# CS 228T: HW 3

Marco Cusumano-Towner

April 26, 2012

## 1 Forwards vs reverse KL divergence

The forwards KL between $p(x, y)$ and $q(x, y) = q(x)q(y)$ is given by:

$$\begin{aligned}
\mathbb{KL}\left(p(x,y)||q(x)q(y)\right) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x)q(y)} \\
&= \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x \sum_y p(x,y) \log q(x) - \sum_x \sum_y p(x,y) \log q(y) \\
&= \text{const.} - \sum_x \sum_y p(x)p(y|x) \log q(x) - \sum_x \sum_y p(y)p(x|y) \log q(y) \\
&= \text{const.} - \sum_x p(x) \log q(x) - \sum_y p(y) \log q(y) \\
&= \text{const.} - H(p(x)) + \mathbb{KL}\left(p(x)||q(x)\right) - H(p(y)) + \mathbb{KL}\left(p(y)||q(y)\right) \\
&= \text{const.} + \mathbb{KL}\left(p(x)||q(x)\right) + \mathbb{KL}\left(p(y)||q(y)\right)
\end{aligned}$$

where $H(p)$ is the entropy $-\sum_x p(x) \log p(x)$. Therefore, minimizing the total KL divergence is achieved by matching the marginals: $q(x) = p(x)$ and $q(y) = p(y)$.

In order for the $\mathbb{KL}\left(q||p\right)$ to be bounded, we require that every $x, y$ where $p(x, y) = 0$ has $q(x)q(y) = 0$. This captures the zero-forcing behavior of the $\mathbb{KL}\left(q||p\right)$. There are three ways to do this and still result in a valid $q(x)$ and $q(y)$:

1. $q(x) = q(y) = (q_x \; 1 - q_x \; 0 \; 0)$
2. $q(x) = q(y) = (0 \; 0 \; 1 \; 0)$
3. $q(x) = q(y) = (0 \; 0 \; 0 \; 1)$

The best acheiveable KL with the first structure uses $q_x = 0.5$, and gives KL value $\log 2$. The 2nd and 3rd also give KL $\log 2$. If we set $q(x) = p(x)$ and $q(y) = p(y)$ (both uniform), then there will be terms of the form $\log \frac{1/16}{0}$ for each $x, y$ with $p(x, y) = 0$, and the KL will go to $\infty$.

## 2 Structured Variational Methods

We start with the figure in 11.17a, in which $Q$ is represented by pairwise marginals $\psi_{ij}$ for $i = 1, \ldots, 4$ and $j = 1, \ldots, 3$. First, we apply the factorization theorem (KF 11.13) to see if we can get away with simpler updates. In particular, the theorem tells us that each $\psi_{ij}(x_{ij}, x_{ij+1})$ will factor into the fully contained $\phi$ and the interfaces with other $\phi$ or $\psi$. In this case, the horizontal $\phi^{-}_{ij}(x_{ij}, x_{ij+1})$ is fully contained in $\psi_{ij}$. The interfaces with the vertical $\phi^{|}(x_{i-1,k}, x_{ik})$ and $\phi^{|}(x_{ik}, x_{i+1,k})$ for $k = 1, \ldots, 4$ is either $x_{ij}$ or $x_{i,j+1}$. The interfaces with the other $\psi_{ik}$ for $k = 1, \ldots, 3$ are also these singletons. Therefore, the $\psi_{ij}(x_{ij}, x_{i,j+1})$ factors according to:

$$\psi_{ij}(x_{ij}, x_{i,j+1}) = \phi^{-}_{ij}(x_{ij}, x_{i,j+1})\psi'_{ij}(x_{ij})\psi'_{i,j+1}(x_{i,j+1})$$

Therefore, we only need to update each $\psi'_{ij}(x_{ij})$. The $\phi^{-}_{ij}$ cancel and the simplified update is (ignoring edge cases where one of the two listed $\phi^{|}$ doesn't exist):

$$\psi'_{ij}(x_{ij}) \propto \exp\left\{ E_Q\left[\sum_{k=1}^{N} \ln \phi^{|}_{i-1,k}(x_{i-1,k}, x_{ik}) + \ln \phi^{|}_{i,k}(x_{ik}, x_{i+1,k})\right] - E_Q\left[\sum_{k \neq j}^{N} \ln \psi'_{ik}(x_{ik})\right]\right\}$$

For the terms involving the vertical $\phi^{|}_{ij}(x_{ij}, x_{i+1,})$, we note that the two variables are independent in $Q$. Therefore, we only need the singleton marginals $Q(x_{ij})$ (for the rows above and below the row we are updating). For the $\psi'_{ij}(x_{ij})$ terms, we need the singleton marginals of all $x_{ij}$ in the same row we are updating.

We can cache the these marginals when we compute them. For example, when we are updating a $\psi'_{ij}$ in row $i$, we can just use the cached $Q(x_{i-1,k})$, and $Q(x_{i+1,k})$ marginals from the rows above and below, and just run the clique-tree algorithm in row $i$. We still need to do clique-tree inference in row $i$, since the setting of the $x_{ij}, x_{i,j+1}$ in $\psi'_{ij}$ during the update (which corresponds to reducing the potential $\psi'_{ij}$ to the setting $x_{ij}, x_{i,j+1}$) can change the marginals in this row.

Second, you can cache the clique messages themselves within a row, and only recompute them when required. Suppose $\psi'_{ij}(x_{ij})$ is being updated. First, you don't need to compute the messages coming towards $\psi'_{ij}$ for each setting of $x_{ij}$, these don't change (you still need to compute the messages coming from $\psi'_{ij}$ however). Additionally, suppose we only update potentials in a row in back-and-forth order $1, 2, 3, 4, 3, 2, 1, 2, 3, 4, 3, 2, 1, ...$ (we can update potentials in other rows interspersed in between, just the order for any given row must be fixed). If we follow this ordering, we can re-use all the messages coming towards the potential being updated (we will still have to compute all the messages emanating from the node being updated). This will reduce the number of messages computed by 1/2. If we use a bad ordering, then we would have to recompute all the messages.

## 3 Cluster variational methods

The original joint distribution for the DBN can be written

$$P(x) = \prod_{i=1}^{m} p(x_i^{(0)}) \prod_{t=1}^{T} p(x_i^{(t)} | x_{pa_i}^{(t)}, x_i^{(t-1)})$$

where $x_{pa_i}^{(t)}$ are the parents of $x_i^{(t)}$ in the tree of time slice $t$ (each of these parents is from a different chain $j$). After some evidence is observed, we use the evidence to reduce the CPT's to un-normalized factors, and the un-normalized distribution becomes

$$\tilde{P}(x) = \prod_{i=1}^{m} \phi_{i0}(x_i^{(0)}) \prod_{t=1}^{T} \phi_{it}(x_i^{(t)}, x_{pa_i}^{(t)}, x_i^{(t-1)})$$

(a) Cluster for each chain:
Suppose $i = 1$ is the root of the tree. Treating each chain $(x_i^{(0)}, \ldots, x_i^{(T)})$ as a cluster, the Factorization theorem (11.13 from KF) tells us that the factor $\psi_i$ for this cluster will factorize into (i) the set of $\phi$'s that are fully contained in $\psi_i$, and (ii) the sets of interface variables with other factors $\phi$ and $\psi_j$. For the fully contained factors, for $i = 1$ (the root of the tree), all the $\phi_{1t} : t \geq 0$ are fully contained. For $i > 1$, only $\phi_{i0}$ is fully contained. Since the $\psi_j$ for each chain are completely separated, no there are no factors for the interface variabels with other clusters. The other $\phi$ involved are the factors corresponding to the CPT of each $x_i^{(t)}$ ($\phi_{it}$ for $t \geq 1$) and the factors corresponding to the CPT's of the children ($\phi_{jt}$ for $j \in ch_i$ and $t \geq 1$). For $i = 1$, the CPT's of the $x_1^{(t)}$ have already been included (the $\phi_{1t}$ are fully contained), and only the children factors are involved. For $\psi_i$, the interface sets to the $\phi_{it}$ involve the pair of variables $(x_i^{(t)}, x_i^{(t-1)})$, and the interface sets to the children CPT's ($\phi_{jt}$) only include the one variable $x_i^{(t)}$. Therefore, we can reduce the set of factors for $\psi_i$ to a set of pairwise factors $\psi_{it}(x_i^{(t)}, x_i^{(t-1)}) : t \geq 1$:

$$\psi_i \propto \prod_{t=1}^{T} \psi'_{it}(x_i^{(t-1)}, x_i^{(t)}) \implies Q(x) \propto \prod_{i=1}^{m} \prod_{t=1}^{T} \psi'_{it}(x_i^{(t-1)}, x_i^{(t)})$$

For the update equations for each potential $\psi_{it}(x_i^{(t)}, x_i^{(t-1)})$), we determine the other potentials $\phi$ and $\psi$ that are not independent of the variables $x_i^{(t)}$ and $x_i^{(t-1)}$ under $Q$. Specifically, any potential involving variables in the $i$ chain will remain. For the $\phi$, this consists of all $\phi_{is}$ potentials for $s = 0, \ldots, T$ as well as the $\phi_{js}$ potentials for all $j \in ch_i$ (the children of $i$) and $s = 1, \ldots, T$. For the $\psi$, this consists only of the $\psi'_{is}$ for all times $s$:

$$\psi'_{it}(x_i^{(t)}, x_i^{(t-1)}) \propto \exp\{E_Q[\ln \psi i0(x_i^{(0)}) + \sum_{s=1}^{T} \ln \phi_{is}(x_i^{(s)}, x_i^{(s-1)}, x_{pa_i}^{(s)}) + \sum_{s=1}^{T} \sum_{j \in ch_i} \ln \phi_{js}(x_j^{(s)}, x_j^{(s-1)}, x_{pa_j}^{(s)}) - \sum_{s=1}^{T} E_Q[\ln \psi_{is}(x_i^{(s)}, x_i^{(s-1)})]\}$$

(b) Cluster for each tree:
If we use a cluster for each time slice tree, then we have a potential $\psi_t$ for each time slice tree. Again using the factorization theorem, we decompose this into smaller factors: (i) there are no fully contained $\phi$ factors. (ii) for other $\phi$, we have all the $\phi_{it}(x_i^{(t)}, x_i^{(t-1)}, x_{pa_i}^{(t)})$, and these each have interface variables $(x_i^{(t)}, x_{pa_i}^{(t)})$). We also have the $\phi_{i,t+1}(x_i^{(t+1)}, x_i^{(t)}, x_{pa_i}^{(t+1)})$, with interface variable $x_i^{(t)}$. As before, there are no dependent other $\psi_s$ for other times $s$, since the clusters are disconnected in $Q$. Therefore, absorbing the singleton factors and the pairwise $\phi$ into pairwise factors $\psi'_{it}$, we have $\psi_t \propto \prod_{i=1}^{m} \psi'_{it}(x_i^{(t)}, x_{pa_i}^{(t)})$, where we have $pa_0 = \{\}$. When updating a given $\psi'_{it}$, all factors $\phi_{jt}$ and $\phi_{j,t+1}$ are involved, as well as all the other $\psi'_{jt}$ (all for all $j$), since these involve variables that are dependent on the $(x_i^{(t)}, x_{pa_i}^{(t)})$ under $Q$:

$$\psi'_{it}(x_i^{(t)}, x_{pa_i}^{(t)}) \propto \exp\{E_Q[\sum_{j=1}^{m} \ln \phi_{jt}(x_j^{(t)}, x_{pa_j}^{(t)}) + \ln \phi_{j,t+1}(x_j^{(t+1)}, x_{pa_j}^{(t+1)}, x_j^{(t)}) - \sum_{j \neq i}^{m} \ln \psi'_{jt}(x_j^{(t)}, x_{pa_j}^{(t)}))]\}$$

# 4 Programming: Gaussian Mixture Models

See code. My collapsed gibbs is a little slow.