Lecture Notes in Probability and Stochastic Process

Kaizhao Liu

January 22, 2026

Contents

Part I

Probability

Chapter 1

Independence

1.1 Basic Definitions

Definition 1.1.1 (independence: events)**.** Two events A, B are independent if $P(AB) = P(A)P(B)$.

Remark 1.1.2. A sequence of events $(A_n) \subset \mathcal{F}$ is said to be independent if $\mathbb{P}(\cap_{n\in\mathcal{I}} A_n) = \prod_{n\in\mathcal{I}} \mathbb{P}(A_n)$ for every finite set $\mathcal{I} \subset \mathbb{N}$.

Definition 1.1.3 (independence: random variables)**.** Two random variables X, Y are independent if for all $C, D \in \mathcal{R}$, the events $A = \{X \in C\}$ and $B = \{Y \in D\}$ are independent.

Definition 1.1.4 (independence: σ-fields)**.** Two σ-fields $\mathcal{F}$ and $\mathcal{G}$ are independent if for all $A \in \mathcal{F}$ and $B \in \mathcal{G}$, the events A and B are independent.

Remark 1.1.5. A sequence of σ-algebra $(\mathcal{A}_n) \subset \mathcal{F}$ is said to be independent if any $A_n \in \mathcal{A}_n$ we have A_n is independent.

Actually, the first definition is a special case of the second, which is a special case of the third. This can be summarized in the following theorem.

Theorem 1.1.6.
(i) If A and B are independent, then so are A^c and B , A and B^c, and A^c and B^c.
(ii) Events A and B are independent if and only if 1_A and 1_B are independent.
(iii) If X and Y are independent then $\sigma(X)$ and $\sigma(Y)$ are.
(iv) If $\mathcal{F}$ and $\mathcal{G}$ are independent, $X \in \mathcal{F}$, and $Y \in \mathcal{G}$, then X and Y are independent.

We can extend this definition in an evident way for finitely many objects. Then, an infinite collection of objects is said to be independent if every finite subcollection is.

1.2 Sufficient Conditions for Independence

Theorem 1.2.1 (π-λ theorem)**.** *If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system that contains $\mathcal{P}$, then $\sigma(\mathcal{P}) \subset \mathcal{L}$.*

Theorem 1.2.2. *Suppose $\mathcal{A}_1$ and $\mathcal{A}_2$ are π-systems on $\mathcal{F}$. If $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$, $\forall A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2$, then $\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2)$ are independent.*

Proof. Fix $A_1 \in \mathcal{A}_1$. We define the measures $\mu(A) = \mathbb{P}(A \cap A_1)$, $\nu(A) = \mathbb{P}(A)\mathbb{P}(A_1)$ $\forall A \in \mathcal{F}$. Then $\mu|_{\mathcal{A}_1} = \nu|_{\mathcal{A}_2}$ and μ, ν are finite. This shows that $\mu|_{\sigma(\mathcal{A}_1)} = \nu|_{\sigma(\mathcal{A}_2)}$ □

Chapter 2

Law of Large Numbers

2.1 Stochastic Orders

In calculus, two sequence of real numbers, $\{a_n\}$ and $\{b_n\}$, satisfy $a_n = O(b_n)$ if and only if $|a_n| \leq c\,|b_n|$ for all n and a constant c; and $a_n = o(b_n)$ if and only if $\frac{a_n}{b_n} \to 0$ as $n \to 0$.

Definition 2.1.1. Let $X_1, X_2, \cdots$ be random vectors and $Y_1, Y_2, \cdots$ be random variables defined on a common probability space.

(i) $X_n = O(Y_n)$ a.s. if and only if $\mathbb{P}(\|X_n\| = O(|Y_n|)) = 1$.

(ii) $X_n = o(Y_n)$ a.s. if and only if $\frac{X_n}{Y_n} \to 0$ a.s..

(iii) $X_n = O_p(Y_n)$ if and only if for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$ s.t. $\sup_n P(\|X_n\| \geq C_\epsilon\,|Y_n|) < \epsilon$.

(iv) $X_n = o_p(Y_n)$ if and only if $\frac{X_n}{Y_n} \to_p 0$.

2.2 WLLN

In this section we study convergence in probability and the laws of large numbers associated with this type of convergence.

Lemma 2.2.1 (moments and tails). *Let $\xi > 0$ be an random variable with $\mathbb{E}\xi \in (0, \infty)$. Then*

$$(1-r)^2 \frac{(\mathbb{E}\xi)^2}{\mathbb{E}\xi^2} \leq \mathbb{P}(\xi > rE\xi) \leq \frac{1}{r}, \quad r > 0$$

Theorem 2.2.2 (convergence in L^p implies convergence in probability). *If $p > 0$, then*

$$\mathbb{E}\,|Z_n|^p \to 0 \Longrightarrow Z_n \longrightarrow 0 \textit{ in probability.}$$

Proof. $\mathbb{P}(|Z_n| \geq \epsilon) \leq \frac{\mathbb{E}|Z_n|^p}{\epsilon^p} \to 0$ □

Theorem 2.2.3 (L^2 weak law). *Let $X_1, X_2, ...,$ be uncorrelated random variables with $EX_i = \mu$ and $\mathrm{Var}(X_i) < C < \infty$. If $S_n = X_1 + \cdots + X_n$, then as $n \to \infty$, $\frac{S_n}{n} \longrightarrow \mu$ in L^2.*

Proof. $\mathbb{E}(\frac{S_n}{n} - \mu)^2 = \mathrm{Var}(\frac{S_n}{n}) = \frac{1}{n^2}(\sum \mathrm{Var}(X_i)) \leq \frac{Cn}{n^2} \to 0$ □

2.3 Borel-Cantelli Lemmas

Borel-Cantelli lemmas are the ladders from convergence in probability to a.s. convergence if the sequence of events are not decreasing. If the sequence of events are decreasing, then convergence in probability is the same as a.s. convergence, and there is no need for Borel-Cantelli lemma.

Theorem 2.3.1 (Borel-Cantelli lemma). $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \Longrightarrow \mathbb{P}(A_n \; i.o.) = 0$.

Proof. Let $N = \sum_{n=1}^{\infty} 1_{A_n}$. $EN < \infty$ implies $N < \infty$ a.s. □

Theorem 2.3.2 (The second Borel-Cantelli lemma). *If the events A_n are independent, then*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \Longrightarrow \mathbb{P}(A_n \; i.o.) = 1$$

Proof. Let $M < N < \infty$. $1-x \le e^{-x}$ and independence imply $\mathbb{P}(\bigcap_{n=M}^{N} A_n^c) = \prod_{n=M}^{N}(1-\mathbb{P}(A_n)) \ge \exp(-\sum_{n=M}^{N} P(A_n)) \to 0$ as $N \to \infty$, so $\mathbb{P}(\bigcup_{n=M}^{N} A_n) = 1, \forall M$. Therefore $\mathbb{P}(\limsup A_n) = 1$. □

Theorem 2.3.3 (Kochen-Stone lemma)**.** *Suppose* $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. *If*

$$\limsup_{n\to\infty} \frac{(\sum_{k=1}^{n} P(A_k))^2}{(\sum_{1\le i,j\le n} \mathbb{P}(A_i \cap A_k))} = \alpha > 0$$

then $\mathbb{P}(A_n \text{ i.o.}) \ge \alpha$.

Remark 2.3.4. This is a generalization of 2.3.2.

Theorem 2.3.5. *If* $A_1, A_2, ...$ *are pairwise independent and* $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, *then* ♡

2.4 SLLN

2.5 0-1 Laws

Theorem 2.5.1 (Kolmogorov's 0-1 law)**.** *If* $X_1, X_2, \cdots$ *are independent and* $A \in \mathcal{T}$, *then* $\mathbb{P}(A) = 0$ *or* 1.

Proof. The key point is to show that A is independent of itself.

To show this, we can procede by two limiting steps. □

Theorem 2.5.2 (Hewitt-Savage 0-1 law)**.** *If* $X_1, X_2, \cdots$ *are i.i.d. and* $A \in \mathbb{E}$, *then* $\mathbb{P}(A) = 0$ *or* 1.

Lemma 2.5.3.

2.6 Convergence of Random Series

Theorem 2.6.1 (Kolmogorov's maximal inequality)**.** *Suppose* $X_1, \cdots, X_n$ *are independent with* $EX_i = 0$ *and* $Var(X_i) < \infty$. *If* $S_n = X_1 + \cdots + X_n$, *then*

$$\mathbb{P}(\max_{1\le k\le n} |S_k| \ge x) \le \frac{Var(S_n)}{x^2}$$

Proof. There is a proof by Doob's inequality. □

Theorem 2.6.2. *Suppose* $X_1, X_2, \cdots$ *are independent and have* $EX_n = 0$. *If*

$$\sum_{n=1}^{\infty} Var(X_n) < \infty$$

then with probability one $\sum_{n=1}^{\infty} X_n(\omega)$ *converges.*

Proof. Let $S_N = \sum_{n=1}^{N} X_n$. From Kolmogorov's maximal inequality, we get

$$\mathbb{P}(\max_{M\le m\le N} |S_m - S_M| > \epsilon) \le \epsilon^{-2}\mathrm{Var}(S_N - S_M) = \epsilon^{-2} \sum_{n=M+1}^{N} \mathrm{Var}(X_n)$$

Letting $N \to \infty$, we get

$$\mathbb{P}(\sup_{M\le m} |S_m - S_M| > \epsilon) \le \epsilon^{-2} \sum_{n=M+1}^{\infty} \mathrm{Var}(X_n)$$

If we let $w_M = \sup_{m,n\ge M} |S_m - S_n|$, then

$$\mathbb{P}(w_M > 2\epsilon) \le \mathbb{P}(\sup_{M\le m} |S_m - S_M| > \epsilon) \to 0 \quad \text{as } M \to \infty$$

As w_M decreases as M increases, $w_M \downarrow 0$ a.s.. But $w_M(\omega) \downarrow 0$ implies $S_n(\omega)$ is a Cauchy sequence and hence $\lim_{n\to\infty} S_n(\omega)$ exists. □

Theorem 2.6.3 (Kolmogorov's three series theorem)**.** *Let $X_1, X_2, \cdots$ be independent. Let $A > 0$ and let $Y_i = X_i 1_{|X_i| \leq A}$. In order that $\sum_{n=1}^{\infty} X_n$ converges a.s., it is necessary and sufficient that:*
(i) $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A) < \infty$
(ii) $\sum_{n=1}^{\infty} EY_n$ converges
(iii) $\sum_{n=1}^{\infty}$ Var$(Y_n) < \infty$

Proof. To prove sufficiency, let $\mu_n = EY_n$. By the above theorem, $\sum_{n=1}^{\infty}(Y_n - \mu_n)$ converges a.s.. Using (ii), $\sum_{n=1}^{\infty} Y_n$ converges a.s.. (i) and Borel-Cantelli lemma imply $\mathbb{P}(X_n \neq Y_n \, i.o.) = 0$, so $\sum_{n=1}^{\infty} X_n$ converges a.s..
For necessity, if the sum of (i) is infinite, $\mathbb{P}(|X_n| > A \, i.o.) > 0$ and $\lim_{m\to\infty} \sum_{n=1}^{m} X_n$ can not converge. Suppose next (i) is finite but the sum □

One of the advantage of the random series proof is that it provides estimates on the rate of convergence.

Theorem 2.6.4. *Let $X_1, X_2, \cdots$ be i.i.d. random variables with $EX_i = 0$ and $EX_i^2 = \sigma^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$. If $\epsilon > 0$ then*

$$\frac{S_n}{\sqrt{n(\log n)^{1+\epsilon}}} \to 0 \ a.s.$$

The next result, show that when $\mathbb{E}\,|X_1| = \infty$, $\frac{S_n}{a_n}$ cannot converge almost surely to a nonzero limit.

Theorem 2.6.5. *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}\,|X_1| = \infty$ and let $S_n = X_1 + \cdots + X_n$. Let a_n be a sequence of positive numbers with $\frac{a_n}{n}$ increasing*

2.7 Large Deviations

Let $X_1, X_2, \cdots$ be i.i.d. and let $S_n = X_1 + \cdots + X_n$. We will investigate the rate at which $\mathbb{P}(S_n \geq na) \to 0$ for $x > \mu = EX_i$.

Lemma 2.7.1. *If $\gamma_{m+n} \geq \gamma_m + \gamma_n$, then as $n \to \infty$, $\frac{\gamma_n}{n} \to \sup_m \frac{\gamma_m}{m}$.*

Theorem 2.7.2. *$\gamma(x) = \lim_{n\to\infty} \frac{\log \mathbb{P}(S_n \geq nx)}{n}$ exists ≤ 0.*

Proof. Let $\pi_n = \mathbb{P}(S_n \geq nx)$, then $\pi_{m+n} \geq \mathbb{P}(S_m \geq mx, S_{n+m} - S_m \geq nx) = \pi_m \pi_n$. Therefore, letting $\gamma_n = \log \pi_n$, from the lemma we conclude the existence of the limit. □

Next we want to determine the limit function $h(x)$. To do this, we need to introduce the cumulant-generating function of a random varaible ξ.

$$\phi(t) = \log \mathbb{E} e^{t\xi}, \quad t \in \mathbb{R}$$

and the Legendre transform of ϕ, given by

$$\phi^*(x) = \sup_{t\in\mathbb{R}} (tx - \phi(t)), \quad x \in \mathbb{R}$$

Lemma 2.7.3. *$\phi(t)$ and $\phi^*(x)$ are convex.*

Proof. The convexity of $\phi(t)$ comes from Holder's inequality, and the convexity for $\phi^*(x)$ is a property of Legendre transform. □

Chapter 3

Central Limit Theorems

3.1 Distributions

Definition 3.1.1 (distribution)**.** If X is a random variable, then X induces a probability measure on $\mathbb{R}$ called its **distribution**.

Remark 3.1.2. The distribution of a random variable X is usually described by giving its **distribution function** $F(x) = P(X \leq x)$.

Theorem 3.1.3 (properties of distribution fucntions)**.**
(i) F is nondecreasing
(ii) $\lim_{x\to-\infty} F(x) = 0, \lim_{x\to\infty} F(x) = 1$
(iii) F is right continuous
(iv) If $F(x-) = \lim_{y\to x^-}$, then $F(x-) = P(X < x)$
(v) $P(X = x) = F(x) - F(x-)$

Proof. Directly follows from the definitions and the inclusion of sets. □

Theorem 3.1.4. *If F satisfies (i),(ii),(iii) in 3.1.3, then it is the distribution function of some random variable.*

Proof. Let $\Omega = (0,1), \mathcal{F}$=the Borel sets, and P=Lebesgue measure. If $\omega \in (0,1)$, construct

$$X(\omega) = \sup\{y : F(y) < \omega\}$$

We need to show:

$$\{\omega : X(\omega) \leq x\} = \{\omega : \omega \leq F(x)\}$$

For $\{\omega : X(\omega) \leq x\} \supseteq \{\omega : \omega \leq F(x)\}$, observe if $\omega \leq F(x)$, then $X(\omega) \leq x$.

For $\{\omega : X(\omega) \leq x\} \subseteq \{\omega : \omega \leq F(x)\}$, observe if $\omega > F(x)$, then since F is right continuous, $\exists \epsilon > 0$ s.t. $F(x+\epsilon) < \omega$. Therefore, $X(\omega) \geq x + \epsilon > x$. □

3.2 weak convergence

Definition 3.2.1 (weak convergence: distribution functions)**.** A sequence of distribution functions F_n is said to **converge weakly** to a limit F if $F_n(y) \to F(y)$ for all y that are continuity points of F.

Remark 3.2.2. Denoted by $F_n \Longrightarrow F$.

Definition 3.2.3 (weak convergence: random variable)**.** A sequence of random variables X_n is said to **converge weakly (converge in distribution)** to a limit X_∞ if their distribution functions converge weakly.

Remark 3.2.4. Denoted by $X_n \Longrightarrow X_\infty$.

Theorem 3.2.5 (Skorokhod)**.** *If $F_n \Longrightarrow F_\infty$, then $\exists$ r.v. Y_n with distribution F_n s.t. $Y_n \longrightarrow Y_\infty$ a.s.*

Proof. As in the proof of 3.1.4, let $\Omega = (0,1), \mathcal{F}$=the Borel sets, and P=Lebesgue measure. If $\omega \in (0,1)$, construct

$$Y_n(\omega) = \sup\{y : F_n(y) < \omega\}$$

We want to show:

$$Y_n(x) \longrightarrow Y_\infty(x)$$

for all but a countable number of x.

We begin by identifying the exceptional set. Let $a_x = \sup\{y : F_\infty(y) < x\}, b_x = \inf\{y : F_\infty(y) > x\}$, and $\Omega_0 = \{x : (a_x, b_x) = \emptyset\}$. Then $\Omega - \Omega_0$ is countable. If $x \in \Omega_0$, then $F_\infty(y) < x$ for $y < Y_\infty(x)$ and $F_\infty(y) > x$ for $y > Y_\infty(x)$.

Now we show $\liminf_{n\to\infty} Y_n(x) \geq Y_\infty(x)$. Choose $y < Y_\infty(x)$ s.t. F_∞ is continuous at y. Then $F_\infty(y) < x$ and $F_n(y) \longrightarrow F_\infty(y)$, so $F_n(y) < x$ for n sufficient large, that is, $Y_n(x) \geq y$. This is true for all such y's so the result follows.

The reverse inequality $\limsup_{n\to\infty} Y_n(x) \leq Y_\infty(x)$ is true by symmetry. □

Theorem 3.2.6. $X_n \Longrightarrow X_\infty \Longleftrightarrow \forall$ *bounded continuous function* $g, Eg(X_n) \longrightarrow Eg(X_\infty)$

Proof. $\Longrightarrow$: By 3.2.5, let Y_n have the same distribution as X_n and converge a.s. Since g is continuous, $g(Y_n) \longrightarrow g(Y_\infty)$ a.s. so by the bounded convergence theorem $Eg(X_n) \longrightarrow Eg(X_\infty)$.

$\Longleftarrow$: construct a bounded and continuous function

$$g_{x,\epsilon}(y) = \begin{cases} 1 & y \leq x \\ 0 & y \geq x+\epsilon \\ \text{linear} & x < y < x+\epsilon \end{cases}$$

Therefore, $\limsup_{n\to\infty} P(X_n \leq x) \leq \limsup_{n\to\infty} Eg_{x,\epsilon}(X_n) = Eg_{x,\epsilon}(X_\infty) \leq P(X_\infty \leq x+\epsilon)$. Letting $\epsilon \to 0$ gives $\limsup_{n\to\infty} P(X_n \leq x) \leq P(X_\infty \leq x)$. The reverse inequality can be proved in the same way. □

Theorem 3.2.7 (continuous mapping theorem)**.**

Theorem 3.2.8. *TFAE:*

(i) $X_n \Longrightarrow X_\infty$

(ii) For all open sets G, $\liminf_{n\to\infty} P(X_n \in G) \geq P(X_\infty \in G)$

(iii) For all closed sets K, $\limsup_{n\to\infty} P(X_n \in K) \leq P(X_\infty \in K)$

(iv) For all Borel sets A *with* $P(X_\infty \in \partial A) = 0$, $\lim_{n\to\infty} P(X_n \in A) = P(X_\infty \in A)$

vague convergence From the test function viewpoint, $C_0(X)$ instead of $C_b(X)$.

Remark 3.2.9. Vague convergence has a nice functional analytical interpretation. Let X be a locally compact space. By the Riesz representation theorem, the space $M(X)$ of Radon measures

The constant function is in C_b but not C_0, this is why vague convergence

Theorem 3.2.10 (Helly's selection theorem)**.** *For every sequence* F_n *of distribution functions, there is a subsequence* $F_{n(k)}$ *and a right continuous nondecreasing function* F *s.t.* $F_{n(k)} \Longrightarrow_v F$.

Remark 3.2.11. The limit may not be a distribution function. This type of convergence is called vague convergence.

Proof. To construct the function F, we adopt the standard diagonal argument. Let $\{q_i\}$ be an enumeration of the rationals. Since $F_m(q_k) \in [0,1]$ is bounded for all m, there is a subsubsequence $m_k(i)$ that is a subsequence of $m_{k-1}(i)$ s.t. $F_{m_k(i)}(q_k) \longrightarrow G(q_k)$. Select the diagonal sequence $n(k) = m_k(k)$, then by construction, $F_{n(k)}(q) \longrightarrow G(q)$ for all rational q.

Now we need to consruct F from G. Let

$$F(x) = \inf\{G(q) : q \in \mathbb{Q}, q > x\}$$

then $F(x)$ is right continuous and nondecreasing.

Let x be a continuity point of F. Pick rational $s > x$ s.t. $F(x) \leq F(s) < F(x) + \epsilon$, then as $F_{n(k)}(s) \longrightarrow G(s) \leq F(s)$, for k sufficient large, we have $F_{n(k)}(x) \leq F_{n(k)}(s) < F(x) + \epsilon$. On the other hand, pick rational $r_1 < r_2 < x$ s.t. $F(x) - \epsilon < F(r_1) \leq F(r_2) \leq F(x)$, then as $F_{n(k)}(r_2) \longrightarrow G(r_2) \geq F(r_1)$, so $F_{n(k)}(x) \geq F_{n(k)}(r_2) > F(x) - \epsilon$ for k sufficient large. Thus as $\epsilon \to 0$, we have the weak convergence. □

Theorem 3.2.12. *Every subsequential limit is the distribution function of a probability measure $\iff$ the sequence is* ***tight****, i.e. $\forall \epsilon > 0, \exists M_\epsilon$ s.t.*

$$\limsup_{n\to\infty} 1 - F_n(M_\epsilon) + F_n(-M_\epsilon) \le \epsilon$$

Proof. First note that for vague convergence $0 \le F(x) \le 1$.

$\Longleftarrow$: Suppose the sequence is tight and $F_{n(k)} \Longrightarrow_v F$. Let $r < -M_\epsilon, s > M_\epsilon$ be continuity points of F, then $1-F(s)+F(r) = \lim_{k\to\infty} 1-F_{n(k)}(s)+F_{n(k)}(r) \le \limsup_{n\to\infty} 1-F_n(M_\epsilon)+F_n(M_\epsilon) \le \epsilon$. Letting $r \to -\infty$ and $s \to \infty$ gives $\limsup_{n\to\infty} 1 - F(x) + F(-x) \le \epsilon$.

$\Longrightarrow$: Suppose F_n is not tight. Then there is an $\epsilon > 0$ and a subsequence $n(k) \to \infty$ s.t.

$$1 - F_{n(k)}(k) + F_{n(k)}(-k) \ge \epsilon$$

for all k. By passing to a further subsequence $F_{n(k_j)}$ we can suppose $F_{n(k_j)} \Longrightarrow_v F$. Let $r < 0 < s$ be continuity points of F. Then $1-F(s)+F(r) = \lim_{j\to\infty} 1-F_{n(k_j)}(s)+F_{n(k_j)}(r) \ge \liminf_{j\to\infty} 1-F_{n(k_j)}(k_j) + F_{n(k_j)}(-k_j) \ge \epsilon$. Letting $s \to \infty$ and $r \to -\infty$, we see that F is not the distribution function of a probability measure. □

Corollary 3.2.13. *If there is a $\varphi \ge 0$ s.t. $\varphi(x) \to \infty$ as $|x| \to \infty$ and*

$$\sup_n \int \varphi(x) \mathrm{d}F_n(x) = C < \infty$$

then F_n is tight.

Proof. $C \ge \int \varphi(x)\mathrm{d}F_n(x) \ge \inf_{|x|\ge M} \varphi(x)(F_n(-M) + 1 - F_n(M))$ □

Lemma 3.2.14. *If $X_n \longrightarrow X$ in probability, then $X_n \Longrightarrow X$. Conversely, if $X_n \Longrightarrow c$ where c is a constant, then $X_n \longrightarrow c$ in probability.*

Theorem 3.2.15 (slutsky)**.** *If $X_n \Longrightarrow X$ and $Y_n \Longrightarrow c$, where c is a constant, then:*
(i) $X_n + Y_n \Longrightarrow X + c$
(ii) $X_nY_n \Longrightarrow cX$

3.3 Characteristic Functions

Definition 3.3.1 (characteristic function)**.** If X is a random variable, we define its characteristic function by $\varphi(t) = Ee^{itX}$.

Theorem 3.3.2 (properties of ch.f.)**.** *All ch.f.s have the following properties:*
(i) $\varphi(0) = 1$
(ii) $\varphi(-t) = \overline{\varphi(t)}$
(iii) $|\varphi(t)| \le 1$
(iv) $\varphi(t)$ is uniformly continuous on $(-\infty, \infty)$
(v) $Ee^{it(aX+b)} = e^{itb}\varphi(at)$

Theorem 3.3.3. *If X_1 and X_2 are independent and have ch.f.'s φ_1 and φ_2, then $X_1 + X_2$ has ch.f. $\varphi_1(t)\varphi_2(t)$.*

Lemma 3.3.4. *If $F_1, \cdots, F_n$ have ch.f. $\varphi_1, \cdots, \varphi_n$ and $\lambda_i \ge 0$ have $\lambda_1 + \cdots + \lambda_n = 1$, then $\sum_{i=1}^n \lambda_i F_i$ has ch.f. $\sum_{i=1}^n \lambda_i \varphi_i$.*

Theorem 3.3.5 (Continuity theorem)**.** *Let μ_n, $1 \le n \le \infty$ be probability measures with ch.f. φ_n.*
(i) If $\mu_n \Longrightarrow \mu_\infty$, then $\varphi_n(t) \to \varphi_\infty(t)$ for all t.
(ii) If $\varphi_n(t)$ converges pointwise to a limit $\varphi(t)$ that is continuous at 0, then the associated sequence of distributions μ_n is tight and converges weakly to the measure μ with characteristic function φ.

The next result is useful for constructing examples of ch.f.'s.

Example 3.3.6 (Polya's distribution)**.**

$$\begin{aligned} \text{Density} \quad & \frac{1-\cos(x)}{\pi x^2} \\ \text{Ch.f.} \quad & (1-|t|)^+ \end{aligned}$$

Theorem 3.3.7 (Polya's criterion)**.** *Let $\varphi(t)$ be real nonnegative and have $\varphi(0)=1$, $\varphi(t)=\varphi(-t)$, and φ is decreasing and convex on $(0,\infty)$ with $\lim_{t\downarrow 0}\varphi(t)=1, \lim_{t\uparrow\infty}\varphi(t)=0$. Then there is a probability measure ν on $(0,\infty)$, so that*

$$\varphi(t)=\int_0^\infty (1-\left|\frac{t}{s}\right|^+)\nu(\mathrm{d}s)$$

and hence φ is a characteristic function.

3.4 The Moment Problem

Example 3.4.1 (Heyde(1963))**.** Consider the lognormal density

$$f_0(x)=\frac{1}{\sqrt{(2\pi)}}\frac{1}{x}\exp^{-\frac{(\log x)^2}{2}}1_{x\geq 0}$$

and for $-1\leq a\leq 1$ let

$$f_a(x)=f_0(x)(1+a\sin(2\pi\log x))$$

We claim that f_a is a density and has the same moment as f_0

Example 3.4.2.

A usual sufficient condition for a distribution to be determined by its moments is:

Theorem 3.4.3. *If $\limsup_{n\to\infty}\frac{\mu_{2n}^{\frac{1}{2n}}}{2n}=r<\infty$, then there is at most one d.f. F with $\mu_n=\int x^n\mathrm{d}F(x)$ for all positive integers n.*

Proof. First we explain why the condition only consider $2n$. Let F be any d.f. with the moment μ_n and let $\nu_n=\int|x|^n\,\mathrm{d}F(x)$. The Cauchy-Schwarz inequality implies $\nu_{2n+1}^2\leq\mu_{2n}\mu_{2n+2}$, so

$$\limsup_{n\to\infty}\frac{\nu_n^{\frac{1}{n}}}{n}=r<\infty$$

Next, we have

$$\left|e^{i\theta X}(e^{itX}-\sum_{m=0}^{n-1}\frac{(itX)^m}{m!})\right|\leq\frac{|tX|^n}{n!}$$

Taking expected value, we have

$$\left|\varphi(\theta+t)-\varphi(\theta)-t\varphi'(\theta)-\cdots-\frac{t^{n-1}}{(n-1)!}\varphi^{(n-1)}(\theta)\right|\leq\frac{|t|^n}{n!}\nu_n$$

So we see that for any θ,

$$\varphi(\theta+t)=\varphi(\theta)+\sum_{m=1}^{\infty}\frac{t^m}{m!}\varphi^{(m)}(\theta)\quad\forall\,|t|<\frac{1}{er}$$

Let G be another distribution with the given moments and ψ its ch.f.. Since $\psi(0)=\varphi(0)=1$, it follows from the above equation and induction that $\psi(t)=\varphi(t)$ for $|t|\leq\frac{k}{3r}$ for all k, so the two ch.f. coincide and the distributions are equal. □

Here is an application.

Theorem 3.4.4 (Semi-Circle Law)**.**

3.5 Central Limit Theorems

Theorem 3.5.1. *Let $X_1,X_2,\cdots$ be i.i.d. with $EX_i=\mu$, $Var(X_i)=\sigma^2\in(0,\infty)$. If $S_n=X_1+\cdots+X_n$, then*

$$\frac{S_n-n\mu}{\sigma\sqrt{n}}\Longrightarrow\mathcal{N}(0,1)$$

Proof. WLOG suppose $\mu = 0$. $\varphi(t) = Ee^{itX_1} = 1 - \frac{\sigma^2 t^2}{2} + o(t^2)$, so $Ee^{itS_n/\sigma n^{\frac{1}{2}}} = (1 - \frac{t^2}{2n} + o(\frac{1}{n}))^n$. The last quantity $\to e^{-\frac{t^2}{2}}$ as $n \to \infty$, and the conclusion follows from the continuity theorem. □

Theorem 3.5.2 (Lindeberg-Feller theorem)**.** *For each n, let $X_{n,m}, 1 \le m \le n$ be independent random variables with $EX_{n,m} = 0$. Suppose*
(i) $\sum_{m=1}^n EX_{n,m}^2 \to \sigma^2 > 0$
(ii) $\forall \epsilon > 0$, $\lim_{n\to\infty} \sum_{m=1}^n E(|X_{n,m}|^2; |X_{n,m}| > \epsilon) = 0$
Then $S_n = X_{n,1} + \cdots + X_{n,n} \Longrightarrow \mathcal{N}(0, \sigma^2)$.

Proof. □

3.6 Local Limit Theorems

Local limit theorems are subtly different from central limit theorems. The story is this:

Example 3.6.1.

Definition 3.6.2 (lattice distribution)**.** A random variable has a lattice distribution if there are constant $b, h > 0$ so that $P(X \in b + h\mathbb{Z}) = 1$. The largest h for which the last statement holds is called the span of the distribution.

Theorem 3.6.3. *Let $\varphi(t) = Ee^{itX}$. Regarding to the relationship between $|\varphi(t)|$ and 1, there are only three possiblities.*
(i) $|\varphi(t)| < 1$ for all $t \neq 0$.
(ii) There is a $\lambda > 0$ so that $|\varphi(\lambda)| = 1$ and $|\varphi(t)| < 1$ for $0 < t < \lambda$. In this case, X has a lattice distribution with span $\frac{2\pi}{\lambda}$.
(iii) $|\varphi(t)|$ for all t. In this case, $X = b$ a.s. for some b.

Proof. □

Theorem 3.6.4 (LLT for the lattice case)**.** *Let $X_1, X_2, \cdots$ be i.i.d. with $EX_i = 0, EX_i^2 = \sigma^2 \in (0, \infty)$, and having a common lattice distribution with span h. If $S_n = X_1 + \cdots + X_n$ and $P(X_i \in b + h\mathbb{Z}) = 1$. We put*

$$p_n(x) = P(\frac{S_n}{\sqrt{n}} = x) \text{ for } x \in \mathcal{L}_n = \left\{ \frac{nb + hz}{\sqrt{n}} : z \in \mathbb{Z} \right\}$$

and

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

Then as $n \to \infty$,

$$\sup_{x \in \mathcal{L}_n} \left| \frac{\sqrt{n}}{h} p_n(x) - n(x) \right| \to 0$$

Proof. Recall the inversion formula for lattice r.v. Y with $P(Y \in a + \theta\mathbb{Z}) = 1$ and $\psi(t) = Ee^{itY}$:

$$P(Y = x) = \frac{\theta}{2\pi} \int_{-\frac{\pi}{\theta}}^{\frac{\pi}{\theta}} e^{-itx} \psi(t) \mathrm{d}t$$

Use this formula for $\frac{S_n}{\sqrt{n}}$ gives

$$\frac{\sqrt{n}}{h} p_n(x) = \frac{1}{2\pi} \int_{-\frac{\pi\sqrt{n}}{h}}^{\frac{\pi\sqrt{n}}{h}} e^{-itx} \varphi^n(\frac{t}{\sqrt{n}}) \mathrm{d}t$$

and we have

$$n(x) = \frac{1}{2\pi} \int e^{itx} e^{-\frac{\sigma^2 t^2}{2}}$$

Substracting the last two equations and doing some estimation gives

$$\left| \frac{\sqrt{n}}{h} p_n(x) - n(x) \right| \le \int_{-\frac{\pi\sqrt{n}}{h}}^{\frac{\pi\sqrt{n}}{h}} \left| \varphi^n(\frac{t}{\sqrt{n}}) - e^{-\frac{\sigma^2 t^2}{2}} \right| \mathrm{d}t + \int_{\frac{\pi\sqrt{n}}{h}}^{\infty} e^{-\frac{\sigma^2 t^2}{2}} \mathrm{d}t$$

So we are left to estimate the integrals. □

3.7 Poisson Convergence

Theorem 3.7.1. *For each n let $X_{n,m}$, $1 \le m \le n$ be independent random variables with $P(X_{n,m} = 1) = p_{n,m}, P(X_{n,m} = 0) = 1 - p_{n,m}$. Suppose*
(i) $\sum_{m=1}^{n} p_{n,m} \to \lambda \in (0, \infty)$.
(ii) $\max_{1 \le m \le n} \to 0$.
If $S_n = X_{n,1} + \cdots + X_{n,n}$, then $S_n \Longrightarrow$ Poisson(λ).

Here is a second proof of this theorem which provides new insight.

Definition 3.7.2 (total variation distance)**.** The total variation distance between two measures on a countable set S. $\|\mu - \nu\| = \frac{1}{2}\sum_z |\mu(z) - \nu(z)|$.

Lemma 3.7.3. $\|\mu - \nu\| = \sup_{A \subset S} |\mu(A) - \nu(A)|$

Lemma 3.7.4. *$d(\mu, \nu) = \|\mu - \nu\|$ defines a metric on probability measures on $\mathbb{Z}$. furthermore*

Lemma 3.7.5. *Consider measures on $\mathbb{Z}$. Then $\|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\| \le \|\mu_1 - \nu_1\| + \|\mu_2 - \nu_2\|$.*

Lemma 3.7.6. *Consider measures on $\mathbb{Z}$.Then $\|\mu_1 * \mu_2 - \nu_1 * \nu_2\| \le \|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\|$.*
Here $*$ stands for the convolution.

3.8 Stable Laws

In this section, we will investigate the case $EX_1^2 = \infty$ and give necessary and sufficient conditions for the existence of constants a_n and b_n so that

$$\frac{S_n - b_n}{a_n} \Longrightarrow Y$$

where Y is nondegenerate.

Definition 3.8.1 (slowly varying)**.** L is said to be slowly varying if $\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1\ \forall t > 0$.

Theorem 3.8.2. *Suppose $X_1, X_2, \cdots$ are i.i.d. with a distribution that satisfies:*
(i) $\lim_{x\to\infty} \frac{P(X_1 > x)}{P(|X_1| > x)} = \theta \in [0, 1]$
(ii) $P(|X_1| > x) = x^{-\alpha}L(x)$ where $\alpha < 2$ and L is slowly varying
Let $S_n = X_1 + \cdots + X_n$, $a_n = \inf\left\{x : P(|X_1| > x) \le \frac{1}{n}\right\}$ and $b_n = nE(X_1 1_{(|X_1| \le a_n)})$.

Definition 3.8.3. The distributions whose ch.f are given by the following family with parameters κ, α, b, c are called stable laws.

$$\exp\left(itc - b\,|t|^{\alpha}\,(1 + i\kappa \mathrm{sgn}(t) w_\alpha(t))\right)$$

where $\kappa \in [-1, 1]$, $\alpha \in (0, 2)$,

$$w_\alpha(t) = \begin{cases} \tan(\frac{\pi}{2}\alpha) & \alpha \ne 1 \\ \frac{2}{\pi}\log|t| & \alpha = 1 \end{cases}$$

Theorem 3.8.4. *Y is the limit of $\frac{X_1 + \cdots + X_k - b_k}{a_k}$ for some i.i.d. sequence X_i if and only if Y has a stable law.*

3.9 Infinitely Divisible Distributions

Definition 3.9.1. Z has an infinitely divisible distribution if for each n there is an i.i.d. sequence $Y_{n,1}, \cdots, Y_{n,n}$ so that $Z =_d Y_{n,1} + \cdots + Y_{n,n}$.

Theorem 3.9.2. *Z is a limit of sums of type $Z = X_{n,1} + \cdots + X_{n,n}$ if and only if Z has an infinitely divisible distribution.*

Proof. □

Theorem 3.9.3 (Levy-Khinchin Theorem)**.** *Z has an infinitely divisible distribution if and only if its characteristic function has*

$$\log\varphi(t) = ict - \frac{\sigma^2 t^2}{2} + \int (e^{itx} - 1 - \frac{itx}{1 + x^2})\mu(\mathrm{d}x)$$

where μ is a measure with $\mu(\{0\}) = 0$ and $\int \frac{x^2}{1+x^2}\mu(\mathrm{d}x) < \infty$.

The theory of infinitely divisible distributions is simpler in the case of finite variance. In this case, we have:

Theorem 3.9.4 (Kolmogorov's Theorem)**.** *Z has an infinitely divisible distribution with mean 0 and finite variance if and only if its ch.f. has the form*

$$\log \varphi(t) = \int \frac{(e^{itx} - 1 - itx)}{x^2} \nu(\mathrm{d}x)$$

ν is called the canonical measure, and $Var(Z) = \nu(\mathbb{R})$.

3.10 Limit Theorems in $\mathbb{R}^d$

Theorem 3.10.1 (Convergence theorem)**.** *Let $X_n, 1 \leq n \leq \infty$ be random vectors with ch.f. φ_n. A necessary and sufficient condition for $X_n \Longrightarrow X_\infty$ is that $\varphi_n(t) \to \varphi_\infty(t)$.*

Theorem 3.10.2 (Cramer-Wold Device)**.** *A sufficient condition for $X_n \Longrightarrow X_\infty$ is that $\theta \cdot X_n \Longrightarrow \theta \cdot X_\infty$ for all $\theta \in \mathbb{R}^d$.*

3.11 Stein's method

There is a lack of calculus in probability theory. We have only used Fourier transform, i.e. the characteristic function. Stein invented a exotic way of using calculus to derive the convergence rate of CLT.

Definition 3.11.1.

Stein's method is related to Slepian's interpolation, which is in turn related to Lindeberg's telescopic interpolation.

Chapter 4

Martingales

4.1 Conditional Expectation

Definition 4.1.1 (conditional expectation)**.** Given a probability space $(\Omega, \mathcal{F}_o, \mathbb{P})$, a σ-field $\mathcal{F} \subset \mathcal{F}_o$, and a random varaible $X \in \mathcal{F}_o$ with $\mathbb{E}|X| < \infty$. The conditional expectation of X given $\mathcal{F}$ is any random variable Y that satisfies:
(i) $Y \in \mathcal{F}$
(ii) $\forall A \in \mathcal{F}, \int_A X \mathrm{d}\mathbb{P} = \int_A \mathrm{d}\mathbb{P}$.

Lemma 4.1.2. *If Y satisfies (i) and (ii), then it is integrable.*

Proof. Let $A = \{Y > 0\} \in \mathcal{F}$. We have $\int_A Y \mathrm{d}\mathbb{P} = \int_A X \mathrm{d}\mathbb{P} \leq \int_A |X| \mathrm{d}\mathbb{P}$ and $\int_{A^c} -Y \mathrm{d}\mathbb{P} = \int_{A^c} -X \mathrm{d}\mathbb{P} \leq \int_{A^c} |X| \mathrm{d}\mathbb{P}$, therefore we have $\mathbb{E}|Y| \leq |X|$. □

Theorem 4.1.3 (uniqueness of conditional expectation)**.** *The conditional expecation of X given $\mathcal{F}$ is unique, denoted by $\mathbb{E}(X|\mathcal{F})$.*

Proof. Suppose Y' also satisfies (i)&(ii). Taking $A = \{Y - Y' \geq \epsilon > 0\}$, we see $0 = \int_A X - X \mathrm{d}\mathbb{P} = \int_A Y - Y' \mathrm{d}\mathbb{P} \geq \epsilon \mathbb{P}(A)$ so $\mathbb{P}(A) = 0$. Since this holds for all ϵ, we have $Y \leq Y'$ a.s., and switching the role of $Y \& Y'$ gives the desiered result. □

Theorem 4.1.4 (existence of conditional expectation)**.** *$\mathbb{E}(X|\mathcal{F})$ exists.*

Proof. The proof is based on Radon-Nikodym Theorem. Suppose first that $X \geq 0$. Construct a measure $\nu(A) = \int_A X \mathrm{d}\mathbb{P}$ for $A \in \mathcal{F}$. Then $\nu \ll \mathbb{P}$, so by Radon-Nikodym Theorem, there exists $Y \in \mathcal{F}$ satisfying $\nu(A) = \int_A Y \mathrm{d}\mathbb{P}$.

To treat the general case, write $X = X^+ - X^-$, let $Y_1 = \mathbb{E}(X^+|\mathcal{F})$ and $Y_2 = \mathbb{E}(X^-|\mathcal{F})$, then verify condition (i)&(ii). □

Now we investigate the properties of conditional expectation.

Theorem 4.1.5.

Theorem 4.1.6. *If φ is convex and $\mathbb{E}|X|, \mathbb{E}|\varphi(X)| < \infty$, then*

$$\varphi(\mathbb{E}(X|\mathcal{F})) \leq \mathbb{E}(\varphi(X)|\mathcal{F})$$

Proof. □

Corollary 4.1.7. *Conditional expectation is a contraction in L^p, $p \geq 1$.*

Theorem 4.1.8. *If $\mathcal{F}_1 \subset \mathcal{F}_2$, then:*
(i) $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_1)|\mathcal{F}_2) = \mathbb{E}(X|\mathcal{F}_1)$
(ii) $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_2)|\mathcal{F}_1) = \mathbb{E}(X|\mathcal{F}_1)$

Proof. Directly follows from the definition. □

Remark 4.1.9. This theorem shows that whatever the order of conditioning is, the result is always conditioning on the smallest σ-field.

Theorem 4.1.10. *If $X \in \mathcal{F}$ and $\mathbb{E}|Y|, \mathbb{E}|XY| < \infty$, then*

$$\mathbb{E}(XY|\mathcal{F}) = XE(Y|\mathcal{F})$$

Proof. Approximate X by the standard process as in the construction of Lebesgue integral. □

Theorem 4.1.11 (LSE)**.** *Suppose $EX^2 < \infty$. $\mathbb{E}(X|\mathcal{F})$ is the variable $Y \in \mathcal{F}$ that minimizes $\mathbb{E}(X-Y)^2$.*

4.2 Martingales

Definition 4.2.1 (filtration)**.** An increasing sequence of σ-fields is called a filtration.

Definition 4.2.2 (adapted)**.** A sequence X_n is said to be adapted to $\mathcal{F}_n$ if $X_n \in \mathcal{F}_n$ for all n.

Definition 4.2.3 (martingale)**.** If X_n is a sequence with:
(i) $\mathbb{E}\,|X_n| < \infty$
(ii) X_n is adapted to $\mathcal{F}_n$
(iii) $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ for all n
then X is said to be a martingale.

Remark 4.2.4. If in (iii) is replaced by $\leq$ or $\geq$, then X is said to be a supermartingale or submartingale respectively.

Theorem 4.2.5. *If X_n is a supermartingale, then for $n > m$, $\mathbb{E}(X_n|\mathcal{F}_m) \leq X_m$.*
If X_n is a submartingale, then for $n > m$, $\mathbb{E}(X_n|\mathcal{F}_m) \geq X_m$.
If X_n is a martingale, then for $n > m$, $\mathbb{E}(X_n|\mathcal{F}_m) = X_m$.

Proof. By definition and induction. □

Theorem 4.2.6. *If X_n is a supermartingale w.r.t. $\mathcal{F}_n$ and φ is an increasing concave function with $\mathbb{E}\,|\varphi(X_n)| < \infty$ for all n, then $\varphi(X_n)$ is a supermartingale w.r.t $\mathcal{F}_n$.*

If X_n is a submartingale w.r.t. $\mathcal{F}_n$ and φ is an increasing convex function with $\mathbb{E}\,|\varphi(X_n)| < \infty$ for all n, then $\varphi(X_n)$ is a submartingale w.r.t $\mathcal{F}_n$.

If X_n is a martingale w.r.t. $\mathcal{F}_n$ and φ is a convex function with $\mathbb{E}\,|\varphi(X_n)| < \infty$ for all n, then $\varphi(X_n)$ is a submartingale w.r.t $\mathcal{F}_n$.

Proof. Directly follows from the definition of martingale and Jensen's inequality. □

Corollary 4.2.7. *If X_n is a submartingale, then $(X_n - a)^+$ is a submartingale.*

Corollary 4.2.8. *If X_n is a supermartingale, then $X_n \wedge a$ is a supermartingale.*

Definition 4.2.9 (predictable)**.** Let $\mathcal{F}_n$, $n \geq 0$ be a filtration. H_n, $n \geq 1$ is said to be a predictable sequence if $H_n \in \mathcal{F}_{n-1}$ for all $n \geq 1$.

Example 4.2.10.

Theorem 4.2.11. *Let X_n, $n \geq 0$, be a supermartingale. If $H_n \geq 0$ is predictable and each H_n is bounded, then $(H \cdot X)_n = \sum_{m=1}^{n} H_m(X_m - X_{m-1})$ is a supermartingale.*

The same fact is true for submartingales and for martingales, while in the latter case we can relax the restriction $H_n \geq 0$.

Theorem 4.2.12 (Doob's decomposition)**.** *Any submartingale $X_n, n \geq 0$, can be written in a unique way as $X_n = M_n + A_n$, where M_n is a martingale and A_n is a predictable increasing sequence with $A_0 = 0$.*

Proof. We want $X_n = M_n + A_n, \mathbb{E}(M_n|\mathcal{F}_{n-1}) = M_{n-1}$, and $A_n \in \mathcal{F}_{n-1}$. So we must have

$$\begin{aligned}\mathbb{E}(X_n|\mathcal{F}_{n-1}) &= \mathbb{E}(M_n|\mathcal{F}_{n-1}) + \mathbb{E}(A_n|\mathcal{F}_{n-1})\\ &= M_{n-1} + A_n\\ &= X_{n-1} - A_{n-1} + A_n\end{aligned}$$

So $A_n - A_{n-1} = \mathbb{E}(X_n|\mathcal{F}_{n-1}) - X_{n-1}$. Since $A_0 = 0$, we have

$$A_n = \sum_{m=1}^{n} E(X_n - X_{n-1}|\mathcal{F}_{n-1})$$

The last step is to check what we have constructed above indeed satisfies the desired properties. □

4.3 Stopping Times

Definition 4.3.1 (stopping time)**.** A random variable N is said to be a stopping time if $\{N = n\} \in \mathcal{F}_n$ for all n.

Corollary 4.3.2. *If N is a stopping time and X_n is a supermartingale, then $X_{N \wedge n}$ is a supermartingale.*

Proof. Let $H_n = 1_{N \geq n}$. Verify that H_n is predictable. It follows from the theorem that $(H \cdot X)_n = X_{N\wedge n} - X_0$ is a supermartingale. Thus $X_{N\wedge n}$ is a supermartingale as a sum of two supermartingale. □

4.4 Almost Sure Convergence

Suppose X_n, $n \geq 0$, is a submartingale. Let $a < b$ and $N_0 = -1$, and for $k \geq 1$ let

$$N_l = \begin{cases} \inf\{m > N_{2k-2} : X_m \leq a\}, & l = 2k-1 \\ \inf\{m > N_{2k-1} : X_m \geq b\}, & l = 2k \end{cases}$$

The N_j are stopping times and $\{N_{2k-1} < m \leq N_{2k}\} = \{N_{2k-1} \leq m-1\} \cap \{N_{2k} \leq m-1\}^c \in \mathcal{F}_{m-1}$, so

$$H_m = \begin{cases} 1 & \text{if } N_{2k-1} < m \leq N_{2k} \text{ for some } k \\ 0 & \text{otehrwise} \end{cases}$$

defines a predictable sequence.

Note that $X_{N_{2k-1}} \leq a$ and $X_{N_{2k}} \geq b$. We can regard H_m as a gambling system taking advantage of these upcrossings. In stock market terms, we buy when $X_m \leq a$ and sell when $X_m \geq b$, so every time an upcrossing is completed, we make a profit of $\geq (b-a)$.

Finally, let $U_n = \sup\{k : N_{2k} \leq n\}$ be the number of upcrossings completed by time n.

Theorem 4.4.1 (upcrossing inequality)**.** *If X_m, $m \geq 0$, is a submartingale, then*

$$(b-a)EU_n \leq \mathbb{E}(X_n - a)^+ - \mathbb{E}(X_0 - a)^+$$

Proof. Let we introduce $Y_n = a + (X_n - a)^+$ to fix the final incomplete upcrossing, then Y_n is a submartingale that upcrosses $[a, b]$ the same number of times that X_m does. Each upcross results in a profit $\geq (b-a)$ and a final incomplete upcrossing of Y_n (instead of X_n) results in a nonnegative profit, therefore we have $(b-a)U_n \leq (H \cdot Y)_n$.

Let $K_m = 1 - H_m$, then $Y_n - Y_0 = (H \cdot Y)_n + (K \cdot Y)_n$. $(K \cdot Y)_n$ is a submartingale as well, so $\mathbb{E}(K \cdot Y)_n \geq \mathbb{E}(K \cdot Y)_0 = 0$. Therefore $\mathbb{E}(H \cdot Y)_n \leq \mathbb{E}(Y_n - Y_0)$. □

Whether the number of upcrossing is finite or infinite characterize convergence.

Theorem 4.4.2 (martingale convergence theorem)**.** *If X_n is a submartingale with $\sup EX_n^+ < \infty$, then as $n \to \infty$, X_n converges a.s. to a limit X with $\mathbb{E}|X| < \infty$.*

Proof. Since $(X-a)^+ \leq X^+ + |a|$, upcrossing inequality implies that

$$EU_n \leq \frac{EX_n^+ + |a|}{b-a}$$

As $n \uparrow \infty$, $U_n \uparrow U$, where U is the number of upcrossings of $[a, b]$ by the whole sequence, so if $\sup EX_n^+ < \infty$, then $EU < \infty$ and hence $U \leq \infty$ a.s..

Since the last conclusion holds for all rational a and b,

$$\mathbb{P}\Big(\bigcup_{a,b \in \mathbb{Q}} \{\liminf X_n < a < b < \limsup X_n\}\Big) = 0$$

and hence $\lim X_n$ exists a.s..

Fatou's lemma guarantees $EX^+ \leq \liminf EX_n^+ < \infty$. For EX^-, we observe that $EX_n^- = EX_n^+ - EX_n \leq EX_n^+ - EX_0$ since X_n is a submartingale, so another application of Fatou's lemma shows $EX^- \leq \liminf EX_n^- \leq \sup EX_n^+ - EX_0$ and completes the proof. □

4.5 Convergence in L^p

Lemma 4.5.1 (bounded optional stopping)**.** *If X_n is a submartingale and N is a stopping time with $\mathbb{P}(N \le k) = 1$, then*

$$EX_0 \le EX_N \le EX_k$$

Proof. $X_{N\wedge n}$ is a submartingale, so $EX_0 = EX_{N\wedge n} \le EX_{N\wedge k} = EX_N$.

To prove the other inequality, let $K_n = 1_{\{N<n\}} = 1_{\{N\le n-1\}}$. K_n is predictable, so $(K\cdot X)_n = X_n - X_{n\wedge N}$ is a submartingale, and it follows that $EX_k - EX_N = \mathbb{E}(K\cdot X)_k \ge \mathbb{E}(K\cdot X)_0 = 0$. □

Lemma 4.5.2. *If X_n is a submartingale and $M \le N$ are stopping times with $\mathbb{P}(N \le k) = 1$, then $EX_M \le EX_N$.*

Proof. Let $K_n = 1_{\{M<n\le N\}}$ and modify the above proof. □

Lemma 4.5.3 (bounded Doob's stopping)**.** *If X_n is a submartingale and $M \le N$ are stopping times with $\mathbb{P}(N \le k) = 1$, then $X_M \le \mathbb{E}(X_N|\mathcal{F}_M)$.*

Proof. Let $A \in \mathcal{F}_M$. Define a random time $L = M1_A + N1_{A^c}$. Actually, this is a stopping time. So $EX_M \le EX_L \le EX_N$ by the above lemma. Thus $\mathbb{E}(X_M 1_A) \le \mathbb{E}(X_N 1_A)$ and the result follows. □

Corollary 4.5.4. *An adapted and integrable process X_t is a martingale if and only if*

$$\mathbb{E}(X_M) = \mathbb{E}(X_N)$$

for every such pair of stopping times.

Proof. Let $A \in \mathcal{F}_{n-1}$ and $L = (n-1)1_A + n1_{A^c}$. By $EX_L = EX_n$, we have $EX_n 1_A = EX_{n-1}1_A$, so $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = X_{n-1}$. □

Theorem 4.5.5 (Doob's inequality)**.** *Let X_m be a submartingale, $\lambda > 0$, and $A = \{\max_{0\le m\le n} X_m^+ \ge \lambda\}$, then*

$$\lambda\mathbb{P}(A) \le EX_n 1_A \le EX_n^+$$

Proof. Let $N = \inf\{m : X_m \ge \lambda\} \wedge n$, then $X_N \ge \lambda$ on A. Therefore $\lambda\mathbb{P}(A) \le EX_N 1_A \le EX_n 1_A$, where the second inequality follows from the lemma above.

The other inequality is obvious. □

Example 4.5.6 (Kolmogorov's maximal inequality)**.** If we let $S_n = \xi_1 + \cdots + \xi_n$, where the ξ_m is independent and have $\mathbb{E}\xi_m = 0$, $\sigma_m^2 = \mathbb{E}\xi_m^2 < \infty$. S_n is a martingale, so S_n^2 is a submartingale. If we let $\lambda = x^2$ and apply Doob's inequality, we get Kolmogorov's maximal inequality:

$$\mathbb{P}(\max_{1\le m\le n} |S_m| \ge x) \le x^{-2}\text{var}(S_n)$$

Theorem 4.5.7 (L^p maximum inequality)**.** *If X_n is a submartingale, and $\bar{X}_n = \max_{0\le m\le n} X_m^+$, then for $1 < p < \infty$,*

$$\mathbb{E}(\bar{X}_n^p) \le (\frac{p}{p-1})^p E(X_n^+)^p$$

Proof. The ingredients are Doob's inequality and Hölder's inequality. To avoid dividing infinity, we will work with $\bar{X}_n \wedge M$ rather than $\bar{X}_n$. This does not change the application of Doob's inequality.

$$\begin{aligned}
\mathbb{E}((\bar{X}_n \wedge M)^p) &= \int_0^\infty p\lambda^{p-1}\mathbb{P}(\bar{X}_n \wedge M \ge \lambda)\mathrm{d}\lambda \\
&\le \int_0^\infty p\lambda^{p-1}(\lambda^{-1}\int X_n^+ 1_{\{\bar{X}_n\wedge M\ge\lambda\}}\mathrm{d}\mathbb{P})\mathrm{d}\lambda \\
&= \int X_n^+ \int_0^{\bar{X}_n\wedge M} p\lambda^{p-2}\mathrm{d}\lambda\mathrm{d}\mathbb{P} \\
&= \frac{p}{p-1}\int X_n^+(\bar{X}_n \wedge M)^{p-1}\mathrm{d}\mathbb{P}
\end{aligned}$$

If we let $q = \frac{p}{p-1}$ be the conjugate to p and apply Hölder's inequality, we see that

$$\le (\frac{p}{p-1})(\mathbb{E}\left|X_n^+\right|^p)^{1/p}(\mathbb{E}\left|\bar{X}_n \wedge M\right|^p)^{1/q}$$

If we divide both sides of the last inequality by $(\mathbb{E}\left|\bar{X}_n \wedge M\right|^p)^{1/q}$, which is finite thanks to $\wedge M$, then take the pth power of each side, and letting $M \to \infty$ and using the monotone convergence theorem gives the desired result. □

Theorem 4.5.8 (L^p convergence theorem). *If X_n is a martingale with $\sup \mathbb{E}\left|X_n\right|^p < \infty$ where $p > 1$, then $X_n \to X$ a.s. and in L^p.*

Proof. $(EX_n^+)^p \leq (\mathbb{E}\left|X_n\right|)^p \leq \mathbb{E}\left|X_n\right|^p$, so it follows from the martingale convergence theorem that $X_n \to X$ a.s..

Applying L^p maximum inequality to $|X_n|$ implies

$$\mathbb{E}(\sup_{0\leq m\leq n} |X_n|)^p \leq (\frac{p}{p-1})^p E(|X_n|)^p$$

Letting $n \to \infty$ and using the monotone convergence theorem implies $\sup|X_n| \in L^p$. Since $|X_n - X|^p \leq (2\sup|X_n|)^p$, it follows from the dominated convergence theorem that $\mathbb{E}\left|X_n - X\right|^p \to 0$. □

4.6 Square Integrable Martingales

In this section, we will suppose

$$X_n \text{ is a martingale with } X_0 = 0 \text{ and } EX_n^2 < \infty \text{ for all } n$$

Thus, X_n^2 is a submartingale. It follows from Doob's decomposition that we can write $X_n^2 = M_n + A_n$, where M_n is a martingale, and

$$A_n = \sum_{m=1}^{n} E(X_n^2 - X_{n-1}^2|\mathcal{F}_{n-1}) = \sum_{m=1}^{n} E((X_n - X_{n-1})^2|\mathcal{F}_{n-1})$$

A_n is called the increasing process associated with X_n.

Theorem 4.6.1. *$\lim_{n\to\infty} X_n$ exists and is finite a.s. on $\{A_\infty < \infty\}$.*

Lemma 4.6.2. $\mathbb{E}(\sup_m \left|X_m\right|^2) \leq 4EA_\infty$

Proof. By L^2 maximum inequality, $\mathbb{E}(sup_{0\leq m\leq n} \left|X_m\right|^2) \leq 4EX_n^2 = 4EA_n + 4EM_n = 4EA_n + 4EM_0 = 4EA_n + 4EX_0^2 = 4EA_n$. Using the monotone convergence theorem now gives the desired result. □

Proof. Let $a > 0$. Since $A_{n+1} \in \mathcal{F}_n$, $N = \inf\left\{n : A_{n+1} > a^2\right\}$ is a stopping time. Applying the lemma to $X_{N\wedge n}$ and noticing $A_{N\wedge n} \geq a^2$ gives $\mathbb{E}(\sup_n \left|X_{N\wedge n}\right|^2) \leq 4a^2$, so the L^2 convergence theorem implies that $\lim X_{N\wedge n}$ exists and is finite a.s.. Since a is arbitary, the desired result follows. □

Theorem 4.6.3. *Let $f \geq 1$ be increasing with $\int_0^\infty f(t)^{-2}\mathrm{d}t < \infty$. Then $\frac{X_n}{f(A_n)} \to 0$ a.s. on $\{A_\infty = \infty\}$.*

4.7 Convergence in L^1

Now we seek the necessary and sufficient conditions for a martingale to converge in L^1. This leads to the definition of uniformly integrability.

Definition 4.7.1 (uniformly integrablity). A collection of random variables $X_i, i \in I$ is said to be uniformly integrable if

$$\lim_{M\to\infty} (\sup_{i\in I} \mathbb{E}(|X_i|\,; |X_i| > M)) = 0$$

Example 4.7.2. A collection of random variables that are dominated by an integrable random variable is uniformly integrable.

Example 4.7.3. A collection of bounded random variables is uniformly integrable.

Below we give an interesting example of a uniformly integrable family.

Theorem 4.7.4. *Given a probability space $(\Omega, \mathcal{F}_o . \mathbb{P})$ and an $X \in L^1$, then $\{\mathbb{E}(X|\mathcal{F}) : \mathcal{F}$ is a σ-field $\subset \mathcal{F}_o\}$ is uniformly integrable.*

A common way to check uniform integrability is to use:

Lemma 4.7.5. *Let $\varphi \geq 0$ be any function with $\frac{\varphi(x)}{x} \to \infty$ as $x \to \infty$. If $\mathbb{E}\varphi(|X_i|) \leq C$ for all $i \in I$, then $X_i, i \in I$ is uniformly integrable.*

Proof. Write $\mathbb{E}(|X_i|\,; |X_i| > M) = \mathbb{E}(\frac{|X_i|}{\varphi(|X_i|)}\varphi(|X_i|); |X_i| > M)$ and notice that $\frac{|X_i|}{\varphi(|X_i|)} \to 0$. □

Lemma 4.7.6. *If integrable random variables $X_n \to X$ in L^1, then $\mathbb{E}(X_n; A) \to \mathbb{E}(X; A)$.*

Proof. The difference is smaller than $\mathbb{E}\,|X_n - X|$. □

Lemma 4.7.7. *If a martingale $X_n \to X$ in L^1, then $X_n = \mathbb{E}(X|\mathcal{F}_n)$.*

Proof. $\mathbb{E}(X_m|\mathcal{F}_n)$ for $m > n$, so if $A \in \mathcal{F}_n$, $\mathbb{E}(X_m; A) = \mathbb{E}(X_n; A)$. By the lemma above, we have $\mathbb{E}(X_n; A) = \mathbb{E}(X; A)$ for all $A \in \mathcal{F}_n$. By the definition of condition expectation, $X_n = \mathbb{E}(X|\mathcal{F}_n)$. □

Theorem 4.7.8. *For a martingale, TFAE:*
(i) It is uniformly integrable.
(ii) It converges a.s. and in L^1.
(iii) It converges in L^1.
(iv) There is an integrable random variable X s.t. $X_n = \mathbb{E}(X|\mathcal{F}_n)$.

4.8 Backwards Martingales

Definition 4.8.1 (backwards martingale)**.** A backwards martingale is a martingale indexed by the negative integers, i.e., $X_n, n \leq 0$, adapted to an increasing sequence of σ-fields $\mathcal{F}_n$ with $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ for $n \leq -1$.

Theorem 4.8.2. *$X_{-\infty} = \lim_{n \to -\infty} X_n$ exists a.s. and in L^1.*

Proof. Let U_n be the number of upcrossings of $[a, b]$ by $X_{-n}, \cdots, X_0$. The upcrossing inequality implies $(b-a)EU_n \leq \mathbb{E}(X_0 - a)^+$. Letting $n \to \infty$ and using the monotone convergence theorem, we have $EU_\infty < \infty$, so $X_{-\infty} = \lim_{n \to -\infty} X_n$ exists a.s..
The martingale property implies $X_n = \mathbb{E}(X_0|\mathcal{F}_n)$, so X_n is uniformly integrable and the convergence occurs in L_1.
□

Now we identify the limit.

Theorem 4.8.3. *If $X_{-\infty} = \lim_{n \to -\infty} X_n$ and $\mathcal{F}_{-\infty} = \bigcap_n \mathcal{F}_n$, then $X_{-\infty} = \mathbb{E}(X_0|\mathcal{F}_{-\infty})$.*

Proof. Clearly $X_{-\infty} \in \mathcal{F}_{-\infty}$. $X_n = \mathbb{E}(X_0|\mathcal{F}_n)$, so if $A \in \mathcal{F}_{-\infty} \subset \mathcal{F}_n$, then $\int_A X_n \mathrm{d}\mathbb{P} = \int_A X_0 \mathrm{d}\mathbb{P}$, so $\int_A X_{-\infty} \mathrm{d}\mathbb{P} = \int_A X_0 \mathrm{d}\mathbb{P}$ for all $A \in \mathcal{F}_{-\infty}$, proving the desired conclusion. □

Theorem 4.8.4. *If $\mathcal{F}_n \downarrow \mathcal{F}_{-\infty}$ as $n \downarrow -\infty$, then $\mathbb{E}(Y|\mathcal{F}_n) \to \mathbb{E}(Y|\mathcal{F}_{-\infty})$ a.s. and in L^1.*

Proof. $X_n = \mathbb{E}(Y|\mathcal{F}_n)$ is a backwards martingale, so $X_n \to X_{-\infty}$ a.s. and in L^1, where $X_{-\infty} = \mathbb{E}(X_0|\mathcal{F}_{-\infty}) = \mathbb{E}(Y|\mathcal{F}_{-\infty})$. □

4.9 Optional Stopping Theorems

Recall that in Lemma 4.5.1, we have already established optional stopping theorem for bounded stopping times, so in this section we mainly focus on unbounded stopping time.

Theorem 4.9.1. *If X_n is a uniformly integrable submartingale, them for any stopping time N, $X_{n \wedge N}$ is uniformly integrable.*

Proof. As X_n^+ is a submartingale, so by bounded optional stopping $EX_{N\wedge n}^+ \leq EX_n^+$. Since X_n^+ is uniformly integrable, $\sup_n X_{N\wedge n}^+ \leq \sup_n EX_n^+ < \infty$. So by martingale convergence, $X_{N\wedge n} \to X_N$ a.s. and $\mathbb{E}|X_N| < \infty$. Now

$$\mathbb{E}(|X_{N\wedge n}|; |X_{N\wedge n}| > K) = \mathbb{E}(|X_{N\wedge n}|; |X_{N\wedge n}| > K, N \leq n) + \mathbb{E}(|X_{N\wedge n}|; |X_{N\wedge n}| > K, N > n)$$

Since $\mathbb{E}|X_N| < \infty$ and X_n is uniformly integrable, if K is large, then each term is controlled.
□

Remark 4.9.2. From the last computation above, we actually get that if $\mathbb{E}|X_N| < \infty$ and $X_n 1_{\{N>n\}}$ is uniformly integrable, then $X_{N\wedge n}$ is uniformly integrable. And this is the requirement of the optional sampling theorem we usually see.

Theorem 4.9.3. *If X_n is a uniformly integrable submartingale, then for any stopping time $N \le \infty$, we have $EX_0 \le EX_N \le EX_\infty$.*

Proof. By bounded optional stopping, $EX_0 \le EX_{N\wedge n} \le EX_n$. Let $n \to \infty$ and use the $L1$ convergence of uniformly integrable submartingale. □

Theorem 4.9.4. *If X_n is a uniformly integrable submartingale and $L \le M$ are stopping times and $Y_{M\wedge n}$ is uniformly integrable submartingale, then we have $Y_L \le \mathbb{E}(Y_M|\mathcal{F}_L)$.*

For a nonnegative supermartingale, we do not require uniform integrability.

Theorem 4.9.5. *If X_n is a nonnegative supermartingale and $N \le \infty$ is a stopping time, then $EX_0 \ge EX_N$*

Proof. □

Theorem 4.9.6. *Suppose X_n is a submartingale and $\mathbb{E}(|X_{n+1} - X_n||\mathcal{F}_n) \le B$ a.s.. If N is a stopping time with $EN < \infty$, then $X_{N\wedge n}$ is uniformly integrable and hence $EX_N \ge EX_0$.*

Proof. We begin by observing that

$$|X_{N\wedge n}| \le |X_0| + \sum_{m=0}^{\infty} |X_{m+1} - X_m| 1_{(N>m)}$$

To prove uniform integrability, it suffices to show that the right-hand side has finite expectation for then $|X_{N\wedge n}|$ is dominated by an integrable r.v..

$$\mathbb{E}(|X_{m+1} - X_m| 1_{\{N>m\}}) = \mathbb{E}(\mathbb{E}(|X_{m+1} - X_m||\mathcal{F}_m) 1_{\{N>m\}}) \le BP(N > m)$$

and hence the expectation of RHS$\le \mathbb{E}|X_0| + BEN$. □

Now we come to Doob's stopping theorem.

Theorem 4.9.7. *If X_n is a uniformly integrable submartingale, then*

Applications

Example 4.9.8.

Example 4.9.9. Let S_n be symmetric random walk with $S_0 = 0$ and let $T_1 = \min\{n : S_n = 1\}$. Find $\mathbb{P}(T_1 = 2n - 1)$.

Proof. First $\mathbb{P}(T_1 < \infty) = 1$ Use the exponential martingale $X_n = \frac{\exp \theta S_n}{\mathbb{E} \exp \theta S_n}$. $\mathbb{E} \exp \theta S_n = \frac{e^\theta + e^{-\theta}}{2}$. □

Example 4.9.10. Let S_n be a symmetric random walk starting at 0, and let $T = \inf\{n : S_n \notin (-a, a)\}$, where a is an integer. Compute ET^2.

Proof. □

Example 4.9.11. Consider a favorable game in which the payoffs are -1,1,2 with probability $\frac{1}{3}$ each. Compute the probability we ever go broke when we start with $i > 0$.

Lemma 4.9.12. *Let $S_n = \xi_1 + \cdots + \xi_n$ be a random walk. Suppose*

Proof. □

Proof. The original problem is the case where $\theta_0 = \ln(\sqrt{2} - 1)$. So the probability is $(\sqrt{2} - 1)^i$. □

Chapter 5

High Dimensional Probability

The goal of HDP is to quantify the convergence rate of limit theorems such as CLT (Now I think this should be considered as large deviation theory instead). So it is of non-asymptotic nature.

High-Dimensional track the dependence on the dimension, overcome curse of dimensionality. Thus tensorization is of vital importance.

5.1 Concentration with Independence

Theorem 5.1.1 (Hoeffding's inequality)**.** *Let $X_1, \cdots, X_n$ be i.i.d. symmetric Bernoulli random variable. Then*

Definition 5.1.2 (sub-gaussian random variable)**.**

Definition 5.1.3 (sub-exponential random variable)**.**

Lemma 5.1.4 (Hoeffding's lemma)**.** *Assume $a \le X \le b$. Then $\phi(t) \le \frac{t^2(b-a)^2}{8}$.*

Proof. WLOG assume $EX = 0$. Recall that $\phi(t) = \log E(e^{tX})$. Then

$$\phi'(t) = \frac{E(Xe^{tX})}{E(e^{tX})}, \quad \phi''(t) = \frac{E(X^2e^{tX})}{E(e^{tX})} - (\frac{E(Xe^{tX})}{E(e^{tX})})^2$$

Let Q denote the distribution with $\frac{\mathrm{d}Q}{\mathrm{d}P} = \frac{e^{tX}}{Ee^{tX}}$, where P is the distribution of X. Then $\phi''(t) = \mathrm{Var}_Q(X) \le \frac{(b-a)^2}{4}$. □

Lemma 5.1.5 (maximal inequality)**.** *Assume that $X_1, \cdots, X_n$ be n sub-gaussian random variables*

$$E \max_{i\in[n]} X_i \le \sqrt{2\log n}$$

Proof. LogSumExp Trick. □

Remark 5.1.6. The bound is sharp even though we do not assume independence.

Example 5.1.7. Let $X_1, \cdots, X_n$ be independent $N(0,1)$ random variables. Then

5.2 Concentration without Independence

The approach to concentration inequality we developed so far relies crucially on independence of random variables. We now pursue some alternative approaches to concentration, which are not based on independence.

5.3 Subgaussian Concentration

5.3.1 The Entropy Method

First we define the entropy of a nonnegative random variable.

Definition 5.3.1. For a random variable $Z \geq 0$, its entropy is defined as the following:

$$\operatorname{Ent}(Z) := \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z].$$

Remark 5.3.2. When $Z(\omega) = \frac{\mathrm{d}\mu(\omega)}{\mathrm{d}\nu(\omega)}$ and ω with measure ν, this is the KL divergence (or relative entropy). However, note that this definition is different from entropy (or differential entropy) for a probability distribution.

One can immediately see by Jensen's ineqaulity that

$$\operatorname{Ent}(Z) \leq \operatorname{Cov}(Z, \log Z)$$

Lemma 5.3.3 (Herbst). *Suppose that*

$$\operatorname{Ent}[e^{\lambda X}] \leq \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda X}], \quad \forall \lambda \geq 0.$$

Then

$$\phi(\lambda) := \log \mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] \leq \frac{\lambda^2 \sigma^2}{2}, \quad \forall \lambda \geq 0.$$

Proof. The key observation is the following calculation for $\phi(\lambda)$. As $\phi(\lambda) = \log \mathbb{E}[e^{\lambda X}] - \lambda \mathbb{E}X$, we can divide by λ and make the following calculation to get rid of $\mathbb{E}X$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \frac{\phi(\lambda)}{\lambda} = \frac{1}{\lambda^2} \frac{\operatorname{Ent}[e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}$$

□

Here we just derive another equivalent formulation of Gaussian
The key is

Theorem 5.3.4. *If $X_1, \ldots, X_n$ are independent, then*

$$\operatorname{Ent} f(X_1, \ldots, X_n) \leq \mathbb{E} \sum_{i=1}^{n} \operatorname{Ent}_i f(X_1, \ldots, X_n).$$

Proof. □

Analogous to Donsker-Varadhan, we have a variational

5.4 Lipschitz Concentration

Isoperimetirc inequality
transport inequality, Bobkov-Gotze

Theorem 5.4.1 (Bobkov-Gotze). *Let*

Proof. Use variational representation of Wasserstein distance and KL divergence (Gibbs variational principle) □

Using the transportation and the chain rule for KL divergence, we have the following tool for tensorization

Theorem 5.4.2 (Marton). *Let $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ be a convex function, and let $w_i : \mathbb{X}_i \times \mathbb{X}_i \to \mathbb{R}^+$ be positive weight function. Suppose that for $i = 1, \ldots, d$*

But the RHS is not a Wasserstein distance itself. weighted ℓ_1-metric

$$d_c(x, y) := \sum_{i=1}^{d} c_i d_i(x_i, y_i),$$

5.4.1 Talagrand's Concentrartion Inequality

One-Sided Lipschitz property

5.5 Kernel Trick

Theorem 5.5.1 (Grothendick's inequality). *Consider an $m \times m$ matrix (a_{ij}) of real numbers. Assume that for numbers $x_i, y_j \in \{0, 1\}$, we have*

$$\left| \sum_{i,j} a_{ij} x_i y_j \right| \leq 1$$

Then, for any Hilbert space H and any vector $u_i, v_j \in H$ satisfying $\|u_i\| = \|v_j\| = 1$, we have

$$\left| \sum_{i,j} a_{ij} x_i y_j \right| \leq K$$

where $K \leq 1.783$ is an absolute constant.

5.6 Decoupling

In the beginning of HDP, we studied independent random variables of the type $\sum_{i=1}^{n} a_i X_i$. Now we want to study quadratic forms of the type $\sum_{i,j=1}^{n} a_{ij} X_i X_j$ where X_i's are independent. Such a quadratic form is called a chaos in probability theory. It is harder to establish a concentration of a chaos. The main difficulty is that the terms of the sum are not independent. This difficulty can be overcome by the decoupling technique.

The purpose of decoupling is to replace the quadratic form with the bilinear form $\sum_{i,j=1}^{n} a_{ij} X_i X_j'$ where X' is an independent copy of X.

Theorem 5.6.1 (decoupling). *Let A be an $n \times n$ diagonal-free matrix. Let X be a random vector with independent mean zero coordinates X_i. Then for every convex function $F : \mathbb{R} \to \mathbb{R}$,*

$$EF(X^T AX) \leq EF(4X^T AX')$$

where X' is an independent copy of X.

Lemma 5.6.2. *Let Y and Z be independent random variables s.t. $EZ = 0$. Then for every convex function $F : \mathbb{R} \to \mathbb{R}$,*

$$EF(Y) \leq EF(Y + Z)$$

Proof. First condition on Y and use Jensen's inequality. Then take expecation w.r.t Y. □

Remark 5.6.3. Intuitively, this lemma tells us, under some conditions, adding a mean zero disturbance increases the value.

Now we can prove the Hason-Wright inequality.

Theorem 5.6.4. *Let $X = (X_1, \cdots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean zero, sub-gaussian coordinates.Let A be an $n \times n$ matrix. Then, for every $t \geq 0$, we have*

5.7 Symmetrization

In this section we develop the simple and useful technique of symmetrization. It allows one to reduce problems about arbitary distributions to symmetric distributions.

Definition 5.7.1 (Rademacher random variable).

Throuhout this section, we denote by $\xi_1, \xi_2, \cdots$ a sequence of independent Rademacher random variables. We assume that they are independent not only with each other, but also of any other random variable in question.

Lemma 5.7.2 (symmetrization). *Let $X_1, \cdots, X_N$ be independent, mean zero random vectors in a normed space. Then*

$$\frac{1}{2} E \left\| \sum_{i=1}^{N} \xi_i X_i \right\| \leq E \left\| \sum_{i=1}^{N} X_i \right\| \leq 2E \left\| \sum_{i=1}^{N} \xi_i X_i \right\|$$

Proof. The proof relies on introducing an independent copy X_i''s of X_i to symmetrize the expression, and noticing that $X =^d \xi X$ if X is symmetric.

$$\begin{aligned} E\left\|\sum_{i=1}^N X_i\right\| &\le E\left\|\sum_{i=1}^N (X_i - X_i')\right\| \\ &= E\left\|\sum_{i=1}^N \xi_i(X_i - X_i')\right\| \\ &= 2E\left\|\sum_{i=1}^N \xi_i X_i\right\| \end{aligned}$$

$$\begin{aligned} E\left\|\sum_{i=1}^N \xi_i X_i\right\| &\le E\left\|\sum_{i=1}^N \xi_i(X_i - X_i')\right\| \\ &= E\left\|\sum_{i=1}^N (X_i - X_i')\right\| \\ &\le 2E\left\|\sum_{i=1}^N X_i\right\| \end{aligned}$$

□

5.8 Chaining

This section should be moved to stochastic analysis notes because it considers a continuum of random variables. Chaining is a multi-scale version of the ϵ-net argument.

Lemma 5.8.1 (discrete Dudley's inequality). *Let $(X_t)_{t\in T}$ be a mean zero random process on a metric space (T, d) with sub-gaussian increments. Then*

$$E \sup_{t\in T} X_t \le CK \sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}(T, d, 2^{-k})}$$

In the single-scale ϵ-net argument, we discretize T by choosing an ϵ-net $\mathcal{N}$ of T. Then every point $t \in T$ can be approximated by a closest point from the net $\pi(t) \in \mathcal{N}$ with accuracy ϵ. The increment condition yields $\left\|X_t - X_{\pi(t)}\right\|_{\phi_2} \le K\epsilon$. This gives $E\sup_{t\in T} X_t \le E\sup_{t\in T} X_{\pi(t)} + E\sup_{t\in T}(X_t - X_{\pi(t)})$. The first term can be controlled by a union bound over $|\mathcal{N}| = \mathcal{N}(T, d, \epsilon)$ points $\pi(t)$. But for the second term, it is not clear how to control the supremum over $t \in T$.

Chapter 6

The Probabilistic Method

6.1 The Method

6.2 123 Theorem

As a sort of 'inverse' of the probabilistic method, combinatorial techniques can also apply to probabilistic statement.

6.3 The Local Lemma

Definition 6.3.1 (dependency digraph)**.** Let $A_i(i \in [n])$ be n events. A directed graph $D = (V, E)$ on the set of vertices $V = \{1, \cdots, n\}$ is called a dependency digraph for the events if for each i the event is mutually independent of all the events $\{A_j : (i, j) \notin E\}$.

Theorem 6.3.2 (the local lemma)**.** *Let $A_i(i \in [n])$ be n events. Suppose $D = (V, E)$ is a dependency digraph for the above events and suppose there are real numbers $x_1, \cdots, x_n$ s.t. $1 \leq x_i < 0$ and $P(A_i) \leq x_i \prod_{(i,j)\in E}(1 - x_j)$ for all $1 \leq i \leq n$. Then*

$$P(\bigcap_{i=1}^{n} \overline{A_i}) \geq \prod_{i=1}^{n}(1 - x_i)$$

Remark 6.3.3. In particular, with positive probability, no event A_i holds.

Corollary 6.3.4 (the local lemma: symmetric case)**.**

Remark 6.3.5. The constant e is the best possible constant.

6.4 Correlation Inequalities

Theorem 6.4.1 (the four functions theorem)**.**

Theorem 6.4.2 (FKG inequality)**.** *Let L be a finite distributive lattice, and let $\mu : L \to \mathbb{R}^+$ be a log-supermodular function. Then for any two increasing functions f, g, we have*

$$E(fg) \geq (Ef)(Eg)$$

where the expectation is taken w.r.t.

Remark 6.4.3. If both f and g are decreasing, the the result still holds. If one is increasing and one is decreasing, then the inequality is reversed.

Lemma 6.4.4 (Kleitman's lemma)**.** *Let*

Theorem 6.4.5.

Definition 6.4.6 (linear extension)**.**

Theorem 6.4.7 (XYZ theorem)**.** *Let*

Chapter 7

Random Matrices

The goal of this section is to extend previous results for random variables to random matrices.

Part II

Stochastic Process

Chapter 8

Discrete Space Discrete Time Markov Chain

We first study the simpliest stochastic process.

8.1 Basic Theory

We begin by summarizing the concepts mathematicans are most interested in, and introducing their notations.

Definition 8.1.1 (initial distribution)**.**

Definition 8.1.2 (transition matrix)**.**

Remark 8.1.3. Transition matrix provides a way to calculate the probability that after n steps the Markov chain is in a given state.

Next we explain the markov property.

Definition 8.1.4 (markov property)**.**

Definition 8.1.5 (strong markov property)**.**

Next we explain the class structure of markov chain.

Definition 8.1.6 (lead to)**.** We say i leads to j and write $i \to j$ if $P_i(X_n = j \text{ for some } n \geq 0) > 0$.

Definition 8.1.7 (communicate with)**.** We say i communicate with j and write $i \leftrightarrow j$ if both $i \to j$ and $j \to i$.

Theorem 8.1.8 (communicating classes)**.** *Communication is an equivalence relation on I, thus partition I into communicating classes.*

Definition 8.1.9 (closed class)**.** We say a class C is closed if

$$i \in C, i \to j \Longrightarrow j \in C$$

A state i is called absorbing if $\{i\}$ is a closed class.

Definition 8.1.10 (irreducibility)**.** A markov chain with only one class is called irreducible.

8.2 Recurrence and Transience

A markov chain starting from a state can visit other state. We are interested in the timing of a visit.

Definition 8.2.1 (hitting time)**.** Let $(X_n)_{n\geq 0}$ be a Markov chain with transition matrix P. The hitting time of a subset A of I is the random variable

$$H^A(\omega) = \inf \{n \geq 0 : X_n(\omega) \in A\}$$

.

Definition 8.2.2 (first passage time)**.** The first passage time to state i is the random variable T_i defined by

$$T_i(\omega) = \inf\{n \geq 1 : X_n(\omega) = i\}$$

Definition 8.2.3 (rth passage time)**.** We define rth passage time inductively by $T_i^{(0)}(\omega) = 0, T_i^{(1)}(\omega) = T_i(\omega)$ and for $r = 0, 1, 2, \cdots$,

$$T_i^{(r+1)}(\omega) = \inf\left\{n \geq T_i^{(r)}(\omega) + 1 : X_n(\omega) = i\right\}$$

A state can never be visited as well.

Definition 8.2.4 (absorption probability)**.** The probability starting from i that $(X_n)_{n\geq 0}$ ever hits A is then

$$h_i^A = P_i(H^A < \infty)$$

When A is a closed class, h_i^A is called the absorption probability.

Remark 8.2.5. A less formal notation is $h_i^A = \mathbb{P}(\text{hit } A)$.

Absorption probability can be calculated by first-step analysis.

Definition 8.2.6. The mean time taken for $(X_n)_{n\geq 0}$ to reach A from i is given by

$$k_i^A = E_i(H^A) = \sum_{n<\infty} nP(H^A = n) + \infty\mathbb{P}(H^A = \infty)$$

Remark 8.2.7. A less formal notation is $k_i^A = E_i(\text{time to hit } A)$.

We can classify the states into three categories according to whether a state is visited i.o. by a markov chain.

Definition 8.2.8 (recurrent)**.** We say a state i is recurrent if $P_i(X_n = i \text{ i.o. }) = 1$.

Definition 8.2.9 (transient)**.** We say a state i is transient if $P_i(X_n = i \text{ i.o. }) = 0$.

Definition 8.2.10 (positive recurrent)**.** A state i is postive recurrent if the expected return time $m_i = E_i(T_i)$ is finite. A recurrent state which fails to have this stronger property is called null recurrent.

The next result is immediate once we combine the strong markov property and the property of the geometric distribution.

Theorem 8.2.11. *y is recurrent if and only if $\mathbb{E}_y N(y) = \infty$.*

This is also called Green's function.

The next fact help us identify recurrent state when the state space is finite.

Theorem 8.2.12. *Let C be a finite closed set. Then C contains a recurrent state. If C is irreducible then all states in C are recurrent.*

Theorem 8.2.13 (Decomposition theorem)**.**

8.3 Invariant Measures

Definition 8.3.1 (invariant measure)**.** A measure is any row vector $(\lambda_i : i \in I)$ with non-negative entries. We say λ is invariant if

$$\lambda P = \lambda$$

Remark 8.3.2. Invariant measure is also called stationary measure.

Definition 8.3.3 (detailed balance)**.**

Definition 8.3.4 (reversible measure)**.** A measure that satisfies the detailed balance condition is said to be a reversible measure.

These quantities can be computed by definition.

Example 8.3.5 (Chip-Firing/Sand-Pile)**.** Every time, pick a random τ such that $a_\tau \geq 2$ and do the update

$$\begin{cases} a_\tau^{t+1} = a_\tau^t - 2 \\ a_{\tau+1}^{t+1} = a_{\tau+1}^t + 1 \\ a_{\tau-1}^{t+1} = a_{\tau-1}^t + 1 \end{cases}$$

The next theorem explain why reversible measure is reversible.

Theorem 8.3.6 (dual transition probability)**.**

A necessary and sufficient condition for a chain to have a reversible measure is given below. This condition can be checked if the transition probability is given.

Theorem 8.3.7 (Kolmogorov's cycle condition)**.**

Only special chains have reversible measures, but the next result shows that many markov chains have stationary measures by relating it to the expected number of visits during an excursion.

Theorem 8.3.8. *Let x be an recurrent state and let $T = \inf\{n \geq 1 : X_n = x\}$. Then*

$$\mu_x(y) = \mathbb{E}_x(\sum_{n=0}^{T-1} 1_{X_n = y})$$

defines a stationary measure.

Theorem 8.3.9. *If p is irreducible and recurrent, then the stationary measure is unique up to constant multiples.*

Having examine the existence and uniqueness of stationary measures, we turn our attention to stationary distributions. Stationary measures may exist for transient chains, e.g. random walks in $d \geq 3$, but stationary distribution only exist for recurrent chains, as the following theorem shows:

Theorem 8.3.10. *If there is a stationary distribution, then all states y with $\pi(y) > 0$ are recurrent.*

We can relate stationary distributions to the expected time of an excursion

Theorem 8.3.11. *If p is irreducible and has stationary distribution π, then*

$$\pi(x) = \frac{1}{\mathbb{E}_x T_x}$$

Recall the concept of positive recurrent and null recurrent. The next result

Theorem 8.3.12. *If p is irreducible, then TFAE:*

- *Some x is positive recurrent.*
- *There is a stationary distribution.*
- *All states are positive recurrent.*

8.4 Ergodic Theorems

The first topic in this section is to investigate

Let

$$N_n(y) = \sum_{m=1}^{n} 1_{X_m = y}$$

be the number of visits to y by time n.

Theorem 8.4.1. *Suppose y is recurrent. For any $x \in S$, as $n \to \infty$*

$$\frac{N_n(y)}{n} \to \frac{1}{E_y T_y} 1_{T_y < \infty}$$

Proof. a

Step 1 Suppose first we start at $x = y$. Let $R(k) = \min\{n \geq 1 : N_n(y) = k\}$ and $t_k = R(k) - R(k-1)$, where $R(0) = 0$. Then t_i are i.i.d. and SLLN implies

$$\frac{R(k)}{k} \to \mathbb{E}_y T_y$$

Step 2 Then we generalize to $x \neq y$. The result is obviously true if $T_y = \infty$. □

Example 8.4.2.

Definition 8.4.3 (period)**.** let d_x by the greatest common divisor of I_x. d_x is called the period of x.

The next lemma says that period is a class property.

Lemma 8.4.4. *If $\rho_{xy} > 0$, then $d_y = d_x$.*

Theorem 8.4.5. *Suppose p is irreducible, aperiodic, and has stationary distribution π. Then as $n \to \infty$, $p^n(x, y) \to \pi(y)$.*

Proof. The proof technique is called **coupling**. Let $S^2 = S \times S$. Define a transition probability $\bar{p}$ on $S \times S$ by

$$\bar{p} = p((x_1, y_1), (x_2, y_2)) = p(x_1, x_2)p(y_1, y_n)$$

i.e. each coordinate moves independently.

$\bar{p}$ **is irreducible.** This is the only step that requires aperiodicity.

$\bar{p}$ **is recurrent.** This is becasue $\bar{\pi}(a, b) = \pi(a)\pi(b)$ defines a stationary distribution for $\bar{p}$.

Now let (X_n, Y_n) denote the cahin on $S \times S$, and let T be the first time that this chain hits the diagonal $\{(y, y) : y \in S\}$. Let $T_{(x,x)}$ be the hitting time of (x, x). Since $\bar{p}$ is irreducible and recurrent, $T_{(x,x)} < \infty$ a.s. and hence $T < \infty$ a.s..

Now we want to show the following key identity:

$$\sum_y |\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)| \leq 2\mathbb{P}(T > n).$$

This is because on $\{T \leq n\}$, the two coordinates X_n and Y_n have the same distribution.

Finally we let $X_0 = x$ and $Y_0 \sim \pi$, then Y_n has distribution π, and it follows that

$$\sum_y |p^n(x, y) - \pi(y)| \leq 2\mathbb{P}(T > n) \to 0.$$

□

Remark 8.4.6. Some shortcuts exist for helping to determine when a Markov chain is ergodic, i.e. irreducible, aperiodic, and positive recurrent.

1. A Markov chain with a finite number of states has only transient and positive recurrent states. Only a Markov chain with an infinite number of states can be null recurrent.
2. A sufficient test for a state to be aperiodic is that it has a self-loop.
3. In an irreducible, finite state Markov chain, the presence of one aperiodic state guarantees ergodicity.

8.5 Introduction to Markov Chain Mixing

Contraction

Canonical Path

Cheeger's Constant

8.6 Introduction to Markov Random Fields

8.6.1 Basics

8.6.2 Coloring

Glauber Dynamics Every time pick a random vertex and assign a random color which is not used by any neighbors.

8.6.3 1D Ising Model without Magnetic Field

The Ising model is a theoretical model in statistical physics that was originally developed to describe ferromagnetism, a property of certain materials such as iron. A system of magnetic particles can be modeled as a linear chain in one dimension or a lattice in two dimension, with one molecule or atom at each lattice site i. To each molecule or atom a magnetic moment is assigned that is represented in the model by a discrete variable σ_i. Each 'spin' can only have a value of $\sigma_i = \pm 1$. The two possible values indicate whether two spins σ_i and σ_j are align and thus parallel $(\sigma_i \cdot \sigma_j = +1)$ or anti-parallel $(\sigma_i \cdot \sigma_j = +1)$

A system of two spins is considered to be in a lower energetic state if the two magnetic moments are aligned. If the magnetic moments points in opposite directions they are consider to be in a higher energetic state. Due to this interaction the system tends to align all magnetic moments in one direction in order to minimise energy. If nearly all magnetic moments point in the same direction the arrangement of molecules behaves like a macroscopic magnet.

A phase transition in the context of the Ising model is a transition from an ordered state to a disordered state. A ferromagnet above the critical temperature T_C is in a disordered state. In the Ising model this corresponds to a random distribution of the spin values. Below the critical temperature T_C (nearly) all spins are aligned, even in the absence of an external applied magnetic field H. If we heat up a cooled ferromagnet, the magnetization vanishes at T_C and the ferromagnet switches from an ordered to a disordered state. This is a phase transition of second order.

8.6.4 1D Ising Model with Magnetic Field

8.6.5 2D Ising Model: Peirls Proof

The Hamiltonion of the system is

The idea of Peirls proof is very similar to the idea of reflection principle, which map

8.7 Electrical Networks

8.7.1 The Correspondence

Assume a unit voltage is charged on a and b is grounded, i.e. $\varphi(a) = 1$ and $\varphi(b) = 0$.

The Correspondence between Reversible Markov Chains and Electrical Networks Given an electrical network, we can define a corresponding reversible Markov chain as follows. Let the **transition probability** be

$$p(x, y) = \frac{C_{xy}}{C_x},$$

where

$$C_x = \sum_y C_{xy}.$$

This chain is indeed reversible because we can define a measure as

$$\pi(x) = C_x,$$

then the detailed balance condition is satisdied as

$$\pi(x)p(x, y) = C_{xy} = C_{yx} = \pi(y)p(y, x).$$

Conversely, given a reversible Markov chain with reversible measure $\pi(x)$, i.e. $\pi(x)p(x,y) = \pi(y)p(y,x)$, we can define a corresponding electrical network as follows. Let

$$C_{xy} = \pi(x)p(x,y),$$

then it is indeed a well-defined electrical network as

$$C_{xy} = \pi(x)p(x,y) = \pi(y)p(y,x) = C_{yx}.$$

Voltage

Theorem 8.7.1 (Voltage as)**.** $\varphi(x) = \mathbb{P}_x(\tau_a < \tau_b)$

Proof. $\varphi(x)$ satisfies the same harmonic equation as $\mathbb{P}_x(\tau_a < \tau_b)$. Moreover, the boundary condition is also the same. □

Theorem 8.7.2 (Electrical Current as Edge Crossings)**.** $I(x,y) = ?$

Proof.

$$I(x,y) = C_{xy}(\varphi(x) - \varphi(y))$$

□

Effective Resistance

Definition 8.7.3 (effective resistance)**.** $R_{\text{eff}} = \frac{1}{I(a^+)}$

Theorem 8.7.4. $\mathbb{P}_b(\tau_a < \sigma_b) = \frac{1}{C_b R_{eff}}$

Proof.

$$\begin{aligned}\mathbb{P}_b(\tau_a < \sigma_b) &= \sum_x p(b,x)\varphi(x)\\ &= \sum_x \frac{C_{bx}\varphi(x)}{C_b}\\ &= \sum_x \frac{I(b,x)}{C_b}\\ &= \frac{1}{C_b R_{\text{eff}}}\end{aligned}$$

□

Thomson Principle Nash-Williams Criterion

Dirichlet Principle

8.7.2 Random Walks

8.8 Introduction to Random Walks

8.8.1 1D Simple Random Walk

Let $S_n = \sum_{i=1}^n \xi_i$ where ξ_i's are i.i.d. Rademacher random varaibles.

Reflection Principle

Lemma 8.8.1. $\mathbb{P}_0(\tau_i < n, S_n = i + j) = \mathbb{P}_0(\tau_i < n, S_n = i - j)$

Proof. Note that when $\tau_i < n$ and $S_n = i + j$, the trajectory must cross the line i. Reflect the trajectory with respect to i yields a trajectory which satisfies $\tau_i < n$ and $S_n = i - j$, and vice versa. Therefore a bijection between these two events is constructed. Noting that the weights of all these trajectories induced by the probability distribution are equal. □

Remark 8.8.2. Note that $\mathbb{P}_0(\tau_i < n, S_n = i + j) = \mathbb{P}_0(S_n = i + j)$.

Theorem 8.8.3 (Ballot Theorem)**.** $\mathbb{P}_0(\tau_i = n|S_n = i) = \frac{i}{n}$

Proof.

$$\begin{aligned}
\mathbb{P}_0(\tau_i = n) &= \mathbb{P}_0(S_n = i) - \mathbb{P}_0(\tau_i < n, S_n = i)\\
&= \mathbb{P}_0(S_n = i) - \frac{1}{2}\mathbb{P}_0(\tau_i < n, S_{n-1} = i - 1) - \frac{1}{2}\mathbb{P}_0(\tau_i < n, S_{n-1} = i + 1)\\
&= \mathbb{P}_0(S_n = i) - \mathbb{P}_0(\tau_i < n, S_{n-1} = i + 1)\\
&= \mathbb{P}_0(S_n = i) - \mathbb{P}_0(S_{n-1} = i + 1)\\
&= C_n^{\frac{n+i}{2}}\frac{1}{2^n} - C_{n-1}^{\frac{n+i}{2}}\frac{1}{2^{n-1}}\\
&= \frac{i}{n}C_n^{\frac{n+i}{2}}\frac{1}{2^n}\\
&= \frac{i}{n}\mathbb{P}_0(S_n = i)
\end{aligned}$$

□

Remark 8.8.4. Another interpretation of this result is that consider the trajectory of. The first time that it reaches There is exactly i such time. Therefore,

Theorem 8.8.5. $\mathbb{P}_0(\tau_1 > 2n - 1) = \mathbb{P}_0(S_{2n} = 0)$

Proof.

$$\begin{aligned}
\mathbb{P}_0(\tau_1 \le 2n) &= \sum_i \mathbb{P}_0(\tau_1 \le 2n, S_{2n} = i)\\
&= \sum_{i\ge 1} \mathbb{P}_0(\tau_1 \le 2n, S_{2n} = i) + \sum_{i\le 0} \mathbb{P}_0(\tau_1 \le 2n, S_{2n} = i)\\
&= 2\sum_{i\ge 1} \mathbb{P}_0(S_{2n} = i)\\
&= 1 - \mathbb{P}_0(S_{2n} = 0)
\end{aligned}$$

□

Remark 8.8.6. $\mathbb{P}_0(S_1 \neq 0, \cdots, S_{2n} \neq 0) = \mathbb{P}_0(S_{2n} = 0)$

Theorem 8.8.7.

Corollary 8.8.8. $\mathbb{P}_0(\tau_1 < \infty) = 1, \mathbb{E}_0\tau_1 = \infty$

Arcsin Laws Let $u_{2n} = \mathbb{P}_0(S_{2n} = 0)$. We first describe the arcsin law for the last visit to 0.

Lemma 8.8.9. *Let* $\sigma_{2n} = \sup\{m \le 2n : S_m = 0\}$*. Then*

$$\mathbb{P}(\sigma_{2n} = 2k) = u_{2k}u_{2n-2k}$$

Proof. $\mathbb{P}(\sigma_{2n} = 2k) = \mathbb{P}(S_{2k} = 0, S_{2k+1} \neq 0, \cdots, S_{2n} \neq 0)$. □

Theorem 8.8.10. *For* $0 < a < b < 1$,

$$\mathbb{P}_0(a \le \frac{\sigma_{2n}}{2n} \le b) \to \int_a^b \frac{1}{\pi}\frac{1}{\sqrt{x(1-x)}}$$

Corollary 8.8.11. $\lim_{n\to\infty} \mathbb{P}_0(S_r \neq 0 \quad \forall \delta n < r \leq n) = \frac{2}{\pi} \arcsin \sqrt{\delta}$

Remark 8.8.12. Anti-concentration Ineqaulity

Next, we prove the arcsin law for the time above 0.

Lemma 8.8.13. *Let π_{2n} be the number of segments $(k-1, S_{k-1}) \to (k, S_k)$ that lie above the axis, i.e. in $\{(x,y) : y \geq 0\}$. Then*

$$\mathbb{P}_0(\pi_{2n} = 2k) = u_{2k} u_{2n-2k}.$$

Corollary 8.8.14. *For $0 < a < b < 1$,*

$$\mathbb{P}_0(a \leq \frac{\pi_{2n}}{2n} \leq b) \to \int_a^b \frac{1}{\pi} \frac{1}{\sqrt{x(1-x)}}$$

Maringale Now we want to calculate the moments of τ. Our method involves differentiating the exponential martingales. For a simple random walk,

$$Y_n(t) := \frac{e^{tS_n}}{(\cosh t)^n}$$

is a martingale. Thus its derivatives are also martingales.

$$\frac{\mathrm{d}Y_n(t)}{\mathrm{d}t} = \frac{e^{tS_n}}{(\cosh t)^{n+1}} (S_n \cosh t - n \sinh t)$$

so $\frac{\mathrm{d}Y_n(t)}{\mathrm{d}t}|_{t=0} = S_n$ is a martingale, as expected.

$$\frac{\mathrm{d}^2 Y_n(t)}{\mathrm{d}t^2} = \frac{e^{tS_n}}{(\cosh t)^{n+1}} ((S_n^2 - n) \cosh t - 2nS_n \sinh t) + n(n+1) \frac{e^{tS_n} \sinh^2 t}{(\cosh t)^{n+2}}$$

so $\frac{\mathrm{d}^2 Y_n(t)}{\mathrm{d}t^2}|_{t=0} = S_n^2 - n$ is a martingale, as expected.

8.8.2 Lamplighter

Example 8.8.15. Consider

reversibility

transience

recurrence

8.9 Introduction to Branching Process

8.9.1 Galton-Watson Process

Let $\{\xi_{ni} : n \geq 0, i \geq 0\}$ be a set of independent and identically-distributed natural number-valued random variables. A Galton-Watson process is a stochastic process $\{X_n\}$ which evolves according to the recurrence formula $X_0 = 1$ and

$$X_{n+1} = \sum_{i=1}^{X_n} \xi_{ni}.$$

Our goal is to analysis the properties of this process.

Mean and Variance As the Galton-Watson process is tree-like, it possesses many recurrence structure. The first one we would utilize is the martingale property. We have

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n \mathbb{E}\xi,$$

so that

$$\frac{X_n}{(\mathbb{E}\xi)^n}$$

is a martingale. By the property of martingales, we have

$$\mathbb{E}X_n = (\mathbb{E}\xi)^n.$$

If we further assume that $\operatorname{Var}\xi < \infty$, we can calculate the variance of X_n similarly. We have

$$\begin{aligned}\operatorname{Var}(X_{n+1}) &= \operatorname{Var}(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) + \mathbb{E}\operatorname{Var}(X_{n+1}|\mathcal{F}_n)\\ &= (\mathbb{E}\xi)^2 \operatorname{Var}(X_n) + \mathbb{E}X_n \operatorname{Var}\xi\\ &= (\mathbb{E}\xi)^2 \operatorname{Var}(X_n) + (\mathbb{E}\xi)^n \operatorname{Var}\xi\end{aligned}$$

and we can solve this update formula by noting that $\operatorname{Var}X_1 = \operatorname{Var}\xi$, which yields

$$\operatorname{Var}X_n = \operatorname{Var}\xi \sum_{i=n}^{2n-1} (\mathbb{E}\xi)^n.$$

The Extinction Probability The key quantity we are interested in is the extinction probability (i.e. the probability of final extinction), which is given by

$$\lim_{n\to\infty} \mathbb{P}(X_n = 0).$$

Note that once $X_n = 0$, then $X_{n+k} = 0$ for all $k \geq 1$, so 0 is an absorbing state.

Theorem 8.9.1 (subcritical). *If $\mu < 1$, then X_n extincts.*

Proof. $\mathbb{P}(X_n > 0) = \mathbb{P}(X_n \geq 1) \leq \mathbb{E}X_n = \mu^n \to 0$. □

Remark 8.9.2. The extinction rate when $\mu < 1$ is exponentially fast.

Theorem 8.9.3 (critical). *If $\mu = 1$ and $\mathbb{P}(\xi = 1) < 1$, then X_n extincts.*

Proof. When $\mu = 1$, X_n itself is a nonnegative martingale. So X_n converges to a finite limit X_∞ a.s.. As X_n is integer valued, we must have $X_n = X_\infty$ for large n. However, for any $k > 0$, $\mathbb{P}(X_n = k \quad \forall n \geq N) = 0$ for any N because $\mathbb{P}(\xi = 1) < 1$. So we must have $X_\infty = 0$. □

In this case, the extinction rate is at most $\frac{1}{n}$ by an easy second moment estimate.

Theorem 8.9.4. *If $\mu = 1$, then $\mathbb{P}(X_n \geq 1) \geq \frac{1}{1+n\operatorname{Var}(\xi)}$.*

Proof. $\mathbb{P}(X_n \geq 1) \geq \frac{(\mathbb{E}X_n)^2}{\mathbb{E}X_n^2} = \frac{1}{1+n\operatorname{Var}(\xi)}$. □

This rate cannot be improved.

Example 8.9.5. Let $\xi \sim \operatorname{Geo}(\frac{1}{2})$. Then $\mathbb{P}(X_n \geq 1) = \frac{1}{n+1}$.

Proof. The distribution of X_n can be described by its generating function $f_{X_n}(s) = f^{(n)}(s)$ where $f(s) = \frac{1}{2-s}$. So $f(0) = \frac{1}{2}$, $f^{(2)}(0) = \frac{2}{3}$, and $f^{(n)} = \frac{n}{n+1}$. Now $\mathbb{P}(X_n = 0) = f_{X_n}(0) = f^{(n)}(0)$, so $\mathbb{P}(X_n \geq 1) = 1 - \mathbb{P}(X_n = 0) = \frac{1}{n+1}$. □

Remark 8.9.6. Another view of this result. This is the probability that a simple random walk

For $s \in [0,1]$, let $f(s) = \sum_{k\geq 0} p_k s^k$ be the generating function of ξ. Then $f(1) = 1$ and $f'(s) \geq 0$.

Lemma 8.9.7. *The generating function for X_n*

Theorem 8.9.8 (supercritical). *If $\mu > 1$, then $\lim_{n\to\infty} \mathbb{P}(X_n = 0) = \rho$, where ρ is the only solution of $f(\rho) = \rho$ in $[0,1)$.*

Proof. This is basically because a tree extinct if and only if all children trees extinct. □

8.9.2 Biased Random Walk on a Galton-Watson Tree

Let $\mathbb{Q}(\cdot) = \mathbb{P}(\cdot||\mathbb{T}| = \infty)$ be the measure conditioned on all Galton-Watson trees that survive.

Lemma 8.9.9. *On the event of nonextinction, $X_n \to \infty$ almost surely, provided $p_1 \neq 1$.*

Proof. If 0 is the only nontransient state state of X_n, then the result follows. Thus we want to show that any state other than 0 is transient.

Now for any $k > 0$, eventually returning to k requires not immediately being extinct, which has probability $\leq 1 - p_0^k$. If $p_0 > 0$, then k is transient. If $p_0 = 0$, as $p_1 \neq 1$, k is also transient. □

Definition 8.9.10 (inherited property)**.** Call a property of trees inherited if
(i) every finite tree has this property and
(ii) if whenever a tree has this property, so do all the descendant trees of the children of the root.

Theorem 8.9.11 (0-1 law for inherited properties)**.** *Every inherited property has conditional probability either 0 or 1 given nonextinction.*

Proof. Let A be the set of trees possessing a given inherited property. For a tree T with k children of the root, let $T(1), \ldots, T(k)$ be the descendant trees of these children. Then

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{P}[T \in A \,|\, X_1]] \leq \mathbb{E}[\mathbb{P}[T(1) \in A, \ldots, T(X_1) \in A \,|\, Z_1]]$$

by definition of inherited. Since $T(1), \ldots, T(Z_1)$ are i.i.d. given Z_1, the last quantity above is equal to $\mathbb{E}[\mathbb{P}(A)^{Z_1}] = f(\mathbb{P}(A))$. Thus, $\mathbb{P}(A) \leq f(\mathbb{P}(A))$. On the other hand, $\mathbb{P}(A) \geq q$ since every finite tree is in A. It follows upon inspection of a graph of f that $\mathbb{P}(A) \in [q, 1]$, from which the desired conclusion follows. □

Corollary 8.9.12. *Either $W = 0$ a.s. or $W > 0$ a.s. on nonextinction*

Theorem 8.9.13 (The Seneta-Heyde Theorem)**.** *If $1 < \mu < 1$, then there exist constants c_n such that*

(i) $\lim \frac{Z_n}{c_n}$ *exists almost surely in* $[0, \infty)$*;*

(ii) $\mathbb{P}[\lim \frac{Z_n}{c_n} = 0] = \rho$*;*

(iii) $\frac{c_{n+1}}{c_n} \to m$*.*

Proof. □

Let η be the (corresponds to ξ)

$$q_k = \sum_{l=0}^{\infty} p_{k+l} C_{k+l}^k \rho^{k-1} (1-\rho)^l \quad k \geq 1$$

We will often want to consider random trees produced by a Galton-Watson branching process. For a precise formulation of tree-valued random variables, one is referred to.

$$\frac{1}{R} = \sum_{i=1}^{\eta} \frac{1}{\lambda + \lambda R^{(i)}}$$

therefore $R = \infty$ if and only if $R^{(i)} = 0\ \forall i$.

Lemma 8.9.14 (0-1 law)**.** $\mathbb{Q}(R = \infty) = 0$ *or* 1

Theorem 8.9.15. *When $\lambda \geq m$, recurrent*

Theorem 8.9.16. *When $\lambda < m$, transient*

Example 8.9.17 (3-1 tree)**.**

Definition 8.9.18 (Branching Number)**.** The branching number of a tree T is the supremum of those λ that admit a positive total amount of water to flow through T ; denote this critical value of λ by br$T$

Precolation

Theorem 8.9.19. *Let T be the family tree of a Galton-Watson process with mean $\mu > 1$. Then $p_c(T) = \frac{1}{\mu}$ almost surely, given nonextinction.*

The basic intuition goes like this. If T is an n-ary tree, then the cluster of the root under percolation is a Galton-Watson tree with progeny distribution Binomial(n, p). Thus, this cluster is infinite with positive probability if and only if $np > 1$, whence $p_c(T) = \frac{1}{n}$.

Proof. Let T be a given tree, and write K for the cluster of the root of T after percolation on T with the survival parameter p. When T has the law of a Galton-Watson tree with mean μ, we claim that K has the law of another Galton-Watson tree having mean μp: if Y_i represent i.i.d. Bin$(1, p)$ random variables that are also independent of L, then

$$\mathbb{E}\left[\sum_{i=1}^{L} Y_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{L} Y_i \,|\, L\right]\right] = \mathbb{E}\left[\sum_{i=1}^{L} \mathbb{E}[Y_i]\right] = \mathbb{E}\left[\sum_{i=1}^{L} p\right] = \mu p.$$

Hence, K is finite almost surely if and only if $mp \leq 1$. Since

$$\mathbb{E}[\mathbb{P}(|K| < \infty \,|\, T)] = \mathbb{P}(|K| < \infty), \tag{8.1}$$

this means that for almost every Galton-Watson tree T, the cluster of its root is finite almost surely if $p \leq \frac{1}{\mu}$. On the other hand, for fixed p, the property $\{T; P_p(|K| < \infty) = 1\}$ is inherited, so has probability ρ or 1. If it has probability 1, then Equation (8.1) shows that $mp \leq 1$. That is, if $mp > 1$, this property has probability ρ, so that the cluster of the root of T will be infinite with positive probability almost surely on the event of nonextinction. Considering a sequence $p_n \to \frac{1}{\mu}$, we see that this holds almost surely on the event of nonextinction for all $p > \frac{1}{\mu}$ at once, not just for a fixed p. We conclude that $p_c(T) = \frac{1}{\mu}$ almost surely on nonextinction.

□

Corollary 8.9.20. *If T is a Galton-Watson tree with mean $\mu > 1$, then $br(T) = \mu$ almost surely, given nonextinction.*

Theorem 8.9.21. *For an independent percolation and adapted conductances on the same tree, we have*

$$\frac{C(o \leftrightarrow 1)}{1 + C(o \leftrightarrow 1)} \leq \mathbb{P}[o \leftrightarrow 1].$$

Corollary 8.9.22. *For any locally finite infinite tree T,*

$$p_c(T) = \frac{1}{brT}$$

Chapter 9

Measurable Space Discrete Time Markov Chain

Now we develop a more formal theory for discrete time Markov Chains by means of measure theory. Let $(S, \mathcal{S}) \to \mathbb{R}$ be a measurable space. This is the state space for our Markov chain.

Chapter 10

Jump Process

Definition 10.0.1 (continuous-time random process)**.** Let I be a countable set. A continuous-time random process

$$(X_t)_{t\geq 0} = (X_t : 0 \leq t \leq \infty)$$

with values in I is a family of random variables $X_t : \Omega \to I$.

We are going to consider ways in which we might specify the probabilistic behavior of $(X_t)_{t\geq 0}$. To avoid uncountable union, we shall restrict our attention to processes $(X_t)_{t\geq 0}$ which are right-continuous.

Definition 10.0.2 (right continuous)**.** In the context of discrete space continuous time, a right-continuous process means $\forall \omega \in \Omega$ and $t \geq 0$, $\exists \epsilon > 0$ s.t.

$$X_s(\omega) = X_t(\omega) \quad t \leq s \leq t + \epsilon$$

Definition 10.0.3 (increment)**.** If $(X_t)_{t\geq 0}$ is a real-valued process, we can consider its increment $X_t - X_s$ over any interval $(s, t]$.

Definition 10.0.4 (stationary)**.** We say that $(X_t)_{t\geq 0}$ has stationary increments if the distribution of $X_{s+t} - X_s$ depends only on $t \geq 0$.

Definition 10.0.5 (independent)**.** We say that $(X_t)_{t\geq 0}$ has independent increments if its increments over amy finite collection of disjoint intervals are independent.

Definition 10.0.6 (Q-matrix)**.** A Q-matrix on I is a matrix $Q = (q_{ij} : i, j \in I)$ satisfying the following conditions:
(i) $\forall i \quad 0 \leq -q_{ii} < \infty$
(ii) $\forall i \neq j \quad q_{ij} \geq 0$
(iii) $\forall i \quad \sum_{j\in I} q_{ij} = 0$

10.0.1 Review: Properties of Exponential Distribution

Definition 10.0.7. A random variable $T : \Omega \to [0, \infty]$ has an exponential distribution of parameter λ $(0 \leq \lambda < \infty)$ if

$$\mathbb{P}(T > t) = e^{-\lambda t} \quad \text{for all } t \geq 0.$$

We write $T \sim \mathrm{E}(\lambda)$ for short. If $\lambda > 0$, then T has a density function

$$f_T(t) = \lambda e^{-\lambda t} 1_{\{t\geq 0\}}.$$

Remark 10.0.8. The mean of T is given by

$$\mathbb{E}(T) = \int_0^\infty P(T > t)\, dt = \lambda^{-1}.$$

Theorem 10.0.9 (memoryless property)**.** *A random variable $T : \Omega \to (0, \infty]$ has an exponential distribution if and only if it has the following memoryless property:*

$$\mathbb{P}(T > s + t \,|\, T > s) = \mathbb{P}(T > t) \quad \textit{for all } s, t \geq 0.$$

Proof. Suppose $T \sim \mathrm{E}(\lambda)$, then

$$\mathbb{P}(T > s+t \,|\, T > s) = \frac{\mathbb{P}(T > s+t)}{\mathbb{P}(T > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}(T > t).$$

On the other hand, suppose T has the memoryless property whenever $P(T > s) > 0$. Then $g(t) = P(T > t)$ satisfies

$$g(s+t) = g(s)g(t) \quad \text{for all } s, t \geq 0.$$

We assumed $T > 0$ so that $g\left(\frac{1}{n}\right) > 0$ for some n. Then, by induction,

$$g(1) = g(\frac{1}{n} + \cdots + \frac{1}{n}) = g\left(\frac{1}{n}\right)^n > 0,$$

so $g(1) = e^{-\lambda}$ for some $0 \leq \lambda < \infty$. By the same argument, for integers $p, q \geq 1$,

$$g\left(\frac{p}{q}\right) = g\left(\frac{1}{q}\right)^p = g(1)^{p/q},$$

so $g(r) = e^{-\lambda r}$ for all rationals $r > 0$. For real $t > 0$, choose rationals $r, s > 0$ with $r \leq t \leq s$. Since g is decreasing,

$$e^{-\lambda r} = g(r) \geq g(t) \geq g(s) = e^{-\lambda s}$$

and, since we can choose r and s arbitrarily close to t, this forces $g(t) = e^{-\lambda t}$, so $T \sim \mathrm{E}(\lambda)$. □

Theorem 10.0.10 (infimum)**.** *Let I be a countable set and let T_k $k \in I$ be independent random variables with $T_k \sim E(q_k)$ and $0 < q := \sum_{k \in I} q_k < \infty$. Set $T = \inf_k T_k$. Then this infimum is attained at a unique random value K of k a.s.. Moreover, T and K are independent, with $T \sim E(q)$ and $\mathbb{P}(K = k) = \frac{q_k}{q}$.*

Proof. Set $K = k$ if $T_k < T_j$ for all $j \neq k$, otherwise, let K be undefined. Then

$$\begin{aligned}
\mathbb{P}(K = k \text{ and } T \geq t) &= \mathbb{P}(T_k \geq t \text{ and } T_j > T_k \text{ for all } j \neq k) \\
&= \int_t^\infty q_k e^{-q_k s} \mathbb{P}(T_j > s \text{ for all } j \neq k)\, ds \\
&= \int_t^\infty q_k e^{-q_k s} \prod_{j \neq k} e^{-q_j s}\, ds \\
&= \int_t^\infty q_k e^{-qs}\, ds = \frac{q_k}{q} e^{-qt}.
\end{aligned}$$

Hence, $\mathbb{P}(K = k \text{ for some } k) = 1$, and T and K have the claimed joint distribution. □

Theorem 10.0.11. *Let $S_1, S_2, \ldots$ be a sequence of independent random variables with $S_n \sim E(\lambda_n)$ and $0 < \lambda_n < \infty$ for all n.*
(i) If $\sum_{n=1}^\infty \frac{1}{\lambda_n} < \infty$, then $\mathbb{P}\left(\sum_{n=1}^\infty S_n < \infty\right) = 1$.
(ii) If $\sum_{n=1}^\infty \frac{1}{\lambda_n} = \infty$, then $\mathbb{P}\left(\sum_{n=1}^\infty S_n = \infty\right) = 1$.

Proof. (i) Suppose $\sum_{n=1}^\infty \frac{1}{\lambda_n} < \infty$. Then, by monotone convergence

$$\mathbb{E}\left(\sum_{n=1}^\infty S_n\right) = \sum_{n=1}^\infty \frac{1}{\lambda_n} < \infty$$

so

$$\mathbb{P}\left(\sum_{n=1}^\infty S_n < \infty\right) = 1.$$

(ii) Suppose instead that $\sum_{n=1}^\infty \frac{1}{\lambda_n} = \infty$. Then $\prod_{n=1}^\infty (1 + \frac{1}{\lambda_n}) = \infty$. By monotone convergence and independence

$$\mathbb{E}\left[\exp\left\{-\sum_{n=1}^\infty S_n\right\}\right] = \prod_{n=1}^\infty \mathbb{E}\left[\exp\{-S_n\}\right] = \prod_{n=1}^\infty (1 + \lambda_1 n)^{-1} = 0$$

so

$$\mathbb{P}\left(\sum_{n=1}^{\infty} S_n = \infty\right) = 1.$$

□

Theorem 10.0.12. *For independent random variables $S \sim E(\lambda)$ and $R \sim E(\mu)$ and for $t \geq 0$, we have*

$$\mu\mathbb{P}(S \leq t < S + R) = \lambda\mathbb{P}(R \leq t < R + S).$$

Proof. We have

$$\mu\mathbb{P}(S \leq t < S + R) = \int_0^t \int_{t-s}^{\infty} \lambda\mu e^{-\lambda s} e^{-\mu r}\, dr\, ds = \lambda\mu \int_0^t e^{-\lambda s} e^{-\mu(t-s)}\, ds$$

from which the identity follows by symmetry.

□

10.0.2 Poisson Process

We begin with a definition of Poisson process in terms of jump chain and holding times, and then relate it to the infinitesimal definition and transition probability definition.

Definition 10.0.13. A right-continuous process $(X_t)_{t\leq 0}$ with values in $\mathbb{N}_{\geq 0}$ is a Poisson process of rate $\lambda \in (0, \infty)$ if its holding times $S_1, S_2, \cdots$ are i.i.d. exponential random variables of mean λ and its jump chain is given by $Y_n = n$.

Theorem 10.0.14. *Let $(X_t)_{t\geq 0}$ be an increasing, right-continuous integer-valued process starting from 0. Let $\lambda \in (0, \infty)$. TFAE:*
(i) (jump chain holding time definition) the holding times $S_1, S_2, \cdots$ of $(X_t)_{t\geq 0}$ are i.i.d. exponential random variables of mean λ and the jump chain is given by $Y_n = n$.
(ii) (infinitesimal definition) $(X_t)_{t\geq 0}$ has independent increments and as $h \downarrow 0$, uniformly in t,

$$\mathbb{P}(X_{t+h} - X_t = 0) = 1 - \lambda h + o(h), \quad \mathbb{P}(X_{t+h} - X_t = 1) = \lambda h + o(h)$$

(iii) (incremental definition) $(X_t)_{t\geq 0}$ has stationary independent increments and for each t, X_t has Poisson distribution of parameter λt.

Theorem 10.0.15. *Let $(X_t)_{t\geq 0}$ be a Poisson process. Then, conditional on $(X_t)_{t\geq 0}$ having exactly one jump in the interval $[s, s+t]$, the time at which that jump occurs is uniformly distributed on $[s, s+t]$.*

Theorem 10.0.16. *Let $(X_t)_{t\geq 0}$ be a Poisson process. Then, conditional on the event $\{X_t = n\}$, the jump times $J_1, \cdots, J_n$ have joint density function*

$$f(t_1, \cdots, t_n) = n! 1_{0 \leq t_1 \leq \cdots \leq t_n \leq t}$$

Remark 10.0.17. Thus, conditional on $\{X_t = n\}$, the jump times $J_1, \cdots, J_n$ have the same distribution as an ordered sample of size n from the uniform distribution on $[0, t]$.

An Approximation Scheme for Poisson Process In the same spirit as Donsker's invariance principle,

Chapter 11

Brownian Motion

11.1 Construction

Definition 11.1.1 (d-dimensional Brownian motion)**.** A d-dimensional Brownian motion $B = (B_t)_{t \geq 0}$ is a stochastic process indexed by $[0, \infty)$ taking values in $\mathbb{R}^d$ s.t.
(i) $B_0(\omega) = 0$

11.2 Sample Path Properties

Theorem 11.2.1. *Almost surely, for all $0 < a < b < \infty$, BM is not monotone on the interval $[a, b]$.*

Lemma 11.2.2. *Almost surely,*

$$\limsup_{t \to \infty} \frac{B(t)}{\sqrt{t}} = +\infty \quad \liminf_{t \to \infty} \frac{B(t)}{\sqrt{t}} = -\infty$$

Lemma 11.2.3. *Fix $t \geq 0$. Then almost surely, BM is not differentiable at t. Moreover, $D^* B(t) = +\infty$ and $D_* B(t) = -\infty$.*

These two lemmas can be strengthened to the law of iterated logarithm and nowhere differentiability.

Theorem 11.2.4. *Almost surely,*

$$\limsup_{t \to \infty} \frac{B(t)}{\sqrt{2t \log \log t}} = 1$$

Proof. □

Theorem 11.2.5. *Almost surely, BM is nowhere differentiable. Furthermore, almost surely, for all t, either $D^* B(t) = +\infty$ or $D_* B(t) = -\infty$ or both.*

Next we consider modulus of continuity.

Lemma 11.2.6. *There exists a constant $C > 0$ such that, almost surely, for every sufficiently small $h > 0$ and all $0 < t \leq 1 - h$,*

$$|B(t+h) - B(t)| \leq C \sqrt{h \log \frac{1}{h}}$$

Lemma 11.2.7. *For every constant $c < \sqrt{2}$, almost surely, for every $\epsilon > 0$ there exist $0 < h < \epsilon$ and $t \in [0, 1-h]$ with*

$$|B(t+h) - B(t)| \geq c \sqrt{h \log \frac{1}{h}}$$

Theorem 11.2.8. *Almost surely,*

$$\limsup_{h \downarrow 0} \sup_{0 \leq t \leq 1-h} \frac{|B(t+h) - B(t)|}{\sqrt{2h \log \frac{1}{h}}} = 1$$

11.3 Brownian Local Time

We denote by $D(a,b,t)$ the number of downcrossings of the interval $[a,b]$ before time t. Note that $D(a,b,t)$ is almost surely finite by the uniform continuity of Brownian motion on the compact interval $[0,t]$.

Theorem 11.3.1. *There exists a stochastic process $\{L(t) : t \geq 0\}$ called the local time at zero such that for all sequence $a_n \uparrow 0$ and $b_n \downarrow 0$ with $a_n < b_n$, a.s.,*

$$\lim_{n\to\infty} 2(b_n - a_n)D(a_n, b_n, t) = L(t) \quad \forall t > 0.$$

Moreover, this process is almost surely locally γ-Holder continuous for any $\gamma < \frac{1}{2}$.

Theorem 11.3.2. $\{L(t) : t \geq 0\} \overset{d}{=} \{M(t) : t \geq 0\}$

Theorem 11.3.3 (Occupation time representation of the local time at zero)**.** *For all sequence $a_n \uparrow 0$ and $b_n \downarrow 0$ with $a_n < b_n$, almost surely,*

$$\lim_{n\to\infty} \frac{1}{b_n - a_n} \int_0^t 1_{a_n \leq B(s) \leq b_n} \mathrm{d}s = L(t) \quad \forall t > 0$$

Theorem 11.3.4 (Ray-Knight theorem)**.** *Suppose $a > 0$ and $\{B(t) : 0 \leq t \leq T\}$ is a liinear BM started at a and stopped at time $T = \int\{t \geq 0 : B(t) = 0\}$. Then*

$$\{L^x(T) : 0 \leq x \leq a\} \overset{d}{=} \{|W(x)|^2 : 0 \leq x \leq a\},$$

where $\{W(x) : x \geq 0\}$ is a standard planar BM.

Chapter 12

Stochastic Calculus

12.1 Continuous Time Martingale

12.1.1 Stopping Times

Definition 12.1.1 (stopping time)**.** Let τ be a random time. If $\{\tau \leq t\} \in \mathcal{F}_t$ for every $t \geq 0$, then τ is called a stopping time.

Definition 12.1.2 (optional time)**.** Let T be a random time. If $\{T < t\} \in \mathcal{F}_t$ for every $t \geq 0$, then T is called a stopping time.

Lemma 12.1.3. *T is an optional time of the filtration $\{\mathcal{F}_t\}$ if and only if it is a stopping time of the right-continuous filtration $\{\mathcal{F}_{t^+}\}$.*

Corollary 12.1.4. *Every stopping time is optional, and the two concepts coincide if the filtration is right-continuous.*

Lemma 12.1.5. *If T is optional and θ is a positive constant, then $T + \theta$ is a stopping time.*

Lemma 12.1.6. *If τ, σ are stopping times, then so are $\tau \wedge \sigma$, $\tau \vee \sigma$, $\tau + \sigma$.*

Proof. The first two assertions are trivial.
For the third, start with the decomposition □

Lemma 12.1.7. *Let T, S be optional times; then $T + S$ is optional.*
Moreover, it is a stopping time if

Lemma 12.1.8. *Let $\{T_n\}_{n=1}^{\infty}$ be a sequence of optional times; then the random times*

$$\sup_{n\geq 1} T_n \quad \inf_{n\geq 1} T_n \quad \limsup_{n\to\infty} T_n \quad \liminf_{n\to\infty} T_n$$

are all optional.
Moreover, if the T_n's are stopping times, then so is $\sup_{n\geq 1} T_n$.

Definition 12.1.9 (σ-field of events determined prior to a stopping time)**.** Let τ be a stopping time of the filtration $\{\mathcal{F}_t\}$. The σ-field of events determined prior to the stopping time T consists of those events $A \in \mathcal{F}$ for which $A \cap \{\tau \leq t\} \in \mathcal{F}_t$ for every $t \geq 0$.

Lemma 12.1.10. *τ is $\mathcal{F}_\tau$-measurable.*

Proof. $\{\tau \leq t\} \cap \{\tau \leq t\} = \{\tau \leq t\} \in \mathcal{F}_t$, so $\{\tau \leq t\} \in \mathcal{F}_\tau$. □

Theorem 12.1.11. *For any two stopping time and τ, σ a random time s.t. $\sigma \leq \tau$ on Ω, we have $\mathcal{F}_\sigma \subset \mathcal{F}_\tau$.*

Proof. For every stopping time τ and positive constant t, $\tau \wedge t$ is an $\mathcal{F}_t$-measurable random variable because $\mathcal{F}_{\tau\wedge t} \subset \mathcal{F}_t$. Therefore, $\{\sigma \wedge t \leq \tau \wedge t\} \in \mathcal{F}_t$. Then for any $A \in \mathcal{F}_\sigma$ we have $A \cap \{\sigma \leq \tau\} \in \mathcal{F}_\tau$, because

$$A \cap \{\sigma \leq \tau\} \cap \{\tau \leq t\} = (A \cap \{\sigma \leq t\}) \cap \{\tau \leq t\} \cap \{\sigma \wedge t \leq \tau \wedge t\}$$

Finally notice that $\{\sigma \leq \tau\} = \Omega$. □

Remark 12.1.12. We have proved a stronger result, namely for any $A \in \mathcal{F}_\sigma$ we have $A \cap \{\sigma \leq \tau\} \in \mathcal{F}_\tau$.

Theorem 12.1.13. *Let σ and τ be stopping times. Then $\mathcal{F}_{\tau\wedge\sigma} = \mathcal{F}_\tau \cap \mathcal{F}_\sigma$.*
Moreover, $\{\tau < \sigma\}$, $\{\tau > \sigma\}$, $\{\tau \leq \sigma\}$, $\{\tau \geq \sigma\}$, $\{\tau = \sigma\}$ belongs to $\mathcal{F}_\tau \cap \mathcal{F}_\sigma$.

Proof. From the above theorem, $\mathcal{F}_{\tau\wedge\sigma} \subset \mathcal{F}_\tau \cap \mathcal{F}_\sigma$.
For $A \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$, $A \cap \{\tau \wedge \sigma \leq t\} = A \cap (\{\tau \leq t\} \cup \{\sigma \leq t\}) \in \mathcal{F}_t$. □

Theorem 12.1.14. *Let τ, σ be stopping times and X an integrable random variable. We have*
(i) $\mathbb{E}(X|\mathcal{F}_\tau) = \mathbb{E}(X|\mathcal{F}_{\sigma\wedge\tau})$ a.s. on $\{\tau \leq \sigma\}$.
(ii) $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_\tau)|\mathcal{F}_\sigma) = \mathbb{E}(X|\mathcal{F}_{\sigma\wedge\tau})$ a.s.. (i) $\mathbb{E}(X|\mathcal{F}_\tau) = \mathbb{E}(X|\mathcal{F}_{\sigma\wedge\tau})$ a.s. on $\{\tau \leq \sigma\}$.
(ii) $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_\tau)|\mathcal{F}_\sigma) = \mathbb{E}(X|\mathcal{F}_{\sigma\wedge\tau})$ a.s..

Proof. (i) Let $A \in \mathcal{F}_\tau$, then $A \cap \{\tau \leq \sigma\}$ belongs to both $\mathcal{F}_\tau$ and $\mathcal{F}_\sigma$, and therefore to $\mathcal{F}_\tau \cap \mathcal{F}_\sigma$. So

$$\int_A 1_{\tau\leq\sigma}\mathbb{E}(X|\mathcal{F}_{\tau\wedge\sigma})\mathrm{d}\mathbb{P} = \int \mathbb{E}(1_A 1_{\tau\leq\sigma} X|\mathcal{F}_{\tau\wedge\sigma})\mathrm{d}\mathbb{P} = \int_A 1_{\tau\leq\sigma} X \mathrm{d}\mathbb{P}$$

(ii) On $\{\tau \leq \sigma\}$ we have $\mathbb{E}(X|\mathcal{F}_\tau) = \mathbb{E}(X|\mathcal{F}_{\sigma\wedge\tau})$ a.s. by (i), so $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_\tau)|\mathcal{F}_\sigma) = \mathbb{E}(\mathbb{E}(X|\mathcal{F}_{\sigma\wedge\tau})|\mathcal{F}_\sigma) = \mathbb{E}(X|\mathcal{F}_{\sigma\wedge\tau})$. Similarly on $\{\sigma \leq \tau\}$ we have $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_\tau)|\mathcal{F}_\sigma) = \mathbb{E}(\mathbb{E}(X|\mathcal{F}_\tau)|\mathcal{F}_{\sigma\wedge\tau}) = \mathbb{E}(X|\mathcal{F}_{\sigma\wedge\tau})$. □

Theorem 12.1.15. *Let $X = \{X_t, \mathcal{F}_t\}$ be a progressively measurable process, and let τ be a stopping time of the filtration $\mathcal{F}_t$. Then the random variable X_τ defined on $\{\tau < \infty\}$ is $\mathcal{F}_\tau$-measurable, and the stopped process $\{X_{\tau\wedge t}, \mathcal{F}_t\}$ is progressively measurable.*

12.1.2 From Discrete to Continuous

In this subsection, we generalize inequalities and convergence results for discrete time martingales to continuous time martingales.
Let X_t be a submartingale adapted to $\{\mathcal{F}_t\}$ whose paths are right-continuous. Let $[\sigma, \tau]$ be a subinterval of $[0, +\infty)$, and let $a < b$, $\lambda > 0$ be real numbers.

Theorem 12.1.16 (Doob's inequality)**.** *Let $A = \left\{\sup_{\sigma\leq t\leq\tau} X_t^+ \geq \lambda\right\}$, then*

$$\lambda\mathbb{P}(A) \leq EX_\tau 1_A \leq EX_\tau^+$$

Proof. Let the finite set $\mathcal{S}$ consist of σ, τ and a finite subset of $[\sigma, \tau] \cap \mathbb{Q}$.

By considering an increasing sequence $\{\mathcal{S}_n\}_{n=1}^\infty$ of finite sets whose union is the whole of $([\sigma, \tau] \cap \mathbb{Q}) \cup \{\sigma, \tau\}$, we may replace S by this union in the preceding discrete version of the inequality. □

Theorem 12.1.17 (upcrossing inequality)**.**

$$(b-a)EU_{[\sigma,\tau]} \leq \mathbb{E}(X_\tau - a)^+ - \mathbb{E}(X_\sigma - a)^+$$

Theorem 12.1.18 (L^p maximum inequality)**.** *$\bar{X}_\tau = \sup_{\sigma\leq t\leq\tau} X_t^+$, then for $1 < p < \infty$,*

$$\mathbb{E}(\bar{X}_\tau^p) \leq (\frac{p}{p-1})^p E(X_\tau^+)^p$$

For the remainder of this subsection, we deal only with right-continuous processes, usually imposing no condition on the filtration $\mathcal{F}_t$.

Theorem 12.1.19 (submartingale convergence)**.** *Assume $\sup_{t\geq0} \mathbb{E}(X_t^+) < \infty$. Then $X_\infty = \lim_{t\to\infty} X_t$ exists a.s., and $\mathbb{E}|X_\infty| < \infty$. Assume $\sup_{t\geq0} \mathbb{E}(X_t^+) < \infty$. Then $X_\infty = \lim_{t\to\infty} X_t$ exists a.s., and $\mathbb{E}|X_\infty| < \infty$.*

Proof. □

Theorem 12.1.20 (optional sampling)**.** *Assume the submartingale has a last element X_∞, and let $S \leq T$ be two optional times of the filtration. We have*

$$\mathbb{E}(X_T|\mathcal{F}_{S+}) \geq X_S \quad a.s.$$

$$\mathbb{E}(X_T|\mathcal{F}_{S+}) \geq X_S \quad a.s.$$

If S is a stopping time, then $\mathcal{F}_S$ can replace $\mathcal{F}_{S+}$ above.

Proof. Consider the sequence of random times

$$S_n(\omega) = \begin{cases} +\infty & S(\omega) = +\infty \\ \frac{k}{2^n} & \frac{k-1}{2^n} \leq S(\omega) < \frac{k}{2^n} \end{cases}$$

and similarly defined sequences $\{T_n\}$. These are stopping times. For every fixed integer $n \geq 1$, both S_n and T_n take on a countable number of values and we also have $S_n \leq T_n$. □

12.1.3 Doob-Meyer Decomposition

Definition 12.1.21 (increasing process)**.** An adapted process A is called increasing if for $\mathbb{P}$-a.e. $\omega \in \Omega$ we have
(i) $A_0(\omega) = 0$
(ii) $t \mapsto A_t(\omega)$ is a nondecreasing, right-continuous function, and $EA_t < \infty$ holds for every $t \in [0, \infty)$.
An increasing process is called integrable if $EA_\infty < \infty$.

Definition 12.1.22. An increasing process A is called natural if for every bounded, right-continuous martingale $\{M_t, \mathcal{F}_t; 0 \le t < \infty\}$ we have

$$\mathbb{E}\int_{(0,t]} M_s \mathrm{d}A_s = \mathbb{E}\int_{(0,t]} M_{s^-} \mathrm{d}A_s \quad \forall 0 < t < \infty$$

$$\mathbb{E}\int_{(0,t]} M_s \mathrm{d}A_s = \mathbb{E}\int_{(0,t]} M_{s^-} \mathrm{d}A_s \quad \forall 0 < t < \infty$$

Lemma 12.1.23. *If A is an increasing process and $\{M_t, \mathcal{F}_t; 0 \le t < \infty\}$ is a bounded right-continuous martingale, then*

$$\mathbb{E}(M_t A_t) = \mathbb{E}\int_{(0,t]} M_s \mathrm{d}A_s$$

$$\mathbb{E}(M_t A_t) = \mathbb{E}\int_{(0,t]} M_s \mathrm{d}A_s$$

The following concept is a strengthening of the notion of uniform integrablity for submartingales.

Definition 12.1.24 (class DL)**.**

Theorem 12.1.25. *Let $\{\mathcal{F}_t\}$ satisfies the usual conditions. If the right-continuous submartingale $X =$ is of class DL, then it admits the decomposition as the summation if a right-continuous martingale*

12.1.4 Square Integrable Martingales

12.2 Stochastic Integration

12.2.1

12.2.2 Martingale Characterization of BM

Theorem 12.2.1 (Levy)**.**

12.2.3 Representations of Martingales by BM

Theorem 12.2.2 (time-change for martingales)**.**

Theorem 12.2.3 (representation of square-integrable martingales by BM via Ito's integral)**.**

12.2.4 The Girsanov Theorem

12.3 The PDE Connection

12.4 Stochastic Differential Equations

Chapter 13

Markov Semigroup Theory

The core idea of Markov semigroup theory is to encode the behavior of a Markov process $(X_t)_{t\geq 0}$ via operators which act on functions. We can then develop calculus rules for working with these operators, and study them using tools from functional analysis. This is analogous to how the linear algebraic study of the transition matrix of a discrete-time Markov chain reveals properties (e.g., ergodicity, convergence) of the chain.

13.1 Diffusions

13.1.1 Kolmogorov's Theory???

13.1.2 Ito's theory???

Part III

Information Theory

Chapter 14

Information Metrics

14.1 Variational Principle

For KL divergence,

Defnie the tilting of $\mathbb{Q}$ along f by $\mathrm{d}\mathbb{Q}^f = e^f \mathrm{d}\mathbb{Q}$. Now,

$$\begin{aligned} \mathrm{KL}(\mathbb{P}\|\mathbb{Q}) &= \mathbb{E}_P \log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} \\ &= \mathbb{E}_P \log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}^f} \frac{\mathrm{d}\mathbb{Q}^f}{\mathrm{d}\mathbb{Q}} \end{aligned}$$

Theorem 14.1.1 (Donsker-Varadhan)**.**

Theorem 14.1.2 (Gibbs variational principle)**.** $\log \mathbb{E}_{\mathbb{Q}} e^f = \sup_{\mathbb{P}} \left(\mathbb{E}_{\mathbb{P}} f - \mathrm{KL}(\mathbb{P}\|\mathbb{Q})\right)$

In general,

Theorem 14.1.3. $\mathrm{D}_f(\mathbb{P}\|\mathbb{Q})$

Part IV

Optimal Transport

Chapter 15

Optimal Transport

15.1 The Optimal Transport Problem

$$\inf_{\gamma\in\Gamma_{\mu,\nu}} \int c(x,y)\,\gamma(\mathrm{d}x,\mathrm{d}y)\,. \tag{KOT}$$

Proposition 15.1.1. *Let μ,ν be two probability measures on $\mathbb{R}^d$. The set $\Gamma_{\mu,\nu}$ of couplings between μ and ν is* non-empty, convex, *and* compact *with respect to the topology of weak convergence.*

15.1.1 Wasserstein Distances

Definition 15.1.2. The p-Wasserstein distance between two probability measures $\mu,\nu\in\mathcal{P}_p(\mathbb{R}^d)$ is defined by

$$W_p(\mu,\nu) = \min_{\gamma\in\Gamma_{\mu,\nu}} \left(\int \|x-y\|^p\,\gamma(\mathrm{d}x,\mathrm{d}y)\right)^{1/p}.$$

15.1.2 p=2

Recall that

$$W_2^2(\mu,\nu) = \min_{\gamma\in\Gamma_{\mu,\nu}} \int \|x-y\|^2\,\gamma(\mathrm{d}x,\mathrm{d}y)\,. \tag{W$_2^2$}$$

Theorem 15.1.3 (Brenier)**.** *Let $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$ be two probability measures such that μ has a density and let $X\sim\mu$. If $\bar\gamma$ is an optimal coupling for (W$_2^2$),*

$$\int \|x-y\|^2\,\bar\gamma(\mathrm{d}x,\mathrm{d}y) = \min_{\gamma\in\Gamma_{\mu,\nu}} \int \|x-y\|^2\,\gamma(\mathrm{d}x,\mathrm{d}y) = W_2^2(\mu,\nu)\,,$$

then there exists a convex function $\varphi:\mathbb{R}^d\to\mathbb{R}$ such that $(X,\nabla\varphi(X))\sim\bar\gamma\in\Gamma_{\mu,\nu}$.

The dual Kantorovich problem is given by

$$\sup_{\substack{f\in L^1(\mu),\,g\in L^1(\nu)\\ f(x)+g(y)\le c(x,y)}} \left\{\int f\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu\right\}. \tag{D-W$_2^2$}$$

Since the dual problem is a supremum, we want to make g as large as possible, but we must respect the constraint $f(x)+g(y)\le c(x,y)$. The optimal function g is therefore given by

$$g(y) = \inf_{x\in\mathbb{R}^d}\{c(x,y)-f(x)\}\,. \tag{15.1}$$

The function defined in (15.1) is called the *c-conjugate* or *c-transform* of f, denoted f^c, associated with the cost $c(x,y)$. This reasoning shows that we can reformulate the dual as

$$(\text{D-W}_2^2) = \sup_{f\in L^1(\mu)} \left\{\int f\,\mathrm{d}\mu + \int f^c\,\mathrm{d}\nu\right\}. \tag{15.2}$$

This is a version of the *semidual* problem. For the quadratic cost, we can go one step further and explicitly link the semidual with convex analysis. In this case, the semidual is given by

$$\inf_{\phi \in L^1(\mu)} \left\{ \int \phi \, \mathrm{d}\mu + \int \phi^* \, \mathrm{d}\nu \right\} \tag{SD}$$

where ϕ^* denotes the *convex conjugate* of ϕ.

The following proposition proves the strong duality for the quadratic cost. In general, strong duality holds for cost function that is lower semicontinuous and bounded from below.

Proposition 15.1.4. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be probability measures. Then, the dual problem (D-W_2^2) is equivalent to the semidual problem (SD) in the following sense:*

1. **Objective values:** *Write S and D for the optimal objective values of (SD) and (D-W_2^2) respectively. Then*
$$\mathsf{D} = \int \| \cdot \|^2 \, \mathrm{d}\mu + \int \| \cdot \|^2 \, \mathrm{d}\nu - 2 \cdot \mathsf{S} \, .$$

2. **Solutions:** *A pair of functions (f, g) is optimal for (D-W_2^2) if and only if $f = \| \cdot \|^2 - 2\varphi$ and $g = \| \cdot \|^2 - 2\varphi^*$ where φ is optimal for (SD).*

Proof. □

what else is needed to be proved?

Theorem 15.1.5 (Fundamental theorem of optimal transport)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be two probability measures such that μ has a density and let $X \sim \mu$. Then the following are equivalent:*

(i) $\bar{\gamma} \in \Gamma_{\mu,\nu}$ is an optimal coupling in the sense that:

$$\int \|x - y\|^2 \, \bar{\gamma}(\mathrm{d}x, \mathrm{d}y) = W_2^2(\mu, \nu) \, .$$

(ii) There exists a proper convex function φ such that $(X, \nabla\varphi(X)) \sim \bar{\gamma} \in \Gamma_{\mu,\nu}$.

(iii) Strong duality holds between (W_2^2) and (D-W_2^2):

$$\int \|x - y\|^2 \, \bar{\gamma}(\mathrm{d}x, \mathrm{d}y) = \sup_{\substack{f \in L^1(\mu),\, g \in L^1(\nu) \\ f(x) + g(y) \leq \|x - y\|^2}} \left\{ \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\nu \right\} .$$

Moreover, the above supremum is achieved for

$$\bar{f}(x) \overset{\mathrm{d}}{=} \|x\|^2 - 2\varphi(x) \qquad \text{and} \qquad \bar{g}(y) \overset{\mathrm{d}}{=} \|y\|^2 - 2\varphi^*(y) \, .$$

With the fundamental theorem, we can state an improved version of Brenier's theorem.

Theorem 15.1.6 (Improved Brenier)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be two probability measures such that μ has a density and let $X \sim \mu$. Then there exists a convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that $(X, \nabla\varphi(X)) \sim \bar{\gamma} \in \Gamma_{\mu,\nu}$ and $\bar{\gamma}$ is an optimal coupling for (W_2^2):*

$$\int \|x - y\|^2 \, \bar{\gamma}(\mathrm{d}x, \mathrm{d}y) = \min_{\gamma \in \Gamma_{\mu,\nu}} \int \|x - y\|^2 \, \gamma(\mathrm{d}x, \mathrm{d}y) = W_2^2(\mu, \nu) \, .$$

Moreover, $\nabla\varphi$ is unique in the sense that if there exists a convex function ψ such that $\nabla\psi(X) \sim \nu$, then $\nabla\psi(X) = \nabla\varphi(X)$, almost surely.

In particular, any valid coupling $\gamma \in \Gamma_{\mu,\nu}$ of the form $(X, \nabla\psi(X)) \sim \gamma$ for some convex function ψ, must be the unique *optimal coupling between μ and ν.*

15.2 Entropic Optimal Transport

The basic principle of entropic optimal transport is to modify the definition of optimal transport to include a penalization term based on the entropy of the coupling, that is, to consider the optimization problem

$$\inf_{\gamma \in \Gamma_{\mu,\nu}} \left\{ \int \|x-y\|^2 \, \gamma(\mathrm{d}x, \mathrm{d}y) - \varepsilon \operatorname{Ent}(\gamma) \right\}, \tag{15.3}$$

where $\operatorname{Ent}(\gamma)$ denotes the differential entropy $\int \gamma(x) \log \frac{1}{\gamma(x)} \, \mathrm{d}x$ for an absolutely continuous probability measure γ.

Define

$$\iota^\varepsilon(f,g) = \varepsilon \iint \left(e^{(f(x)+g(y)-\|x-y\|^2)/\varepsilon} - 1 \right) \mu(\mathrm{d}x)\, \nu(\mathrm{d}y) \,.$$

The function ι^ε is convex and continuous on the space $C_b(\Omega)$ of bounded, continuous functions on Ω.

$$\sup_{f,g \in C_b(\Omega)} \left\{ \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\nu - \iota^\varepsilon(f,g) \right\} \tag{ε-D-W_2^2}$$

$$\inf_{\gamma \in \Gamma_{\mu,\nu}} \left\{ \int \|x-y\|^2 \, \gamma(\mathrm{d}x, \mathrm{d}y) + \varepsilon \operatorname{KL}(\gamma \,\|\, \mu \otimes \nu) \right\}. \tag{ε-W_2^2}$$

15.3 Wasserstein Gradient Flow

$$\dot{X}_t = v_t(X_t) \,. \tag{15.4}$$

Proposition 15.3.1 (Continuity equation)**.** *Suppose that $X_0 \sim \mu_0$, and that $(X_t)_{t\ge 0}$ evolves according to the dynamics (15.4), which we assume is well-posed. Let μ_t denote the law of X_t for all $t \ge 0$. Then, $(\mu_t)_{t \ge 0}$ satisfies the following equation in the weak sense,*

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0 \,, \tag{15.5}$$

i.e., for all compactly supported and smooth test functions $\varphi : \mathbb{R}^d \to \mathbb{R}$, it holds that

$$\partial_t \int \varphi \, \mathrm{d}\mu_t = \int \langle \nabla \varphi, v_t \rangle \, \mathrm{d}\mu_t \,. \tag{15.6}$$

15.4 Otto Calculus

Definition 15.4.1 (First variation)**.** Let $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}$ be a functional. The *first variation* of $\mathcal{F}$ at μ, denoted $\delta\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}$, is the function defined by

$$\lim_{\varepsilon \searrow 0} \frac{\mathcal{F}(\mu + \varepsilon\chi) - \mathcal{F}(\mu)}{\varepsilon} = \int \delta\mathcal{F}(\mu) \, \mathrm{d}\chi \,, \tag{15.7}$$

for all signed measures χ such that $\mu + \varepsilon\chi \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ for all sufficiently small ε.

Proposition 15.4.2. *Let $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}$ be a functional with first variation $\delta\mathcal{F}$. Then, the Wasserstein gradient of $\mathcal{F}$ is the vector field $\mathbb{W}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d$ defined by*

$$\mathbb{W}\mathcal{F}(\mu) = \nabla \delta\mathcal{F}(\mu) \,,$$

where ∇ on the right-hand side denotes the usual Euclidean gradient.

Proof. Let $(\mu_t)_{t\ge 0}$ be a curve of measures with tangent vectors $(v_t)_{t \ge 0}$. The fact that v_t is the tangent vector at time t means that it solves the continuity equation (15.5). Using the definition of the first variation,

$$\begin{aligned} \partial_t \mathcal{F}(\mu_t) &= \int \delta\mathcal{F}(\mu_t) \, \partial_t \mu_t = -\int \delta\mathcal{F}(\mu_t) \operatorname{div}(\mu_t v_t) \\ &= \int \langle \nabla \delta\mathcal{F}(\mu_t), v_t \rangle \, \mathrm{d}\mu_t = \langle \nabla \delta\mathcal{F}(\mu_t), v_t \rangle_{\mu_t} \,. \end{aligned}$$

Moreover, since $\nabla \delta\mathcal{F}(\mu_t)$ is the gradient of a function, from Definition **??** we have $\nabla\delta\mathcal{F}(\mu_t) \in T_{\mu_t}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. From this, we conclude that $\nabla\delta\mathcal{F}(\mu_t)$ is indeed the Wasserstein gradient of $\mathcal{F}$ at μ_t. □