Proposal Summary

Over the past few decades an uncomfortable truth has set in that worst case analysis is not the right framework for studying machine learning: every model that is interesting enough to use in practice leads to computationally hard problems. The goal of the PI's research agenda is to move beyond worst case analysis. This involves formalizing when and why heuristics – such as alternating minimization and Gibbs sampling – work as well as designing fundamentally new algorithms for some of the basic tasks in machine learning. This project has already had a number of successes such as provable algorithms for *nonnegative matrix factorization*, *topic modeling*, *learning mixture models*, *tensor decomposition*, *dictionary learning* and *independent component analysis*. In this proposal, the PI will highlight four inter-related projects that significantly advance this research agenda:

- **Alternating minimization** is a popular heuristic for minimizing certain types of non-convex functions. The PI proposes to view it instead as trying to minimize an *unknown* convex function given an approximation to its gradient. This suggests a new set of questions, whereby tools from optimization could lead to faster and more sample efficient algorithms for *dictionary learning* and offer valuable insights into how these tasks are accomplished in nature.
- The PI will investigate inference in **Bayesian networks**, where we only know of provable algorithms in quite limited settings. In contrast, Gibbs sampling is widely successful, but has defied theoretical explanation. The PI will use recent probabilistic tools from the constructive proof of the local lemma to analyze Gibbs sampling and offer insights into why it works.
- One of the major recent advances in machine learning is the development of a broad range of algorithms for **linear inverse problems**. But there are important gaps in our understanding for problems such as tensor completion. The PI will design new algorithms for it that utilize higher levels of the sum-of-squares hierarchy and need many fewer observations.
- The PI will explore the application of **semi-random models** to machine learning and extend existing algorithms to work in this challenging setting. Such algorithms would likely be more robust to distributional assumptions, and generalize to a broader range of domains.

These projects all involve expanding the reach of theory into areas where there is a serious gap in our current understanding, and will open up new avenues for further exploration.

Intellectual Merits This work will lead to more efficient algorithms for basic machine learning problems, and these algorithms will come with provable guarantees on their performance. This work has the potential to have a major impact on theory and practice by introducing new problems and agendas to theory as well as developing new tools for reasoning about why these algorithms work in practice. And as an additional benefit, it will further contribute to our burgeoning understanding of alternating minimization, graphical models, semi-random models and the sum-of-squares hierarchy.

Broader Impact This program could develop into a central area of focus in theoretical computer science, and lay the groundwork for many more fruitful collaborations between theory and machine learning. This project will involve training the next wave of students, and equipping them with the necessary tools to work in this area. Additionally, the PI plans to continue his outreach activities designed to facilitate communication between theory and practice. This involves giving tutorials, and introducing practitioners to the latest algorithmic developments, as well as teaching a graduate seminar and creating new course materials on modern machine learning models, most of which are not currently studied within theory.

CAREER: Algorithmic Aspects of Machine Learning

Ankur Moitra

Massachusetts Institute of Technology

1 Introduction

Algorithms and complexity are the theoretical foundation and backbone of machine learning. Yet there is a large divide between theory and practice because the optimization problems that arise in machine learning are in large part solved using heuristics that have no provable guarantees. To elaborate, a machine learning system is composed of a *model* and an *algorithm* and we have many rich models that describe aspects of the world around us – e.g. Judea Pearl recently won the A.M. Turing Award in large part for introducing *Bayesian networks* for modeling causal relationships (see Section 2.4). But a model is no good without an algorithm to determine how to set its parameters or to perform inference. Those same aspects that make a model appealing – that it is general enough to describe many interesting processes – also seem to come hand in hand with optimization problems that are NP-hard or worse!

Over the past few decades an uncomfortable truth has set in that worst case analysis is not the right framework for studying machine learning: every model that is interesting enough to use in practice leads to computationally hard problems. The goal of our research agenda is to move beyond worst case analysis. This involves two complementary directions. Firstly, can we formalize when and why heuristics – such as alternating minimization and Gibbs sampling – work? The hope is that if we can analyze heuristics instead of dismissing them, we can add something new and surprising to our algorithmic toolkit. Secondly, can we design fundamentally new algorithms for some of the basic tasks in machine learning? As we will see, there is a rich set of tools available from theory that are a natural fit for some of the outstanding problems in machine learning.

This project has already had a number of successes such as provable algorithms for *nonnegative matrix factorization* [AGK$^+$12], *topic modeling* [AGM12, AFH$^+$12], *learning mixture models* [KMV10, MV10, BS10], *tensor decomposition* [BCM$^+$14], *dictionary learning* [AGM14, AAJ$^+$14] and *independent component analysis* [AGM$^+$12, GVX14]. This line of work not only gave new algorithms for well-studied problems, but in many cases it significantly contributed to our understanding of what are the right instances of these problems to study. In this proposal, we will highlight four inter-related projects that significantly advance this research agenda:

- **Alternating minimization** is a widely used heuristic for minimizing certain types of non-convex functions. We can analyze its performance in some settings (e.g. for matrix completion [JNS13, Ha14]) but for others it remains a mystery. We propose instead to view it as trying to minimize an *unknown* convex function given an approximation to its gradient. This suggests an interesting, new set of questions, whereby tools from optimization could lead to faster and more sample efficient algorithms for *dictionary learning* and other related problems. Moreover, this framework can be used to analyze even simpler variants of alternating minimization that have been proposed in the neuroscience community [OF97, LS00], and would offer the first algorithmic explanation of how these tasks are accomplished in nature. The research proposed here will serve as a bridge between two disparate communities and foster new collaborations. See Section 2.3.

- A **Bayesian network** compactly encodes a joint probability distribution in high dimensions; such networks are ubiquitous in modern machine learning. The basic task is to perform

inference – we want to update our beliefs about the latent variables based on our observations. Gibbs sampling is widely successful in practice, but has defied theoretical explanation. The research proposed here will leverage recent probabilistic tools from the constructive proof of the local lemma [MT10, HSS11] to analyze Gibbs sampling and offer new insights into why it works, finally moving beyond the crutch of assuming that the underlying graph has low tree-width. See Section 2.4.

- One of the major advances in machine learning in recent years is the development of a broad range of algorithms for **linear inverse problems** through semidefinite programming. The most famous example is *matrix completion* [CR08]. However in many cases the natural relaxation (based on the *atomic norm* [CRP$^+$12]) is itself hard to solve and introduces a new source of difficulty. This is the case for *tensor completion* which is a natural generalization that touches on an important issue: To what extent does adding additional structure to an inverse problem make it information theoretically easier – in the sense that we need fewer samples – but computationally more difficult to utilize the higher-order structure? This particular problem turns out to be related to the *quantum separability* problem [BCY11] and suggests an approach for rounding higher levels of the sum-of-squares hierarchy to give algorithms for tensor completion that need many fewer observations. See Section 3.2. See also Section 4.1 for further background on tensors and their use in statistics and learning.

- **Semi-random models** have great explanatory power and are an elegant model originating from theoretical computer science [BS95], [FK01] that sits between worst-case and average-case analysis. These models have not been used in the context of learning, although they are a natural fit, and can help address a foundational question: Are algorithms that we prove to work under various distributional assumptions in fact over-exploiting statistical properties of the instances they are given? We plan to extend existing learning algorithms – such as alternating minimization for matrix completion – to work in this more challenging setting. Such algorithms would likely be more robust to distributional assumptions, and would have a better chance of succeeding across a broad range of application areas. See Section 4.2.

Together, these projects would expand the reach of theory into a number of areas where there is a serious gap in our current understanding. In particular, it would make new connections between alternating minimization and approximate gradient descent, leverage modern probabilistic tools [MT10, HSS11] to analyze Gibbs sampling, connect tensor completion and quantum complexity, and rethink the standard distributional models used in machine learning. This proposal cuts across several areas of computer science and applied mathematics and has the potential to enrich these areas by building new bridges between them. I will also describe my previous research accomplishments that are most relevant to this proposal – in particular, nonnegative matrix factorization (Section 2.1), topic modeling (Section 2.2), learning mixture models (Section 3.1) and tensor decomposition (Section 4.1).

The intellectual merits of this proposal are described above, and its broader impact will be achieved in two ways. Firstly, we will integrate the research activities with an education plan that will create new graduate and undergraduate courses and train the next wave of students, equipping them with the necessary tools to work in this area (Section 6.2). I recently wrote a monograph based on my graduate seminar that is freely available on my website and has already served as a valuable reference both within MIT and at other universities, and I plan to expand this into a textbook. More broadly, I have participated in a number of programs that are specifically designed to encourage students from underrepresented groups to pursue computer science (Section 5).

Secondly, this proposal will foster connections between theoretical computer science, machine learning, statistics and signal processing, and bring new perspectives and new agendas into these

fields. Many of the models and heuristics used in machine learning have been for the most part unexplored by theory. This work has the potential to set new directions in theory as well as lead us to a new understanding of when various heuristics work and where they should be applied. Moreover we expect that some of the algorithms developed in this proposal will lead to better algorithms in practice too. Finally, the projects in this proposal will offer us an opportunity to rethink some of the statistical foundations of machine learning and signal processing, and will contribute important new estimators that can be used in place of the maximum likelihood estimator and have provable algorithms to compute them. These projects cut across many areas, and have the potential to build exciting new bridges between them.

We will organize this proposal into a three-front attack on intractability in machine learning and related issues. In Section 2 we describe an approach based on leveraging structural assumptions to get around known hardness results. In Section 3 we suggest ways to analyze new estimators in cases where the standard ones are hard to compute. And in Section 4 we explore various models in between worst-case and average-case analysis in the context of machine learning.

2 Representation Learning and Inference

Often we can get around known intractability results by introducing structural assumptions that are motivated by applications. Indeed we got around NP-hardness [Va09] and fixed-parameter intractability results [AGK$^+$12] in nonnegative matrix factorization (Section 2.1) and NP-hardness results for computing the maximum likelihood estimator [AGM12] in topic modeling (Section 2.2) through the notion of separability [DS03] and gave new algorithms that have already had an impact on practice. In future work, we will analyze alternating minimization on incoherent dictionaries and we discuss the implications of this project for neuroscience (Section 2.3). Also in future work, we propose to analyze Gibbs sampling on bipartite Bayesian networks and we believe that the right assumptions in this context involve its parameters and not its topology (Section 2.4).

2.1 Prior Work: Nonnegative Matrix Factorization

Nonnegative matrix factorization is a fundamental problem in linear algebra which has a rich history spanning quantum mechanics, probability theory, data analysis, polyhedral combinatorics, communication complexity, demography, chemometrics, etc (see references in [AGK$^+$12]). In this problem, the input is an entry-wise nonnegative matrix $M \in \Re^{n \times m}$ and an integer $r > 0$ and the goal is to write M as the product of $A \in \Re^{n \times r}$ and $W \in \Re^{r \times m}$ where these are entry-wise nonnegative matrices too. In the past decade, this problem has become enormously popular in machine learning. Due to space limitations we will not describe the full range of its applications in detail, but will instead focus on applications to text analysis to illustrate the main ideas.

A grand challenge in machine learning is to design automated tools to organize and reason about large collections of documents; a number of foundational papers have suggested that we think of documents as each being described as a convex combination of "topics" which in turn are distributions on words [PRT$^+$00, Hof99]. This is precisely the nonnegative matrix factorization problem since we can interpret the columns of A as topics, and the columns of W as a representation of each document as a convex combination of topics. (after an appropriate renormalization) In this way, the goal is to extract latent relationships in the data by solving nonnegative matrix factorization.

Unfortunately nonnegative matrix factorization is NP-hard [Va09], and even worse Arora, Ge, Kannan and the PI proved that this problem is fixed parameter intractable [AGK$^+$12]. However

one of the most important contributions of this work was in providing a path to move beyond worst case analysis. To this end, we studied the *separability* condition, which was introduced in [DS89] and is believed to hold in a number of natural settings. In the example above it has a simple interpretation: we require that for each topic there is an *unknown* word called an *anchor word* that is a strong indicator for the given topic. For example, consider the word "401k". There are certainly many articles about "personal finance" that do not contain this word. Yet when "401k" does occur in an article it gives a strong indication that the article is at least partially about "personal finance" and so we call it an anchor word. We proved that there is a polynomial time algorithm to find anchor words and used this to give an efficient algorithm for nonnegative matrix factorization when A is separable. This gave the first algorithm for nonnegative matrix factorization that provably works under a non-trivial condition on the input. This research direction has since been taken up by a number of researchers who have sought even faster and more practical algorithms and has spurred many collaborations across theory and machine learning [BRR$^+$12, KSK13, GV14, ZBG14].

2.2 Prior Work: Topic Models

Topic modeling is closely related to nonnegative matrix factorization, but it has an important stochastic twist [PRT$^+$00, Hof99, BNJ03, Bl12]. The difference is that we assume there is a distribution that generates the columns of W and moreover even though a document is associated with a distribution over words, what we actually observe is not the full distribution but a rather samples from it. When the length of a typical document is much shorter than the number of terms in the vocabulary, these can be quite different. The question is: Can we still give provable algorithms that work in the presence of such sparse and incomplete data?

In [AGM12], Arora, Ge and the PI used the separability assumption to give a provably correct algorithm for learning the parameters of a topic model to any accuracy. Most approaches for topic modeling are based on the singular value decomposition, however what we are trying to compute in this context is necessarily a nonnegative matrix and our work was the first to leverage nonnegative matrix factorization as the main tool. Moreover ours was the first algorithm that provably works for correlated topic models, which are much more realistic and versatile models in practice [BL07, LM07]. These algorithms are theoretically appealing (due to their generality) and in fact turn out to be highly practical. In [AGH$^+$13], the PI and collaborators gave a highly efficient implementation of these algorithms, and tested it against state-of-the-art topic modeling toolkits such as MALLET with the help of one of the maintainers of this package. The experiments showed that this algorithm runs a hundred times faster than previous approaches, while producing better quality topics according to a variety of metrics that have been suggested in the topic modeling community. This is an auspicious example where new models can lead to new theoretical questions, and ultimately much better performance in practice.

2.3 Research Direction: Dictionary Learning via Alternating Minimization

Sparse representations play an essential role in many fields including statistics, signal processing and machine learning [El10, Ma98]. But can we efficiently learn an unknown basis that enables a sparse representation, if one exists? This problem is usually called either *dictionary learning* or *sparse coding* and it is a problem of central importance. Algorithms for this task often serve as a tool for feature extraction and ultimately a building block for more complex tasks such as classification, compression, segmentation and de-noising. More recently, it has also been used in some deep learning architectures [RBL07].

In the usual stochastic model there is an unknown dictionary $A \in \Re^{n \times m}$ and our goal is to learn A from random examples of the form $Y_i = AX_i$ where X_i has at most k non-zeros and is

chosen from an appropriate distribution. This is a natural stochastic generalization of the classic *sparse recovery* problem in which we are given both Y_i and A and our goal is to find X_i [DS89, DH99, DE03, GN03, Do06, CRT06]. Moreover in sparse recovery we impose conditions on A that ensure X_i is the uniquely sparsest solution to the linear system $Y_i = AX_i$. The twist here is that in dictionary learning we do not have the luxury of knowing A or choosing it; we have to discover it.

The standard approach in practice is *alternating minimization* where the general idea is to maintain a guess for both A and the representations $\{X_i\}$ and at every step either update $\{X_i\}$ using an algorithm for sparse recovery or update A by, say, solving a least-squares problem [AEB06, EAH99]. However why these approaches work so well is still a mystery – they are usually derived as a heuristic for minimizing some non-convex function. Recently a number of researchers have sought new algorithms – with provable guarantees – for dictionary learning. Spielman, Wang and Wright gave an algorithm that works provided A has full column rank and $k \approx \widetilde{\Theta}(\sqrt{m})$ [SWW12]. Recall that k is the number of non-zeros in X_i. However this does not encompass the usual applications of dictionary learning where we set m to be much larger than n because it allows us to work with a much richer family of signals. This is called the *overcomplete* case.

Recently, Arora, Ge and the PI and independently Agarwal, Anandkumar, Jain, Netrapalli and Tandon gave the first provable algorithms that work in the overcomplete setting [AGM12, AAJ$^+$14]. These algorithms work for μ-*incoherent* dictionaries where the columns of A_i are unit vectors and $\langle A_i, A_j \rangle \leq \mu/\sqrt{n}$ for all $i \neq j$. In fact, ours works up to $k = \widetilde{\Theta}(\sqrt{n}/\mu)$ which is the information theoretic threshold for sparse recovery [DS89, DH99, DE03, GN03]. We remark that the algorithms in [AGM12, AAJ$^+$14] both make use of alternating minimization, but do so only after they already have an estimate that is extremely close to the true dictionary. Subsequently, Barak, Kelner and Steurer gave an algorithm utilizing the sum-of-squares hierarchy that allows nearly linear sparsity [BKS14b].

However there is still a wide gap in our understanding. In principle, alternating minimization itself could work from weak starting conditions and provide an algorithm that achieves the best of all of these worlds in terms of the sparsity it can tolerate, how rapidly it converges to the true dictionary and its sample complexity. Often, though, one needs the right initialization procedure to get it started:

Question 1. *Is there a variant of alternating minimization (coupled with a good initialization procedure) that provably works from a weak starting condition?*

The usual derivation for alternating minimization is to start with some non-convex energy function that we would like to minimize, such as the reconstruction error

$$E(\widetilde{A}, \widetilde{X}_i) = \sum_{i=1}^{p} \|Y_i - \widetilde{A}\widetilde{X}_i\|$$

where we impose the constraint that each X_i is k-sparse. If we knew the X_i's then we could set $\widetilde{X}_i = X_i$ and the above energy function would be convex. Moreover we could compute its gradient with respect to $\widetilde{A}$ and take a step in that direction. This is obviously a cheat because the whole point is to learn the X_i's in the first place.

The key point is that this is still the right way to think about alternating minimization and we should think of it as trying to minimize an *unknown* convex function $E(\widetilde{A}, X_i)$. Instead we have access to the gradient of $E(\widetilde{A}, \widetilde{X}_i)$ with respect to $\widetilde{A}$ which we hope should be a good enough proxy for us to minimize it.

Question 2. *Is there a general connection between alternating minimization and approximate gradient descent?*

There seem to be many important questions to explore here: if we have access to an approximate oracle for the gradient of a convex function, what are the minimal conditions we need to prove that we converge to something close to its minimum?

Finally, let us turn to neuroscience. Historically, dictionary learning was introduced in a foundational paper of Olshausen and Field [OF97] who considered applying it to a collection of natural images. They discovered that (in contrast to principal component analysis) it finds a relatively interpretable basis that shares many qualitative properties with the receptive field in the mammalian visual cortex. It is by now widely accepted that sparse representations play a key role in neural coding. But how is dictionary learning accomplished in nature?

Question 3. *Are there* neurally plausible *algorithms for dictionary learning?*

The term *neurally plausible* is itself not precisely defined anywhere in the literature but appears in many places [OF97, Se14]. Indeed there are variants of alternating minimization that make use of only very simple operations such as matrix-vector products and thresholding. All of these operations are believed to be implementable by groups of neurons, and so if we could analyze it this would give us the first *neurally plausible* algorithms for dictionary learning with provable guarantees. This would introduce tools from theoretical computer science and optimization into neuroscience would suggest that non-convexity need not be a hurdle to a rigorous mathematical theory of neural computation. It would also pave the way for more interaction between algorithms researchers and neuroscientists. This section is based on ongoing discussions with Sanjeev Arora, Rong Ge and Tengyu Ma.

2.4 Research Direction: Inference in Bayesian Networks

Judea Pearl recently won the A. M. Turing Award for introducing *Bayesian networks* and developing their mathematical foundations [Pe88]. These networks have become a mainstay in robotics, computer vision, computational biology, natural language processing and many other fields [KF09, JB02] and have changed the way we approach reasoning in uncertain environments. There are by now countless surveys, workshops and courses dedicated to this topic and yet we are sorely lacking in algorithms with provable guarantees.

Formally, a Bayesian network is a directed acyclic graph $G = (V, E)$ where we associate a random variable with each node; the key property is that the value at any particular node is conditionally independent of all the other nodes once its parents are fixed. Hence once we fix the topology, we only need to specify the conditional distribution for each node for each configuration of values for its parents. For a more thorough discussion see [KF09, JB02]. In this way, a Bayesian network compactly describes a high-dimensional joint probability distribution. Moreover we can make use of both *causal* (from cause to effect) and *diagnostic* (from effect to cause) reasoning once we are working with probabilities, and this is what distinguishes it from some of the rule-based systems it has supplanted. To be concrete, we will use the Quick Medical Reference (QMR-DT) model as a running example. This is a large bipartite Bayesian network where each node represents either a symptom or a disease. Additionally, each symptom is the weighted noisy-or of some set of diseases. This is one of the most important Bayesian networks and its parameters were hand-tuned by experts and took the equivalent of fifteen man-years of work!

One of the central problems is to take some given Bayesian network (like the one above) and perform inference – we want to update our beliefs about the latent variables based on our observations so that we can make predictions. Unfortunately, all of the known algorithms work in limited settings. There are algorithms such as *belief propagation* [Pe82] and the *junction tree algorithm* that work when the underlying graph has low tree-width. There are also approaches based on rejection

sampling that work in a range of parameters where you could accomplish almost the same task by ignoring the Bayesian network altogether [DL97].

Question 4. *Are there provable algorithms that work on interesting families of Bayesian networks with large tree-width?*

Almost all interesting Bayesian networks – including QMR-DT – have large tree-width. It is a crutch to focus on graphs with low tree-width since it effectively gives up on doing anything more interesting than dynamic programming.

Arguably the most successful algorithm in practice is Gibbs sampling [GG84], which has defied theoretical explanation. The basic strategy is to design a Markov chain in such a way that its steady-state distribution is exactly the posterior distribution on the latent variables we want to compute. We can think of the random walk as exploring the peaks and valleys of some implicitly defined energy function. Some energy functions are easy to get stuck in – for example, ones with many deep modes. The QMR-DT network is a natural starting point for exploring these issues, and we plan to study how the parameters of the network influence the behavior of Gibbs sampling. See also [HS13, JHS13] for work on fitting the parameters of some classes of noisy-or networks.

Question 5. *Does Gibbs sampling mix rapidly for noisy-or networks?*

Markov chain Monte Carlo (MCMC) algorithms have been successfully analyzed in a number of settings in approximate counting, most notably the work of Jerrum, Sinclair and Vigoda on approximating the permanent [JSV04]. But applications in machine learning seem to require genuinely new sorts of tools.

We can take inspiration from the recent progress on constructive versions of the local lemma [MT10] and use this work to give a proof of concept for some of the research directions that we proposed here. In particular let us take a k-SAT formula ϕ with the right regularity conditions so that a satisfying assignment is guaranteed to exist using the local lemma. Moreover, there are many such examples where a random assignment is extremely unlikely to satisfy ϕ and so using rejection sampling to find one would not work. We can construct a (highly artificial) Bayesian network where the observed nodes represent the clauses and the hidden nodes represent the variables. Note that its topology is in general quite complex. Nevertheless, there are efficient algorithms based on setting up an appropriate Markov chain that generate a satisfying assignment [MT10].

Question 6. *Are there efficient algorithms for sampling a uniformly random satisfying assignment of ϕ, provided that it meets the conditions of the local lemma?*

The best algorithms can generate a random satisfying assignment from a distribution that satisfies certain marginal constraints that the uniform distribution on satisfying assignments also satisfies [HSS11]. But this falls short of answering the above question, which is natural in its own right. Moreover the above question would be an important stepping stone towards analyzing Gibbs sampling more generally on bipartite Bayesian networks where one layer represents observed variables and the other represents latent variables. In such cases, one can think of the observations as defining soft constraints and the hope is that many of the intuitions that hold in the constraint satisfaction case (where there are hard constraints) should carry over. This section is based on ongoing discussions with Sam Eldar, David Karger, Jonathan Kelner and David Rolnick.

3 Method of Moments

In many learning problems, the moments of a distribution can be used to construct good estimators in cases where the natural one is hard to compute. Indeed, this is the case for mixtures of Gaussians

where it was known that computing the maximum likelihood estimator is APX-hard [AK05] and we gave an alternate estimator based on the method of moments that yielded the first polynomial time algorithm that works even when the components almost entirely overlap [KMV10, MV10] (Section 3.1). We face a similar difficulty for the tensor completion problem where the natural convex program [CRP$^+$12] is hard to compute. We propose to round the natural sum-of-squares relaxation (which places constraints on a distribution through its moments) using connections to the quantum separably problem. This would yield better algorithms that need many fewer observations (Section 3.2).

3.1 Prior Work: Learning Mixtures of Gaussians

Mixtures of Gaussians are ubiquitous in learning and statistics and are used whenever data is believed to be generated from multiple sources (see [Li95]). A fundamental problem is to learn the parameters of the mixture given what we believe are random samples from its distribution. In fact, the study of these types of problems dates back to one of the founders of statistics – Karl Pearson – who was interested in evolution and believed that a particular species of crab called the Naples crab was not one but actually two species. He postulated that the data he observed could be explained as a mixture of two Gaussians [Pe94]. Since his foundational work, these mixture models have found applications in numerous other areas including physics, geology and genetics.

This problem is a good example of the tension between sample complexity and computational complexity. The maximum likelihood estimator is known to require few samples to converge, but is hard to compute [AK05]. So to find a polynomial time algorithm we would need to rely on different estimators. Starting with Dasgupta [Da99], a long line of works in theoretical computer science has sought a polynomial time algorithm for this problem [Da99, DS00, AK05, VW04, AM05, BV08]. The existing strategy was clustering, which at the very least needs to assume that the components are almost entirely disjoint in order to succeed.

In [KMV10, MV10], Kalai, the PI and Valiant gave an approach based on the method of moments that requires no separation assumptions whatsoever. Our algorithm was the first to learn the parameters of a mixture of two Gaussians with provably minimal assumptions, thus resolving this long-standing open question. Our approach was based on reducing an n-dimensional learning problem to a series of one-dimensional learning problems, and then analyzing the method of moments by proving properties about the associated system of polynomial equations through the heat equation. In [MV10] we gave a polynomial time algorithm for any constant number of components and independently Belkin and Sinha [BS10] gave an algorithm with similar guarantees.

Hsu and Kakade recently gave an algorithm that works for mixtures of many spherical Gaussians, provided that the means are linearly independent. Their approach is based on tensor decompositions [HK13]. See also Section 4.1. In [MV10] we gave a lower bound that showed even in one dimension there are two different mixtures of k Gaussians that are not close on a component by component basis but as mixtures are exponentially close in statistical distance. This leaves open a fundamental question in between these upper bounds and lower bounds:

Question 7. *Are there algorithms that run in time polynomial in n, $1/\epsilon$ and k that learn mixtures of k Gaussians to accuracy ϵ provided that the means of the components are linearly independent?*

The case where the components are identical seems to be very different, and the above question needs completely new ideas.

3.2 Research Direction: Tensor Completion

One of the major advances in machine learning in recent years is the development of a broad range of algorithms for linear inverse problems based on semidefinite programming. The most famous example is the matrix completion problem where there is an unknown matrix $M \in \Re^{n \times n}$ and we observe a subset $\Omega \subseteq [n] \times [n]$ of its entries and our goal is to recover M exactly. Candes and Recht [CR08] introduced this problem and studied it under various natural assumptions – namely that (a) M is low rank (b) the singular vectors of M are not aligned with the standard basis vectors (i.e. M is *incoherent*) and (c) the observations Ω are distributed uniformly at random. The authors showed a stunning result: there is a simple convex program that recovers M exactly with high probability with only $Cn^{1.2}r \log n$ observations, where r is the rank of M. The number of observations was improved in a sequence of works to $Cnr \log^{O(1)} n$ [KMO10, CT10, Re11], which is close to the information theoretic lower bound.

The approach in the above works is to study the following relaxation:

$$\min \|X\|_* \text{ s.t. } X_{i,j} = M_{i,j} \text{ for all } (i,j) \in \Omega$$

where $\|X\|_*$ is called the *nuclear norm* and is the sum of the singular values of X. We remark that the number of non-zero singular values of X is precisely its rank and so the hope is that the nuclear norm is a good proxy for the rank in much the same way that the ℓ_1-norm is a good proxy for the sparsity in compressed sensing.

The above relaxation turns out to be a semidefinite program (this follows from the dual characterization of the nuclear norm). In fact, various other semidefinite programs play a crucial role in designing algorithms for other linear inverse problems such as matrix sensing [RFP10], phase retrieval [CES$^+$13] and sparse principal component analysis [CLM$^+$11]. See also [CRP$^+$12] for a general framework that captures some of these examples as a special case; following their work, the general recipe is to define an *atomic norm* $\|X\|_{\mathcal{A}}$ and study conditions under which the solution to

$$\min \|X\|_{\mathcal{A}} \text{ s.t. } L_i(X) = L_i(M) \text{ for all } i \in \Omega$$

recovers M exactly. Here L_i represents a linear function but for our purposes we will think of it as revealing a particular entry of a matrix or a tensor.

However this is not the end of the story. Consider the *tensor completion* problem where we have the same model as before but our observations now come from a low-rank tensor T [CRP$^+$12]. This extension arises whenever you work with data that has more than two types of attributes. There has been a considerable amount of work on this problem, but the main difficulty is that for this problem and many others the natural choice for the atomic norm is itself NP-hard to compute [Gu03, HM13b]. In such a case we could ask:

Question 8. *Can we use the full power of higher levels of the sum-of-squares hierarchy [Pa00, La01] to design better algorithms for linear inverse problems?*

This touches upon an important issue: The more structured the object is that our observations are coming from, the fewer observations we should need. But it also makes the underlying optimization problem harder, and it is natural to look to the sum-of-squares hierarchy to understand the allowable tradeoffs between the sample complexity and the running time. This question is part of a burgeoning area of research [BR13, CJ13, HM13b, ZWJ14]. To keep the exposition as simple as possible, we will focus on the fourth-order tensor completion problem (the third-order version is just as interesting, but there are more technicalities to deal with).

Question 9. *How many observations of a low-rank, incoherent tensor $T \in \Re^{n \times n \times n \times n}$ do we need in order to be able to fill in the rest of its entries?*

The crucial point is that we could always flatten T to get a low-rank matrix instead as follows:

$$\text{flat}(T) = \sum_i (u_i \otimes_{KR} v_i) \otimes (w_i \otimes_{KR} x_i)$$

Here $a \otimes_{KR} b$ is the Khatri-Rao product which for n-dimensional vectors a and b results in an n^2-dimensional vector whose entries are the product of entries in a and b. It is easy to see that if T is incoherent and has rank r then flat(T) inherits these properties. Hence we can complete T by flattening it and appealing to known results about matrix completion and forgetting entirely about its tensor structure. The trouble is that this requires too many observations. Let us take $r = O(1)$ from now on. We would need about n^2 observations to follow this strategy, even though in principle it should be possible to complete T (if we exploit the fact that it is a tensor) using as few as about n observations.

Question 10. *Are there efficient algorithms for fourth-order tensor completion that need only $n^{2-\delta}$ observations for some $\delta > 0$?*

In fact, this question has a number of intriguing parallels to well-studied problems in quantum complexity. We will attempt to minimize the amount of background needed for this discussion, and give a general idea of the connections. It makes sense to ask an even simpler question: How many entries do we need to observe to be able to predict the sign of the other entries in T with non-trivial advantage? This version of the question has been studied for matrix completion [SS05] and if it sounds more like a generalization bound, that's because it is! The approach in [SS05] is to bound the Rademacher complexity to prove a generalization bound, and this pattern gives us a new way to look for relaxations that work for tensor completion: let's look for ones where we can bound their Rademacher complexity.

In particular, it comes down to a question about random matrices: Suppose M has s non-zero entries that are chosen uniformly at random and are equally likely to be $+1$ or -1. Then how large (as a function of s) is the maximum of $u^T M v$ over unit vectors u and v? It is easy to show that the answer is roughly $\max(1, \sqrt{s/n})$. It turns out that we need this bound to be smaller than s/n in order to get a non-trivial generalization bound, and this happens when s is about n [SS05]. In tensor completion we are faced with a similar problem: Suppose T is a tensor that has s non-zero entries which are chosen in the same manner as above. Again we are interested in a particular norm of T which in this case is the maximum over unit vectors u, v, w and x of $T(u, v, w, x)$. This is called the *injective norm* and is denoted by $\|T\|_{inj}$.

Question 11. *How large does s need to be before we can algorithmically certify that $\|T\|_{inj}$ is at most s/n^2?*

Again, we could always prove an upper bound by flattening T and computing the spectral norm of flat(T) which is at most s/n^2 once s is at about n^2. But the point is that if we restrict the maximization of $T(u, v, w, x)$ to be over spread out vectors then the maximum should be about s/n^2 even when s is about n. If we had a relaxation that could certify this, it would give us an estimator that could predict the entries of a fourth-order tensor T from a *linear* instead of a *quadratic* number of observations.

In fact, understanding the approximability of the injective norm has been a central question in quantum complexity precisely because it is related to the problem of deciding whether or not a given density matrix is close to the set of separable states (see [Gu03, HM13b] and references therein). Brandao, Chrisandl and Yard recently gave an exciting quasi-polynomial time algorithm for the quantum separability problem. It is one of the rare examples of an algorithm that makes use of higher levels of the sum-of-squares hierarchy and has led to progress on related questions such

as approximating the injective norm itself (within an additive term) [BKS14a] and new algorithms for dictionary learning [BKS14b]. Barak, Brandao, Harrow, Kelner, Steurer and Zhou also gave an algorithm for certifying non-trivial upper bounds on the injective norm of random tensors [BFH$^+$12] but here we need the tensors to be sparse since s is what governs how many observations we need in order to solve tensor completion.

There seem to be many potential connections between hierarchies and learning with interesting directions going both ways. Here we gave an example where hierarchies may provide new algorithms for various learning problems (not just linear inverse problems). Moreover this connection could be fruitful in the other direction too where learning problems may provide examples of average-case problems that fool the sum-of-squares hierarchy. This section is based on ongoing discussions with Aram Harrow.

4 Hybrids of Worst-Case and Average-Case

In Section 2 we considered the prospect of making new structural assumptions to get around known intractability results. A complementary approach is to instead assume that an adversary has imprecise control over the instances he gives us, as in the smoothed analysis model [ST04, ST09]. In Section 4.1, we describe our work on analyzing tensor decompositions in this setting as well as some of the applications of this result in learning. In future work, we will explore the prospect of using semi-random models to weaken the standard distributional assumptions. The goal is to extend existing algorithms to this challenging setting to get more robust algorithms that perform better in practice (Section 4.2).

4.1 Prior Work: Smoothed Analysis of Tensor Decompositions

In *factor analysis* the goal is to take many variables and explain them away using fewer unobserved variables, called *factors*. It was introduced in a pioneering study by psychologist Charles Spearman, who used it to test his theory that there are fundamentally two types of intelligence – *verbal* and *mathematical* [Sp04]. This study has had a deep influence on modern psychology. However there is a mathematical complication that is called the *rotation problem* that for our purposes comes from the fact that a matrix decomposition $M = \sum_{i=1}^{R} a_i \otimes b_i$ is unique only if we add rather restrictive assumptions such as requiring the factors $\{a_i\}_i$ and $\{b_i\}_i$ to be orthonormal.

Tensor decompositions were first explored in the psychometrics community [Ha70, Kr77] because they are unique under much weaker conditions and offer a solution to the rotation problem. In particular, given a tensor $T = \sum_{i=1}^{R} a_i \otimes b_i \otimes c_i$ it suffices for the factors $\{a_i\}_i$, $\{b_i\}_i$ and $\{c_i\}_i$ to be linearly independent in order to ensure that the decomposition is unique up to reordering and rescaling. In such a case, there are a number of algorithms to construct this decomposition [Ha70] and this basic fact has been rediscovered many times. Tensor decompositions have found numerous applications in statistics [AMR09] and in learning latent variable models on phylogenetic trees [MR05], HMMs [MR05], topic models [AHK12, AFH$^+$12], mixture models [HK13] and also yield algorithms for community detection [AGH$^+$13]. Throughout this section let n denote the dimension of the factors.

What if we want to design algorithms that work in the overcomplete setting where R is much larger than n? The standard approach is to do so by flattening a higher order tensor T (whose factors are a_i, b_i, c_i, d_i, e_i and f_i):

$$\text{flat}(T) = \sum_{i=1}^{R} \Big(\underbrace{a_i \otimes_{KR} b_i}_{\text{factor}} \Big) \otimes \Big(\underbrace{c_i \otimes_{KR} d_i}_{\text{factor}} \Big) \otimes \Big(\underbrace{e_i \otimes_{KR} f_i}_{\text{factor}} \Big)$$

Here $\otimes_{KR}$ is the Khatri-Rao product that takes the tensor product of two n-dimensional vectors and flattens the result to get an n^2-dimensional vector. Hence flattening results in an $n^2 \times n^2 \times n^2$ tensor (see also Section 3.2). The key point is that the set of vectors $\{a_i \otimes_{KR} b_i\}$ can be linearly independent even if $R = n^2$ and consequently the Khatri-Rao product can be used to take algorithms that work with third-order tensors and boost them to work with flattened higher-order tensors to get algorithms that work in the overcomplete case [AMR09, BCV14]. Recently, Bhaskara, Charikar, the PI and Vijayaraghavan [BCM+14] introduced a smoothed analysis model to study these problems and showed that in this model boosting is extremely effective. This paved the way for getting new algorithms for learning mixtures of axis-aligned Gaussians and multi-view models that work when the number of components is any fixed polynomial in the dimension of the problem [BCM+14]. See also prior work of [GVX14] that studies overcomplete *independent component analysis*, and concurrent work of [ABG+14] that presents alternate approaches. The distinction is in [BCM+14] we establish that the properties of tensors we need hold in the smoothed analysis model with exponentially small failure probability.

4.2 Research Direction: Semi-Random Models for Matrix Completion

The usual recipe for designing algorithms with provable guarantees is to make an assumption about the *structure* of the solution (e.g. incoherence) as well as a *distributional* assumption about how our observations are generated. Are these latter assumptions reasonable? There is a serious danger that algorithms might work because they are over-exploiting statistical properties of the instances they are given. Algorithms that work on a particular type of distribution may fail completely if we slightly change the distribution, as we expect to be the case when we take a successful algorithm from one domain and use it in another.

Let us illustrate these issues by continuing our discussion of the matrix completion problem. The original motivation for studying it comes from collaborative filtering. In this problem, there is an unknown matrix $M \in \Re^{m \times n}$ that describes user preferences, and we assume that it is (approximately) low-rank and incoherent (Section 3.2). These are defensible assumptions and make sense. But we also assume that our observations $\Omega \subseteq [m] \times [n]$ of M are chosen uniformly at random.

In contrast, suppose we adopt another model (whose inspiration comes from the work on semi-random models for graph partitioning [FK01]). Suppose Ω is chosen uniformly at random, but what we actually observe is a superset $\widehat{\Omega} \supseteq \Omega$ chosen by an adversary after the fact. Intuitively, this should make the problem *easier* and not *harder*. Surprisingly all of the algorithms based on alternating minimization break, but those based on semidefinite programming do not! We remark that alternating minimization is still the algorithm of choice in large-scale applications, because it is faster and requires less space; making it work in this more challenging setting could lead to more robust but still practical algorithms for matrix completion.

Question 12. *Are there variants of alternating minimization for matrix completion that provably work in semi-random models?*

This is just one possible problem for which it makes sense to revisit existing algorithms from the perspective of semi-random models. We could do the same for learning mixtures of Gaussians (Section 3.1) and some other distribution learning problems, but the natural way to adapt semi-random models is considerably different and we omit the details.

There is an interesting parallel between the above questions and those that have been studied in the context of graph partitioning. Indeed there was a sequence of works studying the *stochastic block model* where the goal is to partition a random graph and recover the underlying clustering from which it was generated [BCL+84, Bo85, DF86, JS93, Mc01]. Feige and Kilian [FK01] introduced

a semi-random model where an adversary is allowed to add edges inside clusters and delete edges crossing between clusters. This breaks previous algorithms. Nevertheless, Feige and Kilian [FK01] gave a semidefinite program that succeeds in this more challenging setting. See also recent work [MMV12, MMV14] on related problems, albeit in a different model.

In fact, there is by now a standard recipe for designing algorithms that work in the semi-random model. We start with a semidefinite program and the main technical step is in showing that with high probability there is a solution to the dual program that certifies the true solution is optimal. But do we really need to solve a semidefinite program to get these type of robustness guarantees?

Question 13. *Are there algorithms for solving semi-random clustering problems that do not use semidefinite programming?*

In order to get the best of both worlds – algorithms that are robust and scale up to very large problems – it is natural to look for qualitatively new types of algorithms that also work in the semi-random setting. These types of algorithms could have considerable practical impact, because they could give us new ways to do things we thought we already knew how to do.

5 Broader Impact

Research: Our work has already been successful at bringing researchers in theory and machine learning closer together and has resulted in a number of joint workshops and meetings where ideas can cross freely over traditional boundaries that separate the fields. We are starting to make lasting inroads and I expect that this research program along with efforts of my collaborators will develop into a central area of research in both communities.

I am fully committed to popularizing the algorithmic perspective in other fields, not just machine learning but also in statistics, signal processing and applied mathematics more broadly. I believe that models like smoothed analysis and semi-random models that interpolate between worst-case and average-case analysis, can be applied even more broadly than they currently are. Conversely, in many of the applications in machine learning there are fundamentally different sorts of assumptions that are made and that we can learn from and use to revisit areas of theory that have been stagnant. Ultimately I hope that these projects will help theorists and practitioners find common ground, so that there will be significant interactions and diffusion of ideas between them.

Dissemination: A critical component of this research agenda is in disseminating results in other communities and building bridges between theory, machine learning, statistics and signal processing. I have organized many events that bring researchers from disparate areas together, and participated in many more. In Summer 2012, I co-organized a workshop with Sanjeev Arora and Moses Charikar called "Provable Bounds in Machine Learning" that had well over a hundred participants and we made all of these talks freely available online. In Fall 2012, I helped Sanjeev Arora design a new graduate course called "Is Machine Learning Easy?" which later served as a starting point when I designed and taught my own graduate course at MIT in Fall 2013. I gave an invited tutorial at UAI 2013 and co-organized a workshop at NIPS 2013 on topic modeling with experts in the field. This summer I was a long term visitor at Centre de Recerca Matematica, and I gave an invited talk at Curves and Surfaces 2014, which is a bi-annual meeting organized around signal processing and approximation theory.

In the near future, I plan on writing a mini-workshop proposal with some of the experts on the practical aspects of dictionary learning – Karin Schnass, Remi Gribonval and Martin Kleinsteuber – where we will bring together researchers from different fields to review recent progress and set new directions going forward. I will also be giving a summer school at MADALGO with Amr Ahmed,

Mikhail Belkin and Stefanie Jegelka. I also plan to considerably update the graduate seminar I taught in Fall 2013 and teach it again in Spring 2015, which will give me the opportunity to expand the monograph I wrote into a textbook. I am currently in talks with Cambridge University Press about publishing it. In addition, I plan to write a survey aimed at dispelling some of the common misconceptions of computer scientists about statistics. I believe that there are subtle mismatches in the definitions – arising from the difference between *asymptotic* and *effective* bounds.

Finally, I am a firm believer that the best way to disseminate results outside of theory is to actually test out your algorithms and work with researchers in the field to get new insights about real-world problems. Some of these attempts to export theory have been successful – most notably our work on topic modeling where we gave algorithms that outperformed state-of-the-art toolkits – and others have instead inspired me in the other direction, where often I find myself surprised at how well simple heuristics like alternating minimization and Gibbs sampling work and I instead set out to analyze them. I have worked with a number of undergraduates already, and a handful have gone on to graduate school to study machine learning, and I hope that they will take with them some of the perspectives of theory that we discussed and thought about together.

Outreach: I have participated in a number of programs that are specifically designed to encourage students from underrepresented groups to pursue computer science, and I plan on becoming involved in similar such programs at MIT. Last summer, I taught a three-week intensive course called "The Math Behind the Machine" in the NJ Governor's School held in Rutgers, where the audience consisted of hand-selected high school students from across the state who had just finished their junior year. These students started applying to colleges just afterwards, so it is was an ideal time to introduce them to theoretical computer science and put it on their minds. This program encourages participation from underrepresented groups in particular, and I had many such students in my class. Programs like this give them a unique opportunity to be exposed to material that they would otherwise would not have access to.

In previous summers, I also gave lectures to undergraduate students as part of Rajiv Ghandi's summer school held in Princeton. This program is unlike many of its peers in that it targets *only* students from institutions where there is no opportunity to do research at the undergraduate level. The goal is to introduce these students to research level topics and encourage them to apply to graduate school by giving them a taste of what is out there beyond their classwork. This summer I will be involved in several such outreach programs; in particular I will be a speaker and a judge (for final projects) in SPUR/RSI. And in the Fall, I will speak at the opening dinner for the Society of Women Engineers (SWE).

6 Educational Plan

6.1 Mentoring

I have already had the opportunity to collaborate with several graduate and undergraduate students during my postdoc, and I fully believe that training the next wave of students and equipping them with the right tools is an integral part of the success of this research plan. I plan to take on a student in the fall, and have been working with several students who are currently supervised by other faculty members. In addition I am looking forward to further interactions with students outside of theory too. In the past, such projects seem to go in different directions than I would normally think of exploring, and are rewarding in a unique way. Finally, I am eager to take on undergraduate students, and I believe it is my responsibility (and privilege) to give motivated students a glimpse of how exciting research can be! When I was an undergraduate at Cornell, getting the chance to work directly with Eva Tardos was the single biggest reason that I decided

to switch fields and go on to graduate school in theoretical computer science. It added a level of excitement that was something else entirely from my experiences doing coursework.

6.2 Course Development

In Fall 2013, I developed and taught a new graduate course titled "Algorithmic Aspects of Machine Learning" which covered numerous topics in this proposal and many more and had over sixty enrolled students from a wide range of backgrounds. Much of the material in this area is interdisciplinary and draws liberally from theoretical computer science, mathematics, optimization and machine learning and the course was aimed at equipping students with the right tools from these interconnected areas. I wrote a monograph that is freely available on my webpage, which in the meantime has been used at a number of universities as a source for independent reading projects (MIT, Cornell, UT Austin) as well as subject material for qualifying exams (MIT, Princeton). I plan to considerably update this course and I will offer it again in Spring 2015. This will also give me the opportunity to expand my monograph into a textbook, although I will keep it online.

Ultimately, I plan to develop a new undergraduate course on machine learning that emphasizes the theoretical foundations such as boosting, experts, bandits, Markov decision processes, PAC learning and property testing. In my experience, some of our best undergraduates – particularly those majoring in $18C$: Mathematics with Computer Science – are theory-minded and could greatly benefit from a version of the standard undergraduate course that fits better with their mathematical training. These students often go on to heavily use machine learning in industry, and such a course would much better equip them to not just be users of machine learning but also to be able to think critically about it and make different sorts of contributions.

7 Results of Prior Support

From September 2011 to August 2013 I was supported by an NSF Computing and Innovation Fellowship (CIF-D-013, CIF-E-013) while I was a postdoc at the Institute for Advanced Study. The grant covered my salary, health insurance and travel expenses and resulted in nine publications which were presented at top conferences, and numerous other prestigious venues. This work covers a broad range of topics including nonnegative matrix factorization ([AGK$^+$12], STOC 2012, [Mo13], SODA 2013), topic modeling ([AGM12], FOCS 2012, [AGH$^+$13], ICML 2013), independent component analysis ([AGM$^+$12], NIPS 2012), extended formulations ([BM13], STOC 2013), robust statistics ([HM13b], COLT 2013), combinatorics ([AMS12], STOC 2012) and population recovery ([MS13], FOCS 2013), and most of it was described at length in the body of the proposal.

Intellectual Merit The questions that I worked on while supported by an NSF CI Fellowship are of widespread importance. Our work led to new algorithms for basic problems in machine learning, some of which outperform the state-of-the-art. And yet these algorithms come with provable guarantees and hence we know when and why they work, something that is often lacking for many popular machine learning approaches.

Broader Impact Our work has already been successful at bringing researchers in theory and machine learning closer together and there has been considerable follow-up from both communities. I gave a number of invited tutorials and organized cross-disciplinary workshops (see Section 5). These have been an excellent opportunity to disseminate the viewpoints of theoretical computer science to a broader audience.

8 References Cited

[AM05] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT)*, pages 458–469, 2005.

[AAJ+14] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, R. Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. In *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, pages 123–137, 2014.

[AEB06] M. Aharon, M. Elad and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing*, 54(11):4311–4322, 2006.

[AMR09] E. Allman, C. Matias and J. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A):3099–3132, 2009.

[AMS12] N. Alon, A. Moitra, and B. Sudakov. Nearly complete graphs decomposable into large induced matchings and their applications. In *Proceedings of the 44th ACM Symposium on Theory of Computing (STOC)*, pages 1079–1090, 2012.

[AFH+12] A. Anandkumar, D. Foster, D. Hsu, S. Kakade, Y. Liu. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 926–934, 2012.

[AGH+13] A. Anandkumar, R. Ge, D. Hsu and S. Kakade. A tensor spectral approach to learning mixed membership community models. In *Proceedings of the 26th Conference on Learning Theory (COLT)*, pages 867–881, 2013.

[AHK12] A. Anandkumar, D. Hsu and S. Kakade. A method of moments for mixture models and hidden markov models. In *Proceedings of the 25th Conference on Learning Theory (COLT)*, pages 1–33, 2012.

[ABG+14] J. Anderson, M. Belkin, N. Goyal, L. Rademacher and J. Voss. The more, the merrier: The blessing of dimensionality for learning large gaussian mixtures. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, pages 1135–1164, 2014.

[AGH+13] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 280–288, 2013.

[AGK+12] S. Arora, R. Ge, R. Kannan and A. Moitra. Computing a nonnegative matrix factorization – provably. In *Proceedings of the 44th ACM Symposium on Theory of Computing (STOC)*, pages 145–162, 2012.

[AGM12] S. Arora, R. Ge and A. Moitra. Learning topic models - going beyond SVD. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1–10, 2012.

[AGM14] S. Arora, R. Ge and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, pages 779–806, 2014.

[AGM+12] S. Arora, R. Ge, A. Moitra and S. Sachdeva. Provable ICA with unknown gaussian noise, and implications for gaussian mixtures and autoencoders. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 2384–2392, 2012.

[AK05] S. Arora and R. Kannan. Learning mixtures of separated nonspherical gaussians. *Annals of Applied Probability*, 15(1A):69-92, 2005.

[BBG13] M. Balcan, A. Blum and A. Gupta. Clustering under approximation stability. *Journal of the ACM*, 60(2):1–34, 2013.

[BFH+12] B. Barak, F. Brandao, A. Harrow, J. Kelner, D. Steurer and Y. Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the 44th ACM Symposium on Theory of Computing (STOC)*, pages 307–326, 2012.

[BKS14a] B. Barak, J. Kelner and D. Steurer. Rounding sum-of-squares relaxations. In *Proceedings of the 46th ACM Symposium on Theory of Computing (STOC)*, pages 31–40, 2014.

[BKS14b] B. Barak, J. Kelner and D. Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. *Manuscript*, 2014.

[BS10] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 103–112, 2010.

[BR13] Q. Berthet and P. Rigollet. Computational lower bounds for sparse principal component detection. In *Proceedings of the 26th Conference on Learning Theory (COLT)*, pages 1046–1066, 2013.

[BCM+14] A. Bhaskara, M. Charikar, A. Moitra and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the 46th ACM Symposium on Theory of Computing (STOC)*, pages 594–603, 2014.

[BCV14] A. Bhaskara, M. Charikar and A. Vijayaraghavan. Uniqueness of tensor decompositions with applications to polynomial identifiability. In *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, pages 742–778, 2014.

[BRR+12] V. Bittorf, B. Recht, C. Re and J. Tropp. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 1223–1231, 2012.

[Bl12] D. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[BL07] D. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

[BNJ03] D. Blei, A. Ng and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[BS95] A. Blum and J. Spencer. Coloring random and semi-random k-colorable graphs. *Journal of Algorithms*, 19(2):204–234, 1995.

[Bo85] R. Bopanna. Eigenvalues and graph bisection: An average-case analysis. In *Proceedings of the 26th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 280–295, 1985.

[BCY11] F. Brandao, M. Christandl and J. Yard. A quasipolynomial-time algorithms for the quantum separability problem. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, pages 343–352, 2011.

[BM13] M. Braverman and A. Moitra. An information complexity approach to extended formulations. In *Proceedings of the 44th ACM Symposium on Theory of Computing (STOC)*, pages 161–170, 2013.

[BV08] S. C. Brubaker and S. Vempala. Isotropic PCA and affine-invariant clustering. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 551–560, 2008.

[BCL+84] T. Bui, S. Chaudhuri, T. Leighton and M. Sipser. Graph bisection algorithms with good average case behavior. In *Proceedings of the 25th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 181–192, 1984.

[CES+13] E. Candes, Y. Eldar, T. Strohmer and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[CLM+11] E. Candes, X. Li, Y. Ma and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

[CR08] E. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Math.*, 9(6):717–772, 2008.

[CRT06] E. Candes, J. Romberg and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[CT10] E. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053-2080, 2010

[CJ13] V. Chandrasekaran and M. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13)E1181–E1190, 2013.

[CRP+12] V. Chandrasekaran, B. Recht, P. Parrilo and A. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Math.*, 12(6)805–849, 2012.

[Ch96] J. Chang. Full reconstruction of markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences*, 137(1):51–73, 1996.

[CR93] J. Cohen and U. Rothblum. Nonnegative ranks, decompositions and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190(1):149–168, 1993.

[Co94] P. Comon. Independent component analysis: A new concept? *Signal Processing*, 36(3):287–314, 1994.

[Co90] G. Cooper. The computational complexity of probabilistic influence using bayesian belief networks. *Artificial Intelligence*, 42(2-3):393–405, 1990.

[DL93] P. Dagum and M. Luby. Approximate probabilistic inference in bayesian networks is NP hard. *Artificial Intelligence*, 60(1):141–153, 1993.

[DL97] P. Dagum and M. Luby. An optimal approximation algorithm for bayesian inference. *Artificial Intelligence*, 93(1-2)1–27, 1997.

[Da99] S. Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 634–644, 1999.

[DS00] S. Dasgupta and L. J. Schulman. A two-round variant of EM for gaussian mixtures. In *16th Conference on Uncertainty and Artificial Intelligence (UAI)*, pages 152–159, 2000.

[DDL+90] S. Deerwester, S. Dumais, T. Landauer, G. Furnas and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38, 1977.

[Do06] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[DE03] D. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1-minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[DH99] D. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 1999.

[DS89] D. Donoho and P. Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989.

[DS03] D. Donoho and V. Stodden. When does nonnegative matrix factorization give the correct decomposition into parts? In *Advances in Neural Information Processing Systems 16 (NIPS)* pages 1141–1148, 2003.

[DF86] M. Dyer and A. Frieze. Fast solution of some random NP-hard problems. In *Proceedings of the 27th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 331–336, 1986.

[El10] M. Elad. *Sparse and Redundant Representations*. Springer, 2010.

[EAH99] K. Engan, S. Aase and J. Hakon-Husoy. Method of optimal directions for frame design. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5:2443–2446, 1999.

[Fa02] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[FK01] U. Feige and J. Kilian. Heuristics for semi random graph problems. *Journal of Computer and System Sciences*, 63(4):639–671, 2001.

[FJK96] A. Frieze, M. Jerrum, R. Kannan. Learning linear transformations. In *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 359–368, 1996.

[GG84] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[GV14] N. Gillis and S. Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4)698–714, 2014.

[GVX14] N. Goyal, S. Vempala and Y. Xiao. Fourier PCA. In *Proceedings of the 46th ACM Symposium on Theory of Computing (STOC)*, pages 584–593, 2014.

[GN03] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

[Gu03] L. Gurvits. Classical deterministic complexity of Edmonds' problem and quantum entanglement. In *Proceedings of the 35th ACM Symposium on Theory of Computing (STOC)*, pages 10–19, 2003.

[HSS11] B. Haeupler, B. Saha and A. Srinivasan. New constructive aspects of the lovász local lemma. *Journal of the ACM*, 58(6):1–28, 2011.

[HS13] Y. Halpern and D. Sontag. Unsupervised learning of noisy-or bayesian networks. In *29th Conference on Uncertainty and Artificial Intelligence (UAI)*, pages 272–281, 2013.

[Ha14] M. Hardt. Understanding alternating minimization for matrix completion. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2014, to appear.

[HM13a] M Hardt and A. Moitra. Algorithms and hardness for robust subspace recovery. In *Proceedings of the 26th Conference on Learning Theory (COLT)*, pages 354–375, 2013.

[HM13b] A. Harrow and A. Montanaro. Testing product states, quantum merlin-arthur games and tensor optimization. *Journal of the ACM*, 60(1):1–44, 2013.

[Ha70] R. Harshman. Foundations of the PARFAC procedure: model and conditions for an 'explanatory' multi-mode factor analysis. *UCLA Working Papers in Phonetics*, pages 1–84, 1970.

[HL13] C. Hillar and L-H. Lim. Most tensor problems are NP-hard. *Journal of the ACM*, 60(6):1–39, 2013.

[Hof99] T. Hofmann. Probabilistic latent semantic analysis. In *15th Conference on Uncertainty and Artificial Intelligence (UAI)*, pages 289–296, 1999.

[HK13] D. Hsu and S. Kakade. Learning mixtures of spherical gaussians: Moment methods and spectral decompositions. In *4th Annual Innovations in Theoretical Computer Science (ITCS)*, pages 11–20, 2013.

[JNS13] P. Jain, P. Netrapalli and S. Sanghavi. Low rank matrix completion using alternating minimization. In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*, pages 665–674, 2013.

[JHS13] Y. Jernite, Y. Halpern and D. Sontag. Discovering hidden variables in noisy-or networks using quartet tests. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 2355–2363, 2013.

[JSV04] M. Jerrum, A. Sinclair and E. Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM*, 51(4):671–697, 2004.

[JS93] M. Jerrum and G. Sorkin. Simulated annealing for graph bisection. In *Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 94–103, 1993.

[JB02] M. Jordan and C. Bishop. *An Introduction to Graphical Models*. MIT Press, 2002.

[KMV10] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 553-562, 2010.

[KMV12] A. T. Kalai, A. Moitra, and G. Valiant. Disentangling gaussians. *Communications of the ACM*, 55(2):113–120, 2012.

[KMO10] R. Keshavan, A. Montanari and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980-2998, 2010

[KF09] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[Kr77] J. Kruskal. Three-way Arrays: Rank and uniqueness of trilinear decompositions. *Linear Algebra and Applications*, 18(2):95–138, 1977.

[KSK13] A. Kumar, V. Sindhwani and P. Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 231–239, 2013.

[La01] J. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[LBR+06] H. Lee, A. Battle, R. Raina and A. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems 19 (NIPS)*, pages 801–808, 2006.

[LS99] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788-791, 1999.

[LRA93] S. Leurgans, R. Ross and R. Abel. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993.

[LS00] M. Lewicki and T. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

[LM07] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 633–640, 2007.

[Li95] B. Lindsay. *Mixture Models: Theory, Geometry and Applications*. Institute for Mathematical Statistics, 1995.

[MMV12] K. Makarychev, Y. Makarychev and A. Vijayaraghavan. Approximation algorithms for semirandom graph partitioning problems. In *Proceedings of the 44th ACM Symposium on Theory of Computing (STOC)*, pages 367–384, 2012.

[MMV14] K. Makarychev, Y. Makarychev and A. Vijayaraghavan. Constant factor approximation for balanced cut in the PIE model. In *Proceedings of the 46th ACM Symposium on Theory of Computing (STOC)*, pages 41–49, 2014.

[Ma98] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic-Press, 1998.

[Mc01] F. McSherry. Spectral partitioning of random graphs. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 529–537, 2001.

[Mo13] A. Moitra. An almost optimal algorithm for computing nonnegative rank. In *Proceedings of the 24th ACM-SIAM Symposium on Discrete Algorithms*, pages 1454–1464, 2013.

[MS13] A. Moitra and M. Saks. A polynomial time algorithm for lossy population recovery. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 110–116, 2013.

[MV10] A. Moitra and G. Valiant. Setting the polynomial learnability of mixtures of gaussians. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 93–102, 2010.

[MT10] R. Moser and G. Tardos. A constructive proof of the general lovász local lemma. *Journal of the ACM*, 57(2):1–15, 2010.

[MR05] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden markov models. In *Proceedings of the 37th ACM Symposium on Theory of Computing (STOC)*, pages 366–375, 2005.

[OF97] B. Olshausen and B. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):331–3325, 1997.

[PRT$^+$00] C. Papadimitriou, P. Raghavan, H. Tamaki and S. Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.

[Pa00] P. Parrilo. *Structured Semidefinite Programs and Semialgebraic Geometry Method in Robustness and Optimization*. PhD thesis, California Institute of Technology, 2000.

[Pe82] J. Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second National Conference on Artificial Intelligence (AAAI)*, pages 133–136, 1982.

[Pe88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[Pe94] K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*, 185:71–110, 1894.

[RBL07] M. Ranzato, Y. Boureau and Y. LeCun. Sparse feature learning for deep belief networks. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 1185–1192, 2007

[Re11] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[RFP10] B. Recht, M. Fazel and P. Parrilo. Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[Se14] S. Seung. Personal Communication, 2014.

[Sp04] C. Spearman. General intelligence. *American Journal of Psychology*, 15(2):201–293, 1904.

[ST04] D.A. Spielman, S.H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004.

[ST09] D.A. Spielman, S.H. Teng. Smoothed analysis: An attempt to explain the behavior of algorithms in practice. *Communications of the ACM*, 52(10):76-84, 2009.

[SWW12] D. Spielman, H. Wang and J. Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Conference on Learning Theory (COLT)*, pages 1–37, 2012.

[SS05] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *Proceedings of the 18th Conference on Learning Theory (COLT)*, pages 545–560, 2005.

[Va09] S. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364-1377, 2009.

[VW04] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[Ve13] S. Venkatasubramanian. Computational geometry column 55: New developments in nonnegative matrix factorization. *SIGACT News*, 44(1):70–78, 2013.

[WJ08] M. Wainwright and M. Jordan. *Graphical Models, Exponential Families, and Variational Inference.* Foundations and Trends in Machine Learning, pages 1–305, 2008.

[Ya91] M. Yannakakis. Expressing combinatorial optimization problems by linear programs. *Journal of Computer and System Sciences*, 43(3):441–466, 1991.

[ZWJ14] Y. Zhang, M. Wainwright and M. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, pages 921–948, 2014.

[ZBG14] T. Zhou, J. Bilmes and C. Guestrin. Divide-and-conquer learning by anchoring a conical hull. *arXiv:1406:5752*, 2014.