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) References 17

References 17
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 \( \mu : \mathcal{P}([0, 1]) \to [0, 1] \) satisfying (i) \( \mu([a, b]) = b - a \), (ii) \( \mu \) is countably additive, (iii) \( \mu \) is translation invariant. See [Dur19, §A.2] for a similar example.

3. Formal definition of a measure space \((\Omega, \mathcal{F}, \mu)\) (state space, \(\sigma\)-field or \(\sigma\)-algebra, measure). Some consequences of the definition: \( \{\emptyset, \Omega\} \in \mathcal{F} \); \( \mu(\emptyset) = 0 \); \( \mu \) is monotone (if \( A, B \in \mathcal{F} \) with \( A \subseteq B \) then \( \mu(A) \leq \mu(B) \)); continuity from below; continuity from above. Moreover \( \mu \) is countably subadditive over \( \mathcal{F} \): if \( A, A_i \in \mathcal{F} \) and \( A \) is contained in the countable union of the \( A_i \), then

\[
\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i).
\]

Check for yourself that you can prove all these properties!

4. The Borel \(\sigma\)-field over \(\mathbb{R}\), denoted \(\mathcal{B}_\mathbb{R}\), is the smallest \(\sigma\)-field over \(\mathbb{R}\) that contains the open intervals, \(\mathcal{I} = \{(a, b) : -\infty \leq a < b \leq \infty\}\). We say that \(\mathcal{B}_\mathbb{R}\) is the \(\sigma\)-field generated by \(\mathcal{I}\), denoted \(\mathcal{B}_\mathbb{R} = \sigma(\mathcal{I})\). Check for yourself that it is equivalent to \(\mathcal{B}_\mathbb{R} = \sigma(\mathcal{I})\) where

\[
\mathcal{I} = \left\{(a, b] \cap \mathbb{R} : -\infty \leq a < b \leq \infty\right\}.
\]

(Note that \(\mathcal{I}\) contains unbounded intervals of the form \((-\infty, b]\) and \((a, \infty)\).) Check that it is also equivalent to \(\mathcal{B}_\mathbb{R} = \sigma(\mathcal{I})\) where \(\mathcal{I}\) 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 \((\mathbb{R}, \mathcal{B}_\mathbb{R})\) satisfying

\[
\mu((a, b] \cap \mathbb{R}) = F(b) - F(a)
\]

for all \(-\infty \leq a \leq b \leq \infty\), where the values of \(F(\infty)\) and \(F(-\infty)\) are defined by continuity. In the case \(F(x) = x\), the corresponding \(\mu\) is called the Lebesgue measure on \((\mathbb{R}, \mathcal{B}_\mathbb{R})\).

2. Note that condition (1) defines \(\mu : \mathcal{I} \to [0, \infty]\). Let

\[
\mathcal{A} \equiv \left\{ A \subseteq \mathbb{R} : A \text{ is a disjoint union of finitely many elements of } \mathcal{I} \right\}.
\]

This is an algebra (closed under complementation and finite union); it is the smallest algebra containing \(\mathcal{I}\).

**Part I** of the proof of Theorem 1: there is a unique \(\mu : \mathcal{A} \to [0, \infty]\) which extends \(\mu : \mathcal{I} \to [0, \infty]\), and is countably additive over \(\mathcal{A}\).

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 \(\mu : \mathcal{B}_\mathbb{R} \to [0, \infty]\) which extends \(\mu : \mathcal{A} \to [0, \infty]\) and is a measure on \(\mathcal{B}_\mathbb{R}\) (i.e., is countably additive over \(\mathcal{B}_\mathbb{R}\)). In the case that \(\mu\) is a finite measure (\(\mu(\Omega) < \infty\)), this is a special case of a more general theorem:

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

2. Proof of uniqueness in Theorem 2: this is based on Dynkin’s \(\pi\)-\(\lambda\) 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

\[ \mu^*(A) = \inf \left\{ \sum_{i=1}^{\infty} \mu(E_i) : E_i = (a_i, b_i] \text{ and } A \subseteq \bigcup_{i=1}^{\infty} E_i \right\}, \]

where \( \mu(E_i) = F(b_i) - F(a_i) \). Then let

\[ \mathcal{F}_F \equiv \left\{ A \subseteq \Omega : \mu^*(S) = \mu^*(S \cap A) + \mu^*(S \setminus A) \text{ for all } S \subseteq \Omega \right\}. \]

We showed in class that \( \mathcal{F}_F \) is a \( \sigma \)-algebra, and the restriction of \( \mu^* \) to \( \mathcal{F}_F \) is a measure \( \mu = \mu_F \). In the case \( F(x) = x \), \( \mathcal{F}_F \) is the Lebesgue \( \sigma \)-algebra, commonly denoted \( \mathcal{L} \); and \( \mu_F \) is the Lebesgue measure, commonly denoted \( \lambda \) or Leb. Note \( \mathcal{L} \subseteq \mathcal{B} \), so \( \lambda \) further restricts to a measure on \((\mathbb{R}, \mathcal{B})\) which is also called Lebesgue measure.

4. The precise relation between \( \mathcal{L} \) and \( \mathcal{B} \) is as follows: any \( A \in \mathcal{L} \) can be expressed as \( A = B \cup N \) where \( B \in \mathcal{B} \) and \( N \subseteq B' \in \mathcal{B} \) 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 \((\Omega, \mathcal{F})\) be a measurable space. We say \( f \) is a simple function on \((\Omega, \mathcal{F})\) if \( f : \Omega \to \mathbb{R} \) can be expressed as

\[ f = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i} \]

for \( c_i \in \mathbb{R} \) and \( A_i \in \mathcal{F} \). A measurable function on \((\Omega, \mathcal{F})\) is any pointwise limit of simple functions: \( f = \lim_{n} f_n \) where \( f_n \) are simple functions on \((\Omega, \mathcal{F})\). A measurable mapping from \((\Omega, \mathcal{F})\) to \((S, \mathcal{G})\) is a map \( g : \Omega \to S \) such that for all \( B \in \mathcal{G} \), the preimage \( g^{-1}(B) \) belongs to \( \mathcal{F} \). To emphasize measurability of \( g \), we often write \( g : (\Omega, \mathcal{F}) \to (S, \mathcal{G}) \). On Homework 1: \( f \) is a measurable function on \((\Omega, \mathcal{F})\) if and only if it is a measurable mapping from \((\Omega, \mathcal{F})\) to \((\mathbb{R}, \mathcal{B})\).

2. If \( \mu \) is any measure on \((\Omega, \mathcal{F})\) and \( g : (\Omega, \mathcal{F}) \to (S, \mathcal{G}) \), we obtain a pushforward measure \( \nu \) on \((S, \mathcal{G})\),

\[ \nu(B) = \mu(g^{-1}(B)) = \mu \left( \left\{ \omega \in \Omega : g(\omega) \in B \right\} \right). \]

Some standard notations: \( \nu = g_* \mu = g_* \mu = \mu \circ g^{-1} \).

3. Let \((\Omega, \mathcal{F})\) be a \( \sigma \)-finite measure space. If \( f \) is a simple function on \((\Omega, \mathcal{F})\) as in (2), with the additional condition that \( \mu(A_i) < \infty \) for all \( i \), then we define its Lebesgue integral

\[ \int f \, d\mu = \sum_{i=1}^{n} c_i \mu(A_i). \]

If \( f \) is a nonnegative measurable function on \((\Omega, \mathcal{F})\), we define its Lebesgue integral

\[ \int f \, d\mu = \sup \left\{ \int h \, d\mu : 0 \leq h \leq f, h \text{ simple function with } \mu(h > 0) < \infty \right\}. \]

Finally, if \( f \) is a general measurable function on \((\Omega, \mathcal{F})\), we define its Lebesgue integral

\[ \int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu \]

with the caveat that \( \infty - \infty = 0 \). This completes the general definition of the Lebesgue integral

\[ \int f \, d\mu = \int f(\omega) \, d\mu(\omega). \]

It can be finite or \( \pm \infty \). Note from the definition that the Lebesgue integral of \( f \) is finite if and only if

\[ \int |f| \, d\mu < \infty, \]
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 \( (\Omega, \mathcal{F}, \mathbb{P}) \). Its distribution or law is \( \mathcal{L}_X = X_\# \mathbb{P} \). A useful notation: for \( B \in \mathcal{B}_\mathbb{R}, \)

\[
X^{-1}(B) \equiv \{ \omega \in \Omega : X(\omega) \in B \} \equiv \{ X \in B \}.
\]

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, \( \mu \)-almost everywhere (\( \mu \)-a.e.), in \( \mu \)-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

\[
Y = f(X) : (\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (S, \mathcal{G}, \mu) \xrightarrow{f} (\mathbb{R}, \mathcal{B}, \nu)
\]

where \( \mu = X_\# \mathbb{P} \) and \( \nu = f_\# \mu = Y_\# \mathbb{P} \). Provided either \( f \geq 0 \) or \( Y \) is integrable, we have \( I_\Omega = I_S \) where

\[
I_\Omega \equiv \mathbb{E}Y \equiv \int_{\Omega} Y \, d\mathbb{P} = \int_{\Omega} f(X(\omega)) \, d\mathbb{P}(\omega),
\]

\[
I_S \equiv \int_{S} f \, d\mu = \int_{S} f(x) \, d(\mathbb{P} \circ f^{-1})(x).
\]

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 \( \sigma \)-finite measure spaces \( (S, \mathcal{G}, \lambda) \) and \( (T, \mathcal{H}, \rho) \), we defined the product measure space \( (\Omega, \mathcal{F}, \mu) = (S \times T, \mathcal{G} \otimes \mathcal{H}, \lambda \otimes \rho) \).
2. Tonelli and Fubini theorems: for measurable \( f : \Omega \to \mathbb{R} \), provided either \( f \geq 0 \) or \( \int |f| \, d\mu < \infty \), we have

\[
\int_S \int_T f(x, y) \, d\rho(y) \, d\lambda(x) = \int_T \int_S f(x, y) \, d\mu(x, y) = \int_T \int_S f(x, y) \, d\lambda(x) \, d\rho(y).
\]

(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 \( (\Omega_{\alpha}, \mathcal{F}_{\alpha}) \) be measurable spaces indexed by \( \alpha \in I \) (index set, possibly uncountable). Their product:

\[
(\Omega, \mathcal{F}) = \left( \prod_{\alpha \in I} \Omega_{\alpha}, \bigotimes_{\alpha \in I} \mathcal{F}_{\alpha} \right)
\]

where \( \mathcal{F} \) is the minimal \( \sigma \)-field such that the coordinate projections \( \pi_{\alpha} : \Omega \to \Omega_{\alpha} \) are measurable. For \( J \subseteq I \) denote the partial products

\[
\Omega_J \equiv \prod_{\alpha \in J} \Omega_{\alpha}, \quad \mathcal{F}_J \equiv \bigotimes_{\alpha \in J} \mathcal{F}_{\alpha}.
\]
2. If \( \mathcal{P} \) is a probability measure on this \((\Omega, \mathcal{F}), \) its \textbf{finite-dimensional marginals} are the measures \((\pi_j)_k \mathcal{P}.\) This means that the marginal probability of \( E_j \in \mathcal{F}_j \) is
\[
\left( (\pi_j)_k \mathcal{P} \right)(E_j) = \mathcal{P}\left( (\pi_j)^{-1}(E_j) \right) = \mathcal{P}(E_j \times \Omega_{\backslash J}).
\]
They have to be \textbf{consistent}: if \( J' \subseteq J \subseteq I \) and \( \pi_{j \to J'} \) is the projection from \( \Omega_j \) to \( \Omega_{J'} \), then
\[
(\pi_{j \to J'})_k \mathcal{P} = (\pi_{j \to J'})_k \mathcal{P}.
\]
Both the theorems from this lecture go in reverse: given a consistent family of finite-dimensional marginals \( \{\mathcal{P}_j : \text{finite } J \subseteq I\} \), they construct a measure \( \mathcal{P} \) on \((\Omega, \mathcal{F})\) with these finite-dimensional marginals. Note: \( \{\mathcal{P}_j : \text{finite } J \subseteq I\} \) is often also called the \textbf{finite-dimensional distributions}, abbreviated \textbf{f.d.d.}

3. A useful variant of the criterion for the Carathéodory extension theorem (Theorem 2 above):

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

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 \( B_n \uparrow \emptyset \) of the form
\[
B_n = \tilde{B}_n \times \prod_{i=1}^{n+1} \Omega_i \times \prod_{a \in \Lambda \setminus Q} \Omega_a = \tilde{B}_n \times \Omega_{Q \setminus [n]} \times \Omega_{\text{rest}}
\]
where \([n] \equiv \{1, \ldots, n\} \) and \( \tilde{B}_n \in \mathcal{F}_{[n]} \).

4. Product measures on an infinite-dimensional spaces:

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

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

**Theorem 5** (Kolmogorovextension theorem). Let \((\Omega, \mathcal{F})\) be metric spaces with the Borel \( \sigma \)-field. Suppose \( \{\mathcal{P}_j : \text{finite } J \subseteq I\} \) is a consistent family of finite-dimensional distributions, and that each \( \mathcal{P}_j \) is an inner regular measure on \((\Omega_j, \mathcal{F}_j).\) Then there is a unique probability measure \( \mathcal{P} \) on \((\Omega, \mathcal{F})\) with finite-dimensional marginals \((\pi_j)_k \mathcal{P} = \mathcal{P}_j.\)

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
   \[
   (\Omega, \mathcal{F}, \mathcal{P}) = \left( \prod_{\alpha \in I} \Omega_\alpha, \bigotimes_{\alpha \in I} \mathcal{F}_\alpha, \prod_{\alpha \in I} \mathcal{P}_\alpha \right),
   \]
   and \( X_\alpha(\omega) = f_\alpha(\omega_\alpha) \) where \( f_\alpha \) is a measurable function on \( \Omega_\alpha, \) then \( (X_\alpha : \alpha \in I) \) is a collection of independent random variables on \((\Omega, \mathcal{F}, \mathcal{P}).\)
   b. If \((\Omega, \mathcal{F}, \mathcal{P})\) 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 \((\Omega, \mathcal{F}, \mathcal{P}),\) we can define \( X(\omega) \equiv \{X_\alpha(\omega) : \alpha \in I \}. \) This gives a mapping
   \[
   (\Omega, \mathcal{F}, \mathcal{P}) \rightarrow \left( \prod_{\alpha \in I} \mathbb{R}, \bigotimes_{\alpha \in I} \mathcal{B}_{\mathbb{R}}, \mathcal{L}_X \right)
   \]
   where \( \mathcal{L}_X \equiv X_\alpha \mathcal{P} = \mathcal{P} \circ X^{-1} \) 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
\[
\|X\|_p = \left\| \|X\|_{L^p(\Omega, \mathcal{F}, \mathbb{P})} = \left\{ \int |X(\omega)|^p \, d\mathbb{P}(\omega) \right\}^{1/p} = \mathbb{E}(|X|^p)^{1/p}. \]
We write \( L^p = L^p(\Omega, \mathcal{F}, \mathbb{P}) \) for the collection of \( X \) with \( \|X\|_p < \infty \). \( L^p \) monotonicity: if \( r \leq p \) then \( \|X\|_r \leq \|X\|_p \), so \( L^r(\Omega, \mathcal{F}, \mathbb{P}) \subseteq L^p(\Omega, \mathcal{F}, \mathbb{P}) \). Be careful that for \( L^p \) monotonicity it is essential to be on a probability space. The \( \ell_p \) sequence spaces are nested in the opposite direction: if \( r \leq p \) then
\[
\|X\|_r = \left( \sum_i |x_i|^r \right)^{1/r} \geq \left( \sum_i |x_i|^p \right)^{1/p} = \|X\|_p, \]
and so the \( \ell_p \) unit ball contains the \( \ell_r \) unit ball. Finally, the classical \( L^p \) function spaces are not nested at all: if
\[
\|f\|_p = \|f\|_{L^p(\mathbb{R})} = \left\{ \int |f(x)|^p \, dx \right\}^{1/p}
\]
and \( L^p(\mathbb{R}) = \{f : \|f\|_p < \infty\} \), 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 \( X \in L^2(\Omega, \mathcal{F}, \mathbb{P}) \), we define its variance \( \text{Var} X = \mathbb{E}(X - \mathbb{E}X)^2 \). Cauchy–Schwarz inequality: if \( X, Y \in L^2 \) then \( \|XY\|_1 = \mathbb{E}|XY| \leq \|X\|_2 \|Y\|_2 \). Consequently, for \( X, Y \in L^2 \) we can define their covariance
\[
\text{Cov}(X, Y) = \mathbb{E}\left\{ (X - \mathbb{E}X)(Y - \mathbb{E}Y) \right\}, \quad |\text{Cov}(X, Y)| \leq \left( \text{Var}(X)\text{Var}(Y) \right)^{1/2}.
\]
We say \( X \) and \( Y \) are uncorrelated if \( \text{Cov}(X, Y) = 0 \).

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 \( S_n = X_1 + \ldots + X_n \), 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 \( (X_{n,k} : 1 \leq k \leq n) \). 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
\[
S_n = \frac{1}{n} \sum_{k=1}^n X_{n,k}.
\]
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,” SLNN stands for “strong law of large numbers.”

1. \( L^2 \) WLLN for triangular arrays. If \( \text{Var} X_{n,k} \leq C \) uniformly, then \( (S_n - \mathbb{E}S_n)/b_n \to 0 \) 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
\[
(i) \lim_{n \to \infty} \sum_{k=1}^n \mathbb{P}(\{|X_{n,k}| > b_n\}) = 0, \\
(ii) \lim_{n \to \infty} \sum_{k=1}^n \mathbb{E}(|X_{n,k}|^2; |X_{n,k}| \leq b_n)/(b_n^2) = 0.
\]
Let \( Y_{n,k} \equiv \mathbb{1}_{\{|X_{n,k}| \leq b_n\}} \) 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 $\lim_{x \to \infty} x P(|X| \geq x) = 0$. Define $\mu_n \equiv E(X; |X| \leq n)$. Then $(S_n/n - \mu_n) \to 0$ in probability. Note that $E_X$ 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 $E_X = \mu$ 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 $E_X = \mu$ 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 $\mu$. In this lecture we study the chance of a large deviation, $S_n \geq na$ for $a > \mu$.

2. If the $X_i$ are i.i.d. standard gaussians, then $S_n \sim N(0, n)$, and

   $$P(S_n \geq na) = \int \frac{1(z \geq \sqrt{n}a)}{\sqrt{2\pi}} \exp(\frac{z^2}{2}) dz = \exp \left\{ - \frac{na^2}{2} + o(n) \right\}.$$  

3. If the $X_i$ are i.i.d. $\text{Ber}(p)$ for $p \in (0, 1)$, then $S_n \sim \text{Bin}(n, p)$, and we used Stirling’s formula to estimate

   $$P(S_n = nx) = \left( \frac{n}{nx} \right)^{nx} (1-p)^{n(1-x)} = \exp \left\{ -n I_p(x) + o(n) \right\}$$

   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

   $$P \left( \frac{S_n - np}{\sqrt{np(1-p)}} \in [u, v] \right) \to \int_{u}^{v} \frac{dz}{\sqrt{2\pi}} \exp(\frac{z^2}{2})$$

   for any fixed $u, v \in \mathbb{R}$. However, for any fixed $a \in (p, 1)$, the large deviations probability $P(S_n \geq na)$ is outside the regime of the gaussian approximation, and

   $$P(S_n \geq na) = \exp \left\{ -n \mathcal{H}(a|p) + o(n) \right\} \neq \exp \left\{ - \frac{n(a-p)^2}{2p(1-p)} + o(n) \right\}$$

   (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 $nk p$ entries uniformly at random, what is the probability of the event $F$ that every column has at least one filled entry?

   $$P(F) = P_{\theta} \left( \sum_{i=1}^{n} X_i = nkp \right) \geq 1$$

   where under $P_{\theta}$ we let $X, X_i$ be i.i.d. $\text{Bin}(k, \theta)$. Then

   $$P(F) = \frac{P_{\theta}(X \geq 1)^n}{P(\text{Bin}(nk, \theta) = nkp) \sum_{i=1}^{n} X_i = nkp \left\{ \sum_{i=1}^{n} X_i \geq 1 \forall 1 \leq i \leq n \right\}}$$

   We discussed how to calculate each of the three factors for a good choice of $\theta$.  

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 $N(0, I_{d \times d})$; 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 $\mathbb{R}^d$-valued random variable $X = \mu + AZ$ is a multivariate gaussian, and we denote its law by $N(\mu, AA^t)$. 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

$$D \equiv \left\{ \theta \in \mathbb{R} : m(\theta) < \infty \right\},$$

and note $0 \in D$ always. It is possible that $D = \{0\}$.

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 $\mu < a < \inf \sup X$, the sum $S_n = X_1 + \ldots + X_n$ satisfies the large deviations estimate

$$-\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na) = I(a) \equiv \sup_{\theta \geq 0} \left\{ \theta a - \kappa(\theta) \right\} \sup_{\theta \in \mathbb{R}} \left\{ \theta a - \kappa(\theta) \right\} \equiv \kappa^*(a),$$

the Legendre dual or Fenchel–Legendre transform of $\kappa$.

Note the large deviations estimate can be written equivalently as $\mathbb{P}(S_n/n \geq a) = \exp(-n I(a) + o(n))$.

3. The upper bound of Theorem 6 is an easy consequence of Markov’s inequality: for any $\theta \geq 0$,

$$\mathbb{P}(S_n \geq na) \leq \mathbb{P}(e^{n\kappa'} \geq e^{n\theta a}) \leq \frac{\mathbb{E}(e^{n\kappa'})}{e^{n\theta a}} = \exp \left\{ -n \left[ \theta a - \kappa(\theta) \right] \right\}.$$ (The exponential form of Markov’s inequality is often called a Chernoff bound.)

4. For $\theta \in D$ (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 $\theta$ in the interior of $D$, we have $\kappa'(\theta) = \mathbb{E}_\theta X$ and $\kappa''(\theta) = \text{Var}_\theta X > 0$ (assuming the law of $X$ is nondegenerate). Thus $\kappa$ is strictly convex in the interior of $D$. For $\theta > 0$ small enough we have $\mu < \kappa'(\theta) < a$, so $\theta a - \kappa(\theta)$ is increasing in $\theta$ for $\theta > 0$ small enough; and this implies the equality marked $\theta$ in Theorem 6.

5. We proved the lower bound of Theorem 6 in the special case that the supremum over $\theta$ in $\kappa^*(a)$ is achieved by $\theta_a$ in the interior of the set of $\theta$ where $m(\theta) < \infty$. This is done by a change of measure to $\mathbb{P}_\theta$ in which $\{S_n \geq na\}$ becomes a typical rather than rare event. (Read in [Dur19, §2.6] for the proof of Theorem 6 in its full generality.)

6. Informal interpretation of Cramér’s theorem: “the most efficient way to achieve a large deviation $S_n \geq na$ is for $X_1, \ldots, X_n$ to ‘behave like’ a sample from $\mathbb{P}_\theta$, with $\theta$ 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 $1 \leq j \leq k$. The empirical measure of $X_1, \ldots, X_n$ is the random measure

$$L_n^X = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}.$$ 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 $L_n^X$ as the $k$-tuple of empirical fractions

$$L_n^X = \left( \frac{\left\{ 1 \leq i \leq n : X_i = x_j \right\}}{n} \right)_{1 \leq j \leq k} \in [0, 1]^k.$$
Moreover, in this special case we can directly calculate
\[ P(L_n^X = \nu) = \frac{n!(\pi_1)^{\nu_1} \cdots (\pi_k)^{\nu_k}}{(n\nu_1)! \cdots (n\nu_k)!} \equiv \left( \frac{n}{\nu} \right)^{\pi^\nu} = \exp \left\{ -nH(\nu|\pi) + o(n) \right\} \]
where \( H(\nu|\pi) \) is the relative entropy or Kullback–Leibler divergence between \( \nu \) and \( \pi \):
\[ H(\nu|\pi) \equiv D_{\text{KL}}(\nu|\pi) \equiv \sum_{j=1}^{k} \nu_j \log \frac{\nu_j}{\pi_j} \geq 0, \]
a convex function of \( \nu \). The empirical mean of \( X_1, \ldots, X_n \) can be expressed in terms of the empirical measure as \( S_n/n = (x, L_n^X) \). Then, for \( \mathbb{E}X < a < \text{ess sup } X = \max\{x_j : 1 \leq j \leq k\} \), we can calculate
\[ P(S \geq na) = \sum_{\nu} 1((x, \nu) \geq a) P(L_n^X = \nu) = \exp \left\{ -n \inf \left\{ H(\nu|\pi) : \langle x, \nu \rangle \geq a \right\} + o(n) \right\}. \]
The last approximation uses that the set of all possible values \( L_n^X = \nu \) 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
\[ L(\nu, \theta) = H(\nu|\pi) + \rho \left( 1 - \sum_i \nu_i \right) + \theta \left( 1 - \sum_i x_i \nu_i \right), \]
and setting \( \partial L/\partial \nu_j = 0 \) gives the exponentially tilted measure
\[ \nu = \frac{\pi_j \exp(\theta x_j)}{m(\theta)} = \pi \theta. \]
Check for yourself that \( H(\pi \theta|\pi) = \theta \mathbb{E} \theta X - \kappa(\theta) = \theta \kappa'(\theta) - \kappa(\theta) \). For \( \mathbb{E}X < a < \text{ess sup } X \), there is a unique \( \theta(a) \) such that \( \kappa'(\theta(a)) = a \). Therefore
\[ P(S_n \geq na) = \exp \left\{ -n \inf \left\{ \theta \kappa'(\theta) - \kappa(\theta) : \kappa'(\theta) \geq a \right\} + o(n) \right\} \]
\[ = \exp \left\{ -n \inf_{b \geq a} \left\{ \theta(b) \kappa'(\theta(b)) - \kappa(\theta(b)) \right\} + o(n) \right\} \]
\[ = \exp \left\{ -n \left( \theta(a) \kappa'(\theta(a)) - \kappa(\theta(a)) \right) + o(n) \right\} = \exp \left\{ -n \kappa'(a) + o(n) \right\} \]
(in particular, check in the above that the infimum over \( b \geq a \) 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 \( P(X \geq x) \) decays like \( 1/x^p \) for any finite \( p \). Convince yourself in this case that \( P(S_n \geq na) \) 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 \( (\Omega, \mathcal{F}, \mathbb{P}) \). Recall that its distribution or law \( \mu = L_X = X_\# \mathbb{P} \) is a probability measure on \( (\mathbb{R}, \mathcal{B}_\mathbb{R}) \). Its cumulative distribution function (cdf) is the function
\[ F(x) \equiv P(X \leq x) = \mu((-\infty, x]) \]. Clearly, \( \mu \) uniquely determines \( F \). The converse is also true: \( F \) uniquely determines \( \mu \), which can be proved by a \( \pi\lambda \)-argument. It is common to write \( "dF(x)" \) which has the same meaning as \( "d\mu(x)" \) where \( \mu \) is the measure determined by \( F \).

2. Let \( X \) and \( Y \) be (real-valued) random variables on \( (\Omega, \mathcal{F}, \mathbb{P}) \), with laws \( \mu = L_X \) and \( \nu = L_Y \), and cdf’s \( F(x) = P(X \leq x) \) and \( G(y) = P(Y \leq y) \). Assume \( X \) and \( Y \) are independent, i.e., their joint law \( L_{(X,Y)} \) 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 \( \mu \) and \( \nu \), denoted \( \mu * \nu \). The associated cdf is denoted \( F * G \), and is given explicitly by
\[ (F * G)(z) = P(X + Y \leq z) = \int \int 1\{x + y \leq z\} \, d\mu(x) \, d\nu(y) = \int F(z - y) \, dG(y), \]
having used the change of variables formula and Tonelli’s theorem, and recalling that “$dG(y)$” is equivalent notation for “$dν(y)$.” Symmetrically, we also have the formula

$$(F * G)(z) = \mathbb{P}(X + Y \leq z) = \int \int 1\{x + y \leq z\} \, dν(y) \, dμ(x) = \int G(z - x) \, dF(x).$$

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 $μ * ν$ also has a density:

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

so the density for $μ * ν$ 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 $(\mathbb{R}, \mathcal{B}_\mathbb{R})$, the characteristic function (chf) — equivalently, the Fourier transform — is the function $φ_μ: \mathbb{R} \to \mathbb{C}$ defined by

$$φ_μ(t) = \int e^{itx} \, dμ(x). \quad (4)$$

The function $φ_μ$ is well-defined, continuous, and satisfies $|φ_μ(t)| ≤ 1$ 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 $φ_μ(t) = \mathbb{E}(\exp(itX))$. 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

$$φ_{μν}(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}(e^{itX})\mathbb{E}(e^{itY}) = φ_μ(t)φ_ν(t).$$

Reading: PTE 2.1.3, and PTE Theorem 3.3.1.

5. Some of the basic mechanics of the Fourier transform can be worked out very easily and explicitly by considering the discrete space $Ω = \mathbb{Z}/n\mathbb{Z}$ (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 $f : Ω \to \mathbb{C}$. Then $V \cong \mathbb{C}^Ω$, a finite-dimensional complex vector space. It is naturally equipped with the hermitian inner product

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

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 $V \equiv L^2(Ω)$ to emphasize the inner product structure.

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

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

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

$$λ ∈ Φ_n \equiv \{z ∈ \mathbb{C} : z^n = 1\} = \left\{1, \exp\left(\frac{2πi}{n}\right), \ldots, \exp\left(\frac{2πi(n-1)}{n}\right)\right\}.$$
In the general case where \( n \) is not prime, then it is possible to have \( k z \equiv 0 \) 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 \( \Phi_n \) 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

\[
\chi_y(x) = \frac{1}{\sqrt{n}} \exp \left( \frac{2\pi i x y}{n} \right)
\]

is an eigenvector of \( T_z \) with eigenvalue \( \lambda_{z,y} \equiv \exp \left( -\frac{2\pi i y z}{n} \right) \) ∈ \( \Phi_n \).

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 \( \Omega \) is a field, so for any \( x, y \in \Omega \setminus \{0\} \) we have \( y = r x \) for \( r \neq 0 \), so \( T_y = T_{rx} = (T_z)^r \). Thus, if \( \chi \) is an eigenvector of \( T_z \) with eigenvalue \( \lambda \), then it must also be an eigenvector of \( T_y \) with eigenvalue \( \lambda^r \). 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 \( \Phi_n \). (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 \( \Omega \). The \( \chi_y \) 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

\[
\langle \chi_x, \chi_y \rangle = \sum_{z \in \Omega} \chi_x(z) \overline{\chi_y(z)} = \sum_{z \in \Omega} \frac{1}{n} \exp \left( \frac{2\pi i (x - y) z}{n} \right) = 1 \{ x = y \},
\]

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 \( \Lambda_z \) 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: \( H = U D U^* \) 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 \( (T_z)_{x,y} = 1 \{ y = x - z \} \), from which we can work out that \( (T_z)^* = (T_z)^t = T_{-z} = (T_z)^{-1} \). If \( \chi \) and \( \psi \) are eigenvectors of \( T_z \) with distinct eigenvalues \( \lambda \neq \gamma \), then

\[
\langle T_z \chi, \psi \rangle = \langle \lambda \chi, \psi \rangle = \lambda \langle \chi, \psi \rangle, \\
\langle T_z \chi, \psi \rangle = \langle \chi, (T_z)^t \psi \rangle = \langle \chi, (T_z)^{-1} \psi \rangle = \langle \chi, \gamma^{-1} \psi \rangle = \gamma^{-1} \langle \chi, \psi \rangle = \gamma \langle \chi, \psi \rangle,
\]
where the last relation uses that \( \gamma \) 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 transform 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
\[
C_g = \sum_{z \in \Omega} g(z)\sum_{\ell \in \Omega} \lambda_{z, \ell} \hat{\chi}_\ell(\hat{\chi}_\ell)' = \sum_{\ell \in \Omega} \left( \sum_{z \in \Omega} \lambda_{z, \ell} g(z) \right) \hat{\chi}_\ell(\hat{\chi}_\ell)',
\]
where the last equality uses that \( \lambda_{z, \ell} = (U^*)_{z, \ell} \). In more succinct form, if \( \text{diag}(U^* g) \) denotes the diagonal matrix with diagonal entries given by \( U^* g \), then the above shows that
\[
C_g = U \text{diag}(U^* g)U^*.
\]

It follows from this that
\[
\hat{f} * \hat{g} = U^* (f * g) = U^* C_g f = U^* \left( U \text{diag}(U^* g)U^* \right) f = \text{diag}(U^* g)U^* f = \text{diag}(\hat{g}) \hat{f} = \hat{f} \circ \hat{g},
\]
where \( \circ \) 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 \( \ell \neq \frac{n}{2} \) then
\[
(\chi_\ell \quad \chi_{-\ell}) \left( \begin{array}{cc} 1/2 & 1 \/i \\ 1/2 & -1/2 \end{array} \right) = (c_\ell \quad s_\ell)
\]
where \( c_\ell \) and \( s_\ell \) 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 \( \ell = \frac{n}{2} \) then \( \chi_\ell \) is itself a cosine function,
\[
\chi_{\ell}(x) = \frac{\exp(\pi i x)}{\sqrt{n}} = \frac{(-1)^x}{\sqrt{n}} = \cos(\pi k) \sqrt{n}
\]

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 \( \left( \chi_\ell : \ell \in \Omega \right) \). 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
\[
f^*(t) = \int \text{e}^{itx} f(t) \, dt.
\]

You can use Jensen’s inequality to show that \( \|f^*\|_\infty \leq \|f\|_1 \). A key result is the following: if \( f, h \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \), then \( f^*, g^* \in L^\infty(\mathbb{R}) \cap L^2(\mathbb{R}) \), and
\[
\int_{\mathbb{R}} f(x)\overline{h(x)} \, dx = \langle f, h \rangle = \frac{\langle f^*, h^* \rangle}{2\pi} = \frac{1}{2\pi} \int f^*(t)\overline{h^*(t)} \, dt.
\]
This can be used to show that the mapping
\[ UI : L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \to L^{\infty}(\mathbb{R}) \cap L^2(\mathbb{R}), \quad f \mapsto Uf \equiv \frac{f^*}{\sqrt{2\pi}} \]
has a unique continuous extension to a **unitary isometry** \( UI : L^2(\mathbb{R}) \to L^2(\mathbb{R}) \). The \( L^2(\mathbb{R}) \) **Fourier transform** refers to the mapping from \( f \in L^2(\mathbb{R}) \) to \( f^* = \sqrt{2\pi} Uf \in L^2(\mathbb{R}) \). The identity (6) holds for all \( f, h \in L^2(\mathbb{R}) \). The "conjugate transpose" is the map
\[ U^* : L^2(\mathbb{R}) \to L^2(\mathbb{R}), \quad (U^* f)(t) = \frac{f^*(-t)}{\sqrt{2\pi}}. \]
For any \( f \in L^2(\mathbb{R}) \) we have the **Fourier inversion formula**
\[ f(t) = (U^* U f)(t) = \frac{(U^* f^*)(t)}{\sqrt{2\pi}} = \frac{f^*(-t)}{2\pi} = \frac{1}{2\pi} \int e^{-itx} f^*(t) \, dt. \]
For details see [LL01, Ch. 5].

2. If \( \mu \) 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 \( \mu \) may not have a density at all. For this we have:

**Theorem 7** (Fourier inversion theorem for probability measures on \( \mathbb{R} \)). If \( \mu \) is a probability measure on \( (\mathbb{R}, \mathcal{B}) \) with Fourier transform (characteristic function) \( \varphi \), then for all \( -\infty < a < b < \infty \)
\[ \left[ I_T \equiv \frac{1}{2\pi} \int_{-T}^{T} \varphi(t) \frac{e^{-ita} - e^{-itb}}{2\pi i} \, dt \right]_{T \to \infty} \tilde{\mu}(a, b) \equiv \mu((a, b)) + \frac{\mu\{a, b\}}{2}. \]
In particular, \( \varphi \) uniquely determines \( \tilde{\mu} \), which in turn uniquely determines \( \mu \).

Intuition for the theorem: note that if \( h(t) \equiv 1 \{t \in (a, b)\} \) and \( k_T(t) \equiv 1\{|t| \leq T\} \) then
\[ I_T = \frac{1}{2\pi} \int \varphi(t) h_T^*(t) k_T(t) \, dt \]
In view of (6) it is natural to expect (although it does not directly follow) that
\[ I_T = \int h_T(x) \, d\mu(x) \quad (7) \]
where \( h_T \) is the function such that \( (h_T)^*(t) = h^*(t) k_T(t) \). The key steps in the proof of Theorem 7 are to show the identity (7), and to show that
\[ \lim_{T \to \infty} h_T(t) = \tilde{h}(t) = 1\{t \in (a, b)\} + \frac{1\{a \in \{a, b\}\}}{2}. \]
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 \( \sigma \)-algebra \( \mathcal{B} \). If \( \mu, \mu_n \) are probability measures on \( (S, \mathcal{B}) \), we say that \( \mu_n \) **converges weakly** to \( \mu \) (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 \( L_{X_n} \Rightarrow L_X \).

2. Any probability measure \( \mu \) on \( (S, \mathcal{B}) \) is **regular**: for all \( A \in \mathcal{B} \),
\[ \mu(A) = \sup \left\{ \mu(F) : F \text{ closed, } F \subseteq A \right\} = \inf \left\{ \mu(G) : G \text{ open, } A \subseteq G \right\}. \]
This implies that $\mu$ is completely determined by $\{\mu(F) : F \text{ closed}\}$. This can further be used to show that $\mu$ is completely determined by the values of
\[ \int f \, d\mu \]
for bounded continuous $f$. This shows that a sequence $\mu_n$ cannot converge weakly to two different limits.

3. Let $\mathcal{P}$ be the space of probability measures on $(S, \mathcal{B})$. Let $\mathcal{T}$ be the topology on $\mathcal{P}$ generated by the sets
\[ \left\{ \nu : \left| \int f_i \, d\nu - \int f_i \, d\mu \right| < \epsilon \text{ for all } 1 \leq i \leq k \right\} \]
where $f_i$ are bounded continuous functions. Then weak convergence is equivalent to convergence in the topology $\mathcal{T}$. If $S$ is a complete separable metric space (Polish space), then $\mathcal{T}$ is a metric topology, and $\mathcal{P}$ 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 $\epsilon$ only) such that
\[ \inf \left\{ \mu_\alpha(K) : \alpha \in I \right\} \geq 1 - \epsilon. \]

**Theorem 8** (Prohorov’s theorem). If $\{\mu_\alpha : \alpha \in I\}$ is tight, then it is relatively compact (has compact closure) in $\mathcal{P}$. (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 $\mu_n$: first use Theorem 8 to show $\{\mu_n\}_{n \geq 1}$ is confined within a compact subset of $\mathcal{P}$, 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 $\mu_n$ be probability measures on $\mathbb{R}$ with characteristic functions $\varphi_n$.

- a. If $\mu_n \Rightarrow \mu$ then $\varphi_n \to \varphi$ pointwise where $\varphi$ is the characteristic function of $\mu$.
- b. If $\varphi_n$ converges pointwise to a function $\varphi$ that is continuous at $t = 0$, then $\varphi$ is the characteristic function of a probability measure $\mu$, 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 $\pi$ be a random permutation of $[n]$, and let $S_n$ be the number of disjoint cycles in $\pi$. 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 $\pi$ of length $k$, so that
\[ S_n = \sum_{k \geq 1} C_{n,k}. \]

Then $(C_{n,k})_{k \geq 1}$ converges in law to $(Y_k)_{k \geq 1}$ where $Y_k$ are independent $\text{Pois}(1/k)$ 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 $S_n \sim \text{Bin}(n, \lambda/n)$ then $S_n \xrightarrow{d} Y_\lambda \sim \text{Pois}(\lambda)$ 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 $I_k \sim \text{Ber}(p)$ for $k \geq 1$. The number of successes by time $m$ is
\[ B_m = \sum_{k=1}^{m} I_k \sim \text{Bin}(m, p), \quad \mathbb{P}(B_m = \ell) = \binom{m}{\ell} p^\ell (1-p)^{m-\ell}. \]
The time of the first success is
\[ G = \min\left\{ k : I_k = 1 \right\} \sim \text{Geo}(p), \quad \mathbb{P}(G = k) = (1 - p)^{k-1}p \text{ for } k \in \{1, 2, \ldots\}. \]

Let \( G_1, G_2, \ldots \) be i.i.d. copies of \( G \). The time of the \( r \)-th success is
\[ X_r = \sum_{i=1}^{r} G_i \sim \text{NegBin}(r, p), \quad \mathbb{P}(X_r = t) = \binom{t - 1}{r - 1} (1 - p)^{t-r}p^r. \]

4. Now take \( p = 1/n \) and scale time by \( n \), so \( \text{Ber}(1/n) \) trials happen at times \( 1/n, 2/n, \ldots \). The number of successes by time \( t \) is now
\[ B_{nt} \sim \text{Bin}\left(nt, \frac{1}{n}\right) \xrightarrow{d} \text{Pois}(t). \]

The time of the first success is
\[ \frac{G}{n} \sim \frac{\text{Geo}(1/n)}{n} \xrightarrow{d} E \sim \text{Exp}, \quad \mathbb{P}(E \in [a, b]) = \int_a^b e^{-t} dt. \]

The time of the \( r \)-th success is
\[ \frac{1}{n} \sum_{i=1}^{r} G_i \sim \frac{\text{NegBin}(r, 1/n)}{n} \xrightarrow{d} Y = \sum_{i=1}^{r} E_i \sim \Gamma(r) = \text{Exp}^r. \]

The gamma density can be derived by directly taking limits from the negative binomial distribution:
\[ \mathbb{P}(Y \in [a, b]) = \int_a^b e^{-x} x^{-r-1} (r-1)! \, dx. \]

More generally, for any \( \alpha > 0 \), we write \( \Gamma(\alpha) \) for the probability measure on \( \mathbb{R} \) with density
\[ f(x) = \mathbf{1}\{x \geq 0\} \frac{e^{-x} x^{\alpha-1}}{\Gamma(\alpha)}, \quad \Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} \, dx. \]

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 \((\mathbb{R}, \mathcal{B})\). (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
\[ \Psi(x) = \mathbb{P}(Z \geq x) = \int_x^\infty g(z) \, dz = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \, dz, \]
where we use \( g \) to denote the standard gaussian density. We saw the following estimate for all \( x > 0 \):
\[ \frac{g(x)}{x} - \frac{1 - 1}{x^2} \leq \Psi(x) \leq \frac{g(x)}{x} \tag{8} \]

2. Two examples of extremal statistics: if \( X_i \) are i.i.d. standard exponential random variables, then
\[ \max_{1 \leq i \leq n} X_i - \log n \xrightarrow{d} \text{Gumbel} \]
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 \( \tag{8} \) can be used to show that
\[ b_n \left( \max_{1 \leq i \leq n} Z_i - b_n \right) \xrightarrow{d} \text{Gumbel} \]
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 $(\Omega, \mathcal{F}, \mathbb{P})$, if $A \in \mathcal{F}$ 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 $\mathbb{E}M_0 = \mathbb{E}M_\tau$ for all $n$. In upcoming lectures we will prove that $\mathbb{E}M_0 = \mathbb{E}M_\tau$ for stopping times $\tau$, under suitable conditions on $M$ and $\tau$. 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 $\tau$ is a stopping time, but $\mathbb{E}M_0 \neq \mathbb{E}M_\tau$.

4. Why we expect OST to hold in certain circumstances: note that if $M_n$ is a martingale and $\tau$ is a stopping time, then the stopped process $X_\tau = M_{\tau \wedge n}$ is also a martingale: we can decompose
\[
X_\tau = M_{\tau \wedge n} = \sum_{i=1}^{n-1} \mathbf{1}_{\{\tau = i\}} M_i + \mathbf{1}_{\{\tau \geq n\}} M_n,
\]
and then the martingale property of $M_n$ implies
\[
\mathbb{E}(X_\tau \mid \mathcal{F}_{\tau-1}) = \sum_{i=1}^{n-1} \mathbf{1}_{\{\tau = i\}} M_i + \mathbf{1}_{\{\tau \geq n\}} M_{n-1} = X_{\tau-1}.
\]
Therefore $\mathbb{E}X_0 = \mathbb{E}X_n$ for all $n$, which means $\mathbb{E}M_0 = \mathbb{E}M_{\tau \wedge n}$ for all $n$. In the limit $n \to \infty$ we have $n \wedge \tau \to n$ a.s., and $M_{\tau \wedge n} \to M_\tau$ a.s. Based on this, we expect under some integrability conditions that $\mathbb{E}M_\tau \to \mathbb{E}M_\tau$, 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 $\emptyset \subseteq Z \subseteq V$. 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
\[
h : V \to \mathbb{R}, \quad h(x) = \mathbb{E}\left[f(X_\tau) \mid X_0 = x\right]
\]
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).
References