5 The TAP Approach to Intensive and Extensive Connectivity Systems

Yoshiyuki Kabashima and David Saad

The *Thouless-Anderson-Palmer* (TAP) approach was originally developed for analysing the Sherrington-Kirkpatrick model in the study of spin glass models and has been employed since then mainly in the context of extensively connected systems whereby each dynamical variable interacts weakly with the others. Recently, we extended this method for handling general intensively connected systems where each variable has only $\mathcal{O}(1)$ connections characterised by strong couplings. However, the new formulation looks quite different with respect to existing analyses and it is only natural to question whether it actually reproduces known results for systems of extensive connectivity. In this chapter, we apply our formulation of the TAP approach to an extensively connected system, the Hopfield associative memory model, showing that it produces identical results to those obtained by the conventional formulation.

1 Introduction

The Bayesian approach has been successfully and efficiently employed in various inference problems, especially in cases where the data set provided is small with respect to the number of parameters to be determined. Some of the more successful applications have been in the areas of neural networks [11; 22], image restoration [13; 21], error correcting codes [29; 32; 20; 12; 6; 7; 8; 9; 16; 34] etc. There is growing interest in these methods within the physics community, leading to the formation of links between the Bayesian approach and methods that have developed independently in the various sub-disciplines, and in particular in the field of statistical physics [4],

A major difficulty associated with the application of Bayesian methods is the huge computational cost when the number of dynamical variables is large. Since exact computation becomes practically infeasible in such cases, it is inevitable to resort to approximations. One of the most commonly used approximation methods is the Monte Carlo sampling technique, in which the true posterior distribution is approximated by a sampling procedure generated by the appropriate stochastic process. However, the necessary sample size may also prove problematic rendering the method impractical. The quest for more efficient approximations, which are practicable in a broad range of scenarios, is now an important research a variety of research fields.

The family of mean field approximations (MFA) represent one of the most promising approaches. The spirit of the MFA is simple; to approximate a true intractable distribution with a tractable one, which is factorizable with respect to

dynamical variables. Since the factorized model can usually be calculated quite easily, mostly by a deterministic algorithm, the required computation is usually significantly less than that of sampling techniques. Mean field approaches have been developed within the physics community and include a large number of variations, depending on the objectives of the calculation and the properties of the system examined. As the similarity between Bayesian statistics and statistical physics has been identified [10; 35], and the benefits of using MFA methods has been widely recognized, they have been employed in a variety of inference problems formulated within the Bayesian framework. One of the most popular and well known approach is the *Thouless-Anderson-Palmer* (TAP) approximation [33], which will be the focus of the current chapter.

The TAP approach has been originated in the physics community as a refinement of the mean field approximation in analyzing a specific type of disordered systems, where dynamical variables are interacting with each other via randomly predetermined (quenched) couplings. In contrast to the replica method [15], the main approach for analysing disordered systems where one obtains expressions for the typical macroscopic properties averaged over the quenched randomness, the TAP approach enables one to compute thermal averages of the dynamical variables for a given realization of the randomness.

Originally, the TAP approach was introduced for studying the Sherrington-Kirkpatrick (SK) model [30] of spin glass; numerous experiments validated the results obtained by this approach, showing that it reproduces results predicted by the replica method, which are considered exact in the thermodynamic limit [18]. Later on, the TAP approach was employed in other problems of a similar nature, such as the analysis of the Hopfield model [15; 17], the perceptron capacity calculation [14] etc, where it again showed consistency with the predictions obtained by the replica method.

These studies point to the potential use of the TAP approach as a practical algorithm which provides exact thermal averages of quantities depending on the dynamical variables in general disordered systems; this can be carried out in in practical time scales in spite of the fact that the averaging itself might be computationally hard. It is somewhat surprising that the potential of the TAP approach had not been fully appreciated until 1996 when Opper and Winther [23] employed it as a learning algorithm for determining the perceptron weights, in its role as a Bayesian classifier. Using the TAP approach as an efficient algorithm within the Bayesian approach methods is highly promising and has been drawing much attention in recent years.

Historically, the TAP approach has been developed mainly in the context of extensively connected systems where each dynamical variable interacts weakly with all the others. Recently, we extended this method to handle general intensively connected systems where each variable has only $\mathcal{O}(1)$ connections characterized by strong couplings [6]. However, the relation between the new formulation and the existing analyses (for extensively connected systems) is unclear; and raises a question about its ability to reproduce known results obtained for systems of extensive connectivity. The aim of the current article is to bridge the two approaches

and to answer this question.

For this purpose, we will apply the new formulation to the Hopfield model of associative memory, a non-trivial example of an extensively connected system, showing that it reproduces the known results obtained from conventional methods in the limit of extensive connectivity. This implies that the new approach provides a more general framework that covers both intensively and extensively connected systems.

This chapter is organized as follows: In the next section, we introduce the general framework of the problem considered. In section 3, we provide a general formulation of the TAP approach, which can be used for both intensively and extensively connected systems. In this formulation, we derive self-consistent equations between auxiliary distributions; the derivation is based on a tree approximation, which is considered as a generalization of the conventional *cavity method* [15]. It is also shown that the same equations can be derived from a variational principle with respect to a certain functional. In section 4, the new formulation is applied to investigate the Hopfield model of associative memory. We compare the results obtained using several methods, and discuss the conditions under which the TAP approach provides a good approximation. The final section is devoted to summarising the results and for suggesting future research directions.

2 The general framework

The approach presented is applicable to a variety of systems including variables of both binary and continuous representations. However, for simplicity and transparency, we will restrict the analysis presented here to systems comprising N Ising spins $S_{i=1,\ldots,N} \in [-1,+1]$. We represent the Hamiltonian of this system by

$$\mathcal{H}(\boldsymbol{S}|\mathcal{D}) = h_0(\boldsymbol{S}) + \sum_{\mu=1}^{P} h(\boldsymbol{S}|\boldsymbol{d}_\mu), \tag{1}$$

where $\mathcal{D} = \{\boldsymbol{d}_{\mu=1,\ldots,P}\}$ are the predetermined (or quenched, fixed) random variables whose correlations are supposed to be sufficiently weak. Within the statistical physics framework, this representation of the Hamiltonian leads to the following Boltzmann distribution

$$\mathcal{P}_B(\boldsymbol{S}|\mathcal{D},\beta) = \frac{e^{-\beta\mathcal{H}(\boldsymbol{S}|\mathcal{D})}}{\mathcal{Z}(\mathcal{D},\beta)} \tag{2}$$

where $\mathcal{Z}(\mathcal{D},\beta) = \mathrm{Tr}_{\boldsymbol{S}} e^{-\beta\mathcal{H}(\boldsymbol{S}|\mathcal{D})}$ is termed the partition function. Then, our problem may be defined as the computation of the averages

$$m_l = \mathop{\mathrm{Tr}}_{\boldsymbol{S}} S_l \, \mathcal{P}_B(\boldsymbol{S}|\mathcal{D},\beta), \quad (l = 1,\ldots,N), \tag{3}$$

in practical time scales.

Many problems considered in statistical physics of disordered systems are represented in this form by choosing a specific expression for the Hamiltonian. For

example, the SK model is obtained by setting the elements of the Hamiltonian (1) to

$$h_0(\boldsymbol{S}) = -h\sum_{l=1}^{N} S_i, \quad h(\boldsymbol{S}|J_{\langle ij\rangle}) = -J_{\langle ij\rangle} S_i S_j, \tag{4}$$

where $h, J > 0$ and the components of J are taken from a normal distribution of zero mean and J^2/N variance, $J_{\langle ij\rangle} \sim \mathcal{N}(0, J^2/N)$. The Hopfield model, which will be at the focus of the current analysis, corresponds to the case

$$h_0(\boldsymbol{S}) = -h\sum_{l=1}^{N} \xi_i^0 S_i, \quad h(\boldsymbol{S}|\boldsymbol{\xi}^\mu) = -\frac{1}{2}\left(\frac{\boldsymbol{\xi}^\mu \cdot \boldsymbol{S}}{\sqrt{N}}\right)^2 \tag{5}$$

where $h > 0$ is a positive field and $\boldsymbol{\xi}^{\mu=0,\ldots,P}$ are uncorrelated binary random patterns generated according to distribution $\mathcal{P}(\xi_i^\mu = \pm 1) = 1/2$, $\forall i$. Notice that the Hamiltonian of the Hopfield model seemingly becomes similar to that of the SK model by first defining the couplings as $J_{\langle ij\rangle} = (1/N)\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu (1-\delta_{ij})$ and then taking the gauge transformation $\xi_i^0 S_i \to S_i$, $J_{\langle ij\rangle}\xi_i^0\xi_j^0 \to J_{\langle ij\rangle}$. However, the assumption about the weak correlations among the quenched variables $\boldsymbol{d}_\mu$, which are the couplings $J_{\langle ij\rangle}$ in the SK model and the patterns $\boldsymbol{\xi}^\mu$ in the Hopfield model, prevents us from moving freely between the two models, as it should obey the restriction of the Hamiltonian decomposition (1).

Although we have presented the model within the framework of statistical physics and used the corresponding terminology, the same framework is applicable to a wide range of more general models in the framework of Bayesian statistics. Considering general statistical models of the form

$$\mathcal{P}_0(\boldsymbol{S}) \sim e^{-\beta h_0(\boldsymbol{S})}, \quad \mathcal{P}(\boldsymbol{d}|\boldsymbol{S}) \sim e^{-\beta h(\boldsymbol{S}|\boldsymbol{d})}, \tag{6}$$

one can easily link the Boltzmann distribution (2) to posterior distribution of the *parameter* $\boldsymbol{S}$ having observed the *data set* $\mathcal{D}$

$$\mathcal{P}_B(\boldsymbol{S}|\mathcal{D},\beta) = \frac{e^{-\beta h_0(\boldsymbol{S})}\prod_{\mu=1}^{P} e^{-\beta h(\boldsymbol{S}|\boldsymbol{d}_\mu)}}{\mathcal{Z}(\mathcal{D},\beta)} = \frac{\mathcal{P}_0(\boldsymbol{S})\prod_{\mu=1}^{P}\mathcal{P}(\boldsymbol{d}_\mu|\boldsymbol{S})}{\mathcal{P}(\mathcal{D})}, \tag{7}$$

where $\mathcal{P}(\mathcal{D}) = \mathrm{Tr}_{\boldsymbol{S}}\mathcal{P}_0(\boldsymbol{S})\prod_{\mu=1}^{P}\mathcal{P}(\boldsymbol{d}_\mu|\boldsymbol{S})$.

One might feel that the Ising spin assumption on the parameter $\boldsymbol{S}$ is rather artificial within the framework of Bayesian statistics. However, one can find examples which naturally satisfy this assumption, for instance in the area of *error-correcting codes*. It has been shown [31; 32; 7; 8], that the decoding problem in a family of error-correcting codes, termed *low-density parity check codes* [3; 12], may be formulated in the current framework by setting

$$h_0(\boldsymbol{S}) = -\frac{F}{\beta}\sum_{l=1}^{N} S_l, \quad h(\boldsymbol{S}|J_\mu) = -J_\mu S_{i_{\mu,1}}\ldots S_{i_{\mu,K}}, \tag{8}$$

where the additive field F represents prior knowledge about the possibly sparse message and J_μ is a coupling indicator used in examining the parity check conditions among the connected message bits $S_{i_{\mu,1}},\ldots,S_{i_{\mu,K}}$, represented by Ising spins. As is shown in [32], the optimal parameter β is determined by the channel noise, taking the value of Nishimori's temperature [19] which becomes

$\beta = (1/2)\ln(\mathcal{P}(+1|+1)/\mathcal{P}(+1|-1))$ for the binary symmetric channel. In the next chapter we will show how the TAP approach may be employed as a decoding algorithm in this scenario and will analyse its performance and its relation to the commonly used Belief Propagation (BP) algorithm [2].

3 The TAP approach

We now introduce a general formulation of the TAP approach to the system characterized by a Hamiltonian of the form (1). Conventionally, there have been three approaches for deriving the same self-consistent equations known as the TAP equations. The first approach is the *cavity method* [33; 15]. This is based on a correction of the naive MFA by subtracting the self-induced field, referred to as the Onsager's reaction term, in a set of self-consistent equations. The second approach is *Plefka's expansion* [28], which first evaluates the free energy using a Taylor expansion with respect to random couplings, and then derives the TAP equations from a variational condition imposed on the approximated free energy. The final one is the *Parisi-Potters's heuristics* [26; 24], which is another strategy to evaluate the free energy, based on a strong assumption that the contribution from the Onsager's reaction field in the free energy is independent of the prior employed.

The formulation that we will introduce below can be considered as a generalization of the cavity method [6]. However, the strategy used in our approach is not based on refining the result obtained by the naive MFA, i.e., by evaluating Onsager's reaction terms via an expansion with respect to the small couplings; this strategy cannot be extended to intensively connected systems as the influence of each coupling is significant and its removal cannot be regarded as a correction. Instead, we introduce auxiliary distributions to eliminate the self-induced fields, assuming a local tree-like structure representing the interaction at each spin site; we then determine the distributions in a self-consistent way by iteratively solving the equations obtained.

Cavity method

Given a Hamiltonian of the form (1), we start our formulation by assigning a Boltzmann weight to each quenched variable (or data) $\boldsymbol{d}_{\mu=1,\ldots,P}$ as

$$e^{-\beta h(\boldsymbol{S}|\boldsymbol{d}_\mu)}. \tag{9}$$

The remaining term will be restricted to the case of a factorizable prior

$$e^{-\beta h_0(\boldsymbol{S})} = e^{\sum_{l=1}^{N} F_l S_l}, \tag{10}$$

as is appropriate in many of the relevant cases.

Furthermore, we assume the following three properties for the objective system; these are required for constructing a valid MFA.

1. The Boltzmann distribution (2) can be approximated by a factorizable distribution with respect to dynamical variables $S_{l=1,\ldots,N}$.
2. The influence of the data set $\mathcal{D}$ on a specific site S_l is also factorizable with

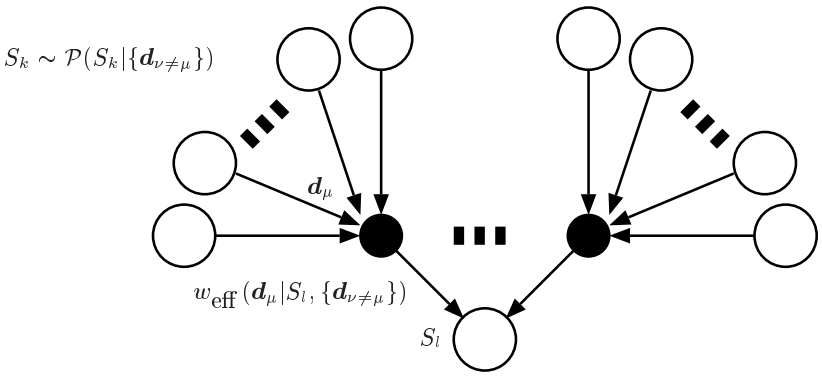

Figure 5.1
The tree-like architecture assumed locally at each spin site. White circles represent the dynamical variables S_l while black circles stand for the quenched variables $\boldsymbol{d}_\mu$. It should be emphasized that this architecture does not represent the actual connectivity but the decomposition of the Hamiltonian in the right hand side of Eq.(1), following the weak correlation assumption on the quenched variables $\boldsymbol{d}_{\mu=1,\ldots,P}$. Dynamical variables $\{S_{k\neq l}\}$ which are connected to different quenched variables are considered as components of different systems.

respect to the quenched variables $\boldsymbol{d}_{\mu=1,\ldots,P}$.

3. The secondary contribution of a single variable S_l or $\boldsymbol{d}_\mu$, other than the one estimated directly, is small and can be isolated. Therefore, at each spin S_l, we can assume a tree-like architecture, depicted in Fig.5.1 describing the influence of neighboring spins on a particular site.

These assumptions are used to provide the TAP equations

$$w_{\rm eff}\,(\boldsymbol{d}_\mu \mid S_l,\{\boldsymbol{d}_{\nu\neq\mu}\}) = \mathop{\rm Tr}_{\{S_{k\neq l}\}} e^{-\beta h(\boldsymbol{S}|\boldsymbol{d}_\mu)} \prod_{k\neq l}\mathcal{P}(S_k|\,\{\boldsymbol{d}_{\nu\neq\mu}\}),$$
$$\mathcal{P}(S_l|\,\{\boldsymbol{d}_{\nu\neq\mu}\}) = a_{\mu l}\; e^{F_l S_l} \prod_{\nu\in\mathcal{M}(l)\mu} w_{\rm eff}\,(\boldsymbol{d}_\nu \mid S_l,\{\boldsymbol{d}_{\sigma\neq\nu}\})\,, \tag{11}$$

where $a_{\mu l}$ is a normalization factor.

Notice that the first equation evaluates the average influence of the newly added element $\boldsymbol{d}_\mu$ to S_l when $S_{k\neq l}$ obeys a posterior distribution determined by the "leave-one-out" data set $\{\boldsymbol{d}_{\nu\neq\mu}\}$. This represents the effective field $w_{\rm eff}\,(\boldsymbol{d}_\mu \mid S_l,\{\boldsymbol{d}_{\nu\neq\mu}\})$ produced by the data $\boldsymbol{d}_\mu$, in which the self-induced contribution from S_l and $\boldsymbol{d}_\mu$ is eliminated by assuming the tree-like description for each interaction; this corresponds to the cavity field in the conventional TAP approach [15]. In addition, note that the second equation is similar to the Bayes formula. This indicates that the stack of the cavity fields determines the posterior distribution $\mathcal{P}(S_l|\,\{\boldsymbol{d}_{\nu\neq\mu}\})$ on the basis of the leave-one-out data set $\{\boldsymbol{d}_{\nu\neq\mu}\}$. The variables $w_{\rm eff}\,(\boldsymbol{d}_\mu \mid S_l,\{\boldsymbol{d}_{\nu\neq\mu}\})$ or $\mathcal{P}(S_l|\,\{\boldsymbol{d}_{\nu\neq\mu}\})$ do not directly correspond to the true posterior distribution although they facilitate the formulation of equations (11), thus providing a closed set of self-consistent which can be solved iteratively. By taking the full set of the

cavity fields, determined self-consistently by (11), into account, one can compute the approximated marginal posterior

$$\mathcal{P}_B(S_l|\mathcal{D}) = a_l e^{F_l S_l} \prod_{\mu\in\mathcal{M}(l)} w_{\rm eff}\left(\boldsymbol{d}_\mu \mid S_l, \{\boldsymbol{d}_{\nu\neq\mu}\}\right), \tag{12}$$

where a_l is a normalization constant.

The difference between the physical distribution $\mathcal{P}_B(S_l|\mathcal{D})$ and the auxiliary one $\mathcal{P}(S_l|\{\boldsymbol{d}_{\nu\neq\mu}\})$ corresponds to Onsager's reaction field. This can be evaluated as a small correction to the self-consistent equations of the physical distributions, expanded with respect to the small couplings, in the conventional TAP approach for extensively connected systems [33; 15]. However, this difference becomes of $\mathcal{O}(1)$ in intensively connected systems, which cannot be regarded as a small perturbation, unlike the case of extensively connected systems. It is therefore difficult to derive the TAP equations (11) directly with respect to the physical distributions $\mathcal{P}_B(S_l|\mathcal{D})$ in the case of intensively connected systems.

It has been known for several cases [6] that similar equations to (11) can be derived within the framework of belief propagation, which is another convenient mathematical tool for calculating high dimensional distributions developed in the field of graphical models [27; 12; 2]. Actually, the argument used to derive the self-consistent equations (11), assuming local tree-like structures (Fig.5.1), is very similar in the TAP and BP frameworks. However, it should be emphasized that unlike in the BP approach, the tree structure in the TAP framework does not necessarily represent the actual connection architecture but is determined through the weak correlation assumption with respect to the quenched variables $\boldsymbol{d}_{\mu=1,\ldots,P}$. In this sense, the approximation used in the BP framework may be more similar to the Bethe approximation [1] which is a naive tree approximation based on the actual connectivity.

Variational principle

Some of the other MFAs can also be derived from a variational extremization with respect to a certain functional, identified as the *free energy* [25]. The existence of an expression for the free energy is useful for studying the convergence properties and the performance of MFAs by analyzing the landscape of the free energy without directly dealing with the dynamics.

Our TAP equation (11) can also be derived from a variational extremization of some cost function that we will identify as the free energy. One can easily verify that Eqs.(11) extremize a functional (TAP free energy) of the form

$$\begin{aligned}\mathcal{F}\left[\{\mathcal{P}\},\{w_{\rm eff}\}\right] = &-\sum_{\mu=1}^{P}\ln\left[\mathop{\rm Tr}_{\boldsymbol{S}} e^{-\beta h(\boldsymbol{S}|\boldsymbol{d}_\mu)}\prod_{l=1}^{N}\mathcal{P}(S_l|\{\boldsymbol{d}_{\nu\neq\mu}\})\right]\\ &+\sum_{l,\mu}\ln\left[\sum_{S_l=\pm1} w_{\rm eff}\left(\boldsymbol{d}_\mu \mid S_l,\{\boldsymbol{d}_{\nu\neq\mu}\}\right)\mathcal{P}(S_l|\{\boldsymbol{d}_{\nu\neq\mu}\})\right]\end{aligned}$$

$$- \sum_{l=1}^{N} \ln \left[\sum_{S_l=\pm 1} e^{F_l S_l} \prod_{\mu=1}^{P} w_{\rm eff} \left(\boldsymbol{d}_\mu \mid S_l, \{\boldsymbol{d}_{\nu\neq\mu}\} \right) \right]. \tag{13}$$

In other cases, the value of the free energy is linked to the distance between the true distribution and the mean field one [25; 5]. Therefore, it can be used as a measure for evaluating the accuracy of the approximation. We have not identified, so far, some distance which is linked to the TAP free energy (13); therefore, it currently cannot be linked to some performance measure of the approximation provided. This makes the motive for the extremization unclear.

To gain insight into the meaning of the TAP free energy extremization we present an alternative derivation of the coupled equations, based on the identity

$$\delta \left(\boldsymbol{S}, \widehat{\boldsymbol{S}} \right) = \underset{\{\rho(\cdot), \widehat{\rho}(\cdot)\}}{\rm ext} \left\{ \frac{\rho(\boldsymbol{S}) \widehat{\rho}(\widehat{\boldsymbol{S}})}{{\rm Tr}_{\boldsymbol{S}'} \rho(\boldsymbol{S}') \widehat{\rho}(\boldsymbol{S}')} \right\}, \tag{14}$$

where $\delta \left(\boldsymbol{S}, \widehat{\boldsymbol{S}} \right)$ represents the Kronecker tensor over all the vectors elements and extremization is taken over the full space of functions with respect to $\boldsymbol{S}$ and $\widehat{\boldsymbol{S}}$ under appropriate normalization constraints. Using this identity, calculating the logarithm of partition function $\mathcal{Z}(\mathcal{D}, \beta) = {\rm Tr}_{\boldsymbol{S}} e^{\sum_{l=1}^{N} F_l S_l} \prod_{\mu=1}^{P} e^{-\beta h(\boldsymbol{S}|\boldsymbol{d}_\mu)}$ can be formulated as a variational problem

$$\begin{aligned} - \ln \mathcal{Z}(\mathcal{D}, \beta) \;=\; & \underset{\{\rho\}, \{\widehat{\rho}\}}{\rm ext} \left\{ - \sum_{\mu=1}^{P} \ln \left[\underset{\boldsymbol{S}}{\rm Tr}\, e^{-\beta h(\boldsymbol{S}|\boldsymbol{d}_\mu)} \rho_\mu(\boldsymbol{S}) \right] \right. \\ & + \sum_{\mu=1}^{P} \ln \left[\underset{\boldsymbol{S}}{\rm Tr}\, \rho_\mu(\boldsymbol{S}) \widehat{\rho}_\mu(\boldsymbol{S}) \right] \\ & \left. - \ln \left[\underset{\boldsymbol{S}}{\rm Tr}\, e^{\sum_{l=1}^{N} F_l S_l} \prod_{\mu=1}^{P} \widehat{\rho}_\mu(\boldsymbol{S}) \right] \right\}. \end{aligned} \tag{15}$$

Functional extremization with respect to $\boldsymbol{S}$ and $\widehat{\boldsymbol{S}}$ leads to the solutions

$$\rho_\mu(\boldsymbol{S}) \propto e^{\sum_{l=1}^{N} F_l S_l} \prod_{\nu\neq\mu} e^{-\beta h(\boldsymbol{S}|\boldsymbol{d}_\nu)}, \quad \widehat{\rho}_\mu \propto e^{-\beta h(\boldsymbol{S}|\boldsymbol{d}_\nu)}. \tag{16}$$

Namely, one can reconstruct true posterior distribution as

$$\mathcal{P}_B(\boldsymbol{S}|\mathcal{D}) = \frac{e^{\sum_{l=1}^{N} F_l S_l} \prod_{\mu=1}^{P} \widehat{\rho}_\mu(\boldsymbol{S})}{{\rm Tr}_{\boldsymbol{S}'} e^{\sum_{l=1}^{N} F_l S'_l} \prod_{\mu=1}^{P} \widehat{\rho}_\mu(\boldsymbol{S}')}, \tag{17}$$

after determining $\widehat{\rho}_{\mu=1,\ldots,P}$ from eq. (15).

The current variational formulation is still general and lacks an important ingredient of our formulation: the factorized dependence of $\widehat{\rho}_\mu(\boldsymbol{S})$ on the spin variables $S_{l=1,\ldots,P}$. Restricting the test functions to those of a factorizable form one obtains the TAP equations (11)

$$\rho_\mu(\boldsymbol{S}) = \prod_{l=1}^{N} \mathcal{P}(S_l | \{\boldsymbol{d}_{\nu\neq\mu}\}), \quad \widehat{\rho}_\mu(\boldsymbol{S}) = \prod_{l=1}^{N} w_{\rm eff} \left(\boldsymbol{d}_\mu \mid S_l, \{\boldsymbol{d}_{\nu\neq\mu}\} \right), \tag{18}$$

as well as the TAP free energy (13).

An important question is to identify the characteristics of the functions which

are successfully approximated by the current method. In the case of extensively connected systems, it has been shown that the TAP approach provides reasonable results when the correlations among $\boldsymbol{d}_{\mu=1,\ldots,P}$ as well as those of $\boldsymbol{S}_{l=1,\ldots,N}$ are sufficiently small [14]. However, it is still unclear what are the necessary conditions in the case of intensively connected systems. Interestingly, one can show that TAP free energy reproduces the expression obtained from the replica method in the thermodynamics limit, which is considered as exact, in the cases of intensively connected random network [16; 34].

4 Example - the Hopfield model

In contrast to the conventional approach our formulation (11) can be applied to both intensively and extensively connected systems. However, the new formulation appears to be quite different with respect to the existing analyses. Here we apply the new formulation to the Hopfield model of associative memory, showing that it reproduces the existing results.

The reason for the choice of the Hopfield model is twofold. First, it is relatively simple to analyse, and second, it provides an instructive example showing the importance of the Hamiltonian decomposition (1), in this formulation, following the weak correlation assumption on the quenched variables $\boldsymbol{d}_{\mu=1,\ldots,P}$. As is already mentioned, the Hopfield model has a similar architecture, in terms of connectivity, to that of the SK model. However, it will be shown later that different statistical properties of the quenched variables yield different solutions to the TAP equations.

Deriving the TAP equations – the new formulation

Consider a Hopfield network in which $P+1$ random patterns $\xi^P = \{\boldsymbol{\xi}^0, \ldots, \boldsymbol{\xi}^P\}$, independently generated with probability $\mathcal{P}(\xi_i^\mu = \pm 1) = 1/2$, are stored. For simplicity, we only consider the system with no external fields, where the Hamiltonian becomes

$$\mathcal{H}(\boldsymbol{S}|\xi^P) = \sum_{\mu=0}^{P} h(\boldsymbol{S}|\boldsymbol{\xi}^\mu), \tag{19}$$

and $h(\boldsymbol{S}|\boldsymbol{\xi}^\mu)$ is given as Eq.(5).

To proceed further, we have to specify a phase to focus on, for instance, the retrieval phase with respect to the pattern $\boldsymbol{\xi}^0$, which is characterized by the conditions

$$\frac{\boldsymbol{\xi}^0 \cdot \boldsymbol{m}}{N} \sim O(1) \ , \ \frac{\boldsymbol{\xi}^{\mu\geq 1} \cdot \boldsymbol{m}}{N} \sim O(N^{-1/2}) \ , \tag{20}$$

where the vector $\boldsymbol{m}$ represents the expectation value of the dynamical variables.

Since this phase strongly depends on $\boldsymbol{\xi}^0$, we have to deal with the contribution of this pattern separately from the others. For this purpose, it is convenient to assign a latent variable ϕ for $\boldsymbol{\xi}^0$ and rewrite the Boltzmann distribution as

$$\mathcal{P}_B(\boldsymbol{S}|\xi^P) \quad = \quad \frac{e^{-\beta\mathcal{H}(\boldsymbol{S}|\xi^P)}}{\mathcal{Z}(\xi^P, \beta)}$$

$$= \int d\phi \mathcal{P}_B(\phi|\xi^P)\mathcal{P}_B(\boldsymbol{S}|\xi^P,\phi), \tag{21}$$

where

$$\begin{aligned}
\mathcal{P}_B\left(\phi|\xi^P\right) &= \sqrt{\frac{N}{2\pi\beta}}e^{-\frac{N\phi^2}{2\beta}} \times \frac{\Xi(\xi^P,\phi,\beta)}{\mathcal{Z}(\xi^P,\beta)}, \\
\mathcal{P}_B\left(\boldsymbol{S}|\xi^P,\phi\right) &= \frac{e^{-\beta\sum_{\mu\geq 1}h(\boldsymbol{S}|\xi^\mu)+\phi\sum_{l=1}^N \xi_l^0 S_l}}{\Xi(\xi^P,\phi,\beta)},
\end{aligned} \tag{22}$$

and

$$\begin{aligned}
\mathcal{Z}(\xi^P,\beta) &= \mathop{\mathrm{Tr}}_{\boldsymbol{S}} e^{-\beta\mathcal{H}(\boldsymbol{S}|\xi^P)}, \\
\Xi(\xi^P,\phi,\beta) &= \mathop{\mathrm{Tr}}_{\boldsymbol{S}} e^{-\beta\sum_{\mu\geq 1}h(\boldsymbol{S}|\boldsymbol{\xi}^\mu)+\phi\sum_{l=1}^N \xi_l^0 S_l}.
\end{aligned} \tag{23}$$

For calculating Boltzmann distribution (21), we first employ the TAP approach to evaluate $\mathcal{P}_B(\boldsymbol{S}|\xi^P,\phi)$. Then, the latent variable ϕ can be determined by the saddle point method from $\mathcal{P}_B(\phi|\xi^P)$.

For Ising spin systems, it is convenient to introduce parameterizations of the form

$$\begin{aligned}
w_{\text{eff}}\left(\boldsymbol{\xi}^\mu \mid S_l, \left\{\boldsymbol{\xi}^{\nu\neq\mu}\right\}\right) &\propto \frac{1}{2}\left(1+\widehat{m}_{\mu l}S_l\right), \\
\mathcal{P}(S_l|\left\{\boldsymbol{\xi}^{\nu\neq\mu}\right\}) &= \frac{1}{2}\left(1+m_{\mu l}S_l\right).
\end{aligned} \tag{24}$$

To proceed with the calculation of the TAP equation (11) we note that since all patterns $\boldsymbol{\xi}^{\nu\neq\mu}$ are uncorrelated with the pattern $\boldsymbol{\xi}^\mu$, so are the dynamical variables S_l which are drawn from the probability distribution $\mathcal{P}(S_l|\{\boldsymbol{\xi}^{\nu\neq\mu}\})$; so that each variable S_l is uncorrelated with ξ_l^μ. This implies that following property for the overlaps

$$\frac{1}{\sqrt{N}}\sum_{k\neq l}\xi_k^\mu S_k \sim \mathcal{N}\left(\frac{1}{\sqrt{N}}\sum_{k\neq l}\xi_k^\mu m_{\mu k}\ ,\ 1-q_{\mu l}\right)\ , \tag{25}$$

where $\mathcal{N}$(mean, variance) represents the normal distribution and $q_{\mu l}=\frac{1}{N}\sum_{k\neq l}m_{\mu k}^2$; this results directly from the central limit theorem and holds for large N values and as long as the conditional probability of the variables S_k is of the form $\mathcal{P}(S_k|\{\boldsymbol{\xi}^{\nu\neq\mu}\})$. Employing the property (25) in Eq.(11) one can derive the TAP equation for $\mathcal{P}_B(\boldsymbol{S}|\xi^P,\phi)$ in the current system

$$\begin{aligned}
\widehat{m}_{\mu l} &= \frac{\beta}{N(1-\beta(1-q_{\mu l}))}\sum_{k\neq l}\xi_l^\mu\xi_k^\mu m_{\mu k}, \\
m_{\mu l} &= \tanh\left(\phi\xi_l^0+\sum_{\nu\neq\mu,0}\tanh^{-1}\widehat{m}_{\nu l}\right),
\end{aligned} \tag{26}$$

where $\mu=1,2,\ldots,P$ is the pattern index and $l=1,2,\ldots,N$ is the site index. Solving these equations enables one to compute (approximately), the averages m_l

$$m_l=\mathop{\mathrm{Tr}}_{\boldsymbol{S}} S_l\,\mathcal{P}_B(\boldsymbol{S}|\xi^P,\phi)=\tanh\left(\phi\xi_l^0+\sum_{\mu=1}^P\tanh^{-1}\widehat{m}_{\mu l}\right), \tag{27}$$

for all sites $l = 1, 2, \ldots, N$.

Comparison to known results

Eqs. (26) have been derived using our formulation to the TAP approach, and appear to be quite different from the known result [15; 17], which determines the physical averages m_l directly.

However, one can show that Eqs.(26) provide the known results in the thermodynamic limit where $N, P \to \infty$ with keeping $\alpha = P/N$ finite. Notice that the scaling assumption $\widehat{m}_{\mu l} \sim \mathcal{O}(N^{-1/2})$ implies that the auxiliary variables $m_{\mu l}$, $\widehat{m}_{\mu l}$ and $q_{\mu l}$ can be represented using only the physical averages m_l in this limit,

$$\begin{aligned} q_{\mu l} &= \frac{1}{N}\sum_{k\neq l} m_{\mu k}^2 \simeq q \equiv \frac{1}{N}\sum_{l=1} m_l^2 \ , \\ \hat{m}_{\mu l} &\simeq \frac{\beta}{N}\sum_{k=1}^{N} \xi_l^\mu \xi_k^\mu m_k - \frac{\beta m_l}{N(1-\beta(1-q))} \ , \\ m_{\mu l} &\simeq m_l - (1-m_l^2)\hat{m}_{\mu l} \ . \end{aligned} \tag{28}$$

In addition, the saddle point equation for ϕ, $(1/N)\partial \ln \mathcal{P}_B(\phi|\xi^P)/\partial\phi = 0$, provides the condition

$$\phi = \frac{\beta}{N}\sum_{l=1}^{N} \xi_l^0 m_l, \tag{29}$$

which is also determined using only physical averages m_l. Substituting relations (28) and (29) into Eqs.(26), we finally obtain the known TAP equations for the Hopfield model

$$m_l = \tanh\left(\beta\sum_{k\neq l} J_{lk} m_k - \frac{\alpha\beta^2(1-q)}{1-\beta(1-q)} m_l\right), \tag{30}$$

where $J_{lk} = \sum_{\mu=0}^{P} \xi_l^\mu \xi_k^\mu$, as given in [15; 17].

Method comparison

To investigate the accuracy of the solutions provided by the TAP equations when applied to the Hopfield model, we have numerically evaluated the overlap $M = (\sum_{l=1}^{N} \xi_l^0 m_l)/N$ by solving Eq.(30) for systems of size $N = 10000$ storing $P = 500$ patterns ($\alpha = 0.05$) with varying temperature T from 0.4 to 0.54.

For comparison, we evaluated the same quantity using three other different methods:

1. *Naive MFA* - in this case the physical averages m_l are represented as

$$m_l = \tanh\left(\beta\sum_{k\neq l} J_{lk} m_k\right), \tag{31}$$

disregarding Onsager's reaction terms.

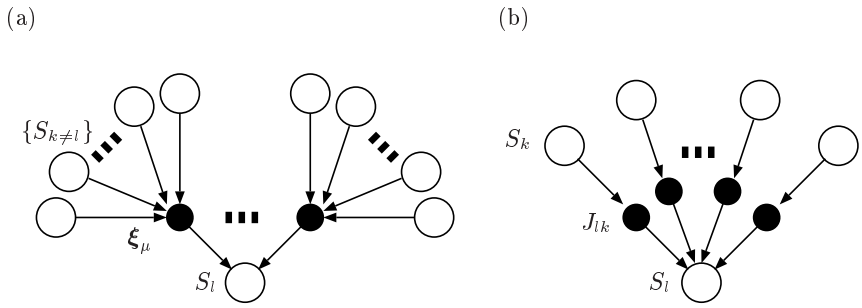

Figure 5.2
The local tree-like architecture assumed at each spin site in the TAP approach to (a) the Hopfield model and (b) the SK model.

2. *TAP equations for the SK model* [33] - which are of the form

$$m_l = \tanh\left(\beta \sum_{k\neq l} J_{lk} m_k - \sum_{k\neq l} \beta^2 J_{lk}^2 (1-m_k^2) m_l\right), \tag{32}$$

and are derived under an assumption that couplings J_{kl} are uncorrelated with one another. In the current context, this implies that Onsager's reaction is eliminated inaccurately by employing the local tree approximation depicted in Fig.5.2 (b), while the correct TAP approach (30) is derived by assuming the more appropriate tree architecture shown in Fig.5.2 (a).

3. *The replica method* - which under the replica symmetry ansatz provides exact results in the thermodynamic limit $N \to \infty$ (the AT stability is not broken in this phase for the parameter region considered $\alpha = 0.05$).

Data obtained from 100 experiments by solving Eqs.(30), (31) and (32) iteratively, together with the solution obtained from the replica symmetric theory, are shown in Fig.5.3. In solving the equations iteratively we set the initial state to $\boldsymbol{\xi}^0$ in order to verify that the solution obtained is within the correct phase. Typically $\mathcal{O}(10)$ iterations were sufficient for convergence in most cases, which implies that an approximate calculation can be performed in $\mathcal{O}(N^2)$ time steps while $\mathcal{O}(2^N)$ computation is necessary for exact calculation (except in the vicinity of the spinodal point $T \simeq 0.54$).

From Fig.5.3, it is clear that the naive MFA yields the largest overlap over all the temperature range considered. This is because of the Onsager's reaction fields, which are not compensated for in this case, and stabilize the retrieval state. This effect becomes stronger for higher temperatures as the reaction fields are proportional to thermal fluctuations. Compared to the naive MFA result, the SK's TAP equations provides smaller overlaps due to the local suppression of the reaction term assuming the tree-like architecture at each spin as shown in Fig. 5.2(b). However, the tree architecture used is not appropriate for the current system resulting in some residual contribution from the reaction terms.

Finally, we present the result of the correct TAP approach, which can also be derived from our formulation of the problem. Of course, this approach is also no

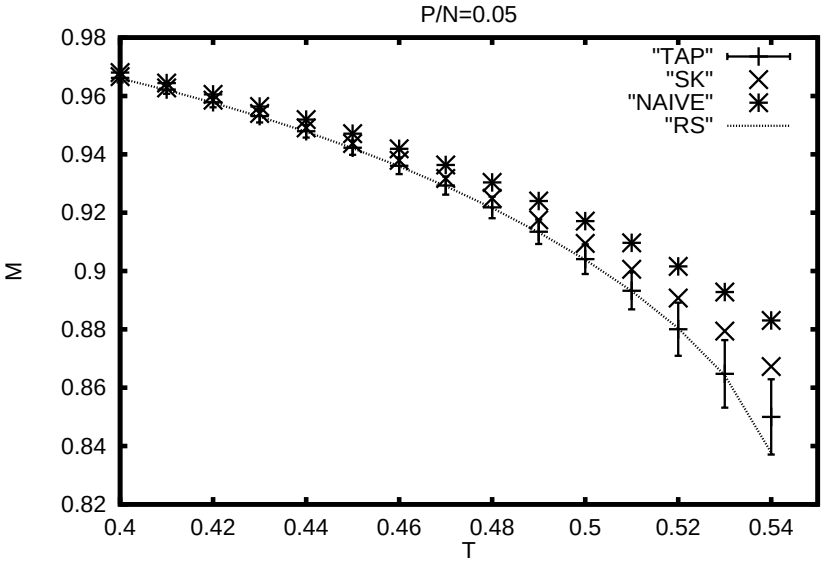

Figure 5.3
The overlap $M = (\sum_{l=1}^{N} \xi_l^0 m_l)/N$ calculated from several methods for a system of size $N = 10000$ storing $P = 500$ examples ($\alpha = 0.05$). Each marker ($*$: the naive MFA, $\times$: the TAP solution to the SK model, $+$: using the correct TAP equation) indicates the average over 100 experiments. Error bars have been added only for the correct TAP data for clarity, the error bars are similar for the other cases. The dotted curve represents the replica symmetric solution, which is considered exact in this case for $N \to \infty$.

more than an approximation for any finite system. However, the data in Fig. 5.3 indicates that the solutions obtained by iterating Eq.(30) are highly similar to the predictions by the replica method, perceived to be exact for $N \to \infty$, even for finite systems of $N = 10000$. This implies that one can construct an efficient algorithm for computing averages (for a specific phase) that runs in polynomial time (in the limit $N \to \infty$) using the TAP approach, as claimed in [23]. At the same time, the results obtained from the TAP equations derived for the SK model suggest that the correct prior knowledge about the underlying statistical structure (in the quenched variables) is important for making full use of the remarkable properties of the TAP approach.

5 Summary

In summary, we have described a formulation of the TAP approach, which can be used for intensively connected systems, based on the cavity method. The given formulation appears to be quite different from the conventional one, derived in the case of extensively connected systems. However, we have demonstrated that the known result can be reproduced from our formulation in the limit of extensive connectivity by examining the Hopfield model of associative memory. This implies that our new formulation provides a more general scheme covering both intensively

and extensive connected systems. In addition, we have showed via numerical experiments, that the correct prior knowledge about the underlying statistical structure in the quenched variables, representing the observed data in many cases of Bayesian statistics, is important for obtaining high quality approximation using the TAP approach.

Future directions of the current research include an alternative derivation of our formulation based on the methods of Plefka [28] and of Parisi-Potters [26; 24] as well as how the treatment of phases with replica symmetry breaking [15] in the current approach; both tasks are are interesting and challenging.

Acknowledgments

This work was partially supported by the program "Research For the Future" (RFTF) of the Japanese Society for the Promotion of Science (YK), by EPSRC grant GR/N00562, and a Royal Society travel grant (DS). We would also like to thank Manfred Opper for critical reading of the manuscript.

References

[1]Bethe, H.A., Proc. R. Soc. London, Ser A, **151**:552, 1935.

[2]Frey, B.J., Graphical Models for Machine Learning and Digital Communication (MIT Press, Cambridge, MA.), 1998.

[3]Gallager, R.G., IRE Trans. Info. Theory **8**:21; 1963, Low Density Parity Check Codes (MIT Press, Cambridge, MA.), 1962.

[4]Iba, Y., J. Phys. A: Math. and Gen. **32**:3875, 1999.

[5]Jordan, M.I., Gharamani, Z., Jaakkola, T.S., and Saul, L.K., in Learning in Graphical Models (ed. Jordan, MI., MIT Press, Cambridge, MA), 105, 1999.

[6]Kabashima, Y., and Saad, D., Europhys. Lett. **45**:668, 1998.

[7]Kabashima, Y., and Saad, D., Europhys. Lett. **45**:97, 1999.

[8]Kabashima, Y., Murayama, T., and Saad, D., Phys. Rev. Lett. **84**:1355, 2000.

[9]Kanter, I., and Saad, D., Phys. Rev. Lett. **83**:2660, 1999.; J. Phys. A: Math. and Gen. **33**:1675, 2000.

[10]Levin, E.N., Tishby, N., and Solla, SA. 1989, in Proceedings of the 2nd Workshop on Computational Learning Theory (ed. Warmth, M.K., and Valliant, L.G., Morgan Kaufmann, San Mateo, CA).

[11]MacKay, D.J.C., Neural Compt. **4**, 415; 1992, Neural Compt. **4**:448, 1992.

[12]MacKay, D.J.C., and Neal, R.M., Electr. Lett. **32**, 1645.; MacKay, DJC. 1999, IEEE Trans. Info. Theory **45**:399, 1996.

[13]Marroquin JL., Mitter, S., and Poggio, T., Journal of the American Statistical Institute **82**:76, 1987.

[14]Mezard, M. 1989, J. Phys. A: Math. and Gen. **22**:2181, 1989.

[15]Mezard, M., Parisi, G., and Virasoro, M.A., Spin Glass Theory and Beyond (World Scientific, Singapore), 1987.

[16]Murayama, T., Kabashima, Y., Saad, D., and Vicente, R., Phys. Rev. E. **62**:1577, 2000.

[17]Nakanishi, K., and Takayama, H., J. Phys. A: Math. and Gen. **30**:8085, 1997.

[18]Nemoto, K., and Takayama, H., J. Phys. C: Solid State Phys. **18**:L529, 1985.

[19]Nishimori, H., 1980, J. Phys. C: Solid State Phys. **13**:4071, 1980; Prog. Theor. Phys. **69**:1169, 1981.

[20]Nishimori, H., J. Phys. Soc. Jpn. **62**:2793, 1993.

[21]Nishimori, H., and Wong, K.Y.M., Phys. Rev. E **60**:132, 1999.

[22]Opper, M., and Haussler D., Phys. Rev. Lett. **66**:2677, 1991.

[23]Opper, M., and Winther O., Phys. Rev. Lett. **76**:1964, 1996.

[24]Opper, M., and Winther O., in Advances in Neural Information Processing Systems **11** (ed.

Kearns, MS. et al, MIT Press, Cambridge, MA), 309, 1999.

[25]Parisi, G., Statistical Field Theory (Addison Wesley, Redwood City, CA), 1988.

[26]Parisi, G., and Potters M., J. Phys. A: Math. and Gen. **28**, 5267, 1995.

[27]Pearl, J., Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (Morgan Kaufmann, San Francisco, CA), 1988.

[28]Plefka, T., J. Phys. A: Math. and Gen. **15**:1971, 1982.

[29]Ruján, P., Phys. Rev. Lett. **70**:2698, 1993.

[30]Sherrington, D., and Kirkpatrick, S., Phys. Rev. Lett. **35**:1792, 1975.

[31]Sourlas, N., Nature **339**:693, 1989.

[32]Sourlas, N., Europhys. Lett. **25**:159, 1994.

[33]Thouless, D.J., Anderson, P.W., and Palmer, R.G., Phil. Mag. **35**:593, 1977.

[34]Vicente, R., Saad, D., and Kabashima, Y., Phys. Rev. E **60**:5352, 1999; Vicente, R., Saad, D., and Kabashima, Y., J. Phys. A, **33**:1577, 2000.

[35]Watkin, T.L.H., Rau, A., and Biehl, M., Rev. Mod. Phys. **65**:499, 1993.