Algorithmic Statistics Final Project

Instructor: Sam Hopkins

Due: December 10th

Acknowledgements: This project description is largely borrowed from Ankur Moitra's 2024 edition of the sister course, Algorithmic Aspects of Machine Learning.

For your end-of-the-semester project you will be asked to write a $4-6$ page final paper. You can work in groups or you can work alone, and there are three options:

(1) You can write a literature review on some topic related to the material that we covered in class. In this class, we only focused on problems where there are provable guarantees so you should certainly choose a topic where there are provable guarantees. The goal of the writeup is to survey what is known, some main ideas of the proof in the papers you choose to read, and identify important open questions. You should think of this project as if you were going to give $1-2$ lectures in the course, and how you would explain some topic beyond what we covered in class, to other students. We covered a lot of material in class, but there is still much more out there and I hope to make your literature reviews available on the course website so that they can be a useful reference for others. Also, you should choose a somewhat focused topic because if you choose a topic that is too broad, it will be difficult to get into the details.

(2) You can do original research. This is the more challenging type of project, but it is also the most open ended. You should choose some open question either explicitly mentioned in class, or connected to any of the topics we covered. You should try to give new provable guarantees for some problem, either by giving a new algorithm or giving an improved analysis of an existing algorithm. As with all research, it is never clear whether you will reach the goal that you set out. And so you should make sure that there is some partial progress that you can make along the way, that you can write up in your final paper. Even if you end up not proving what you set out to prove, you can still write up some preliminary ideas along the way or at the very least write up a literature review focused on the open question you worked on. If you choose this type of project, it is important that you come talk to me about it so that I can talk to you about your ideas or point you to other papers that might be relevant.

(3) You can do an empirical project, where you take some algorithms we have discussed in class, or ones in-scope for this class (ideally having provable guarantees) and invent a way to assess their performance against some baselines.

Please email me and Ittai your final paper by December 10th. You are most welcome to come up with your own directions projects, especially if they connect to your other research endeavors. Please do come talk to me about your project ideas. If you do not have an idea, I am here to help. Here are some suggested topics, for writing a literature review on, and a single paper for each to get you started browsing the literature.

- Spectral Clustering: “Spectral Clustering in the Gaussian Mixture Block Model”, Li, Schramm.
- Stability: “Are Stable Instances Easy?”, Bilu, Linial
- Semi-random Models: “Semirandom Planted Clique and the Restricted Isometry Property”, Blasiok, Buhai, Kothari, Steurer.
- Smoothed Analysis I: “Polynomial-Time Power-Sum Decomposition of Polynomials”, Bafna, Hsieh, Kothari, Xu.
- Smoothed Analysis II: “Smoothed Analysis for Learning Concepts with Low Intrinsic Dimension”, Chandrasekaran, Klivans, Kontonis, Meka, Stavropouos.
- Lower Bounds in RL: “Computational-Statistical Gaps in Reinforcement Learning”, Kane, Liu, Lovett, Mahajan.
- Low Degree Lower Bounds: “The Algorithmic Phase Transition of Random k-SAT for Low Degree Polynomials”, Bresler, Huang.
- Expectation Maximization: “Statistical Guarantees for the EM Algorithm: From Population to Samplebased Analysis”, Balakrishnan, Wainwright, Yu.
- Score Matching: “Statistical Efficiency of Score Matching: The View from Isoperimetry”, Heckett, Koehler, Risteski.
- Diffusion Models: “Sampling Is as Easy as Learning the Score: Theory for Diffusion Models With Minimal Data Assumptions”, Chen, Chewi, Li, Li, Salim, Zhang.
- Multimodality: “Sampling Multimodal Distributions with the Vanilla Score: Benefits of DataBased Initialization”, Kohler, Vuong.
- Privacy: “Robustness Implies Privacy in Statistical Estimation”, Hopkins, Kamath, Majid, Narayanan.
- Calibration: “Omnipredictors”, Gopalan, Kalai, Reingold, Sharan, Wieder.
- Hallucinations: “Calibrated Language Models Must Hallucinate”, Kalai, Vempala.
- Nonconvex optimization with provable guarantees: “Matrix Completion has No Spurious Local Minimum”
- Machine unlearning: “Remember What You Want to Forget: Algorithms for Machine Unlearning”
- Overlap gap property: “Some easy optimization problems have the overlap-gap property”
- Message passing algorithms and robustness: “Fast, robust approximate message passing”