CS 224 - Algorithms for Data Science (Fall '23)

Schedule	MW 2:15-3:30 in SEC LL2-224
Instructor	Sitan Chen ⟨sitan@seas.harvard.edu⟩
	Office hours: Thursday 4-5, SEC 3.325
TFs	David Brewster ⟨dbrewster@g.harvard.edu⟩
	Depen Morwani ⟨dmorwani@g.harvard.edu⟩
	Rosie Zhao ⟨rosiezhao@g.harvard.edu⟩
	Weekly OH's: Monday 11-12, see course page for location
	Pset deadline OH's: TBD
Course page	https://sitanchen.com/cs224-f23.html
Canvas (lecture videos)	https://canvas.harvard.edu/courses/131597
Gradescope	https://www.gradescope.com/courses/569639

COURSE DESCRIPTION AND LEARNING OBJECTIVES

This is a graduate-level topics class on algorithmic challenges arising in modern machine learning and data science more broadly. The course will touch upon a number of well-studied problems (generative modeling, deep learning theory, adversarial robustness, inverse problems, inference) and frameworks for algorithm design (gradient descent, spectral methods, tensor/moment methods, message passing, convex programming hierarchies, Markov chains). As we will see, proving rigorous guarantees for these problems often draws upon a wide range of techniques from stochastic calculus, harmonic analysis, random matrix theory, algebra, and statistical physics. We will also explore the myriad modeling challenges that go into building theory for ML and discuss prominent paradigms (semi-random models, smoothed analysis, oracles) for going beyond traditional worst-case analysis.

The following is a tentative schedule for the course:

(1) **Tensor methods**:
   (a) Intro to tensors, Jennrich's, tensor power method
   (b) Applications (cryo-EM, ICA, superresolution)
   (c) Overcomplete tensor decomposition, smoothed analysis

(2) **Sum-of-squares algorithms for learning**
   (a) SDPs, pseudodistributions, SoS proofs
   (b) Application 1: tensor decomposition
   (c) Application 2: outlier-robust regression

(3) **Robustness**:
   (a) Robust mean estimation, iterative filtering
   (b) List-decodable learning
   (c) Semirandom noise models, robust classification

(4) **Supervised learning**:
   (a) Discrete domains: Fourier analysis, noise stability
   (b) Continuous domains: Hermite analysis, tensor problems, CSQ lower bounds
   (c) Filtered PCA
   (d) Analyzing gradient dynamics: NTK and beyond

(5) **Computational complexity**:
   (a) Statistical query lower bounds, parities, moment-matching
   (b) Cryptographic lower bounds for supervised problems (public key, pseudorandom functions, Daniely-Vardi lifting)
   (c) Cryptographic lower bounds for unsupervised problems (continuous learning with errors)

(6) **Inference and statistical physics**
   (a) Community detection, intro to belief propagation, free energy
   (b) Kesten-Stigum, nonbacktracking operator, computational-statistical gaps
   (c) From belief propagation to approximate message passing
   (d) State evolution, connection to variational inference
(7) **Generative modeling**:
   (a) Langevin dynamics, diffusion model basics, Girsanov's theorem
   (b) Bayesian posterior sampling with diffusion models
   (c) Deterministic samplers, acceleration

The goal is that by the end of the course, students will be sufficiently up to date with the modern literature on theory of ML that they are ready to engage in original research.

PREREQUISITES

Strong mathematical maturity and proficiency with proofs, probability (especially over continuous domains), and linear algebra are highly recommended. Prior coursework at the level of Math 22 and Stat 110 (or equivalent) is required. The course attendance will be capped, so a lottery based on a "pset zero" and a course application form, to be released during the course registration period, will be used to determine placement in the course.

COURSE MATERIALS.

The schedule on the course webpage will include a list of relevant papers for each lecture. We will not be following any particular textbook, but the student may find the following courses previously offered at other institutions helpful for or complementary to various parts of the material:

- Ankur Moitra. Algorithmic Aspects of Machine Learning
- Tselil Schramm. The Sum-of-Squares Algorithmic Paradigm in Statistics
- Prasad Raghavendra. Efficient Algorithms and Computational Complexity in Statistics
- Sanjeev Arora. Theory of Deep Learning
- Song Mei. Mean Field Asymptotics in Statistical Learning
- Lenka Zdeborova & Florent Krzakala. Statistical Physics For Optimization and Learning
- Ahmed El-Alaoui. Topics in High-Dimensional Inference
- Subhabrata Sen. STAT 217: Topics in High-Dimensional Statistics - Methods from Statistical Physics

COURSE FORMAT

Each class will be a 75-minute whiteboard lecture by the instructor, with occasional use of powerpoint slides.

COURSEWORK AND GRADING

- 0% problem set 0 (math background check)
- 15% class participation
- 15% scribing requirement (2 lectures)
- 40% problem sets 1-4 (biweekly)
- 30% final project.

Class participation will be evaluated holistically based on how actively the student engages in discussion in class and on Ed.

Students must sign up to scribe one lecture. Instructions will be posted on the course page at the start of the semester.

The problem sets will be challenging, so you are encouraged to start them early. There will be a primary office hour, to be hosted concurrently by the teaching fellows, the week of every problem set deadline.

For the final project, you will have the option of an expository project or an original theoretical research project. For the former, you have the additional option of writing a traditional survey paper, or writing a series of technical blog posts. The instructor will provide a list of suggested topics, but students are welcome to pick a topic not listed, subject to instructor approval.

EXPECTATIONS AND COURSE POLICIES

- In-personal attendance at **lectures** is strongly recommended.
- **Office hours** are intended for students both to get help with coursework and to engage more deeply with the material.
- **Assignments and collaboration policy:** All assignments will be submitted on Gradescope, so students should either have a scanner available or become familiar with LaTeX.

  The student is responsible for understanding Harvard policies on academic integrity. For problem sets, students are encouraged to collaborate with each other, but they must write their final solutions independently and list their collaborators in their submissions. Final projects are to be completed independently without collaboration. For all assignments, it is acceptable to consult outside sources, but the student must cite whatever is used and synthesize the information in these references in their own words. Use of generative AI for assignments is prohibited.
- **Late days:** Students have 5 late days in total for the semester, which may be used for the problem sets. For exceptions, students must have their senior tutor (for undergrads) or their advisor (for graduate students) contact me.
- **Accommodation requests:** We acknowledge the value of every individual's unique perspective and experiences. If you ever feel hesitant to share your thoughts openly in class, or if something was said in class that made you uncomfortable (either by us or anyone else), please do not hesitate to reach out. The same goes if you find that external experiences are impacting or have the potential to impact your performance in the course. The University Disability Office also offers accommodations and services for students with documented disabilities. We will do our best to create an inclusive and supporting learning environment that respects accessibility and promotes diversity, inclusion, and belonging, and we welcome any and all feedback if you feel there are areas in which we can improve.
- **Student well-being:** We deeply care about your physical and mental well-being. In case you run into any problems in this course or feel that you are falling behind due to external circumstances, please don't hesitate to reach out to the course staff or your resident dean. Other resources we recommend taking advantage of in such cases include Harvard services such as Counseling and Mental Health Services, Room 13, and the Academic Resource Center. Additionally, if you have a serious emergency, medical or otherwise, please contact the instructor. In all of these cases, we will make sure to accommodate you as best as possible.