IOE 516 class Note

Xuyuan Zhang

Update on March 11, 2024

Contents

List of Theorems

List of Definitions

Part I

Baisc knowledge

Chapter 1

Law of Large Numbers

1.1 Metric Space

Definition 1.1 (metric space)**.** A metric space is a pair (S, ρ) of a set S and a function $\rho : S \times S \to \mathbb{R}_+$ such that for all $x, y, z \in S$ the following holds:

- $\rho(x, y) = 0$ if and only if $x = y$
- $\rho(x, y) = \rho(y, x)$ (symmetry)
- $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$

Definition 1.2 (Cauchy Sequence)**.** A sequence $x_n \in S$ is said to converge to a limit $x \in S$ ($x_n \to x$) if $\lim_n \rho(x_n, x) = 0$. A sequence $x_n \in S$ is Cauchy if for every $\varepsilon > 0$ there exists n_0 such that for all $n, n' > n_0, \rho(x_n, x'_n) < \varepsilon$. A metric space is defined to be complete if every Cauchy sequence converges to some limit x.

The space $\mathbb{R}^d$ is a complete space under all three metrics $\mathbb{L}_1, \mathbb{L}_2, \mathbb{L}_\infty$. The space $\mathbb{Q}$ is not complete. a subset $A \subset S$ is called dense if for every $x \in S$ there exists a sequence of points $x_n \in A$ such that $x_n \to x$. The set of rational values in $\mathbb{R}$ is dense and is countable. The set of irrational points in $\mathbb{R}$ is dense but not countable. The set of points $(q_1, \ldots, q_d) \in \mathbb{R}^d$ such that q_i is rational for all $1 \leq i \leq d$ is a countable dense subset of $\mathbb{R}^d$.

Definition 1.3 (separable metric space)**.** A metric space is defined to be separable if it contains a dense countable subset A. A metric space is defined to be a Polish space if it is complete and separable.

$\mathbb{R}^d$ is Polish. Given $x \in S$ and $r > 0$ define a ball with radius r to be $B(x, r) = \{y \in S : \rho(x, y) \leq r\}$. A set $A \subset S$ is defined to be open if for every $x \in A$ there exists: ε such that $(B, \varepsilon) \subset A$. A set A is defined to be closed if $A^c = S \backslash A$ is open. The empty space is both open and closed. Every open interval (a, b) is and every countable union of intervals $\bigcup_{i \geq 1}(a_i, b_i)$ is an open set.

For every set A define its interior A^o as the union of all open sets $U \subset A$. This set is open. For every set A define its closure $\bar{A}$ as the intersection of all closed sets $V \supset A$. This set is closed. For every A define its boundary ∂A as $\bar{A} \backslash A^o$. A set $K \subset S$ is defined to be compact if every sequence $x_n \in K$ contains a converging subsequence $x_{n_k} \to x$ and $x \in K$. It can be shown that $K \subset \mathbb{R}^d$ is compact if and only if K is closed and bounded.

Proposition 1.1 (Heine-Borel Theorem)**.** *Given a metric space (S, ρ) a set K is compact iff every cover of K by open sets contains a finite subcover. Namely, if $U_r, r \in \mathbb{R}$ is a (possibly uncountable) family of sets such that $K \subset \bigcup_r U_r$, then there exists a finite subset $r_1, \ldots, r_m \in \mathbb{R}$ such that $K \subset \bigcup_{1 \leq i \leq m} U_r$.*

Proof. **Compact $\implies$ Closed and Bounded**

Assume K is compact, and we know that a set is closed if its complement is open. Take an arbitrary point $x \notin K$. For each $y \in K$. Since x and y are distinct, there is an $\varepsilon_y > 0$ such that the open balls $B(x, \frac{\varepsilon}{2})$ and $B(y, \frac{\varepsilon_y}{2})$ are disjoint. The collection of all $B(y, \frac{\varepsilon_y}{2})$ for $y \in K$ forms an open cover of K. By compactness, there exists a finite subcover. The corresponding ball $B(x, \frac{\varepsilon_y}{2})$ for this finite subcover will form an open set around x that does not intersect with K, showing that K^c is open, and thus K is closed.

Assume K is compact, suppose for contradiction that K is not bounded. Then for every $n \in \mathbb{N}$, there exists an $x_n \in K$ such that $\rho(x_n, x) > n$ for some fixed $x_0 \in K$. The collection of open balls $\{B(x_0, n)\}$ for $n \in \mathbb{N}$ forms an open cover of K, but no finite subcollection of the balls can cover K, contradicting the assumption that K is compact.

Closed and Bounded $\implies$ Compact

If $K \subset \mathbb{R}^d$, we use the Bolzano-Weierstrass theorem which states that every bounded sequence in $\mathbb{R}^d$ has a convergent subsequence. Then to prove that every sequence has a convergent subsequence. Take any sequence $\{x_n\}$ in K and since K is bounded, so is $\{x_n\}$. By Bolzano-Weierstrass, there exists a convergent subsequence $\{x_{n_k}\}$ of $\{x_n\}$. Since K is closed, the limit of this convergent subsequence lies in K. Other details are trivial. □

Definition 1.4 (continuous metric space). Given two metric spaces $(S_1, \rho_1), (S_2, \rho_2)$ a mapping $f : S_1 \to S_2$ is defined to be continuous in $x \in S_1$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $f(B(x, \delta)) \subset B(f(x), \varepsilon)$. Equivalence for every y such that $\rho_1(x, y) < \delta$ we must have $\rho_2(f(x), f(y)) < \varepsilon$. And again, equivalently, if for every sequence $x_n \in S_1$ converging to $x \in S$ it is also true that $f(x_n)$ converges to $f(x)$.

A mapping f is defined to be continuous if it is continuous in every $x \in S_1$. A mapping is uniformly continuous if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\rho_1(x, y) < \delta$ implies $\rho_2(f(x), f(y)) < \varepsilon$.

Proposition 1.2 (compact; continuous $\implies$ uniform continuous). *Suppose $K \subset S_1$ is compact. If $f : K \to \mathbb{R}^d$ is continuous then it is also uniformly continuous. Also there exists $x_0 \in K$ satisfying $\|f(x_0)\| = \sup_{x\in K} \|f(x)\|$, for every norm $\|\cdot\| = \|\cdot\|_p$.*

Proof. Fix $\varepsilon > 0$, for every $x \in K$ find $\delta = \delta(x)$ such that $f(B(x, \delta(x))) \subset B(f(x), \varepsilon)$. This possibly by continuity of f. Then $K \subset \bigcup_{x\in K} B_o(x, \delta(x)/2)$ Namely, we have an open cover of K. By Proposition 1, there exists a finite subcover $K \subset \bigcup_{1\leq i\leq k} B_o(x_i, \delta(x_i)/2)$. Let $\delta = \min_{1\leq i\leq k} \delta(x_i)$. This value is positive since k is finite. Consider any two points $y, z \in K$ such that $\rho_1(y, z) < \delta/2$. We just showed that there exists $i, 1 \leq i \leq k$ such that $\rho_1(x_i, y) \leq \delta(x_i)/2$. By triangle inequality $\rho_1(x_i, z) < \delta(x_i)/2 + \delta/2 \leq \delta(x_i)$. Namely, both y and z belong to $B_o(x_i, \delta(x_i))$. Then $f(y), f(z) \in B_o(f(x_i), \varepsilon)$. By triangle inequality we have: $\|f(y) - f(z)\| \leq \|f(y) - f(x_i)\| + \|f(x_i) - f(z)\| < 2\varepsilon$. We conclude that every two points y, z such that $\rho(y, z) < \delta/2$ we have $\|f(y) - f(z)\| < 2\varepsilon$. The uniform continuity is established. Notice that in this proof that only property of the target space $\mathbb{R}^d$ we used is that it is a metric space.

Now let us show that the existence of $x_o \in K$ satisfying $\|f(x_o)\| = \sup_{x\in K} \|f(x)\|$ First let us show that $\sup_{x\in K} \|f(x)\| < \infty$. If this is not true, identify a sequence $x_n \in K$ such that $\|f(x_n)\| \to \infty$. Since K is compact, there exists a subsequence x_{n_k} which converges to some point $y \in K$. Since f is continuous, then $f(x_{n_k}) \to f(y)$, but this contradicts $\|f(x_n)\| \to \infty$. Thus, $\sup_{x\in K} \|f(x)\| < \infty$. Find a sequence x_n satisfying $\lim_n \|f(x_n)\| = \sup_{x\in K} \|f(x)\|$. Since K is compact, there exists a converging subsequence $x_{n_k} \to x_0$. Again using continuity of f we have $f(x_{n_k}) \to f(x_0)$. But $\|f(x_{n_k})\| \to \sup_{k\in K} \|f(x)\|$. We conclude that $f(x_0) = \sup_{x\in K} \|f(x)\|$. □

Given $x \in C[0, T]$ and $\delta > 0$, define $w_x(\delta) = \sup_{s,t:|s-t|<\delta} |x(t) - x(s)|$. The quantity $w_x(\delta)$ is called modulus of continuity. Since $[0, T]$ is compact, then by Proposition 2 every $x \in C[0, T]$ is uniformly continuous on $[0, T]$. This may be restated as for $\varepsilon > 0$, there exists $\delta > 0$ such that $w_x(\delta) < \varepsilon$.

Theorem 1.1 (Arezela-Ascoli Theorem). *A set $A \subset C[0,T]$ is compact if and only if it is closed and*

$$\sup_{x \in A} |x(0)| < \infty, \tag{1.1}$$

and

$$\lim_{\delta \to 0} \sup_{x \in A} w_x(\delta) = 0 \tag{1.2}$$

Proof. We only show that if A is compact then 1.1 and 1.2 holds. The converse is established similarly.

We already know that if A is compact, it needs to be closed. The assertion 1.1 follows from Proposition 2. We now show 1.2. For $s, t \in [0,T]$ we have:

$$|y(t) - y(s)| \leq |y(t) - x(t)| + |x(t) - x(s)| + |x(s) - y(s)| \leq |x(t) - x(s)| + 2\|x - y\|$$

Similarly, we show that $|x(t) - x(s)| \leq |y(t) - y(s)| + 2\|x - y\|$. Therefore, for every $\delta > 0$.

$$|w_x(\delta) - w_y(\delta)| < 2\|x - y\| \tag{1.3}$$

1.2 is equivalent to:

$$\lim_n \sup_{x \in A} w_x(\frac{1}{n}) = 0. \tag{1.4}$$

Suppose A is compact but 1.4 does not hold. Then we can find a subsequence $x_{n_i} \in A, i \geq 1$ such that $w_{x_{n_i}}(1/n_i) \geq c$ for some $c > 0$. Since A is compact, then there is further subsequence of x_{n_i}, which converges to $x \in A$. To ease the notation we denote the subsequence again by x_{n_i}. Thus $\|x_{n_i} - x\| \to 0$. From 1.3 we obtain:

$$\left| w_x(1/n_i) - w_{x_{n_i}}(1/n_i) \right| < 2\|x - x_{n_i}\| < c/2$$

for all i larger than some i_0. This implies that:

$$w_x(1/n_i) \geq c/2 \tag{1.5}$$

for all sufficiently large i. But x is continuous on $[0,T]$, which implies it is uniformly continuous, as $[0,T]$ is compact. This contradicts 1.5. □

1.2 Convergence of mappings

Given two metric spaces $(S_1, \rho_1), (S_2, \rho_2)$ a sequence of mappings $f_n : S_1 \to S_2$ is defined to be point-wise converging to $f : S_1 \to S_2$ if for every $x \in S_1$ we have $\rho_2(f_n(x), f(x)) \to 0$. A sequence f_n is defined to converge to f uniformly if:

$$\lim_n \sup_{x \in S_1} \rho_2(f_n(x), f(x)) = 0$$

Also given $K \subset S_1$, sequence f_n is said to converge to f uniformly on K if restriction of f_n, f onto K gives a uniform convergence. A sequence f_n is said to converge to f uniformly on compact sets $u.o.c.$ if f_n converges uniformly to f on every compact set $K \subset S_1$.

Point-wise convergence does not imply uniform convergence even on compact sets. Moreover, if f_n is continuous and f_n converges to f point-wise, this does not imply in general that f is continuous.

Proposition 1.3 (continuous uniform convergence $\implies$ continuous). *Suppose $f_n : S_1 \to S_2$ is a sequence of continuous mappings which converges uniformly to f. Then f is continuous as well.*

Proof. Fix $x \in S_1, \varepsilon > 0$. There exists n_0 such that for all $n > n_0$, $\sup_z \rho_2(f_n(z), f(z)) < \varepsilon/3$. Fix any such $n > n_0$. Since by assumption f_n is continuous, then there exists $\delta > 0$ such that $\rho_2(f_n(x), f_n(y)) < \varepsilon/3$ for all $y \in B_o(x, \delta)$ .Then for any y we have:

$$\rho_2(f(x), f(y)) \leq \rho_2(f(x), f_n(x)) + \rho_2(f_n(x), f_n(y)) + \rho_2(f_n(y), f(x)) < \varepsilon 3/3 = \varepsilon.$$

This proves continuity of f. □

Theorem 1.2 (Polish space). *The spaces $C[0, T], C[0, \infty]$ are Polish.*

1.3 Skorohod space and Skorohod metric

Definition 1.5 (skorohod metric). Let ∇ be the space of strictly increasing continuous functions λ from $[0, T]$ onto $[0, T]$. A skorohod metric on $D[0, T]$ is defined by:

$$\rho_s(x, y) \inf_{\lambda \in \nabla}(\|\lambda - I\| \cup \|x - y\lambda\|)$$

for all $x, y \in D[0, T]$, where $I \in \nabla$ is the identity transformation, and $\|\cdot\|$ is the uniform metric on $D[0, T]$.

Thus, per this definition, the distance between x and y is less than ε if there exists $\lambda \in \Lambda$ such that $\sup_{0 \leq t \leq T} |\lambda(t) - t| < \varepsilon$ and $\sup_{0 \leq t \leq T} |x(t) - y(\lambda(t))| < \varepsilon$.

Proposition 1.4. *he Skorohod metric and uniform metric are equivalent on $C[0, T]$ in a sense that for $x_n, x \in C[0, T]$, the convergence $x_n \to x$ holds under skorohod metric if and only if it holds under the uniform metric.*

1.4 Probability Space

Probability space is defined via $(\Omega, \mathcal{F}, P)$ - Ω is the sample space, $\mathcal{F}$ is the set of events, $P(A)$ is for each event $A \in \mathcal{F}$.

$0 \leq P(A) \leq 1$ for any $A \in \mathcal{F}$, $P(\Omega) = 1$. If $A_1, A_2, \ldots$ is a sequence of mutually exclusive events, then:

$$P(\bigcup_{i=1}^{\infty} A_n) = \sum_{i=1}^{\infty} P(A_i)$$

$\mathcal{F}$ can be the collection of all subsets of Ω. if Ω is not countable, event can be difficult to define. Then so called σ-algebra condition is a formal way to define all the ”measurable events”.

A σ-algebra $\mathcal{F}$ is a collection of subsets of Ω satisfying the conditions:

- $\Omega \in \mathcal{F}$
- If $A \in \mathcal{F}$ then $A^C \in \mathcal{F}$
- If a sequence $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_i A_i \in \mathcal{F}$.

Continuity of probability:

If A_n is increasing, i.e., $A_1 \subset A_2 \subset \ldots$, then we have:

$$\bigcup_n A_n = \lim_n A_n.$$

Similarly, if A_n is decreasing, then $\bigcap_n A_n = \lim_n A_n$.

If either A_n is increasing or decreasing, we have: $P(\lim_n A_n) = \lim_n P(A_n)$.

Proof. Suppose A_n is increasing, and we define $B_n = A_n \backslash A_{n-1}$. Then: $\bigcup_n A_n = \bigcup_n B_n$. We need to show that $\bigcup_n A_n = \bigcup_n B_n$ and then we use this to establish $P(\lim_n A_n) = \lim_n P(A_n)$.

First, we prove $\bigcup_n A_n = \bigcup_n B_n$:

Since $A_k = B_1 \cup B_2 \cup \ldots \cup B_k$, then each A_k is in $\bigcup_k B_k$. Simultaneously, $\bigcup_k B_k \subseteq \bigcup_k A_k$ holds. Therefore, $\bigcup_n A_n = \bigcup_n B_n$.

Now we establish $P(\lim_n A_n) = \lim_n P(A_n)$.

Since A_n is increasing, $\lim_n A_n = \bigcup_n A_n$. By the countable additivity of probability measures, we have $P(\bigcup_n B_n) = \sum_n P(B_n)$. Note that $P(B_n) = P(A_n) - P(A_{n-1})$, because B_n and A_{n-1} are disjoint, and $B_n \cup A_{n-1} = A_n$.

we can write the sum as:

$$\begin{aligned}\sum_n P(B_n) &= P(B_1) + \sum_{n=2}^{\infty} P(B_n) \\ &= P(A_1) + \sum_{n=2}^{\infty} (P(A_n) - P(A_{n-1})) \\ &= \lim_n P(A_n).\end{aligned}$$

Therefore, $P(\lim_n A_n) = P(\bigcup_n A_n) = P(\bigcup_n B_n) = \lim_n P(A_n)$. □

Let $(\Omega, \mathcal{F}, P)$ be the probability space. A random variable, say X, is a variable whose value depends on the outcome $\omega \in \Omega$. So formally, we can also write it as $X(\omega), \omega \in \Omega$. $X(\omega) : \Omega \to \mathbb{R}$.

1.5 Convergence

Definition 1.6 (Convergence in probability). X_n is said to converge to X in probability if, for any $\varepsilon > 0$, it holds that:

$$P(|X_n - X| > \varepsilon) \to 0, \text{ as } n \to \infty$$

Definition 1.7 (Almost sure convergence). X_n is said to converge to X almost surely if:

$$P(\omega : X_n(\omega) \to X(\omega)) = 1.$$

Definition 1.8 (Convergence in L^p). X_n is said to converge to X in L^p if:

$$E[|X_n - X|^p] = 0 \text{ as } n \to \infty$$

Definition 1.9 (Convergence in distribution). Suppose X_n has cdf F_n and X has cdf F. X_n is said to converge to X in distribution if,

$$F_n(x) \to F(x) \text{ as } n \to \infty$$

on every point x at which $F(x)$ is continuous.

1.6 Limit Theorems

Proposition 1.5 (Markov inequality). *For nonnegative random variable X, for any $a > 0$, we have:*

$$P(X > a) \leq \frac{E[X]}{a}.$$

Proof. Suppose it has pdf $f(x)$, and then: $E[X] = \int_0^\infty xf(x)dx \geq \int_a^\infty xf(x)dx \geq a\int_a^\infty f(x)dx = aP(X > a)$. thus: $P(X > a) \leq \frac{E[X]}{a}$. □

Proposition 1.6 (Chebyshev inequality). *Let X be a random variable with finite μ and variance σ^2, then for any $\varepsilon > 0$,*

$$P(|X - \mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

Proof. We can directly implement the Markov inequality by the random variable $Y = (X - \mu)^2$ with $a = (k\sigma)^2$.

Another proof is that:

Let X be a random variable and $a \in \mathbb{R}^+$. we assume X has density function f_X. Then:

$$\begin{aligned} E(X^2) &= \int_{\mathbb{R}} x^2 f_X(x)dx \\ &\geq \int_{|x|\geq a} x^2 f_X(x)dx \\ &\geq a^2 \int_{|x|\geq a} f_X(x)dx = a^2 P(|X| \geq a) \end{aligned}$$

We can generalize the moment $p > 0$:

$$E(|X|^p) \geq a^p P(|X| \geq a)$$

□

Proposition 1.7 (Weakly Law of Large Numbers (WLLN)). *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2. Then:*

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

is called the sample mean. then $\bar{X}_n$ converges to the mean μ in probability, i.e., for any $\varepsilon > 0$, we have:

$$P(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| > \varepsilon) \to 0, \text{ as } n \to \infty.$$

Proof. $E[\bar{X}_n] = \mu$, and $Var(\bar{X}_n) = \sigma^2/n$, and we can apply the Chebyshev inequality. □

Proposition 1.8 (Strong Law of Large Numbers (SLLN)). *$\bar{X}_n$ converges to the mean μ almost surely. That is, for almost every sample ω, we have:*

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \to \mu, \text{ as } n \to \infty$$

The proof of SLLN relies on Borel-Cantelli lemma Let $A_1, A_2, \ldots$ be an infinite sequence of events in Ω. Consider the sequence of events,

$$\bigcup_{n=1}^{\infty} A_n, \bigcup_{n=2}^{\infty} A_n, \bigcup_{n=3}^{\infty} A_n, \ldots$$

Observe that this is a decreasing sequence in the sense that:

$$\bigcup_{n=m+1}^{\infty} A_n \subseteq \bigcup_{n=m}^{\infty} A_n$$

for all $m = 1, 2, \ldots$. We are interested in those events ω that lie in infinitely many A_n. Such ω would lie in $\bigcup_{m=n}^{\infty}$ for every m. Thus we define:

$$\limsup A_n = \lim_{n\to\infty} \bigcup_{m=n}^{\infty} A_m = \omega \text{ that are in infinitely many } A_n$$

We write this event as:

$$\limsup A_n = \{\omega \in A_n, i.o.\}$$

where $i.o.$ is ”infinitely often”. We can state the Borel Cantelli Lemma as:

Lemma 1.1 (Borel-Cantelli Lemma)**.** *If* $\sum_{n=1}^{\infty} P(A_n) < \infty$ *then* $P(\omega \in A_n, i.o.) = 0$.

Proof. First observe that:

$$0 \leq \limsup A_n \subseteq \bigcup_{m=n}^{\infty} A_n$$

for every m since the sequence is a decreasing sequence of events. Then:

$$0 \leq P(\limsup A_n) \leq \sum_{n=m}^{\infty} P(A_n)$$

for every m. But we are assuming that the series $\sum_{n=1}^{\infty} P(A_n)$ converges. This means that:

$$\sum_{n=m}^{\infty} P(A_n) \to 0 \text{ as } m \to \infty.$$

Taking $m \to \infty$ and we have: $P(\limsup A_n) = 0$. □

the strong law of larger number ask the question in what sense can we say: $\lim_{n\to\infty} \frac{S_n(\omega)}{n} = \mu$.

Proof. We will prove this under the addditional restriction that $\sigma^2 = E(X_j^2) < \infty, E(X_j^4) < \infty$. It is without loss of generality to assume $\mu = 0$, now if:

$$\lim_{n\to\infty} \frac{S_n(\omega)}{n} \neq 0$$

then there exists $\varepsilon > 0$ such that for infinitely many n,

$$\left| \frac{S_n(\omega)}{n} \right| > \varepsilon$$

Thus to prove the theorem we prove that for every $\varepsilon > 0$,

$$P(|S_n| > n\varepsilon i.o) = 0$$

This then shows that:

$$P(\xi) = P(\frac{S_n}{n} = 0) = 1.$$

we use the Borel-Cantelli lemma applied to the events:

$$A_n = \{\omega \in \Omega : |S_n| \geq n\varepsilon\}.$$

To estimate $P(A_n)$ we use the generalized Chebyshev inequality with $p = 4$, and thus we must compute $E(S_n^4)$ which equals:

$$E\left(\sum_{1\leq i,j,k,l\leq n} X_i X_j X_k X_l\right)$$

when we sums are multiplied out there will be terms of the form:

$$E(X_i^3 X_j), E(X_i^2 X_j X_k), E(X_i X_j X_k X_l)$$

with i, j, k, l all distinct. These terms are all equal to zero since $E(X_i) = 0$ and the random variables are independent. Thus the nonzero terms in the above sums are:

$$E(X_i^4) \text{ and } E(X_i^2 X_j^2) = (E(X_i^2))^2$$

These are n terms of the form: $E(X_i^4)$. The number of terms of the form $E(X_i^2 X_j^2)$ is $3n(n-1)$. Thus we can show:

$$E(S_n^4) = nE(X_1^4) + 3n(n-1)\sigma^2.$$

For n sufficiently large there exists a constant C such that:

$$3\sigma^4 n^2 + (E(X_1^4) - 3\sigma^4)n \leq Cn^2.$$

That is,

$$E(S_n^4) \leq Cn^2.$$

Then the Chebyshev inequality together with that gives:

$$P(|S_n| \geq n\varepsilon) \leq \frac{1}{(n\varepsilon)^4} E(S_n^4) \leq \frac{C}{\varepsilon^4 n^2}$$

Thus,

$$\sum_{n\geq n_0} P(|S_n| \geq n\varepsilon) \leq \sum_{n\geq n_0} \frac{C}{\varepsilon^4 n^2} < \infty$$

Thus by the Borel-Cantelli lemma we have $P(|S_n| \geq n\varepsilon, i.o.) = 0$. since this holds for every $\varepsilon > 0$ we have proved the strong LLN. □

Strong law of large number shows that:

$$\frac{X_1 + \ldots + X_n - nE[X_1]}{n}$$

converges to zero almost surely.

Definition 1.10 (Moment Generating Function (MGF)). A moment generating function (MGF) of a random variable X is defined as, for any $\theta \in \mathbb{R}$,

$$M_X(\theta) = E[e^{\theta X}]$$

Claim 1.1. *If the MGF of a random variable X is finite in a small neighborhood of 0, then for any positive integer n,*

$$E[X^n] = M_X^{(n)}(\theta) \mid_{\theta=0}$$

Proof.

$$M_X(\theta) = E[e^{\theta X}] = 1 + \theta E[X] + \frac{(\theta E[X])^2}{2!} + \ldots$$

Therefore, we take the n-th derivative of $M_X(\theta)$ with respect to θ and evaluate it at $\theta = 0$: $M_X^{(n)}(0) = \frac{d^n}{d\theta^n} M_X(\theta) \mid_{\theta=0}$ □

The previous result indicates that we need MGF to be finite near 0, and this, however, may not be satisfied.

Example 1.1. Let X be lognormal $X = e^Z$, where Z is standard normal. What is the n-th moment?

$$E[X^n] = E[e^{nZ}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{nz} e^{-\frac{z^2}{2}} dz = e^{n^2/2}$$

However, by change of variable $y = e^z$, we have for any $\theta > 0$,

$$M_X(\theta) = E[e^{\theta X}] = E[e^{\theta e^Z}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\theta e^z} e^{-z^2/2} dz = \int_0^{\infty} \frac{1}{y} e^{\theta y} e^{-(\log(y))^2} dy = \infty.$$

The characteristic function: for any random variable X, its characteristic function is defined:

$$\phi_X(\theta) = E[e^{i\theta X}]$$

Some properties:

- $\phi_X(0) = 1$
- $|\phi_X(\theta)| \leq 1, \forall \theta$
- $\phi_X(\theta)$ is continuous, and
- $\phi_X^{(n)}(\theta) \mid_{\theta=0} = i^n E[X^n]$

Why MGF and/or characteristic function?
This is the time domain versus frequency domain analysis. Using the transform, it allows us to analyze a problem that is otherwise difficult to study in time domain.
This is just like we use Laplace transform and z-transform.

Theorem 1.3 (Levy's Convergence Theorem). *Let $\{X_n\}$ be a sequence of $r.v.$'s with cdf's F_n and characteristic functions ϕ_n. Let X be a random variable with cdf F and characteristic function ϕ that is continuous at $\theta = 0$. Then X_n converge to X in distribution if and only if:*

$$\phi_n(\theta) \to \phi(\theta) \text{ as } n \to \infty, \forall \theta.$$

Theorem 1.4 (Levy's Continuity' Theorem (Simple form)). $X^n \xrightarrow{d} X$ *iff* $\phi_{X^{(n)}}(t) \xrightarrow{d} \phi_X(t), \forall t \in \mathbb{R}^d$.

Proof. We assume that $X^{(n)} \xrightarrow{d} X$. Since $\exp\left(it^T X\right) = \cos\left(t^T X\right) + i\sin\left(t^T X\right)$, we have ϕ is continuous and bounded as a function of X, which together with implication, we have pointwise convergence(convergence almost surely) of the characteristic function.

Conversely, we assume that $\forall t \in \mathbb{R}^d, \phi_{X^{(n)}}(t) \to \phi_X(t)$ and show that for any continuous function g that is zero outside a bounded and closed set, we have $E(g(X^{(n)})) \to E(g(X))$. Using the Portmanteau theorem, this implies that $X^{(n)} \xrightarrow{d} X$. □

Theorem 1.5 (Central Limit Theorem (CLT)). *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. $r.v.$'s with mean μ and $\sigma^2 < \infty$, then:*

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0,1)$$

WLLN and SLLN state that $S_n = \sum_{i=1}^n X_i$ is approximately equal to $S_n \approx n\mu$. CLT claims that $S_n \approx n\mu + \sqrt{n}\sigma N(0,1)$.

By Levy's theorem, it suffices to show that the characteristic function of

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

converges to that standard normal when $n \to \infty$. WLOG, assume $E[X_1] = 0$, we will have $Var(X_1) = \sigma^2$. Let $\phi(\theta) = E[e^{i\theta X_1}]$. Then:

$$E[e^{i\theta\frac{S_n}{\sqrt{n}}}] = (\phi(\frac{\theta}{\sqrt{n}}))^n.$$

Since $\theta/\sqrt{n}$ is small with n is large, we have: $\phi(\frac{\theta}{\sqrt{n}}) = \phi(0) + \phi'(0)\frac{\theta}{\sqrt{n}} + o(\frac{\theta}{\sqrt{n}})$ and therefore,

$$E[e^{i\theta\frac{S_n}{\sqrt{n}}}] = (\phi(\theta/\sqrt{n}))^n = \ldots.$$

Part II

Generation via martingale

Chapter 2

deviance and concentration inequality

Preliminary notes

The goal of large deviation theory is to show that in many interesting cases the decay rate is in fact exponential e^{-cn}. The exponent c is called the large deviation rate, and in many cases it can be computed explicitly or numerically.

2.1 Large deviations upper bound (Chernoff bound)

Consider an i.i.d. sequence with a common probability distribution function $F(x) = P(X \leq x), x \in \mathbb{R}$. Fix a value $a > \mu$, where μ is an expectation corresponding to the distribution F. the WLLN tells us that this hapens with probability converging to zero as n increases, and now we obtain an estimate on this probability. Fix a positive parameter $\theta > 0$. We have:

$$\begin{aligned} P(\frac{\sum_{1\leq i\leq n} X_i}{n} > a) = P(\sum_{1\leq i\leq n} X_i > na) = P(e^{\theta \sum_{1\leq i\leq n} X_i} > e^{\theta na}) \\ \leq \frac{E[e^{\theta \sum_{1\leq i\leq n} X_i}]}{e^{\theta na}} = \frac{E[\Pi_i e^{\theta X_i}]}{(e^{\theta a})^n} \end{aligned}$$

but recall that X_i's are i.i.d. Therefore, $E[\Pi_i e^{\theta X_i}] = (E[e^{\theta X_1}])^n$. Thus we obtain an upper bound:

$$P(\sum_{1\leq i\leq n X_i} n > a) \leq (\frac{E[e^{\theta X_1}]}{e^{\theta a}})^n$$

This bound is meaningful if the ratio $E[e^{\theta X_1}]/e^{\theta a}$ is less than unity. We recognize $E[e^{\theta X_1}]$ as the moment generating function of X_1 and denote it by $M(\theta)$. For the bound to be useful, we need $E[e^{\theta X_1}]$ to be at least finite. If we could show that this ratio is less than unity, we would be done. Similarly, suppose we want to estimate:

$$P(\frac{\sum_{1\leq i\leq n} X_i}{n} < a)$$

for some $a < \mu$. Fixing now a negative $\theta < 0$ we obtain:

$$P(\frac{\sum_{1\leq i\leq n} X_i}{n} < a) = P(e^{\theta \sum_{1\leq i\leq n} X_i} > e^{\theta na}) \leq (\frac{M(\theta)}{e^{\theta a}})^n,$$

and now we need to find a negative θ such that $M(\theta) < e^{\theta a}$. In particular, we need to focus on θ for which the moment generating function is finite. For this purpose let $D(M) \equiv \{\theta : M(\theta) < \infty\}$ Thus we call D the domain of M.

Before we proceed, we give some examples about moment generating functions:

Example 2.1. Poisson distribution: Suppose X has a Poisson distribution with parameter λ, then:

$$M(\theta) = E[e^{\theta X}] = \sum_{m=0}^{\infty} e^{\theta m}\frac{\lambda^m}{m!}e^{-\lambda} = \sum_{m=0}^{\infty} \frac{(e^{\theta}\lambda)^m}{m!}e^{-\lambda} = e^{\lambda(e^{\theta}-1)}$$

where $\sum_{m\geq 0}\frac{t^m}{m!} = e^t$.

2.2 Implications of CLT

One application of CLT is in computing tail probabilities of sum of random variables. Let Φ be the cdf of standard normal.

Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and variance σ^2, and $\sum_{i=1}^{n} X_i$. by CLT, $(S_n - n\mu)/(\sigma\sqrt{n})$ is approximately standard normal when n is large. Thus we can estimate the probability for $S_n > x$:

$$P(S_n > x) = P(\frac{S_n - n\mu}{\sigma\sqrt{n}} > \frac{x - n\mu}{\sigma\sqrt{n}}) \approx 1 - \Phi(\frac{x - n\mu}{\sigma\sqrt{n}})$$

Example 2.2. 10,000 consumers. premium is \$800, The claim for each customer over the year is random and historical data show that the mean is \$700 and the standard deviation is \$500. What is the probability that the total claim for the year exceeds \$1MM from these consumers.

Let X_i denote the net contribution of consumer i which is the premium minus claim, then it has mean $\mu = 100, \sigma = 500$. $S = \sum_{i=1}^{10000} X_i$. By the result from previous:

$$P(S > 1000000) = 1 - \Phi(\frac{1000000 - 10,000 \times 100}{500 \times \sqrt{10000}}) = 1 - \Phi(0) = 0.5$$

Ordinary deviation: CLT implies that, for large n we can estimate $P(S_n - n\mu > \sqrt{n}a)$ as follows:

$$P(S_n - n\mu > \sqrt{n}a) \approx P(Z > a/\sigma) = 1 - \Phi(a/\sigma)$$

Deviations of S_n from its mean by the order of $\sqrt{n}$ is known as ordinary deviation. Thus, the probability for ordinary deviation is can be easily estimated.

We need some property of GMF $M_X(\theta)$. we focus on $\theta \geq 0$ and recall that $M_X(0) = 1$ for any X. we already know that, for some $r.v.$'s, their MGF $M_X(\theta)$ is infinity for any $\theta > 0$. It can be argued that, if there exists $\theta_0 > 0$ such that $M_X(\theta_0)$ if finite, then $M_X(\theta) < \infty$ for all $0 \leq \theta < \theta_0$. This shows that the range of parameter θ such that $M_X(\theta) < \infty$ is an interval containing 0.

Recall that $S_n = \sum_{i=1}^{n} X_i$. Large deviation is concerned with the event that S_n deviates from its mean by the order n. That is, what is $P(S_n - n\mu > na)$ for $a > 0$.

The probability should be small for moderate or large n, but how small and how to evaluate that?
there is an entire subject area in probability, known as Large Deviation Theory.
Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and variance σ^2. MGF $M(\theta) = E[e^{\theta X_1}]$, assuming finite near 0.
Properties of $M(\theta)$: (i) $M(0) = 1$ and (ii) $M'(0) = \mu$.

Definition 2.1 (Chernoff bound). let $s = \mu + a$. For $\theta > 0$, we have:

$$\begin{aligned} P(S_n \geq ns) = P(e^{\theta S_n} > e^{\theta ns}) &\leq \frac{E[e^{\theta S_n}]}{e^{\theta na}} \\ &= e^{-\theta ns}(E[e^{\theta X_1}])^n = e^{-n(\theta s - \log M(\theta))} \end{aligned}$$

$\theta s - \log M(\theta)$ is 0 when $\theta = 0$. Assuming finite on $[0, \theta_0)$, then:

$$(\theta s - \log M(\theta))' \mid_{\theta^-} = s - \frac{M'(0)}{M(0)} = s - \mu > 0.$$

This confirms that there exists $\theta > 0$ such that $\theta s - \log M(\theta) > 0$. Thus $P(S_n \geq na)$ goes to zero exponentially fast!

Theorem 2.1 (Chernoff bound)**.** *Given an i.i.d. sequence $X_1, \ldots, X_n$ sequence the MGF $M(\theta)$ is finite in some interval $(\theta_1, \theta_2) \ni 0$. Let $a > \mu = E[X_1]$ Then there exists $\theta > 0$ such that $M(\theta)/e^{\theta a} < 1$ and*

$$P(\frac{\sum_{1 \leq i \leq n} X_i}{n} > a) \leq (\frac{M(\theta)}{e^{\theta a}})^n$$

Similarly, if $a < \mu$, then there exists $\theta < 0$ such that $M(\theta)/e^{\theta a} < 1$ and

$$P(\frac{\sum_{1 \leq i \leq n} X_i}{n} < a) \leq (\frac{M(\theta)}{e^{\theta a}})^n$$

How small can we make the ratio $M(\theta)/\exp(\theta a)$? the large deviations theory is very often that such a minimizing value θ^* exists and is tight. Namely it provides the correct decay rate.

Remark 2.1. $P(S_n \geq na)$ goes to zero at exponential rate $\theta s - \log M(\theta)$ when it is greater than 0. The laregst $\theta s - \log M(\theta)$ is the fastest rate. We need to identify that θ that maximizes $\theta s - \log M(\theta)$.

Definition 2.2 (Fenchel-Legendre transform)**.** The Legendre transform of a r.v. is defined as:

$$\nabla^*(s) = \sup_{\theta \geq 0}(\theta s - \log M(\theta)) > 0$$

Properties of $\nabla(\theta) = \log M(\theta)$:

- $\nabla(0) = 0$
- $\nabla'(0) = \mu$
- $\nabla(\theta)$ is convex

Thus $\theta s - \nabla(\theta)$ is a concave function. Let its maximizer be θ^*.

Theorem 2.2 (Cramer-Chernoff Theorem)**.** *By Chernoff bound, we have:*

$$P(S_n \geq sn) \leq e^{-\nabla^*(s)n}.$$

This implies that $\frac{1}{n} \log P(S_n \geq sn) \leq -\nabla^*(s)$. Large deviation theory, known as Cramer-Chernoff Theorem, shows that the upper bound is actually tight. That is,

$$\frac{1}{n} \cdot \log P(S_n \geq sn) \to -\nabla^*(s) \text{ as } n \to \infty.$$

Example 2.3. We use Poisson with parameter λ to illustrate: GMF is:

$$E[e^{\theta X}] = \sum_{n=0}^{\infty} e^{\theta n} e^{-\lambda} \frac{\lambda^n}{n!} = e^{\lambda(e^\theta - 1)}$$

where $\nabla^*(s) = \sup_{\theta \geq 0}\{\theta s - \lambda(e^\theta - 1)\} = s\log(s/\lambda) - s + \lambda$ with $\theta^* = \log(s/\lambda)$.
By LDT, $P(S_n > sn) \approx e^{-(s\log(s/\lambda) - s + \lambda)n}$. The LHS is:

$$P(S_n > ns) = \sum_{k \geq sn} e^{-\lambda n} \frac{(\lambda n)^k}{k!}.$$

This decay rate is hard to assess but LDT has provided the solution.

Remark 2.2. How strong is "MGF is finite near 0"?

It is basically those random variables whose tail function decays at least exponentially fast. This shows that, large deviation result is quite natural and expected.

Exponential decay:

Claim 2.1. *MGF is finite near* 0 *iff its tail is exponentially bounded, i.e., there exists* $\mu > 0, \lambda > 0$ *such that:*

$$P(X > t) \leq \mu e^{-\lambda t}$$

If there exists $\theta_0 > 0$ such that $E[e^{\theta_0 X}] = \mu$ is finite, then by markov inequality:

$$P(X > t) = P(e^{\theta_0 X} > e^{\theta_0 t}) \leq \frac{E[e^{\theta_0}]}{e^{\theta_0 t}} = \mu e^{-\theta_0 t}$$

On the other hand, if $P(X > t) \leq \mu e^{-\lambda t}$, then for any $\theta_0 < \lambda$, we have:

$$E[e^{\theta_0 X}] = -\int_0^\infty e^{\theta_0 t} d\bar{F}(t) = -e^{\theta_0 t} \bar{F}(t) \mid_0^\infty + \theta_0 \int_0^\infty \bar{F}(t) e^{\theta_0 t} dt = \dots$$

LDT considers the probability that S_n deviates from its mean by an. When λ_n lies in between n (large deviation) and $\sqrt{n}$ (ordinary deviation), it is called medium deviation.

In the following we assume that the mean of X_i is zero.

if λ_n grows slower than $n^{2/3}$ but faster than $\sqrt{n}$, then under the condition of finite MGF near 0, we have:

$$P(S_n > \lambda_n) \approx \frac{\sigma}{\lambda_n} \sqrt{\frac{n}{2\pi}} \cdot e^{-\frac{\lambda_n^2}{2\sigma^2 n}}$$

Of particular importance is the special case that $\lambda_n = c\sqrt{\sigma n \log n}$.

$$P(S_n > c\sqrt{\sigma n \log n}) \approx \frac{1}{c\sqrt{\pi \log n}} \frac{1}{n}$$

Example: $c = \sqrt{2}, 2$, we observe:

$$P(S_n > \sqrt{2\sigma n \log n}) \approx \frac{1}{2\sqrt{\pi \log n}} \frac{1}{n}$$

$$P(S_n > 2\sqrt{\sigma n \log n}) \approx \frac{1}{2\sqrt{2\pi \log n}} \frac{1}{n^2}$$

A very useful result

In general, we have for any $\alpha > 0$,

$$P(S_n > \sqrt{2\alpha\sigma n \log n}) \approx \frac{1}{2\sqrt{\alpha\pi \log n}} \frac{1}{n^\alpha}$$

Remark 2.3. the result above does not require MGF be finite near zero. A sufficient condition is $E[|X_1|^{2(\alpha+1)+\delta}] < \infty$ for small $\delta > 0$. this condition is much weaker than those imposed in the operations.

Concentration inequalities:

A random variable concentrates its mass around the mean. How fast does it spread out, especially when the random variable is sum of random variables?

Ordinary deviation, medium deviation, and large deviation are all concentration inequalities. But there are many more.

Together, they play a central role in analyzing learning and approximation algorithms in the operations literature.

Concentration inequalities take the form: $P(|X - E[X]| \geq \varepsilon)$ for any given $\varepsilon > 0$. It usually suffices to study $P(X - E[X] > \varepsilon)$ and it is WLOG to assume that $E[X] = 0$.

Chebyshev inequality is the simplest general result on concentration inequality.

In our results, the upper bounds for $P(|X - E[X]| > \varepsilon)$ is twice of $P(X - E[X] > \varepsilon)$.

Example 2.4. X is normal $N(0, 1)$. Find upper bound for $P(X > \varepsilon)$ for $\varepsilon > 0$.

Applying Chebyshev's inequality, we get:

$$P(|X| \geq k) \leq \frac{1}{k^2} \implies P(X > \varepsilon) = \frac{1}{2}P(|X| \geq \varepsilon) \leq \frac{1}{2}\frac{1}{\varepsilon^2}$$

Sharper result: $P(X > \varepsilon) \leq \frac{e^{-\frac{\varepsilon^2}{2}}}{\varepsilon}$
because:

$$P(X > \varepsilon) = \int_\varepsilon^\infty \phi(s)ds \leq \frac{1}{\varepsilon}\int_\varepsilon^\infty s\phi(s)ds = \frac{e^{-\varepsilon^2/2}}{\sqrt{2\pi}\varepsilon}$$

Example 2.5. X_i is normal $N(0, 1)$ and $S_n = \sum_{i=1}^n X_i$. upper bound for $P(S_n > \varepsilon)$ for $\varepsilon > 0$.

Chebyshev's inequality: Mean of $S_n = 0$ variance of $S_n : \sigma^2_{S_n} = n$.

$$P(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2} \implies P(|S_n - 0| \geq \varepsilon) \leq \frac{1}{\varepsilon^2/n} = \frac{n}{\varepsilon^2}$$

Sharper result:

$$P(S_n > \varepsilon) \leq \frac{\sqrt{n}}{\sqrt{2\pi}\varepsilon}e^{-\varepsilon^2/(2n)}$$

Definition 2.3 (Hoeffding inequality)**.** Suppose $X_1, X_2, \ldots$ are i.i.d., on a bounded support $[a, b]$ and $E[X_1] = \mu$. Then for $\varepsilon > 0$:

$$P(S_n - n\mu > \varepsilon) \leq e^{-\frac{2\varepsilon^2}{n(b-a)^2}}$$

Important special case: Bernoulli with mean p then:

$$P(S_n - np > \varepsilon) \leq e^{-2\varepsilon^2/n}$$

WLOG we assume $\mu = 0$. By Chernoff bound: for any $\theta > 0$,

$$P(S_n > \varepsilon) = P(e^{\theta S_n} > e^{\theta\varepsilon}) \leq e^{-\theta\varepsilon}(E[e^{\theta X_1}])^n = e^{-\theta + n\log E[e^{\theta X_1}]}$$

Now, since X_1 has support on $[a, b]$, we can say more about $E[e^{\theta X_1}]$ and we can show the following:

$$E[e^{\theta X_1}] \leq e^{\theta^2(b-a)^2/8}$$

First, note that, by $X \in [a, b]$, $X = \frac{b-X}{b-a}a + \frac{X-a}{b-a}b$

Since $e^{\theta X}$ is convex in X, we have:

$$e^{\theta X} \leq \frac{b-X}{b-a}e^{\theta a} + \frac{X-a}{b-a}e^{\theta b}$$

Thus:

$$\begin{aligned}
E[e^{\theta X}] &\leq \frac{b}{b-a}e^{\theta a} - \frac{a}{b-a}e^{\theta b} \\
&= e^{\theta a}(\frac{b}{b-a} - \frac{a}{b-a}e^{\theta(b-a)}) \\
&= e^{\theta a + \log\left(1 + \frac{a}{b-a} - \frac{a}{b-a}e^{\theta(b-a)}\right)} \\
&= e^{g(u)}
\end{aligned}$$

where

$$\begin{aligned}
g(u) &= -\gamma u + \log(1-\gamma+\gamma e^u) \\
\gamma &= -\frac{a}{b-a} \\
u &= \theta(b-a).
\end{aligned}$$

We can have $g(0) = g'(0) = 0$ and $g''(x) \leq \frac{1}{4}$ for all $x \geq 0$. By taylor Theorem, we have:

$$g(u) = g(0) + g'(0)u + g''(\xi)\frac{u^2}{2} \leq \frac{u^2}{8} = (b-a)^2\theta^2/8$$

Therefore, we have: $P(X_n \geq \varepsilon) \leq \inf_\theta e^{n(b-a)^2\theta^2/8 - \varepsilon\theta}$ The RHS is minimized when $\theta = 4\varepsilon/n(b-a)^2$ to obtain:

$$P(S_n \geq \varepsilon) \leq e^{-\frac{2\varepsilon^2}{n(b-a)^2}}.$$

Detail proofs can be related to the next section.

2.3 Large deviation theorey. Cramer Theorem

We have established under assumptions on MGF $M(\theta)$, an i.i.d. sequence of random variables $X_i, 1 \leq i \leq n$ with mean μ satisfies $P(S_n \geq a) \leq \exp(-nI(a))$, where $S_n = n^{-1}\sum_{1\leq i\leq n} X_i$ and $I(a) \equiv \sup_\theta(\theta a - \log M(\theta))$ is the Legendre transform. The function $I(a)$ is also commonly called the rate function in the theory of Large Deviation. The bound implies $\lim_n \sup \frac{\log P(S_n \geq a)}{n} \leq -I(a)$. We would like to establish the limit $\lim_{n\to\infty} \sup \frac{\log P(S_n \geq a)}{n} = -I(a)$.

Theorem 2.3 (Cramer Theorem). *Given a sequence of i.i.d. real value random variables $X_i, i \geq 1$ with a common MGF $M(\theta) = E[\exp(\theta X_1)]$ the following holds:*

- *For any closed set $F \subseteq \mathbb{R}$*

$$\lim_{n\to\infty} \sup \frac{1}{n}P(S_n \in F) \leq -\inf_{x\in F} I(x)$$

- *For any open set $U \subseteq \mathbb{R}$*

$$\lim_{n\to\infty} \inf \frac{1}{n}P(S_n \in U) \leq -\inf_{x\in U} I(x)$$

We only prove $D(M) = \mathbb{R}$.

To see the power of the theorem, let us apply it to the tail of S_n. We will first establish that $I(x)$ is a non-decreasing function on the interval $[\mu, \infty)$. Furthermore, we will establish that if it is finite in some interval containing x it is also continuous at x. Thus fix a and suppose I is finite in an interval containing a. Taking to be the clsed set $[a, \infty)$ with $a > \mu$, we obtain from the:

$$\lim_{n\to\infty} \sup \frac{1}{n} P(S_n \in [a, \infty)) \leq -\min_{a \geq a} I(x) = -I(a)$$

Applying the second part of Cramer Theorem, we obtain:

$$\lim_{n\to\infty} \inf \frac{1}{n} P(S_n \in [a, \infty)) \leq -\lim_{n\to\infty} \inf \frac{1}{n} \log P(S_n \in (a, \infty)) \geq -\inf_{x>a} I(x) = -I(a)$$

Thus in this special case indeed the large deviations limit exists and is equal to $-I(a)$. The limit is insensitive to whether the inequality is strict, in the sense that we also have the same limit for $P(S_n \geq a)$.

Proof. Part (a). Fix a closed set $F \subset \mathbb{R}$ let $\alpha_+ = \min\{x \in [\mu, \infty) \cap F\}$ and $\alpha_- = \max\{x \in (-\infty, \mu] \cap F\}$. Note that α_+ and α_- exist since F is closed. If $\alpha_+ = \mu$ then $I(\mu) = 0 = \min_{x\in\mathbb{R}} I(x)$. Note that $\log P(S_n \in F) \leq 0$, and the statement (a) follows trivially. Similarly, if $\alpha_- = \mu$, we also have statement (a). Thus we assume $\alpha_- < \mu < \alpha_+$ THen:

$$P(S_n \in F) \leq P(S_n \in [a, \infty)) + P(S_n \in (-\infty, \alpha_-])$$

Define:

$$x_n \equiv P(S_n \in [\alpha_+, \infty)), y_n \equiv P(S_n \in (-\infty, \alpha_-])$$

We already showed that:

$$P(S_n \geq \alpha_+) \leq \exp(-n(\theta\alpha_+ - \log M(\theta))), \forall \theta \geq 0$$

from which we have:

$$\begin{aligned} &\frac{1}{n} \log P(S_n \geq \alpha_+) \leq -(\theta\alpha_+ - \log M(\theta)), \forall \theta \geq 0. \\ \Longrightarrow &\frac{1}{n} \log P(S_n \geq \alpha_+) \leq -\sup_{\theta\geq 0}(\theta\alpha_+ - \log M(\theta)) = -I(\alpha_+) \end{aligned}$$

The second equality in the last equation is due to the fact that the supremum in $I(x)$ is achieved at $\theta \geq 0$, which was established as a part of Proposition 1. Thus we have:

$$\lim_n \sup \frac{1}{n} \log P(S_n \geq \alpha_+) \leq -I(\alpha_+) \tag{2.1}$$

Similarly, we have:

$$\lim_n \sup \frac{1}{n} P(S_n \leq \alpha_-) \leq -I(\alpha_-) \tag{2.2}$$

Applying Proposition 1 we have $I(\alpha_+) = \min_{x\geq\alpha_+} I(x)$ and $I(\alpha_-) = \min_{x\leq\alpha_-} I(x)$ Thus:

$$\min\{I(\alpha_+), I(\alpha_-)\}' = \inf_{x\in F} I(x) \tag{2.3}$$

From 2.1 to 2.3 we have that:

$$\limsup_n \frac{1}{n}\log x_n \leq -\inf_{x\in F} I(x), \limsup_n \frac{1}{n}\log y_n \leq -\inf_{x\in F} I(x) \tag{2.4}$$

which implies that:

$$\limsup_n \frac{1}{n}\log(x_n+y_n) \leq -\inf_{x\in F} I(x).$$

and therefore, we established:

$$\limsup_n \frac{1}{n}P(S_n\in F) \leq -\inf_{x\in F} I(x)$$

□

Proof. Part (b): Fix an open set $U\subset\mathbb{R}$ and fix $\varepsilon>0$ and find y such that $I(y)\leq\inf_{x\in U} I(x)$. It is sufficient to show that:

$$\lim_{n\to\infty}\frac{1}{n}P(S_n\in U)\geq -I(y) \tag{2.5}$$

since it will imply:

$$\lim_{n\to\infty}\inf\frac{1}{n}P(S_n\in U)\geq -\inf_{x\in U} I(x)+\varepsilon$$

and since $\varepsilon>0$ was arbitrary, it will imply the result.

Thus we now establish the 2.5. Assume $y>\mu$, and the case $y<\mu$ is treated similarly. Find $\theta_0=\theta_0(y)$ such that:

$$I(y)=\theta_0 y-\log M(\theta_0)$$

Such θ_0 exists by Proposition 1. Since $y>\mu$, then again by Proposition 1 we may assume $\theta_0\geq 0$.

We will use the change-of-measure technique to obtain the cover bound. For this, consider a new random variable let X_{θ_0} be a random variable defined by:

$$P(X_{\theta_0}\leq z)=\frac{1}{M(\theta_0)}\int_{-\infty}^{z}\exp(\theta_0 x)dP(x)$$

Now,

$$\begin{aligned}E[X_{\theta_0}]&=\frac{1}{M(\theta_0)}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x\exp(\theta_0 x)dP(x)\\&=\frac{\dot{M}(\theta_0)}{M(\theta_0)}=y.\end{aligned}$$

where the second equality we established in the previous lecture, and the last equality follows by the choice of θ_0 and Proposition 1. Since U is open we can find $\delta>0$ be small enough so that $(y-\delta,y+\delta)\subset U$. Thus we have:

$$\begin{aligned}P(S_n\in U)\geq P(S_n\in(y-\delta,y+\delta))&=\int_{|\frac{1}{n}\sum x_i-y|<\delta}dP(x_1)\dots dP(x_n)\\&=\int_{|\frac{1}{n}\sum x_i-y|<\delta}\exp\left(-\theta_0\sum_i x_i\right)M^n(\theta_0)\Pi_{1\leq i\leq n}M^{-1}(\theta_0)\exp(\theta_0 x_i)dP(x_i)\end{aligned}$$

Since θ_0 is non-negative, we obtain a bound:

$$P(S_n \in (y-\delta, y+\delta)) \geq \exp(-\theta_0 yn - \theta_0 n\delta)M^n(\theta_0)\int_{|\frac{1}{n}\sum x_i - y|<\delta} \Pi_{1\leq i\leq n} M^{-1}(\theta_0)\exp(\theta_0 x_i)dP(x_i)$$

However, we recognize the integral on the right hand side of the inequality above as the that the average $n^{-1}\sum_{1\leq i\leq n} Y_i$ of n i.i.d. random variables $Y_i, 1\leq i\leq n$ distributed according to the distribution of X_{θ_0} belongs to the interval $(y-\delta, y+\delta)$. Recall, however that $E[Y_i] = E[X_{\theta_0}] = y$. Thus by the WLLN, this probability converges to unity. As a result:

$$\lim_{n\to\infty} n^{-1}\log\int_{|\frac{1}{n}\sum x_i - y|<\delta} \Pi_{1\leq i\leq n} M^{-1}(\theta_0)\exp(\theta_0 x_i)dP(x_i) = 0$$

we obtain:

$$\begin{aligned}\lim_{n\to\infty}\inf n^{-1}\log P(S_n\in U) &\geq -\theta_0 y - \theta_0\delta + \log M(\theta_0)\\ &= -I(y) - \theta_0\delta\end{aligned}$$

Recalling that θ_0 depends on y only and sending δ to zero, we obtain 2.5. This completes the proof of part (b). □

2.4 Large Deviations in $\mathbb{R}^d$

Let $X_n \in \mathbb{R}^d$ be i.i.d. random variables and $A \subset \mathbb{R}^d$. Let $S_n = \sum_{1\leq i\leq n} X_n$. the large deviation question is now regarding the existence of the limit:

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(\frac{S_n}{n}\in A)$$

given $\theta\in\mathbb{R}^d$, define $M(\theta) = \mathbb{E}[\exp(<\theta, X_1)]$ where $<\cdot,\cdot>$ represents the inner product of two vectors: $<a,b>=\sum_i a_i b_i$. Define $I(x) = \sup_{\theta\in\mathbb{R}^d}(<\theta, x> - \log M(\theta))$, where again $I(x)=\infty$ is possibility.

Theorem 2.4 (Cramer's theorem in multiple dimensions). *Suppose $M(\theta) < \infty$ for all $\theta^d \in \mathbb{R}^d$. Then:*

- *for all closed set $F\subset\mathbb{R}^d$,*

$$\lim_{n\to\infty}\sup\frac{1}{n}\mathbb{P}(\frac{S_n}{S}\in F) \leq -\inf_{x\in F} I(x)$$

- *for all open set $U\subset\mathbb{R}^d$,*

$$\lim_{n\to\infty}\inf\frac{1}{n}\log\mathbb{P}(\frac{S_n}{S}\in U) \geq -\inf_{x\in U} I(x)$$

Unfortunately, the theorem does not hold in full generality, and the additional condition such as $M(\theta) < \infty$ for all θ is needed.

Example 2.6. Let us consider an example of application of this Theorem. Let $X_n = N(0,\Sigma)$ where $d=2$ and:

$$\Sigma = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1\end{bmatrix}, F = \{(x_1,x_2) : 2x_1 + x_2 \geq 5\}.$$

Goal: prove that the limit $\lim_n \frac{1}{n}\log\mathbb{P}(\frac{S_n}{n}\in F)$ exists and compute it.

By the upper bound part,

$$\limsup_n \frac{1}{n} \log \mathbb{P}(\frac{S_n}{n} \in F) \leq - \inf_{x \in F} I(x)$$

we have $M(\theta) = \mathbb{E}[\exp(<\theta, X>)]$ Let $=^d$ denote equality in distribution. Then we have:

$$<\theta, X> =^d N(0, \theta^T \Sigma \theta) = N(0, \theta_1^2 + \theta_1\theta_2 + \theta_2^2),$$

where $\theta = (\theta_1, \theta_2)$. Thus:

$$M(\theta) = \exp\left(\frac{1}{2}(\theta_1^2 + \theta_1\theta_2 + \theta_2^2)\right)$$
$$I(x) = \sup_{\theta_1, \theta_2} (\theta_1 x_1 + \theta_2 x_2 - \frac{1}{2}(\theta_1^2 + \theta_1\theta_2 + \theta_2^2))$$

let:

$$g(\theta_1, \theta_2) = \theta_1 x_1 + \theta_2 x_2 - \frac{1}{2}(\theta_1^2 + \theta_1\theta_2 + \theta_2^2)$$

From $\frac{\partial}{\partial \theta_j} g(\theta_1, \theta_2) = 0$ we obtain: $x_1 - \theta_1 - \frac{1}{2}\theta_2 = 0, x_2 - \theta_2 - \frac{1}{2}\theta_1 = 0$.
From which we have $\theta_1 = \frac{4}{3}x_1 - \frac{2}{3}x_2, \theta_2 = \frac{4}{3}x_2 - \frac{2}{3}x_1$ Then:

$$I(x_1, x_2) = \frac{2}{3}(x_1^2 + x_2^2 - x_1 x_2)$$

So we need to find: $\inf_{x_1, x_2} \frac{2}{3}(x_1^2 + x_2^2 - x_1 x_2)$ subject to $2x_1 + x_2 \geq 5 (x \in F)$. This becomes a non-linear optimization problem. Applying the Karush-Kuhn-Tucker condition we obtain:

$$\min f, s.t., g \leq 0 \quad \nabla f + \mu \nabla g = 0, \quad \mu g = 0, \quad \mu < 0$$

which gives:

$$(\frac{4}{3}x_1 - \frac{2}{3}x_2, \frac{4}{3}x_2 - \frac{2}{3}x_1) + \mu(2, 1) = 0, \mu(2x_1 + x_2 - 5) = 0$$

If $2x_1 + x_2 - 5 \neq 0$ then $\mu = 0$ and further $x_1 = x_2 = 0$ but this violates $2x_1 + x_2 \geq 5$. Therefore, we have $x_2 = 5 - 2x_1$. Thus we have a one dimensional unconstrained minimization problem:

$$\min \frac{2}{3}x_1^2 + \frac{2}{3}(5 - 2x_1)^2 - x_1(5 - 2x_1)$$

which gives $x_1 = \frac{10}{11}, x_2 = \frac{35}{11}$ Thus $I = 5.37$. Applying the lower bound part of the Cramer's Theorem we obtain the result is -5.37.

2.4.1 Gartner-Ellis Theorem

The GE Theorem deals with large deviations when the sequence X_n is not necessarily independent. Let X_n be a sequence of not necessarily independent random variables in $\mathbb{R}^d$. Then in general for $S_n = \sum_{1 \leq k \leq n} X_k$ the identity $\mathbb{E}[\exp(<\theta, S_n>)] = (\mathbb{E}[\exp(<\theta, X_1>)])^n$ does not hold. Nevertheless there exists a broad set of conditions set of conditions under which the large deviations bonuds hold. Thus consider a general sequence of random variables $Y_n \in \mathbb{R}^d$ which stands for $(1/n)S_n$ in the i.i.d. case. Let $\phi_n(\theta) = \frac{1}{n} \log \mathbb{E}[\exp(n <\theta, Y_n>)]$. Note that for the i.i.d. case:

$$\begin{aligned}\phi_n(\theta) &= \frac{1}{n}\log\mathbb{E}[\exp\big(n < \theta, n^{-1}S_n >\big)] = \frac{1}{n}\log M^n(\theta)\\ &= \log M(\theta) = \log\mathbb{E}[\exp(< \theta, X_1 >)].\end{aligned}$$

Loosely speaking GE theorem says that when convergence

$$\phi_n(\theta) = \phi(\theta)$$

takes place for some limiting function ϕ, then under certain additional technical assumptions, the large deviations principle holds for rate function

$$I(x) =^{\Delta} \sup_{\theta\in\mathbb{R}}(< \theta, x > -\phi(\theta))$$

Theorem 2.5 (Gartner-Ellis Theorem)**.** *Given a sequence of random variables Y_n, suppose the limit $\phi_n(\theta) \to \phi(\theta)$ exists for all $\theta \in \mathbb{R}^d$. Furthermore, suppose that $\phi(\theta)$ is finite and differentiable everywhere on $\mathbb{R}^d$. Then the large deviation principle holds for I defined by above:*

$$\limsup_n \frac{1}{n}\log\mathbb{P}(Y_n \in F) \leq -\inf_{x\in F} I(x),\ \textit{for all closed set } F \subset \mathbb{R}^d$$

$$\liminf_n \frac{1}{n}\log\mathbb{P}(Y_n \in U) \geq -\inf_{x\in U} I(x),\ \textit{for all open set } U \subset \mathbb{R}^d$$

This theorem is not proved because the Markov chain is not covered right now, switch back to prove later.

Chapter 3

Generalized Hoeffding inequality and Apps

3.1 Relationships Between the Modes of Convergences

Definition 3.1 (infinitely often (i.o.))**.** We say that events in the sequence occur "infinitely often" if A_n holds true for an infinite number of indices $n \in \{1, 2, 3, \ldots\}$.

Theorem 3.1 (a.s. convergence $\implies$ convergence in distribution)**.** *$X^{(n)} \xrightarrow{as} X$ if and only if $P(\left\|X^{(n)} - X\right\| \geq \varepsilon, i.o.) = 0, \forall \varepsilon > 0$.*

Proof. The event $(X^{(n)} \xrightarrow{d} X)^c$ is equivalent to the event $\bigcup_\varepsilon \{\left\|X^{(n)} - X\right\| \geq \varepsilon, i.o.\}$. It follows that the event $X^{(n)} \xrightarrow{d} X$ is equivalent to $P(\left\|X^{(n)} - X\right\| \geq \varepsilon, i.o.) = 0, \forall \varepsilon > 0$. □

Proposition 3.1 (a.s. $\implies$ p $\implies$ d)**.**

$$X^{(n)} \xrightarrow{as} X \implies X^{(n)} \xrightarrow{p} X$$
$$X^{(n)} \xrightarrow{p} X \implies X^{(n)} \rightsquigarrow X$$

$\rightsquigarrow$ means convergence in distribution.

Proof. We first prove that convergence with probability one implies convergence in probability. Since:

$$\{\left\|X^{(n)} - X\right\| \geq \varepsilon, i.o.\} = \lim_n \sup\{\left\|X^{(n)} - X\right\| \geq \varepsilon\}.$$

the event $X^{(n)} \xrightarrow{d} X$ implies:

$$\lim_n P(\left\|X^{(n)} - X\right\| \geq \varepsilon) \leq \lim_n \sup(\left\|X^{(n)} - X\right\| \geq \varepsilon) \leq P(\lim_n \sup(\left\|X^{(n)} - X\right\| \geq \varepsilon)) = 0$$

The inequality $\lim\sup P(A_n) \leq P(\lim\sup A_n)$ follows from Fatou's lemma applied to the sequence of indicator functions $f_n = I_{A_n}$ and the measure $\mu = P$. The last equality follows from the previous proposition.

We next show that the convergence in probability implies convergence in distribution. Denoting $\mathbf{1} = (\mathbf{1}, \ldots, \mathbf{1})$, and we have that $\mathbf{X^{(n)}} \leq \mathbf{x}$ then either $\mathbf{X} \leq \mathbf{x} + \varepsilon\mathbf{1}$, or $\left\|\mathbf{X} - \mathbf{X^{(n)}}\right\| > \varepsilon$ or both. Similarly, if $\mathbf{X} \leq \mathbf{x} - \varepsilon\mathbf{1}$ then either $\mathbf{X^{(n)}} \leq \mathbf{x}$ or $\left\|\mathbf{X} - \mathbf{X^{(n)}}\right\| > \varepsilon$ or both. This implies that for all n,

$$F^{(n)}_{\mathbf{X}}(\mathbf{x}) \leq \mathbf{P}(\mathbf{X} \leq \mathbf{x} + \varepsilon\mathbf{1}) + \mathbf{P}(\left\|\mathbf{X} - \mathbf{X^{(n)}}\right\| > \varepsilon) = \mathbf{F_X}(\mathbf{x} + \varepsilon\mathbf{1}) + \mathbf{P}(\left\|\mathbf{X} - \mathbf{X^{(n)}}\right\| > \varepsilon)$$

and

$$F_{\mathbf{X}}^{(n)}(\mathbf{x}-\varepsilon\mathbf{1}) \leq \mathbf{P}(\mathbf{X}\leq\mathbf{x}) + \mathbf{P}(\left\|\mathbf{X}-\mathbf{X^{(n)}}\right\| > \varepsilon) = \mathbf{F_{X^{(n)}}}(\mathbf{x}+\varepsilon\mathbf{1}) + \mathbf{P}(\left\|\mathbf{X}-\mathbf{X^{(n)}}\right\| > \varepsilon)$$

Since $\mathbf{X^{(n)}} \xrightarrow{\mathbf{p}} \mathbf{X}$ we have $P(\left\|\mathbf{X}-\mathbf{X^{(n)}}\right\| > \varepsilon) \to 0$ and letting $n \to \infty$, we get:

$$F_X(\mathbf{x}-\varepsilon\mathbf{1}) \leq \liminf \mathbf{F_{X^{(n)}}}(\mathbf{x}) \leq \limsup \mathbf{F_{X^{(n)}}}(\mathbf{x}) \leq \mathbf{F_X}(\mathbf{x}+\varepsilon\mathbf{1})$$

The LHS and RHS converge to $F_X(x)$ as $\varepsilon \to 0$ and points x where F_X is continuous, implying $F_{\mathbf{X^{(n)}}}(\mathbf{x}) \to \mathbf{F_X}(\mathbf{x})$. □

Proposition 3.2 (multi-dimensional convergence in distribution condition). *If* $\mathbf{c} \in \mathbb{R}^{\mathbf{d}}$ *then:* $\mathbf{X^{(n)}}$ *converges in distribution to* c *if and only if* $\mathbf{X} \xrightarrow{\mathbf{d}} \mathbf{c}$.

Proof. It suffices to prove that convergence in distribution to a constant vector implies probability in probability (convergence in probability always implies convergence in distribution). We prove the result below for two dimensions $d = 2$.

We have $P(\left\|X^{(n)}-\mathbf{c}\right\| \leq \sqrt{2}\varepsilon) \geq P(c-\varepsilon(1,1) < \mathbf{X^{(n)}} \leq \mathbf{c}+\varepsilon(\mathbf{1},\mathbf{1})) = \mathbf{P}(\mathbf{X^{(n)}} \leq \mathbf{c}+\varepsilon(\mathbf{1},\mathbf{1})) - \mathbf{P}(\mathbf{X^{(n)}} \leq \mathbf{c}+\varepsilon(\mathbf{1},-\mathbf{1})) - \mathbf{P}(\mathbf{X^{(n)}} \leq \mathbf{c}+\varepsilon(-\mathbf{1},\mathbf{1})) + \mathbf{P}(\mathbf{X^{(n)}} \leq \mathbf{c}+\varepsilon(-\mathbf{1},-\mathbf{1}))$ □

Proposition 3.3 (muti-dimensional convergence in a.s. condition). *The convergence* $\mathbf{X^{(n)}} \xrightarrow{\mathbf{p}} \mathbf{X}$ *occurs if and only if every sequence of natural numbers* $n_1, n_2, \ldots \in \mathbb{N}$ *has a subsequence* $r_1, r_2, \ldots \in \{n_1, n_2, \ldots\}$ *such that* $\mathbf{X^{(r_k)}} \xrightarrow{\mathbf{as}} \mathbf{X}$ *as* $k \to \infty$.

Proof. We assume that $\mathbf{X^{(n)}} \xrightarrow{\mathbf{p}} \mathbf{X}$ and consider a sequence of positive numbers ε_i such that $\sum_i \varepsilon_i < \infty$. For each ε_i we can find a natural number n_i' such that $P(\left\|X^{(n)}-X\right\| \geq \varepsilon_i) < \varepsilon_i$ for all $n_i > n_i'$ we can assume without loss of generality, $n_1' < n_2' < \ldots$.

Defining A_i to be the event $\{\left\|X^{(n)}-X\right\| \geq \varepsilon_i\}$, we have $\sum_i P(A_i) \leq \sum_i \varepsilon_i < \infty$ and by the first Borell-Cantelli Lemma, we have $P(A_i, i.o.) = 0$ since $\lim_{k\to\infty} \varepsilon_k = 0$, this implies that for all $\varepsilon > 0, P(\{\left\|X^{(n)}-X\right\| \geq \varepsilon_i\}, i.o.) = 0$, which implies $\mathbf{X^{n_i'}} \xrightarrow{\mathbf{as}} \mathbf{X}$ as $i \to \infty$.

Consider now an arbitrary sequence $n_1, \ldots$ of natural numbers we have $\mathbf{X^{(n_i)}} \xrightarrow{\mathbf{p}} \mathbf{X}$ as $i \to \infty$ and repeating the above argument with $n_1, \ldots$ replacing $1, 2, 3, \ldots$, we can find a sequence $r_1, \ldots$ of $n_1, \ldots$ along which $\mathbf{X^{(r_i)}} \xrightarrow{\mathbf{as}} \mathbf{X}$ as $i \to \infty$.

To show the converse, we assume the original argument is wrong, then there exists $\varepsilon > 0$ and $\delta > 0$ such that $P(\left\|X^{(n)}-X\right\| \geq \varepsilon) > \delta$ for infinitely many k, which we denote $n_1, n_2, \ldots$ This implies that there exists no subsequence of $n_1, n_2, \ldots$ along which $\mathbf{X^{(n_i)}} \xrightarrow{\mathbf{as}} \mathbf{X}$. □

Theorem 3.2 (Slutsky's Theorem). *Let* X_n, Y_n *be sequences of scalar/vector/matrix random elements. If* X_n *converges in distribution to a random element* X *and* Y_n *converges in probability to a constant* c *then:*

- $X_n + Y_n \xrightarrow{d} X + c$;
- $X_n Y_n \xrightarrow{d} Xc$;
- $X_n / Y_n \xrightarrow{d} X/c$, *provided that* c *is invertible.*

Proof. $X_n \xrightarrow{d} X \implies X_n \xrightarrow{p} X$ then $(X_n, Y_n) \xrightarrow{p} (X, Y)$; and continuous mapping theorem.

□

Theorem 3.3 (continuous mapping theorem). *Let* X *be random elements and* $g : S \to S'$ *has the set of discontinuity points* D_g *such that* $Pr[X \in D_g] = 0$. *Then:*

- $X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X);$
- $X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X);$
- $X_n \xrightarrow{as} X \implies g(X_n) \xrightarrow{as} g(X);$

3.2 Generalized Hoeffding inequality

Recap:

From the last chapter, we discussed the regular, deviation, medium deviation, and large deviation theory, and standard concentration inequality. we focused on i.i.d. case, but the result, and almost identical proof, extends to independent case: Suppose X_i's are independent with mean μ_i and suppose $[a_i, b_i]$, and $E[X_i] = \mu_i$. Then for any $\varepsilon > 0$:

$$P(S_n - \sum_{i=1}^{n} \mu_i > \varepsilon) \leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$$

However, there are some drawbacks of Hoeffding inequality:

- Can we extend the results to unbounded randon variables?
- Can we extend the results to dependent random variables?

Yes, but we need some assumptions.

Definition 3.2 (sub-Gaussian random variable)**.** The probability distribution of a random variable X is called sub-Gaussian if there is a positive constant C such that for all $t \geq 0, P(|X| \geq t) \leq 2\exp\left(-t^2/C^2\right)$. Alternatively, a random variable is considered sub-Gaussian if its distribution function is upper bounded (up to a constant) by the distribution function of a Gaussian. Specifically, we say X is sub-Gaussian if for all $s \geq 0$ we have:

$$P(|X| \geq s) \leq cP(|Z| \geq s)$$

where $c \geq 0$ is constant, and Z is mean zero Gaussian random variable.

Definition 3.3 (alternative definition of sub-Gaussian)**.** A random variable X is said to be sub-Gaussian if there exists a positive constant X such that its moment generating function (MGF) satisfies the following inequality for all t in some neighborhood of 0:

$$E[e^{tX}] \leq e^{\frac{K^2t^2}{2}}$$

Claim 3.1. *X is sub-Gaussian if and only if the MGF of X^2 is finite near* 0.

Proof. $\implies$

Assume X is sub-Gaussin, then for some $K > 0$, and t near 0, we have:

$$E[e^{tX}] \leq e^{\frac{K^2t^2}{2}}$$

Consider the MGF of X^2 at s:

$$M_{X^2}(s) = E[e^{sX^2}]$$

By using a Taylor Series of e^{sX^2} and propery of expectation, we can express $M_{X^2}(s)$ in terms of moments of X, which is bounded by moments of Gaussian random variable, implying finite around 0.

The other side:

There exists some $\varepsilon > 0$ such that $M_{X^2}(s) = E[e^{sX^2}]$ is finite for $|s| < \varepsilon$ From the finiteness of $M_{X^2}(s)$ and the properties of MGFs, it follows from the tails of X^2 are controlled in a manner similar to Gaussian random variable. □

Proposition 3.4. *Let $X = (X_1, \dots, X_n)$ be a vector of independent σ-sub-Gaussian random variables. Then, the random vector X is σ-sub-Gaussian.*

note that every bounded variable is sub-Gaussian.

Example 3.1. Normal, and bounded random variables are sub-Gaussian.

Counter example: exponential, Poisson, are not sub-Gaussian.

Theorem 3.4 (Generalized Hoeffding inequality)**.** *If $X_1, X_2, \dots$ are i.i.d. sub-Gaussian, then there exists $c > 0$ such that:*

$$P(|S_n - n\mu| \geq \varepsilon) \leq 2e^{-\frac{c\varepsilon^2}{n}}$$
$$P(S_n \geq n\mu + \varepsilon) \leq e^{-\frac{c\varepsilon^2}{n}}$$

Application of Generalized Hoeffding inequality

let $\varepsilon = \sqrt{\frac{\alpha}{c}}n$ then we obtain:

$$P(S_n \geq n\mu + \frac{\alpha}{c}n) \leq e^{-\alpha n}$$

Let $\varepsilon = \sqrt{\frac{\alpha}{c}\log n}$, then we obtain:

$$P(S_n \geq n\mu + \varepsilon) \leq \frac{1}{n^\alpha}$$

These are “one-direction” result of large and medium deviation theory.

Definition 3.4 (subexponential r.v.)**.** A random variable X is said to be sub-exponential if its tail satisfies, for some $k > 0$:

$$P(|X| \geq t) \leq 2e^{-kt}, \forall t \geq 0$$

Claim 3.2. *It can be argued that X is sub-exponential if and only if the MGF of $|X|$ is finite near 0. Also, X is sub-exponential if and only if X^2 is sub-Gaussian.*

Theorem 3.5 (Bernstein inequality)**.** *Let $X_1, X_2, \dots$ be a sequence i.i.d. sub-exponential random variable. Then there exist constants $c_1, c_2, c_3 > 0$ and $K > 0$, such that for any $\varepsilon > 0$:*

$$P(|S_n - n\mu| \geq \varepsilon) \leq 2e^{-c_3 \min\{\frac{c_1\varepsilon^2}{n}, c_2\varepsilon\}}, \forall \varepsilon \geq 0,$$
$$P(S_n \geq n\mu + \varepsilon) \leq e^{-c_3 \min\{\frac{c_1\varepsilon^2}{n}, c_2\varepsilon\}}, \forall \varepsilon \geq 0$$

Totally, we are interested in the case where ε grows slower than $O(n)$. Then the first term prevail for large n.

Concentration inequality can be extended to general function of independent random variables, say $g(x_1, \dots, x_n)$.

Theorem 3.6 (McDiarmid inequality)**.** *Suppose $X_1, X_2, \dots$ are independent random variables. if g satisfies:*

$$\sup_{x_1,\dots,x_n,x_i'} |g(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n) - g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)| \le c_i$$

Then we have:

$$P(g(X_1,\dots,X_n) - E[g(X_1,\dots,X_n)] > \varepsilon) \le e^{-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}}$$

McDiarmid inequality is also known as Bounded Difference inequality.

3.3 Revenue management

A flight departs in period T. The demand in period $t \le T$, denoted by $D_t(p_t)$ is random and price sensitive that depends on selling price p_t. Let $d_t(p_t) = E[D_t(p_t)]$ and it is decreasing in p_t. There is a total of N seats available.

Standard formulation in revenue management is that, we divide time into T small time intervals, each is short enough so that, in each time interval, we have either one potential demand or no demand.

Then, $D_t(p_t)$ becomes Bernoulli random variables with success probability $d_t(p_t)$.

the dynamic optimization problem is:

$$\max_{p_t} E[\sum_{t=1}^T p_t D_t(p_t)], \quad s.t., \sum_{t=1}^T D_t(p_t) \le N, \quad p_t \ge 0, t = 1,\dots,T.$$

This dynamic optimization problem can be formulated and solve using DP, and computation of optimal policy can be complex.

Fluid approximation $\implies$ deterministic optimization problem and use concentration inequality.

Actually there are two set of effective approximation: one is reoptimize, when we have new information, we reoptimize the remaining time. The other is policy adjustment, we adjust the policy based on the new information.

There are simple approximate solutions. The result is that, using fixed price policy, the total loss of revenue is in the order $O(\sqrt{T\log(T)})$ or put in differently, the average loss per period is $O(T^{-1/2}(\log(T))^{1/2})$, which is small when T is large.

Fluid approximation

Fluid model is:

$$\max_{p_t} \sum_{t=1}^T p_t d_t(p_t) \quad s.t., \sum_{t=1}^T d_t(p_t) \le N, \quad p_t \ge 0, t = 1,\dots,T.$$

Reformulation: As there is a one-to-one correspondence between $d_t(p_t)$ and p_t, we can change the decision variable to d_t. let $p(d_t)$ be the pricing corresponding to d_t:

$$\max_{p_t} \sum_{t=1}^T p(d_t) d_t \quad s.t., \sum_{t=1}^T d_t \le N, \quad d_t \in \{0,1\}, t = 1,\dots,T.$$

it is typical to assume revenue function $d_t p(d_t)$ is concave in d_t. This is a convex optimization problem.

Let $d^* = \arg\max_d dp(d)$ and $p^* = d(p^*)$. Then we have:

Claim 3.3. *Optimal solution to the fluid model is $d^*(N/T) = \min\{d^*, N/T\}$*

Claim 3.4. *The optimal objective value of fluid model:*

$$T_p^*(N/T)d^*(N/T)$$

is an upper bound for the original problem.

we can use loss of regress to evaluate the performance of fluid model. The gap between policy we have and upper bound (look at the difference).

The loss, or regret is defined as the difference between the value function of the said policy and that of the true optimal solution. The smaller the loss, the better the policy. However, we do not know what is the true optimal value function. We use upper bound of the optimal revenue.

Theorem 3.7 (bounded condition of total loss)**.** *Let $p^*(N/T) := \max\{p^*, p(N/T)\}$ then using the static policy $p^*(N/T)$ in each period has a total loss bounded by:*

$$L(T) = O(\sqrt{T\log(T)})$$

Chapter 4

DP Formulation and Martingales

4.1 Dynamic Pricing

Recall that the DP formulation of the RM problem is

$$\max_{p_t} E[\sum_{t=1}^{T} p_t D_t(p_t)], \quad s.t., \sum_{t=1}^{T} D_t(p_t) \leq N, \quad p_t \geq 0, \quad \forall t, \dots, T$$

And the fluid model is, after changing the decision variable to d_t and using the inverse function $p(d_t)$ (assumed to be bounded), (usually not the same across periods, but we use the fliud model which removes all the uncertainties)

$$\max_{d_t} \sum_{t=1}^{T} d_t p(d_t), \quad s.t., \sum_{t=1}^{T} d_t \leq N, \quad d_t \geq 0, \quad \forall t, \dots, T$$

Assuming concavity of $d_t p(d_t)$, this is a convex optimization problem (convert to minimization so this is a convex problem) whose optimal solution is easy to obtain. This is easy to solve since the constraint is linear.

Lemma 4.1. *Let $d^* = \arg\max_d dp(d)$ and $p^* = d(p^*)$. Then we have the following two claims:*

Claim 4.1. *optimal solution to the fliud model is: $d^*(N/T) = \min\{d^*, N/T\}$ and $p^*(N/T) = \max\{p^*, p(N/T)\}$. In other words, if $Td^* > N$ we can adjust to raise the price.*

Claim 4.2. *The optimal objective value of the fluid model, $Tp^*(N/T)d^*(N/T)$, is an upper bound for the origianl model (averaging selling strategy).*

fluid model almost all of them is the highest revenue that you can obtainl

The loss, or regret is defined as the difference between the value function of the said policy and that of the true optimal solution. Since we do not know the true optimal value function, we evaluate the performance of the policy by, if our policy has price p_t in period t,

$$R(T) = Tp^*(N/T)d^*(N/T) - E[\sum_{t=1}^{T} p_t D_t(p_t)].$$

The loss per period is $R(T)/T$. (we hope this sequence would converge to 0 or at least constant)

Theorem 4.1 (average loss under bounded constraint)**.** *let $p^*(N/T) \equiv \max\{p^*, p(N/T)\}$ and then using the static policy $p^*(N/T)$ in each period has a total loss bounded by: $L(T) = O(\sqrt{T \log T})$ and then the average loss per period is $O(\sqrt{\frac{\log T}{T}})$*

Proof. Consider the proposed static policy $p^*(N/T)$ and expected demand per period is $d^*(N/T)$.

Denote the demand under policy $p_t = p^*(N/T), t = 1, \dots, T$, by $D_1, \dots, D_T$, which is bounded by $\bar{D}$. Then $E[D_t] = d^*(N/T)$. Let $S_T = \sum_{t=1}^T D_t$. By Concentration Inequality, with appropriate choice of α, we have:

$$P(|S_T - Td^*(N/T)| > \sqrt{\alpha T \log T}) \leq \frac{1}{T}$$

Define the good event $A = \{|S_T - Td^*(N/T)| \leq \sqrt{\alpha T \log T}\}$. Then we have:

$$P(A) \geq 1 - \frac{2}{T}, P(A^c) \leq 2/T$$

We have:

$$\begin{aligned}\left|E[\sum_{t=1}^T p^*(N/T)D_t] - Tp^*(N/T)d^*(N/T)\right| &\leq p^*(N/T)E[\left|\sum_{t=1}^T D_t - Td^*(N/T)\right| \mid A]P(A) \\ &+ p(N/T)E[\left|\sum_{t=1}^T D_t - Td^*(N/T)\right| \mid A^c]P(A^c) \\ &\leq p^*(N/T)\sqrt{\alpha T \log T} + T\bar{D}\frac{2}{T} = O(\sqrt{T \log T})\end{aligned}$$

this is because the bad event happens only at constant loss and we can ignore that. □

4.2 Multi-Armed Bandit (MAB) problem

There are m arms. Arm a generates i.i.d. reward with mean $\mu(a)$ when playing arm a. There DM has no information about the reward from each arm, and needs to determine a strategy to maximize the expected total reward up to any time T. Assume that reward is bounded (and WLOG, by 1).

Let $n_t(a)$ denote the number of plays of arm a before time t, then we can compute the sample mean $\bar{\mu}_t(a)$. Let:

$$r_t(a) = \sqrt{\frac{\alpha \log n_t(a)}{n_t(a)}}$$

You can call $[\bar{\mu}_t(a) - r_t(a), \bar{\mu}_t(a) + r_t(a)]$ the confidence interval of true mean $\mu(a)$ by t.

4.2.1 Simple Algorithms: Uniform Exploration

1. Exploration phase: try each arm N times;
2. Select the arm $\hat{a}$ with the highest average reward (break ties arbitrarily);
3. Exploration phase: play arm $\hat{a}$ in all remaining rounds.

By Hoeffding Inequality we have:

$$Pr[|\bar{\mu}(a) - \mu(a)| < \text{ rad }] \geq 1 - 2/T^4, \text{ where rad } \equiv \sqrt{2\log(T)/N}$$

We define the clean event to be the event that satisfies the constraint above for all arms simultaneously. Similarly, we have the bad event.

We start with $K = 2$ arms, consider the clean event. We will show that if we choose the worse arm, it is not so bad because the expected rewards for the two arms would be close. Let the best arm be a^* and suppose the algorithm chooses the other arm $a \neq a^*$. This must have been because its average reward was better than that of $a^* : \bar{\mu}(a) > \bar{\mu}(a^*)$ Since this is a clean event, we have:

$$\mu(a) + \text{ rad } \geq \bar{\mu}(a) > \bar{\mu}(a^*) \geq \mu(a^*) - \text{ rad } .$$

Re-arranging the terms, it follows that $\mu(a^*) - \mu(a) \leq 2$ rad.

Thus each round in the exploitation phase contributes at most 2rad to regret. Each round in exploration trivially contributes at most 1. We derive the upper bound on the regret, which consists of two parts: for exploration, when each arm is chosen N times, and then for the remaining $T - 2N$ rounds of exploration:

$$R(T) \leq N + 2 \text{ rad } \cdot (T - 2N) < N + 2 \text{ rad } \cdot T.$$

The value of N is given to in advance. We can choose N so as to minimize the right-hand side. Since the two summands are, resp., monotonically increasing and monotonically decreasing in N, we can set N so that they are approximately equal. For $N = T^{2/3}(\log T)^{1/3}$, and we obtain:

$$R(T) \leq O(T^{2/3}(\log T)^{1/3}).$$

It remains to analyze the "bad event". Since regret can be at most T, and the bad event happens with a very small probability, regret from this event can be neglected. Formally,

$$\begin{aligned} E[R(T)] &= E[R(T) \mid \text{ clean event}] \times Pr(\text{ clean event}] + E[R(T) \mid \text{ bad event}] \times Pr[\text{ bad event}] \\ &\leq E[R(T) \mid \text{ clean event}] + T \times O(T^{-4} \leq O((\log T)^{1/3} \times T^{2/3}) \end{aligned}$$

For $K > 2$ the argument follows the same ($T \geq K$).

4.2.2 Improvement: Epsilon-greedy algorithm

For each round $t = 1, 2, \ldots$ do:

Toss a coin with success probability ε_t:

If success then select explore: choose an arm uniformly at random; else: choose the arm with the highest average reward so far.

Theorem 4.2 (Epsilon-greedy algorithm). *Epsilon-greedy algorithm with exploration probabilities $\varepsilon_t = t^{-1/3} \cdot (K \log t)^{1/3}$ achieves regret bound $E[R(t)] \leq t^{2/3} \cdot O(K \log t)^{1/3}$ for each round t.*

Explore-first and Epsilon-greedy do not adapt their exploration schedule to the history of the observed rewards. We refer to this property as *non-adaptive exploration*, and formalizes:

Definition 4.1 (non-adaptive exploration). A round t is an exploration round if the data (a_t, r_t) from this round is used by the algorithm in the future rounds. A deterministic algorithm satisfies non-adaptive exploration if the set of all exploration rounds and the choice of arms therein is fixed before round 1. A randomized algorithm satisfies *non-adaptive exploration* if it does so far each realization of its random seed.

4.2.3 Advanced algorithms: adaptive exploration

let us start with the case of $K = 2$ arms, one natural idea is as:

alternate the arms until we are confident which arm is better, and play this arm thereafter.

Fix round t and arm a. Let $n_t(a)$ be the number of rounds before t in which this arm is chosen, and let $\bar{\mu}_t(a)$ be the average reward in these rounds. We have the Hoeffding Inequality:

$$Pr[|\bar{\mu}_t(a) - \mu(a)| \leq r_t(a)] \geq 1 - \frac{2}{T^4}, \text{ where } r_t(a) = \sqrt{2\log(T)/n_t(a)}$$

This equation does not follow immediately. This is because Hoeffding Inequality only applies to a fixed number of independent random variables, whereas here we have $n_t(a)$ random samples from reward distribution $\mathcal{D}_a$, where $n_t(a)$ is itself is a random variable. We present an elementary version of this argument (can also use martingale to prove this). For each arm a, an $1 \times T$ table with each cell independently sampled from $\mathcal{D}_a$. WLOG, the reward tapes encodes rewards as: the j-th time a given arm a is chosen by the algorithm, its reward is taken from the j-th cell in this arm's tape. Let $\bar{v}_j(a)$ represents the average reward at arm a from first j times that arm a is chosen. Now we can use Hoeffding Inequality to derive that:

$$\forall j, Pr[|\bar{v}_j(a) - \mu(a)| \leq r_t(a)] \geq 1 - 2/T^4.$$

Taking a union bound, it follows that:

$$Pr[\varepsilon] \geq 1 - 2/T^2, \text{ where } \varepsilon \equiv \{\forall a \forall t \mid |\mu_t(a) - \mu(a)| \leq r_t(a)\}.$$

The event ε will be the clean event for the subsequent analysis.

Back to the $k = 2$ arms, and the full algorithm for the two arms is:

Alternate two arms until $UCB_t(a) < LCB_t(a')$ after some even round t; Abandom arm a, and use arm a' before since.

Let t be the last round we did not invoke the stopping rule, when the confidence intervals of the two arms still overlap, then: $\Delta \equiv |\mu(a) - \mu(a')| \leq 2(r_t(a) + r_t(a'))$.

Since the algorithm has been alternating the two arms before time t, we have $n_t(a) = t/2$ which yields,

$$\Delta \leq 2(r_t(a) + r_t(a')) \leq 4\sqrt{2\log(T)/\lceil t/2 \rceil} = O(\sqrt{\log(T)/t}).$$

Then the total regre accumulated till round t is:

$$R(t) \leq \Delta \times t \leq O(t \cdot \sqrt{\frac{\log T}{t}}) = O(\sqrt{t \log T}).$$

Since we've chosen the best arm from then on, we have $R(t) \leq O(\sqrt{t \log T})$ and we need to argue that the "bad event" conributes to negligible amount to regret t, like:

$$\begin{aligned} E[R(T)] &= E[R(T) \mid \text{ clean event}] \times Pr(\text{ clean event}] + E[R(T) \mid \text{ bad event}] \times Pr[\text{ bad event}] \\ &\leq E[R(T) \mid \text{ clean event}] + t \times O(T^{-2})) \leq O(\sqrt{t \log T}). \end{aligned}$$

Lemma 4.2. *For two arms, the algorithm achieves regret $E[R(t)] \leq O(\sqrt{t \log T})$ for each round $t \leq T$.*

For $K > 2$: alternate the arms until some arm a is worse than some other arm with high probability. (Successive Elimination):

All arms have initially designated as active: and play each active once, deactive all arms a such that, letting t be the current round and $UCB_t(a) < LCB_t(a')$ for some other arm a'.

Let a^* be an optimal arm and note that it cannot be deactivated. Fix any arm a such that $\mu(a) < \mu(a^*)$ Consider the last round $t \leq T$ when the deactivation was invoked and arm a remained active. as in the argument $K = 2$ arms, the confidence interval of a and a^* must overlap at round t. Hence, $\Delta(a) = \mu(a^*) - \mu(a) \leq 2(r_t(a^*) + r_t(a)) = 4r_t(a)$.

By the choice of t, arm a can be played at most once afterward $n_T(a) \leq 1 + n_t(a)$. Thus, we have:

$$\Delta(a) \leq O(r_T(a)) = O(\sqrt{\log(T)/n_T(a)}) \text{ for each arm } a \text{ with } \mu(a) < \mu(a^*)$$

Informally: an arm played many times cannot be too bad. The contribution of arm a to regret at round t, denoted $R(t;a)$, can be expressed as $\Delta(a)$ for each round this arm is played, by the above result we can bound this quantity as:

$$R(t;a) = n_t(a)\cdot\Delta(a) \le n_t(a)\cdot O(\sqrt{\log(T)/n_t(a)}) = O(\sqrt{n_t(a)\log T}).$$

summing up over all arms, we obtain that:

$$R(t) = \sum_{a\in A} R(t;a) \le O(\sqrt{\log T})\sum_{a\in A}\sqrt{n_t(a)}.$$

since the function is concave and we have the Jensen's inequality that:

$$\frac{1}{K}\sum_{a\in A}\sqrt{n_t(a)} \le \sqrt{\frac{1}{K}\sum_{a\in A} n_t(a)} = \sqrt{\frac{t}{K}}.$$

Therefore, we have $R(t) \le O(\sqrt{Kt\log T})$. Thus we have proved:

Theorem 4.3 (Successive Elimination algorithm)**.** *Successive Elimination algorithm achieves regret:*

$$E[R(t)] = O(\sqrt{Kt\log T})\forall t \le T$$

A bad arm cannot play too often, i.e., $R(T,a) = \Delta(a)\cdot n_T(a) \le \Delta(a)\cdot O(\frac{\log T}{[\Delta(a)]^2}) = O(\frac{\log T}{\Delta(a)})$.

Theorem 4.4 (Successive Elimination algorithm Alternative Representation)**.** *successive Elmination algorithm achieves regret:*

$$E[R(T)] \le O(\log T)\left[\sum_{\text{arms } a \text{ with } \mu(a)<\mu(a^*)} \frac{1}{\mu(a^*)-\mu(a)}.\right]$$

The regret bound is logarithmic in T, with a constant that can be arbitrarily large depending on a problem.

4.2.4 Optimism Under Uncertainty

We can conduct the upper confidence bound method (UCB) algorithm:

First, play each arm once. After that, in each period t, always play the arm that has the highest UCB:

$$UCB_t(a) = \bar{\mu}_t(a) + r_t(a) = \bar{\mu}_t(a) + \sqrt{\frac{\alpha\log n_t(a)}{n_t(a)}}$$

where α is some parameter.

The algorithm in the literature typically use a larger bound, defined by:

$$UCB_t(a) = \bar{\mu}_t(a) + \sqrt{\frac{\alpha\log t}{n_t(a)}}$$

or

$$UCB_t(a) = \bar{\mu}_t(a) + \sqrt{\frac{\alpha\log T}{n_t(a)}}$$

in the discussion later, we only use the second definition.

UCB uses the so-called ”optimistic estimate” to avoid chance of under-play.

The same apporach has been used to develop algorithms for general RL problems.

Techinical Details

Probability of good event can be defiend as: $A_t(a) = \{|\bar{\mu}_t(a) - \mu(a)| > r_t(a)\}$. then the concentration inequality shows for every choice of α:

$$P(A_t(a)) \leq \frac{2}{T^4}$$

Let $\mathcal{A} = \bigcup_{t,a} A_t(a)$, then by union bound (assuming $m \leq T$)

$$P(\mathcal{A}) \leq \frac{2m}{T^3} \leq \frac{2}{T^2}$$

We shall call $\mathcal{A}$ ”bad event” and $\mathcal{A}^c$ ”good event”.

regret:

Clearly, the optimal policy is to play $a^* = \arg\max_a \mu(a)$. WLOG, suppose $a^* = 1$. For convenience we let $\Delta(a) = \mu(1) - \mu(a)$.

If a policy plays arm a_t in period t, then it regret is:

$$R_T = E\left[\sum_{t=1}^{T} (\mu(1) - \mu(a_t))\right]$$

Theorem 4.5 (UCB algorithm). *We have two regret bounds for UCB algorithm:*

$$R_T \leq \sum_{a:\Delta(a)>0} \frac{4\alpha}{\Delta(a)} \log T + 2,$$

$$R_T \leq (1 + 4\alpha)\sqrt{mT \log T} + 2$$

The first bound is known as “instance-dependent regret”, and the second regret bound that is “instance-independent regret”.

Proof. When will you play a wrong arm in period t? Answer $\bar{\mu}_t(a) + r_t(a) \geq \bar{\mu}_t(1) + r_t(1)$ for some $a \neq 1$.

Under good event $\mathcal{A}^c$, we have for each t and a:

$$\mu(a) + 2r_t(a) \geq \bar{\mu}_t(a) + r_t(a); \text{ and } \bar{\mu}_t(1) + r_t(1) \geq \mu(1).$$

Combining we obtain, if an arm $a \neq 1$ is played at t, then:

$$\mu(a) + 2r_t(a) \geq \mu(1)$$

This implies $r_t(a) \geq \Delta(a)/2$, or:

$$n_t(a) \leq \frac{4\alpha}{\Delta^2(a)} \log T$$

The analysis above shows that, any arm $a \neq 1$ is played at most $4\alpha\Delta^{-2}(a)\log T$ times under $\mathcal{A}^c$.

What is the expected number of times $a \neq 1$ is played by T?

It can be computed as follows:

$$\begin{aligned} E[n_T(a)] &= E[n_T \mid \mathcal{A}^c]P(\mathcal{A}^c) + E[n_T \mid \mathcal{A}]P(\mathcal{A}) \\ &\leq 4\alpha\Delta^{-2}(a)\log T + T \times 2/T^2 \leq 4\alpha\Delta^{-2}(a)\log T + 2/T \end{aligned}$$

The first regret bound is obtained as follows.

$$\begin{aligned} R(T) = E[\sum_{a:\Delta(a)\neq 0} \Delta(a)n_T(a)] &= E[\sum_{a:\Delta(a)\neq 0} \Delta(a)n_T(a) \mid \mathcal{A}^c]P(\mathcal{A}^c) \\ &+ E[\sum_{a:\Delta(a)\neq 0} \Delta(a)n_T(a) \mid \mathcal{A}]P(\mathcal{A}) \leq \sum_{a:\Delta(a)\neq 0} 4\alpha\Delta^{-1}\log T + 1, \end{aligned}$$

where the first term in the inequality follows from:

$$E[n_T(a) \mid \mathcal{A}^c] \leq 4\alpha\Delta^{-2}(a)\log T,$$

and the second term follows from $P(\mathcal{A}) \leq 2/T$, $\Delta(a) \leq 1$ and $\sum_a n_T(a) \leq T$.
The second regret is to divide the arms into two categories:

$$G_1 = \{i : \Delta(i) < \sqrt{\frac{m}{T}\log T}; \text{ and } G_2 = \{i : \Delta(i) \geq \sqrt{\frac{m}{T}\log T}$$

The total regret can be written as:

$$\sum_{i\in G_1} E[n_T(i)]\Delta_i + \sum_{i\in G_2} E[n_T(i)]\Delta_i$$

We shall evalutate these two parts separately.
The first part is bounded as follows:

$$\begin{aligned} \sum_{i\in G_1} E[n_T(i)]\Delta(i) &\leq \sqrt{\frac{m}{T}\log T}\sum_{i\in G_1} E[n_T(i)] \\ &\leq \sqrt{\frac{m}{T}\log T} \times T = \sqrt{mT\log T} \end{aligned}$$

The second part is for $i \in G_2$, we have $\Delta(i) \geq \sqrt{\frac{m}{T}\log T}$ thus: $\Delta(i)^{-1} \leq \sqrt{\frac{T}{m\log T}}$
Therefore,

$$\begin{aligned} \sum_{i\in G_2} E[n_T(i)]\Delta(i) &\leq \sum_{i\in G_1} (4\alpha\Delta^{-2}(a)\log T + 1/T)\Delta(i) \\ &= \sum_{i\in G_1} (4\alpha\Delta^{-1}(a)\log T + \Delta(i)/T) \leq 4\alpha\sqrt{mT\log T} + 1. \end{aligned}$$

To sum up, we obtain:

$$R(T) \leq (1+\alpha)\sqrt{mT\log T} + 1 = O(mT\log T)$$

□

4.3 Martingale

The result we discussed up to now assume that the stochastic sequence is independent, and many of our applications do not satisfy the independence condition.

Definition 4.2. A stochastic sequence $\{X_n : n \geq 0\}$ is called a martingale if (i) $E[|X_n|] < \infty$ and (ii)

$$E[X_{n+1} \mid X_0, X_1, \ldots, X_n] = X_n, n = 1, \ldots.$$

Remark 4.1. From the definition and law of total expectation, we immediate have $E[X_{n+1}] = E[X_n]$ and thus, $E[X_n] = E[X_0], \forall n \geq 1$.

A martingale is a generalization of a fair game, if we interpret X_n as a gambler's fortune after the n-th gamble, then the definition states that his expected fortune after the $(n+1)$-st game is equal to his fortune after the n-th gamble no matter what may have previous occured.

We can replace $\{X_0, \dots, X_n\}$ by $\mathcal{F}_n$ represeting the σ-algebra generated by $\{X_0, \dots, X_n\}$ or simply the information up to time n. E.g., if $s < t$, then $\mathcal{F}_s \subseteq \mathcal{F}_t$.

Definition 4.3 (submartingale and supermartingale)**.** An adapted process $(X_t, \mathcal{F}_t)$ is a submartingale if $X_t \leq E[X_{t+1} \mid \mathcal{F}_t], \forall t$ and a super-martingale if $X_t \geq E[X_{t+1} \mid \mathcal{F}_t], \forall t$.

To formally define a martingle we should have:

- S_n is $\mathcal{F}_n$ measurable for all n;
- $E[|S_n|] < \infty$ for all n;
- $E[S_{n+1} \mid \mathcal{F}_n] = S_n$ for all n.

Example 4.1. Let X_n be a sequence of indendent r.v.'s with $E[X_n] = \mu_n$ and then

$$S_n = \sum_{i=1}^{n} (X_i - \mu_i)$$

and $S_0 = 0$ is a martingale process.

$$\begin{aligned} E[S_{n+1}|\mathcal{F}_n] &= E\left[\sum i = 1^{n+1}(X_i - \mu_i)|\mathcal{F}_n\right] \\ &= E\left[\sum i = 1^{n}(X_i - \mu_i) + (X_{n+1} - \mu_{n+1})|\mathcal{F}_n\right] \\ &= \sum i = 1^{n}(X_i - \mu_i) + E[X_{n+1} - \mu_{n+1}|\mathcal{F}_n] \\ &= S_n + E[Xn + 1 - \mu_{n+1}] \\ &= S_n + (E[X_{n+1}] - \mu_{n+1}) \\ &= S_n. \end{aligned}$$

Example 4.2. Let X_n be a sequence of i.i.d. r.v.'s with mean μ and variance σ^2 then:

$$Y_n = (\sum_{i=1}^{n} X_i - n\mu)^2 - n\sigma^2$$

and $Y_0 = 0$ is a martingale process.

We have: $Y_{n+1} = (\sum_{i=1}^{n} X_i - n\mu)^2 + 2(X_{n+1} - \mu)(\sum_{i=1}^{n} X_i - n\mu) + (X_{n+1} - \mu)^2 - (n+1)\sigma^2$ Taking expectation we have $E[X_{n+1} - \mu \mid \mathcal{F}_n] = 0$ and $E[(X_{n+1} - \mu)^2 \mid \mathcal{F}_n] = \sigma^2$ then we have

$$E[Y_{n+1} \mid \mathcal{F}_n] = (\sum_{i=1}^{n} X_i - n\mu)^2 - n\sigma^2 + \sigma^2 = Y_n$$

Example 4.3 (Wald's martingale)**.** Let X_n be a sequence of i.i.d. r.v.'s with $MGF : \phi(\theta) = E[e^{\theta X_1}]$ then:

$$Y_n = (\phi(\theta))^{-n} \cdot e^{\theta \sum_{i=1}^{n} X_i}$$

with $Y_0 = 1$ is a martingale process.

Similar to the above approach, we can write $E[Y_{n+1} \mid \mathcal{F}_n] = E[(\phi(\theta))^{-(n+1)} \cdot e^{\theta \sum_{i=1}^{n+1} X_i} \mid \mathcal{F}_n]$ using updated property we have $E[e^{\theta X_{n+1}}] = \phi(\theta)$ then we have proved the result.

Chapter 5

Martingale

This chapter formally defines and complete the martingale process.

5.1 Conditional expectations, filtration and martingales

We can define $E[X \mid A] = \frac{E[X1\{A\}]}{P(A)}$ given a random variable X and an event A.

Definition 5.1 (conditional expectation of X given $\mathcal{G}$)**.** Given Ω, two σ-fields $\mathcal{G} \subset \mathcal{F}$ on Ω, and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$. Suppose X is a random variable with respect to $\mathcal{F}$ but not necessarily with respect to $\mathcal{G}$, and suppose X has a finite L_1 norm that is $(E[|X|] < \infty)$. Then the conditional expectation of X given $\mathcal{G}$, denoted by $E[X|\mathcal{G}]$, is a random variable Y such that:

- Y is mesurable with respect to $\mathcal{G}$.
- For any $A \in \mathcal{G}$, we have $E[X1\{A\}] = E[Y1\{A\}]$.

We can write $Z \in \mathcal{F}$ to indicate that Z is measurable with respect to $\mathcal{F}$. Also let $\mathcal{F}(Z)$ denote the smallest σ-field such with respect to which Z is measurable

Theorem 5.1 (conditional expectation)**.** *The conditional expectation $E[X|\mathcal{G}]$ exists and is unique.*

Example 5.1. Consider the trivial case when $\mathcal{G} = \{\emptyset, \Omega\}$ We claim that the constant value $c = E[X]$ is $E[X|\mathcal{G}]$. This is because c is measurable with respect to $\mathcal{G}$, so (a) holds. For (b), we have $E[X1\{\Omega\}] = E[X] = c$ and $E[c1\{\Omega\}] = E[c] = c$; and $E[X1\{\emptyset\}] = 0$ and $E[c1\{\emptyset\}] = 0$.

Example 5.2. Given two random variables $X, Y : \Omega \to \mathbb{R}$ suppose both $\in \mathcal{F}$. Let $\mathcal{G} = \mathcal{G}(Y) \subset \mathcal{F}$ be the field generated by Y. We define $E[X|Y]$ to be $E[X|\mathcal{G}]$.

Proof. Given two probability measures $\mathbb{P}_1, \mathbb{P}_2$ defined on the same $(\Omega, \mathcal{F}), \mathbb{P}_2$ is defined to be absolutely continuous with respect to $\mathbb{P}_1$ if for every $A \in \mathcal{F}$, $\mathbb{P}_1(A) = 0$ implies $\mathbb{P}_2(A) = 0$. □

Theorem 5.2 (Radon-Nikodym Theorem)**.** *Suppose $\mathbb{P}_2$ is absolutely continuous with respect to $\mathbb{P}_1$. Then there exist a non-negative random variable $Y : \Omega \to \mathbb{R}_+$ such that for every $A \in \mathcal{F}$:*

$$\mathbb{P}_2(A) = E_{\mathbb{P}_1}[Y1\{A\}]$$

Function Y is called Radon-Nikodym (RN) derivative and sometimes is denoted $d\mathbb{P}_2/d\mathbb{P}_1$.

We now use this theorem to establish the existence of conditional expectations. Thus we have $\mathcal{G} \subset \mathcal{F}$, $\mathbb{P}$ is a probability measure on $\mathcal{F}$ and X is measurable with respect to $\mathcal{F}$. We will only consider the case $X \geq 0$ such that $E[X] < \infty$. We will assume that X is not constant, so that $E[X] > 0$. Consider a new probability measure $\mathbb{P}_2$ on $\mathcal{G}$ defined as follows:

$$\mathbb{P}_2(A) = \frac{E_{\mathbb{P}}[X1\{A\}]}{E_{\mathbb{P}}[X]}, A \in \mathcal{G}$$

where we write $E_{\mathbb{P}}$ in place of E to emphasize that the expectation operator is with respect to the original measure $\mathbb{P}$. We claim that $\mathbb{P}_2$ is absolutely continuous with respect to $\mathbb{P}$. By the Radon-Nikodym THeorem then there exists Z which is measurable with respect to $\mathcal{G}$ such that for any $A \in \mathcal{G}$:

$$\mathbb{P}_2(A) = E_{\mathbb{P}}[Z1\{A\}]$$

We now take $Y = ZE_{\mathbb{P}}[X]$. Then Y satisfies the condition (b) of being a conditional expectation, since for every set B:

$$E_{\mathbb{P}}[Y1\{B\}] = E_{\mathbb{P}}[X]E_{\mathbb{P}}[Z1\{B\}] = E_{\mathbb{P}}[X1\{B\}]$$

The second part, corresponding to the uniqueness property is proved similarly to the uniqueness of the RN derivative.

Here are some additional properties of conditional expectations.

Linearity: $E[aX + bY|\mathcal{G}] = aE[X|\mathcal{G}] + bE[Y|\mathcal{G}]$

Monotonicity: If $X_1 \leq X_2$, then $E[X_1|\mathcal{G}] \leq E[X_2|\mathcal{G}]$

Independence: If X is independent from $\mathcal{G}$, then for every measurable $A \subset \mathbb{R}, B \in \mathcal{G}\mathbb{P}(\{X \in A\} \cap B) = \mathbb{P}(X \in A)\mathbb{P}(B)$. Then: $E[X|\mathcal{G}] = E[X]$

ensen's Inequality: let ϕ be a convex function and $E[|X|], E[|\phi(X)|] < \infty$, then $E[\phi(X)|\mathcal{G}] \geq \phi(E[X|\mathcal{G}])$

Tower property: Suppose $G_1 \subset G_2 \subset \mathcal{F}$. Then $E[E[X \mid \mathcal{G}_1] \mid \mathcal{G}_2] = E[X \mid \mathcal{G}_1]$ and $E[E[X \mid \mathcal{G}_2] \mid \mathcal{G}_1] = E[X \mid \mathcal{G}_1]$. That is the smaller field wins.

Proof. We first prove the conditional Jensen's Inequality part:

We use the following representation of a convex function, let:

$$A = \{(a, b) \in \mathbb{Q} : ax + b \leq \phi(x)\forall x\}.$$

Then $\phi(x) = \sup_{ax+b:(a,b)\in A}$. now we prove the Jensen's inequality. For any pair of rationals $a, b \in \mathbb{Q}$ satisfying the bound above, we have, by monotonicity that $E[\phi(X)|\mathcal{G}] \geq aE[X|\mathcal{G}] + b$, a.s., implying $E[\phi(X)|\mathcal{G}] \geq \sup\{aE[X|\mathcal{G}] + b : (a, b) \in A\} = \phi(E[X|\mathcal{G}])$ a.s.

Prove the Tower Property.

By definition $E[X|\mathcal{G}_1]$ is $\mathcal{G}_1$ measurable. Therefore, it is $\mathcal{G}_2$ measurable. Then the first equality follows from the fact $E[X|\mathcal{G}] = X$ when $X \in \mathcal{G}$ which we established earlier. Now fix any $A \in \mathcal{G}_1$. Denote $E[X|\mathcal{G}_1]$ by Y_1 and $E[X|\mathcal{G}_2]$ by Y_2. Then $Y_1 \in \mathcal{G}_1, Y_2 \in \mathcal{G}_2$. Then:

$$E[Y_1 1\{A\}] = E[X1\{A\}]$$

simply by the definition of $Y_1 = E[X|\mathcal{G}_1]$. On the other hand, we also have $A \in \mathcal{G}_2$. Therefore,

$$E[X1\{A\}] = E[Y_2 1\{A\}]$$

Combining the two equalities we see that $E[Y_2 1\{A\}] = E[Y_1 1\{A\}]$ for every $A \in \mathcal{G}_1$. Therefore, $E[Y_2|\mathcal{G}_1] = Y_1$, which is the desired result. □

5.1.1 Filtration

A family of σ-fileds $\{\mathcal{F}_t\}$ is defined to be a filtration if $\mathcal{F}_{t_1} \subset \mathcal{F}_{t_2}$ whenever $t_1 \leq t_2$. A stochastic process $\{X_t\}$ is said to be adapted to filtration $\{F_t\}$ if $X_t \in \mathcal{F}_t$ for every t.

Definition 5.2 (Filtration)**.** A stochastic process $\{X_t\}$ adapted to a filtration $\{\mathcal{F}_t\}$ is defined to be a martingale if:

- $E[|X_t|] < \infty, \forall t.$
- $E[X_t|\mathcal{F}_s] = X_s$ for all $s \leq t$.

When equality is substituted with $\leq$, the process is called super-martingale. When it is substituted with $\geq$, the proecss is called sub-martingale.

Suppose we have a stochastic process $\{X_t\}$ adapted to filtration $\{\mathcal{F}_t\}$ and suppose for some $s' < s < t$ we have $E[X_t|\mathcal{F}_s] = X_s$ and $E[X_s|\mathcal{F}_{s'}] = X_{s'}$. Then using Tower Property of conditional expectations:

$$E[X_t|\mathcal{F}_{s'}] = E[E[X_t|\mathcal{F}_s]|\mathcal{F}_{s'}] = E[X_s|\mathcal{F}_{s'}] = X_{s'}$$

This means that when the stochastic process $\{X_n\}$ is discrete time it suffices to check $E[X_{n+1}|\mathcal{F}_n] = X_n$ for all n.

Example 5.3. Let X_n be an i.i.d. sequence with mean μ and variance $\sigma^2 < \infty$. Let $\mathcal{F}_n$ be the borel σ-algebra on $\mathbb{R}^n$. then $S_n - n\mu = \sum_{0 \leq k \leq n} X_k - n\mu$ is a martingale. Indeed S_n is adapted to $\mathcal{F}_n$ and:

$$\begin{aligned} E[S_{n+1} - (n+1)\mu|\mathcal{F}_n] &= E[X_{n+1} - \mu + S_n - n\mu|\mathcal{F}_n] = E[X_{n+1} - \mu|\mathcal{F}_n] + E[S_n n\mu|\mathcal{F}_n] \\ &= E[X_{n+1} - \mu] + S_n - n\mu = S_n - n\mu. \end{aligned}$$

Here X_{n+1} is independent from $\mathcal{F}_n$ and $S_n \in \mathcal{F}_n$.

Remark 5.1. The martingale assumes the mean value is finite, and this is also consistent with the real world application.

Lemma 5.1. *Recall from the Law of Total Expectation: $E[X] = E[E[X|Y]]$. the result holds, on a conditional space as well: if we have the probability measure is the conditional probability space given some Z then:*

$$E[X|Z] = E[E[X|Y, Z]|Z]$$

Example 5.4. Suppose $Y_1, Y_2, \ldots$ is a sequence of observations. Then for any random quantity of interest X, the process $X_n = E[X|Y_1, \ldots, Y_n]$ is a martingale.

Proof. $E[X_{n+1}|Y_1, \ldots, Y_n] = E[E[X|Y_1, \ldots, Y_n, Y_{n+1}]|Y_1, \ldots, Y_n] = E[X|Y_1, \ldots, Y_n] = X_n$. □

Example 5.5 (Doob martingale)**.** The martingale $X_n = E[X|Y_1, \ldots, Y_n]$ is known as Doob Martingale. It is well known from prediction theory that, given $Y_1, \ldots, Y_n$, the best prediction of a random variable X is $E[X|Y_1, \ldots, Y_n]$.

Example 5.6. Suppose $X_1, X_2, \ldots$ is a sequence of random variables, neither independent nor identically distributed, then:

$$Y_n = \sum_{i=1}^{n} (X_i - E[X_i \mid X_1, \ldots, X_{i-1}])$$

and $Y_0 = 0$ is a martingale.

Proof. We can check $Y_{n+1} = Y_n + X_{n+1} - E[X_{n+1} \mid Y_1, \ldots, Y_n]$ and then it is easy to verify that it satisfied the martingale. □

Definition 5.3 (martingale differencen)**.** If martingale X_n is written as: $X_n = X_0 + \sum_{i=1}^{n} Z_i$, then Z_i is called the martingale difference. Any martingale can be represented in the form of martingale difference $Z_i = X_i - X_{i-1}, i = 1, 2, \ldots$.

Martingale difference satisfies, for all $n \geq 0$, by $X_{n+1} = X_n + Z_{n+1}$ and by $E[X_{n+1}|\mathcal{F}_n] = X_n$, then $E[Z_{n+1}|\mathcal{F}_n] = 0$.

Thus $E[Z_n] = 0$ for all n. Furthermore, for any $i < j$, $E[Z_i Z_j] = 0$. Because $E[Z_i Z_j] = E[E[Z_i Z_j|\mathcal{F}_{j-1}]] = E[Z_i E[Z_j \mathcal{F}_{j-1}]] = 0$.

Thus, martingale differences are uncorrelated, though they may not be independent. This shows that, martingale process is a generalization of sum of i.i.d. random variables.

Proposition 5.1. *Suppose X_n is a martingale and ϕ is convex function such that $E[|\phi(X_n)|] < \infty$, then $\phi(X_n)$ is a sub-martingale.*

Proof. We apply conditional Jensen's inequality:

$$E[\phi(X_{n+1})|\mathcal{F}_n] \geq \phi(E[X_{n+1}|\mathcal{F}_n]) = \phi(X_n)$$

□

We obtain that if X_n is a martingale and $E[|X_n|^p] < \infty$ for all n for some $p \geq 1$, then $|X_n|^p$ is a sub-martingale. Note that if X_n was sub-martingale and ϕ was non-decreasing, then the same result applies.

Definition 5.4 (particular random variable)**.** A sequence of random variables H_n is defined to be particular if $H_n \in \mathcal{F}_{n-1}$.

Let X_n be adapted to filtration $\mathcal{F}_n$. Then $H_n = X_{n-1}$, $H_0 = H_1 = X_0$ is predictable. Consider $H_n = E[X_n|\mathcal{F}_{n-1}]$. By the definition of conditional expectations, $H_n \in \mathcal{F}_{n-1}$, so it is predictable.

Theorem 5.3 (Doob's decomposition)**.** *Every sub-martingale $X_n, n \geq 0$ adapted to filtration $\mathcal{F}_n$, can be written in a unique way as $X_n = M_n + A_n$, where M_n is a martingale and A_n is an a.s. non-decreasing sequence, predictable with respect to $\mathcal{F}_n$.*

Proof. Set $A_0 = 0$ and define A_n recursively by $A_n = A_{n-1} + E[X_n|\mathcal{F}_{n-1}] - X_{n-1} \geq A_{n-1}$. Let $M_n = X_n - A_n$. By induction, we have that $A_n \in \mathcal{F}_{n-1}$ since $A_{n-1} \in \mathcal{F}_{n-2} \subset \mathcal{F}_{n-1}$ and $E[X_n|\mathcal{F}_{n-1}], X_{n-1} \in \mathcal{F}_{n-1}$. Therefore, A_n is predictable. we now need to show that M_n is martingale. To check that $E[|M_n|] < \infty$, it suffices to show the same for A_n, since by assumption X_n is a sub-martingale and therefore, $E[|X_n|] < \infty$. Now we establish finiteness of $E[|A_n|]$ by induction, for which it suffies to have finiteness of $E[|A_n|]$ by induction, for which it suffices to have finiteness of $E[|E[X_n|\mathcal{F}_{n-1}]|]$ which follows by conditional Jensen's inequality which gives $|E[X_n|\mathcal{F}_{n-1}]| \leq E[|X_n||\mathcal{F}_{n-1}]$ and the tower property which gives $E[E[|X_n||\mathcal{F}_{n-1}]] = E[|X_n|] < \infty$. We now establish that $E[M_n|\mathcal{F}_{n-1}] = M_{n-1}$. we have:

$$\begin{aligned} E[M_n \mid \mathcal{F}_{n-1}] &= E[X_n - A_n \mid \mathcal{F}_{n-1}] = E[X_n \mid \mathcal{F}_{n-1}] - A_n \\ &= X_{n-1} - A_{n-1} = M_{n-1}. \end{aligned}$$

Thus M_n is indeed a martingale. This completes the proof of existence part.

To prove uniqueness, we assume that $X_n = M_n' + A_n'$ is any such decomposition. Then:

$$E[X_n|\mathcal{F}_{n-1}] = E[M_n' + A_n'|\mathcal{F}_{n-1}] = M_{n-1}' + A_n' = X_{n-1} - A_{n-1}' + A_n'$$

since by assumption, M_n' is a martingale and A_n' is predictable. Then we see that A_n' satisfies the same recursion as A_n, implying $A_n' = A_n$. Then $M_n = M_n'$. □

Theorem 5.4 (predictivity under supermartingale). *Suppose X_n is a super-martingale and $H_n \geq 0$ is predictable. Then $Z_n = \sum_{1\leq m\leq n} H_m(X_m - X_{m-1})$ is also a super-martingale.*

Proof. We have:

$$E[Z_{n+1} \mid \mathcal{F}_n] = E[H_{n+1}(X_{n+1} - X_n) \mid \mathcal{F}_n] + E[\sum_{1\leq m\leq n} H_m(X_m - X_{m-1}) \mid \mathcal{F}_n].$$

Since H is predictable, then $H_{n+1} \in \mathcal{F}_n$ implying that the first summand is equal to:

$$H_{n+1}E[(X_{n+1} - X_n) \mid \mathcal{F}_n] \leq 0$$

where inequality follows since X_n is super-martingale. On the other hand,

$$\sum_{1\leq m\leq n} H_m(X_m - X_{m-1}) = Z_n \in \mathcal{F}_n$$

implying that its expectation is $E[Z_n \mid \mathcal{F}_n] = Z_n$. □

5.1.2 Hoeffding Inequality

Theorem 5.5 (Azuma-hoeffding inequality). *If $X_0, X_1, X_2, \ldots$ is a martingale with $X_0 = \mu$ and there exists $a_i \leq b_i$ such that:*

$$a_i \leq X_i - X_{i-1} \leq b_i, i = 1, 2, \ldots$$

Then for any $\varepsilon > 0$,

$$P(X_n - \mu > \varepsilon) \leq e^{-2\varepsilon^2/\sum_{i=1}^n (b_i - a_i)^2}$$

Thus Hoeffding inequality extends to martingale process.

Recall from the MGF function, we have if X has support on $[a, b]$ and has mean 0, then:

$$E[e^{\theta X_1}] \leq e^{\theta^2(b-a)^2/8}$$

We can prove this using: $e^{\theta X} \leq \frac{b-a}{b-a}e^{\theta a} + \frac{X-a}{b-a}e^{\theta b}$ then: we have: $E[e^{\theta X}] \leq \frac{b}{b-a}e^{\theta a} - \frac{a}{b-a}e^{\theta b} = e^{\theta a + \log\left(1 + \frac{a}{b-a} - \frac{a}{b-a}e^{\theta(b-a)}\right)}$

Now we prove the result:

Proof. WLOG, we assume $\mu_i = 0$ for all i, and then for any $\theta > 0$ we have:

$$\begin{aligned}
P(X_n > \varepsilon) = P(e^{\theta X_n} > e^{\theta\varepsilon}) &\leq -e^{\theta\varepsilon}E[e^{\theta X_n}] \\
&= e^{-\theta\varepsilon}E[E[e^{\theta X_n} \mid X_1, \ldots, X_{n-1}]] \\
&\leq e^{\theta\varepsilon}E[e^{\theta X_{n-1}}E[e^{\theta Z_n} \mid X_1, \ldots, X_{n-1}]] \\
&\leq e^{\theta\varepsilon}E[e^{\theta Z_{n-1}}]e^{\theta^2(b_n - a_n)^2/8} \\
&\leq e^{-\theta\varepsilon + \theta^2\sum_{i=1}^n (b_i - a_i)^2/8}
\end{aligned}$$

Choosing $\theta = 4\varepsilon/\sum_{i=1}^n (b_i - a_i)^2$ yields:

$$P(X_n > \varepsilon) \leq e^{-2\varepsilon^2/\sum_{i=1}^n (b_i - a_i)^2}$$

□

Proposition 5.2 (Mcdiarmid Inequality). *With the machinery of martingale, we are now ready to prove McDiarmid inequality, suppose $X_1, X_2, \ldots$ are independent random variables. If g satisfies:*

$$\sup_{x_1, x_2, \ldots, x_n, x_i'} |g(x_1, \ldots, x_i, \ldots, x_n) - g(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

Then we have:

$$P(g(X_1, \ldots, X_n) - E[g(X_1, \ldots, X_n)] > \varepsilon) \leq e^{\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}}$$

Proof. For $i = 1, \ldots, n$, consider Doob martingale:

$$Y_i = E[g(X_1, \ldots, X_n)|X_1, \ldots, X_i]$$

Clearly, $g(X_1, \ldots, X_n) = E[g(X_1, \ldots, X_n)] + \sum_{i=1}^n (Y_i - Y_{i-1})$. □

Note that, for any $X_1 = x_1, \ldots, X_i = x_i$, by independence of X_i's

$$\begin{aligned} Y_i - Y_{i-1} &= E[g(X_1, \ldots, X_{i-1}, X_i, \ldots, X_n)|X_1, \ldots, X_i] \\ &- E[g(X_1, \ldots, X_{i-1}, X_i, \ldots, X_n)|X_1, \ldots, X_{i-1}] \\ &= E[g(x_1, \ldots, x_{i-1}, x_i, \ldots, X_n)] - E[g(x_1, \ldots, x_{i-1}, X_i, \ldots, X_n)] \\ &\leq c_i. \end{aligned}$$

The above is true for any $X_1, \ldots, X_i$. It follows from Azuma-Hoeffding inequality that:

$$P(g(X_1, \ldots, X_n) - E[g(X_1, \ldots, X_n)] > \varepsilon) \leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}}$$

5.2 Stopping Time

Definition 5.5 (stopping time). Given a filtration $\{\mathcal{F}_t\}_{t \in T}$ on a sample space Ω, a random variable $\tau : \Omega \to T$ is called a stopping time, if the event $\{\tau \leq t\} = \{\omega \in \Omega : \tau(\omega) \leq t\} \in \mathcal{F}_t$ for every t.

Definition 5.6 (stopping time (alternative definition)). Let $X_1, X_2, \ldots$ be a stochastic sequence. An integer valued random variable N is called a stopping time with respect to $X_1, X_2, \ldots$ if the event $\{N = n\}$ is completed determined by $X_1, X_2, \ldots, X_n$.

Typically, when we discuss random variables, we only consider r.v.'s that are a.s. finite, or $P(X < \infty) = 1$. Such random variables are called "regular". By convention, when discussing stopping times, we usually include the cases that can take "infinity", i.e., we allow $P(N < \infty) < 1$.

Example 5.7. Let $P(X_n = 1) = \frac{1}{2} = 1 - P(X_n = -1)$ and let $S_n = \sum_{i=1}^n X_i$ and $S_0 = 0$ and define: $N = \inf\{n : S_n = 10\}$. N is a stopping time.

Consider the filtration corresponding to a sequence of random variables $X_1, \ldots, X_{,} \ldots$ Namely $\Omega = \mathbb{R}^\infty$ and $\mathcal{F}_n$ is Borel σ-field of $\mathbb{R}^n$. Fix some real value x. Given any sample $\omega = (\omega_1, \ldots, \omega_n, \ldots) \in \mathbb{R}^\infty$ define:

$$\tau(\omega) = \inf\{n : \sum_{1 \leq k \leq n} \omega_k \geq x\}$$

Namely, $\tau(\omega)$ is the smallest index at which the sum of the components is at least x. Then τ is a stopping time. Indeed, the event $\tau \leq n$ is completely specified by the portion $\omega_1, \ldots, \omega_n$ of the sample. In particular,

$$\{\omega : \tau(\omega) \leq n\} = \bigcup_{1\leq k\leq n} \{\omega : \sum_{1\leq i\leq k} \omega_i \geq x\}$$

Each of the event in the union on the right hand side is measurable with respect to $\mathcal{F}_n$.

Theorem 5.6 (non-positive expectation under supermartingale). *Suppose X_n is a super-martingale and τ is a stopping time, which is a.s. bounded: $\tau \leq M$ a.s. for some M. Then $E[X_\tau] \leq E[X_0]$. In other words, if there exists a bound on the number of rounds for betting, then the expected net gain is non-positive, provided that in each round the expected gain is non-positive.*

This theorem will be established as a result of several short lemmas.

Lemma 5.2. *Suppose τ is a stopping time corresponding to the filtration $\mathcal{F}_n$. Then the sequence of random variables $H_n = 1\{\tau \geq n\}$ is predictable.*

Proof. H_n is a random variable which takes value 0 and 1. Note that the event $\{H_n = 0\} = \{\tau < n\} = \{\tau \leq n-1\}$. Since τ is a stopping time, then the event $\{\tau \leq n-1\} \in \mathcal{F}_{n-1}$. Thus H_n is predictable. □

Corollary 5.7. *Suppose X_n is super-martingale and τ is a stopping time. Then $Y_n = X_{\min(n,\tau)}$ is also a super-martingale*

Proof. Define $H_n = 1\{\tau \geq n\}$. Observe that:

$$\sum_{1\leq m\leq n} H_m(X_m - X_{m-1}) = -H_0X_0 + \sum_{0\leq m\leq n-1} X_m(H_m - H_{m+1}) + H_nX_n.$$

Note, $H_0 = 1\{\tau \geq 0\} = 1.H_m - H_{m+1} = 1\{\tau \geq m\} - 1\{\tau \geq m+1\} = 1\{\tau = m\}$. Therefore, the expression on the right hand side of the above equation is equal to $X_{\min(n,\tau)} - X_0$. The left hand side of the equation is a super-martingale. We conclude that $Y_n = X_{\min(n,\tau)}$ is a super-martingale. □

Proof. Proof of the Theorem: The process $Y_n = X_{\min(n,\tau)}$ is a super-martingale by corollary. Therefore, $E[Y_M] \leq E[Y_0]$.

But $Y_M = X_{\min(M,\tau)} = X_\tau$ and $Y_0 = X_{\min(0,\tau)} = X_0$. we conclude $E[X_\tau] \leq E[X_0]$. □

Remark 5.2. We know that if N is stopping time, then $N+1$ is also a stopping time.

Remark 5.3. Suppose $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. We know that, for given number n, we have: $E[\sum_{i=1}^n X_i] = nE[X_1]$. Let N be a stopping time, do we have $E[\sum_{i=1}^N X_i] = E[N]E[X_i]$?

Here is a counter-example:

Let $P(X_n = 1) = \frac{1}{2} = 1 - P(X_n = -1)$ and define $S_n = \sum_{i=1}^n X_i$ and $N = \inf\{n : S_n = 10\}$. we do not have $E[\sum_{i=1}^N X_i] = E[N]E[X_1]$.

Proposition 5.3 (Wald's equation). *Let $X_1, X_2, \ldots$ be i.i.d. with $E[|X_1|] < \infty$ and N is a stopping time with $E[N] < \infty$, then:*

$$E[\sum_{i=1}^N X_i] = E[N]E[X_1]$$

Proof. If X_n is such that $\sum_{n=1}^\infty E[|X_n|] < \infty$ then: $E[\sum_{n=1}^\infty X_n] = \sum_{n=1}^\infty E[X_n]$ This is called Fibini's theorem.

$$\begin{aligned} E[\sum_{i=1}^{N} X_i] &= E[\sum_{i=1}^{\infty} X_i 1[i \le N]] = \sum_{i=1}^{\infty} E[X_i 1[i \le N]] \\ &= \sum_{i=1}^{\infty} E[E[X_i 1[i \le N]|\mathcal{F}_{i-1}]] = \sum_{i=1}^{\infty} E[1[i \le N]E[X_i|\mathcal{F}_{i-1}]] \\ &= \sum_{i=1}^{\infty} E[1[i \le N]]E[X_1] = E[\sum_{i=1}^{\infty} 1[i \le N]]E[X_1] = E[N]E[X_1] \end{aligned}$$

□

The previous result is also known as Wald's first equation. There is a Wald's seond equation: If X_i's are i.i.d. with mean 0 and finite variance, $E[X_1^2] < \infty$ and N is a stopping time with $E[N] < \infty$. There is

$$Var(\sum_{i=1}^{N} X_i) = E[N]Var(X_1)$$

Definition 5.7 (stopped time)**.** Let N be a stopping time with respect to $X_1, X_2, \ldots$ and n is given integer, then:

$$n \wedge N := \min\{n, N\}$$

is called a stopped time.

The process $\{n \wedge N : n = 1, 2, \ldots\}$ stops at N.

Claim 5.1.

- *Given n, stopping time is a stopped time.*
- *Need to show that, for $k = 1, 2, \ldots, n \wedge N = k$ is determined by the process $X_1, \ldots, X_k$;*
- *if $k < n$ then it is equivalent to $N = k$, so it is ...*
- *If $k \ge n$, then it is equivalent to $N \ge n$, hence it is ...*

5.2.1 Doob-Kolmogorov Inequality

Theorem 5.8 (Second stopping theorem)**.** *Suppose X_n is a super-martingale that is uniformly bounded. That is $|X_n| \le M, a.s.$, for some M. Suppose τ is a stopping time. Then $E[X_\tau] \le X_0$. If in addition X_n is a martingale, then $E[X_\tau] = E[X_0]$*

The gambling interpretation of this theorem is as follows: suppose we tried to use the "double the stakes" algorithm, which we know guarantees winning a dollar, when there are no restrictions. But now suppose that there is a limit on how "negative" we can go. Say this limit is M. Consider a modified process $Y_n = X_{n\wedge\tau}$. Then from our description $-M \le Y_n \le 1 < M$. Also we remember from the previous lecture that Y_n is a super-martingale. That it is a bounded super-martingale. This theorem then tells us that $E[Y_\tau] = E[X_\tau] \le E[X_0]$, namely the scheme does not work anymore.

Proof. Observe that $E[|X_\tau|] \le M < \infty$. Consider $Y_n = X_{n\wedge\tau}$. Then Y_n converges to X_τ a.s. as $n \to \infty$. Since $|Y_n| \le M$ a.s. then using the bounded Convergence Theorem, $\lim_{n\to\infty} E[Y_n] = E[X_\tau]$. On the other hand, we established in Corollary in the previous lecture, that Y_n is super-martingale. Therefore $E[Y_n] \le E[Y_0] = E[X_0]$. Combining together, we have $E[X_\tau] \le E[X_0]$ □

Theorem 5.9 (Doob-Kolmogorov inequality). *Suppose X_n is a non-negative sub-martingale adapted to $\{\mathcal{F}_n\}$ and $\varepsilon > 0$ Then for every $n \in \mathbb{N}$*

$$P(\max_{1\leq m\leq n} X_n \geq \varepsilon) \leq \frac{E[X_n^2]}{\varepsilon^2}$$

If X_n is a martingale, then the non-negativity condition can be dropped.

The convenience of this result is that we can bound the worst case deviation of a submartingale using its value at the end of time interval.

Proof. Using Jensen's inequality we established that if X_n is a martingale then $|X_n|$ is a sub-martingale. Since $|X_n|$ is non-negative, the second part follows from the first.

To establish the first part, consider the events $A = \{\max_{1\leq m\leq n} X_m \leq \varepsilon\}$ and $B_m = \{\max_{1\leq i\leq m-1} X_i \leq \varepsilon, X_m > \varepsilon\}$. Namely, A is the event that the sub-martingale never exceeds ε and B_m is the event that it does so at time m for the first time. We have $\Omega = A \cup \bigcup_{1\leq m\leq n} B_m$ and the event A, B_m are mutually exclusive. Then:

$$E[X_n^2] = E[X_n^2 1\{A\}] + \sum_{1\leq m\leq n} X[X_n^2 1\{B_m\}] \geq \sum_{1\leq m\leq n} E[X_n^2 1\{B_m\}].$$

Note

$$\begin{aligned} E[X_n^2 1\{B_m\}] &= E[(X_n - X_m + X_m)^2 1\{B_m\}] \\ &= E[(X_n - X_m)^2 1\{B_m\}] + 2E[(X_n - X_m)X_m 1\{B_m\}] + E[X_m^2 1\{B_m\}] \end{aligned}$$

The first of the summands is non-negative. The last is at least $\varepsilon^2 P(B_m)$, since the event B_m we have $X_m > \varepsilon$. We now analyze the second term and here we use the tower property:

$$\begin{aligned} E[(X_n - X_m)X_m 1\{B_m\}] &= E[E[(X_n - X_m)X_m 1\{B_m\}|\mathcal{F}_m]] \\ &= E[X_m 1\{B_m\} E[(X_n - X_m)|\mathcal{F}_m]] \geq 0 \end{aligned}$$

where the second equality follows since $1\{B_m\} \in \mathcal{F}_m$ and the last inequality follows since X_n is a sub-martingale and $X_m 1\{B_m\} \geq \varepsilon > 0$ on $\omega \in B_m$ and $= 0$ on $\omega \notin B_m$. we conclude that:

$$E[X_n^2] \geq \sum_{1\leq m\leq n} \varepsilon^2 P(B_m) = \varepsilon^2 P(\bigcup_m B_m) = \varepsilon P(\max_{1\leq m\leq n} X_n > \varepsilon)$$

□

Corollary 5.10 (extension of Doob-Kolmogorov inequality). *Suppose X_n is a martingale and $p \geq 1$. Then for every $\varepsilon > 0$*

$$P(\max_{1\leq m\leq n} |X_n| \geq \varepsilon) \leq \frac{E[|X_n|^p]}{\varepsilon^p}$$

Proof. The proof of the general caase is more complicated, but when $p \geq 2$ we almost immediately obtain the result. Using conditional Jensen's inequality we know that $|X_n|$ is a sub-martingale as $|\cdot|$ is a convex function. It is also non-negative. Function $x^{\frac{p}{2}}$ is convex increasing when $p \geq 2$ and $x \geq 0$. Recall from the previous lecture that this implies $|X_n|^{\frac{p}{2}}$ is also a sub-martingale. Applying Doob-Kolmogorov inequality we obtain:

$$P(\max_{1\leq m\leq n} |X_n| \geq \varepsilon) = P(\max_{1\leq m\leq n} |X_n|^{\frac{p}{2}} \geq \varepsilon^{\frac{p}{2}}) \leq \frac{E[|X_n|^p]}{\varepsilon^p}$$

□

Theorem 5.11 (generalized Doob-Kolmogorov inequality). *Suppose* $\{X_t\}_{t\in\mathbb{R}_+}$ *is a martingale which has a.s. continuous sample paths. Then for every* $p \geq 1, T > 0, \varepsilon > 0$

$$P(\sup_{0\leq t\leq T} |X_t| \geq \varepsilon) \leq \frac{E[|X_T|^p]}{\varepsilon^p}$$

Chapter 6

Martingale Convergence Theorem

6.1 Naive Version Introduction

Definition 6.1 (stopped process)**.** Let N be a stopping time, and n is given integer, then $X_{n\wedge N}$ is called a stopped process. In other words, the process $\{X_{n\wedge N} : n = 1, 2, \ldots\}$ stops at X_N.

Proposition 6.1. *Stopped martingale is a martingale. That is, if $\{X_n : n \geq 0\}$ is a martingale and N a stopping time, then, $\{X_{n\wedge N} : n \geq 0\}$ is also a martingale. Thus we always have $E[X_{n\wedge N}] = E[X_{0\wedge N}] = E[X_0]$*

Proof.

$$\begin{aligned} E[X_{(n+1)\wedge N}|\mathcal{F}_n] &= E[X_{(n+1)\wedge N}1[N \leq n]|\mathcal{F}_n] + E[X_{(n+1)\wedge N}1[N > n]|\mathcal{F}_n] \\ &= E[X_N 1[N \leq n]|\mathcal{F}_n] + E[X_{n+1}1[N > n]|\mathcal{F}_n] \\ &= E[X_{N\wedge n}1[N \leq n]|\mathcal{F}_n] + 1[N > n]E[X_{n+1}|\mathcal{F}_n] \\ &= X_{N\wedge n}1[N \leq n] + 1[N > n]X_n \\ &= X_{N\wedge n}1[N \leq n] + 1[N > n]X_{n\wedge N} \\ &= X_{n\wedge N} \end{aligned}$$

□

Thus we always have $E[X_{N\wedge n}] = E[X_0]$. If N is a stopping time that can not take value ∞, then $X_{n\wedge N}$ converges to X_N a.s. This means that, whether we have $E[X_N] = E[X_0]$ or not depends on whether we can interchange mean and limit:

$$\lim_{n\to\infty} E[X_{n\wedge N}] = E[\lim_{n\to\infty} X_{n\wedge N}].$$

Recall the conditions for a.s. convergence implying L^1 convergence. Then we can find the folloinwg theorem:

Theorem 6.1 (Martingale Stopping Theorem)**.** *Suppose X_n is a martingale and N is a stopping time. Under any of the folloinwg conditions, we have $E[X_N] = E[X_0]$:*

- *$X_{N\wedge n}$ is bounded*
- *N is bounded*
- *$E[N] < \infty$ and $E[|X_n - X_{n-1}|]$ is bounded.*

Proof. Part (a) follows from bounded convergence theorem, that $X_{n\wedge N}$ is bounded.

Part (b) is trivial: $N \leq m$ for some number n^* such $X_{n\wedge N} = X_N$ when $n \geq n^*$, therefore, $E[X_{n\wedge N}] = E[X_N]$ when $n \geq n^*$;

Part (c) is an extension of Wald's equation, which we discuss in detail next.

we want to show that if (i) stopping time N has finite mean, and (ii) $X_n = X_0 + \sum_{i=1}^n Z_i$, where $E[|Z_i|]$ is bounded and $E[Z_i|\mathcal{F}_{i-1}] = 0$, then we have:

$$E[X_N] = E[X_0]$$

This is equivalent to, for martingale difference difference Z_i, it holds that:

$$E[\sum_{i=1}^N Z_i] = E[N]E[Z_1](= 0)$$

Proof is similar to that of Wald equation:

$$\begin{aligned} E[\sum_{i=1}^N Z_i] &= E[\sum_{i=1}^\infty Z_i 1[i \leq N]] = \sum_{i=1}^\infty E[Z_i 1[i \leq N]] \\ &= \sum_{i=1}^\infty E[E[Z_i 1[i \leq N]|\mathcal{F}_{i-1}]] = \sum_{i=1}^\infty E[1[i \leq N]E[Z_i|\mathcal{F}_{i-1}]] = 0 \end{aligned}$$

Thus martingale stopping theorem extends Wald's equation to the non-independent case. □

Remark 6.1. Martingale stopping theorem is also known as Optional Sampling Theorem. This same result extends to sub-martingale and super-martingale, i.e., under any of the three conditions, it holds that: $E[X_N] \geq E[X_0]$ for sub-martingale; $E[X_n] \leq E[X_0]$ for super-martingale.

Proposition 6.2. *From Jensen's inequality, we have $E[\phi(X)] \geq \phi(E[X])$ if ϕ is convex. It extends to conditional probability, i.e., for any Z we have:*

$$E[\phi(X)|Z] \geq \phi(E[X|Z])$$

result: If X_n is a martingale and ϕ is convex, then $Y_n = \phi(X_n)$ is a sub-martingale. This is because: $E[\phi(X_{n+1})|X_1, \ldots, X_n] \geq \phi(E[X_{n+1}|X_1, \ldots, X_n]) = \phi(X_n)$.

If X_n is a sub-martingale and $Y_n = \phi(X_n)$ be a sub-martingale, it may not hold because the function need to be increasing convex.

Lemma 6.1. *If $X_i, i \geq 0$ is a sub-martingale and N is a stopping time satisfying $1 \leq N \leq n$. Then*

$$E[X_1] \leq E[X_N] \leq E[X_n]$$

Proof. Since N is bounded, we have $E[X_N] \geq E[X_0]$. Now, for any $1 \leq k \leq n$, we condition on $N = k$,

$$\begin{aligned} E[X_n|X_1, \ldots, X_N, N = k] &= E[X_n|X_1, \ldots, X_k, N = k] = E[X_n|X_1, \ldots, X_k] \\ &\geq X_k = X_N \end{aligned}$$

Taking expectation on both sides gives: $E[X_n] \geq E[X_N]$. □

Theorem 6.2 (Doob's maximal inequality-formal edition)**.** *Suppose $X_n, n \geq 0$ is a non-negative sub-martingale. Then:*

$$P(\max\{X_1, \ldots, X_n\} > \varepsilon) \leq \frac{E[X_n]}{\varepsilon}$$

The result is called Kolmogorov's maximal inequality.

Proof. Define a stopping time, for give n, $N = \inf\{1 \leq i \leq n : X_i > \varepsilon\}$, where $\inf \emptyset = n$;

By the Lemma, we have: $P(\max\{X_1, X_2, \ldots, X_n\} > \varepsilon) = P(X_N > \varepsilon) \leq \frac{E[X_N]}{\varepsilon} \leq \frac{E[X_n]}{\varepsilon}$. □

Corollary 6.3. *Let $X_n, n \geq 0$ be a martingale. Then for any $\theta > 0$,*

$$P(\max\{|X_1|, \ldots, |X_n|\} > \varepsilon) \leq \frac{1}{\varepsilon} E[|X_n|]$$
$$P(\max\{X_1^2, \ldots, X_n^2\} > \varepsilon) \leq \frac{E[X_n^2]}{\varepsilon},$$
$$P(\max\{X_1, \ldots, X_n\} > \varepsilon) \leq e^{-\theta\varepsilon} E[e^{\theta X_n}]$$

Example 6.1. Let $X_n, n \geq 0$ be an independent sequence with mean μ_n and variance σ_n^2. Let: $S_n = \sum_{i=1}^n (X_i - \mu_i), n = 1, 2, \ldots,$ Then Kolmogorov inequality holds:

$$P(\max_{1 \leq i \leq n} |S_i| > \varepsilon) \leq \frac{\sum_{i=1}^n \sigma_i^2}{\varepsilon^2}$$

Example 6.2. Let $X_i, i \geq 1$ is an independent sequence with mean μ_i and variance σ_i^2 and bounded on $[a_i, b_i]$. Let:

$$S_n = \sum_{i=1}^n (X_i - \mu_i), n = 1, 2, \ldots$$

For any $\varepsilon > 0$, $P(\max_{1 \leq i \leq n} S_i > \varepsilon) \leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}}$

This is because for any $\theta > 0$, $P(\max_{1 \leq i \leq n} S_i > \varepsilon) = P(\max_{1 \leq i \leq n} e^{\theta S_i} > e^{\theta\varepsilon}) \leq e^{-\theta\varepsilon} E[e^{\theta S_n}]$

where the inequality follows from $e^{\theta S_n}$ is a sub-martingale and Doob's maximal inequality. The rest of the proof follows from the argument of standard Azuma-Hoeffding inequality and the choice of: $\theta = 4\varepsilon / \sum_{i=1}^n (b_i - a_i)^2$.

6.2 Doob's Inequality Revisited

Theorem 6.4 (Doob)**.** *Suppose X_n is a super-martingale which satisfies $\sup_n E[|X_n|] < \infty$ Then almost surely $X_\infty = \lim_n X_n$ exists and is finite in expectation. That is, define $X_\infty = \limsup X_n$. Then $X_n \to X_\infty$ and $E[|X_\infty|] < \infty$*

Proof. The proof relies "Doob's Upcrossing Lemma". For that consider

$$\begin{aligned} \Lambda &= \{\omega : X_n(\omega) \text{ does not converge to a limit in } \mathbb{R}\} \\ &= \{\omega : \liminf_n X_n(\omega) < \limsup_n X_n(\omega)\} \\ &= \bigcup_{a<b, a,b \in \mathbb{Q}} \{\omega : \liminf_n X_n(\omega) < a < b < \limsup_n X_n(\omega)\}, \end{aligned} \tag{6.1}$$

where $\mathbb{Q}$ is the set of rational values. Let $U_N[a, b](\omega) =$ largest k such that it satisfies the following: there exists

$$0 \leq s_1 < t_1 < \ldots < s_k < t_k \leq N$$

such that

$$X_{s_i}(\omega) < a < b < X_{t_i}(\omega), 1 \leq i \leq k$$

That is, $U_N[a,b]$ is the number of up-crossings of $[a,b]$ up to N. Clearly, $U_N[a,b](\omega)$ is non-decreasing in N. Let $N_\infty[a,b](\omega) = \lim_{N\to\infty} U_N[a,b](\omega)$ Then the equation 6.1 can be written as:

$$\Lambda = \bigcup_{a<b,a,b\in\mathbb{Q}} \{\omega : U_\infty[a,b](\omega) = \infty\} = \bigcup_{a<b;a,b\in\mathbb{Q}} \Lambda_{a,b}. \tag{6.2}$$

Doob's upcrossing lemma proves that $P(\Lambda_{a,b}) = 0$ for every $a < b$. Then we have from 6.2 that $P(\Lambda) = 0$. Thus, $X_n(\omega)$ converges in $[-\infty, \infty]$ a.s. That is,

$$X_\infty = \lim_n X_n \text{ exists a.s.}$$

now,

$$E[|X_\infty|] = E[\liminf_n |X_n|] \leq \liminf_n E[|X_n|] \leq \sup_n E[|X_n|] < \infty$$

where we have used Fatou's lemma in the first inequality. Thus X_∞ is in L_1. In particular, X_∞ is finite a.s. □

Lemma 6.2 (Doob's Upcrossing)**.** *Let X_n be a super-MG. Let $U_N[a,b]$ be the number of upcrossing of $[a,b]$ until time N with $a < b$/ Then*

$$(b-a)E[U_N[a,b]] \leq E[(X_N - a)^-]$$

where

$$(X_N - a)^- = \begin{cases} a - X_N & \text{if } X_N \leq a \\ 0 & \text{otherwise} \end{cases}$$

Proof. Define a predictable sequence C_n as follows:

$$C_1(\omega) = \begin{cases} 1 & \text{if } X_0(\omega) < a \\ 0 & \text{otherwise} \end{cases}$$

Inductively,

$$C_n(\omega) = \begin{cases} 1 & \text{if } C_{n-1}(\omega) = 1 \text{ and } X_{n-1}(\omega) \leq b \\ 1 & \text{if } C_{n-1}(\omega) = 0 \text{ and } X_{n-1}(\omega) < a \\ 0 & \text{otherwise} \end{cases}$$

By definition, C_n is predictable. The sequence C_n has the following property:

If $X_n < a$ then $C_1 = 1$. Then the sequence C_n remains one until the first time X_n exceeds b. It then remains zero until the first time it becomes smaller than a at which time it switches back to 1, etc. If instead $X_0 > a$, then $C_1 = 0$ and it remains zero until the first time X_n becomes smaler than a, at which point C_n switches to 1 and then continues as above. Consider:

$$Y_n = (C \cdot X)_n = \sum_{1\leq k\leq n} C_k(X_k - X_{k-1})$$

We claim that:

$$Y_N(\omega) \geq (b-a)U_N[a,b] - (X_N(\omega)-a)^-$$

let $U_N[a,b] = k$. Then there is $0 \leq s_1 < t_1 < \ldots < s_k < t_k \leq N$ such that $X_{s_i}(\omega) < a < b < X_{t_i}(\omega), i = 1, \ldots, k$. By definition $C_{s_i+1} = 1$ for all $i \geq 1$. Further, $C_t(\omega) = 1$ for $s_i + 1 \leq t \leq l_i \leq t_i$ where $l_i \leq t_i$ is the smallest time $t \geq s_i$ such that $X_t(\omega) > b$. Without loss of generality, assume that $s_1 = \min\{n : X_n < a\}$. Let $s_{k+1} = \min\{n > t_k : X_n(\omega) < a\}$. Then

$$\begin{aligned}
Y_N(\omega) &= \sum_{j\leq N} C_j(\omega)(X_j(\omega) - X_{j-1}(\omega)) = \sum_{1\leq i\leq k} [\sum_{s_i\leq t\leq l_i} C_{t+1}(X_{t+1}(\omega) - X_t(\omega))] \\
&= \sum_{t>s_{k+1}} C_{t+1}(\omega)(X_{t+1}(\omega) - X_t(\omega)) \text{ (Because otherwise } C_t(\omega) = 0.) \\
&= \sum_{1\leq i\leq k} (X_{l_i}(\omega) - X_{s_i}(\omega)) + X_N(\omega) - X_{s_{k+1}}(\omega)
\end{aligned}$$

where the term $X_N(\omega) - X_{s_{k+1}}(\omega)$ is defined to be zero if $s_{k+1} > N$. Now $X_{l_i}(\omega) - X_{s_i}(\omega) \geq b - a$. now if $X_N(\omega) \geq X_{s_{k+1}}$ then:

$$X_N(\omega) - X_{s_{k+1}} \geq 0$$

Otherwise,

$$|X_N(\omega) - X_{s_{k+1}}(\omega)| \leq |X_N(\omega) - a|.$$

Therefore, we have

$$Y_N(\omega) \geq U_N[a,b](b-a) - (X_N(\omega)-a)^-$$

as claimed.

Now, as we have established earlier, $Y_n = (C \cdot X)_n$ is super MG since $C_n \geq 0$ is predictable. That is,

$$E[Y_N] \leq E[Y_0] = 0$$

By claim,

$$(b-a)E[U_N[a,b]] \leq E[(X_N-a)^-]$$

This completes the proof of Doob's Lemma. □

Lemma 6.3. *For any $a < b$, $P(\Lambda_{a,b}) = 0$*

Proof. By definition $\Lambda_{a,b} = \{\omega : U_\infty[a,b] = \infty\}$. Now by Doob's Lemma

$$(b-a)E[U_N[a,b]] \leq E[(X_N-a)^-] \leq \sup_n E[|X_n|] + |a| < \infty$$

Now, $U_N[a,b] \nearrow U_\infty[a,b]$. Hence, by the Monotone Convergence Theorem, $E[U_N[a,b]] \nearrow E[U_\infty[a,b]]$. That is, $E[U_\infty[a,b]] < \infty$. Hence $P(U_\infty[a,b] = \infty) = 0$. □

Theorem 6.5 (Doob's inequality)**.** *Let X_n be a sub-MG and let $X_n^* = \max_{0\leq m\leq n} X_m^+$. Given $\lambda > 0$ let $A = \{X_n^* \geq \lambda\}$ then:*

$$\lambda P(A) \leq E[X_n 1(A)] \leq E[X_n^+]$$

Proof. Define the stopping time

$$N = \min\{m : X_m^* \geq \lambda \text{ or } m = n\}$$

Thus $P(N \leq n) = 1$. Now, by the Optional Stopping Theorem we have that $X_{N \wedge n} = X_N$. Thus

$$E[X_N] \leq E[X_n] \tag{6.3}$$

We have:

$$\begin{aligned} E[X_N] &= E[X_N 1(A)] + E[X_N 1(A^c)] = E[X_N 1(A)] + E[X_n 1(A^c)] \\ E[X_n] &= E[X_n 1(A)] + E[X_n 1(A^c)] \\ E[X_N 1(A)] &\leq E[X_n 1(A)] \end{aligned} \tag{6.4}$$

where the last inequality follows from 6.3. but

$$\lambda P(A) \leq E[X_N 1(A)]$$

From then we have: $\lambda P(A) \leq E[X_n 1(A)] \leq E[X_n^+ 1(A)] \leq E[X_n^+]$. Suppose X_n is a non-negative sub-MG, then:

$$P(\max_{0 \leq k \leq n} X_k \geq \lambda) \leq \frac{1}{\lambda} E[X_n]$$

if it were MG, then we also obtain:

$$P(\max_{0 \leq k \leq n} X_k \geq \lambda) \leq \frac{1}{\lambda} E[X_n] = \frac{1}{\lambda} E[X_0]$$

□

6.2.1 L^p maximal inequality and L^p convergence

Theorem 6.6 (L^p version Doob's inequality)**.** *Let X_n be a sub MG. Suppose $E[(X_n^+)^p] < \infty$ for some $p > 1$. Then,*

$$E[(\max_{0 \leq k \leq n} X_k^+)^p]^{\frac{1}{p}} \leq q E[(X_n^+)^p]^{\frac{1}{p}}$$

where $\frac{1}{q} + \frac{1}{p} = 1$. In particular, if X_n is a MG then $|X_n|$ is a sub-MG and hence:

$$E[(\max_{0 \leq k \leq n} |X_k|)^p]^{\frac{1}{p}} \leq q E[(|X_n|)^p]^{\frac{1}{p}}$$

Theorem 6.7 (supremum version L^p max inequality)**.** *If X_n is a martingale with $\sup_n E[|X_n|^p] < \infty$ where $p > 1$, then $X_n \to X$ a.s. and L^p, where $X = \limsup_n X_n$.*

Proof. Since $\sup_n E[|X_n|^p] < \infty, p > 1$, by MG-convergence theorem, we have that:

$$X_n \to X, a.s., \text{ where } X = \limsup_n X_n$$

For L^p convergence, we will use L^p-inequality of Theorem 3. That is,

$$E[(\sup_{0 \leq m \leq n} |X_m|)^p] \leq q^p E[|X_n|^p]$$

Now $\sup_{0\leq m\leq n}|X_m| \nearrow \sup_{0\leq m}|X_m|$. Therefore, by the Monontone Convergence Theorem we obtain that:

$$E[\sup_{0\leq m}|X_m|^p] \leq q^p \sup_n E[|X_n|^p] < \infty$$

Thus $\sup_{0\leq m}|X_m| \in L^p$. Now,

$$|X_n - X| \leq 2\sup_{0\leq m}|X_m|$$

Therefore, by the Dominated Convergence Theorem $E[|X_n - X|^p] \to 0$. □

Proof. Proof of the previous theorem: We will use the truncation of X_n^* to prove the result. let M be the truncation parameter: $X_n^{*,M} = \min(X_n^*, M)$. Now, consider the following:

$$\begin{aligned} E[(X_n^{*,M})^p] &= \int_0^\infty p\lambda^{p-1}P(X_n^{*,M} \geq \lambda)d\lambda \\ &\leq \int_0^\infty p\lambda^{p-1}[\frac{1}{\lambda}E[X_n^+1(X_n^{*,M} \geq \lambda)]]d\lambda \end{aligned}$$

THe above inequality follows from:

$$P(X_n^{*,M} \geq \lambda) = \begin{cases} 0, & \text{if } M < \lambda \\ P(X_n^* \geq \lambda), & \text{if } M \geq \lambda. \end{cases}$$

By an application of Fubini for non-negative integrands, we have:

$$\begin{aligned} pE[X_n^+\int_0^{X_n^{*,M}}\lambda^{p-2}d\lambda] &= \frac{p}{p-1}E[X_n^+(X_n^{*,M})^{p-1}] \\ &\leq \frac{p}{p-1}E[(X_n^+)^p]^{\frac{1}{p}}E[(X_n^{*,M})^{(p-1)q}]^{\frac{1}{q}}, \text{ by Holder's inequality} \end{aligned}$$

Hence $\frac{1}{q} = 1-\frac{1}{p} \implies q(p-1) = p$. Thus we have the right hand side simplifies to $= qE[(X_n^+)^p]^{\frac{1}{p}}E[(X_n^{*,M})^p]^{\frac{1}{q}}$. Thus:

$$||X_n^{*,M}||_p^p \leq q||X_n^+||_p||X_n^{*,M}||_p^{\frac{q}{p}} \implies ||X_n^{*,M}||_p^{p(1-\frac{1}{q})} \leq q||X_n^+||_p$$

□

6.3 Backward Martingale

Definition 6.2 (backward martingale). Let $\mathcal{F}_n$ be increasing sequence of σ-algebra, $n \leq 0$, such that $\ldots \subset F_{-3} \subset F_{-2} \subset F_{-1} \subset F_0$. Let X_n be $\mathcal{F}_n$ adapted, $n \leq 0$ and:

$$E[X_{n+1}|\mathcal{F}_n] = X_n, n < 0$$

Then X_n is called backward MG.

Theorem 6.8 (backward MG convergence). *Let X_n be backward MG. Then*

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

Compare with standard MG convergence results: (a): we need $\sup E[|X_n|] < \infty$ or non-negative MG in Doob's convergence theorem, which gives a.s. convergence not L^1. (b): For L^1, we need UI. And, it is necessary because if $X_n \to X_\infty$ a.s. and L^1 then there exists $X \in F_\infty$ s.t. $X_n = E[X|\mathcal{F}_n]$; and hence, X_n is UI.

Proof. Recall Doob's convergence theorem's proof. Let

$$\begin{aligned}\Lambda &:= \{w : X_n(w) \text{ does not converge to a limit in } [-\infty, \infty] \\ &= \{w : \liminf_n X_n(\omega) < \limsup_n X_n(\omega)\} \\ &= \bigcup_{a,b:a,b\in\mathbb{Q}} \{w : \liminf_n X_n(w) < a < b < \limsup_n X_n(w)\} \\ &= \bigcup_{a,b:a,b\in\mathbb{Q}} \Lambda_{a,b}\end{aligned}$$

Now, recall $U_n[a, b]$ is the number of upcrossing of $[a, b]$ in $X_n, X_{n+1}, \ldots, X_0$ as $n \to -\infty$. By upcrossing inequality, it follows that:

$$(b - a)E[U_n[a, b]] \leq E[|X_0|] + |a|$$

Since $U_n[a, b] \nearrow U_\infty[a, b]$ and By Monotone Convergence Theorem, we have: $E[U_\infty[a, b]] < \infty \implies P(\Lambda_{a,b}) = 0$ This implies X_n converges a.s.

Now $X_n = E[X_0|\mathcal{F}_n]$. Therefore, X_n is UI. This implies $X_n \to X_{-\infty}$ in L^1. □

Theorem 6.9 (convergence of backward martingale). *If $X_{-\infty} = \lim_{n\to-\infty} X_n$ and $\mathcal{F}_{-\infty} = \bigcap_n \mathcal{F}_n$. Then $X_{-\infty} = E[X_0|\mathcal{F}_{-\infty}]$*

Proof. Let $X_n = E[X_0|\mathcal{F}_n]$ If $A \in \mathcal{F}_{-\infty} \subset F_n$, then $E[X_n; A] = E[X_0; A]$. Now,

$$\begin{aligned}|E[X_n; A] - E[X_{-\infty}; A]| = |E[X_n - X_{-\infty}; A]| &\leq E[|X_n - X_{-\infty}|; A] \\ &\leq E[|X_n - X_{-\infty}|] \to 0, \text{ as } n \to -\infty.\end{aligned}$$

Hence, $E[X_{-\infty}; A] = E[X_0; A]$. Thus $X_{-\infty} = E[X_0|\mathcal{F}_\infty]$. □

Theorem 6.10 (backward convergence in L^1). *Let $\mathcal{F}_n \searrow \mathcal{F}_{-\infty}$ and $Y \in L^1$. Then, $E[Y|\mathcal{F}_n] \to E[Y|\mathcal{F}_{-\infty}]$ a.s. in L^1*

Proof. $X_n = E[Y|\mathcal{F}_n]$ is backward MG by definition. Therefore, $X_n \to X_{-\infty}$ a.s. and in L^1

By Theorem above, $X_{-\infty} = E[X_0|\mathcal{F}_{-\infty} = E[Y|\mathcal{F}_{-\infty}]$. Thus $E[Y|\mathcal{F}_n] \to E[Y|\mathcal{F}_{-\infty}]$ □

6.3.1 Strong Law of Large Number

Theorem 6.11 (Backward MG SLLN). *Let ξ be i.i.d. with $E[|\xi_i|] < \infty$. Let $S_n = \xi_1 + \ldots + \xi_n$ Let $X_{-n} = \frac{S_n}{n}$ and $\mathcal{F}_{-n} = \sigma(S_n, \xi_{n+1}, \ldots)$. Then:*

$$\begin{aligned}E[X_{-n}|\mathcal{F}_{-n-1}] = E[\frac{S_n}{n}|\mathcal{F}_{-n-1}] &= \frac{1}{n}\sum_{i=1}^{n} E[\xi_i|S_{n+1}] \\ &= E[\xi_1|S_{n+1}] = \frac{1}{n+1}S_{n+1} = X_{-n+1}\end{aligned}$$

Then X_{-n} is backward MG.

Proof. By Theorem above, we have $X_{-n} \to X_{-\infty}$ a.s. and in L^1, with $X_{-\infty} = E[\xi_1|\mathcal{F}_{-\infty}]$. Now $\mathcal{F}_{-\infty}$ is in ξ. By Hewtti-Savage 0-1 law, ξ is trivial. That is, $E[\xi_1|\mathcal{F}_{-\infty}]$ is a constant. Therefore, $E[X_{-\infty}] = E[\xi_1]$ is also a constant. Thus,

$$X_{-\infty} = \lim_{n\to\infty} \frac{S_n}{n} = E[\xi_1]$$

□

Theorem 6.12 (Hewitt-Savage 0-1 law)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. and ξ be the exchangable σ-algebra:*

$$\xi_n = \{A : \pi_n A = A; \forall \pi_n \in S_n\}; \xi = \bigcup_n \xi_n$$

If $A \in \xi$ then $P(A) \in \{0, 1\}$.

Intuition: Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent and identically-distributed random variables taking values in a set $\mathcal{E}$. The Hewitt-Savage zero–one law says that any event whose occurrence or non-occurrence is determined by the values of these random variables and whose occurrence or non-occurrence is unchanged by finite permutations of the indices, has probability either 0 or 1 (a "finite" permutation is one that leaves all but finitely many of the indices fixed).

The proof is based on this Lemma:

Lemma 6.4. *Let $X_1, X_2, \ldots, X_k$ be i.i.d. and define*

$$A_n(\phi) = \frac{1}{n_{p_k}} \sum_{(i_1,\ldots,i_k)\in\{1,\ldots,n\}} A_n(\phi(X_{i_1}, \ldots, X_{i_k}))$$

If ϕ is bounded then

$$A_n(\phi) \to E[\phi(X_1, \ldots, X_k)]a.s.$$

Proof. $A_n(\phi) \in \xi_n$ by definition. So,

$$\begin{aligned} A_n(\phi) = E[A_n(\phi)|S_n] &= \frac{1}{n_{p_k}} \sum_{i_1,\ldots,i_k} E[\phi(X_{i_1}, \ldots, X_{i_k}|\xi_n] \\ &= \frac{1}{n_{p_k}} \sum_{i_1,\ldots,i_k} E[\phi(X_1, \ldots, X_k)|\xi_n] = E[\phi(X_1, \ldots, X_k)|\xi_n] \end{aligned}$$

Let $\mathcal{F}_{-n} = \xi_n$. Then $\mathcal{F}_{-n} \to \mathcal{F}_{-\infty} = \xi$. Then for $Y = \phi(X_1, \ldots, X_k)$. $E[Y|\mathcal{F}_{-n}]$ is backward MG. Therefore,

$$E[Y|\mathcal{F}_{-n}] \to E[Y|\mathcal{F}_{-\infty}] = E[\phi(X_1, \ldots, X_k)|\xi]$$

Thus $A_n(\phi) \to E[\phi(X_1, \ldots, X_k)|\xi]$. We want to show that indeed $E[\phi(X_1, \ldots, X_k)|\xi]$ is $E[\phi(X_1, \ldots, X_n)]$.

First, we show that $E[\phi(X_1, \ldots, X_n)|\xi] \in \sigma(X_{k+1}, \ldots)$ since ϕ is bounded. Then we find that if $E[X|\mathcal{G}] \in \mathcal{F}$ where X is independent of $\mathcal{F}$ then $E[X|\mathcal{G}]$ is constant, equal to $E[X]$. This will complete the proof of Lemma.

□

First step, consider $A_n(\phi)$. It has n_{p_k} terms in which there are $k(n-1)_{p_{k-1}}$ terms containing X_1. Therefore, the effect of terms containing X_1 is:

$$\begin{aligned}T_n(1) &\equiv \frac{1}{n_{p_k}} \sum_{(i_1,\ldots,i_k)} \phi(X_{i_1},\ldots,X_{i_k}) \leq \frac{1}{n_{p_k}} k((n-1)_{p_{k-1}}||\phi||_\infty \\ &= \frac{(n-k)!}{k!} k \frac{(n-1)!}{(n-k)!}||\phi||_\infty \\ &= \frac{k}{n}||\phi||_\infty \to 0, \text{ as } n \to \infty\end{aligned}$$

Let $A_n^{-1}(\phi) = A_n(\phi) - T_n(1)$. Then we have: $A_n^{-1}(\phi) \to E[\phi(X_1,\ldots,X_n)|\xi]$ from the above two equations. Thus $E[\phi(\phi(X_1,\ldots,X_n))|\xi]$ is independent on X_1. Similarly, repeating argument for $X_2,\ldots,X_n$ we obtain that:

$$E[\phi(X_1,\ldots,X_n)|\xi] = \sigma(X_{n+1},\ldots)$$

Second step: if $E[X^2] \leq \infty$, $E[X|\mathcal{G}] \in \mathcal{F}$, X is independent of $\mathcal{F}$ then $E[X|\mathcal{G}] = E[X]$

Proof. Let $Y = E[X|\mathcal{G}]$. Now $Y \in \mathcal{F}$ and X is independent on $\mathcal{F}_1$ we have that: $E[XY] = E[X]E[Y] = E[Y]^2$, since $E[Y] = E[X]$. Now, by definition of conditional expectation for any $Z \in \mathcal{G}$, $E[XZ] = E[YZ]$. Hence, for $Z = Y$. we have $E[XY] = E[Y^2]$. Thus $E[Y^2] = E[Y]^2 \implies Var(Y) = 0 \implies Y = E[Y]$ a.s. □

We have proved that $A_n(\phi) \to E[\phi(X_1,\ldots,X_n)]$ a.s. for all bounded ϕ dependent on finitely many components.

By the first step, ξ is independent on $\mathcal{G}_k = \sigma(X_1,\ldots,X_k)$. This is true for all k. $\bigcup_k \mathcal{G}_k$ is a π-system which contains Ω. Therefore, ξ is independent of $\sigma(\bigcup_k \mathcal{G}_k)$ and $\xi \subset \sigma(\bigcup_k \mathcal{G}_k)$. Thus for all $A \in \xi$, A is independent of itself. Hence,

$$P(A \cap A) = P(A)P(A) \implies P(A) \in \{0,1\}.$$

6.3.2 De Finetti's Theorem

Theorem 6.13 (De Finetti's Theorem). *Given $X_1, X_2, \ldots$ sequence of exchangeable, that is, for any n and $\pi_n \in S_n$, $(X_1, X_2, \ldots, X_n) \equiv (X_{\pi_n(1)},\ldots,X_{\pi_n(n)})$, then conditional on ξ, $X_1,\ldots,X_n,\ldots$ are i.i.d.*

Proof. As in H-S's proof and Lemma, define $A_n(\phi) = \frac{1}{n_{p_k}} \sum_{(i_1,\ldots,i_k)} \phi(X_{i_1},\ldots,X_{i_k})$. Then, due to exchangeability,

$$\begin{aligned}A_n(\phi) &= E[A_n(\phi)|\xi_n] = E[\phi(X_1,\ldots,X_n)|\xi_n] \\ &\to E[\phi(X_1,\ldots,X_n)|\xi] \text{ by backward MG convergence theorem}\end{aligned}$$

Since $X_1,\ldots$ may not be i.i.d. ξ can be nontrivial. Therefore, the limit need not be constant. Consider a $f:\mathbb{R}^{k-1} \to \mathbb{R}$ and $g:\mathbb{R}\to\mathbb{R}$. Let $I_{n,k}$ be set of all distinct $1 \leq i_1,\ldots,i_k \leq n$ then:

$$\begin{aligned}n_{p_{k-1}} A_n(f) n A_n(g) &= \sum_{i \in I_{n,k-1} f(X_{i_1},\ldots,X_{i_{k-1}})} \sum_{m \leq n} g(X_m) \\ &= \sum_{i \in I_{n,k}} f(X_1,\ldots,X_{i_{k-1}}) g(X_{i_k}) + \sum_{i \in I_{n,k-1}} \left[f(X_{i_1},\ldots,X_{i_{k-1}}) \sum_{j=1}^{k-1} g(X_{i_j}) \right]\end{aligned}$$

Let $\phi_j(X_1,\ldots,X_{k-1}) = f(X_1,\ldots,X_{k-1})g(X_j), 1 \leq j \leq k-1$ and $\phi(X_1,\ldots,X_k) = f(X_1,\ldots,X_{k-1})g(X_k)$. Then:

$$n_{p_{k-1}} A_n(f) n A_n(g) = n_{p_k} A_n(\phi) + n_{p_{k-1}} \sum_{j=1}^{k-1} A_n(\phi_j)$$

Dividing by n_{p_k}, we have:

$$\frac{n}{n-k+1} A_n(f) A_n(g) = A_n(\phi) + \frac{1}{n-k+1} \sum_{j=1}^{k} A_n(\phi_j)$$

From the fact that $||f||_\infty, ||g||_\infty < \infty$ we have:

$$E[f(X_1, \ldots, X_{k-1}|\xi)] E[g(X_1)|\xi] = E[f(X_1, \ldots, X_{k-1}) g(X_k)|\xi]$$

Thus we have for any collection of bounded functions $f_1, \ldots, f_k$:

$$E[\Pi_{i=1}^{k} f_i(X_i)|\xi] = \Pi_{i=1}^{k} E[f_i(X_i)|\xi]$$

□

Chapter 7

Martingale and Crossing Theory

7.1 Naive Version Introduction

Definition 7.1 (Random Walk)**.** Let X_n be i.i.d. random variables with $E[X_1] = \mu < 0$. Then:

$$S_n = \sum_{i=1}^{n} X_i, S_0 = 0$$

is called a random walk process.

Two associated martingales are:

- $S_n - n\mu$ is a martingale;
- Let $\phi(\theta) = E[e^{\theta X_1}]$, then the second martingale is:

$$Y_n = (\phi(\theta))^{-n} e^{\theta S_n}$$

Proposition 7.1 (Properties of MGF $\phi(\theta)$)**.**

- $\phi(\theta)$ *is convex;*
- $\phi(0) = 1$ *and* $\phi'(0) = \mu < 0$*;*
- *Under mild conditions, there exists* $\theta^* > 0$ *satisfying* $\phi(\theta^*) = 1$*. Then* $e^{\theta^* S_n}$ *is a martingale. (Check homework 2 for reference)*

Definition 7.2 (Hitting Probability)**.** Let $A < 0 < B$ and define:

$$N = \min\{n : S_n \leq A \text{ or } S_n \geq B\}.$$

Then N is a stopping time and $Y_{n\wedge N}$ is bounded (between A and B). Applying Martingale Stopping Theorem on $Y_n = e^{\theta^* S_n}$, we have:

$$E[Y_N] = E[Y_0] = 1.$$

That is, $E[Y_N|S_N \geq B]P_B + E[Y_N|S_N \leq A]P_A = 1$. When A and B are large compared with X_i, then:

$$E[Y_N|S_N \geq B] \approx e^{\theta^* B}, E[Y_N|S_N \leq A] \approx e^{\theta^* A}.$$

Hence, $e^{\theta^* A}P_A + e^{\theta^* B}P_B \approx 1$. By $P_A = 1 - P_B$, we obtain:

$$P_A \approx \frac{1 - e^{\theta^* B}}{e^{\theta^* A} - e^{\theta^* B}}; P_B \approx \frac{e^{\theta^* A} - 1}{e^{\theta^* A} - e^{\theta^* B}}$$

Remark 7.1. The $\approx$ sign would be " $=$ " if S_A and S_B are reached exactly. This is the case of simple random walk when X_i takes value 1 and -1 (with probability p and $1-p$).

We can conduct the similar analysis for S_N: since $E[S_N] \approx AP_A + BP_B$ and since $E[S_N] = E[N]E[X_1]$, we obtain:

$$E[N] \approx \frac{A(1-e^{\theta^* B}) - B(e^{\theta^* B} - 1)}{E[X_1](e^{\theta^* A} - e^{\theta^* B})}$$

Consider the special case of simple random walk: $P(X_1 = 1) = p = 1 - P(X_1 = -1)$ then the formula become exact, and by letting $q = 1-p$ we have:

$$P_A = \frac{1-(q/p)^B}{(q/p)^A - (q/p)^B}$$

The expected stopping time is then:

$$E[N] = \frac{B((q/p)^A - 1) - A((q/p)^B - 1)}{((q/p)^A - (q/p)^B)(p-q)}$$

In the analysis above, P_B is the probability that the random walk reaches B before reaching A and if $A = -\infty$ then P_B is the probability that the random walk reaches B.

Definition 7.3 (One-Side Crossing)**.** Consider $E[X_1] < 0$ and the process starts at 0. What is the chance of the process goes above B? let $\phi(\theta^*) = E[e^{\theta^* X_1}] = 1$ then $\theta^* > 0$.

Since $Y_n = e^{\theta^* S_n}$ is a martingale, we have:

$$\begin{aligned} 1 = E[Y_0] &= E[Y_N] \\ &= E[e^{\theta^* S_N} | S_N \le A]P_A + E[e^{\theta^* S_N} | S_N \ge B]P_B \\ &\ge E[e^{\theta^* S_N} | S_N \le A]P_A + E[e^{\theta^* S_B} | S_N \ge B]P_B \end{aligned}$$

letting $A \to -\infty$, then the first term goes to 0 and we obtain:

$$P(S_n \text{ ever reaches } B) \le e^{-\theta^* B}.$$

Remark 7.2. For simple random walk with $p = P(X_1 = 1) < \frac{1}{2}$ and we have: $e^{\theta^*} = q/p > 1$. Hence $\theta^* > 0$

In that case, B is reached, hence the approximation becomes exact. $P(S_n \text{ ever reaches } B) = (p/q)^B$.

7.1.1 Application in Actuarial Science

Claims arriving at an insurance company are i.i.d. The claim in period n is X_n with mean $E[X_1]$ and variance $Var(X_1)$. The company receives premium for the periods are i.i.d. $Y_1, Y_2, \ldots$. The company is endowed with initial capital $x > 0$. We want to find the minimum x such that the bankruptcy probability is no more than $\delta(= 0.1\%)$

The Bankruptcy problem is the wealth of the company at time n is:

$$x + \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i$$

The bankruptcy is defined as:

$$x + \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i < 0$$

for some n in the future.

Let $S_n = \sum_{i=1}^{n}(X_i - Y_i)$. Then ban is $P(\max_{i\geq 0} S_i > x)$.

If $E[X_1] > E[Y_1]$, then the probability is always 1 for any $x > 0$. It can be shown that it is true even $E[X_1] = E[Y_1]$ (unless $X_i - Y_1 = 0, a.s.$). We consider the more interesting case $E[X_1] < E[Y_1]$. Let $\theta^* > 0$ be $E[e^{\theta^*(X_1 - Y_1)}] = 1$. The previous result showed that:

$$P(\max_{i\geq 0} S_i > x) \leq e^{-\theta^* x}.$$

In particular, if X_1 is exponentially distributed (say with parameter μ), then $(S_n - A|S_n > A)$ is exponentially distributed with parameter μ. In this case:

$$P(\max_{i\geq 0} S_i > x) = \frac{\mu - \theta^*}{\mu} \cdot e^{-\theta^* x}$$

The last important result we discuss in martingale theory is an a.s. convergence result.

Theorem 7.1 (Martingale Convergence Theorem)**.** *If X_n is a martingale with bounded absolute mean, i.e., there exists $M > 0$ such that $E[|X_n|] \leq M$, then M converges almost surely to some random variable say X.*

Almost surely convergence is always based on Borel-Contelli lemma, and a simple proof for the case where $E[X_n^2]$ is bounded is available in Karlin and Taylor Book.

Proof. Need to show that $\{X_n : n \geq 0\}$ is a Cauchy sequence, or $|X_{N'} - X_N| \to 0$ as $N' \geq N \to \infty$.

This can be expressed as the event: $B = \bigcap_k \bigcup_N B_N(k)$, where $B_N(k) = \{|X_{N'} - X_N| < \frac{1}{k}$ for all $N' \geq N\}$. We want to show that $P(B) = 1$ or $P(B^c) = 0$ where:

$$B^c = \bigcup_k \bigcap_N B_N^c(k).$$

where $B_n^c(k) = \{|X_{N+l} - X_N| \geq 1/k$ for some $l \geq 1\}$. We want to show that $P(B) = 1$ or $P(B^c) = 0$ where $B^c = \bigcup_k \bigcap_N B_N^c(k)$; where

$$B_N^c(k) = \{|X_{N+l} - X_N| \geq 1/k \text{ for some } l \geq 1\}$$

Since $P(B^c) \leq \sum_k P(\bigcap_N B_N^c(k))$, it suffices to show, for any k,

$$P(\bigcap_N B_N^c(k)) = 0.$$

It is sufficient that: $P(B_N^c(k)) \to 0$ as $N \to \infty$. This is so because $P(\bigcap_N B_N^c(k)) \leq P(B_N^c(k))$ for any N.

Assume $E[X_n^2]$ is bounded. Then, since X_n^2 is a sub-martingale, $E[X_n^2]$ is increasing. Thus $E[X_n^2] \to \mu < \infty$. By Kolmogorov inequality,

$$P(|X_{m+l} - X_m|) > \varepsilon \text{ for some } 1 \leq l \leq n) \leq \frac{E[(X_{m+n} - X_m)^2]}{\varepsilon}$$

Note that

$$\begin{aligned} E[(X_{m+n} - X_m)^2] &= E[X_{m+n}^2] - 2E[X_{m+n}X_n] + E[X_n^2] \\ &= E[X_{m+n}^2] - 2E[E[X_{m+n}X_n|\mathcal{F}_n]] + E[X_n^2] \\ &= E[X_{m+n}^2] - 2E[X_n E[X_{m+n}|\mathcal{F}_n]] + E[X_n^2] \\ &= E[X_{m+n}^2] - E[X_n^2] \end{aligned}$$

Letting $n \to \infty$, we obtain, by probability continuity,

$$P(|X_{m+l} - X_m| > \varepsilon \text{ for some } l \geq 1) \leq \frac{\mu - E[X_m^2]}{\varepsilon}$$

Hence, by $E[X_m^2] \to \mu$ we conclude that $P(B_N^c(k)) \to 0$ as $N \to \infty$.
The Martingale Convergence Theorem is thus proved. □

Corollary 7.2. *If X_n is a non-negative martingale. Then X_n converges a.s.*

Proof. Since $E[|X_n|] = E[X_n] = E[X_1] = M < \infty$. Hence X_n converges. □

7.2 Martingale Concentration Inequalities and Applications

Suppose X_n is a martingale wrt filtration $\mathcal{F}_n$ such that $X_0 = 0$. THe goal is to obtain bounds of the form $P(|X_n| \geq \delta n) \leq \exp(-\Theta(n))$ under some condition on X_n.

Theorem 7.3 (Azuma-Hoeffding Inequality Revisit)**.** *Suppose $X_n, n \geq 1$ is a martingale such that $X_0 = 0$ and $|X_i - X_{i-1}| \leq d_i, 1 \leq i \leq n$ almost surely for some constants $d_i, 1 \leq i \leq n$. Then for every $t > 0$,*

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

Notice that in the special case when $d_i = d$, we can take $t = xn$ and obtain an upper bound $2\exp\bigl(-x^2 n/(2d^2)\bigr)$ - which is of the form promised above.

Proof. $f(x) \equiv \exp(\lambda x)$ is a convex function in x for any $\lambda \in \mathbb{R}$. Then we have $f(-d_i) = \exp(-\lambda d_i)$ and $f(d_i) = \exp(\lambda d_i)$. Using convexity we have that when $|x/d_i| \leq 1$

$$\begin{aligned}
\exp(\lambda x) = f(x) &= f(\frac{1}{2}(\frac{x}{d_i} + 1)d_i + \frac{1}{2}(1 - \frac{x}{d_i})(-d_i)) \\
&\leq \frac{1}{2}f(\frac{x}{d_i} + 1)f(d_i) + \frac{1}{2}f(1 - \frac{x}{d_i})f(-d_i) \\
&= \frac{f(d_i) + f(-d_i)}{2} + \frac{f(d_i) - f(-d_i)}{2}x
\end{aligned}$$

Furthermore, for every a:

$$\begin{aligned}
\frac{\exp(a) + \exp(-a)}{2} &= \sum_{k=0}^{\infty} \frac{a^k}{k!} + \sum_{k=0}^{\infty} \frac{(-1)^k a^k}{k!} = \sum_{k=0}^{\infty} \frac{a^{2k}}{(2k)!} \\
&\leq \sum_{k=0}^{\infty} \frac{a^{2k}}{2^k k!} \\
&= \sum_{k=0}^{\infty} \frac{(\frac{a^2}{2})^k}{k!} = \exp\left(\frac{a^2}{2}\right)
\end{aligned}$$

We conclude that for every x such that $|x/d_i| \leq 1$:

$$\exp(\lambda x) \leq \exp\left(\frac{d_i^2}{2}\right) + \frac{\exp(\lambda d_i) - \exp(-\lambda d_i)}{2}x$$

We now turn to our martingale sequence X_n. For every $t > 0$ and every $\lambda > 0$ we have:

$$\begin{aligned}P(X_n \geq t) &= P(\exp(\lambda X_n) \geq \exp(\lambda t)) \\ &\leq \exp(-\lambda t) E[\exp(\lambda X_n)] \\ & \exp(-\lambda t) E[\exp\left(\lambda \sum_{1 \leq i \leq n} (X_i - X_{i-1})\right)]\end{aligned}$$

where $X_0 = 0$ was used in the last equality. Applying the tower property of conditional expectation, we have:

$$\begin{aligned}E[\exp\left(\lambda \sum_{1 \leq i \leq n} (X_i - X_{i-1})\right)] &= E[E[\exp(\lambda(X_n - X_{n-1})) \exp\left(\lambda \sum_{1 \leq i \leq n-1} (X_i - X_{i-1})\right) |\mathcal{F}_{n-1}]] \\ &= E[\exp(\lambda(X_n - X_{n-1})) \exp\left(\lambda \sum_{1 \leq i \leq n} (X_i - X_{i-1})\right) |\mathcal{F}_{n-1}] \\ &= \exp\left(\lambda \sum_{1 \leq i \leq n-1} (X_i - X_{i-1})\right) E[\exp(\lambda(X_n - X_{n-1}))|\mathcal{F}_{n-1}] \\ &\leq \exp\left(\lambda \sum_{1 \leq i \leq n-1} (X_i - X_{i-1})\right) \times \\ &(\exp\left(\frac{d_n^2 \lambda^2}{2}\right) + \frac{\exp(\lambda d_i) - \exp(-\lambda d_i)}{2} E[X_n - X_{n-1}|\mathcal{F}_{n-1}])\end{aligned}$$

Martingale property implies $E[X_n - X_{n-1}|\mathcal{F}_{n-1}] = 0$ then we have the upper bound:

$$E[\exp\left(\lambda \sum_{1 \leq i \leq n} (X_i - X_{i-1})\right)] \leq E[\exp\left(\lambda \sum_{1 \leq i \leq n-1} (X_i - X_{i-1})\right)] \exp\left(\frac{\lambda^2 d_n^2}{2}\right)$$

Iterating further we obtain the following upper bound on $P(X_n \geq t)$:

$$\exp(-\lambda t) \exp\left(\frac{\sum_{1 \leq i \leq n} \lambda^2 d_i^2}{2}\right)$$

Optimizing over the choice of λ, we see that the tightest bound is obtained by setting $\lambda = t / \sum_i d_i^2 > 0$ leading to an upper bound:

$$P(X_n \geq t) \leq \exp\left(-\frac{t^2}{2 \sum_i d_i^2}\right)$$

A similarly approach using $\lambda < 0$ gives for every $t > 0$:

$$P(X_n \leq -t) \leq \exp\left(-\frac{t^2}{2 \sum_i d_i^2}\right)$$

Combining, we obtain the required result. □

7.3 Application to Lipshitz continuous functions of i.i.d. random variables

Suppose $X_1, \dots, X_n$ are independent random variables. SUppose $g : \mathbb{R}^n \to \mathbb{R}$ is a function and $d_1, \dots, d_n$ are constants such that for any two vectors $x_1, \dots, x_n$ and $y_1, \dots, y_n$ in $\mathbb{R}^n$:

$$|g(x_1, \dots, x_n) - g(y_1, \dots, y_n)| \leq \sum_{i=1}^{n} d_i 1\{x_i \neq y_i\}$$

In particular, when a vector x changes value only in the ith-coordinate the amount of change in function g is at most d_i. Suppose g is Lipschitz continuous with constant K and consider a subset of vectors $x = (x_1, \dots, x_n)$ such that $|x_i| \leq c_i$. Then for every x, y, $|g(x) - g(y)| \leq K|x - y|$, where $|x - y| = |x_i - y_i|_i$. Then for every two such vectors:

$$|g(x) - g(y)| \leq K|x - y| \leq K \sum_i 2c_i |x_i - y_i|$$

and therefore this fits into a previous framework with $d_i = Kc_i$.

Theorem 7.4 (McDiarmid inequality alternate version)**.** *Suppose $X_i, 1 \leq i \leq n$ are i.i.d. and function $g : \mathbb{R}^n \to \mathbb{R}$ satisfies the condition $|g(x_1, \dots, x_n) - g(y_1, \dots, y_n)| \leq \sum_{i=1}^{n} d_i 1\{x_i \neq y_i\}$. Then for every t:*

$$P(|g(X_1, \dots, X_n) - E[g(X_1, \dots, X_n)]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} d_i^2}\right)$$

Proof. Let $\mathcal{F}_i$ be the σ-field generated by variables $X_1, \dots, X_i : \mathcal{F}_i = \sigma(X_1, \dots, X_i)$. For convenience, we also set $\mathcal{F}_0$ to be the trivial σ-field consisting of $\emptyset, \Omega$, so that $E[Z|\mathcal{F}_0] = E[Z]$ for every r.v. Z. Let $M_0 = E[g(X_1, \dots, X_n)], M_1 = E[g(X_1, \dots, X_n)|\mathcal{F}_1], \dots, M_n = E[g(X_1, \dots, X_n)|\mathcal{F}_n]$. Obsercve that M_n is simply $g(X_1, \dots, X_n)$ since $X_1, \dots, X_n$ are measurable wrt $\mathcal{F}_n$. Thus we by tower property:

$$E[M_n|\mathcal{F}_{n-1}] = E[E[g(X_1, \dots, X_n)|\mathcal{F}_n]|\mathcal{F}_{n-1}] = M_{n-1}$$

Thus M_i is a martingale. We have:

$$\begin{aligned} M_{i+1} - M_i &= E[E[g(X_1, \dots, X_n)|\mathcal{F}_{i+1}] - E[g(X_1, \dots, X_n)|\mathcal{F}_i]] \\ &= E[E[g(X_1, \dots, X_n) - E[g(X_1, \dots, X_n)]|\mathcal{F}_i]|\mathcal{F}_{i+1}] \end{aligned}$$

Since X_i's are independent, then M_i is a r.v. which on any vector $x = (x_1, \dots, x_n) \in \Omega$ takes value:

$$M_i = \int_{x_{i+1}, \dots, x_n} g(x_1, \dots, x_i, x_{i+1}, \dots, x_n) dP(x_{i+1}) \cdots dP(x_n)$$

Similarly,

$$M_{i+1} = \int_{x_{i+2}, \dots, x_n} g(x_1, \dots, x_{i+1}, x_{i+2}, \dots, x_n) dP(x_{i+2}) \cdots dP(x_n)$$

Thus

$$\begin{aligned} |M_{i+1} - M_i| &= |\int_{x_{i+2}, \dots, x_n} (g(x_1, \dots, x_n) - \int_{x_{i+1}} g(x_1, \dots, x_n) dP(x_{i+1}) \cdots dP(x_n)| \\ &\leq d_{i+1} \int_{x_{i+2}, \dots, x_n} dP(x_{i+1}) \cdots dP(X_n) = d_{i+1} \end{aligned}$$

This derivation represents a simple diea that M_i and M_{i+1} only differ in "averaging out" X_{i+1} in M_i. Applying the Azuma-Hoeffding inequality to the martingale M_i we obtain the result □

Example 7.1. Suppose we have a distribution function F and i.i.d. sequence $X_1, \ldots, X_n$ with distribution F. We can build the empirical distribution function $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq x\}$. An important theorem called Glivenko-Cantelli says that $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ converges to zero and in expectation, the latter meaning of course that $E[\sup_{x\in\mathbb{R}} |F_n(x) - F(x)] \to 0$. Applying the martingale Concentration Inequality we obtain that the deviation of $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ from its expectation is exponentially small. Let $L_n = L_n(X_1, \ldots, X_n) = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$. We need to bound $P(|L_n - E[L_n]| > t)$. let $d_i = 1/n$. From McDiarmid inequality we have:

$$P(|L_n - E[L_n] > t) \leq 2\exp\left(-\frac{t^2}{2n(1/n)^2}\right) = 2\exp\left(-\frac{nt^2}{2}\right)$$

Thus we obtain a large deviations type bound on the difference $L_n - E[L_n]$.

Consider a simple undirected graph $G = (V, E)$. V is the set of nodes, and E is the set of edges which as a list of pairs $(i_1, j_1), \ldots, (i_{|E|}, j_{|E|})$ where $i_1, \ldots, i_{|E|}, j_1, \ldots, j_{|E|}$ are nodes. We can represent the graph as $n \times n$ zero-one matrix A where $A_{i,j} = 1$ if $(i, j) \in E$. Then A is symmetric matrix. A cut in this graph is a partition σ of nodes into two groups, encoded by $\sigma : V \to \{0, 1\}$. The value $MC(\sigma) = |\{(i, j) \in E : \sigma(i) \neq \sigma(j)\}|$. Clearly $MC(\sigma) \leq |E|$. At the same time, a random assignment $\sigma(i) = 0$ with probability $1/2$ and $= 1$ with probability $1/2$ gives a cut with expected value $MC(\sigma) \geq (1/2)|E|$. Now denote $MC(G)$ the maximum possible value of the cut: $MC(G) = \max_\sigma MC(\sigma)$. Thus $1/2 \leq MC(G)/|E| \leq 1$. Further, suppose we delete an arbitrary edge from graph G and obtain a new graph G'. $MC(G') \geq MC(G) - 1$ - the Max-Cut value either stasy the same or goes down by at most one. Similarly, when adding one edge, the Max-Cut value increases by at most one.

Example 7.2 (Max-Cut problem)**.** Suppose $G = G(n, dn)$ is a random Erdos-Renyi graph with $|E| = dn$ edges. Suppose we choose every edges $E_1, \ldots, E_{d_n}$ uniformly at random from the total set $\binom{n}{2}$ edges. Denote by MC_n the value of the maximum cut $MC(G(n, dn))$ on the random graph. Since the graph is random, we have that MC_n is a random variable. Furthermore, $d/2 \leq MC_n/n \leq d$. We can compute the scaling limit $E[MC_n]/n$ as $n \to \infty$. We can easily obtian bounds by Azuma-Hoeffding inequality. Let $g(E_1, \ldots, E_{dn}) = MC_n$ g is a function of dn i.i.d. random variables. Replacing one edge E_i by a different edge E_i' chagnes MC_n by at most one. Thus we can have:

$$P(|MC_n - E[MC_n]| \geq t) \leq 2\exp\left(-\frac{t^2}{2dn}\right)$$

Taking $t = rn$ where r is a constant we obtain a large deviations type bound $2\exp\left(-\frac{r^2 n}{2d}\right)$. Taking $t = r\sqrt{n}$ we obtain a Gaussian type bound $2\exp\left(-\frac{r^2}{2d}\right)$. Namely, $MC_n = E[MC_n] + \Theta(\sqrt{n})$. This is a meaningful Concentration around the mean since we have $E[MC_n] = \Theta(n)$.

7.4 Talagrand's inequality

Let $(\Omega_i, \mathcal{F}_i, \mu_i)$ be probability spaces $(i = 1, 2, \ldots, n)$. Let $\mu = \mu_1 \bigotimes \ldots \bigotimes \mu_n$ be product measure on $X = \Omega_1 \ldots \times \Omega_n$. Let $x = (x_1, \ldots, x_n) \in X$ be a point in this product space.

Hamming distance over X:

$$d(x, y) = |\{1 \leq i \leq n : x_i \neq y_i\}| = \sum_{i=1}^{n} 1_{\{x_i \neq y_i\}}.$$

α-weighted Hamming distance over X for $a\alpha\mathbb{R}_+^n$:

$$d_a(x,y) = \sum_{i=1}^{n} a_i 1_{\{x_i \neq y_i\}}.$$

Also $|\alpha| = \sqrt{\sum a_i^2}$.
control distance from a set: for set $A \subseteq X$ and $x \in X$:

$$D_A^c(x) = \sup_{|a|=1} d_a(x,A) = \inf\{d_a(x,y) : y \in A\}.$$

Theorem 7.5 (Talagrand)**.** *For every measurable non-empty set A and product-measure μ:*

$$\int \exp\left(\frac{1}{4}(D_A^c)^2\right) d\mu \leq \frac{1}{\mu(A)}$$

In particular,

$$\mu(\{D_A^c \geq t\}) \leq \frac{1}{\mu(A)} \exp\left(-\frac{t^2}{4}\right)$$

Chapter 8

Renewal Processes

8.1 Naive Introduction

Definition 8.1. A renewal process is a counting process where the inter-arrival times are i.i.d. (proper) random variables $T_1, T_2, \ldots$ We shall assume the non-trivial case that (i) $T_1 \neq 0$ a.s., and (ii) $E[T_1] \leq \infty$.

Let S_n denote the arrival time of consumer n, then:

$$S_n = \sum_{i=1}^{n} T_i$$

For given time t, let $N(t)$ denote the number of arrivals in the interval $[0, t]$. It can be expressed as: $N(t) = \max\{n : S_n \leq t\}$. Typically, we call $\{N(t), t \geq 0\}$ the renewal process. An alternative is: $N(t) = \sum_{n=1}^{\infty} 1[S_n \leq t]$.

The reason we call this renewal process is that at every renewal point, the process restarts itself, or renewal itself. Looking forward at any renewal point, the process is probabilistically the same.

Remark 8.1. If $T_1 \geq 0$ and $\neq 0$ a.s. then (i) $E[T_1] > 0$ (ii) there exists $\delta > 0$ such that $P(T_1 \geq \delta) > 0$.

Remark 8.2. It follows from SLLN that $\frac{S_n}{n} \to E[T_1] > 0$ This implies that $S_n \to \infty, a.s.$.

Remark 8.3.

- $N(t) < \infty$ a.s. for any $t < \infty$.
- $N(\infty) = \infty$ a.s. That is, $t \to \infty$, we have:

$$N(t) \to N(\infty) = \infty a.s.$$

Proof. For (ii), we have $P(N(\infty) < \infty) = P(\bigcup_{n=1}^{\infty}\{T_n = \infty\}) \leq \sum_{n=1}^{\infty} P(T_n = \infty) = 0.$ □

Proposition 8.1. *With probability 1* $\frac{N(t)}{t} \to \frac{1}{E[T_1]}; t \to \infty$. *Thus,* $1/E[T_1]$ *is the rate at which renewal occurs. We call it the rate of the renewal process. Let* $E[T_1] = \mu$.

Proof. $\frac{N(t)}{S_{N(t)+1}} \leq \frac{N(t)}{t} \leq \frac{N(t)}{S_{N(t)}}$ □

Example 8.1. A sequence of trivials. Each trivial has outcome i with probability $p_i, i = 1, \ldots, n$. The process ends until the same outcome occurs for a consecutive of k times.

We let the process continue without ending. A cycle is defined when any of the outcome appears k times. This gives rise to a renewal process. Let X_i denote the first time there is a consecutive of k outcome i. We have:

$$\begin{aligned}E[X_i] &= \sum_{l=1}^{\infty} E[X_i| \text{ first non-i outcome is at time } k](1-p_i)^{l-1}p_i \\ &= \sum_{l=1}^{k}(l+E[X_i](1-p_i)^{l-1}p_i)+(1-p_i)^k).\end{aligned}$$

Solving for $E[X_i]$ we have: $E[X_i] = \sum_{l=1}^{k} p_i^{-l}$. The rate at which i wins is:

$$\frac{1}{E[X_i]} = \frac{1}{\sum_{l=1}^{k} p_i^{-l}}$$

The rate of the renewal process is:

$$\sum_{j=1}^{n} \frac{1}{\sum_{l=1}^{k} p_j^{-l}}$$

The expected number of trials to end the game is:

$$\frac{1}{\sum_{j=1}^{n} \frac{1}{\sum_{l=1}^{k} p_j^{-l}}}$$

The rate for i win is the total rate multipled by the probability i win. Hence the probability that i wins is:

$$\frac{\frac{1}{\sum_{l=1}^{k} p_i^{-l}}}{\sum_{j=1}^{n} \frac{1}{\sum_{l=1}^{k} p_j^{-l}}}$$

Remark 8.4. $N(t)/t$ converges to $1/\mu$ a.s.

The average number of renewals is $m(t) = E[N(t)]$ is called the renewal function. Note the following is $\{N(t) \geq n\}$ iff $S_n \leq t$.

We hold the claim:

Claim 8.1. *For any $t \geq 0$, we have $m(t) < \infty$*

The elementary renewal Theorem says that $\frac{m(t)}{t} \to 1/\mu$ This is L^1 convergence, we cannot directly expect it is true from $\frac{N(t)}{t} \to \frac{1}{\mu} a.s.$

Renewal processes can be specified in three standard ways, first, by the joint distributions of the arrival epochs $S_1, S_2, \ldots$ second by the joint distributions of the interarrival times $X_1, X_2, \ldots$ and third, by the joint distribution of the counting r.v.'s $N(t)$ for $t > 0$. Each arrival epoch S_n is the sum $X_1 + X_2 + \ldots + X_n$ of n IID rv's. The arrival epochs and the counting rv's are related in each of the following ways:

$$\{S_n \leq t\} = \{N(t) \geq n\}; \quad \{S_n > t\} = \{N(t) < n\}$$

The $j$ths subsequent arrival epoch is at $S_{n+j} - S_n = X_{n+1} + \ldots + X_{n+j}$.

The strong law for renewal processes states taht this limiting time-average renewal rate exists for a set of ω that has probability 1, and that this limiting value is $1/\bar{X}$. We shall often refer to this result by the less precise statement that the time-average renewal rate is $1/\bar{X}$. This result is a direct consequence of the strong law of large number for IID rv's.

8.2 The strong law of large numbers and convergence WP1

8.2.1 Convergence with probability 1

Recall that a sequence $\{Z_n : n \geq 1\}$ of rv's on a simple space Ω is defined to converge WP1 to a rv Z on Ω if:

$$Pr\{\omega \in \Omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\} = 1$$

i.e., if the set of sample sequences $\{Z_n(\omega) : n \geq 1\}$ that converge to $Z(\omega)$ has probability 1. We can understand this by defining $Y_n = Z_n - Z$ for each n. This sequence $\{Y_n : n \geq 1\}$ then converges to 0 WP1 if and only if the sequence $\{Z_n : n \geq 1\}$ converges to Z WP1.

Lemma 8.1. *Let $\{Y_n : n \geq 1\}$ be a sequence of rv's, each with finite expectation. If $\sum_{n=1}^{\infty} E[|Y_n|] < \infty$, then $Pr\{\omega : \lim_{n\to\infty} Y_n(\omega) = 0\} = 1$*

Proof. For any $\alpha, 0 < \alpha < \infty$ and any integer $m \geq 1$, the Markov inequality says that:

$$Pr\{\sum_{n=1}^{m} |Y_n| > \alpha\} \leq \frac{E[\sum_{n=1}^{m} |Y_n|]}{\alpha} = \frac{\sum_{n=1}^{m} E[|Y_n|]}{\alpha}$$

Since $|Y_n$ is non-negative, $\sum_{n=1}^{m} |Y_n| > \alpha$ implies that $\sum_{n=1}^{m+1} |Y_n| > \alpha$. Thus the left hand side of the equation is non-decreasing in m and we can go to the limit:

$$\lim_{m\to\infty} Pr\{\sum_{n=1}^{m} |Y_n| > \alpha\} \leq \frac{\sum_{n=1}^{\infty} E[|Y_n|]}{\alpha}$$

now let $A_m = \{\omega : \sum_{n=1}^{m} |Y_n(\omega)| > \alpha\}$. The sequence $\{A_m : m \geq 1\}$ is nested, $A_1 \subseteq A_2 \subseteq \dots$, so from the axioms of probability:

$$\begin{aligned}\lim_{m\to\infty} Pr\{\sum_{n=1}^{m} |Y_n| > \alpha\} &= Pr(\bigcup_{m=1}^{\infty} A_m) \\ &= Pr(\omega : \sum_{n=1}^{\infty} |Y_n(\omega)| > \alpha)\end{aligned}$$

where we have used the fact that for any given $\omega : \sum_{n=1}^{\infty} |Y_n(\omega)| > \alpha$ if and only if $\sum_{n=1}^{m} |Y_n(\omega)| > \alpha$ for some $n \geq 1$. Combining together we have:

$$Pr\{\omega : \sum_{n=1}^{\infty} |Y_n(\omega)| > \alpha\} \leq \frac{\sum_{n=1}^{\infty} E[|Y_n|]}{\alpha}$$

Looking at the complementary set and assuming $\alpha > \sum_{n=1}^{\infty} E[|Y_n|]$ we have:

$$Pr\{\omega : \sum_{n=1}^{\infty} |Y_n(\omega)| \leq \alpha\} \geq 1 - \frac{\sum_{n=1}^{\infty} E[|Y_n|]}{\alpha}$$

For any ω such that $\sum_{n=1}^{\infty} |Y_n(\omega)| \leq \alpha$, we see that $\{|Y_n(\omega)|; n \geq 1\}$ is simply a sequence of non-negative numbers with a finite sum. Thus the individual numbers in that sequence must approach 0, i.e., $\lim_{n\to\infty} |Y_n(\omega)| = 0$ for each such ω. It follows then that:

$$Pr\{\omega : \lim_{n\to\infty} |Y_n(\omega)| = 0\} \geq Pr\{\omega : \sum_{n=1}^{\infty} |Y_n(\omega)| \leq \alpha\}$$

Combining together we have:

$$Pr\{\omega : \lim_{n\to\infty} |Y_n(\omega)| = 0\} \geq 1 - \frac{\sum_{n=1}^{\infty} E[|Y_n|]}{\alpha}$$

This is true for all α, so $Pr\{\omega : \lim_{n\to\infty} |Y_n| = 0\} = 1$ and thus $Pr\{\omega : \lim_{n\to\infty} Y_n = 0\} = 1$. □

Theorem 8.1 (Strong Law of Large Numbers (SLLN)). *For each integer $n \geq 1$, let $S_n = X_1 + \ldots + X_n$ where $X_1, X_2, \ldots$ are i.i.d. rv's satisfying $E[|X|] < \infty$ then:*

$$Pr\{\omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \bar{X}\} = 1.$$

Proof. Assume that $\bar{X} = 0$ and $E[X^4] < \infty$. Denote $E[X^4]$ by γ. For real number x, if $|x| \leq 1$ then $x^2 \leq 1$ and if $|x| > 1$ then $x^2 < x^4$ Thus $x^2 \leq 1 + x^4$ for all x. It follows that $\sigma^2 = E[X^2] \leq 1 + E[X^4]$ Thus σ^2 is finite if $E[X^4]$ is.

Now let $S_n = X_1 + \ldots + X_n$ where $X_1, \ldots, X_n$ are i.i.d. with the distribution of X.

$$\begin{aligned} E[S_n^4] &= E[(X_1 + \ldots + X_n)(X_1 + \ldots + X_n)\ldots] \\ &= E[(\sum_{i=1}^{n} X_i)(\sum_{j=1}^{n} X_j)(\sum_{k=1}^{n} X_k)(\sum_{l=1}^{n} X_l)] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} E[X_i X_j X_k X_l] \end{aligned}$$

For each $i, 1 \leq i \leq n$ there is a term in this sum with $i = j = k = l$. For each such term, $E[X^4] = \gamma$ There are n such terms and they collectively contribute $n\gamma$ to the sum $E[S_n^4]$. Also, for each $i, k \neq i$, there is a term with $j = i$ and $l = k$. For each of these $n(n-1)$ terms, $E[X_i X_j X_k X_l] = \sigma^4$. There are another $n(n-1)$ terms with $j \neq i$ and $k = i, l = j$. Each such term contributes σ^4 to the sum. Finally, for each $i \neq qj$ there is a term with $l = i, k = j$. Collectively all of these terms contribute $3n(n-1)\sigma^4$ to the sum. Each of the remaining terms is 0 since at least one of i, j, k, l is different from all the others. Thus we have:

$$E[S_n^4] = n\gamma + 3n(n-1)\sigma^4$$

now consider the sequence of rv's $\{S_n^4/n^4 : n \geq 1\}$.

$$\sum_{n=1}^{\infty} E[|\frac{S_n^4}{n^4}|] = \sum_{n=1}^{\infty} \frac{n\gamma + 3n(n-1)\sigma^4}{n^4} < \infty$$

where we have used the fact that the series $\sum_{n\geq 1} 1/n^2$ and the series $\sum_{n\geq 1} 1/n^3$ converges.

Using Lemma applied to $\{S_n^4/n^4; n \geq 1\}$ we see that $\lim_{n\to\infty} S_n^4/n^4 = 0$ WP1. For each ω such that $\lim_{n\to\infty} S_n^4(\omega)/n^4 = 0$, the non-negative fourth root of that sequence of non-negative numbers are also approaches 0. Thus $\lim_{n\to\infty} |S_n/n| = 0$ WP1 □

8.3 Strong law for renewal processes

Note that for any given t, $N(t)/t$ is the slope of a straight line from the origin to the point $(t, N(t))$. As t increases, this slope decreases in the interval between each adjacent pair of arrival epochs and then jumps up at the next arrival epoch. Note that t lies between the $N(t)$ arrival and the $N(t) + 1$th arrival. Thus for full sample points,

$$\frac{N(t)}{S_{N(t)}} \geq \frac{N(t)}{t} > \frac{N(t)}{S_{N(t)+1}}$$

We want to show that intuitively why the slope $N(t)/t$ in the figure approaches $1/\bar{X}$ as $t \to \infty$. As t increases, we would guess that $N(t)$ increases without bound. Since S_n/n converges to $\bar{X}$ WP1 from the strong law of large numbers, we would be brave enough to guess that n/S_n converges to $1/\bar{X}$.

Theorem 8.2 (Stong law for Renewal Process). *For a renewal process with mean inter-renewal interval $\bar{X} < \infty, \lim_{t\to\infty} N(t)/t = 1/\bar{X}$ WP1.*

Lemma 8.2. *Let $\{N(t) : t > 0\}$ be a renewal counting process with inter-renewal rv's $\{X_n : n \geq 1\}$. Then (whether or not $\bar{X} < \infty$), $\lim_{t\to\infty} N(t) = \infty$ WP1 and $\lim_{t\to\infty} E[N(t)] = \infty$.*

Proof. Note that for each sample point $\omega : N(t, \omega)$ is a non-decreasing real-valued function of t and thus either has a finite limit or an infinite limit. Using the result, the probability that this limit is finite with value less than any given n is:

$$\lim_{t\to\infty} Pr\{N(t) < n\} = \lim_{t\to\infty} Pr\{S_n > t\} = 1 - \lim_{t\to\infty} Pr\{S_n \leq t\}.$$

Since the X_i are r.v.'s the sums S_n are also rv's for each n and thus $\lim_{t\to\infty} Pr\{S_n \leq t\} = 1$ for each n. Thus $\lim_{t\to\infty} Pr\{N(t) < n\} = 0$ for each n. This shows that the set of sample points ω for which $\lim_{t\to\infty} N(t(\omega)) < n$ has probability 0 for all n. Thus the set of sample points for which $\lim_{t\to\infty} N(t, \omega)$ is finite has probability 0 and $\lim_{t\to\infty} N(t) = \infty$ WP1.

Next, $E[N(t)]$ is non-decreasing in t and thus has either a finite or infinite limit as $t \to \infty$. For each n, $Pr\{N(t) \geq n\} \geq 1/2$ for large enough t and therefore, $E[N(t)] \geq n/2$ for such t. Thus $E[N(t)]$ can have no finite limit as $t \to \infty$ and $\lim_{t\to\infty} E[N(t)] = \infty$. □

Lemma 8.3. *Let $\{Z_n : n \geq 1\}$ be a sequence of rv's such that $\lim_{n\to\infty} Z_n = \alpha$ WP1. Let f be a real valued function of a real variable that is continuous at α. Then:*

$$\lim_{n\to\infty} f(Z_n) = f(\alpha) WP1.$$

Proof. First, let $z_1, z_2, \ldots$ be a sequence of real numbers such that $\lim_{n\to\infty} z_n = \alpha$. Continuity of f at α means that for every $\varepsilon > 0$, there is a $\delta > 0$ usch that $|f(z) - f(\alpha)| < \varepsilon$ for all z such that $|z - \alpha| < \delta$. Also, since $\lim_{n\to\infty} z_n = \alpha$, we know that for every $\delta > 0$, there is an m such that $|z_n - \alpha| \leq \delta$ for all $n \geq m$. Putting these two statements, we know that for every $\varepsilon > 0$, there is an m such that $|f(z_n) - f(\alpha)| < \varepsilon$ for all $n \geq m$. Thus $\lim_{n\to\infty} f(z_n) = f(\alpha)$.

If ω is any sample point such that $\lim_{n\to\infty} Z_n(\omega) = \alpha$, then $\lim_{n\to\infty} f(Z_n(\omega)) = f(\alpha)$. Since this set of sample points has probability 1, the result holds. □

Proof. Proof of Strong law of renewal processes: Since $Pr\{X > 0\} = 1$ for a renewal process, we see that $\bar{X} > 0$. Choosing $f(x) = 1/x$, we see that $f(x)$ is continuous at $x = \bar{X}$. It follows from above lemma that:

$$\lim_{n\to\infty} \frac{n}{S_n} = \frac{1}{\bar{X}} WP1.$$

From above lemma, we know that $\lim_{t\to\infty} N(t) = \infty$ with probability 1, so, with probability 1, $N(t)$ increases through all the non-negative integers at t increases from 0 to ∞. Thus:

$$\lim_{t\to\infty} \frac{N(t)}{S_{N(t)}} = \lim_{n\to\infty} \frac{n}{S_n} = \frac{1}{\bar{X}} WP1.$$

Recall that $N(t)/t$ is sandwiched between $N(t)/S_{N(t)}$ and $(Nt)/S_{N(t)+1}$, so we can compute the proof by showing that $\lim_{t\to\infty} N(t)/S_{N(t)+1} = 1/\bar{X}$. To show this:

$$\lim_{t\to\infty} \frac{N(t)}{S_{N(t)+1}} = \lim_{n\to\infty} \frac{n}{S_{n+1}} = \lim_{n\to\infty} \frac{n+1}{S_{n+1}} \frac{n}{n+1} = \frac{1}{\bar{X}} WP1.$$

□

Given the strong law for $N(t)$, one would hypothesize that $E[N(t)/t]$ approaches $1/\bar{X}$ as $t \to \infty$. One might also hypothesize that $\lim_{t\to\infty} E[N(t+\delta) - N(t)]/\delta = 1/\bar{X}$, subject to some minor restrictions on δ.

Note that in order to equate time-averages and limiting ensemble-averages, quite a few conditions are required. First, the time-average must exist in the limit $t \to \infty$ with probability 1 and also have a fixed value with probability 1; second, the ensemble-averages must approach a limit as $t \to \infty$; and third, the limits must be the same.

Example 8.2. Let $\{X_i; i \geq 1\}$ be a sequence of binary i.i.d. rv, each taking the value 0 with probability 1/2 and 2 with probability 1/2. Letting $\{M_n : n \geq 1\}$ be the product process in which $M_n = X_1 X_2 \cdots X_n$. Since $M_n = 2^n$ if X_1 to X_n each take the value 2 and $M_n = 0$ otherwise, we see that $\lim_{n\to\infty} M_n = 0$ with probability 1. Also $E[M_n] = 1$ for all $n \geq 1$. Thus the time average exists and equals 0 with probability 1 and the ensamble-average exists and equal 1 for all n, but the two are diferent. The problem is that as n increases, $M_n = 2^n$ has a probability approaching 0, but still has a significant effect on the ensemble-averages.

Theorem 8.3 (Central Limit Theorem (CLT) for $N(t)$)**.** *Assume that the inter-renewal intervals for a renewal counting process $\{N(t) : t \geq 0\}$ have finite standard deviation $\sigma > 0$. Then:*

$$\lim_{t\to\infty} Pr\{\frac{N(t) - t/\bar{X}}{\sigma \bar{X}^{-3/2}\sqrt{t}} < \alpha\} = \Phi(\alpha)$$

where $\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\big(-x^2/2\big) dx$.

This says that the distribution function of $N(t)$ tends to the Gaussian distribution with mean $t/\bar{X}$ and standard deviation $\sigma\bar{X}^{-3/2}\sqrt{t}$.

This theorem can be proved by applying by the fact that CLT for a sum of IID rv's to S_n and then using the identity $\{S_n \leq t\} = \{N(t) \geq n\}$. For any real α, the CLT states that:

$$Pr\{S_n \leq n\bar{X} + \alpha\sqrt{n}\sigma\} \approx \Phi(\alpha),$$

where $\Phi(\alpha) = \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} \exp\big(-x^2/2\big) dx$ and where the approximation becomes exact in the limit $t \to \infty$. Letting:

$$t = n\bar{X} + \alpha\sqrt{n}\sigma,$$

and using $\{S_n \leq t\} = \{N(t) \geq n\}$ we have:

$$Pr\{N(t) \geq n\} \approx \Phi(\alpha)$$

Since t is monotonic in n for fixed α, we can express n in terms of t, getting:

$$n = \frac{t}{\bar{X}} - \frac{\alpha\sigma\sqrt{n}}{\bar{X}} \approx \frac{t}{\bar{X}} - \alpha\sigma t^{1/2}(\bar{X})^{-3/2}$$

Substituting back and establishes the theorem for $-\alpha$ which estbalishes the theorem since α is arbitrary.

8.4 Renewal-reward processes; time-averages

Along with a renewal counting processes $\{N(t); t \geq 0\}$, there is another randomly varying function of time, called a reward function $\{R(t); t \geq 0\}$. $R(t)$ models a rate at which the process is accumulating a reward. The important restriction on these reward functions is that $R(t)$ at a given t depends only on the location of t within the inter-reward interval containing t and perhaps other random variables local to that interval.

Example 8.3 (Time-average residual life)**.** For a renewal counting process $\{N(t), t > 0\}$, let $Y(t)$ be the residual life at time t. The residual life is defined as the interval from t until the next renewal epoch, i.e., as $S_{N(t)+1} - t$. We interpret $\{Y(t); t \geq 0\}$ as a reward function. The time-average of $Y(t)$, over the interval $(0, t]$ is given by $(1/t)\int_0^t Y(\tau)d\tau$. We are interested in the limit of this average as $t \to \infty$. Note that, for a given sample function $\{Y(t) = y(t)\}$, the integral $\int_0^t y(\tau)d\tau$ is simply a sum of isoceles right triangles, with a part of a final triangle at the end. Thus it can be expressed as:

$$\int_0^t y(\tau)d\tau = \frac{1}{2}\sum_{i=1}^{n(t)} x_i^2 + \int_{\tau=s_{n(t)}}^t y(\tau)d\tau.$$

where $\{x_i : 0 < i < \infty\}$ is the set of sample values for the inter-renewal intervals. Since the relationship holds for every sample point, we see that the random variable $\int_0^t Y(\tau)d\tau$ can be expressed in terms of the inter-renewal random variable X_n as:

$$\int_{\tau=0}^t Y(\tau)d\tau = \frac{1}{2}\sum_{n=1}^{N(t)} X_n^2 + \int_{\tau=S_{N(t)}}^t Y(\tau)d\tau.$$

Although the final term above can be easily evaluated for a given $S_{N(t)}(t)$, it is more convenient to use the following bound:

$$\frac{1}{2t}\sum_{n=1}^{N(t)} X_n^2 \leq \frac{1}{2}\int_{\tau=0}^t Y(\tau)d\tau \leq \frac{1}{2}\sum_{n=1}^{N(t)+1} X_n^2.$$

The term on the left can be evaluated in the limit $t \to \infty$ as follows:

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)} X_n^2}{2t} = \lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)} X_n^2}{N(t)} \frac{N(t)}{2t}.$$

Consider each term on the right side separately. For the first term, recall that $\lim_{t\to\infty} N(t) = \infty$ with probability 1. Thus as $t \to \infty$, $\sum_{n=1}^{N(t)} X_n^2/N(t)$ goes through the same set of values as $\sum_{n=1}^{k} X_n^2/k$ as $k \to \infty/$ Thus using the SLLN,

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)} X_n^2}{N(t)} = \lim_{k\to\infty} \frac{\sum_{n=1}^{k} X_n^2}{k} = E[X^2]WP1.$$

The second term on the right side is simply $N(t)/2t$ and using the strong law for renewal processes, $\lim_{t\to\infty} N(t)/2t = 1/(2E[X])$ WP1. Thus both limits exist WP1 and:

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)} X_N^2}{2t} = \frac{E[X^2]}{2E[X]}WP1$$

The right hand term is handled almost the same way:

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)+1} X_n^2}{2t} = \lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)+1} X_n^2}{N(t)+1} \frac{N(t)+1}{N(t)} \frac{N(t)}{2t} = \frac{E[X^2]}{2E[X]}$$

Combining these two results, we see that, with probability 1, the time-average residual life is given by:

$$\lim_{t\to\infty} \frac{\int_{\tau=0}^{t} Y(\tau)d\tau}{t} = \frac{E[X^2]}{2E[X]}$$

Example 8.4 (time-average Age). Let $Z(t)$ be the age of a renewal process at time t where age is defined as the interval from the most recent arrival before t until t, i.e., $Z(t) = t - S_{N(t)}$. By convention, if no arrivals have occured by time t, we take the age to be t. (i.e., in this case $N(t) = 0$ and we take S_0 to be 0)

The same analysis as before can be used to show that the time average of $Z(t)$ is the same as the time-average of the residual life:

$$\lim_{t\to\infty} \frac{\int_{\tau=0}^{t} Z(\tau)d\tau}{t} = \frac{E[X^2]}{2E[X]} WP1.$$

Example 8.5 (Time-average Duration). Let $\tilde{X}(t)$ be the duration of the inter-renewal interval containing time t, i.e., $\tilde{X}(t) = X_{N(t)+1} = S_{N(t)+1} - S_{N(t)}$. it is clear that $\tilde{X}(t) = Z(t) + Y(t)$ and thus the time-average of the duration is given by:

$$\lim_{t\to\infty} \frac{\int_{\tau=0}^{t} \tilde{X}(\tau)d\tau}{t} = \frac{E[X^2]}{E[X]} WP1.$$

In each of these examples, and in many other situations, we have a random function of time $(Y(t), Z(t), \tilde{X}(t))$ whose value at time t depends only on where t is the current inter-renewal interval. We now find the time-average value of $R(t)$, namely, $\lim_{t\to\infty} \frac{1}{t}\int_0^t R(\tau)d\tau$. Define R_n as the accumulated reward in the nth renewal interval:

$$R_n = \int_{S_{n-1}}^{S_n} R(\tau)d\tau = \int_{S_{(n-1)}}^{S_n} R[Z(\tau), \tilde{X}(\tau)]d\tau.$$

In general, since $Z(\tau) = \tau - S_{n-1}$ we have:

$$R_n = \int_{S_{n-1}}^{S_n} R(\tau - S_{n-1}, X_n)d\tau = \int_{z=0}^{X_n} R(z, X_n)dz$$

Note that R_n is a function only of X_n where the form of the function is determined by $R(Z, X)$. From this, it is clear that $\{R_n; n \geq 1\}$ is essentially a set of i.i.d. random variables. For residual life, $R(z, X_n) = X_n - z$, so the integral is $X_n^2/2$ as calculated before. In general, the expected value of R_n is given by:

$$E[R_n] = \int_{x=0}^{\infty}\int_{z=0}^{\infty} R(z, x)dz dF_X(x)$$

Breaking $\int_0^t R(\tau)d\tau$ into the reward over the successive renewal periods, we get:

$$\begin{aligned}\int_0^t R(\tau)d\tau &= \int_0^{S_1} R(\tau)d\tau + \int_{S_1}^{S_2} R(\tau)d\tau + \ldots + \int_{S_{(t)-1}}^{S_{N(t)}} R(\tau)d\tau + \int_{S_{N(t)}}^{t} R(\tau)d\tau \\ &= \sum_{n=1}^{N(t)} R_n + \int_{S_{N(t)}}^{t} R(\tau)d\tau.\end{aligned} \tag{8.1}$$

Theorem 8.4 (expected renewal process theorem). *Let $\{R(t); t \geq 0\} \geq 0$ be a nonnegative renewal reward function for a renewal process with expected inter-renewal time $E[X] = \bar{X} < \infty$. If $E[R_n] < \infty$, then with probability* 1*:*

$$\lim_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} R(\tau) d\tau = \frac{E[R_n]}{\bar{X}}$$

Proof. Using 8.1, the accumulated reward up to time t can be bounded between the accumulated reward up to the renewal before t and that to the next renewal after t,

$$\frac{\sum_{n=1}^{N(t)} R_n}{t} \leq \frac{\int_{\tau=0}^{t} R(\tau)d\tau}{t} \leq \frac{\sum_{n=1}^{N(t)+1} R_n}{t}.$$

The left hand side can be broken into: $\frac{\sum_{n=1}^{N(t)} R_n}{t} = \frac{\sum_{n=1}^{N(t)} R_n}{N(t)} \frac{N(t)}{t}$ Each R_n is given function of X_n, so the R_n are i.i.d. As $t \to \infty, N(t) \to \infty$, and thus, as we have seen before, the strong law of large numbers can be used on the first term on the right side, getting $E[R_n]$ WP1. Also the second term approaches $1/\bar{X}$ by the strong law for renewal processes. Since $0 < \bar{X} < \infty$ and $E[R_n]$ is finite, and the product of the two terms approaches the limit $E[R_n]/\bar{X}$. The right side of inequality is handled in almost the same way:

$$\frac{\sum_{n=1}^{N(t)+1} R_n}{t} = \frac{\sum_{n=1}^{N(t)+1} R_n}{N(t)+1} \frac{N(t)+1}{N(t)} \frac{N(t)}{t}.$$

It is sen that this term on the right side approaches limits as before and thus the term on the left side approaches $E[R_n]/\bar{X}$ WP1. □

Corollary 8.5. *Let $\{R(t); t > 0\}$ be a renewal-reward function for a renewal process with expected inter-renewal time $E[X] = \bar{X} < \infty$. If $E[R_n]$ exists, then with probability 1*

$$\lim_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} R(\tau) d\tau = \frac{E[R_n]}{\bar{X}}$$

Definition 8.2 (stopping trial)**.** A stopping trial (stopping time) J for a sequence of rv's $X_1, X_2, \ldots$ is a positive integer-valued rv such that for each $n \geq 1$, the indicator $I_{\{J=n\}}$ is a function of $\{X_1, \ldots, X_n\}$

Definition 8.3 (Generalized stopping trial)**.** A generalized stopping trial J for a sequence of pairs rv's $(X_1, V_1), \ldots$ is a positive integer-valued rv such that, for each $n \geq 1$, the indicator rv $I_{\{J=n\}}$ is a function of $X_1, V_1, X_2, V_2, \ldots, X_n, V_n$.

Theorem 8.6 (Generalized Wald's equality)**.** *Let $\{(X_n, V_n); n \geq 1\}$ be a sequence of pairs of rv's where each pair is independent and identically distributed to all other pairs. Assume that each X_i has finite mean $\bar{X}$. If J is a stopping trial for $\{(X_n, V_n); n \geq 1\}$ and if $E[J] < \infty$, then the sum $S_J = X_1 + X_2 + \ldots X_J$ satisfies:*

$$E[S_J] = \bar{X}E[J]$$

It is important here, as in many applications, to avoid the confusion created by viewing J as a stopping time. We have seen that J is the number of the first customer to see an empty queue, and S_J is the time until that customer arrives.

8.4.1 Little's Theorem

Example 8.6. Consider a $G/G/1$ queue and assume that the customers are served in First-Come-First-Served (FCFS) order. Both the interarrival intervals $\{X_i; i \geq 1\}$ and the service times $\{V_i; i \geq 0\}$ are assumed to be IID and several times are assumed to be independent of the interarrival intervals. A sample path for which $X_1 < V_0$, so arrival number 1 waits in queue for $W_1^q = V_0 - X_1$. if $X_1 \geq V_0$, on the other hand, then customer one enters service immediately, i.e., customer one "sees an empty system". In general, then

$W^q_1 = \max(V_0 - X_1, 0)$. In the same way if $W^q_1 > 0$, then customer 2 waits for $W^q_1 + V_1 - X_2$ is positive and 0 otherwise. In general:

$$W^q_i = \max(W^q_{i-1} + V_{i-1} - X_i, 0)$$

We use the G/G/1 queue with FCFS service as a specific example. The customers arrives in $[0, t]$, specifically including customer number arriving at $t = 0$. It illustrates the departure process $D(t)$, which is the number of departures up to time t, again including customer 0. The difference, $L(t) = A(t) - D(t)$ is then the customer in the system at time t. The essense of Little's Theorem can be seen by observing that $\int_0^{S^\tau_1} L(\tau)d\tau$ is the area between the upper and lower step functions, integrated out to the first time that the two step functions become equal. For the sample value, the integral is equal to $w_0 + w_1 + w_2$. In terms of the rv's

$$\int_0^{S^\tau_1} L(\tau)d\tau = \sum_{i=0}^{N(S^\tau_1)-1} W_i.$$

The same relationship exists in each inter-renewal interval, and in particular we can define L_n for each $n \geq 1$ as:

$$L_n = \int_{S^\tau_{n-1}}^{S^\tau_n} L(\tau)d\tau = \sum_{i=N(S^\tau_{n-1})}^{N(S^\tau_n)-1} W_i.$$

Renewals occur when the system goes from empty to busy, so the nth renewal is at the beginning of the nth busy period. Then L_n is the area of region between the two step functions over the nth busy period. Since the interarrival intervals and service times in each busy period are IID with respect to those in each other busy period, the sequence $L_1, L_2, \ldots$ is a sequence of IID rv's. The function $L(\tau)$ has the same behavior as a renewal reward function, but it is slightly more general, being a function of more than the age and duration of the renewal counting process $\{N^\tau(t); t > 0\}$ at $t = \tau$ However, the fact that $\{L_n; n \geq 1\}$ is an iid.

Theorem 8.7 (Little). *For a FCFS G/G/1 queue in which the expected inter-renewal interval is finite, the limiting time-average number of customers in the system is equal, with probability 1, to a constant denoted as $\bar{L}$. The sample-path-average waiting time per customer is also equal, with probability 1, to a constant denoted as $\bar{W}$. Finally, $\bar{L} = \lambda \bar{W}$ where λ is the customer arrival rate, i.e., the reciprocal of the expected interarrival time.*

Proof. Note that for any $t > 0$, $\int_0^t L(\tau)d\tau$ can be expressed as the sum over the busy periods completed before t plus a residual term involving the busy period including t. The residual term can be upper bound by the integral over that complete busy period. Using this we have:

$$\sum_{n=1}^{N^r(t)} L_n \leq \int_{\tau=0}^t L(\tau)d\tau \leq \sum_{i=0}^{N(t)} W_i \leq \sum_{n=1}^{N^r(t)+1} L_n.$$

Assuming the expected inter-renewal interval, $E[X^r]$ is finite, we can divide both sides by t and then we have $\lim_{t\to\infty} \frac{\sum_{i=0}^{N(t)} W_i}{t} = \frac{E[L_n]}{E[X^r]} WP1$.

The limit is called the arrival rate λ and is equal to the reciprocal of the term interarrival interval for $\{N(t)\}$. □

Chapter 9

Renewal Processes

Recap: Renewal process is a counting process with interarrival times being i.i.d. random variables that are not identically equal to 0. Let $T_1, T_2, \ldots$ denote the interarrival times and $S_1, S_2, \ldots$ the arrival times, then:

$$N(t) = \max\{t : S_n \leq t\} = \sum_{n=1}^{\infty} 1[S_n \leq t]$$

Let $E[T_1] = \mu$, we have shown that almost surely:

$$\frac{N(t)}{t} \to \frac{1}{\mu} \text{ (which follows from SLLN)}$$

An important question is: Does it mean that $E[N(t)]/t$ also converges to $1/\mu$? This is what we study next.

9.1 Naive Introduction

The average number of renewals, $m(t) = E[N(t)]$ is called the renewal function. Note the following $\{N(t) \geq n\}$ iff $S_n \leq t$.

We have the first claim that for any $t \geq 0$ $m(t) < \infty$)

The typical way to prove this is construct another renewal process and show that the new process which is greater than the first process is strictly less than ∞. We can also prove by definition: $m(t) = \sum_{n=1}^{\infty} P(N(t) \geq n)$ and under the assumption $\mu = E[X_i] < \infty$, it follows from the law of large number that $S_n/n \to \mu$ as $n \to \infty$. For large n if $n\mu > t$ the probability $P(N(t) \geq n)$ becomes very small, and eventually approaches to 0, i.e., it is decreasing with n. Therefore, it converges to a finite value.

Recall that $N(t)$ is the number of renewals by time t, $N(t) = \sum_{n=1}^{\infty} 1[S_n \leq t]$. The average number of renewals,

$$m(t) = E[N(t)] = \sum_{n=1}^{\infty} P(S_n \leq t)$$

is called the renewal function.

Proposition 9.1 (The renewal of equation). *The renewal function $m(t)$ satisfies the so-called renewal equation*

$$m(t) = F(t) + \int_0^t m(t-s)dF(s)$$

Proof. We can prove this by conditioning on the first renewal event.

$$\begin{aligned} m(t) &= \int_0^\infty E[N(t)\big|T_1 = s]dF(s) \\ &= \int_0^t E[N(t)\big|T_1 = s]dF(s) + \int_t^\infty E[N(t)\big|T_1 = s]dF(s) \\ &= \int_0^t (1 + m(t-s))dF(s) + 0 = F(t) + \int_0^t m(t-s)dF(s) \end{aligned}$$

□

Example 9.1. The inter-arrival times of a renewal process is uniform $[0, 1]$. Compute $m(t)$ for $0 \leq t \leq 1$:

$$\begin{aligned} & m(t) = t + \int_0^t m(t-s)ds = t + \int_0^t m(s)ds \\ & \implies m'(t) = m(t) + 1 \\ & \implies e^{-t}[m'(t) - m(t)] = 1 \\ & \implies m(t) = ce^t - 1 \end{aligned}$$

since $m(0) = 0$ we have $m(t) = ce^t - 1$

Very few renewal processes have closed form solution of $m(t)$ Another one is Poisson distribution. In a Poisson process, the expected number of renewals (or events) in a time interval of length t is directly proportional to the length of the interval, and the expected number of events by time t is $E[N(t)] = \lambda t$.

Any combination of uniform and Poisson distribution can be solved because they are all exponential growth.

Theorem 9.1 (Elementary Renewal Theorem)**.** *We have:*

$$\frac{m(t)}{t} \to \frac{1}{\mu}$$

As discussed earlier, this is L_1 convergence, we cannot directly expect it is true from $\frac{N(t)}{t} \to 1/\mu a.s.$

Proof. Recall that $m(t) < \infty$ for all $t \geq 0$, and thus by Wald's equation:

$$E[S_{N(t)+1}] = E[\sum_{i=1}^{N(t)+1} X_i] = (m(t) + 1)E[X_1]$$

On the other hand, $S_{N(t)+1} > t$ and hence, $E[S_{N(t)+1}] \geq t$. We obtain $(m(t) + 1)\mu \geq t$ or:

$$\frac{m(t)}{t} \geq \frac{1}{\mu} - \frac{1}{t}$$

Letting $t \to \infty$, we obtain:

$$\liminf_{t\to\infty} \frac{m(t)}{t} \geq \frac{1}{\mu}$$

To prove the other direction, consider another renewal process with $\bar{X}_i = \min\{X_i, M\}$ and the corresponding renewal process clearly satisfies $\bar{N}(t) \geq N(t)$, $\bar{m}(t) \geq m(t)$ and $\bar{S}_n(t) \leq S(t)$. In addition, $\bar{S}_{N(t)+1} \leq t + M$. Thus for any $t \geq 0$,

$$(\bar{m}(t) + 1)\bar{\mu}_M \leq t + M.$$

Letting $t \to \infty$, we obtain:

$$\limsup_{t\to\infty} \frac{\bar{\mu}(t)}{t} \leq \frac{1}{\bar{\mu}_M} \implies \limsup_{t\to\infty} \frac{m(t)}{t} \leq \frac{1}{\bar{\mu}_M}$$

Since this is for any $M > 0$, letting $M \to \infty$, we obtain: $\limsup_{t\to\infty} \frac{m(t)}{t} \leq \frac{1}{\mu}$. Putting the two inequalities together, we prove the **ELEMENTARY RENEWAL THEOREM** □

Definition 9.1 (Renewal reward process)**.** There is a renewal process, with inter-arrival times $T_1, T_2, \dots$ Associated with renewal cycle n is an reward R_n, which may depend on T_n, but $(T_1, R_1), (T_2, R_2), \dots$ are i.i.d. Let $R(n)$ denote the total rewards received by time t, that is, $R(t) = \sum_{n=1}^{N(t)} R_n$.

Remark 9.1. In the definition above, we assume that the reward is received at the end of the cycle. All the results presented below hold when the reward is received gradually during the cycle.

Theorem 9.2 (Renewal Reward Theorem)**.** *Suppose $E[T_1] < \infty$ and $E[R_1] < \infty$. Then:*
(i) With probability 1, $\frac{R(t)}{t} \to \frac{E[R_1]}{E[T_1]}$;
(ii) Its mean converges as well $\frac{E[R(t)]}{t} \to \frac{E[R_1]}{E[T_1]}$

Note that if each reward is identically 1, then this is just we have shown $\frac{N(t)}{t} \to \frac{1}{\mu}$.

Proof. The first follows from

$$\frac{R(t)}{t} = \frac{\sum_{n=1}^{\infty} R_n}{N(t)} \cdot \frac{N(t)}{t}$$

and SLLN. The proof of the second is similar to that of Elementary Renewal theorem, that uses Wald's equation. □

Example 9.2 (Car buying Model)**.** Car lifetime is random with cdf $H(\cdot)$ and pdf $h(\cdot)$. New car costs C_1 and displacement of a broken car costs C_2. Policy is to buy a new car either when the current one fails or it reaches T. What is the average cost per period?

From the Renewal Reward Theorem, we do not need to look at the whole process, and we only need to care about the first cycle.

$$E[R_n] = C_1 + C_2 P(\text{Life Time} < T) = C_1 + C_2 H(T)$$
$$E[T_1] = \int_0^{\infty} E[T_1 | h(t) < t] dt = T \cdot (1 - H(T)) + \int_0^T th(t)dt.$$

We can also consider a numerical case: $C_1 = 2000, C_2 = 500, h(t) = \frac{t}{10}, H(t) = \frac{1}{10}$ then the average time is 9.25

The following two examples are illustrated in Figure 9.1

Example 9.3 (Average Age of Renewal)**.** $A(t) = t - S_{N(t)}$ and the time average of the age of renewal process by time T is:

$$\frac{\int_0^T A(t)dt}{T} (\text{ we can also express it as: } \frac{\sum_{i=1}^{N(t)} A_i + \square}{T}$$

What is the long run time average of the age of renewal process?

We are interested in the limit of $\lim_{T\to\infty} \frac{\int_0^T A(t)dt}{T}$. By renewal reward theorem, the limit is equal to the ration of average reward during a cycle divided by the average length of a cycle, which is,

$$\frac{E[\int_0^{T_1} tdt]}{E[T_1]} = \frac{E[T_1^2]}{2E[T_1]}$$

Example 9.4 (Average excess of renewal)**.** $Y(t) = S_{N(t)+1} - t$. The time average of the residual life of the renewal process is:

$$\frac{\int_0^T Y(t)dt}{T}$$

What is the long run time average of the residual life of renewal process?

We are interested in the limit of $\lim_{T\to\infty} \frac{\int_0^T Y(t)dt}{T}$.

By Renewal Reward theorem, the limit is equal to the ration of average reward during a cycle divided by the average length of a cycle, which is

$$\frac{E[\int_0^{T_1} (T_1 - t)dt]}{E[T_1]} = \frac{E[T_1^2]}{2E[T_1]}$$

The average inter-arrival time containing the time is average age plus average residual life time

$$\frac{E[T_1^2]}{2E[T_1]} + \frac{E[T_1^2]}{2E[T_1]} = \frac{E[T_1^2]}{E[T_1]} \geq E[T_1]$$

The LHS is actually twice the RHS when T1 is exponentially distributed. Hence, the inter-arrival time containing t is longer than a standard inter-arrival time!

The intuition is that:

At any random point in time, you're more likely to find yourself in a longer inter-arrival period simply because longer intervals cover more time and thus have a higher probability of containing the random point. This is sometimes referred to as the **inspection paradox**: longer periods are "overrepresented" in random samples because they are more likely to be "inspected" or observed.

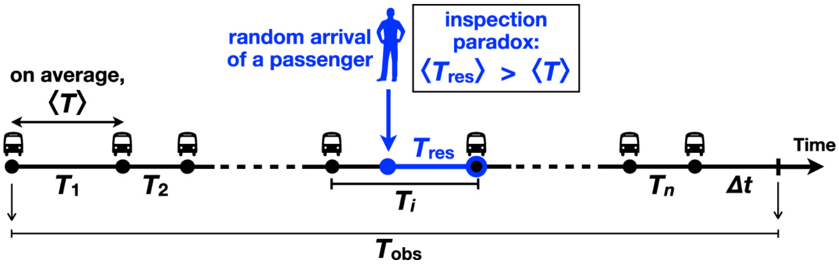

Figure 9.1: An illustration of the inspection paradox.

See the paper: The inspection paradox in stochastic resetting; 10.1088/1751-8121/ac3cdf

Appendix A

Mathematical Appendix

1. Exponential function e^x:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

2. Sine function $\sin(x)$:

$$\sin(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

3. Cosine function $\cos(x)$:

$$\cos(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots$$

4. Natural logarithm $\ln(1 + x)$:

$$\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$$

We say A is a π-system if it is closed under intersection, i.e., if $A, B \in \mathcal{A}$ then $A \cap B \in \mathcal{A}$. We say that $\mathcal{L}$ is a λ-system if: (i) $\Omega \in \mathcal{L}$; (ii): If $A, B \in \mathcal{L}$ and $A \subset B$, then $B - A \in \mathcal{L}$; (iii) if $A_n \in \mathcal{L}$, $A_n \uparrow A$ then $A \in \mathcal{L}$

Theorem A.1 (($\pi - \lambda$) 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 A.2 (Dynkin's π-λ Theorem). *Suppose $A_1, \ldots, A_n$ are independent and each A_i is a π-system, then $\sigma(A_1), \ldots, \sigma(A_n)$ are independent*

Proof. let $A_1, \ldots, A_n$ be sets with $A_i \in \mathcal{A}_i$ let $F = A_2 \cap \ldots \cap A_n$ and let $\mathcal{L} = \{A : P(A \cap F) = P(A)P(F)\}$. Since $P(\Omega \cap F) = P(\Omega)P(F)$, $\Omega \in \mathcal{L}$. To check (ii) of the definition of a λ-system we note that if $A, B \in \mathcal{L}$ with $A \subset B$ then $(B - A) \cap F = (B \cap F) - (A \cap F)$. So we have:

$$P((B - A) \cap F) = P(B \cap F) - P(A \cap F) = P(B)P(F) - P(A)P(F) = P(B - A)P(F)$$

and we have $B - A \in \mathcal{L}$. To check (iii) let $B_k \in \mathcal{L}$ with $B_k \uparrow B$ and note that $(B_k \cap F) \uparrow (B \cap F)$ we have:

$$P(B \cap F) = \lim_k P(B_k \cap F) = \lim_k P(B_k)P(F) = P(B)P(F)$$

Again the $\pi - \lambda$ theorem now gives $\mathcal{L} \supset \sigma(A_1)$. It follows that if $A_1 \in \sigma(A_1)$ and $A_i \in \sigma(A_2), 2 \leq i \leq n$, then:

$$P(\bigcap_{i=1}^{n} A_i) = P(A_1)P(\bigcap_{i=2}^{n} A_i) = \Pi_{i=1}^{n} P(A_i)$$

We therefore have:
If $\mathcal{A}_1, \ldots \mathcal{A}_n$ are independent then $\sigma(A_1), \sigma(A_2), \ldots \sigma(A_n)$ are independent. □

Sums of independent random variables:

Theorem A.3 (PDF Convolution Theorem). *If X and Y are independent, $F(x) = P(X \leq x), G(y) = P(Y \leq y)$ then:*

$$P(X + Y \leq z) = \int F(z - y) dG(y)$$

The integral on the right hand side is called the convolution of F and G and is denoted $F * G(z)$.

Proof. Let $h(x, y) = 1_{(x+y \leq z)}$. Let μ and v be the probability measures with distribution functions F and G respectively. Then:

$$\int h(x, y)\mu(dx) = \int 1_{(-\infty, z-y)}(x)\mu(dx) = F(z - y)$$

Then we have:

$$\begin{aligned} P(X + Y \leq z) &= \int\int 1_{(x+y \leq z)}\mu(dx)v(dy) \\ &= \int F(z - y)v(dy) = \int F(z - y)dG(y) \end{aligned}$$

The last equality is just a change of notation: We regard $dG(y)$ as a shorthand for ”integrate with respect to the measure v with distribution function G”. □

Theorem A.4 (Convolution of Probability Densities). *Suppose that X with density f and Y with distribution function G are independent. Then $X + Y$ has density*

$$h(x) = \int f(x - y)dG(y)$$

When Y has density g, the last formula can be written as:

$$h(x) = \int f(x - y)g(y)dy$$

Theorem A.5 (Kolmogorov's extension theorem). *Suppose we are given probability measures μ_n on $(\mathbb{R}^n, \mathcal{R}^n)$ that are consistent, that is:*

$$\mu_{n+1}((a_1, b_1] \times \ldots \times (a_n, b_n] \times \mathbb{R}) = \mu_n((a_1, b_1] \times \ldots \times (a_n, b_n])$$

Then there is a unique probability measure P on $(\mathbb{R}^N, \mathcal{R}^N)$ with:

$$P(\omega : \omega_i \in (a_i, b_i], 1 \leq i \leq n) = \mu_n((a_1, b_1] \times \ldots \times (a_n, b_n])$$

Probability Cheat Sheet

Distributions

Unifrom Distribution

`notation`	$U\left[a,b\right]$
`cdf`	$\frac{x-a}{b-a}$ for $x\in\left[a,b\right]$
`pdf`	$\frac{1}{b-a}$ for $x\in\left[a,b\right]$
`expectation`	$\frac{1}{2}\left(a+b\right)$
`variance`	$\frac{1}{12}\left(b-a\right)^2$
`mgf`	$\frac{e^{tb}-e^{ta}}{t\left(b-a\right)}$

`story:` all intervals of the same length on the distribution's support are equally probable.

Gamma Distribution

`notation`	$Gamma\left(k,\theta\right)$
`pdf`	$\frac{\theta^k x^{k-1}e^{-\theta x}}{\Gamma\left(k\right)}\mathbb{I}_{x>0}$ $\Gamma\left(k\right)=\int_0^\infty x^{k-1}e^{-x}dx$
`expectation`	$k\theta$
`variance`	$k\theta^2$
`mgf`	$\left(1-\theta t\right)^{-k}$ for $t<\frac{1}{\theta}$
`ind. sum`	$\sum_{i=1}^{n}X_i\sim Gamma\left(\sum_{i=1}^{n}k_i,\theta\right)$

`story:` the sum of k independent exponentially distributed random variables, each of which has a mean of θ (which is equivalent to a rate parameter of θ^{-1}).

Geometric Distribution

`notation`	$G\left(p\right)$
`cdf`	$1-\left(1-p\right)^k$ for $k\in\mathbb{N}$
`pmf`	$\left(1-p\right)^{k-1}p$ for $k\in\mathbb{N}$
`expectation`	$\frac{1}{p}$
`variance`	$\frac{1-p}{p^2}$
`mgf`	$\frac{pe^t}{1-\left(1-p\right)e^t}$

`story:` the number X of Bernoulli trials needed to get one success. Memoryless.

Poisson Distribution

`notation`	$Poisson\left(\lambda\right)$
`cdf`	$e^{-\lambda}\sum_{i=0}^{k}\frac{\lambda^i}{i!}$
`pmf`	$\frac{\lambda^k}{k!}\cdot e^{-\lambda}$ for $k\in\mathbb{N}$
`expectation`	λ
`variance`	λ
`mgf`	$\exp\left(\lambda\left(e^t-1\right)\right)$
`ind. sum`	$\sum_{i=1}^{n}X_i\sim Poisson\left(\sum_{i=1}^{n}\lambda_i\right)$

`story:` the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event.

Normal Distribution

`notation`	$N\left(\mu,\sigma^2\right)$
`pdf`	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/\left(2\sigma^2\right)}$
`expectation`	μ
`variance`	σ^2
`mgf`	$\exp\left(\mu t+\frac{1}{2}\sigma^2t^2\right)$
`ind. sum`	$\sum_{i=1}^{n}X_i\sim N\left(\sum_{i=1}^{n}\mu_i,\sum_{i=1}^{n}\sigma_i^2\right)$

`story:` describes data that cluster around the mean.

Standard Normal Distribution

`notation`	$N\left(0,1\right)$
`cdf`	$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}dt$
`pdf`	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
`expectation`	$\frac{1}{\lambda}$
`variance`	$\frac{1}{\lambda^2}$
`mgf`	$\exp\left(\frac{t^2}{2}\right)$

`story:` normal distribution with $\mu=0$ and $\sigma=1$.

Exponential Distribution

`notation`	$exp\left(\lambda\right)$
`cdf`	$1-e^{-\lambda x}$ for $x\geq0$
`pdf`	$\lambda e^{-\lambda x}$ for $x\geq0$
`expectation`	$\frac{1}{\lambda}$
`variance`	$\frac{1}{\lambda^2}$
`mgf`	$\frac{\lambda}{\lambda-t}$
`ind. sum`	$\sum_{i=1}^{k}X_i\sim Gamma\left(k,\lambda\right)$
`minimum`	$\sim exp\left(\sum_{i=1}^{k}\lambda_i\right)$

`story:` the amount of time until some specific event occurs, starting from now, being memoryless.

Binomial Distribution

`notation`	$Bin(n,p)$
`cdf`	$\sum_{i=0}^{k}\binom{n}{i}p^i\left(1-p\right)^{n-i}$
`pmf`	$\binom{n}{i}p^i\left(1-p\right)^{n-i}$
`expectation`	np
`variance`	$np\left(1-p\right)$
`mgf`	$\left(1-p+pe^t\right)^n$

`story:` the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p.

Basics

Comulative Distribution Function

$F_X\left(x\right)=\mathbb{P}\left(X\leq x\right)$

Probability Density Function

$F_X\left(x\right)=\int_{-\infty}^{\infty}f_X\left(t\right)dt$

$\int_{-\infty}^{\infty}f_X\left(t\right)dt=1$

$f_X\left(x\right)=\frac{d}{dx}F_X\left(x\right)$

Quantile Function

The function $X^*:[0,1]\to\mathbb{R}$ for which for any $p\in[0,1]$, $F_X\left(X^*\left(p\right)^-\right)\leq p\leq F_X\left(X^*\left(p\right)\right)$

$F_{X^*}=F_X$

$\mathbb{E}\left(X^*\right)=\mathbb{E}\left(X\right)$

Expectation

$\mathbb{E}\left(X\right)=\int_0^1X^*(p)dp$

$\mathbb{E}\left(X\right)=\int_{-\infty}^{0}F_X\left(t\right)dt+\int_0^\infty\left(1-F_X\left(t\right)\right)dt$

$\mathbb{E}\left(X\right)=\int_{-\infty}^{\infty}xf_Xxdx$

$\mathbb{E}\left(g\left(X\right)\right)=\int_{-\infty}^{\infty}g\left(x\right)f_Xxdx$

$\mathbb{E}\left(aX+b\right)=a\mathbb{E}\left(X\right)+b$

Variance

$\mathrm{Var}\left(X\right)=\mathbb{E}\left(X^2\right)-\left(\mathbb{E}\left(X\right)\right)^2$

$\mathrm{Var}\left(X\right)=\mathbb{E}\left(\left(X-\mathbb{E}\left(X\right)\right)^2\right)$

$\mathrm{Var}\left(aX+b\right)=a^2\mathrm{Var}\left(X\right)$

Standard Deviation

$\sigma\left(X\right)=\sqrt{\mathrm{Var}\left(X\right)}$

Covariance

$\mathrm{Cov}\left(X,Y\right)=\mathbb{E}\left(XY\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)$

$\mathrm{Cov}\left(X,Y\right)=\mathbb{E}\left(\left(X-\mathbb{E}\left(x\right)\right)\left(Y-\mathbb{E}\left(Y\right)\right)\right)$

$\mathrm{Var}\left(X+Y\right)=\mathrm{Var}\left(X\right)+\mathrm{Var}\left(Y\right)+2\mathrm{Cov}\left(X,Y\right)$

Correlation Coefficient

$\rho_{X,Y}=\frac{\mathrm{Cov}\left(X,Y\right)}{\sigma_X,\sigma_Y}$

Moment Generating Function

$M_X\left(t\right)=\mathbb{E}\left(e^{tX}\right)$

$\mathbb{E}\left(X^n\right)=M_X^{(n)}\left(0\right)$

$M_{aX+b}\left(t\right)=e^{tb}M_{aX}\left(t\right)$

Joint Distribution

$\mathbb{P}_{X,Y}(B) = \mathbb{P}((X,Y) \in B)$
$F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y)$

Joint Density

$$\mathbb{P}_{X,Y}(B) = \iint_B f_{X,Y}(s,t)\,dsdt$$

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)\,dtds$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(s,t)\,dsdt = 1$$

Marginal Distributions

$\mathbb{P}_X(B) = \mathbb{P}_{X,Y}(B \times \mathbb{R})$
$\mathbb{P}_Y(B) = \mathbb{P}_{X,Y}(\mathbb{R} \times Y)$

$$F_X(a) = \int_{-\infty}^{a}\int_{-\infty}^{\infty} f_{X,Y}(s,t)\,dtds$$

$$F_Y(b) = \int_{-\infty}^{b}\int_{-\infty}^{\infty} f_{X,Y}(s,t)\,dsdt$$

Marginal Densities

$$f_X(s) = \int_{-\infty}^{\infty} f_{X,Y}(s,t)dt$$

$$f_Y(t) = \int_{-\infty}^{\infty} f_{X,Y}(s,t)ds$$

Joint Expectation

$$\mathbb{E}(\varphi(X,Y)) = \iint_{\mathbb{R}^2} \varphi(x,y)\,f_{X,Y}(x,y)\,dxdy$$

Independent r.v.

$\mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X \le x)\,\mathbb{P}(Y \le y)$
$F_{X,Y}(x,y) = F_X(x)\,F_Y(y)$
$f_{X,Y}(s,t) = f_X(s)\,f_Y(t)$
$\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$
$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
Independent events:
$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$

Conditional Probability

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

`bayes` $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$

Conditional Density

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

$$f_{X|Y=n}(x) = \frac{f_X(x)\,\mathbb{P}(Y=n \mid X=x)}{\mathbb{P}(Y=n)}$$

$$F_{X|Y=y} = \int_{-\infty}^{x} f_{X|Y=y}(t)\,dt$$

Conditional Expectation

$$\mathbb{E}(X \mid Y=y) = \int_{-\infty}^{\infty} x f_{X|Y=y}(x)\,dx$$

$\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}(X)$
$\mathbb{P}(Y=n) = \mathbb{E}(\mathbb{I}_{Y=n}) = \mathbb{E}(\mathbb{E}(\mathbb{I}_{Y=n} \mid X))$

Sequences and Limits

$$\limsup A_n = \{A_n \text{ i.o.}\} = \bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n$$

$$\liminf A_n = \{A_n \text{ eventually}\} = \bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n$$

$\liminf A_n \subseteq \limsup A_n$
$(\limsup A_n)^c = \liminf A_n^c$
$(\liminf A_n)^c = \limsup A_n^c$

$$\mathbb{P}(\limsup A_n) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{n=m}^{\infty} A_n\right)$$

$$\mathbb{P}(\liminf A_n) = \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{n=m}^{\infty} A_n\right)$$

Borel-Cantelli Lemma

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \Rightarrow \mathbb{P}(\limsup A_n) = 0$$

And if A_n are independent:

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \Rightarrow \mathbb{P}(\limsup A_n) = 1$$

Convergence

Convergence in Probability

`notation` $X_n \xrightarrow{p} X$

`meaning` $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$

Convergence in Distribution

`notation` $X_n \xrightarrow{D} X$

`meaning` $\lim_{n\to\infty} F_n(x) = F(x)$

Almost Sure Convergence

`notation` $X_n \xrightarrow{a.s.} X$

`meaning` $\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$

Criteria for a.s. Convergence

- $\forall\varepsilon\exists N\forall n > N : \mathbb{P}(|X_n - X| < \varepsilon) > 1 - \varepsilon$
- $\forall\varepsilon\mathbb{P}(\limsup(|X_n - X| > \varepsilon)) = 0$
- $\forall\varepsilon \sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty$ (by B.C.)

Convergence in L_p

`notation` $X_n \xrightarrow{L_p} X$

`meaning` $\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0$

Relationships

$$\begin{array}{ccccc} \xrightarrow{L_q} & \underset{q>p\ge1}{\Rightarrow} & \xrightarrow{L_p} & & \\ & & \Downarrow & & \\ \xrightarrow{a.s.} & \Rightarrow & \xrightarrow{p} & \Rightarrow & \xrightarrow{D} \end{array}$$

If $X_n \xrightarrow{D} c$ then $X_n \xrightarrow{p} c$
If $X_n \xrightarrow{p} X$ then there exists a subsequence n_k s.t. $X_{n_k} \xrightarrow{a.s.} X$

Laws of Large Numbers

If X_i are i.i.d. r.v.,

`weak law` $\overline{X_n} \xrightarrow{p} \mathbb{E}(X_1)$

`strong law` $\overline{X_n} \xrightarrow{a.s.} \mathbb{E}(X_1)$

Central Limit Theorem

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

If $t_n \to t$, then

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le t_n\right) \to \Phi(t)$$

Inequalities

Markov's inequality

$$\mathbb{P}(|X| \ge t) \le \frac{\mathbb{E}(|X|)}{t}$$

Chebyshev's inequality

$$\mathbb{P}(|X - \mathbb{E}(X)| \ge \varepsilon) \le \frac{\mathrm{Var}(X)}{\varepsilon^2}$$

Chernoff's inequality

Let $X \sim Bin(n,p)$; then:
$\mathbb{P}(X - \mathbb{E}(X) > t\sigma(X)) < e^{-t^2/2}$
Simpler result; for every X:
$\mathbb{P}(X \ge a) \le M_X(t)\,e^{-ta}$

Jensen's inequality

for φ a convex function, $\varphi(\mathbb{E}(X)) \le \mathbb{E}(\varphi(X))$

Miscellaneous

$$\mathbb{E}(Y) < \infty \iff \sum_{n=0}^{\infty} \mathbb{P}(Y > n) < \infty \ (Y \ge 0)$$

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} \mathbb{P}(X > n) \ (X \in \mathbb{N})$$

$X \sim U(0,1) \iff -\ln X \sim exp(1)$

Convolution

For ind. X, Y, $Z = X + Y$:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(s)\,f_Y(z-s)\,ds$$

Kolmogorov's 0-1 Law

If A is in the tail σ-algebra $\mathcal{F}^t$, then $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$

Ugly Stuff

cdf of Gamma distribution:

$$\int_0^t \frac{\theta^k x^{k-1} e^{-\theta k}}{(k-1)!}dx$$

This cheatsheet was made by Peleg Michaeli in January 2010, using LaTeX.
version: 1.01
comments: peleg.michaeli@math.tau.ac.il