Expectation Propagation Explained

Shoichiro Yamanishi

May 15, 2020

1 Introduction

This is an expository document for expectation propagation for my future self. It is aimed at a self contained document. It converts the following three topics

- general expectation propagation with the exponential family
- detailed explanation of the clutter problem
- detailed explanation of loopy belief propagation

I have found the following subtle but important points during my own learning, which are not well explained in the existing literature. The emphasis are given on those points in this document.

- KL-divergence takes proper (normalized) density functions. The algorithm depends on the minimization of the KL divergenace to which two proper (normalized) density must be given, but we approximate a conditional $q(\theta) \approx p(x|\theta)$ where x is observed. This is not mormalized and a careful conversion is needed when applying the KL divergence.
- distinction between *moments* and *natural* parametes: The algorithm operates on the moments such as $\mathbb{E}[\boldsymbol{x}]$, $\mathbb{E}[\boldsymbol{x}^T\boldsymbol{x}]$ which are not necessarily the natural parameters for the underlying model. For example, Gaussian distribution takes $\mathbb{E}[\boldsymbol{x}]$ as the mean parameter but $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$ is different from the covariance matrix.
- careful treatment of normalization coefficients (partition function). Throughout the algorithm factors are added and removed from the current approximations. For those operations the normalizations coefficients are carefully maintaned.
- *moment matching* requires some tricks. The moment matching for the example clutter problem requires some tricks, which are not explained well in the existing literature.

Minka [6] is the original and seminal article of the expectation propagation. That is too concise as a study material as a lot of details are omitted. It presents the clutter problem, but the updated momments are presented without details. PRML [3] follows the same style as Minka [6] but the details of update of approximation maintaining the normalization coefficient (partition function) is omitted. Barber [1] briefly touches on the belief propagation in relation to expectation propagation in section 28.7. The course notes [4] by Honkela at Helsinki Univ. gives a very nice explanation. However the treatment of the normalization coefficients is not thorough. The lecture video by Simon Barthelmé [2], at Centre International de Rencontres Mathématiques gives a good explanation for cavity, hybrid, narutal parameters and moment parameters.

None of the materials above are detailed enough for normies like me to study this topic, and that was the motivation for me to write this up for my furuter self and possibly others.

2 Model and Problem Formation

We use the notation in Section 10.7 of PRML[3]. Expectation propagation is applicable for the following case. We have the following probability distribution for $\boldsymbol{x}_i$ with the latent variables $\boldsymbol{z}_i$, and a set of parameters $\boldsymbol{\theta}$. Let $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$.

$$p(\boldsymbol{\theta}, \mathcal{D}) = p_0(\boldsymbol{\theta}) \prod_{i=1}^{N} p_i(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{1}$$

Please note that it is factrized into $N+1$ factors conditioned on $\boldsymbol{\theta}$. We want to approximate the following two density estimations, which are intractable.

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{p(\mathcal{D})} p_0(\boldsymbol{\theta}) \prod_{i=1}^{N} p_i(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{2}$$

$$p(\mathcal{D}) = \int p_0(\boldsymbol{\theta}) \prod_{i=1}^{N} p_i(\boldsymbol{x}_i|\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{3}$$

The approximators will be:

$$p(\boldsymbol{\theta}, \mathcal{D}) \approx q(\boldsymbol{\theta}) = \prod_{i=0}^{N} \hat{f}_i(\boldsymbol{\theta}) \tag{4}$$

$$q(\mathcal{D}) \approx \int \prod_{i=0}^{N} \hat{f}_i(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{5}$$

where $\hat{f}_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, and $\hat{f}_i(\boldsymbol{\theta}) \approx p_i(\boldsymbol{x}_i|\boldsymbol{\theta})$ for $1 \leq i \leq N$. Please note that $\hat{f}_i(\boldsymbol{\theta}), 1 \leq i$ are not probability distribution of $\boldsymbol{\theta}$ and hence it is not normalized over $\boldsymbol{\theta}$.

and we choose $\hat{f}_i(\boldsymbol{\theta})$ from the exponential family $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}) = g(\boldsymbol{\eta}_i)\exp\left(\boldsymbol{\eta}_i^T \boldsymbol{u}(\boldsymbol{\theta})\right)$, i.e.,

$$\hat{f}_i(\boldsymbol{\theta}) = c_i g(\boldsymbol{\eta}_i)\exp\left(\boldsymbol{\eta}_i^T \boldsymbol{u}(\boldsymbol{\theta})\right) \tag{6}$$

where $g(\boldsymbol{\eta}_i) = \frac{1}{Z(\boldsymbol{\eta}_i)}$ is a normalization (partition) function, i.e.,

$$\begin{aligned} \int \hat{f}_i(\boldsymbol{\theta}) d\boldsymbol{\theta} &= c_i \int g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{\theta})\right) d\boldsymbol{\theta} \\ &= c_i \end{aligned} \tag{7}$$

and $\boldsymbol{\eta}$ is a set of *moment* parameters. Please note that $\boldsymbol{\eta}$ is not identical to the *natural* parameters such as mean $\boldsymbol{\mu}$ and covariance Σ of a normal distribution. The benefit of using the moment parameter is the ease of the following operations.

$$\begin{aligned} \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_1)\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_2) &\propto \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2) \\ \frac{\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_1)}{\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_2)} &\propto \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2) \end{aligned} \tag{8}$$

This means the joint approximation $q(\boldsymbol{\theta})$ is expressed in the same way with $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta})$ as follows.

$$\begin{aligned} \hat{q}(\boldsymbol{\theta}) &= \prod_{i=0}^{N} \hat{f}_i(\boldsymbol{\theta}) \\ &= \prod_{i=0}^{N} c_i g(\boldsymbol{\eta}_i)\exp\left(\boldsymbol{\eta}_i^T \boldsymbol{u}(\boldsymbol{\theta})\right) \\ &= \left(\prod_{i=0}^{N} c_i g(\boldsymbol{\eta}_i)\right)\exp\left(\left(\sum_{i=0}^{N} \boldsymbol{\eta}_i^T\right)\boldsymbol{u}(\boldsymbol{\theta})\right) \\ &= \frac{\left(\prod_{i=0}^{N} c_i g(\boldsymbol{\eta}_i)\right)}{g\left(\sum_{i=0}^{N} \boldsymbol{\eta}_i\right)} \mathcal{E}(\boldsymbol{\theta}|\sum_{i=0}^{N} \boldsymbol{\eta}_i) \\ &= \frac{\left(\prod_{i=0}^{N} c_i g(\boldsymbol{\eta}_i)\right)}{g\left(\boldsymbol{\eta}_{joint}\right)} \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{joint}) \\ &= C_{joint}\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{joint}) \end{aligned} \tag{9}$$

The exponential family has also the following property.

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}} \int \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}) d\boldsymbol{\theta} &= \frac{\partial}{\partial \boldsymbol{\eta}} g(\boldsymbol{\eta}) \int \exp\left(\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{\theta})\right) d\boldsymbol{\theta} \\
&= \nabla_{\boldsymbol{\eta}} g(\boldsymbol{\eta}) \int \exp\left(\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{\theta})\right) d\boldsymbol{\theta} + g(\boldsymbol{\eta}) \int \boldsymbol{u}(\boldsymbol{\theta}) \exp\left(\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{\theta})\right) d\boldsymbol{\theta} \\
&= \frac{\nabla_{\boldsymbol{\eta}} g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{\theta})] \\
&= \nabla_{\boldsymbol{\eta}} \ln g(\boldsymbol{\eta}) + \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{\theta})]
\end{aligned} \tag{10}$$

If the LHS is 0 at $\boldsymbol{\eta}^*$, i.e., the at the maximum in KL divergeance etc, then,

$$-\nabla_{\boldsymbol{\eta}} \ln g(\boldsymbol{\eta}^*) = \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}^*)}[\boldsymbol{u}(\boldsymbol{\theta})] \tag{11}$$

Now we form the KL divergence.

$$D_{KL}(p\left(\boldsymbol{\theta}|\mathcal{D}\right)||\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta})) = \int p\left(\boldsymbol{\theta}|\mathcal{D}\right) \left(\ln p\left(\boldsymbol{\theta}|\mathcal{D}\right) - \ln \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta})\right) d\boldsymbol{\theta} \tag{12}$$

We take the derivative w.r.t $\boldsymbol{\eta}$ to find $\boldsymbol{\eta}^*$ that maximizes the KL divergence.

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}} D_{KL}(p\left(\boldsymbol{\theta}|\mathcal{D}\right)||\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta})) &= -\frac{\partial}{\partial \boldsymbol{\eta}} \int p\left(\boldsymbol{\theta}|\mathcal{D}\right) \ln \mathcal{E}\left(\boldsymbol{\theta}|\boldsymbol{\eta}\right) d\boldsymbol{\theta} \\
&= -\frac{\partial}{\partial \boldsymbol{\eta}} \int p\left(\boldsymbol{\theta}|\mathcal{D}\right) \left(\ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{\theta})\right) d\boldsymbol{\theta} \\
&= -\frac{\partial}{\partial \boldsymbol{\eta}} \left(\ln g(\boldsymbol{\eta}) \int p\left(\boldsymbol{\theta}|\mathcal{D}\right) d\boldsymbol{\theta} + \int p\left(\boldsymbol{\theta}|\mathcal{D}\right) \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{\theta}) \right) d\boldsymbol{\theta} \\
&= -\frac{\partial}{\partial \boldsymbol{\eta}} \ln g(\boldsymbol{\eta}) - \int p\left(\boldsymbol{\theta}|\mathcal{D}\right) \boldsymbol{u}(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= -\frac{\partial}{\partial \boldsymbol{\eta}} \ln g(\boldsymbol{\eta}) - \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathcal{D})}[\boldsymbol{u}(\boldsymbol{\theta})] \\
&= 0
\end{aligned} \tag{13}$$

From equation 11, and 13,

$$\mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}^*)}[\boldsymbol{u}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathcal{D})}[\boldsymbol{u}(\boldsymbol{\theta})] \tag{14}$$

This means that to find the optimum $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}^*)$, we find $\boldsymbol{\eta}^*$ that makes both expectations of the moments equal.

3 Algorithm Overview

Our aim is to find the joint approximator $\hat{q}(\boldsymbol{\theta})$ of $p(\boldsymbol{x}_i|\boldsymbol{\theta})$ that minimizes the following KL divergence. Recall $\hat{q}(\boldsymbol{\theta}) = C_{joint}\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{joint})$, then

$$D_{KL}(p\left(\boldsymbol{\theta}|\mathcal{D}\right)||\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta})) = \int p\left(\boldsymbol{\theta}|\mathcal{D}\right)\left(\ln\frac{p\left(\boldsymbol{\theta}|\mathcal{D}\right)}{\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta})}\right)d\boldsymbol{\theta} \tag{15}$$

Please note that this is a reverse of the variational inference.

$$D_{KL}\left(f(\boldsymbol{\theta}|\boldsymbol{\eta})||p(\boldsymbol{\theta}|\mathcal{D})\right) = \int f(\boldsymbol{\theta}|\boldsymbol{\eta})\left(\ln\frac{f(\boldsymbol{\theta}|\boldsymbol{\eta})}{p\left(\boldsymbol{\theta}|\mathcal{D}\right)}\right)d\boldsymbol{\theta} \tag{16}$$

In case of the variational inference, high $f(\boldsymbol{\theta}|\boldsymbol{\eta})$, where the real $p(\boldsymbol{\theta})$ is low, is heavily penalized. On the other hand, in the expectation maximization, where $p(\boldsymbol{\theta})$ is high, a low $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta})$ is heavily penalized.

This minimization problem is assumed to be intractable, but optimizing each $\hat{f}_i(\boldsymbol{\theta})$ individually to $p_i(\boldsymbol{x}_i|\boldsymbol{\theta})$ leads to a poor approximation. Hence, in the similar spirit to the variational bayes, we optimize $\hat{f}_i(\boldsymbol{\theta})$ in turn keeping the other approximators.

3.1 Initialization

First we initialize the approximators as follows.

$$\hat{f}_0(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) = c_0 g(\boldsymbol{\eta}_0)\exp({\boldsymbol{\eta}_0}^T\boldsymbol{u}(\boldsymbol{\theta})) \tag{17}$$

where $c_0 = 1$, since $p_0(\boldsymbol{\theta})$ is a proper probability density function. We assume $p_0(\boldsymbol{\theta})$ is expressed in term of $\boldsymbol{\eta}_0$.

For the rest of $1 \leq i \leq N$, we initialize $\hat{f}_i(\boldsymbol{\theta}) = 1$, i.e., $c_i = 1$, and $\boldsymbol{\eta}_i = 0$. The the joint approximation $\hat{q}(\boldsymbol{\theta})$ is

$$\begin{aligned}
\hat{q}(\boldsymbol{\theta}) &= \prod_{i=0}^{N}\hat{f}_i(\boldsymbol{\theta}) \\
&= \prod_{i=0}^{N} c_i g(\boldsymbol{\eta}_i)\exp\left(\boldsymbol{\eta}_i^T\boldsymbol{u}(\boldsymbol{\theta})\right) \\
&= \left(\prod_{i=0}^{N} c_i g(\boldsymbol{\eta}_i)\right)\exp\left(\left(\sum_{i=0}^{N}\boldsymbol{\eta}_i^T\right)\boldsymbol{u}(\boldsymbol{\theta})\right) \\
&= g(\boldsymbol{\eta}_0)\exp\left(\boldsymbol{\eta}_0^T\boldsymbol{u}(\boldsymbol{\theta})\right) \\
&= C_q\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_q)
\end{aligned} \tag{18}$$

where $C_q = 1$ and $\boldsymbol{\eta}_q = \boldsymbol{\eta}_0$.

3.2 Steps in One Iteration

We first pick $1 \leq j \leq N$ to improve $\hat{f}_j(\boldsymbol{\theta})$. For that we make something called *cavity* to $q(\boldsymbol{\theta})$ by removing $\hat{f}_j(\boldsymbol{\theta})$.

$$
\begin{aligned}
\hat{q}_{\setminus j}(\boldsymbol{\theta}) &= \frac{\hat{q}(\boldsymbol{\theta})}{\hat{f}_j(\boldsymbol{\theta})} \\
&= \frac{C_q \exp(\boldsymbol{\eta}_q^T \boldsymbol{u}(\boldsymbol{\theta}))}{c_j g(\boldsymbol{\eta}_j) \exp(\boldsymbol{\eta}_j^T \boldsymbol{u}(\boldsymbol{\theta}))} \\
&= \frac{C_q}{c_j g(\boldsymbol{\eta}_j)} \exp((\boldsymbol{\eta}_q^T - \boldsymbol{\eta}_j^T)\boldsymbol{u}(\boldsymbol{\theta})) \\
&= \frac{C_q}{c_j g(\boldsymbol{\eta}_j) g(\boldsymbol{\eta}_q - \boldsymbol{\eta}_j)} \frac{\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_q - \boldsymbol{\eta}_j)}{g(\boldsymbol{\eta}_j) g(\boldsymbol{\eta}_q - \boldsymbol{\eta}_j)} \\
&= C_{\setminus j} \frac{\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{\setminus j})}{g(\boldsymbol{\eta}_{\setminus j})}
\end{aligned}
\tag{19}
$$

where

$$
\begin{aligned}
\boldsymbol{\eta}_{\setminus j} &= \boldsymbol{\eta}_q^T - \boldsymbol{\eta}_j^T \\
C_{\setminus q} &= \frac{C_q}{c_j g(\boldsymbol{\eta}_j)}
\end{aligned}
\tag{20}
$$

Then we add the true distribution $p(\boldsymbol{x}_j|\boldsymbol{\theta})$ to it to make a *hybrid*.

$$
\begin{aligned}
\hat{q}_{hyb}(\boldsymbol{\theta}) &= p(\boldsymbol{x}_j|\boldsymbol{\theta}) \hat{q}_{\setminus j}(\boldsymbol{\theta}) \\
&= p(\boldsymbol{x}_j|\boldsymbol{\theta}) C_{\setminus j} \frac{f(\boldsymbol{\theta}|\boldsymbol{\eta}_{\setminus j})}{g(\boldsymbol{\eta}_{\setminus j})}
\end{aligned}
\tag{21}
$$

We assume it is tractable to find a normalized version $q_{hyb}(\boldsymbol{\theta})$ such that

$$
\hat{q}_{hyb}(\boldsymbol{\theta}) = Z_{hyb} q_{hyb}(\boldsymbol{\theta}) \tag{22}
$$

Then we try to find $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{new})$ that minimizes the KL divergence as follows.

$$
\boldsymbol{\eta}_{new} = \underset{\boldsymbol{\eta}}{\operatorname{argmin}} \left(D_{KL}(q_{hyb}(\boldsymbol{\theta}) || \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta})) \right) \tag{23}
$$

This is achieved by the moment matching stated in equation 14.

If $\hat{q}_{hyb}(\boldsymbol{\theta})$ is expressed in the same exponential family, then we can find $\boldsymbol{\eta}_{hyb}$ and then equate as follows

$$
\boldsymbol{\eta}_{new} = \boldsymbol{\eta}_{hyb}. \tag{24}
$$

The improved $\hat{f}_j(\boldsymbol{\theta})$ is found as follows.

$$
\begin{aligned}
\hat{f}_j(\boldsymbol{\theta}) &= \frac{Z_{hyb}\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{new})}{\hat{q}_{\setminus j}(\boldsymbol{\theta})} \\
&= \frac{Z_{hyb}g(\boldsymbol{\eta}_{new})\exp(\boldsymbol{\eta}_{new}^T\boldsymbol{u}(\boldsymbol{\theta}))}{C_{\setminus q}\exp((\boldsymbol{\eta}_{\setminus j}^T)\boldsymbol{u}(\boldsymbol{\theta}))} \\
&= \frac{Z_{hyb}g(\boldsymbol{\eta}_{new})}{C_{\setminus q}}\exp((\boldsymbol{\eta}_{new}^T - \boldsymbol{\eta}_{\setminus j}^T)\boldsymbol{u}(\boldsymbol{\theta})) \\
&= \frac{Z_{hyb}g(\boldsymbol{\eta}_{new})}{C_{\setminus q}}\frac{\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{new} - \boldsymbol{\eta}_{\setminus j})}{g(\boldsymbol{\eta}_{new} - \boldsymbol{\eta}_{\setminus j})} \\
&= \frac{Z_{hyb}g(\boldsymbol{\eta}_{new})}{C_{\setminus q}g(\boldsymbol{\eta}_{new} - \boldsymbol{\eta}_{\setminus j})}\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{new} - \boldsymbol{\eta}_{\setminus j})
\end{aligned} \tag{25}
$$

and finally,

$$
\boldsymbol{\eta}_j = \boldsymbol{\eta}_{new} - \boldsymbol{\eta}_{\setminus j} \tag{26}
$$

and

$$
c_j = \frac{Z_{hyb}g(\boldsymbol{\eta}_{new})}{C_{\setminus q}g(\boldsymbol{\eta}_j)} \tag{27}
$$

This concludes one iteration.

3.3 Finding an Approximation to $p(\mathcal{D})$

After finding $\hat{q}(\boldsymbol{\theta})$, it is trivial to find an approximation to the marginal distribution $p(\mathcal{D})$ as follows.

$$
\begin{aligned}
p(\mathcal{D}) &= \int p(\mathcal{D}, \boldsymbol{\theta})d\boldsymbol{\theta} \\
&\approx \int \hat{q}(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \int \frac{C_q}{g(\boldsymbol{\eta}_q)}\mathcal{E}\left(\boldsymbol{\theta}|\boldsymbol{\eta}_q\right)d\boldsymbol{\theta} \\
&= \frac{C_q}{g(\boldsymbol{\eta}_q)}\int \mathcal{E}\left(\boldsymbol{\theta}|\boldsymbol{\eta}_q\right)d\boldsymbol{\theta} \\
&= \frac{C_q}{g(\boldsymbol{\eta}_q)}
\end{aligned} \tag{28}
$$

4 The Clatter Problem Explained

I think this is the most detailed explanation of the clutter problem, as this is an outcome of my own struggle to understand it. We basically apply the process described in the previous section. The clatter model is defined as follows.

$$
\begin{aligned}
p_0(\boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, b\boldsymbol{I}) \\
p_1(\boldsymbol{x}_i|\boldsymbol{\theta}) &= (1-\omega)\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{I}) + \omega\mathcal{N}(\boldsymbol{x}_i|\mathbf{0}, a\boldsymbol{I})
\end{aligned} \tag{29}
$$

4.1 Approximator Factor $\hat{f}_i(\boldsymbol{\theta}_i)$

And the approximator has the following form in both the natural parameters of the mean $\boldsymbol{m}_i$ and the covariance $v_i\boldsymbol{I}$, and with moment parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_i^T\boldsymbol{\theta}_i$.

$$\begin{aligned}
\hat{f}_i(\boldsymbol{\theta}) &= c_i\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}_i, v_i\boldsymbol{I}) \\
&= c_i\frac{1}{(2\pi v_i)^{D/2}}\exp(-\frac{1}{2v_i}(\boldsymbol{\theta}-\boldsymbol{m}_i)^T(\boldsymbol{\theta}-\boldsymbol{m}_i)) \\
&= c_i\frac{1}{(2\pi v_i)^{D/2}}\exp(-\frac{1}{2v_i}(\boldsymbol{\theta}^T\boldsymbol{\theta}-2\boldsymbol{\theta}^T\boldsymbol{m}_i+\boldsymbol{m}_i^T\boldsymbol{m}_i)) \\
&= c_i\frac{1}{(2\pi v_i)^{D/2}}\exp(-\frac{1}{2v_i}\boldsymbol{m}_i^T\boldsymbol{m}_i)\exp(-\frac{1}{2v_i}(\boldsymbol{\theta}^T\boldsymbol{\theta}+\frac{1}{v_i}(\boldsymbol{m}_i^T\boldsymbol{\theta})) \\
&= c_i\frac{\exp(-\frac{1}{2v_i}\boldsymbol{m}_i^T\boldsymbol{m}_i)}{(2\pi v_i)^{D/2}}\exp(-\frac{1}{2v_i}(\boldsymbol{\theta}^T\boldsymbol{\theta}+\frac{1}{v_i}(\boldsymbol{m}_i^T\boldsymbol{\theta})) \\
&= c_i\frac{\exp(-\frac{1}{2v_i}\boldsymbol{m}_i^T\boldsymbol{m}_i)}{(2\pi v_i)^{D/2}}\exp(\begin{bmatrix}\frac{1}{v_i}\boldsymbol{m}_i^T & -\frac{1}{2v_i}\end{bmatrix}\begin{bmatrix}\boldsymbol{\theta}\\ \boldsymbol{\theta}^T\boldsymbol{\theta}\end{bmatrix}) \\
&= c_i g(\boldsymbol{\eta}_i)\exp(\boldsymbol{\eta}_i^T\boldsymbol{u}(\boldsymbol{\theta})) \\
&= c_i\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_i)
\end{aligned} \tag{30}$$

where the moment parameters are

$$\boldsymbol{\eta}_i = \begin{bmatrix}\frac{1}{v_i}\boldsymbol{m}_i \\ -\frac{1}{2v_i}\end{bmatrix} \tag{31}$$

and

$$\boldsymbol{u}(\boldsymbol{\theta}) = \begin{bmatrix}\boldsymbol{\theta}\\ \boldsymbol{\theta}^T\boldsymbol{\theta}\end{bmatrix} \tag{32}$$

and finally the normalization function $g(\boldsymbol{\eta})$ is

$$g(\boldsymbol{\eta}_i) = \frac{\exp(-\frac{1}{2v_i}\boldsymbol{m}_i^T\boldsymbol{m}_i)}{(2\pi v_i)^{D/2}} \tag{33}$$

4.2 Initialization of the Approximators

From the definitions in 29,

$$\begin{aligned}
\hat{q}_0(\boldsymbol{\theta}) &= \hat{p}_0(\boldsymbol{\theta}) \\
&= \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{0}, b\boldsymbol{I}) \\
&= c_0 g(\boldsymbol{\eta}_0)\exp(\boldsymbol{\eta}_0^T\boldsymbol{u}(\boldsymbol{\theta})) \\
&= c_0\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_0)
\end{aligned} \tag{34}$$

where $c_0 = 1$, $\boldsymbol{\eta}_0 = \begin{bmatrix}\boldsymbol{0}\\ -\frac{1}{2b}\end{bmatrix}$,

And for $1 \leq j \leq N$,

$$\begin{aligned}\hat{q}_j(\boldsymbol{\theta}) &= 1 \\ &= c_j g(\boldsymbol{\eta}_j) \exp(\boldsymbol{\eta}_j^T \boldsymbol{u}(\boldsymbol{\theta})) \\ &= c_j \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_j)\end{aligned} \tag{35}$$

where $c_j = 1$, $\boldsymbol{\eta}_j = \begin{bmatrix}\mathbf{0} \\ 0\end{bmatrix}$, altogether

4.3 Making a Cavity

This is a straight forward application of the equation 19.

4.4 Making a Hybrid

According to equations 22 and 29,

$$\begin{aligned}\hat{q}_{hyb}(\boldsymbol{\theta}) &= p(\boldsymbol{x}_j|\boldsymbol{\theta})\hat{q}_{\setminus j}(\boldsymbol{\theta}) \\ &= \left((1-\omega)\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{I}) + \omega\mathcal{N}(\boldsymbol{x}_i|\mathbf{0}, a\boldsymbol{I})\right)\hat{q}_{\setminus j}(\boldsymbol{\theta}) \\ &= \left((1-\omega)\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{x}_i, \boldsymbol{I}) + \omega\mathcal{N}(\boldsymbol{x}_i|\mathbf{0}, a\boldsymbol{I})\right)\hat{q}_{\setminus j}(\boldsymbol{\theta}) \\ &= (1-\omega)\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{x}_i, \boldsymbol{I})\hat{q}_{\setminus j}(\boldsymbol{\theta}) + \omega\mathcal{N}(\boldsymbol{x}_i|\mathbf{0}, a\boldsymbol{I})\hat{q}_{\setminus j}(\boldsymbol{\theta}) \\ &= (1-\omega)\mathcal{E}(\boldsymbol{\theta}|\begin{bmatrix}\boldsymbol{x}_i \\ -\frac{1}{2}\end{bmatrix})\hat{q}_{\setminus j}(\boldsymbol{\theta}) + \omega\mathcal{N}(\boldsymbol{x}_i|\mathbf{0}, a\boldsymbol{I})\hat{q}_{\setminus j}(\boldsymbol{\theta}) \\ &= (1-\omega)\mathcal{E}(\boldsymbol{\theta}|\begin{bmatrix}\boldsymbol{x}_i \\ -\frac{1}{2}\end{bmatrix})C_{\setminus j}g(\boldsymbol{\eta}_{\setminus j})\exp(\boldsymbol{\theta}|\boldsymbol{\eta}_{\setminus j}) + \omega\mathcal{N}(\boldsymbol{x}_i|\mathbf{0}, a\boldsymbol{I})C_{\setminus j}\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{\setminus j}) \\ &= (1-\omega)g(\begin{bmatrix}\boldsymbol{x}_i \\ -\frac{1}{2}\end{bmatrix})\exp(\begin{bmatrix}\boldsymbol{x}_i \\ -\frac{1}{2}\end{bmatrix}^T \boldsymbol{u}(\boldsymbol{\theta}))C_{\setminus j}g(\boldsymbol{\eta}_{\setminus j})\exp(\boldsymbol{\eta}_{\setminus j}^T\boldsymbol{u}(\boldsymbol{\theta})) + \omega\mathcal{N}(\boldsymbol{x}_i|\mathbf{0}, a\boldsymbol{I})C_{\setminus j}\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{\setminus j}) \\ &= (1-\omega)g(\begin{bmatrix}\boldsymbol{x}_i \\ -\frac{1}{2}\end{bmatrix})C_{\setminus j}g(\boldsymbol{\eta}_{\setminus j})\exp\left((\begin{bmatrix}\boldsymbol{x}_i \\ -\frac{1}{2}\end{bmatrix} + \boldsymbol{\eta}_{\setminus j})^T\boldsymbol{u}(\boldsymbol{\theta})\right) + \omega\mathcal{N}(\boldsymbol{x}_i|\mathbf{0}, a\boldsymbol{I})C_{\setminus j}\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{\setminus j}) \\ &= \frac{(1-\omega)g(\begin{bmatrix}\boldsymbol{x}_i \\ -\frac{1}{2}\end{bmatrix})C_{\setminus j}g(\boldsymbol{\eta}_{\setminus j})}{g(\begin{bmatrix}\boldsymbol{x}_i \\ -\frac{1}{2}\end{bmatrix} + \boldsymbol{\eta}_{\setminus j})}\mathcal{E}\left(\boldsymbol{\theta}|\begin{bmatrix}\boldsymbol{x}_i \\ -\frac{1}{2}\end{bmatrix} + \boldsymbol{\eta}_{\setminus j}\right) + \omega\mathcal{N}(\boldsymbol{x}_i|\mathbf{0}, a\boldsymbol{I})C_{\setminus j}\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{\setminus j})\end{aligned} \tag{36}$$

where we have used

$$\begin{aligned}\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{x}_i, \boldsymbol{I}) &= \frac{1}{(2\pi)^{D/2}}\exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{x}_i)^T(\boldsymbol{\theta}-\boldsymbol{x}_i)\right) \\ &= \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{I})\end{aligned} \tag{37}$$

Now we intfoduce the following shorthand notations.

$$\begin{aligned}\alpha_j &= \frac{(1-\omega)g(\begin{bmatrix}\boldsymbol{x}_i\\-\frac{1}{2}\end{bmatrix})C_{\backslash j}g(\boldsymbol{\eta}_{\backslash j})}{g(\begin{bmatrix}\boldsymbol{x}_i\\-\frac{1}{2}\end{bmatrix}+\boldsymbol{\eta}_{\backslash j})}\\ \beta_j &= \omega\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{0}, a\boldsymbol{I})C_{\backslash j}\end{aligned} \tag{38}$$

then,

$$\begin{aligned}\hat{q}_{hyb}(\boldsymbol{\theta}) &= \alpha_j\mathcal{E}\left(\boldsymbol{\theta}|\begin{bmatrix}\boldsymbol{x}_i\\-\frac{1}{2}\end{bmatrix}+\boldsymbol{\eta}_{\backslash j}\right)+\beta_j\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{\backslash j})\\ &= (\alpha_j+\beta_j)\left(\frac{\alpha_j}{\alpha_j+\beta_j}\mathcal{E}\left(\boldsymbol{\theta}|\begin{bmatrix}\boldsymbol{x}_i\\-\frac{1}{2}\end{bmatrix}+\boldsymbol{\eta}_{\backslash j}\right)+\frac{\beta_j}{\alpha_j+\beta_j}\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{\backslash j})\right)\end{aligned} \tag{39}$$

This means $Z_{hyb}=\alpha_j+\beta_j$, and let $\omega'=\frac{\alpha_j}{\alpha_j+\beta_j}$,

$$q_{hyb}(\boldsymbol{\theta})=\omega'\mathcal{E}\left(\boldsymbol{\theta}|\begin{bmatrix}\boldsymbol{x}_i\\-\frac{1}{2}\end{bmatrix}+\boldsymbol{\eta}_{\backslash j}\right)+(1-\omega')\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\eta}_{\backslash j}) \tag{40}$$

4.5 Moment Matching

From equation 40, $q_{hyb}(\boldsymbol{\theta})$ is a mixture of 2 Gaussians, and the by the linearity of moments [5], the combined moments are given as follows.

$$\boldsymbol{\eta}_{new}=\omega'\left(\begin{bmatrix}\boldsymbol{x}_i\\-\frac{1}{2}\end{bmatrix}+\boldsymbol{\eta}_{\backslash j}\right)+(1-\omega')\boldsymbol{\eta}_{\backslash j} \tag{41}$$

4.6 Updating a Factor $\hat{q}_j(\boldsymbol{\theta}|\boldsymbol{\eta}_j)$

This is a straight forward appliation of equations 25, 26, and 27.

5 Loopy Belief Propagation

Section 4 of Minka[6] states the relationship between the expectation propagation and the graphical belief propagation by using the same KL divergence. Let $\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_N\}=Z$ be the random variables of 1-of-K vectors defined on the graph, and $\mathcal{D}$ be the set of observe variables, for example $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$, but it could be anything. Our objective is to find the posterior $p(\boldsymbol{z}_i|\mathcal{D})$, and evidence $p(\mathcal{D})$. We assume $p(\boldsymbol{z}_i|\mathcal{D})$ factorizes as follows.

$$p(\boldsymbol{z}_i|\mathcal{D})=\prod_{i=1}^{F}p_i(Z_i|\mathcal{D}) \tag{42}$$

where Z_i is a subset of Z.

We assume finding the posterior and the evidence is not a tractable problem and hence we propose an approximation.

$$q_i(\boldsymbol{z}_i) \approx p_i(Z_i|\mathcal{D}) \tag{43}$$

Our objective is to find a good approximation by minimizing

$$D_{KL}\left(p_j(Z_j|\mathcal{D})\prod_{i\neq j} q(Z_i)\,\Big|\Big|\, q_{new}(Z)\right) \tag{44}$$

factor-by-factor just like the expectation propagation stated above. However, this time, we don't use the moment matching but the following property.

Assume we have a disjoint partition of $Z = \mathcal{Z}_1 \cup \mathcal{Z}_2 \cup \cdots \cup \mathcal{Z}_L$ where $\mathcal{Z}_i \cap \mathcal{Z}_j = \emptyset$ for $i \neq j$. We have a probabily distribution $p(Z)$ and we try to approximate it by factoring accoding to the disjoint partition of Z as follows.

$$p(Z) \approx \prod_{i=1}^{L} q_i(\mathcal{Z}_i) \tag{45}$$

Then for each $q_j(\mathcal{Z}_j)$ the best approximation in terms of minimization of the KL divergence above is given by

$$q_j(\mathcal{Z}_j) = \int p(Z) dZ_{\setminus j} \tag{46}$$

where $Z_{\setminus j} = (Z \setminus \mathcal{Z}_j)$. Hence the best approximator $q_(\mathcal{Z}_j)$ is given by the marginalization of $p(Z)$ by $dZ_{\setminus j}$.

This can be proven as follows. First, expand the KL divergence as follows.

$$\begin{aligned} D_{KL}\left(p(Z)\Big|\Big|\prod_{i=1}^{L} q_i(\mathcal{Z})\right) &= \int p(Z)\ln\frac{p(Z)}{\prod_{i=1}^{L} q_i(\mathcal{Z})}dZ \\ &= \int p(Z)\ln p(Z) - \sum_{i=1}^{L}\ln q_i(\mathcal{Z}_i)dZ \end{aligned} \tag{47}$$

We form a Lagrangian multiplier on $q_j(\mathcal{Z})$ with the constraint $\int q_j(\mathcal{Z})d\mathcal{Z}_j = 1$ as follows.

$$\mathcal{L}(q_j(\mathcal{Z}_j), \lambda) = \int p(Z)\ln p(Z) - \sum_{i=1}^{L}\ln q_i(\mathcal{Z}_i)dZ + \lambda(\int q_j(\mathcal{Z}_i)d\mathcal{Z}_j - 1) \tag{48}$$

Take the variational derivative of $q_j(\mathcal{Z})$ and equate it with 0.

$$
\begin{aligned}
\frac{\partial}{\partial q_j(\mathcal{Z})}\mathcal{L}(q_j(\mathcal{Z}_j),\lambda) &= \frac{\partial}{\partial q_j(\mathcal{Z})}\left(-\int p(Z)\ln q_j(\mathcal{Z}_j)dZ + \lambda(\int q_j(\mathcal{Z}_i)d\mathcal{Z}_j - 1)\right) \\
&= \frac{\partial}{\partial q_j(\mathcal{Z})}\left(-\int\int p(Z)\ln q_j(\mathcal{Z}_j)d\mathcal{Z}_{\setminus j}d\mathcal{Z}_j + \lambda(\int q_j(\mathcal{Z}_i)d\mathcal{Z}_j - 1)\right) \\
&= \frac{\partial}{\partial q_j(\mathcal{Z})}\left(-\int \ln q_j(\mathcal{Z}_j)\left(\int p(Z)d\mathcal{Z}_{\setminus j}\right)d\mathcal{Z}_j + \lambda(\int q_j(\mathcal{Z}_i)d\mathcal{Z}_j - 1)\right) \\
&= -\frac{1}{q_j(\mathcal{Z}_j)}\left(\int p(Z)d\mathcal{Z}_{\setminus j}\right) + \lambda = 0
\end{aligned}
\tag{49}
$$

hence,

$$
q_j(\mathcal{Z}_j) = \frac{1}{\lambda}\left(\int p(Z)d\mathcal{Z}_{\setminus j}\right) \tag{50}
$$

as $p(Z)$ is a proper (normalized) density and hence $\lambda = 1$, and

$$
q_j(\mathcal{Z}_j) = \int p(Z)d\mathcal{Z}_{\setminus j} \tag{51}
$$

Q.E.D.

Next, we apply this propery of $q_j(\mathcal{Z}_j) = \int p(Z)d\mathcal{Z}_{\setminus j}$ in the belief propagation framework. This can be best presented by going through the following example.

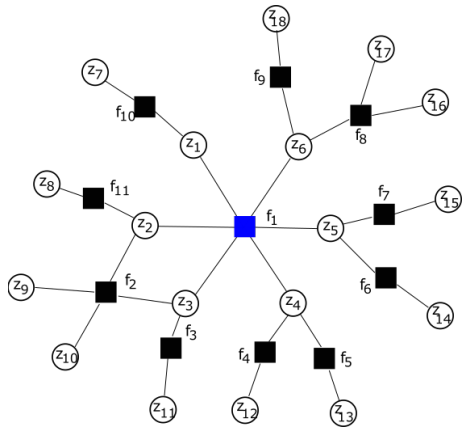

Figure 1: Subgraph of a Factor Graph

Please see figure 1, which is a node-induced subgraph of a factor graph. We assume we have corresponding arpproximators $q_i(Z_i)$ for each factor $f_i(Z_i)$. We want to improve $q_1(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4, \boldsymbol{z}_5, \boldsymbol{z}_6)$ for the factor $f_1(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4, \boldsymbol{z}_5, \boldsymbol{z}_6)$. Let $q(Z) = \prod_{i=1}^{L}(Z_i)$ be the current approximation of $p(Z|\mathcal{D})$. We apply the expectation propagation framework. First, we make a cavity $q_{\setminus 1}(Z) = \frac{q(Z)}{q_1(Z_j)}$, and then a hybrid $f_1(Z_1)q_{\setminus 1}(Z)$. Then we form a KL divergence as

$D_{KL}(f_1(Z_1)q_{\setminus 1}(Z)||q_{new}(Z))$, but this time, we don't apply the moment matching but apply the results from equation 46. Please note that equation 46 applies when the partition of Z is disjoint. For that we further factorize each $q_i(Z_i)$ as in figure 2

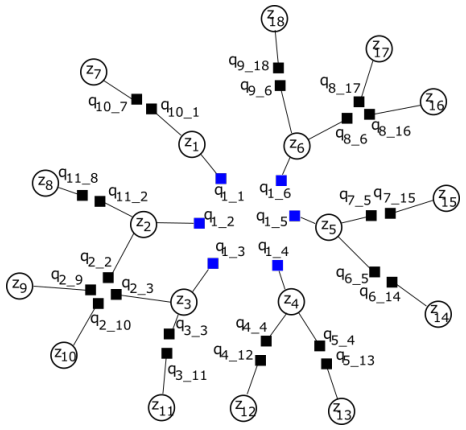

Figure 2: Fully Factored Factor Graph of $q(Z)$

The corresponding hybrid is shown in figure 3.

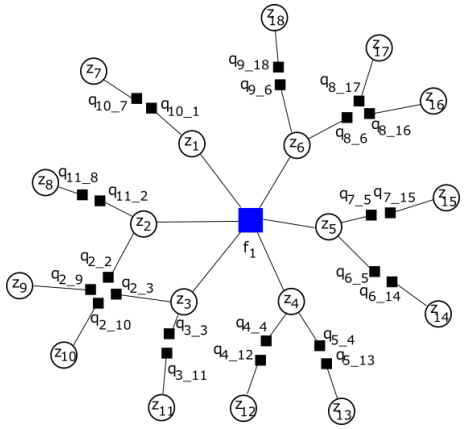

Figure 3: Hybrid $f_1(Z_1)q_{\setminus 1}(Z)$

Now we apply equation 46 for each of the variables of $f_1(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4, \boldsymbol{z}_5, \boldsymbol{z}_6)$. Take $\boldsymbol{z}_4$ for an example.

Please see figure 4. This corresponds to the following marginalization.

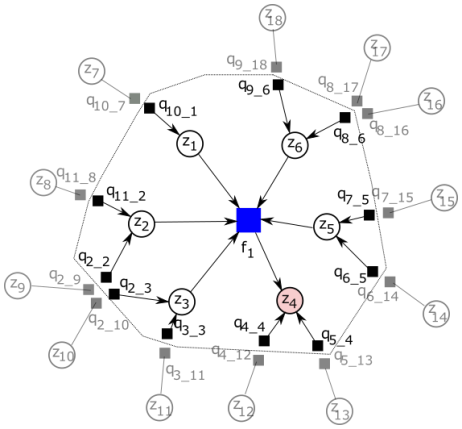

Figure 4: Message Passing to Marginalize $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$, $\boldsymbol{z}_5$, and $\boldsymbol{z}_6$ for $\boldsymbol{z}_4$

$$
\begin{aligned}
q_{new4}(\boldsymbol{z}_4) &= \sum_{\backslash \boldsymbol{z}_4} f_1(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4, \boldsymbol{z}_5, \boldsymbol{z}_6) q_{\backslash 1}(Z) \\
&= \sum_{\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_5, \boldsymbol{z}_6} f_1(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4, \boldsymbol{z}_5, \boldsymbol{z}_6) \\
&\qquad q_{4\_4}(\boldsymbol{z}_4) q_{5\_4}(\boldsymbol{z}_4) q_{10\_1}(\boldsymbol{z}_1) \\
&\qquad q_{11\_2}(\boldsymbol{z}_2) q_{2\_2}(\boldsymbol{z}_2) q_{2\_3}(\boldsymbol{z}_3) q_{3\_3}(\boldsymbol{z}_3) \\
&\qquad q_{6\_5}(\boldsymbol{z}_5) q_{7\_5}(\boldsymbol{z}_5) q_{8\_6}(\boldsymbol{z}_6) q_{9\_6}(\boldsymbol{z}_6)
\end{aligned} \tag{52}
$$

$q_{new4}(\boldsymbol{z}_4)$ corresponds to the improved $q_{1\_4}(\boldsymbol{z}_4) q_{4\_4}(\boldsymbol{z}_4) q_{5\_4}(\boldsymbol{z}_4)$. Hence, the improved $q_{1\_4}(\boldsymbol{z}_4)$ will be

$$
q_{1\_4}(\boldsymbol{z}_4) = \frac{q_{new4}(\boldsymbol{z}_4)}{q_{4\_4}(\boldsymbol{z}_4) q_{5\_4}(\boldsymbol{z}_4)} \tag{53}
$$

In the same way, we find $q_{new1}(\boldsymbol{z}_1)$, $q_{new2}(\boldsymbol{z}_2)$, $q_{new3}(\boldsymbol{z}_3)$, $q_{new5}(\boldsymbol{z}_5)$, and $q_{new6}(\boldsymbol{z}_6)$ by message passing, and then we can find

$$
\begin{aligned}
q_{1\_1}(\boldsymbol{z}_1) &= \frac{q_{new1}(\boldsymbol{z}_1)}{q_{10\_1}(\boldsymbol{z}_1)} \\
q_{1\_2}(\boldsymbol{z}_2) &= \frac{q_{new2}(\boldsymbol{z}_2)}{q_{11\_2}(\boldsymbol{z}_2) q_{2\_2}(\boldsymbol{z}_2)} \\
q_{1\_3}(\boldsymbol{z}_3) &= \frac{q_{new3}(\boldsymbol{z}_3)}{q_{2\_3}(\boldsymbol{z}_3) q_{3\_3}(\boldsymbol{z}_3)} \\
q_{1\_5}(\boldsymbol{z}_5) &= \frac{q_{new5}(\boldsymbol{z}_5)}{q_{6\_5}(\boldsymbol{z}_5) q_{7\_5}(\boldsymbol{z}_5)} \\
q_{1\_6}(\boldsymbol{z}_6) &= \frac{q_{new6}(\boldsymbol{z}_6)}{q_{8\_6}(\boldsymbol{z}_6) q_{9\_6}(\boldsymbol{z}_6)}
\end{aligned} \tag{54}
$$

Then finally, the improved $q_1(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4, \boldsymbol{z}_5, \boldsymbol{z}_6)$ will be:

$$q_1(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4, \boldsymbol{z}_5, \boldsymbol{z}_6) = q_{1\_1}(\boldsymbol{z}_1) q_{1\_2}(\boldsymbol{z}_2) q_{1\_3}(\boldsymbol{z}_3) q_{1\_4}(\boldsymbol{z}_4) q_{1\_5}(\boldsymbol{z}_5) q_{1\_6}(\boldsymbol{z}_6) \tag{55}$$

This concludes the iteration for the approximator $q_1(Z_1)$ for factor $f_1(Z_1)$. We can repeat the process for all the factors until it converges (if it converges).

References

[1] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2011.

[2] Simon Barthelmé. Youtube: The expectation-propagation algorithm: a tutorial - part1. https://youtu.be/0tomU1q3AdY. Accessed: 2020-03-30.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[4] Antti Honkela. Introduction to expectation propagation ; helsinki university of technology, espoo, finland. http://www.cis.hut.fi/ahonkela/. Accessed: 2020-03-30.

[5] Arnab Laha. Stats stackexchange : Central moments of a gaussian mixture density? https://stats.stackexchange.com/questions/32699/central-moments-of-a-gaussian-mixture-density. Accessed: 2020-03-30.

[6] Thomas P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI01, page 362369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.