F_n(x_{m,k-1})|\},$$

then for any m satisfying $1/m < \varepsilon/2$, there is N_m such that

$$D_n(m) \leqslant \varepsilon,\ n \geqslant N_m,\ \omega \in E^c.$$

Any $x \in \mathbb{R}$ is contained in some interval $[x_{m,k-1}, x_{m,k})$, so by step **1**, as soon as $n \geqslant N_m$,

$$|F(x) - F_n(x)| \leqslant D_n(m) \leqslant \varepsilon,\ \omega \in E^c,$$

completing the proof. □

Another application of the strong law, now to renewal theory.

Thm 93 (Renewal theory). *Imagine a number of lightbulbs produced by the same manufacturer are available. At time* 0*, a lightbulb is lit up and replaced by a new one when it burns out. The lifetime of these lightbulbs are modeled by i.i.d random variables* $X_1, X_2, \ldots.$ *with* $0 < EX_1 = \lambda^{-1} \leqslant \infty$*.We are interested in the number of lightbulbs that have burned out by time* $t > 0$*,*

$$N_t = \sup\{n : T_n \leqslant t\} = \sum_{n=1}^{\infty} 1_{T_n \leqslant t},$$

where $T_n = X_1 + \cdots + X_n$. *Then as* $t \to \infty$,

$$\frac{N_t}{t} \to \lambda, \ a.s.$$

If additionally, X_1 *(hence every* X_k*) does not concentrate on a single point, i.e. for all* $a \geqslant 0$, $P(X_1 = a) \neq 1$, *then we obtain the elementary limit theorem,*

$$\frac{EN_t}{t} \to \lambda.$$

PROOF. **1**. For $t > 0$, by the definition, $T_{N_t} \leqslant t < T_{N_t+1}$, so

$$\frac{N_t + 1}{T_{N_t+1}} \cdot \frac{N_t}{N_t + 1} = \frac{N_t}{T_{N_t+1}} < \frac{N_t}{t} \leqslant \frac{N_t}{T_{N_t}}. \tag{10.12}$$

Since $X_1, X_2, \ldots$ are i.i.d with finite expectation $EX_1 = \lambda^{-1}$, we get from the strong law Theorem 89 that

$$\frac{T_n}{n} \to \lambda^{-1}, \ a.s. \tag{10.13}$$

Moreover, we infer from the finite expectation assumption that outside an exception set, $T_n < \infty$ for all n. This together with (10.13) implies that almost surely $T_n \to \infty$ as $n \to \infty$, hence almost surely

$$N_t < \infty \text{ for all } t > 0.$$

We claim that almost surely $N_t \to \infty$ as $t \to \infty$. If this is not true for ω in a set of positive probability, then $\{N_t(\omega) : t > 0\}$ is bounded by some finite $n(\omega)$, this implies that $T_n(\omega) > t$ for all t whenever $n > n(\omega)$. Sending $t \to \infty$ would contradict that almost surely $T_n < \infty$ for all n. Now almost surely, both of the extreme terms of (10.12) are well-defined and converge to λ. Thus $N_t/t \to \lambda$ almost surely.

2. To prove the second conclusion, it sufficies to show that N_t/t, $t \geqslant 1$, is uniformly integrable (by Theorem 45). Note we are only interested in $t \to \infty$, so we have excluded $0 < t < 1$. Let

$$X_{k,a} = a1_{X_k > a} \text{ for } a > 0,\ k = 1, 2, ...$$

This may be thought of as ignoring the lightbulb whose lifetime is no greater than a threshold a, while longer lifetime is simply counted as a. As a result, the renewal time of these "modified lightbulbs" can only happen at the times that are multiples of a, $\{na : n \in \mathbb{N}\}$. So each renewal can be regarded as a geometric random variable waiting for the event $\{X_k > a\}$ to occur. To make the geometric random variable meaningful, we require that $a > 0$ satisfy

$$0 < p \triangleq P(X_1 > a) < 1.$$

We can do this since X_1 does not concentrate on any single point. The frequency of the modified renewal is higher than its unmodified counterpart. So, if

$$N_{t,a} = \sup\{n : T_{n,a} \leqslant t\},$$

where $T_{n,a} = X_{1,a} + \cdots + X_{n,a}$, then $N_t \leqslant N_{t,a}$. Since $N_{t,a}$ is less than the sum of $[t/a]$ independent geometric random variables with parameter p, hence

$$EN_{t,a}^2 \leqslant E\left(\sum_{i=1}^{[t/a]} \mathrm{Geom}(p)\right)^2 \leqslant c\big(1 + t + t^2\big),$$

where c is a constant depending on a and p only. Now for $y > 0$,

$$P\left(\frac{N_t}{t} > y\right) \leqslant P\left(\frac{N_{t,a}}{t} > y\right) \leqslant \frac{EN_{t,a}^2}{y^2t^2} \leqslant \frac{3c}{y^2}, \ \forall t \geqslant 1.$$

Denote $Y_t = N_t/t$. Using Lemma 80, we obtain that

$$\begin{aligned} EY_t 1_{Y_t > y} &= \int_0^\infty P(Y_t 1_{Y_t > y} > s) ds \\ &= \int_0^y P(Y_t > y) ds + \int_y^\infty P(Y_t > s) ds \\ &\leqslant \frac{3c}{y} + + \int_y^\infty \frac{3c}{s^2} ds \to 0 \text{ as soon as } y \to \infty. \end{aligned}$$

Whence N_t/t, $t \geqslant 1$, is uniformly integrable. $\square$

REMARK 7. *Another approach to the second conclusion of Theorem 93 is to modify the lightbulbs by cutoff,*

$$X_k^M = X_k \wedge M \ \textit{for} \ M > 0, \ \ k = 1, 2, \ldots$$

The proof will rely on Wald's equation, but not need the assumption that the lifeime is not a fixed value. However the assumption that $EX_1 > 0$ *is still in place so that* $P(X_1 = 0) = 1$ *does not happen.*

REMARK 8. *A particular application of the renewal theory assumes that the lifetime of a lightbulb is exponentially distributed with parameter* λ*. So on average, during a time period of length* t*, the number* N_t *of lightbulbs that burns out by time* t *approximates* λt*. In this sense, the parameter of exponential distribution is interpreted as the rate of events, i.e., number of events per unit time.*

11. 中心极限定理 Central limit theorem

11.1. Introduction. The central limit theorem and the law of large numbers answer two seemly disparate but related questions. Let $X_1, \ldots, X_n$ be i.i.d with $E|X_1| < \infty$, $S_n = X_1 + \cdots + X_n$. Then the strong law says that

$$\left|\frac{S_n}{n} - EX_1\right| \to 0, \ \ a.s.$$

Now the question is how large should n be so that $|S_n/n - EX_1|$ is less than a given tolerance. This concerns the speed of convergence. An answer has already been hinted by the proof of the L_2 weak law (Theorem 76) which goes as this

$$E\left|\frac{S_n}{n} - \mu\right|^2 = \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\mathrm{Var}(S_n)}{n^2} = \frac{\mathrm{Var}(X_1)}{n}.$$

This means $E(S_n/n - EX_1)^2$ grows at the same speed as n^{-1}, or roughly $|S_n/n - EX_1|$ grows at the speed $n^{-1/2}$. The statement of course is very vague, but the central limit theorem will tell us more.

11.2. From Poisson distribution to Stirling formula. Before we start, we will review a basic limit theorem and on the way provide an intuitive proof of Stirling formula.

Suppose $\lambda > 0$ is the number of events in a unit time interval, thus λ is the rate of events. The unit time interval is divided into n subintervals. Assume that *the probability of an event is approximated by the rate of the event and the probability of two events occurring in a small interval is negligible.* Hence if n is large, the probability of an event occurring on a subinterval can well be thought of as λ/n, and the total number of events that occur approximately follows the binomial distribution $Bin(n, \lambda/n)$. The following limit theorem makes this precise. Since the occurrence of an event on a subinterval of small size is rare, the theorem is commonly referred to as the **law of small numbers**.

Thm 94 (**Poisson approximation to Binomial**). *Let* $p = p(n) \to 0$, $n \to \infty$ *so that* $np \to \lambda > 0$. *Then for* $0 \leqslant k \leqslant n$,

$$\lim_{n\to\infty} C_n^k p^k (1-p)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Proof. Write

$$C_n^k p^k (1-p)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} p^k (1-p)^{n-k}.$$

Since

$$p = \frac{\lambda}{n} + o\left(\frac{1}{n}\right) = \frac{1}{n}(\lambda + o(1)),$$

we have

$$n(n-1)\cdots(n-k+1)p^k = \frac{n(n-1)\cdots(n-k+1)}{n^k}(\lambda + o(1))^k \to 1,$$

and note $k/n \to 0$, so

$$(1-p)^{n-k} = \left[1 - \left(\frac{\lambda}{n} + o\left(\frac{1}{n}\right)\right)\right]^{n-k} \to e^{-\lambda}.$$

Therefore the conclusion is proved. □

The limit theorem indicates that Poisson distribution has the same bell shape as binomial distribution.

Thm 95. *Let $X \sim Poisson(\lambda)$ with $\lambda > 0$. Write $p_k = P(X = k)$, $k \geqslant 0$. Then*

(1) *if λ is an integer,*

$$p_0 < \cdots < p_{\lambda-1} = p_\lambda > p_{\lambda+1} \cdots$$

(2) *if λ is not an integer,*

$$p_0 < \cdots < p_{[\lambda]} > p_{[\lambda]+1} \cdots$$

PROOF. Compute $p_{k+1}/p_k = \lambda/(k+1)$. $\square$

Thm 96 (Stirling formula).

$$n! \sim n^n e^{-n}\sqrt{2\pi n}$$

AN INTUITIVE PROOF. Let $X \sim \text{Poisson}(\lambda)$. Recall that $EX = \lambda$, $\text{Var}(X) = \lambda$. The probability mass function of X is bell-shaped and peaks at λ (if it is an integer) which is close to the normal density with mean λ and variance λ, at least near the peak. So

$$\frac{\lambda^\lambda}{\lambda!}e^{-\lambda} \approx \frac{1}{\sqrt{2\pi\lambda}}e^{-\frac{(\lambda-\lambda)^2}{2\lambda}} = \frac{1}{\sqrt{2\pi\lambda}},$$

which gives

$$\lambda! \approx \lambda^\lambda e^{-\lambda}\sqrt{2\pi\lambda}.$$

$\square$

11.3. De Moivre-Laplace limit theorem. Let $X_1, ..., X_n$ be i.i.d with

$$P(X_1 = -1) = P(X_1 = 1) = \frac{1}{2}.$$

Then $EX_1 = 0$, $\mathrm{Var}(X_1) = 1$. Let $S_n = X_1 + \cdots + X_n$. We intend to show that, through proper standardization, the distribution of S_{2n} is close to the standard normal distribution $\mathcal{N}(0,1)$. Since

$$ES_{2n} = 0,\ \mathrm{Var}(S_{2n}) = 2n,$$

in the spirit of standardization we should compute the distribution of $S_{2n}/\sqrt{2n}$. Note S_{2n} can only take even numbers between $-2n$ and $2n$. So we need to compute the probability of the form

$$P\left(\frac{S_{2n}}{\sqrt{2n}} = \frac{2k}{\sqrt{2n}}\right) = P(S_{2n} = 2k).$$

For each $2k$ in S_{2n}'s range (assuming $2k \geqslant 0$, the nonpositive case is symmetric), denote by u the number of X_ks that are $+1$, and d the number of X_ks that are -1, then

$$u - d = 2k,\ u + d = 2n.$$

Hence

$$u = n + k,\ d = n - k.$$

So

$$P(S_{2n} = 2k) = C_{2n}^{n+k} \cdot \frac{1}{2^{n+k}} \cdot \frac{1}{2^{n-k}} = \frac{(2n)!}{(n+k)!(n-k)!} \cdot \frac{1}{2^{2n}}.$$

Now using Stirling formula, $P(S_{2n} = 2k)$ approximates

$$\frac{(2n)^{2n} e^{-2n} \sqrt{2\pi(2n)}}{(n+k)^{n+k} e^{-n-k} \sqrt{2\pi(n+k)} \cdot (n-k)^{n-k} e^{-n+k} \sqrt{2\pi(n-k)}} \cdot \frac{1}{2^{2n}}$$
$$= \frac{n^{2n}}{(n+k)^{n+k}(n-k)^{n-k}} \cdot \frac{\sqrt{2\pi(2n)}}{\sqrt{2\pi(n+k)}\sqrt{2\pi(n-k)}}$$

which we rewrite as

$$\left(1-\frac{k^2}{n^2}\right)^{-n}\left(1+\frac{k}{n}\right)^{-k}\left(1-\frac{k}{n}\right)^{k}\cdot\frac{1}{\sqrt{\pi n}}\frac{\sqrt{n^2}}{\sqrt{(n+k)(n-k)}}$$

It follows that, if $2k/\sqrt{2n} \approx x$, then $k^2/n \approx -x^2/2$, the above product asymptotically equals

$$e^{x^2/2}e^{-x^2/2}e^{-x^2/2}\cdot\frac{1}{\sqrt{\pi n}}.$$

Therefore we have

LEMMA 97. *Let $X_1, ..., X_n$ be i.i.d with*

$$P(X_1=-1)=P(X_1=1)=\frac{1}{2}.$$

Let $S_n = X_1 + \cdots + X_n$. If $2k/\sqrt{2n} \to x$, then

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

Now we are in a position to compute, for $a < b$,

$$P\left(a \leqslant \frac{S_{2n}}{\sqrt{2n}} \leqslant b\right) = \sum_{x \in [a,b] \cap (2\mathbb{Z})/\sqrt{2n}} P\left(\frac{S_{2n}}{\sqrt{2n}} = x\right)$$
$$\approx \sum_{x \in [a,b] \cap (2\mathbb{Z})/\sqrt{2n}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{\sqrt{2}}{\sqrt{n}}.$$

The RHS approximates a Riemann sum of the function $e^{-x^2/2}/\sqrt{2\pi}$ on $[a, b]$ with grid size $\sqrt{2}/\sqrt{n}$. Since $S_{2n+1} = S_{2n} \pm 1$, the probability of $S_{2n+1}/\sqrt{2n+1} \in [a, b]$ can be approximated by the same Riemann sum, thus we have proved

Thm 98 (**De Moivre-Laplace**). *If* $a < b$, $m \to \infty$, *then*

$$P\left(a \leqslant \frac{S_m}{\sqrt{m}} \leqslant b\right) \to \frac{1}{\sqrt{2\pi}} \int_b^a e^{-\frac{x^2}{2}} dx.$$

11.4. Weak convergence. Recall that a sequence of distribution functions F_n converges weakly to a function F if $F_n(x) \to F(x)$ for all

x where F is continuous. A sequence of random variables X_n converges weakly to X if the associated distribution functions converge weakly.

It is often the case that weak convergence is defined in terms of probability measures. A sequence of probability measures μ_n converges weakly to a *probability measure* μ means the corresponding distribution function of μ_n converges weakly to the distribution function of μ.

Example 43 (**Exponential approximation to Geometric**). *Suppose that a success trial happens with a probability proportional to time length and the probability is p $(0 < p < 1)$ per unit time. For $t \geqslant 0$, divide the interval $[0, t]$ into pieces of equal size $\delta > 0$. Then the waiting time X for the first success follows Geom(δp). Whence*

$$P(X > t) = (1 - \delta p)^{t/\delta} \to e^{-pt}, \text{ for } t \geqslant 0.$$

So geometric random variable converges weakly to exponential. See Remark 8 for the interpretation of the parameter of exponential distribution.

Example 44 (Birthday problem). *Let $X_1, X_2, \ldots$ be i.i.d with uniform distribution on $\{1, \ldots, n\}$, and*

$$T_n = \min\{k : X_k = X_j \text{ for some } j < k\}$$

the first time some X_k repeats itself. Imagine X_ks are birthdays of people or coupons you collect (see Example 38), then T_n represents the first time people's birthday repeats or a type of coupon that you already have is collected. Clearly T_n takes values in $\{2, \ldots, n+1\}$. For $1 \leqslant k \leqslant n$, the occurrence of the event $T_n > k$ amounts to the first k observed values $X_1, \ldots, X_k$ being distinct, so

$$P(T_n > k) = \prod_{j=1}^{k} \left(1 - \frac{j-1}{n}\right).$$

It follows[7] *that for $x \geqslant 0$, $k \approx n^{1/2}x$, we have*

$$P\left(T_n > n^{1/2}x\right) \to e^{-x^2/2} \text{ as } n \to \infty.$$

[7]Use Exercise 3.1.1 with $c_{jn} = -\dfrac{j-1}{n}$ if $j \leqslant k = \left[n^{1/2}x\right]$, and 0 if $j > k$.

Thm 99 (Scheffé Theorem). *Suppose that $p_n(x)$, $p(x)$ are probability densities, and $p_n \to p$, a.s. Let*

$$\nu_n(A) = \int_A p_n(x)dx,\ \nu(A) = \int_A p(x)dx$$

and

$$\|\nu_n - \nu\|_{TV} \triangleq \sup_{A \in \mathscr{B}(\mathbb{R})} |\nu_n(A) - \nu(A)|.$$

Then

$$\|\nu_n - \nu\|_{TV} \leqslant \int_{\mathbb{R}} |p_n - p| dx \to 0.$$

Proof. The inequality follows from the property of integral. To see $\displaystyle\int |p_n - p| \to 0$, we note that, since

$$\int_{\mathbb{R}} (p - p_n)dx = 0,$$

we have

$$\int_{\mathbb{R}} (p - p_n)^+ dx = \int_{\mathbb{R}} (p - p_n)^- dx.$$

It follows that

$$\int_{\mathbb{R}} |p_n - p| dx = 2 \int_{\mathbb{R}} (p - p_n)^+ dx.$$

The RHS converges to zero, since $(p - p_n)^+ \leqslant p$, p is integrable and $(p - p_n)^+ \to 0$, $a.s.$ so that the dominated convergence theorem is applicable. □

REMARK 9. *Given measures ν_n, ν, $\|\nu_n - \nu\|_{TV}$ defined in Theorem 99 is referred to as the **total variation norm** which gauges the discrepancy between ν_n and ν. Since*

$$\sup_{A=(-\infty,x],x\in\mathbb{R}} |\nu_n(A) - \nu(A)| \leqslant \|\nu_n - \nu\|_{TV},$$

so convergence in total variation norm implies weak convergence of distributions. But the converse is not necessarily true. Consider the *example,*

$$\nu_n = \delta_{1/n}, \ \nu = \delta_0,$$

where δ_a is the Dirac measure concentrated on a. Then $\|\nu_n - \nu\|_{TV} = 1$ but for any x,

$$\nu_n((-\infty, x]) = 1_{[1/n,\infty)}(x) \to 1_{[0,\infty)}(x) = \nu_0((-\infty, x]).$$

Example 45 (**Order statistics**). *A sample of n points are picked randomly from the interval $(0,1)$, i.e. the locations $X_1, X_2, ..., X_n$ are i.i.d with common uniform distribution on $(0,1)$. The density of the k-th largest $X_{(k)}$ is*

$$p_k(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}.$$

Proof. As in Example 41, we may assume that $X_1, X_2, ..., X_n$ are distinct. As a matter of fact, the probabality of more than one points landing in a small interval of size δx is $O\big((\delta x)^2\big)$, which is already negligible. If $X_{(k)} \in (x, x+\delta x)$, then there are exactly $k-1$ points whose location $< x$, and $n-k$ points $> x + \delta x$. Imagine arranging n

balls into the interval $(0,1)$, one in the infinitesimal interval $(x, x+\delta x)$ (there are n ways to pick one ball), $k-1$ to the left of it (there are C_{n-1}^{k-1} ways to pick $k-1$ balls out of $n-1$) and $n-k$ to the right. The probability is

$$\begin{aligned} P\big(X_{(k)} \in (x, x+\delta x)\big) &= nC_{n-1}^{k-1} \cdot \delta x \cdot (1-x-\delta x)^{n-k} \\ &= \frac{n!}{(k-1)!(n-k)!} x^{k-1} \cdot \delta x \cdot (1-x-\delta x)^{n-k}. \end{aligned}$$

Therefore

$$p_k(x) = \lim_{\delta x \to 0} \frac{P\big(X_{(k)} \in (x, x+\delta x)\big)}{\delta x} = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}.$$

□

The preceding example shows that $X_{(k)}$ has Beta distribution (Definition 46) with parameter k and $n-k+1$, i.e.

$$p_k(x) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} x^{k-1}(1-x)^{n-k},\ 0 < x < 1.$$

Now assume that $2n+1$ points are picked randomly from $(0,1)$. We will find the weak limit of the **central order statistics** $X_{(n+1)}$ via Theorem 99. By Example 45, $X_{(n+1)}$ has density

$$p_{n+1}(x) = \frac{(2n+1)!}{n!n!} x^n (1-x)^n \sim \text{Beta}(n+1, n+1).$$

Then by Example 36,

$$EX_{(n+1)} = \frac{1}{2},\ \text{Var}\big(X_{(n+1)}\big) = \frac{1}{4(2n+3)}.$$

Consider the standardization of $X_{(n+1)}$,

$$Y_n = 2\sqrt{2n}\left(X_{(n+1)} - \frac{1}{2}\right).$$

Since we are interested in large n asymptotics, we have replaced $(2n+3)$ with $2n$. Through a change of variable $x = 1/2 + y/\left(2\sqrt{2n}\right)$, the density of Y_n is found to be

$$\begin{aligned} p_{Y_n}(y) &= \frac{(2n+1)!}{n!n!}\left(\frac{1}{2}+\frac{y}{2\sqrt{2n}}\right)^n\left(\frac{1}{2}-\frac{y}{2\sqrt{2n}}\right)^n\frac{1}{2\sqrt{2n}} \\ &= C_{2n}^n 2^{-2n}\left(1-\frac{y^2}{2n}\right)^n\frac{2n+1}{2n}\frac{\sqrt{n}}{\sqrt{2}}. \end{aligned}$$

The factor $C_{2n}^n 2^{-2n}$ is identified to be $P(S_{2n}=0)$ from Lemma 97, thus asymptotic to $1/\sqrt{\pi n}$, hence

$$p_{Y_n}(y) \to \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}.$$

It follows from Theorem 99 that Y_n converges weakly to the normal distribution $\mathcal{N}(0,1)$.

LEMMA 100. *Let g be a measurable function from $\mathbb{R}^n$ to $\mathbb{R}$, D the set of discontinuous points of g. Then D is Borel measurable, i.e. $D \in \mathscr{B}(\mathbb{R}^n)$.*

PROOF. For any positive rationals ε, ρ, let

$$D(\varepsilon, \rho) = \{x \in \mathbb{R}^n : \exists\, y, z \in B_\rho(x) \text{ such that } |g(y) - g(z)| \geqslant \varepsilon\},$$

where $B_\rho(x) = \{x' : |x' - x| < \rho\}$. Then

$$D = \bigcup_{\varepsilon} \bigcap_{\rho} D(\varepsilon, \rho).$$

We claim that $D(\varepsilon, \rho)$ is open so that the conclusion follows immediately. Indeed, let $x \in D(\varepsilon, \rho)$, then there are $y, z \in B_\rho(x)$ such that $|g(y) - g(z)| \geqslant \varepsilon$. There is an open ball $B_{\rho_1}(x)$ such that every point in it has a distance less than ρ from y,

$$|y - b| < \rho, \ \forall b \in B_{\rho_1}(x).$$

and there is another open ball $B_{\rho_2}(x)$ such that every point in it has a distance less than ρ from z,

$$|z - b| < \rho, \ \forall b \in B_{\rho_2}(x).$$

Now take the intersection $B_{\rho_1 \wedge \rho_2}(x)$ of the two open balls. Clearly for every x' in $B_{\rho_1 \wedge \rho_2}(x)$, we have that y, $z \in B_\rho(x')$ and $|g(y) - g(z)| \geqslant \varepsilon$. This confirms that $D(\varepsilon, \rho)$ is an open set. $\square$

Thm 101 (Continuous mapping theorem). *Let g be a measurable function from $\mathbb{R}$ to $\mathbb{R}$, D_g the set of discontinuities of g. If $X_n \to X$ weakly and $P(X \in D_g) = 0$, then*

$$g(X_n) \to g(X) \text{ weakly.}$$

If additionally g is bounded, then $Eg(X_n) \to Eg(X)$.

REMARK 10. *By Lemma 100, it makes sense to write $P(X \in D_g)$.*

PROOF. **1**. In view of Skorohod theorem 26, there are Y_n, Y defined on a common probability space $(\Omega, \mathscr{F}, \mu)$ so that

$$Y_n \stackrel{d}{=} X_n, Y \stackrel{d}{=} X \text{ and } Y_n \to Y, a.s.$$

Since Y and X have identical distribution, we get that

$$\mu(Y \in B) = P(X \in B) \text{ for any } B \in \mathscr{B}(\mathbb{R}).$$

Setting $B = D_g$ gives $\mu(Y \in D_g) = 0$. Thus the almost sure convergence $Y_n \to Y$, $a.s.$ remains true outside D_g. But g is continuous in the complement of D_g, so overall we have $g(Y_n) \to g(Y)$ $a.s.$ which implies that $g(Y_n) \to g(Y)$ weakly (Theorem 24). If z is a continuity point of $g(X)$, i.e. $P(g(X) = z) = 0$, then using that $g(Y) \stackrel{d}{=} g(X)$, we see that z is also a continuity point of $g(Y)$. Hence

$$\mu(g(Y_n) \leqslant z) \to \mu(g(Y) \leqslant z) = P(g(X) \leqslant z).$$

Now using that $g(Y_n) \stackrel{d}{=} g(X_n)$, we have

$$\mu(g(Y_n) \leqslant z) = P(g(X_n) \leqslant z).$$

It follows that

$$P(g(X_n) \leqslant z) \to P(g(X) \leqslant z).$$

Therefore $g(X_n) \to g(X)$ weakly.

2. If g is bounded, then bounded convergence theorem gives $Eg(Y_n) \to Eg(Y)$. By the identical distribution construction, we have $Eg(Y_n) = Eg(X_n)$ and $Eg(Y) = Eg(X)$, thus $Eg(X_n) \to Eg(X)$. $\square$

Weak convergence can be characterized through dual actions with bounded continuous functions.

Thm 102. *Let X_n and X be random variables. The following are equivalent.*

(1) $X_n \to X$ *weakly.*

(2) *For every bounded continuous function g, it holds that*

$$Eg(X_n) \to Eg(X).$$

(3) *For every P**-continuity set** A, i.e.*

$$A \in \mathscr{B}(\mathbb{R}) \text{ and } P(X \in \partial A) = 0,$$

it holds that

$$P(X_n \in A) \to P(X \in A).$$

Proof. **1**. (1) $\Rightarrow$ (2) and (1) $\Rightarrow$ (3) follow from the continuous mapping theorem (Theorem 101).

2. (3) $\Rightarrow$ (1). For every $A = (-\infty, x]$, $P(X \in A) = 0$ means that x is a continuous point of the distribution function of X, thus $P(X_n \in A) \to P(X \in A)$ is the same thing as the distribution function of X_n being convergent to that of X at x.

3. (2) $\Rightarrow$ (1). Fix $x \in \mathbb{R}$. For $y > x$, consider the continuous piecewise linear function $g(s)$ which equal $1_{(-\infty,x]}$ if $s \leqslant x$, linear on $s \in [x, y]$ and equals 0 on $s \geqslant y$. Then

$$P(X_n \leqslant x) = E1_{(-\infty,x]}(X_n) \leqslant Eg(X_n).$$

Hence

$$\limsup_n P(X_n \leqslant x) \leqslant \limsup_n Eg(X_n) = Eg(X) \leqslant P(X \leqslant y).$$

Now let $y \downarrow x$ gives

$$\limsup_n P(X_n \leqslant x) \leqslant P(X \leqslant x).$$

Similarly, for $z < x$, consider the continuous piecewise linear function $h(s)$ which equal $1_{(-\infty,z]}$ if $s \leqslant z$, linear on $s \in [z, x]$ and equals 0 on $s \geqslant x$. Then

$$P(X_n \leqslant x) \geqslant Eh(X_n).$$

Hence

$$\liminf_n P(X_n \leqslant x) \geqslant \liminf_n Eh(X_n) = Eh(X) \geqslant P(X \leqslant z).$$

Now let $z \uparrow x$ gives

$$\liminf_n P(X_n \leqslant x) \geqslant P(X < x).$$

Putting these inequalities together, we have

$$P(X < x) \leqslant \liminf_n P(X_n \leqslant x) \leqslant \limsup_n P(X_n \leqslant x) \leqslant P(X \leqslant x).$$

So the weak convergence follows. $\square$

Thm 103 (Helly's selection theorem). *Every sequence of distribution functions F_n has a subsequence F_{n_k} which converges to a* *nondecreasing, right-continuous function F at every continuity point of F.*

Proof. Note $0 \leqslant F_n \leqslant 1$, $\forall n$. Following a standard diagonal procedure, we may find a subsequence F_{n_k} which converges at every rational number $q \in \mathbb{Q}$ to some $G(q)$. Clearly G is nondecreasing. Let

$$F(x) = \inf\{G(q) : x < q\},\ \forall x.$$

Then F is nondecreasing.

1. F is right continuous at any x. For $\varepsilon > 0$, there is $q \in \mathbb{Q}$, $x < q$ such that

$$F(x) \leqslant G(q) < F(x) + \varepsilon.$$

If $x \leqslant y < q$, then $F(y) \leqslant G(q)$. It follows that

$$F(y) \leqslant G(q) < F(x) + \varepsilon.$$

Hence F is right continuous.

2. If F is continuous at x, then $F_{n_k}(x) \to F(x)$ hence the conclusion follows. To see this, first note that by continuity, for $\varepsilon > 0$ there is $y < x$ such that

$$F(x) - \varepsilon < F(y).$$

Now choose rational numbers q, r so that $y < q < x < r$ and $G(r) < F(x) + \varepsilon$. Then

$$F(x) - \varepsilon < F(y) \leqslant G(q) \leqslant G(r) < F(x) + \varepsilon.$$

Therefore by the definition of G and the monotonicity of F_n, for large k,

$$F(x) - \varepsilon \leqslant F_{n_k}(q) \leqslant F_{n_k}(x) \leqslant F_{n_k}(r) \leqslant F(x) + \varepsilon.$$

It follows that

$$F(x) - \varepsilon \leqslant \liminf_k F_{n_k}(x) \leqslant \limsup_k F_{n_k}(x) \leqslant F(x) + \varepsilon.$$

Finally let $\varepsilon \to 0$. $\square$

REMARK 11. *The limit function F of Helly's selection theorem necessarily has $0 \leqslant F \leqslant 1$, but the theorem does not claim that F is a distribution function, the reason being that probability mass could escape to infinity. A necessary and sufficient condition to avoid this situation is **tightness**.*

Def 47. *A sequence of probability measures μ_n on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is* ***tight*** *if $\forall \varepsilon$, there is an interval $(a, b]$ such that*

$$\mu_n((a, b]) > 1 - \varepsilon, \ \forall n.$$

In terms of the corresponding distribution functions F_n, this means that $\forall \varepsilon$, there is $M_\varepsilon > 0$ such that

$$\limsup_n F_n(-M_\varepsilon) + 1 - F_n(M_\varepsilon) \leqslant \varepsilon.$$

Thm 104. *A sequence of probability measures μ_n on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is tight if and only if every subsequence has a further weakly convergent subsequence.*

Proof. **1**. Suppose that $\{\mu_n\}$ is tight. Denote by F_n the distribution function corresponding to μ_n. In view of Helly's selection theorem, there is a subsequence F_{n_k} which converges to some nondecreasing, right-continuous function F at every continuity point of F. Then there is a unique measure μ such that $\mu((a, b]) = F(b) - F(a)$, $\forall a, b$. Due to tightness, $\forall \varepsilon > 0$ we can find $(a, b]$ so that $\mu_n((a, b]) > 1 - \varepsilon$,

$\forall n$. Additionally we may increase b and decrease a so that they are continuity points of F. Now taking limit in the equation

$$F_{n_k}(b) - F_{n_k}(a) = \mu_{n_k}((a,b]) > 1 - \varepsilon$$

yields $\mu((a,b]) \geqslant 1 - \varepsilon$. But ε is arbitrary, hence μ is a probability measure.

2. Suppose the stated subsequential property holds but $\{\mu_n\}$ is not tight. Then there is $\varepsilon_0 > 0$ so that for every interval $(a,b]$, there is an index n' having

$$\mu_{n'}((a,b]) \leqslant 1 - \varepsilon_0.$$

This is particularly true for the sequence of intervals $(-k,k]$, i.e. there is a subsequence n_k so that

$$\mu_{n_k}((-k,k]) \leqslant 1 - \varepsilon_0, \ \forall k.$$

According to the stated subsequential property, there is a subsequence of $\{n_k\}$, say $\{n_{k_l}\}$, such that $\mu_{n_{k_l}}$ converges weakly to some probability measure μ. Choose $(a,b]$ with $\mu(a,b] > 1 - \varepsilon_0$ and a, b being the continuity point of μ, i.e. $\mu(\{a\}) = \mu(\{b\}) = 0$. Hence $\mu_{n_{k_l}}(a,b] \to$

$\mu(a, b]$ as $l \to \infty$. Therefore as soon as l is large, we can have the interval $(a, b]$ contained in $(-k_l, k_l]$ and

$$1 - \varepsilon_0 < \mu_{n_{k_l}}(a, b] \leqslant \mu_{n_{k_l}}((-k, k]) \leqslant 1 - \varepsilon_0.$$

A contradiction. $\square$

Below is a sufficient condition for tightness.

Thm 105. *Suppose* $\varphi \geqslant 0$ *and* $\lim_{|x| \to \infty} \varphi = \infty$. *If*

$$C = \sup_n \int \varphi(x) dF_n(x) < \infty,$$

then F_n *is tight.*

PROOF. Write μ_n for the probability measure corresponding to F_n (Recall Remark 3). Consider the interval $(a,b]$. We have

$$\mu_n((a,b]^c) = \int 1_{(a,b]^c}(x)d\mu_n(x) \leqslant \int \frac{1_{(a,b]^c}(x)}{\varphi(x)}\varphi(x)d\mu_n(x)$$
$$\leqslant \frac{1}{\inf_{x\in(a,b]^c}\varphi}\int \varphi(x)d\mu_n(x) \leqslant \frac{C}{\inf_{x\in(a,b]^c}\varphi}.$$

By the assumption, the rightmost term converges to zero as $a,b\to\infty$. Therefore for $\varepsilon>0$, there are a, b so that

$$\mu_n((a,b]) > 1-\varepsilon,\ \forall n.$$

$\square$

Example 46. *Let $k\geqslant 1$, and X_n have distribution function F_n. If $C=\sup\limits_n EX_n^k<\infty$, then F_n is tight.*

11.5. Characteristic function.

Def 48. *For the complex valued random variable*

$$X = Y + iZ,$$

its expectation is defined to be

$$EX = EY + iEZ.$$

The **Euler's formula** is frequently used: for all real number x,

$$e^{ix} = \cos x + i \sin x.$$

Def 49. *The characteristic function of a probability measure μ on $\mathbb{R}$ is defined to be*

$$\varphi(t) = \int_{\mathbb{R}} e^{itx} d\mu(x) = \int_{\mathbb{R}} \cos(tx) d\mu(x) + i \int_{\mathbb{R}} \sin(tx) d\mu(x).$$

The characteristic function of a random variable X with distribution μ is

$$\varphi(t) = Ee^{itX} = \int_{\mathbb{R}} e^{itx} d\mu(x).$$

Note the complex number e^{itX} has a bounded modulus: $\left|e^{itX}\right| \leqslant 1$, hence the characteristic function of a probability measure or random variable always exists.

Thm 106. *All characteristic functions have the following properties.*

(1) $\varphi(0) = 0$; $\varphi(-t) = \overline{\varphi(t)}$;

(2)

$$|\varphi(t)| = \left|Ee^{itX}\right| \leqslant E\left|e^{itX}\right| \leqslant 1;$$

(3)

$$Ee^{it(aX+b)} = e^{itb}\varphi(at);$$

(4) $\varphi(t)$ *is uniformly continuous on* $\mathbb{R}$, *moreover*

$$|\varphi(t+h) - \varphi(t)| \leqslant E\left|e^{ihX} - 1\right|.$$

Proof. Omitted. □

Thm 107 (Convolution). *If X and Y are independent with respective characteristic functions φ_X, φ_Y, then $X + Y$ has the characteristic function $\varphi_{X+Y} = \varphi_X \cdot \varphi_Y$.*

Proof.

$$Ee^{it(X+Y)} = E\left(e^{itX} \cdot e^{itY}\right) = Ee^{itX} \cdot Ee^{itY}.$$

$\square$

The inversion formula below indicates that probability measures with identical characterisitc function must be equal. But before proceeding to prove the formula, it is useful to recall the facts from anaysis,

$$\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}.$$

A variant of this formula will be used in the proof of the inversion formula. For $a \in \mathbb{R}$,

$$\int_{-\infty}^\infty \frac{\sin ax}{x} dx = 2 \int_0^\infty \frac{\sin ax}{x} dx = \operatorname{sign}(a) \cdot \pi, \tag{11.1}$$

where

$$\operatorname{sign}(a) = \begin{cases} -1, & \text{if } a < 0, \\ 0, & \text{if } a = 0, \\ 1, & \text{if } a > 0. \end{cases}$$

Thm 108 (The inversion formula). *Let φ be the characteristic function of the probability measure μ. Then for $a < b$,*

$$\mu((a,b)) + \frac{1}{2}\mu(\{a,b\}) = \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt.$$

PROOF. Let

$$I_T = \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{2\pi it}\varphi(t)dt = \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{2\pi it}\left(\int_{\mathbb{R}} e^{itx} d\mu(x)\right)dt.$$

Since the integrand is bounded,

$$\left|\frac{e^{-ita} - e^{-itb}}{it} e^{itx}\right| \leqslant \left|\frac{e^{-ita} - e^{-itb}}{it}\right| = \left|\int_a^b e^{-itx} dx\right| \leqslant |b - a|,$$

we can employ Fubini theorem to write

$$I_T = \int_{\mathbb{R}}\left(\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{2\pi it} dt\right) d\mu(x).$$

Since

$$e^{it(x-a)} - e^{it(x-b)}$$
$$= \cos(t(x-a)) + i\sin(t(x-a)) + \cos(t(x-b)) - i\sin(t(x-b)),$$

the real part is even in t. It follows that

$$I_T = \int_{\mathbb{R}} \left(\int_{-T}^{T} \frac{\sin(t(x-a)) - \sin(t(x-b))}{2\pi t} dt \right) d\mu(x).$$

From Equation (11.1), we see that, as soon as T is larger than some $T_0 > 0$, the bracketed integrand is bounded by

$$\left| \int_{-T}^{T} \frac{\sin(t(x-a))}{2\pi t} dt \right| + \left| \int_{-T}^{T} \frac{\sin(t(x-b))}{2\pi t} dt \right|$$
$$\leqslant \frac{1}{2}|\mathrm{sign}(x-a)| + \frac{1}{2}|\mathrm{sign}(x-b)| + 1.$$

Hence we can invoke the bounded convergence theorem to conclude that

$$\lim_{T\to\infty} I_T = \int_{\mathbb{R}} \left(\frac{1}{2}\mathrm{sign}(x-a) - \frac{1}{2}\mathrm{sign}(x-b) \right) d\mu(x).$$

Now

$$\frac{1}{2}\operatorname{sign}(x-a) - \frac{1}{2}\operatorname{sign}(x-b) = \begin{cases} 0, & \text{if } x < a \text{ or } x > b, \\ 1/2, & \text{if } x = a \text{ or } x = b, \\ 1, & \text{if } a < x < b. \end{cases}$$

Therefore

$$\lim_{T\to\infty} I_T = \mu((a,b)) + \frac{1}{2}\mu(\{a,b\}).$$

$\square$

Remark 12. *The inversion formula implies uniqueness. If μ, ν have identical characteristic function φ. Then*

$$\mu((a,b]) = v((a,b]) = \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt, \ \forall a < b,$$

provided $\mu(\{a,b\}) = \nu(\{a,b\}) = 0$. So μ and ν coincide on intervals whose extreme points do not hold mass,

$$\mathscr{S}_0 = \{(a,b] : a,b \in \mathbb{R},\ a,b \notin A\}$$

where

$$A = \{a \in \mathbb{R} : \mu(\{a\}) \textit{ or } \nu(\{a\}) \neq 0\}.$$

But A is at most countable, hence A^c is dense. It follows that μ and ν coincide on the π-system $\{(a,b] : a,b \in \mathbb{R}\}$. Hence μ equals ν by uniqueness (Theorem 10).