IFT 6269, Homework 5

Adel Nabli, ID: 20121744

December 11 2018

1 Cautionary tale about importance sampling

1. By linearity of the expectation, we have:

$$\mathbb{E}[\hat{Z}] = \frac{Z_p}{N}\sum_{i=1}^{N}\mathbb{E}\Big[\frac{p(X_i)}{q(X_i)}\Big]$$

But as the X_i are i.i.d, we can write that $\forall i \in [\![1, N]\!]$, $\mathbb{E}\Big[\frac{p(X_i)}{q(X_i)}\Big] = \mathbb{E}\Big[\frac{p(X)}{q(X)}\Big]$.

Let's name $h : x \mapsto \frac{p(x)}{q(x)}$. Then, we have that $\int_{\mathbb{R}} |h(x)|q(x)dx < +\infty$, and knowing that $X \sim q$, by the *"transfer theorem"* we can write:

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)q(x)dx = \int_{\mathbb{R}} \frac{p(x)}{q(x)}q(x)dx = \int_{\mathbb{R}} p(x)dx = 1$$

This leads to:

$$\mathbb{E}[\hat{Z}] = \frac{Z_p}{N}\sum_{i=1}^{N} 1 = Z_p$$

Hence, we can say that $\hat{Z}$ is an unbiased estimator of Z_p.

2. The X_i being i.i.d and f being continuous, we have that the $f(X_i)$ are also i.i.d. Thus, we can write:

$$\forall\, \sigma_p\ s.t\ Var\big(f(X)\big) < +\infty,\ Var(\hat{Z}) = Var\Big(\frac{1}{N}\sum_{i=1}^{N} f(X_i)\Big) = \frac{1}{N^2}N \times Var\big(f(X)\big) = \frac{1}{N}Var\big(f(X)\big)$$

3. $Var\big(f(X)\big)$ is defined if $f \in \mathbb{L}^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), q)$.
This condition is verified i.i.f we have $\int_{\mathbb{R}} f(x)^2 q(x)dx < +\infty$. But we also have:

$$\begin{aligned}\int_{\mathbb{R}} f(x)^2 q(x)dx &= \int_{\mathbb{R}} \Big(\frac{\tilde{p}(x)}{q(x)}\Big)^2 q(x)dx = \int_{\mathbb{R}} \frac{\tilde{p}(x)^2}{q(x)}dx \\ &= \int_{\mathbb{R}} exp\Big(\frac{x^2}{2} - \frac{2x^2}{2\sigma_p^2}\Big)\sqrt{2\pi}dx \\ &= \sqrt{2\pi}\int_{\mathbb{R}} exp\Big(x^2\big(\frac{1}{2} - \frac{1}{\sigma_p^2}\big)\Big)dx\end{aligned}$$

Which means that $\mathbb{E}\big[f(X)^2\big] < +\infty$ i.i.f $\sigma_p^2 \in]0, 2[$.

2 Gibbs sampling and mean field variational inference

1. We want to have an estimation $\hat{\mu}_s$ of the moments $\mathbb{E}[X_s]$ at each node s using a Gibbs sampling. Our estimate is thus: $\hat{\mu}_s = \dfrac{1}{T-T_0}\sum_{t=T_0+1}^{T} X_s$ with $T_0 = 1000$ the burn-in time and $T = 6000$ the total number of epochs.

   Thus, at each epoch, we will draw an example of each random variable X_s in the UGM, knowing that $X_s \sim Bernoulli(p_s)$ and $p_s = p(X_s = 1|X_{\neg s})$ with $\neg s = [\![1, 49]\!] \backslash \{s\}$.

   But, for the Ising model, we have that

$$p(X_s|X_{\neg s}) \propto p(X_s, X_{\neg s}) = exp\big(\eta_s x_s + \sum_{t \in N(s)} \eta_{st} x_s x_t + rest\big)$$

   with $N(s)$ the Markov blanket at node s. In our case, having a grid, we can refer to each node s using its coordinates (i, j) s.t:

   - $s = (i-1) \times 7 + j$ with $i, j \in [\![1, 7]\!]^2$
   - $N(s) = \{(i-1 \mod 7, j), (i+1 \mod 7, j), (i, j-1 \mod 7), (i, j+1 \mod 7)\}$

   Finally, we can write:

$$p_s = p(X_s = 1|X_{\neg s}) = \frac{exp\big(\eta_s x_s + \sum_{t \in N(s)} \eta_{st} x_t\big)}{1 + exp\big(\eta_s x_s + \sum_{t \in N(s)} \eta_{st} x_t\big)} = sigmoid(\eta_s x_s + \sum_{t \in N(s)} \eta_{st} x_t)$$

   After having implemented the Gibbs Sampling updates (*cf Notebook*), we find:

$$\hat{\mu} = \begin{pmatrix} 0.7072 & 0.912 & 0.7198 & 0.9082 & 0.7284 & 0.907 & 0.7098 \\ 0.9124 & 0.746 & 0.9006 & 0.7288 & 0.9082 & 0.7328 & 0.908 \\ 0.712 & 0.902 & 0.735 & 0.9058 & 0.72 & 0.8946 & 0.7142 \\ 0.912 & 0.7274 & 0.9008 & 0.7342 & 0.9086 & 0.7384 & 0.9092 \\ 0.7254 & 0.8966 & 0.7408 & 0.9058 & 0.7416 & 0.902 & 0.718 \\ 0.9136 & 0.7426 & 0.9028 & 0.7278 & 0.8964 & 0.7344 & 0.9184 \\ 0.7136 & 0.9092 & 0.7114 & 0.9088 & 0.7158 & 0.9152 & 0.7062 \end{pmatrix}$$

   With a standard deviation after running 10 experiments of:

$$\hat{\sigma} = \begin{pmatrix} 0.0053 & 0.0031 & 0.0043 & 0.0046 & 0.0044 & 0.0055 & 0.0032 \\ 0.0047 & 0.0054 & 0.0048 & 0.0050 & 0.0051 & 0.0072 & 0.0035 \\ 0.0057 & 0.0039 & 0.0054 & 0.0045 & 0.0067 & 0.0064 & 0.0061 \\ 0.0039 & 0.0043 & 0.0060 & 0.0052 & 0.0035 & 0.00450 & 0.0047 \\ 0.0047 & 0.0073 & 0.0049 & 0.0031 & 0.0032 & 0.0037 & 0.0039 \\ 0.0040 & 0.0028 & 0.0034 & 0.0050 & 0.0050 & 0.0032 & 0.00465 \\ 0.0038 & 0.0033 & 0.0071 & 0.0027 & 0.0060 & 0.0030 & 0.0037 \end{pmatrix}$$

2. We have:

$$\begin{aligned} KL(q||p) &= \mathbb{E}_q\big[\log\frac{q}{p}\big] = \sum_x q(x)\log\frac{q(x)}{p(x)} \\ &= \sum_x q(x)\log q(x) - \sum_x q(x)\log p(x) \end{aligned}$$

As we have a fully factorized approximation, we know that, $\forall x$:

$$\begin{cases} q(x) = \prod_s q_s(x_s) \ with \ q_s(x_s) = \tau_s \ if x_s = 1 \ and \ 1-\tau_s \ if \ x_s = 0 \\ p(x) = \dfrac{1}{Z_p} \exp\left(\tilde{\eta}^T T(x)\right) \ with \ \tilde{\eta} = \begin{pmatrix} (\eta_s)_{s\in V} \\ (\eta_{st})_{s,t\in E} \end{pmatrix} \ and \ T(x) = \begin{pmatrix} (x_s)_{s\in V} \\ (x_s x_t)_{s,t\in E} \end{pmatrix} \end{cases}$$

Thus, we can re-write $KL(q||p)$:

$$KL(q||p) = \log(Z_p) - \mathbb{E}_q\left[\tilde{\eta}^T T(x)\right] + \sum_x q(x)\log q(x)$$

But, we also have:

$$\sum_x q(x)\log q(x) = \sum_x \prod_t q_t(x_t) \log \prod_s q_s(x_s) = \sum_x \sum_s q_s(x_s)\log q_s(x_s) \prod_{t\neq s} q_t(x_t)$$

And as we have finite sums, we can change the order of summation:

$$\begin{aligned} \sum_x q(x)\log q(x) &= \sum_s \sum_x q_s(x_s)\log q_s(x_s) \underbrace{\prod_{t\neq s} q_t(x_t)}_{=q_{\neg x_s}(\neg x_s)} \\ &= \sum_s \sum_{x_s} \sum_{\neg x_s} q_s(x_s)\log q_s(x_s) q_{\neg x_s}(\neg x_s) \\ &= \sum_s \sum_{x_s} q_s(x_s)\log q_s(x_s) \\ &= \sum_s \tau_s \log \tau_s + (1-\tau_s)\log(1-\tau_s) \end{aligned}$$

Which leads to:

$$KL(q||p) = \log(Z_p) - \mathbb{E}_q\left[\tilde{\eta}^T T(x)\right] + \sum_s \tau_s \log \tau_s + (1-\tau_s)\log(1-\tau_s) \tag{1}$$

For the coordinate descent, we want to fix one q_i, write $q = q_i q_{\neg i}$ and minimize $KL(q_i q_{\neg i}||p)$ with respect to each $q_i(x_i)$. Knowing that, we can re-write $\mathbb{E}_q\left[\tilde{\eta}^T T(x)\right] = \mathbb{E}_{q_i}\left[\tilde{\eta}^T \mathbb{E}_{q_{\neg i}}\big(T(x)\big)\right]$. As we have the condition $\sum_{x_i} q_i(x_i) = 1$, we add the Lagrange multiplier to our KL divergence and solve $\dfrac{\partial}{\partial q_i(x_i)}\Big(KL(q||p) + \lambda\big(1 - \sum_{x_s} q_s(x_s)\big)\Big) = 0$.

This equation leads to solving $\log q_i(x_i) + 1 - \lambda - \tilde{\eta}^T \mathbb{E}_{q_{\neg i}}\big(T(x)\big) = 0$, which means that we have:

$$q_i^{(t+1)} \propto \exp\Big(\tilde{\eta}^T \mathbb{E}_{q_{\neg i}}\big(T(x)\big)\Big)$$

But as we have:

$$\tilde{\eta}^T \mathbb{E}_{q_{\neg i}^{(t)}}\big(T(x)\big) = \eta_i x_i + \sum_{j\neq i} \underbrace{\mathbb{E}_{q_{\neg i}^{(t)}}\big(x_j\big)}_{=\tau_j^{(t)}} + \sum_{j\in N(i)} \eta_{ij} \underbrace{\mathbb{E}_{q_{\neg i}^{(t)}}\big(x_i x_j\big)}_{=x_i\tau_j^{(t)}} + g(x_{\neg i})$$

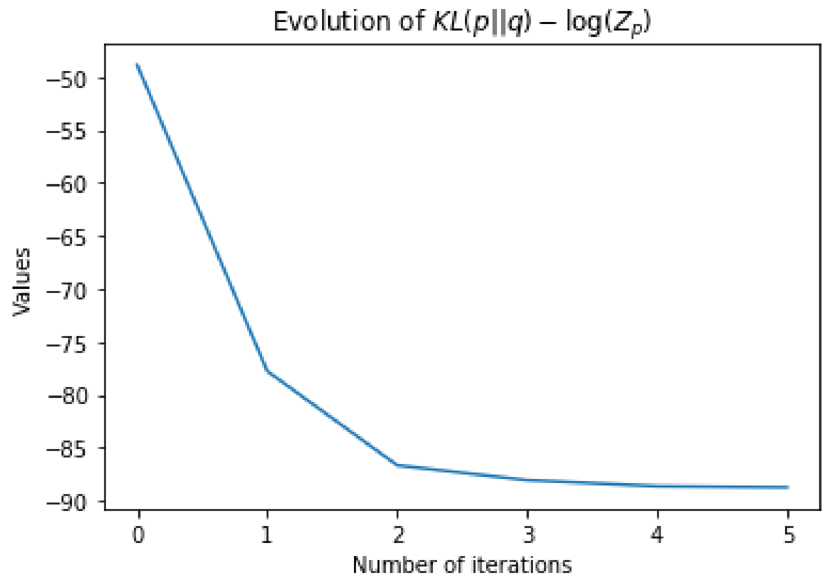
We can derive the update rule:

$$\tau_i^{(t+1)} = sigmoid(\eta_i + \sum_{j\in N(i)} \eta_{ij}\tau_j^{(t)})$$

By taking (1), we also show that:

$$
\begin{aligned}
KL(q||p) - \log(Z_p) &= -\mathbb{E}_q\big[\sum_s \eta_s X_s + \sum_{s,t} \eta_{st} X_s X_t\big] + \sum_s \tau_s \log \tau_s + (1-\tau_s)\log(1-\tau_s) \\
&= -\sum_i \tau_i\big(\eta_i + \sum_{j\in N(i)} \eta_{ij}\tau_j\big) + \sum_s \tau_s \log \tau_s + (1-\tau_s)\log(1-\tau_s)
\end{aligned}
$$

By implementing the method (*cf Notebook*), we show that this method converges very quickly:



By running 5 times with different initialization, we find the values of $d(\tau, \hat{\mu})$ being 0.028, 0.029, 0.029, 0.029, 0.029.

Thus, the mean field seems to be a good approximation here (only off by 3%), which doesn't get stuck in different local minima and converging very quickly compared to the Gibbs sampling method.