CS 228T: HW 2

Marco Cusumano-Towner

April 19, 2012

1 Annealed importance sampling

(a) $T_i^{-1}(X \to X')$ is a valid transition model if $\sum_{x'} T_i^{-1}(x \to x') = 1$ for all x. This follows from the fact that T_i is a transition model with stationary distribution $p_i(X) \propto f_i(X)$:

$$\begin{aligned}
\sum_{x'} T_i^{-1}(x \to x') &= \sum_{x'} T_i(x' \to x) \frac{f_i(x')}{f_i(x)} \\
&= \sum_{x'} T_i(x' \to x) \frac{p_i(x')}{p_i(x)} \\
&= \frac{1}{p_i(x)} \sum_{x'} T_i(x' \to x) p_i(x') \\
&= \frac{p_i(x)}{p_i(x)} = 1
\end{aligned}$$

where the fourth equality follows from the fact that T_i has stationary distribution $p_i(x)$.
(b)

$$\begin{aligned}
p^*(x_1) &= \sum_{x_2,\ldots,x_k} p^*(x_1, \ldots, x_k) \\
&= \frac{1}{\sum_{x_1,\ldots,x_k} f^*(x_1, \ldots, x_k)} \sum_{x_2,\ldots,x_k} f^*(x_1, \ldots, x_k) \\
&= \frac{1}{\sum_{x_1'} f_0(x_1')} f_0(x_1) \sum_{x_2} T_1^{-1}(x_1 \to x_2) \sum_{x_3} T_2^{-1}(x_2 \to x_3) \cdots \sum_{x_k} T_{k-1}^{-1}(x_{k-1} \to x_k) \\
&= \frac{f_0(x_1)}{\sum_{x_1'} f_0(x_1')} \\
&= p(x_1)
\end{aligned}$$

(c) The sampling distribution $g^*(x_1, \ldots, x_k)$ is given by first sampling x_k, then x_{k-1} conditioned on x_k, and so on until x_1 is sampled given x_2:

$$g^*(x_1, \ldots, x_k) = p_k(x_k) \prod_{i=1}^{k-1} T_i(x_{i+1} \to x_i)$$

$$\begin{aligned}
\frac{f^*(x_1, \ldots, x_k)}{g^*(x_1, \ldots, x_k)} &= \frac{f_0(x_1) \prod_{i=1}^{k-1} T_i^{-1}(x_i \to x_{i+1})}{p_k(x_k) \prod_{i=1}^{k-1} T_i(x_{i+1} \to x_i)} \\
&= \frac{f_0(x_k) \prod_{i=1}^{k-1} T_i(x_{i+1} \to x_i) \frac{f_i(x_{i+1})}{f_i(x_i)}}{p_k(x_k) \prod_{i=1}^{k-1} T_i(x_{i+1} \to x_i)} \\
&= \frac{f_0(x_k) \prod_{i=1}^{k-1} f_i(x_{i+1})}{p_k(x_k) \prod_{i=1}^{k-1} f_i(x_i)} \\
&= \prod_{i=1}^{k} \frac{f_{i-1}(x_i)}{f_i(x_i)}
\end{aligned}$$

where in the last step, we noted that $p_k(x_k) = f_k(x_k)$, and we shifted the above-below pairings in the product.

2 Sampling for the correspondence problem

Let $D \subseteq U$ be the set of nodes $u \in U$ that had their assignments changed between A and A'. The ratio $p(A')/p(A)$ is given by:

$$\begin{aligned}
\frac{p(A')}{p(A)} &= \frac{\exp\left(-\sum_{u \in U} w(u, A'(u))\right)}{\exp\left(-\sum_{u \in U} w(u, A(u))\right)} \\
&= \frac{\prod_{u \in U} \exp\left(-w(u, A'(u))\right)}{\prod_{u \in U} \exp\left(-w(u, A(u))\right)} \\
&= \prod_{u \in D} \frac{\exp\left(-w(u, A'(u))\right)}{\exp\left(-w(u, A(u))\right)} \\
&= \prod_{u \in D} \frac{r(u, A'(u)))}{r(u, A(u))}
\end{aligned}$$

For the proposal distribution we consider the cycle C that resulted in the change from A to A'. Note that for each $A \to A'$, all the cycles that result in this change have the same set of forward edges (the stochastic edges). Therefore, their probabilities are identical (it doesn't matter which u was the first node in the cycle):

$$q(A \to A') = \sum_{u \in C} p(C| \text{ start in } u) p(\text{start in } u)$$

Since all C have the same forward edges, the $p(C|$ start in $u)$ is always the same regardless of the starting node, and is given by the product of transition probabilities:

$$p(C) = \prod_{u \in C} r(u, A'(v)) \implies q(A \to A') = \prod_{u \in C} r(u, A'(v))$$

Therefore using the same equation for $q(A' \to A)$, the ratio is

$$\frac{q(A' \to A)}{q(A' \to A)} = \prod_{u \in C} \frac{r(u, A(u))}{r(u, A'(u))}$$

and the acceptance probability α is therefore 1.

3 Auxiliary variable methods and log-linear models

1. If f_{ij}^{kl} is inactive (0), then $u_{ij}^{kl} \in [0, 1]$. If f_{ij}^{kl} is active (1), then $u_{ij}^{kl} \in [0, \exp(\beta_{ij}^{kl})]$. Since $\beta_{ij}^{kl} > 0$, the second range is strictly bigger than the first.

2. If $u_{ij}^{kl} > 1$, then the feaure f_{ij}^{kl} must be active and x_i and x_j must take the values $x_i^k = 1$ and $x_j^l = 1$. The sampling distribution is

$$\pi(X|u) \propto \pi(u|X)\pi(X) \tag{1}$$

$$= \pi(u|X)\pi_0(X) \exp\left(\sum_{(i,j)\in E} \sum_{k,l} \beta_{ij}^{kl} f_{ij}^{kl}(x_i, x_j)\right) \tag{2}$$

$$= \prod_i \phi_i(x_i) \prod_{(i,j)\in E} \prod_{kl} p(u_{ij}^{kl}|x_i, x_j) \prod_{(i,j)\in E} \prod_{k,l} \exp\left(\beta_{ij}^{kl} f_{ij}^{kl}(x_i, x_j)\right) \tag{3}$$

$$= \prod_i \phi_i(x_i) \prod_{(i,j)\in E} \prod_{kl} \frac{1\left(0 \le u_{ij}^{kl} \le \exp\left(\beta_{ij}^{kl} f_{ij}^{kl}(x_i, x_j)\right)\right)}{\exp\left(\beta_{ij}^{kl} f_{ij}^{kl}(x_i, x_j)\right)} \exp\left(\beta_{ij}^{kl} f_{ij}^{kl}(x_i, x_j)\right) \tag{4}$$

$$= \prod_i \phi_i(x_i) \prod_{(i,j)\in E} \prod_{kl} 1\left(0 \le u_{ij}^{kl} \le \exp\left(\beta_{ij}^{kl} f_{ij}^{kl}(x_i, x_j)\right)\right) \tag{5}$$

The indicator function serves to set the probability to zero of any setting X where for some i, j we have $u_{ij}^{kl} > 1$ but $\neg(x_i^k = 1 \wedge x_j^l = 1)$. Therefore the distribution becomes

$$\pi(X|u) = \begin{cases} \prod_i \phi(x_i) & X \text{ agrees with all } u_{ij}^{kl} > 1 \\ 0 & \text{otherwise} \end{cases}$$

Sampling from this $\pi(X|u)$ is accomplished by setting the values in X for which there is an $u_{ij}^{kl} > 1$ to $x_i^k = 1$ and $x_j^l = 1$, and drawing the remaining x from their independent distributions $\phi_i(x_i)$. The acceptance probability is 1, because we are sampling directly from $\pi(X|u)$ and so this is an instance of Gibbs sampling. At each iteration, this algorithm allows a currently soft-constraint to be ignored with probability $1/\exp(\beta_{ij}^{kl})$. A soft-constraint only becomes active randomly.

3. The Swendsen-Wang algorithm forces some sets of variables to continue to have a consistent state, but it allows the set to switch state together. The algorithm we derived when applied to the Ising model forces some sets of variables that share the same state (say 1) to remain in that state (1) and does not perform large state-flips like Swendsen Wang. Therefore, SW will likely give rise to better mixing.

4 Sampling methods for the marginal likelihood

1. (a)

$$\begin{aligned}
\hat{p}(D) &= \frac{\sum_{m=1}^{M} w_m p(D|\theta_m)}{\sum_{m=1}^{M} w_m} \\
&= \frac{\sum_{m=1}^{M} \frac{p(\theta_m)}{p(\theta_m|D)} p(D|\theta_m)}{\sum_{m=1}^{M} \frac{p(\theta_m)}{p(\theta_m|D)}} \\
&= \frac{\sum_{m=1}^{M} p(D)}{\sum_{m=1}^{M} \frac{p(D)}{p(D|\theta_m)}} \\
&= \frac{M}{\sum_{m=1}^{M} \frac{1}{p(D|\theta_m)}} \\
&= \frac{1}{\frac{1}{M}\sum_{m=1}^{M} \frac{1}{p(D|\theta_m)}}
\end{aligned}$$

(b) The posterior is the optimal proposal distribution $q(x) \propto |f(x)|p(x)$, where $f(x) = P(D|\theta)$ and $p(x) = p(\theta)$

$$p(\theta|D) \propto p(D, \theta) = p(\theta)p(D|\theta)$$

(c) If we have overfit to the data, then $p(\theta_m|D)$ will be very spiky, and all our samples θ_m will have very high likelihood $P(D|\theta_m)$. We won't get any θ_m samples that don't have huge likelihood, so the estimator $\hat{p}(D)$ will be over-estimated.

2. (a) Note that

$$\begin{aligned}
\frac{d}{dt} \log Z_t &= \frac{\frac{d}{dt} \int_\Theta p(D|\theta)^t p(\theta) d\theta}{Z_t} \\
&= \frac{\int_\Theta \frac{d}{dt} p(D|\theta)^t p(\theta) d\theta}{Z_t} \\
&= \frac{\int_\Theta \log p(D|\theta) p(D|\theta)^t p(\theta)}{Z_t}
\end{aligned}$$

Now we have

$$\int_0^1 E_{p_t}[\log p(D|\theta)]dt = \int_0^1 \frac{1}{Z_t} \int_\Theta \log p(D|\theta) p(D|\theta)^t p(\theta) d\theta dt$$

Comparing this to above, we see that

$$\begin{aligned}\int_0^1 E_{p_t}[\log p(D|\theta)]dt &= \int_0^1 \frac{d}{dt}\log Z_t dt \\ &= \log Z_1 - \log Z_0 \\ &= \log\left(\frac{\int_\Theta p(D|\theta)p(\theta)d\theta}{\int_\Theta p(\theta)d\theta}\right) \\ &= \log\frac{p(D)}{1} = \log P(D)\end{aligned}$$

(b) Suppose we have a transition model T_i for each t_i with stationary distribution $p_{t_i}(\theta|D)$. We then follow the procedure in the first problem (annealed importance sampling), and start with a sample from $p_0(\theta|D) = p(\theta)$, and then succesively sample from the T_i in order, using the previous sample as the input to each transition model in turn. However, we don't use the same weighting scheme as in annealed importance sampling, because this would cause us to be sampling from $p_1(\theta|D)$, and we want to be uniformly sampling across all the t_i.