ISyE 8803: Special Topics in Modern Mathematical Data Science

Spring 2025

Instructor: Dmitrii M. Ostrovskii – `ostrov@gatech.edu` – Skiles 263

Schedule & location: M/W 11:00–12:15, Groseclose 118

Canvas: https://gatech.instructure.com/courses/451798

Disclaimer. This course is targeted primarily at graduate and senior undergraduate students whose research interests are steering towards the interfaces between statistical theory, mathematics of data, information theory and/or optimization theory. Being more of a "problem-solver" than a "theory-builder," I do not intend to give a comprehensive overview of these areas. Instead, we shall first get equipped with a set of techniques and ideas that prove to be useful for solving interesting and challenging problems to optimality or near-optimality. We shall then explore several such problems, each concerning one or more of the following topics:

- self-concordant functions and local Gaussian approximations;
- information-theoretic tools, e.g. the Gibbs duality and f-divergences;
- estimation under structural constraints, such as shift- and affine-invariance;
- solving variational problems via the method of extreme points.

Some open problems shall be discussed along the way, along with promising approaches to them. In a nutshell, the goal of this course is to put one at the cutting edge of research in some of the most interesting, deep, and important topics in the mathematical foundations of data science.

Prerequisites. Multivariate calculus at the level of Taylor expansions in $\mathbb{R}^d$ (MATH 2550 or 1712); linear algebra (MATH 1554 or 1564); differential equations (MATH 2552 or 2562; only homogeneous linear equations with constant coefficients); introductory probability and statistics (MATH 3215, but preferably 3225, 3235, 3236, or 3670). Convex optimization basics (ISyE 3133) are desirable, but a crash course will be given.
You may lack some of these credentials, especially the latter ones, if you are highly motivated.

Contact and office hours. The best way to contact me is by email (please add the class name ISyE 8803 to the subject line) or in person in the office hours (Mon 2-3pm, Tue 1-2pm).

Outline of topics. The course is taught for the first time, so the list of topics might change.

I. **Toolkit**

- Probability: common distributions; key concentration inequalities; matrix Bernstein.
- Statistics: central limit theorem; least squares; asymptotic theory of maximum likelihood estimators; exponential families; generalized linear models; minimax framework.
- Optimization: convexity; subgradients; first-order optimality conditions; Fenchel dual.
- Information theory: f-divergences; application to hypotheses testing; Gibbs duality.

II. **Gaussian approximation**

1. Self-concordant functions: key properties and geometry; analysis of Newton's method.
   - Application: Sharp finite-sample analysis of logistic regression with random design.
2. Local approximation of the Gibbs distribution with log-affine potential on a polytope.
   - Application: Online portfolio selection.

III. **Estimation under structural constraints**

0. General notion of group-equivariant estimators and their optimality.
1. Affine-equivariant estimation.
   - Iterative refining technique.
   - Application to robust covariance estimation.
2. Estimation under shift-invariance.
   - Connection with super-resolution methods.
   - Connection with reproducing filters.
   - Near-optimal oracle construction.

IV. **Extreme points method**

0. Maximizing a convex function over a convex set. Krein-Milman and Dubins theorems.
1. Polytopal ball and sparse vectors.
   - Application: Estimation under shift-invariance, revisited.
2. Log-concave distributions under moment constraints.
   - Application: Optimal self-concordance constant.
   - Application: Sharp lower bound for differential entropy.

Literature. The first thematic block consists of some classical material; detailed lecture notes will be provided. The other blocks incorporate modern results that have never been taught, and were published only recently. For these, I will provide slides and handwritten lecture notes.

Assignments and evaluation. There will be 3 homework assignments pertaining to the first thematic block of the course ("Toolkit") and designed to complement and broaden the lecture material. Some problems are multi-step and guided (here and here are sample homework assignments for a related course). This will be complemented with an informal discussion of a classical paper pertaining to one of the application topics listed above, or with a written exam, at your preference. Grading is generally expected to be very soft: the main goal is to encourage your research curiosity. On the other hand, depending on the enrollment, I anticipate giving favorable recommendations for a postdoc (or PhD program, if appropriate) to the top students. Some advice:

- Start to work on homework assignments as early as possible.
- It's okay to work in a small group. However, always try to solve the problem yourself first. When blocked, use office hours to ask me the question, but meditate on it yourself first.
- Write solutions on your own; suspected plagiatrism will receive zero credit, and there will be further sanctions in case of a relapse. If you found a solution online – this is possible in some cases – study it carefully, rephrase it, and be ready to explain it if asked.

Grading. 24% each of 3 homework assignments; 18% final exam/presentation; 10% scribbling.

Attendance. In general, I do not keep track of attendance. However, if attendance drops below a certain point, I reserve the right to start controlling it with modest bonuses and penalties. Scribbling will be done on the rolling basis; you should be present in class when it's your turn.