MIT 18.675 FALL 2019. LECTURE TOPICS

The notes below give a brief (not guaranteed to be exhaustive) summary of the main topics covered in each lecture, and relevant reading. It will be updated throughout the semester.

Contents

1. (09/04) Introduction to measure theory 2
2. (09/09) Lebesgue–Stieltjes measures on the real line 2
3. (09/11) Lebesgue–Stieltjes measures on the real line, continued 2
4. (09/16) Random variables (measurable functions) and expectation (Lebesgue integral) 3
5. (09/18) Integral convergence theorems, change of variables formula 4
6. (09/23) Product measures; Tonelli and Fubini theorems 4
7. (09/25) Measures on infinite product spaces 4
8. (09/30) Independence; basic moment inequalities; L^2 weak law of large numbers 5
9. (10/02) Laws of large numbers 6
10. (10/07) Introduction to large deviations theory 7
11. (10/09) Large deviations for the empirical mean: Cramér's theorem 8
12. (10/16) Convolutions; introduction to the Fourier transform 9
13. (10/23) Fourier inversion for probability measures on the real line 12
14. (10/28) Weak convergence and the central limit theorem 13
15. (10/30) Weak convergence: some more examples and theory 14
16. (11/04) Conclusion of weak convergence; introduction to conditional expectation 15
17. (11/06) Conditional expectation; introduction to martingales 16
18. (11/13) Upcrossing inequality, submartingale convergence theorem 17
19. (11/18) L^p martingale convergence theorem, branching processes example 17
20. (11/20) Uniform integrability and L^1 convergence; Doob martingales 18
21. (11/25) Optional stopping theorems 18
22. (11/27) Reverse martingales; Kolmogorov and Hewitt–Savage zero-one laws 19
23. (12/02) Martingale perspective on Radon–Nikodym derivatives 19
24. (12/09) Martingale-based proof of the Kesten–Stigum "$L \log L$ criterion" 20
25. (12/11) The $\zeta(2)$ limit in the random assignment problem 20

References 21

1. (09/04) Introduction to measure theory

1. Example: a game with an infinite sequence of boxes, axiom of choice, and probabilities. (This specific example was told to me by Persi Diaconis, and I don't know the original source. You can see discussion online at *https://mathoverflow.net/questions/151286/probabilities-in-a-riddle-involving-axiom-of-choice*.)
2. Vitali's construction, which shows that there exists no $\mu : \mathscr{P}([0,1]) \to [0,1]$ satisfying (i) $\mu([a,b]) = b-a$, (ii) μ is countably additive, (iii) μ is translation invariant. See [Dur19, §A.2] for a similar example.
3. Formal definition of a measure space $(\Omega, \mathscr{F}, \mu)$ (state space, σ-**field** or σ-**algebra**, **measure**). Some consequences of the definition: $\{\varnothing, \Omega\} \in \mathscr{F}$; $\mu(\varnothing) = 0$; μ is monotone (if $A, B \in \mathscr{F}$ with $A \subseteq B$ then $\mu(A) \le \mu(B)$); continuity from below; continuity from above. Moreover μ is **countably subadditive** over $\mathscr{F}$: if $A, A_i \in \mathscr{F}$ and A is contained in the countable union of the A_i, then
$$\mu(A) \le \sum_{i=1}^{\infty} \mu(A_i)\,.$$
Check for yourself that you can prove all these properties!
4. The **Borel** σ-**field over** $\mathbb{R}$, denoted $\mathscr{B}_{\mathbb{R}}$, is the smallest σ-field over $\mathbb{R}$ that contains the open intervals, $\mathscr{I} = \{(a,b) : -\infty \le a < b \le \infty\}$. We say that $\mathscr{B}_{\mathbb{R}}$ is the σ-field **generated** by $\mathscr{I}$, denoted $\mathscr{B}_{\mathbb{R}} = \sigma(\mathscr{I})$. Check for yourself that it is equivalent to $\mathscr{B}_{\mathbb{R}} = \sigma(\mathscr{S})$ where
$$\mathscr{S} \equiv \Big\{ (a,b] \cap \mathbb{R} : -\infty \le a < b \le \infty \Big\}\,.$$
(Note that $\mathscr{S}$ contains unbounded intervals of the form $(-\infty, b]$ and (a, ∞).) Check that it is also equivalent to $\mathscr{B}_{\mathbb{R}} = \sigma(\mathscr{T})$ where $\mathscr{T}$ is the collection of all open sets in the standard topology on $\mathbb{R}$.

Reading: [Dur19, §A.2], and first part of [Dur19, §1.1].

2. (09/09) Lebesgue–Stieltjes measures on the real line

This lecture was given by Prof. Subhrabata Sen.

1. A **Stieltjes measure function** on $\mathbb{R}$ is a function $F : \mathbb{R} \to \mathbb{R}$ which is nondecreasing and right-continuous.

   Theorem 1. *For any Stieltjes measure function F on $\mathbb{R}$, there is a unique measure $\mu = \mu_F$ on $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ satisfying*
$$\mu((a,b] \cap \mathbb{R}) = F(b) - F(a) \tag{1}$$
   for all $-\infty \le a \le b \le \infty$, where the values of $F(\infty)$ and $F(-\infty)$ are defined by continuity. In the case $F(x) = x$, the corresponding μ is called **Lebesgue measure** *on* $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$.

2. Note that condition (1) defines $\mu : \mathscr{S} \to [0, \infty]$. Let
$$\mathscr{A} \equiv \Big\{ A \subseteq \mathbb{R} : A \text{ is a disjoint union of finitely many elements of } \mathscr{S} \Big\}\,.$$
This is an **algebra** (closed under complementation and finite union); it is the smallest algebra containing $\mathscr{S}$. **Part I** of the proof of Theorem 1: there is a unique $\mu : \mathscr{A} \to [0,\infty]$ which extends $\mu : \mathscr{S} \to [0,\infty]$, and is countably additive over $\mathscr{A}$.

Reading: [Dur19, §1.1].

3. (09/11) Lebesgue–Stieltjes measures on the real line, continued

This lecture was given by Prof. Subhrabata Sen.

1. **Part II** of the proof of Theorem 1: there is a unique $\mu : \mathscr{B}_{\mathbb{R}} \to [0,\infty]$ which extends $\mu : \mathscr{A} \to [0,\infty]$ and is a measure on $\mathscr{B}_{\mathbb{R}}$ (i.e., is countably additive over $\mathscr{B}_{\mathbb{R}}$). In the case that μ is a finite measure ($\mu(\Omega) < \infty$), this is a special case of a more general theorem:

   Theorem 2 (Carathéodory extension theorem)**.** *Suppose $\mathscr{A}$ is an algebra over Ω, and $\mu : \mathscr{A} \to [0,\infty)$ (in particular, $\mu(\Omega) < \infty$) is countably additive over $\mathscr{A}$. Then there is a unique $\mu : \sigma(\mathscr{A}) \to [0,\infty)$ which extends $\mu : \mathscr{A} \to [0,\infty)$ and is a measure on $\sigma(\mathscr{A})$ (i.e., is countably additive over $\sigma(\mathscr{A})$).*

2. Proof of uniqueness in Theorem 2: this is based on **Dynkin's** π-λ **theorem**.

3. Proof of existence in Theorem 2: this is based on the construction of **outer measure**. In the particular setting of Theorem 1 (which is less general than the setting of Theorem 2), the outer measure can be defined as
$$\mu^\star(A) = \inf\left\{ \sum_{i=1}^\infty \mu(E_i) : E_i = (a_i, b_i] \text{ and } A \subseteq \bigcup_{i=1}^\infty E_i \right\},$$
where $\mu(E_i) = F(b_i) - F(a_i)$. Then let
$$\mathscr{F}_F \equiv \Bigg\{ A \subseteq \Omega : \mu^\star(S) = \mu^\star(S \cap A) + \mu^\star(S \setminus A) \text{ for all } S \subseteq \Omega \Bigg\}.$$
We showed in class that $\mathscr{F}_F$ is a σ-algebra, and the restriction of $\mu^\star$ to $\mathscr{F}_F$ is a measure $\mu = \mu_F$. In the case $F(x) = x$, $\mathscr{F}_F$ is the **Lebesgue σ-algebra**, commonly denoted $\mathscr{L}_\mathbb{R}$; and μ_F is the **Lebesgue measure**, commonly denoted λ or Leb. Note $\mathscr{L}_\mathbb{R} \supseteq \mathscr{B}_\mathbb{R}$, so λ further restricts to a measure on $(\mathbb{R}, \mathscr{B}_\mathbb{R})$ which is also called Lebesgue measure.
4. The precise relation between $\mathscr{L}_\mathbb{R}$ and $\mathscr{B}_\mathbb{R}$ is as follows: any $A \in \mathscr{L}_\mathbb{R}$ can be expressed as $A = B \cup N$ where $B \in \mathscr{B}$ and $N \subseteq B' \in \mathscr{B}$ with $\lambda(B') = 0$. This statement is a consequence of [Dur19, Theorem A.2.2]. We did not cover it in lecture (although we may see it in a future lecture or homework).

Reading: [Dur19, §A.1].

4. (09/16) Random variables (measurable functions) and expectation (Lebesgue integral)

1. Let $(\Omega, \mathscr{F})$ be a measurable space. We say f is a **simple function** on $(\Omega, \mathscr{F})$ if $f : \Omega \to \mathbb{R}$ can be expressed as
$$f = \sum_{i=1}^n c_i \mathbf{1}_{A_i} \tag{2}$$
for $c_i \in \mathbb{R}$ and $A_i \in \mathscr{F}$. A **measurable function** on $(\Omega, \mathscr{F})$ is any pointwise limit of simple functions: $f = \lim_n f_n$ where f_n are simple functions on $(\Omega, \mathscr{F})$. A **measurable mapping** from $(\Omega, \mathscr{F})$ to $(S, \mathscr{G})$ is a map $g : \Omega \to S$ such that for all $B \in \mathscr{G}$, the **preimage** $g^{-1}(B)$ belongs to $\mathscr{F}$. To emphasize measurability of g, we often write $g : (\Omega, \mathscr{F}) \to (S, \mathscr{G})$. On Homework 1: f is a measurable function on $(\Omega, \mathscr{F})$ if and only if it is a measurable mapping from $(\Omega, \mathscr{F})$ to $(\mathbb{R}, \mathscr{B}_\mathbb{R})$.
2. If μ is any measure on $(\Omega, \mathscr{F})$ and $g : (\Omega, \mathscr{F}) \to (S, \mathscr{G})$, we obtain a **pushforward measure** ν on $(S, \mathscr{G})$,
$$\nu(B) = \mu(g^{-1}(B)) = \mu\Bigg(\Bigg\{ \omega \in \Omega : g(\omega) \in B \Bigg\}\Bigg).$$
Some standard notations: $\nu = g_\star \mu = g_\# \mu = \mu \circ g^{-1}$.
3. Let $(\Omega, \mathscr{F})$ be a σ-**finite** measure space. If f is a simple function on $(\Omega, \mathscr{F})$ as in (2), with the additional condition that $\mu(A_i) < \infty$ for all i, then we define its Lebesgue integral
$$\int_\Omega f\, d\mu \equiv \sum_i c_i \mu(A_i).$$
If f is a nonnegative measurable function on $(\Omega, \mathscr{F})$, we define its Lebesgue integral
$$\int_\Omega f\, d\mu \equiv \sup\left\{ \int_\Omega h\, d\mu : 0 \le h \le f, h \text{ simple function with } \mu(h > 0) < \infty \right\}.$$
Finally, if f is a general measurable function on $(\Omega, \mathscr{F})$, we define its Lebesgue integral
$$\int_\Omega f\, d\mu = \int_\Omega f_+\, d\mu - \int_\Omega f_-\, d\mu$$
with the caveat that $\infty - \infty = 0$. This completes the general definition of the **Lebesgue integral**
$$\int_\Omega f\, d\mu = \int_\Omega f(\omega)\, d\mu(\omega).$$
It can be finite or $\pm\infty$. Note from the definition that the Lebesgue integral of f is finite if and only if
$$\int_\Omega |f|\, d\mu < \infty,$$

and in this case we call f **integrable**. Basic properties of the integral (monotonicity, linearity, etc.): see the results Lemma 1.4.3, Lemma 1.4.5, Theorem 1.4.7 in [Dur19].

4. A **random variable** is a measurable function X on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Its **distribution** or **law** is $\mathcal{L}_X = X_\# \mathbb{P}$. A useful notation: for $B \in \mathscr{B}_\mathbb{R}$,
$$X^{-1}(B) \equiv \Big\{\omega \in \Omega : X(\omega) \in B\Big\} \equiv \{X \in B\}\,.$$
This allows us to write $\mathcal{L}_X(B)$ in a more intuitively natural way as simply $\mathbb{P}(X \in B)$. The **expectation** or **mean** of X is its Lebesgue integral,
$$\mathbb{E}X = \int_\Omega X\, d\mathbb{P} = \int_\Omega X(\omega)\, d\mathbb{P}(\omega)\,.$$
Note from the definitions that $\mathbb{E}X$ is finite if and only if $\mathbb{E}|X|$ is finite.

Reading: [Dur19, §1.2–1.4]. The Lebesgue integral subsumes the Riemann integral: see for instance the statement of [Tao11, Exercise 1.3.17] (the book gives plenty of hints to prove this statement, which we will take for granted in this class).

5. (09/18) Integral convergence theorems, change of variables formula

1. Modes of convergence: $f_n \to f$ pointwise, μ-almost everywhere (μ-a.e.), in μ-measure.
2. Integral convergence theorems: bounded convergence theorem, Fatou's lemma, monotone convergence theorem, dominated convergence theorem.
3. Application of the monotone convergence theorem to a **change of variables formula**: suppose
$$Y = f(X) : (\Omega, \mathscr{F}, \mathbb{P}) \xrightarrow{X} (S, \mathscr{G}, \mu) \xrightarrow{f} (\mathbb{R}, \mathscr{B}, \nu)$$
where $\mu = X_\#\mathbb{P}$ and $\nu = f_\#\mu = Y_\#\mathbb{P}$. Provided either $f \ge 0$ or Y is integrable, we have $I_\Omega = I_S$ where
$$I_\Omega \equiv \mathbb{E}Y \equiv \int_\Omega Y\, d\mathbb{P} \equiv \int_\Omega f(X(\omega))\, d\mathbb{P}(\omega)\,,$$
$$I_S \equiv \int_S f\, d\mu \equiv \int_S f(x)\, d(\mathbb{P} \circ f^{-1})(x)$$
Proof uses a standard technique: start with indicators, then extend to simple functions, then nonnegative functions (using the monotone convergence theorem), then general functions.

Reading: [Dur19, §1.5–1.6].

6. (09/23) Product measures; Tonelli and Fubini theorems

1. Given two σ-finite measure spaces $(S, \mathscr{G}, \lambda)$ and $(T, \mathscr{H}, \rho)$, we defined the product measure space $(\Omega, \mathscr{F}, \mu) = (S \times T, \mathscr{G} \otimes \mathscr{H}, \lambda \otimes \rho)$.
2. Tonelli and Fubini theorems: for measurable $f : \Omega \to \mathbb{R}$, provided either $f \ge 0$ or $\int |f|\, d\mu < \infty$, we have
$$\int_S \int_T f(x,y)\, d\rho(y)\, d\lambda(x) = \int_\Omega f(x,y)\, d\mu(x,y) = \int_T \int_S f(x,y)\, d\lambda(x)\, d\rho(y)\,.$$
(It is part of the content of the theorem that the left-hand side and right-hand side are well-defined quantities.)

Reading: [Dur19, §1.7].

7. (09/25) Measures on infinite product spaces

1. Let $(\Omega_\alpha, \mathscr{F}_\alpha)$ be measurable spaces indexed by $\alpha \in I$ (index set, possibly uncountable). Their product:
$$(\Omega, \mathscr{F}) = \left(\prod_{\alpha \in I} \Omega_\alpha\,, \bigotimes_{\alpha \in I} \mathscr{F}_\alpha \right) \tag{3}$$
where $\mathscr{F}$ is the minimal σ-field such that the coordinate projections $\pi_\alpha : \Omega \to \Omega_\alpha$ are measurable. For $J \subseteq I$ denote the partial products
$$\Omega_J \equiv \prod_{\alpha \in J} \Omega_\alpha\,, \quad \mathscr{F}_J \equiv \bigotimes_{\alpha \in J} \mathscr{F}_\alpha\,.$$

2. If $\mathbb{P}$ is a probability measure on this $(\Omega, \mathscr{F})$, its **finite-dimensional marginals** are the measures $(\pi_J)_\#\mathbb{P}$. This means that the marginal probability of $E_J \in \mathscr{F}_J$ is
$$\Big((\pi_J)_\#\mathbb{P}\Big)(E_J) = \mathbb{P}\Big((\pi_J)^{-1}(E_J)\Big) = \mathbb{P}\Big(E_J \times \Omega_{I\setminus J}\Big)\,.$$
They have to be **consistent**: if $J' \subseteq J \subseteq I$ and $\pi_{J\to J'}$ is the projection from Ω_J to $\Omega_{J'}$, then
$$(\pi_{J'})_\#\mathbb{P} = (\pi_{J\to J'})_\#\Big((\pi_J)_\#\mathbb{P}\Big)\,.$$
Both the theorems from this lecture go in reverse: given a consistent family of finite-dimensional marginals $\{\mathbb{P}_J : \text{finite } J \subseteq I\}$, they construct a measure $\mathbb{P}$ on $(\Omega, \mathscr{F})$ with these finite-dimensional marginals. Note: $\{\mathbb{P}_J : \text{finite } J \subseteq I\}$ is often also called the **finite-dimensional distributions**, abbreviated **f.d.d.**
3. A useful variant of the criterion for the Carathéodory extension theorem (Theorem 2 above):

   Lemma 3. *Suppose $\mathscr{A}$ is an algebra of sets over Ω, and that $\mu : \mathscr{A} \to [0,\infty)$ is finitely additive over $\mathscr{A}$. If for any sequence $B_n \in \mathscr{A}$ with $B_n \downarrow \varnothing$ we have $\mu(B_n) \downarrow \varnothing$, then μ is countably additive over $\mathscr{A}$.*

   Specialization of Lemma 3 to the infinite product setting (display (3) above): it is enough to prove the criterion in the scenario that we have $Q = \{1, 2, \ldots, \} \subseteq I$ countable and $B_n \downarrow \varnothing$ of the form
$$B_n = \bar{B}_n \times \prod_{i=n+1}^{\infty} \Omega_i \times \prod_{\alpha\in I\setminus Q} \Omega_\alpha = \bar{B}_n \times \Omega_{Q\setminus[n]} \times \Omega_{\text{rest}}$$
where $[n] \equiv \{1, \ldots, n\}$ and $\bar{B}_n \in \mathscr{F}_{[n]}$.
4. Product measures on an infinite-dimensional spaces:

   Theorem 4 (Ionescu–Tulcea theorem). *Let $(\Omega_\alpha, \mathscr{F}_\alpha, \mathbb{P}_\alpha)$ be probability spaces indexed by $\alpha \in I$. There is a unique probability measure $\mathbb{P}$ on the product space $(\Omega, \mathbb{P})$ (as in (3)) with finite-dimensional marginals $(\pi_J)_\#\mathbb{P} = \bigotimes_{\alpha\in J} \mathbb{P}_\alpha$.*

5. Non-product measures on an infinite-dimensional spaces:

   Theorem 5 (Kolmogorov extension theorem). *Let $(\Omega_\alpha, \mathscr{F}_\alpha)$ be metric spaces with the Borel σ-field. Suppose $\{\mathbb{P}_J : \text{finite } J \subseteq I\}$ is a consistent family of finite-dimensional distributions, and that each $\mathbb{P}_J$ is an inner regular measure on $(\Omega_J, \mathscr{F}_J)$. Then there is a unique probability measure $\mathbb{P}$ on $(\Omega, \mathscr{F})$ with finite-dimensional marginals $(\pi_J)_\#\mathbb{P} = \mathbb{P}_J$.*

Reading: [Dur19, Thm. 2.1.14 and §A.3], and [Kal02, Ch. 5] (especially Thm. 5.14, Thm. 5.17 and Cor. 5.18).

8. (09/30) Independence; basic moment inequalities; L^2 weak law of large numbers

1. Independence (of events, of collections of events, and of random variables). Connection with previous week:
   a. If the probability space comes with a product structure
$$(\Omega, \mathscr{F}, \mathbb{P}) = \left(\prod_{\alpha\in I} \Omega_\alpha, \bigotimes_{\alpha\in I} \mathscr{F}_\alpha, \bigotimes_{\alpha\in I} \mathbb{P}_\alpha\right),$$
   and $X_\alpha(\omega) = f_\alpha(\omega_\alpha)$ where f_α is a measurable function on Ω_α, then $(X_\alpha : \alpha \in I)$ is a collection of independent random variables on $(\Omega, \mathscr{F}, \mathbb{P})$.
   b. If $(\Omega, \mathscr{F}, \mathbb{P})$ is a general probability space (i.e., not necessarily equipped with an explicit product structure) and we are told that $(X_\alpha : \alpha \in I)$ is a collection of random variables on $(\Omega, \mathscr{F}, \mathbb{P})$, we can define $X(\omega) \equiv (X_\alpha(\omega) : \alpha \in I)$. This gives a mapping
$$(\Omega, \mathscr{F}, \mathbb{P}) \xrightarrow{X} \left(\prod_{\alpha\in I} \mathbb{R}, \bigotimes_{\alpha\in I} \mathscr{B}_\mathbb{R}, \mathcal{L}_X\right)$$
   where $\mathcal{L}_X \equiv X_\#\mathbb{P} = \mathbb{P} \circ X^{-1}$ is the law of X (as defined previously). Then $(X_\alpha : \alpha \in I)$ is a collection of independent random variables if and only if $\mathcal{L}_X$ is a product measure.

2. Moments of random variables: for $p \in (0,\infty)$ and X a random variable we define
$$\|X\|_p \equiv \|X\|_{L^p(\Omega,\mathscr{F},\mathbb{P})} \equiv \left\{ \int |X(\omega)|^p \, d\mathbb{P}(\omega) \right\}^{1/p} \equiv \mathbb{E}(|X|^p)^{1/p} .$$
We write $L^p \equiv L^p(\Omega,\mathscr{F},\mathbb{P})$ for the collection of X with $\|X\|_p < \infty$. L^p monotonicity: if $r \le p$ then $\|X\|_r \le \|X\|_p$, so $L^p(\Omega,\mathscr{F},\mathbb{P}) \subseteq L^r(\Omega,\mathscr{F},\mathbb{P})$. Be care that for L^p monotonicity it is essential to be on a *probability* space. The ℓ_p sequence spaces are nested in the opposite direction: if $r \le p$ then
$$\|x\|_r \equiv \left(\sum_i |x_i|^r \right)^{1/r} \ge \left(\sum_i |x_i|^p \right)^{1/p} \equiv \|x\|_p ,$$
and so the ℓ_p unit ball contains the ℓ_r unit ball. Finally, the classical L^p function spaces are not nested at all: if
$$\|f\|_p \equiv \|f\|_{L^p(\mathbb{R})} \equiv \left\{ \int |f(x)|^p \, dx \right\}^{1/p}$$
and $L^p(\mathbb{R}) \equiv \{f : \|f\|_p < \infty\}$, both $L^p(\mathbb{R}) \setminus L^r(\mathbb{R})$ and $L^r(\mathbb{R}) \setminus L^p(\mathbb{R})$ are nonempty for $r \neq p$.
3. If $X \in L^2(\Omega,\mathscr{F},\mathbb{P})$, we define its **variance** $\operatorname{Var} X = \mathbb{E}[(X-\mathbb{E}X)^2]$. **Cauchy–Schwarz inequality**: if $X,Y \in L^2$ then $\|XY\|_1 = \mathbb{E}|XY| \le \|X\|_2\|Y\|_2$. Consequently, for $X,Y \in L^2$ we can define their **covariance**
$$\operatorname{Cov}(X,Y) \equiv \mathbb{E}\Big\{(X-\mathbb{E}X)(Y-\mathbb{E}Y)\Big\}, \quad |\operatorname{Cov}(X,Y)| \le \Big\{(\operatorname{Var} X)(\operatorname{Var} Y)\Big\}^{1/2} .$$
We say X and Y are **uncorrelated** if $\operatorname{Cov}(X,Y) = 0$.
4. Markov inequality; Chebychev inequality; and a simple version of the L^2 **weak law of large numbers**: if X, X_i are pairwise uncorrelated and identically distributed, and $S_n \equiv X_1 + \ldots + X_n$, then $S_n/n \to \mathbb{E}X$ in L^2, hence also in probability (by Chebychev). We noted that the L^2 convergence is a restatement of a simple geometric fact: if $(v_1,\ldots,v_n)$ are orthonormal vectors, then their average has small euclidean norm:
$$\left\| \frac{1}{n} \sum_{i=1}^n v_i \right\|_2 = \frac{1}{n^{1/2}}$$
(v_i corresponds to $X_i - \mathbb{E}X$).

Reading: PTE 2.1.

9. (10/02) Laws of large numbers

Setting for entirety of this lecture: triangular array with n-th row given by $(X_{n,k} : 1 \le k \le n)$. Assume independence within each row. Important special case is $X_{n,k} = X_k$ where X_k are i.i.d. Sum of n-th row is
$$S_n \equiv \frac{1}{n} \sum_{k=1}^n X_{n,k} .$$
Interested in convergence of S_n/b_n or $(S_n - \mathbb{E}S_n)/b_n$ for b_n deterministic. WLLN stands for "weak law of large numbers," SLLN stands for "strong law of large numbers."

1. L^2 **WLLN for triangular arrays.** If $\operatorname{Var} X_{n,k} \le C$ uniformly, then $(S_n - \mathbb{E}S_n)/n \to 0$ in L^2 as $n \to \infty$, and hence also in probability (by Chebychev).
2. **WLLN for triangular arrays.** Suppose $b_n \to \infty$ such that
$$\text{(i)} \lim_{n\to\infty} \sum_{k=1}^n \mathbb{P}(|X_{n,k}| > b_n) = 0 ,$$
$$\text{(ii)} \lim_{n\to\infty} \sum_{k=1}^n \frac{\mathbb{E}((X_{n,k})^2; |X_{n,k}| \le b_n)}{(b_n)^2} = 0 .$$
Let $Y_{n,k} \equiv X_{n,k}\mathbf{1}\{|X_{n,k}| \le b_n\}$ and define the truncated row sums $T_n \equiv Y_{n,1} + \ldots + Y_{n,n}$. Then $(S_n - \mathbb{E}T_n)/b_n \to 0$ in probability. *Proof synopsis:* $\mathbb{P}(S_n \neq T_n) \to 0$ *by (i).* $(T_n - \mathbb{E}T_n)/b_n \to 0$ *in* L^2 *by (ii), hence also in probability by Chebychev.*

3. **WLLN for i.i.d. sequences.** Let X, X_k i.i.d. with $\lim_{x\to\infty} x\mathbb{P}(|X| \ge x) = 0$. Define $\mu_n \equiv \mathbb{E}(X; |X| \le n)$. Then $(S_n/n - \mu_n) \to 0$ in probability. *Note that* $\mathbb{E}X$ *need not be defined. The proof is an application of the previous result together with the dominated convergence theorem.*
4. **SLLN for i.i.d. sequences.** Let X, X_k i.i.d. with $\mathbb{E}X = \mu$ finite, then $S_n/n \to \mu$ almost surely.

Reading: PTE 2.2–2.4.

10. (10/07) Introduction to large deviations theory

1. Suppose X, X_i i.i.d. with $\mathbb{E}X = \mu$ finite. The SLLN from the previous lecture tells us that the empirical mean
$$\bar{X}_n = \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^n X_i$$
converges almost surely to μ. In this lecture we study the chance of a **large deviation**, $S_n \ge na$ for $a > \mu$.
2. If the X_i are i.i.d. standard gaussians, then $S_n \sim \mathcal{N}(0, n)$, and
$$\mathbb{P}(S_n \ge na) = \int \frac{\mathbf{1}\{z \ge \sqrt{n}a\}}{\sqrt{2\pi}\exp(z^2/2)}\,dz = \exp\left\{ -\frac{na^2}{2} + o(n)\right\}.$$
3. If the X_i are i.i.d. $\mathrm{Ber}(p)$ for $p \in (0,1)$, then $S_n \sim \mathsf{Bin}(n,p)$, and we used Stirling's formula to estimate
$$\mathbb{P}(S_n = nx) = \binom{n}{nx} p^{nx}(1-p)^{n(1-x)} = \exp\left\{ -nI_p(x) + o(n)\right\}$$
where the exponent is the binary relative entropy function,
$$I_p(x) \equiv \mathcal{H}(x|p) \equiv x\log\frac{x}{p} + (1-x)\log\frac{1-x}{1-p}\,.$$
For x near p we made the change of variables
$$x = p + \frac{\sqrt{p(1-p)}z}{\sqrt{n}}$$
to see that the distribution of $(S_n - np)/\sqrt{np(1-p)}$ is approximately standard gaussian, in the sense that
$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \in [u,v]\right) \to \int_u^v \frac{dz}{\sqrt{2\pi}\exp(z^2/2)}$$
for any **fixed** $u, v \in \mathbb{R}$. However, for any fixed $a \in (p,1)$, the large deviations probability $\mathbb{P}(S_n \ge na)$ is outside the regime of the gaussian approximation, and
$$\mathbb{P}(S_n \ge na) = \exp\left\{ -n\mathcal{H}(a|p) + o(n)\right\} \ne \exp\left\{ -\frac{n(a-p)^2}{2p(1-p)} + o(n)\right\}$$
(where the expression in gray is the naive gaussian approximation).
4. Lastly, in this lecture we saw how to answer the following question: if in a $k \times n$ table we fill nkp entries uniformly at random, what is the probability of the event F that every column has at least one filled entry?
$$\mathbb{P}(F) = \mathbb{P}_\theta\left(X_i \ge 1\,\forall 1 \le i \le n \,\middle|\, \sum_{i=1}^n X_i = nkp\right)$$
where under $\mathbb{P}_\theta$ we let X, X_i be i.i.d. $\mathsf{Bin}(k,\theta)$. Then
$$\mathbb{P}(F) = \frac{\mathbb{P}_\theta(X \ge 1)^n}{\mathbb{P}(\mathsf{Bin}(nk,\theta) = nkp)}\mathbb{P}_\theta\left(\sum_{i=1}^n X_i = nkp \,\middle|\, X_i \ge 1\,\forall 1 \le i \le n\right).$$
We discussed how to calculate each of the three factors for a good choice of θ.

Reading: first part of PTE 2.6. Please also review the definition and basic properties of (multivariate) gaussian random variables. If $Z_1, \ldots, Z_d$ are i.i.d. standard gaussian, then $Z \equiv (Z_1, \ldots, Z_d)$ is an $\mathbb{R}^d$-valued random variable, and we denote its law by $\mathcal{N}(\mathbf{0}, I_{d\times d})$; this is the standard gaussian in $\mathbb{R}^d$. For any $\mu \in \mathbb{R}^k$ and $A \in \mathbb{R}^{k\times d}$, the $\mathbb{R}^k$-valued random variable $X = \mu + AZ$ is a multivariate gaussian, and we denote its law by $\mathcal{N}(\mu, AA^{\mathsf{t}})$. For a quick introduction see the chapters on the normal distribution in *Probability* by J. Pitman (Springer, 1999).

11. (10/09) Large deviations for the empirical mean: Cramér's theorem

1. If X is a (real-valued) random variable, we define its **moment-generating function** (mgf) as $m(\theta) = \mathbb{E}(\exp(\theta X)) \in (0, \infty]$. See [Dur19, Lem. 2.6.2] for its key properties. The **cumulant-generating function** (cgf) is $\kappa(\theta) = \log m(\theta) \in (-\infty, \infty]$. Let
$$\mathcal{D} \equiv \left\{ \theta \in \mathbb{R} : m(\theta) < \infty \right\},$$
and note $0 \in \mathcal{D}$ always. It is possible that $\mathcal{D} = \{0\}$.
2. Large deviations estimate for the empirical mean of i.i.d. random variables:

   Theorem 6 (Cramér's theorem). *Let X, X_i i.i.d. with cgf $\kappa(\theta)$. If $\kappa(\theta) < \infty$ for some $\theta > 0$, then $\mu = \mathbb{E}X \in [-\infty, \infty)$ is well-defined. For any $\mu < a < \operatorname{ess\,sup} X$, the sum $S_n = X_1 + \ldots + X_n$ satisfies the large deviations estimate*
$$-\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \ge na) = I(a) \equiv \sup_{\theta \ge 0} \left\{ \theta a - \kappa(\theta) \right\} \overset{\odot}{=} \sup_{\theta \in \mathbb{R}} \left\{ \theta a - \kappa(\theta) \right\} \equiv \kappa^\star(a),$$
the **Legendre dual** *or* **Fenchel–Legendre transform** *of κ.*

   Note the large deviations estimate can be written equivalently as $\mathbb{P}(S_n/n \ge a) = \exp\{-nI(a) + o(n)\}$.
3. The upper bound of Theorem 6 is an easy consequence of Markov's inequality: for any $\theta \ge 0$,
$$\mathbb{P}(S_n \ge na) \le \mathbb{P}(e^{\theta S_n} \ge e^{n\theta a}) \le \frac{\mathbb{E}(e^{\theta S_n})}{e^{n\theta a}} = \exp\left\{ -n \left[\theta a - \kappa(\theta) \right] \right\}.$$
(The exponential form of Markov's inequality is often called a **Chernoff bound**.)
4. For $\theta \in \mathcal{D}$ (as defined above), we can define the **change of measure**
$$\mathbb{P}_\theta(X \in A) = \mathbb{E}\left[\mathbf{1}\{X \in A\} \frac{e^{\theta X}}{m(\theta)} \right].$$
The probability measure $\mathbb{P}_\theta$ is sometimes called an **exponential tilt** of $\mathbb{P}$. For θ in the interior of $\mathcal{D}$, we have $\kappa'(\theta) = \mathbb{E}_\theta X$ and $\kappa''(\theta) = \operatorname{Var}_\theta X > 0$ (assuming the law of X is nondegenerate). Thus κ is strictly convex in the interior of $\mathcal{D}$. For $\theta > 0$ small enough we have $\mu < \kappa'(\theta) < a$, so $\theta a - \kappa(\theta)$ is increasing in θ for $\theta > 0$ small enough; and this implies the equality marked $\odot$ in Theorem 6.
5. We proved the lower bound of Theorem 6 in the special case that the supremum over θ in $\kappa^\star(a)$ is achieved by θ_a in the interior of the set of θ where $m(\theta) < \infty$. This is done by a change of measure to $\mathbb{P}_\theta$ in which $\{S_n \ge na\}$ becomes a **typical** rather than **rare** event. (Read in [Dur19, §2.6] for the proof of the Theorem 6 in its full generality.)
6. Informal interpretation of Cramér's theorem: "the most efficient way to achieve a large deviation $S_n \ge na$ is for $X_1, \ldots, X_n$ to 'behave like' a sample from $\mathbb{P}_\theta$, with θ chosen such that $\mathbb{E}_\theta X = a$." The exponential change of measure in Cramér's theorem is "optimal" in the *a posteriori* sense that the upper and lower bounds match.
7. In the special case that the law of X has finite support, we can make explicit combinatorial calculations that yield an exponentially tilted measure. Suppose $\mathbb{P}(X = x_j) = \pi_j$ for $1 \le j \le k$. The **empirical measure** of $X_1, \ldots, X_n$ is the random measure
$$L_n^{\mathbf{X}} \equiv \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$
This is well-defined in general, but in the special case where the law of X is supported on $\{x_1, \ldots, x_k\}$, we can equivalently regard $L_n^{\mathbf{X}}$ as the k-tuple of empirical fractions
$$L_n^{\mathbf{X}} = \left(\frac{|\{1 \le i \le n : X_i = x_j\}|}{n} \right)_{1 \le j \le k} \in [0,1]^k.$$

Moreover, in this special case we can directly calculate

$$\mathbb{P}(L_n^{\boldsymbol{X}} = \nu) = \frac{n!(\pi_1)^{n\nu_1}\cdots(\pi_k)^{n\nu_k}}{(n\nu_1)!\cdots(n\nu_k)!} \equiv \binom{n}{n\nu}\pi^{n\nu} = \exp\left\{-n\mathcal{H}(\nu|\pi) + o(n)\right\}$$

where $\mathcal{H}(\nu|\pi)$ is the **relative entropy** or **Kullback–Leibler divergence** between ν and π:

$$\mathcal{H}(\nu|\pi) \equiv D_{\mathrm{KL}}(\nu|\pi) \equiv \sum_{j=1}^k \nu_j \log\frac{\nu_j}{\pi_j} \ge 0\,,$$

a convex function of ν. The empirical mean of $X_1,\ldots,X_n$ can be expressed in terms of the empirical measure as $S_n/n = \langle x, L_n^{\boldsymbol{X}}\rangle$. Then, for $\mathbb{E}X < a < \operatorname{ess\,sup} X = \max\{x_j : 1 \le j \le k\}$, we can calculate

$$\mathbb{P}(S \ge na) = \sum_\nu \mathbf{1}\{\langle x,\nu\rangle \ge a\}\mathbb{P}(L_n^{\boldsymbol{X}} = \nu) = \exp\left\{-n\inf\left\{\mathcal{H}(\nu|\pi) : \langle x,\nu\rangle \ge a\right\} + o(n)\right\}.$$

The last approximation uses that the set of all possible values $L_n^{\boldsymbol{X}} = \nu$ has cardinality $n^{O(1)}$, and is dense in the simplex $\Delta_k \equiv \{p \in [0,1]^k : p_1 + \ldots + p_k = 1\}$. The Lagrangian for the constrained optimization problem is

$$\mathcal{L}(\nu,\theta) = \mathcal{H}(\nu|\pi) + \rho\left(1 - \sum_j \nu_j\right) + \theta\left(1 - \sum_j x_j\nu_j\right),$$

and setting $\partial\mathcal{L}/\partial\nu_j = 0$ gives the exponentially tilted measure

$$\nu = \frac{\pi_j\exp(\theta x_j)}{m(\theta)} = \pi_\theta\,.$$

Check for yourself that $\mathcal{H}(\pi_\theta|\pi) = \theta\mathbb{E}_\theta X - \kappa(\theta) = \theta\kappa'(\theta) - \kappa(\theta)$. For $\mathbb{E}X < a < \operatorname{ess\,sup} X$, there is a unique $\theta(a)$ such that $\kappa'(\theta(a)) = a$. Therefore

$$\begin{aligned}\mathbb{P}(S_n \ge na) &= \exp\left\{-n\inf\left\{\theta\kappa'(\theta) - \kappa(\theta) : \kappa'(\theta) \ge a\right\} + o(n)\right\}\\ &= \exp\left\{-n\inf_{b\ge a}\left\{\theta(b)\kappa'(\theta(b)) - \kappa(\theta(b))\right\} + o(n)\right\}\\ &= \exp\left\{-n\left\{\theta(a)\kappa'(\theta(a)) - \kappa(\theta(a))\right\} + o(n)\right\} = \exp\left\{-n\kappa^\star(a) + o(n)\right\}\end{aligned}$$

(in particular, check in the above that the infimum over $b \ge a$ is achieved at $b = a$).

8. Finally, we remark that the key assumption of Theorem 6, that $m(\theta) < \infty$ for some $\theta > 0$, is a very strong assumption. For instance it fails if $\mathbb{P}(X \ge x)$ decays like $1/x^p$ for any finite p. Convince yourself in this case that $\mathbb{P}(S_n \ge na)$ generally does not decay exponentially in n.

Reading: rest of PTE 2.6, and PTE 3.1.

12. (10/16) Convolutions; introduction to the Fourier transform

In this lecture we reviewed some miscellaneous topics (in particular, the convolution of measures) and introduced the Fourier transform (to be resumed after the first exam):

1. Let X be a (real-valued) random variable on $(\Omega,\mathscr{F},\mathbb{P})$. Recall that its distribution or law $\mu = \mathcal{L}_X = X_\#\mathbb{P}$ is a probability measure on $(\mathbb{R},\mathscr{B}_\mathbb{R})$. Its cumulative distribution function (cdf) is the function $F(x) \equiv \mathbb{P}(X \le x) = \mu((-\infty,x])$. Clearly, μ uniquely determines F. The converse is also true: F uniquely determines μ, which can be proved by a π-λ argument. It is common to write "$dF(x)$" which has the same meaning as "$d\mu(x)$ where μ is the measure determined by F."
2. Let X and Y be (real-valued) random variables on $(\Omega,\mathscr{F},\mathbb{P})$, with laws $\mu \equiv \mathcal{L}_X$ and $\nu \equiv \mathcal{L}_Y$, and cdf's $F(x) \equiv \mathbb{P}(X \le x)$ and $G(y) \equiv \mathbb{P}(y \le y)$. Assume X and Y are **independent**, i.e., their joint law $\mathcal{L}_{(X,Y)}$ is given by the product of their marginal laws, $\mu\otimes\nu$. The resulting law of $Z = X + Y$ is then defined to be the **convolution** of μ and ν, denoted $\mu * \nu$. The associated cdf is denoted $F * G$, and is given explicitly by

$$(F * G)(z) = \mathbb{P}(X + Y \le z) = \int\int \mathbf{1}\{x + y \le z\}\,d\mu(x)\,d\nu(y) = \int F(z-y)\,dG(y)\,,$$

having used the change of variables formula and Tonelli's theorem, and recalling that "$dG(y)$" is equivalent notation for "$d\nu(y)$." Symmetrically, we also have the formula

$$(F * G)(z) = \mathbb{P}(X + Y \le z) = \int\int \mathbf{1}\{x + y \le z\}\, d\nu(y)\, d\mu(x) = \int G(z-x)\, dF(x)\,.$$

The above definitions make sense for any probability measures μ, ν on the real line.

3. In the special case that μ and ν have **densities** — meaning that

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \quad \text{and} \quad G(y) = \int_{-\infty}^{y} g(t)\, dt$$

for nonnegative measurable functions f and g, the convolution $\mu * \nu$ also has a density:

$$(F * G)(z) = \int F(z-y) g(y)\, dy = \int \int_{-\infty}^{z} f(t-y)\, dt\, g(y)\, dy = \int_{-\infty}^{z} \left[\int f(t-y) g(y)\, dy \right] dt\,,$$

so the density for $\mu * \nu$ is given by the last expression above in square brackets:

$$(f * g)(z) = \int f(z-y) g(y)\, dy = \int g(z-x) f(x)\, dx$$

(the last identity holds by symmetry).

4. For a probability measure μ on $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$, the **characteristic function** (chf) — equivalently, the **Fourier transform** — is the function $\varphi_\mu : \mathbb{R} \to \mathbb{C}$ defined by

$$\varphi_\mu(t) \equiv \int e^{itx}\, d\mu(x)\,. \tag{4}$$

The function φ_μ is well-defined, continuous, and satisfies $|\varphi_\mu(t)| \le 1$ for all $t \in \mathbb{R}$. If X is a random variable with law μ, then (using the change of variables formula) we can express $\varphi_\mu(t) = \mathbb{E}(\exp(itX))$. **The Fourier transform takes convolution to multiplication**: if X has law μ, Y has law ν, and X and Y are independent, then (using Fubini's theorem) we have

$$\varphi_{\mu*\nu}(t) = \mathbb{E}\left[e^{it(X+Y)} \right] = \mathbb{E}(e^{itX}) \mathbb{E}(e^{itY}) = \varphi_\mu(t) \varphi_\nu(t)\,.$$

Reading: PTE 2.1.3, and PTE Theorem 3.3.1.

ANY MATERIAL ABOVE THIS LINE CAN APPEAR ON EXAM 1

5. Some of the basic mechanics of the Fourier transform can be worked out very easily and explicitly by considering the discrete space $\Omega = \mathbb{Z}/n\mathbb{Z}$ (the integers modulo n). Below is an outline of some of the basic calculations, which is essentially an exercise in linear algebra. Please review it, especially if you are not very familiar with the Fourier transform.

   a. Let V denote the space of functions $f : \Omega \to \mathbb{C}$. Then $V \cong \mathbb{C}^\Omega$, a finite-dimensional complex vector space. It is naturally equipped with the **hermitian inner product**

   $$\langle f, g \rangle \equiv \sum_{x \in \Omega} f(x) \overline{g(x)} \equiv g^* f\,,$$

   where in the last expression we regard f and g as $n \times 1$ column vectors, and denote their conjugate transposes by f^* and g^* (these are then $1 \times n$ row vectors). For $f \in V$ we define the L^2 norm $\|f\|_2 = \langle f, f \rangle^{1/2}$. This is finite for all $f \in V$ (since it is a finite-dimensional space), and we sometimes also write $V \equiv L^2(\Omega)$ to emphasize the inner product structure.

   b. For $z \in \Omega$, define the **translation operator** $T_z : V \to V$ by

   $$(T_z f)(x) = f(x - z)\,.$$

   If $z \equiv$ (modulo n) then T_z is the identity operator. If $z \in \Omega \setminus \{0\}$ then T_z acts nontrivially on the space V. If f is an eigenfunction of T_z with eigenvalue λ, then $f(x-z) = \lambda f(x)$ for all x, so $f(x - kz) = \lambda^k f(x)$ for all integers k. If $kz \equiv 0$ modulo n, then we must have $\lambda^k = 1$. In particular, we always have $nz \equiv 0$ modulo n, so λ must be an n-th root of unity:

   $$\lambda \in \Phi_n \equiv \{z \in \mathbb{C} : z^n = 1\} = \left\{ 1, \exp\left(\frac{2\pi i}{n}\right), \ldots \exp\left(\frac{2\pi i(n-1)}{n}\right) \right\}.$$

(In general, if n is not prime, then it is possible to have $kz \equiv 0$ modulo n for $1 < k < n$, depending on the common factors between z and n.)

c. In the simplest case where n is **prime**, we see for any $z \in \Omega \setminus \{0\}$, the operator T_z has n **distinct** eigenvalues, given exactly by the set Φ_n of n-th roots of unity. Once we know the eigenvalues, it is easy to solve for the corresponding eigenvectors: for any $y, z \in \Omega$, the vector $\chi_y \in V$ defined by
$$\chi_y(x) \equiv \frac{1}{\sqrt{n}} \exp\left(\frac{2\pi i x y}{n}\right)$$
is an eigenvector of T_z with eigenvalue
$$\lambda_{z,y} \equiv \exp\left(-\frac{2\pi i y z}{n}\right) \in \Phi_n\,.$$
Let us emphasize that the vectors $\chi_0, \ldots, \chi_{n-1}$ form an eigenbasis for T_z for **every** $z \in \Omega$, that is to say, the operators T_z are **simultaneously diagonalizable**. It is not surprising that this occurs: if n is prime then Ω is a **field**, so for any $x, y \in \Omega \setminus \{0\}$ we have $y = rx$ for $r \neq 0$, so $T_y = T_{rx} = (T_x)^r$. Thus, if χ is an eigenvector of T_x with eigenvalue λ, then it must also be an eigenvector of T_y with eigenvalue λ^r. It follows that any eigenbasis for T_x is also an eigenbasis for T_y. The set of eigenvalues is also the same, since the mapping $\lambda \mapsto \lambda^r$ gives an automorphism of the set Φ_n. (Of course, the correspondence between eigenvectors and eigenvalues is permuted when we compare T_x with T_y.)

d. In the general case where n **need not be prime**, it is still the case that the vectors $\chi_0, \ldots, \chi_{n-1}$ are eigenvectors of T_z with eigenvalues $\lambda_{z,y}$ — the only difference in the general case is that this need not be the unique eigenbasis, since it is no longer necessarily the case that $\lambda_{z,y}$ goes over n distinct values as y goes over Ω. The χ_y are the canonical **Fourier basis** for the space $V = L^2(\Omega)$. Let U be the $n \times n$ matrix with columns given by the Fourier basis vectors:
$$U_{x,y} = \chi_y(x) = \frac{1}{\sqrt{n}} \exp\left(\frac{2\pi i x y}{n}\right).$$
Note that U is symmetric, so $U^* = \bar{U}$ (the entrywise conjugate of U). One can check that
$$\langle \chi_x, \chi_y \rangle = \sum_{z \in \Omega} \chi_x(z)\overline{\chi_y(z)} = \sum_{z \in \Omega} \frac{1}{n} \exp\left(\frac{2\pi i (x-y) z}{n}\right) = \mathbf{1}\{x = y\}\,,$$
so U is a **unitary** matrix: $U^* U = I_{n \times n} = U U^*$. The diagonalization of T_z is given by
$$T_z = U \Lambda_z U^* = \sum_{\ell \in \Omega} \lambda_{z,\ell} \chi_\ell (\chi_\ell)^* \tag{5}$$
where Λ_z denotes the diagonal matrix with diagonal entries $(\lambda_{z,\ell} : \ell \in \Omega)$.

e. The **Fourier transform** on the space $\Omega = \mathbb{Z}/n\mathbb{Z}$ is nothing but the **change of basis** operation that sends $f \in V = L^2(\Omega)$ to $\hat{f} \equiv U^* f$, which gives its coordinates in the Fourier basis:
$$f = U U^* f = U \hat{f} = \sum_{\ell \in \Omega} \hat{f}(\ell) \chi_\ell\,.$$
The **Fourier coefficients** $\hat{f}(\ell)$ are given explicitly by
$$\hat{f}(\ell) = (U^* f)_\ell = (\chi_\ell)^* f = \langle f, \chi_\ell \rangle = \frac{1}{\sqrt{n}} \sum_{x \in \Omega} f(x) \exp\left(-\frac{2\pi i \ell x}{n}\right)$$
— note the similarity between this expression and the definition (4) of the characteristic function.

f. Recall that a **hermitian** matrix $H = H^*$ can always be diagonalized by a unitary matrix: $H = UDU^\star$ where D is diagonal with **real** entries. This does not apply above, since T_z is in general **not** hermitian: T_z is a real-valued matrix with entries $(T_z)_{x,y} = \mathbf{1}\{y = x - z\}$, from which we can work out that $(T_z)^* = (T_z)^{\mathrm{t}} = T_{-z} = (T_z)^{-1}$. If χ and ψ are eigenvectors of T_z with distinct eigenvalues $\lambda \neq \gamma$, then
$$\begin{aligned}
\langle T_z \chi, \psi \rangle &= \langle \lambda \chi, \psi \rangle = \lambda \langle \chi, \psi \rangle\,,\\
\langle T_z \chi, \psi \rangle &= \langle \chi, (T_z)^* \psi \rangle = \langle \chi, (T_z)^{-1} \psi \rangle = \langle \chi, \gamma^{-1} \psi \rangle = \overline{\gamma^{-1}} \langle \chi, \psi \rangle = \gamma \langle \chi, \psi \rangle\,,
\end{aligned}$$

where the last relation uses that γ lies on the unit circle in $\mathbb{C}$. If $\lambda \neq \gamma$, the above is a contradiction unless $\langle \chi, \psi \rangle = 0$. This calculation shows why we can expect the Fourier basis to be orthonormal. (This argument simply mimics the usual proof that a hermitian matrix has an orthonormal basis.)

g. For $f, g \in V = L^2(\Omega)$, we define their (discrete) convolution $f * g \in V$ as

$$(f * g)(x) = \sum_{z \in \Omega} g(z) f(x - z) = \sum_{z \in \Omega} g(z)(T_z f)(x) = (C_g f)(x) .$$

The above calculation shows that the **convolution operator** $C_g : f \mapsto f * g$ can be expressed as a **linear combination of translation operators**:

$$C_g = \sum_{z \in \Omega} g(z) T_z ,$$

Since we saw in above that the T_z are simultaneously diagonalizable by the Fourier basis matrix U, it follows that C_g is also diagonalizable by U: explicitly, it follows using (5) that

$$C_g = \sum_{z \in \Omega} g(z) \sum_{\ell \in \Omega} \lambda_{z,\ell} \chi_\ell (\chi_\ell)^* = \sum_{\ell \in \Omega} \left[\sum_{z \in \Omega} \lambda_{z,\ell} g(z) \right] \chi_\ell (\chi_\ell)^* = \sum_{\ell \in \Omega} (U^* g)_\ell \chi_\ell (\chi_\ell)^* ,$$

where the last equality uses that $\lambda_{z,\ell} = (U^*)_{z,\ell}$. In more succinct form, if $\operatorname{diag}(U^* g)$ denotes the diagonal matrix with diagonal entries given by $U^* g$, then the above shows that

$$C_g = U \operatorname{diag}(U^* g) U^* .$$

It follows from this that

$$\widehat{f * g} = U^*(f * g) = U^* C_g f = U^* \Big(U \operatorname{diag}(U^* g) U^* \Big) f = \operatorname{diag}(U^* g) U^* f = \operatorname{diag}(\hat{g}) \hat{f} = \hat{f} \odot \hat{g} ,$$

where $\odot$ denotes the Hadamard product (or entrywise product).

h. In previous classes, you may have learned about Fourier sums involving sines and cosines. This is related to the above by another simple (unitary) change of basis: if $\ell \neq n/2$ then

$$\begin{pmatrix} \chi_\ell & \chi_{-\ell} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1/i & -1/i \end{pmatrix} \frac{1}{\sqrt{2}} = \begin{pmatrix} c_\ell & s_\ell \end{pmatrix}$$

where c_ℓ and s_ℓ are sine and cosine functions:

$$c_\ell(x) = \sqrt{\frac{2}{n}} \cos\left(\frac{2\pi \ell x}{n} \right), \quad s_\ell(x) = \sqrt{\frac{2}{n}} \sin\left(\frac{2\pi \ell x}{n} \right).$$

If $\ell = n/2$ then χ_ℓ is itself a cosine function,

$$\chi_\ell(x) = \frac{\exp(\pi i x)}{\sqrt{n}} = \frac{(-1)^x}{\sqrt{n}} = \frac{\cos(\pi k)}{\sqrt{n}} .$$

Summary of the above: on the hermitian space $L^2(\Omega)$ for $\Omega = \mathbb{Z}/n\mathbb{Z}$, the translation operators T_z ($z \in \Omega$) are simultaneously diagonal by an orthonormal basis, which is the Fourier basis ($\chi_\ell : \ell \in \Omega$). Any convolution operator $C_g : f \mapsto f * g$ is a linear combination of translation operators, so it is also diagonalizable by the Fourier basis. This gives an explanation, using only linear algebra, as to why the Fourier transform behaves so nicely together with convolution.

13. (10/23) Fourier inversion for probability measures on the real line

1. A brief review of the basic theory for the Fourier transform of **functions** on the real line: if $f \in L^1(\mathbb{R})$ then we can define its Fourier transform by the integral formula

$$f\hat{}(t) = \int e^{itx} f(t)\, dt .$$

You can use Jensen's inequality to show that $\|f\hat{}\|_\infty \leq \|f\|_1$. A key result is the following: if $f, h \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, then $f\hat{}, g\hat{} \in L^\infty(\mathbb{R}) \cap L^2(\mathbb{R})$, and

$$\int_{\mathbb{R}} f(x) \overline{h(x)}\, dx = \langle f, h \rangle = \frac{\langle f\hat{}, h\hat{} \rangle}{2\pi} = \frac{1}{2\pi} \int f\hat{}(t) \overline{h\hat{}(t)}\, dt . \qquad (6)$$

This can be used to show that the mapping

$$U : L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \to L^\infty(\mathbb{R}) \cap L^2(\mathbb{R}), \quad f \mapsto Uf \equiv \frac{f^\wedge}{\sqrt{2\pi}}$$

has a unique continuous extension to a **unitary isometry** $U : L^2(\mathbb{R}) \to L^2(\mathbb{R})$. The $L^2(\mathbb{R})$ **Fourier transform** refers to the mapping from $f \in L^2(\mathbb{R})$ to $f^\wedge = \sqrt{2\pi}Uf \in L^2(\mathbb{R})$. The identity (6) holds for all $f, h \in L^2(\mathbb{R})$. The "conjugate transpose" is the map

$$U^* : L^2(\mathbb{R}) \to L^2(\mathbb{R}), \quad (U^*f)(t) = \frac{f^\wedge(-t)}{\sqrt{2\pi}}.$$

For any $f \in L^2(\mathbb{R})$ we have the **Fourier inversion formula**

$$f(t) = (U^*Uf)(t) = \frac{(U^*f^\wedge)(t)}{\sqrt{2\pi}} = \frac{(f^\wedge)^\wedge(-t)}{2\pi} = \frac{1}{2\pi}\int e^{-itx} f^\wedge(t)\, dt.$$

For details see [LL01, Ch. 5].

2. If μ is a probability measure on $\mathbb{R}$ with a density f (with respect to Lebesgue measure), then $f \in L^1(\mathbb{R})$. However, f need not be in $L^2(\mathbb{R})$, and moreover we want to consider the general case that μ may not have a density at all. For this we have:

Theorem 7 (Fourier inversion theorem for probability measures on $\mathbb{R}$). *If μ is a probability measure on $(\mathbb{R}, \mathscr{B})$ with Fourier transform (characteristic function) φ, then for all $-\infty < a < b < \infty$*

$$\left[I_T \equiv \frac{1}{2\pi}\int_{-T}^{T} \varphi(t)\, \frac{e^{-ita} - e^{-itb}}{2\pi it}\, dt\right] \overset{T\to\infty}{\longrightarrow} \tilde{\mu}(a,b) \equiv \mu((a,b)) + \frac{\mu(\{a,b\})}{2}.$$

In particular, φ uniquely determines $\tilde{\mu}$, which in turn uniquely determines μ.

Intuition for the theorem: note that if $h(t) \equiv \mathbf{1}\{t \in (a,b)\}$ and $k_T(t) \equiv \mathbf{1}\{|t| \le T\}$ then

$$I_T = \frac{1}{2\pi}\int \varphi(t)\, \overline{h^\wedge(t) k_T(t)}\, dt$$

In view of (6) it is natural to expect (although it does not directly follow) that

$$I_T = \int h_T(x)\, d\mu(x) \tag{7}$$

where h_T is the function such that $(h_T)^\wedge(t) = h^\wedge(t)k_T(t)$. The key steps in the proof of Theorem 7 are to show the identity (7), and to show that

$$\lim_{T\to\infty} h_T(t) = \tilde{h}(t) = \mathbf{1}\{t \in (a,b)\} + \frac{\mathbf{1}\{a \in \{a,b\}\}}{2}.$$

The h_T are bounded uniformly in T, and the result follows.

Reading: [Dur19, §3.3.1]

14. (10/28) Weak convergence and the central limit theorem

1. Let S be a **metric space** with Borel σ-algebra $\mathscr{B}$. If μ, μ_n are probability measures on $(S, \mathscr{B})$, we say that μ_n **converges weakly** to μ (denoted $\mu_n \Rightarrow \mu$) if

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

for all bounded continuous $f : S \to \mathbb{R}$. If X_n, X are random variables taking values in a metric space S, we say X_n **converges in distribution** (or **converges in law**) to X if and only if $\mathcal{L}_{X_n} \Rightarrow \mathcal{L}_X$.

2. Any probability measure μ on $(S, \mathscr{B})$ is **regular**: for all $A \in \mathscr{B}$,

$$\mu(A) = \sup\left\{\mu(F) : F \text{ closed}, F \subseteq A\right\} = \inf\left\{\mu(G) : G \text{ open}, A \subseteq G\right\}.$$

This implies that μ is completely determined by $\{\mu(F) : F \text{ closed}\}$. This can further be used to show that μ is completely determined by the values of

$$\int f \, d\mu$$

for bounded continuous f. This shows that a sequence μ_n cannot converge weakly to two different limits.

3. Let $\mathscr{P}$ be the space of probability measures on $(S, \mathscr{B})$. Let $\mathscr{T}$ be the topology on $\mathscr{P}$ generated by the sets

$$\left\{\nu : \left| \int f_i \, d\nu - \int f_i \, d\mu \right| < \epsilon \text{ for all } 1 \le i \le k \right\}$$

where f_i are bounded continuous functions. Then weak convergence is equivalent to convergence in the topology $\mathscr{T}$. If S is a **complete separable metric space** (**Polish space**), then $\mathscr{T}$ is a **metric topology**, and $\mathscr{P}$ is also a **Polish space**.

4. We say that a family of probability measures $\{\mu_\alpha : \alpha \in I\}$ is **tight** if for all $\epsilon > 0$ there exists a compact subset $K \subseteq S$ (depending on ϵ only) such that

$$\inf \left\{\mu_\alpha(K) : \alpha \in I\right\} \ge 1 - \epsilon .$$

Theorem 8 (Prohorov's theorem). *If $\{\mu_\alpha : \alpha \in I\}$ is tight, then it is relatively compact (has compact closure) in $\mathscr{P}$. (If S is a Polish space, then the converse also holds, but this is the less useful direction.)*

The case $S = \mathbb{R}$ is easier to prove and is called **Helly's selection theorem**.

5. A common strategy for showing weak convergence of μ_n: first use Theorem 8 to show $\{\mu_n\}_{n \ge 1}$ is confined within a compact subset of $\mathscr{P}$, and then show that all subsequential limits coincide. One application of this strategy is a general characterization of weak convergence via characteristic function convergence:

Theorem 9 (continuity theorem). *Let μ_n be probability measures on $\mathbb{R}$ with characteristic functions φ_n.*
a. If $\mu_n \Rightarrow \mu$ then $\varphi_n \to \varphi$ pointwise where φ is the characteristic function of μ.
b. If φ_n converges pointwise to a function φ that is **continuous at** $t = 0$, *then φ is the characteristic function of a probability measure μ, and $\mu_n \Rightarrow \mu$.*

We can apply Theorem 9 to prove the central limit theorem for i.i.d. sequences, and more generally the **Lindeberg–Feller central limit theorem** (for triangular arrays).

Reading: [Dur19, §3.2, §3.3.2-3, §3.4.1-2]. Optional reading: [Bil99, Ch. 1].

15. (10/30) Weak convergence: some more examples and theory

1. Let π be a random permutation of $[n]$, and let S_n be the number of disjoint cycles in π. In the limit $n \to \infty$,

$$\frac{S_n - \log n}{\sqrt{\log n}} \xrightarrow{d} Z$$

where Z is a standard gaussian random variable. Let $C_{n,k}$ be the number of cycles in π of length k, so that

$$S_n = \sum_{k \ge 1} C_{n,k} .$$

Then $(C_{n,k})_{k \ge 1}$ converges in law to $(Y_k)_{k \ge 1}$ where Y_k are independent $\mathsf{Pois}(1/k)$ random variables (see [AT92] and references therein — this result is slightly beyond the scope of this class; however, you can easily calculate that $\mathbb{E}C_{n,k} = 1/k$).

2. If $S_n \sim \mathsf{Bin}(n, \lambda/n)$ then $S_n \to_d Y_\lambda \sim \mathsf{Pois}(\lambda)$ as $n \to \infty$. As $\lambda \to \infty$ we have

$$\frac{Y_\lambda - \lambda}{\sqrt{\lambda}} \xrightarrow{d} Z$$

where Z is a standard gaussian random variable.

3. Suppose we have i.i.d. Bernoulli trials $I_k \sim \mathsf{Ber}(p)$ for $k \ge 1$. The number of successes by time m is

$$B_m = \sum_{k=1}^m I_k \sim \mathsf{Bin}(m, p) , \quad \mathbb{P}(B_m = \ell) = \binom{m}{\ell} p^\ell (1-p)^{m-\ell} .$$

The time of the first success is

$$G = \min\left\{k : I_k = 1\right\} \sim \mathsf{Geo}(p)\,, \quad \mathbb{P}(G = k) = (1-p)^{k-1}p \text{ for } k \in \{1,2,\ldots\}\,.$$

Let $G_1, G_2, \ldots$ be i.i.d. copies of G. The time of the r-th success is

$$X_r \stackrel{d}{=} \sum_{i=1}^r G_i \sim \mathsf{NegBin}(r,p)\,, \quad \mathbb{P}(X_r = t) = \binom{t-1}{r-1}(1-p)^{t-r}p^r\,.$$

4. Now take $p = 1/n$ and scale time by n, so $\mathsf{Ber}(1/n)$ trials happen at times $1/n, 2/n, \ldots$. The number of successes by time t is now

$$B_{nt} \sim \mathsf{Bin}\left(nt, \frac{1}{n}\right) \xrightarrow{d} \mathsf{Pois}(t)\,.$$

The time of the first success is

$$\frac{G}{n} \sim \frac{\mathsf{Geo}(1/n)}{n} \xrightarrow{d} E \sim \mathsf{Exp}\,, \quad \mathbb{P}(E \in [a,b]) = \int_a^b \frac{\mathbf{1}\{t \ge 0\}\,dt}{e^t}\,.$$

The time of the r-th success is

$$\frac{1}{n}\sum_{i=1}^r G_i \sim \frac{\mathsf{NegBin}(r,1/n)}{n} \xrightarrow{d} Y = \sum_{i=1}^r E_i \sim \mathsf{Gamma}(r) = \mathsf{Exp}^{*r}\,.$$

The gamma density can be derived by directly taking limits from the negative binomial distribution:

$$\mathbb{P}(Y \in [a,b]) = \int_a^b \mathbf{1}\{x \ge 0\}\frac{e^{-x}x^{r-1}}{(r-1)!}\,dx\,.$$

More generally, for any $\alpha > 0$, we write $\Gamma(\alpha)$ for the probability measure on $\mathbb{R}$ with density

$$f(x) = \mathbf{1}\{x \ge 0\}\frac{e^{-x}x^{\alpha-1}}{\Gamma(\alpha)}\,, \quad \Gamma(\alpha) \equiv \int_0^\infty e^{-x}x^{\alpha-1}\,dx\,.$$

5. Weak convergence in general (separable complete) metric spaces: you are expected to know the statements of the **portmanteau theorem**, the **Prohorov theorem**, and the **Skorohod representation theorem**. You should be able to prove all these theorems in the case that the underlying measure space is $(\mathbb{R}, \mathscr{B})$. (On the real line, the Prohorov theorem is usually called the **Helly selection theorem**; and the Skorohod representation theorem is a consequence of the **probability integral transform**.)

Reading: [Dur19, Ch. 3 up to and including §3.6 (except sections marked ∗)].

16. (11/04) Conclusion of weak convergence; introduction to conditional expectation

1. For Z a standard gaussian random variable, define the complementary cdf

$$\Psi(x) = \mathbb{P}(Z \ge x) = \int_x^\infty g(z)\,dz = \int_x^\infty \frac{1}{\sqrt{2\pi}\exp(z^2/2)}\,dz\,,$$

where we use g to denote the standard gaussian density. We saw the following estimate for all $x > 0$:

$$\frac{g(x)}{x}\left[1 - \frac{1}{x^2}\right] \le \Psi(x) \le \frac{g(x)}{x} \tag{8}$$

2. Two examples of extremal statistics: if X_i are i.i.d. standard exponential random variables, then

$$\left(\max_{1\le i\le n} X_i\right) - \log n \xrightarrow{d} \mathsf{Gumbel}$$

where the **Gumbel distribution** is supported on the real line with cdf $F(x) = \exp(-\exp(-x))$. If Z_i are i.i.d. standard gaussian random variables, then (8) can be used to show that

$$b_n\left[\left(\max_{1\le i\le n} Z_i\right) - b_n\right] \xrightarrow{d} \mathsf{Gumbel}$$

for b_n defined by the equation $\Psi(b_n) = 1/n$. We can also use (8) to estimate

$$b_n = \sqrt{2\log n} - \frac{\log\log n + O(1)}{\sqrt{2\log n}}\,.$$

3. We discussed weak convergence for probability measures on $\mathbb{R}^d$. This is a special case of our general discussion of weak convergence for probability measures on metric spaces. Read [Dur19] for the details: pay special attention to the $\mathbb{R}^d$ **Fourier inversion formula** and **continuity theorem**, the **Cramér–Wold theorem**, and the **CLT for i.i.d. sequences on** $\mathbb{R}^d$.
4. We introduced the topic of conditional expectation. On a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, if $A \in \mathscr{F}$ with $\mathbb{P}(A) > 0$, and X is an integrable random variable, then we can define

$$\mathbb{E}(X \mid A) = \frac{\mathbb{E}(X; A)}{\mathbb{P}(A)}\,.$$

In particular, if Y is another random variable with $\mathbb{P}(Y = y) > 0$, then we can define

$$\mathbb{E}(X \mid Y = y) = \frac{\mathbb{E}(X; Y = y)}{\mathbb{P}(Y = y)} = h(y)\,.$$

If Y has *countable* support, then we can define $\mathbb{E}(X \mid Y) = h(Y)$. The goal of the following lecture is to define $\mathbb{E}(X \mid Y)$ in a general setting.

Reading: [Dur19, §3.10, §4.1].

17. (11/06) Conditional expectation; introduction to martingales

1. We covered the formal definition and basic properties of **conditional expectation** — read [Dur19, §4.1].
2. We introduced the formal definitions of **filtration**, **martingale**, **submartingale**, **supermartingale**, and **stopping time**. If M_n is a martingale then $\mathbb{E}M_0 = \mathbb{E}M_n$ for all n. In upcoming lectures we will prove that $\mathbb{E}M_0 = \mathbb{E}M_\tau$ for stopping times τ, **under suitable conditions on** M **and** τ. Results of this kind are called **optional stopping theorems** (OSTs).
3. Conditions **are** necessary for OSTs to hold: if M_n is simple random walk on $\mathbb{Z}$ started from $M_0 = 0$, and $\tau = \min\{n : M_n = 1\}$, then M_n is a martingale and τ is a stopping time, but $M_0 = 0$ and $M_\tau = 1$ almost surely so $\mathbb{E}M_0 \neq \mathbb{E}M_\tau$.
4. Why we expect OST to hold in certain circumstances: note that if M_n is a martingale and τ is a stopping time, then the **stopped process** $X_n = M_{n\wedge\tau}$ is also a martingale: we can decompose

$$X_n = M_{n\wedge\tau} = \underbrace{\sum_{i=1}^{n-1} \mathbf{1}\{\tau = i\} M_i}_{\text{in } \mathscr{F}_{n-1}} + \overbrace{\mathbf{1}\{\tau \ge n\}}^{\text{in } \mathscr{F}_{n-1}}\, M_n\,,$$

and then the martingale property of M_n implies

$$\mathbb{E}(X_n \mid \mathscr{F}_{n-1}) = \sum_{i=1}^{n-1} \mathbf{1}\{\tau = i\} M_i + \mathbf{1}\{\tau \ge n\} M_{n-1} = X_{n-1}\,.$$

Therefore $\mathbb{E}X_0 = \mathbb{E}X_n$ for all n, which means $\mathbb{E}M_0 = \mathbb{E}M_{n\wedge\tau}$ for all n. In the limit $n \to \infty$ we have $n \wedge \tau \to n$ a.s., and $M_{n\wedge\tau} \to M_\tau$ a.s.. Based on this, we expect under some integrability conditions that $\mathbb{E}M_{n\wedge\tau} \to \mathbb{E}M_\tau$, which would imply an OST.
5. An example that we will discuss in more detail in a later lecture: let $G = (V, E)$ be a finite graph with boundary $\varnothing \subsetneq Z \subsetneq V$. Suppose we have some **Dirichlet boundary condition** $f : Z \to \mathbb{R}$. Let X_n be simple random walk on G started at $X_0 \in V \setminus Z$, and let $\tau = \min\{n : X_n \in Z\}$. Then

$$h : V \to \mathbb{R}\,, \quad h(x) = \mathbb{E}\Big[f(X_\tau) \Big| X_0 = x \Big]$$

is the **harmonic interpolation** of f to V. The process $M_n = f(X_{n\wedge\tau})$ is a martingale. More generally, if $h : V \to \mathbb{R}$ is **subharmonic** on $V \setminus Z$, then the process $Y_n = f(X_{n\wedge\tau})$ is a **submartingale**.

Reading: [Dur19, §4.2]. Note that there is no lecture on Monday 11/11 (university holiday).

18. (11/13) Upcrossing inequality, submartingale convergence theorem

1. The key technical result of this lecture is the following bound which controls $U_n(a,b)$, the number of upcrossings of an interval $[a,b]$ by a submartingale X_k over the time interval $0 \le k \le n$:

   Lemma 10 (upcrossing inequality). *If X_n is a submartingale, then*
   $$\mathbb{E}U_n(a,b) \le \frac{\mathbb{E}[(X_n-a)_+ - (X_n-b)_+]}{b-a} \le \frac{\mathbb{E}[(X_n-a)_+]}{b-a} \le \frac{|a| + \mathbb{E}[(X_n)_+]}{b-a}$$
   for all $-\infty < a < b < \infty$.

2. The main consequence of Lemma 10 is the following:

   Theorem 11 (submartingale convergence theorem). *If X_n is a submartingale with $\sup_n \mathbb{E}[(X_n)_+] < \infty$, then X_n converges almost surely to a limit X_∞ with $\mathbb{E}|X_\infty| < \infty$.*

3. An important special case of Theorem 11:

   Theorem 12 (nonnegative supermartingale convergence theorem). *If X_n is a nonnegative supermartingale, then X_n converges almost surely to a limit X_∞ with $\mathbb{E}X_\infty \le \liminf_{n\to\infty} \mathbb{E}X_n \le \mathbb{E}X_0$.*

Reading: [Dur19, §4.2].

19. (11/18) L^p martingale convergence theorem, branching processes example

1. Note that neither Theorem 11 nor Theorem 12 says anything about whether $\mathbb{E}X_n$ converges to $\mathbb{E}X_\infty$. We are often interested in this question, especially for the case that $X_n = M_{n\wedge\tau}$ where M_n is a martingale and τ is a finite stopping time. As we discussed in Lecture 17, we have $\mathbb{E}M_0 = \mathbb{E}M_{n\wedge\tau}$ and $X_n = M_{n\wedge\tau} \to M_\tau = X_\infty$, so *if* $\mathbb{E}X_n \to \mathbb{E}X_\infty$ then in this case we *would* obtain the optional stopping identity $\mathbb{E}M_\tau = \mathbb{E}M_0$. We also saw in Lecture 17 examples showing that such an identity does not hold in full generality, so conditions are needed. The main goal of this lecture is to give a simple *sufficient condition* for a martingale X_n to satisfy $\mathbb{E}X_n \to \mathbb{E}X_\infty$, in the form of an L^p criterion for $p \in (1,\infty)$ (Theorem 15 below). In the next lecture we will see a *necessary and sufficient condition* in the form of a *uniform integrability* criterion (Theorem 19).
2. The main technical result of this lecture is the following:

   Lemma 13 (maximal inequality). *Let X_n be a submartingale and $\bar{X}_n = \max\{(X_i)_+ : 0 \le i \le n\}$. Then*
   $$\lambda\mathbb{P}\Big(\max_{0\le i\le n} X_i \ge \lambda\Big) = \lambda\mathbb{P}\Big(\bar{X}_n \ge \lambda\Big) \le \mathbb{E}\Big(X_n; \bar{X}_n \ge \lambda\Big) \le \mathbb{E}\Big((X_n)_+; \bar{X}_n \ge \lambda\Big) \le \mathbb{E}\Big((X_n)_+\Big)$$
   for all $\lambda \ge 0$.

   The key inequality in Lemma 13 is the first one, compared with the trivial bound $\lambda\mathbb{P}(\bar{X}_n \ge \lambda) \le \mathbb{E}(\bar{X}_n; \bar{X}_n \ge \lambda)$ implied by Markov's inequality. Integrating Lemma 13 gives:

   Lemma 14 (L^p maximal inequality). *If X_n is a submartingale and $\bar{X}_n = \max\{(X_i)_+ : 0 \le i \le n\}$, then for all $p \in (1,\infty)$ we have*
   $$\|\bar{X}_n\|_p \le \frac{p}{p-1}\|(X_n)_+\|_p\,.$$
   (If $\|(X_n)_+\|_p = \infty$ the bound is vacuous.)

3. The following is a straightforward consequence of Theorem 11 combined with Lemma 14:

   Theorem 15 (L^p martingale convergence theorem). *If X_n is a martingale with $\sup_n \|X_n\|_p < \infty$,*

4. Application to branching processes: let $\xi, \xi_{n,i}$ be i.i.d. nonnegative integer random variables with law p. Define the filtration $\mathscr{F}_n = \sigma(\xi_{\ell,i} : i \ge 1, 1 \le \ell \le n)$. The **Galton–Watson (GW) process with offspring law** p is defined by $(Z_n)_{n\ge0}$ where $Z_0 = 1$ and
   $$Z_n = \sum_{i=1}^{Z_{n-1}} \xi_{n,i}\,.$$
   In particular, if $Z_n = 0$ then $Z_\ell = 0$ for all $\ell \ge n$, so we can define the **extinction event**
   $$\text{extinction} = \{Z_n = 0 \text{ ev.}\} = \bigcup_{n\ge0}\{Z_n = 0\}\,.$$

If the offspring law has finite mean $\mathbb{E}\xi = \mu \in (0,\infty)$, then $M_n = Z_n/\mu^n$ is a martingale with respect to $\mathscr{F}_n$. It follows from Theorem 12 that $M_n \to M_\infty$ almost surely, with $\mathbb{E}M_\infty \le \mathbb{E}M_n = 1$. We discussed three cases:

a. For $\mu \in (0,1)$ (subcritical GW) it follows from Markov's inequality that $\mathbb{P}(Z_n > 0) \le \mathbb{E}Z_n = \mu^n \to 0$, so the process Z_n goes extinct almost surely. It follows that $M_\infty = 0$ a.s., so $\mathbb{E}M_\infty = 0$.
b. For $\mu = 1$ (critical GW), outside of the trivial case $\xi = 1$ a.s., we used Theorem 12 to argue that the process $Z_n = M_n$ goes extinct almost surely. It follows that $M_\infty = 0$ a.s., so $\mathbb{E}M_\infty = 0$.
c. For $\mu > 1$ (supercritical GW), we argued that $\mathbb{P}(\text{extinction}) = \lim_{n\to\infty} \psi^{\circ n}(0)$ where
$$\psi(s) = \mathbb{E}(s^\xi) = \sum_{k\ge 0} p_k s^k$$
and $\psi^{\circ n}$ denotes the n-fold composition of ψ. It follows that $\mathbb{P}(\text{extinction}) = \rho$, the unique fixed point of ψ in the interval $[0,1)$. On the event of extinction, clearly $M_\infty = 0$. The behavior of M_∞ on the event of nonextinction is characterized by the **Kesten–Stigum "$L \log L$ criterion,"** which is discussed in more detail in Assignment 6 and Lecture 24.

Reading: [Dur19, §4.3–4.4].

20. (11/20) Uniform integrability and L^1 convergence; Doob martingales

1. The L^p martingale convergence theorem gives a simple sufficient condition for a martingale X_n to converge to a limit X_∞ both almost surely and in L^1. In this lecture we see a *necessary and sufficient condition*, given by Theorem 19 below.
2. First we covered the definition of **uniform integrability** (u.i.) for a family $(X_i)_{i\le I}$ of random variables on $(\Omega,\mathscr{F},\mathbb{P})$. Observation: if $(X_i)_{i\le I}$ is u.i. then $\sup_{i\in I}\mathbb{E}|X_i| < \infty$.

   Lemma 16. *If $X \in L^1(\Omega,\mathscr{F},\mathbb{P})$ then the family $\{\mathbb{E}(X\,|\,\mathscr{G}) : \mathscr{G} \text{ is a sub-}\sigma\text{-field of } \mathscr{F}\}$ is u.i.*

3. Main theorem relating uniform integrability to L^1 convergence:

   Theorem 17. *Under the assumption that $X_n \to X$ in probability, the following are equivalent: (a) $\{X_n\}_{n\ge 0}$ is u.i.; (b) $X_n \to X$ in L^1; (c) $\mathbb{E}|X_n| \to \mathbb{E}|X|$.*

4. Consequences of Theorem 17:

   Theorem 18 (submartingale L^1 convergence theorem)**.** *For a submartingale X_n, the following are equivalent: (a) u.i.; (b) convergence almost surely and in L^1; (c) convergence in L^1.*

   Theorem 19 (martingale L^1 convergence theorem)**.** *For a martingale X_n, the following are equivalent: (a) u.i.; (b) convergence almost surely and in L^1; (c) convergence in L^1; (d) existence of $X \in L^1$ such that $X_n = \mathbb{E}(X\,|\,\mathscr{F}_n)$.*

5. A sequence $X_n = \mathbb{E}(X\,|\,\mathscr{F}_n)$ (for $X \in L^1$) is called a Doob martingale. Identification of the limit:

   Theorem 20 (Lévy convergence theorem)**.** *If $\mathscr{F}_n$ is a filtration and $\mathscr{F}_\infty$ is the σ-field generated by their union, then for any $X \in L^1$ we have $\mathbb{E}(X\,|\,\mathscr{F}_n) \to \mathbb{E}(X\,|\,\mathscr{F}_\infty)$ both a.s. and in L^1.*

Reading: [Dur19, §4.6].

21. (11/25) Optional stopping theorems

1. In this lecture we saw optional stopping theorems (OST) for (sub/super)martingales. The OST for nonnegative supermartingales is especially easy, and does not involve uniform integrability:

   Theorem 21 (nonnegative supermartingale OST)**.** *If X_n is a nonnegative supermartingale and τ is any stopping time, then $\mathbb{E}X_0 \ge \mathbb{E}X_\tau$.*

2. Optional stopping theorems for submartingales with uniform integrability conditions:

   Lemma 22 (stopping preserves u.i.)**.** *If X_n is a u.i. submartingale and τ is any stopping time, then $Y_n = X_{n\wedge\tau}$ is also a u.i. submartingale.*

   Theorem 23 (u.i. submartingale OST)**.**
   a. If X_n is a u.i. submartingale and τ is any stopping time, then $\mathbb{E}X_0 \le \mathbb{E}X_\tau \le \mathbb{E}X_\infty$.
   b. If $X_{n\wedge\tau}$ is a u.i. submartingale, then $\mathbb{E}X_0 \le \mathbb{E}X_\tau$.

(Lemma 22 shows that the conditions of (a) are stronger than those of (b).)

The following is less general than Theorem 23, but has more concrete conditions that are sometimes easier to check (see "ABRACADABRA" example below):

Theorem 24. *Suppose X_n is a submartingale with $\mathbb{E}[|X_{n+1} - X_n| \,|\, \mathscr{F}_n] \le B$ a.s. for all n, and τ is a stopping time with $\mathbb{E}\tau < \infty$. Then $X_{n\wedge\tau}$ is u.i., therefore (by Theorem 23b) $\mathbb{E}X_0 \le \mathbb{E}X_\tau$.*

3. Suppose on $(\Omega, \mathscr{F}, \mathbb{P})$ that we have a filtration $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots \subseteq \mathscr{F}$, and that τ is a stopping time with respect to $\mathscr{F}_n$. We define the associated **stopping time σ-algebra**

$$\mathscr{F}_\tau \equiv \left\{ A \in \mathscr{F} : A \cap \{\tau = n\} \in \mathscr{F}_n \text{ for all } n \right\}. \tag{9}$$

The reason for defining $\mathscr{F}_\tau$ in this way is the following:

Lemma 25. *Suppose on $(\Omega, \mathscr{F}, \mathbb{P})$ that we have a filtration $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots \subseteq \mathscr{F}$, and that τ is a stopping time with respect to $\mathscr{F}_n$. Then $\mathscr{F}_\tau$, as defined by (9), is the minimal σ-algebra such that for any process $(Y_n)_{n\ge0}$ adapted to $(\mathscr{F}_n)_{n\ge0}$ we will have $Y_\tau \in \mathscr{F}_\tau$.* (Check as an exercise that you can prove this lemma!)

We then saw a generalization of Theorem 23b with two stopping times:

Theorem 26. *If σ, τ are stopping times (with respect to $\mathscr{F}_n$) with $\sigma \le \tau$ almost surely, and $X_{n\wedge\tau}$ is a u.i. submartingale, then $X_\sigma \le \mathbb{E}(X_\tau \,|\, \mathscr{F}_\sigma)$ a.s.*

4. Examples:
   a. Analysis of biased simple random walk on $\mathbb{Z}$.
   b. Calculation of expected time to see a given pattern ('ABRACADABRA') in a random string of letters.

Reading: [Dur19, §4.8].

22. (11/27) Reverse martingales; Kolmogorov and Hewitt–Savage zero-one laws

1. A **reverse martingale** (or **backward martingale**) is a process $(M_n)_{n\ge0}$ where $M_n \in L^1(\mathscr{F}_n)$ with $\mathscr{F} \supseteq \mathscr{F}_0 \supseteq \mathscr{F}_1 \supseteq \ldots$, and $M_n = \mathbb{E}(M_{n-1} \,|\, \mathscr{F}_n) = \mathbb{E}(M_0 \,|\, \mathscr{F}_n)$. We write $\mathscr{F}_n \downarrow \mathscr{F}_\infty$ where $\mathscr{F}_\infty$ is the intersection of all the $\mathscr{F}_n$. We saw that the upcrossing inequality (Lemma 10) and the characterization of uniform integrability (Theorem 17) can be combined to prove:

Theorem 27. *If $(M_n)_{n\ge0}$ is a reverse martingale with respect to $\mathscr{F}_n \downarrow \mathscr{F}_\infty$, then $M_n \to M_\infty = \mathbb{E}(M_0 \,|\, \mathscr{F}_\infty)$, both almost surely and in L^1.*

Corollary 28. *If $Y \in L^1(\mathscr{F})$ and $\mathscr{F}_n \downarrow \mathscr{F}_\infty$ then $\mathbb{E}(Y \,|\, \mathscr{F}_n) \to \mathbb{E}(Y \,|\, \mathscr{F}_\infty)$, both almost surely and in L^1.*

2. Two proofs of the Kolmogorov zero-one law: [Dur19, Thm. 2.5.3; and comments following Thm. 4.6.9]
3. Two proofs of the Hewitt–Savage zero-one law [Dur19, Thm. 2.5.4; Example 4.7.6].
4. Proof of the strong law of large number using reverse martingale [Dur19, Example 4.7.4].

Reading: [Dur19, Thm. 2.5.3 and Thm. 2.5.4; §4.7].

23. (12/02) Martingale perspective on Radon–Nikodym derivatives

1. Review of basic definitions: let μ, ν be finite measures on $(\Omega, \mathscr{F})$.
   a. We say that ν **is absolutely continuous with respect to** μ, denoted $\nu \ll \mu$, if for all $A \in \mathscr{F}$ such that $\mu(A) = 0$ we have $\nu(A) = 0$ also.
   b. If there is a measurable function $f : \Omega \to [0, \infty)$ such that

$$\nu(A) = \int_A f \, d\mu \tag{10}$$

   for all $A \in \mathscr{F}$, then we call f the **Radon–Nikodym derivative** of ν with respect to μ, denoted $f = d\nu/d\mu$. Note that f is unique μ-a.e.
   c. If a Radon–Nikodym derivative $d\nu/d\mu$ exists, then $\nu \ll \mu$. The converse is given by the following:

Theorem 29 (Radon–Nikodym theorem)**.** *If μ, ν are finite measures on $(\Omega, \mathscr{F})$ with $\nu \ll \mu$, then there exists a function $f = d\nu/d\mu : \Omega \to [0, \infty)$ satisfying (10) for all $A \in \mathscr{F}$.*

Basic properties of the Radon–Nikodym derivative: if $\pi \ll \nu \ll \mu$ then

$$\frac{d\pi}{d\mu} = \frac{d\pi}{d\nu}\frac{d\nu}{d\mu}\,.$$

If $\nu \ll \mu$ and $\mu \ll \nu$ then $d\mu/d\nu = 1/[d\nu/d\mu]$.

d. In Lecture 17 we stated Theorem 29 without proof, and used it to construct the conditional expectation: if $X \in L^1(\Omega, \mathscr{F}, \mathbb{P})$ and $\mathscr{G} \subseteq \mathscr{F}$, then

$$\mathbb{E}(X\,|\,\mathscr{G}) = \frac{d\nu_+}{d\mu} - \frac{d\nu_-}{d\mu}$$

where $\mu = \mathbb{P}|_{\mathscr{G}}$, and $\nu_\pm$ are the measures on $(\Omega, \mathscr{G})$ defined by setting $\nu_\pm(A) = \mathbb{E}(X_\pm; A)$ for all $A \in \mathscr{G}$.

2. The proof of Theorem 29 is given in [Dur19, §A.4] and is slightly outside the scope of this class. However, there is one special case where the theorem is very easy to prove: suppose Ω has a **countable partition**

$$\Omega = \bigsqcup_{i=1}^{\infty} \Omega_i$$

such that $\mathscr{F} = \sigma(\Omega_1, \Omega_2, \ldots)$. Suppose μ, ν are finite measures on $(\Omega, \mathscr{F})$ with $\nu \ll \mu$. Define

$$f(\omega) = \begin{cases} \nu(\Omega_i)/\mu(\Omega_i) & \text{if } \omega \in \Omega_i \text{ with } \mu(\Omega_i) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Then $f = d\nu/d\mu$ (you should check this). This proves Theorem 29 in the special case that $\mathscr{F}$ is generated by a countable partition of Ω.

3. A more general situation is that there exists $\mathscr{F}_n \uparrow \mathscr{F}$ such that each $\mathscr{F}_n$ is generated by a countable partition of Ω. For instance, if $\Omega = \mathbb{R}$ and

$$\mathscr{F}_n = \sigma\left(\frac{[i, i+1)}{2^n} : i \in \mathbb{Z}\right),$$

then $\mathscr{F}_n \uparrow \mathscr{F} = \mathscr{B}_{\mathbb{R}}$. Suppose μ, ν are finite measures on $(\Omega, \mathscr{F})$ with $\nu \ll \mu$. Suppose $\mathscr{F}_n \uparrow \mathscr{F}$ and denote $\mu_n = \mu|_{\mathscr{F}_n}$ and $\nu_n = \nu|_{\mathscr{F}_n}$. Note that $\nu_n \ll \mu_n$ for all n, so $d\nu_n/d\mu_n$ can be defined by (11). It is then natural to ask whether we can obtain $d\nu/d\mu$ by taking the limit $n \to \infty$. The answer is yes, and in fact it is possible to prove a stronger statement:

Theorem 30. *Let μ, ν be finite measures on $(\Omega, \mathscr{F})$,* **not** *assuming $\nu \ll \mu$. Suppose $\mathscr{F}_n \uparrow \mathscr{F}$ and denote $\mu_n = \mu|_{\mathscr{F}_n}$ and $\nu_n = \nu|_{\mathscr{F}_n}$. Suppose for all n that $\nu_n \ll \mu_n$, with $X_n = d\nu_n/d\mu_n$. Let $X = \limsup_{n\to\infty} X_n$. Then*

$$\nu(A) = \underbrace{\int_A X\,d\mu}_{\nu_{\mathrm{cts}}(A)} + \underbrace{\nu\Big(A \cap \{X = \infty\}\Big)}_{\nu_{\mathrm{sing}}(A)}\,.$$

This decomposes $\nu = \nu_{\mathrm{cts}} + \nu_{\mathrm{sing}}$ where $\nu_{\mathrm{cts}} \ll \mu$ and $\nu_{\mathrm{sing}} \perp \mu$ (the measures ν_{sing} and μ are **mutually singular***).*

A key observation in the proof of Theorem 30 is that X_n is a martingale with respect to $\mathscr{F}_n$ on $(\Omega, \mathscr{F}, \mu)$. In particular, this implies (by Theorem 12) that X_n converges μ-a.s. to a finite limit X_∞. It follows that $\mu(X = \infty) = \mu(X_\infty = \infty) = 0$, which explains why ν_{sing} and μ are mutually singular.

Reading: [Dur19, §4.3.3]. **Lectures 24 and 25: two applications of ideas from the class.**

24. (12/09) Martingale-based proof of the Kesten–Stigum "$L \log L$ criterion"

This lecture is based on [LPP95] (see also [KS66b, KS66a, KS67]). See earlier discussion in Lecture 19. An important ingredient in the proof is Theorem 30 from Lecture 23.

25. (12/11) The $\zeta(2)$ limit in the random assignment problem

This lecture is based on [Ald92, Ald01] (see also [AS04]). Ingredients in the Aldous proof: Kolmogorov extension theorem (Theorem 5 from Lecture 7), local weak convergence (see homework), Poisson processes (Lecture 15). Later improvements on the result: [LW04, NPS05] (independent works, involving complicated inductions), and subsequently [Wä09] (simpler proof).

References

[Ald92] D. Aldous. Asymptotics in the random assignment problem. *Probab. Theory Related Fields*, 93(4):507–534, 1992.

[Ald01] D. J. Aldous. The $\zeta(2)$ limit in the random assignment problem. *Random Structures Algorithms*, 18(4):381–418, 2001.

[AS04] D. Aldous and J. M. Steele. The objective method: probabilistic combinatorial optimization and local weak convergence. In *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci.*, pages 1–72. Springer, Berlin, 2004.

[AT92] R. Arratia and S. Tavaré. The cycle structure of random permutations. *Ann. Probab.*, 20(3):1567–1591, 1992.

[Bil99] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[Dur19] R. Durrett. *Probability—theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. Fifth edition of [MR1068527].

[Kal02] O. Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[KS66a] H. Kesten and B. P. Stigum. Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *Ann. Math. Statist.*, 37:1463–1481, 1966.

[KS66b] H. Kesten and B. P. Stigum. A limit theorem for multidimensional Galton-Watson processes. *Ann. Math. Statist.*, 37:1211–1223, 1966.

[KS67] H. Kesten and B. P. Stigum. Limit theorems for decomposable multi-dimensional Galton-Watson processes. *J. Math. Anal. Appl.*, 17:309–338, 1967.

[LL01] E. H. Lieb and M. Loss. *Analysis*, volume 14 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2001.

[LPP95] R. Lyons, R. Pemantle, and Y. Peres. Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *Ann. Probab.*, 23(3):1125–1138, 1995.

[LW04] S. Linusson and J. Wästlund. A proof of Parisi's conjecture on the random assignment problem. *Probab. Theory Related Fields*, 128(3):419–440, 2004.

[NPS05] C. Nair, B. Prabhakar, and M. Sharma. Proofs of the Parisi and Coppersmith-Sorkin random assignment conjectures. *Random Structures Algorithms*, 27(4):413–444, 2005.

[Tao11] T. Tao. *An introduction to measure theory*, volume 126 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2011.

[W0̈9] J. Wästlund. An easy proof of the $\zeta(2)$ limit in the random assignment problem. *Electron. Commun. Probab.*, 14:261–269, 2009.