IOE 516

Stochastic Processes II

Winter Term, 2024

Prof. Xiuli Chao

Email: xchao@umich.edu

Course info

- IOE 516 is "Stochastic Processes II", or "Advanced Stochastic Processes".

- It can be considered as continuation of IOE 515, but ayou do not need to take IOE 515 before taking IOE 516 (though that will definitely help).

- In contrast to IOE 515, IOE 516 is for doctoral students and methodological oriented Masters/undergrad students.

- There was almost no proofs in IOE 515, but in IOE 516, there will be technical proofs, even though I emphasize intuition and insights.

Course info

- **Day and time:** Friday 9:00qm - 12:00pm.
- **Classroom:** Classroom IOE 1680
- **Instructor:** Xiuli Chao, IOE 2895, email: xchao@umich.edu
- **Office hour:** Thursday 2pm - 3pm.
- **GSI:** Sogand Soghrati Ghasbeh
- **GSI office hour:** (i) IOE 1824 (Flex lab) on Wednesdays, 2-3pm, and (ii) IOE 2858 on Thursdays, 3:30-4:30pm. You can also join remotely on zoom https://umich.zoom.us/j/4680722391

Grading

- There will be about 6 sets of homework problems, and you are typically given 2 weeks to finish each of them.

- There will be two exams, midterm exam and final exam. Midterm exam will be in the week right after winter break, on Friday March 8. Final exam is during the exam week. Midterm exam is in class, open book and open notes, but closed to internet, while the final exam is take home.

- The grading is 40-30-30. That is, your final score = Homework $\times$ 40% + Midterm exam $\times$ 30% + Final exam $\times$ 30%.

- Late homework submission has 50% penalty.

Course outline

- Let us go over the syllabus
- No mandate textbook, but I have listed a number of good references.
- Main topics:
  - Some fundamentals of probability theory (convergence, WLLN, SLLN, CLT, large deviation, and concentration inequalities, etc.)
  - Renewal Theory
  - DTMC
  - Basics of martingale theory
  - CTMC, and continuous state space Markov processes (Brownian motion, etc.)

Any questions before we proceed?

A motivating example

- Multi-armed bandit (MAB) problem, with tons of applications.

- A popular method for solving MAB is UCB (upper confident bound). Similar solution has been developed for general reinforcement learning problems.

- Why would UCB work and what's the intuition? how to prove it works, The machinery needed is precisely what we study in this course. We will get back to this problem.

Probability space $(\Omega, \mathcal{F}, P)$

- Probability space was rigorously defined via axiom by Kolmogorov in the 1930's. It consists of a triplet $(\Omega, \mathcal{F}, P)$.
  - Ω: Sample space
  - $\mathcal{F}$: Set of events
  - $P(A)$: for each event $A \in \mathcal{F}$.
- The triplet $(\Omega, \mathcal{F}, P)$ is called a probability space if the three conditions are satisfied.

Probability axioms

(i) $0 \leq P(A) \leq 1$ for any $A \in \mathcal{F}$.

(ii) $P(\Omega) = 1$.

(iii) If $A_1, A_2, \ldots$ is a sequence of mutually exclusive events, then

$$P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} P(A_n).$$

Remark

- $\mathcal{F}$ is the so-called set of events.

- In the case Ω is a finite set, $\mathcal{F}$ can be, e.g., the collection of all subsets of Ω, that contains $2^{|\Omega|}$ events. That is why it is also called power set.

- If Ω is not countable, event can be difficult to define. The so-called σ-algebra condition is a formal way to define all the "measurable events".

σ-algebra

- A σ-algebra $\mathcal{F}$ is a collection of subsets of Ω satisfying the following conditions:

  (i) $\Omega \in \mathcal{F}$

  (ii) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$

  (iii) If a sequence $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_i A_i \in \mathcal{F}$.

- Intuitively, σ-algebra represents the information cumulated up to a point in time.

Continuity of probability

- If A_n is increasing, i.e., $A_1 \subset A_2 \subset \cdots$, then we write

$$\bigcup_n A_n = \lim_n A_n.$$

- Similarly, if A_n is decreasing, i.e., $A_1 \supset A_2 \supset \cdots$, then we write

$$\bigcap_n A_n = \lim_n A_n.$$

Continuity of probability

- If either A_n is increasing or decreasing, we have

$$P(\lim_n A_n) = \lim_n P(A_n).$$

Proof

- Suppose A_n is increasing. Define $B_n = A_n \setminus A_{n-1}$. Then,

$$\bigcup_n A_n = \bigcup_n B_n.$$

- Thus ...

What is random variable?

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- A random variable, say X, is a variable whose value depends on the outcome $\omega \in \Omega$. So formally we can also write it as $X(\omega)$, $\omega \in \Omega$.

- Therefore, random variable is nothing but function, and if we consider real random variable, then

$$X(\omega) : \Omega \to R$$

Preliminaries

- We will first review some important and useful limit theorems in probability theory.

- That will involve several concepts of convergence, in particular, almost surely convergence, convergence in probability, and convergence in distribution.

- These are essential for concentration inequalities, that is a central tool in analyzing/developing learning algorithms (machine learning, statistical learning, on-line learning, off-line learning, reinforcement learning, etc.) and in data science, including the MAB problem discussed earlier.

- Wes first review the concepts. Let $\{X_n; n \geq 1\}$ be a sequence of random variables

Convergence

- **Convergence in probability**: X_n is said to converge to X in probability if, for any $\epsilon > 0$, it holds that

$$P(|X_n - X| > \epsilon) \to 0 \text{ as } n \to \infty.$$

- **Almost sure convergence** (or converge with probability 1): X_n is said to converge to X almost surely (or with probability 1) if

$$P\Big(\omega : X_n(\omega) \to X(\omega)\Big) = 1.$$

- **Convergence in L^p**: X_n is said to converge to X in L^p if

$$E[|X_n - X|^p] = 0 \text{ as } n \to \infty.$$

- **Convergence in distribution**: Suppose X_n has cdf F_n and X has cdf F. X_n is said to converge to X in distribution if,

$$F_n(x) \to F(x) \text{ as } n \to \infty$$

on every point x at which $F(x)$ is continuous.

Remark

- There are relationships among these convergence concepts, i.e., some imply the other. We will discuss them later after learning techniques to prove them.

Preliminaries
Markov inequality

- Before discussing limit theorems, we need some basic probability inequalities.

- The simplest of all is the so-called **Markov inequality**: For non-negative random variable X, for any $a > 0$, we have

$$P(X > a) \leq \frac{E[X]}{a}.$$

Why?

- We prove the continuous r.v. case. Suppose it has pdf $f(x)$. Then

$$E[X] = \int_0^\infty xf(x)dx \geq \int_a^\infty xf(x)dx \geq a\int_a^\infty f(x)dx = aP(X > a).$$

- Thus

$$P(X > a) \leq \frac{E[X]}{a}.$$

Chebyshev inequality

- Let X be a random variable with finite μ and variance σ^2, then for any $\epsilon > 0$,

$$P(|X - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

- **Proof.**

Weakly Law of Large Numbers (WLLN)

- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2. Then

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

  is called the sample mean.

- Then, $\bar{X}_n$ converges to the mean μ in probability, i.e., for any $\epsilon > 0$,

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

Why?

- $E[\bar{X}_n] = \mu$, and $Var(\bar{X}_n) = \sigma^2/n$. Apply Chebyshev inequality.

Strong Law of Large Numbers (SLLN)

- $\bar{X}_n$ converges to the mean μ almost surely. That is, for almost everyu sample ω, we have

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \to \mu \text{ as } n \to \infty$$

- The proof of SLLN relies on Borel-Cantelli lemma, and we will discuss it depending on time.

Discussions (1/3)

- Strong Law of Large Number shows that

$$\frac{X_1 + \cdots + X_n - nE[X_1]}{n}$$

converges to 0 almost surely.

- **Question.** How fast does it go to 0?

Discussions (2/3)

- **Question:** Suppose $\alpha > 0$. Does

$$\frac{X_1 + \cdots + X_n - nE[X_1]}{n^\alpha}$$

converge as $n \to \infty$?

- For $\alpha = 1$, it converges a.s. That implies that the same is true for any $\alpha \geq 1$.

- What happens when $\alpha < 1$? In particular, what if $\alpha = 1/2$? CLT will address this. How about other value of α?

- I will design some coding (Excel fine too) assignment for you to demonstrate this.

Discussions (3/3)

- **Question.** How about limit theorems for random variables are not i.i.d.? In ML, RL, etc., the random variables in the next periods depend on what happened in earlier periods, hence the sequence is typically not i.i.d.

- An extremely important class of dependent random variables, which which we can extend the limit theorems for i.i.d. case, is martingale process.

Remark

- We next discuss Central Limit Theorem (CLT).
- But first, we need some preliminaries.
- MGF and Characteristic function.

Moment Generating Function (MGF)

- A moment generating function (MGF) of a random variable X is defined as, for any $\theta \in R$,

$$M_X(\theta) = E[e^{\theta X}].$$

- **Example.** MGF of exponential r.v. with parameter λ.

Claim

- If the MGF of a random variable X is finite in a small neighborhood of 0, then for any positive integer n,

$$E[X^n] = M_X^{(n)}(\theta)\Big|_{\theta=0}$$

Drawback of MGF

- The previous slides indicates that we need MGF to be finite near 0. This, however, may not be satisfied.

- **Example**. Let X be lognormal $X = e^Z$, where Z is standard normal. What's its n-th moment?

$$E[X^n] = E[e^{nZ}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{nz} e^{-z^2/2} dz = e^{n^2/2}.$$

- However, by change of variable $y = e^z$, we have for any $\theta > 0$,

$$M_X(\theta) = E[e^{\theta X}] = E[e^{\theta e^Z}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\theta e^z} e^{-z^2/2} dz$$
$$= \int_0^{\infty} \frac{1}{y} e^{\theta y} e^{-(\log y)^2} dy = \infty$$

What comes to the rescue?

- It is characteristic function. For any random variable X, its characteristic function is defined as

$$\phi_X(\theta) = E[e^{i\theta X}].$$

- Some properties:

  (i) $\phi_X(0) = 1$, (ii) $|\phi_X(\theta)| \leq 1$ for all θ,

  (iii) $\phi_X(\theta)$ is continuous, and (iv) $\phi_X^{(n)}(\theta)\Big|_{\theta=0} = i^n E[X^n]$.

Why MGF and/or characteristic function?

- This is the time domain versus frequency domain analysis. Using the transform, it allows us to analyze a problem that is otherwise difficult to study in time domain.

- This is just like we use Laplace transform and z-transform a lot in engineering.

Levy's Convergence Theorem

- Let $\{X_n\}$ be a sequence of r.v.'s with cdf's F_n and characteristic functions ϕ_n. Let X be a random variable with cdf F and characteristic function ϕ that is continuous at $\theta = 0$. Then, X_n converges to X in distribution if and only if

$$\phi_n(\theta) \to \phi(\theta), \text{ as } n \to \infty, \forall \theta.$$

Central Limit Theorem (CLT)

- Let $X_1, X_2, \ldots$ be a sequence of i.i.d.r.v.'s with mean μ and variance $\sigma^2 < \infty$. Then,

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \rightarrow^d \quad N(0,1).$$

- **Note.** It can also be written as

$$\frac{\bar{X}_n - \mu}{Var(\bar{X}_n)} \rightarrow^d \quad N(0,1).$$

- This claims that $\bar{X}_n$ is approximately normal.

Implication

- WLLN and SLLN state that $S_n = \sum_{i=1}^{n} X_i$ is approximately equal to $S_n \approx n\mu$.

- CLT claims that $S_n \approx n\mu + \sqrt{n}\sigma N(0,1)$.

- This is a refinement.

Why true?

- By Levy's theorem, it suffices to show that the characteristic function of

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

converges to that of standard Normal when $n \to \infty$.

- What is the characteristic function of standard normal?

- WLOG, assume $E[X_1] = 0$ (otherwise we let X_i be $X_i - \mu$). We still have $Var(X_1) = \sigma^2$.

- Let $\phi(\theta) = E[e^{i\theta X_1}]$. Then

$$E\left[e^{i\theta \frac{S_n}{\sqrt{n}}}\right] = \left(\phi\left(\frac{\theta}{\sqrt{n}}\right)\right)^n.$$

- Since $\theta/\sqrt{n}$ is small when n is large, we apply Taylor expansion to obtain

$$\phi\left(\frac{\theta}{\sqrt{n}}\right) = \phi(0) + \phi'(0)\frac{\theta}{\sqrt{n}} + o\left(\frac{\theta}{\sqrt{n}}\right).$$

- Thus

$$E\left[e^{i\theta\frac{S_n}{\sqrt{n}}}\right] = \left(\phi\left(\frac{\theta}{\sqrt{n}}\right)\right)^n = \cdots$$