NOTES ON COMPUTATIONAL-TO-STATISTICAL GAPS: PREDICTIONS USING STATISTICAL PHYSICS

AFONSO S. BANDEIRA, AMELIA PERRY, AND ALEXANDER S. WEIN

In memory of Amelia Perry, and her love for learning and teaching.

Abstract. In these notes we describe heuristics to predict computational-to-statistical gaps in certain statistical problems. These are regimes in which the underlying statistical problem is information-theoretically possible although no efficient algorithm exists, rendering the problem essentially unsolvable for large instances. The methods we describe here are based on mature, albeit non-rigorous, tools from statistical physics.

These notes are based on a lecture series given by the authors at the Courant Institute of Mathematical Sciences in New York City, on May 16$^{\text{th}}$, 2017.

1. Introduction

Statistics has long studied how to recover information from data. Theoretical statistics is concerned with, in part, understanding under which circumstances such recovery is possible. Oftentimes recovery procedures amount to computational tasks to be performed on the data that may be computationally expensive, and so prohibitive for large datasets. While computer science, and in particular complexity theory, has focused on studying hardness of computational problems on worst-case instances, time and time again it is observed that computational tasks on data can often be solved far faster than the worst case complexity would suggest. This is not shocking; it is simply a manifestation of the fact that instances arising from real-world data are not adversarial. This illustrates, however, an important gap in fundamental knowledge: the understanding of **"computational hardness of statistical estimation problems"**.

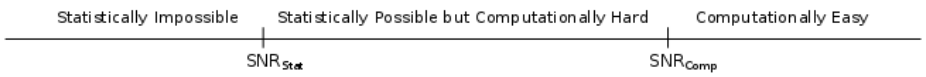
For concreteness we will focus on the case where we want to learn a set of parameters from samples of a distribution, or estimate a signal from noisy measurements (often two interpretations of the same problem). In the problems we will consider, there is a natural notion of signal-to-noise ratio (SNR) which can be related to the variance of the distribution of samples, the strength of the noise, the number of samples or measurements obtained, the size of a hidden planted structure buried in noise, etc. Two "phase transitions" are often studied. Theoretical statistics and information theory often study the critical SNR below which it is statistically impossible to estimate the parameters (or recover the signal, or find the hidden

ASB was partially supported by NSF DMS-1712730 and NSF DMS-1719545.

AP was supported in part by NSF CAREER Award CCF-1453261 and a grant from the MIT NEC Corporation.

ASW received Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

structure), and we call this threshold $\mathrm{SNR}_{\mathrm{Stat}}$. On the other hand, many algorithm development fields propose and analyze efficient algorithms to understand for which SNR levels different algorithms work. Despite significant effort to develop ever better algorithms, there are various problems for which no efficient algorithm is known to achieve recovery close to the statistical threshold $\mathrm{SNR}_{\mathrm{Stat}}$. Thus we are interested in the critical threshold $\mathrm{SNR}_{\mathrm{Comp}} \geq \mathrm{SNR}_{\mathrm{Stat}}$ below which it is fundamentally impossible for an efficient (polynomial time) algorithm to recover the information of interest.



There are many problems believed to exhibit computational-to-statistical gaps. Examples include community detection [HLL83, DKMZ11, AS15, BMNN16], planted clique [AKS98, DM15, BHK$^+$16], sparse principal component analysis [BR12, BR13, LKZ15b], structured spiked matrix models [LKZ15a, PWBM16b, KXZ16, BDM$^+$16, LM16], spiked tensor models [RM14, HSS15, PWB16, LML$^+$17, KBG17], and synchronization problems over groups [Sin11, PWBM16b, PWBM16a].

In these notes we will be concerned with predicting the locations of the thresholds $\mathrm{SNR}_{\mathrm{Stat}}$ and $\mathrm{SNR}_{\mathrm{Comp}}$ for Bayesian inference problems. In particular, we will focus on a couple of heuristics borrowed from statistical physics and illustrate them on two example problems: the Rademacher spiked Wigner problem (Example 2.1) and the related problem of community detection in the stochastic block model (Example 2.2). While we focus on these problems, we will try to cover the techniques in a way that conveys how they are broadly applicable.

At first glance, it may seem surprising that statistical physics has anything to do with Bayesian inference problems. The connection lies in the *Gibbs (or Boltzmann) distribution* that is widely used in statistical physics to model disordered systems such as magnets. It turns out that in many Bayesian inference problems, the posterior distribution of the unknown signal given the data also follows a Gibbs distribution, and thus many techniques from statistical physics can be applied. More specifically, many inference problems follow similar equations to *spin glasses*, which are physical systems in which the interaction strength between each pair of particles is random. The techniques that we borrow from statistical physics are largely non-rigorous but yield extremely precise predictions of both the statistical and computational limits. Furthermore, the predictions made by these heuristics have now been rigorously verified for many problems, and thus we have good reason to trust them on new problems. See the survey [ZK16] for more on the deep interplay between statistical physics and inference.

Many techniques have been developed in order to understand computational-to-statistical gaps. We now give a brief overview of some of these methods, both the ones we will cover in these notes and some that we will not.

Reductions. A natural approach to arguing that a task is computationally hard is via reductions, by showing that a problem is computationally hard conditioned on another problem being hard. This technique is extremely effective when studying the worst-case hardness of computational problems (a famous example being the list of 21 NP-hard combinatorial problems of Karp [Kar72]). There are also some remarkable successes in using this idea in the context of average-case problems (i.e. statistical inference on random models), starting with the work of Berthet and Rigollet on sparse PCA [BR12, BR13] and including also some conditional lower bounds for community detection with sublinear sized communities [MW13, HWX14]. These works show conditional hardness by reduction to the planted clique problem, which is widely believed to be hard in certain regimes. Unfortunately this method of reductions has so far been limited to problems that are fairly similar to planted clique.

Sum-of-squares hierarchy. Sum-of-squares [Las01, Par00, Nes00, Sho87, BS14] is a hierarchy of algorithms to approximate solutions of combinatorial problems, or more generally, polynomial optimization problems. For each positive integer d, the algorithm at level d of the hierarchy is a semidefinite program that relaxes the notion of a distribution over the solution space by only keeping track of moments of order $\leq d$. As you go up the hierarchy (increasing d), the algorithms get more powerful but also run slower: the runtime is $n^{O(d)}$. The level-2 relaxation coincides with the algorithms in the seminal work of Goemans and Williamson [GW95] and Lovasz [Lov79]. The celebrated unique games conjecture of Khot implies that the level-2 algorithm gives optimal worst-case approximation ratio for a wide class of problems [Kho02, Rag08, Kho10]. Sum-of-squares algorithms have also seen many success stories for average-case inference problems such as planted sparse vector [BKS13, HSSS16], dictionary learning [BKS15], tensor PCA [HSS15], tensor decomposition [BKS15, GM15, HSSS16, MSS16], and tensor completion [BM16, PS17]. One way to argue that an inference problem is hard is by showing that the sum-of-squares hierarchy fails to solve it at a particular level d (or ideally, at every constant value of d). Such lower bounds have been shown for many problems such as planted clique [BHK$^+$16] and tensor PCA [HSS15]. There is also recent work that gives evidence for computational hardness by relating the power of sum-of-squares to the low-degree moments of the posterior distribution [HS17].

Belief propagation, approximate message passing, and the cavity method. Another important heuristic to predict computational thresholds is based on ideas from statistical physics and is often referred to as the *cavity method* [MPV86]. It is based on analyzing an iterative algorithm called *belief propagation* (BP) [Pea86], or its close relative *approximate message passing* (AMP) [DMM09]. Specifically, BP has a trivial fixed point wherein the algorithm fails to perform inference. If this fixed point is stable (attracting) then we expect inference to be computationally hard. In these notes we will cover this method in detail. For further references, see [MM09, ZK16].

Replica method and the complexity of the posterior. Another method borrowed from statistical physics is the *replica method* (see e.g. [MM09]). This is a mysterious non-rigorous calculation from statistical physics that can produce many of the same predictions as the cavity method. One way to think about this method is as a way to measure the complexity of the posterior distribution. In particular, we

are interesting in whether the posterior distribution resembles one big connected region or whether it fractures into disconnected clusters (indicating computational hardness). We will cover the replica method in Section 5 of these notes.

Complexity of a random objective function. Another method for investigating computational hardness is through the lens of non-convex optimization. Intuitively, we expect that "easy" optimization problems have no "bad" local minima and so an algorithm such as gradient descent can find the global minimum (or at least a point whose objective value is close to the global optimum). For Bayesian inference problems, maximum likelihood estimation amounts to minimizing a particular random non-convex function. One tool to study critical points of random functions is the *Kac-Rice formula* (see [AT07] for an introduction). This has been used to study optimization landscapes in settings such as spin glasses [ABAČ13], tensor decomposition [GM17], and problems arising in community detection [BBV16]. There are also other methods to show that there are no spurious local minima in certain settings, e.g. [GJZ17, BVB16, LV18].

2. Setting and vocabulary

Throughout, we'll largely focus on Bayesian inference problems. Here we have a signal $\sigma^* \in \mathbb{R}^n$ viewed through some noisy observation model. We present two examples, and examine them simultaneously through the parallel language of machine learning and statistical physics.

Example 2.1 (Rademacher spiked Wigner)**.** The signal σ^* is drawn uniformly at random from $\{\pm 1\}^n$. We observe the $n \times n$ matrix

$$Y = \frac{\lambda}{n}\sigma^*(\sigma^*)^\top + \frac{1}{\sqrt{n}}W,$$

where λ is a signal-to-noise parameter, and W is a GOE matrix[1]. We wish to approximately recover σ^* from Y, up to a global negation (since σ^* and $-\sigma^*$ are indistinguishable).

This problem is motivated by the statistical study of the *spiked Wigner model* from random matrix theory (see e.g. [PWBM16b]). This model has also been studied as a Gaussian variant of community detection [DAM16] and as a model for synchronization over the group $\mathbb{Z}/2$ [JMRT16].

Example 2.2 (Stochastic block model)**.** The signal σ^* is drawn uniformly at random from $\{\pm 1\}^n$. We observe a graph G with vertex set $[n] = \{1, \dots, n\}$, with edges drawn independently as follows: for vertices u, v, we have $u \sim v$ with probability a/n if $\sigma_u\sigma_v = 1$, and probability b/n if $\sigma_u\sigma_v = -1$. We will restrict ourselves to the case $a > b$. We imagine the entries σ^*_u as indicating membership of vertex u in either the $+1$ or -1 'community'; thus vertices in the same community are more likely to share an edge. We wish to approximately recover the community structure σ^* (up to global negation) from G.

This is a popular model for community detection in graphs. See e.g. [Abb17, Moo17] for a survey. Here we consider the sparse regime, but other regimes are also considered in the literature.

[1]Gaussian orthogonal ensemble: symmetric with the upper triangle drawn i.i.d. as $\mathcal{N}(0, 1)$.

There is a key difference between the two models above. The Rademacher spiked Wigner model is *dense* in the sense that we are given an observation for every pair of variables. On the other hand, the stochastic block model is *sparse* (at least in the regime we have chosen) because essentially all the useful information comes from the observed edges, which form a sparse graph. We will see that different tools are needed for dense and sparse problems.

2.1. **Machine learning view.** We are interested in inferring the signal σ^*, so it is natural to write down the posterior distribution. For the Rademacher spiked Wigner problem (Example 2.1), we can compute the posterior distribution explicitly as follows:

$$\begin{aligned}\Pr[\sigma \mid Y] \propto \Pr[Y \mid \sigma] &\propto \prod_{i<j} \exp\left(-\frac{n}{2}\left(Y_{ij} - \frac{\lambda}{n}\sigma_i\sigma_j\right)^2\right) \\ &= \prod_{i<j} \exp\left(-\frac{n}{2}Y_{ij}^2 + \lambda Y_{ij}\sigma_i\sigma_j - \frac{\lambda^2}{2n}\sigma_i^2\sigma_j^2\right) \\ &\propto \prod_{i<j} \exp\left(\lambda Y_{ij}\sigma_i\sigma_j\right).\end{aligned}$$

(Here $\propto$ hides a normalizing constant which depends on Y but not σ; it is chosen so that $\sum_{\sigma\in\{\pm1\}^n} \Pr[\sigma \mid Y] = 1$.) The above factorization over edges defines a **graphical model**: a probability distribution factoring in the form $\Pr[\sigma] = \prod_{S\subset[n]} \psi_S(\sigma_S)$ into **potentials** ψ_S that each only depend on a small (constant-size) subset S of the entries of σ. (For instance, in our example above, S ranges over all subsets of size 2.)

2.2. **Statistical physics view.** The observation Y defines a **Hamiltonian**, or energy function, $H(\sigma) = \sum_{i<j} Y_{ij}\sigma_i\sigma_j$, consisting of two-spin interactions; we refer to each entry of σ as a **spin**, and to σ as a **state**. A Hamiltonian together with a parameter $T = \frac{1}{\beta}$, called the **temperature**, defines a **Gibbs distribution** (or **Boltzmann distribution**):

$$\Pr[\sigma] \propto e^{-\beta H(\sigma)}.$$

Thus low-energy states are more likely than high-energy states; moreover, at low temperature (large β), the distribution becomes more concentrated on lower energy states, becoming supported entirely on the minimum energy states (**ground states**) in the limit as $\beta \to \infty$. On the other hand, in the high-temperature limit ($\beta \to 0$), the Gibbs distribution becomes uniform.

Connecting the ML and physics languages, we observe that the posterior distribution on σ is precisely the above Gibbs distribution, at the particular inverse-temperature $\beta = \lambda$. (This is often referred to as lying on the **Nishimori line** [Nis80, Nis81, Nis01], or being at Bayes-optimal temperature.)

2.3. **Optimization and statistical physics.** A common optimization viewpoint on inference is **maximum likelihood estimation**, or the maximization task of finding the state σ that maximizes the posterior likelihood. This optimization problem is frequently computationally hard, but convex relaxations or surrogates may be studied. To rephraze this optimization task in physical terms, we wish to

minimize the energy $H(\sigma)$ over states σ, or equivalently sample from (or otherwise describe) the low-temperature Gibbs distribution in the limit $\beta \to \infty$.

This viewpoint is limited, in that the MLE frequently lacks any *a priori* guarantee of optimality. On the other hand, the Gibbs distribution at the true temperature $\beta = \lambda$ enjoys optimality guarantees at a high level of generality:

Claim 2.3. *Suppose we are given some observation Y leading to a posterior distribution on σ. For any estimate $\widehat{\sigma} = \widehat{\sigma}(Y)$, define the (expected) mean squared error (MSE) $\mathbb{E}\|\widehat{\sigma} - \sigma\|_2^2$. The estimator that minimizes the expected MSE is given by $\widehat{\sigma} = \mathbb{E}[\sigma \mid Y]$, the posterior expectation (and thus the expectation under the Gibbs distribution at Bayes-optimal temperature).*

Remark 2.4. In the case of the Rademacher spiked Wigner model, there is a caveat here: since σ^* and $-\sigma^*$ are indistinguishable, the posterior expectation is zero. Our objective is not to minimize the MSE but to minimize the error between σ and either $\widehat{\sigma}$ or $-\widehat{\sigma}$ (whichever is better).

Thus the optimization approach of maximum likelihood estimation aims for too low a temperature. aggregate likelihood. Intuitively, MLE searches for the single state with highest individual likelihood, whereas the optimal Bayesian approach looks for a large cluster of closely-related states with a high aggregate likelihood.

Fortunately, the true Gibbs distribution has an optimization property of its own:

Claim 2.5. *The Gibbs distribution with Hamiltonian H and temperature $T > 0$ is the unique distribution minimizing the* **(Helmholtz) free energy**

$$F = \mathbb{E}H - TS,$$

where S denotes the Shannon entropy $S = -\mathbb{E}_\sigma \log \Pr(\sigma)$.

Proof. Entropy is concave with infinite derivative at the edge of the probability simplex, and the expected Hamiltonian is linear in the distribution, so the free energy is convex and minimized in the interior of the simplex. We find the unique local (hence global) minimum with a Lagrange multiplier:

$$\begin{aligned} \text{const} \cdot \mathbb{1} = \nabla F &= \nabla \sum_\sigma p(\sigma)(H(\sigma) + T \log p(\sigma)) \\ \text{const} &= H(\sigma) + T \log p(\sigma) \\ \text{const} \cdot e^{-H(\sigma)/T} &= p(\sigma), \end{aligned}$$

which we recognize as the Gibbs distribution. □

This optimization approach is willing to trade off some energy for an increase in entropy, and can thus detect large clusters of states with a high aggregate likelihood, even when no individual state has the highest possible likelihood. Moreover, the free energy is convex, but it is a function of an arbitrary probability distribution on the state space, which is typically an exponentially large object.

We are thus led to ask the question: is there any way to reduce the problem of free energy minimization to a tractable, polynomial-size problem? Can we get a theoretical or algorithmic handle on this problem?

3. The cavity method and belief propagation

3.1. **BP as an algorithm for inference. Belief propagation** (BP) is a general algorithm for inference in graphical models, generally credited to Pearl [Pea86] (see e.g. [MM09] for a reference). As we've seen above, the study of graphical models is essentially the statistical physics of Hamiltonians consisting of interactions that each only depend on a few spins. Quite often, we care about the **average case** study of random graphical models that describe a posterior distribution given some noisy observation of a signal, such as in the Rademacher spiked Wigner example discussed above. Much of statistical physics is concerned with disorder and random systems, and indeed the concept of belief propagation appeared in physics as the **cavity method**—not only as an algorithm but as a theoretical means to make predictions about systems such as spin glasses [MPV86].

To simplify the setting and notation, let us consider sparse graphical models with only pairwise interactions:

$$\Pr[\sigma] \propto \prod_{u \sim v} \psi_{uv}(\sigma_u, \sigma_v),$$

where each vertex v has only relatively few "neighbors" u (denoted $u \sim v$).

Belief propagation is an iterative algorithm. We think of each spin σ_u as a vertex and each pair of neighbors as an edge. Each vertex tracks a "belief" about its own spin (more formally, an estimated posterior marginal). These beliefs are often initialized to something like a prior distribution, or just random noise, and then iteratively refined to become more consistent with the graphical model. This refinement happens as follows: each vertex u transmits its belief to each neighbor, and then each vertex updates its belief based on the incoming beliefs of its neighbors. If we let $m_{u \to v}$ denote the previous beliefs sent from neighbors u to a vertex v, we can formulate a new belief for v in a Bayesian way, assuming that the incoming influences of the neighbor vertices are independent (more on this assumption below):

$$m_v(\sigma_v) \propto \prod_{u \sim v} \sum_{\sigma_u} \psi_{uv}(\sigma_u, \sigma_v) m_{u \to v}(\sigma_u)$$

Each message $m_{u \to v}$ is a probability distribution (over the possible values for σ_u), with the proportionality constant being determined by probabilities summing to 1 over all values of σ_u.

This is almost a full description of belief propagation, except for one detail. If the belief from vertex v at time $t-2$ influences the belief of neighbor u at time $t-1$, then neighbor u should not parrot that influence back to neighbor v, reinforcing its belief at time t without any new evidence. Thus we ensure that the propagation of messages does not immediately backtrack:

$$m^{(t)}_{v \to w}(\sigma_v) \propto \prod_{\substack{u \sim v \\ u \neq w}} \sum_{\sigma_u} \psi_{uv}(\sigma_u, \sigma_v) m^{(t-1)}_{u \to v}(\sigma_u). \tag{1}$$

This formula is the iteration rule for belief propagation.

The most suspicious aspect of the discussion above is the idea that neighbors of a vertex v exert probabilistically independent influences on v. If the graphical model is a tree, then the neighbors are independent after conditioning on v, and in this setting it is a theorem (see e.g. [MM09]) that belief propagation converges to the exact posterior marginals. On a general graphical model, this independence fails, and belief propagation is heuristic. In many sparse graph models, neighborhoods of

most vertices are trees, with most loops being long, so that independence might approximately hold. BP certainly fails in the worst case; outside of special cases such as trees it is certainly only suitable in an average-case setting. However, on many families of random graphical models, belief propagation is a remarkably strong approach; it is general, efficient, and often yields a state-of-the-art statistical estimate. It is conjectured in many models that belief propagation achieves asymptotically optimal inference, either among all estimators or among all polynomial-time estimators, but most rigorous results in this direction are yet to be established.

To connect to the previous viewpoint of free energy minimization: belief propagation is intimately connected with the **Bethe free energy**, a heuristic proxy for the free energy which may be described in terms of the messages $m_{u\to v}$ (see [ZK16], Section III.B). It can be shown that the fixed points of BP are precisely the critical points of the Bethe free energy, justifying the view that BP is roughly a minimization procedure for the free energy. Again, rigorously the situation is much worse: the Bethe free energy is non-convex, and BP is not guaranteed to converge, let alone guaranteed to find the global minimum.

3.2. **The cavity method for the stochastic block model.** The ideas of belief propagation above appear as the **cavity method** in statistical physics, owing to the idea that the Bethe free energy is believed to be essentially an accurate model for the true (Helmholtz) free energy on a variety of models of interest. In passing to the Bethe free energy, we can pass from studying a general distribution (an exponentially complicated object) to studying node and edge marginals, which are theoretically much simpler objects and, crucially, can be studied locally on the graph. Local neighborhoods of sparse graphs as in the SBM (stochastic block model) look like trees, and so we are drawn to studying message passing on a tree.

Much as for the Rademacher spiked Wigner model above, we derive a Hamiltonian from the block model posterior:

$$H(\sigma) = \sum_{i\sim j} \theta_+ \sigma_i \sigma_j + \sum_{i \not\sim j} \theta_- \sigma_i \sigma_j,$$

where $u \sim v$ denotes adjacency in the observed graph, and $\theta_+ > 0 > \theta_-$ are constants depending on a and b; θ_+ is of constant order, while θ_- is of order $1/n$. In expressing belief propagation, we will make a small notational simplification: instead of passing messages m that are distributions over $\{+,-\}$, it suffices to pass the expectation $m(+) - m(-)$. The reader can verify that rewriting the belief propagation equations in this notation yields

$$m_{u\to v}^{(t)} = \tanh\left(\sum_{\substack{w\sim u \\ w\neq v}} \operatorname{atanh}(\theta_+ m_{w\to u}^{(t-1)}) + \sum_{\substack{w\not\sim u \\ w\neq v}} \operatorname{atanh}(\theta_- m_{w\to u}^{(t-1)}) \right)$$

where tanh is the hyperbolic tangent function $\tanh(z) = (e^z - e^{-z})/(e^z + e^{-z})$, and atanh is its inverse.

The first term inside the tanh represents strong, constant-order attractions with the few graph neighbors, while the second term represents very weak, low-order repulsions with the multitude of non-neighbors. The value of the second term thus depends very little on any individual spin, but rather on the overall balance of positive and negative spins in the graph, with the tendency to cause the global

spin configuration to become balanced. As we are only interested in a local view of message passing, we will assume here that the global configuration is roughly balanced and neglect the second term:

$$m_{u\to v}^{(t)} \approx \tanh\left(\sum_{\substack{w\sim u\\ w\neq v}} \operatorname{atanh}(\theta_+ m_{w\to u}^{(t-1)})\right).$$

As this message-passing only involves the graph edges, it now makes sense to study this on a tree-like neighborhood. We now discuss a generative model for (approximate) local neighborhoods under the stochastic block model.

Model 3.1 (Galton–Watson tree)**.** Begin with a root vertex, with spin + or − chosen uniformly. Recursively, each vertex gives birth to a Poisson number of child nodes: $\mathrm{Pois}((1-\varepsilon)k)$ vertices of the same spin and $\mathrm{Pois}(\varepsilon k)$ vertices of opposite spin, up to a total tree depth of d.

As shown in [MNS12], the Galton–Watson tree with $k = (a+b)/2$ and $\varepsilon = b/(a+b)$ is distributionally very close to the radius-d neighborhood of a vertex in the SBM with its true spins, so long as $d = o(\log n)$. Thus we will study belief propagation on a random Galton–Watson tree.

Let us consider only the BP messages passing toward the root of the tree. The upward message from any vertex v is computed as:

$$m_v = \tanh\left(\sum_u \operatorname{atanh}((1-2\varepsilon)m_u)\right) \tag{2}$$

where u ranges over the children of v. We now imagine that the child messages m_u are independently drawn from some distribution $D_+^{(t-1)}$ for children with spin +, and (leveraging symmetry) from the distribution $D_-^{(t-1)} = -D_+^{(t-1)}$ for children with spin −; this distribution represents the randomness of our BP calculation below each child, over the random subtree hanging off each one. Then, from equation (2), together with the fact that there are $\mathrm{Pois}((1-\varepsilon)k)$ same-spin children and $\mathrm{Pois}(\varepsilon k)$ opposite-spin children, the distribution $D_\pm^{(t)}$ of the parent message m is determined! Thus we obtain a distributional recurrence for $D_+^{(t)}$.

The calculation above is independent of n, and the radius of validity of the tree approximation grows with n, so we are interested in the behavior of the recurrence above as $t\to\infty$, i.e. fixed points of the distributional recurrence above and their stability.

Typically one initializes BP with small random messages, a perturbation of the trivial all-0 fixed point that represents our prior. For small messages, we can linearize tanh and atanh, and write $m_v \approx (1-2\varepsilon)\sum_u m_u$. Then if the child distribution $D_+^{(t-1)}$ has mean μ and variance σ^2, it is easily computed that the parent distribution $D_+^{(t)}$ has mean $k(1-2\varepsilon)^2\mu$ and variance $k(1-2\varepsilon)^2\sigma^2$. Thus if $k(1-2\varepsilon)^2 < 1$, then perturbations of the all-0 fixed point decay, or in other words, this fixed point is stable, and BP is totally uninformative on this typical initialization. If $k(1-2\varepsilon)^2 > 1$, then small perturbations do become magnified under BP dynamics, and one imagines that BP might find a more informative fixed point (though this remains an open question!).

This transition is known as the **Kesten–Stigum threshold** [KS66], and calculations of this form are loosely conjectured to describe the **computational threshold** beyond which no efficient algorithm can perform inference, for sparse models such as the SBM. In this particular form of the SBM, this idea has been rigorously vindicated using techniques slightly different from BP: inference is known to be statistically impossible when $k(1-2\varepsilon)^2 < 1$ (meaning that any estimator has zero correlation with the truth as $n \to \infty$) [MNS12], and efficiently possible when $k(1-2\varepsilon)^2 > 1$ (meaning that asymptotically nonzero correlation is possible) [Mas14, MNS13].

One might also endeavor to study the other fixed points of BP, not just the trivial fixed points. This is a difficult undertaking in most situations, as the BP recurrence lacks convexity properties, but it is expected to give an understanding of the **statistical threshold** of the problem, i.e. the limit below which even inefficient inference techniques fail. This has been rigorously proven for some variants of the stochastic block model [COKPZ16]. Intuitively, exploring the BP landscape by brute force for the best (in terms of Bethe free energy) BP fixed point is a statistically optimal inference technique. For more general stochastic block models with 4 or more communities, there exists a gap between the statistical threshold and the analogous Kesten–Stigum bound [DKMZ11, AS15, BMNN16].

4. Approximate message passing

4.1. **AMP as a simplification of BP.** Our cavity analysis of the block model above was well-adapted to sparse models, in which the analysis localizes onto a tree of constant average degree. But many models, such as the Rademacher spiked Wigner model, are dense and their analysis cannot be local. Thankfully, many of these models are amenable to analysis for different reasons: as each vertex is acted on by a large number of individually weak influences, the quantities of interest in belief propagation are subject to central limit theorems and concentration of measure. In this section we will demonstrate this on the Rademacher spiked Wigner example.

Recall the Hamiltonian $H = \sum_{i<j} Y_{ij}\sigma_i\sigma_j$ and inverse temperature λ. As in the SBM discussion above, we can summarize BP messages by the expectation $m(+) - m(-)$. Then BP for this model reads as

$$m_{u\to v}^{(t)} = \tanh\left(\sum_{w\neq v} \operatorname{atanh}(\lambda Y_{wu} m_{w\to u}^{(t-1)})\right).$$

We next exploit the weakness of individual interactions. Note that the values $m_{w\to u}^{(t-1)}$ lie in $[-1,1]$, while Y_{wu} is of order $n^{-1/2}$ in probability. Taylor-expanding atanh, we simplify:

$$m_{u\to v}^{(t)} = \tanh\left(\left(\sum_{w\neq v} \lambda Y_{wu} m_{w\to u}^{(t-1)}\right) + O(n^{-1/2})\right) \quad \text{w.h.p.}$$

We next simplify the non-backtracking nature of BP. Naïvely, one might expect that we can simply drop the condition $w \neq v$ from the sum above, as the contribution from vertex v in the above sum should be only of size $n^{-1/2}$. As our formula for $m_{u\to v}^{(t)}$ would then no longer depend on v, we could write down messages indexed

by a single vertex:

$$m_u^{(t)} = \tanh\left(\sum_w \lambda Y_{wu} m_w^{(t-1)}\right),$$

or in vector notation,

$$m^{(t)} = \tanh(\lambda Y m^{(t-1)}), \tag{3}$$

where tanh applies entrywise. This resembles the "power iteration" iterative algorithm to compute the leading eigenvector of Y:

$$m^{(t)} = Y m^{(t-1)},$$

but with $\tanh(\lambda \bullet)$ providing some form of soft projection onto the interval $[-1, 1]$, exploiting the entrywise ± 1 structure.

Unfortunately, the non-backtracking simplification above is flawed, and equation (3) does not accurately summarize BP or provide as strong an estimator. The problem is that the terms we have neglected add up constructively over two iterations. Specifically: consider that vertex v exerts an influence $\lambda Y_{vu} m_v^{(t-2)}$ on each neighbor u; this small perturbation translates directly to a perturbation of $m_u^{(t-1)}$ (scaled by a derivative of tanh). At the next iteration, vertex u influences $m_v^{(t)}$ according to $\lambda Y_{vu} m_u^{(t-1)}$; the total contribution from backtracking here is thus $\lambda^2 Y_{vu}^2 m_v^{(t-2)}$, scaled through some derivatives of tanh. This influence is a random, positive, order $1/n$ multiple of $m_v^{(t-2)}$. Summing over all neighbors u, we realize that the aggregate contribution of backtracking over two steps is in fact of order 1.

Thankfully, this contribution is also a sum of small random variables, and exhibits concentration of measure. The solution is thus to subtract off this aggregate backtracking term in expectation, adding a correction called the **Onsager reaction term**:

$$m^{(t)} = \tanh\left(Y m^{(t-1)} - \lambda^2 (1 - \|m\|_2^2/n) m^{(t-2)}\right). \tag{4}$$

This iterative algorithm is known as **approximate message passing** (AMP). The simplifications above to BP first appeared in the work of Thouless, Anderson, and Palmer [TAP77], who used it to obtain a theoretical handle on spin glasses at high temperature. The first AMP algorithm [DMM09] appeared in the context of compressed sensing. The AMP algorithm (4) for this problem can be found in [DAM16], and AMP has been applied to many other problems such as rank-one matrix estimation [FR12], sparse PCA [DM14], non-negative PCA [MR16], planted clique [DM15], and synchronization over groups [PWBM16a] (just to name a few).

4.2. **AMP state evolution.** In contrast to belief propagation, approximate message passing (AMP) algorithms tend to be amenable to exact analysis in the limit $n \to \infty$. Here we introduce *state evolution*, a simple heuristic argument for the analysis of AMP that has been proven correct in many settings. The idea of state evolution was first introduced by [DMM09], based on ideas from [Bol12]; it was later proved correct in various settings [BM11, JM13].

We will focus again on the Rademacher spiked Wigner model: we observe

$$Y = \frac{\lambda}{n} x x^\top + \frac{1}{\sqrt{n}} W$$

where $x \in \{\pm 1\}^n$ is the true signal (drawn uniformly at random) and the $n \times n$ noise matrix W is symmetric with the upper triangle drawn i.i.d. as $\mathcal{N}(0,1)$. In this setting, the AMP algorithm and its analysis are due to [DAM16].

We have seen above that the AMP algorithm for this problem takes the form

$$v^{t+1} = Yf(v^t) + [Onsager]$$

where $f(v)$ denotes entrywise application of the function $f(v) = \tanh(\lambda v)$. (Here we abuse notation and let f refer to both the scalar function and its entrywise application to a vector.) The superscript t indexes timesteps of the algorithm (and is not to be confused with an exponent). The details of the Onsager term, discussed previously, will not be important here.

The state evolution heuristic proceeds as follows. Postulate that at timestep t, AMP's iterate v^t is distributed as

$$v^t = \mu_t x + \sigma_t g \quad \text{where } g \sim \mathcal{N}(0, I). \tag{5}$$

This breaks down v^t into a signal term (recall x is the true signal) and a noise term, whose sizes are determined by parameters $\mu_t \in \mathbb{R}$ and $\sigma_t \in \mathbb{R}_{\geq 0}$. The idea of state evolution is to write down a recurrence for how the parameters μ_t and σ_t evolve from one timestep to the next. In performing this calculation we will make two simplifying assumptions that will be justified later: (1) we drop the Onsager term, and (2) we assume the noise W is independent at each timestep (i.e. there is no correlation between W and the noise g in the current iterate). Under these assumptions we have

$$\begin{aligned} v^{t+1} = Yf(v^t) &= \left(\frac{\lambda}{n} xx^\top + \frac{1}{\sqrt{n}} W\right) f(v^t) \\ &= \frac{\lambda}{n} \langle x, f(v^t)\rangle \, x + \frac{1}{\sqrt{n}} W f(v^t) \end{aligned}$$

which takes the form of (5) with a signal term and a noise term. We therefore have

$$\begin{aligned} \mu_{t+1} = \frac{\lambda}{n}\langle x, f(v^t)\rangle &= \frac{\lambda}{n}\langle x, f(\mu_t x + \sigma_t g)\rangle \\ &\approx \lambda \mathop{\mathbb{E}}_{X,G}[X f(\mu_t X + \sigma_t G)] \quad \text{with scalars } X \sim \mathrm{Unif}\{\pm 1\}, G \sim \mathcal{N}(0,1) \\ &= \lambda \mathop{\mathbb{E}}_{G}[f(\mu_t + \sigma_t G)] \quad \text{since } f(-v) = -f(v). \end{aligned}$$

For the noise term, think of $f(v^t)$ as fixed and consider the randomness over W. Each entry of the noise term $\frac{1}{\sqrt{n}} W f(v^t)$ has mean zero and variance

$$\begin{aligned} (\sigma^{t+1})^2 = \sum_i \frac{1}{n} f(v_i^t)^2 &= \sum_i \frac{1}{n} f(\mu_t x_i + \sigma_t g_i)^2 \\ &\approx \lambda \mathop{\mathbb{E}}_{X,G}[f(\mu_t X + \sigma_t G)^2] \quad \text{with scalars } X, G \text{ as above} \\ &= \mathop{\mathbb{E}}_{G}[f(\mu_t + \sigma_t G)^2] \quad \text{again by symmetry of } f. \end{aligned}$$

We now have "state evolution" equations for μ_{t+1} and σ_{t+1} in terms of μ_t and σ_t. Since we could arbitrarily scale our iterates v^t without adding or losing information, we really only care about the parameter $\gamma \triangleq (\mu/\sigma)^2$. It is possible (see [DAM16]) to reduce the state evolution recurrence to this single parameter:

$$\gamma_{t+1} = \lambda^2 \mathop{\mathbb{E}}_{G \sim \mathcal{N}(0,1)} \tanh(\gamma_t + \sqrt{\gamma_t}\, G) \tag{6}$$

(where we have substituted the actual expression for f).

We can analyze AMP as follows. Choose a small positive initial value γ_0 and iterate (6) until we reach a fixed point γ_∞. We then expect the output v^∞ of AMP to behave like

$$v^\infty = \mu_\infty x + \sigma_\infty g \tag{7}$$

where $g \sim \mathcal{N}(0, I)$, $\mu_\infty = \gamma_\infty/\lambda$, and $\sigma_\infty^2 = \gamma_\infty/\lambda^2$. For the Rademacher spiked Wigner model, this has in fact been proven to be correct in the limit $n \to \infty$ [BM11, JM13]. Namely, when we run AMP (with the Onsager term and without fresh noise W at each timestep), the output behaves like (7) in a particular formal sense.

State evolution reveals a phase transition at $\lambda = 1$: when $\lambda \leq 1$ we have $\gamma_\infty = 0$ (so AMP has zero correlation with the truth as $n \to \infty$) and when $\lambda > 1$ we have $\gamma_\infty > 0$ (so AMP achieves nontrivial correlation with the truth). Furthermore, from (7) we can deduce the value of any performance metric (e.g. mean squared error) at any signal-to-noise ratio λ. It has in fact been shown (for Rademacher spiked Wigner) that the mean squared error achieved by AMP is information-theoretically optimal [DAM16].

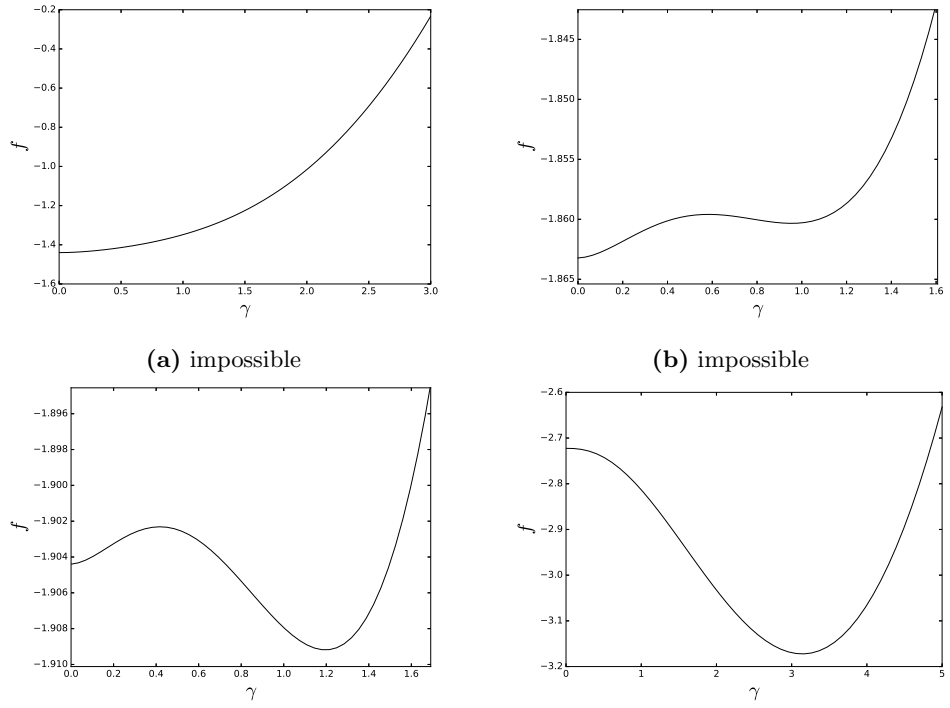
It is perhaps surprising that state evolution is correct, given the seemingly-questionable assumptions we made in deriving it. This can be understood as follows. Recall that we eliminated the Onsager term and assumed independent noise W at each timestep. Also recall that the Onsager term is a correction that makes the update step non-backtracking: a message sent across an edge at one iteration does not affect the message sent back across the edge (in the opposite direction) at the next iteration. It turns out that to leading order, using fresh noise at each timestep is equivalent to using a non-backtracking update step. This is because the largest effect of fresh noise is to terms where a particular noise entry W_{ij} is used twice in a row, i.e. backtracking steps. So the two assumptions we made actually cancel each other out! Note that both of the two assumptions are crucial in making the state evolution analysis tractable, so it is quite spectacular that we are able to make both of these assumptions for free (and still get the correct answer)!

One caveat in the rigorous analysis of AMP is that it assumes an initialization that has some nonzero correlation with the truth [DAM16]. In other words, we need to assume that we start with some nonzero γ because if we start with $\gamma = 0$ we will remain there forever. In practice this is not an issue; a small random initialization suffices.

4.3. **Free energy diagrams.** In this section we will finally see how to predict computational-to-statistical gaps (for dense problems)! Above we have seen how to analyze a particular algorithm: AMP. In various settings it has been shown that AMP is information-theoretically optimal. More generally, it is believed that AMP is optimal among all efficient algorithms (for a wide class of problems). We will now show how to use AMP to predict whether a problem should be easy, (computationally) hard or (statistically) impossible. The ideas here originate from [LKZ15a, LKZ15b].

Recall that the state of AMP is described by a parameter γ, where larger γ indicates better correlation with the truth and $\gamma = 0$ means that AMP achieves zero correlation with the truth. Also recall that the *Bethe free energy* is the quantity that belief propagation (or AMP) is locally trying to minimize. It is possible to

analytically write down the function $f(\gamma)$ which gives the (Bethe) free energy of the AMP state corresponding to γ; in the next section, we will see one way to compute $f(\gamma)$. AMP can be seen as starting near $\gamma = 0$ and naively moving in the direction of lowest free energy until it reaches a local minimum; the γ value at this minimum characterizes AMP's output. The information-theoretically optimal estimator is instead described by the global minimum of the free energy (and this has been proven rigorously in various cases [BDM$^+$16, LM16]); this corresponds to the inefficient algorithm that uses exhaustive search to find the AMP state which globally minimizes free energy. Figure 1 illustrates how the free energy landscape $f(\gamma)$ dictates whether the problem is easy, hard, or impossible at a particular λ value.



(a) impossible **(b)** impossible

(c) hard **(d)** easy

FIGURE 1. **(a)** The global minimizer is $\gamma = 0$ so no estimator achieves nontrivial recovery. **(b)** A new local minimum in the free energy has appeared, but the global minimum is still at $\gamma = 0$ and so nontrivial recovery remains impossible. **(c)** AMP is stuck at $\gamma = 0$ but the (inefficient) statistically optimal estimator achieves a nontrivial γ (the global minimum). AMP is not statistically optimal. **(d)** AMP achieves nontrivial (in fact optimal) recovery. The above image is adapted from [PWBM16a] (used with permission).

For Rademacher spiked Wigner, we have phase (a) (from Figure 1) when $\lambda \leq 1$ and phase (d) when $\lambda > 1$, so there is no computational-to-statistical gap. However, for some variants of the problem (for instance if the signal x is sparse, i.e. only a

small constant fraction of entries are nonzero) then we see phases (a),(b),(c),(d) appear in that order as λ increases; in particular, there is a computational-to-statistical gap during the hard phase (c).

Although many parts of this picture have been made rigorous in certain cases, the one piece that we do not have the tools to prove is that no efficient algorithm can succeed during the hard phase (c). This is merely conjectured based on the belief that AMP should be optimal among efficient algorithms.

There are a few different ways to compute the free energy landscape $f(\gamma)$. One method is to use the replica method discussed in the next section. Alternatively, there is a direct formula for Bethe free energy in terms of the BP messages, which can be adapted to AMP (see e.g. [LKZ15a]).

5. The replica method

The replica method is an alternative viewpoint that can derive many of the same results shown in the previous section. We will again use the example of Rademacher spiked Wigner to illustrate it. A general introduction to the replica method can be found in [MM09]. The calculations of this section are carried out in somewhat higher generality in Appendix B of [PWB16].

Recall again the setup: we observe $Y = \frac{\lambda}{n} xx^\top + \frac{1}{\sqrt{n}} W$ with $x \in \{\pm 1\}^n$ and $W_{ij} = W_{ji} \sim \mathcal{N}(0,1)$.

The posterior distribution of x given Y is

$$\Pr[x \,|\, Y] \propto \prod_{i<j} \exp\left(-\frac{n}{2}\left(\frac{\lambda}{n} x_i x_j - Y_{ij}\right)^2\right) \propto \exp\left(\lambda \sum_{i<j} Y_{ij} x_i x_j\right)$$

and so we are interested in the Gibbs distribution over $\sigma \in \{\pm 1\}^n$ given by $\Pr[\sigma \,|\, Y] \propto \exp(-\beta H(\sigma))$ with energy (Hamiltonian) $H(\sigma) = -\sum_{i<j} Y_{ij}\sigma_i\sigma_j$ and inverse temperature $\beta = \lambda$.

The goal is to compute the free energy density, defined as $f = -\frac{1}{\beta n}\mathbb{E}\log Z$ where

$$Z = \sum_{\sigma \in \{\pm 1\}^n} \exp(-\beta H(\sigma)).$$

(This can be shown to coincide with the notion of free energy introduced earlier.) The idea of the replica method is to compute the moments $\mathbb{E}[Z^r]$ of Z for $r \in \mathbb{N}$ and perform the (non-rigorous) analytic continuation

$$\mathbb{E}[\log Z] = \lim_{r \to 0} \frac{1}{r} \log \mathbb{E}[Z^r]. \tag{8}$$

Note that this is quite bizarre – we at first assume r is a positive integer, but then take the limit as r tends to zero! This will require writing $\mathbb{E}[Z^r]$ in an analytic form that is defined for all values of r. An informal justification for the correctness of (8) is that when r is close to 0, Z^r is close to 1 and so we can interchange log and $\mathbb{E}$ on the right-hand side.

The moment $\mathbb{E}[Z^r]$ can be expanded in terms of r 'replicas' $\sigma^1, \dots, \sigma^r$ with $\sigma^a \in \{\pm 1\}^n$:

$$\mathbb{E}[Z^r] = \sum_{\{\sigma^a\}} \mathbb{E} \exp\left(\beta \sum_{i<j} Y_{ij} \sum_{a=1}^{r} \sigma_i^a \sigma_j^a\right).$$

After applying the definition of Y and the Gaussian moment-generating function (to compute expectation over the noise W) we arrive at

$$\mathbb{E}[Z^r] = \sum_{\{\sigma^a\}} \exp\left[n\left(\frac{\lambda^2}{2}\sum_a c_a^2 + \frac{\lambda^2}{4}\sum_{a,b} q_{ab}^2\right)\right]$$

where $q_{ab} = \frac{1}{n}\sum_i \sigma_i^a \sigma_i^b$ is the correlation between replicas a and b, and $c_a = \frac{1}{n}\sum_i \sigma_i^a x_i$ is the correlation between replica a and the truth.

Without loss of generality we can assume (by symmetry) the true spike is $x = \mathbb{1}$ (all-ones). Let Q be the $(r+1)\times(r+1)$ matrix of overlaps (q_{ab} and c_a), including x as the zeroth replica. Note that Q is the average of n i.i.d. matrices and so by the theory of large deviations (Cramér's Theorem in multiple dimensions), the number of configurations $\{\sigma^a\}$ corresponding to given overlap parameters q_{ab}, c_a is asymptotically

(9)
$$\inf_{\mu,\nu} \exp\left[n\left(-\sum_a \nu_a c_a - \frac{1}{2}\sum_{a\neq b}\mu_{ab}q_{ab} + \log \sum_{\sigma\in\{\pm 1\}^r} \exp\left(\sum_a \nu_a\sigma_a + \frac{1}{2}\sum_{a\neq b}\mu_{ab}\sigma_a\sigma_b\right)\right)\right].$$

We now apply the saddle point method: in the large n limit, the expression for $\mathbb{E}[Z^r]$ should be dominated by a single value of the overlap parameters q_{ab}, c_a. This yields

$$\frac{1}{n}\log \mathbb{E}[Z^r] = -G(q_{ab}^*, c_a^*, \mu_{ab}^*, \nu_a^*)$$

where $(q_{ab}^*, c_a^*, \mu_{ab}^*, \nu_a^*)$ is a critical point of

$$\begin{aligned} G(q_{ab}, c_a, \mu_{ab}, \nu_a) = &-\frac{\lambda^2}{2}\sum_a c_a^2 - \frac{\lambda^2}{4}\sum_{a,b} q_{ab}^2 + \sum_a \nu_a c_a + \frac{1}{2}\sum_{a\neq b}\mu_{ab}q_{ab} \\ &- \log \sum_{\sigma\in\{\pm 1\}^r} \exp\left(\sum_a \nu_a\sigma_a + \frac{1}{2}\sum_{a\neq b}\mu_{ab}\sigma_a\sigma_b\right). \end{aligned}$$

We next assume that the dominant saddle point takes a particular form: the so-called *replica symmetric ansatz*. The replica symmetric ansatz is given by $q_{aa} = 1$, $c_a = c$, $\nu_a = \nu$, and for $a \neq b$, $q_{ab} = q$ and $\mu_{ab} = \mu$ for constants q, c, μ, ν. This yields

(10)
$$\lim_{r\to 0}\frac{1}{r}G(q,c,\mu,\nu) = -\frac{\lambda^2}{2}c^2 - \frac{\lambda^2}{4} + \frac{\lambda^2}{4}q^2 + \nu c - \frac{1}{2}\mu(q-1) - \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)} \log(2\cosh(\nu + \sqrt{\mu}z))$$

where the last term is handled as follows:

$$\begin{aligned}
&\lim_{r\to 0}\frac{1}{r}\log\sum_{\sigma\in\{\pm1\}^r}\exp\left(\sum_a \nu_a\sigma_a+\frac{1}{2}\sum_{a\neq b}\mu_{ab}\sigma_a\sigma_b\right)\\
&=\lim_{r\to 0}\frac{1}{r}\log\sum_{\sigma\in\{\pm1\}^r}\exp\left(\nu\sum_a \sigma_a+\frac{\mu}{2}\sum_{a\neq b}\sigma_a\sigma_b\right)\\
&=\lim_{r\to 0}\frac{1}{r}\log\sum_{\sigma\in\{\pm1\}^r}\exp\left(\nu\sum_a \sigma_a+\frac{\mu}{2}\sum_{a,b}\sigma_a\sigma_b-\frac{r\mu}{2}\right)\\
&=\lim_{r\to 0}\frac{1}{r}\log\sum_{\sigma\in\{\pm1\}^r}\exp(-r\mu/2)\exp\left(\nu\sum_a \sigma_a+\frac{\mu}{2}\left(\sum_a\sigma_a\right)^2\right)\\
&=-\frac{\mu}{2}+\lim_{r\to 0}\frac{1}{r}\log\sum_{\sigma\in\{\pm1\}^r}\exp\left(\nu\sum_a \sigma_a+\frac{\mu}{2}\left(\sum_a\sigma_a\right)^2\right)\\
&\overset{(a)}{=}-\frac{\mu}{2}+\lim_{r\to 0}\frac{1}{r}\log\sum_{\sigma\in\{\pm1\}^r}\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}\exp\left(\nu\sum_a \sigma_a+\sqrt{\mu}z\sum_a\sigma_a\right)\\
&=-\frac{\mu}{2}+\lim_{r\to 0}\frac{1}{r}\log\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}\sum_{\sigma\in\{\pm1\}^r}\exp\left((\nu+\sqrt{\mu}z)\sum_a\sigma_a\right)\\
&=-\frac{\mu}{2}+\lim_{r\to 0}\frac{1}{r}\log\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}\left[\exp(\nu+\sqrt{\mu}z)+\exp(-(\nu+\sqrt{\mu}z))\right]^r\\
&=-\frac{\mu}{2}+\lim_{r\to 0}\frac{1}{r}\log\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}(2\cosh(\nu+\sqrt{\mu}z))^r\\
&\overset{(b)}{=}-\frac{\mu}{2}+\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}\log(2\cosh(\nu+\sqrt{\mu}z))
\end{aligned}$$

where (a) uses the Gaussian moment-generating function and (b) uses the replica trick (8).

We next find the critical points by setting the derivatives of (10) (with respect to all four variables) to zero, which yields

$$\nu=\lambda^2 c,\qquad \mu=\lambda^2 q,\qquad c=\mathbb{E}_z\tanh(\nu+\sqrt{\mu}z),\qquad q=\mathbb{E}_z\tanh^2(\nu+\sqrt{\mu}z).$$

Recall that the replicas are drawn from the posterior distribution $\Pr[x\,|\,Y]$ and so the truth x behaves as if it is a replica; therefore we should have $c=q$. Using the identity $\mathbb{E}_z\tanh(\gamma+\sqrt{\gamma}z)=\mathbb{E}_z\tanh^2(\gamma+\sqrt{\gamma}z)$ (see e.g. [DAM16]), we obtain the solution $c=q$ and $\nu=\mu$ where q and μ are solutions to

$$\mu=\lambda^2 q,\qquad q=\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}\tanh(\mu+\sqrt{\mu}z). \tag{11}$$

The solution q to this equation tells us about the structure of the posterior distribution; namely, if we take two independent draws from this distribution, their overlap will concentrate about q. (Equivalently, the true signal x and a draw from the posterior distribution will also have overlap that concentrates about q.) Note

that (11) exactly matches the state evolution fixed-point equation (6) with μ in place of γ and $q = \gamma/\lambda^2$.

The free energy density of a solution to (11) is given by

$$f = \frac{1}{\beta} \lim_{r \to 0} \frac{1}{r} G(q, c, \mu, \nu) = \frac{1}{\lambda} \left[-\frac{\lambda^2}{4}(q^2 + 1) + \frac{1}{2}\mu(q+1) - \mathbb{E}_z \log(2\cosh(\mu + \sqrt{\mu} z)) \right].$$

This is how one can derive the free energy curves such as those shown in Figure 1. If there are multiple solutions to (11), we should take the one with minimum free energy.

Above, we had a Gibbs distribution corresponding to the posterior distribution of a Bayesian inference problem. In this setting, the replica symmetric ansatz is always correct; this is justified by a phenomenon in statistical physics: "there is no static replica symmetry breaking on the Nishimori line" (see e.g. [ZK16, Nis01]).

More generally, one can apply the replica method to a Gibbs distribution that does not correspond to a posterior distribution (e.g. if the 'temperature' of the Gibbs distribution does not match the signal-to-noise of the observed data). This is important when investigating computational hardness of random non-planted or non-Bayesian problems. In this case, the optimal (lowest free energy) saddle point can take various forms, which are summarized below; the form of the optimizer reveals a lot about the structure of the Gibbs distribution. An important property of a Gibbs distribution is its overlap distribution: the distribution of the overlap between two independent draws from the Gibbs distribution (in the large n limit).

- RS (replica symmetric): The overlap matrix is $q_{aa} = 1$ and $q_{ab} = q$ for some $q \in [0, 1]$. The overlap distribution is supported on a single point mass at value q. The Gibbs distribution can be visualized as having one large cluster where any two vectors in this cluster have overlap q. This case is "easy" in the sense that belief propagation can easily move around within the single cluster and find the true posterior distribution.
- 1RSB (1-step replica symmetry breaking): The $r \times r$ overlap matrix takes the following form. The r replicas are partitioned into blocks of size m. We have $q_{aa} = 1$, $q_{ab} = q_1$ if a, b are in the same block, and $q_{ab} = q_2$ otherwise (for some $q_1, q_2 \in [0, 1]$). The overlap distribution is supported on q_1 and q_2. The Gibbs distribution can be visualized as having a constant number of clusters. Two vectors in the same cluster have overlap q_1 whereas two vectors in different clusters have overlap q_2. This case is "hard" for belief propagation because it gets stuck in one cluster and cannot correctly capture the posterior distribution. The idea of replica symmetry breaking was first proposed in a groundbreaking work of Parisi [Par79].
- 2RSB (2-step replica symmetry breaking): Now we have "clusters of clusters." The overlap matrix has sub-blocks within each block. The overlap distribution is supported on 3 different values (corresponding to "same sub-block", "same block (but different sub-block)", "different blocks"). The Gibbs distribution has a constant number of clusters, each with a constant number of sub-clusters. This is again "hard" for belief propagation.
- FSRB (full replica symmetry breaking): We can define kRSB for any k as above (characterized by an overlap distribution supported on $k+1$ values); FRSB is the limit of kRSB as $k \to \infty$. Here the overlap distribution is a continuous distribution.

- d1RSB (dynamic 1RSB): This phase is similar to RS and (unlike kRSB for $k \geq 1$) can appear in Bayesian inference problems. The overlap matrix is the same as in the RS phase (and so the replica calculation proceeds exactly as in the RS case). However, the Gibbs distribution has exponentially-many small clusters. The overlap distribution is supported on a single point mass because two samples from the Gibbs distribution will be in different clusters with high probability. This phase is "hard" for BP (or AMP) because it cannot easily move between clusters. For a Bayesian inference problem, you can tell whether you are in the RS (easy) phase or 1dRSB (hard) phase by looking at the free energy curve; 1dRSB corresponds to the "hard" phase (c) in Figure 1.

Acknowledgements. The authors would like to thank the engaging audience at the Courant Institute when this material was presented, and their many insightful comments. The authors would also like to thank Soledad Villar and Lenka Zdeborová for feedback on earlier versions of this manuscript.

References

[ABAČ13] Antonio Auffinger, Gérard Ben Arous, and Jiří Černỳ. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.

[Abb17] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*, 2017.

[AKS98] N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *SODA*, 1998.

[AS15] Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015.

[AT07] R. J. Adler and J. E. Taylor. *Random Fields and Geometry*. Springer, 2007.

[BBV16] A. S. Bandeira, N. Boumal, and V. Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. *COLT*, 2016.

[BDM+16] Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Advances in Neural Information Processing Systems*, pages 424–432, 2016.

[BHK+16] B. Barak, S. B. Hopkins, J. Kelner, P. K. Kothari, A. Moitra, and A. Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *Available online at arXiv:1604.03084 [cs.CC]*, 2016.

[BKS13] B. Barak, J. Kelner, and D. Steurer. Rounding sum-of-squares relaxations. *Available online at arXiv:1312.6652 [cs.DS]*, 2013.

[BKS15] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151. ACM, 2015.

[BM11] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[BM16] Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*, pages 417–445, 2016.

[BMNN16] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416, 2016.

[Bol12] Erwin Bolthausen. An iterative construction of solutions of the tap equations for the sherrington-kirkpatrick model. *arXiv preprint arXiv:1201.2891*, 2012.

[BR12] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 2012.

[BR13] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. *Conference on Learning Theory (COLT)*, 2013.

[BS14] B. Barak and D. Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *Survey, ICM 2014*, 2014.

[BVB16] N. Boumal, V. Voroninski, and A. S. Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. *NIPS*, 2016.

[COKPZ16] Amin Coja-Oghlan, Florent Krzakala, Will Perkins, and Lenka Zdeborova. Information-theoretic thresholds from the cavity method. *arXiv preprint arXiv:1611.00814*, 2016.

[DAM16] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the binary stochastic block model. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 185–189. IEEE, 2016.

[DKMZ11] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):065701, 2011.

[DM14] Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse PCA. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 2197–2201. IEEE, 2014.

[DM15] Yash Deshpande and Andrea Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Foundations of Computational Mathematics*, 15(4):1069–1128, 2015.

[DMM09] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[FR12] Alyson K Fletcher and Sundeep Rangan. Iterative reconstruction of rank-one matrices in noise. *Information and Inference: A Journal of the IMA*, 2012.

[GJZ17] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

[GM15] Rong Ge and Tengyu Ma. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. *arXiv preprint arXiv:1504.05287*, 2015.

[GM17] R. Ge and T. Ma. On the optimization landscape of tensor decompositions. *Available online at arXiv:1706.05598 [cs.LG]*, 2017.

[GW95] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefine programming. *Journal of the Association for Computing Machinery*, 42:1115–1145, 1995.

[HLL83] P. W. Holland, K. Blackmond Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[HS17] Samuel B Hopkins and David Steurer. Bayesian estimation from few samples: community detection and related problems. *arXiv preprint arXiv:1710.00264*, 2017.

[HSS15] S. B. Hopkins, J. Shi, and D. Steurer. Tensor principal component analysis via sum-of-squares proofs. *Available at arXiv:1507.03269 [cs.LG]*, 2015.

[HSSS16] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191. ACM, 2016.

[HWX14] B. Hajek, Y. Wu, and J. Xu. Computational lower bounds for community detection on random graphs. *Available online at arXiv:1406.6625*, 2014.

[JM13] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[JMRT16] Adel Javanmard, Andrea Montanari, and Federico Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *Proceedings of the National Academy of Sciences*, 113(16):E2218–E2223, 2016.

[Kar72] R. M. Karp. Reducibility among combinatorial problems. *Complexity of Computer Computation*, 1972.

[KBG17] C. Kim, A. S. Bandeira, and M. X. Goemans. Community detection in hypergraphs, spiked tensor models, and sum-of-squares. *SampTA 2017: Sampling Theory and Applications, 12th international conference*, 2017.

[Kho02] S. Khot. On the power of unique 2-prover 1-round games. *Thiry-fourth annual ACM symposium on Theory of computing*, 2002.

[Kho10] S. Khot. On the unique games conjecture (invited survey). In *Proceedings of the 2010 IEEE 25th Annual Conference on Computational Complexity*, CCC '10, pages 99–121, Washington, DC, USA, 2010. IEEE Computer Society.

[KS66] Harry Kesten and Bernt P Stigum. A limit theorem for multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, 1966.

[KXZ16] Florent Krzakala, Jiaming Xu, and Lenka Zdeborová. Mutual information in rank-one matrix estimation. In *Information Theory Workshop (ITW), 2016 IEEE*, pages 71–75. IEEE, 2016.

[Las01] J. B. Lassere. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[LKZ15a] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *53rd Annual Allerton Conference on Communication, Control, and Computing*, pages 680–687. IEEE, 2015.

[LKZ15b] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Phase transitions in sparse pca. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 1635–1639. IEEE, 2015.

[LM16] Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *arXiv preprint arXiv:1611.03888*, 2016.

[LML+17] T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. *arXiv preprint arXiv:1701.08010*, 2017.

[Lov79] L. Lovasz. On the shannon capacity of a graph. *IEEE Trans. Inf. Theor.*, 25(1):1–7, 1979.

[LV18] J. Bruna L. Venturi, A. S. Bandeira. Neural networks with finite intrinsic dimension have no spurious valleys. *arXiv:1802.06384 [math.OC]*, 2018.

[Mas14] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM, 2014.

[MM09] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[MNS12] Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

[MNS13] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

[Moo17] Cristopher Moore. The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *arXiv preprint arXiv:1702.00467*, 2017.

[MPV86] Marc Mézard, Giorgio Parisi, and MA Virasoro. SK model: The replica solution without replicas. *EPL (Europhysics Letters)*, 1(2):77, 1986.

[MR16] Andrea Montanari and Emile Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Transactions on Information Theory*, 62(3):1458–1484, 2016.

[MSS16] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 438–446. IEEE, 2016.

[MW13] Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *Available online at arXiv:1309.5914*, 2013.

[Nes00] Y. Nesterov. Squared functional systems and optimization problems. *High performance optimization*, 13(405-440), 2000.

[Nis80] Hidetoshi Nishimori. Exact results and critical properties of the ising model with competing interactions. *Journal of Physics C: Solid State Physics*, 13(21):4071, 1980.

[Nis81] Hidetoshi Nishimori. Internal energy, specific heat and correlation function of the bond-random ising model. *Progress of Theoretical Physics*, 66(4):1169–1181, 1981.

[Nis01] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[Par79] Giorgio Parisi. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23):1754, 1979.

[Par00] P. A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000.

[Pea86] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, 29(3):241–288, 1986.

[PS17] Aaron Potechin and David Steurer. Exact tensor completion with sum-of-squares. *arXiv preprint arXiv:1702.06237*, 2017.

[PWB16] Amelia Perry, Alexander S Wein, and Afonso S Bandeira. Statistical limits of spiked tensor models. *arXiv preprint arXiv:1612.07728*, 2016.

[PWBM16a] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Message-passing algorithms for synchronization problems over compact groups. *Communications on Pure and Applied Mathematics (to appear). arXiv preprint arXiv:1610.04583*, 2016.

[PWBM16b] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of PCA for spiked random matrices and synchronization. *arXiv preprint arXiv:1609.05573*, 2016.

[Rag08] P. Raghavendra. Optimal algorithms and inapproximability results for every CSP? In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 245–254. ACM, 2008.

[RM14] Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

[Sho87] N. Shor. An approach to obtaining global extremums in polynomial mathematical programming problems. *Cybernetics and Systems Analysis*, 23(5):695–700, 1987.

[Sin11] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.*, 30(1):20 – 36, 2011.

[TAP77] David J Thouless, Philip W Anderson, and Robert G Palmer. Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

[ZK16] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

(Bandeira) Department of Mathematics and Center for Data Science, Courant Institute of Mathematical Sciences, New York University
E-mail address: bandeira@cims.nyu.edu

(Perry) Department of Mathematics, Massachusetts Institute of Technology

(Wein) Department of Mathematics, Massachusetts Institute of Technology
E-mail address: awein@mit.edu