RESEARCH STATEMENT

SAMUEL B. HOPKINS

Statistics in high dimensions with large data sets forms the basis for modern machine learning and data science. Despite extraordinary successes in practice and the potential for tremendous further impact, we have limited theoretical understanding of the basic algorithmic building blocks for statistics in high dimensions. Traditional statistics, high-dimensional probability, and the worst-case approach to algorithms all provide tall shoulders to stand on, but each suffers limitations. Statistics and probability often *disregard computation*: the algorithms they suggest typically have exponential running times. The worst-case approach to algorithms *disregards niceness of data*. It incorrectly predicts that problems routinely solved in practice are computationally hard. Building a theory of algorithms for high-dimensional statistics is a grand challenge for computer science, statistics, and mathematics, to answer the questions:

Which problems in high-dimensional statistics have efficient algorithms?
What are the best algorithms for such problems?

My work bridges computation and statistics with the aim of discovering foundational principles for algorithm design and the origins of computational hardness in statistics. Towards this goal, I use a powerful problem-independent algorithmic toolkit arising from convex optimization (the *Sum of Squares method*, or *SoS*) to design algorithms with provable guarantees and to shed light on hardness [L01; P03]. Problem-independence leads to breadth: my work so far spans clustering, robust and heavy-tailed statistics, community detection, sparse recovery, tensor methods, optimization in noisy landscapes, and even cryptography. Along the way, in joint and single-author works, I have solved major algorithmic problems in statistics, giving:

(1) The first algorithm to beat greedy **clustering** for high-dimensional **Gaussian mixture models** [**H**L18], breaking a 20-year barrier and leading to an explosion of algorithms [KKM18; RY20; KKK19; CLM19].[1]
(2) A polynomial-time computable **high-dimensional median** [**H**19], leading to information-theoretically optimal estimators in **heavy-tailed statistics** and a flurry of algorithmic developments [CFB19; LD19; LLVZ19; C**H**KRT19].

A second line of my work studies computational hardness in statistics. My work introduced *pseudocalibration*, a new technique for proving lower bounds against convex programming-based algorithms, which led to the strongest-known **lower bounds** for planted clique and sparse PCA [B**H**KKMP16; **H**KPRSS17].

Finally, I leverage insights from algorithms with strong guarantees but slow running times to design **algorithms which run fast on large data sets**. For example, my work gives the first nearly-linear time algorithm for high-dimensional mean estimation with corrupted samples [D**H**L19] and fast algorithms for tensor decomposition [**H**SS19; **H**SSS16]. I run experiments to validate the speed, robustness, and accuracy of these algorithms [**H**SS19; **H**SS15].

I take a **theory of computing** approach by focusing on problem-independent algorithm design strategies arising from convex optimization and pursuing hardness results in addition to algorithms. In the rest of this statement, I will describe this approach in more detail, address its broader consequences outside of algorithms for statistics, and discuss future work on graphical models, connections to statistical physics, and optimal hypothesis tests.

Beyond a Problem-by-Problem Approach. A theory of algorithms is more than just a list of algorithms: it should offer unifying principles, ideally leading both to problem-independent algorithm-design recipes and rigorous evidence for hardness. The worst-case theory of algorithms has principles like this, such as

[1] Similar results were obtained simultaneously by [DKS18; KSS18].

convexity and NP-hardness, suggesting broad algorithm-design paradigms and explaining computational hardness. Practitioners also have re-usable algorithms, such as MCMC and stochastic gradient descent.

By contrast, at present, algorithms in statistics with rigorous analyses are largely designed on a problem-by-problem basis. This leaves little opportunity to re-use ideas or to establish hardness, especially because NP-hardness is rarely meaningful in statistics – right now, often the best evidence that no efficient algorithm exists for a given statistics problem is just that we have not found one yet.

Problem-Independence, Hardness, and Sum of Squares. My work pursues the beginnings of a broad theory of algorithms for high-dimensional statistics via the SoS method. SoS offers problem-independence in algorithm design, as witnessed by its breadth: SoS has found applications across theoretical computer science, statistics, optimization, operations research, control theory, quantum information, and more [H19; CSB11; NPA07]. At its heart, SoS is a family of powerful convex programs which generalize classic algorithms tools like linear programs and spectral methods. A collection of general-purpose algorithmic building blocks has grown around it, together with re-useable design patterns to put them together into sophisticated algorithms. My work leverages these strategies and has made major contributions back to the SoS toolkit [**H**S17; **H**L18].

SoS also provides an avenue to establish computational hardness: it captures so many algorithmic techniques that showing no SoS algorithm solves a given problem is often the strongest form of evidence within reach for computational hardness in statistics. Pseudocalibration, introduced in my work [B**H**KKMP16], is currently the only problem-independent approach to rule out SoS algorithms for many statistics problems.

Beyond Statistics: Optimization, Cryptography, and More. SoS is a fundamental tool for convex optimization. By honing our understanding of what it can achieve, my work sheds light on computational questions well beyond statistics. Works of mine and others witness this, leveraging new approaches to analyze SoS to obtain new algorithms in quantum computing, combinatorial optimization, cryptography, and more. For example, my collaborators and I used SoS-derived ideas to give the first nontrivial linear programming-based approximation algorithm for the **Max-Cut** problem [**H**ST19]. In cryptography, my work rules out a major potential approach to indistinguishability obfuscation [B**H**JKS19]. SoS is also the most promising avenue to disprove the infamous Unique Games conjecture [BBHKSZ12].

FUTURE DIRECTIONS

I will continue to investigate algorithms and computational complexity for fundamental problems in high-dimensional statistics. Significant near-term algorithmic progress via SoS appears particularly possible in **supervised learning**, where I will focus my algorithm-design efforts in the next few years. In addition, below are broader directions I aim to take in the future.

Classification and Optimal Algorithms. Problem-independence alone is not enough: we ultimately need results which tell us which problems are easy, which are hard, and *why*, by identifying simple criteria which trace out the boundary between the tractable and the intractable. Classification theorems like this are crowning achievements in other areas of computer science; they also often identify *optimal polynomial time algorithms*. For example, in constraint satisfaction, SoS algorithms are known to be optimal among polynomial-time ones [R08; B17] (under relatively standard complexity hypotheses).

My work on lower bounds already suggests a **classification conjecture** for a wide range of high-dimensional statistics problems. By describing optimal polynomial-time hypothesis tests, this would answer the main question at the beginning of this statement – *Which problems in high-dimensional statistics have efficient algorithms?* – for at least a substantial class of hypothesis testing problems. A major goal of my future work is to prove such a conjecture. This will be a challenging endeavor, but there are already interesting results along the way. Using the pseudocalibration lens from my work on lower bounds, [**H**S17] obtains new algorithms and unifies existing ones for community detection in stochastic blockmodels. [**H**KPRSS17] views SoS hypothesis testing algorithms through the same lens to give a problem-independent characterization of SoS algorithms in terms of spectral algorithms.

Graphical Models. Graphical models are key tools in Bayesian statistics, to express prior distributions with complex dependency structures. *Marginalization* is the task of computing the conditional (posterior) distribution of hidden variables in the model from observed values of the rest. It is hard in the worst case, but practitioners use heuristic algorithms with great success – Markov-Chain Monte Carlo, Variational Bayes, and more. I aim to build a theory of marginalization based around convex programming and the SoS method to understand when marginalization is possible in polynomial time. As a first step, in ongoing work joint work I am investigating SoS-based algorithms for marginalization in simple graphical models arising from random constraint satisfaction.

A Bridge to Statistical Physics. A line of work using ideas from phase transitions in statistical physics also holds promise to predict and explain algorithmic (in)tractability in statistics [MMM09]. So far, our SoS-based theory makes exactly the same predictions about algorithmic tractability as the physics-based theory for a wide range of problems: this suggests a hidden connection between the two. Building a bridge between these apparently unrelated families of ideas – convex programming and phase transitions – is one of my long-term aims.

References

[BBHKSZ12] Boaz Barak, Fernando GSL Brandao, Aram W Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. "Hypercontractivity, sum-of-squares proofs, and their applications". In: *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM. 2012, pp. 307–326.

[B**H**KKMP16] Boaz Barak, Samuel B Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. "A Nearly Tight Sum-of-Squares Lower Bound for the Planted Clique Problem". In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2016, pp. 428–437.

[B**H**JKS19] Boaz Barak, Samuel B Hopkins, Aayush Jain, Pravesh Kothari, and Amit Sahai. "Sum-of-squares meets program obfuscation, revisited". In: *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, Cham. 2019, pp. 226–250.

[B17] Andrei A Bulatov. "A dichotomy theorem for nonuniform CSPs". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 319–330.

[CFB19] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. "Fast mean estimation with sub-Gaussian rates". In: *Conference on Learning Theory* (2019).

[C**H**KRT19] Yeshwanth Cherapanamjeri, Samuel B Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. "Algorithms for heavy-tailed statistics: regression, covariance estimation, and beyond". In: *In submission* (2019).

[CLM19] Sitan Chen, Jerry Li, and Ankur Moitra. "Efficiently Learning Structured Distributions from Untrusted Batches". In: *arXiv preprint arXiv:1911.02035* (2019).

[CSB11] Abhijit Chakraborty, Peter Seiler, and Gary J Balas. "Susceptibility of F/A-18 flight controllers to the falling-leaf mode: Nonlinear analysis". In: *Journal of guidance, control, and dynamics* 34.1 (2011), pp. 73–85.

[D**H**L19] Yihe Dong, Samuel B Hopkins, and Jerry Li. "Quantum Entropy Scoring for Fast Robust Mean Estimation and Improved Outlier Detection". In: *33rd Conference on Neural Information Processing Systems (to appear)* (2019).

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. "List-decodable robust mean estimation and learning mixtures of spherical gaussians". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2018, pp. 1047–1060.

[H19] Georgina Hall. "Engineering and Business Applications of Sum of Squares Polynomials". In: *arXiv preprint arXiv:1906.07961* (2019).

[**H**L18] Samuel B Hopkins and Jerry Li. "Mixture models, robustness, and sum of squares proofs". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2018, pp. 1021–1034.

[**H**SSS16] Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. "Fast Spectral Algorithms from Sum-of-squares Proofs: Tensor Decomposition and Planted Sparse Vectors". In: *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*. STOC '16. ACM, 2016, pp. 178–191.

[**H**KPRSS17] Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. "The power of sum-of-squares for detecting hidden structures". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 720–731.

[**H**19] Samuel B Hopkins. "Mean Estimation with Sub-Gaussian Rates in Polynomial Time". In: *Annals of Statistics (to appear)* (2019).

[**H**S17] Samuel B Hopkins and David Steurer. "Bayesian estimation from few samples: community detection and related problems". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 379–390.

[**H**SS15] Samuel B Hopkins, Jonathan Shi, and David Steurer. "Tensor principal component analysis via sum-of-square proofs". In: *Conference on Learning Theory*. 2015, pp. 956–1006.

[**H**SS19] Samuel B Hopkins, Tselil Schramm, and Jonathan Shi. "A Robust Spectral Algorithm for Overcomplete Tensor Decomposition". In: *Conference on Learning Theory*. 2019, pp. 1683–1722.

[**H**ST19] Samuel B Hopkins, Tselil Schramm, and Luca Trevisan. "Subexponential LPs Approximate Max-Cut". In: *In submission* (2019).

[KKK19] Sushrut Karmalkar, Pravesh Kothari, and Adam Klivans. "List-Decodable Linear Regression". In: *33rd Conference on Neural Information Processing Systems, Spotlight Presentation (to appear)* (2019).

[KKM18] Adam Klivans, Pravesh K Kothari, and Raghu Meka. "Efficient Algorithms for Outlier-Robust Regression". In: *Conference On Learning Theory*. 2018, pp. 1420–1430.

[KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. "Robust moment estimation and improved clustering via sum of squares". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2018, pp. 1035–1046.

[L01] Jean B Lasserre. "Global optimization with polynomials and the problem of moments". In: *SIAM Journal on optimization* 11.3 (2001), pp. 796–817.

[LD19] Guillaume Lecué and Jules Depersin. "Robust subgaussian estimation of a mean vector in nearly linear time". In: *arXiv preprint arXiv:1906.03058* (2019).

[LLVZ19] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. "A Fast Spectral Algorithm for Mean Estimation with Sub-Gaussian Rates". In: *arXiv preprint arXiv:1908.04468* (2019).

[MMM09] Marc Mezard, Marc Mezard, and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[NPA07] Miguel Navascués, Stefano Pironio, and Antonio Acín. "Bounding the set of quantum correlations". In: *Physical Review Letters* 98.1 (2007), p. 010401.

[P03] Pablo A Parrilo. "Semidefinite programming relaxations for semialgebraic problems". In: *Mathematical programming* 96.2 (2003), pp. 293–320.

[R08] Prasad Raghavendra. "Optimal algorithms and inapproximability results for every CSP?" In: *Proceedings of the fortieth annual ACM symposium on Theory of computing*. ACM. 2008, pp. 245–254.

[RY20] Prasad Raghavendra and Morris Yau. "List Decodable Learning via Sum of Squares". In: *ACM-SIAM Symposium on Discrete Algorithms (SODA), to appear* (2020).