21-721 Probability

Based on lectures by Tikhomirov

Notes taken by Kadin Zhang

Spring 2024

Contents

1 Probability spaces

Random variables X on probability space $(\Omega, \Sigma, \mathbf{P})$ are Σ-measurable maps $X : \Omega \to \mathbb{R}$. The prototypical motivation for Σ-measurability is allowing us to integrate over $\mathbb{R}$ with respect to the distribution (a probability measure) $\mu_X : \Sigma \to \mathbb{R}$ defined by $\mu_X(A) := \mathbf{P}(X \in A)$ instead of Ω:

$$\mathbf{E}X = \int_\Omega X(\omega)\, d\,\mathbf{P}(\omega) = \int_\mathbb{R} t\, d\mu_X(t).$$

Radon Nikodym says when μ_X is absolutely continuous to Lebesgue measure, there exists density p

$$\mu_X(A) = \int_A p(t)\, dt,$$

then we may extend change of variables to

$$\int_\mathbb{R} t d\mu_X(t) = \int_\mathbb{R} tp(t)\, dt.$$

1.1 1/17 - Measure theory review

Definition (Field/Algebra): Let Ω be a set, Σ a collection of subsets of Ω. Then Σ is a *field* (algebra) if

(a) $\varnothing, \Omega \in \Sigma$.

(b) If $A \in \Sigma$, then $A^c \in \Sigma$.

(c) For all $A, B \in \Sigma$, $A \cap B \in \Sigma$.

Definition (σ-field): Let Ω be a set. Σ is a *σ-field* if it is a field and closed under countable intersections.

Example (σ-fields):

(a) $\{\varnothing, \Sigma\}$.

(b) 2^Σ.

(c) Let Ω be a finite set with $\Omega = \bigcup_{i=1}^m A_i$ disjoint. Then the set

$$\{\bigcup_{i \in I} A_i : I \subset \{1, 2, \ldots, m\}\}$$

is a σ-field.

(d) Let $\Omega = \mathbb{Z}$.

Example: Let $\Sigma = \mathbb{Z}$, Σ the sets A in Ω such that either A is finite or A^c is finite. This is a field but not a σ-field: consider

$$A_i = \{2i\}.$$

Then $\bigcup_{i \in \mathbb{Z}} A_i$, the even numbers, is not in Σ.

Definition (Generation of σ-field): Let $T \subset 2^\Omega$. Then Σ is a σ-field *generated* by T, denoted $\Sigma = \sigma(T)$, if Σ is the intersection of all σ-fields containing T.

Example: Let (T, τ) be a topological space, we define the Borel σ-field on T as the σ-field generated by the open sets. In particular, $\mathcal{B}(\mathbb{R}^N)$ is generated by the sets $(-\infty, a] : a \in \mathbb{R}$ (for each dimension).

Definition (Measurable space): A *measurable space* is a pair (Ω, Σ) where Σ is a σ-field of subsets of Ω.

Definition (Product measure space): Let I be an index set and for each $i \in I$, (Ω_i, Σ_i) is a measurable space. Then the *product measure space* (Ω, Σ) is defined on $\Omega = \prod_i \Omega_i$ where Σ is generated by sets of the form

$$\prod_i A_i$$

where $A_i \in \Sigma_i$ and $A_i = \Omega_i$ for all but at most countably many indices $i \in I$.

Remark: $\mathbb{R}^N$ can be viewed as a product topological space of $(\mathbb{R}, \tau)$ and $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$ is the corresponding product measurable space.

Definition (Measure): Let (Ω, Σ) be a measurable space and $\mu : \Sigma \to [0, \infty]$. Then μ is a *measure* if

(a) $\mu(\varnothing) = 0$.

(b) For disjoint subsets $A_i \in \Sigma$,

$$\mu(\bigcup_i A_i) = \sum_i \mu(A_i).$$

Definition (σ-finite measure): A measure μ is σ-finite if there is countable collection of measurable sets $\bigcup_i A_i = \Omega$ such that for all i, $\mu(A_i) < \infty$. For example, the Lebesgue measure in $\mathbb{R}^N$.

Definition (Probability measure): μ is *probability measure* if $\mu(\Omega) = 1$.

Definition: A triple (Ω, Σ, μ) is *measure space* if μ is a measure on (Ω, Σ). It is called a *probability space* if μ is a probability measure.

1.2 1/19 - π, λ systems, random variables

Definition (π-system): Let P be a collection of subsets of Ω closed under intersections. Then P is called a *π-system*.

Definition (λ-system): Let $\Omega, L \subset 2^\Omega$ such that

(a) $\varnothing, \Omega \in L$.

(b) $A \in L \implies A^c \in L$.

(c) L is closed under disjoint countable union.

Remark: There are λ-systems that are not σ-fields. Every σ-field is a λ-system.

Theorem (Dynkin's $\pi\lambda$-theorem): Suppose $P, L \subset 2^\Omega$ are π and λ systems respectively and assume $P \subset L$. Then $\sigma(P) \subset L$.

Example (λ-system but not σ-field): Let $\Omega = \{HH, HT, TH, TT\}$, $\Sigma = 2^\Omega$. Let P_1 be the uniform probability measure. Let $P_2(\{HH\}) = P_2(\{TT\}) = 1/2$. Define L as the collection of elements of Σ such that $P_1(A) = P_2(A)$. L is a λ-system (easy to verify). But L is not a σ-field:

$$L = \{\varnothing, \Omega, \{HH, HT\}, \{HH, TH\}, \{TT, TH\}, \{TT, HT\}\}.$$

Example (Probability spaces):

(a) Suppose we have 52 cards and we want to model uniform random shuffling. We let Ω be the set of permutations and define Σ as the powerset.

Definition (Random variable): Let (Ω, Σ) and (Ω', Σ') be measurable spaces. We call $X : \Omega \to \Omega'$ *measurable* if for all $A' \in \Sigma'$,

$$\{\omega \in \Omega : X(\omega) \in A'\} \in \Sigma.$$

When $(\Omega', \Sigma') = (\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$, this means that the preimage of every Borel subset is measurable. If in the above definition (Ω, Σ, P) is a probability space, then the mapping X is called *random*. If moreover the target space is $(\mathbb{R}, \mathcal{B})$ or $(\mathbb{C}, \mathcal{B}_\mathbb{C})$ then X is *random variable*.

A *random vector* is a $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ or $(\mathbb{C}^n, \mathcal{B}_{\mathbb{C}^n})$-valued random variable.

Proposition (Random variables are preserved under usual transformations and operations): Suppose $X_1, \ldots, X_m$ are random variables on (Ω, Σ, P). Then:

(a) If $f : \mathbb{R}^m \to \mathbb{R}^n$ is a Borel measurable function then $f(X_1, \ldots, X_m)$ is a random vector in $\mathbb{R}^n$.

(b) $(X_1, \ldots, X_m)$ is a random vector. We can show this with $\pi\lambda$ theorem considering preimages of rectangles.

(c)

$$\limsup_n X_i, \liminf_i X_i$$

are random variables.

1.3 1/22 - Distribution, law

Definition (Law): Let $X : (\Omega, \Sigma, \mathbf{P}) \to (\mathbb{R}^n, \mathcal{B}_n)$ be a random vector. Then μ_X is the *law or induced Borel probability measure* on $(\mathbb{R}^n, \mathcal{B}_n)$ defined by

$$\mu_X(A) = \mathbf{P}(\{\omega : X(\omega) \in A\}).$$

Remark: μ_X is well defined on $\mathcal{B}_n$ since X is a random vector. $\mu_X(\mathbb{R}^n) = 1$ follows since $X^{-1}(\mathbb{R}) = \Omega$ and $\mathbf{P}$ is a probability measure. For disjoint Borel sets B_i, additivity follows:

$$\mathbf{P}(X^{-1}(\bigcup_i B_i)) = \mathbf{P}(\bigcup_i X^{-1}(B_i)) = \sum_i \mathbf{P}(X^{-1}(B_i)) = \sum_i \mu_X(B_i).$$

Definition (Distribution function): The *distribution function* F_X of a real-valued R.V. X is

$$F_X(t) = \mathbf{P}(\{\omega : X(\omega) \leq t\}) = \mu_X((-\infty, t]).$$

We will also use the notation F_μ to denote distribution function specified by an arbitrary Borel probability measure.

Proposition (Characterization of distribution functions): A function $F : \mathbb{R} \to [0, 1]$ is the distribution function of some random variable if and only if:

(a) F is non-decreasing.

(b) $\lim_{t\to\infty} F(t) = 1$ and $\lim_{t\to-\infty} F(t) = 0$.

(c) F is right-continuous, i.e. for all $t \in \mathbb{R}$,

$$F(t) = \lim_{s\to 0^+} F(t + s).$$

Proof. Let F_X be the distribution function of X. F_X is monotonic by monotonicity of measure. (a) and (b) follow by writing $\varnothing = \bigcap_n (-\infty, n]$ and $\mathbb{R} = \bigcup_n (-\infty, n]$ and using measure of increasing/decreasing sets. Similarly for right continuity we may write $(-\infty, t + s] \downarrow (-\infty, t]$ as $s \downarrow 0$.

On the other hand, suppose we have (a), (b), and (c). Consider the R.V.

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

on $((0, 1], \mathcal{B}_{(0,1]}, \mathcal{L})$. Then, for $y \in \mathbb{R}$,

$$F_{X^-}(y) = \mathcal{L}((0, F(y)]) = F(y).$$

□

Example (Durrett 1.2.7): Let $F(x) = \mathbf{P}(X \leq x)$ be continuous. Then the RV $Y = F(X)$ is a uniform distribution on $(0, 1)$ "percentile of a random test score from the class is uniform".

Proposition (Probability measures with same cdf are equal):

Proof. Assume $F_{\mu_1} = F_{\mu_2}$. Then for all $t \in \mathbb{R}$,

$$\mu_1((-\infty, t]) = \mu_2((-\infty, t]).$$

So the measures agree on the π-system generating $\mathcal{B}$. Now let L be the sets on which the measures agree. L is a λ-system (direct by properties of measure). Thus by $\pi - \lambda$ theorem the measures agree on $\mathcal{B}$.

We can show that in fact distributions with the same probability measures are equal by Caratheodory. □

Corollary: If X is a random variable then μ_X is completely determined by the distribution function of X.

Definition (Characterization of distribution functions in $\mathbb{R}^n$): A multivariate distribution function $F : \mathbb{R}^n \to [0, 1]$ is

(a) Monotonically non-decreasing in each coordinate.

(b) Right continuous, i.e. for all sequences of vectors $x_n \downarrow x$,

$$F(x_n) \to F(x).$$

(c) For all $i \leq n$,

$$\lim_{t_i \to -\infty} F(t_1, \dots, t_n) = 0.$$

(d)

$$\lim_{t_1, t_2, \dots, t_n \to \infty} F(t_1, \dots, t_n) = 1.$$

(e)

$$0 \leq F(t_1, \dots, t_n) \leq 1.$$

1.4 1/24 - Distribution functions cont.

Suppose we have distribution function F. Let μ be the unique Borel probability measure with $F_\mu = F$ (by bijection from last lecture). Define X as the identity mapping on $(\mathbb{R}^n, \mathcal{B}_n, \mu)$. Then,

$$F_X(t_1, \dots, t_n) = \mu(X^{-1}(\prod_i \{(-\infty, t_i]\})) = \mu(\prod_i \{(-\infty, t_i]\}) = F(t_1, \dots, t_n).$$

Now, if we wanted to construct a random variable with the Lebesgue measure,

Proposition: Let $F : \mathcal{R}^n \to [0, 1]$ be a distribution function. There is a random mapping

$$X : ([0, 1], \mathcal{B}([0, 1]), \mathcal{L}) \to \mathbb{R}^n$$

with cdf F.

Proof. For 1D case, this is proved in 1/22 notes. □

1.5 1/26 - Integration, density, expectation

Recall: Suppose we have a measurable function f on (Ω, Σ, μ). Then there exists a sequence of measurable simple functions f_n such that $\lim_{n\to\infty} f_n = f$. Moreover if f is bounded then this convergence can be monotone. We then define for nonnegative f

$$\int f d\mu = \{\sup_s \int s d\mu : s \leq f, s \text{ simple}\}.$$

For general measurable $f = f^+ - f^-$, we write

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

Theorem (Monotone convergence theorem): Suppose $f_n \uparrow f$ is a sequence of nonnegative measurable functions increasing to f. Then

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu.$$

Theorem (Dominated convergence theorem): Let f_n be a sequence of measurable functions such that $\lim_{n\to\infty} f_n = f$ μ a.e. Suppose there is integrable g such that $f_n \leq g$ for all n. Then

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu$$

Theorem (Fatou's lemma): Let f_n nonnegative measurable functions. Then

$$\int \liminf_n f_n d\mu \leq \lim_{n\to\infty} \int f_n d\mu.$$

Theorem (Markov-Chebyshev inequality): Let $X : (\Omega, \Sigma, \mathbf{P}) \to [0, \infty]$. For all $t \geq 0$,

$$P(X \geq t) \leq \frac{1}{t} \int X d\,\mathbf{P}\,.$$

Corollary ("Chebyshev inequality"): For not necessarily nonnegative X we can plug in $|X - \mu|$. Then

$$\mathbf{P}(|X-\mu| \geq t) = \mathbf{P}((X-\mu)^2 \geq t^2) \leq \frac{\mathbf{E}(X-\mu)^2}{t^2} = \frac{\mathbf{Var} X}{t^2}.$$

Theorem (Jensen's inequality): Let $X : (\Omega, \Sigma, \mathbf{P}) \to \mathbb{R}$. Let ϕ convex. Assume $\int X d\,\mathbf{P}$ exists. Then

$$\phi\Big(\int X d\,\mathbf{P}\Big) \leq \int \phi(X) d\,\mathbf{P}.$$

Proof. Consider the tangent line to ϕ at $\int X d\,\mathbf{P}$, ℓ, such that $\ell \leq \phi$. Then,

$$\int \phi(X) d\,\mathbf{P} \geq \int \ell(X) d\,\mathbf{P} = \ell\Big(\int X d\,\mathbf{P}\Big) = \phi\Big(\int X d\,\mathbf{P}\Big).$$

Though this doesn't really need proof. "Weighted average of $\phi(X)$ is greater than ϕ over the weighted average of X, directly by convexity". □

Definition (Density): Let X be a random vector such that μ_X is absolutely continuous with respect to the Lebesgue measure (in $\mathbb{R}^n$ or $\mathbb{R}$). By Radon-Nikodym there exists a nonnegative measurable function $p_X : \mathbb{R}^n \to \mathbb{R}$ such that for Borel sets B,

$$\mu_X(B) = \int_B p_X(x) dx.$$

We call p_X, the Radon-Nikodym derivative of μ_X with respect to Lesbesgue measure, the *probability density function*.

Comments: Recall that μ_X is absolutely continuous w.r.t. Lebesgue measure if every set E of Lebesgue measure 0 has $\mu_X(E) = 0$.

Remark: For all $A \in \mathcal{B}_{\mathbb{R}^n}$,

$$\int_A p(z) dz = \mathbf{P}(X \in A).$$

This follows by noting that all subsets A satisfying the statement form a λ-system and applying Dynkin's theorem.

Definition: Let X be a random variable on $(\Omega, \Sigma, \mathbf{P})$ with $\int X d\,\mathbf{P}$ well defined. Then this integral is the *mean* or *expected value* or *first moment* notated

$$\mathbf{E}X = \int_\Omega X(\omega) d\,\mathbf{P}(\omega).$$

The nth moment is

$$\mathbf{E}X^n = \int_\Omega X^n(\omega) d\,\mathbf{P}(\omega).$$

If $p > 0$ the p-th absolute moment of X is

$$\mathbf{E}|X|^p.$$

Proposition: For $p < q$,

$$(\mathbf{E}|X|^p)^{1/p} \leq (\mathbf{E}|X|^q)^{1/q}.$$

Proof. Let $\phi(t) = t^{q/p}$. Apply Jensen's inequality. □

Theorem (Change of variables): Let f be Borel measurable. Then,

$$\mathbf{E}f(X) = \int_{\mathbb{R}} f(t) d\mu_X.$$

In other words, instead of taking a weighted average of X over Ω, we take a weighted average of X over $\mathbb{R}$.

Proof. We first show this is true for simple f. Then use MCT to show it holds in general. □

1.6 1/29 - Product spaces, moments, moment generating functions

Definition (Fubini-Tonelli): Let $(\Omega_1, \Sigma_1, \mu_1), (\Omega_2, \Sigma_2, \mu_2)$ be σ-finite measure spaces. Define (Ω, Σ, μ) by $\Omega = \Omega_1 \times \Omega_2, \Sigma = \Sigma_1 \times \Sigma_2, \mu = \mu_1 \otimes \mu_2$. Let $f : \Omega \to \mathbb{R}$ measurable either nonnegative or integrable. Then

$$\int_{\Omega} f d\mu = \int_{\Omega_1} \left(\int_{\Omega_2} f(\omega_1, \omega_2) d\mu_2 \right) d\mu_1.$$

Theorem (Expectation of nonnegative RV): Let X be a nonnegative random variable.

$$\mathbf{E}X = \int_0^\infty \mathbf{P}(X \geq t)\, dt.$$

Proof.

$$\begin{aligned}
\mathbf{E}X &= \int_\Omega X(\omega)\, d\mathbf{P}(\omega) \\
&= \int_\Omega \left(\int_0^\infty \mathbb{1}_{\{X(\omega) \geq t\}}(\omega, t)\, dt \right) d\mathbf{P}(\omega) \\
&= \int_0^\infty \left(\int_\Omega \mathbb{1}_{\{X(\omega) \geq t\}}(\omega, t)\, d\mathbf{P}(\omega) \right) dt \\
&= \int_0^\infty (1 - F_X(t))\, dt \\
&= \int_0^\infty \mathbf{P}(X > t)\, dt = \int_0^\infty \mathbf{P}(X \geq t)\, dt.
\end{aligned}$$

So if X is arbitrary random variable, we can write

$$\mathbf{E}X = \mathbf{E}X^+ - \mathbf{E}X^- = \int_0^\infty \mathbf{P}(X \geq t)\, dt - \int_0^\infty \mathbf{P}(-X \geq t)\, dt.$$

□

Proposition (Density in integral):

$$\int_{\mathbb{R}} t\, d\mu_X = \int_{\mathbb{R}} p(t)\, dt.$$

Proof. Intuitively $p(t)$ is $\mu_X([t-\varepsilon, t+\varepsilon])$. Proof? □

Definition (Moment generating function): Let X be a RV. Define

$$M_X(t) = \mathbf{E} \exp(tX).$$

This is always well-defined but can take on ∞.

Proposition: If $X_1, \dots, X_n$ are independent,

$$M_{X_1 + \cdots + X_n} = \prod_i M_{X_i}.$$

Proof. Easy. □

Theorem (Generating moments): Suppose $M_X(t)$ is finite in a neighborhood of 0. Then

$$M_X'(0) = \lim_{\delta \to 0} \mathbf{E} \frac{\exp(\delta X) - 1}{\delta} = \mathbf{E} \frac{1 + \delta X - 1}{\delta} = \mathbf{E} X.$$

Similarly we can find by induction that for all $n \geq 1$,

$$M_X^{(n)}(0) = \mathbf{E} X^n.$$

1.7 1/31 - Subexponential random variables, independence

Definition (Subexponential random variable): A random variable X is called *subexponential* if M_X is finite in a neighborhood of 0.

Proposition: X is subexponential $\iff$ $\limsup_p (\mathbf{E}|X|^p)^{1/p} \frac{1}{p} < \infty$.

Proof. Suppose $t > 0$. Note that we can bound $\exp(tX^+), \exp(tX^-)$ by

$$\exp(tX^+) \leq \max(1, \exp(tX)) \leq 1 + \exp(tX),$$

so

$$\begin{aligned} \mathbf{E} \exp(t|X|) &= \mathbf{E} \exp(tX^+ + tX^-) \\ &\leq 2 + \mathbf{E} \exp(tX) + \mathbf{E} \exp(-tX). \end{aligned}$$

Since X is subexponential there exists some neighborhood of 0 where $|X|$ is bounded by L, so for small ε,

$$\mathbf{E} \exp(\varepsilon |X|) \leq 2 + 2L < \infty.$$

By Taylor expansion e^x, for all $z > 0$, integer $p \geq 1$,

$$e^{\varepsilon z} \geq \frac{(\varepsilon z)^p}{p!} \implies \mathbf{E} \exp(\varepsilon |X|) \geq \mathbf{E} \frac{(\varepsilon |X|)^p}{p!}.$$

So for some $\widetilde{L}$,

$$\sup_p \frac{\varepsilon^p}{p!}\mathbf{E}|X|^p < \widetilde{L}.$$

Apply Stirlings do some algebra...

$$(\mathbf{E}|X|^p)^{1/p} \leq L'p\ldots$$

Conversely, suppose for some $1 \leq L < \infty$,

$$\lim\sup_p \frac{(\mathbf{E}|X|^p)^{1/p}}{p} < L.$$

Take small $\varepsilon > 0$, then

$$\sum_i \frac{(\varepsilon|X|^i)}{i!} \uparrow \exp(\varepsilon|X|).$$

Note that for all $i \geq 1$, by Sterlings,

$$\mathbf{E}\frac{(\varepsilon|X|)^i}{i!} \leq \frac{\varepsilon^i}{i!}(Li)^i \leq (\frac{1}{2}(1+O(1)))^i.$$

Aply MCT do some algebra...

$$\mathbf{E}\exp(\varepsilon|X|) = \lim_{i\to\infty}$$

□

Definition: Let X have well defined finite mean $\mathbf{E}X$. Then define the *variance* of X as

$$\mathbf{Var}X = \mathbf{E}[(X - \mathbf{E}X)^2] = \mathbf{E}X^2 - (\mathbf{E}X)^2.$$

Standard deviation is $\sqrt{\mathbf{Var}X}$.

Definition (Covariance): Let X, Y random variables and assume $\mathbf{Var}X$, $\mathbf{Var}Y$ well defined and finite. Then the *covariance* is

$$\operatorname{cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)] = \mathbf{E}XY - \mathbf{E}X\mathbf{E}Y.$$

This is well defined since

$$\mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)] \leq \mathbf{E}[(X - \mathbf{E}X)^2/2 + (Y - \mathbf{E}Y)^2/2] < \infty.$$

Definition (Correlation): We call X, Y *uncorrelated* if $\operatorname{cov}(X,Y) = 0$.

Definition (Independent events): Let $e_1, e_2, \ldots, e_n$ be events ($\in \Sigma$) on $(\Omega, \Sigma, \mathbf{P})$. The events are *independent* if for every subset $I \subset \{1, 2, \ldots, n\}$,

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

An arbitrary collection of events is independent if every finite subset is independent. A set is *k-wise independent* if any k-subset is independent.

Example (Pairwise independent but not independent): Let $\Omega = \{1, 2, 3, 4\}$ and $\Sigma = 2^\Omega$. Let $\mathbf{P}(\{1\}) = \cdots = \mathbf{P}(\{4\}) = \frac{1}{4}$. Then, the events

$$\{1, 2\}, \{2, 3\}, \{1, 3\}$$

are pairwise independent but not independent.

Definition (Independent fields): Suppose we have a collection of subfields of Σ, $(\Sigma_\alpha)_{\alpha \in A}$. (Σ) *is independent* if for all choices $e_\alpha \in \Sigma_\alpha$, $(e_\alpha)_{\alpha \in A}$ are independent.

Definition (Independent RVs): Let $(X_\alpha)_{\alpha \in A}$ be a collection of random variables/vectors on $(\Omega, \Sigma, \mathbf{P})$. (X_α) *is independent* if the σ-fields $(\sigma(X_\alpha))_{\alpha \in A}$ are mutually independent, where

$$\sigma(X_\alpha) = \{X_\alpha \in A, A \in \mathcal{B}\}.$$

1.8 2/2 - Independence

Proposition (Sigma algebras generated by independent π-systems are independent): If $P_1, \ldots, P_n$ are independent π-systems then $\sigma(P_1), \ldots, \sigma(P_n)$ are independent.

Proof. Fixing elements from $P_2, \ldots, P_n$, $F = A_2 \cap \cdots \cap A_n$, we can show that the set of elements A such that $\mathbf{P}(A \cap F) = \mathbf{P}(A)\,\mathbf{P}(F)$ is a λ-system, thus $\sigma(P_1), P_2, \ldots, P_n$ are independent. Repeat. □

Theorem (Condition for independence): $X_1, \ldots, X_n$ are independent if and only if

$$\mathbf{P}(X_1 \le t_1, \ldots, X_n \le t_n) = \prod_i \mathbf{P}(X_i \le t_i),$$

for all choices of $t_1, \ldots, t_n \in (-\infty, \infty]$.

Proof. If $X_1, \ldots, X_n$ are independent, the statement follows directly from definition of independent RVs. The other direction follows directly from the previous proposition. □

Theorem (Change of variables for product RVs): Suppose X, Y are independent and have distributions μ, ν. Suppose $h : \mathbb{R}^2 \to \mathbb{R}$ is measurable with either $h \ge 0$ or $\mathbf{E}|h(X, Y)| < \infty$. Then

$$\mathbf{E}h(X, Y) = \iint h(x, y)\, d\mu\, d\nu.$$

Proof. Let $\widetilde{X}, \widetilde{Y}$ be coordinate projections on $(\mathbb{R}^2, \mathcal{B}_2, \mu \times \nu)$ (we can put them under the product measure by independence). Then

$$\mathbf{E}h(X,Y) = \mathbf{E}h(\widetilde{X}, \widetilde{Y}) = \int_{\mathbb{R}^2} h(\widetilde{X}, \widetilde{Y}) d(\mu \times \nu).$$

By Fubini/Tonelli (as applicable by assumption), this expands to

$$\iint h(x,y)\, d\mu\, d\nu.$$

□

Corollary: If X, Y are independent random variables on $(\Omega, \Sigma, \mathbf{P})$ and $\mathbf{E}|XY| < \infty$, then

$$\mathbf{E}(XY) = \mathbf{E}X \cdot \mathbf{E}Y.$$

Remark: These results extend to n random variables or random vectors.

Remark: If $X_1 \dots X_m$ are independent and $(I_k)_{k \in \ell}$ is a partition of the index set $\{1, \dots, m\}$, and h_k is a measurable function, then

$$\{h_k((X_j)_{j \in I_k})\}_{k=1}^{\ell}$$

are independent.

1.9 2/5 - Khintchine's inequality, elementary conditioning

Theorem (Khintchine's inequality): Let $X_1, \dots, X_n$ be independent Rademacher variables, i.e.

$$\mathbf{P}(X_i = 1) = \mathbf{P}(X_i = -1) = \frac{1}{2}.$$

Let $(a_1, \dots, a_n)$ be a vector of unit length. Then for $t > 0$,

$$\mathbf{P}\left(\left|\sum_i a_i X_i\right| > t\right) \leq 2\exp\bigl(-t^2/2\bigr),$$

Proof. Exponential Markov:

$$\begin{aligned} \mathbf{P}\left(\sum_i a_i X_i > t\right) &\leq \frac{\mathbf{E}\exp(\lambda \sum_i a_i X_i)}{\exp(\lambda t)} \\ &= \frac{\prod_i \mathbf{E}\exp(\lambda a_i X_i)}{\exp(\lambda t)} \end{aligned}$$

Expand the expectation as

$$\begin{aligned} \mathbf{E}\exp(\lambda a_i X_i) &= \frac{1}{2}(\exp(\lambda a_i) + \exp(-\lambda a_i)) \\ &= \frac{1}{2}(2 + 2 \cdot \frac{\lambda^2 a_i^2}{2!} + 2 \cdot \frac{\lambda^4 a_i^4}{4!} + \dots) \\ &\leq \exp\left(\frac{1}{2}\lambda^2 a_i^2\right). \end{aligned}$$

Now,

$$\begin{aligned}\frac{\prod_i \mathbf{E}\exp(\lambda a_i X_i)}{\exp(\lambda t)} &\leq \frac{\prod_i \exp\left(\frac{1}{2}\lambda^2 a_i^2\right)}{\exp(\lambda t)} \\ &= \exp\left(\frac{1}{2}\lambda^2 - \lambda t\right).\end{aligned}$$

Optimizing over λ, we get the desired bound $\exp\left(-\frac{t^2}{2}\right)$. □

Definition (Conditional probability): Let $(\Omega, \Sigma, \mathbf{P})$ and events $e_1, e_2 \in \Sigma$. Assume $\mathbf{P}(e_2) > 0$. The conditional probability of e_1 given e_2 is

$$\mathbf{P}(e_1 \mid e_2) = \frac{\mathbf{P}(e_1 \cap e_2)}{\mathbf{P}(e_2)}.$$

When e_1, e_2 are independent we must have $\mathbf{P}(e_1 \mid e_2) = \mathbf{P}(e_1)$.

Proposition (Conditioning as new probability space): We can treat conditioning as redefining the probability space:

$$\widetilde{\Omega} = e_2, \widetilde{\Sigma} = \Sigma \cap e_2, \widetilde{\mathbf{P}}(.) = \mathbf{P}(. \mid e_2).$$

If X is a random variable we can define the conditional distribution given e_2 as

$$\widetilde{F}(t) = \mathbf{P}(X \leq t \mid e_2).$$

This is equivalent to the distribution function of $\widetilde{X}$ (restriction of X to $\widetilde{\Omega}$) on $(\widetilde{\Omega}, \widetilde{\Sigma}, \widetilde{\mathbf{P}})$.

Definition (Conditional independence): If $e_1, e_2, \ldots, e_n$ are events on $(\Omega, \Sigma, \mathbf{P})$, $e \in \Sigma, \mathbf{P}(e) > 0$, then, $e_1, \ldots, e_n$ are *independent given* e if for all $I \subset [n]$,

$$\mathbf{P}\left(\bigcap_{i \in I} e_i \mid e\right) = \prod_{i \in I} \mathbf{P}(e_i \mid e).$$

Example: Let X, Y independent on $(\Omega, \Sigma, \mathbf{P})$ and $h : \mathbb{R}^2 \to \mathbb{R}$. Let $A, B \in \mathcal{B}_{\mathbb{R}}$. Then

$$\mathbf{P}(X \in A, h(X, Y) \in B) \leq \mathbf{P}(X \in A) \sup_{z \in A} \mathbf{P}(h(z, Y) \in B).$$

Proof. WLOG $\mathbf{P}(X \in A) > 0$. Then, the inequality is equivalent to

$$\begin{aligned}&\mathbf{P}(h(X, Y) \in B \mid X \in A) \leq \sup_{z \in A} \mathbf{P}(h(z, Y) \in B) \\ \implies\ &\widetilde{\mathbf{P}}(h(\widetilde{X}, \widetilde{Y}) \in B) \leq \sup_{z \in A} \widetilde{\mathbf{P}}(h(z, \widetilde{Y}) \in B),\end{aligned}$$

where $\widetilde{X}, \widetilde{Y}$ are restrictions of X, Y to $\{x \in A\}$. Let $(\widetilde{\Omega}, \widetilde{\Sigma}, \widetilde{\mathbf{P}})$ be the product space where $\widetilde{X}, \widetilde{Y}$ are coordinate projections.

$$\begin{aligned}\widetilde{\mathbf{P}}(h(\widetilde{X}, \widetilde{Y}) \in B) &= \int_{\widetilde{\Omega}} \chi_B(h(\omega_1, \omega_2)) d\,\mathbf{P} \\ &= \int_{\omega_1} \int_{\omega_2} \chi_B(h(\omega_1, \omega_2)) d\,\mathbf{P} \\ &\leq \sup_{\omega_1} \int_{\omega_2} \chi_B(h(\omega_1, \omega_2)) d\,\mathbf{P} \\ &= \sup_{\omega_1} \mathbf{P}(h(\omega_1, \widetilde{Y}) \in B).\end{aligned}$$

□

Example: The inequality

$$\mathbf{P}(h(X, Y) \in B \mid X \in A) \leq \sup_{z \in A} \mathbf{P}(h(z, Y) \in B)$$

fails if $X = Y$ and $\mathbf{P}(X = 1) = \mathbf{P}(X = 0) = \frac{1}{2}$, $h^{-1}(B) = \{(0,0), (1,1)\}$.

2 Convergence

2.1 2/7 - Convolution, weak convergence

Definition (Convolution): Let F, G be distribution functions. Then

$$(F * G)(t) = \int_{-\infty}^{\infty} F(t-s)\,\mathrm{d}\mu_G(s).$$

Proposition:

(a) Convolution is communative.

(b) If F, G have densities p_F, p_G, then $F * G$ has density

$$p_{F*G}(t) = p_F(t) * p_G(t) = \int_{-\infty}^{\infty} p_F(t-s)p_G(s)\,\mathrm{d}s.$$

(c) If X, Y are independent then

$$F_{X+Y} = F_X * F_Y.$$

Proof.

(a) easy

(b) easy

(c) By change of variables,

$$\begin{aligned} F_{X+Y}(t) &= \int_{\mathbb{R}^2} \chi_{\{\omega_1+\omega_2 \le t\}} d\mu_X(\omega_1) d\mu_Y(\omega_2) \\ &= \int_{\omega_1} \left(\int_{-\infty}^{t-\omega_1} \mathrm{d}\mu_Y(\omega_2) \right) d\mu_X(\omega_1) \\ &= \int_{\omega_1} F_Y(t-\omega_1) d\mu_X(\omega_1) \\ &= (F_X * F_Y)(t). \end{aligned}$$

□

Definition (Weak convergence of measures): Let $(\mu_n), \mu$ be probability measures in $\mathbb{R}^m$. Then μ_n *weakly converges* to μ if for every bounded continuous $h : \mathbb{R}^m \to \mathbb{R}$,

$$\int h d\mu_n \to \int h d\mu.$$

Theorem (Portmanteau): The following are equivalent:

(a) $\mu_n \rightharpoonup \mu$.

(b) $\int h d\mu_n \to \int h d\mu$ for every bounded Lipschitz function.

(c) $\int h d\mu_n \to \int h d\mu$ for every compactly supported infinitely differentiable function.

(d) $\mu_n(S) \to \mu(S)$ for every continuity set S (boundary has measure 0) of μ.

(e) $\limsup_n \mu_n(F) \leq \mu(F)$ for every closed subset $F \subset \mathbb{R}^m$.

(f) $\liminf_n \mu_n(U) \geq \mu(U)$ for every open subset $U \subset \mathbb{R}^m$.

Definition (Convergence in distribution): Let $(X_n), X$ be random vectors in $\mathbb{R}^m$. Then X_n converges to X *in distribution*, notated $X_n \xrightarrow{\text{d}} X$, if

$$\mu_{X_n} \rightharpoonup \mu_X.$$

We notate $X_n \to_d X$. These do not have to be defined on a common probability space. Equivalently, $X_n \xrightarrow{\text{d}} X$ if for all continuous bounded h,

$$\mathbf{E}h(X_n) \to \mathbf{E}h(X).$$

2.2 2/9 - Weak convergence cont.

Proposition: $X_n \to_d X$ if and only if $F_{X_n}(t) \to F_X(t)$ for every point t where F is continuous.

Proof. (One dimensional case). Let $X_n \to_d X$, i.e. $\mu_{X_n}(S) \to \mu_X(S)$ for every continuity set S. Then

$$\mu_{X_n}((-\infty, t]) \to \mu_X((-\infty, t]) \implies F_{X_n}(t) \to F_X(t),$$

whenever F_X is continuous at t (so that the interval is a continuity set).

Conversely, assume $F_{X_n}(t) \to F_X(t)$ whenever F_X is continuous at t. Pick an open interval (a, b). We want to show that

$$\liminf_n \mu_{X_n}((a, b)) \geq \mu_X((a, b)).$$

Note that by continuity of measure, for all $\varepsilon > 0$ there is $\delta > 0$ such that

$$\mu_X((a + \delta, b - \delta)) \geq \mu_X((a, b)) - \varepsilon.$$

Then since there are at most countably many discontinuities of F_X, we can choose continuous points $t_1 \in (a, a + \delta)$ and $t_2 \in (b - \delta, b)$. So

$$F_{X_n}(t_1) \to F_X(t_1), F_{X_n}(t_2) \to F_X(t_2) \implies \mu_{X_n}((t_1, t_2]) \to \mu_X((t_1, t_2]).$$

Taking $\varepsilon \to 0$ completes the desired statement. Since all open sets in $\mathbb{R}$ can be represented as a countable disjoint union of intervals, we can easily extend this to all open sets. □

Proposition (Convergence in distribution for discrete RVs): Suppose X_n, X take integer values. Then $X_n \xrightarrow{\text{d}} X$ if and only if $\mathbf{P}(X_n = i) \to \mathbf{P}(X = i)$ for all $i \in \mathbb{Z}$.

Proof. Suppose $X_n \xrightarrow{\text{d}} X$. Then $F_{X_n} \to F_X$ on all non integral values, so

$$\mathbf{P}(X_n = i) = F_{X_n}(i+\varepsilon) - F_{X_n}(i-\varepsilon) \to F_X(i+\varepsilon) - F_X(i-\varepsilon) = \mathbf{P}(X = i).$$

□

Example: Let $(\Omega, \Sigma, \mathbf{P}) = ([0,1], \mathcal{B}_{[0,1]}, \mathcal{L})$ and X be the identity (uniform on $[0,1]$). Can we construct a random variable Y uniform on $[0,1]$ and independent from X?

Proof. The answer is no. Suppose it were true. Note that $\sigma(X) = \mathcal{B}$ so $\sigma(Y) \subset \sigma(X)$. Then, for $A \in \sigma(Y)$,

$$\mathcal{L}(A \cap A) = \mathcal{L}(A)\mathcal{L}(A),$$

which means $\mathcal{L}(A) = 0$ or $\mathcal{L}(A) = 1$. □

Example:

(a) $X \sim B(p)$ if X is an indicator of an event with probability p

(b) $Y \sim Bino(n,p)$ if $\mathbf{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$. $\mu Y = np, \mathbf{Var} Y = n(p - p^2)$.

2.3 2/12 - CLT type stuff

Example: Let $Y_n \sim \text{Bin}(n, 1/2)$, so $\sum_i^n b_i \sim_d Y_n$, where $b_i \sim \text{Bin}(1/2)$. Let

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp(-s^2/2)\, \mathrm{d}s.$$

Then

$$\mathbf{P}\left(\frac{y - n/2}{\sqrt{n/4}} \le t \right) \to \Phi(t).$$

Proof. First,

$$\begin{aligned} \mathbf{P}(Y_n \le n/2 + t\sqrt{n/4}) &= 2^{-n} \sum_{m \le n/2 + t\sqrt{n/4}} \binom{n}{m} \\ &= 2^{-n}(1 + \pm O(1)) \end{aligned}$$

We use the approximation

$$\begin{aligned}\binom{n}{m} &= (1 \pm O(1)) \frac{\sqrt{2\pi n}(n/e)^n}{\sqrt{2\pi m}(m/e)^m \sqrt{2\pi(n-m)}((n-m)/e)^{n-m}} \\ &= (1 \pm O(1)) \frac{\sqrt{n}}{\sqrt{2\pi}(n/2)}(n/m)\end{aligned}$$

□

Example: Let $X_n \to_d X$ and Y independent from $X_1, X_2, \ldots, X$. Prove that $X_n + Y \to_d X + Y$.

Proof. Let h be a Lipschitz bounded function. We want to show

$$\mathbf{E}h(X_n + Y) \to \mathbf{E}h(X + Y)$$

Let μ_n, ν be distributions of X_n, Y. By change of variables,

$$\begin{aligned}\mathbf{E}h(X_n + Y) &= \iint h(\omega_1 + \omega_2) d\mu_n(\omega_1) d\nu(\omega_2) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(\omega_1 + \omega_2) d\mu_n(\omega_1) \right) d\nu(\omega_2) \\ &\to \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(\omega_1 + \omega_2) d\mu(\omega_1) \right) d\nu(\omega_2) \\ &= \mathbf{E}h(X + Y),\end{aligned}$$

where the convergence follows from DCT and assumption on h. □

2.4 2/14 - Examples

Proposition (Condition for independence): Suppose $X_1, \ldots, X_n$ are random variables. Then the following are equivalent:

(a) $X_1, \ldots, X_n$ independent

(b) For all $1 \leq k \leq n-1$, the vector $(X_1, \ldots, X_k)$ is independent from X_{k+1}.

Proof. We show (b) implies (a) by induction. Suppose $X_1, \ldots, X_k$ are mutually independent and $(X_1, \ldots X_k)$ independent from X_{k+1}. Suppose $A_1, A_2, \ldots, A_{k+1} \in \mathcal{B}$. Then

$$\begin{aligned}\mathbf{P}(X_i \in A_i \forall i \leq k+1) &= \mathbf{P}((X_1, \ldots, X_k) \in \prod_i A_i, X_{k+1} \in A_{k+1}) \\ &= \mathbf{P}((X_1, \ldots, X_k) \in \prod_i A_i) \, \mathbf{P}(X_{k+1} \in A_{k+1}) \\ &= \prod_i \mathbf{P}(X_i \in A_i).\end{aligned}$$

Conversely, suppose $X_1, \ldots, X_n$ are independent. Fix $1 \leq k \leq n-1$. Let

$$P = \left\{ \prod_i^k (-\infty, a_i] : a_i \in \mathbb{R} \right\}.$$

This is a π-system that generates $\mathcal{B}_k$. For $A \in P$, $B \in \mathcal{B}$,

$$\begin{aligned} \mathbf{P}((X_i, \ldots, X_k) \in A, X_{k+1} \in B) &= \prod_i^k \mathbf{P}(X_i \in (-\infty, a_i])\, \mathbf{P}(X_{k+1} \in B) \\ &= \mathbf{P}((X_1, \ldots, X_k) \in A)\, \mathbf{P}(X_{k+1} \in B). \end{aligned}$$

A being a rectangle allows us to split up vector into its components. Then we can apply independence of $X_1, \ldots, X_{k+1}$ to pull out the X_{k+1}. Now we construct a lambda system L containing P:

$$L = \{A \in \mathcal{B}_k : \text{above events are independent}\}.$$

We check that $\varnothing$ is a limit of sets in P so $\varnothing \in L$. When $A \in L$, $A^c \in L$ by writing $\mathbf{P}((X_1, \ldots, X_k) \in A^c, X_{k+1} \in B)$ as

$$\mathbf{P}(X_{k+1} \in B) - \mathbf{P}((X_1, \ldots, X_k) \in A, X_{k+1} \in B).$$

Finally, for disjoint $A_i \in L$, we may write $\mathbf{P}((X_1, \ldots, X_k) \in \bigcup_i A_i, X_{k+1} \in B)$ as

$$\sum_i \mathbf{P}((X_1, \ldots, X_k) \in A_i, X_{k+1} \in B).$$

It follows easily that L is a λ system and so $\sigma(P) = \mathcal{B} \subset L$ as required. □

Example: Let X RV and assume for $a_3, \ldots, a_k \in \mathbb{R}, k \geq 3$,

$$M_X(t) = \exp\left(\sum_{i=3}^k a_i t^i \right).$$

Show $\mathbf{P}(X = 0) = 1$.

Proof.

$$\begin{aligned} M_X(t) &= \mathbf{E} \exp(tX) \\ &= 1 + t\mathbf{E}X + \frac{t^2}{2}\mathbf{E}X^2 + \ldots \\ &= 1 + \sum_{i=3}^k a_i t^i + \frac{1}{2}\left(\sum_{i=3}^k a_i t^i \right)^2 + \ldots \end{aligned}$$

In particular we must have $\mathbf{E}X = 0$ and $\mathbf{E}X^2 = 0$, meaning $\mathbf{P}(X = 0) = 1$. □

2.5 2/19 - Gaussian distribution

Definition (Gaussian): A *Gaussian* random variable of mean μ and variance σ^2 is defined with its density

$$p(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \frac{(t-\mu)^2}{\sigma^2} \right).$$

This is a valid density that integrates to 1 by the two dimensional integral argument. A *standard* Gaussian has $\mu = 0, \sigma^2 = 1$, notated $N(0,1)$. We notate the standard Gaussian distribution function as Φ.

Remark: Recall that when $b_1, b_2, \ldots$ are mutually independent Ber(1/2) random variables, then

$$\frac{\sum_i b_i - \frac{n}{2}}{\sqrt{n/4}} \to_d N(0,1).$$

Proposition: If g_1, g_2 are *independent* Gaussians then $g_1 + g_2$ is Gaussian.

Proof 1. Use the above remark to write the sum of Bernoulli's in two ways, one way getting you $N(0,1) + N(0,1)$ and the other getting you $\sqrt{2}N(0,1)$. □

Proof 2. Consider the convolution:

$$\begin{aligned} p_{g_1+g_2}(t) &= (p_{g_1} * p_{g_2})(t) \\ &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma_1} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2} \frac{(t-s-\mu_1)^2}{\sigma_1^2} \right) \exp\left(-\frac{1}{2} \frac{(s-\mu_2)^2}{\sigma_2^2} \right) ds \end{aligned}$$

□

Remark: This proposition says that the Gaussian distribution is *stationary*, meaning linear combinations of independent draws end up in the same distribution.

2.6 2/21 - Poisson, more notions of convergence

As motivation consider the calls a support line receives over a fixed time frame. We can model this as the sum of the probabilities each customer calls in. Suppose further that we have infinitely many customers each with infinitely small probabilities of calling in, but we expect λ total calls across all customers. Then we say the number of calls is modeled by the Poisson distribution.

Definition (Poisson distribution): Let $\lambda > 0$. Define the *Poisson distribution* with parameter λ by the probability mass function

$$f_\lambda(m) = \frac{\lambda^m}{m!} \exp(-\lambda), m \in \mathbb{N}.$$

Theorem (Law of rare events): Fix n. Let $X_{n,1}, \ldots, X_{n,n}$ be independent Bernoullis with $P(X_{n,m} = 1) = p_{n,m}$. Suppose

(a) $\lim_{n\to\infty} \sum_{i=1}^n p_{n,i} = \lambda \in (0, \infty)$.

(b) $\lim_{n\to\infty} \max_{1\le i\le n} p_{n,i} = 0$.

If $S_n = X_{n,1} + \cdots + X_{n,n}$ then $S_n \overset{\mathrm{d}}{\to} Z$, where $Z \sim \text{Poisson}(\lambda)$.

Proof. (When $p_{n,j} = \lambda/n$). In this case $S_n \sim \text{Bin}(n, \lambda/n)$. As $n \to \infty$,

$$\begin{aligned}\mathbf{P}(S_n = k) &= \binom{n}{k}(\lambda/n)^k(1-\lambda/n)^{n-k} \\ &\sim \binom{n}{k}(\lambda/n)^k e^{-\lambda/n(n-k)} \\ &\sim \frac{n!}{k!(n-k)!}\frac{\lambda^k}{n^k}e^{-\lambda} \\ &= \frac{\lambda^k}{k!}e^{-\lambda}.\end{aligned}$$

□

Proposition (Poisson is stationary): Let X, Y be two independent Poisson RVs with parameters λ_1 and λ_2. Then $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Proof.

$$\begin{aligned}\mathbf{P}(X+Y=k) &= \sum_{i=0}^{k}\mathbf{P}(X=i)\,\mathbf{P}(Y=k-i) \\ &= \sum_{i=0}^{k}\frac{\lambda_1^i}{i!}e^{-\lambda_1}\cdot\frac{\lambda_2^{k-i}}{(k-i)!}e^{-\lambda_2} \\ &= e^{-(\lambda_1+\lambda_2)}\sum_{i=0}^{k}\frac{\lambda_1^i\lambda_2^{k-i}}{i!(k-i)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{i=0}^{k}\frac{k!}{i!(k-i)!}\lambda_1^i\lambda_2^{k-i} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{i=0}^{k}\binom{k}{i}\lambda_1^i\lambda_2^{k-i} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}(\lambda_1+\lambda_2)^k.\end{aligned}$$

□

Definition (Convergence almost surely): $X_n \xrightarrow{\text{a.s.}} X$ if $\mathbf{P}(\lim_{n\to\infty} X_n = X) = 1$, i.e. $X_n(\omega)$ does not converge to $X(\omega)$ on at most a set of measure 0.

Definition (Convergence in probability): $X_n \xrightarrow{\text{p}} X$ if for all $\varepsilon > 0$, $\mathbf{P}(|X_n - X| \geq \varepsilon) \to 0$.

Proposition:

(a) $X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\text{p}} X$.

(b) $X_n \xrightarrow{\text{p}} X \implies X_n \xrightarrow{\text{d}} X$.

(c) $X_n \xrightarrow{\text{d}} c \implies X_n \xrightarrow{\text{p}} c$ for constant $c \in \mathbb{R}$.

Proof.

(a) The set version of Fatou's lemma tells us

$$\mathbf{P}(\liminf_n A_n) \le \liminf_n \mathbf{P}(A_n) \implies \mathbf{P}(\limsup_n A_n) \ge \limsup_n \mathbf{P}(A_n).$$

Then $\mathbf{P}(\limsup_n \{|X_n - X| > \varepsilon\}) \ge \limsup_n \mathbf{P}(\{|X_n - X| > \varepsilon\})$.

(b) Let $t \in \mathbb{R}$ such that F_X is continuous at t. Fix $\varepsilon > 0$. We note that

$$\{X_n \le t\} \subset \{X \le t + \varepsilon\} \cup \{|X - X_n| > \varepsilon\}.$$

Thus

$$\begin{aligned}\mathbf{P}(X_n \le t) &\le \mathbf{P}(X \le t + \varepsilon) + \mathbf{P}(|X - X_n| > \varepsilon) \\ &\le \mathbf{P}(X \le t + \varepsilon) + \varepsilon && (n \ge n_0(\varepsilon))\end{aligned}$$

Choose ε by right continuity F_X. On the other hand, we can similarly show with continuity that for all $\delta > 0$ and large n:

$$F_{X_n}(t) \ge F_X(t) + \delta.$$

Alternatively, for continuous and bounded f, $X_n \xrightarrow{\text{p}} X$ implies $f(X_n) \xrightarrow{\text{p}} f(X)$, so

$$\mathbf{E}|f(X_n) - f(X)| \to 0 \implies \mathbf{E}f(X_n) \to \mathbf{E}f(X).$$

(c) Since $F_{X_n}(t) \to F_X(t)$ for all $t \ne c$,

$$\mathbf{P}(|X_n - c| > \varepsilon) \le F_{X_n}(c - \varepsilon) + (1 - F_{X_n}(c + \varepsilon)) \implies \mathbf{P}(|X_n - c| > \varepsilon) \to 0.$$

□

Example ($\xrightarrow{\text{p}}$ but not $\xrightarrow{\text{a.s.}}$): (2.3.11) Consider independent bernoulli RV $b_n \sim \text{Ber}(1/n)$. Then $b_n \xrightarrow{\text{p}} 0$ but not a.s. by BC2.

Theorem (Bernoulli SLLN): Let $b_1, b_2, \dots$ iid $\text{Ber}(p)$.

$$\frac{1}{n}\sum_{i=1}^n b_i \xrightarrow{\text{a.s.}} p.$$

This is the basis for frequentist interpretation of probability.

Proof. Fix $\varepsilon > 0$. We show

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n b_i \ge p + \varepsilon\right) \to 0 \text{ and } \mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n b_i \le p - \varepsilon\right) \to 0.$$

$$\begin{aligned}\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} b_i \geq p+\varepsilon\right) &= \sum_{m\geq pn+\varepsilon n}\binom{n}{m}p^m(1-p)^{n-m}\\ &= (1+o(1))\left(\frac{n}{n-m}\right)^{n-m}\left(\frac{n}{m}\right)^m p^m(1-p)^{n-m}.\end{aligned}$$

We can prove this is exponentially small in n, i.e. for $\delta > 0$ only depending on p and ε, the above is bounded by

$$\exp(-\delta n).$$

□

2.7 2/26 - Borel Cantelli lemmas

Definition (Infinitely often and eventually sets): Define (A_n, infinitely often) as the set of ω that show up in every tail:

$$(A_n, \text{i. o.}) \coloneqq \limsup A_n \coloneqq \bigcap_n \bigcup_{m\geq n} A_n.$$

Define (A_n, eventually) as the set of ω such that there is n_0 where $\omega \in A_n$ for $n \geq n_0$:

$$(A_n, \text{ev}) \coloneqq \liminf A_n \coloneqq \bigcup_n \bigcap_{m\geq n} A_n.$$

Theorem (Borel-Cantelli lemma 1): Let A_n be a sequence of events on a common probability space with $\sum_i \mathbf{P}(A_i) < \infty$. Then

$$\mathbf{P}(A_n, \text{i. o.}) = 0.$$

Proof. Using $E = \bigcap_i \bigcup_{j\geq i} A_j$, we see that for each i,

$$\mathbf{P}(E) \leq \sum_{j\geq i}\mathbf{P}(A_j),$$

so by assumption $\sum_i \mathbf{P}(A_i) < \infty$, $\mathbf{P}(E) = 0$.

Alternatively, let $N = \sum_i \mathbb{1}_{A_i}$ be the number of events that occur. By MCT,

$$\mathbf{E}\sum_i \mathbb{1}_{A_i} = \sum_i \mathbf{P}(A_i) < \infty.$$

So $N < \infty$ almost everywhere as required. □

Remark: The converse is false. Consider $((0,1), \mathcal{B}, \mathcal{L})$ and events $A_n = (0, 1/n)$.

Example: Consider a game where on the nth round you have a $\frac{1}{2^n+1}$ chance of losing 2^n dollars and a $\frac{2^n}{2^n+1}$ chance of gaining a dollar. Even though the expected value of each round is 0, the probability we lose infinitely many times by BC1 is 0, meaning we can profit.

I wonder what happens in practice. someday python it?

Theorem (Borel-Cantelli lemma 2): Suppose A_n are independent and $\sum_i \mathbf{P}(A_i) = \infty$. Then

$$\mathbf{P}(A_n, \mathrm{i.\,o.}) = 1.$$

Proof 1. Fix i. It suffices to show $\mathbf{P}(\cup_{j\geq i} A_j) = 1$.

$$\mathbf{P}((\cup_{j\geq i} A_i)^c) = \prod_{j\geq i}(1 - \mathbf{P}(A_j)) \leq \exp\left(-\sum_{j\geq i}\mathbf{P}(A_j)\right) = 0.$$

□

Proof 2. Let

$$A_n = \left\{\left|\frac{1}{n}\sum_i^n b_i - p\right| > \varepsilon\right\}.$$

We showed that for fixed $\varepsilon > 0$, $\sum_i \mathbf{P}(A_i) < \infty$. So by BC1, $\mathbf{P}(A_n, \mathrm{i.\,o.}) = 0$. □

Example: The probability we will find any arbitrarily large number of heads in a row within infinite coin flips is 1 by BC2.

2.8 2/28 - Coupon collector's problem

Example (Coupon collector's problem): For all $n \in \mathbb{N}$, define a sequence of random variables $(X_{i,n})$ i.i.d. uniform on $\{1, 2, \ldots, n\}$. Consider the minimum time to "collect all coupons",

$$T_n \coloneqq \min\{m \geq 1 : \forall k \in [n], \exists i \leq m, X_{i,n} = k\}.$$

Show $\mathbf{E}T_n \sim n \log n$ and $\frac{T_n}{n \log n} \xrightarrow{\mathrm{p}} 1$.

Lemma (Waiting times are independent): Define the minimum time to collect k distinct coupons,

$$T_k = \inf\left\{m \geq 1 : \left|\cup_{j=1}^m \{X_{j,n}\}\right| = k\right\}.$$

Let $T_0 = 0$. We claim that the differences $(T_j - T_{j-1})$ for $j = 1, \ldots, n$ are mutually independent. By earlier proposition it is sufficient to show $T_j - T_{j-1}$ is independent from $(T_1 - T_0, \ldots, T_{j-1} - T_{j-2})$. This is the same as showing independence from $(T_1, \ldots, T_{j-1})$, for which it suffices to show that conditioned on the event

$$e = \{T_1 = t_1, \ldots, T_{j-1} = t_{j-1}\}$$

for fixed $t_1, \ldots, t_{j-1}$, the distribution of $T_j - T_{j-1}$ remains geometric with parameter $\frac{n-j+1}{n}$ (this is more or less obvious, but we prove it for pedagogical value).

We partition e by the order in which we collect the coupons, $a = a_1, \ldots, a_n$. Conditioned on e_a for any a, the distribution of $T_j - T_{j-1}$ remains the same. So conditioned on e it also remains the same, and we are done.

Proof. Since we have shown $T_j - T_{j-1}$ is geometric with parameter $\frac{n-j+1}{n}$, with expected value $\frac{n}{n-j+1}$ and variance bounded by $(\frac{n}{n-j+1})^2$, we compute $\mathbf{E}T_n$ as

$$\mathbf{E}T_n = \sum_{j=1}^{n} \frac{n}{n-j+1} = n\sum_{k=1}^{n}\frac{1}{k} \sim n\log n.$$

We similarly compute $\mathbf{Var}T_n$ as

$$\mathbf{Var}T_n = \sum_{j=1}^{n}\left(\frac{n}{n-j+1}\right)^2 = n^2\sum_{k=1}^{n}\frac{1}{k^2} = O(n^2).$$

So

$$\mathbf{Var}\frac{T_n}{n\log n} = o(1),$$

meaning $\frac{T_n}{n\log n} \xrightarrow{\text{p}} 1$:

$$\mathbf{P}\left(\left|\frac{T_n}{n\log n} - 1\right| > \varepsilon\right) \le \frac{o(1)}{\varepsilon^2} \to 0.$$

□

2.9 3/1 - WLLN and more convergence theorems

Theorem: Let W_n iid bounded by some $L > 0$, i.e.

$$\mathbf{P}(|W_i| > L) = 0.$$

Then

$$\frac{1}{n}\sum_{i=1}^{n} W_i \xrightarrow{\text{a.s.}} \mathbf{E}W_1.$$

Proof. WLOG recenter to $\mathbf{E}W_i = 0$. Consider expanding the fourth moment,

$$\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} W_i\right)^4.$$

When there is an index that appears only once, the expectation of the product is 0. Otherwise either all indices are the same or we have two pairs. So

$$\begin{aligned}\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} W_i\right)^4 &\le \frac{1}{n^4}\left(\sum_{i=1}^{n}\mathbf{E}W_i^4 + 12\sum_{i,j}^{n}\mathbf{E}W_i^2W_j^2\right)\\ &\le \frac{1}{n^4}(nL^4 + 12n^2L^4)\\ &= O\left(\frac{1}{n^2}\right).\end{aligned}$$

Thus by Markov

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} W_i\right|^4 > \varepsilon^4\right) \le O(n^{-2})/\varepsilon^4 = O(n^{-2}).$$

Define $A_n = \left\{\left|\frac{1}{n}\sum_{i=1}^{n} W_i\right|^4 > \varepsilon^4\right\}$, then $\sum_n \mathbf{P}(A_n) < \infty$ so by Borel Cantelli for large enough n the quantity will be within ε^4. Since this holds for all ε,

$$\mathbf{P}\left(\lim_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} W_i\right| = 0\right) = 1 \implies \frac{1}{n}\sum_{i=1}^{n} W_i \xrightarrow{\text{a.s.}} 0.$$

□

Theorem (Weak law of large numbers): Suppose X_n are iid with finite mean μ. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{p}} \mu.$$

Proof. For $M > 0$ define $y_i = X_i \mathbb{1}_{\{|X_i| \le M\}}$ and $z_i = X_i \mathbb{1}_{\{|X_i| > M\}}$. By Markov,

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i\right| > \varepsilon\right) \le \frac{\mathbf{E}|z_1|}{\varepsilon}.$$

Consider z_1 as parameterized by M. Define $M_0(\varepsilon^2)$ such that for $M \ge M_0$, $\mathbf{E}|z_1| \le \varepsilon^2$. By Markov,

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i\right| \ge \varepsilon\right) \le \varepsilon.$$

For all sufficiently large n,

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} y_i - \mathbf{E}y_1\right| > \varepsilon\right) \le \varepsilon \implies \mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} y_i - \mathbf{E}X_1\right| > 2\varepsilon\right) \le \varepsilon.(*)$$

So combining,

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X_1\right| > 3\varepsilon\right) \le 2\varepsilon.$$

□

Theorem: Suppose b_n be independent Bernoulli variables with $\mathbf{P}(b_i = 1) = p_i \in (0,1)$. Then

$$\frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} p_i} \xrightarrow{\text{p}} 1$$

if and only if $\sum_i p_i = \infty$.

Proof. We observe that the variance of the LHS converges to 0:

$$\mathbf{Var}\frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n p_i} = \frac{\sum_{i=1}^n p_i(1-p_i)}{(\sum_{i=1}^n p_i)^2} \le \frac{1}{\sum_{i=1}^n p_i} \to 0.$$

Therefore we can apply Markov-Chebyshev: for fixed ε,

$$\mathbf{P}\left(\left|\frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n p_i} - 1\right| \ge \varepsilon\right) \le \frac{o(1)}{\varepsilon} \to 0.$$

Now suppose the convergence holds. TODO □

2.10 3/11 - SLLN

Theorem (SLLN): Suppose X_n i.i.d. with finite mean. Then

$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\text{a.s.}} \mathbf{E}X_1.$$

Proof. Etemadi's proof was discussed in class. Omitted. □

2.11 3/13 - Almost sure convergence review

To show convergence almost surely, we have used:

(a) Moment method

(b) Truncation

(c) Partitioning into geometric increasing blocks

Theorem: Suppose b_i independent Bernoulli with parameters p_i Then

$$\frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n p_i} \xrightarrow{\text{a.s.}} 1$$

if and only if $\sum_i p_i = \infty$.

Proof. Since $b_i \in [0,1]$, $\mathbf{Var} b_i \le \mathbf{E} b_i^2 \le \mathbf{E} b_i$, so the convergence holds easily by Chebyshev in probability.

Fix $\alpha > 1$ and for $k \ge 1$ let n_k be the smallest index such that

$$\sum_{i=1}^{n_k} p_i \ge \alpha^k.$$

Fix $\varepsilon > 0$.

$$\begin{aligned}\mathbf{P}\left(\left|\sum_{i=1}^{n_k} b_i - \sum_{i=1}^{n_k} p_i\right| \geq \varepsilon \sum_{i=1}^{n_k} p_i\right) &\leq \frac{\mathbf{Var}\sum_{i=1}^{n_k} b_i}{\varepsilon^2 \left(\sum_{i=1}^{n_k} p_i\right)^2} \\ &= \frac{\sum_{i=1}^{n_k} p_i(1-p_i)}{\varepsilon^2 \left(\sum_{i=1}^{n_k} p_i\right)^2} \\ &\leq \frac{1}{\varepsilon^2}\alpha^{-k}.\end{aligned}$$

So by BC2 ... TODO □

Example: Let X_n, X defined on $(\Omega, \Sigma, \mathbf{P})$ with Ω countable and $\Sigma = 2^\Omega$. Suppose $X_n \xrightarrow{\text{p}} X$. Prove $X_n \xrightarrow{\text{a.s.}} X$.

Proof. It suffices to show that $X_n(\omega) = X(\omega)$ for all ω with $\mathbf{P}(\omega) > 0$. So fix such an ω. By convergence in probability for each $\varepsilon > 0$ we can choose n_0 such that for $n \geq n_0$,

$$\mathbf{P}(|X_n - X| \geq \varepsilon) < \mathbf{P}(\omega),$$

so $|X_n(\omega) - X(\omega)| < \varepsilon$. In particular $X_n(\omega) \to X(\omega)$ as required. □

Theorem: Let X RV and X_n independent variables equidistributed with X (infinite samples).

For every n, define EDF (empirical distribution function)

$$F_n(t) := \frac{1}{n}\#\{i \leq n : X_i \leq t\}.$$

Then if F is the distribution function of X,

$$\lim_{n\to\infty} \|F_n - F\|_\infty = 0, \text{a.s.}$$

This tells us that our convergence is uniform, so we can approximate our true distribution with arbitrary precision.

3 Central limit theorems

3.1 3/18 - Characteristic functions

Definition (Characteristic function): Let X be RV, then its *characteristic function* is

$$\phi_X(t) = \mathbf{E}\exp(itX).$$

We define expectation of complex numbers in the natural way:

$$\mathbf{E}Z = \mathbf{E}(\mathrm{Re}\ Z) + i\mathbf{E}(\mathrm{Im}\ Z).$$

Proposition:

(a) $\phi_X(0) = 1$, $|\phi_X(t)| \leq 1, t \in \mathbb{R}$, and $\phi_{aX+b}(t) = \exp(itb)\phi_X(at)$.

(b) ϕ_X is uniformly continuous.

(c) If $\mathbf{E}|X|^n < \infty$ for some n, then $\phi_X^{(n)}(t)$ exists, equals $i^n \mathbf{E}X^n \exp(itX)$, and is uniformly continuous. In particular, we can get the moments:

$$\mathbf{E}X^n = \frac{1}{i^n}\phi_X^{(n)}(0).$$

(d) (Main motivation) If X, Y are independent then $\phi_{X+Y} = \phi_X\phi_Y$.

Proof.

(a) Easy to verify.

(b) Fix $t, h \in \mathbb{R}$. Then

$$|\phi_X(t+h) - \phi_X(t)| = |\exp(itX)(\exp(ihX) - 1)| \leq |\exp(ihX) - 1| \to 0.$$

(c) Omitted.

(d) $\phi_{X+Y}(t) = \mathbf{E}\exp(it(X+Y)) = \mathbf{E}\exp(itX)\exp(itY) = \phi_X(t)\phi_Y(t)$.

□

Recall that $X_n \xrightarrow{\mathrm{d}} X$ if and only if $\mathbf{E}h(X_n) \to \mathbf{E}h(X)$ for *all* "nice functions" (ex. continuous bounded, Lipschitz, infinitely differentiable) h.

Theorem (Levy continuity theorem): Let X_n have associated characteristic functions ϕ_n.

(a) Suppose X_n converges in distribution to some RV X. Then $\phi_n \to \phi_X$ pointwise.

(b) Suppose ϕ_n converges pointwise to a function ϕ which is continuous at 0. Then ϕ is characteristic function of some RV X and $X_n \xrightarrow{\mathrm{d}} X$.

Proof.

(a) Fix t. Since $h(y) \coloneqq \exp(ity)$ is a continuous bounded function, by characterization of convergence in distribution we have

$$\mathbf{E}h(X_n) \to \mathbf{E}h(X) \implies \phi_{X_n}(t) \to \phi_X(t).$$

(b) Omitted.

□

Theorem (Inversion formula): Let $\phi(t) = \int \exp(itx) d\mu(x)$ where μ is a distribution. Then for $a < b$,

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{\exp(-ita) - \exp(-itb)}{it} \phi(t)\, dt = \mu(a, b) + \frac{1}{2}\mu(\{a, b\}).$$

Proof. Omitted. □

Corollary: Characteristic function uniquely determines distribution.

3.2 3/20 - Multivariate Gaussian

Definition (Multivariate Gaussian): A random vector $X = (X_1, \ldots, X_n)$ in $\mathbb{R}^n$ is said to have *Gaussian distribution* if every linear combination of components of X is Gaussian. A *standard Gaussian* in $\mathbb{R}^n$ has zero mean and identity covariance matrix.

Proposition: The covariance matrix of a zero mean random vector X is $\mathbf{E}XX^\top$.

Theorem (Characterization of multivariate Gaussians): Suppose X is Gaussian in $\mathbb{R}^m$ with mean μ. Then there is matrix A such that

$$X = AG + \mu,$$

where G is a standard Gaussian. So multivariate Gaussians are a parametric family.

Proof. WLOG let $\mathbf{E}X = 0$. Let Σ be the covariance matrix of X. We want to find A such that the covariance matrix of AG is Σ:

$$\mathbf{E}((AG)(AG)^\top) = A\mathbf{E}(GG^\top)A^\top = AA^\top.$$

So we can choose A as $\Sigma^{1/2}$ (defined by diagonalization). □

Proposition (Uniqueness): Let $\mu \in \mathbb{R}^n$ and Σ a positive semi-definite matrix. There is unique Gaussian distribution with mean μ and covariance Σ. Further, if $\det \Sigma > 0$ then the distribution has well defined density

$$p(t) = \frac{1}{(2\pi)^{n/2} \det \Sigma^{1/2}} \exp\left(-\frac{1}{2}(t-\mu)^\top \Sigma^{-1} (t-\mu)\right).$$

Example: Let $g \sim N(0,1)$ and r Rademacher independent from g. Let $X := (g, rg)$. This is not Gaussian since the linear combination of components $g + rg$ takes value 0 with probability 1/2.

Definition (Characteristic function of random vector): If X is random vector in $\mathbb{R}^n$, define

$$\phi_X(t) = \mathbf{E} \exp(it^\top X).$$

Proposition (Characteristic function for Gaussian): The characteristic function of a random vector $X \sim N(\mu, M)$ is

$$\phi_X(t) = \exp(it^\top \mu - t^\top Mt/2).$$

Proof. Omitted. □

3.3 3/22-3/27 - Proofs of CLT

Theorem (Multivariate CLT): Let X_n be sequence of i.i.d. random vectors in $\mathbb{R}^n$ with mean μ and covariance Σ. Then the sequence of random vectors

$$\frac{X_1 + \cdots + X_m - m\mu}{\sqrt{m}}$$

converges to a centered multivariate normal with covariance Σ.

Proof. We prove the theorem for $\mu = 0$, $\Sigma = I$, as the general result holds by linear transformation. First consider the one dimensional case. Let $Z_m = \frac{X_1+\cdots+X_m}{\sqrt{m}}$. Then, using first the product rule then Taylor's theorem around 0,

$$\begin{aligned}
\phi_{Z_m}(t) &= \phi_X(t/\sqrt{m})^m \\
&= \left(1 + \frac{t}{\sqrt{m}}\phi'_X(0) + \frac{t^2}{2m}\phi''_X(0) + o(1/m)\right)^m \\
&= \left(1 - \frac{t^2}{2m} + o(1/m)\right)^m && (\phi_X \text{ gives moments}) \\
&\to \exp(-t^2/2).
\end{aligned}$$

It follows that $\phi_{Z_m}(t)$ converges pointwise to the characteristic function of a standard Gaussian. Thus Z_m converges in distribution to Z, where $Z \sim N(0,1)$.

In the general case, fix $t \in \mathbb{R}^n$ and define $Y_m = t^\top X_m$. So Y_m are i.i.d. random variables with mean 0 and variance

$$\mathbf{E}(t^\top X)^2 = t^\top \mathbf{E}(Xt^\top X) = t^\top \mathbf{E}(XX^\top t) = t^\top t.$$

Apply the one dimensional CLT to Y_m to get

$$\frac{Y_1 + \cdots + Y_m}{\sqrt{m}} \xrightarrow{\mathrm{d}} Z,$$

where $Z \sim N(0, t^\top t)$. Then, by Levy continuity theorem, for $s \in \mathbb{R}$, letting $W_m = \frac{Y_1 + \cdots + Y_m}{\sqrt{m}}$,

$$\phi_{W_m}(s) = \phi_{Y_m}(s/\sqrt{m})^m \to \phi_Z(s).$$

Note that

$$\phi_{X_m}(t) = \exp(it^\top X_m) = \exp(iY_m) = \phi_{Y_m}(1).$$

Therefore, setting $s = 1$, and letting $Z_m = \frac{X_1 + \cdots + X_m}{\sqrt{m}}$,

$$\phi_{Z_m}(t) = \phi_{X_m}(t/\sqrt{m})^m \to \exp(-t^\top t/2).$$

So again using Levy continuity theorem, $Z_m \xrightarrow{\text{d}} Z$, where $Z \sim N(0, I)$.

□

4 Martingales

4.1 3/29 - Conditional expectation

The conditional expectation of X w.r.t. $\mathcal{F}$ gives us the best approximation of X that is constant on the minimal sets of $\mathcal{F}$.

Example: Consider a square lattice on the plane $\mathbb{Z}^2$. Let $B_n = [-n, n]^2$. Let X_n, Y_n independent uniform on $B_n \cap (\mathbb{Z} \times \mathbb{Z})$. Given that X_n, Y_n are on the same unit square, let p_n be conditional probability $X_n = Y_n$. What is $\lim_{n\to\infty} p_n$?

Proof. The answer is 1/9, importantly not 1/4. □

Example (Conditional expectation, finite Ω): Consider a partition of finite Ω into atoms $A_1, \ldots, A_k$, and let Σ_0 be all possible unions of atoms.

Let $\Sigma \subset \Sigma_0$, so Σ_0 is a refinement where Σ is unions of atoms $B_1, \ldots, B_n$, where each B_i is union of some A_js. Then, the *conditional expectation* of X given Σ is defined for $\omega \in B_{i(\omega)}$ as

$$\mathbf{E}(X \mid \widetilde{\Sigma})(\omega) = \frac{\sum_{A_j \subset B_{i(\omega)}} X(\{\omega' \in A_j\})\, \mathbf{P}(A_j)}{\mathbf{P}(B_{i(\omega)})}.$$

Note this is $\widetilde{\Sigma}$-measurable and if $X = \mathbb{1}_E$ then $\mathbf{P}(E \mid \widetilde{\Sigma}) = \mathbf{E}[\mathbb{1}_E \mid \widetilde{\Sigma}]$.

Definition (Conditional expectation): Let $(\Omega, \mathcal{F}_0, \mathbf{P})$, $\mathcal{F} \subset \mathcal{F}_0$. Let X have well defined finite expectation. Then the *conditional expectation* $\mathbf{E}(X \mid \mathcal{F})$ is a $\mathcal{F}$ measurable RV such that

$$\int_B \mathbf{E}(X \mid \mathcal{F})d\,\mathbf{P} = \int_B X d\,\mathbf{P},$$

for every $B \in \mathcal{F}$. This is the canonical way to approximate X over $\mathcal{F}$. We additionally define

(a) For set A, $\mathbf{E}(X \mid A) \coloneqq \mathbf{E}(X \mid \{\varnothing, A, A^c, \Omega\})$.

(b) For RV Z, $\mathbf{E}(X \mid Z) \coloneqq \mathbf{E}(X \mid \sigma(Z))$.

(c) The *conditional probability* of event A given $\mathcal{F}$ as

$$\mathbf{P}(A \mid \mathcal{F}) \coloneqq \mathbf{E}(\mathbb{1}_A \mid \mathcal{F}).$$

Intuitively, how rich $\mathcal{F}$ is tells us how much we know about X. The extreme ends are full information, $\mathbf{E}(X \mid \mathcal{F}_0) = X$, and no information, $\mathbf{E}(X \mid \{\varnothing, \Omega\}) = \mathbf{E}X$.

Theorem: The conditional expectation exists and is uniquely defined up to a set of $\mathbf{P}$-measure zero.

Proof. Existence is by Radon-Nikodym. First suppose $X \geq 0$. For each $A \in \mathcal{F}$, let

$$\nu(A) = \int_A X d\,\mathbf{P}\,.$$

It follows by DCT that this is a measure. Then, for $\mu = \mathbf{P}$ there exists a $\mathcal{F}$-measurable Radon-Nikodym derivative $\frac{d\nu}{d\mu}$ such that

$$\nu(A) = \int_A \frac{d\nu}{d\mu} d\,\mathbf{P}\,.$$

For uniqueness, let $Y = \mathbf{E}(X \mid \mathcal{F})$ and suppose there is some other $\mathcal{F}$-measurable Y' whose integral agrees with X over $\mathcal{F}$. Then letting $A = \{Y - Y' \geq \varepsilon\}$,

$$0 = \int_A X - X d\,\mathbf{P} = \int_A Y - Y' d\,\mathbf{P} \geq \varepsilon\,\mathbf{P}(A),$$

so along with the symmetric argument, $Y = Y'$. □

Proposition:

(a) If $\mathcal{F}$ is independent from X, then

$$\mathbf{E}(X \mid \mathcal{F}) = \mathbf{E}X.$$

(b) If $\mathcal{F}_1 \subset \mathcal{F}_2$ then

$$\mathbf{E}(\mathbf{E}(X \mid \mathcal{F}_1) \mid \mathcal{F}_2) = \mathbf{E}(\mathbf{E}(X \mid \mathcal{F}_2) \mid \mathcal{F}_1) = \mathbf{E}(X \mid \mathcal{F}_1).$$

Intuitively, the smaller σ-field always wins.

Proof.

(a) First $\mathbf{E}X$ is $\mathcal{F}$-measurable since $\{\mathbf{E}X \leq t\} \in \{\varnothing, \Omega\}$. Then for all $A, B \in \mathcal{F}$, we have $\mathbf{P}(\{X \in B\} \cap A) = \mathbf{P}(X \in B)\,\mathbf{P}(A)$. So

$$\int_A X d\,\mathbf{P} = \mathbf{E}(\mathbb{1}_A X) = \mathbf{P}(A)\mathbf{E}X = \int_A \mathbf{E}X d\,\mathbf{P}\,.$$

(b) Omitted.

□

4.2 4/1 - Conditional expectation cont.

Proposition (Monotonicity): If $X \leq Y$ a.s. then $\mathbf{E}(X \mid \Sigma) \leq \mathbf{E}(Y \mid \Sigma)$ a.s. It follows that if $0 \leq X_n \uparrow X$ then

$$\mathbf{E}(X_n \mid \Sigma) \uparrow \mathbf{E}(X \mid \Sigma).$$

Proposition: If Y is Σ measurable and $\mathbf{E}|XY| < \infty, \mathbf{E}|X| < \infty$, then

$$\mathbf{E}(XY \mid \Sigma) = Y\mathbf{E}(X \mid \Sigma),$$

a.s.

Proof. First suppose $Y = \mathbb{1}_B$, $B \in \Sigma$. Then for $C \in \Sigma$,

$$\int_C Y \mathbf{E}(X \mid \Sigma) d\mathbf{P} = \int_{C\cap B} \mathbf{E}(X \mid \Sigma) d\mathbf{P} = \int_{C\cap B} X d\mathbf{P} = \int_C \mathbb{1}_B X d\mathbf{P}.$$

By linearity the property extends to all simple functions. When Y is nonnegative and we can approximate $0 \leq Y_n \uparrow Y$ and invoke earlier proposition to say

$$\mathbf{E}(XY \mid \Sigma) = Y\mathbf{E}(X \mid \Sigma),$$

a.s. For general functions we consider Y^+, Y^-. □

Definition (Conditional expectation, product space): Let $(\Omega, \Sigma, \mathbf{P})$ and X, Y independent. Let $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ Borel, $\mathbf{E}|f(X,Y)| < \infty$, and $\mathbf{E}|f(x,Y)| < \infty$ for μ_X almost every $x \in \mathbb{R}$. Define

$$g(x) \coloneqq \mathbf{E}f(x,Y).$$

Then

$$\mathbf{E}(f(X,Y) \mid X) = g(X).$$

Proof. Intuitively, $g(X)$ is $\sigma(X)$-measurable. Now let $A \in \sigma(X)$ so that $A = \{X \in C\}$ for some Borel C. Then, by change of variables,

$$\begin{aligned}\int_A f(X,Y) d\mathbf{P} &= \iint \mathbb{1}_C f(x,y)\, d\mu(x) d\nu(y) \\ &= \int_C \mathbf{E}f(x,Y) d\mu(x) \\ &= \int_A g(X) d\mathbf{P}.\end{aligned}$$

□

Example: Prove that X has symmetric distribution ($\mathbf{P}(X \leq t) = \mathbf{P}(X \geq -t)$) if and only if $\mathbf{E}(X \mid |X|) = 0$ a.s.

Proof. Suppose X is symmetric. Let $A \in \sigma(|X|)$ so that $A = \{|X| \in C\}$. Then,

$$\begin{aligned}\int_A X d\mathbf{P} &= \mathbf{E}(\mathbb{1}_A X) \\ &= \mathbf{E}(\mathbb{1}_A X^+) - \mathbf{E}(\mathbb{1}_A X^-) \\ &= \int_0^\infty \mathbf{P}(\mathbb{1}_A X^+ \geq t)\, dx - \int_0^\infty \mathbf{P}(\mathbb{1}_A X^- \geq t)\, dx \\ &= \int_0^\infty \mathbf{P}(X \geq t, |X| \in C)\, dx - \int_0^\infty \mathbf{P}(X \leq -t, |X| \in C)\, dx \\ &= 0.\end{aligned}$$

Now suppose $\mathbf{E}(X \mid |X|) = 0$. Writing $X = |X| \operatorname{sgn}(X)$,

$$\mathbf{E}(|X| \operatorname{sgn}(X) \mid |X|) = |X|\mathbf{E}(\operatorname{sgn} X \mid |X|) = 0.$$

So where $|X| > 0$, the conditional expectation is 0 almost surely. So for $t > 0$,

$$\int_{|X|\geq t} \operatorname{sgn}(X) d\mathbf{P} = \mathbf{P}(X \geq t) - \mathbf{P}(X \leq t) = 0.$$

□

4.3 4/3 - Martingales

Martingales are the fortune of a gambler betting on a fair game. The prototypical martingale is the sum of i.i.d. mean zero random variables.

Example: Let X, Y independent standard Gaussian in $\mathbb{R}^2$. Let Z be orthogonal projection of Y onto X. What is the covariance matrix of Z?

Proof. First note that Z is rotationally invariant, since (intuitively) X and Y are rotationally invariant. So the covariance matrix of Z is a multiple of the identity matrix cI.

Condition on a realization of X, we can assume X is on the x-axis. Then the projection of Y onto X is its x-component, distributed $N(0,1)$. Thus the length of the projection squared is distributed $N(0,1)^2$, which has mean 1. So $Z = (z_1, z_2)$ is a rotationally invariant vector with expected length 1: $z_2^2 + z_2^2 = 1$. TODO

□

Definition (Filtration): Let $(\Omega, \Sigma, \mathbf{P})$. A *filtration* is a sequence of σ-algebras $\{\varnothing, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \Sigma$.

X_n is *adapted* to the filtration if X_n is $\mathcal{F}_n$-measurable for all $n \geq 0$.

Definition (Martingale): X_n is *martingale* with respect to $\mathcal{F}_n$ if

(a) X_n is adapted to $\mathcal{F}_n$,

(b) $\mathbf{E}X_n < \infty$,

(c) $\mathbf{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$.

It is *submartingale* (favorable game) if $\mathbf{E}(X_{n+1} \mid \mathcal{F}_n) \geq X_n$ and *supermartingale* (unfavorable game) if $\mathbf{E}(X_{n+1} \mid \mathcal{F}_n) \leq X_n$.

Example: Let Y_n independent mean zero RV. Then the partial sums S_n are martingale with $\mathcal{F}_i = \sigma(Y_j, j \leq i)$.

4.4 4/5 - Martingale concentration inequalities

Theorem (Azuma's inequality): Let M_n martingale with $M_0 = 0$ and $|M_k - M_{k-1}| \leq a_k$ for some sequence of integers a_k. For $n \geq 1$ and any $t > 0$,

$$\mathbf{P}(|M_n| > t) \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n a_k^2}\right).$$

Proof. Fix a parameter λ and use typical exponential Markov:

$$\mathbf{P}(M_n > t) \leq \frac{\mathbf{E}(\exp(\lambda M_n))}{\exp(\lambda t)}.$$

We examine $\mathbf{E}(\exp(\lambda M_n))$ inductively:

$$\begin{aligned}\mathbf{E}(\exp(\lambda M_n)) &= \mathbf{E}(\exp(\lambda(M_n - M_{n-1}))\exp(\lambda M_{n-1})) \\ &= \mathbf{E}(\mathbf{E}(\exp(\lambda(M_n - M_{n-1})) \mid \mathcal{F}_{n-1})\exp(\lambda M_{n-1}))\end{aligned}$$

Now using the identity $e^x \leq x + e^{x^2}$, applying martingale property, and bound on differences:

$$\begin{aligned}\mathbf{E}(\exp(\lambda M_n) \mid \mathcal{F}_{n-1}) &\leq \mathbf{E}(\lambda(M_n - M_{n-1}) + \exp(\lambda^2(M_n - M_{n-1})^2) \mid \mathcal{F}_{n-1}) \\ &\leq \exp(\lambda^2 a_n^2).\end{aligned}$$

So inductively we may show

$$\mathbf{E}(\exp(\lambda M_n)) \leq \prod_i^n \exp(\lambda^2 a_i^2) = \exp\left(\lambda^2 \sum_{i=1}^n a_i^2\right).$$

Now we may minimize our bound over λ.

$$\exp\left(\lambda^2 \sum_{i=1}^n a_i^2 - \lambda t\right) \to \exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right).$$

Finally, applying this to the symmetric negative case gives the extra factor of two in the result, completing the proof. □

Theorem: Let X RV and $\{0, \Omega\} = \mathcal{F}_0 \subset \cdots \subset \mathcal{F}_n = \Sigma$. Suppose for $i = 1, \ldots, n$,

$$|\mathbf{E}(X \mid \mathcal{F}_i) - \mathbf{E}(X \mid \mathcal{F}_{i-1})| \leq 1.$$

Then for all $t \geq 0$,

$$\mathbf{P}(|X - \mathbf{E}X| \geq t) \leq 2\exp\left(-\frac{t^2}{4n}\right).$$

Proof. By exponential Markov,

$$\mathbf{P}(X - \mathbf{E}X \geq t) \leq \frac{\mathbf{E}\exp(\lambda(X - \mathbf{E}X))}{\exp(\lambda t)}.$$

We note that $\mathbf{E}(X \mid \mathcal{F}_n)$ is a martinagle and so we may write $X - \mathbf{E}X$ as the $n+1$th element and mimic the proof of Azuma's inequality:

$$X - \mathbf{E}X = \mathbf{E}(X \mid \mathcal{F}_n) - \mathbf{E}(X \mid \mathcal{F}_0) = \sum_{i=1}^n (\mathbf{E}(X \mid \mathcal{F}_i) - \mathbf{E}(X \mid \mathcal{F}_{i-1})).$$

□

Example: Consider a discrete cube $\{-1,1\}^n$. Let $A \subset \{-1,1\}^n$ contain at least half the vertices. For $v \in \{-1,1\}^n$ define the distance to A as the minimum Hamming distance from points in A:

$$d(v,A) = \min_{y \in A} |\{i \le n : v_i \ne y_i\}|.$$

Then, for $t \ge 0$,

$$\left|\left\{v \in \{-1,1\}^n : d(v,A) \ge t\sqrt{n}\right\}\right| \le 2^{n+1} \exp\left(-\frac{t^2}{16}\right),$$

and for X a uniformly random vertex,

$$\mathbf{P}(d(X,A) \ge t\sqrt{n}) \le 2\exp\left(-\frac{t^2}{16}\right).$$

Proof. Define the probability space as

$$(\{-1,1\}^n, 2^{\{-1,1\}^n}, \text{Uniform}).$$

Consider filtrations

$$\mathcal{F}_i = \sigma(\{v \in \{-1,1\}^n : v_j = a_j, j \le i\}, a_1, \ldots, a_i \in \{-1,1\}),$$

satisfying $(\Omega, \varnothing) = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = 2^{\{-1,1\}^n}$. Let X be a uniformly random vertex and $f(X) \coloneqq d(X,A)$. Then it's possible to show

$$|\mathbf{E}(f(X) \mid \mathcal{F}_i) - \mathbf{E}(f(X) \mid \mathcal{F}_{i-1})| \le 1,$$

and

$$\mathbf{E}d(X,A) \le 3\sqrt{n}.$$

From this we can use Azuma's inequality to complete the proof.

$$\mathbf{P}(|d(X,A) - \mathbf{E}d(X,A)| \ge t) \le 2\exp\left(-\frac{t^2}{4n}\right).$$

□

4.5 4/8 - Transforms and stopping times

Definition (Predictable process): Let $\mathcal{F}_n$. We call H_n a predictable process if H_n is $\mathcal{F}_{n-1}$-measurable.

Definition (Martingale transform): Let X_n be an adapted process. The *martingale transform* of X_n by a predictable prcoess H_n is the process $(H \bullet X)_n$ defined by

$$(H \bullet X)_n \coloneqq \sum_{i=1}^{n} H_i (X_i - X_{i-1}).$$

Remark: $(H \bullet X)_n$ is the discrete analogue of the Ito integral $\int H dX$.

Definition (Stopping time): A nonnegative random integer T is a stopping time wrt a filtration $\mathcal{F}_n$ if the event $\{T = n\}$ is $\mathcal{F}_n$ measurable for all n.

Proposition: Let $\mathcal{F}_n$, X_n, and H_n. Suppose H_n is uniformly bounded. Then

(a) If X_n is (super)submartingale and H_n is nonnegative then $(H \cdot X)$ is (super)submartingale.

(b) If X_n is martingale then $(H \cdot X)$ is martingale.

Proof. We show the proof for supermartingales, which similarly extends to the other two cases.

$$\begin{aligned}\mathbf{E}((H \bullet X)_{n+1} \mid \mathcal{F}_n) &= \sum_{i=1}^{n} H_i(X_i - X_{i-1}) + \mathbf{E}(H_{n+1}(X_{n+1} - X_n) \mid \mathcal{F}_n) \\ &\leq \sum_{i=1}^{n} H_i(X_i - X_{i-1}) && (X_n \text{ s.m., } H_n \geq 0) \\ &= (H \bullet X)_n.\end{aligned}$$

□

Example: Let X_n be the price of a stock we buy at time 0. Let T be the time to sell. Then T is stopping time so long as our strategy is based only on historical price data. The *stopped martingale* $X_{\min(n,T)}$ can be modeled as a martingale transform:

$$X_{\min(n,T)} = \sum_{i=1}^{\infty} \mathbb{1}_{\{T \geq i\}}(X_i - X_{i-1}) + X_0.$$

So in particular we aren't making money by the last proposition. We formalize this in the optional stopping theorem.

4.6 4/10 - Optional stopping time and applications

Consult https://people.eecs.berkeley.edu/ sinclair/cs271/n21.pdf

Theorem (Optional stopping time theorem): Let X_n be (sub/super) martingale and T a stopping time. Let $M < \infty$.

$$X_{\min(T,n)}$$

and

$$X_{\min(\max(T,n),M)}$$

are (sub/super) martingale.

Proof.

$$X_{\min(T,n)} - X_0 = \sum_{i=1}^{n} \mathbb{1}_{\{T \geq i\}}(X_i - X_{i-1})$$

allows us to apply the previous proposition as $\mathbb{1}_{\{T\geq i\}}$ is a nonnegative bounded predictable process. The second part is omitted. □

Theorem (Kolmogorov maximal inequality): Let (X_n) submartingale. Then for all $t > 0$,

$$\mathbf{P}\left(\max_{i\leq n} X_n \geq t\right) \leq \frac{\mathbf{E}|X_n|}{t}.$$

In particular, if Y_j are i.i.d. with zero mean,

$$\mathbf{P}\left(\max_{i\leq n} \sum_{j=1}^{i} Y_j \geq t\right) \leq \frac{\mathbf{E}\left|\sum_{j=1}^{n} Y_j\right|}{t}.$$

Proof. Let $T = \inf\{m \geq 0 : X_m \geq t\}$. Consider submartingale $X_{\min(\max(T,m),n)}$ (indexed by m), the first element is $X_{\min(T,n)}$ and the nth element is X_n. So by optional stopping time theorem, $\mathbf{E}X_{\min(T,n)} \leq \mathbf{E}X_n$, thus

$$\begin{aligned}\mathbf{P}\left(\max_{i\leq n} X_i \geq t\right) &= \mathbf{P}\big(\mathbb{1}_{\{\max_{i\leq n} X_i \geq t\}} X_{\min(T,n)} \geq t\big)\\ &\leq \frac{1}{t}\mathbf{E}(\mathbb{1}_{\{\max_{i\leq n} X_i\geq t\}} X_{\min(T,n)})\\ &\leq \frac{1}{t}\mathbf{E}(\mathbb{1}_{\{\max_{i\leq n} X_i\geq t\}} X_n)\\ &\leq \frac{\mathbf{E}|X_n|}{t}.\end{aligned}$$

□

4.7 4/15 - Upcrossing inequality and martingale convergence theorem

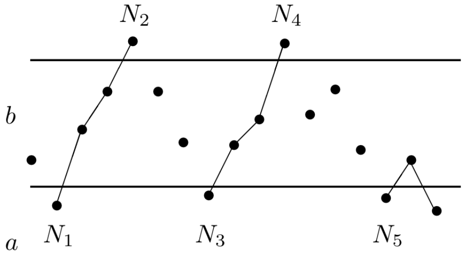
Definition (Upcrossing): Let X_n be a submartingale, define stopping times N_k with $N_0 = -1$,

$$\begin{aligned}N_{2k-1} &= \min\{m > N_{2k-2} : X_m \leq a\}\\ N_{2k} &= \min\{m > N_{2k-1} : X_m \geq b\}.\end{aligned}$$

Consider the predictable sequence

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

that “buys” starting at the next time step when $X_n \leq a$ and “sells” when $X_n \geq b$:



Let $U_n = \max\{k : N_{2k} \le n\}$ be the number of upcrossings completed by time n.

Theorem (Upcrossing inequality): If X_n is submartingale then

$$(b-a)\mathbf{E}U_n \le \mathbf{E}(X_n - a)^+ - \mathbf{E}(X_0 - a)^+.$$

Proof. Define $Y_n = a + (X_n - a)^+$. By Jensen for cet,

$$\begin{aligned}\mathbf{E}(\max(X_{n+1} - a, 0) \mid \mathcal{F}_n) &\ge \max(\mathbf{E}(X_{n+1} \mid \mathcal{F}_n) - a, 0)\\ &\ge \max(X_n - a, 0).\end{aligned}$$

So Y_n is submartingale (it generally holds that for increasing convex φ that doesn't break the finite expectation property, $\varphi(X_n)$ is submartingale). Moreover Y_n upcrosses at the same points as X_n. Since the final potentially incomplete upcrossing by definition has nonnegative contribution,

$$(b-a)U_n \le (H \bullet Y)_n$$

Let $K_n = 1 - H_n$, then

$$Y_n - Y_0 = (H \bullet Y)_n + (K \bullet Y)_n.$$

Since $\mathbf{E}(K \bullet Y)_n \ge \mathbf{E}(K \bullet Y)_0 = 0$,

$$\begin{aligned}(b-a)\mathbf{E}U_n &\le \mathbf{E}(H \bullet Y)_n\\ &\le \mathbf{E}(Y_n - Y_0)\\ &= \mathbf{E}(X_n - a)^+ - \mathbf{E}(X_0 - a)^-.\end{aligned}$$

□

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

Proof. By upcrossing inequality, for any choice of a, b,

$$\mathbf{E}U_n \le \mathbf{E}(X_n - a)^+ \le \mathbf{E}X_n^+ < \infty.$$

so the total number of expected crossings $\mathbf{E}U$ is also finite. Consider the set where X_n does not converge:

$$\left\{\omega : \liminf_n X_n < \limsup_n X_n\right\} = \cup_{a<b\in\mathbb{Q}}\left\{\omega : \liminf_n X_n < a < b < \limsup_n X_n\right\}.$$

Since for each choice of a, b the set on the RHS implies infinitely many crossings, it has measure 0 and thus the union has measure 0. By Fatou's lemma,

$$\mathbf{E}X^+ = \mathbf{E}\liminf X_n^+ \leq \liminf \mathbf{E}X_n^+ < \infty.$$

On the other hand, using again Fatou and $\mathbf{E}X_n^- = \mathbf{E}X_n^+ - \mathbf{E}X_n \leq \mathbf{E}X_n^+ - \mathbf{E}X_0$,

$$\mathbf{E}X^- \leq \liminf \mathbf{E}X_n^- \leq \liminf(\mathbf{E}X_n^+ - \mathbf{E}X_0) < \infty.$$

□

4.8 4/17 - Uniform integrability and Levy's 0-1 law

Definition (Uniform integrability): The sequence X_n is *uniformly integrable* if

$$\lim_{M\to\infty} \sup_n \mathbf{E}(|X_n|\mathbb{1}_{\{X_n\}\geq M}) = 0.$$

Theorem: Suppose $\mathbf{E}|X_n| < \infty$ and $X_n \to X$ in probability. Then the following are equivalent:

(a) $\{X_n, n \geq 0\}$ uniformly integrable.

(b) $X_n \to X$ in L^1.

(c) $\mathbf{E}|X_n| \to \mathbf{E}|X| < \infty$.

Proof. Omitted. □

Theorem: Given $X \in L^1$ on $(\Omega, \mathcal{F}_0, \mathbf{P})$,

$$\{\mathbf{E}(X \mid \mathcal{F}) : \mathcal{F} \subset \mathcal{F}_0\}$$

is uniformly integrable.

Proof. Omitted. □

Theorem (Martingale convergence): Let X RV with $\mathbf{E}|X| < \infty$. Let $\mathcal{F}_n$ be filtration and define $\mathcal{F}_\infty := \sigma(\mathcal{F}_n, n \geq 1)$. Then

$$\mathbf{E}(X \mid \mathcal{F}_n)$$

converges to $\mathbf{E}(X \mid \mathcal{F}_\infty)$ almost surely and in L^1.

Proof. It's easy to check that $\mathbf{E}(X \mid \mathcal{F}_n)$ is a martingale satisfying the condition of the submartingale convergence theorem. So there is ($\mathcal{F}_\infty$-measurable) random variable Y such that $\mathbf{E}(X \mid \mathcal{F}_n) \xrightarrow{\text{a.s.}} Y$. Then, by the previous lemma, this convergence is further in L^1.

To show $Y = \mathbf{E}(X \mid \mathcal{F}_\infty)$, we must show for all $E \in \mathcal{F}_\infty$, $\mathbf{E}(\mathbb{1}_E X) = \mathbf{E}(\mathbb{1}_E Y)$. Fix an $E \in \mathcal{F}_m$, then for all $n \geq m$,

$$\mathbb{1}_E \mathbf{E}(X \mid \mathcal{F}_n) = \mathbf{E}(X\mathbb{1}_E \mid \mathcal{F}_n) \implies \mathbf{E}(\mathbb{1}_E \mathbf{E}(X \mid \mathcal{F}_n)) = \mathbf{E}(X\mathbb{1}_E).$$

By L^1 convergence to Y,

$$\mathbf{E}(\mathbb{1}_E \mathbf{E}(X \mid \mathcal{F}_n)) \to \mathbf{E}(\mathbb{1}_E Y).$$

So $\mathbf{E}(\mathbb{1}_E Y) = \mathbf{E}(\mathbb{1}_E X)$ for all $E \in \cup_n \mathcal{F}_n$. So we are done by minimality of $\mathcal{F}_\infty$. □

As a corollary it is immediate to see the intuitive fact that we will gradually become certain of outcomes of events with more information:

Corollary (Levy's 0-1 law): Let $\mathcal{F}_n$ be filtration and $\mathcal{F}_\infty = \sigma(\mathcal{F}_n, n \geq 1)$. Then for any event $E \in \mathcal{F}_\infty$,

$$\mathbf{P}(E \mid \mathcal{F}_n) \xrightarrow{\text{a.s.}} \mathbb{1}_E.$$

In particular, the random variable $\lim_{n\to\infty} \mathbf{P}(E \mid \mathcal{F}_n)$ exists and belongs to $\{0, 1\}$ almost surely.

4.9 4/19 - Approximation lemma and 0-1 laws

Lemma: Let $\mathcal{F}_n$ filtration and $\mathcal{F}_\infty = \sigma(\mathcal{F}_i, i \geq 1)$. For any event $E \in \mathcal{F}_\infty$, we can approximate E with arbitrary precision from some $\mathcal{F}_n$: for all $\varepsilon > 0$ there exists $n \in \mathbb{N}, E_n \in \mathcal{F}_n$ such that $\mathbf{P}(E \triangle E_n) < \varepsilon$.

Proof. We use $\pi - \lambda$. Let L be the set of $E \in \mathcal{F}_\infty$ satisfying this property. Show L is λ-system. □

Theorem (Kolmogorov's 0-1 law): Let X_n independent and finite everywhere on probability space. Define terminal σ-field (events in every tail) as

$$\mathcal{F}_T := \cap_n \sigma(X_n, X_{n+1}, \dots).$$

Then for all $F \in \mathcal{F}_T$, $\mathbf{P}(F) \in \{0, 1\}$.

Proof. Take $E \in \mathcal{F}_\infty$. For fixed $\varepsilon > 0$ use the lemma to find E_n such that $\mathbf{P}(E_n \triangle E) < \varepsilon$. By definition of $\mathcal{F}_\infty$, E_n is independent from F. So

$$\begin{aligned}\mathbf{P}(F \cap E_n) = \mathbf{P}(F)\,\mathbf{P}(E_n) &\implies \mathbf{P}(F \cap E) \pm \varepsilon = \mathbf{P}(F)(\mathbf{P}(E) \pm \varepsilon) \\ &\implies |\mathbf{P}(F \cap E) - \mathbf{P}(F)\,\mathbf{P}(E)| < 2\varepsilon.\end{aligned}$$

Thus F, E are independent, which means that the terminal σ-field is independent from itself ($\mathcal{F}_T \subset \mathcal{F}_\infty$). This means

$$\mathbf{P}(F) = \mathbf{P}(F \cap F) = \mathbf{P}(F)\,\mathbf{P}(F) \implies \mathbf{P}(F) \in \{0, 1\}.$$

□

Corollary: Let $Y = \limsup \frac{1}{n} \sum_{i=1}^n X_i$ (allowed to take values $\pm\infty$). There exists constant $c \in \overline{\mathbb{R}}$ such that $Y = c$ a.s.

Proof. It suffices to show that Y is $\mathcal{F}_T := \cap_m \sigma(X_m, X_{m+1}, \dots)$-measurable and apply Kolmogorov's 0-1 law. For any fixed m let $Y_m = \frac{1}{n} \sum_{i=m}^n X_i$. It's clear to see that $Y_m = Y$ everywhere on Ω. As we take $m \to \infty$, it follows that Y is $\mathcal{F}_T$ measurable. □

4.10 4/22 - Backwards martingales

Definition: Let $\mathcal{F}_n$ for $n \le 0$ and X_n adapted to the filtration. If

(a) $\mathbf{E}|X_n| < \infty$,

(b) $\mathbf{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$ for $n \le -1$.

then X_n is called *backwards martingale* w.r.t. $\mathcal{F}_n$.

Theorem (Convergence of backwards martingales): Let $\mathcal{F}_{-\infty} = \cap_n \mathcal{F}_n$. Then $\lim_{n\to-\infty} X_n$ exists and is equal to $\mathbf{E}(X_{-1} \mid \mathcal{F}_{-\infty})$.

Proof. Omitted, proof by upcrossing lemma. □

Definition (Exchangeable σ-field): Given X_n i.i.d., the *exchangeable* σ-field, denoted $\mathcal{E}$, is the collection of events whose occurences are not affected by permuting a finite number of the X_n. Note that this contains the terminal σ-field.

Theorem (Hewitt-Savage 0-1 law): Let X_n i.i.d. and $A \in \mathcal{E}$. Then $\mathbf{P}(A) \in \{0, 1\}$.

Proof. Omitted. □

Example (SLLN revisited): Let X_n i.i.d. and $\mathbf{E}|X_i| < \infty$. Show $\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\text{a.s.}} \mathbf{E}X_1$.

Proof. Assume WLLN. Define backwards martingale Y by the partial sums of X_n, $Y_0 = 0$, $Y_{-n} := S_n$, and let

$$\mathcal{F}_{-n} = \sigma(Y_{-n}, Y_{-n-1}, \dots).$$

To check Y_n is backwards martingale w.r.t. $\mathcal{F}_n$, we must show $\mathbf{E}(Y_{-n} \mid \mathcal{F}_{n-1}) = Y_{-n-1}$. Write $Y_{-n} = \frac{S_{n+1} - X_{n+1}}{n}$, and note that by symmetry

$$\mathbf{E}(X_{n+1} \mid \mathcal{F}_{-n-1}) = \frac{1}{n+1} \sum_{i=1}^n \mathbf{E}(X_i \mid \mathcal{F}_{-n-1}) = \frac{S_{n+1}}{n+1}.$$

$$\begin{aligned}\mathbf{E}(Y_{-n} \mid \mathcal{F}_{-n-1}) &= \mathbf{E}(S_{n+1}/n \mid \mathcal{F}_{-n-1}) - \mathbf{E}(X_{n+1}/n \mid \mathcal{F}_{-n-1}) \\ &= \frac{S_{n+1}}{n} - \frac{S_{n+1}}{n(n+1)} \\ &= \frac{S_{n+1}}{n+1} \\ &= Y_{-n-1}.\end{aligned}$$

By backwards martingale convergence,

$$\lim_{m\to-\infty} Y_m = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mathbf{E}(Y_{-1} \mid \mathcal{F}_{-\infty}).$$

We can show $\mathcal{F}_{-\infty} \subset \mathcal{E}$, and it follows by Hewitt-Savage 0-1 law that $\mathcal{F}_{-\infty}$ is trivial and $\mathbf{E}(Y_{-1} \mid \mathcal{F}_{-\infty}) = \mathbf{E}Y_{-1}$. □

Example: Let X_n, Y_n adapted to $\mathcal{F}_n$ and nonnegative. Assume $\sum_n Y_n$ is finite almost everywhere, and

$$\mathbf{E}(X_{n+1} \mid \mathcal{F}_n) \le X_n + Y_n$$

almost everywhere. Show $\lim_{n\to\infty} X_n$ exists a.s.

Proof. Define $W_n = X_n - \sum_{i=1}^{n-1} Y_i$. Then

$$\mathbf{E}(W_{n+1} \mid \mathcal{F}_n) = \mathbf{E}(X_{n+1} \mid \mathcal{F}_n) - \sum_{i=1}^{n} Y_i \le X_n - \sum_{i=1}^{n-1} Y_i = W_n,$$

so W_n is supermartingale. Fix a truncation level $M > 0$. Define stopping time $N := \inf\left\{k \le \infty : \sum_{j=1}^{k} Y_j > M\right\}$ (this can take ∞ with positive probability). Then the process

$$W_{\min(n,N)}$$

is supermartingale by optional stopping time theorem, and also $W_{\min(n,N)} \ge -M$ a.s. for $n \ge 1$. TODO □

4.11 4/24 - Important examples

Example: Let X, Y on $(\Omega, \Sigma, \mathbf{P})$ and X', Y' on $(\Omega', \Sigma', \mathbf{P}')$. If $(X,Y) \sim (X', Y')$ and $\mathbf{E}|X| < \infty$, then $\mathbf{E}(X \mid Y) \sim \mathbf{E}(X' \mid Y')$.

Proof. Enumerate rationals as q_n. Let $\mathcal{F}_n := \sigma(\{Y < q_i\}, i \le n)$ and $\mathcal{F}'_n := \sigma(\{Y' < q_i\}, i \le n)$. Note that

$$\mathcal{F}_\infty = \sigma(\mathcal{F}_n, n \ge 1),\ \mathcal{F}'_\infty = \sigma(\mathcal{F}'_n, n \ge 1).$$

By martingale convergence theorem,

$$\mathbf{E}(X \mid \mathcal{F}_n) \xrightarrow{\text{a.s.}} \mathbf{E}(X \mid Y),\ \mathbf{E}(X' \mid \mathcal{F}'_n) \xrightarrow{\text{a.s.}} \mathbf{E}(X' \mid Y').$$

We claim that for all n,

$$\mathbf{E}(X \mid \mathcal{F}_n) \sim \mathbf{E}(X' \mid \mathcal{F}'_n).$$

There is bijection between atoms of $\mathcal{F}_n$ and $\mathcal{F}'_n$, and on corresponding atoms the conditional distributions of X and X' are the same. Thus the values of the conditional expectations are the same. □

Example: Let X, Y on $(\Omega, \Sigma, \mathbf{P})$. Assume there is well defined density $f(x, y)$ for (X, Y) which implies $f_Y(y)$ is well defined. What is $\mathbf{E}(X \mid Y)$?

Proof. Let $(X', Y')(z_1, z_2) := (z_1, z_2)$ on the space $(\mathbb{R}^2, \mathcal{B}_2, \mu_{(X,Y)})$, where X' and Y' themselves are projections. By last example,

$$(X', Y') \sim (X, Y) \implies \mathbf{E}(X' \mid Y') \sim \mathbf{E}(X \mid Y).$$

Claim that

$$\mathbf{E}(X' \mid Y')(z_1, z_2) = \frac{1}{f_Y(z_2)} \int_{-\infty}^{\infty} x f(x, z_2)\, dx,$$

when $f_Y(z_2) > 0$ and the integral exists, and arbitrary value otherwise. This is $\sigma(Y')$-measurable. Now consider Y'-measurable event $E = \mathbb{R} \times B$ where B is Borel subset of $\mathbb{R}$ such that on B, $f_Y > 0$ and the prior integral is well defined.

$$\begin{aligned}
&\int_B \int_{\mathbb{R}} \mathbf{E}(X' \mid Y')(z_1, z_2) d\mu(z_1, z_2) \\
&= \int_B \int_{\mathbb{R}} \mathbf{E}(X' \mid Y')(z_1, z_2) f(z_1, z_2) dz_1 dz_2 \\
&= \int_B \frac{1}{f_Y(z_2)} \int_{\mathbb{R}} \left(\int_{-\infty}^{\infty} x f(x, z_2)\, dx \right) f(z_1, z_2) dz_1 dz_2 \\
&= \int_B \int_{-\infty}^{\infty} x f(x, z_2)\, dx dz_2 \\
&= \int_E X' d\mathbf{P}.
\end{aligned}$$

□

Example: Let Y with $\mathbf{E}|Y|^2 < \infty$ and $\mathcal{F}$. Define V as linear space of square integrable $\mathcal{F}$-measurable random variables. Then $\mathbf{E}(Y \mid \mathcal{F}) = Proj_V(Y)$ a.s. Equivalently, this conditional expectation minimizes mean squared error $\mathbf{E}(\mathbf{E}(Y \mid \mathcal{F}) - Y)^2 = \min_{Z \in \mathcal{F}} \mathbf{E}(Z - Y)^2$.

5 Appendix

Definition (Big O, small O): We write

$$f(x) = O(g(x))$$

if for large enough x, there is constant c such that $0 \leq f(x) \leq cg(x)$. We write

$$f(x) = o(g(x))$$

if $\frac{f(x)}{g(x)} \to 0$. We write

$$f(x) \sim g(x)$$

if $\frac{f(x)}{g(x)} \to 1$. Is is normally assumed limits are taken as $x \to \infty$ but it can also be to a finite limit $x \to a$.

Theorem (Stirling's approximation): As $n \to \infty$,

$$n! = n^n e^{-n} \sqrt{2\pi n} \cdot (1 + O(1/n)).$$

Proof. (Up to constants) Approximate $\ln(n!) = \sum_{k=1}^{n} \ln k$ as a Riemann sum. □

Theorem (Taylor series approximations): As $x \to 0$,

$$e^x = 1 + x + O(x^2) = 1 + x + o(1).$$