Expectation Propagation Algorithm

Shuang Wang
School of Electrical and Computer Engineering
University of Oklahoma, Tulsa, OK, 74135
Email: {shuangwang}@ou.edu

This note contains three parts. First, we will review some preliminaries for EP. Then, EP algorithm will be described in the next section. Finally, the relationship between EP and other variational methods will be discussed.

I. Preliminaries

A. Exponential family

The exponential family of distributions over $\mathbf{x}$ is a set of distributions with the form

$$p(\mathbf{x};\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta})\exp\left(\boldsymbol{\theta}^T\mathbf{u}(\mathbf{x})\right), \tag{1}$$

where measurement $\mathbf{x}$ may be scalar or vector, discrete or continuous, $\boldsymbol{\theta}$ are parameters of the distribution, $h(\mathbf{x})$ and $\mathbf{u}(\mathbf{x})$ are some functions of $\mathbf{x}$, and the function $g(\boldsymbol{\theta})$ is a normalization factor as

$$g(\boldsymbol{\theta})\int h(\mathbf{x})\exp\left(\boldsymbol{\theta}^T\mathbf{u}(\mathbf{x})\right)d\mathbf{x} = 1. \tag{2}$$

In addition, if the variables are discrete, just simply replace the integration with summation.

Exponential family has many properties, which may simplify computations. For example, if a likelihood function is one of members in the exponential family, the posterior can be expressed in a closed-form expression by choosing a conjugate prior within the exponential family. Moreover, exponential family has a wide range of members such as Gaussian, Bernoulli, discrete multinomial, Poisson and so on, thus it is applicable to many different inference models.

B. Kullback-Leibler divergence

Kullback-Leibler (KL) divergence [1] is a measure to quantify the difference between a probabilistic distributions $p(\mathbf{x})$ and an approximate distribution $q(\mathbf{x})$. For the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ over continuous variables, KL divergence is defined as

$$D_{KL}(p(\mathbf{x})\|q(\mathbf{x})) = \int p(\mathbf{x})\log\frac{p(\mathbf{x})}{q(\mathbf{x})}d\mathbf{x}, \tag{3}$$

where for discrete variables, just replace integration with summation. Moreover, KL divergence is a non-symmetric measure, which means $D_{KL}(p(\mathbf{x})\|q(\mathbf{x})) \neq D_{KL}(q(\mathbf{x})\|p(\mathbf{x}))$. To give readers an intuitive view about the difference between the above two forms of KL divergence, we assume that the true distribution $p(\mathbf{x})$ is multimodal and the candidate distribution $q(\mathbf{x})$ is unimodal. By minimizing $D_{KL}(q(\mathbf{x})\|p(\mathbf{x}))$, the approximate distribution $q(\mathbf{x})$ will pick one of the modes in $p(\mathbf{x})$, which is usually used in variational Bayes method. However, the best approximate distribution $q(\mathbf{x})$ obtained by minimizing $D_{KL}(p(\mathbf{x})\|q(\mathbf{x}))$ will be the average of all modes. The later case is used in the approximate inference procedure of EP. Since this report focus on the review of EP algorithm, we will study the property of minimizing $D_{KL}(p(\mathbf{x})\|q(\mathbf{x}))$ first. Regarding the difference between minimizing $D_{KL}(p(\mathbf{x})\|q(\mathbf{x}))$ and $D_{KL}(q(\mathbf{x})\|p(\mathbf{x}))$, we will discuss it later in this chapter.

To ensure a tractable solution for minimizing KL divergence $D_{KL}(p(\mathbf{x})\|q(\mathbf{x}))$, the approximate distribution $q(\mathbf{x})$ is usually restricted within a member of the exponential family. Thus, according to (1), $q(\mathbf{x})$ can be written as

$$q(\mathbf{x};\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta})\text{exp}\left(\boldsymbol{\theta}^T\mathbf{u}(\mathbf{x})\right), \tag{4}$$

where $\boldsymbol{\theta}$ are the parameters of the given distribution.

By substituting $q(\mathbf{x};\boldsymbol{\theta})$ into the KL divergence $D_{KL}(p(\mathbf{x})\|q(\mathbf{x}))$, we get

$$D_{KL}(p(\mathbf{x})\|q(\mathbf{x})) = -\text{ln}g(\boldsymbol{\theta}) - \boldsymbol{\theta}^T\mathbb{E}_{p(\mathbf{x})}[\mathbf{u}(\mathbf{x})] + \text{const}, \tag{5}$$

where the const represents all the terms that are independent of parameters $\boldsymbol{\theta}$. To minimize KL divergence, take the gradient of $D_{KL}(p(\mathbf{x})\|q(\mathbf{x}))$ with respect to $\boldsymbol{\theta}$ to zero, we get

$$-\bigtriangledown\text{ln}g(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{x})}[\mathbf{u}(\mathbf{x})]. \tag{6}$$

Moreover, for (2), taking the gradient of both sides respect to $\boldsymbol{\theta}$, we get

$$\bigtriangledown g(\boldsymbol{\theta})\int h(\mathbf{x})\text{exp}\left\{\boldsymbol{\theta}^T\mathbf{u}(\mathbf{x})\right\}d\mathbf{x} + g(\boldsymbol{\theta})\int h(\mathbf{x})\text{exp}\left\{\boldsymbol{\theta}^T\mathbf{u}(\mathbf{x})\right\}\mathbf{u}(\mathbf{x})d\mathbf{x} = 0. \tag{7}$$

Then by rearranging and reusing (2) again, we get

$$-\bigtriangledown\text{ln}g(\boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{x})}[\mathbf{u}(\mathbf{x})]. \tag{8}$$

By comparing (6) and (8), we obtain

$$\mathbb{E}_{p(\mathbf{x})}[\mathbf{u}(\mathbf{x})] = \mathbb{E}_{q(\mathbf{x})}[\mathbf{u}(\mathbf{x})]. \tag{9}$$

Thus, from (9), we see that the minimization of KL divergence is equivalent to matching the expected sufficient statistics. For example, for minimizing KL divergence with a Gaussian distribution $q(\mathbf{x};\boldsymbol{\theta})$, we only need to find the mean and covariance of $q(\mathbf{x};\boldsymbol{\theta})$ that are equal to the mean and covariance of $p(\mathbf{x};\boldsymbol{\theta})$, respectively.

C. *Assumed-density filtering (ADF)*

ADF is a technique to construct tractable approximation to complex probability distribution. EP can be viewed as an extension on ADF. Thus, we first provide a concise review of ADF and then extend it to EP algorithm.

Let us consider the Bayes's rule and suppose that the factorization of posterior distribution has the following form

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \frac{p(\mathbf{x})p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} \\ &= \frac{1}{Z}p_0(\mathbf{x})\prod_{i=1}^{N} p(y_i|\mathbf{x}), \\ &= \frac{1}{Z}\prod_{i=0}^{N} p_i(\mathbf{x}), \end{aligned} \tag{10}$$

where Z is a normalization constant, $p_i(\mathbf{x})$ is a simplified notation of each corresponding factor in (10), where $p_0(\mathbf{x}) = p_0(\mathbf{x})$ and $p_i(\mathbf{x}) = p_i(y_i|\mathbf{x})$ for $i > 0$. If we assume that the likelihood function $p(y_i|\mathbf{x})$ has a complex form, the direct evaluation of the posterior distribution would be infeasible. For example, if each likelihood function is a mixture of two Gaussian distributions and there are total $N = 100$ number of observed data. Then to get the posterior distribution, we need to evaluate mixture of 2^{100} Gaussians.

To solve this problem, approximate inference methods try to seek an approximate posterior distribution that can be very close to the true posterior distribution $p(\mathbf{x}|\mathbf{y})$. Usually, the approximate distributions are chosen within the exponential family to ensure the computational feasibility. Then the best approximate distribution can be found by minimizing KL divergence:

$$\boldsymbol{\theta}^* = \arg\min_{\theta} D_{KL}(p(\mathbf{x})\|q(\mathbf{x};\boldsymbol{\theta})). \tag{11}$$

However, we can see that it is difficult to solve (11) directly. ADF solves this problem by iteratively including each factor function in the true posterior distribution. Thus, at first, ADF chooses $q(\mathbf{x};\boldsymbol{\theta}^0)$ to best approximate factor function $p_0(\mathbf{x})$ through

$$\boldsymbol{\theta}^0 = \arg\min_{\theta} D_{KL}(p_0(\mathbf{x})\|q(\mathbf{x};\boldsymbol{\theta})). \tag{12}$$

Then ADF will update the approximation by incorporating the next factor function $p_i(y_i|\mathbf{x})$ until all the factor functions have been involved, which gives us the following update rule

$$\boldsymbol{\theta}^i = \arg\min_{\theta} D_{KL}(p_i(\mathbf{x})q(\mathbf{x};\boldsymbol{\theta}^{i-1})\|q(\mathbf{x};\boldsymbol{\theta})). \tag{13}$$

As shown in Section I-B, if $q(\mathbf{x};\boldsymbol{\theta})$ is chosen from the exponential family, the optimal solution of (13) is matching the expected sufficient statistics between the approximate distribution $q(\mathbf{x};\boldsymbol{\theta}^i)$ and the target distribution $p_i(\mathbf{x})q(\mathbf{x};\boldsymbol{\theta}^{i-1})$. Moreover, according to (13), we can see that the current best approximation is based on the previous best approximation. For this reason, the estimation performance of ADF may be sensitive to the process order of factor functions, which may produce extremely poor approximation in some cases. In the next section, we will provide another perspective of the ADF update rule, which results the EP algorithm and provides a way to avoid the drawback associated with ADF algorithm.

II. Expectation Propagation

By taking another perspective, ADF can be seen as sequentially approximating the factor function $p_i(\mathbf{x})$ by the approximate factor function $\tilde{p_i}(\mathbf{x})$, which is restricted within the exponential family, and then exactly updating the approximate distribution $q(\mathbf{x};\boldsymbol{\theta})$ by multiplying these approximate factor functions. This alternative view of ADF can be described as:

$$\tilde{p_i}(\mathbf{x}) \propto \frac{q(\mathbf{x};\boldsymbol{\theta}^i)}{q(\mathbf{x};\boldsymbol{\theta}^{i-1})}, \tag{14}$$

which also produces the EP algorithm. EP algorithm initializes each factor function $p_i(\mathbf{x})$ by a corresponding approximate factor function $\tilde{p_i}(\mathbf{x})$. Then, at later iterations, EP revisits each approximated factor function $\tilde{p_i}(\mathbf{x})$ and refined it by multiplying all the current best estimate but one true factor function $p_i(\mathbf{x})$ of the revisiting term. After multiple iterations, the approximation is obtained according (15).

$$q(\mathbf{x};\boldsymbol{\theta}^*) \propto \prod_i \tilde{p_i}(\mathbf{x}). \tag{15}$$

Since this procedure does not depend on the process order of the factor function, EP provides a more accurate approximation than ADF.

The general process of EP is given as follows:

1) Initialize the term approximation $\tilde{p_i}(\mathbf{x})$, which can be chosen from one of members in the exponential family based on the problem.
2) Compute the approximate distribution
$$q(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z}\prod_i \tilde{p_i}(\mathbf{x}), \tag{16}$$
where $Z = \int \prod_i \tilde{p_i}(\mathbf{x})d\mathbf{x}$.
3) Until all $\tilde{p_i}(\mathbf{x})$ converge:
   a) Choose $\tilde{p_i}(\mathbf{x})$ to refine the approximate.
   b) Remove $\tilde{p_i}(\mathbf{x})$ from the current approximated distribution $q(\mathbf{x};\boldsymbol{\theta})$ with a normalization factor:
   $$q(\mathbf{x};\boldsymbol{\theta}^{\backslash i}) \propto \frac{q(\mathbf{x};\boldsymbol{\theta})}{\tilde{p_i}(\mathbf{x})}. \tag{17}$$
   c) Update $q(\mathbf{x};\boldsymbol{\theta})$, where we first combine $q(\mathbf{x};\boldsymbol{\theta}^{\backslash i})$, current $p_i(\mathbf{x})$ and a normalizer Z_i, and then minimize the KL divergence through moment matching projection (9) (i.e. the $\textbf{Proj}(\cdot)$ operator):
   $$q(\mathbf{x};\boldsymbol{\theta}) = \textbf{Proj}\left(\frac{1}{Z_i}q(\mathbf{x};\boldsymbol{\theta}^{\backslash i})p_i(\mathbf{x})\right). \tag{18}$$
   d) Update $\tilde{p_i}(\mathbf{x})$ as
   $$\tilde{p_i}(\mathbf{x}) = Z_i\frac{q(\mathbf{x};\boldsymbol{\theta})}{q(\mathbf{x};\boldsymbol{\theta}^{\backslash i})}. \tag{19}$$
4) Get the final approximate distribution through
$$p(\mathbf{x}) \approx q(\mathbf{x};\boldsymbol{\theta}^*) \propto \prod_i \tilde{p_i}(\mathbf{x}). \tag{20}$$

A. Relationship with BP

This section shows that BP algorithm is a special case of EP, where EP provides an improved approximation for models in which BP is generally intractable.

Let us first take a quick review of BP algorithm. The procedure of BP algorithm is iteratively updating all variables nodes, then updating all factor nodes through sending messages in parallel, and finally update the belief of each variable after each iteration. By taking another viewpoint, BP can be viewed as updating the belief over a variable x_i by incorporating one factor node at each time. Under this perspective, the belief of variable x_i will be updated as

$$b(x_i) = \frac{m_{X_i \to f_s}(x_i)\, m_{f_s \to X_i}(x_i)}{Z_i}, \tag{21}$$

where $Z_i = \int m_{X_i \to f_s}(x_i)\, m_{f_s \to X_i}(x_i)\, dx_i$ is the normalization factor. Moreover, we can loosely interpret $m_{X_i \to f_s}(x_i)$ and $m_{f_s \to X_i}(x_i)$ as the prior and likelihood message, respectively.

Let us suppose that each likelihood message $m_{f_s \to X_i}(x_i)$ has a complex form, e.g. a mixture of multiple Gaussian distributions. Then the computational complexity to evaluate the exact beliefs over all variables would be infeasible. Instead of propagating exact likelihood message $m_{f_s \to X_i}(x_i)$, EP passes an approximate message $\tilde{m}_{f_s \to X_i}(x_i)$, where $\tilde{m}_{f_s \to X_i}(x_i)$ is obtained by using the projection operation as shown in the general process of EP. Moreover, $\tilde{m}_{f_s \to X_i}(x_i)$ is usually chosen from exponential family to make the problem tractable. Thus, the approximate belief in EP has the following form

$$b(x_i) \approx q(x_i) \propto \prod_{s \in N(X_i)} \tilde{m}_{f_s \to X_i}(x_i). \tag{22}$$

To show BP as a special case of EP, we further define the partial belief of a variable node as

$$b(x_i)^{\backslash f_s} = \frac{b(x_i)}{\tilde{m}_{f_s \to X_i}(x_i)} \propto \prod_{s' \in N(X_i) \backslash s} \tilde{m}_{f_{s'} \to X_i}(x_i), \tag{23}$$

and the partial belief of a factor node as

$$b(f_s)^{\backslash X_i} = \frac{b(f_s)}{\tilde{m}_{X_i \to f_s}(x_i)}, \tag{24}$$

where $b(f_s) = \prod_{j \in N(f_s)} \tilde{m}_{X_j \to f_s}(x_j)$ is define as the belief of the factor node f_s. By comparing to (18) and (19), the factor node message updating rule in EP can be written as

$$\begin{aligned} \tilde{m}_{f_s \to X_i}(x_i) &= \frac{\textbf{Proj}\left(b(x_i)^{\backslash f_s} m_{f_s \to X_i}(x_i)\right)}{b(x_i)^{\backslash f_s}} \\ &= \frac{\textbf{Proj}\left(b(x_i)^{\backslash f_s} \int_{\mathbf{x}_s \backslash x_i} f_s(\mathbf{x}_s)\, b(f_s)^{\backslash X_i}\right)}{b(x_i)^{\backslash f_s}} \end{aligned} \tag{25}$$

where the integral works over continuous variable. For discrete variable, one can simply replace integral with summation. Furthermore, the new belief $b(x_i)$ will be approximated as

$$b(x_i) \approx q_i(x_i) = \frac{b(x_i)^{\backslash f_s} \tilde{m}_{f_s \to X_i}(x_i)}{Z_i}, \tag{26}$$

where $Z_i = \int_{x_i} b(x_i)^{\backslash f_s} \tilde{m}_{f_s \to X_i}(x_i)$.

Now, if the integral in (25) is tractable (e.g. a linear Gaussian model) even without using the projection to approximate $m_{f_s \to X_i}$. Then $b(x_i)^{\backslash f_s}$ in (25) can be canceled. Finally, the factor node message update rule in EP reduces to the standard BP case.

III. Relationship With Other Variational Inference Methods

In this section, we will describe the relationship between EP and other variational inference algorithms, e.g. variational Bayes (VB). The Bayesian probabilistic model specifies the joint distribution $p(\mathbf{x}, \mathbf{y})$, where all the hidden variables in $\mathbf{x}$ are given prior distributions. The goal is to find the best approximation for the posterior distribution $p(\mathbf{x}|\mathbf{y})$. Let us take a look at the decomposition of the log joint distribution as follows

$$\log p(\mathbf{x}, \mathbf{y}) = \log p(\mathbf{x}|\mathbf{y}) + \log p(\mathbf{y}). \tag{27}$$

By rearranging (27) and taking the integral of the both side of the rearranged equation with respect to a given distribution $q(\mathbf{x})$, we get the log model evidence

$$\begin{aligned} \log p(\mathbf{y}) &= \int q(\mathbf{x}) \log(p(\mathbf{y})) d\mathbf{x} \\ &= \int q(\mathbf{x}) \log(p(\mathbf{x}, \mathbf{y})) - \int q(\mathbf{x}) \log(p(\mathbf{x}|\mathbf{y})) d\mathbf{x}, \end{aligned} \tag{28}$$

where $\int q(\mathbf{x}) d\mathbf{x} = 1$. Then, by reformatting (28), we get

$$\log p(\mathbf{y}) = \mathcal{L}(q(\mathbf{x})) + D_{KL}(q(\mathbf{x})||p(\mathbf{x})), \tag{29}$$

where we define

$$\mathcal{L}(q(\mathbf{x})) = \int q(\mathbf{x}) \log(\frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{x})}) d\mathbf{x}, \tag{30}$$

$$D_{KL}(q(\mathbf{x})||p(\mathbf{x})) = \int q(\mathbf{x}) \log(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})}) d\mathbf{x}. \tag{31}$$

Since $D_{KL}(q(\mathbf{x})||p(\mathbf{x}))$ is a nonnegative functional, $\mathcal{L}(q(x))$ gives the lower bound of $\log p(\mathbf{y})$. Then the maximization of the lower bound $\mathcal{L}(q(x))$ with respect to the distribution $q(\mathbf{x})$ is equivalent to minimizing $D_{KL}(q(\mathbf{x})||p(\mathbf{x}))$, which happens when $q(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})$. However, working with the true posterior distribution $p(\mathbf{x}|\mathbf{y})$ may be intractable. Thus, we assume that the elements of $\mathbf{x}$ can be partitioned into M disjoint groups $\mathbf{x}_i$, $i = 1, 2, \cdots, M$. We then further assume that the factorization of the approximate distribution $q(\mathbf{x})$ with respect to these group has the form

$$q(\mathbf{x}) = \prod_{i}^{M} q_i(\mathbf{x}_i). \tag{32}$$

Please note that the factorized approximation corresponds to the mean filed theory, which was developed in physics. Given the aforementioned assumptions, we now try to find any possible distribution $q(\mathbf{x})$ over which the lower bound $\mathcal{L}(q(\mathbf{x}))$ is largest. Since the direct maximization of (30) with respect to $q(\mathbf{x})$ is difficult, we instead to optimize (30) with respect to each of the factors in (32). By substituting (32) into (30), we get

$$\begin{aligned}\mathcal{L}(q(\mathbf{x})) &= \int q_j(\mathbf{x}_j) \left(\int \log\left(p(\mathbf{x}, \mathbf{y})\right) \prod_{i \neq j} q_i(\mathbf{x}_i) d\mathbf{x}_i \right) d\mathbf{x}_j - \int q_j(\mathbf{x}_j) \log(q_j(\mathbf{x}_j)) d\mathbf{x}_j + \text{const} \\ &= - \int q_j(\mathbf{x}_j) \log(\frac{q_j(\mathbf{x}_j)}{\tilde{p}(\mathbf{x}_j, \mathbf{y})}) d\mathbf{x_j} + \text{const} \\ &= -D_{KL}(q_j(\mathbf{x}_j) || \tilde{p}(\mathbf{x}_j, \mathbf{y})) + \text{const},\end{aligned} \tag{33}$$

where we define $\tilde{p}(\mathbf{x}_j, \mathbf{y})$ as

$$\begin{aligned}\tilde{p}(\mathbf{x}_j, \mathbf{y}) &= \exp\left(\int \log\left(p(\mathbf{x}, \mathbf{y})\right) \prod_{i \neq j} q_i(\mathbf{x}_i) d\mathbf{x}_i \right) \\ &= \exp\left(\mathbb{E}_{i \neq j}[\log p(\mathbf{x}, \mathbf{y})]\right).\end{aligned} \tag{34}$$

Thus, if we keep all the factors $q_i(\mathbf{x}_i)$ for $i \neq j$ fixed, then the maximization of (33) with respect to $q_j(\mathbf{x}_j)$ is equivalent to the minimization of $D_{KL}(q_j(\mathbf{x}_j)||\tilde{p}(\mathbf{x}_j, \mathbf{y}))$. In practices, we need to initialize all of the factors $q_i(\mathbf{x}_i)$ first, and then iteratively update each of the factor $q_j(\mathbf{x}_j)$ by minimizing the $D_{KL}(q_j(\mathbf{x}_j)||\tilde{p}(\mathbf{x}_j, \mathbf{y}))$, until the algorithm convergences.

Now we can see the key difference between EP and VB is the way to minimizing the KL divergence. The advantage of VB is that it provides a lower bound during each optimizing step, thus the convergence is guaranteed. However, VB may cause under-estimate for variance. In EP, minimizing $D_{KL}(p(\mathbf{x})||q(\mathbf{x}))$ is equivalence to the "moment matching", but convergence is not guaranteed. However, EP has a fix point and if it does converge, the approximation performance of EP usually outperforms VB.

References

[1] S. Kullback and R. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[2] C. Bishop, *Pattern recognition and machine learning*. Springer New York, 2006, vol. 4.

[3] T. Minka, "Expectation propagation for approximate Bayesian inference," in *Uncertainty in Artificial Intelligence*, vol. 17. Citeseer, 2001, pp. 362–369.

[4] ——, *http://research.microsoft.com/en-us/um/people/minka/papers/ep/roadmap.html.*