USC Marshall School of Business

DSO 699: Special Topics in Data Sciences and Operations
Spring 2024 – Stochastic Modeling for Optimization and Learning

Professor: Vishal Gupta
Meeting Time: T/Th 10 am – 11:30 am
Office: BRI 401
Email: None! Use Slack only
Office Hours: *By appointment only*

Course Description

This is a Ph.D.-level lecture course covering the core probability theory and stochastic analysis necessary for modeling and analysis in data-driven optimization, sometimes called "prescriptive analytics" or "decision-aware learning." As examples, consider the following questions:

- Suppose we fit a parametric demand model to some transaction data, and then optimize revenue to find a price. How suboptimal do we expect our price to be? Would doubling the amount of data substantively help?
- What if we fit a non-parametric demand model to our data instead? (How do we even do this?!?) When will this be better or worse than the parametric model?
- Suppose we have access to large collection of patient health data. Can we use this data to design personalized treatment plans for patients with specific diseases? How? How much data do we need to ensure our plans aren't unintentionally hurting our patients?
- Suppose we design a custom, blended ``estimation and optimization" methodology for prescriptive analytics. Can we say anything about the statistical properties of our approach? Is it better than the state of practice?

This course aims to provide students with a rigorous theoretical background to enable them to answer questions like these and pursue their own research in these areas. It will cover what has become a standard set of tools used to analyze these methods, including concentration inequalities, maximal deviations, generalization bounds, Reproducing Kernel Hilbert Space Methods (RKHS), and the basics of policy learning in causal inference settings.

This course is **not** meant as a first course in probability theory – students are expected to be familiar with concepts like conditional probability, tail bounds, and independence. Similarly, it is **not** a formal treatment of measure-theoretic probability. Rather, the focus is on the probability tools most commonly used to analyze algorithms in data-driven optimization, machine learning, and personalization. In this sense, the course is also application focused, with special attention to the common problems in these fields.

Learning Objectives

By the end of the course, student should be able to

- Use standard tail bounds and concentration results to analyze randomized algorithms and data-driven methods. In particular, they should be able to analyze "stable" data-driven optimization algorithms.
- Prove fundamental results in causal inference and policy learning.
- Prove fundamental results about sample average approximation (empirical risk minimization) including generalization guarantees and uniform laws of large numbers. In particular, students will analyze algorithms using metric entropy and VC-dimension.
- Analyze algorithms for contextual stochastic optimization based on data-driven estimates of conditional expectation.

Prerequisites and/or Recommended Preparation:

While there are no formal prerequisites for this course, students are strongly encouraged to take "DSO 699: Fundamentals of Probability for Data Science and Operations Research" in the fall semester or a similarly rigorous, proof-based introduction to probability. Measure theoretic probability is **not** strictly required. Any students concerned about their background ability should contact the instructor to discuss their situation.

Required Materials

There is **no** required textbook for the course. Lecture notes and recordings will be distributed through Blackboard/Slack.

That said, as a Ph.D. Class, you are highly encouraged to consult outside sources to supplement your learning as necessary. Some works I personally recommend:

- Asymptotic of Random Variables
  - *Asymptotic Statistics* by Van der Vaart (Esp Chapt. 1-2.5). This is a classic book and a good reference on these basics.
- Tail Bounds and Concentration of Measure
  - *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* by Martin Wainright. This treatment in this book is one of my favorites. I highly recommend it because it is fairly intuitive.
  - *Concentration Inequalities: A Nonasymptotic Theory of Independence* by Boucheron, Lugosi and Massart. (Esp. Chapt 1-2) This book is a bit more technical/terse than the Wainright book above, but also has some other techniques results that are worth learning for more advanced students. It also serves as an excellent reference.
- Empirical Process Theory
  - Pollard's Iowa Notes (Esp. Chapt. 1-7) - Available here: http://www.stat.yale.edu/~pollard/Books/Iowa/Iowa-notes.pdf. My treatment of uniform laws will largely follow this presentation with some deviations around the development of pseudo-dimension.
  - *Asymptotic Statistics (listed above)* (Chapt. 19): This is very terse, but covers the basics.
  - *Concentration Inequalities (listed above)*. This book gives a more "classical probability" treatment of empirical processes. Consequently, it covers ``more cases" than the Pollard treatment but is also more technical and a little less "user friendly." I recommend it as a reference or for more advanced students, but not as a first read.
  - Lecture Notes: There are also TONS of lecture notes online on this topic. Here are some that I like and use: https://www.shivani-agarwal.net/Teaching/E0370/Aug-2011/ and https://ambujtewari.github.io/teaching/LearningTheory-Spring2008/

Office Hours/Contacting Me:

I will not be using email for this course. We will use Slack to replace email. Hence, I will NOT respond to any emails sent to me about course materials. Please instead send me a Slack Message. Slack is available (for free) to all USC students, and you should automatically be enrolled in the class channel. If you aren't, send me a slack message, and I will add you. Familiarize yourself with Slack and how to use it on IT's website: https://cio.usc.edu/digital-campus-slack/

Occasionally, you may need to contact me about a private matter. In that case, please use the direct message feature. If upon reading your message, I deem it should be public, I might ask you to resend it to the public class slack channel so that all students can benefit from the question.

I will do my best to respond promptly to Slack messages. Some common emoji's we will use in this course are listed on our first pinned Slack message.

1-on-1 office hours are available at any time by appointment. Slack me and we'll schedule a time.

Please keep in mind that Slack is as much part of the academic environment of this course as is class time. Hence, please keep the language professional (but fun!). You know how to be a good citizen. Just do what you know.

Course Notes:

The course will be using its Slack channel extensively to distribute materials, lecture slides, make announcements. Part of the rationale is to encourage discussion among students to coordinate working on homework together, sharing materials, and in general, building their academic community.

The majority of the course will be lecture based. A rough list of topics and outline of the material is at the end of the syllabus; however, depending on the speed of the class and discussion, these topics are open to change. However, the precise dates of homework and exams will only change with substantive notice.

Grading Policies:

There will be several graded deliverables for the class:

- In-Class Participation: At the beginning of most classes I will write a simple problem on the board for you to think about, discuss, and informally write-up. You CAN and SHOULD work with classmates on this and turn in your own individual response on a slip of paper. I will use these slips to assign an in-class participation grade (based on effort).

- Slack Participation: Participation in discussion on Slack will contribute to your final grade.

- Homework Assignments will be approximately every two weeks. These are meant to be challenging and proof based. The clarity of your proof matters. You can work in groups on homework (see below for the group-work policy). Some homework may involve coding small simulations. **You cannot receive an A for the course unless you turn in every homework.**

- Homework: Assignments will be approximately every two weeks. These are meant to be challenging and proof based. The clarity of your proof matters. You can work in groups on homework (see below for the group-work policy). Some homework may involve coding small simulations. **You cannot receive an A for the course unless you turn in every homework.**

- Midterm: There will be one in-class midterm, open-book. The Midterm will consist entirely of 1) *Slightly* modified homework questions 2) Exercises that were given in class. If you've been following along with the homeworks, solutions and lectures, the midterm will be very straightforward. You may NOT work in groups on the Midterm.

- Final Exam: The final exam will have two parts. You may NOT work in groups for either part. The first part will be in-class and again consist entirely of 1) *Slightly* modified homework questions 2) Exercises that were given in class. This part should be very straightforward. The second part will be a take-home final exam (24 hours).

Policy on Group Work

Group discussion is STRONGLY encouraged throughout this class with other students in the class, but not with tutors or students outside the class. Throughout your PhD, your peers will always be your best resource. Use them. You may collaborate with other students on ANY of the above deliverables except the midterm and final.

However, you MUST always write up your own assignments individually and separately. (Thus, you can talk about a paper together, or even get a peer to read through your report and give you feedback, but you must incorporate that feedback on your own.) Please also list the names of students you collaborated with on the deliverable under your name, with a brief description of their contribution (if you deem it necessary).

For example, on my homework, I might write:

> Collaborated with: John Snow (Problem 1 and 2), Sansa Stark (Problem 3), Tyrion Lannister (entire assignment)

Grading Breakdown:

Assignment	**% of Total Grade**
Participation/Discussion	10%
Homework	50%
Midterm	20%
Final Exam	20%
Total	100%

Synchronous session recording notice

All sessions of the course will be recorded and provided to all students enrolled in the course (and officially auditing non-USC students) via Slack. Consequently, it is also important that students respect USC's policy and do NOT share any of the course content outside the course. This includes recordings, lecture notes, or other materials. For clarity, from SCampus:

SCampus Section 11.12(B)

Distribution or use of notes or recordings based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study is a violation of the USC Student Conduct Code. This includes, but is not limited to, providing materials for distribution by services publishing class notes. This restriction on unauthorized use also applies to all information, which had been distributed to students or in any way had been displayed for use in relationship to the class, whether obtained in class, via email, on the Internet or via any other media. (SeeSection C.1 Class Notes Policy).

ADDITIONAL INFORMATION

USC Statement on Academic Conduct and Support Systems

Explanation - This section, or an enhanced version, is required by the University. You are free to enhance the content as you deem necessary within the structure of the following.

Academic Conduct:

Students are expected to make themselves aware of and abide by the University community's standards of behavior as articulated in the Student Conduct Code. Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Part B, Section 11, "Behavior Violating University Standards" https://policy.usc.edu/scampus-part-b/. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

Support Systems:

It is important to recognize that distance learning is hard. The online platform is difficult, but being isolated, especially in a process as challenging as Ph.D., is also hard. Please look out for one another. If you are feeling overwhelmed, reach out. You may always reach out to me or to your classmates. In other circumstances, you might feel more comfortable reaching out to one of the resources below.

Counseling and Mental Health - (213) 740-9355– 24/7 on call
https://studenthealth.usc.edu/counseling/
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL), press "0" after hours – 24/7 on call
https://studenthealth.usc.edu/sexual-assault/
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED)- (213) 740-5086 | Title IX – (213) 821-8298
equity.usc.edu, titleix.usc.edu
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
https://usc-advocate.symplicity.com/care_report/
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC is committed to making reasonable accommodations to assist individuals with disabilities in reaching their academic potential. If you have a disability which may impact your performance, attendance, or grades in this course and require accommodations, you must first register with the Office of Disability Services and Programs (www.usc.edu/disability). DSP provides certification for students with disabilities and helps arrange the relevant accommodations. Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me (or to your TA) as early in the semester as possible. DSP is located in GFS (Grace Ford Salvatori Hall) 120 and is open 8:30 a.m.–5:00 p.m., Monday through Friday. The phone number for DSP is (213) 740-0776. Email: ability@usc.edu.

USC Campus Support and Intervention - (213) 821-4710
https://uscsa.usc.edu/
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101
diversity.usc.edu
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
dps.usc.edu Non-emergency assistance or information.

COURSE CALENDAR

The details of the course calendar are subject to change depending on the pace of the class.

Sessions 1-2: Basic Probability Review

- Independence, Properties of Normal and Poisson Random Variables
- Central Limit Theorem and Dominated Convergence Theorems
- Big O_p and Little o_p

Sessions 3-7: Tail Bounds and Concentration

- Chernoff Technique
- Properties of SubGaussian Random Variables
- Convergence of SAA/ERM for Finite Discrete Feasible Regions
- Symmetrization
- SubGamma Random Variables, Bernstein's Inequality
- "Fast Rates" of Convergence for Classification

Sessions 8-10: Functions of Independent Random Variables and Algorithmic Stability

- Efron-Stein, McDiarmid, McCombes Inequalities
- Uniform Hypothesis Stability
- Applications to ERM for Convex Optimization and Combinatorial Optimization Problems
- Advanced Stability Results via the Entropy Method

Session 11, 12, 14: Stochastic Gradient Descent

- Basic Convergence Results and Differences to ERM
- SGD in the "Interpolable" Setting
- SGD and Algorithmic Stability

Session 15-21: Uniform Laws of Large Numbers

- Packing/Covering in Euclidean Spaces
- Pseudo-dimension: Definition, Examples
- Relationship between Pseudo-Dimension and Packing #'s
- Rademacher Complexity
- ULLN's for the Optimizer
- Applications to ERM, Plug-In Methods, Empirical Bayes

Session 22 – 25: Contextual Stochastic Optimization

- The Modeling Paradigm and Differences from ERM
- Structured Learning and Associated Challenges
- Learning Conditional Distributions Directly
- The Predict-then-Optimize Framework
- Calibration and Plug-In Policies
- Applications

Session 26 – 29: Causal Inference and Policy Learning

- The Neyman-Pearson Paradigm
- The Optimizer's Curse and Other Common Fallacies
- Direct Estimators, IPW and Doubly Robust Estimators
- Applications

Homework Due Dates

- HW 1: 18 Jan
- HW 2: 1 Feb
- HW 3: 15 Feb
- HW 4: 12 Mar
- HW 5: 2 April
- HW6: 16 April

Homework due at beginning of class, LaTeXed. Remember the policy above on group work.

Midterm Exam (In-Class): 27 February

Midterm only covers material up through and including 15 Feb, assuming we stick to syllabus. If we adjust topics slightly, coverage may (slightly) change.

Final Exam (In-Class): 30 April

Final Exam will only cover material up to and including 23 April, assuming we stick to syllabus. If we adjust topics slightly, coverage may (slightly) change.

Final Exam (Take-Home, 24 hours): Date TBD

We'll decide jointly as a class on a good date for this.