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

Authored by:

Florent Krzakala
Lenka Zdeborov?
Jean Barbier
Mohamad Dia
Nicolas Macris
Thibault Lesieur

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.

1 Paper Body

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 w = (wij)ni,j=1 of the pair-wise product of the components of a vector s = (s1 , . . . , sn)— ? Rn with i.i.d components distributed as Si ? P0 , i = 1, . . . , n. The entries of w are observed ? through a noisy element-wise (possibly non-linear) output probabilistic channel Pout (wij —si sj / n). The goal is to estimate the vector s from w assuming that both P0 and Pout are known and independent of n (noise is symmetric so that wij = wji). Many important problems in statistics and machine learning can be expressed in this way, such as sparse PCA [1], the Wigner spiked model [2, 3], community detection [4] or matrix completion [5]. Proving a result initially derived by a heuristic method from statistical physics, we give an explicit expression for the mutual information (MI) and the information theoretic minimal mean-square-error (MMSE) in the asymptotic n ? ? limit. Our results imply that for a large region of parameters, the posterior marginal expectations of the underlying signal components (often assumed intractable 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

to compute) can be obtained in the leading order in n using a polynomial-time algorithm called approximate message-passing (AMP) [6, 3, 4, 7]. We also demonstrate the existence of a region where both AMP and spectral methods [8] 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

The additive white Gaussian noise setting: A standard and natural setting ? is the case of additive ? white Gaussian noise (AWGN) of known variance ?, wij = si sj / n+zij ?, where z = (zij)ni,j=1 is a symmetric matrix with i.i.d entries Zij ? N (0, 1), 1 ? i ? j ? 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 [9] (already proven for community detection in [4] and conjectured in [10]). This theorem states that given an

output channel Pout (w—y), such that (s.t) log Pout (w—y = 0) is three times differentiable with bounded second and third derivatives, then the MI satisfies I(S; W) = ? ? — ? I(S; SS / n+Z ?)+O(n), where ? is the inverse Fisher information (evaluated at y = 0) of ?1 := the output channel: ? EPout (w—0) [(?y log Pout (W —y)—y=0)2]. Informally, this means that we only have to compute the MI 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 P0 , an explicit one-letter formula for the MI per variable I(S; W)/n in the asymptotic limit n ? ?. Main result: Our central result is a proof of the expression for the asymptotic n ? ? MI per variable via the so-called replica symmetric (RS) potential iRS (E; ?) defined as Z

x2 S Z (v ? E)2 + v 2 ? 2?(E;?) 2 +x ?(E;?)2 + ?(E;?) iRS (E; ?) := ? ES,Z ln , (1) dx P0 (x)e 4? with Z ? N (0, 1), S ? P0 , E[S 2] = v and ?(E; ?)2 := ?/(v?E), E ? [0, P v]. Here we will assume ? that P0 is a discrete distribution over a finite bounded real alphabet P0 (s) = ?=1 p? ?(s?a?). Thus the only continuous integral in (1) 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 theoretic and algorithmic AMP thresholds are determined by the set of stationary points of (1) (w.r.t E). It is possible to show that for all ? ¿ 0 there always exist at least one stationary minimum. Note E = 0 is never a stationary point (except for P0 a single Dirac mass) and E = v is stationary only if 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 1.1 (RS formula for the mutual information) Fix ? ¿ 0 and let P0 be a discrete distribution s.t (1) has at most three stationary points. Then limn?? I(S; W)/n = minE?[0,v] iRS (E; ?). The proof of the existence of the limit does not require the above hypothesis on P0 . Also, it was first shown in [9] that for all n, I(S; W)/n ? minE?[0,v] iRS (E; ?), an inequality that we will use in the proof section. It is conceptually useful to define the following threshold: Definition 1.2 (Information theoretic threshold) Define ?Opt as the first non-analyticity point of the MI as ? increases: ?Opt := sup{?— limn?? I(S; W)/n is analytic in]0, ?[}. When P0 is s.t (1) has at most three stationary points, as discussed below, then minE?[0,v] iRS (E; ?) has at most one non-analyticity point denoted ?RS (if minE?[0,v] iRS (E; ?) is analytic over all R+ we set ?RS = ?). Theorem 1.1 gives us a mean to compute the information theoretic threshold ?Opt = ?RS . A basic application of theorem 1.1 is the expression of the MMSE: Corollary 1.3 (Exact formula for the MMSE) For all ? 6= ?RS , the matrix-MMSE Mmmsen := ES,W [kSS— ? E[XX— —W]k2F]/n2 (k ? kF being the Frobenius norm) is asymptotically limn?? Mmmsen (??1) = v 2 ?(v ?argminE?[0,v] iRS (E; ?))2 . Moreover, if ? ¡ ?AMP (where ?AMP is the algorithmic threshold, see definition 1.4) or ? ¿ ?RS , then the usual vector-MMSE Vmmsen := ES,W [kS?E[X—W]k22]/n satisfies limn?? Vmmsen = argminE?[0,v] iRS (E; ?). 2

It is natural to conjecture that the vector-MMSE is given by argminE?[0,v]

iRS (E; ?) for all ? 6= ?RS , but our proof does not quite yield the full statement. A fundamental consequence concerns the performance of the AMP algorithm [6] for estimating s. AMP has been analysed rigorously in [11, 12, 4] where it is shown that its asymptotic performance is tracked by state evolution (SE). Let E t := limn?? ES,Z [kS??st k22]/n be the asymptotic average vector-MSE of the AMP estimate ?st at time t. Define mmse(??2) := ES,Z [(S ?E[X—S +?Z])2] as the usual scalar mmse function associated to a scalar AWGN channel of noise variance ?2 , with S ? P0 and Z ? N (0, 1). Then E t+1 = mmse(?(E t ; ?)?2),

E 0 = v,

(2)

is the SE recursion. Monotonicity properties of the mmse function imply that E t is a decreasing sequence s.t limt?? E t = E ? exists. Note that when E[S] = 0 and v is an unstable fixed point, as such, SE ?does not start?. While this is not really a problem when one runs AMP in practice, for analysis purposes one can slightly bias P0 and remove the bias at the end of the proofs. Definition 1.4 (AMP algorithmic threshold) For ? ¿ 0 small enough, the fixed point equation corresponding to (2) has a unique solution for all noise values in]0, ?[. We define ?AMP as the supremum of all such ?. Corollary 1.5 (Performance of AMP) In the limit n ? ?, AMP initialized without any knowledge other than P0 yields upon convergence the asymptotic matrix-MMSE as well as the asymptotic vector-MMSE iff ? ¡ ?AMP or ? ¿ ?RS , namely E ? = argminE?[0,v] iRS (E; ?). ?AMP can be read off the replica potential (1): by differentiation of (1) one finds a fixed point equation that corresponds to (2). Thus ?AMP is the smallest solution of ?iRS /?E = ? 2 iRS /?E 2 = 0; in other words it is the ?first? horizontal inflexion point appearing in iRS (E; ?) when ? increases. Discussion: With our hypothesis on P0 there are only three possible scenarios: ?AMP ¡ ?RS (one ?first order? phase transition); ?AMP = ?RS ¡ ? (one ?higher order? phase transition); ?AMP = ?RS = ? (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 [3] and follow as a special case from our proof. The only cases that would require more work are those where P0 is s.t (1) develops more than three stationary points and more than one phase transition is present. For ?AMP ¡ ?RS the structure of stationary points of (1) is as follows1 (figure 1). There exist three branches Egood (?), Eunstable (?) and Ebad (?) s.t: 1) For 0 ¡ ? ¡ ?AMP there is a single stationary point Egood (?) which is a global minimum; 2) At ?AMP a horizontal inflexion point appears, for ? ? [?AMP , ?RS] there are three stationary points satisfying Egood (?AMP) ¡ Eunstable (?AMP) = Ebad (?AMP), Egood (?) ¡ Eunstable (?) ¡ Ebad (?) otherwise, and moreover iRS (Egood ; ?) ? iRS (Ebad ; ?) with equality only at ?RS ; 3) for ? ¿ ?RS there is at least the stationary point Ebad (?) which is always the global minimum, i.e. iRS (Ebad ; ?) ¡ iRS (Egood ; ?). (For higher ? the Egood (?) and Eunstable (?) branches may merge and disappear); 4) Egood (?) is analytic for ? ?]0,

?0 [, ?0 ¿ ?RS , and Ebad (?) is analytic for ? ¿ ?AMP . We note for further use in the proof section that E ? = Egood (?) for ? ¡ ?AMP and E ? = Ebad (?) for ? ¿ ?AMP . Definition 1.4 is equivalent to ?AMP = sup{?—E ? = Egood (?)}. Moreover we will also use that iRS (Egood ; ?) is analytic on]0, ?0 [, iRS (Ebad ; ?) is analytic on]?AMP , ?[, and the only non-analyticity point of minE?[0,v] iRS (E; ?) is at ?RS . Relation to other works: Explicit single-letter characterization of the MI in the rank-one problem has attracted a lot of attention recently. Particular cases of theorem 1.1 have been shown rigorously in a number of situations. A special case when si = ?1 ? Ber(1/2) already appeared in [13] where an equivalent spin glass model is analysed. Very recently, [9] has generalized the results of [13] and, notably, obtained a generic matching upper bound. The same formula has been also rigorously computed following the study of AMP in [3] for spiked models (provided, however, that the signal was not too sparse) and in [4] for strictly symmetric community detection. 1

We take E[S] 6= 0. Once theorem 1.1 is proven for this case a limiting argument allows to extend it to E[S] = 0.

3
iRS(E)
0.125 0.12 0.115 0.11 0.105 0.1 0.095
0.086 0.085 0.084 0.083 ?=0.0008 0
0.005
0.01
0.082 0.08 0
0.005
0.01 E
0.015
?=0.0012
0.082 0.02
?=0.00125
0.084 iRS(E)
0.015
0.02
0.08 0.078 0.076 0.074 0.072 0.07 0.068 0.066
0
0.005
0.01
0.015
0.02
?=0.0015
0
0.005
0.01 E
0.015
0.02

Figure 1: The replica symmetric potential iRS (E) for four values of ? in the Wigner spiked model. The MI is min iRS (E) (the black dot, while the black cross corresponds to the local minimum) and the asymptotic matrix-MMSE is v 2 ?(v?argminE iRS (E))2 , where v = ? in this case with ? = 0.02 as in the inset of figure 2. From top left to bottom right: (1) For low noise values, here ? = 0.0008 ¡ ?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 ?AMP ¡ ? = 0.0012 ¡ ?RS . (3) For ? = 0.00125 ¿ ?RS , the ?bad? minimum becomes the global one and the MMSE suddenly deteriorates. (4) For larger values of ?, only the ?bad? minimum exists. AMP 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 ?AMP ¡ ? ¡ ?RS , it is trapped by the local minimum with large MSE instead of reaching the global one corresponding to the MMSE.

For rank-one symmetric matrix estimation problems, AMP has been introduced by [6], who also computed the SE formula to analyse its performance, generalizing techniques developed by [11] and [12]. SE was further studied by [3] and [4]. In [7, 10], the generalization to larger rank was also considered. The general formula proposed by [10] 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 [9]). 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 capacityachieving spatially coupled codes [14, 15, 16, 17] with other progress, coming from developments in mathematical physics putting on a rigorous basis predictions of spin glass theory [18]. 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 spiked model and community detection

In order to illustrate the consequences of our results we shall present two examples. Wigner spiked model: In this model, the vector s is a Bernoulli random vector, Si ? Ber(?). For large enough densities (i.e. ? ¿ 0.041(1)), [3] 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 ?. Our results allow to close the gap left open by [3]: on one hand we now obtain rigorously the MMSE for ? ? 0.041(1), and on the other one we observe that for such values of ?, and as ? decreases, there is a small region where two local minima coexist in iRS (E; ?). In particular for ?AMP ¡ ? ¡ ?Opt = ?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 ? our 4

Wigner Spike model 0.005
Asymmetric Community Detection
matrix-MSE(?) at ?=0.02 0.0004
0.004
MMSE AMP ?AMP
0.003
0.0003 ?opt
0.0002 0.0001
? 0.002
0
0.001
?
0 0.002
0.001
?AMP ?Opt ?spectral
0 0
0.01
0.02
0.03
0.04
0.05
3 2.8 2.6 2.4 2.2 2 1.8 1.6 1.4 1.2 1 0.8 0.6 0.4 0.2 0
matrix-MSE(?) at ?=0.05 1 0.8 MMSE 0.6 AMP 0.4 ?AMP 0.2 0 0
0.5
1
1.5
2
2.5
?AMP ?Opt ?spectral 0
?
?opt
0.1
0.2
0.3
0.4

0.5
?

Figure 2: Phase diagram in the noise variance ? versus density ? plane for the rank-one spiked Wigner model (left) and the asymmetric community detection (right). Left: [3] proved that AMP achieves the matrix-MMSE for all ? as long as ? ¿ 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 ? = 0.02 as a function of ?. AMP (dashed red) provably achieves the matrix-MMSE except in the region ?AMP ¡ ? ¡ ?Opt = ?RS . We conjecture that no polynomial-time algorithm will do better than p AMP in this region. Right: Asymmetric community detection problem with two communities. For ? ¿ 1/2? 1/12 (black point) and when ? ¿ 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 ? ¡ 1. In this region, AMP is always achieving the matrix-MMSEp and spectral methods can find a non-trivial overlap with the truth as well, starting from ? ¡ 1. For ? ¡ 1/2? 1/12, however, it is information theoretically possible to find an overlap with the hidden communities for ? ¿ 1 (below the blue line) but both AMP and spectral methods miss this information. Inset: matrix-MMSE (blue) at ? = 0.05 as a function of ?. AMP (dashed red) again provably achieves the matrix-MMSE except in the region ?AMP ¡ ? ¡ ?Opt .

results are consistent with the known optimal and algorithmic thresholds predicted in sparse PCA [19, 20], that treats the case of sub-extensive ? = O(1) values. Another interesting line of work for such probabilistic models appeared in the context of random matrix theory (see [8] and references therein) and predicts that a sharp phase transition occurs at a critical value of the noise ?spectral = ?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. Asymmetric balanced community detection: We now consider the problem of detecting two communities (groups) with different sizes ?n and (1 ? ?)n, that generalizes the one considered in [4]. One is given ? a graph where the probability to have a link between? nodes in the first group is p + ?(1 ? ?)/(? n), between those in the ? second group is p + ??/(n(1 ? ?)), while interconnections appear with probability p ? ?/ 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 the first section, this is equivalent to a 2 model with AWGN according to p of variance ? = p(1p? p)/? where each variable si is chosen 2 P0 (s) = ??(s? (1??)/?)+(1??)?(s+ ?/(1??)). Our results for this problem are summarized p on the right hand side of figure 2. For ? ¿ ?c = 1/2 ? 1/12 (black point), it is asymptotically information theoretically possible to get an estimation better than chance if and only if ? ¡ 1. When ? ¡ ?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 ? ¡ 1, regardless of the value of ?. Again, a region [?AMP , ?Opt = ?RS] exists where a computational gap appears when ? ¡ ?c . One can investigate the very low ? regime where we find that the information theoretic transition goes as ?Opt (? ? 0) = 1/(4?— log ?—). Now if we assume that this result ? stays true even for ? = O(1) (which is a speculation at this point), we can choose ? ? (1?p)? n such that the small group is a clique. Then the problem corresponds to a ?balanced? version of the famous planted clique problem [21]. We find that the AMP/spectral approach finds the 2

Note that here since E = v = 1 is an extremum of iRS (E; ?), one must introduce a small bias in P0 and let it then tend to zero at the end of the proofs.

5

p hidden clique when it is larger than 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 pclassical planted clique problem at p = 1/2 with its gap between log(n) (information theoretic), n/e (AMP [22]) ? and n (spectral [21]). 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 theoretic limits is eliminated, while at the same time the MI is maintained unchanged for the coupled and original models. Roughly speaking, this means that it is possible to algorithmically compute the information theoretic limit of the original model because a suitable algorithm is optimal on the coupled system. The present spatially coupled construction is similar to the one used for the coupled Curie-Weiss model [14]. Consider a ring of length L+1 (L even) with blocks positioned at ? ? {0, . . . , L} and coupled to neighboring blocks {??w, . . . , ?+w}. Positions ? are taken modulo L+1 and the integer w ? {0, . . . , L/2} equals the size of the coupling window. The coupled model is r ? ??? wi? j? = si? sj? (3) + zi? j? ?, n where the index i? ? {1, . . . , n} (resp. j?) belongs to the block ? (resp. ?) along the ring, ? is an (L+1)?(L+1) matrix which describes the strength of the coupling between blocks, and Zi? j? ? N (0, 1) are i.i.d. For the proof to work, the matrix elements have to be chosen appropriately. We assume that: i) ? is a doubly stochastic matrix; ii) ??? depends on —???—; iii) ??? is not vanishing for —???— ? w and vanishes for —???— ¿ w; iv) ? is smooth in the sense —??? ???+1? — = O(w?2); v) ? 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 {si? } for ? ? B := {?w?1, . . . , w?1} mod L+1. This seed is what allows to close the gap between the algorithmic and information theoretic limits and therefore plays a crucial role. Note it can also be viewed as an ?opening? of the chain with fixed boundary conditions. Our first crucial result states that the MI Iw,L (S; W) of the coupled and original systems are the same in a suitable limit. Lemma 3.1 (Equality of mutual informations) For any fixed w the following limits exist and are equal: limL?? limn?? Iw,L (S; W)/(n(L+1)) = limn?? I(S; W)/n. An immediate corollary is that non-analyticity points (w.r.t ?) of the MIs are the same in the coupled and original models. In particular, defining ?Opt,coup := sup{? — limL?? limn?? Iw,L (S; W)/(n(L+1)) is analytic in]0, ?[}, we have ?Opt,coup = ?Opt . The second crucial result states that the AMP threshold of the spatially coupled system is at least as good as ?RS . The analysis of AMP applies to the coupled system as well [11, 12] and it can be shown that the performance of AMP is assessed by SE. Let E?t := limn?? ES,Z [kS? ??st? k22]/n be the asymptotic average vector-MSE of the AMP estimate ?st? at time t for the ?-th ?block? of S. We associate to each position ? ? {0, . . . , L} an independent scalar system with AWGN of the form PL Y = S +?? (E; ?)Z, with ?? (E; ?)2 := ?/(v ? ?=0 ??? E?) and S ? P0 , Z ? N (0, 1). Taking into account knowledge of the signal components in B, SE reads: E?t+1 = mmse(?? (Et ; ?)?2), E?0 = v for ? ? {0, . . . , L} B, E?t = 0 for ? ? B, t ? 0,

(4)

where the mmse function is defined as in section 1. From the monotonicity of the mmse function we have E?t+1 ? E?t for all ? ? {0, . . . , L}, a partial order which implies that limt?? Et = E? exists. This allows to define an algorithmic threshold for the coupled system: ?AMP,w,L := sup{?—E?? ? Egood (?) ? ?}. We show (equality holds but is not directly needed): Lemma 3.2 (Threshold saturation) Let ?AMP,coup := lim inf w?? lim inf L?? ?AMP,w,L . We have ?AMP,coup ? ?RS . 6

Proof sketch of theorem 1.1: First we prove the RS formula for ? ? ?Opt . It is known [3] that the matrix-MSE of AMP when n ? ? is equal to v 2 ?(v?E t)2 . This cannot improve the matrix-MMSE, hence (v 2 ?(v ?E ?)2)/4 ? lim supn?? Mmmsen /4. For ? ? ?AMP we have E ? = Egood (?) which is the global minimum of (1) so the left hand side of the last inequality equals the derivative of minE?[0,v] iRS (E; ?) w.r.t ??1 . Thus using the matrix version of the I-MMSE relation [23] we get d 1 dI(S; W) min iRS (E; ?) ? lim sup . (5) ?1 d??1 E?[0,v] n?? n d? Integrating this relation on [0, ?] ? [0, ?AMP] and checking that minE?[0,v] iRS (E; 0) = H(S) (the Shannon entropy of P0) we obtain minE?[0,v] iRS (E; ?) ? lim inf n?? I(S; W)/n. But we know I(S; W)/n ? minE?[0,v] iRS (E; ?) [9], thus we already get theorem 1.1 for ? ? ?AMP . We notice that ?AMP ? ?Opt . While this might seem intuitively clear, it follows from ?RS ? ?AMP (by their definitions) which together with ?AMP ¿ ?Opt would imply from theorem 1.1 that limn?? I(S; W)/n is analytic at ?Opt , a contradiction. The next step is to extend theorem 1.1 to the range [?AMP , ?Opt]. Suppose for a moment ?RS ? ?Opt . Then both functions on each side of the RS formula are analytic on the whole range]0, ?Opt [and since

they are equal for ? ? ?AMP , they must be equal on their whole analyticity range and by continuity, they must also be equal at ?Opt (that the functions are continuous follows from independent arguments on the existence of the n ? ? limit of concave functions). It remains to show that ?RS ?]?AMP , ?Opt [is impossible. We proceed by contradiction, so suppose this is true. Then both functions on each side of the RS formula are analytic on]0, ?RS [and since they are equal for]0, ?AMP [?]0, ?RS [they must be equal on the whole range]0, ?RS [and also at ?RS by continuity. For ? ¿ ?RS the fixed point of SE is E ? = Ebad (?) which is also the global minimum of iRS (E; ?), hence (5) is verified. Integrating this inequality on]?RS , ?[?]?RS , ?Opt [and using I(S; W)/n ? minE?[0,v] iRS (E; ?) again, we find that the RS formula holds for all ? ? [0, ?Opt]. But this implies that minE?[0,v] iRS (E; ?) is analytic at ?RS , a contradiction. We now prove the RS formula for ? ? ?Opt . Note that the previous arguments showed that necessarily ?Opt ? ?RS . Thus by lemmas 3.1 and 3.2 (and the sub-optimality of AMP as shown as before) we obtain ?RS ? ?AMP,coup ? ?Opt,coup = ?Opt ? ?RS . This shows that ?Opt = ?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 ? ¿ ?RS we have E ? = Ebad (?) which is the global minimum of iRS (E; ?). Therefore we again have (5) in this range and the proof can be completed by using once more the integration argument, this time over the range [?RS , ?] = [?Opt , ?]. Proof sketch of corollaries 1.3 and 1.5: Let E? (?) = argminE iRS (E; ?) for ? 6= ?RS . By explicit calculation one checks that diRS (E? , ?)/d??1 = (v 2 ?(v?E? (?))2)/4, so from theorem 1.1 and the matrix form of the I-MMSE relation we find Mmmsen ? v 2 ?(v?E? (?))2 as n ? ? which is the first part of the statement of corollary 1.3. Let us now turn to corollary 1.5. For n ? ? the vectorMSE of the AMP estimator at time t equals E t , and since the fixed point equation corresponding to SE is precisely the stationarity equation for iRS (E; ?), we conclude that for ? ? / [?AMP , ?RS] we must have E ? = E? (?). It remains to prove that E? (?) = limn?? Vmmsen (?) at least for ? ? / [?AMP , ?RS] (we believe this is in fact true for all ?). This will settle the second part of corollary 1.3 as well as 1.5. Using (Nishimori) identities ES,W [Si Sj E[Xi Xj —W]] = ES,W [E[Xi Xj —W]2] (see e.g. [9]) and using the law of large numbers we can show limn?? Mmmsen ? limn?? (v 2 ? (v ? Vmmsen (?))2). Concentration techniques similar to [13] suggest that the equality in fact holds (for ? 6= ?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? (?) = limn?? Vmmsen (?) for all ? 6= ?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 ?)2 . On the other hand the left hand side of the inequality is equal to v 2?(v?E? (?))2 . Since E? (?) = E ? when ? ? / [?AMP , ?RS], we can conclude limn?? Vmmsen (?) = argminE iRS (E; ?) for this range of ?. Proof sketch of lemma 3.1: 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 MI per variable when L ? ?. The statistical physics formulation is convenient: up to the trivial additive term n(L+1)v 2 /4, the MI Iw,L (S; W) equals the R := free energy

?ES,Z [ln Zw,L], where Zw,L dxP0 (x) exp(?H(x, z, ?)) and ?+w L

X X X X 1 H(x, z, ?) = ??? Ai? j? (x, z, ?) + ??? Ai? j? (x, z, ?) , (6) ? ?=0 ?=?+1 i ,j i? ?j?

?

7

?

? p with Ai? j? (x, z, ?) := (x2i? x2j?)/(2n)?(si? sj? xi? xj?)/n?(xi? xj? zi? j? ?)/ n??? . Consider a pair of systems with coupling matrices ? and ?0 and i.i.d noize realizations z, z0 , an interpolated Hamiltonian H(x, z, t?)+H(x, z0 , (1?t)?0), t ? [0, 1], and the corresponding partition function Zt . d ES,Z,Z0 [ln Zt] ? 0 for all The main idea of the proof is to show that for suitable choices of matrices, ? dt t ? [0, 1] (up to negligible terms), so that by the fundamental theorem of calculus, we get a comparison between the free energies of H(x, z, ?) and H(x, z0 , ?0). Performing the t-derivative brings down a Gibbs average of a polynomial in all variables si? , xi? , zi? j? and zi0? j? . This expectation over S, Z, Z0 of this Gibbs average is simplified using integration by parts over the Gaussian noise zi? j? , zi0? j? and Nishimori identities (see e.g. proof of corollary 1.3 for one of them). This algebra leads to ?

1 d 1 ES,Z,Z0 [ln Zt] = ES,Z,Z0 [hq— ?q ? q— ?0 qit] + O(1/(nL)), n(L + 1) dt 4?(L + 1)

(7)

where P h?it is the Gibbs average w.r.t the interpolated Hamiltonian, q is the vector of overlaps n q? := i? =1 si? xi? /n. If we can choose matrices s.t ?0 ¿ ?, 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 ? w ¡ L/2 and then between 1 ? w ¡ L/2 and the fully connected system w = L/2. The w = 0 system has ??? = ??? with eigenvalues (1, 1, . . . , 1). For the 1 ? 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 ??? = 1/(L + 1) which is a projector with eigenvalues (0, 0, . . . , 1). With these choices we deduce that the free energies and MIs are ordered as Iw=0,L + O(1) ? Iw,L + O(1) ? Iw=L/2,L + O(1). To conclude the proof we divide by n(L+1) and note that the limits of the leftmost and rightmost MIs are equal, provided the limit exists. Indeed the leftmost term equals L times I(S; W) and the rightmost term is the same MI for a system of n(L+1) variables. Existence of the limit follows by subadditivity, proven by a similar interpolation [18]. Proof sketch of lemma 3.2: Fix ? ¡ ?RS . We show that, for w large enough, the coupled SE recursion (4) must converge to a fixed point E?? ? Egood (?) for all ?. 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 Egood (?) [16, 17]. The relevant potential function iw,L (E, ?) is in fact the replica potential of the coupled system (a generalization of (1)). The stationarity condition for this potential is precisely (4) (without the seeding condition). Monotonicity properties of SE ensure that

any fixed point has a ?unimodal? shape (and recall that it vanishes for ? ? B = {0, . . . , w?1} ? {L?w, . . . , L}). Consider a position ?max ? {w, . . . , L?w?1} where it is maximal and suppose that E??max ¿ Egood (?). We associate to the fixed point E? a so-called saturated profile Es defined on the whole of Z as follows: E?s = Egood (?) for all ? ? ?? where ?? +1 is the smallest position s.t E?? ¿ Egood (?); E?s = E?? for ? ? {?? +1, . . . , ?max ?1}; E?s = E??max for all ? ? ?max . We show that Es cannot exist for w large enough. To this end define a shift operator by s [S(Es)]? := E??1 . On one hand the shifted profile is a small perturbation of Es 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 —iw,L (S(Es); ?)?iw,L (Es ; ?)— = O(1/w) uniformly in L. On the other hand, by explicit cancellation of telescopic sums iw,L (S(Es); ?)?iw,L (Es ; ?) = iRS (Egood ; ?)?iRS (E??max ; ?). Now one can show from monotonicity properties of SE that if E? is a non trivial fixed point of the coupled SE then E??max cannot be in the basin of attraction of Egood (?) for the uncoupled SE recursion. Consequently as can be seen on the plot of iRS (E; ?) (e.g. figure 1) we must have iRS (E??max ; ?) ? iRS (Ebad ; ?). Therefore iw,L (S(Es); ?)?iw,L (Es ; ?) ? ?—iRS (Ebad ; ?)?iRS (Egood ; ?)— 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 SNSF (grant 200021-156672). Part of this research received funding from the ERC under the EU?s 7th Framework Programme (FP/2007-2013/ERC Grant Agreement 307087-SPARCS). F.K and L.Z thank the Simons Institute for its hospitality.

8

2 References

[1] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. Journal of computational and graphical statistics, 15(2):265?286, 2006. [2] 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. [3] Y. Deshpande and A. Montanari. Information-theoretically optimal sparse pca. In IEEE Int. Symp. on Inf. Theory, pages 2197?2201, 2014. [4] Y. Deshpande, E. Abbe, and A. Montanari. Asymptotic mutual information for the two-groups stochastic block model. arXiv:1507.08685, 2015. [5] E.J. Cand?s and B. Recht. Exact matrix completion via convex optimization. Foundations of Computational mathematics, 9(6):717?772, 2009. [6] 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. [7] T. Lesieur, F. Krzakala, and L. Zdeborov?. Phase transitions in sparse pca. In IEEE Int. Symp. on Inf. Theory, page 1635, 2015. [8] J. Baik, G. Ben Arous, and S. P?ch?. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Annals of Probability, page 1643, 2005. [9] F. Krzakala,

J. Xu, and L. Zdeborov?. Mutual information in rank-one matrix estimation. arXiv:1603.08447, 2016. [10] T. Lesieur, F. Krzakala, and L. Zdeborov?. Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In Annual Allerton Conference, 2015. [11] 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. [12] A. Javanmard and A. Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. J. Infor. & Inference, 2:115, 2013. [13] 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. [14] S.H. Hassani, N. Macris, and R. Urbanke. Coupled graphical models and their thresholds. In IEEE Information Theory Workshop (ITW), 2010. [15] 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. Th., 57, 2011. [16] 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. [17] J. Barbier, M. Dia, and N. Macris. Threshold saturation of spatially coupled sparse superposition codes for all memoryless channels. CoRR, abs/1603.04591, 2016. [18] F. Guerra. An introduction to mean field spin glass theory: methods and results. Mathematical Statistical Physics, pages 243?271, 2005. [19] 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. [20] Q. Berthet and P. Rigollet. Computational lower bounds for sparse pca. arXiv:1304.0828, 2013. [21] 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. p [22] Y. Deshpande and A. Montanari. Finding hidden cliques of size N/e in nearly linear time. Foundations of Computational Mathematics, 15(4):1069?1128, 2015. [23] D. Guo, S. Shamai, and S. Verd?. Mutual information and minimum mean-square error in gaussian channels. IEEE Trans. on Inf. Theory, 51, 2005. 9