Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Dishay Mehta
Roll Number: 200341
Date: May 2, 2023

QUESTION 1

We have to do limit $\alpha \to 1$ of $D_\alpha(p||q)$

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

Taking α=1-ϵ and changing the limits to $\epsilon \to 0$,

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

Ignoring the ϵ^2 term in denominator since its much smaller than ϵ and using $p^\epsilon = 1 + \epsilon(logp)$

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

Using binomial expansion of $(1+\epsilon/2(\log p))^{-1}$ as $(1-\epsilon/2(\log p))$

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

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

Ignoring the ϵ^2 term in numerator since its much smaller than ϵ

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

Using $\int p(z)dz = 1$ since its a true distribution

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

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

$$\lim_{\epsilon \to 0} D_\epsilon(p||q) = KL(p||q)$$

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

Now doing the same for the limit $\alpha \to$ -1 of $D_\alpha(p||q)$

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

Taking α=-1+ϵ and changing the limits to $\epsilon \to 0$,

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

Ignoring the ϵ^2 term in denominator since its much smaller than ϵ and using $p^\epsilon = 1 + \epsilon(logp)$

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

Using binomial expansion of $(1 + \epsilon/2(\log\, p))^{-1}$ as $(1 - \epsilon/2(\log\, p))$

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

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

Ignoring the ϵ^2 term in numerator since its much smaller than ϵ

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

Using $\int q(z)dz = 1$ since its a true distribution

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

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

$$\lim_{\epsilon \to 0} D_\epsilon(p||q) = KL(q||p)$$

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

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Dishay Mehta
Roll Number: 200341
Date: May 2, 2023

QUESTION 2

Given:

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

$$\mathbf{X} \sim \mathcal{N}(\mu, \tau^{-1})$$

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

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

Using mean-field VI, we approximate the posterior,

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

We are also given that,

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

Here the latent variables $\mathbf{Z}$ are μ and τ.
Thus we can write, We are also given that,

$$\log q^*_\mu(\mu) = \mathop{\mathbb{E}}_{q_\tau}[\log p(\mathbf{X}, \mu, \tau)] + \text{const} \tag{1}$$

$$\log q^*_\tau(\tau) = \mathop{\mathbb{E}}_{q_\mu}[\log p(\mathbf{X}, \mu, \tau)] + \text{const} \tag{2}$$

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

Thus,

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

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

$$\log p(\mathbf{X}|\mu, \tau) = \prod_{n=1}^{N} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(x_n - \mu)^2} \tag{3}$$

$$\log p(\mu) = -\log \sigma_\mu \tag{4}$$

$$\log p(\mu) = -\log \tau \tag{5}$$

Substituting the value of (3) and (4) in (1)

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

Only keeping the terms containing μ

$$= -\frac{\mathbb{E}_{q_\tau}}{2}[\tau]\left\{\sum_{n=1}^{N}(x_n-\mu)^2\right\} + \text{const}$$

The above is the log of Gaussian - $q^*_\mu = \mathcal{N}(\mu|\mu_N, \lambda_N)$
Using the results from lecture 14 below,

- Substituting $p(\mathbf{X}|\mu,\tau) = \prod_{n=1}^{N} p(x_n|\mu,\tau)$ and $p(\mu|\tau)$, we get

$$\begin{aligned}\log q^*_\mu(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathbf{X}|\mu,\tau) + \log p(\mu|\tau)] + \text{const}\\ &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2}\left\{\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right\} + \text{const}\end{aligned}$$

- (Verify) The above is log of a Gaussian. This $q^*_\mu = \mathcal{N}(\mu|\mu_N, \lambda_N)$ with

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N} \quad \text{and} \quad \lambda_N = (\lambda_0 + N)\mathbb{E}_{q_\tau}[\tau]$$

This update depends on q_τ

Putting $\lambda_0 = 0$

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

where $\bar{x}$ is the **mean** of all values of given x's.

$$\lambda_N = N \underset{q_\tau}{\mathbb{E}}[\tau]$$

For the optimal distribution of μ is a Gaussian with mean and variance as shown above and we know that the Gaussian distribution is maximum at its mean and hence,

$$\boxed{\mu_{\text{opt}} = \frac{\sum_{n=1}^{N} x_i}{N}}$$

Substituting the value of (3) and (5) in (2)

$$\log\, q^*_\tau(\tau) = \underset{q_\mu}{\mathbb{E}}[\frac{N}{2}\log\,\tau - \frac{N}{2}\log(2\pi) - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2 - \log\,\tau] + \text{const}$$

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

The above is the log of Gamma - $q^*_\tau = \text{Gamma}(\tau|a_N, b_N)$ where a_N is the shape parameter, b_N is the rate parameter
Using the results from lecture 14,

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

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

τ_{opt} for which the above distribution is maximised is,

$$\boxed{\tau_{\text{opt}} = \frac{N-2}{\mathbb{E}_{q_\mu}[\sum_{n=1}^{N}(x_n-\mu)^2]}}$$

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 3

Student Name: Dishay Mehta
Roll Number: 200341
Date: May 2, 2023

Consider the Latent Direchlet Allocation(LDA) model

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

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

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

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

In the above, ϕ_k denoted the V dim topic vector for topic k (assuming vocabulary of V unique words), θ_d denotes the K dim topic proportion vector for document d, and the number of words in document d in N_d.
Deriving a Gibbs sampler for the word-topic assignment $z_{d,n}$,
The CP is

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

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

$$= \int \theta_{d,k} p(\theta_d|\mathbf{Z}_{-d,n})d\theta_d$$

$$= \mathop{\mathbb{E}}_{p(\theta_d|\mathbf{Z}_{-d,n})}[\theta_{d,k}]$$

Now,

$$p(\theta_d|\mathbf{Z}_{-d,n}) \propto p(\mathbf{Z}_{-d,n}|\theta_d)p(\theta_d)$$

$$\propto \text{Dirichlet}(\alpha, \alpha, ...\alpha) \prod_{i=1, i\neq n}^{N_d} \text{multinoulli}(\theta_d)$$

$$\propto (\theta_{d,k})^{\alpha-1} \prod_{i=1, i\neq n}^{N_d} (\theta_{d,k})^{\mathbb{I}[z_{d,i}=k]}$$

$$\propto (\theta_{d,k})^{\alpha+\sum_{i=1,i\neq n}^{N_d} \mathbb{I}[z_{d,i}=k]-1}$$

So,

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

Therefore,

$$p(z_{d,n} = k|\mathbf{Z}_{-d,n}, \mathbf{W}) = \mathop{\mathbb{E}}_{p(\theta_d|\mathbf{Z}_{-d,n})}[\theta_{d,k}]$$

$$= \frac{\alpha + \sum_{i=1,i\neq n}^{N_d} \mathbb{I}[z_{d,i} = k]}{K\alpha + N_d - 1}$$

Now,

$$p(w_{d,n} = v|Z_{-d,n}, \mathbf{Z}_{-d,n}, \mathbf{W}_{-d,n}) = \int p(w_{d,n} = v|\phi_k)p(\phi_k|\mathbf{z}_{d,n}, \mathbf{W}_{-d,n})d\phi_k$$

$$= \int \phi_{k,v} p(\phi_k|\mathbf{z}_{d,n}, \mathbf{W}_{-d,n})d\phi_k$$

$$= \underset{p(\phi_k|\mathbf{z}_{d,n}, \mathbf{W}_{-d,n})}{\mathbb{E}} [\phi_{k,v}]$$

Also,

$$p(\phi_k|\mathbf{z}_{d,n}, \mathbf{W}_{-d,n}) \propto p(\mathbf{W}_{-d,n}|\phi_k, \mathbf{z}_{d,n})p(\theta_k)$$

$$\propto (\phi_k)^\eta \prod_{i=1,i\neq n}^{N_d} \prod_{j=1,j\neq d}^{D} p(w_{i,j}|\phi_k, z_{i,j})$$

$$\propto (\phi_k)^\eta \prod_{i=1,i\neq n}^{N_d} \prod_{j=1,j\neq d}^{D} (\phi_k)^{\mathbb{I}[w_{i,j}=v]\mathbb{I}[z_{i,j}=k]}$$

$$\propto (\phi_k)^{\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[w_{i,j}=v]\mathbb{I}[z_{i,j}=k]-1}$$

Therefore,

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

And,

$$p(w_{d,n} = v|z_{d,n} = k, \mathbf{Z}_{-d,n}, \mathbf{W}_{-d,n}) = \underset{p(\phi_d|\mathbf{z}_{d,n}, \mathbf{W}_{-d,n})}{\mathbb{E}} [\phi_{k,v}]$$

$$= \frac{\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[w_{i,j} = v]\mathbb{I}[z_{i,j} = k]}{V\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[z_{i,j} = k]}$$

Therefore, finally we get

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

$$\propto \frac{\alpha + \sum_{i=1,i\neq n}^{N_d} \mathbb{I}[z_{d,i} = k]}{K\alpha + N_d - 1} \times \frac{\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[w_{i,j} = v]\mathbb{I}[z_{i,j} = k]}{V\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[z_{i,j} = k]}$$

So, $p(z_{d,n}|\mathbf{Z}_{dn}, \mathbf{W})$is a multinoulli$(\pi_1, \pi_2, ...\pi_K)$
The expression implies that the the probability of the word $w_{d,n}$ belonging to topic k depends on the proportion of the number of times the words across the document belonged to topic k (excluding current occurrence) and proportion of the number of times the word $w_{d,n}$ across the corpus belonged to topic k (excluding current occurrence). $z_{d,n}$ which is drawn from θ_d depends on the document d, so we look across the document d,whereas for word $w_{d,n}$ we look across entire corpus because it depends on topic vectors which are kind of support for the entire corpus.

Sketch of Gibbs sampler:

- Initialize the latent variable matrix $\mathbf{Z} = \mathbf{Z}^{(0)}$ randomly. Note that for each $z_{d,n}$, the possible values are 1 to K. Set t = 1
-
$$
\begin{aligned}
\pi_k^{(t)} &= p(z_{d,n}^{(t)} = k|\mathbf{Z}_{-dn}^{(t-1)}, \mathbf{W}) \\
&\propto \frac{\alpha + \sum_{i=1,i\neq n}^{N_d} \mathbb{I}[z_{d,i} = k]}{K\alpha + N_d - 1} \times \frac{\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[w_{i,j} = v]\mathbb{I}[z_{i,j} = k]}{V\eta + \sum_{i=1,i\neq n}^{N_d} \sum_{j=1,j\neq d}^{D} \mathbb{I}[z_{i,j} = k]} \\
z_{d,n}^{(t)} &\sim \text{multinoulli}(\pi^{(t)})
\end{aligned}
$$
- t = t + 1. Go to step 2 if t $\neq$ T

NOTE:
Using S samples of $\mathbf{Z}$, we can compute the expected values of θ_d and ϕ_k by applying Monte-Carlo approximation.

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

Therefore $\mathbb{E}[\theta_{d,k}]$ depends on number of words in document d assigned to topic k based on samples $\mathbf{Z}^{(s)}$. Please note that the information of which topic a word belongs to is given by $\mathbf{Z}^{(s)}$.
Also, for $\mathbb{E}[\phi_{k,v}]$ depends on the number of times the word v belongs to the topic k in the entire corpus, and the number of words belonging to topic k accross the corpus, both wrt sample $\mathbf{Z}^{(s)}$.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Dishay Mehta
Roll Number: 200341
Date: May 2, 2023

QUESTION 4

Given:

$$p(r_{ij}|\mathbf{u}_i, \mathbf{v}_i) = \mathcal{N}(r_{ij}|\mathbf{u}_i^\intercal \mathbf{v}_j, \beta^{-1})$$

$$p(r_{ij}|\mathbf{R}) = \int p(r_{ij}|\mathbf{u}_i, \mathbf{v}_j) p(\mathbf{u}_i, \mathbf{v}_j|\mathbf{R}) d\mathbf{u}_i d\mathbf{v}_j$$

Using Monte Carlo sampling, we can find approximation of mean and varianceof any entry r_{ij} as follows:

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

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 $\mathbf{V}^{(s)} = \{\mathbf{v}_i^{(s)}\}_{i=1}^N$

$$r_{ij} = \mathbf{u}_i^\intercal \mathbf{v}_j + \epsilon_{ij} \text{ where } \epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}|0, \beta^1)$$

Expected value of r_{ij}:

$$\mathbb{E}[r_{ij}] = \int_{r_{ij}} p(r_{ij}|\mathbf{R}) r_{ij} dr_{ij}$$

$$= \int_{r_{ij}} \frac{1}{S}\sum_{s=1}^{S} p(r_{ij}|\mathbf{u}_i^{(s)}, \mathbf{v}_j^{(s)}) r_{ij} dr_{ij}$$

Simplifying this, we have

$$\mathbb{E}[r_{ij}] = \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{p(r_{ij}|\mathbf{u}_i^{(s)}, \mathbf{v}_j^{(s)})}[r_{ij}]$$

Using linearity of expectation,

$$\mathbb{E}[r_{ij}] = \mathbb{E}[\mathbf{u}_i^\intercal \mathbf{v}_j] + \mathbb{E}[\epsilon_{ij}]$$
$$= \mathbf{u}_i^\intercal \mathbf{v}_j$$

Thus, we have the expectation as follows

$$\boxed{\mathbb{E}[r_{ij}] = \frac{1}{S}\sum_{s=1}^{S} (\mathbf{u}_i^{(s)})^\intercal \mathbf{v}_j^{(s)}}$$

Variance of r_{ij}

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

We calculate $\mathbb{E}[r_{ij}^2]$ as follows,

$$\mathbb{E}[r_{ij}^2] = \int_{r_{ij}} p(r_{ij}|\mathbf{R}) r_{ij}^2 dr_{ij}$$

$$= \int_{r_{ij}} \frac{1}{S} \sum_{s=1}^{S} p(r_{ij}|\mathbf{u}_i^{(s)}, \mathbf{v}_j^{(s)}) r_{ij}^2 dr_{ij}$$

Simplifying this, we have

$$\mathbb{E}[r_{ij}^2] = \frac{1}{S} \sum_{s=1}^{S} \mathbb{E}_{p(r_{ij}|\mathbf{u}_i^{(s)}, \mathbf{v}_j^{(s)})}[r_{ij}^2]$$

$$r_{ij}^2 = (\mathbf{u}_i^\intercal \mathbf{v}_j)^2 + 2\epsilon_{ij}(\mathbf{u}_i^\intercal \mathbf{v}_j) + \epsilon_{ij}^2$$

As $(\mathbf{u}_i^\intercal \mathbf{v}_j)$ is independent of ϵ_{ij}

$$\mathbb{E}[r_{ij}^2] = (\mathbf{u}_i^\intercal \mathbf{v}_j)^2 + \beta^{-1}$$

Thus the variance is as follows,

$$\boxed{\mathrm{Var}(r_{ij}) = \frac{1}{S} \sum_{s=1}^{S} ((\mathbf{u}_i^{(s)})^\intercal \mathbf{v}_j^{s})^2 + \beta^{-1} - (\frac{1}{S} \sum_{s=1}^{S} (\mathbf{u}_i^{(s)})^\intercal \mathbf{v}_j^{(s)})^2}$$

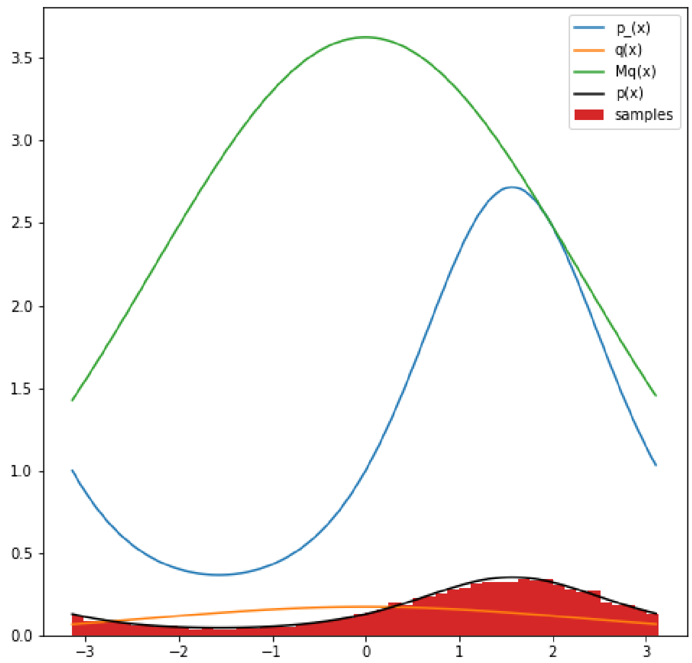
Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Dishay Mehta
Roll Number: 200341
Date: May 2, 2023

QUESTION 5

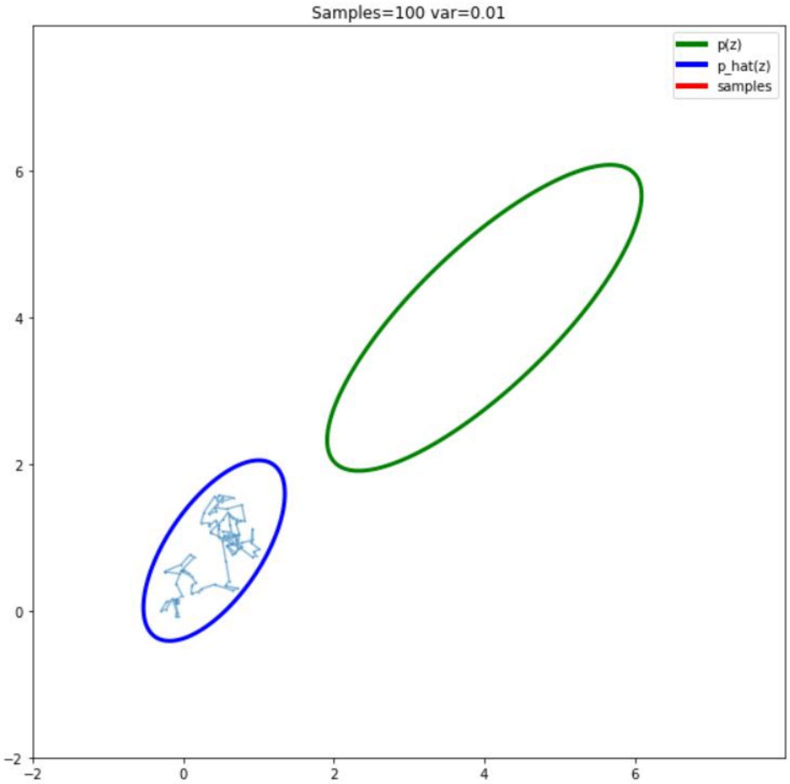
Part 1: Implementing a Rejection Sampler

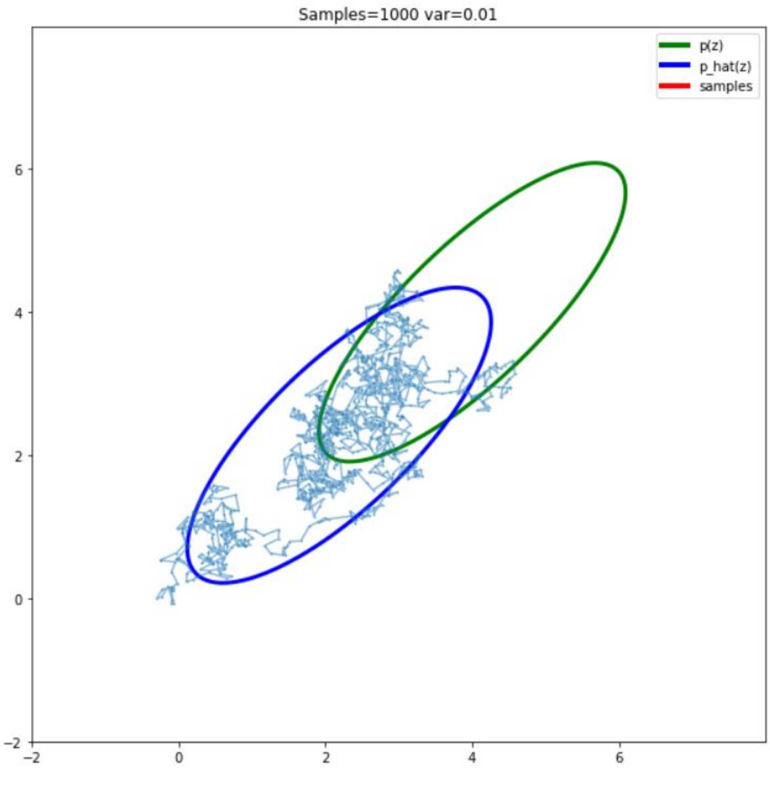
For $\sigma = 2.3$, the optimal value of M = 20.90
This is a histogram of 10000 samples based on rejection sampling by choosing an appropriate Gaussian proposal distribution.

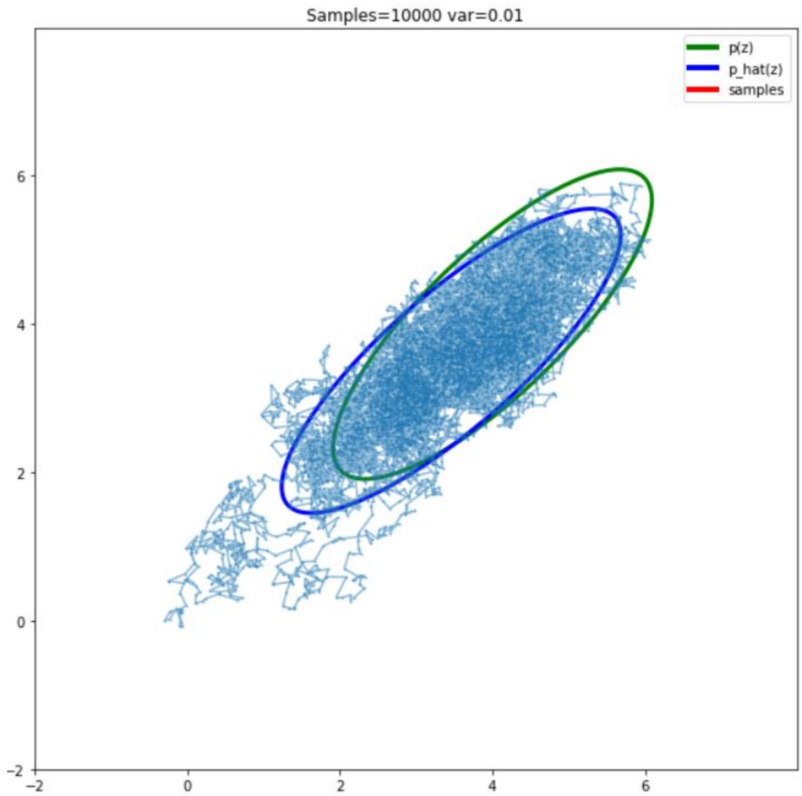


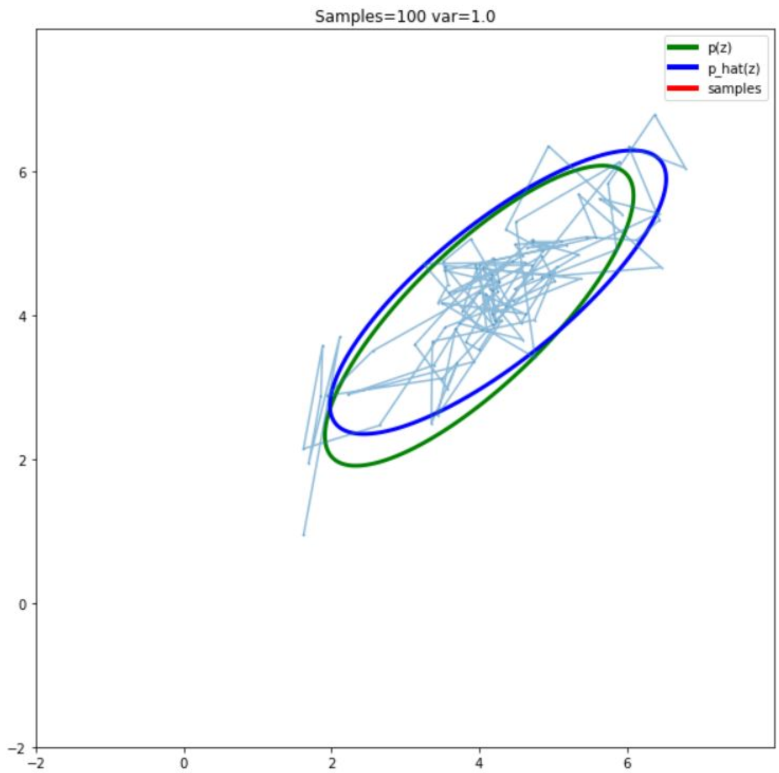
We can see from the above diagram that the $Mq(x)$ covers the $\tilde{p}(x) \propto \exp(\sin(x))$.
The red histograms are the samples generated by rejection sampling and the black curve is an approximate $p(x)$ distribution.
The acceptance rate comes out to be 52.5%.

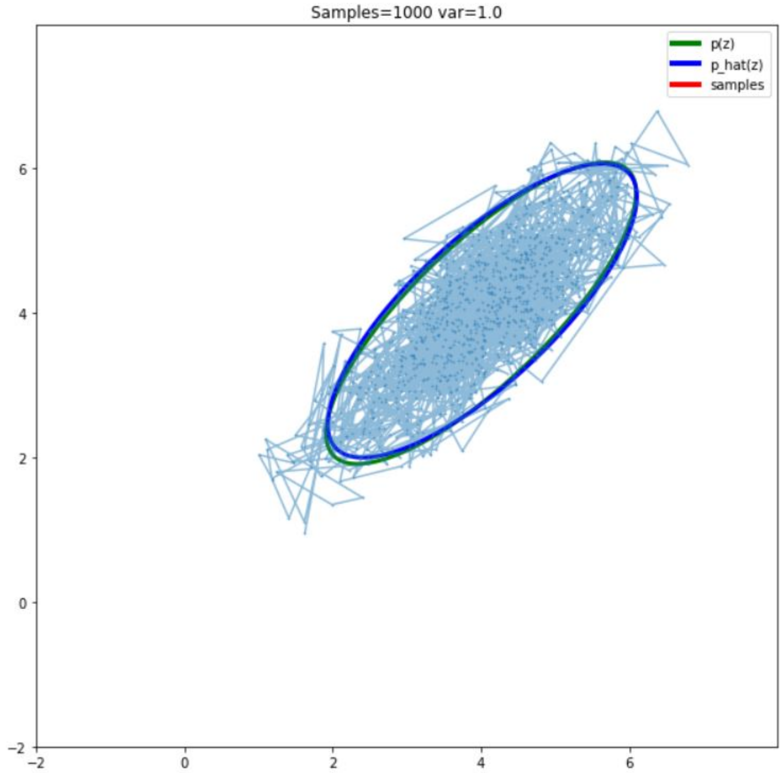
Part 2: Implementing MH Sampling for 2-D Gaussian

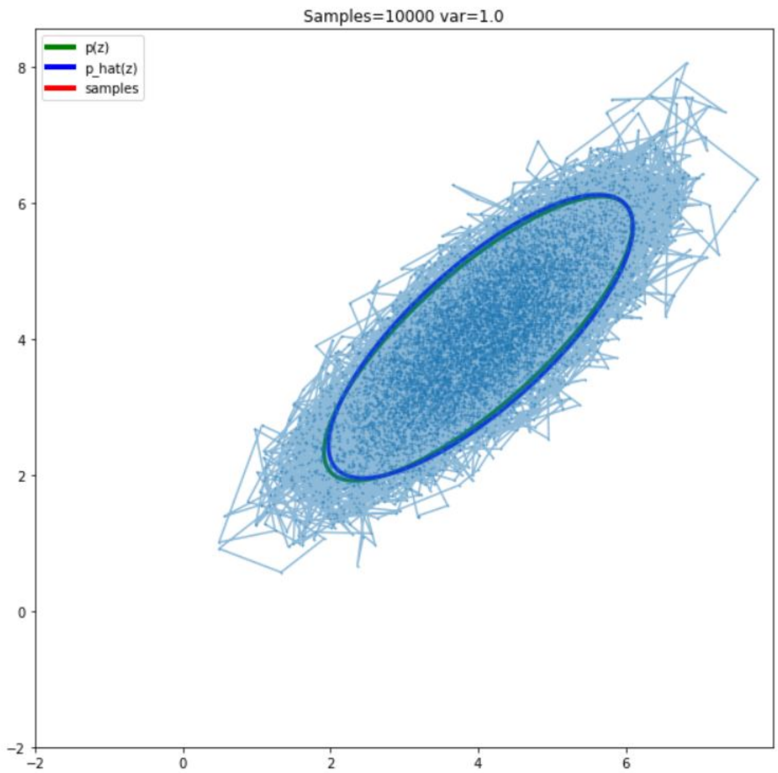

Samples=100 var=0.01
p(z)
p_hat(z)
samples
6
4
2
0
−2
−2
0
2
4
6

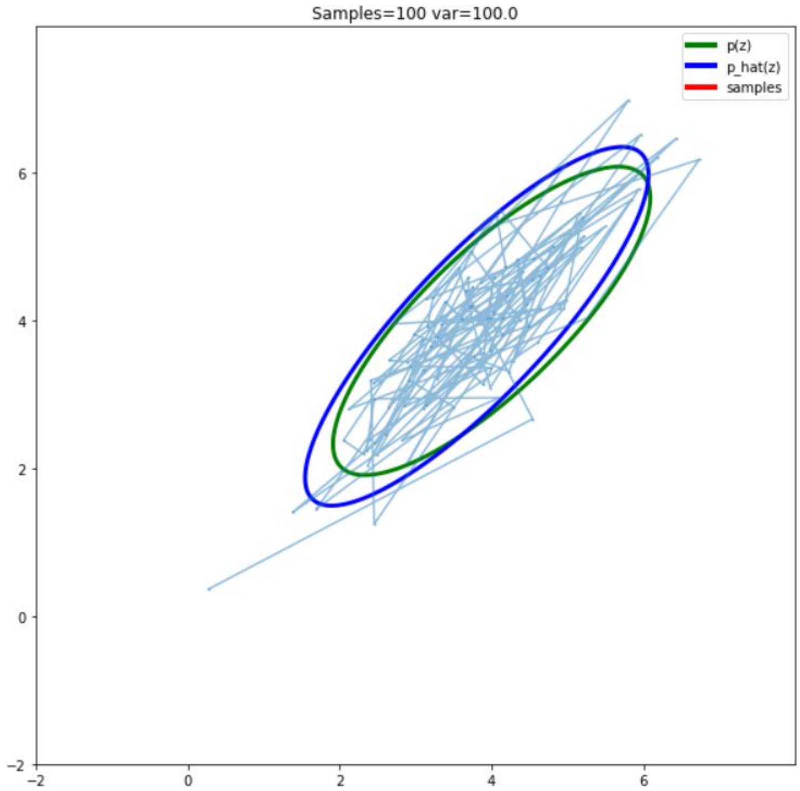

Samples=1000 var=0.01
p(z)
p_hat(z)
samples
6
4
2
0
−2
−2
0
2
4
6

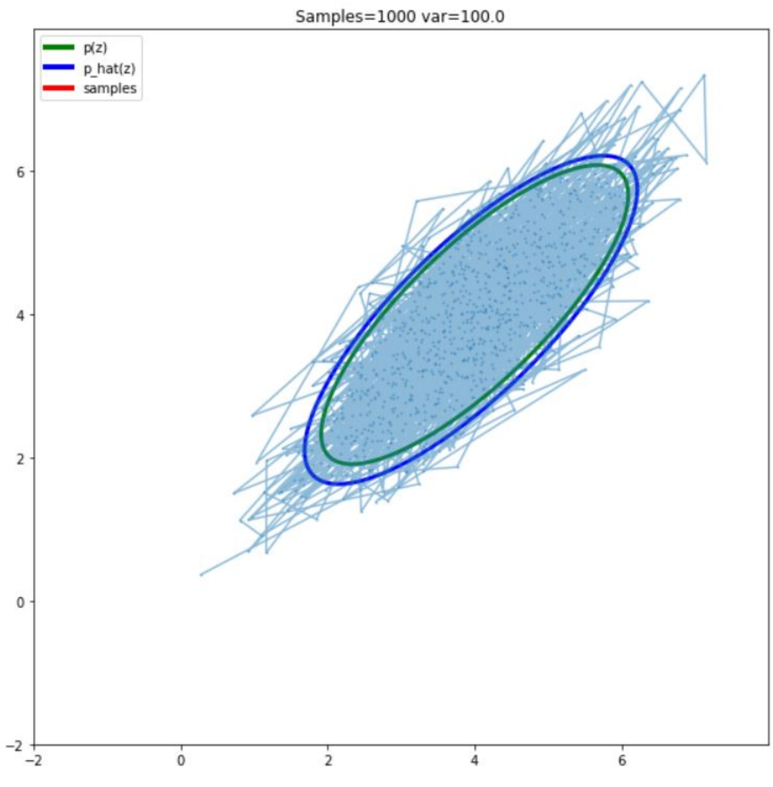

Samples=10000 var=0.01
p(z)
p_hat(z)
samples

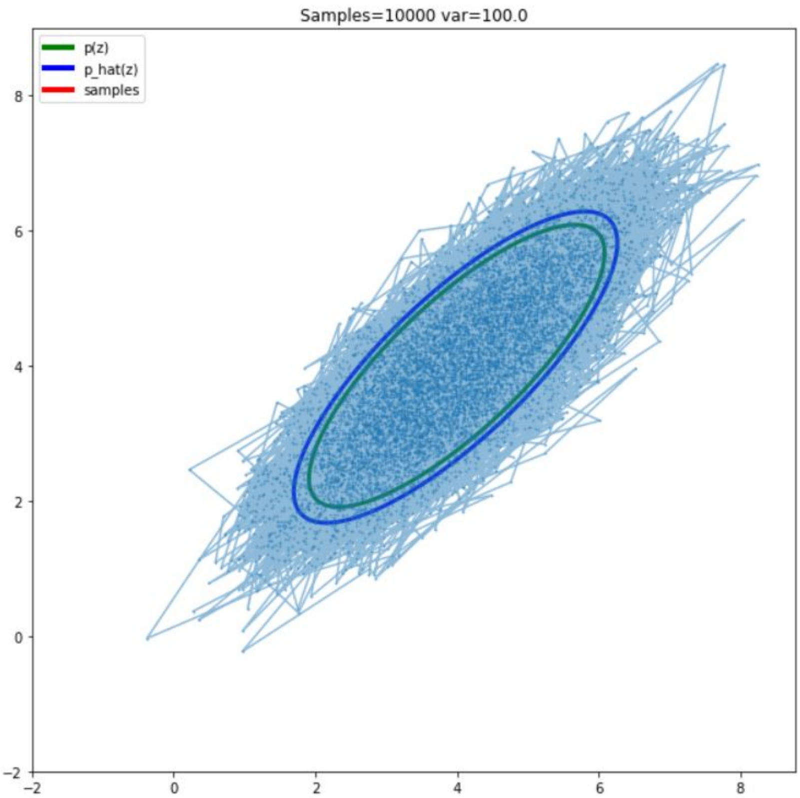

Samples=100 var=1.0
p(z)
p_hat(z)
samples


Samples=1000 var=1.0
p(z)
p_hat(z)
samples
6
4
2
0
−2
−2
0
2
4
6


Samples=10000 var=1.0
p(z)
p_hat(z)
samples
8
6
4
2
0
−2
−2
0
2
4
6


Samples=100 var=100.0
p(z)
p_hat(z)
samples


Samples=1000 var=100.0
p(z)
p_hat(z)
samples



NOTE: The sampling for var = 100 and 10000 samples takes about 10 to 15 minutes.

Looking at the plots, it seems that var = 1 seems to be the best choice because it strikes a balance between exploration and exploitation and allows the algorithm to efficiently explore the distribution while avoiding getting stuck in local optima or experiencing too many rejections. The fitting is also good enough. var = 100 has a higher diffusion rate while variance of 0.01 causes a lot of errors.

We get these results from sampling for different variances
Rejection rate for (var = 0.01) = 0.07620543136892666
Rejection rate for (var = 1.0) = 0.5973021944835917
Rejection rate for (var = 100.0) = 0.9884603757387506