scientific reports

OPEN

Inverse problems for structured datasets using parallel TAP equations and restricted Boltzmann machines

Aurelien Decelle[1,2], Sungmin Hwang[3], Jacopo Rocchi[3] & Daniele Tantari[4]✉

We propose an efficient algorithm to solve inverse problems in the presence of binary clustered datasets. We consider the paradigmatic Hopfield model in a teacher student scenario, where this situation is found in the retrieval phase. This problem has been widely analyzed through various methods such as mean-field approaches or the pseudo-likelihood optimization. Our approach is based on the estimation of the posterior using the Thouless–Anderson–Palmer (TAP) equations in a parallel updating scheme. Unlike other methods, it allows to retrieve the original patterns of the teacher dataset and thanks to the parallel update it can be applied to large system sizes. We tackle the same problem using a restricted Boltzmann machine (RBM) and discuss analogies and differences between our algorithm and RBM learning.

Inverse problems consist in inferring information about the structure of a system from the observation of its configurations. Cases where the system's variables s_i are binary can be studied in the framework of the inverse Ising model, whose parameters $\{J_{ij}, h_i\}$ are tuned in order to describe the observed configurations according to the Boltzmann weight $P(s) \sim \exp(\sum_{i<j} J_{ij} s_i s_j + \sum_i h_i s_i)$. This is the simplest distribution emerging when using the maximum entropy approach in order to reproduce exactly the one and two points statistics of the data. Successful applications of this method arise in biology[1], immunology[2], neurosciences[3,4] as in the study of collective behaviors[5] and financial time series[6–8]. In general, inferring the parameters of the model is a challenging problem since maximizing the likelihood involves the computation of the partition function $Z = \sum_s P(s)$, which is intractable in most realistic cases. On the other hand, when dealing with time-series, it is possible to use a simpler approach based on the dynamic (kinetic) version of the Ising model analysed in[9], optimized in[10] and generalized in[11–13]. A recent review on this subject can be found in[14].

The original attempt to solve the problem is a gradient descent algorithm known as Boltzmann learning[15]. This method is unpractical on large systems unless heuristic methods, like Monte Carlo (MC) sampling, are used to estimate correlations[16]. Nevertheless MC is slow and thus many sophisticated techniques coming from statistical mechanics and machine learning have been proposed as alternative approaches[17–28]. These methods, however, share one or both of the following shortcomings: (1) they require a large number of observations and (2) the overall performance drops significantly when the dataset is structured. This is often the case when data is produced from a (sub)set of many attractive states or is collected in different regimes, e.g. quiescent and spiking regimes in neural networks. This problem becomes particularly relevant at low temperatures and it has already been studied both in the sparse[29] and in the dense case[30,31]. Pseudo-likelihood[32] based methods[31] were shown to be the best options in a wide range of temperatures. Here, we present two algorithms to compete with the existing state-of-the-art by posing the problem in a Bayesian framework using the Thouless-Anderson-Palmer (TAP) equations[33] and the Restricted Boltzmann Machine (RBM)[34,35]. Our TAP-based algorithm will be shown to achieve a better quality of the results by observing far fewer configurations in the clusterized phase. Moreover, it allows to consider larger system size with respect to those studied in[30,31].

[1]Laboratoire Interdisciplinaire des Sciences du Numérique, Université Paris-Saclay, CNRS, INRIA TAU team, 91190 Gif-sur-Yvette, France. [2]Departamento de Física Téorica I, Universidad Complutense, 28040 Madrid, Spain. [3]LPTMS, Université Paris-Sud 11, UMR 8626 CNRS, Bat. 100, 91405, Orsay, Cedex, France. [4]Mathematics Department, University of Bologna, Piazza di Porta S. Donato 5, 40126 Bologna, Italy. ✉email: daniele.tantari@unibo.it

Results

We consider a dataset with many clusters by drawing configurations from the Hopfield model[36,37]. Given a set of N-dimensional binary independent patterns $\{\zeta^{\mu}\}_{\mu=1,\ldots,P}$, teacher's patterns, the coupling matrix of the associated Hopfield model is defined as $J_{ij} = N^{-1}\sum_{\mu}\zeta_i^{\mu}\zeta_j^{\mu}$ and its Hamiltonian is $H_{\zeta}(\underline{s}) = -1/2\sum_{ij}J_{ij}s_i s_j$. We construct a set of equilibrium configurations $\mathscr{D} = \{\underline{s}^a\}_{a=1,\ldots,M}$ sampled from the Boltzmann distribution

$$P(\underline{s}) = Z^{-1}\exp[-\beta H_{\zeta}(\underline{s})], \tag{1}$$

being β the inverse temperature. The task of a student is to infer the teacher's patterns from the observation of $\mathscr{D}$. This task differs from the one in[30,31], whose focus was the inference of the coupling matrix J only.

For $P = 1$, the Hopfield model is nothing but a Curie–Weiss model. In this case the posterior distribution is

$$P(\underline{\xi}|\mathscr{D}) = Z(\underline{\xi})^{-M}\exp^{\frac{\beta}{2N}\sum_{ij}\sum_{a=1}^{M}s_i^a s_j^a \xi_i \xi_j}, \tag{2}$$

where $\underline{\xi}$ denote the student's pattern and the problem is called *dual* Hopfield model[38,39]. This is readily established by absorbing the ξ-dependence of the partition function into a redefined set of variables $\underline{s}$ via $s_i' = \xi_i s_i$. On the other hand, for $P > 1$, this transformation is not feasible and the posterior comes from the log-likelihood

$$\mathscr{L}(\{\underline{\xi}^{\mu}\}_{\mu=1,\ldots,P}|\mathscr{D}) = \log P(\{\underline{\xi}^{\mu}\}_{\mu=1,\ldots,P}|\mathscr{D}) = \frac{\beta}{2N}\sum_{ij}\sum_{\mu=1}^{P}\sum_{a=1}^{M}s_i^a s_j^a \xi_i^{\mu}\xi_j^{\mu} - M\log Z(\{\xi_i^{\mu}\}_{\mu=1,\ldots,P}). \tag{3}$$

We propose an algorithm based on TAP equations to estimate the posterior associated with Eqs. (2) and (3). In the direct problem, i.e. the study of the Boltzmann distribution of Eq. (1), TAP equations[40–44] describe the local magnetizations $m_i = \langle s_i\rangle$ of the equilibrium states and their use as an inference method has been pioneered in[43,45–49] since they can be used to approximate the intractable partition function appearing in the likelihood. These works paved the way to their applications in a number of problems such as error correcting codes, compressed sensing and learning in neural networks, as discussed in the recent review[50].

Even if TAP and mean field methods have already been used to solve inverse problems[23,29–31,51], the present approach is completely different since we directly apply TAP to the posterior distribution (*dual* model) to improve the quality of the reconstructed network. On the *dual* model, the role of spins and patterns is exchanged: the variables (spins) are now the ξ's and the M sampled configurations play the role dual patterns, thus we use TAP equations to study the local magnetizations $m_i = \langle \xi_i\rangle$. We notice that a similar approach has been independently proposed in[52] for an RBM with 2 hidden binary units, using Belief Propagation.

Single pattern. We start by considering the simplest case $P = 1$. We introduce a *naive* time indexing for an iterative scheme of the TAP equations,

$$m_i^{t+1} = \tanh\left(\beta\sum_{j=1}^{N}J_{ij}m_j^t - \frac{\alpha\beta}{1-\beta(1-q^t)}m_i^t\right), \tag{4}$$

where $J_{ij} = N^{-1}\sum_a s_i^a s_i^a$, $\alpha = P/N$ and $Nq^t = \sum_i (m_i^t)^2$. The entire set of magnetizations $\underline{m}^t$ are updated to achieve $\underline{m}^{t+1}$ in a parallel way. In principle, any sophisticated time indexing schemes can be employed as long as it achieves the convergence to a physical state. Particularly, the so-called Approximate Message Passing (AMP) equations has been the focus of many studies in inference problems[50]. This scheme is inspired by the convergence issues of naive indexing in SK model even in the replica symmetric phase[53]. An explanation to this behavior can be found in[54], where a less trivial time index setting is shown to improve convergence properties outside the glassy phase. The AMP equations exhibit convergence issues for the case of the Hopfield model in the direct problem, when the initial condition is chosen at random, thus in the following we adopt Eq. (4). More details about these issues are discussed in detail in the Methods through simulations and analytical insights.

Once solved Eq. (4), we use m_i as the student estimator for the pattern ζ_i. In Fig. 1, we present the teacher-student overlap $q(\underline{m}, \zeta) = |N^{-1}\sum_i m_i\zeta_i|$, where $\underline{m}$ is the solution of Eq. (4). We observe that in the ferromagnetic-retrieval phase $\bar{\beta} > 1$, a perfect reconstruction may be realized already with a small number of samples. This is due to the large signal contained in the correlation matrix of the data $\bar{c}$. In particular, we notice that in the ferromagnetic phase the student is able to find a pattern correlated with the teacher's one even at $M = 1$. On the other hand in the paramagnetic phase the signal in $\bar{c}$ is weaker and reconstruction is possible only exploiting finite size effects, at the price of observing an extensive number of samples. As discussed in[39], the critical fraction M/N of samples necessary for reconstruction corresponds to the paramagnetic-spin glass transition of the direct problem.

Multiple patterns. The $P > 1$ case is more difficult because of the presence of the term in Eq. (3) coming from the partition function. However, we argue (see Methods) that this term is effectively a (soft) orthogonality constraint over inferred patterns. This observation allows us to design an inference algorithm accordingly. First, let us construct a time evolution of the coupling matrix J_{ij}^{τ} with its initial condition given by $J_{ij}^{\tau=0} = N^{-1}\sum_a s_i^a s_j^a$. At each time step τ, we consider P' TAP students trying to learn the teacher's patterns *independently*. Namely, the magnetizations $m_i^{\mu} = \left\langle \xi_i^{\mu}\right\rangle$ for each student evolve according to Eq. (4) from a randomly initialized configuration. To escape the (unstable) fixed point at $m = 0$, the absolute value of the local magnetization is chosen to be in the interval [0.1, 1]. Upon convergence, we evaluate the P' solutions with the

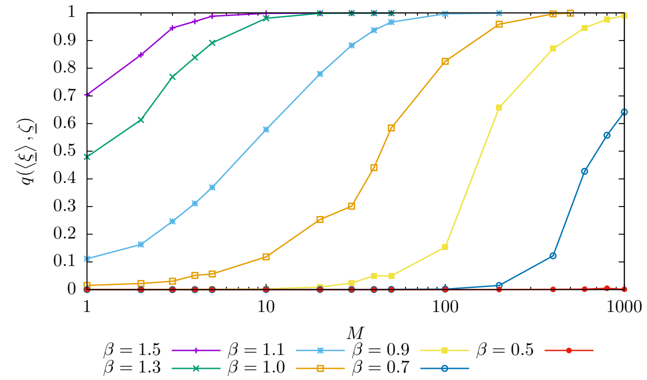

Figure 1. Overlap between the teacher's pattern and pattern recovered by the student when using Eq. (4) with $P = 1$, as a function of the number of samples M, at different temperatures. System size is $N = 1000$. When $\beta < 1$, there exists a critical value of $M \sim O(N)$ below which it is impossible to infer the pattern, whereas above only a finite set of samples is needed.

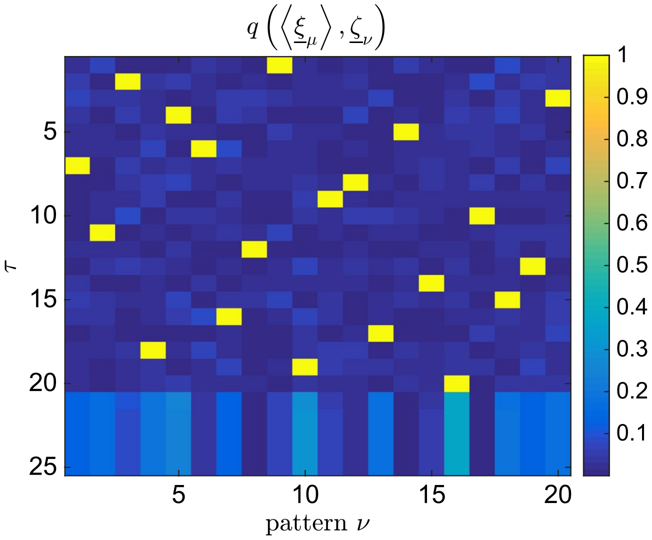

Figure 2. Overlap between the best TAP solutions and the teacher's patterns. The system size is $N = 1000$, the teacher generates $P = 20$ patterns at $\beta = 2$. Inference is done with $P' = 25$ students observing only $M = 200$ samples, i.e. 10 per state. At each iteration step $\tau = 1\ldots, P'$, we pick the best TAP solution and we plot its overlap with all of the teacher's patterns. We observe clearly that the students are able to retrieve all the patterns from the teacher.

score $S_\mu = \sum_{ij}\sum_{a=1}^{M} s_i^a s_j^a m_i^\mu m_j^\mu$. These scores characterize the quality of the TAP fixed points and we pick the one with the largest score. The corresponding magnetization selected by this criterion at time τ are denoted by $\underline{m}^\tau$. This trick is closely reminiscent of the algorithm presented in[28], where the iterative steps are performed by evaluating the likelihood gain obtained moving in different directions and choosing the one with the largest payoff. Finally, in order to learn the remaining contributions, we remove the rank-1 part associated to the retrieved state $\underline{m}^\tau$ from the coupling matrix. When the student knows the actual number of patterns, this correspond to the rule $J_{ij}^{\tau+1} = J_{ij}^{\tau} - \gamma N^{-1} m_i^\tau m_j^\tau$, where $\gamma = M/P$ (assuming that different states are uniformly sampled in the dataset). We repeat these steps until no further patterns are found.

We stress that our algorithm finds solutions correlated with the patterns without any prior information, i.e. we start iterating the TAP equations from a random initial configuration. This is a rather remarkable property in comparison to the method used in[29], where BP equations were guided to converge to the fixed points associated with the patterns using a reinforcement term aligned with the magnetizations of the states. In Fig. 2 we compare the P teacher's patterns with the P' TAP fix points learned from data generated in the retrieval phase[44]. We clearly see that all the P patterns are successfully retrieved from the first $P < P'$ students. In addition, let us define two performance measures, the *simplified* likelihood

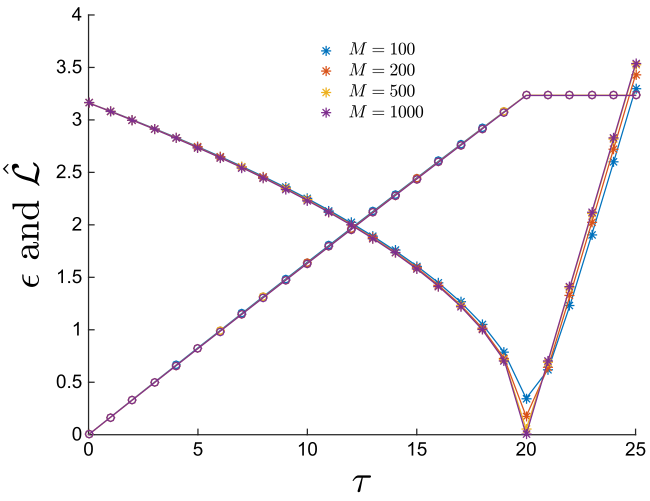

Figure 3. Evolution of the error ϵ and of the simplified Likelihood $\hat{\mathscr{L}}$, as defined in the text, with the iteration of the algorithm. Different lines refer to different values of $M = 100, 200, 500, 1000$ at $\beta = 2$, $P = 20$, $N = 1000$. The error decreases with M and it reaches zero for $M = 1000$, although we observe that even with very few samples, the errors are very small and, as shown on Fig. 2 the patterns are perfectly recovered. The dependency of $\hat{\mathscr{L}}$ on M is negligible. $\hat{\mathscr{L}}$ is rescaled in order to fit in the figure.

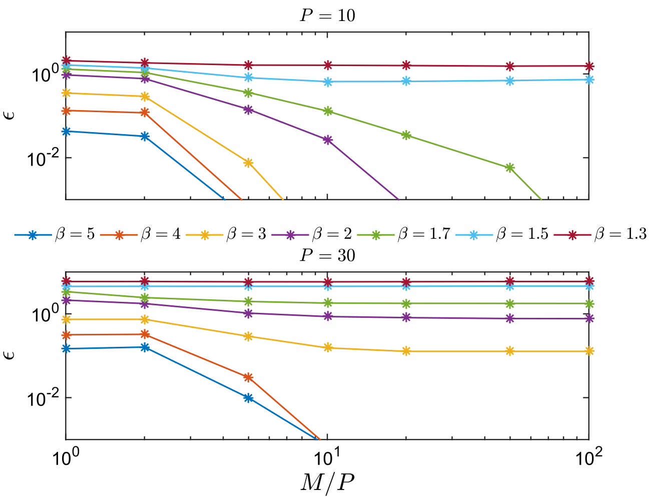

Figure 4. Average error as a function of M/P for a system of size $N = 1000$ with a number of patterns $P = 10$ (top panel) and $P = 30$ (bottom panel). The reconstruction is done using $P' = 2P$ students. The error is computed stopping the algorithm with the criterion described in the text. Each point represent an average over 100 independent trials. In the retrieval phase, the error goes to zero with M.

$$\hat{\mathscr{L}} = \frac{1}{2N}\sum_{ij}\sum_{\mu=1}^{P'}\sum_{a=1}^{M} s_i^a s_j^a m_i^\mu m_j^\mu, \tag{5}$$

and the reconstruction error $\epsilon = [N(N-1)/2]^{-1}\sqrt{\sum_{i<j}(J^\tau_{ij} - J^*_{ij})^2 / \sum_{i<j}(J^*_{ij})^2}$, where J^* denotes the teacher's coupling matrix, and J^τ is the inferred matrix at time τ, $J^\tau_{ij} = N^{-1}\sum_{t=1}^{\tau} m_i^t m_j^t$. The simplified likelihood is defined by neglecting the partition function in Eq. (3). In Fig. 3 their behaviors are reported as a function of iteration time. As expected, ϵ decreases as the students learn the patterns but then increases when the students start to learn the remaining noise. Similarly, the simplified likelihood $\hat{\mathscr{L}}$ develops a kink at the point where students learn all the patterns, that can be used as a stopping condition of the algorithm. In Fig. 4 we study the behavior of ϵ for different values of the temperature. As a function of β, the system sweeps through different regions of the phase diagram. Data is generated with a sequential Glauber dynamics and states are equally sampled. In Fig. 4, we show the behavior of the error computed using the criterion discussed above with the number

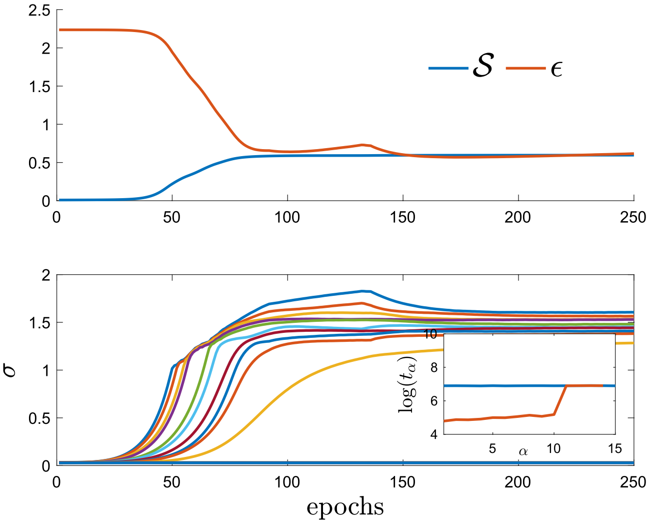

Figure 5. Upper panel: Pseudo-likelihood $\mathscr{S}$ and error ϵ for M during learning. Data is produced by a teacher Hopfield model with $N = 1000$, $P = 10$, $M = 1000$ at $\beta = 2$. Learning is done with an RBM with $N_v = N$ visible units and $N_h = 15$ hidden units. Lower Panel: emergence of singular values σ during learning for the same dataset. $P = 10$ modes emerge. Inset: error t_α for different modes at the beginning of learning (blue line) and at the end of learning (orange line).

of observations in different regions of the phase diagram. As expected, perfect reconstruction is obtained only in the retrieval phase.

Another approach to perform inference is using the equivalence between Hopfield model and RBM[39,44,55–60]. In fact, the likelihood Eq. (3) can be rewritten as

$$P(\{W_i^\mu\}_{\mu=1,\dots,P}|\mathscr{D}) \propto \prod_{a=1}^{M} Z^{-1}(\underline{W}) \int d\underline{\lambda} e^{-\sum_\mu \lambda_\mu^2/2+\sum_{i,\mu} W_i^\mu s_i^a \lambda_\mu}, \tag{6}$$

with $W_i^\mu = \sqrt{\frac{\beta}{N}}\xi_i^\mu$, which defines a RBM with $N_v = N$ binary visible units and $N_h = P$ Gaussian hidden units. Following the standard practice[61], the weights W_i^μ are learned maximizing the log-likelihood using the Persistent Constrastive Divergence (PCD-10) algorithm[62], with 10 Monte Carlo steps to estimate the part of the log-likelihood derivative involving the partition function. Compared to existing methods[19,21,23,24,26–28], the RBM is both time and space efficient as the number of parameters scales as $N_v N_h$ rather than N^2. The number of hidden units N_h plays the role of P', thus we consider the general setting $N_h \geq P$ in the following. RBM learns a set of weights $J_{ij}^r = \beta^{-1} \sum_\mu W_i^\mu W_j^\mu$ that we can compare with the teacher coupling matrix. The error between J_{ij}^r and J_{ij}^* decreases during learning but it never achieves the values found with TAP. In order to monitor learning, we study the pseudo-likelihood $\mathscr{S}$[32], i.e a proxy for the likelihood that can be easily computed (see Methods). In Fig. 5, we show the behavior of these quantities for a dataset generated by a teacher with P patterns and using an RBM with $N_h = P' > P$ hidden units. The minimum of ϵ is achieved when the pseudo-likelihood flattens. This happen when all of the relevant (P) modes of the data have been learned.

Unless learning starts in the vicinity of the teacher's patterns, final RBM weights do not reproduce them, contrarily to the TAP-based algorithm discussed above. In fact, the Hopfield model is invariant under a rotation in the pattern space[30] and the student RBM can learn, at most, the subspace spanned by teacher's patterns. To prove it, we consider the Singular Value Decomposition (SVD) of the dataset, and the SVD of the weights. We denote by $\{\sigma^\alpha\}$ the singular values of the matrix W_i^μ and by t^α the error in reconstructing the singular vector of the data, indexed by α, using only the singular vectors of the weight matrix. In Fig. 5, we show the emergence of different modes during learning. When the singular values σ^α of the coupling matrix emerge, the error t^α decreases. The first P principal modes of the dataset are well represented by the subspace spanned by the singular vectors of the weight matrix W.

Methods

Linear stability analysis of TAP and AMP in the paramagnetic state. Here we present the linear stability analysis of TAP and AMP equations in the paramagnetic state. We will focus on the direct problem where J_{ij} is constructed from M random patterns. While the complete analysis is possible for arbitrary α, we find it more instructive to focus on the limit $\alpha \to 0$ as it greatly simplifies the discussion. As will be shown below, our results are valid if $|1-\beta| \sim O(1)$, which is larger than $O(\alpha)$.

From now on, we denote by TAP the simple iterative updating scheme discussed in the text, reported here for convenience

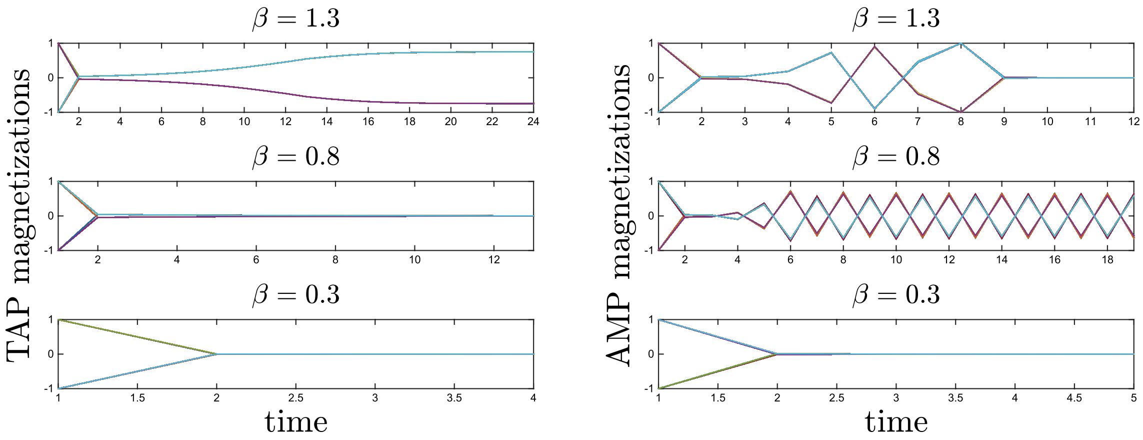

Figure 6. Trajectories of the N magnetizations m_i^t in the updating schemes of TAP, Eq. (7), and AMP, Eq. (8), for three different temperatures at $N = 1000$, $P = 1$. Most of the trajectories are very similar, thus they are indistinguishable. The critical value is at $\beta = 1$. The starting point is chosen at random with the absolute value of the local magnetization equal to one. In the first steps, both TAP and AMP destroy the initial condition and create very small magnetization values. Then, once close to the paramagnetic fixed point $m = 0$, AMP eqs. escape from it for $\beta > 0.5$ while TAP eqs. do not until $\beta > 1$. Moreover, when leaving the paramagnetic state, the direction chosen by AMP is completely random, while TAP moves towards the pattern.

$$m_i^{t+1} = \tanh\left(\beta \sum_{j=1}^{N} J_{ij} m_j^t - \frac{\alpha\beta}{1-\beta(1-q^t)} m_i^t\right) \tag{7}$$

and by AMP the iterative scheme derived in[44],

$$H_i^{t+1} = \frac{1}{1-u^t}\left[\sum_{j\neq i} J_{ij} m_j^t - u^t H_i^t - \frac{\alpha u^t}{1-u^{t-1}} m_i^{t-1}\right], \tag{8}$$

where $m_i^t = \tanh(\beta H_i^t)$, $u^t = \beta(1-q^t)$ and $Nq^t = \sum_i (m_i^t)^2$. This time index setting naturally emerges from the expansion of the BP equations in the large connectivity limit[43].

Let us first consider the linear stability of Eq. (7). Near the paramagnetic state $M_i \sim 0$, this equation may be expanded into

$$m_i^{t+1} \simeq \beta \sum_{j=1}^{N} J_{ij} m_j^t + O(\alpha) \tag{9}$$

where the second term is neglected as it is of $O(\alpha)$. Performing the coordinate change with the eigenvectors of J_{ij} as its basis, one obtains

$$\tilde{m}_\lambda^{t+1} \simeq \beta\lambda\tilde{m}_\lambda^t, \tag{10}$$

where λ is an eigenvalue of J_{ij}. This implies that the paramagnetic solution becomes unstable when $\beta\lambda_{\max} > 1$. The spectrum of coupling matrix follows the Marchenko–Pastur law[44]. Namely, P eigenvalues are $1 + O(\sqrt{\alpha})$ and their eigenvectors span the same space spanned from the set of patterns. The remaining $N - P$ eigenvalues are zero. Thus we find that the critical temperature is $T_c = 1 + O(\sqrt{\alpha})$ (the true value is $T_c = 1/(1+\sqrt{\alpha})$, found expanding TAP equation beyond the $\alpha \to 0$ limit, which is identical to the result of replica theory[44].

Similarly, AMP Eq. (8) can be expanded as follows:

$$\tilde{m}_\lambda^{t+1} = \frac{\beta}{1-\beta}\left[(\lambda-1)\tilde{m}_\lambda^t - O(\alpha)\right]. \tag{11}$$

Because of the $\lambda - 1$ term, in the limit $\alpha \to 0$, the $N - P$ eigenvalues equal to zero give the largest $O(1)$ contribution to the instability of the paramagnetic fixed point. In particular, the modes associated with patterns, with eigenvalues $1 + O(\sqrt{\alpha})$, give a vanishing contribution. From the infinite temperature limit, the first T where this equation becomes unstable is given by $-\frac{\beta}{1-\beta} = -1$, i.e. $T_c = 1/2$. Nevertheless, this unstable direction is orthogonal to the patterns and the magnetization either converges to an unphysical state or never converge (see Fig. 6). The negative value of the leading eigenvalue for $\beta \in [1/2, 1]$ leads to an oscillating behavior starting from the paramagnetic solution, as can be seen in the second plot in Fig. 6. Similar issues with parallel updating of the AMP equations were discussed in[50], and they can be alleviated by updating spins sequentially and introducing a strong dumping. Nevertheless their sequential updating leads to a much slower algorithm, without showing any improvement in the quality of inference in comparison to the parallel updating scheme of Eq. (7).

A different updating scheme of the TAP equations has been recently proposed by[63]. This approach does not require to consider the fully connected limit of the BP equations and it is suitable to be applied in systems with dense random coupling matrices. It is based on a dynamical mean field theory which allows to study the dynamics of iterative algorithms in the thermodynamic limit by averaging over the noise contained in the couplings. For the Hopfield model, the updating scheme turns out to be

$$m_i^{t+1} = \tanh\left(z_i^t + A_t m_i^t\right) \tag{12}$$

$$z_i^t = A_t\left(\sum_j J_{ij} m_j^t - m_i^t\right) + \alpha(1-q^t)A_t z_i^{t-1} \tag{13}$$

where $A_t = \beta/(1+\alpha u_t)$ and u_t is the same quantity introduced in the AMP Eq. (8). It is possible to see that this updating scheme does not present the issues of the AMP algorithm by repeating the same $\alpha \to 0$ analysis presented above.

In Fig. 7, we compare the performances of these three algorithms for different system sizes. We define P_c as the probability to converge to one of the patterns of the system with overlap greater than 0.7 when the initial condition is chosen at random. Sequential AMP were iterated with a damping term d, i.e. $m_i^{t+1} = (1-d)\tanh \beta H_i^{t+1} + dm_i^t$, and $d = 0.95$. For the two parallel TAP equations, (Eqs. 7–13), the iteration is stopped when the average difference between m_i^{t+1} and m_i^t is smaller than 0.001. For the AMP sequential algorithm the iteration is stopped when the average between $\tanh \beta H_i^{t+1}$ and $\tanh \beta H_i^t$ is smaller than 0.001. In all the cases we observe that convergence to patterns is achieved in the retrieval phase. For small values of N, due to finite size effect, convergence regime extends in the metastable retrieval phase too.

The instability issue of the AMP equations presented above holds for the direct problem, but it can be extended also to the inverse, *dual* problem. In this last case, where $J_{ij} = N^{-1}\sum_{a=1}^M s_i^a s_j^a$, if there is enough signal in the data and $\lambda_{\max} > 2$, inference is possible also with parallel AMP equations. Nevertheless the analysis shows that obtaining time indexes from BP does not necessarily lead to good algorithms. TAP, as well as BP, equations describe only fixed points of the associated free energy and, in principle, any updating scheme could be used to solve these equations in an iterative manner, as shown in[63]. The relevance of this observation for other problems requires further analysis and, given that the AMP convergence issues are usually mitigated by considering a sequential updating with a strong dumping, it would be interesting to study whether a similar improvements is achieved when iterating TAP equations with the *naive* time indexing sequentially and with a strong damping, in problems where their parallel updating was failing.

Posterior for P>1.

We discuss the role of the difficult term arising in the posterior distribution when $P > 1$. We show that for the simple case $P = 2$, it has a clear interpretation in terms of a constraint on the orthogonality of the inferred patterns. In fact, let us consider

$$Z(\{\xi_i^1, \xi_i^2\}) = \sum_s e^{\frac{\beta}{N}\sum_{ij}(\xi_i^1\xi_j^1 + \xi_i^2\xi_j^2)s_i s_j}, \tag{14}$$

and let us define $S = \{i : \xi_i^1 = \xi_i^2\}$, such that $|S| = N(1+q)/2$, where q is the mutual overlap between the two patterns, $Nq = \sum_i \xi_i^1\xi_i^2$. The exponent in Eq. (14) reads

$$H_\xi(s) = \frac{2\beta}{N}\sum_{i\in S, j\in S}\xi_i^1\xi_j^1 s_i s_j + \frac{2\beta}{N}\sum_{i\in\overline{S}, j\in\overline{S}}\xi_i^1\xi_j^1 s_i s_j, \tag{15}$$

where we indicate with $\overline{S}$ the complement of set S. Using again the gauge transformation $s_i' = \xi_i^1 s_i$, Eq. (15) leads to

$$Z(\{\xi_i^1, \xi_i^2\}) = Z_{cw,\beta(1+q)}^{N(1+q)/2} Z_{cw,\beta(1-q)}^{N(1-q)/2}, \tag{16}$$

where we indicate with $Z_{cw,\beta}^N$ the partition function of a ferromagnetic Curie–Weiss model at inverse temperature β. We observe that the interaction depends only on their mutual overlap. If we define $\phi = -N\log Z$, we obtain

$$\phi(q) = \frac{1+q}{2}f_{cw}(\beta(1+q)) + \frac{1-q}{2}f_{cw}(\beta(1-q)) \tag{17}$$

where $f_{cw}(\beta)$ is the free energy of the Curie–Weiss model at inverse temperature β. It is easy to check that $\phi(q)$ is a convex function with a minimum in $q = 0$. Thus the term $-M\log Z(\underline{\xi})$ in the posterior can be interpreted as a soft regularizer for patterns orthogonality.

Restricted Boltzmann machine.

A restricted Boltzmann machine (RBM) is a particular kind of Boltzmann machines in which units are divided in two layers, formed by visible $\{s_i\}$ and hidden $\{\lambda_\mu\}$ units, and only interactions W_i^μ between units of different layers are allowed, such that the proxy probability distribution reads

$$P(\underline{s}, \underline{\lambda}|\{W\}) = Z^{-1}(\{W\})e^{-E_W(\underline{s},\underline{\lambda})} \tag{18}$$

where $E_W(\underline{s}, \underline{\lambda}) = -\sum_{i,\mu} W_i^\mu s_i \lambda_\mu$, and $Z(\{W\})$ is the partition function,

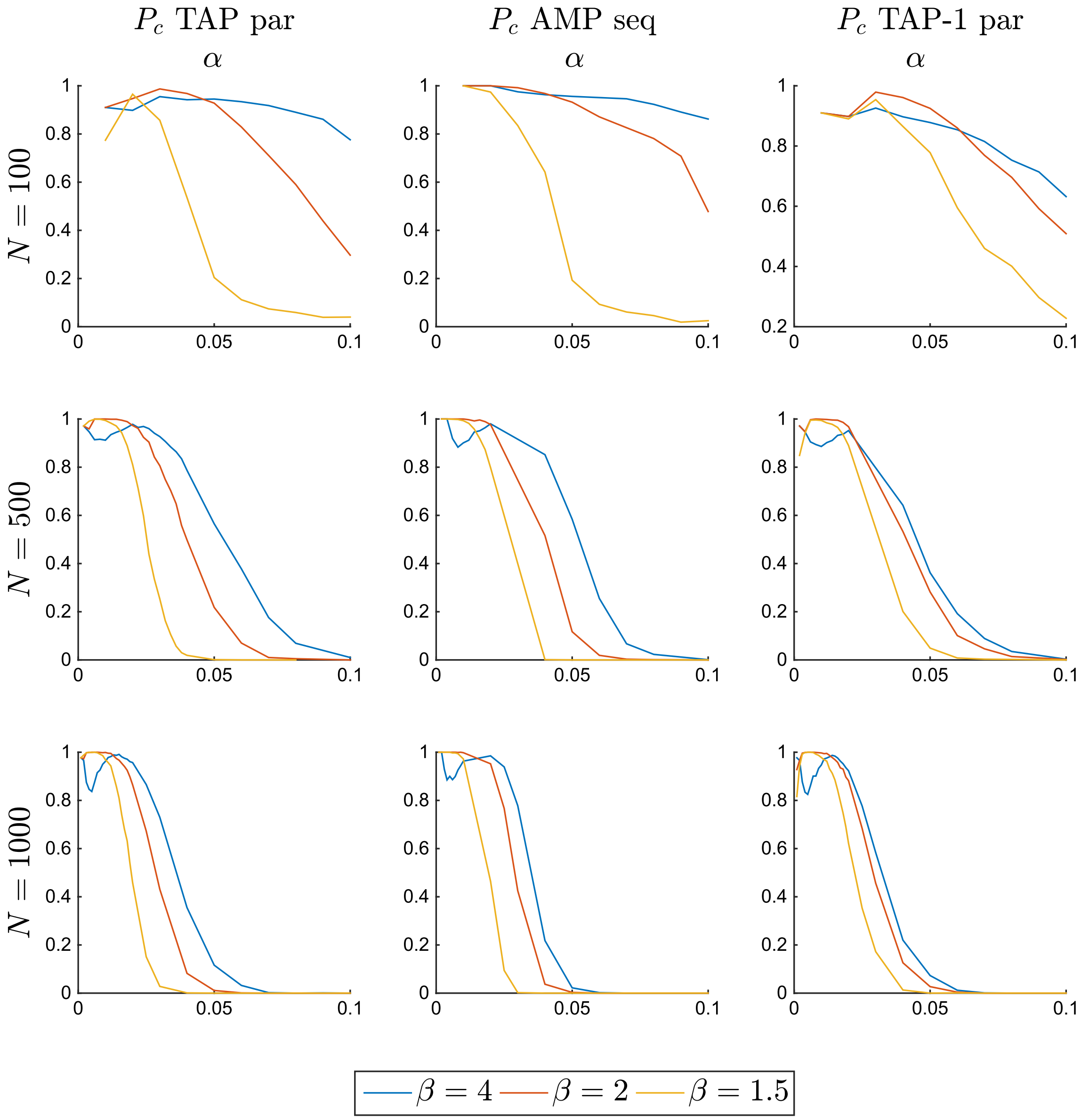

Figure 7. Probability of converging to a pattern when iterating Eq. (7) (left), Eq. (8) (center), Eq. (13) (right), starting from a random initial condition in the direct problem. This probability is estimated running 1000 independent experiments from different realizations of the patterns and different initial conditions and counting the number of times that the equations converged to one of the patterns of the system with overlap greater than 0.7, in order to exclude mixture states. The sequential updating of the AMP equations is done with a dumping term equal to 0.95. The performance of all these algorithms is similar, with the second one being much slower. The initial absolute value of the local magnetizations are mostly irrelevant in the first two cases (and it is chosen to be 1), but needs to be chosen small at low temperatures in the third case (and it is chosen to be 0.1).

$$Z(\{W\}) = \int \prod_{\mu=1}^{N_h} dP(\lambda_\mu) \sum_{\underline{v}} e^{-E_W(\underline{s},\underline{\lambda})}. \tag{19}$$

For our purposes, $P(\underline{\lambda})$ denotes a generic distribution over hidden units, while visible units are ±1 binary variables. We indicate with N_v the number of visible units and with N_h the number of hidden units. RBM has the property that the two conditional probabilities, $P(\underline{s}|\underline{\lambda}, \{W\})$ and $P(\underline{\lambda}|\underline{s}, \{W\})$, factorize over the visible (resp. hidden) units. These machines are used to learn weights such that the distribution over the visible units reproduce the distribution of the data. In other words

$$P(\underline{s}|\{W\}) = \int \prod_{\mu=1}^{N_h} dP(\lambda_\mu) P(\underline{s}, \underline{\lambda}|\{W\}) \tag{20}$$

should reproduce as close as possible $P_D(\underline{s}) = M^{-1}\sum_{a=1}^{M}\delta_{\underline{s},\underline{s}^a}$. Weights can be found minimizing the KL distance between the two distribution, which is equivalent to maximizing the likelihood $\prod_{a=1}^{M} P(\underline{s}^a|\{W\})$ or the log-likelihood

$$\mathscr{L} = \frac{1}{M}\sum_{a=1}^{M}\left(-\log Z(\underline{W}) + \log\int dP(\underline{\lambda})e^{-E_W(\underline{s}^a,\underline{\lambda})}\right). \tag{21}$$

Optimal weights can be learned by gradient ascent:

$$W_i^{\mu} = W_i^{\mu} + \left(\langle\lambda_\mu s_i\rangle_D - \langle\lambda_\mu s_i\rangle_{RBM}\right) \tag{22}$$

where the first average, usually referred to as positive phase, is

$$\langle\lambda_\mu s_i\rangle_D = M^{-1}\sum_a \int dP(\underline{\lambda})P(\underline{\lambda}|\underline{s}^a,\{W\})\lambda_\mu s_i^a \tag{23}$$

and the second average, usually referred to as negative phase, is

$$\langle\lambda_\mu s_i\rangle_{RBM} = \frac{\partial}{\partial W_i^{\mu}}\log Z(\underline{W}). \tag{24}$$

The second one is known to be difficult and it can be computed with approximate methods. One way to estimate it is to use a Monte Carlo (MC). Depending on the number of steps T of the MC Markov chain, this method is referred to as CD-T, where CD stands for Contrastive Divergence. In the text, we discuss results obtained with $T = 10$. When the positive term is computed over a sub set (mini-batch) of the dataset, the direction indicated by the gradient does not correspond to the correct one obtained considering the whole dataset. This trick introduces a source of randomness in the path to the solution, and the associate learning algorithm is called Stochastic Gradient ascent. In our experiments we use a mini-batch size equal to 100. Since mini-batch samples are independent, different parallel MC can be used. In our experiments we used 100 MC chains, one per mini-batch sample. Their initial conditions can be chosen to be the considered samples, but this quickly results in over fitting the parameters, since the MC dynamics spend all the time in the phase space regions close to the samples. When the initial condition of the MC dynamics is chosen at random and we keep track of their positions through different batches and epochs, this method is called Persistent CD (PCD). Our results are obtained using PCD.

As stated above, the likelihood function cannot be easily computed. Thus, we introduce the pseudo-likelihood that, for a model with hidden units, is defined by $\mathscr{S} = \sum_{r=1}^{N_v}\mathscr{S}_r$, where

$$\mathscr{S}_r = \frac{1}{M}\sum_{a=1}^{M}\log(\langle p(s_r^a|\underline{\lambda}\langle_{P(\underline{\lambda}|\underline{s}^a,\{W\})}), \tag{25}$$

where the term inside the log is defined by

$$\langle p(s_r^a|\underline{\lambda}\rangle_{P(\underline{\lambda}|\underline{s}^a,\{W\})} = \int dP(\underline{\lambda})p(s_r^a|\underline{\lambda})P(\underline{\lambda}|\underline{s}^a,\{W\}) \tag{26}$$

and it is equal to

$$\langle p(s_r^a|\underline{\lambda}\rangle_{P(\underline{\lambda}|\underline{s}^a,\{W\})} = \mathscr{N}_a^{-1}\int dP(\underline{\lambda})e^{\sum_{\mu k}W_k^{\mu}s_k^a\lambda_\mu}, \tag{27}$$

where $\mathscr{N}_a$ is a sample dependent normalization factor,

$$\mathscr{N}_a = \sum_{s_r^a}\int dP(\underline{\lambda})e^{\sum_{k\mu}W_k^{\mu}s_k^a\lambda_\mu}. \tag{28}$$

The Pseudo-likelihood is optimized by the same set of parameters $\{W\}$ that optimize the likelihood. In order to show this property, we can take derivatives of $\mathscr{S}$:

$$\frac{\partial\mathscr{S}_r}{\partial W_r^{\nu}} = \frac{1}{M}\sum_{a=1}^{M}\left(\frac{\int\prod_\mu dP(\lambda_\mu)s_r^a\lambda_\nu e^{\sum_{k\mu}W_k^{\mu}s_k^a\lambda_\mu}}{\int dP(\underline{\lambda})e^{\sum_{k\mu}W_k^{\mu}s_k^a\lambda_\mu}} - \frac{\partial}{\partial W_r^{\nu}}\log\sum_{s_r^a}\int dP(\underline{\lambda})e^{\sum_{k\mu}W_k^{\mu}s_k^a\lambda_\mu}\right). \tag{29}$$

The definition

$$P(\underline{\lambda}|\underline{s}^a,\{W\}) = \frac{e^{\sum_{k,\mu}W_k^{\mu}\lambda_\mu s_k^a}}{\int dP(\underline{\lambda})e^{\sum_{k,\mu}W_k^{\mu}\lambda_\mu s_k^a}} \tag{30}$$

allows to write the first term of Eq. (29) as

$$\langle \lambda_\nu s_r \rangle_D = \frac{1}{M} \sum_{a=1}^{M} \int dP(\underline{\lambda}) \lambda_\nu s_r^a P(\underline{\lambda} | \underline{s}^a, \{W\}). \tag{31}$$

The second term is given by the average over samples of

$$\mathscr{N}_a^{-1} \frac{\partial \mathscr{N}_a}{\partial W_r^\nu} = \frac{\sum_{s_r^a} \int dP(\underline{\lambda}) s_r^a \lambda_\nu e^{\sum_{k\mu} W_k^\mu s_k^a \lambda_\mu}}{\sum_{s_r^a} \int dP(\underline{\lambda}) e^{\sum_{k\mu} W_k^\mu s_k^a \lambda_\mu}} \tag{32}$$

and similar manipulations on the second term lead to

$$\frac{\partial \mathscr{S}}{\partial W_r^\nu} = \frac{1}{M} \sum_{a=1}^{M} s_r^a \langle \lambda_\nu \rangle_{P(\underline{\lambda}|\underline{s}^a,\{W\})} - \frac{1}{M} \sum_{a=1}^{M} \left\langle \lambda_\nu \tanh \sum_\mu W_r^\mu \lambda_\mu \right\rangle_{P(\underline{\lambda}|\underline{s}^a,\{W\})} = \langle \lambda_\nu s_r \rangle_D - \left\langle \lambda_\nu \tanh \sum_\mu W_r^\mu \lambda_\mu \right\rangle_D . \tag{33}$$

In the infinite sampling limit,

$$\lim_{M \to \infty} \left\langle \lambda_\nu \tanh \sum_\mu W_r^\mu \lambda_\mu \right\rangle_D = \langle \lambda_\nu s_r \rangle_{RBM}. \tag{34}$$

In fact, it is easy to show that

$$\left\langle \lambda_\nu \tanh \sum_\mu W_r^\mu \lambda_\mu \right\rangle_{RBM} = \langle \lambda_\nu s_r \rangle_{RBM} \tag{35}$$

and, on the other hand, that

$$\lim_{M \to \infty} \left\langle \lambda_\nu \tanh \sum_\mu W_r^\mu \lambda_\mu \right\rangle_D = \left\langle \lambda_\nu \tanh \sum_\mu W_r^\mu \lambda_\mu \right\rangle_{RBM} . \tag{36}$$

Thus the gradient of the pseudo-likelihood $\mathscr{S}$ vanishes on the same set of parameters $\{W\}$ that solve $0 = \partial_W \mathscr{L}$, and this is the reason the pseudo-likelihood can be used to control the learning state. In practice, the probability in Eq. (26), can be estimated after one step of Monte Carlo:

$$\left\langle p(\underline{s_r}^a | \underline{\lambda}) \right\rangle_{P(\underline{\lambda}|\underline{s}^a,\{W\})} \sim \frac{e^{s_r^a \sum_\mu W_i^\mu \lambda_\mu}}{2 \cosh \sum_\mu W_i^\mu \lambda_\mu} \tag{37}$$

where $\underline{\lambda}$ is sampled from the distribution $P(\underline{\lambda} | \underline{s}^a, \{W\})$.

Finally, we discuss the learning of the RBM compared to our TAP based algorithm. As mentioned in the text, unless learning starts in the vicinity of the teacher's patterns, final RBM weights do not reproduce them. In fact, the Hopfield model is invariant under a rotation in the pattern space. The dataset $\mathscr{D}$ analyzed by the student could have been produced by another set of patterns $\{\hat{\underline{\xi}}^\mu\}_{\mu=1,\ldots,P}$ given by $\hat{\xi}_i^\mu = \sum_{\nu=1}^P O_{\mu,\nu} \xi_i^\nu$ where O is an orthogonal matrix. This symmetry implies that the student RBM cannot learn exactly the teacher's patterns. One could think that the singular vectors of W should learn at least the principal vectors of the data (that, given the spherical symmetry, are not necessarily aligned along the teacher's patterns), as discussed in[59]. Nevertheless this is true only during the initial steps of learning, when couplings are small. This is reminiscent of the results discussed in[30], where the posterior of the problem is analyzed in a perturbative expansion. At the first order, corresponding to the small couplings regime, the student's patterns are aligned along the singular vectors of the data at zero order. Anyway, computing higher order corrections, this relation breaks down.

In the following we show that RBM is learning the subspace spanned by the singular vectors of the data. To prove it, we consider the Singular Value Decomposition SVD of the dataset, $D = U_D \Sigma_D V_D^T$, where, considering $N < M$, D is a $N \times M$ matrix, U_D is a orthogonal $N \times N$ matrix, Σ_D is a $N \times M$ matrix, with only N diagonal elements different from zero, and V_D is a orthogonal $M \times M$ matrix. D represent the matrix of the dataset $\mathscr{D}$, where each column is a sample. Similarly, we consider the SVD of the weight matrix, $W^T = U_W \Sigma_W V_W^T$, where W^T is $N_v \times N_h$, U_W is a $N_v \times N_v$ orthogonal matrix, Σ_W is a $N_v \times N_h$ matrix, with N_h diagonal elements different from zero, and V_W is a $N_h \times N_h$ orthogonal matrix. We consider $N_v = N$ and we decompose all of the data modes $u_i^{(\alpha),D} = [U_D]_{i\alpha}$ onto the subspace spanned by the first N_h singular vectors of the weights, $u_i^{(\mu),W} = [U_W]_{i\mu}$, $\mu = 1, \ldots, N_h$:

$$\begin{aligned} \vec{u}^{(\alpha),D} &= \sum_{\mu=1}^{N_h} c_\mu^\alpha \vec{u}^{(\mu),W} + e^\alpha(W), \\ c_\mu^\alpha &= \left\langle \vec{u}^{(\mu),W}, \vec{u}^{(\alpha),D} \right\rangle \end{aligned} \tag{38}$$

where $\{\vec{u}^{(\mu),W}\}_{\mu=1,\ldots,N_h}$ are orthogonal vectors normalized to one. We measure the behavior of

$$t_\alpha = \sum_i |e_i^\alpha| / \sum_i |u_i^{(\alpha),D}| \tag{39}$$

at the beginning and at the end of learning. This quantity measures the difference between the original vector and its projection onto the subspace spanned by the basis $\{\vec{u}^{(\mu),W}\}_{\mu=1,\ldots N_h}$. The results of this analysis are found in the insets of Fig. 5, where we plot these quantities at the initial stage of learning and at the end.

Conclusions

In summary, we discussed a new method to solve inverse problems with a clusterized dataset. We analyzed the fully connected Hopfield model in a teacher–student scenario and proposed an inference method based on the TAP equations working directly on the posterior distribution, i.e. the *dual* problem. We discussed a retrieval algorithm based on the parallel updating of the TAP equations with a naive indexing, showing that in our case it gives good results. Contrarily to previous methods, our algorithm is able at retrieving patterns, besides couplings, because TAP equations allows to reduces the continuous symmetry under rotation to a simple symmetry under permutation over the pattern labels. Finally we compare these results with those obtained with RBM, exploiting their analogies with the Hopfield model. RBM is a good candidate model to perform inference with many variables, a task that would require a much longer execution time to methods based on the optimization of the pseudo-likelihood of an associate pairwise Ising model. Their ability to perform inference tasks systematically, as well as their performance on inferring sparse models, will be addressed elsewhere.

Received: 8 February 2021; Accepted: 7 September 2021
Published online: 07 October 2021

References

1. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.* **108**, E1293–E1301 (2011).
2. Wood, K., Nishida, S., Sontag, E. D. & Cluzel, P. Mechanism-independent method for predicting response to multidrug combinations in bacteria. *Proc. Natl. Acad. Sci.* **109**, 12254–12259 (2012).
3. Schneidman, E., Berry, M. J., Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007–1012 (2006).
4. Cocco, S., Leibler, S. & Monasson, R. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proc. Natl. Acad. Sci.* **106**, 14058–14062 (2009).
5. Bialek, W. *et al.* Social interactions dominate speed control in poising natural flocks near criticality. *Proc. Natl. Acad. Sci.* **111**, 7212–7217 (2014).
6. Bury, T. Market structure explained by pairwise interactions. *Phys. A* **392**, 1375–1385 (2013).
7. Campajola, C., Lillo, F., Mazzarisi, P. & Tantari, D. On the equivalence between the kinetic ising model and discrete autoregressive processes. *J. Stat. Mech. Theory Exp.* **2021**, 033412 (2021).
8. Campajola, C., Lillo, F. & Tantari, D. Unveiling the relation between herding and liquidity with trader lead-lag networks. *Quant. Financ.* **20**, 1765–1778 (2020).
9. Roudi, Y. & Hertz, J. Mean field theory for nonequilibrium network reconstruction. *Phys. Rev. Lett.* **106**, 048702 (2011).
10. Decelle, A. & Zhang, P. Inference of the sparse kinetic ising model using the decimation method. *Phys. Rev. E* **91**, 052136 (2015).
11. Dunn, B. & Roudi, Y. Learning and inference in a nonequilibrium ising model with hidden nodes. *Phys. Rev. E* **87**, 022127 (2013).
12. Campajola, C., Lillo, F. & Tantari, D. Inference of the kinetic ising model with heterogeneous missing data. *Phys. Rev. E* **99**, 062138 (2019).
13. Campajola, C., Di Gangi, D., Lillo, F. & Tantari, D. Modelling time-varying interactions in complex systems: The score driven kinetic ising model. arXiv:2007.15545 (arXiv preprint) (2020).
14. Nguyen, H. C., Zecchina, R. & Berg, J. Inverse statistical problems: From the inverse ising problem to data science. *Adv. Phys.* **66**, 197–261 (2017).
15. Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. A learning algorithm for boltzmann machines. *Cogn. Sci.* **9**, 147–169 (1985).
16. Huang, H. Reconstructing the hopfield network as an inverse ising problem. *Phys. Rev. E* **81**, 036104 (2010).
17. Kappen, H. J. & Rodríguez, F. D. B. Efficient learning in Boltzmann machines using linear response theory. *Neural Comput.* **10**, 1137–1156 (1998).
18. Tanaka, T. Information geometry of mean-field approximation. *Neural Comput.* **12**, 1951–1968 (2000).
19. Sohl-Dickstein, J., Battaglino, P. B. & DeWeese, M. R. New method for parameter estimation in probabilistic models: Minimum probability flow. *Phys. Rev. Lett.* **107**, 220601 (2011).
20. Cocco, S. & Monasson, R. Adaptive cluster expansion for inferring Boltzmann machines with noisy data. *Phys. Rev. Lett.* **106**, 090601 (2011).
21. Aurell, E. & Ekeberg, M. Inverse ising inference using all the data. *Phys. Rev. Lett.* **108**, 090201 (2012).
22. Ricci-Tersenghi, F. The bethe approximation for solving the inverse ising problem: A comparison with other inference methods. *J. Stat. Mech. Theory Exp.* **2012**, P08015 (2012).
23. Nguyen, H. C. & Berg, J. Mean-field theory for the inverse ising problem at low temperatures. *Phys. Rev. Lett.* **109**, 050602 (2012).
24. Cocco, S. & Monasson, R. Adaptive cluster expansion for the inverse ising problem: Convergence, algorithm and tests. *J. Stat. Phys.* **147**, 252–314 (2012).
25. Raymond, J. & Ricci-Tersenghi, F. Mean-field method with correlations determined by linear response. *Phys. Rev. E* **87**, 052111 (2013).
26. Decelle, A. & Ricci-Tersenghi, F. Pseudolikelihood decimation algorithm improving the inference of the interaction network in a general class of ising models. *Phys. Rev. Lett.* **112**, 070603 (2014).
27. Lokhov, A. Y., Vuffray, M., Misra, S. & Chertkov, M. Optimal structure and parameter learning of ising models. *Sci. Adv.* **4**, e1700791 (2018).
28. Franz, S., Ricci-Tersenghi, F. & Rocchi, J. A fast and accurate algorithm for inferring sparse ising models via parameters activation to maximize the pseudo-likelihood. arXiv:1901.11325 (arXiv preprint) (2019).
29. Braunstein, A., Ramezanpour, A., Zecchina, R. & Zhang, P. Inference and learning in sparse systems with multiple states. *Phys. Rev. E* **83**, 056114 (2011).
30. Cocco, S., Monasson, R. & Sessak, V. High-dimensional inference with the generalized hopfield model: Principal component analysis and corrections. *Phys. Rev. E* **83**, 051123 (2011).

31. Decelle, A. & Ricci-Tersenghi, F. Solving the inverse ising problem by mean-field methods in a clustered phase space with many states. *Phys. Rev. E* **94**, 012112 (2016).
32. Besag, J. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika* **20**, 616–618 (1977).
33. Thouless, D. J., Anderson, P. W. & Palmer, R. G. Solution of 'solvable model of a spin glass'. *Phil. Mag.* **35**, 593–601 (1977).
34. Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14**, 1771–1800 (2002).
35. Decelle, A. & Furtlehner, C. Restricted boltzmann machine, recent advances and mean-field theory. *Chin. Phys. B* **30**(4), 040202 (2020).
36. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.* **79**, 2554–2558 (1982).
37. Amit, D. J., Gutfreund, H. & Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.* **55**, 1530 (1985).
38. Barra, A., Genovese, G., Sollich, P. & Tantari, D. Phase transitions in restricted Boltzmann machines with generic priors. *Phys. Rev. E* **96**, 042156 (2017).
39. Barra, A., Genovese, G., Sollich, P. & Tantari, D. Phase diagram of restricted Boltzmann machines and generalized hopfield networks with arbitrary priors. *Phys. Rev. E* **97**, 022310 (2018).
40. Mézard, M., Parisi, G. & Virasoro, M.-A. *Spin Glass Theory and Beyond* (World Scientific Publishing Co., 1990).
41. Nakanishi, K. & Takayama, H. Mean-field theory for a spin-glass model of neural networks: Tap free energy and the paramagnetic to spin-glass transition. *J. Phys. A Math. Gen.* **30**, 8085 (1997).
42. Shamir, M. & Sompolinsky, H. Thouless–Anderson–Palmer equations for neural networks. *Phys. Rev. E* **61**, 1839 (2000).
43. Kabashima, Y. & Saad, D. The tap approach to intensive and extensive connectivity systems. *Adv. Mean Field Methods Theory Pract.* **6**, 65–84 (2001).
44. Mézard, M. Mean-field message-passing equations in the hopfield model and its generalizations. *Phys. Rev. E* **95**, 022117 (2017).
45. Opper, M. & Winther, O. Mean field approach to Bayes learning in feed-forward neural networks. *Phys. Rev. Lett.* **76**, 1964 (1996).
46. Kappen, H. J. & Rodríguez, F. B. Efficient learning in Boltzmann machines using linear response theory. *Adv. Neural Inf. Process. Syst.* **280–286**, 20 (1998).
47. Tanaka, T. Mean-field theory of boltzmann machine learning. *Phys. Rev. E* **58**, 2302 (1998).
48. Kabashima, Y. & Saad, D. *Europhys. Lett.* **44**, 668 (1998).
49. Saad, D. *On-Line Learning in Neural Networks* Vol. 17 (Cambridge University Press, 2009).
50. Zdeborová, L. & Krzakala, F. Statistical physics of inference: Thresholds and algorithms. *Adv. Phys.* **65**, 453–552 (2016).
51. Gabrié, M., Tramel, E. W. & Krzakala, F. Training restricted Boltzmann machine via the Thouless–Anderson–Palmer free energy. *Adv. Neural Inf. Process. Syst.* **1**, 640–648 (2015).
52. Hou, T., Wong, K. & Huang, H. Minimal model of permutation symmetry in unsupervised learning. arXiv:1904.13052 (arXiv preprint) (2019).
53. Kabashima, Y. Propagating beliefs in spin-glass models. *J. Phys. Soc. Jpn.* **72**, 1645–1649 (2003).
54. Bolthausen, E. An iterative construction of solutions of the tap equations for the Sherrington–Kirkpatrick model. *Commun. Math. Phys.* **325**, 333–366 (2014).
55. Barra, A., Bernacchia, A., Santucci, E. & Contucci, P. On the equivalence of hopfield networks and Boltzmann machines. *Neural Netw.* **34**, 1–9 (2012).
56. Agliari, E., Migliozzi, D. & Tantari, D. Non-convex multi-species hopfield models. *J. Stat. Phys.* **172**, 1247–1269 (2018).
57. Genovese, G. & Tantari, D. Legendre equivalences of spherical Boltzmann machines. *J. Phys. A. Math. Theor.* **53**, 094001 (2020).
58. Barra, A., Genovese, G., Guerra, F. & Tantari, D. How glassy are neural networks?. *J. Stat. Mech. Theory Exp.* **2012**, P07009 (2012).
59. Decelle, A., Fissore, G. & Furtlehner, C. Thermodynamics of restricted Boltzmann machines and related learning dynamics. *J. Stat. Phys.* **172**, 1576–1608 (2018).
60. Sollich, P., Tantari, D., Annibale, A. & Barra, A. Extensive parallel processing on scale-free networks. *Phys. Rev. Lett.* **113**, 238106 (2014).
61. Hinton, G. E. A practical guide to training restricted Boltzmann machines. In *Neural Networks: Tricks of the Trade* 599–619 (Springer, 2012).
62. Tieleman, T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, 1064–1071 (ACM, 2008).
63. Opper, M., Cakmak, B. & Winther, O. A theory of solving tap equations for ising models with general invariant random matrices. *J. Phys. A. Math. Theor.* **49**, 114002 (2016).

Acknowledgements

We thank S. Franz, F. Ricci Tersenghi and D. Saad for interesting discussions. J.R. and S.H. acknowledge the support of a Grant from the Simons Foundation (no. 454941, Silvio Franz). A.D. was supported by the Comunidad de Madrid and the Complutense University of Madrid (Spain) through the Atracción de Talento program (Ref. 2019-T1/TIC-13298). D.T. acknowledges GNFM-Indam for financial support of the project "Unsupervised Learning with Multi-Layer Boltzmann Machines".

Author contributions

All authors have equally contributed in developing the theory, making numerical experiments and writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.