Contents

Contents **1**

2 Laws of Large Numbers **2**

2.1 Independence . . . 2
2.2 Weak Laws of Large Numbers . . . 6
2.3 Triangular Arrays . . . 10
2.4 Borel-Cantelli Lemmas . . . 12
2.5 Kolmogorov 0-1 Law . . . 15
2.6 Strong Law of Large Numbers . . . 16
2.7 Large Deviations . . . 21

3 Weak Convergence and CLT **24**

3.1 Weak Convergence . . . 25
3.2 Characteristic Functions . . . 31
3.3 Weak Convergence . . . 33
3.4 Central Limit Theorem . . . 37
3.5 Poisson Convergence & Poisson Processes . . . 40
3.6 Conditional Probabilities . . . 44

Chapter 2

Laws of Large Numbers

2.1 Independence

Beginning of Sept.12, 2022

First, some definitions/recaps on independence of events and σ-fields:

- Independence of two events: we say events A, B are independent, $A \perp B$, if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
- Independence of two σ-fields: $\mathcal{F}, \mathcal{G}$ are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all $A \in \mathcal{F}$ and $\mathcal{B} \in \mathcal{G}$.
- For *more than* 2 *events*: $A_1, \ldots, A_n$ are **mutually independent** if

$$\mathbb{P}(\bigcap_{i\in I} A_i) = \prod_{i\in A} \mathbb{P}(A_i) \qquad \text{for all } I \in \{1, \ldots, n\}. \tag{*}$$

- Similarly, for more than 2 sigma fields, we say $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent if the above product identity holds for all $A_i \in \mathcal{A}_i$.
- We say events $A_1, \ldots, A_n$ are **pairwise independent** if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \qquad \text{for all } i \neq j.$$

Example: $\mathbb{P}(\bigcap A_i) = \prod \mathbb{P}(A_i)$ is insufficient. Consider two coin tosses. Let $A := \{\text{first is head}\}$, $B := \{\text{second is head}\}$, and $C := \{\text{both tosses are the same}\}$. Then $A \cap B \subset C$, so they are not mutually independent, but

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{8}.$$

In fact, we also have A, B, C pairwise independent here.

- For an *infinite sequence* of A_i's, we say they are independent if (*) holds for any *finite* $I \subset \mathbb{N}$.

Moving to independence of two random variables:

- Two random variables X, Y are independent if

$$\mathbb{P}(X \in Y, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \tag{**}$$

for all A, B in their corresponding σ-fields. It can be shown that this definition is equivalent to requiring $\sigma(X)$ and $\sigma(Y)$ to be independent.

- To show independence, it is sufficient to check (**) for $(-\infty, x] \times (-\infty, x]$ for all x, y. That is,

$$F_{(X,Y)}(x, y) = F_X(x)F_Y(y) \qquad \text{for all } x, y.$$

Example: $\mathcal{A} \perp \mathcal{B}$ does not imply $\sigma(\mathcal{A}) \perp \sigma(\mathcal{B})$. (The example given in lecture relies heavily on drawings so I will replace it with one easier to type in LaTeX.) Let $\mathcal{A} := \{\{1,2\},\{3,4\}\}$ and let $\mathcal{B} := \{\{2,4\}\}$. Then $\{2,4\} \in \sigma(\mathcal{A})$.

Definition: π-system and λ-system

A collection $\mathcal{G}$ is called a **π-system** if it is nonempty and closed under finite intersections (two suffice):

- $\mathcal{G} \neq \varnothing$, and
- For $A, B \in \mathcal{G}$, $A \cap B \in \mathcal{G}$.

A collection $\mathcal{G}$ is called a **λ-system** if $\mathcal{G}$ contains Ω, is closed under set subtraction, and is closed under countable increasing union:

- $\Omega \in \mathcal{G}$,
- If $A \subset B$ and $A, B \in \mathcal{G}$ then $B \backslash A \in \mathcal{G}$, and
- If $A_n \in \mathcal{G}$ and $A_n \uparrow A$ then $A \in \mathcal{G}$.

The **$\pi - \lambda$ theorem** states that if $\mathcal{P}$ is a π-system and $\mathcal{L}$ a λ-system with $\mathcal{P} \subset \mathcal{L}$, then $\sigma(\mathcal{P}) \subset \mathcal{L}$.
We will skip the proof and directly use the result to prove the following (the proof of which we again omit):

Theorem: D2.1.7

If $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent σ-fields and each $\mathcal{A}_i$ a π-system, then the $\sigma(\mathcal{A}_i)$'s are independent.

We now discuss the independence of functions of random variable in greater generality. Suppose we have an array of independent random varaibles

$$\{X_{i,j} : i \leq n, j \leq m(i)\}$$

and n functions

$$X_{1,1}, \ldots, X_{1,m(1)} \mapsto f_1(X_{1,1}, \ldots, X_{1,m(1)})$$
$$X_{2,1}, \ldots, X_{2,m(2)} \mapsto f_2(x_{2,1}, \ldots, X_{2,m(2)})$$

and so on, where each $f_i : \mathbb{R}^{m(i)} \to \mathbb{R}$. **Question**: are these random variables $f_i(\cdot)$ independent? The answer is yes, and we will formulate the question in terms of σ-fields:

Theorem: D2.1.10

Given an independent collection of σ-fields $\{\mathcal{F}_{i,j} : i \leqslant n, j \leqslant m(i)\}$, let $\mathcal{B}_i := \sigma(\mathcal{F}_{I,1}, \dots, \mathcal{F}_{i,m(i)})$ (i.e., the i^{th} row listed above). Then $\mathcal{B}_1, \dots, \mathcal{B}_n$ are independent.

Proof. For each row, let

$$A_i := \{\text{all} \bigcap_{i=1}^{m(i)} A_{i,j} \text{ with } A_{i,j} \in \mathcal{F}_{i,j}\}.$$

Then $\mathcal{A}_i$ is a π-system that contains Ω (intersection of all $\Omega \in \mathcal{F}_{i,j}$) and also all $\mathcal{F}_{i,j}$ (intersection of $\mathcal{F}_{i,j}$ with a bunch of Ω's). Therefore $\mathcal{A}_i$ generates $\mathcal{B}_i$. Finally, the $\mathcal{A}_i$'s are independent:

$$\mathbb{P}(\bigcap_{i=1}^{n}(\bigcap_{j=1}^{m(i)} A_{i,j})) = \prod_{i=1}^{n}\prod_{j=1}^{m(i)} \mathbb{P}(A_{i,j}) = \prod_{i=1}^{n}\mathbb{P}(\bigcap_{j=1}^{m(i)} A_{i,j}).$$

Therefore, by (D2.1.7) the $\mathcal{B}_i$'s are independent. □

Beginning of Sept. 14, 2022

Theorem

Let $\{X_{i,j} : i \leqslant n, j \leqslant m(i)\}$ be independent. Then $f(X_{i,1}, \dots, X_{i,m(i)})$, $i \leqslant n$, are independent random variables.

Proof. Let $\mathcal{F}_{i,j} = \sigma(X_{i,j})$ and $\mathcal{B}_i := \sigma(\mathcal{F}_{i,1}, \dots, \mathcal{F}_{i,m(i)})$. By the previous theorem each $\mathcal{B}_i$'s are independent. Each f_i is $\mathcal{B}_i$-measurable so the random variables f_i are independent. □

Fubini theorem says for $f(x, y)$ on $\Omega \times \Omega_2$,

$$\int f \, \mathrm{d}(\mu_1 \times \mu_2) = \int_{\Omega_1}\int_{\Omega_2} f \, \mathrm{d}\mu_2 \, \mathrm{d}\mu_1$$

provided $f \geqslant 0$ or f is integrable (i.e., $\int |f| \, \mathrm{d}(\mu_1 \times \mu_2) < \infty$). (Here since μ_i's are probability measures they are assumed to be σ-finite.) For random variables:

Theorem: D2.1.12

Let X, Y be independent with distributions μ_X and μ_Y on $\mathbb{R}$. Let $h : \mathbb{R}^2 \to \mathbb{R}$ satisfy either $h \geqslant 0$ or $\mathbb{E}|h(X,Y)| < \infty$. Then

$$\mathbb{E}h(X,Y) = \iint_{\mathbb{R}^2} h(X,Y) \, \mathrm{d}\mu_X(\mathrm{d}x)\mu_Y(\mathrm{d}y) \tag{*}$$

and the other of integration does not matter.
In particular, for $h(x,y) = f(x)g(y)$ with either $f, g \geqslant 0$ or $\mathbb{E}|f(X)|, \mathbb{E}|g(Y)| < \infty$, we have

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}f(X)\mathbb{E}g(Y), \tag{**}$$

i.e., independence $\Rightarrow$ (product of $\mathbb{E}$ = $\mathbb{E}$ of product).

Proof. (*) follows from Fubini since the distribution of (X,Y) is $\mu_X \times \mu_Y$ by independence.

For (**),

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \iint_{\mathbb{R}^2} f(x)g(y)\ \mu_X(\mathrm{d}x)\mu_Y(\mathrm{d}y)\\ &= \int_{\mathbb{R}} g(y) \int_{\mathbb{R}} f(x)\mu_X(\mathrm{d}x)\mu_Y(\mathrm{d}y)\\ &= \left(\int_{\mathbb{R}} f(x)\mu_X(\mathrm{d}x)\right)\int_{\mathbb{R}} g(y)\mu_Y(\mathrm{d}y) = \mathbb{E}f(X)\mathbb{E}g(Y).\end{aligned}$$

□

By induction, we may generalize the above result into any finite number of random variables. That is, for independent $X_1, \dots, X_n$ with $\mathbb{E}|\prod X_i| < \infty$, we have $\mathbb{E}[\prod X_i] = \prod \mathbb{E}X_i$.

Sums of Independent Random Variables

Let X, Y be independent with distribution functions F and G. The d.f. of $X + Y$ is the **convolution**

$$H(z) = \mathbb{P}(X + Y \leqslant z) = \int_{-\infty}^{\infty} F(z-y)\mathrm{d}(G(y)) =: (F * G)(z).$$

To see this, we apply Fubini to $1_{x+y\leqslant z}$:

$$H(z) = \mathbb{E}1_{\{X+Y\leqslant z\}} = \iint 1_{\{x\leqslant z-y\}}\ \mathrm{d}F(x)\mathrm{d}G(y) = \int F(z-y)\ \mathrm{d}G(y).$$

Note by doing $\{y \leqslant x - z\}$ first we obtain see that $F * G \equiv G * F$.

Example. Let X be uniform on $[0, 2]$ and Y is exponential with parameter λ. That is, X has $1/2$ on $[0, 2]$ and Y has $\lambda e^{-\lambda y}$ on $[0, \infty)$. The distribution function of Y is $1 - e^{-\lambda y}$ for $y \geqslant 0$. Then

$$H(z) = \int_{-\infty}^{\infty} \underbrace{(1 - e^{-\lambda(z-y)})1_{\{z-y\geqslant 0\}}}_{F(z-y)} \underbrace{\frac{1}{2}1_{[0,2]}\ \mathrm{d}y}_{\mathrm{d}G(y)}.$$

That is,

$$H(z) = \begin{cases} 0 & z < 0 \\ \int_0^z (1 - e^{-\lambda z}e^{\lambda y})/2\ \mathrm{d}y & 0 \leqslant z \leqslant 2 \\ \int_0^2 (1 - e^{-\lambda z}e^{\lambda y})/2\ \mathrm{d}y & z > 2 \end{cases} = \begin{cases} 0 & z < 0 \\ \frac{z}{2} - \frac{1 - e^{-\lambda z}}{2\lambda} & 0 \leqslant z \leqslant 2 \\ 1 - e^{-\lambda z}\frac{e^{2\lambda} - 1}{2\lambda} & z > 2. \end{cases}$$

Let μ_n be a probability measure on $(\mathbb{R}^n, \mathcal{R}^n)$. We can make a random vector with distribution μ_n:we take $(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}^n, \mathcal{R}^n, \mu_n)$ and $(X_1, \dots, X_n)$ to be identity.

Infinite Sequence of Random Variables

We say **finite-dimensional** distributions of $\{X_n, n \geqslant 1\}$ are all distributions of form $\{X_i, i \leqslant I\}$ for $I \subset \mathbb{N}$ finite. By using marginals is sufficient to consider $I = \{1, \dots, n\}$. Suppose we are given μ_n on $(\mathbb{R}^n, \mathcal{R}^n)$ for every n.
Question: is there a $\mathbb{P}$ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}})$ with distribution μ_n for the first n coordinates? That is,

$$\mathbb{P}(A_1 \times \dots \times A_n \times \mathbb{R} \times \dots) = \mu_n(A_1 \times \dots \times A_n)?$$

(Well of course no, since if $n > m$, μ_n determines what μ_m would be.) What if this consistency is satisfied?

Theorem: Kolmogorov Extension Theorem

Let μ_n be a p.m. on $(\mathbb{R}^n, \mathcal{R}^n)$ for all n. Suppose consistency holds among the μ_n's, i.e.,

$$\mu_{n+1}(A \times \mathbb{R}) = \mu_n(A) \qquad \text{for all } A = \prod_{i=1}^{n}(a_i, b_i] \text{ and } n \geqslant 1.$$

(In reality the above choice of A can be anything in $\mathcal{R}^n$; we just picked the most canonical one.) Then there exists a unique $\mathbb{P}$ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}})$ with

$$\mathbb{P}(\prod_{i=1}^{n}(a_i, b_i] \times \mathbb{R} \times \mathbb{R} \times \ldots) = \mu_n(\prod_{i=1}^{n}(a_i, b_i]).$$

(Sets of form $A \times \mathbb{R} \times \mathbb{R} \times \ldots$ with $A \in \mathcal{R}^n$ is a **cylinder set** in $\mathbb{R}^{\mathbb{N}}$. They form a σ-field.)

2.2 Weak Laws of Large Numbers

Some recaps:

- We say $X_n \to X$ **in probability** (i.e. in measure) if $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.
- We say $X_n \to X$ **in L^p** (where $p > 0$) if $\mathbb{E}(|X_n - X|^p) \to 0$ as $n \to \infty$.
  - For $p \geqslant 1$, this is equivalent to $\|X_n - X\|_p \to 0$.
  - For $p < 1$ this does not hold as such $\|\cdot\|_p$ does not define a norm.
  - In principle the X_n can have infinite p^{th} moment but the definition still makes sense.

Beginning of Sept. 16, 2022

Theorem

Convergence in L^p implies convergence in probability.

Proof. Suppose $X_n \to X$ in L^p. Then for all $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|X_n - X|^p > \epsilon^p) \leqslant \frac{\mathbb{E}|X_n - X|^p}{\epsilon^p} \to 0$$

The converse fails due to mass escaping. For example, consider a coin with probability of heads $1/n$. Let U be uniform on $[0, 1]$. Define

$$X_n = \begin{cases} U & \text{if tails} \\ U + n^{1/p} & \text{if heads.} \end{cases}$$

Then

$$\mathbb{P}(|X_n - U| > \epsilon) = \mathbb{P}(\text{tails}) = \frac{1}{n} \to 0$$

whereas for all n,

$$\mathbb{E}(|X_n - U|^p) = \frac{1}{n}(n^{1/p})^p = 1.$$

More recaps:

- If $\mathbb{E}X^2, \mathbb{E}Y^2 < \infty$, we defined the **covariance** $\operatorname{cov}(X,Y) := \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}X)) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y$.
- X, Y are **uncorrelated** if $\operatorname{cov}(X,Y) = 0$. Notation: $\sigma_X := \operatorname{cov}(X,X) = \operatorname{var}(X)^{1/2}$.
- The **correlation** coefficient of $X - Y$ is invariant under affine mappings of X, Y (but cov is not, which makes it dependent on units). In particular it computes the covariance of standardized X, Y:
$$\rho(X,Y) := \operatorname{cov}\left(\frac{(X - \mathbb{E}X)}{\sigma_X}, \frac{(Y - \mathbb{E}Y)}{\sigma_Y}\right) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \in [-1, 1].$$
- Given X, Y, by minimizing $\mathbb{E}[(Y - (aX + b))^2]$, the quantity $\mathbb{E}[(Y - (aX + b))^2]/\sigma_Y^2$ is the "fraction of Y variance due to deviation from the best fit line (i.e., not caused by X)", and it is $1 - \rho(X,Y)^2$.
- Variance of sums $S_n = X_1 + \ldots + X_n$:
$$\begin{aligned}\operatorname{var}(S_n) &= \mathbb{E}\left[\sum_{i=1}^n (X_i - \mathbb{E}X_i)\right]^2 \\ &= \mathbb{E}\sum_{i=1}^n \sum_{j=1}^n (X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j) \\ &= \sum_{i=1}^n \operatorname{var}(X_i) + 2\sum_{i<j} \operatorname{cov}(X_i, X_j).\end{aligned}$$
- Independence implies zero correlation, but not conversely: consider the distribution of (X, Y) defined uniformly on $\{(0,0)\} \cup \{\pm 1\} \times \{\pm 1\}$ (i.e. each point with $1/5$ probability). By symmetry, the correlation of (X, Y) is 0, but $Y = 0$ only if $X = 0$.

Theorem: L^2 weak Law, D2.2.3

Suppose $X_1, X_2, \ldots$ are uncorrelated with $\mathbb{E}X_i = \mu$ for each i and $\operatorname{var}(X_i) \leqslant C < \infty$. Then $S_n/n \to \mu$ in L^2 (and therefore in probability).

Proof. A one liner proof:
$$\mathbb{E}\left(\frac{S_n}{n} - \mu\right)^2 = \operatorname{var}(S_n/n) = \frac{1}{n^2}\sum_{i=1}^n \operatorname{var}(X_i) \leqslant \frac{C}{n} \to 0. \qquad \square$$

Beginning of Sept. 19, 2022

Theorem

If $X \geqslant 0$ and $p > 0$, then
$$\mathbb{E}X^p = \int_0^\infty p x^{p-1} \mathbb{P}(X > x)\, \mathrm{d}x.$$

Proof. We take $g(x) = x^p$ for $x \geqslant 0$ and 0 otherwise. Then
$$\begin{aligned}\mathbb{E}(X^p) = \mathbb{E}g(X) &= \lim_{b\to\infty} \int_{[0,b]} g(x)\, \mathrm{d}F(x) \\ &= \lim_{b\to\infty} -\mathbb{P}(X > b) b^p + \lim_{b\to\infty} \int_{[0,b]} p x^{p-1} \mathbb{P}(X > x)\, \mathrm{d}x.\end{aligned}$$

- If $\mathbb{E}(X^p) < \infty$, from homework we know $\mathbb{P}(X > b) b^p \to 0$, so the claim is true.

- If $\mathbb{E}(X^p) = \infty$, then

$$\infty = \mathbb{E}(X^p) \leqslant \lim_{b\to\infty} \int_{[0,b]} px^{p-1}\mathbb{P}(X > x)\,\mathrm{d}x = \int_0^\infty px^{p-1}\mathbb{P}(X > x)\,\mathrm{d}x. \qquad \square$$

Recall from L^2 weak law, if $X_1, X_2, \ldots$ are i.i.d. with $\mathbb{E}X_1 = \mu$ and $\mathbb{E}X_1 < \infty$, then

$$\mathbb{P}(|S_n/n - \mu| > \epsilon) \to 0 \qquad \text{for all } \epsilon > 0.$$

What if we weaken the assumptions? What if $\mathbb{E}X_1 = \infty$ or undefined? Is there $\{\mu_n\}$ such that $\mathbb{P}(|S_n/n - \mu_n| > \epsilon) \to 0$, or does the sequence S_n/n retain its randomness?
Intuitively, if S_n/n settles down, no particular X_i should contribute much to this quantity. To formulate, we require that

$$\mathbb{P}(|X_j|/n > \delta \text{ for some } j \geqslant 0) \to 0 \qquad \text{for all } \delta.$$

This is the same as requiring

$$1 - \mathbb{P}(|X_n| \leqslant \delta_n)^n = 1 - (1 - \mathbb{P}(|X_1| > \delta_n)^n \to 1.$$

We use the fact that if for $a_n \in (0,1)$ with $a_n \to 0$ and $b_n \to \infty$, $(1-a_n)^{b_n} \to 1$ if and only if $a_n b_n = 0$. To see this: for small a_n (or equivalently large n),

$$e^{-2a_n} \leqslant 1 - a_n \leqslant e^{-a_n} \implies e^{-2a_n b_n} \leqslant (1-a_n)^{b_n} \leqslant e^{-a_n b_n}.$$

Therefore,

$$\begin{aligned}\mathbb{P}(|X_j|/n > \delta \text{ for some } j \leqslant n) \to 0 &\Leftrightarrow n\mathbb{P}(|X_1| > \delta_n) \to 0 \text{ for all } \delta \\ &\Leftrightarrow \delta n\mathbb{P}(|X_1| > \delta_n) \to 0 \text{ for all } \delta \\ &\Leftrightarrow x\mathbb{P}(|X_1| > x) \to 0 \text{ as } x \to \infty.\end{aligned}$$

When is there $\{\mu_n\}$ with $\mathbb{P}(|S_n/n - \mu_n| > \epsilon) \to 0$ for all ϵ?

Truncation

A **truncation** of X is $\overline{X} = X1_{\{|X|\leqslant M\}}$ for some M, so in particular it is bounded. For some proofs about S_n/n, below is a roadmap:

- prove the result for $\overline{S} = \overline{X}_1 + \ldots + \overline{X}_n$,
- show $S_n - \overline{S}_n$ is small, e.g., $\mathbb{P}(S_n - \overline{S}_n) \to 0$ or $\mathbb{E}[(S_n - \overline{S}_n)^2] \to 0$.

The Weak Law of Large Numbers

Theorem: WLLN, D2.2.12

Let $X_1, X_2, \ldots$ be i.i.d. In order that there exists $\{\mu_n\}$ such that $S_n/n - \mu_n$ in probability, it is necessary and sufficient that

$$x\mathbb{P}(|X_1| > x) \to 0 \qquad \text{as} \qquad x \to \infty.$$

If so, $\mu_n = \mathbb{E}[X_1 1_{\{|X_1|\leqslant n\}}]$ works.

Proof. We prove the sufficiency part only; the necessity part is beyond the scope even in Durrett's book.

- We first truncate the variables and define $X_{n,k} := X_k 1_{\{|X_k| \leqslant n\}}$. Let $S'_n = \sum_{k=1}^{n} X_{n,k}$.
- We show truncation "does little:"

$$\begin{aligned}\mathbb{P}(S'_n \neq S_n) &= \mathbb{P}(|X_k| > n \text{ for some } k \leqslant n)\\ &\leqslant n\mathbb{P}(|X_1| > n) \to 0 \text{ by union bound.}\end{aligned}$$

- We show the theorem holds for truncated random variables: by Chebyshev,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S'_n}{n} - \mu_n\right| > \epsilon\right) &\leqslant \frac{\operatorname{var}(S'_n/n)}{\epsilon^2} = \frac{\operatorname{var}(X_{n,1})}{\epsilon^2 n} \leqslant \frac{\mathbb{E}X_{n,1}^2}{\epsilon^2 n}\\ &= \epsilon^{-2} n^{-1} \int_0^\infty 2y\mathbb{P}(|X_{n,1}| > y)\,\mathrm{d}y\\ &\leqslant \epsilon^{-2} \underbrace{n^{-1} \int_0^n 2y\mathbb{P}(|X_1| > y)\,\mathrm{d}y}_{\text{average of } 2y\mathbb{P}(|X_1|>y) \text{ on } [0,n]} \to 0.\end{aligned}$$

- Combine and QED:

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu_n\right| > \epsilon\right) \leqslant \mathbb{P}(S_n \neq S'_n) + \mathbb{P}\left(\left|\frac{S'_n}{n} - \mu_n\right| > \epsilon\right) \to 0.$$ □

Given a random variable, we consider the **standardized random variable** $(X - \mathbb{E}X)/\sigma_X$ whenever this makes sense. For $b > \sigma_X$,

$$\mathbb{P}\left(\left|\frac{E - \mathbb{E}X}{b}\right| > \epsilon\right) = \mathbb{P}\left(\left|\frac{X - \mathbb{E}X}{\sigma_X}\right| > \frac{\epsilon b}{\sigma_X}\right) \leqslant \frac{\operatorname{var}((X - \mathbb{E}X)/\sigma_X)}{\epsilon^2 b^2/\sigma_X^2} = \frac{\sigma_X^2}{\epsilon^2 b^2},$$

which is small for $b \gg \sigma_X$. This proves the following theorem:

Theorem: D2.2.6

Let $\{T_n\}$ be random variables. If $\operatorname{var}(T_n)/b_n^2 \to 0$, then $\dfrac{T_n - \mathbb{E}T_n}{b_n} \to 0$ in probability.

Beginning of Sept.92022

Consider a geometric distribution with parameter p:

- $\mathbb{P}(X = n) = (1-p)^{n-1}p$.
- $\mathbb{E}X = 1/p$.
- $\mathbb{E}(X(X-1)) = \sum_{n=1}^{\infty} n(n-1)(1-p)^{n-2}(1-p)p = \dfrac{2-2p}{p^2}$, so
- $\operatorname{var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \dfrac{2-2p}{p^2} + \dfrac{1}{p} - \dfrac{1}{p^2} = \dfrac{1}{p^2} - \dfrac{1}{p}$.

Example: The coupon collector's problem. Suppose each cereal box has one of the n coupons equally likely. Let T_n be the time to get all n.

Let R be repeats and N be new coupons. The outcome is a sequence of R's and N's. Let $X_{n,k}$ be the time from getting the $(k-1)^{\text{th}}$ coupon to the k^{th} new coupon. It follows immediately that the $X_{n,k}$'s are independent from each other, with $T_n = \sum_{k=1}^{n} X_{n,k}$. In particular,

$$X_{n,k} \sim \text{geometric}\left(\frac{n-k+1}{n}\right),$$

so

$$\mathbb{E}T_n = 1 + \frac{n}{n-1} + \frac{n}{n-2} + \ldots + n = n(1 + 1/2 + \ldots + 1/n) \sim n \log n.$$

On the other hand,

$$\text{var}(T_n) = \sum_{k=1}^{n} \text{var}(X_{n,k}) \leqslant \sum_{k=1}^{n} \left(\frac{n}{n-k+1}\right)^2 = \frac{n^2\pi^2}{6}.$$

Since $\text{var}(T_n)/(n \log n)^2 \to 0$, by D2.2.6, $(T_n - \mathbb{E}T_n)/(n \log n) \to 0$ in probability, i.e.,

$$\frac{T_n}{n \log n} \to 1 \qquad \text{in probability.}$$

2.3 Triangular Arrays

Consider a **triangular array** $\{X_{n,k} : n \geqslant k, k \leqslant k_n\}$ where the n^{th} row has k_n variables.

Theorem: D2.2.11, WLLN for triangular arrays

Let $\{X_{n,k}\}$ be given. Let $b_n \to \infty$ and

$$a_n := \sum_{k=1}^{k_n} \mathbb{E}(X_{n,k} 1_{\{|X_{n,k}| \leqslant b_n\}}).$$

Assume

$$\sum_{k=1}^{k_n} \mathbb{P}(|X_{n,k}| > b_n) \to 0 \qquad \text{as } n \to \infty \tag{1}$$

and

$$b_n^{-2} \mathbb{E}(X_{n,k}^2 1_{\{|X_{n,k}| \leqslant b_n\}}) \to 0 \qquad \text{as } n \to \infty, \tag{2}$$

then $(S_n - a_n)/b_n$ converges to 0 in probability.

In the i.i.d. case, where $X_{n,k} = X_k$ and $k_n = b_n = n$, (1) says $n\mathbb{P}(|X_1| > n) \to 0$ and (2) says $n^{-1}\mathbb{E}(X_1^2 1_{\{|X_1| \leqslant n\}}) \to 0$.

Theorem: D2.2.14, Finite mean of WLLN

Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}X_1 = \mu$. Then $S_n/n \to \mu$ in probability without any assumption on the second moment.

Proof. We use WLLN 2.2.12. Let $\mu_n := \mathbb{E}(X_1 1_{\{|X_1| \leqslant n\}})$. We know $\mu_n \to \mu$ by DCT. Also,

$$x\mathbb{P}(|X_1| > x) = \mathbb{E}(x 1_{\{|X_1| > x\}}) = \mathbb{E}(|X_1| 1_{\{|X_1| > x\}}) \to 0$$

again using DCT. Therefore by 2.2.12 $S_n/n - \mu_n \to 0$ in probability, so $S_n/n \to \mu$ in probability. □

If $X_1 \geqslant 0, \mathbb{E}X_1 = \infty$, we can compare X_1 with the truncated variables to see $S_n/n \to \infty$. Nevertheless, we can still ask if there exist a_n, b_n such that $(S_n - a_n)/b_n \to 0$ in probability.

Example: D2.2.16 St. Petersburg paradox. Game: win 2^j if first heads toss is trial j, $j \geqslant 1$. Note that $S_n/n \to \mu$ implies that μ is the "fair price" to pay to play one game. Let X_k be the r.v. describing the among of games won by game k. Then

$$\mathbb{E}X_1 = \sum_{j \geqslant 1} 2^j 2^{-j} = o.$$

Then a_n is the "fair price for n games." By 2.2.11 (triangular array WLLN), we take $X_{n,k} = X_k$ for $k \leqslant n$ and $\{b_n\}$ to be determined. Let

$$a_n = n\mathbb{E}(X_1 1_{\{X_1 \leqslant b_n\}}).$$

We want b_n to satisfy two things:

- the truncation probability $n\mathbb{P}(X_1 > b_n) \to 0$,
- $b_n^{-2} n\mathbb{E}(X_1^2 1_{\{X_1 \leqslant b_n\}}) \to 0$, and
- $b_n \leqslant ca_n$.

For tails:

$$\mathbb{P}(X_1 \geqslant 2^m) = \mathbb{P}(\text{first } m-1 \text{ all tails}) = 2^{-m+1}.$$

Example: D2.2.16 St. Petersburg paradox.
Consider $X_1, X_2, \ldots$ i.i.d. with $X_1 = 2^{-j}$ with probability 2^{-j}. Then $\mathbb{E}X_1 = \infty$. Treat this as a game, but the paradox is the expected value is infinite and we cannot play an infinite amount of times. The question: how much we should pay to play this game n times?
We construct a_n, b_n such that

- $n\mathbb{P}(X_1 > b_n) \to 0$,
- $b_n^{-2} n\mathbb{E}(X_1^2 1_{\{X_1 \leqslant b_n\}}) \to 0$, and
- $b_n \leqslant ca_n$.

If so, by WLLN (2.2.11), $(S_n - a_n)/b_n \to 0$ in probability.
From 2.2.11, we simply pick

$$a_n = n\mathbb{E}(X_1 1_{\{X_1 \leqslant b_n\}})$$

and $P(X_1 \geqslant 2^m) = 2^{-m+1}$. We take b_n of form $2^{m(n)}$.
In order for the first condition to satisfied, $n2^{-m(n)+1} \to 0$ implies the candidate $m(n) = \log_2 n + K(n)$ with $K(n) \to \infty$. Then $2^{m(n)} = n2^{-K(n)}$. For the truncation condition,

$$\mathbb{E}(X_1^2 1_{X_1 \leqslant 2^{m(n)}}) = \sum_{j=1}^{m(n)} 2^{2j}\mathbb{P}(X_1 = 2^j) = 2^{m(n)+1}.$$

Therefore $b_n^{-2} n\mathbb{E}(X_1^2 1_{\{X_1 \leqslant b_n\}}) = 2^{-2m(n)} n 2^{m(n)+1} = 2^{-K(n)+1}$. Letting $n \to \infty$ this term does converge to 0.

Finally, to check the third condition, we want

$$a_n = n\mathbb{E}(X_1 1_{\{X_1 \leqslant b_n\}}) = n \sum_{j=1}^{m(n)} 2^j \mathbb{P}(X_1 = 2^j) = nm(n).$$

That is, we want

$$\frac{a_n}{b_n} = \frac{nm(n)}{2^{m(n)}} = \frac{m(n)}{2^{K(n)}} = \frac{\log_2(n) + K(n)}{2^{K(n)}}.$$

If we take $K(n) = \log_2 \log_2 n$ then the fraction will converge to 1, and we are finally done:

$$\frac{S_n - n(\log_2 n + \log_2 \log_2 n)}{n \log_2 n} \to 0 \qquad \text{in probability.}$$

Since $\log_2 \log_2 n / \log_2 n \to 0$, we have

$$\frac{S_n - n \log_2 n}{n \log_2 n} \to 0 \qquad \text{in probability,}$$

so

$$\frac{S_n}{n \log_2 n} \to 1 \qquad \text{in probability.}$$

That is, a fair price for playing n games is paying $\log_2 n$ per play.

2.4 Borel-Cantelli Lemmas

Some very quick recap: if $\{A_n\}$ are subsets of Ω, then

$$\limsup A_n = \bigcap_{m \geqslant 1} \bigcup_{n \geqslant m} A_n = \{\omega : \omega \in A_n \text{ i.o. (infinitely often)}\}$$

and

$$\liminf A_n = \bigcup_{m \geqslant 1} \bigcap_{n \geqslant m} A_n = \{\omega : \omega \in A_n \text{eventually / for all but finitely many } A_n\}$$

Theorem: First Borel-Cantelli Lemma

Let $\{A_n\}$ be events with $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. Then $\mathbb{P}(\limsup A_n) =: \mathbb{P}(A_n \text{ i.o.}) = 0$.

Proof. For all m, $\mathbb{P}(A_n$ i.o.) has to occur after m, so

$$\mathbb{P}(A_n \text{ i.o.}) \leqslant \mathbb{P}(\bigcup_{n \geqslant m} A_n) \leqslant \sum_{n \geqslant m} \mathbb{P}(A_n) \to 0 \qquad \text{as } m \to \infty.$$ □

The converse does not hold, as illustrated by $A_n = (0, 1/n)$ on the unit interval equipped with the uniform probability. With independence of events, however, we have the following result:

Theorem: Second Borel-Cantelli Lemma

Let $\{A_n\}$ be independent with $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. Then $\mathbb{P}(A_n \text{ i.o.}) = 1$.

We first need a lemma: if $0 < a_n < 1$, $\prod_{i=1}^{\infty}(1-a_n) > 0$ if and only if $\sum_{n=1} x_i < \infty$. When x is small, $e^{-2x} < 1 - x < e^{-x}$ and we obtain the claim after some algebra.

Proof of Borel-Cantelli. Fix m. Using continuity of probability on decreasing events,

$$\begin{aligned}\mathbb{P}(\text{no } A_n\text{'s after some index } m) &= \lim_{k\to\infty} \mathbb{P}(\text{no } A_n\text{'s from index } m \text{ to } k) \\ &= \lim_{k\to\infty} \prod_{n=m}^{k} (1-\mathbb{P}(A_n)) = \prod_{n\geqslant m} (1-\mathbb{P}(A_n)) = 0.\end{aligned}$$

Therefore, $\mathbb{P}(\text{some } A_n \text{ after index } m, \text{ for all } m) = 1$. □

Example. Let $X_1, X_2, \dots$ be i.i.d. exponential with parameter λ, i.e., with density $\lambda e^{-\lambda x}$ on $[0, \infty)$. Goal: find $\{c_n\}$ with $\limsup X_n/c_n = 1$ a.s.; we call c_n the max growth rate of X_n. That is, for all ϵ, we want

$$\mathbb{P}(X_n/c_n > 1+\epsilon \text{ i.o.}) = 0 \qquad \text{and} \qquad \mathbb{P}(X_n/c_n > 1-\epsilon \text{ i.o.}) = 1.$$

By first and second B-C, it is sufficient to show that

$$\sum_{n\geqslant 1} \mathbb{P}(X_n > (1+\epsilon)c_n) = \sum_{n\geqslant 1} e^{-\lambda(1+\epsilon)c_n} < \infty$$

and

$$\sum_{n\geqslant 1} \mathbb{P}(X_n > (1-\epsilon)c_n) = \sum_{n\geqslant 1} e^{-\lambda(1-\epsilon)c_n} = \infty$$

for all ϵ. We let c_n be such that $e^{-\lambda c_n} = 1/n$, i.e., $c_n = \log n/\lambda$. And this works.

Beginning of Sept. 26, 2022

Some quick recap of convergence a.s. and in probability:

- If $X_n \to X$ a.s. then $1_{\{|X_n - X| > \epsilon\}} \to 0$ a.s. for all $\epsilon > 0$, so

$$\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{E}1_{\{|X_n - X|>\epsilon\}} \to 0$$

  by bounded convergence theorem, so $X_n \to X$ in probability.

- The converse is false, as illustrated by the scanning intervals. Let $\mathbb{P}$ be uniform on $[0,1]$ and consider $[0,1], [0,1/2], [1/2,1], [0,1/3], [1/3,2/3], [2/3,1]$, and so on.

However, the following does hold:

Proposition

If $X_n \to X$ in probability then there exists a subsequence $X_{n_k} \to X$ a.s.

Proof. Take a sequence of increasing indices n_k such that

$$\mathbb{P}(|X_{n_k} - X| > 1/k) < 2^{-k}.$$

Using Borel-Cantelli,

$$\sum_k \mathbb{P}(|X_{n_k} - X| > 1/k) < \infty$$

so $\mathbb{P}(|X_{n_k} - X| > 1/k \text{ i.o.}) = 0$ and $X_{n_k} \to X$ a.s. □

Recall from analysis that in a metric space, $y_n \to y$ iff every subsequence y_{n_k} has a further subsequence converging to y. Using this fact we obtain a stronger characterization of convergence in probability:

Theorem: D2.3.2

$X_n \to X$ in probability iff for every subsequence $\{X_{n_k}\}$ there exists a further subsequence converging a.s. to X.

In particular, this theorem implies that there is no metric $d(X,Y)$ such that $X_n \to X$ a.s. iff $d(X_n, X) \to 0$. On the other hand, $d(X,Y) := \mathbb{E}(|Y-X|)/(1+|Y-X|)$ satisfies $X_n \to X$ in probability iff $d(X_n, X) \to 0$. More generally, any g bounded, invertible, concave, with $g(0) = 0$ works, like $g(t) = t/(1+t)$.

Proof. The forward direction follows from the previous proposition.

Conversely, fix $\epsilon > 0$ and let $y_n = \mathbb{P}(|X_n - X| > \epsilon)$. The assumption implies that for any $\{y_{n_k}\}$, there exists a further subsequence $\{y_{n_{k(\ell)}}\}$ converging to 0. Using the previous remark, $y_n \to 0$, i.e., $X_n \to X$ a.s. □

Corollary: D2.3.4

If $X_n \to X$ in probability and $f : \mathbb{R} \to \mathbb{R}$ is continuous, then $f(X_n) \to f(X)$ in probability.

Proof. We use the previous theorem twice. For all $\{X_{n_k}\}$ there exists a further subsequence $\{X_{n_{k(\ell)}}\}$ converging to X a.s., and by continuity $f(X_{n_{k(\ell)}}) \to f(X)$ a.s. Now using D2.3.2 again, $f(X_n) \to f(X)$ in probability. □

Theorem: D2.3.8

Suppose $X_1, X_2, \ldots$ are i.i.d. with $\mathbb{E}|X_1| = \infty$. Then

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists and is finite}\right) = 0.$$

Proof. We first show that $\mathbb{P}(|X_n|/n > 1 \text{ i.o.}) = 1$ and that

$$\mathbb{P}\left(\left|\frac{S_{n+1}}{n+1} - \frac{S_n}{n}\right| > \frac{1}{2} \text{ i.o.}\right) = 1.$$

- For the first claim:

$$\infty = \mathbb{E}|X_1| = \int_0^\infty \mathbb{P}(|X_1| > n)\, dx \leqslant \sum_{n\geqslant 1} \mathbb{P}(|X_1| > n) = \sum_{n \geqslant 1} \mathbb{P}(|X_n| > n)$$

  so by the second B-C, $\mathbb{P}(|X_n| > n \text{ i.o.}) = 1$ and in particular $\mathbb{P}(|X_n|/n > 1 \text{ i.o.}) = 1$.

- To show the second claim, define

$$C := \{\omega : \lim_{n\to\infty} S_n/n \text{ exists and is finite}\}.$$

Note that

$$-\frac{S_{n+1}}{n+1}+\frac{S_n}{n}=\frac{S_n}{n}-\frac{S_n}{n+1}+\frac{X_{n+1}}{n+1}.$$

Therefore, for a.e. $\omega \in C$, $S_n/(n(n+1)) \to 0$ (since S_n/n is finite) and $|X_{n+1}|/(n+1) > 1$ i.o., so

$$\left|\frac{S_n}{n}-\frac{S_{n+1}}{n+1}\right| > \frac{1}{2} \text{ i.o.}$$

- Therefore $\mathbb{P}(C) = \mathbb{P}\left(C \cap \left\{\left|\frac{S_n}{n}-\frac{S_{n+1}}{n+1}\right| > \frac{1}{2} \text{ i.o.}\right\}\right) = 0$ and we are done.

□

2.5 Kolmogorov 0-1 Law

Previously, we have shown:

(1) If A_n's are independent then $\mathbb{P}(A_n \text{ i.o.}) = 0$ or 1 by the first and/or second B-C.

(2) We have also shown that if $X_1, X_2, \ldots$ are i.i.d. exponential r.v.'s with parameter λ then $\limsup_{n\to\infty} X_n/\log n = 1/\lambda$ a.s., so

$$\mathbb{P}\left(\limsup_{n\to\infty}\frac{X_n}{\log n} > c\right) = \begin{cases} 1 & \text{if } c < 1/\lambda \\ 0 & \text{if } c > 1/\lambda. \end{cases}$$

Beginning of Sept. 29, 2022

We define $F_n := \sigma(X_n, X_{n+1}, \ldots)$ for each n, and we define $\mathcal{T} := \bigcap_{n\geqslant 1} \mathcal{F}_n$, the **tail σ-field.**

Example: $\{S_n/n \to \mu\}$ is a tail event. To see this, we fix $m < n$ and get

$$\frac{S_n}{n} = \frac{S_m}{n} + \frac{X_{m+1}+\ldots+X_n}{n}.$$

Now letting $n \to \infty$ we see $S_n/n \to \mu$ iff $(X_{m+1}+\ldots+X_n)/n \to \mu$. In particular, the right side quantity does not depend on which m we start with.

Theorem: Kolgomorov 0-1 Law

Let $X_1, X_2, \ldots$ be random variables. Then $\mathbb{P}(A) = 0$ or 1 for $A \in \mathcal{T}$.

Proof. Idea: it suffices to show that if $A \in \mathcal{T}$ then $\mathbb{P}$ is independent of itself, i.e., $\mathbb{P}(A \cap A) = \mathbb{P}(A)^2$. In fact, we'll show $A \perp B$ for every event $B \in \sigma(X_1, X_2, \ldots)$.

Preliminaries. For events A, B, we define a distance $d(A, B) := \mathbb{P}(A\Delta B) = \mathbb{E}|1_A - 1_B|$. (This is a pseudo-metric but can be 0 when $A \neq B$.) An important property: $d(A\cup B, G\cup H) = d(A, G) + d(B, H)$. Same for intersections. More formally, given $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space and $\mathcal{G}$ a collection of events, we say $A \in \mathcal{F}$ is **approximable** by $\mathcal{G}$ if for every $\epsilon > 0$, there exists $G \in \mathcal{G}$ with $d(A, G) < \epsilon$.

Idea, continued: we approximate any B by $\tilde{B} \in \sigma(X_1, \ldots, X_n)$ for some n w.r.t. our distance defined above. We also approximate A by some $\tilde{A} \in \sigma(X_{n+1}, X_{n+2}, \ldots)$. Note that these two σ-fields are indeed independent, and $\tilde{A}$ and $\tilde{B}$ are independent. Intuitively, this gives

$$\mathbb{P}(A\cap B) \approx \mathbb{P}(\tilde{A}\cap\tilde{B}) = \mathbb{P}(\tilde{A})\mathbb{P}(\tilde{B}) \approx \mathbb{P}(A)\mathbb{P}(B).$$

Lemma. If $\mathcal{G}$ is a field then {all events approximable by $\mathcal{G}$} is a σ-field.

Proof of subclaim. Closedness under countable union is immediate by using $\epsilon/2^{-n}$ along with the fact that $\mathcal{G}$ is a field. If $A_1, A_2, \ldots$ are approxmaeble by $G_1, G_2, \ldots$ with errors $< \epsilon/2^n$, we have

$$\mathbb{P}\left((\bigcup_{n\geqslant 1} G_n)\Delta(\bigcup_{n\geqslant 1} A_n)\right) \leqslant \mathbb{P}\left((\bigcup_{n\geqslant 1} G_n)\Delta(\bigcup_{n\geqslant 1}^{k} A_n)\right) + \mathbb{P}(\bigcup_{n\geqslant 1} A_n) - \mathbb{P}(\bigcup_{n=1}^{k} A_n) \to \sum_{n=1}^{k} \epsilon 2^{-n} < \epsilon \tag{*}$$

as $k \to \infty$. END OF CLAIM OF LEMMA.

Now we prove the Kolgomorov 0-1 law. We apply the lemma to $\mathcal{G} = \bigcup_{n\geqslant 1} \sigma(X_1, \ldots, X_n)$ which is a field. We approximate $B \in \sigma(X_1, X_2, \ldots)$ by $\tilde{B} \in \sigma(X_1, \ldots, X_n)$ for some n with error $< \epsilon$. For $A \in \mathcal{T}$, we apply the lemma to $\bigcup_{k\geqslant n} \sigma(X_{n+1}, \ldots X_k)$ and get $\tilde{A} \in \sigma(X_{n+1}, \ldots, X_{n_k})$, also with error $< \epsilon$. Then by (*) we are done! □

Using Kolgomorov 0-1 law on $A = \{S_n/n \to \mu\}$ can be approximated in $\sigma(X_1, \ldots, X_m)$, even though it doesn't depend on $X_1, \ldots, X_m$ for any given m, as shown before stating the theorem. The approximation will be by something like

$$\tilde{A} = \{|S_m/m - \mu| < \epsilon \text{ for all } m \in [k, n]\}$$

for some n, k, m.

Related Result

We consider permutable events. A **finite permutation** π of $\mathbb{N}$ is one with $\pi(i) = i$ for all but finitely many i's. Here we have $\Omega = \mathbb{R}^{\mathbb{N}}$ and $X_1, X_2, \ldots$ random variables (coordinates in Ω). Let $\omega = (\omega_i, i \geqslant 1)$. Consider $\pi\omega = (\omega_{\pi(1)}, \omega_{\pi(2)}, \ldots)$, i.e., $(\pi\omega) - i = \omega_{\pi(i)}$. We say event A is **permutable** if $\pi^{-1}A = A$ for every finite permutation π, and we let $\mathcal{E}$ be the collection of all permutable events.

It is easy to check that $\mathcal{E}$ is a σ-field. Also, if $A \in \sigma(X_{n+1}, X_{n+2}, \ldots)$, then the occurrence of A is unaffected by the permutation of $X_1, \ldots, X_n$. In particular, any $A \in \mathcal{J}$ is permutable, so $\mathcal{T} \subset \mathcal{E}$.

An example of permutable sets: $\{S_n \in B \text{ i.o.}\}$: if the sum is in B then mixing the first (finitely many) coordinates does not change the fact that S_n is still in B. However, $\{S_n \in B \text{ i.o.}\}$ is not a tail event: if we change the value of $X_1(\omega)$ dramatically, every $S_n(\omega)$ will be affected.

Theorem: Hewitt-Savage 0-1 Law

If $X_1, X_2, \ldots$ are i.i.d. and $A \in \mathcal{E}$ then $\mathbb{P}(A) = 0$ or 1.

2.6 Strong Law of Large Numbers

Beginning of Sept. 30, 2022

Theorem: WLLN, D2.4.1

Let $X_1, X_2, \ldots$ be i.i.d. (pairwise in fact suffice) with $\mathbb{E}|X_1| < \infty$. Then $S_n/n \to \mu = \mathbb{E}X_1$ almost surely.

Proof. Idea: we assume $X_1 \geqslant 0$ or otherwise we use $X = X^+ - X^-$. Then S_n and n are both increasing in n. Consider a subsequence, say $k(n) = \lfloor \alpha^n \rfloor$ with $\alpha > 1$ but close to 1. For the indices in between the subsequences, i.e., for $k(n) \leqslant m \leqslant k(n+1)$,

$$\frac{S_{k(n)}}{k(n)}\frac{k(n)}{k(n+1)} = \frac{S_{k(n)}}{k(n+1)} \leqslant \frac{S_m}{m} \leqslant \frac{S_{k(n+1)}}{k(n)} = \frac{S_{k(n+1)}}{k(n+1)}\frac{k(n+1)}{k(n)}.$$

As $n \to \infty$, $k(n)/k(n+1) \to 1/\alpha$ and $k(n+1)/k(n) \to \alpha$. Therefore if we show show convergence of the subsequence $S_{k(n)}/k(n) \to \mu$, then

$$\frac{\mu}{\alpha} \leqslant \liminf_{m\to\infty} \frac{S_m}{m} \leqslant \limsup_{m\to\infty} \frac{S_m}{m} \leqslant \alpha\mu,$$

and since α is arbitrary, we are done.

Proof of SLLN. **Step 1**. We truncate as usual: let $Y_n = X_n 1_{\{X_n \leqslant n\}}$ and let $T_n = \sum_{i=1}^{n} Y_i$. Then

$$\sum_{n\geqslant 1} \mathbb{P}(X_n \neq Y_n) = \sum_{n\geqslant 1} \mathbb{P}(X_n > n) = \sum_{n\geqslant 1} \mathbb{P}(X_1 > n) < \infty$$

since $\mathbb{E}X_1 < \infty$. Therefore, by B-C, $\mathbb{P}(X_n \neq Y_n \text{ i.o.}) = 0$. Since there are only finite number of different terms between S_n and T_n, $(S_n/n) - (T_n/n) \to 0$. Therefore it suffices to show $T_n/n \to \mu$ almost surely.

Step 2. We apply B-C to T_n/n. Using Chebyshev,

$$\mathbb{P}\left(\left|\frac{T_k - \mathbb{E}T_k}{k}\right| > \epsilon\right) \leqslant \frac{\operatorname{var}(T_k)}{k^2\epsilon^2} = \frac{1}{k^2\epsilon^2}\sum_{i=1}^{k} \operatorname{var}(Y_i).$$

But $\operatorname{var}(Y_i)$ may not $\to 0$. and then the terms on the RHS is bounded from below by some constant divided by k, not summable. Remedy:

Step 3. Apply step 2 to a subsequence $k(n) = \lfloor \alpha^n \rfloor \geqslant \alpha^n/2$. Then

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{|T_{k(n)} - \mathbb{E}T_{k(n)}|}{k(n)} > \epsilon\right) &\leqslant \sum_{n=1}^{\infty} \frac{1}{\epsilon k(n)^2} \sum_{i=1}^{k(n)} \operatorname{var}(Y_i) \\
&= \sum_{i=1}^{\infty} \epsilon^{-2} \operatorname{var}(Y_i) \sum_{k(n)\geqslant i} \frac{1}{k(n)^2} \\
&\leqslant \sum_{i=1}^{\infty} 4\epsilon^{-2} \operatorname{var}(Y_i) \sum_{k(n)\geqslant i} \alpha^{-2n} \\
&\leqslant \sum_{j=1}^{\infty} 4\epsilon^{-2} \operatorname{var}(Y_j) \frac{1}{j^2}\frac{1}{1-\alpha^{-2}} \\
&= \frac{4\epsilon^{-2}}{1-\alpha^{-2}} \sum_{j=1}^{\infty} \frac{\operatorname{var}(Y_j)}{j^2} \leqslant \frac{4\epsilon^{-2}}{1-\alpha^{-2}} \sum_{j=1}^{\infty} \frac{\mathbb{E}Y_j^2}{j^2} \qquad (*)
\end{aligned}$$

Since

$$\mathbb{E}Y_j^2 = \int_0^{\infty} 2y\mathbb{P}(Y_j > y)\,\mathrm{d}y \leqslant \int_0^{j} 2y\mathbb{P}(X_1 > y)\,\mathrm{d}y,$$

the sum in (*) becomes

$$\begin{aligned}
\sum_{j=1}^{\infty} \frac{\mathbb{E}Y_j^2}{j^2} &= \sum_{j=1}^{\infty} j^{-2} \int_0^{\infty} 1_{\{Y<j\}} 2y\mathbb{P}(X_1 > y)\,\mathrm{d}y \\
&= \int_0^{\infty} \Big(\sum_{j>y} j^{-2}\Big) 2y \underbrace{\mathbb{P}(X_1 > y)\,\mathrm{d}y}_{\text{integrable}} \qquad (**)
\end{aligned}$$

Since $\sum_{j>y} j^{-2} \approx y^{-1}$, it can be shown that (D2.4.4)

$$\Big(\sum_{j>y} j^{-2}\Big) 2y \leqslant 4 \qquad \text{for all } y.$$

Hence (**) $\leqslant 4\mathbb{E}X_1 < \infty$. Then (*) and B-C says

$$\mathbb{P}\left(\frac{|T_{k(n)} - \mathbb{E}T_{k(n)}|}{k(n)} > \epsilon \text{ i.o.}\right) = 0 \qquad \text{for all } \epsilon,$$

so

$$\frac{T_{k(n)} - \mathbb{E}T_{k(n)}}{k(n)} \to 0 \text{ a.s.} \qquad \text{and} \qquad \frac{T_{k(n)}}{k(n)} \text{ and } \frac{S_{k(n)}}{k(n)} \to \mu \text{ a.s.}$$

We have shown the a.s. convergence of a subsequence of $S_{k(n)}/k(n)$. By a remark made earlier we are done. □

Beginning of Oct. 3, 2022

Example: Renewal theory. Let $X_1, X_2, \ldots$ be i.i.d. with $0 < X_i < \infty$. Let $T_n = X_1 + \ldots + X_n$ and think of T_n as the time of n^{th} occurrence of some event. Let $N_t := \sup\{n : T_n \leqslant t\}$. Think of X_i as the lifespans of light bulbs and a person replaces a light bulb right when it burns out. Then N_t is the number of light bulbs that have burnt out by time t.

Theorem: D2.4.7

If $\mathbb{E}X_1 = \mu < \infty$ and $X_1, X_2, \ldots$ are i.i.d. then

$$N_t/t \to 1/\mu \text{ a.s.}$$

Proof. Let $T(N_t)$ be the time of last renewal up to time t. Then $T(N_t) \leqslant t < T(N_t + 1)$, so

$$\frac{T(N_t)}{N_t} \leqslant \frac{t}{N_t} < \frac{T(N_t+1)}{N_t} = \frac{T(N_t+1)}{N_t+1}\frac{N_t+1}{N_t}.$$

Since $T(N_t)/N_t \to \mu$ a.s. and $(N_t+1)/N_t \to 1$, we have $t/N_t \to \mu$ a.s. □

SLLN when $\mathbb{E}X_1 = \infty$: we know

$$\frac{1}{n}\sum_{i=1}^{n} \min(X_i, M) \to \mathbb{E}\min(X_i, M) \text{ a.s.}$$

Since $\mathbb{E}\min(X_i, M) \to \mathbb{E}X_1 = \infty$ as $M \to \infty$, we also have

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \infty \text{ a.s.}$$

Example: Empirical d.f.'s, D2.4.8. Let $X_1, X_2, \ldots$ be i.i.d. with distribution F. We let

$$F_n(x) := \frac{1}{n}\sum_{m=1}^{n} 1_{\{X_m \leqslant \frac{x}{\}}}.$$

Namely, $F_n(x)$ is the observed frequency of values that are $\leqslant x$. For fixed x, $1_{\{X_i \leqslant x\}}$ are i.i.d. with mean $F(x)$, so SLLN says $F_n(x) \to F(x)$. Simiilarly $F_n(x-) \to F(x-)$ a.s.

Theorem: Gilvenko-Cantelli, D2,4,9

We have "almost sure" uniform convergence:

$$\sup_x |F_n(x) - F(x)| \to 0 \text{ a.s.}$$

Proof. *Idea*: if F_n is close to F at two points a, b where $F(b) - F(a)$ is small, then F_n is close to F in all $[a, b]$ by monotonicity.

Fix $k \geqslant 1$. We let

$$I_j := \{x : 1/k \leqslant F(x) \leqslant (k+1)/k\}, 0 \leqslant i \leqslant k.$$

This is either an empty set or an interval, so say $I_j = [a_j, b_j]$. If $I_j \neq \varnothing$, then $F(a_j), F(b_j-)$ are in $[j/k, (j+1)/k]$. For all j, there exists $n_0(j)$ such that $n \geqslant n_0(j)$ implies (almost surely)

$$\begin{cases} |F_n(a_j) - F(a_j)| \leqslant 1/k \\ |F_n(b_j-) - F(b_j-)| \leqslant 1/k. \end{cases}$$

That is, on the endpoints, we have convergence.

What about in-between? For $x \in I_j$, we have

$$F_n(x) \geqslant F_n(a_j) \geqslant F_n(a_j) - \frac{1}{k} \geqslant \frac{j-1}{k} \geqslant F(x) - \frac{2}{k}.$$

The other direction is similar:

$$F_n(x) \leqslant F_n(b_j-) \leqslant F(b_j-) + \frac{1}{k} \leqslant \frac{j+2}{k} \leqslant F(x) + \frac{2}{k}.$$

Therefore the supremum is bounded by $2/k \to 0$, and we are done! □

An alternate proof of SLLN uses $k(n) = n^2$, where the goal is to bound

$$\mathbb{P}\left(\left|\frac{S_m - S_{k(n)}}{k(n)}\right| > \epsilon \text{ for some } k(n) \leqslant m \leqslant k(n+1)\right).$$

To do so, we need the following theorem:

Theorem: D2.5.5, Kolmogorov's maximal inequality

If $X_1, \ldots, X_n$ are independent (not requiring i.i.d.) with $\mathbb{E}X_i = 0$ and $\mathrm{var}(X_i) < \infty$. Then

$$\mathbb{P}(\max_{1 \leqslant k \leqslant n} |S_k| \geqslant x) \leqslant \frac{\mathrm{var}(S_n)}{x^2}.$$

Note that Chebyshev gives $\mathbb{P}(|S_k| \geqslant x) \leqslant \mathrm{var}(S_n)/x^2$ so this is strictly stronger.

Proof. We decompose the events according to the k^{th} occurrence:

$$A_k := \{|S_k| \geqslant x \text{ but } |S_j| < x \text{ for } j < k\}.$$

It is clear that the A_i's are disjoint. We show that $\mathrm{var}(S_n) = \mathbb{E}(S_n)^2 \geqslant x^2 \mathbb{P}(\max \geqslant x)$. For this:

$$\begin{aligned} \mathbb{E}S_n^2 &\geqslant \sum_{k=1}^n \int_{A_k} S_n^2 \, \mathrm{d}\mathbb{P} \\ &= \sum_{k=1}^n \int_{A_k} (S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2) \, \mathrm{d}\mathbb{P} \\ &\geqslant \sum_{k=1}^n \left[\int_{A_k} x^2 \, \mathrm{d}\mathbb{P} + \int_\Omega 1_{A_k} 2S_k(S_n - S_k) \, \mathrm{d}\mathbb{P}\right]. \end{aligned}$$

The first $\geqslant$ is because the A_k's are disjoint. We use x^2 as a lower bound for S_k over A_k by definition, and we note that $(S_n - S_k)^2 \geqslant 0$. Finally, we note that $1_{A_k} 2S_k$ depends only on what happens on the first k sets and $S_n - S_k$

depends on something else, so they are independent. Therefore

$$\int_{\Omega} 1_{A_k} 2S_k(S_n - s_K)\, d\mathbb{P}$$

the product of expected values, is the expected value of products, and $\mathbb{E}(S_n - S_k) = 0$. Therefore $\mathbb{E}S_n^2 \geqslant x^2 \sum_{k-1}^{n} \mathbb{P}(A_k) = x^2\mathbb{P}(\max|S_k| \geqslant x)$. □

Beginning of Oct. 5, 2022

Proposition: Ottaviani's inequality

Let $X_1, X_2, \ldots$ be independent and $a > 0$. Then

$$\mathbb{P}\left(\max_{j\leqslant n}|S_j| > 2a\right) \cdot \min_{j\leqslant n} \mathbb{P}(|S_n - S_k| \leqslant a) \leqslant \mathbb{P}(|S_n| > a).$$

To apply this theorem, suppose we know $\min_{j\leqslant n} \mathbb{P}(|S_n - S_k| \leqslant a) \leqslant c$, then

$$\mathbb{P}\left(\max_{j\leqslant n}|S_j| > 2a\right) \leqslant \frac{1}{c}\mathbb{P}(|S_n| > a).$$

Proof. We define

$$A_j = \{|S_i| \leqslant 2a \text{ for all } i \leqslant j \text{ and } |S_j| \geqslant 2a\}.$$

Then

$$\begin{aligned}
\mathbb{P}(|S_n| > a) &\geqslant \sum_{k=1}^{n} \mathbb{P}(|S_n| > a \text{ and } A_k) \\
&\geqslant \sum_{k=1}^{n} \mathbb{P}(|S_n - S_k| \leqslant a \text{ and } A_k) \\
&= \sum_{k=1}^{n} \mathbb{P}(|S_n - S_k| \leqslant a)\mathbb{P}(A_k) \\
&\geqslant \min_{k\leqslant n} \mathbb{P}(|S_n - S_k| \leqslant a) \cdot \mathbb{P}\left(\bigcup_{k\leqslant n} A_k\right) \\
&= \min_{k\leqslant n} \mathbb{P}(|S_n - S_k| \leqslant a)\mathbb{P}\left(\max_{k\leqslant n}|S_k| > 2a\right).
\end{aligned}$$

□

Example. If $x\mathbb{P}(|X_1| > x) \to 0$ but $\mathbb{E}|X_1| = \infty$, and if X and $-X$ have the same distribution (i.e., X is symmetric), then by weak law, $S_n/n \to$ the truncated mean in probability, which is always 0. However, since the mean is infinite, S_n/n will not converge to 0 almost surely.

Theorem: D2.5.8 Kolmogorov's three-series theorem

Let $X_1, X_2, \ldots$ be independent. Let $A > 0$ and let Y_i be $X_i 1_{\{|X_i|\leqslant A\}}$. Then $\sum_{n=1}^{\infty} X_i$ converges almost surely if and only if the following are all satisfied:

- $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A) < \infty$,
- $\sum_{n=1}^{\infty} \mathbb{E}Y_n$ converges, and
- $\sum_{n=1}^{\infty} \mathrm{var}(Y_n)$ converges.

Note that (i) A is chosen arbitrarily, so the three conditions cannot depend on A; (ii) if $\sum_{n=1}^{\infty} \mathrm{var}(Y_n) = \infty$, then $\mathrm{var}(\sum_{j=m}^{n} Y_n) \to \infty$ as $n \to \infty$, for all m, which implies the variate is staying big for the tail, and so we don't expect the tail of $\sum_{n=1}^{\infty} X_i$ to converge to 0.

2.7 Large Deviations

Let $X_1, X_2, \ldots$ be i.i.d. and let $S_n = X_1 + \ldots + X_n$, as usual. SLLN says if $\mathbb{E}|X_1| < \infty$ then $S_n/n \to \mu$ a.s. How big is $\mathbb{P}(S_n/n) > a)$, for some $a > \mu$?

Main idea: consider the **exponential moment**: $\varphi(t) = \mathbb{E}(\exp(tX)) < \infty$ for some t. Then $\mathbb{P}(S_n > na) \to 0$ decays exponentially.

Beginning of Oct. 7, 2022

For a sequence $b_n \to 0$ converging to 0, we say b_n decays like e^{-cn} if

$$-c = \lim \frac{1}{n} \log b_n,$$

or equivalently, for $\epsilon > 0$,

$$e^{-(c+\epsilon)n} \leqslant b_n \leqslant e^{-(c-\epsilon)n}$$

for sufficiently large n.

Similarly, we want to show that $\mathbb{P}(S_n/\mu > a)$ decays like $e^{-I(a)n}$ for some $I(a) > 0$. Question: what is

$$\gamma(a) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n/n > a)?$$

We define $\pi_a := \mathbb{P}(S_n/n \geqslant a)$. We claim that $\log \pi_n$ is **superadditive**: $\pi_{m+n} \geqslant \pi_m \pi_n$, so $\log \pi_{m+n} \geqslant \log \pi_n + \log \pi_m$. This is true because

$$\mathbb{P}(S_{m+n} \geqslant (m+n)a) \geqslant \mathbb{P}(S_n \geqslant na, S_{n+m} - S_n \geqslant ma) = \mathbb{P}(S_n \geqslant na, S_m \geqslant ma) = \mathbb{P}(S_n \geqslant na)\mathbb{P}(S_m \geqslant ma).$$

Lemma: D2.7.1

If γ_n is superadditive, then $\gamma/n \to \sup_m \gamma_m/m$.

Proof. Call the supremum limit c. It suffices to show $0 \leqslant \liminf \leqslant \limsup \leqslant c$.

$\limsup \leqslant c = \sup$ is trivial by definition.

Conversely, we need to show $\liminf \gamma_n/n \geqslant \gamma_m/m$ for all m. Induction says if $n = n_1 + \ldots + n_k$ then $\gamma_n \geqslant \gamma_{n_1} + \ldots + \gamma_{n_k}$.

In particular, if we fix m, then we can write n as $n = km + \ell$ with $0 \leqslant \ell < m$. Then,

$$\frac{\gamma_n}{n} \geqslant \frac{k\gamma_m + \gamma_\ell}{km + \ell} = \frac{km}{km + \ell}\frac{\gamma_m}{m} + \frac{\gamma_\ell}{km + \ell}.$$

As $n \to \infty$ so $k \to \infty$, ℓ is bounded, so $km/(km + \ell) \to 1$. So does $\gamma_\ell/(km + \ell)$. Therefore,

$$\liminf_{n\to\infty} \frac{\gamma_n}{n} \geqslant \frac{\gamma_m}{m}. \qquad \square$$

Therefore, $\mathbb{P}(S_n/n \leqslant a) \leqslant e^{\gamma(a)n}$ in particular, since $\gamma(a) \geqslant n^{-1}\log \mathbb{P}(S_n/n \geqslant a)$ as shown above. That is, this exponential decay rate is also an upper bound for $\mathbb{P}(S_n/n \leqslant a)$.

Suppose MGF $\varphi(\theta) = \mathbb{E}e^{\theta X}$ is finite in $(-\delta, \delta)$. In this interval,

$$\frac{X^k e^{\theta X}}{e^{(\theta+\epsilon)X}} = X^k e^{-\epsilon X} \to 0$$

as $X \to \infty$. In particular, if $\theta \in (-\delta, \delta)$, so does the new quantity when ϵ is small, so $\mathbb{E}(X^k e^{\theta X})$ is finite, for all k. In particular, for $k = 1$,

$$\lim_{h\to 0} \mathbb{E}\left(\frac{e^{(\theta+h)X} - e^{\theta X}}{h}\right) = \lim_{h\to 0} \mathbb{E}\left(e^{\theta X}\frac{e^{hX} - 1}{h}\right).$$

Assuming h positive,

$$\left|\frac{e^{hX} - 1}{h}\right| \leqslant \begin{cases} |X| & \text{if } X < 0 \\ Xe^{hX} & \text{if } X \geqslant 0 \end{cases}$$

We have shown that Xe^{hX} is integrable for small h, so indeed we can apply DCT and obtain $\varphi'(\theta) = \mathbb{E}(Xe^{\theta X})$. Similarly, if we differentiate twice, we obtain $\varphi''(\theta) = \mathbb{E}(X^2 e^{\theta X})$, and so on. Also,

$$(\log\varphi)'(\theta) = \frac{\varphi'(\theta)}{\varphi(\theta)} = \frac{\int Xe^{\theta X}\,\mathrm{d}\mathbb{P}}{\int e^{\theta X}\,\mathrm{d}\mathbb{P}}.$$

Given $g \geqslant 0$, we can define a probability measure by

$$\nu(A) = \frac{\int_A g\,\mathrm{d}\mathbb{P}}{\int g\,\mathrm{d}\mathbb{P}},$$

"$\mathbb{P}$ weighted by g", and equivalently

$$\mathbb{E}_\nu 1_A = \frac{\int 1_A g\,\mathrm{d}\mathbb{P}}{\int g\,\mathrm{d}\mathbb{P}}.$$

Using standard measure theory argument, we obtain

$$\mathbb{E}_\nu f = \frac{\int fg\,\mathrm{d}\mathbb{P}}{\int g\,\mathrm{d}\mathbb{P}}.$$

Therefore, $(\log\varphi)'(\theta)$ can be thought of as $\mathbb{E}_{\nu_\theta}X$, under the tilted distribution of ν_θ.
Also,

$$(\log\varphi)''(\theta) = \frac{\varphi(\theta)\varphi''(\theta) - \varphi'(\theta)^2}{\varphi(\theta)^2} = \frac{\int X^2 e^{\theta X}\,\mathrm{d}\mathbb{P}}{\int e^{\theta X}\,\mathrm{d}\mathbb{P}} - \left(\frac{\int Xe^{\theta X}\,\mathrm{d}\mathbb{P}}{\int e^{\theta X}\,\mathrm{d}\mathbb{P}}\right)^2,$$

namely $\mathrm{var}_{\nu_\theta}(X)$, which is nonnegative. Therefore, $\log\varphi$ is convex. Also note $(\log\varphi)(0) = 0$ with $(\log\varphi)'(0) = \mathbb{E}X$. What about MGF of sums S_n for i.i.d. random varaibles?

$$\varphi_{S_n}(\theta) = \mathbb{E}e^{\theta(X_1 + \ldots + X_n)} = \varphi(\theta)^n.$$

Now we fix $\theta > 0$. Then

$$\mathbb{P}(S_n/n > a) = \mathbb{P}(e^{\theta S_n} > e^{\theta na}),$$

so by Markov, this is bounded from above by

$$\mathbb{P}(S_n/n > a) \leqslant \frac{\mathbb{E}e^{\theta S_n}}{e^{\theta na}} = \frac{\varphi(\theta)^n}{e^{\theta na}} = \exp(-(a\theta - \log\varphi(\theta))n).$$

To show exponential decay, it suffices to show the above exponent is positive. In particular, if

$$I(a) := \sup_{\theta>0}(a\theta - \log\varphi(\theta)) > 0$$

we are done. Indeed! $a > \mathbb{E}X$, and $\log\varphi$ is convex, so if we start at the origin and draw a line $a\theta$, it is steeper than $\log\varphi$ so it will go above the graph of $\log\varphi$, resulting in a positive supremum.

Beginning of Oct. 17, 2022

Large Deviations Regime

We define $X_i = \pm 1$, $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$, and consider the trajectory $\{S_j, j \leqslant n\}$. We define a new probability measure $Q(\{S_j, j \leqslant n\})$ as

$$Q(\{S_j, j \leqslant n\}) := \frac{\mathbb{P}(\{S_j, j \leqslant n\})\exp(\beta N_n)}{Z_n(\beta)}$$

where N_n is the number of times the trajectory touches the x-axis and $Z_n(\beta)$ is the scaling constant to make Q a probability measure.

One can show that $\mathbb{P}(S_j = 0) \approx C/\sqrt{j}$, so $\mathbb{E}(N_n) = \sum_{j=1}^{n} \mathbb{P}(S_j = 0) \sim c\sqrt{n}$.

We first assume that $N_n \approx \lambda n$. What λ is optimal under such assumption? Note that $\{N_n \geqslant \lambda n\}$ is a large deviation event since $\mathbb{E}\tau = \infty$, and $N_n \geqslant \lambda n$ is asking for finite gap between returns. Then $\mathbb{P}(N_n \geqslant \lambda n) \approx e^{-I(\lambda)n}$, so

$$\mathbb{P}(N_n \geqslant \lambda n)e^{\beta\lambda n} \approx e^{(\beta\lambda - I(\lambda))n}.$$

The optimal λ is therefore the quantity that maximizes the above expression.

Furthermore, for all $\beta > 0$, there is λ for which the ma is positive. (For $\geqslant 3$ dimensions, need $\beta > \beta_c$ for some $\beta_c > 0$.)

Chapter 3

Weak Convergence and CLT

Notation: we use $\sigma(X)$ to denote the standard deviation of X.

Let $X_1, X_2, \ldots$ be i.i.d. with $\sigma^2 = \mathrm{var}(X_1) < \infty$. Then $\sigma(S_n) = \sigma\sqrt{n}$, so $S_n - \mathbb{E}S_n$ "grows like $\sqrt{n}$." What happens to

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{n}}$$

as $n \to \infty$? This quantity always has zero mean and variance σ, so in particular it does not converge in probability. We will show that this quantity converges **in distribution** to a standard normal Z:

$$\mathbb{P}\left(\frac{S_n - \mathbb{E}S_n}{\sigma\sqrt{n} \leqslant x}\right) \to \mathbb{P}(Z \leqslant x).$$

For triangular arrays $X_{n,k}(n \geqslant 1, k \leqslant k_n)$, the row sums $S_n = \sum_{k=1}^{k_n} X_{n,k}$ connects to this quantity:

$$\frac{S_n - \mathbb{E}S_n}{\sigma\sqrt{n}} = \sum_{k=1}^{n} \frac{X_k - \mathbb{E}X_k}{\sigma\sqrt{n}}.$$

General principle: *$S_n \to Z$ if X_n, k's are approximately independent and with high probability, no one $X_{n,k}$ contributes much to S_n. We will expand on this more rigorously later.*

Example: Coin toss. Let $X_i = \pm 1$ with probability $1/2$ each. Let $+1$ be heads and -1 tails. Then $S_n =$ number of heads − number of tails. The **DeMoivre-Laplace limit theorem** states that

$$\mathbb{P}(S_n/\sqrt{n} \in [a,b]) \to \mathbb{P}(Z \in [a,b]).$$

Proof sketch: consider even indices $\mathbb{P}(S_{2n} = 2k)$ for some k. That is, we get $n+k$ heads and $n-k$ tails in $2n$ tosses. This probability is

$$\mathbb{P}(S_{2n} = 2k) = \binom{2n}{n+k} 2^{-2n} \approx \frac{1}{\sqrt{\pi n}} e^{-k^2/n}$$

uniformly over k with $(k/n)^3 n \to 0$, i.e., $k \ll n^{2/3}$. Then, for $x = 2k/\sqrt{2n}$,

$$\mathbb{P}(S_{2n}/\sqrt{2n} = x) = \frac{1}{\sqrt{\pi n}} e^{-x^2/2} = \frac{2}{\sqrt{2n}} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

The last two terms reminds of standard Gaussian. Now note that

$$\mathbb{P}\left(\frac{S_{2n}}{\sqrt{2n}} = x\right) = \mathbb{P}\left(\frac{S_{2n}}{\sqrt{2n}} \in (x - 1/\sqrt{2n}, x + 1/\sqrt{2n}]\right)$$

since S_{2n} is discrete. On the other hand,

$$\mathbb{P}(z \in (x - 1/\sqrt{2n}, x + 1/\sqrt{2n}]) = \int_{x-1/\sqrt{2n}}^{x+1/\sqrt{2n}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t \sim \frac{2}{\sqrt{2n}} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

Now sum over all $x = 2k/\sqrt{2n}$ in $[a, b]$, and fill in all the $\epsilon - \delta$ proof.

3.1 Weak Convergence

Beginning of Oct. 19, 2022

Given a sequence of distribution functions F_n, we say $F_n \to F$ **weakly** if $F_n(x) \to F(x)$ at all continuity points of F. Why continuity points only? Consider $F_n(x) = 1_{[1/n,\infty)}$, which should converge to a point mass at 0, with $F(x) = 1_{[0,\infty)}$. However, $F_n(0) = 0$.

We say $X_n \to X$ **in distribution** if the distribution functions converge weakly. Our previous De Moivre-Laplace theorem then states that $S_n/\sqrt{n}$ converges in distribution to $\mathcal{N}(0, 1)$.

Note that this simply says $\mathbb{P}(X_n \in (-\infty, x]) \to \mathbb{P}(X \in (-\infty, x])$ for continuity points x, but does not require $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$ for all Borel A. We will discuss more on what A satisfies such limit equation.

Example: Geometric r.v.'s. Let X_p be such that $\mathbb{P}(X_p = n) = (1-p)^{n-1}p$. What happens when $p \to 0$?

First note that $\mathbb{E}X_p = 1/p$, so $\mathbb{E}(pX_p) = 1$. Natural question: does pX_p has a limit in distribution?

For fixed x, $x/p \sim \lfloor x/p \rfloor$ (meaning ratio $\to 1$) as $p \to 0$. What about $\mathbb{P}(pX_p > x)$?

First, $\mathbb{P}(X_p > n) = (1-p)^n$ (i.e., first n all tails). Therefore,

$$\mathbb{P}(pX_p > x) = \mathbb{P}(X_p > x/p) = \mathbb{P}(X_p > \lfloor x/p \rfloor) = (1-p)^{\lfloor x/p \rfloor}.$$

Taking log, we obtain

$$\log(1-p)^{\lfloor x/p \rfloor} = \left\lfloor \frac{x}{p} \right\rfloor \log(1-p) \sim \frac{x}{p}(-p) = -x.$$

Therefore $\mathbb{P}(pX_p > x) \to e^{-x}$, an exponential with parameter 1.

Example: Density functions. If $F_n \to F$ weakly, it is not necessarily true that their derivatives $f_n \to f$ weakly.

Consider $f_n = 2$ on $(j - 1/2^n, j/2^n]$ for odd j and 0 for even j. Then F_n almost looks like diagonal and in fact it converges to $F(x) = x$. But clearly $f_n \nrightarrow f \equiv 1$.

Proposition: Scheffe's Theorem

If f_n, f are densities of μ_n and μ, and if $f_n \to f$ pointwise, then $\sup_{B \in \mathcal{B}} |\mu_n(B) - \mu(B)| \to 0$.

Proof. Let $B_n := \{x : f_n(x) > f(x)\}$. Then

$$\sup_{B \in \mathcal{B}} (\mu_n(B) - \mu(B)) = \mu_n(B_n) - \mu(B_n) = \int (f_n - f)^+ \,\mathrm{d}x.$$

Similarly,

$$\sup_{B\in\mathcal{B}}(\mu(B) - \mu_n(B)) = \int (f_n - f)^- \,\mathrm{d}x.$$

Since f_n and f are densities, the two lines above are equal. It suffices to show $\int (f_n - f)^- \,\mathrm{d}x \to 0$ as $n \to \infty$. To do so we use DCT: $(f_n - f)^- \to 0$ a.s. and is bounded by f, so by DCT, the integral converges to 0. □

Lemma

For all distribution function F, there exists a random variable Y on $([0,1], \mathcal{B}, \mathbb{P})$ with $\mathbb{P}$ uniform, such that Y has distribution function F.

Proof. If F is continuous and strictly increasing, let $Y(\omega) = F^{-1}(\omega)$. Then $Y(\omega) \leqslant t$ iff $\omega \leqslant F(t)$ iff $\omega \in [0, F(t)]$, so $\mathbb{P}(Y \leqslant t) = \mathbb{P}(Y \in [0, F(t)]) = F(t)$.
More generally, let $Y(\omega) := \sup\{y : F(y) < \omega\}$. Then $Y(\omega) \leqslant t$ iff $\omega \leqslant F(t)$ iff $w \in [0, F(t)]$, and we are done. □

If X_n and Y_n have the same distribution, X and Y have the same distribution, and $X_n \to X$ a.s., is it true that $Y_n \to Y$ a.s.? The answer is no.

Example. Let $X \sim \mathcal{N}(0,1)$, and let $X_n = X$ for all n. Then $X_n \to X$ trivially. Let Y_n be i.i.d. standard normals, and clearly $Y_n \nrightarrow \mathcal{N}(0,1) := Y$.

Beginning of Oct. 21, 2022

Theorem: Convergence in distribution vs a.s.

If $F_n \to F$ in distribution, then there exist Y_n, Y with distribution functions F_n, F such that $Y_n \to Y$ almost surely.

Proof. The existence of Y_n, Y have been shown above. We need to only consider $\omega \in [0,1]$ for which $F^{-1}(\omega)$ contains 0 or 1 point. Fix ω and let $t = Y(\omega)$. Then

$$F^{-1}(\omega) = \varnothing \text{ or } \{t\}.$$

Therefore, for such points, for all $\delta > 0$,

$$F(t - \delta) < F(t) < F(t + \delta).$$

Choose δ such that $t \pm \delta$ are continuity points of F. Then, for large n, $F_n(t - \delta) < F(t) < F_n(t + \delta)$, so $t - \delta \leqslant Y_n(\omega) \leqslant t + \delta$, and similarly $t - \delta \leqslant Y(\omega) \leqslant t + \delta$. Since δ is arbitrary, $Y_n \to Y$ a.s., as there can only be countably many exceptions (countable jumps). □

Theorem: D3.2.9, Characterization of Weak Convergence

$X_n \to X$ in distribution iff $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ for all bounded continuous g.

Proof. Suppose $X_n \to X$ weakly. Take Y_n with the same distribution of X_n and Y similarly, with $Y_n \to Y$ almost surely. Let g be bounded and continuous. Then $g(Y_n) \to g(Y)$ a.s., so $\mathbb{E}g(Y_n) \to \mathbb{E}g(Y)$ by bounded convergence theorem.

Conversely, suppose $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ for all bounded continuous functions. We want

$$\mathbb{E}1_{(-\infty,x]}(X_n) \to \mathbb{E}1_{(-\infty,x]}(X)$$

for all continuity points x.

$1_{(-\infty,x]}$ isn't continuous, but it can be approximated by 1 on $(-\infty, x-\epsilon)$, 0 on (x,∞), and linear in between. We call this function $g_{x-\epsilon,x}$ and define $g_{x,x+\epsilon}$ similarly. By assumption,

$$\mathbb{E}g_{x-\epsilon,x}(X_n) \to \mathbb{E}g_{x-\epsilon,x}(X) \geqslant F(x-\epsilon)$$

and

$$\mathbb{E}g_{x,x+\epsilon}(X_n) \to \mathbb{E}g_{x,x+\epsilon}(X) \leqslant F(x+\epsilon).$$

Then since $F(x-\epsilon) \leqslant \liminf_{n\to\infty} F_n(x) \leqslant \limsup_{n\to\infty} F_n(x) \leqslant F(x+\epsilon)$, if x is a continuity point, we obtain the claim. □

Remark. Note that we can weaken the assumption and only require g to be continuous a.e.: denote the discontinuity set as D_g; if $\mathbb{P}(X \in D_g) = 0$ and $X_n \to X$ in distribution, then $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$.

Corollary

If $X_n \to X$ in distribution and f is continuous, then $f(X_n) \to f(X)$ in distribution too.

Proof. If g is bounded, then $g \circ f$ is bounded, so $\mathbb{E}g(f(X_n)) \to \mathbb{E}g(f(X))$. Using the previous theorem once more, $f(X_n) \to f(X)$ in distribution. □

Corollary

If $X_n \to X$ almost surely, then $X_n \to X$ in distribution.

We have shown that there exists a metric w.r.t. convergence in probability: $|X-Y|/(1+|X-Y|)$. There also exists metrics (one example is Lévy metric) for convergence in distribution.

Proposition: Convergence in probability ⇒ in distribution

Slick proof. It suffices to show that for all subsequence, there exists a further subsequence converging almost surely (then such sub-subsequence converges in distribution). And this is true as shown previously. Finally, since there is a metric for convergence in distribution, the full sequence indeed $\to X$ in distribution.

More revealing proof. Let g be bounded continuous, with $|g| \leqslant K$. By uniform continuity on compact sets, given

M and ϵ, there exists δ satisfying the uniform continuity criterion on $[-M, M]$. Then

$$\begin{aligned}|\mathbb{E}g(X_n) - \mathbb{E}g(X)| &= \int_\Omega |g(X_n) - g(X)| \, d\mathbb{P} \\ &\leqslant \int_{|X|\leqslant M, |X_n - X| < \delta} |g(X_n) - g(X)| \, d\mathbb{P} + \int_{|X|>M} |...| \, d\mathbb{P} + \int_{|X_n - X|\geqslant\delta} ... \, d\mathbb{P} \\ &\leqslant \int_{|X|\leqslant M, |X_n - X| < \delta} \epsilon \, d\mathbb{P} + \int_{|X|>M} 2K \, d\mathbb{P} + \int_{|X_n - X|\geqslant\delta} 2K \, d\mathbb{P} \\ &\leqslant \epsilon + 2K\mathbb{P}(|X| > M) + 2K\mathbb{P}(|X_n - X| \geqslant \delta).\end{aligned}$$

Hence,

$$\limsup_{n\to\infty} |\mathbb{E}g(X_n) - \mathbb{E}g(X)| \leqslant \epsilon + 2K\mathbb{P}(|X| > M) \qquad \text{for all } M, \epsilon.$$

Since M, ϵ are arbitrary, we see $\limsup |\mathbb{E}g(X_n) - \mathbb{E}g(X)| = 0$, as claimed. □

Remark: Converse is false. Let X_n, X be i.i.d. $\mathcal{N}(0,1)$. Then clearly $X_n \to X$ in distribution, but not in probability.
However, (shown in HW), if $X_n \to c$ for some constant c, then indeed $X_n \to c$ in probability.

Beginning of Oct. 24, 2022

Let X be any random variable. Then for all $\epsilon > 0$ there exists M such that $\mathbb{P}(|X| > M) < \epsilon$. We say $\{X_n\}$ is **tight** if given ϵ, there exists M such that $\mathbb{P}(|X_n| > M) < \epsilon$ for all n. One easy example: if $\{\mu_n\}$ is bounded, and $X_n \sim \mathcal{N}(\mu_n, 1)$, then $\{X_n\}$ are tight.

Remark. In general, if $X_n \to X, Y_n \to Y$ in distribution, $X_n + Y_n \nrightarrow X + Y$ in distribution. Example: let $X_n = X = Y = 1$ if heads and 0 if tails, and let $Y_n = 0$ if heads and 1 if tails. Then clearly $X_n + Y_n$ is constantly 1 whereas $X + Y$ is either 2 or 0.

Theorem: Slutsky's Theorem

If $X_n \to X$ in distribution and $Y_n \to 0$ in distribution, then $X_n + Y_n \to X$ in distribution.

Proof. Using the bounded function characterization of convergence in distribution, let g be bounded with $|g| \leqslant K$. Given $M, \epsilon > 0$, there exists δ satisfying the uniform continuity criterion on $[-M, M]$. Then

$$\mathbb{E}|g(X_n + Y_n) - g(Y_n)| \leqslant \int_{|X_n|\leqslant M, |Y_n| < \delta}^{\infty} \epsilon \, d\mathbb{P} + \int_{|X_n|>M} 2K \, d\mathbb{P} + \int_{|Y_n|\geqslant\delta} 2K \, d\mathbb{P}$$

just like in the proof of D3.2.9, characterization of weak convergence. □

Proposition: Tightness lemma

If $X_n \to X$ in distribution then $\{X_n\}$ is tight.

Proof. Let F_n be the d.f. of X_n and f that of X. Let $\epsilon > 0$. Clearly,

$$\mathbb{P}(|X_n| > M) \leqslant F_n(-M) + 1 - F_n(M).$$

We take M_0 such that $\pm M_0$ are continuity points of F, and

$$F(-M_0) + 1 - F(M_0) < \frac{\epsilon}{2}.$$

Hence by assumption there exists n_0 such that if $n \geqslant n_0$,

$$F_n(-M_0) + 1 - F_n(M) < \epsilon.$$

For each one in the first finitely many terms, there exists M_i with $F_i(-M_i) + 1 - F_i(M_i) < \epsilon$. Take the maximum among $M_i, i = 1, 2, \ldots, n_0 - 1$, and M_0, we finish the proof. ☐

If $X_n \to X$ in distribution, for what A does $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$? Intuitively, for an open set, X_n's distribution may converge to the boundary, resulting in a loss of probability.

Theorem: D3.2.11

The following are equivalent:

(1) $X_n \to X$ in distribution,

(2) For all open sets G, $\liminf_{n\to\infty} \mathbb{P}(X_n \in G) \geqslant \mathbb{P}(X \in G)$,

(3) For all closed sets K, $\limsup_{n\to\infty} \mathbb{P}(X_n \in K) \leqslant \mathbb{P}(X \leqslant k)$, and

(4) For every Borel A with $\mathbb{P}(X \in \partial A) = 0$, $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$.

Proof. $(1) \Rightarrow (2)$. Let X_n and Y_n have the same distribution, and same for X, Y. Assume $Y_n \to Y$ a.s. Let G be open. Then $Y \in G$ means $Y_n \in G$ "eventually." That is,

$$1_G(Y) = \liminf_{n\to\infty} 1_G(Y_n).$$

By Fatou, taking expectation gives

$$\mathbb{P}(Y \in G) \leqslant \mathbb{E}(\liminf_{n\to\infty} 1_G(Y_n)) \leqslant \liminf_{n\to\infty} \mathbb{P}(Y_n \in G).$$

$(2) \Rightarrow (3)$. Take complements.

$(2), (3) \Rightarrow (4)$. Suppose $\mathbb{P}(X \in \partial A = 0)$. We denote the interior as A° and closure $\overline{A}$. Then $\mathbb{P}(X \in A^\circ) = \mathbb{P}(X \in A) = \mathbb{P}(X \in \overline{A})$. We apply (2) to A° and (3) to $\overline{A}$ and obtain the claim.

$(4) \Rightarrow (1)$. Let A take form $(-\infty, x]$. If $\mathbb{P}(X = x) = 0$ then the d.f. is continuous at x. Done. ☐

Example. Let X_n be uniform on $[-n-1, -n] \cup [-1, 1] \cup [n, n+1]$. For $x \geqslant 1$, the distribution function $F_n(x) \to 3/4 =: F(x)$. Note that F is a measure but not a probability measure anymore — mass escapes at infinity!

In general: if F is right-continuous and nondecreasing, if $F_n(x) \to F(x)$ for every continuity point of F, we say $F_n \to F$ **vaguely**.The above example shows that if F_n are distribution functions and $F_n \to F$ vaguely, it is still *not* necessarily true that F is a CDF.

Theorem: Helly selection theorem

Every sequence $\{F_n\}$ of distribution functions has a subsequence $\{F_{n_k}\}$ converging vaguely to F for some F, again, not necessarily a probability measure. "Almost compactness, but not quite."

Beginning of Oct. 26, 2022

Proposition

Suppose $\{F_n\}$ are distribution functions and $F_n(q) \to G(q)$ for all $q \in Q$, Let $F(x) = G(x+)$. Then $F_n \to F$ vaguely.

Proof. From analysis, F is continuous. Also $F \geqslant G$. If $r > s$, then $G(r) \geqslant G(s+) = F(s)$.
Let x be a continuity point of F and let $\epsilon > 0$. Take $r_1 < r_2 < x < s$ with $r_1, r_2, s \in \mathbb{Q}$, and

$$F(x) - \epsilon < F(r_1) \leqslant F(r_2) \leqslant F(x) \leqslant F(s) < F(x) + \epsilon.$$

By definition/assumption $F_n(r_2) \to G(r_2) \geqslant F(r_1)$ by the previous observation. Also, $F_n(s) \to G(s) \leqslant F(s)$. Therefore $F_n(x)$ is sandwiched between $F(x) - \epsilon$ and $F(x) + \epsilon$. □

Proof of Helly selection theorem. We use the diagonal method. Enumerate Q by $\{q_i\}$.
There exists a subsequence S_1 on which $F_n(q_1) \to$ some constant $G(q_1)$ by compactness of $[0, 1]$ (we are defining the values of G at rationals using compactness). We can pick a further subsequence S_2 on which $F_n(q_1) \to G(q_1)$ and $F_n(q_2) \to G(q_2)$. So on and so forth. We now take the i^{th} element in S_i, and the nearly formed sequence $n(k)$ satisfies $F_{n(k)}(q_i) \to G(q_i)$ for all $q_i \in \mathbb{Q}$, and by the previous remark we are done. □

Theorem: D3.2.13

Let $\{F_n\}$ be a sequence of distribution functions. Then every subsequential limit is a d.f. iff $\{F_n\}$ is tight.

Proof. Let μ_n be the probability measure corresponding to F_n.
First suppose $\{F_n\}$ is tight. Let $\epsilon > 0$. By assumption there exists M such that $\mu_n([-M, M]) > 1 - \epsilon$ for all n. If a subsequence $\mu_{n_k} \to \mu$ vaguely, we want to show that $\mu(\mathbb{R}) = 1$. Indeed, assuming F is a continuity point (which we can always choose so),

$$\mu(\mathbb{R}) \geqslant \mu([-M, M]) \geqslant \limsup \mu_{n_k}([-M, M]) > 1 - \epsilon.$$

Conversely, suppose $\{F_n\}$ is not tight. That is, there exists $\epsilon > 0$ such that for all M, there exists $\mu_{n(M)}$ with $\mu_{n(M)} \leqslant 1 - \epsilon$. WLOG assume $n(1) < n(2) < \ldots$. Then there exists a further sbsequence $n(M_k)$ on which $F_{n(M_k)}$ converges vaguely to some μ by Helly. Thnen for all continuity points a of μ, $\mu((-a, a]) = \lim_k \mu_{n(M_k)}((-a, a]) \leqslant \liminf_k \mu_{n(M_k)}((-M_k, M_k])$ for large M_k. Then the quantity is bounded by $1 - \epsilon$, and so $\mu(\mathbb{R}) \leqslant 1 - \epsilon$, and we are done. □

Beginning of Oct. 28, 2022

Theorem: D3.2.14, Sufficient condition for tightness

Suppose there exists $\varphi : \mathbb{R} \to [0, \infty)$ with $\varphi \to \infty$ as $|x| \to \infty$ and $\mathbb{E}\varphi(X_n)$ is (uniformly) bounded. Then $\{X_n\}$ is tight.

For example, if $\mathbb{E}|X_n|$ or $\mathbb{E}\log(1 + |X_n|)$ is bounded, then $\{X_n\}$ is tight.

Proof. Define $\varphi_0(x) := \inf\{\varphi(t) : |t| \geqslant |x|\}$. By assumption φ_0 is symmetric/even and monotonically $\to \infty$ on $[0, \infty)$. Also, $\varphi_0 \leqslant \varphi$, so $\mathbb{E}\varphi_0(X_n)$ also has (uniformly) bounded expectation, say by some K. WLOG we can further assume φ_0 to be strictly increasing on $[0, \infty)$ by adding something strictly increasing and also bounded bounded to it. Then

$$\mathbb{P}(|X_n| \geqslant M) = \mathbb{P}(\varphi_0(X_n) \geqslant \varphi_0(M)) \leqslant \frac{\mathbb{E}\varphi_0(X_n)}{\varphi_0(M)} \leqslant \frac{K}{\varphi_0(M)}.$$

Given $\epsilon > 0$, choose M large with $K/\varphi_0(M) < \epsilon$, and we are done. □

3.2 Characteristic Functions

Let X be a random varaible. We define the complex function $\varphi_X(t) := \mathbb{E}e^{itX} = \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX)$ to be its **characteristic function**. Note immediately that

$$|\varphi_X(t)| \leqslant \mathbb{E}|e^{itX}| = 1 \qquad \text{with} \qquad \varphi_X(0) = 1.$$

Example. Let X be uniform on $[-1, 1]$. Then

$$\varphi_X(t) = \int_{-1}^{1} \frac{1}{2}(\cos tx + i \sin tx)\, \mathrm{d}x = \frac{1}{2}\frac{\sin tx}{t}\bigg|_{-1}^{1} + 0 = \frac{\sin t}{t}.$$

Example 3.2.1. We will show later that if $\mathbb{E}|X| < \infty$ then $\varphi_X'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\int e^{itX}\, \mathrm{d}\mathbb{P} = \int iXe^{itX}\, \mathrm{d}\mathbb{P}$. For example if X is standard normal, then

$$\varphi_X'(t) = \int iXe^{itx} f(x)\, \mathrm{d}x.$$

Note that $f(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$ satisfies $f'(x) = -xf(x)$, so the cosine part is an odd function, and so

$$\varphi_X'(t) = -\int x\sin(tx) f(x)\, \mathrm{d}x = \int \sin(tx) f'(x)\, \mathrm{d}x.$$

IBP and we obtain $\varphi_X'(t) = -t\varphi_X(t)$ with initial condition $\varphi_X(0) = 1$. This gives $\varphi_X(t) = \exp(-t^2/2)$.

Proposition

For all X, $\varphi_X(t)$ is uniformly continuous. (We will drop the subscript X for convenience.)

Proof. Since

$$|\varphi(t+h) - \varphi(t)| = \mathbb{E}|e^{i(t+h)X} - e^{itX}| = \mathbb{E}|e^{ihX} - 1|,$$

$|e^{ihX} - 1| \leqslant 2$, and $e^{ihX} \to 0$ as $h \to 0$, by DCT the limit is 0, uniform in t. □

Remark. If X is symmetric, i.e., X and $-X$ have the same distribution, $\varphi_X(t) = \overline{\varphi_X(t)}$ so $\varphi_X(t) \in \mathbb{R}$.

Question. Does φ_X determine the distribution of X. Furthermore, can we calculate the distribution of μ from φ_X?

Relation to Fourier transform: given f, we define

$$\hat{f}(t) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\theta x} f(x)\, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \varphi(-t).$$

If the density $f \in L^2$ and $\varphi \in L^1$ (i.e. integrable), then

$$f(x) = \hat{\hat{f}}(-x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} \hat{f}(t)\, \mathrm{d}t = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itx} \varphi(-t)\, \mathrm{d}t = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t)\, \mathrm{d}t.$$

Hence (since $\varphi \in L^1$, Fubini applies)

$$\begin{aligned} \mu((a,b)) &= \int_a^b f(x)\, \mathrm{d}x = \frac{1}{2\pi} \int_a^b \int_{-\infty}^{\infty} e^{-itx} \varphi(t)\, \mathrm{d}t\, \mathrm{d}x \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_a^b e^{-itx} \varphi(t)\, \mathrm{d}x\, \mathrm{d}t = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, \mathrm{d}t. \end{aligned}$$

Theorem: Inversion Formula

If μ is a probability measure with ch.f. φ, then

$$\lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, \mathrm{d}t = \mu((a,b)) + \mu(\{a\})/2 + \mu(\{b\})/2.$$

Beginning of Halloween, 2022

Proof. For convenience, define $I_T := \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, \mathrm{d}t = \int_{-T}^{T} \int_{\mathbb{R}} e^{itX}\, \mu(\mathrm{d}x)\, \mathrm{d}t$. By Fubini,

$$I_T = \int_{\mathbb{R}} \int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}\, \mathrm{d}t\, \mu(\mathrm{d}x).$$

Define $R(\theta, T) := \int_{-T}^{T} \frac{e^{it\theta} - 1}{it}\, \mathrm{d}t$. Then the inner integral above is $R(x-a, T) - R(x-b, T)$. Note that R is real since R is odd, and so only the real part remains:

$$R(\theta, T) = \int_{-T}^{T} \frac{\sin(\theta t)}{t}\, \mathrm{d}t = 2\,\mathrm{sgn}(\theta) \int_{\theta}^{T|\theta|} \frac{\sin u}{u}\, \mathrm{d}u.$$

Because of the sign function, $R(x-a, T) - R(x-b, T)$ depends on the relative position of x to a and b.

Since $\int_0^{\infty} \frac{\sin u}{u}\, \mathrm{d}u = \pi/2$, $R(\theta, T) \to g(x) := \begin{cases} 2\pi & x \in (a,b) \\ \pi & x \in \{a,b\} \\ 0 & \text{otherwise.} \end{cases}$ Then, by bounded convergence,

$$\frac{1}{2\pi} I_T = \frac{1}{2\pi} \int_{\mathbb{R}} (R(x-a,T) - R(x-b,T))\, \mu(\mathrm{d}x) \to \frac{1}{2\pi} \int_{\mathbb{R}} g(x)\, \mu(\mathrm{d}x) = \mu((a,b)) + \frac{1}{2}\mu(\{a\}) + \frac{1}{2}\mu(\{b\}).$$

□

Theorem: D3.3.14

If $\varphi \in L^1$, then μ has a bounded continuous density

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\varphi(y)\,\mathrm{d}y.$$

Proof. We want to show that for $a < b$, $\mu((a,b)) = \int_a^b \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\varphi(y)\,\mathrm{d}y\,\mathrm{d}x$. Note that

$$\left|\frac{e^{-ita} - e^{-itb}}{it}\right| = \left|\int_a^b e^{-ity}\,\mathrm{d}y\right| \leqslant |b-a|,$$

so (defining I_T as above), $I_T \to I_\infty$. Therefore

$$\begin{aligned}\mu((a,b)) + \mu(\{a\})/2 + \mu(\{b\})/2 &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-ita} - e^{-itb}}{it}\varphi(t)\,\mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_a^b e^{-itx}\,\mathrm{d}x\varphi(t)\,\mathrm{d}t \\ &= \int_a^b \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)\,\mathrm{d}t\,\mu(\mathrm{d}x).\end{aligned}$$

Call the last quantity $h(a,b)$. We want to show that $h(a,b) = \mu((a,b))$. This is because as an integral, h is continuous in (a,b). The absense of jumps imply that $\mu(\{a\}), \mu(\{b\}) = 0$ for a, b. So $h(a,b) = \mu((a,b))$ for all a, b. □

3.3 Weak Convergence

If $X_n \to X$ weakly, then $\varphi_{X_n}(t) = \mathbb{E}(r\cos(tX_n) + i\sin(tX_n))$, a bounded continuous function, must also converge to $\varphi_X(t)$ pointwise.

Conversely, if $\varphi_n \to \varphi$ pointwise, does the measures μ_n associated with φ_n necessarily to some μ weakly? The answer is no.

Example. Let $U_1, U_2, \ldots$ be i.i.d. over $[-1,1]$, and let $S_n = \sum_{i=1}^n U_i$. Then $\varphi_{U_1}(t) = \sin t/t$, and $\varphi_{S_n}(t) = \varphi_{U_1}(t)^n$. As $n \to \infty$, φ equals 0 only when $t = 1$ and $\to 0$ otherwise. However, S_n is not converging in distribution. Furthermore, the limit $\varphi = 1_{\{0\}}$ is not continuous, so it cannot be a ch.f. anyway. Therefore $\{S_n\}$ does not converge weakly.

Theorem: Continuity theorem, D3.3.17

Let μ_n be probability measures with ch.f. φ_n. Let μ be a probability measure with ch.f. φ.

(1) If $\mu_n \to \mu$ weakly, then $\varphi_n(t) \to \varphi(t)$ pointwise.

(2) Conversely, if $\varphi_n \to \varphi$ pointwise, *and* φ is continuous at 0, then $\mu_n \to \mu$ weakly, and μ has ch.f. φ.

Remark: "General principle". The behavior of φ near 0 is related (in various ways) to "the measure μ near ∞," e.g. moments, tail probabilities, etc.

Intuitively, for small t, e^{itX} is close to 1 unless X is big, pushing the value away from 1 significantly.

Proposition: Tightness from char functions

Let φ be the ch.f. of μ. Then for all $u > 0$, $\mu(\{x : |x| > 2/n\}) \leqslant u^{-1} \int_{-u}^{u} (1 - \varphi(t))\, \mathrm{d}t$.

Proof. We plug in the definition of $\varphi(t)$.

$$\begin{aligned}\frac{1}{u}\int_{-u}^{u}(1-\varphi(t))\,\mathrm{d}t &= \frac{1}{u}\int_{-u}^{u}\int_{\mathbb{R}}(1-e^{-itx})\,\mu(\mathrm{d}x)\,\mathrm{d}t \\ &= \int_{\mathbb{R}}\frac{1}{u}\int_{-u}^{u}(1-e^{itx})\,\mathrm{d}t\,\mu(\mathrm{d}x) \\ &= 2\int_{\mathbb{R}}\left(\frac{1-(\sin ux)}{ux}\right)\mu(\mathrm{d}x).\end{aligned}$$

For $|y| \geqslant 2$, $|\sin y/y| \leqslant 1/2$, so integrating over $\{|x| \geqslant 2/u\}$ gives an lower bound $2\int_{|x|\geqslant 2/u} 1/2\, \mu(\mathrm{d}x)$. □

Beginning of Nov. 2, 2022

Proof of continuity theorem. (i) True. (ii) By the lemma, for $u > 0$,

$$\mu_n(\{|x| \geqslant 2/u\}) = \frac{1}{u}\int_{-u}^{u}(1-\varphi_n(t))\,\mathrm{d}t.$$

By bounded convergence theorem, this converges to the integral with integrand φ_n replaced by φ, as $n \to \infty$. Given $\epsilon > 0$, choose u such that last integral $< \epsilon$. Then beyond some N_0,

$$\frac{1}{u}\int_{-u}^{u}(1-\varphi_n(t))\,\mathrm{d}t < \epsilon \qquad \text{for all } n \geqslant N_0,$$

so μ_n is tight, as there exists $M \geqslant 2/u$ such that $\mu_n(\{|x| > M\}) < \epsilon$ for all $n \geqslant N$, and there are only finitely many early terms, which we can bound individually.

Finally, we show that the full sequence converges. If $\mu_n \to \mu$ weakly then there exists a subsequence $\varphi_{n_k} \to$ (ch.f. of μ), so φ must be the ch.f. of μ. Thus the full sequence μ_n converges to μ. □

Differentiation and Moments

When can we differentiate

$$\varphi(t) = \int_{\mathbb{R}} e^{itx}\,\mu(\mathrm{d}x)?$$

Note that

$$\frac{\varphi(x+h)-\varphi(x)}{h} = \int_{\mathbb{R}} e^{itx}\frac{e^{ihx}-1}{h}\,\mu(\mathrm{d}x).$$

The term $(e^{ihx}-1)/h$ is bounded by $|x|$, so it is integrable as $h \to 0$. Thus it is sufficient to require $\mathbb{E}|X| = \int |x|\ \mu|\mathrm{d}x| < \infty$. Then

$$\varphi'(t) = \int_{\mathbb{R}} ixe^{itx}\,\mu(\mathrm{d}x).$$

More generally, to take the n^{th} derivative, it suffices to require $\mathbb{E}|X|^n < \infty$, with $\varphi^{(n)}(t) = \int_{\mathbb{R}}(ix)^n e^{itx}\,\mu(\mathrm{d}x)$. Note that the expression holds independent of t.

Theorem

(Conversely,) if $\varphi^{2n}(0)$ exists and is finite, then $\mathbb{E}|X|^{2n} < \infty$. (Does not necessarily work for odd powers.)

Proof. We will prove the special $k = 1$ case: assume $\varphi''(0)$ is finite. We write it out:

$$\begin{aligned}\varphi''(0) &= \lim_{h\to 0} \frac{\varphi(x) - 2\varphi(0) + \varphi(-h)}{h^2} \\ &= \lim_{h\to 0} \int_{\mathbb{R}} \frac{e^{itx} - 2 + e^{-ihx}}{h^2}\, \mu(\mathrm{d}x) \\ &= -2\lim_{h\to 0} \int_{\mathbb{R}} \frac{1-\cos(hx)}{h^2}\, \mu(\mathrm{d}x).\end{aligned}$$

By Fatou's lemma (and since limit exists),

$$\varphi''(0) \leqslant -2\int_{\mathbb{R}} x^2\, \mu(\mathrm{d}x) = -\mathbb{E}X^2$$

so $\mathbb{E}|X|^2$ is finite. For $k \geqslant 2$, we need to induct on k and apply the above argument to $X^2/\mathbb{E}X^2\mu(\mathrm{d}x)$. □

Taylor-type Expansions

Expanding φ around 0 gives

$$e^{itX} = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} X^k = \sum_{k=0}^{n} \frac{(it)^k}{k!} X^k + \mathcal{O}(|tX|^{n+1})$$

as $|tX| \to 0$. Clearly we can take the expected value of the first finite sum, when $\mathbb{E}|X|^n < \infty$. But what about the remainder? A lemma from Durrett —

Proposition: D3.3.19

$$\left|e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!}\right| \leqslant \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!}\right).$$

Proof sketch. Integrate by parts and iterate:

$$e^{ix} = 1 + ix + \sum_{k=0}^{n} \frac{i^k}{k!} x^k + \frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is}\, \mathrm{d}s.$$

This will give

$$\left|\frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is}\, \mathrm{d}s\right| \leqslant \frac{|x|^{n+1}}{(n+1)!}.$$ □

Now looking back at Taylor expansions:

$$\left|\mathbb{E}e^{itX} - \sum_{m=0}^{n} \mathbb{E}\frac{(itX)^m}{m!}\right| \leqslant \mathbb{E}\min(|tX|^{n+1}, 2|tX|^n).$$

Beginning of Nov. 4, 2022

Convex Combinations of r.v.'s

If $\mu_1, \ldots, \mu_n$ are probability measures and $\sum_{i=1}^n \lambda_n = 1$, then the weighted sum $\sum_{i=1}^n \lambda_i \mu_i$ is also a probability measure. Similarly, if μ_s is a probability measure for all $s \in I$, and ν is a measure on I, then (assuming measurability)

$$\int_I \mu_s\, \nu(\mathrm{d}s)$$

is also a probability measure, with ch.f. $\int_I \varphi_s(t)\, \nu(\mathrm{d}s)$ where φ_s is the ch.f. of μ_s.

Example. Let $f_1(x) = (1-\cos x)/(\pi x^2)$, which has ch.f. $\varphi 1(t) = \max(1-|t|, 0)$. If s is any scalar then X/s has ch.f. $\varphi_s(t_0 = \varphi 1(t/s)$.
Then we can consider combinations of φ_i's. For example, we note that $f_1/3 + 2f_5/3$ is a density with ch.f. $\varphi_1/3 + 2\varphi_5/3$.

Theorem: Polya's criterion, D3.3.22

Let $\varphi \geqslant 0$ with $\varphi(0) = 1$, and $\varphi(t) = \varphi(-t)$. Furthermore assume φ is decreasing and convex on $(0, \infty)$ with $\lim_{t\to 0} \varphi(t) = 1$ and $\lim_{t\to\infty} \varphi(t) = 0$. Then φ is a ch.f.

Proof. Idea: express $\varphi(t)$ *as* $\int_0^\infty \varphi_s(t)\ \nu(\mathrm{d}s)$ *as defined in the example above for some* ν. *If we can differentiate inside this integral, then* φ *convex implies* φ' *exists a.e. and increasing, and* $\varphi'(t) = \int_0^\infty \varphi_s'(t)\ \nu(\mathrm{d}s)$. *But note that* $\varphi_s'(t) = 0$ *if* $s \leqslant t$, *so this equals* $-\int_t^\infty s^{-1}\ \nu(\mathrm{d}s)$. *So,as a measure,* $\mathrm{d}\varphi'(t) = t^{-1}\nu(\mathrm{d}t)$, *and* $\nu(\mathrm{d}t) = t\mathrm{d}\varphi'(t)$ *is our candidate for* ν.

We may assume $\varphi(\infty) = 0$. Assume φ' is right-continuous (if not, replace it with $F(t) = \varphi'(t+)$. Define ν by $\nu([0,t]) = \int_0^t s\ \mathrm{d}\varphi'(s)$. Then

$$\mathrm{d}\varphi'(t) = t^{-1}\nu(\mathrm{d}t)$$

as a measure. For $t > 0$,

$$\varphi'(\infty) - \varphi'(t) = \int_t^\infty \mathrm{d}\varphi'(s) = \int_{(t,\infty)} s^{-1}\ \nu(\mathrm{d}s),$$

so

$$\varphi'(t) = -\int_{(t,\infty)} s^{-1}\ \nu(\mathrm{d}s) \text{ a.e.}$$

Since φ is convex,

$$\begin{aligned}
\varphi(\infty) - \varphi(t) = \int_t^\infty \varphi'(u)\ \mathrm{d}u &= -\int_t^\infty \int_{(u,\infty)} s^{-1}\ \nu(\mathrm{d}s)\ \mathrm{d}u \\
&= -\int_{(t,\infty)} \int_{(t,s)} s^{-1}\ \mathrm{d}u\ \nu(\mathrm{d}s) \\
&= -\int_{(t,\infty)} (1 - t/s)\ \nu(\mathrm{d}s) \\
&= -\int_{(t,\infty)} \varphi_s(t)\ \nu(\mathrm{d}s)
\end{aligned}$$

and we are done, following our previous observation. □

Back to Taylor series: if *all* moments $\mathbb{E}|X|^k$ are finite, can we conclude

$$\mathbb{E}e^{itX} = \mathbb{E}\sum_{k=0}^\infty \frac{i^n X^n}{n!} t^n = \sum_{k=0}^\infty \frac{i^n \mathbb{E}|X|^n}{n!} t^n?$$

The answer is still no in general.
All $\mathbb{E}(X^k)$ finite implies all $\varphi^{(k)}(t)$ exist for all t, with

$$\varphi^{(k)}(\theta) = \mathbb{E}((iX)^k e^{i\theta X}) \text{ and } \varphi^{(k)}(0) = i^k \mathbb{E}X^k.$$

If the full Taylor series

$$\sum_{k=0}^{\infty} \frac{\varphi^{(k)}(\theta)}{k!}(1-\theta)^k = \sum_{k=0}^{\infty} \frac{1}{k!}(\mathbb{E}(iX)^k e^{i\theta X})t^k$$

has a positive radius of convergence $r > 0$ for some θ, then the values $\mathbb{E}((iX)^k e^{i\theta X})$ determine φ in an interval around θ, $(\theta - r, \theta + r)$. Further, $r > 0$ iff

$$\limsup_{k\to\infty} \left| \frac{1}{k!}\mathbb{E}((iX)^k e^{i\theta X}) \right|^{1/k} < \infty.$$

By Stirling, $k! \geqslant (k/e)^k$, so this is true for all θ, if

$$\limsup_{k\to\infty} \frac{(\mathbb{E}|X|^k)^{1/k}}{k} < \infty.$$

That is, it suffices to require that "moments don't grow too fast:"

$$\mathbb{E}|X|^k \leqslant (ck)^k \text{ for some } c.$$

Theorem: D3.3.25

If $\{\mu_{2k}\} > 0$ with $\limsup_{k\to\infty} \frac{\mu_{2k}^{1/k}}{k} < \infty$, then there exists at most one distribution with moments $\mathbb{E}X^k = \mu_k$, for all k.

3.4 Central Limit Theorem

Beginning of Nov. 7, 2022

From Taylor's theorem, $|\log(1+z) - z| = \mathcal{O}(|z|^2)$ as $z \to 0$, so if $c_n \to c$ in $\mathbb{C}$, then $(1 + c_n/n)^n \to e^c$ as $n \to \infty$, and

Theorem: i.i.d. CLT

Let $X_1, X_2, \ldots$ be i.i.d., with $\mathbb{E}X_1 = \mu$ and $\text{var}(X_1) = \sigma^2 \in (0, \infty)$, then

$$\frac{S_n - n\mu}{\sigma n^{1/2}} \to \mathcal{N}(0,1)$$

in distribution.

Proof. We first assume $\mu = 0$ From D3.3.20 $\varphi_{X_1}(t) = \mathbb{E}\exp(itX_1) = 1 - \sigma^2 t^2/2 + o(t^2)$ as $t \to 0$. Hence

$$\varphi_{S_n/(\sigma\sqrt{n})}(t) = \mathbb{E}\exp(itS_n/(\sigma\sqrt{n}) = \varphi_{S_n}(1/(\sigma\sqrt{n})) = \varphi_{X_1}(1/(\sigma\sqrt{n}))^n.$$

For t fixed, this quantity becomes $(1 - t^2/(2n) + o(1/\sigma))^n = (1 - (t^2 - nt_n)/(2n))^n$. The numerator is converging to t^2, so by the previous observation, the entire quantity converges to $\exp(-t^2/2)$, and we are done. □

Example. A business rounds all transformations to the nearest integer, so the error X in one transaction is a uniform distribution (though unrealistic) on $[-0.5, 0.5)$. Let $n = 100$ be the number of transactions. Then

$\mathbb{E}X_1 = 0$ and $\text{var}(X_1) = 1/12$.

$$\mathbb{P}(|\text{total error}| > 20) = \mathbb{P}\left(\frac{S_n - 0}{\sqrt{n}1/\sqrt{12}} > \frac{20 - 0}{\sqrt{n}1/\sqrt{12}}\right) \approx \mathbb{P}(z > 2.19) \approx 0.14.$$

How about triangular arrays? Suppose the varaibles on the n^{th} row are independent, $S_n = \sum_{k=1}^{k(n)} X_{n,k}$, and $\mathbb{E}X_{n,m} = 0$. When does $S_n \to \mathcal{N}(0, \sigma^2)$ in distribution?

The basic condition is to require the triangular array to be **uniformly asymptotically negligible** (UAN):

$$\sum_{k=1}^{k(n)} \mathbb{P}(|X_{n,m}| > \epsilon) \to 0 \text{ as } n \to \infty.$$

Note that

$$\begin{aligned} \text{UAN} &\iff \prod_{k=1}^{k(n)} (1 - \mathbb{P}(|X_{n,k}| > \epsilon) \to 1 \\ &\iff \prod_{k=1}^{k(n)} \mathbb{P}(|X_{n,k}| \leqslant \epsilon) \to 1 \\ &\iff \mathbb{P}(|X_{n,m}| = \epsilon \text{ for all } n \leqslant k(n)) \to 1. \end{aligned}$$

To get the result from CLT we want $\varphi_{S_n}(t) = \prod_{k=1}^{k(n)} \varphi_{X_{n,k}}(t) = \prod_{k=1}^{k(n)} \left(1 - \frac{t^2\sigma_{n,k}^2}{2} + \text{ some error}\right) \to e^{-t^2\sigma^2/2}$.

Proposition

Let $\lambda_{n,m} \in \mathbb{C}$ form a triangular array. If

(1) $\sum_{k=1}^{k(n)} \lambda_{n,k} \to \lambda$,

(2) $\max_{\text{row}} |\lambda_{n,m}| \to 0$ as $n \to \infty$, and

(3) $\sup_{n \geqslant 1} \sum_{k=1}^{k(n)} |\lambda_{n,k}| < \infty$,

then $\prod_{k=1}^{k(n)} (1 + \lambda_{n,k}) \to e^{\lambda}$.

Proof. We consider $\left|\log \prod_{k=1}^{k(n)} (1 + \lambda_{n,k}) - \sum_{k=1}^{k(n)} \lambda_{n,k}\right|$:

$$\begin{aligned} \text{LHS} &= \sum_{k=1}^{k(n)} |\log(1 + \lambda_{n,k}) - \lambda_{n,k}| \\ &\leqslant K \sum_{k=1}^{k(n)} |\lambda_{n,k}|^2 \\ &\leqslant K \left(\sup_n \sum_{k=1}^{k(n)} |\lambda_{n,k}|\right) \max_{m \geqslant k_n} |\lambda_{n,m}| \to 0. \end{aligned}$$

Finally since $\sum_{k=1}^{k(n)} \lambda_{n,k} \to \lambda$, we are done. □

Proposition: D3.4.3

Let $z_1, \dots, z_n, w_1, \dots, w_n \in \mathbb{C}$, all with modulus $\leqslant K$. Then

$$\left| \prod_{m=1}^{n} z_m - \prod_{m=1}^{n} w_m \right| \leqslant K^{n-1} \sum_{m=1}^{n} |z_m - w_m|.$$

Proof. Trivial for $n = 1$. If $n > 1$, we use triangle inequality to remove z_1 and w_1 to get

$$\begin{aligned} \left| \prod_{m=1}^{n} z_m - \prod_{m=1}^{n} w_m \right| &\leqslant \left| z_1 \prod_{m=2}^{n} z_m - z_1 \prod_{m=2}^{n} w_m \right| + \left| z_1 \prod_{m=2}^{n} w_m - w_1 \prod_{m=2}^{n} w_m \right| \\ &\leqslant K \left| \prod_{m=2}^{n} z_m - \prod_{m=2}^{n} w_m \right| + K^{n-1} |z_1 - w_1|. \end{aligned}$$

□

Theorem: D3.4.10, Lindeberg-Feller Theorem

Let $X_{n,m}$ be a triangular array with independent random varaibles and zero mean. If

(1) $\sum_{m=1}^{n} \mathbb{E} X_{n,m}^2 \to \sigma^2 > 0$, and

(2) $\sum_{m=1}^{n} \mathbb{E}(X_{n,m}^2 1_{\{|X_{n,m}| > \epsilon\}}) \to 0$ for all ϵ,

then the row sums S_n converges to $\mathcal{N}(0, \sigma^2)$ in distribution.

Beginning of Nov. 9, 2022

Note that (ii) implies UAN by Chebyshev. Also, L-F CLT covers the i.i.d. case which we are already familiar with: for $X_1, X_2, \dots$ i.i.d., we simply let $X_{n,m} = X_m / \sqrt{n}$.

Proof of L-F CLT. We let $\varphi_{n,m}$ be the ch.f. of $X_{n,m}$, and similarly $\sigma_{n,m}^2$ the variance $= \mathbb{E} X_{n,m}^2 = \operatorname{var}(X_{n,m})$. First observation:

$$\begin{aligned} \max_{m \leqslant n} \sigma_{n,m}^2 &= \max_{m \leqslant n} [\mathbb{E}(X_{n,m}^2 1_{|X_{n,m}| \leqslant \epsilon}) + \mathbb{E}(X_{n,m}^2 1_{|X_{n,m}| > \epsilon})] \\ &\leqslant \epsilon^2 + \sum_{k=1}^{n} \mathbb{E}(X_{n,k}^2 1_{|X_{n,k}| > \epsilon}). \end{aligned}$$

The sum $\to 0$ by (ii), so $\max_{m \leqslant n} \sigma_{n,m}^2 \to 0$.

Next, we note that $\varphi_{S_n}(t) = \prod_{m=1}^{n} \varphi_{X_{n,m}}(t)$ and compare this to $\prod_{m=1}^{n} (1 - (t^2 \sigma_{n,m}^2)/2)$. Note that $|\varphi_{n,m}| \leqslant 1$, and $|1 - (t^2 \sigma_{n,m}^2)/2| \leqslant 1$ for n large and t fixed by the observation above. Therefore, by D3.4.3 (the inequality just shown above)

$$\left| \varphi_{S_n}(t) - \prod_{m=1}^{n} (1 - t^2 \sigma_{n,m}^2 / 2) \right| \leqslant \sum_{m=1}^{n} |\varphi_{X_{n,m}}(t) - 1 + t^2 \sigma_{n,m}^2 / 2|.$$

From D3.3.20 we can bound the error of expansions by $|\varphi_{X_{n,m}}(t) - 1 + t^2 \sigma_{n,m}^2 / 2| \leqslant t^2/6 \cdot \mathbb{E} \min(|t||X_{n,m}|^3, 6|X_{n,m}|^2)$. For $|X_{n,m}| \leqslant \epsilon$ we consider $|X_{n,m}|^2$; otherwise we consider the latter. We thus obtain the following bound:

$$|\varphi_{X_{n,m}}(t) - 1 + t^2 \sigma_{n,m}^2 / 2| \leqslant \frac{t^2}{6} \mathbb{E}(|t||X_{n,m}|^3 1_{|X_{n,m}| \leqslant \epsilon}, 6|X_{n,m}|^2 1_{|X_{n,m}| > \epsilon}).$$

Bounding $|t||X_{n,m}|^3$ by $t\epsilon|X_{n,m}|^2$ and using assumption (ii) yield

$$\limsup_{n\to\infty}\sum_{m=1}^{n}|\varphi_{X_{n,m}}(t)-1+t^2\sigma_{n,m}^2/2|\leqslant\frac{t^3}{6}\epsilon\sum_{m=1}^{n}\sigma_{n,m}^2+0.$$

Convergence of $\sum\sigma_{n,m}^2$ imply in particular that they are bounded in n, regardless of ϵ. Thus the upper limit is 0. It remains to notice that $\prod_{m=1}^{n}(1-t^2\sigma_{n,m}^2/2)\to\exp(-t^2\sigma^2/2)$, the ch.f. of $\mathcal{N}(0,\sigma^2)$. □

Example: Normal approximation to binomial. Let S_n be the number of success in n independent trials, each with success probability p. If A_i is the event of success on trial i, then $S_n=\sum_{i=1}^n 1_{A_i}$. We have $\mathbb{E}1_{A_i}=p$ and $\operatorname{var}(1_{A_i})=p(1-p)$, so $(S_n-np)/\sqrt{np(1-p)}\to\mathcal{N}(0,1)$ in distribution. For integer valued distributions, we often apply continuity corrections to obtain better approximation results.

3.5 Poisson Convergence & Poisson Processes

Beginning of Nov. 14, 2022

Theorem: D3.6.1

Let $\{A_{n,m}\}$ be a triangular array with n on n^{th} row. Assume it is row-independent (within each row). Let $S_n=\sum_{m=1}^n 1_{A_{n,m}}$. If $\sum_{m=1}^n\mathbb{P}(A_{n,m})\to\lambda$ as $n\to\infty$, and if $\max_{m\leqslant n}\mathbb{P}(A_{n,m})\to 0$ as $n\to\infty$, then $S_n\to$ Poisson with parameter λ as $n\to\infty$.

Proof. Note that

$$\varphi_{S_n}(t)=\prod_{m=1}^{n}(1+\mathbb{P}(A_{n,m})(e^{it}-1)).$$

Let $\lambda_{n,m}=\mathbb{P}(A_{n,m})(e^{it}-1)$. By assumption, $\sum_{m=1}^n\lambda_{n,m}\to\lambda(e^{it}-1)$, and $\sum_{m=1}^n|\lambda_{n,m}|\leqslant 2\sum_{m=1}^n\mathbb{P}(A_{n,m})\to 2\lambda$ and is in particular bounded. Finally, $\max_{m\leqslant n}|\lambda_{n,m}|\to 0$ as $n\to\infty$. By a previous proposition,

$$\prod_{m=1}^{n}(1+\lambda_{n,m})\to e^{\lambda(e^{it}-1)},$$

the ch.f. of a parameter λ Poisson. □

Theorem: D3.7.1

Let $\{X_{n,m}\}$ be a row-independent triangular array with $m\leqslant n$, $n\geqslant 1$. Assume X are integer valued random variables. Let $p_{n,m}=\mathbb{P}(X_{n,m}=1)$, $\epsilon_{n,m}=\mathbb{P}(X_{n,m}\geqslant 2)$, and $S_n=\sum_{m=1}^n X_{n,m}$. If $\sum_{m=1}^n p_{n,m}\to\lambda\in(0,\infty)$, $\max_{m\leqslant n}p_{n,m}\to 0$, and $\sum_{m=1}^n\epsilon_{m,n}\to 0$, then $S_n\to\text{Poisson}(\lambda)$. Namely, if $\mathbb{P}(X_{n,m}\geqslant 2)$ is sufficiently small, the result still holds.

Proof. Let $X'_{n,m}=1_{\{X_{n,m}=1\}}$ and S'_n the row sum of $X'_{n,m}$. By D3.6.1 $S'_n\to\text{Poisson}(\lambda)$. It remains to apply Slutsky's theorem to obtain convergence in distribution of S_n as well. This is indeed true:

$$\mathbb{P}(S_n\neq S'_n)\leqslant\sum_{m=1}^{n}\epsilon_{n,m}\to 0.$$ □

Example: Birthdays.... Consider n friends, and let N be the umber of days (out of 365) with no birthdays among these friends. Then $\mathbb{P}(\text{no one on a fixed day}) = (1-1/365)^n \approx e^{-n/365}$, so $\mathbb{E}N = 365e^{-n/365}$. Taking $n = 365\log(365/\lambda)$ gives $\mathbb{E}N \approx \lambda$.
This, however, does not take into dependency into account. If no one has birthday on Jan 1, then the probability of no one having birthday on Jan 2 is slightly smaller with this prior information. We can show that if we replace 365 with r_n and N with r_n, and if $r_n/n \cdot \log(r_n/\lambda) \to 1$, then $N_n \to \text{Poisson}(\lambda)$.

Poisson Processes

We now consider an arrival problem. Let λ be the arrival rate. Let $N(s,t)$ be the $\mathbb{Z}$-valued number of arrivals in $(s,t]$.
Suppose the following:

(1) disjoint time intervals are independent,

(2) distribution of $N(s,t)$ depend only on $t-s$ (once λ is fixed),

(3) $\mathbb{P}(N(0,h)=1) = \lambda h + o(h)$ as $h \to 0$, and

(4) $\mathbb{P}(N(0,h) \geqslant 2) = o(h)$.

Theorem: D3.7.2

If $N(\cdot,\cdot)$ satisfies the above assumptions, then $N(s,t)$ is poisson distributed with parameter $\lambda(t-s)$, for all $s<t$.

Proof. WLOG assume $s=0$. We divide $[0,\lambda t]$into n equal subintervals and let $X_{n,m} = N((m-1)/n\cdot\lambda t, m/n\cdot\lambda t)$. The row sums are just

$$S_n = \sum_{m=1}^{n} X_{n,m} = N(0,\lambda t).$$

Observe that $\mathbb{P}(X_{n,m}=1) = \lambda t/n + o(1/n)$ as $n\to\infty$, keeping λ, t fixed. Also, $\mathbb{P}(X_{n,m}\geqslant 2) = o(1/n)$. Applying D3.7.1, we see $S_n \to \text{Poisson}(\lambda t)$. This holds for all n, so $N(0,\lambda t) \sim \text{Poisson}(\lambda t)$. ☐

If T_1 is the time of the first arrival, then $\mathbb{P}(T_1 > t) = \mathbb{P}(N(0,t)=0) p \mathbb{P}(\text{Poisson}(\lambda t) = 0) = e^{-\lambda t}$, so $T_1 \sim \text{exponential}(\lambda)$. We will later show that the gaps between different arrivals are also i.i.d. exponential(λ).

Multivariate Normal

Beginning of Nov. 18, 2022

Let $X = (X_1, \ldots, X_\ell)$ be a random vector with $\text{var}(X_i) < \infty$ and covariance matrix $\Sigma_{i,j} = \text{cov}(X_i, X_j)$. Then for any vector θ,

$$\text{var}(\theta\cdot X) = \text{var}(\sum_{i=1}^{\ell}\theta_i X_i) = \sum_{i,j}\theta_i\theta_j \,\text{cov}(X_i,X_j) = \theta^T\Sigma\theta \in \mathbb{R}.$$

This shows Σ is PSD and symmetric. If T is a linear transformation of X, then

$$\text{cov}((TX)_i, (TX)_j) = \text{cov}(\sum_k T_{i,k}X_k + \sum_\ell T_{j,\ell}X_\ell) = (T\Sigma T^T)_{i,j}.$$

From linear algebra, since Σ is PSD, there exists an unitary U ($U^{-1} = U^*$ and orthonormal) such that $U^T\Sigma U$ is diagonal. From our remark above, viewing U^T as a linear transformation, the resulting random vector U^TX has uncorrelated components.

If X has density f_X and T a multivaraible linear transformation, then TX has density

$$f_{TX}(x) = \frac{1}{|\text{det}T|} f_X(T^{-1}x).$$

Finally, we are ready to talk about multivariate normal distribution.
Of course, the standard multivariate normal has each coordinate as an independent $\mathcal{N}(0,1)$. The density

$$f(X) = (2\pi)^{-d/2}\exp(-\sum_{i=1}^d x_i^2/2) = (2\pi)^{-1/2}\exp(-x^T I x/2) =: \mathcal{N}(0, I).$$

If T is invertible then

$$f_{TX}(x) = (2\pi)^{-d/2}|\text{det}T|^{-1}\exp(-x^T T^{-T}T^{-1}x/2) = (2\pi)^{-d/2}|\text{det}T|^{-1}\exp(-x(TT^*)^{-1}x/2).$$

This gives rise to a more general multivariable normal $\mathcal{N}(\mu, \Sigma)$, whose density is

$$f(x) = (2\pi)^{-d/2}||\Sigma|^{-1/2}\exp(-(x-\mu)^T\Sigma^{-1}(x-\mu)/2)$$

where $\mu \in \mathbb{R}^k$ and Σ PSD.

Beginning of Nov. 21, 2022

For a degenerate mnultivariate normal, consider $\mu = 0$ and $r < d$ (rank). We take $X_1, \ldots, X_r \sim \mathcal{N}(0, \tilde{S})$ where $\tilde{S}$ is invertible. we put $X = (X_1, \ldots, X_r, 0, \ldots, 0)$ corresponding to block diagonal $\tilde{S}, 0$. Then TX has covariance matrix TST^T. Given Σ, we want to choose $\tilde{S}, T$ so that $TST^T = \Sigma$.
We know there exists a unitary matrix T with $T\Sigma T$ = diagonal $(\lambda_1^2, \ldots, \lambda_r^2, 0, \ldots)$. Let $\tilde{D}$ = diagonal $(\lambda_1^2, \ldots, \lambda_r^2)$ and let $\tilde{X} \sim \mathcal{N}(0, \tilde{D})$, $X = (\hat{X}, 0, \ldots, 0)$. Then TX has covariance matrix TDT^T since $T^T = T^{-1}$.

Proposition

If $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ nonsingular, then the marginals X_i are normal.

Proof. WLOG $\mu = 0$ and we are looking at the first coordinate, X_1.
The claim is easy if Σ has first row $(\sigma_1^2, 0, \ldots, 0)^T$ and column $(\sigma_1^2, 0, \ldots, 0)$. In this case,

$$f_X(x) = C\exp(-x^T\Sigma^{-1}x/2) = C\exp(-\frac{x_1^2}{2\sigma_1^2} - g(x_2, \ldots, x_d)).$$

Therefore

$$\begin{aligned} f_{X_1}(x) &= \int_{x_2,\ldots,x_d} f_X(x_1, \ldots, x_d)\, \mathrm{d}x_2 \ldots x_d \\ &= \text{Const}\exp(-\frac{x_1^2}{2\sigma_1^2}). \end{aligned}$$

Since f_{X_1} integrates to 1 the constant must match up, so $X_1\mathcal{N}(0,\sigma_1^2)$.
For the general case, we need to find T so that TX's covariance has the special form with $(TX)_1 = X_1$.
We take unitary U such that the 1^{st} row of U is perpendicular to the j^{th} column of $\Sigma^{-1/2}$ $j \geqslant 2$. Take $T = U\Sigma^{-1/2}$.
Then TX has covariance matrix

$$T\Sigma T^T = U\Sigma^{-1/2}\Sigma\Sigma^{-1/2}U^T = UU^T = I.$$

[To be fixed] □

Example. Multivariate normal implies normal marginals, but not the converse. For example let $X = \mathcal{N}(0,1)$ and $\xi = \pm 1$ with probably 0.5 each, independent of X.
Let $Y = (X, \xi X)$, so it's on either diagonal with probability 0.5. Clearly Y is not a bivariate normal, even if its covariance matrix is I.

If $X, X_2, \ldots, X_d$ are independent $\mathcal{N}(\mu_i, \sigma_i^2)$, then $X = (X_1, \ldots, X_d) \sim \mathcal{N}(\mu, \Sigma)$ with Σ = diagonal $(\sigma_1^2, \ldots, \sigma_d^2)$. Since

$$f_X(x) = \prod_{i=1}^d f_{X_i}(x_i) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_i}\exp(-x_i/(2\sigma_i^2)) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp(-x^T\Sigma^{-1}x/2),$$

we obtain the following result:

Proposition

While not necessarily true for other distributions, for $(X_1, \ldots, X_n)$ multivariate normal, X_i's are uncorrelated iff X_i's are independent.
(The previous example we have shows that if (X_1, X_2) is not multivariate normal, even if it has normal marginals, then (uncorrelated but dependent) can happen.)

If $X \sim \mathcal{N}(0, \Sigma)$, and T is invertible, then TX has density

$$\begin{aligned} f_{TX}(x) &= \frac{1}{|T|}f_X(T^{-1}x) = \frac{1}{(2\pi)^{d/2}|T||\Sigma|^{1/2}}\exp(x^TT^{-T}\Sigma^{-1}T^{-1}x/2) \\ &= \frac{1}{(2\pi)^{d/2}|T\Sigma T^T|^{1/2}}\exp(-x^T(T\Sigma T^T)^{-1}x/2) \sim \mathcal{N}(0, T\Sigma T^T). \end{aligned}$$

Therefore the marginals are normal, in particular $(TX)_1$. Since T is arbitrary, $\theta \cdot X$ is normal for all $\theta \in \mathbb{R}^d$, with $\operatorname{var}(\theta \cdot X) = \theta^T\Sigma\theta$.
More generally, if $X \sim \mathcal{N}(\mu, \Sigma)$ and T is invertible then $TX \sim \mathcal{N}(T\mu, T\Sigma T^T)$.

Characteristic functions of $\mathcal{N}(\mu, \Sigma)$

$$\varphi_X(\theta) = \mathbb{E}e^{i\theta\cdot X} = \varphi_{\theta\cdot X}(1) = \exp(-\operatorname{var}(\theta\cdot X)/2) = \exp(-\theta^T\Sigma\theta/2).$$

CLT in $\mathbb{R}^d$

Theorem

Let $X_1, X_2, \ldots$ be i.i.d. in $\mathbb{R}^d$ with finite mean $\mathbb{E}X_1 = \mu$ and finite covaraince matrix Σ. Let $S_n = X_1 + \ldots + X_n$. Then $(S_n - n\mu)/\sqrt{n}$ converges in distribution to $\mathcal{N}(0, \Sigma)$.

Proof. By Cramer-Wold it suffices to show

$$\theta \cdot \frac{S_n - n\mu}{\sqrt{n}} \to \theta \cdot X \text{ in distribution for all } \theta \in \mathbb{R}^d.$$

Note $\mathrm{var}(\theta \cdot X) = \theta^T \Sigma \theta$ so $\theta \cdot X \sim \mathcal{N}(0, \theta^T \Sigma \theta)$. This is a one-dimensional distribution, so

$$\theta \cdot \frac{S_n - n\mu}{\sqrt{n}} = \frac{\sum_{i=1}^n X_i \cdot \theta - n\theta \cdot \mu}{\sqrt{n}} \to \mathcal{N}(0, \mathrm{var}(X_i \cdot \theta)) = \mathcal{N}(0, \theta^T \Sigma \theta).$$

□

3.6 Conditional Probabilities

It can be shown that given n coin tosses, $X = (\text{number of heads})^2$ has mean $(n + n^2)/2$. Viewing n as a variable we therefore obtain

$$\mathbb{E}(X \mid N = n) = \frac{n + n^2}{4} \text{ or more geneerally } \mathbb{E}(X \mid N) = \frac{N + N^2}{4}.$$

Now consider the integral over an event in $\sigma(\mathbb{N})$, say $\{N \leqslant 2\}$. Average of X on $\{N = n\}$ is

$$\frac{1}{\mathbb{P}(N = n)} \int_{\{N=n\}} X \, \mathrm{d}\mathbb{P}$$

so

$$\int_{\{N \leqslant 2\}} X \, \mathrm{d}\mathbb{P} = \sum_{n=0}^{2} \int_{\{N=n\}} X \, \mathrm{d}\mathbb{P} = \int_{\{N \leqslant 2\}} \frac{1}{4}(N^2 + N) \, \mathrm{d}\mathbb{P}$$

and more generally, for any $\{N \in A\} \in \sigma(\mathbb{N})$,

$$\int_{\{N \in A\}} X \, \mathrm{d}\mathbb{P} = \int_{\{N \in A\}} \frac{N + N^2}{4} \, \mathrm{d}\mathbb{P}.$$

Another example: consider $\Omega = [0, 1] = A_1 \cup A_2 \cup A_3$ disjoint, and let X be a r.v. and $\mathcal{Y} = \sigma(A_1, A_2, A_3)$. Let Y be constant on each A_j with value $\frac{1}{\mathbb{P}(A_j)} \int_{A_j} X \, \mathrm{d}\mathbb{P}$. Then Y is (the only r.v.) measurable w.r.t. $\mathcal{Y}$ and that

$$\int_B Y \, \mathrm{d}\mathbb{P} = \int_B X \, \mathrm{d}\mathbb{P} \qquad \text{for all } B \in \mathcal{Y}.$$

General case

Let X be a r.v. with $\mathbb{E}|X| < \infty$ on $(\Sigma, \mathcal{F}_0, \mathbb{P})$

Suppose we have partial information to

$$\mathcal{F} = \{\text{all events known to occur or not}\}.$$

$\mathcal{F}$ is a σ-algebra. We want to formalize $\mathbb{E}(X \mid \mathcal{F})$:

Definition

Let $(\Omega, \mathcal{F}_0, \mathbb{P})$ be a probability space and let $\mathcal{F} \subset \mathcal{F}_0$ be a σ-algebra. Let X be a r.v. with $\mathbb{E}|X| < \infty$. We define $\mathbb{E}(X \mid \mathcal{F})$ to be any r.v. Y with

(1) $Y \in \mathcal{F}$, and

(2) $\int_A Y \, \mathrm{d}\mathbb{P} = \int_A X \, \mathrm{d}\mathbb{P}$, for all $A \in \mathcal{F}$.

Lemma

If Y, Y' satisfy (1) and (2) above, then $Y = Y'$ a.s.

Proof. Fix $\epsilon > 0$. Consider $A = \{Y - Y' \geqslant \epsilon\}$. By (2)

$$0 = \int_A (Y - Y') \, \mathrm{d}\mathbb{P} \geqslant \epsilon \mathbb{P}(A)$$

so $\mathbb{P}(A) = 0$. □

Lemma

Y satisfying (1) and (2) exists.

Proof. Consider measures on $\mathcal{F}$ *only*:

$$\tilde{P} = \mathbb{P}\big|_{\mathcal{F}} \qquad \text{and} \qquad \nu(A) = \int_A X \, \mathrm{d}\mathbb{P}, \text{ for } A \in \mathcal{F}.$$

Clearly if $\tilde{P}(A) = 0$ we have $\nu(A) = 0$, so $\nu \ll \tilde{P}$ (absolute continuity). By Radon-Nikodym there exists a density $\frac{\mathrm{d}\nu}{\mathrm{d}\tilde{P}}$, $\mathcal{F}$-measurable, such that for all $A \in \mathcal{F}$,

$$\int_A X \, \mathrm{d}\mathbb{P} = \nu(A) = \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\tilde{P}} \, \mathrm{d}\tilde{P} = \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\tilde{P}} \, \mathrm{d}\mathbb{P}.$$

That is, setting Y to the Radon-Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\tilde{P}}$ works. □

Properties of conditional probabilities:

(1) $\mathbb{P}(A \mid \mathcal{F}) = \mathbb{E}(1_A \mid \mathcal{F})$ (definition),

(2) $\mathbb{P}(A \mid \mathcal{F}) \in \mathcal{F}$, and

(3) $\int_B \mathbb{P}(A \mid \mathcal{F}) \, \mathrm{d}\mathbb{P} = \int_B 1_A \, \mathrm{d}\mathbb{P} = \mathbb{P}(A \cap B)$ for all $B \in \mathcal{F}$.