An Introduction to Probability Theory: Outline

1. Definitions: sample space and (measurable) random variables
2. σ-algebras
3. Expectation (integration)
4. Conditional expectation
5. Useful inequalities
6. Independent random variables
7. The central limit theorem
8. Laws of large numbers: Borel-Cantelli lemma
9. Uniform integrability
10. Kolmogorov's extension theorem for consistent finite-dimensional distributions

Sample space and events

- Consider a random experiment resulting in an *outcome* (or "sample"), ω.
- *E.g.*, the experiment is a pair of dice thrown onto a table and the outcome is the exact orientation of the dice and their position on the table when they stop moving.
- The space of all outcomes, Ω, is called the *sample space*, *i.e.*, $\omega \in \Omega$.
- An *event* is merely a subset of Ω, *e.g.*, "the sum of the dots on the upward facing surfaces of the dice is 7".
- We say that an event $A \subset \Omega$ *has occurred* if the outcome ω of the random experiment belongs to A, *i.e.*, $\omega \in A$, so
  - events A *and* B occurred if $\omega \in A \cap B$, and
  - events A *or* B occurred if $\omega \in A \cup B$.
- A sample space Ω is an abstract, unordered set in general.
- Let $\mathcal{F}$ be the set of events, *i.e.*, $A \in \mathcal{F} \Rightarrow A \subset \Omega$.

Probability on a sample space

- A *probability measure* P maps each event $A \subset \Omega$ to a real number between zero and one inclusive, *i.e.*, $\mathsf{P}(A) \in [0,1]$.

- A probability measure has certain properties:

  1. $\mathsf{P}(\Omega) = 1$ and
  2. $\mathsf{P}(A) = 1 - \mathsf{P}(A^c)$ $\forall$ events A, where $A^c = \{\omega \in \Omega \mid \omega \notin A\}$ is the complement of A.

- Moreover, if the events $\{A_i\}_{i=1}^n$ are disjoint (*i.e.*, $A_i \cap A_j = \emptyset$ for all $i \neq j$), then

$$\mathsf{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathsf{P}(A_i),$$

  i.e., P is *finitely additive*.

- Formally, a probability measure is defined to be *countably* additive:

  3. For any disjoint $\{A_i\}_{i=1}^\infty$,

$$\mathsf{P}\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mathsf{P}(A_i),$$

Probability measures on σ-algebras

- On large sample spaces Ω (*e.g.*, $\Omega = \mathbb{R}$), a formal probability measure may be impossible to construct if *all* subsets of Ω are defined as events, *i.e.*, if $\mathcal{F} = 2^\Omega$ (the power set of all subsets of Ω), *cf.*, Caratheodory's Extension Theorem.

- So, the set of events $\mathcal{F}$ is restricted to a *σ-algebra* (or *σ-field*) of subsets of Ω formally satisfying the following properties:

  1. $\Omega \in \mathcal{F}$ (possesses intersection identity)
  2. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ (closed under complementation)
  3. if $A_1, A_2, A_3, ... \in \mathcal{F}$, $\cap_{n=1}^\infty A_n \in \mathcal{F}$ (closed under countable intersections)

- The probability measure P is defined only on the σ-algebra $\mathcal{F} \subset 2^\Omega$,

$$\mathsf{P} : \mathcal{F} \to [0,1].$$

- We have thus identified a fundamental probability (measure) space: $(\Omega, \mathcal{F}, \mathsf{P})$.

- Note: Equivalently by De Morgan's theorem, can use $\emptyset \in \mathcal{F}$ (union identity) and closed under countable unions instead of conditions 1 and 3 above.

Conditioned events

- The probability that A occurred *conditioned on* (or "given that") another event B occurred is $\mathsf{P}(A|B) := \mathsf{P}(A \cap B)/\mathsf{P}(B)$, where $\mathsf{P}(B) > 0$ is assumed.

- A group of events $A_1, A_2, ..., A_n$ are said to be *mutually independent* if

$$\mathsf{P}\left(\bigcap_{i\in\mathcal{I}} A_i\right) = \prod_{i\in\mathcal{I}} \mathsf{P}(A_i) \quad \forall\, \mathcal{I} \subset \{1,2,...,n\}.$$

- Note if events A and B are independent and $\mathsf{P}(B) > 0$, then $\mathsf{P}(A|B) = \mathsf{P}(A)$, *i.e.*, knowledge that the event B has occurred has no bearing on the probability that the event A has occurred as well.

- Given that B has occurred with $\mathsf{P}(B) > 0$:
  - The set of events $\mathcal{F}_B := \{A \cap B \mid A \in \mathcal{F}\}$ is itself a σ-algebra, and
  - $\mathsf{P}(\cdot|B)$, also a probability measure for $(\Omega, \mathcal{F})$ and $(B, \mathcal{F}_B)$, addresses the residual uncertainty in the random experiment given that the event B has occurred.
  - On $(\Omega, \mathcal{F})$, $\mathsf{P}(A) = 0 \Rightarrow \mathsf{P}(A|B) = 0\ \forall A \in \mathcal{F}$, *i.e.*, $\mathsf{P}(\cdot|B)$ is *absolutely continuous* w.r.t. P.

Random variables

- A *random variable* X is a real-valued function with domain Ω, $X : \Omega \to \mathbb{R}$.

- So, $X(\omega)$ is a real number representing some feature of the outcome ω.

- *E.g.*, in a dice-throwing experiment, $X(\omega)$ could be defined as just the sum of the dots on the upward-facing surfaces of outcome ω (which is the configuration of the dice on the table when they stop moving).

- For random variables, we are typically interested in the probability of the event that X takes values in a contiguous interval B of the real line (including singleton points), or some union of such intervals, *i.e.*,

$$\mathsf{P}(X \in B) \ := \ \mathsf{P}(\{\omega \in \Omega \mid X(\omega) \in B\}) \ =: \ \mathsf{P}(X^{-1}(B)).$$

- To ensure that the fundamental probability space $(\Omega, \mathcal{F}, \mathsf{P})$ is capable of evaluating the probabilities of such events, we formally define random variables as being *measurable*.

- To explain measurability, we need to first define the *Borel* σ-algebra of subsets of $\mathbb{R}$ that is *generated* by contiguous intervals of $\mathbb{R}$.

The Borel σ-algebra on $\mathbb{R}$

- Consider contiguous intervals of the real line, *e.g.*,

$$\begin{aligned}[x,\infty) &= \{z \in \mathbb{R} \mid z \geq x\} \text{ or} \\ (x,y] &= \{z \in \mathbb{R} \mid x < z \leq y\} \textit{ etc.}\end{aligned}$$

- Define $\sigma(\mathcal{A})$ as the *smallest* σ-algebra containing all elements of elements $\mathcal{A}$, *i.e.*, *generated* by $\mathcal{A}$.

- The Borel σ-algebra is

$$\mathcal{B} := \sigma([x,\infty) \mid x \in \mathbb{R}).$$

- Note that the singleton sets $\{x\} \in \mathcal{B} \quad \forall x \in \mathbb{R}$ and that, *e.g.*,

$$\begin{aligned}\mathcal{B} &= \sigma((-\infty,x] \mid x \in \mathbb{R}) \\ &= \sigma([x,y) \mid x \leq y,\ x,y \in \mathbb{Q}) \text{ etc}\end{aligned}$$

  To see the first equality, note that $[x,\infty)^c = (-\infty,x)$ and that we can define a monotonically decreasing sequence x_n converging to x so that $\cap_{n=1}^{\infty}(-\infty,x_n) = (-\infty,x]$.

- The Vitali subset of $\mathbb{R}$ is *not* in the Borel σ-algebra, *i.e.*, $\mathcal{B} \neq 2^{\mathbb{R}}$.

- Indeed, the cardinality of $\mathcal{B}$ is only that of $\mathbb{R}$.

Measurability of random variables

- Formally, random variables are defined to be *measurable* with respect to $(\Omega,\mathcal{F})$, *i.e.*,

$$X^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{B},$$

  so that $\mathsf{P}(X \in B)$ is well-defined $\forall B \in \mathcal{B}$.

- A random variable X *induces* a probability measure P_X on $(\mathbb{R},\mathcal{B})$ (the *distribution* of X),

$$\mathsf{P}_X(B) := \mathsf{P}(X \in B),$$

  so that $(\mathbb{R},\mathcal{B},\mathsf{P}_X)$ is also a probability space.

- Note: If a function $g:\mathbb{R}\to\mathbb{R}$ is $(\mathbb{R},\mathcal{B})$-measurable (*i.e.*, $g^{-1}(B) \in \mathcal{B}\ \forall B \in \mathcal{B}$) and X is a random variable, then $g(X)$ is a random variable too.

- Note: The cumulative distribution function (CDF) of X is just $F_X(x) := \mathsf{P}_X((-\infty,x]) = \mathsf{P}(X \leq x)$.

Measurable compositions of random variables

If Y, X and $X_1, X_2, X_3, \ldots$ are all extended random variables, then the following are also random variables:

- $\min\{X, Y\}$, $\max\{X, Y\}$, XY, $\mathbf{1}\{X \neq 0\}/X$ where $\frac{0}{0} := 1$.
- $\alpha X + \beta Y\ \forall \alpha, \beta \in \mathbb{R}$.
- $\sup_{n \geq 1} X_n$, $\inf_{n \geq 1} X_n$, $\limsup_{n \to \infty} X_n$, $\liminf_{n \to \infty} X_n$.

σ-algebra generated by a random variable

- Define $\sigma(X) := \sigma(\{X^{-1}(B) \mid B \in \mathcal{B}\})$, *i.e.*, the smallest σ-algebra of events for which the random variable X is measurable.
- Note: One can directly show that
$$\sigma(X) = \sigma(\{X^{-1}([x, \infty)) \mid x \in \mathbb{R}\}),$$
i.e., considering only a "generating" subset of $\mathcal{B}$.
- $\sigma(X)$ captures the "information" gained by the knowledge of $X(\omega)$ about the outcomes ω.
- *E.g.*, If X is constant then $\sigma(X) = \{\emptyset, \Omega\}$.
- *E.g.*, If $X = \mathbf{1}_B$ for an event $B \in \mathcal{F}$,
  - where the the *indicator function* $\mathbf{1}_B(\omega) = 1$ if $\omega \in B$ and $\mathbf{1}_B(\omega) = 0$ else (also may write $\mathbf{1}B := \mathbf{1}_B$),
  - *i.e.*, X is a *Bernoulli distributed* random variable,
  - then $\sigma(X) = \{\emptyset, B, B^c, \Omega\}$.
  - If the scalars $a \neq b$, then $Y := a\mathbf{1}_B + b\mathbf{1}_{B^c}$ also indicates whether B or B^c has occurred,
  - *i.e.*, $\sigma(X) = \sigma(Y)$ in this case.
- **Doob's Theorem:** If Y is $\sigma(X)$-measurable (so that $\sigma(Y) \subset \sigma(X)$), then $\exists$ Borel measurable g such that $Y = g(X)$ a.s.
- If g is one-to-one, then $\sigma(Y) = \sigma(X)$.

Independent random variables

- The random variables $X_1, X_2, ..., X_n$ are said to be *mutually independent* (or just "independent") if and only if

$$\mathsf{P}(\cap_{i=1}^n \{X_i \in B_i\}) = \prod_{i=1}^n \mathsf{P}(X_i \in B_i) \quad \forall B_1, ..., B_n \in \mathcal{B}.$$

- Clearly mutual independence implies that the *joint* CDF

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) \quad := \quad \mathsf{P}(\cap_{i=1}^n \{X_i \le x_i\})$$

  satisfies

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \prod_{i=1}^n F_{X_i}(x_i) \quad \forall x_1, ..., x_n \in \mathbb{R}.$$

- To prove the converse statement, we will now discuss monotone class theorems.

- Note: The *marginal* $F_{X_1}(x_1) = F_{X_1,...,X_n}(x_1, \infty, \infty, ..., \infty)$.

Monotone class theorems

- $\mathcal{C} \subset 2^\Omega$ is a π-*class* over Ω if $A, B \in \mathcal{C} \;\Rightarrow\; A \cap B \in \mathcal{C}$.

- $\mathcal{C} \subset 2^\Omega$ is a λ-*class* over Ω when:
  (i) $\Omega \in \mathcal{C}$;
  (ii) if $A, B \in \mathcal{C}$ and $A \subset B \Rightarrow B \backslash A := B \cap A^c \in \mathcal{C}$; and
  (iii) if $A_1, A_2, ... \in \mathcal{C}$ is monotonically increasing $(A_n \subset A_{n+1} \; \forall n)$, then $\cup_{n=1}^\infty A_n \in \mathcal{C}$.

- **Proposition:** If $\mathcal{C}$ is a λ-class over Ω, then
  (a) $A \in \mathcal{C} \;\Rightarrow\; A^c \in \mathcal{C}$ (by (i) and (ii), $\Omega \backslash A \equiv A^c \in \mathcal{C}$).
  (b) $A, B \in \mathcal{C}$ and $A \cap B = \emptyset$ (*i.e.*, they're disjoint), then $A \cup B \in \mathcal{C}$ ($A \subset B^c \Rightarrow (B^c \backslash A)^c \equiv B \cup A \in \mathcal{C}$ by (ii) and (a)).
  (c) if $\mathcal{C}$ is also a π-class, then $\mathcal{C}$ is a σ-algebra.

  The proof of (c) is left as an exercise.

- Because of the conditions on (ii) or (b), a λ-class seems less inclusive than a σ-algebra.

Dynkin's theorem

If $\mathcal{D}$ is a π-class, $\mathcal{C}$ a λ-class and $\mathcal{D} \subset \mathcal{C}$, then $\sigma(\mathcal{D}) \subset \mathcal{C}$.
Proof:

- Define $\mathcal{G}$ as the *smallest* λ-class such that $\mathcal{D} \subset \mathcal{G}$; thus $\mathcal{D} \subset \mathcal{G} \subset \mathcal{C}$.
- We now prove $\mathcal{G}$ is also a π-class; the theorem then follows by the previous proposition (c).
- To this end, define $\mathcal{H} := \{A \subset \Omega \mid A \cap D \in \mathcal{G} \ \forall D \in \mathcal{D}\}$:
  - Since $\mathcal{D} \subset \mathcal{G}$ and $\mathcal{D}$ is a π-class, $\mathcal{D} \subset \mathcal{H}$.
  - Check that $\mathcal{H}$ is a λ-class $\Rightarrow$ (minimal) $\mathcal{G} \subset \mathcal{H}$.
  - Thus, $A \in \mathcal{G}$ ($\Rightarrow A \in \mathcal{H}$) and $D \in \mathcal{D} \Rightarrow A \cap D \in \mathcal{G}$.
- Now define $\mathcal{F} := \{B \subset \Omega \mid B \cap A \in \mathcal{G} \ \forall A \in \mathcal{G}\}$:
  - By the previous step, $\mathcal{D} \subset \mathcal{F}$.
  - Check that $\mathcal{F}$ is a λ-class $\Rightarrow$ (minimal) $\mathcal{G} \subset \mathcal{F}$.
  - Thus, $\mathcal{G}$ is a π-class. $\square$

Note: So, a λ-class can be much larger than a π-class.

Classical **monotone class theorem**: if $\mathcal{D}$ is an algebra, $\mathcal{C}$ a monotone class (contains all limits of its monotone sequences) and $\mathcal{D} \subset \mathcal{C}$, then $\sigma(\mathcal{D}) \subset \mathcal{C}$.

Independence in probability space $(\Omega, \mathcal{F}, \mathsf{P})$

- **Lemma:** If $\mathcal{D}, \mathcal{C} \subset \mathcal{F}$ are independent classes of events and $\mathcal{D}$ is a π-class, then $\sigma(\mathcal{D})$ and $\mathcal{C}$ are independent.
  Proof:
  - Take an arbitrary $B \in \mathcal{C}$ and define $\mathcal{D}_B = \{A \in \sigma(\mathcal{D}) \mid \mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)\}$
  - $\mathcal{D} \subset \mathcal{D}_B$.
  - Check $\mathcal{D}_B$ is a λ-class.
  - Apply Dynkin's theorem. $\square$
- **Theorem:** If the joint CDF $F_{X_1,...,X_n} \equiv \prod_{i=1}^n F_{X_i}$ (*i.e.*, the LHS and RHS are equal at all points in $\mathbb{R}^n$), then the n random variables $X_1, ..., X_n$ are independent.
  Proof:
  - Define $\mathcal{D}_k = \{\{X_k \le x\} \mid -\infty \le x \le \infty\}$.
  - Note: $x = -\infty \Rightarrow \emptyset \in \mathcal{D}_k$.
  - Check that $\mathcal{D}_k$ is a π-class $\forall k$.
  - Finally use the lemma to obtain independence of the $\sigma(\mathcal{D}_k) = \sigma(X_k)$. $\square$

Conditional Independence

- Events A and C are said to be independent *given* B if

$$\mathsf{P}(A \mid B, C) \;=\; \mathsf{P}(A \mid B).$$

- Note that this implies $\mathsf{P}(C \mid B, A) = \mathsf{P}(C \mid B)$.

- This is a natural extension of the unqualified notion of independent events, *i.e.*, events A and C are (unconditionally) independent if $\mathsf{P}(A \mid C) = \mathsf{P}(A)$.

- Similarly, random variables X and Y are conditionally independent given Z if

$$\mathsf{P}(X \in A \mid Z \in B,\ Y \in C) \;=\; \mathsf{P}(X \in A \mid Z \in B)$$

  for all Borel $A, B, C \subset \mathbb{R}$, *cf.*, Markov processes.

Expectation

- The *expectation* $\mathsf{E}X$ of a random variable X is simply its average or mean value, which can be expressed as the Riemann-Stieltjes integral:

$$\mathsf{E}X \;=\; \int_{-\infty}^{\infty} x \; \mathrm{d}F_X(x),$$

  recall that the CDF F_X is nondecreasing on $\mathbb{R}$.

- In the special case of a differentiable F_X with probability density function (PDF) $f_X = F_X'$, *i.e.*, X is *continuously* distributed, we can use the Riemann integral

$$\mathsf{E}X \;=\; \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x.$$

- In the case of a discretely distributed random variable X with *countable* state-space R_X,
  - $F_X'(x) = \sum_{\xi \in R_X} p_X(\xi)\delta(x - \xi)$, where
  - δ is the Dirac unit impulse and the probability mass function (PMF) $p_X(\xi) := \mathsf{P}(X = \xi) > 0$ for all $\xi \in R_X$, so that

$$\mathsf{E}X \;=\; \sum_{\xi \in R_X} \xi \; p_X(\xi).$$

Lebesgue integration

- Formally, the Lebesgue integral is used to define expectation:

$$\mathsf{E}X \;=\; \int_\Omega X(\omega)\mathsf{d}\mathsf{P}(\omega) \;=\; \int_\Omega X\ \mathsf{dP}.$$

- Recall that Ω is generally abstract, unordered.

- If X is *simple* (discretely distributed with $|R_X| = M < \infty$),
  - define the state-space $R_X = \{\xi_1, \xi_2, ..., \xi_M\}$ and
  - the events $A_i := \{X = \xi_i\} := \{\omega \in \Omega \mid X(\omega) = \xi_i\}$, which a.s. partition Ω,
  - so that the (well-defined) Lebesgue integral is

$$\int_\Omega X(\omega)\mathsf{d}\mathsf{P}(\omega) \;=\; \sum_{i=1}^{M} \xi_i \mathsf{P}(A_i),$$

  i.e., $\mathsf{P}(A_i) = p_X(\xi_i)$ for all i.

Lebesgue integration (cont)

- To develop the general Lebesgue integral, we need to consider "extended" random variables $X : \Omega \to \overline{\mathbb{R}}$, where
  - the extended reals $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$, and
  - measurability involves Borel sets that include $\pm\infty$.

- For any sequence $X_1, X_2, X_3, ...$ of random variables, the following are extended random variables:
  - $\lim_{n\to\infty} X_n$ (assuming the limit in n of $X_n(\omega)$ exists $\forall\omega \in \Omega$) and
  - $\sup_{n\to\infty} X_n$.

Approximating random variables

- Proposition: For any non-negative extended random variable X, there is a sequence of *simple* random variables X_n such that
  (a) $\mathsf{P}(0 \le X_n \le X_{n+1} \le X) = 1$ for all n (*i.e.*, monotonicity), and
  (b) $X_n \to X$ almost surely (a.s.), *i.e.*, **almost sure convergence:**

$$\mathsf{P}(\lim_{n\to\infty} X_n = X) \;=\; 1.$$

- Proof:
  1. Define the n^{th} partition of $\overline{\mathbb{R}^+}$ (*i.e.*, of the y-axis unlike Riemannian integration) into a finite collection of contiguous intervals
  $$\{[b_n^k, b_n^{k+1})\}_{k=0}^{K_n},$$
  where $\forall n$: $b_n^0 = 0$, $b_n^{K_n+1} = \infty+$ (last interval includes ∞).
  2. $\forall k$, define $X_n(\omega) = b_n^k \;\forall \omega \in X^{-1}[b_n^k, b_n^{k+1})$, *i.e.*, $X_n \le X$ a.s.
  3. The $(n+1)^{\text{st}}$ partition is finer than the n^{th} (*i.e.*, $X_n \le X_{n+1}$), in such a way that $\lim_{n\to\infty} K_n \uparrow \infty$ to achieve (b). □

Construction of the Lebesgue integral

- So for a non-negative extended random variable, the Lebesgue integral is defined as

$$\int_\Omega X \;\mathrm{dP} \;=\; \lim_{n\to\infty} \int_\Omega X_n \;\mathrm{dP},$$

i.e., $\mathsf{E}X = \lim_{n\to\infty} \mathsf{E}X_n$.

- For a signed extended random variable:
  1. Note that $X = X^+ - X^-$ where the *non-negative* extended random variables
  $$X^+ := \max\{0, X\} \quad \text{and} \quad X^- := \max\{0, -X\}.$$
  2. If $\mathsf{E}X^+ < \infty$ or $\mathsf{E}X^- < \infty$, then define the Lebesgue integral
  $$\mathsf{E}X \;=\; \mathsf{E}X^+ - \mathsf{E}X^-,$$
  otherwise the Lebesgue integral $\mathsf{E}X$ is not defined. Note: $|X| = X^+ + X^-$.

- *E.g.*, $\mathbf{1}_{\mathbb{Q}^c}$ is Lebesgue but not Riemann integrable:

$$\int_0^1 \mathbf{1}_{\mathbb{Q}^c}(x)\mathrm{d}x \;=\; 1 \cdot \mathsf{P}(\mathbb{Q}^c \cap [0,1]) + 0 \cdot \mathsf{P}(\mathbb{Q} \cap [0,1]) \;=1,$$

where here P is *Lebesgue measure* on $\Omega = [0,1]$, so that $\mathsf{P}(\mathbb{Q} \cap [0,1]) = 0$ as the rationals are countable.

Integration theorems (1 variable integrand)

Consider a sequence $X_1, X_2, \ldots$ of random variables:

- **Bounded convergence theorem:** if $\sup_n |X_n| \leq K < \infty$ a.s. (where K is constant) and $X = \lim_{n\to\infty} X_n$ a.s., then $\mathsf{E}X = \lim_{n\to\infty} \mathsf{E}X_n$ and $\mathsf{E}|X| \leq K$.

- **Proof:** Define $A_n = \{X - X_n > \varepsilon\}$ for arbitrary positive $\varepsilon \ll 1$ and note

$$\begin{aligned} |\mathsf{E}X_n - \mathsf{E}X| &\leq \mathsf{E}|X_n - X| \\ &= \mathsf{E}|X_n - X|\mathbf{1}_{A_n} + \mathsf{E}|X_n - X|\mathbf{1}_{A_n^c} \\ &\leq 2K\mathsf{P}(A_n) + \varepsilon. \quad \square \end{aligned}$$

- Now note

$$(\liminf_{n\to\infty} X_n)(\omega) \quad := \quad \lim_{n\to\infty} \inf_{k\geq n} X_k(\omega)$$

always exists (though possibly not finite) since $Y_n := \inf_{k\geq n} X_k$ is a.s. monotonically nondecreasing in n.

- **Fatou's lemma:** $\liminf_{n\to\infty} \mathsf{E}X_n \geq \mathsf{E}(\liminf_{n\to\infty} X_n)$.

- **Proof:**
  - Let $X := \liminf_{n\to\infty} X_n$.
  - For any $K > 0$, invoke the bounded convergence theorem on $\min\{Y_n, K\} \uparrow \min\{X, K\}$.
  - Approximating with simple RVs and using monotonicity, $\lim_{K\to\infty} \mathsf{E}\min\{X, K\} = \mathsf{E}X$. $\square$

Integration theorems (cont)

Let X be the extended RV such that $X_n \to X$ a.s.

- **Monotone convergence theorem:** if $\lim_{n\to\infty} X_n \uparrow X$ a.s., then

$$\lim_{n\to\infty} \mathsf{E}X_n \quad \uparrow \quad \mathsf{E}X.$$

- **Lebesgue's dominated convergence theorem:** If there exists a random variable Y such that $|X| \leq |Y|$ a.s. and $\mathsf{E}|Y| < \infty$, then

$$\lim_{n\to\infty} \mathsf{E}|X - X_n| \quad = \quad 0.$$

- **Corollary** (Scheffe):

$$\lim_{n\to\infty} \mathsf{E}|X_n - X| = 0 \quad \Leftrightarrow \quad \lim_{n\to\infty} \mathsf{E}|X_n| = \mathsf{E}|X|$$

Conditional expected value and event-conditional distributions

Consider a random variable X and an event A such that $\mathsf{P}(A) > 0$.

- The *conditional expected value of X given A*, denoted $\mu(X|A)$ is

$$\mu(X|A) = \int_{-\infty}^{\infty} x \mathsf{d}F_{X|A}(x), \quad \text{where} \quad F_{X|A}(z) := \mathsf{P}(X \le z|A).$$

- For a discretely distributed random variable X with $R_X = \{a_j\}_{j=1}^{\infty}$,

$$\mu(X|A) \;=\; \sum_{j=1}^{\infty} a_j \mathsf{P}(X = a_j|A).$$

- The conditional PMF of X given A is $p_{X|A}(a_j) = \mathsf{P}(X = a_j|A)$ for all j.

- Event-conditional PDF of a continuously distributed X is

$$f_{X|A}(x) := \frac{\mathsf{d}}{\mathsf{d}x} F_{X|A}(x) \quad \Rightarrow \quad \mu(X|A) = \int_{-\infty}^{\infty} x f_{X|A}(x) \mathsf{d}x.$$

Conditional expectation

- Consider now two discretely distributed X and Y.

- The *conditional expectation of X given the random variable Y*, denoted $\mathsf{E}(X|Y)$, is a random variable itself.

- Indeed, suppose $\{b_j\}_{j=1}^{\infty} = R_Y$ and, for *all* samples

$$\omega_j \;\in\; \{\omega \in \Omega \mid Y(\omega) = b_j\} \;=:\; B_j$$

  define

$$\mathsf{E}(X|Y)(\omega_j) \;:=\; \mu(X|B_j) := \mu(X|Y = b_j),$$

- That is, $\mathsf{E}(X|Y)$ maps all samples in the event B_j to the conditional expected value $\mu(X|B_j)$, *i.e.*, $\mathsf{E}(X|Y)$ is "smoother" (less uncertain) than X.

- Therefore, the random variable $\mathsf{E}(X|Y)$ is a.s. a *function of* Y, *i.e.*, $\mathsf{E}(X|Y)$ is $\sigma(Y)$-*measurable*.

- So, $\mathsf{E}(X|Y) = \mathsf{E}(X|Z)$ a.s. whenever $\sigma(Z) = \sigma(Y)$ allowing for differences involving P-null events.

Conditional densities

- Now consider two random variables X and Y which are continuously distributed with joint PDF

$$f_{X,Y} \quad = \quad \frac{\partial^2 F_{X,Y}}{\partial x \partial y}.$$

- For $f_Y(y) > 0$, we can define the conditional density:

$$f_{X|Y}(x|y) \quad := \quad \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

for all $x \in \mathbb{R}$.

- Note that $f_{X|Y}(\cdot|y)$ is itself a PDF and,

$$\mu(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\mathrm{d}x, \quad \text{where} \quad \mathsf{P}(Y=y) = 0.$$

Conditional expectation and MSE

- In general, $\mathsf{E}(X|Y)$ is the function of Y which minimizes the *mean-square error* (MSE),

$$\mathsf{E}[(X - h(Y))^2],$$

among all (measurable) functions h.

- So, $\mathsf{E}(X|Y)$ is the best approximation of X given Y.

- In particular, $\mathsf{E}(X|Y)$ and X have the same expectation,

$$\mathsf{E}(\mathsf{E}(X|Y)) = \mathsf{E}X.$$

- Note: if X and Y are independent,

$$\mathsf{E}(X|Y) \quad = \quad \mathsf{E}X \text{ a.s.}$$

Some useful inequalities

- If event $A_1 \subset A_2$, then $\mathsf{P}(A_1) \leq \mathsf{P}(A_2) = \mathsf{P}(A_1) + \mathsf{P}(A_2 \backslash A_1)$.

- For any group of events $A_1, A_2, ..., A_n$, Boole's inequality holds:

$$\mathsf{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathsf{P}(A_i).$$

- Note that when the A_i are disjoint, equality holds simply by the additivity property of a probability measure P and recall the inclusion-exclusion set identities.

- If two random variables X and Y are such that $X \geq Y$ a.s., then $\mathsf{E}X \geq \mathsf{E}Y$.

- Recall Fatou's lemma.

Markov's Inequality

- Consider a random variable X with $\mathsf{E}|X| < \infty$ and a real number $x > 0$.

- Since $|X| \geq |X|\mathbf{1}\{|X| \geq x\} \geq x\mathbf{1}\{|X| \geq x\}$ a.s., we arrive at Markov's inequality:

$$\begin{aligned}
\mathsf{E}|X| &\geq \mathsf{E}x\mathbf{1}\{|X| \geq x\} \\
&= x\mathsf{E}\mathbf{1}\{|X| \geq x\} \\
&= x\mathsf{P}(|X| \geq x).
\end{aligned}$$

- An alternative explanation for continuously distributed random variables X (with PDF f) is

$$\begin{aligned}
\mathsf{E}|X| &= \int_{-\infty}^{\infty} |z| f(z)\mathsf{d}z \\
&\geq \int_{-\infty}^{-x} (-z) f(z)\mathsf{d}z + \int_{x}^{\infty} z f(z)\mathsf{d}z \\
&\geq \int_{-\infty}^{-x} x f(z)\mathsf{d}z + \int_{x}^{\infty} x f(z)\mathsf{d}z \\
&= x\mathsf{P}(|X| \geq x).
\end{aligned}$$

Chebyshev and Cramer's Inequalities

- Take $x = \varepsilon^2$, where $\varepsilon > 0$, and argue Markov's inequality with $(X - \mathsf{E}X)^2$ in place of $|X|$ to get Chebyshev's inequality

$$\mathsf{var}(X) \ := \ E[(X-\mathsf{E}X)^2] \ \geq \ \varepsilon^2 \mathsf{P}(|X-\mathsf{E}X| \geq \varepsilon),$$

  i.e.,

$$\mathsf{P}(|X-\mathsf{E}X| \geq \varepsilon) \ \leq \ \varepsilon^{-2}\mathsf{var}(X).$$

- Noting that, for all $\theta > 0$, $\{X \geq x\} = \{\mathsf{e}^{\theta X} \geq \mathsf{e}^{\theta x}\}$ and arguing as for Markov's inequality gives the Chernoff (or Cramer) inequality:

$$\begin{aligned} \mathsf{E}\mathsf{e}^{\theta X} &\geq \mathsf{e}^{\theta x}\mathsf{P}(X \geq x) \\ \Rightarrow \mathsf{P}(X \geq x) &\leq \exp\left(-[x\theta - \log \mathsf{E}\mathsf{e}^{\theta X}]\right) \\ &\leq \exp\left(-\max_{\theta > 0}[x\theta - \log \mathsf{E}\mathsf{e}^{\theta X}]\right), \end{aligned}$$

  where we have simply sharpened the inequality by taking the maximum over the free parameter $\theta > 0$.

- Note the Legendre transform of the log moment-generating function of X in the Chernoff bound.

Inequalities of Minkowski, Holder, and Cauchy-Schwarz-Bunyakovsky

- Minkowski's inequality: if $\mathsf{E}|X|^q, \mathsf{E}|Y|^q < \infty$ for $q \geq 1$, then

$$(\mathsf{E}|X+Y|^q)^{1/q} \ \leq \ (\mathsf{E}|X|^q)^{1/q} + (\mathsf{E}|Y|^q)^{1/q},$$

  i.e., triangle inequality in the L^q space of random variables.

- Holder's inequality: if $\mathsf{E}|X|^r, \mathsf{E}|Y|^q < \infty$ for $r > 1$ and $q^{-1} := 1 - r^{-1}$, then

$$\mathsf{E}|XY| \ \leq \ (\mathsf{E}|X|^r)^{1/r}(\mathsf{E}|Y|^q)^{1/q}.$$

- CBS inequality ($q = 2$): if $\mathsf{E}X^2, \mathsf{E}Y^2 < \infty$, then

$$\mathsf{E}|XY| \ \leq \ \sqrt{\mathsf{E}(X^2)}\sqrt{\mathsf{E}(Y^2)}.$$

- CBS is strict whenever $X \neq cY$ or $Y = 0$ a.s. for some constant c.

- CBS is an immediate consequence of the fact that whenever $X \neq 0$ a.s. and $Y \neq 0$ a.s.,

$$\mathsf{E}\left(\frac{X}{\sqrt{\mathsf{E}(X^2)}} - \frac{Y}{\sqrt{\mathsf{E}(Y^2)}}\right)^2 \ \geq \ 0.$$

Jensen's Inequality

- Note that if we take $Y = 1$ a.s., the Cauchy-Schwarz-Bunyakovsky inequality simply states that $\text{var}(X) \geq 0$, *i.e.*,
$$\mathsf{E}(X^2) - (\mathsf{E}X)^2 \geq 0.$$
- This is also an immediate consequence of Jensen's inequality.
- A real-valued function g on $\mathbb{R}$ is said to be *convex* if
$$g(px + (1-p)y) \leq pg(x) + (1-p)g(y)$$
for any $x, y \in \mathbb{R}$ and any real fraction $p \in [0,1]$.
- If the inequality is reversed, g is said to be *concave*.
- For any convex function g and random variable X, we have Jensen's inequality:
$$g(\mathsf{E}X) \leq \mathsf{E}(g(X)).$$

Inequalities: Conditioned versions

- These inequalities and integration theorems have straightforward "conditional" extensions.
- *E.g.*, the conditional Jensen's theorem: if g is convex and $\mathcal{G} \subset \mathcal{F}$ is a σ-algebra,
$$\mathsf{E}(g(X) \mid \mathcal{G}) \geq g(\mathsf{E}(X \mid \mathcal{G}))$$
- Applying conditional Jensen's with $g(x) = |x|^q$ for real $q \geq 1$, and using the linearity of conditional expectation, we get the following result.
- If X and $X_1, X_2, X_3, \ldots$ are random variables such that, for real $q \geq 1$, $\mathsf{E}|X|^q < 1$ and $\mathsf{E}|X_n|^q < \infty$ (*i.e.*, $X, X_n \in \mathsf{L}^q$) $\forall n$, then:
  (a) $||\mathsf{E}(X \mid \mathcal{G})||_q \leq ||X||_q := (\mathsf{E}(|X|^q))^{1/q}$, and, therefore,
  (b) If $\lim_{n\to\infty} ||X - X_n||_q = 0$, *i.e.*, **convergence in** L^q, then
$$\lim_{n\to\infty} ||\mathsf{E}(X \mid \mathcal{G}) - \mathsf{E}(X_n \mid \mathcal{G})||_q = 0.$$

Sums of independent random variables

- Consider two independent random variables X_1 and X_2 with PDFs f_1 and f_2 respectively; so, $f_{X_1,X_2} = f_1 f_2$.
- The CDF of the sum is

$$F(z) = \mathsf{P}(X_1 + X_2 \le z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z-x_1} f_1(x_1)f_2(x_2)\mathrm{d}x_2\mathrm{d}x_1.$$

- Exchanging the first integral on the RHS with a derivative w.r.t. z gives the PDF of $X_1 + X_2$:

$$f(z) = \frac{\mathrm{d}}{\mathrm{d}z}F(z) = \int_{-\infty}^{\infty} f_1(x_1)f_2(z-x_1)\mathrm{d}x_1 \quad \text{for all } z \in \mathbb{R}.$$

- Thus, f is the *convolution* of f_1 and f_2 which is denoted $f = f_1 * f_2$.

Sums of independent random variables

- In this context, moment generating functions can be used to simplify calculations.
- Let the MGF (bilateral Laplace transform) of X_i be

$$m_i(\theta) = \mathsf{E}\mathrm{e}^{\theta X_i} = \int_{-\infty}^{\infty} f_i(x)\mathrm{e}^{\theta x}\mathrm{d}x.$$

- The MGF of $X_1 + X_2$ is, by independence,

$$m(\theta) = \mathsf{E}\mathrm{e}^{\theta(X_1+X_2)} \;=\; \mathsf{E}\mathrm{e}^{\theta X_1}\mathrm{e}^{\theta X_2} = m_1(\theta)m_2(\theta).$$

- So, convolution of PDFs corresponds to simple multiplication of MGFs (and to addition of independent random variables).

Example: exponential and gamma distributions

Consider independent random variables that are all exponentially distributed with mean $1/\lambda$.

- The PDF of X_1+X_2 is f, where $f(z)=0$ for $z<0$ and, for $z\geq 0$,
$$f(z)=\int_0^z f_1(x_1)f_2(z-x_1)\mathrm{d}x_1=\lambda^2 z\mathrm{e}^{-\lambda z},$$
i.e., the (n,λ) gamma distribution with $n=2$ (a.k.a. Erlang distribution when $n\in\mathbb{Z}^+$).

- So, the MGF of X_1+X_2 is
$$m(\theta)=\left(\frac{\lambda}{\lambda-\theta}\right)^2,$$
which is consistent with the PDF just computed.

- There is a 1-to-1 relationship between PDFs and MGFs of nonnegative random variables (unilateral Laplace transform).

- So, for a sum of n random variables
$$m(\theta)=\left(\frac{\lambda}{\lambda-\theta}\right)^n \quad\Leftrightarrow\quad f_n(z)=\frac{\lambda^n z^{n-1}\mathrm{e}^{-\lambda z}}{(n-1)!}\ \forall z\geq 0.$$

- Note: Construction of continuous-time Markov chains is based on the memoryless property that is unique to the exponential distribution.

The Gaussian distribution

Assume X_i is Gaussian (normally) distributed with mean μ_i and variance σ_i^2, *i.e.*, $X_i\sim \mathsf{N}(\mu_i,\sigma_i^2)$.

- If independent RVs, the MGF of X_1+X_2 is
$$\begin{aligned} m(\theta) &= \exp\big(\mu_1\theta+\tfrac{1}{2}\sigma_1^2\theta^2\big)\times\exp\big(\mu_2\theta+\tfrac{1}{2}\sigma_2^2\theta^2\big)\\ &= \exp\big((\mu_1+\mu_2)\theta+\tfrac{1}{2}(\sigma_1^2+\sigma_2^2)\theta^2\big),\end{aligned}$$
which we also recognize as a Gaussian MGF.

- Even if dependent, $\alpha_1X_1+\alpha_2X_2$, for scalars α_i, is Gaussian distributed with mean $\alpha_1\mu_1+\alpha_2\mu_2$ and variance $\alpha_1^2\sigma_1^2+\alpha_2^2\sigma_2^2+2\alpha_1\alpha_2\mathsf{cov}(X_1,X_2)$, where the *covariance* $\mathsf{cov}(X_1,X_2):=\mathsf{E}X_1X_2-\mathsf{E}X_1\mathsf{E}X_2$.

- $\underline{X}=(X_1,X_2,...,X_n)$ are jointly Gaussian if
$$f_{\underline{X}}(\underline{x}) = \frac{1}{[2\pi\det(\mathbf{C})]^{n/2}}\exp\left(-\tfrac{1}{2}(\underline{x}-\mathsf{E}\underline{X})^{\mathsf{T}}\mathbf{C}^{-1}(\underline{x}-\mathsf{E}\underline{X})\right),$$
where the (symmetric) *covariance matrix* is $\mathbf{C}=\mathsf{E}(\underline{X}-\mathsf{E}\underline{X})(\underline{X}-\mathsf{E}\underline{X})^{\mathsf{T}}$.

- Note: $\mathsf{E}(X_1\mid X_2)=\mathsf{E}X_1+(X_2-\mathsf{E}X_2)\mathsf{E}(X_1X_2)/\mathsf{E}X_2^2$ is a.s. *linear* in X_2, and

- if X_1,X_2 are *uncorrelated* (diagonal covariance matrix), then Gaussian distributed with mean μ_i and variance σ_i^2. X_1,X_2 are *independent* (the converse is always true).

de Moivre's formula

There is a constant $\beta > 0$ such that $n!\mathrm{e}^n \sim \beta n^n \sqrt{n}$.

Proof:

- Define $B(n) = n!\mathrm{e}^n/(n^n\sqrt{n})$.
- $\log B(n) = 1 + \sum_{j=2}^{n}[\log B(j) - \log B(j-1)]$ where $1 = \log B(1)$.
- By Taylor's theorem,

$$\log(1-x) \;=\; -x - x^2/2 - x^3/3 + o(x^3) \quad \text{where } \lim_{y\to 0}\frac{o(y)}{y} = 0.$$

- So,

$$\begin{aligned}\log B(j) - \log B(j-1) &= 1 + (j - \frac{1}{2})\log(1 - \frac{1}{j}) \\ &= -\frac{1}{12j^2} + o(\frac{1}{j^2})\end{aligned}$$

- Since j^{-2} is summable, $B(n)$ converges to a finite β.

de Moivre-Laplace Central Limit Theorem (CLT)

- Consider a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables $X_1, X_2, ...$, where

$$p := \mathrm{P}(X_i = 1) \quad \text{and} \quad q := 1 - p = \mathrm{P}(X_i = 0).$$

- Define the sum $S_n = X_1 + X_2 + ... + X_n$.
- S_n is binomially distributed with parameters (n, p), *i.e.*,

$$\mathrm{P}(S_n = k) \;=\; \binom{n}{k} p^k q^{n-k}$$

  for all $k \in \{0, 1, 2, ..., n\}$.

- $\mathrm{E}S_n = np$ and variance $\mathrm{var}(S_n) = \mathrm{E}S_n^2 - (\mathrm{E}S_n)^2 = npq$.
- Thus $Y_n := (S_n - np)/\sqrt{nqp}$ is centered ($\mathrm{E}Y_n = 0$) and has unit variance $\mathrm{var}(Y_n) = 1$ for all n.
- **Theorem** (de Moivre-Laplace CLT): If X_i are i.i.d. Bernoulli random variables, then Y_n defined above **converges in distribution** to a standard normal (Gaussian), *i.e.*,

$$\lim_{n\to\infty} \mathrm{P}(Y_n > y) \;=\; \Phi(y) := \int_y^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \mathrm{d}x$$

de Moivre-Laplace CLT: Proof

$$\begin{aligned}
&\mathsf{P}(a < Y_n := \frac{S_n - np}{\sqrt{npq}} \leq b) \\
&= \sum_{np+a\sqrt{nqp}<k\leq np+b\sqrt{nqp}} \binom{n}{k} p^k q^{n-k} \\
&= \sum_{a\sqrt{nqp}<k'\leq b\sqrt{nqp}} \binom{n}{k'+np} p^{k'+np} q^{nq-k'}
\end{aligned}$$

where the sums are over integers k, k'. Using de Moivre's formula to uniformly approximate $\binom{n}{k'+np}$ over k' as $n \to \infty$:

$$\begin{aligned}
&\mathsf{P}(a < Y_n \leq b) \\
&\sim \quad \frac{1}{\beta\sqrt{npq}} \sum_{a\sqrt{nqp}<k'\leq b\sqrt{nqp}} (1+\frac{k'}{np})^{-k'-np}(1-\frac{k'}{nq})^{-nq+k'} \\
&\sim \quad \frac{1}{\beta\sqrt{npq}} \sum_{a\sqrt{nqp}<k'\leq b\sqrt{nqp}} \exp\left(-\frac{(k')^2}{2npq}\right) \\
&\to_{n\to\infty} \quad \int_a^b \frac{\mathrm{e}^{-x^2/2}}{\beta} \mathrm{d}x \\
&= \quad \frac{\sqrt{2\pi}}{\beta}(\Phi(a) - \Phi(b))
\end{aligned}$$

where the second step is $\log(1-x) = -x - x^2/2 + o(x^2)$ and the third is the Riemann integral.

Taking $-a, b \to \infty$, gives Wallis' identity: $\beta = \sqrt{2\pi}$. $\square$

Stirling's formula

- de Moivre's formula with Wallis' identity gives Stirling's formula:

$$n!\mathrm{e}^n \quad \sim \quad n^n\sqrt{2\pi n}.$$

- In the following, we prove a more general sequential CLT.

An i.i.d. CLT

Theorem: If $X_1, X_2, ...$ are i.i.d. with $\mathsf{E}|X_1| < \infty$ and $0 < \sigma^2 := \mathsf{var}(X_1) < \infty$, then

$$Y_n := \frac{S_n - n\mu}{\sigma\sqrt{n}} \quad \to_d \quad N(0,1),$$

i.e., converges in distr'n to a standard normal.

Proof:

- Taylor's thm: $\mathsf{e}^{ix} = 1 + ix - \frac{1}{2}x^2 + R(x)$ where $|R(x)| \leq |x|^3$.

- For $|x| > 4$, $|R(x)| \leq |\mathsf{e}^{ix}| + 1 + |x| + \frac{1}{2}x^2 \leq x^2$, $\Rightarrow |R(x)| \leq \min\{|x|^3, x^2\}$.

- So, if $X \sim (X_1 - \mathsf{E}X_1)/\sigma$ then the **characteristic function** of Y_n is $\mathsf{E}\mathsf{e}^{itY_n} = \left(\mathsf{E}\mathsf{e}^{itX/\sqrt{n}}\right)^n$

$$\begin{aligned}\Rightarrow \mathsf{E}\mathsf{e}^{itY_n} &= \left(1 + i(\mathsf{E}\frac{tX}{\sqrt{n}}) - \frac{\mathsf{E}(tX)^2}{2n} + \mathsf{E}R(\frac{tX}{\sqrt{n}})\right)^n \\ &= \left(1 - \frac{t^2}{2n} + \mathsf{E}R(\frac{tX}{\sqrt{n}})\right)^n,\end{aligned}$$

where, by the dominated convergence theorem,

$$|\mathsf{E}R(tX/\sqrt{n})| \leq n^{-1}\mathsf{E}\min\{\frac{|tX|^3}{\sqrt{n}}, (tX)^2\} = n^{-1}o(1).$$

$$\Rightarrow \lim_{n\to\infty} \mathsf{E}\mathsf{e}^{itY_n} = \lim_{n\to\infty}(1 - t^2/(2n) + o(1)/n)^n = \mathsf{e}^{-t^2/2}. \qquad \square$$

Trotter's proof of the i.i.d. CLT Preliminaries

- Let C be the set of bounded uniformly continuous functions on $\mathbb{R}$, *i.e.*, if $f \in C$ then $\forall \varepsilon > 0$, $\exists \delta > 0$ such that: $\forall x, y \in \mathbb{R}$, $|x - y| < \varepsilon \;\Rightarrow\; |f(x) - f(y)| < \delta$.

- A transformation (function, operator) $T : C \to C$ is said to be *linear* if $T(af + bg) = aTf + bTg$ $\forall f, g \in C$ and $\forall a, b \in \mathbb{R}$. Note that $\forall x \in \mathbb{R}$, $(aTf + bTg)(x) := a(Tf)(x) + b(Tg)(x)$.

- Define the *supremum norm* $\| f \| := \sup_{x\in\mathbb{R}} |f(x)|$.

- T is said to be a *contraction* operator if $\| Tf \| \leq \| f \|$ $\forall f \in C$.

- For a random variable X, define $T_X : C \to C$ as

$$(T_X f)(y) \;:=\; \mathsf{E}f(X + y) = \int_{-\infty}^{\infty} f(x + y)\mathsf{d}F_X(x) \quad y \in \mathbb{R}.$$

- Note: $f \in C \;\Rightarrow\; T_X f \in C$, T_X is a linear contraction, and $(T_X f)(0) = \mathsf{E}f(X)$.

- Note: $T_{X_1}T_{X_2} = T_{X_2}T_{X_1}$ (commutation), furthermore if X_1, X_2 are independent then (as characteristic functions)

$$T_{X_1+X_2} = T_{X_1}T_{X_2} = T_{X_2}T_{X_1},$$

c.f., Fubini's theorem.

- Define $C^2 = \{f \in C \mid f', f'' \in C\}$.

Trotter's proof of the i.i.d. CLT
Preliminaries (cont)

Lemma 1: If $\lim_{n\to\infty} \mathsf{E}f(X_n) = \mathsf{E}f(X)$ $\forall f \in C^2$, then $X_1, X_2, ...$ converges in distribution to X.
Note: Hypothesis is satisfied if $\| T_{X_n}f - T_X f \| \to 0$.
Proof of Lemma 1:

- Consider any y at which F_X is *continuous*.
- Fix $\varepsilon > 0$ arbitrarily and take $\delta > 0$ small enough so that $F_X(y+\delta) - F_X(y-\delta) < \varepsilon$.
- Define $f, g \in C^2$ such that
  (i) $f(x) = 1$ for $x \le y - \delta$,
  (ii) $g(x) = 1$ for $x \le y$,
  (iii) $f(x) = 0$ for $x \ge y$, and
  (iv) $g(x) = 0$ for $x \ge y + \delta$;
  so that $0 \le f \le g \le 1$ in particular.
- So, since $f(X) \le \mathbf{1}\{X \le y - \delta\}$ *etc.*,
$$F_X(y-\delta) \le \mathsf{E}f(X) = \lim_{n\to\infty} \mathsf{E}f(X_n) \le \liminf_{n\to\infty} F_{X_n}(y)$$
$$\le \limsup_{n\to\infty} F_{X_n}(y) \le \lim_{n\to\infty} \mathsf{E}g(X_n) = \mathsf{E}g(X) \le F_X(y+\delta)$$
  where the equalities are by hypothesis.
- Since this holds $\forall \varepsilon > 0$, $\lim_{n\to\infty} F_{X_n}(y) = F_X(y)$. □

Trotter's proof of the i.i.d. CLT
Preliminaries (cont)

Lemma 2: If $A, B : C \to C$ are linear, contraction operators that commute, then $\| A^n f - B^n f \| \le n \| Af - Bf \|$ $\forall n \in \mathbb{Z}^+, f \in C$.
Proof:

- Factor
$$A^n f - B^n f = \sum_{i=0}^{n-1} A^{n-i-1}(A-B)B^i f = \sum_{i=0}^{n-1} A^{n-i-1}B^i(A-B)f.$$
  where the second equality is by commutativity.
- Now take norm of both sides, use the triangle inequality, and finally repeatedly use the contraction hypotheses. □

Trotter's proof of the i.i.d. CLT
Preliminaries (cont)

Lemma 3: If $\mathsf{E}X = 0$ and $\mathsf{E}X^2 = 1$, then $\forall f \in C^2$ and $\forall \varepsilon > 0$: $\exists N < \infty$ such that $\| T_{n^{-1/2}X} f - f - \frac{1}{2n} f'' \| \le \frac{\varepsilon}{n}$ $\forall n \ge N$.
Proof:

- Fix y such that F_X is continuous at y.

- By Taylor's theorem, $\exists z(x) \in [y, y+x]$ such that

$$f(y+x) \;=\; f(y) + xf'(y) + \tfrac{1}{2}x^2 f''(y) + \tfrac{1}{2}x^2[f''(z(x)) - f''(y)].$$

- By uniform continuity of f'' ($f \in C^2$), $\forall \varepsilon > 0$, $\exists \delta > 0$ such that $|z(x) - y| < \delta \;\Rightarrow\; |f''(z(x)) - f''(y)| < \varepsilon$. Thus,

$$\begin{aligned}
&(T_{n^{-1/2}X} f)(y)\\
&= \int f(y + n^{-1/2}x)\mathrm{d}F_X(x)\\
&= f(y)\int \mathrm{d}F_X(x) + \tfrac{1}{\sqrt{n}} f'(y)\int x\mathrm{d}F_X(x) + \tfrac{1}{2n} f''(y)\int x^2\mathrm{d}F_X(x)\\
&\quad + \frac{1}{2n}\int [f''(z(n^{-1/2}x)) - f''(y)]x^2\mathrm{d}F_X(x)\\
&= f(y) + \tfrac{1}{2n} f''(y)\\
&\quad + \frac{1}{2n}\left(\int_{|x|<\delta\sqrt{n}} + \int_{|x|\ge\delta\sqrt{n}}\right)[f''(z(n^{-1/2}x)) - f''(y)]x^2\mathrm{d}F_X(x)
\end{aligned}$$

Trotter's proof of the i.i.d. CLT
Lemma 3's proof (cont)

- Now, $|x| < \delta\sqrt{n} \;\Rightarrow\; |z(n^{-1/2}x) - y| \le |n^{-1/2}x| \le \delta$.

- Thus,

$$\begin{aligned}
&\left|\frac{1}{2n}\int_{|x|<\delta\sqrt{n}} [f''(z(n^{-1/2}x)) - f''(y)]x^2\mathrm{d}F_X(x)\right|\\
&\le \left|\frac{1}{2n}\int_{|x|<\delta\sqrt{n}} \varepsilon x^2\mathrm{d}F_X(x)\right|\\
&\le \frac{\varepsilon}{n}.
\end{aligned}$$

- Since $|f''(z) - f''(x)| \le 2 \| f'' \| < \infty$ and $\mathsf{E}X^2 < \infty$,

$$\begin{aligned}
&\frac{1}{2n}\left|\int_{|x|\ge\delta\sqrt{n}} [f''(z(n^{-1/2}x)) - f''(y)]x^2\mathrm{d}F_X(x)\right|\\
&\le \frac{1}{n} \| f'' \| \left|\int_{|x|\ge\delta\sqrt{n}} x^2\mathrm{d}F_X(x)\right|\\
&\le \frac{\varepsilon}{n} \;\forall \text{ suff. large } n.
\end{aligned}$$

- Finally, substitute the last two estimates into the expression for $(T_{n^{-1/2}X} f)(y)$ of the previous slide. $\square$

Trotter's proof of the i.i.d. CLT

Theorem: If $X_1, X_2, ...$ are i.i.d. with $\mathsf{E}|X_1| < \infty$ and $0 < \mathsf{E}X_1^2 < \infty$, then $n^{-1/2}S_n := n^{-1/2}(X_1 + ... + X_n) \to_d \mathsf{N}(\mu, \sigma^2)$ where $\mu := \mathsf{E}X_1$ and $\sigma^2 := \mathsf{var}(X_1)$.
Proof:

- Let $Y \sim \mathsf{N}(\mu, \sigma^2)$.
- By Lemma 1, theorem follows if $\lim_{n\to\infty} \| T_{n^{-1/2}S_n} f - T_Y f \| = 0$.
- w.l.o.g., $\mu = 0$ and $\sigma^2 = 1$, *i.e.*, $Y \sim$N(0,1) if we restate the theorem in terms of $(S_n - n\mu)/(\sigma\sqrt{n})$.
- Since $Y \sim$N(0,1), $T_Y = T^n_{n^{-1/2}Y}$ and, by IBP, $T^n_{n^{-1/2}Y} f = f + \frac{1}{2n} f''$.
- Since the X_i are i.i.d., $T_{n^{-1/2}S_n} = T^n_{n^{-1/2}X_1}$.
- By Lemma 2,
$$\| T_{n^{-1/2}S_n} f - T_Y f \| \quad \leq \quad n \, \| T_{n^{-1/2}X_1} f - T_{n^{-1/2}Y} f \| \, .$$
- Applying Lemma 3 we get
$$\| T_{n^{-1/2}X_1} f - T_{n^{-1/2}Y} f \| \quad \leq \quad 2\varepsilon.$$
for all sufficiently large n. □

Lindeberg's CLT for independent random variables

- Consider an independent sequence of random variables with $\mathsf{E}X_i = 0 \ \forall i$ w.l.o.g.
- Let $\sigma_i^2 = \mathsf{E}X_i^2$, *i.e.*, they are not necessarily identically distributed.
- The CLT can be generalized to a sequence of random variables that assuming only their mutual independence under **Lindeberg's condition:**
$$\lim_{n\to\infty} s_n^{-2} \sum_{i=1}^{n} \int_{|x|\geq \delta s_n} x^2 \mathrm{d}F_{X_i}(x) \quad = \quad 0 \quad \forall \delta > 0,$$
where $s_n := \sqrt{\sum_{i=1}^n \sigma_i^2}$, *i.e.*, the X_i are not identically distributed (recall the last inequality of the proof of Lemma 3).
- A simple proof of Lindeberg's CLT follows that of Trotter's for the i.i.d. case [Trotter'59].
- Feller proved that Lindeberg's condition is necessary.

Modes of convergence

- A CLT involves convergence *in distribution*.
- This is the weakest class of convergence results, in which no limiting random variable need exist.
- $F_{Y_n}(y) \to F_Y(y)$ $\forall y$ that are points of continuity of F_Y implies $Y_n \to Y$ in distr'n.
- In the following, we will see that convergence:
  in distr'n $\Leftarrow$ in prob. $\Leftarrow$ (a.s. or in L^2).

Weak law of large numbers (WLLN): assumptions

- Assume random variables $X_1, X_2, X_3, ...$ are i.i.d.
- Also suppose that the common distribution has finite variance, *i.e.*, $\sigma^2 := \text{var}(X) := \mathsf{E}(X - \mathsf{E}X)^2 < \infty$, where $X \sim X_i$ $\forall i$.
- Finally, suppose that the mean exists and is finite, *i.e.*, $\mu := \mathsf{E}X < \infty$.
- Recall the sum $S_n := X_1 + X_2 + \cdots + X_n$ for $n \geq 1$, $\mathsf{E}S_n = n\mu$ and $\text{var}(S_n) = n\sigma^2$.
- The quantity S_n/n is called the *empirical* mean of X after n samples and is an *unbiased* estimate of μ, *i.e.*,

$$\mathsf{E}\left(\frac{S_n}{n}\right) = \mu.$$

A WLLN: Statement and Proof

Theorem: If $X_1, X_2, \ldots$ i.i.d. with var$(X_1) < \infty$,

$$\lim_{n\to\infty} \mathsf{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \;=\; 0 \quad \forall \varepsilon > 0.$$

- By Chebyshev's inequality,

$$\mathsf{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \;\leq\; \frac{\mathrm{var}(S_n/n)}{\varepsilon^2} \;=\; \frac{\sigma^2}{n\varepsilon^2}.$$

$\square$

- Note: So, S_n/n is said to be a **weakly consistent** estimator of μ.

- Consequently, L^2 convergence, $\mathsf{E}(Y_n - Y)^2 \to 0$, implies weak convergence, $\mathsf{P}(|Y_n - Y| > \varepsilon) \to 0\ \forall \varepsilon > 0$.

- Example: Discretely distributed $Y_n \to 0$ in probability but *not* in L^2 when:

  $$\mathsf{P}(Y_n = -n) = \mathsf{P}(Y_n = n) = p_n = (1 - \mathsf{P}(Y_n = 0))/2$$

  such that $p_n \to 0$ as $n \to \infty$ but $n^2 p_n \not\to 0$, *e.g.*, $p_n = 1/n$ for $n > 1$ so that $\mathsf{E}Y_n^2 = 2n \not\to 0$.

- Example: $Y_n \to c$ (a constant) in probability $\Leftrightarrow$ $Y_n \to c$ in distribution.

Strong Law of Large Numbers (SLLN)

- Again, a sequence of random variables $X_1, X_2, \ldots$ is said to **converge almost surely (a.s.)** to a random variable X if

$$\mathsf{P}\left(\lim_{n\to\infty} X_n \neq X\right) \;=\; 0.$$

- Kolmogorov's *strong* LLN: if $X_1, X_2, \ldots$ i.i.d. and $\mathsf{E}|X_1| < \infty$, then

$$\mathsf{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \;=\; \mu\right) = 1 \quad \textit{i.e.,} \quad \frac{S_n}{n} \to \mu := \mathsf{E}X_1 \text{ a.s.}$$

- Formally, the limit states that $\forall r \in \mathbb{Z}^+$, $\exists n$ such that $\forall k \geq n$: $|S_n/n - \mu| < 1/r$ a.s.; *i.e.*,

$$\mathsf{P}\left(\bigcap_{r=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left\{\left|\frac{S_k}{k} - \mu\right| < \frac{1}{r}\right\}\right) \;=\; 1.$$

- But $\mathsf{P}(\bigcap_{r=1}^{\infty} B_r) = 1 \Leftrightarrow \mathsf{P}(B_r) = 1\ \forall r$, so $S_n/n \to \mu$ a.s. if and only if

$$\mathsf{P}\left(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left\{\left|\frac{S_k}{k} - \mu\right| < \frac{1}{r}\right\}\right) \;=\; 1\ \forall r.$$

SLLN statement (cont)

- Now fix $r \in \mathbb{Z}^+$ arbitrarily and let
$$A_k^c = \left\{|S_k/k - \mu| < r^{-1}\right\}.$$
- The event
$$\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c$$
is denoted A_n^c *almost always* (a.a.), *i.e.*,
$$\omega \in A_n^c \text{ a.a.} \iff \exists n^*(\omega) \text{ such that } \omega \in A_n^c \ \forall n \geq n^*(\omega).$$
- Note: $\mathsf{P}(\bigcap_{k=n}^{\infty} A_k^c) \uparrow \mathsf{P}(A_n^c \ a.a.)$.
- Note: equivalently express above in terms of A_n where the event
$$\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$
is denoted A_n *infinitely often* (i.o.) $= (A_n^c \text{ a.a.})^c$, *i.e.*,
$$\omega \in A_n \text{ i.o.} \iff \forall n \ \exists m > n \text{ such that } \omega \in A_m^c.$$

SLLN and Borel-Cantelli Lemmas

So, $S_n/n \to \mu$ a.s. $\Leftrightarrow \mathsf{P}(A_n^c \text{ a.a.}) = 1 \Leftrightarrow \mathsf{P}(A_n \text{ i.o.}) = 0$ for all $r \in \mathbb{Z}^+$ where $A_n = \{|S_n/n - \mu| \geq r^{-1}\}$.

- **First BC Lemma:** Generally for events $A_1, A_2, \ldots,$
$$\sum_{n=1}^{\infty} \mathsf{P}(A_n) < \infty \ \Rightarrow \ \mathsf{P}(A_n \text{ i.o.}) = 0.$$
- **Proof:**
$$\begin{aligned} \mathsf{P}(A_n \text{ i.o.}) &\leq \mathsf{P}(\bigcup_{n=k}^{\infty} A_n) \ \ \forall k \\ &\leq \mathsf{P}(A_k) + \mathsf{P}(A_{k+1}) + \cdots \to_k 0 \quad \square \end{aligned}$$
- **Second BC Lemma:** For *independent* events $A_1, A_2, \ldots,$
$$\sum_{n=1}^{\infty} \mathsf{P}(A_n) = \infty \ \Rightarrow \ \mathsf{P}(A_n \text{ i.o.}) = 1.$$
- **Proof:** $\lim_{m\to\infty} \sum_{k=n}^{m} \mathsf{P}(A_k) = \infty$ implies
$$0 \leftarrow \mathrm{e}^{-\sum_{k=n}^{m} \mathsf{P}(A_k)} = \prod_{k=n}^{m} \mathrm{e}^{-\mathsf{P}(A_k)} \geq \prod_{k=n}^{m} (1 - \mathsf{P}(A_k)) = \mathsf{P}(\bigcap_{k=n}^{m} A_k^c)$$
where the equalities are by independence and $\mathrm{e}^{-x} \geq 1 - x$ was used. $\square$
- Note: $\mathsf{P}(A_n \text{ i.o.}) \in \{0, 1\}$ by the "zero-one law".

Kolmogorov's maximal inequality

If $X_1, X_2, ...$ are independent and $\mathsf{E}X_n^2 < \infty\ \forall n$,

$$\mathsf{P}(\max_{1\leq k\leq n} |S_k - \mathsf{E}S_k| \geq \lambda) \quad \leq \quad \frac{\mathsf{var}(S_k)}{\lambda^2} \quad \forall n \geq 1, \lambda > 0.$$

Proof:

- W.l.o.g. assume $\mathsf{E}X_n = 0\ \forall n\ \Rightarrow \mathsf{E}S_n = 0\ \forall n$. Fix $\lambda > 0$.
- Define *disjoint* $B_k = \{|S_i| < \lambda\ \forall i < k,\ |S_k| \geq \lambda\}$.

$$\begin{aligned}
\mathsf{E}S_n^2 &\geq \sum_{k=1}^{n} \mathsf{E}S_n^2 \mathbf{1}_{B_k} \\
&\geq \sum_{k=1}^{n} \mathsf{E}(2(S_n - S_k)S_k + S_k^2)\mathbf{1}_{B_k} \\
&= \sum_{k=1}^{n} [2\mathsf{E}(S_n - S_k)\mathsf{E}S_k\mathbf{1}_{B_k} + \mathsf{E}S_k^2\mathbf{1}_{B_k}] \\
&= \sum_{k=1}^{n} \mathsf{E}S_k^2\mathbf{1}_{B_k}
\end{aligned}$$

  where the second-to-last inequality is by independence.

- Thus, $\mathsf{E}S_n^2 \geq \lambda^2 \sum_{k=1}^{n} \mathsf{P}(B_k) = \lambda^2 \mathsf{P}(\bigcup_{k=1}^{n} B_k)$. □
- Note how this generalizes Chebyshev's inequality.

SLLN: proof of bounded second moment case

Assume $\mathsf{E}X_1^2 < \infty$.

- Again, assume centered X_n w.l.o.g., and apply the maximal inequality with $\lambda = 2^n\varepsilon$ and First BC Lemma to get that

$$\mathsf{P}(\{\max_{1\leq k\leq 2^n} |S_k| \leq 2^n\varepsilon\} \text{ a.a.}) \quad = \quad 1\ \forall \varepsilon > 0.$$

- Now, $\forall m$ such that $2^{n-1} < m \leq 2^n$,

$$\max_{1\leq k\leq 2^n} |S_k| \leq 2^n\varepsilon \quad \text{implies} \quad |S_m| \leq 2m\varepsilon$$

- This leads to $\mathsf{P}(|S_m|/m \leq 2\varepsilon \text{ a.a.}) = 1\ \forall \varepsilon > 0$. □
- Note: Kolmogorov's SLLN only requires $\mathsf{E}|X_1| < \infty$, *i.e.*, bounded first moment.

Weak and strong LLNs

- SLLN $\Rightarrow$ WLLN since P(A_n^c i.o.) $= 0 \Rightarrow$ P$(A_n) \to 0$.

- Example of persistently shrinking pulse on $\Omega = [0,1]$ with P Lebesgue measure:
  - $\forall m \in \mathbb{Z}^+, k \in \{1, 2, \ldots, m\}$, define

$$Y_{k+m(m-1)/2}(\omega) \quad := \quad \mathbf{1}\left\{\frac{k-m}{m} < \omega \le \frac{k}{m}\right\}.$$

  - The random variables Y_n converge weakly but not strongly to zero because P($\{Y_m > \varepsilon\}$ i.o.) $= 1$ for all $\varepsilon = r^{-1} > 0$.

- **Theorem:** If $X_1, X_2, \ldots$ converges to X in probability then there is a *subsequence* $X_{n_1}, X_{n_2}, \ldots$ that converges to X a.s.

Uniform integrability: motivation

- Persistently shrinking pulse example *also* showed that convergence in L^q, for $q \ge 1$, does not generally imply convergence a.s.

- Example (Dirac/Heaviside impulse):
  - $\Omega = [0,1]$ and P is Lebesgue measure.
  - $X_n(\omega) := n\mathbf{1}_{[0,\frac{1}{n}]}$ $\forall n \ge 1$.
  - Clearly, $X_n \to 0$ a.s.
  - But, E$X_n = 1$ $\forall n \ge 1$.

- So, convergence a.s. (*i.e.*, "pointwise") does not generally imply convergence in L^q either.

- Under what conditions does a.s. convergence imply convergence in L^q for $q \ge 1$?

Uniform integrability: preliminaries

Consider the probability space $(\Omega, \mathcal{F}, \mathsf{P})$.

- **Theorem:** If $\mathsf{E}|X| < \infty$ then $\forall \varepsilon \in (0, \infty)\ \exists c(\varepsilon) \in [0, \infty)$ such that

$$\mathsf{E}(|X|\mathbf{1}_{\{|X| \geq c\}}) \ < \ \varepsilon \quad \forall c \in [c(\varepsilon), \infty).$$

  Proof:

  - Lebesgue integrals are uniformly continuous, *i.e.*, $\exists \delta(\varepsilon) \in (0, \infty)$ such that $\mathsf{E}X\mathbf{1}_A < \varepsilon\ \forall A \in \mathcal{F}$ such that $\mathsf{P}(A) < \delta(\varepsilon)$
    (exercise: prove by contradiction of the assumption that $\mathsf{E}|X| < \infty$).
  - By Markov's inequality $\mathsf{P}(|X| \geq c) \leq c^{-1}\mathsf{E}|X|\ \forall c \in (0, \infty)$.
  - Since $\mathsf{E}|X| < \infty$, $\exists c(\varepsilon) \in (0, \infty)$ such that $\mathsf{P}(|X| \geq c) < \delta(\varepsilon)\ \forall c \in [c(\varepsilon), \infty)$.
  - Finally, take $A = \{X \geq c\}$ for $c \in (c(\varepsilon), \infty)$. □

Uniform integrability: definition and sufficient conditions

- A collection of random variables $\mathcal{C}$ on $(\Omega, \mathcal{F}, \mathsf{P})$ is **uniformly integrable** if: $\forall \varepsilon \in (0, \infty)\ \exists c(\varepsilon) \in [0, \infty)$ such that

$$\sup_{X \in \mathcal{C}} \mathsf{E}(X\mathbf{1}_{\{|X| \geq c\}}) \ < \ \varepsilon \quad \forall c \in [c(\varepsilon), \infty) \text{ and } \forall X \in \mathcal{C}.$$

- If $|X| \leq Y$ a.s. $\forall X \in \mathcal{C}$ with $\mathsf{E}Y < \infty$, then $\mathcal{C}$ is uniformly integrable (note that $\mathsf{E}X_n\mathbf{1}\{|X_n| > Y\} = 0$ and recall Lebesgue's dominated convergence theorem); the converse is not true.

- For uniformly integrable $\mathcal{C}$, if $c \in [c(\varepsilon), \infty)$ then

$$\mathsf{E}|X| \ = \ \mathsf{E}|X|\mathbf{1}_{\{|X|<c\}} + \mathsf{E}|X|\mathbf{1}_{\{|X| \geq c\}} \ < \ c + \varepsilon \quad \forall X \in \mathcal{C},$$

  i.e., $\mathcal{C}$ is *uniformly* L^1 *bounded*; the converse is not true.

- If $\mathcal{C} = \{X_0, X_1, ...\}$ such that $0 \leq X_n \leq X_{n+1}\ \forall n \in \mathbb{Z}^+$ (*i.e.*, an *increasing* sequence) and $\mathcal{C}$ is L^1 bounded, then by monotone convergence theorem, $\mathcal{C}$ is uniformly integrable.

- **Theorem**: If $X \in (\Omega, \mathcal{F}, \mathsf{P})$ and $\mathsf{E}|X| < \infty$ and $\{\mathcal{G}_\lambda,\ \lambda \in \Lambda\}$ is a collection of sub σ-algebras of $\mathcal{F}$, then $\{\mathsf{E}(X \mid \mathcal{G}_\lambda),\ \lambda \in \Lambda\}$ is uniformly integrable.
  Proof: exercise.

Uniform integrability: main theorem prelim

- Define the ramp $\theta_c(x) = x\mathbf{1}_{\{|x|<c\}} + c\mathbf{1}_{\{x\geq c\}} - c\mathbf{1}_{\{x\leq -c\}}$.

- **Theorem:** If $\mathcal{C}$ is uniformly integrable then $\forall \varepsilon \in (0,\infty)$ $\exists c(\varepsilon) \in [0,\infty)$ such that

$$\mathsf{E}|X - \theta_c(X)| \quad < \quad \varepsilon \quad \forall X \in \mathcal{C},\ c \in [c(\varepsilon),\infty).$$

- **Proof:** Arbitrarily fix $c \in (0,\infty)$ and note that

$$\begin{aligned} |x - \theta_c(x)| &= (x-c)^+ + (x+c)^- \quad \forall x \in \mathbb{R} \\ \Rightarrow\ \mathsf{E}(X-c)^+ &= \mathsf{E}(X-c)\mathbf{1}_{\{X\geq c\}} \ \leq \mathsf{E}|X|\mathbf{1}_{\{|X|\geq c\}} \\ \text{and } \mathsf{E}(X+c)^- &= -\mathsf{E}(X+c)\mathbf{1}_{\{X\leq -c\}} \ \leq \mathsf{E}|X|\mathbf{1}_{\{|X|\geq c\}}. \quad \square \end{aligned}$$

Swapping limits & expectation (integration)

Theorem: If (a) $\lim_{n\to\infty} X_n = X$ a.s. and (b) $\{X_0, X_1, ...\}$ are uniformly integrable, then

$$\mathsf{E}|X| < \infty \quad \text{and} \quad \lim_{n\to\infty} \mathsf{E}|X_n - X| = 0.$$

Proof:

- By (b) and previous "uniformly L^1 boundedness" result, $\exists B < \infty$ such that $\mathsf{E}|X_n| < B$ $\forall n \in \mathbb{Z}^+$.

- So, by (a) and Fatou's lemma,

$$\mathsf{E}|X| \quad = \quad \mathsf{E}\liminf_{n\to\infty} |X_n| \quad \leq \quad \liminf_{n\to\infty} \mathsf{E}|X_n| \quad < \quad B.$$

- To get L^1 convergence of X_n to X, arbitrarily fix $\varepsilon \in (0,\infty)$. By the previous "ramp" result and (b), $\exists c(\varepsilon) \in (0,\infty)$ such that, $\forall c \in (c(\varepsilon),\infty),\ n \in \mathbb{Z}^+$,

$$\mathsf{E}|X_n - \theta_c(X_n)| < \varepsilon/3 \quad \text{and} \quad \mathsf{E}|X - \theta_c(X)| < \varepsilon/3.$$

- Fix $c \in (c(\varepsilon),\infty)$. Since θ_c is continuous and bounded in magnitude by c, by the dominated convergence theorem, $\exists N(\varepsilon) \in \mathbb{Z}^+$ such that

$$\mathsf{E}|\theta_c(X_n) - \theta_c(X)| \quad < \quad \varepsilon/3 \quad \forall n \geq N(\varepsilon).$$

- By these three "$\varepsilon/3$" inequalities and the triangle inequality,

$$\begin{aligned} \mathsf{E}|X - X_n| &\leq \mathsf{E}|X_n - \theta_c(X_n)| + \mathsf{E}|X - \theta_c(X)| + \mathsf{E}|\theta_c(X_n) - \theta_c(X)| \\ &< \varepsilon. \quad \square \end{aligned}$$

Completeness of L^q

- **Definition**: $\mathsf{L}^q(\Omega, \mathcal{F}, \mathsf{P})$, or just L^q, is the set of *random variables* X on $(\Omega, \mathcal{F}, \mathsf{P})$ such that $||X||_q := (\mathsf{E}(|X|^q))^{1/q} < \infty$.

- **Definition**: $X_1, X_2, ...$ is said to be a **Cauchy sequence** in L^q if $\forall \varepsilon > 0 \ \exists N_\varepsilon$ such that $||X_n - X_m||_q < \varepsilon$ whenever $n, m > N_\varepsilon$.

- **Definition**: A set is said to be **complete** if all of its Cauchy sequences converge to an element inside it.

- For $q \geq 1$, $||\cdot||_q$ is a **norm** by Minkowski's inequality, $||X + Y||_q \leq ||X||_q + ||Y||_q$, so that L^q is a (complete) **Banach space**.

- To show that L^q is complete:
  - Consider a Cauchy sequence $X_1, X_2, ...$ in L^q.
  - Let $N^*(\varepsilon) := \inf\{N_\xi \mid \xi \leq \varepsilon\}$ and define $X^* = X_{N^*(\varepsilon)}$; so, by Minkowski's inequality, $\forall n \geq N^*$,
  $$||X_n||_q \leq ||X^*||_q + ||X_n - X^*||_q \leq ||X^*||_q + \varepsilon,$$
  i.e., the sequence $\{X_n, \ n \geq N^*\}$ is bounded in L^q.
  - Then, one can show that a subsequence X_{n_k} a.s. converges to a (measurable) random variable X (use the completeness of $\mathbb{R}$ and argue by contradiction).
  - So by Fatou's lemma, $||X||_q^q < \infty$, *i.e.*, $X \in \mathsf{L}^q$.
  - Finally, use Fatou's lemma on $||X - X_n||_q^q$ to establish L^q-convergence to X.

Caratheodory's extension theorem

- **Theorem:** If $\mathcal{A}$ is an algebra on Ω and P is countably additive on $\mathcal{A}$, then there exists $\bar{\mathsf{P}}$ on $\sigma(\mathcal{A})$ such that $\mathsf{P} = \bar{\mathsf{P}}$ on $\mathcal{A}$.

- In addition, if $\exists \ \Omega_1 \subset \Omega_2 \subset \Omega_3 ... \in \mathcal{A}$ such that $\Omega_n \uparrow \Omega$, then the extension $\bar{\mathsf{P}}$ is unique.

- On $\mathbb{R}$, Caratheodory's theorem extends a countably additive probability measure on the algebra $\mathcal{A}$ containing all intervals and their finite unions, to a σ-field that is a strict subset of $2^{\mathbb{R}}$ but contains $\mathcal{B} = \sigma(\mathcal{A})$.

Product probability spaces

- Consider two probability spaces $(\Omega_i, \mathcal{F}_i, \mathsf{P}_i)$, $i = 1, 2$.
- Define the product sample space
$$\Omega := \Omega_1 \times \Omega_2 := \{(\omega_1, \omega_2) \mid \omega_i \in \Omega_i,\ i = 1, 2\}$$
- Note that $\mathcal{A}_0 := \{A_1 \times A_2 \mid A_i \in \mathcal{F}_i,\ i = 1, 2\}$ is closed under intersections but *not* under finite unions, *e.g.*, cannot express
$$(A_1 \times A_2) \cup (B_1 \times B_2) \quad \text{as} \quad C_1 \cup C_2$$
where $A_i, B_i, C_i \in \mathcal{F}_i$.
- So, add all finite disjoint unions of elements of $\mathcal{A}_0$ to $\mathcal{A}_0$ and call the result $\mathcal{A}$, an algebra.
- *Denote* $\mathcal{F}_1 \times \mathcal{F}_2 = \sigma(\mathcal{A})$.

Product probability space extension

Theorem: There exists a unique P such that

(a) $\mathsf{P}(A_1 \times A_2) = \mathsf{P}_1(A_1)\mathsf{P}_2(A_2)\ \forall A_1 \times A_2 \in \mathcal{A}_0$, and uniquely extending P to $\mathcal{A}$ with finite unions, and

(b) $(\Omega = \Omega_1 \times \Omega_2, \mathcal{F} = \sigma(\mathcal{A}), \mathsf{P})$ is a probability space.

Proof:

- *Sections of measurable sets are measurable*, where a section of $A \subset \Omega$ along ω_2 is
$$A_{\omega_2} \ := \ \{\omega_1 \in \Omega_1 \mid (\omega_1, \omega_2) \in A\} \text{ for } \omega_2 \in \Omega_2,$$
because $\forall \omega_2 \in \Omega_2$, $\mathcal{M} := \{A \in \mathcal{F} \mid A_{\omega_2} \in \mathcal{F}_1\}$ is a monotone class $\Rightarrow \mathcal{M} = \mathcal{F}$.
- $A = A_1 \times A_2 \in \mathcal{A}_0 \ \Rightarrow A_{\omega_2} = A_1$ if $\omega_2 \in A_2$ otherwise $A_{\omega_2} = \emptyset \ \Rightarrow$
$$\mathsf{P}_1(A_{\omega_2}) = \mathsf{P}_1(A_1)\mathbf{1}_{A_2}(\omega_2) \quad \Rightarrow \quad \mathsf{P}(A) = \int \mathsf{P}_1(A_{\omega_2})\mathrm{d}\mathsf{P}_2(\omega_2),$$
where $\mathsf{P}_1(A_{\omega_2})$ is a $(\Omega_2, \mathcal{F}_2, \mathsf{P}_2)$ random variable.
- Such *disintegration* extends to $A \in \mathcal{A}$ by finite additivity and P is countably additive on $\mathcal{A}$ by dominated convergence.
- So, P (and disintegration) extend uniquely to $\mathcal{F}$ by Caratheodory. $\square$

Fubini-Tonelli Theorem

- **Theorem:** If
  (i) $X : \Omega = \Omega_1 \times \Omega_2 \to \mathbb{R}$ is $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$ measurable and,
  (ii) $\omega_{3-i} \to \int X(\underline{\omega})\mathrm{dP}(\omega_i)$ is a.s. finite and $\mathcal{F}_{3-i}$-measurable $\forall i \in \{1,2\}$,

  then
  $$\mathrm{E}X := \int X\mathrm{dP} = \int \left(\int X\mathrm{dP}_1\right)\mathrm{dP}_2 = \int \left(\int X\mathrm{dP}_2\right)\mathrm{dP}_1.$$
  Proof:
  – By disintegration, $\int X\mathrm{dP}_i$ are $\mathcal{F}_{3-i}$-measurable and the theorem holds for $f = \mathbf{1}_A$, $A \in \mathcal{F}$.
  – Extend to simple functions and take limits via dominated convergence to prove for the case where $E|X| < \infty$. $\square$

- Considering hypothesis (ii), recall how absolute summability of a sequence implies its (unique) summability in any order.

Consistency of Probability Measures

- Consider the product space $(\mathbb{R}^{\mathbb{Z}^+}, \mathcal{B}^{\mathbb{Z}^+}) =: (\mathbb{R}^\infty, \mathcal{B}^\infty)$, where $\mathbb{Z}^+ := \{0,1,2,3,...\}$.

- Again, underlying probability space $(\Omega, \mathcal{F}, \mathrm{P})$.

- A *cylinder event* $A \in \mathcal{B}^\infty$ is of the form
  $$A \;=\; A_0 \times A_1 \times A_2 \times ...$$
  where all but a finite number of $A_i = \Omega$, *i.e.*, there is a finite index $I_A \subset \mathbb{Z}^+$ such that $A_i = \mathbb{R}\ \forall i \notin I_A$.

- A family of probability measures $\{\mathrm{P}^n\}_{n\in\mathbb{Z}^+}$, P_n on $(\mathbb{R}^n, \mathcal{B}^n)$, is said to be *consistent* if
  $$\mathrm{P}^n(A_0 \times A_1 \times \ldots A_{n-1}) \;=\; \mathrm{P}^{n+1}(A_0 \times A_1 \times \ldots A_{n-1} \times \mathbb{R})$$
  for all cylinder sets $A_0 \times A_1 \times \ldots A_{n-1}$.

Kolmogorov's Extension Theorem

For each consistent family of probability measures P^n on $(\mathbb{R}^n, \mathcal{B}^n)$, $\exists!$ consistent P^∞ on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$.

- Clearly, require that
$$\mathsf{P}^\infty(A) = \mathsf{P}^n(A_0 \times A_1 \times ... \times A_{n-1})$$
for all cylinder sets $A = A_0 \times A_1 \times ...$ and all $n \in \mathbb{Z}^+$.

- Since P^∞ is specified for all cylinder sets, P^∞ is unique on the algebra generated by them and, by the monotone class theorem, unique on $\mathcal{B}^\infty$ too.

- For existence:
  - Let $\mathcal{A}$ be the set of finite unions of cylinder sets, including $\emptyset$, so that $\sigma(\mathcal{A}) = \mathcal{B}^\infty$.
  - Show P^∞ is finitely additive on $\mathcal{A}$ and apply Caratheodory's extension theorem.

Consistency and FDDs

- Consider a *discrete-time/parameter stochastic process*
$$X := \{X_t \mid t \in \mathbb{Z}^+\}$$
where each X_t is itself a random variable.

- Let $F_{t_1,t_2,...,t_n}$ be the joint CDF of $X_{t_1}, X_{t_2}, ..., X_{t_n}$ for some finite n and different $t_k \in \mathbb{Z}^+$ for all $k \in \{1, 2, ..., n\}$, *i.e.*,
$$F_{t_1,...,t_n}(x_1, ..., x_n) = \mathsf{P}(X_{t_1} \le x_1, ..., X_{t_n} \le x_n),$$
where P is the underlying probability measure.

- This is called an n-*dimensional distribution* of X.

KET and Consistent FDDs

- A family of such joint CDFs is called a set of *finite-dimensional distributions* (FDDs).

- The FDDs are *consistent* if one can marginalize (reduce the dimension) of one and obtain another, *e.g.*,

$$F_{t_1,t_4}(x_1,x_4) \quad := \quad F_{t_1,t_2,t_3,t_4}(x_1,\infty,\infty,x_4).$$

  then using KET one can prove $\exists!$ a discrete-time *stochastic process* X on $\mathbb{R}^\infty$ (with distribution P^∞), *i.e.*,

$$\mathsf{dP}^n := \mathsf{d}F_{0,1,\ldots,n-1} \quad \text{and} \quad \mathsf{dP}^\infty := \mathsf{d}F_{\mathbb{Z}^+}$$

- Samples ω of the underlying probability space space are actually *sample paths* of the stochastic process, *i.e.*, $X_t(\omega)$.

- KET can be extended to *continuous-time* stochastic processes, *i.e.*, sample paths in $X_t(\omega) \in \mathbb{R}^{\mathbb{R}^+}$ instead of $\in \mathbb{R}^{\mathbb{Z}^+}$.

- In the following, we will focus on the *underlying* probability space Ω and the σ-algebras $\sigma(X_s \mid s \le t)$ for $t \in \mathbb{R}^+ = [0,\infty)$, *i.e.*, in continuous-time...

Uncountable products

Suppose I is an uncountably infinite index set and $\forall t \in I$: $(\Omega, \mathcal{F}_t)$ is a sample space and σ-algebra of events.

- **Theorem:** If $A \in \sigma(\mathcal{F}_t,\ t \in I) =: \mathcal{G}_I$, then there is some *countable* $J \subset I$ (depending on A) such that $A \in \mathcal{G}_J$.
  Proof:
  - Define $\mathcal{H} = \{A \in \mathcal{G}_I \mid \exists \text{ countable } J \subset I \text{ s.t. } A \in \mathcal{G}_J\}$.
  - Clearly, $\Omega \in \mathcal{H}$ and $\mathcal{H}$ is closed under countable unions so that $\mathcal{H}$ is a σ-algebra.
  - Finally since $\mathcal{F}_t \subset \mathcal{H}\ \forall t \in I$, $\mathcal{H} = \mathcal{G}_I$. □

- **Corollary:** If $Y : \Omega \to \bar{\mathbb{R}}$ is $\mathcal{G}_I$-measurable, then $\exists$ a countable $J \subset I$ such that Y is $\mathcal{G}_J$-measurable.
  Proof:
  - Use previous theorem if $Y = \mathbf{1}_A$ and easily extended to simple (discretely distributed) Y.
  - Extend to any (measurable) random variable Y by approximating with simple functions. □

- For the special case where $\mathcal{G}_[0,t] = \sigma(X_s \mid s \le t)$, *i.e.*, $\mathcal{F}_t = \sigma(X_t)$ for random variables X_t: if Y is $\mathcal{F}_t$-measurable then $\exists$ countable $\{t_0, t_1, ...\} \subset [0,t]$ and a $\mathcal{B}^{\mathbb{Z}^+}$-measurable mapping Ψ such that

$$Y \quad = \quad \Psi(X_{t_0}, X_{t_1}, \ldots) \text{ a.s.}$$

  Recall Doob's theorem.