TAP Gibbs Free Energy, Belief Propagation and Sparsity

Lehel Csató and **Manfred Opper**
Neural Computing Research Group
Divison of Electronic Engineering and Computer Science
Aston University Birmingham B4 7ET, UK.
`[csatol,opperm]@aston.ac.uk`

Ole Winther
Center for Biological Sequence Analysis, BioCentrum
Technical University of Denmark, B208, 2800 Lyngby, Denmark.
`winther@cbs.dtu.dk`

Abstract

The adaptive TAP Gibbs free energy for a general densely connected probabilistic model with qaudratic interactions and arbitrary single site constraints is derived. We show how a specific sequential minimization of the free energy leads to a generalization of Minka's expectation propagation. Lastly, we derive a sparse representation version of the sequential algorithm. The usefulness of the approach is demonstrated on classification and density estimation with Gaussian processes and on an independent component analysis problem.

1   Introduction

There is an increasing interest in methods for approximate inference in probabilistic (graphical) models. Such approximations may usually be grouped in three classes. In the first case we approximate *self-consistency relations* for marginal probabilities by a set of nonlinear equations. Mean field (MF) approximations and their advanced extensions belong to this group. However, it is not clear, how to solve these equations efficiently. This latter problem is of central concern to the second class, the *Message passing algorithms*, like Bayesian online approaches (for references, see e.g. [1]) and belief propagation (BP) which dynamically update approximations to conditional probabilities. Finally, approximations based on *Free Energies* allow us to derive marginal moments by a minimising entropic loss measures. This method introduces new possibilities for algorithms and also gives approximations for the log-likelihood of observed data. The *variational method* is the most prominent member of this group.

One can gain important insight into an approximation, when it can be derived by different approaches. Recently, the fixed points of the BP algorithm were identified as the stable minima of the *Bethe* Free Energy, an insight which led to improved approximation schemes [2]. While BP is good and efficient on sparse tree-like structures, one may look for an approxi-

mation that works well in the opposite limit of densely connected graphs where individual dependencies are weak but their overall effect cannot be neglected. A interesting candidate is the adaptive TAP (ADATAP) approach introduced in [3] as a set of self-consistency relations. Recently, a message passing algorithm of Minka (termed *expectation propagation*) [1] was found to solve the ADATAP equations efficiently for models with Gaussian Process (GP) priors.

The goal of this paper is three-fold. We will add a further derivation of ADATAP using an approximate free energy. A sequential algorithm for minimising the free energy generalises Minka's result. Finally, we discuss how a sparse representation of ADATAP can be achieved for GP models, thereby extending previous sparse on-line approximations to the batch case [4].

We will specialize to probabilistic models on undirected graphs with nodes i that are of the type

$$P_\rho(\mathbf{S}) = \frac{\rho(\mathbf{S})}{Z} \exp \left[\sum_{i<j} S_i J_{ij} S_j \right] \tag{1}$$

The set of J_{ij}'s encodes the dependencies between the random variables $\mathbf{S} = (S_1, \ldots, S_N)$, whereas the factorising term $\rho(\mathbf{S}) = \prod_j \rho_j(S_j)$ (called *likelihood* in the following) usually encodes observed data at sites i and also incorporates all local constraints of the S_i (the range, discreteness, etc). Eq. (1) is a sufficiently rich and interesting class of models containing Boltzmann machines, models with Gaussian process priors [3], probabilistic independent component analysis [5] as well as Bayes belief networks and probabilistic neural networks (when the space of variables is augmented by auxiliary integration variables).

2   ADATAP approach from Gibbs Free Energy

We use the minimization of an approximation to a *Gibbs Free Energy* G in order to re-derive the ADATAP approximation.

The Gibbs Free Energy provides a method for computing marginal moments of P as well as of $-\ln Z$ within the same approach. It is defined by a constrained relative entropy minimization which is, for the present problem defined as

$$G_\rho(\mathbf{m}, \mathbf{M}) = \min_Q \left\{ KL(Q, P_\rho) \mid \langle \mathbf{S} \rangle_Q = \mathbf{m}, \langle \mathbf{S}^2 \rangle_Q = \mathbf{M} \right\} - \ln Z , \tag{2}$$

where the brackets denote expectations with respect to the distribution Q and $\langle \mathbf{S}^2 \rangle_Q$ is shorthand for a vector with elements $\langle S_i^2 \rangle_Q$. Since at the total minimum of G (with respect to its arguments) the minimizer in (2) is just $Q = P_\rho$, we conclude that $\min_{\mathbf{m},\mathbf{M}} G(\mathbf{m}, \mathbf{M}) = -\ln Z$ and the desired marginal moments of P are $(\langle \mathbf{S} \rangle, \langle \mathbf{S}^2 \rangle) = \operatorname{argmin}_{\mathbf{m},\mathbf{M}} G(\mathbf{m})$.

We will search for an approximation to G_ρ which is based on splitting $G_\rho = G_\rho^0 + \Delta G_\rho$, where G^0 is the Gibbs free energy for a factorising model that is obtained from (1) by setting all $J_{ij} = 0$. Previous attempts [6, 7] were based on a truncation of the power series expansion of ΔG_ρ with respect to the J_{ij} at second order. While this truncation leads to the correct TAP equations for the large N limit of so-called SK-model in statistical physics, its general significance is unclear. In fact, it will *not be exact* for a *simple* model with *Gaussian* likelihood. To make our approximation exact for such a case, we define (generalizing an idea of [8]) for an *arbitrary Gaussian* likelihood ρ_i^g $\Delta G^g(\mathbf{m}, \mathbf{M}) \doteq G_{\rho^g}(\mathbf{m}, \mathbf{M}) - G_{\rho^g}^0(\mathbf{m})$. The main reason for this definition is the fact that $\Delta G^g(\mathbf{m}, \mathbf{M})$ is *independent of* the actual Gaussian likelihood ρ_i^g chosen to compute G_ρ! This result depends crucially on

the moment constraints in (2). Changes in a Gaussian likelihood can always be absorbed within the Lagrange-multipliers for the constraints. We use this *universal* form ΔG^g to define The ADATAP approximation as $G_\rho^{TAP} = G_\rho^0 + \Delta G^g$, which by construction is exact for any Gaussian likelihood ρ. Introducing appropriate Lagrange multipliers γ and λ, we get

$$\Delta G^g(\mathbf{m}, \mathbf{M}) = \max_{\lambda, \gamma} \left\{ -\ln Z^g(\gamma, \lambda) + \mathbf{m}^T \gamma - \frac{1}{2}\mathbf{M}^T\lambda \right\} - \frac{1}{2}\sum_i \ln\left(M_i - m_i^2\right) \quad (3)$$

with $Z^g(\gamma, \lambda) = \int d\mathbf{S} \; \exp\left[\sum_i(\gamma_i S_i - \frac{1}{2}\lambda_i S_i^2) + \sum_{i<j} S_i J_{ij} S_j\right]$. Finally, setting $Z_i^0(\gamma_i^0, \lambda_i^0) = \int dS \rho_i(S) \; \exp[\gamma_i^0 S - \frac{1}{2}\lambda_i^0 S^2]$, we have

$$G_0 = \max_{\lambda^0, \gamma^0} \left\{ -\sum_i \ln Z_i^0(\gamma_i^0, \lambda_i^0) + \mathbf{m}^T \gamma^0 - \frac{1}{2}\mathbf{M}^T\lambda^0 \right\} . \quad (4)$$

3 Sequential Algorithm

The expression of G_ρ^{TAP} in terms of moments $(\mathbf{m}, \mathbf{M})$ and Lagrange parameters γ, λ and γ^0, λ^0 suggests that we may find local minima of G_ρ^{TAP} by iteratively alternating between updates of moments and Lagrange multipliers. Of special interest is the following sequential algorithm, which is a generalization of Minka's EP [1] for Gaussian process classification to an arbitrary model of the type eq. (1).

We choose a site i and define the updates by using the saddle points of G_ρ with respect to the moments and Lagrange multipliers in the following sequential order (where Λ is a diagonal matrix with elements λ_i):

$$\begin{array}{lll}
\partial_{\gamma_i, \lambda_i} G_\rho = 0 & \Rightarrow & m_i := \sum_j \left[(\Lambda - J)^{-1}\right]_{ij} \gamma_j \quad \& \quad M_i - m_i^2 := \left[(\Lambda - J)^{-1}\right]_{ii} \\
\partial_{m_i, M_i} G_\rho = 0 & \Rightarrow & \gamma_i^0 := -\gamma_i + \frac{m_i}{M_i - m_i^2} \quad \& \quad \lambda_i^0 := -\lambda_i - \frac{1}{M_i - m_i^2} \\
\partial_{\gamma_i^0, \lambda_i^0} G_\rho = 0 & \Rightarrow & m_i := \partial_{\gamma_i^0} \ln Z_i^0 \quad \& \quad M_i := -2\partial_{\lambda_i^0} \ln Z_i^0 \\
\partial_{m_i, M_i} G_\rho = 0 & \Rightarrow & \gamma_i := -\gamma_i^0 + \frac{m_i}{M_i - m_i^2} \quad \& \quad \lambda_i := -\lambda_i^0 - \frac{1}{M_i - m_i^2} .
\end{array}$$

The algorithm proceeds then by choosing a new site. The computation of $(\Lambda - J)^{-1}$ can be performed efficiently using the Sherman-Woodbury formula because *only one element* λ_i is changed in each update.

3.1 Cavity interpretation

At the fixed point, we may take $P_i(S) \equiv \frac{\rho_i(S)}{Z_i} \exp[\gamma_i^0 S - \frac{1}{2}\lambda_i^0 S^2]$ as the ADATAP approximation to the true marginal distribution of S_i [3]. The sequential approach may thus be considered as a belief propagation algorithm for ADATAP.

Although P_i is usually not Gaussian, we can also derive the moments $\mathbf{m}$ and $\mathbf{M}$ from the Gaussian distribution corresponding to Z^g. This auxiliary Gaussian model $P^g(\mathbf{S})$ has a likelihood $\rho_i^g(S) \propto \exp[-\frac{1}{2}\lambda_i S^2 + \gamma_i S]$ and provides us also with an additional approximation to the matrix of covariances via $\chi = (\Lambda - J)^{-1}$. This is useful when the coupling matrix J must be adapted to a set of observations by maximum likelihood II. We will give an example of this for independent component analysis below.

It is important to understand the role of γ^0 and λ^0 within the 'cavity' approach to the TAP equations. Defining $h_i = \sum_j J_{ij} S_j$, it is easy to show that $\gamma_i^0 = \langle h_i \rangle_{\backslash i}$ and $\lambda_i^0 = \langle h_i^2 \rangle_{\backslash i} - \langle h_i \rangle_{\backslash i}^2$ where the brackets denote an expectation with respect to the distribution of

all remaining variables $P^g(\mathbf{S}\backslash S_i) \propto \prod_{j,j\neq i} \rho_j^g(S_j) \exp[\sum_{k<l\neq i} J_{kl}S_kS_l]$ when node i is deleted from the graph. This statistics of h_i corresponds to the empty "cavity" at site i. The marginal distribution $P_i(S)$ as computed by ADATAP is equivalent to the approximation that the cavity distribution is Gaussian.

4 Examples

4.1 Models with Gaussian Process Priors

For this class of models, we assume that the graph is embedded in R^D, where the vector $\mathbf{S}$ is the restriction of a Gaussian process (random field) $\phi(\mathbf{x})$ with $\mathbf{x} \in R^D$, to a set of training inputs via $S_i = \phi(\mathbf{x}_i)$. $P(\mathbf{S})$ is the posterior distribution corresponding to a local likelihood model, when we set $\mathbf{J} = -\mathbf{K}^{-1}$ and the matrix $\mathbf{K}$ is obtained from a positive definite covariance kernel as $K_{ij} = K_0(\mathbf{x}_i, \mathbf{x}_j)$.

Our ADATAP approximation can be extended from the finite set of inputs to the entire space R^D by extending the auxiliary Gaussian distribution P^g with its likelihoods $\rho_i^g(S)$ to a Gaussian process with mean $\langle\phi(\mathbf{x})\rangle$ and covariance kernel $K_t(\mathbf{x}, \mathbf{x}')$ which approximates the posterior process. A calculation similar to [4] leads to the representation

$$\langle\phi(\mathbf{x})\rangle = \sum_j K_0(\mathbf{x}, \mathbf{x}_j)\gamma_j \tag{5}$$

$$K_t(\mathbf{x}, \mathbf{x}') = K_0(\mathbf{x}, \mathbf{x}') + \sum_{j,k} K_0(\mathbf{x}, \mathbf{x}_j)\chi_{jk}K_0(\mathbf{x}_k, \mathbf{x}') \tag{6}$$

Algorithms for the update of γ's and χ's will usually suffer from time consuming matrix multiplications when N is large. This common problem for GP models can be overcome by a *sparsity* approximation which extends previous on-line approaches [4] to the batch ADATAP approach. The idea is to replace the current version P^g of the approximate Gaussian with a further approximation $\hat{P}^g$ for which both the the corresponding $\hat{\gamma}_j$ as well as $\hat{\chi}_{jk}$ are nonzero only, when the nodes j and k belong to a smaller subset of nodes called "basis vectors" (BV) of size n [4]. For fixed BV set, the parameters of $\hat{P}^g$ are determined by minimizing the relative entropy $KL(\hat{P}^g, P^g)$. This yields $\hat{\gamma} = \boldsymbol{\pi}\gamma$ and $\hat{\mathbf{\Lambda}} = \boldsymbol{\pi}\mathbf{\Lambda}\boldsymbol{\pi}^T$ with the $n \times N$ projection matrix $\boldsymbol{\pi} = \mathbf{K}_{BV}^{-1}\mathbf{K}^+$. Here $\mathbf{K}$ is the kernel matrix between BVs and and $\mathbf{K}^+$ the kernel matrix between BVs and all nodes. The new distribution $\hat{P}^g$ can be written in the form (1) with a likelihood that contains only BVs

$$\hat{\rho}^g(\mathbf{S}^{BV}) = \exp[\sum_i \gamma_i(\boldsymbol{\pi}^T\mathbf{S}^{BV})_i - \frac{1}{2}\sum_i \lambda_i\{(\boldsymbol{\pi}^T\mathbf{S}^{BV})_i\}^2] \,. \tag{7}$$

Eq. (7) can be used to compute the sparse approximation within the sequential algorithm. We will only give a brief discussion here. In order to recompute the appropriate "cavity" parameters γ_i^0 and λ_i^0 when a new node is chosen by the algorithm, one removes a "pseudo-variable" $(\boldsymbol{\pi}^T\mathbf{S}_{BV})_i$ from the likelihood and recomputes the statistics of the remaining ones. When i is in the BV set, then simply $(\boldsymbol{\pi}^T\mathbf{S}^{BV})_i = Sb_i^{BV}$ and the computation reduces to the previous one. We will demonstrate the significance of this approach for two examples.

4.2 Independent Component Analysis

We consider a measured signal $\mathbf{X}_t$ which is assumed to be an instantaneous linear mixing of the sources corrupted with additive white Gaussian noise $\mathbf{\Gamma}$ that is,

$$\mathbf{X}_t = \mathbf{A}\mathbf{S}_t + \mathbf{\Gamma}_t \,, \tag{8}$$

where $\mathbf{A}$ is a (time independent) mixing matrix and the noise is assumed to be without temporal correlations nd with time independent covariance matrix $\mathbf{\Sigma}$. We thus have the following likelihood for parameters and sources at time t

$$P(\mathbf{X}_t|\mathbf{A}, \mathbf{\Sigma}, \mathbf{S}_t) = (\det 2\pi\mathbf{\Sigma})^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{X}_t - \mathbf{A}\mathbf{S}_t)^T \mathbf{\Sigma}^{-1}(\mathbf{X}_t - \mathbf{A}\mathbf{S}_t)} . \tag{9}$$

and for all times $P(\mathbf{X}|\mathbf{A}, \mathbf{\Sigma}, \mathbf{S}) = \prod_t P(\mathbf{X}_t|\mathbf{A}, \mathbf{\Sigma}, \mathbf{S}_t)$. The aim of independent component analysis is to recover the unknown quantities: the sources $\mathbf{S}$, the mixing matrix $\mathbf{A}$ and the noise covariance $\mathbf{\Sigma}$ from the observed data using the further assumption of statistical independence of the sources $P(\mathbf{S}_t) = \prod_i P(S_{it})$. We are thus back to the model eq. (1) with a we have to solve a set of mean field equations for each time step. Maximum Likelihood II (MLII) estimation of $\mathbf{A}$ and $\mathbf{\Sigma}$, i.e. maximizing the Likelihood $P(\mathbf{X}|\mathbf{A}, \mathbf{\Sigma}) = \int d\mathbf{S} P(\mathbf{X}|\mathbf{A}, \mathbf{\Sigma}, \mathbf{S}) P(\mathbf{S})$ leads to [5] $\mathbf{A}_{\mathrm{MLII}} = \sum_t \mathbf{X}_t \langle \mathbf{S}_t \rangle^T \left(\sum_{t'} \langle \mathbf{S}_{t'} \mathbf{S}_{t'}^T \rangle \right)^{-1}$ and $\mathbf{\Sigma}_{\mathrm{MLII}} = \frac{1}{N} \langle (\mathbf{X} - \mathbf{A}\mathbf{S})(\mathbf{X} - \mathbf{A}\mathbf{S})^T \rangle$. The sufficient statistics of the model is the first and second moments of the sources. As we shall see in the next section these can be effectively estimated using the sequential approach.

5 Simulations

5.1 Classification with GPs

This problem has been studied before [9, 4] using a sequential, sparse algorithm, based on a single sweep through the data only. Within the ADATAP approach we are able to perform multiple sweeps in order to achieve a self-consistent solution. For the likelihood we choose the probit model. The outputs are binary $y \in \{-1, 1\}$ and the probability is given by the error function (where $u = y\phi(x)/\sigma_0$): $P(y|\phi(x)) = \mathrm{Erf}(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} dt \exp\left[-\frac{t^2}{2}\right]$. The predictive distribution for an unseen x is $\mathrm{Erf}(y\langle\phi(x)\rangle_t/\sigma_x)$ with $\sigma_x^2 = \sigma_0^2 + K_t(x, x)$ and these are easily rewritten in terms of the parameters γ's and χ's according to eqs. (5). We used the USPS dataset[1] of gray-scale handwritten digit images of size 16×16 with 7291 training patterns and 2007 test patterns. For the kernel we choose the RBF kernel $K_0(x, x') = a_K \exp(\|x - x'\|^2/(m\sigma_K^2))$ where m is the dimension of the inputs – 256 in this case, and a_K and σ_K are the free parameters of the model. In the simulations we used 7000 random training examples. We performed simulations for different sizes of the BV set and also allowed multiple iterations throught the dataset, results displayed in Fig. 1. The lines show the average results of 5 runs where the task was to classify the images in fours/non-fours. The results show that, in contrast to the online learning, the fluscuations caused by the order of presentation are diminished (marked with bars on the figure).

5.2 Density estimation with GPs

Bayesian non-parametric models for density estimation can be defined [10] by parametrising densities p as $p(x|\phi) = \frac{\phi^2(x)}{\int \phi^2(x)\,dx}$ and putting a Gaussian process prior over the space of functions ϕ. Observing N data points $D = x_1, \ldots, x_N$, we can express the predictive distribution (again, E denotes the expectation over the GP prior) as

$$\begin{aligned}
p(x|D) &= \frac{1}{Z} E\left[\phi^2(x) \prod_{i=1}^{N} p(x_i|\phi)\right] = \frac{1}{ZN!} \int_0^\infty dl\, l^N\, E\left[\phi^2(x) \prod_{i=1}^{N} \phi^2(x_i)\, e^{-l \int \phi^2(x)dx}\right] \\
&\propto \int_0^\infty dl\, Z_l\, l^N\, E_l\left[\phi^2(x) \prod_{i=1}^{N} \phi^2(x_i)\right] .
\end{aligned}$$

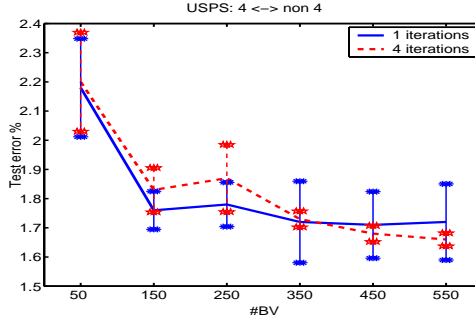[1] Available from `http://www.kernel-machines.org/data/`

Figure 1: Results for classification for different BV sizes (x-axis) and multiple sweeps through the data.

In the last expression, we have introduced an expectation over a new, effective Gaussian obtained by multiplying the old prior and the term $e^{-l \int \phi^2(x) dx}$ and normalizing by Z_l. We assume that for sufficiently large N the integral over l can be performed by Laplace's method, leaving us with an approximate predictor of the form $p(x|D) \propto \langle \phi^2(x) \rangle_l$, where the brackets denote posterior expectation for a GP model with a kernel that is a solution to $K^{(l)}(x,y) = K_0(x,y) - l \int dz\ K_0(x,z) K^{(l)}(z,y)$. The likelihood of the fields $S_i \doteq \phi(x_i)$ at the observation points is $\rho_i(S) = S^2$. For any fixed l, we can apply the sparse ADATAP algorithm to this problem. After convergence of this inner loop, a new value of l must be determined from (following a Laplace argument) $\frac{N}{\lambda} = \langle \phi^2(x) \rangle_l$ until global convergence is achieved. To give a simplified toy example, we choose the kernel $K_0(x,y)$ which reproduces itself after convolution. Hence, the l dependence is scaled out and we work with $l = 1$ and normalise at the end. We used a periodic kernel and assumed that the input is from $[0,1]$. For this case the kernel is

$$K_0(x,y) = -\cos(2\pi k_0(x-y)) + \sin(2\pi k_0(x-y)) \cot(\pi(x-y))$$

corresponding to a Fourier coefficients up to the cutoff frequency k_0 given a-priori (6 in simulations).

For the experiment we are using artificial data from a mixture of two Gaussians (dotted line in Fig. 2). We apply the sparse algorithm with multiple sweeps through the data. The sparse algorithm, an extension of the online GP learning [4], also considers the geometry "induced" by the kernel, limiting the number of basis vectors, avoiding numerical problems caused by a possible singular Gram matrix. For the experiments we did not have any upper limit for the size of the BV set, we used the geometry [4] when deciding about the inclusion.

Using 500 training data, only 10 are retained for prediction (continuous line in Fig. 2), nevertheless giving an accurate estimation of the pdf. This application shows clearly the benefit of using the sparse algorithm: with only a fraction of the data retained, we can recover the density underlying the data.

5.3 Independent Component Analysis

We consider the problem of local feature extraction with positive encoding which is also considered in [5] and compare the parallel and sequential algorithm. As a simplification, we take the naive mean field approximation that is setting $\boldsymbol{\lambda}^0 = 0$. This gives the parallel algorithm an advantage since this eliminates the need for inverting the matrix of covariances in the inner loop of the algorithm. However, we find the sequential algorithm needs only on average 7 sweeps through the sites to reach the desired accuracy whereas the parallel fails to reach the desired accuracy in 100 sweeps. The algorithms have almost the same time complexity in this case and the reach almost the solution despite the difference in accuracy.
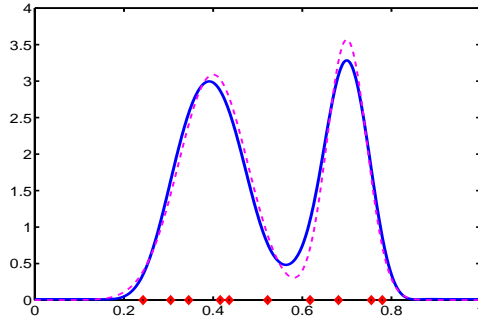
Figure 2: The GP estimation (continuous line) of a mixture of Gaussians (dotted line) using 10 BVs. The limitation is due to the geometry of the kernel.

6   Conclusion and Outlook

An obvious future direction for the ADATAP approach is the investigation of minimization algorithms as an alternative to the EP approach outlined before. Also an extension of the sparse approximation to other non-GP models will be interesting. A highly important but hard problem is the assessment of the accuracy of the approximation.

References

[1] T.P. Minka. Expectation propagation for approximate Bayesian inference. PhD thesis, Dep. of Electrical Eng. and Comp. Sci.; MIT, 2000.

[2] J. S. Yedidia, W. T. Freeman and Y. Weiss, Generalized Belief Propagation, to appear in Advances in Neural Information Processing Systems (NIPS'2000), MIT Press (2001).

[3] M. Opper and O. Winther, Tractable approximations for probabilistic models: The adaptive TAP approach, Phys. Rev. Lett. **86**, 3695 (2001).

[4] L. Csató and M. Opper. Sparse Gaussian Processes. Neural Computation accepted (2001).

[5] P.A.d.F.R. Højen-Sørensen, O. Winther, and L. K. Hansen, Mean Field Approaches to Independent Component Analysis, Neural Computation accepted (2001). Available from http://www.cbs.dtu.dk/winther/

[6] T. Plefka, Convergence condition of the TAP equations for the infinite-ranged Ising spin glass model, J. Phys. A **15**, 1971 (1982).

[7] T. Tanaka, Mean-Field Theory of Boltzmann Machine Learning, Phys. Rev. E **58**, 2302(1998).

[8] G. Parisi and M. Potters, Mean-Field Equations for Spin Models with Orthogonal Interaction Matrices, J. Phys. A (Math. Gen.) **28**, 5267 (1995).

[9] L. Csató, E. Fokoué, M. Opper, B. Schottky, and O. Winther. Efficient approaches to Gaussian process classification. In *Advances in Neural Information Processing Systems*, volume 12, (2000).

[10] D.M. Schmidt. Continuous probability distributions from finite data. *arXiv:physics/9808005* (1998)