CS 228 Problem Set 1

Hugh Zhang

March 4, 2017

Problem 1

We want to compute $P(C \mid v_1 \dots v_k)$. For our purposes, our sampler Q is uniform on the possible, so assuming a permutation x' is possible to reach from a state x, then the probabilities are equal. Thus, we can cancel the Qs out.

Acceptance probability is

$$\begin{aligned} & A(c' \mid c, v_1 \dots v_k) \\ = \; & min(1, \frac{P(c' \mid v_1 \dots v_k) Q(c \mid x', v_1 \dots v_k)}{P(c \mid v_1 \dots v_k) Q(c' \mid x, v_1 \dots v_k}) \\ = \; & min(1, \frac{P(c' \mid v_1 \dots v_k)}{P(c \mid v_1 \dots v_k)}) \end{aligned}$$

How do we calculate

$$\frac{P(c' \mid v_1 \dots v_k)}{P(c \mid v_1 \dots v_k)}$$

$$\begin{aligned} & \frac{P(c' \mid v_1 \dots v_k)}{P(c \mid v_1 \dots v_k)} \\ = \; & \frac{P(c', v_1 \dots v_k)}{P(c, v_1 \dots v_k)} \\ = \; & \frac{P(v_1 \dots v_k \mid c') P(c')}{P(v_1 \dots v_k \mid c) P(c)} \end{aligned}$$

We are given $P(v_1 \dots v_k \mid C)$, and although we don't quite know P(c), we are given the assumption that it is uniform a priori, so for our approximation it cancel out.

1.2

The samples are directly from $P(C \mid v_1 \dots v_k)$. Thus,

$$P(C_i = k \mid v_1 \dots v_k) = \frac{\sum_{m=1}^{M} 1(C_i[m] == k)}{M}$$

, or in English, just count the number of times you see it in the sample and divide it by the total number of samples.

1.3

Gibbs sampling does not work. When we sample C, we take two elements C_i and C_j and swap them. If we were to try to Gibbs sample this and try to sample each C_i independently and in order for all i, we would get invalid samples that were not permutations.

Problem 2

From lecture, we have.

$$\begin{aligned} & \log P(y_i \mid x_i, \theta) \\ = \; & \log\left(\frac{1}{Z(x^i,\theta)} \prod_{n \in N} exp(\theta_n f_n(x^i, y_n^i))\right)) \\ = \; & \sum_{n \in N} \theta_n f_n(x^i, y_n^i)) - log(Z(x^i, \theta)) \\ = \; & \sum_{n \in N} \theta_n f_n(x^i, y_n^i)) - log \sum_n \sum_y exp(\theta_n * f_n(y, x^i)) \end{aligned}$$

Since the likelihood is just the log of the probability summed over all the data, letting M be the number of examples in D, we have

$$g(\theta, D) = (1-\alpha)\ell_{Y|X}(\theta, D) + \alpha\ell_{X|Y}(\theta, D)$$

Where as per above,

$$\ell_{Y|X}(\theta, D) = \frac{1}{M}\sum_i^M (\sum_{n \in N} \theta_n f_n(x_n^i, y_n^i)) - log \sum_n \sum_y exp(\theta_n * f_n(y_n, x_n^i)))$$

$$\ell_{X|Y}(\theta, D) = \frac{1}{M}\sum_i^M (\sum_{n \in N} \theta_n f_n(x_n^i, y_n^i)) - log \sum_n \sum_x exp(\theta_n * f_n(y_n^i, x_n)))$$

2.2

We want to calculate

$$
\begin{aligned}
& \frac{\partial}{\partial\theta} g(\theta, D) \\
= \; & (1-\alpha)\frac{\partial}{\partial\theta}\ell_{Y|X}(\theta, D) + \alpha\frac{\partial}{\partial\theta}\ell_{X|Y}(\theta, D)
\end{aligned}
$$

We calculate

$$
\begin{aligned}
& \frac{\partial}{\partial\theta}\ell_{Y|X}(\theta, D) \\
= \; & \frac{1}{M}\frac{\partial}{\partial\theta}\sum_{i}^{M}(\sum_{n\in N}\theta_n f_n(x_n^i, y_n^i)) - log\sum_{n}\sum_{y} exp(\theta_n * f_n(y_n, x_n^i))) \\
= \; & \frac{1}{M}\sum_{i}^{M}(\sum_{n\in N} f_n(x_n^i, y_n^i)) - \frac{\partial}{\partial\theta}\log\sum_{n}\sum_{y} exp(\theta_n * f_n(y_n, x_n^i))) \\
= \; & \frac{1}{M}\sum_{i}^{M}(\sum_{n\in N} f_n(x_n^i, y_n^i)) - \frac{1}{\sum_n\sum_y exp(\theta_n * f_n(y_n, x_n^i)))} * \frac{\partial}{\partial\theta}\left(\sum_{n}\sum_{y} exp(\theta_n * f_n(y_n, x_n^i))\right)) \\
= \; & \frac{1}{M}\sum_{i}^{M}(\sum_{n\in N} f_n(x_n^i, y_n^i)) - \frac{\sum_n\sum_y exp(\theta_n * f_n(y_n, x_n^i)) * f_n(y_n, x_n^i))}{\sum_n\sum_y exp(\theta_n * f_n(y_n, x_n^i)))} \\
= \; & \frac{1}{M}\sum_{i}^{M}(\sum_{n\in N} f_n(x_n^i, y_n^i)) - \frac{\sum_y\sum_n exp(\theta_n * f_n(y_n, x_n^i))) * f_n(y_n, x_n^i))}{Z(\theta, x^i)} \\
= \; & \frac{1}{M}\sum_{i}^{M}(f(x^i, y^i)) - \sum_{y} P(y, x^i) f(y, x^i)) \\
= \; & E_Q[f(x,y)] - E_{p(y|x)}[f(x^i, y^i)]
\end{aligned}
$$

Where Q is the distribution that you got while sampling

so our answer is XCXCXCXCXC IS THE M FACTOR RIGHT?????

$$
\begin{aligned}
& E_Q[f(x,y)] - E_{p(y|x)}[f(x,y)] - \alpha E_Q[f(x,y)] + \alpha E_{p(y|x)}[f(x,y)] + \alpha E_Q[f(x,y)] - E_{p(x|y)}[f(x,y)] \\
= \; & E_Q[f(x,y)] - E_{p(y|x)}[f(x,y)] + \alpha E_{p(y|x)}[f(x,y)] - \alpha E_{p(x|y)}[f(x,y)]
\end{aligned}
$$

Problem 4

Let A denote the set of variables not in either Y or Y's markov blanket. I'm going to split up $Y_M(i,j)$ to include the 2-4 neighboring Y variables, and separately include x_{ij}, which would ordinarily be in Y's markov blanket.

$$
\begin{aligned}
& P(y_{ij} \mid y_{M(i,j)}) = \frac{P(y_{ij}, y_{M(i,j)})}{P(y_{M(i,j)}))} \\
= & \frac{\sum_A P(y_{ij}, y_{M(i,j)}, A)}{\sum_A \sum_{y_{ij}} P(y_{ij} y_{M(i,j)}), A)} \\
= & \frac{\sum_A exp(\eta \sum_N \sum_M y_{ij} x_{ij} + \beta \sum_{edges} y_{ij} y_{i'j'})}{\sum_A \sum_{y_{ij}} exp(\eta \sum_N \sum_M y_{ij} x_{ij} + \beta \sum_{edges} y_{ij} y_{i'j'})} \\
= & \frac{exp(\eta y_{ij} x_{ij} + \beta \sum_{Y_{M(i,j)}} y_{ij} y_{i'j'}) * \sum_A exp(\eta \sum_A y_a x_a + \beta \sum_{restofedges} y_{ij} y_{i'j'})}{\sum_{y_{ij}} exp(\eta y_{ij} x_{ij} + \beta \sum_{Y_{M(i,j)}} y_{ij} y_{i'j'}) * \sum_A exp(\eta \sum_A y_a x_a + \beta \sum_{restofedges} y_{ij} y_{i'j'})} \\
= & \frac{exp(\eta y_{ij} x_{ij} + \beta \sum_{Y_{M(i,j)}} y_{ij} y_{i'j'})}{\sum_{y_{ij}} exp(\eta y_{ij} x_{ij} + \beta \sum_{Y_{M(i,j)}} y_{ij} y_{i'j'})} \\
= & \frac{exp(\eta y_{ij} x_{ij} + \beta \sum_{Y_{M(i,j)}} y_{ij} y_{i'j'})}{exp(\eta x_{ij} + \beta \sum_{y_{i'j'} \in Y_{M(i,j)}} y_{i'j'}) + exp(-\eta x_{ij} - \beta \sum_{y_{i'j'} \in Y_{M(i,j)}} y_{i'j'})}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& P(y_{ij} = 1 \mid Y_{M(i,j)}, x_{ij}) \\
= & \frac{exp(\eta x_{ij} + \beta \sum_{Y_{M(i,j)}} y_{i'j'})}{exp(\eta x_{ij} + \beta \sum_{y_{i'j'} \in Y_{M(i,j)}} y_{i'j'}) + exp(-\eta x_{ij} - \beta \sum_{y_{i'j'} \in Y_{M(i,j)}} y_{i'j'})} \\
= & \frac{1}{1 + exp(-2\eta x_{ij} - 2\beta \sum_{y_{i'j'} \in Y_{M(i,j)}} y_{i'j'})}
\end{aligned}
$$

4.2

You are given the X's and randomly initialize the Ys. Then, go through the Y's in order and draw from the $P(y_{ij} = 1 \mid Y_{M(i,j)}, x_{ij})$ that we gave above (using the newest values of Y that we just got if necessary). When you have gone through all the Y's you have gone through one "iteration" of Gibbs Sampling. After burning through the burn in samples, Gibbs Sampling guarantees that we converge to $P(Y \mid X)$ eventually.