Phase Diagram and Approximate Message Passing for Blind Calibration and Dictionary Learning

Florent Krzakala
ESPCI and CNRS UMR 7083
10 rue Vauquelin,
Paris 75005 France
fk@espci.fr

Marc Mézard
Ecole Normale Supérieure
45 rue d'Ulm, Paris France
and LPTMS-CNRS Univ. Paris Sud
Orsay, France

Lenka Zdeborová
Institut de Physique Théorique
IPhT, CEA Saclay
and URA 2306, CNRS
91191 Gif-sur-Yvette, France.

***Abstract*—We consider dictionary learning and blind calibration for signals and matrices created from a random ensemble. We study the mean-squared error in the limit of large signal dimension using the replica method and unveil the appearance of phase transitions delimiting impossible, possible-but-hard and possible inference regions. We also introduce an approximate message passing algorithm that asymptotically matches the theoretical performance, and show through numerical tests that it performs very well, for the calibration problem, for tractable system sizes.**

I. Introduction

Matrix factorization $Y = FX$, where Y is known and one seeks F and X, with requirements (e.g. sparsity, positivity, probability distribution of elements etc.) on the properties of X and F, is a relatively new but a generic problem that appears in many fields of science and engineering. A long but incomplete survey can be found in [1]. Theoretical limits on when matrix factorization is possible and tractable are still very poorly understood. In this work we make a step towards this understanding by determining the limits of matrix factorization when Y is created using randomly generated matrices F and X.

Consider a set of P sparse N-dimensional vectors ("signals"), with iid components generated as follows. Each vector component x^0_{il} (where $i = 1, \dots, N$, $l = 1, \dots, P$) is chosen independently, equal to 0 with probability$1 - \rho$, and drawn with probability ρ from a continuous density distribution ϕ. The parameter ρ controls the sparsity of the vectors x^0.

For each of the P vectors we perform M linear measurements, summarized by a $M \times N$ measurement matrix with iid elements $\mathcal{F}^0_{\mu i} = F^0_{\mu i}/\sqrt{N}$, where $F^0_{\mu i}$ are iid random variables drawn from a Gaussian distribution of zero mean and unit variance. We have only access to the (noisy) results of these measurements, that is, to the P vectors y_l such that

$$y_{\mu l} = \sum_i \mathcal{F}^0_{\mu i} x^0_{il} + \xi_{\mu l} \, , \tag{1}$$

where $\xi_{\mu l}$ is a Gaussian additive noise with variance Δ.

Is it possible to find both the vectors x^0 and the matrix (dictionary) F^0 (up to a permutation of N elements and their signs)? This is the *dictionary learning* problem. A related situation is when one knows at least a noisy version of the matrix F^0, defined by $F' = (F^0 + \sqrt{\eta} W)/\sqrt{1+\eta}$ where W is a random matrix with the same statistics as F^0. $P(F^0|F')$ then reads, for each matrix element

$$P(F^0_{\mu i}|F'_{\mu i}) = \mathcal{N}(\frac{F'_{\mu i}}{\sqrt{1+\eta}}, \frac{\eta}{1+\eta}) \, . \tag{2}$$

Recovering F^0 and x^0, knowing this time F' and the P vectors y_l is a problem that we shall refer to as *blind calibration*. It becomes equivalent to dictionary learning when $\eta \to \infty$.

Our goal here is to analyse optimal Bayes inference (that provides the MMSE (minimal MSE)) where the signal x^0_{il} and the dictionary $F^0_{\mu i}$ are estimated from the marginals of the posterior probability

$$P(x_{il}, F_{\mu i}|y_{\mu l}, F'_{\mu i}) = \frac{1}{Z} \prod_{\mu i} P(F_{\mu i}|F'_{\mu i})$$

$$\prod_{il} P(x_{il}) \prod_{\mu l} \left[\frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{(y_{\mu l} - \sum_i F_{\mu i} x_{il}/\sqrt{N})^2}{2\Delta}} \right] \, . \tag{3}$$

A. Related works

There are several algorithms suggested and tested for dictionary learning, see e.g. [2], [3], [4], [5]. The algorithm we derive in this paper is closely related (but different) to the bilinear AMP proposed by [6] as explained in Sec. III.

The question of how many samples P are necessary for the dictionary to be identifiable in the noiseless case has a straightforward lower bound $MP > N(M + P\rho)$, otherwise there is more unknown variables than measurements and hence exact recovery is clearly impossible. Several works analyzed what is a sufficient number of samples for exact recovery. While early rigorous results were able to show learnability from only exponential many samples [7], more recent analysis of convex relaxation based approaches suggests that $O(N \log N)$ samples are needed [8], [9], [10], [11]. A very recent non-rigorous work suggested that $P = O(N)$ samples should be sufficient to identify the dictionary [12]. That work was based on the replica analysis of the problem, but did not analyze the Bayes-optimal approach.

Several works also considered blind calibration, where only an uncertain version of the matrix F is known, on the other hand one has the access to many signals and their measurements such that calibration of the matrix F is possible, see e.g. [13] and reference therein. Cases when both the signal and the

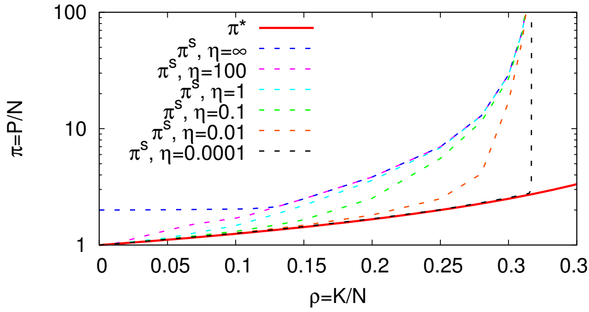

Fig. 1. Phase diagram for $\alpha = 0.5$ and $\Delta = 0$. For both blind calibration and dictionary learning, exact learning is possible by the Bayes optimal approach above the full red line $\pi^* = \alpha/(\alpha - \rho)$. However, such a sampling procedures will not be tractable below the spinodal transition $\pi^s(\eta)$ shown here in dotted line for dictionary learning ($\eta = \infty$) and blind calibration.

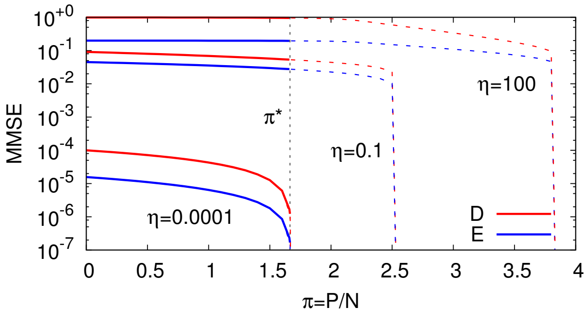

Fig. 2. MMSE D (for the matrix F) and E (for the signal x) corresponding to $\rho = 0.2$, $\alpha = 0.5$, $\Delta = 0$, for three values of η. The MMSE jumps abruptly from a finite value to zero at π^*. However, sampling should remain intractable until the spinodal transition arises at a larger value $\pi^s(\eta)$ and we denote the corresponding MSE in dotted line. The figure remains qualitatively the same for small value of the additive noise $\Delta > 0$, where the MMSE at large π is not zero but rather $O(\Delta)$. If Δ is large enough, however, the sharp transition disappears and the MMSE is continuous (see e.g. Fig. 4).

dictionary are sparse are also considered in the literature, e.g. [14], and our theory can be applied to these as well.

B. *Main results*

The present paper has three main results. First, using the replica method [15] we estimate the Bayes optimal MMSE in the limit of large signals $N \to \infty$. In particular, we define $\alpha = M/N$, $\pi = P/N$ and show that for the noiseless case, $\Delta = 0$, exact reconstruction is possible if $\pi > \pi^* = \alpha/(\alpha - \rho)$ and $\alpha > \rho$. In this regime, it is thus possible to recover the matrix and the signal exactly if one can compute the marginals of the posterior probability distribution (3). This result is striking, all the more because it is independent of η.

Computing the marginals of the posterior probability distribution is an extremely hard problem, all the more when there is a phase transition in the problem. We determine the value $\pi^s(\eta)$ (the spinodal transition) below which iterative sampling (using for instance Monte Carlo Markov chains or message passing) is believed to be intractable in polynomial time.

Finally, we introduce an AMP-like message passing algorithm designed to perform such sampling, and show that it performs very well for the calibration problem. However, at the moderate size we are able to study numerically, finite-size deviations from the asymptotic behavior become large as η grows and prevent our algorithm to function nicely in the dictionary learning limit. However, we believe that this still sets a very promising stage for new algorithmic development for dictionary learning.

II. Asymptotic analysis with the replica method

A. *Replica analysis for matrix factorization*

The MMSE obtained from the Bayes-optimal approach can be computed exactly in the limit of large N via the replica method. Although this method is in general non-rigorous, it is sometimes possible to prove that the results derived from it are exact. We shall leave out details of the derivation and refer instead to [16], [17] for a very similar computation in the case of compressed sensing. We estimate the Bayes optimal MMSE by computing the marginals of the matrix and signal elements. Our computation is also very similar to the one of [12], who however did not analyze the Bayes optimal MMSE.

We now compute $\Phi = \mathbb{E}_{F^0,F',x^0}(\log Z(F^0,F',x^0))/NP$, where Z, the so-called partition sum (or evidence), is the normalization in eq. (3) for a given instance of the problem. In order to do so we compute $\mathbb{E}_{F^0,F',x^0}(Z^n)$ and then use the replica trick $\log Z = \lim_{n\to 0}(Z^n - 1)/n$. We use the replica symmetric ansatz which is correct for inference problems with prior distributions corresponding to the generative process. The final result is that in the large signal limit, $N \to \infty$, the MMSE D (on the matrix) and E (on the signals) are given by

$$\mathrm{MMSE}(\alpha,\pi,\rho,\Delta,\eta) = \underset{E,D}{\mathrm{argmax}}\,\Phi(E,D)\,, \tag{4}$$

where the so-called "potential" is given by

$$\Phi(E,D) = -\frac{\alpha}{2}\log\left(\Delta + E + D(\rho - E)\right) - \frac{\alpha(\Delta+\rho)}{\Delta + E + D(\rho - E)} + \frac{\alpha}{2} + \left[\int \mathcal{D}z \log\left\{\left[e^{-\frac{\hat{m}_x}{2}x^2 + \hat{m}_x x x^0 + z\sqrt{\hat{m}_x}x}\right]_{P(x)}\right\}\right]_{P(x^0)} + \frac{\alpha}{\pi}\left[\int \mathcal{D}z \log\left[e^{-\frac{\hat{m}_F F^2}{2} + \hat{m}_F F F^0 + z\sqrt{\hat{m}_F}F}\right]_{P(F|F')}\right]_{P(F^0,F')} .$$

Here $[f(u)]_{Q(u)}$ denotes an average of a function f of the random variable u with distribution $Q(u)$, $\mathcal{D}z$ a Gaussian measure with zero mean and unit variance. The probability distributions $P(x)$, and $P(F|F')$ are the single-element distributions introduced in eqs. (**??**) and (2). Finally we denoted

$$\hat{m}_x = \frac{\alpha(1-D)}{\Delta + E + \rho D - ED}, \quad \hat{m}_F = \frac{\pi(\rho - E)}{\Delta + E + \rho D - ED}\,. \tag{5}$$

Note that the present expression for the potential is very general and can be used to study many similar problems, such as matrix completion, or sparse matrix factorization, by changing the distribution of the matrix and of the signal.

B. Gaussian matrix and Gauss-Bernoulli signal

When the matrix elements $F^0_{\mu i}$ are generated from a Gaussian with zero mean and unit variance, and x^0_{il} from Gauss-Bernoulli distribution, the potential simplifies to

$$\begin{aligned}\Phi(E,D)=&-\frac{\alpha}{2}\log\left(\Delta+E+D(\rho-E)\right)-\frac{\alpha(\Delta+\rho)}{\Delta+E+D(\rho-E)}\\ +\;& \frac{\alpha}{2}+(1-\rho)\int\mathcal{D}z\,\log\left(1-\rho+\frac{\rho}{\sqrt{\hat{m}_x+1}}e^{\frac{z^2\hat{m}_x}{2(\hat{m}_x+1)}}\right)\\ +\;& \rho\int\mathcal{D}z\log\left(1-\rho+\frac{\rho}{\sqrt{\hat{m}_x+1}}e^{\frac{z^2\hat{m}_x}{2}}\right)+\frac{\alpha}{2\pi}\hat{m}_F\\ -\;& \frac{\alpha}{2\pi}\log\left(1+\frac{\eta\hat{m}_F}{1+\eta}\right).\end{aligned}\tag{6}$$

The Bayes-optimal MMSE is obtained by maximizing $\Phi(E,D)$. Analyzing the above expression in the zero-noise limit ($\Delta=0$) allows to demonstrate our first main result: in both the blind calibration and dictionary learning problems, the global maximum is given by $D=E=0$, with $\Phi\to\infty$, as long as $\alpha>\rho$ and $\pi>\pi^*=\alpha/(\alpha-\rho)$. Hence for $\pi>\pi^*$ it is possible to learn the matrix and the signal exactly from the measurements. This result is striking since it is independent of η as long as $\eta>0$; and coincides with the simple counting lower bound. When $\eta\to 0$, more precisely when $\eta\ll\Delta$, then the compressed sensing phase transition —that goes to $\alpha=\rho$ when $\Delta\to 0$— is recovered independently of π.

Bayes-optimal learning, however, requires exact sampling from the measure (3), and this remains an extremely hard computational problem. In this regard, another important transition (the "spinodal"), that can be studied from the form of the potential function $\Phi(E,D)$, is the appearance of a high-MSE local maxima. This phenomenon marks the downfall of many sampling strategy, such as Gibbs sampling or message-passing strategy (that performs a steepest ascent in the $\Phi(E,D)$ function). We determine the value $\pi^s(\eta)$ (the so-called "spinodal" transition) above which $\Phi(E,D)$ does not have the spurious secondary maxima. As shown in Fig. 1 and Fig. 2, it depends strongly on the value η. In fact, when $\eta\to 0$, we recover again the compressed sensing limit we have obtained in [17]. Figs. 2 and 4 depict the values of MMSE for the signal E and the matrix D reached for large systems $N\to\infty$ by the Bayes optimal sampling and by local sampling strategies that reach $D=E=0$ discontinuously at π^s.

The behavior with finite noise $\Delta<\eta$ is also interesting and we observe a phenomenology similar to that described in [17] in the case of compressed sensing. Because of a two-maxima shape of the function $\Phi(E,D)$ for moderate Δ, the MMSE displays a sharp transition separating a region of parameters with a small MMSE, comparable to Δ, from a region with a larger $O(\eta)$ MMSE. For larger value of Δ we do not see any abrupt transition, and the MMSE continuously decays with π (see e.g. Fig. 4).

To conclude, there exist three regions in the phase diagram, corresponding to impossible (below π^*), intractable (below π^s), and perhaps-tractable learning. We will now study algorithmically the later one.

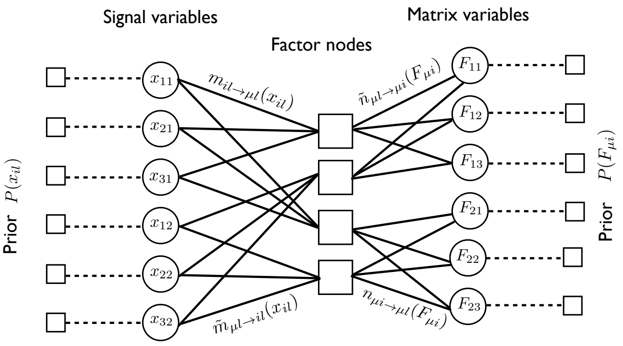

Fig. 3. Factor graph used for the belief propagation inference, here drawn using $N=3$, $P=2$ and $M=2$. The factor nodes ensure (in probability) the condition $y_{\mu l}=\sum_i \mathcal{F}_{\mu i}x_{il}+\xi_{\mu l}$.

III. Message-passing algorithm

To make the Bayes optimal sampling tractable we have to resort to approximations. In compressed sensing, the Bayesian approach combined with a belief propagation (BP) reconstruction algorithm leads to the so-called approximate message passing (AMP) algorithm. It was first derived in [18] for the minimization of ℓ_1, and subsequently generalized in [19], [20]. We shall now adapt this strategy to the present case.

A. From BP to AMP

The factor graph corresponding to the posterior probability (3) is depicted in Fig. 3. The canonical BP iterative equations are written for messages $m,n,\hat{m},\hat{n}$ and read

$$m_{il\to\mu l}(x_{il}) \propto P(x_{il})\prod_{\nu\neq\mu}^{M}\hat{m}_{\nu l\to il}(x_{il})\,,\tag{7}$$

$$n_{\mu i\to\mu l}(F_{\mu i}) \propto P(F_{\mu i}|F'_{\mu i})\prod_{n\neq l}^{P}\hat{n}_{\mu n\to\mu i}(F_{\mu i})\,,\tag{8}$$

$$\begin{aligned}\hat{m}_{\mu l\to il}(x_{il}) \propto& \int\prod_{j\neq i}dx_{jl}dF_{\mu k}e^{-\frac{(y_{\mu l}-\sum_i F_{\mu i}x_{il}/\sqrt{N})^2}{2\Delta}}\\ &\prod_k n_{\mu k\to\mu l}(F_{\mu k})\prod_{j\neq i}m_{jl\to\mu l}(x_{jl})\,,\end{aligned}\tag{9}$$

$$\begin{aligned}\hat{n}_{\mu l\to\mu i}(F_{\mu i}) \propto& \int dx_{jl}\prod_{k\neq i}dF_{\mu k}e^{-\frac{(y_{\mu l}-\sum_i F_{\mu i}x_{il}/\sqrt{N})^2}{2\Delta}}\\ &\prod_{k\neq i} n_{\mu k\to\mu l}(F_{\mu k})\prod_{j}m_{jl\to\mu l}(x_{jl})\,.\end{aligned}\tag{10}$$

A major simplification of these iterative equations arises when one uses the central limit theorem and realizes that only the two first moments of the above distributions are important for the leading contribution when $N\to\infty$. This "Gaussian" approximation is at the basis of approximate message passing as used in compressed sensing. The next step of the derivation again neglects $O(1/N)$ terms and allows to reduce the number of messages to be iterated from $O(N^4)$ to $O(N^2)$. This leads to a TAP-like set of equations [21]. Finally if the matrix F^0 and signal x^0 have known distribution of elements this further simplifies the algorithm. A full derivation will be given elsewhere and we instead refer the reader to the re-derivation of AMP in [17] where we followed essentially the very same steps.

The final form of our algorithm for blind calibration and dictionary learning follows. We denote the mean and variance of the BP estimates of marginals over x_{il} as a_{il} and c_{il}, and those over $\mathcal{F}_{\mu i}$ as $r_{\mu i}$ and $s_{\mu l}$. We define several auxiliary values (all of them are of order $O(1)$):

$$\begin{aligned}
\overline{a^2} &= \frac{1}{NP}\sum_{jl} a_{jl}^2, \quad \overline{c} = \frac{1}{NP}\sum_{jl} c_{jl}\,, \\
\overline{r^2} &= \frac{1}{M}\sum_{\nu j} r_{\nu j}^2, \quad \overline{s} = \frac{1}{M}\sum_{\nu j} c_{\nu j}\,, \\
\overline{(y-\omega)^2} &= \frac{1}{MP}\sum_{\nu l}(y_{\nu l}-\omega_{\nu l})^2\,.
\end{aligned}$$

Then our AMP algorithm reads

$$\begin{aligned}
\omega_{\nu l}^{t+1} &= \sum_j r_{\nu j} a_{jl}^t - \frac{y_{\nu l}-\omega_{\nu l}^t}{\overline{(y-\omega)^2}^t}(\overline{c}^t\overline{r^2}^t + \overline{a^2}^t\overline{s}^t), && (11) \\
(\Sigma_R^{t+1})^2 &= \left[\alpha\frac{\overline{r^2}^t}{\overline{(y-\omega)^2}^{t+1}}\right]^{-1}, && (12) \\
(\Sigma_S^{t+1})^2 &= \left[\pi\frac{\overline{a^2}^t}{\overline{(y-\omega)^2}^{t+1}}\right]^{-1}, && (13) \\
R_{il}^{t+1} &= a_{il}^t - a_{il}^t\frac{\overline{s}^t}{\overline{r^2}^t} + \frac{\sum_\nu (y_{\nu l}-\omega_{\nu l}^{t+1})r_{\nu i}^t}{\alpha\overline{r^2}^t}, && (14) \\
S_{\nu i}^{t+1} &= r_{\nu i}^t - r_{\nu i}^t\frac{\overline{c}^t}{\overline{a^2}^t} + \frac{\sum_l (y_{\nu l}-\omega_{\nu l}^{t+1})a_{il}^t}{N\pi\overline{a^2}^t}, && (15) \\
a_{il}^{t+1} &= f_a\left((\Sigma_R^{t+1})^2, R_{il}^{t+1}\right), && (16) \\
c_{il}^{t+1} &= f_c\left((\Sigma_R^{t+1})^2, R_{il}^{t+1}\right), && (17) \\
r_{\mu i}^{t+1} &= f_r\left((\Sigma_S^{t+1})^2, S_{\mu i}^{t+1}\right), && (18) \\
s_{\mu i}^{t+1} &= f_s\left((\Sigma_S^{t+1})^2, S_{\mu i}^{t+1}\right). && (19)
\end{aligned}$$

where only the following functions are prior-dependent:

$$\begin{aligned}
f_a(\Sigma^2,T) &= \frac{\rho\, e^{-\frac{T^2}{2(\Sigma^2+1)}}\frac{\Sigma}{(\Sigma^2+1)^{\frac{3}{2}}}(\Sigma^2+T)}{(1-\rho)e^{-\frac{T^2}{2\Sigma^2}} + \rho\frac{\Sigma}{\sqrt{\Sigma^2+1}}e^{-\frac{T^2}{2(\Sigma^2+1)}}}, && (20) \\
f_c(\Sigma^2,T) &= \Sigma^2\frac{\mathrm{d}}{\mathrm{d}T}f_a(\Sigma^2,T)\,, && (21) \\
f_r(\Sigma^2,T) &= \frac{T+\Sigma^2 F'_{\mu i}\frac{\sqrt{1+\eta}}{\sqrt{N}\eta}}{(1+\frac{1}{\eta})\Sigma^2+1}\,, && (22) \\
f_s(\Sigma^2,T) &= \frac{1}{N}\frac{\Sigma^2}{(1+\frac{1}{\eta})\Sigma^2+1}\,. && (23)
\end{aligned}$$

Initial conditions are set so that the marginals correspond to the means and variances of the prior, and $\omega_{\mu l} = y_{\mu l}$. One iteration of the algorithm takes $O(N^3)$ steps. In practice, we also damp the expressions (16, 17, 18, 19) to ensure convergence. If the matrix elements are not learned, this algorithm reduces to the AMP for matrix uncertainty from [22], and is asymptotically equivalent to the MU-AMP of [23]. The authors of [6] suggested a *bilinear AMP* algorithm for matrix factorization. Their algorithm does not include the second terms from eq. (14) and (15). Whereas the difference in performances between the present algorithm and the one of [6] is yet to be studied, it is not clear if the later implements asymptotically the Bayes optimal inference, and neither if the state evolution (next paragraph) applies to it.

B. State evolution

The AMP approach is amenable to asymptotic ($N \to \infty$) analysis using a method known as "cavity method" in statistical physics [15], or "state evolution" in compressed sensing [18]. Given the parameters ρ, α,π η, Δ, the MSE given by our approach follows, in the infinite size limit:

$$\begin{aligned}
E^{t+1} &= (1-\rho)\int \mathcal{D}z f_c(\hat{m}_x^t, z\sqrt{\hat{m}_x^t}) \\
&+ \rho\int \mathcal{D}z f_c(\hat{m}_x^t, z\sqrt{\hat{(}m_x^t)^2+\hat{m}_x^t})\,, && (24) \\
D^{t+1} &= \frac{1}{\hat{m}_F^t + \frac{1+\eta}{\eta}}\,, && (25)
\end{aligned}$$

where $\hat{m}_F^t$ and $\hat{m}_x^t$ follow from eq. (5), with E^t, D^t on the right hand side, $E^{t=0} = \rho$, $D^{t=0} = 1$, $\mathcal{D}z$ is a Gaussian integral, and f_c is defined by eq. (21).

From eqs. (24, 25), one can show that the evolution of the algorithm is equivalent to a steepest ascent of the potential $\phi(E,D)$ obtained in eq. (6). This explains the peculiar meaning of the spinodal transition arising at π^s and shows that for $\pi > \pi^s$ our algorithm should approximates correctly the Bayes optimal inference for matrix factorization for large systems, $N \to \infty$, as AMP does for compressed sensing. Note that in the later case, the state evolution approach has been proven rigorously [24] and would be interesting to see if this could be generalize to the present, arguably more complex, case.

C. Numerical tests

We have tested our algorithm on instances of tractable sizes. The results are shown in Fig. 4 for two noisy cases of matrix calibration. The agreement with the theoretical large N prediction is excellent for small π. In the large π region, however, we observe finite-size corrections going roughly as η/N. Despite these finite size effect, the MSE reached by the algorithm is excellent in both case. To appreciate the performance of the algorithm, note that a ℓ_1-minimization would give very poor results even with a perfectly known matrix, as the values of α and ρ are above the Donoho-Tanner transition [25].

The presence of $O(\eta/N)$ corrections prevents us from using our algorithm successfully for large η. This means that so far we are not able to solve efficiently the dictionary learning. More work will be needed to reduce these effects.

IV. Perspectives

It would be interesting to see if our result for the MMSE and the exactness of the state evolution can be proven rigorously,

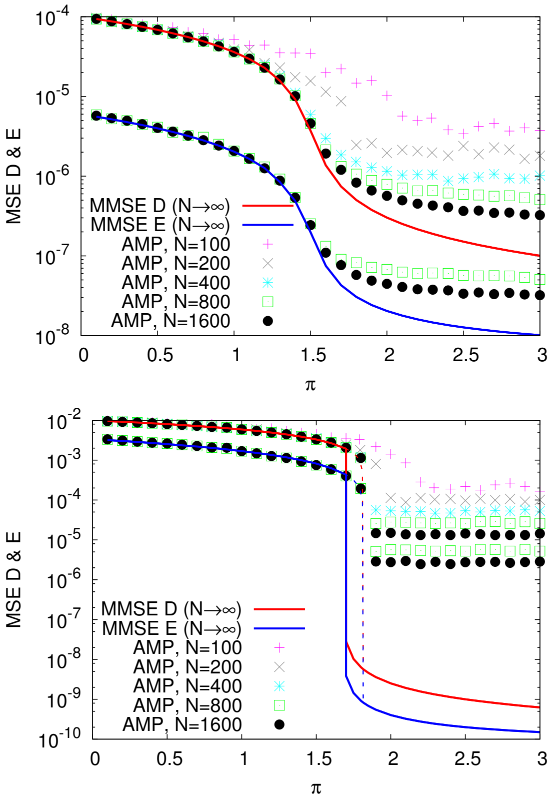

Fig. 4. Comparison of the performance of the AMP algorithm with the MMSE D (for the matrix) and E (for the signal, fewer points are shown for visibility) for different system sizes N. Top: A case with continuous decay of the MMSE for $\Delta = 10^{-8}$, $\eta = 10^{-4}$, $\alpha = 0.3$, $\rho = 0.1$. The initial error on the matrix is about 1%. Bottom: A case with a jump in the MMSE for $\Delta = 10^{-8}$, $\eta = 10^{-2}$, $\alpha = 0.5$, $\rho = 0.2$, the initial error on the matrix is about 10%. As N increases, the MSE found by the algorithm approaches the MMSE computed theoretically. In the large π region, corrections to the asymptotic behavior are roughly proportional to η/N.

as in compressed sensing [26], [24]. Further, it is important to investigate if our algorithm can be improved and the finite size effects reduced. One can also generalize our approach to other matrix factorization problems and their applications.

Acknowledgment

After completing this study, we became aware that [27] also analyzed the Bayes-optimal inference for dictionary learning and reached the same conclusions as in our Sec. II for $\eta \to \infty$.

This work has been supported in part by the ERC under the European Unions 7th Framework Programme Grant Agreement 307087-SPARCS, by the EC Grant STAMINA, No. 265496, and by the Grant DySpaN of Triangle de la Physique.

References

[1] I. Carron, "The matrix factorization jungle," https://sites.google.com/site/igorcarron2/matrixfactorizations.

[2] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1," *Vision Research*, vol. 37, p. 3311, 1997.

[3] K. Engan, S. O. Aase, and H. J. Husoy, "Method of optimal directions for frame design," *IEEE Acoustic, Speech and Signal Processing*, vol. 5, p. 2443, 1999.

[4] M. Aharon, M. Elad, and A. M. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, p. 4311, 2006.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 689–696.

[6] P. Schniter, J. Parker, and V. Cevher, "Bilinear generalized approximate message passing (big-amp) for matrix recovery problem," in *Workshop on Information Theory and Applications (ITA), San Diego CA*, 2012.

[7] A. M. B. Michal Aharon, Michael Elad, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear Algebra and its Applications*, vol. 416, pp. 48–67, 2006.

[8] R. Gribonval and K. Schnass, "Dictionary identification - sparse matrix-factorisation via ℓ_1-minimisation," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[9] Q. Geng, H. Wang, and J. Wright, "On the local correctness of ℓ_1-minimization for dictionary learning," 2011, arXiv:1101.5672.

[10] D. Vainsencher, S. Mannor, and A. M. Bruckstein, "The sample complexity of dictionary learning," *Journal of Machine Learning Research*, vol. 12, pp. 3259–3281, 2011.

[11] R. Jenatton, R. Gribonval, and F. Bach, "Local stability and robustness of sparse dictionary learning in the presence of noise," *arXiv:1210.0685*, 2012.

[12] A. Sakata and Y. Kabashima, "Statistical mechanics of dictionary learning," 2012, arXiv:1203.6178.

[13] R. Gribonval, G. Chardon, and L. Daudet, "Blind calibration for compressed sensing by convex optimization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2713 –2716.

[14] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Transactions on Signal Processing*, vol. 58, p. 1553, 2010.

[15] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin-Glass Theory and Beyond*. Singapore: World Scientific, 1987, vol. 9.

[16] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Statistical physics-based reconstruction in compressed sensing," *Phys. Rev. X*, vol. 2, p. 021005, 2012.

[17] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," *J. Stat. Mech.*, 2012.

[18] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[19] D. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *IEEE Information Theory Workshop (ITW)*, 2010, pp. 1 –5.

[20] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2011, pp. 2168 –2172.

[21] D. J. Thouless, P. W. Anderson, and R. G. Palmer, "Solution of 'solvable model of a spin-glass'," *Phil. Mag.*, vol. 35, pp. 593–601, 1977.

[22] F. Krzakala, M. Mézard, and L. Zdeborová, "Compressed sensing under matrix uncertainty: Optimum thresholds and robust approximate message passing," 2012, arXiv:1301.0901 [cs.IT], ICASSP 2013.

[23] J. T. Parker, V. Cevher, and P. Schniter, "Compressive sensing under matrix uncertainties: An approximate message passing approach," in *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2011, pp. 804–808.

[24] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764 –785, 2011.

[25] D. L. Donoho and J. Tanner, "Sparse nonnegative solution of underdetermined linear equations by linear programming," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9446–9451, 2005.

[26] Y. Wu and S. Verdu, "Optimal phase transitions in compressed sensing," 2011, arXiv:1111.6822v1 [cs.IT].

[27] A. Sakata and Y. Kabashima, "Sample complexity of bayesian optimal dictionary learning," 2013, arXiv:1301.6199.