Approximate Inference in Probabilistic Models

Manfred Opper[1] and Ole Winther[2]

[1] ISIS
School of Electronics and
Computer Science
University of Southampton
SO17 1BJ, United Kingdom
mo@ecs.soton.ac.uk
[2] Informatics and
Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark
owi@imm.dtu.dk

Abstract. We present a framework for approximate inference in probabilistic data models which is based on free energies. The free energy is constructed from two approximating distributions which encode different aspects of the intractable model. Consistency between distributions is required on a chosen set of moments. We find good performance using sets of moments which either specify factorized nodes or a spanning tree on the nodes.
The abstract should summarize the contents of the paper using at least 70 and at most 150 words. It will be set in 9-point font size and be inset 1.0 cm from the right and left margins. There will be two blank lines before and after the Abstract. ...

1 Introduction

Probabilistic data models explain the dependencies of complex observed data by a set of hidden variables and the joint probability distribution of all variables. The development of tractable approximations for the statistical inference with these models is essential for developing their full potential. Such approximations are necessary for models with a large number of variables, because the computation of the marginal distributions of hidden variables and the learning of model parameters requires high dimensional summations or integrations.

The most popular approximation is the *Variational Approximation* (VA) [2] which replaces the true probability distribution by an optimized simpler one, where multivariate Gaussians or distributions factorizing in certain groups of variables [1] are possible choices. The neglecting of correlations for factorizing distributions is of course a drawback. On the other hand, multivariate Gaussians allow for correlations but are restricted to *continuous random variables* which have the entire real space as their natural domain (otherwise, we get an infinite relative entropy which is used as a measure for comparing exact and approximate

S. Ben-David, J. Case, A. Maruoka (Eds.): ALT 2004, LNAI 3244, pp. 494–504, 2004.

densities in the VA). In this paper, we will discuss approximations which allow to circumvent these drawbacks. These will be derived from a Gibbs Free Energy (GFE), an entropic quantity which (originally developed in Statistical Physics) allows us to formulate the statistical inference problem as an optimization problem. While the true GFE is usually not exactly tractable, certain approximations can give quite accurate results. We will specialise on an *expectation consistent* (EC) approach which requires consistency between *two* complimentary approximations (say, a factorizing or tree with a Gaussian one) to the same probabilistic model.

The method is a generalization of the *adaptive TAP* approach (ADATAP) [16,15] developed for inference on densely connected graphical models which has been applied successfully to a variety of relevant problems. These include Gaussian process models [17,14,10,11], probabilistic independent component analysis [6], the CDMA coding model in telecommunications [4], bootstrap methods for kernel machines [7,8], a model for wind field retrieval from satellite observations [3] and a sparse kernel approach [12]. For a different, but related approximation scheme see [10,9].

2 Approximative Inference

Inference on the hidden variables $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ of a probabilistic model usually requires the computation of expectations, ie of certain sums or integrals involving a probability distribution with density

$$p(\mathbf{x}) = \frac{1}{Z} f(\mathbf{x}) \ . \tag{1}$$

This density represents the *posterior* distribution of $\mathbf{x}$ conditioned on the observed data, the latter appearing as parameters in p. $Z = \int d\boldsymbol{x} f(\mathbf{x})$ is the normalizing *partition function*.

Although some results can be stated in fairly general form, we will mostly specialize on densities (with respect to the Lebesgue measure in R^N) of the form

$$p(\mathbf{x}) = \prod_i \Psi_i(x_i) \ \exp\left(\sum_{i<j} x_i J_{ij} x_j\right) \ , \tag{2}$$

where the Ψ_i's are *non-Gaussian* functions. This also includes the important case of Ising variables $x_i = \pm 1$ by setting

$$\Psi(x_i) = (\delta(x_i + 1) + \ \delta(x_i - 1)) \ e^{\theta_i x_i} \ . \tag{3}$$

The type of density (2) appears as the posterior distribution for all models cited at the end of the introduction chapter.

$p(\mathbf{x})$ is a product of two functions $p(\mathbf{x}) = f_1(\mathbf{x}) f_2(\mathbf{x})$, where both the factorizing part $f_1 = \prod_i \Psi_i(x_i)$ and the Gaussian part $f_2 = \exp\left(\sum_{i<j} x_i J_{ij} x_j\right)$

individually are simple enough to allow for exact computations. Hence, as an approximation, we might want to keep f_1 but replace f_2 by a function which also factorizes in the components x_i. As an alternative, one may keep f_2 but replace f_1 by a Gaussian function to make the whole distribution *Gaussian*. Both choices are not ideal. The first completely neglects correlations of the variables but leads to marginal distributions of the x_i, which may share non Gaussian features (such as multimodality) with the true marginal. The second one neglects such features but incorporates nontrivial correlations. We will later develop an approach for combining these two approximations.

3 Gibbs Free Energies

Gibbs free energies (GFE) provide a convenient formalism for dealing with probabilistic approximations. In this framework, the *true, intractable* distribution $p(\mathbf{x})$ is *implicitly* characterized as the solution of an *optimization problem* defined through the the relative entropy (KL divergence)

$$KL(q,p) = \int d\mathbf{x}\; q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} \tag{4}$$

between p and other trial distributions q. We consider a *two stage optimization* process, where in the first step, the trial distributions q are constrained by fixing a set of values $\boldsymbol{\mu} = \langle \mathbf{g}(\mathbf{x}) \rangle_q$ for a set of generalized moments. The Gibbs Free Energy $G(\boldsymbol{\mu})$ is defined as

$$G(\boldsymbol{\mu}) = \min_q \{ KL(q,p) \mid \langle \mathbf{g}(\mathbf{x}) \rangle_q = \boldsymbol{\mu} \} - \ln Z \;, \tag{5}$$

where the term $\ln Z$ is subtracted to make the expression independent of the intractable partition function Z. In a second stage, both the true values $\boldsymbol{\mu} = \langle \mathbf{g}(\mathbf{x}) \rangle_p$ and the partition function Z are found by relaxing the constraints ie by minimizing $G(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$:

$$\min_{\boldsymbol{\mu}} G(\boldsymbol{\mu}) = -\ln Z \qquad \text{and} \qquad \langle \mathbf{g} \rangle = \operatorname*{argmin}_{\boldsymbol{\mu}} G(\boldsymbol{\mu}) \;. \tag{6}$$

A variational upper bound to G is obtained by restricting the minimization in (5) to a subset of densities q.

It can be easily shown that the optimizing distribution (5) is of the form

$$q(\mathbf{x}) = \frac{f(\mathbf{x})}{Z(\boldsymbol{\lambda})} \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}) \right) \;, \tag{7}$$

where the set of *Lagrange parameters* $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\mu})$ is chosen such that the conditions $\langle \mathbf{g}(\mathbf{x}) \rangle_q = \boldsymbol{\mu}$ are fulfilled, i.e. $\boldsymbol{\lambda}$ satisfies

$$\frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \boldsymbol{\mu} \;, \tag{8}$$

where $Z(\boldsymbol{\lambda})$ is a normalizing partition function.

Inserting the optimizing distribution eq. (7) into eq. (5) yields the *dual representation* of the Gibbs free energy

$$G(\boldsymbol{\mu}) = -\ln Z(\boldsymbol{\lambda}(\boldsymbol{\mu})) + \boldsymbol{\lambda}^T(\boldsymbol{\mu})\boldsymbol{\mu} = \max_{\boldsymbol{\lambda}} \left\{ -\ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T\boldsymbol{\mu} \right\} , \tag{9}$$

showing that G is the *Legendre transform* of $-\ln Z(\boldsymbol{\lambda})$ making G a convex function of its arguments.

We will later use the following simple result for the derivative of the GFE with respect to a parameter t contained in the probability density $p(\mathbf{x}|t) = \frac{f(\mathbf{x},t)}{Z_t}$. This can be calculated using (9) and (8) as

$$\frac{dG_t(\boldsymbol{\mu})}{dt} = -\frac{\partial \ln Z(\boldsymbol{\lambda}, t)}{\partial t} + \left(\boldsymbol{\mu} - \frac{\partial \ln Z(\boldsymbol{\lambda}, t)}{\partial \boldsymbol{\lambda}} \right) \frac{d\boldsymbol{\lambda}^T}{dt} = -\frac{\partial \ln Z(\boldsymbol{\lambda}, t)}{\partial t} . \tag{10}$$

Hence, we can keep $\boldsymbol{\lambda}$ fixed upon differentiation.

3.1 Simple Models

We give results for Gibbs free energies of three tractable models and choices of moments $\langle \mathbf{g}(\mathbf{x}) \rangle$. These will be used later as building blocks for the free energies of more complicated models.

Independent Ising variables. The Gibbs free energy for a set of independent Ising variables each with a density of the form (3) and fixed first moments $\boldsymbol{\mu} = \langle \mathbf{x} \rangle = \mathbf{m} = (m_1, m_2, \ldots, m_N)$ is $G(\mathbf{m}) = \sum_i G_i(m_i)$ where

$$G_i(m_i) = \frac{(1+m_i)}{2} \ln \frac{(1+m_i)}{2} + \frac{(1-m_i)}{2} \ln \frac{(1-m_i)}{2} - \theta_i m_i . \tag{11}$$

It will be useful to introduce a more complicated set of moments for this simple noninteracting model. We choose a tree graph $\mathcal{G}$ out of all possible sets of edges linking the variables $\mathbf{x}$ and fix the second moments $M_{ij} = \langle x_i x_j \rangle$ along theses edges as constraints. In this case, it can be shown that the free energy is represented in terms of single- and two-node free energies

$$G(\mathbf{m}, \{M_{ij}\}_{(ij)\in\mathcal{G}}) = \sum_{(ij)\in\mathcal{G}} G_{ij}(m_i, m_j, M_{ij}) + \sum_i (1-n_i) G_i(m_i) , \tag{12}$$

where $G_{ij}(m_i, m_j, M_{ij})$ is the two-node free energy computed for a single pair of variables, and n_i is the number of links to node i.

Multivariate Gaussians. The Gaussian model is of the form (2) with $\Psi_i(x_i) \propto \exp[a_i x_i - \frac{b_i}{2} x_i^2]$. Here, we fix $\boldsymbol{\mu} = (\mathbf{m}, \mathbf{M})$ where $\mathbf{m}$ is the set of all first moments and $\mathbf{M}$ is an arbitrary subset of second moments $\langle x_i x_j \rangle = M_{ij} = M_{ji}$. We get

$$\begin{aligned} G(\mathbf{m}, \mathbf{M}) = &-\frac{1}{2}\mathbf{m}^T\mathbf{J}\mathbf{m} - \mathbf{m}^T\mathbf{a} + \frac{1}{2}\sum_i M_{ii} b_i \\ &+ \max_{\Lambda} \left\{ \frac{1}{2} \ln \det(\Lambda - \mathbf{J}) - \frac{1}{2} \operatorname{Tr} \Lambda(\mathbf{M} - \mathbf{m}\mathbf{m}^T) \right\} , \end{aligned} \tag{13}$$

where Λ is a matrix of Lagrangemultipliers conjugate to the values of $\mathbf{M}$.

3.2 Complex Models: A Perturbative Representation

We will now concentrate on the more complex model (2) together with a suitably chosen set of moments. We will represent the Gibbs free energy of this model as the GFE for a tractable "noninteracting" part $f_1 = \prod_i \Psi_i(x_i)$ plus a correction for the "interaction" term $f_2 = \exp\left(\sum_{i<j} x_i J_{ij} x_j\right)$. We fix as constraints all the first moments $\mathbf{m}$ and a subset of second moments $\mathbf{M}$ which is chosen in such a way that the Gibbs free energy for f_1 remains still tractable. Different choices of second moments will allow later for more accurate approximations. If *all second moments* are fixed, our result will be exact, but for most models of the form (2) this leads again to intractable computations.

We define $f_2(\mathbf{x}, t)$ to be a smooth *interpolation* between the trivial case $f_2(\mathbf{x}, t = 0) = 1$ and the "full" intractable case $f_2(\mathbf{x}, t = 1) = f_2(\mathbf{x})$. For the model (2) we can set

$$f_2(\mathbf{x}, t) = \exp\left(t \sum_{i<j} x_i J_{ij} x_j\right) .$$

Differentiating the Gibbs free energy with respect to t, using eq. (10), we get

$$G(\boldsymbol{\mu}, 1) - G(\boldsymbol{\mu}, 0) = -\int_0^1 dt \left\langle \frac{d \ln f_2(\mathbf{x}, t)}{dt} \right\rangle_{q(\boldsymbol{x}|t)} . \tag{14}$$

where $q(\mathbf{x}|t) = \frac{1}{Z_q(\boldsymbol{\lambda}, t)} f_1(\mathbf{x}) f_2(\mathbf{x}, t) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right)$. For the model (2) this can be written as

$$G(\mu) \equiv G(\boldsymbol{\mu}, 1) = G(\boldsymbol{\mu}, 0) - \int_0^1 dt \sum_{i<j} J_{ij} \langle x_i x_j \rangle_{q(\boldsymbol{x}|t)} . \tag{15}$$

4 Approximations to the Free Energy

4.1 Mean Field Approximation

If we restrict ourselves to fixed diagonal second moments M_{ii} only, the simplest approximation is obtained by replacing the expectation over $q(\boldsymbol{x}|t)$ by the factorizing distribution $q(\boldsymbol{x}|0)$ giving

$$G(\boldsymbol{\mu}) \approx G(\boldsymbol{\mu}, 0) - \sum_{i<j} J_{ij} m_i m_j . \tag{16}$$

This result is equivalent to the *variational mean field* approximation, obtained by restricting the minimization in (5) to densities of the form $q(\mathbf{x}|0)$. Hence, it gives an *upper bound* to the true GFE.

4.2 Perturbative Expansion

One can improve on the mean field result by turning the exact expression (15) into a series expansion of the free energy in powers of t, setting $t = 1$ at the end. It is easy to see that the term linear in t corresponds to the mean field result. The second order term of this so-called Plefka expansion can be found in [18], see also several contributions in [13]. While the second order term seems to be sufficient for models with random independent couplings J_{ij} in a "thermodynamic limit", more advanced approximations are necessary in general [16,15].

4.3 A Lower Bound to the Gibbs Free Energy for Ising Variables

This was recently found by Wainwright & Jordan [20,19] and can be obtained by specifying *all second moments* $\mathbf{M}$. Then it is easy to see from the definition of the free energy that

$$G(\boldsymbol{\mu}) + \sum_{i<j} J_{ij} M_{ij} = -H[\mathbf{x}] \ ,$$

where $H[\mathbf{x}]$ equals the (discrete) negative entropy of the random variable $\mathbf{x}$. Wainwright and Jordan construct a *continuous random variable* $\tilde{\mathbf{x}}$ (a noisy version of $\mathbf{x}$) which has the same *differential entropy* $h[\tilde{\mathbf{x}}] = H[\mathbf{x}]$. Now they can apply a Maximum -Entropy argument and upper bound $h[\tilde{\mathbf{x}}]$ by the differential entropy $h_{\text{Gauss}}[\tilde{\mathbf{x}}]$ of a Gaussian with the same moments:

$$\begin{aligned} -H[\mathbf{x}] = -h[\tilde{\mathbf{x}}] \geq -h_{\text{Gauss}}[\tilde{\mathbf{x}}] &= \frac{1}{2}\log\det[\text{Cov}(\tilde{\mathbf{x}})] + \frac{N}{2}\log(\frac{ne}{2}) \\ &= \frac{1}{2}\log\det[\frac{1}{4}\text{Cov}(\mathbf{x}) + \frac{1}{3}I_N] + \frac{N}{2}\log(\frac{ne}{2}) \ . \end{aligned} \tag{17}$$

The approximate free energy comes out a convex function of its arguments.

4.4 Bethe–Kikuchi Type of Approximations

These are usually applied to discrete random variables and become exact if the graph which is defined by the edges of nonzero couplings $J_{ij} \neq 0$ is a tree or (for the Kikuchi approximation) a more generalized cluster of nodes. For tree connected graphs, the joint density of variables can always be expressed through single and two node marginals (similar to (12)). Using this structure within the optimization (5), one can calculate the Gibbs free energy exactly and efficiently when all first moments and the second moments along the edges of the graph are fixed. The approximation [21,22,5] is obtained when the graph of nonzero couplings is not a tree, but the simple form of the tree type distribution is still used in the optimization (5). Although the variation is over a subset of distributions, the Bethe–Kikuchi approximations do not lead to an upper bound to the free energy. This is because the constraints are no longer along trees and are thus not consistent with the distribution assumed.

5 Expectation Consistent Approximations

Our goal is to come up with another approximation which improves over the mean field result (16) by making a more clever approximation to $q(\mathbf{x}|t)$ in (15). We will use our assumption that we may approximate the density (2) by alternatively discarding the factor $f_1(\mathbf{x})$ as intractable, replacing the density $q(\mathbf{x}|t)$ by

$$r(\mathbf{x}|t) = \frac{1}{Z_r(\boldsymbol{\lambda}, t)} f_2(\mathbf{x}, t) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \ , \tag{18}$$

where the parameters $\boldsymbol{\lambda}$ are chosen to have *consistency for the expectations* of $\mathbf{g}$, i.e. $\langle \mathbf{g}(\mathbf{x}) \rangle_{r(\boldsymbol{x}|t)} = \boldsymbol{\mu}$.

$r(\mathbf{x}|t)$ defines another Gibbs free energy with a dual representation eq. (9)

$$G_r(\boldsymbol{\mu}, t) = \max_{\boldsymbol{\lambda}} \left\{ -\ln Z_r(\boldsymbol{\lambda}, t) + \boldsymbol{\lambda}^T \boldsymbol{\mu} \right\} \ . \tag{19}$$

We will use $r(\mathbf{x}|t)$ to treat the integral in eq. (14), writing

$$\int_0^1 dt \ \left\langle \frac{d \ln f_2(\mathbf{x}, t)}{dt} \right\rangle_{q(\boldsymbol{x}|t)} \approx \int_0^1 dt \ \left\langle \frac{d \ln f_2(\mathbf{x}, t)}{dt} \right\rangle_{r(\boldsymbol{x}|t)} \ . \tag{20}$$

Using the relations eqs. (10) for the free energy eq. (19) we get

$$\int_0^1 dt \ \left\langle \frac{d \ln f_2(\mathbf{x}, t)}{dt} \right\rangle_{r(\boldsymbol{x}|t)} = G_r(\boldsymbol{\mu}, 1) - G_r(\boldsymbol{\mu}, 0) \ . \tag{21}$$

and arrive at the *expectation consistent (EC)* approximation:

$$G(\boldsymbol{\mu}) \approx G(\boldsymbol{\mu}, 0) + G_r(\boldsymbol{\mu}, 1) - G_r(\boldsymbol{\mu}, 0) \equiv G^{\mathrm{EC}}(\boldsymbol{\mu}) \ . \tag{22}$$

6 Results for Ising Variables

We will now apply our EC framework to the model (2) with Ising variables $x_i = \pm 1$. We will discuss two types of approximations which differ by the set of fixed second moments M_{ij}. By fixing more and more second moments, we reduce the number of interaction terms of the form $J_{ij} x_i x_j$ which are not fixed and have to be approximated.

Since $r(\mathbf{x}|t)$ is a multivariate Gaussian, we have

$$\begin{aligned} G^{\mathrm{EC}}(\mathbf{m}, \mathbf{M}) = G(\mathbf{m}, \mathbf{M}, 0) - \frac{1}{2} \mathbf{m}^T \mathbf{J} \mathbf{m} \\ + \max_{\Lambda} \left\{ \frac{1}{2} \ln \det(\Lambda - \mathbf{J}) - \frac{1}{2} \operatorname{Tr} \Lambda (\mathbf{M} - \mathbf{m}\mathbf{m}^T) \right\} \\ - \max_{\Lambda} \left\{ \frac{1}{2} \ln \det \Lambda - \frac{1}{2} \operatorname{Tr} \Lambda (\mathbf{M} - \mathbf{m}\mathbf{m}^T) \right\} \ . \end{aligned} \tag{23}$$

To obtain estimates for the second moments which are not fixed in the free energy, we take derivatives of the free energy with respect to coupling parameters J_{ij} yielding

$$\langle \mathbf{x}\mathbf{x}^T \rangle - \langle \mathbf{x} \rangle \langle \mathbf{x}^T \rangle = (\Lambda - \mathbf{J})^{-1} \ . \tag{24}$$

This result is also consistent with the fixed values $\mathbf{M}$ for the second moments.

6.1 Diagonal Approximation

When we fix only the trivial diagonal second moments $M_{ii} \equiv \langle x_i^2 \rangle = 1$ (Ising constraints), $\mathbf{M}$ does not appear as a variable in the free energy. The EC approximation eq. (23) is then given by

$$\begin{aligned} G^{\mathrm{D}}(\mathbf{m}) = G^{\mathrm{Is}}(\mathbf{m}) - \frac{1}{2}\mathbf{m}^T \mathbf{J}\mathbf{m} \\ + \max_{\Lambda} \left\{ \frac{1}{2} \ln \det(\Lambda - \mathbf{J}) - \frac{1}{2} \sum_{i=1}^{N} \Lambda_i (1 - m_i^2) \right\} \\ + \frac{1}{2} \sum_{i=1}^{N} \ln(1 - m_i^2) + \frac{N}{2} \ , \end{aligned} \tag{25}$$

where $G^{\mathrm{Is}}(\mathbf{m})$ is given by eq. (11) and Λ is a diagonal matrix of Lagrange parameters. This result coincides with the older *adaptive TAP approximation* [16, 15].

6.2 Tree Approximation

A more complex, but still tractable approximation is obtained by selecting an arbitrary tree connected subgraph of pairs of nodes and fixing the second moments of the Ising variables along the edges of this graph. The free energy is again of the form eq. (23) but now with $G(\mathbf{m}, \mathbf{M}, 0)$ given by eq. (12), the Lagrange parameters Λ_{ij} are restricted to be non-zero on the tree graph only. If the tree is chosen in such a way as to include the most important couplings (defined in a proper way), one can expect that the approximation will improve significantly over the diagonal case.

7 Simulations

We compare results of the EC approximation with those of the Bethe–Kikuchi approaches on a toy problem suggested in [5]. We use $N = 10$ nodes, constant "external fields" $\theta_i = \theta = 0.1$. The J_{ij}'s are drawn independently at random, setting $J_{ij} = \beta w_{ij}/\sqrt{N}$, with Gaussian w_{ij}'s of zero mean and unit variance. We study eight different scaling factors $\beta = [0.10, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00, 10.00]$. The results are summarized in figures 1 and 2. Figure 1

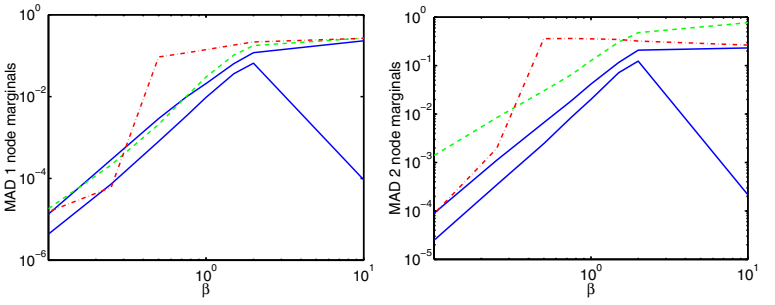

Fig. 1. Maximal absolute deviation (MAD) for one- (left) and two-variable (right) marginals. Blue upper full line: EC factorized, blue lower full line EC tree, green dashed line: Bethe and red dash-dotted line: Kikuchi.

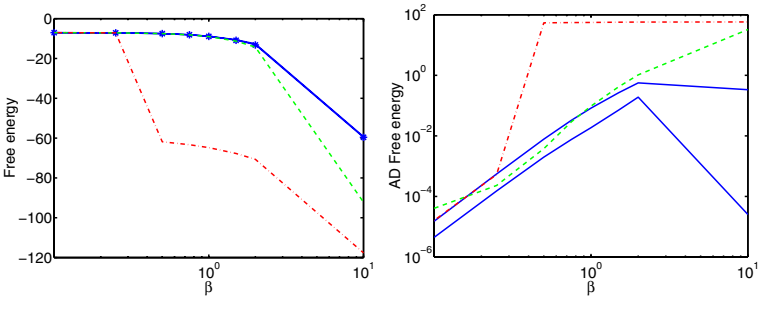

Fig. 2. Left plot: free energy for EC factorized and tree (blue full line), Bethe (green dashed line), Kikuchi (red dash-dotted) and exact (stars). Right: Absolute deviation (AD) for the three approximations, same line color and type as above. Lower full line is for the tree EC approximation.

gives the maximum absolute deviation (MAD) of our results from the exact marginals for different scaling parameters. We consider one-variable marginals $p(x_i) = \frac{1+x_i m_i}{2}$ and the two-variable marginals $p(x_i, x_j) = \frac{x_i x_j C_{ij}}{4} + p(x_i)p(x_j)$ with the approximate covariance $C_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$ given by eq. (24). Figure 2 gives estimates for the free energy. The results show that the simple diagonal EC approach gives performance similar to (and in many case better than) the more structured Bethe and Kikuchi approximations.

For the EC tree approximation, we construct a spanning tree of edges by choosing as the next edge, the (so far unlinked) pair of nodes with strongest absolute coupling $|J_{ij}|$ that will not cause a loop in the graph. The EC tree version is almost always better than the other approximations. A comparison with the Wainwright–Jordan approximation (corresponding to (17)) and details of the algorithm will be given elsewhere.

8 Outlook

We have introduced a scheme for approximate inference with probabilistic models. It is based on a free energy expansion around an exactly tractable substructure (like a tree) where the remaining interactions are treated in a Gaussian approximation thereby retaining nontrivial correlations. In the future, we plan to combine our method with a perturbative approach which may allow for a systematic improvement together with an estimate of the error involved. We will also work on an extension of our framework to more complex types of probabilistic models beyond the pairwise interaction case.

References

[1] H. Attias. A variational Bayesian framework for graphical models. In T. Leen et al., editor, *Advances in Neural Information Processing Systems 12*. MIT Press, Cambridge, 2000.

[2] C. M. Bishop, D. Spiegelhalter, and J. Winn. VIBES: A variational inference engine for Bayesian networks. In *Advances in Neural Information Processing Systems 15*, 2002.

[3] Dan Cornford, Lehel Csató, David J. Evans, and Manfred Opper. Bayesian analysis of the scatterometer wind retrieval inverse problem: Some new approaches. *Journal Royal Statistical Society B 66*, pages 1–17, 2004.

[4] T. Fabricius and O. Winther. Correcting the bias of subtractive interference cancellation in cdma: Advanced mean field theory. *Submitted to IEEE trans. Inf. Theory*, 2004.

[5] T. Heskes, K. Albers, and H. Kappen. Approximate inference and constrained optimization. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 313–320, San Francisco, CA, 2003. Morgan Kaufmann Publishers.

[6] P. A.d.F.R. Hojen-Sorensen, O. Winther, and L. K. Hansen. Mean field approaches to independent component analysis. *Neural Computation*, 14:889–918, 2002.

[7] Dórthe Malzahn and Manfred Opper. An approximate analytical approach to resampling averages. *Journal of Machine Learning Research 4*, pages 1151–1173, 2003.

[8] Dórthe Malzahn and Manfred Opper. Approximate analytical bootstrap averages for support vector classifiers. In *Proceedings of NIPS2003*, 2004.

[9] T. Minka and Y. Qi. Tree-structured approximations by expectation propagation. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[10] T. P. Minka. Expectation propagation for approximate bayesian inference. In *UAI 2001*, pages 362–369, 2001.

[11] T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT Media Lab, 2001.

[12] Qui nonero Candela and Ole Winther. Incremental gaussian processes. In *Advances in Neural Information Processing Systems 15*, pages 1001–1008, 2003.

[13] M. Opper and D. Saad. *Advanced Mean Field Methods: Theory and Practice*. MIT Press, 2001.

[14] M. Opper and O. Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12:2655–2684, 2000.

[15] M. Opper and O. Winther. Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling. *Physical Review E*, 64:056131, 2001.

[16] M. Opper and O. Winther. Tractable approximations for probabilistic models: The adaptive Thouless-Anderson-Palmer mean field approach. *Physical Review Letters*, 64:056131, 2001.

[17] Manfred Opper and Ole Winther. Mean field methods for classification with gaussian processes. In *Advances in Neural Information Processing Systems 11*, pages 309–315, 1999.

[18] T. Plefka. Convergence condition of the tap equation for the infinite-range ising spin glass. *J. Phys. A*, 15:1971, 1982.

[19] M. Wainwright and M. I. Jordan. Semidefinite relaxations for approximate inference on graphs with cycles. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[20] M. J. Wainwright and M. I. Jordan. Semidefinite methods for approximate inference on graphs with cycles. Technical Report UCB/CSD-03-1226, UC Berkeley CS Division, 2003.

[21] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation 13. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *NIPS*, pages 689–695, 2001.

[22] A. L. Yuille. Cccp algorithms to minimize the bethe and kikuchi free energies: convergent alternatives to belief propagation. *Neural Comput.*, 14(7):1691–1722, 2002.