Expectation propagation

Lloyd Elliott

May 17, 2011

Suppose $p(\mathbf{x})$ is a pdf and we have a factorization

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{i=1}^{n} f_i(\mathbf{x}). \tag{1}$$

Expectation propagation is an inference algorithm designed to approximate the factors f_i. In doing so, we may recover approximations of the marginals and joints of p, or we may find the normalizing constant for p. EP involves parameterising an approximation $\tilde{f}_i$ of each factor f_i and iteratively including each factor into the approximation by minimising a KL-divergence.

For each factor f_i, fix an approximating family of distributions Ω_i. Given (1) and Ω_i, the EP algorithm is as follows:

initialize approximations $\tilde{f}_i$
repeat
  for $i = 1, \ldots, n$ **do**

$$\tilde{f}_i \leftarrow \operatorname*{argmax}_{\hat{f}_i \in \Omega_i} \mathrm{KL}\left(\frac{1}{B} f_i \prod_{j \neq i} \tilde{f}_i \middle\| \frac{1}{C} \hat{f}_i \prod_{j \neq i} \tilde{f}_i \right) \tag{2}$$

  end for
until stopping condition reached

Here, B and C are normalising constants.

Writing $\tilde{p}^{\neg i} = \prod_{j \neq i} \tilde{f}_i$, we see that the update in the EP algorithm sets $\tilde{f}_i$ to:

$$\operatorname*{argmin}_{\hat{f}_i \in \Omega_i} \int \frac{1}{B}(f_i \tilde{p}^{\neg i})(\mathbf{x}) \log \frac{C f_i(\mathbf{x})}{B \hat{f}_i(\mathbf{x})} \mathrm{d}\mathbf{x}, \text{ such that } \int (\hat{f}_i \tilde{p}^{\neg i})(\mathbf{x}) \mathrm{d}\mathbf{x} = C. \tag{3}$$

From this equation, we see that if $\hat{f}_i$ were unconstrained (*i.e.* if Ω_i were all functions on the range of $\mathbf{x}$), then $\hat{f}_i = \frac{C}{B} f_i$ would be a solution. Unfortunately, the computation of B and C are often intractable. Therefore, to make progress in EP, we must place constraints on $\tilde{f}_i$ so that minimising (3) is tractable.

There are two main sorts of constraints on $\tilde{f}_i$ that we will examine:

1. Exponential family constrains,
2. Fully factorised constraints.

In what follows we will see the general implication of these assumptions in detail, making reference to the formulation of EP updates as minimising (2). Other constraints are possible: any choice of Ω_i for which the computation of (3) is tractable leads to an EP algorithm.

Exponential family constraints

Suppose $f(x) = h(x)\exp(\eta^T u(x) - A(\eta))$ and $p(x)$ is any distribution. We want to find the sufficient statistic η that minimises the following KL-divergence:

$$\begin{aligned}\mathrm{KL}(p||q) &= \int p(x)\log\frac{p}{f(x)}\mathrm{d}\mathbf{x}, \\ &= \mathbb{E}_p[p(x)] + \mathbb{E}_p[h(x)] - A(\eta) + \eta^T\mathbb{E}_p[u(x)].\end{aligned}$$

We proceed by equating the derivative of with respect to η to zero:

$$\nabla_\eta A(\eta) = \mathbb{E}_p[u(x)]. \tag{4}$$

But, because f is from an exponential family, $\nabla_\eta A(\eta) = \mathbb{E}_f[u(x)]$. Thus, (9) is minimised when $\mathbb{E}_f[u(x)] = \mathbb{E}_p[u(x)]$. This is why EP is sometimes called 'moment matching.'

Returning to the situation of EP, suppose we restrict $\tilde{f}_i$ to be proportional to a distribution in a a given exponential family:

$$\Omega_i = \{f(x) : f(x) \propto h_i(x) \exp(\eta^T u(x) - A_i(\eta)) \forall \eta\}.$$

Without loss of generality, we have assumed the same form of the sufficient statistics $u(x)$ for each approximating distribution. Suppose $\tilde{f}_i \propto \exp(\eta_i^T u(x) - A_i(\tilde{\eta}_i))$ are the current site approximations (proportionality in $\tilde{\eta}_i$). The EP minimisation step for f_i (2) is:

$$\tilde{f}_i \leftarrow \underset{\hat{f}_i \in \Omega_i}{\operatorname{argmax}} \operatorname{KL}\left(\frac{1}{B} f_i \tilde{p}^{\neg i} \middle|\middle| \frac{1}{C} \hat{f}_i \tilde{p}^{\neg i}\right).$$

Collecting terms in the exponent, the second argument in the KL-divergence is exponential family with (proportionality in $\hat{\eta}_i$):

$$\hat{f}_i \tilde{p}^{\neg i} \propto \exp\left((\hat{\eta}_i^T + \sum_{j \neq i} \tilde{\eta}_j^T) u(x) - A_i(\hat{\eta}_i) - \sum_{j \neq i} A_i(\tilde{\eta}_i)\right). \quad (5)$$

Suppose $\tilde{\eta}_j$ agree given for all $j \neq i$. We will use (5) to write $\mathbb{E}_{\hat{f}_i \tilde{p}^{\neg i}}[u(x)]$ as a function of $\hat{\eta}_i$: Suppose $\Phi_i(\hat{\eta}_i) = \mathbb{E}_{\hat{f}_i \tilde{p}^{\neg i}}[u(x)]$. To proceed, we must be able to compute $\mathbb{E}_{f_i \tilde{p}^{\neg i}}[u(x)]$ for the fixed $\tilde{\eta}_j$. In this case, the update (2) is given by the following:

$$\hat{\eta}_i \leftarrow \Phi_i^{-1}(\mathbb{E}_{f_i \tilde{p}^{\neg i}}[u(x)]). \tag{6}$$

Fully factorised constraints

Suppose $\mathbf{x} = (x_1, \ldots, x_k)$ and

$$p(\mathbf{x}) = \frac{1}{B} \prod_{i=1}^{n} f_i(C_i),$$

where $C_1, \ldots, C_n$ are subsets of $\mathbf{x}$. (*N.b.* that the C_i might overlap.) This model has the same expressive power as factor graphs: If G is a factor graph then the terms $f_i(C_i)$ correspond to the factors of G. In particular, if G is an undirected graphical model, then we can choose $C_1, \ldots, C_n$ so that C_i is the pair of vertices conencted by the i-th edge of G.

The fully factorised constraint on $\tilde{f}_i(C_i)$ is:

$$\tilde{f}_i(C_i) = \prod_{x_\ell \in C_i} \tilde{f}_{i\ell}(x_\ell)$$

We will also assume that $\tilde{f}_{i\ell}(x_\ell)$ are restricted to functions proportional to exponential families with base measure, sufficient statistics, and partition functions $h_{i\ell}, \eta_{i\ell}, A_{i\ell}$ respectively. As above: $\tilde{f}_{i\ell}(x_\ell) \propto \exp(\tilde{\eta}_{i\ell}^T u_\ell(x_\ell) - A_{i\ell}(\tilde{\eta}_{i\ell}))$. Note that as $\tilde{f}_i$ splits, we write seperate sufficient statistics for each component of $\mathbf{x}$. We have constrained Ω_i to be an exponential family that splits over the random variables contained in C_i.

Under these constraints, we find factors in the KL-divergence (3) that depend on $\hat{f}_i$ for a fixed i:

$$\begin{aligned}
\mathrm{KL}\left(\frac{1}{B} f_i \tilde{p}^{\neg i} \middle|\middle| \frac{1}{C} \tilde{f}_i \tilde{p}^{\neg i}\right) =& \frac{1}{B} \int (f_i \tilde{p}^{\neg i})(\mathbf{x}) \log(f_i/\hat{f}_i)(\mathbf{x}) \mathrm{d}\mathbf{x} \\
=& \frac{1}{B} \int f_i(C_i) \prod_{j \neq i} \prod_{x_\ell \in C_j} \tilde{f}_{j\ell}(x_\ell) \log(f_i/\hat{f}_i)(\mathbf{x}) \mathrm{d}\mathbf{x} \\
=& \frac{1}{B} \left(\int_{\mathbf{x} \setminus C_i} \prod_{j \neq i} \prod_{x_\ell \in C_j \setminus C_i} \tilde{f}_{j\ell}(x_\ell) \right) \leftarrow \overset{\text{no } \hat{\eta}_i}{\text{dependence}} \\
& \cdot \int_{C_i} f_i(C_i) \prod_{j \neq i} \prod_{x_\ell \in C_j \cap C_i} \tilde{f}_{j\ell}(x_\ell) \log(f_i/\hat{f}_i)(\mathbf{x}) \mathrm{d}\mathbf{x} \\
=& \mathrm{KL}\left(\frac{1}{B'} f_i \tilde{p}_{C_i}^{\neg i} \middle|\middle| \frac{1}{C'} \hat{f}_i \tilde{p}_{C_i}^{\neg i}\right),
\end{aligned}$$

where $\tilde{p}_{C_i}^{\neg i} = \prod_{j, x_\ell : x_\ell \in C_j \cap C_i} \tilde{f}_{j\ell}(x_\ell)$. Expectations with respect to the first argument of this KL are integrals over C_i which are tractable.

In particluar, $\hat{f}_i = \prod_{x_\ell \in C_i} \hat{f}_{i\ell}(x_\ell)$, and so the above KL is optimised when the following KL-divergences are minimised for each ℓ:

$$\mathrm{KL}\left(\frac{1}{B'} f_i \tilde{p}_{C_i}^{\neg i} \middle|\middle| \frac{1}{D'} \hat{f}_{i\ell} \tilde{p}_{C_i}^{\neg i}\right).$$

By the exponential family derivation above,

$$(\hat{f}_{i\ell} \tilde{p}_{C_i}^{\neg i})(x_\ell) \propto \exp\left(\left(\hat{\eta}_{i\ell}^T + \sum_{j \neq i : x_\ell \in C_j} \tilde{\eta}_{j\ell}^T\right) u_\ell(x_\ell) - A_{i\ell}(\hat{\eta}_{i\ell}) - \sum_{j \neq i : x_\ell \in C_j} A_{j\ell}(\tilde{\eta}_{i\ell})\right) \tag{7}$$

So the EP update for $\hat{f}_{i\ell}$ is found as follows:

1. Use equation (7) above to write $\mathbb{E}_{\hat{f}_{i\ell}\tilde{p}_{C_i}^{\neg i}}[u_\ell(x_\ell)]$ as a function of $\hat{\eta}_{i\ell}$: suppose the function is $\Phi_{i\ell}(\hat{\eta}_{i\ell}) = \mathbb{E}_{\hat{f}_{i\ell}\tilde{p}_{C_i}^{\neg i}}[u_\ell(x_\ell)]$
2. Compute $\mathbb{E}_{f_{i\ell}\tilde{p}_{C_i}^{\neg i}}[u_\ell(x_\ell)]$.
3. Set $\hat{f}_{i\ell} \leftarrow \Phi_{i\ell}^{-1}\left(\mathbb{E}_{f_{i\ell}\tilde{p}_{C_i}^{\neg i}}[u_\ell(x_\ell)]\right)$.

These first two steps involve integration over C_i which is tractable if the sizes of C_i are small. Every named exponential family admits an analytic form for Φ^{-1}.

Example: Graphical models on binary variables

Suppose G is an undirected graphical model on binary random variables $V(G) = \{x_1, \ldots, x_n\}$:

$$p(G) \propto \frac{1}{Z} \prod_{xy \in E(G)} f_{xy}(x, y). \tag{8}$$

Here, $E(G)$ are the edges of G. We have absorbed the factors involving just one variable into the factors on the edges. We can write f_{xy} as the following exponential family with sufficient statistics x, y, xy:

$$\begin{aligned} f_{xy}(xy) &= \mu_{xy;00}^{(1-x)(1-y)} \mu_{xy;10}^{x(1-y)} \mu_{xy;01}^{(1-x)y} \mu_{xy;11}^{xy} \\ &= \exp(\sigma_x x + y\sigma_y + \sigma_{xy} xy + b_{xy}). \end{aligned} \tag{9}$$

In (9), the sufficient statistics for f_{xy} are:

$$\begin{aligned}\sigma_x &= \log(\mu_{xy;10}/\mu_{xy;00}),\\ \sigma_y &= \log(\mu_{xy;01}/\mu_{xy;00}),\\ \sigma_{xy} &= \log\frac{\mu_{xy;11}\mu_{xy;00}}{\mu_{xy;10};\,\mu_{xy;01}}\end{aligned}$$

And the partition function is:

$$b_{xy} = \log \mu_{xy;00}.$$

We will apply the fully factorized constraint to the approximate site potentials:

$$\begin{aligned}\tilde{f}_{xy}(xy) &= \tilde{f}_{xy:x}(x) f_{xy:y},\\ &\propto \exp(\delta_{xy:x} x)\exp(\delta_{xy:y} y).\end{aligned} \tag{10}$$

The sufficient statistics of this approximation are x and y.

We derive the update (6) for $\hat{f}_{xy}$ assuming that $\tilde{f}_{x'y'}$ are given for all $x'y' \neq xy$. We must find the expected values of the sufficient statistics of $f_{xy} p^{\neg xy}_{\{xy\}}$. As in (7), with $C_i = \{xy\}$:

$$f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}(x,y) \propto \exp(\sigma_x x + \sigma_y y + \sigma_{xy} xy + b_{xy} + \sum_{y' \in N(x)\setminus y} \tilde{\sigma}_{xy';x} x + \sum_{x' \in N(y)\setminus x} \tilde{\sigma}_{x'y,y} y). \qquad (11)$$

We compute the expected value of x under (11). $\mathbb{E}_{f_{xy}\tilde{\boldsymbol{p}}^{\neg xy}_{\{xy\}}}[x]$ is:

$$\exp\left(\sigma_x + \sum_{y' \in N(x)\setminus y} \tilde{\sigma}_{xy';x}\right)\left(1 + \exp(\sigma_y + \sigma_{xy} + \sum_{x' \in N(y)\setminus x} \tilde{\sigma}_{x'y;y})\right)$$

$$/\left(1 + \exp(\sigma_x + \sum_{y' \in N(x)\setminus y} \tilde{\sigma}_{xy';x}) + \exp(\sigma_y + \sum_{x' \in N(y)\setminus x} \tilde{\sigma}_{x'y;y})\right.$$

$$\left. + \exp(\sigma_x + \sigma_y + \sigma_{xy} + \sum_{x' \in N(y)\setminus x} \tilde{\sigma}_{x'y;y} + \sum_{y' \in N(x)\setminus y} \tilde{\sigma}_{xy';x})\right),$$

$$= \rho_x. \tag{12}$$

The expression for (12) in the previous slide can be calculated directly from (11) by expanding $\mathbb{E}_{f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}}[x]$ as:

$$\frac{0 * (f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}(0,0) + f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}(0,1)) + 1 * (f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}(1,0) + f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}(1,1))}{f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}(0,0) + f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}(0,1) + f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}(1,0) + f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}(1,1)}.$$

Next,

$$\begin{aligned}
\mathbb{E}_{\tilde{f}_{xy}}[x] = &(0 * (\exp(0\tilde{\sigma}_{xy:x} + 0\tilde{\sigma}_{xy:y}) + \exp(0\tilde{\sigma}_{xy:x} + 1\tilde{\sigma}_{xy:y})) \\
&+1 * (\exp(1\tilde{\sigma}_{xy:x} + 0\tilde{\sigma}_{xy:y}) + \exp(1\tilde{\sigma}_{xy:x} + 1\tilde{\sigma}_{xy:y}))) \\
&/ (\exp(0\tilde{\sigma}_{xy:x} + 0\tilde{\sigma}_{xy:y}) + \exp(1\tilde{\sigma}_{xy:x} + 0\tilde{\sigma}_{xy:y}) \\
&+ \exp(0\tilde{\sigma}_{xy:x} + 1\tilde{\sigma}_{xy:y}) + \exp(1\tilde{\sigma}_{xy:x} + 1\tilde{\sigma}_{xy:y}))) \\
&= \frac{\exp(\tilde{\delta}_{xy;x})}{1 + \exp(\tilde{\delta}_{xy;x})}. \qquad (13)
\end{aligned}$$

Equating (12) and (13) yields the update for $\tilde{\delta}_{xy;x}$:

$$\begin{aligned} \mathbb{E}_{\tilde{f}_{xy}}[x] &= \mathbb{E}_{f_{xy}\tilde{p}^{\neg xy}_{\{xy\}}}[x], \\ \Leftrightarrow \frac{\exp(\tilde{\delta}_{xy;x})}{1+\exp(\tilde{\delta}_{xy;x})} &= \rho_x, \\ \Leftrightarrow \tilde{\delta}_{xy;x} &= \log\frac{\rho_x}{1-\rho_x}. \end{aligned} \tag{14}$$

Thus, the update for $\tilde{\delta}_{xy;x}$ is:

$$\tilde{\delta}_{xy;x} \leftarrow \log\frac{\rho_x}{1-\rho_x},$$

and the update for $\delta_{xy;y}$ is by symmetry. This completes the EP algorithm for arbitrary undirected graphs of binary random variables. Note that (14) is found by inverting the expected value as a function of the natural parameter. This is the Φ^{-1} function from (6).