Solution 7 (Graded homework 3): 7^{th} of May 2019
CS-526 Learning Theory

Problem 1: MCMC, Gibbs sampling, and application to the Ising model

1) The detailed balance condition has to be checked only for $\underline{x}_{t+1} \neq \underline{x}_t$. On one hand we have

$$q(\underline{x}_{t+1} \mid \underline{x}_t)p(\underline{x}_t) = \widetilde{q}(\underline{x}_{t+1} \mid \underline{x}_t)A(\underline{x}_{t+1}, \underline{x}_t)p(\underline{x}_t) = \min\big(\widetilde{q}(\underline{x}_{t+1} \mid \underline{x}_t)p(\underline{x}_t), \widetilde{q}(\underline{x}_t \mid \underline{x}_{t+1})p(\underline{x}_{t+1})\big)$$

and on the other hand we have

$$q(\underline{x}_t|\underline{x}_{t+1})p(\underline{x}_{t+1}) = \widetilde{q}(\underline{x}_t|\underline{x}_{t+1})A(\underline{x}_t, \underline{x}_{t+1})p(\underline{x}_{t+1}) = \min\big(\widetilde{q}(\underline{x}_t|\underline{x}_{t+1})p(\underline{x}_{t+1}), \widetilde{q}(\underline{x}_t|\underline{x}_{t+1})p(\underline{x}_t)\big)$$

2) We have

$$\begin{aligned} A(\underline{x}', \underline{x}) &= \min\left(1, \frac{\widetilde{q}(\underline{x}|\underline{x}')p(\underline{x}')}{\widetilde{q}(\underline{x}'|x)p(\underline{x})}\right) = \min\left(1, \frac{p(x_i|\{x'_j\}_{j\neq i})p(x'_i, \{x'_j\}_{j\neq i})}{p(x'_i|\{x_j\}_{j\neq i})p(x_i, \{x_j\}_{j\neq i})}\right) \\ &= \min\left(1, \frac{p(x_i|\{x'_j\}_{j\neq i})p(x'_i|\{x'_j\}_{j\neq i})p(\{x'_j\}_{j\neq i})}{p(x'_i|\{x_j\}_{j\neq i})p(x_i|\{x_j\}_{j\neq i})p(\{x_j\}_{j\neq i})}\right) \end{aligned}$$

Recalling that $x'_j = x_j$ for $j \neq i$ we find that the last expression is $\min(1, 1) = 1$.

3) From the probability distribution of the Ising model:

$$\begin{aligned} p(s'_i|\{s_j\}_{j\neq i}) &= p(s'_i|\{s_j\}_{j\in MB(i)}) \\ &= \frac{\exp(s'_i(\sum_{j\in MB(i)} J_{ij}s_j + h_i)}{\sum_{s'_i=\pm 1} \exp(s'_i \sum_{j\in MB(i)} J_{ij}s_j + h_i)} \\ &= \frac{\cosh(\sum_{j\in MB(i)} J_{ij}s_j + h_i) + s'_i \sinh(\sum_{j\in MB(i)} J_{ij}s_j + h_i)}{2\cosh(\sum_{j\in MB(i)} J_{ij}s_j + h_i)} \\ &= \frac{1}{2}(1 + s'_i \tanh(\sum_{j\in MB(i)} J_{ij}s_j + h_i)) \end{aligned}$$

Problem 2: KL divergence (Barber 8.42)

1) Recall that $KL(p|q) = \mathbb{E}_p[\log p] - \mathbb{E}_p[\log q]$. Since U enters only in q,

$$\begin{aligned} \arg\min_U KL(p|q) &= \arg\min_U(-\mathbb{E}_p[\log q]) = \arg\max_U \mathbb{E}_p[\log q] \\ &= \arg\max_U\big\{\mathbb{E}_p[\mathbf{x}^T\mathbf{U}\mathbf{x}] - \log Z_q(U)\big\} \end{aligned}$$

Since $\mathbf{x}^T\mathbf{U}\mathbf{x}$ is a scalar, we can rewrite it as $\mathrm{Tr}(\mathbf{x}^T\mathbf{U}\mathbf{x}) = \mathrm{Tr}(\mathbf{U}\mathbf{x}\mathbf{x}^T)$. Trace is a linear operation, so we can write $\mathbb{E}_p[\mathrm{Tr}(\mathbf{U}\mathbf{x}\mathbf{x}^T)] = \mathrm{Tr}(\mathbf{U}\mathbb{E}_p[\mathbf{x}\mathbf{x}^T]) = \mathrm{Tr}(\mathbf{U}\mathbf{C})$

2) We know that $KL(p|q) \geq 0$ and $KL(p|q) = 0$ holds only for $q = p$. Hence, if the matrix C is given, we can always plug it into expression from the previous optimization problem and solve (in theory at least). Optimal solution gives us W and thus p is specified, in other words

$$W = \arg\max_{U} \{ \mathrm{Tr}(\mathbf{UC}) - \log Z_q(\mathbf{U}) \}$$

Problem 3: Naive Bayes classifier. Learning by counting. (Barber 10.4)

1) Naive Bayes classifies $\mathbf{x}^*$ as class 1 if $p(class = 1|\mathbf{x}^*) > p(class = 0|\mathbf{x}^*)$. Using Bayes rule and taking log of both sides gives

$$\sum_i \log p(x_i^*|class = 1) + \log p_1 > \sum_i \log p(x_i^*|class = 0) + \log p_0$$

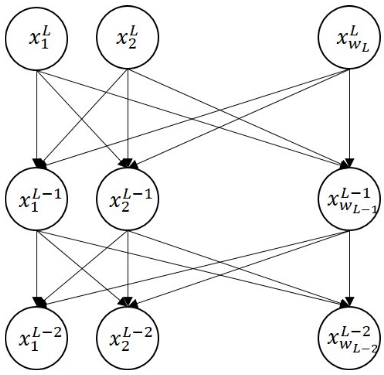
Since $x_i^* \in \{0, 1\}$, we can rewrite the expression above as

$$\sum_i (x_i^* \log \frac{\theta_i^1}{\theta_i^0} + (1 - x_i^*) \log \frac{1 - \theta_i^1}{1 - \theta_i^0}) + \log \frac{p_1}{p_0}$$
$$= \sum_i x_i^* \log \frac{\theta_i^1 (1 - \theta_i^0)}{\theta_i^0 (1 - \theta_i^1)} + \sum_i \log \frac{1 - \theta_i^1}{1 - \theta_i^0} + \log \frac{p_1}{p_0} > 0$$

Setting $a_i = \log \frac{\theta_i^1(1-\theta_i^0)}{\theta_i^0(1-\theta_i^1)}$ and $b = \sum_i \log \frac{1-\theta_i^1}{1-\theta_i^0} + \log \frac{p_1}{p_0}$ gives the desired result

Problem 4: Sigmoid Belief Network (Barber 11.7)

1) The structure is length-L chain of vector variables $\mathbf{x}^l$, where individual entries are fully connected



2) It is possible to use message-passing "along a chain" from top to bottom layer to compute the marginal $p(\mathbf{x}_0)$. Equivalently one can organize the marginalization as a product of $2^w \times 2^w$ "matrices" $p(\mathbf{x}^{l-1}|\mathbf{x}^l)$. Each product costs $O(2^{2w})$ time and since the belief network is a chain of variables $\mathbf{x}^l$ of length L, total time complexity is $O(L2^{2w})$

3) Energy term can be written as

$$\sum_{l=1}^{L}\sum_{i=1}^{w}\mathbb{E}_{q(x_i^{l-1},\mathbf{x}^l)}[\log p(x_i^{l-1}|x^l)] = \sum_{l=1}^{L}\sum_{i=1}^{w}\mathbb{E}_{q(x_i^{l-1},\mathbf{x}^l)}[\log \sigma((2x_i^{l-1}-1)\mathbf{w}_{i,l}^T\mathbf{x}^l)]$$

The expectation $\mathbb{E}_{q(x_i^{l-1},\mathbf{x}^l)}$ involves a sum over 2^{w+1} terms (all possible binary assignments for $w+1$ variables $x_i^{l-1},\mathbf{x}^l$). Moreover the product $w_{i,l}^T\mathbf{x}^l$ is a sum of w terms. So to compute each term for given i,l costs $O(w2^{w+1})$. Hence the total complexity is $O(Lw^2 2^{w+1})$.

Problem 5: EM algorithm for mixtures of Gaussians

1) For the M-step, we need to consider the energy which is given by

$$\sum_{n=1}^{N}\mathbb{E}_{p^{old}(i|\mathbf{x}^n)}[\log p(\mathbf{x}^n|i)p(i)],$$

which in our case can be rewritten as

$$\sum_{n=1}^{N}\sum_{i=1}^{H}p^{old}(i|\mathbf{x}^n)\{-\frac{1}{2\sigma_i^2}\|\mathbf{x}^n-\mathbf{m}_i\|^2-\frac{D}{2}\log 2\pi\sigma_i^2+\log p(i)\}.$$

(a) Optimizing w.r.t. $\mathbf{m}_i$: minimize $\sum_{n=1}^{N}p^{old}(i|\mathbf{x}^n)\|\mathbf{x}^n-\mathbf{m}_i\|^2/\sigma_i^2$. Differentiating w.r.t. $\mathbf{m}_i$ and equaling to zero gives

$$\mathbf{m}_i = \frac{\sum_{n=1}^{N}p^{old}(i|\mathbf{x}^n)\mathbf{x}^n}{\sum_{n=1}^{N}p^{old}(i|\mathbf{x}^n)}$$

(b) Optimizing w.r.t. σ_i^2: minimize $\sum_{n=1}^{N}p^{old}(i|\mathbf{x}^n)\|\mathbf{x}^n-\mathbf{m}_i\|^2/\sigma_i^2 + D\log\sigma^2$. Differentiating w.r.t. $1/\sigma_i^2$ and equaling to zero gives

$$\sigma_i^2 = \frac{\sum_{n=1}^{N}p^{old}(i|\mathbf{x}^n)\|\mathbf{x}^n-\mathbf{m}_i\|^2}{D\sum_{n=1}^{N}p^{old}(i|\mathbf{x}^n)}$$

Problem 6: On gradient ascent for RBM's

1) We have

$$\begin{aligned}L(W) &= \sum_{n=1}^{N}\log p(\underline{v}^{(n)}\mid W) = \sum_{n=1}^{N}\log\sum_{\underline{h}}p(\underline{v}^{(n)},\underline{h})\mid W)\\ &= \sum_{n=1}^{N}\log\left(\sum_{\underline{h}}\exp(\sum_{k,l}v_k^{(n)}W_{kl}h_l)\right)-N\log Z\end{aligned}$$

The partial derivative with respect to W_{ij} is

$$
\begin{aligned}
\frac{\partial L(W)}{\partial W_{ij}} &= \sum_{n=1}^{N} \frac{\sum_{\underline{h}} v_i^{(n)} h_j \exp(\sum_{k,l} v_k^{(n)} W_{kl} h_l)}{\sum_{\underline{h}} \exp(\sum_{k,l} v_k^{(n)} W_{kl} h_l)} - N \frac{1}{Z} \frac{\partial Z}{\partial W_{ij}} \\
&= \sum_{n=1}^{N} v_i^{(n)} \frac{\sum_{\underline{h}_j} h_j \exp(h_j \sum_k v_k^{(n)} W_{kj})}{\sum_{\underline{h}_j} \exp(h_j \sum_k v_k^{(n)} W_{kj})} - N \frac{\sum_{\underline{h}} v_i h_j \exp(\sum_{k,l} v_k W_{kl} h_l)}{\sum_{\underline{v},\underline{h}} \exp(\sum_{k,l} v_k W_{kl} h_l)} \\
&= \sum_{n=1}^{N} (v_i^{(n)} \mathbb{E}_{p(h_j|\underline{v}^{(n)},W)}[h_j] - \langle v_i h_j \rangle)
\end{aligned}
$$

Note that in the second inequality we use that the Markov blanket of node hidden variable h_j is just the set of all visible variables.

2) For binary variables $h_j \in \{-1, +1\}$:

$$
\begin{aligned}
\mathbb{E}_{p(h_j|\underline{v}^{(n)},W)}[h_j] &= \frac{\sum_{\underline{h}_j} h_j \exp(h_j \sum_k v_k^{(n)} W_{kj})}{\sum_{\underline{h}_j} \exp(h_j \sum_k v_k^{(n)} W_{kj})} \\
&= \frac{e^{\sum_k v_k^{(n)} W_{kj}} e^{-\sum_k v_k^{(n)} W_{kj}}}{e^{\sum_k v_k^{(n)} W_{kj}} + e^{\sum_k v_k^{(n)} W_{kj})}} \\
&= \tanh\Big(\sum_k v_k^{(n)} W_{kj}\Big)
\end{aligned}
$$

which proves the claim by replacing in the result of 1).