# Expectation Propagation for Exponential Families

Matthias Seeger
Department of EECS
University of California at Berkeley
485 Soda Hall, Berkeley CA
*mseeger@cs.berkeley.edu*

April 5, 2008

**Abstract**

This is a tutorial describing the Expectation Propagation (EP) algorithm for a general exponential family. Our focus is on simplicity of exposition. Although the overhead of translating a specific model into its exponential family representation can be considerable, many apparent complications of EP can simply be sidestepped by working in this canonical representation.

**Note:** This material is extracted from the Appendix of my PhD thesis (see `www.kyb.tuebingen.mpg.de/bs/people/seeger/papers/thesis.html`).

## 1 Exponential Families

**Definition 1 (Exponential Family)** *A set $\mathcal{F}$ of distributions with densities*

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x}) - \Phi(\boldsymbol{\theta})\right), \quad \boldsymbol{\theta} \in \Theta,$$
$$\Phi(\boldsymbol{\theta}) = \log \int \exp\left(\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})\right)\, d\mu(\boldsymbol{x})$$

*w.r.t. a base measure $\mu$ is called an* exponential family*. Here, $\boldsymbol{\theta}$ are called* natural parameters*, $\Theta$ the* natural parameter space*, $\boldsymbol{\phi}(\boldsymbol{x})$ the* sufficient statistics*, and $\Phi(\boldsymbol{\theta})$ is the* log partition function*. Furthermore, $\boldsymbol{\eta} = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\phi}(\boldsymbol{x})]$ are called* moment parameters*, where $\mathrm{E}_{\boldsymbol{\theta}}[\cdot]$ denotes expectation w.r.t. $P(\boldsymbol{x}|\boldsymbol{\theta})$.*

One of the important reasons for considering exponential families is that the likelihood function for i.i.d. data from $\mathcal{F}$ is a function of the sample average of the sufficient statistics $\boldsymbol{\phi}(\boldsymbol{x})$ which has the fixed dimensionality of $\boldsymbol{\theta}$, independent of the sample size. Even if a model does not give rise to posteriors in an exponential family, members of $\mathcal{F}$ can be used as approximating distributions, since new information can be incorporated without increasing the size of the parametric representation. Many familiar distributions form exponential families, such as Gaussians, multinomials, gammas, etc.

The natural parameter space $\Theta$ for $\boldsymbol{\theta}$ is always convex. If there are linear or affine dependencies between the components of $\boldsymbol{\phi}(\boldsymbol{x})$, then some components in $\boldsymbol{\theta}$ are redundant, and

the representation is called *overcomplete*. Otherwise, it is called *minimal*. Note that many useful properties hold only (in general) for minimal representations, which are also most useful in practice, however sometimes notationally clumsy to work with. Our approach here is to state general properties for minimal representations only, however use these properties for special overcomplete representations occasionally. This can be justified by adding linear constraints on $\boldsymbol{\theta}$, which do not destroy the convexity of $\Theta$. In the remainder of this section, we assume that the representation of $\mathcal{F}$ is minimal.

The log partition function $\Phi(\boldsymbol{\theta})$ is closely related to the cumulant generating function of $\boldsymbol{\phi}(\boldsymbol{x})$, $\boldsymbol{x} \sim P(\boldsymbol{x}|\boldsymbol{\theta})$:

$$\log \mathrm{E}_{\boldsymbol{\theta}} \left[\exp(\boldsymbol{\varepsilon}^T \boldsymbol{\phi}(\boldsymbol{x}))\right] = \Phi(\boldsymbol{\theta} + \boldsymbol{\varepsilon}) - \Phi(\boldsymbol{\theta})$$

which exists iff $\boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \Theta$. Thus, if $\boldsymbol{\theta}$ is in the interior of $\Theta$, the cumulants of $\boldsymbol{\phi}(\boldsymbol{x})$ are obtained as derivatives of $\Phi(\boldsymbol{\theta})$, especially $\nabla_{\boldsymbol{\theta}} \Phi = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\phi}(\boldsymbol{x})] = \boldsymbol{\eta}$ and $\nabla\nabla_{\boldsymbol{\theta}} \Phi = \mathrm{Var}_{\boldsymbol{\theta}}[\boldsymbol{\phi}(\boldsymbol{x})]$. Since the representation is minimal, we see that $\Phi(\boldsymbol{\theta})$ is strictly convex, and using Legendre duality [3] we obtain the following

**Lemma 1 (Natural and Moment Parameters)** *If $\mathcal{F}$ is an exponential family with minimal representation, then there is a bijective mapping between the natural parameters $\boldsymbol{\theta}$ and the moment parameters $\boldsymbol{\eta}$. The log partition function $\Phi(\boldsymbol{\theta})$ is strictly convex and has the Legendre dual*

$$\Psi(\boldsymbol{\eta}) = \mathrm{E}_{\boldsymbol{\eta}} \left[\log P(\boldsymbol{x}|\boldsymbol{\eta})\right],$$

*where $\mathrm{E}_{\boldsymbol{\eta}}[\cdot]$ denotes expectation w.r.t. $P(\boldsymbol{x}|\boldsymbol{\eta}) = P(\boldsymbol{x}|\boldsymbol{\theta}(\boldsymbol{\eta}))$. Conversions between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are done as follows:*

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \Phi, \quad \boldsymbol{\theta}(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \Psi.$$

$\Psi$ is a convex function of the moment parameters $\boldsymbol{\eta}$ (strictly convex for a minimal representation), but in many cases there is no simple explicit form for $\Psi$ so that the Legendre pair $\boldsymbol{\theta}$ has to be found in order to evaluate $\Psi(\boldsymbol{\eta})$. Note that $\boldsymbol{\theta}$ are sometimes called *expnential parameters*, and $\boldsymbol{\eta}$ are also known as *mean parameters*.

Note that the class of all exponential family distributions is *not* closed w.r.t. marginalization. For example, if $P(\boldsymbol{x}|\boldsymbol{\theta})$ is a joint of a continuous Gaussian and a discrete multinomial variable, marginalizing over the latter results in a mixture of Gaussians which is in general not in an exponential family. A number of exponential subfamilies such as the (multivariate) Gaussian or multinomial ones *are* however closed under marginalization.

**Lemma 2 (Product of Exponential Distributions)** *A product of densities from $\mathcal{F}$ is an unnormalised member of $\mathcal{F}$:*

$$\prod_{j=1}^{m} P(\boldsymbol{x}|\boldsymbol{\theta}_j) = P\left(\boldsymbol{x} \middle| \sum_{j=1}^{m} \boldsymbol{\theta}_j\right) \exp\left(\Phi\left(\sum_{j=1}^{m} \boldsymbol{\theta}_j\right) - \sum_{j=1}^{m} \Phi(\boldsymbol{\theta}_j)\right),$$

*given that $\sum_j \boldsymbol{\theta}_j$ lies in $\Theta$.*

If $P(\boldsymbol{x}|\boldsymbol{\theta}) \in \mathcal{F}$ for an exponential family $\mathcal{F}$, then the convexity of $\Phi$ implies that $\log P(\boldsymbol{x}|\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$ (strictly so for a minimal representation). In other words, exponential family densities are *log-concave*:

$$P(\boldsymbol{x}|\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\theta}') \geq P(\boldsymbol{x}|\boldsymbol{\theta})^{\lambda} P(\boldsymbol{x}|\boldsymbol{\theta}')^{1-\lambda}, \quad \lambda \in [0,1].$$

Given a positive function $f(\boldsymbol{x})$, we can induce a *tilted exponential family* from $\mathcal{F}$ by modifying the base measure $\mu$ and recomputing the log partition function.

**Definition 2 (Tilted Exponential Family)** *If $\mathcal{F}$ is an exponential family with natural parameter $\boldsymbol{\theta}$ and $f(\boldsymbol{x})$ is a positive function such that*

$$\Phi_f(\boldsymbol{\theta}) = \log \mathrm{E}_{\boldsymbol{\theta}}\left[f(\boldsymbol{x})\right] + \Phi(\boldsymbol{\theta})$$

*exists for every $\boldsymbol{\theta}$, then the* tilted exponential family $\mathcal{F}_f$ *induced by $f(\boldsymbol{x})$ from $\mathcal{F}$ contains the densities*

$$P_f(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x}) - \Phi_f(\boldsymbol{\theta})\right)$$

*w.r.t. the modified base measure $d\mu_f(\boldsymbol{x}) = f(\boldsymbol{x})d\mu(\boldsymbol{x})$. $\mathcal{F}_f$ has the same natural parameter space $\Theta$ than $\mathcal{F}$.*

Since $\mathcal{F}_f$ is a proper exponential family, the moment parameter of $P_f(\boldsymbol{x}|\boldsymbol{\theta})$ can be computed as derivatives of $\Phi_f(\boldsymbol{\theta})$, i.e.

$$\mathrm{E}_{P_f(\cdot|\boldsymbol{\theta})}[\boldsymbol{\phi}(\boldsymbol{x})] = \nabla_{\boldsymbol{\theta}} \log \mathrm{E}_{\boldsymbol{\theta}}\left[f(\boldsymbol{x})\right] + \boldsymbol{\eta}. \tag{1}$$

If we "update" a distribution from $\mathcal{F}$ by multiplying with a positive factor and renormalising, we will end up in $\mathcal{F}$ iff the update factor has the structure of a ratio of members of $\mathcal{F}$.

**Definition 3 (Unnormalised Exponential Family)** *If $\mathcal{F}$ is an exponential family with natural parameter $\boldsymbol{\theta} \in \Theta$, the set of functions*

$$P^U(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})\right), \quad \boldsymbol{\theta} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2,\ \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta,$$

*is referred to as* unnormalised exponential family $\mathcal{F}^U$ *associated with $\mathcal{F}$.*

Note that members of $\mathcal{F}^U$ are in general no probability densities, and some of them may not be normalisable at all. If $P(\boldsymbol{x}|\boldsymbol{\theta}) \in \mathcal{F}$, $P^U(\boldsymbol{x}|\tilde{\boldsymbol{\theta}}) \in \mathcal{F}^U$, then $P(\boldsymbol{x}|\boldsymbol{\theta})P^U(\boldsymbol{x}|\tilde{\boldsymbol{\theta}})$ is proportional to a member of $\mathcal{F}$ (namely, to $P(\boldsymbol{x}|\boldsymbol{\theta}+\tilde{\boldsymbol{\theta}})$) iff $\boldsymbol{\theta}+\tilde{\boldsymbol{\theta}} \in \Theta$. Note also that $1 \equiv P^U(\boldsymbol{x}|\boldsymbol{0}) \in \mathcal{F}^U$.

## 2 Expectation Propagation

*Expectation Propagation* (EP) [11] provides a general-purpose framework for approximating posterior beliefs by exponential family distributions. The Gaussian special case has been proposed by Opper and Winther [16] as *ADATAP*, see also Section 6. General papers on EP include [10, 9]. The tutorial description of EP here does not include any new material, but tries to simplify earlier expositions by consequently working with the exponential family framework introduced in Section 1.

Suppose we are given some statistical model with observables $S$ and latent variables $\boldsymbol{u}$, and with a prior distribution[1] $P^{(0)}(\boldsymbol{u})$ from an exponential family $\mathcal{F}$. We do not assume here

[1] In some applications of EP, the tractable $P^{(0)}$ is the likelihood, so our nomenclature could be misleading.

that the parameterization of $\mathcal{F}$ is minimal, but allow for overcomplete parameterizations as well. The likelihood function $P(S|\boldsymbol{u})$ often factors in a particular way,

$$P(S|\boldsymbol{u}) = \prod_{i=1}^{n} t_i(\boldsymbol{u}),$$

for example in the case of i.i.d. data $S$ or Bayesian networks. We refer to the $t_i(\boldsymbol{u})$ as *sites*. If the true posterior

$$P(\boldsymbol{u}|S) \propto P^{(0)}(\boldsymbol{u}) \prod_{i=1}^{n} t_i(\boldsymbol{u}) \tag{2}$$

is analytically intractable, we may approximate it by a distribution $Q(\boldsymbol{u})$ from $\mathcal{F}$:

$$Q(\boldsymbol{u}) = Q(\boldsymbol{u}|\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{u}) - \Phi(\boldsymbol{\theta})\right),\ \boldsymbol{\theta} \in \Theta.$$

An often tractable way for choosing $Q$ is to start from $Q(\boldsymbol{u}) = P^{(0)}(\boldsymbol{u})$ and incorporate the sites $t_i(\boldsymbol{u})$ one after the other, following some sequential ordering. Namely, in order to incorporate $t_i(\boldsymbol{u})$, first compute the true Bayesian update

$$\hat{P}(\boldsymbol{u}) = Z_i^{-1} Q(\boldsymbol{u}) t_i(\boldsymbol{u}), \quad Z_i = \mathrm{E}_{\boldsymbol{u}\sim Q}[t_i(\boldsymbol{u})].$$

$\hat{P}$ lies in the tilted exponential family $\mathcal{F}_{t_i}$ which is different from $\mathcal{F}$ in general. In order to approximate $\hat{P}$, we choose $Q^{new}$ to have the same $\mathcal{F}$-moments than $\hat{P}$:

$$Q^{new}(\boldsymbol{u}) = \operatorname*{argmin}_{\tilde{Q}\in\mathcal{F}} \mathrm{D}\left[\hat{P}(\boldsymbol{u}) \,\|\, \tilde{Q}(\boldsymbol{u})\right] \quad \Leftrightarrow \quad \boldsymbol{\eta}^{new} = \mathrm{E}_{\hat{P}}[\boldsymbol{\phi}(\boldsymbol{u})].$$

We refer to this process as *inclusion* of site $t_i(\boldsymbol{u})$ into the belief $Q$. An inclusion is different from a true Bayesian update, since the full updated belief $\hat{P}$ is "collapsed" to $Q^{new}$, a member of $\mathcal{F}$, which allows inclusions to be chained. The moments of $\hat{P}$ can be computed via Eq. 1 which is often feasible (or amenable to numerical approximations) even though the moments of the full posterior $P(\boldsymbol{u}|S)$ remain intractable. This simple idea has been used extensively, for example in the context of Bayesian on-line learning [15] or switching linear dynamical systems [1, 4, 14], see [11] for more exhaustive references. It is known as *assumed density filtering (ADF)*. Nevertheless, each site may be included only once, and in the context of dynamical systems we are restricted to updates in one direction along the backbone chain (filtering), while bidirectional smoothing would maybe improve the approximation.

In [11], a new view on ADF is established which allows these shortcomings to be removed. The process of including $t_i(\boldsymbol{u})$ results in $Q(\boldsymbol{u})$ being replaced by $Q^{new}(\boldsymbol{u})$, which can also be seen as multiplying $Q(\boldsymbol{u})$ by the ratio $\tilde{t}_i(\boldsymbol{u}) \propto Q^{new}(\boldsymbol{u})/Q(\boldsymbol{u})$ and renormalising. This operation becomes particularly simple in the natural parameters: $\boldsymbol{\theta}^{new} = \boldsymbol{\theta} + (\boldsymbol{\theta}^{new} - \boldsymbol{\theta})$. The ratio $\tilde{t}_i(\boldsymbol{u})$ is a member of the unnormalised exponential family $\mathcal{F}^U$ associated with $\mathcal{F}$ (Definition 3): it has a form very similar to $Q$ and $Q^{new}$, but $\boldsymbol{\theta}^{new} - \boldsymbol{\theta}$ will not in general lie in the natural parameter space $\Theta$ of $\mathcal{F}$.[2] This view motivates *representing* $Q$ as

$$Q(\boldsymbol{u}) \propto P^{(0)}(\boldsymbol{u}) \prod_{i=1}^{n} \tilde{t}_i(\boldsymbol{u}), \tag{3}$$

[2]For example, if $\mathcal{F}$ is the family of Gaussians, then $\tilde{t}_i$ may correspond to a "Gaussian with negative variance".

where the $\tilde{t}_i(\boldsymbol{u}) = \tilde{t}_i(\boldsymbol{u}|\boldsymbol{\theta}^{(i)}) \in \mathcal{F}^U$ are referred to as *site approximations*, and their natural parameters $\boldsymbol{\theta}^{(i)}$ as *site parameters*. If $\boldsymbol{\theta}^{(0)}$ denotes the parameters of $P^{(0)}(\boldsymbol{u})$, then

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)} + \sum_{i=1}^{n} \boldsymbol{\theta}^{(i)}.$$

Note that we allow $\boldsymbol{\theta}^{(i)} = \mathbf{0}$, $\tilde{t}_i(\boldsymbol{u}) \equiv 1$, in fact in the beginning all site approximations are constant, leading to $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$, i.e. $Q(\boldsymbol{u}) = P^{(0)}(\boldsymbol{u})$. An *ADF update (inclusion)* w.r.t. site $t_i$ can now be seen as follows (note that $\tilde{t}_i \equiv 1$):

**Definition 4 (ADF Update, Inclusion)**

1. *Compute moments of* $\hat{P}(\boldsymbol{u}) \propto Q(\boldsymbol{u})t_i(\boldsymbol{u})$ *and pick* $Q^{new}(\boldsymbol{u}) \in \mathcal{F}$ *with these moments.*
2. *In order to replace* $Q(\boldsymbol{u})$ *by* $Q^{new}(\boldsymbol{u})$*, we replace* $\tilde{t}_i(\boldsymbol{u}) \equiv 1$ *by* $\tilde{t}_i^{new}(\boldsymbol{u}) \propto Q^{new}(\boldsymbol{u})/Q(\boldsymbol{u})$*.*

From this viewpoint, it becomes clear how ADF can be generalised to a full-fledged iterative approximation scheme, allowing for multiple iterations over the sites. An *EP update (inclusion-deletion)* w.r.t. site $t_i$ works as follows:

**Definition 5 (EP Update, Inclusion-Deletion)**

1. *Delete the site approximation* $\tilde{t}_i(\boldsymbol{u})$ *from* $Q(\boldsymbol{u})$ *by renormalising* $Q(\boldsymbol{u})/\tilde{t}_i(\boldsymbol{u})$*, obtaining*

   $$Q^{\setminus i}(\boldsymbol{u}) \propto P^{(0)}(\boldsymbol{u}) \prod_{j \neq i} \tilde{t}_j(\boldsymbol{u}).$$

   *In natural parameters:* $\boldsymbol{\theta}^{\setminus i} = \boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}$*.*
2. *Let* $\hat{P}(\boldsymbol{u}) = Z_i^{-1} t_i(\boldsymbol{u}) Q^{\setminus i}(\boldsymbol{u})$ *and compute*

   $$\boldsymbol{\eta}^{new} = \mathrm{E}_{\hat{P}}[\boldsymbol{\phi}(\boldsymbol{u})] = \nabla_{\boldsymbol{\theta}^{\setminus i}} \log Z_i + \boldsymbol{\eta}^{\setminus i}, \quad Z_i = \mathrm{E}_{\boldsymbol{\theta}^{\setminus i}}[t_i(\boldsymbol{u})],$$

   *and pick* $Q^{new} \in \mathcal{F}$ *with these moments.*
3. *Replace* $\tilde{t}_i$ *by* $\tilde{t}_i^{new}(\boldsymbol{u}) \propto Q^{new}(\boldsymbol{u})/Q^{\setminus i}(\boldsymbol{u})$*.*
   *In natural parameters:* $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{new} - \boldsymbol{\theta}^{\setminus i}$*.*

In line with [16], we will refer to $Q^{\setminus i}$ as *cavity distribution.*

On networks with discrete nodes or Gaussian Markov random fields, EP can be seen as generalisation of loopy belief propagation, allowing for a more flexible choice of approximating structure and distribution family (see [11] and [26], Chap. 6). The algorithm does not always converge, but if it does, the fixed point must be a saddle points of an approximation to the free energy which is a generalisation of the Bethe free energy [9, 7, 26]. Double-loop concave-convex algorithms can be applied in order to ensure convergence [7]. Problems of

convergence can sometimes be overcome by using "damped" updates: instead of $\boldsymbol{\theta} \to \boldsymbol{\theta}^{new}$, we update $\boldsymbol{\theta}$ to a convex combination of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{new}$. Also, updates which lead to $\boldsymbol{\theta}^{new}$ outside of $\Theta$ (or very close to its boundary) should be rejected. In practice, it is important to address the issue of numerical stability: the conversion between natural and moment parameters is typically not a stable operation.[3] If possible, an implementation should remain in the moment parameters entirely and fold conversions with update operations into a single mapping which can then be stabilised. This is of course less generic than the presentation above, and it is also not clear how to do damped updates (which are convex combinations in the natural parameters) in this way.

It is important to note that EP is not simply a local approximation of the sites $t_i(\boldsymbol{u})$ by corresponding $\tilde{t}_i(\boldsymbol{u})$, but a global fit by $Q \in \mathcal{F}$ to the distribution obtained by replacing $\tilde{t}_i(\boldsymbol{u})$ by $t_i(\boldsymbol{u})$ in the current belief $Q(\boldsymbol{u})$. In fact, the sites may not even be continuous functions of $\boldsymbol{u}$. EP has been applied for approximate inference in two very different regimes: sparsely connected Bayesian or Markov networks and models with fully connected Gaussian prior $P^{(0)}(\boldsymbol{u})$. In the former regime, every single $t_i(\boldsymbol{u})$ depends on a small number of components of $\boldsymbol{u}$ only, e.g. in the Markov network case on small cliques of the underlying graph. By choosing a special structure of the approximating distribution $Q$, based on a tractable subgraph of the decomposable extension of the model graph, *and* requiring that the prior $P^{(0)}(\boldsymbol{u})$ follows this structure, one can run EP as a message-passing scheme, updating the parameters of $Q$ and certain small extensions thereof. This notion is developed and generalised in [26], Chap. 6 (for discrete variables), and in [9] (does not address the issue of how to represent $Q$). In the second regime, $\mathcal{F}$ is the family of Gaussians, and $P^{(0)}(\boldsymbol{u})$ is typically densely connected, while the likelihood factors $t_i(\boldsymbol{u})$ are local again. The special case for a completely factorised likelihood $P(S|\boldsymbol{u})$ has been given in [16, 5].

### 2.1 Marginal Likelihood Approximation

The marginal likelihood

$$P(S) = \int \prod_{i=1}^{n} t_i(\boldsymbol{u}) P^{(0)}(\boldsymbol{u})\, d\mu(\boldsymbol{u})$$

can be approximated within EP as well which allows to optimize over free hyperparameters. As long as EP is used only to approximate the posterior $P(\boldsymbol{u}|S)$, it does not matter how the site approximations $\tilde{t}_i$ are normalized, but now we make the normalization explicit by using the site approximations

$$C_i \tilde{t}_i(\boldsymbol{u}), \quad \tilde{t}_i(\boldsymbol{u}) = \exp\left(\boldsymbol{\theta}^{(i)T} \boldsymbol{\phi}(\boldsymbol{u})\right).$$

The idea is to match the normalization constants in the same way as the moments making use of the cavity distributions. Let

$$Z_i = \mathrm{E}_{\boldsymbol{\theta}^{\setminus i}}\left[t_i(\boldsymbol{u})\right], \quad \tilde{Z}_i = \mathrm{E}_{\boldsymbol{\theta}^{\setminus i}}\left[\tilde{t}_i(\boldsymbol{u})\right] = \exp\left(\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}^{\setminus i})\right).$$

We require the cavity expectations of $t_i$ and $C_i \tilde{t}_i$ to be the same for all $i$, which means that $Z_i = C_i \tilde{Z}_i$ or

$$\log C_i = \log Z_i - \Phi(\boldsymbol{\theta}) + \Phi(\boldsymbol{\theta}^{\setminus i}).$$

[3] For example, for the Gaussian family we have to invert a matrix.

An approximation to $\log P(S)$ is obtained by replacing the sites $t_i$ by their approximations $C_i\tilde{t}_i$:

$$\begin{aligned} L &= \log \int \exp\left(\sum_{i=1}^n \log C_i + \boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{u}) - \Phi(\boldsymbol{\theta}^{(0)})\right) d\mu(\boldsymbol{u}) \\ &= \sum_{i=1}^n \log C_i + \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}^{(0)}). \end{aligned}$$

In order to maximize $L$ we require its gradient w.r.t. hyperparameters. Note that $L$ depends on these directly as well as through the site parameters $\boldsymbol{\theta}^{(i)}$. We assume that there exists an open region around the current hyperparameters within which the $\boldsymbol{\theta}^{(i)}$ are continuously differentiable[4]. The computation can be simplified greatly by making use of the fixed point conditions which hold at convergence of EP. After an update of site $i$ we have $\mathrm{E}_{\hat{P}_i}[\boldsymbol{\phi}(\boldsymbol{u})] = \mathrm{E}_Q[\boldsymbol{\phi}(\boldsymbol{u})]$, where $\hat{P}_i(\boldsymbol{u}) \propto t_i(\boldsymbol{u})Q^{\backslash i}(\boldsymbol{u})$ and $Q(\boldsymbol{u}) \propto \tilde{t}_i(\boldsymbol{u})Q^{\backslash i}(\boldsymbol{u})$. Using our remarks on tilted exponential families in Section 1 we have

$$\nabla_{\boldsymbol{\theta}^{(0)}} \log Z_i = \left(\frac{\partial\boldsymbol{\theta}^{\backslash i}}{\partial\boldsymbol{\theta}^{(0)}}\right)^T \left(\mathrm{E}_{\hat{P}_i}[\boldsymbol{\phi}(\boldsymbol{u})] - \boldsymbol{\eta}^{\backslash i}\right) = \left(\frac{\partial\boldsymbol{\theta}^{\backslash i}}{\partial\boldsymbol{\theta}^{(0)}}\right)^T \left(\boldsymbol{\eta} - \boldsymbol{\eta}^{\backslash i}\right),$$

because $\mathrm{E}_{\hat{P}_i}[\boldsymbol{\phi}(\boldsymbol{u})] = \mathrm{E}_Q[\boldsymbol{\phi}(\boldsymbol{u})] = \boldsymbol{\eta}$. Furthermore,

$$\nabla_{\boldsymbol{\theta}^{(0)}} \log \tilde{Z}_i = \left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\theta}^{(0)}}\right)^T \boldsymbol{\eta} - \left(\frac{\partial\boldsymbol{\theta}^{\backslash i}}{\partial\boldsymbol{\theta}^{(0)}}\right)^T \boldsymbol{\eta}^{\backslash i},$$

so that

$$\nabla_{\boldsymbol{\theta}^{(0)}} \sum_{i=1}^n \log C_i = -\boldsymbol{J}^T\boldsymbol{\eta}, \quad \boldsymbol{J} = \frac{\partial\boldsymbol{\theta}^{\backslash 0}}{\partial\boldsymbol{\theta}^{(0)}}$$

where $\boldsymbol{\theta}^{\backslash 0} = \sum_{i=1}^n \boldsymbol{\theta}^{(i)}$. Also,

$$\nabla_{\boldsymbol{\theta}^{(0)}} \left(\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}^{(0)})\right) = \left(\boldsymbol{I} + \boldsymbol{J}^T\right)\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)},$$

since $(\partial\boldsymbol{\theta})/(\partial\boldsymbol{\theta}^{(0)}) = \boldsymbol{I} + \boldsymbol{J}$. The Jacobian $\boldsymbol{J}$ (which could not be computed in general) drops out and

$$\nabla_{\boldsymbol{\theta}^{(0)}} L = \mathrm{E}_Q[\boldsymbol{\phi}(\boldsymbol{u})] - \mathrm{E}_P[\boldsymbol{\phi}(\boldsymbol{u})] = \boldsymbol{\eta} - \boldsymbol{\eta}^{(0)} = \mathrm{E}_Q[\nabla_{\boldsymbol{\theta}^{(0)}} \log P^{(0)}(\boldsymbol{u})].$$

In other words, although the dependence of $L$ on $\boldsymbol{\theta}^{(0)}$ is direct as well as through the site parameters $\boldsymbol{\theta}^{(i)}$, the second one can be ignored for the purpose of the gradient computation, as long as the EP fixed point conditions hold. For the gradient computation, the site parameters can be considered fixed.

It is important to note the following consistency property. We may ask what the true gradient $\nabla_{\boldsymbol{\theta}^{(0)}} \log P(D)$ is:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}^{(0)}} \log P(D) &= P(D)^{-1} \int \prod_i t_i(\boldsymbol{u}) P^{(0)}(\boldsymbol{u}) \nabla_{\boldsymbol{\theta}^{(0)}} \left(\boldsymbol{\theta}^{(0)T}\boldsymbol{\phi}(\boldsymbol{u}) - \Phi(\boldsymbol{\theta}^{(0)})\right) d\boldsymbol{u} \\ &= \mathrm{E}_{P(\boldsymbol{u}|D)}[\boldsymbol{\phi}(\boldsymbol{u})] - \mathrm{E}_P[\boldsymbol{\phi}(\boldsymbol{u})]. \end{aligned}$$

[4]We do not know whether this is guaranteed in general. The problem is that for EP, the site parameters are not unique solutions of some smooth optimization problem.

Therefore, the two approaches of either approximating $\log P(D)$ by $L$ and deriving the gradient of $L$, or of approximating the gradient of $\log P(D)$ (by replacing $P(\cdot|D)$ by $Q$) do lead to the same result.

If $\alpha$ is a parameter of the site $t_j(\boldsymbol{u})$, the dependence of $L$ on $\alpha$ is direct or through $\boldsymbol{\theta}^{\backslash 0}$ (we assume that $\partial\boldsymbol{\theta}^{(0)}/\partial\alpha = \mathbf{0}$. We show that the dependence through $\boldsymbol{\theta}^{\backslash 0}$ can be ignored. If we ignore the direct dependence,

$$\frac{\partial}{\partial\alpha}\log Z_i = \left(\boldsymbol{\eta} - \boldsymbol{\eta}^{\backslash i}\right)^T \frac{\partial\boldsymbol{\theta}^{\backslash i}}{\partial\alpha}, \quad \frac{\partial}{\partial\alpha}\log\tilde{Z}_i = \boldsymbol{\eta}^T\frac{\partial\boldsymbol{\theta}}{\partial\alpha} - \boldsymbol{\eta}^{\backslash iT}\frac{\partial\boldsymbol{\theta}^{\backslash i}}{\partial\alpha},$$

therefore

$$\frac{\partial}{\partial\alpha}\sum_{i=1}^{n}\log C_i = -\boldsymbol{\eta}^T\frac{\partial\boldsymbol{\theta}^{\backslash 0}}{\partial\alpha}.$$

Also,

$$\frac{\partial}{\partial\alpha}\left(\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}^{(0)})\right) = \boldsymbol{\eta}^T\frac{\partial\boldsymbol{\theta}}{\partial\alpha} = \boldsymbol{\eta}^T\frac{\partial\boldsymbol{\theta}^{\backslash 0}}{\partial\alpha},$$

because $\boldsymbol{\theta}^{(0)}$ does not depend on $\alpha$. Therefore, only the direct dependence counts:

$$\frac{\partial L}{\partial\alpha} = \frac{\partial}{\partial\alpha}\log C_j = \frac{\partial}{\partial\alpha}\log Z_j = \mathrm{E}_{\hat{P}_j}\left[\frac{\partial}{\partial\alpha}\log t_j(\boldsymbol{u})\right]$$

where $\hat{P}_j(\boldsymbol{u}) \propto t_j(\boldsymbol{u})Q^{\backslash j}(\boldsymbol{u})$.

It is also possible in principle to compute higher-order derivatives of $L$. The results just shown can be summed up as follows. Let $\alpha$ be some parameter of the true joint distribution $P^{(0)}(\boldsymbol{u})\prod_i t_i(\boldsymbol{u})$ (need not be normalized, see Section 2.2), so that the log joint is differentiable w.r.t. $\alpha$. Then,

$$\frac{\partial L}{\partial\alpha} = \mathrm{E}_{\boldsymbol{\theta}}[g(\boldsymbol{u})], \quad g(\boldsymbol{u}) := \frac{d}{d\alpha}\log P^{(0)}(\boldsymbol{u})\prod_i t_i(\boldsymbol{u}).$$

This is consistent, in that

$$\frac{\partial\log P(D)}{\partial\alpha} = \mathrm{E}_{P(\boldsymbol{u}|D)}[g(\boldsymbol{u})].$$

For the second derivative, let $H(\boldsymbol{u}) := (d^2/d\alpha^2)\log P^{(0)}(\boldsymbol{u})\prod_i t_i(\boldsymbol{u})$. It is easy to see that

$$\frac{\partial^2\log P(D)}{\partial\alpha^2} = \mathrm{E}_{P(\boldsymbol{u}|D)}[H(\boldsymbol{u})] + \mathrm{Cov}_{P(\boldsymbol{u}|D)}[g(\boldsymbol{u})].$$

Now, $L = L(\alpha, \boldsymbol{\theta}^{\backslash 0})$. We use subindices to denote direct derivatives of $L$. The EP fixed point conditions translate to

$$L_{\boldsymbol{\theta}^{\backslash 0}} = \mathbf{0}, \quad \boldsymbol{\theta}^{\backslash 0} = \boldsymbol{\theta}^{\backslash 0}(\alpha).$$

Note that in this section, we assume that $\alpha \to \boldsymbol{\theta}^{\backslash 0}$ is unique in an environment of the current $\alpha$. Since $\boldsymbol{\theta}^{\backslash 0}$ is not the maximum point of a concave criterion, this condition may not hold for all $\alpha$.

The second derivative of $L$ is *not* consistent in the sense used above for the first one. We have that

$$L_{\alpha,\alpha} = \mathrm{E}_{\boldsymbol{\theta}}[H(\boldsymbol{u})] + \mathrm{Cov}_{\boldsymbol{\theta}}[g(\boldsymbol{u})],$$

but then

$$\frac{\partial^2 L}{\partial \alpha^2} = L_{\alpha,\alpha} + L^T_{\boldsymbol{\theta}^{\backslash 0},\alpha}(d\boldsymbol{\theta}^{\backslash 0})/(d\alpha).$$

Since the fixed point condition holds in an environment around $\alpha$, we have that

$$\mathbf{0} = \frac{\partial}{\partial \alpha} L_{\boldsymbol{\theta}^{\backslash 0}} = L_{\boldsymbol{\theta}^{\backslash 0},\alpha} + L_{\boldsymbol{\theta}^{\backslash 0},\boldsymbol{\theta}^{\backslash 0}}(d\boldsymbol{\theta}^{\backslash 0})/(d\alpha).$$

Assuming that $L_{\boldsymbol{\theta}^{\backslash 0},\boldsymbol{\theta}^{\backslash 0}}$ is invertible, we have that

$$\frac{\partial^2 L}{\partial \alpha^2} = L_{\alpha,\alpha} - L^T_{\boldsymbol{\theta}^{\backslash 0},\alpha} L^{-1}_{\boldsymbol{\theta}^{\backslash 0},\boldsymbol{\theta}^{\backslash 0}} L_{\boldsymbol{\theta}^{\backslash 0},\alpha}.$$

Thus, in general, $(\partial^2 L)/(\partial \alpha^2) \neq L_{\alpha,\alpha}$, and the consistency valid for the gradient does not hold for higher-order derivatives. However, they can still be computed exactly in principle.

## 2.2 Unnormalized Prior Measure

We assumed above that $P^{(0)}(\boldsymbol{u})$ is a proper exponential family distribution. However, in some situations, $P^{(0)}(\boldsymbol{u})$ will rather be a member of an *unnormalized* family: $P^{(0)}(\boldsymbol{u}) = \exp(\boldsymbol{\theta}^{(0)T}\boldsymbol{\phi}(\boldsymbol{u}) + C)$, where $\Phi[\boldsymbol{\theta}^{(0)}]$ need not be finite. For example, EP can be applied to models where $\prod_i t_i(\boldsymbol{u})$ corresponds to the prior distribution, and $P^{(0)}(\boldsymbol{u})$ is a likelihood function.

Running EP for such a case is slightly more challenging. For example, we cannot start from $\boldsymbol{t} = \boldsymbol{\theta}^{(0)}$, but have to initialize the $\boldsymbol{\theta}^{(i)}$ to non-zero values such that $\Phi[\boldsymbol{\theta}]$ is finite. Furthermore, computing EP updates can be numerically less stable. One useful remedy in such situations is to do *fractional* updates [12], see Section 2.3.

The marginal likelihood approximation works in the same way as described in Section 2.1. First, we have $L = \sum_i \log C_i + \Phi[\boldsymbol{\theta}] + C$ in this case. Furthermore, the arguments surrounding the gradient computation remain valid, so that $\nabla_{\boldsymbol{\theta}^{(0)}} L = \boldsymbol{\eta} + \nabla_{\boldsymbol{\theta}^{(0)}} C$. If $\alpha$ is a parameter of $t_j$, the result of Section 2.1 remains valid if $C$ does not depend on $\alpha$.

## 2.3 Fractional EP

Fractional EP (or Power EP) is introduced in [12] as an extension to EP. For a fraction $\eta \in (0, 1]$, we redefine $Q^{\backslash i} \propto Q\tilde{t}_i^{-\eta}$ and $\hat{P} \propto Q^{\backslash i} t_i^{\eta}$, but otherwise proceed as before. Standard EP is obtained by setting $\eta = 1$.

Minka [12] gives a different interpretation in terms of $\alpha$-divergences. Recall that in standard EP, $\hat{P}$ is projected onto the new $Q^{new}$ by matching $\mathcal{F}$-moments, *i.e.* by minimizing $\mathrm{D}[\hat{P} \,\|\, \cdot]$. Fractional EP can be seen as keeping the original definition of $\hat{P}$, but instead projecting back into $\mathcal{F}$ w.r.t. an $\alpha$-divergence in place of the "exclusive" relative entropy. We will not follow this different view in the sequel.

An important principal advantage of fractional EP with $\eta < 1$ versus standard EP has been mentioned in [25]. In some practically important cases, standard EP seems inherently numerically unstable to run, but fractional EP with $\eta < 1$ works fine. In the setting treated in [25], the prior measure $P^{(0)}$ is unnormalized (in the Gaussian case, $\boldsymbol{\Pi}^{(0)}$ is (numerically)

singular, see Section 5). Moreover, the site $t_i(\boldsymbol{u})$ depends on $u_i$ only, as does $\tilde{t}_i(\boldsymbol{u})$ (see Section 3). In this case, $Q^{\setminus i}(u_i)$ has extremely large variance, and the computation of moments of $\hat{P} \propto Q^{\setminus i} t_i$ is numerically unstable. In this setup, it is principally $\tilde{t}_i$ which keeps the variance of $Q$ along $u_i$ in a sensible range. Standard EP requires to remove $\tilde{t}_i$ temporarily, leading to an almost ill-defined $Q^{\setminus i}$. If $t_i$ is not Gaussian, there seems no simple way to obtain the moments of $\hat{P}$ without the intermediate $Q^{\setminus i}$.

In contrast, in fractional EP we have that $Q^{\setminus i} \propto Q\tilde{t}_i^{-\eta}$, so a fraction $\tilde{t}_i^{1-\eta}$ still remains in $Q^{\setminus i}$, keeping the variance small. Moreover, $\hat{P} \propto Q^{\setminus i} t_i^{\eta}$, so the wider $t_i^{\eta}$ is factored in instead of $t_i$ as in standard EP. Together, this leads to a stabilization of the EP update. In the application considered in [25], even a careful implementation of standard EP diverges in an erratic way, while fractional EP converges quickly to a useful approximate posterior.

Moreover, the log marginal likelihood approximation $L$ of Section 2.1 can easily be adapted to the fractional case. In general,

$$\boldsymbol{\theta}^{\setminus i} = \boldsymbol{\theta} - \eta \boldsymbol{\theta}^{(i)}, \quad Z_i = \mathrm{E}_{\boldsymbol{\theta}^{\setminus i}}\left[t_i(\boldsymbol{u})^{\eta}\right], \quad \tilde{Z}_i = \mathrm{E}_{\boldsymbol{\theta}^{\setminus i}}\left[\tilde{t}_i(\boldsymbol{u})^{\eta}\right].$$

Matching local normalization constants in the fractional case leads to $\log C_i = \eta^{-1}(\log Z_i - \log \tilde{Z}_i)$. With these more general definitions, the form of $L$ given in Section 2.1 remains valid for $\eta < 1$. Importantly, the analytical gradient computation remains the same as well. In fact, the proof given in Section 2.1 applies essentially without modifications.

Finally, note that another idea of making EP run smoother on hard problems is *damping* [11]. There, the full standard EP update is computed, but the site parameters $\boldsymbol{\theta}^{(i)}$ are updated to a convex combination of old and proposed new values. This addresses a quite contrary problem to ours here. Damping is useful if single EP update computations are stable, but lead to an improper new posterior, or the propagation of the updated information fails. If EP is viewed as finding a saddle point of a free energy approximation (see Section 6), damping can be understood as a step-size rule within this process. The stability problem mentioned above is not solved by damping, since proposed new values for the site parameters cannot even be computed.

## 3 Locality Property. Feasibility of EP

The general EP method discussed above can be used with any exponential family $\mathcal{F}$. However, only under further restrictions do we actually end up with a feasible method, given that exact inference is intractable. In this section, we discuss these additional properties of $\mathcal{F}$.

In our setting, intractability of inference means that the moments $\mathrm{E}_{P(\boldsymbol{u}|D)}[\boldsymbol{\phi}(\boldsymbol{u})]$ cannot be computed directly. EP will only be tractable if we can compute these moments for the cavity distributions $\hat{P}(\boldsymbol{u}) \propto Q^{\setminus i}(\boldsymbol{u}) t_i(\boldsymbol{u})$, or more specifically if we can find $\boldsymbol{\theta}^{new}$ for $Q^{new} \in \mathcal{F}$ with these moments. In general, this places two restrictions on $\mathcal{F}$ as well as the $t_i$. First, the $t_i$ must be "local", in the sense that although $\hat{P} \notin \mathcal{F}$, we can compute $\mathrm{E}_{\hat{P}}[\boldsymbol{\phi}(\boldsymbol{u})]$ efficiently, given that we know $\mathrm{E}_Q[\boldsymbol{\phi}(\boldsymbol{u})]$, for example requiring an integration over few variables only. Second, $\mathcal{F}$ must be a tractable family itself: marginal inference within $\mathcal{F}$ must be tractable. Given these two properties, we can perform EP updates based on $Q$ marginals only, which requires local moment matching only. These updates affect $Q$ in the same way as local

evidence, and the feasibility of $\mathcal{F}$ allows us to propagate these changes towards marginals for the next EP update.

It is not insightful to formalize these requirements further, and we prefer to give some examples. First, let $\mathcal{F}$ be the family of multivariate Gaussians. $\mathcal{F}$ is closed under marginalization, and marginal inference is clearly tractable. Furthermore, suppose that $t_i(\boldsymbol{u})$ is a function of $v_i = \boldsymbol{c}_i^T \boldsymbol{u}$ only. Then, it is easy to see that the EP update can be done using the marginal $Q(v_i)$ only, and it is sufficient to use $\tilde{t}_i(\boldsymbol{u}) = \tilde{t}_i(v_i)$, which is parameterized by two scalars only. To see this, assume w.l.o.g. that $\boldsymbol{c}_i = \boldsymbol{\delta}_1$, so that $v_i = u_1$. This is achieved by extending $\boldsymbol{c}_i$ to a nonsingular transformation, noting that $\mathcal{F}$ is closed under such as well. Now, $\hat{P}(\boldsymbol{u}) = \hat{P}(u_1) Q^{\backslash i}(\boldsymbol{u}_{\backslash 1} | u_1)$ and

$$\begin{aligned}\mathrm{D}[\hat{P}(\boldsymbol{u}) \,\|\, Q^{new}(\boldsymbol{u})] = \mathrm{D}[\hat{P}(u_1) \,\|\, Q^{new}(u_1)] \\ + \mathrm{E}_{u_1 \sim \hat{P}} \left[ \mathrm{D}[Q^{\backslash i}(\boldsymbol{u}_{\backslash 1} | u_1) \,\|\, Q^{new}(\boldsymbol{u}_{\backslash 1} | u_1)] \right].\end{aligned}$$

In order to minimise this expression, we set $Q^{new}(\boldsymbol{u}_{\backslash 1} | u_1) = Q^{\backslash i}(\boldsymbol{u}_{\backslash 1} | u_1)$ and match moments between the marginals $\hat{P}(u_1)$ and $Q^{new}(u_1)$. It follows that $\tilde{t}_i(\boldsymbol{u}) = \tilde{t}_i(u_1)$, having the form of $Q^{new}/Q$, i.e. the site approximations inherit the locality of the corresponding sites and can be parameterised economically, in the sense that many of the components in $\boldsymbol{\theta}^{(i)}$ are clamped to zero. Furthermore, since $Q(\boldsymbol{u}) \propto Q^{\backslash i}(\boldsymbol{u}) \tilde{t}_i(u_1)$, we see that $Q(u_1) \propto Q^{\backslash i}(u_1) \tilde{t}_i(u_1)$, so that in order to update the site approximation $\tilde{t}_i$, we only need to access the marginal $Q(u_1)$. Note however that a change of $\tilde{t}_i(u_1)$ in general affects all marginals of $Q(\boldsymbol{u})$, due to the densely connected prior.

This locality property extends straightforwardly to the case where each $t_i$ depends on a small number of linear degrees of freedom of $\boldsymbol{u}$. We have used three properties of the Gaussian family: closedness under marginalization, closedness under nonsingular linear transformations, and tractable marginal inference.

A similar property holds for the multinomial family $\mathcal{F}$, which is closed under marginalization as well. Here, the $t_i$ may depend on small subsets of components of $\boldsymbol{u}$. However, inference is not tractable in general in the multinomial family, so that a subfamily has to be chosen. Minka *et.al.*[13] suggest using a fixed tree-structured approximating distribution $Q$, so that $\mathcal{F}$ is the family of all multinomial distributions on this tree. For a target undirected graphical model, potentials coinciding with cliques of the tree are collected in $P^{(0)}(\boldsymbol{u})$, while all others become $t_i$ factors. Note that $\mathcal{F}$ is once more closed under marginalization and allows for tractable inference through Pearl's belief propagation method. The $t_i$ may depend on few variables only. For an update at $t_i$, let $V_i$ be the components $t_i$ depends upon. The form of $\tilde{t}_i$ is deduced by checking which tree potentials within $Q$ are affected by multiplying with $t_i$ and projecting back onto the tree (moment matching): these are the ones of the smallest subtree containing all nodes in $V_i$. To see this, note that the multiplication with $t_i$ is equivalent to introducing evidence on the nodes in $V_i$ of potentially arbitrary form. $\tilde{t}_i$ is thus represented by the potentials on this subtree. If $V_i$ is small for all $t_i$ (they use $|V_i| = 2$), Pearl's cutset conditioning method can be used to incorporate the "evidence" $t_i$. The idea is to partially instantiate $t_i$ in a minimal way, so that multiplying with the instantiated potentials does not introduce any cycles. The true new tree marginals can be computed by averaging the results for all partly instantiated $t_i$ variants. Note that between EP updates, it is not necessary to update all tree marginals, this only needs to be done on the subtree for the next $t_i$ potential. Their paper gives the details. Note that this example differs from

the Gaussian ones above, in that we do not have to compute marginals on $V_i$ in order to do the EP update for $t_i$, the reason being that we are only interested in the modifications of tree marginals this update will enduce.

Note that for all examples in this section, the underlying exponential family $\mathcal{F}$ is closed under marginalization. This requirement somewhat restricts the use of EP to subfamilies of the Gaussian or the multinomial family, based on a fixed structure. A similar restriction holds in principle for other approximate inference techniques as well. However, closedness under arbitrary marginalizations is not necessarily a binding requirement for being able to apply EP efficiently. In the common situation where the $t_i$ are local factors which are coupled by overlaps and/or a joint factor within $\mathcal{F}$, we need to be able to propagate local evidence changes coming from an EP update towards the marginal required for the next one. This is certainly possible if $\mathcal{F}$ is closed under marginalizations and allows for tractable marginal inference, but lesser requirements may be sufficient. For example, suppose that all $t_i$ potentials depend on one component of $\boldsymbol{u}$, say $u_1$. In this case, we never need to marginalize over $u_1$ in order to run EP, and consequently $\mathcal{F}$ need not be closed under $u_1$ marginalization.

# 4 Invariance of EP

By looking at the primitives EP iterates on, one guesses that the algorithm should be invariant to invertible transformations of the variables $\boldsymbol{u}$. This is true, as is shown here. It is an important property of EP, not shared by several other approximate inference methods, and it should be helpful for analyzing EP.

Let $\breve{\boldsymbol{u}} = T(\boldsymbol{u})$ be an invertible transformation satisfying the requirements for changing measures from $\boldsymbol{u}$ to $\breve{\boldsymbol{u}}$ and back (via $T^{-1}$). If $Q(\boldsymbol{u})$ lives in the exponential family $\mathcal{F}$, then $\boldsymbol{u} \sim Q$ iff $\breve{\boldsymbol{u}} \sim Q_T$, where

$$Q_T(\breve{\boldsymbol{u}}) = \exp\left(\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{u}) - \Phi(\boldsymbol{\theta})\right) d\breve{\mu}(\breve{\boldsymbol{u}}),$$

where $d\breve{\mu}(\breve{\boldsymbol{u}})$ and $d\mu(\boldsymbol{u})$ are related through the Jacobian of $T$. Let $\mathcal{F}_T$ be the class of all $Q_T$, $Q \in \mathcal{F}$. Note that the log partition function only depends on the distribution, not on the parameterization of $\boldsymbol{u}$, so is the same for $\mathcal{F}$ and $\mathcal{F}_T$. The latter has sufficient statistics $\boldsymbol{\phi} \circ T^{-1}$ and a modified base measure.

If we do EP in $\mathcal{F}_T$, based on $(P^{(0)})_T$ and sites $t_i \circ T^{-1}$, the site approximations have the same form as before, only that the sufficient statistics are the ones of $\mathcal{F}_T$. Therefore, $(Q^{\backslash i})_T = (Q_T)^{\backslash i}$. The step to $\hat{P}$ is a Bayesian update, which naturally is invariant to transformations of $\boldsymbol{u}$. Now, $\boldsymbol{\eta}^{new} = \mathrm{E}_{\hat{P}(\breve{\boldsymbol{u}})}[\boldsymbol{\phi}(T^{-1}(\breve{\boldsymbol{u}}))] = \mathrm{E}_{\hat{P}(\boldsymbol{u})}[\boldsymbol{\phi}(\boldsymbol{u})]$ is the same in both cases. Since the log partition function is invariant and determines the conversion to $\boldsymbol{\theta}^{new}$, we have shown that EP is invariant to transformations $T$. Note that this is a strong form of invariance, in that every single intermediate $Q$ does not depend on the form of $\boldsymbol{u}$, not just the final result.

This is certainly a desirable property. For a method which is not invariant in this sense, inference results may depend on the particular representation chosen for $\boldsymbol{u}$. For example, variational mean field approximations make specific factorization assumptions and are not invariant to transformations which couple variables in different factors. Common MCMC techniques share the invariance property with EP, as of course does exact Bayesian inference.

Does this property render real advantages in practice? This has not been well understood in general. We have shown that $Q$ is the same distribution after each step, no matter what $T$ is. Convergence of EP is therefore in theory not affected by $T$, as long as it is assessed by a criterion independent of the $\boldsymbol{u}$ representation (for example, the relative entropy between successive $Q$). This is an advantage of EP over certain MCMC techniques such as coordinate-wise Gibbs sampling or Metropolis-Hastings with a fixed proposal, where speed of convergence can depend significantly on the $\boldsymbol{u}$ representation.

However, our argument assumes that EP updates can be done exactly, or to very high accuracy, while in practice often approximate computations such as numerical quadrature are used. Such rules are invariant to linear transformations, but in general not to non-linear ones, and their accuracy might well depend on the exact form of $\boldsymbol{u}$. Furthermore, the update of the $Q$ representation may be more numerically stable for some forms of $\boldsymbol{u}$ than others. In fact, we can even convert between different forms of $\boldsymbol{u}$ between local updates and evidence distribution, or even use different forms depending on the site to be updated. Here, the conversion should of course itself be a stable operation.

We are also not free in choosing every transformation $T$. The family $\mathcal{F}$ may be chosen with a specific structure of the sufficient statistcs for efficiency properties. Any $T$ destroying this structure would make it much harder to run EP[5].

## 5 The Gaussian Case

As noted in Section 2, an important application of EP is concerned with networks over continuous variables with Gaussian prior $P^{(0)}(\boldsymbol{u})$. In this case, the posterior approximation $Q(\boldsymbol{u})$ is Gaussian as well. The locality property of EP within the Gaussian family was discussed in Section 3. In this section, we elaborate this important special case in detail.

We assume that $P^{(0)}$ is fully coupled, and $\mathcal{F}$ is the family of all multivariate Gaussians. Furthermore, $\boldsymbol{u} \in \mathbb{R}^n$, and $t_i(\boldsymbol{u}) = t_i(u_i)$, so there are as many sites as variables. These assumptions are made for simplicity and can easily be generalized to each $t_i$ depending on a linear function of $\boldsymbol{u}$ (see Section 3), without almost no further complications. In the underdetermined case of fewer sites than variables, additional measures have to be taken to ensure numerical stability of the method, see [25]. The locality property of EP means that in this case, we have to perform one-dimensional non-Gaussian integrals only in order to compute the tilted moments, which can usually be done using Gaussian quadrature. If $t_i$ depends on a (small) number of components of $\boldsymbol{u}$, these quadratures may still be possible, but become quite hard to do accurately[6].

The unnormalized Gaussian family is given by

$$N^U(\boldsymbol{x}|\boldsymbol{b}, \boldsymbol{\Pi}) = \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Pi}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\right),$$

where $\boldsymbol{\Pi}$ is symmetric. The sufficient statistics $\boldsymbol{\phi}(\boldsymbol{x})$ consist of $-(1/2)\boldsymbol{x}\boldsymbol{x}^T$ and $\boldsymbol{x}$, the

[5]It should be clear that by "running EP on $\breve{\boldsymbol{u}}$" we do not mean that conversions to the full $\boldsymbol{u}$ and back are done all the time.

[6]The numerical stability, convergence properties, and approximation accuracy of EP all depend significantly on the accuracy of the moment matching quadratures.

natural parameters are $\mathbf{\Pi}$ and $\boldsymbol{b}$. Note that this parameterization is not minimal. It would be if we used the lower triangle of $\mathbf{\Pi}$ only.

The site approximations are $\tilde{t}_i(u_i) = N^U(u_i|b_i, \pi_i)$. In order to do an EP update for site $i$, we first need the marginal $Q(u_i) = N(h_i, a_i)$. The cavity marginal $Q^{\backslash i}(u_i) = N(h_{\backslash i}, a_{\backslash i})$ is obtained as

$$a_{\backslash i} = \frac{a_i}{1 - a_i\pi_i}, \quad h_{\backslash i} = \frac{h_i - a_i b_i}{1 - a_i\pi_i}.$$

If $\hat{P}_i(u_i) \propto t_i(u_i)Q^{\backslash i}(u_i)$ is the tilted marginal, we need to compute its first and second moments, which is equivalent to minimizing $\mathrm{D}[\hat{P}_i \,\|\, Q']$ over Gaussian $Q'(u_i) = N(h'_i, a'_i)$. A simple way of computing these moments is via the log partition function. Define $Z_i = \mathrm{E}_{\backslash i}[t_i(u_i)]$, where $\mathrm{E}_{\backslash i}[\cdot]$ is w.r.t. $Q_{\backslash i}$. Note that $Z_i$ is the normalization constant for $\hat{P}_i$. Let

$$\alpha_i = \frac{\partial}{\partial h_{\backslash i}} \log Z_i, \quad \nu_i = -\frac{\partial^2}{\partial h_{\backslash i}^2} \log Z_i.$$

Now, it is easy to see that

$$h'_i = h_{\backslash i} + a_{\backslash i}\alpha_i, \quad a'_i = (1 - a_{\backslash i}\nu_i)a_{\backslash i}.$$

Furthermore, we can obtain $Q'$ by updating the site parameters as

$$\pi'_i = \frac{\nu_i}{1 - a_{\backslash i}\nu_i}, \quad b'_i = \pi'_i(h_{\backslash i} + \alpha_i/\nu_i) = \frac{h_{\backslash i}\nu_i + \alpha_i}{1 - a_{\backslash i}\nu_i}.$$

Another possibility is to compute $Z_i$, $h'$, $a'$ directly, which amounts to computing the moments $I_k = \int u_i^k t_i(u_i)Q^{\backslash i}(u_i)\, du_i,\ k = 0, 1, 2$. If we cannot compute $Z_i$ analytically, we can use *Gaussian quadrature* (of the Gauss-Hermite type) [6] in order to approximate the $I_k$, from which $b'_i$, $\pi'_i$ can be obtained easily.

As an example, consider the binary classification probit noise model $P(y_i|u_i) = \Phi(y_i(u_i+\beta))$, $\Phi$ the c.d.f. of $N(0, 1)$, $y_i \in \{-1, +1\}$. We have

$$Z_i = \int \Phi(y_i(u_i + \beta))Q_{\backslash i}(u_i)\, du_i = \Phi\left(\frac{y_i(h_{\backslash i} + \beta)}{\sqrt{1 + a_{\backslash i}}}\right),$$

and

$$z_i = \frac{y_i(h_{\backslash i} + \beta)}{\sqrt{1 + a_{\backslash i}}}, \quad \alpha_i = \frac{y_i N(z_i|0, 1)}{\Phi(z_i)\sqrt{1 + a_{\backslash i}}}, \quad \nu_i = \alpha_i\left(\alpha_i + \frac{h_{\backslash i} + \beta}{1 + a_{\backslash i}}\right).$$

In practice, some care has to be taken towards numerical error. First, we compute $\log Z_i$ instead of $Z_i$. For the probit noise, we use code to compute $\log \Phi(z)$ directly. We also need to take care of "0/0" situations, an example is $N(z_i|0, 1)/\Phi(z_i) = (d/dz_i) \log \Phi(z_i)$ in the term for $\alpha_i$ above. Another case where numerical stability is much more of a critical issue, is elaborated in [25] (with Laplacian sites $t_i(u_i) = e^{-\tau|u_i|}$).

Note that in general we cannot guarantee that an EP update actually can be done. Depending on $t_i$ and the current $Q(\boldsymbol{u})$, it may be that the cavity distribution or $\hat{P}_i$ are degenerate and cannot be normalized. We now show that such a breakdown cannot occur for a large class of frequently used site functions $t_i$, and that EP is a numerically stable algorithm in such cases. Namely, assume that $t_i(u_i)$ is *log-concave*, in that $\log t_i(u_i)$ is concave in $u_i$. A

powerful theorem states that if $f(\boldsymbol{x})$ is a log-concave function of $\boldsymbol{x} \in \mathbb{R}^m$, then any marginal of $f$ (obtained by integrating out some components of $\boldsymbol{x}$) is log-concave again [2]. Now, if $Z_i = \mathrm{E}_{\setminus i}[t_i(u_i)]$, as long as $Q_{\setminus i}$ is a proper Gaussian, then $Z_i$ is log-concave as a function of $h_{\setminus i}$. Namely, in this case both $t_i$ and $Q_{\setminus i}$ are log-concave in $(u_i, h_{\setminus i})$, and the product of log-concave functions is log-concave. This means that $\nu_i \geq 0$ (as second derivative of the convex function $-\log Z_i$). Now, $a'_i = (1 - a_{\setminus i}\nu_i)a_{\setminus i}$ is the variance of $\hat{P}_i$, which exists and is positive (because the support of $t_i$ has positive Lebesgue measure), therefore $1 - a_{\setminus i}\nu_i \in (0, 1]$. Now, $\pi'_i = \nu_i/(1 - a_{\setminus i}\nu_i) \geq \nu_i \geq 0$. We see that if we start with all $\pi_i = 0$, they remain nonnegative throughout for log-concave $t_i$, which in turn means that the update of the EP representation is numerically stable. The implications of log-concavity are the same in the case of the $t_i$ depending on more than a single $\boldsymbol{u}$ value, or depeding on a linear mapping of $\boldsymbol{u}$ (if $f$ is log-concave, $A$ linear, then $f \circ A$ is log-concave).

In order to implement EP, we need a representation of the posterior $Q(\boldsymbol{u})$ from which we can extract the required marginals $Q(u_i)$ efficiently, and which can be updated efficiently and in a numerically stable manner. The details of such a representation depends on the exact setup, namely the form of the prior $P^{(0)}(\boldsymbol{u})$. In some applications, many of the site parameters $b_i, \pi_i$ remain clamped at zero, which leads to a more efficient representation, an example is the Informative Vector Machine (IVM) [8, 22, 24].

We finally turn to the marginal likelihood approximation of Section 2.1. We have that

$$L = \sum_i \log C_i + \Phi[Q] - \Phi[P], \quad \Phi[N(\boldsymbol{\mu}, \boldsymbol{\Sigma})] = \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}| + \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.$$

Note that for most concrete situations, many terms cancel out in the difference $\Phi[Q] - \Phi[P]$. An illustrative example is the case of the IVM, described in [24], Sect. C.2. We have that $\log C_i = \log Z_i - \log \tilde{Z}_i$, where

$$\log \tilde{Z}_i = \Phi[Q(u_i)] - \Phi[Q_{\setminus i}(u_i)] = \frac{1}{2}\left(\log(1 - \pi_i a_i) - \frac{\pi_i h_i^2 - 2h_i b_i + a_i b_i^2}{1 - \pi_i a_i}\right).$$

Some algebra gives

$$\log \tilde{Z}_i = \frac{1}{2}\left(\log(1 - \pi_i a_i) - \frac{\pi_i h_i^2 - 2h_i b_i + a_i b_i^2}{1 - \pi_i a_i}\right).$$

The gradient of $L$ is required to drive hyperparameter estimation by empirical Bayes. In general, the prior $P^{(0)}(\boldsymbol{u}) = N^U(\boldsymbol{b}^{(0)}, \boldsymbol{\Pi}^{(0)})$, and the arguments in Section 2.1 result in

$$\nabla_{\boldsymbol{b}^{(0)}} L = \mathrm{E}_Q[\boldsymbol{u}] - \mathrm{E}_P[\boldsymbol{u}], \quad \nabla_{\boldsymbol{\Pi}^{(0)}} L = \frac{1}{2}\left(\mathrm{E}_P[\boldsymbol{u}\boldsymbol{u}^T] - \mathrm{E}_Q[\boldsymbol{u}\boldsymbol{u}^T]\right),$$

which can be used to obtain the gradient for parameters determining the prior. The reader may wonder what happens if $P^{(0)}(\boldsymbol{u})$ is not a proper Gaussian in itself, because it cannot be normalized. Recall that EP can still be used in such cases, given that the posterior $Q(\boldsymbol{u})$ is always proper. In this case, the mean of $P$ might not even exist. However, neither are all coefficients of $\boldsymbol{b}^{(0)}$, $\boldsymbol{\Pi}^{(0)}$ independent in this case, so if we derive the gradient expression for the real independent prior parameters, the problematic terms will be projected to become expressions whose prior expectations do exist.

## 6 Expectation Consistent Approximate Inference

Opper and Winther [19] proposed *expectation consistent approximate inference* (EC) as a generalization of their previous ADATAP framework [16], which in turn was the basis for expectation propagation. In this section, we introduce EC and describe its relationship to EP. It turns out that EC is equivalent to a certain parallel way of running EP, while previous applications of EP have used sequential local updates only. A parallel scheme like EC can converge faster, and to a fixed point of higher approximation quality, because each step makes use of global properties of the combination of all sites. Moreover, as pointed out by Opper and Winther, EC is more general, in that exponential families with Lebesgue and counting base measure can be used at the same time, which allows the application of EC to models with discrete variables and Gaussian-like dense couplings.

The goal of EC is to approximate moments of a target distribution $Z^{-1} f_q(\boldsymbol{u}) f_r(\boldsymbol{u})$, which is not tractable. To this end, an exponential family with sufficient statistics $\boldsymbol{g}(\boldsymbol{u})$ is chosen, with the requirement that both tilted families $\propto f_q(\boldsymbol{u}) \exp(\boldsymbol{\lambda}_q^T \boldsymbol{g}(\boldsymbol{u}))$ and $\propto f_r(\boldsymbol{u}) \exp(\boldsymbol{\lambda}_r^T \boldsymbol{g}(\boldsymbol{u}))$ are tractable, in the sense that the log partition functions and $\boldsymbol{g}(\boldsymbol{u})$ moments can be computed analytically, or through tractable quadrature. Note that $\boldsymbol{g}$ does not play exactly the same role as $\boldsymbol{\phi}$ above. The relationship is clarified below.

The idea is to keep *two* tilted distributions around,

$$q(\boldsymbol{u}) = Z_q^{-1} f_q(\boldsymbol{u}) \exp\left(\boldsymbol{\lambda}_q^T \boldsymbol{g}(\boldsymbol{u})\right), \quad r(\boldsymbol{u}) = Z_r^{-1} f_r(\boldsymbol{u}) \exp\left(\boldsymbol{\lambda}_r^T \boldsymbol{g}(\boldsymbol{u})\right),$$

with the aim of matching their moments, $\mathrm{E}_q[\boldsymbol{g}(\boldsymbol{u})] = \mathrm{E}_r[\boldsymbol{g}(\boldsymbol{u})]$. At convergence, the moments are the same, so $q$ and $r$ are expectation-consistent. Just as in other approximate inference schemes, $\boldsymbol{g}(\boldsymbol{u})$ should be chosen to represent dominating moments of the original distribution. Of course, there is a trade-off between accuracy (many moments, large cliques) and computational tractability.

Opper and Winther derive EC as log partition function (or free energy) approximation. Namely,

$$\log Z = \log Z_r + \log \mathrm{E}_r \left[ f_q(\boldsymbol{u}) \exp\left(-\boldsymbol{\lambda}_r^T \boldsymbol{g}(\boldsymbol{u})\right) \right].$$

The expectation over $r$ is intractable, but we may replace $r$ by the (non-tilted) exponential family distribution $s(\boldsymbol{u}) = Z_s^{-1} \exp(\boldsymbol{\lambda}_s^T \boldsymbol{g}(\boldsymbol{u}))$, where $\boldsymbol{\lambda}_s = \boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r$. Plugging this in, we obtain

$$\begin{aligned} \log Z &\approx \log Z_r + \log Z_s^{-1} \int f_q(\boldsymbol{u}) \exp\left((\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_r)^T \boldsymbol{g}(\boldsymbol{u})\right) \, d\boldsymbol{u} \\ &= \log Z_r + \log Z_q - \log Z_s =: \log Z_{EC}. \end{aligned} \tag{4}$$

The right hand side is the EC approximation to the negative free energy $\log Z$. Note that the derivation is symmetric w.r.t. $q$, $r$, so we could just as well start with $q$. The replacement $r \to s$ in the expectation above requires matching these distributions, and in EC we require them to be consistent on $\boldsymbol{g}(\boldsymbol{u})$: $\mathrm{E}_r[\boldsymbol{g}(\boldsymbol{u})] = \mathrm{E}_s[\boldsymbol{g}(\boldsymbol{u})]$. Since $s(\boldsymbol{u})$ is defined by these moments, there is no stronger sense of a match we could use. By symmetry, we should also require that $\mathrm{E}_q[\boldsymbol{g}(\boldsymbol{u})] = \mathrm{E}_s[\boldsymbol{g}(\boldsymbol{u})]$, so that our final fixed point conditions must be $\mathrm{E}_q[\boldsymbol{g}(\boldsymbol{u})] = \mathrm{E}_r[\boldsymbol{g}(\boldsymbol{u})]$. Another way of arriving at these is to note that the left hand side in Eq. 4 does not depend on $\boldsymbol{\lambda}_q$, $\boldsymbol{\lambda}_r$, to that the EC approximation should better be stationary w.r.t. variations in either of these variables. In other words, our final choice of $(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r)$ should be a saddle point of

$\log Z_{EC}$. We re-derive the fixed point conditions through $\nabla_{\boldsymbol{\lambda}_r} \log Z_{EC} = \mathrm{E}_r[\boldsymbol{g}(\boldsymbol{u})] - \mathrm{E}_s[\boldsymbol{g}(\boldsymbol{u})]$ and $\nabla_{\boldsymbol{\lambda}_q} \log Z_{EC} = \mathrm{E}_q[\boldsymbol{g}(\boldsymbol{u})] - \mathrm{E}_s[\boldsymbol{g}(\boldsymbol{u})]$.

Note that $\log Z_q$, $\log Z_r$ are convex in $(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r)$, while $\log Z_s$ is concave, so that $\log Z_{EC}$ is a sum of convex and concave parts, reminescent of other free energy approximations (for example, the Bethe free energy). Opper and Winther define a "single-loop" algorithm, which consists of iterating the following two steps:

1. Determine $\boldsymbol{\lambda}_s$ so that $\mathrm{E}_s[\boldsymbol{g}(\boldsymbol{u})] = \mathrm{E}_r[\boldsymbol{g}(\boldsymbol{u})]$. Update $\boldsymbol{\lambda}_q \leftarrow \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_r$.
2. Determine $\boldsymbol{\lambda}_s$ so that $\mathrm{E}_s[\boldsymbol{g}(\boldsymbol{u})] = \mathrm{E}_q[\boldsymbol{g}(\boldsymbol{u})]$. Update $\boldsymbol{\lambda}_r \leftarrow \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q$.

If this algorithm converges, we have found a saddle point of $\log Z_{EC}$ so that $\mathrm{E}_q[\boldsymbol{g}(\boldsymbol{u})] = \mathrm{E}_r[\boldsymbol{g}(\boldsymbol{u})]$. We are not aware of further general results ensuring such convergence, or proving uniqueness properties of the saddle point attained. Opper and Winther propose a "double-loop" algorithm whose inner loop minimizes convex upper bounds $\log Z_q + \log Z_r - B$ to $\log Z_{EC}$, where $B$ is linear in $(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r)$. This algorithm is provably convergent, but for most applications, the single-loop method is convergent and runs much faster.

At convergence, $\log Z_{EC}$ can be used as approximation to the intractable $\log Z$, the log marginal likelihood (see Section 2.1). If $\tau$ is some (hyper-)parameter of $f_q$ or $f_r$, the derivative $d \log Z_{EC}/d\tau$ is the sum of a direct part and a part involving $d(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r)/d\tau$. However, at convergence, the latter part vanishes, precisely because $\nabla_{(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r)} \log Z_{EC} = \mathbf{0}$. This is the same as we have shown in Section 2.1, albeit somewhat more direct.

A key insight and novelty in EP and the earlier ADATAP framework is that the (tilted) exponential families for $q$ and $r$ need not be defined w.r.t. the same base measure. As explained in Section 6.2, EC can be applied to the Ising model with counting base measure. In that case, $r$ is defined w.r.t. counting and $q$ w.r.t. Lebesgue measure. It is in such applications where the perfect symmetry of EC is broken, in that we need to decide on the base measure for the intermediate $s$ (which, by definition, is tractable to handle w.r.t. either base measure of $q$, $r$). We are reminded by such "extreme" cases that generic EC is an abstract framework whose application to a specific model has to be motivated by separate arguments. This point will become clearer for the Ising model application.

Finally, note a subtle issue in the formulation of $\log Z_{EC}$ in terms of $\boldsymbol{\lambda}_q$, $\boldsymbol{\lambda}_r$. As will be seen in Section 6.1, $\boldsymbol{\lambda}_q$ corresponds to what is called "site parameters" above. Given $\boldsymbol{\lambda}_q$, we could eliminate $\boldsymbol{\lambda}_r$ by the (unique) moment conditions $\mathrm{E}_q[\boldsymbol{g}] = \mathrm{E}_s[\boldsymbol{g}]$. However, Opper and Winther argue that in general, $-\log Z_{EC}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r(\boldsymbol{\lambda}_q))$ need not be lower bounded, and an EC fixed point need not be a local minimum of this function.

### 6.1 Equivalence to EP

Opper and Winther [19] considered models with $n$ sites $t_i(u_i)$, one for each component of $\boldsymbol{u}$. For these cases, EC is equivalent to a special parallel variant of EP, and the difference between the two is only in the method for attaining a fixed point[7]. EC is more generally applicable, as motivated by the example in Section 6.2, where Lebesgue and counting base

[7]Since the approximate free energy is not convex or concave, parallel and sequential EC can converge to different fixed points, moreover for the sequential variant, the fixed point can depend on the update ordering.

measure are used at the same time. Finally, EP has been applied more generally to models where sites and components of $\boldsymbol{u}$ are not in a one-to-one correspondence, for example to generalized linear models with more datapoints than variables [23], or to tree-structured core distributions [13]. A slight (formal) generalization of EC encompasses these cases as well, as shown in Section 6.3, so that EP and EC can be considered equivalent for all practical purposes, with EC offering a somewhat more general perspective.

To see the equivalence, let $f_q(\boldsymbol{u}) = P^{(0)}(\boldsymbol{u})$ and $f_r(\boldsymbol{u}) = \prod_i t_i(\boldsymbol{u})$. The choice of $\boldsymbol{g}(\boldsymbol{u})$ follows the same rules as the choice of $\mathcal{F}$ in EP above. In EP, we need to ensure that any titled distribution for a single site extension can be projected back onto $\mathcal{F}$ efficiently. Here, we need to ensure that both $q(\cdot)$ and $r(\cdot)$ are tractable, the latter using the product of *all* sites $t_i$. For the remainder of this section, we focus on the case $t_i(\boldsymbol{u}) = t_i(u_i)$, with the more general case being treated in Section 6.3.

Suppose we are in the setting of Section 5, with $t_i(\boldsymbol{u}) = t_i(u_i)$. In this case, we choose a factorizing Gaussian as $\boldsymbol{g}(\boldsymbol{u})$ family, and EC becomes equivalent to a parallel variant of the scheme described in Section 5. We see that the choice of $\boldsymbol{g}(\boldsymbol{u})$ in EC is different from the choice of $\mathcal{F}$ in local EP. In the latter, the tractable factor $P^{(0)}$ (fully coupled Gaussian in the Gaussian prior examples) is contained in $\mathcal{F}$, while in EC the underlying family given by $\boldsymbol{g}(\boldsymbol{u})$ has the form of $\prod_i \tilde{t}_i$, *without* the factor $P^{(0)}$. In EC, $f_q(\boldsymbol{u})$ and $f_r(\boldsymbol{u})$ are treated symmetrically. For example, the approximating distribution $Q(\boldsymbol{u})$ is $q(\boldsymbol{u})$ here, *not* $s(\boldsymbol{u})$. In fact, Opper and Winther recommend using $q(\boldsymbol{u})$ *and* $r(\boldsymbol{u})$ at convergence, depending on what prediction question is in fact asked. These agree on the $\boldsymbol{g}(\boldsymbol{u})$ moments, but apart from that are very different distributions. Couplings between variables should be extracted from $q(\boldsymbol{u})$ ($r(\boldsymbol{u})$ factorizes in their application), while higher-order marginal cumulants have to be taken from $r(\boldsymbol{u})$ ($q(\boldsymbol{u})$ is Gaussian in their example).

For completeness, we give the details of the correspondence. $\boldsymbol{g}(\boldsymbol{u})$ are the marginal Gaussian statistics, in that $\boldsymbol{g}(\boldsymbol{u}) = (\boldsymbol{g}_i(u_i))_i$, $\boldsymbol{g}_i(u_i) = (u_i, -u_i^2/2)^T$. $\boldsymbol{\lambda}_q$ are the site parameters, so that $\exp(\boldsymbol{\lambda}_q^T \boldsymbol{g}(\boldsymbol{u})) = \prod_i \tilde{t}(u_i)$. $q(\boldsymbol{u})$ is $Q(\boldsymbol{u})$, and $\log Z_q$ is $\Phi(\boldsymbol{\theta})$. $\boldsymbol{\lambda}_r$ are the cavity marginal parameters, so that $\exp(\boldsymbol{\lambda}_r^T \boldsymbol{g}(\boldsymbol{u})) \propto \prod_i Q^{\backslash i}(u_i)$.

## 6.2 EC for the Ising Model

The main application in [19] is a method for obtaining correlation estimates in the Ising model improving on the standard mean field ones. Here, $f_q(\boldsymbol{u})$ is Gaussian (the inverse covariance matrix is given), and $f_r(\boldsymbol{u}) = \prod_i t_i(u_i)$, $t_i(u_i) = (1/2)(\mathrm{I}_{\{u_i=-1\}} + \mathrm{I}_{\{u_i=+1\}})$. The Ising model is a Gaussian restricted to hypercube edges, and the restriction is done by the factorizing $f_r(\boldsymbol{u})$. Importantly, the model is defined w.r.t. the discrete counting base measure, not the continuous Lebesgue measure.

Opper and Winther consider two approximations. The simpler one uses a factorizing $s$, with $\boldsymbol{g}(\boldsymbol{u})$ as in Section 6.1. In the more advanced one, a fixed tree is chosen, and $\boldsymbol{g}(\boldsymbol{u})$ contains pairwise potentials on the egdes of the tree. Their method then iterates between the two domains that EP (and maybe most other variational approximations) has previously been applied to separately, namely continuous Gaussian and discrete tree-structured distributions. The symmetry of EC is broken, in that $s$ is a continuous Gaussian as well. We come back to this point at the end of the section.

The EC updates work as follows. First, $q(\boldsymbol{u})$ is a dense Gaussian distribution, and the tree moments can be obtained in $O(n^3)$ at most, or through Gaussian loopy belief propagation[8]. Second, $r(\boldsymbol{u})$ is a discrete tree-structured distribution, and discrete belief propagation on the tree is used to update the $s(\boldsymbol{u})$ moments[9] in $O(n)$.

Especially updates of the second kind should be compared to what we would do in local EP on this model. Opper and Winther compare against EC with fully factorized $s(\boldsymbol{u})$ family (called "EC factorized" in their work), and their tree-based approach clearly outperforms this simpler factorized variant. A more appropriate method to compare against would probably be Tree-EP [13], which could be used in the Ising model setting as well (the authors present experiments on densely connected models in [13]).

As alluded to already above, there is a remarkable assymetry in this EC application. The Ising model is defined w.r.t. counting measure, as is the $r$ distribution, yet $q$ is a Gaussian w.r.t. Lebesgue measure. What should $s$ be defined w.r.t.? This is where additional arguments are required in order to justify bringing in continuous Gaussian distributions, and Opper and Winther do that within their ADATAP approach [17, 18]. Only these arguments justify the rather drastic step of replacing the discrete $r$ by the continuous structured Gaussian $s$ in order to correct the simple approximation of $\log Z$ by $\log Z_r$ (itself motivated by Gaussian approximations to cavity marginals). Therefore, $s$ is a continuous Gaussian, and $\log Z_s$ as well as $\mathrm{E}_s[\boldsymbol{g}(\boldsymbol{u})]$ is computed w.r.t. Lebesgue measure.

### 6.3 A Generalization of EC

Strictly speaking, EC, as described in [19], does not encompass EP variants such as Tree-EP [13], or applications of EP to generalized linear models with more datapoints than variables [23]. While the authors in [19] recommend applying the EC approximation recursively, there is simpler way to obtain a generalization of EC which covers these cases[10]. Let $\boldsymbol{v}_j := \phi_j(\boldsymbol{u})$, so that the exponential family $\mathcal{F}$ on $\boldsymbol{u}$ induces an exponential family $\mathcal{F}_j$ on $\boldsymbol{v}_j$ with sufficient statistics $\boldsymbol{g}_j(\boldsymbol{v}_j)$. More precisely, if $Q(\boldsymbol{u}) \in \mathcal{F}$, then $Q(\boldsymbol{v}_j) \in \mathcal{F}_j$. This induces a map $\mathcal{F} \to \prod_j \mathcal{F}_j$. The posterior we are interested in approximating, is

$$P(\boldsymbol{u}|D) \propto P^{(0)}(\boldsymbol{u}) \prod_j t_j(\boldsymbol{v}_j).$$

We need a requirement[11], which is a restriction on the families $\mathcal{F}_j^U$. Namely, for any $Q \in \mathcal{F}$ and any $\boldsymbol{\lambda}_j$ (natural parameter for $\mathcal{F}_j^U$), whenever $Q \exp(\sum_j \boldsymbol{\lambda}_j^T \boldsymbol{g}_j(\boldsymbol{v}_j))$ can be normalized, then the corresponding distribution lies in $\mathcal{F}$.

Let us look at some examples. First, in the Gaussian case (Section 5), $\mathcal{F}$ is the family of Gaussians over $\boldsymbol{u}$. Moreover, $\boldsymbol{v}_i = u_i$, and $\mathcal{F}_i$ is the family of univariate Gaussians over

[8] Note that we need both the means and edge covariances. While the former are correct if loopy BP converges, the latter may not be. If $O(n^3)$ is prohibitive, the moments can be obtained approximately by a Lanczos method [20].

[9] The potentials are then given locally via these moments, by the junction tree theorem.

[10] It is possible that the recursive EC approximation is equivalent to this notion.

[11] It is possible that one can generalize this point even further. We require that the normal product $Q \prod_j e^{\boldsymbol{\lambda}_j^T \boldsymbol{g}_j(\boldsymbol{v}_j)}$ as a function of $\boldsymbol{u}$ lies in $\mathcal{F}$ (if normalizable). This requirement is weak enough in order to capture all variants we are interested in here. For other cases, it may be necessary to generalize this *product* in some formal way.

$u_i$. Next, in a generalized linear model [23], we have sites of the form $t_j(\boldsymbol{w}_j^T \boldsymbol{u})$, so that $\boldsymbol{v}_j = \boldsymbol{w}_j^T \boldsymbol{u}$, and $\mathcal{F}_j$ is the family of univariate Gaussians over $\boldsymbol{v}_j$. Finally, in Tree-EP [13], $\mathcal{F}$ is the family of (multinomial or Gaussian) distributions on a fixed spanning tree $T$. The sites are factors $t_j(\boldsymbol{v}_j)$, where $\boldsymbol{v}_j$ is a subset of $\boldsymbol{u}$ (set of all nodes). It is assumed that none of the $t_j$ contain egde potentials on $T$ itself. For each $j$, let $T_j$ be the smallest subtree of $T$ containing all nodes $\boldsymbol{v}_j$. Now, $\mathcal{F}_j$ is the family of distributions on $T_j$. Since all $\mathcal{F}_j$ are naturally embedded in $\mathcal{F}$, the requirement we just stated holds.

We now consider the formal family $\prod_j \mathcal{F}_j$ with statistics $\boldsymbol{g} = (\boldsymbol{g}_j)_j$. $\boldsymbol{\lambda}_s$, $\boldsymbol{\lambda}_q$, $\boldsymbol{\lambda}_r$ are natural parameters for this family, where only $\boldsymbol{\lambda}_s = \boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r$ needs to correspond to a normalizable member. Note that $q$ lives in $\mathcal{F}$, and $r$ lives in $\prod_j (\mathcal{F}_j)_{t_j}$.

Once more, it becomes clear that $r$ should not be interpreted as distribution over $\boldsymbol{u}$, reiterating the point made at the end of Section 6.2. The different distributions in the EC formulation may live in different exponential families with different base measures and different statistics. $\boldsymbol{g}(\boldsymbol{u})$ is a common ground between these different statistics, and information between the families is exchanged by moment matching, where "moments" are defined in general as parameterization of a family. Readers familiar with loopy belief propagation are reminded to "pseudomarginals" which are consistent on overlaps, but do not in general come from a proper joint distribution. In some very concrete sense, inference is a reparameterization of a target distribution [27], and many current schemes work by iterating local reparameterizations along the graph until convergence. The example of ADATAP and EC shows that it can make sense to take this abstract view literally, since it in principle allows to cross the boundaries between Lebesgue and counting measure by relying on formal exponential family mathematics.

### 6.4 Relation to Mean Field Lower Bound

Opper and Winther noted an interesting relationship of EC to standard mean field methods. For the moment, assume that the model as well as both $q$, $r$ are defined w.r.t. Lebesgue measure. We will comment on the Ising model application below.

We can start with $q(\boldsymbol{u})$ and its parameterization in terms of the site parameters $\boldsymbol{\lambda}_q$, and consider the standard mean field lower bound $\log Z = \log Z_{MF} + \mathrm{D}[q \,\|\, P(\boldsymbol{u}|D)]$, where

$$\log Z_{MF} = \log Z_q + \mathrm{E}_q[\log f_r(\boldsymbol{u})] - \boldsymbol{\lambda}_q^T \mathrm{E}_q[\boldsymbol{g}(\boldsymbol{u})].$$

A variational mean field method is obtained by maximizing the lower bound $\log Z_{MF}$ w.r.t. $\boldsymbol{\lambda}_q$. It is interesting to note that in the Gaussian case of Section 5, the optimization of the mean field lower bound over *all* joint Gaussian distributions is equivalent to the optimization of $\log Z_{MF}$ over $\boldsymbol{\lambda}_q$, the latter only being $2n$ parameters (see [21], Sect. 5.2.1).

Opper and Winther [19] relate EC to this procedure by noting that

$$\log Z_{EC} = \log Z_q + \mathrm{E}_s[\log f_r(\boldsymbol{u})] - \boldsymbol{\lambda}_q^T \mathrm{E}_s[\boldsymbol{g}(\boldsymbol{u})] + \mathrm{D}[s \,\|\, r].$$

After each EC step, we have that $\mathrm{E}_s[\boldsymbol{g}] = \mathrm{E}_q[\boldsymbol{g}]$. Keeping this constraint valid at all times is equivalent to eliminating $\boldsymbol{\lambda}_r$, leaving us with $\boldsymbol{\lambda}_q$ only. Moreover, as is discussed in Section 6.3, in all cases we are interested in, the statistics that $f_r(\boldsymbol{u})$ depends upon have the *same distribution* under $s$ and $q$. In these cases, $\log Z_{EC} = \log Z_{MF} + \mathrm{D}[s \,\|\, r]$, so the mean field

lower bound on $\log Z$ is also a lower bound to $\log Z_{EC}$, where both are taken as functions of $\boldsymbol{\lambda}_q$.

We need to be careful not to interpret this result in a wrong way. For example, the difference $\mathrm{D}[s \,\|\, r]$ between the free energy approximations is tractable to compute, but is not convex in $\mathrm{E}_q[\boldsymbol{g}]$. This would be the case for fixed $r$, but $r$ depends on $\boldsymbol{\lambda}_q$ through the (implicitly defined) $\boldsymbol{\lambda}_r$. One might think about improving on the $\log Z$ estimate given by the mean field method by plugging the final $\boldsymbol{\lambda}_q^*$ into $\log Z_{EC}$. However, this could just as well result in a value larger than $\log Z$. Moreover, even if EC on many models results in a better approximation to $\log Z$ than the mean field method, this holds only if $\log Z_{EC}$ is evaluated at an EC fixed point.

The relationship may be useful for proving that the EC free energy approximation $-\log Z_{EC}$ is lower bounded, at least if the range of site parameters is sensibly restricted. Namely, $\log Z_{EC} \leq \log Z + \mathrm{D}[s \,\|\, r]$, so we only need to show that the relative entropy $\mathrm{D}[s \,\|\, r]$ remains bounded. In the application considered above, both $s$ and $r$ are factorizing, with

$$\mathrm{D}[s \,\|\, r] = \sum_i \mathrm{D}[Q(u_i) \,\|\, \hat{P}_i(u_i)].$$

In other words, $\log Z_{EC}$ can only grow very large if at least for one $i$, the approximate marginal $Q(u_i)$ becomes very different from the tilted marginal $\hat{P}_i(u_i)$, despite the fact that they eventually have to agree on their Gaussian moments. Depending on the $t_i$, it may be possible to bound $\mathrm{D}[Q(u_i) \,\|\, \hat{P}_i(u_i)]$ in terms of some norm of $\mathrm{E}_Q[\boldsymbol{g}_i] - \mathrm{E}_{\hat{P}_i}[\boldsymbol{g}_i]$. If this holds, then the optimization could be constrained by imposing an upper bound on the latter, which would render $\log Z_{EC}$ upper bounded.

What about the application to the Ising model? In this case, the relationship between $\log Z_{EC}$ and $\log Z_{MF}$ is not of any use. Specifically above, both $\mathrm{D}[q(\boldsymbol{u}) \,\|\, P(\boldsymbol{u}|D)]$ and $\mathrm{D}[s \,\|\, r]$ are infinite, since $q$, $s$ are defined w.r.t. Lebesgue measure, $P(\boldsymbol{u}|D)$ and $r$ w.r.t. counting measure.

## References


[1] Y. Bar-Shalom and X. Li. *Estimation and Tracking: Principles, Techniques, and Software.* Artech House, 1993.

[2] V. Bogachev. *Gaussian Measures.* Mathematical Surveys and Monographs. American Mathematical Society, 1998.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2002.

[4] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In G. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence 14.* Morgan Kaufmann, 1998.

[5] Lehel Csató, Manfred Opper, and Ole Winther. TAP Gibbs free energy, belief propagation and sparsity. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 657–663. MIT Press, 2002.

[6] P. Davis and P. Rabinovitz. *Methods of Numerical Integration*. Academic Press, 1984.

[7] Tom Heskes and Onno Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In A. Darwiche and N. Friedman, editors, *Uncertainty in Artificial Intelligence 18*. Morgan Kaufmann, 2002.

[8] N. D. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process methods: The informative vector machine. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 609–616. MIT Press, 2003.

[9] T. Minka. The EP energy function and minimization schemes. See `www.stat.cmu.edu/~minka/papers/learning.html`, August 2001.

[10] T. Minka. Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller, editors, *Uncertainty in Artificial Intelligence 17*. Morgan Kaufmann, 2001.

[11] T. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, January 2001.

[12] T. Minka. Power EP. Technical report, Microsoft Research, Cambridge, 2004.

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

[14] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2003.

[15] M. Opper. A Bayesian approach to on-line learning. In D. Saad, editor, *On-Line Learning in Neural Networks*. Cambridge University Press, 1998.

[16] M. Opper and O. Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

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

[18] M. Opper and O. Winther. Tractable approximations for probabilistic models: The adaptive Thouless-Anderson-Palmer mean field approach. *Physical Review Letters*, 86(3695), 2001.

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

[20] M. Schneider and A. Willsky. Krylov subspace estimation. *SIAM Journal on Scientific Computing*, 22(5):1840–1864, 2001.

[21] M. Seeger. Bayesian methods for support vector machines and Gaussian processes. Master's thesis, University of Karlsruhe, Germany, 1999. See `www.kyb.tuebingen.mpg.de/bs/people/seeger`.

[22] M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, July 2003. See `www.kyb.tuebingen.mpg.de/bs/people/seeger`.

[23] M. Seeger, S. Gerwinn, and M. Bethge. Bayesian inference for sparse generalized linear models. In J. Kok, J. Koronacki, R. Lopez, S. Matwin, D. Mladenic, and A. Skowron, editors, *European Conference on Machine Learning 18*. Springer, 2007.

[24] M. Seeger, N. Lawrence, and R. Herbrich. Efficient nonparametric Bayesian modelling with sparse Gaussian process approximations. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2006. See `www.kyb.tuebingen.mpg.de/bs/people/seeger`.

[25] M. Seeger, F. Steinke, and K. Tsuda. Bayesian inference and optimal design in the sparse linear model. In M. Meila and X. Shen, editors, *Workshop on Artificial Intelligence and Statistics 11*, 2007.

[26] M. J. Wainwright. *Stochastic Processes on Graphs with Cycles*. PhD thesis, Massachusetts Institute of Technology, January 2002.

[27] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, Dept. of Statistics, 2003.