Summer 2022 – (Some) Stochastic Foundations of Decision-Aware Learning
Prof. Vishal Gupta (Visiting)

Professor: Vishal Gupta
Meeting Time: T/Th 9:30 – 11:30 am
Room: See below for details
Email: guptavis@usc.edu
Office Hours: *By appointment only*
Course Materials: https://bit.ly/3dqqz1c

Meeting Time:
There was some confusion about the timing of the course, and I apologize. Going forwards, we will start at 9:30 am. Although we've scheduled 2 hours for each session, I am hoping to only speak for 1 hour to 1 hour 30 minutes. (The remaining time is for you to ask questions, meet with me, or work together on homework problems.) Notice, because of the Analytics for X conference, we are not having a session on 22 Sept. Also, because of scheduling conflicts, our meeting room changes in some instances. Please consult schedule below as needed:

Date	Room
20 Sept.	BIZ1 0302
27 Sept.	BIZ2 0509
29 Sept.	BIZ1 0307
4 Oct.	BIZ2 0509
6 Oct.	BIZ1 0307

Background

The "estimate-then-optimize" framework is perhaps the most fundamental paradigm for utilizing data in operations research and analytics. The idea is simple: First, use the best statistical methods available to fit a good model to your data. Second, solve an optimization problem based on that model to identify the best decision. Note, the first estimation step is typically completely agnostic to the structure of the second step's optimization problem. This setup often mirrors operations in real firms, where one team conducts model fitting tasks (e.g., demand forecasting) whereas a second downstream team uses those models for tactical decision-making.

Although intuitive, recent work has highlighted shortcomings of the estimate-then-optimize framework. In particular, when data are too limited to fit a *very* accurate model in the first step, the induced decisions in the second step often perform poorly in practice.

Consequently, there has been a recent explosion in research around so-called ``decision-aware" learning methods that tailor the first-stage fitting to the second stage optimization problem. Such methods go by a variety of names – end-to-end learning, decision-focused learning, optimization-aware, blended estimation and optimization, operational statistics, etc. Preliminary empirical results strongly suggest that such methods can vastly outperform estimate-then-optimize approaches – especially when data are scarce, limited, or very high-dimensional. However, a full theoretical understanding of the potential benefits (and drawbacks) of these decision-aware approaches is still forming.

This is a Ph.D. level mini-course surveying the core probability theory and statistical tools necessary for theoretically analyzing many popular data-driven approaches, which particular emphasis on decision-aware learning. Given the short format, we will likely focus more on developing deep intuition and learning how to apply these tools to prove performance guarantees for commonly used methods and real-world applications and may defer some technical development to the cited (excellent) references.

Consequently, this course is **not** meant as a first course in probability theory – students are expected to have strong familiarity with concepts like random variables, expectation, and conditional probability. Similarly, it is **not** a formal treatment of measure-theoretic probability. Rather, the focus is on the probability tools most used to analyze algorithms in data-driven optimization. Finally, students are expected to have a good working knowledge of convex optimization (duality, first order conditions, etc.)

Learning Objectives

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

- Use standard tail bounds and concentration results to analyze data-driven methods.
- Prove fundamental results for uniform laws of large numbers based on metric entropy and VC-dimension.
- Use these tools to prove performance guarantees for sample average approximation (empirical risk minimization) and several (more recent) decision-aware learning methods.

Required Materials

There is no required textbook for the course. Lecture notes will be distributed and I may provide citations to certain papers for technical results.

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

- *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* by Martin Wainright (Esp. Chapt. 2, 4, and 5).
  This treatment in this book is my favorite and the exposition is ideal for researchers somewhat new to concentration and uniform laws of large numbers.
- Pollard's Iowa Notes (Esp. Chapt. 1-7) - Available here:
  http://www.stat.yale.edu/~pollard/Books/Iowa/Iowa-notes.pdf .
  Pollard's beautiful exposition goes a long way to "demystifying" some of the otherwise unintuitive Rademacher complexity results in uniform laws of large numbers. His approach is somewhat different than the standard approach in the Wainright book, however, I feel it is a bit more flexible (e.g., he treats empirical processes formed by independent, but not identically distributed, random variables). Indeed, whenever I need to prove a result outside the ``typical" assumptions used in the field, I consult this work.
- *Concentration Inequalities: A Nonasymptotic Theory of Independence* by Boucheron, Lugosi and Massart. (Chapt 1-3, 12, 13)
  It is an **excellent** advanced reference for researchers intending to work in the field. This book is a bit more technical/terse than the Wainright book above making it difficult for the beginner. I wouldn't describe it as "user-friendly," but it has a broad variety of techniques, useful if you just need to quote a standard result. The treatment of the entropic method is infinitely better than anything else I've seen, and the treatment of empirical processes that is far more indepth than the Wainwright book, although I've rarely needed the extra generality.
- *Asymptotic Statistics* by Van der Vaart (Chapt. 19): This is very terse, but covers the basics.
- Other Class 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/

Prerequisites and/or Recommended Preparation:

There are no formal prerequisites for the course but students should be familiar with basics of probability theory (random variables, expectation, probability density functions, moment generating functions, combinatorial probability). Measure theoretic probability is **not** required. A good working knowledge of optimization (first order conditions, duality, Lipschitz objectives, etc.) and elementary analysis (limits of functions, big Oh notation, linear algebra) is needed. Any students concerned about their background ability should reach out to the instructor to discuss their particular situation.

Homework:

I will assign a small homework assignment after *EVERY* session. Collaborating with other students in the class on the homework is STRONGLY encouraged with other students in 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 homework assignments.

However, you MUST always write up your own assignments individually and separately. (Thus, you can talk about a problem together, or even get a peer to read through your solution 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)

Homework will be graded as “Credit /No Credit.”

Grading

This course will be graded Pass/Fail.

COURSE CALENDAR

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

Sept 20: Session 1: Motivating Proxy-Objectives in Decision-Aware Learning

- Course Overview and Some Motivating Examples
- Types of Performance Guarantees for Optimization

Sept 27: Session 2: Basic Concentration Inequalities

- Review of Chernoff Bounding Techniques
- Hoeffding, SubGaussian and Bernstein Inequalities
- McDiarmid's Inequality

Sept 29: Session 3: Maximal Discrepancy Approach to Analyzing Surrogate Losses

- Application: SAA with Discrete, Finite Feasible Regions
- Application: Unbiased Model-Selection for Denoising

Oct 4: Session 5: Rademacher Complexity and ULLN

- Sums of Independent Random Variables
- Classical Rademacher Complexity Bounds on Suprema of Processes
- Application: Revisiting SAA in the Large-Sample Regime

Oct 6*:* Session 6: Small-Data, Large-Scale Linear Optimization

- Challenges of Small-Data Regime
- Debiasing In-Sample Performance
- Best-in-Class Performance Guarantee