UNDERSTANDING THE DYNAMICS OF MESSAGE PASSING ALGORITHMS: A FREE PROBABILITY HEURISTICS*

Manfred Opper, Burak Çakmak

Department of Artificial Intelligence, Technische Universität Berlin
Berlin 10587, Germany

(Received February 3, 2020)

We use freeness assumptions of random matrix theory to analyze the dynamical behavior of inference algorithms for probabilistic models with dense coupling matrices in the limit of large systems. For a toy Ising model, we are able to recover previous results such as the property of vanishing effective memories and the analytical convergence rate of the algorithm.

DOI:10.5506/APhysPolB.51.1673

1. Introduction

Probabilistic inference plays an important role in statistics, signal processing and machine learning. A major task is to compute statistics of unobserved random variables using distributions of these variables conditioned on observed data. An exact computation of the corresponding expectations in the multivariate case is usually not possible except for simple cases. Hence, one has to resort to methods which approximate the necessary high-dimensional sums or integrals and which are often based on ideas of statistical physics [1]. A class of such approximation algorithms is often termed *message passing*. Prominent examples are *belief propagation* [2] which was developed for inference in probabilistic Bayesian networks with sparse couplings and *expectation propagation* (EP) which is also applicable for networks with dense coupling matrices [3]. Both types of algorithms make assumptions on weak dependencies between random variables which motivate the approximation of certain expectations by Gaussian random variables invoking central limit theorem arguments [4]. Using ideas of the statistical physics of disordered systems, such arguments can be justified for the *fixed points* of such algorithms for large network models where couplings

* Presented at the conference *Random Matrix Theory: Applications in the Information Era*, Kraków, Poland, April 29–May 3, 2019.

are drawn from random, rotation-invariant matrix distributions. This extra assumption of randomness allows for further simplifications of message passing approaches [5, 6], leading *e.g.* to the *approximate message passing* AMP or VAMP algorithms, see [7–9].

Surprisingly, random matrix assumptions also facilitate the analysis of the *dynamical* properties of such algorithms [8–10] allowing *e.g.* for exact computations of convergence rates [10, 11]. This result might not be expected, because, mathematically, the updates of message passing algorithms somewhat resemble the dynamical equations of spin-glass models or of recurrent neural networks which often show a complex behavior in the large system limit [12]. This manifests itself *e.g.* in a slow relaxation towards equilibrium [13] with a possible long-time memory on initial conditions [14]. Such properties would definitely not be ideal to the design of a numerical algorithm. So a natural question is: which properties of the dynamics enable both their analytical treatment and guarantee fast convergence? In this paper, we give a partial answer to this question by interpreting recent results on the dynamics of algorithms for a toy inference problem for an Ising network. We develop a heuristics based on freeness assumptions on random matrices which lead to an understanding of the simplifications in the analytical treatment and provide a simple way for predicting the convergence rate of the algorithm.

The paper is organized as follows: In Section 2, we introduce the motivating Ising model and provide a brief presentation on the Thouless–Anderson–Palmer (TAP) mean-field equations. In Section 3 and Section 4, we present the message passing algorithm of [10] (to solve the TAP equations), and provide a brief discussion on its dynamical properties in the thermodynamic limit, respectively. In Section 5 and Section 6, we recover the property of vanishing memories and analytical convergence speed of the message passing algorithm using a free probability heuristic. Comparisons of our results with simulations are given in Section 7. Section 8 presents a summary and outlook.

2. Motivation: Ising models with random couplings and TAP mean field equations

We consider a model of a multivariate distribution of binary units. This is given by an Ising model with pairwise interactions of the spins $\boldsymbol{s} = (s_1, \dots, s_N)^\top \in \{-1, 1\}^N$ described by the Gibbs distribution

$$p(\boldsymbol{s}|\boldsymbol{J}, \boldsymbol{h}) \doteq \frac{1}{Z} \exp\left(\frac{1}{2}\boldsymbol{s}^\top \boldsymbol{J} s + \boldsymbol{s}^\top \boldsymbol{h}\right), \tag{1}$$

where Z stands for the normalizing partition function. While such models have been used for data modeling where the couplings $\boldsymbol{J}$ and fields $\boldsymbol{h}$ are

adapted to data sets [15], we will restrict ourselves to a toy model where all external fields are equal

$$h_i = h \neq 0\,, \qquad \forall i\,. \tag{2}$$

The coupling matrix $\boldsymbol{J} = \boldsymbol{J}^\top$ is assumed to be drawn at random from a rotation invariant matrix ensemble, in order to allow for nontrivial and rich classes of models. This means that $\boldsymbol{J}$ and $\boldsymbol{V}\boldsymbol{J}\boldsymbol{V}^\top$ have the same probability distributions for any orthogonal matrix $\boldsymbol{V}$ independent of $\boldsymbol{J}$. Equivalently, $\boldsymbol{J}$ has the spectral decomposition [16]

$$\boldsymbol{J} = \boldsymbol{O}^\top \boldsymbol{D}\boldsymbol{O}\,, \tag{3}$$

where $\boldsymbol{O}$ is a random Haar (orthogonal) matrix that is independent of a diagonal matrix $\boldsymbol{D}$. This class of models generalizes the well-known SK (Sherrington–Kirkpatrick) model [17] of spin glasses for which $\boldsymbol{J}$ is a symmetric Gaussian random matrix.

The simplest goal of probabilistic inference would reduce to the computation of the magnetizations

$$\boldsymbol{m} = \mathbb{E}[\boldsymbol{s}]\,, \tag{4}$$

where the expectation is taken over the Gibbs distribution. For random matrix ensembles, the so-called TAP equations [17] were developed in statistical physics to provide approximate solutions to $\boldsymbol{m}$. Moreover, these equations can be assumed (under certain conditions) to give exact results (for a rigorous analysis in the case of the SK model, see [18]) for the magnetizations in the thermodynamic limit [12] $N \to \infty$ for models with random couplings. For general rotation invariant random coupling matrices, the TAP equations are given by

$$\boldsymbol{m} = \mathrm{Th}(\boldsymbol{\gamma})\,, \tag{5a}$$

$$\boldsymbol{\gamma} = \boldsymbol{J}\boldsymbol{m} - \mathrm{R}(\chi)\boldsymbol{m}\,, \tag{5b}$$

$$\chi = \mathbb{E}\left[\left(\mathrm{Th}'(\sqrt{(1-\chi)\mathrm{R}'(\chi)}u\right)\right]\,. \tag{5c}$$

Here, u denotes the normal Gaussian random variable and, for convenience, we define the function

$$\mathrm{Th}(x) \doteq \tanh(h + x)\,.$$

Equation (5) provides corrections to the simpler naive mean field method. The latter, ignoring statistical dependencies between spins, would retain only the term $\boldsymbol{J}\boldsymbol{m}$ as the "mean field" acting on spin i. The so-called *Onsager*

reaction term $-\mathrm{R}(\chi)\boldsymbol{m}$ models the coherent small changes of the magnetisations of the other spins due to the presence of spin i. Furthermore, χ coincides with static susceptibility computed by the replica-symmetric Ansatz. The Onsager term for a Gaussian matrix ensemble was developed in [19] and later generalized to general ensembles of rotation-invariant coupling matrices in [20] using a free energy approach. For alternative derivations, see [4] and [6].

The only dependency on the random matrix ensemble in (5) is via the R-transform $\mathrm{R}(\chi)$ and its derivative $\mathrm{R}'(\chi)$. The R-transform is defined as [21]

$$\mathrm{R}(\omega) = \mathrm{G}^{-1}(\omega) - \frac{1}{\omega}\,, \tag{6}$$

where G^{-1} is the functional inverse of the Green function

$$\mathrm{G}(z) \doteq \mathrm{Tr}\left((z\boldsymbol{I} - \boldsymbol{J})^{-1}\right)\,. \tag{7}$$

Here, for an $N \times N$ matrix $\boldsymbol{X}$, we define its limiting (averaged) normalized-trace by

$$\mathrm{Tr}(\boldsymbol{X}) \doteq \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\boldsymbol{X}}\mathrm{tr}(\boldsymbol{X})\,. \tag{8}$$

From a practical point of view, for a concrete N-dimensional coupling matrix $\boldsymbol{J}$, the R-transform term can be approximated using the spectral decomposition (3). The Green function (7) is then replaced by its empirical approximation as

$$\mathrm{G}(z) \simeq \frac{1}{N}\mathrm{tr}\left((z\boldsymbol{I} - \boldsymbol{D})^{-1}\right)\,. \tag{9}$$

The R-transform $\mathrm{R} \doteq \mathrm{R}(\chi)$ (for short) and its derivative $\mathrm{R}' \doteq \mathrm{R}'(\chi)$ are then obtained by solving the fixed-point equations

$$\lambda = \mathrm{R} + \frac{1}{\chi}\,, \tag{10a}$$

$$\mathrm{R} = \lambda - \frac{1}{\mathrm{G}(\lambda)}\,, \tag{10b}$$

$$\mathrm{R}' = \frac{1}{\mathrm{G}(\lambda)^2} + \frac{1}{\mathrm{G}'(\lambda)}\,, \tag{10c}$$

$$\chi = \mathbb{E}\left[\mathrm{Th}'\left(\sqrt{(1-\mathrm{G}(\lambda))\mathrm{R}'}u\right)\right]\,. \tag{10d}$$

3. Approximate message passing algorithm for TAP equations

In this section, we reconsider an iterative algorithm for solving the TAP equations (5) which was introduced in [10] and was motivated by the so-

called VAMP algorithms of [8, 9]. We introduce a vector of auxiliary variables $\boldsymbol{\gamma}(t)$, where t denotes the discrete time index of the iteration. We then proceed by iterating a nonlinear dynamics which is of the simple form of

$$\boldsymbol{\gamma}(t) = \boldsymbol{A} f(\boldsymbol{\gamma}(t-1)) \tag{11}$$

for $t = 1, 2, 3, \ldots$ Here, f is a nonlinear function which is applied component-wise to the vector $\boldsymbol{\gamma}(t-1)$ and $\boldsymbol{A}$ is a fixed $N \times N$ matrix. Before we specify the dynamical system (11) for the TAP equations and its parameters, we should mention that the point-wise nonlinear operation followed by a matrix multiplication is typical of the dynamics of a (single layer) *recurrent neural network* [22]. Hence, the analysis of (11) could also be of interest to these types of models.

For the current application to the TAP equations, we specialize to the function

$$f(x) \doteq \frac{1}{\chi}\mathrm{Th}(x) - x\,, \tag{12}$$

where χ was defined in (5c). The *time-independent* random matrix is given by

$$\boldsymbol{A} \doteq \frac{1}{\chi}\left[\left(\frac{1}{\chi} + \mathrm{R}(\chi)\right)\boldsymbol{I} - \boldsymbol{J}\right]^{-1} - \boldsymbol{I}\,. \tag{13}$$

The initialization of dynamics (11) is given by $\boldsymbol{\gamma}(0) = \sqrt{(1-\chi)\mathrm{R}'(\chi)}\boldsymbol{u}$ where $\boldsymbol{u}$ is a vector of independent normal Gaussian random variables. It is easy to see that the fixed points of $\boldsymbol{\gamma}(t)$ coincide with the solution of the TAP equations for $\boldsymbol{\gamma}$, (5), if we identify the corresponding magnetizations by $\boldsymbol{m} = \chi(\boldsymbol{\gamma} + f(\boldsymbol{\gamma}))$.

We have the following important properties of the dynamics:

$$\mathrm{Tr}(\boldsymbol{A}) = 0 \quad \text{and} \quad \mathrm{Tr}(\boldsymbol{E}(t)) = 0 \quad \text{with} \quad [\boldsymbol{E}(t)]_{ij} \doteq f'(\gamma_i(t))\delta_{ij}\,, \quad \forall t\,. \tag{14}$$

Here, the first and second equalities follow by the constructions of the random matrix $\boldsymbol{A}$ and random initialization $\boldsymbol{\gamma}(0)$, respectively [10]. It is also worth mentioning that we have the freedom to replace the function f with an appropriate sequence of functions, say f_t, in such a way that the conditions $\mathrm{Tr}[\mathrm{diag}(f_t'(\boldsymbol{\gamma}(t)))] = 0$ and $f_t \to f$ as $t \to \infty$ are fulfilled, see [10, Section VIII.B].

4. Dynamics in the thermodynamic limit

Dynamical properties of fully connected disordered systems can be analyzed by a discrete time version of the dynamical functional theory (DFT) of statistical physics originally developed by Martin, Siggia and Rose [23]

and later used for the study of spin-glass dynamics, see *e.g.* [14, 24, 25], and neural network models [26]. Using this approach, it is possible to perform the average over the random matrix ensemble of $\boldsymbol{A}$ and initial conditions for $N \to \infty$, and marginalize out all degrees of freedom $\gamma_j(t)$ for $j \neq i$ and all times t to obtain the statistical properties of trajectories of length T for an arbitrary single node $\{\gamma_i(t)\}_{t=1}^{T}$. Since the nodes are exchangeable random variables under the random matrix assumption, one can obtain the convergence properties of the algorithm by studying a single node.

For a rotation-invariant matrix $\boldsymbol{A}$ and an arbitrary function f, the DFT yields an "effective" stochastic dynamics for $\gamma_i(t)$ which is of the universal form (we skip the index i, since it is the same for all nodes)

$$\gamma(t) = \sum_{s<t} \hat{\mathcal{G}}(t,s) f(\gamma(s-1)) + \phi(t)\,, \qquad t \leq T\,. \tag{15}$$

Here, $\phi(t)$ is a colored Gaussian noise term. This dynamics is of a "mean field" type because the statistics of the noise must be computed from averages over the process itself which involves the function f and the R transform [25]. In general, the explicit analysis of the single node statistics becomes complicated by the presence of the additional memory terms $\hat{\mathcal{G}}(t,s)$ which can be explicitly represented as a function of the $T \times T$ order parameter matrix

$$\mathcal{G}(t,s) \doteq \mathbb{E}\left[\frac{\partial f(\gamma(t-1))}{\partial \phi(s)}\right], \qquad t,s \leq T \tag{16}$$

which again must be computed from the entire ensemble of trajectories of $\gamma(t)$. $\mathcal{G}(t,s)$ represents the average (linear) response of the variable $f(\gamma(t-1))$ to a small perturbation of the driving force $\phi(s)$ at previous times. Hence, by causality, $\mathcal{G}$ is an upper triangular matrix (*i.e.* $\mathcal{G}(t,s) = 0$ for $s \geq t$). In addition, the case of zero response matrix $\mathcal{G} = \mathbf{0}$ leads to $\hat{\mathcal{G}} = \mathbf{0}$. The combination of the Gaussian noise and the response function in the dynamics has an intuitive meaning: The Gaussian can be understood as a representation of the incoherent addition of random variables arising from the multiplication of the vector $f(\boldsymbol{\gamma}(t-1))$ with the random matrix $\boldsymbol{A}$. On the other hand, by treating the typically small matrix elements A_{ij} in a perturbative way [12, Chapter 6], one can estimate the influence of a node i (using a linear response argument) on the $N-1$ neighboring nodes $j \neq i$, which by the symmetry of the matrix, will lead to a coherent, retarded influence of all nodes j back on node i at later times. This explains why memory terms were found to be absent for neural network dynamics with i.i.d. *nonsymmetric* random couplings [26]. This has made a complete analytical treatment of the effective dynamics in such a case possible.

Surprisingly, for the nonlinear function f given in Eq. (12) and the *symmetric* matrix $\boldsymbol{A}$, we have shown in [10] that the response functions (16) vanish, *i.e.* $\mathcal{G}(t,s) = 0$ for all t, s. As a result, also the memory terms vanish; $\gamma(t)$ in (15) simply becomes a Gaussian field. Hence, an analytical treatment is possible as was also shown in the previous studies [8, 9]. In the following section, we will use the freeness argument of random matrix theory to explain this result.

5. Absence of memory terms and asymptotic freeness

To analyze the average response (16) for a single node, we use the chain rule in the dynamical susceptibility for the original N node dynamics (see (11))

$$G_{ij}(t,s) \doteq \frac{\partial f(\gamma_i(t-1))}{\partial \gamma_j(s)} = \left[(\boldsymbol{E}(s)\boldsymbol{A}\boldsymbol{E}(s+1)\boldsymbol{A}\cdots\boldsymbol{E}(t-2)\boldsymbol{A}\boldsymbol{E}(t-1))\right]_{ij} , \quad s < t . \tag{17}$$

By its construction, we can argue that the derivative w.r.t. $\gamma_i(s)$ acts in the same way as the derivative w.r.t. $\phi(s)$ and thus we will have (as $N \to \infty$)

$$\mathbb{E}\left[G_{ii}(t,s)\right] \to \mathcal{G}(t,s) . \tag{18}$$

Here, $\{G_{ii}(t,s)\}_{i\leq N}$ are random w.r.t. the random matrix $\boldsymbol{A}$ and random initialization $\boldsymbol{\gamma}(0)$. By exchangeability $G_{ii}(t,s) \sim G_{jj}(t,s), j \neq i$, the condition $\mathrm{Tr}(\boldsymbol{E}(t)) = 0$ (see (14)) implies vanishing single-step memories, *i.e.* $\mathbb{E}[G_{ii}(t,t-1)] \to 0$. We next argue that for further time-lags, the memories do vanish *in a stronger sense*. Specifically, we will show that

$$\epsilon(t,s) \doteq \lim_{N\to\infty} \mathbb{E}\left[G_{ii}(t,s)^2\right] = 0 , \qquad s < t-1 . \tag{19}$$

To this end, we introduce an auxiliary random diagonal $N \times N$ matrix $\boldsymbol{Z}$ which is independent of $\boldsymbol{A}$ and $\{\boldsymbol{E}(t)\}$. The diagonal entries of $\boldsymbol{Z}$ are independent and composed of ± 1 with equal probabilities. Note that $\mathbb{E}[Z_{nn}Z_{kk}] = \delta_{nk}$. Hence, we can write

$$\frac{1}{N}\mathbb{E}\left[\mathrm{tr}\left((\boldsymbol{Z}\boldsymbol{G}(t,s))^2\right)\right] = \frac{1}{N}\sum_{i,j\leq N}\mathbb{E}[Z_{ii}Z_{jj}]\mathbb{E}[G_{ij}(t,s)G_{ji}(t,s)] \tag{20}$$

$$= \frac{1}{N}\sum_{j\leq N}\mathbb{E}\left[G_{jj}(t,s)^2\right] = \mathbb{E}\left[G_{ii}(t,s)^2\right] . \tag{21}$$

Then, we have

$$\begin{aligned}\epsilon(t,s) &= \mathrm{Tr}(\boldsymbol{Z}\boldsymbol{E}(s)\boldsymbol{A}\boldsymbol{E}(s+1)\cdots\boldsymbol{A}\boldsymbol{E}(t-1)\boldsymbol{Z}\boldsymbol{E}(s)\boldsymbol{A}\boldsymbol{E}(s+1)\cdots\boldsymbol{A}\boldsymbol{E}(t-1)) \\ &= \mathrm{Tr}(\boldsymbol{E}_Z(t,s)\boldsymbol{A}\boldsymbol{E}(s+1)\cdots\boldsymbol{A}\boldsymbol{E}_Z(t,s)\boldsymbol{A}\boldsymbol{E}(s+1)\cdots\boldsymbol{A}) .\end{aligned} \tag{22}$$

Here, we have defined the diagonal matrix $\boldsymbol{E}_Z(t,s) \doteq \boldsymbol{E}(t-1)\boldsymbol{Z}\boldsymbol{E}(s)$. To simplify (22), we will make use of the concept of *asymptotic freeness* of random matrices.

Definition 1 *[21] For the two families of matrices $\mathcal{A} \doteq \{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_a\}$ and $\mathcal{E} \doteq \{\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_e\}$, let $\boldsymbol{P}_i(\mathcal{A})$ and $\boldsymbol{Q}_i(\mathcal{E})$ stand for (noncommutative) polynomials of the matrices in $\mathcal{A}$ and the matrices in $\mathcal{E}$, respectively. Then, we say the families $\mathcal{A}$ and $\mathcal{E}$ are asymptotically free if for all $i \in [1, K]$ and for all polynomials $\boldsymbol{P}_i(\mathcal{A})$ and $\boldsymbol{Q}_i(\mathcal{E})$, we have*

$$\mathrm{Tr}(\boldsymbol{P}_1(\mathcal{A})\boldsymbol{Q}_1(\mathcal{E})\boldsymbol{P}_2(\mathcal{Q})\boldsymbol{Q}_2(\mathcal{E})\cdots\boldsymbol{P}_K(\mathcal{A})\boldsymbol{Q}_K(\mathcal{E})) = 0 \tag{23}$$

given that all polynomials in (23) are centered around their limiting normalized traces, i.e.

$$\mathrm{Tr}(\boldsymbol{P}_i(\mathcal{A})) = \mathrm{Tr}(\boldsymbol{Q}_i(\mathcal{E})) = 0\,, \qquad \forall i\,.$$

Namely, the limiting normalized trace of any adjacent product of powers of matrices — which belong to different free families and are centered around their limiting normalized traces — vanishes.

In product (22), the matrices belong to two families: rotation-invariant and diagonal. Under certain technical conditions — which includes the independence of matrix families — these two matrix families can be treated as asymptotically free [21]. *E.g.* $\boldsymbol{A}$ is asymptotically free of $\boldsymbol{Z}$. Our **heuristic assumption** is that $\boldsymbol{A}$ is also free of the diagonals $\{\boldsymbol{E}(t)\}$. A subtle point should be noted here: Being outcomes of the dynamical system, the diagonal matrices $\{\boldsymbol{E}(t)\}$ are not independent of $\boldsymbol{A}$. Nevertheless, since we expect that the diagonals $\boldsymbol{E}(t)$ have limiting spectral distributions, we consider that asymptotic freeness is a fair heuristic here.

Result (19) follows immediately from the asymptotic freeness assumption: we have that all adjacent factors in product (22) are polynomials belonging to the different free families and all matrices in the product are centered around their limiting normalized-traces.

6. Asymptotic of the local convergence

We will analyze the convergence rate of dynamics (11) in terms of the following measure:

$$\mu_\gamma \doteq \lim_{t\to\infty}\lim_{N\to\infty} \frac{\mathbb{E}\|\boldsymbol{\gamma}(t+1)-\boldsymbol{\gamma}(t)\|^2}{\mathbb{E}\|\boldsymbol{\gamma}(t)-\boldsymbol{\gamma}(t-1)\|^2}\,. \tag{24}$$

To this end, we will assume that one starts the iterations at a point which is close enough to the fixed point of $\boldsymbol{\gamma}(t)$, denoted by $\boldsymbol{\gamma}^*$ such that a linearisation of the dynamics is justified. We conjecture (in accordance with our

simulations) that the initialization does not affect the asymptotic rates. This means that we can substitute $\boldsymbol{\gamma}(t)$ by the following "effective" dynamics:

$$\boldsymbol{\gamma}(t) = \boldsymbol{\gamma}^* + \boldsymbol{\epsilon}(t) \tag{25}$$

with $\boldsymbol{\epsilon}(t)$ small enough to justify the linearised dynamics

$$\boldsymbol{\epsilon}(t) \simeq \boldsymbol{AE}\boldsymbol{\epsilon}(t-1) = (\boldsymbol{AE})^t \boldsymbol{\epsilon}(0) \quad \text{with} \quad [\boldsymbol{E}]_{ij} \doteq f'(\gamma_i^*)\delta_{ij} \,. \tag{26}$$

Moreover, we consider a random initialization $\boldsymbol{\epsilon}(0)$ with $\mathbb{E}[\boldsymbol{\epsilon}(0)\boldsymbol{\epsilon}(0)^\top] = \sigma^2 \boldsymbol{I}$. Then, one can write

$$\mu_\gamma = \lim_{t\to\infty} \frac{\operatorname{Tr}\left[(\boldsymbol{EA}-\boldsymbol{I})(\boldsymbol{EA})^t(\boldsymbol{AE}-\boldsymbol{I})(\boldsymbol{AE})^t\right]}{\operatorname{Tr}\left[(\boldsymbol{EA}-\boldsymbol{I})(\boldsymbol{EA})^{t-1}(\boldsymbol{AE}-\boldsymbol{I})(\boldsymbol{AE})^{t-1}\right]} \,. \tag{27}$$

Similar to the response function, we encounter the same product of two (asymptotic) trace free matrices. We then assume that $\boldsymbol{A}$ and $\boldsymbol{E}$ can be treated as free matrices. Doing so leads to

$$\operatorname{Tr}\left[(\boldsymbol{EA})^{t\mp 1}(\boldsymbol{AE})^t\right] = 0 \quad \text{and} \quad \operatorname{Tr}\left[(\boldsymbol{EA})^t(\boldsymbol{AE})^t\right] = \operatorname{Tr}\left(\boldsymbol{A}^2\right)^t \operatorname{Tr}\left(\boldsymbol{E}^2\right)^t \,. \tag{28}$$

So that we get the simple expression for the convergence rate as

$$\mu_\gamma = \operatorname{Tr}\left(\boldsymbol{A}^2\right) \operatorname{Tr}\left(\boldsymbol{E}^2\right) \,. \tag{29}$$

This shows that when $\operatorname{Tr}(\boldsymbol{A}^2)\operatorname{Tr}(\boldsymbol{E}^2) < 1$, we obtain local convergence of the algorithm towards the fixed point. Moreover, a straightforward calculation shows that

$$\operatorname{Tr}\left(\boldsymbol{A}^2\right) \operatorname{Tr}\left(\boldsymbol{E}^2\right) = 1 - \frac{1 - \operatorname{Tr}\left(\boldsymbol{E}^2\right)\mathrm{R}'(\chi)}{1 - \chi^2 \mathrm{R}'(\chi)} \tag{30}$$

which exactly agrees with the result of the more complex DFT calculation [10]. In the following section, we will support our heuristics by simulations on two instances of random matrices.

7. Simulations

In the sequel, we illustrate the results of the free probability heuristics, *i.e.* (19) and (29). Since we expect that these results are self-averaging in the large-system limit, our simulations are based on single instances of a large random matrix $\boldsymbol{A}$ and random initialization $\boldsymbol{\gamma}(0)$. In particular, we consider the empirical approximation of the limit (19) as

$$\epsilon_N(t,s) \doteq \frac{1}{N}\sum_{i=1}^{N} G_{ii}(t,s)^2 \,. \tag{31}$$

In Fig. 1 (a) and (b), we illustrate the vanishing memory property and the convergence rate of dynamics (11) for the SK model

$$\boldsymbol{J} = \beta \boldsymbol{G}\,, \tag{32}$$

where G_{ij}, $1 \le i < j \le N$, are i.i.d. centered Gaussian random variables with variance $1/N$.

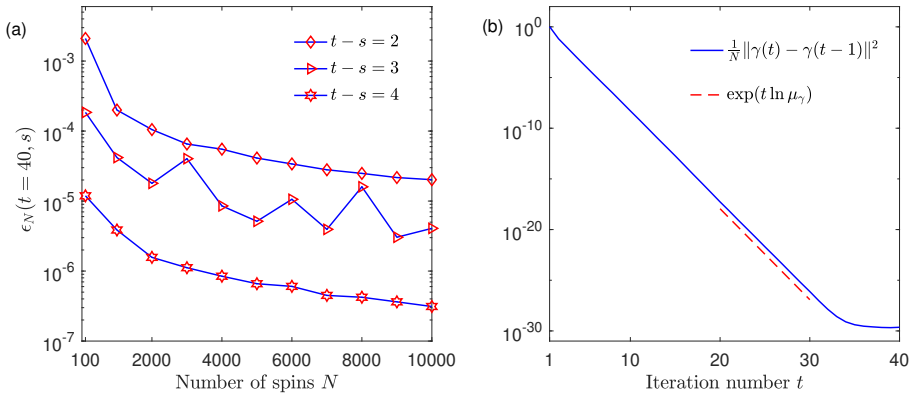

Fig. 1. SK model with the model parameters $h = 1$ and $\beta = 1$: (a) Illustration of vanishing memories (w.r.t. $\epsilon_N(t,s)$ in (31)) for different time lags; (b) Asymptotic of the algorithm with $N = 10^4$ (where the flat line around 10^{-30} is the consequence of the machine precision of the computer which was used).

Second, motivated by a recent study [27] in random matrix theory, we consider a nonrotation invariant random coupling matrix model. The model is related to the random orthogonal model discussed by Parisi and Potters [20] which is defined as

$$\boldsymbol{J} = \beta \boldsymbol{O}^\top \boldsymbol{D} \boldsymbol{O}\,, \tag{33}$$

where $\boldsymbol{O}$ is a Haar matrix and $\boldsymbol{D} = \mathrm{diag}(d_1, \cdots, d_N)$ has random binary elements $d_i = \mp 1$ with $|\{d_i = 1\}| = N/2$. Specifically, we substitute the Haar basis of the random orthogonal model with a randomly-signed DCT (discrete-cosine-transform) matrix as

$$\boldsymbol{J} = \beta \tilde{\boldsymbol{O}}^\top \boldsymbol{D} \tilde{\boldsymbol{O}} \quad \text{with} \quad \tilde{\boldsymbol{O}} \doteq \boldsymbol{\Theta}_N \boldsymbol{Z}\,. \tag{34}$$

Here, $\boldsymbol{Z}$ is an $N \times N$ diagonal matrix whose diagonal entries are independent and composed of binary ∓ 1 random variables with equal probabilities and $\boldsymbol{\Theta}$ is $N \times N$ (deterministic) DCT matrix. The simulation results for the latter model are illustrated in Fig. 2.

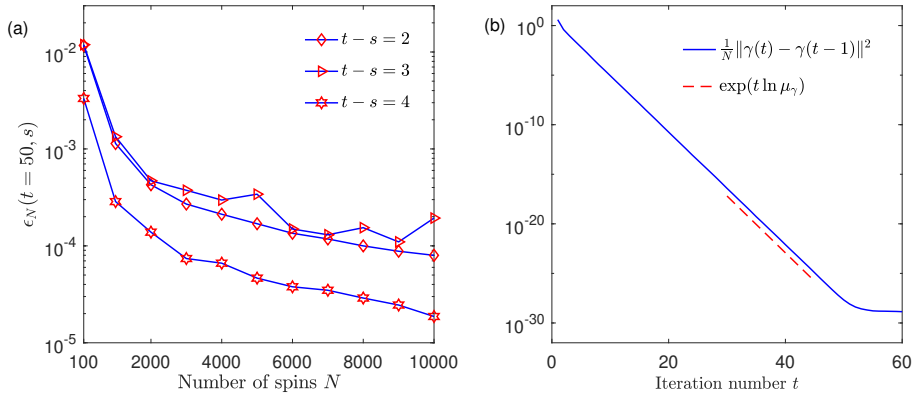

Fig. 2. Randomly signed DCT model with the model parameters $h = 2$ and $\beta = 2$: (a) Illustration of vanishing memories for different time lags; (b) Asymptotic of the algorithm with $N = 10^4$.

They indicate that the free probability heuristics are also very accurate for randomly signed (deterministic) DCT matrix (which contains considerably less randomness compared to the rotation invariant case). As a mater of fact, this is not surprising because for a random permutation matrix $\boldsymbol{P}$ and diagonal matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ such that all matrices are mutually independent, it is proved that the matrices $\boldsymbol{P}^\top \tilde{\boldsymbol{O}}^\top \boldsymbol{D}_1 \tilde{\boldsymbol{O}} \boldsymbol{P}$ and $\boldsymbol{D}_2$ are asymptotically free [27].

8. Summary and outlook

In this paper, we have presented a free probability heuristics for understanding and recovering analytical results for the dynamical behavior of so-called message passing algorithms for probabilistic inference. Such algorithms have the form of a discrete time, recurrent neural network dynamics. We were able to show for a toy Ising model with random couplings that parts of previous results which were obtained by more complicated techniques can be understood and re-derived under the heuristic hypothesis of asymptotic freeness of two matrix families. Under this assumption together with the condition that the matrices are (asymptotically) trace-free, the diagonal elements of the response function which determine the effective memories in the dynamics vanish. This property also yields an analytical result for the exponential convergence of the algorithm towards its fixed point. We have tested these predictions successfully on two types of random matrix ensembles.

We expect that similar arguments can be applied to the analysis of more general types of inference algorithms of the expectation propagation type. It would also be interesting to design novel algorithms that can be analyzed

assuming the freeness heuristics. Of course, the heuristics should eventually be replaced by more rigorous arguments. While our results indicate that message passing algorithms could be analyzed under somewhat weaker conditions on random matrices (compared to explicit assumptions on rotational invariant ensembles), the applicability of these concepts to real data needs to be shown.

The authors would like to thank Yue M. Lu for inspiring discussions. This work was supported by the German Research Foundation, Deutsche Forschungsgemeinschaft (DFG), under grant No. OP 45/9-1 and BMBF (German Ministry of Education and Research) joint project 01 IS18037 A: BZML-Berlin Center for Machine Learning.

REFERENCES

[1] M. Mezard, A. Montanari, «Information, Physics, and Computation», *Oxford University Press*, 2009.

[2] J. Pearl, «Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference», *Elsevier*, 2014.

[3] T.P. Minka, «Expectation Propagation for Approximate Bayesian Inference», Proceedings of the 17$^{\text{th}}$ Conference in Uncertainty in Artificial Intelligence, UAI'01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[4] M. Opper, O. Winther, *Phys. Rev. E* **64**, 056131 (2001).

[5] B. Çakmak, M. Opper, B.H. Fleury, O. Winther, arXiv:1608.06602 [cs.IT].

[6] B. Çakmak, M. Opper, «Expectation Propagation for Approximate Inference: Free Probability Framework», 2018 IEEE International Symposium on Information Theory (ISIT), Piscataway, NJ, USA, 2018, pp. 1276–1280, ISSN 2157-8117.

[7] J. Ma, L. Ping, *IEEE Access* **5**, 2020 (2017).

[8] S. Rangan, P. Schniter, A.K. Fletcher, *IEEE Trans. Inf. Theory* **65**, 6664 (2019).

[9] K. Takeuchi, *IEEE Trans. Inf. Theory* **66**, 368 (2020).

[10] B. Çakmak, M. Opper, *Phys. Rev. E* **99**, 062140 (2019).

[11] B. Çakmak, M. Opper, arXiv:2001.04918 [cs.LG].

[12] M. Mézard, G. Parisi, M. Virasoro, «Spin Glass Theory and Beyond. An Introduction to the Replica Method and Its Applications», *World Scientific Lecture Notes in Physics: Volume 9*, 1987.

[13] L.F. Cugliandolo, J. Kurchan, *Phys. Rev. Lett.* **71**, 173 (1993).

[14] H. Eisfeller, M. Opper, *Phys. Rev. Lett* **68**, 2094 (1992).

[15] G.E. Hinton, *Scholarpedia* **2**, 1668 (2007).

[16] B. Collins, T. Kemp, *J. Funct. Anal.* **266**, 1988 (2014).

[17] D. Sherrington, S. Kirkpatrick, *Phys. Rev. Lett.* **35**, 1792 (1975).

[18] S. Chatterjee, *Probab. Theory Relat. Fields* **148**, 567 (2010).

[19] D.J. Thouless, P.W. Andersen, R.G. Palmer, *Philos. Mag.* **35**, 593 (1977).

[20] G. Parisi, M. Potters, *J. Phys. A: Math. Gen.* **28**, 5267 (1995).

[21] F. Hiai, D. Petz, «The Semicirle Law, Free Random Variables and Entropy», *American Mathematical Society*, Providence, Rhode Island 2006.

[22] I. Goodfellow, Y. Bengio, A. Courville, «Deep Learning», *MIT Press*, 2016, `http://www.deeplearningbook.org`

[23] P.C. Martin, E.D. Siggia, H.A. Rose, *Phys. Rev. A* **8**, 423 (1973).

[24] H. Sompolinsky, A. Zippelius, *Phys. Rev. B* **25**, 6860 (1982).

[25] M. Opper, B. Çakmak, O. Winther, *J. Phys. A: Math. Theor.* **49**, 114002 (2016).

[26] H. Sompolinsky, A. Crisanti, H.J. Sommers, *Phys. Rev. Lett.* **61**, 259 (1988).

[27] G.W. Anderson, B. Farrell, *Adv. Math.* **255**, 381 (2014).