Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula

Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, Lenka Zdeborova

Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula

Jean Barbier, Mohamad Dia and Nicolas Macris FIRSTNAME.LASTNAME@EPFL.CH
Laboratoire de Théorie des Communications, Faculté Informatique et Communications, Ecole Polytechnique Fédérale de Lausanne, 1015, Suisse.

Florent Krzakala FLORENT.KRZAKALA@ENS.FR
Laboratoire de Physique Statistique, CNRS, PSL Universités et Ecole Normale Supérieure, Sorbonne Universités et Université Pierre & Marie Curie, 75005, Paris, France.

Thibault Lesieur and Lenka Zdeborová LESIEUR.THIBAULT,LENKA.ZDEBOROVA@GMAIL.COM
Institut de Physique Théorique, CNRS, CEA, Université Paris-Saclay, F-91191, Gif-sur-Yvette, France.

Abstract

Factorizing low-rank matrices has many applications in machine learning and statistics. For probabilistic models in the Bayes optimal setting, a general expression for the mutual information has been proposed using heuristic statistical physics computations, and proven in few specific cases. Here, we show how to rigorously prove the conjectured formula for the symmetric rank-one case. This allows to express the minimal mean-square-error and to characterize the detectability phase transitions in a large set of estimation problems ranging from community detection to sparse PCA. We also show that for a large set of parameters, an iterative algorithm called approximate message-passing is Bayes optimal. There exists, however, a gap between what currently known polynomial algorithms can do and what is expected information theoretically. Additionally, the proof technique has an interest of its own and exploits three essential ingredients: the interpolation method introduced in statistical physics by Guerra, the analysis of the approximate message-passing algorithm and the theory of spatial coupling and threshold saturation in coding. Our approach is generic and applicable to other open problems in statistical estimation where heuristic statistical physics predictions are available.

Consider the following probabilistic rank-one matrix estimation problem: one has access to noisy observations $\mathbf{w}=(w_{ij})_{i,j=1}^{n}$ of the pair-wise product of the components of a vector $\mathbf{s}=(s_1,\dots,s_n)^\intercal\in\mathbb{R}^n$ with i.i.d components distributed as $S_i\sim P_0$, $i=1,\dots,n$. The entries of $\mathbf{w}$ are observed through a noisy element-wise (possibly non-linear) output probabilistic channel $P_{\rm out}(w_{ij}|s_is_j/\sqrt{n})$. The goal is to estimate the vector $\mathbf{s}$ from $\mathbf{w}$ assuming that both P_0 and $P_{\rm out}$ are known and independent of n (noise is symmetric so that $w_{ij}=w_{ji}$). Many important problems in statistics and machine learning can be expressed in this way, such as sparse PCA [Zou et al. (2006)], the Wigner spike model [Johnstone and Lu (2012); Deshpande and Montanari (2014)], community detection [Deshpande et al. (2015)] or matrix completion [Candès and Recht (2009)].

Proving a result initially derived by a heuristic method from statistical physics, we give an explicit expression for the mutual information and the information theoretic minimal mean-square-error (MMSE) in the asymptotic $n\to+\infty$ limit. Our results imply that for

a large region of parameters, the posterior marginal expectations of the underlying signal components (often assumed intractable to compute) can be obtained in the leading order in n using a polynomial-time algorithm called approximate message-passing (AMP) [Rangan and Fletcher (2012); Deshpande and Montanari (2014); Deshpande et al. (2015); Lesieur et al. (2015b)]. We also demonstrate the existence of a region where both AMP and spectral methods [Baik et al. (2005)] fail to provide a good answer to the estimation problem, while it is nevertheless information theoretically possible to do so. We illustrate our theorems with examples and also briefly discuss the implications in terms of computational complexity.

1. Setting and main results

1.1 The additive white Gaussian noise setting

A standard and natural setting is the case of additive white Gaussian noise (AWGN) of known variance Δ,

$$w_{ij} = \frac{s_i s_j}{\sqrt{n}} + z_{ij}\sqrt{\Delta}, \tag{1}$$

where $\mathbf{z}=(z_{ij})_{i,j=1}^{n}$ is a symmetric matrix with i.i.d entries $Z_{ij}\sim\mathcal{N}(0,1)$, $1\leq i\leq j\leq n$. Perhaps surprisingly, it turns out that this Gaussian setting is sufficient to completely characterize all the problems discussed in the introduction, even if these have more complicated output channels. This is made possible by a theorem of channel universality [Krzakala et al. (2016)] (already proven for community detection in [Deshpande et al. (2015)] and conjectured in [Lesieur et al. (2015a)]). This theorem states that given an output channel $P_{\rm out}(w|y)$, such that $\log P_{\rm out}(w|y=0)$ is three times differentiable with bounded second and third derivatives, then the mutual information satisfies $I(\mathbf{S};\mathbf{W})=I(\mathbf{S};\mathbf{S}\mathbf{S}^\intercal/\sqrt{n}+\mathbf{Z}\sqrt{\Delta})+\mathcal{O}(\sqrt{n})$, where Δ is the inverse Fisher information (evaluated at $y=0$) of the output channel: $\Delta^{-1} := \mathbb{E}_{P_{\rm out}(w|0)}[(\partial_y \log P_{\rm out}(W|y)|_{y=0})^2]$. Informally, this means that we only have to compute the mutual information for an AWGN channel to take care of a wide range of problems, which can be expressed in terms of their Fisher information. In this paper we derive rigorously, for a large class of signal distributions P_0, an explicit one-letter formula for the mutual information per variable $I(\mathbf{S};\mathbf{W})/n$ in the asymptotic limit $n\to+\infty$.

1.2 Main result

Our central result is a proof of the expression for the asymptotic $n\to+\infty$ mutual information per variable via the so-called *replica symmetric potential function* $i_{\rm RS}(E;\Delta)$ defined as

$$i_{\rm RS}(E;\Delta) := \frac{(v-E)^2+v^2}{4\Delta} - \mathbb{E}_{S,Z}\left[\ln\left(\int dx\, P_0(x) e^{-\frac{x^2}{2\Sigma(E;\Delta)^2}+x\left(\frac{S}{\Sigma(E;\Delta)^2}+\frac{Z}{\Sigma(E;\Delta)}\right)}\right)\right], \tag{2}$$

with $Z\sim\mathcal{N}(0,1)$, $S\sim P_0$, $\mathbb{E}[S^2]=v$ and $\Sigma(E;\Delta)^2:=\Delta/(v-E)$, $E\in[0,v]$. Here we will assume that P_0 is a discrete distribution over a finite bounded real alphabet $P_0(s)=\sum_{\alpha=1}^{\nu} p_\alpha \delta(s-a_\alpha)$. Thus the only continuous integral in (2) is the Gaussian over z. Our results can be extended to mixtures of discrete and continuous signal distributions at the expense of technical complications in some proofs.

It turns out that both the information theoretical and algorithmic AMP thresholds are determined by the set of stationary points of (2) (w.r.t E). It is possible to show that for all $\Delta > 0$ there always exist at least one stationary minimum. Note $E = 0$ is never a stationary point (except for P_0 a single Dirac mass) and $E = v$ is stationary only if $\mathbb{E}[S] = 0$. In this contribution we suppose that at most three stationary points exist, corresponding to situations with at most one phase transition. We believe that situations with multiple transitions can also be covered by our techniques.

Theorem 1 (One letter formula for the mutual information) *Fix $\Delta > 0$ and assume P_0 is a discrete distribution such that $i_{\rm RS}(E;\Delta)$ given by (2) has at most three stationary points. Then*

$$\lim_{n\to+\infty} \frac{1}{n} I(\boldsymbol{S};\boldsymbol{W}) = \min_{E\in[0,v]} i_{\rm RS}(E;\Delta). \tag{3}$$

The proof of the *existence of the limit* does not require the above hypothesis on P_0. Also, it was first shown in [Krzakala et al. (2016)] that for all n, $I(\mathbf{S};\mathbf{W})/n \leq \min_{E\in[0,v]} i_{\rm RS}(E;\Delta)$, an inequality that *we will use* in the proof section. It is conceptually useful to define the following threshold:

Definition 2 (Information theoretic threshold) *Define $\Delta_{\rm Opt}$ as the first non-analyticity point of the asymptotic mutual information per variable as Δ increases, that is formally $\Delta_{\rm Opt} := \sup\{\Delta |\, \lim_{n\to+\infty} I(\boldsymbol{S};\boldsymbol{W})/n$ is analytic in $]0,\Delta[\}$.*

When P_0 is such that (2) has at most three stationary points, as discussed below, then $\min_{E\in[0,v]} i_{\rm RS}(E;\Delta)$ has at most one non-analyticity point denoted $\Delta_{\rm RS}$ (if $\min_{E\in[0,v]} i_{\rm RS}(E;\Delta)$ is analytic over all $\mathbb{R}_+$ we set $\Delta_{\rm RS} = +\infty$). Theorem 1 gives us a mean to *compute* the information theoretical threshold $\Delta_{\rm Opt} = \Delta_{\rm RS}$. A basic application of theorem 1 is the expression of the MMSE:

Corollary 3 (Exact formula for the MMSE) *For all $\Delta \neq \Delta_{\rm RS}$, the matrix-MMSE $\mathrm{Mmmse}_n := \mathbb{E}_{\boldsymbol{S},\boldsymbol{W}} \|\boldsymbol{S}\boldsymbol{S}^\intercal - \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\intercal|\,\boldsymbol{W}]\|_{\rm F}^2/n^2$ ($\|-\|_{\rm F}$ being the Frobenius norm) is asymptotically $\lim_{n\to+\infty} \mathrm{Mmmse}_n(\Delta^{-1}) = v^2 - (v - argmin_{E\in[0,v]} i_{\rm RS}(E;\Delta))^2$. Moreover, if $\Delta < \Delta_{\rm AMP}$ (where $\Delta_{\rm AMP}$ is the algorithmic threshold, see definition 4) or $\Delta > \Delta_{\rm RS}$, then the usual vector-MMSE $\mathrm{Vmmse}_n := \mathbb{E}_{\boldsymbol{S},\boldsymbol{W}} \|\boldsymbol{S} - \mathbb{E}[\boldsymbol{X}|\,\boldsymbol{W}]\|_2^2/n$ satisfies $\lim_{n\to+\infty} \mathrm{Vmmse}_n = argmin_{E\in[0,v]} i_{\rm RS}(E;\Delta)$.*

It is natural to conjecture that the vector-MMSE is given by $\mathrm{argmin}_{E\in[0,v]} i_{\rm RS}(E;\Delta)$ for all $\Delta \neq \Delta_{\rm RS}$, but our proof does not quite yield the full statement.

A fundamental consequence concerns the performance of the AMP algorithm [Rangan and Fletcher (2012)] for estimating $\mathbf{s}$. AMP has been analysed rigorously in [Bayati and Montanari (2011); Javanmard and Montanari (2013); Deshpande et al. (2015)] where it is shown that its asymptotic performance is tracked by *state evolution*. Let $E^t := \lim_{n\to+\infty} \mathbb{E}_{\mathbf{S},\mathbf{Z}}[\|\mathbf{S} - \hat{\mathbf{s}}^t\|_2^2]/n$ be the asymptotic average vector-MSE of the AMP estimate $\hat{\mathbf{s}}^t$ at time t. Define $\mathrm{mmse}(\Sigma^{-2}) := \mathbb{E}_{S,Z}[(S - \mathbb{E}[X|S + \Sigma Z])^2]$ as the usual scalar mmse function associated to a scalar AWGN channel of noise variance Σ^2, with $S \sim P_0$ and $Z \sim \mathcal{N}(0,1)$. Then

$$E^{t+1} = \mathrm{mmse}(\Sigma(E^t;\Delta)^{-2}), \qquad E^0 = v, \tag{4}$$

is the state evolution recursion. Monotonicity properties of the mmse function imply that E^t is a decreasing sequence such that $\lim_{t\to+\infty} E^t = E^\infty$ exists. Note that when $\mathbb{E}[S]=0$ and v is an unstable fixed point, as such, state evolution "does not start". While this is not really a problem when one runs AMP in practice, for analysis purposes one can slightly bias P_0 and remove the bias at the end of the proofs.

Definition 4 (AMP algorithmic threshold) *For $\Delta > 0$ small enough, the fixed point equation corresponding to (4) has a unique solution for all noise values in $]0,\Delta[$. We define Δ_{AMP} as the supremum of all such Δ.*

Corollary 5 (Performance of AMP) *In the limit $n\to+\infty$, AMP initialized without any knowledge other than P_0 yields upon convergence the asymptotic matrix-MMSE as well as the asymptotic vector-MMSE iff $\Delta<\Delta_{\text{AMP}}$ or $\Delta>\Delta_{\text{RS}}$, namely $E^\infty = argmin_{E\in[0,v]} i_{\text{RS}}(E;\Delta)$.*

Δ_{AMP} can be read off the replica potential (2): by differentiation of (2) one finds a fixed point equation that corresponds to (4). Thus Δ_{AMP} is the smallest solution of $\partial i_{\text{RS}}/\partial E = \partial^2 i_{\text{RS}}/\partial E^2 = 0$; in other words it is the "first" horizontal inflexion point that appears in $i_{\text{RS}}(E;\Delta)$ when we increase Δ.

1.3 Discussion

With our hypothesis on P_0 there are only three possible scenarios: $\Delta_{\text{AMP}} < \Delta_{\text{RS}}$ (one "first order" phase transition); $\Delta_{\text{AMP}} = \Delta_{\text{RS}} < +\infty$ (one "higher order" phase transition); $\Delta_{\text{AMP}} = \Delta_{\text{RS}} = +\infty$ (no phase transition). In the sequel we will have in mind the most interesting case, namely *one first order phase transition*, where we determine the gap between the algorithmic AMP and information theoretic performance. The cases of no phase transition or higher order phase transition, which present no algorithmic gap, are basically covered by the analysis of [Deshpande and Montanari (2014)] and follow as a special case from our proof. The only cases that would require more work are those where P_0 is such that (2) develops more than three stationary points and more than one phase transition is present.

For $\Delta_{\text{AMP}} < \Delta_{\text{RS}}$ the structure of stationary points of (2) is as follows[1] (figure 1). There exist three branches $E_{\text{good}}(\Delta)$, $E_{\text{unstable}}(\Delta)$ and $E_{\text{bad}}(\Delta)$ such that: **1)** For $0 < \Delta < \Delta_{\text{AMP}}$ there is a single stationary point $E_{\text{good}}(\Delta)$ which is a global minimum; **2)** At Δ_{AMP} a *horizontal inflexion point* appears, for $\Delta\in[\Delta_{\text{AMP}},\Delta_{\text{RS}}]$ there are three stationary points satisfying $E_{\text{good}}(\Delta_{\text{AMP}}) < E_{\text{unstable}}(\Delta_{\text{AMP}}) = E_{\text{bad}}(\Delta_{\text{AMP}})$, $E_{\text{good}}(\Delta) < E_{\text{unstable}}(\Delta) < E_{\text{bad}}(\Delta)$ otherwise, and moreover $i_{\text{RS}}(E_{\text{good}};\Delta) \le i_{\text{RS}}(E_{\text{bad}};\Delta)$ *with equality only at* Δ_{RS}; **3)** for $\Delta > \Delta_{\text{RS}}$ there is *at least* the stationary point $E_{\text{bad}}(\Delta)$ which is always the global minimum, i.e. $i_{\text{RS}}(E_{\text{bad}};\Delta) < i_{\text{RS}}(E_{\text{good}};\Delta)$. (For higher Δ the $E_{\text{good}}(\Delta)$ and $E_{\text{unstable}}(\Delta)$ branches may merge and disappear); **4)** $E_{\text{good}}(\Delta)$ is analytic for $\Delta\in]0,\Delta'[$, $\Delta' > \Delta_{\text{RS}}$, and $E_{\text{bad}}(\Delta)$ is analytic for $\Delta > \Delta_{\text{AMP}}$.

We note for further use in the proof section that $E^\infty = E_{\text{good}}(\Delta)$ for $\Delta < \Delta_{\text{AMP}}$ and $E^\infty = E_{\text{bad}}(\Delta)$ for $\Delta > \Delta_{\text{AMP}}$. Definition 4 is equivalent to $\Delta_{\text{AMP}} = \sup\{\Delta | E^\infty = E_{\text{good}}(\Delta)\}$. Moreover we will also use that $i_{\text{RS}}(E_{\text{good}};\Delta)$ is analytic on $]0,\Delta'[$, $i_{\text{RS}}(E_{\text{bad}};\Delta)$ is analytic on $]\Delta_{\text{AMP}},+\infty[$, and the only non-analyticity point of $\min_{E\in[0,v]} i_{\text{RS}}(E;\Delta)$ is at Δ_{RS}.

1. We take $\mathbb{E}[S]\neq 0$. Once theorem 1 is proven for this case a limiting argument allows to extend it to $\mathbb{E}[S]=0$.

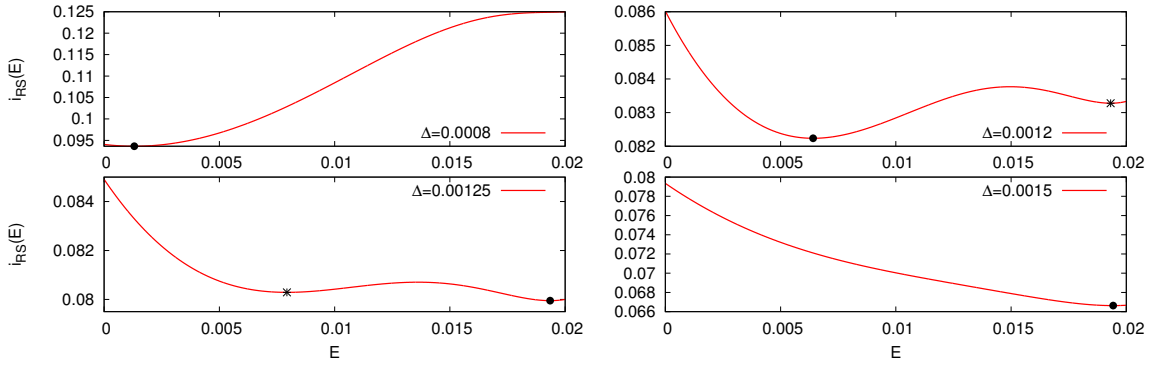

Figure 1: The replica formula $i_{\mathrm{RS}}(E)$ for four values of Δ in the Wigner spike model. The mutual information is $\min i_{\mathrm{RS}}(E)$ (the black dot, while the black cross corresponds to the local minimum) and the asymptotic matrix-MMSE is $v^2-(v-\mathrm{argmin}_E i_{\mathrm{RS}}(E))^2$, where $v=\rho$ in this case with $\rho=0.02$ as in the inset of figure 2. From top left to bottom right: **(1)** For low noise values, here $\Delta=0.0008<\Delta_{\mathrm{AMP}}$, there exists a unique "good" minimum corresponding to the MMSE and AMP is Bayes optimal. **(2)** As the noise increases, a second local "bad" minimum appears: this is the situation at $\Delta_{\mathrm{AMP}}<\Delta=0.0012<\Delta_{\mathrm{RS}}$. **(3)** For $\Delta=0.00125>\Delta_{\mathrm{RS}}$, the "bad" minimum becomes the global one and the MMSE suddenly deteriorates. **(4)** For even larger values of Δ, only the "bad" minimum exists. The AMP algorithm can be seen as a naive minimizer of this curve starting from $E=v=0.02$. It reaches the global minimum in situations (1), (3) and (4), but in (2), when $\Delta_{\mathrm{AMP}}<\Delta<\Delta_{\mathrm{RS}}$, it is trapped by the local minimum with large MSE instead of reaching the global one corresponding to the MMSE.

1.4 Relation to other works

Explicit single-letter characterization of the mutual information in the rank-one problem has attracted a lot of attention recently. Particular cases of (3) have been shown rigorously in a number of situations. A special case when $s_i=\pm 1\sim\mathrm{Ber}(1/2)$ already appeared in [Korada and Macris (2009)] where an equivalent spin glass model is analysed. Very recently, [Krzakala et al. (2016)] has generalized the results of [Korada and Macris (2009)] and, notably, obtained a generic matching upper bound. The same formula has been also rigorously computed following the study of AMP in [Deshpande and Montanari (2014)] for spike models (provided, however, that the signal was not *too* sparse) and in [Deshpande et al. (2015)] for strictly symmetric community detection.

For rank-one symmetric matrix estimation problems, AMP has been introduced by [Rangan and Fletcher (2012)], who also computed the state evolution formula to analyse its performance, generalizing techniques developed by [Bayati and Montanari (2011)] and [Javanmard and Montanari (2013)]. State evolution was further studied by [Deshpande and Montanari (2014)] and [Deshpande et al. (2015)]. In [Lesieur et al. (2015b,a)], the generalization to larger rank was also considered.

The general formula proposed by [Lesieur et al. (2015a)] for the conditional entropy and the MMSE on the basis of the heuristic cavity method from statistical physics was not demonstrated in full generality. Worst, all existing proofs could not reach the more interesting regime where a gap between the algorithmic and information theoretic perfomances appears,

leaving a gap with the statistical physics conjectured formula (and rigorous upper bound from [Krzakala et al. (2016)]). Our result closes this conjecture and has interesting non-trivial implications on the computational complexity of these tasks.

Our proof technique combines recent rigorous results in coding theory along the study of capacity-achieving spatially coupled codes [Hassani et al. (2010); Kudekar et al. (2011); Yedla et al. (2014); Barbier et al. (2016)] with other progress, coming from developments in mathematical physics putting on a rigorous basis predictions of spin glass theory [Guerra (2005)]. From this point of view, the theorem proved in this paper is relevant in a broader context going beyond low-rank matrix estimation. Hundreds of papers have been published in statistics, machine learning or information theory using the non-rigorous statistical physics approach. We believe that our result helps setting a rigorous foundation of a broad line of work. While we focus on rank-one symmetric matrix estimation, our proof technique is readily extendable to more generic low-rank symmetric matrix or low-rank symmetric tensor estimation. We also believe that it can be extended to other problems of interest in machine learning and signal processing, such as generalized linear regression, features/dictionary learning, compressed sensing or multi-layer neural networks.

2. Two examples: Wigner spike model and community detection

In order to illustrate the consequences of our results we shall present two examples. In the first one we are given data distributed according to the spiked Wigner model where the vector $\mathbf{s}$ is a Bernoulli random vector, $S_i \sim \mathrm{Ber}(\rho)$. For large enough densities (i.e. $\rho > 0.041(1)$), [Deshpande and Montanari (2014)] computed the matrix-MMSE and proved that AMP is a computationally efficient algorithm that asymptotically achieves the matrix-MMSE for *any* value of the noise Δ. Our results allow to close the gap left open by [Deshpande and Montanari (2014)]: on one hand we now obtain rigorously the MMSE for $\rho \leq 0.041(1)$, and on the other one, we observe that for such values of ρ, and as Δ decreases, there is a small region where two local minima coexist in $i_{\mathrm{RS}}(E;\Delta)$. In particular for $\Delta_{\mathrm{AMP}} < \Delta < \Delta_{\mathrm{Opt}} = \Delta_{\mathrm{RS}}$ the global minimum corresponding to the MMSE differs from the local one that traps AMP, and a computational gap appears (see figure 1). While the region where AMP is Bayes optimal is quite large, the region where is it not, however, is perhaps the most interesting one. While this is by no means evident, statistical physics analogies with physical phase transitions in nature suggest that this region should be hard for a very broad class of algorithms.

For small ρ our results are consistent with the known optimal and algorithmic thresholds predicted in sparse PCA [Amini and Wainwright (2008); Berthet and Rigollet (2013)], that treats the case of sub-extensive $\rho = \mathcal{O}(1)$ values. Another interesting line of work for such probabilistic models appeared in the context of random matrix theory (see [Baik et al. (2005)] and references therein) and predicts that a sharp phase transition occurs at a critical value of the noise $\Delta_{\mathrm{spectral}} = \rho^2$ below which an outlier eigenvalue (and its principal eigenvector) has a positive correlation with the hidden signal. For larger noise values the spectral distribution of the observation is indistinguishable from that of the pure random noise.

We now consider the problem of detecting two communities (groups) with different sizes ρn and $(1-\rho)n$, that generalizes the one considered in [Deshpande et al. (2015)]. One is given a graph where the probability to have a link between nodes in the first group is $p + \mu(1-\rho)/(\rho\sqrt{n})$, between those in the second group is $p + \mu\rho/(\sqrt{n}(1-\rho))$, while

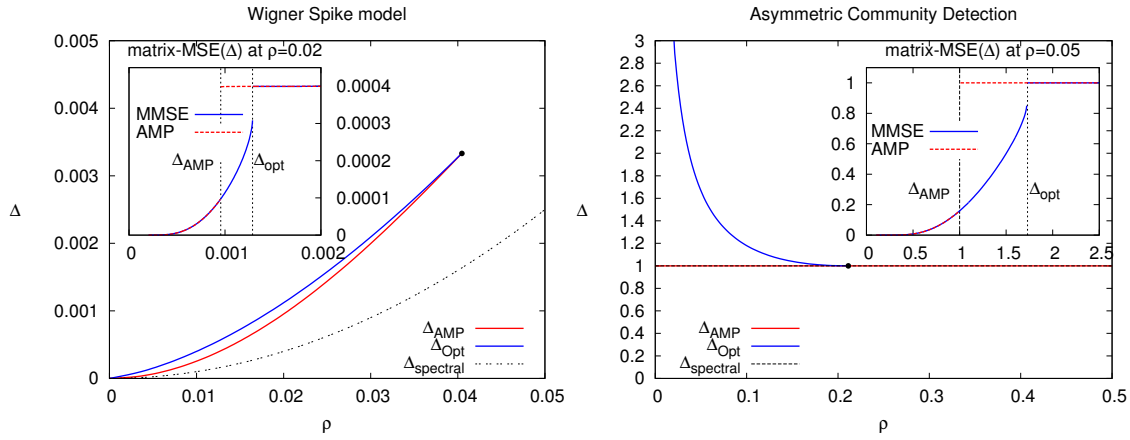

Figure 2: Phase diagram in the noise variance Δ versus density ρ plane for the rank-one spiked Wigner model (left) and the asymmetric community detection (right). **Left:** [Deshpande and Montanari (2014)] proved that AMP achieves the matrix-MMSE for all Δ as long as $\rho > 0.041(1)$. Here we show that AMP is actually achieving the optimal reconstruction in the whole phase diagram except in the small region between the blue and red lines. Notice the large gap with spectral methods (dashed black line). **Inset:** matrix-MMSE (blue) at $\rho = 0.02$ as a function of Δ. AMP (dashed red) provably achieves the matrix-MMSE except in the region $\Delta_{\rm AMP} < \Delta < \Delta_{\rm Opt} = \Delta_{\rm RS}$. We conjecture that no polynomial-time algorithm will do better than AMP in this region. **Right:** Asymmetric community detection problem with two communities. For $\rho > 1/2 - \sqrt{1/12}$ (black point) and when $\Delta > 1$, it is information theoretically impossible to find any overlap with the true communities and the matrix-MMSE is 1, while it becomes possible for $\Delta < 1$. In this region, AMP is always achieving the matrix-MMSE and spectral methods can find a non-trivial overlap with the truth as well, starting from $\Delta < 1$. For $\rho < 1/2 - \sqrt{1/12}$, however, it is information theoretically possible to find an overlap with the hidden communities for $\Delta > 1$ (below the blue line) but both AMP and spectral methods miss this information. **Inset:** matrix-MMSE (blue) at $\rho = 0.05$ as a function of Δ. AMP (dashed red) again provably achieves the matrix-MMSE except in the region $\Delta_{\rm AMP} < \Delta < \Delta_{\rm Opt}$.

interconnections appear with probability $p - \mu/\sqrt{n}$. With this peculiar "balanced" setting, the nodes in each group have the same degree distribution with mean pn, making them harder to distinguish. According to the universality property described in section 1.1, this is equivalent to a model with AWGN of variance $\Delta = p(1-p)/\mu^2$ where each variable s_i is chosen according to $P_0(s) = \rho\delta(s - \sqrt{(1-\rho)/\rho}) + (1-\rho)\delta(s + \sqrt{\rho/(1-\rho)})$. Our results for this problem[2] are summarized on the right hand side of figure 2. For $\rho > \rho_c = 1/2 - \sqrt{1/12}$ (black point), it is asymptotically information theoretically *possible* to get an estimation better than chance if and only if $\Delta < 1$. When $\rho < \rho_c$, however, it becomes possible for much larger values of the noise. Interestingly, AMP and spectral methods have the same transition and can find a positive correlation with the hidden communities for $\Delta < 1$, regardless of the value of ρ. Again, a region $[\Delta_{\rm AMP}, \Delta_{\rm Opt} = \Delta_{\rm RS}]$ exists where a computational gap appears when $\rho < \rho_c$.

One can investigate the very low ρ regime where we find that the information theoretic transition goes as $\Delta_{\rm Opt}(\rho \to 0) = 1/(4\rho|\log\rho|)$. Now if we assume that this result stays

2. Note that here since $E = v = 1$ is an extremum of $i_{\rm RS}(E; \Delta)$, one must introduce a small bias in P_0 and let it then tend to zero at the end of the proofs.

true even for $\rho=\mathcal{O}(1)$ (which is a speculation at this point), we can choose $\mu\to(1-p)\rho\sqrt{n}$ such that the small group is a clique. Then the problem corresponds to a "balanced" version of the famous planted clique problem [d'Aspremont et al. (2007)]. We find that the AMP/spectral approach finds the hidden clique when it is larger than $\sqrt{np/(1-p)}$, while the information theoretic transition translates into size of the clique $4p\log(n)/(1-p)$. This is indeed reminiscent of the more classical planted clique problem at $p=1/2$ with its gap between $\log(n)$ (information theoretic), $\sqrt{n/e}$ (AMP [Deshpande and Montanari (2015)]) and $\sqrt{n}$ (spectral [d'Aspremont et al. (2007)]). Since in our balanced case the spectral and AMP limits match, this suggests that the small gain of AMP in the standard clique problem is simply due to the information provided by the distribution of local degrees in the two groups (which is absent in our balanced case). We believe this correspondence strengthens the claim that the AMP gap is actually a fundamental one.

3. Proofs

The crux of our proof rests on an auxiliary "spatially coupled system". The hallmark of spatially coupled models is that one can tune them so that the gap between the algorithmic and information theoretical limits can be eliminated, while at the same time the mutual information is maintained unchanged for the coupled and original models. Roughly speaking, this means that it is possible to algorithmically compute the information theoretical limit of the original model because a suitable algorithm is optimal on the coupled system.

The spatially coupled construction used here is very similar to the one used for the coupled Curie-Weiss model [Hassani et al. (2010)]. We consider a ring of length $L+1$ (L even) with *blocks* positioned at $\mu\in\{0,\dots,L\}$ and coupled to neighboring blocks $\{\mu-w,\dots,\mu+w\}$. The positions μ are taken modulo $L+1$ and $w\in\{0,\dots,L/2\}$ is an integer equal to the size of the *coupling window*. The coupled model is

$$w_{i_\mu j_\nu} = s_{i_\mu} s_{j_\nu}\sqrt{\frac{\Lambda_{\mu\nu}}{n}} + z_{i_\mu j_\nu}\sqrt{\Delta}, \tag{5}$$

where the index $i_\mu\in\{1,\dots,n\}$ (resp. j_ν) belongs to the block μ (resp. ν) along the ring, $\mathbf{\Lambda}$ is an $(L+1)\times(L+1)$ matrix which describes the strength of the coupling between blocks, and $Z_{i_\mu j_\nu}\sim\mathcal{N}(0,1)$ are i.i.d. For the proof to work, the matrix elements have to be chosen appropriately. We assume that: *i)* $\mathbf{\Lambda}$ is a doubly stochastic matrix; *ii)* $\Lambda_{\mu\nu}$ depends on $|\mu-\nu|$; *iii)* $\Lambda_{\mu\nu}$ is not vanishing for $|\mu-\nu|\le w$ and vanishes for $|\mu-\nu|>w$; *iv)* $\mathbf{\Lambda}$ is *smooth* in the sense $|\Lambda_{\mu\nu}-\Lambda_{\mu+1\nu}|=\mathcal{O}(w^{-2})$; *v)* $\mathbf{\Lambda}$ has a non-negative Fourier transform. All these conditions can easily be met, the simplest example being a triangle of base $2w+1$ and height $1/(w+1)$. The construction of the coupled system is completed by introducing a *seed* in the ring: we assume perfect knowledge of the signal components $\{s_{i_\mu}\}$ for $\mu\in\mathcal{B}:=\{-w-1,\dots,w-1\}$ mod $L+1$. This seed is what allows to close the gap between the algorithmic and information theoretical limits and therefore plays a crucial role. Note it can also be viewed as an "opening" of the chain with pinned boundary conditions.

Our first crucial result states that the mutual information $I_{w,L}(\mathbf{S};\mathbf{W})$ of the coupled and original systems are the same in a suitable asymptotic limit.

Lemma 6 (Equality of mutual informations) *For any $w\in\{0,\dots,L/2\}$ the following limits exist and are equal:* $\lim_{L\to+\infty}\lim_{n\to+\infty} I_{w,L}(\boldsymbol{S};\boldsymbol{W})/(n(L+1))=\lim_{n\to+\infty} I(\boldsymbol{S};\boldsymbol{W})/n$.

An immediate corollary is that non-analyticity points (w.r.t Δ) of the mutual informations are the same in the coupled and original models. In particular, defining $\Delta_{\text{Opt,coup}} := \sup\{\Delta \mid \lim_{L\to+\infty}\lim_{n\to+\infty} I_{w,L}(\mathbf{S};\mathbf{W})/(n(L+1)) \text{ is analytic in }]0,\Delta[\}$, we have $\Delta_{\text{Opt,coup}} = \Delta_{\text{Opt}}$.

The second crucial result states that the AMP threshold of the spatially coupled system is at least as good as Δ_{RS}. The analysis of AMP applies to the coupled system as well [Bayati and Montanari (2011); Javanmard and Montanari (2013)] and it can be shown that the performance of AMP is assessed by state evolution. Let $E_\mu^t := \lim_{n\to+\infty} \mathbb{E}_{\mathbf{S},\mathbf{Z}}[\|\mathbf{S}_\mu - \hat{\mathbf{s}}_\mu^t\|_2^2]/n$ be the asymptotic average vector-MSE of the AMP estimate $\hat{\mathbf{s}}_\mu^t$ at time t for the μ-th "block" of $\mathbf{S}$. We associate to each position $\mu \in \{0,\dots,L\}$ an *independent* scalar system with AWGN noise of the form $Y = S + \Sigma_\mu(\mathbf{E};\Delta)Z$ with $\Sigma_\mu(\mathbf{E};\Delta)^2 := \Delta/(v - \sum_{\nu=0}^L \Lambda_{\mu\nu}E_\nu)$ and $S \sim P_0$, $Z \sim \mathcal{N}(0,1)$. Taking into account knowledge of the signal in $\mathcal{B}$, state evolution reads:

$$E_\mu^{t+1} = \text{mmse}(\Sigma_\mu(\mathbf{E}^t;\Delta)^{-2}),\ E_\mu^0 = v \text{ for } \mu \in \{0,\dots,L\}\setminus\mathcal{B},\ E_\mu^t = 0 \text{ for } \mu \in \mathcal{B}, t \geq 0, \quad (6)$$

where the mmse function is defined as in section 1.2. From the monotonicity of the mmse function we have $E_\mu^{t+1} \leq E_\mu^t$ for all $\mu \in \{0,\dots,L\}$, a partial order which implies that $\lim_{t\to+\infty}\mathbf{E}^t = \mathbf{E}^\infty$ exists. This allows to define an algorithmic threshold: $\Delta_{\text{AMP},w,L} := \sup\{\Delta | E_\mu^\infty \leq E_{\text{good}}(\Delta)\ \forall\ \mu\}$. We show (equality holds but is not directly needed)

Lemma 7 (Threshold saturation) *Let* $\Delta_{\text{AMP,coup}} := \liminf_{w\to+\infty}\liminf_{L\to+\infty}\Delta_{\text{AMP},w,L}$. *We have* $\Delta_{\text{AMP,coup}} \geq \Delta_{\text{RS}}$.

Proof sketch of theorem 1 *First we prove* (3) *for* $\Delta \leq \Delta_{\text{Opt}}$. It is known [Deshpande and Montanari (2014)] that the matrix-MSE of AMP when $n\to+\infty$ is equal to $v^2 - (v-E^t)^2$. This cannot improve the matrix-MMSE, hence

$$\frac{1}{4}(v^2 - (v-E^\infty)^2) \geq \limsup_{n\to+\infty}\frac{1}{4n^2}\mathbb{E}_{\mathbf{S},\mathbf{W}}\|\mathbf{S}\mathbf{S}^\intercal - \mathbb{E}[\mathbf{X}\mathbf{X}^\intercal|\mathbf{W}]\|_{\text{F}}^2. \quad (7)$$

For $\Delta \leq \Delta_{\text{AMP}}$ we have $E^\infty = E_{\text{good}}(\Delta)$ which is the global minimum of (2) so the left hand side of (7) is equal to the derivative of $\min_{E\in[0,v]} i_{\text{RS}}(E;\Delta)$ w.r.t Δ^{-1}. Thus using a matrix version of the well known I-MMSE relation [Guo et al. (2005)] we get

$$\frac{d}{d\Delta^{-1}}\min_{E\in[0,v]} i_{\text{RS}}(E;\Delta) \geq \limsup_{n\to+\infty}\frac{1}{n}\frac{dI(\mathbf{S};\mathbf{W})}{d\Delta^{-1}}. \quad (8)$$

Integrating this relation on $[0,\Delta] \subset [0,\Delta_{\text{AMP}}]$ and checking that $\min_{E\in[0,v]} i_{\text{RS}}(E;0) = H(S)$ (the Shannon entropy of P_0) we obtain $\min_{E\in[0,v]} i_{\text{RS}}(E;\Delta) \leq \liminf_{n\to+\infty} I(\mathbf{S};\mathbf{W})/n$. But we know $I(\mathbf{S};\mathbf{W})/n \leq \min_{E\in[0,v]} i_{\text{RS}}(E;\Delta)$ [Krzakala et al. (2016)], thus we already get (3) for $\Delta \leq \Delta_{\text{AMP}}$. We notice that $\Delta_{\text{AMP}} \leq \Delta_{\text{Opt}}$. While this might seem intuitively clear, it follows from $\Delta_{\text{RS}} \geq \Delta_{\text{AMP}}$ (by their definitions) which together with $\Delta_{\text{AMP}} > \Delta_{\text{Opt}}$ would imply from (3) that $\lim_{n\to+\infty} I(\mathbf{S};\mathbf{W})/n$ is analytic at Δ_{Opt}, a contradiction. The next step is to extend (3) to the range $[\Delta_{\text{AMP}},\Delta_{\text{Opt}}]$. Suppose for a moment $\Delta_{\text{RS}} \geq \Delta_{\text{Opt}}$. Then *both* functions on each side of (3) are analytic on the whole range $]0,\Delta_{\text{Opt}}[$ and since they are equal for $\Delta \leq \Delta_{\text{AMP}}$, they *must* be equal on their whole analyticity range and by continuity, they must also be equal at Δ_{Opt} (that the functions are continuous follows from independent arguments on the existence of the $n\to+\infty$ limit of concave functions). It remains to show that $\Delta_{\text{RS}} \in]\Delta_{\text{AMP}},\Delta_{\text{Opt}}[$ is impossible. We proceed by contradiction, so suppose this is

true. Then both functions on each side of (3) are analytic on $]0,\Delta_{\rm RS}[$ and since they are equal for $]0,\Delta_{\rm AMP}[\subset]0,\Delta_{\rm RS}[$ they must be equal on the whole range $]0,\Delta_{\rm RS}[$ and also at $\Delta_{\rm RS}$ by continuity. For $\Delta>\Delta_{\rm RS}$ the fixed point of state evolution is $E^\infty=E_{\rm bad}(\Delta)$ which is also the global minimum of $i_{\rm RS}(E;\Delta)$, hence (8) is verified. Integrating this inequality on $]\Delta_{\rm RS},\Delta[\subset]\Delta_{\rm RS},\Delta_{\rm Opt}[$ and using $I(\mathbf{S};\mathbf{W})/n\le\min_{E\in[0,v]}i_{\rm RS}(E;\Delta)$ again, we find that (3) holds for all $\Delta\in[0,\Delta_{\rm Opt}]$. But this implies that $\min_{E\in[0,v]}i_{\rm RS}(E;\Delta)$ is analytic at $\Delta_{\rm RS}$, a contradiction.

We now prove (3) *for* $\Delta\ge\Delta_{\rm Opt}$. Note that the previous arguments showed that necessarily $\Delta_{\rm Opt}\le\Delta_{\rm RS}$. Thus by lemmas 6 and 7 (and the sub-optimality of AMP as shown as before) we obtain $\Delta_{\rm RS}\le\Delta_{\rm AMP,coup}\le\Delta_{\rm Opt,coup}=\Delta_{\rm Opt}\le\Delta_{\rm RS}$. This shows that $\Delta_{\rm Opt}=\Delta_{\rm RS}$ (this is the point where spatial coupling came in the game and we do not know of other means to prove such an equality). For $\Delta>\Delta_{\rm RS}$ we have $E^\infty=E_{\rm bad}(\Delta)$ which is the global minimum of $i_{\rm RS}(E;\Delta)$. Therefore we again have (8) in this range and the proof can be completed by using once more the integration argument, this time over the range $[\Delta_{\rm RS},\Delta]=[\Delta_{\rm Opt},\Delta]$.

Proof sketch of corollaries 3 and 5 Let $E_*(\Delta)=\mathrm{argmin}_E i_{\rm RS}(E;\Delta)$ for $\Delta\neq\Delta_{\rm RS}$. By explicit calculation one checks that $di_{\rm RS}(E_*,\Delta)/d\Delta^{-1}=(v^2-(v-E_*(\Delta))^2)/4$, so from theorem 1 and the matrix form of the I-MMSE relation we find $\mathrm{Mmmse}_n\to v^2-(v-E_*(\Delta))^2$ as $n\to+\infty$ which is the first part of the statement of corollary 3. Let us now turn to corollary 5. For $n\to+\infty$ the vector-MSE of the AMP estimator at time t equals E^t, and since the fixed point equation corresponding to state evolution is precisely the stationarity equation for $i_{\rm RS}(E;\Delta)$, we conclude that for $\Delta\notin[\Delta_{\rm AMP},\Delta_{\rm RS}]$ we must have $E^\infty=E_*(\Delta)$. It remains to prove that $E_*(\Delta)=\lim_{n\to+\infty}\mathrm{Vmmse}_n(\Delta)$ at least for $\Delta\notin[\Delta_{\rm AMP},\Delta_{\rm RS}]$ (we believe this is in fact true for all Δ). This will settle the second part of corollary 3 as well as 5. Using (Nishimori) identities $\mathbb{E}_{\mathbf{S},\mathbf{W}}[S_iS_j\mathbb{E}[X_iX_j|\mathbf{W}]]=\mathbb{E}_{\mathbf{S},\mathbf{W}}[\mathbb{E}[X_iX_j|\mathbf{W}]^2]$ (see e.g. [Krzakala et al. (2016)]) and the law of large numbers we can show $\lim_{n\to+\infty}\mathrm{Mmmse}_n\le\lim_{n\to+\infty}(v^2-(v-\mathrm{Vmmse}_n(\Delta))^2)$. Concentration techniques similar to [Korada and Macris (2009)] suggest that the equality in fact holds (for $\Delta\neq\Delta_{\rm RS}$) but there are technicalities that prevent us from completing the proof of equality. However it is interesting to note that this equality would imply $E_*(\Delta)=\lim_{n\to+\infty}\mathrm{Vmmse}_n(\Delta)$ for all $\Delta\neq\Delta_{\rm RS}$. Nevertheless, another argument can be used when AMP is optimal. On one hand the right hand side of the inequality is necessarily smaller than $v^2-(v-E^\infty)^2$. On the other hand the left hand side of the inequality is equal to $v^2-(v-E_*(\Delta))^2$. Since $E_*(\Delta)=E^\infty$ when $\Delta\notin[\Delta_{\rm AMP},\Delta_{\rm RS}]$, we can conclude $\lim_{n\to+\infty}\mathrm{Vmmse}_n(\Delta)=\mathrm{argmin}_E i_{\rm RS}(E;\Delta)$ for this range of Δ.

Proof sketch of lemma 6 Here we prove the lemma for a ring that is not seeded. An easy argument shows that a seed of size w does not change the mutual information per variable when $L\to+\infty$. The statistical physics formulation is convenient: up to a trivial additive term equal to $n(L+1)v^2/4$, the mutual information $I_{w,L}(\mathbf{S};\mathbf{W})$ is equal to the free energy $-\mathbb{E}_{\mathbf{S},\mathbf{Z}}[\ln\mathcal{Z}_{w,L}]$, where $\mathcal{Z}_{w,L}:=\int d\mathbf{x}P_0(\mathbf{x})\exp(-\mathcal{H}(\mathbf{x},\mathbf{z},\mathbf{\Lambda}))$ is the partition function

with Hamiltonian

$$\mathcal{H}(\mathbf{x},\mathbf{z},\mathbf{\Lambda}) = \frac{1}{\Delta}\sum_{\mu=0}^{L}\Lambda_{\mu\mu}\sum_{i_\mu\le j_\mu}\left(\frac{x_{i_\mu}^2 x_{j_\mu}^2}{2n} - \frac{s_{i_\mu}s_{j_\mu}x_{i_\mu}x_{j_\mu}}{n} - \frac{x_{i_\mu}x_{j_\mu}z_{i_\mu j_\mu}\sqrt{\Delta}}{\sqrt{n\Lambda_{\mu\mu}}}\right) + \frac{1}{\Delta}\sum_{\mu=0}^{L}\sum_{\nu=\mu+1}^{\mu+w}\Lambda_{\mu\nu}\sum_{i_\mu,j_\nu}\left(\frac{x_{i_\mu}^2 x_{j_\nu}^2}{2n} - \frac{s_{i_\mu}s_{j_\nu}x_{i_\mu}x_{j_\nu}}{n} - \frac{x_{i_\mu}x_{j_\nu}z_{i_\mu j_\nu}\sqrt{\Delta}}{\sqrt{n\Lambda_{\mu\nu}}}\right). \quad (9)$$

Consider a pair of systems with coupling matrices $\mathbf{\Lambda}$ and $\mathbf{\Lambda'}$ and i.i.d noize realizations $\mathbf{z},\mathbf{z'}$, an *interpolated Hamiltonian* $\mathcal{H}(\mathbf{x},\mathbf{z},t\mathbf{\Lambda})+\mathcal{H}(\mathbf{x},\mathbf{z'},(1-t)\mathbf{\Lambda'})$, $t\in[0,1]$, and the corresponding partition function $\mathcal{Z}_t$. The main idea of the proof is to show that for suitable choices of matrices, $-\frac{d}{dt}\mathbb{E}_{\mathbf{S},\mathbf{Z},\mathbf{Z'}}[\ln\mathcal{Z}_t]$ is negative for all $t\in[0,1]$ (up to negligible terms), so that by the fundamental theorem of calculus, we get a comparison between the free energies of $\mathcal{H}(\mathbf{x},\mathbf{z},\mathbf{\Lambda})$ and $\mathcal{H}(\mathbf{x},\mathbf{z'},\mathbf{\Lambda'})$. Performing the t-derivative brings down a Gibbs average of a polynomial in all variables s_{i_μ}, x_{i_μ}, $z_{i_\mu j_\nu}$ and $z'_{i_\mu j_\nu}$. This expectation over $\mathbf{S}$, $\mathbf{Z}$, $\mathbf{Z'}$ of this Gibbs average can be greatly simplified using integration by parts over the Gaussian noise $z_{i_\mu j_\nu}, z'_{i_\mu j_\nu}$ and Nishimori identities (see e.g. proof of corollary 3 for one of them). This algebra leads to

$$-\frac{1}{n(L+1)}\frac{d}{dt}\mathbb{E}_{\mathbf{S},\mathbf{Z},\mathbf{Z'}}[\ln\mathcal{Z}_t] = \frac{1}{4\Delta(L+1)}\mathbb{E}_{\mathbf{S},\mathbf{Z},\mathbf{Z'}}[\langle\mathbf{q}^\intercal\mathbf{\Lambda}\mathbf{q}-\mathbf{q}^\intercal\mathbf{\Lambda'}\mathbf{q}\rangle_t] + \mathcal{O}(1/(nL)), \quad (10)$$

where $\langle-\rangle_t$ is the Gibbs average w.r.t the interpolated Hamiltonian, $\mathbf{q}$ is the vector of overlaps $q_\mu := \sum_{i_\mu=1}^n s_{i_\mu}x_{i_\mu}/n$. If we can choose matrices such that $\mathbf{\Lambda'}>\mathbf{\Lambda}$, the difference of quadratic forms in the Gibbs bracket is negative and we obtain an inequality in the large size limit. We use this scheme to interpolate between the fully decoupled system $w=0$ and the coupled one $1\le w<L/2$ and then between $1\le w<L/2$ and the fully connected system $w=L/2$. The $w=0$ system has $\Lambda_{\mu\nu}=\delta_{\mu\nu}$ with eigenvalues $(1,1,\dots,1)$. For the $1\le w<L/2$ system, we take any stochastic translation invariant matrix with non-negative discrete Fourier transform (of its rows): such matrices have an eigenvalue equal to 1 and all others in $[0,1[$ (the eigenvalues are precisely equal to the discrete Fourier transform). For $w=L/2$ we choose $\Lambda_{\mu\nu}=1/(L+1)$ which is a projector with eigenvalues $(0,0,\dots,1)$. With these choices we deduce that the free energies and mutual informations are ordered as $I_{w=0,L}+\mathcal{O}(1)\le I_{w,L}+\mathcal{O}(1)\le I_{w=L/2,L}+\mathcal{O}(1)$. To conclude the proof we divide by $n(L+1)$ and note that the limits of the leftmost and rightmost mutual informations are equal, provided the limit exists. Indeed the leftmost term equals L times $I(\mathbf{S};\mathbf{W})$ and the rightmost term is the *same* mutual information for a system of $n(L+1)$ variables. Existence of the limit follows by a subadditivity inequality which itself is proven by a similar interpolation [Guerra (2005)].

Proof sketch of lemma 7 Fix $\Delta<\Delta_{\rm RS}$. We show that, for w large enough, the coupled state evolution recursion (6) must converge to a fixed point $E_\mu^\infty\le E_{\rm good}(\Delta)$ for all μ. The main intuition behind the proof is to use a "potential function" whose "energy" can be lowered by small perturbation of a fixed point that *would* go above $E_{\rm good}(\Delta)$ [Yedla et al. (2014); Barbier et al. (2016)]. The relevant potential function $i_{w,L}(\mathbf{E},\Delta)$ is in fact the replica potential of the coupled system, and equals up to a constant $(2w+1)Lv^2/4\Delta$

$$\sum_\mu\left\{\sum_{\nu=\mu-w}^{\mu+w}\frac{\Lambda_{\mu\nu}}{4\Delta}(v-E_\mu)(v-E_\nu) - \mathbb{E}_{S,Z}\left[\ln\left(\int dx\, P_0(x)e^{-\frac{x^2}{2\Sigma_\mu(\mathbf{E};\Delta)^2}+x\left(\frac{S}{\Sigma_\mu(\mathbf{E};\Delta)^2}+\frac{Z}{\Sigma_\mu(\mathbf{E};\Delta)}\right)}\right)\right]\right\}.$$

We note that the stationarity condition for this potential is precisely (6) (without the seeding condition). Monotonicity properties of state evolution ensure that any fixed point has a "unimodal" shape (and recall that it vanishes for $\mu \in \mathcal{B} = \{0, \dots, w-1\} \cup \{L-w, \dots, L\}$). Consider a position $\mu_{\max} \in \{w, \dots, L-w-1\}$ where it is maximal and suppose that $E^{\infty}_{\mu_{\max}} > E_{\text{good}}(\Delta)$. We associate to the *fixed point* $\mathbf{E}^{\infty}$ a so-called *saturated profile* $\mathbf{E}^{\text{s}}$ defined on the whole of $\mathbb{Z}$ as follows: $E^{\text{s}}_{\mu} = E_{\text{good}}(\Delta)$ for all $\mu \leq \mu_{\infty}$ where $\mu_{\infty}+1$ is the smallest position such that $E^{\infty}_{\mu} > E_{\text{good}}(\Delta)$; $E^{\text{s}}_{\mu} = E^{\infty}_{\mu}$ for $\mu \in \{\mu_{\infty}+1, \dots, \mu_{\max}-1\}$; $E^{\text{s}}_{\mu} = E^{\infty}_{\mu_{\max}}$ for all $\mu \geq \mu_{\max}$. We show that $\mathbf{E}^{\text{s}}$ cannot exist for w large enough. To this end define a shift operator by $[\mathcal{S}(\mathbf{E}^{\text{s}})]_{\mu} := E^{\text{s}}_{\mu-1}$. On one hand the shifted profile is a *small* perturbation of $\mathbf{E}^{\text{s}}$ which matches a fixed point, except where it is constant, so if we Taylor expand, the first order vanishes and the second order and higher orders can be estimated as $|i_{w,L}(\mathcal{S}(\mathbf{E}^{\text{s}}); \Delta) - i_{w,L}(\mathbf{E}^{\text{s}}; \Delta)| = \mathcal{O}(1/w)$ uniformly in L. On the other hand, by explicit cancellation of telescopic sums $i_{w,L}(\mathcal{S}(\mathbf{E}^{\text{s}}); \Delta) - i_{w,L}(\mathbf{E}^{\text{s}}; \Delta) = i_{\text{RS}}(E_{\text{good}}; \Delta) - i_{\text{RS}}(E^{\infty}_{\mu_{\max}}; \Delta)$. Now one can show from monotonicity properties of state evolution that if $\mathbf{E}^{\infty}$ is a fixed point then $E^{\infty}_{\mu_{\max}}$ *cannot be in the basin of attraction of* $E_{\text{good}}(\Delta)$ for the uncoupled recursion. Consequently as can be seen on the plot of $i_{\text{RS}}(E; \Delta)$ (e.g. figure 1) we must have $i_{\text{RS}}(E^{\infty}_{\mu_{\max}}; \Delta) \geq i_{\text{RS}}(E_{\text{bad}}; \Delta)$. Therefore $i_{w,L}(\mathcal{S}(\mathbf{E}^{\text{s}}); \Delta) - i_{w,L}(\mathbf{E}^{\text{s}}; \Delta) \leq -|i_{\text{RS}}(E_{\text{bad}}; \Delta) - i_{\text{RS}}(E_{\text{good}}; \Delta)|$ which is an energy gain independent of w, and for large enough w we get a contradiction with the previous estimate coming from the Taylor expansion.

Acknowledgments

J.B and M.D acknowledge funding from the Swiss National Science Foundation (grant num. 200021-156672). Part of the research has received funding from the European Research Council under the European Union's 7th Framework Programme (FP/2007-2013/ERC Grant Agreement 307087-SPARCS). This work was done in part while F.K and L.Z were visiting the Simons Institute for the Theory of Computing.

References

A.A. Amini and M.J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *IEEE Int. Symp. on Inf. Theory*, page 2454, 2008.

J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, page 1643, 2005.

J. Barbier, M. Dia, and N. Macris. Threshold saturation of spatially coupled sparse superposition codes for all memoryless channels. *CoRR*, abs/1603.04591, 2016. URL `http://arxiv.org/abs/1603.04591`.

M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. on Inf. Theory*, 57(2):764 –785, 2011.

Q. Berthet and P. Rigollet. Computational lower bounds for sparse pca. *arXiv:1304.0828*, 2013.

E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G.RG. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434, 2007.

Y. Deshpande and A. Montanari. Information-theoretically optimal sparse pca. In *IEEE Int. Symp. on Inf. Theory*, pages 2197–2201, 2014.

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

Y. Deshpande, E. Abbe, and A. Montanari. Asymptotic mutual information for the two-groups stochastic block model. *arXiv:1507.08685*, 2015.

F. Guerra. An introduction to mean field spin glass theory: methods and results. *Mathematical Statistical Physics*, pages 243–271, 2005.

D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE Trans. on Inf. Theory*, 51, 2005.

S.H. Hassani, N. Macris, and R. Urbanke. Coupled graphical models and their thresholds. In *IEEE Information Theory Workshop (ITW)*, 2010.

A. Javanmard and A. Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *J. Infor. & Inference*, 2:115, 2013.

I.M. Johnstone and A.Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 2012.

S.B. Korada and N. Macris. Exact solution of the gauge symmetric p-spin glass model on a complete graph. *Journal of Statistical Physics*, 136(2):205–230, 2009.

F. Krzakala, J. Xu, and L. Zdeborová. Mutual information in rank-one matrix estimation. *arXiv:1603.08447*, 2016.

S. Kudekar, T.J. Richardson, and R. Urbanke. Threshold saturation via spatial coupling: Why convolutional ldpc ensembles perform so well over the bec. *IEEE Trans. on Inf. Theory*, 57, 2011.

T. Lesieur, F. Krzakala, and L. Zdeborová. Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *Annual Allerton Conference*, 2015a.

T. Lesieur, F. Krzakala, and L. Zdeborová. Phase transitions in sparse pca. In *IEEE Int. Symp. on Inf. Theory*, page 1635, 2015b.

S. Rangan and A.K. Fletcher. Iterative estimation of constrained rank-one matrices in noise. In *IEEE Int. Symp. on Inf. Theory*, pages 1246–1250, 2012.

A. Yedla, Y.Y. Jian, P.S. Nguyen, and H.D. Pfister. A simple proof of maxwell saturation for coupled scalar recursions. *IEEE Trans. on Inf. Theory*, 60(11):6943–6965, 2014.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.