Analysis of Random Sequential Message Passing Algorithms for Approximate Inference

Burak Çakmak[1], Yue M. Lu[2] and Manfred Opper[3]

[1] Artificial Intelligence Group, Technische Universität Berlin, 10587, Germany

[2] John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA

[3] Centre for Systems Modelling and Quantitative Biomedicine, University of Birmingham, B15 2TT, United Kingdom

Abstract. We analyze the dynamics of a random sequential message passing algorithm for approximate inference with large Gaussian latent variable models in a student-teacher scenario. To model nontrivial dependencies between the latent variables, we assume random covariance matrices drawn from rotation invariant ensembles. Moreover, we consider a model mismatching setting, where the teacher model and the one used by the student may be different. By means of dynamical functional approach, we obtain exact dynamical mean-field equations characterizing the dynamics of the inference algorithm. We also derive a range of model parameters for which the sequential algorithm does not converge. The boundary of this parameter range coincides with the *de Almeida Thouless* (AT) stability condition of the *replica symmetric* ansatz for the static probabilistic model.

Keywords: Bayesian Inference, Iterative Algorithms, Approximate message passing, TAP Equations, Random Matrices, Dynamical Functional Theory

1. Introduction

The analysis of the dynamics of message passing algorithms for inference in large probabilistic models has attracted considerable interest in the fields of statistical physics and information sciences [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. From a statistical physics point of view, the fixed points of such algorithms correspond to solutions of TAP mean field equations for disordered systems [12, 13, 14, 15]. The latter, under some conditions on the statistics of the disorder, can lead to *exact solutions* to thermal averages in the large system limit. Hence, message passing algorithms provide efficient computation methods for obtaining accurate solutions to high–dimensional statistical inference problems.

So far, most of the theoretical works on the dynamics of message passing consider a parallel update scheme, where all dynamical nodes are updated *simultaneously* at each iteration of the algorithm. For large classes of the random interaction matrices, the exact temporal progress of the algorithm can then be described by the so-called *state-evolution* equations [1, 2, 16].

In many applications, the parallel dynamics of the algorithm is often replaced by a sequential version, where only a subset of nodes is updated per iteration. For example, Minka's EP (expectation propagation) algorithm [17], which is one of the motivations behind the so-called VAMP (vector approximate message passing) approach [18, 9], is originally formulated in terms of sequential iterations. This type of sequential algorithms have lower computational complexity per iteration. They can also be more memory efficient as they only need to have access to a small batch of the available data at any given time. Moreover, in certain situations, they were found to improve the convergence properties [19]. Our goal in this paper is to extend the theoretical analysis of message passing dynamics from the parallel update setting to the sequential setting. Specifically, we address the following issues:

(i) We analyze the dynamics of a random sequential message passing algorithm for approximate inference with a large Gaussian latent variable model. At each iteration, a random selection of nodes are updated by the algorithm. The probability for a given node to be included in an update is a free parameter. Varying this parameter allows for an interpolation between a full parallel update of all nodes and the case where on average only a single node is updated. Relying on the technique of the *dynamical functional* approach of statistical mechanics [20], we decoupled the degrees of freedom and derive an effective *single node* evolution equation that characterizes the limiting dynamics of the sequential algorithm.

(ii) In practice, the probabilistic model assumed by the inference algorithm may differ significantly from the real data generating process. We take into account this issue by allowing for a possible mismatch between the data generating teacher model and the model used by the student. From a technical point of view, this more general scenario requires a larger number of time dependent order parameters to describe the dynamics of the algorithm. In addition, unlike the case of perfect match between the student and teacher models [21], the message-passing algorithm is no longer guaranteed to converge in the mismatched case. We have identified a range of model parameters for which the convergence of the sequential algorithm is impossible. Interestingly, the boundary of this parameter range coincides with the *de Almeida Thouless* (AT) stability condition of the *replica symmetric* ansatz for the probabilistic model [14, 15].

There have been several earlier studies of sequential dynamics for solving various statistical physics and inference problems [22, 23, 24]. The effective single node dynamics obtained in these studies often contain memory terms that make it difficult to evaluate the two-time correlation functions. Remarkably, due to the construction of our message passing algorithm, its single node dynamics has no *memory term*. As a result, the corresponding two–time correlation functions can be obtained by tractable recursion formulas. A similar "memory-free" property of sequential algorithms was observed in our previous paper [25] on solving the TAP equations for the Sherrington–Kirkpatrick model. Finally, the issue of data-model mismatch has also been previously considered in [26] for parallel-updating message passing algorithms. Unlike in [26] where the analysis is focused on the "single-

time" statistics of the algorithm, we characterize the full effective single-node dynamics. This characterization provides information about the joint statistics of the algorithm over multiple time steps, which is crucial for analysing the convergence properties of the message passing algorithm.

The paper is organized as follows: Section 2 presents the details of the Bayesian probabilistic model considered in this work. We introduce in Section 3 a random sequential iterative algorithm for solving the inference problem. Its thermodynamic properties are studied in Section 4 by using the method of dynamical functional theory. Comparisons of the theory with simulations are given in Section 5. We conclude the paper in Section 6 with a summary and some discussions. The derivations of our results can be found in the Appendix.

2. Latent Gaussian variable models

Message passing algorithms have been successfully applied to latent Gaussian variable models [17, 27, 28]. This class of models finds widespread applications in statistics, machine learning and signal processing. A typical scenario is to infer an unobserved latent vector $\boldsymbol{\theta} \in \Re^{N\times 1}$ by using the Bayesian *posterior distribution*

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{K}) \doteq \frac{1}{Z}\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{0}, \boldsymbol{K}) \prod_{i\leq N} p(y_i|\theta_i) \tag{1}$$

where Z is a normalization constant. This model assumes that the components of the vector $\boldsymbol{y}$ of N real data values are assumed to be generated independently from a likelihood $p(y|\theta)$ based on a vector of unknown parameters $\boldsymbol{\theta}$. Prior statistical knowledge about $\boldsymbol{\theta}$ is introduced by the correlated Gaussian with covariance $\boldsymbol{K} \in \Re^{N\times N}$.

We will later illustrate our theory on the well known example of Bayesian learning of a noisy perceptron—also known as *probit regression* [29]. This corresponds to a binary classification problem with class labels $y_i = \pm 1$. For this model, one assumes a training set given by $\{(\boldsymbol{x}_i, y_i)\}_{i\leq N}$ where $\boldsymbol{x}_i \in \mathbb{R}^{P\times 1}$ stands for a vector of inputs. Class labels y_i are generated according to the observation model

$$y_i = \epsilon_i \,\mathrm{sign}(\boldsymbol{x}_i^\top \boldsymbol{w} + n_i). \tag{2}$$

Here, we allow for additive i.i.d. Gaussian noises n_i for all i with $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\mathbf{I})$ as well as i.i.d. multiplicative flip noises $\epsilon_i = \pm 1$ for all i with $\beta \doteq \Pr(\epsilon_i = -1)$. We assume a latent vector $\boldsymbol{w}$ with Gaussian prior distribution $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})$. To map this problem onto the model (1), we introduce the latent vector $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\omega}$ with $\boldsymbol{X} \doteq [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_N^\top]$. Hence, the prior covariance of $\boldsymbol{\theta}$ equals $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top$ and we have the data likelihood function

$$p(y|\theta) = (1-\beta)\Phi\left(\frac{y\theta}{\sigma}\right) + \beta\Phi\left(-\frac{y\theta}{\sigma}\right) \tag{3}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

3. The random sequential VAMP algorithm

Typical prediction tasks based on observed data involve the computations of *expectations* of components of $\boldsymbol{\theta}$ (or of functions of these components) using the posterior (1). Unfortunately, except for simple Gaussian likelihoods or simple diagonal covariance matrices, such expectations lead to multi–dimensional integrals which cannot be computed analytically. Hence, one has to resort to approximations. To be able to obtain reliable results in the case of high–dimensional vectors $\boldsymbol{\theta}$, so–called message passing algorithms have been developed which provide efficient iterative computations of generalized mean field approximations to the desired expectations.

Given the auxiliary single-site partition function

$$Z_\nu(\gamma, y) \doteq \int \mathrm{d}\theta \; p(y|\theta) e^{-\frac{\nu}{2}\theta^2 + \gamma\theta} \tag{4}$$

the logarithmic derivatives

$$m_\nu(\gamma, y) \doteq \frac{\partial \ln Z_\nu(\gamma, y)}{\partial \gamma} \approx \mathbb{E}[\theta|\boldsymbol{y}, \boldsymbol{K}] \tag{5}$$

$$m'_\nu(\gamma, y) \doteq \frac{\partial m_\nu(\gamma, y)}{\partial \gamma} \approx \mathbb{E}[\theta^2|\boldsymbol{y}, \boldsymbol{K}] - \mathbb{E}^2[\theta|\boldsymbol{y}, \boldsymbol{K}] \tag{6}$$

provide approximations to posterior mean and variances of single components $\theta \equiv \theta_i$ upon convergence of the algorithm. The mean γ_i of the *cavity fields* of a node [14, 30] can be computed iteratively by the VAMP algorithm [18, 9]. In the following, we introduce a *random sequential* version of the usual parallel VAMP. Before the iteration starts, we compute the spectral decomposition

$$\boldsymbol{K} = \boldsymbol{O}\boldsymbol{D}\boldsymbol{O}^\top \tag{7}$$

where $\boldsymbol{D}$ is diagonal with the diagonal entries being the eigenvalues of $\boldsymbol{K}$. We define the iterative algorithm in discrete time by the vector updates for $t = 1, 2, \ldots T$ by

$$\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^{(t-1)} + \boldsymbol{P}^{(t)}[\boldsymbol{\phi}^{(t)} - \boldsymbol{\gamma}^{(t-1)}] \tag{8a}$$

$$\boldsymbol{\phi}^{(t)} = \frac{1}{\tau^{(t)}} \boldsymbol{O}\boldsymbol{D}(\lambda^{(t)}\boldsymbol{D} + \mathbf{I})^{-1}\boldsymbol{O}^\top \tilde{\boldsymbol{\gamma}}^{(t)} - \tilde{\boldsymbol{\gamma}}^{(t)} \tag{8b}$$

$$\tilde{\boldsymbol{\gamma}}^{(t)} = \frac{m_{\nu^{(t-1)}}(\boldsymbol{\gamma}^{(t-1)}, \boldsymbol{y})}{\chi^{(t)}} - \boldsymbol{\gamma}^{(t-1)}. \tag{8c}$$

The scalar quantities $\chi^{(t)}$, $\lambda^{(t)}$, $\tau^{(t)}$, and $\nu^{(t)}$ are updated as

$$\chi^{(t)} = \langle m'_{\nu^{(t-1)}}(\boldsymbol{\gamma}^{(t-1)}, \boldsymbol{y}) \rangle \tag{9a}$$

$$\lambda^{(t)} = \frac{1}{\chi^{(t)}} - \nu^{(t-1)} \tag{9b}$$

$$\tau^{(t)} = \frac{1}{N} \sum_{i \le N} \frac{D_{ii}}{\lambda^{(t)} D_{ii} + 1} \tag{9c}$$

$$\nu^{(t)} = \frac{1}{\tau^{(t)}} - \lambda^{(t)} \tag{9d}$$

where the brackets $\langle \ldots \rangle$ denote an *empirical average* over the sites. Moreover, we consider a random initialization for $\boldsymbol{\gamma}^{(0)}$ from an i.i.d. normal Gaussian distribution. The diagonal matrix $\boldsymbol{P}^{(t)}$ in (8a) is composed of binary decision variables $p_i^{(t)} \doteq P_{ii}^{(t)} \in \{0, 1\}$. The original parallel version of the VAMP algorithm is obtained when $\boldsymbol{P}^{(t)}$ is equal to the unit matrix. Random *sequential* updates are introduced by making the $p_i^{(t)}$ random variables which decide if node i is updated ($p_i^{(t)} = 1$) at time k or not ($p_i^{(t)} = 0$). We assume that the $p_i^{(t)}$ are independent for all i, t and that $\Pr(p_i^{(t)} = 1) = \eta$. The case $\eta = 1/N$ corresponds to updating only a single node on average.

4. The dynamical mean-field equations

We consider an average case analysis of the algorithm in the limit $N \to \infty$ assuming that the data $\boldsymbol{y}$ is generated from a given likelihood model, where for generality, we consider a data-model mismatching scenario. In general, we assume that the components of the vector $\boldsymbol{y}$ are generated independently from a likelihood $p_0(y|\theta)$ which is not necessarily equal to $p(y|\theta)$. For the noisy perception model this would correspond to different sets of hyperparameters β and σ in (3). But we also assume that the prior covariance matrix $\boldsymbol{K}$ is the same for both true parameter and the parameter in the inference (student) model. We choose $\boldsymbol{K}$ to be a random matrix with a rotational invariant distribution. This means that the matrix $\boldsymbol{O}$ in (8) is assumed to be a *Haar* matrix, i.e. a random rotation. In this way, it is possible to model matrices $\boldsymbol{K}$ with nontrivial (weak) dependencies between entries.

Following previous studies [5, 21], we derive an effective dynamics of a *single node*. This is obtained by averaging the generating functional of the dynamics over the randomness of $\boldsymbol{y}$, $\boldsymbol{O}$ and $\{\boldsymbol{P}^{(t)}\}$ and a subsequent decoupling of the degrees of freedom. This involves order parameter functions which are self-averaging for $N \to \infty$. Generating functionals are partition functions for the computation of expectations of dynamical variables where the dynamics is included in terms of Dirac δ functions. To avoid cluttered notation, and with $\boldsymbol{J} \doteq \boldsymbol{K}^{-1}$ ‡, the dynamical functionals corresponding to (8) and (9) for T discrete time steps can be written in the form

$$Z_i(\{l^{(t)}\}) = \int \prod_{t=1}^{T} \Big\{ \mathrm{d}\boldsymbol{\psi}^{(t)} \mathrm{d}\boldsymbol{m}^{(t)} \; \delta\left[\boldsymbol{m}^{(t)} - f_t\left(\{\boldsymbol{\psi}^{(t)}, \boldsymbol{m}^{(l)}, \boldsymbol{P}^{(l)}\}_{l=1}^{t}; \boldsymbol{y}\right)\right] \times \delta(\boldsymbol{\psi}^{(t)} - \boldsymbol{J}\boldsymbol{m}^{(t)}) \mathrm{e}^{i\psi_i^{(t)} l^{(t)}} \Big\} \tag{10}$$

where $\{f_t\}$ is an appropriate sequence of non-linear scalar functions. Using the Fourier

‡ Unless the covariance matrix $\boldsymbol{K}$ has an inverse, one can consider the substitution $\boldsymbol{K} \to \boldsymbol{K} + \epsilon \mathbf{I}$ for $\epsilon > 0$ and perform the limit $\epsilon \to 0$ at the end of the analysis. In any case, the need for the inverse $\boldsymbol{K}^{-1}$ will be bypassed in the analysis. Hence, without loss of generality, we can assume that $\boldsymbol{K}$ has an inverse.

representation of the Dirac measures the averaged generating functional is of the form

$$\mathbb{E}[Z_i(\{l^{(t)}\})] = \int \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{y}\mathrm{dP}(\boldsymbol{O})\; p_0(\boldsymbol{y}|\boldsymbol{\theta})\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{0},\boldsymbol{K})\prod_{t\leq T}\mathrm{dP}(\boldsymbol{P}^{(t)})\; Z_i(\{l^{(t)}\}) \tag{11}$$

$$= c\int \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{y}\; p_0(\boldsymbol{y}|\boldsymbol{\theta})\prod_{t\leq T}\mathrm{dP}(\boldsymbol{P}^{(t)})\mathrm{d}\boldsymbol{\psi}^{(t)}\mathrm{d}\boldsymbol{m}^{(t)}\mathrm{d}\hat{\boldsymbol{\psi}}^{(t)}\delta\left[\boldsymbol{m}^{(t)} - f_t\left(\{\boldsymbol{\psi}^{(l)},\boldsymbol{m}^{(l)},\boldsymbol{P}^{(l)}\}_{l=1}^{t};\boldsymbol{y}\right)\right]$$
$$\times\, \mathrm{e}^{\mathrm{i}\sum_k(\hat{\boldsymbol{\psi}}^{(t)})^\top\boldsymbol{\psi}^{(t)}}\mathrm{e}^{\mathrm{i}\sum_t\psi_i^{(t)}l^{(t)}}\mathbb{E}_{\boldsymbol{O}}\left[\mathrm{e}^{-\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{J}\boldsymbol{\theta}-\mathrm{i}\sum_{t\leq T}(\hat{\boldsymbol{\psi}}^{(t)})^\top\boldsymbol{J}\boldsymbol{\psi}^{(t)}}\right] \tag{12}$$

where $\mathrm{dP}(\boldsymbol{O})$ stands for the Haar invariant measure of the orthogonal group $O(N)$ and c stands for a nonrandom term to ensure the normalization property $\mathbb{E}[Z(\{l^{(t)}=0\})]=1$.

Appendix A gives a short summary of details and references needed for the computations of the expectations and the subsequent decoupling of the degrees of freedom. We find that the effective statistics of an arbitrary single node $\gamma^{(t)} \equiv \gamma_i^{(t)}$ (with similar definitions for other variables) of the algorithm (8) and (9) is given by the stochastic process

$$\left(\theta, y, \phi^{(1:T)}\right) \sim \mathcal{N}(\theta|0,q)p_0(y|\theta)\mathcal{N}(\phi^{(1:T)}|\theta\hat{\mathcal{B}},\mathcal{C}) \tag{13a}$$

$$\gamma^{(t)} = \gamma^{(t-1)} + p^{(t)}(\phi^{(t)} + \gamma^{(t-1)}) \tag{13b}$$

where for short $\phi^{(1:T)} \doteq (\phi^{(1)},\phi^{(2)},\ldots,\phi^{(T)})$ and $q \doteq \lim_{N\to\infty}\frac{1}{N}\mathrm{tr}(\boldsymbol{K})$.

Luckily, similar to previous results [21] obtained for the simpler scenario of parallel dynamics and matching teacher–student models, the effective dynamics does *not contain memory terms*. These terms are often encountered for the stochastic dynamics of disordered systems [23, 24] and would render the driving process $\phi^{(1:T)}$ non Gaussian. This would preclude the computation of explicit analytical results for averages at finite time t and one would have to resort to Monte–Carlo simulations [31] of the effective process (13).

The entries of the $T\times 1$ vector $\hat{\mathcal{B}}$ and the $T\times T$ covariance matrix $\mathcal{C}$ are recursively computed according to

$$\hat{\mathcal{B}}^{(t)} = \frac{\tau^{(t)}\zeta^{(t)}\mathbb{E}[\theta\tilde{\gamma}^{(t)}]}{1-q\tau^{(t)}\zeta^{(t)}} \tag{14}$$

$$\mathcal{C}^{(t,t')} = \frac{\mathcal{D}^{(t,t')} + \mathcal{Q}^{(t,t')}\left(q\hat{\mathcal{B}}^{(t)}\hat{\mathcal{B}}^{(t')} + \hat{\mathcal{B}}^{(t)}\mathbb{E}[\theta\tilde{\gamma}^{(t')}] + \hat{\mathcal{B}}^{(t')}\mathbb{E}[\theta\tilde{\gamma}^{(t)}] + \mathbb{E}[\tilde{\gamma}^{(t)}\tilde{\gamma}^{(t')}]\right)}{1-\mathcal{Q}^{(t,t')}} \tag{15}$$

where we have introduced the auxiliary dynamical order parameters

$$\zeta^{(t)} = \frac{\mathrm{R}(-\tau^{(t)}) - 1/q}{q-\tau^{(t)}} \tag{16a}$$

$$\mathcal{Q}^{(t,t')} = \tau^{(t)}\tau^{(t')}\begin{cases}\frac{\mathrm{R}(-\tau^{(t')})-\mathrm{R}(-\tau^{(t)})}{\tau^{(t)}-\tau^{(t')}} & t\neq t'\\ \mathrm{R}'(-\tau^{(t)}) & \text{else}\end{cases} \tag{16b}$$

$$\mathcal{D}^{(t,t')} = \frac{\hat{\mathcal{B}}^{(t)}\hat{\mathcal{B}}^{(t')}}{\zeta^{(t)}\zeta^{(t')}}\begin{cases}\frac{\zeta^{(t)}-\zeta^{(t')}}{\tau^{(t)}-\tau^{(t')}} & t\neq t'\\ \frac{\zeta^{(t)}-\mathrm{R}'(-\tau^{(t)})}{q-\tau^{(t)}} & \text{else.}\end{cases} \tag{16c}$$

Here, the function $\mathrm{R}(\omega)$ is defined as

$$\mathrm{R}(\omega) \doteq \mathrm{G}^{-1}(\omega) - \frac{1}{\omega} \tag{17}$$

where G^{-1} denotes the functional inverse (w.r.t. decomposition) of the function

$$\mathrm{G}(z) \doteq \lim_{N\to\infty} \frac{1}{N}\mathrm{tr}(\boldsymbol{K}(z\boldsymbol{K} - \mathbf{I})^{-\mathbf{1}}). \tag{18}$$

Note that, when $\boldsymbol{K}$ has an inverse, the function R stands for the R-transform [32] of the limiting spectral distribution of $\boldsymbol{K}^{-1}$. In general, $\mathrm{R}(\omega)$ is well-defined and it is related to the limiting distribution of the non-zero eigenvalues of $\boldsymbol{K}$. Finally, the random field $\tilde{\gamma}^{(t)}$ stands for the effective stochastic process of an arbitrary component of $\tilde{\boldsymbol{\gamma}}^{(t)}$ in (8). Specifically, we have

$$\tilde{\gamma}^{(t)} = \frac{m_{\nu^{(t-1)}}(\gamma^{(t-1)}, y)}{\chi^{(t)}} - \gamma^{(t-1)} \tag{19}$$

where for convenience we have replaced the empirical averages in representing the dynamical order parameter in the algorithm, such as $\chi^{(t)}$, $\lambda^{(t)}$ and etc, by the averages w.r.t. the effective stochastic process (13), e.g. $\chi^{(t)} = \mathbb{E}[m'_{\nu^{(t-1)}}(\gamma^{(t-1)}, y)]$. We can see that the explicit computations of order parameter functions require expectations of nonlinear functions of *pairs* of correlated Gaussian random variables. These can be performed easily by numerical quadrature. To obtain a recursion for such order parameters, we note that the first line of (13) implies

$$\mathcal{C}_\phi^{(t,t')} = \mathcal{C}^{(t,t')} + q\hat{\mathcal{B}}^{(t)}\hat{\mathcal{B}}^{(t')} \tag{20}$$

where $\mathcal{C}_\phi^{(t,t')} \doteq \mathbb{E}[\phi^{(t)}\phi^{(t')}]$. Finally, the covariance of the $\gamma^{(t)}$ variables can be obtained from the second line of (13) by averaging over the decision variables $p^{(t)}$

$$\begin{aligned}\mathcal{C}_\gamma^{(t,t')} = (1-\eta)^2\mathcal{C}_\gamma^{(t-1,t'-1)} + \eta^2 \Big[\mathcal{C}_\phi^{(t,t')} \\ + \sum_{l'=1}^{t'-1}(1-\eta)^{t'-l'}\mathcal{C}_\phi^{(t,l')} + \sum_{l=1}^{t-1}(1-\eta)^{t-l}\mathcal{C}_\phi^{(t',l)}\Big] .\end{aligned} \tag{21}$$

Combined with (15) and (19), we obtain a closed set of equations for the iterative computation of two time correlation functions. We give explicit results of such computations together with comparisons to simulations of the algorithm for the perceptron model (3) in section 5. In the following section, we will analyse the local convergence properties of the algorithm based on a recursion for the necessary *single time* order parameters.

4.1. The fixed point solution

We assume in the following that parameters of the probabilistic model and initial conditions are chosen in such a way that asymptotically for large times, the algorithm will converge to a fixed point. We will then analyze the consistency of this assumption and establish

a necessary criterion for convergence and show its relation to the AT line of the static learning model. Translated to the case of the single node dynamics of the dynamical mean field approach, we assume that $\gamma^{(t)}$ converges to a *static random variable* $\gamma^\star$ for $t \to \infty$. It follows from (13), that $\gamma^\star = \phi^\star$ which is the limit of the Gaussian random variable $\phi^{(t)}$ which drives the dynamics. Specifically, we have

$$(\theta, y, \gamma^\star) \sim \mathcal{N}(\theta|0, q) p_0(y|\theta) \mathcal{N}(\gamma^\star|\theta\hat{\mathcal{B}}^\star, \mathcal{C}^\star) \tag{22}$$

where $\hat{\mathcal{B}}^\star$ stands for the stationary solution of $\hat{\mathcal{B}}^{(t)}$ etc. It is easy to see that

$$\hat{\mathcal{B}}^\star = \zeta^\star \mathbb{E}[\theta m_{\nu^\star}(\gamma^\star, y)]. \tag{23}$$

Then, it follows from (15) that

$$\mathcal{C}^\star = \frac{(\hat{\mathcal{B}}^\star)^2}{\zeta^\star(q - \chi^\star)} + \left(\mathbb{E}[m_{\nu^\star}(\gamma^\star, y)^2] - \frac{(\hat{\mathcal{B}}^\star)^2}{(\zeta^\star)^2(q - \chi^\star)} \right) \mathrm{R}'(-\chi^\star). \tag{24}$$

For example, in the teacher–student matching case, i.e. when $p_0(y|\theta) = p(y|\theta)$, we have

$$\mathbb{E}[m_{\nu^\star}(\gamma^\star, y)^2] = \mathbb{E}[\theta m_{\nu^\star}(\gamma^\star, y)] = q - \chi^\star \tag{25}$$

and the general solutions (23) and (24) simplify to $\hat{\mathcal{B}}^\star = \mathcal{C}^\star = \mathrm{R}(-\chi^\star) - 1/q$. which agrees with our previous results [21].

4.2. Single-time recursion of dynamics and the AT instability criteria

In order to study the convergence towards the fixed point over time, we need the covariances between the random variables $\gamma^{(t)}$, $\phi^{(t)}$ and their asymptotic limits. This can be obtained from recursions of the single time order parameters defined from the limits $\mathcal{C}^{(t)}_{\phi,\gamma} \doteq \lim_{t' \to \infty} \mathcal{C}^{(t,t')}_{\gamma,\phi}$ (assuming the limits exist). Using (20), (21) and (15), we get

$$\mathcal{C}^{(t)}_\gamma = (1 - \eta)\mathcal{C}^{(t-1)}_\gamma + \eta \mathcal{C}^{(t)}_\phi \tag{26}$$

$$\mathcal{C}^{(t)}_\phi = \mathcal{C}^{(t)} + q\hat{\mathcal{B}}^{(t)}\hat{\mathcal{B}}^\star \tag{27}$$

$$\mathcal{C}^{(t)} = \frac{\mathcal{D}^{(t)} + \mathcal{Q}^{(t)} \left(q\hat{\mathcal{B}}^{(t)}\hat{\mathcal{B}}^\star + \hat{\mathcal{B}}^{(t)}\mathbb{E}[\theta\tilde{\gamma}^\star] + \hat{\mathcal{B}}^\star \mathbb{E}[\theta\tilde{\gamma}^{(t)}] + \mathbb{E}[\tilde{\gamma}^{(t)}\tilde{\gamma}^\star] \right)}{1 - \mathcal{Q}^{(t)}} \tag{28}$$

where e.g. $\mathcal{Q}^{(t)} \doteq \lim_{t' \to \infty} \mathcal{Q}^{(t,t')}$. In a similar way, we can show that

$$\mathbb{E}[\theta\tilde{\gamma}^{(t)}] = (1 - \eta)\mathbb{E}[\theta\tilde{\gamma}^{(t-1)}] + \eta\mathbb{E}[\theta f_t(\phi^{(t-1)})] \tag{29}$$

$$\mathbb{E}[\tilde{\gamma}^{(t)}\tilde{\gamma}^\star] = (1 - \eta)\mathbb{E}[\tilde{\gamma}^{(t-1)}\tilde{\gamma}^\star] + \eta\mathbb{E}[f_t(\phi^{(t-1)}) f_\star(\phi^\star)] \tag{30}$$

where we have introduced the function

$$f_{t,\star}(x, y) \doteq \frac{m_{\nu^{(t-1)},\nu^\star}(x, y)}{\mathbb{E}[m'_{\nu^{(t-1)},\nu^\star}(x, y)]} - x. \tag{31}$$

Based on these recursions, we will derive a condition on the parameters of the model for which the assumption of convergence leads to a contradiction. To this end, we study the asymptotic speed of convergence $\gamma^{(t)} \to \gamma^\star$ which we define as

$$\mu_\gamma \doteq \lim_{t\to\infty} \frac{\mathcal{C}_\gamma^\star - \mathcal{C}_\gamma^{(t)}}{\mathcal{C}_\gamma^\star - \mathcal{C}_\gamma^{(t-1)}} \,. \tag{32}$$

The condition $\mu_\gamma \geq 1$ implies that the algorithm no longer converges. Note, that a divergence of the algorithm was not observed for the teacher–student matching scenario discussed in [21]. In Appendix B, we derive the explicit formula

$$\mu_\gamma = 1 - \eta \frac{1 - \mathbb{E}[(m'_{\nu^\star}(\gamma^\star, y))^2]\mathrm{R}'(-\chi^\star)}{1 - (\chi^\star)^2 \mathrm{R}'(-\chi^\star)}. \tag{33}$$

Here, from definition of the function $\mathrm{R}(\omega)$ (17) it follows that the term $1 - (\chi^\star)^2\mathrm{R}'(-\chi^\star)$ is always positive. Hence, $\mu_\gamma \geq 1$ if and only if

$$\mathbb{E}[(m'_{\nu^\star}(\gamma^\star, y))^2]\mathrm{R}'(-\chi^\star) \geq 1. \tag{34}$$

Following the arguments in [14, 15] we conclude that equation (34) coincides with the stability condition of the *replica symmetric* ansatz for the static probabilistic model-known as the *de Almeida Thouless* (AT) criterion. Remarkably, the stability criterion is *independent* of the update parameter η . This indicates that (at least within our theoretical setting), a diverging parallel iterated algorithm cannot be made convergent by reverting to a random sequential version.

5. Simulation results

In the following, we compare our analytical results to numerical simulations of the algorithm for the perceptron model. We assume that the teacher model from which data are generated is of the general form (3) with teacher parameters denoted by β and σ. For the student likelihood used in the inference algorithm we restrict ourselves to the simple *noise free* likelihood model

$$p(y|\theta) = \Theta(y\theta) \tag{35}$$

where Θ stands for the unit-step function. This is a special case of (3) corresponding to the limits $\beta, \sigma \to 0$. We specialise on the following random matrix models for $\boldsymbol{X}$: (i) the entries of $\boldsymbol{X}$ are independent Gaussian with zero mean variance $1/N$; (ii) $\boldsymbol{X} = \boldsymbol{O}\boldsymbol{S}$ where $\boldsymbol{S}$ is the $N \times P$ projection matrix with $N \geq P$ and $\boldsymbol{S}_{ij} = \delta_{ij}$ for all i, j, and $\boldsymbol{O}$ is an $N \times N$ Haar random matrix. The function $\mathrm{R}(\omega)$ in (17) for these models reads as

$$\mathrm{R}(\omega) = \begin{cases} \frac{q-1-\sqrt{(q-1)^2-4\omega}}{2\omega} & \text{model } (i) \\ 1 + \frac{q-1}{\omega} & \text{model } (ii) \end{cases} \,. \tag{36}$$

We will first present non asymptotic (finite times) results. In order to demonstrate that our analytical approach also applies to non convergent dynamics of the algorithm, we consider

model parameter settings from the *unstable* region. Specifically, we set $\beta_0 = \frac{1}{2}$ and $\alpha = 2$ and $\eta = 0.8$. In this case, we obtain the following results for two–time correlations over 5 time steps with

$$\mathbb{E}[\phi^{(1:5)}(\phi^{(1:5)})^\top] = \begin{bmatrix} 6.47 & 11.73 & 17.45 & 23.55 & 29.98 \\ 11.73 & 24.03 & 36.60 & 50.15 & 64.46 \\ 17.45 & 36.60 & 58.6041 & 80.29 & 103.80 \\ 23.55 & 50.15 & 80.2957 & 114.26 & 146.86 \\ 29.98 & 64.46 & 103.80 & 146.86 & 194.73 \end{bmatrix}$$

$$\frac{1}{N}\boldsymbol{\phi}^{(1:5)}(\boldsymbol{\phi}^{(1:5)})^\top = \begin{bmatrix} 6.48 & 11.74 & 17.38 & 23.36 & 29.62 \\ 11.74 & 24.09 & 36.50 & 49.82 & 63.80 \\ 17.38 & 36.50 & 58.15 & 79.39 & 102.27 \\ 23.36 & 49.82 & 79.39 & 112.72 & 144.46 \\ 29.62 & 63.80 & 102.27 & 144.46 & 190.87 \end{bmatrix}.$$

We clearly see a strong increase in autocorrelations over time, indicating the divergence of the algorithm. Here, the simulation result is based on a *single instance* of the model with $N = 2^{14}$. The results were obtained from the random matrix (i). For the random matrix model (ii) we have similar theory-experiment agreement.

Secondly, we present results on the error of estimating the true teacher parameter $\boldsymbol{\theta}$ at each iteration step in Figure 1. The parameters correspond to the region of convergence. In contrast to typical results for cases of teacher–student model matching (with optimally chosen variance of initial conditions), the prediction error turns out to be non-monotonic. Finally, we illustrate the asymptotic speed of convergence predicted by the theory compared

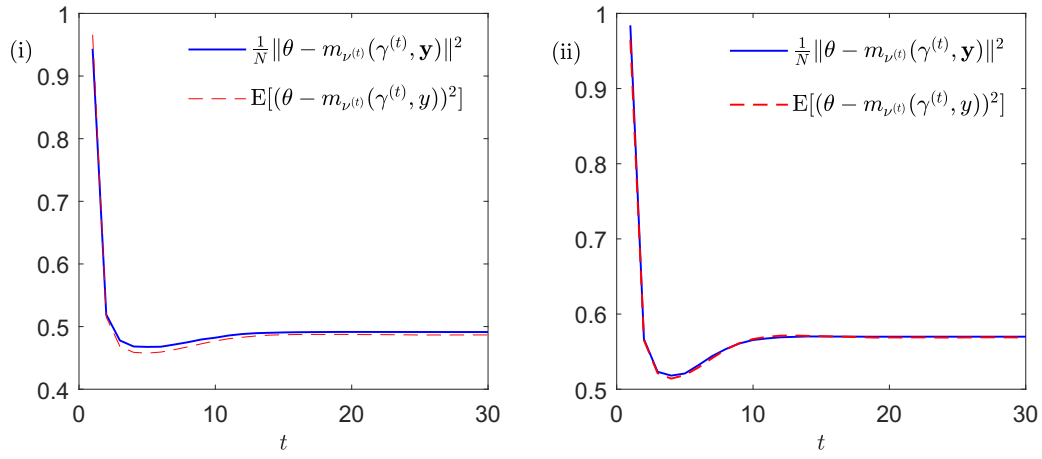

Figure 1: Predicting of the estimation error for a given iteration time-step t. E.g. the figure with label (i) is for the random matrix model (i). The model parameters are chosen as $\sigma_0^2 = 10^{-2}$, $\beta_0 = 0.2$, $\eta = 0.5$, $N = 3P/2$ and $P = 2^{12}$.

to a single simulation of the algorithm. To show the robustness of our results, we have chosen the parameters yielding large values of static order parameters. Nevertheless, we find a remarkably good prediction of the exponential convergence.

6. Summary and discussion

We have analysed the dynamics of a message passing algorithm for inference in large latent Gaussian variable models. Our analysis is based on a teacher-student scenario

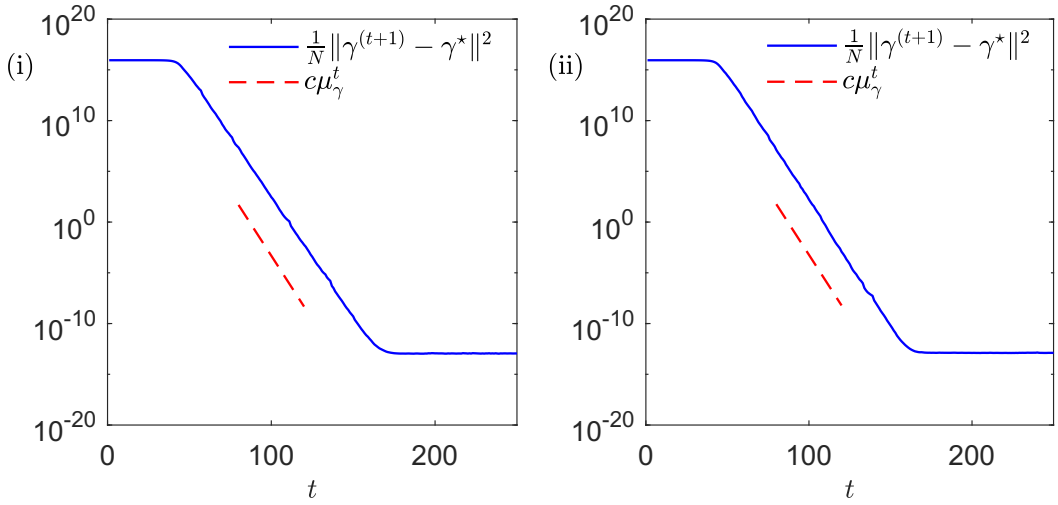

Figure 2: Asymptotic of the algorithm. The model parameters are chosen as $\sigma_0^2 = 0.01$, $\beta_0 = 0.3$, $\eta = 0.5$, $N = 3P$ and $P = 2^{12}$. In this case, we have $\mathcal{C}_\phi^\star \approx 8 \times 10^{15}$ and $\hat{\mathcal{B}}^\star \approx 10^7$.

together with random matrix assumptions for data. We have focused on the problem of student-teacher mismatch and random sequential updates. Using a dynamical functional approach we have decoupled the degrees of freedom and have derived an effective stochastic dynamics for single nodes. The absence of memory terms in the single node dynamics leads to tractable recursions for two–time correlation functions. Comparison between our theory and simulations on single instances of large systems show excellent agreement.

We have shown that a teacher-student mismatch opens the possibility of a divergence of the algorithm. We have identified the range of model parameters for which convergence to a fixed point is impossible. Our main result is that the critical set of parameters is identified as the AT line of the static replica symmetric solution. It would be interesting to see if one could prove *global convergence* in the stable region. For this one would have to go beyond the local stability analysis presented in this paper and study the full temporal development of a set of coupled order parameters. A possible simplification could be the construction of a Lyapunov function for the single node dynamics.

Since the static AT stability criterion is independent of the update schedule of the dynamics, we were not able to show that a divergent (parallel) algorithm can be made convergent using random sequential updates. One might argue that this negative result could be related to the random matrix distributions used in the modeling of the data. A second possibility is the simplicity of the node update used in our model. As an alternative one could define random updates of the $\tilde{\boldsymbol{\gamma}}^{(t)}$ in last line of (8) variables instead. The dynamical mean field analysis of the corresponding model will be given elsewhere.

Acknowledgment

This work was supported by the German Research Foundation, Deutsche Forschungsgemeinschaft (DFG), under Grant "RAMABIM" with No. OP 45/9-1, by the US National Science Foundation under grant CCF-1910410, and by the Harvard FAS Dean's Competitive Fund for Promising Scholarship.

Appendix A. The dynamical functional analysis

The disorder average in (12) can be computed using the saddle-point method. Specifically, we can follow the steps [21, Eq. (B.6)–(B.34)], by essentially replacing all averages over the matrix $\boldsymbol{A}$ by averages over $\boldsymbol{J}$ and read off the result. In our case, the variables $\boldsymbol{\theta}$, $\boldsymbol{m}^{(k)}$ and $\hat{\boldsymbol{\psi}}^{(k)}$ play the roles of $\boldsymbol{u}$, $\boldsymbol{\gamma}(k)$ and $\hat{\boldsymbol{\rho}}(k)$ in [21], respectively. Doing so leads to the large N limit approximation of the the averaged generating functional as

$$\mathbb{E}[Z_i(\{l(t)\})] \simeq \int \mathrm{d}\theta \mathrm{d}y \mathrm{d}\phi^{(1:T)}\; p_0(y|\theta)\mathcal{N}(\theta|0,q)\mathcal{N}(\phi^{(1:T)}|\theta\hat{\mathcal{B}},\mathcal{C}) \prod_{t\leq K} \mathrm{dP}(p^{(t)})\mathrm{d}\psi^{(t)}\mathrm{d}m^{(t)}$$
$$\times\, \delta\left[m^{(t)} - f_t\left(\{\psi^{(l)}, m^{(l)}, p^{(l)}\}_{l=1}^{t}; y\right)\right] \delta\left[\psi^{(t)} - \phi^{(t)} - \sum_{t\leq T} \hat{\mathcal{G}}^{(t,t')} m^{(t)}\right] \mathrm{e}^{\mathrm{i}\psi^{(t)} l^{(t)}}. \tag{A.1}$$

Here, $\hat{\mathcal{G}}^{(t,l)}$ denotes the (t,l)th indexed entries of the $T \times T$ memory matrix $\hat{\mathcal{G}}$ which is defined in terms of the R-transform and its power series expansion as

$$\hat{\mathcal{G}} = \mathrm{R}(\mathcal{G}) = \sum_{n=1}^{\infty} c_n \mathcal{G}^{n-1}. \tag{A.2}$$

Te entries of the $T \times T$ *response matrix* $\mathcal{G}$ are given by

$$\mathcal{G}^{(t,t')} \doteq \mathbb{E}\left[\frac{\partial m^{(t)}}{\partial \phi^{(t')}}\right]. \tag{A.3}$$

Moreover, the Gaussian process $\{\phi^{(t)}\}$ has the $T \times 1$ mean vector $\theta\hat{\mathcal{B}}$ and $T \times T$ covariance matrix $\mathcal{C}$ which are computed by

$$\hat{\mathcal{B}} = \left(\sum_{n=2}^{\infty} c_n \sum_{n'=0}^{n-2} (-q)^{n'} \mathcal{G}^{n-n'-2}\right) \mathcal{B} \tag{A.4}$$

$$\mathcal{C} = \sum_{n=2}^{\infty} c_n \sum_{n'=0}^{n-2} \mathcal{G}^{n'} \mathcal{C}_m (\mathcal{G}^{\top})^{n-2-n'}$$
$$- \sum_{n=3}^{\infty} c_n \sum_{n'=0}^{n-2} (-q)^{n'} \sum_{l=0}^{n-n'-3} (\mathcal{G}^{\top})^{l} \mathcal{B}\mathcal{B}^{\top} \mathcal{G}^{n-n'-l-3} \tag{A.5}$$

where we have defined

$$\mathcal{C}_m^{(t,t')} \doteq \mathbb{E}[m^{(t)} m^{(t')}] \;\text{ and }\; \mathcal{B}^{(t)} \doteq -\mathbb{E}[\theta m^{(t)}]. \tag{A.6}$$

Appendix A.1. The analysis of sequential dynamics

By using the property of Dirac-delta function $\delta(\boldsymbol{y}) = |\boldsymbol{X}|\delta(\boldsymbol{X}\boldsymbol{y})$ we note that

$$\delta\left[\boldsymbol{\phi}^{(t)} - \left(\frac{1}{\tau^{(t)}}(\lambda^{(t)}\mathbf{I} + \boldsymbol{J})^{-1} - \mathbf{I}\right)\tilde{\boldsymbol{\gamma}}^{(t)}\right] = \frac{1}{c^{(t)}}\delta\left[\tilde{\boldsymbol{\gamma}}^{(t)} - \tau^{(t)}(\lambda^{(t)}\mathbf{I} + \boldsymbol{J})(\boldsymbol{\phi}^{(t)} + \tilde{\boldsymbol{\gamma}}^{(t)})\right] \quad \text{(A.7)}$$

$$= \frac{1}{c^{(t)}}\int \mathrm{d}\boldsymbol{m}^{(t)}\mathrm{d}\boldsymbol{\psi}^{(t)}\delta[\tilde{\boldsymbol{\gamma}}^{(t)} + \lambda^{(t)}\boldsymbol{m}^{(t)} + \boldsymbol{\psi}^{(t)}]\delta[\boldsymbol{m}^{(t)} + \tau^{(t)}(\boldsymbol{\phi}^{(t)} + \tilde{\boldsymbol{\gamma}}^{(t)})]\delta[\boldsymbol{\psi}^{(t)} - \boldsymbol{J}\boldsymbol{m}^{(t)}] \quad \text{(A.8)}$$

where for short we define the dynamical determinant $c^{(t)} \doteq \left|\tau^{(t)}(\lambda^{(t)}\mathbf{I} + \boldsymbol{J})\right|$ which do not depend on the disorder variables and thereby they solely play the role of appropriate constant terms in the disorder average. Indeed, one can express the dynamics $\boldsymbol{\phi}^{(t)}$ in (8b) in terms of the system of equations

$$\boldsymbol{m}^{(t)} = -\tau^{(t)}(\boldsymbol{\phi}^{(t)} + \tilde{\boldsymbol{\gamma}}^{(t)}) \quad \text{(A.9a)}$$

$$\boldsymbol{\phi}^{(t)} = \boldsymbol{J}\boldsymbol{m}^{(t)} - \nu^{(t)}\boldsymbol{m}^{(t)}. \quad \text{(A.9b)}$$

Consequently, by the general results of the dynamical functional theory (A.1), $\{\phi_i^{(t)}\}_{t=1}^{T}$ (for an arbitrary component i) can be transformed into a Gaussian random sequence by appropriate *subtractions*. The subtractions define an auxiliary dynamical system which is obtained by replacing the variable $\boldsymbol{\phi}^{(t)}$ by

$$\boldsymbol{\phi}_{aux}^{(t)} = \boldsymbol{J}\boldsymbol{m}^{(t)} - \sum_{l\leq t}\hat{\mathcal{G}}^{(t,l)}\boldsymbol{m}^{(t)} \quad \text{(A.10)}$$

for $t = 1, 2, \ldots T$. The entries of the *response matrix* $\mathcal{G}$ read

$$\mathcal{G}^{(t,t')} \doteq \mathbb{E}\left[\frac{\partial m^{(t)}}{\partial \phi_{aux}^{(t')}}\right]. \quad \text{(A.11)}$$

Moreover, by construction we have

$$\frac{\partial \gamma^{(t-1)}}{\partial \phi_{aux}^{(t')}} = \underbrace{p^{(t')}\prod_{l=t'+1}^{t-1}(1-p^{(l)})}_{\doteq p^{(t,t')}} \qquad t' < t. \quad \text{(A.12)}$$

Hence, the response terms read

$$\begin{aligned}
\mathcal{G}^{(t,t')} &= -\tau^{(t)}\delta_{tt'} - \tau^{(t)}\mathbb{E}\left[f_t(\gamma^{(t-1)}; y)p^{(t,t')}\right] \\
&= -\tau^{(t)}\delta_{tt'} - \tau^{(t)}\mathrm{Pr}(p^{(t,t')} = 1)\mathbb{E}\left[f_t'(\gamma^{(t-1)})p^{(t,t')}|p^{(t,t')} = 1\right] \\
&= -\tau^{(t)}\delta_{tt'} - \tau^{(t)}\eta(1-\eta)^{t-1-t'}\mathbb{E}\left[f_t'(\phi_{aux}^{(t')})\right] \\
&= -\tau^{(t)}\delta_{tt'}
\end{aligned} \quad \text{(A.13)}$$

where for convenience we have introduced the function

$$f_t(x,y) \doteq \frac{m_{\nu^{(t-1)}}(x,y)}{\mathbb{E}[m'_{\nu^{(t-1)}}(x,y)]} - x \tag{A.14}$$

which fulfills the divergence-free property $\mathbb{E}[f'_t(x,y)] = 0$. Thereby, we get

$$\boldsymbol{\phi}^{(t)}_{aux} = \boldsymbol{J}\boldsymbol{m}^{(t)} - \hat{\mathcal{G}}^{(t,t)}\boldsymbol{m}^{(t)} \tag{A.15}$$

$$= \boldsymbol{J}\boldsymbol{m}^{(t)} - \mathrm{R}(-\tau^{(t)})\boldsymbol{m}^{(t)} = \boldsymbol{\phi}^{(t)}. \tag{A.16}$$

The effective stochastic process (w.r.t. dynamical functional analysis) of $\phi_i^{(1:T)}$ becomes then a Gaussian process as

$$\phi^{(1:K)} \sim \mathcal{N}(\theta\hat{\mathcal{B}}, \mathcal{C}) \tag{A.17}$$

We next use the result $\mathcal{G}^{(t,t')} = -\tau^{(t)}\delta_{tt'}$ to compute the the necessary order parameters $\hat{\mathcal{B}}$ and $\mathcal{C}$ in (A.4) and (A.5), respectively.

Appendix A.1.1. Computation of $\hat{\mathcal{B}}$ We have

$$\hat{\mathcal{B}}^{(t)} = \left(\sum_{n=2}^{\infty} c_n \sum_{n'=0}^{n-2} (-q)^{n'}(-\tau^{(t)})^{n-n'-2}\right)\mathcal{B}^{(t)} \tag{A.18}$$

$$= \frac{\mathcal{B}^{(t)}}{\tau^{(t)} - q}\sum_{n=2}^{\infty} c_n[(-q)^{n-1} - (-\tau^{(t)})^{n-1}] \tag{A.19}$$

$$= \frac{\mathcal{B}^{(t)}}{\tau^{(t)} - q}\sum_{n=1}^{\infty} c_n[(-q)^{n-1} - (-\tau^{(t)})^{n-1}] \tag{A.20}$$

$$= \frac{\mathcal{B}^{(t)}}{\tau^{(t)} - q}(\mathrm{R}(-q) - \mathrm{R}(-\tau^{(t)})) \tag{A.21}$$

$$= \frac{\mathcal{B}^{(t)}}{\tau^{(t)} - q}(1/q - \mathrm{R}(-\tau^{(t)})) \tag{A.22}$$

On the other hand, we have

$$\mathcal{B}^{(t)} = \tau^{(t)}\mathbb{E}[\theta(\phi^{(t)} + \tilde{\gamma}^{(t)})] \tag{A.23}$$

$$= \tau^{(t)}(q\hat{\mathcal{B}}^{(t)} + \mathbb{E}[\theta\tilde{\gamma}^{(t)}]). \tag{A.24}$$

Combining both result we easily obtain that

$$\hat{\mathcal{B}}^{(t)} = \frac{\tau^{(t)}\zeta^{(t)}\mathbb{E}[\theta\tilde{\gamma}^{(t)}]}{1 - q\tau^{(t)}\zeta^{(t)}} \quad \text{with} \quad \zeta^{(t)} \doteq \frac{\mathrm{R}(-\tau^{(t)}) - 1/q}{q - \tau^{(t)}}. \tag{A.25}$$

Appendix A.1.2. Computation of $\mathcal{C}$ Recall that

$$\mathcal{C} = \sum_{n=2}^{\infty} c_n \sum_{n'=0}^{n-2} \mathcal{G}^{n'}\mathcal{C}_m(\mathcal{G}^\top)^{n-2-n'}$$

$$\underbrace{-\sum_{n=3}^{\infty} c_n \sum_{n'=0}^{n-2}(-q)^{n'} \sum_{l=0}^{n-n'-3} (\mathcal{G}^\top)^l \mathcal{B}\mathcal{B}^\top \mathcal{G}^{n-n'-l-3}}_{\mathcal{D}}. \tag{A.26}$$

We then write

$$\mathcal{C}^{(t,t')} - \mathcal{D}^{(t,t')} = \mathcal{C}_m^{(t,t')} \sum_{n=2}^{\infty} c_n \sum_{n'=0}^{n-2} (-\tau^{(t)})^{n'} (-\tau^{(t')})^{n-2-n'} \tag{A.27}$$

$$= \mathcal{C}_m^{(t,t')} \sum_{n=2}^{\infty} c_n \frac{(-\tau^{(t')})^{n-1} - (-\tau^{(t)})^{n-1}}{\tau^{(t)} - \tau^{(t')}} \tag{A.28}$$

$$= \mathcal{C}_m^{(t,t')} \frac{\mathrm{R}(-\tau^{(t')}) - \mathrm{R}(-\tau^{(t)})}{\tau^{(t)} - \tau^{(t')}} \tag{A.29}$$

On the other hand, by construction we have

$$\mathcal{C}_m^{(t,t')} = \tau^{(t)}\tau^{(t')}(\mathbb{E}[\phi^{(t)}\phi^{(t')}] + \mathbb{E}[\phi^{(t)}\tilde{\gamma}^{(t')}] + \mathbb{E}[\phi^{(t')}\tilde{\gamma}^{(t)}] + \mathbb{E}[\tilde{\gamma}^{(t)}\tilde{\gamma}^{(t')}]) \tag{A.30}$$

$$= \tau^{(t)}\tau^{(t')}(\mathcal{C}^{(t,t')} + q\hat{\mathcal{B}}^{(t)}\hat{\mathcal{B}}^{(t')} + \mathbb{E}[\phi^{(t)}\tilde{\gamma}^{(t')}] + \mathbb{E}[\phi^{(t')}\tilde{\gamma}^{(t)}] + \mathbb{E}[\tilde{\gamma}^{(t)}\tilde{\gamma}^{(t')}]) \tag{A.31}$$

$$= \tau^{(t)}\tau^{(t')}(\mathcal{C}^{(t,t')} + q\hat{\mathcal{B}}^{(t)}\hat{\mathcal{B}}^{(t')} + \hat{\mathcal{B}}^{(t)}\mathbb{E}[\theta\tilde{\gamma}^{(t')}] + \hat{\mathcal{B}}^{(t')}\mathbb{E}[\theta\tilde{\gamma}^{(t)}] + \mathbb{E}[\tilde{\gamma}^{(t)}\tilde{\gamma}^{(t')}]) \tag{A.32}$$

where in the last line we have invoked the results

$$\mathbb{E}[\phi^{(t)}\tilde{\gamma}^{(t')}] = \mathbb{E}[\phi^{(t)} f_{t'}(\gamma^{(t'-1)})] \tag{A.33}$$

$$= (1-\eta)\mathbb{E}[\phi^{(t)} f_{t'}(\gamma^{(t'-2)})] + \eta\mathbb{E}[\phi^{(t)} f_{t'}(\phi^{(t'-1)})] \tag{A.34}$$

$$= (1-\eta)\mathbb{E}[\phi^{(t)} f_{t'}(\gamma^{(t'-2)})] + \eta\hat{\mathcal{B}}^{(t)}\mathbb{E}[\theta f_{t'}(\phi^{(t'-1)})] \tag{A.35}$$

$$= \hat{\mathcal{B}}^{(t)}\mathbb{E}[\theta\tilde{\gamma}^{(t')}]. \tag{A.36}$$

The equation (A.35) follows from the Stein' lemma. Then, by invoking (A.32) in (A.29) we get

$$\mathcal{C}^{(t,t')} = \frac{\mathcal{D}^{(t,t')} + \mathcal{Q}^{(t,t')}\left(q\hat{\mathcal{B}}^{(t)}\hat{\mathcal{B}}^{(t')} + \hat{\mathcal{B}}^{(t)}\mathbb{E}[\theta\tilde{\gamma}^{(t')}] + \hat{\mathcal{B}}^{(t')}\mathbb{E}[\theta\tilde{\gamma}^{(t)}] + \mathbb{E}[\tilde{\gamma}^{(t)}\tilde{\gamma}^{(t')}]\right)}{1 - \mathcal{Q}^{(t,t')}} \tag{A.37}$$

We complete the derivation by simplifying $\mathcal{D}^{(t,t')}$: Firstly, for $\tau^{(t)} \neq \tau^{(t')}$ we have

$$\mathcal{D}^{(t,t')} = -\mathcal{B}^{(t)}\mathcal{B}^{(t')} \sum_{n=3}^{\infty} c_n \sum_{n'=0}^{n-2} (-q)^{n'} \sum_{l=0}^{n-n'-3} (-\tau^{(t)})^{l} (-\tau^{(t')})^{n-n'-l-3} \tag{A.38}$$

$$= \frac{\mathcal{B}^{(t)}\mathcal{B}^{(t')}}{\tau^{(t)} - \tau^{(t')}} \sum_{n=3}^{\infty} c_n \sum_{n'=0}^{n-2} (-q)^{n'} \{(-\tau^{(t)})^{n-n'-2} - (-\tau^{(t')})^{n-n'-2}\} \tag{A.39}$$

$$= \frac{\mathcal{B}^{(t)}\mathcal{B}^{(t')}}{\tau^{(t)} - \tau^{(t')}} \left[\frac{\mathrm{R}(-\tau^{(t)}) - 1/q}{q - \tau^{(t)}} - \frac{\mathrm{R}(-\tau^{(t')}) - 1/q}{q - \tau^{(t')}}\right] \tag{A.40}$$

where we used the fact that

$$\sum_{n=3}^{\infty} c_n \sum_{n'=0}^{n-2} (-q)^{n'} (-\tau^{(t)})^{n-n'-2} = \frac{1}{\tau^{(t)} - q} \sum_{n=3}^{\infty} c_n [(-q)^{n-1} - (-\tau^{(t)})^{n-1}] \tag{A.41}$$

$$= \frac{1}{\tau^{(t)} - q} \sum_{n=1}^{\infty} c_n [(-q)^{n-1} - (-\tau^{(t)})^{n-1}] - 1 \tag{A.42}$$

$$= \frac{\mathrm{R}(-\tau^{(t)}) - \mathrm{R}(-q)}{q - \tau^{(t)}} - c_2 \tag{A.43}$$

$$= \frac{\mathrm{R}(-\tau^{(t)}) - 1/q}{q - \tau^{(t)}} - c_2 \tag{A.44}$$

Moreover, in the equal-time case, we have

$$\mathcal{D}^{(t,t)} = \frac{\mathcal{B}^{(t)} \mathcal{B}^{(t)}}{q - \tau^{(t)}} \left[\frac{\mathrm{R}(-\tau^{(t)}) - 1/q}{q - \tau^{(t)}} - \mathrm{R}'(-\tau^{(t)}) \right]. \tag{A.45}$$

Appendix B. Derivation of (33)

For convenience, we define the single time deviations $\Delta_{\gamma,\phi}^{(t)} \doteq \mathcal{C}_{\gamma}^{\star} - \mathcal{C}_{\gamma,\phi}^{(t)}$. Furthermore, from (26) we write the recursion

$$\Delta_{\gamma}^{(t)} = (1 - \eta) \Delta_{\gamma}^{(t-1)} + \eta \Delta_{\phi}^{(t)}. \tag{B.1}$$

Moreover, we introduce the rate $\mu_{\phi} \doteq \lim_{t \to \infty} \frac{\Delta_{\phi}^{(t)}}{\Delta_{\phi}^{(t-1)}}$. From (B.1) it follows that $\mu_{\gamma} = \mu_{\phi}$. We next compute μ_{ϕ}. To this end, from (27) and (28) we firstly write $\Delta_{\phi}^{(t)}$ in the form

$$\Delta_{\phi}^{(t)} = \mathcal{C}_{\gamma}^{\star} - c^{(t)} \hat{\mathcal{B}}^{\star} \hat{\mathcal{B}}^{(t)} - \frac{\mathcal{Q}^{(t)}}{1 - \mathcal{Q}^{(t)}} \mathbb{E}[\tilde{\gamma}^{(t)} \tilde{\gamma}^{\star}] \tag{B.2}$$

for an appropriately computed constant sequence $c^{(t)}$ (which does not depend on $\Delta_{\phi}^{(t-1)}$). Then, from (30) we further write

$$\begin{aligned} \Delta_{\phi}^{(t)} = {} & \mathcal{C}_{\gamma}^{\star} - c^{(t)} \hat{\mathcal{B}}^{\star} \hat{\mathcal{B}}^{(t)} - (1 - \eta) \frac{\mathcal{Q}^{(t)} (1 - \mathcal{Q}^{(t-1)})}{(1 - \mathcal{Q}^{(t)}) \mathcal{Q}^{(t-1)}} (\mathcal{C}_{\gamma}^{\star} - c^{(t-1)} \hat{\mathcal{B}}^{\star} \hat{\mathcal{B}}^{(t-1)}) \\ & + (1 - \eta) \frac{\mathcal{Q}^{(t)} (1 - \mathcal{Q}^{(t-1)})}{(1 - \mathcal{Q}^{(t)}) \mathcal{Q}^{(t-1)}} \Delta_{\phi}^{(t-1)} - \eta \frac{\mathcal{Q}^{(t)}}{1 - \mathcal{Q}^{(t)}} \mathbb{E}[f_t'(\phi^{(t-1)}) f_{\star}'(\phi^{\star})]. \end{aligned} \tag{B.3}$$

Thereby, we get the derivative

$$g_t(\Delta_{\phi}^{(t-1)}) \doteq \frac{\partial \Delta_{\phi}^{(t)}}{\partial \Delta_{\phi}^{(t-1)}} = (1 - \eta) \frac{\mathcal{Q}^{(t)} (1 - \mathcal{Q}^{(t-1)})}{(1 - \mathcal{Q}^{(t)}) \mathcal{Q}^{(t-1)}} + \eta \frac{\mathcal{Q}^{(t)}}{1 - \mathcal{Q}^{(t)}} \mathbb{E}[f_t'(\phi^{(t-1)}) f_{\star}'(\phi^{\star})]. \tag{B.4}$$

We then obtain the rate as

$$\mu_\phi = \lim_{t\to\infty} g_t(0) = (1-\eta) + \eta \frac{(\chi^\star)^2 \mathrm{R}'(-\chi^\star)}{1-(\chi^\star)^2\mathrm{R}'(-\chi^\star)} \mathbb{E}[(f'_\star(\phi^\star))^2] \tag{B.5}$$

$$= (1-\eta) + \eta \frac{\mathrm{R}'(-\chi^\star)}{1-(\chi^\star)^2\mathrm{R}'(-\chi^\star)} (\mathbb{E}[(m'_{\nu^\star}(\phi^\star))^2] - (\chi^\star)^2) \tag{B.6}$$

$$= 1 - \eta \frac{1-\mathbb{E}[(m'_{\nu^\star}(\phi^\star))^2]\mathrm{R}'(-\chi^\star)}{1-(\chi^\star)^2\mathrm{R}'(-\chi^\star)}. \tag{B.7}$$

References

[1] Kabashima Y 2003 *Journal of Physics A: Mathematical and General* **36** 11111

[2] Bolthausen E 2014 *Communications in Mathematical Physics* **325** 333–366.

[3] Bayati M and Montanari A 2011 *IEEE Transactions on Information Theory* **57** 764–785

[4] Mimura K and Okada M 2014 *IEEE Transactions on Information Theory* **60** 3645–3670

[5] Opper M, Çakmak B and Winther O 2016 *Journal of Physics A: Mathematical and Theoretical* **49** 114002

[6] Çakmak B and Opper M 2019 *Phys. Rev. E* **99**(6) 062140

[7] Çakmak B and Opper M 2020 *Journal of Statistical Mechanics: Theory and Experiment* **2020** 103303

[8] Takeuchi K 2020 *IEEE Transactions on Information Theory* **66** 368–386

[9] Rangan S, Schniter P and Fletcher A K 2019 *IEEE Transactions on Information Theory* **65** 6664–6684

[10] Fletcher A K, Rangan S and Schniter P 2018 Inference in deep networks in high dimensions *2018 IEEE International Symposium on Information Theory (ISIT)* (IEEE) pp 1884–1888

[11] Fan Z 2020 *arXiv preprint arXiv:2008.11892*

[12] Thousless D J, Andersen P W and Palmer R G 1977 *Philosophical Magazine* **35** 593–601

[13] Parisi G and Potters M 1995 *Journal of Physics A: Mathematical and General* **28** 5267

[14] Opper M and Winther O 2001 *Physical Review E* **64** 056131–(1–14)

[15] Kabashima Y 2008 *Journal of Physics: Conference Series* **95**

[16] Donoho D L, Maleki A and Montanari A 2009 *Proceedings of the National Academy of Sciences* **106** 18914–18919

[17] Minka T P 2001 Expectation propagation for approximate Bayesian inference *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*

[18] Ma J and Ping L 2017 *IEEE Access* **5** 2020–2033

[19] Manoel A, Krzakala F, Tramel E and Zdeborová L 2015 Swept approximate message passing for sparse estimation *Proceedings of the 32nd International Conference on Machine Learning* (*Proceedings of Machine Learning Research* vol 37) ed Bach F and Blei D (Lille, France: PMLR) pp 1123–1132 URL `https://proceedings.mlr.press/v37/manoel15.html`

[20] Martin P C, Siggia E D and Rose H A 1973 *Physical Review A* **8** 423

[21] Çakmak B and Opper M 2020 *Journal of Physics A: Mathematical and Theoretical* **53** 274001

[22] Sommers H J 1987 *Physical review letters* **58** 1268

[23] Sollich P and Barber D 1997 *EPL (Europhysics Letters)* **38** 477

[24] Mignacco F, Krzakala F, Urbani P and Zdeborová L 2020 Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification (*Preprint* `2006.06098`)

[25] Çakmak B and Opper M 2021 *Phys. Rev. E* **103**(3) L030101 URL `https://link.aps.org/doi/10.1103/PhysRevE.103.L030101`

[26] Takahashi T and Kabashima Y 2020 Macroscopic analysis of vector approximate message passing in a model mismatch setting *2020 IEEE International Symposium on Information Theory (ISIT)* (IEEE) pp 1403–1408

[27] Opper M and Winther O 2000 *Neural computation* **12** 2655–2684

[28] Rasmussen C E and Williams C K 2006 *Gaussian processes for machine learning* vol 1 (Cambridge: MIT press)

[29] Neal R M 1997 *arXiv preprint physics/9701026*

[30] Çakmak B and Opper M 2018 Expectation propagation for approximate inference: Free probability framework *2018 IEEE International Symposium on Information Theory (ISIT)* (Piscataway, NJ, USA: IEEE) pp 1276–1280 ISSN 2157-8117

[31] Eisfeller H and Opper M 1992 *Physical Review Letters* **68** 2094

[32] Akemann G, Baik J and Di Francesco P (eds) 2011 *The Oxford Handbook of Random Matrix Theory* (Oxford University Press)