Approximate Inference Approaches

Amir Globerson

In many cases we have a graphical model which describes our data well but in which exact inference is intractable (e.g., NP hard). Luckily there are approximation methods that work rather well in many practical applications. Here we discuss some of those.

1 Mean Field Methods

For simplicity we consider a pairwise undirected model:

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{ij \in E} \phi_{ij}(x_i, x_j) \prod_i \phi_i(x_i) \tag{1}$$

It will be easier for us to introduce $\theta_{ij}(x_i, x_j) = \log \phi_{ij}(x_i, x_j)$ and $\theta_i(x_i) = \log \phi_i(x_i)$ and write:

$$p(\boldsymbol{x}) = \frac{1}{Z} e^{\sum_{ij} \theta_{ij}(x_i, x_j) + \sum_i \theta_i(x_i)} \tag{2}$$

Say we want to calculate the marginals $p(x_i)$, but for this graph structure it is hard to do. Here's an idea for approximation: we will approximate $p(x)$ with a distribution $q(\boldsymbol{x})$ for which we can calculate the marginals. The simplest example of such $q(\boldsymbol{x})$ is the fully factored (independent) distribution:

$$q(\boldsymbol{x}) = \prod_i q_i(x_i) \tag{3}$$

defined by the singleton distributions $q_i(x_i)$. Clearly $q_i(x_i)$ are the singleton marginals of $q(\boldsymbol{x})$ (i.e., $q(x_i) = q_i(x_i)$). Now there are two key questions:

1. How can we find the distribution $q(\boldsymbol{x})$ that best approximates $p(\boldsymbol{x})$?
2. Why should we expect this to be a good approximation?
3. Can the result be used to approximate the partition function?

Regarding the second question, clearly there are cases for which it will be a bad approximation. But, one case in which this would be a good approximation is when $p(\boldsymbol{x})$ is sharply peaked around one assignment $\boldsymbol{x}^*$. For example, in the extreme case where $p(\boldsymbol{x}^*) = 1$ then clearly it is of the same form of $q(\boldsymbol{x})$. This sharp peak often happens when $p(\boldsymbol{x})$ is obtained after conditioning on evidence that makes one particular assignment very likely. There are other cases where it can be shown that this approximation is exact (e.g., some instances of Gaussian models)

Regarding the first question, it seems somewhat unintuitive that we could approximate a distribution which we can't evaluate. In fact, we will see that in some sense we can't find the best approximation, but we will be able to progressively improve it.

Lets see how the approximation is done. A natural measure of similarity between distributions is the KL divergence:

$$D_{KL}[p|q] = \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \tag{4}$$

It has several useful properties: it is non-negative and equal to zero iff $p(x) = q(x)$ for all x (assuming a finite set of x). Also, it is non-symmetric ($D_{KL}[p|q] \neq D_{KL}[q|p]$).

We want to find a factorized q (see Eq. 3) that is as close as possible to p. So it makes sense to look for q that is closest to p in KL. Because of asymmetry of KL there are two ways of doing that: namely minimizing $D_{KL}[q|p]$ w.r.t. q or minimizing $D_{KL}[p|q]$ w.r.t q. Turns out the first is easier to handle (**Exercise:** What is the problem with minimizing $D_{KL}[p|q]$ w.r.t. q?).

$$D_{KL}[q|p] = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log q(\boldsymbol{x}) - \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log p(\boldsymbol{x})$$

Because of the structure of q we can considerably simplify the above.

$$\sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log q(\boldsymbol{x}) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \sum_i \log q_i(x_i) = \sum_i \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log q_i(x_i) = \sum_i \sum_{x_i} q(x_i) \log q_i(x_i) \quad (5)$$

(note that the LHS is minus the entropy of q and the RHS is minus the sum of entropies of $q(x_i)$). Also, we have (by definition of $p(\boldsymbol{x})$ that:

$$\begin{aligned}
\sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log p(\boldsymbol{x}) &= \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \left[\sum_{ij} \theta_{ij}(x_i, x_j) + \sum_i \theta_i(x_i) - \log Z \right] \\
&= \sum_{ij} \sum_{x_i, x_j} q(x_i, x_j) \theta_{ij}(x_i, x_j) + \sum_i \sum_{x_i} q(x_i) \theta_i(x_i) - \log Z \\
&= \sum_{ij} \sum_{x_i, x_j} q_i(x_i) q_j(x_j) \theta_{ij}(x_i, x_j) + \sum_i \sum_{x_i} q(x_i) \theta_i(x_i) - \log Z
\end{aligned}$$

where the last equality follows from the structure of $q(x)$. Putting it together we have:

$$D_{KL}[q|p] = -\sum_{ij} \sum_{x_i, x_j} q_i(x_i) q_j(x_j) \theta_{ij}(x_i, x_j) - \sum_i \sum_{x_i} q(x_i) \theta_i(x_i) + \sum_i \sum_{x_i} q(x_i) \log q_i(x_i) + \log Z$$

Our goal is to find the q that minimize it, namely $q_{MF} = \arg\min D_{KL}[q|p]$. The key thing to note is that we can easily evaluate the above function except for the constant $\log Z$. So we define a new function:

$$F(q, \theta) = -\sum_{ij} \sum_{x_i, x_j} q_i(x_i) q_j(x_j) \theta_{ij}(x_i, x_j) - \sum_i \sum_{x_i} q(x_i) \theta_i(x_i) + \sum_i \sum_{x_i} q(x_i) \log q_i(x_i) \quad (6)$$

This function is sometimes called the free energy because of its relation to statistical mechanics. Clearly:

$$\min_q F(q, \theta) = \min_q D_{KL}[q|p] \quad (7)$$

Since they differ only by a constant. So we can define:

$$q_{MF} = \arg \min_{\substack{q_i(x_i) \geq 0 \\ \sum_{x_i} q_i(x_i) = 1}} F(q, \theta) \quad (8)$$

The key thing to observe again is that we now have a constrained minimization problem over a function that is easy to evaluate. There is one caveat though: the function $F(q, \theta)$ is not convex in its variables, and generally we can only find its local optima (in fact you can convince yourself, as an exercise, that if you could solve this problem, you would have been able to calculate MAP efficiently, which we know cannot be done).

What can be done is to find local optima of $F(q, \theta)$? One simple approach is to change only a subset of the q variables at each iteration while keeping the others fixed. Here's one way of doing this. At each iteration:

- Pick some k.
- Fix all values in q except $q_k(x_k)$ (for all x_k). Now seek the values of $q_k(x_k)$ that minimize $F(q, \theta)$.

This scheme is known as block coordinate descent or block coordinate minimization. What is particularly nice is that the above optimal value of $q_k(x_k)$ can be found in closed form. We are now viewing F as a function only of q_k (with the other being fixed). So we can write:

$$F(q_k) = -\sum_{j \in N(k)} \sum_{x_j, x_k} q_j(x_j) q_k(x_k) \theta_{kj}(x_k, x_j) - \sum_{x_k} q_k(x_k) \theta_k(x_k) + \sum_{x_k} q_k(x_k) \log q_k(x_k) \tag{9}$$

We would like to minimize this subject to non-negativity and normalization of q_k. Lets forget about non-negativity for now. So the Lagrangian is:

$$\mathcal{L}(q_k, \lambda) = F(q_k) + \lambda \left(\sum_{x_k} q_k(x_k) - 1 \right) \tag{10}$$

Deriving wrt $q_k(x_k)$ (i.e., for a particular value of x_k. e..g, $x_k = 1$) we have:

$$\frac{\partial \mathcal{L}(q_k, \lambda)}{\partial q_k(x_k)} = -\sum_{j \in N(k)} \sum_{x_j} q_j(x_k) \theta_{kj}(x_k, x_j) - \theta_k(x_k) + 1 + \log q_k(x_k) + \lambda \tag{11}$$

Yielding:

$$q_k(x_k) = e^{\theta_k(x_k) + \sum_{j \in N(k)} \sum_{x_j} q_j(x_k) \theta_{kj}(x_k, x_j) - \lambda - 1} \tag{12}$$

To solve for λ we need to write substitute the above into the normalization constraint:

$$1 = \sum_{x_k} q_k(x_k) = \sum_{x_k} e^{\theta_k(x_k) + \sum_{j \in N(k)} \sum_{x_j} q_j(x_k) \theta_{kj}(x_k, x_j) - \lambda - 1} \tag{13}$$

Yielding:

$$e^{1+\lambda} = \sum_{x_k} e^{\theta_k(x_k) + \sum_{j \in N(k)} \sum_{x_j} q_j(x_k) \theta_{kj}(x_k, x_j)} \tag{14}$$

So:

$$q_k(x_k) = \frac{e^{\theta_k(x_k) + \sum_{j \in N(k)} \sum_{x_j} q_j(x_k) \theta_{kj}(x_k, x_j)}}{\sum_{x'_k} e^{\theta_k(x'_k) + \sum_{j \in N(k)} \sum_{x_j} q_j(x'_k) \theta_{kj}(x'_k, x_j)}} \tag{15}$$

Or simply:

$$q_k(x_k) \propto e^{\theta_k(x_k) + \sum_{j \in N(k)} \sum_{x_j} q_j(x_k) \theta_{kj}(x_k, x_j)} \tag{16}$$

This is the mean field update, and should be performed for all k. The order in which you select k should be chosen. A natural choice is just $k = 1, \ldots, n$ but others are possible.

The updates decrease the function $F(q, \theta)$ at every iteration. They will not necessarily reach its global minimum because the function F is not convex.

1.1 Mean-field as a lower bound on the partition function

We now show that mean-field can be used to find lower bounds on the partition function. To see this we note that:

$$F(q, \theta) = D_{KL}[q|p] - \log Z \tag{17}$$

So that $F(q, \theta) \geq -\log Z$ or $\log Z \geq -F(q, \theta)$. So, for every value of q we get a lower bound on $\log Z$. Every iteration decreases F and thus increases the lower bound (makes it tighter). **Exercise:** when is the bound tight?