**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 3**

*Student Name:* Ritesh Kumar
*Roll Number:* 160575
*Date:* April 2, 2019

# QUESTION 1

Consider approximating an expectation $\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}$ using $S$ samples $\mathbf{z}^{(1)}, ..., \mathbf{z}^{(L)}$ drawn iid from $p(\mathbf{z})$. Denote the approximated expectation as $\hat{f} = \frac{1}{S}\sum_{s=1}^{S} f\left(\mathbf{z}^{(s)}\right)$

**To Show**: This approximation is unbiased, ie $\mathbb{E}[\hat{f}] = \mathbb{E}[f]$.

**Proof**:

We know that,

$$\hat{f} = \frac{1}{S}\sum_{s=1}^{S} f\left(\mathbf{z}^{(s)}\right)$$

So

$$\begin{aligned}
\mathbb{E}[\hat{f}] &= \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f\left(\mathbf{z}^{(s)}\right)\right] \\
&= \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}\left[f\left(\mathbf{z}^{(s)}\right)\right] \text{ by linearity of expectation} \\
&= \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}\left[f\right] \\
&= \mathbb{E}\left[f\right]
\end{aligned}$$

Therefore,

$$\mathbb{E}[\hat{f}] = \mathbb{E}\left[f\right]$$

**To Show**: The variance of this approximation is given by $var[\hat{f}] = \frac{1}{S}\mathbb{E}\left[(f - \mathbb{E}[f])^2\right]$.

**Proof**:

We know that,

$$
\begin{aligned}
var[\hat{f}] &= \mathbb{E}\left[\hat{f}^2\right] - \mathbb{E}\left[\hat{f}\right]^2 \\
&= \mathbb{E}\left[\left(\frac{1}{S}\sum_{s=1}^{S} f\left(\mathbf{z}^{(s)}\right)\right)^2\right] - \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f\left(\mathbf{z}^{(s)}\right)\right]^2 \\
&= \mathbb{E}\left[\frac{1}{S^2}\sum_{s=1}^{S} f^2\left(\mathbf{z}^{(s)}\right) + \frac{1}{S^2}\sum_{s=1}^{S}\sum_{k=1}^{S} f\left(\mathbf{z}^{(s)}\right) f\left(\mathbf{z}^{(k)}\right)\right] - \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f\left(\mathbf{z}^{(s)}\right)\right]^2 \\
&= \frac{1}{S^2}\sum_{s=1}^{S}\mathbb{E}\left[f^2\left(\mathbf{z}^{(s)}\right)\right] + \frac{1}{S^2}\sum_{s=1}^{S}\sum_{k=1}^{S}\mathbb{E}\left[f\left(\mathbf{z}^{(s)}\right) f\left(\mathbf{z}^{(k)}\right)\right] - \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f\left(\mathbf{z}^{(s)}\right)\right]^2 \\
&= \frac{1}{S^2}\sum_{s=1}^{S}\mathbb{E}\left[f^2\left(\mathbf{z}\right)\right] + \frac{1}{S^2}\sum_{s=1}^{S}\sum_{k=1}^{S}\mathbb{E}\left[f\left(\mathbf{z}\right)\right]^2 - \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f\left(\mathbf{z}\right)\right]^2 \\
&= \frac{1}{S^2}S\mathbb{E}\left[f^2\left(\mathbf{z}\right)\right] + \frac{1}{S^2}S(S-1)\mathbb{E}\left[f\left(\mathbf{z}\right)\right]^2 - \mathbb{E}\left[f\left(\mathbf{z}\right)\right]^2 \\
&= \frac{1}{S}\mathbb{E}\left[f^2\left(\mathbf{z}\right)\right] - \frac{1}{S}\mathbb{E}\left[f\left(\mathbf{z}\right)\right]^2 \\
&= \frac{1}{S}var[f]
\end{aligned}
$$

Therefore

$$
\begin{aligned}
var[\hat{f}] &= \frac{1}{S}var[f] \\
&= \frac{1}{S}\mathbb{E}\left[\left(f - \mathbb{E}[f]\right)^2\right]
\end{aligned}
$$

**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 3**

*Student Name:* Ritesh Kumar
*Roll Number:* 160575
*Date:* April 2, 2019

# QUESTION 2

Consider linear regression with likelihood defined by Student t distribution $p(y_n|\mathbf{x}_n, \mathbf{w}, \sigma^2, \nu) = \tau(y_n|\mathbf{w}^\top\mathbf{x}_n, \sigma^2, \nu)$ and a Gaussian prior on the weights $\mathbf{w}$ ie $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|0, \rho^2\mathbf{I}_D\right)$. A Student t likelihood is often better than a Gaussian likelihood since it models outliers better. Assume we are given $N$ training examples, $(\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ to infer $\mathbf{w}$.
Student t distribution can be expressed in the following "infinite mixture" form

$$\tau(y|\mu, \sigma^2, \nu) = \int \mathcal{N}(y|\mu, \sigma^2/z) Gamma\left(z|\frac{\nu}{2}, \frac{\nu}{2}\right) dz$$

This is called Gaussian scale mixture. In this augmented model, we consider the joint distribution of the output $y_n$ and the augmented variable $z_n$ eg. instead of $\tau(y|\mu, \sigma^2, \nu)$ we can consider $p(y, z|\mu, \sigma^2, \nu) = \mathcal{N}(y|\mu, \sigma^2/z) Gamma\left(z|\frac{\nu}{2}, \frac{\nu}{2}\right)$.

**Gibbs sampler**: Construct a Gibbs sampler for $p(\mathbf{w}, \mathbf{z}|\mathbf{X}, \mathbf{y})$. Derive the conditional posteriors of all the unknowns and clearly write down their expressions of their parameters. Assume all other unknown $(\sigma^2, \nu, \rho^2)$ to be known.
**Solution**:
Derive the joint probability

$$\begin{aligned}
p(\mathbf{y}, \mathbf{w}, \mathbf{z}|\mathbf{X}, \sigma^2, \nu, \rho^2) &= p(\mathbf{y}, \mathbf{z}|\mathbf{X}, \mathbf{w}, \sigma^2, \nu, \rho^2) p(\mathbf{w}|\mathbf{X}, \sigma^2, \nu, \rho^2) \\
&= p(\mathbf{w}|\mathbf{X}, \sigma^2, \nu, \rho^2) \prod_{n=1}^N p(y_n, z_n|\mathbf{x}_n, \mathbf{w}, \sigma^2, \nu, \rho^2) \\
&= \mathcal{N}\left(\mathbf{w}|0, \rho^2\mathbf{I}_D\right) \prod_{n=1}^N \mathcal{N}(y_n|\mathbf{w}^\top\mathbf{x}_n, \sigma^2/z) Gamma\left(z_n|\frac{\nu}{2}, \frac{\nu}{2}\right)
\end{aligned}$$

From above we can find CP(taking terms that contain the conditional variable). The CPs will be as follows

$$p(\mathbf{w}|\mathbf{z}, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{w}|0, \rho^2\mathbf{I}_D) \prod_{n=1}^N \mathcal{N}(y_n|\mathbf{w}^\top\mathbf{x}_n, \sigma^2/z)$$

Therefore,

$$\begin{aligned}
p(\mathbf{w}|\mathbf{z}, \mathbf{y}, \mathbf{X}) &= \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\sigma}) && \text{where} \\
\mathbf{\Sigma} &= \left(\frac{\mathbf{I}_D}{\rho^2} + \frac{\sum_{n=1}^N z_n\mathbf{x}_n\mathbf{x}_n^\top}{\sigma^2}\right)^{-1} && \text{and} \\
\boldsymbol{\mu} &= \left(\frac{\mathbf{I}_D}{\rho^2} + \frac{\sum_{n=1}^N z_n\mathbf{x}_n\mathbf{x}_n^\top}{\sigma^2}\right)^{-1} \left(\frac{\sum_{n=1}^N z_n\mathbf{x}_n\mathbf{y}_n}{\sigma^2}\right)
\end{aligned}$$

and

$$p(z_n|\mathbf{w}, \mathbf{z}_{-n}, \mathbf{y}, \mathbf{X}) \propto \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{w}^\top \mathbf{x}_n, \sigma^2/z) Gamma\left(z_n|\frac{\nu}{2}, \frac{\nu}{2}\right)$$

Therefore,

$$\begin{aligned} p(z_n|\mathbf{w}, \mathbf{z}_{-n}, \mathbf{y}, \mathbf{X}) &= Gamma\,(z_n|\alpha, \beta) && \text{where} \\ \alpha &= \frac{\nu+1}{2} && \text{and} \\ \beta &= \frac{\nu}{2} + \frac{(y_n - \mathbf{w}^\top \mathbf{x}_n)^2}{2\sigma^2} \end{aligned}$$

Gibbs Sampler

- Initialize $\mathbf{w}$. Set t = 1
- Sample $z_n^{(t)}$ from $Gamma(z_n|\alpha, \beta^{(t-1)})$ for $n = 1, \ldots, N$
- Sample $\mathbf{w}^{(t)}$ from $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$
- Go to step 2 if $t \neq T$

**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 3**

# QUESTION 3

*Student Name:* Ritesh Kumar
*Roll Number:* 160575
*Date:* April 2, 2019

Consider the Latent Dirichlet Allocation(LDA) model

$$\begin{aligned}
\phi_k &\sim Dirichlet(\eta, ..., \eta), \;\; k = 1, ..., K \\
\theta_d &\sim Dirichlet(\alpha, ..., \alpha), \;\; d = 1, ..., D \\
\mathbf{z}_{d,n} &\sim multinoulli(\theta_d), \;\; n = 1, ..., N_d \\
\mathbf{w}_{d,n} &\sim multinoulli(\phi_{\mathbf{z}_{d,n}})
\end{aligned}$$

In the above, $\phi_k$ denoted the $V$ dim. topic vector for topic $k$ (assuming vocabulary of $V$ unique words), $\theta_d$ denotes the $K$ dim. topic proportion vector for document $d$, and the number of words in document $d$ in $N_d$.

**Gibbs Sampler**: Derive a Gibbs sampler for the word-topic assignment variable $z_{dn}$.

**Solution**:

The CP is

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

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

Now

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

So

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

Therefore,

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

Now

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

Also

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

Therefore,

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

And

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

Therefore, finally we get

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

Normalize the above by summing numerator over all $k$ to obtain the required conditional probability.
The idea is that, the probability of the word $w_{dn}$ belonging to topic $k$ depends on the number of times the word $w_{dn}$ across the corpus belonged to topic $k$ (excluding the current occurrence), and the number of times the words across the document belonged to topic $k$ (excluding current occurrence). We are looking across the corpus for word $w_{dn}$ because it depends on topic vectors which are for the entire corpus. Whereas, $z_{dn}$ which is drawn from $\theta_d$ depends on the document $d$, so we look across the document $d$.
Sketch of Gibbs sampler:

- Initialize the latent variable matrix $\mathbf{Z} = \mathbf{Z}^{(0)}$ randomly. Note that for each $z_{dn}$ the possible values are 1 to $K$. Set t = 1.

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

- $t = t + 1$. Go to step 2 if $t \neq T$

NOTE

Using $S$ samples of $\mathbf{Z}$, we can compute the expected values of $\theta_d$ and $\phi_k$ applying Monte-Carlo approximation.

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

Therefore, $\mathbb{E}\left[\theta_{dk}\right]$ 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 $\mathbb{E}\left[\phi_{kv}\right]$ depends on the number of times the word $v$ belongs to topic $k$ in the entire corpus, and the number of words belonging to topic $k$ across the corpus, both wrt sample $\mathbf{Z}^{(s)}$

**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 3**

# QUESTION 4

*Student Name:* Ritesh Kumar
*Roll Number:* 160575
*Date:* April 2, 2019

---

Consider an $N \times M$ matrix $\mathbf{X}$ with each entry $X_{nm}$ a count value, modeled as

$$\begin{aligned} p(X_{nm}|\mathbf{u}_n, \mathbf{v}_m) &= Poisson(X_{nm}|\mathbf{u}_n \top \mathbf{v}_m) \\ p(u_{nk}|a, b) &= Gamma(u_{nk}|a, b) \\ p(v_{mk}|c, d) &= Gamma(v_{mk}|c, d) \end{aligned}$$

In the above, $\mathbf{u}_n \in \mathbb{R}_+^K, \mathbf{v}_m \in \mathbb{R}_+^K$, and the Gamma distribution is assumed to have the shape and rate parameterization. The above is essentially a gamma-Poisson matrix factorization model for count data.

**Useful Result**: Given $K$ independent Poisson rv's $x_1, ..., x_K$ st $x_k$ ~Poisson($\lambda_k$), their sum $x = \sum_{k=1}^K x_k$ is also Poisson distributed ie $x \sim$ Poisson($\lambda$) where $\lambda = \sum_{k=1}^K \lambda_k$. The converse is also true. Based on this, a count valued rv $x$ can be thought of as a sum of smaller count valued rv $x_1, ..., x_K$.

**Gibbs Sampler**: Derive a Gibbs sampler for the above model. Assume the hyperparameters $a, b, c, d$ to be known.

**Solution**:

We can $X_{nm}$ as $X_{nm} = \sum_{k=1}^K X_{nmk}$. Therefore, using the above result, we get that, $p(X_{nmk} = Poisson(X_{nmk}|u_{nk}v_{mk})$ Derive the joint probability

$$\begin{aligned} p(\mathbf{X}, \mathbf{U}, \mathbf{V}) &= p(\mathbf{X}|\mathbf{U}, \mathbf{V})p(\mathbf{U})p(\mathbf{V}) \\ &= \prod_{n=1}^M \prod_{m=1}^M \prod_{k=1}^K p(X_{nmk}|u_{nk}, v_{mk})p(u_{nk})p(v_{mk}) \\ &= \prod_{n=1}^M \prod_{m=1}^M \prod_{k=1}^K Poisson(X_{nmk}|u_{nk}v_{mk})Gamma(u_{nk}|a, b)Gamma(v_{mk}|c, d) \end{aligned}$$

Using the above expression the CP are as follows

$$\begin{aligned} p(u_{nk}|\mathbf{U}_{-nk}, \mathbf{V}, \mathbf{X}) &\propto Gamma(u_{nk}|a, b) \prod_{m=1}^M Poisson(X_{nmk}|u_{nk}v_{mk}) \\ &\propto u_{nk}^{a-1} \exp(-bu_{nk}) \prod_{m=1}^M (u_{nk}v_{mk})^{X_{nmk}} \exp(-u_{nk}v_{mk}) \\ &\propto u_{nk}^{\sum_{m=1}^M X_{nmk}+a-1} \exp\left(-u_{nk}\left(b + \sum_{m=1}^M v_{mk}\right)\right) \end{aligned}$$

Therefore

$$p(u_{nk}|\mathbf{U}_{-nk}, \mathbf{V}, \mathbf{X}) = Gamma\left(u_{nk}|a + \sum_{m=1}^M X_{nmk}, b + \sum_{m=1}^M v_{mk}\right) \quad \text{For } n = 1, ..., N \text{ and } k = 1, ..., K$$

Similarly

$$p(v_{mk}|\mathbf{V}_{-mk}, \mathbf{U}, \mathbf{X}) = Gamma\left(v_{mk}|a + \sum_{n=1}^{N} X_{nmk}, b + \sum_{n=1}^{N} v_{nk}\right) \quad \text{For } m = 1, ..., M \text{ and } k = 1, ..., K$$

Note that the $\mathbf{X}$ in above two conditional posteriors denotes the part of $X$ that is generated by the unknown whose posterior is being calculated in terms of latent counts $X_{nmk}$'s. Now we need a posterior over the latent counts as well. Due to property ii given in question we can write it as follows

$$p(X_{nm1}, \dots, X_{nmK}|\mathbf{X}, \mathbf{u}, \mathbf{v}) = multinomial\big(X_{nm}; \big[\frac{u_{n1}v_{m1}}{\mathbf{u}_n^\top \mathbf{v}_m}, \frac{u_{n2}v_{m2}}{\mathbf{u}_n^\top \mathbf{v}_m}, \dots, \frac{u_{nK}v_{mK}}{\mathbf{u}_n^\top \mathbf{v}_m}\big]\big)$$
$$n = 1, \dots, N \quad m = 1, \dots, M$$

The above expression, thought it is conditioned on complete $\mathbf{X}$, will only depend of data which is created by $X_{nm1}, \dots, X_{nmK}$ i.e only on $X_{nm}$

Gibbs Sampler

- Initialize $u_{nk}$ and $X_{nm}$ randomly, for $n = 1, ..., N$ $m = 1, ..., M$ $k = 1, ...K$. Set t = 1
- Sample $v_{mk}^{(t)}$ from $Gamma\left(v_{mk}|a + \sum_{n=1}^{N} X_{nmk}, b + \sum_{n=1}^{N} v_{nk}\right)$ for $m = 1, ..., M$ $k = 1, ..., K$
- Sample $u_{nk}^{(t)}$ from $Gamma\left(u_{nk}|a + \sum_{m=1}^{M} X_{nmk}, b + \sum_{m=1}^{M} v_{mk}\right)$ for $n = 1, ..., N$ $k = 1, ..., K$
- Sample $X_{nm}^{(t)}$ from $multinomial\big(X_{nm}; \big[\frac{u_{n1}v_{m1}}{\mathbf{u}_n^\top \mathbf{v}_m}, \frac{u_{n2}v_{m2}}{\mathbf{u}_n^\top \mathbf{v}_m}, \dots, \frac{u_{nK}v_{mK}}{\mathbf{u}_n^\top \mathbf{v}_m}\big]\big)$ for $n = 1, ..., N$ $m = 1, ..., M$
- Go to step 2 if $t \neq T$

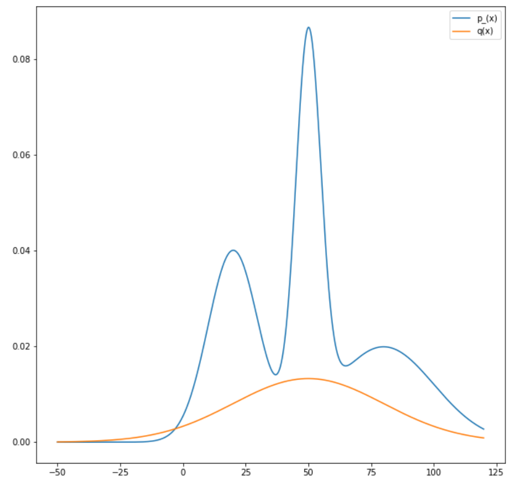
**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 3**

*Student Name:* Ritesh Kumar
*Roll Number:* 160575
*Date:* April 2, 2019

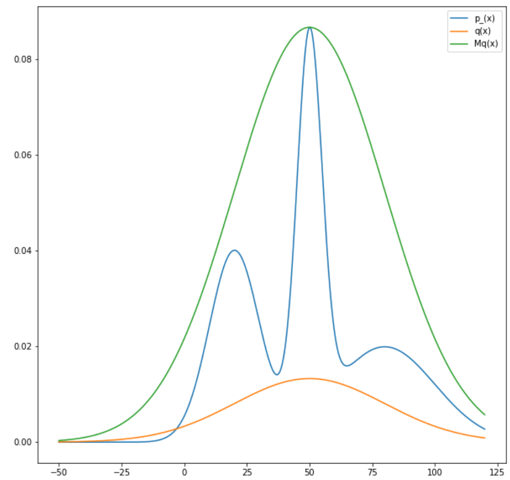
# QUESTION 5

**Part A**

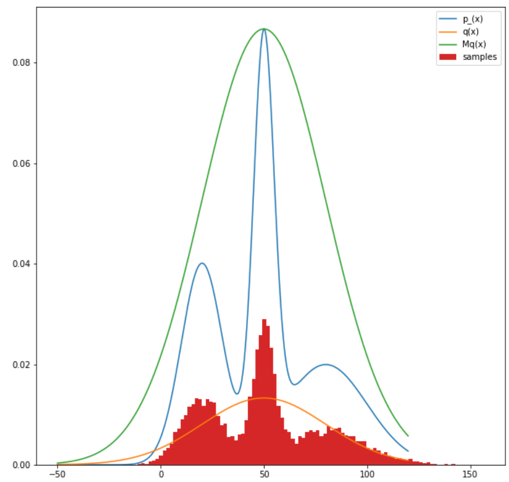
- Maximizing over a predefined range we get $M \approx 6.5$
- Acceptance Rate $\approx 0.454$
- $p(accept) = \frac{Z_p}{M}$. Using the above acceptance rate as probability we get $Z_p \approx 3$. So the plot of $p(z)$ and histogram become similar. Plot is attached



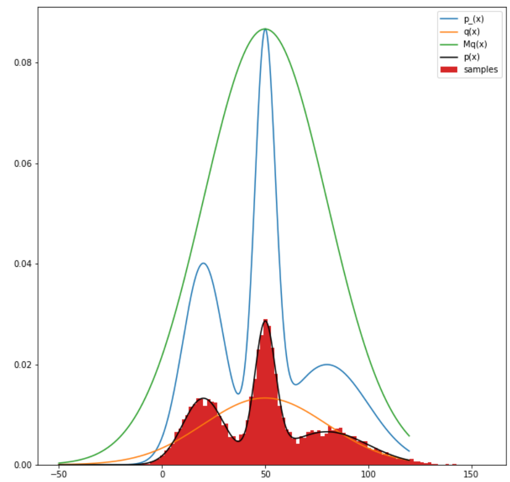

Plot for $\widetilde{p}(x)$ and $q(x)$



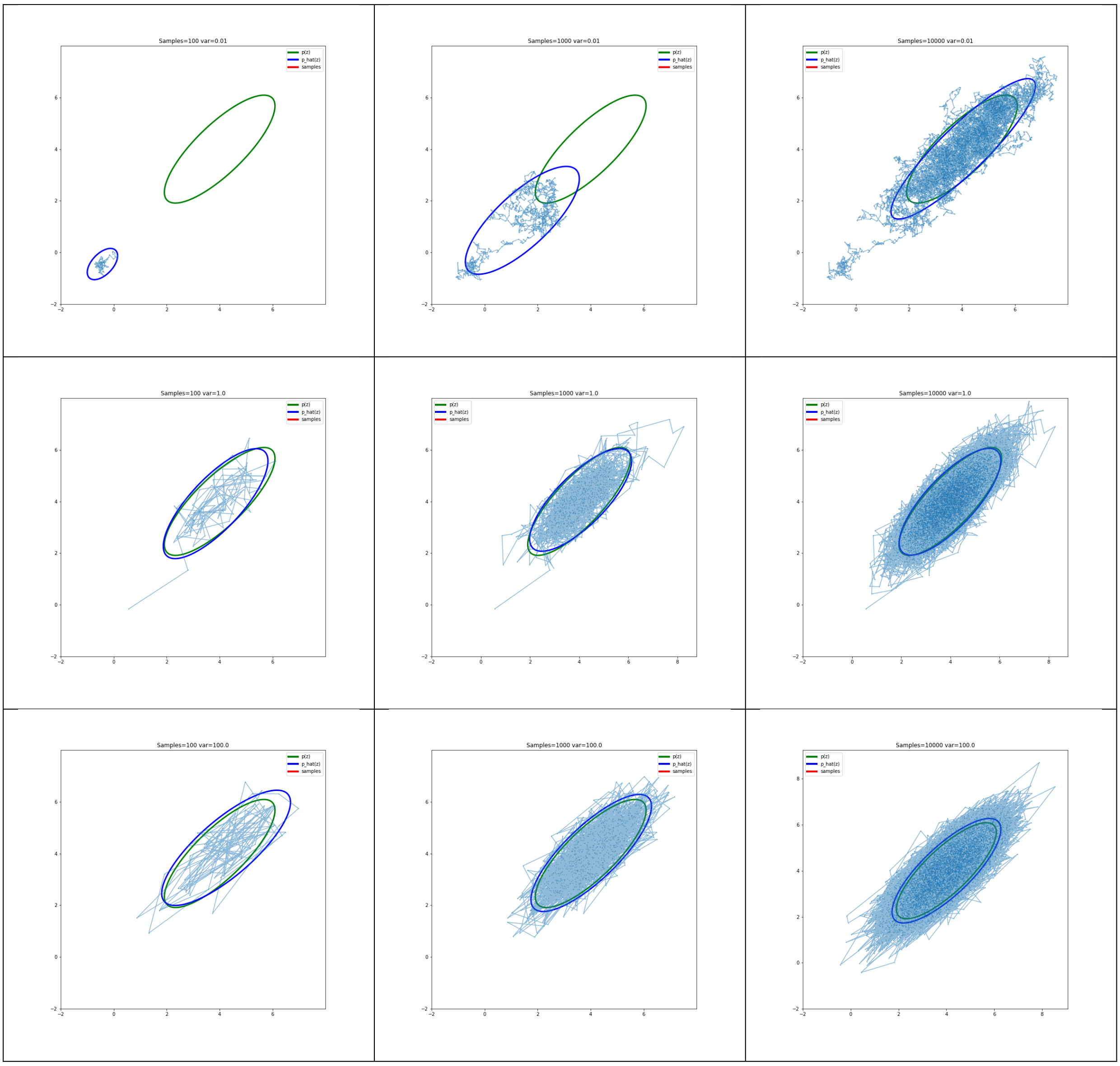

Plot for $\widetilde{p}(x)$, $q(x)$ and envelope $Mq(x)$




Plot for $\widetilde{p}(x)$, $q(x)$, envelope $Mq(x)$ and samples




Plot for $\widetilde{p}(x)$, $q(x)$, envelope $Mq(x)$, samples & distr.




### Part B

- The plots are submitted along with the code
- For $\sigma^2 = 0.01$; Rejection Rate $\approx 0.088$
- For $\sigma^2 = 1$; Rejection Rate $\approx 0.600$
- For $\sigma^2 = 100$; Rejection Rate $\approx 0.988$
- Given 10000 samples we can see from the plots that $\sigma^2 = 1$ works fairly well in terms of convergence. It is able to traverse the entire space relatively fast. We see that for $\sigma^2 = 0.01$ rejection rate is less but traversal of space is poor. For $\sigma^2 = 100$ traversal of space is good but convergence rate is very slow. It is for $\sigma^2 = 1$ that traversal is good and convergence rate is not bad either. Therefore, $\sigma^2 = 1$ is preferred choice.