

Expectation Propagation for Exponential Families

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

July 20, 2007

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*

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

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

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 $\boldsymbol{\phi}(\mathbf{x})$ 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 Θ for $\boldsymbol{\theta}$ is always convex. If there are linear or affine dependencies between the components of $\boldsymbol{\phi}(\mathbf{x})$, 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 Θ . 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 $\boldsymbol{\phi}(\mathbf{x})$, $\mathbf{x} \sim P(\mathbf{x}|\boldsymbol{\theta})$:

$$\log E_{\boldsymbol{\theta}} [\exp(\boldsymbol{\varepsilon}^T \boldsymbol{\phi}(\mathbf{x}))] = \Phi(\boldsymbol{\theta} + \boldsymbol{\varepsilon}) - \Phi(\boldsymbol{\theta})$$

which exists iff $\boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \Theta$. Thus, if $\boldsymbol{\theta}$ is in the interior of Θ , the cumulants of $\boldsymbol{\phi}(\mathbf{x})$ are obtained as derivatives of $\Phi(\boldsymbol{\theta})$, especially $\nabla_{\boldsymbol{\theta}} \Phi = E_{\boldsymbol{\theta}}[\boldsymbol{\phi}(\mathbf{x})] = \boldsymbol{\eta}$ and $\nabla \nabla_{\boldsymbol{\theta}} \Phi = \text{Var}_{\boldsymbol{\theta}}[\boldsymbol{\phi}(\mathbf{x})]$. 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*

$$\Psi(\boldsymbol{\eta}) = E_{\boldsymbol{\eta}} [\log P(\mathbf{x}|\boldsymbol{\eta})],$$

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

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

Ψ 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 Ψ 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 *exponential 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 $P(\mathbf{x}|\boldsymbol{\theta})$ 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} :*

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

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

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

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

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

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

$$\Phi_f(\boldsymbol{\theta}) = \log \mathbb{E}_{\boldsymbol{\theta}} [f(\mathbf{x})] + \Phi(\boldsymbol{\theta})$$

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

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

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

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

$$\mathbb{E}_{P_f(\cdot|\boldsymbol{\theta})}[\boldsymbol{\phi}(\mathbf{x})] = \nabla_{\boldsymbol{\theta}} \log \mathbb{E}_{\boldsymbol{\theta}} [f(\mathbf{x})] + \boldsymbol{\eta}. \quad (1)$$

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*

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

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 $P(\mathbf{x}|\boldsymbol{\theta}) \in \mathcal{F}$, $P^U(\mathbf{x}|\tilde{\boldsymbol{\theta}}) \in \mathcal{F}^U$, then $P(\mathbf{x}|\boldsymbol{\theta})P^U(\mathbf{x}|\tilde{\boldsymbol{\theta}})$ is proportional to a member of \mathcal{F} (namely, to $P(\mathbf{x}|\boldsymbol{\theta} + \tilde{\boldsymbol{\theta}})$) iff $\boldsymbol{\theta} + \tilde{\boldsymbol{\theta}} \in \Theta$. Note also that $1 \equiv P^U(\mathbf{x}|\mathbf{0}) \in \mathcal{F}^U$.

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 \mathbf{u} , and with a prior distribution¹ $P^{(0)}(\mathbf{u})$ from an exponential family \mathcal{F} . We do not assume here

¹In 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 $P(S|\mathbf{u})$ often factors in a particular way,

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

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

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

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

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

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

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

\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} :

$$Q^{new}(\mathbf{u}) = \operatorname{argmin}_{\tilde{Q} \in \mathcal{F}} D[\hat{P}(\mathbf{u}) \parallel \tilde{Q}(\mathbf{u})] \quad \Leftrightarrow \quad \boldsymbol{\eta}^{new} = \mathbb{E}_{\hat{P}}[\boldsymbol{\phi}(\mathbf{u})].$$

We refer to this process as *inclusion* of site $t_i(\mathbf{u})$ 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 $P(\mathbf{u}|S)$ 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 $t_i(\mathbf{u})$ results in $Q(\mathbf{u})$ being replaced by $Q^{new}(\mathbf{u})$, which can also be seen as multiplying $Q(\mathbf{u})$ by the ratio $\tilde{t}_i(\mathbf{u}) \propto Q^{new}(\mathbf{u})/Q(\mathbf{u})$ and renormalising. This operation becomes particularly simple in the natural parameters: $\boldsymbol{\theta}^{new} = \boldsymbol{\theta} + (\boldsymbol{\theta}^{new} - \boldsymbol{\theta})$. The ratio $\tilde{t}_i(\mathbf{u})$ 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 Θ of \mathcal{F} .² This view motivates *representing* Q as

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

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

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

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

Note that we allow $\boldsymbol{\theta}^{(i)} = \mathbf{0}$, $\tilde{t}_i(\mathbf{u}) \equiv 1$, in fact in the beginning all site approximations are constant, leading to $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$, i.e. $Q(\mathbf{u}) = P^{(0)}(\mathbf{u})$. 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 $\hat{P}(\mathbf{u}) \propto Q(\mathbf{u})t_i(\mathbf{u})$ and pick $Q^{new}(\mathbf{u}) \in \mathcal{F}$ with these moments.
2. In order to replace $Q(\mathbf{u})$ by $Q^{new}(\mathbf{u})$, we replace $\tilde{t}_i(\mathbf{u}) \equiv 1$ by $\tilde{t}_i^{new}(\mathbf{u}) \propto Q^{new}(\mathbf{u})/Q(\mathbf{u})$.

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 $\tilde{t}_i(\mathbf{u})$ from $Q(\mathbf{u})$ by renormalising $Q(\mathbf{u})/\tilde{t}_i(\mathbf{u})$, obtaining

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

In natural parameters: $\boldsymbol{\theta}^{\setminus i} = \boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}$.

2. Let $\hat{P}(\mathbf{u}) = Z_i^{-1}t_i(\mathbf{u})Q^{\setminus i}(\mathbf{u})$ and compute

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

and pick $Q^{new} \in \mathcal{F}$ with these moments.

3. Replace \tilde{t}_i by $\tilde{t}_i^{new}(\mathbf{u}) \propto Q^{new}(\mathbf{u})/Q^{\setminus i}(\mathbf{u})$.
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 [21], 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, 21]. 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 $\theta \rightarrow \theta^{new}$, we update θ to a convex combination of θ and θ^{new} . Also, updates which lead to θ^{new} outside of Θ (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 $t_i(\mathbf{u})$ by corresponding $\tilde{t}_i(\mathbf{u})$, but a global fit by $Q \in \mathcal{F}$ to the distribution obtained by replacing $\tilde{t}_i(\mathbf{u})$ by $t_i(\mathbf{u})$ in the current belief $Q(\mathbf{u})$. In fact, the sites may not even be continuous functions of \mathbf{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 $P^{(0)}(\mathbf{u})$. In the former regime, every single $t_i(\mathbf{u})$ depends on a small number of components of \mathbf{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 $P^{(0)}(\mathbf{u})$ 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 [21], 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 $P^{(0)}(\mathbf{u})$ is typically densely connected, while the likelihood factors $t_i(\mathbf{u})$ are local again. The special case for a completely factorised likelihood $P(S|\mathbf{u})$ has been given in [16, 5].

2.1 Marginal Likelihood Approximation

The marginal likelihood

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

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 $P(\mathbf{u}|S)$, 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

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

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

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

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

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

³For 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 $C_i \tilde{t}_i$:

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

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 differentiable⁴. 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 $E_{\hat{P}_i}[\boldsymbol{\phi}(\mathbf{u})] = E_Q[\boldsymbol{\phi}(\mathbf{u})]$, where $\hat{P}_i(\mathbf{u}) \propto t_i(\mathbf{u})Q^{\setminus i}(\mathbf{u})$ and $Q(\mathbf{u}) \propto \tilde{t}_i(\mathbf{u})Q^{\setminus i}(\mathbf{u})$. Using our remarks on tilted exponential families in Section 1 we have

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

because $E_{\hat{P}_i}[\boldsymbol{\phi}(\mathbf{u})] = E_Q[\boldsymbol{\phi}(\mathbf{u})] = \boldsymbol{\eta}$. Furthermore,

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

so that

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

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

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

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

$$\nabla_{\boldsymbol{\theta}^{(0)}} L = E_Q[\boldsymbol{\phi}(\mathbf{u})] - E_P[\boldsymbol{\phi}(\mathbf{u})] = \boldsymbol{\eta} - \boldsymbol{\eta}^{(0)} = E_Q[\nabla_{\boldsymbol{\theta}^{(0)}} \log P^{(0)}(\mathbf{u})].$$

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:

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

⁴We 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 α is a parameter of the site $t_j(\mathbf{u})$, the dependence of L on α is direct or through $\boldsymbol{\theta}^{\setminus 0}$ (we assume that $\partial\boldsymbol{\theta}^{(0)}/\partial\alpha = \mathbf{0}$). We show that the dependence through $\boldsymbol{\theta}^{\setminus 0}$ can be ignored. If we ignore the direct dependence,

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

therefore

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

Also,

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

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

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

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

2.2 Unnormalized Prior Measure

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

Running EP for such a case is slightly more challenging. For example, we cannot start from $\mathbf{t} = \boldsymbol{\theta}^{(0)}$, 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], but this is not discussed here.

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 α is a parameter of t_j , the result of Section 2.1 remains valid if C does not depend on α .

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 $\mathbb{E}_{P(\mathbf{u}|D)}[\phi(\mathbf{u})]$ cannot be computed directly. EP will only be tractable if we can compute these moments for the cavity distributions $\hat{P}(\mathbf{u}) \propto Q^{\setminus i}(\mathbf{u})t_i(\mathbf{u})$, or more specifically if we can find θ^{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 $\mathbb{E}_{\hat{P}}[\phi(\mathbf{u})]$ efficiently, given that we know $\mathbb{E}_Q[\phi(\mathbf{u})]$, 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 $t_i(\mathbf{u})$ is a function of $v_i = \mathbf{c}_i^T \mathbf{u}$ 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 $\tilde{t}_i(\mathbf{u}) = \tilde{t}_i(v_i)$, which is parameterized by two scalars only. To see this, assume w.l.o.g. that $\mathbf{c}_i = \delta_1$, so that $v_i = u_1$. This is achieved by extending \mathbf{c}_i to a nonsingular transformation, noting that \mathcal{F} is closed under such as well. Now, $\hat{P}(\mathbf{u}) = \hat{P}(u_1)Q^{\setminus i}(\mathbf{u}_{\setminus 1}|u_1)$ and

$$\begin{aligned} \mathbb{D}[\hat{P}(\mathbf{u}) \parallel Q^{new}(\mathbf{u})] &= \mathbb{D}[\hat{P}(u_1) \parallel Q^{new}(u_1)] \\ &\quad + \mathbb{E}_{u_1 \sim \hat{P}} \left[\mathbb{D}[Q^{\setminus i}(\mathbf{u}_{\setminus 1}|u_1) \parallel Q^{new}(\mathbf{u}_{\setminus 1}|u_1)] \right]. \end{aligned}$$

In order to minimise this expression, we set $Q^{new}(\mathbf{u}_{\setminus 1}|u_1) = Q^{\setminus i}(\mathbf{u}_{\setminus 1}|u_1)$ and match moments between the marginals $\hat{P}(u_1)$ and $Q^{new}(u_1)$. It follows that $\tilde{t}_i(\mathbf{u}) = \tilde{t}_i(u_1)$, 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 $\theta^{(i)}$ are clamped to zero. Furthermore, since $Q(\mathbf{u}) \propto Q^{\setminus i}(\mathbf{u})\tilde{t}_i(u_1)$, we see that $Q(u_1) \propto Q^{\setminus i}(u_1)\tilde{t}_i(u_1)$, 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 $Q(\mathbf{u})$, 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 \mathbf{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 \mathbf{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 $P^{(0)}(\mathbf{u})$, 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 induce.

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

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

where $d\check{\mu}(\check{\mathbf{u}})$ and $d\mu(\mathbf{u})$ 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 \mathbf{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, $(Q^{\setminus i})_T = (Q_T)^{\setminus i}$. The step to \hat{P} is a Bayesian update, which naturally is invariant to transformations of \mathbf{u} . Now, $\boldsymbol{\eta}^{new} = E_{\hat{P}(\check{\mathbf{u}})}[\boldsymbol{\phi}(T^{-1}(\check{\mathbf{u}}))] = E_{\hat{P}(\mathbf{u})}[\boldsymbol{\phi}(\mathbf{u})]$ is the same in both

cases. Since the log partition function is invariant and determines the conversion to θ^{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 \mathbf{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 \mathbf{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 \mathbf{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 \mathbf{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 \mathbf{u} . Furthermore, the update of the Q representation may be more numerically stable for some forms of \mathbf{u} than others. In fact, we can even convert between different forms of \mathbf{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 statistics for efficiency properties. Any T destroying this structure would make it much harder to run EP⁵.

5 The Gaussian Case

As noted in Section 2, an important application of EP is concerned with networks over continuous variables with Gaussian prior $P^{(0)}(\mathbf{u})$. In this case, the posterior approximation $Q(\mathbf{u})$ 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, $\mathbf{u} \in \mathbb{R}^n$, and $t_i(\mathbf{u}) = t_i(u_i)$, 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 \mathbf{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 [20]. 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

⁵It should be clear that by “running EP on $\check{\mathbf{u}}$ ” we do not mean that conversions to the full \mathbf{u} and back are done all the time.

depends on a (small) number of components of \mathbf{u} , these quadratures may still be possible, but become quite hard to do accurately⁶.

The unnormalized Gaussian family is given by

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

where $\mathbf{\Pi}$ is symmetric. The sufficient statistics $\phi(\mathbf{x})$ consist of $-(1/2)\mathbf{x}\mathbf{x}^T$ and \mathbf{x} , the natural parameters are $\mathbf{\Pi}$ and \mathbf{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 $Q_{\setminus i}(u_i) = N(h_{\setminus i}, a_{\setminus i})$ is obtained as

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

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

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

Now, it is easy to see that

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

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

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

Another possibility is to compute Z_i, h', a' directly, which amounts to computing the moments $I_k = \int u_i^k t_i(u_i) Q_{\setminus i}(u_i) du_i$, $k = 0, 1, 2$. 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, π'_i can be obtained easily.

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

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

and

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

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

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 α_i above. Another case where numerical stability is much more of a critical issue, is elaborated in [20] (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 $Q(\mathbf{u})$, 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 $f(\mathbf{x})$ is a log-concave function of $\mathbf{x} \in \mathbb{R}^m$, then any marginal of f (obtained by integrating out some components of \mathbf{x}) is log-concave again [2]. Now, if $Z_i = \mathbb{E}_{\lambda_i}[t_i(u_i)]$, as long as Q_{λ_i} is a proper Gaussian, then Z_i is log-concave as a function of h_{λ_i} . Namely, in this case both t_i and Q_{λ_i} are log-concave in (u_i, h_{λ_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 $-\log Z_i$), so that if $1 - a_{\lambda_i}\nu_i > 0$, we have that $\pi'_i \geq 0$. We can give an argument why we should typically have $\nu_i < a_{\lambda_i}^{-1}$. Namely, if t_i is Gaussian with variance σ^2 , then $\nu_i = 1/(a_{\lambda_i} + \sigma^2) < 1/a_{\lambda_i}$. If t_i is a Delta distribution at some point, we have $\nu_i = 1/a_{\lambda_i}$. Therefore, we would argue that if t_i is not concentrated infinitely, then $1 - a_{\lambda_i}\nu_i > 0$. 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 \mathbf{u} value, or depending on a linear mapping of \mathbf{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 $Q(\mathbf{u})$ 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 $P^{(0)}(\mathbf{u})$. In some applications, many of the site parameters b_i, π_i remain clamped at zero, which leads to a more efficient representation, an example is the Informative Vector Machine (IVM) [8, 18, 19].

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 [19], Sect. C.2. We have that $\log C_i = \log Z_i - \log \tilde{Z}_i$, where

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

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 $P^{(0)}(\mathbf{u}) = N^U(\mathbf{b}^{(0)}, \mathbf{\Pi}^{(0)})$, and the arguments in Section 2.1 result in

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

which can be used to obtain the gradient for parameters determining the prior. The reader may wonder what happens if $P^{(0)}(\mathbf{u})$ 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 $Q(\mathbf{u})$ is always proper. In this case, the mean of P might not even exist. However, neither are all coefficients of $\mathbf{b}^{(0)}$, $\mathbf{\Pi}^{(0)}$ 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 [17] 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. On the other hand, local EP is feasible to run in situations where the global EC updates are not feasible, for example for setups with more sites than variables.

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

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

with the aim of matching their moments, $\mathbb{E}_q[\mathbf{g}(\mathbf{u})] = \mathbb{E}_r[\mathbf{g}(\mathbf{u})]$. At convergence, the moments are the same, so q and r are expectation-consistent. Just as in other approximate inference schemes, $\mathbf{g}(\mathbf{u})$ 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,

$$\log Z = \log Z_q + \log \mathbb{E}_q [f_r(\mathbf{u}) \exp(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{u}))].$$

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

obtain

$$\begin{aligned} \log Z &\approx \log Z_q + \log Z_s^{-1} \int f_r(\mathbf{u}) \exp((\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)^T \mathbf{g}(\mathbf{u})) d\mathbf{u} \\ &= \log Z_q + \log Z_r - \log Z_s =: \log Z_{EC}. \end{aligned} \quad (4)$$

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

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, reminiscent 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 $E_s[\mathbf{g}(\mathbf{u})] = E_r[\mathbf{g}(\mathbf{u})]$. Update $\boldsymbol{\lambda}_q \leftarrow \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_r$.
2. Determine $\boldsymbol{\lambda}_s$ so that $E_s[\mathbf{g}(\mathbf{u})] = E_q[\mathbf{g}(\mathbf{u})]$. 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 $E_q[\mathbf{g}(\mathbf{u})] = E_r[\mathbf{g}(\mathbf{u})]$. 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 τ 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.

6.1 Equivalence to EP

On certain models, EC is equivalent to a special parallel variant of EP. The global EC updates, however, may lead to a better fixed point than applying a local EP scheme, and this point may be reached faster.

To see the equivalence, let $f_q(\mathbf{u}) = P^{(0)}(\mathbf{u})$ and $f_r(\mathbf{u}) = \prod_i t_i(\mathbf{u})$. The choice of $\mathbf{g}(\mathbf{u})$ 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 . This requirement is violated for example in a model with Gaussian prior, but more sites than variables (see Section 3), and the parallel EC updates cannot be done in this case.

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

6.2 EC for the Ising Model

The main application in [17] is a method for obtaining correlation estimates in the Ising model improving on the standard mean field ones. Here, $f_q(\mathbf{u})$ is Gaussian (the inverse covariance matrix is given), and $f_r(\mathbf{u}) = \prod_i t_i(u_i)$, $t_i(u_i) = (1/2)(\mathbb{I}_{\{u_i=-1\}} + \mathbb{I}_{\{u_i=+1\}})$. The Ising model is a Gaussian restricted to hypercube edges, and the restriction is done by the factorizing $f_r(\mathbf{u})$. For a fixed tree structure, $\mathbf{g}(\mathbf{u})$ contains pairwise potentials on the tree edges. They propose a mixture of the two domains EP has previously been applied to separately, namely continuous Gaussian and discrete structured distributions. In consequence, $s(\mathbf{u})$ strictly speaking cannot even be seen as a single distribution anymore, but in fact is merely a product of potentials on tree edges, and only the multiplication with $f_q(\mathbf{u})$ or $f_r(\mathbf{u})$ supplies the underlying dominating measure (Lebesgue volume in the continuous case, counting on $\{-1, +1\}^n$ in the discrete case).

The EC updates work as follows. First, $q(\mathbf{u})$ is a dense Gaussian distribution, and the tree moments can be obtained in $O(n)$ using Gaussian belief propagation on the tree (recall that the inverse covariance matrix of $f_q(\mathbf{u})$ is given). Second, $r(\mathbf{u})$ is a discrete tree-structured distribution, and discrete belief propagation on the tree is used to update the $s(\mathbf{u})$ moments⁷.

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 $s(\mathbf{u})$ 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 (although it was previously applied to sparsely connected discrete models only).

⁷The potentials are then given locally via these moments, by the junction tree theorem.

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. Available online at www.stanford.edu/~boyd/cvxbook.html.
- [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] Manfred Opper and Ole Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.
- [17] Manfred Opper and Ole Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6:2177–2204, 2005.
- [18] 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.
- [19] M. Seeger, N. Lawrence, and R. Herbrich. Efficient nonparametric Bayesian modelling with sparse gaussian process approximations. Technical report, Max Planck Institute for Biologic Cybernetics, Tübingen, Germany, 2006. See www.kyb.tuebingen.mpg.de/bs/people/seeger.
- [20] 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.
- [21] M. J. Wainwright. *Stochastic Processes on Graphs with Cycles*. PhD thesis, Massachusetts Institute of Technology, January 2002.