Expectation Propagation

October 28, 2020

1 Objective

Expectation propagation [3] aims to provide tractable approximations to complex probability density functions of the form

$$p(x) \propto \prod_{i=1}^{T} \psi_i(x) \tag{1}$$

where $\psi_i(x)$ are called *factors* or *compatibility functions*. Note that the form in (2) is similar to the one in Forney style factor graph [1]. Densities of the form (2) appear in the context of Bayesian inference

$$\pi_n(\theta) \propto \pi(\theta) \prod_{i=1}^{n} f(x_i \mid \theta) \tag{2}$$

where π_n is the posterior obtained with π as the prior and $f(x \mid \theta)$ as the likelihood.

Similar to variational approximation, the approximating family is restricted to a class of probability distributions and a proper divergence measure is chosen to obtain the best approximation. Unlike variational inference, EP solves the optimization problem

$$\hat{q} := \arg\ \min_{q \in \Gamma} \mathrm{D}(p \,\|\, q), \tag{3}$$

where $\mathrm{D}(p \,\|\, q) := \int p(x) \log\{p(x)/q(x)\} dx$. In this note, we review algorithms to solve (3) under appropriate assumptions on Γ. We also review connections of EP with Belief propagation and Bethe free energy minimization in the context of probabilistic graphical models.

2 EP algorithms

2.1 Γ is an exponential family

In this case, substantial simplifications can be made towards solving (3). Consider the following parameterization of q

$$q(x \mid \theta) = \exp\left\{\langle \theta, t(x) \rangle - A(\theta)\right\}$$

where θ is the natural exponential family and $A(\cdot)$ is the log-partition function. Then (3) is equivalent to solving

$$\begin{aligned}\hat{\theta} &:= \arg\max_{\theta\in\Theta} \int p(x)\log q(x \mid \theta)dx \\ &= \arg\max_{\theta\in\Theta} \Big[\int \langle\theta, t(x)\rangle p(x)dx - A(\theta)\Big].\end{aligned}$$

If $\Theta \subset \mathbb{R}^p$, a stationary equation of the above maximization problem is given by

$$\int t(x)p(x)dx = \frac{\partial A(\theta)}{\partial\theta} = \int t(x)q(x \mid \theta)dx$$

which is equivalent to *moment-matching.* Note that in the above, we did not use the specific form of the target function (2). In the following, we shall derive an iterative procedure that aims to utilize this structure.

2.2 Special case: Assumed density filtering for parametric model fitting

One can iteratively refine the solution by choosing $q(x, \theta^1)$ to best approximate the first compatibility function $\psi_1(x)$ in the sense

$$\theta^1 = \arg\min_\theta \mathrm{D}(\psi_1(x) \,\|\, q(x;\theta)).$$

The subsequent approximates are obtained as

$$\theta^i = \arg\min_\theta \mathrm{D}(\psi_i(x)q(x;\theta^{i-1}) \,\|\, q(x;\theta)),$$

Intuitively, the algorithm first approximates $\psi_1(x)$ by $q(x;\theta^1)$ and then approximates $\psi_2(x)q(x;\theta^1)$ by $q(x;\theta^2)$ and so on. While approximating this cascading product may be preferable to constructing independent approximates to each term, it has the undesirable property of being very sensitive to the order in which the compatibility terms are processed.

2.3 EP algorithm: Assumed density filtering for fitting a factor graph to another factor graph

Instead of fitting parametric family, we assume

$$q(x) \propto \prod_{i=1}^{T} \tilde{t}_i(x).$$

So the problem (3) is equivalent to minimizing

$$\mathrm{D}\Big\{\frac{\prod_{i=1}^T \psi_i(x)}{Z_0} \,\Big\|\, \frac{\prod_{i=1}^T \tilde{t}_i(x)}{Z}\Big\}$$

Starting with an initial value of $\tilde{t}_i(x)$ for $i = 1, \ldots, T$, we run through the following. For fixed i, we remove the effect of the factor $\tilde{t}_i(x)$ from q forming the *cavity distribution*,

$$q^{\backslash i}(x) = \frac{q(x)}{\tilde{t}_i(x)}. \tag{4}$$

and update the factor $\tilde{t}_i(x)$ by

$$\text{arg min}_q \quad \mathrm{D}\Big\{ \frac{q^{\backslash i}(x)\psi_i(x)}{Z_{0i}} \Big\| q(x) \Big\}.$$

We rename this operation using the projection operation

$$\hat{q}(x) = \text{proj}\{q^{\backslash i}(x)\psi_i(x)\}. \tag{5}$$

Then update $\tilde{t}_i$ by an *anti-cavitating* operation:

$$\tilde{t}_i^{\text{new}}(x) = K_i \frac{\hat{q}(x)}{q^{\backslash i}(x)} \tag{6}$$

where the coefficient K is determined by multiplying both sides of (6) by $q^{\backslash i}(x)$

$$\int \tilde{t}_i^{\text{new}}(x) q^{\backslash i}(x) = K_i \int \hat{q}(x) dx = K_i. \tag{7}$$

The division operation in (4) amounts to subtracting natural parameters for exponential family. The projection operation (5) is an M-projection operation in information geometry [2]. The projection is equivalent to matching the sufficient statistics of $\tilde{t}_i(x) q^{\backslash i}(x)$ to those of $t_i(x) q^{\backslash i}(x)$. In particular for Gaussian $\tilde{t}_i(x)$, matching the sufficient statistics is equivalent to matching the zeroth, first and the second moments. Thus

$$K_i = \int \tilde{t}_i^{\text{new}}(x) q^{\backslash i}(x) dx = \int \psi_i(x) q^{\backslash i}(x) dx = Z_{0i}. \tag{8}$$

3 Summary of the algorithm

Input: $\pi(x) = \prod_i \psi_i(x)$.

Desired output: Approximation for $Z_0 = \int \prod_i \psi_i(x) dx$.

Approximator: $q(x) \propto \prod_i \tilde{t}_i(x)$.

Initialization:

1. Initialize all factors $\tilde{t}_i(x)$

2. Initialize the EP approximation $q(x) = \prod_i \tilde{t}_i(x)$

Iterate until convergence of $\hat{Z}_0$:

1. Choose a factor $\tilde{t}_i(x)$ to update
2. Construct cavity distribution according to (4): $q^{\backslash i}(x) = \frac{q(x)}{\tilde{t}_i(x)}$.
3. Construct the projection: $\hat{q}(x) = \text{proj}\{q^{\backslash i}(x)\psi_i(x)\}$, i.e., evaluate the new approximation $\hat{q}$ by setting the sufficient statistics of $\hat{q}$ equal to that of $\psi_i(x)q^{\backslash i}(x)$ and evaluate $Z_{0i} = K_i = \int \psi_i(x)q^{\backslash i}(x)dx$.
4. Construct updated factor according to (6): $\tilde{t}_i^{\text{new}}(x) = K_i \frac{\hat{q}(x)}{q^{\backslash i}(x)}$
5. Compute $\hat{Z}_0 = \int \prod_i \tilde{t}_i(x)dx$.

3.1 An example

Consider the clutter problem, where

$$p(x \mid \theta) = (1-\pi)\text{N}(x;\theta,\mathbb{I}) + \pi\text{N}(x;0,a\mathbb{I}), x \in \mathbb{R}^d$$

where the goal is to recover the mean $\theta \in \mathbb{R}^d$. We assume (a,π) to be known. A Bayesian inference proceeds with specifying $p(\theta) = \text{N}(\theta;0,b\mathbb{I})$ and obtaining the posterior distribution of θ from the Bayes theorem

$$p(\theta \mid \mathcal{D}) \propto p(\theta)\prod_{i=1}^{n} p(x_i \mid \theta).$$

To apply EP, note that

$$\psi_0(\theta) = p(\theta), \quad \psi_i(\theta) = (1-\pi)\text{N}(x_i;\theta,\mathbb{I}) + \pi\text{N}(x_i;0,a\mathbb{I}), \quad i = 1,\ldots,n.$$

We approximate the posterior distribution by $q(\theta) = \text{N}(\theta;m,\nu\mathbb{I})$, where we employ ADF to find optimal values of the parameters m and ν. We assume that the factors take the form

$$\boxed{\tilde{t}_i(\theta) = s_i\text{N}(\theta;m_i,\nu_i\mathbb{I})}.$$

where we take $s_i = (2\pi\nu_i)^{d/2}$ so that $\tilde{t}_i(\theta)$ is unnormalized. Note that ν_i can be negative. Next, note that the cavity distribution is given by

$$\boxed{q^{\backslash i}(\theta) = \text{N}(\theta;m^{\backslash i},\nu^{\backslash i}\mathbb{I})},$$

where $m^{\backslash i} = m + \nu^{\backslash i}\nu_i^{-1}(m - m_i)$ and $(\nu^{\backslash i})^{-1} = \nu^{-1} - \nu_i^{-1}$. We also obtained an expression for Z_{0i} for $i = 1, \ldots, n$ below.

$$\begin{aligned} Z_{0i} &= \int \psi_i(\theta) q^{\backslash i}(\theta) d\theta = \int \mathrm{N}(\theta; m^{\backslash i}, \nu^{\backslash i}\mathbb{I})[(1-\pi)\mathrm{N}(x_i; \theta, \mathbb{I}) + \pi \mathrm{N}(x_i; 0, a\mathbb{I})]d\theta. \\ &= (1-\pi)\mathrm{N}(x_i; m^{\backslash i}, (\nu^{\backslash i} + 1)\mathbb{I}) + \pi \mathrm{N}(x_i; 0, a\mathbb{I}) \\ &:= (1-\pi)Z_{0i}^{(1)} + \pi Z_{0i}^{(2)}. \end{aligned}$$

Now to perform the projection operation, we find the mean and variance of the density $Z_{0i}^{-1}\psi_i(\theta)q^{\backslash i}(\theta)$.

$$\begin{aligned} q^{\backslash i}(\theta)\psi_i(\theta) &= Z_{0i}^{(1)}(1-\pi)\frac{\mathrm{N}(x_i; \theta, \mathbb{I})\mathrm{N}(\theta; m^{\backslash i}, \nu^{\backslash i})}{Z_{0i}^{(1)}} + \pi Z_{0i}^{(2)}\frac{\mathrm{N}(x_i; 0, a\mathbb{I})\mathrm{N}(\theta; m^{\backslash i}, \nu^{\backslash i})}{Z_{0i}^{(2)}} \\ &:= Z_{0i}^{(1)}(1-\pi)\pi_{1i}(\theta \mid x_i) + \pi Z_{0i}^{(2)}\pi_{2i}(\theta \mid x_i). \end{aligned}$$

where

$$\pi_{1i}(\theta \mid x_i) \equiv \mathrm{N}\left\{\theta; \frac{\nu^{\backslash i}x_i + m^{\backslash i}}{\nu^{\backslash i} + 1}, \frac{\nu^{\backslash i}}{\nu^{\backslash i} + 1}\mathbb{I}\right\}, \quad \pi_{2i}(\theta \mid x_i) \equiv \mathrm{N}(\theta; m^{\backslash i}, \nu^{\backslash i}).$$

$\hat{q}(\theta)$ is then a Gaussian distribution with mean given by

$$\begin{aligned} m &= Z_{0i}^{-1}Z_{0i}^{(1)}(1-\pi)\Big[\frac{\nu^{\backslash i}x_i + m^{\backslash i}}{\nu^{\backslash i} + 1}\Big] + Z_{0i}^{-1}Z_{0i}^{(2)}\pi \mathrm{N}(x_i; 0, a\mathbb{I})m^{\backslash i}, \\ &= \hat{\pi}_i\Big[\frac{\nu^{\backslash i}x_i + m^{\backslash i}}{\nu^{\backslash i} + 1}\Big] + m^{\backslash i}(1 - \hat{\pi}_i) \\ &= m^{\backslash i} + \hat{\pi}_i\frac{\nu^{\backslash i}}{\nu^{\backslash i} + 1}(x_i - m^{\backslash i}), \end{aligned}$$

where $\hat{\pi}_i$ can be interpreted as the posterior exclusion probability of not being in the clutter, given by

$$\hat{\pi}_i = \frac{Z_{0i}^{(1)}(1-\pi)}{Z_{0i}}.$$

Similarly, the variance is obtained as

$$\nu = \nu^{\backslash i} - \hat{\pi}_i\frac{(\nu^{\backslash i})^2}{\nu^{\backslash i} + 1} + \hat{\pi}_i(1 - \hat{\pi}_i)\frac{(\nu^{\backslash i})^2\|x_i - m^{\backslash i}\|^2}{d(\nu^{\backslash i} + 1)}.$$

Lemma 3.1. *If p and q are the densities of $N(\mu_p, \sigma_p^2\mathbb{I})$ and $N(\mu_q, \sigma_q^2\mathbb{I})$ respectively, then $r = p/q$ is proportional to $N(\mu_r, \sigma_r^2\mathbb{I})$, where*

$$(\sigma_r^2)^{-1} = (\sigma_p^2)^{-1} - (\sigma_q^2)^{-1}, \quad \begin{aligned} \mu_r &= \mu_p\sigma_r^2(\sigma_p^2)^{-1} - \mu_q\sigma_r^2(\sigma_q^2)^{-1} \\ &= \mu_p\sigma_r^2\{(\sigma_r^2)^{-1} + (\sigma_q^2)^{-1}\} - \mu_q\sigma_r^2(\sigma_q^2)^{-1} \\ &= \mu_p + \sigma_r^2(\sigma_q^2)^{-1}(\mu_p - \mu_q). \end{aligned}$$

References

[1] G David Forney. Codes on graphs: Normal realizations. *IEEE Transactions on Information Theory*, 47(2):520–548, 2001.

[2] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[3] Thomas P Minka. Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.