Topics in Stochastic Optimization
ENG EC 700 Spring 2025
Instructor: Ashok Cutkosky email: cutkosky@bu.edu office: PHO 420

Class meets:
Course website:
Piazza signup link:
Gradescope course ID:

Office Hours

Course description:
This course will explore recent advancements in optimization theory, with a particular focus on "gradient-based" optimization techniques that are prevalent in stochastic and online optimization problems arising in machine learning as well as techniques for tuning hyperparameters and dataset properties for improved convergence. The class will emphasize connections between optimization algorithms and disparate areas such as statistics and high-dimensional probability and architecture choices in deep learning. Students will read contemporary literature throughout the class. This course will focus on mathematical fundamentals rather than implementation details.

Prerequisites:
A strong background in linear algebra and general familiarity with basic concepts in machine learning such as kernel methods and neural networks. The course will also make use of some concepts from high dimensional statistics and concentration of measure, such as the Johnson-Lindenstrauss and Hoeffding lemmas.

Courses that provide these requirements include: EC 719, SE 714, EC 505, EC 503, EK 500

The class will include a brief review of background concepts accessible to mathematically mature students.

Coursework:
During the semester we will assign weekly readings that must be completed before lecture.
There will be no regular homework, but there will be a final project turned in at the end of class.
Final projects may be completed in groups of up to three students.
A list of potential final project ideas will be provided during class.

More detailed Course, Project and Grading Information:

- Each week, several papers will be assigned for reading.

- Sometimes, papers will be accompanied by notations suggesting to only read a certain subset of the paper.
- Several questions will be provided for each paper.
- I expect you to read the papers and think about the questions. You are not required to write up answers to the questions.
- During the next week's classes, we will discuss the papers.
    - of class grade will be determined by participation in discussion about the papers.
- I will call upon students to give answers to the provided questions.
- answering the questions is a straightforward way to earn participation credit.
- You can also earn credit by other meaningful participation.
- The rest of the grade will be for a project.
- The project can be done in teams. You can choose teams however you like, but for groups of size greater than 3, you must have a plan for how everyone will contribute.
- Group projects can be along the following lines:
    - Choose a paper, and improve upon it in some way.
    - Implement the algorithm in a paper. Find out if it performs well in practice. Propose and test some alterations to the algorithm that you hope will make it better.
    - Come up with a problem in stochastic optimization and solve it.
    - Provide a simplified proof of a result in a paper, or explain the results in a simpler manner. This is best used for papers that present important results, but are very complicated/hard to read. Such papers frequently are more complicated than they need to be.
    - Anything else, with instructor approval.
- Projects and teams must be approved by the instructor. Current target is to make these decisions 3/18 (just after spring break).
- A completed project consists of: (1) a written project report, (2) an oral presentation that will be presented to the rest of the class. The time alloted to each presentation will depend on the total number of teams.
- The presentations will take place in the final class. Reports will be due at the same time.

List of Potential Topics:

- Stochastic Gradient Descent
- Online optimization
- Empirical concentration bounds
- Hyperparameter optimization
- Learning rate tuning and scheduling
- Beyond first-order optimization with stochastic and/or non-convex losses
- Universal adaptive optimizers.
- Dataset selection
- Robust optimization
- Connections between deep learning architectures as optimization

Textbook:
There is no textbook for this course.

Other resources that will be helpful at times in the class:
“Convex Optimization: Algorithms and Complexity”: https://arxiv.org/pdf/1405.4980.pdf
Covers convex optimization algorithms for the non-stochastic case. Primarily only the first few chapters will be useful to us.

“A Modern Introduction to Online Learning”: https://arxiv.org/abs/1912.13213
Covers convex optimization for stochastic settings (in fact, for a more difficult setting called the “online” setting). Again, primarily only the first few chapters are covered in this class.

“Distributionally robust learning”:
https://www.nowpublishers.com/article/DownloadSummary/OPT-026

Example papers and algorithms that may be discussed include:

Online Convex Programming and Generalized Infinitesimal Gradient Ascent

Black-Box Reductions for Parameter-free Online Learning in Banach Spaces

Acceleration by Stepsize Hedging II: Silver Stepsize Schedule for Smooth Convex Optimization

Tight Concentrations and Confidence Sequences From the Regret of Universal Portfolio
Estimating means of bounded random variables by betting
Online-to-PAC Conversions: Generalization Bounds via Regret Analysis

A simpler approach to accelerated optimization: iterative averaging meets optimism

“Convex Until Proven Guilty”: Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions

Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer

Random Features for Large-Scale Kernel Machines

Training Compute-Optimal Large Language Models

Shampoo: Preconditioned Stochastic Tensor Optimization

Second-Order Information in Non-Convex Stochastic Optimization: Power and Limitations

The Road Less Scheduled

DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining

Variance-based Regularization with Convex Objectives

Transformers learn to implement preconditioned gradient descent for in-context learning

Curriculum learning: A regularization method for efficient and stable billion-scale gpt model pre-training

Stacking as Accelerated Gradient Descent

Online Control with Adversarial Disturbances