Journal of Statistical Mechanics:
Theory and Experiment

PAPER • OPEN ACCESS

Disordered systems insights on computational hardness

To cite this article: David Gamarnik *et al J. Stat. Mech.* (2022) 114015

OPEN ACCESS

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: ML 2022

Disordered systems insights on computational hardness

David Gamarnik[1], **Cristopher Moore**[2] **and Lenka Zdeborová**[3,*]

[1] Operations Research Center and Sloan School of Management, MIT, Cambridge, MA 02139, United States of America
[2] Santa Fe Institute, Santa Fe, NM 87501, United States of America
[3] SPOC Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Route Cantonale, CH-1015 Lausanne, Switzerland
E-mail: gamarnik@mit.edu, moore@santafe.edu and lenka.zdeborova@epfl.ch

Received 18 October 2022
Accepted for publication 18 October 2022
Published 24 November 2022

Online at stacks.iop.org/JSTAT/2022/114015
https://doi.org/10.1088/1742-5468/ac9cc8

Abstract. In this review article we discuss connections between the physics of disordered systems, phase transitions in inference problems, and computational hardness. We introduce two models representing the behavior of glassy systems, the spiked tensor model and the generalized linear model. We discuss the random (non-planted) versions of these problems as prototypical optimization problems, as well as the planted versions (with a hidden solution) as prototypical problems in statistical inference and learning. Based on ideas from physics, many of these problems have transitions where they are believed to jump from easy (solvable in polynomial time) to hard (requiring exponential time). We discuss several emerging ideas in theoretical computer science and statistics that provide rigorous evidence for hardness by proving that large classes of algorithms fail in the conjectured hard regime. This includes the overlap gap property, a particular mathematization of clustering or dynamical symmetry-breaking, which can be used to show that many algorithms that are local or robust to changes in their input fail. We also discuss the sum-of-squares hierarchy, which places bounds

*Author to whom any correspondence should be addressed.

on proofs or algorithms that use low-degree polynomials such as standard spectral methods and semidefinite relaxations, including the Sherrington–Kirkpatrick model. Throughout the manuscript we present connections to the physics of disordered systems and associated replica symmetry breaking properties.

Keywords: cavity and replica method, message-passing algorithms, statistical inference, typical-case computational complexity

Contents

J. Stat. Mech. (2022) 114015

1. Introduction

Computational complexity theory [1] aims to answer the question of what problems can be solved by computers. More specifically, it aims to classify computational problems according to the resources (usually time or memory) needed to solve them, and how these resources scale with the problem size. Computationally hard problems are those that can be solved in principle but require prohibitively large amounts of resources, such as a running time that grows exponentially with the problem size.

The most iconic result of computational complexity theory is the existence of so-called NP-complete problems [2]. These problems, of which hundreds have been identified, are all hard unless P = NP, in which case they are all easy. But if P = NP, anything which is easy to check would be easy to find. All modern cryptosystems would be breakable; it would be easy to find short proofs of unsolved mathematics problems or elegant theories to explain empirical data, without any need for insight or intuition. Even evolution would gain shortcuts: it would be easy to design proteins with certain structures, rather than having to search for them by exploring a vast space of possible amino acid sequences. This would violate many of our deepest beliefs about the nature of search, proof, and even creativity. For these and other reasons, resolving the P $\neq$ NP conjecture is considered the most important problem of theoretical computer science, and one of the most important open problems in mathematics more generally.

Since we believe some problems are computationally hard, the question becomes the nature of this hardness. What is it about a problem's structure that defeats polynomial-time algorithms? Since the late 1980s and early 1990s (e.g. [3–5]), some researchers have looked to the physics of disordered systems as one source of hardness. This comes very naturally since, for many canonical models such as spin glasses, finding a ground state is easily shown to be NP-hard (i.e. at least as hard as any NP-complete problem).

Physical dynamics is itself computationally limited by the locality of interactions, and physics-based algorithms such as Markov chain Monte Carlo and simulated annealing are subject to the same limits. In glassy systems these algorithms often get stuck in metastable states, or take exponential time to cross free energy barriers. Unless there is some miraculous algorithmic shortcut for exploring glassy landscapes—which seems unlikely, except for a few isolated cases—it seems likely that no polynomial-time algorithms for these problems exist.

In this paper we review some current areas of research on the connections between theory of disordered systems and computational hardness, and attempts to make this physical intuition mathematically rigorous. We will discuss two types of computational problems: *optimization problems* where one aims to minimize an objective function (such as the energy) over a set of variables, and *signal recovery* or *inference problems* where a signal is observed but obscured by noise, and the task is to reconstruct it (at least approximately) from these observations. In section 2 we define canonical examples of both these problems, stressing their relationship to disordered

J. Stat. Mech. (2022) 114015

systems studied in physics as well as their broad applicability to modelling various computational tasks. In section 3 we discuss recent results on computational hardness of optimization problems based on the *overlap gap property*, which formalizes the idea that solutions are widely separated from each other by energy barriers. Section 4 switches to signal recovery/inference problems and presents a rather generic picture that emerges from the study of phase transition in those problems. Finally, section 5 discusses the *sum-of-squares hierarchy*, another approach to proving computational lower bounds.

2. Two problems in optimization and inference: definitions

2.1. The spiked tensor model and spin glasses

One of the models we will consider from the statistics and computational perspective is a natural variant of the spin glass model with a 'planted signal' to be learned or reconstructed—physically, a low-energy state built into the landscape. It is called the *spiked tensor model* or *tensor PCA*, and is defined as follows. Given a hidden vector $u \in \mathbb{R}^N$, we observe the following tensor:

$$Y = \lambda u^{\otimes p} + J. \tag{1}$$

Here $u^{\otimes p}$ is the p-fold tensor outer product of u, and J is a $N \times \cdots \times N$ tensor describing the noise. We will assume that the entries $J_{i_1,\ldots,i_p}$ with $1 \leqslant i_1 < i_2 < \ldots < i_p \leqslant N$ are drawn i.i.d. from some common distribution with mean zero and variance σ^2, such as the normal distribution $\mathcal{N}(0,1)$. The other entries of J are fixed by a symmetry assumption, $J_{i_{\sigma(1)},\ldots,i_{\sigma(p)}} = J_{i_1,\ldots,i_p}$ for all permutations σ of $[p] = \{1, 2, \ldots, p\}$.

We can think of λ as a signal-to-noise ratio, parameterizing how strongly the signal u affects the observation Y compared to the noise J. In order to look for phase transitions in the hardness of reconstructing the planted vector u, we will allow λ to scale in various ways with N. We can also let J's variance σ^2 vary with N, but in most of the paper we will take it to be 1.

We can consider variants of this problem where different types of restrictions are placed on u. One is to take $u \in S_N$ where S_N is the N-dimensional sphere $\{u : \|u\|^2 = N\}$. Another choice is to take Boolean values on the N-dimensional hypercube or equivalently Ising spins, $u \in B_N$ where $B_N = \{\pm 1\}^N$. We can also impose sparsity by demanding that a fraction ρ of u's entries are nonzero, writing $u \in B_{N,\rho}$ where $B_{N,\rho} = \{u \in \{\pm 1, 0\}^N : \|u\|_1 = N\rho\}$. In terms of Bayesian inference, we take the uniform measure on each of these sets to be a prior on u.

The variant $p = 2$, i.e. the spiked matrix model, is particularly widely studied. It is also known as the spiked covariance model, or as low-rank matrix estimation, since $u \otimes u$ is a rank-1 approximation of Y [6, 7].

The general questions to be addressed in this model are (a) can we learn, or reconstruct, the planted vector u from the observation Y? and (b) can we do this with an efficient algorithm, i.e. one whose running time is polynomial in N? (We assume p is a constant, so polynomial in N is equivalent to polynomial in the size N^p of the observed data.) Since reconstructing u exactly is often impossible, we are interested in

J. Stat. Mech. (2022) 114015

approximate reconstruction, i.e. producing an algorithmic estimate $\hat{u} = \hat{u}(Y)$ which has a nontrivial correlation with the ground truth u: for instance, by having an overlap $(1/N)\langle \hat{u}, u\rangle$ bounded above zero with high probability.

Question (a) is an information-theoretic or statistical question, unconcerned with computational resources. Using the theory of Bayesian inference we can write the posterior distribution,

$$P(z|Y) = \frac{1}{\mathcal{Z}} P(z)\, P(Y|z) \quad \text{where} \tag{2}$$

$$P(Y|z) = \prod_{1\leqslant i_1<i_2<\ldots<i_p\leqslant N} \mathcal{N}(Y_{i_1,\ldots,i_p} - \lambda z_{i_1} z_{i_2} \ldots z_{i_p}, 1), \tag{3}$$

where for concreteness we considered the elements of the noise J to be Gaussian with variance 1. (Due to universality properties, e.g. [7], this is not very restrictive for what follows.) Note that the partition function or normalization factor $\mathcal{Z}$ depends both on the observed tensor Y, the prior $P(z)$, and the parameters λ, σ of the likelihood $P(Y|z)$. In our notation we drop this explicit dependence.

The posterior distribution $P(z|Y)$ is an exponentially complicated object. However, for several natural loss functions including the overlap $\langle \hat{u}, u\rangle$ and the ℓ_2 error $\|\hat{u} - u\|^2$, the best possible estimator $\hat{u}$ depends only on the marginals $P(z_i|Y)$. Thus question (b) boils down to whether, given Y, we can approximate these marginals with a polynomial-time algorithm.

Another common approach in statistics is the maximum likelihood estimator[4] (MLE) where we set $\hat{u}$ to the z that maximizes $P(Y|z)$. In the Gaussian case (3), we have

$$\begin{aligned} P(Y|z) &\propto \exp\left[-\frac{1}{2} \sum_{1\leqslant i_1<i_2<\ldots<i_p\leqslant N} \left(Y_{i_1,\ldots,i_p} - \lambda z_{i_1} z_{i_2} \ldots z_{i_p}\right)^2\right] \\ &= \exp\left[-\frac{1}{p!}\left(\frac{1}{2}\|Y\|^2 + \frac{\lambda^2}{2}\|z\|^{2p} - 2\langle Y, z^{\otimes p}\rangle\right)\right], \end{aligned} \tag{4}$$

where in the limit of large N we ignore terms with repeated indices, and where

$$\langle Y, z^{\otimes p}\rangle = \sum_{1\leqslant i_1<i_2<\ldots<i_p\leqslant N} Y_{i_1,\ldots,i_p} z_{i_1} z_{i_2} \ldots z_{i_p}. \tag{5}$$

Since $\|Y\|^2$ is fixed by the observed data, and since $\|z\|^2 = N$ if $z \in S_N$ or B_N (or ρN if it is in $B_{N,\rho}$) then the MLE is the z that maximizes (5). But this is exactly the ground state of a p-spin model with coupling tensor Y, with spherical or Ising spins if z is in S_N or B_N respectively.

In particular, if $\lambda = 0$ so that $Y = J$, we have a p-spin model with Gaussian random couplings and Hamiltonian

$$E(z) = -\sum_{1\leqslant i_1<i_2<\ldots<i_p\leqslant N} J_{i_1,\ldots,i_p} z_{i_1} z_{i_2} \ldots z_{i_p}. \tag{6}$$

[4] It should be noted that while the MLE and similar extremization-based approaches are very popular in statistics, they are typically suboptimal in high-dimensional settings: that is, they do not optimize the overlap or minimize the ℓ_2 error.

J. Stat. Mech. (2022) 114015

Studying the optimization landscape of this un-planted problem may seem irrelevant to the inference problem of reconstructing u from Y. But in addition to being physically natural, as a generalization of the Sherrington–Kirkpatrick model [8] which corresponds to the case $p = 2$ and $z \in B_N$, it serves both as a starting point for the inference problem and as a null model where there is no signal at all.

Thus in addition to the *reconstruction problem* where we assume that Y is drawn from the planted model (1) and we want to learn u, we will also consider the *detection* problem. That is, given Y, we want to determine whether it is drawn from the planted model, or the un-planted model where $Y = J$. Like reconstruction, this hypothesis testing problem may or may not be information-theoretically possible. If it is, it may or may not have a polynomial-time algorithm that succeeds with high probability.

In the literature there are many variants of the spiked tensor model. The signal can be of higher rank, i.e. $\sum_j u_j^{\otimes p}$ for multiple planted vectors u_j, or one can plant a subspace rather than a vector. In addition to being non-Gaussian, the noise can be nonadditive, binary or sparse. And the observation could consist of multiple tensors with different p rather than a single Y. All these variants have their own interest and applications; see examples in e.g. [7, 9]. In what follows we will also sometimes refer to sparse versions of the spiked matrix model, such as the stochastic block model which is popular in network science as a model of community structure (see e.g. [10]).

2.2. The generalized linear model and perceptrons

Another class of problems we will consider in this paper is the generalized linear model (GLM). Again, a planted vector $u \in \mathbb{R}^N$ is observed through a set of noisy observations, but this time through approximate linear combinations $Y_1, \ldots, Y_P$:

$$Y_i \sim P_{\text{out}}\left(Y_i \,|\, \sum_{a=1}^{N} J_{ia} u_a\right). \tag{7}$$

Here $J \in \mathbb{R}^{P \times N}$ is a known matrix whose entries are i.i.d. with zero mean and variance σ^2, and P_{out} is some noisy channel. In other words, $f(j) = \langle j, u \rangle$ is an unknown linear function from $\mathbb{R}^N$ to $\mathbb{R}$, and our goal is to learn this function—that is, to reconstruct u—from noisy observations of its values $f(j_1), \ldots, f(j_P)$ at P random vectors where j_i is the ith row of J. In machine learning we would say that the set of tuples (j_i, Y_i) are the training data, and by learning u we can generalize to $f(j)$ for new values of j.

The main questions for the GLM are the same as for the spiked tensor model: (a) whether it is information-theoretically possible to learn the signal u given J and Y, and (b) whether there are efficient algorithms that do that. Again Bayesian inference aims at computing the marginals of a posterior

$$P(z|Y, J) = \frac{1}{\mathcal{Z}} P(z) \prod_{i=1}^{P} P_{\text{out}}\left(Y_i \,|\, \sum_{a=1}^{N} J_{ia} u_a\right). \tag{8}$$

Here the partition function $\mathcal{Z}$ depends implicitly on the matrices Y and J as well as on the parameters of the probability P_{out} and of the prior $P(z)$. As in tensor PCA, u can

J. Stat. Mech. (2022) 114015

be restricted to S_N, B_N or some other set, and we will assume that its Bayesian prior is uniform over this set.

Another family of estimators minimize some loss function ℓ, perhaps with a regularization term with strength λ:

$$\mathcal{L}(z) = \sum_{i=1}^{P} \ell\left(Y_i, \sum_{a=1}^{N} J_{ia} z_a\right) + \lambda \sum_{a=1}^{N} r(z_a). \tag{9}$$

In a linear regression context, J is the observed data and Y the observed dependent variable, and (9) seeks to minimize the empirical risk ℓ. A typical regularization term might be $r(z_a) = |z_a|$, giving the 'lasso' or L_1 regularization $\lambda\|z\|_1$ which pushes z towards sparse vectors.

The GLM captures many versions of high-dimensional linear regression, and covers a broad range of applications and situations. In signal processing or imaging u would be the N-dimensional signal/image to be reconstructed from measurements Y, where J is the measurement matrix and the channel P_{out} typically consists of additive Gaussian noise. In compressed sensing we consider the under-determined case $N > P$, but with a sparse prior on the signal u.

Just as for the spiked tensor model the signal u can be seen as a planted solution to recover from Y and J. The version of the model where the distribution of Y is independent of u is well known in the statistical physics literature as the perceptron. The variant with $z \in S_N$ is the spherical perceptron [11], and $z \in B_N$ gives the binary perceptron [11, 12]. The perceptron model is particularly important as its study started the line of work applying physics of disordered systems to understanding supervised learning in artificial neural networks. The recent major success of methods based on deep learning [13] only added importance and urgency to this endeavour.

3. Hardness of optimizing *p*-spin models: the overlap gap property and implications

In this section we discuss the algorithmic hardness of the problem (6) of finding near ground states of p-spin models using the overlap gap property (OGP). The OGP is a property of solution space geometry which roughly speaking says that near optimal solutions should be either close or far from each other. It is intimately related to the replica symmetry breaking (RSB) property and the clustering (also sometimes called shattering) property exhibited by some constraint satisfaction problems. In fact it emerged directly as way to establish the presence of the shattering property in constraint satisfaction problems [14, 15]. There are important distinctions, however, between RSB, clustering and OGP, which we will discuss as well. A survey of OGP-based methods is in [16]. Our main focus is to illustrate how OGP presents a barrier to a certain class of algorithms as potential contenders for finding near ground states. Loosely speaking, it is the class of algorithms exhibiting input stability (noise insensitivity), thus revealing deep and intriguing connections with a rich field of Fourier analysis of Boolean functions [17]. Many important algorithms are special

J. Stat. Mech. (2022) 114015

cases of this class, including approximate message passing (AMP) [18], low-degree polynomials [19, 20], Langevin dynamics [19], and low-depth Boolean circuits [21]. OGP was also established to be a barrier for certain types of quantum algorithms, specifically quantum approximate optimization algorithms (QAOA) [22–24], using a slightly different implementation of the stability argument. We will therefore conclude that the values produced by these algorithms are bounded away from optimality. We will discuss various extensions of the OGP, including the multi-overlap gap property (m-OGP), which will allow us to bring the algorithmic barriers to the known algorithmic thresholds. In the case of the p-spin models these thresholds are achieved by AMP. It is entirely possible that models in the OGP regime do not admit any polynomial time algorithms, which at this stage is evidenced by just the lack of those. Proving this say modulo P $\neq$ NP assumption does not yet appear to be within the reach of the known techniques.

3.1. *p*-spin model, ground states and algorithms

We recall that our focus is the optimization problem (6). The optimization is over choice of z in some space Θ_N which for the purposes of this section is either S_N or B_N. The former is referred to as spherical p-spin model and the latter is called the Ising p-spin model. We assume that the variance σ_N^2 of the i.i.d. entries of the tensor J is $N^{-(p+1)}$. A series of groundbreaking works by Parisi [25, 26], followed by Guerra-Toninelli [27], Talagrand [28], and Panchenko [29–31] led to proof of the existence and a method for computing a deterministic limit of (6) in probability as $N \to \infty$. We denote this limit by $\eta_{\mathrm{p,OPT}}$ in either case, where the choice of Θ_N will be clear from the context. The value of this limit arises as a solution of a certain variational problem over the space of one-dimensional probability measures. The measure which provides the solution to this variational problem is called the Parisi measure which we denote by μ.

The algorithmic goal under consideration is the goal of constructing a solution $z \in \Theta_N$ which achieves near optimality, namely the value close to $\eta_{\mathrm{p,OPT}}$ when the tensor J is given as an input. Ideally, we want an algorithm $\mathcal{A}$ which for every constant $\epsilon > 0$ produces a solution $\hat{z} \triangleq \mathcal{A}(J)$ satisfying $\langle J, \hat{z}^{\otimes p}\rangle \geqslant (1-\epsilon)\eta_{\mathrm{OPT}}$ in polynomial (in N) time. This was achieved in a series of important recent developments [32–34], when the associated Parisi measure μ is strictly increasing. This monotonicity property is related to the OGP as we will discuss below.

3.2. OGP and its variants

The following result states the presence of the OGP for the p-spin models.

Theorem 1. *For every even $p \geqslant 4$, $\Theta_N = B_N$ or $\Theta_N = S_N$, there exists $\eta_{\mathrm{p,OGP}} < \eta_{\mathrm{p,OPT}}$, $0 < \nu_1 < \nu_2 < 1$ and $c > 0$ such that with probability at least $1 - \exp(-cN)$ for large enough N the following holds. For every $z_1, z_2 \in \Theta_N$ satisfying $\langle J, z_j^{\otimes p}\rangle \geqslant \eta_{\mathrm{p,OGP}}, j = 1, 2$*

$$\frac{1}{N}|\langle z_1, z_2\rangle| \notin (\nu_1, \nu_2).$$

J. Stat. Mech. (2022) 114015

Here $\langle x, y\rangle$ denotes the inner product $\sum_{1\leqslant i\leqslant N} x_i y_i$. Namely, modulo an exponentially in N unlikely event, the normalized angle (overlap) between any two solutions with value at least $\eta_{\mathrm{p,OGP}}$ cannot fall into the interval (ν_1, ν_2). The model exhibits an overlap gap.

The values $\eta_{\mathrm{p,OGP}}$ and ν_j (and in fact the optimal values $\eta_{\mathrm{p,OPT}}$ themselves) are in general different for Ising and spherical models and their precise values are of no algorithmic significance. While the result is only known to hold for even $p \geqslant 4$, it is expected to hold for all $p \geqslant 3$. It is conjectured not to hold when $p = 2$ [26] for the Ising case and the AMP algorithm achieving the near ground state value in this case is effective modulo this conjecture [33]. It does not hold when $p = 2$ for the spherical case for a trivial reason as in this case the problem corresponds to optimizing a quadratic form over sphere S_N. The proof of this theorem 1 for the Ising case can be found in [35], and is obtained by a detailed analysis of the variational problem associated with pairs of solutions z_1, z_2 within a certain proximity to optimality. The proof for the spherical case can be found in [36].

In order to use this result as an algorithmic barrier, we need to extend this theorem to the following *ensemble* variant of the OGP which we dub e-OGP. For this purpose it will be convenient to assume that the distribution of the entries of J is Gaussian. Consider an independent pair of tensors $J, \tilde{J} \in \mathbb{R}^{N^{\otimes p}}$ with Gaussian entries. Introduce the following natural interpolation between the two: $J(t) = \sqrt{1-t}J + \sqrt{t}\tilde{J},\ t \in [0, 1]$. The distribution of $J(t)$ is then identical to one of J and $\tilde{J}$ for every t.

Theorem 2. *For every even $p \geqslant 4$, $\Theta_N = B_N$ or $\Theta_N = S_N$, for the same choice of parameters $\eta_{\mathrm{p,OGP}}, \nu_1, \nu_2$ as in theorem 1 the following holds with probability at least $1 - \exp(-cN)$ for some c and large enough N. For every $t_1, t_2 \in [0, 1]$ and every $z_1, z_2 \in \Theta_N$ satisfying $\langle J(t_j), z_j^{\otimes p}\rangle \geqslant \eta_{\mathrm{p,OGP}}, j = 1, 2$ we have*

$$\frac{1}{N}|\langle z_1, z_2\rangle| \notin (\nu_1, \nu_2).$$

Furthermore, when $t_1 = 0, t_2 = 1$, it holds $\frac{1}{N}|\langle z_1, z_2\rangle| \in [0, \nu_1]$.

The probability event above is with respect to the joint randomness of J and $\tilde{J}$. Theorem 2 says that the OGP holds for pairs of solutions with values above $\eta_{\mathrm{p,OGP}}$ *across the entire* interpolated sequence of instances $J(t)$. Furthermore, at the extremes, that is for the pair of instances J and $\tilde{J}$, these solutions must have overlap at most ν_1. We note that the overlap value 1 is trivially achievable when $t_1 = t_2$ by taking two identical solutions $z_1 = z_2$ with value at least $\eta_{\mathrm{p,OGP}}$. The proof for the Ising case can be found in [18], and for the spherical case in [19], and it is a rather straightforward extension of theorem 1 by appealing to the chaos property exhibited by many glassy models [37, 38].

3.3. e-OGP as an algorithmic barrier to stable algorithms

We now discuss how the presence of the e-OGP presents an algorithmic barrier to a class of algorithms we loosely define as *stable* (noise-insensitive) algorithms. This part will be discussed rather informally, as each concrete instantiation of the arguments is model and algorithm dependent. We think of algorithms as mappings of the form $\mathcal{A}(J) \to \Theta_N$ which

J. Stat. Mech. (2022) 114015

map instances (tensors) J into a solution $z = \mathcal{A}(J)$ in the solution space Θ_N. In some cases the algorithms can take advantage of an additional randomization with functions now taking the form $\mathcal{A}(J, \omega)$, where ω is a sample corresponding to the randomization seed. For simplicity, we stick with non-randomized versions $\mathcal{A} : \mathbb{R}^{N^{\otimes p}} \to \Theta_N$. Informally, we say that the algorithm $\mathcal{A}$ is stable (noise-insensitive), if a small change in J results in a small change in the output. Namely, $\|\mathcal{A}(J_1) - \mathcal{A}(J_2)\|$ is likely to be small with respect to the natural metric on Θ_N when $\|J_1 - J_2\|_2$ is small. The choice of metric on Θ_N is driven by the space itself and can be Hamming distance when $\Theta_N = B_N$ or $\mathbb{L}_2$ norm when it is S_N. The 'likely' is in reference to the randomness of the tensor J. The following theorem stated informally shows why the presence of the e-OGP presents a barrier to stable algorithms.

Theorem 3 (Informal). *For every stable algorithm $\mathcal{A}$ and every $\epsilon > 0$, $\langle J, (\mathcal{A}(J))^{\otimes p} \rangle \leqslant \eta_{\mathrm{p,\,OGP}} + \epsilon$ w.h.p. as $N \to \infty$.*

Namely, this theorem states that stable algorithm cannot overcome the OGP barrier.

Proof sketch: we provide an outline of a simple proof of this theorem. The stability of the algorithm can sometimes be used to establish the concentration of its value around expectation, namely that $\langle J, (\mathcal{A}(J))^{\otimes p} \rangle \approx \mathbb{E}\langle J, (\mathcal{A}(J))^{\otimes p} \rangle$ as $N \to \infty$. This is not the case universally, but for simplicity let us assume this for now. Then it suffices to establish the claim $\mathbb{E}\langle J, (\mathcal{A}(J))^{\otimes p} \rangle \leqslant \eta_{\mathrm{p,\,OGP}} + \epsilon$. Suppose not. Then we have $\mathbb{E}\langle J, (\mathcal{A}(J))^{\otimes p} \rangle \geqslant \eta_{\mathrm{p,\,OGP}} + \epsilon$ implying $\mathbb{E}\langle J(t), \mathcal{A}(J(t)) \rangle \geqslant \eta_{\mathrm{p,\,OGP}} + \epsilon$ for every t in the interpolation path. We will obtain a contradiction.

By the second part of theorem 2 we then must have w.h.p. and in expectation

$$\frac{1}{N} |\langle \mathcal{A}(J(0)), \mathcal{A}(J(1)) \rangle| \leqslant \nu_1,$$

namely

$$\frac{1}{N} \|\mathcal{A}(J(0)) - \mathcal{A}(J(1))\|_2 \geqslant \sqrt{2 - 2\nu_1}.$$

Here we assume that we use $\mathbb{L}_2$ for Θ_N and the norm of every solution produced by the algorithm is $\sqrt{N}$ (which is the case when say $\Theta_N = B_N$). On the other hand trivially $\frac{1}{N} |\langle \mathcal{A}(J(0)), \mathcal{A}(J(0)) \rangle| = 1 > \nu_2$, implying

$$\frac{1}{N} \|\mathcal{A}(J(0)) - \mathcal{A}(J(1))\|_2 = 0 \leqslant \sqrt{2 - 2\nu_2}.$$

Stability of the algorithm $\mathcal{A}$ implies then the existence of time τ such that

$$\frac{1}{N} |\langle \mathcal{A}(J(0)), \mathcal{A}(J(\tau)) \rangle| \in (\nu_1, \nu_2),$$

which is a contradiction to the first part of theorem 2 (figure 1). □

The proof above is just an outline of the main ideas that have different specific implementations for specific problems. The earliest application of this idea was in [39], in a different context of finding large independent sets in sparse random graphs. The method was used to show that local algorithms, appropriately defined, are stable, where

J. Stat. Mech. (2022) 114015

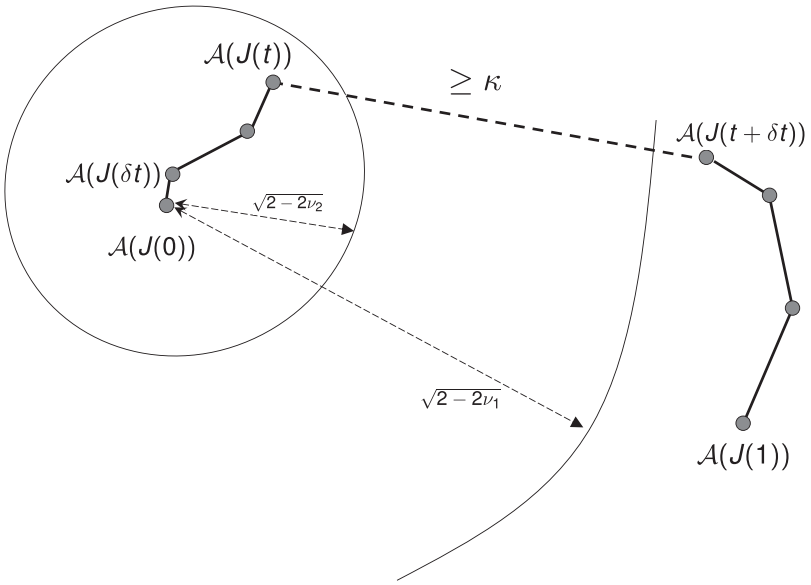

Figure 1. The smaller circle represents $\eta_{\text{p,OGP}}$-optimal solutions at distance $\leqslant\sqrt{2-2\nu_2}$ from $\mathcal{A}(J(0))$. The complement to the larger circle represents $\eta_{\text{p,OGP}}$-optimal solutions at distance $\geqslant\sqrt{2-2\nu_2}$ from $\mathcal{A}(J(0))$. As distance between the circle boundaries is $\sqrt{2-2\nu_1}-\sqrt{2-2\nu_2}\triangleq\kappa$, at some instance t the distance between 'successive' solutions $\mathcal{A}(J(t))$ and $\mathcal{A}(J(t+\delta t))$ has to be at least κ, contradicting stability.

J denotes random graph connectivities. In the context of spin glasses, it was shown in [18] that the AMP algorithm is stable and thus cannot overcome $\eta_{\text{p,OGP}}$ barrier. This was generalized in [19] where algorithms based on low-degree polynomials were shown to be stable. In the same paper Langevin dynamics was shown to be stable for spherical spin models when the running time is linear in N. Extending the limitation of the Langevin dynamics beyond linear bound is an interesting open problem. A natural conjecture is that the Langevin dynamics produces a value at most $\eta_{\text{p,OGP}}$ when run for $N^{O(1)}$ time.

By leveraging the multi-e-OGP method, which involves studying overlap patterns of more than two solutions, the barrier $\eta_{\text{p,OGP}}$ and its analogues for other models can be pushed to the value achievable by the state of the art algorithms. These algorithms are AMP in the p-spin Ising case [34] and the spherical p-spin model case [32], simple greedy algorithms for the case of random K-SAT problem and the case of independent sets in sparse random graphs. The implementation of the multi-e-OGP for spin glass models was done by Huang and Sellke [40], who have implemented a very ingenious version of the multi-OGP, called branching-OGP. This version was motivated by the ultrametric structure of the solution space of p-spin models, widely conjectured to hold. The implementation for the random K-SAT was done by Bresler and Huang [41], and for independent sets in sparse random graphs by Wein in [20].

Arguably the strongest implication of the OGP as an algorithmic barrier is its usage for establishing the state of the art lower bounds on depth of polynomial size Boolean

J. Stat. Mech. (2022) 114015

circuits. There is a long history in the theoretical computer science literature on establishing such lower bounds for various problems. In the context of constraint satisfaction problems, the prior state of the art result was achieved by Rossman [42, 43] (see also extensions in [44, 45]), who established a depth lower bound $\Theta(\log n/(\kappa_n \log\log n))$ for poly-size circuits deciding the presence of an independent set of size k_n in graphs with n nodes. When the depth of the circuit is bounded by an n-independent constant, he showed that the size of the circuit has to be at least $n^{\Omega(\log n)}$. This was done in the regime of random graphs where the typical value of k_n grows at most logarithmically in n. Using the OGP method this bound was improved to $\Theta(\log n/\log\log n)$, though for the search as opposed to the decision problem [21]. Similarly, when the depth of the circuit is at most a constant, a stretched exponential lower bound $\exp(n^{\Omega(1)})$ on the size was established as well. It is in the context of this problem where the concentration around expectation adopted in the proof sketch does not hold, and furthermore, the stability property does not hold w.h.p. Instead the idea was to establish that circuits with small depth have stability property with *at least* sub-exponentially small probability. On the other hand, the stability can occur only for the event which is complementary to the OGP, and this complement event holds with exponentially small probability, thus leading to a contradiction.

A similar application of the OGP based method shows that poly-size circuits producing solutions larger than $\eta_{\mathrm{p,OGP}}$ in p-spin models also have depth at least $\Theta(\log n/\log\log n)$. Pushing this result towards the value algorithmically achievable by the AMP, say using the Huang and Sellke [40] is not immediate due to the overlap Lipschitz concentration assumption required in [40]. This extension is an interesting open problem.

Broadly speaking a big outstanding challenge is the applicability of OGP or similar methods for models with a planted signal, which we discuss in the following sections. While a version of OGP takes place in many such models, its algorithmic implication is far narrower than in the settings discussed above, such as p-spin models and random constraint satisfaction problems. This presents an interesting and rather non-trivial challenge for future.

3.4. Connections with replica symmetry, symmetry breaking and the clustering (shattering) property

We discuss these connections rather informally now, leaving the technical aspects to other sources which we reference here.

The OGP arose in connection with studying the replica symmetry, RSB and related properties of spin glasses and their variants. Specifically, it arose as a method of proving that the set of satisfying solutions of a random constraint satisfaction problem is clustered (sometimes called shattered), meaning that it can be partitioned into 'connected' components with order $\Theta(N)$ distance between them. How can one establish the existence of such a clustering picture? If the model exhibits the OGP say with parameters $\nu_1 < \nu_2$, then clustering follows immediately, provided that solutions at distances $\sqrt{2-2\nu_1}$ or larger exist, as in this case one defines clusters as the set of solutions which can be reached from each other by paths in the underlying Hamming cube. The fact that

J. Stat. Mech. (2022) 114015

distances between $\sqrt{2-2\nu_2}$ and $\sqrt{2-2\nu_1}$ do not exist between the pairs of solutions imply that at least two (but in fact many) clusters exist.

There are several caveats associated with this connection between the OGP and the clustering property. First this connection is one directional, in the sense that the presence of clustering does not necessarily imply the OGP, for a very simple reason: the diameter of the cluster can in principle be larger than the distances between the clusters. In this case, while the clustering property takes place, the set of all normalized pairwise distances could potentially span the entire interval [0, 1] without any gaps. Therefore the path towards establishing algorithmic lower bounds is not entirely clear.

Second, as it turns out in some models and in some regimes, the clustering picture has been established for the 'majority' of the solution space, and not for the entire solution space. We will call it the weak clustering property, to contrast with the strong clustering property, which refers to a clustering property without exceptions. For example, for the random K-SAT problem the onset of the clustering property is known to take place close to the threshold $(2^K/K)\log K$ for the clauses to variables densities, when K is large, but only in the weak clustering sense discussed above: most but not necessarily all of the solutions can be split into clusters [46].

As it turns out, these exceptions are not just a minor nuisance, and can have profound algorithmic implications. The so-called symmetric perceptron model is a good demonstration of this [47–51]. For this model, the weak clustering property is known to take place at all constraints to variables densities, yet polynomial time algorithms exist at some strictly positive density values [49]. The multi-OGP analysis conducted in [50] reveals that the gaps in the overlaps occur at densities *higher* than the known algorithmic thresholds and thus the thresholds for the weak clustering property and the OGP do not coincide and, furthermore, the weak clustering property is apparently not a signature of an algorithmic hardness. Whether the strong clustering property can be used as a 'direct' evidence of algorithmic hardness remains to be seen. For the further discussion of the connection between the OGP, the weak and strong clustering properties, and the algorithmic ramifications, we refer the reader to [16].

Next we discuss the connection between the OGP, replica symmetry, symmetry breaking and the Parisi measure μ. The Parisi measure μ arises in studying the Gibbs measure associated with Hamiltonian H. (Very) roughly speaking, it describes an overlap structure of two nearly optimal solutions σ and τ chosen uniformly at random. This can be formalized by introducing a small positive temperature parameter in the Gibbs distribution, but we skip this formalism. The idea is that $(1/N)|\langle\sigma,\tau\rangle|$ has the cumulative distribution function (CDF) described by μ in the large N limit. The support of μ is naturally some subset of [0, 1]. The source of randomness is dual here, one arising from the randomness of the Hamiltonians, and one arising from the sampling procedure. Whether μ is indeed the limit the CDF of the overlaps in the limit remains a conjecture, which has been confirmed only for the spherical case. Loosely speaking the model is defined to be in the replica symmetric regime (RS) if μ is just a δ mass at zero. Namely, the overlap $(1/N)\langle\sigma,\tau\rangle$ is approximately zero with high probability, implying that typical pairs of solutions are nearly orthogonal to each other.

Replica symmetry breaking (RSB) then refers to μ being distinct from this singleton structure. Now if the model exhibits OGP, then a part of μ is flat: the CDF of

J. Stat. Mech. (2022) 114015

the overlaps is constant on (ν_1, ν_2). Namely, the CDF *is not* strictly increasing. The absence of this flat part of μ is exactly what was used in constructions of near optimal solutions in [32–34], (and the presence of the OGP is an algorithmic obstruction as we have discussed). So presumably, we could have used the flatness of the Parisi measure as a 'certificate' of hardness. However, there are challenges associated with this alternative. First, as we have discussed, whether μ indeed describes the distribution of overlaps remains an open question, whereas the presence of the OGP has been confirmed. More importantly though, even modulo the μ being the accurate descriptor of the overlaps, the connection between OGP and the flatness of μ is one-directional. The flatness of μ in some intervals (ν_1, ν_2) means only that the density of the overlaps falling into this interval is asymptotically zero after taking N to infinity. It does not imply the absence of such overlaps. This is similar to the distinction between the weak and strong clustering property: most of the overlaps are outside of the flat parts, but exceptions might exist. The presence of such exceptions is bad news for the efforts of establishing algorithmic lower bounds. Not only the argument for proving the algorithmic lower bounds appears to break down, but also the presence of exceptions, namely a small number of overlaps falling into this interval, might be potentially a game changer, as we saw in the case of the symmetric perceptron model.

4. Statistical and computational trade-offs in inference and learning

In this section we move from optimization problems to statistical inference, in other words from the non-planted problems to the planted ones. We recall our working examples defined in section 2, that cover a large range of settings and applications, the spiked tensor model and the GLM.

In order to describe the conjectured results on the algorithmic hardness of the planted problems we will first discuss the Bayes-optimal inference of the planted configuration from observations. We will then show how to analyze the performance of the Bayes-optimal inference in the large size limit $N \to \infty$ and under the stated randomness of the generative model. We will then show that phase transitions in the capability of the Bayes-optimal estimator to reconstruct the signal have an intriguing algorithmic role as a suitable type of message passing algorithms are able to reach optimal performance for all parameters except in the metastable region of first order phase transitions. This metastable region is then conjectured to be algorithmically hard—the hard phase. Section 5 will then present the currently strongest known method for showing evidence of such hardness in some cases.

4.1. The minimum mean-squared error

In both the spiked tensor model and the GLM as defined in section 2 the optimal inference of the planted signal u can be achieved by computing the marginals of the

J. Stat. Mech. (2022) 114015

posterior probability distribution

$$P(z|Y) = \frac{1}{\mathcal{Z}} P(z) P(Y|z). \tag{10}$$

Concretely, when aiming to find an estimator $\hat{z}$ that would minimize the mean-squared error to the signal u

$$\mathrm{MSE}(\hat{z}) = \frac{1}{N} \sum_{i=1}^{N} (u_i - \hat{z}_i)^2 \tag{11}$$

we conclude that from all the possible estimators we should take $\hat{z}$ to be the marginal of the posterior

$$\hat{z}_i = \mathbb{E}_{P(z|Y)}(z_i). \tag{12}$$

We will call the MSE achieved by this estimator the minimum-MSE, abbreviated MMSE. In the large size limit $N \to \infty$ computing marginals over $P(z|\, Y)$ with $z \in \mathbb{R}^N$ is in general exponentially costly in N, and thus potentially computationally hard even in the specific probabilistic generative models from section 2.

However, for the spiked tensor model as well as for the GLM tools from the theory of spin glasses come to the rescue and allow us to analyze the value of the MMSE in the larger size limit as well as design message passing algorithms with properties closely related to the approach to obtain the MMSE. Let us start by describing the form in which we obtain the asymptotic value of the MMSE. Replica theory allows us to derive an explicit formula for a function $\Phi_{\mathrm{RS}}(m)$, $m \in \mathbb{R}$, called the replica symmetric free entropy such that

$$\lim_{N\to\infty} \mathbb{E}_{Y,u,J} \log \mathcal{Z} = \max_m \Phi_{\mathrm{RS}}(m). \tag{13}$$

We note that in physics it is more common to define the free energy which is just the negative of the free entropy. The average over Y, u, J applies to the GLM. In the spiked matrix model the Y can be dropped as in the definition we gave it explicitly depends on u and J. The function $\Phi_{\mathrm{RS}}(m)$ explicitly depends on the parameters of the prior, the likelihood and the ratio $\alpha = N/P$, but in our notation we omit this dependence. We then call

$$m^* = \mathrm{argmax}\, \Phi_{\mathrm{RS}}(m) \tag{14}$$

and state a generic result for the MMSE that is given by the global maximizer of the replica symmetric free entropy

$$\lim_{N\to\infty} \mathrm{MMSE} = \rho - m^* \tag{15}$$

where the constant $\rho = \mathbb{E}(u_i^2)$ is simply the second moment of the signal components.

The derivations of these result and the explicit formulas for $\Phi_{\mathrm{RS}}(m)$ were given in the spin glass literature for many special cases and mostly without a rigorous justification. In the general form considered in this paper and including rigorous proofs they were given for the spiked tensor model in [52], and for the GLM in [53]. For the purpose of

J. Stat. Mech. (2022) 114015

this paper we will stay on the abstract level expressed above because on this level the discussion applies to a broad range of settings and we do not want to obfuscate it with setting-dependent details.

An important comment needs to be made here about the very generic validity of the replica symmetric result for the free entropy in the Bayes-optimal setting, i.e. when the prior and likelihood match the corresponding distributions in the model that generated the data. By the very nature of the Bayes' formula the signal u has properties interchangeable with properties of a random sample from the posterior $P(z|Y)$. This is true even at finite size N and even for models where J is not random and where the likelihood and the prior are not separable. A consequence of the interchangeability is that under the averages over the posterior measure and the signal u we can replace the signal u for a random sample from the posterior and vice versa. This is called the Nishimori condition in the statistical physics literature [54, 55]. A direct consequence of the Nishimori condition is that the magnetization (correlation between the signal and a random sample) and the overlap (correlation of two random samples) have to be equal, which in return means that the overlap distribution needs to be concentrated on a delta function and thus no RSB is possible in the Bayes-optimal setting. The Nishimori conditions also play a key role in the proof techniques used to establish the above results rigorously in [52, 53].

It also important to note that what we discuss in this section is limited to the large size limit $N \to \infty$ with parameters scaling in such a way with N for the MMSE to go from ρ to 0 as the signal-to-noise ratio α increases from 0 to large $O(1)$ values. This imposes scaling on the λ_N for the spiked tensor model that is $O(N^{(1-p)/2})$. This will be in particular important for our claims about the optimality of the AMP algorithm that will be restricted to this regime and will not necessarily apply to performance of AMP for much larger signal to noise ratios.

4.2. AMP and its state evolution

In the previous section we analyzed the MMSE as it would be achieved by the exact computation of the posterior average. This is, however, in general computationally demanding and thus a next natural question is whether we can reach this MMSE computationally efficiently. Message passing algorithms provide an algorithmic counter-part of the replica method. In particular, the approximate message passing algorithm (AMP) that is an extension of the TAP equations [56] to the general setting of the spiked tensor model and the GLM is of interest to us in this paper. AMP is an iterative algorithm that aims to compute the Bayes-optimal estimator $\hat{z}$. Schematically the update of AMP at time step t for the AMP's estimate $z^t_{\rm AMP} \in \mathbb{R}^N$ can be written for both the considered models as

$$z^{t+1}_{\rm AMP} = \mathcal{F}(z^t_{\rm AMP}) \tag{16}$$

for an update function $\mathcal{F}(.)$ that depends on Y, parameters of the prior and the likelihood, and for the GLM also on J.

The key property that makes AMP so theoretically attractive is that in the large size limit the accuracy of the AMP estimator can be tracked via low-dimensional set

J. Stat. Mech. (2022) 114015

of equations called state evolution. To state this we introduce the correlation between AMP estimate and the signal at iteration t

$$m_N^t = \frac{1}{N}\sum_{i=1}^{N} u_i \, (z_{\rm AMP}^t)_i \tag{17}$$

The state evolution implies that this quantity in the large size limit $m^t = \lim_{N\to\infty} m_N^t$ behaves as

$$m^{t+1} = f_{\rm SE}\ (m^t), \tag{18}$$

for a function $f_{\rm SE}$ that depends on the parameters of the models, but not any longer of any high-dimensional quantity. The state evolution of AMP is a crucial contribution that came from mathematical developments of the theory [57, 58] and was not known in its current form in the statistical physics literature before that. The proofs of state evolution have been extended to a broader setting [59–61].

What makes the state evolution particularly appealing in the statistical physics context is its connection to the computation of the MMSE. The fixed points of the expression (18) can be expressed at the stationary points of the replica symmetric free entropy

$$m = f_{\rm SE}\ (m) \quad \Leftrightarrow \quad \frac{\partial \Phi_{\rm RS}(m)}{\partial m} = 0 \tag{19}$$

where $\Phi_{\rm RS}(m)$ is indeed the same free entropy as in equation (13).

Since the signal u is unknown the corresponding initialization is $m^{t=0} = 0$ (this is for prior distribution with zero mean) and thus the performance of AMP is given by the stationary point of the free entropy that is reached by iterating (18) initialized at $m^{t=0} = 0$. The performance of AMP at convergence thus corresponds to the local maximum $m_{\rm AMP}$ of the free entropy $\Phi_{\rm RS}(m)$ that has the largest error. The corresponding MSE is then

$$\text{MSE}_{\rm AMP} = \rho - m_{\rm AMP}. \tag{20}$$

4.3. The phase diagrams and the hard phase

We have seen in the previous two subsections that the values of the MMSE as well as the MSE obtained by the AMP algorithm can both be deduced from the extremizers of the free entropy function $\Phi_{\rm RS}(m)$.

While the MMSE is given by the global maximizer of $\Phi_{\rm RS}(m)$, the MSE reached by the AMP algorithm is given by the maximizer having the smallest m. In the following we will consider all the extremizers of $\Phi_{\rm RS}(m)$ as this will allow us to understand the resulting overall picture. We will discuss how the extremizers depend on some kind of signal to noise ratio α. This signal to noise ratio can be simply the value of $\alpha = \lambda$ in the spiked matrix model, or the sample complexity ratio $\alpha = P/N$ in the GLM.

Depending on the other parameters of the model we can observe a number of scenarios, we will discuss several of them below and refer to examples where they appear. In

J. Stat. Mech. (2022) 114015

the following sketches all the colored curves are extremizers of $\Phi_{\rm RS}(m)$. Those in blue are the global maximizers of the free entropy corresponding to the MMSE. No algorithmic procedure can achieve an error lower than the MMSE. When the AMP algorithm does not achieve the MMSE, the MSE it reaches at its fixed point corresponds to a maximizer of the free entropy of a higher error $\text{MSE}_{\rm AMP}$ depicted in green. In red we depict the other extremizers of the free entropy, in dashed red the minimizers, and in full red the other maximizers.

The region of error between the green and the blue curve are values of the MSE that are information-theoretically reachable, but the AMP algorithm does not reach them. We call this region the **hard phase**, and its boundaries on the signal-to-noise ratio axes: $\alpha_{\rm IT}$ for the information theoretic threshold where the values of the two maximizers of $\Phi_{\rm RS}(m)$ switch order, and $\alpha_{\rm alg}$ above which AMP reaches the MMSE. The hard phase exists in between these two thresholds, $\alpha_{\rm IT} < \alpha < \alpha_{\rm alg}$. A third threshold $\alpha_{\rm s}$ marks the spinodal point at which the lower-error maximizer of the free entropy ceases to exist, this point does not have significant algorithmic consequences for finding the signal. In other cases there may be no phase transition at all or a second order (continuous) phase transition marked by $\alpha_{\rm c}$.

The physical interpretation of the cases where the hard phase exists is the one of first order phase transition in a high-dimensional (mean-field) system. The $\alpha_{\rm IT}$ corresponds to the thermodynamic phase transition while $\alpha_{\rm s}$ and $\alpha_{\rm alg}$ are the spinodals, i.e. the boundaries of the metastable regions. In the hard phase the thermodynamic equilibrium corresponds to the higher free entropy branch depicted in blue, and the green fixed point corresponds to the metastable state. In the region $\alpha_{\rm s} < \alpha < \alpha_{\rm IT}$ the AMP algorithm finds the thermodynamic equilibrium, but this state is split into exponentially many separated states, each corresponding to the metastable branch (full red). In the language of replica-symmetry breaking this phase corresponds to the dynamical-1RSB phase (d-1RSB). In the d-1RSB phase the AMP algorithm reached optimal performance in terms of finding the signal, however, sampling the posterior measure in the d-1RSB region is conjectured computationally hard.

In figure 2 we depict one possible structure of extremizers of the free entropy $\Phi_{\rm RS}(m)$ for models where neither $m = 0$ nor $m = \rho$ are fixed points for $\alpha > 0$. On the left-hand side of figure 2 we depict a case without a phase transition. This situation arises for instance in generalizes linear models with Gaussian prior and a sign activation function, corresponding to the spherical teacher-student perceptron, see e.g. center of figure 2 in [53] for a concrete example. On the right-hand side of figure 2 we depict a case with a first order phase transitions. Such as situation arises for instance spiked matrix model where the prior is sparse with non-zero mean, see e.g. rhs of figure 4 in [7] for a concrete example.

In figure 3 we depict another possible structure of extremizers of the free entropy $\Phi_{\rm RS}(m)$ for models where $m = 0$ is a fixed point. On the left of figure 3 there is a situation with a second order phase transition as is the case for instance in the symmetric stochastic block model with two groups, see e.g. figure 1 in [62] for a specific example. On the right of figure 3 there is a situation with a first order phase transition as is the

J. Stat. Mech. (2022) 114015

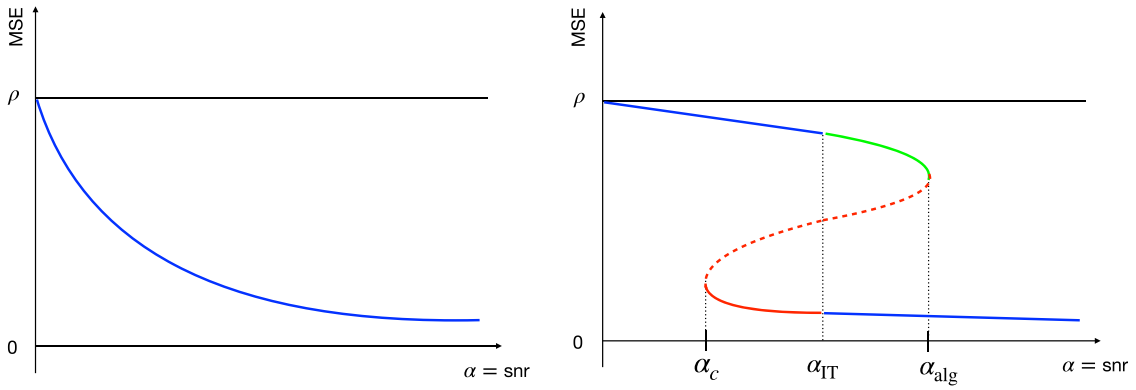

Figure 2. Extremizers of the replica symmetric free entropy when neither $m = 0$ nor $m = \rho$ are stationary points. Colors explained in the text. (Left) A case without a phase transition. (Right) A case with a first order phase transition.

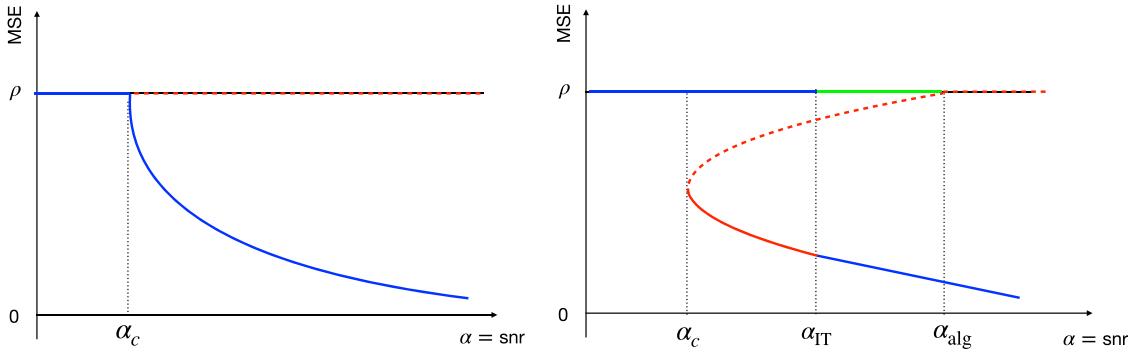

Figure 3. Extremizers of the replica symmetric free entropy when $m = 0$ is a stationary point for all α. Colors explained in the text. (Left) A case with a (continuous) 2nd order phase transition. (Right) A case with a (discontinuous) first order phase transition.

case for instance in the symmetric stochastic block model with more than 4 groups, see e.g. figure 3 in [62] for a specific example. In this case the threshold at which the fixed point at $m = 0$ ceases to be a maximum and start to be a minimum is the well-known Kesten–Stigum threshold [63], marked $\alpha_{\rm c}$ on the lhs of the figure, and $\alpha_{\rm alg}$ on the rhs of the figure. When $m = 0$ and MMSE $= \rho$ is the thermodynamic equilibrium no correlation with the signal can be obtained and the phase $\alpha < \alpha_{\rm IT}$ is in this case referred to as the undetectable region. In this phase the planted model is contiguous to the non-planted model in the sense that all high-probability properties in the planted model are the same in the non-planted one other [64]. This is the setting that is most often explored in the sum-of-squares approach of section 5.

In figure 4 we depict yet another possible structure of extremizers of the free entropy $\Phi_{\rm RS}(m)$ for models where $m = \rho$ is a fixed point and thus where exact recovery of the signal with MMSE $= 0$ is possible for sufficiently large signal-to-noise ratios. On the right of figure 4 we depict a case with a first order phase transition. Such a situation

J. Stat. Mech. (2022) 114015

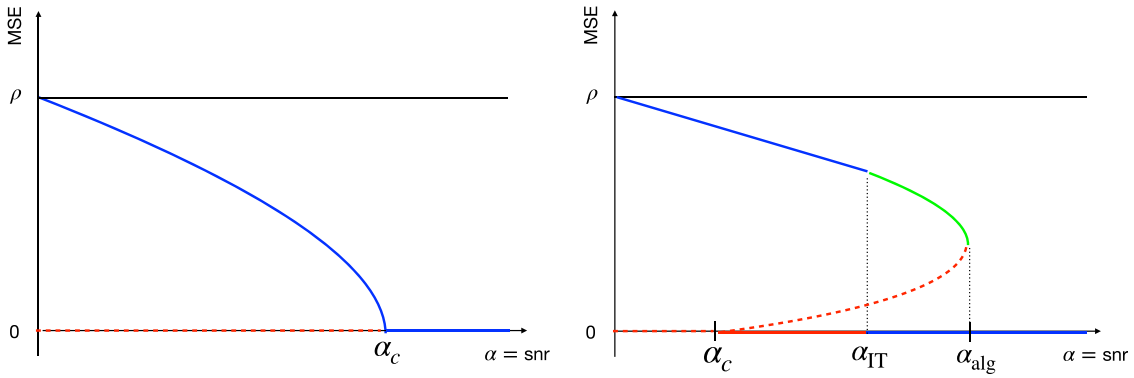

Figure 4. Extremizers of the replica symmetric free entropy when $m = \rho$ is a stationary point for all α. Colors explained in the text. (Left) A case with a 2nd order phase transition. (Right) A case with a first order phase transition.

arises e.g. in the GLM with binary prior and sign activations, corresponding to the teacher-student binary perceptron, see left-hand side of figure 2 in [53]. On the left of figure 4 we depict a case with a second order phase transition, this arises e.g. in the GLM with Laplace prior and no noise, corresponding to the minimization of the ℓ_1 regularization, see e.g. figure 3 in [65].

The examples we depict in this section do not exhaust all the possible scenarios one encounters in computational problems. Some of those we did not cover include the planted locked constraint satisfaction problems where both $m = 0$ and $m = \rho$ fixed points exist and an all-to-nothing first order phase transition happens between these two fixed points [66]. Both $m = 0$ and $m = \rho$ fixed point also exist for instance in the GLM with Gaussian prior and absolute value activation corresponding to the phase retrieval problem. In that case there is a second order phase transition from the undetectable phase to a detectable one and later on a first order phase transition to exact recovery, see e.g. left-hand side of figure 5 in [53].

Another interesting and very generic case is depicted e.g. in figure 6 of [7] for the spiked matrix model with a symmetric Rademacher–Bernoulli prior. In this case the undetectable phase ($m = 0$ fixed point) is followed by a phase where a correlation with the signal is detectable but small, and where AMP reaches a small but suboptimal correlation to the signal. The position of the first order phase transition can be either before or after the detectability threshold (as in the left or right of the lower part of figure 6 in [7]). While this may seem a rare scenario, results in [67] (see figure 2) actually indicate that it is likely very generic and that often the size of the region where detection is possible but sub-optimal is very thin.

Yet another interesting example of a phase transition in a planted problem is the planted matching problem where the phase transition is infinite order, i.e. all the derivatives of the order parameter m exist at the transition from partial recovery phase to exact recovery phase [68].

4.4. Is the hard phase really hard?

A fundamental question motivating the discussion of this paper is for what class of algorithms is the hard phase computationally inaccessible?

An important evidence towards the hardness is summarized in [69] where it is shown that a very broad range of algorithms related structurally to the AMP cannot improve over the AMP that uses the Bayes-optimal parameters. Efforts to prove lower bounds are considerable, as discussed in section 5. A number of authors put forward a conjecture that in settings where the large-size limit and randomness is taken in such a way that AMP and the Bayes-optimal solution are related in the way we describe above, then AMP is optimal among a large class of algorithms. But could this possibly be all polynomial algorithms?

It is important to note that there are problems with the phenomenology leading to the hard phase yet for which polynomial algorithms to find the signal exist never-theless. One of them is that planted XOR-SAT problem [66, 70] that is mathematically a linear problem in the Boolean algebra and can thus always be solved using Gaussian elimination. Gaussian elimination, however, runs with time larger than linear in the size of the system and is not robust to noise where we plant a solution that violates a small fraction of clauses. A more surprising and recent example is given by the noiseless phase retrieval problem for Gaussian matrix J where the so-called LLL algorithm also works in polynomial time down to the information-theoretic threshold [71, 72]. The phase retrieval problem is NP-hard, unlike the planted XOR-SAT. Again the LLL is based on linear algebra and thus in some sense related to Gaussian eliminations, it is not robust to noise, or runs in time that is polynomial with an exponent considerably larger than one.

The existence of these examples makes it clear that in some cases other algorithms can perform better than AMP with the Bayes-optimal parameters in the high-dimensional limit. It is thus more reasonable to conjecture that the AMP algorithm may be optimal among those polynomial ones that are required to be robust to noise? Or among those that run with resources linear with the input size of the problem (i.e. quadratic in N)?

We also want to note here another case that is often cited as an example where other algorithms beat AMP. This is the spiked tensor model for $p \geqslant 3$. However, in this case the algorithmic threshold happens at $\lambda_N \sim N^{-p/4}$ while the information theoretic one at $\lambda_N \sim N^{(1-p)/2}$. We do not expect AMP to be in general optimal for other scalings than the information-theoretic one, we thus do not consider this as a counter-example to the conjecture of optimality of AMP. Our conjectures about optimality of AMP restrict to the information-theoretic scaling.

4.5. The hard phase is glassy, causing hurdles to gradient-based algorithms

From the physics point of view the conjecture of optimality of AMP is very intriguing. It needs to be stressed that the state evolution that rigorously tracks the performance of the AMP algorithm corresponds to the replica symmetric branch of the free entropy while RSB is needed to describe the physical properties of the metastable state [73].

Physically, and following the success of survey propagation [74] in solving the random K-SAT problem, one may have hoped that including the glassiness in the form of the algorithm, as done in [75], would improve the performance. This is, however, not happening and is rigorously precluded by the proof of [76]. So in a sense while AMP follows the non-physical solution for the metastable state, this solution has fundamental meaning in terms of being the best solution achievable by a computationally tractable algorithm.

It is interesting to note that early work in statistical physics indeed dismissed the replica symmetric spinodal as non-physical, see [77], and presumed that algorithms will be stopped by the glassiness of the metastable phase. This is a nice example where the later state-evolution proof takes over the early physical intuition about what is the relevant algorithmic threshold.

At the same time, the physics intuition of the glassiness stopping the dynamics for signal-to-noise ratios larger than where the replica symmetric appears was not wrong. It simply does not apply to the AMP algorithm that does not correspond to a physical dynamics as it does not perform a walk in the space of possible signals but rather iterates marginals over the signal components. If we consider now instead physical dynamics such and Monte-Carlo Markov chains (MCMC) or algorithms updating the signal estimate based on possibly noisy gradient descent the early intuition of [77] turned out to be completely correct in the sense that these algorithms actually perform considerably worse than AMP when the hard phase is present. Interestingly this was not expected in some works, e.g. [62] conjectured that MCMC performs as well as message passing in the stochastic block model, which turns out to be wrong [78]. Very clear-cut examples of gradient-based Langevin algorithms performing worse than AMP are given for the mixed spiked matrix-tensor model in [79] and for the phase retrieval in [80].

The phase retrieval example is particularly relevant due to its interpretation as a neural network and given that gradient descent is the working horse of the current machine learning revolution. One may ask whether some key parts of the current machine learning tool-box such as over-parameterization and stochasticity in gradient descent are not a consequence of mitigation of the hurdles that gradient descent encounters due to glassiness of the landscape. Some resent works on the phase retrieval problem do point in that direction [81, 82].

5. Polynomial proofs: the sum-of-squares hierarchy

In the absence of a proof that $\mathrm{P} \neq \mathrm{NP}$, we have no hope of proving that problems in a certain parameter range truly require exponential time. There may in fact be no hard regimes. But we can try to gather the efficient algorithms we know of into large families—each characterized by a particular strategy or kind of reasoning, or which can only 'understand' certain things about their input—and show that no algorithm in these families can succeed. In the previous section, we discussed how the OGP can be used to defeat algorithms that are stable to noise or small perturbations in their input.

J. Stat. Mech. (2022) 114015

Here we discuss classes of algorithms that have an algebraic flavor. We will focus on the sum-of-squares hierarchy, and briefly discuss its cousin the low-degree likelihood ratio. Many of the best algorithms we know of are captured by these classes, including powerful generalizations of spectral algorithms and classic approximation algorithms. Thus if we can show that they fail to solve certain problems, or more precisely that they require polynomial 'proofs' or 'likelihood ratios' of high degree, this constitutes additional evidence that these problems are hard.

There are types of reasoning that these systems have difficulty with, as our first example will illustrate. This leaves open the possibility that some very different algorithm could efficiently solve problems in what we thought was a hard regime. However, these other types of reasoning seem fine-tuned and fragile, and only work in noise-free settings. For a wide variety of noisy problems, algorithms associated with sum-of-squares are conjectured to be optimal [83].

5.1. Proofs and refutations

At its heart, the sum-of-squares (SoS) hierarchy is a way of constructing *refutations* of constraint satisfaction or optimization problems: proofs that a solution does not exist, or that no solution achieves a certain value of the objective function. It comes with a dual problem, of evading refutation by finding a *pseudoexpectation*: a fictional distribution of solutions that looks reasonable as long as we only ask about polynomials up to a certain degree. If a pseudoexpectation can be constructed that 'fools' polynomials up to degree d, then any refutation must have degree greater than d.

Let us look at an example. Consider three variables $x, y, z \in \{\pm 1\}$. Is it possible for them to sum to zero? This problem may seem trivial, but bear with us. Algebraically, we are asking whether the following system of polynomials has a solution,

$$\begin{aligned} x^2 - 1 &= 0 \\ y^2 - 1 &= 0 \\ z^2 - 1 &= 0 \\ x + y + z &= 0. \end{aligned} \tag{21}$$

Here is a proof, that the motivated reader can verify, that no solution exists:

$$\begin{aligned} &\frac{1}{8}\Big[\left(x^2 + 3(y^2+z^2) + 4(xy+xz+3yz) - 3\right)(x^2-1) + \left(y^2 + 3(x^2+z^2)\right. \\ &\quad \left. + 4(yz+xy+3xz) - 3\right)(y^2-1) + \left(z^2 + 3(x^2+y^2) + 4(xz+yz+3xy) - 3\right)(z^2-1)\Big] \\ &\quad + (x+y+z)^2 = \frac{1}{8}\left((x+y+z)^2 - 1\right)^2 + 1. \end{aligned} \tag{22}$$

If the constraints (21) hold, then the left-hand side of (22) is identically zero. On the other hand, the right-hand side is the square of a polynomial plus 1, giving the contradiction $0 \geqslant 1$. We will reveal below how we constructed this proof.

More generally, suppose we have a set of polynomials $f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})$ over n variables $x_1, \ldots, x_n$. We wish to prove that there is no $\mathbf{x} \in \mathbb{R}^n$ such that $f_i(\mathbf{x}) = 0$ for all i. A sum-of-squares proof consists of additional polynomials $g_1, \ldots, g_k$ and $h_1, \ldots, h_t$ such that

$$\sum_{i=1}^{k} g_i(\mathbf{x}) f_i(\mathbf{x}) = \sum_{j=1}^{t} h_j(\mathbf{x})^2 \; + \; 1, \tag{23}$$

where 1 on the right-hand side can be replaced by any positive constant. In other words, we find a linear combination of the f_i that is strictly positive everywhere, so they can never be zero simultaneously. Any unsatisfiable system of polynomial equations $\{f_i(\mathbf{x}) = 0\}$ has a refutation of this form [84, 85]. A logician would say that the SoS proof system is *complete*.

Now, we say a SoS proof is of degree d if the polynomials $g_i f_i$ and h_j^2 on the left and right sides of (23) have maximum degree d. Thus our example (22) is a proof of degree $d = 4$. (By convention d is always even: the h_j have degree at most $d/2 = 2$.) As we increase d, we obtain a hierarchy of increasingly powerful proof systems.

In some cases the lowest possible degree of an SoS proof is much larger than the degree of the original constraints f_i, since we may need high-degree coefficients g_i to create the right cancellations so that the sum can be written as a sum of squares. As we will see below, if the necessary degree grows with the size of the problem, we can interpret this as evidence that the problem is computationally hard.

5.2. From proofs to algorithms: semidefinite programming

Of course, the existence of an SoS proof does not necessarily make it easy to find. Algorithmically, how would we search for these polynomials? If we choose some ordering for the monomials up to some degree, writing a symbolic vector $\mathbf{m} = (1, x, y, z, x^2, xy, xz, y^2, \ldots)$, then we can represent a polynomial q as a vector $\mathbf{q}$ of its coefficients and write $q(\mathbf{x})$ as an inner product $\langle \mathbf{q} \,|\, \mathbf{m} \rangle$. Multiplying two polynomials is a bilinear operation, and the sum on the right-hand side of (23) can be written

$$\sum_{j=1}^{t} h_j(\mathbf{x})^2 = \sum_{j=1}^{t} \langle \mathbf{m} \,|\, \mathbf{h}_j \rangle \langle \mathbf{h}_j \,|\, \mathbf{m} \rangle = \langle \mathbf{m} | \mathbf{H} | \mathbf{m} \rangle \qquad \text{where} \quad \mathbf{H} = \sum_{j=1}^{t} |\mathbf{h}_j\rangle\langle\mathbf{h}_j|. \tag{24}$$

This bilinear form $\mathbf{H}$ is positive semidefinite, which we denote $\mathbf{H} \succeq 0$.

With this abstraction, the problem of finding SoS proofs asks for a positive semidefinite matrix that matches the left-hand side of (23). To nail this down, for a polynomial q let q_u denote the coefficient of each monomial u. Then summing over all the cross-terms in the product of two polynomials p, q gives

$$(pq)_u = \sum_{v,w \,:\, vw = u} p_v q_w. \tag{25}$$

J. Stat. Mech. (2022) 114015

Since for any two monomials s, t the entry $H_{s,t} = \langle s|\mathbf{H}|t\rangle$ must equal the coefficient of $u = st$ on the left-hand side of (23), for any s, t such that $st \neq 1$ we have

$$\sum_i \sum_{v,w\,:\,vw=st} (g_i)_s (f_i)_t = H_{s,t}, \tag{26}$$

and for $s = t = 1$ we have

$$\sum_i (g_i)_1 (f_i)_1 = 1 + H_{1,1}. \tag{27}$$

For a given set $\{f_i\}$, these constraints are linear in the coefficients of the $\{g_i\}$. Adding the semidefiniteness constraint $\mathbf{H} \succeq 0$ to this linear system of equations makes this a case of *semidefinite programming* or SDP [86–89].

SDP can be solved up to arbitrarily small error in polynomial time whenever the number of constraints and the dimension of the matrices is polynomial. (There is an important caveat, namely that the coefficients of the SoS proof need to be polynomially bounded [90, 91].) Since the number of monomials over n variables of degree d is $\binom{n+d-1}{d} = O(n^d)$, this means that SoS proofs are easy to find whenever the degree d is constant.

On the other hand, if we can somehow prove that the lowest degree of any SoS proof grows with n, this rules out a large class of polynomial-time algorithms. When we can prove them, these SoS lower bounds are thus evidence of computational hardness.

5.3. Sum-of-squares lower bounds: enter the charlatan

To see how we might prove such a lower bound, let us return to our earlier problem. A Charlatan[5] comes along and claims that the system (21) has not just one solution, but many. That is, they claim to know a joint probability distribution over reals x, y, z such that $x^2 = y^2 = z^2 = 1$ and $x + y + z = 0$. To convince you, they offer to tell you the expectation $\mathbb{E}[q]$ of any polynomial $q(x, y, z)$ you desire—but only for q of degree d or less, where in this case $d = 2$.

Let us call the Charlatan's claimed value for $\mathbb{E}[q]$ the *pseudoexpectation*, and denote it $\widetilde{\mathbb{E}}[q]$. How might you catch them in a lie? You are no fool; you know that the expectation of a sum is the sum of the expectations. Since the constraints $f_i(\mathbf{x}) = 0$ must hold identically, you also know that any q that has f_i as a factor must have zero expectation. Finally, you are well aware that the square of any polynomial is everywhere nonnegative, and thus has nonnegative expectation.

Putting this together, the pseudoexpectation must be a linear operator from the space of polynomials of degree d to $\mathbb{R}$ with the following properties:

(a) $\widetilde{\mathbb{E}}[1] = 1$.
(b) $\widetilde{\mathbb{E}}[f_i q] = 0$ for any polynomial $q(x)$ of degree $d - \deg(f_i)$ or less.
(c) $\widetilde{\mathbb{E}}[q^2] \geqslant 0$ for any polynomial $q(x)$ of degree $d/2$ or less.

[5] Many concepts in theoretical computer science have become personified over the years: the Adversary, the Oracle, Arthur and Merlin, Alice, Bob, and Eve, and so on. We propose that the Charlatan be added to this cast of characters.

J. Stat. Mech. (2022) 114015

Let us think of $\widetilde{\mathbb{E}}$ as a bilinear form that takes two polynomials p, q of degree up to $d/2$ and returns $\widetilde{\mathbb{E}}[pq] = \langle p|\widetilde{\mathbb{E}}|q\rangle$. Then condition (3) corresponds to $\widetilde{\mathbb{E}}$ being positive semidefinite, just as for $\mathbf{H}$ above. Since conditions (1) and (2) are linear, finding a pseudoexpectation is another case of semidefinite programming.

In our example, since $d = 2$, the monomials that $\widetilde{\mathbb{E}}$ needs to deal with are just $1, x, y, z$. Without further ado, we present the Charlatan's claim as a multiplication table of pseudoexpectations:

$\widetilde{\mathbb{E}}$	1	x	y	z
1	1	0	0	0
x	0	1	$-1/2$	$-1/2$
y	0	$-1/2$	1	$-1/2$
z	0	$-1/2$	$-1/2$	1

(28)

That is, they claim that x, y, z each have expectation $\widetilde{\mathbb{E}}[x] = \langle 1|\widetilde{\mathbb{E}}|x\rangle = 0$; they each have variance $\widetilde{\mathbb{E}}[x^2] = \langle x|\widetilde{\mathbb{E}}|x\rangle = 1$; and each distinct pair is negatively correlated, with $\widetilde{\mathbb{E}}[xy] = \langle x|\widetilde{\mathbb{E}}|y\rangle = -1/2$. As a result, $\widetilde{\mathbb{E}}[x + y + z] = 0$, and $\widetilde{\mathbb{E}}[(x + y + z)p] = 0$ for any linear function p, satisfying condition (2) above.

It is easy to check that this matrix of pseudomoments is positive semidefinite. Indeed its 3×3 part is the Gram matrix of three unit vectors that are 120° apart. This is impossible for three real-valued variables in $\{\pm 1\}$, but as far as quadratic polynomials of x, y, z are concerned, there is no contradiction.

On the other hand, we already know that we can debunk the Charlatan's claims if we ask about degree-4 polynomials. The left-hand side of (22) must have zero expectation since it is a linear combination of the f_i. By linearity, this would imply that

$$\widetilde{\mathbb{E}}\left[\frac{1}{8}\left((x + y + z)^2 - 1\right)^2\right] = -1 < 0. \tag{29}$$

Thus there is no way to extend the pseudoexpectation in (28) from degree 2 to degree 4 without violating positive semidefiniteness. More generally, an SoS proof of the form (23) would imply

$$\widetilde{\mathbb{E}}\left[\sum_j h_j^2\right] = -1 < 0. \tag{30}$$

Thus for each degree d, there is an SoS proof if and only if there is no pseudoexpectation. These two problems are dual SDPs; a solution to either is a certificate that the other has no solution. In particular, any degree at which the Charlatan can succeed is a lower bound on the degree a refuter needs to prove that no solution exists. In this example, we have shown that degree 4 is both necessary and sufficient to prove that no three variables in $\{\pm 1\}$ can sum to zero.

J. Stat. Mech. (2022) 114015

5.4. What does sum-of-squares understand?

The reader is probably wondering how the SoS framework performs on larger versions of our example. Suppose we have n variables $x_1, \ldots, x_n$. If n is odd, clearly it is impossible to satisfy the system

$$\begin{aligned} x_i^2 - 1 &= 0 \quad \text{for all } i = 1, \ldots, n \\ \sum_{i=1}^{n} x_i &= 0. \end{aligned} \tag{31}$$

To put it differently, if you take an odd number of steps in a random walk on the integers, moving one unit to the left or right on each step, there is no way to return to the origin.

It turns out [92–94] that any SoS proof of this fact requires degree $n+1$. That is, the Charlatan can construct a pseudoexpectation for polynomials of degree d up to $n-1$. This includes the case $n = 3$ we studied above.

How can the Charlatan do this? Since $x_i^2 = 1$ for all i, it suffices for them to construct pseudoexpectations for the multilinear monomials, i.e. those of the form $x_S = \prod_{i \in S} x_i$ for some set $S \subset \{1, \ldots, n\}$. Furthermore, we can symmetrize over all permutations of the x_i, and assume that $\widetilde{\mathbb{E}}[x_S]$ only depends on their degree $|S|$: semidefinite programming is a convex problem, so symmetric problems have symmetric solutions if any.

Now let a_k denote $\widetilde{\mathbb{E}}[x_S]$ for $|S| = k$. Equivalently, $a_k = \widetilde{\mathbb{E}}[x_1 x_2 \ldots x_k]$. We can compute a_k as follows. Suppose I tell you that $n/2$ of the x_i are $+1$, and $n/2$ are -1. (Do not ask whether $n/2$ is an integer.) If we choose a uniformly random set of k distinct variables from among the x_i, then a_k is the average parity of their product. An enjoyable combinatorial exercise gives, for k even,

$$a_k = (-1)^{k/2} \frac{\binom{n/2}{k/2}}{\binom{n}{k}} = (-1)^{k/2} \frac{(k-1)(k-3)(k-5)\ldots 1}{(n-1)(n-3)(n-5)\ldots(n-k+1)} \tag{32}$$

and $a_k = 0$ for k odd.

Again using the fact that $x_i^2 = 1$ for all i, for any two sets S, T we have $x_S x_T = x_{S \triangle T}$ where $\triangle$ denotes the symmetric difference. Thus we define the pseudoexpectation as a bilinear operator that takes monomials x_S, x_T where $|S|, |T| \leqslant d/2$, with matrix elements

$$\langle x_S | \widetilde{\mathbb{E}} | x_T \rangle = \widetilde{\mathbb{E}}[x_S \, x_T] = \widetilde{\mathbb{E}}[x_{S \triangle T}] = a_{|S \triangle T|}, \tag{33}$$

which generalizes (28) above. As long as $d \leqslant n-1$, it turns out that this $\widetilde{\mathbb{E}}$ is positive semidefinite [94]; its spectrum can be analyzed using representation theory [95]. Thus any SoS refutation of the system (31) must be of degree at least $d = n+1$.

This lower bound is tight: any pseudoexpectation on Boolean variables $x_1, \ldots, x_n \in \{\pm 1\}$ of degree $n+1$ must be a true expectation, i.e. must correspond to an actual distribution over the hypercube [96]. Thus at degree $n+1$, the Charlatan can no longer produce a convincing pseudoexpectaton unless solutions actually exist. If n is odd, there are no solutions, so by SDP duality there is a refutation of degree $n+1$.

J. Stat. Mech. (2022) 114015

One way to construct a refutation is as follows. Let w denote $\sum_i x_i$. First we 'prove' that w is an odd integer between $-n$ and n by finding polynomials $g_1, \ldots, g_n$ such that

$$\sum_{i=1}^{n} g_i(\mathbf{x})\,(x_i^2 - 1) = \prod_{t=-n,-n+2,\ldots}^{\ldots,n-2,n} (w - t). \tag{34}$$

For instance, the reader can check that the three terms inside the square brackets in (22) sum to $(w+3)(w+1)(w-1)(w-3)$ where $w = x + y + z$. The polynomials g_i in (34) are guaranteed to exist because, in the ring of polynomials, the set $\{x_i^2 - 1\}$ spans the set of all polynomials that vanish on $\{\pm 1\}^n$. For the experts, $\{x_i^2 - 1\}$ is a Gröbner basis for this ideal.

Now we wish to show that some polynomial with w as a factor, say w^2, is nonzero. To do this, we find a polynomial $q(w)$ that is everywhere positive and that coincides with w^2 at the odd integers between $-n$ and n. By polynomial interpolation, we can take $q(w)$ to be even and of degree $n+1$. For $n = 3$, for instance, we have

$$q(w) = \frac{1}{8}(w^2 - 1)^2 + 1 \;\geqslant\; 1, \tag{35}$$

which we have already written as a sum of squares.

Since the polynomial $q(w) - w^2$ has these odd integers as roots, it is a multiple of the expression in (34). Putting this together for $n = 3$ gives

$$\frac{1}{8}(w+3)(w+1)(w-1)(w+3) \;+\; w^2 = q(w), \tag{36}$$

which is exactly what we wrote in (22).

Now recall that SoS refutations of degree d can be found in polynomial time only if d is a constant. This means that as far as SoS is concerned, proving that (31) is unsatisfiable is hard. Clearly SoS does not understand parity arguments very well.

Morally, this is because the matrix elements (32) are analytic functions of n: they cannot tell whether n is odd or even, or even whether n is an integer or not. To put it differently, binomials like those in the numerator of a_k in (32) will happily generalize to half-integer inputs with the help of the Gamma function. After all, there are $\binom{3}{3/2} = 32/(3\pi) = 3.395\ldots$ ways to take three steps of a random walk and return to the origin.

The 'hardness' of this example may make SoS look like a very weak proof system. But parity is a very delicate thing. If n Boolean variables are represented as $\{0, 1\}$, then their parity is merely their sum mod 2; but if we represent them as spins ± 1, the parity is their product, which is of degree n. When n is large, we would be amazed to find such a term in the Hamiltonian of a physical system. No observable quantity depends on whether the number of atoms in a block of iron is odd or even.

The situation seems similar to XORSAT, whose clauses are linear equations mod 2. Its energy landscape has many of the hallmarks of algorithmic hardness, with clusters, frozen variables, and large barriers between solutions [97]. See also the discussion in subsection 3.4 of section 3. In the noise-free case it can be solved in polynomial time using Gaussian elimination over $\mathbb{Z}_2$. But if we add any noise, for instance only requiring

J. Stat. Mech. (2022) 114015

that 99% of the XORSAT clauses be satisfied, its algebraic structure falls apart and this algorithmic shortcut disappears. So while parity and XORSAT are good cautionary tales, we should not think of them as representative of more generic problems. As we will see next, for many problems with noise, including those involving random matrices and tensors with planted structure, the SoS framework is associated with many algorithms that are conjectured to be optimal.

5.5. Relaxation and the Sherrington–Kirkpatrick model

Above we referred to the pseudoexpectation as the work of a charlatan who falsely claims that an unsatisfiable problem has many solutions. But there is another, less adversarial way to describe this character: rather than trying to fool us, they are a *Relaxer* who honestly solves a less-constrained problem, and thus proves bounds on the optimum of the original problem.[6]

To celebrate the 40th anniversary that inspired this book, let us consider the Sherrington–Kirkpatrick model. Given a coupling matrix J we can write the ground state energy of an Ising spin glass as

$$E_0 = -\max_{\mathbf{x}\in\{\pm 1\}^n} \sum_{i<j} J_{ij} x_i x_j = -\frac{1}{2}\max_{X\in\mathcal{C}} \text{ tr } JX \tag{37}$$

(where we take J to be symmetric and zero on the diagonal). In other words, the energy is quadratic in the spins, but linear in the products $X_{ij} = x_i x_j$. So we just have to maximize a linear function! This is exactly the maximization problem (6) when $p = 2$, ignoring the $-1/2$ factor.

The tricky part is that we have to maximize $\text{tr } JX$ over a complicated set. In (37), $\mathcal{C}$ is the set of matrices $X = |x\rangle\langle x|$ corresponding to actual spin configurations, namely symmetric rank-1 matrices with ± 1 entries and $+1$s on the diagonal. We would get the same maximum if we defined $\mathcal{C}$ to be the polytope of all convex linear combinations of such matrices. But this so-called *cut polytope* has exponentially many facets, making this maximization computationally infeasible [98]. In the worst case where J is designed by an adversary, it is NP-hard since, for instance, it includes Max Cut as a special case [99].

We can relax this problem by allowing X to range over some superset $\mathcal{C}'$ of $\mathcal{C}$. Then the maximum of $\text{tr } JX$ will be greater than or equal to the true maximum over $\mathcal{C}$, providing a lower bound on E_0. A hopeful goal is to find a set $\mathcal{C}'$ whose structure is simple enough to perform this maximization efficiently, while giving a bound that is not too far from the truth.

The first attempt we might make is to allow X to range over all positive semidefinite matrices with trace n. Call this set $\mathcal{C}_0$:

$$\mathcal{C}_0 = \{X : X \succeq 0 \text{ and } \text{tr } X = n\}. \tag{38}$$

[6] Thanks to Tselil Schramm for suggesting the name 'relaxer' for this rehabilitated version of the Charlatan. Perhaps 'slacker' would also work in contemporary English.

Then

$$\max_{x \in \mathcal{C}_0} \operatorname{tr} JX = n\lambda_{\max} \tag{39}$$

where $\lambda_{\max}$ is J's most positive eigenvalue. For the SK model where the J_{ij} are Gaussian with mean 0 and variance $1/n$, the Wigner semicircle law tells us that, in the limit of large n, the spectrum of J is supported on $[-2, 2]$. Thus

$$\lim_{n\to\infty} E_0/n \geqslant -\frac{\lambda_{\max}}{2} = -1. \tag{40}$$

This is fairly far from Parisi's solution $E_0/n = -0.7632$ [100, 101]. Can we get a better bound with some other choice of $\mathcal{C}'$?

We can tighten our relaxation by adding any constraint that holds for the true set of matrices $\mathcal{C}$. Let us start with the constraint that X's diagonal entries are 1. This gives a set of matrices sometimes called the *elliptope* [102],

$$\mathcal{C}_2 = \{X : X \succeq 0 \text{ and } X_{ii} = 1 \text{ for all } i\}. \tag{41}$$

We might hope that maximizing $\operatorname{tr} JX$ over $\mathcal{C}_2$ rather than $\mathcal{C}_0$ gives a better bound on the energy. Unfortunately, this is not the case: for any constant $\varepsilon > 0$, with high probability there is an $X \in \mathcal{C}_2$ such that $\operatorname{tr} JX \geqslant 2 - \varepsilon$. We will sketch the proof of [103].

First let v_λ denote the eigenvector of J with eigenvalue λ, normalized so that $|v_\lambda|^2 = 1$. Let m denote the number of eigenvalues in the interval $[2 - \varepsilon, 2]$. These eigenvalues span a low-energy subspace where $E_0 \approx -1$. Now define Y as

$$Y = \frac{n}{m} \sum_{\lambda \in [2-\varepsilon, 2]} |v_\lambda\rangle\langle v_\lambda|. \tag{42}$$

That is, Y is n/m times the projection operator onto this subspace. Thus $Y \succeq 0$ and $\operatorname{tr} JY \geqslant (2 - \varepsilon)n$.

We can write Y's diagonal entries as

$$Y_{ii} = \frac{n}{m} \sum_{\lambda} (v_\lambda)_i^2. \tag{43}$$

Since the distribution of Gaussian random matrices is rotationally invariant, the v_λ are distributed as a uniformly random set of m orthonormal vectors in n dimensions. Thus the $(v_\lambda)_i^2$ are asymptotically independent, and are $1/n$ on average. As a result, each Y_{ii} is concentrated around 1.

To turn Y into an X such that $X_{ii} = 1$ holds exactly, define D as the diagonal matrix $D_{ii} = Y_{ii}$ and let

$$X = D^{-1/2} Y D^{-1/2}. \tag{44}$$

Clearly $X \succeq 0$. Moreover, since D itself is close to the identity, we have $\operatorname{tr} JX = \operatorname{tr} JY$ up to a vanishing error term. Since $X \in \mathcal{C}_2$, we have shown that $\mathcal{C}_2$ does not give a bound any better than the simple spectral bound provided by $\mathcal{C}_0$.

The alert reader will note that $\mathcal{C}_2$ is exactly the set of pseudoexpectations $\widetilde{\mathbb{E}}$ that a degree-2 charlatan can choose from. If $X_{ij} = \widetilde{\mathbb{E}}[x_i x_j]$, then $X \succeq 0$ and $X_{ii} = \widetilde{\mathbb{E}}[x_i^2] = 1$.

J. Stat. Mech. (2022) 114015

So whether we regard X as the solution to a relaxed problem or a false claim about the covariances $\mathbb{E}[x_i x_j]$, we have shown that degree-2 SoS proofs cannot establish a bound better than $E_0/n > -1$ on the SK ground state energy. That is, they are incapable of refuting the claim that there are states with energy $-1+\varepsilon$ or below, for arbitrarily small ε.

(There is a subtlety here. The refuter's goal is not to understand the typical ground state energy of the SK model, but to provide ironclad proofs for individual realizations J that their ground state energy is above a certain point. What we have shown is that, for most realizations J, there is no degree-2 proof that its ground state energy is noticeably above -1.)

We should also note that, just as $\mathcal{C}$ is the set of matrices $X = |x\rangle\langle x|$ where the $x_i = \pm 1$ are Ising spins, $\mathcal{C}_2$ is the set of matrices $X = |x\rangle\langle x|$ where the x_i are n-dimensional vectors with $|x_i|^2 = 1$. So while Ising spins cannot achieve the covariances $X_{ij} = \widetilde{\mathbb{E}}[x_i x_j]$ that the Charlatan claims, these vector-valued spins can achieve them in the sense that $X_{ij} = \langle x_i \,|\, x_j \rangle$.

This is the heart of the Goemans–Williamson approximation algorithm for Max Cut [104]—or, in physics terms, bounding the ground-state energy of an antiferromagnet. In Max Cut, our goal is to assign a spin $x_i = \pm 1$ to each vertex, and maximize the number w of edges whose spins are opposite. For a graph with m edges and adjancency matrix A, this is

$$w = \frac{1}{2}(m - \langle x|A|x\rangle). \tag{45}$$

If we relax this problem by letting the x_i be unit-length vectors in $\mathbb{R}^n$ instead of just ± 1, this becomes an SDP that we can solve in polynomial time. It can be shown that this relaxation increases w by a factor of at most $1/0.878 = 1.138\ldots$, so the optimum of this relaxation is not too far from that of the original problem.

We do not know whether going to higher-degree SoS improves this approximation ratio. If we assume the unique games conjecture (a plausible strengthening of P $\neq$ NP) then no polynomial-time algorithm can do better than Goemans–Williamson [105].[7] This suggests that going to degree 4, 6, and so on does not give a better algorithm, but even for degree 4 this is an open question.

On the other hand, for the SK model it was recently shown [106] that higher-degree SoS does not improve our bounds on the ground state energy, as we will see next.

5.6. Beyond degree 2

Can SoS proofs of some constant degree $d > 2$ prove a tighter bound on the ground state energy E_0 of the Sherrington–Kirkpatrick model? Do higher-degree polynomials help us go beyond the simple spectral bound $E_0 \geqslant -1$?

The Charlatan's job for $d = 4$ is already quite interesting. In addition to providing $X \in \mathcal{C}_2$, they now have to provide an $\binom{n}{2}$-dimensional matrix $X^{(4)}$, with rows and

[7] This is usually presented the other way around. If we round the relaxed solution to ± 1 spins by cutting $\mathbb{R}^n$ with a random hyperplane, the Goemans–Williamson algorithm gives a cut that is at least 0.878 times the optimum, and the unique games conjecture implies that this cannot be improved. The same argument [105] implies an upper bound on the relaxed solution. (Thanks to Tim Kunisky for pointing this out).

J. Stat. Mech. (2022) 114015

columns for each pair (i, j), such that

$$X^{(4)}_{(i,j),(k,\ell)} = \widetilde{\mathbb{E}}[x_i x_j x_k x_\ell]. \tag{46}$$

Thus $X^{(4)}$ must have the symmetries of a symmetric four-index tensor,

$$X^{(4)}_{(i,j),(k,\ell)} = X^{(4)}_{(i,k),(j,\ell)} = X^{(4)}_{(i,\ell),(j,k)}. \tag{47}$$

In addition, $X^{(4)}$ needs to be consistent with the degree-2 pseudexpectations and the constraint $x_i^2 = 1$. Thus

$$X^{(4)}_{(i,j),(i,k)} = \widetilde{\mathbb{E}}\left[x_i^2 x_j x_k\right] = \widetilde{\mathbb{E}}[x_j x_k] = X_{jk} \tag{48}$$

$$X^{(4)}_{(i,j),(i,j)} = \widetilde{\mathbb{E}}\left[x_i^2 x_j^2\right] = 1. \tag{49}$$

(We saw these relations in section 5.4 where we wrote $x_S x_T = x_{S \triangle T}$.) Finally, as always $X^{(4)}$ must be positive semidefinite,

$$X^{(4)} \succeq 0. \tag{50}$$

The energy $E = -(1/2)\mathrm{tr}\, JX$ is still a function of the second-order pseudoexpectation X. But not all matrices X in $\mathcal{C}_2$ can be extended to fourth order in this way: the set

$$\mathcal{C}_4 = \{X \in \mathcal{C}_2 : \exists X^{(4)} \text{ such that (47)–(50) holds}\} \tag{51}$$

is a proper subset of the elliptope $\mathcal{C}_2$. In other words, armed with degree-4 SoS proofs, a refuter can prove some new constraints on the covariances $X_{ij} = x_i x_j$ that go beyond $X_{ii} = 1$ and $X \succeq 0$.

For example, consider any three Ising spins, x_i, x_j, and x_k. Their products $(x_i x_j, x_j x_k, x_i x_k)$ can only take the values $(1, 1, 1)$, $(1, -1, -1)$, $(-1, 1, -1)$, and $(-1, -1, 1)$. Thus the expectation of their products (X_{ij}, X_{jk}, X_{ik}) must lie in the convex hull of these four vectors, namely the tetrahedron with these four vertices. The facets of this tetrahedron are the linear inequalities

$$X_{ij} + X_{jk} + X_{ik} + 1 \geqslant 0 \tag{52}$$

$$X_{ij} - X_{jk} - X_{ik} + 1 \geqslant 0 \tag{53}$$

$$-X_{ij} + X_{jk} - X_{ik} + 1 \geqslant 0 \tag{54}$$

$$-X_{ij} - X_{jk} + X_{ik} + 1 \geqslant 0. \tag{55}$$

We have already seen a pseudoexpectation in $\mathcal{C}_2$ that violates the first of these inequalities—namely (28) where $X_{ij} = X_{jk} = X_{ik} = -1/2$. Thus we cannot prove these inequalities with degree-2 sum-of-squares. But we can prove them with degree 4, and we already have! After all, we can rewrite (52) as

$$\widetilde{\mathbb{E}}[x_i x_j + x_j x_k + x_i x_k + 1] \geqslant 0. \tag{56}$$

J. Stat. Mech. (2022) 114015

But if $x_i^2 = x_j^2 = x_k^2 = 1$ this is equivalent to

$$\widetilde{\mathbb{E}}\big[(x_i + x_j + x_k)^2\big] \geqslant 1. \tag{57}$$

Looking again at our proof (22) that no three spins can sum to zero, the reader will see that we in fact proved that $(x+y+z)^2 \geqslant 1$ whenever $x^2 = y^2 = z^2 = 1$. The symmetry operations $x \mapsto -x$, $y \mapsto -y$, and $z \mapsto -z$ give similar proofs of (53)–(55).

Thus any matrix X that violates these 'triangle inequalities' can be refuted by degree-4 sum-of-squares. More generally, since any $t+1$ pseudoexpectation on t spin variables is a true expectation [96], any linear inequality on the covariances of t spins—or equivalently any inequality that involves a $t \times t$ principal minor of X—can be proved with degree $t+1$ sum-of-squares.

Perhaps these and other degree-4 constraints will finally give a better bound on E_0? Sadly—or happily if you love computational hardness—they do not. In fact, no constant degree can refute the claim that some spin configuration lies in the low-energy subspace, and thus prove a bound tighter than $\operatorname{tr} JX \leqslant 2$ or $E_0 \geqslant -1$.

One intuition for this is that for natural degree-2 pseudoexpectations, like the X we constructed above (44) by projecting onto the low-energy subspace, triangle inequalities and their generalizations already hold with room to spare. In the SK model we typically have $\widetilde{\mathbb{E}}[x_i x_j] = O(1/\sqrt{n})$, so (52)–(55) all read $1 + O(1/\sqrt{n}) \geqslant 0$. Thus, with perhaps a slight perturbation to make it positive definite and full rank, X is already deep inside the elliptope $\mathcal{C}_2$, and is not refuted by the additional inequalities we can prove with low-degree SoS proofs.

There are several ways to make this intuition rigorous. One is to explicitly construct higher-degree pseudoexpectations $X^{(4)}$, $X^{(6)}$, and so on that extend X in a natural way, somewhat like a cluster expansion in physics. For instance, we could define

$$X^{(4)}_{(i,j),(k,\ell)} = X_{ij}X_{k\ell} + X_{ik}X_{j\ell} + X_{i\ell}X_{jk} - 2\sum_{m=1}^{n} X_{im}X_{jm}X_{km}X_{\ell m}. \tag{58}$$

This expression has the permutation symmetry of (47). The first three terms look like Wick's theorem or Isserlis' theorem for the moments of Gaussian variables [107]; the reader can check that by cancelling two of these terms when $k = \ell$, the sum over m ensures the consistency relations (48) and (49) to leading order. A small perturbation then satisfies these conditions exactly [108] and it is relatively easy to show that the result is positive semidefinite; see also [109]. A similar approach works for degree 6 [110].

5.7. Pseudocalibration and clever planted models

While constructions like (58) could probably be carried out for higher degree, the recent proof [106] that no constant degree of SoS can improve the bound on E_0 comes from a different direction called *pseudocalibration* [111, 112].

In pseudocalibration, the Charlatan claims that the data is generated by a planted model where the claimed solution is built in, rather than the (true) null model. In the Sherrington–Kirkpatrick model this means pretending that the couplings J have been chosen so that some Boolean vector $x \in \{\pm 1\}^n$ achieves the spectral bound $E_0 = -1$.

J. Stat. Mech. (2022) 114015

If we can construct a pseudoexpectation around this idea, then low-degree SoS cannot tell the difference between the null model and the planted model. In particular, it cannot prove that the planted solution does not exist.

Following [112], we can briefly describe pseudocalibration as follows. We consider two joint distributions on a signal x and observed data Y. In both cases, we choose x from a prior $P(x)$. In the null model, we choose Y independently of x with probability $P_0(Y)$; in the planted model, we choose Y with probability $P_1(Y|x)$. Thus

$$P_0(x, Y) = P_0(Y)\, P(x)$$
$$P_1(x, Y) = P_1(Y \mid x)\, P(x) = P_1(Y)\, P_1(x \mid Y),$$

where $P_1(Y) = \mathbb{E}_{x\sim P(x)} P_1(Y \mid x)$ is Y's likelihood in the planted model.

In the Charlatan's first attempt, they define the pseudoexpectation of a function $q(x)$ as its true expectation given Y, but reweighted to change the null model into the planted one:

$$\begin{aligned}\widetilde{\mathbb{E}}[q(x) \mid Y] &= \mathbb{E}_{x\sim P(x)}\left[\frac{P_1(x, Y)}{P_0(x, Y)}\, q(x)\right] \\ &= \mathbb{E}_{x\sim P(x)}\left[\frac{P_1(Y)\, P_1(x \mid Y)}{P_0(Y)\, P(x)}\, q(x)\right] \\ &= \frac{P_1(Y)}{P_0(Y)}\, \mathbb{E}_{x\sim P_1(x \mid Y)} q(x). \end{aligned} \tag{59}$$

That is, the pseudoexpectation of $q(x)$ is its true expectation in the posterior distribution $P_1(x|Y)$, multiplied by the likelihood ratio $P_1(Y)/P_0(Y)$.

This pseudoexpectation is proportional to a true expectation, albeit over another distribution. Thus it is positive semidefinite, $\widetilde{\mathbb{E}}[q^2] \geqslant 0$. Similarly, if $P(x)$ and therefore $P(x|Y)$ are supported on x satisfying some constraint $f_i(x) = 0$, then $\widetilde{\mathbb{E}}[f_i q] = 0$ for any q.

Moreover, (59) gives any function of x and Y the expectation over the null model that it would have in the planted model,

$$\mathbb{E}_{Y\sim P_0} \widetilde{\mathbb{E}}[q(x, Y)] = \mathbb{E}_{(x,Y)\sim P_0}\left[\frac{P_1(x, Y)}{P_0(x, Y)}\, q(x, Y)\right] = \mathbb{E}_{(x,Y)\sim P_1} q(x, Y). \tag{60}$$

where we took the average over Y as well as x.

On the other hand, for individual Y we have some trouble. For instance, (59) gives $\widetilde{\mathbb{E}}[1 \mid Y] = P_1(Y)/P_0(Y)$, the likelihood ratio instead of 1. This would make it easy to catch the Charlatan whenever the null and planted models can be distinguished information-theoretically. Moreover, while the planted model guarantees that Y has a solution x, most Y drawn from the null model have no such solution. In that case we have $P_1(Y) = 0$, and the posterior distribution $P_1(x|Y)$ is undefined.

We can fix both these problems by projecting $\widetilde{\mathbb{E}}[q(x) \mid Y]$ into the space of low-degree polynomials, both in x and in Y. In other words, we take its Taylor series in x and Y up to some degree. For Boolean variables, this is equivalent to keeping just the low-frequency part of the Fourier spectrum; in some cases, we might project onto a suitable

J. Stat. Mech. (2022) 114015

set of orthogonal polynomials. This preserves the appearance (60) of the planted model for functions of low degree in x and Y.

If all goes well, this projection smooths the likelihood ratio, keeping it concentrated around its expectation 1. It also smooths the posterior distribution $P_1(x|Y)$ as a function of Y, extending it from the small set of Y produced by the planted model (for instance, the few instances of the SK model where $E_0 = -1$) to the more generic Y produced by the null model.

However, the Charlatan has to preserve enough of the dependence on Y to make $\widetilde{\mathbb{E}}[q \,|\, Y]$ convincing. To do this for $q(x)$ of degree d, they typically need to preserve terms in Y up to some sufficient degree $D > d$.

Showing that $\widetilde{\mathbb{E}}$ remains positive semidefinite after this projection, and that it continues to satisfy the constraints $\widetilde{\mathbb{E}}[f_i] = 0$, can involve summing over many combinatorial terms. This was first done for the planted Clique problem [111]. While each application since then has involved special-purpose calculations, several conjectures [112] offer general principles by which this program might be extended.

The projection of $\widetilde{\mathbb{E}}[1 \,|\, Y] = P_1(Y)/P_0(Y)$ into low-degree polynomials in Y is of its own interest: it is the *low-degree likelihood ratio*. If it is usually close to 1 in the null model but is large in the planted model, then it provides a polynomial-time hypothesis test for distinguishing between these two. Thus showing that it has bounded variance in the null model is in itself evidence of computational hardness [113]. In particular, [114] showed that the degree-D likelihood ratio fails to improve the bound on the SK model for any $D = o(n/\log n)$. This does not in itself prove that SoS fails up to this degree, but the two approaches are closely related.

We conclude this section by discussing the choice of planted model. Proving that refutation is hard might require a clever way to hide a solution, as opposed to the standard spiked matrices and tensors. For instance, to prove their SoS lower bounds on the Sherrington–Kirkpatrick model, [106] related a planted model proposed by [109] where a random subspace (i.e. the low-energy subspace) contains a Boolean vector to a model of Gaussian random vectors, where in the planted case these vectors belong to two parallel hyperplanes.

More generally, there is a long history in physics and computer science of 'quiet' planting, in order to make the solution as difficult as possible to detect [66, 115]. The quieter the planting, the harder it is to distinguish from the null model. In this case, we want the planting to be *computationally* quiet [114], and in particular to match the low-degree moments of the null distribution. For instance, rather than the usual spiked model where we add a rank-1 perturbation to a Gaussian random matrix J—which disturbs the entire spectrum—we can plant a large eigenvalue more quietly by increasing the eigenvalue of a specific eigenvector [116].

5.8. Optimal algorithms and the curious case of tensor PCA

We have talked a lot about what SoS algorithms cannot do. But for many problems they seem to be optimal, performing as well as any polynomial-time algorithm can. For Max Cut and the Sherrington–Kirkpatrick model, we have seen evidence that this is the case even at degree 2.

J. Stat. Mech. (2022) 114015

Thus in many cases, SoS algorithms seem to succeed or fail at the same place where physics suggests a hard/easy transition. Even when these thresholds do not coincide exactly, they often have the same scaling and thus differ by a constant. For example, degree-2 SoS—also known as the Lovász ϑ function—can refute graph colorings in random regular graphs within a factor of 4 of the Kesten–Stigum transition [117], and it is possible that higher-degree SoS does better.

While refuting the existence of a planted solution lets SoS solve the detection problem—distinguishing the null from a planted model—a refinement of this idea often yields algorithms for reconstruction as well. Roughly speaking, if we can refute the existence of a solution when it does not exist, we can often find it when it does [112].

To see how this works, consider a planted model, and let x^* denote the ground truth. Let $\phi(x)$ be some polynomial for which $\phi(x^*) \leqslant \phi^*$: for instance, in PCA, $\phi(x)$ could be the ℓ_2 distance between the signal matrix $|x\rangle\langle x|$ and the observed matrix Y. Now suppose there is a degree-d refutation of the claim that there are any good solutions far from the ground truth: that is, a proof that if $\phi(x) \leqslant \phi^*$ then $|x - x^*|^2 \leqslant \varepsilon$. Then any degree-$d$ pseudoexpectation must claim that $|\widetilde{\mathbb{E}}[x] - x^*|^2 \leqslant \varepsilon$, and $\widetilde{\mathbb{E}}[x]$ is a good estimate of x^*.

This approach yields efficient algorithms for many problems [118, 119], including tensor PCA [120]. But for tensor PCA in particular, a curious gap appeared between algorithms and physics. Recall from section 2.1 that tensor PCA, a.k.a. the spiked tensor model, is a planted model of p-index tensors defined by

$$Y = \lambda u^{\otimes p} + J. \tag{61}$$

Here λ is the signal-to-noise ratio, the planted vector u is normalized so that $|u|^2 = n$, and the noise tensor J is permutation-symmetric with Gaussian entries $\mathcal{N}(0, 1)$. The information-theoretic transition occurs at $\lambda = \lambda_c n^{-(p-1)/2}$ for a constant λ_c depending on p and u's prior [52, 121].

The best known polynomial-time algorithms, on the other hand, require a considerably larger signal-to-noise ratio, $\lambda \gtrsim n^{-p/4}$. One such algorithm, called 'tensor unfolding,' reinterprets Y as a matrix and iteratively applies PCA to it. For $p = 4$, for instance, we treat Y as an $n^2 \times n^2$ matrix $Y_{ij,k\ell}$ and find its leading eigenvector v. Since $v \approx u \otimes u$, we then treat v as an $n \times n$ matrix and estimate u as its leading eigenvector. At each stage we unfold the tensor into a matrix which is as square as possible.

Other algorithms, that also succeed for $\lambda \gtrsim n^{-p/4}$, can be derived directly from sum-of-squares [122]. Conversely, SoS lower bounds suggest that there is no polynomial-time algorithm if $\lambda \lesssim n^{-p/4}$, so this appears to be the algorithmic threshold [123].[8]

On the other hand, physics-based algorithms such as belief propagation and its asymptotic cousin AMP, as well as Langevin dynamics, all fail unless $\lambda \gtrsim n^{-1/2}$, making these algorithms suboptimal whenever $p \geqslant 3$ [121, 124]. This does not contradict conjectures of optimality from section 4.4 as those were restricted to the scaling of parameters corresponding to the information-theoretical regime which in this case is

[8] Our notation $\gtrsim$ and $\lesssim$ suppresses logarithmic factors. These are consequences of matrix Chernoff bounds, and could probably be removed.

J. Stat. Mech. (2022) 114015

$\lambda \approx n^{-(p-1)/2}$. Never-the-less, focusing on the regime discussed here, does sum-of-squares know something that physics does not?

This conundrum has a satisfying answer [125]: in the scaling regime $\lambda \gtrsim n^{-(p-1)/2}$ we were using the wrong physics. Belief propagation keeps track of pairwise correlations. When we compute the Bethe free energy, we pretend that the Gibbs distribution, i.e. the posterior distribution $P(x|\,Y)$, has the form

$$P(x) = \prod_i \mu_i(x_i) \times \prod_{(i,j)} \frac{\mu_{ij}(x_i, x_j)}{\mu_i(x_i)\,\mu_j(x_j)} \tag{62}$$

where μ_i and μ_{ij} are one- and two-point marginals. Minimizing the resulting free energy is equivalent to finding fixed points of belief propagation [126].

But when $p \geqslant 3$, it becomes vital to consider correlations between clusters of p variables. This gives rise to a hierarchy of free energies due to [127]. For $p = 3$, for instance, we assume that the Gibbs distribution has the form

$$P(x) = \prod_i \mu_i \times \prod_{(i,j)} \frac{\mu_{ij}}{\mu_i\,\mu_j} \times \prod_{(i,j,k)} \frac{\mu_{ijk}\,\mu_i\,\mu_j\,\mu_k}{\mu_{ij}\,\mu_{jk}\,\mu_{ik}} \tag{63}$$

(where for readability we suppress (x_i), (x_i, x_j), and so on). At each level of this approximation, we correct for overcounting smaller clusters. Taking the logarithm of this expression and averaging over x gives an inclusion-exclusion-like formula for the entropy.

There are several ways one might turn this into a spectral algorithm. One is to write an iterative algorithm to minimize the free energy. This gives rise to a generalization of belief propagation in which each variable sends messages to clusters of up to $p-1$ variables with which it interacts [128, 129]. One could then linearize this message-passing algorithm around a trivial fixed point, producing a operator analogous to the non-backtracking operator for belief propagation [130, 131].

An alternate approach is to compute the Hessian of the free energy at a trivial fixed point, generalizing the use of the Bethe Hessian for spectral clustering in graphs [132]. This gives rise to the following operator. For a set $U = \{s_1, \ldots, s_p\}$ with $|U| = p$, let Y_U denote $Y_{s_1,\ldots,s_p}$. Fix $\ell \geqslant p/2$. Then define the following $\binom{n}{\ell}$-dimensional operator, whose rows and columns are indexed by sets S, T with $|S| = |T| = \ell$:

$$M_{S,T} = \begin{cases} Y_{S \triangle T} & \text{if } |S \triangle T| = p \\ 0 & \text{otherwise,} \end{cases} \tag{64}$$

where $\triangle$ again denotes the symmetric difference.

The spectral norm of M can be used as a test statistic to distinguish the planted model from the null model where $\lambda = 0$. In addition, the leading eigenvector of M points approximately to the minimum of the free energy, and a voting procedure yields a good estimate of the signal u. This yields polynomial-time algorithms for detection and reconstruction whenever $\lambda \gtrsim n^{-p/4}$, matching the SoS threshold. Thus the marriage of algorithms and statistical physics is redeemed [125].

J. Stat. Mech. (2022) 114015

The same analysis matches a continuum of subexponential-time algorithms at smaller values of λ [133] and yields a simpler refutation of random constraint satisfaction problems at high clause densities [134]. These 'Kikuchi matrices' have additional applications, e.g. [135].

6. Conclusion

What does the future hold? As our understanding of algorithms deepens, we hope to understand the universal characteristics that make problems easy or hard, unifying larger and larger classes of polynomial-time algorithms and connecting them rigorously with physical properties of the energy landscape. Very recently, [136] connected the low-degree likelihood ratio with the Franz–Parisi potential, adding to the evidence that free energy barriers imply computational hardness. We will know much more in a few years than we know now.

Acknowledgments

We are deeply grateful to Tim Kunisky, Tselil Schramm, and Alex Wein for helpful comments on section 5. We also thank Freya Behrens, Giovanni Piccioli, Paula Mürmann, Yatin Dandi, Emanuele Troiani for their useful comments on the manuscript. C M is supported by NSF grant BIGDATA-1838251, while D G acknowledges the funding from grant DMS-2015517.

References

[1] Moore C and Mertens S 2011 *The Nature of Computation* (Oxford: Oxford University Press)
[2] Cook S A 1971 The complexity of theorem-proving procedures *Proc. 3rd Annual Symp. Theory of Computing* pp 151–8
[3] Fu Y and Anderson P W 1986 *J. Phys. A: Math. Gen.* **19** 1605
[4] Cheeseman P C *et al* 1991 Where the really hard problems are *IJCAI* **91** 331–7
[5] Monasson R, Zecchina R, Kirkpatrick S, Selman B and Troyansky L 1999 *Nature* **400** 133–7
[6] Donoho D L, Gavish M and Johnstone I M 2018 *Ann. Stat.* **46** 1742
[7] Lesieur T, Krzakala F and Zdeborová L 2017 *J. Stat. Mech.* 073403
[8] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
[9] Babacan S D, Luessi M, Molina R and Katsaggelos A K 2012 *IEEE Trans. Signal Process.* **60** 3964–77
[10] Moore C 2017 Bull. EATCS **121** (http://eatcs.org/beatcs/index.php/beatcs/article/view/480)
[11] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[12] Krauth W and Mézard M 1989 *J. Phys.* **50** 3057–66
[13] LeCun Y, Bengio Y and Hinton G 2015 *Nature* **521** 436–44
[14] Achlioptas D and Ricci-Tersenghi F 2006 On the solution-space geometry of random constraint satisfaction problems *Proc. 38th Annual Symp. Theory of Computing* pp 130–9
[15] Mézard M, Mora T and Zecchina R 2005 *Phys. Rev. Lett.* **94** 197205
[16] Gamarnik D 2021 *Proc. Natl Acad. Sci.* **118**
[17] O'Donnell R 2014 *Analysis of Boolean Functions* (Cambridge: Cambridge University Press)
[18] Gamarnik D and Jagannath A 2021 *Ann. Probab.* **49** 180–205
[19] Gamarnik D, Jagannath A and Wein A S 2020 Low-degree hardness of random optimization problems *61st Annual Symp. Foundations of Computer Science*
[20] Wein A S 2020 *Math. Stat. Learning* **4** 221

J. Stat. Mech. (2022) 114015

[21] Gamarnik D, Jagannath A and Wein A S 2020 arXiv:2004.12063
[22] Farhi E, Gamarnik D and Gutmann S 2020 arXiv:2004.09002
[23] Chou C N, Love P J, Sandhu J S and Shi J 2021 arXiv:2108.06049
[24] Basso J, Gamarnik D, Mei S and Zhou L 2022 arXiv:2204.10306
[25] Parisi G 1980 *J. Phys. A: Math. Gen.* **13** L115
[26] Mézard M, Parisi G and Virasoro M A 1987 *Spin-Glass Theory and Beyond Lecture Notes in Physics* vol 9 (Singapore: World Scientific)
[27] Guerra F and Toninelli F L 2002 *Commun. Math. Phys.* **230** 71–9
[28] Talagrand M 2006 *Ann. Math.* **163** 221–63
[29] Panchenko D 2013 *Ann. Math.* **177** 383–93
[30] Panchenko D 2013 *The Sherrington-Kirkpatrick Model* (Berlin: Springer)
[31] Crisanti A and Sommers H J 1992 *Z. Phys.* B **87** 341–54
[32] Subag E 2021 *Commun. Pure Appl. Math.* **74** 1021–44
[33] Montanari A 2021 *SIAM J. Comput.* FOCS19–1
[34] El Alaoui A, Montanari A and Sellke M 2021 *Ann. Probab.* **49** 2922–60
[35] Chen W K, Gamarnik D, Panchenko D and Rahman M 2019 *Annals of Probability.* **47** 1587–618
[36] Auffinger A and Chen W K 2018 *Adv. Math.* **330** 553–88
[37] Chatterjee S 2009 arXiv:0907.3381
[38] Chen W K, Panchenko D *et al* 2018 *Ann. Appl. Probab.* **28** 1356–78
[39] Gamarnik D and Sudan M 2017 *Ann. Probab.* **45** 2353–76
[40] Huang B and Sellke M 2021 arXiv:2110.07847
[41] Bresler G and Huang B 2021 FOCS 2021
[42] Rossman B 2008 On the constant-depth complexity of k-clique *Proc. 40th Annual ACM Symp. Theory of Computing* pp 721–30
[43] Rossman B 2010 Average-case complexity of detecting cliques *PhD Thesis* Massachusetts Institute of Technology
[44] Li Y, Razborov A and Rossman B 2017 *SIAM J. Comput.* **46** 936–71
[45] Rossman B 2018 Lower bounds for subgraph isomorphism *Proc. Int. Congress of Mathematicians: Rio de Janeiro 2018* (Singapore: World Scientific) pp 3425–46
[46] Achlioptas D, Coja-Oghlan A and Ricci-Tersenghi F 2011 *Random Struct. Alg.* **38** 251–68
[47] Aubin B, Perkins W and Zdeborová L 2019 *J. Phys. A: Math. Theor.* **52** 294003
[48] Abbe E, Li S and Sly A 2022 Proof of the contiguity conjecture and lognormal limit for the symmetric perceptron *IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)* pp 327–38
[49] Abbe E, Li S and Sly A 2021 arXiv:2111.03084
[50] Gamarnik D, Kizildag E, Perkins W and Xu C 2021 arXiv:2203.15667
[51] Perkins W and Xu C 2021 Frozen 1-RSB structure of the symmetric Ising perceptron *Proc. 53rd Annual ACM SIGACT Symp. Theory of Computing* pp 1579–88
[52] Lesieur T, Miolane L, Lelarge M, Krzakala F and Zdeborová L 2017 Statistical and computational phase transitions in spiked tensor estimation *2017 IEEE Int. Symp. Information Theory (ISIT)* (IEEE) pp 511–5
[53] Barbier J, Krzakala F, Macris N, Miolane L and Zdeborová L 2019 *Proc. Natl Acad. Sci. USA* **116** 5451–60
[54] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: An Introduction* vol 111 (Oxford: Clarendon)
[55] Zdeborová L and Krzakala F 2016 *Adv. Phys.* **65** 453–552
[56] Thouless D J, Anderson P W and Palmer R G 1977 *Phil. Mag.* **35** 593–601
[57] Bolthausen E 2014 *Commun. Math. Phys.* **325** 333–66
[58] Bayati M and Montanari A 2011 *IEEE Trans. Inf. Theory* **57** 764–85
[59] Javanmard A and Montanari A 2013 *Inf. Inference* **2** 115–44
[60] Bayati M, Lelarge M and Montanari A 2015 *Ann. Appl. Probab.* **25** 753–822
[61] Gerbelot C and Berthier R 2021 arXiv:2109.11905
[62] Decelle A, Krzakala F, Moore C and Zdeborová L 2011 *Phys. Rev.* E **84** 066106
[63] Kesten H and Stigum B P 1966 *Ann. Math. Stat.* **37** 1211–23
[64] Mossel E, Neeman J and Sly A 2012 arXiv:1202.1499
[65] Krzakala F, Mézard M, Sausset F, Sun Y and Zdeborová L 2012 *J. Stat. Mech.* P08009
[66] Zdeborová L and Krzakala F 2011 *SIAM J. Discrete Math.* **25** 750–70
[67] Ricci-Tersenghi F, Semerjian G and Zdeborová L 2019 *Phys. Rev.* E **99** 042109
[68] Semerjian G, Sicuro G and Zdeborová L 2020 *Phys. Rev.* E **102** 022304

J. Stat. Mech. (2022) 114015

[69] Celentano M, Montanari A and Wu Y 2020 The estimation error of general first order methods *Conf. Learning Theory* (PMLR) pp 1078–141
[70] Franz S, Mézard M, Ricci-Tersenghi F, Weigt M and Zecchina R 2001 *Europhys. Lett.* **55** 465
[71] Gamarnik D, Kizildag E C and Zadik I 2021 *IEEE Trans. Inf. Theory* **67** 8109–39
[72] Song M J, Zadik I and Bruna J 2021 *Adv. Neural Inf. Process. Syst.* **34** 29602–15
[73] Antenucci F, Franz S, Urbani P and Zdeborová L 2019 *Phys. Rev.* X **9** 011020
[74] Braunstein A, Mézard M and Zecchina R 2005 *Random Struct. Alg.* **27** 201–26
[75] Antenucci F, Krzakala F, Urbani P and Zdeborová L 2019 *J. Stat. Mech.* 023401
[76] Celentano M and Montanari A 2019 arXiv:1903.10603
[77] Sompolinsky H, Tishby N and Seung H S 1990 *Phys. Rev. Lett.* **65** 1683
[78] Chiara Angelini M and Ricci-Tersenghi F 2022 arXiv:2206.04760
[79] Mannelli S S, Biroli G, Cammarota C, Krzakala F, Urbani P and Zdeborová L 2020 *Phys. Rev.* X **10** 011057
[80] Sarao Mannelli S, Biroli G, Cammarota C, Krzakala F, Urbani P and Zdeborová L 2020 *Adv. Neural Inf. Process. Syst.* **33** 3265–74
[81] Sarao Mannelli S, Vanden-Eijnden E and Zdeborová L 2020 *Adv. Neural Inf. Process. Syst.* **33** 13445–55
[82] Mignacco F, Urbani P and Zdeborová L 2021 *Mach. Learn.: Sci. Technol.* **2** 035029
[83] Barak B and Steurer D 2014 Sum-of-squares proofs and the quest toward optimal algorithms *Proc. Intl. Congress of Mathematicians (ICM)*
[84] Krivine J L 1964 *J. Anal. Math.* **12** 307–26
[85] Stengle G 1974 *Math. Ann.* **207** 87–97
[86] Shor N Z 1987 *Cybernetics* **23** 695–700
[87] Nesterov Y 2000 Squared functional systems and optimization problems *High Performance Optimization* (Berlin: Springer) pp 405–40
[88] Parrilo P A 2000 Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization *PhD Thesis* California Institute of Technology
[89] Lasserre J B 2001 *SIAM J. Optim.* **11** 796–817
[90] O'Donnell R 2017 SOS is not obviously automatizable, even approximately *8th Innovations in Theoretical Computer Science Conf. (ITCS 2017)*
[91] Raghavendra P and Weitz B 2017 On the bit complexity of sum-of-squares proofs *44th Int. Colloq. Automata, Languages, and Programming (ICALP 2017)* 80 pp 80:1–80:13
[92] Grigoriev D 2001 *Comput. Complexity* **10** 139–54
[93] Grigoriev D 2001 *Theor. Comput. Sci.* **259** 613–22
[94] Laurent M 2003 *Math. OR* **28** 871–83
[95] Kunisky D and Moore C 2022 arXiv:2203.05693
[96] Fawzi H, Saunderson J and Parrilo P A 2016 *Math. Program.* **160** 149–91
[97] Cocco S, Dubois O, Mandler J and Monasson R 2003 *Phys. Rev. Lett.* **90** 047205
[98] Deza M M and Laurent M 2009 *Geometry of Cuts and Metrics* (Berlin: Springer)
[99] Karp R M 1972 *Reducibility Among Combinatorial Problems Complexity of Computer Computations* (Berlin: Springer) pp 85–103
[100] Parisi G 1979 *Phys. Rev. Lett.* **43** 1754–6
[101] Parisi G 1980 A sequence of approximated solutions to the S-K model for spin glasses *J. Phys. A: Math. Gen.* **13** L115–21
[102] Laurent M and Poljak S 1995 On a positive semidefinite relaxation of the cut polytope **223–224** 439–61
[103] Montanari A and Sen S 2016 Semidefinite programs on sparse random graphs and their application to community detection *Proc. 48th Annual Symp. Theory of Computing* pp 814–27
[104] Goemans M X and Williamson D P 1995 *J. ACM* **42** 1115–45
[105] Khot S, Kindler G, Mossel E and O'Donnell R 2007 *SIAM J. Comput.* **37** 319–57
[106] Ghosh M, Jeronimo F, Jones C, Potechin A and Rajendran G 2020 Sum-of-squares lower bounds for Sherrington–Kirkpatrick via planted affine planes *Proc. 61st Annual Symp. Foundations of Computer Science (FOCS)* pp 954–65
[107] Isserlis L 1918 *Biometrika* **12** 134–9
[108] Kunisky D and Bandeira A S 2021 *Math. Program.* **190** 721–59
[109] Mohanty S, Raghavendra P and Xu J 2020 Lifting sum-of-squares lower bounds: degree-2 to degree-4 *Proc. 52nd Annual Symp. Theory of Computing* pp 840–53
[110] Kunisky D 2020 arXiv:2009.07269
[111] Barak B, Hopkins S, Kelner J, Kothari P K, Moitra A and Potechin A 2019 *SIAM J. Comput.* **48** 687–735

J. Stat. Mech. (2022) 114015

[112] Raghavendra P, Schramm T and Steurer D 2018 Statistical inference and the sum of squares method *High Dimensional Estimation via Sum-Of-Squares Proofs Proc. Int. Congress of Mathematicians* (Rio de Janeiro 2018) pp 3389–423
[113] Hopkins S 2018 *PhD Thesis* Cornell University
[114] Bandeira A S, Kunisky D and Wein A S 2020 Computational hardness of certifying bounds on constrained PCA problems *11th Innovations in Theoretical Computer Science Conf.* vol 151 (LIPIcs) pp 78:1–78:29
[115] Krzakala F and Zdeborová L 2009 *Phys. Rev. Lett.* **102** 238701
[116] Bandeira A S, Banks J, Kunisky D, Moore C and Wein A 2021 Spectral planting and the hardness of refuting cuts, colorability, and communities in random graphs *Proc. 34th Conf. Learning Theory* pp 410–73
[117] Banks J, Kleinberg R and Moore C 2019 *SIAM J. Comput.* **48** 1098–119
[118] Barak B, Kelner J A and Steurer D 2014 Rounding sum-of-squares relaxations *Proc. 46th Annual ACM Symp. Theory of Computing STOC '14* pp 31–40
[119] Barak B and Moitra A 2016 Noisy tensor completion via the sum-of-squares hierarchy *29th Annual Conf. Learning Theory* pp 417–45
[120] Hopkins S B, Shi J and Steurer D 2015 Tensor principal component analysis via sum-of-square proofs *Proc. 28th Conf. Learning Theory* pp 956–1006
[121] Richard E and Montanari A 2014 A statistical model for tensor PCA *Adv. Neural Inf. Process. Syst.* 2897–905
[122] Hopkins S B, Schramm T, Shi J and Steurer D 2016 Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors *Proc. 48th Annual ACM Symp. Theory of Computing* pp 178–91
[123] Hopkins S B, Kothari P K, Potechin A, Raghavendra P, Schramm T and Steurer D 2017 The power of sum-of-squares for detecting hidden structures *2017 IEEE 58th Annual Symp. Foundations of Computer Science (FOCS)* (IEEE) pp 720–31
[124] Anandkumar A, Ge R and Janzamin M 2017 *J. Mach. Learn. Res.* **18** 752–91
[125] Wein A S, Alaoui A E K and Moore C 2019 *Proc. 60th Annual Symp. Foundations of Computer Science* pp 1446–68
[126] Yedidia J, Freeman W and Weiss Y 2003 Understanding belief propagation and its generalizations *Exploring Artificial Intelligence in the New Millennium* vol 8 pp 236–9
[127] Kikuchi R 1951 *Phys. Rev.* **81** 988–1003
[128] Yedidia J S, Freeman W and Weiss Y 2000 Generalized belief propagation *Adv. Neural Inf. Process. Syst.*
[129] Yedidia J, Freeman W and Weiss Y 2001 Bethe free energy, Kikuchi approximations, and belief propagation algorithms *Advances in Neural Information Processing Systems* vol 13 p 689
[130] Krzakala F, Moore C, Mossel E, Neeman J, Sly A, Zdeborová L and Zhang P 2013 *Proc. Natl Acad. Sci. USA* **110** 20935–40
[131] Bordenave C, Lelarge M and Massoulie L 2015 Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs *Proc. 56th Annual Symp. Foundations of Computer Science* pp 1347–57
[132] Saade A, Krzakala F and Zdeborová L 2014 Spectral clustering of graphs with the Bethe hessian *Adv. Neural Inf. Process. Syst.* 406–14
[133] Bhattiprolu V, Guruswami V and Lee E 2017 Sum-of-squares certificates for maxima of random tensors on the sphere *Proc. APPROX/RANDOM* pp 31:1–20
[134] Raghavendra P, Rao S and Schramm T 2017 Strongly refuting random CSPs below the spectral threshold *Proc. 49th Annual Symp. Theory of Computing* pp 121–31
[135] Guruswami V, Kothari P K and Manohar P 2022 Algorithms and certificates for Boolean CSP refutation: smoothed is no harder than random *Proc. 54th Annual Symp. Theory of Computing* pp 678–89
[136] Bandeira A S, El Alaoui A, Hopkins S B, Schramm T, Wein A S and Zadik I 2022 (arXiv:2205.09727)

J. Stat. Mech. (2022) 114015