Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 1

Student Name: Nidhi Hegde
Roll Number: 180472
Date: May 9, 2021

The KL- Divergence between 2 distributions $p(z)$ and $q(z)$ is given as:

$$KL(q(z) \mid\mid p(z)) = -\int \log\left(\frac{p(z)}{q(z)}\right) q(z)dz \tag{1}$$

Expanding the KL-term in the objective, it becomes:

$$\begin{aligned}
&\arg\min_{q(\theta)} - \left[\sum_{n=1}^{N} \int q(\theta)\log(p(x_n|\theta))d\theta + \int \log\left(\frac{p(\theta)}{q(\theta)}\right) q(\theta)d\theta\right] \\
= \quad &\arg\max_{q(\theta)} \left[\sum_{n=1}^{N} \int q(\theta)\log(p(x_n|\theta))d\theta + \int \log\left(\frac{p(\theta)}{q(\theta)}\right) q(\theta)d\theta\right]
\end{aligned}$$

Exchanging log and summation in the first term, and using properties of log, we can re-write it as:

$$\begin{aligned}
= \quad &\arg\max_{q(\theta)} \left[\int q(\theta)\log(\prod_{n=1}^{N} p(x_n|\theta))d\theta + \int \log\left(\frac{p(\theta)}{q(\theta)}\right) q(\theta)d\theta\right] \\
= \quad &\arg\max_{q(\theta)} \left[\int q(\theta)\log\left(\frac{\prod_{n=1}^{N} p(x_n|\theta)p(\theta)}{q(\theta)}\right) d\theta\right] \\
= \quad &\arg\max_{q(\theta)} \quad \mathbb{E}_{q(\theta)}\left[\log\left(\frac{p(\mathbf{X},\theta)}{q(\theta)}\right)\right]
\end{aligned} \tag{2}$$

The last expression is the ELBO, which is a lower bound on the marginal likelihood as discussed in class. The relation between ELBO and marginal likelihood is given as(derived using Bayes Rule at the end):

$$\log(p(\mathbf{X}|m)) = \mathbb{E}_{q(\theta)}\left[\log\left(\frac{p(\mathbf{X},\theta)}{q(\theta)}\right)\right] + KL(q(\theta)||p(\theta|\mathbf{X})) \tag{3}$$

Where $p(\mathbf{X}|m)$ can be further expanded as:

$$p(\mathbf{X}|m) = \int p(\mathbf{X},\theta|m)p(\theta|m)d\theta$$

We see that the marginal likelihood is independent of the variational parameters. Thus, we can write the above argmax problem as a minimization of KL Divergence between $p(\theta|\mathbf{x})$ and the distribution $q(\theta)$.

$$\begin{aligned}
q(\hat{\theta}) = \quad &\arg\max_{q(\theta)} \quad \mathbb{E}_{q(\theta)}\left[\log\left(\frac{p(\mathbf{X},\theta)}{q(\theta)}\right)\right] \\
= \quad &\arg\min_{q(\theta)} \; KL(q(\theta)||p(\theta|\mathbf{X})) \\
\implies \quad &q(\hat{\theta}) = p(\theta|\mathbf{X})
\end{aligned}$$

Now, arguing that $p(\theta|\mathbf{X})$ in Eq[3] is the posterior obtained by Bayes rule is sufficient to show that the solution obtained by solving the above objective is same as that obtained by Bayes Rule.
Derivation of Eq[3] from Bayes Rule:

$$\begin{aligned}
p(\mathbf{X}|m) &= \frac{p(\mathbf{X},\theta)}{p(\theta|\mathbf{X})} \\
\log(p(\mathbf{X}|m)) &= \log(p(\mathbf{X},\theta)) - \log(p(\theta|\mathbf{X})) && (Taking\ log) \\
log(p(\mathbf{X}|m)) &= \log(p(\mathbf{X},\theta)) - \log(q(\theta)) - \log(p(\theta|\mathbf{X})) + \log(q(\theta)) && (Add,\ subtract\ \log(q(\theta))) \\
log(p(\mathbf{X}|m)) &= \log\left(\frac{p(\mathbf{X},\theta)}{q(\theta)}\right) - \log\left(\frac{p(\theta|\mathbf{X})}{q(\theta)}\right) \\
\int log(p(\mathbf{X}|m))q(\theta)d\theta &= \int \log\left(\frac{p(\mathbf{X},\theta)}{q(\theta)}\right) q(\theta)d\theta - \int \log\left(\frac{p(\theta|\mathbf{X})}{q(\theta)}\right) q(\theta)d\theta && (Marginalize\ wrt\ \theta) \\
log(p(\mathbf{X}|m)) &= \mathbb{E}_{q(\theta)}\left[\log\left(\frac{p(\mathbf{X},\theta)}{q(\theta)}\right)\right] + KL(q(\theta)||p(\theta|\mathbf{X})) && (From\ [2]\ and\ [1])
\end{aligned}$$

Intuitive Explanation
The form of the objective as presented in question, consists of 2 terms.
The first term is like maximizing the probability of data,i.e. $\mathbb{E}_q[p(\mathbf{X}|\theta)]$.
The second term acts as a regularizer, minimizes the KL-divergence b/w $p(\theta)$ and $q(\theta)$, i.e. keeps the posterior probability low where the prior probability is low.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 2

Student Name: Nidhi Hegde
Roll Number: 180472
Date: May 9, 2021

Using the Mean-Field assumption, the variational distribution can be written as:

$$q(\mathbf{w}, \beta, \boldsymbol{\alpha}) = q(\mathbf{w})q(\beta)\prod_{d=1}^{D} q(\alpha_d) \tag{4}$$

The updates for the variational distributions of each parameter can be written in terms of expectations(w.r.t remaining unknowns) of the logarithm of the joint distribution as:

$$\begin{aligned}
\log(q^*(\mathbf{w})) &= \mathbb{E}_{q_\beta, q_{\boldsymbol{\alpha}}}[\log(p(\mathbf{y}, \mathbf{w}, \beta, \boldsymbol{\alpha}|\mathbf{X}))] + const \\
\log(q^*(\beta)) &= \mathbb{E}_{q_{\mathbf{w}}, q_{\boldsymbol{\alpha}}}[\log(p(\mathbf{y}, \mathbf{w}, \beta, \boldsymbol{\alpha}|\mathbf{X}))] + const \\
\log(q^*(\alpha_d)) &= \mathbb{E}_{q_\beta, q_{\mathbf{w}}, q_{\boldsymbol{\alpha}_{-d}}}[\log(p(\mathbf{y}, \mathbf{w}, \beta, \boldsymbol{\alpha}|\mathbf{X}))] + const \qquad \forall d \in [1, D]
\end{aligned}$$

The joint distribution can be obtained using Chain Rule as:

$$\begin{aligned}
p(\mathbf{y}, \mathbf{w}, \beta, \boldsymbol{\alpha}|\mathbf{X}) &= p(\mathbf{y}|\mathbf{w}, \beta, \boldsymbol{\alpha}, \mathbf{X})p(\mathbf{w}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\beta) \\
&= \left(\prod_{n=1}^{N} p(y_n|\mathbf{w}, \beta, \boldsymbol{\alpha}, \mathbf{x}_n)\right) p(\mathbf{w}|\boldsymbol{\alpha}) \left(\prod_{d=1}^{D} p(\alpha_d)\right) p(\beta)
\end{aligned}$$

where the distributions on the RHS are given as:

$$\begin{aligned}
p(y_n|\mathbf{w}, \beta, \boldsymbol{\alpha}, \mathbf{x}_n) &= \mathcal{N}(\mathbf{w}^T\mathbf{x}_n,\ \beta^{-1}), \qquad \forall n \\
p(\mathbf{w}|\boldsymbol{\alpha}) &= \mathcal{N}(0,\ \mathbf{D}), \qquad where\ \mathbf{D} = diag(\alpha_1^{-1}, \ldots, \alpha_D^{-1}) \\
p(\alpha_d) &= Gamma(\alpha_d \mid e_0, f_0), \quad \forall d \\
p(\beta) &= Gamma(\beta \mid a_0, b_0)
\end{aligned}$$

Thus, taking log and expanding the joint distribution, we get:

$$\begin{aligned}
\log(p(\mathbf{y}, \mathbf{w}, \beta, \boldsymbol{\alpha}|\mathbf{X})) &= \sum_{n=1}^{N} \log\left(\frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp(-\frac{\beta(y - \mathbf{w}^T\mathbf{x}_n)^2}{2})\right) + \log\left(\frac{\sqrt{\alpha_1\alpha_2 ... \alpha_D}}{\sqrt{(2\pi)^D}} \exp(-\frac{\mathbf{w}^T\mathbf{D}\mathbf{w}}{2})\right) \\
&\quad + \log\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\beta^{a_0-1}\exp(-b_0\beta)\right) + \sum_{d=1}^{D}\log\left(\frac{f_0^{e_0}}{\Gamma(e_0)}\alpha_d^{e_0-1}\exp(-f_0\alpha_d)\right) \\
&\propto \frac{N}{2}\log(\beta) - \sum_{n=1}^{N}\frac{\beta(y - \mathbf{w}^T\mathbf{x}_n)^2}{2} + \frac{1}{2}\sum_{d=1}^{D}\log(\alpha_d) - \frac{\mathbf{w}^T\mathbf{D}\mathbf{w}}{2} \\
&\quad + (a_0 - 1)\log(\beta) - b_0\beta \\
&\quad + (e_0 - 1)\sum_{d=1}^{D}\log(\alpha_d) - f_0\sum_{d=1}^{D}\alpha_d
\end{aligned}$$

Update of $q(\mathbf{w})$:
We keep the terms that involve $\mathbf{w}$ and take the expectation wrt $q_\beta, q_{\boldsymbol{\alpha}}$, where $q_{\boldsymbol{\alpha}} = [q_{\alpha_1}, \dots, q_{\alpha_D}]$.

$$\begin{aligned}
\log(q^*(\mathbf{w})) &= \mathbb{E}_{q_\beta, q_{\boldsymbol{\alpha}}}\left[-\sum_{n=1}^{N}\frac{\beta(y-\mathbf{w}^T\mathbf{x}_n)^2}{2} - \frac{\mathbf{w}^T\mathbf{D}\mathbf{w}}{2}\right] + const \\
&= -\frac{\mathbb{E}_{q_\beta}[\beta]}{2}\left(\sum_{n=1}^{N}(y-\mathbf{w}^T\mathbf{x}_n)^2\right) - \frac{\mathbf{w}^T\mathbb{E}_{q_{\boldsymbol{\alpha}}}[\mathbf{D}]\mathbf{w}}{2} + const \\
&= -\frac{\mathbb{E}_{q_\beta}[\beta]}{2}\left(\sum_{n=1}^{N}y_n^2 + \mathbf{w}^T\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T\mathbf{w} - 2\mathbf{w}^T\sum_{n=1}^{N}y_n\mathbf{x}_n\right) - \frac{\mathbf{w}^T\mathbb{E}_{q_{\boldsymbol{\alpha}}}[\mathbf{D}]\mathbf{w}}{2} + const
\end{aligned}$$

Rearranging and ignoring constant terms wrt $\mathbf{w}$, we get:

$$\log(q^*(\mathbf{w})) = -\frac{1}{2}\mathbf{w}^T\left(\left(\sum_{n=1}^{N}\mathbb{E}_{q_\beta}[\beta]\mathbf{x}_n\mathbf{x}_n^T\right) + \mathbb{E}_{q_{\boldsymbol{\alpha}}}[\mathbf{D}]\right)\mathbf{w} - \mathbb{E}_{q_\beta}[\beta]\mathbf{w}^T\sum_{n=1}^{N}y_n\mathbf{x}_n$$

which has the form of log of a gaussian distribution. Hence, $q^*(\mathbf{w})$ is given by a Gaussian distribution with parameters:

$$\begin{aligned}
&q^*(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \qquad s.t. \qquad (5)\\
&\quad \boldsymbol{\mu}_N = \left(\mathbb{E}_{q_\beta}[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T + \mathbb{E}_{q_{\boldsymbol{\alpha}}}[\mathbf{D}]\right)^{-1}\mathbb{E}_{q_\beta}[\beta]\sum_{n=1}^{N}y_n\mathbf{x}_n \\
&\quad \boldsymbol{\Sigma}_N = \mathbb{E}_{q_\beta}[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T + \mathbb{E}_{q_{\boldsymbol{\alpha}}}[\mathbf{D}], \qquad where \;\; \mathbb{E}[\mathbf{D}] = diag(\mathbb{E}[\alpha_1], \dots, \mathbb{E}[\alpha_D])
\end{aligned}$$

Update of $q(\beta)$:
We keep the terms that involve β and take the expectation wrt $q_{\mathbf{w}}, q_{\boldsymbol{\alpha}}$, where $q_{\boldsymbol{\alpha}} = [q_{\alpha_1}, \dots, q_{\alpha_D}]$.

$$\begin{aligned}
\log(q^*(\beta)) &= \mathbb{E}_{q_{\mathbf{w}}, q_{\boldsymbol{\alpha}}}\left[\frac{N}{2}\log(\beta) - \sum_{n=1}^{N}\frac{\beta(y-\mathbf{w}^T\mathbf{x}_n)^2}{2} + (a_0-1)\log(\beta) - b_0\beta\right] + const \\
&= (\frac{N}{2} + a_0 - 1)\log(\beta) - b_0\beta - \frac{\beta}{2}\left(\sum_{n=1}^{N}y_n^2 - 2\mathbb{E}[\mathbf{w}^T]\sum_{n=1}^{N}y_n\mathbf{x}_n + \sum_{n=1}^{N}\mathbf{x}_n^T\mathbb{E}[\mathbf{w}\mathbf{w}^T]\mathbf{x}_n\right)
\end{aligned}$$

Rearranging and ignoring constant terms wrt β, we get:

$$\log(q^*(\beta)) = \log(\beta)\left(\frac{N}{2} + a_0 - 1\right) - \beta\frac{\left(\sum_{n=1}^{N}y_n^2 - 2\mathbb{E}[\mathbf{w}^T]\sum_{n=1}^{N}y_n\mathbf{x}_n + \sum_{n=1}^{N}\mathbf{x}_n^T\mathbb{E}[\mathbf{w}\mathbf{w}^T]\mathbf{x}_n + 2b_0\right)}{2}$$

Thus, it has the form of a Gamma distribution:

$$\begin{aligned}
&q^*(\beta) = Gamma(\beta|a_0', b_0') \qquad s.t. \qquad (6)\\
&\quad a_0' = \frac{N}{2} + a_0 \\
&\quad b_0' = \frac{\left(\sum_{n=1}^{N}y_n^2 - 2\mathbb{E}[\mathbf{w}^T]\sum_{n=1}^{N}y_n\mathbf{x}_n + \sum_{n=1}^{N}\mathbf{x}_n^T\mathbb{E}[\mathbf{w}\mathbf{w}^T]\mathbf{x}_n + 2b_0\right)}{2}
\end{aligned}$$

Update of $q(\boldsymbol{\alpha})$:
For computing the variational distribution of α_d, we only keep the terms that involve α_d and take the expectation wrt $q_{\mathbf{w}}, q_\beta, q_{\boldsymbol{\alpha}_{-d}}$, where $q_{\boldsymbol{\alpha}_{-d}} = [q_{\alpha_1}, \ldots, q_{\alpha_D}] - \{q_{\alpha_d}\}$

$$\begin{aligned}\log(q^*(\alpha_d)) &= \mathbb{E}_{q_{\mathbf{w}}, q_\beta, q_{\boldsymbol{\alpha}_{-d}}}\left[\left(\frac{1}{2} + e_0 - 1\right)\log(\alpha_d) - f_0\alpha_d - \frac{w_d^2\alpha_d}{2}\right] + const \\ &= \left(\frac{1}{2} + e_0 - 1\right)\log(\alpha_d) - f_0\alpha_d - \frac{\mathbb{E}[w_d^2]\alpha_d}{2} + const\end{aligned}$$

Rearranging and ignoring constant terms wrt α_d, we get:

$$\log(\alpha_d)\left(\frac{1}{2} + e_0 - 1\right) - \alpha_d\left(f_0 + \frac{\mathbb{E}[w_d^2]}{2}\right)$$

This also has the form of a Gamma distribution given as:

$$\begin{aligned}q^*(\alpha_d) &= Gamma(\beta|e'_d, f'_d) \quad \forall d \quad s.t. \\ e'_d &= \frac{1}{2} + e_0 \\ f'_d &= f_0 + \frac{\mathbb{E}[w_d^2]}{2}\end{aligned} \tag{7}$$

The above updates for $q(\mathbf{w}), q(\beta), q(\boldsymbol{\alpha})$ are performed in an alternating fashion to yield the final variational distribution upon convergence.

The entire **Mean-Field VI Algorithm** can thus be summarized as:

Input : Data $\mathbf{X}, \mathbf{y}$, Joint Distribution $p(\mathbf{y}, \mathbf{w}, \beta, \boldsymbol{\alpha}|\mathbf{X})$ and parameters a_0, b_0, e_0, f_0
Output : A variational distribution $q(\mathbf{w}, \beta, \boldsymbol{\alpha}) = q(\mathbf{w})q(\beta)\prod_{d=1}^{D} q(\alpha_d)$
Initialization : $a'_0 = a_0, b'_0 = b_0$ and $\{e'_d = e_0, f'_d = f_0\}\forall d$

Repeat the following steps until convergence:

- Compute $\mathbb{E}[\beta], \mathbb{E}[\alpha_1], \ldots, \mathbb{E}[\alpha_D]$ as below:

$$\begin{aligned}\mathbb{E}[\beta] &= \frac{a'_0}{b'_0}, \\ \mathbb{E}[\alpha_d] &= \frac{e'_d}{f'_d} \quad \forall d \in [1..D]\end{aligned}$$

- Update $q(\mathbf{w})$ by computing $\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N$ using the above expectation values in formula [5]
- Compute $\mathbb{E}[\mathbf{w}^T], \mathbb{E}[\mathbf{w}\mathbf{w}^T]$ using updated value of $\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N$ as:

$$\begin{aligned}\mathbb{E}[\mathbf{w}^T] &= \boldsymbol{\mu}_N^T, \\ \mathbb{E}[\mathbf{w}\mathbf{w}^T] &= \boldsymbol{\Sigma}_N + \boldsymbol{\mu}_N\boldsymbol{\mu}_N^T\end{aligned}$$

- Update $q(\beta)$ by computing the values of a'_0, b'_0 using formula [6].
- Compute $\mathbb{E}[w_d^2]$ using using updated value of $\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N$ as:

$$\mathbb{E}[w_d^2] = \boldsymbol{\Sigma}_{N_{dd}} + \boldsymbol{\mu}_{N_d}^2 \quad \forall d \in [1..D]$$

- Update $q(\alpha_d)$ by computing e'_d, f'_d for each d using [7]
- Go back to step 1 if not converged

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 3

Student Name: Nidhi Hegde
Roll Number: 180472
Date: May 9, 2021

We first write down the joint distribution (Using Chain rule and i.i.d. property of x_n's as) $p(\mathbf{x}, \lambda_1 \ldots, \lambda_N, \alpha, \beta)$ as:

$$\begin{aligned}
p(\mathbf{x}, \lambda_1 \ldots, \lambda_N, \alpha, \beta) &= \left(\prod_{n=1}^{N} p(x_n|\lambda_n) p(\lambda_n|\alpha, \beta) \right) p(\alpha|a, b) p(\beta|c, d) \\
&= \left(\prod_{n=1}^{N} Poisson(x_n|\lambda_n) Gamma(\lambda_n|\alpha, \beta) \right) Gamma(\alpha|a, b) Gamma(\beta|c, d) \\
&= \left(\prod_{n=1}^{N} \frac{\lambda_n^{x_n} \exp(-\lambda_n)}{x_n!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_n^{\alpha-1} \exp(-\beta\lambda_n) \right) \frac{b^a}{\Gamma(a)} \alpha^{a-1} \exp(-b\alpha) \frac{d^c}{\Gamma(c)} \beta^{c-1} \exp(-d\beta)
\end{aligned}$$

Separating out the terms that contain the variables we need conditional posteriors on, we get:
For $\boldsymbol{\lambda}$
The Markov Blanket of λ_n includes α, β, x_n.
Due to Poisson-Gamma conjugacy, we get a closed form expression for the CP(same form as the prior) being another Gamma distribution with updated hyperparameters:

$$\begin{aligned}
p(\lambda_n|x_n, \alpha, \beta) &\propto Poisson(x_n|\lambda_n) Gamma(\lambda_n|\alpha, \beta) \\
&\propto \frac{\lambda_n^{x_n} \exp(-\lambda_n)}{x_n!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_n^{\alpha-1} \exp(-\beta\lambda_n) \\
&= Gamma(\lambda_n|\alpha + x_n, \beta + 1) \qquad \forall\ n \in [1...N]
\end{aligned}$$

For α:
The Markov Blanket of α includes $\lambda_1, \ldots, \lambda_n, a, b$. This CP is not available in a closed form.

$$\begin{aligned}
p(\alpha|\lambda_1, \ldots, \lambda_n, a, b) &\propto \left(\prod_{n=1}^{N} \frac{\lambda_n^{x_n} \exp(-\lambda_n)}{x_n!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_n^{\alpha-1} \exp(-\beta\lambda_n) \right) \frac{b^a}{\Gamma(a)} \alpha^{a-1} \exp(-b\alpha) \\
&\propto \frac{\beta^{N\alpha} \left(\prod_{n=1}^{N} \lambda_n \right)^{\alpha-1}}{\Gamma(\alpha)^N} \alpha^{a-1} \exp(-b\alpha)
\end{aligned}$$

For β:
The Markov Blanket of β includes $\lambda_1, \ldots, \lambda_n, c, d$.

$$\begin{aligned}
p(\beta|\lambda_1, \ldots, \lambda_n, c, d) &\propto \left(\prod_{n=1}^{N} \frac{\lambda_n^{x_n} \exp(-\lambda_n)}{x_n!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_n^{\alpha-1} \exp(-\beta\lambda_n) \right) \frac{d^c}{\Gamma(c)} \beta^{c-1} \exp(-d\beta) \\
&\propto \beta^{(N\alpha+c-1)} \exp\left(-\beta \sum_{n=1}^{N} \lambda_n - d\beta \right) \\
&\propto \beta^{(N\alpha+c-1)} \exp\left(-\beta \left(\sum_{n=1}^{N} \lambda_n + d \right) \right)
\end{aligned}$$

The form of CP for β is thus a gamma distribution, and hence available in closed form as:

$$p(\beta|\lambda_1, \ldots, \lambda_n, c, d) = Gamma(\beta \mid N\alpha + c,\ \sum_{n=1}^{N} \lambda_n + d)$$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 4

Student Name: Nidhi Hegde
Roll Number: 180472
Date: May 9, 2021

The PPD is given as:

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

We are also given with the following quantities:

$$\begin{aligned} p(r_{ij}|\mathbf{u}_i, \mathbf{v}_j) &= \mathcal{N}(\mathbf{u}_i^T\mathbf{v}_j, \beta^{-1}) \\ \implies \quad r_{ij} &= \mathbf{u}_i^T\mathbf{v}_j + \epsilon \\ Also, \quad \mathbb{E}[\epsilon] = 0,\ &var(\epsilon) = \beta^{-1} \end{aligned}$$

Since the samples $\{\mathbf{U}^{(s)}, \mathbf{V}^{(s)}\}_{s=1}^{S}$ generated by Gibb's sampler can be assumed to be drawn from the joint posterior $p(\mathbf{U}, \mathbf{J} \mid \mathbf{R})$, we can use these samples to obtain the mean and variance of r_{ij}.

Using Monte Carlo averaging, we obtain the sample based approximation for the following quantities which will be needed later for mean and variance computation:

$$\begin{aligned} \mathbb{E}[\mathbf{u}_i^T\mathbf{v}_j] &\approx \frac{1}{S}\sum_{s=1}^{S}\mathbf{u}_i^{(s)^T}\mathbf{v}_j^{(s)} \\ \mathbb{E}[(\mathbf{u}_i^T\mathbf{v}_j)^2] &\approx \frac{1}{S}\sum_{s=1}^{S}(\mathbf{u}_i^{(s)^T}\mathbf{v}_j^{(s)})^2 \end{aligned}$$

We also use the following identity.

$$var(x) = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \tag{8}$$

Now we calculate the mean and variance of r_{ij} using the approximated quantities above as:
For Mean:

$$\begin{aligned} \mathbb{E}[r_{ij}] &= \mathbb{E}[\mathbf{u}_i^T\mathbf{v}_j + \epsilon] && (Using\ r_{ij} = \mathbf{u}_i^T\mathbf{v}_j + \epsilon) \\ &= \mathbb{E}[\mathbf{u}_i^T\mathbf{v}_j] + \mathbb{E}[\epsilon] && (By\ Linearity\ of\ Expectation) \\ &= \frac{1}{S}\sum_{s=1}^{S}\mathbf{u}_i^{(s)^T}\mathbf{v}_j^{(s)} + 0 && (Using\ value\ of\ \mathbb{E}[(\mathbf{u}_i^T\mathbf{v}_j)]\ and\ \mathbb{E}[\epsilon]) \\ \mathbb{E}[r_{ij}] &= \frac{1}{S}\sum_{s=1}^{S}\mathbf{u}_i^{(s)^T}\mathbf{v}_j^{(s)} \end{aligned}$$

For Variance:

$$\begin{aligned}
var(r_{ij}) &= E[r_{ij}^2] - (E[r_{ij}])^2 && (Using\ [8]\ for\ r_{ij}) \\
E[r_{ij}^2] &= \mathbb{E}[(\mathbf{u}_i^T \mathbf{v}_j + \epsilon)^2] \\
&= \mathbb{E}[(\mathbf{u}_i^T \mathbf{v}_j)^2] + \mathbb{E}[(\epsilon)^2] + 2\mathbb{E}[\mathbf{u}_i^T \mathbf{v}_j]\mathbb{E}[\epsilon] && (By\ Linearity\ of\ Expectation) \\
&= \mathbb{E}[(\mathbf{u}_i^T \mathbf{v}_j)^2] + var(\epsilon) + (\mathbb{E}[\epsilon])^2 + 2\mathbb{E}[\mathbf{u}_i^T \mathbf{v}_j]\mathbb{E}[\epsilon] && (Using\ [8]\ for\ \epsilon) \\
&= \mathbb{E}[(\mathbf{u}_i^T \mathbf{v}_j)^2] + \beta^{-1} + 0 + 2 \cdot \mathbb{E}[\mathbf{u}_i^T \mathbf{v}_j] \cdot 0 && (Using\ \mathbb{E}[\epsilon] = 0,\ var(\epsilon) = \beta^{-1}) \\
\mathbb{E}[r_{ij}^2] &= \left(\frac{1}{S} \sum_{s=1}^{S} (\mathbf{u}_i^{(s)^T} \mathbf{v}_j^{(s)})^2 \right) + \beta^{-1} && (Using\ value\ of\ \mathbb{E}[(\mathbf{u}_i^T \mathbf{v}_j)^2]) \\
\therefore \quad var(r_{ij}) &= \frac{1}{S} \sum_{s=1}^{S} (\mathbf{u}_i^{(s)^T} \mathbf{v}_j^{(s)})^2 - \frac{1}{S^2} \left(\sum_{s=1}^{S} \mathbf{u}_i^{(s)^T} \mathbf{v}_j^{(s)} \right)^2 + \beta^{-1}
\end{aligned}$$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

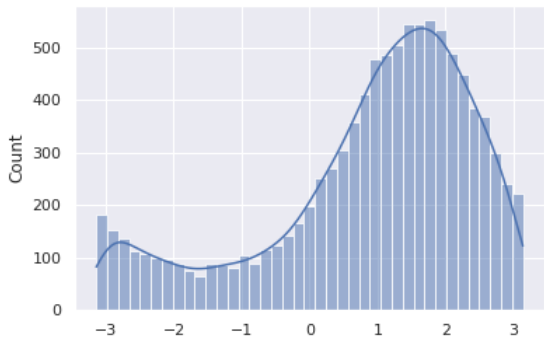
Student Name: Nidhi Hegde
Roll Number: 180472
Date: May 9, 2021

QUESTION 5

The optimal value for M can be obtained as:

$$
\begin{aligned}
&\tilde{p(x)} = \exp\left(sin(x)\right) \qquad \forall x \in [-\pi, \pi] \\
&q(x) = \mathcal{N}(0, \sigma^2) \\
&Mq(x) \geq \tilde{p(x)} \\
&\qquad \implies M \geq \frac{\sqrt{2\pi\sigma^2}\exp\left(sin(x)\right)}{\exp\left(\frac{-x^2}{2\sigma^2}\right)} \\
&\qquad \implies M \geq \sqrt{2\pi\sigma^2}\exp\left(sin(x) + \frac{x^2}{2\sigma^2}\right) \qquad \forall x \in [-\pi, \pi] \\
&\qquad \implies M = \max\left\{\sqrt{2\pi\sigma^2}\exp\left(sin(x) + \frac{x^2}{2\sigma^2}\right)\right\}, \qquad x \in [-\pi, \pi]
\end{aligned}
$$

On solving the above equation for the specified range of x, we get $M = 348.54$.
Using this value of M, and $\sigma^2 = 1$, the resulting histogram plot of samples obtained through Rejection Sampling is:



The code for the rejection sampler is present in the submitted notebook.