Cumulant expansions for improved inference with EP in discrete Bayesian networks

Manfred Opper
Technical University Berlin
Berlin, Germany
opperm@cs.tu-berlin.de

Ulrich Paquet
Microsoft Research
Cambridge, United Kingdom
ulripa@microsoft.com

Ole Winther
Technical University of Denmark
Lyngby, Denmark
owi@imm.dtu.dk

Abstract

Inference in discrete graphical models is often computationally intractable, requiring the summation of exponentially many terms. Amongst relaxations to this problem, we consider an Expectation Propagation (EP) approximation, and derive a general framework for corrections to EP and apply the framework to inference in binary Bayesian networks (Ising models). We show how to systematically compute higher order cumulant corrections of marginal likelihoods, marginal distributions, and moments, and use it to show state-of-the accuracy in difficult benchmarks.

1 Introduction

Reliable estimation of marginal likelihoods, predictive or marginal distributions and moments are crucial for the practical application of Bayesian inference. Expectation propagation (EP) has proven to be especially well-suited for Gaussian latent variable models. What is less well known is that EP can also be applied to discrete inference problems. In this contribution we will derive a general framework for corrections to the EP and apply the framework to inference in binary Bayes networks (Ising models).

EP can be considered as the zeroth order approximation in a specific cumulant (Edgeworth) expansion, and has the pleasing property amongst variational methods that at the stationary point of the marginal likelihood approximation, it is exact up to the second order cumulants. The corrections we derive here incorporate the remaining non-Gaussian cumulants that are neglected when tractable approximations to latent Gaussian models are made. The typical $\mathcal{O}(N^3)$ complexity with system size N of EP is retained in this approach because it lowest order corrections are computed after convergence of EP in $\mathcal{O}(N^3)$. We investigate how the structure of the variational approximation can be chosen within different tractable families to give factorized and tree-structured approximations with state-of-the accuracy in difficult benchmarks.

This paper specifically addresses corrections to the Gaussian approximating family, and follows on earlier work by the authors [4]. It is by no means unique in its approach to correcting the approximation, as is evinced by cluster-based expansions [6], marginal corrections for EP [2] and corrections to Loopy Belief Propagation [1, 7].

2 Ising model

In this workshop paper, we consider an Ising model over $\mathbf{x} = (x_1, \ldots, x_N)$ as a specific case of a *Gaussian latent variable model*, which we generally define as a product of terms $t_n(x_n)$ with a quadratic exponential $f_0(\mathbf{x})$, i.e. $p(\mathbf{x}) = \frac{1}{Z} \prod_n t_n(x_n)\, f_0(\mathbf{x})$ with partition function (normalizer) $Z = \int \prod_n t_n(x_n)\, f_0(\mathbf{x})\, d\mathbf{x}$.

An Ising model can be constructed by letting the terms t_n restrict x_n to ± 1 (through Dirac delta functions), and introducing the symmetric coupling matrix $\mathbf{J}$ and field $\boldsymbol{\theta}$ into f_0 with

$$p(\mathbf{x}) = \frac{1}{Z} \prod_n \left[\frac{1}{2}\delta(x_n + 1) + \frac{1}{2}\delta(x_n - 1) \right] \exp\left\{ \frac{1}{2}\mathbf{x}^T \mathbf{J}\mathbf{x} + \boldsymbol{\theta}^T \mathbf{x} \right\} . \tag{1}$$

In the Ising model, the partition function Z is intractable, as it sums $f_0(\mathbf{x})$ over 2^N binary values of $\mathbf{x}$. In the variational approaches, the intractability is addressed by allowing approximations to Z and other marginal distributions, decreasing the computational complexity from being exponential to polynomial in N, and typically cubic for EP.

3 Expectation Propagation

An approximation to Z or other marginalizations can be made by allowing $p(\mathbf{x})$ in to factorize into a product of *factors* f_a. This factorization is not unique; for example, a three-term product may be factorized as $(t_1)(t_2)(t_3)$, but could equally factorize as $(t_1 t_2)(t_2 t_3)/(t_2)$, when the resulting free energy is be that of the tree-structured EC approximation [5]. To therefore allow for regrouping, combining, splitting, and dividing terms, a power D_a is associated with each f_a, such that

$$p(\mathbf{x}) = \frac{1}{Z} \prod_a f_a(\mathbf{x})^{D_a} \tag{2}$$

with intractable normalization (or partition function) $Z = \int d\mathbf{x} \prod_a f_a(\mathbf{x})^{D_a}$. To define an approximation to p, terms g_a, which take an exponential family form, are chosen such that

$$q(\mathbf{x}) = \frac{1}{Z_q} \prod_a g_a(\mathbf{x})^{D_a} \tag{3}$$

has the same structure as p's factorization. Although not shown explicitly, f_a and g_a have a dependence on the *same* subset of variables $\mathbf{x}_a$. The optimal parameters of the g_a-term approximations are found through a set of auxiliary *tilted* distributions, defined by

$$q_a(\mathbf{x}) = \frac{1}{Z_a} \left(\frac{q(\mathbf{x}) f_a(\mathbf{x})}{g_a(\mathbf{x})} \right) . \tag{4}$$

Here a *single* approximating term g_a is replaced by an original term f_a. Assuming that this replacement leaves q_a still tractable, the parameters in g_a are determined by the condition that $q(\mathbf{x})$ and all $q_a(\mathbf{x})$ should be made as similar as possible. This is usually achieved by requiring that these distributions share a set of generalised moments which usually coincide with the sufficient statistics of the exponential family. For example with sufficient statistics $\phi(\mathbf{x})$ we require that

$$\langle \phi(\mathbf{x}) \rangle_{q_a} = \langle \phi(\mathbf{x}) \rangle_q \quad \text{for all } a . \tag{5}$$

The partition function associated with this approximation is $Z_{\mathrm{EP}} = Z_q \prod_a Z_a^{D_a}$.

Continuous approximations to discrete problems. As $p(\mathbf{x})$ is a latent Gaussian model, the g-terms in eq. (3) are chosen in this paper to give a Gaussian approximation

$$q(\mathbf{x}) = \frac{1}{Z_q} \exp\{\boldsymbol{\lambda}^T \phi(\mathbf{x})\} = \mathcal{N}(\mathbf{x}\,;\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) .$$

The sufficient statistics $\phi(\mathbf{x})$ and natural parameters $\boldsymbol{\lambda}$ of the Gaussian are defined as

$$\phi(\mathbf{x}) = (\mathbf{x}, -\tfrac{1}{2}\mathbf{x}\mathbf{x}^T) \quad \text{and} \quad \boldsymbol{\lambda} = (\boldsymbol{\gamma}, \boldsymbol{\Lambda}) ,$$

where $\boldsymbol{\lambda}^T \phi(\mathbf{x}) = \boldsymbol{\gamma}^T \mathbf{x} - \frac{1}{2} \operatorname{tr}[\boldsymbol{\Lambda}\mathbf{x}\mathbf{x}^T] = \boldsymbol{\gamma}^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \boldsymbol{\Lambda}\mathbf{x}$. We define $g_0(\mathbf{x}) = \exp\{\boldsymbol{\lambda}_0^T \phi(\mathbf{x})\}$, where $\boldsymbol{\lambda}_0 = (\boldsymbol{\gamma}^{(0)}, \boldsymbol{\Lambda}^{(0)})$, such that it is essentially a rescaling of factor f_0. In the Ising model in eq. (1), this means that $\boldsymbol{\Lambda}^{(0)} = -\mathbf{J}$ and $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\theta}$.

Tree structured example. Let $\mathcal{G}$ define a spanning tree structure over all $\mathbf{x}$, and let $\boldsymbol{\tau} = (m, n) \in \mathcal{G}$ define the edges in the tree. Let d_n be the number of edges emanating from node x_n in the graph. Through a clever regrouping of terms into a "junction tree" form with

$$\prod_n t_n(x_n) = \frac{\prod_{\boldsymbol{\tau}=(m,n)}[\, t_m(x_m)t_n(x_n)\,]}{\prod_n t_n(x_n)^{d_n-1}} = \frac{\prod_{\boldsymbol{\tau}} f_{\boldsymbol{\tau}}(\mathbf{x})}{\prod_n f_n(\mathbf{x})^{d_n-1}}\ , \tag{6}$$

the term-approximation will be tree-structured. In this example the D_a powers are 1 for edge factors $f_{\boldsymbol{\tau}}$ and $(1 - d_n)$ for node factors f_n. Suppressing $g_0 = f_0$, we define the approximation through

$$\frac{\prod_{\boldsymbol{\tau}} g_{\boldsymbol{\tau}}(\mathbf{x})}{\prod_n g_n(\mathbf{x})^{d_n-1}} = \frac{\prod_{\boldsymbol{\tau}} \exp\{\boldsymbol{\lambda}_{\boldsymbol{\tau}}^\top \phi(\mathbf{x})\}}{\prod_n \exp\{\boldsymbol{\lambda}_n^\top \phi(\mathbf{x})\}^{d_n-1}}\ .$$

The natural parameters of $g_n(\mathbf{x}) = \exp\{\boldsymbol{\lambda}_n^T \phi(\mathbf{x})\}$ are chosen to be $\boldsymbol{\lambda}_n = (\gamma_n^{(n)}, \Lambda_{nn}^{(n)})$, corresponding to $\phi_n(x_n) = (x_n, -\frac{1}{2}x_n^2)$.[1] Furthermore, let $g_{\boldsymbol{\tau}}(\mathbf{x})$ be similarly parameterized by $\boldsymbol{\lambda}_{\boldsymbol{\tau}} = (\gamma_m^{(\boldsymbol{\tau})}, \gamma_n^{(\boldsymbol{\tau})}, \Lambda_{mm}^{(\boldsymbol{\tau})}, \Lambda_{nn}^{(\boldsymbol{\tau})}, \Lambda_{mn}^{(\boldsymbol{\tau})})$, with by symmetry $\Lambda_{nm}^{(\boldsymbol{\tau})} = \Lambda_{mn}^{(\boldsymbol{\tau})}$. The resulting $q(\mathbf{x})$ therefore has parameter vector $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 + \sum_{\boldsymbol{\tau}} \boldsymbol{\lambda}_{\boldsymbol{\tau}} - \sum_n (d_n - 1)\boldsymbol{\lambda}_n$.[2]

4 Corrections to EP

We present a derivation of the cumulant expansion of R in eq. (16). *The strategy can be re-used to correct other quantities of interest, like marginal distributions or the predictive density of new data when $p(\mathbf{x})$ is a Bayesian probabilistic model.*

Exact expression for correction. We define the (intractable) correction R as $Z = RZ_{\mathrm{EP}}$. We can derive a useful expression for R in a few steps as follows: First we solve f_a in eq. (4), and substituting this into eq. (2) obtain

$$\prod_a f_a(\mathbf{x})^{D_a} = \prod_a \left(\frac{Z_a q_a(\mathbf{x}) g_a(\mathbf{x})}{q(\mathbf{x})}\right)^{D_a} = Z_{\mathrm{EP}}\ q(\mathbf{x}) \prod_a \left(\frac{q_a(\mathbf{x})}{q(\mathbf{x})}\right)^{D_a}\ . \tag{7}$$

We introduce $F(\mathbf{x}) \equiv \prod_a (q_a(\mathbf{x})/q(\mathbf{x}))^{D_a}$ to derive the expression for the correction $R = Z/Z_{\mathrm{EP}}$ by integrating eq. (7)

$$R = \int d\mathbf{x}\ q(\mathbf{x}) F(\mathbf{x}) \tag{8}$$

and using $Z = \int d\mathbf{x}\ \prod_a f_a(\mathbf{x})^{D_a}$. Similarly we can write:

$$p(\mathbf{x}) = \frac{1}{Z}\prod_a f_a(\mathbf{x})^{D_a} = \frac{Z_{\mathrm{EP}}}{Z}\ q(\mathbf{x})F(\mathbf{x}) = \frac{1}{R}\ q(\mathbf{x})F(\mathbf{x})\ . \tag{9}$$

Corrections to the marginal and predictive densities of $p(\mathbf{x})$ can be computed from this formulation. This expression will become especially useful because the terms in $F(\mathbf{x})$ turn out to be "local", and only depend on the marginals of the variables associated with factor a. Let $f_a(\mathbf{x})$ depend on the subset $\mathbf{x}_a$ of $\mathbf{x}$, and let $\mathbf{x}_{\setminus a}$ ("$\mathbf{x}$ without a") denote the remaining variables. The distributions in eqs. (3) and (4) differ only with respect to their marginals on $\mathbf{x}_a$, $q_a(\mathbf{x}_a)$ and $q(\mathbf{x}_a)$, and therefore $q_a(\mathbf{x})/q(\mathbf{x}) = q_a(\mathbf{x}_a)/q(\mathbf{x}_a)$. Now we can rewrite $F(\mathbf{x})$ in terms of marginals:

$$F(\mathbf{x}) = \prod_a \left(\frac{q_a(\mathbf{x}_a)}{q(\mathbf{x}_a)}\right)^{D_a}\ . \tag{10}$$

The key quantity, then, is F, after which the key operation is to compute its expected value. The rest of this section is devoted to the task of obtaining a "handle" on F.

[1] For clarity the other γ and Λ parameters in $\boldsymbol{\lambda}_n$ are not shown, as they are clamped at zero.

[2] Only the term approximation is tree-structured through $\boldsymbol{\lambda}_{\boldsymbol{\tau}}$ and $\boldsymbol{\lambda}_n$, whilst $\boldsymbol{\lambda}_0$ from g_0 is not.

Characteristic functions and cumulants. The distributions present in each of the ratios in $F(\mathbf{x})$ in eq. (10) share their first two cumulants, mean and covariance. As the $q(\mathbf{x}_a)$'s are Gaussian and do not contain any higher order cumulants (three and above), F can be expressed in terms of the higher cumulants of the *marginals* $q_a(\mathbf{x}_a)$. When the term-product approximation is fully factorized, these are simply cumulants of *one*-dimensional distributions.

Let N_a be the number of variables in subvector $\mathbf{x}_a$—in our results N_a is one, or two for a tree-structured approximation. We let $\mathbf{k}_a$ be an N_a-dimensional vector $\mathbf{k}_a = (k_1, \ldots, k_{N_a})_a$. The characteristic function of q_a is $\chi_a(\mathbf{k}_a) = \int d\mathbf{x}_a \; e^{i\mathbf{k}_a^T \mathbf{x}_a} \, q_a(\mathbf{x}_a) = \left\langle e^{i\mathbf{k}_a^T \mathbf{x}_a} \right\rangle_{q_a}$, and is obtained through its Fourier transform. Inversely, $q_a(\mathbf{x}_a) = \frac{1}{(2\pi)^{N_a}} \int d\mathbf{k}_a \; e^{-i\mathbf{k}_a^T \mathbf{x}_a} \chi_a(\mathbf{k}_a)$. The cumulants $c_{\boldsymbol{\alpha} a}$ of q_a are the coefficients that appear in the Taylor expansion of $\log \chi_a(\mathbf{k}_a)$ around the $\mathbf{k}_a = \mathbf{0}$ zero vector[3]

$$\log \chi_a(\mathbf{k}_a) = \sum_{l=1}^{\infty} i^l \sum_{|\boldsymbol{\alpha}|=l} \frac{c_{\boldsymbol{\alpha} a}}{\boldsymbol{\alpha}!} \, \mathbf{k}_a^{\boldsymbol{\alpha}} \; .$$

There are two characteristic functions that come into play in $F(\mathbf{x})$ and R in eq. (9). The first is that of the tilted distribution, $\log \chi_a(\mathbf{k}_a)$, and the other is the characteristic function of the EP marginal $q(\mathbf{x}_a)$, defined as $\chi(\mathbf{k}_a) = \langle e^{i\mathbf{k}_a^T \mathbf{x}_a} \rangle_q$. By virtue of matching the first two moments, and $q(\mathbf{x}_a)$ being Gaussian,

$$r_a(\mathbf{k}_a) = \log \chi_a(\mathbf{k}_a) - \log \chi(\mathbf{k}_a) = \sum_{l \geq 3} i^l \sum_{|\boldsymbol{\alpha}|=l} \frac{c_{\boldsymbol{\alpha} a}}{\boldsymbol{\alpha}!} \, \mathbf{k}_a^{\boldsymbol{\alpha}} \tag{11}$$

contains the remaining higher-order cumulants where the tilted and approximate distributions *differ*.

The correction as a complex expectation. The expected value of F, which is required for the correction, has a dependence on a product of ratios of distributions $q_a(\mathbf{x}_a)/q(\mathbf{x}_a)$, which simplifies to $q_a(\mathbf{x}_a)/q(\mathbf{x}_a) = \langle \exp r_a(\mathbf{k}_a) \rangle_{\mathbf{k}_a|\mathbf{x}_a}$, with $\mathbf{k}_a$ shifted into the complex plane with $p(\mathbf{k}_a|\mathbf{x}_a) = \mathcal{N}(\mathbf{k}_a \, ; -i\boldsymbol{\Sigma}_a^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_a) \, , \, \boldsymbol{\Sigma}_a^{-1})$. We find that R, from eq. (8), is equal to

$$R = \left\langle F(\mathbf{x}) \right\rangle_{\mathbf{x} \sim q(\mathbf{x})} = \left\langle \prod_a \left\langle \exp r_a(\mathbf{k}_a) \right\rangle_{\mathbf{k}_a|\mathbf{x}_a}^{D_a} \right\rangle_{\mathbf{x}} \; . \tag{12}$$

When $\mathbf{x}$ is given, the $\mathbf{k}_a$-variables are independent, but when the uncertainty in $q(\mathbf{x})$ is taken into account, the $\mathbf{k}_a$-variables are zero-mean *complex* Gaussian $\langle \mathbf{k}_a \rangle = \mathbf{0}$ and are coupled with a zero self-covariance! In other words, if $\boldsymbol{\Sigma}_{ab} = \mathrm{cov}(\mathbf{x}_a, \mathbf{x}_b)$, the covariance $\mathrm{cov}(\mathbf{k}_a, \mathbf{k}_b)$ between the variables in the set $\{\mathbf{k}_a\}$ is

$$\mathrm{cov}(\mathbf{k}_a, \mathbf{k}_b) = \begin{cases} \mathbf{0} & \text{if } a = b \\ -\boldsymbol{\Sigma}_a^{-1} \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_b^{-1} & \text{if } a \neq b \end{cases} \; . \tag{13}$$

When $D_a = 1$, the above expectation can be written directly over $\{\mathbf{k}_a\}$ and expanded. In the general case, the inner expectation is first expanded (to treat the D_a powers) before computing an expectation over $\{\mathbf{k}_a\}$. In both cases the expectation will involve polynomials in k-variables, and the expected values of Gaussian polynomials can be evaluated with Wick's theorem [3].

5 Factorized approximations

In the fully factorized approximation, with $f_n(x_n) = t_n(x_n)$, the exact distribution in eq. (9) depends on the *single node marginals* $F(\mathbf{x}) = \prod_n q_n(x_n)/q(x_n)$. Following eq. (12), the correction to the free energy is taken directly over the centered complex-valued Gaussian random variables $\mathbf{k} = (k_1, \ldots, k_N)$, which have a covariance $\langle k_m k_n \rangle = 0$ if $m = n$ and $\langle k_m k_n \rangle = -\Sigma_{mn}/(\Sigma_{mm}\Sigma_{nn})$ if $m \neq n$. In the section to follow, all expectations shall be with respect to $\mathbf{k}$, which will be dropped where it is clear from the context.

[3] We introduced some notation to facilitate manipulating a multivariate series. The vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{N_a})$, with $\alpha_j \in \mathbf{N}_0$ denotes a multi-index on the elements of $\mathbf{k}_a$. Other notational conventions that employ $\boldsymbol{\alpha}$ (writing k_j instead of k_{aj}) are: $|\boldsymbol{\alpha}| = \sum_j \alpha_j$, $\mathbf{k}_a^{\boldsymbol{\alpha}} = \prod_j k_j^{\alpha_j}$, and $\boldsymbol{\alpha}! = \prod_j \alpha_j!$. For example, when $N_a = 2$, say for the edge-factors in a spanning tree, the set of multi-indices $\boldsymbol{\alpha}$ where $|\boldsymbol{\alpha}| = 3$ are $(3, 0)$, $(2, 1)$, $(1, 2)$, and $(0, 3)$.

Table 1: Average absolute deviation (AAD) of marginals *and* absolute deviation log partition function in a Wainwright-Jordan set-up. Results compare loopy belief propagation (LBP), log-determinant relaxation (LD), EC, EC with $l = 4$ second order correction (EC c), an EC tree (EC t), and EC tree with $l = 4$ second order correction (EC tc). Results in **bold** face highlight best results, while *italics* indicate where the cumulant expression is less accurate than the original approximation.

Problem type			AAD marginals					Absolute deviation log Z			
Graph	Coupling	d_{coup}	LBP	LD	EC	EC c	EC t	EC	EC c	EC t	EC tc
Full	Repulsive	0.25	.037	.020	.003	**.0006**	.0017	.0310	.0061	.0104	**.0010**
		0.50	.071	.018	.031	.0157	**.0143**	.3358	.0697	.1412	**.0440**
	Mixed	0.25	.004	.020	.002	**.0004**	.0013	.0235	.0013	.0129	**.0009**
		0.50	.055	.021	.022	.0159	**.0151**	.3362	.0655	.1798	**.0620**
	Attractive	0.06	.024	.027	.004	**.0023**	.0025	.0236	.0028	.0166	**.0006**
		0.12	.435	.033	.117	.1066	**.0211**	.8297	**.1882**	.2672	.2094
Grid	Repulsive	1.0	.294	.047	.153	*.1693*	**.0031**	1.7776	.8461	.0279	**.0115**
		2.0	.342	.041	.198	*.4244*	**.0021**	4.3555	2.9239	.0086	**.0077**
	Mixed	1.0	.014	.016	.011	*.0122*	**.0018**	.3539	.1443	.0133	**.0039**
		2.0	.095	.038	.082	*.0984*	**.0068**	1.2960	.7057	.0566	**.0179**
	Attractive	1.0	.440	.047	.125	*.1759*	**.0028**	1.6114	.7916	.0282	**.0111**
		2.0	.520	.042	.177	*.4730*	**.0002**	4.2861	2.9350	.0441	**.0433**

Second order perturbation expansion. Thus far, R is re-expressed in terms of site contributions. The expression in eq. (12) is exact, albeit still intractable, and will be treated through a power series expansion. Assuming that the r_n are small, eq. (12) is expanded and the lower order terms kept:

$$\begin{aligned}\log R = \log\left\langle \exp\left[\sum_n r_n(k_n)\right]\right\rangle &= \sum_n \langle r_n\rangle + \frac{1}{2}\left\langle\left(\sum_n r_n\right)^2\right\rangle - \frac{1}{2}\left(\sum_n \langle r_n\rangle\right)^2 + \cdots \\ &= \frac{1}{2}\sum_{m\neq n}\langle r_m r_n\rangle + \cdots \end{aligned} \tag{14}$$

The simplification in the second line is a result of the variance terms being zero from eq. (13). The single marginal terms also vanish (and hence EP is correct to first order) because both $\langle k_n\rangle = 0$ and $\langle k_n^2\rangle = 0$. The expectation $\langle r_m r_n\rangle$, as it appears in eq. (14), is treated by substituting r_n with its cumulant expansion $r_n(k_n) = \sum_{l\geq 3} i^l c_{ln} k_n^l/l!$ from eq. (11). Wick's theorem now plays a pivotal role in evaluating the expectations that appear in the expansion:

$$\langle r_m\, r_n\rangle = \sum_{l,s\geq 3} i^{l+s}\frac{c_{ln}\, c_{sm}}{l!s!}\langle k_m^s k_n^l\rangle = \sum_{l\geq 3} i^{2l} l!\frac{c_{ln}\, c_{sm}}{(l!)^2}\langle k_m k_n\rangle^l = \sum_{l\geq 3}\frac{c_{lm}\, c_{ln}}{l!}\left(\frac{\Sigma_{mn}}{\Sigma_{mm}\Sigma_{nn}}\right)^l . \tag{15}$$

The second last simplification above follows from contractions in Wick's theorem. All the *self-pairing terms*, when for example one of the l k_n's is paired with another k_n, are zero because $\langle k_n^2\rangle = 0$. To therefore get a non-zero result for $\langle k_m^s k_n^l\rangle$, *each* factor k_n has to be paired with some factor k_m, and this is possible only when $l = s$. Wick's theorem sums over all pairings, and there are $l!$ ways of pairing a k_n with a k_m, giving the result in eq. (15). Finally, plugging eq. (15) into eq. (14) gives the second order correction

$$\log R = \frac{1}{2}\sum_{m\neq n}\sum_{l\geq 3}\frac{c_{lm}\, c_{ln}}{l!}\left(\frac{\Sigma_{mn}}{\Sigma_{mm}\Sigma_{nn}}\right)^l + \cdots . \tag{16}$$

6 Ising model results

This section discusses various aspects of corrections to EP as applied to the Ising model, eq. (1), that is a Bayesian network with binary variables and pairwise potentials.

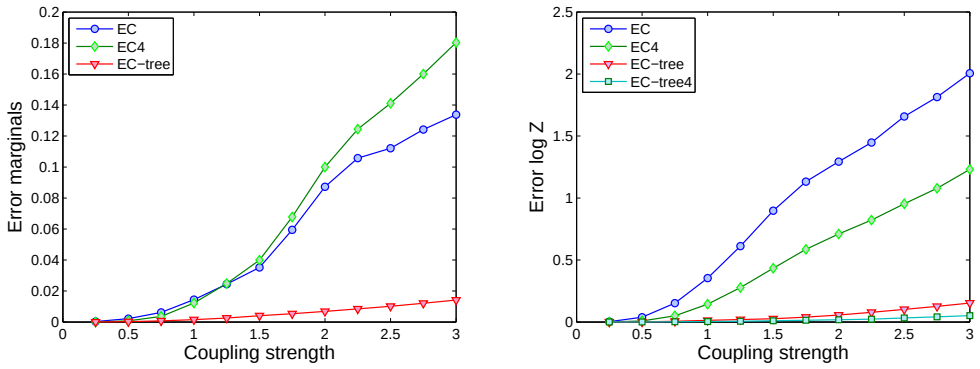

Figure 1: Error on marginal (left) and $\log Z$ (right) for grid and mixed couplings as a function of coupling strength.

We consider the set-up proposed by [8] in which $N = 16$ 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) $J_{ij} \sim \mathcal{U}[0, +2d_{\text{coup}}]$.

Cumulant expansion. For the factorized approximation we use eq. (14) for the $\log Z$ correction. Expressions for corrections to marginal means and the corresponding tree structured correction will be presented elsewhere.

Table 1 gives the average absolute deviation (AAD) of marginals $\text{AAD} = \frac{1}{N}\sum_i |p(x_i = 1) - p(x_i = 1|\text{method})| = \frac{1}{2N}\sum_i |m_i - m_i^{\text{est}}|$, as well as the absolute deviation of $\log Z$ averaged of 100 repetitions. We observe that for the Grid simulations, the corrected marginals in factorized approximation are less accurate than the original approximation. In Figure 1 we vary the coupling strength for a specific set-up (Grid Mixed) and observe a cross-over between the correction and original for the error on marginals as the coupling strength increases. We conjecture that when the error of the original solution is high then the number of terms needed in the cumulant correction increases. The estimation of the marginal seems more sensitive to this than the $\log Z$ estimate. The tree approximation is very precise for the whole coupling strength interval considered and the fourth order cumulant in the second order expansion is therefore sufficient to get often quite large improvements over the original tree approximation.

References

[1] M. Chertkov and V. Y. Chernyak. Loop series for discrete statistical models on graphs. *Journal of Statistical Mechanics: Theory and Experiment*, 2006:P06009, 2006.

[2] B. Cseke and T. Heskes. Approximate marginals in latent Gaussian models. *Journal of Machine Learning Research*, 12:417–457, 2011.

[3] S. Janson. *Gaussian Hilbert spaces*. Cambridge Tracts in Mathematics 129. Cambridge University Press, 1997.

[4] M. Opper, U. Paquet, and O. Winther. Improving on expectation propagation. In *Advances in Neural Information Processing Systems*, pages 1241–1248. 2009.

[5] M. Opper and O. Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6:2177–2204, 2005.

[6] U. Paquet, M. Opper, and O. Winther. Perturbation corrections in approximate inference: Mixture modelling applications. *Journal of Machine Learning Research*, 10:935–976, 2009.

[7] E. Sudderth, M. Wainwright, and A. Willsky. Loop series and Bethe variational bounds in attractive graphical models. In *Advances in Neural Information Processing Systems*, pages 1425–1432. MIT Press, Cambridge, MA, 2008.

[8] M. J. Wainwright and M. I. Jordan. Log-determinant relaxation for approximate inference in discrete Markov random fields. *IEEE Transactions on Signal Processing*, 54(6):2099–2109, 2006.