Tractable approximations for probabilistic models: The adaptive TAP mean field approach

Manfred Opper[1] and Ole Winther[2,3]
[1] *Neural Computing Research Group, Department of Computer Science and Applied Mathematics, Aston University, Birmingham B4 7ET, United Kingdom*
[2] *Theoretical Physics II, Lund University, Sölvegatan 14 A, S-223 62 Lund, Sweden*
[3] *Department of Mathematical Modelling, Technical University of Denmark B321, DK-2800 Lyngby, Denmark*

We develop an advanced mean field method for approximating averages in probabilistic data models that is based on the TAP approach of disorder physics. In contrast to conventional TAP, where the knowledge of the distribution of couplings between the random variables is required, our method adapts to the concrete couplings. We demonstrate the validity of our approach by replica calculations for a wide class of models as well as by simulations for a real data set.

Probabilistic models (for a review see e.g. [1]) find widespread applications in many areas of data modeling. Their goal is to explain complex observed data by a set of unobserved, *hidden* random variables based on the joint distribution of both sets of variables. The price that a modeler has to pay for the high degree of flexibility of these models is the vast increase in computational complexity when the number of hidden variables is large.

Both statistical inference about hidden variables and training usually require computation of marginal distributions of the hidden variables which for exact calculation demands infeasible high dimensional sums or integrals. Since similar types of calculations are ubiquitous in the computations of thermal averages, there is a great deal of interest in adopting approximation techniques from statistical physics. For a variety of cases, when a standard tool, the Monte Carlo sampling technique reaches its limits, a simple mean field (MF) method, which neglects correlations of random variables has yielded good results in a variety of probabilistic data models. The MF approximation yields a closed set of nonlinear equations for the approximate expectation values of random variables which usually can be solved in a time that only grows polynomially in the number of variables. At present, there is a growing research activity trying to overcome the limitations of the simple MF method by partly including the dependencies of variables but still keeping the approximation tractable (for a review see [2]).

Various researchers [3–12] have discussed applications of the so-called TAP MF approach, originating in the statistical physics of disordered systems, first introduced by Thouless, Anderson and Palmer (TAP) [13] to treat the Sherrington-Kirkpatrick (SK) model of disordered magnetic materials [14]. Under the assumption that the couplings (interactions) between random variables are themselves drawn at random from certain classes of distributions, the TAP equations become *exact* in the thermodynamic limit of infinitely many variables. Unfortunately, the *Onsager correction* to the simple, naive MF theory will explicitly depend on the distribution of these couplings. Two models with the same connectivities but different distributions for the couplings, like e.g. the SK model and the Hopfield model [15] have different expressions for the Onsager corrections (see e.g. [5], chapter XIII).

In order to use the TAP method as a good approximation for models of real data, the lack of knowledge of the underlying distribution of the couplings (which are usually functions of the observed data) should be compensated by an algorithm which *adapts* the Onsager correction to the *concrete* set of couplings. Simply taking the correction from a theory that *assumes* a specific distribution may lead to suboptimal performance. This letter presents a solution to this problem for an important class of probabilistic models. As a check of the validity of the approach, we show that our method leads to the exact results in the thermodynamic limit for large classes of probability distributions over the couplings.

We will consider probabilistic models of the type

$$P(\mathbf{S}) = \frac{\rho(\mathbf{S})}{Z(\boldsymbol{\theta},\mathbf{J})} \exp\left[\sum_{i<j} S_i J_{ij} S_j + \sum_i S_i \theta_i\right] \qquad (1)$$

where the set $\mathbf{S} = (S_1, \ldots, S_N)$ denotes the (hidden) random variables of the model. Any observed (i.e. fixed) quantities are assumed to be encoded in the matrix $\mathbf{J}$ and the fields $\boldsymbol{\theta}$. The term $\rho(\mathbf{S}) \equiv \prod_j \rho_j(S_j)$ is a product distribution which also contains all constraints of the S_i (the range, discreteness, etc). In its simplest version, when $\mathbf{S}$ is a real variable with positive measure ρ, the class of models (1) contains Ising models (such the SK and Hopfield models), Gaussian process models [3], probabilistic independent component analysis [16] and combinatorial optimization problems [5]. If we lift the restrictions that all variables must be real random variables, we can treat a variety of important models with dependencies between the S_i that are defined through a set of fields $\sum_{i=1}^{N} x_{ij} S_i$. We will give two examples. Bayesian learning in single layer neural networks is described by a Gibbs distribution $P(\mathbf{S}) \propto P_0(\mathbf{S}) \prod_{j=1}^m F(\sum_{i=1}^N x_{ij} S_i)$, where $\mathbf{S}$ is a weight vector of the network being trained on a number of m data vectors with components x_{ij} in a N dimensional space. P_0 is a prior distribution of the weights and F is the Likelihood quantifying the goodness of fit to the data [8]. A second example is given by the class of Bayesian belief networks on a directed graph which are promising models for adaptive expert systems. They are defined by $P(\mathbf{S}) = \prod_i P(S_i|\mathrm{pa}(S_i))$ where $S_i \in \{0,1\}$ and pa denotes the parents of S_i, i.e. the variables in the graph that feed their information into S_i via directed bonds. A specific type is the sigmoid belief networks [17], where $P(S_i|\mathrm{pa}(S_i)) = \frac{e^{S_i h_i}}{1+e^{h_i}}$ with $h_i = \sum_{j \in pa(S_i)} x_{ij} S_j$. The latter two models can be easily brought into the form (1) by the standard 'field-theoretic' trick of introducing Dirac δ-functions and their exponential representations using purely imaginary conjugate variables $\hat{\mathbf{S}} = (\hat{S}_1, \ldots, \hat{S}_m)$. This leads to an augmentation of the space of variables to the set $(\mathbf{S}, \hat{\mathbf{S}})$. The hatted variables have the complex single variable distributions $\hat{\rho}(\hat{S}) = \int \frac{dh}{2\pi i} e^{-\hat{S}h} e^{-H(h)}$ in case of the neural network model and $\hat{\rho}(\hat{S}) = \int \frac{dh}{2\pi i} e^{-\hat{S}h}/(1+e^h)$ for the belief network (where $m = N$). The augmented coupling matrix is of the form $\mathbf{J} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix}$, where $B_{ij} = x_{ij}$ and $\mathbf{A} = 0$ for the neural network and $A_{ij} = B_{ij} = x_{ij}$ for the belief net.

We will derive both an adaptive TAP-like approximation for the marginal distribution $P_i(S) \equiv \int \prod_{j \neq i} dS_j P(\mathbf{S})$ and the free energy $F(\mathbf{J}, \boldsymbol{\theta}) = -\ln Z(\mathbf{J}, \boldsymbol{\theta})$. The free energy corresponds to the negative log probability of the observed data which can be used as a yardstick for deciding which model best fits to the data.

Our derivation will be based on the cavity approach introduced by [5]. We will assume that we are not dealing with a glassy system with its many ergodic components, but that all averages are for a single state. This is (as shown for many of the teacher-student scenarios studied in the statistical mechanics of neural networks) usually expected to hold when the probabilistic model is well matched to the data. Defining the field $h_i = \sum_j J_{ij} S_j$, the marginal distribution of S_i can be written as

$$P_i(S) = \int \prod_{j \neq i} dS_j P(\mathbf{S}) = \frac{\rho_i(S)}{Z_i} e^{-H_i(S)} \ , \tag{2}$$

where we have introduced an effective single variable Hamiltonian $H_i(S)$ with corresponding partition function Z_i. Defining an auxiliary average over the distribution of the system with variable S_i left out by $\langle \ldots \rangle_{\backslash i}$, we get

$$-H_i(S) = \ln \left\langle e^{S h_i} \right\rangle_{\backslash i} = \sum_k \frac{\kappa_k^{(i)}}{k!} S^k \tag{3}$$

where $\kappa_k^{(i)}$ are the cumulants of this *cavity* distribution, i.e. $\kappa_1^{(i)} = \langle h_i \rangle_{\backslash i}$ and $\kappa_2^{(i)} = \langle h_i^2 \rangle_{\backslash i} - \langle h_i \rangle_{\backslash i}^2$ etc.

The basic physical assumption, which is the major ingredient of all cavity derivations of the TAP mean field theory [5], is that all variables S_j have only weak mutual dependencies. Mathematically expressed within the so–called clustering hypothesis [5] this becomes equivalent to the *vanishing* of all cumulants $\kappa_k^{(i)}$ with $k > 2$ for fully connected systems. In the case, where the S_j are real variables with positive measure, this corresponds to a central limit theorem for the cavity fields. Under this assumption, setting $V_i = \kappa_2^{(i)}$, we get

$$\langle h_i \rangle = \frac{1}{Z_i} \int dS \rho_i(S) \frac{\partial}{\partial S} e^{-H_i(S)} = \langle h_i \rangle_{\backslash i} + V_i \langle S_i \rangle \tag{4}$$

$$P_i(S) = \frac{\rho_i(S)}{Z_i} \exp\left[(\sum_j J_{ij} \langle S_j \rangle - V_i) S + \frac{1}{2} V_i S^2 \right] \tag{5}$$

for $i = 1, \ldots, N$. For comparison we note that the naive mean field approximation (for real random variables) is obtained by setting $V_i = 0$. So far, the approach is well known. The new aspect of our paper is in the way we compute the V_i's. Since these reaction terms account for the weak influence between random variables, they can be computed self-consistently from the matrix of susceptibilities $\chi_{ij} \equiv \frac{\partial \langle S_i \rangle}{\partial \theta_j}$. We make the approximation that upon differentiation, the V_i's are held constant which is consistent with the fact that the V_i's are expected to be selfaveraging quantities in the thermodynamic limit. Under this assumption we get from eq. (5) $\chi_{ij} = \chi_{ii} (\delta_{ij} + \sum_k (J_{ik} - V_k \delta_{ik}) \chi_{kj})$ which can be solved with respect to $\boldsymbol{\chi}$ and yields $\boldsymbol{\chi} = (\boldsymbol{\Lambda} - \mathbf{J})^{-1}$, where $\boldsymbol{\Lambda} = \mathrm{diag}\{V_i + 1/\chi_{ii}\}$ is a diagonal matrix. The *Fluctuation Dissipation Theorem* (again assuming that we deal with a single state), shows that $\boldsymbol{\chi}$ also equals the matrix of correlations $C_{ij} = \langle S_j S_k \rangle - \langle S_j \rangle \langle S_k \rangle$. By specializing to the diagonal elements, we can compute V_i as a function of $\langle S_i^2 \rangle - \langle S_i \rangle^2$ by solving

$$\langle S_i^2 \rangle - \langle S_i \rangle^2 = \frac{\partial^2 \ln Z_i}{\partial \theta_i^2} = \left[(\boldsymbol{\Lambda} - \mathbf{J})^{-1} \right]_{ii} \tag{6}$$

for $i = 1, \ldots, N$. The sets of equations (5) together with (6) constitute the first main result of this letter. They yield closed sets of equations for the first and second moments of S_i which in turn enables us to approximate the full marginal distribution of S_i and the correlation functions. A sanity check of the internal consistency of our approach is obtained by the fact that the matrix $\boldsymbol{\chi}$ must be positive definite. (If a group of the variables are complex, this has to hold for the submatrix of the real random variables).

The next task is to compute the adaptive TAP approximation to the free energy $F(\mathbf{J}, \boldsymbol{\theta}) = -\ln Z(\mathbf{J}, \boldsymbol{\theta})$. It is useful to generalize our model eq. (1) to a one parameter class of models where the interaction $\mathbf{J}$ is replaced by $s\mathbf{J}$ with $0 \leq s \leq 1$ and to define the Legendre transform (Gibbs free energy) by

$$\Phi_s(\mathbf{m},\mathbf{M}) = F(s\mathbf{J}+\boldsymbol{\lambda},\boldsymbol{\theta}+\boldsymbol{\gamma}) + \sum_i \gamma_i m_i + \sum_i \frac{\lambda_i}{2} M_i,$$

where γ_i and λ_i are external fields conjugate to S_i and S_i^2 which must be chosen to extremize the right hand side and $\boldsymbol{\lambda}$ is a diagonal matrix with entries λ_i. The solutions $\mathbf{m}^e$ and $\mathbf{M}^e$ of the sets of equations $\partial_{m_i}\Phi_s = \partial_{M_i}\Phi = 0$, determine the correct equilibrium expectation values $\langle S_i\rangle_s = m_i^e$ and $\langle S_i^2\rangle_s = M_i^e$ (the index indicates that the expectation is taken with parameter s). Our desired approximation to the free energy is finally obtained as $F(\mathbf{J},\boldsymbol{\theta}) = \Phi_1(\mathbf{m}^e,\mathbf{M}^e)$. To compute Φ_1 we differentiate Φ_s with respect to s, to show that

$$\Phi_1 = \Phi_0 - \frac{1}{2}\int_0^1 ds\, \left\{ \sum_{i,j} m_i J_{ij} m_j + \mathrm{Tr}(\boldsymbol{\chi}_s \mathbf{J}) \right\} \qquad (7)$$

with $\chi_{s,ij} = \langle S_i S_j\rangle_s - \langle S_i\rangle_s \langle S_j\rangle_s$. Inserting our TAP approximation $\boldsymbol{\chi}_s = (\boldsymbol{\Lambda}_s - s\mathbf{J})^{-1}$ and integrating, we obtain

$$\Phi_1 = \Phi_0 - \frac{1}{2}\sum_{ij} m_i J_{ij} m_j + \Delta\Phi \qquad (8)$$

$$\Delta\Phi = \frac{1}{2}\mathrm{Tr}\ln(\boldsymbol{\Lambda}-\mathbf{J}) - \frac{1}{2}\sum_i V_i \chi_{ii} + \frac{1}{2}\sum_i \ln\chi_{ii}$$

with $\chi_{ii} = M_i - m_i^2$. The first two terms constitute the naive mean field approximation to Φ and the last term $\Delta\Phi$ is the Onsager correction. Note, that this result is not equivalent to a truncation of a power series expansion of Φ to second order in s (a Plefka expansion [18]) but contains terms of all orders. A different way to derive this result is obtained from the observation that the functional form of the Onsager term V_i in the TAP equations does not depend on the specific single variable densities $\rho(\mathbf{S})$. Hence, we may compute this universal form by calculating Φ for an exactly solvable model, i.e. for a Gaussian ρ and subtract the naive mean field part. This is related to the strategy used by Parisi and Potters [19] in order to derive the TAP equations for a spin glass model with orthogonal random matrix $\mathbf{J}$.

To check the significance of our approach, we will next show that it will give the correct results for the statistical mechanics in the thermodynamic limit $N \to \infty$ for a large class of distributions of the random matrix $\mathbf{J}$. For simplicity, we specialize to models with only one type of single variable distribution $\rho_i(S) = \rho(S)$. Selfaveraging properties of the models can be computed within the replica framework by averaging the free energy over the distribution of the random matrix $\mathbf{J}$. This requires the calculation of the asymptotic scaling of the function $K_N(\mathbf{A}) \equiv \frac{1}{N}\ln\left[e^{\frac{1}{2}\mathrm{Tr}(\mathbf{AJ})}\right]_{\mathbf{J}}$ for the matrix $A_{ij} = \sum_{a=1}^n S_{ia}S_{ja}$, where the $\mathbf{S}_a$, are n replicas of the variables. Following Ref. [19] and assuming the scaling $K_N(\mathbf{A}) \simeq \mathrm{Tr}\, G(\mathbf{A}/N)$ as $N \to \infty$ where the function G characterizes the random matrix ensemble, the averaged free energy will depend only on the *single* set of orderparameters given by $q_{ab} \equiv \frac{1}{N}\sum_i S_{ia}S_{ib}$. This is characteristic for models with matrices $\mathbf{J}$ of extensive connectivity. E.g., the SK-model with coupling matrix of independent components of variance $\frac{\beta}{N}$ has $G(r) = \frac{(\beta r)^2}{4}$ and the Hopfield model with $J_{ij} = \sum_{\mu=1}^{\alpha N} x_i^\mu x_j^\mu$ and independent x_i^μ with variance $\frac{\beta}{N}$ leads to $G(r) = -\frac{\alpha}{2}(\ln(1-\beta r) + \beta r)$. Under the assumption of replica symmetry, the averaged free energy $f = -\frac{1}{N}[\ln Z]_{\mathbf{J}}$ is obtained by extremizing

$$f(q,\Delta) = -G(\Delta) + \Delta\left(qG''(\Delta) + G'(\Delta)\right) \qquad (9)$$
$$- \int Dz \ln \int dS\, \rho(S) \exp\left[\sqrt{2qG''(\Delta)}\, zS + G'(\Delta)S^2\right]$$

with respect to the off-diagonal orderparameter $q = q_{ab}$ and to $\Delta = q_{aa} - q$, where $Dz = \frac{dz}{\sqrt{2\pi}}\, e^{-z^2/2}$.

We can show the correspondence for $N \to \infty$ of the adaptive TAP method and replica theory. A disorder average gives the conventional TAP result for the Onsager coefficients: $V_i = V = 2G'(\overline{\chi})$, with $\overline{\chi} = \frac{1}{N}\sum_i [\chi_{ii}]_{\mathbf{J}}$. To compare the TAP Gibbs free energy eq. (8) with the replica symmetric free energy (9), we compute $\hat{f} = -\lim_{N,\gamma\to\infty}\frac{1}{\gamma N}\left[\ln\int d\mathbf{m}\, d\mathbf{M}\exp(-\gamma\Phi)\right]_{\mathbf{J}}$, where the paths of integration must be chosen such that the integral converges. The integral will be dominated by the values for $\mathbf{m}$ and $\mathbf{M}$ which fulfill the TAP equations. Evaluating this expression using the replica method shows that both free energies coincide, i.e. $\hat{f} = f$. It is also possible to translate the condition of positive definiteness of the susceptibility matrix $\boldsymbol{\chi}$ into the thermodynamic limit. We can show that this stability is satisfied for $1 - 2G''(\overline{\chi})\frac{1}{N}\sum_i [\chi_{ii}]^2_{\mathbf{J}} > 0$, which coincides with the well known AT stability condition of replica theory [5].

We have performed two types of simulations of the TAP approaches on Bayesian neural network learning problems. For the first case (Fig. 1) we test the self–consistency of our method on a real data set, 'Sonar – Mines versus Rocks' [20] of size $m = 104$ with with binary class labels $y_j = \pm 1$ and a $N = 60$ dimensional input space. The prior is $P_0(\mathbf{S}) \propto \exp(-\mathbf{S}\cdot\mathbf{S}/2)$ and the likelihood $F(\hat{h}_j) = \phi(y_j\hat{h}_j/\sigma)$, with $\hat{h}_j = \sum_i x_{ij}S_i$, $\phi(t) = \int_{-t}^{\infty} Dz$ and $\sigma^2 = 0.5$. We compute the prediction for the average (conjugate) cavity field $\langle\hat{h}_j\rangle_{\backslash j} = \langle\hat{h}_j\rangle - \hat{V}_j\langle\hat{S}_j\rangle$, using eq. (4). The fraction of negative terms $y_j\langle\hat{h}_j\rangle_{\backslash j}$ equals the 'leave-one-out' estimate ϵ_{loo} which provides an important practical estimator for the generalization error of the network. If our theory takes the reaction of the remaining variables correctly into account, this prediction should be close to the 'exact' average cavity field

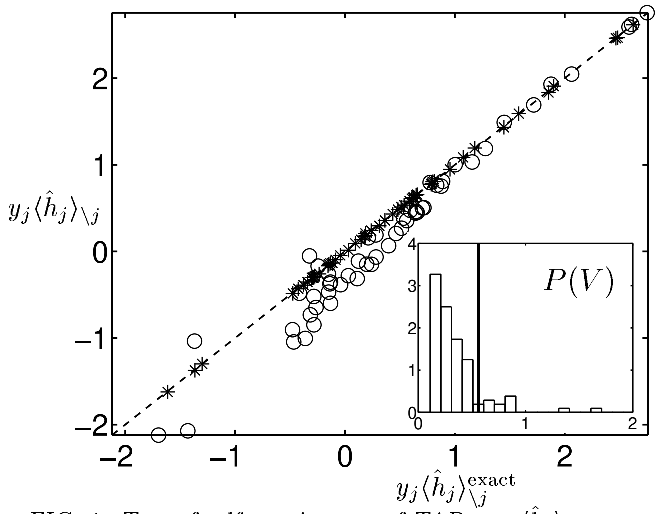

FIG. 1. Test of self–consistency of TAP – $y_j\langle\hat{h}_j\rangle_{\setminus j}$ versus $y_j\langle\hat{h}_j\rangle_{\setminus j}^{\mathrm{exact}}$. The stars/circles are for adaptive/conventional TAP. The inset shows the distribution of V_i with the thick line indicating the conventional TAP solution.

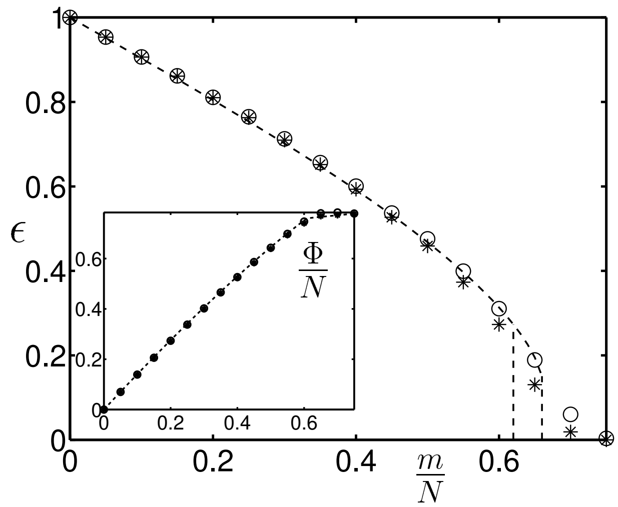

FIG. 2. Learning curve for the linear Ising perceptron – test error rate against number of training examples. The stars/circles are for adaptive/conventional TAP. The dashed lines are the replica result where the first vertical is the thermodynamic transition and the second the spinodal point where the meta-stable solution vanishes. The inset shows the corresponding normalized free energy Φ/N. The simulations are averaged over 100 runs with error bars of the size of the symbols.

obtained by leaving one example out and solving the TAP eqs. for the remaining $m-1$ examples. Fig. 1 shows excellent agreement between the two computations and we find $\epsilon_{\mathrm{loo}} = \epsilon_{\mathrm{loo}}^{\mathrm{exact}} = 33/104$. For comparison, the conventional TAP approach [8], which assumes a distribution of input data vectors with independent components, leads to a wrong result, $\epsilon_{\mathrm{loo}} = 41/104$ and $\epsilon_{\mathrm{loo}}^{\mathrm{exact}} = 33/104$.

In the second set of simulations Fig. 2 we demonstrate that the adaptive TAP method yields the correct statistical physics for the case of the linear Ising perceptron [22]. This has prior distribution $P(S) = \frac{1}{2}\delta(S-1) + \frac{1}{2}\delta(S+1)$ and likelihood $F(\hat{h}_j) \propto \exp(-(y_i - \hat{h}_j)^2/2\sigma^2)$, where we have chosen $\sigma^2 = 0.2$ and $N = 60$ in the simulations. See Ref. [21] for a discussion of this model in the context of demodulation in communications systems. To compare with the replica results [22], we have generated inputs at random and outputs using a noise free teacher perceptron sampled from the same prior. The small deviations between theory and TAP simulations close to the first order transition are attributed to hysteresis effects.

Acknowledgments. We thank T. Tanaka for providing us with his preprint [21]. This research is supported by the Swedish Foundation for Strategic Research as well as the Danish Research Councils through the Computational Neural Network Center (CONNECT) and the THOR Center for Neuroinformatics.

[1] M. Jordan, ed.,*Learning in Graphical Models*, MIT Press (1999).
[2] M. Opper and D. Saad, eds. *Advanced Mean Field Methods, Theory and Practice*, MIT Press (2001).
[3] M. Opper and O. Winther, Neural Computation **12**, 2655 (2000).
[4] M. Mézard and G. Parisi, Europhys. Lett. **2**, 913 (1986).
[5] M. Mézard, G. Parisi and M. A. Virasoro, *Spin Glass Theory and Beyond*, Lecture Notes in Physics, 9, World Scientific (1987).
[6] M. Mézard, J. Phys. A (Math. Gen. **22**, 2181 (1989).
[7] K. Y. M. Wong, Europhys. Lett. **30**, 245 (1995).
[8] M. Opper and O. Winther, Phys. Rev. Lett. **76**, 1964
[9] Y. Kabashima and D. Saad, Euro. Phys. Lett. **44**, 668 (1998).
[10] H. J. Kappen and F. B. Rodríguez, Neural Computation **10**, 1137 (1998).
[11] T. Tanaka, Phys. Rev. E **58**, 2302-2310 (1998).
[12] C. Bhattacharyya and S. S. Keerthi, J. Phys. A (Math. Gen.) **10**, 1307 (2000).
[13] D. J. Thouless, P. W. Anderson and R. G. Palmer, Phil. Mag. **35**, 593 (1977).
[14] D. Sherrington, and K. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[15] J. J. Hopfield, Proc. Nat. Acad. Sci. USA, **79** 2554 (1982).
[16] P.A.d.F.R. Højen-Sørensen, O. Winther, and L. K. Hansen, to appear in (NIPS'2000), MIT Press (2001).
[17] R. Neal, Artificial Intelligence **56**, 71-113 (1992).
[18] T. Plefka, J. Phys. A **15**, 1971 (1982).
[19] G. Parisi and M. Potters, J. Phys. A (Math. Gen.) **28**, 5267 (1995).
[20] R. P. Gorman and T. J. Sejnowski, Neural Networks **1**, 75 (1988).
[21] T. Tanaka, to appear in (NIPS'2000), MIT Press (2001).
[22] H. S. Seung, H. Sompolinsky and N. Tishby, Phys. Rev. A. **45**, 6056 (1992).