CS281: Advanced ML **October 30, 2017**

Lecture 15: Mean Field

Lecturer: Sasha Rush *Scribes: Alex Lin, Wei Zhang, Daniel Giebisch, Richard Zhang, Fahad Alhasoun*

15.1 Variational Inference

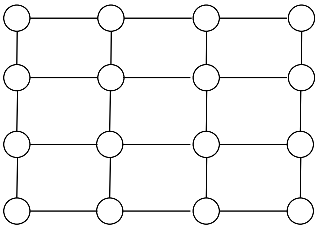
In inference, we're interested in computing $p(z|D)$, the marginal distribution of a node z, given D, where D is all of the evidence in the model. However, very often, direct computation with exact methods is too difficult. As a result the most feasible approach may be to find a $q(z)$ that reasonably approximates $p(z)$ where q is easy to compute. Thus, we define

$$q^* = argmin_{q \in EASY} d(q,p)$$

where EASY is the set of functions that are easy to compute and find the q* such that by some metric d, q* is the closest to p (or smallest gap). Note that the function p can be anything.

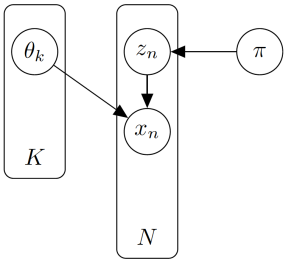
15.2 Hard models

- Ising model $P(y)$



  Recall the Ising model from a few lectures ago, a grid of nodes connected horizontally and vertically to adjacent nodes on 4 sides (or 3 on the edges and 2 on the corners). Things in which we may be interested include the marginal distributions $p(y_{ij} = 1)$ and the partition function. Unlike for a tree structure, direct computation using loopy belief propagation may not converge well for this graph with cycles, so we will look for a different method.

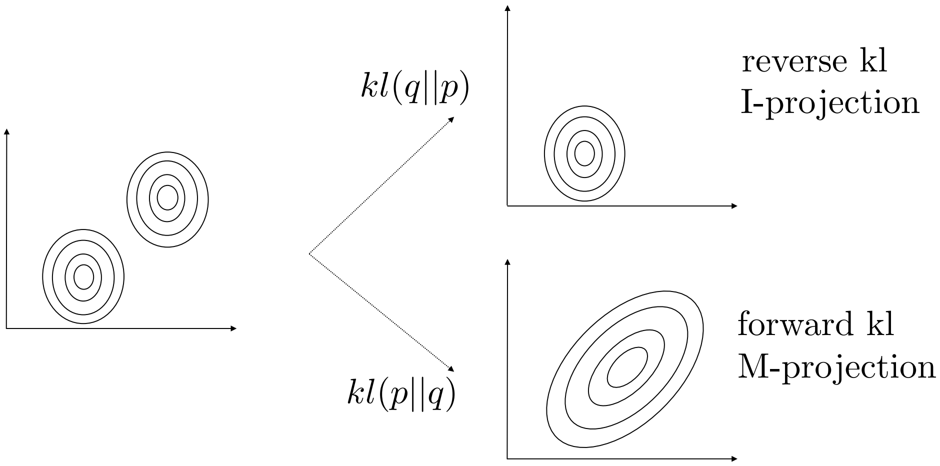
- Gaussian Mixture Model



$$\pi \sim dir(\alpha)$$
$$\Theta_k = (\mu_k, \Sigma_k)$$
$$Z_n \sim \pi$$
$$X_n|(Z_n, \Theta) = N(\mu_k, \Sigma_k)$$

Recall the Gaussian Mixture Model from last lecture, an unsupervised learning method that attempts to categorize data points into clusters. This is nearly the same graphical model as Naive Bayes classification, except that the classes are not observed, but rather inferred. Again, computation of $p(x_n)$ is rather difficult because z_n are not observed. Therefore, one must find a method that infers the z_n in order to infer $p(x_n)$.

15.3 Variational Idea



$$
\begin{aligned}
min_q d(p,q) &= KL(q\|p) \\
&= \int q(z) log \frac{q(z)}{p(z|D)}
\end{aligned}
$$

In trying find the gap function $d(p,q)$ that measures our gap from p to q, one good choice is the KL divergence, but because it is asymmetric, we essentially have 2 options, both of which lead to valid methods:

- $KL(q\|p)$
  The reverse KL, I-projection
- $KL(p\|q)$
  The forward KL, M-projection

The forward KL method is also known as the Expectation Propagation (EP) method.

15.4 Relationship to EM

Recall the Expectation-Maximization (EM) method from last lecture. Much of the mathematics we will use resembles that which we have already studied.

$$
\begin{aligned}
\log p(D) &= \log \int_z p(D,z)dz && \text{(D is the observed data set)} \\
&= \log \int_z q(z)p(D,z)/q(z)dz \\
&= \log E_{z\sim q}\left[\frac{p(D,z)}{q(D)}\right] \\
&\geq E_{z\sim q}\left[\log \frac{p(D,z)}{q(D)}\right] && \text{(the lower bound, by Jensen's inequality)}
\end{aligned}
$$

$$\begin{aligned}\log p(D) - E_q \log \frac{p(D,z)}{q(z)} &= E_q \left[\log p(D) - \log \frac{p(D,z)}{q(z)}\right] \\ &= E_q \left[\log \frac{q(z)}{p(z|D)}\right] \\ &= KL(q\|p)\end{aligned}$$

One more remaining issue is that

$$p(z|D) = \frac{p(z,D)}{p(D)}$$

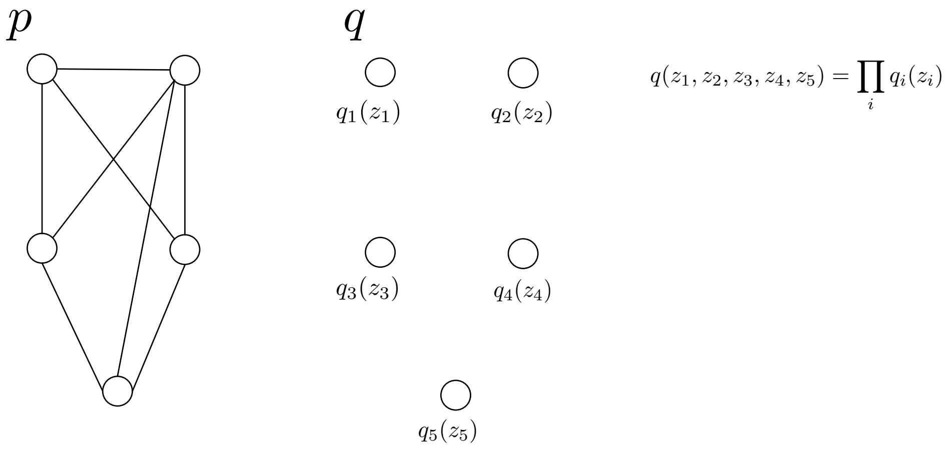
where $p(D)$ is generally very difficult to compute and the main reason why computing p exactly becomes an intractable problem. However, for the purposes of finding q, we can ignore $p(D)$ as just a constant, since it is independent of z. Thus,

$$\min_q KL(q\|p) = \min_q E\left[\log \frac{q(z)}{p(z,D)}\right]$$

Comparison to EM

- In variational inference, we have coordinate ascent just like in EM, except we try to optimize a lower bound.
- In variational inference, we pick q from the EASY set. As a result, we don't have to select point estimates, but rather we can use entire distributions.
- This is very useful in Bayesian setups.
- We can also combine this with EM and sampling to obtain more sophisticated techniques as well.

15.5 Mean Field



This is an algorithm that optimizes one particular q_i, via an expectation over the adjacent variables, while fixing all of the other q and then iterates this procedure over all of the q. It proceeds as follows:

- We assume we have all q except q_i.
- Select $q(z) = \prod q_i(z_i)$ from our EASY set.

- Recall that the goal is to reduce the gap $min_q KL(q\|p)$. We do this by fitting each q_i as

$$q_i^* \leftarrow argmin_q KL(q\|p)$$

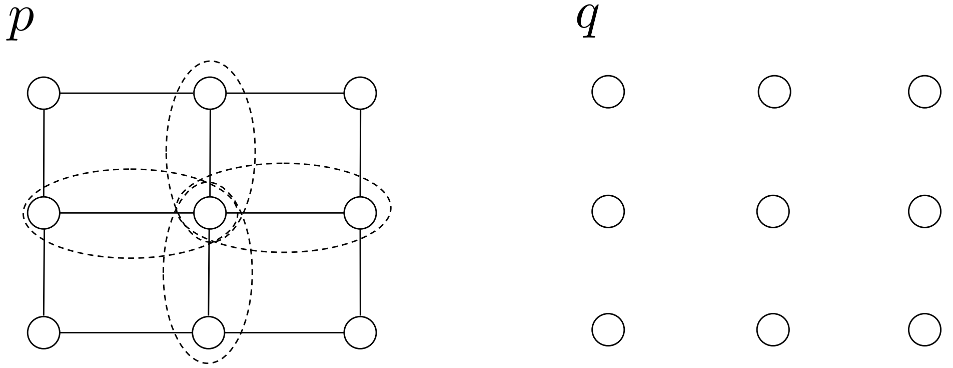
where

$$\begin{aligned}
argmin_q KL(q\|p) &= argmin_{q_i} - H(q_i) - E_q \log(p(z)) \\
&= argmin_{q_i} - H(q_i) - \int_z (\prod_i q_i(z_i)) \log p(z) + other(j \neq i) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(q(z)=}\textstyle\prod q_i(z_i)\text{)} \\
&= argmin_{q_i} - H(q_i) - \int_{z_i} q_i(z_i) \int_{z_j:j\neq i} \prod_{j\neq i} q_i(z_i) \log p(z) + ... \\
&= argmin_{q_i} - H(q_i) - \int_{z_i} q_i(z_i) \log f_i \qquad (\log f_i = E_{q_i}[\log p(z)]) \\
&= argmin_{q_i} KL(q_i\|f_i) \qquad (f_i \text{ is not distribution , but still OK})
\end{aligned}$$

$$q_i \propto exp\{E_{-q_i}(log(p(z)))\}$$

Note that E_{-q_i} denotes an expectation taken over all the variables except z_i.

15.6 Ising Model



We now apply Mean Field Variational Inference to the Ising Model. Note that θ_{v_i} denotes the log-potential of vertex i while $\theta_{E_{i-n}}$ denotes the log-potential of the edge from vertex i to vertex n. We compute q_i as an expectation over the 4 neighboring nodes, the Markov blanket. This is then

$$\begin{aligned}
q_i &\propto exp[E_{qi} \log p(z)] \\
p(z) &\propto exp[\theta_v^T z + z^T \theta_E z] \\
\log q_i(x_i) &\propto E_{-q_i}[\log p(z)] \\
&= E_{-q_i}[\theta_v^T z + z^T \theta_E z] \\
&= E_{-q_i}[z_i \theta_{v_i} + \Sigma_n \theta_{E_{i-n}} Z_i Z_n] \\
&= Z_i \theta_{v_i} + \Sigma_{n \in neighbor} \theta_{E_{i-n}} Z_i E_{q_n}[Z_n] \\
&= Z_i \theta_{v_i} + \Sigma_n \theta_{E_{i-n}} Z_i q_n(1)
\end{aligned}$$

Therefore

$$q_i \propto exp[Z_i \theta_{v_i} + \Sigma_n \theta_{E_{i-n}} Z_i q_n(1)]$$

We can then repeat this procedure over all nodes to update the entire graph. Then we can repeat that over several epochs until we find good convergence of q.