IFT 6135 - Homework 3

Adel Nabli

05/04/2019

1 Question 1

1. Let Z be a gaussian vector of $\mathbb{R}^K$ following $\mathcal{N}(0, I_K)$. Let's pose $G = \mu + \sigma \odot Z = \mu + diag(\sigma)Z$ with $\mu \in \mathbb{R}^K$ and $\sigma \in \mathbb{R}_+^*$. Then, for any measurable set $B \in \mathbb{R}^K$ we have:

$$P(G \in B) = P(Z \in B') = \int_{B'} q(z)dz = \int_{B'} \frac{1}{(2\pi)^{K/2}} \exp\Big[-\frac{1}{2} < z, z > \Big] dz$$

with B' the image of B by the mapping of $\mathbb{R}^K$ in itself $x \mapsto diag(\sigma)^{-1}[x - \mu]$ *($diag(\sigma)$ is invertible as $\sigma > 0$)*. Then, using the change of variable $z = diag(\sigma)^{-1}[g - \mu]$ and as $\frac{\partial z}{\partial g} = diag(\sigma)^{-1}$ we have, using the given formula:

$$\begin{aligned} P(G \in B) &= \int_B \frac{1}{(2\pi)^{K/2}} \exp\Big[-\frac{1}{2} < diag(\sigma)^{-1}(g-\mu), diag(\sigma)^{-1}(g-\mu) > \Big] \Big| \det\big(diag(\sigma)\big)^{-1} \Big| dg \\ &= \int_B \frac{1}{(2\pi)^{K/2}\sqrt{\det\big(diag(\sigma^2)\big)}} \exp\Big[-\frac{1}{2} < (g-\mu), diag(\sigma^2)^{-1}(g-\mu) > \Big] dg \end{aligned}$$

and we recognize the distribution $G \sim \mathcal{N}(\mu, diag(\sigma^2))$ *(the law of a random variable is entirely determined by its CDF and we can chose B to have a CDF in the expression above)*.

2. Using exactly the same derivations with S instead of $diag(\sigma)$ *(which is possible as S is non-singular)*, and as $\sqrt{\det(SS^\top)} = \sqrt{\det(S)^2} = |\det(S)|$, we have:

$$\begin{aligned} P(G \in B) &= \int_B \frac{1}{(2\pi)^{K/2}} \exp\Big[-\frac{1}{2} < S^{-1}(g-\mu), S^{-1}(g-\mu) > \Big] \Big| \det\big(S\big)^{-1} \Big| dg \\ &= \int_B \frac{1}{(2\pi)^{K/2}\sqrt{\det\big(SS^\top\big)}} \exp\Big[-\frac{1}{2} < (g-\mu), (SS^\top)^{-1}(g-\mu) > \Big] dg \end{aligned}$$

and we recognize the distribution $G \sim \mathcal{N}(\mu, SS^\top)$.

2 Question 2

1. We have:

$$\log p_\theta(x|z) = \log \frac{p_\theta(x,z)}{p(z)} = \log p_\theta(x) + \log p_\theta(z|x) - \log p(z)$$

Then, as $\log p_\theta(x)$ is independent of z, taking the expectation w.r.t $z \sim q_\phi$ leads to:

$$\begin{aligned} \mathbb{E}_{z\sim q_\phi}\big[\log p_\theta(x|z)\big] &= \log p_\theta(x) + \mathbb{E}_{z\sim q_\phi}\Big[\log\Big(p_\theta(z|x)\frac{q_\phi(z|x)}{q_\phi(z|x)}\Big)\Big] - \mathbb{E}_{z\sim q_\phi}\big[\log p(z)\big] \\ &= \log p_\theta(x) - \underbrace{\mathbb{E}_{z\sim q_\phi}\Big[\log\frac{q_\phi(z|x)}{p_\theta(z|x)}\Big]}_{=D_{KL}\big(q_\phi(z|x)||p_\theta(z|x)\big)} + \underbrace{\mathbb{E}_{z\sim q_\phi}[\log q_\phi(z|x)] - \mathbb{E}_{z\sim q_\phi}\big[\log p(z)\big]}_{\textit{independent of } \theta} \end{aligned}$$

Then, as we are maximizing w.r.t θ and the two last terms don't depend on θ, we have:

$$\arg\max_\theta \mathbb{E}_{z\sim q_\phi}\big[\log p_\theta(x|z)\big] = \arg\max_\theta \Big\{ \log p_\theta(x) \underbrace{-D_{KL}\big(q_\phi(z|x)||p_\theta(z|x)\big)}_{=B(\theta)} \Big\}$$

We have that $B(\theta)$ is negative as the KL divergence is positive.

2. As ϕ^* maximizes the sum of ELBO, it doesn't necessarily maximize each of them whereas q_i^* is set to maximize the i^{th} one. Then, we have:

$$\forall i,\ \mathcal{L}_{q_{\phi^*}}(\theta, x_i) \leq \mathcal{L}_{q_i^*}(\theta, x_i) \tag{1}$$

But, for any q, we can write:

$$\begin{aligned}
\mathcal{L}_q(\theta, x_i) &= \mathbb{E}_q[\log p_\theta(x_i|z)] - D_{KL}\big(q(z|x_i)||p(z)\big) \\
&= \mathbb{E}_q[\log p_\theta(x_i|z)] - \mathbb{E}_q\Big[\log \frac{q(z|x_i)}{p(z)}\Big] \\
&= \mathbb{E}_q\Big[\log \frac{p_\theta(x_i|z)p(z)}{q(z|x_i)}\Big] \\
p(x|z)p(z) = p(z|x)p(x) \rightarrow &= \mathbb{E}_q\Big[\log \frac{p_\theta(z|x_i)p_\theta(x_i)}{q(z|x_i)}\Big] \\
&= -\mathbb{E}_q\Big[\log \frac{q(z|x_i)}{p_\theta(z|x_i))}\Big] + \mathbb{E}_q[\underbrace{\log p_\theta(x_i)}_{fixed}] \\
&= -D_{KL}\big(q(z|x_i)||p_\theta(z|x_i)\big) + \log p_\theta(x_i)
\end{aligned}$$

In our case, we have, on the one hand $q(z|x_i) = q_{\phi^*}(z|x_i)$ and on the other hand $q(z|x_i) = q_{i^*}(z)$, then using (1) we write:

$$\begin{gathered}
-D_{KL}\big(q_{\phi^*}(z|x_i)||p_\theta(z|x_i)\big) + \log p_\theta(x_i) \leq -D_{KL}\big(q_{i^*}(z)||p_\theta(z|x_i)\big) + \log p_\theta(x_i) \\
D_{KL}\big(q_{\phi^*}(z|x_i)||p_\theta(z|x_i)\big) \geq D_{KL}\big(q_{i^*}(z)||p_\theta(z|x_i)\big)
\end{gathered}$$

which makes sense as q_{i^*} has been optimized to match as much as possible $p_\theta(z|x_i)$, making it "closer" to it than q_{ϕ^*} is.

3. (a) Using the q_{i^*} allows us to lower the bias compared to using q_{ϕ^*} as we showed (the KL divergence is lower).
   (b) From a computational point of view, it's more expensive to compute n different $\arg\max$ than to compute only one for the whole sum. Then, computing the q_{i^*} is more expensive than computing the q_{ϕ^*}.
   (c) In terms of memory, it is also more expensive to store n parameters q_{i^*} than only one ϕ^*.

3 Question 3

1. As log is a concave function, by the Jensen's inequality, $\forall$ r.v X, $\log\big(\mathbb{E}[X]\big) \geq \mathbb{E}[\log(X)]$. Thus, we can write:

$$\mathcal{L}_K = \mathbb{E}_{h_i\sim q(h)}\Big[\log\Big(\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,h_i)}{q(h_i)}\Big)\Big] \leq \log\Big(\mathbb{E}_{h_i\sim q(h)}\Big[\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,h_i)}{q(h_i)}\Big]\Big) \tag{2}$$

$$= \log\Big(\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{h_i\sim q(h)}\Big[\frac{p(x,h_i)}{q(h_i)}\Big]\Big) \tag{3}$$

But, we also have that:

$$\mathbb{E}_{h_i \sim q(h)}\Big[\frac{p(x,h_i)}{q(h_i)}\Big] = \int \frac{p(x,h_i)}{\cancel{q(h_i)}}\cancel{q(h_i)}dh_i = \int p(x,h_i)dh_i = p(x) \tag{4}$$

Then, using (3) in (2) leads to:

$$\mathcal{L}_K \leq \log\Big(\frac{1}{K}\sum_{i=1}^{K} p(x)\Big) = \log p(x)$$

2. Let's recall the Jensen's inequality in the discrete case:

   Lemma 3.1 (Jensen's inequality in the discrete case)**.** *$\forall\ f : I \to \mathbb{R}$ concave, $\forall\ m \geq 1$, $(y_1, ..., y_m) \in I^m$, $\forall\ (t_1, ..., t_m) \in \mathbb{R}_+^m$ s.t $\sum_i t_i = 1$, we have $f(\sum_i t_i y_i) \geq \sum_i t_i f(y_i)$*

   In particular, we can set $m = 2$, $f = \log$ and $t_1 = \frac{K}{K+1}$, $t_2 = \frac{1}{K+1}$. Then, we have:

$$\forall K \geq 0,\ \mathcal{L}_{K+1} = \mathbb{E}_{h_i \sim q(h)}\Big[\log\Big(\frac{1}{K+1}\sum_{i=1}^{K+1}\frac{p(x,h_i)}{q(h_i)}\Big)\Big] \tag{5}$$

$$= \mathbb{E}_{h_i \sim q(h)}\Big[\log\Big(\frac{K}{K+1}\Big\{\sum_{i=1}^{K}\frac{1}{K}\frac{p(x,h_i)}{q(h_i)}\Big\} + \frac{1}{K+1}\frac{p(x,h_{K+1})}{q(h_{K+1})}\Big)\Big] \tag{6}$$

$$\textit{Lemma } 3.1 \to \geq \frac{K}{K+1}\mathbb{E}_{h_i \sim q(h)}\Big[\log\Big(\sum_{i=1}^{K}\frac{1}{K}\frac{p(x,h_i)}{q(h_i)}\Big)\Big] + \frac{1}{K+1}\mathbb{E}_{h_i \sim q(h)}\Big[\log\frac{p(x,h_{K+1})}{q(h_{K+1})}\Big] \tag{7}$$

$$h_i \textit{ are i.i.d} \to = \frac{K}{K+1}\mathcal{L}_K + \frac{1}{K+1}\mathcal{L}_1 \tag{8}$$

$$\textit{by iterating } (8) \to \geq \frac{\cancel{K}}{K+1}\frac{\cancel{K-1}}{\cancel{K}}...\frac{1}{\cancel{2}}\mathcal{L}_1 + \underbrace{\frac{1}{K+1}\mathcal{L}_1 + ... + \frac{1}{K+1}\mathcal{L}_1}_{K\ times} \tag{9}$$

$$= \frac{1}{K+1}\mathcal{L}_1 + \frac{K}{K+1}\mathcal{L}_1 = \mathcal{L}_1 \tag{10}$$

   And then, we have $\forall K \geq 1,\ \mathcal{L}_K \geq \mathcal{L}_1$.

4 Question 4

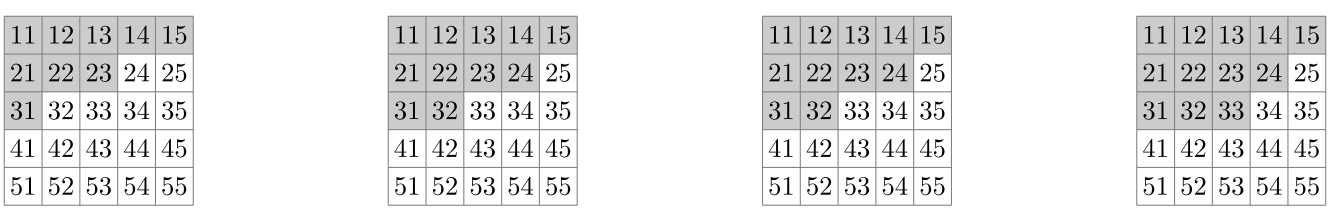

Figure 1: Receptive field under the masking schemes 1,2,3,4, in this order from left to right.

5 Question 5

Let's call S the shared support of f_1 and f_0. Then we have:

$$\mathbb{E}_{x\sim P_1}[\log D(x)] + \mathbb{E}_{x\sim P_0}[\log(1-D(x))] = \int_S \Big[\underbrace{\log D(x)f_1(x) + \log(1-D(x))f_0(x)}_{:=h(D)(x)}\Big]dx \tag{11}$$

As S is shared, then $\forall x \in S$, $h(D)(x)$ must be defined, which leads us to write that $D(S) \subset]0,1[$. Moreover, we have that h is a concave function of D ($u \mapsto \alpha \log u + \beta \log(1-u)$ is concave over $]0,1[$ $\forall \alpha, \beta > 0$), then to find the arg max in (11) it suffices to find the argument $D^* \in]0,1[$ that sets $\frac{\partial h}{\partial D}$ to 0, which leads us to:

$$\begin{aligned}\frac{\partial h}{\partial D} = \frac{f_1}{D} - \frac{f_0}{1-D} = 0 &\Leftrightarrow f_1(1-D^*) - f_0 D^* = 0\\ &\Leftrightarrow f_1 = \frac{f_0 D^*}{1-D^*}\end{aligned}$$

Then, as f_1 is a continuous function of D^*, using a D not too far from D^* will allow us to approximate well f_1.

6 Question 6

1. (a) We have that $f^*(t) = \sup_{u\in dom(f)} \big(\underbrace{ut - u\log u}_{:=g(u)}\big)$. Then, as g is a concave function of u, we can find u^* the argument that maximizes it by setting $\frac{\partial g}{\partial u}$ to 0:

$$\frac{\partial g}{\partial u} = 0 \Leftrightarrow t - \log(u) + 1 = 0 \Leftrightarrow u^* = e^{t-1}$$

We confirm that $u^* \in \mathbb{R}_+^* = dom(f)$ which is good. Then we have:

$$f^*(t) = sup_u g(u) = g(u^*) = e^{t-1}t - e^{t-1}(t-1) = e^{t-1}$$

In a similar way, let's define $f^{**}(t) = sup_{u\in dom(f^*)}(\underbrace{ut - e^{u-1}}_{:=h(u)})$. Again, h is concave in u and:

$$\frac{\partial h}{\partial u} = 0 \Leftrightarrow t - e^{u-1} = 0 \Leftrightarrow u^* = \log(t) + 1$$

Then, we have:

$$f^{**}(t) = h(u^*) = t(\log(t)+1) - e^{\log t} = t\log t$$

(b) Let's recall that $f^{**}(v) = sup_t(tv - e^{t-1}) = v\log v$ and let's define $u := \frac{p}{q}$. Then, we have:

$$\begin{aligned}\sup_T R_1[T] &= \sup_T \int p(x)T(x) - q(x)e^{T(x)-1}dx\\ &= \int \sup_{t\in\mathbb{R}} p(x)t - q(x)e^{t-1}dx\\ &= \int q(x)\Big[\underbrace{\sup_t \frac{p(x)}{q(x)}t - e^{t-1}}_{=f^{**}(u)}\Big]dx\\ &= \int q(x)\frac{p(x)}{q(x)}\log\frac{p(x)}{q(x)}dx\\ &= D_{KL}(p||q)\end{aligned}$$

2. (a) As $\mathbb{E}_q[e^T]$ is a scalar, we have:

$$\int r(x)q(x)dx = \frac{1}{\mathbb{E}_q[e^T]}\int e^{T(x)}q(x)dx = \frac{\mathbb{E}_q[e^T]}{\mathbb{E}_q[e^T]} = 1$$

and rq is a density function.

(b) As $\log r = T - \log \mathbb{E}_q[e^T]$, we have:

$$\begin{aligned} D_{KL}(p||q) &= \mathbb{E}_p\Big[\log\frac{p}{q}\Big] = \mathbb{E}_p\Big[\log\frac{pr}{qr}\Big] \\ &= \mathbb{E}_p\Big[\log\frac{p}{qr}\Big] + \mathbb{E}_p[T] - \mathbb{E}_p\Big[\log \mathbb{E}_q[e^T]\Big] \\ &= \underbrace{D_{KL}(p||qr)}_{\geq 0} + \mathbb{E}_p[T] - \log \mathbb{E}_q[e^T] \\ &\geq \mathbb{E}_p[T] - \log\mathbb{E}_q[e^T] = R_2[T] \end{aligned}$$

And we have equality i.i.f $D_{KL}(p||qr) = 0$ which is the case i.i.f $\forall x,\ p(x) = q(x)r(x)$. But we have:

$$\begin{aligned} \forall x,\ p(x) = q(x)r(x) &\Leftrightarrow \frac{p(x)}{q(x)} = \frac{e^{T(x)}}{\mathbb{E}_q[e^T]} \\ &\Leftrightarrow \log\frac{p(x)}{q(x)} + \underbrace{\log\mathbb{E}_q[e^T]}_{=c} = T(x) \end{aligned}$$

3. We have that the function $u \mapsto \log(u) - u/e$ is negative $\forall u \in \mathbb{R}_+^*$. Setting $u = \mathbb{E}_q[e^T]$ (which is strictly positive) we deduce that:

$$\begin{aligned} \log\mathbb{E}_q[e^T] - \frac{\mathbb{E}_q[e^T]}{e} \leq 0 &\Leftrightarrow -\mathbb{E}_q[e^{T-1}] \leq -\log\mathbb{E}_q[e^T] \\ &\Leftrightarrow \mathbb{E}_p[T] - \mathbb{E}_q[e^{T-1}] \leq \mathbb{E}_p[T] - \log\mathbb{E}_q[e^T] \\ &\Leftrightarrow R_1[T] \leq R_2[T] \end{aligned}$$

7 Question 7

As p, q have disjoint support, $\forall x \in Supp(p)$, $(p+q)(x) = p(x) + 0$ and reciprocally $\forall x \in Supp(q)$, $(p+q)(x) = 0 + q(x)$. Then, we can write:

$$\begin{aligned} D_{JS}(p||q) &= \frac{1}{2}D_{KL}(p||r) + \frac{1}{2}D_{KL}(q||r) \\ &= \int_{Supp(p)\cup Supp(q)} \Big(p(x)\log\frac{p(x)}{r(x)} + q(x)\log\frac{q(x)}{r(x)}\Big)dx \\ &= \frac{1}{2}\int_{Supp(p)\cup Supp(q)} \Big(p\log 2p + q\log 2q - (p+q)\log(p+q)\Big)(x)dx \\ &= \frac{1}{2}\int_{Supp(p)\cup Supp(q)} \Big((p+q)\log(2p+2q) - (p+q)\log(p+q)\Big)(x)dx \\ &= \frac{1}{2}\int_{Supp(p)\cup Supp(q)} \log(2)(p+q)(x)dx \\ &= \frac{1}{2}\times 2\times\log(2) = \log(2) \end{aligned}$$