Probability Theory

Wenda Chu

Probability: *Study of regular patterns that arise from random phenomena*

Statistics: *The inverse problem of making inferences from observations of patterns in random phenomena.*

Contents

1 Preliminary: Measure Theory

Measures are defined to characterize the "mass" of a set, which is crucial in defining the Lebesgue integral, probability measure, etc. In order to define measures, we need first assume some structures on the space of sets, i.e., σ-Algebra.

1.1 σ-Algebra

Definition 1.1 (σ-Algebra). *A σ-algebra on X is a family of subsets of X, $\mathcal{F}$, such that,*

- $\varnothing, X \in \mathcal{F}$
- *If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$*
- *Let $\mathcal{I}$ be a countable set. If $\forall A_i \in \mathcal{F}$, for $i \in \mathcal{I}$, then $\bigcup_{i=1}^{\mathcal{I}} A_i \in \mathcal{F}$ and $\bigcap_{i\in\mathcal{I}} A_i \in \mathcal{F}$.*

Definition 1.2 (Borel σ-Algebra). *Borel σ-Algebra $B(X)$ is the σ-Algebra generated by open sets in X.*

$$B(X) = B(\mathcal{T}_X) \tag{1}$$

where $(X, \mathcal{T}_X)$ is a topological space.

1.2 Measures

Definition 1.3 (Measures). *A measure $\mu : \mathcal{F} \to [0, \infty]$ is defined on a space with σ-Algebra $(X, \mathcal{F})$, such that*

- $\mu(\varnothing) = 0$.
- $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$, *for disjoint sets $\{A_i\}$.*

This space is then called measurable space $(X, \mathcal{F}, \mu)$.

Definition 1.4 (Probability measure). *A probability measure μ on X is a specific measure with $\mu(X) = 1$.*

Other measures:

- Dirac measure:
$$\delta_t(A) := \begin{cases} 1, t \in A \\ 0, t \notin A \end{cases} \tag{2}$$
- Counting measures
$$\#(A) = \#\text{ elements in } A \tag{3}$$

1.3 Some related definitions

Finite measure: If $\mu(X) < \infty$, then μ is finite.

σ-finite measure: If we can cover X by countably many measurable sets each with finite measure, we call it a σ-finite measure.

A straightforward claim: finite measures are σ-finite measures. "Length" on $\mathbb{R}$ is a σ-finite measure but not finite.

Negligible: If a measurable set A has $\mu(A) = 0$ then it is a negligible set for μ.

Almost everywhere: If a measurable set A has $\mu(A^C) = 0$ then it is μ-almost everywhere, denoted as $\mu-$a.e.

1.4 Lebesgue measure

Borel measure: $B(\mathbb{R}) \to [0, \infty]$.

Lebesgue measures are defined on Borel σ-Algebra. For example, on $\mathbb{R}$:

$$\lambda(A) := \inf\{\sum_{i=1}^{\infty} |b_i - a_i| : A \subseteq \bigcup_{i=1}^{\infty} (a_i, b_i)\}, \forall A \in B(\mathbb{R}). \tag{4}$$

Intuition: to find the minimum union of intervals that covers A. (Note: union of intervals on $\mathbb{R}$ can always be written as a countable union of intervals).

- It is a measure
- $\lambda((a, b)) = |b - a|$, for $a < b$.
- Translation invariance: $\lambda(A + t) = \lambda(A)$ for all $A \in B(\mathbb{R})$.

Lebesgue measure is the *only* Borel measure that satisfies the above properties.

Another example of Borel measure: is the *cumulative distribution function*, which is finite and monotone.

1.5 Integration

Riemann Integral. Let $a \leq b \in \mathbb{R}$. We say f is Riemann Integrable if the following limit exists:

$$\lim_{\max \Delta x_k \to 0} \sum_{k=1}^{n} f(x_k^*) \Delta x_k, \tag{5}$$

where x_k^* is an arbitrary point in the interval Δx_k.

However, Riemann integral may not exist. For example, $f_n(q_i) = 1$ for rational numbers in $[0, 1]$ and $f_n(x) = 0$ otherwise. No matter how small an interval Δx_k could be, it contains rational numbers and irrational numbers.

Definition 1.5 (Decreasing Rearrangement). *For a function $f : \mathbb{R} \to \mathbb{R}_+$, the dereasing rearrangement $h : \mathbb{R}_+ \to \mathbb{R}_+$ is defined as following:*

$$h(t) := \lambda\{x \in \mathbb{R} L f(x) > t\} for\ t \geq 0. \tag{6}$$

$h(t)$ is a decreasing function, as measure h is monotonic.

Remark. Given the definition of h, our expectation is to define integration by:

$$\int_{\mathbb{R}} f(x)\lambda(x) = \int_0^\infty \lambda\{x \in \mathbb{R} : f(x) > t\}\mathrm{d}t = \int_0^\infty h(t)\mathrm{d}t. \tag{7}$$

The intuition of this definition is: instead of counting the area by x-axis, we count the area by y-axis.

In order to ensure this integration works, we need $\{x \in \mathbb{R} : f(x) > t\}$ to be measurable. The definition below of measurable function solves this problem.

Definition 1.6 ((Borel) Measurable function). *Let $(X, \mathcal{F})$ be a measurable space. A function $f : X \to \mathbb{R}$ is Borel-measurable if*

$$f^{-1}((t, \infty)) := \{x \in X : f(x) > t\} \in \mathcal{F} \tag{8}$$

Proposition 1.7. *A function $f : X \to \mathbb{R}$ is Borel measurable if and only if*

$$f^{(-1)}(B) \in \mathcal{F}, \tag{9}$$

for all Borel $B \in \mathcal{B}(\mathbb{R})$.

Properties of measurable functions:

- If f, g measurable, $f \circ g$ is measurable (for $X = \mathbb{R}$),
- $f + g$ is measurable,
- $f_+(x) := \max\{f(x), 0\}, f_(x) = \max_{-f(x),0}$ are measurable,
- $|f|$ is measurable,
- fg is measurable.
- If $f_i : X \to \mathbb{R} \cup \{-infty, \infty\}$ are measurable functions, $\sup_{i\in\mathbb{N}} f_i, \inf_{i\in\mathbb{N}} f_i$ are measurable.

Given the rigorous definition above, we can finally define Lebesgue integration. We first define it on positive functions and use it as a tool to define Lebesgue-integrable function in general.

Definition 1.8 (Lebesgue integral of positive function). *Let μ be a Borel measure. Let $f : \mathbb{R} \to \mathbb{R}_+ \cup \{\infty\}$ be a positive, measurable function. Then we define the Lebesgue integral of f with respect to μ with*

$$\int_{\mathbb{R}} f(x)\mu(\mathrm{d}x) = \int_0^\infty \mu\{x \in \mathbb{R} : f(x) > t\}\mathrm{d}t \tag{10}$$

Definition 1.9 (Lebesgue integrable function)**.** *A finite valued measurable function $f : \mathbb{R} \to \mathbb{R}$ is integrable with respect to Borel measure μ if*

$$\int_{\mathbb{R}} |f(x)|\mu(\mathrm{d}x) < \infty. \tag{11}$$

Its Lebesgue integral is defined as

$$\int_{\mathbb{R}} f(x)\mu(\mathrm{d}x) := \int_{\mathbb{R}} f_+(x)\mu(\mathrm{d}x) - \int_{\mathbb{R}} f_-(x)\mu(\mathrm{d}x). \tag{12}$$

notated by $\mu(f)$.

Proposition 1.10. *A bounded function over a bounded domain $f : [a, b] \to \mathbb{R}$ is a Riemann integrable function, then f is Lebesgue integrable with respect to the Lebesgue measure λ.*

$$f_{[a,b]}f(x)\lambda(\mathrm{d}x) = \int_a^b f(x)\mathrm{d}x. \tag{13}$$

Proposition 1.11 (Monotone Convergence)**.** *Let $(f_j : X \to \mathbb{R}_+ \cup \{\infty\})j \in \mathbb{N}$ are positive, measurable functions. $f_j(x)$ increasing in terms of j, and $\lim_{j\to\infty} f_j(x) = f(x)$ for all $x \in X$. Then the integral*

$$\lim_{j\to\infty} \mu(f_j) = \mu(f). \tag{14}$$

2 Mathematical Foundation of Probability

Developed by Kolmogorov in 1933, an axiomatic foundation for the theory of probability.

Definition 2.1 (Probability Space, ANK 1933)**.** *A triple* $(\Omega, \mathcal{F}, \mathbb{P})$*, where*

- Ω *is a set called the sample space.*
- $\mathcal{F}$ *is a* σ*-Algebra of* Ω*.*
- $\mathbb{P}$ *is a probability measure on* $(\Omega, \mathcal{F})$ *assigning probability to events.*

Notes:

- $\omega \in \Omega$ are called sample points, outcomes of a probability experiment in simple cases.
- $\mathcal{F}$ identifies sets of sample points that are assigned probabilities. $E \in \mathcal{F}$ are called events.
- When $\omega_0 \in E$ occurs, we say the event E occurs.
- If Ω is finite or countable, then we can take $\mathcal{F} = P(\Omega)$. However for $\mathbb{R}^n$, the power set is too big for a reasonable measurable σ-Algebra, so we take Borel σ-Algebra $B(\mathbb{R}^n)$.

States in Ω give a complete system description, but it's often too detailed and inaccessible. Instead, we are more interested in observables, or system statistics, which are called random variables in probability theory.

Definition 2.2 (Real Random Variable)**.** *Let* $\Omega, \mathcal{F}, \mathbb{P}$ *be a probability space. A real random variable is a measurable function*

$$X : \Omega \to \mathbb{R} \tag{15}$$

Remarks.

X is a fixed function, so where does the "randomness" come from? In fact, all the randomness comes from the $X(\omega)$ where ω is random in the probability space.

Distributions of probability over the events induce a distribution of probability over the values in $\mathbb{R}$ of the random variable.

The term "measurable function" used above is defined as:

Definition 2.3 (Measurable function)**.** *A measurable function* $X : \Omega \to \mathbb{R}$ *has the property that* $X^{-1}(B) \in \mathcal{F}$ *for all* $B \in \mathcal{B}(\mathbb{R})$*.*

Based on this property, we can compute the probability that X takes a value in $B \in \mathcal{B}(\mathbb{R})$:

$$\mathbb{P}(X \in B) := \mathbb{P}\{\omega \in \Omega : X(\omega \in B\} = \mathbb{P}(X^{-1}(B)). \tag{16}$$

Definition 2.4 (Law of a real random variable)**.** *Let* $\Omega, \mathcal{F}, \mathbb{P}$ *be a probability space. Let* $X : \Omega \to \mathbb{R}$ *be a real random variable.*

The law (or distribution) of X *is the Borel probability measure* μ_X *on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, s.t.,*

$$\mu_X(B) = \mathbb{P}\{X \in B\}, \forall B \in (\mathbb{R}). \tag{17}$$

Notes. The probability that $X \in B$ can be expressed as

$$\mathbb{P}(X \in B) = \int_{\mathbb{R}} \mathbb{1}_B(x)\mu_X(\mathrm{d}x). \tag{18}$$

Definition 2.5 (Distribution function)**.** *Let X be a real random variable on a probability space with law μ_X. Define*

$$F_X(a) := \mathbb{P}\{X \leq a\} = \mu_X((-\infty, a]), \forall a \in \mathbb{R}. \tag{19}$$

This is called a cumulative distribution function (CDF) of X.

Note: It is important to make sure it's a $\leq$ but not a $<$ in the definition.

Properties of CDF:

- Monotonicity
- Asymptotic
- Right-continuous
- Law: $\mu_X(a, b] = F_X(b) - F_X(a)$.

Continuous random variables: Law has a density with respect to the Lebesgue measure λ.

$$\mu_X(B) = \int_B f_X(x)\lambda(\mathrm{d}x). \tag{20}$$

where $f_X(x)$ is the probability density function. In particular, F_X is absolutely continuous.

Singular continuous: F_X is continuous but μ_X has no densit1y functions.

3 Expectation

3.1 Definitions

Definition 3.1 (Expectation). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a real random variable.*

The expectation $\mathbb{E}[X]$ is the real number given by the integral:

$$\mathbb{E}[X] := \int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega) = \int_\Omega X \mathrm{d}\mathbb{P}. \tag{21}$$

More formally, we define it using the Lebesgue integral:

$$\mathbb{E}[X] := \int_0^\infty \mathbb{P}\{X > t\}\lambda(\mathrm{d}t), \text{ when } X \geq 0. \tag{22}$$

For real X (not necessarily positive), if $\mathbb{E}[|X|] < \infty$, we define $\mathbb{E}[X] := \mathbb{E}[X_+] - \mathbb{E}[X_-]$.

Definition 3.2 (Integrable random variable.). *For a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we define*

$$L_1 := L_1\{\Omega, \mathcal{F}_1, \mathbb{P}\} = \{X : \Omega \to \mathbb{R} : \mathbb{E}[|X|] < \infty\}. \tag{23}$$

Note: Not all random variables are integrable, e.g., Cauchy random variables.

3.2 Change of variables

How to compute the expectation using the law? We have the change of variables proposition.

Proposition 3.3 (Law of the unconscious statistician, Lotus). *Let X be a real random variable with law μ_X. Let $h : \mathbb{R} \to \mathbb{R}$ be a function that is μ_X-integrable. Then*

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)\mu_X(\mathrm{d}x) \tag{24}$$

Example. If X is continuous with a density f_X,

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \cdot \mu_X(\mathrm{d}x) = \int_{\mathbb{R}} x f_X(x) \cdot \lambda(\mathrm{d}x) \tag{25}$$

Expectation is linear. $\mathbb{E}[\alpha X_1 + X_2] = \mathbb{E}[\alpha X_1] + \mathbb{E}[X_2]$.

3.3 Convexity

Recall the (informal) definition of convex functions in $\mathbb{R} \to \mathbb{R}$:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2), \forall x_1, x_2 \in dom(f), dom(f) \text{ convex.} \tag{26}$$

Proposition 3.4 (Subgradient property)**.** *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function. Then for all $x, y \in \mathbb{R}^d$, there exists a subgradient vector $g_x \in \mathbb{R}^d$.*

$$f(y) \geq f(x) + \langle y - x, g_x \rangle. \tag{27}$$

Theorem 3.5 (Jensen's inequality)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ convex. Assume f is bounded below. Let X be an integrable random variable, then*

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[X]). \tag{28}$$

Proof Sketch. Use the subgradient property with $x = \mathbb{E}[X]$ and the inequality should pop out directly. □

4 Moments and Tails

How do we collect information about the distribution of a random variable?

4.1 Definitions

Definition 4.1 (Moment)**.** *Let X be a real random variable with law μ_X . A moment is an integral of a real test function (μ_X-integrable) $h : \mathbb{R} \to \mathbb{R}$ against the law:*

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)\mu_X(\mathrm{d}x) \tag{29}$$

Examples:

- *Indicator $h = \mathbb{1}_B$ for Borel B, which gives $\mu_X(B)$.*
- *1^{st} order moment, $h(x) = x$. Then $m_1 = \mathbb{E}[X] = \int_{\mathbb{R}} x\mu_X(\mathrm{d}x)$.*
- *n-th polynomial moment: $h(x) = x^n$ for $n \in \mathbb{N}$. $m_n = \mathbb{E}[X^n]$.*
- *Exponential moment: let $h(x) = \exp^{\theta x}$. $\mathbb{E}[e^{\theta X}]$.*

Definition 4.2 (Tails)**.** *Let X be a real random variable, $t \in \mathbb{R}$. The (right) tail is defined as $\mathbb{P}(X \geq t)$. We often refer to the tail probability $\mathbb{P}(|X| \geq t)$.*

4.2 Moments and Tail Bounds

Several theorems on tail bounds bridge a connection between moments and tails.

Theorem 4.3 (Markov's inequality)**.** *Let X be a real random variable and $X \geq 0$.*

$$\mathbb{P}[X \geq t] \leq \frac{1}{t}\mathbb{E}[X], \forall t > 0. \tag{30}$$

Remark. For positive, increasing $\varphi : \mathbb{R}_+ = \mathbb{R}_+$, we have

$$\mathbb{P}\{X \geq t\} \leq \frac{1}{\varphi(t)}\mathbb{E}[\varphi(X)], \ \forall t > 0. \tag{31}$$

We can use $\varphi(x) = x^p$ to introduce higher-order information of X to the tail bound. $\varphi(x) = \exp^{cx}$ sometimes yields a tighter tail bound of exponential decay, i.e., Chernoff bound.

In other words, polynomial moments control tail probabilities.

Theorem 4.4 (Integral by parts)**.** *Let X be a positive real random variable and $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be an increasing and continuously differentiable function. Then,*

$$\mathbb{E}[\varphi(X)] = \varphi(0) + \int_0^\infty \mathbb{P}\{X \geq t\}\varphi'(t)\mathrm{d}t. \tag{32}$$

Remark.

- When $\varphi(x) = x^q$ for some $q > 0$, then

$$\mathbb{E}[|X|^q] = \int_0^\infty \mathbb{P}\{|X| \geq t\} \cdot q \cdot t^{q-1} \mathrm{d}t. \tag{33}$$

This means tail probability controls moments.

Assume X is a real random variable with $\mathbb{P}\{|X| \geq t\} = O(t^{-p})$ for some $p \geq 1$. Then we can bound the q-th moment for $q < p$.

$$\begin{aligned}
\mathbb{E}|X|^q &= \int_0^\infty \mathbb{P}\{|X| \geq t\} q t^{q-1} \mathrm{d}t && (34)\\
&= \int_0^1 \mathbb{P}\{|X| \geq t\} q t^{q-1} \mathrm{d}t + \int_1^\infty \mathbb{P}\{|X| \geq t\} q t^{q-1} \mathrm{d}t && (35)\\
&\leq \int_0^1 q t^{q-1} \mathrm{d}t + \int_1^\infty C q t^{q-p-1} \mathrm{d}t && (36)\\
&= 1 + \frac{Cq}{p-q} < +\infty && (37)
\end{aligned}$$

5 L_p Spaces

Definition 5.1 (L_p space)**.** *For $p > 0$, the space $L_p := L_p\{\Omega, \mathcal{F}, \mathbb{P}\}$ is*

$$L_p := \{X : \Omega \to \mathbb{R} : \mathbb{E}|X|^p < \infty\}. \tag{38}$$

Roughly, random variables with tail decay rate at least t_{-p}. These random variables are called "p-integrable random variables".

Remarks. L_p is a linear space.

Definition 5.2 (Homogeneous p-th moment.)**.** *For $p > 0$, we define the homogeneous p-th moment of X by*

$$\|X\|_p := (\mathbb{E}|X|^p)^{1/p} \tag{39}$$

Sometimes denoted as $\|X\|_{L_p}$.

Theorem 5.3 (Lyapunov inequality)**.** *For $0 < p \leq q$, real random variables X have*

$$\|X\|_p \leq \|X\|_q. \tag{40}$$

Therefore,

$$\|X\|_q \subseteq \|X\|_p. \tag{41}$$

Proof Sketch. Use Jensor's inequality for $\varphi(t) = |t|^{q/p}$. □

Warning. For sequence spaces, $\ell_p \subseteq \ell_q$. For Lebesgue spaces $\mathcal{L}_p(\mathbb{R}) \not\subset \mathcal{L}_q(\mathbb{R})$.

5.1 Pseudonorm

Question: Is $\|\cdot\|_{L_p}$ a norm? We would need to prove the triangular inequality.

Theorem 5.4 (Hölder's inequality)**.** *Let X, Y be real random variables with $X \in L_p, Y \in L_q$ with $\frac{1}{p} + \frac{1}{q} = 1, p, q > 1$.*

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q. \tag{42}$$

Proof. We use a lemma called Young's inequality. If $\frac{1}{p} + \frac{1}{q} = 1$, $p > 1$, then $|xy| \leq \frac{1}{p}|x|^p + \frac{1}{q}|y|^q$. We consider $X/\|X\|_p$ and $Y/\|Y\|_q$, and take expectation. □

Theorem 5.5 (Minkowski)**.** *Assume $p \geq 1$. Let $X, Y \in L_p$. Then*

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p. \tag{43}$$

Proof.

$$|X+Y|^p = |X+Y| \cdot |X+Y|^{p-1} \leq |X| \cdot |X+Y|^{p-1} + |Y| \cdot |X+Y|^{p-1}. \tag{44}$$

Apply Holden with $\frac{1}{p} + \frac{p-1}{p} = 1$.

$$\mathbb{E}|X+Y|^p \leq (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|X+Y|^p)^{\frac{p-1}{p}} + (\mathbb{E}|Y|^p)^{1/p}(\mathbb{E}(|X+Y|^p)^{p-1/p}) \tag{45}$$

After dividing $\mathbb{E}|X+Y|^{p-1}$, the triangular inequality pops out. □

Proposition 5.6 (L_p pseudonorm)**.** *For random variables in L_p, $\|\cdot\|_p$ is a pseudonorm*

- *Positive semi-definite:* $\|X\|_p \geq 0$ *and* $\|0\|_p = 0$.
- *Positive homogeneous:* $\|\alpha X\| = |\alpha| \cdot \|X\|_p$.
- *Triangle inequality:* $\|X+Y\|_p \leq \|X\|_p + \|Y\|_p$.
- *Almost positive:* $\|X\|_p = 0$ *implies* $X = 0$ *μ-almost sure.*

We can thus define a pseudometric on L_p:

$$D(X,Y) := \|X-Y\|_p, \forall X, Y \in L_p. \tag{46}$$

5.2 Convergence in L_p

Definition 5.7. *A sequence $(X_j : j \in \mathbb{N})$ in L_p converges in L_p if there is a real random variable $Y \in L_p$, such that*

$$\|X_j - Y\|_p \to 0, \text{ as } j \to \infty. \tag{47}$$

Notes. $X_j \to Y$ in L_p implies $X_j \to Y$ in L_q for $q \leq p$.

Question: When does a sequence in L_p converge?

Definition 5.8 (Cauchy sequence)**.** *A sequence $(X_j : j \in \mathbb{N})$ in L_p is called Cauchy if*

$$\sup_{i,j \geq N} \|X_i - X_j\|_p \to 0 \text{ as } N \to \infty. \tag{48}$$

It is trivial that any converging sequence is Cauchy, but the other direction is not trivial.

Theorem 5.9 (L_p space is complete.)**.** *Every Cauchy sequence in L_p converges to a random variable. Moreover, it converges to a random variable in L_p.*

Notes. An example of non-complete space: $\mathbb{Q}$. A sequence of rational numbers may converge to an irrational number.

Remark. Limits in L_p are not necessarily unique! They are only required to be equal μ-almost surely.

5.3 L_2 Space

Next, we restrict our discussion to L_2 space, which is special for setting orthogonality, variance, covariance, and orthogonal projection.

$$L_2 := L_2(\Omega, \mathcal{F}, \mathbb{P}) = \{X : \Omega \to \mathbb{R} : \mathbb{E}|X|^2 < \infty\}. \tag{49}$$

Theorem 5.10 (Cauchy-Schwarz). *If $X, Y \in L_2$, then $XY \in L_1$, and*

$$|\mathbb{E}[XY]| \leq \|XY\|_1 \leq \|X\|_2 \cdot \|Y_2\|. \tag{50}$$

This is a special case of Hölder's theorem when $p = q = 2$. Nevertheless, here is another simpler proof

Proof. Let $t \in \mathbb{R}$.

$$0 \leq \mathbb{E}(t|X| + |Y|)^2 = \mathbb{E}|X|^2 t^2 + 2t\mathbb{E}|XY| + \mathbb{E}|Y|^2, \tag{51}$$

which is a quadratic function of t. Hence $4(E|XY|)^2 - 4\mathbb{E}|X|^2\mathbb{E}|Y|^2 \leq 0$. □

Definition 5.11 (L_2 (pseudo)-inner product). *For random variables $X, Y \in L_2$, we define*

$$\langle X, Y \rangle_{L_2} := \mathbb{E}[XY]. \tag{52}$$

Remarks.

- This is well-defined by Cauchy-Schwarz.
- $\langle X, X \rangle = \mathbb{E}[|X|^2] = \|X\|_{L_2}^2$.

Definition 5.12 (Orthogonality). *If $\langle X, Y \rangle = 0$, then we say X and Y are orthogonal random variables, and we write $X \perp Y$.*

Warning: $X \perp Y$ does NOT imply that X, Y are "independent".

Definition 5.13 (Covariance). *Let $X, Y \in L_2$, the covariance of X, Y is*

$$Cov(X, Y) := \langle X - \mathbb{E}X, Y - \mathbb{E}Y \rangle. \tag{53}$$

If $Cov(X, Y) = 0$, we say X, Y are uncorrelated, not necessarily independent!

Moreover, we define the variance of a random variable $X \in L_2$ by $Var(X) := Cov(X, X)$.

The correlation of X, Y is defined by

$$\rho(X, Y) := \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}} \in [-1, 1] \tag{54}$$

Proposition 5.14 (Pythagorean). *If $X, Y \in L_2$, then $Cov(X, Y) = 0$ implies*

$$Var(X, Y) = Var(X) + Var(Y). \tag{55}$$

Theorem 5.15 (Orthogonal projection)**.** *Let $K \subseteq L_2$ be a complete linear subspace of $L_2(\Omega, \mathcal{F}, \mathbb{P})$. For $X \in L_2$, there is a random variable $Y \in K$ such that*

- *Primal:* $\|X - Y\|_2 = \inf\{\|X - w\|_2 : w \in K\}$.
- *Dual:* $(X - Y) \perp Z$ *for all* $Z \in K$.

These two statements are essentially the same. We call Y a version of the orthogonal projection of X onto K.

Remark. Such Y are not unique. We can have $Y = Y'$ almost surely.

Proof. Consider a minimizing sequence in K: $(Y_i \in K : I \in \mathbb{N})$, such that $\|Y_i - X\|_2$ is decreasing and converges to $d := \inf\{\|w - X\|^2 : w \in K\}$.

Step 1: Show that (Y_i) has a limit in K, which is a candidate for the projection. Check that the sequence (Y_i) is Cauchy.

$$0 \leq \|\frac{1}{2}(Y_i - Y_j)\|^2 = \frac{1}{2}\|Y_i - X\|^2 + \frac{1}{2}\|Y_j - X\|^2 - \|\frac{1}{2}(Y_i + Y_j) - X\|^2 \leq d^2 + (-d^2) = 0. \tag{56}$$

So Y_i is Cauchy in K. Since K is complete, we know $Y_i \to Y$ for some $Y \in K$. Y is a candidate for the orthogonal projection of X onto K.

Step 2: Check that Y is indeed an orthogonal projection. By Minkowski, for each i

$$\|X - Y\|_2 \leq \|X - Y_i\|_2 + \|Y_i - Y\|_2, \tag{57}$$

which converges to d. Moreover, $\|X - Y\|_2 \geq d$, so $\|X - Y\|_2 = d$.

Step 3: Dual characterization. For $Z \in K$, let $t \in \mathbb{R}$, $Y + tZ \in K$ since K is a linear subspace. Then

$$\|X - (Y + tZ)\|^2 \geq \|X - Y\|^2 \tag{58}$$

Expand the equation and it's trivial to see we must have $\langle X - Y, Z\rangle = 0$. □

6 Independence

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. How to update our knowledge about an event A, when another event E has occurred?

$$\mathbb{P}(A|E) := \frac{\mathbb{P}(A \cap E)}{\mathbb{P}(E)} \tag{59}$$

We define $\mathcal{F}|_E$ the σ-algebra $\{A \cap E \mid A \in \mathcal{F}\}$.

Then $(E, \mathcal{F}|_E, \mathbb{P}(\cdot|E))$ is a probability measure space.

Definition 6.1 (Independence of events)**.** *Suppose knowledge that E occurs does not change the probability of A:*

$$\mathbb{P}(A|E) = \mathbb{P}(A) \tag{60}$$

Then we call events A and E are independent. It is equivalent to saying

$$\mathbb{P}(A \cap E) = \mathbb{P}(A) \cdot \mathbb{P}(E). \tag{61}$$

6.1 Independent for random variables

How about random variables being independent? Random variables can take any value in their domain, so we must consider a series of events to be independent.

Definition 6.2 (Independence of random variables)**.** *Let X, Y be real random variables. We say that (X, Y) are independent when*

$$\mathbb{P}\{X \in A \cap Y \in B\} = \mathbb{P}\{X \in A\} \cdot \mathbb{P}\{X \in B\}, \tag{62}$$

for all $A, B \in \mathcal{B}(\mathbb{R})$.

In particular, for all $a, b \in \mathbb{R}$, it is common to see the definition below,

$$\mathbb{P}\{X \leq a \wedge Y \leq b\} = \mathbb{P}\{X \leq a\} \cdot \mathbb{P}\{Y \leq b\}. \tag{63}$$

Fact: These two definitions are equivalent.

This is also equivalent to

$$\mu_{XY}(A \times B) = \mu_X(A) \cdot \mu_X(B), \ \forall \ A, B \in \mathcal{B}(\mathbb{R}). \tag{64}$$

Moreover, it is equivalent to $\mu_{XY} = \mu_X \times \mu_Y$ on $\mathcal{B}(\mathbb{R}^2)$.

Proposition 6.3. *Let X, Y be independent real random variables. Let $f \in L_1(\mu_X)$ and $g \in L_1(\mu_Y)$. Then*

$$\mathbb{E}[f(X) \cdot g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)]. \tag{65}$$

Proof.

$$\mathbb{E}[f(X)\cdot g(Y)] = \int_{\mathbb{R}^2} f(x)g(y)\mu_{XY}(\mathrm{d}x\mathrm{d}y) \tag{66}$$

$$= \int_{\mathbb{R}^2} f(x)g(y)\mu_X(\mathrm{d}x)\mu_Y(\mathrm{d}y) \tag{67}$$

$$= \int_{\mathbb{R}} f(x)\mu_X(\mathrm{d}x)\int_{\mathbb{R}} g(y)\mu_Y(\mathrm{d}y) = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]. \tag{68}$$

□

Question: Can we generate a collection of random variables with specified marginals? For a countable sequence, the answer is yes!

Theorem 6.4 (Kolmogorov extension)**.** *Let $(\mu_i : i \in \mathbb{N})$ be a collection of Borel probability measures on $\mathbb{R}$. Define*

$$\Omega = \mathbb{R}^{\mathbb{N}} = \{\omega = (\omega_1, \omega_2, \dots,) : \omega_i \in \mathbb{R}, i \in \mathbb{N}\}. \tag{69}$$

Consider coordinate random variables $X_i(\omega) = \omega_i$.

We can define product σ-algebra:

$$\mathcal{F} = \sigma(X_i : i \in \mathbb{N}) = \text{ smallest sigma algebra that have coordinate rvs measurable.} \tag{70}$$

There exists a product probability measure $\mathbb{P}$ in $(\Omega, \mathcal{F})$, such that

$$\mathbb{P}\{(x_1, x_2, \dots) \in B_1 \times B_2 \times \dots\} = \prod_{i\in\mathbb{N}} \mathbb{P}(X_i \in B_i) = \prod_{i\in\mathbb{N}} \mu_i(B_i). \tag{71}$$

6.2 Independence for sigma-algebras

Note that σ-algebras carry information. We can also define the independence of σ-algebras.

Definition 6.5 (Independence of σ-algebra)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{F}_1, \mathcal{F}_2$ be sub-σ-algebra of $\mathcal{F}$.*

We say that $\mathcal{F}_1, \mathcal{F}_2$ are independent if for all $E_1 \in \mathcal{F}_1$ and $E_2 \in \mathcal{F}_2$,

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\cdot\mathbb{P}(E_2). \tag{72}$$

For events $A, B.\sigma(\{A\}) = \{\varnothing, A, A^{\mathrm{C}}, \Omega\}$. We can check that A and B are independent if and only if $\sigma(\{A\})$ and $\sigma(\{B\})$ are independent σ-algebras.

For real random variables: $\sigma(X) := \sigma(\{X^{-1}(B), B \in \mathcal{B}(\mathbb{R})\})$. Then real random variables X, Y are independent if and only if $\sigma(X)$,$\sigma(Y)$ are independent σ-algebras.

7 Law of Large Numbers

Definition 7.1 (Stochastic process)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A stochastic or random process is a family $(X_t : t \in T)$ of real random variables, indexed by an abstract set T.*

- Discrete time: $T = \mathbb{N}$.
- Continuous time: $T = \mathbb{R}_{\geq 0}$.
- Space: $T = \mathbb{R}^n$, "random fields".

Key models covered in this note: Independent sum; discrete-time martingales.

Definition 7.2 (Independent sum)**.** *Let $(Y_i : i \in \mathbb{N})$ be an independent sequence of real random variables. The partial sums:*

$$X_n := \sum_{i=1}^{m} Y_i \tag{73}$$

compose a discrete-time $(T = \mathbb{Z}_+)$ random process, called an independent sum process.

Examples.

- Repeated independent experiments.
- Random walk.
- Renewal processes.

Definition 7.3 (Running average process)**.** *Let $(Y_i : i \in \mathbb{N})$ be independent real random variables. $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} Y_u$ is called the running average process.*

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}[Y] \tag{74}$$

$$Var(\bar{X}_n) = \frac{1}{n} Var[Y]. \tag{75}$$

7.1 Weak Law of Large Numbers

Definition 7.4 (Convergence in probability)**.** *A sequence $(W_n : n \in \mathbb{N})$ of real random variables converges in probability to a random variable W, if*

$$\sup_{t>0} \lim_{n\to\infty} \mathbb{P}\{|W_n - W| \geq t\} = 0, \forall t > 0. \tag{76}$$

Theorem 7.5 (Chebyshev's weak LLN)**.** *Let $Y \in L_2$ be a real random variable and $(Y_i : i \in \mathbb{N})$ are iid copies of Y. Let $\bar{X}_n = \sum_{i=1}^{n} Y_i$. Then*

$$\bar{X}_n \to \mathbb{E}Y \text{ in probability } \mathbb{P} \tag{77}$$

7.2 Strong Law of Large Numbers

Definition 7.6 (A.s. Convergence)**.** *Consider a sequence* $(W_n : n \in \mathbb{N})$ *of real random variables. We say that* $W_n \to W$ *almost surely, if when* $n \to \infty$*,*

$$\mathbb{P}\{w \in \Omega : W_n(\omega) \to w(\omega)\} = 1. \tag{78}$$

Equivalently,

$$\mathbb{P}\{\limsup_{n\to\infty} |W_n \to W| > 0\} = 0. \tag{79}$$

Remark.

$$W_n \to W \text{ point-wise } \implies W_n \to W \text{ a.s. } \implies W_n \to W \text{ in } \mathbb{P}. \tag{80}$$

Theorem 7.7 (Kolmogorov Strong LLN)**.** *Let* $Y \in L_1$ *be a real random variable. Let* $(Y_i : i \in \mathbb{N})$ *be iid copies of* Y*. Let* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n Y_i$*. Then* $\bar{X}_n \to \mathbb{E}Y$ *a.s. That is,*

$$\mathbb{P}\{\bar{X}_n \to \mathbb{E}Y\} = 1. \tag{81}$$

Theorem 7.8 (Cantelli Strong LLN)**.** *Assume* $Y \in L_4$*. Then the previous theorem has a simpler proof* :)

Proof. WLOG, assume $\mathbb{E}Y = 0$.

Lemma 1 (Borel Cantelli.)**.** *Let* $(A_n : n \in \mathbb{N})$ *be events.*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \textit{ implies } \mathbb{P}(\limsup_{n\to\infty} A_n) = 0 \tag{82}$$

Proof of the lemma.

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{i\geq n} A_i = \text{event that } A_i \text{ happens infinite times} \tag{83}$$

$$\mathbb{P}(\limsup_{n\to\infty} A_n) \leq \mathbb{P}(\bigcup_{i\geq n} A_i), \forall n \tag{84}$$

$$\leq \sum_{i\geq n} \mathbb{P}(A_i) = 0, \text{as } n \to \infty. \tag{85}$$

□

Lemma 2 (Cartelli tail bound.)**.** *Assume* $\mathbb{E}Y = 0$*,* $\mu = \mathbb{E}|Y|^4$*. Then* $\forall n \in \mathbb{N}$*,*

$$\mathbb{P}\{|\bar{X}_n| \geq n^{-1/8}\} \leq 3\mu n^{-3/2}. \tag{86}$$

Cont. Define event $A_n = \{\omega \in \Omega : |\bar{X}_n(\omega)| \geq n^{-1/8}\}$. Then

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) \leq \sum_{n=1}^{\infty} 3\mu n^{-3/2} < \infty. \tag{87}$$

Then using the Borel-Cantelli lemma.

$$\mathbb{P}\{\limsup_{n\to\infty} |X_n| > 0\} = 0. \tag{88}$$

□

7.3 Concentration

The law of large numbers gives an asymptotic analysis for the limit of $X_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. However, we also want nonasymptotic analysis for a fixed n.

Proposition 7.9 (Chebyshev)**.** *Let $X \in L_2$, and $\sigma = \sqrt{Var(X)}$,*

$$\mathbb{P}\{|X - \mathbb{E}X| \geq \sigma t\} \leq \frac{1}{t^2}, \forall t > 0. \tag{89}$$

- Chebyshev's inequality gives weak controls on tail bound (t^{-2}), but this is the best we could get by assuming $X \in L_2$ only.

Definition 7.10 (Moment generating function (mgf) and cumulant generating function (cgf))**.** *Let X be a real random variable. The moment generating function is defined as*

$$m_X(\theta) := \mathbb{E}[e^{\theta X}], \tag{90}$$

and the cumulant generating function

$$\xi_X(\theta) := \log \mathbb{E}[e^{\theta X}]. \tag{91}$$

Proposition 7.11 (Laplace transform method)**.** *Let X be a real random variable. For all $t \in \mathbb{R}$,*

$$\mathbb{P}\{X \geq t\} \leq \inf_{\theta>0} \exp(-\theta t + \xi_X(\theta)) \tag{92}$$

$$\mathbb{P}\{X \leq t\} \leq \inf_{\theta<0} \exp(-\theta t + \xi_X(\theta)). \tag{93}$$

Proof Sketch. Simply apply Markov's inequality to $\exp(\theta X)$. □

Example. (Normal distribution) Recall that for $Z \sim \mathcal{N}(0, \sigma^2)$, we have $\xi_Z(\theta) = \frac{\sigma^2\theta^2}{2}$. Then

$$\mathbb{E}\{Z \geq t\} \leq \inf_{\theta>0} \exp\big(-\theta t + \sigma^2\theta^2/2\big) = \exp\big(-t^2/2\sigma^2\big). \tag{94}$$

Proposition 7.12 (Additivity of cgf)**.** *Let $X = \sum_i Y_i$ for independent random variables Y_i, then*

$$\xi_X(\theta) = \sum_i \xi_{Y_i}(\theta). \tag{95}$$

Thanks to the additivity of cgf, we can easily derive a tail bound for independent sums,

Theorem 7.13. *Let $X = \sum_{i=1}^n Y_i$ for independent Y_i.*

$$\mathbb{P}\{X \geq t\} \leq \inf_{\theta>0} \exp\left(-\theta t + \sum_{i=1}^n \xi_{Y_i}(\theta)\right) \tag{96}$$

$$\mathbb{P}\{X \leq t\} \leq \inf_{\theta<0} \exp\left(-\theta t + \sum_{i=1}^n \xi_{Y_i}(\theta)\right) \tag{97}$$

Example. (Binomial) Let $X \sim Binomial(n, p)$, so $X = \sum_{i=1}^n Y_i$ where $Y_i \sim Bern(p)$ iid.

$$\xi_{Y_i}(\theta) \leq p(e^\theta - 1). \tag{98}$$

So we get (the Chernoff bound)

$$\mathbb{P}\{X \geq t\mathbb{E}[X]\} \leq \inf_{\theta>0} \exp\Big(-np(\theta t - e^\theta + 1)\Big) = \exp(-np(t\log t - t + 1)) = \Big(\frac{e^{t-1}}{t^t}\Big)^{np}. \tag{99}$$

Theorem 7.14 (Hoeffding's theorem)**.** *For $X = \sum_{i=1}^n Y_i$ for independent bounded random variables $Y_i \in [a_i, b_i]$. Let $\sigma = \sqrt{\sum_{i=1}^n (b_i - a_i)^2}/2$. Then*

$$\mathbb{P}\{|X - \mathbb{E}X| \geq \sigma t\} \leq 2\exp\big(-t^2/2\big) \tag{100}$$

Proof. We need a cgf bound for Y_i. We find that $\xi_{Y_i}(\theta) \leq (b_i - a_i)^2\theta^2/8$. This is because

$$m_{Y_i}(\theta) = \mathbb{E}[e^{\theta Y}] \tag{101}$$

$$\leq \mathbb{E}[f(x)] = \cosh\left(\theta\frac{b-a}{2}\right) \tag{102}$$

where $f(x) = \frac{(e^{\theta b} - e^{\theta a})(x-a)}{2(b-a)} + e^{-\theta a}$ so $f(x) < e^{\theta x}$ on $[a, b]$. Plug in $\xi_{Y_i} = \log m_{Y_i}$ proves the theorem. □

7.4 Weak convergence

Limit theorems in probability tell us when a sequence of random variables converges to a limiting random variable. In this section, we consider describing when the distributions of the random variables converge to a limiting distribution.

What does it mean for the distributions of random variables to converge?

Idea: Moments carry information about distributions.

$$\mu(h) = \int_{\mathbb{R}} h(x)\mu(\mathrm{d}x) \tag{103}$$

If two measures μ, ν are close, then many moments are similar: $\mu(h) \approx \nu(h)$ for "many" h.

Definition 7.15. *A function $h : \mathbb{R} \to \mathbb{R}$ is bounded Lipschitz, if*

$$\|h\|_{BL} = \max\{\|h\|_{sup}, \|h\|_{Lip}\} < \infty, \tag{104}$$

where

$$\|h\|_{sup} = sup\{|h(x)| : x \in \mathbb{R}\} \tag{105}$$

and

$$\|h\|_{Lip} = \inf_{L>0:|h(x)-h(y)|\leq L|x-y|,\forall x,y\in\mathbb{R}} \tag{106}$$

Remark. $\|\cdot\|_{BL}$ is a norm.

Proposition 7.16 (BL functions separate measures.)**.** *Let μ, ν be Borel probability measures on $\mathbb{R}$. Then $\mu = \nu$ if and only if $\mu(h) = \nu(h)$ for all bounded Lipschitz $h : \mathbb{R} \to \mathbb{R}$.*

Proof. The forward direction is trivial.

The reverse direction. Assume $\mu \neq \nu$, then the cdf $F_\mu \neq F_\nu$, so there exists $a \in \mathbb{R}$,

$$F_\mu(a) \neq F_\nu(a). \tag{107}$$

Let us consider a series of BL functions:

$$h_n(x) = \begin{cases} 1, \ x \leq a \\ 0, \ x \geq a + \frac{1}{n} \\ 1 - n(x-a), \ x \in (a, a + \frac{1}{n}. \end{cases} \tag{108}$$

Using bounded convergence theorem (or dominated convergence), $\mu(h_n) \to F_\mu(a)$ and $\nu(h_n) \to F_\nu(a)$. Therefore, there exists n, such that $\mu(h_n) \neq \nu(h_n)$. □

Definition 7.17 (Bounded Lipschitz metric)**.** *Let μ, ν be Borel probability measures on $\mathbb{R}$. Define*

$$d_{BL}(\mu, \nu) := \sup\{|\mu(h) - \nu(h)| : \|h\|_{BL} \leq 1\}. \tag{109}$$

Note: d_{BL} is a metric on the probability measures on $\mathbb{R}$, i.e.,

- $d_{BL}(\mu, \nu) = 0$ if and only if $\mu = \nu$.
- $d_{BL}(\mu, \nu) = d_{BL}(\nu, \mu)$.
- $d_{BL}(\mu, \nu) \leq d_{BL}(\mu, \rho) + d_{BL}(\rho, \nu)$.

Definition 7.18 (Weak convergence)**.** *Let $(\mu_n : n \in \mathbb{N})$ be Borel probability measures on $\mathbb{R}$, μ be a Borel probability. We say μ converges weakly to μ, if*

$$d_{BL}(\mu_n, \mu) \to 0 \text{ as } n \to \infty. \tag{110}$$

Let $(X_n : n \to \mathbb{N})$ be real random variables and X be a real random variable. Then we say X_n weakly converges to X, if μ_{X_n} weakly converges to μ_X.

Theorem 7.19. *Let $(\mu_n : n \in \mathbb{N})$, μ be Borel probability measures on $\mathbb{R}$. The following are equivalent:*

- *μ_n weakly converges to μ*
- *$\mu_n(h) \to \mu(h)$ for all bounded Lipschitz $h : \mathbb{R} \to \mathbb{R}$.*
- *$\mu_n(h) \to \mu(h)$ for all bounded continuous $h : \mathbb{R} \to \mathbb{R}$.*
- *The distribution function $F_{\mu_n}(a)$ converges to $F_\mu(a)$ for all $a \in \mathbb{R}$ where $F_\mu(a)$ us continuous.*
- *The characteristic function $\chi_{\mu_n} \to \chi_\mu$ point-wise, where*

$$\chi_W(\theta) = \mathbb{E}[e^{i\theta W}] \tag{111}$$

Proposition 7.20. *$X_n \to \mathbb{E}Y$ almost sure implies that X_n converges weakly to $\mathbb{E}Y$.*

Definition 7.21 (Integral probability metric)**.** *Let H be a collection of functions from $\mathbb{R} \to \mathbb{R}$. Define the IPM related to H by*

$$d_H(\mu, \nu) = \sup\{|\mu(h) - \nu(h)| : h \in H\}, \tag{112}$$

for all Borel probability measure ν on $\mathbb{R}$.

Remark.

- Different H gives different notions of distance
- If $H \subseteq H'$, then $d_H \leq d_{H'}$
- d_H is a metric if and only if H separates measures.

Examples

- Kolmogorov: $H = \{\mathbb{1}_{(-\infty,a]} : a \in \mathbb{R}\}$, which induces uniform convergence of cdfs.
- Total variation: $H = \{h : \mathbb{R} \to \mathbb{R} : \|h\|_{\sup} \leq 1, h \text{ continuous}\}$ or $H' = \{\mathbb{1}_B : B \text{ Borel}\}$.
- Kantorovich Wasserstein-1 distance

$$H = \{h : \mathbb{R} \to \mathbb{R} : \|h\|_{Lip} \leq 1\} \tag{113}$$

  Convergence weakly and L_1.

7.5 Central Limit Theorem

Let $(Y_i : i \in \mathbb{N})$ be iid copies of $Y \in L_2$, then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ converges almost surely to $\mathbb{E}Y$.

Idea: rescale $\bar{X}_n$ to make its variance nontrivial. We define *standardized sums*

$$T_n := \sqrt{n}\frac{\bar{X}_n - \mathbb{E}Y}{\sqrt{Var(Y)}}. \tag{114}$$

It can be checked that $\mathbb{E}T_n = 0$ and $Var(T_n) = 1$.

Theorem 7.22 (Central Limit Theorem). *Let $(Y_i : i \in \mathbb{N})$ be iid copies of $Y \in L_2$. Define*

$$T_n = \sqrt{n}\Big(\frac{\bar{X}_n - \mathbb{E}Y}{\sqrt{Var(Y)}}\Big) \ \forall n \in \mathbb{N}. \tag{115}$$

Then T_n weakly converges to $\mathcal{N}(0,1)$ as $n \to \infty$.

Remark.

- Historically, the central limit theorem is used to support asymptotic confidence intervals for $\mathbb{E}Y$.
- Equivalent statements:

$$d_{BL}(T_n, Z) = \sup_{\|h\|_{BL} \le 1} |\mathbb{E}h(T_n) - \mathbb{E}h(Z)| \to \infty. \tag{116}$$

  Also equivalent to

$$\mathbb{P}\{T_n \le a\} = F_{T_n}(a) \to \Phi(a), \ \forall a \in \mathbb{R}, \tag{117}$$

  which is often called the convergence in distribution.

7.6 Nonasymptotic counterpart

Theorem 7.23 (Berry-Esseen). *Let $(Y_i : i \in \mathbb{N})$ be iid copies of $Y \in L_3$. Define*

$$\sigma^2 = Var(Y), \text{ and } M_3 = \mathbb{E}|Y - \mathbb{E}Y|^3. \tag{118}$$

Let Y_n be the standardized sum, then the Kolmogorov distance

$$d_{Kol}(T_n, Z) := \sup_{a \in \mathbb{R}} |F_{T_n}(a) - \Phi(a)| \le \frac{M_3}{\sqrt{n}\sigma^3}, \ \forall n \in \mathbb{N}. \tag{119}$$

***Proof** of a weaker version by the Lindeberg exchange principle.* The idea is to show for smooth functions $h : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}h(T_n) = \mathbb{E}h(Z), \text{ by Taylor expansion.} \tag{120}$$

Lemma 3. *Let (Y, Z) be L_3 random variables with $\mathbb{E}Y = \mathbb{E}Z$ and $\mathbb{E}Y^2 = \mathbb{E}Z^2$. Let $h : \mathbb{R} \to \mathbb{R}$ with $\|h^{(3)}\|_{\sup} < \infty$.*

$$\|\mathbb{E}h(Y) - \mathbb{E}h(Z)\| \le \frac{1}{6}\|h^{(3)}\|_{\sup}(\mathbb{E}|Y|^3 + \mathbb{E}|Z|^3). \tag{121}$$

Proof of Lemma.

$$h(t) - h(0) - th'(0) - \frac{t^2}{2}h''(0) = \frac{t^3}{6}h^{(3)}(\xi), \text{ for some } \xi \in [0, t]. \tag{122}$$

Then

$$\left|\mathbb{E}[h(Y) - h(0) - Yh'(0) - \frac{1}{2}Y^2h''(0)]\right| \leq \frac{1}{6}\mathbb{E}|Y^3|\|h^{(3)}\|_{\sup}. \tag{123}$$

The left hand side is1 the same for Y and Z, except for the first term, so

$$|\mathbb{E}h(Y) - \mathbb{E}h(Z)| \leq \frac{1}{6}\|h^{(3)}\|_{\sup} \cdot (\mathbb{E}|Y|^3 + \mathbb{E}|Z|^3). \tag{124}$$

□

Let $Y_1, \dots Y_n$ and $Z_1, \dots, Z_n$ be independent with $\mathbb{E}Y_i = \mathbb{E}Z_i$ and $\mathbb{E}Y_i^2 = \mathbb{E}Z_i^2$. For $f : \mathbb{R}^n \to \mathbb{R}$, such that $\|\partial_{iii}f\|_{\sup} < \infty$, then

$$|\mathbb{E}f(Y_1, \dots, Y_n) - \mathbb{E}f(Z_1, \dots, Z_n)| \leq \frac{1}{6}\sum_{i=1}^{n}\|\partial_{iii}f\|_{\sup}(\mathbb{E}|Y_i|^3 + \mathbb{E}|Z_i|^3). \tag{125}$$

This is called the Lindeberg exchange theorem. To see this, let

$$W_i = (Y_1, \dots, Y_i, Z_{i+1}, \dots, Z_n). \tag{126}$$

Then since $W_0 = (Z_1, \dots, Z_n)$ and $W_n = (Y_1, \dots, Y_n)$,

$$|\mathbb{E}f(Y) - \mathbb{E}f(Z)| \leq \sum_{i=1}^{n}|\mathbb{E}f(W_i) - \mathbb{E}f(W_{i+1})| \leq \frac{1}{6}\sum_{i=1}^{n}\|\partial_{iii}f\|_{\sup}(\mathbb{E}|Y_i|^3 + \mathbb{E}|Z_i|^3). \tag{127}$$

Let $Y_i \sim Y$, $Z_i \sim Z$ in L_3 be independent random variables that are already standardized, i.e., $\mathbb{E}Y = \mathbb{Z} = 0$ and $Var(Y) = Var(Z) = 1$. Let $M = \max\{\mathbb{E}|Y|^3, \mathbb{E}|Z|^3\} < \infty$. Let $h : \mathbb{R} \to \mathbb{R}$ be smooth. Define

$$f(X_1, \dots, X_n) = h(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i), \text{ where } X \text{ means } Y, Z. \tag{128}$$

$\|\partial_{iii}f\|_{\sup} \leq n^{-3/2}\|h^{(3)}\|_{\sup}$. Then applying the Lindeberg exchange principle,

$$|\mathbb{E}h(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Y_i) - \mathbb{E}h(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i)| \leq \frac{M}{3\sqrt{n}}\|h^{(3)}\|_{\sup}. \tag{129}$$

Let

$$H = \{h : \mathbb{R} \to \mathbb{R}, \|h^{(3)}\|_{\sup} \leq 1\} \tag{130}$$

$$d_H(\frac{1}{\sqrt{n}}\sum_i Y_i, \frac{1}{n}\sum_i Z_i) \leq \frac{M}{3\sqrt{n}}. \tag{131}$$

□

Issues:

- Where is the normal random variable?
  Solution: Choose $Z_i \sim \mathcal{N}(0,1)$, then we can compute $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i \sim \mathcal{N}(0,1)$.

- What is d_H?

  Solution: For each $h : \mathbb{R} \to \mathbb{R}$ with $\|h\|_{BL} \leq 1$, we can smooth h to get a function h_σ such that $\|h_\sigma^{(3)}\|_{\sup}$ is bounded and $\|h - h_\sigma\|_{\sup}$ is bounded, which implies

$$d_{BL}(T_n, Y) \leq \frac{C\mathbb{E}(|Y|^3)^{1/3}}{n^{1/6}} \to \infty, \text{ as } n \to \infty. \tag{132}$$

8 Conditional Expectation

If two random variables are dependent, by observing one, we can update our knowledge about probabilities of events involving the other. In particular, we can update our "best guess" for the expectation.

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $X \in L_2(\Omega.\mathcal{F}, \mathbb{P})$.

Recall that $Var(X) = \mathbb{E}[(X - \mathbb{E}X)^2] = \inf_{a \in \mathbb{R}} \|X - a\|^2$.

8.1 Conditional expectation in least squares

Consider a subspace in L_2:

$$K_0 = \{Y \in L_2(\Omega, \mathcal{F}, \mathbb{P}) : Y(\omega) = a, \forall \omega \in \Omega, a \in \mathbb{R}\}. \tag{133}$$

K_0 is a complete subspace in L_2, so there is always a projection from L_2 to K_0, i.e.,

$$Y^* = \arg\min_{Y \in K_0} \mathbb{E}(X - Y)^2 \tag{134}$$

Now suppose we observe a random variable $Z \in L_2$ and want to update our best guess of $\mathbb{E}X$. We first try an affine approximation:

$$\min_Y \|X - Y\|_2, \tag{135}$$

where $Y \in K_1$.

$$K_1 = \{a + bZ : a, b \in \mathbb{R}\}. \tag{136}$$

We can compute the optimal a, b given $Var(X)$, $Var(Z)$ and $Cov(X, Z)$.

As $K_0 \subset K_1$, the error is no worse than $Y = \mathbb{E}X$.

However, this affine function of Z is suboptimal (consider $X = Z^2$). In general, we have to consider a nonlinear function of Z.

Consider $K_Z = \{h(Z) \in L_2,\ h : \mathbb{R} \to \mathbb{R} \text{ measurable}\}$.

$$\min \|X - Y\|_2, \text{ s.t. } Y \in K_Z. \tag{137}$$

Recall that

Lemma 4. *Y is $\sigma(Z)$-measurable, if and only if $Y = h(Z)$ for some $h : \mathbb{R} \to \mathbb{R}$ measurable.*

As a result, K_Z is a complete subspace of L_2, since

$$K_Z = L_2(\Omega, \sigma(Z), \mathbb{P}). \tag{138}$$

Therefore, there exists an orthogonal projection of X to K_Z. Moreover, it is unique almost sure.

We define this projection by $\mathbb{E}[X|Z]$.

8.2 Conditioning on a σ-algebra

Consider conditioning on a σ-algebra: $\mathcal{G} \subseteq \mathcal{F}$. For each event $G \in \mathcal{G}$, we can decide if $\omega \in G$ or not.

Definition 8.1 (Conditional expectation in L_2)**.** *Fix $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra on Ω. For $X \in L_2(\Omega, \mathcal{F}, \mathbb{P})$,*

$$\min_Y \|X - Y\|_2, \text{ s.t. } Y \in L_2(\Omega, \mathcal{G}, \mathbb{P}|_{\mathcal{G}}). \tag{139}$$

A solution Y to this least square problem is the conditional expectation of X, given $\mathcal{G}$, denoted as $Y = \mathbb{E}[X \mid \mathcal{G}]$.

Dual Characterization. $(X - Y) \perp Z$ for all $Z \in L_2(\Omega, \mathcal{G}, \mathbb{P}_{\mathcal{G}})$.

Remarks.

- $Y = \mathbb{E}[X \mid \mathcal{G}]$ is a random variable on the sample space. Value is determined, once given $\mathbb{1}_G$ for all $G \in \mathcal{G}$.
- For random variables Z, $\mathbb{E}[X \mid Z] := \mathbb{E}[X \mid \sigma(Z)]$.
- For an event E, $\mathbb{E}[X \mid E] := \mathbb{E}[X \mid \sigma(E)]$. The conditional expectation must be $Y = a\mathbb{1}_E + b\mathbb{1}_{E^C}$. The optimal value of a, b is $a = \frac{\mathbb{E}[X\mathbb{1}_E]}{\mathbb{P}(E)}$ and $b = \frac{\mathbb{E}[X\mathbb{1}_{E^C}]}{\mathbb{P}[E^C]}$.

8.3 Characteristic properties of conditional expectation

Let $G_1, \ldots, G_n$ be a disjoint cover of Ω, $\mathcal{G} = \sigma(G_1, G_2, \ldots, G_n)$. Let $Y = \mathbb{E}[X \mid \mathcal{G}]$.

Measurability: Y is a constant on each minimal event G_i. Therefore, Y is $\mathcal{G}$-measurable, i.e., $Y^{-1}(B) \in \mathcal{G}$, for all Borel B.

Consistency: Average of Y on events in $\mathcal{G}$ equals the average of X on the event.

$$\mathbb{E}[X\mathbb{1}_G] = \mathbb{E}[Y\mathbb{1}_G], \ \forall G \in \mathcal{G}. \tag{140}$$

This is also often called the coarse-graining property.

Proof.

$$\langle X - Y, W \rangle = 0, \ \forall W \in L_2(\Omega, \mathcal{G}, \mathbb{P}). \tag{141}$$

Then $\mathbb{E}[XW] = \mathbb{E}[YW]$. Since $\mathbb{1}_G$ is measurable in $\mathcal{G}$, the statement is proved. □

8.4 Conditional expectation in L_1

So far we have only defined conditional expectation in L_2. How can we make it work if it only belongs to L_1?

Definition 8.2 (Kolmogorov, 1933)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{G} \in \mathcal{F}$ be a σ-algebra on Ω. Let $X \in L_1(\Omega, \mathcal{F}, \mathbb{P})$. A real random variable Y is a version of $\mathbb{E}[X \mid \mathcal{G}]$ if*

- *Y is integrable: $Y \in L_1(\Omega, \mathcal{F}, \mathbb{P})$*
- *Y is $\mathcal{G}$-measurable*
- *Consistency: $\mathbb{E}[Y\mathbb{1}_G] = \mathbb{E}[X\mathbb{1}_G]$ for all $G \in \mathcal{G}$.*

Theorem 8.3 (Fundamental theorem of conditional expectation)**.** *There exists a version Y of the conditional expectation $\mathbb{E}[X \mid \mathcal{G}]$. If Y' is a another version, then $\mathbb{P}[Y = Y'] = 1$.*

Proof Sketch. **Existence.** WLOG assume $X \geq 0$. Let $X_n(\omega) = \min\{X(\omega), n\}$.

X_n is positive, bounded, $\mathcal{F}$-measurable, and mono-increasingly converges to X point-wise.

X_n bounded implies that $X_n \in L_2$, so we can compute its conditional expectation $Y_n = \mathbb{E}[X_n \mid \mathcal{G}]$. We can check that $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. Define

$$Y(\omega) = \limsup_{n \to \infty} Y_n(\omega) \text{ for } \omega \in \Omega. \tag{142}$$

By monotone convergence theorem, $\mathbb{E}[Y] = \lim_{n\to\infty} \mathbb{E}[Y_n] = \lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X] < \infty$. Following a similar argument, $\mathbb{E}[Y\mathbb{1}_G] = \mathbb{E}[X\mathbb{1}_G]$ for all $G \in \mathcal{G}$.

Moreover, $Y = \limsup_{n\to\infty} Y_n$ is $\mathcal{G}$-measurable, since Y_n are measurable.

Uniqueness. Assume Y, Y' are both versions of conditional expectations. By consistency $\mathbb{E}[(Y - Y')\mathbb{1}_G] = 0$ for all $G \in \mathcal{G}$. For contradiction, assume $\mathbb{P}\{Y > Y'\} > 0$.

Let $E_n = \{Y > Y' + 1/n\}$, which are $\mathcal{G}$-measurable events. E_n mono-increasing converges to $E = \{Y > Y'\}$. This means $\mathbb{P}(E_n)$ mono-increasingly converges to $\mathbb{P}\{Y > Y'\} > 0$. Therefore, there exists $n \in \mathbb{N}$, such that $\mathbb{E}[(Y - Y')E_n] > \frac{1}{n}\mathbb{P}(E_n) > 0$, contradiction. □

Proposition 8.4 (Conditional expectation is an expectation.)**.** *Conditional expectation satisfies the same properties as expectation.*

- *Unital and positive.*
- *Conditional expectation is linear.*
- *Monotonic. If $X \leq Y$ a.s., then $\mathbb{E}[X \mid \mathcal{G}] \leq \mathbb{E}[Y \mid \mathcal{G}]$ a.s.*
- *Jenson's inequality holds: $\mathbb{E}[\varphi(x) \mid \mathcal{G}] \geq \varphi(\mathbb{E}[X \mid \mathcal{G}])$ a.s for convex φ.*

8.5 Convergence theorems

Theorem 8.5 (Conditional Monotonic convergence)**.** *Suppose $0 \leq X_n \uparrow X$ a.s. where $X \in L_1$. Then*

$$\mathbb{E}[X_n \mid \mathcal{G}] \uparrow \mathbb{E}[X \mid \mathcal{G}] \text{ a.s.} \tag{143}$$

Proposition 8.6. *Some properties for conditional expectation:*

- *Expectation: If* $Y = \mathbb{E}[X \mid \mathcal{G}]$*, then* $\mathbb{E}[X] = \mathbb{E}[Y]$
- *Full knowledge: If* X *is* $\mathcal{G}$*-measurable, then* $\mathbb{E}[X \mid \mathcal{G}] = X$ *a.s.*
- *Independence: If* $\sigma(X)$ *independence of* $\mathcal{G}$*, then* $\mathbb{E}[X \mid \mathcal{G}]$*, then* $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ *a.s.*
- *Pull-through: If* Z *is bounded and* $\mathcal{G}$*-measurable, then* $\mathbb{E}[XZ \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{G}]Z$*, a.s.*
- *Tower: If* $\mathcal{G} \subseteq \mathcal{H} \subseteq \mathcal{F}$ *are increasing* σ*-algebras. Then*

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}] \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{G}], a.s. \tag{144}$$

Conditional integration:

$$\mathbb{E}[h(X)|\mathcal{G}](\omega) = \int h(x)\mu_{X|\mathcal{G}}(\mathrm{d}x|\omega), a.s. \tag{145}$$

8.6 Gaussian and conditioning

Definition 8.7 (Multivariate normal random variable)**.** $\mathbf{X} = (X_1, \ldots, X_n)$ *is a multivariate normal random vector if it is an affine function of a standard normal vector* $\mathbf{Z} = (\mathbf{Z_1}, \ldots, \mathbf{Z_n})$*.*

$$\mathbf{X} = m + \Sigma\mathbf{Z}, \tag{146}$$

where $m \in \mathbb{R}^n$ *and* $\Sigma \in \mathbb{R}^{n \times n}$*.*

Remark.

- $\mathbb{E}[X_i] = m_i, Cov(X_i, X_j) = C_{ij}$, where $C = \Sigma\Sigma^* \succeq 0$.
- We denote $\mathbf{X} \sim \mathcal{N}(m, C)$, as m, C uniquely determine a multivariate normal random variable.

Definition 8.8 (Multivariate characteristic function)**.** *Let* $\mathbf{X}$ *be a random vector on* $\mathbb{R}^n$*. The characteristic function of* $\mathbf{X}$ *is* $\chi : \mathbb{R}^n \to \mathbb{C}$

$$\chi_{\mathbf{X}}(\theta) := \mathbb{E}[e^{i\theta^T\mathbf{X}}] \tag{147}$$

We can check that if $\mathbf{Y} = A\mathbf{X} + b$, then

$$\chi_{\mathbf{Y}}(\theta) = e^{i\theta^T b}\chi_{\mathbf{X}}(A^*\theta). \tag{148}$$

If $\mathbf{X}, \mathbf{Y}$ are independent random vectors on $\mathbb{R}^n$, then

$$\chi_{\mathbf{X}+\mathbf{Y}}(\theta) = \chi_{\mathbf{X}}(\theta)\chi_{\mathbf{Y}}(\theta). \tag{149}$$

It is easy to check that for $\mathbf{X} = m + \Sigma\mathbf{Z}$,

$$\chi_{\mathbf{X}}(\theta) = e^{i\theta^T m}e^{(-\|\Sigma^*\theta\|_2^2/2)} = \exp\left(i\theta^T m - \frac{\theta^T C\theta}{2}\right). \tag{150}$$

Theorem 8.9. *Suppose* X *are random vectors taking values in* $\mathbb{R}^n$*, then the laws of* X *and* Y *are the same if and only if the characteristic functions are the same.*

Theorem 8.10 (Linear marginals)**.** *A random vector* X *on* $\mathbb{R}^n$ *is normal if and only if* $\langle a, X\rangle$ *is a real normal random variable for every* $a \in \mathbb{R}^n$*.*

8.6.1 Gaussian conditioning

Let $(X, \mathbf{Y}) \in \mathbb{R}^{n+1}$ be multivariate normal distribution. $\mathbf{Y} = (Y_1, \ldots, Y_n)$. Assume $\mathbb{E}[X] = 0$ and $\mathbb{E}[\mathbf{Y}] = 0$.

We want to find the best approximation of X as a linear function of Y:

$$\min_{a \in \mathbb{R}^n} \|X - \langle a, \mathbf{Y} \rangle\|^2. \tag{151}$$

We define $c_{XX} = Var(X)$, $c_{XY} = Cov(X, \mathbf{Y}) \in \mathbb{R}^n$, and $C_{YY} = Cov(\mathbf{Y}, \mathbf{Y}) \in \mathbb{R}^{n \times n}$.

Then

$$\mathbb{E}[(X - \langle a, \mathbf{Y} \rangle)^2] = c_{XX} - 2\langle a, c_{XY} \rangle + a^T C_{YY} a. \tag{152}$$

By taking the derivative with respect to a,

$$a = C_{YY}^{-1} c_{XY} \tag{153}$$

The best approximation is thus

$$\hat{X} = \langle a, \mathbf{Y} \rangle = c_{XY}^T C_{YY}^{-1} \mathbf{Y}. \tag{154}$$

This means that

$$X - \hat{X} \perp span(Y_1, \ldots, Y_n). \tag{155}$$

Then, $Cov(X - \hat{X}, Y) = 0$. Specifically, for normal distributions, uncorrelation means independence (which can be shown using characteristic functions as normal random variables are determined by covariance).

Theorem 8.11 (Gaussian conditional expectation.)**.** *Different from general random variables, linear functions $\langle a, Y \rangle$ gives full information for conditioning.*

$$\mathbb{E}[X|Y] = \hat{X} = c_{XY} C_{YY}^{-1} Y. \tag{156}$$

Proof.

$$\begin{aligned}
\mathbb{E}[X|Y] &= \mathbb{E}[(X - \hat{X}) + \hat{X}|Y] \\
&= \mathbb{E}[X - \hat{X}|Y] + \mathbb{E}[\hat{X}|Y]. \\
&= \mathbb{E}[X - \hat{X}] + \hat{X}. && \text{(independence and full information)} \\
&= \hat{X}.
\end{aligned}$$

□

9 Martingales

Recall that a σ-algebra captures information about the world. A bigger σ-algebra includes more knowledge.

A filtration is a mathematical model for accumulating information.

9.1 Filtration

Definition 9.1 (Filtration)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A (discrete-time) filtration is a sequence of increasing σ-algebras on Ω:*

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_\infty \subseteq \mathcal{F}. \tag{157}$$

Commonly, define $\mathcal{F}_0 = \{\varnothing, \Omega\}, \mathcal{F}_\infty = \sigma(\bigcup_{k=1}^{\infty} \mathcal{F}_k)$.

Examples.

Let $Z_0, Z_1, \ldots, Z_n$ be real random variables, and $\mathcal{F}_k = \sigma(Z_0, \ldots, Z_k)$. Then $\{\mathcal{F}_k\}$ is a filtration.

Definition 9.2 (Adapted process)**.** *A sequence $(X_0, X_1, \ldots)$ of real random variables is adapted to the filtration $(\mathcal{F}_k : k \in \mathbb{Z}_+)$ if each X_k is $\mathcal{F}_k$-measurable.*

In other words, at time k, the value of X_k is determined by the information we have.

9.2 Defining Martingales

Definition 9.3 (Martingales (Informal))**.** *A martingale is an adapted process that is indifferent about the future.*

That is, for $n \geq k$, $\mathbb{E}[X_n \mid \mathcal{F}_k] = X_k$.

Definition 9.4 (Martingales (Formal))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_\infty$. An sequence of real random variables $(X_i : i \in \mathbb{Z}^+)$ is a martingale if*

- *Adapted: (X_k) is adapted to the filtration $(\mathcal{F}_k)$*
- *Integrable: $\mathbb{E}|X_k| < \infty$, for each k.*
- *Status Quo: $\mathbb{E}[X_{k+1}|\mathcal{F}_k] = X_k$ almost sure for each k.*

We can relax the third property to $\mathbb{E}[X_{k+1}|\mathcal{F}_k] \leq X_k$ a.s., in which (X_k) is called a supermartingale.

$\mathbb{E}[X_{k+1}|\mathcal{F}_k] \geq X_k$ a.s. is called a submartingale.

This seemingly awkward terminology comes from the super/sub-harmonic function of Markov chains.

Examples. Independent sums of centered random variables, Random walks, Levy-Doob process.

9.3 Formal model for Gambling strategies

Definition 9.5 (Previsible process.). *A sequence of real random variables $(C_1, C_2, \dots,)$ is a previsible with respect to the filtration $(\mathcal{F}_k)$ if c_k is $\mathcal{F}_{k-1}$-measurable for all $k \in \mathbb{N}$.*

When you play a game of chance, you bet before the game.

Definition 9.6 (Martingale transform). *Let (X_k) be a martingale and $(C_k : k \in \mathbb{N})$ be a previsible sequence. The martingale transform is the sequence*

$$(C \cdot X)_k = \sum_{i=1}^{k} C_i(X_i - X_{i-1}), \text{ for } k = 1, 2, \dots. \tag{158}$$

Remark. You bet C_i units of capital on a fair game (martingale). Then you win $C_i(X_i - X_{i-1})$ on the i-th day. The total winnings after k plays is

$$(C \cdot X)_k = \sum_{i=1}^{k} C_i(X_i - X_{i-1}). \tag{159}$$

Proposition 9.7 (Martingale transform.). *Let $(X_k : k \in \mathbb{Z}_+)$ be a martingale and $(C_k : k \in \mathbb{N})$ be a previsible process that are bounded. Then*

$$((C \cdot X)_k : k \in \mathbb{Z}_+) \tag{160}$$

is a null martingale, i.e., $(C \cdot X)_0 = 0$.

Proof. Pull-through property:

$$\mathbb{E}[(C \cdot X)_{k+1} - (C \cdot X)_k | \mathcal{F}_k] = \mathbb{E}[C_{k+1}(X_{k+1} - X_k)|\mathcal{F}_k] = C_{k+1}\mathbb{E}[X_{k+1} - X_k|\mathcal{F}_k] = 0. \tag{161}$$

Since $(C \cdot X)_k$ is $\mathcal{F}_k$ measurable,

$$\mathbb{E}[(C \cdot X)_{k+1}|\mathcal{F}_k] = (C \cdot X)_k. \tag{162}$$

□

We interpret this proposition: *No sequence of bounded bets allows you to beat a fair game.*

Another question is: can you quit the game when you are ahead to get advantages in a fair game?

Definition 9.8 (Stopping times). *A random variable $\tau : \Omega \to \mathbb{Z}_+ \cup \{+\infty\}$ is a stopping time, if $\{\tau \leq k\}$ is $\mathcal{F}_k$ measurable for each $k \in \mathbb{Z}_+$.*

Remark. The meaning of this definition is that after game k, you have enough information to decide whether the stopping criterion is triggered.

Definition 9.9 (Stopped process)**.** *Let (X_k) be an adapted process and τ be a stopping time. The stopped process is the sequence*

$$(X_{k\wedge\tau} : k \in \mathbb{Z}_+), \text{ where } k \wedge \tau = \min\{k, \tau\}. \tag{163}$$

The idea of this definition is that your winnings freeze after you stop gambling.

Definition 9.10 (Stopped martingales.)**.** *If (X_k) is a martingale, τ is the stopping time, then the stopped process $(X_{k\wedge\tau})$ is a martingale.*

Remark. This means *you cannot gain an advantage in a fair game by quitting strategically.*

Proof. Consider a previsible process $(C_k : k \in \mathbb{N})$,

$$C_k = \begin{cases} 1, \ k \leq \tau, \\ 0, \ k > \tau. \end{cases} \tag{164}$$

Since C_k is previsible and bounded, the martingale $(C \cdot X)_k$ is a martingale with an initial value 0. This means

$$(C \cdot X)_k = \sum_{i=1}^{k} C_i(X_i - X_{i-1}) = \sum_{i=1}^{k\wedge\tau}(X_i - X_{i-1}) = X_{k\wedge\tau} - X_0. \tag{165}$$

Therefore,

$$\mathbb{E}[X_{k\wedge\tau}|\mathcal{F}_k] = X_0. \tag{166}$$

□

Warning: It is not true that $\mathbb{E}[X_\tau] = \mathbb{E}[X_0]$ as τ is a random variable dependent on the outcomes of (X_k). For example, we stop when $X_k \geq 1$, then $\mathbb{E}[X_\tau] \geq 1$.

However, with the following theorem, we circumvent this paradox by restricting the horizon of our game plays.

Theorem 9.11 (Optional stopping)**.** *Let (X_k) be a martingale. and τ be a stopping time. If $\tau \leq B$ for a fixed B almost surely, then*

$$\mathbb{E}[X_\tau] = \mathbb{E}[X_0]. \tag{167}$$

Proof. By the proposition on the stopped martingales,

$$\mathbb{E}[X_0] = \mathbb{E}[X_{k\wedge\tau}] \text{ for each } k \tag{168}$$

Choose $k = B$, so

$$\mathbb{E}[X_0] = \mathbb{E}[X_{B\wedge\tau}] = \mathbb{E}[X_\tau]. \tag{169}$$

□

9.4 Convergence of martingales

Martingales are flavored by probability theorists because they converge to a stable equilibrium under weak assumptions.

Theorem 9.12 (Doob's convergence theorem)**.** *Let $(X_k : k \in \mathbb{Z}_+)$ be a martingale sequence that is **uniformly bounded in** L_1, i.e., there exists a positive R, such that $\mathbb{E}[|X_k|] \le R$ for all $k \in \mathbb{Z}_+$.*

Then almost surely, $X_\infty(\omega) := \lim_{k\to\infty} X_k(\omega)$ exists, and is finite. That is, X_k converges to X_∞ almost surely.

Remark. It is not true in general that $\mathbb{E}[X_k]$ converges to $\mathbb{E}[X_\infty]$.

Lemma 5 (Interval sandwich)**.** *A nonrandom real-valued sequence (x_k) fails to converge if and only if*

$$\liminf_{k\to\infty} x_k < a < b < \limsup_{k\to\infty} x_k, \tag{170}$$

for some $a, b \in \mathbb{Q}$.

We define "upcrossing" of a sequence to an interval.

Definition 9.13. *Let (x_k) be a nonrandom real sequence. Fix $a < b$. The number $u_N a, b$ of upcrossings before time N is the largest m, such that*

$$0 \le s_1 \le t_1 < s_2 < t_2 < \cdots < s_m < t_m \le N, \tag{171}$$

where $x_{s_i} < a$ and $x_{t_i} > b$.

The total number of upcrossings:

$$u_\infty[a, b] = \lim_{N\to\infty} u_N[a, b]. \tag{172}$$

Lemma 6. *Equation (171) implies $u_\infty[a, b] = \infty$. If $u_\infty[a, b]$ is finite for all rational pairs (a, b) with $a < b$, then (x_k) converges in $\mathbb{R}$.*

9.4.1 Proof of Doob's convergence

Idea: if a martingale crosses an interval $[a, b]$ infinitely times, we can make money by betting on the upcrossings. This is impossible.

Specifically, we keep betting below b, and start betting when x_k gets below a. We formalize it as below.

Fix $a < b$. Let $c_1 = \mathbb{1}\{x_0 < a\}$.

Let $c_k = \mathbb{1}\{c_{k-1} = 1, x_{k-1} \le b\} + \mathbb{1}\{c_{k-1} = 0, x_{k-1} < a\}$.

We can check that (c_k) is positive, bounded, and previsible. Then the martingale transform $Y_k = (C \cdot X)_k$ is a null-martingale.

Define $U_N[a,b](\omega)$ as the number of upcrossings of $[a,b]$ up to time N by the sample path $X_k(\omega)$, and $U_\infty[a,b](\omega) = \lim_{N\to\infty} U_N[a,b](\omega)$.

Note that

$$Y_N(\omega) \geq (b-a)U_N[a,b](\omega) - (X_N(\omega) - a) \tag{173}$$

Since $\mathbb{E}[Y_n] = 0$, so we have the following proposition:

Proposition 9.14 (Snell upcrossing inequality)**.** *Let (X_k) be a martingale and $a < b$ be real numbers.*

$$(b-a)\mathbb{E}[U_N[a,b]] \leq \mathbb{E}[(X_N - a)] \tag{174}$$

As $\mathbb{E}[|X_N|] \leq R$, then for any $a < b$,

$$\mathbb{P}\{U_\infty[a,b] = \infty\} = 0. \tag{175}$$

Define an event $E = \{\omega : X_k(\omega) \text{ does not converge in } \mathbb{R}\}$,

$$E = \bigcup_{a<b,a,b\in\mathbb{Q}} \{\omega : \liminf x_k < a < b < \limsup x_k\} \subseteq \bigcup_{a<b,a,b\in\mathbb{Q}} \{\omega : U_\infty[a,b] = \infty\}. \tag{176}$$

Therefore, $\mathbb{P}(E)$, because rational number set $\mathbb{Q}$ is countable. Thus, $X_\infty(\omega) = \lim_{k\to\infty} X_k(\omega)$ with probability 1. Moreover,

$$\mathbb{E}|X_\infty| = \mathbb{E}[\liminf_{k\to\infty} |X_k|] \leq \liminf_{k\to\infty} \mathbb{E}|X_k| \leq R. \tag{177}$$

9.5 Concentration inequalities for martingales

What is the probability that a martingale ever deviates much from its mean value?

Begin with submartingales: $\mathbb{E}[X_{k+1}|\mathcal{F}_k] \geq X_k$ a.s.

Theorem 9.15 (Doob's maximal inequality for submartingales)**.** *Consider a positive submartingale $(X_k : k \in \mathbb{Z}_+)$. For each $N \in \mathbb{Z}_+$,*

$$\mathbb{P}\{\sup_{0\leq k\leq N} X_k > t\} \leq \frac{\mathbb{E}X_N}{t}. \tag{178}$$

Moreover, if $X_n \to X_\infty$ in L_1, then it's true for $N = \infty$.

Remark. Though this statement looks like Markov's inequality, we cannot derive it directly from Markov's inequality.

Proof. We want to bound the probability that the martingale escapes a band. We define the stopping time $\tau : \Omega \to \mathbb{Z}_+ \cup \{\infty\}$,

$$\tau := \inf\{k \leq N : X_k > t\}. \tag{179}$$

Therefore, the event

$$E := \{\sup_{k\leq N} X_k > t\} = \{\tau < \infty\}. \tag{180}$$

Then,

$$\begin{aligned}\mathbb{E}X_N &\geq \mathbb{E}X_{N\wedge\tau} \\ &\geq \mathbb{E}[X_{N\wedge\tau}\mathbb{1}_E] \\ &= \mathbb{E}[X_\tau \mathbb{1}_E] \\ &> t\mathbb{P}[E] \end{aligned} \tag{181}$$

Therefore,

$$\mathbb{P}(\sup_{k\leq N} X_k > t) \leq \frac{\mathbb{E}[X_N]}{t}. \tag{182}$$

Moreover, $E_N := \{\sup_{k\leq N} X_k > t\}$ mono-increasingly converges to E_∞, and $\mathbb{E}X_N \to \mathbb{E}[X_\infty]$ by L_1 convergence. By dominated convergence theorem the statement holds for $N = \infty$. □

Proposition 9.16 (Martingale: convex transformation.)**.** *Consider a martingale $(M_k : k \in \mathbb{Z}_+)$. For any convex function $\varphi : \mathbb{R} \to \mathbb{R}$, the sequence*

$$X_k := \varphi(M_k) \tag{183}$$

is a submartingale, provided that $\mathbb{E}|X_k| < \infty$.

Proof. By conditional Jensen's inequality. □

We can then generalize Chebyshev's inequality to Martingales,

Theorem 9.17 (Kolmogorov's inequality)**.** *Let $(M_k : k \in \mathbb{Z}_+)$ be a martingale in L_2. For $N \in \mathbb{Z}_+$,*

$$\mathbb{P}\{\max_{k\leq N}(X_k - \mathbb{E}X_k)^2 > t\} \leq \frac{Var[X_N]}{t}. \tag{184}$$

Proof. The sequence $(X_k - \mathbb{E}X_k)$ is an L_2 martingale, so $(X_k - \mathbb{E}X_k)^2$ is a submartingale, due to Proposition 9.16. Using Doob's maximal inequality the statement follows. □

Remark. Define $\Delta_k = X_k - X_{k-1}$. Then

$$Var[X_n] = \sum_{k=1}^{N} \mathbb{E}[\Delta_k^2]. \tag{185}$$

Proposition 9.18 (Exponential maximal inequality)**.** *Let $(X_k : k \in \mathbb{Z}_+)$ be a bounded martingale. For $N \in \mathbb{Z}_+$,*

$$\mathbb{P}\{\max_{k\leq N} X_k > t\} \leq \inf_{\theta>0} \exp(-\theta t + \xi_{X_N}(\theta)), \tag{186}$$

where $\xi_X(\theta) = \log \mathbb{E}e^{\theta X}$ is the cgf of X.

Theorem 9.19 (Azuma-Hoeffding: uniform version)**.** *Consider a martingale $(X_k : k \in \mathbb{Z}_+)$ whose difference sequence $\Delta_k := X_k - X_{k-1}$ is bounded, $|\Delta_k| \leq a_k$. Let the variance proxy $v_N := \sum_{k=1}^N a_k^2$.*

$$\mathbb{P}\{\max_{k\leq N} |X_k - X_0| > t\} \leq 2\exp\big(-t^2/2v_N\big). \tag{187}$$

Proof. Hoeffding lemma: If Y is a real random variable with $\mathbb{E}Y = 0$, $|Y| \leq a$, then its mgf $\xi_Y(\theta) \leq \frac{1}{2}\theta^2 a^2$. As $X_N - X_0 = \sum_{k=1}^N \Delta_k$,

$$\begin{aligned} m_{X_N - X_0}(\theta) &= \mathbb{E}[e^{\theta\Delta_N} \dots e^{\theta\Delta_1}] \\ &= \mathbb{E}[\mathbb{E}[e^{\theta\Delta_N|\mathcal{F}_{N-1}}]e^{\theta\Delta_{N-1}} \dots e^{\theta\Delta_1}] \\ &\leq e^{\theta^2\Delta_N/2}\mathbb{E}[e^{\theta\Delta_{N-1}} \dots e^{\theta\Delta_1}] \\ &\leq \dots \leq e^{\theta^2\nu_N/2}. \end{aligned} \tag{188}$$

Therefore,

$$\xi_{X_N - X_0}(\theta) \leq \frac{1}{2}\theta^2\nu_N. \tag{189}$$

Using this result with $\theta = t/\nu_n$ in the Exponential maximal inequality proves the theorem. □

Theorem 9.20 (Ville's inequality)**.** *Let $(X_k : k \in \mathbb{Z}_+)$ be a positive supermartingale. Then*

$$\mathbb{P}\{\sup_{k\in\mathbb{Z}_+} X_k > t\} \leq \frac{\mathbb{E}X_0}{t}. \tag{190}$$