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

Student Name: Shrey Mehta
Roll Number: 200580
Date: May 4, 2023

QUESTION 1

Standard VI minimizes the KL divergence between the true distribution p(z) and it's variational approximation q(z).
A more general form of divergence is the α-divergence defined as

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

To show: KL(p||q) corresponds to α-divergence as $\alpha \to 1$.
Proof:
Take α as $1-\epsilon$ and changing the limits to $\epsilon \to 0$, we have to prove that

$$\lim_{\epsilon\to 0} D_\epsilon(p||q) = \lim_{\epsilon\to 0} \frac{4}{2\epsilon - \epsilon^2}(1 - \int p(z)^{1-\frac{\epsilon}{2}} q(z)^{1+\frac{\epsilon}{2}} dz)$$

Since we have that $\epsilon \to 0$, so we can ignore the terms with ϵ^2 with respect to ϵ and using the results that we have in the question that $p^\epsilon = 1 + \epsilon log(p)$, we have

$$\lim_{\epsilon\to 0} D_\epsilon(p||q) = \lim_{\epsilon\to 0} \frac{4}{2\epsilon}(1 - \int p(z) \frac{1+\frac{\epsilon}{2}log(q)}{1+\frac{\epsilon}{2}log(p)} dz)$$

We know that in the binomial expansion of $(1+\frac{\epsilon}{2}log(p))^{-1}$, we can ignore the higher order terms and write it as $1-\frac{\epsilon}{2}log(p)$, so we have

$$\lim_{\epsilon\to 0} D_\epsilon(p||q) = \lim_{\epsilon\to 0} \frac{4}{2\epsilon}(1 - \int p(z)(1+\frac{\epsilon}{2}log(q))(1-\frac{\epsilon}{2}log(p))dz)$$

Multiplying the terms in the RHS and ignoring the terms containing ϵ^2, we have

$$\lim_{\epsilon\to 0} D_\epsilon(p||q) = \lim_{\epsilon\to 0} \frac{4}{2\epsilon}(1 - \int p(z)(1+\frac{\epsilon}{2}log(\frac{q(z)}{p(z)}))dz)$$

Also, integral of p(z) over the entire space will be 1, so replacing $\int p(z)dz = 1$, we have

$$\begin{aligned}
\lim_{\epsilon\to 0} D_\epsilon(p||q) &= \lim_{\epsilon\to 0} \frac{4}{2\epsilon}(- \int p(z)(\frac{\epsilon}{2}log(\frac{q(z)}{p(z)}))dz) \\
&= \lim_{\epsilon\to 0}(- \int p(z)(log(\frac{q(z)}{p(z)}))dz) \\
&= KL(p(z)||q(z))
\end{aligned}$$

So, we have that

$$\lim_{\alpha\to 1} D_\alpha(p||q) = KL(p||q)$$

To show: KL(q||p) corresponds to α-divergence as $\alpha \to$ -1.
Proof:
Now, consider the case of $\alpha \to -1$. Take α as $-1+\epsilon$ and changing the limits to $\epsilon \to 0$, we have to prove that

$$\lim_{\epsilon \to 0} D_\epsilon(p||q) = \lim_{\epsilon \to 0} \frac{4}{2\epsilon - \epsilon^2}(1 - \int p(z)^{\frac{\epsilon}{2}} q(z)^{1-\frac{\epsilon}{2}} dz)$$

Since we have that $\epsilon \to 0$, so we can ignore the terms with ϵ^2 with respect to ϵ and using the results that we have in the question that $p^\epsilon = 1 + \epsilon log(p)$, we have

$$\lim_{\epsilon \to 0} D_\epsilon(p||q) = \lim_{\epsilon \to 0} \frac{4}{2\epsilon}(1 - \int q(z) \frac{1 + \frac{\epsilon}{2} log(p)}{1 + \frac{\epsilon}{2} log(q)} dz)$$

We know that in the binomial expansion of $(1 + \frac{\epsilon}{2} log(q))^{-1}$, we can ignore the higher order terms and write it as $1 - \frac{\epsilon}{2} log(q)$, so we have

$$\lim_{\epsilon \to 0} D_\epsilon(p||q) = \lim_{\epsilon \to 0} \frac{4}{2\epsilon}(1 - \int q(z)(1 + \frac{\epsilon}{2} log(p))(1 - \frac{\epsilon}{2} log(q)) dz)$$

Multiplying the terms in the RHS and ignoring the terms containing ϵ^2, we have

$$\lim_{\epsilon \to 0} D_\epsilon(p||q) = \lim_{\epsilon \to 0} \frac{4}{2\epsilon}(1 - \int q(z)(1 + \frac{\epsilon}{2} log(\frac{p(z)}{q(z)})) dz)$$

Also, integral of p(z) over the entire space will be 1, so replacing $\int p(z)dz = 1$, we have

$$\begin{aligned}
\lim_{\epsilon \to 0} D_\epsilon(p||q) &= \lim_{\epsilon \to 0} \frac{4}{2\epsilon}(- \int q(z)(\frac{\epsilon}{2} log(\frac{p(z)}{q(z)})) dz) \\
&= \lim_{\epsilon \to 0}(- \int q(z)(log(\frac{p(z)}{q(z)})) dz) \\
&= KL(q(z)||p(z))
\end{aligned}$$

So, we have that

$$\lim_{\alpha \to -1} D_\alpha(p||q) = KL(q||p)$$

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

Student Name: Shrey Mehta
Roll Number: 200580
Date: May 4, 2023

QUESTION 2

We are given that

$$\mathbf{X} = \{x_1, x_2, ..., x_N\}$$

which are assumed to be generated iid from $\mathcal{N}(\mu, \tau^{-1})$ with the parameters having the priors

$$p(\mu) = \frac{1}{\sigma_\mu}$$
$$p(\tau) = \frac{1}{\tau}$$

where σ_μ is a constant.
Using the mean-field VI procedure, we approximate the posterior considering both the parameters not to be correlated to each other and inferring q over both the parameters separately,

$$\begin{aligned} p(\mu, \tau|\mathbf{X}) &\sim q(\mu, \tau) \\ &= q_\mu(\mu) q_\tau(\tau) \end{aligned}$$

From the lectures, we know the expression for q($\mathbf{Z}$) for all the latent variables $\mathbf{Z}$, given by

$$\log\, q^*(\mathbf{Z}_j) = \mathop{\mathbb{E}}_{i \neq j}[\log\, p(\mathbf{X}, \mathbf{Z})] + \text{c}$$

where c is a constant.
Here the latent variables $\mathbf{Z}$ are μ and τ.
Hence, we have that

$$\log\, q_\tau^*(\tau) = \mathop{\mathbb{E}}_{q_\mu}[\log\, p(\mathbf{X}, \mu, \tau)] + \text{c}$$

$$\log\, q_\mu^*(\mu) = \mathop{\mathbb{E}}_{q_\tau}[\log\, p(\mathbf{X}, \mu, \tau)] + \text{c}$$

$$\log\, p(\mathbf{X}, \mu, \tau) = \log\, p(\mathbf{X}|\mu, \tau) + \log\, p(\mu) + \log\, p(\tau)$$

Thus, putting the values as given in the lectures, we have

$$\log\, q_\tau^*(\tau) = \mathop{\mathbb{E}}_{q_\mu}[\log\, p(\mathbf{X}|\mu, \tau) + \log\, p(\tau)] + \text{c}$$

$$\log\, q_\mu^*(\mu) = \mathop{\mathbb{E}}_{q_\tau}[\log\, p(\mathbf{X}|\mu, \tau) + \log\, p(\mu)] + \text{c}$$

Now,

$$\log\, p(\mu) = -\log\, \sigma_\mu$$
$$\log\, p(\tau) = -\log\, \tau$$

Also, since $\mathbf{X}$ is derived from the Gaussian $\mathcal{N}(\mu, \tau^{-1})$, we can infer the posterior as

$$\log\, p(\mathbf{X}|\mu,\tau) = \prod_{n=1}^{N} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(x_n-\mu)^2}$$

Now, substituting the values in the expression for $q_\tau^*(\tau)$, we have

$$\begin{aligned}\log\, q_\tau^*(\tau) &= \mathop{\mathbb{E}}_{q_\mu}[\frac{N}{2}\log\,\tau - \frac{N}{2}\log(2\pi) - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2 - \log\,\tau] + \mathrm{c} \\ &= \big\{(\frac{N}{2}-1)log(\tau) - \frac{\tau}{2}\mathbb{E}_{q_\mu}\big[\sum_{n=1}^{N}(x_n-\mu)^2\big]\big\} + c\end{aligned}$$

The above is the log of Gamma distribution. So, we can write $q_\tau^* = Gamma(\tau|a_N, b_N)$. Now, from the results in the lectures, we have that

$$a_N = \frac{N}{2}$$

$$b_N = \frac{E_{q_\mu}[\sum_{n=1}^{N}(x_n-\mu)^2]}{2}$$

So, the optimal distribution of μ is a Gamma distribution with the above given shape and rate parameters and the maxima of the Gamma distribution obtained is given by the optimal value of q_τ which is

$$\tau_{\mathbf{opt}} = \frac{N-2}{E_{q_\mu}[\sum_{n=1}^{N}(x_n-\mu)^2]}$$

Now, substituting the values obtained in equation for $q_\mu^*(\mu)$, we can have that

$$\log\, q_\mu^*(\mu) = \mathop{\mathbb{E}}_{q_\tau}[\frac{N}{2}\log\,\tau - \frac{N}{2}\log(2\pi) - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2 - \log\,\sigma_\mu] + \mathrm{c}$$

Only keeping the terms containing μ, we have

$$\log\, q_\mu^*(\mu) = -\frac{\mathbb{E}_{q_\tau}}{2}[\tau]\left\{\sum_{n=1}^{N}(x_n-\mu)^2\right\} + \mathrm{c}$$

The above is the log of Gaussian. So, we can write $q_\mu^* = \mathcal{N}(\mu|\mu_N, \lambda_N)$

$$\mu_N = \bar{x}$$

where $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and

$$\lambda_N = (N\mathbb{E}_{q_\tau}[\tau])^{-1}$$

So, the optimal distribution of μ is a Gaussian distribution with the above given mean and variance and the maxima of a Gaussian is at it's mean, so we can infer the optimal value of μ as

$$\mu_{\mathbf{opt}} = \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

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

Student Name: Shrey Mehta
Roll Number: 200580
Date: May 4, 2023

QUESTION 3

Latent Dirichlet Allocation (LDA) is a probabilistic generative model used for topic modeling. In the LDA model, documents are represented as mixtures of latent topics, and each topic is represented as a distribution over words in the vocabulary. Given a document, the goal is to infer the topic mixture proportions and the topic assignments of each word in the document.
We are given the Latent Dirichlet Allocation (LDA) model as follows:

$$\phi_k \sim \text{Dirichlet}(\eta, ..., \eta), k = 1, ..., K$$

$$\theta_D \sim \text{Dirichlet}(\alpha, ..., \alpha), d = 1, ..., D$$

$$\mathbf{z}_{d,n} \sim \text{multinoulli}(\theta_D), n = 1, ..., N_d$$

$$\mathbf{w}_{d,n} \sim \text{multinoulli}(\phi_{\mathbf{z}_{d,n}})$$

In the above definitons of the LDA model, ϕ_k denotes the topic vector of topic k, θ_D denotes the k-dimensional topic mixing proportion vector and N_d is the number of words in document d. It is assumed that the vocabulary for the model has V distinct words.
The task is to derive a Gibbs sampler for the word-topic assignment variable $\mathbf{z}_{d,n}$.

Gibb's Sampler:
The conditional posterior inferred for $z_{d,n} = k$ is given by:

$$p(z_{d,n} = k|\mathbf{Z}_{-d,n}, \mathbf{W}) = p(w_{d,n}|z_{d,n} = k, \mathbf{Z}_{-d,n}, \mathbf{W}_{-d,n})p(z_{d,n} = k|\mathbf{Z}_{-d,n})$$

$$\begin{aligned} p(z_{d,n} = k|\mathbf{Z}_{-d,n}) &= \int p(x_{d,n} = k|\mathbf{Z}_{-d,n}, \theta_d)p(\theta_d|\mathbf{Z}_{-d,n})d\theta_d \\ &= \int \theta_{d,k}p(\theta_d|\mathbf{Z}_{-d,n})d\theta_d \\ &= \mathbb{E}_{p(\theta_d|\mathbf{Z}_{-d,n})}[\theta_{d,k}] \end{aligned}$$

We have that,

$$\begin{aligned} p(\theta_d|\mathbf{Z}_{-d,n}) &\propto p(\mathbf{Z}_{-d,n}|\theta_d)p(\theta_d) \\ &\propto \text{Dirichlet}(\alpha, ..., \alpha) \prod_{i=1, i\neq n}^{N_d} \text{multinoulli}(\theta_d) \\ &\propto (\theta_{d,k})^{\alpha-1} \prod_{i=1, i\neq n}^{N_d} (\theta_{d,k})^{\mathbb{I}[z_{d,i}=k]} \\ &\propto (\theta_{d,k})^{\alpha-1+\sum_{i=1, i\neq n}^{N_d} \mathbb{I}[z_{d,i}=k]} \end{aligned}$$

So, we have the result that

$$p(\theta_d|\mathbf{Z}_{-d,n}) = \text{Dirichlet}\left(\left\{\alpha + \sum_{i=1,i\neq n}^{N_d} \mathbb{I}[z_{d,i}=k]\right\}_{k=1}^{K}\right)$$

So,

$$\begin{aligned} p(z_{d,n}=k|\mathbf{Z}_{-\mathbf{d},\mathbf{n}}) &= \mathbb{E}_{p(\theta_d|\mathbf{Z}_{-d,n},\mathbf{W})}[\theta_{d,k}] \\ &= \frac{\alpha + \sum_{i=1,i\neq n}^{N_d} \mathbb{I}[z_{d,i}=k]}{K\alpha + N_d - 1} \end{aligned}$$

Hence,

$$\begin{aligned} p(w_{d,n}=v|z_{d,n}=k,\mathbf{Z}_{-d,n},\mathbf{W}_{-d,n}) &= \int p(w_{d,n}=v|\phi_k)p(\phi_k|\mathbf{Z}_{-d,n},\mathbf{W}_{-d,n})d\phi_k \\ &= \int \phi_{k,v}p(\phi_k|\mathbf{Z}_{-d,n},\mathbf{W}_{-d,n})d\phi_k \\ &= \mathbb{E}_{p(\phi_k|\mathbf{Z}_{-d,n},\mathbf{W}_{-d,n})}[\phi_{k,v}] \end{aligned}$$

Also,

$$\begin{aligned} p(\phi_k|\mathbf{Z}_{\mathbf{d},\mathbf{n}},\mathbf{W}_{-\mathbf{d},\mathbf{n}}) &\propto p(\mathbf{W}_{-d,n}|\phi_k,\mathbf{Z}_{d,n})p(\phi_k) \\ &\propto (\phi_k)^{\eta} \prod_{i=1,i\neq n}^{N_d}\prod_{j=1,j\neq d}^{D} p(w_{ij}|\phi_k,z_{ij}) \\ &\propto (\phi_k)^{\eta} \prod_{i=1,i\neq n}^{N_d}\prod_{j=1,j\neq d}^{D} (\phi_k)^{\mathbb{I}[w_{ij}=v]\mathbb{I}[z_{ij}=k]} \\ &\propto (\phi_k)^{\eta+\sum_{i=1,i\neq n}^{N_d}\sum_{j=1,j\neq d}^{D}\mathbb{I}[w_{ij}=v]\mathbb{I}[z_{ij}=k]} \end{aligned}$$

Therefore,

$$p(\phi_k|\mathbf{Z}_{\mathbf{d},\mathbf{n}},\mathbf{W}_{-\mathbf{d},\mathbf{n}}) = \text{Dirichlet}\left(\left\{\eta + \sum_{i=1,i\neq n}^{N_d}\sum_{j=1,j\neq d}^{D}\mathbb{I}[w_{ij}=v]\mathbb{I}[z_{ij}=k]\right\}_{v=1}^{V}\right)$$

And,

$$\begin{aligned} p(w_{dn}=v|z_{d,n}=k,\mathbf{Z}_{-d,n},\mathbf{W}_{-d,n}) &= \mathop{\mathbb{E}}_{p(\phi_d|\mathbf{Z}_{d,n},\mathbf{W}_{-d,n})}[\phi_{k,v}] \\ &= \frac{\eta + \sum_{i=1,i\neq n}^{N_d}\sum_{j=1,j\neq d}^{D}\mathbb{I}[w_{i,j}=v]\mathbb{I}[z_{i,j}=k]}{V\eta + \sum_{i=1,i\neq n}^{N_d}\sum_{j=1,j\neq d}^{D}\mathbb{I}[z_{i,j}=k]} \end{aligned}$$

Therefore, finally we get

$$\begin{aligned} p(z_{d,n}=k|\mathbf{Z}_{-d,n},\mathbf{W}) &= p(z_{d,n}=k|\mathbf{Z}_{-d,n})p(w_{d,n}|z_{d,n}=k,\mathbf{Z}_{-d,n},\mathbf{W}_{-d,n}) \\ &\propto \frac{\eta + \sum_{i=1,i\neq n}^{N_d}\sum_{j=1,j\neq d}^{D}\mathbb{I}[w_{ij}=v]\mathbb{I}[z_{i,j}=k]}{V\eta + \sum_{i=1,i\neq n}^{N_d}\sum_{j=1,j\neq d}^{D}\mathbb{I}[z_{i,j}=k]} \times \frac{\alpha + \sum_{i=1,i\neq n}^{N_d}\mathbb{I}[z_{d,i}=k]}{K\alpha + N_d - 1} \end{aligned}$$

To get the necessary conditional probability, normalise the above by adding the numerator across all k.
According to the theory, the likelihood that a word $(w_{d,n})$ belongs to topic (k) depends on both the frequency with which that word $(w_{d,n})$ appears in the corpus as a whole (excluding current occurrence) and the frequency with which it appears in the document as a whole (excluding current occurrence). Because the word $w_{d,n}$ depends on topic vectors, which comprise the entire corpus, we are searching the entire corpus for it. In contrast, since $z_{d,n}$, which is derived from θ_d, depends on the document d, we search through it.
The Gibbs sampler is a Markov chain Monte Carlo (MCMC) method used to approximate the posterior distribution of the LDA model. It is an iterative algorithm that generates samples of the latent variables, given the observed data. The goal is to estimate the posterior distribution of the latent variables, given the observed data.

Algorithm for Gibb's Sampling:

1. Initialize the latent variable matrix $\mathbf{Z} = \mathbf{Z}^{(0)}$ randomly and set t = 1. Note that for each $z_{d,n}$, the possible values are 1 to K.

2. Compute the following for the next T steps, i.e. while t $\neq$ T

$$\begin{aligned}
\pi_k^{(t)} &= p(z_{d,n}^{(t)} = k | \mathbf{Z}_{-d,n}^{(t-1)}, \mathbf{W}) \\
&\propto \frac{\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[w_{i,j} = v]\mathbb{I}[z_{i,j} = k]}{V\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[z_{i,j} = k]} \times \frac{\alpha + \sum_{i=1,i\neq n}^{N_d} \mathbb{I}[z_{d,i} = k]}{K\alpha + N_d - 1} \\
z_{d,n}^{(t)} &\sim \text{multinoulli}(\pi^{(t)}) \\
t &= t + 1
\end{aligned}$$

We can compute the expected values of θ_d and ϕ_k by applying Monte-Carlo approximation using the S samples of $\mathbf{Z}$ obtained, .

$$\begin{aligned}
\mathbb{E}[\theta_{d,k}] &= \frac{1}{S} \sum_{s=1}^{S} \frac{\alpha + \sum_{i=1,i\neq n}^{N_d} \mathbb{I}[z_{di} = k]}{K\alpha + N_d - 1} \\
\mathbb{E}[\phi_{k,v}] &= \frac{1}{S} \sum_{s=1}^{S} \frac{\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[w_{ij} = v]\mathbb{I}[z_{ij} = k]}{V\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[z_{ij} = k]}
\end{aligned}$$

For $\theta_{d,k}$, we want to estimate the probability of topic k in document d. This can be done by counting the number of times that topic k appears in document d in the set of samples $\mathbf{Z}^{(s)}$ that we have drawn. In other words, we are estimating the probability of topic k in document d based on the number of times we have observed topic k in document d in our set of samples. This estimation is then weighted by the hyperparameter α, which controls the overall sparsity of the document-topic distribution.
Similarly, for $\phi_{k,v}$, we want to estimate the probability of word v in topic k. This can be done by counting the number of times that word v appears in topic k across the entire

corpus in the set of samples $\mathbf{Z}^{(s)}$ that we have drawn. In other words, we are estimating the probability of word v in topic k based on the number of times we have observed word v in topic k in our set of samples. This estimation is then weighted by the hyperparameter β, which controls the overall sparsity of the topic-word distribution.
The intuition behind this estimation is that we assume that the observed data, i.e., the words in the corpus and their assignments to topics, are generated from a probabilistic process defined by the LDA model. By estimating the parameters of this model based on the observed data, we can infer the latent structure of the corpus, i.e., the underlying topics and their distributions over words and documents.

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

Student Name: Shrey Mehta
Roll Number: 200580
Date: May 4, 2023

QUESTION 4

We are given a matrix factorization model for a partially observed N $\times$ M matrix $\mathbf{R}$, where

$$p(r_{ij}|\mathbf{u_i}, \mathbf{v_j}) = \mathcal{N}(r_{ij}|\mathbf{u_i}^T\mathbf{v_j}, \beta^{-1})$$

and $\mathbf{u_i}$ and $\mathbf{v_j}$ denote the latent factors of i-th row and j-th column or $\mathbf{R}$, respectively. The PPD of each r_{ij} is defined a

$$p(r_{ij}|\mathbf{R}) = \int p(r_{ij}|\mathbf{u_i}, \mathbf{v_j})p(\mathbf{u_i}, \mathbf{v_j}|\mathbf{R})d\mathbf{u_i}d\mathbf{v_j}$$

We can use Monte-Carlo sampling for approximating the mean and the variance of the entries r_{ij},

$$p(r_{ij}|\mathbf{R}) = \frac{1}{S}\sum_{s=1}^{S} p(r_{ij}|\mathbf{u_i}^{(s)}, \mathbf{v_j}^{(s)})$$

where $\mathbf{u_i}^{(s)}$ and $\mathbf{v_j}^{(s)}$ are the ones generated by the Gibb's sampler for this matrix factorization model.
Now,

$$r_{ij} = \mathbf{u_i}^T\mathbf{v_j} + \epsilon_{ij}$$
$$\epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}|0, \beta^{-1})$$

So, the expected value of r_{ij} is given by

$$\begin{aligned}
\mathbb{E}[r_{ij}] &= \int_{r_{ij}} r_{ij}p(r_{ij}|\mathbf{R})dr_{ij} \\
&= \int_{r_{ij}} r_{ij}(\frac{1}{S}\sum_{s=1}^{S} p(r_{ij}|\mathbf{u_i}^{(s)}, \mathbf{v_j}^{(s)}))dr_{ij} \\
&= \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{p(r_{ij}|\mathbf{u_i}^{(s)}, \mathbf{v_j}^{(s)}))}[r_{ij}]
\end{aligned}$$

Now, using the linearity of expectation, we can have that

$$\begin{aligned}
\mathbb{E}[r_{ij}] &= \mathbb{E}[\mathbf{u_i}^T\mathbf{v_j}] + \mathbb{E}[\epsilon_{ij}] \\
&= \mathbf{u_i}^T\mathbf{v_j}
\end{aligned}$$

So, the expected value of r_{ij} is given by

$$\mathbb{E}[r_{ij}] = \frac{1}{S}\sum_{s=1}^{S} (\mathbf{u_i}^{(s)})^T\mathbf{v_j}^{(s)}$$

Now, the variance of r_{ij} is given by

$$var(r_{ij}) = \mathbb{E}[r_{ij}^2] - (\mathbb{E}[r_{ij}])^2$$

Now,

$$\begin{aligned}
\mathbb{E}[r_{ij}^2] &= \int_{r_{ij}} r_{ij}^2 p(r_{ij}|\mathbf{R}) dr_{ij} \\
&= \int_{r_{ij}} r_{ij}^2 (\frac{1}{S} \sum_{s=1}^{S} p(r_{ij}|\mathbf{u_i}^{(s)}, \mathbf{v_j}^{(s)})) dr_{ij} \\
&= \frac{1}{S} \sum_{s=1}^{S} \mathbb{E}_{p(r_{ij}|\mathbf{u_i}^{(s)}, \mathbf{v_j}^{(s)}))}[r_{ij}^2]
\end{aligned}$$

Now, since $r_{ij} = \mathbf{u_i}^T \mathbf{v_j} + \epsilon_{ij}$, we have that

$$r_{ij}^2 = (\mathbf{u_i}^T \mathbf{v_j})^2 + \epsilon_{ij}^2 + 2\epsilon_{ij} \mathbf{u_i}^T \mathbf{v_j}$$

Now, considering the fact that $\mathbf{u_i}^T \mathbf{v_j}$ is independent of ϵ_{ij}, we have that

$$\mathbb{E}[r_{ij}^2] = (\mathbf{u_i}^T \mathbf{v_j})^2 + \beta^{-1}$$

So, the variance of r_{ij} is given by

$$var(r_{ij}) = \frac{1}{S} \sum_{s=1}^{S} ((\mathbf{u_i}^{(s)})^T \mathbf{v_j}^{(s)})^2 + \beta^{-1} - (\frac{1}{S} \sum_{s=1}^{S} (\mathbf{u_i}^{(s)})^T \mathbf{v_j}^{(s)})^2$$

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

Student Name: Shrey Mehta
Roll Number: 200580
Date: May 4, 2023

QUESTION 5

1 Part 1: Implementing A Rejection Sampler

We know that

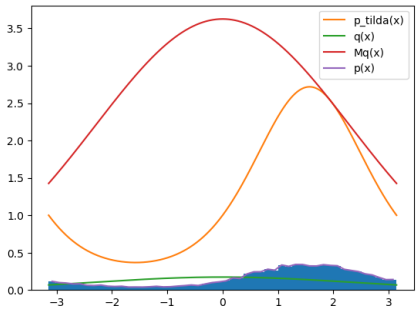
$$Mq(z) \geq \tilde{p}(x)$$

So, we have that

$$M \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \geq e^{sin(x)}$$

$$M \geq \sqrt{2\pi\sigma^2} e^{(sin(x)+\frac{x^2}{2\sigma^2})}$$

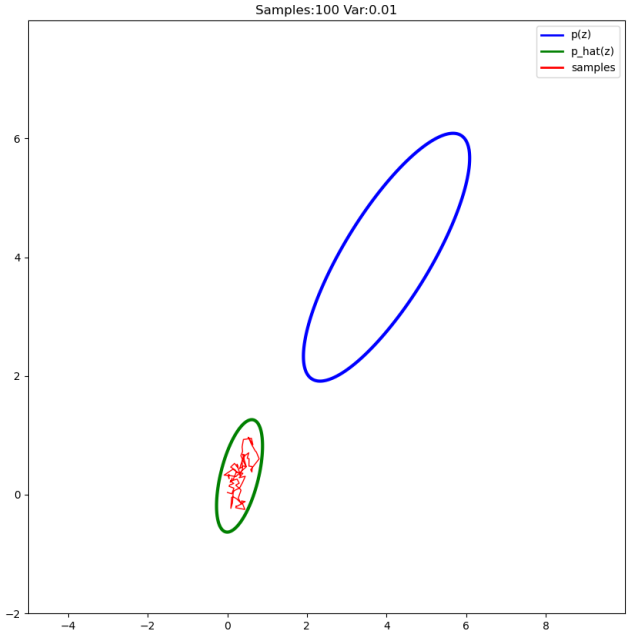
Now, we consider σ to be 2.3, to ease out the calculations, and from that, we get the optimum value of M obtained is 20.90, for which the histogram of the samples obtained is plotted below

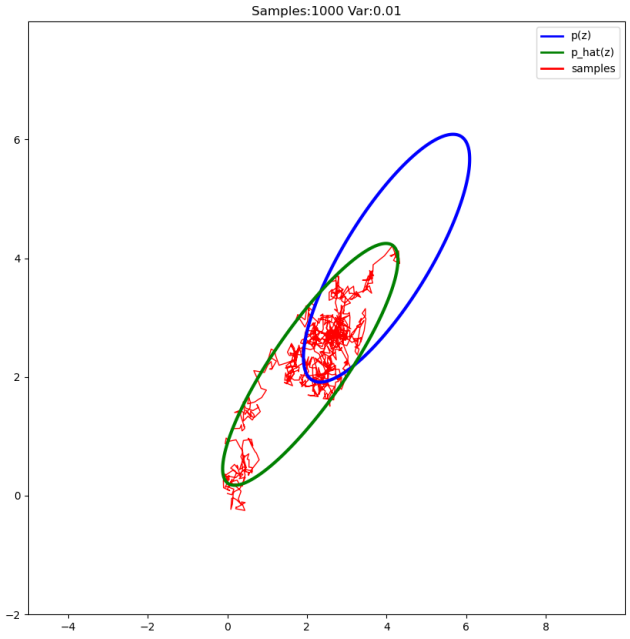


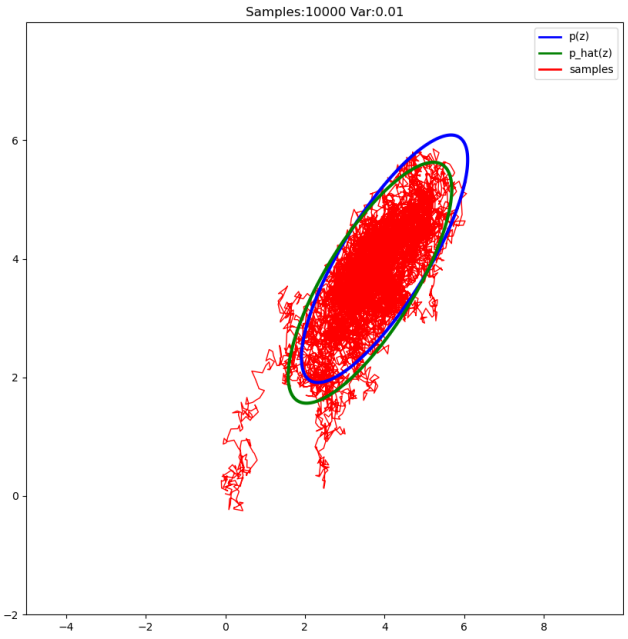
In the figure above, the blue coloured histograms are plotted, which correspond to the samples that are generated using rejection sampling and the green coloured plot is inference of p(x) from the generated samples. The acceptance ratio, in these conditions, is around 50%.

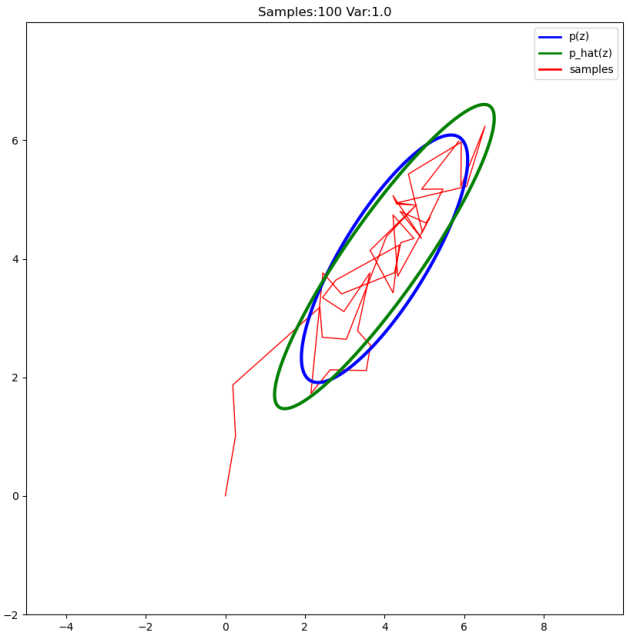
2 Part 2: Implementing MH Sampling for 2-D Gaussian

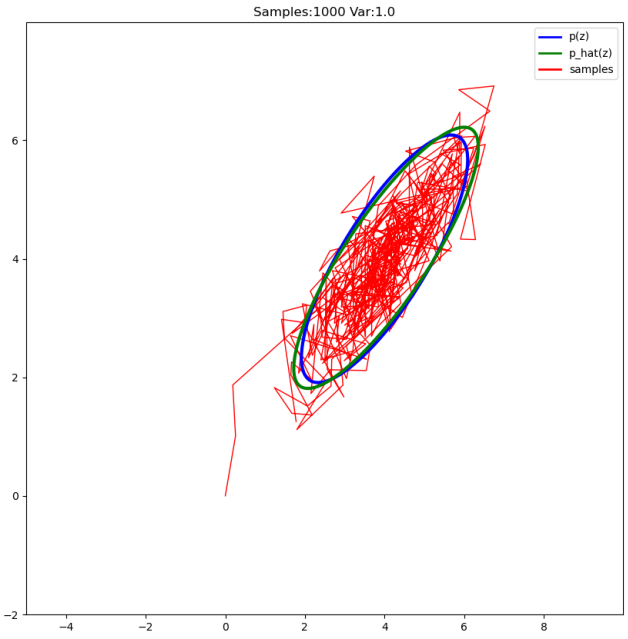
The plots of the generated samples on a 2-D plane for 100 samples, 1000 samples, and 10,000 samples are given below

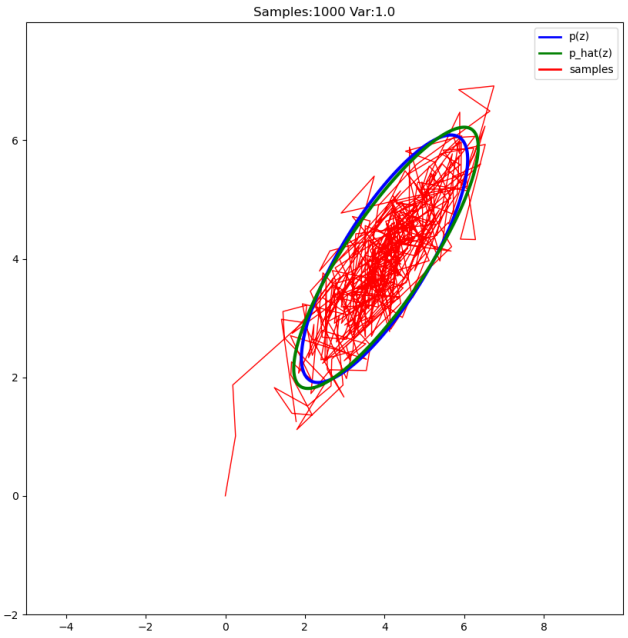


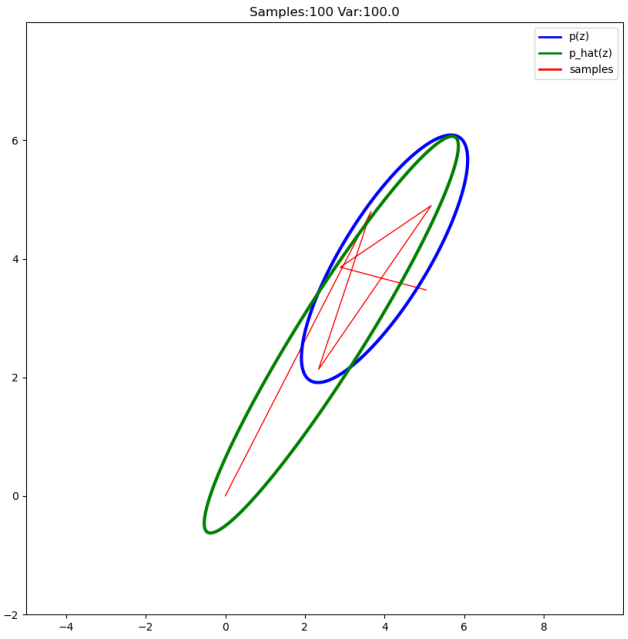


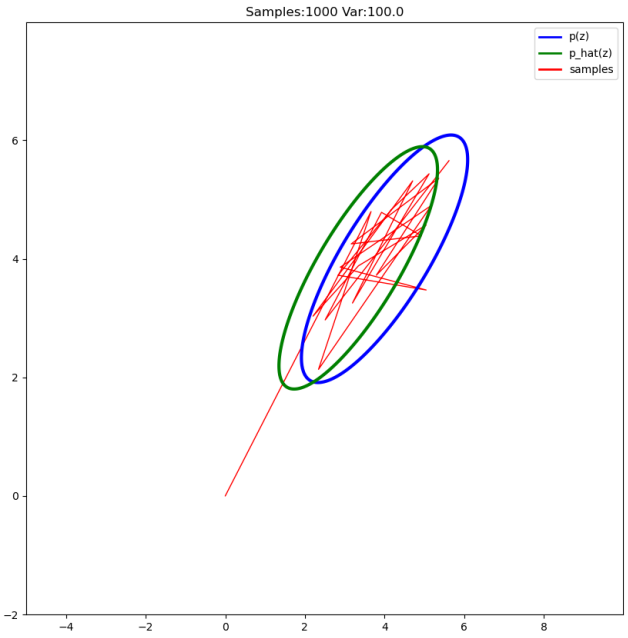

Samples:10000 Var:0.01
p(z)
p_hat(z)
samples
6
4
2
0
−2
−4
−2
0
2
4
6
8

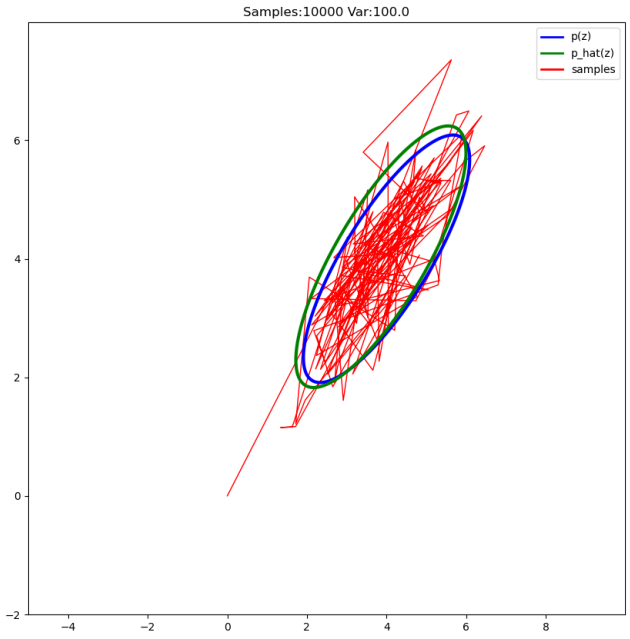

Samples:100 Var:1.0
p(z)
p_hat(z)
samples
6
4
2
0
−2
−4
−2
0
2
4
6
8


Samples:1000 Var:1.0
p(z)
p_hat(z)
samples
6
4
2
0
−2
−4
−2
0
2
4
6
8


Samples:1000 Var:1.0
p(z)
p_hat(z)
samples
6
4
2
0
−2
−4
−2
0
2
4
6
8


Samples:100 Var:100.0
p(z)
p_hat(z)
samples
6
4
2
0
−2
−4
−2
0
2
4
6
8


Samples:1000 Var:100.0
p(z)
p_hat(z)
samples
6
4
2
0
−2
−4
−2
0
2
4
6
8



In the Metropolis-Hastings algorithm, the choice of the proposal distribution variance plays a critical role in the algorithm's efficiency and accuracy. The proposal distribution determines how the Markov chain explores the target distribution. A low variance in the proposal distribution limits the range of values the Markov chain can explore, leading to slow convergence and a high rejection rate. In contrast, a high variance in the proposal distribution leads to a large range of values being explored but also increases the likelihood of the Markov chain moving far from the current state, potentially leading to many rejections and slow convergence.
In the case of the given example, when the proposal variance is set to 0.01, the proposal distribution is too narrow, which limits the range of values the Markov chain can explore, leading to slow convergence, and a very less rejection rate. On the other hand, when the proposal variance is set to 100, the proposal distribution is too wide, which causes the Markov chain to frequently move far from the current state, resulting in many rejections and slow convergence. Therefore, a variance of 1 seems to be the best choice as it balances the trade-off between exploration and exploitation, allowing the algorithm to efficiently explore the distribution while avoiding getting stuck in local optima or experiencing too many rejections.
The rejection rates obtained from the above sampling are as follows:

- **Proposal Variance: 0.01**
  Rejection Rate for 100 samples: 0.14
  Rejection Rate for 1000 samples: 0.098
  Rejection Rate for 10000 samples: 0.0837

- **Proposal Variance: 1.0**
  Rejection Rate for 100 samples: 0.56
  Rejection Rate for 1000 samples: 0.622
  Rejection Rate for 10000 samples: 0.5958

- **Proposal Variance: 100.0**
  Rejection Rate for 100 samples: 0.94
  Rejection Rate for 1000 samples: 0.979
  Rejection Rate for 10000 samples: 0.9867