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

Student Name: Atreya Goswami
Roll Number: 190201
Date: May 4, 2023

QUESTION 1

The general form of divergence, the $\alpha-$ divergence is given by

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

Let us represent $p(z)$ by p and $q(z)$ by q.

1. First, we show that $KL\left(p \parallel q\right)$ corresponds to $\alpha-$ divergence as $\alpha \to 1$. Let us take $\epsilon_1 = \frac{(1-\alpha)}{2}$. As $\alpha \to 1$, $\epsilon_1 \to 0$.

$$\begin{aligned}
D_\alpha\left(p \parallel q\right) &= \frac{4}{1-\alpha^2}\left(1 - \int \frac{p}{p^{(1-\alpha)/2}}\, q^{(1-\alpha)/2} dz\right) \\
&= \frac{1}{\epsilon_1\left(1-\epsilon_1\right)}\left(1 - \int p\, \frac{q^{\epsilon_1}}{p^{\epsilon_1}}\, dz\right) \\
&= \frac{1}{\epsilon_1\left(1-\epsilon_1\right)}\left\{-\int \left(p\, \frac{q^{\epsilon_1}}{p^{\epsilon_1}} - p\right) dz\right\} \qquad \left(\because \int p\, dz = 1\right) \\
&= \frac{1}{\epsilon_1\left(1-\epsilon_1\right)}\left\{-\int p\, \left(\frac{q^{\epsilon_1} - p^{\epsilon_1}}{p^{\epsilon_1}}\right) dz\right\} \\
&= \frac{1}{\epsilon_1\left(1-\epsilon_1\right)}\left[-\int p\, \left\{\frac{\left(1+\epsilon_1 \log q + O(\epsilon_1^2)\right) - \left(1+\epsilon_1 \log p + O(\epsilon_1^2)\right)}{1+\epsilon_1 \log p + O(\epsilon_1^2)}\right\} dz\right] \\
&= \frac{1}{\epsilon_1\left(1-\epsilon_1\right)}\left[-\int p\, \left\{\frac{\epsilon_1\left(\log q - \log p\right) + O(\epsilon_1^2)}{1+\epsilon_1 \log p + O(\epsilon_1^2)}\right\} dz\right] \\
&= \frac{1}{\left(1-\epsilon_1\right)}\left[-\int p\, \left\{\frac{\log q - \log p + O(\epsilon_1)}{1+\epsilon_1 \log p + O(\epsilon_1^2)}\right\} dz\right]
\end{aligned}$$

Taking limit $\epsilon_1 \to 0$

$$\begin{aligned}
&= -\int p\left(\log q - \log p\right) dz \\
&= KL\left(p \parallel q\right)
\end{aligned}$$

2. Now, we show that $KL\left(q \parallel p\right)$ corresponds to $\alpha-$ divergence as $\alpha \to -1$. Let us take $\epsilon_2 = \frac{(1+\alpha)}{2}$.

As $\alpha \to -1$, $\epsilon_2 \to 0$.

$$
\begin{aligned}
D_\alpha(p \,\|\, q) &= \frac{4}{1-\alpha^2}\left(1 - \int p^{(1+\alpha)/2}\,\frac{q}{q^{(1+\alpha)/2}}\,dz\right) \\
&= \frac{1}{\epsilon_2\,(1-\epsilon_2)}\left(1 - \int q\,\frac{p^{\epsilon_2}}{q^{\epsilon_2}}\,dz\right) \\
&= \frac{1}{\epsilon_2\,(1-\epsilon_2)}\left\{-\int\left(q\,\frac{q^{\epsilon_2}}{p^{\epsilon_2}} - q\right)dz\right\} \qquad \left(\because \int q\,dz = 1\right) \\
&= \frac{1}{\epsilon_2\,(1-\epsilon_2)}\left\{-\int q\,\left(\frac{p^{\epsilon_2} - q^{\epsilon_2}}{q^{\epsilon_2}}\right)dz\right\} \\
&= \frac{1}{\epsilon_2\,(1-\epsilon_2)}\left[-\int q\,\left\{\frac{\left(1+\epsilon_2\log p + O(\epsilon_2^2)\right) - \left(1+\epsilon_2\log q + O(\epsilon_2^2)\right)}{1+\epsilon_2\log q + O(\epsilon_2^2)}\right\}dz\right] \\
&= \frac{1}{\epsilon_2\,(1-\epsilon_2)}\left[-\int q\,\left\{\frac{\epsilon_2\,(\log p - \log q) + O(\epsilon_2^2)}{1+\epsilon_2\log q + O(\epsilon_2^2)}\right\}dz\right] \\
&= \frac{1}{(1-\epsilon_2)}\left[-\int q\,\left\{\frac{\log p - \log q + O(\epsilon_2)}{1+\epsilon_2\log q + O(\epsilon_2^2)}\right\}dz\right]
\end{aligned}
$$

Taking limit $\epsilon_2 \to 0$

$$
\begin{aligned}
&= -\int q\,(\log p - \log q)\,dz \\
&= KL\,(q \,\|\, p)
\end{aligned}
$$

QUESTION 2

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

Student Name: Atreya Goswami
Roll Number: 190201
Date: May 4, 2023

Given N scalar observations $\mathbf{X} = \{x_n\}_{n=1}^{N} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(x_n \mid \mu, \tau^{-1}\right)$

Likelihood: $p\left(\mathbf{X} \mid \mu, \tau\right) = \prod_{n=1}^{N} p\left(x_n \mid \mu, \tau\right) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \tau^{-1}\right)$

Prior on μ: $p\left(\mu\right) = \frac{1}{\sigma_\mu}$

Prior on τ: $p\left(\tau\right) = \frac{1}{\tau}$

We approximate the true joint posterior $p\left(\mu, \tau \mid \mathbf{X}\right)$ by a variational distribution $q\left(\mu, \tau\right)$. We use the mean-field assumption to factorise $q\left(\mu, \tau\right)$ as $q\left(\mu, \tau\right) = q_\mu\left(\mu\right) q_\tau\left(\tau\right)$. We will use the mean-field VI recipe as follows to compute the optimal factors $q_\mu^*\left(\mu\right)$ and $q_\tau^*\left(\tau\right)$.

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

For our model, the log-joint is given by

$$\begin{aligned}
\log p\left(\mathbf{X}, \mu, \tau\right) &= \log p\left(\mathbf{X} \mid \mu, \tau\right) + \log p\left(\mu\right) + \log p\left(\tau\right) \\
&= \sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \tau^{-1}\right) - \log \sigma_\mu - \log \tau
\end{aligned}$$

Now, we can write

$$\begin{aligned}
\log q_\tau^*\left(\tau\right) &= \mathbb{E}_{q_\mu}\left[\log p\left(\mathbf{X}, \mu, \tau\right)\right] + \text{const} \\
&= \mathbb{E}_{q_\mu}\left[\sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \tau^{-1}\right) - \log \sigma_\mu - \log \tau\right] + \text{const} \\
&= \mathbb{E}_{q_\mu}\left[\frac{N}{2} \log \tau - \frac{\tau}{2} \sum_{n=1}^{N} \left(x_n - \mu\right)^2 - \log \tau\right] + \text{terms independent of } \tau \\
&= \left(\frac{N}{2} - 1\right) \log \tau - \frac{\tau}{2} \mathbb{E}_{q_\mu}\left[\sum_{n=1}^{N} \left(x_n - \mu\right)^2\right] + \text{terms independent of } \tau
\end{aligned}$$

Hence,

$$\begin{aligned}
q_\tau^*\left(\tau\right) &\propto \tau^{\left(\frac{N}{2} - 1\right)} \exp\left[-\tau\, \mathbb{E}_{q_\mu}\left[\sum_{n=1}^{N} \frac{\left(x_n - \mu\right)^2}{2}\right]\right] \\
&= \text{Gamma}\left(\tau \,\middle|\, \frac{N}{2}, \mathbb{E}_{q_\mu}\left[\sum_{n=1}^{N} \frac{\left(x_n - \mu\right)^2}{2}\right]\right)
\end{aligned}$$

Also,

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

Hence,

$$\begin{aligned}
q_\mu^*(\mu) &\propto \exp\left[-\frac{N\mathbb{E}_{q_\tau}[\tau]}{2}\left(\mu - \frac{\sum_{n=1}^{N}x_n}{N}\right)^2\right] \\
&= \mathcal{N}\left(\mu \,\middle|\, \frac{\sum_{n=1}^{N}x_n}{N}\,,\,\left(N\mathbb{E}_{q_\tau}[\tau]\right)^{-1}\right)
\end{aligned}$$

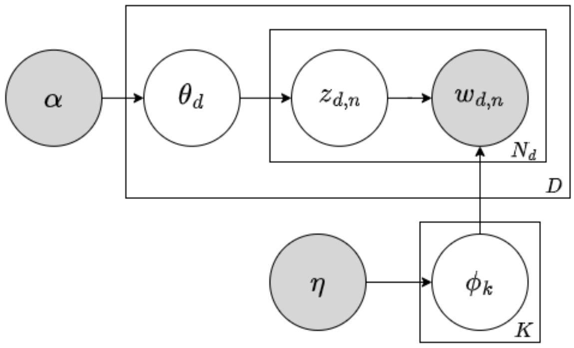
Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 3

Student Name: Atreya Goswami
Roll Number: 190201
Date: May 4, 2023

We are given the LDA model

$$\begin{aligned}
\phi_k &\sim \text{Dir}(\eta, \cdots, \eta), & k &= 1, \cdots, K \\
\theta_d &\sim \text{Dir}(\alpha, \cdots, \alpha), & d &= 1, \cdots, D \\
z_{d,n} &\sim \text{multinoulli}(\theta_d), & n &= 1, \cdots, N_d \\
w_{d,n} &\sim \text{multinoulli}\left(\phi_{z_{d,n}}\right)
\end{aligned}$$



where number of unique words in the vocabulary is V, ϕ_k denotes the V-dim. topic vector for topic k, θ_d denotes the K-dim. topic mixing proportion vector for document d, and the number of words in document d is N_d.
$z_{d,n}$ and $w_{d,n}$ are categorical random variables such that $z_{d,n} \in \{1, \cdots, K\}$ and $w_{d,n} \in \{1, \cdots, V\}$. The required CP for our Gibbs sampler is given by $p(z_{d,n} \mid \mathbf{Z}_{-d,n}, \mathbf{W})$, where $\mathbf{Z}_{-d,n}$ represents all the entries in $\mathbf{Z} = \left\{\{z_{d,n}\}_{n=1}^{N_d}\right\}_{d=1}^{D}$ except $z_{d,n}$, and $\mathbf{W} = \left\{\{w_{d,n}\}_{n=1}^{N_d}\right\}_{d=1}^{D}$. Now, we can write

$$\begin{aligned}
p(z_{d,n} = k \mid \mathbf{Z}_{-d,n}, \mathbf{W}) &= p(z_{d,n} = k \mid \mathbf{Z}_{-d,n}, w_{d,n} = v, \mathbf{W}_{-d,n}, \alpha, \eta) \\
&\propto p(w_{d,n} = v \mid z_{d,n} = k, \mathbf{Z}_{-d,n}, \mathbf{W}_{-d,n}, \eta) \times p(z_{d,n} = k \mid \mathbf{Z}_{-d,n}, \alpha)
\end{aligned}$$

First, we derive the expression for $p(z_{d,n} = k \mid \mathbf{Z}_{-d,n}, \alpha)$. Since only the $z_{d,n}$'s for a single document d are tied together by integrating out the prior on θ_d, hence $p(z_{d,n} = k \mid \mathbf{Z}_{-d,n}, \alpha)$ is same as $p(z_{d,n} = k \mid \mathbf{Z}_{d,-n}, \alpha)$, where $\mathbf{Z}_{d,-n}$ denotes all the entries in $\{z_{d,n}\}_{n=1}^{N_d}$ except $z_{d,n}$.

$$p(z_{d,n} = k \mid \mathbf{Z}_{-d,n}, \alpha) = p(z_{d,n} = k \mid \mathbf{Z}_{d,-n}, \alpha) = \int_{\theta_d} p(z_{d,n} = k \mid \theta_d)\, p(\theta_d \mid \mathbf{Z}_{d,-n}, \alpha)\, d\theta_d$$

where,

$$\begin{aligned}
p(z_{d,n} = k \mid \theta_d) &= \theta_{dk} \\
\text{and,} \quad p(\theta_d \mid \mathbf{Z}_{d,-n}, \alpha) &\propto p(\theta_d \mid \alpha) \times p(\mathbf{Z}_{d,-n} \mid \theta_d) \\
&\propto \text{Dir}(\alpha, \cdots, \alpha) \times \prod_{i=1, i\neq n}^{N_d} p(z_{d,i} \mid \theta_d) \\
&\propto \prod_{k=1}^{K} (\theta_{dk})^{\alpha-1} \times \prod_{i=1, i\neq n}^{N_d} \prod_{k=1}^{K} (\theta_{dk})^{\mathbb{I}[z_{d,i}=k]} \\
&\propto \prod_{k=1}^{K} (\theta_{dk})^{\alpha-1+\sum_{i=1,i\neq n}^{N_d} \mathbb{I}[z_{d,i}=k]}
\end{aligned}$$

Hence,

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

Therefore the collapsed prior probability can be written as,

$$\begin{aligned} p\left(z_{d,n} = k \,|\, \mathbf{Z}_{-d,n}, \alpha\right) &= \int_{\theta_d} \theta_{dk} \, \text{Dir}\left(\theta_d \,\middle|\, \left\{\alpha + \sum_{i=1, i\neq n}^{N_d} \mathbb{I}[z_{d,i} = k]\right\}_{k=1}^{K}\right) d\theta_d \\ &= \frac{\alpha + \sum_{i=1, i\neq n}^{N_d} \mathbb{I}[z_{d,i} = k]}{K\alpha + N_d - 1} \end{aligned} \tag{3.1}$$

Now, we derive the expression for collapsed likelihood factor $p\left(w_{d,n} = v \,|\, z_{d,n} = k, \mathbf{Z}_{-d,n}, \mathbf{W}_{-d,n}, \eta\right)$.

$$p\left(w_{d,n} = v \,|\, z_{d,n} = k, \mathbf{Z}_{-d,n}, \mathbf{W}_{-d,n}, \eta\right) = \int_{\phi_k} p\left(w_{d,n} = v \,|\, \phi_k\right) \, p\left(\phi_k \,|\, z_{d,n} = k, \mathbf{Z}_{-d,n}, \mathbf{W}_{-d,n}, \eta\right) d\phi_k$$

where,

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

Hence,

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

Therefor the collapsed likelihood factor can be written as

$$\begin{aligned} & p\left(w_{d,n} = v \,|\, z_{d,n} = k, \mathbf{Z}_{-d,n}, \mathbf{W}_{-d,n}, \eta\right) \\ = \quad & \int_{\phi_k} \phi_{kv} \text{Dir}\left(\phi_k \,\middle|\, \left\{\eta + \sum_{i=1}^{D} \sum_{j=1}^{N_d} \left(1 - \mathbb{I}[i = d] \cdot \mathbb{I}[j = n]\right) \mathbb{I}[z_{i,j} = k] \cdot \mathbb{I}[w_{i,j} = v]\right\}_{k=1}^{K}\right) d\phi_k \\ = \quad & \frac{\eta + \sum_{i=1}^{D} \sum_{j=1}^{N_d} \left(1 - \mathbb{I}[i = d] \cdot \mathbb{I}[j = n]\right) \mathbb{I}[z_{i,j} = k] \cdot \mathbb{I}[w_{i,j} = v]}{\eta V + \sum_{i=1}^{D} \sum_{j=1}^{N_d} \left(1 - \mathbb{I}[i = d] \cdot \mathbb{I}[j = n]\right) \mathbb{I}[z_{i,j} = k]} \end{aligned} \tag{3.2}$$

Hence the overall expression for conditional posterior probability of $z_{d,n}$ is given by

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

Now, we we can obtain the expression for $p\left(z_{d,n} = k \mid \mathbf{Z}_{-d,n}, \mathbf{W}\right)$ by normalising β_k over all values of $k = 1, 2, \cdots, K$. Hence the collapsed CP

$$
p\left(z_{d,n} \mid \mathbf{Z}_{-d,n}, \mathbf{W}\right) = \text{multinoulli}\left(\left\{\frac{\beta_k}{\sum_{k=1}^{K} \beta_k}\right\}_{k=1}^{K}\right)
$$

The expression for CP of $z_{d,n}$ makes intuitive sense as it is a multinoulli distribution where probability for each k denotes the probability of word $w_{d,n}$ of belonging to topic k. In order to calculate this probability, we consider the frequency of assignment of $w_{i,j}$ to topic k across the corpus (excluding $w_{d,n}$), as well as the frequency of assignment of $w_{d,i}$ to topic k in the current document d (excluding $w_{d,n}$).

Gibbs Sampling algorithm to sample from collapsed posterior $p\left(z_{d,n} \mid \mathbf{Z}_{-d,n}, \mathbf{W}\right)$

1. Initialise $\mathbf{Z}^{(0)}$ randomly with each entry $z_{i,j} \in \{1, 2, \cdots, K\}$, set $t = 1$
2. For word $n \in \{1, 2, \cdots, N_d\}$ in document $d \in \{1, 2, \cdots, D\}$

(a) Compute for $k = 1, 2, \cdots, K$,

$$
\beta_k^{(t)} = \frac{\eta + \sum_{i=1}^{D} \sum_{j=1}^{N_d} \left(1 - \mathbb{I}[i = d] \cdot \mathbb{I}[j = n]\right) \mathbb{I}[z_{i,j}^* = k] \cdot \mathbb{I}[w_{i,j} = v]}{\eta V + \sum_{i=1}^{D} \sum_{j=1}^{N_d} \left(1 - \mathbb{I}[i = d] \cdot \mathbb{I}[j = n]\right) \mathbb{I}[z_{i,j}^* = k]} \times \frac{\alpha + \sum_{i=1, i \neq n}^{N_d} \mathbb{I}[z_{d,i}^* = k]}{K\alpha + N_d - 1}
$$

where $z_{i,j}^*$ denotes the latest value of $z_{i,j}$.

(b) Compute for $k = 1, 2, \cdots, K$, normalised $\beta_k^{(t)} = \gamma_k^{(t)}$ (let) as

$$
\gamma_k^{(t)} = \frac{\beta_k^{(t)}}{\sum_{k=1}^{K} \beta_k^{(t)}}
$$

(c) Sample $z_{d,n}^{(t)} \sim \text{multinoulli}\left(\left\{\gamma_k^{(t)}\right\}_{k=1}^{K}\right)$

3. If $t \neq T$, set $t = t + 1$ and repeat from step 2.

We estimate the expected values $\mathbb{E}[\theta_d]$ and $\mathbb{E}[\phi_k]$ by Monte-Carlo summation using samples from the collapsed CP of **Z**.

$$
\begin{aligned}
\mathbb{E}[\theta_d] &\approx \frac{1}{S} \sum_{s=1}^{S} \frac{\alpha + \sum_{i=1, i \neq n}^{N_d} \mathbb{I}[z_{d,i}^{(s)} = k]}{K\alpha + N_d - 1} \\
\mathbb{E}[\phi_k] &\approx \frac{1}{S} \sum_{s=1}^{S} \frac{\eta + \sum_{i=1}^{D} \sum_{j=1}^{N_d} \left(1 - \mathbb{I}[i = d] \cdot \mathbb{I}[j = n]\right) \mathbb{I}[z_{i,j}^{(s)} = k] \cdot \mathbb{I}[w_{i,j} = v]}{\eta V + \sum_{i=1}^{D} \sum_{j=1}^{N_d} \left(1 - \mathbb{I}[i = d] \cdot \mathbb{I}[j = n]\right) \mathbb{I}[z_{i,j}^{(s)} = k]}
\end{aligned}
$$

Based on samples drawn $Z^{(s)}$, the expected value of θ_{dk} depends on the number of words assigned to topic k in document d, with some prior information incorporated in it. The topic assignment for a particular word is provided by the corresponding entry from samples of $Z^{(s)}$.
Also, the expected value of ϕ_{kv} is determined, using samples drawn $Z^{(s)}$, by the number of times that word v is assigned to topic k across the entire corpus, and the total number of words assigned to topic k in the corpus, along with some prior information incorporated in it. .

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

QUESTION 4

Student Name: Atreya Goswami
Roll Number: 190201
Date: May 4, 2023

We are given a matrix factorization model $p\left(r_{ij} \,|\, \boldsymbol{u}_i, \boldsymbol{v}_j\right) = \mathcal{N}\left(r_{ij} \,|\, \boldsymbol{u}_i^\top \boldsymbol{v}_j, \beta^{-1}\right)$ for a partially observed $N \times M$ matrix $\mathbf{R}$, where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ denote the latent factors of i^{th} row and j^{th} column of $\mathbf{R}$ respectively. The PPD of each r_{ij} is given by

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

which is intractable in general. We are also given a set of S samples $\left\{\mathbf{U}^{(s)}, \mathbf{V}^{(s)}\right\}_{s=1}^{S}$ generated by a Gibbs sampler for the posterior $p\left(\mathbf{U}, \mathbf{V} \,|\, \mathbf{R}\right)$, where $\mathbf{U}^{(s)} = \left\{\boldsymbol{u}_i^{(s)}\right\}_{i=1}^{N}$ and $\mathbf{V}^{(s)} = \left\{\boldsymbol{v}_j^{(s)}\right\}_{j=1}^{M}$.

First, we find the expectation of any term r_{ij} of the matrix $\mathbf{R}$, where the expectation is taken with respect to the PPD $p\left(r_{ij} \,|\, \mathbf{R}\right)$.

$$\begin{aligned}
\mathbb{E}\left[r_{ij}\right] &= \int r_{ij}\, p\left(r_{ij} \,|\, \mathbf{R}\right)\, dr_{ij} \\
&= \int\int\int r_{ij}\, p\left(r_{ij} \,|\, \boldsymbol{u}_i, \boldsymbol{v}_j\right)\, p\left(\boldsymbol{u}_i, \boldsymbol{v}_j \,|\, \mathbf{R}\right)\, d\boldsymbol{u}_i\, d\boldsymbol{v}_j\, dr_{ij}
\end{aligned}$$

We use the samples $\{\boldsymbol{u}_i^{(s)}, \boldsymbol{v}_j^{(s)}\}$ for Monte-Carlo approximation of the expectation.

$$\begin{aligned}
\mathbb{E}\left[r_{ij}\right] &\approx \frac{1}{S}\sum_{s=1}^{S} \int r_{ij}\, p\left(r_{ij} \,|\, \boldsymbol{u}_i^{(s)}, \boldsymbol{v}_j^{(s)}\right) dr_{ij} \\
&= \frac{1}{S}\sum_{s=1}^{S} \int r_{ij}\, \mathcal{N}\left(r_{ij} \,|\, \boldsymbol{u}_i^{(s)\top} \boldsymbol{v}_j^{(s)}, \beta^{-1}\right) dr_{ij} \\
&= \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{u}_i^{(s)\top} \boldsymbol{v}_j^{(s)}
\end{aligned}$$

Now, we find the variance of r_{ij}, given by

$$\text{var}\left[r_{ij}\right] = \mathbb{E}\left[r_{ij}^2\right] - \mathbb{E}\left[r_{ij}\right]^2$$

where,

$$
\begin{aligned}
\mathbb{E}\left[r_{ij}^2\right] &= \int r_{ij}^2\, p\left(r_{ij} \,|\, \mathbf{R}\right)\, dr_{ij} \\
&\approx \frac{1}{S}\sum_{s=1}^{S}\int r_{ij}^2\, p\left(r_{ij} \,|\, \boldsymbol{u}_i^{(s)}, \boldsymbol{v}_j^{(s)}\right) dr_{ij} \\
&= \frac{1}{S}\sum_{s=1}^{S}\int r_{ij}^2\, \mathcal{N}\left(r_{ij} \,|\, \boldsymbol{u}_i^{(s)\top}\boldsymbol{v}_j^{(s)}, \beta^{-1}\right) dr_{ij} \\
&= \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{r_{ij}\sim\mathcal{N}\left(r_{ij} \,|\, \boldsymbol{u}_i^{(s)\top}\boldsymbol{v}_j^{(s)}, \beta^{-1}\right)}\left[r_{ij}^2\right] \\
&= \frac{1}{S}\sum_{s=1}^{S}\left[\mathbb{E}_{r_{ij}\sim\mathcal{N}\left(r_{ij} \,|\, \boldsymbol{u}_i^{(s)\top}\boldsymbol{v}_j^{(s)}, \beta^{-1}\right)}\left[r_{ij}\right]^2 + \mathrm{var}_{r_{ij}\sim\mathcal{N}\left(r_{ij} \,|\, \boldsymbol{u}_i^{(s)\top}\boldsymbol{v}_j^{(s)}, \beta^{-1}\right)}\left[r_{ij}\right]\right] \\
&= \beta^{-1} + \frac{1}{S}\sum_{s=1}^{S}\left(\boldsymbol{u}_i^{(s)\top}\boldsymbol{v}_j^{(s)}\right)^2
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\mathrm{var}\left[r_{ij}\right] &= \mathbb{E}\left[r_{ij}^2\right] - \mathbb{E}\left[r_{ij}\right]^2 \\
&= \beta^{-1} + \frac{1}{S}\sum_{s=1}^{S}\left(\boldsymbol{u}_i^{(s)\top}\boldsymbol{v}_j^{(s)}\right)^2 - \left(\frac{1}{S}\sum_{s=1}^{S}\boldsymbol{u}_i^{(s)\top}\boldsymbol{v}_j^{(s)}\right)^2
\end{aligned}
$$

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

Student Name: Atreya Goswami
Roll Number: 190201
Date: May 4, 2023

QUESTION

5

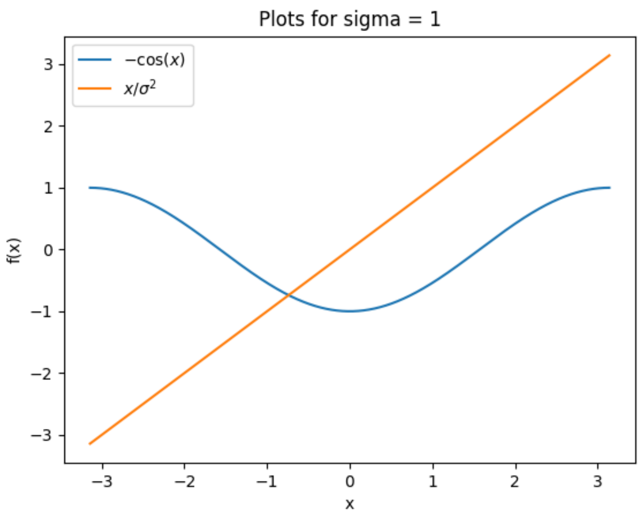
(a) Given $\tilde{p}(x) = \exp(\sin(x))$ for $x \in [-\pi, \pi]$, and proposal $q(x) = \mathcal{N}\left(0, \sigma^2\right)$, we need to choose M such that

$$
\begin{aligned}
Mq(x) &\geq \tilde{p}(x)\; \forall\, x \in [-\pi, \pi] \\
\text{or,} \qquad M &\geq \frac{\exp(\sin(x))}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{x^2}{2\sigma^2})} \\
\text{or,} \qquad M &\geq \sqrt{2\pi\sigma^2}\exp\left(\frac{x^2}{2\sigma^2} + \sin(x)\right)
\end{aligned}
$$

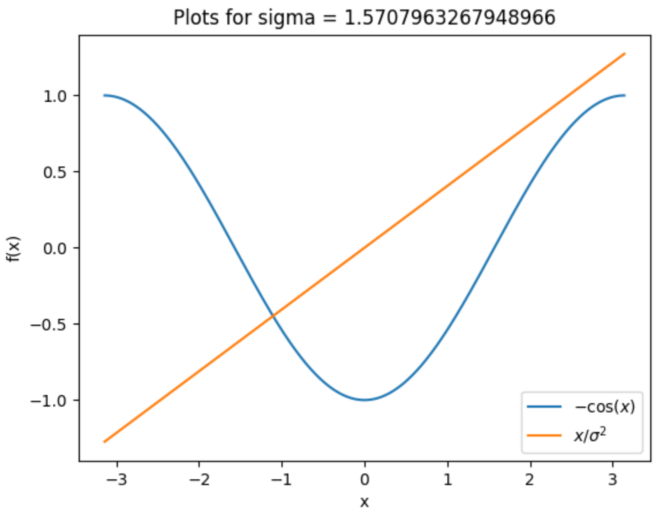
Taking $f(x) = \frac{x^2}{2\sigma^2} + \sin(x)$, we find its maximum by taking $f'(x_m) = 0$ and $f''(x_m) < 0$ at $x = x_m$.

$$
f'(x) = \frac{x_m}{\sigma^2} + \cos(x_m) = 0 \qquad \text{and} \qquad f''(x) = \frac{1}{\sigma^2} - \sin(x_m) < 0
$$

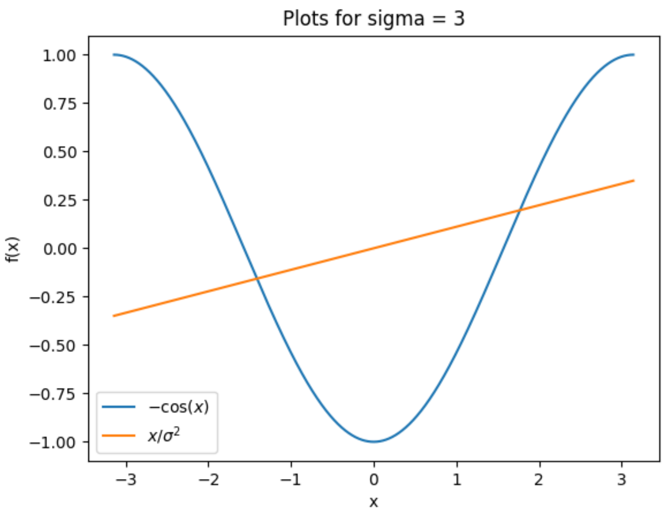
Now, we find approximate x_m graphically (by plotting $\frac{x}{\sigma^2}$ and $-\cos(x)$, and finding points of intersection) for 4 different values of $\sigma = 1, 3, \pi/2$ (since $\sim 95\%$ of the data-points in normal distribution lie within $\mu - 2\sigma$ and $\mu + 2\sigma$), and $\pi^2/2$.



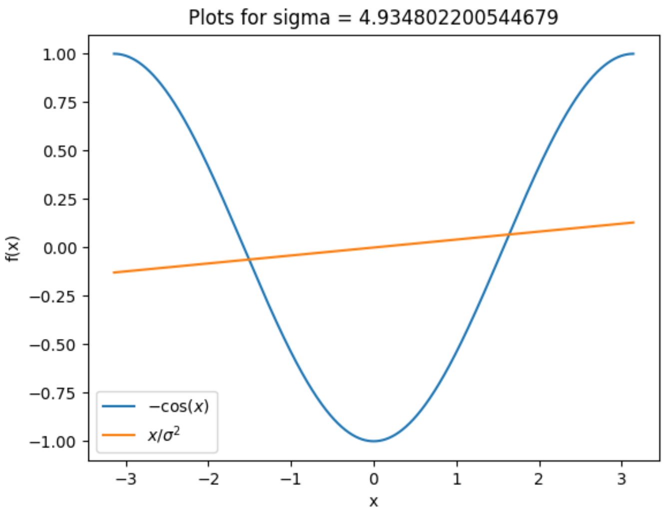
For $\sigma = 1,\ x_m = \pi$. Hence $M = 348.53$.



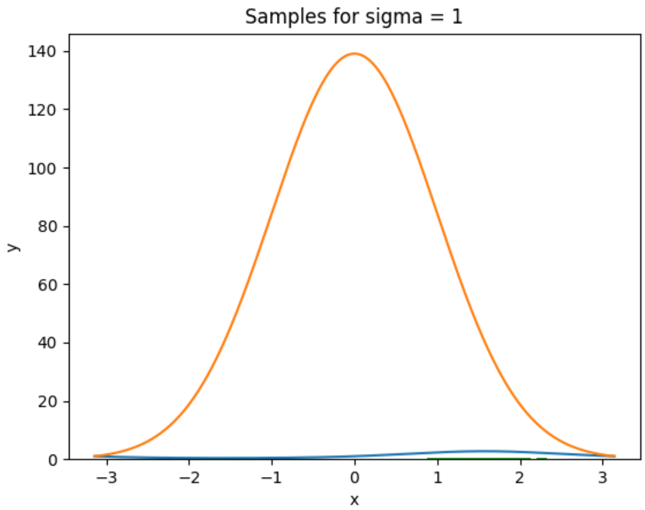
For $\sigma = \pi/2,\ x_m = \pi$. Hence $M = 29.09$.



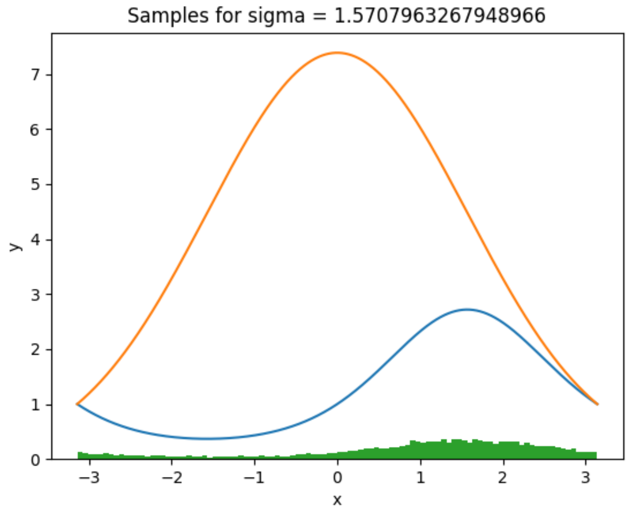
For $\sigma = 3,\ x_m = 1.77$. Hence $M = 23.85$.



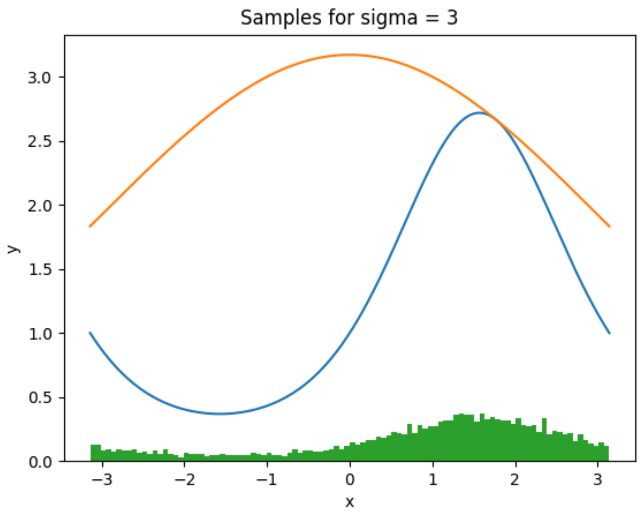
For $\sigma = \pi^2/2,\ x_m = 1.64$. Hence $M = 35.41$.



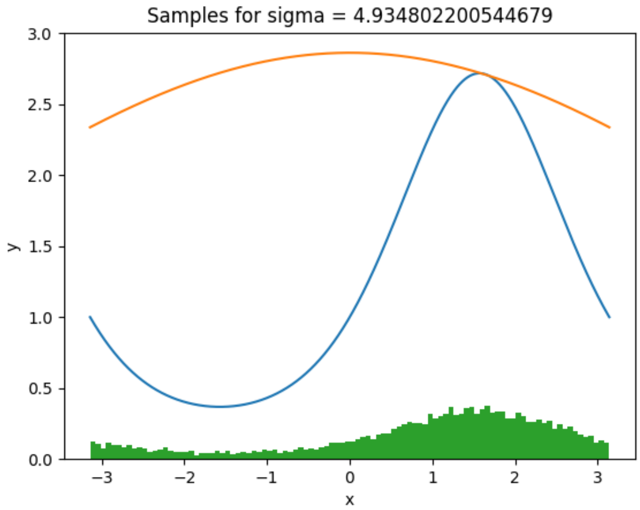
Rejection rate = 97.75%



Rejection rate = 72.40%



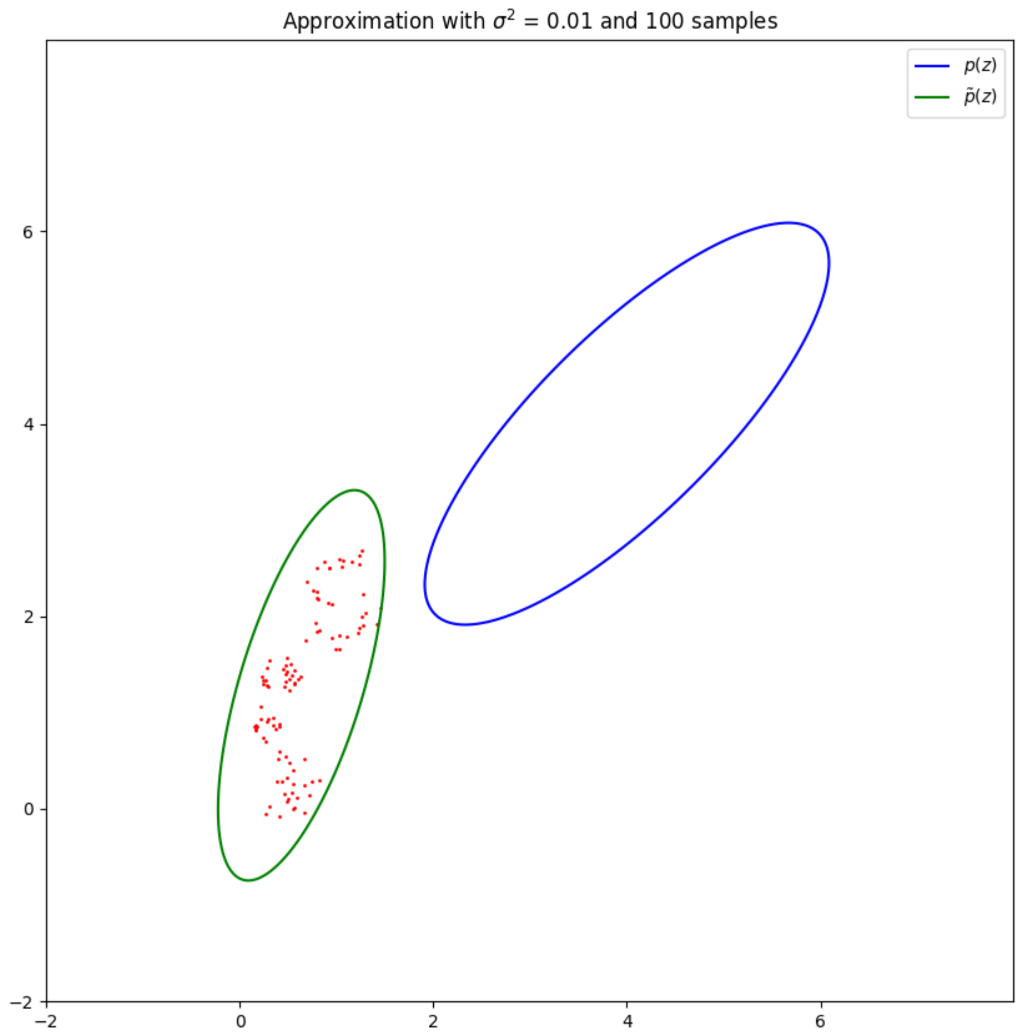
Rejection rate = 66.86%



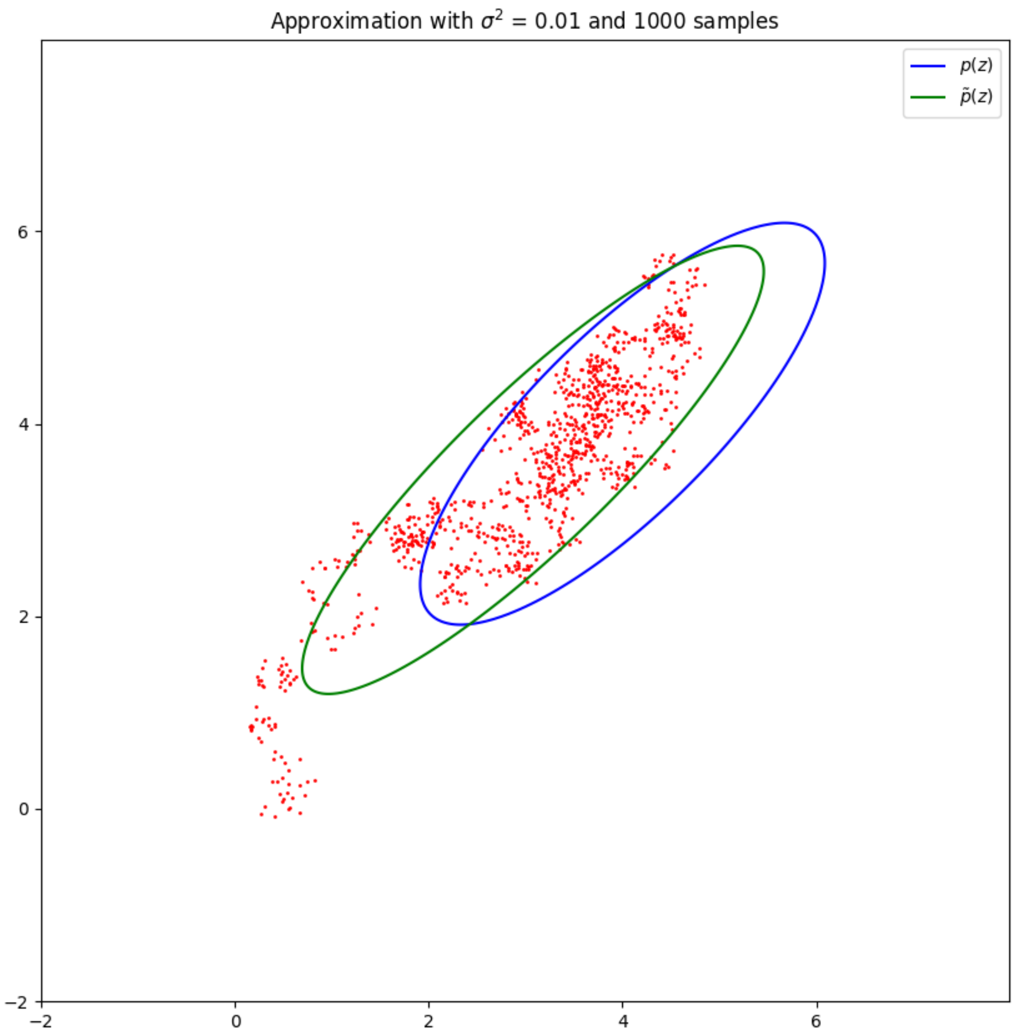
Rejection rate = 77.46%

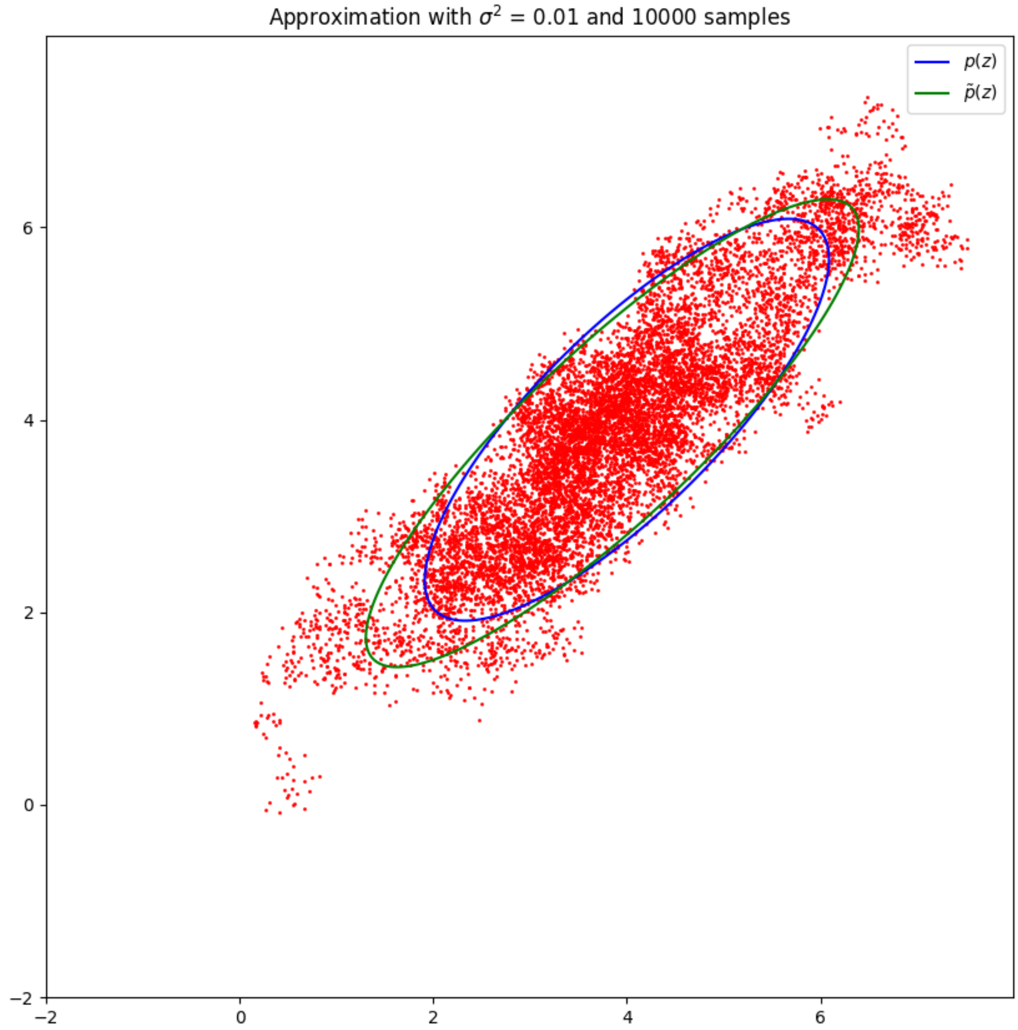
Lowest rejection rate (66.86%) is for $\sigma = 3$. Hence, we choose $\sigma^2 = 9$.

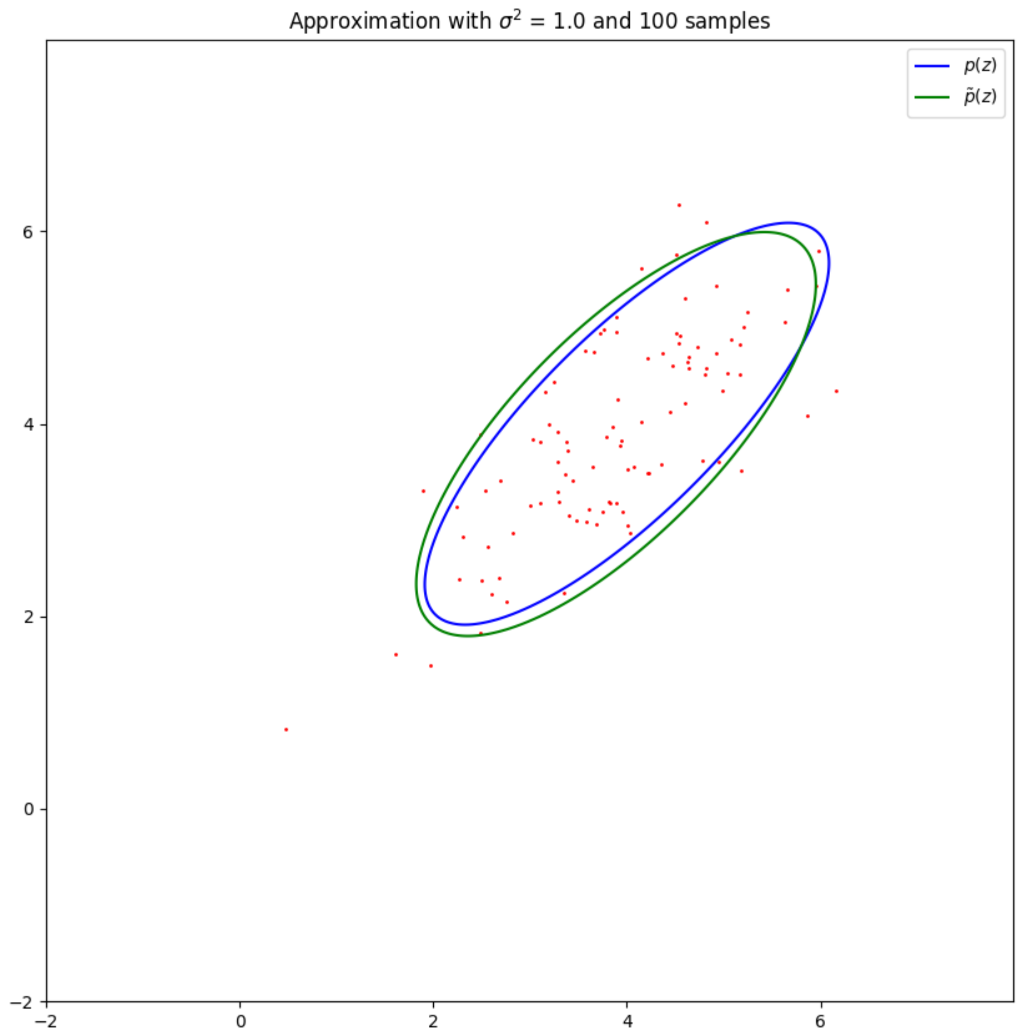
(b) The rejection rates are $8.16\%, 59.76\%$, and 98.84% for proposals with $\sigma^2 = 0.01, 1$, and 100 respectively.
The proposal with $\sigma^2 = 1.0$ seems to be the best choice, as the samples fit the true distribution well, along with having less rejection rate 59.76%.



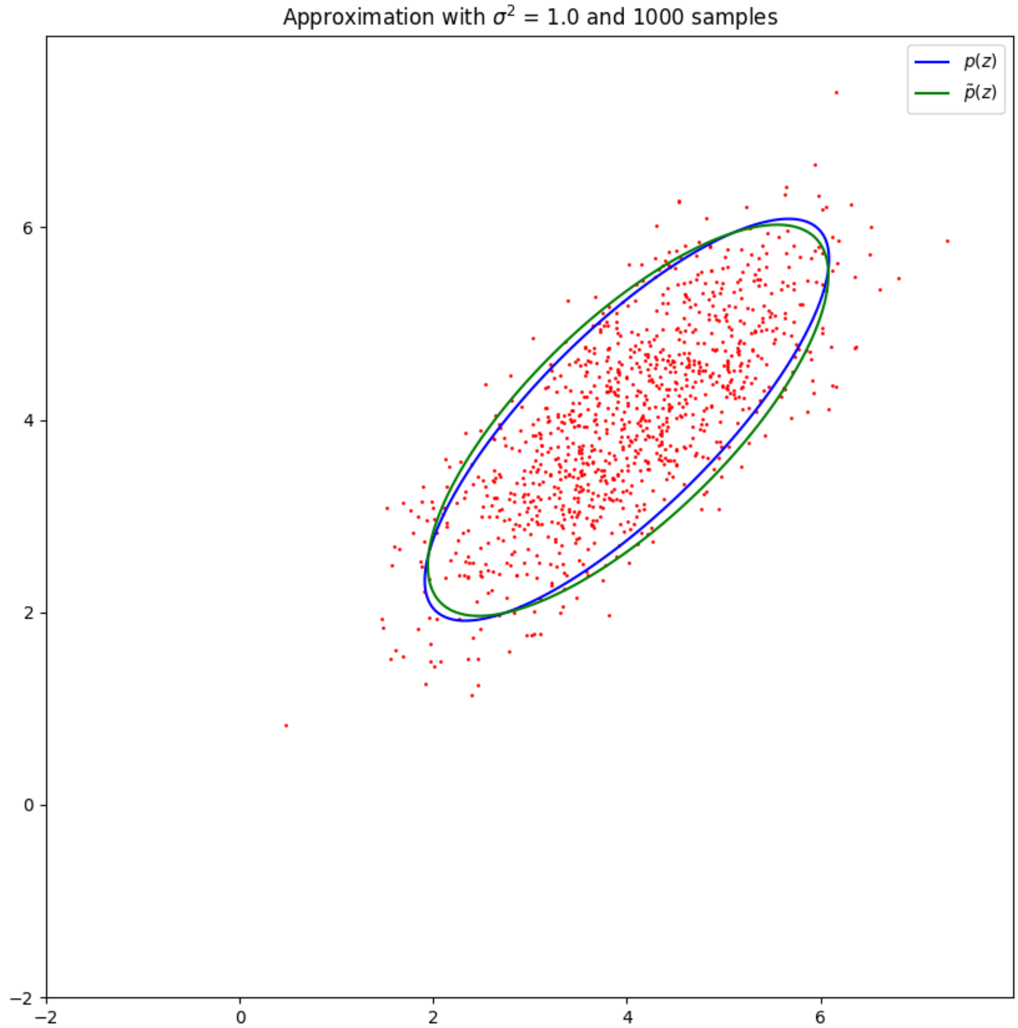
Approximation with σ^2 = 0.01 and 100 samples
$p(z)$
$\tilde{p}(z)$
6
4
2
0
−2
−2
0
2
4
6


Approximation with σ^2 = 0.01 and 1000 samples
$p(z)$
$\tilde{p}(z)$
6
4
2
0
−2
−2
0
2
4
6

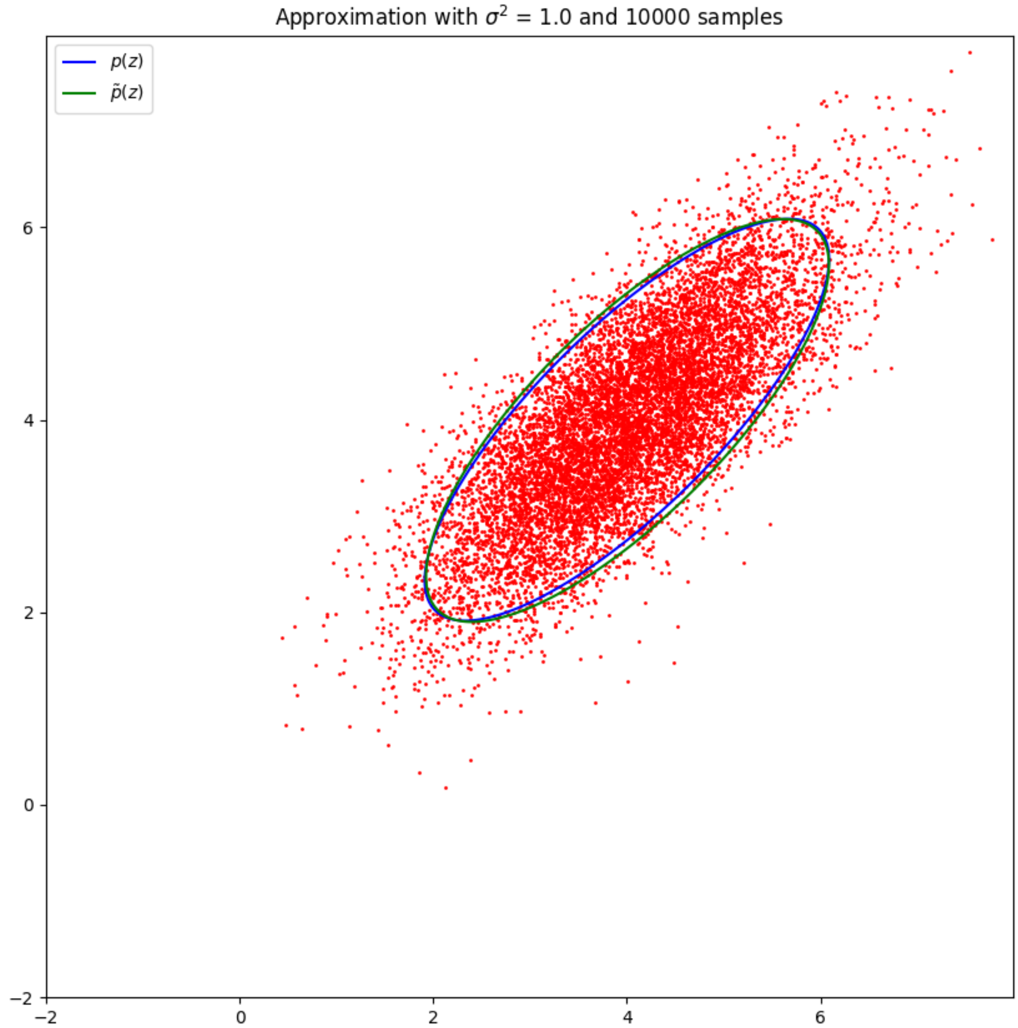

Approximation with σ² = 0.01 and 10000 samples
p(z)
p̃(z)
−2
0
2
4
6
−2
0
2
4
6

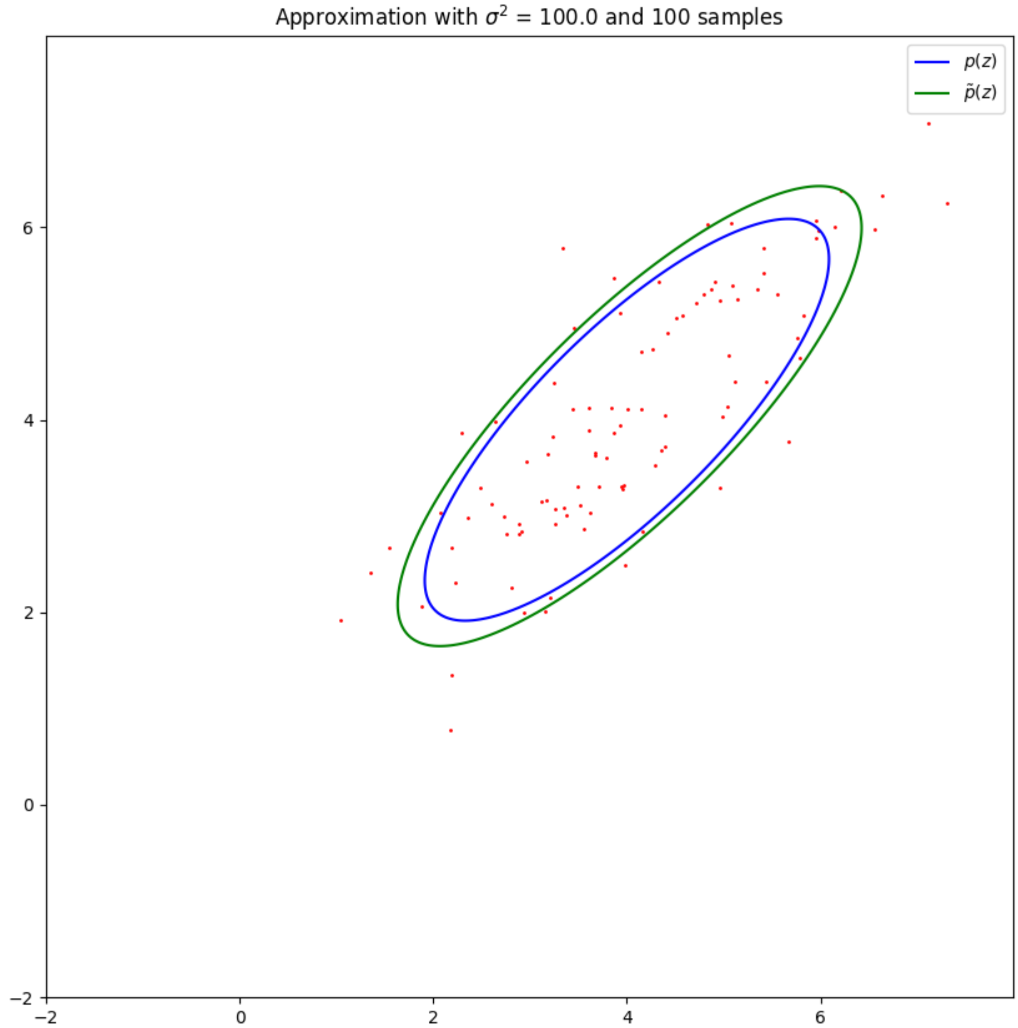


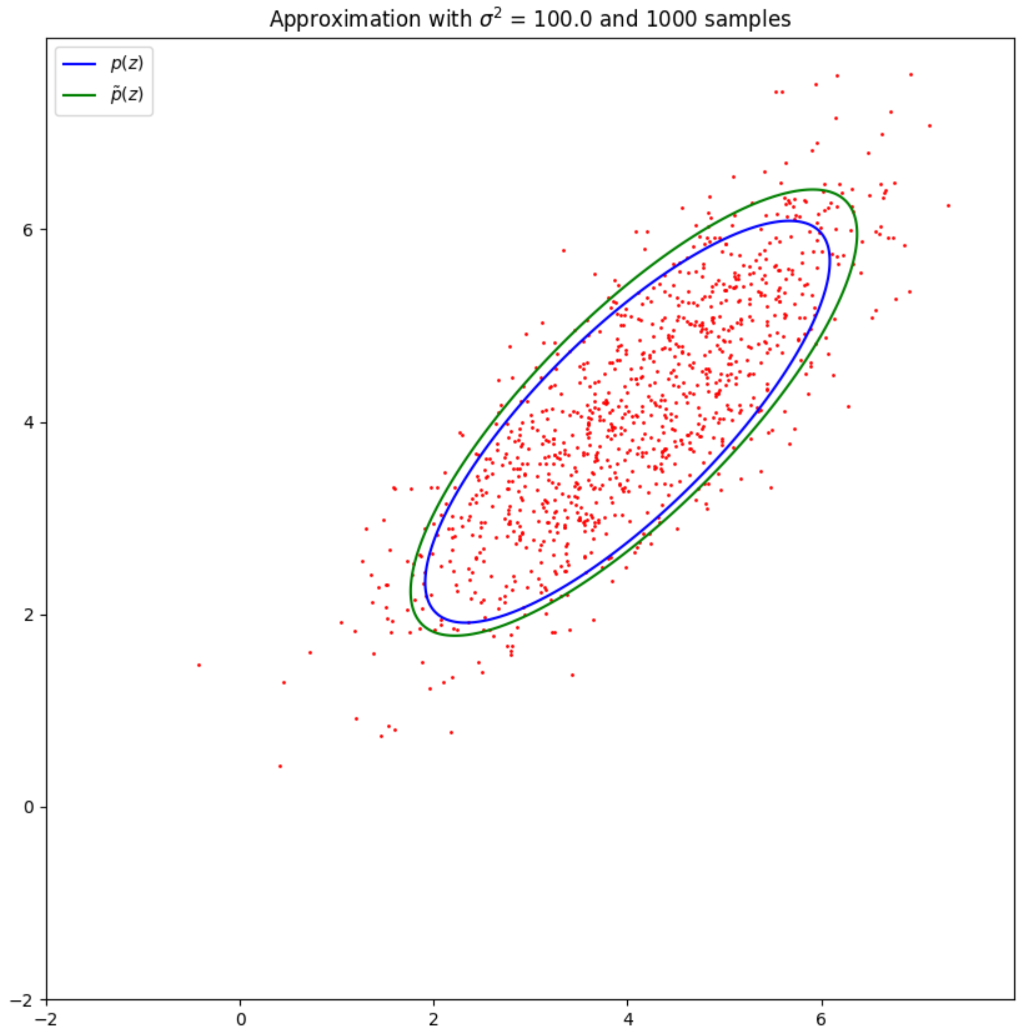
Approximation with σ^2 = 1.0 and 100 samples
$p(z)$
$\tilde{p}(z)$
6
4
2
0
−2
−2
0
2
4
6

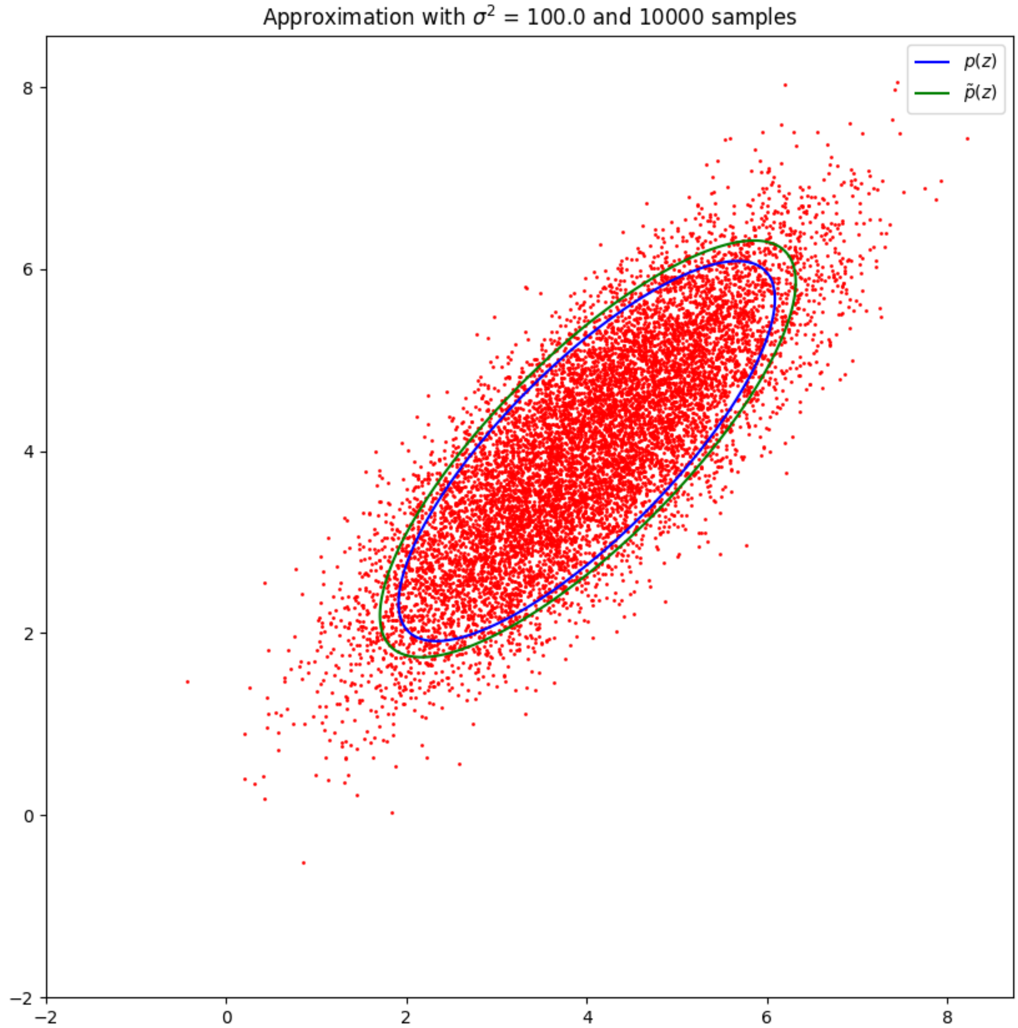


Approximation with σ^2 = 1.0 and 1000 samples
$p(z)$
$\tilde{p}(z)$
6
4
2
0
−2
−2
0
2
4
6


Approximation with σ^2 = 1.0 and 10000 samples
$p(z)$
$\tilde{p}(z)$
−2
0
2
4
6
−2
0
2
4
6


Approximation with σ^2 = 100.0 and 100 samples
$p(z)$
$\tilde{p}(z)$
−2
0
2
4
6


Approximation with σ^2 = 100.0 and 1000 samples
$p(z)$
$\tilde{p}(z)$
−2
0
2
4
6



Approximation with σ^2 = 100.0 and 10000 samples
$p(z)$
$\tilde{p}(z)$
8
6
4
2
0
−2
−2
0
2
4
6
8