Stat 991-302: Mathematics of High-Dimensional Data

Spring 2022

Instructor:	Yuxin Chen
Lecture time:	Mon & Wed, 1:45pm–3:15pm
Location:	F36, Jon M. Huntsman Hall
Office:	TBA
Email:	yuxinc@wharton.upenn.edu

Course Description. This is a graduate level course covering various aspects of mathematical data science, particularly for large-scale problems. We will cover the mathematical foundations of several fundamental learning and inference problems, including clustering, spectral methods, tensor decomposition, graphical models, large-scale numerical linear algebra, matrix concentration inequalities, sparse recovery and compressed sensing, low-rank matrix factorization, shallow neural nets, and so on. Both convex and nonconvex approaches will be discussed. We will focus on designing algorithms that are effective in both theory and practice.

Instructor. Yuxin Chen (email: yuxinc@wharton.upenn.edu). Office hours: by appointment.

Prerequisites. Students should have backgrounds in basic linear algebra, basic probability (measure-theoretic probability is not needed), basic optimization, as well as knowledge of a programming language like MATLAB or Python to conduct simple simulation exercises.

Tentative Topics.

- Week 1-2: Spectral methods
- Week 3: Matrix concentration inequalities
- Week 4: Large-scale eigenvalue problems
- Week 5: Tensor decomposition
- Week 6: Sparse representation
- Week 7-8: Compressed sensing and sparse recovery
- Week 9: Phase transition and convex geometry
- Week 10: Gaussian graphical models and graphical lasso
- Week 11: Low-rank matrix recovery
- Week 12: Robust principal component analysis
- Week 13: Nonconvex matrix factorization
- Week 14: Reinforcement learning

Grading. This is a graduate-level topic class. The grading breakdown is as follows:

- *Attendance.* 20%
- *Final project.* 80%. See the description below.

Project. The term project can either be a literature review or include original research, and you can do it either individually or in groups of 2.

(i) *Literature review.* We will provide a list of related papers not covered in the lectures, and the literature review should involve in-depth summaries and exposition of one of these papers.

(ii) *Original research.* It can be either theoretic or experimental (ideally a mix of the two). You are encouraged to combine your current research with your term project.

There are 2 milestones / deliverables to help you through the process.

- *Proposal.* Submit a short report (no more than 1 page) stating the papers you plan to survey or the research problems that you plan to work on. Describe why they are important or interesting, and provide some appropriate references. If you elect to do original research, please do not propose an overly ambitious project that cannot be completed by the end of the semester, and do not be too lured by generality. Focus on the simplest scenarios that can capture the issues you'd like to address.
- *A written report.* You are expected to submit a final project report—*up to 5 pages with unlimited appendix*—summarizing your findings / contributions. You must submit an electronic copy to my email.

Textbooks. We recommend the following books, although we will not follow them closely.

- *Spectral methods for data science: A statistical perspective*, Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, Foundations and Trends in Machine Learning, vol. 14, no. 5, pp. 566–806, 2021.
- *High-dimensional statistics: A non-asymptotic viewpoint*, Martin Wainwright, Cambridge University Press, 2019.
- *High-dimensional probability: An introduction with applications in data science*, Roman Vershynin, Cambridge University Press, 2018.

References. The following references also contain topics relevant to this course, and you might want to consult them.

- *Nonconvex optimization meets low-rank matrix factorization: An overview*, Yuejie Chi, Yue M. Lu, Yuxin Chen, IEEE Transactions on Signal Processing, vol. 67, no. 20, pp. 5239-5269, October 2019, https://ieeexplore.ieee.org/document/8811622.
- *Mathematics of sparsity (and a few other things)*, Emmanuel Candes, http://statweb.stanford.edu/~candes/papers/ICM2014.pdf, International Congress of Mathematicians, 2014.
- *Statistical Foundations of Data Science*, Jianqing Fan, Runze Li, Cun-Hui Zhang, Hui Zou, Chapman & Hall, 2020.
- *An Introduction to Matrix Concentration Inequalities*, Joel Tropp, Foundations and Trends in Machine Learning, vol. 8, no. 1-2, pp. 1-230, 2015.
- *Graphical models, exponential families, and variational inference*, Martin Wainwright, and Michael Jordan, https://people.eecs.berkeley.edu/~wainwrig/Papers/WaiJor08_FTML.pdf, Foundations and Trends in Machine Learning, 2008.
- *Convex optimization*, Stephen Boyd, and Lieven Vandenberghe, http://stanford.edu/~boyd/cvxbook/, Cambridge University Press, 2004.
- *Topics in random matrix theory*, Terence Tao, https://terrytao.files.wordpress.com/2011/02/matrix-book.pdf, American Mathematical Society, 2012.