ECE 8803: High-dimensional statistics, signal processing, and optimization

Course syllabus

Course summary and philosophy

This graduate course will cover high-dimensional statistics and optimization, and students will be introduced to modern **analyses** of (one-shot and iterative) algorithms in high-dimensional statistical learning and signal processing. We will cover analytical tools from probability and optimization, and then dive into analyzing M-estimators and iterative methods in statistical models. Students will be introduced to convex relaxation and its analysis. The key takeaway for iterative algorithms will be that they often converge much faster in random ensembles than is predicted by worst-case theory. Finally, we will explore some modern algorithmic ideas, some of which have originated in the statistical physics literature. Our focus will be on presenting rigorous guarantees on various methods that are now commonplace in high-dimensional statistics and statistical signal processing, and along the way we will introduce several applications in which they are broadly useful.

Learning outcomes

Upon successful completion of the course, you will have learned:

(a) A variety of modern algorithms for high-dimensional data analysis problems

(b) Several ways to model and exploit structure in high dimensions

(c) The design principles of convex relaxation and iterative optimization

(d) Several analysis tools from optimization and probability—including dual witness techniques and Gaussian process theory—to prove rigorous guarantees on modern methods

Preparation and prerequisites

Students will be expected to have a solid knowledge of undergraduate probability and statistics, linear algebra and multivariable calculus, and basic optimization. The most important prerequisite is mathematical maturity and familiarity with understanding and writing proofs.

Course staff

Instructor:

- **Ashwin Pananjady (ashwinpm@gatech.edu):** Ashwin is a faculty member with a joint appointment between ISyE and ECE at Georgia Tech. He got his Ph.D. in EECS at UC Berkeley, and then spent a semester as a postdoctoral research fellow at the Simons Institute for the Theory of Computing. He studies high-dimensional statistical methods as well as optimization theory and algorithms, and works on problems arising from the intersection of these disciplines with machine learning, reinforcement learning, and signal processing.

Scheduled live meetings (tentative)

Lecture time:	MW 9.30-10.45am
Lecture location:	Ford Environmental Science and Tech L1175
Instructor OH:	TBD

Grading

The course will be graded on the following components:

- **Homework (50%)**: There will be a total of 4-5 homework assignments, each with 3-5 questions.
- **Midterm (20%)**: (Tentative) There will be a **take home midterm** for this class that will be active for 48 hours. A final decision about whether to have a midterm will be made after enrollment stabilizes; if we decide not to have a midterm, points will be transferred to homework.
- **Course project (30%)**: We will have a project in which students study in depth a topic of their choosing related to the course. Students can form groups of up to 2 people.

Course expectations and guidelines

Academic integrity Georgia Tech aims to cultivate a community based on trust, academic integrity, and honor. Students are expected to act according to the highest ethical standards. For information on Georgia Tech's Academic Honor Code, please visit `http://www.catalog.gatech.edu/policies/honor-code`. Any student suspected of cheating or plagiarizing on a quiz, exam, or assignment will be reported to the Office of Student Integrity, who will investigate the incident and identify the appropriate penalty for violations. This is an advanced undergraduate class; we expect that you are here to learn something new and that you will be conscientious and conduct yourself with courtesy and integrity.

Redistributing materials for this course and/or using external sites for assistance (e.g. contributing to test banks, CourseHero, Chegg, or similar sites) is prohibited.

Collaboration and group work Students are strongly encouraged to discuss homework problems with one another. However, *each student must write up and turn in their own solutions, written in their own words/consisting of their own code. Cases where written solutions or code appear to be identical or nearly identical will be immediately referred to the Office of Student Integrity.*

Absences/late submissions Out of fairness to the entire class, *late submission of homework will not be accepted in the absence of a prior agreement between the student and instructor.* In particular, *excused absences* include illnesses, religious observations, career fairs and job interviews. In the event than an excused absence such as above prevents a student from submitting an assignment, their homework grade will be calculated on a prorated basis.

Etiquette for questions: Aside from questions you have in our in-person meetings, please put all technical questions on Piazza. We will make an effort to curate this material with separate threads for each problem so that we enable peer-to-peer question answering to the best extent possible. We will try our best to answer your question during the day in which it is asked, but please help each other out on this front.

For logistical questions, please create a separate thread on Piazza between you and the instructor and ask it there. If you have an urgent question that needs to be answered immediately, please email the instructor with subject line beginning with [8803].

Accommodations for students with disabilities If you are a student with learning needs that require special accommodation, contact the Office of Disability Services at (404) 894-2563 as soon as possible, to make an appointment to discuss your special needs and to obtain an accommodations letter. Please also email the instructor ASAP to discuss your learning needs.

Student-faculty expectations agreement At Georgia Tech, we believe that it is important to strive for an atmosphere of mutual respect, acknowledgement, and responsibility between faculty members and the student body. In the end, simple respect for knowledge, hard work, and cordial interactions will help build the environment we seek. Therefore, I encourage you to remain committed to the ideals of Georgia Tech while in this class. See `http://www.catalog.gatech.edu/rules/22` for an articulation of some basic expectation that we can have of each other.

Support for student health and well-being: This is still an unprecedented time, so please be kind to yourself. I applaud your resolve and drive in choosing to learn something new this semester. Make sure to eat well, exercise, and reach out to your support system or the course staff if you need to.

Homework

There will be a total of 4-5 HW assignments in this class.

Rough schedule. Homework will be assigned approximately once in three weeks, with deadlines typically on Wednesday. You will have a total of 6 late days that you can use on HW throughout the semester (you can use all of them on one HW, or split them between multiple HW). These late days are meant to account for sickness, travel, and other commitments. Beyond these 6, there will be no extra late days.

Solutions and grading. Solutions for the HW will be released 7 days after HW deadline, and you will need to **self-grade** your HW. In particular, you will be expected to cross-check your answers with the released solutions, and hand in on Canvas (one week after the solutions are released) a detailed self-grade for the HW. Your HW and self-grade will be checked to ensure that credit is being assigned appropriately.

Project

Proojects can be done in groups of size *up to two students* (no exceptions). The final course project is intended to give students the opportunity for in-depth exploration of a topic in modern statistical signal processing and data analysis. Course projects could look like one of the following:

- An in-depth survey of one of the topics covered in the class. A survey consists of a rigorous academic review of the literature related to the topic interpreted *in the student's own words*, and a possible discussion of future areas of research.
- An application of the algorithms discussed in class on a data set or an engineering application. *The application can be non-standard; in fact, proposing new applications for the material developed in class is encouraged.* However, the methodology in the implementation needs to involve techniques that we discuss during the class.
- A conceptual (i.e. either theoretical or simulation-based) topic of novel research related to the class. *Initial results and directions for future work are common outcomes.*

Project proposals that do not neatly fall into one of these categories are also welcome. Students will be asked to submit a short abstract on their course project topic in late September.

Course materials

We will maintain course materials on Canvas. This will provide general course information, links to supplementary reading materials, and links to homework assignments. Homework assignments and solutions will also be posted on Canvas as they are made available.

The instructor will make exclusive use of Piazza to make announcements and answer questions. Piazza will also be curated to foster discussion about problems. Piazza is a great forum to discuss problems, find study groups, etc. Please direct any questions you might have about the course to Piazza. To the maximum extent possible, please do not make private Piazza posts: if you have a question, you are probably not the only one!

Resources: The course will cover several deep fields of study and each by itself could make up an entire course. This course is designed as a introductory gateway to these areas. As a result, we will not be using a single "required" textbook, but drawing from multiple references. More references will be made available on Piazza/Canvas as and when we cover related topics.

Some helpful resources are listed below. Others will be given over the course of the class, and you will also have access to notes for the resources that are not covered by existing books or monographs.

- "High-Dimensional Statistics: A Non-Asymptotic Viewpoint" by Martin Wainwright.
- "High-Dimensional Probability: An Introduction with Applications in Data Science" by Roman Vershynin.
- "Spectral Methods for Data Science: A Statistical Viewpoint" by Yuxin Chen, Yueijie Chi, Jianqing Fan and Cong Ma.
- "Convex Optimization: Algorithms and Complexity" by Sebastien Bubeck.

Outline of class

We will attempt to cover the following (approximately ordered) list of topics:

Part I: Analysis of convex optimization for statistical learning problems

- Concentration inequalities for scalar random variables
- Some concentration properties of random matrices and applications
- Sparse linear models in high dimensions
- Geometric properties of convex relaxation
- Exact error analysis using Gaussian processes

Part II: Iterative algorithms in statistical learning

- Basic analysis of first-order and second-order iterative optimization for convex problems
- Fast convergence of iterative algorithms in statistical settings
- The approximate message passing algorithm for structured linear models
- Analysis tools for global convergence in nonconvex optimization (if there is time/interest)
- Optimality guarantees via information theory (if there is time/interest)