# A Quick Introduction to Expectation Propagation

Yingzhen Li

Apr 2016

## 1 Background

### 1.1 Exponential Families

A random variable $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is said to have an *exponential family* distribution if its distribution can be written as

$$p(\boldsymbol{\theta}|\boldsymbol{\lambda}) = h(\boldsymbol{\theta}) \exp\left[\langle \boldsymbol{\lambda}, \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle - \log Z(\boldsymbol{\lambda})\right]. \tag{1}$$

Here $h(\boldsymbol{\theta})$ is the *base measure*, $\boldsymbol{\lambda}$ is called the *natural parameters* or *canonical parameters*, $\boldsymbol{\phi}(\boldsymbol{\theta})$ is the *sufficient statistic*, $\langle \cdot, \cdot \rangle$ denote the inner product and

$$Z(\boldsymbol{\lambda}) = \int_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) \exp\left[\langle \boldsymbol{\lambda}, \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle\right] d\boldsymbol{\theta}$$

is the *partition function* or *normalising constant* of the distribution. In the following we define $\boldsymbol{\lambda}$ in $\boldsymbol{\Lambda} = \{Z(\boldsymbol{\lambda}) < +\infty\}$ and assume this set is open (so the family is *regular*). We also say the exponential family is *minimal* if the coefficients of the sufficient statistic $\boldsymbol{\phi}(\cdot)$ are linear independent. Otherwise there exists $\boldsymbol{\lambda}$ such that $\langle \boldsymbol{\lambda}, \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle$ is constant, then the distribution family is *overcomplete*.

Here are some properties of exponential family distributions.

- $\log Z(\boldsymbol{\lambda})$ as a function of $\boldsymbol{\lambda}$ is convex on $\boldsymbol{\Lambda}$, strictly so if the family is minimal.
- $\nabla_{\boldsymbol{\lambda}} \log Z(\boldsymbol{\lambda}) = \mathbb{E}_p[\boldsymbol{\phi}(\boldsymbol{\theta})] := \boldsymbol{\mu}$. This is easily seen by differentiating the log partition function:

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}} \log Z(\boldsymbol{\lambda}) &= \frac{1}{Z(\boldsymbol{\lambda})} \int_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) \nabla_{\boldsymbol{\lambda}} \exp\left[\langle \boldsymbol{\lambda}, \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle\right] d\boldsymbol{\theta} \\ &= \frac{1}{Z(\boldsymbol{\lambda})} \int_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) \exp\left[\langle \boldsymbol{\lambda}, \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle\right] \boldsymbol{\phi}(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \mathbb{E}_p[\boldsymbol{\phi}(\boldsymbol{\theta})] := \boldsymbol{\mu}. \end{aligned} \tag{2}$$

  The expectation of the sufficient statistic $\boldsymbol{\mu}$ is called the *mean parameter*.

- The convex dual of $\log Z(\boldsymbol{\lambda})$ is $-H(\boldsymbol{\mu})$, where $H(\boldsymbol{\mu})$ is the entropy of distribution $p(\boldsymbol{\theta}|\boldsymbol{\lambda})$ with moments $\mathbb{E}_p[\boldsymbol{\phi}(\boldsymbol{\theta})] = \boldsymbol{\mu}$. Also $-\nabla_{\boldsymbol{\mu}} H(\boldsymbol{\mu}) = \boldsymbol{\lambda} = (\nabla_{\boldsymbol{\lambda}} \log Z(\boldsymbol{\lambda}))^{-1}$.
- In general $\log Z(\boldsymbol{\lambda}) = \max_{\boldsymbol{\mu}} \langle \boldsymbol{\lambda}, \boldsymbol{\mu} \rangle + H(\boldsymbol{\mu})$ where the optimum is attained at $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\lambda}) = \mathbb{E}_p[\boldsymbol{\phi}(\boldsymbol{\theta})]$.

For convenience in the following derivations we assume the base measure $h(\boldsymbol{\theta}) = 1$.

### 1.2 Divergence measure between exponential family distributions

Now we assume two distributions $p(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ belongs to the same exponential family[1] with natural parameters $\boldsymbol{\lambda}_p$ and $\boldsymbol{\lambda}_q$, respectively. In the following we present some useful divergence measure to describe the "closeness" of these two distributions.

- Kullback-Leibler (KL) divergence:

$$\mathrm{KL}[p||q] = \mathbb{E}_p\left[\log \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right] = \langle \boldsymbol{\lambda}_p - \boldsymbol{\lambda}_q, \boldsymbol{\mu}_p \rangle - \log Z(\boldsymbol{\lambda}_p) + \log Z(\boldsymbol{\lambda}_q). \tag{3}$$

[1] WLOG we can define $\boldsymbol{\phi} = \boldsymbol{\phi}_p \cup \boldsymbol{\phi}_q$ if $\boldsymbol{\phi}_p \neq \boldsymbol{\phi}_q$.

– Moment matching: if we minimise the inclusive $\mathrm{KL}[p||q]$ w.r.t. $\boldsymbol{\lambda}_q$:

$$\nabla_{\boldsymbol{\lambda}_q}\mathrm{KL}[p||q] = \boldsymbol{\mu}_q - \boldsymbol{\mu}_p = 0 \Rightarrow \boldsymbol{\mu}_q \leftarrow \boldsymbol{\mu}_p. \tag{4}$$

We write the corresponding natural parameter of a given moment $\boldsymbol{\mu}$ as $\boldsymbol{\lambda}(\boldsymbol{\mu})$, and this means the update of the natural parameter is $\boldsymbol{\lambda}_q \leftarrow \boldsymbol{\lambda}(\boldsymbol{\mu}_p)$.

– Natural parameter matching: if we minimise the exclusive $\mathrm{KL}[q||p]$ w.r.t. $\boldsymbol{\lambda}_q$:

$$\nabla_{\boldsymbol{\lambda}_q}\mathrm{KL}[q||p] = \langle \boldsymbol{\lambda}_q - \boldsymbol{\lambda}_p, \nabla_{\boldsymbol{\lambda}_q}\boldsymbol{\mu}_q \rangle = 0 \Rightarrow \boldsymbol{\lambda}_q \leftarrow \boldsymbol{\lambda}_p. \tag{5}$$

– SVI is natural gradient descent: for exponential families, the Fisher information can also be defined as $\boldsymbol{I}(\boldsymbol{\lambda}) = \nabla^2_{\boldsymbol{\lambda}} \log z(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}}\boldsymbol{\mu}$. So applying Amari's natural gradient descent algorithm to the exclusive KL divergence minimisation returns $\boldsymbol{\lambda}_q \leftarrow \boldsymbol{\lambda}_q - \gamma(\boldsymbol{\lambda}_q - \boldsymbol{\lambda}_p)$.

- Amari's $\alpha$-divergence: for $\alpha \neq 0, 1$,

$$\begin{aligned} \mathrm{D}_\alpha[p||q] &= \frac{1}{\alpha(1-\alpha)}\left(1 - \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta}\right) \\ &= \frac{1}{\alpha(1-\alpha)}\left(1 - \frac{Z(\alpha\boldsymbol{\lambda}_p + (1-\alpha)\boldsymbol{\lambda}_q)}{Z(\boldsymbol{\lambda}_p)^\alpha Z(\boldsymbol{\lambda}_q)^{1-\alpha}}\right). \end{aligned} \tag{6}$$

– By L'Hopital's rule it's easy to show that

$$\lim_{\alpha\to 0} \mathrm{D}_\alpha[p||q] = \mathrm{KL}[q||p],$$

$$\lim_{\alpha\to 1} \mathrm{D}_\alpha[p||q] = \mathrm{KL}[p||q].$$

– Moment matching: when $\alpha \neq 0$, if we minimise $\mathrm{D}_\alpha[p||q]$ w.r.t. $\boldsymbol{\lambda}_q$

$$\nabla_{\boldsymbol{\lambda}_q}\mathrm{D}_\alpha[p||q] = \frac{1}{\alpha}\frac{Z(\boldsymbol{\lambda}_\alpha)}{Z(\boldsymbol{\lambda}_p)^\alpha Z(\boldsymbol{\lambda}_q)^{1-\alpha}}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_\alpha) = 0 \Rightarrow \boldsymbol{\mu}_q \leftarrow \boldsymbol{\mu}_\alpha, \tag{7}$$

where we short-hand $\boldsymbol{\lambda}_\alpha = \alpha\boldsymbol{\lambda}_p + (1-\alpha)\boldsymbol{\lambda}_q$ and $\boldsymbol{\mu}_\alpha$ is the corresponding moment parameter. Note that we cannot recover the fixed point conditions for VI by limiting $\alpha \to 0$, although for the gradients $\lim_{\alpha\to 0} \nabla_{\boldsymbol{\lambda}_q}\mathrm{D}_\alpha[p||q] = \nabla_{\boldsymbol{\lambda}_q}\mathrm{KL}[q||p]$.

### 1.3 Factor Graph

# 2 Expectation Propagation

Consider for simplicity observing a dataset comprising $N$ i.i.d. samples $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$ from a probabilistic model $p(\boldsymbol{x}|\boldsymbol{\theta})$ parametrised by an unknown $D$-dimensional vector $\boldsymbol{\theta}$ that is drawn from a prior $p_0(\boldsymbol{\theta})$. Exact Bayesian inference involves computing the (typically intractable) posterior distribution of the parameters given the data,

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p_0(\boldsymbol{\theta}) \prod_{n=1}^N p(\boldsymbol{x}_n|\boldsymbol{\theta}). \tag{8}$$

Now assume the prior distribution $p_0(\boldsymbol{\theta})$ has an exponential family form as in section 1.1 with natural parameter $\boldsymbol{\lambda}_0$ and sufficient statistic $\boldsymbol{\phi}(\boldsymbol{\theta})$. Also we denote $\boldsymbol{\psi}_n(\boldsymbol{\theta}) = \log p(\boldsymbol{x}_n|\boldsymbol{\theta})$ and write the collection as $\boldsymbol{\Phi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, ..., \boldsymbol{\psi}_N, \boldsymbol{\phi})$. Then if we write $\boldsymbol{\eta} = (\mathbf{1}, \mathbf{1}, ..., \mathbf{1}, \boldsymbol{\lambda}_0)$, we can formulate the true posterior in an exponential family

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{\tilde{Z}(\boldsymbol{\eta})} \exp\left[\langle \boldsymbol{\eta}, \boldsymbol{\Phi}(\boldsymbol{\theta})\rangle\right]. \tag{9}$$

Note that the way we define the sufficient statistic $\boldsymbol{\Phi}$ also indicates the factor graph we select for representation. In this case we treat each likelihood term and the prior as factors, and later we shall see using different factorisation returns different update equations.

We approximate the true posterior with the following distribution

$$q(\boldsymbol{\theta}) = \frac{1}{Z_q} p_0(\boldsymbol{\theta}) \prod_n f_n(\boldsymbol{\theta}), \tag{10}$$

where we define $f_n(\boldsymbol{\theta}) = \exp\left[\langle \boldsymbol{\lambda}_n, \boldsymbol{\phi}(\boldsymbol{\theta})\rangle\right]$. So if we define $\boldsymbol{\lambda}_q = \boldsymbol{\lambda}_0 + \sum_n \boldsymbol{\lambda}_n$, then the approximate posterior distribution also belongs to the same exponential family as $p_0$:

$$q(\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\lambda}_q)} \exp\left[\langle \boldsymbol{\lambda}_q, \boldsymbol{\phi}(\boldsymbol{\theta})\rangle\right]. \tag{11}$$

## 2.1 The algorithm

The goal of EP is to refine the approximate factors so that they capture the contribution of each of the likelihood terms to the posterior i.e. $f_n(\boldsymbol{\theta}) \approx p(\boldsymbol{x}_n|\boldsymbol{\theta})$.

- An idealised approach would be to minimise $\mathrm{KL}[p(\boldsymbol{\theta}|\mathcal{D})||p(\boldsymbol{\theta}|\mathcal{D})f_n(\boldsymbol{\theta})/p(\boldsymbol{x}_n|\boldsymbol{\theta})]$. Unfortunately this is still intractable as it involves computing the full posterior.
- Instead, EP approximates this procedure by replacing the exact leave-one-out posterior $p_{-n}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathcal{D})/p(\boldsymbol{x}_n|\boldsymbol{\theta})$ on both sides of the KL by the approximate leave-one-out posterior (called the cavity distribution) $q_{-n}(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta})/f_n(\boldsymbol{\theta})$. Since this couples the updates for the approximating factors, the updates must now be iterated.
- This EP iteration contains four simple steps.

  1) Compute cavity: $q_{-n}(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta})/f_n(\boldsymbol{\theta})$;
  For exponential families the cavity is $q_{-n}(\boldsymbol{\theta}) \propto \exp\left[\langle \boldsymbol{\lambda}_{-n}, \boldsymbol{\phi}(\boldsymbol{\theta})\rangle\right]$ with $\boldsymbol{\lambda}_{-n} = \boldsymbol{\lambda}_q - \boldsymbol{\lambda}_n$.

  2) Compute the tilted distribution $\tilde{p}_n(\boldsymbol{\theta}) \propto q_{-n}(\boldsymbol{\theta})p(\boldsymbol{x}_n|\boldsymbol{\theta})$;
  For exponential families we can represent the tilted distribution as

$$\tilde{p}_n(\boldsymbol{\theta}) = \frac{1}{\tilde{Z}(\boldsymbol{\eta}_n)} \exp\left[\langle \boldsymbol{\eta}_n, \boldsymbol{\Phi}(\boldsymbol{\theta})\rangle\right],$$

  with $\boldsymbol{\eta}_n = (\mathbf{0}, ..., \mathbf{1}, ..., \mathbf{0}, \boldsymbol{\lambda}_{-n})$ where $\mathbf{1}$ appears at the $\mathrm{n}^{th}$ element.

  3) Update $f_n(\boldsymbol{\theta})$ by minimising $\mathrm{KL}[\tilde{p}_n(\boldsymbol{\theta})||q_{-n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta})]$;
  For exponential families this is done by first minimising $\mathrm{KL}[\tilde{p}_n(\boldsymbol{\theta})||q(\boldsymbol{\theta})]$ w.r.t. $\boldsymbol{\eta}_q = (\mathbf{0}, ..., \mathbf{0}, \boldsymbol{\lambda})$. Using the results in section 1.2 and noticing that $\boldsymbol{\lambda}$ only associates with the sufficient statistic $\boldsymbol{\phi}$, we have

$$\tilde{\boldsymbol{\mu}}_n := \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}(\boldsymbol{\theta})],\ \boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}(\tilde{\boldsymbol{\mu}}_n) \quad \Rightarrow \quad \boldsymbol{\lambda}_n \leftarrow \boldsymbol{\lambda}(\tilde{\boldsymbol{\mu}}_n) - \boldsymbol{\lambda}_{-n}.$$

  We often perform partial update to improve convergence. That is, with step-size $\gamma$, the update is specified as $\boldsymbol{\lambda}_n \leftarrow \boldsymbol{\lambda}_n + \gamma(\boldsymbol{\lambda}(\tilde{\boldsymbol{\mu}}_n) - \boldsymbol{\lambda}_{-n} - \boldsymbol{\lambda}_n) = \boldsymbol{\lambda}_n + \gamma(\boldsymbol{\lambda}(\tilde{\boldsymbol{\mu}}_n) - \boldsymbol{\lambda}_q)$.

  4) Include the new factor: $q(\boldsymbol{\theta}) \leftarrow q_{-n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta})$.
  For exponential families this means

$$\boldsymbol{\lambda}_q \leftarrow \boldsymbol{\lambda}_{-n} + \boldsymbol{\lambda}_n = \boldsymbol{\lambda}_q + \gamma(\boldsymbol{\lambda}(\tilde{\boldsymbol{\mu}}_n) - \boldsymbol{\lambda}_q).$$

## 2.2 The Energy function

There are mainly three different forms of the EP energy function. First we notice that $\boldsymbol{\eta}_n$ only depends on $\boldsymbol{\lambda}_q - \boldsymbol{\lambda}_n$ and $\boldsymbol{\psi}_n$ so we rewrite $\tilde{Z}(\boldsymbol{\eta}_n) = \tilde{Z}(\boldsymbol{\lambda}_q - \boldsymbol{\lambda}_n)$. Using local natural parameters $\{\boldsymbol{\lambda}_n\}$ we write

$$\log Z_{EP} = \sum_{n=1}^{N} \log \tilde{Z}(\boldsymbol{\lambda}_q - \boldsymbol{\lambda}_n) - \log Z(\boldsymbol{\lambda}_0) - (N-1)\log Z(\boldsymbol{\lambda}_q). \tag{12}$$

We show the EP algorithm above, if converges, returns the fixed point of $\log Z_{EP}$. By differentiating $\log Z_{EP}$ we have

$$\nabla_{\boldsymbol{\lambda}_n} \log Z_{EP} = \sum_{m \neq n} \tilde{\boldsymbol{\mu}}_m - (N-1)\boldsymbol{\mu}_q, \quad n = 1, ..., N. \tag{13}$$

So zeroing all the gradients we have all the moments matched: $\boldsymbol{\mu}_q = \tilde{\boldsymbol{\mu}}_n, \forall n$. This is also the fixed point equation of the EP updates where it iterates $\boldsymbol{\mu}_q \leftarrow \tilde{\boldsymbol{\mu}}_n$ for all datapoints.

We also note that there is an equivalent representation of the local natural parameters $\{\boldsymbol{\lambda}_n = \boldsymbol{\lambda}_q - \boldsymbol{\lambda}_{-n}\}$. If we treat $\{\boldsymbol{\lambda}_{-n}\}$ as free parameters, then we need to add in the constraint $\sum_n \boldsymbol{\lambda}_{-n} = (N-1)\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_0$. Now we have the minimax problem

$$\max_{\boldsymbol{\lambda}_q} \min_{\{\boldsymbol{\lambda}_{-n}\}} \log Z_{EP} = \sum_{n=1}^{N} \log \tilde{Z}(\boldsymbol{\lambda}_{-n}) - \log Z(\boldsymbol{\lambda}_0) - (N-1)\log Z(\boldsymbol{\lambda}_q) \quad \text{subject to } \sum_n \boldsymbol{\lambda}_{-n} = (N-1)\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_0. \tag{14}$$

Now we show the connection to the *Bethe free energy*. Recall from section 1.1 that we can write the log-partition function as $\log Z(\boldsymbol{\lambda}) = \max_{\boldsymbol{\mu}} \langle \boldsymbol{\lambda}, \boldsymbol{\mu}\rangle + H(\boldsymbol{\mu})$. This means we can optimise the moment parameters instead and replace the log-partition functions with the one containing entropy terms:

$$\log Z(\boldsymbol{\lambda}) \geq \langle \boldsymbol{\lambda}, \boldsymbol{\mu}\rangle + H(\boldsymbol{\mu}),\ \text{equality holds iff.} \boldsymbol{\mu} = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\lambda})}[\boldsymbol{\phi}] = \boldsymbol{\mu}. \tag{15}$$

So we define the corresponding representation as $\boldsymbol{\mu}_q = \boldsymbol{\mu}(\boldsymbol{\lambda}_q)$ and $\boldsymbol{\nu}_n = \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\psi}_n(\boldsymbol{\theta})]$. Recall the fixed point equations $\boldsymbol{\mu}_q = \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}(\boldsymbol{\theta})], \forall n$, and the constraint $\sum_n \boldsymbol{\lambda}_{-n} = (N-1)\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_0$. Then the EP energy function can be viewed as a generalisation of the Bethe free energy:

$$\log Z_{EP} = -\log Z(\boldsymbol{\lambda}_0) + \langle \boldsymbol{\lambda}_0, \boldsymbol{\mu}_q \rangle + \sum_{n=1}^{N} \langle \mathbf{1}, \boldsymbol{\nu}_n \rangle + \sum_{n=1}^{N} H(\boldsymbol{\mu}_q, \boldsymbol{\nu}_n) - (N-1)H(\boldsymbol{\mu}_q). \tag{16}$$

The connection to Bethe can be seen by denoting $\tilde{\boldsymbol{\mu}}_n = \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}]$ and rewriting the energy

$$-\mathcal{F}_{Bethe}(q, \{\tilde{p}_n\}) = \log Z_{EP} = (N-1) \int_{\boldsymbol{\theta}} q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} d\boldsymbol{\theta} - \sum_{n=1}^{N} \int_{\boldsymbol{\theta}} \tilde{p}_n(\boldsymbol{\theta}) \log \frac{\tilde{p}_n(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta}) p(\boldsymbol{x}_n|\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{17}$$

subject to $\tilde{\boldsymbol{\mu}}_n = \boldsymbol{\mu}_q$ for all $n$, and $\tilde{p}_n$, $q$ are valid distributions.

**The double loop algorithm**

The double loop algorithm in [Heskes and Zoeter, 2002] relaxes the Bethe free energy optimisation by

$$-\mathcal{F}_{Bethe}(q, \{\tilde{p}_n\}, q') = (N-1) \int_{\boldsymbol{\theta}} q(\boldsymbol{\theta}) \log \frac{q'(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} d\boldsymbol{\theta} - \sum_{n=1}^{N} \int_{\boldsymbol{\theta}} \tilde{p}_n(\boldsymbol{\theta}) \log \frac{\tilde{p}_n(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta}) p(\boldsymbol{x}_n|\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{18}$$

subject to the same constraints as before: $\tilde{\boldsymbol{\mu}}_n = \boldsymbol{\mu}_q$. If we write down this relaxation explicitly in terms of natural/moment parameters, this is

$$-\mathcal{F}_{Bethe}(\boldsymbol{\mu}_q, \{\tilde{\boldsymbol{\mu}}_n, \boldsymbol{\nu}_n\}, \boldsymbol{\lambda}_{q'}) = -\log Z(\boldsymbol{\lambda}_0) + \langle \boldsymbol{\lambda}_0, \boldsymbol{\mu}_q \rangle + \sum_{n=1}^{N} \langle \mathbf{1}, \boldsymbol{\nu}_n \rangle + \sum_{n=1}^{N} H(\tilde{\boldsymbol{\mu}}_n, \boldsymbol{\nu}_n) - (N-1)(\log Z(\boldsymbol{\lambda}_{q'}) - \langle \boldsymbol{\lambda}_{q'}, \boldsymbol{\mu}_q \rangle). \tag{19}$$

In [Teh et al. 2015] the relaxation is slightly different:

$$-\mathcal{F}_{Bethe}(\boldsymbol{\mu}_q, \{\tilde{\boldsymbol{\mu}}_n, \boldsymbol{\nu}_n\}, \boldsymbol{\lambda}_{q'}) = -\log Z(\boldsymbol{\lambda}_0) + \langle \boldsymbol{\lambda}_0, \boldsymbol{\mu}_q \rangle + \sum_{n=1}^{N} \langle \mathbf{1}, \boldsymbol{\nu}_n \rangle + \sum_{n=1}^{N} H(\tilde{\boldsymbol{\mu}}_n, \boldsymbol{\nu}_n) - N(\log Z(\boldsymbol{\lambda}_{q'}) - \langle \boldsymbol{\lambda}_{q'}, \boldsymbol{\mu}_q \rangle) + H(\boldsymbol{\mu}_q). \tag{20}$$

But for both cases it is easily seen that $-\mathcal{F}_{Bethe}(q, \{\tilde{p}_n\}, q') \leq -\mathcal{F}_{Bethe}(q, \{\tilde{p}_n\})$ because $\log Z(\boldsymbol{\lambda}_{q'}) - \langle \boldsymbol{\lambda}_{q'}, \boldsymbol{\mu}_q \rangle \geq H(\boldsymbol{\mu}_q)$ by convex duality. Now the optimisation contains two steps:

1) maximise $-\mathcal{F}_{Bethe}$ over $\boldsymbol{\mu}_q, \{\tilde{\boldsymbol{\mu}}_n, \boldsymbol{\nu}_n\}$ under constraints $\tilde{\boldsymbol{\mu}}_n = \boldsymbol{\mu}_q$;
   For both types of relaxations we can do reverse computations, i.e. finding the corresponding natural parameters $\boldsymbol{\eta}_n$ given $(\boldsymbol{\mu}_q, \boldsymbol{\nu}_n)$. Or, we can introduce Lagrange multipliers $\{\boldsymbol{\lambda}_n\}$ and minimise w.r.t. them. Zeroing the gradient w.r.t. $\{\boldsymbol{\lambda}_n\}$ shows that they are the natural parameters of the local approximation $f_n(\boldsymbol{\theta})$. In [Teh et al. 2015] the Lagrange multipliers are further parametrised with the corresponding moment parameters $\boldsymbol{\mu}_n = \boldsymbol{\mu}(\boldsymbol{\lambda}_n)$ and perform natural gradient descent on them.
   (This assumes $\boldsymbol{\mu}_n$ should always be in the marginal polytope, however for general EP this is not required?)

2) minimise $-\mathcal{F}_{Bethe}$ over $\boldsymbol{\lambda}_{q'}$;
   this yields update $\boldsymbol{\lambda}_{q'} \leftarrow \boldsymbol{\lambda}(\boldsymbol{\mu}_q)$.

**EP as a fixed point iteration method to a relaxed Bethe free energy optimisation**

The EP algorithm discussed in section 2.1 is derived following these steps:

1) Relax the problem: rewrite $H(\boldsymbol{\mu}_q, \boldsymbol{\nu}_n) = H(\tilde{\boldsymbol{\mu}}_n, \boldsymbol{\nu}_n)$ and add the constraint $\tilde{\boldsymbol{\mu}}_n = \boldsymbol{\mu}_q$ for all $n$;

2) Define the Lagrange multiplier $\{\boldsymbol{\lambda}_n\}$ for all these constraints;

3) Write down the Lagrangian

$$\mathcal{L}(\boldsymbol{\mu}_q, \{\tilde{\boldsymbol{\mu}}_n, \boldsymbol{\nu}_n, \boldsymbol{\lambda}_n\}) = -\log Z(\boldsymbol{\lambda}_0) + \langle \boldsymbol{\lambda}_0, \boldsymbol{\mu}_q \rangle + \sum_{n=1}^{N} \langle \mathbf{1}, \boldsymbol{\nu}_n \rangle + \sum_{n=1}^{N} (H(\tilde{\boldsymbol{\mu}}_n, \boldsymbol{\nu}_n) - H(\tilde{\boldsymbol{\mu}}_n)) + H(\boldsymbol{\mu}_q) + \sum_{n=1}^{N} \langle \boldsymbol{\lambda}_n, \boldsymbol{\mu}_q - \tilde{\boldsymbol{\mu}}_n \rangle;$$

4) Zeroing all the gradients. This returns the following fixed point equations:

$$\boldsymbol{\lambda}(\boldsymbol{\mu}_q) = \boldsymbol{\lambda}_0 + \sum_{n=1}^{N} \boldsymbol{\lambda}_n, \tag{21}$$

$$\boldsymbol{\lambda}(\tilde{\boldsymbol{\mu}}_n) = \boldsymbol{\lambda}(\tilde{\boldsymbol{\mu}}_n, \boldsymbol{\nu}_n) + \boldsymbol{\lambda}_n, \tag{22}$$

$$\boldsymbol{\mu}_q = \tilde{\boldsymbol{\mu}}_n, \forall n. \tag{23}$$

5) EP finds the fixed point by ensuring the first two conditions and iterating until the third one is satisfied for all indices.

**The minimax problem derivation**

We note that maximising the energy function (16) directly is hard even now we have no constraint on this problem. Here we provide derivations of a double loop algorithm starting from this Bethe-like objective. First using convex duality

$$H(\boldsymbol{\mu}) \leq \log Z(\boldsymbol{\lambda}) - \langle \boldsymbol{\lambda}, \boldsymbol{\mu} \rangle, \text{ equality holds iff. } \boldsymbol{\lambda} \text{ satisfies } \boldsymbol{\mu} = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\lambda})}[\boldsymbol{\phi}] = \boldsymbol{\mu}. \tag{24}$$

This means we can add in free parameters $\boldsymbol{\lambda}_q, \{\boldsymbol{\lambda}_{-n}, \boldsymbol{\xi}_n\}$ to the objective (defining $\boldsymbol{\eta}_n = (\mathbf{0}, ..., \boldsymbol{\xi}_n, ..., \mathbf{0}, \boldsymbol{\lambda}_{-n})$):

$$\log Z_{EP} \approx -\log Z(\boldsymbol{\lambda}_0) + \langle \boldsymbol{\lambda}_0 + (N-1)\boldsymbol{\lambda}_q - \sum_n \boldsymbol{\lambda}_{-n}, \boldsymbol{\mu}_q \rangle + \sum_{n=1}^{N} \langle \mathbf{1} - \boldsymbol{\xi}_n, \boldsymbol{\nu}_n \rangle + \sum_{n=1}^{N} \log \tilde{Z}(\boldsymbol{\eta}_n) - (N-1) \log Z(\boldsymbol{\lambda}_q), \tag{25}$$

where equality holds iff. the RHS is *minimised* over $\{\boldsymbol{\lambda}_{-n}, \boldsymbol{\eta}_n\}$ AND *maximised* over $\boldsymbol{\lambda}_q$. Importantly we note that we have NOT introduced any constraint yet, and the notations $\boldsymbol{\lambda}_{-n}$ at this stage should NOT be interpreted as "cavity parameters" or something similar. Also there's NO coupling between $\boldsymbol{\lambda}_q$ and $\boldsymbol{\mu}_q$, etc. However readers shall see later why we use these notation. In summary, if we denote (25) as $\mathcal{L}(\boldsymbol{\mu}_q, \{\boldsymbol{\nu}_n\}, \boldsymbol{\lambda}_q, \{\boldsymbol{\lambda}_{-n}\}, \{\boldsymbol{\xi}_n\})$, then the optimisation problem turns out to be

$$\max_{\boldsymbol{\mu}_q, \{\boldsymbol{\nu}_n\}} \max_{\boldsymbol{\lambda}_q} \min_{\{\boldsymbol{\lambda}_{-n}, \boldsymbol{\xi}_n\}} \mathcal{L}(\boldsymbol{\mu}_q, \{\boldsymbol{\nu}_n\}, \boldsymbol{\lambda}_q, \{\boldsymbol{\lambda}_{-n}\}, \{\boldsymbol{\xi}_n\}). \tag{26}$$

Now we shall eliminate the moment parameters and $\{\boldsymbol{\xi}_n\}$. Zeroing the gradients w.r.t. $\boldsymbol{\nu}_n$ and $\boldsymbol{\xi}_n$ returns $\boldsymbol{\xi}_n = \mathbf{1}$ and $\boldsymbol{\nu}_n = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\lambda}_{-n}, \boldsymbol{\xi}_n)}[\boldsymbol{\psi}_n]$. We can always assume these two conditions holds because there is no need to keep tracking of them. Then zeroing the gradient w.r.t $\boldsymbol{\mu}_q$ yields $\boldsymbol{\lambda}_0 + (N-1)\boldsymbol{\lambda}_q = \sum_n \boldsymbol{\lambda}_{-n}$. Substituting these conditions recovers the optimisation problem (14).

**Comparing EP and the double loop algorithm**

In summary, EP and the double loop convergent algorithm optimises (25) by assuming some of the fixed point conditions are always satisfied and running until the rest of them to hold. These conditions are:

1) for all $n$, $\boldsymbol{\xi}_n = \mathbf{1}$, $\boldsymbol{\nu}_n = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\lambda}_{-n}, \boldsymbol{\xi}_n)}[\boldsymbol{\psi}_n]$;
   this removes the term $\sum_{n=1}^{N} \langle \mathbf{1} - \boldsymbol{\xi}_n, \boldsymbol{\nu}_n \rangle$ in (25);

2) $(N-1)\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_0 = \sum_{n=1}^{N} \boldsymbol{\lambda}_{-n}$, $\boldsymbol{\mu}_q = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\lambda}_q)}[\boldsymbol{\phi}]$;
   if we parameterise $\boldsymbol{\lambda}_n = \boldsymbol{\lambda}_q - \boldsymbol{\lambda}_{-n}$ then the first constraint changes to $\boldsymbol{\lambda}_q = \boldsymbol{\lambda}_0 + \sum_{n=1}^{N} \boldsymbol{\lambda}_n$;
   this removes the term $\langle \boldsymbol{\lambda}_0 + (N-1)\boldsymbol{\lambda}_q - \sum_n \boldsymbol{\lambda}_{-n}, \boldsymbol{\mu}_q \rangle$ in (25);

3) for all $n$, $\tilde{\boldsymbol{\mu}}_n := \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\lambda}_{-n}, \boldsymbol{\xi}_n)}[\boldsymbol{\phi}] = \boldsymbol{\mu}_q$.