Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Abhishek Pardhi
Roll Number: 200026
Date: May 2, 2023

QUESTION 1

It has been given that α-divergence is

$$D_\alpha(p||q) = \frac{4}{1-\alpha^2}\left(1 - \int p(z)^{\frac{(1+\alpha)}{2}} q(z)^{\frac{(1-\alpha)}{2}} dz\right)$$

Let's find $Lim_{\alpha\to 1} D_\alpha(p||q)$, but before that notice that the numerator as well as the denominator tends to 0, hence we can use L'hospital rule to find this kind of limit:
$\{$using the result $(a^x)' = a^x \, ln\, a\}$
$\Rightarrow Lim_{\alpha\to 1} \frac{-4}{-2\alpha} \int \left(\frac{ln\, p(z)}{2} p(z)^{\frac{(1+\alpha)}{2}} q(z)^{\frac{(1-\alpha)}{2}} - \frac{ln q(z)}{2} p(z)^{\frac{(1+\alpha)}{2}} q(z)^{\frac{(1-\alpha)}{2}}\right) dz$
$\Rightarrow Lim_{\alpha\to 1} \frac{1}{\alpha} \int ln \frac{p(z)}{q(z)}\, p(z)^{\frac{(1+\alpha)}{2}} q(z)^{\frac{(1-\alpha)}{2}} dz$
$\Rightarrow \int p(z) ln \frac{p(z)}{q(z)} dz = -\int p(z) ln \frac{q(z)}{p(z)} dz = KL(p||q)$
Hence, $KL(p||q)$ corresponds to α-divergence as $\alpha \to 1$.
Now similarly, we can find $Lim_{\alpha\to -1} D_\alpha(q||p)$:
$\Rightarrow Lim_{\alpha\to -1} \frac{1}{\alpha} \int ln \frac{p(z)}{q(z)}\, p(z)^{\frac{(1+\alpha)}{2}} q(z)^{\frac{(1-\alpha)}{2}} dz$
$\Rightarrow -\int q(z) ln \frac{p(z)}{q(z)} dz = KL(q||p)$
Hence, $KL(q||p)$ corresponds to α-divergence as $\alpha \to -1$.

Student Name: Abhishek Pardhi
Roll Number: 200026
Date: May 2, 2023

QUESTION 2

Given:

- Likelihood $p(x_n|\mu,\tau) = \mathcal{N}(\mu, \tau^{-1})$
- Priors $p(\mu) = \frac{1}{\sigma_\mu}$, $p(\tau) = \frac{1}{\tau}$
- let $\mathbf{X} = [x_1, x_2, \ldots, x_N]$

The log-joint will be:
$\Rightarrow log\, p(\mathbf{X}, \mu, \tau) = log\, p(\mathbf{X}|\mu,\tau) + log\, p(\mu|\tau) + log\, p(\tau)$
Since μ doesn't depend on τ we can write the above expression as:
$\Rightarrow log\, p(\mathbf{X}, \mu, \tau) = log\, p(\mathbf{X}|\mu,\tau) + log\, p(\mu) + log\, p(\tau)$
Let's find $q^*_\mu(\mu)$ by using the above expression (only keeping terms that involve μ)
$log\, q^*_\mu(\mu) = \mathbb{E}_{q_\tau}[log\, p(\mathbf{X}|\mu,\tau) + log\, p(\mu)] + const$
$\Rightarrow -\frac{\mathbb{E}_{q_\tau}[\tau]}{2}\Sigma_{n=1}^N (x_n - \mu)^2 + const$
Therefore, using [1] and putting $\lambda_0 = 0$, we get:
$q^*_\mu = \mathcal{N}(\mu|\mu_N, \lambda_N)$ where $\mu_N = \bar{x}$ and $\lambda_N = N\mathbb{E}_{q_\tau}[\tau]$
Let's find $q^*_\tau(\tau)$ by using the log-joint expression (only keeping terms that involve τ)
$log\, q^*_\tau(\tau) = \mathbb{E}_{q_\tau}[log\, p(\mathbf{X}|\mu,\tau) + log\, p(\tau)] + const$
$\Rightarrow -\frac{\tau}{2}\Sigma_{n=1}^N \mathbb{E}_{q_\mu}[(x_n - \mu)^2] + \frac{N}{2} log\, \tau - log\, \tau + const$
$q^*_\tau(\tau) \propto \tau^{\frac{N}{2}-1} exp(-\frac{\tau}{2}\Sigma_{n=1}^N \mathbb{E}_{q_\mu}[(x_n - \mu)^2])$
This is a Gamma function, therefore:
$q^*_\tau = Gamma(\tau|a_N, b_N)$ where $a_n = \frac{N}{2}$ and $b_N = \frac{\Sigma_{n=1}^N \mathbb{E}_{q_\mu}[(x_n-\mu)^2]}{2}$

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 3

Student Name: Abhishek Pardhi
Roll Number: 200026
Date: May 2, 2023

The conditional posterior of z_{dn}:
$p(z_{dn} = k|\mathbf{W}, \mathbf{Z}_{-dn}) \propto p(z_{dn} = k|\mathbf{Z}_{-dn}) \times p(w_{dn}|z_{dn} = k, \mathbf{W}_{-dn}, \mathbf{Z}_{-dn})$
The prior on z_{dn} is:
$p(z_{dn} = k|\mathbf{Z}_{-dn}) = \int p(z_{dn} = k|\mathbf{Z}_{-dn}, \theta_d)p(\theta_d|\mathbf{Z}_{-dn})d\theta_d$
$\Rightarrow \mathbb{E}_{p(\theta_d|\mathbf{Z}_{-dn})}[p(z_{dn} = k|\mathbf{Z}_{-dn}, \theta_d)]$
$\Rightarrow \mathbb{E}_{p(\theta_d|\mathbf{Z}_{-dn})}[\theta_{dk}]$
let's find $p(\theta_d|\mathbf{Z}_{-dn})$:
$p(\theta_d|\mathbf{Z}_{-dn}) \propto p(\theta_d)p(\mathbf{Z}_{-dn}|\theta_d) \Rightarrow \text{Dirichlet}(\alpha, \ldots, \alpha)\Pi_{i\neq n}\text{multinoulli}(\theta_d)$
$\Rightarrow \theta_{dk}^{(\alpha-1+\Sigma_{i\neq 1}\mathbb{I}[z_{di}=k])}$
Or we can say:
$p(\theta_d|\mathbf{Z}_{-dn}) = \text{Dirichlet}([\alpha + \Sigma_{i\neq n}\mathbb{I}[z_{di} = k]]_{k=1}^{K})$
Hence, we have the prior on z_{dn}:
$p(z_{dn}|\mathbf{Z}_{-dn}) = \frac{\alpha+\Sigma_{i\neq n}\mathbb{I}[z_{di}=k]}{K\alpha-1+N_d}$

Now let's find $p(w_{dn}|z_{dn} = k, \mathbf{W}_{-dn}, \mathbf{Z}_{-dn})$:
$p(w_{dn}|z_{dn} = k, \mathbf{W}_{-dn}, \mathbf{Z}_{-dn}) = \int p(w_{dn} = v|\phi_k)p(\phi_k|\mathbf{Z}_{dn}, \mathbf{W}_{-dn})d\phi_k$
$\Rightarrow \mathbb{E}_{p(\phi_k|\mathbf{Z}_{dn},\mathbf{W}_{-dn})}[p(w_{dn} = v|\phi_k)]$
$\Rightarrow \mathbb{E}_{p(\phi_k|\mathbf{Z}_{dn},\mathbf{W}_{-dn})}[\phi_{kv}]$
let's find $p(\phi_k|\mathbf{Z}_{dn}, \mathbf{W}_{-dn}) \propto p(\mathbf{W}_{-dn}|\phi_k, \mathbf{Z}_{dn})p(\phi_k)$
$\phi_k^{\eta}\Pi_{i\neq n}\Pi_{j\neq d}p(w_{ij}|\phi_k, z_{ij})$
$\Rightarrow \phi_k^{\eta+\Sigma_{i\neq n}\Sigma_{j\neq d}\mathbb{I}[w_{ij}=v]\mathbb{I}[z_{ij}=k]}$
Or we can say:
$p(\phi_k|\mathbf{Z}_{dn}, \mathbf{W}_{-dn}) = \text{Dirichlet}([\eta + \Sigma_{i\neq n}\Sigma_{j\neq d}\mathbb{I}[w_{ij} = v]\mathbb{I}[z_{ij} = k]]_{v=1}^{V})$
Hence, we have:
$p(w_{dn}|z_{dn} = k, \mathbf{W}_{-dn}, \mathbf{Z}_{-dn}) = \frac{\eta+\Sigma_{i\neq n}\Sigma_{j\neq d}\mathbb{I}[w_{ij}=v]\mathbb{I}[z_{ij}=k]}{V\eta+\Sigma_{i\neq n}\Sigma_{j\neq d}\mathbb{I}[z_{ij}=k]}$
Putting all of these together, we will get the posterior of z_{dn} as follows:
$p(z_{dn} = k|\mathbf{W}, \mathbf{Z}_{-dn}) \propto \frac{\alpha+\Sigma_{i\neq n}\mathbb{I}[z_{di}=k]}{K\alpha-1+N_d} \times \frac{\eta+\Sigma_{i\neq n}\Sigma_{j\neq d}\mathbb{I}[w_{ij}=v]\mathbb{I}[z_{ij}=k]}{V\eta+\Sigma_{i\neq n}\Sigma_{j\neq d}\mathbb{I}[z_{ij}=k]}$

Intuitive sense behind CP of z_{dn}

The probability of a word belonging to a particular topic is calculated by looking at the frequency of that word in the entire corpus, as well as in the current document (excluding current instance). The frequency of a word in the corpus helps to determine the topic vectors for the entire corpus, while the frequency in the document helps to determine the topic distribution for that particular document. Therefore, both the corpus and the document are taken into account to calculate the probability of a word belonging to a certain topic.

Gibbs Sampler

- Initialize $\mathbf{Z}^{(0)}$ and for $t = 1..T$ do:
- $z_{dn}^{(t)} \sim p(z_{dn} = k|\mathbf{W}, \mathbf{Z}_{-dn})$

We can use *Monte Carlo* approximation to compute an approximation of θ_{dk} and ϕ_{kv}:
First we take L samples $\{Z^{(l)}\}_{l=1}^{L}$ from the CP
then we find the approximation:
$\mathbb{E}[\theta_{dk}] = \Sigma_{l=1}^{L} \frac{\alpha + \Sigma_{i \neq n} \mathbb{I}[z_{di}=k]}{K\alpha - 1 + N_d}$
$\mathbb{E}[\phi_{kv}] = \Sigma_{l=1}^{L} \frac{\eta + \Sigma_{i \neq n} \Sigma_{j \neq d} \mathbb{I}[w_{ij}=v] \mathbb{I}[z_{ij}=k]}{V\eta + \Sigma_{i \neq n} \Sigma_{j \neq d} \mathbb{I}[z_{ij}=k]}$

Intuitive meaning of the expectations

The expected value of the topic distribution $\mathbb{E}[\theta_{dk}]$ for a given document d is determined by the number of words in the document assigned to the topic k based on the samples Z(s). Similarly, $\mathbb{E}[\phi_{kv}]$ is determined by the number of times the word v appears in the corpus assigned to topic k, as well as the number of words assigned to topic k across the entire corpus.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 4

Student Name: Abhishek Pardhi
Roll Number: 200026
Date: May 2, 2023

Given: $p(r_{ij}|\mathbf{u}_i, \mathbf{v}_j) = \mathcal{N}(r_{ij}|\mathbf{u}_i^T\mathbf{v}_j, \beta^{-1})$ and $p(r_{ij}|\mathbf{R}) = \int p(r_{ij}|\mathbf{u}_i, \mathbf{v}_j)p(\mathbf{u}_i, \mathbf{v}_j|\mathbf{R})d\mathbf{u}_i d\mathbf{v}_j$
Using Monte Carlo sampling, we can find approximation of mean and variance of any entry r_{ij} as follows:
$p(r_{i,j}|\mathbf{R}) = \frac{1}{S}\Sigma_{s=1}^{S}p(r_{ij}|\mathbf{u}_i^{(s)}, \mathbf{v}_j^{(s)})$ where $\mathbf{u}_i^{(s)}, \mathbf{v}_j^{(s)}$ are known since we know $\mathbf{U}^{(s)} = \{\mathbf{u}_i^{(s)}\}_{i=1}^{N}$ and $\mathbf{V}^{(s)} = \{\mathbf{v}_i^{(s)}\}_{i=1}^{N}$
Let's prove the following two identities:
Suppose we need to find $\mathbb{E}[f] = \int f(z)p(z)dz$, and we approximated this expectation using monte-carlo approximation $\hat{f} = \frac{1}{S}\Sigma_{s=1}^{S}f(z^{(s)})$
Claim 1: $\mathbb{E}[\hat{f}] = \mathbb{E}[f]$
Proof:
$\mathbb{E}[\hat{f}] = \mathbb{E}[\frac{1}{S}\Sigma_{s=1}^{S}f(z^{(s)})]$
$\Rightarrow \frac{1}{S}\Sigma_{s=1}^{S}\mathbb{E}[f(z^{(s)})]$
$\Rightarrow \frac{1}{S}\Sigma_{s=1}^{S}\mathbb{E}[f] = \mathbb{E}[f]$
Hence, proved
Claim 2: $Var[\hat{f}] = \frac{1}{S}\mathbb{E}[(f - \mathbb{E}[f])^2]$
Proof:
$Var[\hat{f}] = \mathbb{E}[\hat{f}^2] - \mathbb{E}[\hat{f}]^2$
$\Rightarrow \mathbb{E}[(\frac{1}{S}\Sigma_{s=1}^{S}f(z^{(s)}))^2] - \mathbb{E}[(\frac{1}{S}\Sigma_{s=1}^{S}f(z^{(s)}))]^2$
$\Rightarrow \mathbb{E}[\frac{1}{S^2}\Sigma_{s=1}^{S}f^2(z^{(s)}) + \frac{1}{S^2}\Sigma_{s=1}^{S}\Sigma_{l=1}^{L}f(z^{(s)})f(z^{(l)})] - \mathbb{E}[\frac{1}{S}\Sigma_{s=1}^{S}f(z^{(s)})]$
$\Rightarrow \frac{1}{S^2}\Sigma_{s=1}^{S}\mathbb{E}[f^2(z)] + \frac{1}{S^2}\Sigma_{s=1}^{S}\Sigma_{l=1}^{S}\mathbb{E}[f(z)]^2 - \mathbb{E}[\frac{1}{S}\Sigma_{s=1}^{S}f(z)]^2$
$\Rightarrow \frac{1}{S}\mathbb{E}[f^2(z)] - \frac{1}{S}\mathbb{E}[f(z)]^2$
$\frac{1}{S}Var[f]$
Hence, proved
Now using above claims we get $E[p(r_{i,j}|\mathbf{R})] = \mathbf{u}_i^T\mathbf{v}_j$
$\Rightarrow E[\hat{r_{ij}}] = E[r_{ij}] = \frac{1}{S}\Sigma_{s=1}^{S}\mathbf{u}_i^{(s)T}\mathbf{v}_i^{(s)}$
$\Rightarrow Var[\hat{r_{ij}}] = \frac{1}{S}Var[r_{ij}] = \frac{1}{\beta S}$
Hence, the required values are $E[\hat{r_{ij}}]$ and $Var[\hat{r_{ij}}]$.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Abhishek Pardhi
Roll Number: 200026
Date: May 2, 2023

QUESTION 5

Part 1:

First we tune the value of σ such that we get as smaller value of M as possible which almost envelopes the distribution. We find the value of M by finding the largest possible value of $\frac{\tilde{p}(x)}{q(x)}$ and assigning it to M. This is done in the python notebook. Below is the graph for Rejection Sampler where it can be seen that the samples from the Rejection Sampler(blue) resembles the actual distribution(green). The black dotted line is $Mq(x)$. The actual distribution is plotted by dividing the pdf values by area under the curve of the distribution.

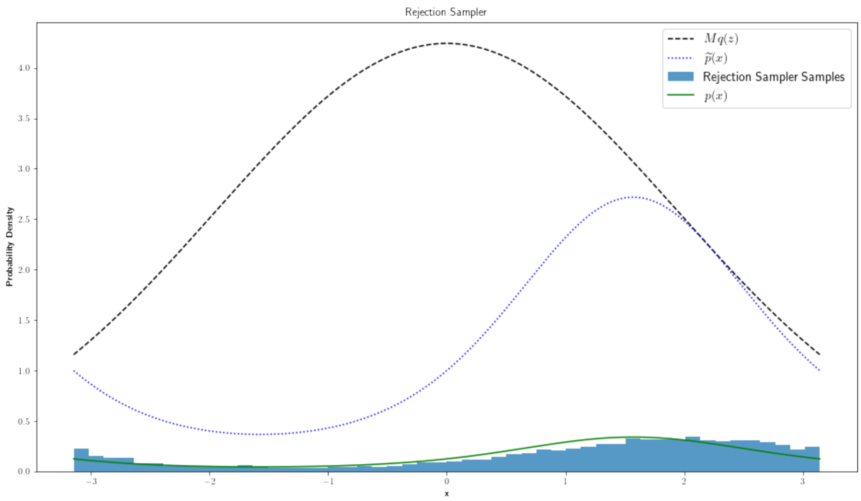

Figure 1: Rejection Sampler

Part 2:

- $\sigma^2 = 1$ is giving good approximation of the 2-D Gaussian where as $\sigma^2 = 0.01$ is performing badly. $\sigma^2 = 100$ is giving almost same approximation as $\sigma^2 = 1$ but took a long time to build due to high rejection rate.
- For $\sigma^2 = 0.01$, Rejection rate = 9.7%
- For $\sigma^2 = 1$, Rejection rate = 59.7%
- For $\sigma^2 = 100$, Rejection rate = 98.8%

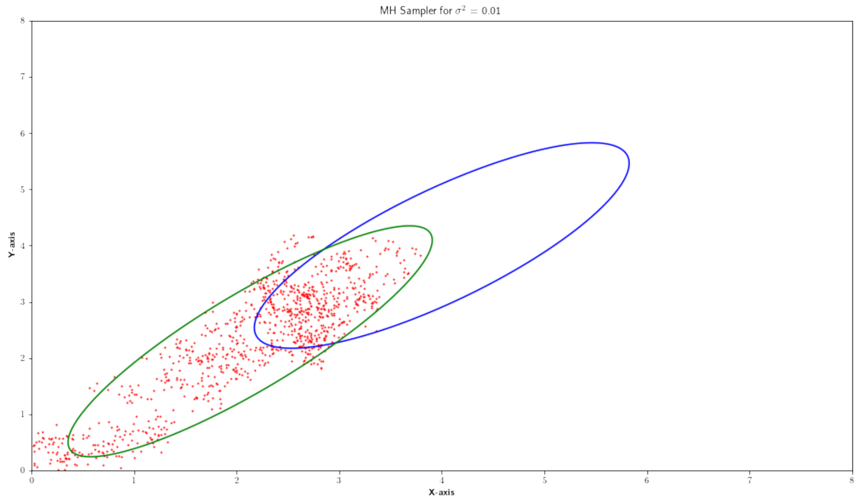

Figure 2: $\sigma^2 = 0.01$

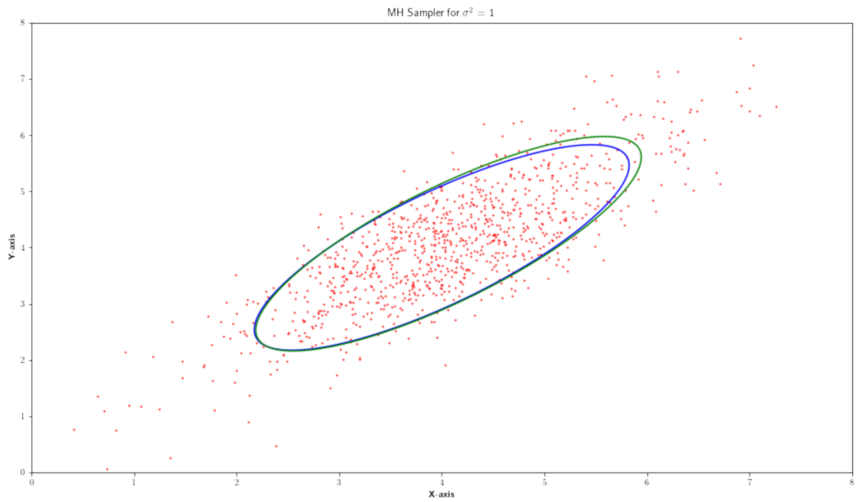

Figure 3: $\sigma^2 = 1$

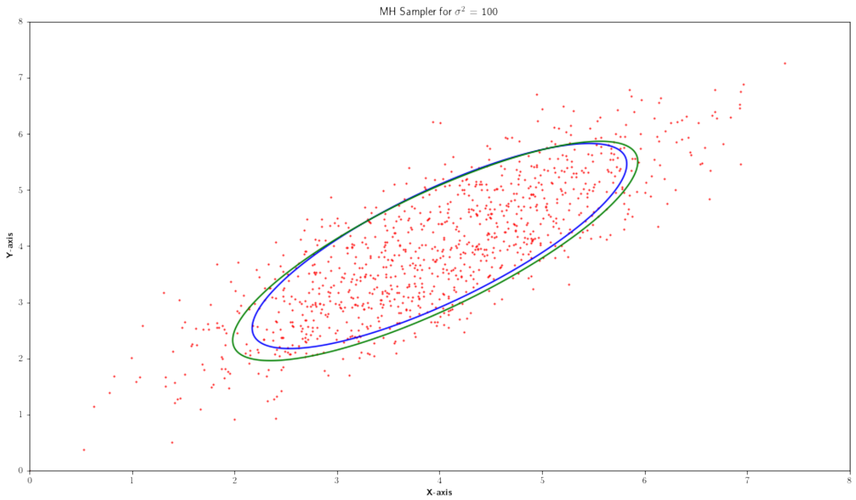

Figure 4: $\sigma^2 = 100$

References

[1] Piyush Rai. *CS772 Lecture*. Slide 13, Page 29.