Assignment Five

Matthew C. Scicluna
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
Montréal, QC H3T 1J4
matthew.scicluna@umontreal.ca

December 7, 2017

1 Importance Sampling

We estimate the normalizing constant Z_p for an un-normalized Gaussian $\tilde{p}(x) = \exp\left(-\frac{1}{2\sigma_p^2}x^2\right)$; i.e. we have $p(\cdot) \sim N(0, \sigma_p^2)$ with $p(x) = \tilde{p}(x)/Z_p$. Given N i.i.d. samples $x^{(1)}, \cdots, x^{(N)}$ from a standard normal $q(\cdot) \sim N(0,1)$, we consider the following importance sampling estimate $\hat{Z} = \frac{1}{N}\sum_{i=1}^N \frac{\tilde{p}(x^{(i)})}{q(x^{(i)})}$.

(a) We can see this estimator is unbiased since:

$$\begin{aligned}
\mathbb{E}_{X\sim q}\left\{\hat{Z}\right\} &= \mathbb{E}_{X\sim q}\left\{\frac{1}{N}\sum_{i=1}^N \frac{\tilde{p}(X^{(i)})}{q(X^{(i)})}\right\} && (1.1)\\
&= \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{X\sim q}\left\{\frac{\tilde{p}(X^{(i)})}{q(X^{(i)})}\right\} && (1.2)\\
&= \int_X \frac{\tilde{p}(X)}{q(X)}q(X)dX && (1.3)\\
&= Z_p && (1.4)
\end{aligned}$$

where (1.4) follows from $\int \tilde{p} = Z_p$

(b) Let $f(X) = \frac{\tilde{p}(X)}{q(X)}$ Then $Var(\hat{Z})$ can be easily computed

$$Var(\hat{Z}) = Var\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\tilde{p}(X^{(i)})}{q(X^{(i)})}\right\} \tag{1.5}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N} Var\left\{\frac{\tilde{p}(X^{(i)})}{q(X^{(i)})}\right\} \tag{1.6}$$

$$= \frac{1}{N} Var(f(X)) \tag{1.7}$$

Provided that $Var(f(X))$ is finite

(c) To find the values of σ_p^2 which make $Var(f(X))$ finite, it is enough to find the values which make $\mathbb{E}(f^2)$ finite.

$$\mathbb{E}(f(X)^2) = \int_X \frac{\tilde{p}(X)^2}{q(X)} dX \tag{1.8}$$

$$= \int_X \sqrt{2\pi}\frac{\exp\left(-\frac{1}{2\sigma_p^2}X^2\right)^2}{\exp\left(-\frac{1}{2}X^2\right)} dX \tag{1.9}$$

$$= \sqrt{2\pi}\int_X \exp\left(-\frac{X^2}{2\sigma_c^2}\right) dX \tag{1.10}$$

Where $\sigma_c^2 = \frac{\sigma_p^2}{2-\sigma_p^2}$. There are three cases. If $\sigma_c^2 < 0$ then the integral clearly diverges. If $\sigma_c^2 = 0$ it is undefined. If $\sigma_c^2 > 0$ then (1.10) is an unnormalized Gaussian, and so:

$$\mathbb{E}(f(X)^2) = 2\pi\sigma_c \tag{1.11}$$

$$= \begin{cases} 2\pi\sqrt{\frac{\sigma_p^2}{2-\sigma_p^2}} & \text{for } 0 < \sigma_p^2 < 2 \\ \infty & \text{o.w.} \end{cases} \tag{1.12}$$

2 Gibbs Sampling and Mean Field Variational Inference

We consider the Ising model with binary variables $X_s \in \{0, 1\}$ and a factorization of the form:

$$p(x; \eta) = \frac{1}{Z_p}\exp\left(\sum_{s\in V}\eta_s x_s + \sum_{\{s,t\}\in E}\eta_{st}x_s x_t\right) \tag{2.1}$$

On the 7×7 2D toroidal grid

(a) We derive the Gibbs sampling updates for this model. The algorithm is detailed in Algorithm 1:

Algorithm 1 Calculate $\mu_s = p(X_s = 1)$ using Gibbs Sampling

```
for states $i, j \in \{1, \cdots, 49\}$ and $\{i, j\} \in E$ do
    $\eta_i \leftarrow (1)^i$                                  ▷ Initialize Nodes as per HW
    $\eta_{i,j} \leftarrow 0.5$                                ▷ Initialize Edges as per HW
    Sample $x_i^{(0)} \sim Bern(0.5)$                          ▷ Initialize variables randomly
for Epoch $t \in \{0, \cdots, 5999\}$ do
    for $i \in \{1, \cdots, 49\}$ do
        Sample $x_i^{(t+1)} \sim p\left(x_i = 1 \mid x_1^{(t+1)}, \cdots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \cdots, x_{49}^{(t)}\right)$
for $i \in \{1, \cdots, 49\}$ do
    $\mu_i \leftarrow \frac{1}{5000} \sum_{t=1001}^{6000} x_i^{(t)}$    ▷ Compute Monte Carlo estimates of Moments
return $\mu_1, \cdots, \mu_{49}$
```

To compute $p(x_i = 1 \mid x_{\neg i})$, we notice that it is enough to compute the odds ratio since:

$$\begin{aligned} p(x_i = 1 \mid x_{\neg i}) &= \frac{p(x_i = 1, x_{\neg i})}{\sum_{k=0}^{1} p(x_i = k, x_{\neg i})} && (2.2) \\ &= \frac{1}{1 + \frac{p(x_i=0, x_{\neg i})}{p(x_i=1, x_{\neg i})}} && (2.3) \\ &= \frac{1}{1 + \exp(-Z)} && (2.4) \\ &= \sigma(Z) && (2.5) \end{aligned}$$

Where $Z = log \frac{p(x_i=1, x_{\neg i})}{p(x_i=0, x_{\neg i})}$. We compute the log joint density with $x_i = 1$ and $x_i = 0$

$$\begin{aligned} logp(x_i = 1, x_{\neg i}) &= \sum_{s \in V} \eta_s x_s + \sum_{\{s,t\} \in E} \eta_{st} x_s x_t && (2.6) \\ &= \eta_i + \sum_{j \neq i} \eta_j x_j + \sum_{j \in N(i)} \eta_{ij} x_j + \sum_{\substack{j \notin N(i) \\ \{k,j\} \in E}} \eta_{kj} x_k x_j && (2.7) \\ &= \eta_i + \sum_{j \in N(i)} \eta_{ij} x_j + C && (2.8) \\ logp(x_i = 0, x_{\neg i}) &= \sum_{j \neq i} \eta_j x_j + \sum_{\substack{j \notin N(i) \\ \{k,j\} \in E}} \eta_{kj} x_k x_j && (2.9) \\ &= C && (2.10) \end{aligned}$$

And so subbing (2.8) and (2.10) into Z gives us

$$Z = \eta_i + \sum_{j \in N(i)} \eta_{ij} x_j \tag{2.11}$$

Hence $p(x_i = 1 \mid x_{\neg i}) = \sigma\left(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j\right)$

Ran the experiment 10 times and computed the estimated moments. We found that the standard deviations were very low, ranging from 0.002 to 0.009. The values are available in the code.

(b) We derive the naive mean field updates. Let $q(X_s = 1) = \tau_s$ We wish to do cyclic coordinate descent on $KL(q||p)$. Note that:

$$\begin{aligned} KL(q||p) &= \mathbb{E}_q\{\log q(x) - \log p(x)\} && (2.12)\\ &= \mathbb{E}_q\{\log q(x)\} - \mathbb{E}_q\{\log p(x)\} && (2.13)\\ &= -H(q) - \sum_{s \in V} \eta_s \mathbb{E}_q\{x_s\} - \sum_{\{s,t\} \in E} \eta_{st} \mathbb{E}_q\{x_s x_t\} + log(Z_p) && (2.14)\\ &= -H(q) - \sum_{s \in V} \eta_s \tau_s - \sum_{\{s,t\} \in E} \eta_{st} \tau_s \tau_t + log(Z_p) && (2.15) \end{aligned}$$

Where (2.14) comes from (2.1) and using the mean field assumption. For the update we compute the derivative. First we compute the derivative of $H(q)$

$$\begin{aligned} \frac{\partial H(q)}{\partial \tau_s} &= \frac{\partial}{\partial \tau_s} \sum_i \mathbb{E}_q\{\log q(x_i)\} && (2.16)\\ &= \frac{\partial}{\partial \tau_s} \mathbb{E}_q\{\log q(x_s)\} && (2.17)\\ &= \frac{\partial}{\partial \tau_s} q(x_s) \log q(x_s) + (1 - q(x_s)) \log(1 - q(x_s)) && (2.18)\\ &= \frac{\partial}{\partial \tau_s} \tau_s \log \tau_s + (1 - \tau_s) \log(1 - \tau_s) && (2.19)\\ &= \log\left(\frac{\tau_s}{1 - \tau_s}\right) && (2.20) \end{aligned}$$

The rest of the derivative is easy:

$$\frac{\partial KL(q||p)}{\partial \tau_s} = \eta_s + \sum_{t \in N(s)} \eta_{st} \tau_t - \log\left(\frac{\tau_s}{1 - \tau_s}\right) = 0 \qquad (2.21)$$

$$\Rightarrow \tau_s = \sigma\left(\eta_s + \sum_{t \in N(s)} \eta_{st} \tau_t\right) \qquad (2.22)$$

We can derive the expression for $KL(q||p) - log(Z_p)$ using (2.15)

$$KL(q||p) - log(Z_p) = -H(q) - \sum_{s \in V} \eta_s \tau_s - \sum_{\{s,t\} \in E} \eta_{st} \tau_s \tau_t \qquad (2.23)$$

$$= -\sum_{s \in V} \eta_s \tau_s - \sum_{\{s,t\} \in E} \eta_{st} \tau_s \tau_t - \sum_s (\tau_s \log \tau_s + (1 - \tau_s) \log(1 - \tau_s)) \qquad (2.24)$$

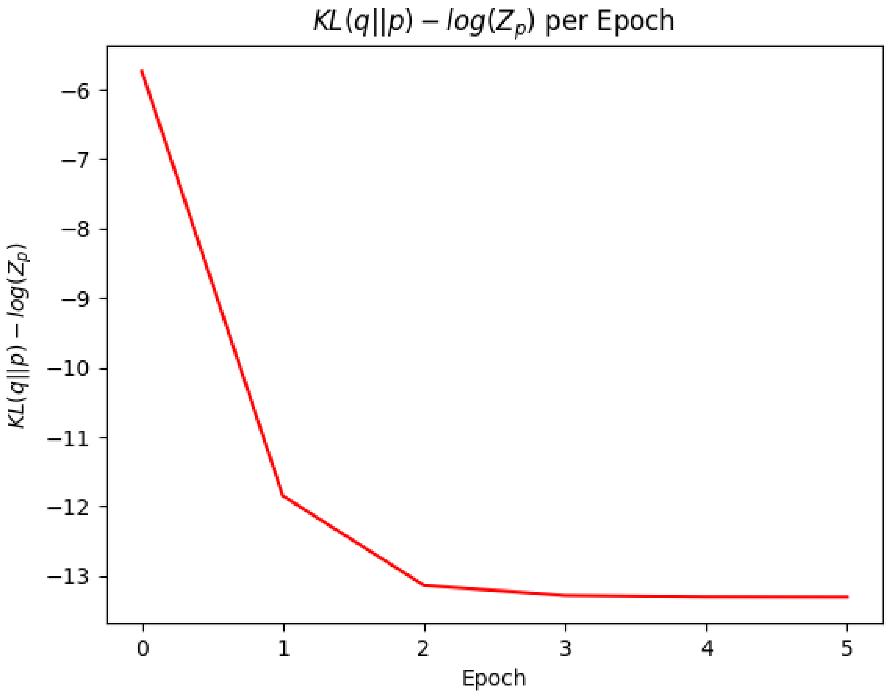

Figure 2.1: $KL(q||p) - log(Z_p)$ computed per epoch

We did cyclic coordinate descent on $KL(q||p)$, sequentially updating each τ_s. We monitored $KL(q||p) - log(Z_p)$ and plotted it per epoch in figure 2.1.

We computed the l_1 distance between τ_s and μ_s. We found that it was only 0.0077. In this sense it was a good approximation. We tried several different initializations but found that the model did not get stuck in different local minima, and converged to the same parameters each time.