Journal of Machine Learning Research 6 (2005) 2177–2204 Submitted 7/04; Revised 6/05; Published 12/05

Expectation Consistent Approximate Inference

Manfred Opper MO@ECS.SOTON.AC.UK
ISIS, School of Electronics and Computer Science
University of Southampton
SO17 1BJ, United Kingdom

Ole Winther OWI@IMM.DTU.DK
Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

Editor: Michael I. Jordan

Abstract

We propose a novel framework for approximations to intractable probabilistic models which is based on a free energy formulation. The approximation can be understood as replacing an average over the original intractable distribution with a tractable one. It requires two tractable probability distributions which are made consistent on a set of moments and encode different features of the original intractable distribution. In this way we are able to use Gaussian approximations for models with discrete or bounded variables which allow us to include non-trivial correlations. These are neglected in many other methods. We test the framework on toy benchmark problems for binary variables on fully connected graphs and 2D grids and compare with other methods, such as loopy belief propagation. Good performance is already achieved by using single nodes as tractable substructures. Significant improvements are obtained when a spanning tree is used instead.

1. Introduction

Recent developments in data acquisition and computational power have spurred an increased interest in flexible statistical Bayesian models in many areas of science and engineering. Inference in probabilistic models is in many cases intractable; the computational cost of marginalization operations can scale exponentially in the number of variables or require integrals over multivariate non-tractable distributions. In order to treat systems with a large number of variables, it is therefore necessary to use approximate polynomial complexity inference methods.

Probably the most prominent and widely developed approximation technique is the so-called *variational* (or *variational Bayes*) approximation (see, e.g. Jordan et al., 1999; Attias, 2000; Bishop et al., 2003). In this approach, the true intractable probability distribution is approximated by another one which is optimally chosen within a given, tractable family minimizing the *Kullback Leibler (KL) divergence* as the measure of dissimilarity between distributions. We will use the name *variational bound* for this specific method because the approximation results in an upper bound to the *free energy* (an entropic quantity related to the KL divergence). The alternative approximation methods discussed in this paper can also be derived from the variation of an approximate free energy which is not necessarily a bound. The most important tractable families of distributions in the variational bound approximation are multivariate Gaussians and distributions often in the exponential family which factorize in the marginals of all or for certain disjoint groups of variables (Attias, 2000) (this is often called a mean–field approximation). The use of multivariate Gaussians makes it possible to retain a significant amount of correlation between variables in the approximation. However, their application in the variational bound approximation is limited to distributions of continuous variables which have the entire real space as their natural domain. This is due to the fact that the KL divergence would diverge for distributions with non-matching support. Hence, in a majority of those applications, where random variables with constrained values (such as Boolean ones) appear, variational distributions of the mean field type have to be chosen. However, such factorizing approximations have the drawback that correlations are neglected and one often observes that fluctuations are underestimated (MacKay, 2003; Opper and Winther, 2004).

Recently, a lot of effort has been devoted to the development of approximation techniques which give an improved performance compared to the variational bound approximation. Thomas Minka's *Expectation Propagation* (EP) approach (Minka, 2001a,b) seems to provide a general framework from which many of these developments can be re-derived and understood. EP is based on a dynamical picture where factors—their product forming a global tractable approximate distribution—are iteratively optimized. In contrast to the variational bound approach, the optimization proceeds *locally* by minimizing KL divergences between appropriately defined *marginal* distributions. Since the resulting algorithm can be formulated in terms of the matching of marginal moments, this would not rule out factorizations where discrete distributions are approximated by multivariate Gaussians. However, such a choice seems to be highly unnatural from the derivation of the EP approximation (by the infinite KL measure) and has to our knowledge not been suggested so far (Minka, private communication). Hence, in practice, the correlations between discrete variables have been mainly treated using tree-based approximations. This includes the celebrated Bethe-Kikuchi approach (Yedidia et al., 2001; Yuille, 2002; Heskes et al., 2003), for EP interpretations see Minka (2001a,b) and Minka and Qi (2004). For a variety of related approximations within statistical physics see Suzuki (1995). However, while tree-type approximations often work well for sparsely connected graphs they become inadequate for inference problems in a dense graph regardless of the type of variables.

In this paper we present an alternative view of local-consistency approximations of the EP–type which we call *expectation consistent* (EC) approximations. It can be understood by requiring consistency between *two* complementary global approximations which may have different support (say, a Gaussian one and one that factorizes into marginals). Our method is a generalization of the *adaptive TAP* approach (ADATAP) (Opper and Winther, 2001a,b) developed for inference on densely connected graphical models. Although it has been applied successfully to a variety of problems ranging from probabilistic ICA (Hojen-Sorensen et al., 2002) over Gaussian process models (Opper and Winther, 2000) to bootstrap methods for kernel machines (Malzahn and Opper, 2003), see Appendix A, its potential as a fairly general scheme has been somewhat overlooked in the Machine Learning community.[1]

[1] This is probably due to the fact that the most detailed description of the method has so far only appeared in the statistical physics literature (Opper and Winther, 2001a,b) in a formulation that is not very accessible to a general audience. Shortly after the method first appeared–in the context of Gaussian

Although one algorithmic realization of our method can be given an EP-style interpretation (Csató et al., 2002), we believe that it is more natural and more powerful to base the derivation on a framework of optimizing a free energy approximation. This not only has the advantage of providing a simple and clear way for adapting the model parameters within the empirical Bayes framework, but also motivates different practical optimization algorithms among which the EP–style may not always be the best choice.

Our paper is organized as follows: Section 2 motivates approximate inference and explains the notation. The expectation consistent (EC) approximation to the free energy is derived in Section 3. Examples for EC free energies are given in Section 4. The algorithmic issues are treated in Section 5, simulations in Section 6 and finally we conclude in Section 7.

2. Motivation: Approximate Inference

We consider the problem of computing expectations, i.e. certain sums or integrals involving a probability distribution with density $p(\mathbf{x})$ for a vector of random variables $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. We assume that such computations are considered *intractable*, either because the necessary sums are over a too large number of variables or because multivariate integrals cannot be evaluated exactly. A further complication might occur when the density itself is expressed by a non-normalized multivariate function $f(\mathbf{x})$, say, equal to the product of a prior and a likelihood, which requires further normalization, i.e.

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

where the *partition function* Z must be obtained by the (intractable) summation or integration of f: $Z = \int d\mathbf{x} f(\mathbf{x})$. In a typical scenario, $f(\mathbf{x})$ is expressed as a product of two functions

$$f(\mathbf{x}) = f_q(\mathbf{x}) f_r(\mathbf{x}) \tag{2}$$

with $f_{q,r}(\mathbf{x}) \geq 0$, where f_q is "simple" enough to allow for tractable computations. The goal is to approximate the "complicated" part $f_r(\mathbf{x})$ by replacing it with a "simpler" function, say of some exponential form

$$\exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \equiv \exp\left(\sum_{j=1}^{K} \lambda_j g_j(\mathbf{x})\right) \ . \tag{3}$$

We have used the same vector notation for $\mathbf{g}$-vectors as for the random variables $\mathbf{x}$, however one should note that vectors will often have different dimensionalities, i.e. $K \neq N$. The vector of functions $\mathbf{g}$ is chosen in such a way that the desired sums or integrals can be calculated in an efficient way and the parameters $\boldsymbol{\lambda}$ are adjusted to optimize certain criteria. Hence, the word tractability should always be understood as relative to some approximating set of functions $\mathbf{g}$.

Our framework of approximation will be restricted to problems, where both parts f_q and f_r can be considered as tractable relative to some suitable $\mathbf{g}$, and the intractability

processes (Opper and Winther, 2000)–Minka introduced his EP framework and showed the equivalence of the fixed points of the two methods for Gaussian process models.

of the density p arises from forming their product.[2] In such a case, one may alternatively retain f_r but replace f_q by an approximation of the form eq. (3). One would then end up with two types of approximations

$$q(\mathbf{x}) = \frac{1}{Z_q(\boldsymbol{\lambda}_q)} f_q(\mathbf{x}) \, \exp\left(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right) \tag{4}$$

$$r(\mathbf{x}) = \frac{1}{Z_r(\boldsymbol{\lambda}_r)} f_r(\mathbf{x}) \, \exp\left(\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{x})\right) \tag{5}$$

for the same density, where $Z_q(\boldsymbol{\lambda}_q) = \int d\mathbf{x} \; f_q(\mathbf{x}) \; \exp\left(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right)$. We will not assume that either choice q and r is a reasonably good approximation for the global joint density $p(\mathbf{x})$ as we do in the variational bound approximation. In fact, later we will apply our method to the case of Ising variables, where the KL divergence between one of them and p is even infinite! Though, suitable different marginalizations of q and r can give quite accurate answers for important marginal statistics.

Take, as an example, the density $p(\mathbf{x}) = f(\mathbf{x})/Z = f_q(\mathbf{x}) f_r(\mathbf{x})/Z$—with respect to the Lebesgue measure in R^N—with

$$f_q(\mathbf{x}) = \prod_i \psi_i(x_i) \tag{6}$$

$$f_r(\mathbf{x}) = \exp\left(\sum_{i<j} x_i J_{ij} x_j + \sum_i \theta_i x_i\right) , \tag{7}$$

where, in order to have a nontrivial problem, ψ_i should be a non-Gaussian function. We will name this the *quadratic model*. Usually there will be an ambiguity in the choice of factorization, e.g. we could have included $\exp\left(\sum_i \theta_i x_i\right)$ as a part of $f_q(\mathbf{x})$. One may approximate $p(\mathbf{x})$ by a factorizing distribution, thereby replacing $f_r(\mathbf{x})$ by some function which factorizes in the components x_i. Alternatively, one can consider replacing $f_q(\mathbf{x})$ by a Gaussian function to make the whole distribution Gaussian. Both approximations are not ideal. The first completely neglects correlations of the variables but leads to marginal distributions of the x_i, which might qualitatively resemble the non-Gaussian shape of the true marginal. The second one neglects the non-Gaussian effects but incorporates correlations which might be used in order to approximate the two variable covariance functions. While within the variational bound approximation, both approximations appear independent from each other we will, in the following develop an approach for combining two complimentary approximations which "communicate" by matching the corresponding expectations of the functions $\mathbf{g}(\mathbf{x})$.

2.1 Notation

Throughout the paper, densities $p(\mathbf{x})$ are assumed relative to the Lebesgue measure $d\mathbf{x}$ in R^N. Other choices, such as the counting measure, may lead to alternative approximations for discrete variables. We will denote the expectation of a function $h(\mathbf{x})$ with respect to a

[2] This excludes many interesting models, for example mixture models, where tractability cannot be achieved with one split. These models can be treated by applying the approximation repeatedly. But for sake of clarity we will limit the treatment here to only one split.

density p by brackets

$$\langle h(\mathbf{x})\rangle = \int d\mathbf{x}\; p(\mathbf{x})\; h(\mathbf{x}) = \frac{1}{Z}\int d\mathbf{x}\; f(\mathbf{x})\; h(\mathbf{x})\ , \tag{8}$$

where, in cases of ambiguity, the density will appear as a subscript, like in $\langle h(\mathbf{x})\rangle_p$. One of the strengths of our formalism is to allow for a treatment of discrete and continuous random variables within the same approach.

Example: Ising variables Discrete random variables can be described using Dirac distributions in the densities. For examples, the density of N independent Ising variables $x_i \in \{-1, +1\}$ which occur with equal probabilities (one-half) has the density

$$p(\mathbf{x}) = \prod_{i=1}^{N}\left[\frac{1}{2}\delta(x_i+1) + \frac{1}{2}\delta(x_i-1)\right]\ . \tag{9}$$

3. Expectation Consistent Free Energy Approximation

In this section we will derive an approximation for $-\ln Z$, the negative log-partition function also called the (Helmholtz) free energy. We will use an approximating distribution $q(\mathbf{x})$ of the type eq. (4) and split the exact free energy into a corresponding part $-\ln Z_q$ plus a rest which will be further approximated. The split is obtained by writing

$$\begin{aligned} Z &= Z_q\frac{Z}{Z_q} = Z_q\frac{\int d\mathbf{x} f_r(\mathbf{x}) f_q(\mathbf{x}) \exp\left((\boldsymbol{\lambda}_q - \boldsymbol{\lambda}_q)^T\mathbf{g}(\mathbf{x})\right)}{\int d\mathbf{x} f_q(\mathbf{x}) \exp \boldsymbol{\lambda}_q^T\mathbf{g}(\mathbf{x})} \\ &= Z_q\left\langle f_r(\mathbf{x})\exp\left(-\boldsymbol{\lambda}_q^T\mathbf{g}(\mathbf{x})\right)\right\rangle_q\ , \end{aligned} \tag{10}$$

where

$$Z_q(\boldsymbol{\lambda}_q) = \int d\mathbf{x}\; f_q(\mathbf{x})\; \exp\left(\boldsymbol{\lambda}_q^T\mathbf{g}(\mathbf{x})\right)\ . \tag{11}$$

This expression can be used to derive a variational bound to the free energy $-\ln Z$. Applying Jensen's inequality $\ln\langle f(\mathbf{x})\rangle \geq \langle \ln f(\mathbf{x})\rangle$ we arrive at

$$-\ln Z \leq -\ln Z^{\text{var}} = -\ln Z_q - \langle \ln f_r(\mathbf{x})\rangle_q + \boldsymbol{\lambda}_q^T\langle \mathbf{g}(\mathbf{x})\rangle_q\ . \tag{12}$$

The optimal value for $\boldsymbol{\lambda}_q$ is found by minimizing this upper bound.

Our new approximation is obtained by arguing that *one may do better by retaining the* $f_r(\mathbf{x})\exp\left(-\boldsymbol{\lambda}_q^T\mathbf{g}(\mathbf{x})\right)$ *expression* in eq. (10) *but instead changing the distribution we use in averaging.* Hence, we replace the average with respect to $q(\mathbf{x})$ with an average using a distribution $s(\mathbf{x})$ containing the same exponential form

$$s(\mathbf{x}) = \frac{1}{Z_s(\boldsymbol{\lambda}_s)}\exp\left(\boldsymbol{\lambda}_s^T\mathbf{g}(\mathbf{x})\right)\ .$$

Given a sensible strategy for choosing the parameters $\boldsymbol{\lambda}_s$ and $\boldsymbol{\lambda}_q$, we expect that this approach in most cases gives a more precise approximation than the corresponding variational bound. Qualitatively, the more one can retain of the intractable function in the averaging the closer the result will to the exact partition function. It is difficult to make this statement quantitative and general. However, the method gives nontrivial results for a variety of cases where the variational bound would be simply infinite! This always happens, when f_q is Gaussian and f_r vanishes on a set which has nonzero probability with respect to the density f_q. Examples are when f_r is discrete or contains likelihoods which vanish in certain regions as in noise-free Gaussian process classifiers (Opper and Winther, 1999). Our approximation is further supported by the fact that for specific choices of f_r and f_q it is equivalent to the adaptive TAP (ADATAP) approximation (Opper and Winther, 2001a,b). ADATAP (unlike the variational bound) gives exact results for certain statistical ensembles of distributions in an asymptotic (thermodynamic) limit studied in statistical physics.

Using s instead of q, we arrive at the approximation for $-\ln Z$ which depends upon two sets of parameters $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_s$:

$$\begin{aligned} -\ln Z^{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s) &= -\ln Z_q - \ln \left\langle f_r(\mathbf{x}) \exp\left(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right)\right\rangle_s \\ &= -\ln \int d\mathbf{x} f_q(\mathbf{x}) \exp\left(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right) - \ln \int d\mathbf{x} f_r(\mathbf{x}) \exp\left((\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)^T \mathbf{g}(\mathbf{x})\right) \\ &\quad + \ln \int d\mathbf{x} \exp\left(\boldsymbol{\lambda}_s^T \mathbf{g}(\mathbf{x})\right) \ . \end{aligned} \tag{13}$$

Here we have utilized our additional assumption, that also f_r is tractable with respect to the exponential family and thus $Z_r = \int d\mathbf{x} f_r(\mathbf{x}) \exp\left((\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)^T \mathbf{g}(\mathbf{x})\right)$ can be computed in polynomial time. Eq. (13) leaves two sets of parameters $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_s$ to be determined. We expect that eq. (13) is a sensible approximation as long as $s(\mathbf{x})$ shares some key properties with q, for which we choose the *matching of the moments* $\langle \mathbf{g}(\mathbf{x}) \rangle_q = \langle \mathbf{g}(\mathbf{x}) \rangle_s$. This will fix $\boldsymbol{\lambda}_s$ as a function of $\boldsymbol{\lambda}_q$. Second, we know that the exact expression eq. (10) is independent of the value of $\boldsymbol{\lambda}_q$. If the replacement of $q(x)$ by $s(x)$ yields a good approximation, one would still expect that eq. (13) is a fairly flat function of $\boldsymbol{\lambda}_q$ (after eliminating $\boldsymbol{\lambda}_s$) in a certain region. Hence, it makes sense to require that an optimized approximation should make eq. (13) *stationary* with respect to variations of $\boldsymbol{\lambda}_q$. This does not imply that we are expecting a local minimum of eq. (13), see also section 3.1, but saddle points could occur. Since we are not after a bound on the free energy, this is not necessarily a disadvantage of the method. Readers who feel uneasy with this argument, might find the alternative, dual derivation (using the Gibbs free energy) in appendix B more appealing.

Both conditions can be summarized by the *expectation consistency* (EC) conditions

$$\frac{\partial \ln Z^{\mathrm{EC}}}{\partial \boldsymbol{\lambda}_q} = 0 \quad : \quad \langle \mathbf{g}(\mathbf{x}) \rangle_q = \langle \mathbf{g}(\mathbf{x}) \rangle_r \tag{14}$$

$$\frac{\partial \ln Z^{\mathrm{EC}}}{\partial \boldsymbol{\lambda}_s} = 0 \quad : \quad \langle \mathbf{g}(\mathbf{x}) \rangle_r = \langle \mathbf{g}(\mathbf{x}) \rangle_s \tag{15}$$

for the three approximating distributions

$$q(\mathbf{x}) = \frac{1}{Z_q(\boldsymbol{\lambda}_q)} f_q(\mathbf{x}) \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})) \tag{16}$$

$$r(\mathbf{x}) = \frac{1}{Z_r(\boldsymbol{\lambda}_r)} f_r(\mathbf{x}) \exp(\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{x})) \qquad \text{with} \qquad \boldsymbol{\lambda}_r = \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q \tag{17}$$

$$s(\mathbf{x}) = \frac{1}{Z_r(\boldsymbol{\lambda}_s)} \exp(\boldsymbol{\lambda}_s^T \mathbf{g}(\mathbf{x})) \ . \tag{18}$$

The corresponding EC approximation of the free energy is then

$$-\ln Z \approx -\ln Z^{\mathrm{EC}} = -\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q) + \ln Z_s(\boldsymbol{\lambda}_s) \qquad (19)$$

where $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_s$ are chosen such that the partial derivatives of the right hand side vanish.

3.1 Properties of the EC approximation

Invariances Although our derivation started with approximating one of the two factors f_q and f_r by an exponential, the final approximation is completely symmetric in the factors f_q and f_r. We could have chosen to define q in terms of f_r and still got the same final result. If f contains multiplicative terms which are of the form $\exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right)$ for some fixed $\boldsymbol{\lambda}$, we are free to include them either in f_q or f_r without changing the approximation. This can be easily shown by redefining $\boldsymbol{\lambda}_q \rightarrow \boldsymbol{\lambda}_q \pm \boldsymbol{\lambda}$.

Derivatives with respect to parameters. The following is a useful result about the derivative of $-\ln Z^{\mathrm{EC}}$ with respect to a parameter t in the density $p(\mathbf{x})$. Setting $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s)$, we get

$$\frac{d \ln Z^{\mathrm{EC}}(t)}{dt} = \frac{\partial \ln Z^{\mathrm{EC}}(\boldsymbol{\lambda}, t)}{\partial t} + \left(\frac{\partial \ln Z^{\mathrm{EC}}(\boldsymbol{\lambda}, t)}{\partial \boldsymbol{\lambda}}\right) \frac{d\boldsymbol{\lambda}^T}{dt} = \frac{\partial \ln Z^{\mathrm{EC}}(\boldsymbol{\lambda}, t)}{\partial t} , \qquad (20)$$

where the second equality holds at the stationary point. The important message is that we only need to take the explicit t dependence into account, i.e. we can keep the stationary values $\boldsymbol{\lambda}$ fixed upon differentiation. This property can also be useful when optimizing the free energy with respect to parameters in the empirical Bayes framework.

Relation to the variational bound. Applying Jensen's inequality to (13) yields

$$\begin{aligned} -\ln Z^{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s) &= -\ln Z_q - \ln \left\langle f_r(\mathbf{x}) \exp\left(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right)\right\rangle_s \\ &\geq -\ln Z_q - \langle \ln f_r(\mathbf{x})\rangle_s + \boldsymbol{\lambda}_q^T \langle \mathbf{g}(\mathbf{x})\rangle_s \ . \end{aligned}$$

Hence, if f_r and $\mathbf{g}(\mathbf{x})$ are defined in such a way that the matching of the moments $\langle \mathbf{g}(\mathbf{x})\rangle_s = \langle \mathbf{g}(\mathbf{x})\rangle_q$ implies $\langle \ln f_r(\mathbf{x})\rangle_q = \langle \ln f_r(\mathbf{x})\rangle_s$ then the rhs of the inequality is equal to the variational (bound) free energy eq. (12) for fixed $\boldsymbol{\lambda}_q$. This will be the case for the models discussed in this paper. Of course, this does not imply any relation between $-\ln Z^{\mathrm{EC}}$ and the true free energy. The similarity of EC to the variational bound approximation should also be interpreted with care. One could be tempted to try solving the EC stationarity conditions by eliminating $\boldsymbol{\lambda}_s$, i.e. enforcing the moment constraints between q and s, and minimizing the free energy approximation $-\ln Z^{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s(\boldsymbol{\lambda}_q))$ with respect to $\boldsymbol{\lambda}_q$, as in the variational bound method. Simple counter examples show however that this function maybe unbounded from below and that the stationary point may not even be a local minimum.

Non-convexity. The log–partition functions $\ln Z_{q,r,s}(\boldsymbol{\lambda})$ are the *cumulant generating functions* of the random variables $\mathbf{g}(\mathbf{x})$. Hence, they are differentiable and convex functions on their domains of definition, i.e.

$$\mathbf{H} = \frac{\partial^2 \ln Z}{\partial \boldsymbol{\lambda}^T \partial \boldsymbol{\lambda}} = \left\langle \mathbf{g}(\mathbf{x})\mathbf{g}(\mathbf{x})^T\right\rangle - \langle \mathbf{g}(\mathbf{x})\rangle \langle \mathbf{g}(\mathbf{x})\rangle^T$$

is positive semi-definite. It follows for fixed $\boldsymbol{\lambda}_s$ that eq. (19) is concave in the variable $\boldsymbol{\lambda}_q$, and there is only a single solution to eq. (14) corresponding to a maximum of $-\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)$. On the other hand, eq. (19) is a sum of a concave and a convex function of $\boldsymbol{\lambda}_s$. Thus, unfortunately there may be more than one stationary point, a property which the EC approach shares with other approximations such as the variational Bayes and the Bethe–Kikuchi methods. Nevertheless, we can use a double loop algorithm which alternates between solving the concave maximization problem for $\boldsymbol{\lambda}_q$ at fixed $\boldsymbol{\lambda}_s$ and updating $\boldsymbol{\lambda}_s$ given the values of the moments $\langle \mathbf{g}(\mathbf{x}) \rangle_r = \langle \mathbf{g}(\mathbf{x}) \rangle_q$ at fixed $\boldsymbol{\lambda}_q$. We will show in Section 5 and in Appendix B that such a simple heuristic leads to convergence to a stationary point assuming that a certain cost function is bounded from below.

4. EC Free Energies – Examples

In this section we derive the EC free energy for a specific model, the quadratic, and discuss several possible choices for the consistent statistics $\langle \mathbf{g}(\mathbf{x}) \rangle$.

4.1 Tractable Free Energies

Our approach applies most naturally to a class of models for which the distribution of random variables $\mathbf{x}$ can be written as a product of a factorizing part eq. (6) and "Gaussian part" eq. (7).[3] The choice of $\mathbf{g}(\mathbf{x})$ is then guided by the need to make the computation of the EC free energy, eq. (19), tractable. The "Gaussian part" stays tractable as long as we take $\langle \mathbf{g}(\mathbf{x}) \rangle$ to contain first and second moments of $\mathbf{x}$. It will usually be a good idea to take all first moments, but we have a freedom in choosing the amount of consistency and the number of second order moments in $\langle \mathbf{g}(\mathbf{x}) \rangle$. To keep Z_q tractable (assuming f_q it is not Gaussian), a restriction to diagonal moments, i.e. $\langle x_i^2 \rangle$ will be sufficient. When variables are discrete, it is also possible to include second moments $\langle x_i x_j \rangle$ for pairs of variables located at the edges $\mathcal{G}$ of a tree.

The following three choices represent approximations of increasing complexity:

- Diagonal restricted: consistency on $\langle x_i \rangle$, $i = 1, \ldots, N$ and $\sum_i \langle x_i^2 \rangle$.

$$\mathbf{g}(\mathbf{x}) = \left(x_1, \ldots, x_N, -\sum_i \frac{x_i^2}{2} \right) \qquad \text{and} \qquad \boldsymbol{\lambda} = (\gamma_1, \ldots, \gamma_N, \Lambda)$$

- Diagonal: consistency on $\langle x_i \rangle$ and $\langle x_i^2 \rangle$, $i = 1, \ldots, N$

$$\mathbf{g}(\mathbf{x}) = \left(x_1, -\frac{x_1^2}{2}, \ldots, x_N, -\frac{x_N^2}{2} \right) \qquad \text{and} \qquad \boldsymbol{\lambda} = (\gamma_1, \Lambda_1, \ldots, \gamma_N, \Lambda_N)$$

- Spanning tree: as above and additional consistency of correlations $\langle x_i x_j \rangle$ defined on a spanning tree $(ij) \in \mathcal{G}$. Since we are free to move the terms $J_{ij} x_i x_j$ with $(ij) \in \mathcal{G}$ from the Gaussian term f_r into the term f_q, without changing the approximation, we find that the number of interaction terms which have to be approximated using the

[3] A generalization where f_q factorizes into tractable "potentials" ψ_α defined on disjoint subsets $\mathbf{x}_\alpha$ of $\mathbf{x}$ is also straightforward.

complementary Gaussian density is reduced. If the tree is chosen in such a way as to include the most important couplings (defined in a proper fashion), one can expect that the approximation will be improved significantly.

It is of course also possible to go beyond a spanning tree to treat a larger part of the marginalization exactly. We will next give explicit expressions for some free energies which will be used later for the EC approximation.

Independent Ising random variables. Here, we consider N independent Ising variables $x_i \in \{-1,+1\}$:

$$f(\mathbf{x}) = \prod_{i=1}^{N} \psi_i(x_i) \qquad \text{with} \qquad \psi_i(x_i) = [\delta(x_i+1) + \ \delta(x_i-1)] \ . \tag{21}$$

For the case of diagonal moments we get $Z(\boldsymbol{\lambda}) = \prod_i Z_i(\boldsymbol{\lambda}_i)$, $\boldsymbol{\lambda}_i = (\gamma_i, \Lambda_i)$:

$$Z_i(\boldsymbol{\lambda}_i) = \int dx_i \ \psi_i(x_i) e^{\gamma_i x_i - \Lambda_i x_i^2/2} = 2\cosh(\gamma_i) e^{-\Lambda_i/2} \ . \tag{22}$$

Multivariate Gaussian. Consider a Gaussian model: $p(\mathbf{x}) = \frac{1}{Z} e^{\mathbf{x}^T \mathbf{J}\mathbf{x} + \boldsymbol{\theta}^T \mathbf{x}}$. We introduce an arbitrary set of first moments $\langle x_i \rangle$ and second moments $-\langle x_i x_j \rangle/2$ with conjugate variables $\boldsymbol{\gamma}$ and $\boldsymbol{\Lambda}$. Here it is understood, that entries of $\boldsymbol{\gamma}$ and $\boldsymbol{\Lambda}$ corresponding to the non-fixed moments are set equal to zero. $\boldsymbol{\Lambda}$ is chosen to be a symmetric matrix, $\Lambda_{ij} = \Lambda_{ji}$, for notational convenience. The resulting free energy is

$$\ln Z(\boldsymbol{\gamma}, \boldsymbol{\Lambda}) = \frac{N}{2} \ln 2\pi - \frac{1}{2} \ln \det(\boldsymbol{\Lambda} - \mathbf{J}) + \frac{1}{2} (\boldsymbol{\gamma} + \boldsymbol{\theta})^T (\boldsymbol{\Lambda} - \mathbf{J})^{-1} (\boldsymbol{\gamma} + \boldsymbol{\theta}) \ .$$

The free energies for binary and Gaussian tree graphs are given in Appendix C.

4.2 EC Approximation

We can now write down the explicit expression for the free energy, eq. (19) for the model eqs. (6) and (7) with diagonal moments using the result for the Gaussian model:

$$\begin{aligned} -\ln Z^{\mathrm{EC}} \quad = \quad & -\sum_i \ln \int dx_i \ \psi_i(x_i) e^{\gamma_{q,i} x_i - \Lambda_{q,i} x_i^2/2} + \frac{1}{2} \ln \det(\boldsymbol{\Lambda}_s - \boldsymbol{\Lambda}_q - \mathbf{J}) \qquad (23) \\ & -\frac{1}{2} (\boldsymbol{\theta} + \boldsymbol{\gamma}_s - \boldsymbol{\gamma}_q)^T (\boldsymbol{\Lambda}_s - \boldsymbol{\Lambda}_q - \mathbf{J})^{-1} (\boldsymbol{\theta} + \boldsymbol{\gamma}_s - \boldsymbol{\gamma}_q) - \frac{1}{2} \sum_i \left(\ln \Lambda_{s,i} - \frac{\gamma_{s,i}^2}{\Lambda_{s,i}} \right) \end{aligned}$$

where $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_s$ are chosen to make $-\ln Z^{\mathrm{EC}}$ stationary. The $\ln Z_s(\boldsymbol{\lambda}_s)$ term is obtained from the general Gaussian model setting $\boldsymbol{\theta} = \mathbf{0}$ and $\mathbf{J} = \mathbf{0}$.

Generating moments. Derivatives of the free energy with respect to parameters provide a simple way for generating expectations of functions of the random variable $\mathbf{x}$. We will explain the method for the second moments $\langle x_i x_j \rangle$ of the model defined by the factorization eqs. (6) and (7). If we consider $p(\mathbf{x})$ as a function of the parameter J_{ij}, we get after a short calculation

$$\frac{d \ln Z(\boldsymbol{\lambda}, J_{ij})}{d J_{ij}} = \frac{1}{2} \langle x_i x_j \rangle \ . \tag{24}$$

Here we assume that the coupling matrix $\mathbf{J}$ is augmented to a full matrix with the auxiliary elements set to zero at the end. Evaluating the left hand side of eq. (24) within the EC approximation eq. (23) and using eq. (20) yields

$$\langle \mathbf{x}\mathbf{x}^T \rangle - \langle \mathbf{x} \rangle \langle \mathbf{x} \rangle^T = \left(\mathbf{\Lambda}_s - \mathbf{\Lambda}_q - \mathbf{J} \right)^{-1} \ . \tag{25}$$

The result eq. (25) could have also obtained by computing the covariance matrix directly from the Gaussian approximating density $r(\mathbf{x})$. We have consistency between $r(\mathbf{x})$ and $q(\mathbf{x})$ on the second order moments included in $\mathbf{g}(\mathbf{x})$, but for those not included, one can argue on quite general grounds that $r(\mathbf{x})$ will be more precise than $q(\mathbf{x})$ (Opper and Winther, 2004). Similarly, one may hope that higher order diagonal moments or even the entire marginal density of variables can be well approximated using the density $q(\mathbf{x})$. An application which shows the quality of this approximation can be found in Malzahn and Opper (2003).

5. Algorithms

This section deals with the task of solving the EC optimization problem, that is solving the consistency conditions eqs. (14) and (15): $\langle \mathbf{g}(\mathbf{x}) \rangle_q = \langle \mathbf{g}(\mathbf{x}) \rangle_r = \langle \mathbf{g}(\mathbf{x}) \rangle_s$ for the three distributions q, r and s, eqs. (16)-(18). As already discussed in section 3, the EC free energy is not a concave function in the parameters $\boldsymbol{\lambda}_q$, $\boldsymbol{\lambda}_s$ and one may have to resort to double loop approaches (Welling and Teh, 2003; Yuille, 2002; Heskes et al., 2003; Yuille and Rangarajan, 2003). Heskes and Zoeter (2002) were the first to apply double loop algorithms EC type of approximations. Since the double loop approaches may be slow in practice it is also of interest to define single loop algorithms that come with no warranty, but in many practical cases will converge fast. A pragmatic strategy is thus to first try a single loop algorithm and switch to a double loop when necessary. In the following we first discuss the algorithms in general and then specialize to the model eqs. (6) and (7).

5.1 Single Loop Algorithms

The single loop approaches typically are of the form of propagation algorithms which send "messages" back and forth between the two distributions $q(\mathbf{x})$ and $r(\mathbf{x})$. In each step the "separator" or "overlap distribution" $s(\mathbf{x})$[4] is updated to be consistent with either q or r depending upon which way we are propagating. This corresponds to an Expectation Propagation style scheme with two terms, see also Appendix D. Iteration t of the algorithm can be sketched as follows:

1. Send message from r to q
   - Calculate separator $s(\mathbf{x})$: Solve for $\boldsymbol{\lambda}_s$: $\langle \mathbf{g}(\mathbf{x}) \rangle_s = \boldsymbol{\mu}_r(t-1) \equiv \langle \mathbf{g}(\mathbf{x}) \rangle_{r(t-1)}$
   - Update $q(\mathbf{x})$: $\boldsymbol{\lambda}_q(t) := \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_r(t-1)$
2. Send message from q to r
   - Calculate separator $s(\mathbf{x})$: Solve for $\boldsymbol{\lambda}_s$: $\langle \mathbf{g}(\mathbf{x}) \rangle_s = \boldsymbol{\mu}_q(t) \equiv \langle \mathbf{g}(\mathbf{x}) \rangle_{q(t)}$

[4] These names are chosen because $s(\mathbf{x})$ plays the same role as the separator potential in the junction tree algorithm and as the overlap distribution in the Bethe approximation.

- Update $r(\mathbf{x})$: $\boldsymbol{\lambda}_r(t) := \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q(t)$.

Here $r(t)$ and $q(t)$ denote the distributions q and r computed with the parameters $\boldsymbol{\lambda}_r(t)$ and $\boldsymbol{\lambda}_q(t)$. Convergence is reached when $\boldsymbol{\mu}_r = \boldsymbol{\mu}_q$ since each parameter update ensures $\boldsymbol{\lambda}_r = \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q$. Several modifications of the above algorithm are possible. First of all a "damping factor" (or "learning rate") η can be introduced on both or one of the parameter updates. Secondly we can abandon the parallel update and solve sequentially for factors containing only subsets of parameters.

5.2 Single Loop Algorithms for Quadratic Model

In the following we will explain details of the algorithm for the quadratic model eqs. (6) and (7) with consistency for first and second diagonal moments, corresponding to the EC free energy eq. (23). We will also briefly sketch the algorithm for moment consistency on a spanning tree. In appendix D we give the algorithmic recipes for a sequential algorithm for the factorized approximation and a parallel algorithm for tree approximation. These are simple, fast and quite reliable.

For the diagonal choice of $\mathbf{g}(\mathbf{x})$, $s(\mathbf{x})$ is simply the product of univariate Gaussians: $s(\mathbf{x}) = \prod_i s_i(x_i)$ and $s_i(x_i) \propto \exp\left(\gamma_{s,i} x_i - \Lambda_{s,i} x_i^2/2\right)$. Solving for $s(\mathbf{x})$ in terms of the moments of q and r, respectively, corresponds to a simple marginal moment matching to the univariate Gaussian $\propto \exp\left(-(x_i - m_i)^2/2v_i\right)$: $\gamma_{s,i} := m_i/v_i$ and $\Lambda_{s,i} := 1/v_i$. $r(\mathbf{x})$ is a multivariate Gaussian with covariance, eq. (25), $\boldsymbol{\chi}_r \equiv (\boldsymbol{\Lambda}_r - \mathbf{J})^{-1}$ and mean $\mathbf{m}_r = \boldsymbol{\chi}_r \boldsymbol{\gamma}_r$. Matching the moments with $r(\mathbf{x})$ gives $m_i := m_{r,i}$ and $v_i := \chi_{r,ii}$. The most expensive operation of the algorithm is the calculation of the moments of $r(\mathbf{x})$ which is $\mathcal{O}(N^3)$ because $\boldsymbol{\chi}_r = (\boldsymbol{\Lambda}_r - \mathbf{J})^{-1}$ has to be recalculated after each update of $\boldsymbol{\lambda}_r$. $q(\mathbf{x})$ is a factorized non-Gaussian distribution for which we have to obtain the mean and variance and match as above.

The spanning tree algorithm is only slightly more complicated. Now $s(\mathbf{x})$ is a Gaussian distribution on a spanning tree. Solving for $\boldsymbol{\lambda}_s$ can be performed in linear complexity in N using the tree decomposition of the free energy, see appendix C. $r(\mathbf{x})$ is still a full multivariate Gaussian and inferring the moments of the spanning tree distribution $q(\mathbf{x})$ is $\mathcal{O}(N)$ using message passing (MacKay, 2003).

5.3 Double Loop Algorithm

Since the EC free energy $-\ln Z^{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s)$ is concave in $\boldsymbol{\lambda}_q$, we can attempt a solution of the stationarity problem eqs. (14) and (15), by first solving the *concave maximization* problem

$$F(\boldsymbol{\lambda}_s) \equiv \max_{\boldsymbol{\lambda}_q} \left\{-\ln Z^{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s)\right\} = \max_{\boldsymbol{\lambda}_q} \left\{-\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)\right\} + \ln Z_s(\boldsymbol{\lambda}_s) \quad (26)$$

and subsequently finding a solution to the equation

$$\frac{\partial F(\boldsymbol{\lambda}_s)}{\partial \boldsymbol{\lambda}_s} = 0 \; . \quad (27)$$

Since $F(\boldsymbol{\lambda}_s)$ is in general neither a convex nor a concave function, there might be many solutions to this equation.

The double loop algorithm aims at finding a solution iteratively. It starts with an arbitrary admissible value $\boldsymbol{\lambda}_s(0)$ and iterates two elementary procedures for updating $\boldsymbol{\lambda}_s$ and $\boldsymbol{\lambda}_q$ aiming at matching the moments between the distribution q, r and s. Assume that at iteration step t $\boldsymbol{\lambda}_s = \boldsymbol{\lambda}_s(t)$, then iterate over the two steps

1. **Solve the concave maximization problem eq. (26)** yielding the update

$$\boldsymbol{\lambda}_q(t) = \underset{\boldsymbol{\lambda}_q}{\operatorname{argmax}} \left\{ -\ln Z^{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s(t)) \right\} \ . \tag{28}$$

With this update, we achieve equality of the moments

$$\boldsymbol{\mu}(t) \equiv \langle \mathbf{g}(\mathbf{x}) \rangle_{q(t)} = \langle \mathbf{g}(\mathbf{x}) \rangle_{r(t)} \ . \tag{29}$$

2. **Update $\boldsymbol{\lambda}_s$** as

$$\boldsymbol{\lambda}_s(t+1) = \underset{\boldsymbol{\lambda}_s}{\operatorname{argmin}} \left\{ -\boldsymbol{\lambda}_s^T \boldsymbol{\mu}(t) + \ln Z_s(\boldsymbol{\lambda}_s) \right\} \tag{30}$$

which is a convex minimization problem. This yields $\langle \mathbf{g}(\mathbf{x}) \rangle_{s(t+1)} = \boldsymbol{\mu}(t)$.

To discuss convergence of these iterations, we prove that $F(\boldsymbol{\lambda}_s(t))$ for $t = 0, 1, 2, \ldots$ is a nondecreasing sequence:

$$\begin{aligned}
F(\boldsymbol{\lambda}_s(t)) &= \max_{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r} \left\{ -\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_r) + \ln Z_s(\boldsymbol{\lambda}_s) + (\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r - \boldsymbol{\lambda}_s(t))^T \boldsymbol{\mu}(t) \right\} \qquad (31) \\
&\geq \max_{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r} \left\{ -\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_r) + (\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r)^T \boldsymbol{\mu}(t) + \min_{\boldsymbol{\lambda}_s} \left(-\boldsymbol{\lambda}_s^T \boldsymbol{\mu}(t) + \ln Z_s(\boldsymbol{\lambda}_s) \right) \right\} \\
&= \max_{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r} \left\{ -\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_r) + \ln Z_s(\boldsymbol{\lambda}_s(t+1)) + (\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r - \boldsymbol{\lambda}_s(t+1)) \boldsymbol{\mu}(t) \right\} \\
&\geq \max_{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r | \boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r = \boldsymbol{\lambda}_s(t+1)} \left\{ -\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_r) \right\} + \ln Z_s(\boldsymbol{\lambda}_s(t+1)) \\
&= F(\boldsymbol{\lambda}_s(t+1)) \ .
\end{aligned}$$

The first equality follows from the fact that $\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r - \boldsymbol{\lambda}_s(t) = 0$ and that at the maximum we have matching moments $\boldsymbol{\mu}(t)$ for the q and r distributions. The next inequality is true because we do not increase $-\boldsymbol{\lambda}_s^T \boldsymbol{\mu}(t) + \ln Z_s(\boldsymbol{\lambda}_s)$ by minimizing. The next equality implements the definition of eq. (30). The final inequality follows because we maximize over a restricted set. Hence, when F is bounded from below we will get convergence.

Hence, the double loop algorithm attempts in fact a minimization of $F(\boldsymbol{\lambda}_s)$. It is not clear a priori why we should search for a minimum rather than a maximum or any other critical value. However, a reformulation of the EC approach given in Appendix B shows that we can interpret $F(\boldsymbol{\lambda}_s)$ as an upper bound on an approximation to the so–called Gibbs free energy which is the Lagrange dual to the Helmholtz free energy from which the desired moments are derived by minimization.

5.4 Double Loop Algorithms for the Quadratic Model

The outer loop optimization problem (step 2 above) for $\boldsymbol{\lambda}_s$ is identical to the one for the single loop algorithm. The concave optimization problem of the inner loop for $\mathcal{L}(\boldsymbol{\lambda}_q) \equiv$

$-\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_s(t) - \boldsymbol{\lambda}_q)$ (step 1 above) can be solved by standard techniques from convex optimization (Vandenberghe et al., 1998; Boyd and Vandenberghe, 2004). Here we will describe a sequential approach that exploits the fact that updating only one element in $\boldsymbol{\Lambda}_r = \boldsymbol{\Lambda}_s(t) - \boldsymbol{\Lambda}_q$ (or in spanning tree case a two-by-two sub-matrix) is a rank one (or rank two) update of $\boldsymbol{\chi}_r = (\boldsymbol{\Lambda}_r - \mathbf{J})^{-1}$ that can be performed in $\mathcal{O}(N^2)$.

Specializing to the quadratic model with diagonal $\mathbf{g}(\mathbf{x})$ we have to maximize

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda}_q) &= -\sum_i \ln \int dx_i \psi_i(x_i) \exp\left[\gamma_{q,i} x_i - \frac{1}{2}\Lambda_{q,i} x_i^2\right] \\ &\quad - \ln \int d\mathbf{x}\ \exp\left[-\frac{1}{2}\mathbf{x}^T(\boldsymbol{\Lambda}_s(t) - \boldsymbol{\Lambda}_q - \mathbf{J})\mathbf{x} + (\boldsymbol{\gamma}_s(t) - \boldsymbol{\gamma}_q)^T\mathbf{x}\right]\end{aligned}$$

with respect to $\boldsymbol{\gamma}_q$ and $\boldsymbol{\Lambda}_q$. We aim at a sequential approach where we optimize the variables for one element in $\mathbf{x}$, say the ith. We can isolate $\gamma_{q,i}$ and $\Lambda_{q,i}$ in the Gaussian term to obtain a reduced optimization problem:

$$\begin{aligned}\mathcal{L}(\gamma_{q,i}, \Lambda_{q,i}) &= \text{const} + \frac{1}{2}\ln[1 - v_{r,i}(\Lambda^0_{q,i} - \Lambda_{q,i})] - \frac{(\gamma^0_{q,i} - \gamma_{q,i} - m_{r,i}/v_{r,i})^2}{2(1/v_{r,i} + \Lambda^0_{q,i} - \Lambda_{q,i})} \\ &\quad - \log \int dx_i \psi_i(x_i) \exp\left[\gamma_{q,i} x_i + \frac{1}{2}\Lambda_{q,i} x_i^2\right] , \qquad (32)\end{aligned}$$

where superscript 0 denotes current values of the parameters and we have set $m_{r,i} = \langle x_i \rangle_r = [(\boldsymbol{\Lambda}^0_r - \mathbf{J})^{-1}\boldsymbol{\gamma}^0_r]_i$ and $v_{r,i} = \langle x_i^2 \rangle_r - m_{r,i}^2 = [(\boldsymbol{\Lambda}^0_{r,i} - \mathbf{J})^{-1}]_{ii}$, with $\boldsymbol{\lambda}^0_r = \boldsymbol{\lambda}_s(t) - \boldsymbol{\lambda}^0_q$. Introducing the corresponding two first moments for $q_i(x_i)$

$$m_{q,i} = m_{q,i}(\gamma_{q,i}, \Lambda_{q,i}) = \langle x_i \rangle_q = \frac{1}{Z_{q_i}} \int dx_i\, x_i\, \psi_i(x_i) \exp\left[\gamma_{q,i} x_i - \frac{1}{2}\Lambda_{q,i} x_i^2\right] \qquad (33)$$

$$v_{q,i} = v_{q,i}(\gamma_{q,i}, \Lambda_{q,i}) = \langle x_i^2 \rangle_q - m_{q,i}^2 \qquad (34)$$

we can write the stationarity condition for $\gamma_{q,i}$ and $\Lambda_{q,i}$ as:

$$\gamma_{q,i} + \frac{m_{q,i}}{v_{q,i}} = \gamma^0_{q,i} + \frac{m_{r,i}}{v_{r,i}} \qquad (35)$$

$$\Lambda_{q,i} + \frac{1}{v_{q,i}} = \Lambda^0_{q,i} + \frac{1}{v_{r,i}} \qquad (36)$$

collecting variable terms and constant terms on the lhs and rhs, respectively. These two equations can be solved very fast with a Newton method. For binary variables the equations decouple since $m_{q,i} = \tanh(\gamma_{q,i})$ and $v_{q,i} = 1 - m_{q,i}^2$ and we are left with a one dimensional problem.

Typically, solving these two non-linear equations are not the most computationally expensive steps because after these have been solved, the first two moments of the r-distribution have to be recalculated. This final step can be performed using the matrix inversion lemma (or Sherman-Morrison formula) to reduce the computation to $\mathcal{O}(N^2)$. The matrix of second moments $\boldsymbol{\chi}_r = (\boldsymbol{\Lambda}_r - \mathbf{J})^{-1}$ is thus updated as:

$$\boldsymbol{\chi}_r := \boldsymbol{\chi}_r - \frac{\Delta\Lambda_{r,i}}{1 + \Delta\Lambda_{r,i}\,[\boldsymbol{\chi}_r]_{ii}}[\boldsymbol{\chi}_r]_i[\boldsymbol{\chi}_r]_i^T , \qquad (37)$$

where $\Delta\Lambda_{r,i} = -\Delta\Lambda_{q,i} = -(\Lambda_{q,i} - \Lambda^0_{q,i}) = \frac{1}{v_{q,i}} - \frac{1}{v_{r,i}}$ and $[\boldsymbol{\chi}_r]_i$ is defined to be the ith row in $\boldsymbol{\chi}_r$.

Note that the solution for $\Lambda_{q,i}$ is a coordinate ascent solution which has the nice property that if we initialize $\Lambda_{q,i}$ with an admissible value, i.e. with $\boldsymbol{\chi}_r$ positive semi-definite then with this update $\boldsymbol{\chi}_r$ will stay positive definite since the objective has an infinite barrier at $\det \chi_r = 0$.

6. Simulations

In this section we apply expectation consistent inference (EC) to the model of pair-wise connected Ising variables introduced in Section 4. We consider two versions of EC: "factorized" with $\mathbf{g}(\mathbf{x})$ containing all first and only diagonal second moments and the structured "spanning tree" version. The tree is chosen as a maximum spanning tree, where the maximum is defined over $|J_{ij}|$, i.e. choose as next pair of nodes to link, the (so far unlinked) pair with strongest absolute coupling $|J_{ij}|$ that will not cause a loop in the graph. The free energy is optimized with the parallel single loop algorithm described in section 5 and appendix D. Whenever non-convergence is encountered we switch to the double loop algorithm. We compare the performance of the two EC approximations with two other approaches for two different set-ups that have previously been used as benchmarks in the literature[5].

In the first set of simulations we compare with the Bethe and Kikuchi approaches (Heskes et al., 2003). We consider $N = 10$ and choose constant "external fields" (observations) $\theta_i = \theta = 0.1$. The "couplings" J_{ij} are fully connected and generated independently at random according to $J_{ij} = \beta w_{ij}/\sqrt{N}$, the w_{ij}s are Gaussian with zero mean and unit variance. We consider eight different scalings $\beta = [0.10, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00, 10.00]$. and compare one-variable marginals $p(x_i) = \frac{1+x_i m_i}{2}$ and the two-variable marginals $p(x_i, x_j) = \frac{x_i x_j C_{ij}}{4} + p(x_i)p(x_j)$ where C_{ij} is the covariance $C_{ij} = \langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle$. For EC, C_{ij} is given by eq. (25). In figure 1 we plot maximum absolute deviation (MAD) of our results from the exact marginals for different scaling parameters:

$$\begin{aligned} \text{MAD1} &= \max_i |p(x_i = 1) - p(x_i = 1|\text{Method})| \\ \text{MAD2} &= \max_{i,j} \max_{x_i=\pm 1, x_j=\pm 1} |p(x_i, x_j) - p(x_i, x_j|\text{Method})| \ . \end{aligned}$$

In figure 2 we compare estimates of the free energy. The results show that the simple factorized EC approach gives performance similar to (and in many case better than) the structured Bethe and Kikuchi approximations. The EC tree version is almost always better than the other approximations. The Kikuchi approximation is not uniquely defined, but depends upon the choice of "cluster-structure". Different types of structures can give rise to quite different performance (Minka and Qi, 2004). The results given above is thus just to be taken as one realization of the Kikuchi method where the clusters are taken as all variable triplets. We expect the Kikuchi approximation to yield better results (probably better than EC in some cases) for an appropriate choice of sub-graphs, for example triangles forming a star for fully connected models and all squares for grids (Yedidia et al., 2001; Minka and Qi, 2004). EC can also be improved beyond trees as discussed in the Conclusion.

[5] All results and programs are available from the authors.

The second test is the set-up proposed by Wainwright and Jordan (2003, 2005). The $N = 16$ nodes are either fully connected or connected to nearest neighbors in a 4-by-4 grid. The external field (observation) strengths θ_i are drawn from a *uniform* distribution $\theta_i \sim \mathcal{U}[-d_{\text{obs}}, d_{\text{obs}}]$ with $d_{\text{obs}} = 0.25$. Three types of coupling strength statistics are considered: repulsive (anti-ferromagnetic) $J_{ij} \sim \mathcal{U}[-2d_{\text{coup}}, 0]$, mixed $J_{ij} \sim \mathcal{U}[-d_{\text{coup}}, +d_{\text{coup}}]$ and attractive (ferromagnetic) $J_{ij} \sim \mathcal{U}[0, +2d_{\text{coup}}]$ with $d_{\text{coup}} > 0$. We compute the average absolute deviation on the marginals:

$$\text{AAD} = \frac{1}{N} \sum_i |p(x_i = 1) - p(x_i = 1|\text{method})|$$

over 100 trials testing the following methods: SP = sum-product (aka loopy belief propagation (BP) or Bethe approximation) and LD = log-determinant maximization (Wainwright and Jordan, 2003, 2005), EC factorized and EC tree. Results for SP and LD are taken from Wainwright and Jordan (2003). Note that instances where SP failed to converge were excluded from the results. A fact that is likely to bias the results in favor of SP. The results are summarized in table 6. The Bethe approximation always gives inferior results compared to EC. This might be a bit surprising for the sparsely connected grids. LD is a robust method which however seems to be limited in it's achievable precision. EC tree is uniformly superior to all other approaches. It would be interesting to compare to the Kikuchi approximation which is known to give good results on grids.

A few comments about complexity, speed and rates of convergence: Both EC algorithms are $\mathcal{O}(N^3)$. For the $N = 16$ simulations typical wall clock times were 0.5 sec. for exact computation, half of that for the single-loop tree and one-tenth for the factorized single-loop. Convergence is defined to be when $||\langle \mathbf{g}(\mathbf{x}) \rangle_q - \langle \mathbf{g}(\mathbf{x}) \rangle_r||^2$ is below 10^{-12}. Double loop algorithms typically were somewhat slower (1-2 sec.) because a lot of outer loop iterations were required. This indicates that the bound optimized in the inner loop is very conservative for these binary problems. For the easy problems (small d_{coup}) all approaches converged. For the harder problems the factorized EP-style algorithms typically converged in 80-90 % of the cases. A greedy single-loop variant of the sequential double-loop algorithm, where the outer loop update is performed after every inner loop update, converged more often without being much slower than the EP-style algorithm. We treated the grid as a fully connected system yielding a complexity of $\mathcal{O}(N^3)$. Exploiting the structure using message passing, one can reduce the complexity of inference, i.e. calculating the covariance on the links, to $\mathcal{O}(N^2)$.

7. Conclusion and Outlook

We have introduced a novel method for approximate inference which tries to overcome limitations of previous approximations in dealing with the correlations of random variables. While we have demonstrated its accuracy in this paper only for a model with binary elements, it can also be applied to models with continuous random variables or hybrid models with both discrete and continuous variables (i.e. cases where further approximations are needed in order to apply Bethe/Kikuchi approaches).

We expect that our method becomes most powerful when certain tractable substructures of variables with strong dependencies can be identified in a model. Our approach would then

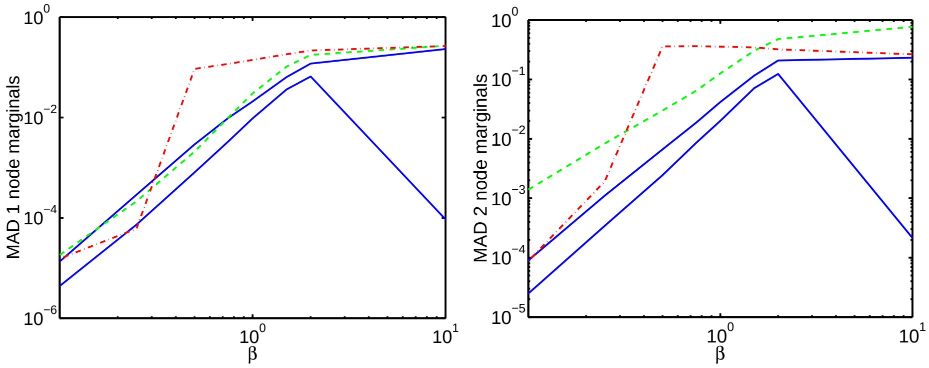

Figure 1: Maximal absolute deviation (MAD) for one- (left) and two-variable (right) marginals. EC factorized: upper full line (blue), EC tree: lower full line (blue), Bethe: dashed line (green) and Kikuchi: dash-dotted line (red).

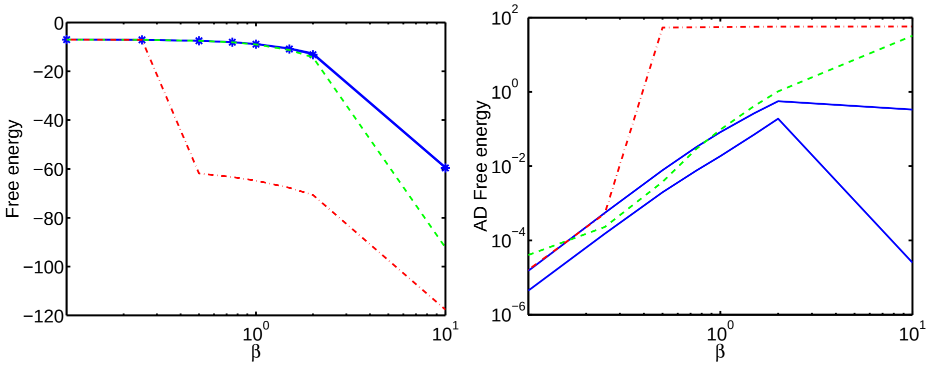

Figure 2: Left plot: free energy exact: stars, EC factorized and tree: full lines virtually on top on each others (blue), Bethe: dashed line (green) and Kikuchi: dash-dotted (red). Right: Absolute deviation (AD) for the three approximations, same line type (and color) as above. Lower full line is for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

allow us to deal well with the weaker dependencies between substructures. Better heuristics for determining the choice of substructures will also be useful for improving the performance (Minka and Qi, 2004). Consider inference on the square grid as a problem where one can introduce tractable substructures without getting a very large increase in complexity. The spanning tree treats approximately half of the links exactly, whereas covering the grid with strips of width L would treat a fraction of $1 - 1/2L$ of the links exactly at a computational increase of a factor of 2^{L-1} compared to the spanning tree for the binary part, but keeping

Problem type			Method							
			SP	LD	EC factorized			EC tree		
Graph	Coupling	d_{coup}	Mean	Mean	Mean±std	Med	Max	Mean±std	Med	Max
Full	Repulsive	0.25	0.037	0.020	0.003 ± 0.002	0.002	0.00	0.0017 ± 0.0011	0.001	0.01
	Repulsive	0.50	0.071	0.018	0.031 ± 0.045	0.016	0.20	0.0143 ± 0.0141	0.010	0.10
	Mixed	0.25	0.004	0.020	0.002 ± 0.002	0.002	0.00	0.0013 ± 0.0008	0.001	0.00
	Mixed	0.50	0.055	0.021	0.022 ± 0.030	0.013	0.17	0.0151 ± 0.0204	0.010	0.16
	Attractive	0.06	0.024	0.027	0.004 ± 0.002	0.004	0.01	0.0025 ± 0.0014	0.002	0.01
	Attractive	0.12	0.435	0.033	0.117 ± 0.090	0.112	0.30	0.0211 ± 0.0307	0.012	0.16
Grid	Repulsive	1.0	0.294	0.047	0.153 ± 0.123	0.124	0.58	0.0031 ± 0.0021	0.003	0.01
	Repulsive	2.0	0.342	0.041	0.198 ± 0.135	0.214	0.49	0.0021 ± 0.0010	0.002	0.01
	Mixed	1.0	0.014	0.016	0.011 ± 0.010	0.009	0.08	0.0018 ± 0.0011	0.002	0.01
	Mixed	2.0	0.095	0.038	0.082 ± 0.081	0.034	0.32	0.0068 ± 0.0053	0.005	0.03
	Attractive	1.0	0.440	0.047	0.125 ± 0.104	0.068	0.36	0.0028 ± 0.0018	0.002	0.01
	Attractive	2.0	0.520	0.042	0.177 ± 0.125	0.198	0.41	0.0002 ± 0.0004	0.000	0.00

Table 1: The average one-norm error on marginals for the Wainwright-Jordan set-up.

the complexity of the most computationally expensive part of the inference—calculating the moments of the Gaussian part—unchanged.

A generalization of our method to treat graphical models beyond pair-wise interaction may be obtained by iterating the approximation. This is useful in cases, where an initial three term approximation $-\ln Z^{EC} = -\ln Z_q - \ln Z_r + \ln Z_s$ still contains non-tractable component free energies. These individual terms can be further approximated using the EC approach. We can show that in such a way a variety of other relevant types of graphical models beyond the pair-wise interaction case (on certain directed graphs and mixture models) become tractable with our method.

For practical applicability of approximate inference techniques improvements in the numerical implementation of the free energy minimization are crucial. In the simulations in this paper we used both single and double loop algorithms. The single loop algorithms often converged very fast, i.e. in $\mathcal{O}(10)$ iterations to achieve a solution close to the machine precision. However, whether convergence could be achieved was instance dependent and depended upon set-up details like parallel/sequential update and damping factor. It seems that there is a lot of room for improvement here and theoretical analysis of convergence properties of algorithms will be important in this respect (Heskes and Zoeter, 2002). In the guaranteed convergent double loop approaches the free energy minimization is formulated in terms of a sequence of convex optimization problems. This allows for the application of theoretically well-founded and powerful techniques of convex optimization (Boyd and Vandenberghe, 2004). Unfortunately, for the problems considered here, convergence is typically quite slow because we have to solve large number of the convex problems. This again underlines the need for further algorithmic development.

There are a couple of ways to improve on the EC approximation itself. One may calculate corrections to the EC free energy and marginals by a perturbative analysis using cumulant expansions of the approximating distributions. This should also enable a kind of sanity check of the theory, i.e. when the corrections are predicted to be comparable to original prediction, it is a signal that the approximation is breaking down. Another possible improvement could come from physics of disordered system where methods have be devised to analyze non-ergodic free energy landscapes (Mézard et al., 1987). This will allow to make improved estimates of the free energy and marginals for example binary variables with large coupling strengths.

Acknowledgments

Discussions with and suggestions by Kees Albers, Bert Kappen, Tom Minka, Wim Wiegerinck, Onno Zoeter and anonymous referees are greatly appreciated. Special thanks to Wim for his contributions to clarifying the single loop algorithm concepts.

Appendix A. Applications

In this appendix we give list of of previous applications of the ADATAP method which is a special case of the EC approach to models with the factorization eqs. (6) and (7).

Application	meaning of x_i	type of x_i	Refs.
Channel Division Multiple Access (CDMA)	source symbol	Ising	a
Gaussian Processes (GP) classification	latent variable	continuous	b
GP for wind retrieval	wind vector	continuous	c
Bootstrap estimates	latent variable	continuous	d
Independent component analysis (ICA)	source variable	arbitrary	e
Sparse kernel method	latent variable	continuous	f

Table 2: Examples of applications of simplest version of EC, ADATAP. The references are a: Fabricius and Winther (2004), b: Opper and Winther (1999, 2000); Minka (2001a,b), c: Cornford et al. (2004), d: Malzahn and Opper (2003, 2004), e: Hojen-Sorensen et al. (2002) and f: Quiñonero-Candela and Winther (2003).

Appendix B. Dual Formulation

In this appendix we present an alternative route to EC free energy approximation using a two stage variational formulation. The result is the so-called Gibbs free energy which is the Lagragian dual of the Helmholtz free energy eq. (19).

B.1 Gibbs Free Energies and Two Stage Inference

In this framework, one starts with the well known fact that the true, intractable distribution $p(\mathbf{x}) = \frac{f(\mathbf{x})}{Z}$ is implicitly characterized as the solution of an optimization problem defined through the relative entropy or KL divergence

$$KL(q, p) = \int d\mathbf{x}\, q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} \tag{38}$$

between p and other trial or approximate distributions q. We introduce the Gibbs free energy (GFE) approach, (see, e.g. Roepstorff, 1994; Csató et al., 2002; Wainwright and Jordan, 2003, 2005) which splits this optimization into a two stage process. One first constrains the trial distributions q by fixing the values of the generalized moments $\langle \mathbf{g}(\mathbf{x})\rangle_q$. We define the Gibbs free energy $G(\boldsymbol{\mu})$ as

$$G(\boldsymbol{\mu}) = \min_q \left\{ KL(q,p) \mid \langle \mathbf{g}(\mathbf{x})\rangle_q = \boldsymbol{\mu} \right\} - \ln Z \ . \tag{39}$$

The term $\ln Z$ has been subtracted to make the resulting expression independent of the intractable partition function Z.

In a second step, the moments of the distribution and also the partition function Z are found within the same approach by relaxing the constraints and further minimizing $G(\boldsymbol{\mu})$ with respect to the $\boldsymbol{\mu}$.

$$\begin{aligned} \min_{\boldsymbol{\mu}} G(\boldsymbol{\mu}) &= -\ln Z & (40) \\ \langle \mathbf{g}(\mathbf{x})\rangle &= \operatorname*{argmin}_{\boldsymbol{\mu}} G(\boldsymbol{\mu}) \ . & (41) \end{aligned}$$

A variational bound approximation is recovered by restricting the minimization in eq. (39) to a tractable family of densities q. Note that the values for $\boldsymbol{\mu}$ in the definition of $G(\boldsymbol{\mu})$ cannot be chosen arbitrarily. For a detailed discussion of this problem, see Wainwright and Jordan (2003, 2005). We will not discuss these constraints here, but leave this, when necessary, to the discussion of concrete models.

Gibbs free energy and duality. The optimization problem eq. (39) is solved by the density given by

$$q(\mathbf{x}) = \frac{f(\mathbf{x})}{Z(\boldsymbol{\lambda})} \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \ . \tag{42}$$

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

$$\frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \boldsymbol{\mu} \ . \tag{43}$$

In the following, it should be clear from the context when $\boldsymbol{\lambda}$ is a free variable or is to be determined from eq. (43). Inserting the optimizing distribution eq. (42) into the definition of the Gibbs free energy eq. (39), we get the simpler expression:

$$G(\boldsymbol{\mu}) = -\ln Z(\boldsymbol{\lambda}(\boldsymbol{\mu})) + \boldsymbol{\lambda}^T(\boldsymbol{\mu})\boldsymbol{\mu} = \max_{\boldsymbol{\lambda}} \left\{ -\ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T \boldsymbol{\mu} \right\} \ . \tag{44}$$

showing that $G(\boldsymbol{\mu})$ is the Lagrangian dual of $\ln Z(\boldsymbol{\lambda})$.

Derivatives with respect to parameters. We will use the following result about the derivative of G with respect to a parameter t in the density. Using the notation $p(\mathbf{x}|t) = \frac{f(\mathbf{x},t)}{Z_t}$ (which should not be confused with a conditional probability), we calculate the derivative of $G(\boldsymbol{\mu},t)$ using (43) and (44) as for fixed $\boldsymbol{\mu}$:

$$\frac{dG(\boldsymbol{\mu},t)}{dt} = -\frac{\partial \ln Z(\boldsymbol{\lambda},t)}{\partial t} + \left(\boldsymbol{\mu} - \frac{\partial \ln Z(\boldsymbol{\lambda},t)}{\partial \boldsymbol{\lambda}} \right) \frac{d\boldsymbol{\lambda}^T}{dt} = -\frac{\partial \ln Z(\boldsymbol{\lambda},t)}{\partial t} \ , \tag{45}$$

where $Z(\boldsymbol{\lambda},t) = \int d\mathbf{x}\, f(\mathbf{x},t) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right)$.

B.2 An Interpolation Representation of Free Energies

If the density p factors into a tractable f_q and an intractable part f_r, according to eq. (2), we can construct a representation of the Gibbs free energy which also separates into two corresponding parts. This is done by treating $f_r(\mathbf{x})$ as a *perturbation* which is smoothly turned on using a parameter $0 \leq t \leq 1$. We define $f_r(\mathbf{x}, t)$ to be a smooth interpolation between the trivial $f_r(\mathbf{x}, t = 0) = 1$ and the "full" intractable $f_r(\mathbf{x}, t = 1) = f_r(\mathbf{x})$. The most common choice is to set $f_r(\mathbf{x}, t) = [f_r(\mathbf{x})]^t$, but a more complicated construction can be necessary, when f_r contains δ-distributions, see appendix E. However, we will see later, that an explicit construction of the interpolation will not be necessary for our approximation.

Next, we define the interpolating density and the associated optimizing distribution for the Gibbs free energy

$$\begin{aligned} p(\mathbf{x}|t) &= \frac{1}{Z_t} f_q(\mathbf{x}) f_r(\mathbf{x}, t) & (46) \\ q(\mathbf{x}|t) &= \frac{1}{Z_q(\boldsymbol{\lambda}, t)} f_q(\mathbf{x}) f_r(\mathbf{x}, t) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \ , & (47) \end{aligned}$$

where

$$Z_q(\boldsymbol{\lambda}, t) = \int d\mathbf{x}\, f_q(\mathbf{x}) f_r(\mathbf{x}, t) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \tag{48}$$

and the corresponding free energy $G_q(\boldsymbol{\mu}, t) = \max_{\boldsymbol{\lambda}} \left\{ - \ln Z_q(\boldsymbol{\lambda}, t) + \boldsymbol{\lambda}^T \boldsymbol{\mu} \right\}$. For later convenience, we have given a subscript to G and $\ln Z$ to indicate which approximating distribution is being used. We can now use the following simple identity for the free energy $G(\boldsymbol{\mu}, t)$

$$G(\boldsymbol{\mu}, 1) - G(\boldsymbol{\mu}, 0) = \int_0^1 dt\, \frac{dG(\boldsymbol{\mu}, t)}{dt} \tag{49}$$

to relate the Gibbs free energy of the intractable model $G(\boldsymbol{\mu}) = G(\boldsymbol{\mu}, t = 1)$ and tractable model $G(\boldsymbol{\mu}, t = 0)$. Using eq. (20), we get

$$\frac{dG(\boldsymbol{\mu}, t)}{dt} = -\frac{\partial \ln Z(\boldsymbol{\lambda}, t)}{\partial t} = -\left\langle \frac{d \ln f_r(\mathbf{x}, t)}{dt} \right\rangle_{q(\mathbf{x}|t)} \ . \tag{50}$$

While this representation can be used to re-derive a variational bound approximation (see Appendix F), we will next re-derive a dual representation of the EC free energy by making an approximation similar in spirit to the one used in Section 3. We again assume that besides the family of distributions eq. (4), there is a second family which can be used as an approximation to the distribution eq. (46). It is defined by

$$r(\mathbf{x}|t) = \frac{1}{Z_r(\boldsymbol{\lambda}, t)} f_r(\mathbf{x}, t) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \ , \tag{51}$$

where, as before the parameters $\boldsymbol{\lambda}$ are chosen in such a way as to guarantee *consistency for the expectations* of $\mathbf{g}$, i.e. $\langle \mathbf{g}(\mathbf{x}) \rangle_{r(\mathbf{x}|t)} = \boldsymbol{\mu}$ and

$$Z_r(\boldsymbol{\lambda}, t) = \int d\mathbf{x}\, f_r(\mathbf{x}, t) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \ . \tag{52}$$

Obviously, $r(\mathbf{x}|t)$ defines another Gibbs free energy which in its dual representation eq. (44) is given by

$$G_r(\boldsymbol{\mu}, t) = \max_{\boldsymbol{\lambda}} \left\{ -\ln Z_r(\boldsymbol{\lambda}, t) + \boldsymbol{\lambda}^T \boldsymbol{\mu} \right\} \ . \tag{53}$$

Using the density $r(\mathbf{x}|t)$ to treat the integral in eq. (49), we make the approximation

$$\int_0^1 dt \ \left\langle \frac{d \ln f_r(\mathbf{x}, t)}{dt} \right\rangle_{q(\mathbf{x}|t)} \approx \int_0^1 dt \ \left\langle \frac{d \ln f_r(\mathbf{x}, t)}{dt} \right\rangle_{r(\mathbf{x}|t)} \ . \tag{54}$$

The fact that both types of densities eqs. (47) and (51) contain the same exponential factor $f_r(\mathbf{x}, t) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right)$ allows us to carry out the integral over the interaction strength t on the right hand side of eq. (54) in closed form without specifying the interpolating term $f_r(\mathbf{x}, t)$ explicitly. We simply use the relations eqs. (49) and (50) again, but this time for the free energy eq. (53) to get

$$\int_0^1 dt \ \left\langle \frac{d \ln f_r(\mathbf{x}, t)}{dt} \right\rangle_{r(\mathbf{x}|t)} = G_r(\boldsymbol{\mu}, 1) - G_r(\boldsymbol{\mu}, 0) \ . \tag{55}$$

Using the approximation eq. (54) and the two exact relation eqs. (49) for q and r we arrive at the *expectation consistent (EC)* approximation:

$$G_q(\boldsymbol{\mu}, 1) \approx G_q(\boldsymbol{\mu}, 0) + G_r(\boldsymbol{\mu}, 1) - G_r(\boldsymbol{\mu}, 0) \equiv G^{\mathrm{EC}}(\boldsymbol{\mu}) \ . \tag{56}$$

Recovering the EC free energy eq. (19) Using the duality expression for the free energies eq. (44), the free energy approximation can be written as

$$\begin{aligned} G^{\mathrm{EC}}(\boldsymbol{\mu}) &= G_q(\boldsymbol{\mu}) + G_r(\boldsymbol{\mu}) - G_s(\boldsymbol{\mu}) \qquad (57) \\ &= \max_{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r} \min_{\boldsymbol{\lambda}_s} \left\{ -\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_r) + \ln Z_s(\boldsymbol{\lambda}_s) + \boldsymbol{\mu}^T (\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r - \boldsymbol{\lambda}_s) \right\} \ , \end{aligned}$$

where we have defined $G_q(\boldsymbol{\mu}) = G_q(\boldsymbol{\mu}, 0)$, $G_r(\boldsymbol{\mu}) = G_r(\boldsymbol{\mu}, 1)$ and $G_s(\boldsymbol{\mu}) = G_r(\boldsymbol{\mu}, 0)$. To obtain the corresponding approximation for the Helmholtz free energy $-\ln Z$, we should minimize this expression with respect to $\boldsymbol{\mu}$. Any local minimum will be characterized by the vanishing of the partial derivative with respect to $\boldsymbol{\mu}$. This yields the following constraint on the Lagrange parameters

$$\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r - \boldsymbol{\lambda}_s = 0 \ , \tag{58}$$

which can be used to eliminate, say $\boldsymbol{\lambda}_r$ and we recover eq. (19).

Recovering the double loop algorithm. Since the free energy given by eq. (44) is a convex function of $\boldsymbol{\mu}$, we can see that the EC approximation eq. (56) appears directly as a sum of a convex (the first two terms) and a concave function of $\boldsymbol{\mu}$. Hence, the approximation is not guaranteed to be convex, and multiple local minima and other stationary points may occur. However, this natural split allows us to develop a double loop algorithm similar to Yuille (2002); Heskes et al. (2003), which is guaranteed to converge to at least one of the stationary points, provided that the EC free energy is bounded from below. Assume that at iteration step t, the current approximation to the minimizer $\boldsymbol{\mu}(t)$, such an algorithm first upper bounds the concave function $-G_s(\boldsymbol{\mu})$ by the linear function $-\left(\boldsymbol{\mu}-\boldsymbol{\mu}(t)\right)^T \left.\frac{\partial G_s(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}\right|_{\boldsymbol{\mu}=\boldsymbol{\mu}(t)}$.

In terms of the corresponding Lagrange-parameter $\boldsymbol{\lambda}_s(t) = \left.\frac{\partial G_s(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}\right|_{\boldsymbol{\mu}=\boldsymbol{\mu}(t)}$, this yields

$$\begin{aligned} G^{\mathrm{EC}}(\boldsymbol{\mu}) &\leq G_q(\boldsymbol{\mu}) + G_r(\boldsymbol{\mu}) - \left(\boldsymbol{\mu}-\boldsymbol{\mu}(t)\right)^T \boldsymbol{\lambda}_s(t) \\ &= \max_{\boldsymbol{\lambda}_q,\boldsymbol{\lambda}_r} \left\{-\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_r) + \boldsymbol{\mu}^T(\boldsymbol{\lambda}_q+\boldsymbol{\lambda}_r) + \ln Z_s(\boldsymbol{\lambda}_s(t))\right\} \equiv G_t^{\mathrm{EC}}(\boldsymbol{\mu}) \end{aligned}$$

Minimizing $G_t^{\mathrm{EC}}(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$, we immediately get

$$\min_{\boldsymbol{\mu}} G_t^{\mathrm{EC}}(\boldsymbol{\mu}) = \max_{\boldsymbol{\lambda}_q} \{-\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)\} + \ln Z_s(\boldsymbol{\lambda}_s(t)) = F(\boldsymbol{\lambda}_s(t)) \ , \qquad (59)$$

where $F(\boldsymbol{\lambda}_s(t))$ was introduced in eq. (26). The new approximation is computed as

$$\boldsymbol{\mu}(t+1) = \langle \mathbf{g}(\mathbf{x}) \rangle_{q(t+1)} \ .$$

Hence, this double loop procedure is equivalent to the one defined in Section 5, demonstrating that the sequence $F(\boldsymbol{\lambda}_s(t))$ yields nondecreasing upper bounds to the minimal EC Gibbs free energy.

Appendix C. Tree-Connected Graphs

For the EC tree approximation we will need to make inference on tree-connected graphs. To handle a problem with binary variables both binary and Gaussian distributed variables on a tree will be needed. We will write the model as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_i \psi_i(\mathbf{x}_i) \exp\left(-\frac{1}{2}\mathbf{x}^T \boldsymbol{\Lambda} \mathbf{x} + \boldsymbol{\gamma}^T \mathbf{x}\right) \ ,$$

where $\psi_i(x_i) = \delta(x_i - 1) + \delta(x_i + 1)$ for binary and $\psi_i(x_i) = 1$ for Gaussian. Assuming that $\boldsymbol{\Lambda}$ defines a tree one can express the free energy in terms of single- and two-node free energies (Yedidia et al., 2001):

$$-\ln Z(\boldsymbol{\lambda}) = -\sum_{(ij)\in\mathcal{G}} \ln Z_{ij}(\boldsymbol{\lambda}^{(ij)}) - \sum_i (1-n_i) \ln Z_i(\boldsymbol{\lambda}^{(i)}) \ , \qquad (60)$$

where $\boldsymbol{\lambda}^{(ij)} = \left(\gamma_i^{(ij)}, \gamma_j^{(ij)}, \Lambda_{ii}^{(ij)}, \Lambda_{ij}^{(ij)}, \Lambda_{jj}^{(ij)}\right)$ are the parameters associated with the moments $\mathbf{g}^{(ij)} = \left(x_i, x_j, -\frac{x_i^2}{2}, -x_i x_j, -\frac{x_j^2}{2}\right)$ and n_i is the number of links to node i. The two-node partition function Z_{ij} is given by

$$Z_{ij}(\boldsymbol{\lambda}^{(ij)}) = \int dx_i dx_j \psi_i(x_i) \psi_j(x_j) e^{\gamma_i x_i + \gamma_j x_j - \Lambda_{ij} x_i x_j - \Lambda_{ii} x_i^2/2 - \Lambda_{jj} x_j^2/2} \ . \qquad (61)$$

The one-node partition function is defined in a similar fashion.

The Gibbs free energy $G(\boldsymbol{\mu}) = \max_{\boldsymbol{\lambda}}\{-\ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T\boldsymbol{\mu}\}$ can be written in terms of one- and two-node Gibbs free energies:

$$\begin{aligned} G(\boldsymbol{\mu}) &= \sum_{(ij)\in\mathcal{G}} \ln G_{ij}(\boldsymbol{\mu}^{(ij)}) - \sum_i (1-n_i) G_i(\boldsymbol{\mu}^{(i)}) \\ G_{ij}(\boldsymbol{\mu}^{(ij)}) &= \max_{\boldsymbol{\lambda}^{(ij)}}\{-\ln Z_{ij}(\boldsymbol{\lambda}^{(ij)}) + (\boldsymbol{\lambda}^{(ij)})^T\boldsymbol{\mu}^{(ij)}\} \ , \end{aligned} \tag{62}$$

where $\boldsymbol{\mu}^{(ij)} = \langle \mathbf{g}^{(ij)}(\mathbf{x})\rangle$. We can write $\boldsymbol{\lambda} = \sum_{(ij)\in\mathcal{G}} \boldsymbol{\lambda}^{(ij)} - \sum_i (1-n_i)\boldsymbol{\lambda}^{(i)}$, where $\boldsymbol{\lambda}^{(ij)}$ here should be understood as a vector of the same length as $\mathbf{g}$ having non-zero elements for moments defined for the pair (ij). By solving the max condition we can write the Lagrange parameters in terms of the mean values $m_i = \langle x_i\rangle$ and covariances $\chi_{ij} = \langle x_i x_j\rangle - m_i m_j$. This will be useful when we derive algorithms for optimizing the free energy in section 5 where we need to solve for $\boldsymbol{\lambda}$ in terms of $\boldsymbol{\mu}$. For binary variables we get:

$$\begin{aligned} \gamma_i^{(i)} &= \tanh^{-1}(m_i) \\ \gamma_i^{(ij)} &= \frac{1}{2}\tanh^{-1}\left(\frac{m_i+m_j}{1+\langle x_i x_j\rangle}\right) + \frac{1}{2}\tanh^{-1}\left(\frac{m_i-m_j}{1-\langle x_i x_j\rangle}\right) \\ \gamma_j^{(ij)} &= \frac{1}{2}\tanh^{-1}\left(\frac{m_i+m_j}{1+\langle x_i x_j\rangle}\right) + \frac{1}{2}\tanh^{-1}\left(\frac{m_j-m_i}{1-\langle x_i x_j\rangle}\right) \\ \Lambda_{ij}^{(ij)} &= -\frac{1}{2}\tanh^{-1}\left(\frac{\langle x_i x_j\rangle+m_i}{1+m_j}\right) - \frac{1}{2}\tanh^{-1}\left(\frac{\langle x_i x_j\rangle-m_i}{1-m_j}\right) \end{aligned}$$

and for Gaussian defining $\mathbf{m}^{(ij)} = \begin{pmatrix} m_i \\ m_j \end{pmatrix}$ and $\boldsymbol{\chi}^{(ij)} \equiv \begin{pmatrix} \chi_{ii} & \chi_{ij} \\ \chi_{ji} & \chi_{jj} \end{pmatrix}$:

$$\begin{aligned} \gamma_i^{(i)} &= m_i/\chi_{ii} \qquad &&\text{and} \qquad \Lambda_i^{(i)} = 1/\chi_{ii} \\ \boldsymbol{\gamma}^{(ij)} &= (\boldsymbol{\chi}^{(ij)})^{-1}\mathbf{m}^{(ij)} \qquad &&\text{and} \qquad \boldsymbol{\Lambda}^{(ij)} = (\boldsymbol{\chi}^{(ij)})^{-1} \ . \end{aligned}$$

Finally, we will also need to make inference about the mean values and covariances on the tree for the binary variables. This can be done effectively by message passing on the tree. The message from link (ij) to node i denoted by $r_{(ij)\to i}$ can be obtained by the following recursion (MacKay, 2003)

$$\begin{aligned} r_{(ij)\to i} &= \tanh(-\Lambda_{ij})\tanh(\theta_{j\backslash i}) \\ \theta_{j\backslash i} &= \theta_j + \sum_{k,(jk)\in\mathcal{G},(jk)\neq(ij)} r_{(jk)\to j} \ . \end{aligned}$$

The recursion converges in one collect and one distribute messages sweep (to/from an arbitrarily chosen root node). Inference is linear because the tree contains $N-1$ links. The mean values and correlations are given by

$$\begin{aligned} m_i &= \tanh\left(\theta_i + \sum_{j,(ij)\in\mathcal{G}} r_{(ij)\to i}\right) \\ \langle x_i x_j\rangle &= \frac{e^{-\Lambda_{ij}}\cosh(\theta_{i\backslash j}+\theta_{j\backslash i}) - e^{\Lambda_{ij}}\cosh(\theta_{i\backslash j}-\theta_{j\backslash i})}{e^{-\Lambda_{ij}}\cosh(\theta_{i\backslash j}+\theta_{j\backslash i}) + e^{\Lambda_{ij}}\cosh(\theta_{i\backslash j}-\theta_{j\backslash i})} \ . \end{aligned}$$

Appendix D. Single Loop Algorithmic Recipes

In this appendix we give the algorithmic recipes for one sequential algorithm for the factorized EC and a parallel algorithm for tree EC. The sequential algorithm is close in spirit to Expectation Propagation with $\psi_i(x_i)$ and $\exp\left(\gamma_{r,i}x_i - \frac{1}{2}\Lambda_{r,i}x_i^2\right)$ being what is called exact and approximate factors, respectively (Minka, 2001b):

- Initialize mean and covariance of r-distribution:

$$\begin{aligned} \mathbf{m}_r &:= (\mathbf{\Lambda}_r - \mathbf{J})^{-1}(\boldsymbol{\gamma}_r + \boldsymbol{\theta}) \\ \boldsymbol{\chi}_r &:= (\mathbf{\Lambda}_r - \mathbf{J})^{-1} \end{aligned}$$

  with $\boldsymbol{\gamma}_r = \mathbf{0}$ and $\mathbf{\Lambda}_r$ set such that the covariance is positive definite.

Run sequentially over the nodes:

1. Send message from r to q_i
   - Calculate separator s_i: $\gamma_{s,i} := m_{r,i}/\chi_{r,ii}$ and $\Lambda_{s,i} := 1/\chi_{r,ii}$.
   - Update q_i: $\gamma_{q,i} := \gamma_{s,i} - \gamma_{r,i}$ and $\Lambda_{q,i} := \Lambda_{s,i} - \Lambda_{r,i}$.
   - Update moments of q_i: $m_{q,i} := \tanh(\gamma_{q,i})$ and $\chi_{q,ii} = 1 - m_{q,i}^2$.
2. Send message from q_i to r
   - Calculate separator s_i: $\gamma_{s,i} := m_{q,i}/\chi_{q,ii}$ and $\Lambda_{s,i} := 1/\chi_{q,ii}$.
   - Update r: $\gamma_{r,i} := \gamma_{s,i} - \gamma_{q,i}$, $\Delta\Lambda_{r,i} := \Lambda_{s,i} - \Lambda_{q,i} - \Lambda_{r,i}$ and $\Lambda_{r,i} := \Lambda_{s,i} - \Lambda_{q,i}$.
   - Update moments of r (see eq. 37):

$$\begin{aligned} \boldsymbol{\chi}_r &:= \boldsymbol{\chi}_r - \frac{\Delta\Lambda_{r,i}}{1 + \Delta\Lambda_{r,i}\,[\boldsymbol{\chi}_r]_{ii}}[\boldsymbol{\chi}_r]_i[\boldsymbol{\chi}_r]_i^T \\ \mathbf{m}_r &:= \boldsymbol{\chi}_r(\boldsymbol{\gamma}_r + \boldsymbol{\theta}) \ . \end{aligned}$$

Convergence is reached when and if $\mathbf{m}_r = \mathbf{m}_q$ and $\chi_{r,ii} = \chi_{q,ii}$, $i = 1, \ldots, N$. The computational complexity of the algorithm is $\mathcal{O}(N^3 N_{\text{ite}})$ because each Sherman-Morrison update is $\mathcal{O}(N^2)$ and we make N of those in each sweep over the nodes.

The tree EC algorithm is very similar. The only difference is that it is parallel and uses inference on a tree graph, see appendix C for details on the tree inference:

- Initialize as above.

Update:

1. Send message from r to q
   - Calculate separator s: $[\boldsymbol{\gamma}_s, \mathbf{\Lambda}_s] :=$ Lagrange_Gauss_tree$(\mathbf{m}_r, \text{tree}(\boldsymbol{\chi}_r))$, where tree() sets all non-tree elements to zero.
   - Update q: $\boldsymbol{\gamma}_q := \boldsymbol{\gamma}_s - \boldsymbol{\gamma}_r$ and $\mathbf{\Lambda}_q := \mathbf{\Lambda}_s - \mathbf{\Lambda}_r$.
   - Update moments of q: $[m_q, \boldsymbol{\chi}_q] :=$ inference_binary_tree$(\boldsymbol{\gamma}_q, \mathbf{\Lambda}_q)$ will only return non-zero elements of the covariance on the tree.

2. Send message from q to r
    - Calculate separator s: $[\boldsymbol{\gamma}_s, \boldsymbol{\Lambda}_s] := \text{Lagrange\_Gauss\_tree}(\mathbf{m}_q, \boldsymbol{\chi}_q)$.
    - Update r: $\boldsymbol{\gamma}_r := \boldsymbol{\gamma}_s - \boldsymbol{\gamma}_q$ and $\boldsymbol{\Lambda}_r := \boldsymbol{\Lambda}_s - \boldsymbol{\Lambda}_q$.
    - Update moments of r: $\boldsymbol{\chi}_r := (\boldsymbol{\Lambda}_r - \mathbf{J})^{-1}$ and $\mathbf{m}_r := \boldsymbol{\chi}_r(\boldsymbol{\gamma}_r + \boldsymbol{\theta})$.

Convergence is reached when $\mathbf{m}_q = \mathbf{m}_r$ and $\boldsymbol{\chi}_q = \text{tree}(\boldsymbol{\chi}_r)$. This algorithm is also $\mathcal{O}(N^3 N_{\text{ite}})$ because of the matrix inverse. All other operations are $\mathcal{O}(N)$ even though these will dominate for small N. Typically when convergent both algorithms converge in $N_{\text{ite}} = \mathcal{O}(10)$ steps.

Appendix E. Interpolation Scheme for Discrete Variables

The Ising case eq. (9) can be treated by defining the bimodal density

$$f_r(\mathbf{x}, t) = \prod_{i=1}^{N} \left(\frac{\exp\left[-\frac{t}{1-t}(x_i^4 - 2x_i^2)\right]}{\sqrt{1-t}} \right)$$

which interpolates between a constant function for $t = 0$ and becomes proportional to the Dirac measures eq. (9) in the limit $t \to 1$. Other discrete variables can be treated in a similar fashion.

Appendix F. Re-deriving the Variational Bound Approximation

The choice $f_r(\mathbf{x}, t) = t \ln f_r(\mathbf{x})$ for the interpolation can be used for a perturbation expansion of the free energy $G(\boldsymbol{\mu}, t)$ in powers of t, where at the end one sets $t = 1$. The lowest nontrivial (first) order term is obtained by replacing $q(\mathbf{x}|t)$ by $q(\mathbf{x}|0)$ in eq. (50). In this case, one obtains an approximation to the Gibbs free energy given by

$$G(\boldsymbol{\mu}) \approx G(\boldsymbol{\mu}, 0) - \int_0^1 dt \left\langle \frac{d \ln f_r(\mathbf{x}, t)}{dt} \right\rangle_{q(\mathbf{x}|0)} = G(\boldsymbol{\mu}, 0) - \langle \ln f_r(\mathbf{x}) \rangle_{q(\mathbf{x}|0)} \ . \qquad (63)$$

For the second order term of this so-called Plefka expansion see, e.g. Plefka (1982) and several contributions in Opper and Saad (2001).

For comparison, we define a variational bound approximation, where the minimization in eq. (39) is restricted to the family $\mathcal{F}$ of densities of the form eq. (4), i.e.

$$G^{\text{var}}(\boldsymbol{\mu}) = \min_{q \in \mathcal{F}} \{ KL(q, p) \mid \langle \mathbf{g}(\mathbf{x}) \rangle_q = \boldsymbol{\mu} \} - \ln Z \ . \qquad (64)$$

Since we are minimizing in a restricted class of distributions, we obtain the upper bound $G(\boldsymbol{\mu}) \leq G^{\text{var}}(\boldsymbol{\mu})$ on the Gibbs free energy. Using the fact that the density eq. (4) is exactly of the form of $q(\mathbf{x}|0)$, we can show that $G^{\text{var}}(\boldsymbol{\mu})$ coincides exactly with eq. (63).

References

H. Attias. A variational Bayesian framework for graphical models. In T. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

C. M. Bishop, D. Spiegelhalter, and J. Winn. Vibes: A variational inference engine for Bayesian networks. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 777–784. MIT Press, 2003.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

D. Cornford, L. Csató, D. J. Evans, and M. Opper. Bayesian analysis of the scatterometer wind retrieval inverse problem: Some new approaches. *Journal Royal Statistical Society B*, 66:1–17, 2004.

L. Csató, M. Opper, and O. Winther. TAP Gibbs free energy, belief propagation and sparsity. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 657–663, Cambridge, MA, 2002. MIT Press.

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.

T. Heskes, K. Albers, and H. Kappen. Approximate inference and constrained optimization. In *Proceedings UAI-2003*, pages 313–320. Morgan Kaufmann, 2003.

T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In A. Darwiche and N. Friedman, editors, *Proceedings UAI-2002*, pages 216–233, 2002.

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.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37:183–233, 1999.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

D. Malzahn and M. Opper. An approximate analytical approach to resampling averages. *Journal of Machine Learning Research*, pages 1151–1173, 2003.

D. Malzahn and M. Opper. Approximate analytical bootstrap averages for support vector classifiers. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16.* MIT Press, 2004.

M. Mézard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*, volume 9 of *Lecture Notes in Physics.* World Scientific, 1987.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In J. S. Breese and D. Koller, editors, *Proceedings UAI-2001*, pages 362–369. Morgan Kaufmann, 2001a.

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

T. P. 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, 2004.

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

M. Opper and O. Winther. Mean field methods for classification with gaussian processes. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 309–315. MIT Press, 1999.

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

M. Opper and O. Winther. Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling. *Phys. Rev. E*, 64:056131, 2001a.

M. Opper and O. Winther. Tractable approximations for probabilistic models: The adaptive Thouless-Anderson-Palmer mean field approach. *Phys. Rev. Lett.*, 86:3695, 2001b.

M. Opper and O. Winther. Variational linear response. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16.* MIT Press, Cambridge, MA, 2004.

T. Plefka. Convergence condition of the TAP equation for the infinite-range Ising spin glass. *J. Phys. A*, 15:1971, 1982.

J. Quiñonero-Candela and O. Winther. Incremental gaussian processes. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1001–1008. MIT Press, 2003.

G. Roepstorff. *Path Integral Approach to Quantum Physics, An Introduction.* Springer - Verlag Berlin Heidelberg, New York, 1994.

M. Suzuki, editor. *Coherent Anomaly Method, Mean Field, Fluctuations and Symmetries.* World Scientific, 1995.

L. Vandenberghe, S. Boyd, and S.-P Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19:499–533, 1998.

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.

M. J. Wainwright and M. I. Jordan. A variational principle for graphical models. In S. Haykin, J. Principe, S. Sejnowski, and J McWhirter, editors, *New Directions in Statistical Signal Processing: From Systems to Brain*. MIT Press, 2005.

M. Welling and Y.W. Teh. Approximate inference in Boltzmann machines. *Artificial Intelligence*, 143:19–50, 2003.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 689–695, 2001.

A. L. Yuille. CCCP algorithms to minimize the Bethe and Kikuchi free energies: convergent alternatives to belief propagation. *Neural Computation*, 14:1691–1722, 2002.

A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15: 915–936, 2003.