CS 2243 – Algorithms for Data Science (Fall '24)

Schedule	MW 2:15 - 3:30, SEC 1.413
Professor	**Sitan Chen** ⟨sitan@seas.harvard.edu⟩
	Office hours: Monday 5:30-6:30, SEC 3.325
TFs	**Weiyuan Gong** ⟨wgong@g.harvard.edu⟩
	Marvin Li ⟨marvinli@college.harvard.edu⟩
	Weekly office hours: Tuesday 4-5, SEC 3.317 (Weiyuan), Friday 4-5, SEC 4.307 + 4.308
Course page	https://sitanchen.com/cs224/f24/index.html
Canvas (lecture videos)	https://canvas.harvard.edu/courses/136969
Gradescope	https://www.gradescope.com/courses/828638
Edstem	https://edstem.org/us/courses/62981/discussion/

Course Description and Learning Objectives

This is a graduate topics class on algorithmic challenges in modern machine learning and data science. We will touch upon a number of domains (generative modeling, deep learning theory, robust statistics, Bayesian inference) and frameworks for algorithm design (spectral/tensor methods, gradient descent, message passing, MCMC, diffusions), focusing on provable guarantees. The theory draws upon a range of techniques from stochastic calculus, harmonic analysis, statistical physics, algebra, and beyond. We will also explore the myriad modeling challenges in building this theory and prominent paradigms (average-case complexity, smoothed complexity, oracles) for going beyond traditional worst-case analysis.

The following is a tentative schedule. Topics labeled with an asterisk are new to this year's offering of the course.

(1) **Tensor methods**:
   (A) Intro to tensors, Jennrich's algorithm, applications (super-resolution, Gaussian mixtures, ICA)
   (B) Iterative methods for tensor decomposition
   (C) Smoothed analysis
(2) **Sum-of-squares**:
   (A) Sum-of-squares proofs, pseudo-distributions
   (B) Applications to robust statistics
   (C) Mixtures of spherical Gaussians, certifiable hypercontractivity
(3) **Supervised learning**:
   (A) PAC learning, generalized linear models, semirandom noise
   (B) Low-degree algorithm, polynomial regression, noise sensitivity
   (C) Hermite analysis, CSQ algorithms
   (D) Filtered PCA
   (E) Linearized networks (NTK, mean field)
(4) **Computational complexity**:
   (A) Hardness basics, CSQ lower bounds, noisy parity
   (B) SQ and statistical dimension
   (C) Cryptographic lower bounds, Daniely-Vardi lifting
(5) **Statistical physics**:
   (A) Graphical models and variational inference
   (B) Belief propagation, Bethe free energy
   (C) Derivation of AMP, state evolution, $\mathbb{Z}_2$ synchronization
   (D) Optimality of AMP for $\mathbb{Z}_2$ synchronization
(6) **Sampling and generative modeling**:
   (A) *Stochastic calculus basics, Langevin dynamics

(B) *Acceleration
(C) Diffusion models, Girsanov's theorem
(D) *Probability flow ODE
(E) *Algorithms for score estimation

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 required. Prior coursework at the level of Stat 210 (or equivalent) is strongly recommended.

Course Materials

The schedule on the course webpage will include a list of relevant papers for each lecture, as well as a set of preliminary lecture notes. 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. `http://people.csail.mit.edu/moitra/408b.html`. Algorithmic Aspects of Machine Learning.
- Tselil Schramm. `https://tselilschramm.org/sos-paradigm/sos-paradigm.html`. The Sum-of-Squares Algorithmic Paradigm in Statistics.
- Sam Hopkins. `https://www.samuelbhopkins.com/teaching/sos-fall-22/sos-fall-22.html`. The Sum of Squares Method.
- Prasad Raghavendra. `http://people.eecs.berkeley.edu/~prasad/fall2022.html`. Efficient Algorithms and Computational Complexity in Statistics.
- Sanjeev Arora. `https://www.cs.princeton.edu/courses/archive/fall19/cos597B/lecnotes/bookdraft.pdf`. Theory of Deep Learning.
- Song Mei. `https://www.stat.berkeley.edu/~songmei/Teaching/STAT260_Spring2021/`. Mean Field Asymptotics in Statistical Learning.
- Lenka Zdeborova & Florent Krzakala. `https://sphinxteam.github.io/EPFLDoctoralLecture2021/`. Statistical Physics For Optimization and Learning.
- Ahmed El-Alaoui. `https://courses.cit.cornell.edu/stsci6940/`. Topics in High-Dimensional Inference.
- Subhabrata Sen. `https://canvas.harvard.edu/courses/119055`. STAT 217: Topics in High-Dimensional Statistics - Methods from Statistical Physics..
- Kevin Tian. `https://kjtian.github.io/cs395t.html`. Continuous Algorithms..
- Mark Sellke. `https://msellke.com/courses/STAT_291/course_page_website.html`. Random High-Dimensional Optimization: Landscapes and Algorithmic Barriers..
- Sam Hopkins & Costis Daskalakis. `https://hackmd.io/@QkEI9EXuQp2xal2TYj7T2w/SklMHfUkT`. Algorithmic Statistics.

Course Format

Each class will feature a combination of powerpoint slides and whiteboard lecture.

Coursework and Grading

- 0% problem set 0 (math background check)
- 15% class participation
- 15% lecture note editing (1 lecture)
- 40% problem sets 1-4 (biweekly)

- 30% final project

Class participation will be evaluated holistically based on the student's level of engagement in discussions in class and on Ed. **In-personal attendance at lectures is mandatory to receive full participation credit.** In the event of unavoidable conflicts (illness, travel, etc.), the student will not be penalized provided that they write to the instructors prior to the lecture in question.

Students must sign up either to scribe one of the lectures marked by an asterisk, or to edit the existing lecture notes of some other 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. In the middle of the semester, the instructor will post a list of suggested topics and papers, but students are welcome to pick a topic not listed, subject to instructor approval.

Expections and Course Policies

- **Office hours** are optional but 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 strongly discouraged, but if the student chooses to use it, they must provide a full transcript of the prompts used and write the final solutions in their own words.
- **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.